

Datafusie – wat wordt ermee bedoeld en wat heb je eraan?

Heb je het met iemand over verkeersdata, dan valt – naast de zo populaire kreet ‘big data’ – bijna altijd de term datafusie. In veel gevallen is het echter helemaal niet zo duidelijk wat daar nu precies mee wordt bedoeld en waarom het zinvol zou zijn. In deze tutorial daarom een korte introductie. We gaan daarbij vooral in op de kwaliteit van data: welke mogelijkheden biedt datafusie om die kwaliteit te verbeteren?

Datafusie is het bij elkaar brengen van data en informatie van verschillende aard of uit verschillende bronnen. Denk bijvoorbeeld aan het combineren van lusdata met reistijddata of *floating car data* (FCD). De verwachting is dat het samenbrengen van verschillende databronnen zal leiden tot een aanzienlijke verbetering van zowel de *kwaliteit* als de *omvang* van de informatie die we uit onze verkeersgegevens halen. Hiermee bedoelen we dat niet alleen de nauwkeurigheid, de betrouwbaarheid en de tijdigheid van de informatie verbeteren (kwaliteit),

maar ook het aantal verkeerskundige grootheden dat in de gegevens is vervat (omvang). Een voorbeeld: van een wegvak met alleen meetlussen weten we de verkeersintensiteit, maar door die data te combineren met bluetooth-waarnemingen kunnen we ook de reistijden en de herkomst-bestemmingsgegevens achterhalen.

Datafusie biedt daarmee interessante mogelijkheden. We komen méér te weten en wat we weten, weten we beter (nauwkeuriger). Een andere optie is om datafusie te gebruiken om een gelijkblijvende kwaliteit te leveren tegen minder kosten. Zo zou een slimme combinatie van FCD uit in-car systemen en de lusdata van enkele lussen dezelfde informatie-kwaliteit kunnen leveren als het huidige dichte meetlussen-net. In een tijdsgewricht waarin we over ‘meer met minder’ en over het ‘verdunnen van het meetnet’ spreken, is dat een welkome optie!

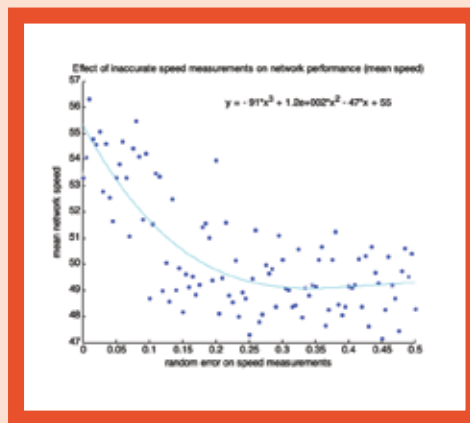
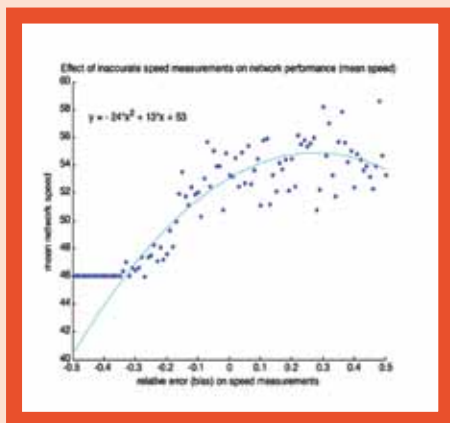
Datakwaliteit

Maar alvorens we verder ingaan op de mogelijkheden van datafusie, willen we kort stilstaan bij wat we eigenlijk bedoelen als we het hebben over de kwaliteit van data. Duidelijk is dat die kwaliteit nog vaak te wensen overlaat. De nauwkeurigheid van locatiebepaling is bijvoorbeeld sterk afhankelijk van de inwinmethode (gps of wifi) en de omgeving. Ook de vertrouwde inductielussen in het Nederlandse hoofdwegennet hebben te kampen met ontbrekende data (8-12% in 2012, bron: Rijkswaterstaat) en ‘outliers’, zoals zeer hoge

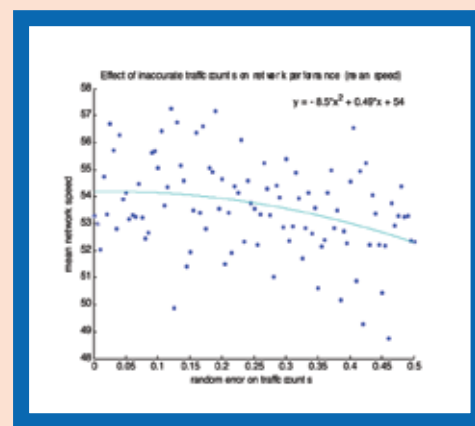
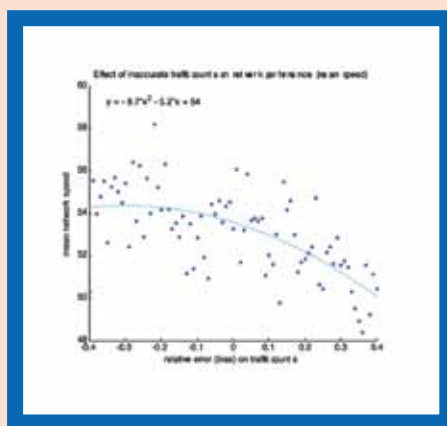
gemeten snelheden en onder- of overschattingen in het aantal getelde voertuigen. De Nationale Databank Wegverkeersgegevens (NDW) heeft een minimaal kwaliteitsniveau bepaald waaraan de data die wordt aangeleverd door de verschillende dataproviders, moet voldoen. Het is evident dat de kwaliteitseisen niet voor elke toepassing en elke situatie hetzelfde zouden moeten zijn. Een tijdkritische maatregel als het automatisch incidentdetectiesysteem (AID), dat weggebruikers waarschuwt voor aankomende files, heeft bijvoorbeeld behoefte aan nauwkeurigere en een hogere resolutie data dan een systeem voor route-informatie of een netwerkbreed verkeersmanagementsysteem. Maar wat de precieze relatie is tussen deze verschillende toepassingen en de kwaliteitseisen van de data, is in de meeste gevallen niet bekend.

Dat is wel een belangrijk punt. Immers, als je met een mindere kwaliteit van data toekunt, levert dit een besparing op, bijvoorbeeld omdat minder detectielussen nodig zijn. Zoals we hierboven al aanstipten, kan ook datafusie hierin een rol spelen. Door twee bronnen van onvoldoende kwaliteit te fuseren, kan nieuwe data gegenereerd worden die wél van voldoende kwaliteit is.

Hoe kun je nagaan of je data van voldoende kwaliteit is? Een belangrijke overweging is wat het doel is van de betreffende verkeersmanagementmaatregel. De data is immers goed genoeg als je het doel van de maatregel kunt bereiken. In de praktijk is dit doel vaak niet helder. Zo wil je met een routeringsmaatregel reistijdverliezen voorkomen, maar hoeveel en wat is haalbaar? Het signaleringssysteem is bedoeld om de verkeersveiligheid te verbeteren, maar hoe kun je dit aantonen en wanneer is het goed genoeg? Om een goede analyse van de benodigde datakwaliteit te kunnen doen, moet deze *Measure of Perfor-*



Figuur 1 en 2
Effect van onnauwkeurige snelheidsmetingen op de netwerkprestatie.



Figuur 3 en 4
Effect van onnauwkeurige intensiteitsmetingen op de netwerkprestatie.

mance vooraf bepaald worden.

Om het nog verder te compliceren, speelt bij veel maatregelen ook de menselijke factor mee. Als reizigers er door ervaring achter komen dat de verschaft reistijden niet correct zijn, zullen ze een volgende keer dan weer hun beslissing baseren op het systeem? Welke fout is nog toelaatbaar? In dit geval zou je al deze menselijke overwegingen moeten meenemen om te bepalen wat de toegestane fout is om een reistijd van voldoende kwaliteit te kunnen garanderen.

Optimalisatie meetconfiguratie in de ontwerpfase

Laten we aan de hand van een specifiek voorbeeld kijken hoe de kwaliteit van de data de prestatie van de verkeersmanagementmaatregel beïnvloedt. In het voorbeeld hebben we met behulp van simulaties bepaald wat het effect is van de nauwkeurigheid van de inwindata voor een toeritdoosinstallatie (TDI) op de prestatie van het netwerk, uitgaande van een gangbare configuratie van lusdetectoren. Vervolgens is aangenomen dat dezelfde metingen ook met camera's gedaan kunnen worden, die minder nauwkeurige metingen opleveren maar wel goedkoper zijn in aanschaf en onderhoud.

Op deze manier is een 'design-time' optimalisatie gedaan waarin aan de hand van een kosten-batenanalyse een investeringsbeslissing kan worden genomen. Er is gebruik gemaakt van een VISSIM-simulatie voor een stuk snelweg met een TDI op de toerit. Zowel de snelheids- als intensiteitsmetingen zijn in de simulatie kunstmatig aangepast door het toevoegen van een ruis of een bias, om zodoende onnauwkeurige metingen te simuleren. Onderzocht is wat het effect van deze onnauwkeurigheden is op het functioneren van de TDI. Twee soorten fouten hebben we op deze wijze bekeken: een structurele fout in het gemiddelde, variërend van -40 tot 40%, en een normaal verdeelde fout met gemiddelde 0 en een standaardafwijking variërend van 0 tot 50%. De figuren 1 tot en met 4 laten het effect van de verschillende fouten zien op de gemiddelde netwerksnelheid, in dit geval de *Measure of Performance*. Uit de figuren blijkt onder andere dat een fout in de intensiteiten een minder grote invloed heeft dan een fout in de snelheidsmetingen.

Met behulp van regressie is voor elk type fout een polynomiaal verband afgeleid: de zogenaamde *utility*-functie (zie de functies in figuur 1-4). Op deze manier is de relatie tussen de nauwkeurigheid van de inwindata en

het uiteindelijke effect op het verkeerssysteem bepaald.

Vervolgens kan, op basis van aannames over de investerings- en onderhoudskosten en de nauwkeurigheden van verschillende meet-systemen, het meest kosteneffectieve meet-systeem gekozen worden. In dit voorbeeld gaan we uit van de volgende kosten voor lusdetectoren: vervangingskosten € 1.548 per paar + € 9.502 per detectorstation, en onderhoudskosten € 633 per jaar per detector. Voor de camera's gaan we uit van een eenvoudige laaggeprijsde camera: vervangingskosten € 3.000, onderhoudskosten € 1.000 per jaar. We nemen vervolgens aan dat beide systemen geen bias hebben op de snelheidsmetingen, maar wel een random error van 5% (lusdetectoren) en 8% (camera's). Beide systemen onderschatten de intensiteit omdat ze voertuigen kunnen 'missen'. Voor de lusdetectoren schatten we de afwijking op de intensiteit op -2% met een *random error* van 5% en voor de camera's nemen we een afwijking op de intensiteit van -10% (door bijvoorbeeld 'overlappende' voertuigen) met een random error van 10%. Nu kunnen we voor elk systeem het effect op de netwerksnelheid en de hoeveelheid voertuigverliezen bepalen, gebruik makend van de *utility*-

ty-functies. In vergelijking met de situatie zonder TDI, worden 14.034 voertuigverliesuren per jaar bespaard met de lusdetectoren en 8.970 voertuigverliesuren per jaar met camera's. Rekenend met een *value of time* van € 11 per uur en uitgaande van de hierboven genoemde kosten per systeem, zijn de camera's € 82.544 per jaar duurder vergeleken met lusdetectoren. Uiteindelijk kunnen we de kosten afzetten tegen de tijd, uitgaande van een configuratie van vier paar lusdetectoren met drie detectorstations of drie camera's. Het resultaat is gevisualiseerd in figuur 5.

Deze figuur laat zien dat het al binnen een jaar kosteneffectiever is om lusdetectoren te gebruiken in plaats van camera's. Een camera heeft echter ook andere eigenschappen die hierbij niet zijn meegenomen, zoals het vermogen om wachtrijlengtes nauwkeurig te meten. Ook kunnen voor een wegbeheerder de daadwerkelijke kosten (aanschaf en onderhoud) belangrijker zijn dan het reduceren van de maatschappelijke kosten.

Voorbeelden van mogelijke effecten

Terug naar het combineren van data, oftewel datafusie. Wat kunnen we daar nu van verwachten? Om dit te illustreren gebruiken we weer voorbeelden: één voor het hoofdwegennet en één voor het stedelijk wegennet.

In de eerste casus is onderzocht tot welke datakwaliteit de combinatie van lusdata en FCD-traces leidt bij verschillende configuraties. Hierbij is met name gekeken naar de afstanden tussen de lusdetectoren en de penetratie van voertuigen die een FCD-trace afgeven. Het simulatiemodel FOSIM is gebruikt om de data te genereren. Meer informatie over datafusiemethoden is te vinden in de thesis *Langrangian Multi-Class State Estimation* van Y. Yuan (2013, TRAIL Thesis Series no. T2013/5).

Figuur 6 toont het resultaat van de berekening voor het schatten van de snelheden op een willekeurige locatie. De grafiek laat duidelijk zien hoe met een beperkte hoeveelheid *floating car data* de schatting van de snelheden aanzienlijk kan worden verbeterd: bij 2% FCD halveert de fout ten opzichte van de 0% FCD-situatie. Tegelijkertijd laat de figuur zien dat in de huidige situatie (alleen lussen om ongeveer iedere 500 m) de fout iets meer dan 6% is. Dezelfde fout kan worden gehaald met veel minder lussen (bijvoorbeeld om de 2500 m), zolang de gegevens worden gecombineerd met FCD (in dit geval, ongeveer 2%). Deze theoretische exercitie geeft daarmee duidelijk weer wat de kansen van datafusie zijn! Maar dergelijke kansen beperken zich niet tot het hoofdwegennet: ook op het onder-

liggende wegennet kan datafusie de huidige beschikbare verkeersgegevens aanzienlijk verbeteren. In een recent onderzoek (Kuwahara e.a., 2013) wordt aangetoond hoe schattingen voor de lengte van wachtrijen, die essentieel zijn voor het goed regelen van het verkeer in een stedelijk wegennet, kunnen worden verbeterd door de lusdata te combineren met FCD. Ook hier blijkt dat de combinatie van de gegevens essentieel is om tot een goede schatting te komen. Figuur 7 geeft een beeld van de toepassing van het concept, waarbij wegwakantlussen (tellingen op de randen) worden gecombineerd met FCD. De trajectoriën van de voertuigen die FCD genereren worden bekend verondersteld, terwijl de tellingen aan de randen van het traject worden gebruikt om in combinatie met een model de dynamiek van de wachtrij te reconstrueren. Dit blijkt tot uitstekende resultaten te leiden waarbij de wachtrijen zeer nauwkeurig kunnen worden geschat.

Tot slot

Er lijkt dus muziek te zitten in datafusie. Aan de ene kant is de noodzaak er: voor goed netwerkmanagement hebben we immers goede data nodig. En er zijn toepassingen waarvoor de huidige datakwaliteit gewoon niet voldoende is! Daarnaast is het in een tijd waarin we op de financiën moeten letten, verstandig om te kijken naar methoden om goedkoper tot zinvolle informatie te komen. Datafusie biedt daar de mogelijkheden toe. Juist nu komen nieuwe databronnen beschikbaar. De bronnen afzonderlijk zijn zeker niet perfect, maar door ze slim te combineren kunnen we toch tot hele nuttige en bruikbare informatie komen.

Deze perspectieven stellen ons niet alleen voor de nodige verkeerskundige uitdagingen. Ook op het gebied van techniek en organisatie moeten we aan de slag om stappen te maken. Maar doen we dat en doen we dat goed, dan vinden we in de wetenschappelijke literatuur voldoende aanleiding om te mogen veronderstellen dat 1+1 inderdaad 3 kan zijn. ●

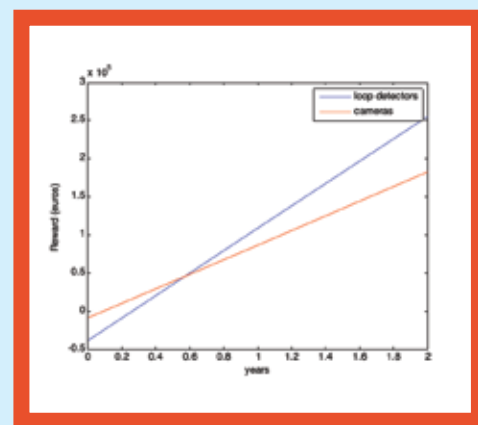
De auteurs

Drs. Gerdien Klunder is PhD-onderzoeker bij TU Delft en ITS-consultant bij TNO.

Prof. dr. ir. Serge Hoogendoorn is hoogleraar Verkeersstromen en Dynamisch Verkeersmanagement, verbonden aan de TU Delft en TrafficQuest.

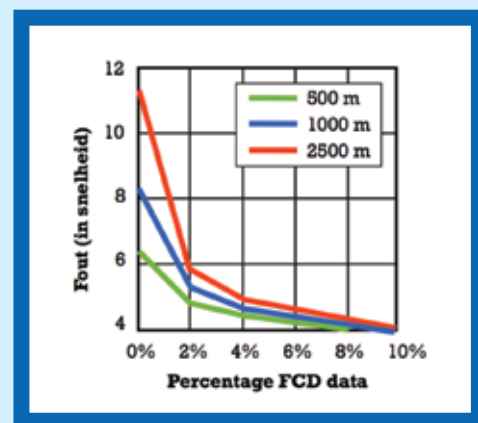
Dr. ir. Leon Kester is senior onderzoeker bij TNO.

Dr. ir. Henk Taale is senior adviseur bij Rijkswaterstaat WVL, docent aan de TU Delft en verbonden aan TrafficQuest.



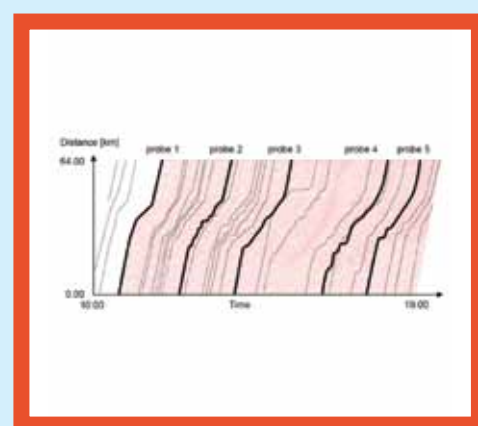
Figuur 5

Een rekenvoorbeeld van de kosteneffectiviteit van lusdetectoren (blauw) en camera's (rood).



Figuur 6

Relatie tussen fout in schatting snelheid en aandeel FCD-voertuigen.



Figuur 7

Voorbeeld van de reconstructie van de voertuigtrajectoriën door het combineren van tellussen en floating car data. De figuur laat duidelijk zien hoe de wachtrijen ter hoogte van een VRI ($x = 48$) wordt gereconstrueerd met slechts een beperkt aantal FCD-traces.