Eindrapport Voorspelling Infectieziekten

Het eerder signaleren en voorspellen van infectieziekten met behulp van Google Trend data en andere open source databronnen.

Dymphie Mioch, MSc (GGD West-Brabant)
Dr. Leonard Vanbrabant (GGD West-Brabant)
Drs. Stijn Raven (GGD West-Brabant)
Drs. Paulien Tolsma (GGD Brabant-Zuidoost)
Drs. Sandra van Dam (GGD Hart voor Brabant)

11 april 2019

JADS Jheronimus Academy of Data Science
GGD Brabant-Zuidoost Hart voor Brabant West-Brabant

Inhoud

Achtergrond

Bij het merendeel van de infectieziekten is het tijdig opmerken van een infectieziekte(uitbraak) van cruciaal belang om effectieve preventieve maatregelen te kunnen nemen. In een vroeg stadium van een infectieziekte bij de index, hebben maatregelen als vaccinatie (o.a. bij bof, mazelen en kinkhoest), profylaxe (o.a. meningokokken, invasieve groep A streptokokken) en algemene hygiënemaatregelen zin om verdere verspreiding naar de omgeving tegen te gaan. Welke tijdsperiode staat voor 'tijdig' en 'vroeg stadium' is per infectieziekte verschillend. Uit recent Nederlands onderzoek blijkt dat voor de infectieziekte kinkhoest er vaak (te) veel tijd zit tussen diagnose en het melden van kinkhoestgevallen bij de GGD. Hier kan één tot enkele weken tussen zitten, waardoor het in veel gevallen niet meer mogelijk is om adequate preventieve maatregelen te nemen (Heil et al., 2017). Door het vroegtijdig detecteren van een toename van kinkhoest in een bepaalde regio, kan gerichte advisering aan de eerste lijn door de GGD mogelijke verspreiding naar kwetsbaren voorkomen. Ook het tijdig detecteren van een uitbraak van bijvoorbeeld scabiës en bof kan mogelijk bijdragen aan betere uitbraakbestrijding van deze ziektebeelden. Enerzijds zijn dit ziektebeelden waarbij de verspreiding en/of ziektelast gereduceerd kan worden door profylaxe/behandeling en vaccinatie, anderzijds belemmert laattijdige opsporing deze maatregelen (Opstelten et al., 2012). De oorzaken van diagnostische vertraging voor scabiës liggen onder andere in het missen van de diagnose en het hebben van aspecifieke klachten (Tsjoe et., 2006). Vroegtijdige signalering zou de gevolgen van het missen van deze diagnose kunnen reduceren.

In mei 2017 is een verkennend onderzoek uitgevoerd door de drie Brabantse GGD'en en de Jheronymus Academy of Data Science (JADS) waarbij kinkhoest-registratiegegevens van de GGD en Google Trend data (zoektermen kinkhoest en koorts) zijn gebruikt om een model te ontwikkelen voor het voorspellen van een uitbraak van kinkhoest. Uit dit verkennende onderzoek kwam naar voren dat zowel de Google Trends data als de registraties van de kinkhoestmeldingen informatie aan elkaar geven. Met andere woorden, het aantal kinkhoestmeldingen heeft een invloed op hoe vaak er gezocht wordt op termen zoals `kinkhoest' en `koorts', en dit zoekgedrag op Google heeft weer invloed op het aantal kinkhoestmeldingen; waarschijnlijk door de grotere media-aandacht. Uit nader onderzoek blijkt bovendien, dat de informatie van het aantal kinkhoestmeldingen op de Google Trend data groter is dan andersom. Kortom, personen gaan vooral na een kinkhoestuitbraak zoeken op Google. Om het model te verbeteren zijn meer relevante databronnen nodig. In dit huidige project zijn verscheidene studies uitgevoerd om het model verder te ontwikkelen om een uitbraak vroegtijdig te voorspellen. Daarnaast is ook gekeken of het model vertaald kan worden naar meldingsplichtige ziekten die minder prevalent zijn, maar waarbij vroegdetectie invloed heeft op de grootte van een uitbraak (e.g., bof en scabiës).

VRAAGSTELLING

De kernvraag van dit onderzoek is:

- Zijn infectieziekten (bijvoorbeeld kinkhoest, scabiës en bof) eerder te signaleren en te voorspellen door middel van het combineren van verschillende databronnen, zodat eerder dan nu preventieve maatregelen te nemen zijn?

Deelvragen bij deze kernvraag zijn:

- Is het prototypemodel voor kinkhoest dat ontwikkeld is voor Noord-Brabant op basis van GGD data en Google Trends te valideren voor andere regio's?

- Kan de voorspellingskracht van het kinkhoestmodel verbeterd worden met aanvullende databronnen?
- Is het mogelijk om de ziekten bof, en scabiës te voorspellen aan de hand van het combineren van GGD data met andere databronnen? En is hiervoor een prototype model te ontwikkelen? Wat zijn ervaringen bij implementatie van het voorspellingsmodel voor kinkhoest en (eventueel) andere infectieziekten in de dagelijkse praktijk van teams Infectieziektebestrijding (IZB) van de GGD'en in Noord-Brabant. Welke aanbevelingen zijn er te geven over implementatie?

Methoden

In het onderzoek zijn verschillende databronnen en verschillende analysetechnieken gebruikt. Het onderzoek bevat 3 onderzoekslijnen:

- Ontwikkeling voorspellend model kinkhoest en de bof in de regio Noord-Brabant
- Ontwikkeling voorspellend model kinkhoest in de regio Zuid-Limburg
- Ontwikkeling voorspellend model scabiës in de regio Noord-Brabant

DATA KINKHOEST EN BOF NOORD-BRABANT

De kinkhoestdata bevatten het aantal gemelde kinkhoest gevallen (N=18.799) voor de GGD-regio's in Noord-Brabant tussen 2003-2017. De data bevatten zowel de datum (inschatting door de patiënt) van het ontstaan (*date of onset)* van kinkhoest als de datum waarop de melding bij de GGD is gedaan (*date of notification)* dat de patiënt inderdaad kinkhoest heeft. De *date of onset* variabele bevat veel ontbrekende waarden. Daarnaast is de inschatting van patiënten wanneer de klachten begonnen onbetrouwbaar. Daarom is voor de analyses de *date of notification* gebruikt. De data van Google Trends is voor iedere provincie beschikbaar vanaf januari 2014 en worden per maand weergegeven. In dit onderzoek is gezocht op de termen 'kinkhoest' en 'koorts'. Verder is gebruik gemaakt van data afkomstig van de Rijksuniversiteit Groningen met informatie over gemeenten in Nederland. Het databestand van de bof bestaat uit het aantal patiënten (N=279) waarbij de bof is geconstateerd en gemeld bij de GGD in de regio Noord-Brabant tussen 2009-2017. Naast Google Trends als externe databron zijn ook de meteorologische data (i.e., maandelijkse gemiddelde temperatuur in Noord-Brabant) van het KNMI gebruikt.

ANALYSES KINKHOEST NOORD-BRABANT

Het ontwikkelen van een model voor het voorspellen van een uitbraak van kinkhoest bestaat uit een aantal fasen. In de eerste fase, is een model ontwikkeld door middel van verschillende *machine learning* technieken op basis van de *date of notification* in combinatie met de Google Trend data. In de tweede fase, is naar zowel de *date of notification* als naar de locatie waar een kinkhoest geval is vastgesteld gekeken. Op basis hiervan is een neuraal netwerk ontwikkeld, dat zowel de tijd en de locatie van de kinkhoest gevallen in het model meeneemt als de Google Trend data. Verder is in deze fase gekeken of extra databronnen, namelijk data over het weer en data over de indeling van de regio in kleinere delen, de prestaties van het model zouden kunnen verbeteren. In de derde fase, is onderzocht of het model dat ontwikkeld is om een kinkhoest uitbraak te voorspellen ook toepasbaar

is op de bof. Ook is gekeken naar hoe de data in de toekomende jaren automatisch kunnen worden voorbewerkt zodat deze eenvoudig in het model kunnen worden opgenomen.

DATA KINKHOEST ZUID-LIMBURG

De dataset in de regio Zuid-Limburg bestaat uit personen waarbij een labaanvraag heeft plaatsgevonden om op kinkhoest te testen tussen januari 2007 en december 2013. Deze dataset bevat in tegenstelling tot de dataset van Noord-Brabant waarin alleen de positieve cases staan, zowel de positieve als negatieve cases voor kinkhoest. Daarmee geeft het ook informatie over testgedrag per regio.

ANALYSES KINKHOEST ZUID-LIMBURG

In de eerste onderzoekslijn op basis van de Noord-Brabant data is het ontwikkelde model gebaseerd op *machine learning* technieken. In deze onderzoekslijn wordt gebruik gemaakt van netwerkanalyse technieken om de dynamische relatie te vinden tussen cases zonder kinkhoest, de cases met kinkhoest en de locatie (postcode).

Eerst is er gekeken naar relaties tussen individuen met een positief testresultaat. Om deze relatie te vinden is de aanname gedaan dat één week voor en twee weken na de datum van de labuitslag de periode is waarin de besmetting van de andere persoon heeft plaatsgevonden. Verder is de aanname gedaan dat de besmetting alleen kan plaatsvinden als mensen zich op minder dan acht kilometer van elkaar bevinden. Nadat de relatie tussen de personen is gevonden, wordt de *influence rate* (IR) berekend tussen patiënt x en patiënt y door de volgende formule, waarbij AVG (average) staat voor het gemiddelde. De IR *is* het gewicht van het effect tussen twee patiënten:

$$IR(x,y) = \frac{AVG(Density\ of\ city\ for\ patient\ x,\quad Density\ of\ city\ for\ patient\ y)}{Distance(patient\ x, patient\ y)}$$

Om de verspreiding van kinkhoest te visualiseren wordt één maand als tijdsvenster gebruikt. Dus voor elke maand is er een apart grafisch netwerk ontwikkeld. Om de grafische netwerken met elkaar te kunnen vergelijken is de indicator *Weighted Graph Density* (WGD) aangemaakt, waarbij de IR wordt gebruikt als gewicht voor de lijnen in het grafische netwerk. De WGD voor elk netwerk wordt gedefinieerd door de volgende formule, waarbij V_i het totaal aantal mogelijke *edges* is tussen alle *nodes* in maand *i*. Een node is gedefinieerd als de postcode van een positief geteste patiënt. Twee *nodes* zijn met elkaar verbonden door middel van een *edge* als de IR niet gelijk is aan nul.

$$WGD_i = \frac{sum(IR\ for\ each\ node)}{V_i * (V_i - 1)} \qquad i: index\ of\ the\ month\ in\ the\ whole\ dataset$$

Gebaseerd op de WGD is er een drempelwaarde bepaald om de grafische netwerken waarbij sprake is van een kinkhoestuitbraak te onderscheiden. Deze drempelwaarde *L* wordt als volgt berekend:

$$L = AVG(WGD) - \frac{std(WGD)}{2}$$

waarbij *std* staat voor de standaarddeviatie. Als de WGD van een grafisch netwerk kleiner is dan de drempelwaarde, spreken we van een uitbraak in dat netwerk.

Het detecteren van een uitbraak bij de laatste stap gebeurt via een lijst van uitbraak netwerken. Het netwerk voor het volgende tijdsvenster wordt vergeleken met alle voorgaande uitbraaknetwerken op basis van hun netwerkgelijkenis. Als het nieuwe grafische netwerk gelijk is aan één van de uitbraak netwerken, kunnen we zeggen dat er een nieuwe uitbraak komt. Daarnaast wordt de waarschijnlijkheid op het vóórkomen van kinkhoest niet alleen op GGD regio maar ook van elke postcode in de toekomst voorspeld.

DATA SCABIËS NOORD-BRABANT

Voor de laatste onderzoekslijn, waarbij een model is ontwikkeld voor het voorspellen van scabiës, waren verschillende databronnen beschikbaar. Ten eerste, de GGD beschikt over een dataset met gemelde gevallen van scabiës (N=689). Deze dataset geeft echter geen compleet beeld van het aantal scabiës gevallen in Noord-Brabant, omdat scabiës enkel een meldingsplichtige infectieziekte is als artikel 26 melding in instellingen. Individuele scabiës meldingen buiten een kwetsbare instelling zijn niet meldingsplichtig. Ten tweede, is een dataset gebruikt met informatie over de verstrekking van medicijnen, specifiek voor de behandeling van scabiës, in apotheken. Deze data zijn afkomstig van Stichting Farmaceutische Kengetallen (SFK). Het aantal verstrekkingen wordt per maand weergegeven voor de periode 2007-2017 en de data zijn opgesplitst in vier regio's namelijk, Noordoost-Brabant, West-Brabant, Midden Brabant en Zuidoost-Brabant. Voor de behandeling van scabiës bestaan drie soorten medicijnen: Permetrine, Ivermectine en Benzylbenzoaat. Ten slotte, zijn Google Trend data gebruikt, waarbij is gezocht op de termen 'schurft' en 'scabiës'.

ANALYSES SCABIËS NOORD-BRABANT

Voor deze onderzoekslijn zijn verschillende data analyse technieken gebruikt. Eerst, is gekeken of er een patroon te herkennen is in de seizoenen en of er trends over maanden/jaren heen naar voren komen. Bovendien is gekeken naar de geografische spreiding van het aantal scabiës gevallen. Ook is er gekeken hoe de verdeling is, rekening houdend met de leeftijd, om zo kwetsbare groepen te kunnen identificeren. Ten tweede, is onderzocht of er mogelijk relaties bestaan tussen het aantal cases en andere variabelen door middel van Pearson *correlaties* en *Kernel Density Estimation*. Als laatste is onderzocht of een scabiësuitbraak voorspeld kan worden. Hiervoor is een algoritme ontwikkeld dat een bepaalde drempelwaarde definieert voor de Google Trends variabele: als de Google Trend variabele een waarde laat zien hoger dan de drempelwaarde, dan is dat een indicatie voor een bepaalde waarschijnlijkheid dat er een scabiës uitbraak zal komen.

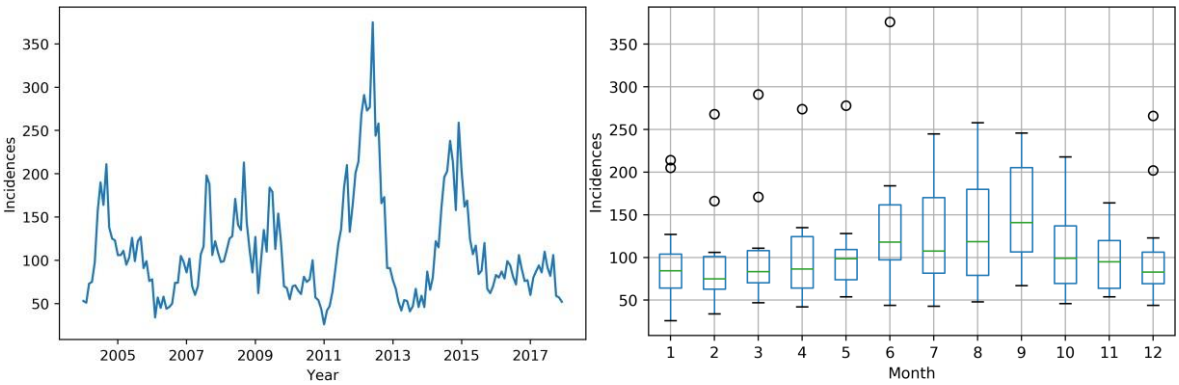
Resultaten

In deze sectie zullen de resultaten uit de hierboven beschreven onderzoekslijnen apart besproken worden.

RESULTATEN KINKHOEST REGIO NOORD-BRABANT

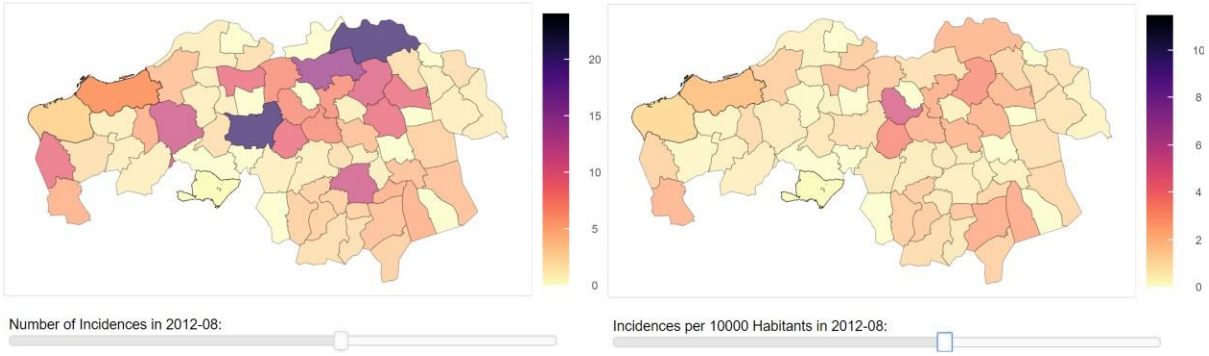
Beschrijvende analyses

Uit de beschrijvende analyses blijkt dat kinkhoest vaker voorkomt bij vrouwen (55,9%) dan bij mannen (44,1%) en dat mensen tussen de tien en twintig jaar oud de meeste kans hebben om geïnfecteerd te raken met kinkhoest. Ook is te zien dat het aantal kinderen jonger dan één jaar oud heel laag was. Dit is positief omdat kinkhoest juist voor deze leeftijdsgroep ernstige consequenties kan hebben (i.e. ziekenhuisopname i.v.m. dyspnoe en hoestaanvallen en incidenteel overlijden). Daarnaast is ook gekeken naar de factor 'tijd'. In Figuur 1 (linker) wordt het aantal kinkhoest gevallen in Noord-Brabant voor de periode 2004-2018 grafisch weergegeven. In Figuur 1 (rechts) zijn de data over de jaren heen geaggregeerd per maand. Op deze manier kan in kaart gebracht worden of kinkhoest een seizoensgebonden infectieziekte is. De meest ernstige uitbraak van kinkhoest vond plaats tussen 2011 en 2013 en bereikte een piek in juni 2012, waarbij 375 cases werden gerapporteerd aan de GGD'en in Noord-Brabant. Hoewel het aantal cases in de zomermaanden hoger is, kunnen we concluderen dat het seizoen geen sterke invloed heeft op het krijgen van kinkhoest.



Figuur 1.

Naast de factor 'tijd' is er ook gekeken naar de 'locatie' van de patiënt.

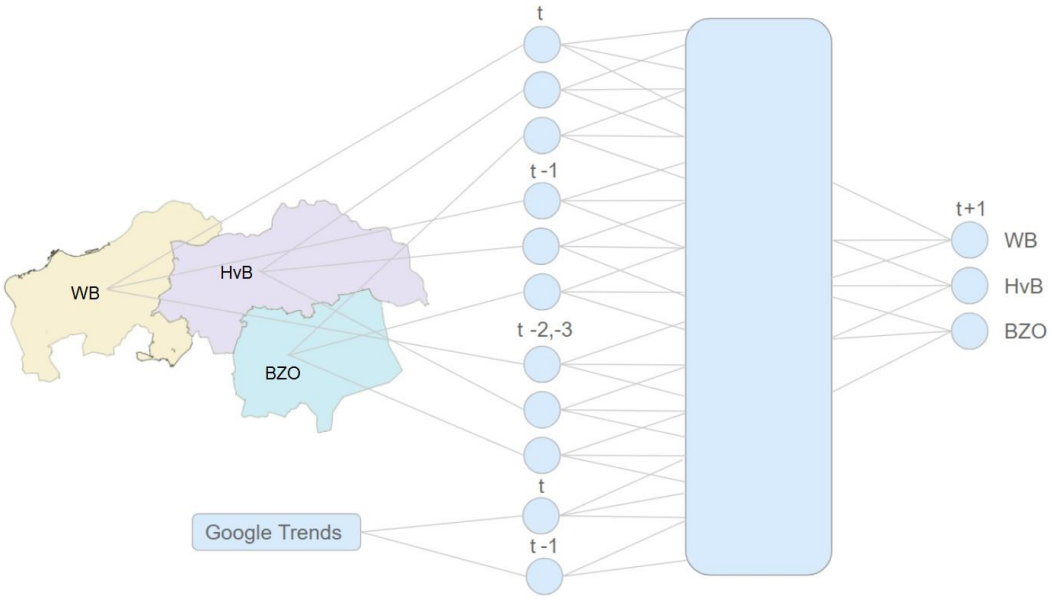


Figuur 2.

In Figuur 2 (links), komt de kleur van de gemeente overeen met het aantal patiënten in die gemeente en het kaartje in Figuur 2 (rechts) laat het aantal patiënten zien per 10.000 inwoners. Via de '*time slider*' aan de onderkant van het kaartje kan eenvoudig de periode worden aangepast om inzichtelijk te maken hoe de ziekte zich verspreid van de ene regio naar de andere regio. Op het linker kaartje wordt inzichtelijk gemaakt dat hoe meer inwoners de gemeente telt hoe donkerder de gemeenten kleuren en dus hoe meer gevallen van kinkhoest daar voorkomen. Op het kaartje rechts is deze verkleuring niet te zien als we de time slider verschuiven. Naar ratio blijven de aantallen dus meer gelijk tussen grote en kleine gemeenten. Een opvallende bevinding is dat vóórdat er daadwerkelijk sprake is van een kinkhoestuitbraak, en er dus een stijging is van het aantal patiënten, dat de stijging het sterkst is in de grotere gemeenten in vergelijking met kleinere gemeenten. Verder is het opvallend dat er in sommige gemeenten geen enkele casus wordt gemeld tijdens een uitbraak en dat in een aantal gemeenten het aantal cases erg hoog is in een periode dat er geen verheffing van kinkhoest plaatsvond in de regio.

Neuraal netwerk

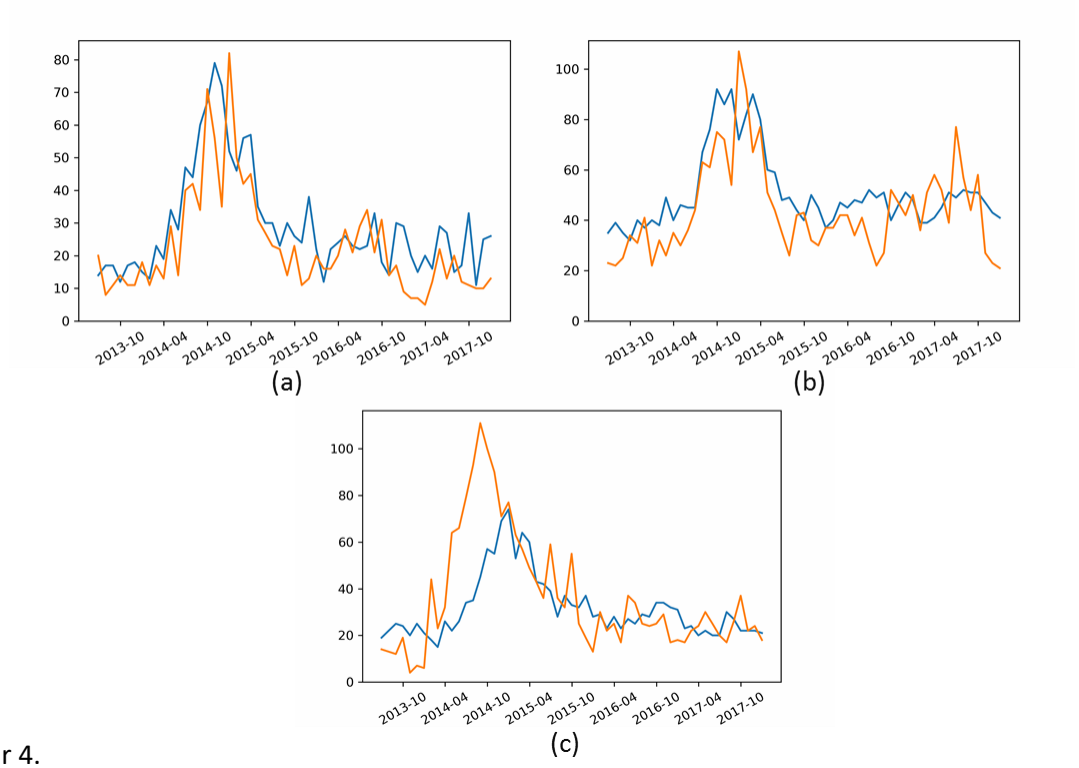
Als eerste benadering om een uitbraak van kinkhoest te voorspellen is gekozen voor *time series analysis.* Bij deze analyse wordt het aantal kinkhoest gevallen voorspeld aan de hand van het aantal huidige kinkhoest gevallen en in de afgelopen maanden, zonder rekening te houden met de locatie van de patiënt. Voor deze analyse is gebruik gemaakt van ARIMA modellen. Uit de resultaten bleek dat dit model niet goed in staat is een uitbraak van kinkhoest te voorspellen. De volgende stap was de ontwikkeling van een neuraal netwerkmodel *(NN),* waarbij naast de factor 'tijd' ook rekening werd gehouden met de 'locatie' van de patiënt.



Figuur 3.

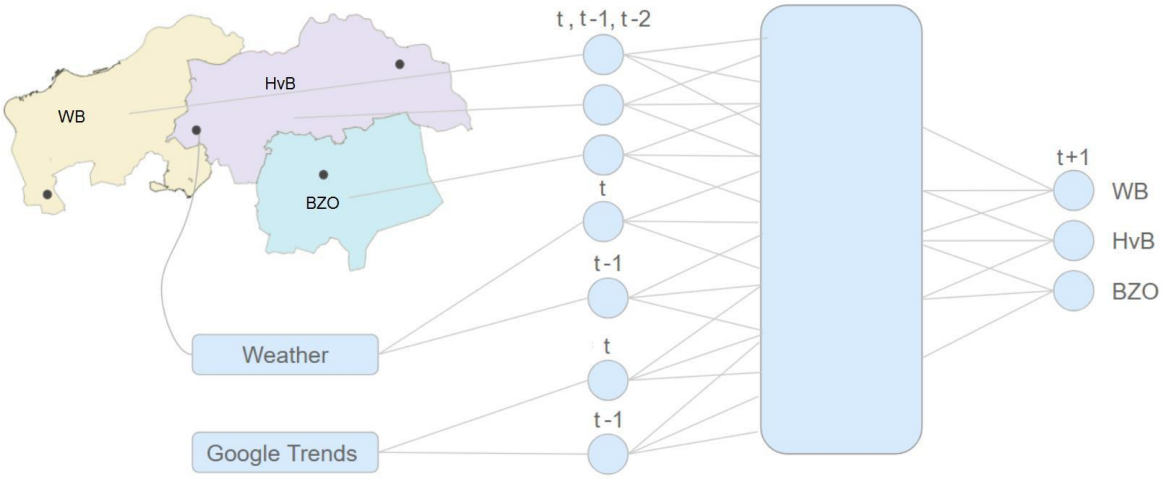
Figuur 3 is een grafische weergave van het neurale netwerk. De provincie Noord-Brabant is onderverdeeld in de drie GGD regio's. Uit Google Trends bleek dat vaker werd gezocht op de term 'kinkhoest' tijdens een uitbraak. Voor de term 'koorts' gold dit niet. De input van dit model bestaat dus

uit het aantal kinkhoest gevallen per regio per maand gecombineerd met de data van Google Trends. Via dit neurale netwerk proberen we het aantal kinkhoest gevallen te voorspellen voor de volgende maand (t+1) per regio, met als input het aantal kinkhoest gevallen per regio in de huidige maand (t), de vorige maand (t-1) en de maanden daarvoor (t-2 en t-3). Om te bepalen wat de prestaties van het neurale netwerk zijn, zijn de data onderverdeeld in een training set en een test set. Het neurale netwerk werd ontwikkeld op basis van de data uit de training set (januari 2004-juni 2013) en getest op de test data (juli 2013-december 2017). De resultaten worden weergegeven in Figuur 4 (a = BZO, b = HvB, c = WB). Hieruit blijkt dat het model voor de factor 'tijd' redelijk in staat is om een uitbraak in de regio Brabant Zuidoost en Hart voor Brabant te voorspellen. Dit geldt echter niet voor de regio West-Brabant. Een verklaring hiervoor is dat in deze regio in 2014 een hele groot uitbraak was met meer dan 110 gevallen, terwijl de meeste uitbraken een aantal had van maximaal 60 tot 80 gevallen. Het model herkent daarom deze uitzonderlijk hoge piek niet, en onderschat de grootte van de uitbraak. Een belangrijke bevinding is wel dat voor alle regio's geldt dat, hoewel het model niet in staat is een exact aantal te voorspellen, het model wel goed in staat is een stijging of een daling te voorspellen.



Figuur 4.

Naast de Google Trends data als extra databron is er ook gekeken naar andere databronnen die de voorspelling van het model mogelijk beter zouden kunnen maken. Uit literatuur blijkt dat voor de infectieziekte kinkhoest het niet geheel duidelijk is of het weer mogelijk invloed heeft op een uitbraak van kinkhoest. Om dat te onderzoeken hebben we het weer als extra databron gebruikt voor het neurale netwerk. Een neuraal netwerk is namelijk een flexibel model dus je kunt gemakkelijk databronnen toevoegen. Als data hebben we de gemiddelde temperatuur afkomstig van vier weerstations in Noord-Brabant toegevoegd aan het model. Het model ziet er dan als volgt uit:



Figuur 5.

Uit de resultaten blijkt dat het toevoegen van een extra databron, data over het weer, aan het neurale netwerk geen verbetering oplevert voor de voorspelling van een kinkhoest uitbraak.

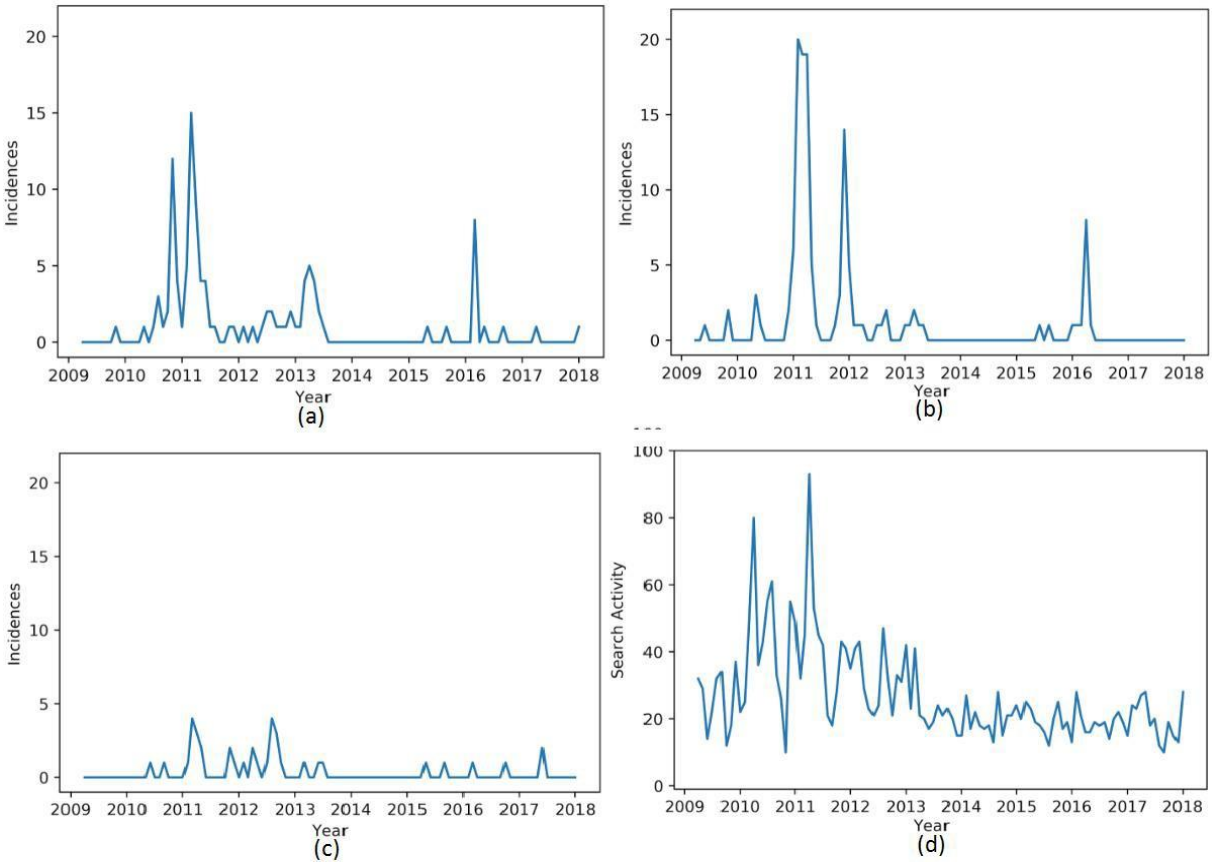
Als laatste is onderzocht of de voorspellingen beter zouden worden als we de 'locatie' specifieker maken. Daarom zijn de GGD regio's verdeeld in twee kleinere subregio's. Elke GGD regio is verdeeld in een oostelijk en westelijk deel en in elk deel bevindt zich minimaal één grote gemeente. Uit de resultaten bleek dat het verdelen van de GGD regio in subregio's leidde tot meer ruimtelijke complexiteit, waardoor het model niet meer zo effectief was in het voorspellen van een kinkhoest uitbraak als wanneer we alleen de GGD regio's gebruiken.

Analyse van de bof

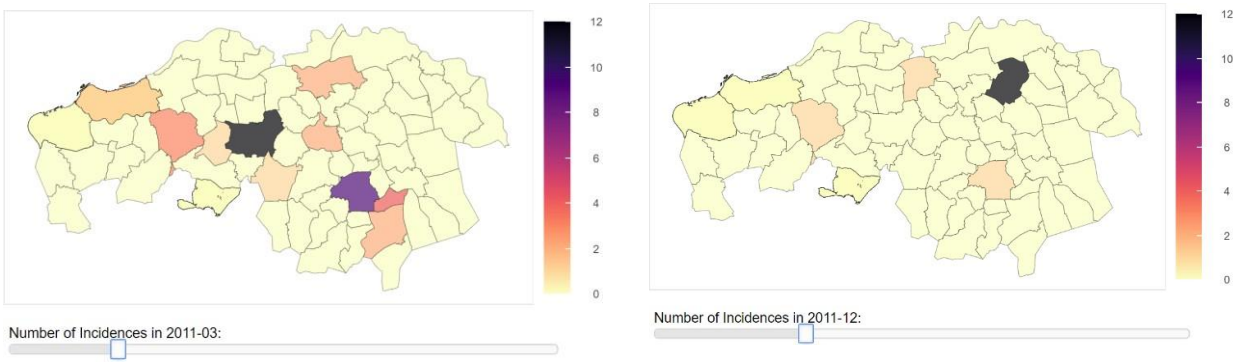
Een doel van het onderzoek is of het model dat ontwikkeld is voor het voorspellen van een kinkhoest uitbraak ook te gebruiken is bij andere infectieziekten. Hiervoor gebruiken we data van de bof. Uit de beschrijvende analyses blijkt dat de bof meer voorkomt bij mannen dan bij vrouwen in de leeftijd van 20 tot 25 jaar oud. In Figuur 6 is weergegeven hoeveel gevallen van bof er waren in de periode van 2009 tot 2018 voor de drie verschillende GGD regio's. (a=BZO, b=HvB, c=WB). In Figuur 6d zien we de Google Trends data waarbij gekeken is naar de zoekterm 'bof'. De resultaten laten grote verschillen in het aantal bof gevallen in de drie regio's.

Bij de volgende stap is opnieuw gebruik gemaakt van de eerdere ontwikkelde interactieve kaarten. Door het verschuiven van de *time slider* werd zichtbaar dat de meeste gevallen van de bof tijdens een uitbraak werden geregistreerd in de grote gemeenten (Figuur 7 links). Verder was het opvallend dat 13 van de 14 gevallen die werden gerapporteerd tijdens de uitbraak afkomstig waren van de gemeente Bernheze (Figuur 7 rechts). Nader onderzoek wees uit dat er in die tijd een groot feest werd gegeven in een dorp in de gemeente en dat daar waarschijnlijk mensen elkaar besmet hebben. Eenzelfde soort situatie is te zien in april 2016. In die periode werd een voetbaltoernooi gehouden in een klein dorpje in de gemeente Landerd, en ook daar hebben de jongeren elkaar waarschijnlijk besmet. Deze clusters waren uiteraard bekend bij de GGD.

Als laatst is geprobeerd om het ontwikkelde model voor het voorspellen van een uitbraak van kinkhoest ook toe te passen voor de bof. Door de kleine dataset (N=279) was het onmogelijk een model te ontwikkelen. Bovendien waren er in veel maanden geen gevallen van de bof. Het lukte daarom niet om voorspellingen te doen over een bofuitbraak met dit model.



Figuur 6.



Figuur 7.

RESULTATEN KINKHOEST REGIO ZUID-LIMBURG

Beschrijvende analyses

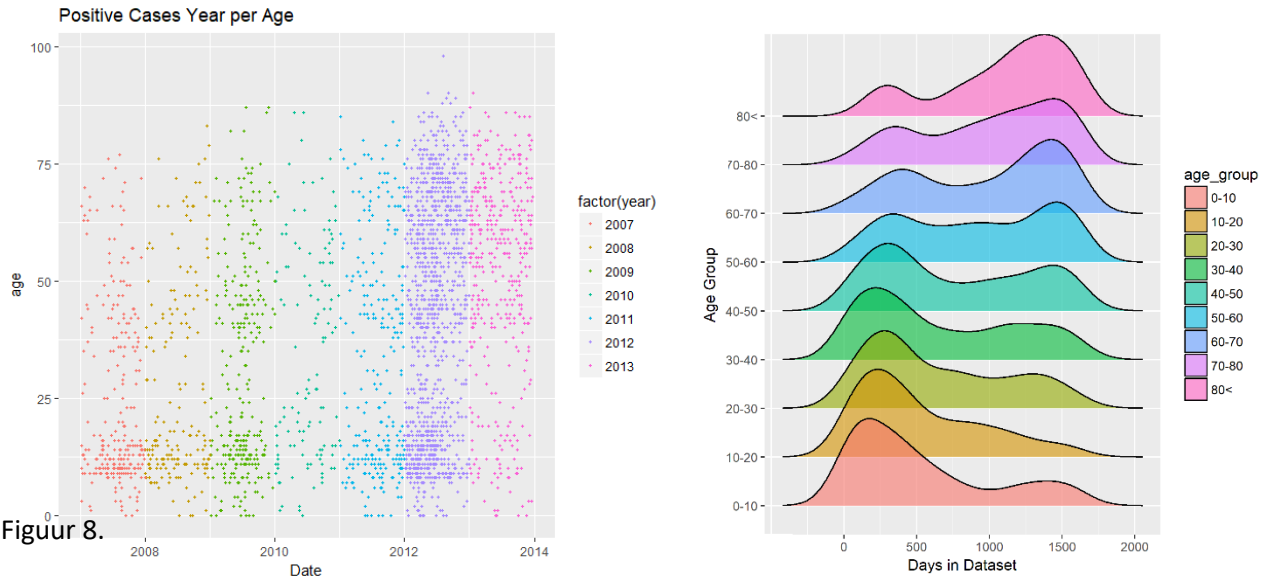
De dataset van Limburg bevat 12.237 cases die via het laboratorium getest zijn om te bepalen of zij wel of geen kinkhoest hebben. In Tabel 1 staan de resultaten weergegeven van de labaanvraag voor kinkhoest met negatief (0), onbekend (-1) of positief (1) resultaat, voor de periode januari 2007 tot december 2013 in de provincie Limburg.

Tabel 1.

Lab Result	Number of Cases	Prob. For Female	Prob. For Male
-1	3262 (%27)	0.62	0.38
0	6154 (%50)	0.63	0.37
1	2821 (%23)	0.55	0.45

23% van de cases werd positief (1) getest, bij 50% was de test negatief (0) en bij 27% is er geen duidelijk testresultaat (-1). Verder zien we dat het aantal vrouwen dat negatief test veel hoger is dan het aantal mannen dat negatief test.

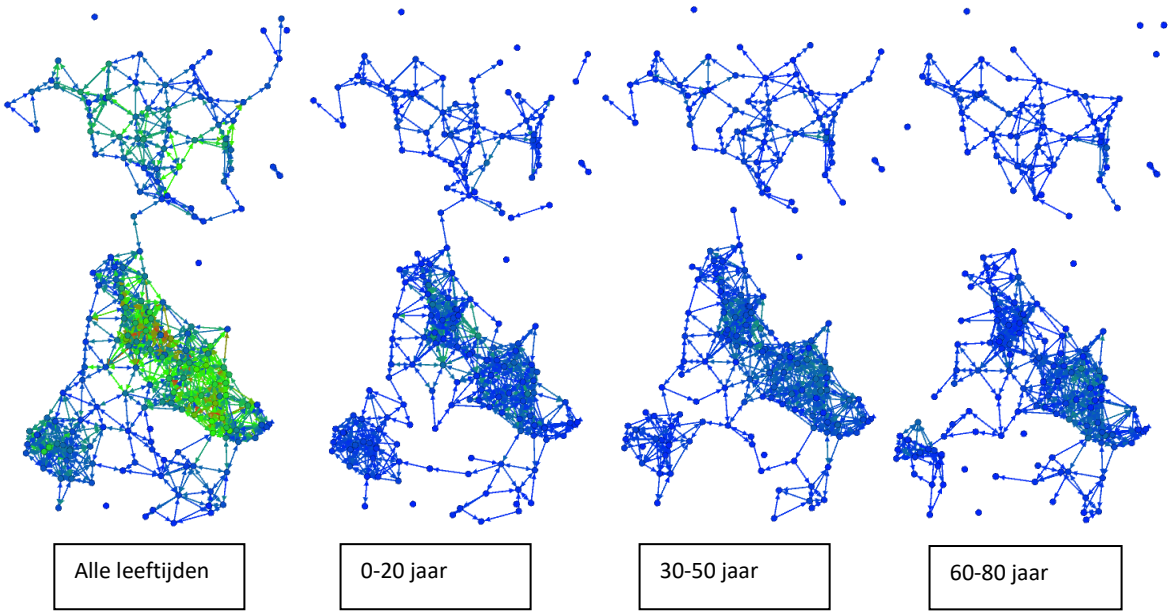
De spreidingsplot in Figuur 8 (linker) laat zien dat er een patroon zichtbaar is gedurende de tijd voor de leeftijd van de positief geteste personen. In 2007 en 2008 betroffen de meeste positieve gevallen kinderen en jongeren terwijl in 2013 de meeste positieve gevallen ouderen waren. Ditzelfde patroon is zichtbaar in Figuur 8 (rechter). Daarnaast laten de resultaten zien dat mensen tussen de 20-40 jaar oud minder kans hebben op het krijgen van kinkhoest dan mensen in andere leeftijdsgroepen.



Figuur 8.

Grafische analyse

Een grafisch netwerk bestaat uit knopen en lijnen. De knopen (nodes) staan voor de postcodes van de positief geteste personen. Twee knopen zijn met elkaar verbonden door een lijn (edges) als de *influence rate* (IR) zoals gedefinieerd in 3.2.2 niet nul is. Het gewicht van een lijn is gelijk aan de IR. De gewogen lijnen bepalen hoeveel effect de punten op elkaar hebben. De netwerken met de punten en lijnen, die worden weergegeven binnen de contouren van de provincie Limburg, staan weergegeven in Figuur 9.



Figuur 9.

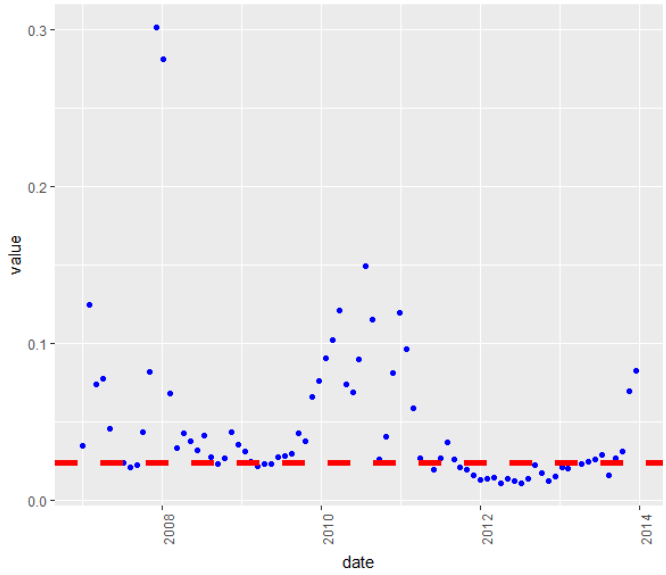
De kleur van de lijnen die twee punten met elkaar verbindt geeft het gewicht weer tussen de twee punten. Rood geeft aan dat het ene punt veel effect heeft op het andere punt. Bij een groene lijn is het effect matig en bij een blauwe lijn is er nauwelijks sprake van effect tussen de twee punten. Gebaseerd op de kleuren in de netwerken is het effect groter tussen leeftijdsgroepen dan binnen leeftijdsgroepen. Daarnaast laten de resultaten zien dat de locatie de grootte van de kans bepaalt dat iemand besmet wordt met kinkhoest. De meeste kinkhoestgevallen komen voor in steden waar de *influence rate* hoog is.

Predictiemodel

In 2009 en in 2012 waren twee grote kinkhoest uitbraken in de provincie Limburg. Om een uitbraak van kinkhoest te kunnen voorspellen is gebruik gemaakt van een tijdsvenster van 1 maand. De hele periode die beschikbaar is in de dataset is verdeeld in 84 maanden. Voor elke maand is een netwerk gemaakt waarin de spreiding van het aantal positieve gevallen staat weergegeven en hun relatie met elkaar. Verder is de *Weighted Graph Density (WGD)* van elk netwerk berekend om de netwerken met elkaar te kunnen vergelijken (Figuur 10). De drempelwaarde waarbij sprake is van een uitbraak is de rode lijn. Deze drempelwaarde is als volgt berekend:

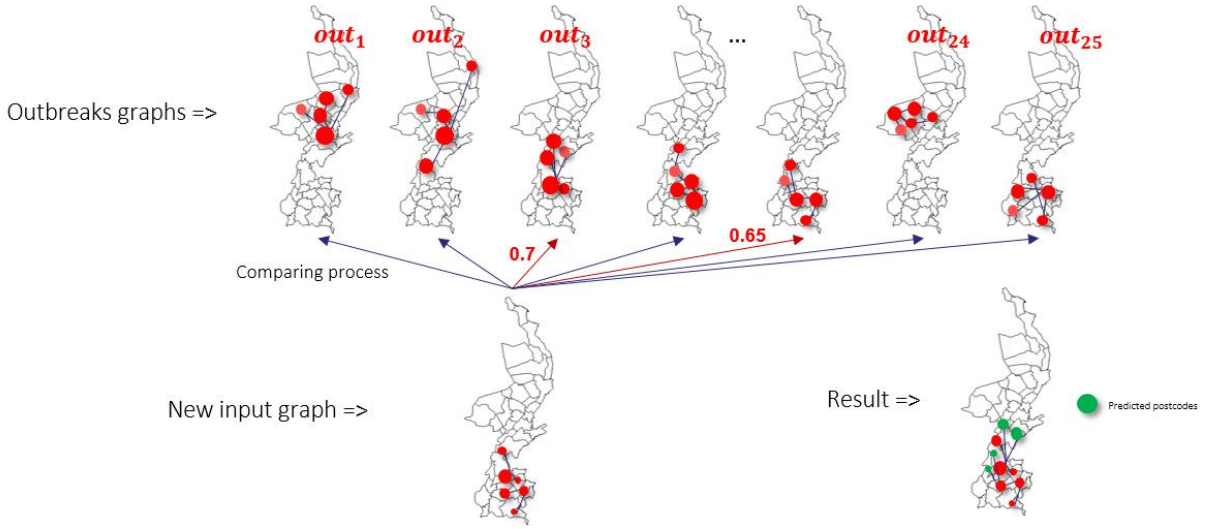
drempelwaarde = gemiddelde (van de waarden van alle blauwe punten) – standaarddeviatie (van de waarden van alle blauwe punten)/2

Zo wordt zichtbaar dat in deze periode 25 keer sprake is geweest van een uitbraak.



Figuur 10.

De volgende stap voor het maken van voorspellingen was om een nieuw netwerk in de volgende tijdsperiode te vergelijken met elk netwerk waarbij sprake is van een uitbraak. Om de netwerken met elkaar te kunnen vergelijken wordt gebruik gemaakt van de *most-common-subgraph* methode waarbij een *similarity score* wordt berekend. Deze score geeft de mate van gelijkenis tussen twee netwerken weer. Figuur 11 laat zien hoe een nieuw netwerk zich volgens de *similarity score* verhoudt tot de netwerken waarbij sprake is van een uitbraak. Bijvoorbeeld, de *similarity score* tussen het netwerk in de nieuwe maand en het netwerk met een uitbraak out_3 is 0,7. Dit is de hoogste waarde van alle *similarity scores.* Op deze manier kunnen we zien of er in de volgende periode een uitbraak van kinkhoest zal zijn of niet. Daarnaast kunnen we een signaal afgeven voor een aantal postcodes die niet in het nieuwe netwerk zitten, maar die volgens het netwerk met uitbraak wel veel gelijkenis vertonen met het nieuwe netwerk in de volgende maand. Daarom kunnen we, naast dat we een uitbraak kunnen voorspellen voor de volgende maand, ook de waarschijnlijkheid voorspellen dat bepaalde postcodes te maken krijgen met kinkhoest.



Figuur 11.

Als laatste testen we het model voor één netwerk. Tabel 2 laat zien dat het netwerk het meest gelijk is aan uitbraaknetwerk 15 (similarity score = 0.75). Vanwege deze hoge *score* kunnen we zeggen dat er hoogstwaarschijnlijk in de volgende maand een uitbraak van kinkhoest zal zijn.

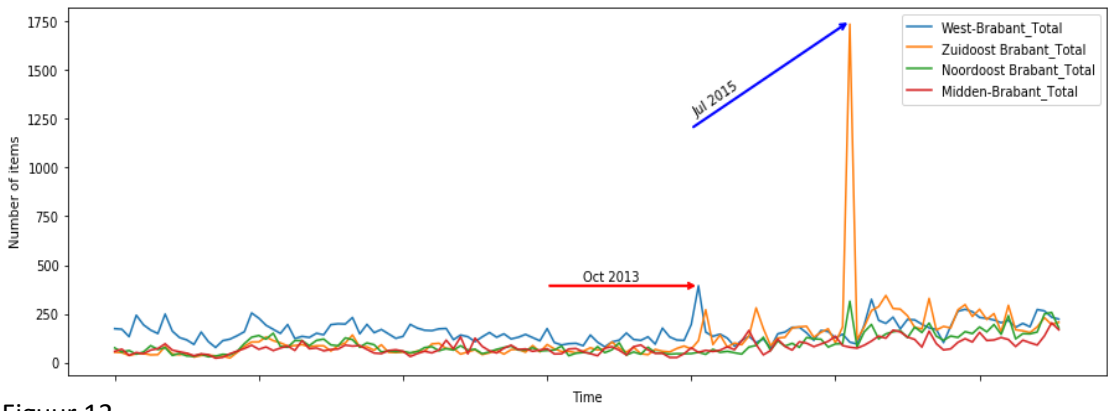
Tabel 2.

Outbreak Graph Index	Similarity score
15	74.8704663
16	69.6891192
13	55.6994819
14	52.8497409
12	42.4870466
17	35.492228

RESULTATEN SCABIËS REGIO NOORD-BRABANT

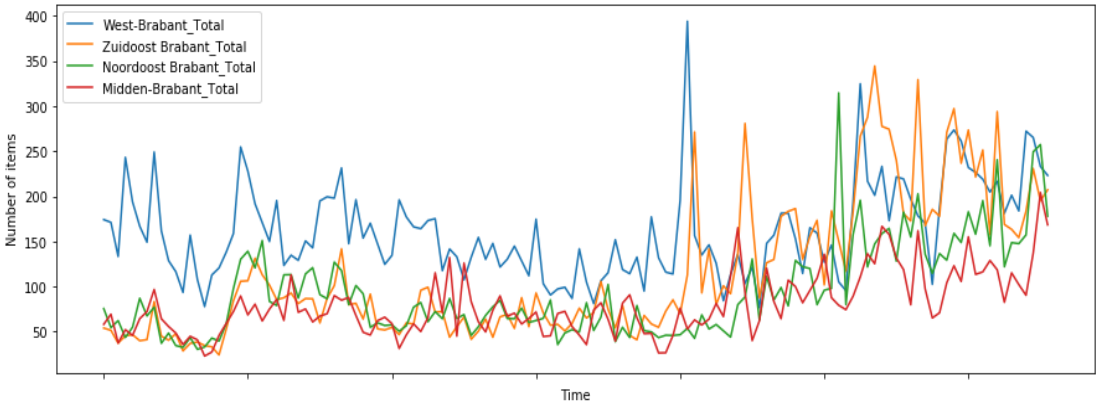
Exploratory data analysis (EDA)

In Figuur 12 zien we cijfers van de verstrekking van medicijnen door apotheken om scabiës te behandelen in Noord-Brabant. In juli 2015 was er een significante piek in de verkoop zichtbaar, in totaal werden 2.333 items door apotheken verstrekt (gemiddeld = 501, standaarddeviatie = 268). Bij navraag aan deskundigen bij de GGD bleken in die periode veel vluchtelingen uit Eritrea naar Nederland te zijn gekomen, waarvan velen gediagnosticeerd werden met scabiës. We beschouwen deze piek als een uitschieter in de verdere analyses.



Figuur 12.

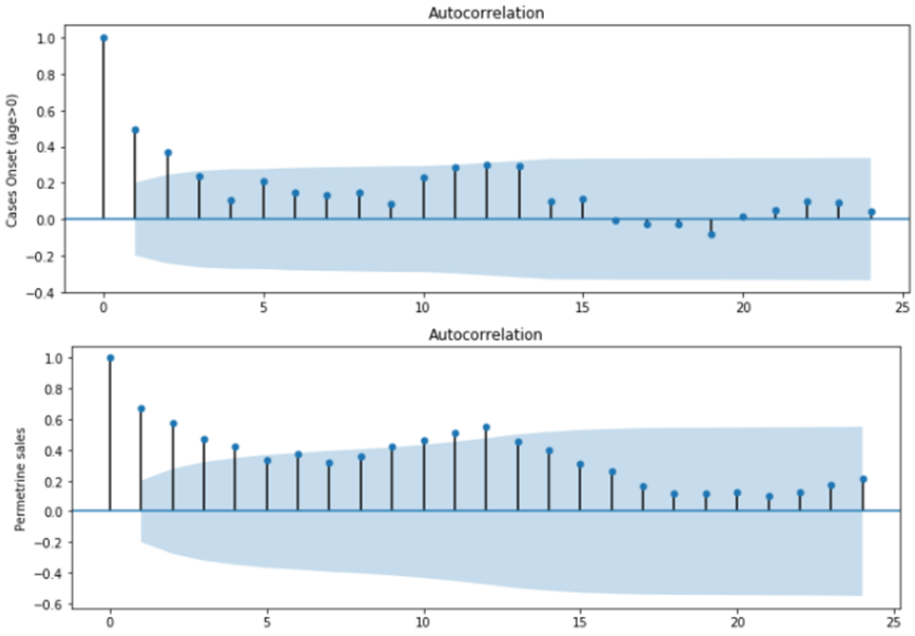
Als we de *uitschieter* verwijderen krijgen we een duidelijkere grafiek van de verstrekking van medicijnen voor de behandeling van scabiës. Figuur 13 laat zien dat van 2007-2014 het aantal medicijn verstrekkingen tegen scabiës in de regio West-Brabant hoger ligt dan in de rest van Brabant, maar vanaf 2014 zien we wel een stijging in alle regio's. De grootste stijging zien we dan in Zuidoost Brabant, waarbij het aantal verstrekkingen doorgaans hoger ligt dan in West-Brabant.



Figuur 13.

Uit de leeftijdsverdeling komt naar voren dat het aantal baby's aan wie het medicijn wordt verstrekt hoog is. Van het totaal aantal cases (699) is 14% jonger dan 1 jaar. Dit hoge aantal lijkt niet te kloppen, aangezien het hoogste percentage verwacht wordt bij studenten. De groep kinderen jonger dan 1 jaar wordt dan ook niet meegenomen in verdere analyses.

Verder blijkt er een relatie tussen het aantal cases en het aantal verstrekkingen van medicijnen, vooral voor het medicijn Permetrine. De grafieken in Figuur 14 laten een harmonische autocorrelatie zien van beide variabelen, weergegeven in maanden.

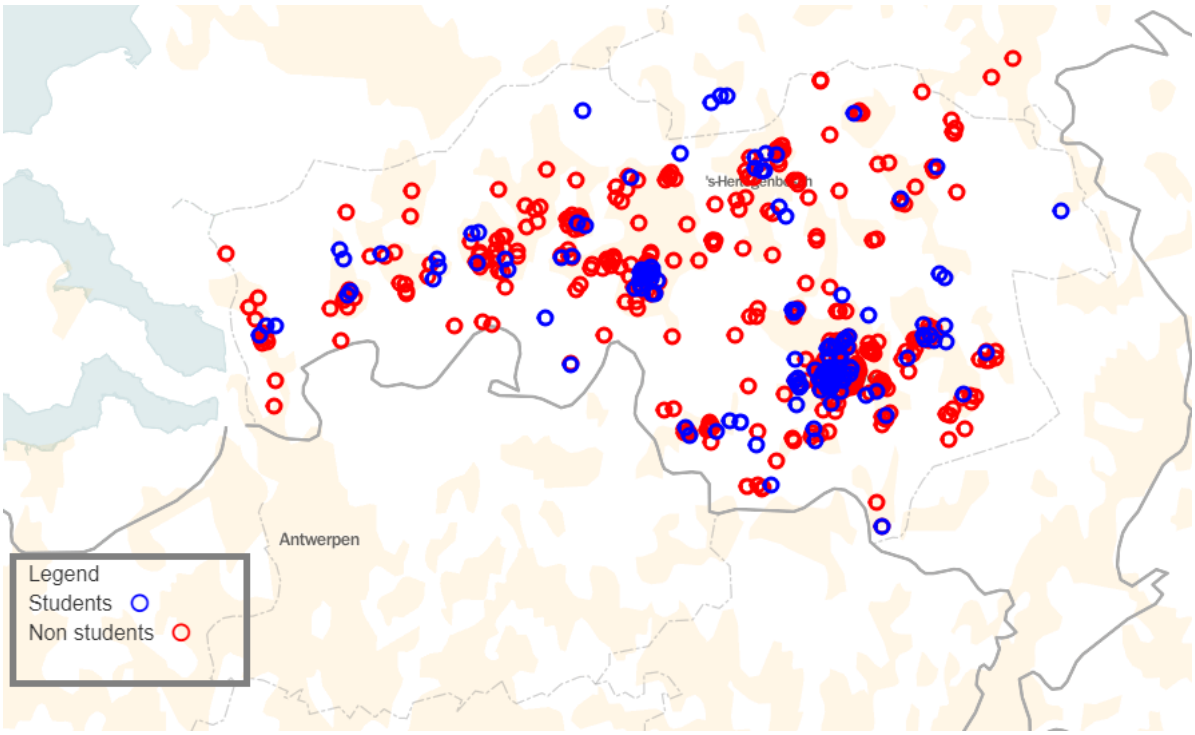


Figuur 14.

De correlatie is het hoogst voor de meest recente maanden (0, 1 en 2) en voor de maanden van ongeveer 12 maanden geleden. Het gemiddeld aantal cases en het gemiddeld aantal verstrekkingen van Permetrine (onderste figuur) is het hoogst in de 11^{e} en 12^{e} maand van het jaar. Het gemiddeld aantal verstrekkingen van Permetrine is eveneens hoog in de eerste drie maanden van het jaar. Dit zou erop kunnen wijzen dat deze variabelen gerelateerd zijn aan het weer, waarbij de verstrekking en het voorkomen van scabiës hoger lijkt te zijn in de koudere seizoenen. Echter, de variantie van deze

variabelen is groot. Daarom verifiëren we of deze instabiliteit van de variabelen kan worden verklaard aan de hand van de dynamiek in de verschillende regio's.

Voor alle vier de regio's is er een hogere vraag naar Permetrine in de periode van oktober tot januari. In de zomermaanden is de vraag relatief laag en in de regio's Zuidoost en Midden-Brabant is het nog lager. Deze daling is mogelijk te verklaren doordat in deze regio's twee steden liggen met een universiteit (Tilburg en Eindhoven) en in de zomermaanden studenten minder het studentenleven leiden door de zomervakantie. Dit heeft tot gevolg dat scabiës dan mogelijk minder voorkomt. Om deze reden is gekeken naar de spreiding van scabiës tussen studenten en niet-studenten. De aanname is dat studenten tussen de 18 en 28 jaar oud zijn. Figuur 15 laat de spreiding zien van scabiës gevallen van studenten en niet-studenten.



Figuur 15.

Op basis van Figuur 15 kunnen we concluderen dat er twee clusters van studenten zijn, één in Eindhoven en één in Tilburg. De niet-studenten bevinden zich meer verspreid over de provincie.

RELATIES ONDERZOEKEN

We onderzoeken in deze sectie relaties tussen verschillende variabelen. Deze relaties worden onderzocht aan de hand van Pearson correlaties en *estimated joint distributions.*

Correlatie

In Tabel 2 zien we de correlaties tussen verschillende variabelen. De meest interessante correlatie is de correlatie tussen het aantal medicijnverstrekkingen van Permetrine en aantal scabiës gevallen per maand in Noord-Brabant (r = 0.55). Hoewel dit niet hoog is, lijkt er wel een positieve relatie te zijn tussen deze twee variabelen. Daarnaast zien we een redelijke correlatie tussen Google Trend data en het aantal verstrekkingen van Permetrine. Verder is het een interessante bevinding dat de correlatie

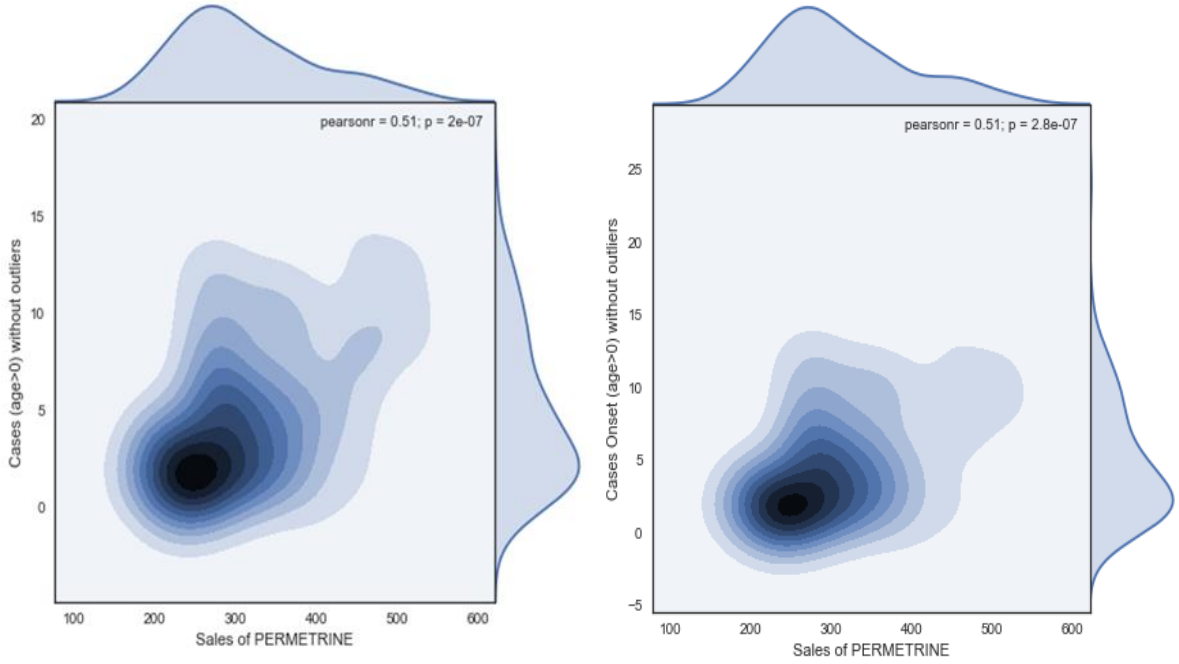
tussen het aantal medicijnverstrekkingen in de volgende maand en het aantal zoekacties naar scabiës volgens Google Trend in de huidige maand bijna hetzelfde is (0.41). Dat zou erop kunnen wijzen dat wanneer er in de huidige maand meer gezocht wordt op scabiës in Google dat er in deze én de volgende maand meer medicijnen worden uitgegeven. Daarmee kun je concluderen dat er meer mensen scabiës hebben in deze periode.

Tabel 3.

	Cases	**Cases Onset**	**Permetrine Sales**	**Google Trends**
Cases	1.00		0.55	0.36
Cases Onset		1.00	0.51	0.40
Permetrine Sales			1.00	0.45
Google Trend				1.00

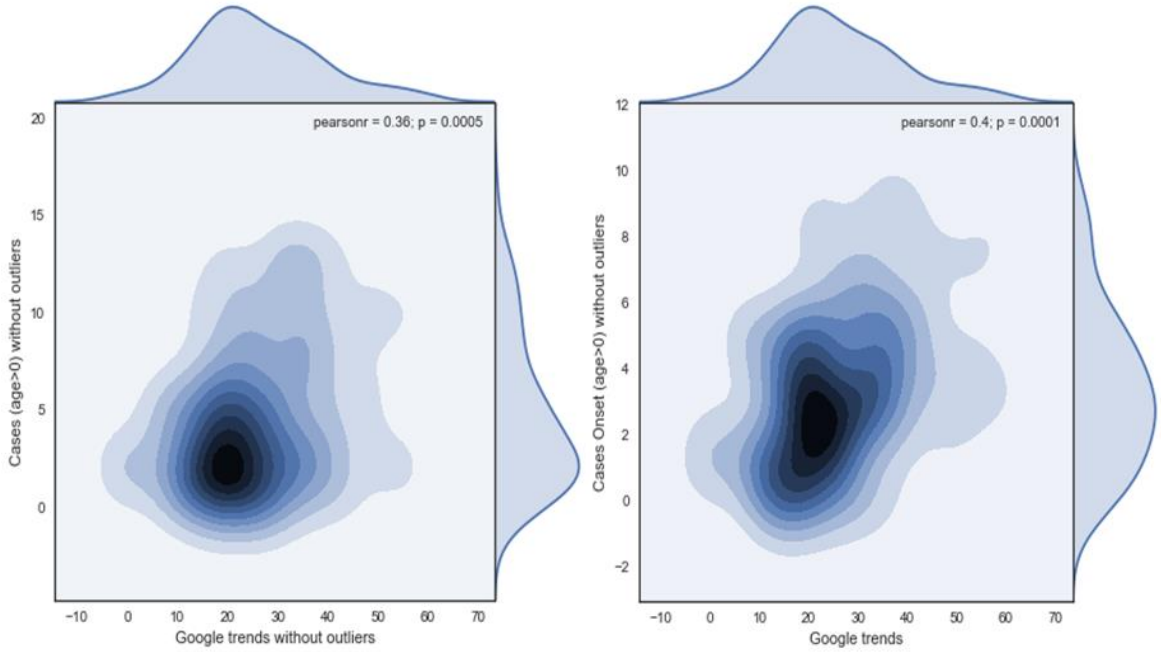
Kernel density estimation

De *joint probability density functions* voor de variabelen *Cases, Case onset, Sales of Permetrine* en *Google Trends* zijn berekend aan de hand van *Kernel density estimation. Case onset* geeft het aantal cases per maand weer gebaseerd op de *date of onset.* De *date of onset* en de datum waarop een casus daadwerkelijk geregistreerd wordt kan maximaal 3 maanden verschillen. De verwachting is dat *Case onset* een sterkere correlatie heeft met *Sales of Permetrine* en *Google Trends.* Echter, de correlatie tussen *Case onset* en *Sales of Permetrine* is precies hetzelfde als de correlatie tussen *Cases* en *Sales of Permetrine. Deze resultaten worden ondersteund door de resultaten uit de estimated density distributions (zie* Figuur 16).



Figuur 16.

De correlatie tussen *Case Onset* en *Google Trends* is wat sterker dan de correlatie tussen *Cases* en *Google Trends.* De *approximated joint distribution* in Figuur 17 laat een *direct proportional relation* zien tussen zowel *Case onset* als *Cases*, waarbij deze relatie voor *Case Onset* iets sterker is.



Figuur 17.

Voorspellen van een uitbraak

Als laatste is onderzocht of er een model ontwikkeld kan worden om een scabiësuitbraak te voorspellen. Hiervoor is een drempelwaarde voor de Google Trends data bepaald die een bepaalde maand als 'kwetsbaar' bestempeld voor een scabiës uitbraak. Om de beste drempelwaarde te vinden is eerst de dataset opgesplitst in een training set (65 *entries)* en een test set *(22 entries)*. Vervolgens wordt een functie gedefinieerd die het aantal *true positives* en het aantal *true negatives* maximaliseert als een gewogen gemiddelde. De *true positives* en *true negatives* zijn als volgt gedefinieerd:

- *True positive:* in het geval dat de methode een signaal voor een uitbraak geeft (positief volgens Google Trends data) en er ook echt sprake is van een uitbraak (volgens de *Cases* data).
- *True negative:* in het geval dat de methode geen signaal voor een uitbraak geeft (negatief, volgens Google Trend data) en er is ook geen sprake van een uitbraak (volgens de *Cases* data).

Ten slotte wordt de *accuracy* van de drempelwaarde berekend als volgt:

$$Accuracy = \alpha_1 \frac{Num.\,of\ True\ positives}{Number\ of\ positive\ cases} + \alpha_2 \frac{Num.\,of\ True\ negatives}{Number\ of\ negative\ cases}$$

De waardes $\alpha_1 \in (0,1)$ en $\alpha_2 \in (0,1)$ (de alpha-waarden liggen tussen 0 en 1) geven weer welk gewicht er gegeven wordt aan respectievelijk de *true positives* en *true negatives*. Verder wordt aangenomen dat er sprake is van een uitbraak als er meer dan 4 gevallen van scabiës in een maand voorkomen. Dit is gebaseerd op de waardes van de *Cases* data en niet op een bestaande definitie van

wanneer we spreken van een uitbraak van scabiës. In Tabel 4 staat de optimale grenswaarde voor de Google Trends data voor diverse *accuracy* waarden.

Tabel 4.

α_1	α_2	**Optimal Threshold**	**Train Accuracy**	**Test Accuracy**
0.33	0.67	33.58	72.2%	66.8%
0.5	0.5	26.41	67.4%	50%
0.66	0.33	16.6	69.4%	66.0%

De resultaten laten zien, dat de accuraatheid onvoldoende hoog is om een scabiësuitbraak goed te voorspellen.

Conclusie

In de eerste studie is een neuraal netwerk ontwikkeld om een kinkhoestuitbraak te voorspellen in de provincie Noord-Brabant. Op basis van de resultaten kunnen we concluderen dat het neurale netwerk een bijdrage levert aan het voorspellen van een uitbraak van kinkhoest. Als we strikt kijken naar de prestaties van het model lijkt het model niet voldoende in staat om een uitbraak correct te voorspellen. Dit is echter te verklaren doordat de prestaties gemeten worden aan de hand van de juiste voorspelling van het exact aantal patiënten met kinkhoest. Echter, voor de praktijk is het voornamelijk van belang dat het model een sterke stijging (i.e., uitbraak) goed kan voorspellen en niet exacte aantallen. Het is uiteindelijk de beslissing van de afdeling infectieziektebestrijding van de GGD om wel of geen preventieve maatregelen te treffen in de regio. Daarbij is de ontwikkelde interactieve kaart van nut om te gebruiken, omdat deze de data visueel in kaart brengt, wat de spreiding inzichtelijk maakt. Zo werd zichtbaar dat, vóórdat een uitbraak plaatsvindt, het aantal patiënten in de gemeenten met een groot aantal inwoners stijgt. Dergelijke signalen kunnen, naast de eigen expertise en kennis van teams infectieziektebestrijding, meegenomen worden bij het nemen van een beslissing voor eventuele preventieve maatregelen. Om het neurale netwerk te verbeteren is gekeken naar extra databronnen. Zo is er gekeken naar meteorologische data van het weer, omdat het seizoen mogelijk invloed heeft op de verspreiding van kinkhoest. Deze data verbeterden de voorspellingen echter nauwelijks.
In deze studie is verder gekeken of het ontwikkelde model ook gebruikt kan worden voor het voorspellen van een uitbraak van een andere infectieziekte, namelijk de bof. Resultaten lieten zien dat dit niet mogelijk was voor de bof, vanwege de lage aantallen. De gebruikte analysetechnieken zijn niet geschikt om met kleine datasets te werken.

In een tweede studie, is gekeken of het model te valideren is voor andere regio's. Hierbij is gebruik gemaakt van data van de provincie Limburg. Zoals eerder genoemd betrof het andere type data dan de data die is gebruikt voor de regio Noord-Brabant. Tijdens het onderzoek bleek dat het ontwikkelde model voor de regio Noord-Brabant niet het beste model leek voor de regio Limburg. Daarom zijn andere analysetechnieken gebruikt om een uitbraak van kinkhoest in de regio Limburg te voorspellen. Uit de resultaten van de grafische analyse bleek dat, gebaseerd op de *positive cases,* het model in staat is een uitbraak te voorspellen voor de volgende periode. Het model werkt effectief, de resultaten zijn betrouwbaar en geeft de waarschijnlijkheid weer voor het krijgen van kinkhoest voor een specifieke locatie. Om het model verder te verbeteren wordt geadviseerd om het huidige model ook toe te

passen in andere provincies. Daarnaast kan er worden gekeken of er met andere aannames dan nu zijn gedaan, zoals dat besmetting alleen plaatsvindt als mensen zich minder dan acht kilometer van elkaar bevinden, de voorspellingen verbeterd kunnen worden. Op dit moment wordt er in samenwerking met de JADS en de drie Brabantse GGD'en verder onderzoek gedaan, waarbij o.a. geprobeerd wordt om het huidige model toe te passen in de provincie Noord-Brabant aan de hand van hun labdata.

In de derde studie is onderzocht of een uitbraak van scabiës te voorspellen is. Voor scabiës is naast de dataset van de GGD (aantal gemelde gevallen van scabiës) vooral gebruik gemaakt van een extra databron, het aantal verstrekkingen door apotheken van medicijnen voor de behandeling van scabiës. Voor deze studie is gekeken of het zoekgedrag op internet (Google Trend) en de verstrekkingen van het aantal medicijnen inzichtelijk maakt hoe scabiës zich op de grote schaal ontwikkelt (bv. een uitbraak, nieuwe besmettingen, clusters). Uit de resultaten kwamen echter geen sterke relaties naar voren tussen de verschillende variabelen. Geadviseerd wordt om het aantal verstrekkingen van medicijnen te gebruiken als uitkomstvariabele. Dit is gebaseerd op de aanname dat het aantal verstrekkingen van medicijnen een directe relatie heeft met het aantal scabiës gevallen. Ook naar dit advies wordt gekeken in de nieuwe studie zoals hierboven genoemd. Omdat veel medicijnen voor de behandeling van scabiës vrij verkrijgbaar zijn, in bv. drogisterijen, wordt geprobeerd deze data te verkrijgen en te gebruiken voor het onderzoek.

Om de ontwikkelde modellen te gebruiken in de praktijk van teams infectieziektebestrijding is het van belang deze modellen in te bouwen in een bruikbare tool. Dit bleek in de praktijk lastiger dan vooraf was ingeschat. Op het moment wordt een nieuwe tool ontwikkeld, die bruikbaar is voor teams infectieziektebestrijding.

Tot slot heeft deze studie ook geleid tot een nieuw project, waarbij gekeken wordt naar het voorspellen van griep op lokaal niveau. We weten dat er elk jaar een griepepidemie plaatsvindt, maar de ernst, duur en piek van de epidemie zijn onbekend. De griepepidemie vormt een risico voor de zorgcontinuïteit (GGD/GHOR is hiervoor verantwoordelijk). Door de duur, ernst, en piek van een griepepidemie vroegtijdig te signaleren, kunnen tijdig preventieve maatregelen worden genomen om de zorgcontinuïteit te garanderen. Ook dit onderzoek is een samenwerking tussen de JADS en de drie Brabantse GGD'en en GHOR's.