Liever rijden dan stilstaan:

Data biedt inzicht in halteertijden bij tram

Natalie in 't Veld – GVB – natalie.intveld@gvb.nl

Bijdrage aan het Colloquium Vervoersplanologisch Speurwerk
22 en 23 november 2018, Amersfoort

Samenvatting

In Amsterdam wordt het steeds drukker. De Vervoerregio Amsterdam (opdrachtgever van het OV) en GVB werken samen om het OV te versnellen, zodat er meer reizigers kunnen worden vervoerd. Onderdeel dat bij GVB ligt, is het verkorten van de halteertijd (halteerduur). Minder halteertijd betekent een kortere rijtijd (en reistijd), zo kun je met hetzelfde aantal voertuigen vaker heen en weer rijden.

Er zijn veel factoren die invloed hebben op de halteertijd: denk aan in- en uitstappers, ligging van de halte, verkeerslichten, voetgangersgebied of vrije trambaan, overig verkeer en het gedrag van de bestuurder zelf. Een aantal van dit soort invloeden is beschikbaar in verschillende databronnen. Deze paper biedt inzicht in hoeverre de halteertijd verklaard kan worden met de beschikbare data.

De databronnen die tot onze beschikking staan zijn de databases met de geplande dienstregeling, de gerealiseerde exploitatie en de chipkaartdatabase. Vanwege de gigantische omvang is voor deze analyse gewerkt met een dataset over de maanden maart en april 2018. In een aantal stappen is een lineair regressiemodel ontwikkeld. Het model verklaart zo'n 40% van de variantie. Dit geeft aan dat er naast de beschikbare gegevens zoals verwacht meer dingen van invloed zijn. In het uiteindelijke model hebben het voertuigtype, het aantal in- en uitstappers, tijdstip op de dag en de zone waar de halte ligt significante en 'belangrijke' invloed. De haltepalen zijn in een clusteranalyse verdeeld in drie groepen die qua kenmerken van elkaar verschillen. Deze clusterindeling heeft (vanzelfsprekend) ook een grote invloed op de halteertijd.

Het meest opvallende resultaat is de invloed van het voertuigtype: de halteertijd van een 2-richtings Combino ligt 8% hoger dan van een gewone Combino, en de halteertijd van oude (11G en 12G) materieel levert zo'n 15% meer halteertijd. Voor de uitspraak of dit een-op-een geldt voor het wel of niet rijden met een conducteur is verder onderzoek nodig.

Keywords: openbaar vervoer, tram, data-analyse

1. Inleiding

1.1 Aanleiding

In Amsterdam wordt het steeds drukker. De Vervoerregio Amsterdam (opdrachtgever van het OV) en GVB werken samen om het OV te versnellen, zodat er meer reizigers kunnen worden vervoerd. Onderdeel dat bij GVB ligt, is het verkorten van de halteertijd (halteerduur). Minder halteertijd betekent een kortere rijtijd (en reistijd), zo kun je met hetzelfde aantal voertuigen vaker heen en weer rijden. Dit project biedt inzicht in de factoren die de halteertijd bepalen.

Dit onderwerp is binnen GVB niet nieuw. Verschillende afdelingen hebben samen gekeken naar mogelijk factoren die invloed hebben op de halteertijd. Voor deze paper beperken we ons tot de factoren die we op basis van beschikbare data kunnen analyseren.

Centrale vraag: welke factoren bepalen de halteertijd bij tram?

1.2 Invloed op halteren

GVB wil reizigers snel en comfortabel naar hun bestemming vervoeren. Daarbij moeten we steeds de afweging maken tussen fijnmazigheid (stoppen op elke hoek) en snelheid (zoveel mogelijk doorrijden). Als je in de tram zit, wil je zo snel mogelijk naar je bestemming. Als je er nog niet in zit, wil je graag dat de tram heel dichtbij een halte heeft. Dat halteren kost tijd, en die tijd hangt af verschillende factoren. In de planning wordt nu 18 seconden aangehouden voor het halteren. In de praktijk zien we afwijkingen zowel naar boven als naar beneden. Het aantal instappers en uitstappers zal van invloed zijn, en of er veel mensen instappen die nog een kaartje moeten kopen. Verder moet GVB zich aan z'n eigen dienstregeling houden: op bepaalde haltes mag je niet te vroeg vertrekken, de juiste tijd moet worden afgewacht en op een klein aantal haltes sturen we op niet te vroeg aankomen (aankomstpunctualiteit). Het laatste om 'af te dwingen' dat er geen reizigers wachten op een tram die al geweest is. GVB bepaalt zelf de hoeveelheid tijd die er nodig is om van a naar b te rijden.

Iets waar we als GVB geen invloed op hebben, is het overige verkeer en de infrastructuur. De haltes zijn van onze opdrachtgever. Idealiter doen we de inrichting van kruispunten en de afstelling van de verkeerslichten gezamenlijk, maar dat is niet altijd het geval en ook niet reëel. Haltes liggen voor of na de kruising en trams moeten soms wachten voor verkeerslichten. Taxi's mogen over de trambaan, en sommige trajecten lopen dwars door een voetgangersgebied.

1.3 Leeswijzer

Hoofdstuk 2 beschrijft welke databronnen er gebruikt zijn en welke bewerkingsstappen er zijn uitgevoerd. In hoofdstuk 3 worden eerst de variabelen besproken, dan enkele bivariate relaties en uiteindelijk hoe het model tot stand is gekomen. De modelresultaten en interpretatie staan in hoofdstuk 4. Hoofdstuk 5 sluit af met de belangrijkste conclusies en aanbevelingen voor verder onderzoek.

2. Data en dataverwerking

2.1 Bronnen

Realisatie van de halteertijden
De belangrijkste bron is de realisatie van rij- en halteertijden met allerlei kenmerken. Deze bron is samengesteld op basis van logfiles van de voertuigen: voertuigen geven elke zoveel seconden aan waar ze zich bevinden en de 'acties' die zich voordoen: deuren open, deuren dicht, stilstaan, optrekken < 3km/u, sneller dan 3km/u etc. Hier is ook alle informatie over het voertuig en de uitgevoerde rit bekend: wagennummer, lijnnummer, richting, halte, datum en tijdstip. Deze gegevens zijn over meerdere jaren beschikbaar in een dashboard. Dit zijn dus gestructureerde, bewerkte gegevens en in principe 'schoon'. Om elke halte ligt een gps-box. De tram rijdt dus de box in, halteert, en rijdt de box weer uit. De halteertijd is het aantal seconden tussen eerste keer deuren open in de box en het eerste van deze twee: voldoende snelheid of de box uit. Wachten voor een kruispunt ín de box betekent dus een hogere halteertijd.

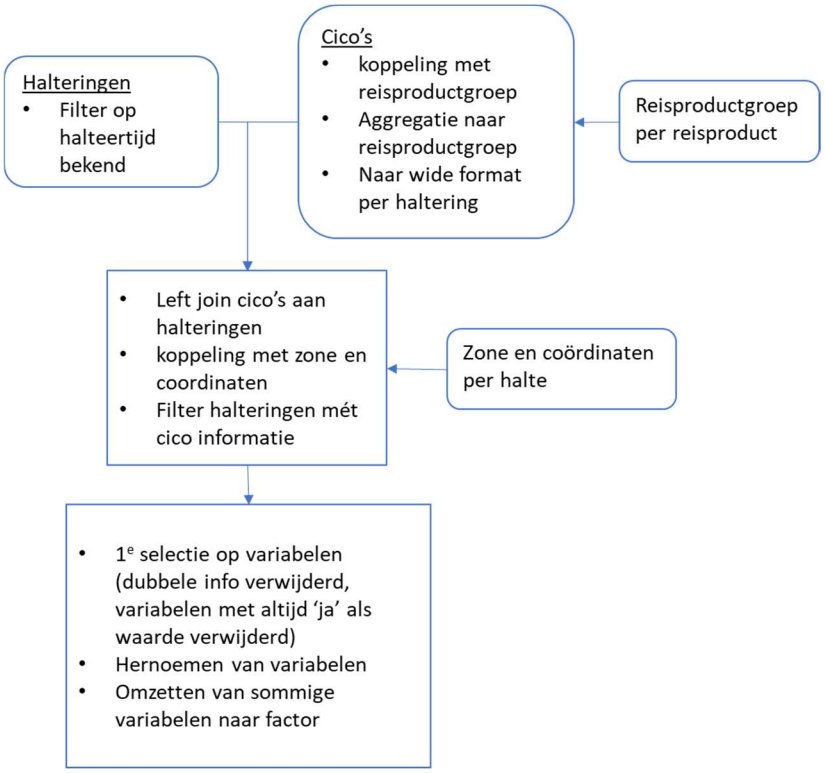
Chipkaartdata voor in- en uitstappers
Voor in- en uitstappers maken we gebruik van onze chipkaartinformatie. In de chipkaartdatabase staan alle checkin- en checkuittransacties, op welke halte die plaatsvonden en met welk reisproduct. Alle transacties van maart en april 2018 bij tram zijn geëxporteerd. De bron is privacygevoelig, de gegevens per transactie zijn dan ook beperkt tot datum-lijn-richting-grootwagennummer-geplande vertrektijd-halte-reisproduct-transactietype. De koppeling tussen chipkaart en de voertuigrit is zeker niet triviaal, en is in deze bron dus al gemaakt. Verder te noemen: cico's (van check-in-check-out).

De geplande dienstregeling
De derde bron is de geplande dienstregeling met daarin de geplande vertrek, rij- en halteertijden van elke voertuigrit. Deze is nodig om te zien of een haltering plaatsvond op een rit die te laat was. De koppeling met de gerealiseerde dienstregeling was al gemaakt.

2.2 Databewerking

De bronnen waren allemaal bij GVB beschikbaar. Er zijn 15 tramlijnen die afhankelijk van weekdag en tijdstip 4-15x per uur een route van 11-32 haltes rijden, dus de omvang wordt snel gigantisch. We hebben daarom een dataset gebruikt van slechts 2 maanden: maart en april 2018. Koningsdag is uit de dataset verwijderd. Alle bewerkingen en analyses zijn uitgevoerd in R.

De bewerkingsstappen zijn in het volgende schema weergegeven:



Figuur 1: stappen om de data samen te voegen

Bij de export van de gerealiseerde halteringen en de geplande dienstregeling zijn alle missings en nullen al verwijderd. In het geval van missings is er niet goed gemeten, nullen krijg je als er niet gehalteerd wordt. Gemiddeld wordt er in 80% van alle halte-passages gehalteerd. De geplande 18 seconden geldt voor alle passages, gecorrigeerd voor daadwerkelijk halteren ligt de geplande halteertijd dus op 22,5 seconden.

Bij de export van de cico's zijn de rijen zonder ritidentificatie niet meegenomen, die konden toch niet worden gekoppeld. Het koppelen van de cico's aan de halteringen lukt in een kleine 5% van de gevallen niet, die regels zijn ook weggelaten.

Sommige halteertijden zijn heel lang. Dat zijn geen meetfouten, dat gebeurt gewoon door de situatie op dat moment, vaak grote drukte en verstopte kruispunten. Deze waarnemingen zijn niet verwijderd.

Het resultaat van de bewerkingsstappen is een analysebestand met 3,3 miljoen observaties en 57 variabelen. Daarvan zijn er initieel 23 meegenomen in de regressies, de andere waren (soms) wel handig als extra informatie.

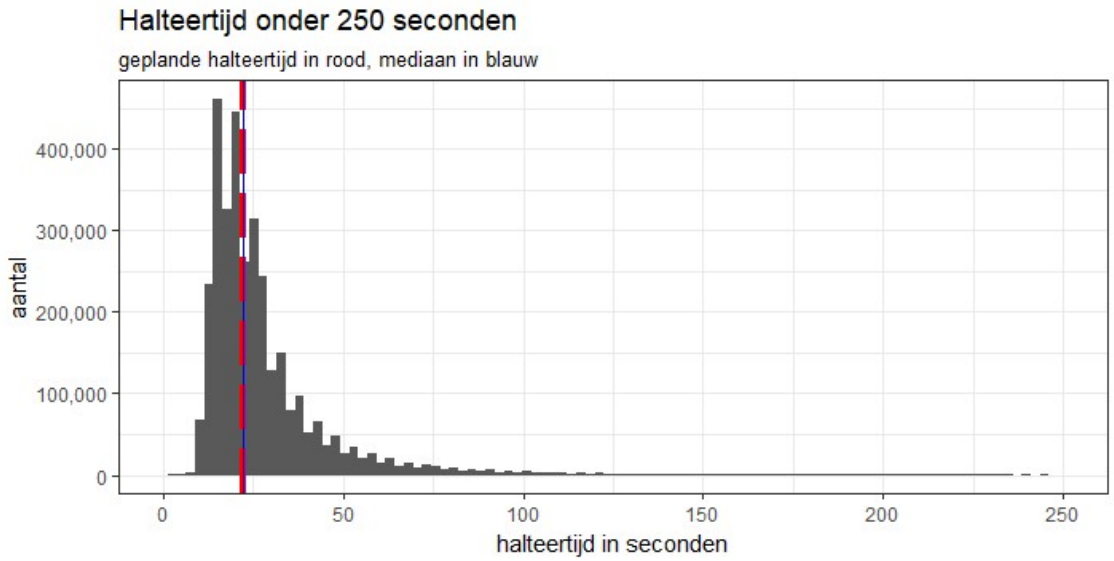
3. Modelontwikkeling

Het doel is het verklaren van de halteertijd uit de andere variabelen. Belangrijk is om te weten te komen hoeveel invloed die variabelen hebben. We zijn gestart met een verkennende analyse op de beschikbare variabelen en de samenhang tussen variabelen. Vervolgens hebben we middels lineaire regressie inzicht gekregen in de factoren die de halteertijd beïnvloeden. Lineaire regressie modelleert het lineaire verband tussen de te

verklaren en verklarende variabelen. De coëfficiënten van de verklarende variabelen geven de invloed weer van die variabelen, rekening houdend met de andere verklarende variabelen in het model.

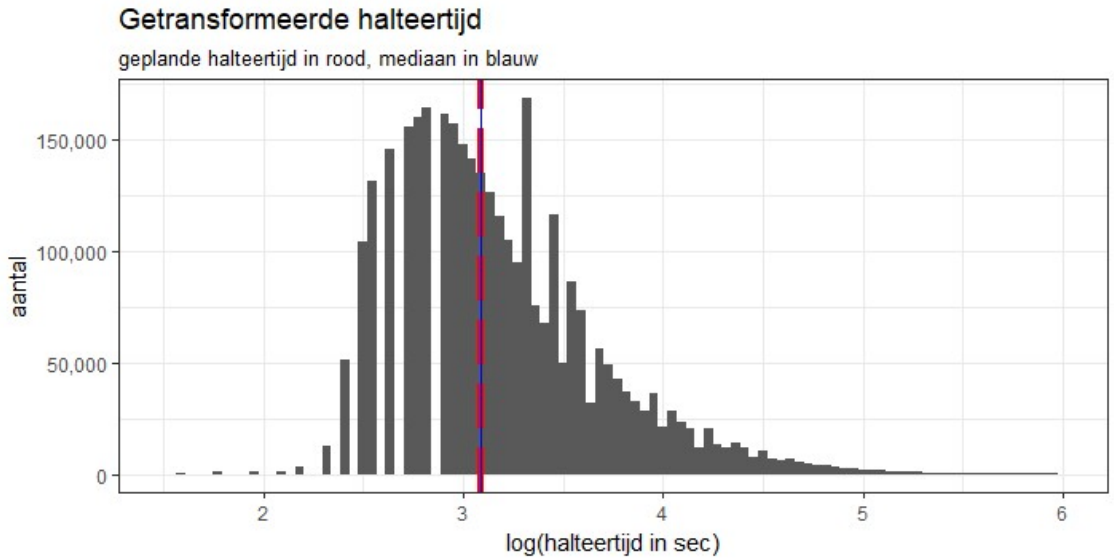
3.1 Variabelen

Te verklaren: de halteertijd
De halteertijd wordt gemeten als het aantal seconden tussen de eerste keer deuren open in de gps-box rond de halte en het minimum van de box weer uit en harder dan 3km/u binnen de box. De verdeling (zonder de extremen) is als volgt:



Figuur 2: verdeling van de halteertijd

Deze verdeling heeft een vrij dikke staart. Om de lineaire fit te verbeteren is een logtransformatie toegepast. Dit leidt tot een iets minder scheve verdeling:



Figuur 3: verdeling van de getransformeerde halteertijd

Verder heeft deze te verklaren variabele in de bestudeerde periode de volgende kenmerken:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00   17.00   22.00   28.14   32.00 1768.00
```

Verklarende variabelen

Als verklarende variabelen hebben we gegevens over de uitgevoerde rit, die invloed kunnen hebben op de halteertijd. Bepaalde vermoedens heeft iedereen: meer in- en uitstappers hebben vast een hogere halteertijd tot gevolg, op zondagmorgen kunnen we sneller over kruispunten heen omdat er verder weinig verkeer is, en als een rit eerder vertraging heeft opgelopen zal de bestuurder zijn best doen om de tijd weer goed te maken.

De dataset bevat de volgende variabelen:

Variabele	Betekenis	Waarden	Format
Lijnnummer	Nummer van de lijn	1,2,3,4,5,7,9,10,12, 13,14,16,17,24,26	Categorie
Richting	Van het eindpunthuisje of naar het eindpunthuisje	Heen of terug	Categorie
Weekdag		Ma-zo	Categorie
Aank_verschil_halte	Aantal seconden tussen gerealiseerde en geplande aankomsttijd op de betreffende halte	-1731,…,1683	integer
Voertuigtype	Uitvoering van de tram (wel/geen vaste instapdeuren, deuren aan 2 kanten)	Combino, Combino-twee-richtingen, 11G-twee-richtingen, 12G	Categorie
Tijdhalte_ind	Moet de juiste vertrektijd worden afgewacht op de halte	Ja/Nee	Categorie
Synchr_ind	Is de halte het eind van een synchronisatietraject	Ja/Nee	Categorie
Rit_vert_verschil	Aantal seconden tussen geplande en gerealiseerde vertrektijd bij de eerste halte	-1696,…,1546	Interval
Ci_*/co_***	Het aantal checkins en checkouts op abonnementen, saldoreizen, dagkaarten, uurkaarten, andere incidentele kaartjes en overige kaarten.	0,1,2,…	Interval
Zone	Geografisch gebied	5700, 5710,5713, 5714, 5715, 5723, 5724 en een aantal grenzen	Categorie
uur	Het hele uur van het halteertijdstip	0,4,5,…,23	Categorie

Tabel 1: verklarende variabelen in de dataset

Er is één lijn die geen keerlus heeft op het eindpunt, dat is lijn 5. Speciaal voor deze lijn is er twee-richting materiaal: deuren aan beide kanten en twee voorkanten. Daarnaast

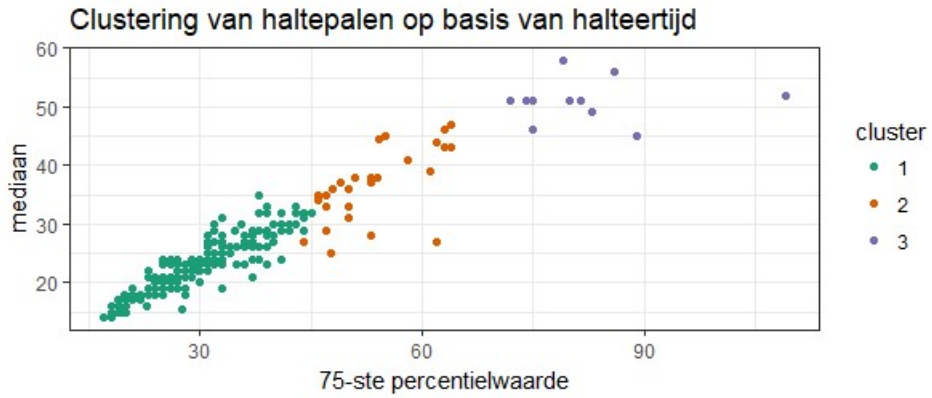
zijn er 3 tramtypes in omloop: 11G, 12G en Combino's. 11G en 12G zijn de oudste types, deze hebben per zijde vier instapmogelijkheden waarvan er twee een trap hebben. Alleen de 'gewone' Combino's hebben een conducteur. Over het algemeen worden de twee-richting-voertuigen alleen op lijn 5 ingezet, en worden lijn 16 en 24 bediend met 12G-voertuigen.

Figuur 4: Combino en 11G

Tijdens het ontwikkelen van het model is een aantal afgeleide variabelen bepaald. Bijvoorbeeld een indicator voor het hebben van een conducteur (geen op lijn 5,16 en 24), het totaal aantal in- en uitstappers en combinaties van soorten in- en uitstappers. De afgeleide variabelen die daadwerkelijk zijn gebruikt, worden verderop toegelicht.

3.2 Clustering van haltes

Uiteraard is de halteertijd op verschillende haltes ook verschillend: het gemiddelde, de mediaan en vooral de spreiding verschilt. Om hier aan de voorkant rekening mee te houden, zijn middels een clusteranalyse drie groepen haltes onderscheiden. De clustering heeft plaatsgevonden op een aggregatie van de data, die bestond uit een aantal indicatoren per haltecode (haltepaal): het gemiddelde, de mediaan, de standaarddeviatie en de 75ste percentielwaarde van de halteertijd en het aantal passerende lijnen. Het meenemen van de clustering levert een winst van bijna 5 procentpunt in verklaarde variantie.

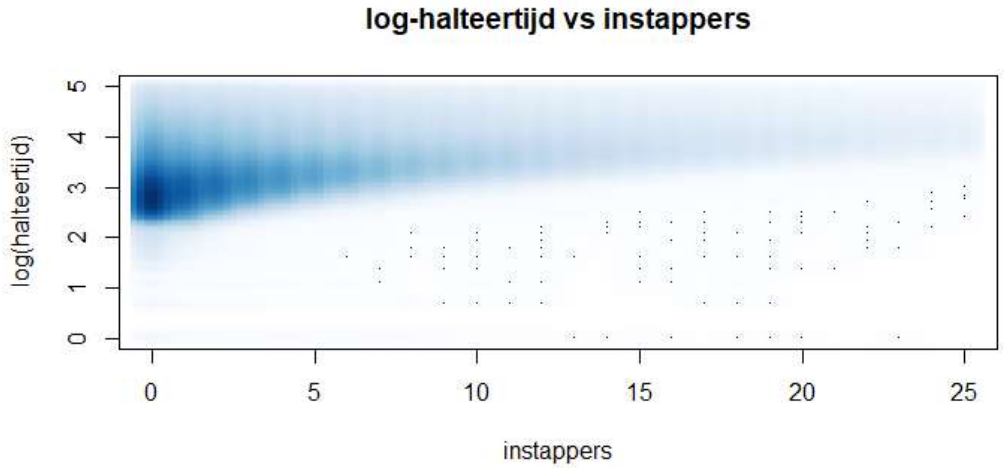


Figuur 5: clustering van haltepalen

3.3 Bivariate relaties

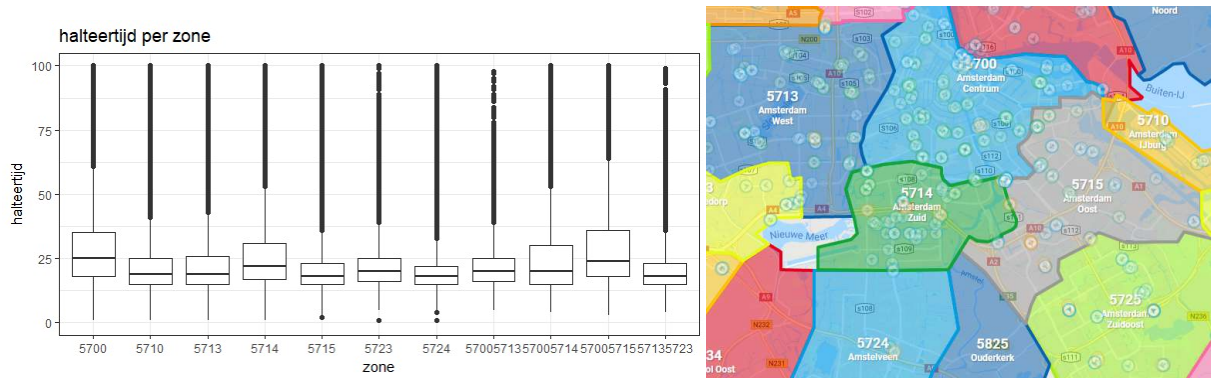
Om een eerste indruk te krijgen van de bivariate relaties tussen de halteertijd en de verklarende variabelen, staat hieronder een aantal voorbeelden.

Er lijkt wel een relatie te zijn tussen de log van de halteertijd en het totaal aantal instappers:



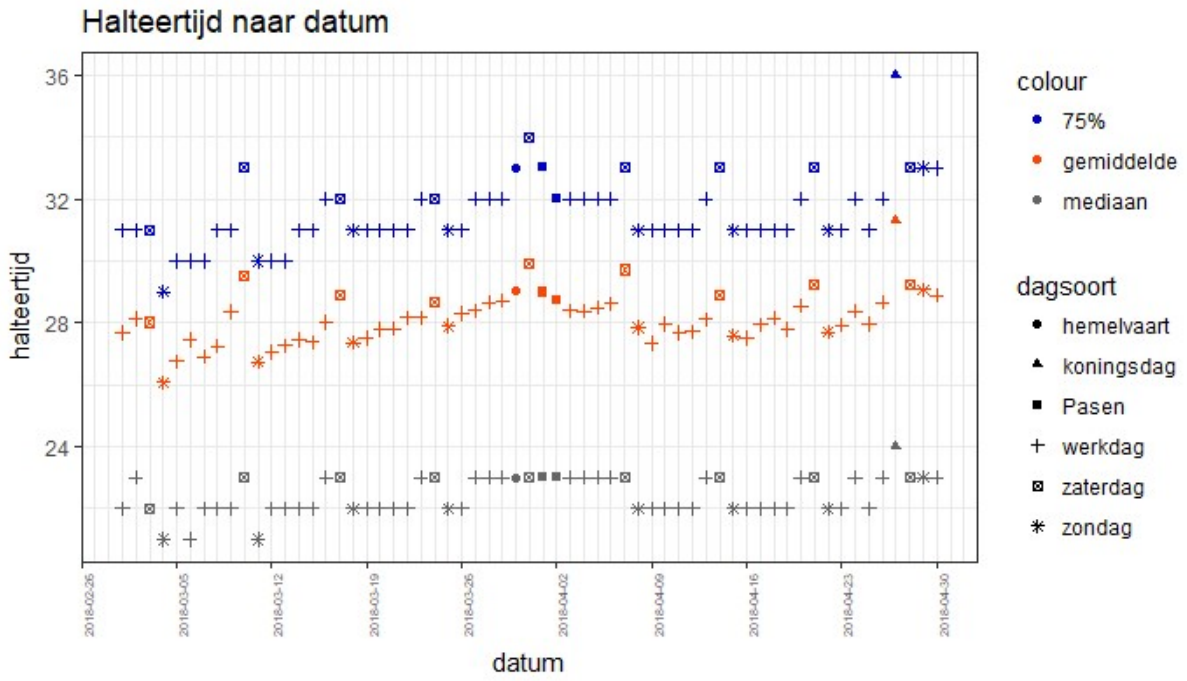
Figuur 6: relatie tussen instappers en de halteertijd

Tussen de zones verschilt vooral de variatie:



Figuur 7: relatie tussen zone en halteertijd

De dataset bestaat uit 2 maanden aan data. Daar zit ook een aantal bijzondere dagen tussen, zoals de Paasdagen en Koningsdag. In onderstaande figuur is te zien dat vooral Koningsdag een uitschieter is. Op die dag worden van een aantal lijnen ook de routes aangepast, dat gebeurt niet op feestdagen. In de dataset zijn de waarnemingen van Koningsdag verwijderd.



Figuur 8: halteertijd per dag: mediaan, gemiddelde en 75ste percentielwaarde

3.4 Multivariate relaties: lineaire regressie

Modelontwikkeling

Om te komen tot 'het beste model' hebben we verschillende modellen met elkaar vergeleken. Daartoe is de dataset gesplitst in een trainingsset (70%) en een testset (30%). Het model is telkens geschat op basis van de trainingsset. Van elk model is de verklaarde variantie (R^2) bewaard. Vervolgens is het model gebruikt om op basis van de testset voorspellingen te doen voor de halteertijd. Die voorspellingen zijn weer vergeleken met de werkelijke waarden uit de testset. De bewaarde indicatoren zijn de median absolute deviation[1] (MAD) en de correlatie tussen de voorspellingen en de werkelijke waarden (COR).

Slag 1: 1 of meer modellen?

Initieel is een model geschat met alle variabelen en alle waarnemingen erin. We hebben bekeken of het iets op zou leveren om verschillende modellen te maken voor verschillende groepen waarnemingen. Voor de hand ligt onderscheid te maken naar verschillende clusters, en voor haltes waar de tijd wel of niet hoeft te worden afgewacht. Tegen de verwachting in was het modelresultaat voor de gesplitste modellen (zonder verdere aanpassing) telkens slechter dan dat van het geïntegreerde model.

Slag 2: selectie van variabelen

Vanuit het basismodel zonder uitsluiting van bepaalde groepen waarnemingen zijn we gaan kijken naar de selectie van de variabelen zelf. Ook zijn verschillende variabelen

[1] De median absolute deviation is een maat voor de grootte van de residuen: MAD = mediaan(|voorspelling – werkelijke waarde|). Deze is minder gevoelig voor de outliers dan de gewoonlijk gebruikte residual standard error.

samengevoegd en ingedikt. In het basismodel zijn de coëfficiënten van alle variabelen significant. Sommige coëfficiënten zijn echter heel klein, zodat ze niet belangrijk zijn. Voor het selecteren van de variabelen is gekeken naar gestandaardiseerde coëfficiënten die onder de -0,05 of boven de 0,05 liggen. Op deze manier wordt rekening gehouden met het verschil in variantie tussen de (categorische en interval) variabelen. Dit betekent dat een eenheid (1 standaarddeviatie) verschil in de bijbehorende variabele leidt tot meer dan 1 seconde verschil in halteertijd. Omdat de coëfficiënten wel significant zijn, is -voor ze helemaal weg te laten- gekeken of samenvoegen toegevoegde waarde had.

Speciaal geval: Lijnnummer, voertuigtype en instapregime
In de geselecteerde tijdsperiode rijden lijn 5,16 en 24 zonder conducteur. Dit betekent dat reizigers bij alle deuren mogen in- en uitstappen. Hierop zijn de voertuigen aangepast, die hebben een ander voertuigtype. Daarnaast zijn de meeste voertuigen zonder conducteur ouder en hebben voor en achter trappen. Alleen in het midden kun je gelijkvloers instappen. De overige lijnen hebben wel een conducteur, de reizigers moeten op die lijnen bij de bestuurder of bij de conducteur instappen en mogen alleen bij de overige deuren uitstappen
In het initiële model zijn dummy's opgenomen voor alle drie de variabelen. Omdat ze heel sterk samenhangen, is het niet handig om ze alle drie in het model te houden. Tijdens de modelontwikkeling is gekeken welke variabele de meeste informatie toevoegt. Uiteindelijk hebben we gekozen voor het voertuigtype. Dit maakt zowel onderscheid naar het instapregime als naar de 'twee-richting' voertuigen.

3.5 Toetsen van statistische aannames

De aannames bij lineaire regressie zijn een lineaire relatie tussen de inputvariabelen en outputvariabele, onderling onafhankelijke inputvariabelen, homoskedasticiteit en normaal verdeelde residuen. De eerste aanname volgt uit het model. Er is gecheckt op multicollineariteit (alle GVIF < 2) en homoskedasticiteit is in orde. De residuen zijn niet helemaal netjes normaal verdeeld, wat tot gevolg heeft dat de betrouwbaarheidsintervallen rond de coëfficiënten mogelijk niet kloppen. Gezien de grote aantallen is dit niet aan de orde.

4. Resultaten

In hoofdstuk 3 is beschreven hoe het beste model tot stand is gekomen. In dit hoofdstuk worden de resultaten van dit model besproken.

4.1 Geselecteerde variabelen

Uiteindelijk hebben we voor het model gekozen dat zo min mogelijk variabelen bevat, maar wel zo goed mogelijk presteert. Dat is het geval met onderstaande verklarende variabelen. Deze zijn in het uiteindelijke model allemaal significant én belangrijk:

- Voertuigtype
- Aantal instappers op saldo
- Aantal instappers met een dagkaart
- Aantal instappers met een uurskaart

- Overige instappers (afgeleid uit instappers met abonnement, overige incidentele kaartjes en overige kaartjes)
- Totaal aantal uitstappers (afgeleid uit uitstappers met een abonnement, overige incidentele kaartjes, overige kaartjes, saldo, dagkaarten en uurkaarten)
- Zone van de halte
- Uur van de dag
- clusteruitkomst

Dat betekent dat deze variabelen geen belangrijke invloed hadden:

- Weekdag, ook niet samengevoegd naar werkdagen/zaterdag/zondag/afzonderlijke feestdagen
- Lijnnummer (ten gunste van het voertuigtype)
- Richting
- Aantal seconden tussen geplande en gerealiseerde vertrektijd van de beginhalte
- Aantal seconden tussen geplande en gerealiseerde aankomsttijd bij de halte
- Of de halte het eind van een synchronisatietraject was
- Of op de halte het goede vertrektijdstip moet worden afgewacht

4.2 Modelprestaties

Het beste model heeft een R^2 van 0,40. Op de testset is de MAD 0,31, wat overeenkomt met iets meer dan 1 seconde halteertijd. De correlatie tussen de voorspelde halteertijd en de werkelijke halteertijd in de testset is 0,63.

4.3 Coëfficiënten

Voor de verklarende variabelen die significant en belangrijk zijn, staan de coëfficiënten in onderstaande tabel. Zie voor de statistische output de bijlage.

Omdat de halteertijd een logtransformatie heeft ondergaan, zijn de coëfficiënten (behalve van de constante) te interpreteren als een percentage.

De categorische variabelen hebben een dummy gekregen voor elke waarde. Per variabele heeft er maar één dummy tegelijk waarde 1. De coëfficiënt geeft het verschil aan ten opzichte van de 'basiswaarde' van die variabele. Een voorbeeld: er zijn vier voertuigtypes: Combino, Combino-twee-richtingen, 11G-twee-richtingen en 12G. De basiswaarde is de Combino, die heeft geen coëfficiënt. De coëfficiënt van voertuigtype Combino-twee-richtingen is 0,08, wat betekent dat een haltering van een Combino-twee-richtingen 8% hogere halteertijd heeft dan eentje van een Combino (mits de andere variabelen hetzelfde blijven). De halteertijd van 11G-twee-richtingen ligt 14% hoger en die van 12G 16%.

Zone 5700 heeft duidelijk de langste halteertijd: de coëfficiënten van alle andere zones zijn negatief. Zone 5724 heeft de kortste halteertijd, dit is zone Amstelveen waar een vrije trambaan ligt.

Variabele	dummy's	coëfficiënt
Constante		2,88
Voertuigtype Basis = Combino	Combino tweerichting	0,08
	Tram G11 Tweerichting	0,14
	Tram G12	0,16
Instappers saldo		0,03
Instappers dagkaarten		0,03
Instappers uurskaarten		0,06
Overige instappers		0,03
Uitstappers		0,02
Zone Basis = 5700	5710	-0,13
	5713	-0,12
	5714	-0,08
	5715	-0,11
	5723	-0,07
	5724	-0,25
	Zonegrens 5700-5713	-0,11
	Zonegrens 5700-5714	-0,05
	Zonegrens 5700-5715	-0,06
	Zonegrens 5713-5723	-0,13
Uur Basis = 00:00 – 01:00	05:00 – 06:00	-0,10
	06:00 – 07:00	-0,08
	07:00 – 08:00	0,00
	08:00 – 09:00	0,02
	09:00 – 10:00	0,06
	10:00 – 11:00	0,09
	11:00 – 12:00	0,10
	12:00 – 13:00	0,11
	13:00 – 14:00	0,12
	14:00 – 15:00	0,12
	15:00 – 16:00	0,12
	16:00 – 17:00	0,11
	17:00 – 18:00	0,09
	18:00 – 19:00	0,07
	19:00 – 20:00	0,06
	20:00 – 21:00	0,04
	21:00 – 22:00	0,03
	22:00 – 23:00	0,01
	23:00 – 00:00	0,00
Cluster Basis = Cluster 1	Cluster 2	0,36
	Cluster 3	0,49

Tabel 2: coëfficiënten van het regressiemodel

5. Conclusies en aanbevelingen

5.1 Conclusies

- Er zijn groepen haltes die nogal een afwijkende halteertijd hebben. Onderscheid hiernaar maken levert een hogere verklaarde variantie
- Het voertuigtype heeft een grote invloed op de halteertijd. De oude types met de trappen hebben een 14-16% hogere halteertijd dan de één-richting Combino. Vergelijking van de Combino met en zonder conducteur levert 8% verschil in halteertijd (hoger zonder conducteur).
- De halteertijd laat zich uit de beschikbare data redelijk goed voorspellen, de correlatie tussen voorspeld en gerealiseerd ligt op 0,63 en de verklaarde variantie op 40%.
- Er blijft dus ook een aanzienlijk deel onverklaarde variantie over. Er zijn ook andere dingen die invloed hebben, zoals de ligging van de halte ten opzichte van de kruising, veel of weinig ander verkeer en verkeerslichten.

5.2 Aanbevelingen voor verder onderzoek

- Een langere periode meenemen: dan zit er ook invloed in van seizoenen, het weer, vakantieperiodes en feestdagen. Daarnaast zijn er door de tijd allerlei aanpassingen gedaan die de halteertijd beïnvloeden. Bij analyse van een langere periode kunnen die verschillende regimes worden vergeleken. Denk aan het schrappen in het assortiment kaartjes dat op de voertuigen beschikbaar is, invoering van pinnen op de tram en uitfaseren van cashbetalingen, beschikbaarheid van verkoopautomaten op de halte etc.
- Per 22 juli 2018 is het lijnennet behoorlijk gewijzigd. Het is interessant om te kijken of bepaalde conclusies (bijvoorbeeld die over de voertuigtypes) blijven gelden. Lijn 5 (twee-richting materieel) heeft een ander eindpunt en lijn 16 is vervallen. Er wordt nu ouder materieel ingezet op lijn 19, die weer niet naar Centraal Station rijdt.
- Om het effect van de conducteur verder uit te diepen, kan er gekeken worden naar haltes waar zowel lijnen met als zonder conducteur komen.

Bijlage: coëfficiënten van het lineaire model met hun standaarddeviaties

```
Coefficients:
                                Estimate Std. Error  t value Pr(>|t|)
(Intercept)                    2.876e+00  2.591e-03 1110.213  < 2e-16 ***
type_omsCombino tweerichting   7.585e-02  3.031e-03   25.024  < 2e-16 ***
type_omsTram G11 Tweerichting  1.356e-01  1.138e-03  119.175  < 2e-16 ***
type_omsTram G12               1.551e-01  1.000e-03  155.089  < 2e-16 ***
ci_saldo                       3.001e-02  1.076e-04  278.837  < 2e-16 ***
ci_dag                         3.121e-02  1.241e-04  251.409  < 2e-16 ***
ci_uur                         5.861e-02  2.811e-04  208.468  < 2e-16 ***
ci_rest                        2.520e-02  1.351e-04  186.515  < 2e-16 ***
co_tot                         1.859e-02  5.178e-05  359.109  < 2e-16 ***
zone5710                      -1.254e-01  1.568e-03  -79.954  < 2e-16 ***
zone5713                      -1.203e-01  6.707e-04 -179.387  < 2e-16 ***
zone5714                      -7.876e-02  1.074e-03  -73.320  < 2e-16 ***
zone5715                      -1.093e-01  1.320e-03  -82.814  < 2e-16 ***
zone5723                      -7.464e-02  2.895e-03  -25.786  < 2e-16 ***
zone5724                      -2.504e-01  2.490e-03 -100.565  < 2e-16 ***
zone57005713                  -1.058e-01  5.611e-03  -18.865  < 2e-16 ***
zone57005714                  -4.999e-02  3.173e-03  -15.755  < 2e-16 ***
zone57005715                  -5.620e-02  2.292e-03  -24.516  < 2e-16 ***
zone57135723                  -1.333e-01  4.035e-03  -33.040  < 2e-16 ***
uur5                          -9.864e-02  5.080e-03  -19.417  < 2e-16 ***
uur6                          -7.885e-02  3.043e-03  -25.915  < 2e-16 ***
uur7                          -1.087e-03  2.817e-03   -0.386    0.699
uur8                           1.557e-02  2.782e-03    5.596 2.19e-08 ***
uur9                           5.789e-02  2.781e-03   20.813  < 2e-16 ***
uur10                          8.656e-02  2.784e-03   31.096  < 2e-16 ***
uur11                          1.023e-01  2.772e-03   36.910  < 2e-16 ***
uur12                          1.140e-01  2.769e-03   41.174  < 2e-16 ***
uur13                          1.193e-01  2.767e-03   43.113  < 2e-16 ***
uur14                          1.171e-01  2.763e-03   42.371  < 2e-16 ***
uur15                          1.205e-01  2.758e-03   43.672  < 2e-16 ***
uur16                          1.133e-01  2.753e-03   41.147  < 2e-16 ***
uur17                          9.141e-02  2.760e-03   33.117  < 2e-16 ***
uur18                          6.935e-02  2.780e-03   24.942  < 2e-16 ***
uur19                          5.838e-02  2.800e-03   20.851  < 2e-16 ***
uur20                          3.787e-02  2.848e-03   13.296  < 2e-16 ***
uur21                          2.530e-02  2.873e-03    8.806  < 2e-16 ***
uur22                          1.166e-02  2.881e-03    4.048 5.16e-05 ***
uur23                         -1.335e-04  2.923e-03   -0.046    0.964
clus.ident2                    3.596e-01  9.859e-04  364.715  < 2e-16 ***
clus.ident3                    4.854e-01  1.493e-03  325.051  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3909 on 2298051 degrees of freedom
  (1076 observations deleted due to missingness)
Multiple R-squared:  0.3965,   Adjusted R-squared:  0.3964
F-statistic: 3.871e+04 on 39 and 2298051 DF,  p-value: < 2.2e-16
```