Titel: De datacleaner
Auteur: G. Derksen; TNO I&T
Datum: 06-04-2009

Inleiding

De datacleaner is een stuk methode waarmee outliers en/of niet werkende sensoren in een sensornetwerk gedetecteerd kunnen worden. Het is een toepassing van een Bayesian Belief Network.

Het specifieke van deze methodiek is dat rekening gehouden wordt met correlaties in de tijd en ruimte. Een meting van een sensor wordt vergeleken met een meting van dezelfde sensor in het verleden en met meetwaarden van naburige sensoren op hetzelfde moment. Met behulp van conditionelekansverdelingen worden de aannemelijkheden van alle mogelijke waarden bepaald. De waarde met de hoogste aannemelijkheid is de meest aannemelijke uitkomst. Indien om een of andere reden een meetwaarde ontbreekt kan de meest aannemelijke uitkomst ook als schatting gebruikt worden.
Is deze aannemelijkheid laag in vergelijking met die van de andere mogelijke uitkomsten dan wordt aangenomen dat er iets mis is met de meting of dat er is iets mis met de sensor. Nader onderzoek/kennis is dan nodig om uitsluitsel te geven. Dit nader onderzoek kan bijvoorbeeld ook met behulp van een BBN, maar dan op een hoger niveau, uitgevoerd worden. De input van dat BBN bestaat uit de aannemelijkheden van aanliggende sensoren en extra informatie zoals mogelijke effecten van gemeenschappelijke oorzaken zoals verkeerd ingestelde parameters een dergelijke. Op deze wijze kan een eventuele gemeenschappelijke oorzaak aangewezen worden.

Deze notitie beschrijft een toepassing van de data cleaner en gaat in op enkele aandachtspunten bij het in de praktijk toepassen hiervan. De toepassing maakt, bij gebrek aan beschikbaarheid van een echt sensor netwerk gebruik van een dataset van Rijkswaterstaat. Deze set bevat stroefheidsgegvens en lijkt qua opbouw enigszins op die van een sensornetwerk.

Aandachtspunten bij een BBN

Een BBN is gebaseerd op (conditionele) kansen die geschat worden met behulp van een representatieve trainingsset. Deze set moet groot genoeg zijn om deze kansen betrouwbaar te kunnen schatten.
Bij een BBN kan gecorrigeerd worden voor een co-variabele. Een co-variabele is een variabele die het effect beschrijft van een externe oorzaak van een verloop in de tijd of ruimte. Voorbeelden zijn veroudering en/of temperatuur.

DVS data

In het kader van onderzoek naar verloop van de stroefheid in relatie tot het aantal voertuigpassages heeft DVS data beschikbaar gesteld die in het kader van de meerjaren onderhoudsplanning van het Rijkswegennet gedurende de periode 2001/2008 ingewonnen zijn. De dataset ziet er globaal als volgt uit

- van elk hm vak van het Rijkswegennet wordt om de 2 jaar de stroefheid bepaald
- van elk vak is het aanlegjaar bekend

- van elk vak is locatie bekend (wegnummer, baan, strook, hm)
- van elk vak is type deklaag bekend
- van elk vak is het globale aantal voertuig passages in een jaar bekend
- van elke meting is de meetdatum bekend

Deze data kan gebruikt worden als simulatie van een sensor netwerk. Het spatiële aspect wordt benaderd door de opeenvolgende hectometervakken terwijl het tijdsaspect overeenkomt met de metingen van de verschillende meetgeneraties
In deze set zijn er twee co-variabelen. De eerste is het totale aantal voertuigpassages en de tweede is het meetmoment in het jaar. In eerder onderzoek is gebleken dat er sprake is
- van een lineair verband tussen de stroefheid en log(totaal aantal voertuig passages)
- van een sinusvormig verloop van de stroefheid binnen een jaar

Analyse DVS data
Om de methodiek te demonstreren zijn de volgende keuzes gemaakt:

- Selectie van de data
    - meetdata van 2001 en 2003
    - type deklaag is ZOAB
- Covariabelen
    - Van de 2 covariabelen wordt gecorrigeerd voor het aantal voertuigpassages.
    - Het seizoenseffect wordt buiten beschouwing gelaten omdat over het algemeen de metingen verricht zijn in de periode dat het seizoenseffect minimaal is
- Trainingsset. Voor deze data is het lastig om een goede trainingsset te selecteren. Bij gebrek aan beter en tijd is er voor gekozen om de gehele dataset als trainingsset te gebruiken. De belangrijkste problemen met betrekking tot de dataset komen voort uit:
    - verschillen in aanlegdatum
        - tussen 2001 en 2003 vindt er soms vernieuwing van de deklaag plaats met als gevolg dat er maar 1 meetgeneratie beschikbaar is en zodoende minder vakken.
        - niet zijn alleen jonge vakken zonder meer te vergelijken met oude vakken. Jonge vakken zijn ook niet echt onderling met elkaar te vergelijken omdat verschillen kunnen ontstaan door verschillende aannemers, omstandigheden ten tijde van productie, verschillen in grondstoffen etc
    - de vakken zijn afkomstig van een groot aantal verschillende wegen. Bovendien worden de wegen niet in zijn geheel gemeten, stukken worden overgeslagen omdat ze bijvoorbeeld in de even jaren worden gemeten.

  Als gevolg van bovenstaande ontstaat er als het ware een lappendeken met als resultaat dat een gedeelte van de ruimtelijke correlatie verdwijnt.
- Kansverdelingen. Het BBN is gebaseerd op 3 kansverdelingen; de eerste geeft de kans op een meetwaarde, de tweede beschrijft de relatie met het verleden en de derde geeft de relatie met de buren

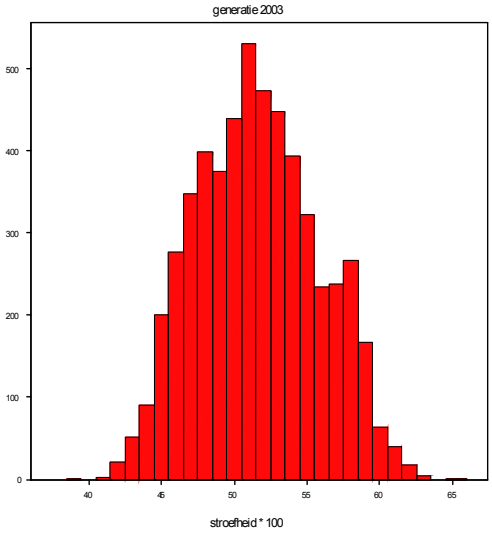
    1) De kans op een meetwaarde.

    De lappendeken heeft tot gevolg dat de set een verzameling is van een groot aantal niet aaneengesloten deelverzamelingen. Zodoende heeft het weinig zin om deze met behulp van één (theoretische)statistische verdeling te beschrijven. Het is dan ook min of meer noodzakelijk om over te gaan op discrete verdelingen en

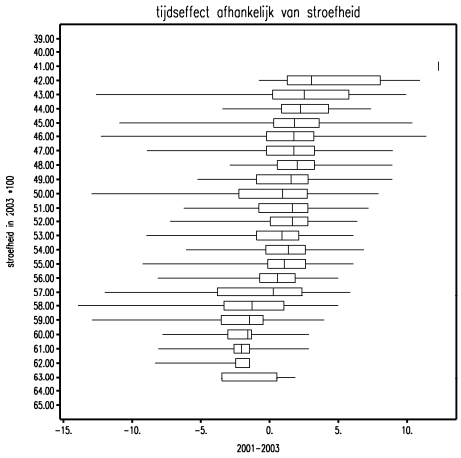
deze te benaderen met behulp van experimentele kansverdelingen. De gegevens van 2003 worden gebruikt om deze kansen te schatten, hierbij worden de klassen gevormd door de grenzen 0.395, 0.405 … 0.655 [1](zie fig 1)

2) De relatie met het verleden

De waarden uit 2001 worden eerst gecorrigeerd naar waarden uit 2003 mbv de totale auto passages in de periode 2001/2003. Vervolgens blijkt dat de verschillen tussen 2003 en de gecorrigeerde waarden uit 2001 afhankelijk zijn van de meetuitkomsten in 2003, zie boxplots in fig 2. Voor dit verloop is een lineair verband aangenomen. Vervolgens wordt aangenomen dat de overgebleven afwijkingen dan het gevolg zijn van meetfouten. Voor de meetfout wordt aangenomen dat deze normaal verdeeld is met een standaardafwijking van 0.017 (blijkt uit onderzoek voor DVS).



figuur 1 experimentele verdeling van 2003 data 2003 en 2001 afhankelijk van stroefheid in 2003



figuur 2:boxplots met verschillen tussen

3) Relatie met de buren

De verdeling van twee aanliggende vakken wordt met behulp van de 2 grenzen -1.5 en 1.5 gediscretiseerd naar 3 klassen. Omdat met beide buren rekening gehouden wordt[2] ontstaan er zodoende 6 klassen om het gedrag van de buren te beschrijven. Uit tabel 1 blijkt dat deze kansen afhankelijk zijn van het middelste vak

[1] De stroefheid wordt in 2 decimalen gegeven

[2] Er wordt geen onderscheid gemaakt tussen het vak date er voor ligt of dat er na ligt.

	Verschil van de twee buren ten opzicht van vak						
stroefheid 2003	<-0.015 <-0.015	<-0.015 -0.015;0.015	<-0.015 >0.015	-0.015;0.015 -0.015;0.015	-0.015;0.015 >0.015	>0.015 >0.015	toaal # vakken
0.39	0	0	0	0	0	100	1
0.40	0	0	0	0	0	100	2
0.41	0	0	0	100	0	0	5
0.42	0	0	4	60	32	4	27
0.43	0	2	0	71	14	12	60
0.44	0	7	2	56	31	4	92
0.45	1	4	1	75	14	3	210
0.46	0	2	1	84	10	2	363
0.47	1	8	2	76	10	3	404
0.48	0	9	2	79	9	1	415
0.49	1	9	1	78	10	1	487
0.50	1	7	2	77	12	1	538
0.51	1	7	2	83	7	1	576
0.52	2	12	2	78	6	1	500
0.53	1	11	2	77	8	1	474
0.54	1	9	1	81	6	1	427
0.55	1	7	1	83	7	1	489
0.56	1	8	0	86	4	0	513
0.57	1	8	1	86	4	0	517
0.58	0	7	0	86	6	0	460
0.59	1	11	1	83	4	0	344
0.60	0	16	0	81	3	1	199
0.61	0	15	2	73	10	0	124
0.62	1	15	1	81	2	0	89
0.63	8	21	0	71	0	0	38
0.64	29	14	0	57	0	0	7
0.65	0	0	0	100	0	0	2

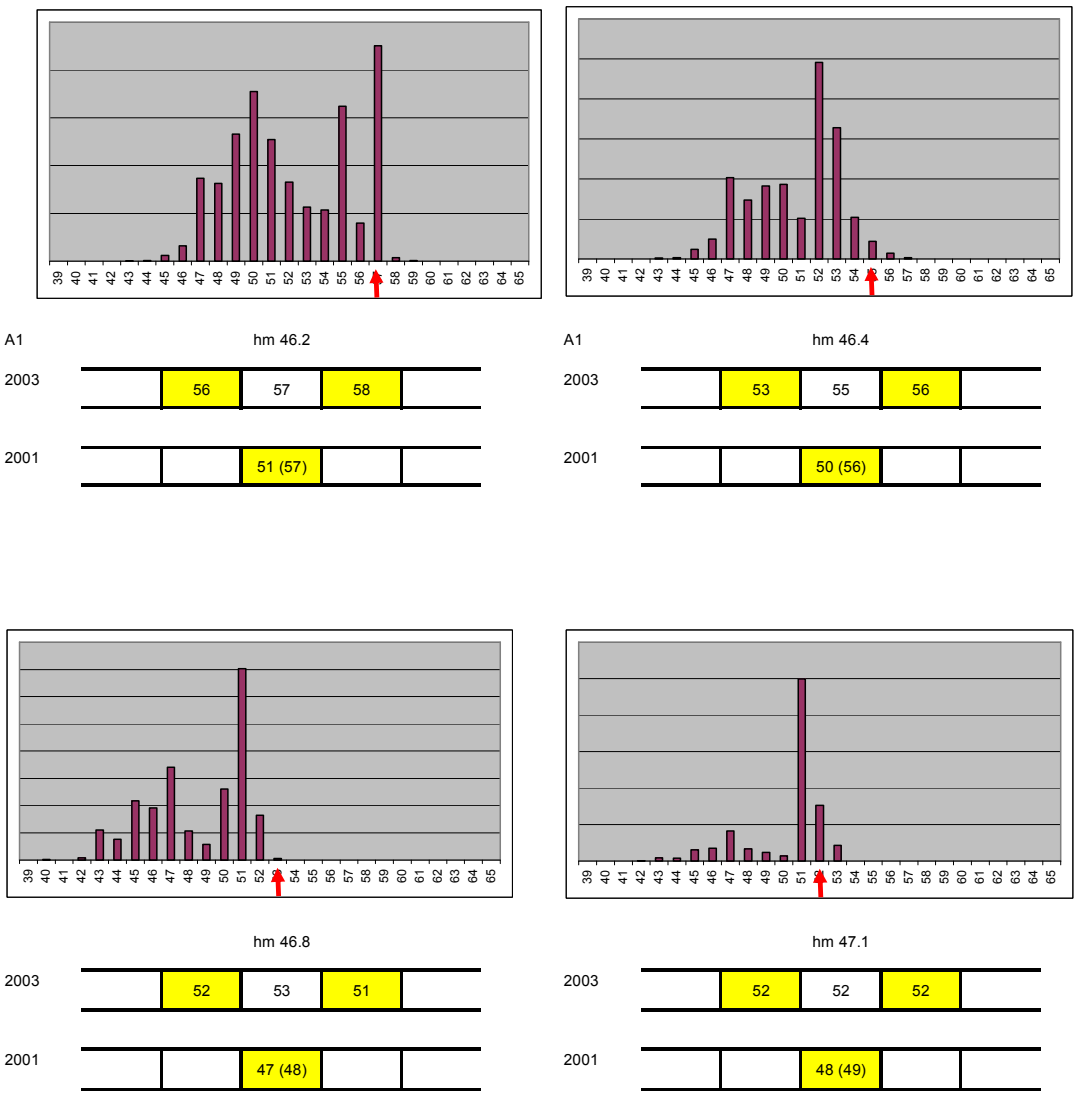
tabel 1: kansen op afwijkende buren en totaal aantal waarden afhankelijk van waarde in 2003

Toelichting:

In 2003 zijn er 474 vakken met een waarde van 0.53. Hiervan heeft 11% een buurpaar waarvan één waarde lager is dan 0.53-0.015 en de ander in het interval [0.53-0.015;0.53+0.015] ligt.

Voor de beschikbare data wordt de aanemelijkheid berekend van elke meting in 2003 en vergeleken meest de meest aannemelijkheid.

Voor een viertal vakken op de A1 is de procedure gevisualeerd in fig 3. Per vak zijn de aannemelijkheden uitgerekend op mogelijke meetwaarden in 2003 gegeven de gecorrigeerde metingen in 2001 (tussen haakjes staat de gemeten waarde in 2001) van dat vak en de metingen van zijn buren in 2003 (de gebruikte data staan in de geel gemarkeerde cellen). De rode pijl in het figuur komt overeen met de meting in 2003. Als de meting ‘correct’ is moet de ligging van de pijl overeen komen met de piek van de verdeling.



figuur 3: De aannemelijkheid van de meetuitkomsten gegeven de gecorrigeerde waarde uit 2001 (tussen haakjes staan de gemeten waardn) en de waarden van de buren in 2001.

Opmerking

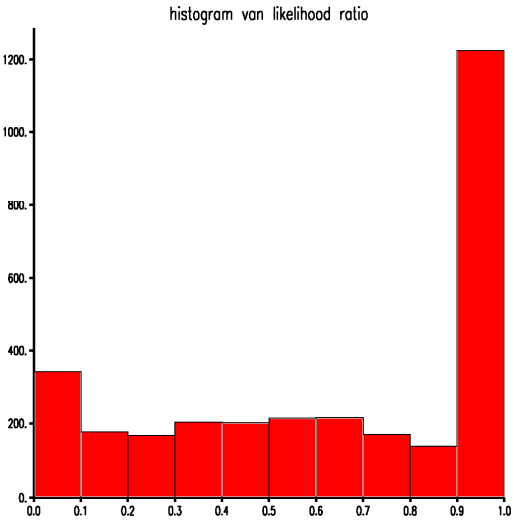
- Voor de vakken 46.2 en 46.4 is er een veschil van 6 tussen de metingen in 2001 en de gecorrigeerde metingen. Bij de vakken 46.8 en 47.1 is dit verschil gelijk aan 1. De oorzaak hiervan is dat de eerste 2 vakken in 2001 aangelegd zijn en de laatste 2 in 1995.
- Op vallend is, voor met name de vakken uit 1995, het onlogische verschil tussen 2003 en 2001.

Bovenstaande opmerkingen moet leiden naar een nader onderzoek van mogelijke oorzaken.

Een maat voor de correctheid van de 2003 meting is de verhouding van de maximale aannemelijkheid en de aannemelijkheid behorend bij de 2003 meting. Een histogram van

deze ratio's is fig 4. In tabel 2 staat de verdeling van de relatieve aannemelijkheid in getalsvorm. Het lijkt erop dat er drie groepen zijn

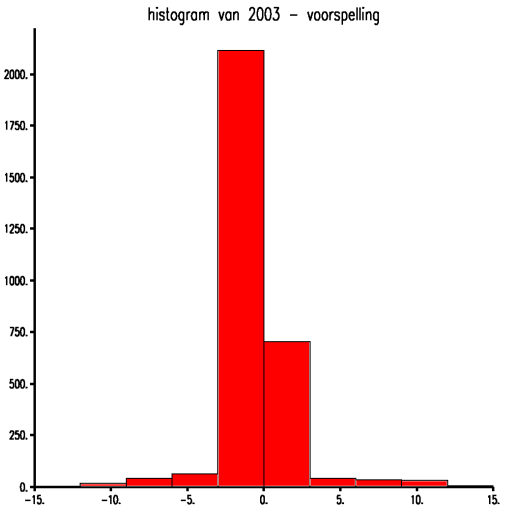
- data komen overeen met de verwachting; aannemelijkheidsratio van 0.9 tot 1
- data komen duidelijk niet overeen met de verwachting; aannemelijkheidsratio van 0.0 tot 0.1
- een minder duidelijk gebied met aannemelijkheidsratios van 0.1 tot 0.9.
-



figuur 4 relatieve aannemelijkheden van 2003 data

aanneme-lijkheid	%	cumulatief
0.0	10	10
0.1	6	16
0.2	5	21
0.3	7	28
0.4	7	35
0.5	7	42
0.6	7	49
0.7	6	55
0.8	5	60
0.9	5	65
1.0	35	100

tabel 2 relatieve aannemelijkheden van 2003 data



figuu 5: histogram van verschillen tussen voorspelling en 2003

[1] E. Elnahrawy: Statistical approaches to cleaning sensor data. In Distributed Sensor Networks; edited by S. Sitharama Iyengar and Richards R. Brooks