

Datamining voor Informatie Gestuurde Politie

ir. R.C.P. van der Veer
Sentient

H.T. Roos MSc
Politie Amsterdam

A. van der Zanden MSc
Politie Amsterdam

Singel 160, 1015 AH, Amsterdam, Nederland
+31 (0)20 5 300 325

rvdveer@sentient.nl

SAMENVATTING

De voordelen van datamining voor politietoepassingen lijken groot, toch zijn er slechts enkele voorbeelden bekend. Dit artikel beschrijft de toepassingsproblemen van datamining bij de politie en introduceert een nieuwe benadering om deze problemen op te lossen. Dit in de vorm van een dataminingsysteem met als belangrijkste techniek het associatief geheugen. Deze techniek maakt dat het systeem makkelijk is in gebruik, weinig gegevensbewerking nodig heeft en vele soorten datatypes ondersteunt. Zo wordt de datavoorbereiding eenvoudiger en bevatten de resultaten uiteindelijk meer informatie. Een aantal politiekorpsen in Nederland gebruikt dit systeem al enkele jaren met een groep van inmiddels meer dan dertig gebruikers. De analyseprocessen binnen de politie zijn erg kennisintensief en vereisen veel domeinexpertise, wat het moeilijk maakt een politie dataminer te vinden die over voldoende politiekennis beschikt en die daarnaast ook technische vaardigheden heeft op het gebied van databases, statistiek en datamining. Ook kent het politiedomein datakwaliteitsproblemen en een zeer diverse behoefte aan informatie. Daarom probeert het systeemontwerp de noodzaak van technische vaardigheden zoveel mogelijk te beperken door gebruik te maken van één datawarehouse, van dataminingstechnieken die automatisch worden ingesteld en van een actieve gebruikersbegeleiding. Het gebruikersgemak is gewaarborgd doordat de vele tools en technieken uit de business intelligence, statistiek en datamining worden geïntegreerd in één interactieve omgeving die geen ingewikkeld ontwerp van een analyseproces op voorhand vereist. De analyse wordt interactief en stap voor stap uitgevoerd. Dit artikel bespreekt de voordelen van datamining voor de politie, het systeemontwerp, een aantal praktische toepassingen, praktijkervaringen en succesverhalen. Experimenten laten een efficiëntiewinst van een factor 20 zien en een factor 2 verbetering van voorspellingsnauwkeurigheid, een daling van 15% in criminaliteit en 50% meer daderherkenning.

Algemene termen

Algoritmes, management, business case, ontwerp, experimentatie, veiligheid, menselijke factoren.

Steekwoorden

Criminaliteit, misdaad, politie, datamining, voorspelling, GIS, hotspots, openbre veiligheid, analyse, business intelligence.

1. INLEIDING

In dit digitale tijdperk kunnen politiekorpsen over steeds meer gegevens beschikken. Gecombineerd met het dynamische karakter en de complexiteit van crimineel gedrag biedt dit mogelijkheden voor succesvolle dataminingtoepassingen. Echter, voorbeelden van consequent gebruikte toepassingen van politie-datamining zijn schaars. In dit artikel bespreken we de praktische

toepassing van een standaard politie-dataminingsysteem dat al in gebruik is bij een groeiend aantal Nederlandse politiekorpsen. Dit systeem is ontwikkeld door de gezamenlijke inspanning van datamining softwarebedrijf Sentient en de Nederlandse politiekorpsen Amsterdam-Amstelland, Midden- en West-Brabant en Brabant-Noord. Het is opgebouwd uit een geïntegreerd datamining softwarepakket DataDetective, dat wordt ontwikkeld en toegepast door Sentient sinds 1992, plus een uitgebreid datawarehouse, dat over data beschikt van verschillende politiestructuren en externe bronnen, zoals weergegevens, geografische gegevens en socio-demografische gegevens. Gedurende de toepassing in de afgelopen acht jaar is het dataminingsysteem voortdurend geëvalueerd, verbeterd en uitgebreid. De voorbeelden in dit artikel laten zien hoe datamining inmiddels een belangrijke rol speelt op strategisch, tactisch en operationeel niveau binnen de besluitvorming.

De belangrijkste sleutel tot het succes is het focussen op de eenvoud. Na een korte training kunnen geselecteerde politiemedewerkers snel patronen en trends ontdekken, voorspellingen maken, verbanden vinden met de mogelijke verklaringen daarvoor, criminele netwerken in kaart brengen en mogelijke daders herkennen. Deskundigheid op het gebied van statistiek of datamining is niet vereist. Daarnaast kan een grote groep medewerkers over dataminingresultaten beschikken door de verspreiding van wekelijkse rapporten met statistieken, voorspellingskaarten, clusters van criminaliteit, trends en lijsten met verdachten. Deze rapporten worden automatisch geproduceerd door het dataminingsysteem.

In hoofdstuk 2 brengen we de discussie op gang waarom het voor de politie belangrijk is om datamining toe te passen, gevolgd door een beschrijving van de tekortkomingen van de traditionele systemen in hoofdstuk 3. Hoofdstuk 4 omschrijft het systeem dat wij hebben ontwikkeld en hoe het gebruik ervan is georganiseerd. In hoofdstuk 5 bespreken we de verschillende toepassingen van het systeem, gevolgd door een aantal succesverhalen in hoofdstuk 6. We eindigen in hoofdstuk 7 met het bespreken van de toegevoegde waarde van datamining, de uitdagingen bij de uitvoering en de toekomstige werkzaamheden.

2. POLITIE:BEHOEFTE AAN DATAMINING

Het toenemend gebruik van het model van *Informatie Gestuurde Politie* [1] maakt analyse de kern van operationele, tactische en strategische besluitvorming. Volgens dit model stuurt informatie de operatie, in plaats van andersom. Daarom is het nu belangrijker dan ooit om te bepalen hoe datamining kan helpen bij het creëren van een beter inzicht en betere voorspellingen.

Traditioneel gezien richten politiestructuren zich op kleine delen van beschikbare gegevens (bijvoorbeeld jaar, maand, type misdrijf) voor een specifiek doel (bijvoorbeeld het strategisch in de gaten houden van criminaliteitscijfers).

Zonder datamining wordt de hoeveelheid data die in de analyse gebruikt wordt beperkt door de tijd die analisten nodig hebben om stap voor stap door de gegevens te gaan. Het is eenvoudigweg niet uitvoerbaar om alle gegevens die potentieel bruikbaar zijn met de hand te analyseren. Voor veel problemen is het belangrijk om zoveel mogelijk data te gebruiken om de politie in staat te stellen te begrijpen, te verklaren, verbanden te leggen en te voorspellen. De verklaring van een fenomeen (bijvoorbeeld de plotselinge groei van de activiteit van zakkenrollers) zit vaak in kleine details, bijvoorbeeld in het feit dat er in die periode veel straatfestivals hebben plaatsgevonden, waardoor er veel potentiële slachtoffers van zakkenrollers op straat waren. Dit laat zien dat bij het gebruik van meer data, patronen meer contextuele informatie laten zien en analisten worden geholpen de juiste conclusies te trekken. Dus om criminaliteit te begrijpen zijn er gegevens nodig die verder gaan dan enkel de simpele kenmerken van een gebeurtenis of persoon. Bijvoorbeeld omgevingstype, de Modus Operandi (MO of manier van werken), getuigenverklaringen, gestolen goederen, de gebruikte vervoersmiddelen en de achtergrond van de betrokkenen (geschiedenis, werkwijze, socio-demografisch profiel). Ook het leggen van verbanden tussen misdrijven op basis van overeenkomsten heeft om diezelfde redenen baat bij een grote hoeveelheid gegevens. Het leggen van verbanden kan hulp bieden bij het ontdekken van misdrijfreeksen en het koppelen en oplossen van zaken. Daarbij is datamining nodig. Geautomatiseerde patroonherkenning is noodzakelijk om de overdaad aan gegevens om te zetten in een hanteerbare informatiestroom.

Los van het omgaan met de hoeveelheid data, helpen dataminingstechnieken ook bij het omgaan met het dynamische karakter en de complexiteit van crimineel gedrag. Complexe patronen kunnen aanwezig zijn in een klein aantal gegevens. De volgende vraag is bijvoorbeeld moeilijk te beantwoorden met conventionele technieken: waar en wanneer vinden er gedurende de week misdrijven plaats, kijkend naar de X- en Y coördinaten van het incident, het tijdstip en de dag van de week. Een methode die zichzelf bewezen heeft om deze vraag te kunnen beantwoorden is het gebruik van slimme clusteringstechnieken (zie 5.1).

Het is een veelvoorkomend misverstand dat datamining grote hoeveelheden gegevens vereist om waarde toe te kunnen voegen, integendeel. Onze ervaring is dat wanneer datamining wordt toegepast het beste met één of twee gegevensbronnen kan worden begonnen om zo kennis te maken met de mogelijkheden en verwachtingen te managen. Veel organisaties zijn verrast over de grote hoeveelheid bruikbare resultaten die datamining levert op basis van een kleine hoeveelheid gegevens. Desalniettemin zijn meer gegevens altijd beter voor meer diepte en meer context.

Concluderend kan datamining de politie in staat stellen om criminaliteit beter te begrijpen en te voorspellen omdat vele gegevensbronnen geanalyseerd en complexe patronen gevonden kunnen worden. In 2004 concludeerde een uitvoerig onderzoek van het programmabureau van de Nederlandse Politie (ABRIO) dat 'door het toepassen van datamining het politiemangement doeltreffender en doelmatiger kan sturen op strategisch, tactisch en operationeel gebied' (intern rapport).

3. TEKORTKOMINGEN VAN POLITIE-ANALYSESYSTEMEN

Traditionele tools (verzamelingen technieken en hulpmiddelen) voor criminaliteitsanalyse kennen de volgende problemen:

1. Gebaseerd op een selectie van variabelen

Traditionele analysetools vragen de analist om één voor één naar de variabelen te kijken. Deze manier van werken is niet toe te passen op grote hoeveelheden gegevens.

2. Statische resultaten

Bestaande systemen produceren typisch statische rapporten die niet of nauwelijks interactie toelaten. Zij kunnen niet worden gebruikt om de verklaringen achter de geleverde cijfers te vinden.

3. Gebaseerd op simpele verbanden

Wanneer een analist zich focust op één of twee variabelen laten de traditionele tools alleen op die individuele variabelen een analyse toe. Zelden wordt de interactie tussen twee gekozen variabelen geanalyseerd, laat staan tussen meerdere variabelen. Hierdoor worden mogelijk nuttige verbanden over het hoofd gezien.

4. Complexe extractie

Het is doorgaans moeilijk om gegevens te extraheren uit bronsystemen van de politie vanwege de lang bestaande en uiteenlopende databasesystemen met gegevensmodellen waar transacties aan de basis liggen in plaats van analyses. Er bestaan verscheidene analytische toepassingen die werken met kleine extracties, maar wanneer de analyse een stap verder moet gaan, wordt de analist geconfronteerd met de uitdaging de juiste overige gegevens uit de systemen te halen. Het verschil in systemen in de organisatie en de matige datakwaliteit bemoeilijken deze uitdaging nog eens extra. Veelal vereist een analyse het uitvoeren van extractie, koppeling, correctie en voorbereiding van brongegevens. Om dit te adresseren heeft een klein aantal van de Nederlandse politiekorpsen datawarehouses geïmplementeerd.

5. Diversiteit aan tools

Analisten hebben de toegang tot een grote verscheidenheid aan tools, met elke tool een eigen focus. Er zijn tools voor geografisch visualiseren, statistisch analyseren, het maken van kaarten, definiëren van queries, rapporten maken, analyseren van criminele netwerken, visualiseren van criminaliteitcijfers, etc. Dit maakt dat de analisten al deze technieken moeten kennen en het kost tijd om data tussen de tools onderling uit te wisselen. Er is geen nauwsluitende koppeling die het mogelijk maakt dat analisten van techniek naar techniek kunnen gaan.

6. Tool complexiteit

De beschikbare tools voor statistische analyse vereisen een speciale training en soms een opleiding in wiskunde of statistiek. Vaak beschikken analisten niet over deze achtergrond.

Lost datamining deze problemen op? Het vorige hoofdstuk beredeneert dat datamining in principe veel kan toevoegen voor de politie. Dataminingtools lossen inderdaad sommige problemen die hierboven genoemd staan op, omdat ze niet vragen om de selectie van variabelen en door het vermogen om complexe patronen te vinden. Maar ze lijken andere problemen weer groter te maken door het toenemende aantal tools (weer een nieuw pakket), meer complexiteit (dataminingexpertise vereist) en problemen met de extractie (dataminingstechnieken stellen nieuwe eisen aan de gegevens). Het resultaat is dat gebruikers van politie-datamining aan hoge eisen moeten voldoen om effectief te kunnen zijn: zij moeten goed opgeleid zijn in IT, datamining, statistiek en over domeinexpertise beschikken. Met andere woorden; zij moeten beschikken over de kennis aangaande politiedatabases, het extraheren van data, data-voorbereiding, het gebruik van verschillende analytische

tools, het ontwerpen van een analyseproces, het selecteren van variabelen, het corrigeren van ontbrekende waarden, het uitkiezen van de juiste technieken, het instellen van de juiste parameters en kennis hebben van psychologie, criminelen, de samenleving en politiewerk. Daarbij wordt de gebruiker doorgaans ook nog uitgedaagd door de datakwaliteitsproblemen.

Wij geloven dat de hoge kwaliteitseisen voor gebruikers van standaard dataminingtools de belangrijkste reden vormen waarom er maar een paar succesvolle datamining toepassingen zijn ([2][3][4][5][6]) en de meeste van deze toepassingen zijn of academische experimenten of kleine toegepaste projecten - geen doorlopende activiteiten. Tevens blijkt dat wanneer de toepassingen wel blijvend zijn, ze slechts beperkt blijven tot een enkel politiekorps .

4. SYSTEEMOVERZICHT

Het vorige hoofdstuk beredeneert dat standaard dataminingtools moeilijk constant inzetbaar zijn te maken voor de politie. Met dit in gedachten is onze ontwerphilosofie voor het dataminingsysteem geweest om van de gebruikers alleen kennis van het domein en analytische vaardigheden te vragen. Meer vakkennis of vaardigheden moeten niet worden vereist. Het ontwikkelde systeem brengt verschillende technieken uit de business intelligence, statistiek, machine learning en geografie tezamen in een veelomvattende datamining-infrastructuur. Deze infrastructuur bevat een datawarehouse, een rapportagemodule plus een desktop-tool met functies voor eenvoudig queries maken, matching, datavisualisatie, statistieken, clustering, voorspellingsmodellen, verklaringsmodellen, netwerkanalyse, geografische profilering en uitgebreide geografische visualisatie.

De volgende paragrafen bespreken de kernonderdelen van het systeem, de inhoud van de database en de manier waarop het dataminingsysteem wordt gebruikt.

4.1 Kernonderdelen van het systeem

De tekortkomingen opgesomd in het voorafgaande hoofdstuk zijn gebruikt als wensen bij het ontwerpen van het dataminingsysteem. Het resulterende systeem beschikt over de volgende karakteristieken:

- **Kant en klare database**

De vereiste deskundigheid op het gebied van IT en databasekennis wordt beperkt door het leveren van een op zichzelf staande allesomvattende database waarin alle gegevens zijn samengevat, gekoppeld, opgeschoond en verrijkt. Het doel is dat de database 99% van de informatiebehoefte dekt. De overgebleven 1% heeft een ad hoc data-extractie en -voorbereiding. De uitgebreide database werkt als een *single point of truth*: alle analisten gebruiken dezelfde standaardgegevens en -definities.

- **Geautomatiseerde datamining**

De vereiste deskundigheid op het gebied van statistiek en datamining wordt beperkt door automatische selectie plus configuratie van dataminingtechnieken. Gebaseerd op de analysevraag en de gegevens kiest het systeem de juiste techniek en optimaliseert de parameters. Verder beschermt het systeem de gebruiker tegen typische valkuilen, zoals onbetrouwbare patronen en ontbrekende waarden. In sommige gevallen zou een volleerde dataminingexpert de geautomatiseerde selectie en configuratie misschien kunnen overtreffen. Dit is het compromis dat gemaakt moet worden om niet-experts in staat te stellen te dataminieren. Toch, wanneer dataminingexperts de tool gebruiken, besparen zij tijd en is hun kwaliteit constanter.

- **Gebruiksvriendelijke interface**

Een intuïtieve grafische gebruikersinterface wordt geboden met een opzet die zich richt op taken in plaats van technieken.

- **Interactieve analyse**

Het systeem werkt als een interactief analyse-instrument met de mogelijkheid om elk deel van de resultaten aan te klikken om in te zoomen. Op deze manier kan de gebruiker eenvoudig een analytische 'reis' maken zonder van tevoren te bedenken wat er moet gebeuren, zoals gevraagd wordt bij de op werkstream gebaseerde dataminingtools. Om dit intuïtieve en ad hoc proces te ondersteunen is visualisatie een belangrijk onderdeel van de gebruikersinterface. Deze interactieve mogelijkheden ondersteunen het ontdekkingsproces en bouwen op de creativiteit en de domeinkennis van de analist. Daaraan toegevoegd laten ze interactieve sessies toe met opdrachtgevers (bijvoorbeeld iemand die een onderzoek leidt). Door samenwerking op kritische momenten in het analyseproces, kan de werkelijke vraag worden verrijkt, kunnen nieuwe vragen direct worden beantwoord en kunnen resultaten worden geselecteerd op basis van relevantie.

- **Traceerbaarheid**

Het is belangrijk om op het spoor te blijven van de stappen die werden genomen om een analyseresultaat te behalen, ook al werkt de gebruiker interactief, vooral omdat die documentatie kan worden vereist in de rechtbank. Daarom houdt het systeem de geschiedenis van elk resultaat bij.

- **Dataflexibiliteit**

Associatief geheugen [7][8] wordt gebruikt als belangrijkste techniek bij voorspellen, clusteren en matchen in het dataminingsysteem. Dit betekent dat de inputgegevens worden gematched met een representatie van referentiegegevens (het geheugen) door middel van een vergelijkingsprincipe, zoals gebruikt in *Self Organizing Maps* [9], maar dan met de mogelijkheid om met veel meer datatypes om te gaan. Op deze manier wordt voorbereiding eenvoudiger en het datagebruik omvangrijker:

1. Er is bijna geen datavoorbereiding nodig omdat het associatief geheugen een breed scala aan datatypes kan hanteren, zoals symbolische gegevens, cyclische ordinalen (bijvoorbeeld dag van de week), lijsten, teksten en categorieën met veel waarden. Zolang er een manier is om gegevens te vergelijken kan een datatype worden gebruikt. Verder hoeven ontbrekende waarden niet te worden verwijderd of geraden omdat ontbrekende waarden simpelweg niet worden meegenomen bij het matchen.
2. Omdat de technologie een breed scala aan datatypes toelaat kan er bij het analyseproces veel meer informatie worden gebruikt dan bij andere technieken die strengere eisen stellen aan gegevens. Bij het gebruik van andere technieken is het werken met niet-standaard datatypes ofwel onmogelijk ofwel vereist het veel werk in de datavoorbereiding.
3. Het associatief geheugen is in staat om te verklaren hoe het tot een resultaat is gekomen door relevante zaken en personen uit het geheugen te presenteren, wat een intuïtieve manier van uitleggen is voor elke gebruiker.
4. Het zelflerende proces van associatieve geheugenmodellen komt goed en snel tot een resultaat vergeleken met bijvoorbeeld *back-propagation* neurale netwerken [8].
5. Associatieve geheugens zijn sterk opgewassen tegen suboptimale parameters (*graceful degradation*) en

daarom goed toe te passen in een situatie waar parameters automatisch worden bepaald en de gebruiker geen expert is in het optimaliseren van de techniek [8].

Met andere woorden: de gebruiker hoeft zich geen zorgen te maken over het nauwkeurig afstellen van parameters, sturen van het zelflerende proces, oplossen van ontbrekende waarden, selecteren van variabelen en decoderen van variabelen in een bruikbare vorm.

Er is een prijs betaald voor de genoemde voordelen. Als eerste: de rekentijd van associatieve voorspellingsmodellen is langer dan die van andere technieken. In de politiepraktijk veroorzaakt dat geen problemen omdat een langere wachttijd zich typisch voordoet aan het eind van het analyseproces, als er bijvoorbeeld tijd is om te wachten op de resultaten van een batch-opdracht. Daarnaast gaat associatief clusteren heel snel. Ten tweede: soms zijn alternatieve technieken nauwkeuriger dan een associatief geheugen en soms andersom. Wij vinden dat deze incidentele kleine verschillen in modelkwaliteit niet opwegen tegen de voordelen: de analist wordt veel tijd bespaard, kan meer informatie gebruiken en non-datamining experts kunnen gebruik maken van datamining.

- **Integratie**

Doordat veel tools en technieken in één tool zijn geïntegreerd, is er meer consistentie in het softwaregebruik; de gebruiker hoeft niet langer verschillende tools te beheersen en resultaten uit te wisselen tussen de tools door middel van importeren en exporteren. Het systeem beschikt over meer functies dan alleen maar dataminingstechnieken, uiteenlopend van het simpelweg bladeren door gegevens tot geavanceerde OLAP analyse. Voor sommige resultaattypes zijn populaire standaardtools op een zodanige manier gekoppeld dat resultaten automatisch uitwisselbaar zijn: ExcelTM, MapInfoTM, Microsoft WordTM, Cognos ReportnetTM, Analyst's NotebookTM, Weka en Google MapsTM. Veel analisten zijn al bekend met deze tools.

- **Geo-spatieële analyse**

Het spatieële aspect van criminaliteit is vanzelfsprekend belangrijk, daarom maakt het dataminingsysteem resultaten zichtbaar in kaarten en is het in staat spatieële aspecten te gebruiken in modellen (bijvoorbeeld coördinaten, grondgebruik en socio-demografie).

- **Geautomatiseerd routinewerk**

Het dataminingsysteem omvat een rapportagemodule die een rapport kan produceren voor ieder gebied en voor elk type speerpuntmisdrijf, met de volgende elementen :

1. Hotspotkaarten van de voorafgaande periode.
2. Trend-hotspotkaarten die laten zien wat er is veranderd.
3. Voorspellingskaarten voor de aankomende periode.
4. Een waar/wanneer-analyse met een beschrijving van de gevonden clusters.
5. Een voorspelling van de week-drukte van misdrijven over de aankomende periode: op welke dagen en tijden worden de hoogste aantallen misdrijven verwacht?
6. Criminiteitcijfergrafieken met basisstatistieken, trends en *key performance indicators*.
7. Hotshot-lijsten met de veelplegers, hun sociale netwerkstatus en foto's.
8. Kaarten die de verblijfplaatsen en werkterreinen van veelplegers laten zien.

- **Gemeenschappelijk gebruik van praktijkervaringen**

Het systeem geeft gebruikers de mogelijkheid om hun werkwijze te hergebruiken en te delen in de vorm van recepten: beschrijvingen van problemen met de stappen die zijn genomen om ze op te lossen. Deze best practices kunnen worden ingezien en hergebruikt door alle andere gebruikers.

4.2 Datawarehouse

Het dataminingsysteem verschaft inzicht in duizenden variabelen uit verschillende politiestructuren, socio-demografische gegevens, spatieële informatie, het weer en levensstijlgegevens. Al deze gegevens worden steeds geëxtraheerd, gekoppeld, opgeschoond, verrijkt en toegankelijk gemaakt in een open datawarehouse door een geautomatiseerde module. Dat betekent dat de data niet opnieuw verzameld en opgeschoond dient te worden voor elke analyse.

De datawarehouse integreert de volgende bronnen:

1. BPS/BVH/GIDS: de hoofdtransactiesystemen bestaande uit incidenten, goederen, personen, vervoermiddelen etc.
2. HKS: een systeem dat beschikt over meer details aangaande daders en een langere geschiedenis.
3. SHERPA: uitvoerig geografisch materiaal, dat gebruikt wordt om meer informatie te leveren over de plaats delict (omgevingstype, omgevingsinfrastructuur etc.) en de verblijfadressen van personen.
4. CBS: Socio-demografische informatiebron op buurniveau.
5. Experian: uitgebreide bron van informatie over socio-demografie en levensstijl op postcodeniveau.
6. Zon/Maan: informatie over de lichtcondities op een gegeven plaats en tijd.
7. Evenementen: evenementen die plaatsvinden op vaststaande plekken (bijvoorbeeld voetbalwedstrijden, festivals, vakanties).
8. KNMI: weerinformatie op een gegeven plaats en tijd.

Van deze bronnen wordt informatie verzameld op het niveau van incidenten en personen (slachtoffers, daders, getuigen, andere betrokkenen). Omdat het systeem in staat is om veel variabelen te hanteren en met veel verschillende datatypes te werken, worden ook relationele gegevens bij elkaar gebracht, bijvoorbeeld een lijst van misdrijftypes bij het dossier van een dader.

4.3 Soorten gebruik

Er kunnen drie soorten gebruik van het dataminingsysteem worden onderscheiden: persoonlijke desktop-analyse, groepsessies en rapporten. De groepsessie is een speciale vorm van de persoonlijke desktop-analyse waarin de analist het systeem bedient in bijzijn van domeinexperts en belanghebbenden om interactief naar een probleem te kijken. Op deze manier kunnen nieuwe vragen en theorieën direct worden geadresseerd en bekrachtigd.

Bij het verspreiden van resultaten van geautomatiseerde dataminingrapporten kan een groot publiek profijt hebben van analyseresultaten zonder zelf met het dataminingsysteem te werken. Als er vragen opkomen naar aanleiding van de rapporten kunnen de lezers een interactieve dataminingssessie aanvragen met een getrainde gebruiker.

4.4 Gebruikersorganisatie

De Amsterdamse politie startte in 2001 met het gebruik van de eerste versie van het dataminingsysteem en heeft inmiddels rond de dertig geautoriseerde en getrainde gebruikers. Iedere gebruiker heeft toegang tot een specifieke selectie van de functies van de tool. In Amsterdam zijn de gebruikers georganiseerd in groepen van twee per district. Deze kleine teams worden bijgestaan door een centraal team van analisten en één persoon is verantwoordelijk voor het aansturen van de functionaliteit van het systeem en de communicatie met de groep van gebruikers. Deze gebruikersgroep bestaat uit domeinexperts: politie-analisten, onderzoekers, en rechercheurs.

De training die deze gebruikers nodig hebben varieert van één tot drie dagen. Bij het beoordelen van nieuwe gebruikers wordt meer gekeken naar analytische en communicatieve vaardigheden, dan naar technische kennis en opleiding. Simpel gezegd is het profiel van een dataminer: een intelligent persoon met analytisch inzicht, goed met cijfers, ervaring met computers, goede presentatievaardigheden, verbaal sterk en met kennis van het politiedomein. Gebruikers hoeven geen achtergrond te hebben op het gebied van statistiek, data-analyse of databases. Het is wel een voordeel wanneer ze kennis hebben van het datamodel.

Het belang van domeinexperts bij datamining

Uitgebreide kennis van het politiedomein is essentieel bij het herkennen van bruikbare patronen en het interpreteren daarvan. Een voorbeeld: een dataminingexpert van buiten de politie was gevraagd om het activiteitenpatroon te analyseren van drugsdelicten in en rondom een winkelcentrum. De expert vond een regelmatigheid maar dacht dat het niet interessant was omdat het leek alsof er geen oorzaak en gevolg waren. Echter, een ervaren politiemedewerker merkte op dat dit specifieke gedrag overeen kwam met het tijdschema van de zogenoemde methadonbus, die voorgeschreven drugsvervangers verstrekt aan verslaafden. Verder, omdat de methadonbus een doeltreffend programma is, lag de oplossing niet in het stopzetten van het langskomen van de bus in die buurt, maar het vinden van de veroorzakers van de drugsdelicten op die gezette tijden en locaties. Het bleek dat een paar verslaafden in het methadonprogramma zich aanhoudend misdroegen in de omgeving van de bus. Het plan van aanpak werd deze individuen te confronteren en te informeren dat de verstrekking van hun drugsvervangers op het spel stond.

5. TOEPASSINGEN

De volgende paragrafen beschrijven de belangrijkste analysemethoden die met het dataminingsysteem worden toegepast in de praktijk.

5.1 Tijd-ruimte clusters: waar, wanneer

Een goede inschatting van wanneer en waar de kans op criminaliteit het hoogst is, maakt een pro-actieve en effectieve inzet van middelen mogelijk. De traditionele methode voor spatiale risicoanalyse maakt ofwel een lijst van gebieden met de hoogste recente activiteit ofwel een kaart, die in essentie hetzelfde doet. Deze methode negeert het feit dat de criminaliteit in een gebied sterk afhangt van de tijd van de dag en de dag van de week. Bijvoorbeeld: sommige gebieden zijn overvol rond het spitsuur, sommige gebieden zijn druk bezet wanneer de café's sluiten en sommige gebieden lijden onder veel inbraken op zaterdag omdat veel bewoners dan niet thuis zijn. Om deze patronen te vinden, kan men een enorme kruistabel creëren op de

combinatie van buurt, tijd en dag van de week, wat lastig zou zijn om te interpreteren. Er is ook meer locatiedetail vereist voor een goede tactische planning. Immers, de aanwezigheid van politie in de ene straat kan problemen in de straat verderop niet uitsluiten. Daarom is het belangrijk om exacte coördinaten te gebruiken, ook omdat daders geen rekening houden met grenzen van gebieden. Wanneer er coördinaten worden gebruikt in plaats van gebieden, schieten traditionele methodes en visuele inspectie te kort.

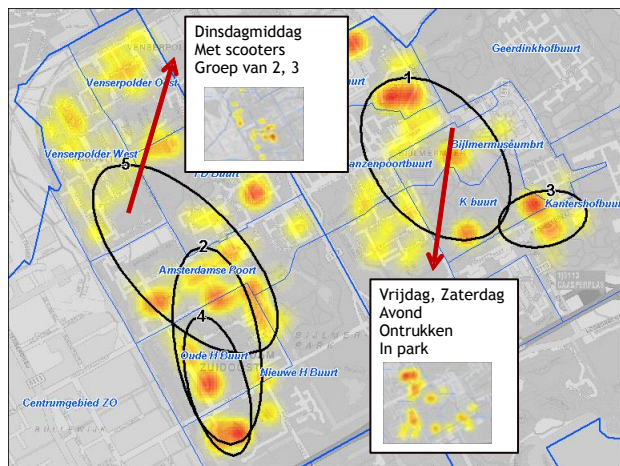
Eén van de meest succesvolle toepassingen van het dataminingsysteem is de zogenaamde 'waar-en-wanneer-analyse' die clusters maakt van recente incidenten op basis van coördinaten, tijdstip en dag van de week. Elk cluster wordt gerapporteerd met een typisch profiel van die incidenten met daaraan toegevoegd een MO en beschrijvingen van de daders. Tezamen met de locatie en tijd hebben politiemedewerkers zo alle informatie om te weten waar en wanneer ze ergens heen kunnen gaan en waar ze op kunnen letten.

De waar-en-wanneer-analyse werkt met een associatieve clustertechniek die we baseren op het principe van *Metric Multi Dimensional Scaling* (MMDS) [10], door de hoogdimensionale ruimte te projecteren op twee dimensies op een niet-lineaire wijze. Dit proces probeert de afstanden (de mate van overeenkomst) te behouden tussen incidenten zoals vastgesteld door het associatief geheugen. In tegenstelling tot factoranalyse, wordt alle variantie verwerkt in het tweedimensionale vlak. Dit leidt tot enige vervorming, maar dat is geen probleem omdat het doel van de techniek is clusters te visualiseren en te herkennen. Het voordeel van de MMDS-benadering is dat de resultaten eenvoudiger te interpreteren zijn voor non-dataminingexperts, omdat het verschijnt als een normale *scatterplot* die de clusters en de relaties onderling visualiseert.

Ons clusteralgoritme past eerst het associatieve geheugen toe om de afstanden tussen de incidenten te bepalen. Deze afstanden worden dan gebruikt om de zwaartekrachten tussen deze incidenten te berekenen: hoe meer incidenten op elkaar lijken, des te sterker willen ze naar elkaar toe bewegen in het tweedimensionale vlak. Verder is er een roterende kracht en een kracht die alle incidenten wegdrijft uit het midden. Deze krachten worden vervolgens gebruikt door middel van iteratie in een optimalisatieproces waarin de incidenten die op elkaar lijken naar elkaar toe bewegen en wolken vormen in het veld (zie figuur 4). Wanneer de beweging daalt tot onder een drempelwaarde, stopt de iteratie en zullen de afstanden in twee dimensies lijken op de multidimensionale afstanden.

De resultaten van de waar-en-wanneer-analyse zijn continu van waarde voor de preventie en repressie van speerpunt misdrijftypes, daarom worden clusteroverzichten automatisch gegenereerd en toegevoegd aan de wekelijkse standaardrapporten. Deze overzichten laten een hotspotkaart zien van het gebied, met ellipsen getekend waar de clusters zijn, plus een lijst van clusters, elk gepresenteerd door een hotspotkaart, een beschrijving van tijd en dag van de week en een profiel van wat typerend is voor het cluster (bijvoorbeeld dadertype, type buit).

In Amsterdam wordt deze analyse ook gebruikt om te bepalen waar en wanneer preventief fouilleren moet worden gepland, met het doel wapens te vinden die mensen op straat bij zich dragen. Dit is een samenwerking van de gemeente met de Amsterdamse Politie. Sinds de start van deze door datamining gestuurde zoekacties is het wapenbezit gedaald met 27%.



Figuur 1: Tijdruimtelijke clusters met profielen

Tabel 1: kwaliteit associatieve geografische voorspelling

Gebied	Techniek	Gemiddelde fout ²
Tilburg Centrum	Kernel density	0,11
Tilburg Centrum	Associatief	0,052
Noord Tilburg	Kernel density	0,086
Noord Tilburg	Associatief	0,044

De tabel laat zien dat het associatieve model de standaard voor deze situaties overtreft met ongeveer de helft van de fout. Als vervolg op dit werk wordt gekeken naar meer situaties en naar een kwaliteitsmeting die rekening houdt met hoe de techniek de effectiviteit van politiewerk in de praktijk verbetert.

5.2 Associatieve geografische voorspelling

Tijdruimtelijke clusters (zie 5.1) vinden patronen in ruimte en tijd die gebruikt kunnen worden in een algemeen tactisch plan voor een specifieke periode, bijvoorbeeld de aanwijzingen voor een andere surveillanceroute voor elke dag van de week gedurende de maand november 2009. Associatieve geografische voorspelling is een andere analysemethode, gericht op het leveren van een kaart met een optimale schatting van criminaliteitsrisico's voor een specifieke korte periode; zeg 1 november 2009 gedurende de dienst van 16:00 tot 20:00. Voor zo'n korte periode is meer context bekend omdat er bijvoorbeeld voor die dag een weersvoorspelling is, in november gaat de zon vroeg onder en het is Allerheiligen. Associatieve geografische voorspelling past een associatief geheugen toe om te zoeken naar relevante situaties in het verleden, deze situaties worden vervolgens gestapeld om een gedetailleerde hotspotkaart te maken die geschatte geografische spreiding van risico's toont voor de gegeven toekomstige situatie.

Deze benadering is gebaseerd op de theorieën van *repeat victimization* [11][12], *routine activity* [13][14] en *prospective hotspotting* [15]. Associatieve geografische voorspelling bouwt voort op dit werk door rekening te houden met meer dan alleen locatie: recentheid, trends, seizoensinvloeden, het weer, tijdstip, dag van de week, evenementen en vakanties. Experimenten hebben laten zien dat voor overvallen de hotspots in de voorspellingskaarten 50% van alle toekomstige incidenten bevatten, waar de traditionele methode (een hotspotkaart van de voorafgaande periode) maar 25% in de hotspots laten zien.

Toetsing

Associatieve geografische voorspelling is getoetst door het specificeren van meerdere willekeurige combinaties van tijdsperiodes voor training en testen. Voor elke combinatie is het voorspellingsmodel getraind op de ene tijdsperiode en getest op de andere. De uitvoering van het model is gemeten door de fout te berekenen tussen de voorspelde aantallen misdrijven en de actuele (toekomstige) aantallen voor elk vak in een 30x30 raster over het totale gebied dat voorspeld is. De totale modelfout is het gemiddelde over alle cellen, gemiddeld over de diverse testsets. De tabel hieronder laat de modelresultaten zien vergeleken met wat beschouwd wordt als de standaardmethode voor geografische anticipatie; een hotspotkaart van de misdrijven in de recente periode (*Kernel density* – zie 5.4):

5.3 Analyseren van trends of gedrag

Deze paragraaf bespreekt hoe het dataminingsysteem wordt gebruikt om trends en gedrag te beschrijven en te verklaren.

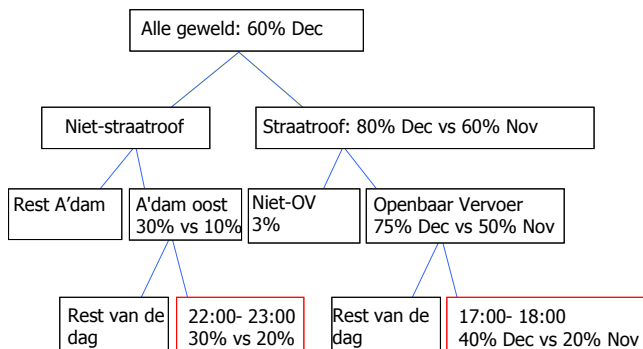
5.3.1 Verklaren van trends

Politiecorpsen reageren op veranderingen in criminaliteit. Het is vaak noodzakelijk om de reden achter een trend te begrijpen om te weten of de trend verklaard kan worden door een bekend fenomeen of dat het speciale aandacht nodig heeft. Ook kan door het vinden van een waarschijnlijke oorzaak, de oorzaak mogelijk worden aangepakt.

Het dataminingsysteem levert trendverklaringen door de recente periode te vergelijken met de periode daarvoor, gebruikmakend van chi kwadraat en *Student testen* voor alle beschikbare datakolommen en vervolgens op de index te sorteren. Het resultaat is een *profielanalyse*; een opsomming van de meest significante verschillen tussen de periodes, als mogelijke verklaringen. Het systeem kan hierin ook een stap verder gaan door de optie te bieden een beslisboom te bouwen die dezelfde testen gebruikt, om combinaties van factoren te vinden als verklaringen voor een trend. Bijvoorbeeld: aan het eind van vorig jaar vonden er meer inbraken plaats in de vroege avond in het zakelijke gebied en op zaterdagochtenden in de westelijke buitenwijken.

Diezelfde methode kan worden gebruikt bij het bepalen van het succes van tegenmaatregelen. Bijvoorbeeld: het effect van het installeren van bewakingscamera's in het publieke domein werd geanalyseerd door de periode na het installeren te vergelijken met de periode ervoor. De resultaten lieten zien dat het vandalisme was afgenomen, maar meldingen van straatroven in parkeergarages (buiten het zicht van de camera's) waren toegenomen.

Het voorbeeld op de volgende pagina laat zien hoe het verklaard kan worden waarom afgelopen december meer geweldsdelicten vertoonde dan november. De boom begint met het kijken naar alle geweldsdelicten in november en december, waar december 60% van het totaal inneemt. Het beslisboomalgoritme verklaart dat straatroven de meeste verschillen laat zien: 80% van alle geweldsdelicten in december betrof een straatroof, wat in november 60% was. Vervolgens blijkt dat overvallen in het openbaar vervoer relatief zijn toegenomen, vooral tussen 17:00 en 18:00. De laatstgenoemde dekt 40% van 75% van 80% is 24% van alle geweldsdelicten in december.



Figuur 2: Trendverklarende beslisboom

5.3.2 Contextuele trends opsporen

Een alternatieve manier om periodes met elkaar te vergelijken is het vinden van trends in de context; niet hoe de trends in criminaliteit veranderen, maar de trends in de aspecten van criminaliteit. Bijvoorbeeld: er is geen sterke toename in de totale cijfers aangaande inbraken, maar er is wel een plotselinge toename in diefstal van flatscreens uit appartementen, of een trend waar een specifiek werktuig wordt gebruikt. Deze trends kunnen worden gebruikt om de activiteit van een specifieke dader of dadergroep aan te duiden, of een meer algemeen fenomeen, zoals een toename in diefstal van motoren omdat het voorjaar zich aandient.

5.3.3 Gedrag verklaren of beschrijven

Diezelfde aanpak kan worden toegepast op het verklaren van gedrag, door activiteiten en/of kenmerken van een persoon of groep te vergelijken met een referentiegroep. Bijvoorbeeld: het vergelijken van bepaalde gewelddadige jonge criminelen in een buurt met alle jonge criminelen in diezelfde buurt. Wat maakt deze gewelddadige individuen anders? Komen ze uit een specifiek deel van de stad, zijn er patronen in hun loopbanen? Wat is hun typische sociale achtergrond? Zulke patronen kunnen worden gevonden door een beslisboomanalyse. Bij het verkregen inzicht in aspecten van dergelijk gedrag kunnen preventieve maatregelen worden getroffen. Een andere toepassing is het maken van een beschrijving van de *handtekening* (typische manier van werken) van een persoon of een groep.

5.3.4 Tijd-ruimtelijke relaties verklaren

Het opsporen van relaties tussen spatiële aspecten (bijvoorbeeld omgevingstype) en gedrag (bijvoorbeeld criminaliteit) wordt traditioneel aangepakt door naar omgevingen te kijken: de criminaliteit in elk gebied wordt gebruikt als variabele die verklaard moet worden door de kenmerken van dat gebied (gemiddeld inkomen etc.). Wij hebben een nieuwe methode ontwikkeld om deze relaties te verklaren, die de dichtheid van de lokale criminaliteit als te verklaren variabele neemt. Op deze manier kan de mate van detail hoger zijn, bijvoorbeeld op adresniveau, zonder de behoefte aan een groot aantal incidenten op dat niveau. Als eerste wordt er een hotspot-analyse uitgevoerd (zie paragraaf 5.4), die de dichtheid van criminaliteit berekent voor elk adres. Vervolgens worden de verklaringstechnieken toegepast. Dit resulteert in betere en meer gedetailleerde verklaringen omdat criminaliteit en spatiële aspecten niet worden gemiddeld over grote gebieden.

5.3.5 Gebruik van teksten

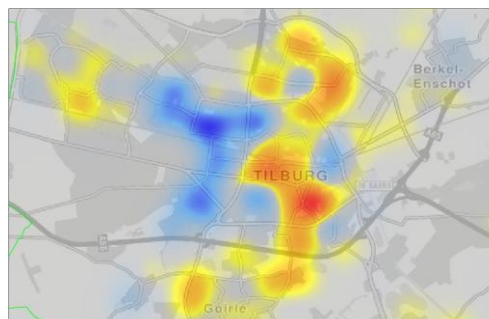
Het gebruik van tekstuele informatie kan veel hulp bieden bij het verkrijgen van inzicht. Een voorbeeld: het systeem werd gevraagd om de ramadan periode (islamitische feestmaand) te analyseren en het ontdekte dat er in deze periode een opvallende toename was van vandalisme. Voordat we moesten denken aan de eventuele relaties tussen dit patroon en het vasten, liet het systeem ook zien dat de woorden *rotje* en *vuurwerk* meer voorkwamen bij incidenten gedurende de ramadan. De ramadan vindt plaats in de negende maand van de islamitische kalender, waardoor het soms samenvalt met de maand december. Dit verklaart het vuurwerkpatroon en dus ook de vernieling van eigendommen voortgekomen uit de incidenten die er zijn op de dagen voor oudejaarsavond, wat werd bevestigd door de resulterende beslisboom.

5.4 Hotspotkaarten

Weten waar misdrijven plaatsvinden is cruciale informatie voor de politie en wordt voor optimale interpretatie het beste weergegeven in een kaart. De basismanier om te visualiseren is het weergeven van losstaande incidenten met behulp van punten. Als de punten heel dicht bij elkaar staan, kunnen ze worden gecombineerd tot grotere punten. Deze methode lijkt voor de hand te liggen. Echter dergelijke puntkaarten kunnen moeilijk te interpreteren zijn, vooral wanneer de concentratie van punten hoog is. Hotspotkaarten leveren de oplossing door incidenten te interpoleren voor elke cel van een gedetailleerd raster op de kaart, resulterend in een kleurgecodeerde hotspotkaart. Het dataminingsysteem gebruikt *kernel density estimation* [13] voor de interpolatie op drie detailniveaus, elk voor een ander zoombereik op de kaart. Op deze manier laat de hotspotkaart meer detail zien wanneer de gebruiker inzoomt. Deze techniek is uitgebreid met de mogelijkheid om ook delen van straten te visualiseren, naast exacte adressen, omdat een groot percentage van incidenten geregistreerd staat met alleen een straatnaam.

5.5 Temporele hotspots

Het dataminingsysteem kan temporele hotspotkaarten maken waarmee spatiële trends in de tijd worden gevisualiseerd. Dit wordt uitgevoerd door een *kernel density* raster aan te maken voor de recente periode en één voor de periode daarvoor, die dan met elkaar worden vergeleken, wat resulteert in een dichtheidkaart met positieve (rood) gebieden met een toename van criminaliteit en negatieve (blauw) gebieden met een afname van criminaliteit.



Figuur 3: Temporele hotspot-analyse van inbraken in Tilburg

5.6 Cluster reeksen van misdrijven

Een typische taak voor een politie-analist is het verbanden leggen tussen misdrijven, voor het oplossen van een zaak of

verdachte zijn voor een zeer overeenkomend incident van een paar weken daarvoor, maar de gelijkenis kan subtiel zijn; het eerdere misdrijf is opgenomen als een zakkenrol-incident en niet als een straatroof; de zakkenrol-locatie was vlak bij de locatie van de beroving, maar niet in dezelfde straat; getuigenverklaringen verklaren blond haar voor de zakkenroller, maar donkerblond voor de straatrover. Er zullen overeenkomsten zijn die de zakkenrol-zaak interessant maken, maar die zullen typisch niet worden gevonden met de gebruikelijke selectiemethoden.

Het politie-dataminingsysteem beschikt over associatieve technieken die gelijksoortige zaken kunnen vinden op basis van een gegeven zaak. Deze aanpak wordt ook gebruikt als onderdeel van de algoritmen voor voorspellingsmodellen en clustering. Een dergelijke zoekactie neemt een zaak als invoer en resulteert in een uitvoer van gelijksoortige zaken, geordend naar overeenkomst. Deze overeenkomstige zaken kunnen dan worden gebruikt om gerelateerde verdachten te vinden. Wanneer één van die verdachten meer dan één gelijksoortige zaak heeft, dan wordt die verdachte interessanter. Dit gehele proces is geautomatiseerd in het systeem en ook kunnen meerdere zaken als invoer worden gebruikt voor een zoekactie. Bijvoorbeeld: een reeks van 10 misdrijven werd gebruikt in een associatieve zoekactie die 33 gelijksoortige zaken in het verleden vond, waarvan 4 zaken dezelfde verdachte hadden en die 4 zaken lijken sterk op 8 van de 10 zaken in de zoekactiereeks.

Er zijn meerdere toepassingen voor associatief zoeken, bijvoorbeeld het zoeken van foto's voor een getuige op basis van diens verklaring, of het kijken naar incidenten die het meest lijken op de gehele criminele geschiedenis van een dader, om op die manier andere misdrijven op te sporen die gepleegd zouden kunnen zijn door diezelfde dader.

6. SUCCESVERHALEN

De volgende paragrafen bespreken succesverhalen van het dataminingsysteem in de praktijk.

6.1 Overval

In het begin van 2009, kende het district Tilburg een ernstige golf van overvallen bij vooral tankstations en restaurants. Analisten gebruikten het dataminingsysteem om de locaties van deze overvallen te visualiseren en pasten associatieve geografische voorspelling toe om te bepalen waar en wanneer politie-acties (bijvoorbeeld wegversperringen) optimaal zouden zijn. Zij pasten netwerkanalyse toe op overvallen in het verleden om te bepalen welke verdachten belangrijk waren om in de gaten te houden. Daaraan toegevoegd produceerde het systeem top 10 lijsten van daders, gevolgd door hun overvalactiviteiten in het verleden. Deze geselecteerde personen werden ontboden op het politiebureau als ze boetes open hadden staan. De rest kreeg een bezoek van de politie, en hun foto's werden gebruikt in briefings. Binnen twee weken was de golf van overvallen gestopt.

6.2 Autodiefstallen

In 2006 gaf een trendanalyse in Amsterdam een toename van diefstallen van motorvoertuigen aan in district 2 voor de maand mei. Een profielanalyse van deze trend toonde aan dat de toename kon worden verklaard door diefstallen in privégarages. Er werden meer agenten dan normaal ingezet en hun acties werden ondersteund door waar-wanneer analyses plus de top 10 lijst van autodieven. Het resultaat was dat binnen een uur nadat de informatie door datamining was geleverd, een veelpleger op heterdaad werd betrapt en dat de criminaliteit in de eerste week daalde met 90%.

6.3 Inbraak

Het volgende scenario laat zien hoe dataminingstechnieken werden gecombineerd voor verklaren en ontrafelen bij het oplossen van een probleem.

1. Het standaardrapport wijst op een plotselinge toename van inbraken in oktober, wat atypisch is voor de tijd van het jaar.
2. De gebruiker maakt een grafiek met een tijdreeks om de trend te visualiseren en de indicatie te toetsen.
3. Met het aanklikken van oktober in de grafiek, vraagt de gebruiker om een verklaring van de trend in de vorm van een beslisboom, om te ontdekken welke combinaties van factoren veranderd zijn (zie 5.3). De beslisboom laat zien dat er meer inbraken zijn aan de achterzijde van huizen dan voorheen, vooral in de vroege avond. Het laat ook zien dat de inbraken vaker plaatsvinden na zonsondergang. Dit leidt tot de theorie dat doordat het vroeger donker wordt er meer gelegenheid is in de vroege avond, vooral in achtertuinen omdat daar nauwelijks straatverlichting is. Dit vindt vooral plaats in een specifieke buurt. Dit patroon kan direct worden gebruikt om de trend te verklaren, maar bijvoorbeeld ook voor het opstarten van een programma voor het stimuleren van gebruik van (bewegingsgevoelige) verlichting in achtertuinen in die buurt.
4. De beslisboom duidt ook op een sterk geografisch verschil. Dit leidt tot een temporele hotspot-analyse (zie 5.5) die een aantal hotspots laat zien waar de criminaliteit is toegenomen.
5. De grootste hotspot wordt gekozen door in te zoomen op de kaart en het systeem te vragen een verklaring te creëren voor deze trend. Dit wijst op een zeer specifiek tijdstip, specifieke straten en een aantal getuigenverklaringen; bruikbare informatie om agenten in dat gebied te ondersteunen in hun surveillance.
6. Om te bepalen waar de agenten op kunnen letten wordt de hotspot-selectie gebruikt om associatief zoeken (zie 5.10) toe te passen. Dit koppelt de incidenten in de hotspot met de gehele gekende misdrijfgeschiedenis. De beste en meest recente matches worden geselecteerd om te zien wie de betrokken verdachten waren, en vervolgens worden hun foto's aan de agenten geleverd.
7. Daarna wordt een netwerkanalyse uitgevoerd om de personen te vinden die direct of indirect betrokken zijn bij de geselecteerde waarschijnlijke daders, om vervolgens ook hun foto's aan te leveren.

7. DISCUSSIE

In dit hoofdstuk bespreken we de toegevoegde waarde van het dataminingsysteem voor het politiewerk, de betreffende uitdagingen en toekomstig werk.

7.1 Toegevoegde waarde

Hoofdstuk 2 bespreekt de verschillende redenen waarom datamining belangrijk is voor criminaliteitsanalyse. Het organisatorische gevolg van het besproken dataminingsysteem is dat een grote groep medewerkers de mogelijkheid krijgt om meer inzicht te verkrijgen en om crimineel gedrag te voorspellen. Deze personen zijn verantwoordelijk voor de informatievoorziening van de organisatie, dus met andere woorden: datamining maakt de organisatie intelligenter. Het toenemend aantal gebruikers van het dataminingsysteem is een duidelijk teken, dat wijst op de acceptatie van datamining als zijnde een belangrijke

toepassing. Maar hoe kan daarnaast de toegevoegde waarde worden gemeten?

Er zijn vijf verschillende methoden om de toegevoegde waarde van datamining te bepalen: modelnauwkeurigheid, experimenten in de praktijk, nabootsen van de praktijk, analist-efficiëntie en kwalitatief vergelijken.

7.1.1 Meten van modelnauwkeurigheid

De nauwkeurigheid van een voorspellingsmodel kan worden gemeten door het model te baseren op een gedeelte van de gegevens en het vervolgens te testen op de rest. De resulterende nauwkeurigheid kan worden vertaald naar de verwachte toename in de effectiviteit van politiewerk. Bijvoorbeeld: we berekenden dat de top 5% hotspotgebieden van associatieve geografische voorspelling uiteindelijk 50% bevat van alle misdrijven in de voorspelde periode, terwijl daarentegen de top 5% gebieden van de traditionele verwachting 25% bevat. We nemen aan dat de politie net voldoende capaciteit heeft om aanwezig te zijn in die 5% gebieden en dat de aanwezigheid van politie de kans op het plaatsvinden van een incident met 50% reduceert. Op deze manier kunnen we de geprojecteerde vermindering van misdrijven berekenen. Voor de spatiële voorspelling is 50% van alle misdrijven gereduceerd met 50%, dus de totale criminaliteit zou gereduceerd zijn met 25%, waar de traditionele benadering een reductie van 50% van 25% laat zien, wat een criminaliteitsvermindering is van 12,5%. Deze methode om effectiviteit te projecteren maakt gebruik van aannames, maar kan bruikbaar zijn bij het illustreren van de impact van deze voorspellingmodellen.

7.1.2 Experimenten in politiepraktijken

Een praktijkproef is de ideale manier om effectiviteit te meten. Echter, omdat publieke veiligheid niet iets is om mee te experimenteren, zijn zuivere testen moeilijk te organiseren. Daarnaast zijn criminaliteit en omgeving zo dynamisch dat een exclusieve succesverklaring niet gebaseerd kan worden op een experiment alleen. Criminaliteit kan tijdens een experiment veranderen om verschillende redenen anders dan het wel of niet toepassen van datamining. Daarom zijn meerdere experimenten vereist voor een goede meting. Een andere uitdaging bij het meten van toegevoegde waarde is het meten van veiligheid en veiligheidsgevoel. Aangiftecijfers zijn bruikbaar maar dekken maar een gedeelte van de lading.

Onlangs pakte politiekorps Midden- en West-Brabant een regiobrede golf van misdrijven aan. Eén district paste datamining toe om te zorgen voor informatie over deze misdrijven, de andere deden dat niet. Na twee maanden reduceerde het district dat datamining gebruikte criminaliteit met meer dan 15%, terwijl de andere districten constante cijfers lieten zien. Er zijn meer aanmoedigende resultaten dan deze, maar dat zijn statistisch geen harde bewijzen. Vanwege de moeilijkheden met het meten van dit type toegevoegde waarde gebruiken we om politieorganisaties te overtuigen ook de overige methoden uit dit hoofdstuk.

7.1.3 Nabootsen van de politiepraktijk

Een alternatieve manier om in de praktijk te experimenteren is het nabootsen van de praktijk. We voerden een veldtest uit met de politiekorpsen van Rotterdam en Haarlem waarbij vrijwilligers werden gevraagd om in een wachtruimte te zitten, waar een laptop werd 'gestolen' door een acteur. De vrijwilligers waren getuige van deze actie en werden gevraagd beschrijvingen te geven die vervolgens werden gebruikt om foto's te zoeken die werden getoond voor identificatie. Dit werd gedaan door twee systemen: het standaardsysteem van de politie en een associatieve

zoektool, met dezelfde techniek als in het huidige dataminingsysteem. Het resultaat was dat 50% meer getuigen de foto's herkenden van de criminelen die werden geselecteerd op basis van het associatieve systeem, dan die van het standaard zoekstelsel van de politie. Het nadeel van dergelijke experimenten is dat ze kostbaar zijn.

7.1.4 Meten van de efficiëntie van analisten

Een alternatief om de toegenomen effectiviteit te meten is het meten van de winst aan *efficiëntie*, als een resultaat van datamining. Het meten van de efficiëntie van de analist is eenvoudiger dan het meten van de effectiviteit van politiewerk. Ook hun werksituatie is vanzelfsprekend makkelijker constant te houden dan criminelen en hun omgeving. Onderzoek in de regio Midden- en West-Brabant liet zien dat analisten in het algemeen hun analysetijd terugbrachten met een factor 20 wanneer ze het dataminingsstelsel gebruikten. De toegevoegde waarde daarvan komt neer op kostenreductie of effectiviteitsverbetering, omdat analisten meer werk kunnen verrichten. Verder zijn de responstijden veel korter en de analyse wordt snel genoeg uitgevoerd om groepsessies te ondersteunen waarin experts en belanghebbenden een onderwerp bespreken en analyseren.

7.1.5 Kwalitatief vergelijken van resultaten

Onze ervaring is dat een kwalitatieve benadering van de meerwaarde van datamining vaak overtuigend is voor beslissers. Zodra analisten en hun meerderen ervaren wat datamining hen kan bieden hebben zij opvallend weinig behoefte aan het kwantitatieve bewijs dat de organisatie effectiever wordt. Hetzelfde is het geval bij bijvoorbeeld geografische informatiesystemen. Zodra medewerkers beginnen te werken met kaarten is de toegevoegde waarde helder en is er geen vraag naar het uitvoeren van experimenten waarin de nieuwe situatie wordt vergeleken met situaties zonder kaarten. Onderzoeken door ABRIO [intern rapport] en Midden- en West-Brabant toonden aan dat datamining productiever is en dat de resultaten beter voorzien in de informatiebehoefte vanwege het detail en het verklarend vermogen.

7.1.6 Kosten en rendement op investering

De kosten voor datamining kunnen worden verdeeld in implementatie- en exploitatiekosten. Implementatie vereist de beschikking over een datawarehouse waarin de gegevensbronnen zijn geëxtraheerd, gekoppeld, opgeschoond en eventueel verrijkt met berekeningen (bijvoorbeeld huiselijk geweld ja/nee). Een dergelijke datawarehouse levert een *single point of truth* op voor de organisatie en de meerwaarde daarvan reikt daarmee verder dan datamining alleen. Daarom zien politiekorpsen in Nederland het beschikken over een datawarehouse als een algemene must en zijn er veel datawarehouse initiatieven. Wij hebben een datawarehouse ontwikkeld dat is gebaseerd op de meest gebruikte standaard politiestelsels in Nederland. Die ontwikkelkosten hoeven korpsen dus niet te dragen. Wat overblijft zijn de kosten voor hardware en software voor de databaseserver, installatie en configuratie.

Andere kosten voor de implementatie van datamining zijn: serverhardware voor de applicatie, softwarelicentie en training. Exploitatiekosten zijn: beheerkosten van de systemen, softwarelicentie, herscholing, functioneel beheer en uitvoeringskosten van batch-modules die de datawarehouse updaten en standaardrapporten produceren.

Wij hebben deze kosten sterk kunnen reduceren door het maken van een dataminingsstelsel dat eenvoudig is te leren

en makkelijk is in gebruik. Training wordt gegeven in twee of drie dagen en er hoeven geen experts te worden ingehuurd.

Het is mogelijk om een positief rendement op investering te bepalen alleen al gebaseerd op het voordeel van de analyse-efficiëntie (7.1.4). De winst aan effectiviteit is een factor 20, dus theoretisch kan de capaciteit van het analysepersoneel sterk worden teruggebracht, terwijl hetzelfde niveau van informatievoorziening gewaarborgd blijft. Dit zou een kostenbesparing creëren die ruimschoots opweegt tegen de kosten gemaakt voor implementatie en exploitatie. Echter, politieorganisaties hebben niet de intentie de capaciteit aan analisten terug te brengen omdat de informatiebehoefte de informatieproductie sterk overstijgt. Met andere woorden: het gebruik van datamining kan worden vergeleken met het inhuren van meer analisten voor een fractie van de kosten. Dit is de business case voor datamining, enkel gebaseerd op de efficiëntie van analisten, alle overige voordelen buiten beschouwing gelaten.

7.2 Uitdagingen

De implementatie en het gebruik van het dataminingsysteem brengen een aantal uitdagingen met zich mee.

7.2.1 Gebruikersvaardigheden managen

Ook al is er veel aandacht besteed aan het gebruiksvriendelijk maken van het systeem, het blijft altijd belangrijk het niveau van de gebruikersvaardigheden op peil te houden. Analyse is nu eenmaal een complexe taak, zelfs wanneer alle keuzes aangaande parameters en technieken automatisch worden bepaald. Het onderhouden van gebruikersniveau is georganiseerd door het pro-actief assisteren van gebruikers door ervaren collega's, uitwisseling van recepten, stimulatie van gebruikersgroepen en reguliere opfris-sessies.

7.2.2 Datakwaliteit

Hoe beter de datakwaliteit, des te beter de resultaten van datamining. Het beheren van de kwaliteit van politiegegevens is een bekend en lastig probleem door de verscheidenheid aan mensen die de gegevens invoeren, de uiteenlopende omstandigheden waarin dit gebeurt en doordat het moeilijk is definities toe te passen bij het registreren van feiten. Dit weerhoudt datamining er niet van om goede resultaten te bereiken, maar het spreekt voor zich dat de resultaten beter zullen zijn wanneer de datakwaliteit toeneemt.

Het is belangrijk te constateren dat door de toepassing van datamining de datakwaliteit verbeterd kan worden door twee effecten: 1) alle gegevens worden gebruikt (wat de organisatie laat zien hoe belangrijk het is om alle gegevens juist te registreren) en 2) sommige patronen die worden gevonden met datamining illustreren dataproblemen. Bijvoorbeeld: een toename in een specifiek MO hoeft niet veroorzaakt te zijn door een toename van dat misdrijftype maar eerder door een veranderde manier van werken bij de mensen die de gegevens invoeren.

Er gelden twee principes bij het dataminingsysteem voor het aanpakken van datakwaliteitsproblemen :

1. Pak potentiële gegevensproblemen bij het creëren van een datawarehouse zo vroeg mogelijk aan door het koppelen van gegevensbronnen en door het toepassen van software die fouten corrigeert, bijvoorbeeld: simpelweg een stuk informatie verwijderen wanneer regels ontdekken dat het onjuist is.

2. Documenteer gegevensproblemen en maak deze documentatie eenvoudig toegankelijk binnen het dataminingsysteem.

7.2.3 Verwachtingsmanagement

Wij hebben ervaren dat soms te veel verwacht wordt van de impact van datamining. Dit maakt het belang duidelijk van heldere communicatie over wat verwacht kan worden. Datamining zal niet alle gegevensproblemen oplossen. Datamining maakt analisten niet overbodig. Datamining zal niet in staat zijn tekstuele informatie perfect te interpreteren. Datamining zal niet eenvoudigweg slagen in het samenbrengen van alle beschikbare gegevens gezien de privacywetten en de hoeveelheid inspanning die het vraagt om herhaalbare gegevensextracties te organiseren.

7.3 Toekomstig werk

Ook al is het dataminingsysteem al jaren operationeel, er komen steeds nieuwe ideeën om het systeem te verbeteren en op andere wijze toe te passen.

7.3.1 Meer voorspellingsmodellen

Momenteel worden voorspellingsmodellen gebruikt om te voorspellen waar misdrijven waarschijnlijk zullen plaatsvinden en om een trend of gedrag te verklaren. Veel andere toepassingen van voorspellingsmodellen worden op dit moment bestudeerd:

1. Criminele loopbaan: Wie maakt de grootste kans een veelpleger te worden?
2. Criminele profilering: Voorspelt het waarschijnlijke motief, leeftijd van de dader, geslacht etc., gebaseerd op een misdrijf of reeksen van misdrijven.
3. Wapenbezit : Hoe groot is de kans dat een persoon een wapen draagt, gebaseerd op het persoonlijk profiel en het verleden?
4. Huiselijk geweld: Hoe groot is de kans dat een gemelde situatie van huiselijk geweld uit de hand loopt? Dit om de beslissing te ondersteunen bij de vraag over extra aandacht aan de zaak.
5. Voorspelde criminaliteit gebaseerd op infrastructuur, gebouwen en socio-demografie in een buurt, door middel van wat-als experimenten.
6. Wat zijn de risico's voor een bepaald misdrijftype voor een bepaald gebied, gebaseerd op de eigenschappen van dat gebied en de situatie (tijd, dag, het weer)? Deze alternatieve benadering van geografische voorspelling (zie 5.2) is bruikbaar wanneer er te weinig voorbeelden zijn om die voorspelling uit te voeren.

7.3.2 Dashboard

De standaardrapporten die worden gegenereerd door het dataminingsysteem worden meestal verspreid in documentformaat, via de transformatie van XML naar HTML. Er wordt gewerkt aan het aanbieden van deze informatie in de vorm van een aanklikbaar dashboard dat start met een overzicht van de criminaliteit waarin informatie-onderdelen aangeklikt kunnen worden om op die manier in te zoomen op gebieden en misdrijftype, waardoor meer details worden getoond, zoals in kaarten en in trends. Op deze manier kan de rijkdom aan beschikbare gerapporteerde informatie effectiever worden ontsloten.

7.3.3 Tekst

Tekst is een belangrijke informatiebron binnen de politie, bijvoorbeeld omdat verklaringen meer details kunnen bevatten dan de informatie die ligt opgeslagen in de structurele gegevens. Het toegepaste dataminingsysteem kan omgaan met tekstuele informatie bij selecties en patroonanalyse. Echter, de Politie werkt nog aan een standaard voorziening die automatisch tekstuele informatie kan verzamelen uit de brongegevens. Een andere manier van werken met tekst is het extraheren van entiteiten (nummerplaten, telefoonnummers, adressen en namen) om ze vervolgens te gebruiken bij het koppelen van mensen aan zaken. Als dergelijke extractiesystemen van entiteiten in de toekomst worden geïmplementeerd in politiekorpsen, zullen we deze koppelinformatie toevoegen aan het dataminingsysteem.

7.4 Conclusie

In dit artikel hebben we laten zien dat datamining belangrijk is voor de politie en dat het tot een continue activiteit is gemaakt, goed uitvoerbaar door en voor politiemedewerkers zonder uitgebreide kennis van databases en datamining. Dit is gerealiseerd door een dataminingsysteem te ontwikkelen in samenwerking met de politie, met daarin samengebracht: interactie, visualisatie, tool-integratie, geautomatiseerde algoritmen, een groot en divers datawarehouse en gebruikersgemak. Vooral dat laatste is de belangrijkste sleutel tot succes gezien de combinatie van factoren die datamining voor de politie relatief moeilijk maakt: analytische diversiteit, sterke behoefte aan domeinexpertise, problemen met de datakwaliteit, en een gecompliceerde data-extractie. Het systeem heeft bewezen dat het meer diepte kan leveren in analyses en dat het een winst aan efficiëntie kan opleveren van een factor 20, wat resulteert in meer analytische capaciteit, kortere responstijden en de mogelijkheid interactief te analyseren gedurende groepsessies. Daarnaast zijn criminaliteitscijfers gedaald in veldproeven. Na acht jaar van gebruik, wordt het systeem nog voortdurend uitgebreid, gebaseerd op nieuwe ideeën, hetgeen laat zien dat er nog voldoende vooruitgang te boeken is in criminaliteitsanalyse.

8. REFERENTIES

- [1] Ratcliffe, J.H. 2003. Intelligence-led Policing, Australian Institute of Criminology, Canberra, Australia, 248.
- [2] Adderley, R. & P.B. Musgrove, P.B. 2001. Data Mining Case Study: Modeling the Behavior of Offenders Who Commit Serious Sexual Assaults, Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining.
- [3] McCue, C. 2007. Data mining and predictive analysis: Intelligence gathering and crime analysis, Butterworth-Heinemann.
- [4] Mena, J. 2003. Investigative Data Mining for Security and Criminal Detection, Elsevier Science (USA).
- [5] Brown D.E. 1998. The Regional Crime Analysis Program (RECAP): A framework for mining data to catch criminals, University of Virginia.
- [6] Bruin, J.S. de, Cocx, T.K., Kusters, W.A., Laros, J.F.J., Kok, J.N. 2006. Data Mining Approaches to Criminal Career Analysis, Proceedings of ICDM '06.
- [7] Uyl, M.J. den. 1986. Representing Magnitude by Memory Resonance, Proceedings of the 6th Annual Conference of the Cognitive Science Society, pp. 63-71.
- [8] Kohonen, T. 1984. Self-organisation and associative memory, Springer series in information sciences, Vol8. Springer Verlag, New York, USA.
- [9] Kohonen, T. 2001. Self-Organizing maps, Springer series in Information Sciences, Vol. 30, Springer, Berlin, Heidelberg, New York
- [10] Borg, I. & Groenen, P. 2005. Modern Multidimensional Scaling: theory and applications (2nd ed.), Springer/Verlag New York.
- [11] Farrell, G. 2005. Progress and prospects in the prevention of repeat victimization.
- [12] Eck, J.E., Chainey, S., Cameron, J.G., Letiner, M. & Wilson, R.E. 2005. Mapping crime: Understanding hot spots, Technical report 209393, National Institute of Justice.
- [13] Chainey, S. & Ratcliffe, J. 2005. GIS and Crime Mapping, Mastering GIS: Technology, applications and management, West Sussex, England: Wiley.
- [14] Clarke, R. V. & M. Felson, Eds. 2004. Routine Activity and Rational Choice (Advances in Criminological Theory), Transaction Publishers, New Brunswick (U.S.A).
- [15] Bowers, K.J., Johnson, S.D. & Pease, K. 2004. Prospective hot-spotting: The future of crime mapping?, Br. J. Criminol 44(5), pp. 641-658.
- [16] Rossmo, D.K. 2000. Geographic profiling, Boca Raton, Fla, CRC Press.
- [17] Freeman, L. 2004. The development of social network analysis, Vancouver, CA: Empirical Press.