TIJDRUIMTELIJK VOORSPELLEN VAN CRIMINELE INCIDENTEN

Een masterthesis waar tijdruimtelijke patronen in crimedata worden gezocht, om voorspellingen van incidenten in tijd en ruimte te verbeteren door gebruik te maken van statistische- en dataminingmethoden.

Amsterdam, 2014

Door

JITSKE SANNE DE GRAAUW

Supervisors

Dick Willems (Politie Amsterdam)
Rob van der Mei (Vrije Universiteit Amsterdam)
Evert Haasdijk (Vrije Universiteit Amsterdam)

4 november 2014

VRIJE UNIVERSITEIT AMSTERDAM

Deze pagina is bewust leeg gelaten.

TIJDRUIMTELIJK VOORSPELLEN VAN CRIMINELE INCIDENTEN

Master thesis Business Analytics

Amsterdam, 2014

Door

JITSKE SANNE DE GRAAUW

Supervisors

Dick Willems (Politie Amsterdam)
Rob van der Mei (Vrije Universiteit Amsterdam)
Evert Haasdijk (Vrije Universiteit Amsterdam)

Vrije Universiteit Amsterdam
MA Business Analytics
De Boelelaan 1105
1081 HV Amsterdam

Politie Amsterdam
Dienst Regionale Informatie
James Wattstraat 84
1097 DJ Amsterdam

4 november 2014

VRIJE UNIVERSITEIT AMSTERDAM

Deze pagina is bewust leeg gelaten.

Samenvatting

Wanneer de politie voorafgaand aan incidenten een inschatting heeft van wanneer en waar deze gaan plaatsvinden, kunnen politiepatrouilles veel doelgerichter en efficiënter te werk gaan. De Politie Amsterdam speelt daarop in door criminele incidenten te voorspellen met het Crime Anticipation System (CAS). CAS verdeelt regio Amsterdam in gebieden van 125x125 meter door het hanteren van een grid over heel Amsterdam. Voor ieder gebied voorspelt CAS de kans op een incident voor een tweewekelijkse peilperiode. De top 3% van de locaties met de hoogste kans op een incident wordt aangeduid als de high risk area en wordt gekleurd op een geografische kaart. Als aanvulling op deze tweewekelijkse kaarten worden op basis van de high risk area aparte kaarten gegenereerd per *weekdag, dagdeel* en *diensttijd* waarbij alleen de high risk area is herzien. Deze werkwijze leunt dus op de aanname dat de geografische verspreiding van incidenten identiek is voor alle onderliggende tijdsintervallen door het gebruik van dezelfde high risk area voor alle tijdsintervallen. Deze aanname is echter nooit theoretisch onderbouwd. Op basis van deze probleemstelling wordt onderzocht in hoeverre de huidige voorspellingen geschikt zijn om woninginbraken en straatroven in onderliggende tijdsintervallen te voorspellen en of dit beter kan dan met de huidige voorspellingen.

Het huidige CAS model kan $36,3\%$ van de woninginbraken en $57,7\%$ van de straatroven voorspellen op basis van het near hits criterium: incidenten zijn goed voorspeld of bijna goed voorspeld[1]. Voor zowel straatroven als woninginbraken is deze performance niet gelijkwaardig over alle onderliggende tijdsintervallen verdeeld. Bij woninginbraken overpresteert het dagdeel avond en bij straatroven kent het dagdeel nacht een veel hogere performance. Dit gaat voor beide gepaard gaat met een slechtere performance op de andere dagdelen. Verschillen in de performance worden mogelijk veroorzaakt door een afwijkende geografische verdeling, doordat vermoedelijk incidenten hebben plaatsgevonden op een andere locatie dan de voorspelling had verwacht. Wanneer tijdsintervallen met een afwijkende verdeling worden ontmaskerd, zou het theoretisch mogelijk moeten zijn voorspellingen van incidenten beter af te stemmen op de verdelingen van incidenten.

Onderzoek bewijst dat de geografische verdeling van zowel woninginbraken als straatroven afhangt van het tijdsinterval. Dit betekent dat incidenten op andere plekken gebeuren afhankelijk van het tijdsinterval en niet elke locatie een gelijke kans op een incident heeft in de gehele tweeweekse periode. Er worden twee duidelijke onderverdelingen in de tijd gevonden die een verschillende geografische verdeling van incidenten kennen: de dagdelen en het onderscheid in week en weekenddagen.

CAS kan voorspellingen genereren op basis van dagdelen en week- en weekenddagen die voor ieder gebied in Amsterdam de kans op een incident voorspeld specifiek voor het gedefinieerde tijdsinterval. Deze kleinere tijdsintervallen kunnen vervolgens worden samengenomen om de hele tweeweekse periode te

[1] Berekend op basis van de peilperioden 177 t/m 197.

omvatten. Tussen deze samengestelde modellen en het huidige CAS model kan geen verschil gevonden worden in performance.

Een belangrijk resultaat is dat bij het verkleinen van de tijdsintervallen het aantal te voorspellen incidenten afneemt. Deze afname in incidenten leidt tot minder verreikende incidenthistorie om nieuwe incidenten te voorspellen wat uiteindelijk wil leiden tot een slechtere performance. Er zijn duidelijke verschillen opgemerkt tussen de week- en weekenddagen en tussen de dagdelen, toch presteert het model dat beide incorporeert slechter dan de modellen die alleen naar dagdelen of alleen naar week- en weekenddagen kijken. Vermoedelijk ligt dit niet aan het feit dat de keuze voor deze tijdsintervallen slecht gekozen is, maar doordat het aantal te voorspellen incidenten en daarmee ook de incidenthistorie laag is. Het model is daarbij niet meer voldoende in staat de juiste patronen te extraheren. Een samengesteld model zal daarbij alleen in staat zijn de CAS performance te verbeteren, wanneer het onderscheidt in de geografische verdeling van incidenten voor beide tijdsintervallen dermate groot is dat dit opweegt tegen de vermindering in incidenthistorie.

De huidige CAS-kaarten kunnen op basis van dit onderzoek worden uitgebreid met specifieke dagdeel of week- en weekendkaarten die meer informatie bieden over de verdeling van incidenten in een kleiner tijdsinterval.

Sleutelwoorden: predictive policing, voorspellen van criminele incidenten, tijdruimtelijke voorspellingen, dataminingtechnieken, criminaliteits anticipatie systeem (CAS), logistische regressie, neuraal netwerk, bayes netwerk, ruimtelijke data analyse

Inhoudsopgave

Deze pagina is bewust leeg gelaten.

Hoofdstuk 1

Introductie

1.1 Aanleiding voor dit onderzoek

Wanneer de politie voorafgaand aan incidenten een inschatting heeft van waar en wanneer deze gaan plaatsvinden, kunnen politiepatrouilles veel doelgerichter en efficiënter te werk gaan. Het klinkt wellicht als toekomstmuziek, maar de eerste stappen in deze richting zijn al gemaakt. Uit analyse blijkt dat criminele incidenten niet volstrekt random plaatsvinden, maar dat tijdruimtelijke patronen te ontdekken zijn [4] [7]. Het ontmaskeren van deze patronen kan leiden tot een goede voorspelling van incidenten in de toekomst. Binnen het politiekorps Amsterdam houdt de afdeling datamining zich o.a. bezig met het voorspellen van incidenten. Hiervoor is het Crime Anticipation System (CAS) ontwikkeld dat voor iedere veertien dagen de kans op een type incident voor iedere gridlocatie[1] in regio Amsterdam voorspeld. Deze voorspellingen worden zichtbaar gemaakt op geografische overzichtskaarten en op die manier worden risicogebieden waarneembaar voor een periode van twee weken met betrekking tot een specifiek type incident. Als aanvulling op deze tweewekelijkse kaarten worden op basis van deze voorspelling 21 diensttijdkaarten gegenereerd waarop de kleuren zijn aangepast naar aanleiding van de incidentintensiteit van een specifieke 8-urige diensttijd, maar de geografische verdeling blijft identiek aan die van de tweewekelijkse voorspelling.

De tweewekelijkse voorspellingen vanuit CAS resulteren in een werkwijze waarbij operationele politiemedewerkers op ieder moment in deze tweeweekse periode dezelfde geografisch kaart met kansen raadplegen. Door de 21 diensttijdkaarten wordt wel inzicht verschaft in de incidentintensiteit tussen de 21 diensttijden maar wordt er vooralsnog gewerkt met één geografische verdeling. Hierbij wordt dus aangenomen dat deze geografische verdeling van de tweeweekse voorspellingen geschikt zijn voor alle onderliggende tijdsintervallen. Deze aanname is echter nooit theoretisch onderbouwd en toch vormen deze voorspellingen in de praktijk de basis voor het uitzenden van flexteams in 8-urige diensttijden. De politie Amsterdam wil daarom meer inzicht in de toepasbaarheid van deze voorspellingen op onderliggende tijdsintervallen, met daaruit voortkomend het doel: de huidige voorspellingen te verbeteren of aan te vullen met extra tijdsindicatieve modellen. Dit onderzoek zal zich richten op deze twee aspecten: (1)

[1] Binnen CAS is Amsterdam in gebieden van 125 × 125 meter verdeeld door het hanteren van een grid over heel Amsterdam. Zie paragraaf 1.2.1, model.

bepalen van de geschiktheid van de tweewekelijkse voorspellingen op onderliggende tijdsintervallen en (2) het onderzoeken van mogelijkheden om de huidige voorspellingen te verbeteren of aan te vullen met deze kennis.

De politie Amsterdam stelt hiervoor een database ter beschikking met alle incidenten van de afgelopen twintig jaar naar *incidenttype*, *locatie* en *tijd*. Daarnaast zijn omgevingskenmerken en CBS gegevens beschikbaar van de verschillende gridlocaties in Amsterdam. De gebruikte technieken en methoden van de huidige voorspellingen zijn eveneens beschikbaar. Deze dataset biedt mogelijkheden om de bruikbaarheid van de huidige voorspellingen te onderzoeken of te experimenteren met verschillende technieken om incidenten te voorspellen.

1.2 Achtergrond: predictive policing bij Politie Amsterdam

De afdeling datamining (politie Amsterdam) houdt zich bezig met het vinden van verbanden in grote hoeveelheden data om het verleden te kunnen beschrijven of juist de toekomst te voorspellen. Binnen dat kader is de afdeling twee jaar geleden begonnen met het voorspellen van criminele incidenten voor de regio Amsterdam, waaruit het Crime Anticipation System (CAS) is ontstaan: een datamining systeem dat criminele incidenten binnen Amsterdam voorspelt.

1.2.1 Crime Anticipation System

CAS staat voor het Crime Anticipation System en wordt in Nederlandse documenten ook wel aangeduid als Criminaliteits Anticipatie Systeem. Met CAS wordt gedoeld op het proces van data extractie, preparatie, het genereren van voorspellingen tot aan de daadwerkelijke weergaven van de output zoals geografische kaarten.

Oorsprong

Binnen de politie Amsterdam werd veelvuldig gewerkt met hotspot- en hottimesinformatie om inzichtelijk te maken waar en wanneer welke vorm van criminaliteit of overlast zich concentreert en werd meestal gebruikt om verwachtingen te onderbouwen [11]. Onder een hotspot wordt door Van Dijk, Van den Handel en Versteegh (2011) verstaan: *"een specifieke geografische locatie waar gedurende langere tijd en/of terugkerend sprake is van een hoge concentratie van criminaliteit"*[19]. Deze hotspots kunnen worden geplot op een geografische kaart en op die manier worden risicogebieden waarneembaar. Hierbij is het uitgangspunt dat patronen in data uit het verleden indicatief zijn voor toekomstige concentraties van criminaliteit. Echter is in de literatuur en ook binnen de politie geen eenduidige afgebakende definitie van een hotspot en geven analisten aan deze term een eigen draai waardoor bij een gelijke vraag verschillende hotspots worden aangemerkt. Daarnaast kent de hotspotmethodiek ook een interpretatieprobleem bij het bepalen van capaciteitallocatie en kunnen door verandering in constanten resultaten naar eigen hand worden gezet. In de zoektocht naar een generieke methode werd vanuit de afdeling datamining het idee aangedragen voor een voorspelmodel op basis van dataminingtechnieken. Na een succesvolle pilot werd CAS realiteit.

Model

Binnen CAS is de regio Amsterdam in gebieden van 125×125 meter verdeeld door het hanteren van een grid over heel Amsterdam. Op deze wijze ontstaan $196 \times 196 = 38.416$ (grid)locaties. Binnen deze grote groep locaties wordt een selectie gemaakt op stedelijk gebied[2] waardoor 'slechts' 11.500 relevante locaties overblijven ($\pm 30\%$). CAS baseert zijn voorspellingen op een grote hoeveelheid gegevens die per locatie worden gemeten: afstand tot bekende verdachten, afstand tot de dichtstbijzijnde snelwegoprit, soort en aantal bedrijven bekend bij de politie, demografische en socio-economische gegevens via het CBS. Daarnaast is een grote hoeveelheid criminaliteitshistorie bekend welke zijn gesommeerd voor de verschillende tijdsintervallen per twee weken, vier weken en half jaar voorafgaand aan de peilperiode.

Vanuit de input dataset wordt gezocht naar verbanden die indicatief zijn voor een verhoogde kans op een incident in de aankomende twee weken. Vanwege de complexiteit van zulke verbanden en de omvang van de dataset, wordt dit gedaan door een multi-layer perceptron (MLP). Een MLP is een neuraal netwerk (NN) dat data projecteert vanuit input nodes via een netwerk van neuronen op passende outputnodes. De aanduiding neuron is afgeleid van de neurons in ons zenuwstelsel. Wanneer zulke zenuwcellen voldoende geprikkeld zijn, versturen ze een signaal. Neuronen zijn dus bijzonder geschikt voor het ontvangen, verwerken en versturen van signalen. Neuronen binnen neurale netwerken zijn geïnspireerd op het gedrag van neuronen in de hersenen en kunnen aan elkaar worden gekoppeld en vervolgens stapsgewijs worden geoptimaliseerd. In iedere stap wordt informatie van een vakje aan het netwerk aangeboden en vervolgens wordt de uitkomst vergeleken met de daadwerkelijke feiten: heeft er ook in de twee weken na het peilmoment een incident plaatsgevonden? Deze uitkomst wordt vervolgens teruggekoppeld aan het netwerk en de neuronen zijn in staat daarop te reageren en verbindingen bij te stellen. Dit proces wordt backpropagation genoemd vanwege het achteraf bijstellen van de neuronen. Het leerproces kent dus een supervised leerproces doordat terugkoppeling vanuit de werkelijkheid het model bijstuurt waarbij het gebruik maakt van een niet-lineaire activatiefunctie. Als output wordt een kanswaarde tussen 0 en 1 per locatie voor de 2 weken naar het peilmoment bepaald. Binnen de huidige richtlijnen wordt een scheidingslijn getrokken na het 97ste percentiel, waardoor de top 3%[3] van de locaties wordt onderscheiden en aangeduid als de *high risk area*.

Bij deze werkwijze moet wel een kanttekening gemaakt worden. Het neuraal netwerk is begin dit jaar (2014) op de server is overgenomen door een logistisch regressie model. Deze overstap is doorgevoerd omdat de serverversie van CAS technische problemen kreeg met het genereren van een neuraal netwerk in SPSS Modeler. De oplossing zou liggen in een nieuwere versie van SPSS Modeler en deze wordt eind 2014/begin 2015 verwacht. Het is nog niet bekend of het systeem weer wordt ingericht met een neuraal netwerk of dat de logistische regressie behouden blijft. Over het algemeen wordt aangenomen dat de performances van beide modellen ongeveer gelijk zijn, maar dat is gabaseerd op de performance van één peilmoment waardoor robustheid tussen de methoden niet is onderzocht.

[2] Door deze selectie worden alle weilanden, open water, grasland etc. verwijderd uit de dataset.

[3] De keuze voor deze 3% ligt bij de hoeveelheid locaties die voor flexteams haalbaar zijn om te surveilleren in de tweeweekse periode.

Toepassingen

De kans op een incident per locatie wordt door middel van CAS voorspeld. Om een eenvoudige interpretatie aan deze grote hoeveelheid kansgegevens te geven, wordt de high risk area van locaties ingekleurd op een geografische kaart. Het inkleuren gebeurt aan de hand van drie kleuren die allen staan voor een specifiek percentiel: 98ste percentiel geel, 99ste percentiel oranje en het 100ste percentiel rood. Op deze manier worden de high risk areas eenvoudig ontmanteld en toepasbaar voor operationele teams zonder statistische kennis. De geografische kaarten die zo ontstaan worden aangeduid als *CAS-kaarten* en worden voornamelijk gebruikt voor de toekenning van operationele flexteams die Amsterdam breed worden ingezet. Daarnaast loopt er een pilot in het district Oost waar gebruik wordt gemaakt van een *CAS-kaart* gespecificeerd op het district. De kaarten worden iedere twee weken automatisch ververst en zijn beschikbaar via het interne politienetwerk.

Software

CAS draait voor een groot deel op IBM SPSS Modeler waarin de datapreparatie en modelleringstappen van CAS zijn ondergebracht. Het systeem wordt daarbij ondersteund door een ORACLE database die toegankelijk is voor het wegschrijven of ophalen van data. Na de modelleringstappen worden de geografische kaarten gecompileerd door middel van MapInfo. Wanneer gerechtigd, zijn deze kaarten via het interne netwerk te laden.

1.2.2 Crime Anticipation System op diensttijdniveau

De huidige CAS-voorspellingen worden gegenereerd op basis van een peilperiode van twee weken, echter speelde bij de operationele politieteams steeds meer de vraag naar gedetailleerdere voorspellingen op basis van de verschillende diensttijden in de week. Aan de hand van die vraag is CAS uitgebreid met voorspellingen op diensttijdniveau.

Model

De high risk locaties (top 3% van de locaties met de hoogste kans op een incident) worden gebruikt als input voor een kohonen clustering en logistische regressie met tijdsvariabelen. Op die manier wordt geprobeerd voor de locaties in de high risk area een nieuwe voorspelling te doen die voor de 21 diensttijden in de week een onafhankelijke kans op een incident voorspelt. Op basis van deze voorspellingen worden wederom CAS-kaarten gegenereerd, waarbij de voorspellingen voor alle diensttijden samen worden genomen en in drie terciles wordt geschaald die opeenvolgend de kleuren geel, oranje en rood krijgen. Deze modelleringstap levert dus 21 additionele kaarten op die inzicht geven in de kans op een incident in elk van de 21 diensttijden, waarbij de top 3% van de locaties opnieuw beoordeeld is op basis van een specifieke diensttijd.

Schaduwzijde

Het bovenstaand model leunt op de aanname dat verschillende tijdsvensters voor een type incident een verschillende intensiteit kennen maar de geografische verspreiding identiek is. Dit omdat voor alle tijds-

vensters dezelfde top 3% locaties met de hoogste kans op een incident als uitgangspunt zijn genomen. Het model leidt dus tot het genereren van CAS-kaarten op diensttijdniveau, waar op iedere kaart dezelfde vakjes gekleurd zijn, namelijk de aanvankelijk ingestelde 3%. Het enige zichtbare verschil tussen twee kaarten is het gebruik van kleuren per vakje die gebaseerd zijn op intensiteit. Het zou theoretisch gezien wel mogelijk zijn op basis van kleurafwijkingen een vorm van geografisch verspreiding te zien wanneer deze zich extreem differentieert van de andere tijdsvensters. De aanname dat incidenten geografisch gelijk verdeeld zijn op alle diensttijden is nooit onderbouwd, waardoor de vraag of deze modellen wel van toegevoegde waarde zijn in twijfel kan worden getrokken.

1.3 Doelstelling

Het doel van dit project is om te onderzoeken *in hoeverre de huidige voorspellingen geschikt zijn om incidenten in onderliggende tijdsintervallen te voorspellen die mogelijk een afwijkende geografische voorspelling hebben.* Daarbij speelt de ruimtelijke verdeling van incidenten over de tijd een grote rol. Wanneer incidenten over de tijd ruimtelijk gezien gelijk verdeeld zijn, kan de aanname dat de huidige voorspellingen geschikt zijn op onderliggende tijdsintervallen worden onderbouwd. Wanneer grote verschillen of trends plaatsvinden binnen deze periode kunnen kaarten specifiek gebonden aan een kleiner tijdsinterval mogelijk een betere indicatie geven van de kansen op een incident. Deze probleemstelling leidt tot de volgende drie onderzoeksvragen:

1. *In hoeverre zijn de huidige tweewekelijkse voorspellingen geschikt om gehanteerd te worden op onderliggende tijdsintervallen?*
2. *In hoeverre zijn incidenten ruimtelijk gezien gelijk verdeeld t.a.v. verschillende onderliggende tijdsintervallen?*
3. *Hoe kan met gebruik van algoritmen de kans op een incident voor iedere gridlocatie m.b.t. een specifiek tijdsinterval worden voorspeld?*

De eerste vraag focust op de huidige stand van de voorspellingen met betrekking tot de probleemstelling, terwijl de tweede vraag zich richt op het verkrijgen van meer inzicht in de verdeling van incidenten over de tijd. De derde vraag combineert beide en onderzoekt de mogelijkheden voor het verbeteren of aanvullen van de huidige CAS omgeving op basis van de huidige voorspellingen (vraag 1) en het verkregen inzicht in de ruimtelijke verdeling van incidenten (vraag 2). Binnen dit onderzoek worden twee type incidenten besproken: *woninginbraken* en *straatroven.* Daarnaast richt het onderzoek zich alleen op incidenten die zijn geregistreerd op basis van aangifte bij de politie Amsterdam.

1.4 Structuur van rapport

De rapport gaat verder met een kort overzicht van de beschikbare literatuur en theoretisch kader waarbinnen het onderzoek plaatsvindt in hoofdstuk 2. Hoofdstuk 3 volgt met een toelichting en bespreking van de beschikbare data en geeft daar de benodigde achtergrondinformatie bij. Het beantwoorden van

de onderzoeksvragen vindt plaats in de hoofdstukken 4 t/m 7. Hoofdstuk 4 begint met het onderzoeken van de toepasbaarheid van de huidige voorspellingen op onderliggende tijdsintervallen om deelvraag één te beantwoorden. Hoofdstuk 5 geeft antwoord op de tweede deelvraag door in te gaan op de ruimtelijke verdelingen die de onderliggende tijdsintervallen kennen. De derde vraag wordt beantwoord aan de hand van twee hoofdstukken die beiden de mogelijkheden onderzoeken van het voorspellen van incidenten in kleinere tijdsintervallen: hoofdstuk 6 dat doet door middel van CAS en in hoofdstuk 7 worden andere technieken toepast. De conclusie en aanbevelingen sluiten het rapport af in hoofdstuk 8.

Hoofdstuk 2

Literatuuronderzoek

In dit hoofdstuk wordt een overzicht gegeven van de beschikbare literatuur gerelateerd aan de centrale vraag binnen dit onderzoek: het voorspellen van criminele incidenten in tijd en ruimte. Door dit literatuuronderzoek wordt een breder perspectief geboden waarbinnen dit onderzoek tot stand is gekomen en daarnaast worden technieken en methoden besproken die in gelijke of verwante onderzoeken bruikbaar zijn gebleken. Achtereenvolgens wordt predictive policing toegelicht (2.1), de voorspelbaarheid van incidenten (2.2), technieken om incidenten te voorspellen (2.3) en tot slot tijdruimtelijke verdelingen (2.4).

2.1 Predictive Policing

Het voorspellen van criminele incidenten valt in zijn geheel onder *predictive policing*. Perry et al. (2013) omschrijft predictive policing als: "*predictive policing is the application of analytical techniques - particularly quantitative techniques - to identify likely targets for police intervention and prevent crime or solve past crimes by making statistical predictions*" [14].

In de 19de eeuw begon Quetelet (1835) al met (statistisch) onderzoek naar de relatie tussen sociale factoren en de crime rate om zo de crime rate in de nabije toekomst te voorspellen [15]. De echte doorbraak van predictive policing is echter pas begonnen na de Tweede Wereldoorlog met de opkomst van de computer en samenhangende toepassingsgebieden zoals datamining en artificial intelligence. Daarbij speelde ook de extreme toename in criminaliteit eind jaren '60 een rol in zowel Europa als de Verenigde Staten [10]. De predictive policing methoden kunnen grofweg worden onderverdeeld in vier categorieën [14]:

1. Voorspellen van incidenten: het voorspellen van tijd en plaats van incidenten in de toekomst.
2. Voorspellen van toekomstige misdadigers: het voorspellen van risico's van latere toetreding tot criminaliteit.
3. Voorspellen van identiteit van daders: profielen van misdadigers matchen.
4. Voorspellen van slachtoffers: het voorspellen en identificeren van groepen of individuele slachtoffers.

Dit onderzoek richt zich alleen op de eerste categorie: het voorspellen van tijd en plaats van criminele incidenten in de toekomst. De overige categorieën worden niet verder toegelicht.

2.2 Wat maakt criminele incidenten voorspelbaar?

Predictive policing is gebaseerd op het idee dat incidenten niet volstrekt random gebeuren. Een voorbeeld van een goed te voorspellen incident is een woninginbraak. Wanneer in een huis wordt ingebroken, neemt de waarschijnlijkheid van een inbraak in dat huis en de nabijgelegen huizen in de opeenvolgende dagen toe. Vaak wordt het tegenovergestelde verwacht, bedenkend dat ze al slachtoffer zijn geweest, de kans op herhaling kleiner is. Incidenten zoals moord of verkrachting zijn daarentegen veel moeilijker te voorspellen: ze komen minder vaak voor en de crime scene van zulke incidenten is niet stationair zoals een huis. Hiermee wordt bedoeld dat bij een moord of verkrachting de samenkomst tussen dader en slachtoffer niet stationair is en daarom veel moeilijker te voorspellen zijn dan een inbraak waar een samenkomst tussen dader en een stationair huis voorspeld wordt.

Cohen (1979) en Block et al. (1987) bewijzen dat criminele incidenten niet volstrekt random plaatsvinden en tijdruimtelijke patronen gevonden kunnen worden [4] [7]. Daarnaast komen Figlio en Sellin (1972) [23] met een onderzoek waaruit blijkt dat een klein deel aan veelplegers verantwoordelijk is voor een groot porportie aan incidenten waardoor de voorspelbaarheid toeneemt. Jeff Brantingham, antropoloog van de University of Californa en betrokken bij het predictive police project van de Los Angeles Police Department, zegt het volgende:

The naysayers want you to believe that humans are too complex and too random - that this sort of math cant be done . . . but humans are not nearly as random as we think. . . . In a sense, crime is just a physical process, and if you can explain how offenders move and how they mix with their victims, you can understand an incredible amount.[16]

De visie van Brantingham sluit aan bij de meeste criminele gedragstheorieën zoals de *routine activity theory, rational choice theory* en de *crime pattern theory.* De routine activity theory (Cohen & Felson, 1979) gaat ervan uit dat een incident bestaat uit drie aspecten: (1) een gemotiveerde pleger, (2) een passend target en (3) de afwezigheid van een bewaker (eventueel politie, buren etc.) [7]. Met deze theorie wordt de aanwezigheid van de bewaker gezien als negatief effect op een mogelijke crime. Oftewel, de aanwezigheid van bijvoorbeeld wetshandhavers op de juiste plek kan criminele incidenten voorkomen.

De rational choice theory (Cornish & Clarke, 1987) ziet een crimineel incident als een costs functie, waarbij de winst significant groter moet zijn dan de kosten en waarbij de pleger alleen denkt aan zijn eigen belangen [9]. Deze theorie geeft onderbouwing en inzicht in de motivatie van de pleger.

De crime pattern theory (Brantingham & Brantingham, 1984) ziet een crimineel incident als een complexe gebeurtenis die pas ontstaat wanneer aan een grote hoeveelheid voorwaarden is voldaan [5]:

1. Criminelen en slachtoffers volgen beide een levenspatroon en pas als deze patronen elkaar overlappen in tijd en ruimte ontstaat een toegenomen kans op een incident.

2. Het criminele incident is in tegenstrijd met de strafwet.

3. Het target is toegankelijk.

4. De afwezigheid van middelen en personen die mogelijk kunnen interfereren met de actie of strafrechtelijke gevolgen kunnen vergemakkelijken.

5. Een gemotiveerde dader die rationele keuzes kan maken.

Deze theorie leidt tot het inzicht dat al deze aspecten niet onmogelijk gelijkmatig in tijd en ruimte kunnen samenkomen en daarmee dat incidenten nooit random in tijd en ruimte kunnen plaatsvinden. Deze theorieën en aannamen passen bij de meeste incidentsoorten zoals inbraak, straatroof en overvallen. Zoals al eerder aangegeven (paragraaf 2.2) zijn type incidenten zoals verkrachtingen en moorden moeilijker te voorspellen. Voor deze incidenten zijn dan ook andere frameworks ontwikkelt die in dit onderzoek niet besproken worden.

2.3 Voorspellen van criminele incidenten

Dit onderzoek focust op voorspellingen van criminele incidenten: wanneer en waar is de kans op een incident het hoogst. Binnen de politie wordt momenteel al veel gefocust op *waar* incidenten plaatsvinden en worden met betrekking op *wanneer* ingedeeld in tweewekelijkse tijdsperioden. In dit onderzoek zal het *wanneer* gedetailleerder worden onderzocht ten aanzien van de *waar*.

De meeste methoden om incidenten te voorspellen, baseren zich op historie van criminele incidenten. Hierbij wordt dus de aanname gemaakt dat recentelijk plaatsgevonden incidenten gelden als voorgeschiedenis op de nog te gebeuren incidenten in de nabije toekomst. Vrijwel alle methoden werken op basis van dit principe, al is de context waarin deze aanname wordt geplaatst vaak anders. Drie type methoden worden gesommeerd door Perry et al. (2013) [14]:

- Hotspot analyses, datamining technieken, near-repeat methoden en statistische regressie worden over het algemeen gebruikt om de *waar* te identificeren van incidenten over een gegeven tijdsinterval.
- Tijdruimtelijke methoden worden gebruikt om de *wanneer* te identificeren van incidenten.
- Risico terrein analyses worden gebruikt om ruimtelijke factoren te identificeren die o.a. op basis van historie de kans op een type incident verhogen (*waar*).

Veel methoden worden ook gebruikt om op basis van de kennis van de waar en/of wanneer ook de *wie* te ontmaskeren. Binnen dit onderzoek wordt geen nadruk gelegd op wie de mogelijke plegers zijn en is dus in dit literatuuronderzoek buiten beschouwing gelaten. Dit hoofdstuk gaat verder met een korte toelichting per bovengenoemde methode.

2.3.1 Hotspot analyses

Eén van de meest populaire methoden om incidenten te voorspellen is het *hotspot model*. Het idee van crime hotspots wordt geïntroduceerd als crime mapping methode door Sherman, Gartin & Buerger (1989). Hierbij worden hotspots gezien als weergave van het verleden en niet als voorspelling voor de toekomst.

Het hotspotmodel als forecast methode door Block (1995) baseert zich volledig op de stelling dat waar incidenten gaan gebeuren, waar ook de incidenten in het verleden gebeurd zijn [3]. Criminele incidenten uit het verleden worden geclusterd over de ruimte ontstaan de zogeheten hotspots. Er zijn in de loop der tijd meerdere modellen ontwikkelt om hotspots te ontmaskeren, zoals ruimtelijke histogrammen die geprojecteerd kunnen worden op een grid, eclipse covering methoden, scan statistieken en kernel dichtheid verwachtingen. Hotspotmodellen hebben als nadeel dat ze zich alleen baseren op de huidige patronen en niet in staat zijn inzicht te geven in de relatie tussen incidenten en omgeving over de tijd heen. Als aspecten in de omgeving veranderen, kan het hotspotmodel daar niet op anticiperen. Ondanks deze nadelen blijft het hotspotmodel onverminderd populair, doordat deze relatief makkelijk te implementeren is en de totstandkoming van de output eenvoudig te begrijpen is.

2.3.2 Regressiemodellen

Regressiemodellen zoeken een wiskundig verband tussen een uitkomstvariabele (bijv. wel of geen incident) en de responsvariabelen. Waar hotspotmodellen zich alleen richten op de historie aan incidenten, kunnen regressiemodellen alle gewenste variabelen gebruiken om mee te nemen in de te generen modellen. Hierdoor wordt de kans op een incident in de toekomst niet alleen gebaseerd op de historie maar ook op eventueel andere (significant) afhankelijke variabelen, zoals aantallen huizen of type inwoners ten aanzien van een specifiek ruimtelijke locatie. Voor de toepassing van regressiemodellen wordt het ruimtelijke aspect ingedeeld in areal locaties (zoals buurten) of verdeeld door het hanteren van een grid over een regio. Gebruikte regressiemodellen zijn lineaire regressie, non-lineare regressie of regressie splits waarbij meerdere regressiemodellen worden gecombineerd.

2.3.3 Datamining technieken

Regressiemodellen zijn wiskundige modellen die in staat zijn voorspellingen te maken op basis van inputdataset. De generalisatie van wiskundige modellen die in staat zijn voorspellingen te maken op basis van een inputdataset wordt doorgaans aangeduid als dataminingmodellen. Een duidelijke definitie van *datamining* wordt gegeven door Statsoft:

"Datamining is an analytic process designed to explore data (usually large amounts of data - typically business or market related - also known as "big data") in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. The ultimate goal of data mining is prediction - and predictive data mining is the most common type of data mining and one that has the most direct business applications."[1]

Over het algemeen kan gesproken worden van classificatie- en clusteringmethoden. Bij classificatiemethoden worden de te voorspellen kansen opgedeeld in intervallen (categorieën) en wordt de meest waarschijnlijke categorie toegekend. Bij clustering methoden worden records onderverdeeld in groepen met gelijke kenmerken. Op basis van die groepen en het die in het verleden als 'hotspot' werden aangemerkt

[1] Definitie van http://www.statsoft.com/Textbook/Data-Mining-Techniques

kunnen nieuwe hotspots worden toegekend. Naast deze technieken kunnen de modellen ook weer worden gecombineerd door middel van ensemble methoden om een uiteindelijke voorspelling te verbeteren.
Dataminingtechnieken zijn toegepast op een brede range aan toepassingsvelden en hebben al een prominente plaats ingenomen als forecast methode. Zo kunnen enkele voorbeelden van een geslaagde implementatie van dataminingtechnieken worden genoemd. Neurale netwerken worden in meerdere artikelen als goede schatters van criminele incidenten aangewezen [8] [13]. Oatley & Ewart (2003) voorspellen incidenten op korte termijn door gebruik te maken van een Bayes Net (classificatie methode) [13]. Binnen de politie Amsterdam wordt gebruik gemaakt van een logistisch regressie model of neuraal netwerk (op basis van een multilayer perceptron) om criminele incidenten te voorspellen.

2.3.4 Near repeat modellen

Near repeat modellen baseren zich op de aanname dat een toekomstig incident in tijd en plaats kort na een gebeurd incident plaatsvindt. Er zijn meerdere studies die deze aanname onderbouwen, al lijkt dit fenomeen het sterkst aanwezig bij woninginbraken. Townsley et al. (2000) ondervinden een 18,7% repeat rate voor woninginbraken in Beenleigh, Australie [18]. Deze repeat rate betekent dat in 18,7% van de woninginbraken werd gevolgd door een nieuw woninginbraak in korte tijd en op korte afstand van de vorige woninbraak. Ook Mohler (2012) komt met gelijke conclusies voor woninginbraken en ontwikkelt op basis van deze kennis een model dat lijkt op een aardbeving model [12]. In een aarbeving model zorgt een aardbeving voor naschokken, maar in het model van Mohler triggert een incident eventuele opvolgende incidenten.

2.3.5 Tijdruimtelijke methoden

Alle bovenstaande modellen gaan uit van variabelen die tijd, plaats en historie kennen. *Tijdruimtelijke methoden* gaan een stapje verder: de voorspelling wordt uitgebreid met de correlatie tussen *tijd* en *ruimte*.

Wang & Brown (2012) presenteren een gridbased tijdruimtelijk model door tijdruimtelijke variabelen toe te voegen aan de dataset waarop het model zich baseert [20] [21]. CAS is ontwikkelt en geïnspireerd op basis van dit model. Wang & Brown gebruiken echter een generalized addictive model (GAM) om de daadwerkelijke voorspelling te genereren, terwijl CAS gebruik maakt van een logistische regressie. Een GAM is een generalized lineair model (GLM) waarbij de uitkomstvariabele lineair afhangt van smooth functies van de responsvariabelen. Een GLM is de generalisatie van een ordinaire lineaire regressie, waarbij de error een verdeling kan hebben anders dan de normale verdeling. De GAM presenteren Wang & Brown zelf als Spatio-Temporal GAM (ST-GAM) om de toevoeging van tijdruimtelijke variabelen kenbaar te maken. Daarnaast laten ze zien dat het gridbased model met ST-GAM betere voorspellingen genereert dan het hotspotmodel.

De reden dat binnen de politie Amsterdam geen gebruik is gemaakt van een ST-GAM, maar wel van logistische regressie is puur een softwarebeperking. Binnen de politie wordt SPSS Modeler gebruikt die geen GAM of GLM ondersteunt. Er is ook zover bekend, geen reden waarom een GAM of GLM beter incidenten kan voorspellen dan het huidige logistische regressiemodel.

2.3.6 Risico terrein modellen

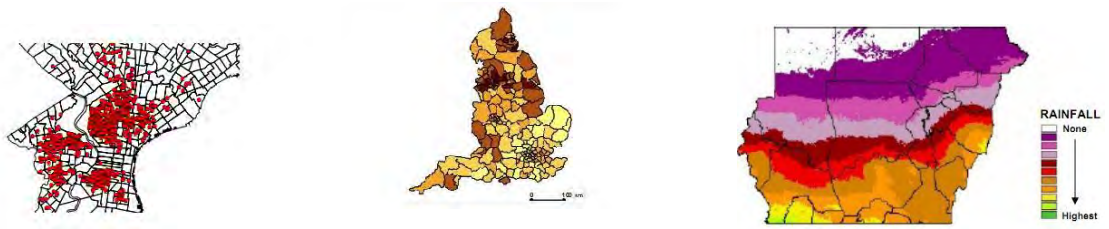
Risico terrein modellen worden ontwikkelt om ruimtelijke factoren te identificeren die de kans op een type incident verhogen. Met risico terrein modellen worden aparte map layers voor iedere risicofactor gegenereerd. Een combinatie van al die layers vormt een risico terrein map. Caplan & Kennedy (2011) presenteren een onderzoek waarbij risico terreinmappen succesvol worden toegepast op crime data [6]. Het voordeel van een risico terreinmap is het inzicht in de factoren die bijdragen aan het risico op een crimineel incident. Daarnaast kunnen risico terrein modellen op basis van trends in factoren hotspots aanwijzen die mogelijk in de toekomst een verhoogd risico kennen zonder dat deze gebieden eerst slachtoffer zijn geworden. Daarbij zal het model wel voldoende factoren moeten bevatten zodat geen cruciale aspecten gemist worden.

2.4 Tijdruimtelijke verdelingen

Naast het voorspellen van incidenten wordt ook geprobeerd meer inzicht te krijgen in de correlatie tussen criminele incidenten en tijdspatronen. Daarvoor zullen twee of meer ruimtelijke patronen van criminele incidenten moeten worden vergeleken of bestudeerd worden. Technieken die hiervoor gebruikt kunnen worden vallen onder *ruimtelijke data-analyse* die zich richt op het toepassen van statistische methoden om informatie te extraheren uit data met een ruimtelijk aspect.

2.4.1 Ruimtelijke datastructuren

Ruimtelijke data-analyse baseert zich op het analyseren van data die ruimtelijke patronen beschrijft. Technieken om ruimtelijke patronen te vergelijken en te analyseren zijn afhankelijk van de datastructuur, oftewel de manier waarop het tijdruimtelijke patroon in data beschikbaar is. Er zijn meerdere datastructuren om ruimtelijke patronen te beschrijven. De drie meest voorkomende zijn *ruimtelijk puntproces, areal ruimtelijke data* en een *continue ruimtelijke data* [17]. Grafische voorbeelden van deze datastructuren zijn weergegeven in 2.1.



Figuur 2.1: Voorbeelden van spatial point pattern (links), areal spatial data (midden) en continuous spatial data (rechts)

In de incidentdatabase zijn incidenten als punt in de ruimte beschikbaar: ieder incident kent een locatie die als punt in de ruimte kan worden weergegeven. De verzameling van zulke tijdruimtelijke punten wordt een *ruimtelijk puntproces* genoemd. Datapunten in de ruimte kunnen vervolgens worden toegekend aan regio's zoals gridlocaties, wijkteams of districten zoals in de huidige CAS omgeving gebeurd wanneer incidenten worden toegekend aan gridlocaties. Op die manier ontstaat een *areal ruimtelijke datastructuur* waarbij

de meetwaarden zijn geaggregeerd per area/gebied. CAS genereert voorspellingen op het niveau van deze areal ruimtelijke datastructuur waarbij voor iedere gridlocatie een kans wordt voorspeld. Een *continu ruimtelijke datastructuur* baseert zich op een ruimtelijk patroon waar voor elke locatie een specifieke meetwaarde kan worden toegekend. In het kader van criminele incident data komt dat overeen met een datastructuur waar voor alle locaties de criminaliteitsintensiteit bekend is. De continue data kan worden gecreëerd door interpolatie van een ruimtelijk punt proces.

De data bij de politie Amsterdam kan dus worden getransformeerd tot één van de drie bovenstaande datastructuren. In dit onderzoek zullen echter alleen ruimtelijke puntprocessen en areal ruimtelijke datastructuren gebruikt worden. Om de data te transformeren tot een continue datastructuur door middel van interpolatie vergt veel tijd en brengt een error met zich mee door het gebruik van interpolatie op een punt proces.

2.4.2 Ruimtelijke analysetechnieken

Incidenten kunnen worden toegekend aan een specifiek tijdsinterval waarbinnen het incident heeft plaatsgevonden. Op die manier ontstaan verschillende ruimtelijke patronen voor ieder tijdsinterval. Deze ruimtelijke patronen zullen met de juiste analyse technieken vergeleken moeten worden om verschillen of juist overeenkomsten aan het licht te brengen.

Ruimtelijke puntprocessen

Om twee ruimtelijke punt processen te vergelijken kunnen de *random shift test* en *random labeling test* worden gebruikt [2]. Deze twee testen werken beide op basis van de cross K-functies als toetsingsgrootheid, waardoor deze test zich eerder uitlaat over de mate van clustering dan over de verdeling van incidenten. Deze methode is niet in staat bij toepassing op complexe verdelingen van criminele incidenten een verschil aantoonbaar te maken. Dit kan worden laten zien door een uitleg van de cross K-functies als toetsingsgrootheid bij deze twee testen. Neem twee punt processen, 1 en 2, die een intensiteit λ_1 en λ_2 kennen. De K-functie voor populatie 1 ten opzichte van populatie 2 kan worden opgesteld als:

$$K_{1,2}(h) = \frac{1}{\lambda_2} E(aantal\ incidenten\ in\ afstand\ h\ van\ een\ willekeurig\ incident\ i) \tag{2.1}$$

In beide testen wordt deze functie omgeschreven naar een sample cross K-functie waarbij ook het totale aantal incidenten wordt meegenomen. Desondanks kan worden gesteld dat deze methode, toegepast op patronen waarbij meerdere clusters van incidenten vindbaar zijn, moeilijk een verschil kan vinden. De toetsingsgrootheid gebaseerd op de cross K-functie met een bereik van 0 tot 1 respectievelijk aantrekking en afstoting, loopt met ingewikkeldere patronen al snel naar een waarde van 0,5. Deze waarde geeft patroon aan waarbij geen verschil in afstoting of aantrekking tussen de patronen gevonden kan worden.

Areal ruimtelijke data

De methoden om areal ruimtelijke datapatronen te vergelijken zijn veelzijdiger en makkelijker toepasbaar dan de methoden van ruimtelijke puntprocessen. Ieder ruimtelijk puntproces kan daarentegen ook wor-

den getransformeerd tot een areal ruimtelijke datastructuur, waardoor de beschreven technieken breed toepasbaar zijn.

Smith (2014) beschrijft de *quadrat methode* voor het testen van spatial randomness [17]. Deze methode verdeeld alle punten van een ruimtelijk punt proces in gridlocaties, de zogeheten quadrats (waardoor een areal ruimtelijke datastructuur ontstaat). Wanneer ruimtelijke randomness geldt, is het totaal aantal punten in een quadrat onafhankelijk en poisson verdeeld. Deze hypothese kan getest worden met gebruik van de Pearson χ^2 goodness-of-fit-test, waarbij het aantal verwachte incidenten in elke cel wordt gegeven door het gemiddelde van de bovenliggende poissonverdeling. Deze methode kan worden hergebruikt waarbij twee ruimtelijke verdelingen worden vergeleken. In principe ontstaan wanneer het aantal incidenten in de verschillende quadrats wordt vergeleken onder meerdere tijdsintervallen, twee categorische variabelen: het tijdsinterval en het quadrat. De χ^2 goodness-of-fit-test is een toets om parametervrij na te gaan of twee of meerdere verdelingen, bestaande uit twee categorische variabelen, van elkaar verschillen [22]. De χ^2 goodness-of-fit-test kan dus ook worden gebruikt voor het vergelijken van aantallen incidenten in districten of wijken (in plaats van quadrats). Een belangrijke nadeel is wel de gevoeligheid van de methode onderhevig aan de keuze van de geografische afbakening. Dat binnen de politie twee wijken worden onderscheiden op naam, betekent niet dat de verdelingen van criminele incidenten zich aan deze grens houdt.

Andersen (2009) beschrijft een *nonparametrische Monte Carlo benadering* die door middel van sampling twee ruimtelijke areal verdelingen vergelijkt [1]. Door één verdeling te kiezen als base verdeling worden uit de andere dataset herhaaldelijk 85% van de incidenten gesampled. Op basis van deze gesampelde verzameling worden percentages per area berekend waar een betrouwbaarheidsinterval uit wordt opgesteld. De percentages uit de base set worden vervolgens getoetst aan de hand van de betrouwbaarheidintervallen om gelijkenis per area vast te stellen. Andersen geeft daarnaast een methode om deze uitkomsten grafisch weer te geven. Deze Monte Carlo benadering is wel onderhevig aan pieken in specifieke gebieden wanneer de regio wordt opgedeeld in een klein aantal gebieden.

2.5 Toepassing literatuuronderzoek

Dit onderzoek baseert zich op de kennis opgedaan vanuit de literatuur en samengevat in dit literatuuronderzoek. Daarnaast is deze theoretische achtergrond gebruikt om een kader te scheppen waarbinnen dit onderzoek plaatsvindt. De methodes en technieken die worden beschreven in dit literatuuronderzoek worden grotendeels toegepast om antwoord te vinden op de gestelde onderzoeksvragen. Hierbij is een selectie gemaakt op methoden en technieken die geïmplementeerd kunnen worden in SPSS Modeler 14.2 omdat de Politie Amsterdam met deze software voorspellingen genereert.

In SPSS Modeler zijn veel algemene technieken beschikbaar zoals neurale netwerken, Bayes net, regressie modellen, beslisbomen en diverse clusteringalgoritmen. Versie 15 (en verder) bevat plug-ins om processen uit te voeren in statistische programma's zoals R en Matlab maar deze zijn in SPSS Modeler 14.2 nog niet beschikbaar. In dit onderzoek zijn alleen technieken en methoden gebruikt die in SPSS Modeler 14.2 beschikbaar zijn of eenvoudig geïmplementeerd konden worden. Dit resulteert in het gebruik van methoden die worden onderbouwd door literatuur en zijn toegepast in SPSS Modeler.

Hoofdstuk 3

Achtergrond

In de hoofdstukken 4 t/m 7 wordt geprobeerd een antwoord te vinden op de gestelde onderzoeksvragen met behulp van databronnen die beschikbaar zijn gesteld door de politie Amsterdam. Vanuit deze databronnen vindt een data-extractie- en datapreparatie proces plaats waarin interpretatie van de begrippen incident, tijd en locatie van belang zijn. In dit hoofdstuk wordt die achtergrondinformatie gegeven. Paragraaf 3.1 gaat over de definitie van criminele incidenten gevolgd door twee paragrafen over de wijk en districtsstructuur (3.2) en peilperioden (3.3). Paragraaf 3.4 beschrijft de volledige dataset en paragraaf 3.5 sluit af met een overzicht van de gebruikte performance measures.

3.1 Incidenten

Voor de ontwikkeling van CAS wordt gebruik gemaakt van diverse databronnen die beschikbaar zijn in een ORACLE database. Door verschillende datapreparatie stappen wordt de data gevormd tot een input dataset om voorspellingen op te baseren. Deze databronnen zijn eveneens beschikbaar om antwoorden te vinden op de gestelde onderzoeksvragen binnen de kaders van dit onderzoek.

De *incident_actie* is de meest cruciale tabel waarop dit onderzoek is gebaseerd. In deze tabel zijn alle gemelde incidenten en acties verzameld. Alle records binnen deze tabellen relateren aan een specifieke actie of incident die allemaal een locatie, begindatum, begintijd, einddatum, eindtijd en een maatschappelijke klasse kennen. De keyvariabele binnen deze tabel is het incidentnummer (INC_ACT_ID) waarop incidenten worden geregistreerd. De maatschappelijke klasse, datum/tijd en locatie van incidenten worden in de volgende drie subparagrafen toegelicht.

3.1.1 Maatschappelijke klasse

Ieder incident wordt gekoppeld aan een maatschappelijke klasse die bepaald in welke klasse het incident valt. Voorbeelden van maatschappelijke klassen zijn bijvoorbeeld brandstichting, fietsendiefstal en joyriding. De maatschappelijke klasse werkt als overkoepelende segmentatie, want bij het aanmaken van een incident kunnen ondersteunende velden het incident verder in kaart brengen. Door het gebruik van deze klasse kan een dataselectie worden gemaakt op specifieke incidenten. Binnen dit onderzoek wordt gekeken naar twee verschillende type incidenten: woninginbraken (WIB), en straatroven (SRF). Deze typen

Maatschap. klasse	Beschrijving	Incidenttype	Beschrijving
A20	Gekwal. diefstal in/uit woning	WIB	Woninginbraak
A30	Diefstal in/uit woning (niet gekwal.)	WIB	Woninginbraak
B20	Gekwal. diefstal met geweld in/uit woning	WIB	Woninginbraak
B30	Diefstal met geweld in/uit woning (niet gekwal.)	WIB	Woninginbraak
B40	Zakkenrollerij/tassenrollerij met geweld	SRF	Straatroof
B70	Straatroof	SRF	Straatroof

Tabel 3.1: Overzicht van maatschappelijke klassen naar incidenttype

incidenten vallen onder meerdere maatschappelijke klassen. Tabel 3.1 geeft de lijst van maatschappelijke klassen weer die worden meegenomen binnen dit onderzoek naar type incident. Er zal in dit onderzoek alleen gesproken worden over de samenvattende incidenttypes en daarbij wordt geen onderscheidt gemaakt naar de onderliggende maatschappelijke klasse. Woninginbraak en straatroven worden binnen de politie allebei gekenmerkt als high impact crimes en kennen daarom een hoge prioriteit.

3.1.2 Datum en tijd

Alle incidenten kennen twee datumvelden die een incident in de tijd plaatst: een einddatum en begindatum. Aan deze data hangen tabellen die een veelzijdige hoeveelheid informatie bieden over de specifieke datum, zoals feestdagen en planperiodes (m.b.t. dienstroosters) waarin de dag valt. De tijd wordt bijgehouden in twee tijdsvelden: een begintijd en een eindtijd. Er geldt: $begindatum{+}tijd \leq einddatum{+}tijd$. Dit betekent concreet dat incidenten geen specifiek pleegmoment kennen maar een interval binnen de tijd waarop een incident heeft plaatsgevonden: zie formule 3.1.

$$pleeginterval = [begindatum + tijd : einddatum + tijd] \tag{3.1}$$

Het is mogelijk dat het begindatumtijd en einddatumtijd gelijk aan elkaar zijn. Dit komt bijvoorbeeld voor bij straatroven omdat het slachtoffer zich vaak bewust is van het exacte moment waarop de straatroof heeft plaatsgevonden. Een woninginbraak heeft vaak plaatsgevonden in een groter tijdsinterval waardoor een specifiek tijdsmoment lastiger te bepalen is. Soms zijn slachtoffers op vakantie en komen er bij thuiskomst pas achter dat een woninginbraak heeft plaatsgevonden. Door deze grote tijdsintervallen is het exacte moment waarop de inbraak heeft plaatsgevonden nauwelijks meer te achterhalen. Het exacte pleegmoment zal dus bij veel incidenten geschat moeten worden. Het schatten van de exacte pleegdatumtijd gebeurt door het rekenkundig gemiddelde van de begindatumtijd en de einddatumtijd te nemen zoals formule 3.2 aangeeft.

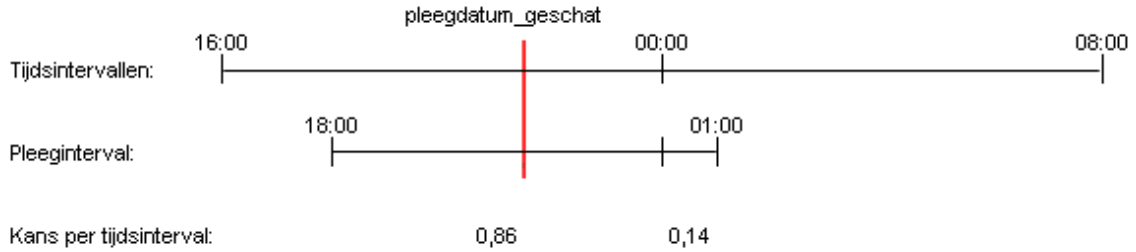
$$pleegdatumtijd\_geschat = \frac{(begindatum + tijd) + (einddatum + tijd)}{2} \tag{3.2}$$

Een voorbeeld van twee incidenten waarbij het pleeginterval en geschatte pleegdatum worden berekend is te zien in tabel 3.2. Naast de berekening op basis van de formules 3.1 en 3.2 wordt ook de lengte van het pleeginterval berekend in dagen.

Inc.	Begindatum	-tijd	Einddatum	-tijd	Pleeginterval	Pleegdatum-tijd_geschat	Pleeginterval duur in dagen
1	11-04-2014	10:00	11-04-2014	16:00	[11-04 2014 10:00 : 11-04-2014 16:00]	11-04-2014 13:00	0,25
2	11-04-2014	18:00	12-04-2014	01:00	[11-04 2014 18:00 : 12-04-2014 01:00]	11-04 2014 21:30	0,29

Tabel 3.2: Voorbeeld van incidenten met een geschatte pleegdatumtijd en pleeginterval

In het algemeen kan gezegd worden dat hoe groter het pleeginterval hoe onbetrouwbaarder de geschatte pleegdatum en des te groter de error tussen de geschatte pleegdatum en de werkelijke pleegdatum. Wanneer echter de incidenten worden toegekend aan een vast tijdsinterval[1] om geanalyseerd te worden op basis van de geschatte pleegdatum, is het mogelijk dat deze error zich verkleint tot 0 wanneer het gehele pleeginterval zich bevind in het vastgestelde tijdsinterval. Dit betekent dus dat de error van deze geschatte pleegdatumtijd zich verhoudt tot de gekozen tijdsintervallen, waarbij het dus van belang is dat incidenten aan de juiste tijdsintervallen worden toegekend. Een voorbeeld van een toekenning van een incident op basis van de geschatte pleegdatum is te zien in figuur 3.1. Wanneer wordt aangenomen dat de kans op een incident binnen het pleeginterval uniform verdeeld is, wordt de kleinste error gevonden bij toekenning aan het eerste tijdsinterval: een kans van 0,14 dat het incident verkeerd is toegekend ten opzichte van een kans van 0,86 dat het incident juist is toegekend. De keuze voor een tijdsinterval van 8 uur is niet willekeurig. Binnen dit onderzoek wordt voornamelijk gewerkt met tijdsintervallen van 8 uur aangezien deze overeenkomen met de diensttijden (zie paragraaf 3.3).



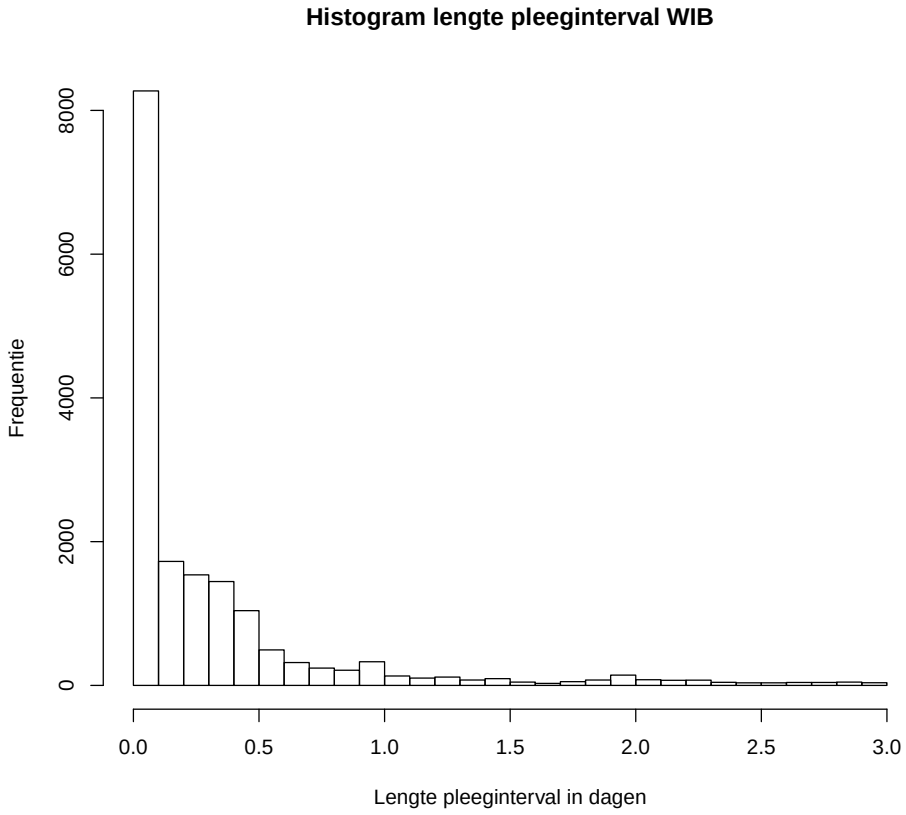
Figuur 3.1: Voorbeeld: Incident 2 uit tabel 3.2 uitgezet tegen vaste tijdsvensters van 8 uur.

Figuur 3.2 geeft de verdeling weer van de lengte van de pleegintervallen van woninginbraken waarbij een selectie is gemaakt op pleegintervallen die kleiner of gelijk zijn aan 3 dagen. Wanneer wederom wordt aangenomen dat de kans op een incident binnen het pleeginterval uniform verdeeld is en de incidenten worden toegekend aan tijdsintervallen van 8 uur, worden alle incidenten met een pleeginterval kleiner dan $0,67$ ($\frac{2\times 8}{24} = 0,67$) aan een tijdsinterval toegekend met een kans $> 0,5$ op juist toekennen[2]. Op basis van data over 2013 en deels 2014[3] valt 89,77% van de woninginbraken binnen dit criterium.

[1] Het is ook mogelijk pleegdatumtijd te gebruiken voor tijdsreeksen of punt proces modellen. Hierbij worden de incidenten als punt in de tijd gezet. Modellen waarbij dat het geval is worden binnen dit onderzoek niet gehanteerd.

[2] De enige uitzondering hierop zijn incidenten met een pleeginterval waarbij de geschatte pleegdatum exact op de grens van een tijdsinterval valt.

[3] Betreft heel 2013 en de maanden januari t/m juli van 2014.



Figuur 3.2: Histogram van de lengtes van het pleeginterval van woninginbraken

In de huidige CAS omgeving worden woninginbraken met een interval van meer dan 2 dagen verwijderd uit de dataset. Bij deze vergelijking wordt alleen gekeken naar de begindatum en einddatum. Wanneer einddatum - begindatum < 2 wordt het incident behouden en daarbij wordt de tijd buiten beschouwing gelaten. Dit is een haalbare aanname wanneer incidenten worden toegekend aan tweewekelijkse perioden, maar in dit onderzoek wordt gebruik gemaakt van tijdsintervallen die 21 keer zo klein zijn. Op basis van de bovenstaande analyse worden incidenten met een pleeginterval groter dan $\frac{2}{3}$ dag verwijderd uit de dataset omdat deze niet met voldoende overtuigingskracht aan een vast tijdsinterval van 8 uur kunnen worden toegekend. Deze aanname is doorgevoerd om analyses te kunnen uitvoeren over de incidenten die zuiver tot een specifiek interval behoren en is toegepast op zowel woninginbraken als straatroven.

3.1.3 Locatie

Binnen CAS is Amsterdam in gebieden van 125 x 125 meter verdeeld door het hanteren van een grid over Amsterdam. Dit levert 196 x 196 = 38.416 gebieden. Dit hele onderzoek baseert zich eveneens op dit gehanteerde grid om twee redenen: (1) op basis van deze gridlocaties zijn CBS gegevens en politiegegevens over bekende veelplegers beschikbaar en (2) om aan te sluiten bij de huidige modellen is het gebruik van dit grid gewenst. Dit betekent dat incidenten moeten worden gekoppeld aan de juiste locatie in het grid en daarnaast zal ook de relevantie van alle gridlocaties worden onderzocht. Deze twee aspecten worden achtereenvolgens behandeld.

Adresregistratie

Alle incidenten in de tabel *incident_actie* zijn gekoppeld aan een adres waarvan de rijksdriehoekscoordinaten bekend zijn. In principe vult de beschikbare databasemart zich met data vanuit de regio Amsterdam, maar wanneer aangifte wordt gedaan in Amsterdam met betrekking tot een incident dat buiten regio Amsterdam heeft plaatsgevonden, wordt deze wel opgenomen in de databasemart. Incidenten moeten dus gefilterd en geschaald worden naar de coördinaten van het grid dat over Amsterdam wordt geplaatst. Dit gebeurt op basis van de rijksdriehoekscoördinaten RC_X en RC_Y die kunnen worden geschaald naar de gridcoördinaten $XCOR$ en $YCOR$. Dit proces gebeurt aan de hand van de formules 3.3 en 3.4.

$$XCOR = \lfloor \frac{RC_X - 106000}{125} \rfloor + 1 \tag{3.3}$$

$$YCOR = \lfloor \frac{RC_X - 470000}{125} \rfloor + 1 \tag{3.4}$$

Op basis van de verkregen variabelen $XCOR$ en $YCOR$ wordt vervolgens een selectie gemaakt op incidenten die hebben plaatsgevonden in de regio Amsterdam; namelijk de incidenten die in het grid vallen (zie formule 3.5).

$$0 \leq COR_X \leq 196,\ 0 \leq COR_Y \leq 196 \tag{3.5}$$

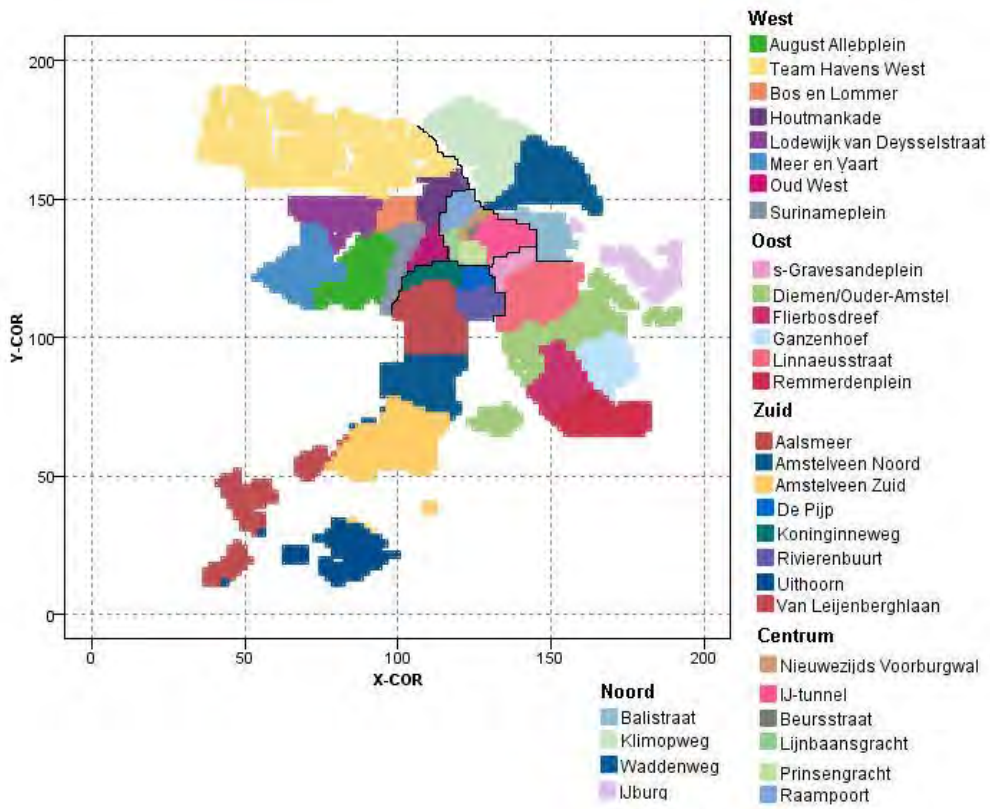
Deze werkwijze resulteert in incidenten die een locatie kennen genoteerd als een combinatie van $XCOR$ en $YCOR$ als referentie aan één gridlocatie.

Selectie van locaties binnen grid

Amsterdam wordt verdeeld in gebieden van 125 x 125 meter door het hanteren van een grid wat 196 x 196 = 38.416 locaties genereert. Binnen deze grote groep gridlocaties wordt een selectie gemaakt van locaties die toebehoren tot het stedelijk gebied van Amsterdam. Dit betekent dat locaties die bestaan uit weiland of open water worden verwijderd uit de dataset, omdat incidenten (woninginbraken en straatroven) doorgaans plaatsvinden in stedelijk gebied. Ook worden alle gridlocaties die een park bevatten verwijderd uit de dataset. Dit wordt gedaan in verband met de wijze waarop het adres van een park is geregistreerd. Ieder park kent één adres waarop alle incidenten die in het park hebben plaatsgevonden worden geregistreerd. Dit betekent dat alle incidenten in één park worden geregistreerd op één adres dat maar gekoppeld kan worden aan één gridlocatie. Als voorbeeld het Vondelpark in Amsterdam: op één gridlocatie worden alle incidenten sommeert die hebben plaatsgevonden in het Vondelpark, terwijl de incidenten daadwerkelijk hebben plaatsgevonden over een oppervlakte van 20 gridlocaties. Door deze wijze van adresregistratie zijn alle locaties die uitsluitend uit park bestaan verwijderd uit de dataset. Als laatste zijn ook de locaties toebehorend tot de wijk Havens West verwijderd uit de dataset. In deze wijk wonen slecht 185 mensen (CBS, 2013) en bestaat voornamelijk uit een industriële haven en is daarom als niet relevant bestempeld voor onderzoek naar woninginbraken en straatroven. Het totale aantal locaties waar dit onderzoek zich op baseert, bestaat door het verwijderen van landelijk gebied, water, park en de wijk Havens West uit 9.376 gridlocaties.

3.2 Districten en wijken

Gridlocaties worden gebruikt als ruimtelijke polynoom waar een incident aan kan worden toegekend (zie paragraaf 3.1.3). In hoofdstuk 5 wordt echter ook gewerkt met wijken en districten als ruimtelijke polynoom waar incidenten aan worden toegekend. In dat geval worden de gridlocaties gebruikt als sleutel voor het toekennen van districten en wijken. Amsterdam bestaat uit 5 districten die weer zijn verdeeld in 32 wijken[4]. Van de 32 wijken is alleen de wijk Havens West niet opgenomen in dit onderzoek (zie paragraaf 3.1.3). Figuur 3.3 geeft de wijken per district weer. De onderverdeling van de districten is bij naastgelegen wijken afgebakend met een zwarte lijn. De wijk Havens West is voor de volledigheid opgenomen in dit totaaloverzicht. Alle witte delen op de kaart zijn niet weergegeven doordat deze bestaan uit open water, weilanden of niet toebehoren aan regio Amsterdam.



Figuur 3.3: Verdeling van wijken en districten in Amsterdam

De toekenning van de gridlocaties aan de wijken en districten gebeurt door de centroides van de locaties te koppelen aan een wijk en district. Deze werkwijze betekent dat het kan voorkomen dat een incident wordt toegekend aan een wijk waarbinnen het incident niet is gevallen, doordat de centroide van de gridlocatie in een andere wijk valt dan het specifieke incident binnen de gridlocatie. Echter kan over het algemeen worden aangenomen dat incidenten juist worden toegekend.

[4]Dit is de onderverdeling ten tijde van dit onderzoek. Deze structuur is mogelijk onderhevig aan reorganisatie van de district- wijkteams in de toekomst

3.3 Peilperioden en onderliggende tijdsintervallen

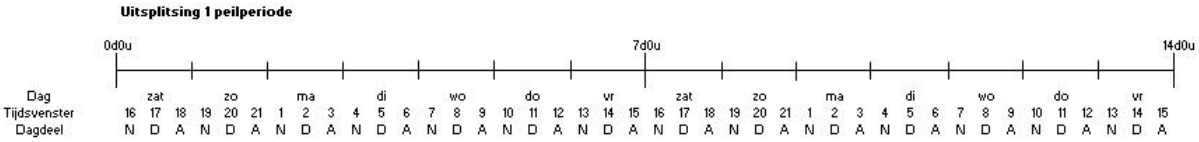
In paragraaf 3.1.2 werd uitgelegd dat ieder incident (o.a.) een pleeginterval kent waarin het incident heeft plaatsgevonden. Wanneer een pleeginterval langer is dan $\frac{2}{3}$ dag, kan een incident niet meer met een kans groter dan 0,5 worden toegekend aan een interval van 8 uur (voor berekening zie paragraaf 3.1.2). De keuze om hier een interval van 8 uur te nemen is niet volstrekt willekeurig. Deze 8 uur zijn afgeleid van de acht uur durende diensten die de operationele politieteams kennen. Een verzameling van 42 diensten behoren weer tot een roosterperiode (BVCM periode genoemd). Eén peilperiode loopt van zaterdag t/m de vrijdag twee weken later. Omdat er politieteams zijn die worden ingepland op basis van de voorspellingen die over de tweewekelijkse peilperioden worden gemaakt, zijn de peilperioden gelijk gesteld aan de roosterperioden. Tabel 3.3 geeft alle roostertechnische peilperioden weer die binnen dit onderzoek relevant zijn.

Peilperiode	Data	Peilperiode	Data
177	05-10-2013 t/m 18-10-2013	188	08-03-2014 t/m 21-03-2014
178	19-10-2013 t/m 01-11-2013	189	22-03-2014 t/m 04-04-2014
179	02-11-2013 t/m 15-11-2013	190	05-04-2014 t/m 18-04-2014
180	16-11-2013 t/m 29-11-2013	191	19-04-2014 t/m 02-05-2014
181	30-11-2013 t/m 13-12-2013	192	03-05-2013 t/m 16-05-2014
182	14-12-2013 t/m 27-12-2013	193	17-05-2014 t/m 30-05-2014
183	28-12-2013 t/m 10-01-2014	194	31-05-2014 t/m 13-06-2014
184	11-01-2014 t/m 24-01-2014	195	14-06-2014 t/m 27-06-2014
185	25-01-2014 t/m 07-02-2014	196	28-06-2014 t/m 11-07-2014
186	08-02-2014 t/m 21-02-2014	197	12-07-2014 t/m 25-07-2014
187	22-02-2014 t/m 07-03-2014		

Tabel 3.3: Relevante peilperioden/roosterperioden binnen de politie

Voor iedere tweewekelijkse peilperiode berekent CAS voor iedere gridlocatie in Amsterdam de kans op een incident. Deze voorspellingen worden vervolgens gebruikt om de kansen voor alle 21 verschillende diensttijden te genereren (zie paragraaf 1.2.2). Elke peilperiode bestaat uit 14 dagen, waarbij iedere dag bestaat uit een nacht- dag en avonddienst. Dit levert 42 diensten per peilperiode, maar diensten waarbij dag en dagdeel overeenkomen worden beschouwd als een 'gelijke' dienst. Elke werkweek bestaat zo uit 21 verschillende diensten en elke peilperiode weer uit twee weken die in totaal 21 verschillende diensten kennen die allen 2 keer voorkomen. Figuur 3.4 geeft deze onderverdeling van één peilperiode van 14 dagen grafisch weer. De belangrijkste onderliggende tijdsintervallen zijn daarbij de diensten, weekdagen en dagdelen.

Op basis van de peilperioden uit tabel 3.3 worden politieteams ingepland en daarom wordt ook binnen dit onderzoek aan deze perioden vastgehouden. Alle peilperioden kunnen daarbij worden opgesplitst in: weekdagen, dagdelen, diensttijden. Door onderzoek te doen naar deze vooraf gedefinieerde tijdsintervallen wordt geprobeerd inzicht te krijgen in geografisch verdeling van incidenten in onderlig-



Figuur 3.4: Onderliggende tijdsintervallen van één peilperiode

gende tijdsintervallen van de overkoepelende twee weken durende peilperiode.

3.4 Input dataset

Voorspellingen van incidenten worden gegenereerd op basis van een input dataset. De basis hiervoor is een record per tijdsinterval-gridlocatie waaraan responsvariabelen worden gekoppeld die mogelijk correleren met het wel of niet plaatsvinden van een incident. Alle responsvariabelen worden gekoppeld via het tijdsinterval en/of de gridlocatie. Aan de gridlocatie worden wijk, district, CBS gegevens, bedrijfsinformatieve gegevens en veelpleger informatie gekoppeld. Aan het tijdsinterval in combinatie met de gridlocatie worden variabelen toegevoegd die de historie van criminaliteit vastleggen. Tot slot wordt de uitkomstvariabele toegevoegd: heeft er uiteindelijk wel of geen incident plaatsgevonden. De volledige lijst met variabelen is in bijlage A gegeven.

3.5 Performance

Voor het evalueren en vergelijken van verschillende voorspellingen in het tijdruimtelijke vlak is een meetwaarde nodig die de performance van een voorspelling kan uitdrukken. Binnen de huidige onderzoeken op het gebied van tijdruimtelijke voorspellingen is geen standaard metriek beschikbaar. Ondanks dit gebrek worden er in de literatuur en intern bij de politie meetwaarden gebruikt die de performance van voorspellingen schatten. Achtereenvolgens worden de relatieve en absolute performance besproken.

3.5.1 Relatieve hits performance

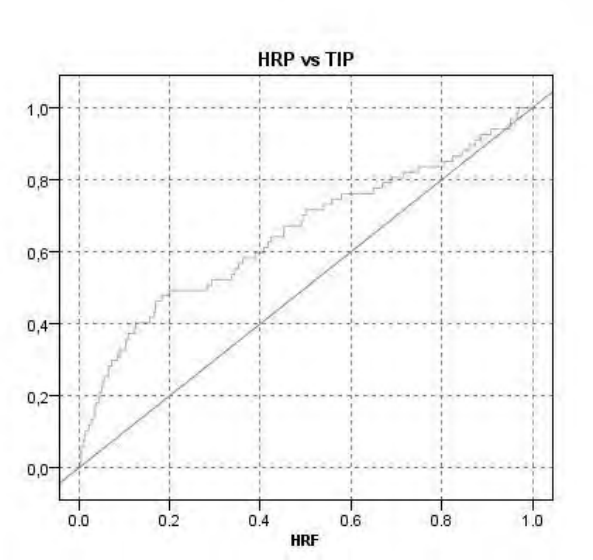
Wang & Brown (2011) vangen performance van modellen in twee criteria: (1) een goed model zou op locaties waar incidenten daadwerkelijk plaatsvinden, een hoge kans op een incident moeten voorspellen; (2) het totale gebied waar hoge kansen worden voorspeld zal klein moeten zijn op ieder willekeurig tijdstip. Op basis van deze twee criteria zijn de volgende performance meetwaarden bepaald:

$$HRP_\delta = \frac{||\{s_i|p(inci_{s_i,t_j} = 1) > \delta\}||}{||\{s_i\}||} \tag{3.6}$$

$$TIP_\delta = \frac{||\{inci_{s_i,t_j} = 1|s_i \subset \{s_i|p(inci_{s_i,t_j} = 1) > \delta\}\}||}{||\{inci_{s_i,t_j} = 1\}||} \tag{3.7}$$

waar $||\cdot||$ de lengte van een vector met getallen is en δ een drempelwaarde. HRP staat voor het percentage locaties die als high risk zijn aangemerkt ten aanzien van het totaal aantal locaties. TIP geeft het percentage incidenten binnen de high risk locaties aan. Wanneer de vectoren HRP en TIP

worden berekend voor verschillende waarden van δ en deze tegen elkaar geplot worden ontstaat een curve zoals in figuur 3.5.



Figuur 3.5: HRF geplot tegen TIP

Hoe beter de modellen performen hoe meer de curve in de richting van de linkerbovenhoek gaat, omdat logisch volgt dat bij deze modellen de meeste incidenten gebeuren in de high risk gebieden met een beperkte omvang. Door de loop van deze curve, kan de oppervlakte onder de curve (AUC, area under curve) worden aangenomen als single performance meetwaarde gebaseerd op zowel de HRP als de TIP. Deze performance measure wordt aangeduid als *relatieve hits performance* RHP:

$$RHP = AUC(HRP_\delta, TIP_\delta) \tag{3.8}$$

De relatieve hitsperformance kan alleen gebruikt worden om twee voor voorspellingen die hetzelfde aantal incidenten voorspellen en daarom niet toepasbaar om twee verschillende patronen van incidenten te vergelijken. Dit komt doordat ieder plot een verschillende maximale mogelijke RHP kan halen. Wanneer een voorspelling wordt gemaakt waarbij in iedere gridlocatie een incident plaatsvindt, kan de RHP nooit een andere waarde aannemen dan 0,5. Dit kan dus niet vergeleken worden met een voorspelling waarbij uiteindelijk één incident heeft plaatsgevonden en de RHP bijna 1 kan aannemen. Een RHP van precies 1 kan alleen worden gehaald wanneer geen incidenten hebben plaatsgevonden.

3.5.2 Absolute hits performance

Binnen de politie Amsterdam wordt gewerkt met een absolute hits performance die is opgedeeld in absolute hits en absolute near hits performance. Deze performance measure is eenvoudiger en beter interpreteerbaar dan de relatieve hits performance. De absolute performance measure gaat uit van een percentiel p_k die binnen de politie gelijk wordt gesteld op p_3 (3%).

Formule 3.9 geeft de *absolute hits performance* weer. Deze formule berekent het proportionele aantal incidenten dat valt in de 3% (k%) van de locaties met de hoogste kans op een incident (de high risk area) tav het totale aantal incidenten.

$$Hits_k = ||\{inci_{s_{x,y},t_j} = 1 | s_i \subset \{s_{x,y} | p(inci_{s_{x,y},t_j} = 1) > p_k\}\}|| \tag{3.9}$$

De *absolute near hits performance* is gebaseerd op de high risk area. De locaties omliggend aan de high risk area wordt aangeduid als near hits locaties. Iedere kaart kent 282 high risk locaties die de high risk area vormen. Het aantal near hits locaties kan dus oplopen tot maximaal 2.256 locaties (24,06% van het totale aantal locaties) en maakt in combinatie met de high risk locaties dat maximaal $27,07\%$ van de locaties kan worden opgenomen als high risk of near hits locaties. In de praktijk zal echter blijken dat veel high risk locaties clusteren waardoor het aantal near hits locaties vaak tussen de 1.000 en 1.500 ligt.

De absolute near hits performance is toepasbaar voor het vergelijken van verschillende verdelingen van incidenten. Doordat het percentage van het totale aantal incidenten hierin leidend is, zal bij minder incidenten de performance ook sneller omhoog gaan bij het goed voorspellen van een incident dan wanneer er veel incidenten voorspelt moeten worden. Doordat ongeacht het aantal incidenten wordt gewerkt met een 100% score kunnen verschillende voorspellingen met een ander onderliggend aantal incidenten worden vergeleken, al wordt daarbij niet het verschil in voorspelbaarheid van de onderliggende verdelingen in acht genomen.

In de volgende hoofdstukken zal het opvallen dat veel modellen die worden vergeleken een significant verschil kennen in hits performance maar wel een gelijke near hits performance kennen. Hierbij speelt parten dat de toekenning van near hits random gebeurd en de onderliggende voorspelling van deze locaties geen rol speelt. Daarbij is dus de kans om een incident juist te voorspellen door middel van de near hits performance groter, wanneer weinig incidenten door de hits performance juist zijn voorspeld. De hits performance is dus leidend wanneer op het moment dat een vergelijking door middel van de absolute performances wordt gemaakt.

Hoofdstuk 4

Toepassing van de huidige voorspellingen op onderliggende tijdsintervallen

CAS voorspelt in de huidige omgeving voor een peilperiode van twee weken de kans op een incident per gridlocatie. Deze tweewekelijkse voorspellingen worden ook gebruikt voor het genereren van voorspellingen op basis van weekdag, dagdeel en diensttijd (zie paragraaf 1.2.2 en 3.3), al worden over deze onderliggende tijdsintervallen binnen de politie geen performances gemeten. Het is dus niet bekend of er specifieke tijdsintervallen zijn die extreem afwijken van de gemiddelde performance. Een afwijking in performance kan mogelijk veroorzaakt worden door een afwijkende ruimtelijke verdeling en is daarom in het kader van dit onderzoek interessant. In dit hoofdstuk wordt de performance van de voorspellingen met betrekking tot de onderliggende (kleinere) tijdsintervallen weekdag, dagdeel en diensttijd onderzocht, om eventuele afwijkende ruimtelijke verdelingen te ontmaskeren. De centrale vraag binnen dit hoofdstuk is:

In hoeverre zijn de huidige tweewekelijkse voorspellingen geschikt om gehanteerd te worden op onderliggende tijdsintervallen?

Dit hoofdstuk vervolgt met een beschrijving van de methode gevolgd door de resultaten voor woninginbraken en straatroven en tot slot de conclusie.

4.1 Methode

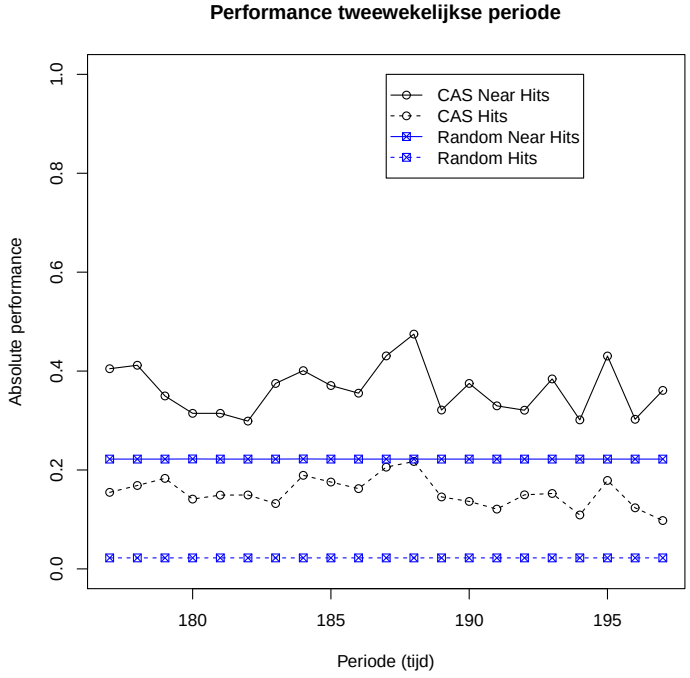
De tweewekelijkse voorspellingen worden toegepast op de onderliggende tijdsintervallen: *weekdagen, dagdelen* en *diensttijden*. Om de geschiktheid te meten worden de voorspellingen getoetst aan de incidenten die binnen een specifiek tijdsinterval hebben plaatsgevonden. Het doel hierbij is kijken of specifieke tijdsintervallen onder- of bovengemiddeld presteren. Een afwijkende prestatie van een specifiek tijdsinterval geeft aanleiding tot het aannemen dat incidenten binnen dat tijdsinterval mogelijk een afwijkende ruim-

telijke verdeling kennen. Daarnaast geeft het toetsen van de voorspellingen op kleinere tijdsintervallen inzicht in de toepasbaarheid van de voorspellingen op onderliggende tijdsintervallen en kan uitsluitsel worden gegeven of deze voorspellingen inderdaad breed over de kleinere tijdsintervallen kunnen worden toegepast. De analyses in dit hoofdstuk baseren zich op voorspellingen die tot stand zijn gekomen met de CAS Flexteam versie 1.0 van 24 januari 2013. Er zijn voorspellingen gemaakt voor de peilperiodes 177 t/m 197. Om de performance van de voorspellingen te kwantificeren is gebruik gemaakt van de absolute hits performance (zie paragraaf 3.4).

Mann-Whitney toets - De Mann-Whitney toets wordt gebruikt om te toetsen of twee performance verdelingen gelijk verdeeld zijn. De Mann-Whitney toets is non-parametrisch en niet gevoelig voor verschil in variantie omdat de waarnemingen op basis van rangorde worden vergeleken. Voor het gebruik van de Mann-Whitneytoets moeten minimaal 20 meetwaarden beschikbaar zijn.

4.2 Performance woninginbraken

Binnen de politie Amsterdam wordt gewerkt met de absolute (near)hits performance measure om de performance van de CAS voorspellingen te kwantificeren (paragraaf 3.4). Deze measure wordt toegepast op de tweewekelijkse voorspellingen en berekent achteraf op basis van de plaatsgevonden incidenten in de periode de performance van de voorspelling. In het algemeen wordt vaak gesproken over een performance van 35 tot 40% wanneer men spreekt over woninginbraken. In dat geval wordt gedoeld op de absolute nearhits performance. Deze kijkt naar het percentage incidenten dat heeft plaatsgevonden in de top 3% van de locaties met de hoogste kans op een incident (hit) of in een direct naastgelegen vakje (near hit).



Figuur 4.1: Performance CAS op basis van de woninginbraken die plaatsvinden op tweewekelijkse basis

Figuur 4.1 geeft de absolute hits en near hits performance weer op basis van de voorspellingen van CAS over een tweewekelijkse periode. De gemiddelde near hits performance over deze periode is $0,3632$ ($\sigma = 0,0496$) en de hits performance is gemiddeld $0,1544$ ($\sigma = 0,0304$). In figuur 4.1 is ook de performance van een random kans generator weergegeven om de voorspellingen te vergelijken met een random trekking. Deze random trekking is tot stand gekomen door iedere locatie een kans op een incident toe te kennen op basis van een trekking uit de uniforme verdeling $[0, 1]$. Hierbij worden alleen de 9.376 locaties meegenomen die CAS ook meeneemt (paragraaf 3.2), locaties die dus door CAS zijn uitgesloten omdat ze bijvoorbeeld alleen open water of een park bevatten, krijgen geen kans toebedeeld. Voor iedere peilperiode worden 200 voorspellingen gedaan waar iedere locatie bij iedere voorspelling een random kans krijgt toegekend. Iedere voorspelling kent daarbij een performance, waarvan het gemiddelde wordt gehanteerd als de random performance voor een specifieke peilperiode.

Conclusie - Het huidige CAS model heeft voor woninginbraken een gemiddelde near hits performance van $36,3\%$ en een gemiddelde hits performance van $15,4\%$.

4.2.1 Performances naar weekdagen

De tweewekelijkse voorspellingen worden ook gebruikt voor het genereren van dagkaarten waarover binnen de politie geen performances worden gemeten. Figuur 4.2 laat de performances van de voorspellingen zien wanneer alleen gekeken wordt naar de daadwerkelijke incidenten op de betreffende weekdagen.

Het lijkt alsof alle performances zich redelijk verhouden tot de tweewekelijkse performances. De variantie van de tweewekelijkse performance is wel zichtbaar kleiner. Dit volgt logisch uit het feit dat het aantal te voorspellen incidenten ook ongeveer zeven keer zo hoog ligt dan bij een weekdag. Met de Mann-Whitney toets kan een significant verschil in performance worden getoetst. Dit gebeurt aan de hand van de volgende hypotheses:

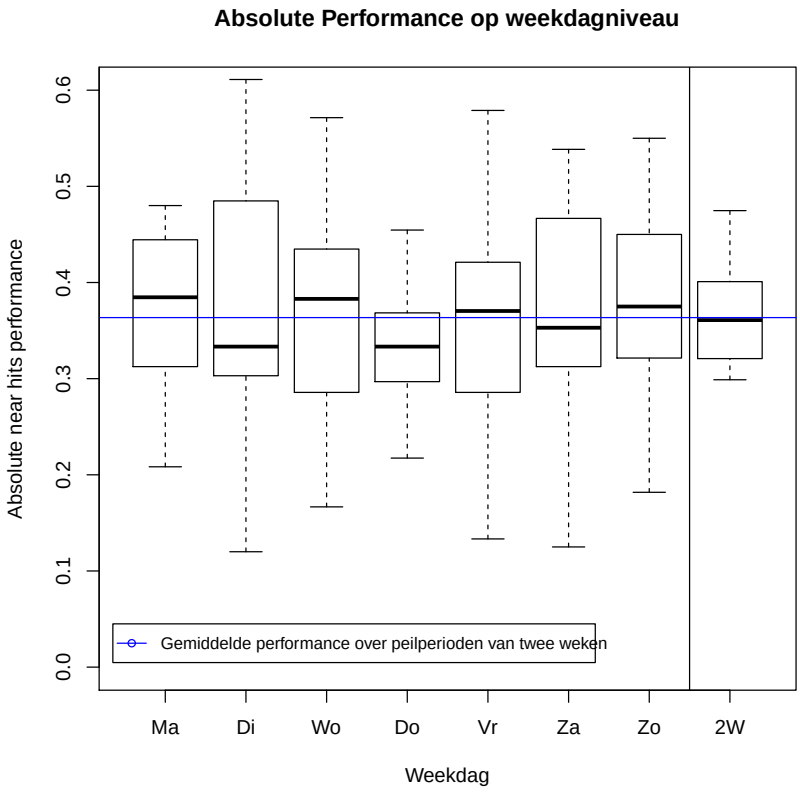
H_0: De performances van dag x en dag y zijn gelijk aan elkaar.
H_1: De performances van dag x en dag y zijn *niet* gelijk aan elkaar.

Op basis van de Mann-Whitney toets kunnen geen significante verschillen ($\alpha = 0,05$) worden gevonden tussen twee verschillende weekdagen. Ook zijn er geen weekdagen waarvan de verdeling van performances significant afwijkt van de tweewekelijkse performances.

Conclusie - De voorspellingen van het huidige CAS model voor woninginbraken zijn op alle weekdagen even goed toepasbaar.

4.2.2 Performances naar dagdeel

De tweewekelijkse voorspellingen worden binnen CAS niet specifiek gebruikt om ook dagdeelkaarten te genereren. Wel worden er diensttijdkaarten gegenereerd die de kans op een incident weergeven op



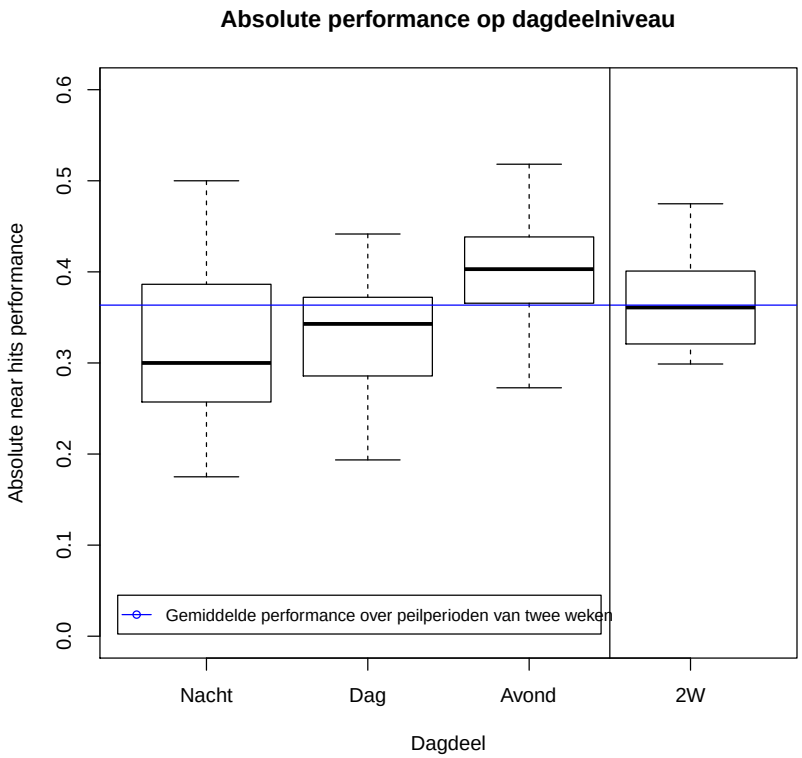
Figuur 4.2: Boxplot van absolute near hits performance op basis van de woninginbraken per weekdag voor de peilperioden 177 t/m 197

dagdeel per weekdag niveau (de performances hiervan komen in paragraaf 4.2.3 aan bod). Door deze uitsplitsing op dagdeel wordt hier (toch) de performance van de tweewekelijkse voorspellingen op de dagdelen onderzocht. Figuur 4.3 laat de performances van de tweewekelijkse voorspellingen zien wanneer alleen gekeken wordt naar de daadwerkelijk incidenten die hebben plaatsgevonden in een betreffend dagdeel.

Zichtbaar is dat de performances in de avond hoger liggen dan de performances van de incidenten ‘s nachts en overdag. Zoals ook al werd opgemerkt in paragraaf 4.2.1 is de variantie van de tweewekelijkse voorspellingen kleiner dan van de dagdelen apart, maar kennen de dagdelen weer kleinere varianties dan de weekdagen zoals te zien was in figuur 4.2. Met de Mann-Whitney toets kan een significant verschil in performance worden getoetst. Dit gebeurt aan de hand van de volgende hypotheses:

H_0: De performances van dagdeel x en dagdeel y zijn gelijk aan elkaar.
H_1: De performances van dagdeel x en dagdeel y zijn *niet* gelijk aan elkaar.

Op basis van de Mann-Whitney toets wordt H_0 verworpen en H_1 aangenomen voor de dagdelen nacht en avond ($W = 115$; p-waarde $= 0,007298$; $\alpha = 0,05$) en dag en avond ($W = 96,5$; p-waarde $0,001891$; $\alpha = 0,05$). Voor de dagdelen dag en nacht wordt H_0 niet verworpen. Op basis van deze analyse kan worden aangenomen dat het dagdeel avond beter aansluit op de huidige tweewekelijkse voorspellingen dan de dagdelen dag en nacht. Daarnaast doet dit resultaat vermoeden dat incidenten binnen de dagdelen op andere locaties plaatsvinden. Het feit dat twee verdelingen van incidenten een significant verschil



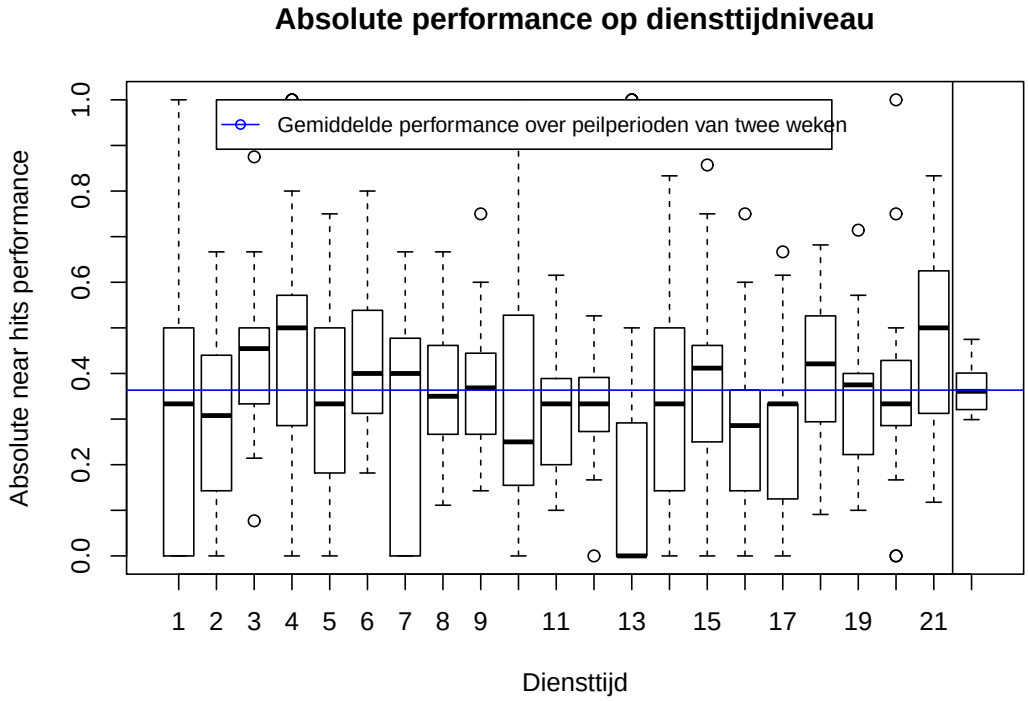
Figuur 4.3: Boxplot van absolute performance op basis van de woninginbraken per dagdeel voor de peilperioden 177 t/m 197

in performance kennen onder dezelfde voorspelling, indicateert dat de twee verdelingen geografisch verschillend verdeeld zijn. Bijvoorbeeld: wanneer incidenten 's avonds veel vaker juist voorspeld worden dan de incidenten in de nacht, lijkt het erop dat de incidenten in de nacht op een andere plek hebben plaatsgevonden. Dat is precies wat bedoeld wordt met een verschillende geografische verdeling. Deze verschillen zijn specifiek gevonden voor de dagdelen avond t.a.v. dag en avond t.a.v. nacht waar de verdelingen van incidenten een significant andere performance kennen. Het feit dat de dagdelen dag en nacht niet significant verschillen in performance zegt niet dat deze dagdelen een gelijke geografische spreiding kennen, aangezien deze analyse alleen naar de performance kijkt en niet naar het tot stand komen van deze performance. Hoofdstuk 6 gaat verder in op het onderzoeken van geografische verschillen.

Conclusie - De tweewekelijkse voorspelling door middel van het huidige CAS model sluit voor woninginbraken significant beter aan op het dagdeel avond dan op de dagdelen nacht en dag.

4.2.3 Performances naar diensttijd

De tweewekelijkse voorspellingen worden ook gebruikt voor het genereren van diensttijdkaarten waarover door de politie geen performances worden gemeten. Figuur 4.4 laat de performances van de voorspellingen zien wanneer alleen gekeken wordt naar de daadwerkelijke incidenten in de betreffende diensttijdintervallen.



Figuur 4.4: Boxplot van absolute performance op basis van de woninginbraken per diensttijd voor de peilperioden 177 t/m 197

Wat opvalt zijn de grote varianties die de verschillende verdelingen van incidenten per diensttijd kennen. Dit komt voornamelijk doordat het aantal incidenten tijdens één diensttijd soms op 1 ligt wat kan leiden tot een 100% performance wanneer dat incident juist wordt voorspeld of een performance van 0% wanneer dat incident niet juist wordt voorspeld. Desalniettemin kunnen uitspraken worden gedaan over de performances per diensttijd. De incidenten waarbij de boxen in het boxplot het laagste liggen (1, 7, 13) zijn allemaal nachten. Het viel in sectie 4.2.2 ook al op dat het dagdeel nacht onderpresteerde ten aanzien van de andere dagdelen. De diensttijdvensters welke horen bij het dagdeel avond (3, 6, 9, 12, 15, 18, 21) lijken het daarnaast (over het algemeen) ook beter te doen. Hierbij moet wel worden meegenomen dat het aantal te voorspellen incidenten ook kleiner is voor de nachtelijke dagdelen dan die van de avond en dag. Met de Mann-Whitney toets kan een significant verschil in performance worden getoetst. Dit gebeurt aan de hand van de volgende hypotheses:

H_0: De performances van diensttijd x en diensttijd y zijn gelijk aan elkaar.
H_1: De performances van diensttijd x en diensttijd y zijn *niet* gelijk aan elkaar.

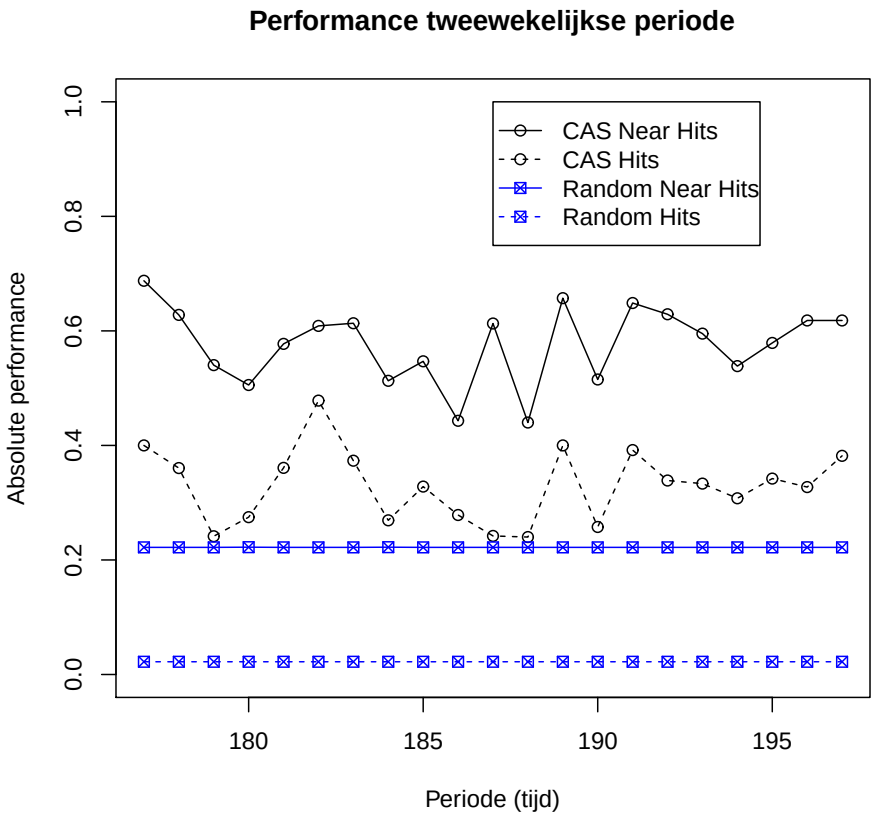
Voor 44 combinaties van dienstijden wordt H_0 verworpen en H_1 aangenomen ($\alpha = 0,05$). Deze 44 combinaties bestaan uit 7 combinaties tussen dienstijden die in hetzelfde dagdeel vallen en 37 tussen twee verschillende dagdelen. In totaal zijn 220 combinaties mogelijk wat het aantal significante diensttijdvensters tot 20,95% brengt. Diensttijdvenster 13, donderdag op vrijdagnacht, kent zelfs een performance die significant afwijkt ($\alpha = 0,05$) van alle andere diensttijdvensters (m.u.v. diensttijd 1 de zondag op maandagnacht), maar ook de minste incidenten in totaal kent (zie voor exacte aantallen tabel 5.3 in paragraaf

5.2.3). Wellicht is dat wel het grootste probleem dat moet worden meegenomen in het interpreteren van deze resultaten. De performances van sommige diensttijden lopen zo wijd uiteen, van 0% naar 100% dat het formuleren van conclusies op basis van diensttijden zal moeten gebeuren met een aanzienlijk grotere steekproef.

Conclusie - De tweewekelijkse voorspelling lijkt voor woninginbraken niet op alle diensttijden even goed aan te sluiten, maar doordat er in sommige diensttijden weinig incidenten gebeuren is het niet mogelijk daarover een sterke conclusie te formuleren.

4.3 Performance straatroven

Binnen de politie Amsterdam wordt gewerkt met de absolute (near)hits performance measure om de performance van de CAS voorspellingen te kwantificeren (paragraaf 3.4). Deze measure wordt toegepast op de tweewekelijkse voorspellingen en berekent achteraf op basis van de plaatsgevonden incidenten in de periode de performance van de voorspelling. In het algemeen wordt vaak gesproken over een performance van 60% wanneer men spreekt over straatroven. In dat geval wordt gedoeld op de absolute nearhits performance. Deze kijkt naar het percentage incidenten dat heeft plaatsgevonden in de top 3% van de locaties met de hoogste kans op een incident (hit) of in een direct naastgelegen vakje (near hit).



Figuur 4.5: Performance op basis van de straatroven die plaatsvinden op tweewekelijkse basis

Figuur 4.5 geeft de absolute hits en near hits performance weer op basis van de voorspellingen van CAS over een tweewekelijkse periode. De gemiddelde near hits performance over deze periode is $0,5769$ ($\sigma = 0,0671$) en de hits performance is gemiddeld $0,3299$ ($\sigma = 0,06370$). In figuur 4.5 is ook de performance van een random kans generator weergegeven om de voorspellingen te vergelijken met een random trekking. Deze random trekking is tot stand gekomen door iedere locatie een kans op een incident toe te kennen op basis van een trekking uit de uniforme verdeling $[0, 1]$. Hierbij worden alleen de 9.376 locaties meegenomen die CAS ook meeneemt (paragraaf 3.2). Locaties die dus door CAS zijn uitgesloten omdat ze bijvoorbeeld alleen open water of een park bevatten krijgen geen kans toebedeeld. Voor iedere peilperiode worden 200 voorspellingen gedaan waar iedere locatie bij iedere voorspelling een random kans krijgt toegekend. Iedere voorspelling kent daarbij een performance, waarvan het gemiddelde wordt gehanteerd als de random performance voor een specifieke peilperiode.

Conclusie - Het huidige CAS model heeft voor straatroven een gemiddelde near hits performance van $57,7\%$ en een gemiddelde hits performance van $33,0\%$.

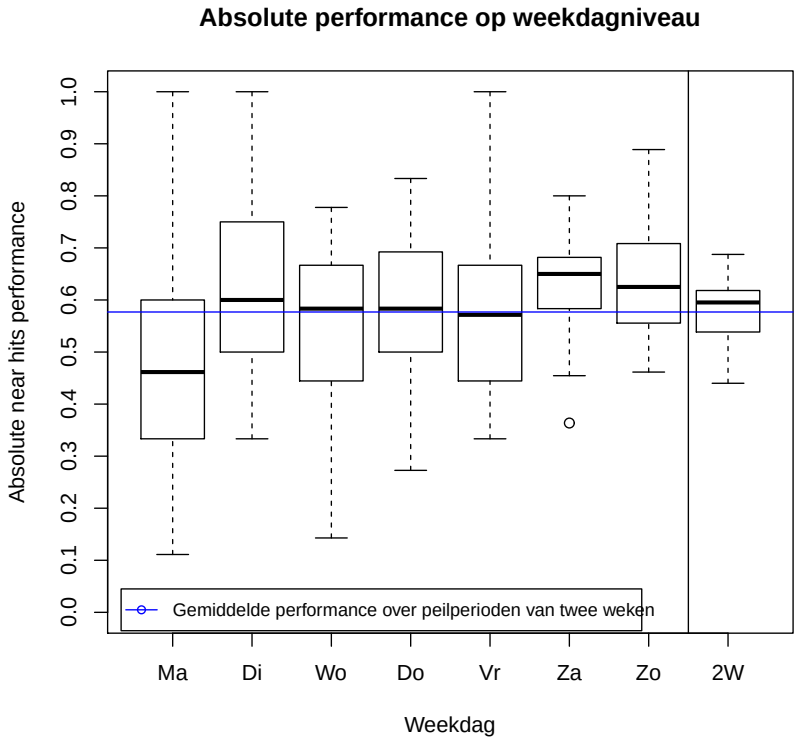
4.3.1 Performances naar weekdagen

De tweewekelijkse voorspellingen worden ook gebruikt voor het genereren van dagkaarten waarover binnen de politie geen performances worden gemeten. Figuur 4.6 laat de performances van de voorspellingen zien wanneer alleen gekeken wordt naar de daadwerkelijke incidenten op de betreffende weekdagen.

Het lijkt alsof alle performances zich redelijk verhouden tot de tweewekelijkse performances. De variantie van de tweewekelijkse performance is wel zichtbaar kleiner. Dit volgt logisch uit het feit dat het aantal te voorspellen incidenten ook ongeveer zeven keer zo hoog ligt dan bij een weekdag. Door de afname van het aantal te voorspellen incidenten neemt de variantie toe (wet van grote getallen). Op basis van de Mann-Whitney toets kan worden getoetst of het aannemelijk is dat twee performance verdelingen gelijk verdeeld zijn. De volgende hypotheses worden aangehouden:

H_0: De performances van dag x en dag y zijn gelijk aan elkaar.
H_1: De performances van dag x en dag y zijn *niet* gelijk aan elkaar.

Op basis van de Mann-Whitney toets wordt H_0 verworpen en H_1 aangenomen voor de weekdagen maandag en dinsdag ($W = 289$; p-waarde$= 0,02076$; $\alpha = 0,05$), maandag en donderdag ($W = 292$; p-waarde$= 0,03014$; $\alpha = 0,05$), maandag en zaterdag ($W = 241$; p-waarde$= 0,003568$; $\alpha = 0,05$) en maandag en zondag ($W = 227$; p-waarde$= 0,001810$; $\alpha = 0,05$). In vergelijking met de tweewekelijkse voorspelling wordt H_0 verworpen en H_1 aangenomen voor de weekdagen maandag ($W = 265$; p-waarde$= 0,01044$; $\alpha = 0,05$) en zaterdag ($W = 130$; p-waarde$= 0,02349$; $\alpha = 0,05$). Hieruit blijkt dat de maandag onderpresteert op basis van de tweewekelijkse voorspellingen en het weekend overpresteert (waarbij alleen zaterdag significant overpresteert).



Figuur 4.6: Boxplot van absolute performance op basis van de straatroven per weekdag voor de peilperioden 177 t/m 197

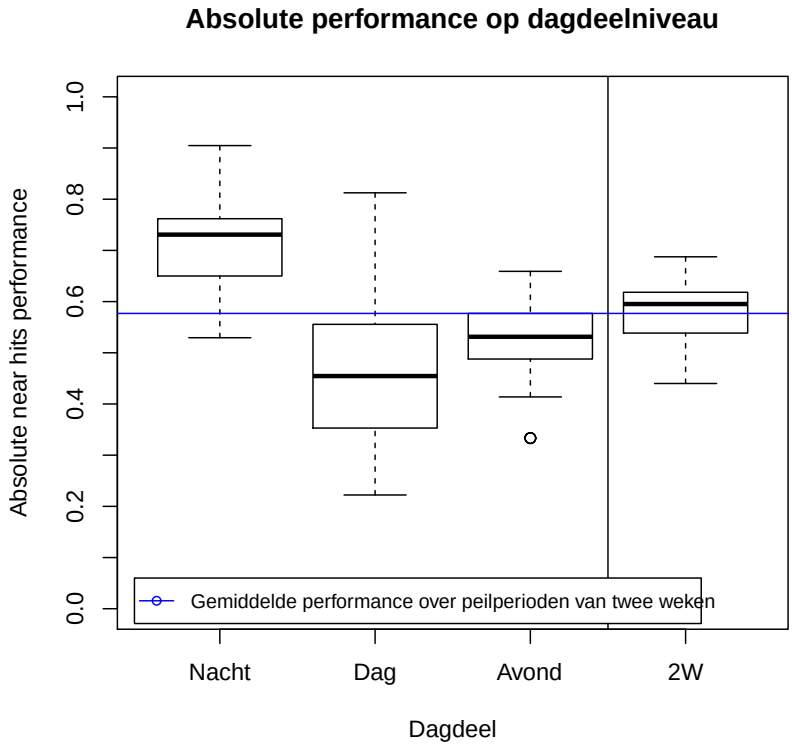
Conclusie - De tweewekelijkse voorspelling door middel van het huidige CAS model sluit voor straatroven beter aan op de weekenddagen zaterdag en zondag en minder goed op de maandag.

4.3.2 Performances naar dagdeel

De tweewekelijkse voorspellingen worden binnen CAS niet specifiek gebruikt om ook dagdeelkaarten te genereren. Wel worden er diensttijdkaarten gegenereerd die de kans op een incident weergeven op dagdeel per weekdag niveau (de performances hiervan komen in paragraaf 4.3.3 aan bod). Door deze uitsplitsing op dagdeel wordt hier (toch) de performance van de tweewekelijkse voorspellingen op de dagdelen onderzocht. Figuur 4.7 laat de performances van de tweewekelijkse voorspellingen zien wanneer alleen gekeken wordt naar de daadwerkelijk incidenten die hebben plaatsgevonden in een betreffend dagdeel.

Zichtbaar is dat de performances in de nacht hoger liggen dan de performances van de incidenten ‘s nachts en overdag. Zoals ook al werd opgemerkt in paragraaf 4.3.1 is de variantie van de tweewekelijkse voorspellingen kleiner dan van de dagdelen apart, maar kennen de dagdelen weer kleinere varianties dan de weekdagen (wet van de grote getallen) zoals te zien was in figuur 4.6. Op basis van de Mann-Whitney toets kan worden getoetst of het aannemelijk is dat twee performance verdelingen gelijk verdeeld zijn. De volgende hypotheses worden aangehouden:

H_0: De performances van dag x en dag y zijn gelijk aan elkaar.



Figuur 4.7: Boxplot van absolute performance op basis van de straatroven per dagdeel voor de peilperioden 177 t/m 197

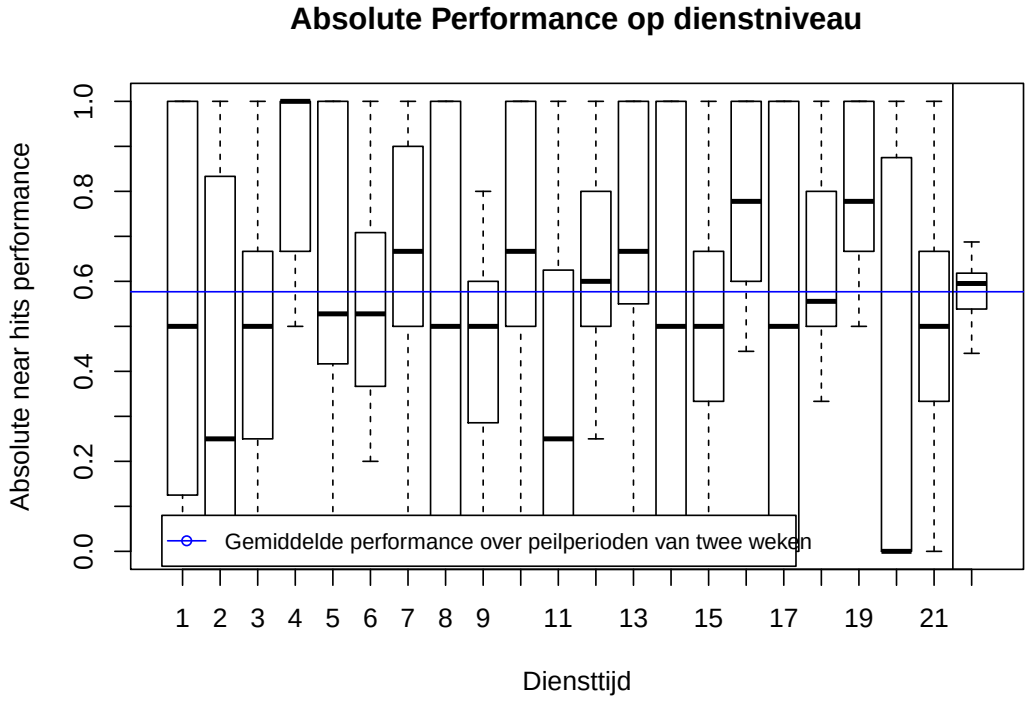
H_1: De performances van dag x en dag y zijn *niet* gelijk aan elkaar.

Op basis van de Mann-Whitneytoets wordt H_0 verworpen en H_1 aangenomen voor de dagdelen nacht en dag ($W = 383$; p-waarde $= 4,577E-5$; $\alpha = 0,05$) en nacht en avond ($W = 413$; p-waarde $= 1,361E-6$; $\alpha = 0,05$). Voor de dagdelen dag en avond wordt H_0 niet verworpen. Op basis van deze analyse kan worden aangenomen dat het dagdeel nacht beter aansluit op de huidige tweewekelijkse voorspellingen dan de dagdelen dag en avond. Daarnaast doet dit resultaat vermoeden dat incidenten binnen de dagdelen op andere locaties plaatsvinden. Het feit dat twee verdelingen van incidenten een significant verschil in performance kennen onder dezelfde voorspellingen, indicateert dat de twee verdelingen geografisch verschillend verdeeld zijn. Als voorbeeld: wanneer incidenten 's avonds veel vaker juist voorspeld worden dan de incidenten in de nacht, lijkt het erop dat de incidenten in de nacht op een andere plek hebben plaatsgevonden. Dat is precies wat bedoeld wordt met een verschillende geografische verdeling. Deze verschillen zijn specifiek gevonden voor de dagdelen avond t.a.v. dag en avond t.a.v. nacht, waar de verdelingen van incidenten een significant andere performance kennen. Het feit dat de dagdelen dag en avond niet significant verschillen in performance zegt niet dat deze dagdelen een gelijke geografische spreiding kennen aangezien deze analyse alleen naar de performance kijkt en niet naar het tot stand komen van deze performance. Hoofdstuk 6 gaat verder in op het onderzoeken van eventuele geografische verschillen.

Conclusie - De tweewekelijkse voorspelling door middel van het huidige CAS model sluit voor straatroven significant beter aan op het dagdeel nacht dan op de dagdelen dag en avond.

4.3.3 Performances naar diensttijd

De tweewekelijkse voorspellingen worden ook gebruikt voor het genereren van diensttijdkaarten waarover door de politie geen performances worden gemeten. Figuur 4.8 laat de performances van de voorspellingen zien wanneer alleen gekeken wordt naar de daadwerkelijke incidenten in de betreffende diensttijdintervallen.



Figuur 4.8: Boxplot van absolute performance op basis van de straatroven per diensttijd voor de peilperioden 177 t/m 197

Wat opvalt zijn de grote varianties die de verschillende verdelingen van incidenten per diensttijd kennen. Dit komt voornamelijk doordat het aantal incidenten tijdens één diensttijd soms op 1 ligt wat kan leiden tot een 100% performance wanneer dat incident juist wordt voorspeld of een performance van 0% wanneer dat incident niet juist wordt voorspeld. De performances van sommige diensttijden lopen zo wijd uiteen, van 0% naar 100%, dat het interpreteren zal moeten gebeuren met een aanzienlijk grotere steekproef. Voor de straatroven zullen verder geen analyses meer worden uitgevoerd op basis van diensttijden, omdat er in één diensttijd te weinig incidenten worden geregistreerd.

Conclusie - Door het gebrek aan incidenten tijdens een diensttijd is het niet mogelijk daarover een sterke conclusie te formuleren. De diensttijden als tijdsinterval worden voor straatroven niet langer geanalyseerd.

4.4 Conclusie

Het huidige CAS model kan $36,3\%$ van de woninginbraken en $57,7\%$ van de straatroven voorspellen. Deze voorspellingen worden in de huidige CAS omgeving voorspelt voor perioden van twee weken, maar ook gebruikt voor het genereren van voorspellingen op basis van *weekdag, dagdeel* en *diensttijd*. Over deze onderliggende tijdsintervallen worden echter geen performances gemeten waardoor niet bekend is of er specifieke tijdsintervallen zijn die extreem afwijken van de gemiddelde performance. Een afwijking in performance kan mogelijk veroorzaakt worden door een afwijkende ruimtelijke verdeling en is daarom in het kader van dit onderzoek interessant. De volgende resultaten zijn gevonden:

Woninginbraken

1. De voorspellingen van het huidige CAS model voor woninginbraken zijn op alle weekdagen even goed toepasbaar.
2. De tweewekelijkse voorspelling door middel van het huidige CAS model sluit voor woninginbraken significant beter aan op het dagdeel avond dan op de dagdelen nacht en dag.
3. De tweewekelijkse voorspelling lijkt voor woninginbraken niet op alle diensttijden even goed aan te sluiten, maar doordat er in sommige diensttijden weinig incidenten gebeuren is het niet mogelijk daarover een sterke conclusie te formuleren.

Straatroven

1. De tweewekelijkse voorspelling door middel van het huidige CAS model sluit voor straatroven beter aan op de weekenddagen zaterdag en zondag en minder goed op de maandag.
2. De tweewekelijkse voorspelling door middel van het huidige CAS model sluit voor straatroven significant beter aan op het dagdeel nacht dan op de dagdelen dag en avond.
3. Door het gebrek aan incidenten tijdens een diensttijd is het niet mogelijk daarover een sterke conclusie te formuleren. De diensttijden als tijdsinterval worden voor straatroven niet langer geanalyseerd.

Hoofdstuk 5

Ruimtelijke verschillen in onderliggende tijdsintervallen

CAS leunt op de aanname dat voor verschillende tijdsintervallen binnen een tweewekelijkse peilperiode incidenten een verschillende intensiteit kennen, maar de geografische spreiding identiek is (zie paragraaf 1.2.2). Om deze aanname te onderzoeken is in hoofdstuk 4 de toepasbaarheid van de huidige tweewekelijkse voorspellingen op verschillende onderliggende tijdsintervallen onderzocht. Daaruit worden significante verschillen in performance gevonden ten aanzien van verschillende onderliggende tijdsintervallen. De meest voor de hand liggende oorzaak hiervan is dat incidenten in de verschillende tijdsintervallen op andere plekken gebeuren, wat duidt op een verschillende geografische verdeling van incidenten. Dit hoofdstuk gaat verder met het onderzoeken of er inderdaad geografische verschillen waarneembaar zijn tussen verschillende tijdsintervallen en probeert antwoord te geven op de onderzoeksvraag:

In hoeverre zijn incidenten ruimtelijk gezien gelijk verdeeld t.a.v. verschillende onderliggende tijdsintervallen?

Dit hoofdstuk vervolgt met een beschrijving van de onderzoeksmethode in paragraaf 5.1 gevolgd door paragraaf 5.2 en 5.3 met de resultaten uitgesplitst naar respectievelijk woninginbraken en straatroven. Paragraaf 5.4 bevat een discussie en tot slot komt in paragraaf 5.5 de conclusie.

5.1 Methode

Om te onderzoeken hoe incidenten ruimtelijk gezien verdeeld zijn, worden incidenten (punten in het tijdruimtelijk vlak) toegekend aan een tijdsinterval en ruimtelijke polynoom zoals een gridlocatie, wijk of district (zie paragraaf 3.1.3 en 3.2). Dit zorgt ervoor dat de geografisch verdeling van incidenten is getransformeerd tot een areal ruimtelijke datastructuur (zie paragraaf 2.4.1). Het voordeel hiervan is dat deze datastructuur eenvoudiger te interpreteren is en er in het gebruik van analysetechnieken aan minder beklemmende aannamen hoeft te worden voldaan. Op basis van de areal ruimtelijke datastructuur kunnen analysetechnieken die van toepassing zijn op twee categorische variabelen worden toegepast.

Pearson χ^2 test De Pearson χ^2 test wordt gebruikt om afhankelijkheid tussen tijd en ruimte te toetsen [17] [22]. Hierbij worden de incidenten binnen een tijdsinterval als één verdeling van incidenten beschouwd en wordt getoetst of de verdelingen van verschillende tijdsintervallen verschillen.

Correspondentieanalyse Als uitbreiding op de Pearson χ^2 test wordt ook de correspondentieanalyse toegepast om de χ^2 toetsingsgrootheid te ontleden in dimensies. Deze dimensies beschrijven in termen de χ^2 afstanden waardoor de afstanden tussen de punten een betekenis hebben maar de assen en bijhorende waarden op zichzelf niet. De twee dimensies die het grootste deel van de χ^2 toetsingsgrootheid omvatten worden weergegeven in een tweedimensionaal plot. Die plot verschaft inzicht in het gedrag van alle categorieën ten aanzien van de χ^2 toetsingsgrootheid. Deze techniek is dus ook vooral beschrijvend en niet toetsend.

Monte Carlo benadering Naast de technieken die zich baseren op de χ^2 statistic wordt er ook gebruik gemaakt van een Monte Carlo benadering [1] waarbij incidenten worden gesampled. Uit de verzameling van incidenten die hebben plaatsgevonden in een specifiek tijdsinterval wordt 85% van de incidenten random getrokken. Voor deze 85% van de incidenten worden de percentages incidenten per gebied bepaald. Dit proces herhaalt zich 200 keer waarna een 95% betrouwbaarheidsinterval voor de percentages per gebied kan worden opgesteld. De percentages incidenten per gebied van een ander tijdsinterval kunnen vervolgens worden getoetst aan de 95% betrouwbaarheidsintervallen. De Monte Carlo benadering wordt in dit onderzoek alleen gebruikt wanneer het aantal te vergelijken tijdsintervallen behapbaar blijft en wordt daarom niet gebruikt wanneer alle diensttijden afzonderlijk worden vergeleken. Als bijvoorbeeld alle 21 diensttijden met elkaar vergeleken worden, zijn er 210 (21de partiele som van $\frac{n(n+1)}{2}$) analyses nodig en zal 210 keer een output moeten worden geevalueerd wat in tijd niet opweegt tegen de informatie die daaruit te verkrijgen is. Daarnaast kennen diensttijden vaak een lage frequentie van incidenten waardoor het sampelen zorgt voor grote betrouwbaarheidsintervallen en deze methodiek weinig kennis kan toevoegen.

5.2 Woninginbraken toegekend aan districten

De regio Amsterdam is binnen de politie onderverdeeld in 5 districten: Centrum, Noord, Oost, West en Zuid[1]. In de peilperioden 177 t/m 197 (zie paragraaf 3.3) zijn er 4.400 woninginbraken geregistreerd die hebben plaatsgevonden in een van deze vijf districten. In deze paragraaf wordt onderzocht of voor verschillende tijdsintervallen, incidenten in dezelfde districten plaatsvinden.

5.2.1 Woninginbraken toegekend aan districten en dagdelen

Categorische benadering

Tabel 5.1 geeft de verdeling van incidenten weer over de dagdelen nacht, dag en avond ten aanzien van de vijf districten. Wanneer het dagdeel geen invloed heeft op de locatie waar incidenten plaatsvinden,

[1] Dit is de onderverdeling ten tijde van dit onderzoek. Deze structuur is mogelijk onderhevig aan reorganisatie van de district- wijkteams in de toekomst

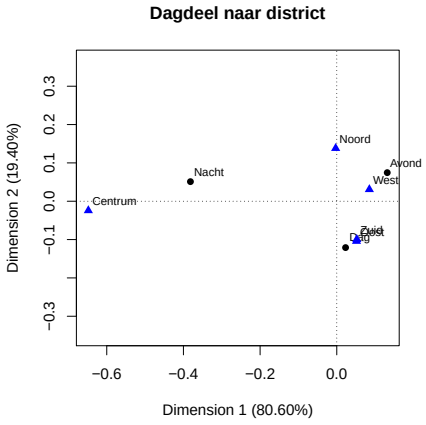
Dagdeel	Centrum	Noord	Oost	Zuid	West	Totaal
Nacht	137	182	146	99	230	794
Dag	115	275	388	261	533	1.572
Avond	80	478	415	301	760	2.034
Totaal	332	935	949	661	1.523	4.400

Tabel 5.1: Aantal woninginbraken naar dagdeel per district

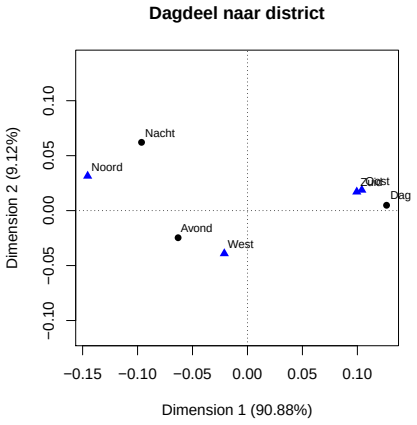
kan inderdaad gezegd worden dat op districtsniveau de ruimtelijke verdeling mogelijk gelijk is voor de verschillende dagdelen. De data in tabel 5.1 is weergegeven als twee categorische variabelen (district en dagdeel) waardoor de Pearson χ^2 test kan nagaan of dagdeel afhankelijk is van het district. De volgende hypotheses worden opgesteld:

H_0: De proportie van incidenten die plaatsvinden in de districten is *niet* afhankelijk van het dagdeel.
H_1: De proportie van incidenten die plaatsvinden in de districten is afhankelijk van het dagdeel.

Op basis van de Pearson χ^2 test wordt H_0 verworpen en H_1 aangenomen ($\chi^2 = 191,23$; $df = 8$; p-waarde$< 2,2e$-16; $\alpha = 0,05$). Dit betekent dat het aannemelijk is dat woninginbraken over de districten heen op verschillende tijdstippen binnen de dag plaatsvinden. Er kan een correspondentieanalyse worden uitgevoerd om de variantie te decompenseren in verschillende dimensies waar figuur 5.1 op grafische wijze de output van weergeeft. Vooral het district Centrum kent een extreme afwijking t.a.v. de overige vier districten. In dezelfde hoek ligt ook het nachtelijke tijdsinterval. In cijfers is dit te onderbouwen: 41,81% van de woninginbraken in het Centrum vindt 's nachts plaats (t.a.v. 34,24% overdag en 23,94% 's avonds), terwijl in alle andere districten het aantal incidenten 's nachts lager is dan overdag en 's avonds.



Figuur 5.1: CA met vijf districten



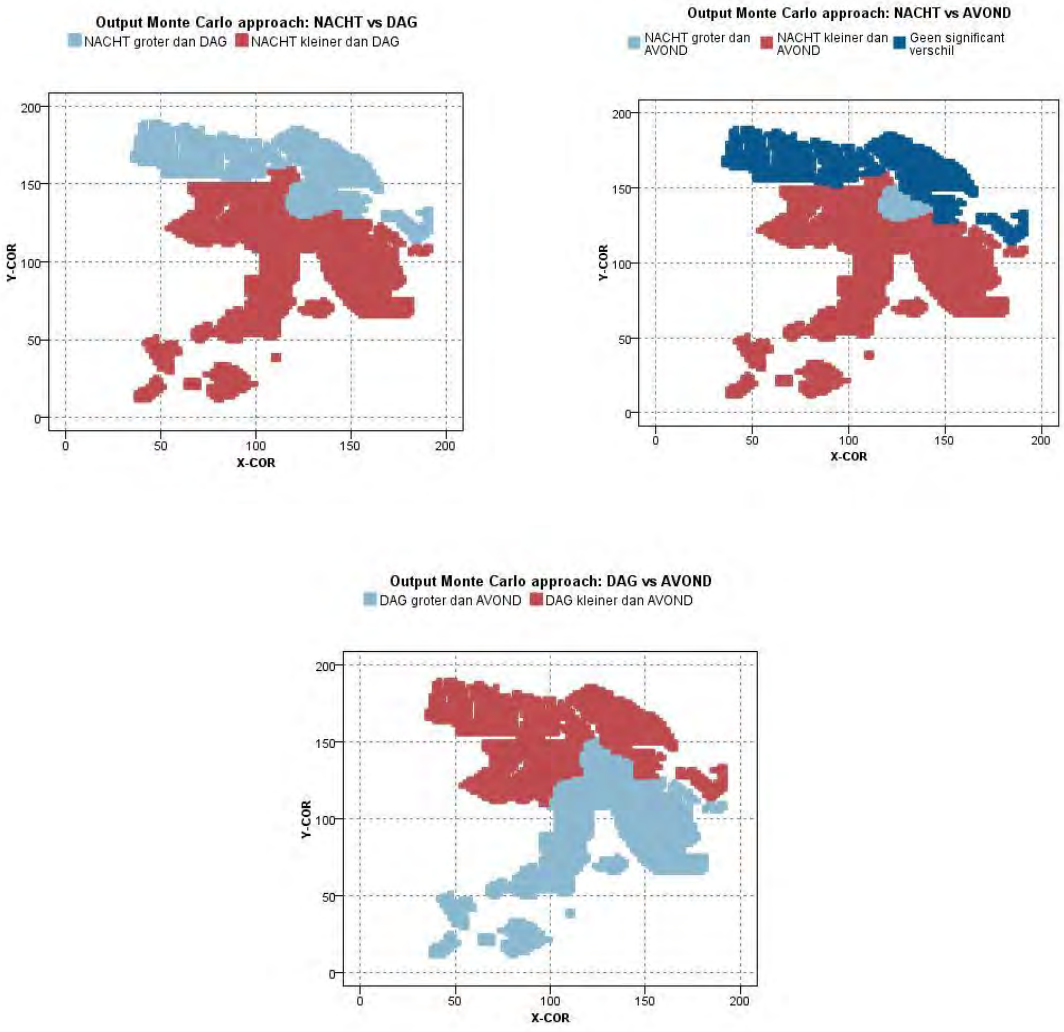
Figuur 5.2: CA met vier districten

Wanneer alleen gekeken wordt naar de districten Noord, West, Zuid en Oost wordt op basis van de Pearson χ^2 test H_0 verworpen en H_1 aangenomen ($\chi^2 = 41,90$; $df = 6$; p-waarde $< 1,92e$-7; $\alpha = 0,05$). Figuur 5.2 geeft de output van de correspondentieanalyse aan. Deze uitkomst geeft aan dat ondanks dat district Centrum het meest lijkt af te wijken, de andere districten onderling ook een significant afwijkende verdeling van incidenten kennen onderliggend aan de dagdelen. Beide correspondentieanalyses kennen

wel zeer kleine afwijkingen tussen de districten Zuid en Oost. Wanneer alleen de districten Zuid en Oost worden onderworpen aan de Pearson χ^2 test kan H_0 niet verworpen worden ($\chi^2 = 0,0093$; $df = 2$; p-waarde $= 0,995$; $\alpha = 0,05$). Voor alle andere combinaties van districten wordt H_0 verworpen met een p-waarde $< 0,05$.

Monte Carlo benadering

De geografisch data uit tabel 5.1 is een cijfermatige weergave van een areal ruimtelijke datastructuur. Een techniek om zulke datapatronen te vergelijken op gelijkenis is de non-parametrische Monte Carlo benadering.



Figuur 5.3: Monte Carlo output van links naar rechts, van boven naar onder: nacht vs dag, nacht vs avond, dag vs avond.

Voor de dagdelen en districten is de Monte Carlo benadering gebruikt, ondanks dat vijf gebieden voor het gebruik van de Monte Carlo benadering aan de lage kant is. De uitkomsten van deze benadering zijn te vinden in de figuur 5.3. Ook uit deze analysetechniek wordt duidelijk dat het district Centrum 's nachts een hogere intensiteit aan incidenten kent dan overdag en 's avonds. Overdag wordt in Oost en Zuid relatief vaak ingebroken en 's avonds kent West relatief veel woninginbraken.

Conclusie - Woninginbraken vinden plaats in verschillende districten wanneer onderscheid wordt gemaakt in de dagdelen nacht, dag, avond. De meest afwijkende verdeling van incidenten wordt waargenomen in district Centrum, waar het hoogtepunt 's nachts is, terwijl alle andere districten 's nachts de minste inbraken plaatsvinden.

5.2.2 Woninginbraken toegekend aan weekdagen

Categorische benadering

Tabel 5.2 geeft de verdeling van incidenten weer over de weekdagen ten aanzien van de vijf districten.

Weekdag	Centrum	Noord	Oost	Zuid	West	Totaal
Maandag	51	135	156	95	216	653
Dinsdag	44	130	147	90	219	630
Woensdag	42	129	158	114	238	681
Donderdag	37	134	178	106	251	706
Vrijdag	38	73	79	56	126	372
Zaterdag	62	201	136	123	292	814
Zondag	58	133	95	77	181	544
Totaal	332	935	949	661	1523	4400

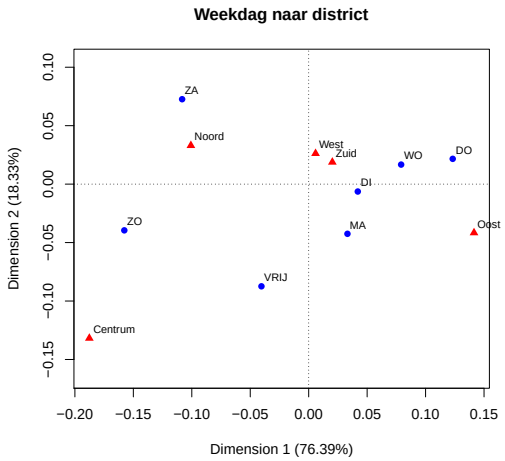
Tabel 5.2: Aantal woninginbraken naar weekdag per district

Wanneer de weekdag geen invloed heeft op de locatie waar incidenten plaatsvinden, kan inderdaad gezegd worden dat er op districtsniveau de ruimtelijke verdeling mogelijk gelijk is voor verschillende weekdagen. De data in tabel 5.2 is evenals in de vorige paragraaf weergegeven als twee categorisch variabelen (district en weekdag) waardoor de Pearson χ^2 test kan nagaan of weekdag afhankelijk is van het district. De volgende hypotheses worden opgesteld:

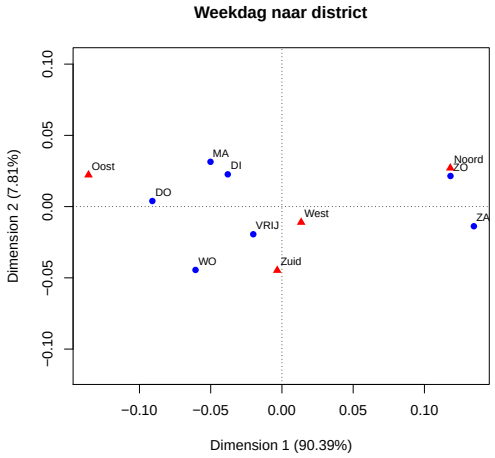
H_0: De proportie van incidenten die plaatsvinden in de districten is *niet* afhankelijk van de weekdag.
H_1: De proportie van incidenten die plaatsvinden in de districten is afhankelijk van de weekdag.

Op basis van de Pearson χ^2 test wordt H_0 verworpen en H_1 aangenomen ($\chi^2 = 53,00$; $df = 24$; p-waarde $< 0,0006$; $\alpha = 0,05$). Dit betekent dat het aannamelijk is dat woninginbraken over de districten heen op verschillende weekdagen plaatsvinden. Paragraaf 4.2.1 probeerde ook de mogelijkheden hiervoor te onderzoeken, maar kwam tot geen overtuigend verschil. Er kan een correspondentieanalyse worden uitgevoerd om de χ^2 toetsingsgrootheid te decompenseren in verschillende dimensies. Figuur 5.4 geeft hier op grafische wijze de output van. Er zijn twee aspecten die hier lijken op te vallen: (1) het district Centrum kent een extreme afwijking ten aanzien van de overige districten en (2) de weekeindedagen zaterdag, zondag en in mindere mate vrijdag kennen een afwijking ten aanzien van de overige dagen. Het feit dat district Centrum zich afwijkend gedraagd werd ook al opgemerkt in paragraaf 5.2.1. De afwijking van de dagen vrijdag, zaterdag en zondag is nog niet eerder opgemerkt. Wanneer alleen gekeken wordt

naar de districten Noord, West, Zuid en Oost wordt op basis van de Pearson χ^2 test H_0 verworpen en H_1 aangenomen ($\chi^2 = 34,10$; $df = 18$; p-waarde $< 0,012$; $\alpha = 0,05$). Figuur 5.5 geeft de output van de bijhorende correspondentieanalyse. Deze correspondentieanalyse laat zien dat de dagen maandag t/m vrijdag een relatief zelfde verdeling kennen in tegenstelling tot de dagen zaterdag en zondag. Doordat het district Centrum is verwijderd, lijkt de afwijking van de vrijdag zich voornamelijk te verhouden tot het district Centrum en met betrekking tot de overige vier districten geen parten te spelen.



Figuur 5.4: CA met vijf districten

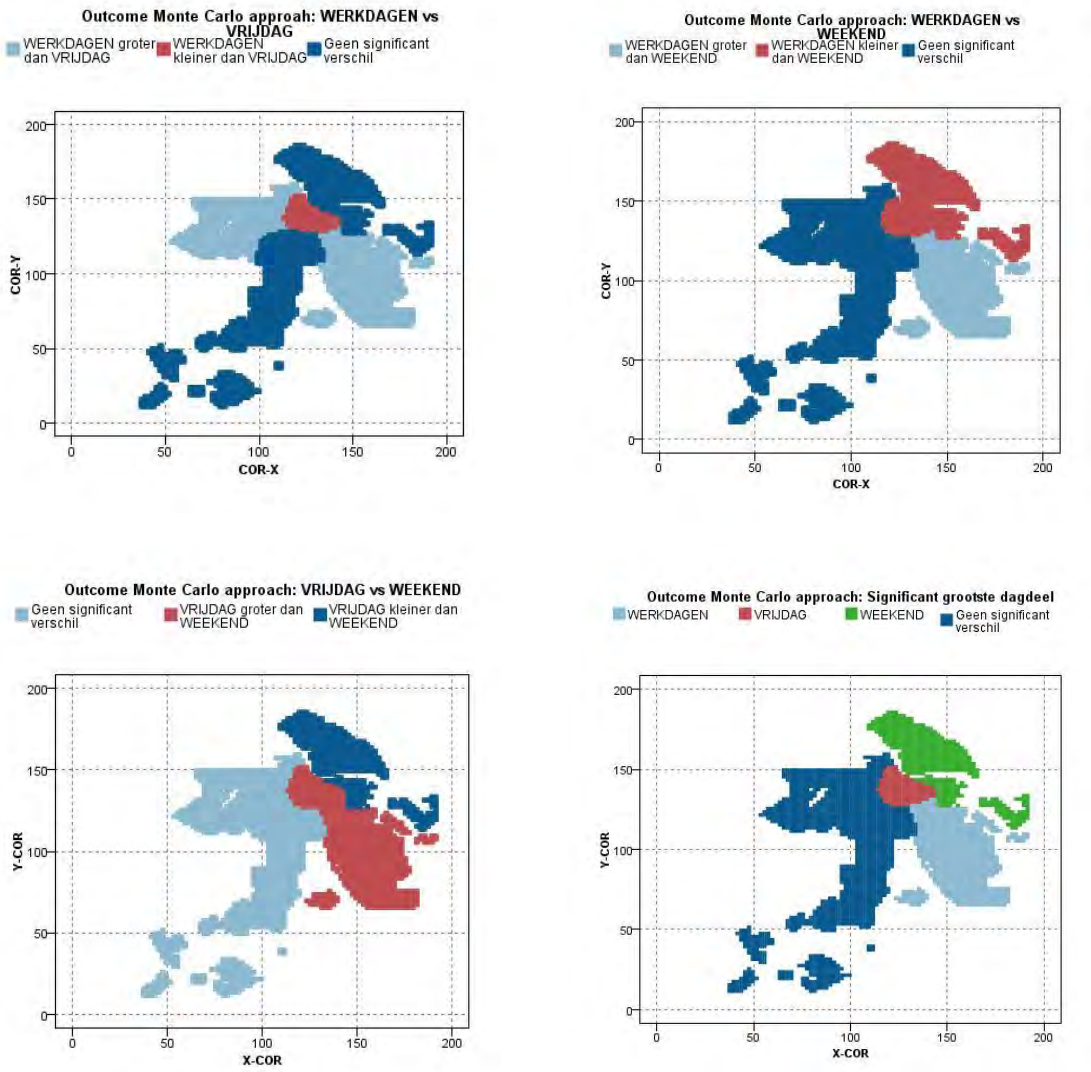


Figuur 5.5: CA met vier districten

Wanneer er een selectie wordt gemaakt op de dagen maandag, dinsdag, woensdag, donderdag en vrijdag onderhevig aan alle vijf districten kan H_0 niet worden verworpen ($\chi^2 = 14,98$; $df = 16$; p-waarde $< 0,53$; $\alpha = 0,05$). Dit betekent dat het lijkt erop dat inbraken op de dagen maandag t/m vrijdag ruimtelijk gelijk zijn verdeeld over de districten. Zaterdag en zondag lijken zich sterk te verhouden tot district Noord: in het weekend vindt 35,72% van het totale aantal inbraken in Noord plaats in tegenstelling tot de 29,55% gemiddeld in andere districten. Bij een homogene intensiteit over alle weekdagen zou een percentage van 28,57% verwacht worden. Dat het gemiddelde in de vier districten (Centrum, Oost, Zuid en West) op 29,55% ligt, komt voornamelijk door de piek op zaterdag en de terugval op zondag in intensiteit: 59,49% van de weekendinbraken vindt plaats op zaterdag.

Monte Carlo benadering

In paragraaf 5.2.1 wordt de Monte Carlo benadering gebruikt om twee geaggregeerde datapatronen te vergelijken. Voor de analyse naar incidenten ten aanzien van de verschillende districten kan deze methode wederom gebruikt worden. Als alle weekdagen met elkaar vergeleken worden, zijn er 21 (6de partiele som van $\frac{n(n+1)}{2}$, met $n = 6$) analyses nodig en komen daar 21 plots uit. De categorische analyse wijst vooral op een verschil tussen de werkdagen en weekenddagen. De rol van de vrijdag lijkt daarbij wat discutabel. Voor deze analyse wordt onderscheid gemaakt tussen drie type dagen: werkdagen (ma t/m do), vrijdag en weekenddagen. Deze drie geaggregeerde datapatronen worden ter vergelijking onderworpen aan de Monte Carlo benadering. De uitkomsten zijn te vinden in figuur 5.6, waar naast de gebruikelijke output wordt ook het significant grootste dagdeel is geplot.



Figuur 5.6: Monte Carlo benadering toegespitst op werkdagen (ma t/m do), vrijdag en weekenddagen (zat en zo)

Uit deze analyse blijkt dat op basis van relatieve percentages Noord zich verhouwd tot het weekend, Oost tot de werkdagen en Centrum tot de vrijdag. Wel moet er bij het interpeteren wel rekening worden gehouden met het lage aantal geografische gebieden, wat de uitkomsten onderhevig maakt aan uitschieters.

Conclusie - Woninginbraken vinden in verschillende districten plaats op basis van de betreffende weekdag. Het grootste verschil kan gevonden worden tussen de werkdagen en weekenddagen. Tussen de werkdagen maandag t/m vrijdag onderling en tussen de weekenddagen zaterdag en zondag onderling lijken geen grote verschillen te zijn, al lijkt de vrijdag zich meer afwijkend te gedragen ten aanzien van de overige werkdagen.

5.2.3 Woninginbraken toegekend aan diensttijden

Categorische benadering

Tabel 5.3 geeft de verdeling van incidenten weer over alle 21 diensttijden ten aanzien van de vijf districten. Wanneer de weekdag geen invloed heeft op de locatie waar incidenten plaatsvinden, kan inderdaad gezegd worden dat er op districtsniveau de ruimtelijke verdeling mogelijk gelijk is voor de verschillende diensttijden. De data in tabel 5.3 is evenals in de vorige paragrafen weergegeven als twee categorisch variabelen (district en diensttijd) waardoor de Pearson χ^2 test kan nagaan of weekdag afhankelijk is van het district. De volgende hypotheses worden opgesteld:

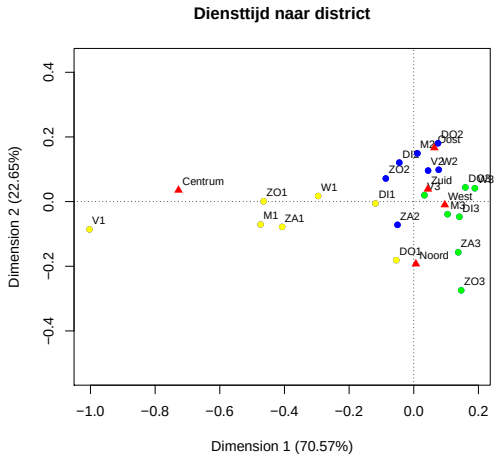
Tijdsvenster	Centrum	Noord	Oost	Zuid	West	Totaal
MA1	16	20	14	12	20	82
MA2	22	51	81	46	96	296
MA3	13	64	61	37	100	275
DI1	10	20	20	11	32	93
DI2	24	45	65	45	89	268
DI3	10	65	62	34	98	269
WO1	18	26	26	16	33	119
WO2	16	52	70	55	95	288
WO3	8	51	62	43	110	274
DO1	7	27	18	9	25	86
DO2	20	48	89	51	120	328
DO3	10	59	71	46	106	292
VR1	16	11	5	8	8	48
VR2	8	20	28	18	44	118
VR3	14	42	46	30	74	206
ZA1	30	38	27	19	51	165
ZA2	14	37	28	29	54	162
ZA3	18	126	81	75	187	487
ZO1	40	40	36	24	61	201
ZO2	11	22	27	17	35	112
ZO3	7	71	32	36	85	231
Totaal	332	935	949	661	1523	4400

Tabel 5.3: Aantal woninginbraken naar diensttijd per district

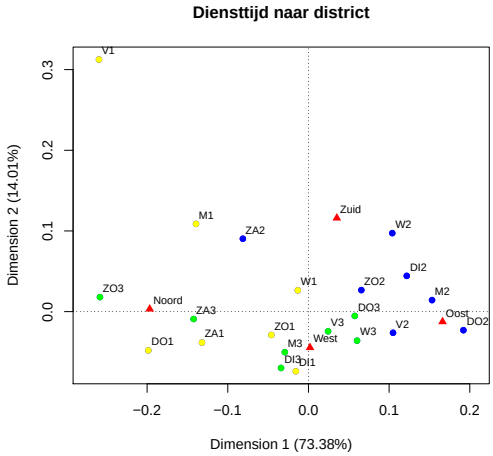
H_0: De proportie van incidenten die plaatsvinden in de districten is *niet* afhankelijk van de diensttijd.
H_1: De proportie van incidenten die plaatsvinden in de districten is afhankelijk van diensttijd.

Op basis van de Pearson χ^2 test wordt H_0 verworpen en H_1 aangenomen ($\chi^2 = 276,24$; $df = 80$; p-waarde $< 2,2e$-16; $\alpha = 0,05$). Dit betekent dat het aannamelijk is dat woninginbraken over de districten

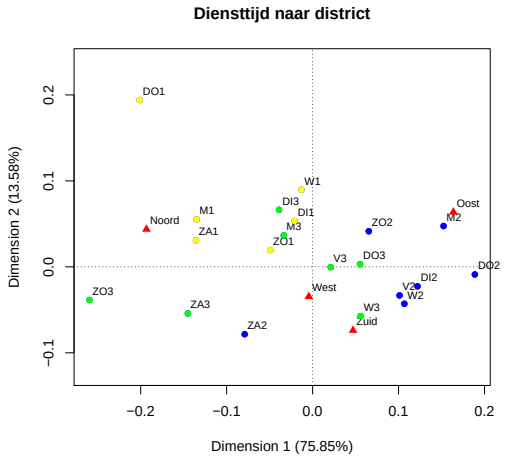
heen op verschillende weekdagen plaatsvinden. Daarbij speelt wel mee dat de frequentie incidenten in sommige tijdsvakken erg laag is en de Pearson χ^2 test voorzichtig gebruikt moet worden. Er kan een correspondentieanalyse worden uitgevoerd om de variantie te decompenseren in verschillende dimensies. Figuur 5.7 geeft hier op grafische wijze de output van. De diensttijden zijn gekleurd naar het dagdeel waarin zij vallen: nacht is geel, dag is blauw en avond is groen. Er zijn twee aspecten die hier lijken op te vallen: (1) het district Centrum kent een extreme afwijking ten aanzien van de overige districten en (2) de nachtelijke diensttijden kennen een extremere afwijking dan de avond en dag diensttijden. Deze twee aspecten werden ook al opgemerkt in paragraaf 5.2.1 waar de dagdelen ten aanzien van districten werd bekeken.



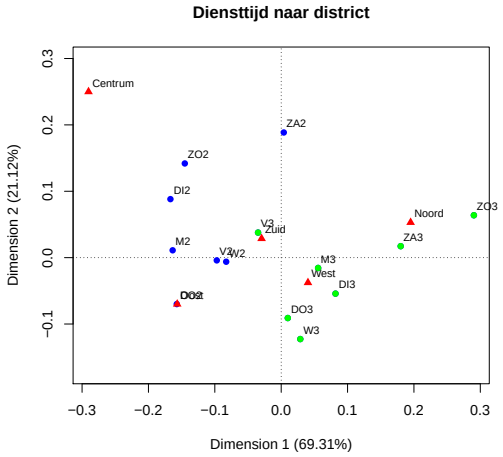
Figuur 5.7: CA met vijf districten, alle diensttijden



Figuur 5.8: CA met vier districten (geen Centrum), alle diensttijden



Figuur 5.9: CA met vier districten (geen Centrum), alle diensttijden zonder V1



Figuur 5.10: CA met vijf districten, alle dag en avond diensttijden

Wanneer alleen gekeken wordt naar de districten Noord, West, Zuid en Oost over alle tijdsvensters wordt op basis van de Pearson χ^2 test H_0 verworpen en H_1 aangenomen ($\chi^2 = 86,21$; $df = 60$; p-waarde $< 0,015$). Figuur 5.8 geeft de output van de bijhorende correspondentieanalyse. Het onderscheid in

de dagdelen is zonder het district Centrum veel minder duidelijk. Het dagdeel dag lijkt nog het meest geclusterd. Vanuit alle tijdsvensters gedraagt het tijdsvenster donderdag op vrijdagnacht zich het meest afwijkend. Wanneer we deze ook verwijderen uit de analyse ontstaat de output die is weergegeven in figuur 5.9. Hier is de onderverdeling op dagdeel al weer meer geclusterd. Een onderscheid op basis van dag wordt niet sterk gevonden. Wel zijn de weekenddagen zaterdag en zondag links onderin geclusterd. Het onderscheid van de afwijkende weekenddagen werd al gevonden in paragraaf 5.2.2. Wanneer alleen gekeken wordt naar de tijdsvensters in het dagdeel dag en avond over alle districten wordt op basis van de Pearsons χ^2 test H_0 verworpen en H_1 aangenomen ($\chi^2 = 97,20$; $df = 52$; p-waarde $< 0,00015$; $\alpha = 0,05$). Figuur 5.10 geeft de output van de bijhorende correspondentieanalyse.

Wanneer er een selectie wordt gemaakt op de diensttijden die vallen in het dagdeel dag, wordt op basis van de Pearson χ^2 test H_0 niet verworpen ($\chi^2 = 17,83$; $df = 24$; p-waarde $= 0,81$; $\alpha = 0,05$). Wanneer er een selectie wordt gemaakt op de diensttijden die vallen in het dagdeel avond, wordt op basis van de Pearson χ^2 test H_0 niet verworpen ($\chi^2 = 33,20$; $df = 24$; p-waarde $= 0,0999$; $\alpha = 0,05$). Wanneer er een selectie wordt gemaakt op de diensttijden die vallen in het dagdeel nacht, wordt op basis van de Pearson χ^2 test H_0 niet verworpen ($\chi^2 = 28,49$; $df = 24$; p-waarde $= 0,24$; $\alpha = 0,05$).

Er is geen gebruik gemaakt van een Monte Carlo benadering door het grote aantal verschillende tijdsvensters en het kleine aantal geografische clusters.

Conclusie - Woninginbraken vinden in verschillende districten plaats op basis van de betreffende diensttijd. Vooral het district Centrum in combinatie met de diensttijden die in de nacht vallen kennen een extreem afwijkende verdeling. De verdeling van incidenten over de diensttijden lijken zich daarnaast te gedragen in clusters van dagdelen en de week- en weekenddagen.

5.3 Woninginbraken toegekend aan wijkteams

De regio Amsterdam is binnen de politie onderverdeeld in 5 districten die weer zijn opgedeeld in 31 wijken[2]. In de peilperioden 177 t/m 197 (zie paragraaf 3.3) zijn er 4.400 woninginbraken geregistreerd die hebben plaatsgevonden in een van deze 31 wijken.

5.3.1 Woninginbraken toegekend aan dagdelen

Categorische benadering

Tabel 5.4 geeft de verdeling van incidenten weer over de dagdelen nacht, dag en avond ten aanzien van de 31 wijken. Wanneer het dagdeel geen invloed heeft op de locatie waar incidenten plaatsvinden, kan inderdaad gezegd worden dat op wijkniveau de ruimtelijke verdeling mogelijk gelijk is voor de verschillende dagdelen. In paragraaf 5.2 werd echter al onderzoek gedaan naar de afhankelijkheid van district ten aanzien van dagdelen, waaruit bleek dat het aannemelijk is dat de locatie waar incidenten plaatsvinden

[2]Dit is de onderverdeling ten tijde van dit onderzoek. Deze structuur is mogelijk onderhevig aan reorganisatie van de district- wijkteams in de toekomst

Wijk	Nacht	Dag	Avond	Totaal
Aalsmeer	6	15	32	53
Amstelveen Noord	16	41	52	109
Amstelveen Zuid	16	55	69	140
August Allebplein	40	110	153	303
Balistraat	27	60	51	138
Beursstraat	8	2	6	16
Bos en Lommer	29	64	102	195
De Pijp	17	30	28	75
Diemen/Ouder-Amstel	28	46	65	139
Flierbosdreef	12	66	65	143
Ganzenhoef	26	112	88	226
Houtmankade	24	52	61	137
IJ-tunnel	25	49	29	103
IJburg	12	19	46	77
Klimopweg	78	104	180	362
Koninginneweg	13	25	31	69
Lijnbaansgracht	24	21	13	58
Linnaeusstraat	37	81	66	184
Lodewijk van Deysselstraat	41	110	167	318
Meer en Vaart	51	86	158	295
Nieuwezijds Voorburgwal	11	4	4	19
Oud West	25	41	37	103
Prinsengracht	24	7	6	37
Raampoort	46	30	21	97
Remmerdenplein	21	64	107	192
Rivierenbuurt	12	30	16	58
Surinameplein	23	72	88	183
Uithoorn	11	25	21	57
Van Leijenberghlaan	14	64	57	135
Waddenweg	68	94	206	368
s-Gravesandeplein	22	30	33	85
Totaal	807	1609	2058	4474

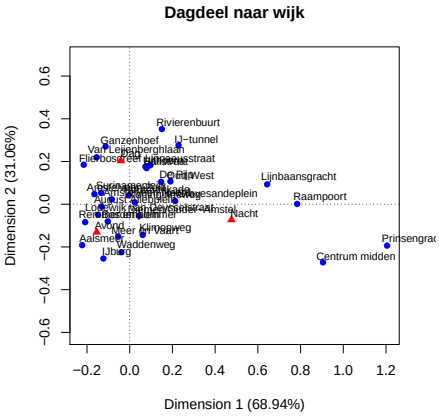
Tabel 5.4: Aantal woninginbraken naar dagdeel per wijk

afhankelijk is van het dagdeel. Hier wordt onderzocht of deze afhankelijkheid ook kan worden gevonden wanneer er gekeken wordt naar wijken in plaats van districten. De data in tabel 5.4 is weergegeven als twee categorische variabelen (wijk en dagdeel) waardoor de Pearson's χ^2 test kan nagaan of dagdeel

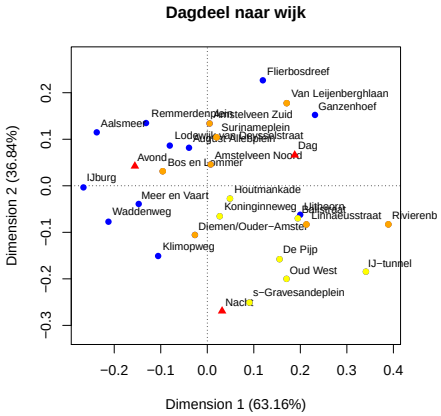
afhankelijk is van de wijk. De volgende hypotheses worden opgesteld:

H_0: De proportie van incidenten die plaatsvinden in de wijken is *niet* afhankelijk van het dagdeel.
H_1: De proportie van incidenten die plaatsvinden in de wijken is afhankelijk van het dagdeel

Om de Pearson's χ^2 test goed te kunnen is het verplicht dat iedere te schatten waarde een minimale freqentie van 5 heeft. Wat niet het geval is voor de wijken Beursstraat en Nieuwezijds Voorburgwal. Geografisch gezien liggen deze wijken recht naast elkaar en worden voor de Pearson's χ^2 test samengevoegd tot één categorie: Centrum Midden. Op basis van de Pearson's χ^2 test wordt H_0 verworpen en H_1 aangenomen ($\chi^2 = 340,85$; $df = 58$; p-waarde $< 2,2e$-16; $\alpha = 0,05$). Dit betekent dat het aannamelijk is dat woninginbraken over de wijken heen op verschillende tijdstippen binnen de dag plaatsvinden. Dit resultaat sluit aan bij de verwachtingen die al werden geschept in paragraaf 4.2.2 en op districtsniveau werden bewezen in 6.2.1. Er kan een correspondentieanalyse worden uitgevoerd om meer inzicht te krijgen in de χ^2 toetsingsgrootheid. Figuur 5.11 geeft hier op grafische wijze de output van. Opvallend is dat de meeste wijken zich clusteren tussen de nacht, dag en avond, met uitzondering van 5 wijken: Lijnbaansgracht, Raampoort, Centrum Midden (Beursstraat, Nieuwzijds Voorburgwal) en Prinsengracht. Deze wijken vormen samen met de wijk IJburg het district Centrum. In paragraaf 5.2.1 werd ook al opgemerkt dat volgens de correspondentieanalyse district Centrum de meest extreme afwijking kent, wat in deze gedetaileerdere analyse op wijk terugkomt. Hieraan kan wel worden toegevoegd dat het aannemelijk is dat de wijk IJburg minder lijkt te passen in het afwijkende gedrag van district Centrum.



Figuur 5.11: CA met 31 wijken



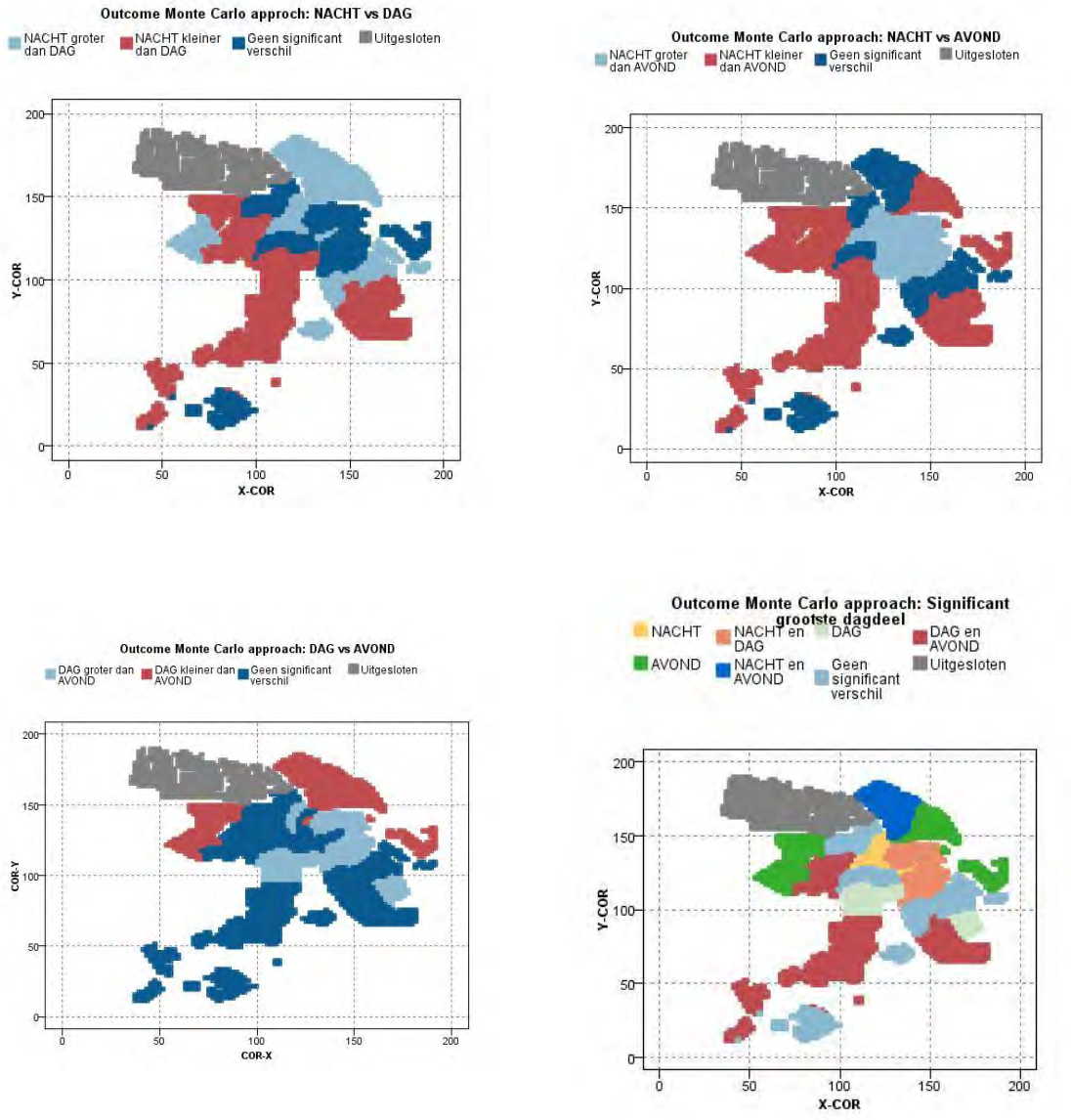
Figuur 5.12: CA met 26 wijken

Wanneer de wijken Lijnbaansgracht, Raampoort, Centrum Midden (Beursstraat, Nieuwzijds Voorburgwal) en Prinsengracht verwijderd worden uit de analyse, wordt wederom op basis van de Pearson's χ^2 test H_0 verworpen en H_1 aangenomen ($\chi^2 = 164,48$; $df = 50$; p-waarde $< 3,98e$-14; $\alpha = 0,05$). Figuur 5.12 geeft ook de output van de correspondentieanalyse aan. De output van de correspondentieanalyse lijkt geen duidelijke informatie te bevatten, maar wanneer de afstand van het centrum naar de desbetreffende wijken wordt meegenomen valt er een verband te zien. Wanneer de districten verdeeld worden op afstand tot het centrum lijkt er een geografisch verband zichtbaar tussen de punten in de correspondentieanalyse. Er zijn suggestief veel mogelijke oorzaken waardoor de geografisch afstand tot het centrum in bepaalde mate de inbraaktrends beinvloed, waarin ook niet de geografisch afstand maar eerder de functie, type wijk

en het type inwoners (etc.) een rol speelt. Suggestief zou dan worden aangenomen dat deze kenmerken correleren met de afstand van de wijk tot aan het centrum. Binnen dit onderzoek wordt daar niet verder op ingegaan.

Monte Carlo benadering

In paragraaf 5.2.1 wordt de Monte Carlo benadering gebruikt om twee geaggregeerde datapatronen te vergelijken. Voor de analyse naar incidenten ten aanzien van de verschillende wijkteams kan deze methode wederom gebruikt worden. De uitkomsten van deze benadering zijn te vinden in de figuur 5.13, waar naast de gebruiktelijk output ook het significant grootste dagdeel is geplot.



Figuur 5.13: Monte Carlo output van links naar rechts, van boven naar onder: nacht vs dag, nacht vs avond, dag vs avond, significant grootste dagdeel

Conclusie - Woninginbraken vinden plaats in verschillende wijken wanneer onderscheid wordt gemaakt in de dagdelen nacht, dag, avond. De meest afwijkende verdeling van incidenten wordt waargenomen bij de wijken toebehorend tot district Centrum (m.u.v. de wijk IJburg), waar het hoogtepunt 's nachts is, terwijl in bijna alle andere wijken 's nachts de minste inbraken plaatsvinden. Het valt op dat de fysieke afstand van de wijken tot het centrum afhangt van de geografische verdeling van incidenten.

5.3.2 Woninginbraken toegekend aan weekdagen

Categorische benadering

Tabel 5.5 geeft de verdeling van incidenten weer over de weekdagen ten aanzien van de 31 wijken. Wanneer de weekdag geen invloed heeft op de locatie waar incidenten plaatsvinden, kan inderdaad gezegd worden dat op wijkniveau de ruimtelijke verdeling mogelijk gelijk is voor de verschillende weekdagen. In paragraaf 5.2 werd echter al onderzoek gedaan naar de afhankelijkheid van district ten aanzien van weekdagen, waaruit bleek dat het aannemelijk is dat de locatie waar incidenten plaatsvinden afhankelijk is van het weekdag, waarbij vooral weekend en werkdagen werden onderscheiden. Hier wordt onderzocht of deze afhankelijkheid ook kan worden gevonden wanneer er gekeken wordt naar wijken in plaats van districten. De data in tabel 5.5 is weergegeven als twee categorische variabelen (wijk en weekdag) waardoor de Pearson's χ^2 test kan nagaan of dagdeel afhankelijk is van de wijk. De volgende hypotheses worden opgesteld:

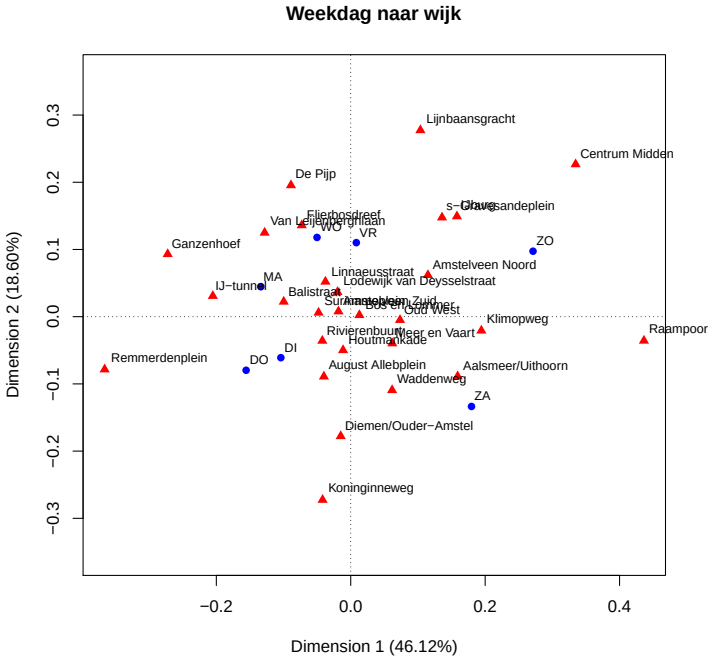
H_0: De proportie van incidenten die plaatsvinden in de wijken is *niet* afhankelijk van de weekdag.
H_1: De proportie van incidenten die plaatsvinden in de wijken is afhankelijk van de weekdag.

Om de Pearson's χ^2 test goed te kunnen is het verplicht dat iedere te schatten waarde een minimale freqentie van 5 heeft. Wat niet het geval is voor alle wijken, waardoor de wijken Aalsmeer en Uithoorn worden samengenomen tot de wijk Alsmeer/Uithoorn en de wijken Beursstraat, Nieuwezijds Voorburgwal en Prinsengracht worden samengenomen tot Centrum Midden. Op basis van de Pearson's χ^2 test wordt H_0 verworpen en H_1 aangenomen ($\chi^2 = 224,73$; $df = 162$; p-waarde $0,0008$; $\alpha = 0,05$). Dit betekent dat het aannemelijk is dat woninginbraken over de wijken heen op verschillende weekdagen plaatsvinden. Dit resultaat sluit aan bij de gevonden resultaten in paragraaf 5.2.2 waar de verdelingen van incidenten op de verschillende weekdagen ruimtelijk worden vergeleken op basis van district. Er kan een correspondentieanalyse worden uitgevoerd om meer inzicht te krijgen in de χ^2 toetsingsgrootheid.
De output van de correspondentieanalyse in 5.14 geeft een redelijke gelijke verdeling weer over alle wijken en dagen alsin, er zijn weinig wijken of dagen die zich extreem differentieren van de rest. De conclusies getrokken in paragraaf 5.2.2 waar de verdelingen van incidenten op de verschillende weekdagen ruimtelijk worden vergeleken op basis van district kunnen eveneens worden onderbouwd door deze analyse op basis van wijk, De wijken van het district Centrum (Centrum Midden, Lijnbaansgracht en Raampoort) gedragen zich afwijkender dan de overige wijken. Deze wijken vormen samen met de wijk IJ-tunnel district Centrum. De wijk IJ-tunnel lijkt zich echter niet te verhouden tot de ruimtelijke afwijkingen die zichtbaar zijn bij de andere wijken van district Centrum. Dat de wijk IJ-tunnel zo afwijkt van de andere

Wijk	MA	DI	WO	DO	VR	ZA	ZO	Totaal
Aalsmeer	6	3	6	8	7	8	4	42
Amstelveen Noord	8	17	20	16	8	18	18	105
Amstelveen Zuid	22	17	22	21	8	24	16	130
August Allebplein	51	47	31	53	26	57	34	299
Balistraat	25	21	20	22	13	22	14	137
Beursstraat	4	3	3	1	0	2	4	17
Bos en Lommer	30	32	29	27	16	36	25	195
De Pijp	16	8	12	12	9	9	10	76
Diemen/Ouder-Amstel	15	22	17	25	11	30	12	132
Flierbosdreef	28	16	24	23	14	20	18	143
Ganzenhoef	38	29	40	47	22	22	16	214
Houtmankade	21	23	15	23	12	24	18	136
IJburg	13	12	11	6	5	12	15	74
IJ-tunnel	22	19	13	16	12	14	8	104
Klimopweg	40	46	52	52	28	81	62	361
Koninginneweg	9	14	8	14	2	16	7	70
Lijnbaansgracht	7	9	11	4	9	8	9	57
Linnaeusstraat	25	29	36	28	13	30	22	183
Lodewijk van Deysselstraat	44	43	63	48	25	59	34	316
Meer en Vaart	38	35	42	48	29	63	35	290
Nieuwezijds Voorburgwal	2	0	4	1	3	3	6	19
Oud West	10	13	22	16	8	23	12	104
Prinsengracht	6	3	3	4	5	9	8	38
Raampoort	10	10	8	11	9	26	23	97
Remmerdenplein	37	40	27	45	12	20	11	192
Rivierenbuurt	9	5	11	10	5	13	4	57
s-Gravesandeplein	13	11	14	10	7	14	16	85
Surinameplein	22	26	36	36	10	30	23	183
Uithoorn	4	6	8	4	3	17	5	47
Van Leijenberghlaan	21	20	27	21	14	18	13	134
Waddenweg	57	51	46	54	27	86	42	363
Totaal	653	630	681	706	372	814	544	4400

Tabel 5.5: Aantal woninginbraken naar weekdag per wijkteam

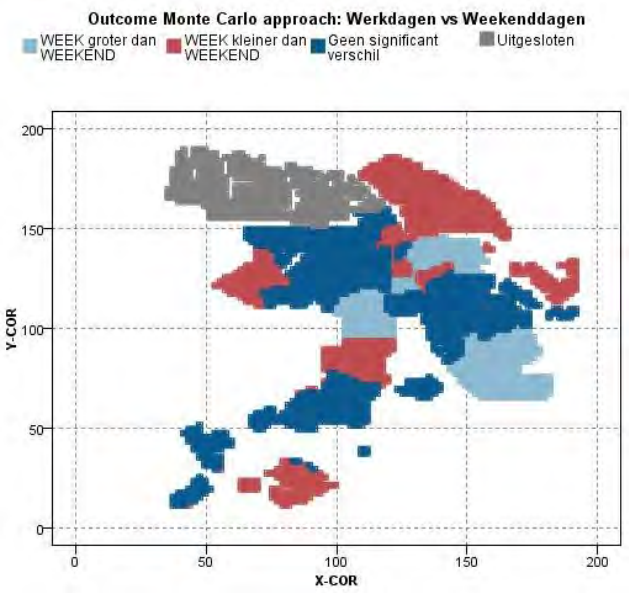
wijken in het district werd ook al aangetoond in paragraaf 5.3.1 Naast het district Centrum kan er over deze weekdaganalyse worden gezegd dat de weekenddagen zaterdag en zondag zich eveneens ruimtelijk anders gedragen dan de werkdagen.



Figuur 5.14: Correspondentieanalyse van wijken en weekdagen

Monte Carlo benadering

In voorgaande paragrafen werd de Monte Carlo benadering gebruikt om twee geaggregeerde ruimtelijke datapatronen te vergelijken. Voor de analyse naar incidenten ten aanzien van de verschillende wijkteams kan deze methode wederom worden gebruikt. In deze analyse worden 7 weekdagen vergeleken wat bij een Monte Carlo benadering 21 simulaties vereist en 21 verschillende plots oplevert. In deze analyse wordt door dit grote aantal verschillende plots alleen een analyse gemaakt op de werkdagen ten aanzien van de weekenddagen. De output hiervan is weergegeven in figuur 5.15.



Figuur 5.15: Monte Carlo: Werkdagen vs weekenddagen op basis van wijk

Deze plot geeft geen eenduidig beeld qua ruimtelijke verdeling. Noord kent wederom een verhoogde incidentrate in de weekenden. Een groot deel van Oost kent juist een verhoogd aantal incidenten op de werkdagen. De wijken in district Centrum zijn niet eenduidig, maar de aantallen incidenten waaruit is gesampled per wijk zijn ook laag.

Conclusie - Woninginbraken vinden in verschillende wijken plaats op basis van de betreffende weekdag. Het grootste verschil kan gevonden worden tussen de werkdagen en weekenddagen.

5.4 Straatroven toegekend aan districten

De regio Amsterdam is binnen de politie onderverdeeld in 5 districten: Centrum, Noord, Oost, West en Zuid[3]. In de peilperioden 177 t/m 197 (zie paragraaf 3.3) zijn er 1.519 straatroven geregistreerd die hebben plaatsgevonden in een van deze vijf districten. In deze paragraaf wordt onderzocht of voor verschillende tijdsintervallen, incidenten in dezelfde districten plaatsvinden.

5.4.1 Straatroven toegekend aan dagdelen

Categorische benadering

Tabel 5.6 geeft de verdeling van incidenten weer over de dagdelen nacht, dag en avond ten aanzien van de vijf districten.

Dagdeel	Centrum	Noord	Oost	Zuid	West	Totaal
Nacht	270	41	67	48	74	500
Dag	49	47	103	27	68	294
Avond	195	106	220	72	132	725
Totaal	514	194	390	147	274	1.519

Tabel 5.6: Aantal straatroven naar dagdeel per district

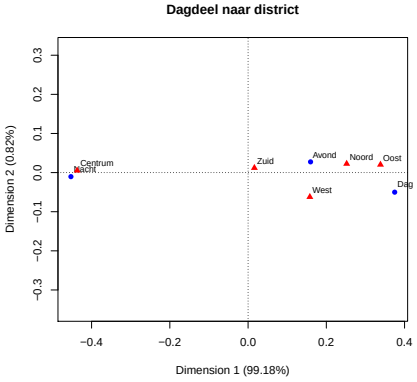
Wanneer het dagdeel geen invloed heeft op de locatie waar incidenten plaatsvinden, kan inderdaad gezegd worden dat op districtsniveau de ruimtelijke verdeling van incidenten mogelijk gelijk is voor de verschillende dagdelen. De data in tabel 5.6 is weergegeven als twee categorische variabelen (district en dagdeel) waardoor de Pearsons χ^2 test kan nagaan of dagdeel afhankelijk is van het district. De volgende hypotheses worden opgesteld:

H_0: De proportie van incidenten die plaatsvinden in de districten is *niet* afhankelijk van het dagdeel.
H_1: De proportie van incidenten die plaatsvinden in de districten is afhankelijk van het dagdeel.

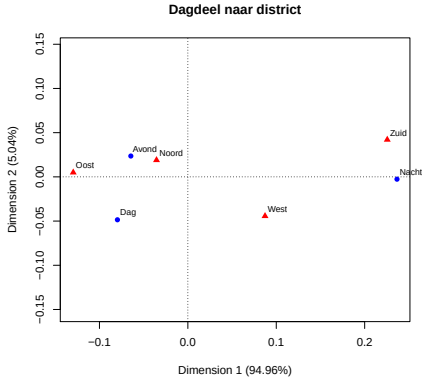
Op basis van de Pearsons χ^2 test wordt H_0 verworpen en H_1 aangenomen ($\chi^2 = 163,43$; $df = 8$; p-waarde $< 2,2e-16$; $\alpha = 0,05$). Dit betekent dat het aannamelijk is, dat straatroven over de districten heen

[3]Dit is de onderverdeling ten tijde van dit onderzoek. Deze structuur is mogelijk onderhevig aan reorganisatie van de district- wijkteams in de toekomst

op verschillende tijdstippen binnen de dag plaatsvinden. Paragraaf 4.3.2 gaf hier ook al indicatie toe. Er kan een correspondentieanalyse worden uitgevoerd om de variantie te decompenseren in verschillende dimensies. Figuur 5.16 geeft hier op grafische wijze de output van. Vooral het district Centrum kent een extreme afwijking t.a.v. de overige vier districten. In dezelfde hoek ligt ook het nachtelijke tijdsinterval. In cijfers is dit te onderbouwen: 52,53% van de straatroven in het Centrum vindt 's nachts plaats (ten aanzien van 9,53% overdag en 37,94% 's avonds), terwijl in alle andere districten het aantal incidenten 's nachts lager is dan overdag en 's avonds (m.u.v. overdag in Zuid en West).



Figuur 5.16: CA met vijf districten



Figuur 5.17: CA met vier districten (zonder Centrum)

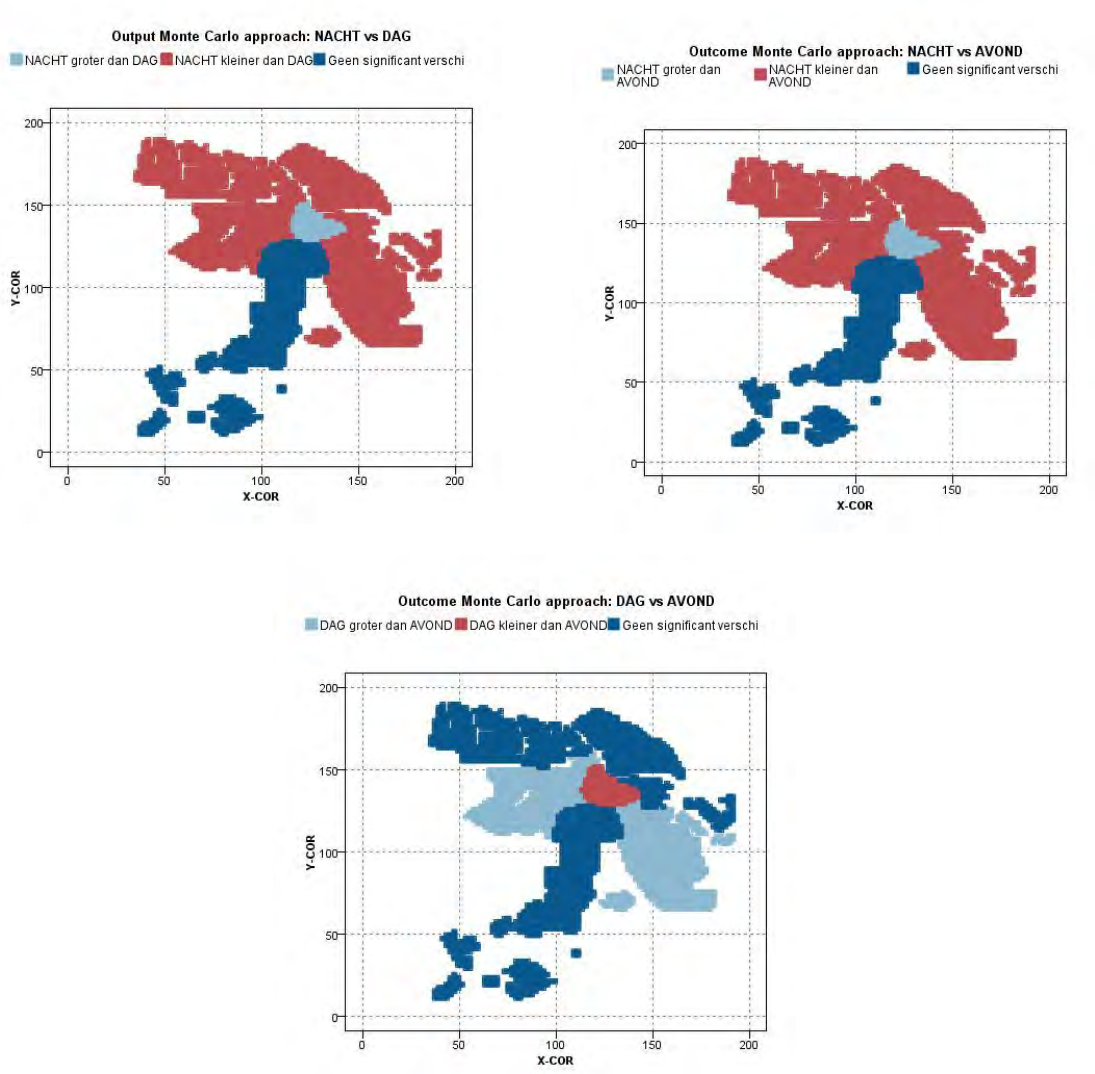
Wanneer alleen gekeken wordt naar de districten Noord, West, Zuid en Oost wordt op basis van de Pearsons χ^2 test H_0 verworpen en H_1 aangenomen ($\chi^2 = 17,34$; $df = 6$; p-waarde $< 0,0081$; $\alpha = 0,05$). Figuur 5.17 geeft de output van de correspondentieanalyse aan. In deze analyse wijken juist nacht en zuid uit. Tijdens de nacht vindt 54% van de incidenten in het Centrum plaats en nu deze uit de selectie is verwijderd lijkt nacht te hangen aan district Zuid. In district Zuid gebeuren relatief de minste straatroven (9,68%) en mag dus wel meer afwijken dan de districten Noord, Oost en West, maar dit betreft ook een laag aantal incidenten. Verder gebeuren de meeste inidenten in aantal 's avonds en overdag in Oost. Wanneer alle districten als duo worden onderworpen aan de χ^2 test, wordt voor alke duo's H_0 verworpen en H_1 aangenomen met een p-waarde $< 0,05$, uitgezonderd de duo's Oost & Noord, Noord & West en West & Zuid. Voor alle duo's in dagdelen wordt H_0 verworpen en H_1 aangenomen met een p-waarde $< 0,05$.

Monte Carlo benadering

De geografisch data uit tabel 5.6 is een cijfermatige weergave van een areal ruimtelijke datastructuur (paragraaf 2.4.1). Een techniek om zulke datapatronen te vergelijken op gelijkenis is de non-parametrische Monte Carlo benadering (paragraaf 2.4.2, 5.1).

Voor de dagdelen en districten is de Monte Carlo benadering gebruikt, ondanks dat vijf gebieden voor het gebruik van de Monte Carlo approach aan de lage kant is. De uitkomsten van deze benadering zijn te vinden in de figuur 5.18. Ook uit deze analysetechniek wordt duidelijk dat het district Centrum 's nachts een hogere intensiteit aan straatroven kent dan overdag en 's avonds. Overdag en avond is daarnaast lastiger. 's Avonds vinden relatief meer straatroven plaats in het Centrum, maar lang niet zo

extreem als in de nacht. Overdag zijn er daardoor relatief meer straatroven in West en Oost, maar dit ligt eerder aan de weinige straatroven in het Centrum, dan aan een toename in West en Oost.



Figuur 5.18: Monte Carlo output van links naar rechts van boven naar onder: nacht vs dag, nacht vs avond, dag vs avond

Conclusie - Straatroven vinden plaats in verschillende districten wanneer onderscheid wordt gemaakt in de dagdelen nacht, dag, avond. De meest afwijkende verdeling van straatroven wordt waargenomen in district Centrum, waar het hoogtepunt 's nachts is, terwijl alle andere districten 's nachts de minste straatroven plaatsvinden.

5.4.2 Straatroven toegekend aan weekdagen

Categorische benadering

Tabel 5.7 geeft de verdeling van incidenten weer over de weekdagen ten aanzien van de vijf districten. Wanneer de weekdag geen invloed heeft op de locatie waar incidenten plaatsvinden, kan inderdaad gezegd

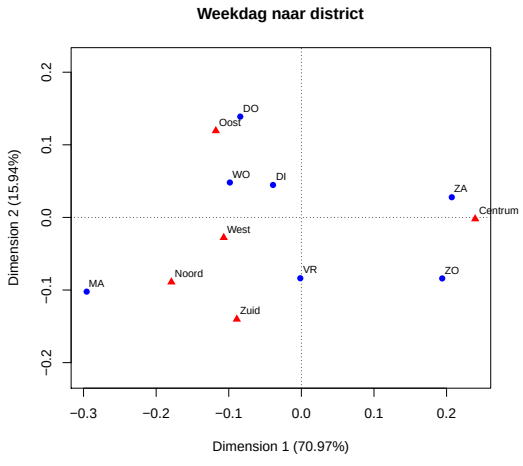
Weekdag	Centrum	Noord	Oost	Zuid	West	Totaal
Maandag	44	40	60	29	46	219
Dinsdag	64	28	55	16	35	198
Woensdag	62	25	66	26	37	216
Donderdag	62	26	65	13	41	207
Vrijdag	42	17	25	10	30	124
Zaterdag	122	26	66	25	42	281
Zondag	118	32	53	28	43	274
Totaal	514	194	390	147	274	1.519

Tabel 5.7: Aantal straatroven naar weekdag per district

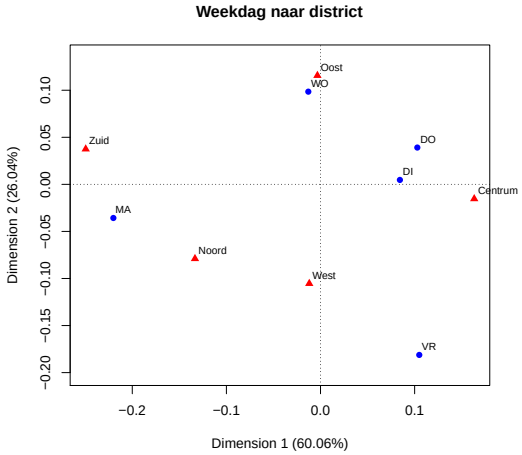
worden dat er op districtsniveau de ruimtelijke verdeling mogelijk gelijk is voor verschillende weekdagen. De data in tabel 5.7 is evenals in de vorige paragraaf weergegeven als twee categorisch variabelen (district en weekdag) waardoor de Pearsons χ^2 test kan nagaan of weekdag afhankelijk is van het district. De volgende hypotheses worden opgesteld:

H_0: De proportie van incidenten die plaatsvinden in de districten is *niet* afhankelijk van de weekdag.
H_1: De proportie van incidenten die plaatsvinden in de districten is afhankelijk van de weekdag.

Op basis van de Pearsons χ^2 test test wordt H_0 verworpen en H_1 aangenomen ($\chi^2 = 63,95$; $df = 24$; p-waarde $= 1,73e\text{-}5$; $\alpha = 0,05$). Dit betekent dat het aannamelijk is, dat straatroven over de districten heen op verschillende weekdagen plaatsvinden. Paragraaf 4.2.1 probeerde ook de mogelijkheden hiervoor te onderzoeken, maar kwam tot een klein zichtbaar verschil tussen de weekenddagen en weekdagen. Er kan een correspondentieanalyse worden uitgevoerd om de variantie te decompenseren in verschillende dimensies.



Figuur 5.19: CA met vijf districten



Figuur 5.20: CA met vier districten

Figuur 5.19 geeft op grafische wijze de output van de correspondentieanalyse. Er zijn twee aspecten die hier lijken op te vallen: (1) het district Centrum kent een extreme afwijking ten aanzien van de overige

districten en (2) de weekeindedagen zaterdag, zondag en in mindere mate vrijdag kennen een afwijking ten aanzien van de overige dagen. Het feit dat district Centrum zich afwijkend gedraagd werd ook al opgemerkt in paragraaf 5.3.1. De afwijking van de dagen vrijdag, zaterdag en zondag is nog niet eerder zo duidelijk opgemerkt. Wanneer alleen gekeken wordt naar de districten Noord, West, Zuid en Oost wordt op basis van de Pearsons χ^2 test kan H_0 niet worden verworpen ($\chi^2 = 18,83$; $df = 18$; p-waarde $= 0,402$; $\alpha = 0,05$). Zonder de weekenddagen zaterdag en zondag, maar met district Centrum kan H_0 niet worden verworpen ($\chi^2 = 25,99$; $df = 16$; p-waarde $= 0,054$; $\alpha = 0,05$). Figuur 5.20 geeft de output van de bijhorende correspondentieanalyse.

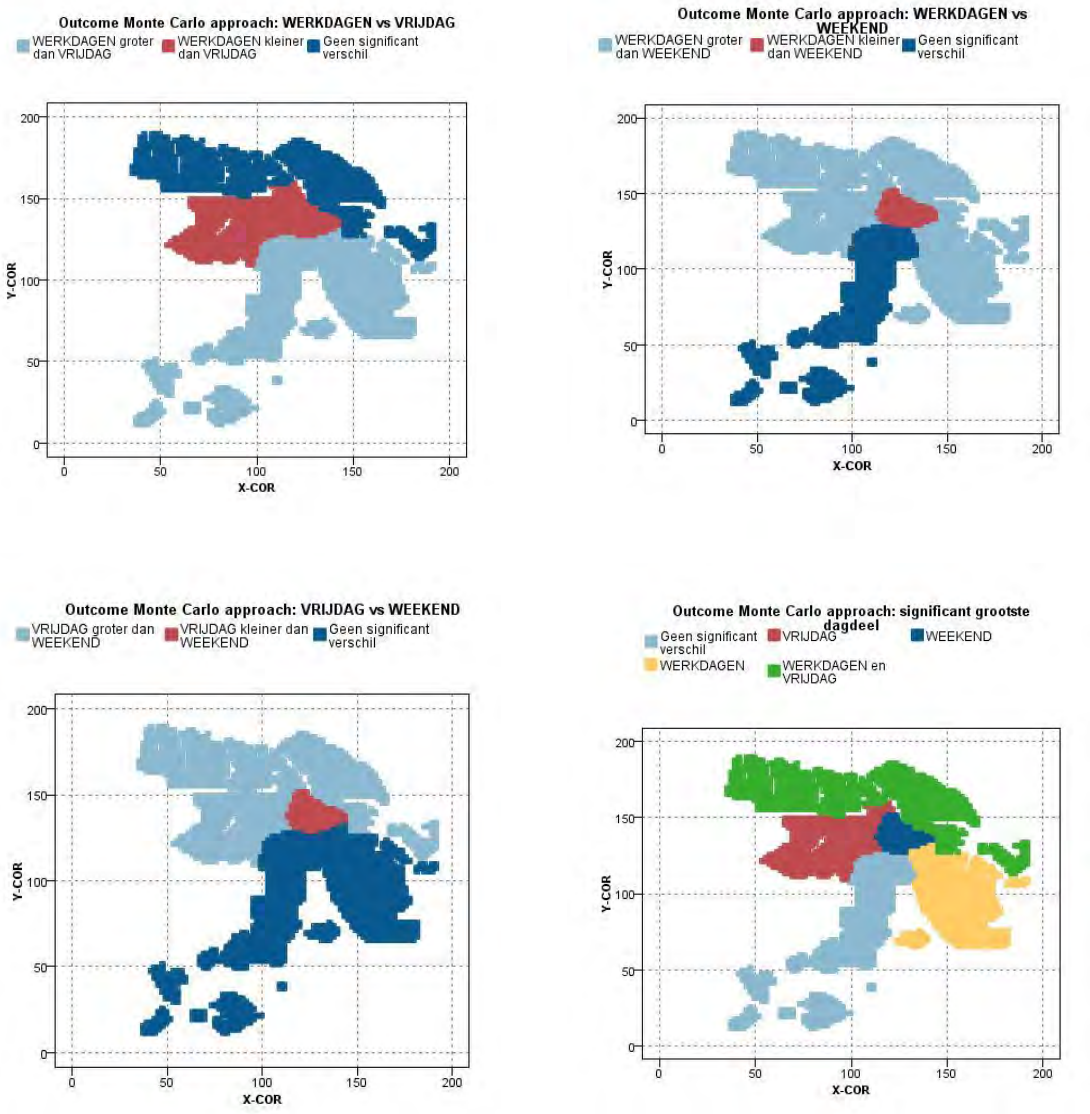
Een aannemelijke verwachting is dat de analyse van de weekdagen zonder vrijdag, zaterdag en zondag de χ^2 waarde verder zal laten dalen waarmee de overtuiging op verwerpen van H_0 afneemt. Wanneer echter alleen de dagen maandag t/m donderdag aan de Pearsons χ^2 test worden onderworpen kan H_0 niet worden verworpen ($\chi^2 = 19,34$; $df = 12$; p-waarde $= 0,081$; $\alpha = 0,05$). Het lijkt er dus op dat vrijdag de duidelijke schakel is tussen de week- en weekenddagen. Voor de dagen vrijdag, zaterdag en zondag kan de Pearsons χ^2 test worden onderworpen kan H_0 niet worden verworpen ($\chi^2 = 10,33$; $df = 8$; p-waarde $= 0,24$; $\alpha = 0,05$).

Monte Carlo benadering

In voorgaande paragrafen wordt de Monte Carlo benadering gebruikt om twee geaggregeerde datapatronen te vergelijken. Voor de analyse naar incidenten ten aanzien van de verschillende districten kan deze methode wederom gebruikt worden. Als alle weekdagen met elkaar vergeleken worden, zijn er 21 (6de partiele som van $\frac{n(n+1)}{2}$, met $n = 6$) analyses nodig en komen daar 21 plots uit. De categorische analyse wijst vooral op een verschil tussen de werkdagen en weekenddagen. De rol van de vrijdag lijkt daarbij wat discutabel. Voor deze analyse wordt onderscheid gemaakt tussen drie type dagen: werkdagen (ma t/m do), vrijdag en weekenddagen. Deze drie geaggregeerde datapatronen worden ter vergelijking onderworpen aan de Monte Carlo benadering. De uitkomsten zijn te vinden in figuur **??**, waar naast de gebruikelijk output ook het significant grootste dagdeel is geplot.

Uit deze analyse blijkt dat op basis van relatieve percentages Centrum zich verhouwd tot het weekend, oost tot de werkdagen en West tot vrijdag. Noord lijkt lastiger te interpreteren en verhoudt zich tot zowel vrijdag als de werkdagen. Wel moet er bij het interpeteren wel rekening worden gehouden met het lage aantal geografische gebieden, wat de uitkomsten onderhevig maakt aan enkele uitschieters. Vooral de straatroven in het Centrum op de weekenddagen zorgen voor een licht vertekent beeld.

Conclusie - Straatroven vinden in verschillende districten plaats op basis van de betreffende weekdag. Het grootste verschil kan gevonden worden tussen de werkdagen en weekenddagen. Tussen de werkdagen maandag t/m vrijdag onderling en tussen de weekenddagen zaterdag en zondag onderling lijken geen grote verschillen te zijn, al lijkt de vrijdag zich meer afwijkend te gedragen ten aanzien van de overige werkdagen.



Figuur 5.21: Monte Carlo output van links naar rechts, van boven naar onder: ma t/m do vs vrijdag, ma t/m do vs weekenddagen, vrijdag vs weekenddagen, significant grootste weekdagsegment

5.5 Straatroven toegekend aan wijkteams

De regio Amsterdam is binnen de politie onderverdeeld in 5 districten die weer zijn opgedeeld in 31 wijken[4]. In de peilperioden 177 t/m 197 (zie paragraaf 3.3) zijn er 1.519 straatroven geregistreerd die hebben plaatsgevonden in een van deze 31 wijken.

5.5.1 Straatroven toegekend aan dagdelen

Categorische benadering

Tabel 5.8 geeft de verdeling van incidenten weer over de dagdelen nacht, dag en avond ten aanzien van de 31 wijken.

[4]Dit is de onderverdeling ten tijde van dit onderzoek. Deze structuur is mogelijk onderhevig aan reorganisatie van de district- wijkteams in de toekomst

Wijk	Nacht	Dag	Avond	Totaal
Aalsmeer	0	1	0	1
Amstelveen Noord	0	0	9	9
Amstelveen Zuid	2	1	4	7
August Allebplein	13	16	24	53
Balistraat	7	6	31	44
Beursstraat	42	7	32	81
Bos en Lommer	10	5	14	29
De Pijp	17	1	10	28
Diemen/Ouder-Amstel	6	4	15	25
Flierbosdreef	16	29	65	110
Ganzenhoef	16	20	42	78
Houtmankade	8	7	15	30
IJburg	1	1	6	8
IJ-tunnel	52	14	60	126
Klimopweg	14	13	25	52
Koninginneweg	18	8	15	41
Lijnbaansgracht	42	7	25	74
Linnaeusstraat	11	13	35	59
Lodewijk van Deysselstraat	11	12	15	38
Meer en Vaart	5	12	17	34
Nieuwezijds Voorburgwal	43	7	29	79
Oud West	16	9	22	47
Prinsengracht	69	5	25	99
Raampoort	22	9	24	55
Remmerdenplein	4	15	35	54
Rivierenbuurt	4	7	13	24
s-Gravesandeplein	14	22	28	64
Surinameplein	11	7	25	43
Uithoorn	1	0	4	5
Van Leijenberghlaan	6	9	17	32
Waddenweg	19	27	44	90
Totaal	500	294	725	1.519

Tabel 5.8: Aantal incidenten naar dagdeel per wijk

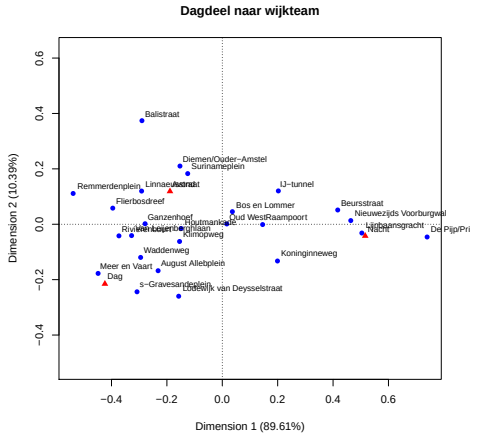
Wanneer het dagdeel geen invloed heeft op de locatie waar incidenten plaatsvinden, kan inderdaad gezegd worden dat op wijkniveau de ruimtelijke verdeling mogelijk gelijk is voor de verschillende dagdelen. In paragraaf 5.3 werd echter al onderzoek gedaan naar de afhankelijkheid van district ten aanzien van

dagdelen, waaruit bleek dat het aannemelijk is dat de locatie waar incidenten plaatsvinden afhankelijk is van het dagdeel. Hier wordt onderzocht of deze afhankelijkheid ook kan worden gevonden wanneer er gekeken wordt naar wijken in plaats van districten. De data in tabel 5.8 is weergegeven als twee categorische variabelen (wijk en dagdeel) waardoor de Pearson's χ^2 test kan nagaan of dagdeel afhankelijk is van de wijk. De volgende hypotheses worden opgesteld:

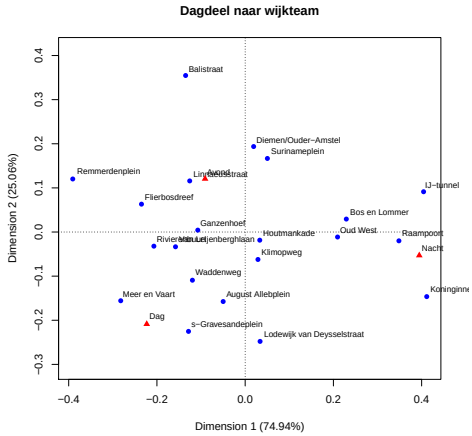
H_0: De proportie van incidenten die plaatsvinden in de wijken is *niet* afhankelijk van het dagdeel.
H_1: De proportie van incidenten die plaatsvinden in de wijken is afhankelijk van het dagdeel.

Om de Pearson's χ^2 test goed te kunnen is het verplicht dat iedere te schatten waarde een minimale freqentie van 5 heeft. Dit is lang niet voor alle wijken het geval. In de voorgaande hoofdstukken werden vaak enkele wijken samengevoegd, maar bij straatroven gebeuren in enkele wijken zo weinig incidenten dat optellen met omliggende wijken ook niet leidt tot een aantal incidenten boven de 5. De wijken Amstelveen Noord, Amstelveen Zuid, Aalsmeer en Uithoon kunnen geografisch gezien bij elkaar worden opgeteld, maar komen samen tot overdag 2 straatroven en 's nachts tot 3 straatroven. Samenvoegen met de bovenliggende wijk Van Leijenberghlaan is mogelijk, maar hier is de relatieve verdeling van incidenten beduidend anders dan van de 4 samengevoegde wijken. Op basis van deze analyse zijn de wijken Amstelveen Nood, Amstelveen Zuid, Aalsmeer en Uithoorn verwijderd uit de dataset, evenals de wijk IJburg. Tot slot zijn de wijken Prinsengracht en de Pijp samengevoegd tot één categorie.

Op basis van de Pearson's χ^2 test wordt H_0 verworpen en H_1 aangenomen ($\chi^2 = 233,71$; $df = 48$; p-waarde $< 2,2e\text{-}16$; $\alpha = 0,05$). Dit betekent dat het aannamelijk is, dat straatroven over de wijken heen op verschillende tijdstippen binnen de dag plaatsvinden. Dit resultaat sluit aan bij de verwachtingen die al werden geschept in paragraaf 4.3.2 en op districtsniveau werden bewezen in 5.4.1. Er kan een correspondentieanalyse worden uitgevoerd om meer inzicht te krijgen in de Pearsons χ^2 statistic.



Figuur 5.22: CA met 26 wijken



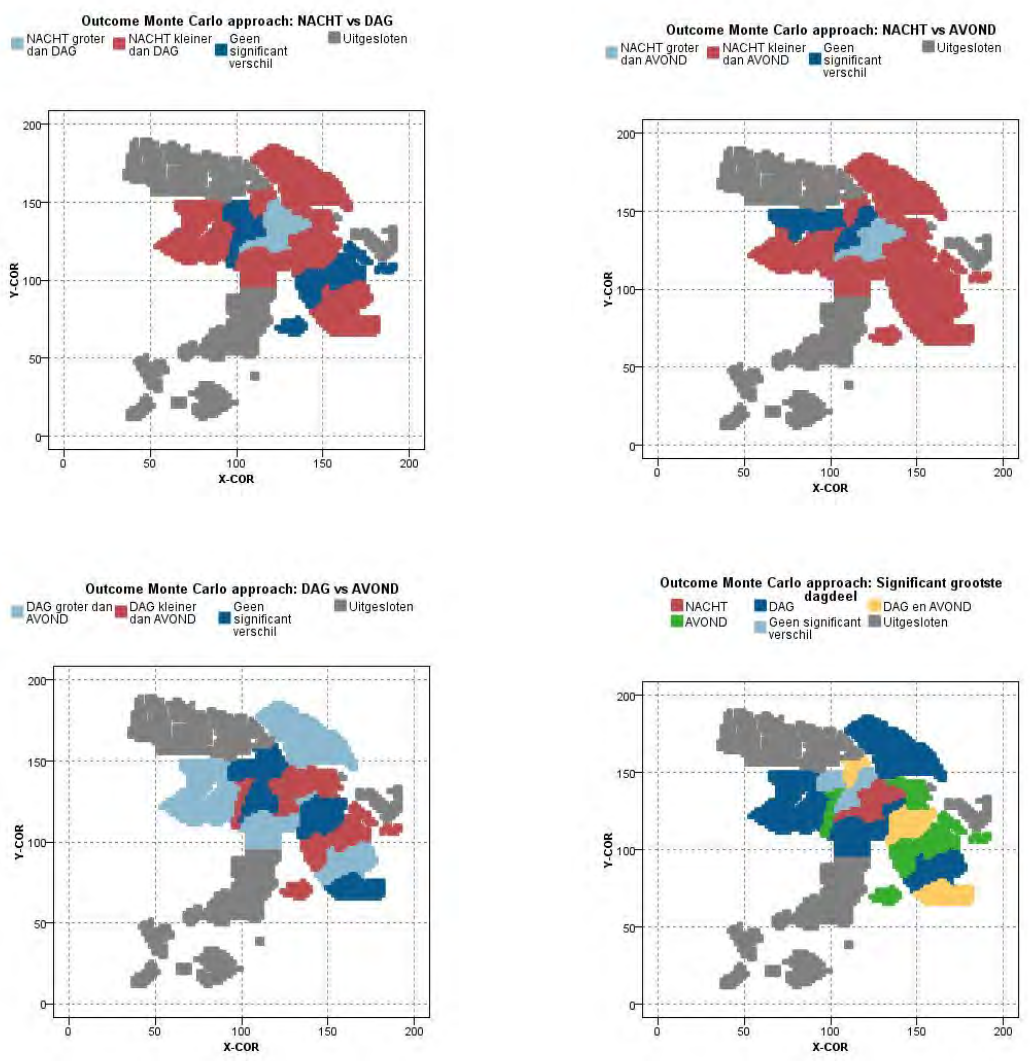
Figuur 5.23: CA met 21 wijken

Figuur 5.22 geeft hier op grafische wijze de output van. De wijken die toebehoren tot het district Centrum (Beursstraat, Nieuwezijds Voorburgwal, Lijnbaasgracht, De Pijp/Prinsengracht) liggen allemaal rondom het dagdeel nacht. De wijken die tussen de dagdelen avond/dag en nacht liggen, bevinden zich geografisch gezien rondom het centrum. De wijken rondom dag en avond liggen juist aan de rand van Amsterdam.

Wanneer de wijken Beursstraat, Nieuwezijds Voorburgwal, Lijnbaansgracht en De Pijp/Prinsengracht verwijderd worden uit de analyse, wordt wederom op basis van de Pearsons χ^2 test H_0 verworpen en H_1 aangenomen ($\chi^2 = 82,74$; $df = 40$; p-waarde $< 8,28e\text{-}5$; $\alpha = 0,05$). Figuur 5.12 geeft ook de output van de correspondentieanalyse aan. Deze output komt volledig overeen met de output uit de vorige correspondentieanalyse, alleen is het dagdeel nacht hiermee opgeschoven de de richting van de oorsprong van de grafiek. Zoals ook bij woninginbraken werd gevonden lijkt de afstand tot het centrum ook bij straatroven ten aanzien van de wijken parten te spelen. Wanneer alleen het dagdeel nacht wordt verwijderd uit de dataset (en de wijken Beursstraat, Nieuwezijds Voorburgwal, Lijnbaansgracht en De Pijp/Prinsengracht weer worden toegevoegd) kan H_0 niet worden verworpen ($\chi^2 = 34,83$; $df = 24$; p-waarde $= 0,071$; $\alpha = 0,05$).

Monte Carlo benadering

De Monte Carlo benadering kan worden gebruikt om twee geaggregeerde datapatronen te vergelijken zoals in dit geval wijkteam en dagdeel.



Figuur 5.24: Monte Carlo output van links naar rechts van boven naar onder: nacht vs dag, nacht vs avond, dag vs avond, het significant grootste dagdeel.

De uitkomsten van deze benadering zijn te vinden in de figuur 5.24 waar naast de gebruikelijk output ook het significant grootste dagdeel is geplot. In de nacht kennen de wijken toebehorend tot stadsdeel centrum het relatief hoogste percentage straatroven. Alle overige wijken zijn relatief hoog aan de dagdelen avond en dag.

Conclusie - Straatroven vinden plaats in verschillende wijken wanneer onderscheid wordt gemaakt in de dagdelen nacht, dag, avond. De meest afwijkende verdeling van straatroven wordt waargenomen in de wijken toebehorend tot district Centrum (m.u.v. Raampoort en IJ-tunnel), waar het hoogtepunt 's nachts is, terwijl alle andere districten 's nachts de minste straatroven plaatsvinden.

5.5.2 Straatroven toegekend aan weekdagen

Categorische benadering

Tabel 5.9 geeft de verdeling van incidenten weer over de weekdagen ten aanzien van de 31 wijken. Wanneer de weekdag geen invloed heeft op de locatie waar straatroven plaatsvinden, kan inderdaad gezegd worden dat op wijkniveau de ruimtelijke verdeling mogelijk gelijk is voor de verschillende weekdagen. In paragraaf 5.3 werd echter al onderzoek gedaan naar de afhankelijkheid van district ten aanzien van weekdagen, waaruit bleek dat het aannemelijk is dat de locatie waar incidenten plaatsvinden afhankelijk is van het weekdag, waarbij vooral weekend en werkdagen werden onderscheiden. Nu wordt onderzocht of deze afhankelijkheid ook kan worden gevonden wanneer er gekeken wordt naar wijken in plaats van districten. De data in tabel 5.9 is weergegeven als twee categorische variabelen (wijk en weekdag) waardoor de Pearson's χ^2 test kan nagaan of dagdeel afhankelijk is van de wijk. De volgende hypotheses worden opgesteld:

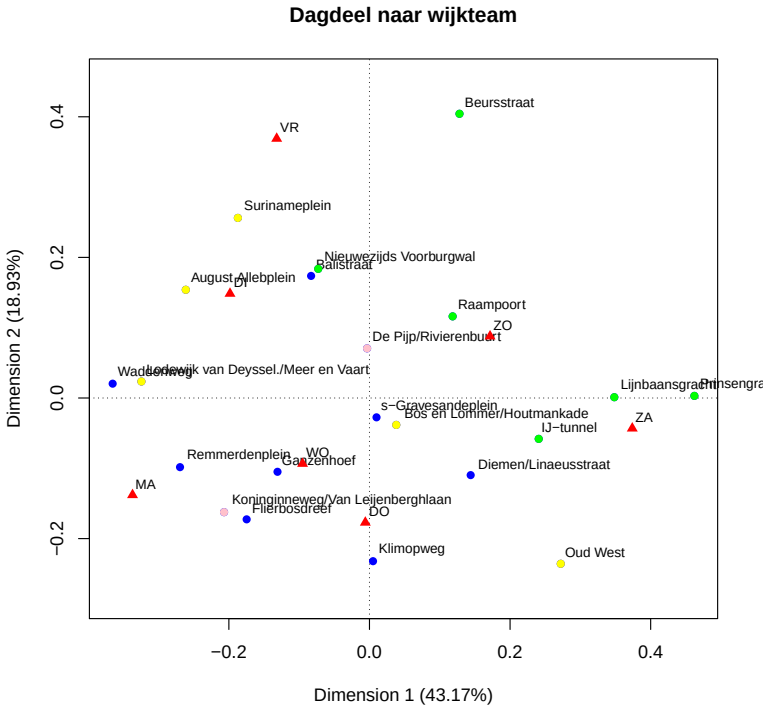
H_0: De proportie van incidenten die plaatsvinden in de wijken is *niet* afhankelijk van de weekdag.
H_1: De proportie van incidenten die plaatsvinden in de wijken is afhankelijk van de weekdag.

Om de Pearson's χ^2 test goed te kunnen is het verplicht dat iedere te schatten waarde een minimale freqentie van 5 heeft. Dit is lang niet voor alle wijken het geval. In de voorgaande hoofdstukken werden vaak enkele wijken samengevoegd, maar bij straatroven gebeuren in enkele wijken zo weinig incidenten dat optellen met omliggende wijken ook niet leidt tot een aantal incidenten boven de 5. De wijken Amstelveen Noord, Amstelveen Zuid, Aalsmeer, Uithoon en IJburg worden verwijderd uit de dataset doordat daar vrijwel geen incidenten hebben plaatsgevonden. De volgende wijken worden samengevoegd: Bos en Lommer en Houtmankade, Pijp en Rivierenbuurt, Konninginneweg en Van Leijenberglaan, Lodewijk van Deyssel. en Meer en Vaart. Op basis van de Pearson's χ^2 test wordt H_0 verworpen en H_1 aangenomen ($\chi^2 = 189,96$; $df = 120$; p-waarde $< 4,94e\text{-}5$; $\alpha = 0,05$). Dit betekent dat het aannemelijk is, dat straatroven over de wijken heen op verschillende weekdagen plaatsvinden. Dit resultaat sluit aan bij de gevonden resultaten in paragraaf 5.3.2 waar de verdelingen van incidenten op de verschillende weekdagen ruimtelijk worden vergeleken op basis van district. Er kan een correspondentieanalyse worden uitgevoerd om meer inzicht te krijgen in de Pearson's χ^2 statistic.

Wijk	MA	DI	WO	DO	VR	ZA	ZO	Totaal
Aalsmeer	0	0	1	0	0	0	0	1
Amstelveen Noord	1	1	1	0	0	4	2	9
Amstelveen Zuid	0	3	1	0	0	2	1	7
August Allebplein	11	7	6	7	9	7	6	53
Balistraat	7	6	5	5	6	7	8	44
Beursstraat	2	14	11	8	13	15	18	81
Bos en Lommer	4	1	3	6	4	5	6	29
De Pijp	6	2	3	2	3	8	4	28
Diemen/Ouder-Amstel	3	3	4	4	1	5	5	25
Flierbosdreef	20	16	20	19	6	16	13	110
Ganzenhoef	9	10	15	18	7	9	10	78
Houtmankade	5	1	6	4	3	5	6	30
IJburg	2	1	0	1	1	2	1	8
IJ-tunnel	11	10	17	22	8	28	30	126
Klimopweg	8	7	7	11	2	11	6	52
Koninginneweg	9	2	5	6	4	5	10	41
Lijnbaansgracht	7	8	7	9	5	24	14	74
Linnaeusstraat	7	6	8	10	4	15	9	59
Lodewijk van Deysselstraat	8	7	8	4	4	3	4	38
Meer en Vaart	7	7	4	4	2	4	6	34
Nieuwezijds Voorburgwal	10	17	8	9	6	10	19	79
Oud West	5	5	6	9	1	14	7	47
Prinsengracht	8	7	10	10	4	30	30	99
Raampoort	6	8	9	4	6	15	7	55
Remmerdenplein	11	13	7	8	2	8	5	54
Rivierenbuurt	2	6	5	2	2	4	3	24
s-Gravesandeplein	10	7	12	6	5	13	11	64
Surinameplein	6	7	4	7	7	4	8	43
Uithoorn	2	0	1	0	0	0	2	5
Van Leijenberghlaan	9	2	9	3	1	2	6	32
Waddenweg	23	14	13	9	8	6	17	90
Totaal	219	198	216	207	124	281	274	1.519

Tabel 5.9: Aantal incidenten naar weekdag per wijkteam

De output van de correspondentieanalyse in 5.25 geeft een redelijke gelijke verdeling weer over alle wijken en dagen alsin, er zijn weinig wijken of dagen die zich extreem differentieren van de rest. De punten in de plot zijn gekleurd naar aanleiding van het district waartoe ze behoren. Er is een duidelijke clustering van



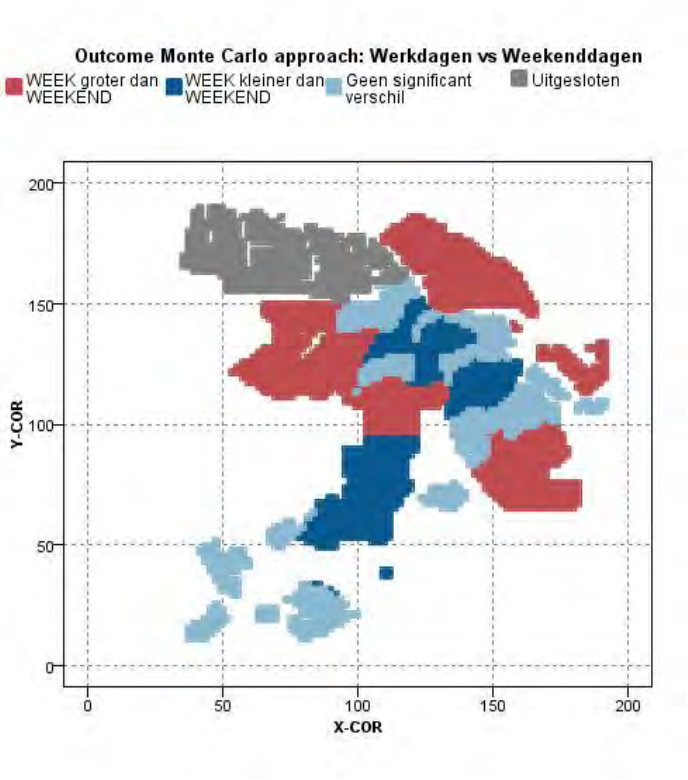
Figuur 5.25: Monte Carlo: Weekdag op basis van wijk

de wijken bijhorend bij district West (blauw) in de linkeronderhoek en een clustering van de centrumwijken (groen) in de rechterbovenhoek. De weekdagen zaterdag en zondag bevinden zich ook in de richting van de centrumwijken. Voor de dagen maandag t/m vrijdag kan H_0 niet worden verworpen ($\chi^2 = 92,45$; $df = 80$; p-waarde$= 0,16$; $\alpha = 0,05$). Voor de dagen zaterdag en zondag kan H_0 niet worden verworpen ($\chi^2 = 27,00$; $df = 20$; p-waarde $= 0,14$). Voor de dagen vrijdag, zaterdag en zondag wordt H_0 verworpen en H_1 aangenomen ($\chi^2 = 61,12$; $df = 40$; p-waarde $= 0,017$; $\alpha = 0,05$). Hieruit blijkt dat de dagen maandag t/m vrijdag vermoedelijk een andere verdeling kennen dan de weekenddagen zaterdag en zondag.

Monte Carlo benadering

In voorgaande paragrafen werd de Monte Carlo benadering gebruikt om twee geaggregeerde ruimtelijke datapatronen te vergelijken. Voor de analyse naar incidenten ten aanzien van de verschillende wijkteams kan deze methode wederom worden gebruikt. In deze analyse worden 7 weekdagen vergeleken wat bij een Monte Carlo benadering 21 simulaties vereist en 21 verschillende plots oplevert. In deze analyse wordt door dit grote aantal verschillende plots alleen een analyse gemaakt op de werkdagen ten aanzien van de weekenddagen. De output hiervan is weergegeven in figuur 5.26.

Deze plot geeft geen eenduidig beeld qua ruimtelijke verdeling. Het Centrum en de noordelijke delen van het district Oost kennen een verhoogd aantal incidenten in het weekend, al zijn er aan de randen enkele wijken waar geen significant verschil waarneembaar is. Alle wijken aan de rand van Amsterdam kennen een groter relatief deel aan straatroven op de weekdagen. De uitzondering hiervan in het deel Amstelveen in het zuiderlijk district.



Figuur 5.26: Monte Carlo: Weekdag vs wijk

Conclusie - Straatroven vinden in verschillende wijken plaats op basis van de betreffende weekdag. Het grootste verschil kan gevonden worden tussen de werkdagen en weekenddagen. Tussen de werkdagen maandag t/m vrijdag onderling en tussen de weekenddagen zaterdag en zondag onderling lijken geen grote verschillen te zijn.

5.6 Conclusie

Woninginbraken en straatroven zijn onderhevig aan verschillende geografische verdelingen wanneer de tweewekelijkse periode wordt onderdeeld in de tijdsintervallen *weekdag, dagdeel* en *diensttijd*. Dit is gebaseerd op de geografische spreiding van incidenten over wijken en districten. De volgende resultaten zijn gevonden:

Woninginbraken

1. Woninginbraken vinden plaats in verschillende districten wanneer onderscheid wordt gemaakt in de dagdelen nacht, dag, avond. De meest afwijkende verdeling van incidenten wordt waargenomen in district Centrum, waar het hoogtepunt 's nachts is, terwijl alle andere districten 's nachts de minste inbraken plaatsvinden. Uit de analyse op basis van wijken komen dezelfde resultaten al blijkt de Centrumwijk IJ-tunnel niet mee te doen in het afwijkende gedrag van het district.

2. Woninginbraken vinden in verschillende districten plaats op basis van de betreffende weekdag. Het grootste verschil kan gevonden worden tussen de werkdagen en weekenddagen. Tussen de werkdagen

maandag t/m vrijdag onderling en tussen de weekenddagen zaterdag en zondag onderling lijken geen grote verschillen te zijn, al lijkt de vrijdag zich meer afwijkend te gedragen ten aanzien van de overige werkdagen.

3. Woninginbraken vinden in verschillende districten plaats op basis van de betreffende diensttijd. Vooral het district Centrum in combinatie met de diensttijden die in de nacht vallen kennen een extreem afwijkende verdeling. De verdeling van incidenten over de diensttijden lijken zich daarnaast te gedragen in clusters van dagdelen en de week- en weekenddagen.

Straatroven

1. Straatroven vinden over de verschillende dagdelen nacht, dag, avond plaats in verschillende districten/wijken. De meest afwijkende verdeling van straatroven wordt waargenomen in district Centrum, waar het hoogtepunt 's nachts is, terwijl alle andere districten 's nachts de minste straatroven plaatsvinden. Het afwijkende gedrag van district Centrum lijkt zich niet te verhouden tot de wijken Centrumwijken IJ-tunnel en Raampoort, maar zijn de wijken Konniginneweg en Pijp in district Zuid hier wel onderhevig aan.

2. Straatroven vinden in verschillende districten/wijken plaats op basis van de betreffende weekdag. Het grootste verschil kan gevonden worden tussen de werkdagen en weekenddagen. Tussen de werkdagen maandag t/m vrijdag onderling en tussen de weekenddagen zaterdag en zondag onderling lijken geen grote verschillen te zijn.

Bij zowel straatroven als woninginbraken vindt 's nachts een enorme toename plaats in het district Centrum. Bij woninginbraken lijkt zich dit te verhouden tot de wijken Nieuwezijds Voorburgwal, Lijnbaansgracht, Prinsengracht en Raampoort. Voor straatroven tot de wijken Prinsengracht, Beursstraat, Nieuwezijds Voorburgwal, IJ-tunnel, Konniginneweg (Zuid) en de Pijp (Zuid).

Hoofdstuk 6

Voorspellen van woninginbraken op tijdsintervalniveau I

CAS voorspelt momenteel voor iedere peilperiode van twee weken de kans op een woninginbraak in de regio Amsterdam. Het huidige CAS model kent een near hits performance van $0,3632$ ($\sigma = 0,0496$) gebaseerd op de peilperioden 177 t/m 197 over tweewekelijkse peilperioden. Deze performance is echter niet uniform verdeeld over alle onderliggende tijdsintervallen. In hoofdstuk 4 werd aangetoond dat er tijdsintervallen bestaan met een significant afwijkende performance. Deze significant afwijkende performance duidt erop dat incidenten in verschillende tijdsintervallen op andere plekken gebeuren en daardoor een verschillende performance kennen. Hoofdstuk 5 gaat door op die aaname en laat zien dat er inderdaad significante geografische verschillen waarneembaar zijn tussen de verschillende tijdsintervallen. Met deze kennis is een logische volgende stap: kunnen incidenten (beter) worden voorspeld wanneer er alleen naar de incidenten in een specifiek tijdsinterval wordt gekeken? Het antwoord op deze vraag is uitgesmeerd over twee hoofdstukken: in dit hoofdstuk worden incidenten voorspeld door een model identiek aan CAS, in het volgende hoofdstuk worden enkele andere modellen gebruikt ter vergelijking. De centrale vraag binnen dit hoofdstuk is:

In hoeverre kan men met gebruik van de methodiek van CAS de kans op een incident voor iedere gridlocatie m.b.t. een specifiek tijdsinterval worden voorspeld?

In dit hoofdstuk worden modellen ontwikkeld voor de verschillende tijdsintervallen waarover vervolgens de performance wordt berekend. Voor deze modellen is gebruik gemaakt van een implementatie identiek aan CAS, waarbij alleen de incidenten worden meegenomen die bij het specfieke tijdsinterval horen waar de voorspelling zich op richt. In paragraaf 6.1 wordt deze methode toegelicht gevolgd door de beschrijving van het logistische regressiemodel in paragraaf 6.2. De paragrafen 6.3, 6.4 en 6.5 geven de resultaten van de voorspellingen met betrekking tot respectievelijk de vooraf vastgestelde tijdsintervallen diensttijden, weekdagen en dagdelen. Paragraaf 6.6 gaat verder met het onderscheid in week- en weekenddagen omdat in hoofdstuk 5 is aangetoond dat deze tijdsintervallen een afwijkende geografisch verdeling van incidenten kennen. In hoofdstuk 5 worden ook sterke geografische verschillen tussen dagdelen gevonden, waar paragraaf 6.6 op aansluit met een combinatie van week-, weekenddagen en dagdelen. Paragraaf 6.8 sluit

af met een verdeling waarbij gekeken is naar de overeenkomsten van de voorspellingen op basis van de diensttijden uit paragraaf 6.4. Paragraaf 6.9 sluit af met de conclusie.

6.1 Methode

In het huidige CAS model wordt voor een peilperiode van twee weken de kansen op een woninginbraak voor iedere locatie voorspeld. In dit hoofdstuk wordt specifiek gekeken naar het voorspellen van incidenten over kleinere tijdsintervallen door gebruik te maken van een model identiek aan CAS. Centraal staat in hoeverre de performances van het huidige CAS model kunnen worden geëvenaard/verbeterd wanneer de peilperiode wordt gesplitst in meerdere kleinere tijdsintervallen. De tweewekelijkse periode wordt daarbij opgeknipt in meerdere tijdsintervallen waar afzonderlijke voorspellingen per locatie de kans op een incident voor het specifieke tijdsinterval voorspellen. Al deze modellen worden gegenereerd door middel van een logistisch regressiemodel zoals beschreven in 6.1. De voorspellingen samen omvatten zo één tweewekelijkse peilperiode, waardoor de performance over de gehele peilperiode kan worden gemeten door alle voorspellingen voor kleinere tijdsintervallen samen te nemen. Hierbij wordt het aantal hits, near hits en incidenten van de verschillende weekdagen gesommeerd om een performance over de gehele periode te kunnen berekenen.

Gepaarde t-test De gepaarde t-test wordt gebruikt om te toetsen de performances van twee verschillende modellen gelijk verdeeld zijn. De t-test is een parametrisch toets waarbij wordt getoetst of het gemiddelde tussen twee normaal verdeelde populaties gelijk aan elkaar zijn. In dit hoofdstuk wordt veelvuldig gebruikt gemaakt van de gepaarde t-test waarbij ieder meetpunt in beide populaties voorkomt. In dat geval test de gepaarde t-test of het verschil tussen alle gepaarde punten gelijk is aan 0, aangenomen dat de verschillen normaal verdeeld zijn. In dit onderzoek worden daarbij de verschillen in performance gemeten voor iedere peilperiode en wordt er getoetst of deze verschillen kunnen toebehoren tot een normale verdeling met een μ van 0.

Overeenkomstpercentage Met het overeenkomstpercentage wordt een maatstaf gegeven om twee gegenereerde voorspellingen te vergelijken op gelijkenis. Het overeenkomstpercentage is het aantal locaties die beide modellen aanwijzen als high risk locaties gerelativeerd naar het totale aantal van 282 locaties waaruit de high risk area uit bestaat.

6.2 Model omschrijving

CAS werkt op basis van een logistische regressie model. Zoals aangegeven in paragraaf 1.2.1 draaide CAS op een neuraal netwerk, maar is CAS vanwege softwarebeperkingen overgegaan op een logistisch regressie model tot in ieder geval december 2014. In deze paragraaf worden voorspellingen dus gegenereerd op basis van een logistisch regressie model.

Een logistische regressie kan gebruikt worden om een dichotome uitkomstvariabele te relateren aan responsvariabelen. In dit geval is het wel of niet plaatsvinden van een incident de uitkomstvariabele

en kunnen alle mogelijke responsvariabelen worden meegenomen om deze te voorspellen. Logistische regressie gaat uit van het idee dat wanneer de uitkomstvariabele wordt getransformeerd, er een lineaire regressie mogelijk is. Lineaire regressie kan worden toegepast wanneer de uitkomstvariabele een continue normale verdeling kent en dus zal de transformatie van de uitkomstvariabele daartoe moeten leiden. Het wel of niet plaatsvinden van het incident wordt daarom niet als een dichotome variabele gemodeleerd, maar als de kans op deze uitkomst. Omdat de kansen echter alleen tussen 0 en 1 een betekenis kennen, maakt de logistische regressie gebruik van de relatieve kans: de kansverhoudingen, ook wel odds genoemd. Formule 6.1 geeft de odds weer wanneer p de kans is op de eerste uitkomst en $1 - p$ de kans is op de tweede uitkomst.

$$odds : \frac{p}{1-p} \tag{6.1}$$

De odds kennen een verdeling van 0 tot oneindig maar zijn niet normaal verdeeld. Om deze odds te transformeren tot een normale verdeling worden ze getransforeert met behulp van het natuurlijk logaritme. Het natuurlijk logaritme van de odds kent een continue en normale verdeling. Het logistische regressiemodel met k variabelen wordt geformuleerd in 6.2, met $X_i(i = 1, 2, ..., k)$ de responsvariabelen en $\beta_i(i = 1, 2, ..., k)$ de logistische regressiecoefficienten.

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k \tag{6.2}$$

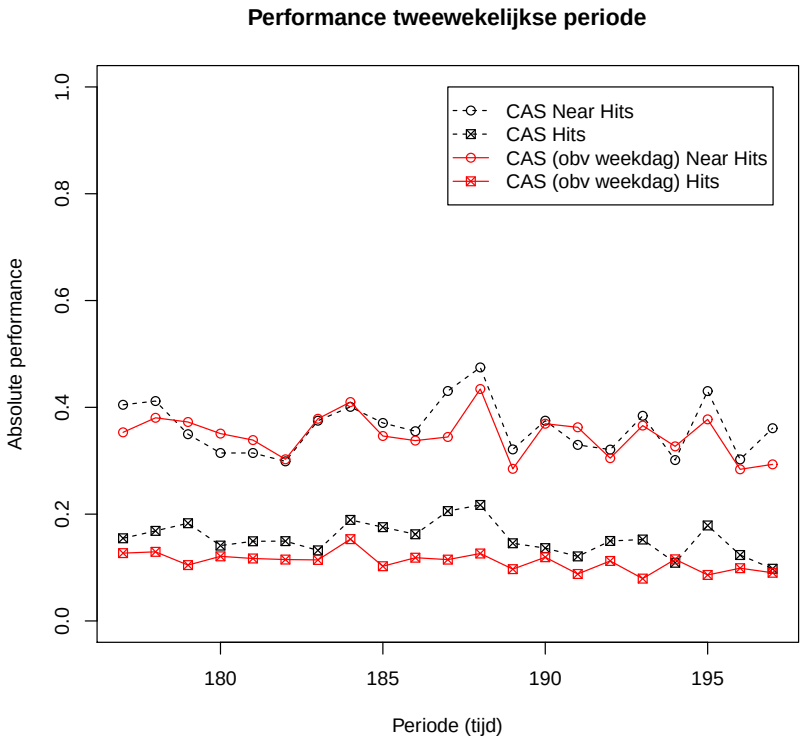
Het gebruik van een logistisch regressie is het enige verschil tussen het huidige CAS model en het gebruikte model in deze paragraaf.

In de resultaten wordt vaak gesproken over het effect van een variabele in plaats van de daadwerkelijke coefficient. Dit effect kan worden afgeleidt door e^{β_k} te berekenen, wat het geschatte effect op de log odds uitdrukt. Dit effect kan worden uitgedrukt in een positief effect, wanneer de $exp(\beta) > 1$, of een negatief effect, wanneer de $exp(\beta) < 1$. Met een positief effect wordt gedoeld op het feit dat de kans op een incident met de toename van de variabele toeneemt, terwijl een negatief effect de toename van de variabele laat afnemen.

6.3 Resultaten woninginbraken per weekdag

De tweeweekse peilperiode kan worden opgeknipt in 7 afzonderlijke voorspellingen die voor iedere weekdag de kans op een incident per locatie voorspellen. Figuur 6.1 geeft de absolute performance weer van dit samengestelde model en ter vergelijking ook de performances van het huidige CAS model.

De gemiddelde hits performance van CAS is $15,44\%$ ($\sigma = 3,04\%$) en het samengestelde model op basis van weekdag $11,09\%$ ($\sigma = 1,73\%$). De gemiddelde near hits performance van CAS is $36,32\%$ ($\sigma = 4,96\%$) en het samengestelde model op basis van weekdag $34,85\%$ ($\sigma = 3,96\%$). Het huidige CAS model kent in $95,24\%$ van de perioden een hogere hits performance dan het samengestelde model op basis van weekdag: 20 van de 21 peilperioden. In $61,90\%$ kent het huidige CAS model ook een hogere near hits performance: 13 van de 21 peilperioden. Dit resultaat geeft indicatie dat het huidige CAS model



Figuur 6.1: Performance van het huidige CAS model en het samengestelde model op basis van losse voorspellingen voor de 7 weekdagen.

beter incidenten kan voorspellen dan het samengestelde model op basis van weekdag. Met de gepaarde t-test kan deze uitspraak worden getoetst. Dit gebeurt aan de hand van de volgende hypotheses:

H_0: Het verschil in performance van de twee modellen is gelijk aan 0.
H_1: Het verschil in performance van de twee modellen *niet* gelijk aan 0.

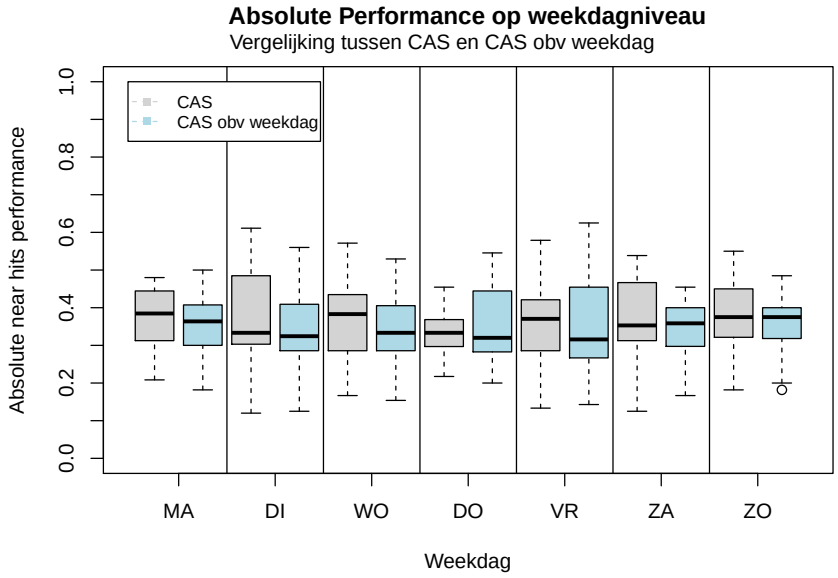
Op basis van de gepaarde t-test wordt H_0 verworpen voor de absolute hits performance (T =-6, 90; df = 20; p-waarde $< 1,05e$-6; $\alpha = 0,05$) en wordt H_0 niet verworpen voor de absolute near hits performance (T =-1, 99; df = 20; p-waarde $< 0,061$; $\alpha = 0,05$). De interpretatie van het aantal near hits hangt echter van het aantal toegewezen near hits locaties. Hierover kan een gewogen gemiddelde worden berekend waarbij het gemiddelde aantal near hits locaties wordt gewogen op basis van het aantal incidenten wat onderhevig was aan het aantal near hits locaties. Het samengestelde model heeft gemiddeld 1.160, 59 near hits locaties (per target) terwijl het huidige CAS model gemiddeld op 1.108 locaties zit.

Dit resultaat betekent dat het aannemelijk is dat de absolute hits performance van de huidige CAS methodiek significant hoger is dan de absolute hits performance van het samengestelde model. Op basis van de absolute near hits performance kan geen verschil in performance voor beide methoden worden gevonden omdat er geen verschil is, of er onvoldoende bewijs is om aan te nemen dat de performances verschillen. Het samengestelde model kent een opmerkelijk hoger aantal near hits dan het huidige model wat mogelijk kan worden verklaard door het een hoger aantal near hits locaties (ruim 52 locaties meer). Het kan ook zijn dat de weekdagvoorspellingen vaak dichtbij (near hit) zitten maar nog niet voldoende de daadwerkelijke incidenten weten te raken (hit).

De relatieve hits performance van het huidige CAS model gespecificeerd naar weekdag is $0,81$ ($\sigma = 0,005$). Bij de voorspellingen gespecificeerd op weekdag ligt dit gemiddelde op $0,77$ ($\sigma = 0,013$). Om 7 werkdagen te voorspellen voor 21 peilperioden worden 147 modellen gegenereerd. In slechts 8 gevallen presteert de voorspelling specifiek op de weekdag beter: 5,44%. Het huidige CAS model over tijdsperioden van twee weken voorspelt beter dan een model waarbij voor iedere weekdag een afzonderlijke voorspelling wordt gemaakt. Dit geldt ook voor alle weekdagen op zichzelf.

Conclusie - Het huidige CAS model over tijdsperioden van twee weken voorspelt beter dan het samengestelde model waarbij voor iedere weekdag een afzonderlijke voorspelling wordt gemaakt.

Figuur 6.2 geeft de absolute nearhit performances weer van de weekdagvoorspellingen op basis van een logistisch regressie model (blauw). Ter vergelijking zijn ook de performances van het huidige CAS model weergegeven gespecificeerd naar de verschillende weekdagen (grijs).



Figuur 6.2: Boxplot van de 7 weekdagvoorspellingen op basis van een logistisch regressie model in combinatie met de performances van de huidige CAS.

Het valt op dat geen enkele op weekdag toegespitste voorspelling een zichtbaar betere performances geeft dan het huidige CAS model. Op basis van de gepaarde t-test kan worden getoetst of het verschil in performance tussen de huidige CAS en de gegenereerde weekdagvoorspellingen per weekdag gelijk is. Op basis van de gepaarde t-test wordt H_0 niet verworpen voor de absolute near hits performance voor alle weekdagcombinaties ($\alpha = 0,05$). Dit betekent dat er geen verschil is in performance voor beide methoden of dat er niet voldoende bewijs is om aan te nemen dat de performances verschillen.

Logistische modellen De logistische modellen die worden gegenereed op basis van een volledig model waarbij geen enkele variabele, significant of niet, wordt verwijderd uit het model. Deze keuze is gemaakt omdat voor een dataset met 65 variabelen een methodiek waarbij variabelen worden geelimineerd te

Conclusie - Het huidige CAS model over tijdsperioden van twee weken voorspelt beter voor iedere afzonderlijke weekdag dan een model waarbij een specifieke voorspelling wordt gemaakt voor de weekdag.

tijdrovend is. Toch kunnen significante variabelen binnen dit model iets zeggen over de variabelen die bijdragen aan een verhoogde kans op een incident op de verschillende weekdagen. Hiervoor worden van elke weekdag drie modellen bekeken: peilperioden 177, 187 en 197. Voor alle onderzochte modellen geldt: significant op basis van Chi-square toets, $\alpha = 0,000$, $df = 65$, voor alle 21 modellen. Dit betekent dat ondanks dat misschien niet alle variabelen significant zijn, het model significant beter alle variabelen kan bevatten dan geen. Voor de volledige lijst van variabelen zie bijlage A. Uitleg over het logistische regressiemodel of de intepetatie hiervan is te vinden in paragraaf 7.1.

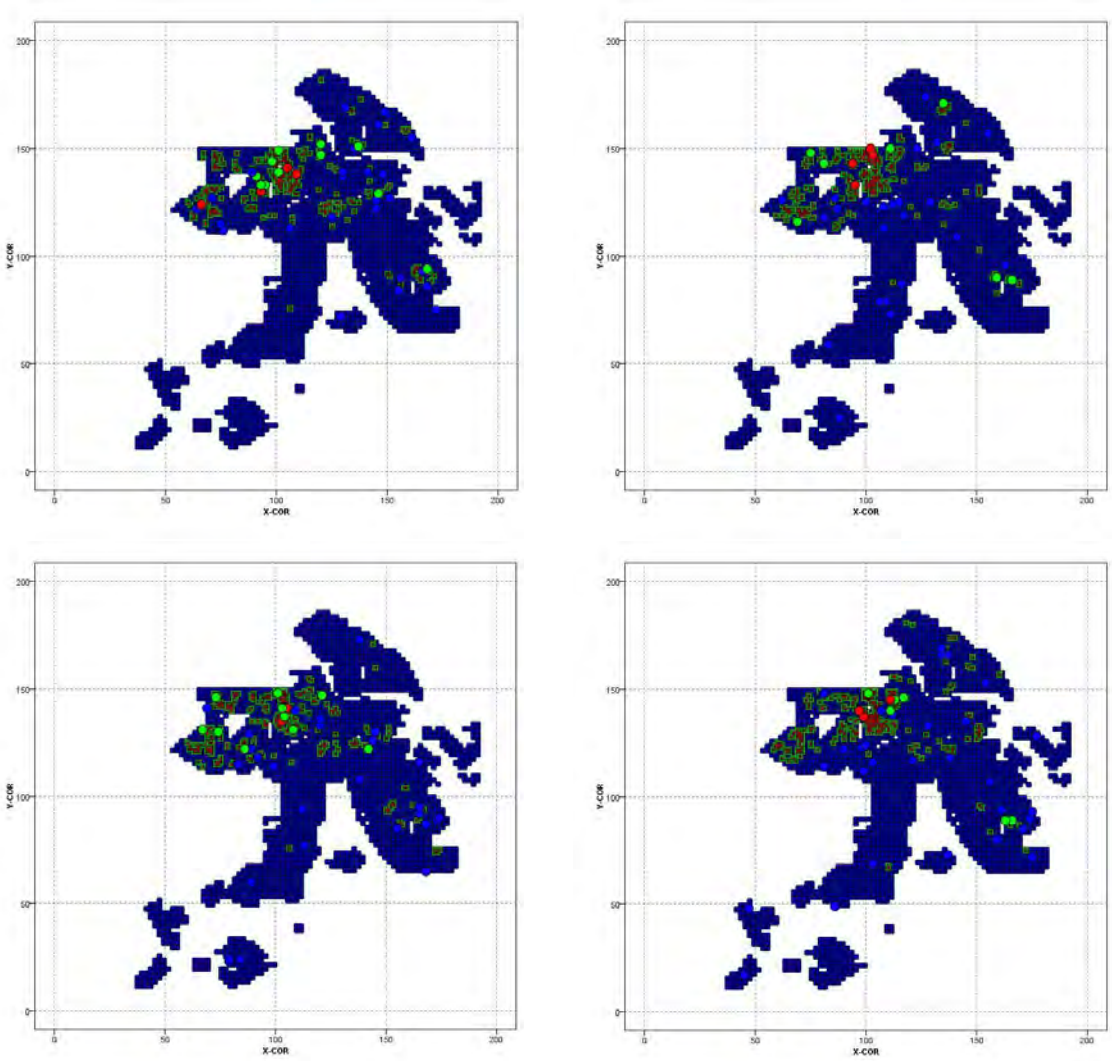
CBS schatters Opmerkelijk zijn de kleine verschillen tussen de modellen die de afzonderlijke weekdagen schatten. Een aantal variabelen en effecten komen in zowel alle weekdagmodellen als het tweewekelijkse model terug. Woningvoorraad kent een significant positief effect (1,093), de x-coordinaat een significant negatief effect (0,996), de y-coordinaat een significant positief effect (1,004), de variabele niet westerse allechtoon kent een positief effect (1,101), aantallen hooginkomen kent een significant negatief effect (0,976) en het aantal zelfstandigen ook (0,916). Met een positief effect word gedoeld op het feit dat de kans op een incident met de toename van de variabele toeneemt, terwijl een negatief effect de toename van de variabele laat afnemen.

Crimehistorie schatters De meeste modellen hebben ongeveer 3 tot 6 significante variabelen die de crimehistorie beschrijven, terwijl in de huidige CAS omgeving vrijwel alle variabelen met betrekking tot de crimehistorie significant zijn. Misschien dat door een gebrek aan daadwerkelijke incidenten in de geschiedenis relaterend aan een weekdag er geen goede significante verschillen ontdekt kunnen worden. Er bestaat daarnaast een klein verschil tussen de weekdagen, waarbij de dagen zaterdag en zondag allebei meer significante historisch variabelen kennen. Het zou daardoor kunnen zijn dat de incidenten in het weekeinde beter voorspelbaar zijn en meer voortborduren op historische patronen, al komt dat laatste niet tot uiting in paragraaf 4.2.1 waar de weekdagperformances zijn onderzocht. Vrijwel alle variabelen die de crimehistorie beschrijven kennen een positief effect.

Bedrijfsinformatieve schatters In de bedrijfsvariabelen kunnen ook significante verschillen gevonden worden. De variabele bejaardenhuis kent in de meeste dagen (m.u.v. vrijdag) een significant positief effect (1,620). De variabelen koffieshop en hotel-motel-botel komen voornamelijk in het weekeinde significant naar voren als positief effect. In paragraaf 5.2.2. bleek dat de weekenddagen afwijkend gedrag vormde ten aanzien van de andere dagen. Wellicht kan deze signifcantie daarbij een rol spelen.

6.3.1 CAS-kaarten woninginbraken per weekdag

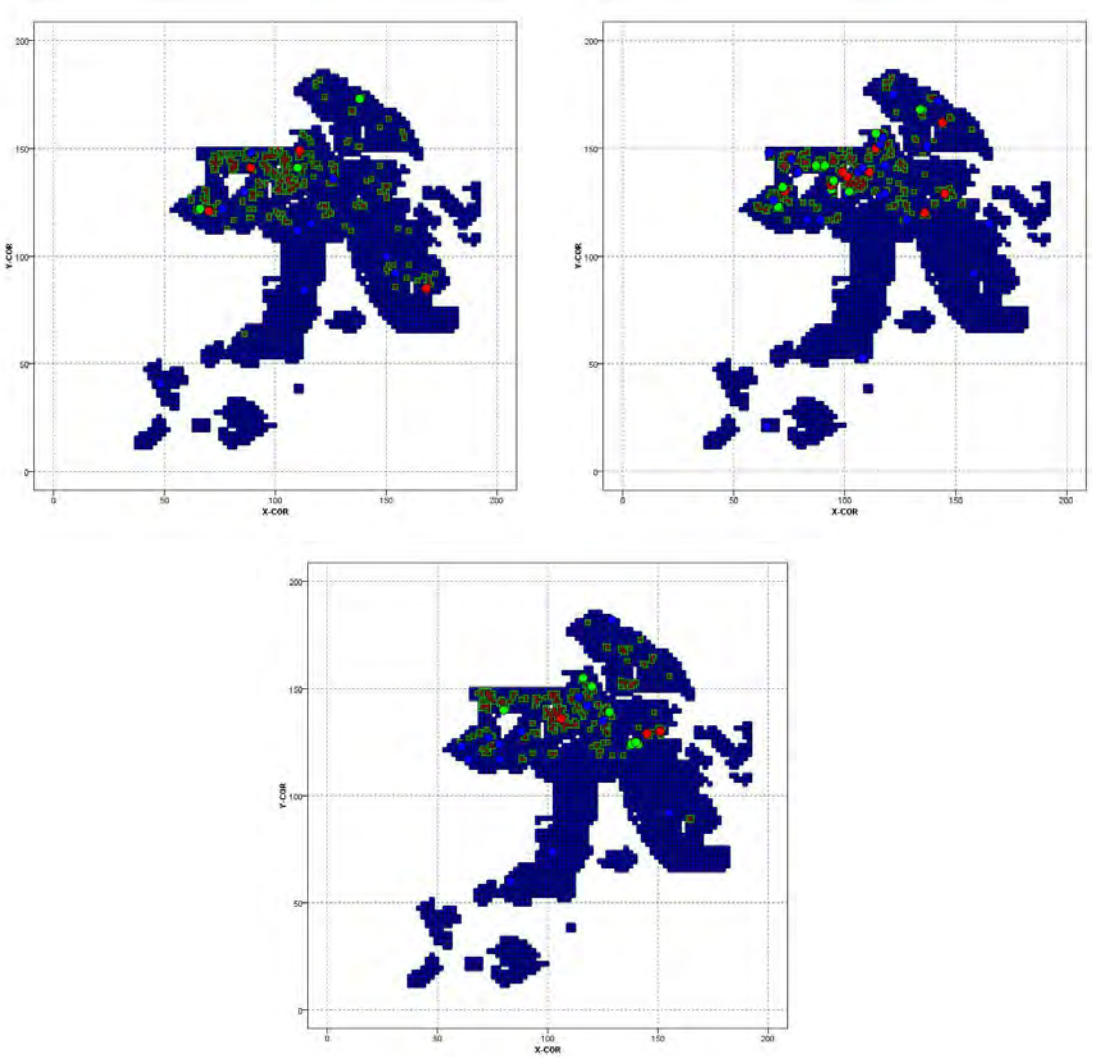
Eén van de belangrijkste vragen binnen dit onderzoek is de geografische spreiding van criminele incidenten tijdens de verschillende tijdsintervallen. Een andere methode om hier inzicht in te krijgen is het vergelijken van de geografische kaarten.



Figuur 6.3: CAS-kaarten maandag, dinsdag, woensdag en donderdag (boven naar onder, links naar rechts) obv logistische regressie voor periode 177.

De figuren 6.3 en 6.4 geven een grafische weergave van de high risk locaties (rood) en de omliggende near hits locaties (groen) van de voorspellingen op basis van weekdag voor de peilperiode 177. Alle locaties zijn weergegeven als rondje op basis van een x en y-coordinaat. De locaties waar achteraf een incident heeft plaatsgevonden in de betreffende periode zijn drie keer uitvergroot. Alle kaarten zijn output van een specifieke weekdagvoorspelling en kennen allemaal een ander patroon van incidenten waarop de voorspelling is gebaseerd. Wanneer twee kaarten dus op elkaar lijken, is dat gebaseerd op een andere verzameling van incidenten.

In paragraaf 3.4 wordt aangegeven dat 282 locaties worden gekenmerkt als high risk area. Om twee kaarten te vergelijken kan gekeken worden hoeveelheid high risk locaties die overeenkomen tussen twee kaarten: hoe meer locaties overeenkomen, hoe groter de gelijkenis. Voor peilperioden 177, 187 en



Figuur 6.4: CAS-kaarten vrijdag, zaterdag en zondag (boven naar onder, links naar rechts) obv logistische regressie voor periode 177.

Weekdag	DI	WO	DO	VR	ZA	ZO
MA	0,429	0,417	0,379	0,338	0,313	0,351
DI		0,372	0,387	0,352	0,348	0,348
WO			0,357	0,338	0,307	0,344
DO				0,304	0,333	0,339
VR					0,299	0,327
ZA						0,384

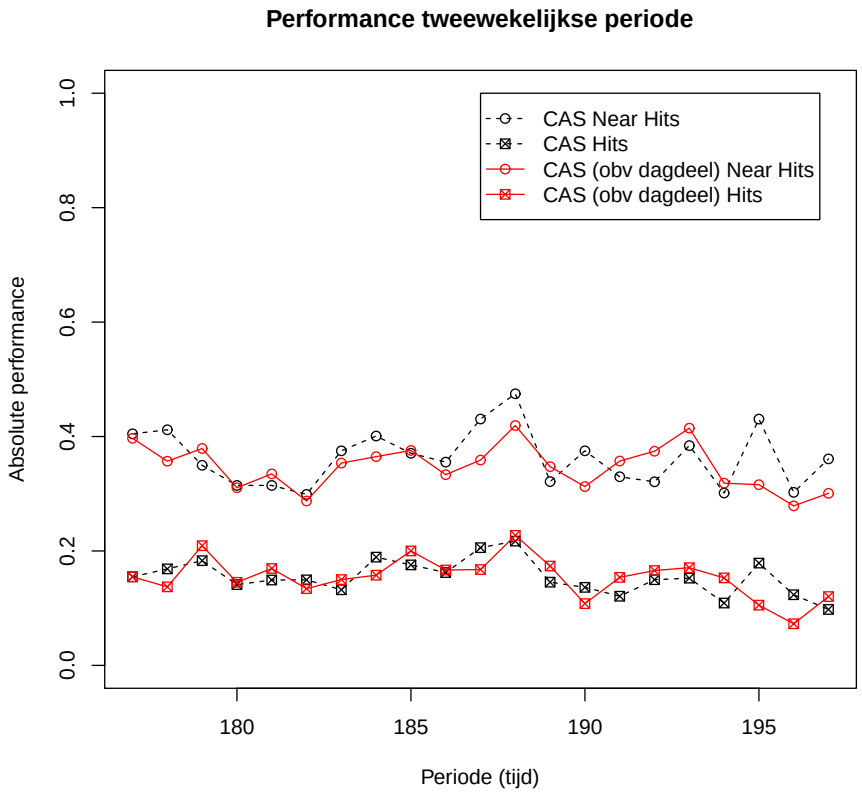
Tabel 6.1: Percentage overeengekomen high risk locaties, gemiddelde over peilperioden 177, 187 en 197.

197 is het percentage overeenkomende high risk locaties berekend. De uitkomsten hiervan zijn opgenomen in tabel 6.1. Uit tabel 6.1 blijkt dat maandag t/m donderdag relatief veel op elkaar lijken. Daarnaast kennen de dagen zaterdag en zondag ook weer een behoorlijke gelijkenis. In paragraaf 6.2.2 en 6.3.2 werden de geografische verdelingen van incidenten onderhevig aan de weekdagen onderzocht en daar werd ook opgemerkt dat de weekenddagen (zaterdag en zondag) en de weekdagen (maandag t/m donderdag) zich

afwijkend van elkaar gedroegen. Over vrijdag kon in die paragrafen moeilijk uitsluitsel worden gegeven, evenals hier blijkt dat vrijdag zich afwijkend gedraagd. Tussen de dagen maandag t/m donderdag is het gemiddelde percentage overeenkomstige high risk locaties $39,03\%$ en tussen zaterdag en zondag $38,42\%$. Het gemiddelde percentage overeenkomstige locaties tussen de weekend en weekdagen is $33,54\%$: zaterdag $32,54\%$ en zondag $34,54\%$. Tussen vrijdag en de werkdagen wordt een percentage van $33,30\%$ gemeten en tussen vrijdag en de weekenddagen $33,63\%$. Vrijdag lijkt dus een geografisch patroon op zichzelf te hebben dat mogelijk een combinatie is tussen de patronen op de werk- en weekenddagen. Het patroon op zondag verhoudt zich daarnaast meer tot de weekdagen maandag t/m donderdag dan zaterdag dat doet.

6.4 Resultaten woninginbraken per dagdeel

De tweeweekse peilperiode kan worden opgeknipt in 3 afzonderlijke voorspellingen die voor ieder dagdeel de kans op een incident per locatie voorspellen. Figuur 6.5 geeft de absolute performance weer van dit samengestelde model en ter vergelijking ook de performances van het huidige CAS model.



Figuur 6.5: Performance van de huidige CAS en een samengestelde CAS op basis van losse voorspellingen voor de 3 dagdelen op basis van een logistisch regressie model

De gemiddelde hits performance van CAS is $15,44\%$ ($\sigma = 3,04\%$) en het samengestelde model op basis van dagdeel $15,44\%$ ($\sigma = 3,53\%$). De gemiddelde near hits performance van CAS is $36,32\%$ ($\sigma = 4,96\%$) en het samengestelde model op basis van dagdeel $34,72\%$ ($\sigma = 3,91\%$). Het huidige CAS model kent in $33,33\%$ van de perioden een hogere hits performance dan het samengestelde model op basis van dagdeel: 7 van de 21 peilperioden. In één periode weten beide modellen een gelijke performance te halen en in

13 perioden presteert het samengestelde model beter. In $57,14\%$ kent het huidige CAS model wel een hogere near hits performance: 12 van de 21 peilperioden. Dit resultaat geeft geen indicatie dat één van de twee modellen beter incidenten kan voorspellen. Met de gepaarde t-test kan worden getoetst of de performances significant verschillen. Dit gebeurt aan de hand van de volgende hypotheses:

H_0: Het verschil in performance van de twee modellen is gelijk aan 0.
H_1: Het verschil in performance van de twee modellen *niet* gelijk aan 0.

Op basis van de gepaarde t-test wordt H_0 niet verworpen voor de absolute hits performance ($T = 0,0012$; $df = 20$; p-waarde= 0.999; $\alpha = 0,05$) en niet voor de near hits performance ($T = 1,72$; $df = 20$; p-waarde= $0,1$; $\alpha = 0,05$). De interpretatie van het aantal near hits hangt echter van het aantal toegewezen near hits locaties. Hierover kan een gewogen gemiddelde worden berekend waarbij het gemiddelde aantal near hits locaties wordt gewogen op basis van het aantal incidenten wat onderhevig was aan het aantal near hits locaties. Het samengestelde model heeft gemiddeld $1.058,87$ near hits locaties (per target) terwijl het huidige CAS model gemiddeld op 1.108 locaties zit.

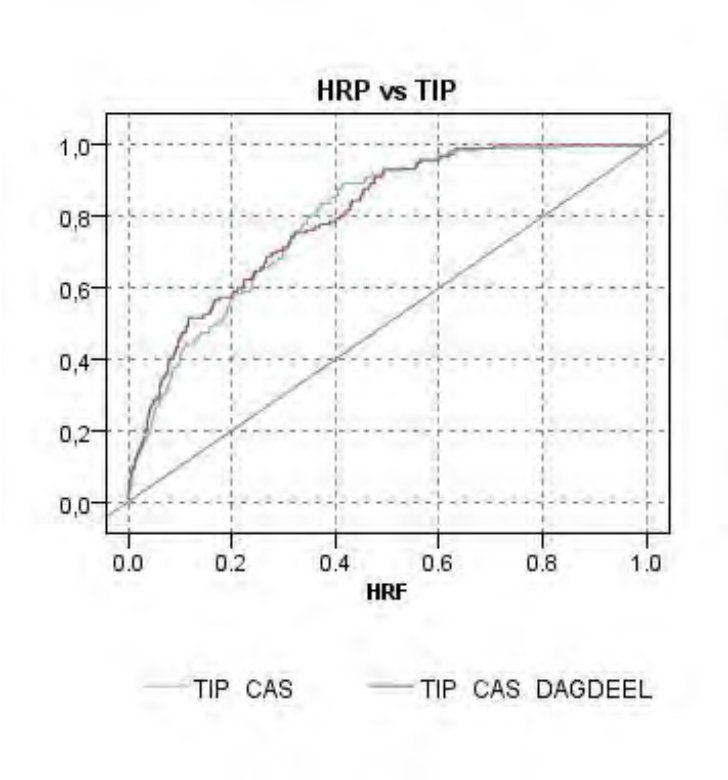
Dit resultaat betekent dat er geen verschil is tussen de hits en near hits performance van de huidige CAS model en het samengestelde model op basis van dagdeel. Daarnaast kent het samengestelde model een lager aantal near hits (ruim 50 minder) wat kan leiden tot de iets minder goede near hits performance ten aanzien van de hits performance.

De relatieve hits performance van het huidige CAS model gespecificeerd naar dagdeel is $0,81$ ($\sigma = 0,010$). Bij de voorspellingen gespecificeerd op dagdeel ligt dit gemiddelde op $0,80$ ($\sigma = 0,015$). Om 3 dagdelen te voorspellen voor 21 peilperioden worden 63 modellen gegenereerd. In slechts 15 gevallen presteert de voorspelling specifiek op de dagdeel beter: 23,81%. Op basis van de relatieve performance measure voorspelt het huidige CAS model over tijdsperioden van twee weken beter dan een model waarbij voor ieder dagdeel een afzonderlijke voorspelling wordt gemaakt, maar de verschillen zijn minimaal. Dit geldt ook voor alle dagdelen op zichzelf, maar het minst voor het dagdeel nacht waar 10 van de 21 modellen beter wordt voorspelt door het samengestelde model. Voor de dagdelen avond en dag kan H_0 verworpen voor de absolute hits performance ($\alpha = 0,05$) en is de relatieve hits performance van het huidige model significant hoger. Omdat de absolute hits performance measure niet onderdoet aan het huidige CAS model kan het verschil in relatieve performance ook worden veroorzaakt in de staart van de curve (zie paragraaf 3.4.1) in plaats van het begin zoals bijvoorbeeld in de curve van figuur 6.6.

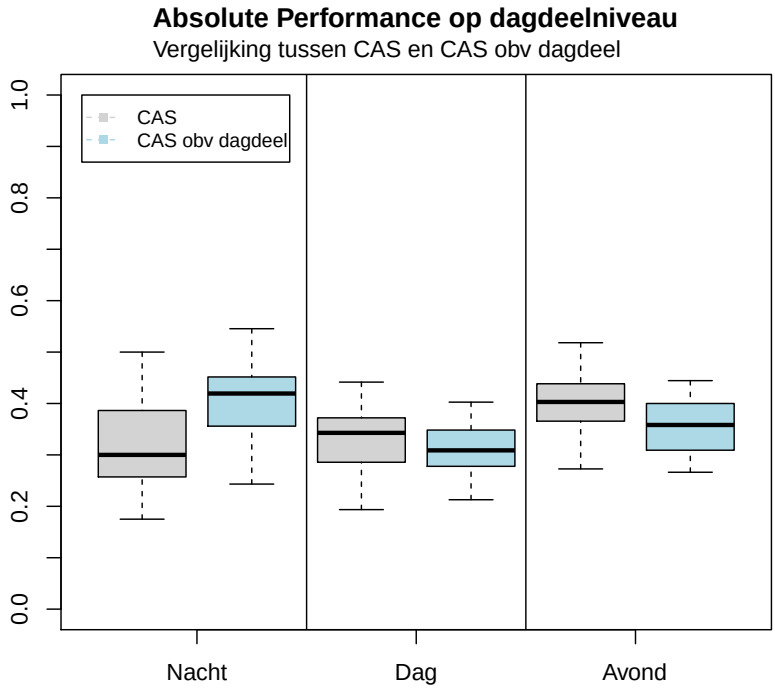
Conclusie - Er kan geen verschil worden gevonden tussen de performance van het samengestelde model op basis van dagdelen en het huidige CAS model.

Figuur 6.7 geeft de absolute nearhit performances weer van de dagdeelvoorspellingen op basis van een logistisch regressie model (blauw). Ter vergelijking zijn ook de performances van het huidige CAS model gespecificeerd naar de verschillende dagdelen weergegeven (grijs).

Het valt op dat de incidenten die 's nachts plaatsvinden beter zijn voorspeld met de samengestelde versie van CAS op basis van het dagdeel. Met de gepaarde t-test kan worden getoetst of het verschil in



Figuur 6.6: Plot relatieve performance measure peilperiode 178 dagdeel avond: huidig model heeft een performance van 0,825 en het samengestelde model op basis van dagdeel 0,799



Figuur 6.7: Performance CAS en een samengestelde CAS op basis van losse voorspellingen voor 21 tijdsvensters op tweewekelijkse basis

performance tussen de huidige CAS en de gegenereerde weekdagvoorspellingen gelijk is. Op basis van de gepaarde t-test wordt H_0 verworpen voor de absolute near hits performance voor het dagdeel nacht ($T = -3,05$; $df = 20$; p-waarde$= 0,0063$; $\alpha = 0,05$). Dit betekent dat de incidenten die plaatsvinden in de nacht beter te voorspellen zijn door middel van een voorspelling specifiek gekoppeld aan de nacht dan wanneer deze zijn opgenomen in de tweewekelijkse voorspellingen. Op basis van de gepaarde t-test wordt H_0 verworpen voor de absolute near hits performance voor het dagdeel avond ($T = 3,16$; $df = 20$; p-waarde$= 0,0049$; $\alpha = 0,05$). Dit betekent dat de incidenten die plaatsvinden in de avond beter te voorspellen zijn met het huidige CAS model dan de samengestelde versie.

Conclusie - Het dagdeel *nacht* wordt beter voorspelt door het samengestelde model op basis van dagdeel. Het dagdeel *avond* wordt beter voorspelt op basis van het huidige CAS model. Met betrekking tot het dagdeel *dag* is het lastiger een verschil te vinden, maar lijkt het huidige CAS model iets beter in het voorspellen van incidenten.

Logistische modellen De logistische modellen die worden gegenereed op basis van een volledig model waarbij geen enkele variabele, significant of niet, wordt verwijderd uit het model. Deze keuze is gemaakt omdat voor een dataset met 65 variabelen een methodiek waarbij variabelen worden geelimineerd te tijdrovend is. Toch kunnen significante variabelen binnen dit model iets zeggen over de variabelen die bijdragen aan een verhoogde kans op een incident in de verschillende dagdelen. Hiervoor worden van elk dagdeel drie modellen bekeken: peilperioden 177, 187 en 197. Voor alle onderzochte modellen geldt: significant op basis van Chi^2 toets, $\alpha = 0,05$; $df = 65$, voor alle 21 modellen. Dit betekent dat ondanks dat misschien niet alle variabelen significant zijn, het model significant beter alle variabelen kan bevatten dan geen. De gegeven effecten zijn altijd gemiddelden over de 3 modellen tenzij anders staat aangegeven.

CBS Schatters Er zijn kleine verschillen in significantie van de CBS schatters onder de dagdeelmodellen. Een aantal variabelen én effecten komen in alle dagdeelmodellen terug. De x-coordinaat kent een significant negatief effect (0,997), de y-coordinaat kent een significant positief effect (1,003), het percentage eenpersoonshuishoudens is ook significant onder 8 van de 9 modellen met een positief effect (1,048) maar dit effect is het minst in het geval in het dagdeel dag (1,033) en het hoogst in dagdeel nacht (1,067). De woningvoorraad kent een significant positief effect (1,087) al is dit effect juist overdag het hoogst (1,102) en ‘s nachts het laagst (1,073). De variabele die aantallen hooginkomen beschrijft kent een significant negatief effect (0,978) en de aantallen zelfstandigen kennen ook een negatief effect (0,919). De variabele aantallen laaginkomen is alleen significant met een negatief effect (0,977) in alle nachtmodellen evenals de variabale aantallen inkomensontvangers die eveneens een significant negatief effect kennen (0,980). Het aantal uitkeringsontvangers is juist ‘s avonds significant negatief (0,960). Ook zijn er wat effecten die in enkele modellen als positief werken en in andere als negatief effect (ongeacht significantie). De variabele huishoudgrootte kent een negatief (niet significant) effect voor nacht en dag, maar juist ‘s avonds kent deze variabele een significant positief effect. Aantal inwoners kent een positief effect ‘s nachts en overdag, maar werkt ‘s avonds juist negatief (beide niet significant). Respectievelijk

is het grootste signifciante positieve effect 1,158, toebehorend aan de avondmodellen van de variabele 'aantallen niet westerse allechtonen'. 's Nachts is deze variabele juist negatief (0,993) en overdag kent deze variabele ook een positief effect (1,071), maar deze is niet significant.

Crimehistorie schatters Gemiddeld zijn alle effecten van de crimehistorievariabelen positief. Dit is ook te verwachten, want een hoge crimehistorie zorgt vermoedelijk niet voor minder inbraken, al kunnen verhoogde politiepatroulilles hier parten in spelen. Het hoogste positieve effect is het aantal inbraken in de betreffende gridlocatie in de afgelopen twee weken (1,466). Na mate de historische variabelen een tijdspad beschrijft verder in het verleden, des te lager het positieve effect. Over het algemeen kennen de modellen die de nacht en dag beschrijven 6 significante variabelen, maar bij de modellen die de avond beschrijven kennen vrijwel alle historische variabelen een significant positief effect. Opmerkelijk is dat dit ook het geval is bij de tweewekelijkse voorspellingen en het dagdeel avond ook significant beter presteerde onder de twee wekelijkse voorspellingen (paragraaf 4.2.2). Dit lijkt erop dat het dagdeel nacht beter te voorspellen is op basis van historie en meer volgens een vast patroon plaatsvindt.

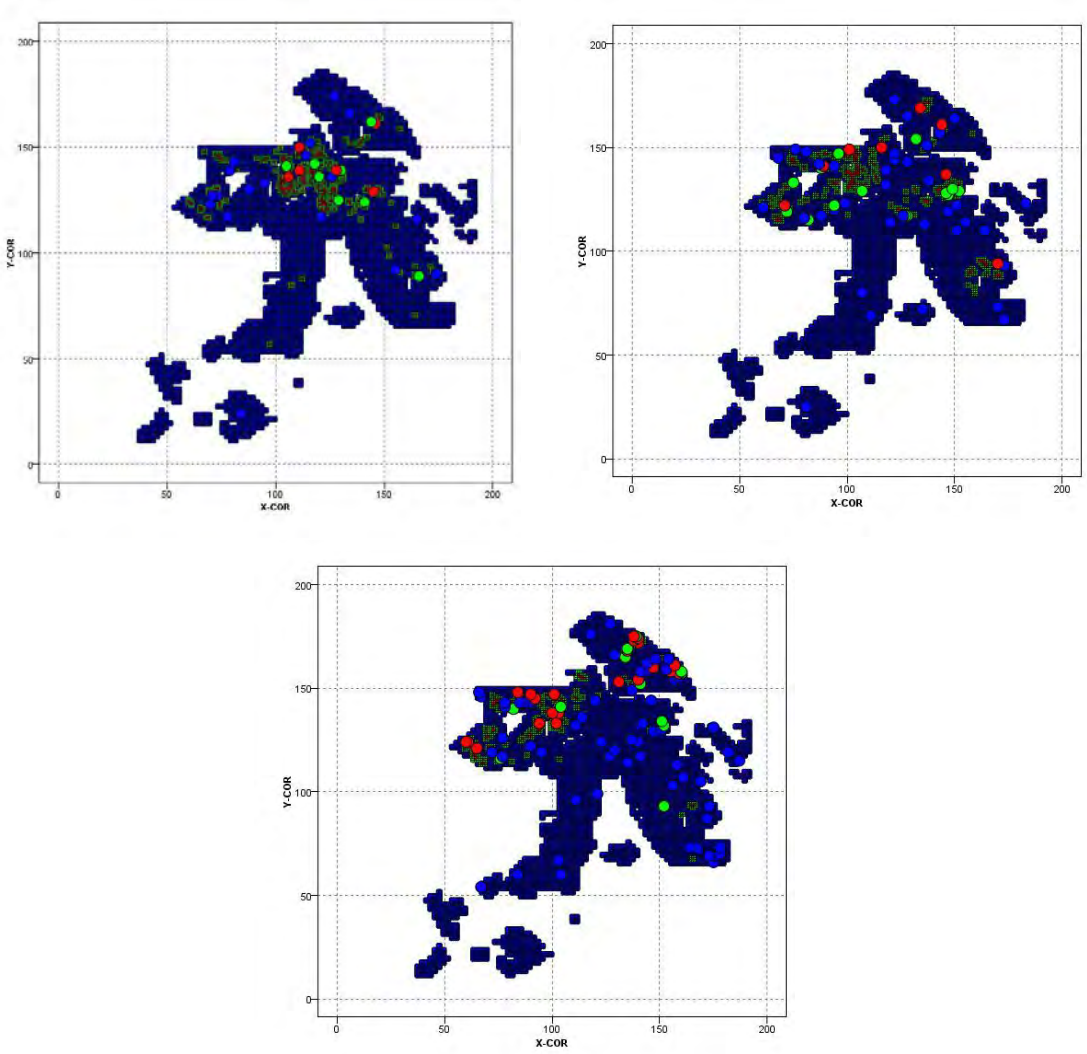
Bedrijfsinformatieve schatters Significante bedrijfsvariabelen zijn erg wisselend per model. Daarnaast zijn over het algemeen weinig variabelen significant, maar kunnen de schaarse significante variabelen wel iets zeggen over het inbraakpatroon in een specifiek dagdeel. 's Nachts is de variabele koffieshop een significant positief effect (1,135) evenals de variabele bejaardenhuis (1,854). Bejaardenhuis kent overdag ook een positief significant effect (1,385), maar 's avonds is deze variabele niet significant en soms ook negatief. In de avond zijn er relatief de meeste significante bedrijfsvariabelen ten aanzien van de andere twee dagdelen: zowel de variabelen koffieshops (+ 1,095), banken (- 0,795) en hotel/motel/botel (+ 1,115) en winkel (+ 1,020) zijn significant. Ongeacht significantie is het grootste positieve effect voor alle dagdelen de aanwezigheid van een bejaardenhuis. Het grootste negatieve effect in de nacht het benzinestation (0,665), overdag de discotheek/dancing/nachtclub (0,852) en in de avond de aanwezigheid van een bank (0,795).

Veelpleger schatters Er zijn twee variabelen die informatie geven over het aantal bekende veelplegers rondom een specifieke gridlocaties. Deze variabelen kennen allemaal een significant effect in de dagdelen avond en dag, maar dit effect is gelijk aan 1,000.

6.4.1 CAS-kaarten woninginbraken per dagdeel

Eén van de belangrijkste vragen binnen dit onderzoek is de geografische spreiding van criminele incidenten tijdens de verschillende tijdsintervallen. Een methode om hier inzicht in te krijgen, is het vergelijken van de geografische kaarten.

De afbeeldingen in figuur 6.8 geven een grafische weergave van de high risk area (rood) en de omliggende near hits area (groen) van de voorspellingen op basis van dagdeel voor de peilperiode 177. Alle locaties zijn weergegeven als rondje op basis van een x en y-coordinaat. De locaties waar achteraf een incident heeft plaatsgevonden in de betreffende periode zijn drie keer uitvergroot. Alle kaarten zijn output van een specifieke dagdeelvoorspelling en kennen allemaal een ander patroon incidenten waarop



Figuur 6.8: CAS-kaarten nacht, dag en avond (boven naar onder, links naar rechts) obv logistische regressie voor periode 177.

de voorspelling is gebaseerd. Wanneer twee kaarten dus op elkaar lijken, is dat gebaseerd op een andere verzameling van incidenten.

In paragraaf 3.4 wordt aangegeven dat 282 locaties worden gekenmerkt als high risk area. Om twee kaarten te vergelijken kan gekeken worden hoeveelheid high risk locaties die overeenkomen tussen twee kaarten: hoe meer locaties overeenkomen, hoe groter de gelijkenis. Voor peilperioden 177, 187 en 197 is het percentage overeenkomende high risk locaties berekend. De uitkomsten hiervan zijn opgenomen in tabel 6.2.

Weekdag	DAG	AVOND
NACHT	0,2104	0,1702
DAG		0,3073

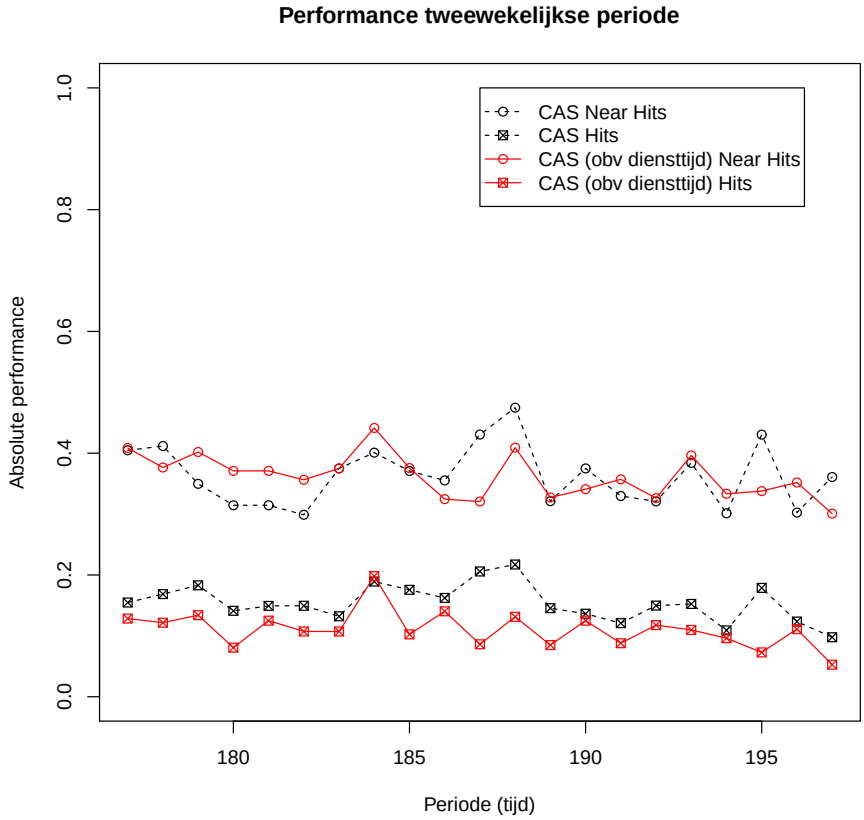
Tabel 6.2: Percentage overeengekomen high risk locaties, gemiddelde over peilperioden 177, 187 en 197.

Uit tabel 6.2 blijkt dat alle dagdelen van elkaar afwijken wanneer gekeken wordt naar het percentage gelijke high risk locaties. De percentuele overeenkomsten zijn ook veel lager dan de overeenkomsten

tussen de verschillende weekdagen. Paragraaf 5.2.1 en 5.3.1 onderzochten de geografische verschillen op basis van dagdeel en komen ook met een significant aantoonbaar verschil. Het dagdeel dat het meeste lijkt af te wijken is het dagdeel nacht, maar ook dagdelen dag en avond hebben slechts een overeenkomst van 30,73%.

6.5 Resultaten woninginbraken per diensttijd

De tweeweekse peilperiode kan worden opgeknipt in 21 afzonderlijke voorspellingen die voor iedere diensttijd de kans op een incident per locatie voorspellen. Figuur Figuur 6.9 geeft de absolute performance weer van dit samengestelde model en ter vergelijking ook de performances van het huidige CAS model.



Figuur 6.9: Performance CAS en een samengestelde CAS op basis van losse voorspellingen voor 21 tijdsvensters op tweewekelijkse basis.

De gemiddelde hits performance van CAS is $15,44\%$ ($\sigma = 3,04\%$) en het samengestelde model op basis van diensttijd $11,05\%$ ($\sigma = 3,02\%$). De gemiddelde near hits performance van CAS is $36,32\%$ ($\sigma = 4,96\%$) en het samengestelde model op basis van diensttijd $36,21\%$ ($\sigma = 3,57\%$). Het huidige CAS model kent in $95,24\%$ van de perioden een hogere hits performance dan het samengestelde model op basis van diensttijden: 20 van de 21 peilperioden. In $38,10\%$ kent het huidige CAS model ook een hogere near hits performance: 8 van de 21 peilperioden. In één periode weten beide modellen een gelijke near hits performance te halen en in 12 perioden presteert het samengestelde model beter. Met de gepaarde t-test kan een significant verschil in performance worden getoetst. Dit gebeurt aan de hand van de volgende

hypotheses:

H_0: Het verschil in performance van de twee modellen is gelijk aan 0.
H_1: Het verschil in performance van de twee modellen *niet* gelijk aan 0.

Op basis van de gepaarde t-test wordt H_0 verworpen voor de absolute hits performance ($T = 6,2962$; $df = 20$; p-waarde$= 3,802E-6$; $\alpha = 0,05$) en wordt H_0 niet verworpen voor de absolute near hits performance ($T = 0,1054$; $df = 20$; p-waarde$= 0,9171$; $\alpha = 0,05$). De interpretatie van het aantal near hits hangt echter van het aantal toegewezen near hits locaties. Hierover kan een gewogen gemiddelde worden berekend waarbij het gemiddelde aantal near hits locaties wordt gewogen op basis van het aantal incidenten wat onderhevig was aan het aantal near hits locaties. Het samengestelde model heeft gemiddeld 1.213,94 near hits locaties (per target) terwijl het huidige CAS model gemiddeld op 1.108 locaties zit.

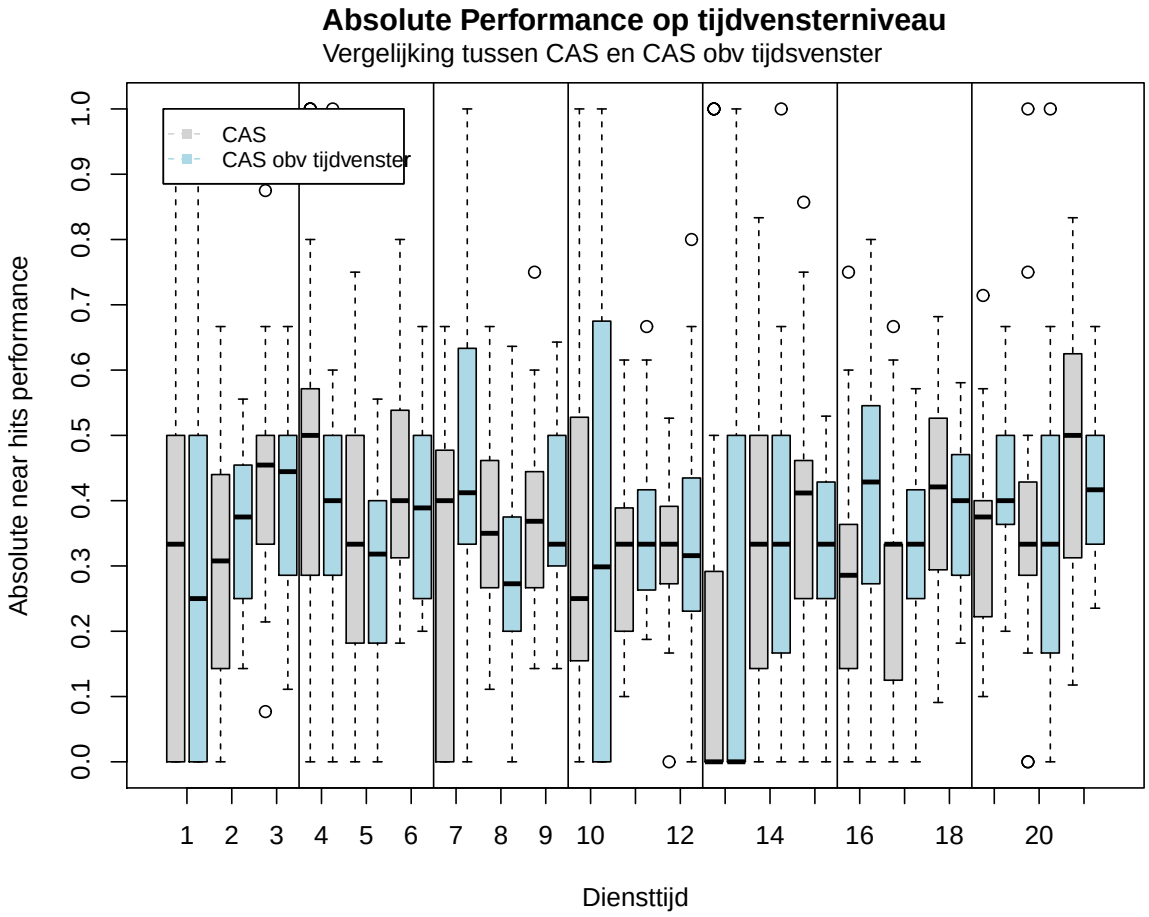
Dit resultaat betekent dat het aannemelijk is dat de absolute hits performance van de huidige CAS methodiek hoger is dan die van de samengestelde performance op basis van 21 aparte diensttijd-voorspellingen. Op basis van de absolute near hits performance kan geen verschil in performance voor beide modellen worden gevonden maar door het hoge aantal near hits locaties (ruim 100 meer) in het samengestelde model, is de hits performance measure bepalender.

De relatieve hits performance van het huidige CAS model gespecificeert naar diensttijd is $0,81$ ($\sigma = 0,021$). Bij de voorspellingen gespecificeerd op diensttijd ligt het gemiddelde op $0,76$ ($\sigma = 0,034$). Om 21 diensttijden te voorspellen voor 21 peilperioden worden 441 modellen gegenereerd. In 101 gevallen presteert de voorspelling specifiek op een dienstttijd beter: 22,90%. Op basis van de relatieve performance measure voorspelt het huidige CAS model over tijdsperioden van twee weken beter dan een model waarbij voor iedere diensttijd een afzonderlijke voorspelling wordt gemaakt. Voor bijna alle diensttijden op zichzelf kan H_0 worden verworpen op basis van de relatieve hits performance ($\alpha = 0,05$) behalve voor de diensttijden 7, 10, 13, 14, 16 en 20.

Conclusie - Het huidige CAS model over de tijdsperioden van twee weken voorspelt beter dan het samengestelde model waarbij voor iedere diensttijd een afzonderlijke voorspelling wordt gemaakt.

Figuur 6.10 geeft de absolute near hits performances weer van de samengestelde diensttijdvoorspellingen op basis van een logistisch regressie model (blauw). Ter vergelijking zijn ook de performances van het huidige CAS model weergegeven gespecificeerd naar de verschillende diensttijden (grijs).

De diensttijden 7 en 19 lijken beter te worden voorspelt door het model op basis van diensttijd. Diensttijd 8 lijkt door het gespecificeerde model juist slechter te voorspellen. Op basis van de gepaarde t-test kan worden getoetst of het verschil in performance tussen de huidige CAS en de gegenereerde weekdagvoorspellingen gelijk is. Op basis van de gepaarde t-test wordt H_0 niet verworpen voor alle combinaties van gelijke diensttijden tussen het huidige CAS model en het samengestelde model op basis van diensttijd. Er is dus onvoldoende bewijs of geen verschil in performances tussen twee gelijke diensttijden wanneer voorspelt door de huidige CAS of een specifiek model.



Figuur 6.10: Performance CAS en een samengestelde CAS op basis van losse voorspellingen voor 21 tijdsvensters op tweewekelijkse basis.

In de paragrafen 6.1.2 en 6.1.4 worden schatters (significante variabelen) van de modellen vergeleken op basis van het tijdsinterval waarvoor de voorspelling is gegenereerd. Voor de 21 diensttijdmodellen is dit niet gedaan. In de voorgaande analyses bleek al dat het vinden van verschillen in schatters lastig is en het aantal significante variabelen achteruit loopt wanneer er minder incidenten te voorspellen zijn. Er is voor gekozen deze analyse niet op diensttijdniveau uit te voeren.

Conclusie - Het huidige CAS model over de tijdsperioden van twee weken voorspelt beter incidenten voor alle diensttijden met uitzondering van de diensttijden 7, 10, 13, 14, 16 en 20 waar geen verschil te vinden is tussen beide modellen.

6.5.1 CAS-kaarten woninginbraken per diensttijd

Er is gekeken naar de verscheidendheid van de verschillende CAS-kaarten voor de uiteenlopende voorspelling per diensttijd. Voor peilperioden 177, 187 en 197 is het percentage overeenkomende high risk locaties berekend en weergegeven in tabel 6.4. Gemiddeld komen twee willekeurige kaarten voor $46,20\%$ overeen

met elkaar. Dit is vergeleken met de overeenkomstigheidspercentages van weekdag en dagdeelkaarten een hoog percentage overeenkomstige locaties. Dit ligt voor een groot deel (vermoedelijk) aan de beschikbare incidenthistorie per diensttijd. Voorspellingen gespecificeerd op diensttijden hebben 21 keer zo weinig incidenten om aan te relateren dan wanneer een tweewekelijkse voorspelling wordt gemaakt. Door dit 'gebrek' aan incidenten bestaan alle historisch variabelen voor een groot deel uit nullen, wat locaties veel minder onderscheidend maakt. Daarnaast betekent dit ook dat er minder historie bekend is om het effect van de responsvariabelen op de uitkomstvariabele te bepalen. De variabelen waarop het model zich dan (significant) het meest gaat baseren, zijn de variabelen die breed over alle diensttijden beschikbaar zijn (zoals CBS gegevens of bedrijfsinformatie) en wat uiteindelijk zorgt voor meer gelijkwaardige CAS-kaarten.

	M2	M3	D1	D2	D3	W1	W2	W3	D1	D2	D3	V1	V2	V3	Z1	Z2	V3	Z1	Z2	Z3
M1	0,43	0,45	0,59	0,48	0,46	0,52	0,43	0,49	0,6	0,44	0,46	0,6	0,52	0,47	0,5	0,52	0,43	0,5	0,57	0,49
M2		0,43	0,46	0,4	0,39	0,45	0,39	0,4	0,45	0,41	0,45	0,45	0,43	0,43	0,4	0,45	0,37	0,41	0,44	0,43
M3			0,45	0,44	0,45	0,43	0,39	0,43	0,45	0,41	0,47	0,45	0,42	0,46	0,39	0,41	0,4	0,45	0,44	0,47
D1				0,49	0,46	0,55	0,45	0,46	0,61	0,46	0,45	0,62	0,55	0,51	0,53	0,55	0,43	0,63	0,58	0,48
D2					0,45	0,47	0,41	0,43	0,48	0,43	0,44	0,49	0,49	0,44	0,43	0,45	0,4	0,44	0,46	0,41
D3						0,42	0,43	0,44	0,44	0,43	0,45	0,46	0,4	0,46	0,4	0,42	0,41	0,42	0,42	0,43
W1							0,43	0,44	0,55	0,45	0,47	0,55	0,55	0,49	0,52	0,52	0,42	0,51	0,52	0,48
W2								0,37	0,46	0,42	0,4	0,47	0,41	0,39	0,4	0,42	0,37	0,38	0,43	0,44
W3									0,46	0,43	0,45	0,48	0,47	0,46	0,42	0,44	0,38	0,44	0,44	0,46
D1										0,46	0,46	0,62	0,55	0,49	0,51	0,54	0,42	0,52	0,57	0,47
D2											0,45	0,48	0,44	0,43	0,41	0,45	0,41	0,45	0,44	0,44
D3												0,46	0,46	0,49	0,43	0,47	0,4	0,46	0,47	0,47
V1													0,53	0,5	0,51	0,55	0,44	0,5	0,6	0,48
V2														0,48	0,47	0,51	0,41	0,45	0,55	0,46
V3															0,49	0,48	0,41	0,48	0,48	0,5
Z1																0,48	0,39	0,5	0,5	0,46
Z2																	0,43	0,47	0,53	0,47
Z3																		0,4	0,43	0,41
Z1																			0,47	0,46
Z2																				0,47

Tabel 6.3: Percentage overeengekomen high risk locaties, gemiddelde over peilperioden 177, 187 en 197.

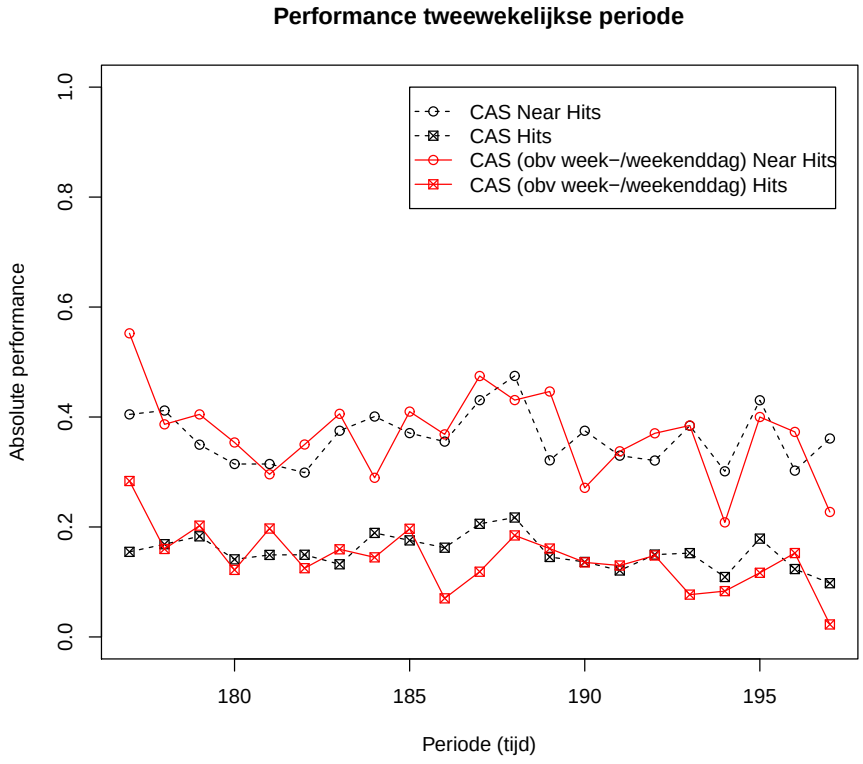
Er zijn geen duidelijke weekend en weekdagen verschillen. Gemiddeld komen de weekdagkaarten $46,4\%$ overeen en weekendkaarten $45,8\%$. De kaarten die wel het meest overeenkomen zijn de kaarten die voor het dagdeel nacht inbraken voorspellen. De diensttijden die op werkdagen en 's avonds plaatsvinden kennen een overeenkomst van $58,1\%$. De weekendkaarten die de avond voorspellen een overeenkomst van $50,0\%$. Alle kaarten die een avond voorspellen op een weekdag ten aanzien van een weekenddag kennen ook nog een overeenkomst van $52,4\%$. De andere 2 dagdelen kennen over het algemeen minder

overeenkomsten. Dagdeel 2 kent over de weekdagen een overeenkomst van $42,4\%$ en een overeenkomst van $53,0\%$ over weekenddagen. Dagdeel 3 kent over de weekdagen een overeenkomst van $45,6\%$ en een overeenkomst van $40,6\%$ over weekenddagen.

In paragraaf 6.7 wordt op basis van de overeenkomstigheidspercentages uit deze tabel gezocht naar het maximaal gemiddelde overeenkomstigheidspercentage wanneer alle diensttijdvensters worden verdeeld over twee modellen. Daar wordt gevonden dat bij een verdeling van 10 over 11 modellen alle nachtelijke dagdelen + de vrijdag, zaterdag en zondag overdag samen worden genomen. De overige 11 diensttijden vormen het andere model. Het maximaal haalbare gemiddelde over de twee modellen is een overeenkomstigheidspercetage van 48,20% wat boven het algemeen gemiddelde ligt voor beide modellen.

6.6 Resultaten woninginbraken per week- en weekenddag

De tweeweekse peilperiode kan worden opgeknipt in 2 afzonderlijke voorspellingen die voor de week- en weekenddagen de kans op een incident per locatie voorspellen. In paragraaf 5.2.3, 5.3.2 en 6.1.3 worden indicatoren gevonden die duiden op een verschillende geografische verdeling van incidenten tussen de week- en weekenddagen. Figuur 6.11 geeft de absolute performance weer van dit samengestelde model en ter vergelijking ook de performances van het huidige CAS model.



Figuur 6.11: Performance van de huidige CAS en een samengestelde CAS op basis van losse voorspellingen voor de 3 dagdelen op basis van een logistisch regressie model

De gemiddelde hits performance van CAS is $15,44\%$ ($\sigma = 3,04\%$) en het samengestelde model op basis van diensttijd $14,24\%$ ($\sigma = 5,56\%$). De gemiddelde near hits performance van CAS is $36,32\%$

($\sigma = 4,96\%$) en het samengestelde model op basis van diensttijd $36,86\%$ ($\sigma = 8,03\%$). Het huidige CAS model kent in $61,90\%$ van de perioden een hogere absolute hits performance dan het samengestelde model op basis van dagdeel: 13 van de 21 peilperioden. In $38,10\%$ van de perioden kent het huidige model ook een hogere near hits performance: 8 van de 21 peilperioden. Dit resultaat geeft geen indicatie dat één van de modellen beter incidenten kan voorspellen. Met de gepaarde t-test kan worden getoetst of de performances daadwerkelijk significant verschillen. Dit gebeurt aan de hand van de volgende hypotheses:

H_0: Het verschil in performance van de twee modellen is gelijk aan 0.
H_1: Het verschil in performance van de twee modellen *niet* gelijk aan 0.

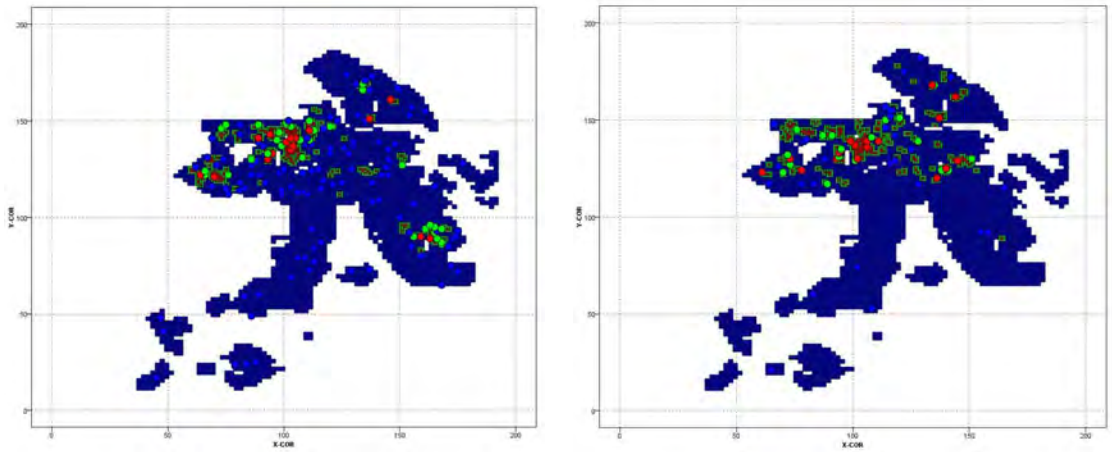
Op basis van de gepaarde t-test wordt H_0 niet verworpen voor de absolute hits performance ($T = 1,049$; $df = 20$; p-waarde$= 0,3067$;$\alpha = 0,05$) en ook niet voor de near hits performance ($T = -0,3342$; $df = 20$; p-waarde$= 0,7417$; $\alpha = 0,05$). De interpretatie van het aantal near hits hangt echter van het aantal toegewezen near hits locaties. Hierover kan een gewogen gemiddelde worden berekend waarbij het gemiddelde aantal near hits locaties wordt gewogen op basis van het aantal incidenten wat onderhevig was aan het aantal near hits locaties. Het samengestelde model heeft gemiddeld $1.039,03$ near hits locaties (per target) terwijl het huidige CAS model gemiddeld op 1.108 locaties zit.

Dit resultaat betekent dat er geen verschil is in performance voor beide methoden, of dat er niet voldoende bewijs is om aan te nemen dat de performances verschillen. Wel is de variantie van het samengestelde model groter dan van het huidige CAS model. Dit komt door het opknippen van het huidige model in twee kleinere modellen waardoor het aantal te voorspellen incidenten afneemt en de variantie inherent toeneemt, maar bij eerdere verknipte modellen is dit niet zo extreem toegenoemen. Door deze variantie vormt het huidge week-/weekendmodel niet direct aanleiding om de huidige CAS voorspellingen te vervangen met dit model, maar zou parallel gebruik hiervan een toevoeging kunnen zijn.

De relatieve hits performance is niet berekent voor het huidige CAS model, omdat deze tijdintervallen niet geen vooraf vastgesteld zijn. Het samengestelde model kent voor de weekdagen een relatieve hits performance van $0,79$ en voor de weekenddagen een gemiddelde van $0,80$.

Conclusie - Er kan geen duidelijk verschil worden gevonden tussen de performance van het samengestelde model op basis van week- en weekenddagen, en het huidige CAS model.

Figuur 6.12 geeft de CAS kaarten weer die toebehoren aan de week- en weekendvoorspelling voor de peilperiode 177. Wanneer gekeken wordt naar 3 peilperioden (177, 187 en 197) komen de kaarten gemiddeld op 34,63% overeenkomstige high risk locaties. Dit is ongeveer gelijk aan de schatting gegeven in paragraaf 6.1.3, waar op basis van alle afzonderlijke weekdagen een overeenkomstigheidspercentage van 33,54% wordt gemeten tussen week en weekenddagen.



Figuur 6.12: CAS-kaarten weekdagen (links) en weekenddagen (rechts) obv logistische regressie voor periode 177.

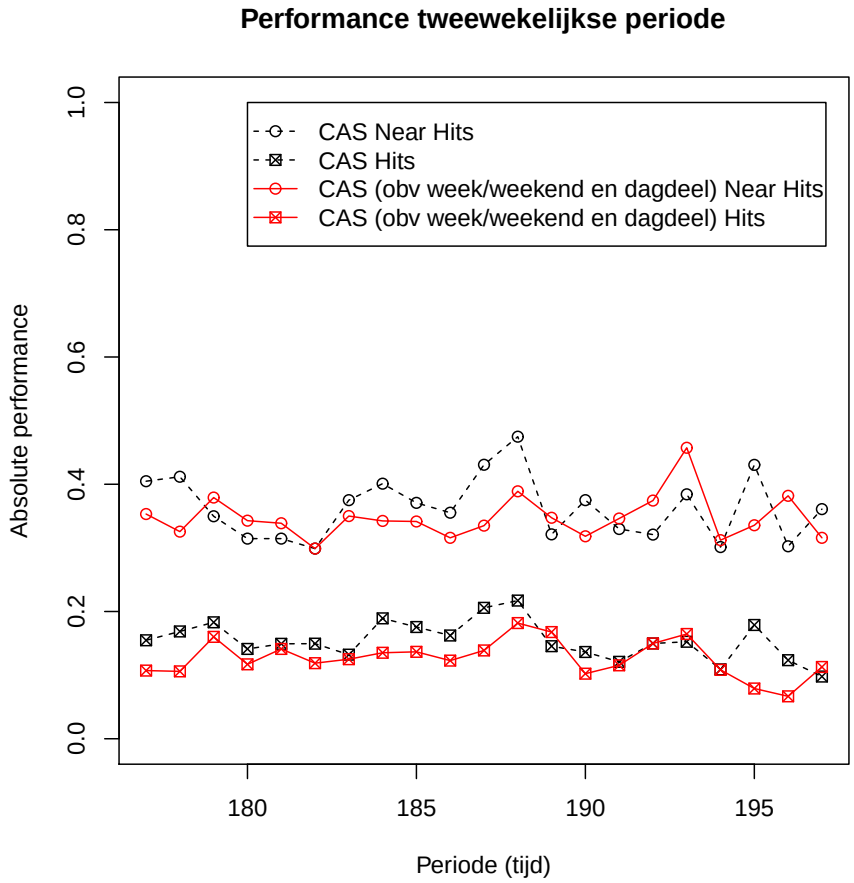
6.7 Resultaten woninginbraken per week-, weekenddag en dagdeel

De tweewekelijkse peilperiode kan worden opgeknipt in 6 afzonderlijke voorspellingen die elk de kans op een incident per locatie voor een specifiek dagdeel onderverdeeld in week- en weekenddagen voorspellen. Figuur 6.13 laat de absolute performance weer van dit samengestelde model en ter vergelijking ook de performances van het huidige CAS model.

De gemiddelde hits performance van CAS is $15,44\%$ ($\sigma = 3,04\%$) en van het samengestelde model $12,64\%$ ($\sigma = 2,87\%$). De gemiddelde near hits performance van CAS is $36,32\%$ ($\sigma = 4,96\%$) en het samengestelde model $34,76\%$ ($\sigma = 3,47\%$). Het huidige CAS model kent in $80,95\%$ van de perioden een hogere hits performance dan het samengestelde model: 17 van de 21 peilperioden. In $52,38\%$ van de perioden kent het huidige CAS model ook een hogere near hits performance: 11 van de 21 modellen. Dit resultaat geeft de indicatie dat het huidige CAS model beter incidenten kan voorspellen dan het samengestelde model op basis van week-, weekendagen en dagdelen. Met de gepaarde t-test kan worden getoetst of de performances daadwerkelijk significant verschillen. Dit gebeurt aan de hand van de volgende hypotheses:

H_0: Het verschil in performance van de twee modellen is gelijk aan 0.
H_1: Het verschil in performance van de twee modellen *niet* gelijk aan 0.

Op basis van de gepaarde t-test wordt H_0 verworpen voor de absolute hits performance ($T = 4,1429$; $df = 20$; p-waarde$= 0,0005037$; $\alpha = 0,05$) en niet verworpen voor de absolute near hits performance ($T = 1,3031$; $df = 20$; p-waarde$= 0,2073$; $\alpha = 0,05$). De interpretatie van het aantal near hits hangt echter van het aantal toegewezen near hits locaties. Hierover kan een gewogen gemiddelde worden berekend waarbij het gemiddelde aantal near hits locaties wordt gewogen op basis van het aantal incidenten wat onderhevig was aan het aantal near hits locaties. Het samengestelde model heeft gemiddeld $1053,60$ near hits locaties (per target) terwijl het huidige CAS model gemiddeld op 1.108 locaties zit. Dit resultaat betekent dat het aannemelijk is dat de absolute hits performance van de huidige CAS methodiek significant hoger



Figuur 6.13: Performance van de huidige CAS en een samengestelde CAS op basis van losse voorspellingen voor de 3 dagdelen op basis van een logistisch regressie model

is dan die van de samengestelde model. Op basis van de near hits performance kan geen verschil in performance worden gevonden. Het samengestelde model kent echter een ruim minder aantal near hits locaties, waardoor het lijkt dat het samengestelde model vaak dichtbij (near hit) zit, maar nog niet voldoende de daadwerkelijke incidenten weet te raken (hit).

De relatieve hits performance is niet berekent voor het huidige CAS model, omdat deze tijdintervallen niet geen vooraf vastgesteld zijn. Het samengestelde model kent in het algemeen een gemiddelde relatieve performance measure van $0,77$.

Conclusie - Het huidige CAS model over tijdsperioden van twee weken voorspeld beter dan het samengestelde model waarbij voor iedere combinatie week-/weekenddag en dagdeel een afzonderlijke voorspelling wordt gemaakt.

De CAS-kaarten die voortkomen uit de gegenereerde 6 voorspellingen kunnen worden vergeleken op basis van overeenkomstige high risk locaties, weergegeven in figuur 6.4.
Op basis van de overeenkomstigheidspercentages kunnen twee clusters worden gevonden die overeenkomen met de gevonden percentages op basis van diensttijd in 6.1.7: De tijdsintervallen week1, weekend1 en weekend2 lijken gemiddeld meer op elkaar (32,26% overeengekomende high risk locaties) en weekend3,

Weekdag	WEEK2	WEEK3	WEEKEND1	WEEKEND2	WEEKEND3
WEEK1	0,251	0,225	0,375	0,316	0,280
WEEK2		0,285	0,234	0,229	0,296
WEEK3			0,194	0,223	0,313
WEEKEND1				0,287	0,278
WEEKEND2					0,287

Tabel 6.4: Percentage overeengekomen high risk locaties, gemiddelde over peilperioden 177, 187 en 197

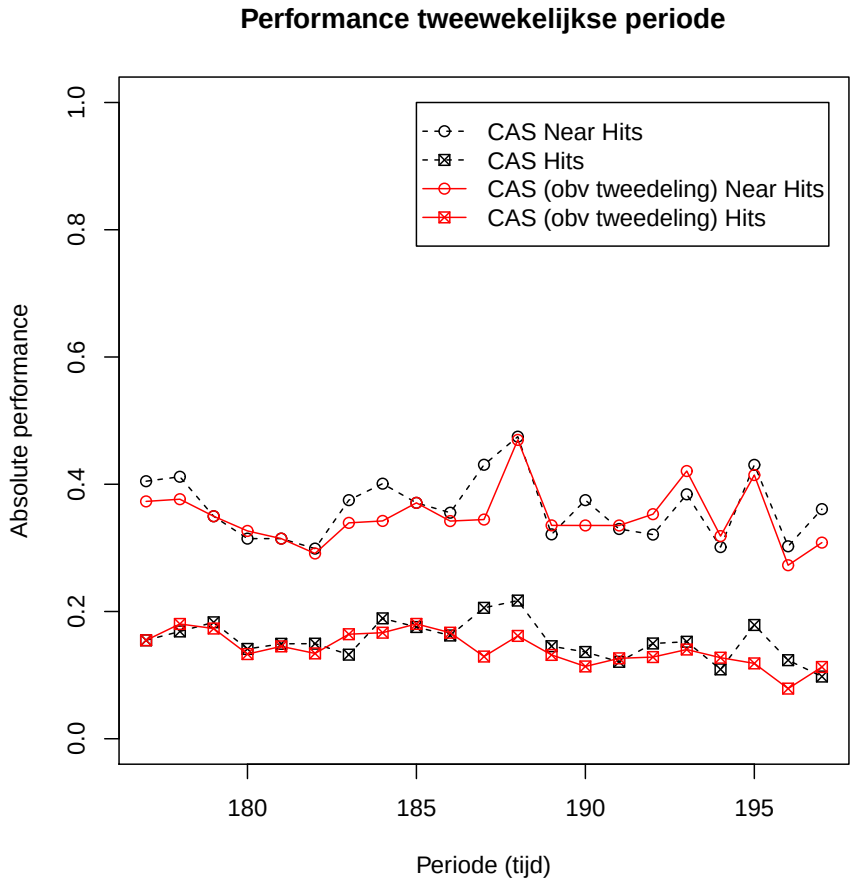
week2 en week3 lijken gemiddeld meer op elkaar (29,80% overeengekomende high risk locaties). Deze uitkomsten liggen allebei hoger dan het algemeen gemiddelde van 27,10%.

6.8 Resultaten woninginbraken op basis van tweedeling obv analyse

In hoofdstuk 5 zijn vooral onderverdelingen gevonden in de week en weekenddagen en tussen de dagdelen. Op basis van de voorspellingen van die uitkomst zijn voorspellingen gegenereerd op basis van week- en weekenddagen (6.5), dagdelen (6.3) en de combinatie tussen beide (6.6). Bij alle voorspellingen is daarnaast de overeenkomstigheid van de voorspellingen gemeten door het aantal overeengekomen high risk locaties tussen twee voorspellingen te bepalen. Op die manier kon worden bepaald in hoeverre twee voorspellingen (met een andere achterliggende tijdreeks van incidenten) dezelfde kaarten genereerden.

In paragraaf 6.1.7 worden percentages overeenkomstige high risk locaties gemeten tussen alle diensttijden. Wanneer hier gezocht wordt naar de maximaal haalbare gemiddelde overeenkomstigheid, wanneer de dienstvensters in twee delen worden verdeeld. Bij een splitsing van 10 om 11 diensttijden wordt bij een maximaal gemiddeld overeenkomstigheidspercentage gemeten van 48,20%. De diensttijden nacht en vrijdag t/m zondag overdag kennen een overeenkomst van gemiddeld 53,80%. De overige delen hanteren in dat geval 42,60% overeenkomst. Deze uitkomsten lijken te worden herhaald in de analyse in paragraaf 6.2.2 waar er hoge overeenkomsten worden gevonden tussen week1, weekend1 en weekend2 en weekend3, week2 en week3. Door de uitkomst in deze paragrafen wordt in deze paragraaf gekozen voor een tweedeling waarbij het eerste deel de tijdintervalen nacht en vrijdag t/m zondag overdag (Part A) bevat en het tweede deel de tijdsintervallen avond en maandag t/m donderdag overdag (Part B) bevat.

De peilperiode van twee weken wordt dus opgesplitst in 2 afzonderlijke voorspellingen die elk de kans op een incident per locatie voor één van de twee delen voorspeld op basis van een logistische regressie. Figuur 6.14 geeft de absolute performance weer van dit samengestelde model en ter vergelijking ook de performances van het huidige CAS model.
De gemiddelde hits performance van CAS is $15,44\%$ ($\sigma = 3,04\%$) en van het samengestelde model $14,13\%$ ($\sigma = 2,59\%$). De gemiddelde near hits performance van CAS is $36,32\%$ ($\sigma = 4,96\%$) en het samengestelde model $34,92\%$ ($\sigma = 4,47\%$). Het huidige CAS model kent in $61,90\%$ van de perioden een hogere hits performance dan het samengestelde model op basis van dagdeel: 13 van de 21 peilperioden. In $57,14\%$ kent het huidige CAS model ook een hogere near hits performance: 12 van de 21 peilperioden. In drie



Figuur 6.14: Performance van de huidige CAS en een samengestelde CAS op basis van 2 losse voorspellingen door een logistisch regressie model

peilperioden halen beide modellen een gelijke performance en in 6 perioden presteert het samengestelde model beter. Dit resultaat geeft indicatie dat het huidige CAS model beter incidenten kan voorspellen dan het samengestelde model op basis van de tweedeling. Opvallend is wel dat zowel het samengestelde model en het huidige model een bijna identieken verdeling en trend kennen, al doet het huidige model het in veel gevallen net iets beter. Met de gepaarde t-test kan worden getoetst of de performances significant verschillen. Dit gebeurt aan de hand van de volgende hypotheses:

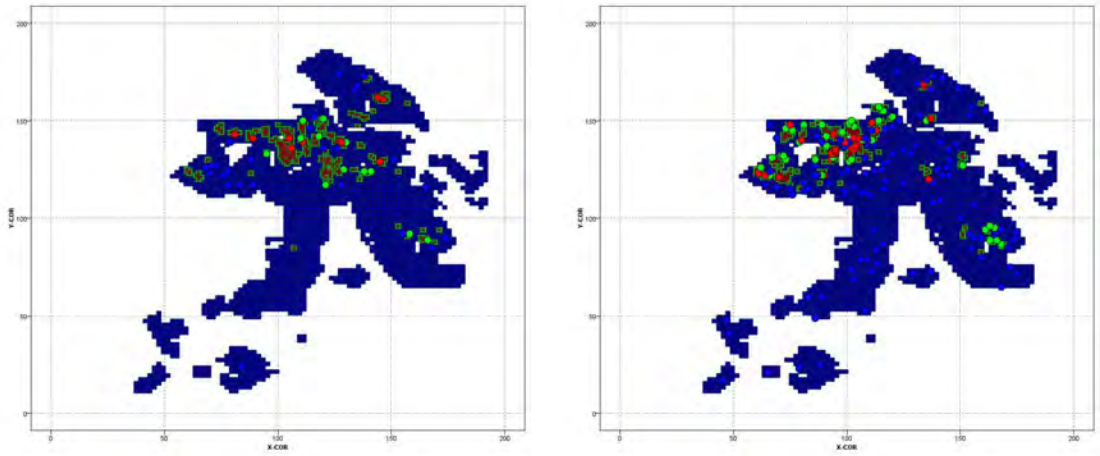
H_0: Het verschil in performance van de twee modellen is gelijk aan 0.
H_1: Het verschil in performance van de twee modellen *niet* gelijk aan 0.

Op basis van de gepaarde t-test wordt H_0 verworpen voor de absolute hits performance ($T = 2.1875$; $df = 20$; p-waarde$= 0,04075$; $\alpha = 0,05$) en niet verworpen voor de near hits performance ($T = 2,0779$; $df = 20$; p-waarde$= 0,05082$; $\alpha = 0,05$). De interpretatie van het aantal near hits hangt echter van het aantal toegewezen near hits locaties. Hierover kan een gewogen gemiddelde worden berekend waarbij het gemiddelde aantal near hits locaties wordt gewogen op basis van het aantal incidenten wat onderhevig was aan het aantal near hits locaties. Het samengestelde model heeft gemiddeld $1003,05$ near hits locaties (per target) terwijl het huidige CAS model gemiddeld op 1.108 locaties zit. Dit resultaat betekent dat

het aannemelijk is dat de absolute hits performance van de huidige CAS methodiek significant hoger is dan die van de samengestelde model. Op basis van de near hits performance kan geen verschil in performance worden gevonden. Het samengestelde model kent echter een ruim minder aantal near hits locaties, waardoor het lijkt dat het samengestelde model vaak dichtbij (near hit) zit, maar nog niet voldoende de daadwerkelijke incidenten weet te raken (hit).

De relatieve hits performance is niet berekent voor het huidige CAS model, omdat deze tijdintervallen niet geen vooraf vastgesteld zijn. Het samengestelde model kent voor part A een gemiddelde relatieve performance measure van $0,800$ en voor part B een gemiddelde van $0,804$. Dit zijn in vergelijking met vorige paragrafen hoge relatieve performance waarden.

Conclusie - Er kan geen duidelijk verschil worden gevonden tussen de performance van het samengestelde model en het huidige CAS model. In veel gevallen lijkt het huidige CAS model iets beter, maar er is geen overweldigend verschil.



Figuur 6.15: CAS-kaarten Part A (links) en Part B (rechts) obv logistische regressie voor periode 177.

Figuur 6.15 geeft de CAS kaarten weer die toebehoren de twee voorspellingen voor de peilperiode 177. Wanneer gekeken wordt naar 3 peilperioden (177, 187 en 197) komen de kaarten gemiddeld op 21,30% overeenkomstige high risk locaties.

6.9 Conclusie

In dit hoofdstuk zijn 5 modellen gepresenteerd die de tweeweekelijkse peilperioden opsplitsen in meerdere kleine perioden waar afzonderlijke voorspellingen voor worden gemaakt. De kleinere perioden kunnen vervolgens worden samengenomen om de hele tweewekelijkse periode te omvatten. De gepresenteerde modellen zijn toepast op 21 peilperioden om de performance van deze samengestelde modellen te vergelijken ten aanzien van het huidige CAS model waar incidenten worden voorspeld voor een periode van twee weken.

Weekdag - Het huidige CAS model dat een voorspelling maakt over twee weken voorspelt beter dan een samengesteld model waarbij voor iedere weekdag een afzonderlijke voorspelling wordt gemaakt. Daar-

naast voorspelt het huidige model ook beter incidenten voor iedere weekdag afzonderlijk, dan een voorspellling gespecificeerd op de weekdag.

Dagdeel - Er kan geen verschil worden gevonden tussen de performance van het samengestelde model op basis van dagdelen en het huidige CAS model. Het dagdeel nacht wordt beter voorspelt daoor het samengestelde model op basis van dagdeel en het dagdeel avond wordt beter voorspeld door het huidige model. Over het dagdeel dag wordt geen uitsluitsel gegeven.

Diensttijd - Het huidige CAS model dat een voorspelling maakt over twee weken voorspelt beter dan een samengesteld model waarbij voor iedere diensttijd een afzonderlijke voorspelling wordt gemaakt. Er zijn 6 diensttijden waar tussen performance van het samengestelde model en het huidige model geen verschil gevonden kan worden. De overige 15 diensttijden worden beter voorspeld met het huidige CAS model.

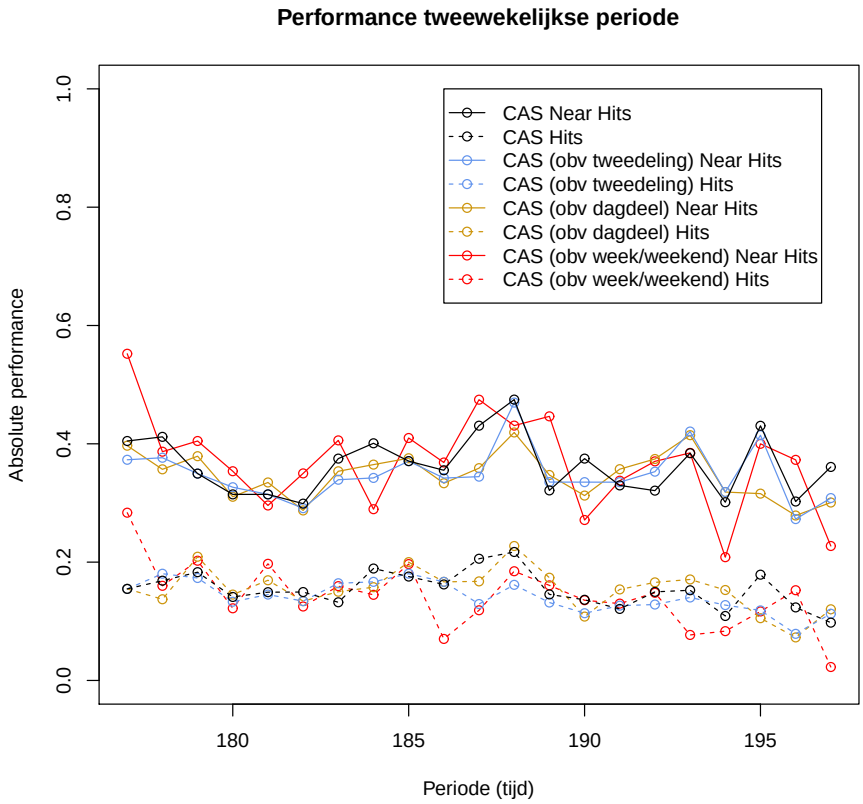
Week- en weekend - Er kan geen verschil gevonden worden tussen de performance van het samengestelde model op basis van week- en weekenddagen en het huidige CAS model.

Week-/weekend en dagdelen - Het huidige CAS model dat een voorspelling maakt over twee weken voorspelt beter dan een samengesteld model waarbij voor iedere week-/weekend en dagdeel combinatie een afzonderlijke voorspelling wordt gemaakt.

Tweedeling obv analyse - Er kan geen verschil gevonden worden tussen de performance van het samengestelde model (opdeling tussen de tijdsintervallen nacht en vrijdag t/m zondag overdag en de tijdsintervallen avond en maandag t/m donderdag overdag) en het huidige CAS model. In veel gevallen lijkt het huidige CAS model iets beter, maar er is geen overweldigend verschil.

Hoe kleiner de tijdsintervallen worden gemaakt, des te minder incidenten er per tijdinterval beschikbaar zijn om toekomstige incidenten op te voorspellen. Dat leidt in bijna alle gevallen tot een lagere performance van het samengestelde model dan het huidge model. De tweewekelijkse periode dus opsplitsen in kleinere intervallen en dezelfde performance halen, is dus alleen mogelijk wanneer de geografisch verschillen tussen de kleinere tijdsintervallen dermate groot zijn dat dit opweegt tegen het verlies in historie en onderscheidend vermogen van alle locaties.

De modellen op basis van dagdeel, week-/weekend en de tweedeling obv analyse worden aangewezen als modellen waar het onderscheidend vermogen van de verschillende tijdsintervallen opweegt tegen het verlies in historie en onderscheidend vermogen. In het volgende hoofdstuk worden alleen die drie modellen verder getest door gebruik te maken van andere voorspellende modellen. Figuur 6.16 geeft een plot van de performances van deze drie modellen.



Figuur 6.16: Performance van het huidige CAS model en de drie best presterende samengestelde modellen.

Hoofdstuk 7

Voorspellen van woninginbraken op tijdsintervalniveau II

CAS voorspelt momenteel voor iedere peilperiode van twee weken de kans op een woninginbraak in de regio Amsterdam. Het huidige CAS model kent een near hits performance van 0,3632 ($\sigma = 0,0496$) gebaseerd op de peilperioden 177 t/m 179 over tweewekelijkse peilperioden. In hoofdstuk 6 zijn incidenten voorspelt voor kleinere tijdsintervallen onderliggend aan de tweeweekse periode. De voorspellingen van deze kleinere tijdsintervallen worden samengenomen tot een periode van twee weken waarover de performance wordt gemeten om deze te vergelijken met de performance van het huidige CAS model. Van alle samengestelde modellen kan er tussen drie modellen geen significant verschil worden gevonden ten aanzien van het huidige CAS model. Deze modellen zullen naar verwachting de grootste kans hebben de performance van CAS te overstijgen wanneer er eventueel gebruik wordt gemaakt van een andere techniek. In dit hoofdstuk wordt daarop ingespeeld, door gebruik te maken van andere technieken dan het logistische regressie model om incidenten te voorspellen voor deze drie modellen. De centrale vraag in dit hoofdstuk is:

In hoeverre kan met gebruik van algoritmen de kans op een incident voor iedere gridlocatie m.b.t. een specifiek tijdsinterval worden voorspellen?

Dit hoofdstuk begint met een beschrijving van de gebruikte technieken in paragraaf 7.1. De resultaten van de nieuw gepresenteerde technieken op de geselecteerde modellen vinden plaats in paragraaf 7.2 t/m 7.4. Het model op basis van dagdeel in paragraaf 7.1, het model op basis van week- en weekend in paragraaf 7.3 en het model op basis van een tweedeling tot stand gekomen door een analyse in paragraaf 7.4. Tot slot volgt de conclusie in paragraaf 7.5.

7.1 Model omschrijving

In de vorige hoofdstukken is gebruik gemaakt van een CAS model waarbij alleen de data is gespecificeerd op een bepaald tijdsinterval onderliggend aan de standaard gebruikte peilperioden van twee weken. In dit hoofdstuk wordt ook gebruik gemaakt van een CAS model waarbij de data is gespecificeerd op basis

van een onderliggend tijdsinterval, maar wordt het logistische regressie model vervangen door een andere model of techniek. De technieken die in dit hoofdstuk worden gebruikt zijn een *neuraal netwerk met multi-layer perceptron*, een *Bayes netwerkmodel* en een *hotspotmodel*.

Neuraal netwerk met multi-layer perceptron Een multi-layer perceptron (MLP) is een neuraal netwerk (NN) dat data projecteert vanuit input nodes via een netwerk van neuronen op passende output-nodes. De aanduiding neuron is afgeleid van de neurons in ons zenuwstelsel. Wanneer zulke zenuwcellen voldoende geprikkeld zijn, versturen ze een signaal. Neuronen zijn dus bijzonder geschikt voor het ontvangen, verwerken en versturen van signalen. Neuronen binnen neurale netwerken zijn geinspireerd op het gedrag van neuronen in de hersenen en kunnen aan elkaar worden gekoppeld en vervolgens stapsgewijs worden geoptimaliseerd. In iedere stap wordt informatie van een vakje aan het netwerk aangeboden en vervolgens wordt de uitkomst vergeleken met de daadwerkelijke feiten: heeft er ook in de twee weken na het peilmoment een incident plaatsgevonden? Deze uitkomst wordt vervolgens teruggekoppeld aan het netwerk en de neuronen zijn in staat daarop te anticiperen en zichzelf bij te stellen. Dit proces wordt backpropagation genoemd vanwege het achteraf bijstellen van de neuronen. Het leerproces kent dus een supervised leerproces doordat terugkoppeling vanuit de werkelijkheid het model bijstuurt waarbij het gebruik maakt van een niet-lineaire activatiefunctie. Als output wordt een kanswaarde tussen 0 en 1 per locatie per gewenst tijdsinterval.

Bayes netwerkmodel Als alternatief voor neurale netwerken wordt ook vaak het Bayes netwerk model genoemd. Dit is een model waarbij gebruik wordt gemaakt van voorwaardelijke kansen uit de statistiek. Het grote voordeel van deze modellen is het niet beschikbaar hoeven hebben van grote hoeveelheden trainingsdata en zijn Bayes modellen oplettend naar uitzonderingen die wel gedifferentieerd kunnen worden. Doordat het een kansmodel genereerd wordt op basis van het voorkomen van combinaties in het verleden speelt de gebruikte hoeveelheid historische data een grote rol. Deze paragraaf bestaat uit een modelbeschrijving, resultaten en tot slot de conclusie.

Een Bayes netwerk model wordt ook wel een probabilistisch netwerk genoemd. Deze netwerken zijn volledig gebasseerd op de kansregel van Bayes. Deze regel geeft de kans weer dat een bepaalde mogelijkheid ten grondslag ligt aan de gebeurtenis uitgedrukt in voorwaardelijke kansen op de gebeurtenis van elk van de mogelijkheden. Formule 7.1 geeft de kansregel van Bayes weer.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} \tag{7.1}$$

De regel van Bayes betekent concreet dat gebeurtenis B kan plaatsvinden wanneer ook A heeft plaatsgevonden of wanneer A niet heeft plaatsgevonden. Wanneer de voorwaardelijke kansen op B zijn gegeven, kan de kans bepaald worden dat wanneer B is gebeurd, dit is gebeurd onder de omstandigheid dat A ook is gebeurd. Bij het toepassen hiervan wordt uitgegaan van a-priori kansen, die op basis van eerder onderzoek zijn verkregen. En dat laatste is precies het uitgangspunt van de verdeling van incidenten in de tijd. Stel een incident heeft plaatsgevonden, in hoeverre is dat dan gebeurd onder de omstandigheid dat A ook is gebeurd, waarbij A een willekeurige responsvariabele is.

Op basis van deze voorwaardelijk kansen kan een Bayes netwerk worden opgesteld. Een Bayes netwerk is een graaf zonder cykels waarbij alle responsvariabelen worden aangeduid als knoop. Er worden vervolgens pijlen tussen de knopen gespannen die de directe invloed weergeven. Met de pijlen mee worden op basis van de voorwaardelijke kansen een totale kans gegenereerd voor de target: wel of geen inbraak.

Hotspotmodel Hotspotmodellen baseren zich op de aanname dat incidenten daar gaan gebeuren waar in het verleden ook incidenten hebben plaatsgevonden. Het gebruikte hotspotmodel in dit hoofdstuk voorspeld voor iedere gridlocatie de kans op een incident door deze gelijk te stellen aan het percentage incidenten dat het *afgelopen jaar* heeft plaatsgevonden in betreffende gridlocatie. Op die manier krijgen alle locaties een kans op een incident in de toekomstige peilperiode toegewezen. Bij het berekenen van de high risk locaties wordt de top 3% van de locaties (282 locaties) met de hoogste kans op een incident geselecteerd. Door het gebruik van een percentage als kans ontstaan veel locaties met een gelijke kans rond de cutoff grens van de 282 locaties. De locaties die wel of niet zijn meegenomen, zijn daarbij random gesampled om toch tot 282 locaties te komen en niet meer of minder.

7.2 Resultaten modellen obv dagdeel

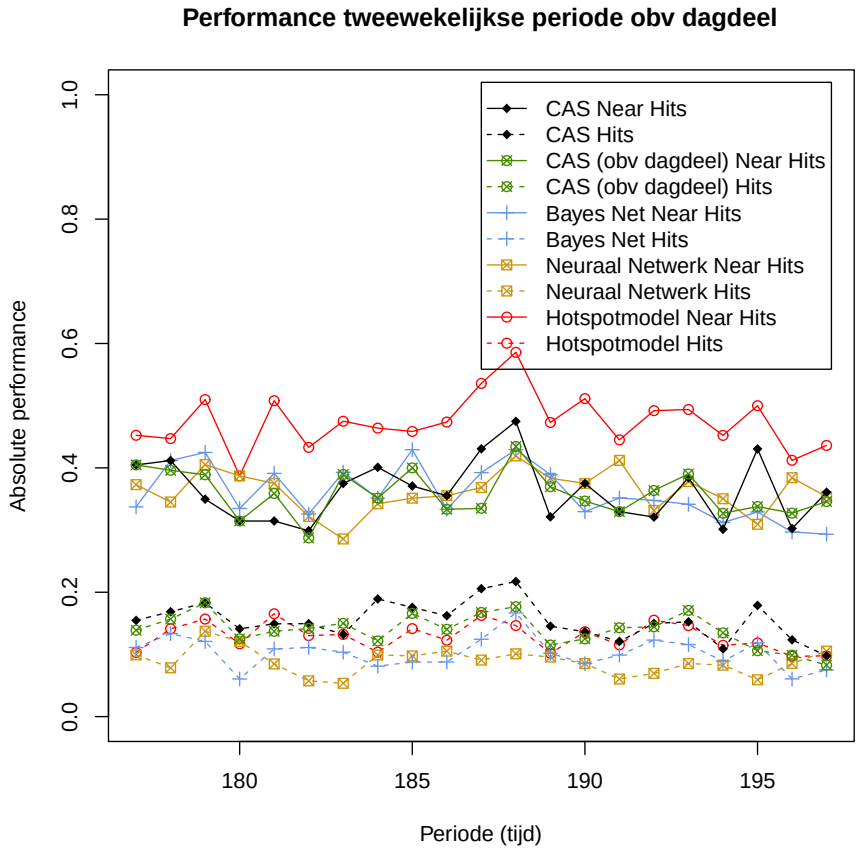
De tweeweekse peilperiode kan worden opgeknipt in drie afzonderlijke voorspellingen die voor ieder dagdeel de kans op een incident voorspellen. In paragraaf 6.4 is deze methodiek toegepast en zijn drie afzonderlijke voorspellingen gemaakt door middel van CAS obv een logistisch regressie model. Tussen de combinatie van deze drie voorspellingen en het huidige CAS model kunnen geen duidelijke verschillen in performance worden gevonden, wat de indruk geeft dat beide modellen gelijk presteren.

De drie afzonderlijke voorspellingen per dagdeel worden ter vergelijking voorspelt door middel van een neuraal netwerk, Bayes netwerk en hotspotmodel. Figuur 7.1 geeft de absolute performance weer van de modellen op basis van CAS, dagdeel specifieke CAS, Bayes netwerk, neurale netwerk en hotspotmodel.

Absolute hits performance De modellen op basis van een neuraal en Bayes netwerk presteren voor alle 21 peilperioden een lagere hits performance dan het huidige CAS model en preseteren beide eenmaal beter dan het CAS model obv dagdeel. Het hotspotmodel kent 3 van de 21 peilperioden een hogere hits performance dan het huidige CAS model en in drie perioden een gelijke performance. Voor 5 modellen weet het hotspotmodel een hogere performance te halen dan het CAS model op basis van dagdeel. Met de gepaarde t-test kan worden getoetst of de performances significant verschillen. Dit gebeurt aan de hand van de volgende hypotheses:

H_0: Het verschil in performance van de twee modellen is gelijk aan 0.
H_1: Het verschil in performance van de twee modellen *niet* gelijk aan 0.

Op basis van de gepaarde t-test wordt H_0 verworpen voor het verschil in absolute hits performance tussen het huidige CAS model en van het hotspotmodel ($T = 4,2904$; $df = 20$; p-waarde$=0,0004$; $\alpha = 0,05$), het neurale netwerk ($T = 9,3205$; $df = 20$; p-waarde$=1,019E-8$; $\alpha = 0,05$) en het Bayes netwerk ($T = 9,6445$; $df = 20$; p-waarde$=5,795E-9$; $\alpha = 0,05$). Op basis van de gepaarde t-test wordt H_0

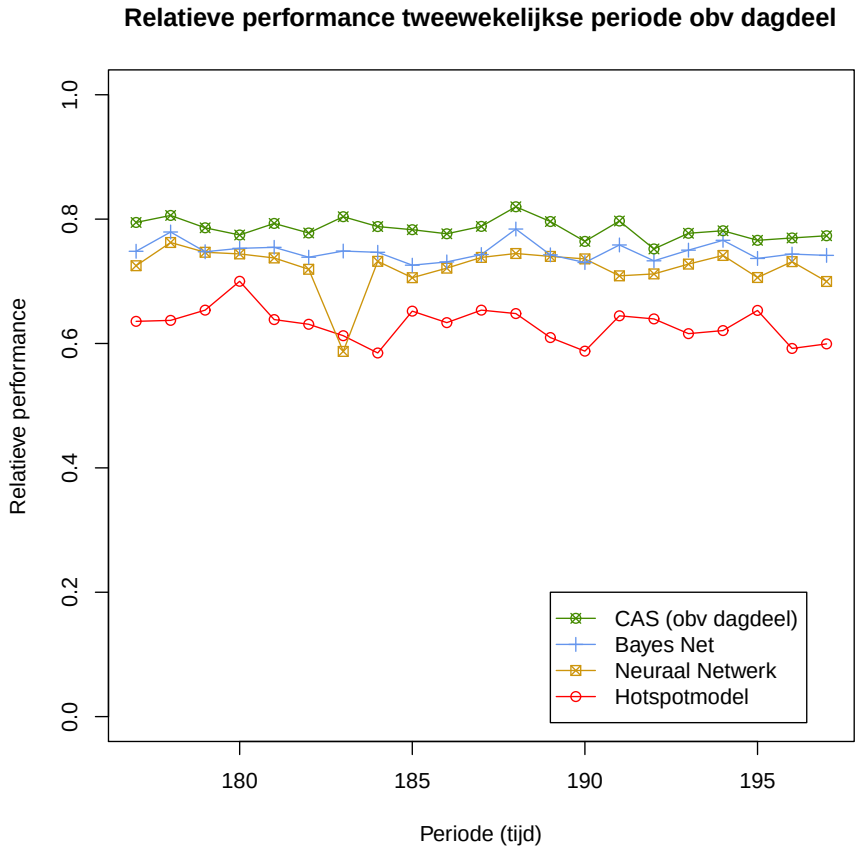


Figuur 7.1: Absolute performance op basis van dagdeel.

verworpen voor het verschil in absolute hits performance tussen het CAS model obv dagdeel en van het hotspotmodel, het neurale netwerk en het Bayes netwerk ($\alpha = 0,05$). Dit betekent dat op basis van de absolute hits performance de andere methoden minder goed presteren dan het dagdeel model op basis van logistische regressie.

Absolute near hits performance De absolute near hits performance is ondergeschikt aan de absolute hits performance omdat deze afhankelijk is van het aantal near hits locaties. Het hotspotmodel lijkt bijvoorbeeld een bijzonder hoge performance te kennen, maar heeft ook een gemiddeld aantal near hits locaties per target van 1.712,04 waar het huidige CAS model op gemiddeld 1.108 locaties zit. Ook het gemiddelde aantal near hit locaties van het Bayes netwerk (1.232,24) en het neurale netwerk (1.463,80) liggen hoog. De grote verschillen in aantallen near hits locaties worden voornamelijk veroorzaakt door het clusteren van high risk locaties waardoor het aantal omringende near hits locaties afneemt. Door de grote verschillen in aantallen locaties wordt geen uitspraak gedaan over de performance van de modellen op basis van de near hits performance.

Relatieve hits performance Op basis van de relatieve hits performance kan een betere uitspraak worden gedaan over de performance van het gegenereerde model over alle targets heen. Figuur 7.2 geeft de gemiddelde relatieve performance weer van de verschillende modellen. Het huidige CAS model is niet toegevoegd omdat deze een ander aantal incidenten per voorspelling kent, waardoor de relatieve



Figuur 7.2: Relatieve performance op basis van dagdeel.

performance measures niet vergelijkbaar zijn (paragraaf 3.4.1).

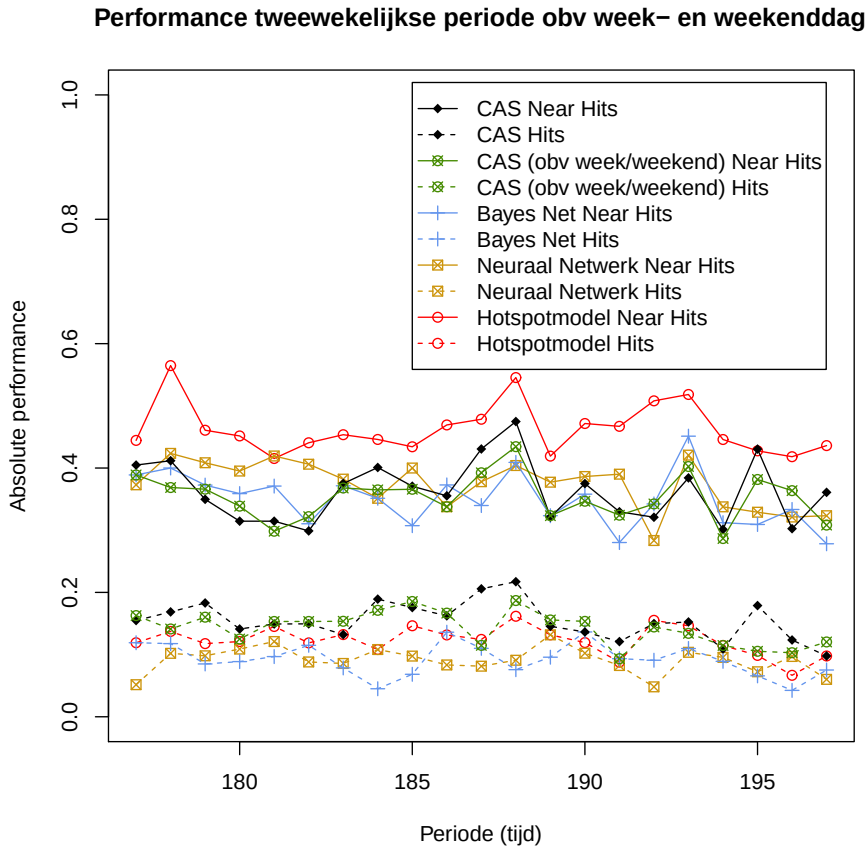
Van alle modellen presteert het dagdeelmodel obv logistische regressie het beste en verslaat daarmee de andere modellen in performance. Na het model obv een logistische regressie presteert het Bayes netwerk het best gevolgd door het neurale netwerk. De performance van het hotspotmodel is beduidend lager en kent ook een hogere variantie.

Conclusie - Een CAS model aangedreven door een logistisch regressie model voor de drie afzonderlijke dagdelen voorspelt beter incidenten dan een gelijk model op basis van een neuraal netwerk, bayes netwerk of hotspotmodel.

7.3 Resultaten modellen obv week- en weekend

De tweeweekse peilperiode kan worden opgeknipt in twee afzonderlijke voorspellingen die voor de week- en weekenddagen de kans op een incident voorspellen. In paragraaf 6.6 is deze methodiek toegepast en zijn twee afzonderlijke voorspellingen gemaakt door middel van CAS obv een logistisch regressie model. Tussen de combinatie van deze twee voorspellingen en het huidige CAS model kunnen geen duidelijke verschillen in performance worden gevonden, wat de indruk geeft dat beide modellen gelijk presteren.

De twee afzonderlijke voorspellingen voor de week- en weekenddagen worden ter vergelijking voorspelt door middel van een neuraal netwerk, Bayes netwerk en hotspotmodel. Figuur 7.3 geeft de absolute performance weer van de modellen op basis van CAS, dagdeel specifieke CAS, Bayes netwerk, neurale netwerk en hotspotmodel.



Figuur 7.3: Absolute performance op basis van week- en weekenddagen.

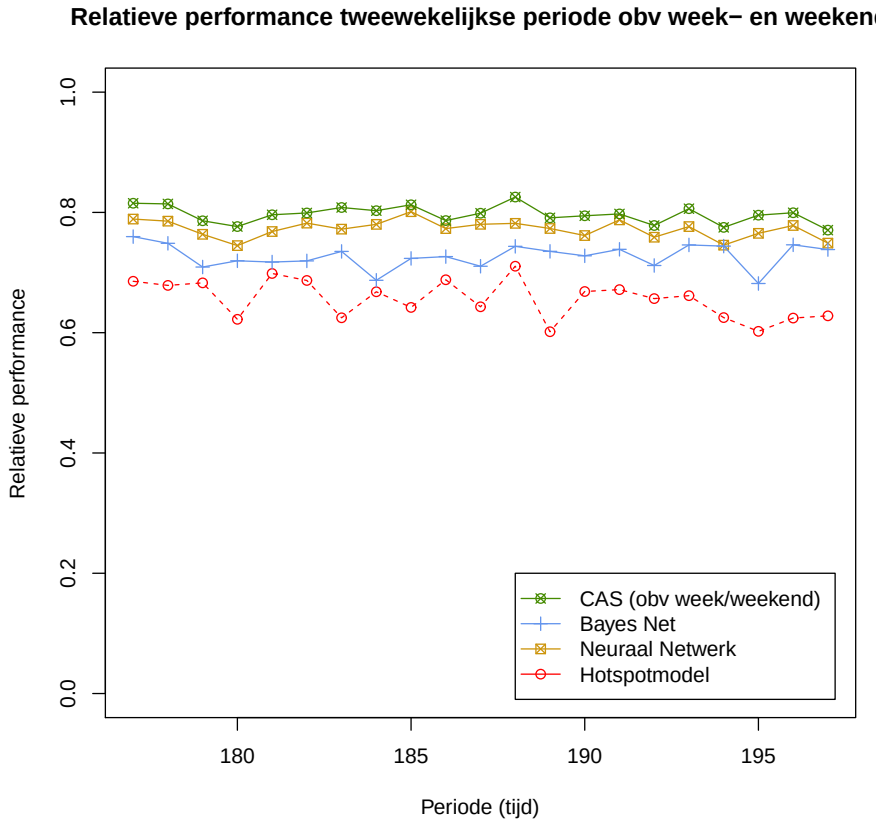
Absolute hits performance De modellen op basis van een neuraal en Bayes netwerk presteren voor 20 peilperioden een lagere hits performance dan het huidige CAS model en 1 keer wordt een gelijke performane gehaald. Beide modellen presteren ook eenmaal gelijk aan het CAS model obv dagdeel enn in 20 perioden wordt een lagere performance gehaald. Het hotspotmodel kent 2 van de 21 peilperioden een hogere hits performance dan het huidige CAS model en in twee perioden een gelijke performance. Voor 3 modellen weet het hotspotmodel een hogere performance te halen dan het CAS model op basis van dagdeel. Met de gepaarde t-test kan worden getoetst of de performances significant verschillen. Dit gebeurt aan de hand van de volgende hypotheses:

H_0: Het verschil in performance van de twee modellen is gelijk aan 0.
H_1: Het verschil in performance van de twee modellen *niet* gelijk aan 0.

Op basis van de gepaarde t-test wordt H_0 verworpen voor het verschil in absolute hits performance tussen het huidige CAS model en van het hotspotmodel ($T = 5,0253$; $df = 20$; p-waarde$=6,485E-5$;

$\alpha = 0,05$), het neurale netwerk ($T = 8,2237$; $df = 20$; p-waarde=$7,593E-8$; $\alpha = 0,05$) en het Bayes netwerk ($T = 7,0769$; $df = 20$; p-waarde=$7,338E-7$; $\alpha = 0,05$). Op basis van de gepaarde t-test wordt H_0 verworpen voor het verschil in absolute hits performance tussen het CAS model obv dagdeel en van het hotspotmodel, het neurale netwerk en het Bayes netwerk ($\alpha = 0,05$). Dit betekent dat op basis van de absolute hits performance de andere methoden minder goed presteren dan het dagdeel model op basis van logistische regressie.

Absolute near hits performance De absolute near hits performance is ondergeschikt aan de absolute hits performance omdat deze afhankelijk is van het aantal near hits locaties. Het hotspotmodel lijkt bijvoorbeeld een bijzonder hoge performance te kennen, maar heeft ook een gemiddeld aantal near hits locaties per target van 1.721,84 waar het huidige CAS model op gemiddeld 1.108 locaties zit. Ook het gemiddelde aantal near hit locaties van het Bayes netwerk (1.273,34) en het neurale netwerk (1.468,61) liggen hoog. De grote verschillen in aantallen near hits locaties worden voornamelijk veroorzaakt door het clusteren van high risk locaties waardoor het aantal omringende near hits locaties afneemt. Door de grote verschillen in aantallen locaties wordt geen uitspraak gedaan over de performance van de modellen op basis van de near hits performance.



Figuur 7.4: Relatieve performance op basis van week- en weekenddagen.

Relatieve hits performance Op basis van de relatieve hits performance kan een betere uitspraak worden gedaan over de performance van het gegenereerde model over alle targets heen. Figuur 7.4 geeft

de gemiddelde relatieve performance weer van de verschillende modellen. Het huidige CAS model is niet toegevoegd omdat deze een ander aantal incidenten per voorspelling kent, waardoor de relatieve performance measures niet vergelijkbaar zijn (paragraaf 3.4.1).

Van alle modellen presteert het week/weekendmodel obv logistische regressie het beste en verslaat daarmee de andere modellen in performance. Na het model obv een logistische regressie presteert het neurale netwerk het best gevolgd door het Bayes netwerk. De performance van het hotspotmodel is beduidend lager en kent ook een hogere variantie. Deze uitkomst is iets afwijkender dan bij de modellen op basis van dagdeel, waar het Bayes netwerk het neurale netwerk overtrof. Wellicht is het neurale netwerk een betere methode om te voorspellen wanneer het aantal incidenten te voorspellen groter is.

Conclusie - Een CAS model aangedreven door een logistisch regressie model voor de twee afzonderlijke week- en weekendmodellen voorspelt beter incidenten dan een gelijk model op basis van een neuraal netwerk, bayes netwerk of hotspotmodel.

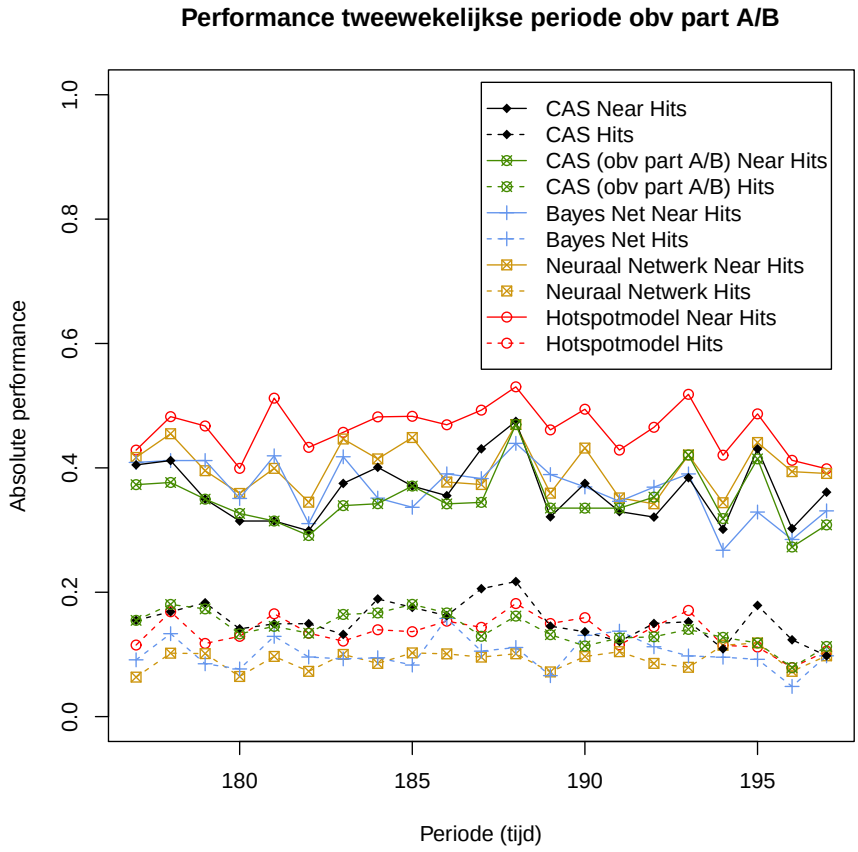
7.4 Resultaten modellen obv tweedeling

De tweeweekse peilperiode kan worden opgeknipt in twee afzonderlijke voorspellingen waarbij het het ene model de kansen voorspelt voor de tijdsintervallen nacht en vrijdag t/m zondag overdag (Part A) en het andere model voor de tijdsintervallen avond en maandag t/m donderdag overdag (Part B) de kans op een incident voorspellen. De keuze voor deze opdeling wordt onderbouwd door een analyse in paragraaf 6.8. In die paragraaf is deze methodiek ook toegepast en zijn twee afzonderlijke voorspellingen gemaakt door middel van CAS obv een logistisch regressie model. Tussen de combinatie van deze twee voorspellingen en het huidige CAS model kunnen geen duidelijke verschillen in performance worden gevonden, wat de indruk geeft dat beide modellen gelijk presteren.

De twee afzonderlijke voorspellingen voor de twee delen worden ter vergelijking voorspelt door middel van een neuraal netwerk, Bayes netwerk en hotspotmodel. Figuur 7.3 geeft de absolute performance weer van de modellen op basis van CAS, dagdeel specifieke CAS, Bayes netwerk, neurale netwerk en hotspotmodel.

Absolute hits performance De modellen op basis van een neuraal en Bayes netwerk presteren voor 19 peilperioden een lagere hits performance dan het huidige CAS model en 1 keer wordt een gelijke performane gehaald en 1 keer een betere performance. Het Bayes netwerk presteert ook tweemaal gelijk aan het CAS model obv dagdeel, waar het neurale netwerk nooit de performance van het CAS model obv dagdeel overtreft. Het hotspotmodel kent 6 van de 21 peilperioden een hogere hits performance dan het huidige CAS model en in twee perioden een gelijke performance. Voor 7 modellen weet het hotspotmodel een hogere performance te halen dan het CAS model op basis van dagdeel. Met de gepaarde t-test kan worden getoetst of de performances significant verschillen. Dit gebeurt aan de hand van de volgende hypotheses:

H_0: Het verschil in performance van de twee modellen is gelijk aan 0.



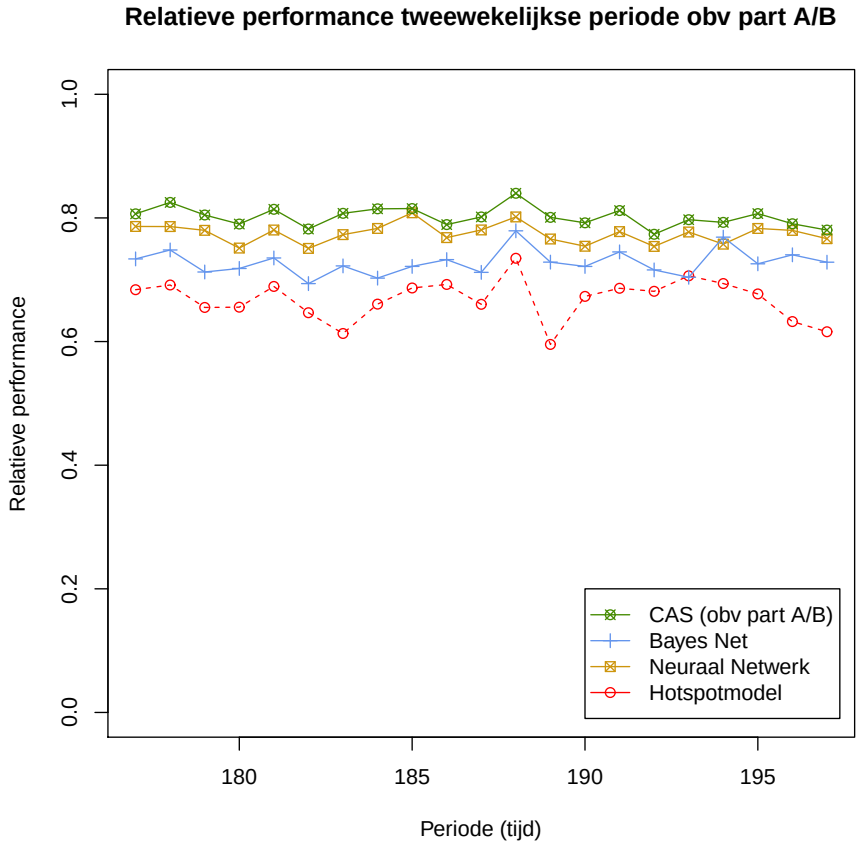
Figuur 7.5: Absolute performance op basis van part A en B.

H_1: Het verschil in performance van de twee modellen *niet* gelijk aan 0.

Op basis van de gepaarde t-test wordt H_0 verworpen voor het verschil in absolute hits performance tussen het huidige CAS model en van het hotspotmodel ($T = 2,9613$; $df = 20$; p-waarde=$0,00772$; $\alpha = 0,05$), het neurale netwerk ($T = 8,8203$; $df = 20$; p-waarde=$2,499E-8$; $\alpha = 0,05$) en het Bayes netwerk ($T = 6,3985$; $df = 20$; p-waarde=$3,051E-6$; $\alpha = 0,05$). Op basis van de gepaarde t-test wordt H_0 verworpen voor het verschil in absolute hits performance tussen het CAS model obv dagdeel en van het neurale netwerk en het Bayes netwerk ($\alpha = 0,05$). Voor het hotspotmodel en CAS obv dagdeel wordt H_0 niet verworpen ($T = 0,9208$; $df = 20$; p-waarde=$0,3681$; $\alpha = 0,05$). Dit betekent dat op basis van de absolute hits performance de andere methoden minder goed presteren dan het huidige CAS model op basis van logistische regressie. Het CAS model obv part A/B presteert ook beter dan het Bayes en neuraal netwerk.

Absolute near hits performance De absolute near hits performance is ondergeschikt aan de absolute hits performance omdat deze afhankelijk is van het aantal near hits locaties. Het hotspotmodel lijkt bijvoorbeeld een bijzonder hoge performance te kennen, maar heeft ook een gemiddeld aantal near hits locaties per target van 1.658,19 waar het huidige CAS model op gemiddeld 1.108 locaties zit. Ook het gemiddelde aantal near hit locaties van het Bayes netwerk (1.471,83) en het neurale netwerk (1.274,99) liggen hoog. De grote verschillen in aantallen near hits locaties worden voornamelijk veroorzaakt door

het clusteren van high risk locaties waardoor het aantal omringende near hits locaties afneemt. Door de grote verschillen in aantallen locaties wordt geen uitspraak gedaan over de performance van de modellen op basis van de near hits performance.



Figuur 7.6: Relatieve performance op basis van part A en B.

Relatieve hits performance Op basis van de relatieve hits performance kan een betere uitspraak worden gedaan over de performance van het gegenereerde model over alle targets heen. Figuur 7.6 geeft de gemiddelde relatieve performance weer van de verschillende modellen. Het huidige CAS model is niet toegevoegd omdat deze een ander aantal incidenten per voorspelling kent, waardoor de relatieve performance measures niet vergelijkbaar zijn (paragraaf 3.4.1).

Van alle modellen presteert het de verdeling van part A/B obv logistische regressie het beste en verslaat daarmee de andere modellen in performance. Na het model obv een logistische regressie presteert het neurale netwerk het best gevolgd door het Bayes netwerk. De performance van het hotspotmodel is beduidend lager en kent ook een hogere variantie. Deze uitkomst is iets afwijkender dan bij de modellen op basis van dagdeel, waar het Bayes netwerk het neurale netwerk overtrof, maar gelijk aan de uitkomsten bij week-/weekendmodellen.

Conclusie - Een CAS model aangedreven door een logistisch regressie model voor de twee modellen waarbij het eerste model een voorspelling maakt voor de tijdsintervallen nacht en vrijdag t/m zondag overdag (Part A) en het andere model voor de tijdsintervallen avond en maandag t/m donderdag overdag (Part B), voorspelt beter incidenten dan een gelijk model op basis van een neuraal netwerk, bayes netwerk of hotspotmodel.

7.5 Conclusie

In hoofdstuk 6 wordt CAS obv logistische regressie gebruikt om incidenten te voorspellen voor kleinere tijdsintervallen dan de standaard peilperioden van twee weken. De performance van de modellen specifiek toegepast op kleinere tijdsintervallen wordt berekend door het samenvoegen van de performance van meerdere kleinere tijdsintervallen tot een standaard periode van twee weken. In dit hoofdstuk is voor de drie samengestelde modellen die obv logistische regressie de beste performance leveren, gekeken naar de prestatie van dezelfde modellen wanneer een andere techniek gebruikt wordt: een neuraal netwerk met multi-layer perceptron, een Bayes netwerkmodel en een hotspotmodel.

Dagdeelmodel - Een CAS model aangedreven door een logistisch regressie model voor de drie afzonderlijke dagdelen voorspelt beter incidenten dan een gelijk model op basis van een neuraal netwerk, bayes netwerk of hotspotmodel.

Week- en weekendmodel - Een CAS model aangedreven door een logistisch regressie model voor de twee afzonderlijke week- en weekendmodellen voorspelt beter incidenten dan een gelijk model op basis van een neuraal netwerk, bayes netwerk of hotspotmodel.

Part A/B model - Een CAS model aangedreven door een logistisch regressie model voor de twee modellen waarbij het eerste model een voorspelling maakt voor de tijdsintervallen nacht en vrijdag t/m zondag overdag (Part A) en het andere model voor de tijdsintervallen avond en maandag t/m donderdag overdag (Part B), voorspelt beter incidenten dan een gelijk model op basis van een neuraal netwerk, bayes netwerk of hotspotmodel.

Hoofdstuk 8

Conclusie en aanbevelingen

8.1 Conclusie

Het huidige CAS model kan $36,3\%$ van de woninginbraken en $57,7\%$ van de straatroven voorspellen[1]. Deze voorspellingen worden in de huidige CAS omgeving gebaseerd op tweeweekse peilperioden en als aanvulling worden op deze voorspellingen ook voorspellingen gebaseerd voor de onderliggende tijdsintervallen *weekdag, dagdeel* en *diensttijd*. Hiervoor wordt echter alleen de high risk area van de tweewekelijkse voorspelling herzien, waardoor het model leunt op de aanname dat de geografische verspreiding van incidenten identiek is voor alle onderliggende tijdsintervallen. Deze aanname is echter nooit theoretisch onderbouwd. De volgende hoofdvraag is op basis van deze probleemstelling geformuleerd:

In hoeverre zijn de huidige tweewekelijkse voorspellingen geschikt om onderliggende tijdsintervallen te voorspellen die mogelijk een afwijkende geografische voorspelling hebben?

Toepassing huidige voorspellingen op onderliggende tijdsintervallen Er is onderzocht of de tweewekelijkse voorspellingen gelijk aansluiten op alle onderliggende tijdsintervallen door de performances van de onderliggende tijdsintervallen weekdag, dagdeel en diensttijd ten aanzien van de overall tweewekelijkse voorspelling te vergelijken. Hieruit zijn de volgende conclusies gevonden:

1. De tweewekelijkse voorspellingen van het huidige CAS model voor woninginbraken kennen voor alle werkdagen een ongeveer gelijke performance en daarmee lijkt de verdeling in weekdagen zich redelijk te verhouden tot de overall tweewekelijkse performance. Bij een splitsing in dagdelen sluiten de tweewekelijkse voorspellingen significant beter aan op het dagdeel avond dan op de dagdelen nacht en dag. Wanneer de performances over diensttijden worden geanalyseerd kunnen geen sterke conclusies getrokken worden doordat er te weinig incidenten plaatsvinden in een diensttijd. De incidenten zijn daarmee ook verdeeld over 21 tijdsintervallen, waar normaal één interval gebruikt werd.

[1] Berekend op basis van de peilperioden 177 t/m 197.

2. De tweewekelijkse voorspelling door middel van het huidige CAS model sluit voor straatroven beter aan op de weekenddagen zaterdag en zondag en minder goed op de maandag. Dit geeft de indruk dat voor straatroven het weekend beter aansluit op de tweewekelijkse voorspellingen. Bij een splitsing in dagdelen sluiten de tweewekelijkse voorspellingen significant beter aan op het dagdeel nacht dan op de dagdelen dag en avond. Door het gebrek aan incidenten tijdens een diensttijd is het niet mogelijk daarover een sterke conclusie te formuleren.

Bij zowel straatroven als woninginbraken is een duidelijke conclusie te trekken op basis van dagdelen. Voor woninginbraken vindt 46,2% van de inbraken plaats in de avond en het lijkt dat de voorspelling zich daar meer op aansluit. Voor straatroven vinden ook de meeste straatroven plaats in de avond (47,7%), maar toch sluit de voorspelling beter aan op de nacht waar 32,9% van de straatroven plaatsvindt. Van alle straatroven in de nacht vindt 54,0% plaats in het centrum wat een relatief klein oppervlak is en daardoor vermoedelijk het makkelijkst te voorspellen is doordat in de nacht de straatroven geclusterd plaatsvinden. Een soortgelijke clustering in minder sterke mate is ook zichtbaar bij woninginbraken waar 37,4% van de inbraken 's avonds in district West plaatsvindt. West is echter groter in oppervlakte en het aantal incidenten relatief tot de andere districten lager dan de verhouding bij straatroven.

Ruimtelijke verschillen in onderliggende tijdsintervallen Incidenten hebben een verschillende geografische verdeling onder verschillende onderliggende tijdsintervallen aan de huidige peilperiode van twee weken. Dit betekent dat zowel woninginbraken als straatroven op andere plekken gebeuren afhankelijk van het tijdsinterval binnen een peilperiode van twee weken en niet elke plek over de gehele tweeweekse peilperiode een gelijke kans op een incident heeft. De volgende gedetailleerdere resultaten met betrekking tot geografische verschillen zijn gevonden:

1. Woninginbraken vinden plaats in verschillende districten wanneer onderscheid wordt gemaakt in de dagdelen nacht, dag en avond. De meest afwijkende geografische verdeling van incidenten wordt waargenomen in district Centrum, waar het hoogtepunt 's nachts is, terwijl alle andere districten 's nachts de minste inbraken plaatsvinden. Uit de analyse op basis van wijken komen dezelfde resultaten al blijkt de centrumwijk IJ-tunnel niet mee te doen in het afwijkende gedrag van het district. Op basis van weekdagen kunnen er een geografisch verschil gevonden worden tussen de weekdagen maandag t/m vrijdag en de weekenddagen zaterdag en zondag. Tussen de werkdagen maandag t/m vrijdag en tussen de weekenddagen zaterdag en zondag onderling lijken geen grote geografische verschillen te zitten. De rol van de vrijdag hierbinnen is discutabel: vrijdag past niet volledig in de verdeling van de weekdagen maar ook niet bij de verdeling van de weekenddagen.

2. Straatroven vinden over de verschillende dagdelen nacht, dag en avond plaats in verschillende districten/wijken. De meest afwijkende verdeling van straatroven wordt waargenomen in district Centrum, waar het hoogtepunt 's nachts is, terwijl alle andere districten 's nachts de minste straatroven plaatsvinden. Het afwijkende gedrag van district Centrum lijkt zich niet te verhouden tot de wijken centrumwijken IJ-tunnel en Raampoort, maar zijn de wijken Konniginneweg en Pijp in district Zuid hier wel onderhevig aan. Op basis van weekdagen kan er een verschil gevonden worden

tussen de weekdagen maandag t/m vrijdag en de weekenddagen zaterdag en zondag. Tussen de werkdagen maandag t/m vrijdag en tussen de weekenddagen zaterdag en zondag onderling lijken geen grote geografische verschillen te zitten.

De geografische verschillen zijn gevonden en onderbouwd door technieken die gebruik maken van kruistabellen tussen twee categorische variabalen, waarvan één variabele de tijd indicateert en één variabele de ruimte. De grootste beperking in het gebruik van deze technieken zit in de afbakening van tijd en ruimte. Zowel tijd als ruimte wordt afgebakend op momenten die logisch zijn aan de hand van het rooster van operationele politiemedewerkers. Tijd is afgebakend op basis van de diensttijden en de ruimte is afgebakend op basis van de wijken en districten waarin de politieteams opereren. Aan de ene kant zijn de grenzen van diensttijden of politieteams niet random gekozen, maar aan de andere kant is er ook niet voorafgaand aan dit onderzoek onderzocht of deze grenzen toepasbaar zijn. Deze methode zorgt ervoor dat er geen andere tijdsindicatieve afbakening gevonden kan worden dan de tijdsgrenzen van de diensttijden. Ook verhouden geografische verschillen zich tot de afgebakende wijken en districten en kunnen niet ontpoppen tot vrije ruimtelijke vormen. Desondanks blijft de conclusie betrouwbaar: er zijn geografisch verschillen, alleen de wijze van detail is niet volledig uitgediept.

Voorspellen van woninginbraken op tijdsintervalniveau I Incidenten in een kleiner onderliggend tijdsinterval kunnen voorspeld worden door CAS. De werkwijze is in dat geval gelijk aan CAS, alleen voorspelt het model zich alleen op de incidenten die hebben plaatsgevonden in een specifiek tijdsinterval. Het CAS model kan op basis van afzonderlijke voorspellingen die gezamelijk de tweeweekse periode opvatten, de performance van het huidige model niet verbeteren maar wel evenaren. Dit betekent dat CAS kan worden vervangen of worden verrijkt met voorspellingen voor kleinere tijdsintervallen met een

Het huidige CAS model op basis van een tweeweekse peilperiode voorspelt 36,3% van de woninginbraken op basis van de *near hits performance* en 15,4% van de woninginbraken op basis van de *hits performance*.

1. *Weekdagmodel* - Het huidige CAS model dat een voorspelling maakt over twee weken voorspelt beter dan een samengesteld model waarbij voor iedere weekdag een afzonderlijke voorspelling wordt gemaakt. Daarnaast voorspelt het huidige model ook beter incidenten voor iedere weekdag afzonderlijk, dan een voorspellling gespecificeerd op de weekdag. Op basis van weekdagen wordt een hits performance gehaald van 11,1% en een near hits performance van 34,9%. De voorspellingen van de weekdagen wijzen gemiddeld 35,1% dezelfde high risk locaties aan.

2. *Dagdeelmodel* - Er kan geen verschil worden gevonden tussen de performance van het samengestelde model op basis van dagdelen en het huidige CAS model. Het dagdeel nacht wordt beter voorspelt dan het samengestelde model op basis van dagdeel en het dagdeel avond wordt beter voorspeld door het huidige model. Over het dagdeel dag wordt geen uitsluitsel gegeven. Op basis van weekdagen wordt een hits performance gehaald van 15,4% en een near hits performance van 34,9%. De voorspellingen van de dagdelen wijzen gemiddeld 22,9% dezelfde high risk locaties aan.

3. *Diensttijdmodel* - Het huidige CAS model dat een voorspelling maakt over twee weken voorspelt beter dan een samengesteld model waarbij voor iedere diensttijd een afzonderlijke voorspelling

wordt gemaakt. Er zijn 6 diensttijden waar tussen performance van het samengestelde model en het huidige model geen verschil gevonden kan worden. De overige 15 diensttijden worden beter voorspeld met het huidige CAS model. Op basis van diensttijden wordt een hits performance gehaald van 11,1% en een near hits performance van 36,2%. De voorspellingen van de dagdelen wijzen gemiddeld 46,2% dezelfde high risk locaties aan.

4. *Week- en weekendmodel* - Er kan geen verschil gevonden worden tussen de performance van het samengestelde model op basis van week- en weekenddagen en het huidige CAS model. Het samengestelde model kent een hits performance van 14,2% en een near hits performance van 38,9%. De afzonderlijke voorspellingen wijzen gemiddeld 34,6% dezelfde high risk locaties aan.

5. *Week/weekend en dagdeelmodel* - Het huidige CAS model dat een voorspelling maakt over twee weken voorspelt beter dan een samengestelde model waarbij voor iedere week-/weekend en dagdeel combinatie een afzonderlijke voorspelling wordt gemaakt. Op basis van de week/weekend en dagdeelopsplitsing wordt een hits performance gehaald van 12,6% en een near hits performance van 34,8%. De voorspellingen van de dagdelen wijzen gemiddeld 27,1% dezelfde high risk locaties aan.

6. *Analytische tweedeling* - Er kan geen verschil gevonden worden tussen de performance van het samengestelde model (opdeling tussen de tijdsintervallen nacht en vrijdag t/m zondag overdag en de tijdsintervallen avond en maandag t/m donderdag overdag) en het huidige CAS model. In veel gevallen lijkt het huidige CAS model iets beter, maar er is geen overweldigend verschil. Het samengestelde model kent een hits performance van 12,6% en een near hits performance van 34,8%. De afzonderlijke voorspellingen wijzen gemiddeld 27,1% dezelfde high risk locaties aan.

Het *dagdeel* en *week- en weekend model* presteren van alle samengestelde modellen het best. Zij weten beiden de performance van het huidige CAS model te evenaren. Het model op basis van de *analytische tweedeling* komt daarbij in de buurt maar is minder overtuigend.

Een belangrijk resultaat is dat bij het verkleinen van de tijdsintervallen het aantal te voorspellen incidenten afneemt. Deze afname in incidenten leidt tot minder verreikende incidenthistorie om nieuwe incidenten te voorspellen wat uiteindelijk wil leiden tot een slechtere performance. Er zijn duidelijke verschillen opgemerkt tussen de week- en weekenddagen en tussen de dagdelen, toch presteert het model dat beide incorporeert slechter dan de modellen die alleen naar dagdelen of alleen naar week- en weekenddagen kijken. Vermoedelijk ligt dit niet aan het feit dat de keuze voor deze tijdsintervallen slecht gekozen is, maar doordat het aantal te voorspellen incidenten en daarmee ook de incidenthistorie laag is. Het model is daarbij niet meer voldoende in staat de juiste patronen te extraheren.

Een samengesteld model zal daarbij alleen in staat zijn de CAS performance te verbeteren, wanneer het onderscheidt in de geografische verdeling van incidenten voor beide tijdsintervallen dermate groot is dat dit opweegt tegen de vermindering in incidenthistorie.

Voorspellen van woninginbraken op tijdsintervalniveau II Het huidige CAS model werkt door middel van een logistisch regressie model. Op basis van een neuraal netwerk, hotspotmodel en Bayes netwerk is geprobeerd voor de drie modellen die de performance van CAS weten te evenaren onderzocht of

deze beter presteren dan het model op basis van logistisch regressie. Het resultaat luidde: de modellen op basis van een neuraal netwerk, hotspotmodel en Bayes netwerk weten de performance van het logistische regressie model *niet* te evenaren, laat staan te verbeteren.

8.2 Aanbevelingen

Aanbeveling 1: De huidige CAS kaarten die worden gebaseerd op de tweewekelijkse modellen kunnen worden uitgebreid met tijdsindicatieve modellen op basis van dagdeel of week/weekendmodellen. Wanneer de huidige CAS kaarten worden uitgebreid met extra tijdsindicatieve modellen is de vorm waarin ook van belang. Wanneer de huidige kaarten aangeboden blijven en de nieuwe tijdsindicatieve modellen als uitbreiding worden aangeboden, kan dit tot verwarring leiden. De huidige tijdsindicatieve kaarten die locaties aanwijzen op basis van de tweewekelijkse voorspelling zullen in tegenstrijd zijn met de nieuw ontwikkelde kaarten. Hierover zal een duidelijk intepetatieverschil moeten worden uitgelegd. De tijdsindicatieve kaarten van het huidige model baseren zich daarbij op de intensiteit van incidenten binnen het gebied waar over het algemeen in twee weken de meeste kans op een incident is. De nieuwe tijdsindicatieve momenten vertellen ongeacht intensiteit waar voor een bepaald tijdsinterval een verhoogde kans op een incident is. Daarnaast is door operationele politieteams gemeld dat niet meer dan 3% van de locaties mag worden uitgelicht omdat de politie op dit moment niet in staat is met het flexteam meer gebieden te patrouileren. Wanneer onderscheidt wordt gemaakt in het aanbieden van deze dagdeel of week/weekenddagmodellen moeten teams wel in staat zijn deze hoeveelheid locaties aan te kunnen ongeacht dat op ieder tijdstip één kaart van kracht is met 3% van de locaties uitgelicht. Over het algemeen zal eerst de vraag naar tijdsindicatieve modellen op basis van week-/weekend of dagdeelmodellen onderzocht moeten worden vanuit de operationele kant van de organisatie.

Aanbeveling 2: Het is aan te bevelen verder onderzoek te verrichten naar tijdsindicatieve modellen om incidenten te voorspellen. In dit onderzoek is bewezen dat er ruimtelijke verschillen zijn in de tijdsintervallen onderliggend aan de tweewekelijkse peilperioden, maar is er op basis van deze kennis nog weinig verder onderzoek gedaan. Ook zouden andere technieken om incidenten te voorspellen een optie zijn zoals het nagaan van bijvoorbeeld near repeat modellen of modellen op basis van tijdruimtelijke patronen.

Bibliografie

[1] M.A. Andresen. Testing for similarity in area-based spatial patterns: A nonparametric monte carlo approach. *Applied Geography*, 29:333–345, 2009.

[2] A. Baddeley. *Spatial Point Processes and their Applications*. Online.

[3] C. Block. Stac: hot-spot areas: A statistical tool for law enforcement decisions. crime analysis through computer mapping. *Police Executive Research Forum*, pages 15–32, 1995.

[4] C.R. Block, S.L. Knight, W.G. Gould, and J.D. Coldren. *Is crime predictable? A test of methodology for forecasting criminal offenses*. Illinois Criminal Justice Information Authority, Chicago.

[5] P.J. Brantingham and P.L. Brantingham. *Patterns in crime*. New york: Macmillan.

[6] J.M. Caplan and L.W. Kennedy. *Risk Terrain Modeling Compendium for Crime Analysis*. Newark, N.J.: Rutgers Center on Public Security.

[7] L.E. Cohen and M. Felsen. Social change and crime rate trends: A routineactivity approach. *American Sociological Review*, 44:588–607, 1979.

[8] J.J. Corcoran, I.D. Wilson, and J.A. Ware. Predicting the geo-temporal variations of crime and disorder. *International Journal of Forecasting*, 19.

[9] D.B. Cornish and R.V. Clarke. Understanding crime displacement: An application of rational choice theory. *Criminology 25*.

[10] M.B. Gordon. A random walk in the literature on criminality: A partial and critical view on some statistical analyses and modelling approaches. *European Journal of Applied Mathematics*, 21.

[11] H. Mietus, S. ter Woerds, and D. Willems. Waar en wanneer het ertoe doet: bepalen en duiden van hotspot- en hottimesinformatie binnen de politie amsterdam. 2012.

[12] G.O. Mohler, M.B. Short, P.J. Brantingham, F.P. Schoenberg, and G.E. Tita. Self-exiting point process modeling of crime. *Journal of the American Statistical Association*, 106(493).

[13] G.C. Oatley and B.W. Ewart. Crimes analysis software: 'pins in maps', clustering and bayes net prediction. *Expert systems with Applications*, 25.

[14] W.L. Perry, B. McInnis, C.C. Price, S.C. Smith, and J.S. Hollywood. *Predictive Policing - The Role of Crime Forecasting in Law Enforcement Operations*. RAND Corporation.

[15] A. Quetelet. *Essai de Physique Sociale*. Bachelier, Parijs.

[16] J. Rubin. Stopping crime before it starts. *Los Angeles Times*.

[17] T.E. Smith. *Notebook for spatial data analysis*. Online.

[18] M. Townsley, R. Homel, and J. Chaseling. Repeat bulgary victimistation: Spatial and temporal patterns. *Australian and New Zeeland Journal of Criminology*, 33(1).

[19] B van Dijk, C van den Handel, and P Versteegh. Hotspotaanpak in vier stappen. 2011.

[20] X. Wang and D.E. Brown. The spatio-temporal generalized addictive model for criminal incidents. *Proceedings of the IEEE International Conference on Intelligence and Security Informatics: 9-12 july 2011, Beijing, China*.

[21] X. Wang and D.E. Brown. The spatio-temporal modeling for criminal incidents. *Security Informatics 1:2*.

[22] E.W. Weisstein. Chi-squared test. *MathWorld–A Wolfram Web Resource*.

[23] M.E. Wolfgang, R.M. Figlio, and T. Sellin. *Delinquency in a birth cohort*. Chicago: University of Chicago Press.

Bijlage A

Overzicht variabelen

Variabele	Omschrijving
PERIODE_ID	Peilperiode (zie paragraaf 3.2)
JAAR	Jaar peilmoment (jaar op eerste dag van peilperiode)
STARTDATE	Datum peilmoment (eerste dag van peilperiode)
DISTRICT	District binnen Amsterdam
WIJKTEAM	Wijk binnen Amsterdam
INC_SUM_WIB	Aantal woninginbraken in vakje plaatsgevonden
TARGET_WIB	Flag: minstens één woninginbraak in vakje op peilmoment?
INC_SUM_STR	Aantal woninginbraken in vakje plaatsgevonden
TARGET_STR	Flag: minstens één straatroof in vakje op peilmoment?
AANTAL_INWONERS	Aantal inwoners in postcodegebied waarin het vakje ligt (in 5 kwantielen)
AANTAL_MANNEN	Aantal mannen in postcodegebied waarin het vakje ligt (in 5 kwantielen)
AANTAL_VROUWEN	Aantal vrouwen in postcodegebied waarin het vakje ligt (in 5 kwantielen)
AANTAL_PARTHH	Aantal part. huishoudens in postcodeg. waarin het vakje ligt (in 5 kwan.)
GEMHHGROOTTE	Gem. huishoudensgrootte in postcodeg. waarin het vakje ligt (in 5 kwan.)
PERC_00_14	Percentage 0-14 jaar van postcodegebied waarin het vakje ligt
PERC_15_24	Percentage 15-24 jaar van postcodegebied waarin het vakje ligt
PERC_25_44	Percentage 25-44 jaar van postcodegebied waarin het vakje ligt
PERC_45_64	Percentage 45-64 jaar van postcodegebied waarin het vakje ligt
PERC_65_75	Percentage 65-74 jaar van postcodegebied waarin het vakje ligt
PERC_75_OUDER	Percentage 75 jaar en ouder van postcodeg. waarin het vakje ligt
NIETWESTERSALLECHTOON	Perc. nietwes. allochtonen in postcodeg. van vakje (in 5 kwan.)

Vervolg op de volgende pagina.

Variabele	Omschrijving
EENPERSOONSHH	Perc. éénpersoons-huish. in postcodeg. van vakje (in 5 kwan.)
EENOUDERHH	Perc. éénouderhuish. in postcodeg. van vakje (in 5 kwan.)
MEERPZONDERKINDEREN	Perc. meerpersoonshuish. z. kinderen in postcodeg. van vakje (in 5 kwan.)
TWEEOUDERHH	Perc. tweeouderhuish. in postcodeg. van vakje (in 5 kwan.)
WONINGVRD	Woningvoorraad in postcodeg. waarin het vakje ligt (in 5 kwan.)
GEMWONINGWAARDE	Gem. woningwaarde in postcodeg. van vakje (in 5 kwan.)
LAAGINKOMEN	Perc. lage inkomens in postcodegebied van vakje (in 10 kwan.)
HOOGINKOMEN	Perc. hoge inkomens in postcodegebied van vakje (in 10 kwan.)
INKOMENSONTVANGERS	Aantal inkomensontvangers in postcodeg. van vakje (in 10 kwan.)
UITKERINGSONTVANGERS	Perc. uitkeringsontvangers in postcodeg. van vakje (in 10 kwan.)
ZELFSTANDIGEN	Perc. zelfstandigen in postcodeg. van vakje (in 10 kwan.)
FISCAALMAANDINKOMEN	Gem. fiscaal maandinkomen in postcodeg. van vakje (in 10 kwan.)
CAFE_BAR	Aantal café's/bars in vakje
RESTAURANT	Aantal restaurants in vakje
ONDERWIJSINSTELLING	Aantal onderwijsinstellingen in vakje
VERENIGING	Aantal verenigingen in vakje
SNACKBAR	Aantal snackbars in vakje
HOTEL_MOTEL_BOTEL	Aantal hotels/motels/botels in vakje
OVERHEIDSINSTELLING	Aantal overheidsinstellingen in vakje
BANK	Aantal banken in vakje
SUPERMARKT	Aantal supermarkten in vakje
KOFFIESHOP	Aantal koffieshops in vakje
SEXSHOP_CLUB_SHOW	Aantal seksshops/-clubs/-shows in vakje
SLIJTERIJ	Aantal slijterijen in vakje
BENZINESTATION	Aantal benzinestations in vakje
DISCO_DANCING_NACHTCLUB	Aantal discotheken/dancings/nachtclubs in vakje
JONGERENCENTRUM	Aantal jongerencentra in vakje
ZIEKENHUIS	Aantal ziekenhuizen in vakje
BEJAARDENHUIS	Aantal bejaardenhuizen in vakje
GOK_SPEELAUTOMATENHAL	Aantal gok-/speelautomaten in vakje
VVV_TOERISTEN_INFORMATIE	Aantal VVV's in vakje
WINKEL	Aantal winkels in vakje
MIN_DIST_WIB	Afst. centroide van vakje tot adres dichtsbijzijndste bekende inbreker
MIN_DIST_SRF	Afst. centroide van vakje tot adres dichtsbijzijndste bekende straatrover
SUBJECTS_CLOSE_WIB	# bekende inbrekers in straal van 1km rond centroide van het vakje

Vervolg op de volgende pagina.

Variabele	Omschrijving
SUBJECTS_CLOSE_SRF	# bekende straatrovers in de straal van 1km rond centroide van het vakje
BINNEN_WERKGEBIED_WIB	# bekende inbrekers waar het vakje in het werkgebied ligt
BINNEN_WERKGEBIED_SRF	# bekende straatrovers waar het vakje in het werkgebied ligt
2W1_VAK_WIB	Aantal woninginbraken in vakje in 2 weken voorafgaand start peilper.
2W2_VAK_WIB	Aantal woninginbraken in vakje in 2 weken voorafgaand start peilper. minus 2 weken
2W3_VAK_WIB	Aantal woninginbraken in vakje in 2 weken voorafgaand start peilper. minus 4 weken
2W4_VAK_WIB	Aantal woninginbraken in vakje in 2 weken voorafgaand start peilper. minus 6 weken
4W1_VAK_WIB	Aantal woninginbraken in vakje in 4 weken voorafgaand start peilper.
4W2_VAK_WIB	Aantal woninginbraken in vakje in 4 weken voorafgaand start peilper. minus 4 weken
2W3_VAK_WIB	Aantal woninginbraken in vakje in 4 weken voorafgaand start peilper. minus 8 weken
2W4_VAK_WIB	Aantal woninginbraken in vakje in 4 weken voorafgaand start peilper. minus 12 weken
26W1_VAK_WIB	Aantal woninginbraken in vakje in 26 weken voorafgaand start peilper.
TREND_2W_VAK_WIB	Hellingscoef. regressielijn woninginb. als functie van tijd in vakje (obv 4*2 weken data)
TREND_4W_VAK_WIB	Hellingscoef. regressielijn woninginb. als functie van tijd in vakje (obv 4*4 weken data)
2W1_BUURT_WIB	Aantal woninginbraken in aangrenzende vakjes in 2 weken voorafgaand start peilper.
2W2_BUURT_WIB	Aantal woninginbraken in aangrenzende vakjes in 2 weken voorafgaand start peilper. minus 2 weken
2W3_BUURT_WIB	Aantal woninginbraken in aangrenzende vakjes in 2 weken voorafgaand start peilper. minus 4 weken
2W4_BUURT_WIB	Aantal woninginbraken in aangrenzende vakjes in 2 weken voorafgaand start peilper. minus 6 weken
4W1_BUURT_WIB	Aantal woninginbraken in aangrenzende vakjes in 4 weken voorafgaand start peilper.
4W2_BUURT_WIB	Aantal woninginbraken in aangrenzende vakjes in 4 weken voorafgaand start peilper. minus 4 weken

Vervolg op de volgende pagina.

Variabele	Omschrijving
2W3_BUURT_WIB	Aantal woninginbraken in aangrenzende vakjes in 4 weken voorafgaand start peilper. minus 8 weken
2W4_BUURT_WIB	Aantal woninginbraken in aangrenzende vakjes in 4 weken voorafgaand start peilper. minus 12 weken
26W1_BUURT_WIB	Aantal woninginbraken in aangrenzende vakjes in 26 weken voorafgaand start peilper.
TREND_2W_BUURT_WIB	Hellingscoef. regressielijn woninginb. als functie van tijd in vakje (obv 4*2 weken data)
TREND_4W_BUURT_WIB	Hellingscoef. regressielijn woninginb. als functie van tijd in vakje (obv 4*4 weken data)
TSLI_WIB	Tijd in maanden sinds laatste woninginbraak in vakje
2W1_VAK_SRF	Aantal straatroven in vakje in 2 weken voorafgaand start peilper.
2W2_VAK_SRF	Aantal straatroven in vakje in 2 weken voorafgaand start peilper. minus 2 weken
2W3_VAK_SRF	Aantal straatroven in vakje in 2 weken voorafgaand start peilper. minus 4 weken
2W4_VAK_SRF	Aantal straatroven in vakje in 2 weken voorafgaand start peilper. minus 6 weken
4W1_VAK_SRF	Aantal straatroven in vakje in 4 weken voorafgaand start peilper.
4W2_VAK_SRF	Aantal straatroven in vakje in 4 weken voorafgaand start peilper. minus 4 weken
2W3_VAK_SRF	Aantal straatroven in vakje in 4 weken voorafgaand start peilper. minus 8 weken
2W4_VAK_SRF	Aantal straatroven in vakje in 4 weken voorafgaand start peilper. minus 12 weken
26W1_VAK_SRF	Aantal straatroven in vakje in 26 weken voorafgaand start peilper.
TREND_2W_VAK_SRF	Hellingscoef. regressielijn straatroof als functie van tijd in vakje (obv 4*2 weken data)
TREND_4W_VAK_SRF	Hellingscoef. regressielijn straatroof als functie van tijd in vakje (obv 4*4 weken data)
2W1_BUURT_SRF	Aantal straatroven in aangrenzende vakjes in 2 weken voorafgaand start peilper.
2W2_BUURT_SRF	Aantal straatroven in aangrenzende vakjes in 2 weken voorafgaand start peilper. minus 2 weken

Vervolg op de volgende pagina.

Variabele	Omschrijving
2W3_BUURT_SRF	Aantal straatroven in aangrenzende vakjes in 2 weken voorafgaand start peilper. minus 4 weken
2W4_BUURT_SRF	Aantal straatroven in aangrenzende vakjes in 2 weken voorafgaand start peilper. minus 6 weken
4W1_BUURT_SRF	Aantal straatroven in aangrenzende vakjes in 4 weken voorafgaand start peilper.
4W2_BUURT_SRF	Aantal straatroven in aangrenzende vakjes in 4 weken voorafgaand start peilper. minus 4 weken
2W3_BUURT_SRF	Aantal straatroven in aangrenzende vakjes in 4 weken voorafgaand start peilper. minus 8 weken
2W4_BUURT_SRF	Aantal straatroven in aangrenzende vakjes in 4 weken voorafgaand start peilper. minus 12 weken
26W1_BUURT_SRF	Aantal straatroven in aangrenzende vakjes in 26 weken voorafgaand start peilper.
TREND_2W_BUURT_SRF	Hellingscoef. regressielijn straatroof als functie van tijd in vakje (obv 4*2 weken data)
TREND_4W_BUURT_SRF	Hellingscoef. regressielijn straatroof als functie van tijd in vakje (obv 4*4 weken data)
TSLI_SRF	Tijd in maanden sinds laatste straatroof in vakje

Einde bijlage.