



Nationale Databank
Wegverkeersgegevens

Auteurs: M.G. Uenk-Telgen, O. Vroom, M.I. Hovestad (NDW)

Samenvatting

Datasets worden vaak voor een enkele toepassing ingekocht en gebruikt. Dit wordt veroorzaakt doordat vanuit een vraag wordt gekeken welke data er nodig is en vervolgens deze data wordt gezocht. Met de reeds beschikbare data is echter vaak veel meer mogelijk dan op het eerste gezicht gedacht wordt. In dit paper laten we aan de hand van twee voorbeelden zien dat attributen die niet aanwezig zijn in een dataset zoals de V85 waarde of de uitstroom in een netwerk een relatie kunnen hebben met een attribuut dat wel aanwezig is. Met deze gevonden relatie is het toch mogelijk om vanuit een bestaande dataset compleet andere toepassingen te bedienen. Dit soort inzichten in het gebruik van data ontstaan wanneer niet langer wordt gezocht naar een dataset die past bij de onderzoeksvraag, maar eerst wordt onderzocht of reeds beschikbare datasets de vraag kunnen beantwoorden. Deze aanpak kan uiteindelijk veel tijd en geld besparen en maakt het mogelijk om datasets voor compleet andere toepassingen te gebruiken dan aanvankelijk was beoogd.

Voorwaarden voor het slagen van een dergelijke aanpak is dat er inzicht is in de beschikbare datasets en inzicht in de inhoud van de dataset.

Ieder zijn eigen dataset?

Inleiding

In het verkeersdomein worden steeds meer data analyse uitgevoerd. Enerzijds om beslissingen te ondersteunen; kan hier een rotonde worden geplaatst, zijn snelheid reducerende maatregelen hier gewenst en anderzijds om inzichten te verwerven in de huidige situatie. Vanuit elk specialisme in het verkeersdomein zoals geluid, beheer en onderhoud, verkeersmanagement, verkeersveiligheid, milieu etc worden andere analyses met (verkeers)data uitgevoerd. Veelal wordt dan ook vanuit elke specialisme gewerkt met een eigen dataset die nieuw wordt ingekocht of opgezet. We zijn dan ook al meer dan eens tegengekomen, dat binnen een overheidsorganisatie er vanuit verschillende specialisme net iets andere datasets zijn ingekocht die allemaal zijn opgebouwd uit dezelfde brongegevens. (Dit betekent dat bijvoorbeeld uw collega van verkeersveiligheid een contract heeft met een leverancier die het aandeel overschrijders rapporteert, terwijl u zelf onafhankelijk van uw collega een contract met deze leverancier heeft en de gemiddelde snelheden gerapporteerd krijgt. Op deze manier betaalt uw organisatie waarschijnlijk twee keer voor dezelfde brongegevens). Het voor elke toepassing opnieuw inkopen dan wel opzetten van datasets is een kostbaar en tijdrovend proces. Door inzicht te hebben in de beschikbare datasets en met enige creativiteit en kennis van data science kan veel tijd en geld worden bespaard. In dit paper beschrijven we welke aspecten van een dataset inzichtelijk moeten zijn en tonen we aan de hand van twee voorbeelden hoe de creativiteit en kennis van data science kan worden gebruikt.

Inzicht in datasets

Het gebruiken van datasets voor andere toepassingen is alleen mogelijk als binnen een organisatie(s) inzichtelijk is welke datasets er beschikbaar zijn. Hierdoor kan er efficiënter gebruik worden gemaakt van de beschikbare datasets. Het is van belang dat er één bekende locatie is waar datasets worden geplaatst. Alleen al door het centraal opslaan van datasets (binnen een organisatie) kan het dubbel inkopen van dezelfde datasets worden voorkomen. Daarnaast dienen datasets te worden voorzien van een bijsluiters. Afhankelijk van de grootte van de organisatie en het aantal beschikbare datasets dient de opslag en bijsluiters gestandaardiseerd te worden zodat met eenvoudige zoekfuncties data kan worden gevonden.

In een bijsluiters dient tenminste informatie beschikbaar te zijn over de betekenis van verschillende attributen, een eigenaar van de dataset en informatie over hoe de dataset is verkregen aanwezig te zijn. Met informatie over de attributen is deze dataset ook toepasbaar voor andere onderzoeken die vergelijkbare informatie nodig hebben. Door het toevoegen van informatie over hoe de data verkregen is, is het ook mogelijk om te kijken naar andere toepassingen die op het eerste gezicht niet voor de hand liggen (zoals onderstaande voorbeelden). Een eigenaar van de data kan informatie geven over de houdbaarheid van data of een datalevering wordt gecontinueerd of dat dit een mogelijkheid is.

Data science

Onder data science verstaan we het halen van kennis uit data of het ondersteunen van het data gedreven beslissingsproces. Voor het goed toepassen van data science is dat de inhoud van de dataset bekend is. Dit benadrukt nogmaals het belang van het inzichtelijk hebben van de beschikbare datasets en hun inhoud. Maar met inzicht in de datasets en kennis van data science is natuurlijk nog niet elke verkeersdata geschikt voor elke verkeers-toepassing. Wel is er met creativiteit en kennis van data science is er veel meer mogelijk dan op het eerste gezicht wordt gedacht. Dit tonen we aan de hand van twee voorbeelden. In deze twee voorbeelden wordt gebruik gemaakt van een en dezelfde dataset. Het betreft de Floating Car Data (FCD) levering aan de Nationale Databank

Wegverkeergegevens (NDW). NDW koopt sinds februari 2017 landelijk FCD snelheden in voor het gehele wegennet. Dit betreft gemiddelde snelheden over segmenten van het wegennet die elke minuut worden geleverd. Hiervoor is het gehele wegennet opgedeeld in segmenten van maximaal 50 meter. NDW ontvangt elke minuut voor elk segment een gemiddelde snelheid, tenzij er de afgelopen 30 minuten geen verkeer is gemeten op het betreffende segment. Het geleverde snelheid is het gemiddelde van de laatste (maximaal) 10 metingen binnen de afgelopen 30 minuten. Het aantal metingen waarop dit gemiddelde is geleverd wordt bijgeleverd.

Deze FCD snelheden zijn ingekocht voor verkeersmanagement toepassingen. Voor deze toepassingen is het detecteren van vertragingen het belangrijkste, daarom waren de kwaliteitscriteria voor deze data gericht op het juist en accuraat detecteren van de vertragingen. Andere aspecten zoals bijvoorbeeld de nauwkeurigheid van de freeflow snelheid waren minder van belang en hadden daarom veel zwakkere kwaliteitseisen. Dit heeft er toe geleid dat er aan deze dataset enkele beperkingen zitten die de dataset op het eerste gezicht minder geschikt maakt voor andere toepassingen. De belangrijkste beperkingen van de deze dataset zijn:

- Gemiddelden: de dataset bevat enkel gemiddelde snelheden en geen individuele snelheden
- Afkapping: alle individuele snelheden zijn afgekapt op de maximum snelheid voor ze zijn gemiddeld
- Metingen kunnen in meerdere minuten worden gebruikt: Omdat de laatste 10 metingen van maximaal 30 minuten geleden worden gebruikt kan het voorkomen dat enkele metingen meerdere keren worden gebruikt. In het slechtste geval wordt er 1 minuut 1 of meerdere voertuigen gemeten en vervolgens 30 minuten niks. In dat geval wordt er 30 minuten lang dezelfde snelheid doorgegeven.

Voorbeeld 1: Verkeerveiligheidsindicator

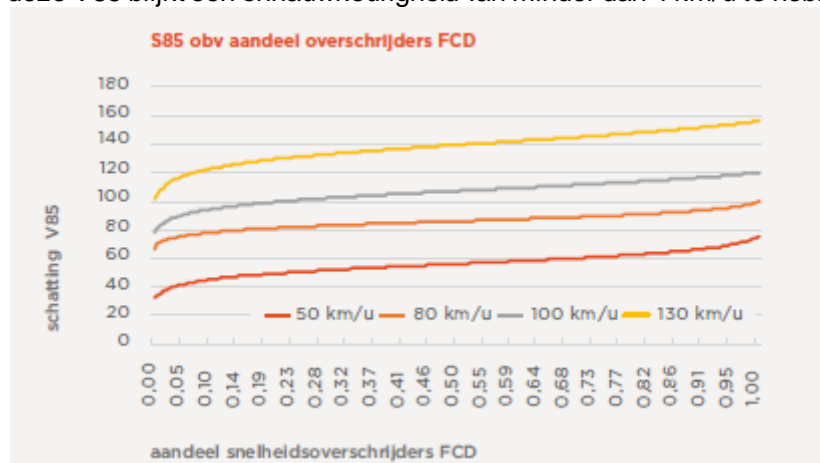
Een belangrijke indicator op het gebied van verkeerveiligheid is de V85 waarde. De V85 is de snelheid waar 85% van het verkeer onder blijft. Normaal gesproken wordt deze waarde bepaald door alle individuele snelheden van laag naar hoog te sorteren en de waarde op 85% van deze rij aan te wijzen. Deze individuele snelheden worden vaak gemeten met lussen. Een nadeel van lussen is dat we alleen informatie hebben van de wegen waar een dergelijke lus aanwezig is. De FCD geeft ons

informatie over het gehele wegennet, daarom is onderzocht of vanuit de FCD levering een V85 te bepalen is.

De FCD is op het eerste gezicht niet direct geschikt om de V85 te bepalen, omdat we ten eerste geen individuele snelheden hebben, ten tweede omdat de snelheden zijn afgekapt op de maximum snelheid en ten derde omdat metingen meerdere minuten worden vastgehouden en hierdoor zwaarder mee kunnen tellen.

Toch hebben we een methode gevonden om V85 te bepalen uit FCD. Na onderzoek blijkt er een relatie te zijn tussen het aantal minuten waarop een FCD snelheid boven de 96% van de maximum snelheid wordt geleverd en de V85. Hiervoor is op meer dan 200 diverse locaties FCD snelheden vergeleken met lusdata en de bijbehorende V85 waarde. Onderstaande figuur toont de gevonden relatie tussen de FCD snelheden en de V85 waarde. Hier is te zien dat op een weg met een maximum snelheid van 50 km/u waar 60% van de tijd een FCD minuut snelheid wordt geleverd boven de $0,96 \cdot 50 = 48$ km/u, dat de V85 naar schatting 60 km/u is.

Met deze relaties is het mogelijk om de V85 voor het gehele wegennet met een vaste snelheidslimiet ≥ 50 km/u in te schatten, zonder dat we hiervoor nieuwe specifieke V85 data hoeven in te kopen. De inschatting van deze V85 blijkt een onnauwkeurigheid van minder dan 4 km/u te hebben.

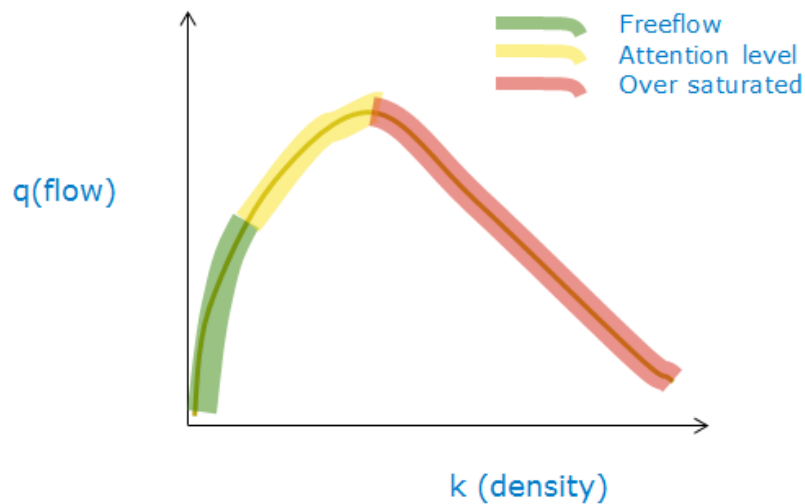


Figuur 1 - Relatie tussen aandeel maximumsnelheid in FCD en de V85

Voorbeeld 2: Netwerk fundamenteel diagram

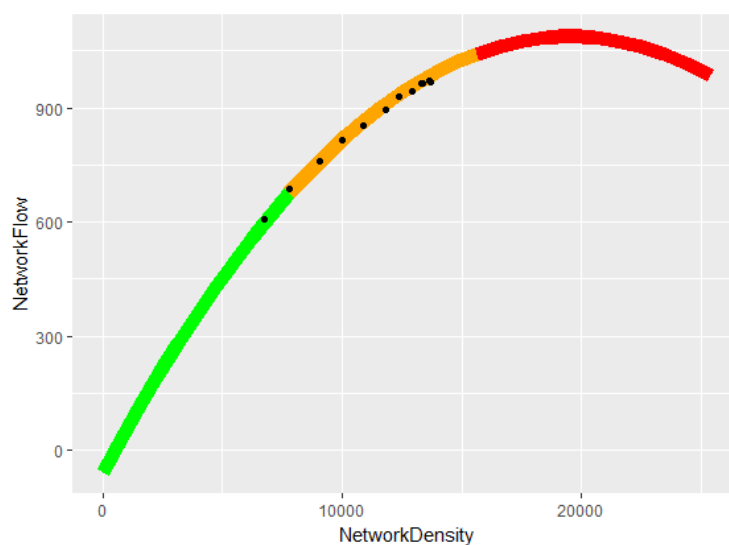
Het netwerk fundamenteel diagram geeft informatie over de huidige situatie op het netwerk en de nog beschikbare capaciteit in het netwerk. Waarbij het netwerk een verzameling wegen is. Figuur 2 geeft een voorbeeld van een dergelijk fundamenteel diagram. Zolang de toestand van het netwerk zich nog aan de linkerkant van de top bevindt presteert het netwerk nog goed, maar hoe dicht er bij de top hoe vaster het netwerk gaat lopen en aan de rechterkant zijn de prestaties van het netwerk slecht door terugslag van verkeer. In het netwerk fundamenteel diagram is de uitstroom van het netwerk uitgezet tegen de dichtheid (verhouding tussen intensiteit en snelheid) in het netwerk.

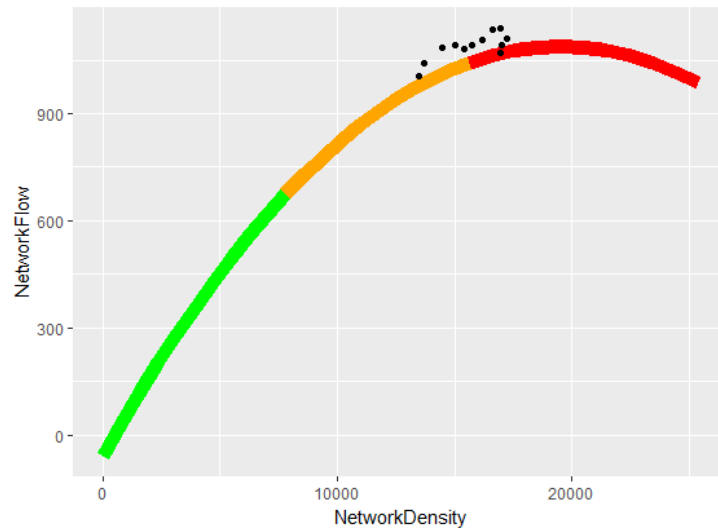
In de FCD levering is geen informatie beschikbaar over de intensiteiten in het netwerk, hierdoor is het op het eerste gezicht niet mogelijk om een fundamenteel diagram op basis van de FCD te bepalen. We zouden informatie vanuit vaste meetlocaties kunnen gebruiken om de intensiteiten in te vullen, maar op deze manier kunnen we alleen netwerk fundamenteel diagrammen opstellen van voor gedefinieerde netwerken (waar meetsystemen aanwezig zijn).



Figuur 2 – Voorbeeld van een netwerk fundamenteel diagram

Toch hebben we een methode gevonden waardoor het mogelijk is om een aangepast netwerk fundamenteel diagram te bepalen. Onderzoek heeft getoond dat het aantal metingen in de FCD levering over het netwerk zich hetzelfde gedraagt als de uitstroom van het netwerk. Ook de verhouding tussen dit aantal metingen en de snelheid gedraagt zich hetzelfde als de dichtheid. Hoewel de absolute waarde van dit totaal aantal metingen en aangepaste dichtheid geen betekenis heeft, hebben de relatieve aantallen binnen een netwerk dat wel. Een hoog aantal metingen betekent daadwerkelijk veel uitstroom en vice versa. Hierdoor ontstaat eenzelfde vorm bij gebruik van het aantal metingen en de aangepaste dichtheid in dit aangepast netwerk fundamenteel diagram. Figuur 3 toont dit aangepaste diagram voor een netwerk in Delft om 7 en 8 uur 's ochtends (de zwarte punten). De vorm van het aangepaste netwerk fundamenteel (de gekleurde boog) is verkregen op basis van slechts 2 dagen historie. Goed te zien is dat het gedurende deze ochtendspits het drukker in het netwerk wordt en minder zal gaan presteren.





Figuur 3 – Voorbeeld van het aangepaste netwerk fundamenteel diagram om 7 en 8 uur ‘s ochtends

Conclusie

Met data is vaak veel meer mogelijk dan op het eerste gezicht gedacht wordt. In elke dataset is maar een beperkt aantal attributen opgenomen, maar vaak hebben de beschikbare attributen ook relaties met niet aanwezige attributen. Zoals bijvoorbeeld het geval was met de V85 en het aantal minuten met de maximum snelheid en het aantal metingen en de uitstroom. Het vinden van dit soort relaties is alleen mogelijk als inzichtelijk is welke datasets beschikbaar zijn en inzichtelijk is wat er in staat. Dit soort inzichten in het gebruik van data ontstaan wanneer niet langer wordt gezocht naar een dataset die past bij de onderzoeksvraag, maar eerst wordt onderzocht of reeds beschikbare datasets de vraag kunnen beantwoorden. Deze aanpak kan uiteindelijk veel tijd en geld besparen en maakt het mogelijk om datasets voor compleet andere toepassingen te gebruiken dan aanvankelijk was beoogd.