

# Identificatie en monitoring van DNA-groepen in ongevallen- en partijdata

Sjoerd Broos, Daan Kolkman, Giedo Donkers · 26 november 2020

**Met data science-technieken leiden we het DNA van natuurlijke groepen ongevallen en partijen af, zodat gemeenten gelijksoortige ongevallen over de jaren heen kunnen monitoren.**

---

*Resultaat en methode tot het groeperen en indelen van ongevallen en partijen in groepen.*

S. Broos (VIA), D. Kolkman (MKB datalab), G. Donkers (VIA)

## Samenvatting

Met behulp van data science-technieken is het DNA van natuurlijke groepen ongevallen en partijen te bepalen. Gemeten kan worden hoe gelijksoortige ongevallen en betrokken partijen zich over de jaren heen ontwikkelen: monitoring. Het kan per gemeente inzichtelijk maken wat voor groep gelijksoortige ongevallen over de jaren heen juist veel of weinig voorkomt. Er kan vergeleken worden met hetzelfde soort gemeentes, gemeentes met dezelfde problemen of met heel Nederland. Daarnaast laat het zien wat voor gelijksoortige partijen zich voordoen bij de ongevallen. Dit onderzoek gebruikt “unsupervised leren” zoals clustering om groepen af te leiden en heuristieken om DNA te bepalen.

## Inleiding

### Data science

Data science (datawetenschap) wordt steeds populairder en Davenport en Patil (2012) hebben het beroep van data wetenschapper zelfs uitgeroepen tot de meest sexy baan van deze eeuw. Met nieuwe technieken kan data gebruikt worden om nieuwe inzichten te krijgen. Deze inzichten zijn divers, bijvoorbeeld het voorspellen van de vraag (bv. brood) voor een bakker of een meer complex vraagstuk als het aanbevelen van producten op een webshop. Data wordt uitgelezen, schoongemaakt, hervormd en vervolgens gebruikt om een specifiek doel te bereiken. Een aantal technieken die gebruikt zijn voor dit onderzoek worden in de volgende paragrafen uitgelicht.

## Supervised leren

Allereerst bestaat er supervised leren; hier probeert men de relatie tussen input en output te leren (Kaptein, 2019a). De input is een set van variabelen (ook wel features of kenmerken genoemd) waar een output tegenover staat. Neem bijvoorbeeld een foto waarvan je weet dat het een hond is. De input zijn dan de pixels van de foto en de output is de groep/categorie hond. Een ander voorbeeld is het voorspellen van huisprijzen. Als input neem je de eigenschappen of kenmerken van het huis en als output neem je de huisprijs. Deze voorbeelden zijn te categoriseren als classificatie- en regressieproblemen. Dit onderzoek focust op een classificatieprobleem.

## Unsupervised leren

Unsupervised leren is een ander type van machine learning. Hier heb je wel een input, maar geen output (Kaptein, 2019b). Hier heb je bijvoorbeeld de kenmerken van gebruikers op een online webshop, maar weet je niet in welke groep gebruikers deze precies vallen. Er kunnen actieve en niet-actieve gebruikers zijn, maar ook meer complexe relaties zoals gebruikers die geïnteresseerd zijn in bepaalde productsegmenten. Deze data kan gebruikt worden om natuurlijke groepen te ontdekken in de data. Hiervoor worden clustering-algoritmes gebruikt. Om het clustering-algoritme te optimaliseren (qua looptijd), wordt regelmatig gebruik gemaakt van dimensiereductie-algoritmes. Deze dimensiereductie-algoritmes kunnen het aantal kenmerken (ofwel: features, variabelen) met slimme technieken reduceren.

## Verkeerskundig probleem

Het idee is om op basis van data-analyse gemeentes de mogelijkheid te geven om gelijksoortige ongevallen en bijbehorende partijen gerichter aan te pakken. Het uiteindelijke doel is om instanties aan te zetten tot maatregelen. Zo kan een gemeente in de toekomst naar andere soortgelijke gemeentes kijken die effectief gelijksoortige ongevallen aanpakken. De algemene term hier is monitoring: door de tijd heen de status van gemeentes bepalen op basis van verschillende kenmerken. Door het in kaart brengen en monitoren van de verschillende soorten DNA (groepen) kan de gemeente niet alleen inzichtelijk krijgen wat voor complexiteit aan ongevallen en partijen er in zijn gemeente plaatsvindt, maar ook effectiever en efficiënter te werk gaan bij het doelmatig reduceren van het aantal ongevallen. Dit leidt tot de volgende onderzoeksvraag die we zullen beantwoorden in deze white-paper:

*Vallen ongevallen & partijen in natuurlijke groepen (DNA)? Zo ja, kunnen we deze groepen (DNA) over de jaren heen monitoren?*

## Data science en het verkeer

### Data

Voor dit onderzoek is gebruik gemaakt van een deel van de STAR-database[1] met ongevallen- en partijdata. Deze data omvat kenmerken van een ongeval[2] en kenmerken van de betrokken partijen[3] (alleen bestuurders, voetgangers en objecten) bij een ongeval. Mocht bij een ongeval een kenmerk onbekend zijn, dan is voor dat ongeval geen DNA af te leiden. Dit geldt ook op het moment dat er een ongeval is waarbij alleen objecten of dieren voorkomen. Hierdoor hebben we een dataset van 205.653 ongevallen en 520.284 partijen voor de jaren 2017 – 2019.

## Data voorbereiding en vervolg

Om de onderzoeksvraag te beantwoorden is de hierboven beschreven data geanalyseerd. Sommige variabelen zijn ordinaal (volgorde is belangrijk) en sommige nominaal (volgorde is niet belangrijk). De nominale variabelen zijn uitgesplitst (one-hot encoding | dummies) en de ordinale variabelen zijn numeriek gemaakt.

Voor het bepalen van DNA in de data is er gekozen voor clustering (unsupervised leren). Deze clustering kan verschillende soorten ongevallen en partijen in groepen onderverdelen. Deze groepen (clusters) hebben ieder een eigen betekenis. Deze betekenis kan later achterhaald worden. Door de uitsplitsing (one-hot encoding) van een aantal kenmerken zijn er veel dimensies ontstaan. Om het clusteren makkelijker te laten verlopen (kortere looptijd van het algoritme) is er gekozen voor een dimensiereductie. Deze dimensiereductie reduceert het aantal kenmerken naar een X-aantal dimensies (waar X staat voor een zelf te kiezen aantal dimensies).

De dimensiereductiemethode is principal component analysis (PCA); deze transformeert en combineert lineaire relaties naar een gespecificeerd X-aantal dimensies. Vervolgens zijn de overgebleven X-dimensies geclusterd. Voor de clustering is gekozen voor het K-means-algoritme (zowel voor de ongevallen- als partijdata) (MacQueen, 1967). Na de clustering is de data getransformeerd om de belangrijkste (dominantste) kenmerken per groep te achterhalen. Hiervoor is een heuristiek gebruikt. De kracht van de heuristiek is om domein- en case-specifieke kennis toe te voegen. In dit geval is de uitkomst de dominante variabelen die een cluster definiëren.

## Resultaten

Deze sectie beschrijft de resultaten van het onderzoek. Allereerst zal de dimensiereductie besproken worden, daarna de clustering, vervolgens het DNA en als laatste de monitoring.

### Dimensiereductie

De eerste stap in het proces is de dimensiereductie. De variabelen zijn gereduceerd naar 25 dimensies. De gekozen 25 dimensies gaven de beste resultaten gekeken naar de uiteindelijke uitkomst (de betekenis van de groepen).

## Clustering

De data van de dimensiereductie wordt vervolgens gebruikt om te clusteren. Het doel is om groepen ongevallen en partijen te kunnen onderscheiden en te monitoren. Allereerst zijn de ongevalclusters afgeleid met het K-means-algoritme. Hiervoor is  $k$  (het aantal clusters) van 40 bepaald aan de hand van verschillende prestatie-indicatoren:

- **Inertia:** geeft de afstand binnen het cluster aan; deze score moet zo laag mogelijk uitvallen.
- **Calinski-Harabasz index:** geeft de ratio tussen de afstand binnen het cluster en tussen clusters. Deze score moet zo hoog mogelijk uitvallen.
- **Davies-Bouldin:** geeft aan hoe ver clusters van elkaar afliggen. Deze score moet zo laag mogelijk uitvallen.

Voor de partijdata is ook het K-means-algoritme gebruikt met  $k=48$ . Deze zijn ook gebaseerd op de hierboven benoemde prestatie-indicatoren. Daarnaast is de dominantie (betekenis) van het DNA ook meegewogen. Belangrijk in de overweging is dat kleine gebieden ook de mogelijkheid moeten hebben om een cluster te monitoren. Het moet natuurlijk niet zo zijn dat er zoveel clusters zijn dat elk cluster een individueel ongeval bevat.

## DNA

Iedere groep (cluster) heeft zijn eigen betekenis. Het ene cluster kan veel ongevallen bevatten op een kruispunt, terwijl het andere cluster veel ongevallen bevat die juist niet op een kruispunt zijn gebeurd. Ditzelfde geldt voor de partijen: het ene cluster kan bestaan uit uitsluitend fietsers en het andere uit uitsluitend voetgangers. Om te achterhalen wat ieder cluster betekent, is er gebruik gemaakt van een heuristiek. Er zijn meerdere dominantiemethodes afgewogen, zoals classificatie (supervised leren) met het random forest-algoritme of alleen visualisaties met bv. een heatmap. Uiteindelijk is de heuristiek ontwikkeld om domein- en case-specifieke kennis mee te kunnen nemen. In de heuristiek zijn "actieve" variabelen en "illustratieve" variabelen meegenomen. De actieve variabelen zijn opgenomen om de clusters af te leiden en de illustratieve zijn alleen bedoeld om de dominantie te bepalen. Twee voorbeelden van de ongevalen- en partijclusters zijn weergegeven in tabel 1 en 2. Belangrijk om hierbij rekening te houden is dat er gekeken is naar hetzelfde soort ongevallen. Het hoeft dus niet te betekenen dat elk ongeval exact hetzelfde is (anders krijg je oneindig veel clusters). Hierdoor zijn termen als "vaak", "heel vaak", "regelmatig", etc. erg belangrijk. Dit laat de dominantie zien van een bepaald kenmerk binnen een cluster. Als voorbeeld: een cluster met de betekenis "vaak 50 km/u-wegen" wil zeggen dat het binnen dit cluster vaak voorkomt dat ongevallen op 50 km/u-wegen plaatsvinden, echter dit is niet altijd het geval.

*Tabel 1: Ongevalcluster vergelijking*

| Cluster 9                                                                     | Cluster 22                                   | Cluster 25                                      |
|-------------------------------------------------------------------------------|----------------------------------------------|-------------------------------------------------|
| Mix kruispunt/geen kruispunt                                                  | Vaak niet op een kruispunt                   | Vaak op een kruispunt                           |
| In de avonden en ochtenden (voor de spits)                                    | Overdag                                      | Overdag                                         |
| Regelmatig heviger ongevallen                                                 | Regelmatig heviger ongevallen                | Vaak UMS-ongevallen                             |
| Vaak 50-wegen                                                                 | Vaak 50 km/u-wegen                           | Regelmatig 60, 80-wegen                         |
| Vaak duisternis                                                               | Heel vaak daglicht                           | Heel vaak daglicht                              |
| Vaak overig asfalt                                                            | Vaak overig asfalt                           | Altijd overig asfalt                            |
| Heel vaak regen                                                               | Heel vaak droog                              | Altijd droog                                    |
| Heel vaak 2 partijen                                                          | Heel vaak 2 partijen                         | Heel vaak 2 partijen                            |
| Botsing langzaam – snelverkeer                                                | Botsing langzaam – snelverkeer               | Botsing snelverkeer met een object              |
| Hele jaar door met pieken in kwartaal 1 en 4                                  | Hele jaar door met pieken in kwartaal 2 en 3 | Hele jaar door met pieken in kwartaal 2         |
| Vaak alle leeftijden bij betrokkenen waarbij 1 van de partijen een jongere is | Altijd volwassenen bij betrokkenen           | Alle leeftijden die tegen een object aan botsen |

Tabel 2: Partijcluster vergelijking

| Cluster 1                        | Cluster 3                        | Cluster 11                       |
|----------------------------------|----------------------------------|----------------------------------|
| Fietsers                         | Motorfietsen                     | Auto's en busjes                 |
| Leeftijd tussen de 58 en 70 jaar | Leeftijd tussen de 18 en 21 jaar | Leeftijd tussen de 69 en 76 jaar |

## Monitoring

De gegeven clusters en hun bijbehorende DNA kunnen over de tijd heen gemeten worden, zodat instanties binnen hun gebied beter zicht krijgen op wat er speelt en dit kunnen analyseren. Dit kan bestempeld worden als monitoring[4]. Tabel 3 geeft inzicht in de cluster grootte (absolute en relatieve waarden) over de tijd heen voor verschillende gebieden: (1) Nederland, (2) 's-Hertogenbosch, (3) Vught. In tabel 3a is te zien dat cluster 22 absoluut en relatief toeneemt. Cluster 25, aan de andere kant, neemt absoluut en relatief af. Cluster 9 is stabiel over de jaren heen.

Voor een instantie kan het interessant zijn om zijn gegevens te vergelijken met een soortgelijke instantie of met de trend in heel Nederland. Tabel 3b toont de cluster grootte voor de gemeente 's-Hertogenbosch. Hierin is te zien dat cluster 9 stabiel is en cluster 22 over de jaren heen schommelt. Cluster 25 is afgenomen. Dus cluster 9 en 25 ontwikkelen zich hetzelfde in 's-Hertogenbosch als in heel Nederland.

In Vught zijn de aantallen ongevallen (veel) lager (tabel 3c). Echter, we kunnen op sommige clusters wel

monitoren. Cluster 9 heeft lage aantallen en schommelt over de jaren heen. Hier kun je (op dit moment) niet goed op monitoren. Echter, cluster 22 en 25 laten een duidelijkere trend zien. Cluster 22 neemt sterk toe, terwijl cluster 25 schommelt. Dus cluster 22 en 25 ontwikkelen zich hetzelfde in Vught als in heel Nederland.

*Tabel 3a: Monitoring ongevallen Nederland per cluster (absoluut: links, relatief: rechts)*

| Cluster | 2017 | 2018 | 2019 | 2017 | 2018 | 2019 |
|---------|------|------|------|------|------|------|
| 9       | 995  | 980  | 1109 | 1,5  | 1,4  | 1,6  |
| 22      | 2448 | 2823 | 2815 | 3,6  | 4,1  | 4,1  |
| 25      | 3212 | 3074 | 2832 | 4,7  | 4,5  | 4,1  |

*Tabel 3b: Monitoring ongevallen instantie 's-Hertogenbosch per cluster (absoluut: links, relatief: rechts)*

| Cluster | 2017 | 2018 | 2019 | 2017 | 2018 | 2019 |
|---------|------|------|------|------|------|------|
| 9       | 10   | 10   | 10   | 1,6  | 1,7  | 1,7  |
| 22      | 28   | 31   | 22   | 4,3  | 5,4  | 3,7  |
| 25      | 30   | 16   | 17   | 4,7  | 2,8  | 2,9  |

*Tabel 3c: Monitoring ongevallen instantie Vught per cluster (absoluut: links, relatief: rechts)*

| Cluster | 2017 | 2018 | 2019 | 2017 | 2018 | 2019 |
|---------|------|------|------|------|------|------|
| 9       | 2    | 0    | 2    | 2,4  | 0,0  | 2,1  |
| 22      | 0    | 4    | 6    | 0,0  | 4,0  | 6,2  |
| 25      | 3    | 2    | 3    | 3,6  | 2,0  | 3,1  |

Op dit moment zijn er 3 “random” clusters gekozen om een gevoel te krijgen voor de betekenis van een cluster. Echter, betere analysestrategieën zijn mogelijk op bijvoorbeeld gemeentelijk niveau. Een instantie zou kunnen focussen op groepen die snel groeien of juist op de grootste groepen over meerdere jaren heen. De strategie is ook afhankelijk van de gemeente. Om het voor kleine gemeentes interessant te houden, is er een balans gezocht tussen het aantal clusters en de hoeveelheid ongevallen binnen het cluster. Met een dergelijke analyse kunnen instanties bepaalde ongeval-typen (DNA) identificeren, monitoren en (uiteindelijk) maatregelen treffen.

## Monitoring met een strategie

### Strategie: grootste groepen, instantie: gemeente Vught

In de gemeente Vught vallen relatief weinig ongevallen en is monitoring af en toe lastig. Echter, een

strategie zoals het pakken van de grootste groepen valt wel onder de mogelijkheden. Clusters 2 (20 ongevallen[5]), 21 (21 ongevallen[5]) en 36 (15 ongevallen[5]) zijn in totaal goed voor 20% van alle ongevallen in de jaren 2017-2019 in de gemeente Vught (ter referentie: dit zijn maar 3 van de 40 clusters, oftewel 7,5%). Cluster 2 zijn voornamelijk ongevallen die vaak niet op een kruispunt voorkomen, overdag met daglicht, vaak 50 km/u-wegen, regelmatig heviger, vaak overig asfalt, heel vaak 2 partijen, en langzaam verkeer dat botst met snelverkeer (figuur 1).

*Figuur 1: Voorbeeld hoe de ongevallen van cluster 2 eruit zouden kunnen zien.*

*Figuur 2: Voorbeeld hoe de ongevallen van cluster 21 eruit zouden kunnen zien.*

*Figuur 3: Voorbeeld hoe de ongevallen van cluster 36 eruit zouden kunnen zien.*

Cluster 21 zijn voornamelijk ongevallen die overdag met daglicht gebeuren, regelmatig heviger, regelmatig 30 km/u-wegen, heel vaak op klinkers, heel vaak droog, vaak 2 partijen, heel vaak een botsing tussen langzaam en snelverkeer (figuur 2). Cluster 36 zijn voornamelijk ongevallen die vaak op kruispunten gebeuren, heel vaak droog, vaak overdag met daglicht, regelmatig 70, 80 of 100 km/u-wegen, vaak UMS, vaak overig asfalt, regelmatig 3 of meer partijen aanwezig, vaak een kettingbotsing/kop-staart (figuur 3).

Door deze strategie kan geconcludeerd worden dat er problemen zijn met langzaam en snelverkeer (botsingen) en botsingen met meerdere snelverkeer-partijen in de gemeente Vught. Hierop zouden maatregelen genomen kunnen worden.

## **Strategie: grootste stijger, instantie: gemeente 's-Hertogenbosch**

In de gemeente 's-Hertogenbosch vallen relatief meer ongevallen waardoor monitoring bijna altijd mogelijk is. In dit geval is gekeken naar de grootste stijger over de jaren 2017 t/m 2019. Hierin zien we 2 clusters: cluster 12 (26 ongevallen in 2017, 37 ongevallen in 2018, en 38 ongevallen in 2019) en cluster 33 (8 ongevallen in 2017, 10 ongevallen in 2018, en 20 ongevallen in 2019).

*Figuur 4: Voorbeeld hoe de ongevallen van cluster 12 eruit zouden kunnen zien.*

*Figuur 5: Voorbeeld hoe de ongevallen van cluster 33 eruit zouden kunnen zien.*

Cluster 12 zijn voornamelijk ongevallen overdag in daglicht, vaak UMS, altijd overig asfalt, vaak droog, vaak 1 partij, vaak snelverkeer, alle leeftijden (figuur 4). Cluster 33 zijn voornamelijk ongevallen overdag in daglicht, heel vaak UMS, vaak overig asfalt, altijd regen, vaak ouderen en/of volwassenen betrokken, vaak 2 partijen, botsing tussen snelverkeer en snelverkeer (figuur 5). Duidelijk is dat er een toenemend probleem is met eenzijdige snelverkeer-ongevallen overdag en botsingen tussen snelverkeer en snelverkeer in de regen in de gemeente 's-Hertogenbosch.

## **Conclusies/aanbevelingen**

In dit paper laten we het gebruik van data science in combinatie met verkeersdata zien. We tonen aan dat het mogelijk is om door middel van unsupervised leren een uniek profiel of “DNA” te onderscheiden in de data over ongevallen en de betrokken partijen. De clusters met vergelijkbaar DNA kunnen door wegbeheerders gemonitord worden, om zo opvallende ontwikkelingen te signaleren, in kaart te brengen en waar nodig te interveniëren. Door de clusters op te nemen in de software maakt VIA Software het mogelijk voor wegbeheerders om hier zelf mee aan de slag te gaan. Dit onderzoek demonstreert de meerwaarde van data science in het verkeersdomein, maar is slechts een eerste stap. De vervolgstappen zijn tweedelig. Ten eerste zal er gekeken worden naar het verbeteren van modellen door bv. het onderzoeken van (nieuwe) modellen, zoals gaussian mixture models (GMMs) of Birch (Zhang, et al. 1996), andere dimensiereducties zoals UMAP (McInnes & Healy, 2018), de verdere specificatie van dominantiebepaling en het gebruik van meer of minder variabelen. Als tweede zal er gekeken worden naar de toepasbaarheid van het algoritme in de VIA-software, zoals het vergelijken met hetzelfde soort gemeentes of vergelijken met gemeentes die hetzelfde probleem hebben.

## Literatuur/referenties/bronnen

Davenport, T. H., & Patil, D. J. (2012). Data scientist. *Harvard Business Review*, 90(5), 70–76.

Kaptein, M. (2019a, 20 mei). *Hoe doen computers slimme dingen? Zo werkt machine learning*. Geraadpleegd van <https://www.emerce.nl/best-practice/machine-learning-werkt-dankzij-supervised-leren>

Kaptein, M. (2019b, 5 juni). *Hoe doen computers slimme dingen? Ze kunnen ook prima data samenvatten*. Geraadpleegd van <https://www.emerce.nl/achtergrond/hoe-doen-computers-slimme-dingen-ze-kunnen-prima-data-samenvatten>

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297.

McInnes, L., Healy, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints* 1802.03426.

Zhang, T., Ramakrishnan, R., & Livny, M. (1996). BIRCH: an efficient data clustering method for very large databases. *ACM Sigmod Record*, 25(2), 103–114.

## Noten

1. <https://www.star-verkeersongevallen.nl>

2. Voorbeelden van meegenomen ongeval-kenmerken: lichtgesteldheid, wegverharding, maximumsnelheid.

3. Voorbeelden van meegenomen partij-kenmerken: leeftijd, vervoerwijze.

4. Monitoring is het systematisch volgen van de ontwikkelingen over de tijd heen van specifieke groepen met vooraf bepaalde indicatoren. Het wordt gebruikt om te zien wat, wanneer en waarom het speelt (analyse).

5. Data: 2017 t/m 2019.