Wat is het gedrag van lichtvervuiling?

Bekker, Zoë

Wat is het gedrag van lichtvervuiling?

Student: Zoë Bekker

Supervisor: Alexander Dürre

Bachelor thesis
Mathematics
17 juli 2025

Inhoudsopgave

1 Samenvatting

Dit onderzoek gaat over lichtvervuiling. Er wordt gekeken naar welke variabelen van invloed zijn op de hoeveelheid lichtvervuiling met behulp van regressie. Hieruit blijkt dat de coëfficiënten van de variabelen van het model dat menselijk gedrag beschrijft, een grote invloed heeft op lichtvervuiling. Ook wordt er gekeken naar het verschil door de jaren heen. Hieruit is te concluderen dat lichtvervuiling sinds 2012 op een niet lineaire manier gestegen is.

Er wordt data gebruikt van NASA's DNB sensor van de VIIRS [25] vanaf het jaar 2012 voor de metingen van lichtvervuiling. Ook wordt data van CBS [12] gebruikt voor de informatie van de gemeenten.

Eerst vindt u een introductie van het probleem lichtvervuiling, met daarop volgend de subproblemen die in dit onderzoek geanalyseerd worden. Vervolgens is er een uitleg over de gebruikte datasets te vinden en belangrijke definities staan uitgelegd. Hierna wordt nog ingegaan op 'modelselectie', omdat dit begrip zeer belangrijk is in dit onderzoek. Vervolgens begint het onderzoek en zijn de observaties te vinden in Hoofdstuk 7.

De structuur van het onderzoek is als volgt. Eerst wordt een model opgesteld. Dit model wordt vervolgens verbeterd en geanalyseerd. Uit de beoordeling van het laatste model blijkt dat er nog een paar factoren missen voor een 'perfect' model. De variabelen die een positieve invloed hebben, zijn over het algemeen variabelen die een geografische, economische of demografische betekenis hebben. Precieze waardes zijn te vinden in Tabel 8 in Appendix: Tabellen 12.1.

2 Introductie probleem

Lichtvervuiling heeft grote invloed op het milieu. Het is vanzelfsprekend dat nachtdieren last hebben van het kunstmatige licht 's nachts, maar niet alleen zij worden belemmerd. Zoals Ron Chepesiuk schreef in 'Missing the dark: health effects of light pollution' zorgt het voor lange tijd bloodstellen aan kunstmatig licht dat bomen niet mee kunnen in de seizoenswisselingen. Dat heeft weer een effect op de dieren die deze bomen nodig hebben om te overleven [13]. Dit probleem bevindt zich niet alleen in steden, maar ook in het platteland omdat licht zich verspreid en niet plaatselijk voor problemen zorgt. In ditzelfde artikel beschreef hij ook op welke manier lichtvervuiling het broedproces van verschillende dieren verstoort. Een goed voorbeeld hiervan zijn zeeschildpadden. De vrouwtjes leggen hun eieren op het strand waar zij zelf geboren zijn. Als hier echter veel licht is, wordt dit nestelen verstoord. Mocht het nestelen gelukt zijn, is er een volgend probleem wat hij beschrijft. Babyzeeschildpadden oriënteren zich door de andere kant op te gaan dan een verhoogde donkere plek, het land. Echter door kunstmatig licht, zijn er andere donkere plekken, omdat er juist op het land licht is. Hierdoor raken de babyschildpadden gedesoriënteerd. Een ander voorbeeld uit ditzelfde artikel is dat trekvogels worden gedesoriënteerd door het kunstmatige licht in de nacht en verstrikken in een stad in plaats van het zuidelijke gebied waar ze heen wilden. Zo zijn er nog veel meer diersoorten die belemmerd worden door lichtvervuiling. Niet alleen op dieren, maar ook op mensen heeft lichtvervuiling negatieve effecten. Omdat er constant ligt is, wordt de circadiaanse klok verstoord, vertelt Ron Chepesiuk tevens als de problemen die dit geeft. De circadiaanse klok beïnvloedt fysiologische processen, zoals hormoonproductie of celregulatie, in ons lichaam. Dit alles zorgt er onder andere voor dat tumorgroei wordt versneld en er andere medische aandoeningen bij mensen vaker optreden zoals insommnia of depressie.

Onderzoek naar lichtvervuiling is dus van groots belang voor alle levensvormen op onze planeet. Daarom wordt er in deze scriptie onderzoek gedaan naar waar lichtvervuiling door ontstaat en of er verandering door de jaren op te merken valt.

3 Waarom is er op sommige plekken licht

Om lichtvervuiling te verminderen, is het van belang om te weten welke factoren van invloed zijn op lichtvervuiling. Daarom is het belangrijk om een model op te stellen, te kijken welke variabelen daarvan op een positieve manier invloed hebben en welke punten outliers zijn volgens dit model. De variabelen die een positieve invloed hebben in het model, dragen bij aan lichtvervuiling. Outliers zijn interessant, omdat deze punten een onverwachte hoeveelheid lichtvervuiling hebben. Aan de hand daarvan kan gekeken worden waarom een punt een andere waarde dan verwacht heeft en daarmee kan gekeken worden naar diens invloed op lichtvervuiling.

De hypothese is dat lichtvervuiling vooral problematisch is in steden en dat de variabelen omtrent inwoners een hoge positieve waarde in het model zullen hebben.

4 Uitleg datasets

4.1 tmap

De dataset tmap, thematic maps [23], is een R package waarin gegevens staan over verschillende landen over de wereld. Voor dit onderzoek wordt in het bijzonder NLD_dist gebruikt. Hierin zijn de gegevens van elke regio te verkrijgen, zoals het oppervlakte of de populatie. Zie hieronder in Tabel 1 de specifieke uitleg over de variabelen.

Variabele	Data type	Uitleg
code	categorisch	Een code die aan een district gegeven wordt. GM staat voor gemeente en WK voor wijk. De eerste vier getallen geven het gemeentelijke identificatienummer weer en de laatste twee getallen die bij een wijk voorkomen zijn het district nummer.
name	categorisch	De naam van het district.
province	categorisch	De naam van de provincie waarin het district zich bevindt.
area	numeriek	Het oppervlakte van het district in km^2.
urbanity	factor met 5 levels	Level van stedelijkheid gebasseert op het aantal adressen in km^2 opgedeeld in 5 categoriën.
population	integer	Totale populatie van het district op 1 januari 2022.
pop_0_14	numeriek	Afgerond percentage van mensen tussen de 0 en 15 in het district.
pop_15_24	numeriek	Afgerond percentage van mensen tussen de 15 en 25 in het district.
pop_25_44	numeriek	Afgerond percentage van mensen tussen de 25 en 45 in het district.
pop_45_64	numeriek	Afgerond percentage van mensen tussen de 45 en 65 in het district.
pop_65plus	numeriek	Afgerond percentage van mensen ouder dan 65 in het district.
dwelling_total	integer	Totaal aantal onderkomens in het district.
dwelling_value	integer	Gemiddelde WOZ-waarde van het district.
dwelling_ownderschip	integer	Percentage onderkomens dat eigendom is van de bewoners.
employment_rate	integer	Aandeel van de werkzame bevolking in de totale bevolking met een leeftijd tussen 15 en 75 jaar.
income_low	numeriek	Percentage individuen uit huishouden dat tot de laagste 40% van het landelijke persoonlijk inkomen behoort.
income_high	numeriek	Percentage individuen uit huishouden dat tot de hoogste 20% van het landelijke persoonlijk inkomen behoort.
edu_appl_sci	numeriek	Persentage van individuen met een leeftijd tussen 15 en 75 jaar met een HBO of universitaire diploma.

Tabel 1: Uitleg variabelen uit dataset tmap

4.2 CBS

De data van NLD_dist komt van het CBS [12]. Op de site van CBS staat nog meer informatie over de variabelen. Zo verstrekken zij informatie vanaf 2004 over wijken en buurten. De dataset van het jaar 2024 bevat 134 variabelen. Aangezien dit dus zeer veel data is, is er voor dit project gekozen om eerst met NLD_dist te werken en daarna het jaartal en de variabelen uit de CBS data te kiezen. Op deze manier kan de code sneller werken en nog steeds een goed resultaat leveren. De variabelen die gebruikt zijn, zijn in de onderstaande Tabel 2 weergegeven.

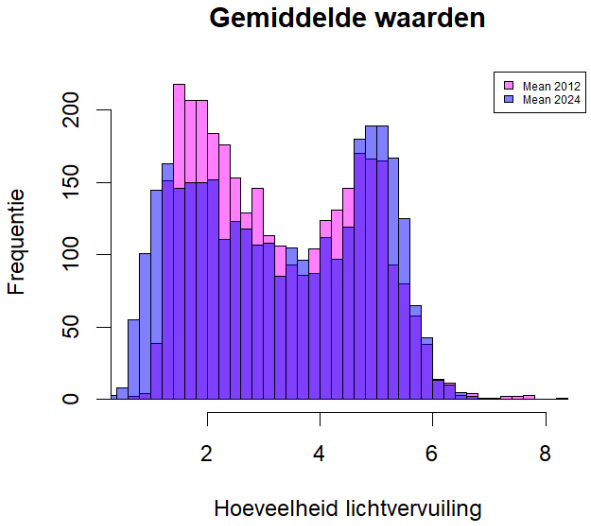
Variabele	Data type	Uitleg
gm_naam	karakter	Gemeentenaam.
a_inw	integer	Aantal inwoners van het district.
a_00_14	integer	Aantal inwoners met leeftijd tussen 0 en 15 in het district.
a_15_24	integer	Aantal inwoners met leeftijd tussen 15 en 25 in het district.
a_25_44	integer	Aantal inwoners met leeftijd tussen 25 en 45 in het district.
a_45_64	integer	Aantal inwoners met leeftijd tussen 45 en 65 in het district.
a_65_oo	integer	Aantal inwoners ouder dan 65 in het district.
a_hh	integer	Het aantal huishouden in totaal in het district.
bev_dich	integer	De bevolkingsdichtheid van het district.
a_1p_hh	integer	Aantal eenpersoonshuishoudens in het district.
a_hh_z_k	integer	Aantal huishoudens zonder kinderen in het district.
a_hh_m_k	integer	Aantal huishoudens met kinderen in het district.
g_hhgro	numeriek	Gemiddeld huishoudensgrootte in het district.
g_wozbag	integer	Gemiddelde WOZ-waarde van de woningen in het district.
a_woning	integer	De woningvoorraad in het district.
a_pau	integer	Totaal aantal personenauto's in het district.
g_pau_hh	numeriek	Gemiddeld aantal personenauto's per huishouden in het district.
g_pau_km	integer	Gemiddeld aantal personenauto's naar oppervlakte in het district.
a_m2w	integer	Aantal motorfietsen in het district.
a_bedv	integer	Totale hoeveelheid bedrijfvestigingen in het district.
a_bed_a	integer	Aantal bedrijven in de landbouw, bosbouw en visserij.
a_bed_bf	integer	Aantal bedrijven in de nijverheid en energie.
a_bed_gi	integer	Aantal bedrijven in de handel en horeca.
a_bed_hj	integer	Aantal bedrijven in het vervoer, informatie en communicatie.
a_bed_kl	integer	Aantal bedrijven in de financiële diensten en onroerend goed.
a_bed_mn	integer	Aantal bedrijven in de zakelijke dienstverlening.
a_bed_oq	integer	Aantal bedrijven in de overheid, onderwijs en zorg.
a_bed_ru	integer	Aantal bedrijven in de cultuur, recreatie en de overige diensten.
a_opp_ha	integer	Totale oppervlakte van het district.
a_lan_ha	integer	Oppervlakte land in het district.
a_wat_ha	integer	Oppervlakte water in het district.
ste_oad	integer	Omgevingsadressendichtheid.

Tabel 2: Uitleg variabelen uit data van CBS

4.3 Data NASA

De Day/Night Band (DNB) [21] sensor van de Visible Infrared Imaging Radiometer Suite, ook wel VIIRS genoemd, zorgt dagelijks voor wereldwijde data die gebruikt kan worden voor wetenschappelijk onderzoek. In 2011 is er een sateliet, genaamd S-NPP, gelanceerd. Hiermee wordt verschillende data verzameld, zoals hoe het weer invloed heeft op de urbanische infrastructuur. Sinds januari 2012 wordt ook de lichtvervuiling met deze sateliet gemeten. Na het meten wordt de data gecorrigeerd aangezien er veel factoren van invloed zijn op de gegevens. Zo worden biases verwijderd en wordt de invloed van maanlicht en sneeuwbedekt oppervlakte ingecalculeerd [25]. Deze data is in dit onderzoek gebruikt.

Aangezien de data per dag op een tijdstip gemeten wordt, en niet continu gemeten wordt, is de data discreet. De eenheid van de data is $\text{nWatt} \cdot \text{cm}^{-2} \cdot \text{sr}^{-1}$. De range van de sensor is 0 tot en met 65534. Voor locaties waar geen data gevonden is, is de waarde 65535 ingevuld. Voor dit project zijn de waarden 65535 vervangen door NA. Verder wordt er in dit project gewerkt met het logaritme van de data. Bij punten die een waarde van 0 hadden, wordt $\log(\frac{1}{2})$ gebruikt. In onderstaande histogram in Figuur 1 is te zien waar de waardes van lichtvervuiling zich gemiddeld genomen bevinden in de jaren 2012 en 2024.



Figuur 1: Histogram over gemiddelde lichtvervuiling van 2012 en 2024

5 Definities

In dit hoofdstuk worden de meest belangrijke definities toegelicht voor dit onderzoek.

5.1 AIC waarde

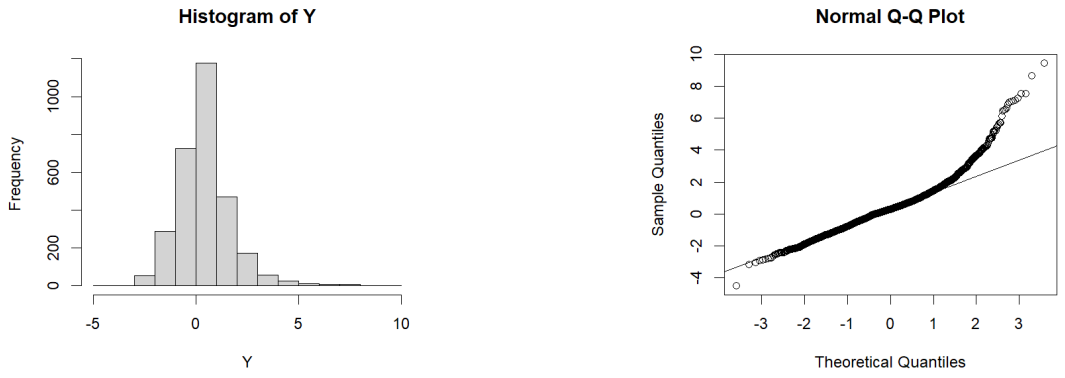
Akaike's informatie criterium, kort gezegd 'AIC', is een vorm van modelselectie. Elk model krijgt een relatieve waarde die wiskundig bepaald wordt. Aan de hand van deze waardes kan een beste model gekozen worden. Naast AIC zijn er meerdere vormen van modelselestie, hierover is meer te lezen in het hoofdstuk 'Modelselectie' 6. De formule die gebruikt wordt voor AIC is AIC $= 2p - 2\log(\hat{\mathcal{L}})$ waarbij p het aantal opgenomen variabelen zijn en $\hat{\mathcal{L}}$ de loglikelihood functie is van het geschatte model. De loglikelihood functie is een functie die helpt te bepalen hoe goed een model de data omschrijft [17]. Een lagere AIC waarde duidt op een beter model [1]. Voor dit project is modelselectie AIC gekozen omdat deze methode het informatieverlies minimaliseert [7].

5.2 Eindig gemixte verdeling

De verdelingsfunctie F_X van een eindig gemixt model kan geschreven worden als $F_X(x) = \sum_{i=1}^{n} w_i F_{X_i}(x)$ met $w_1, \ldots, w_n > 0$ waarbij $\sum_{i=1}^{n} w_i = 1$ en verdelingen $F_{X_1}, \ldots, F_{X_n}$ [20]. Gemixte modellen kunnen gezien worden als een experiment in twee stappen. In de eerste stap wordt een verdeling bepaald met kans w_i. In de tweede stap wordt een random variabele gegenereerd van de gekozen verdeling. Gemixte modellen zijn nuttig om te gebruiken voor het modelleren van afhankelijkheid tussen de waarnemingen of verschillen in de data die niet verklaard kunnen worden door vaste effecten, ook wel heterogeniteit genoemd [8, p. 306]. Om een beeld te krijgen van een gemixte verdeling, wordt er gekeken naar een verdeling Y die wordt gegenereerd in R. Hierin is er met kans $0,7$ de normale verdeling en met kans $0,3$ een exponentiële verdeling. De bijbehorende code ziet er als volgt uit.

```
Index <- rbinom(3000, size=1, prob=0.7)
Y <- Index*rnorm(3000)+(1-Index)*(1.5*rexp(3000))
```

Bij deze verdeling horen de histogram en Q-Q plot uit Figuur 2.



(a) Histogram van verdeling Y

(b) Q-Q plot van verdeling Y

Figuur 2: Visueel voorbeeld van een gemixte verdeling

5.3 Regressie

In statistiek houdt regressie in dat een afhankelijke variabele Y wordt verklaard door een onafhankelijke variabele X waardoor er voorspellingen over bijvoorbeeld opbrengst gemaakt kan worden [8, p. 268]. Binnen regressie zijn er verschillende subcategoriën. De meest gebruikelijke methode is linaire regressie. In lineaire regressie wordt over het algemeen de aanname gemaakt dat model Y normaal verdeeld is [8, pp. 270-271]. Zij het model $Y_j = \sum_{i=1}^{p} \beta_i x_{i,j} + \varepsilon_j$, $j = 1, \ldots, n$. De errors, $\varepsilon_1, \ldots, \varepsilon_n$ zijn identiek en onafhankelijk verdeeld. Voor ε geldt $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. Daarom geldt er voor het model $Y \sim \mathcal{N}(\sum_{i=1}^{p} \beta_i x_{i,j}, \sigma^2)$. Voor de kansdichtheidsfuncties $f_{Y_1,\ldots,Y_n}$ geldt dan $f_{Y_1,\ldots,Y_n}(\beta_1, \ldots, \beta_p) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2}} \exp\left(-\frac{(Y_j - \sum_{i=1}^{p} \beta_i x_{i,j})^2}{2\sigma^2}\right)$.

Dus

$$\hat{\mathcal{L}}(\beta,\sigma) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2}} \exp\left(-\frac{(Y_j - \sum_{i=1}^{p}\beta_i x_{i,j})^2}{2\sigma^2}\right) \tag{1}$$

$$= (2\pi)^{-\frac{n}{2}}(\sigma)^{-n}\exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(Y_j - \sum_{i=1}^{p}\beta_i x_{i,j})^2\right). \tag{2}$$

Voor de log-liklihood geldt dan

$$\log(\hat{\mathcal{L}}(\beta,\sigma)) = \log\left((2\pi)^{-\frac{n}{2}}(\sigma)^{-n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \sum_{i=1}^{p}\beta_i x_{i,j})^2\right)\right) \tag{3}$$

$$= -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \sum_{i=1}^{p}\beta_i x_{i,j})^2. \tag{4}$$

Het model bevat $p+1$ parameters die worden genoteerd in een vector $\theta = (\beta_1, \beta_2, \ldots, \beta_p, \sigma^2)$. De waarden van β_k worden geschat en worden ook wel schatters genoemd. De schatters kunnen op verschillende manieren geschat worden. Zo is er bijvoorbeeld de maximim likelihood-schatting, een momentenschatter en de bayes-schatter [8, p. 95]. De maximumlikelihood schatter voor β en σ^2 uit (3) worden nu afgeleden.
Voor elke β_k, $k = 1, \ldots, p$, geldt

$$\frac{\partial}{\partial\beta_k}\log(\hat{\mathcal{L}}(\beta,\sigma)) = \frac{1}{\sigma^2}\sum_{j=1}^{n} x_{k,j}(Y_j - \sum_{i=1}^{p}\beta_i x_{i,j}). \tag{5}$$

Om het maximum te bepalen, moeten deze afgeleiden gelijk gesteld worden aan 0. Dit geeft voor elke β_k de vergelijking

$$\frac{1}{\sigma^2}\sum_{j=1}^{n} x_{k,j}(Y_j - \sum_{i=1}^{p}\beta_i x_{i,j}) = 0. \tag{6}$$

Schrijf dit nu in matrixnotatie voor elke β. Definieer daarvoor de matrixnotatie $\beta = \begin{pmatrix}\beta_1 & \beta_2 & \ldots & \beta_p\end{pmatrix}$, $X = \begin{pmatrix} x_{1,1} & x_{1,2} & \ldots & x_{1,p} \\ x_{2,1} & x_{2,2} & \ldots & x_{2,p} \\ \vdots & \ddots & \ddots & \vdots \\ x_{p,1} & x_{p,2} & \ldots & x_{p,p}\end{pmatrix}$ en $Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n\end{pmatrix}$. Neem aan dat X^TX inverteerbaar is. Werk nu de vergelijkingen verder uit, dat geeft

$$\frac{1}{\sigma^2}X^T(Y - X\beta) = 0 \tag{7}$$

$$\frac{1}{\sigma^2}X^TY = \frac{1}{\sigma^2}X^TX\beta \tag{8}$$

$$X^TY = X^TX\beta \tag{9}$$

$$\beta = (X^TX)^{-1}X^TY. \tag{10}$$

Bekijk nu de partiele afgeleide naar σ^2 van (3). Schrijf daarvoor eerste σ in (3) naar σ^2.

$$\log(\hat{\mathcal{L}}(\beta,\sigma)) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \sum_{i=1}^{p}\beta_i x_{i,j})^2. \tag{11}$$

Dit afleiden naar σ^2 geeft

$$\frac{\partial}{\partial\sigma^2}\log(\hat{\mathcal{L}}(\beta,\sigma^2)) = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(Y_i - \sum_{i=1}^{p}\beta_i x_{i,j})^2. \tag{12}$$

Wederom om het maximum te krijgen, moet dit gelijk gesteld worden aan 0. Dit geeft

$$-\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(Y_i - \sum_{i=1}^{p}\beta_i x_{i,j})^2 = 0 \tag{13}$$

$$\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \sum_{i=1}^{p}\beta_i x_{i,j})^2 = \frac{n}{2} \tag{14}$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \sum_{i=1}^{p}\beta_i x_{i,j})^2. \tag{15}$$

Hiermee worden de extremen waarden gevonden. Merk op dat deze overeenkomen met de extremen waarden uit [15] en dat dit dus inderdaad maxima zijn.
De maximum likelihood estimator van β is dus gelijk aan $\hat{\beta} = (X^T X)^{-1} X^T Y$ en de maximum likelihood estimator van σ^2 is dus gelijk aan $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \sum_{i=1}^{p}\hat{\beta}_i x_{i,j})^2$.

In dit project wordt gewerkt met een lineair model. Voor de log-likelihood was de formule

$$\log(\hat{\mathcal{L}}(\beta, \sigma^2)) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \sum_{i=1}^{p}\beta_i x_{i,j})^2. \tag{16}$$

bekend. Substitueer nu de gevonden schatters $\hat{\beta}$ en $\hat{\sigma}$. Dit geeft

$$\log(\hat{\mathcal{L}}(\hat{\beta}, \hat{\sigma}^2)) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}(Y_i - \sum_{i=1}^{p}\hat{\beta}_i x_{i,j})^2. \tag{17}$$

Substitueer nu $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \sum_{i=1}^{p}\hat{\beta}_i x_{i,j})^2$ met $\hat{\sigma}^2$, dit geeft

$$\log(\hat{\mathcal{L}}(\hat{\beta}, \hat{\sigma}^2)) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}\cdot n\hat{\sigma}^2 \tag{18}$$

$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\hat{\sigma}^2) - \frac{n}{2}. \tag{19}$$

Dit geeft voor de AIC uiteindelijk

$$\text{AIC} = 2p - 2\log(\hat{\mathcal{L}}) \tag{20}$$

$$= 2p - 2\left(-\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\hat{\sigma}^2) - \frac{n}{2}\right) \tag{21}$$

$$= 2p + n\log(2\pi\hat{\sigma}^2) + n. \tag{22}$$

Vaak wordt er een intercept geïntroduceerd door $x_1 = 1$. Wanneer de verklarende variabele $X = \{x_1, \ldots, x_p\}$ enkelvoudig is, wordt het model een enkelvoudig lineair regressiemodel genoemd [8, p. 272]. Het model wordt dan beschreven door $Y_i = \alpha + \beta x_i + \varepsilon_i$ voor $i = 1, \ldots, n$. De variantie blijft hetzelfde. Merk op dat het model een lineaire vergelijking is als er geen meetfouten gemaakt zouden worden. De waarde van α wordt bepaald naar aanleiding van de intercept.

5.4 Scatterplots

Scatterplots zijn visuele representaties van belangrijke informatie zoals van de relatie tussen twee variabelen of de invloed van ourliers [22]. Door op de horizontale as de waarden van de ene variabele te plaatsen en op de verticale as de waarden van de andere variabele, kan elk datapunt in de grafiek geplaatst worden. Door elk datapunt te plaatsen in deze grafiek, kan er naar aanleiding van hoe de grafiek eruit is komen te zien, gekeken worden of er een verband is tussen de twee variabelen [24]. Als een punt afwijkt van een eventuele trend die op te merken is in de scatterplot, is dit punt een outlier.

5.4.1 Residuals vs. Leverage Plot

De Residuals vs. Leverage plot wordt gebruikt om eventuele invloedrijke waarnemingen te identificeren. In de plot zijn gestippelde lijnen te zien. Deze stellen een afstand van Cook voor, afhankelijk welke waarde erbij staat. De afstand van Cook, D_i, wordt berekend met de formule $D_i = \left(\frac{r_i^2}{p \cdot \text{MSE}} \cdot \frac{h_{ii}}{(1-h_{ii})^2}\right)$ waarbij r_i het i'de residu is, p het aantal coëfficiënten in het regressie model is, MSE de *Mean Squared error* is en h_{ii} de *leverage value* van punt i is [26]. Als datapunten buiten deze lijnen vallen, zijn de betreffende datapunten invloedrijke waarnemingen [27].

5.4.2 Q-Q plot

Een quantile-quantile plot, kortgezegd Q-Q plot, wordt gebruikt om te kijken of een model van een verdeling afstamt [28]. Meestal wordt een Q-Q plot gebruikt om te kijken of het model normaal verdeeld zou kunnen zijn. Op de horizontale as staat de theoretische waarde en op de verticale as de gemeten waarde. Als de punten dan rond de diagonale as, $y = x$, liggen, lijkt de aangenomen verdeling goed te passen. Een Q-Q plot is geen wetenschappelijk bewijs, maar kan inzicht geven in hoe de data in elkaar steekt.

5.4.3 Residuals vs Fitted plot

De Residuals vs. Fitted plot wordt gebruikt om te beoordelen of residuen niet-lineaire patronen vertonen [27]. In een Residuals vs. Fitted plot is een (rode) lijn te zien rond de 0 op de verticale as. Als deze lijn grofweg horizontaal loopt, wordt de aanname gemaakt dat de residuen een lineair patroon volgen [27]. Hiervoor wordt op de verticale as de Pearson Residu gebruikt. Zij o_{ij} de geobserveerde waarde van het datapunt in kolom i en rij j. Zij v_{ij} de verwachte waarde van het datapunt in kolom i en rij j. Dan is het Person Residu r_{ij} van het datapunt in kolom i en rij j gelijk aan $r_{ij} = \frac{o_{ij} - v_{ij}}{\sigma}$ met σ gelijk aan de verwachte standaarddeviatie van de fouttermen [19]. Merk op: Hoe dichter de Person Residu bij 0 ligt, hoe kleiner het verschil is tussen de verwachte en geobserveerde waarde.

6 Modelselectie

In stapsgewijze modelselectie wordt er over het algemeen een keuze gemaakt tussen forward en backward selectie. Deze modelselectiemethoden worden gecombineerd met een informatiecriteria om uiteindelijk een zo goed mogelijk model te bepalen.

In de richting 'forward' [30] begin je met een model zonder predictor variabelen. Hierop pas je het gekozen informatiecriterium op toe, bijvoorbeeld AIC. Vervolgens wordt bij elk van je mogelijke modellen met één predictor de waarde van het informatiecriterium berekend. Hiervan bepaal je welk model een laagste waarde heeft volgens jouw gekozen informatiecriterium en daarmee het model verbetert ten opzichte van het voorgaande model zonder predictoren. Als het model met predictor X_i het beste model geeft, ga je alle mogelijke modellen af met twee predictoren waarvan X_i een van deze predictoren is. Je kiest het model met twee predictoren die verbeterd is ten opzicht van het voorgaande model met één predictor. Hierna check je de modellen met drie predictoren, waarvan er twee al vast staan zo gaat dit proces door totdat de modellen met meer predictoren niet meer leidt tot een beter model.

In de richting 'backward' [29] wordt er met een volledig model met alle mogelijke predictoren begonnen en worden deze stapsgewijs verwijderd. De waarde van het gekozen informatiecriterium van dit volledige model wordt berekend. Vervolgens worden alle modellen met een predictor minder opgesteld en wordt de waarde van het informatiecriterium berekend. Een model die zorgt voor een nieuw beste model wordt voor de volgende stap gebruikt. Van dit model wordt wederom elke mogelijkheid met bijbehorende waarde voor het informatiecriterium bepaald en een beste model wordt gekozen. Dit proces zet zo door totdat het verwijderen van predictoren niet meer zorgt voor een verbetering van het model.

De twee meest voorkomende informatiecriteria zijn Aikane's Information Criterion, AIC, en Bayesian information criterion, BIC [18]. Andere vormen van informatiecriteria zijn bijvoorbeeld Deviance Information Criterium, Focused Information Criterium, Likelihood ratio test en zo gaat de lijst met informatiecriteria nog langer door.
Zoals in de definities al gezegd was, is de formule voor AIC $\text{AIC} = 2p - 2\log(\hat{\mathcal{L}})$ met p het aantal opgenomen variabelen en $\hat{\mathcal{L}}$ de loglikelihood functie 5.1. De formule voor BIC is $\text{BIC} = p\log(n) - 2\log(\hat{\mathcal{L}})$ [1]. Wederom zijn p en $\hat{\mathcal{L}}$ respectievelijk het aantal opgenomen variabelen en de maximum likelihood. De factor $\log(n)$ is het logaritme van het totale aantal variabelen die het model op zou kunnen nemen.
De criteria AIC en BIC hebben beide een penalty. Respectievelijk $2|p|$ en $|p|\log(n)$. De penalty voor BIC is dus met een factor van $\frac{1}{2}\log(n)$ groter dan de penalty van AIC. Voor modellen met veel mogelijke variabelen, en dus een grote n waarde, wordt de penalty van BIC aanzienlijk groter als er ook veel variabelen opgenomen worden. Daarom resulteert BIC vaker in kleinere modellen dan AIC [8, p. 334]. Aangezien er in dit project interesse is in zoveel mogelijk redenen voor lichtvervuiling, is er gekozen voor AIC.

Zowel backward als forward selection geven een zo goed mogelijk model, maar dit hoeft niet het meest optimale model te zijn. In dit onderzoek is daarom gekozen om elk mogelijk model op te stellen met een functie in R. Dit kost meer tijd, maar geeft met zekerheid het beste model. Qua tijd had het effectiever gekund door wel forward of backward selectie te gebruiken. Uiteindelijk werden de modellen gerangschikt op laagste AIC waarde. Hiermee werd het beste model gekozen.

6.1 Afleiding AIC

In 1973 publiceerde Akaike een artikel waarin de Kullback-Leibler divergentie of afstand als basis voor modelselectie werd gebruikt [9]. De Kullback-Leibler afstand is een maat om de afstand tussen twee kansdichtheden te bepalen, bekeken vanuit een van deze kansdichtheden [10]. Met behulp van het boek *Model Selection and Multimodel Inference* geschreven door K.P. Burnham en D.R. Anderson [9], het hoofdstuk *Akaike's Information Criterion: Background, Derivation, Properties, and Refinements* uit encyclopedie *International Encyclopedia of Statistical Science* [11] en de lesaantekeningen van Alexander Dürre van het college Inleiding Mathematische Statistiek op Universiteit Leiden [16], is de formule die Akaike afgeleid heeft, hieronder ook afgeleid.

Er wordt aangenomen dat gegeven een parametrisch structureel model, er een unieke waarde I voor θ is die de K-L afstand, D_{KL}, minimaliseert [9]. Deze unieke waarde I is afhankelijk van de functie f (hieronder verder uitgewerkt), het model M met kansdichtheidsfunctie g, de parameterruimte Θ en de steekproefruimte $\mathbb{X} = \{X_1, \ldots, X_n\} \in \mathbb{R}^n$ met $X_1, \ldots, X_n$ onafhankelijk en identiek. Neem verder ook aan dat er een multivariate verdeling $f : \mathbb{R}^n \to \mathbb{R}$ is. Aangezien $X_1, \ldots, X_n$ onafhankelijk en identiek verdeeld zijn, geldt er $f(X_1, \ldots X_n) = \prod_{i=1}^n \tilde{f}(X_i)$. Zo geldt er ook voor de kansdichtheidsfuncties voor model M dat $g(X_1, \ldots X_n) = \prod_{i=1}^n \tilde{g}_\theta(X_i)$ met $\theta \in \mathbb{R}^p$. Neem aan dat er een $\theta_0 \in \mathbb{R}^p$ bestaat waarvoor geldt $g_{\theta_0}(X_1, \ldots, X_n) = f(X_1, \ldots, X_n)$.

Uit [9] kan de volgende formule voor D_{KL} gevonden worden

$$D_{KL}(f||g_\theta) = \int f(x) \log\left(\frac{f(x)}{g_\theta(x)}\right) dx. \tag{23}$$

Gezocht wordt θ die D_{KL} minimaliseert. Met rekenregels van logaritmen geldt dan

$$D_{KL} = \int f(x) \log(f(x)) dx - \int f(x) \log(g(x|\theta)) dx. \tag{24}$$

Omdat $\int f(x) \log(f(x)) dx$ onafhankelijk is van θ, heeft deze term geen invloed op de gezochte θ die D_{KL} minimaliseert. Dus als modellen M_1 en M_2 vergeleken worden is $\int f(x) \log(f(x)) dx$ irrelevant. Daarom moet er gefocust worden op $-\int f(x) \log(g(x|\theta)) dx = -E_f(g_\theta(\mathbb{X}))$. Omdat f onbekend is, kan $E_f(g_\theta(\mathbb{X}))$ niet gemaximaliseerd worden ten opzichte van θ. Neem aan dat $E_f(g_\theta(\mathbb{X}))$ differentieerbaar is, dan moet

$$\left.\frac{\partial E(g_\theta(\mathbb{X}))}{\partial \theta}\right|_{\theta_0} = 0 \tag{25}$$

gelden. Als de integraal en differentiatie verwisseld kunnen worden, geldt

$$E_f\left(\frac{\partial}{\partial \theta} \log(g_\theta(x))|_{\theta_0} \, dx\right) = 0. \tag{26}$$

Neem aan dat dit kan.

Zij θ_{ML} de maximum likelihood estimator van M. Dan maximaliseert θ_{ML} de term $\log(g_\theta(\mathbb{X}))$ ten opzichte van θ, dus

$$\left.\frac{\partial E(g_\theta(\mathbb{X}))}{\partial \theta}\right|_{\theta_{ML}} = 0. \tag{27}$$

Omdat $\min_\theta E_f(\log(g_\theta(\mathbb{X}))$ niet berekend kan worden, wordt er gekeken naar $E_f(\log(g_\theta(\mathbb{X}))|_{\theta=\theta_{ML}}$. Echter is deze term biased. Noteer, net als in [11], $E(E_f(\log(g_\theta(\mathbb{X}))|_{\theta=\theta_{ML}})$ als volgt

$$\begin{aligned} E(E_f(\log(g_\theta(\mathbb{X}))|_{\theta=\theta_{ML}}) =& E_f(\log(g_{\theta_{ML}}(\mathbb{X}))) - E_f(\log(g_{\theta_{ML}}(\mathbb{X})) + \\ & E_f(\log(g_{\theta_0}(\mathbb{X}))) - E_f(\log(g_{\theta_0}(\mathbb{X}))) + E(E_f(\log(g_\theta(\mathbb{X}))|_{\theta=\theta_{ML}}) \end{aligned} \tag{28}$$

$$\begin{aligned} =& E_f(\log(g_{\theta_{ML}}(\mathbb{X}))) + (E_f(\log(g_{\theta_0}(\mathbb{X}))) - E_f(\log(g_{\theta_{ML}}(\mathbb{X}))) + \\ & (E(E_f(\log(g_\theta(\mathbb{X}))|_{\theta=\theta_{ML}}) - E_f(\log(g_{\theta_0}(\mathbb{X})))) . \end{aligned} \tag{29}$$

Herinner de tweede orde Taylorreeks

$$h(\theta) = h(\theta_0) + \left[\frac{\partial h(\theta_0)}{\partial \theta}\right]' [\theta - \theta_0] + \frac{1}{2} [\theta - \theta_0]^T \left[\frac{\partial^2 h(\theta_0)}{\partial \theta^2}\right] [\theta - \theta_0] + Re. \tag{30}$$

Hierin is Re van orde $O\left(||\theta - \theta_0||^3\right)$ waarbij $||\theta - \theta_0|| = \sqrt{\sum_{i=1}^p (\theta_i - (\theta_0)_i)^2}$ de Euclidische afstand is, zoals in [9]. Hierin is $\theta_0 \in \Theta$ is de minimale K-L waarde over Θ. Beschouw ook de Hessian van $h(\theta)$.
Beschouw nu uit (29) het eerste verschil

$$(E_f(\log(g_{\theta_0}(\mathbb{X}))) - E_f(\log(g_{\theta_{ML}}(\mathbb{X}))) . \tag{31}$$

Benader dit met de tweede orde Taylorreeks, weergegeven in (30). Dit geeft

$$\begin{aligned} & E_f(\log(g_{\theta_0}(\mathbb{X}))) - E_f(\log(g_{\theta_0}(\mathbb{X})) + E_f\left(\left[\left.\frac{\partial \log(g_\theta(\mathbb{X}))}{\partial \theta}\right|_{\theta_0}\right] [\theta_{ML} - \theta_0]\right) + \\ & E_f\left(\frac{1}{2} [\theta_{ML} - \theta_0]^T \mathcal{I}(\theta_0, \mathbb{X}) [\theta_{ML} - \theta_0]\right) + E_f(Re_1). \end{aligned} \tag{32}$$

met $\mathcal{I}_{i,j}(\theta_0, \mathbb{X}) = \left[\frac{\partial^2 \log(g_{\theta_0}(\mathbb{X})}{\partial \theta_i \partial \theta_j}\right]$, $i, j = 1, \ldots, p$, de geobserveerde Fisher informatie matrix. Wegens (26) geldt $E_f\left(\left[\left.\frac{\partial \log(g_\theta(\mathbb{X}))}{\partial \theta}\right|_{\theta_0}\right] [\theta_{ML} - \theta_0]\right) = 0$. Aangenomen wordt dat $E_f(Re_1)$ naar 0 convergeert. Bekend is dat

$[\theta_{ML} - \theta_0]^T \mathcal{I}(\theta_0, \mathbb{X}) [\theta_{ML} - \theta_0]$ in verdeling naar de Chi-verdeling convergeert met p degrees of freedom [11]. Daarom geldt $E_f([\theta_{ML} - \theta_0]^T \mathcal{I}(\theta_0, \mathbb{X}) [\theta_{ML} - \theta_0]) = p$, dus

$$E\left(\frac{1}{2}[\theta_{ML} - \theta_0]^T \mathcal{I}(\theta_0, \mathbb{X}) [\theta_{ML} - \theta_0]\right) = \frac{1}{2}E\left([\theta_{ML} - \theta_0]^T \mathcal{I}(\theta_0, \mathbb{X}) [\theta_{ML} - \theta_0]\right) \tag{33}$$

$$= \frac{1}{2}p \tag{34}$$

Bekijk nu het tweede verschil uit (29)

$$(E(E_f(\log(g_\theta(\mathbb{X}))|_{\theta=\theta_{ML}}) - E_f(\log(g_{\theta_0}(\mathbb{X}))). \tag{35}$$

Benader dit wederom met de tweedeorde Taylorreeks uit (30). Dit geeft

$$E_f(\log(g_{\theta_0}(\mathbb{X}))) + E\left(E_f\left(\frac{\partial}{\partial\theta}\log(g_\theta(\mathbb{X}))\Big|_{\theta_0}\right)[\theta_{ML} - \theta]\right) + \tag{36}$$

$$E_f\left(\frac{1}{2}[\theta_{ML} - \theta]^T I(\theta_0)[\theta_{ML} - \theta]\right) + E_f(Re_2) - E_f(\log(g_{\theta_0}(\mathbb{X})))$$

met $[I(\theta_0)] = E\left(\frac{\partial^2 \log(g_\theta(\mathbb{X}))}{\partial\theta_i \partial\theta_j}\Big|_{\theta_0}\right)$, $\theta_k = \theta_1, \ldots, \theta_p$ voor $k \in \{i, j\}$, de verwachte Fisher informatie matrix. Wederom geldt er wegens (26) dat $E_f\left(\left[\frac{\partial \log(g_\theta(\mathbb{X}))}{\partial\theta}\Big|_{\theta_0}\right][\theta_{ML} - \theta_0]\right) = 0$. En wederom wordt aangenomen dat $E_f(Re_2)$ naar 0 convergeert. Bekend is dat ook $[\theta_{ML} - \theta_0]^T I(\theta_0) [\theta_{ML} - \theta_0]$ in verdeling naar de Chi-verdeling convergeert met p degrees of freedom. Daarom geldt $[\theta_{ML} - \theta_0]^T I(\theta_0) [\theta_{ML} - \theta_0] = p$, dus

$$E\left([\theta_{ML} - \theta_0]^T I(\theta_0) [\theta_{ML} - \theta_0]\right) = \frac{1}{2}E\left([\theta_{ML} - \theta_0]^T I(\theta_0) [\theta_{ML} - \theta_0]\right) \tag{37}$$

$$= \frac{1}{2}p. \tag{38}$$

Dus door (34) en (38) in (29) in te vullen, wordt het volgende verkregen

$$E_f(\log(g_{\theta_{ML}}(\mathbb{X}))) + (E_f(\log(g_{\theta_0}(\mathbb{X}))) - E_f(\log(g_{\theta_{ML}}(\mathbb{X}))) + \tag{39}$$

$$(E(E_f(\log(g_\theta(\mathbb{X}))|_{\theta=\theta_{ML}}) - E_f(\log(g_{\theta_0}(\mathbb{X})))) = E_f(\log(g_{\theta_{ML}}(\mathbb{X}))) + \frac{1}{2}p + \frac{1}{2}p$$

$$= E_f(\log(g_{\theta_{ML}}(\mathbb{X}))) + p. \tag{40}$$

De bias van $\log(\hat{\mathcal{L}})$ is dus p. Om de te zorgen dat $E_f(\log(g_\theta(\mathbb{X})))$ zonder unbiased is, kan $\log(\hat{\mathcal{L}}) - p$ gebruikt worden. Merk op dat er eerder voor de loglikelihood de notatie $\log(\hat{\mathcal{L}})$ gebruikt werd. Om zo dicht bij de chi-square schatter te blijven, wordt de factor -2 toegevoegd [2]. Verkregen wordt dus

$$\text{AIC} = 2p - 2\log(\hat{\mathcal{L}}(\theta)). \tag{41}$$

7 Observaties

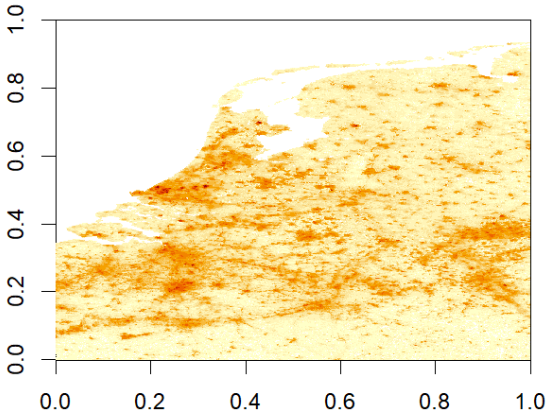
In dit hoofdstuk zullen de observaties vermeld worden. Er is vooral aandacht besteed naar de invloed van verschillende factoren op lichtvervuiling.

7.1 Waarom is er op sommige plekken meer licht dan andere?

Om te kijken waarom er op sommige plekken meer licht is dan op andere plekken, wordt er allereerst een model opgesteld. Dit model wordt vervolgens verbeterd en geobserveerd. Daarna wordt additionele data toegevoegd om het model verder te verbeteren en tot slot wordt er gekeken naar de schatters van het model.

7.1.1 Model opstellen

Het is vanzelf sprekend dat er op sommige plekken meer lichtvervuiling zal zijn dan op andere plekken. Maar hoe zit dit precies? Gevoelsmatig zal men zeggen dat er in steden meer licht is. Als er gekeken wordt naar een visuele weergave van de hoeveelheid licht gemeten op 3 januari 2025 is het op te merken dat er rondom steden inderdaad veel lichtuitstoot is.



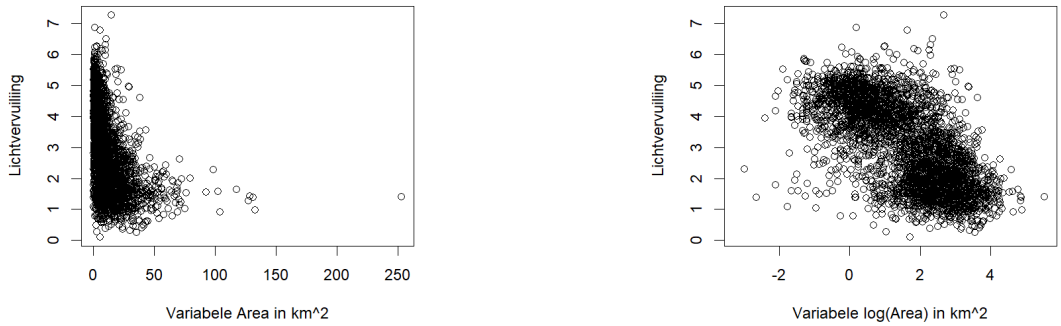
Figuur 3: Hoeveelheid licht 3 januari 2025

Om te kijken hoe dat precies zit, is het belangrijk een model op te stellen. Met R-code 'model opstellen' 11.5 heeft R het beste model berekent door de mogelijke modelen op te stellen en de AIC waarde te vergelijken. De beste drie modellen verschillen niet veel qua AIC waarde 5.1. Zie hiervoor tabel 3.

Variabele	Eerst aangegegeven model	Tweede aangegegeven model	Derde aangegegeven model
area	X	X	X
province	X	X	X
urbanity	X	X	X
population	X		
dwelling_total	X		X
employment_rate	X	X	X
edu_appl_sci	X	X	X
pop_15_24	X	X	X
pop_25_44	X	X	X
pop_65plus	X	X	X
dwelling_value	X	X	X
dwelling_ownership	X	X	X
income_high	X	X	X
AIC	6089, 700	6090, 100	6090, 354

Tabel 3: Weergave van de beste drie modellen die uit de code komen.

Te zien in Tabel 3 is dat een van de verschillen van deze drie modellen betrekking heeft op covariaat *population*. Deze covariaat komt alleen in het eerste model voor. Een ander verschil is dat de covariaat *dwelling_total* niet voorkomt in het tweede model. Verder waren deze drie modellen hetzelfde. Met de uitkomst van het eerste model kan het model verder zo goed mogelijk geoptimaliseerd worden door elke covariaat apart te bekijken. Een voorbeeld van een covariaat die beter logaritmisch benaderd kan worden is *area*. Dit is te zien in de bijbehorende scatterplots in Figuur 4.



(a) Scatter plot area

(b) Scatter plot log(area)

Figuur 4: Scatterplots waarbij het oppervlakte uitgezet wordt tegenover de gemiddelde lichtvervuiling.

Door elke scatterplot van elke opgenomen covariaten uitgezet tegen de gemiddelde lichtvervuiling af te gaan, kan er een keuze gemaakt worden over hoe de covariaat in het model terecht komt. Voor dit onderzoek is gekozen om het logaritme van een covariaat op te nemen, de covariaat zelf of de covariaat in het kwadraat. Ook zijn er twee covariaten gecreëerd en toegevoegd. In de covariaat *population/area* wordt er een concreter getal gegeven aan de bevolkingsdichtheid. Met de covariaat *valown* wordt de gemiddelde woonwaarde vermenigvuldigd met het percentage woningen wat in bezit is van bewoners. Hiermee is uiteindelijk een model opgesteld die heuristisch bepaald is. Dit model wordt model 8 genoemd. In de onderstaande Tabel 4 is te zien hoe model 8 eruit ziet ten opzichte van model 1.

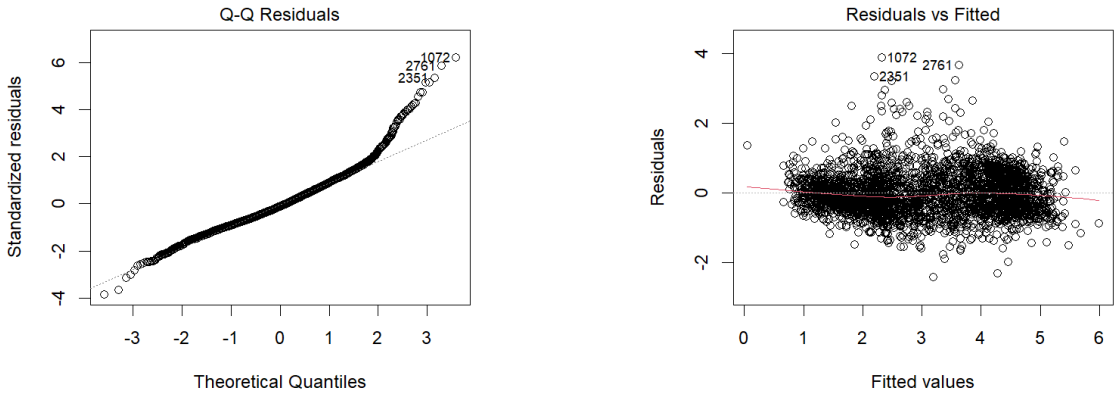
Variabele	Model 1	Model 8
area	X	log(X)
province	X	X
urbanity	X	X
population	X	log(X)
dwelling_total	X	log(X)
employment_rate	X	X^2
edu_appl_sci	X	X
pop_15_24	X	log(X)
pop_25_44	X	log(X)
pop_65plus	X	X
dwelling_value	X	X
dwelling_ownership	X	X
income_high	X	X
population/area		X
valown		X
AIC	6089, 700	5764, 7622

Tabel 4: Weergave van model 1 en dit model verbeterd, genaamd model 8.

De code tot het ontstaan van het vernieuwe model 8 is te vinden in Appendix: R code 11.5.1.
Om verder naar variabelen omtrent lichtvervuiling te kijken, wordt voor nu verder gewerkt met model 8 uit Tabel 4.

7.1.2 Beoordeling model

Om het model te beoordelen, wordt gekeken naar de scatterplots. Deze zijn te zien in Figuur 5.



(a) Q-Q plot

(b) Residuals vs Fitted plot

Figuur 5: Plots over model 8.

In deze plots zijn een paar van de outliers te zien. Met de onderstaande R-code (ook te zien in de Appendix R code 11.5.2) is het gemakkelijk te zien welke regio bij welk datapunt hoort.

```
print(NLD_dist$name[2351])
```

Bovenstaand in Figuur 5 is ook de Q-Q plot van het model weergegeven. Wat opvalt is dat voor hogere verwachte waardes, de waarde nog hoger uitvalt. Een mogelijkheid is dat het model een gemixte verdeling heeft 5.2. Een andere mogelijkheid zou kunnen zijn dat er één of meerdere covariaten missen. Om hierachter te komen, kan er ingezoomd worden op de outliers.

Datapunt	Naam	Regio	Informatie over plaats
2351	'Wijk 33 Bleiswijk Buiten'	Langsingerland, Zuid-Holland	In deze wijk bevinden zich redelijk veel bedrijven [4].
2751	'Wijk 18 Saasveld'	Dinkelland, Overijssel	
1072	'Aalsmeerderbrug/ Oude Meer/ Rozenburg / Schiphol Rijk'	Haarlemmermeer, Noord-Holland	In dit district bevinden zich relatief veel bedrijven en kantoren [5].
100	'Wijk 00'	Zeewolde, Flevoland	
2052	'De Leij'	Tilburg, Noord-Brabant	
3293	'Bedrijventerrein Heerhugowaard'	Gemeente Dijk en Waard,Noord-Holland	Deze wijk is vooral gericht op industrie. Het aantal inwoners is in deze wijk de laatste jaren erg gestegen. [6]

Tabel 5: Outliers

Op te merken uit Tabel 5 valt dat er onder andere een bedrijven terrein en regio Schiphol onderdeel van de outliers zijn. Daarom wordt er gekeken of er iets opvalt aan de rest van de outliers. Het zou mogelijk zijn dat er variabelen over bedrijven mist. Daarom wordt nu de data van CBS uit het jaar 2024 gebruikt.

7.1.3 Model verbeteren

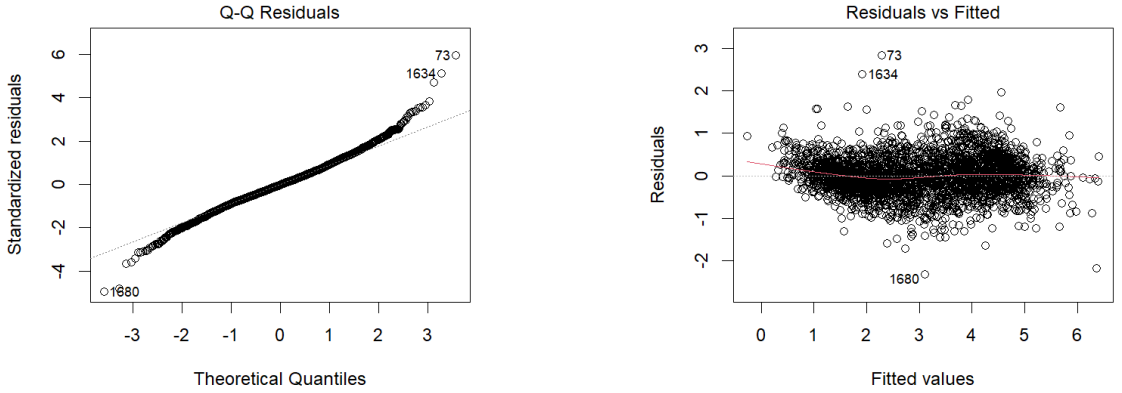
Aangezien de data van het CBS 4.2 134 variabelen bevat, wordt er eerst handmatig gekeken naar welke variabalen interessant zijn. Vervolgens wordt hierop de eerder gebruikte R-code 11.5 toegepast om een begin model op te stellen. Uit voorkennis van het eerder opgestelde model uit 7.1.1, zijn al een aantal variabelen met enige zekerheid in het model opgenomen. Hiervoor is een educatieve keuze gemaakt. Deze variabelen, genoemd 'toevoeging', worden voor nu buiten beschouwing gelaten en achteraf aan het model toegevoegd. De code is te vinden in Appendix: R code 11.5.3. Dit geeft model 1a met een AIC waarde van $4985,49$. Vervolgens zijn er variabelen die overeenkomen met model 8 toegevoegd, dit zorgde voor model 1b. Hierna is wederom handmatig gekeken naar de scatterplots van elke variabele uitgezet tegen de gemiddelde lichtvervuiling om een keuze te maken hoe de variabele in het model opgenomen zou worden. Dit leverde model 2b op. De uitkomsten van deze drie modellen zijn in onderstaande Tabel 6 te vinden. Ook is in de tabel te zien hoe de variabele eventueel in Model 8 heette.

Variabele	Model 8	Model 1a	Model 1b	Model 2b
a_inw	Population		X	log(X)
gm_naam	Province	X	X	X
a_hh_z_k		X	X	log(X)
a_hh_m_k		X	X	X
g_hh_gro		X	X	X
a_pau		X	X	log(X)
g_pau_hh		X	X	X
g_pau_km		X	X	log(X)
a_opp_ha	log(Area)	X	X	X
ste_oad		X	X	log(X)
a_bed_a			X	X
a_bed_bf			X	X
a_bed_gi			X	X
a_bed_hj			X	X
a_bed_kl			X	X
a_bed_mn			X	X
a_bed_oq			X	X
a_bed_ru			X	X
g_wozbag	dwelling_value		X	X
bev_dich	urbanity & population/area		X	X^2
a_hh			X	log(X)
a_00_14			X	log(X)
a_15_24	pop_15_24		X	X
a_25_44	pop_25_44		X	log(X)
a_65_oo	pop_65plus		X	X
AIC	5764, 7622	4985, 49	4789, 515	4388, 295

Tabel 6: Weergave van model 1 en dit model verbeterd, genaamd model 8.

Opmerkingen: De variabele *a_woning* toevoegen, gaf een hogere AIC waarde. Deze variabele kwam het meest overeen met dwelling_total uit model 8 in de tabel. De data over werknemers uit de CBS data waren voids. Deze data is daarom niet meer meegenomen in de vernieuwe modellen met de nieuwe data. Hetzelfde geldt voor het percentage hoger opgeleiden (variabele *edu_appl_sci*) en percentage van het hoogste inkomen (variabele *income_high*). In de CBS data zijn de woningen in veel verschillende categoriën verdeeld. Een categorie die goed aansloot bij variabele *dwelling_ownership* was niet gevonden. Deze is daarom weggelaten. Er is gekozen geen datasets te combineren. Hierdoor zijn er dus verschillen te observeren in de toegevoegde variabelen en de variabelen die in model 8 meegenomen waren.

Bij dit model worden wederom diagnostische plots geplot. Deze staan weergegeven in Fig 6.



(a) Q-Q plot

(b) Residuals vs Fitted plot

Figuur 6: Plots over model 2b.

Uit de scatterplots van model 2b is nieuwe informatie te halen. Zo zijn er nieuwe outliers te zien. Als de scatterplots over model 2b vergeleken worden met de scatterplots over model 8, is op te merken dat de lijn in de Q-Q plot meer lineair loopt in model 2b. Dit duidt op een beter passend model. Ook is op te merken dat

de rode lijn in de Residuals vs Fitted plot in model 8 een meer horizontale rode lijn heeft. R licht in deze twee scatterplots drie outliers uit. Wederom is hier een tabel van gemaakt.

Datapunt	Naam	Regio	Informatie over plaats
73	'Buitenvaart'	Flevoland	Klein gebied naast Almere.
1634	'Niewdorp'	Zeeland	Bevolkingsdichtheid van 96.[1]
1680	'Brigdamme'	Zeeland	Klein dorp naast Middelburg. 575 inwoners in 2024 [3]

Tabel 7: Outliers

```
1 print(Year2024$bev_dich[1634])
```

Opvallend aan de outliers is dat het kleine buurten zijn. Ook zijn twee van deze punten naast een stad. Aangezien het verspreiden van licht niet stopt bij de grens van een district, kan het zo zijn dat het licht van de steden invloed heeft op deze regio's. Om het model te verbeteren zou nog verder gezocht kunnen worden naar andere variabelen, zoals de locatie van broeikassen in een regio of de infrastructuur van een regio. Ook zou er gekeken kunnen worden naar spatial regression. Dit wordt niet meer in dit onderzoek gedaan, maar kan wel gebruikt worden voor vervolgonderzoek. Hier mag op het moment nog geen conclusie over getrokken worden. Wel kan er gekeken worden naar de invloed van de variabelen op het model. In Appendix: Tabellen 12.1 is in Tabel 8 te zien wat de invloed is van de variabelen op het model. Opvallend is dat de gemeenten Westland de grootste positieve schatter heeft in het model, gevolgd door gemeenten Midden-Delfland en Pijnacker-Nootdorp. Op te merken valt dat er veel broeikassen in deze regio's te vinden zijn. De gemeente Smallingerland heeft de grootste negatieve schatter in het model, gevolgd door gemeenten Deurne en Doesburg. Van de andere variabelen heeft het gemiddelde aantal personenauto's per huishouden de grootste positieve geschatte coëfficiënt en het logatritme van het aantal personenauto's de grootste negatieve geschatte coëfficiënt.

7.1.4 Conclusie

Al met al kan er geconcludeerd worden dat de gemeenten met veel lichtvervuiling plaatsen zijn waar veel mensen wonen. Dit is terug te zien in de hogere lichtvervuiling waar in steden sprake van is. Verder zijn locaties met veel bedrijven ook hoger in lichtvervuiling. Alle oorzaken voor hogere lichtvervuiling zijn nog niet gevonden. Hiervoor kan nog onder anderen naar infrastructuur en broeikassen gekeken worden.

8 Conclusie

Dit onderzoek stond in het teken van het gedrag van lichtvervuiling, daarom ook de hoofdvraag; *Wat is het gedrag van lichtvervuiling?* Hiervoor werd in het bijzonder gekeken naar welke variabelen van invloed zijn op lichtvervuiling. Hieruit bleek dat variabelen die invloed hebben veelal demografisch of variabelen met betrekking tot bedrijven zijn. Uit de samenvatting van dit model bleek dat de logaritme van variabele *aantal huishouden zonder kinderen* de grootste negatieve schatter had. De variabele met de grootste positieve geschatte coëfficiënt heeft *gemeente Westland*. Verder is het opvallend dat het meenemen van variabelen over bedrijven het model verbeterd, ondanks dat de schatter van deze variabelen niet groot is. Al met al heeft de locatie waar de mens zich bevindt de meeste invloed op lichtvervuiling, of dit bedrijventerreinen zijn of steden. Echter moet niet vergeten worden dat licht niet ophoudt bij de grens van een district. Lichtvervuiling is overal een probleem, niet alleen in steden, maar op drukkere gebieden vallen wel hogere waarden op te merken. De meeste gemeenten hebben een positieve schatter in het model op lichtvervuiling en variabelen omtrent mensen, zoals aantal inwoners, ook. Echter is er geen direct verband op te merken tussen het aantal inwoners en de hoeveelheid lichtvervuiling. Wel heeft de mate van stedelijkheid grote invloed op de hoeveelheid lichtvervuiling.
Het gedrag van lichtvervuiling valt samen te vatten als toenemend met invloed van waar de mens zich bevindt. Het advies is om de regering een akkoord op te laten stellen die voor elke gemeente geldt om lichtvervuiling te verminderen. Ook is het belangrijk om verschillende mogelijke manieren om lichtvervuiling te verminderen te bestuderen om te kijken hoe er zo min mogelijk vervuiling komt. Dit kan gedaan worden door testregio's een maatregel te laten testen en een andere regio's als controleregio te laten. Vervolgens kan er dan gekeken worden naar de percentele verandering over de jaren heen.

9 Discussie

De data uit dit onderzoek is afkomstig van NASA en CBS. NASA is een grote organisatie die zich bezig houdt met wetenschap en ruimte. De data over lichtvervuiling die hiervan komt is dus betrouwbaar. Het CBS is het centraal bureau voor statistieken. Dit bedrijf houdt zich bezig met vele statistieken van onder andere economie en maatschappij. Het CBS is daarom zeer betrouwbaar als het gaat over inwonergegevens. De data is dus valide. De modellen zijn opgesteld met behulp van programma RStudio en zijn daarom ook zorgvuldig opgesteld. De resultaten en dit onderzoek zijn daarom valide.

Uit het onderzoek kwam dat lichtvervuiling hoger uitkomt in steden en bedrijventerreinen. Dit komt overeen met wat verwacht zou worden.

In dit onderzoek is op een gegeven moment een educatieve keuze gemaakt voor de gebruikte variabelen uit de CBS data naar aanleiding van het resultaat met de NLD_dist data. In vervolgonderzoek kunnen al deze variabelen meegenomen worden om zeker te zijn dat geen van deze weggelaten variabelen van toepassing zouden zijn. Echter is deze keuze gemaakt wegens de tijd die de code nodig heeft om te runnen en de kracht van de laptop waarop gewerkt is.

Het zou nog interessant zijn om in vervolgonderzoek te kijken of er een correlatie zit tussen lucht- en lichtvervuiling. Er kunnen verschillende redenen zijn dat de lichtvervuiling niet blijft groeien. Een mogelijkheid is dat gemeenten maatregelen nemen, een andere mogelijkheid is dat er een verband is met het schone lucht akkoord [14] wat is opgesteld en uitgevoerd moet worden. Helaas moest dit onderzoek afgekaderd worden en is er voor gekozen de cijfers van luchtvervuiling niet mee te nemen. Als dit wel wordt gedaan, kan er gekeken worden of er een verband is tussen lucht- en lichtvervuiling. Dit is interessant omdat vermindering in luchtvervuiling dan ook lichtvervuiling zou kunnen verminderen. De vermindering van luchtvervuiling wordt daarmee extra interessant.

Op een gegeven moment waren de outliers in de buurt van steden. Een mogelijke verklaring hiervoor is dat het licht uit de steden ook de dorpen ernaast beïnvloedt waardoor de hoeveelheid lichtvervuiling in het dorp hoger uitvalt dan het model verwacht. Een andere mogelijke verklaring zou kunnen zijn dat licht van snelwegen invloed heeft op deze gebieden. Kijken naar de invloed van infrastructuur was voor dit project niet mogelijk. In vervolgonderzoek zou meer gefocust kunnen worden op de relatie tussen het licht van steden en de nabijgelegen dorpen en de invloed van de infrastructuur door wegen en de verschillende soorten licht die hier zijn te bestuderen, zoals de intensiteit van het verkeer en de soort verlichting die er is.

De vraag of er door het jaar heen verschil is in lichtvervuiling is in dit project vervallen wegens een te grote hoeveelheid onbekende data. Een manier om hiermee aan het werk te gaan is door een waarde voor deze missende periode te bepalen. Echter mist de data van elk jaar rond dezelfde periode en wordt het daarom lastig om zeker te zijn dat de ingevulde data representatief is. Daarom is de keuze gemaakt dit niet te doen voor een bachelor scriptie.

10 Referenties

[1] Ken Aho, DeWayne Derryberry en Teri Peterson. "Model selection for ecologists: the worldviews of AIC and BIC". In: *Ecology* 95.3 (2014), p. 631–636. ISSN: 00129658, 19399170. URL: http://www.jstor.org/stable/43495189 (bezocht op 13-05-2025).

[2] Hirotugu Akaike. "Akaike's Information Criterion". In: *International Encyclopedia of Statistical Science*. Red. door Miodrag Lovrić. Springer, Berlin, Heidelberg, 2011, p. 25. ISBN: 978-3-642-04898-2. DOI: 10.1007/978-3-642-04898-2_110.

[3] AlleCijfers. *Statistieken buurt Bridgamme*. Geraadpleegd op 30 mei 2025. 2025. URL: https://allecijfers.nl/buurt/brigdamme-middelburg/#:~:text=Verken%20alle%20cijfers%20over%20meer%20dan%20250%20onderwerpen,De%20buurt%20Brigdamme%20telt%20575%20inwoners%20in%202024..

[4] AlleCijfers. *Statistieken wijk Bleiswijk Buiten*. Geraadpleegd op 16 mei 2025. 2025. URL: https://allecijfers.nl/wijk/wijk-33-bleiswijk-buiten-lansingerland/.

[5] AlleCijfers. *Statistieken wijk Bleiswijk Buiten*. Geraadpleegd op 16 mei 2025. 2025. URL: https://allecijfers.nl/wijk/aalsmeerderbrug-oude-meer-rozenburg-schiphol-rijk-haarlemmermeer/.

[6] AlleCijfers. *Statistieken wijk Bleiswijk Buiten*. Geraadpleegd op 16 mei 2025. 2025. URL: https://allecijfers.nl/wijk/bedrijventerrein-heerhugowaard-dijk-en-waard/.

[7] AskAnyDifference. *AIC vs BIC*. Geraadpleegd op 30 mei 2025. 2025. URL: https://askanydifference.com/nl/aic-vs-bic/.

[8] F. Bijma, M.A. Jonker en A.W. van der Vaart. *Inleiding in de Statistiek*. Dutch. Epsilon Uitgaven, 2013. ISBN: 9789050411356.

[9] Kenneth P. Burnham en David R. Anderson. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. 2nd. New York: Springer-Verlag, 2002, p. 352–370.

[10] Kenneth P. Burnham en David R. Anderson. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. 2nd. New York: Springer-Verlag, 2002, p. 51–53.

[11] Joseph E. Cavanaugh en Andrew A. Neath. "Akaike's Information Criterion: Background, Derivation, Properties, and Refinements". In: *International Encyclopedia of Statistical Science*. Red. door Miodrag Lovric. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, p. 26–29. ISBN: 978-3-642-04898-2. DOI: 10.1007/978-3-642-04898-2_111. URL: https://doi.org/10.1007/978-3-642-04898-2_111.

[12] CBS. *Kerncijfers wijken en buurten 2004-2024*. Geraadpleegd op 19 mei 2025. 2025. URL: https://www.cbs.nl/nl-nl/reeksen/publicatie/kerncijfers-wijken-en-buurten.

[13] Ron Chepesiuk. *Missing the dark: health effects of light pollution*. 2009.

[14] Convenant. *Schone Lucht Akkoord*. Geraadpleegd op 4 juli 2025. 2020. URL: https://www.rijksoverheid.nl/documenten/convenanten/2020/01/13/bijlage-1-schone-lucht-akkoord.

[15] Alexander Dürre. *Inleiding Mathematische Statistiek: 6.7. Theorem: ML estimator of multiple linear regression model*. Lesaantekeningen, Universiteit Leiden. 2023.

[16] Alexander Dürre. *Inleiding Mathematische Statistiek: 8.8. Motivation: Relationship between information loss and AIC*. Lesaantekeningen, Universiteit Leiden. 2023.

[17] GeeksforGeeks Contributors. *Log likelihood*. Geraadpleegd op 30 mei 2025. 2025. URL: https://www.geeksforgeeks.org/log-likelihood/.

[18] Jouni Kuha. "AIC and BIC: Comparisons of assumptions and performance". In: *Sociological methods & research* 33.2 (2004), p. 188–229.

[19] P. McCullagh en J.A. Nelder. *Generalized Linear Models, Second Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1989, p. 37. ISBN: 9780412317606. URL: https://books.google.nl/books?id=h9kFH2_FfBkC.

[20] Geoffrey J McLachlan, Sharon X Lee en Suren I Rathnayake. "Finite mixture models". In: *Annual review of statistics and its application* 6.1 (2019), p. 355–378.

[21] M. O. Román e.a. "NASA's Black Marble nighttime lights product suite". In: *Remote Sensing of Environment* 210 (2018), p. 113–143. DOI: 10.1016/j.rse.2018.03.017.

[22] Kristin L Sainani. "The value of scatter plots". In: *PM&R* 8.12 (2016), p. 1213–1217.

[23] Martijn Tennekes. "tmap: Thematic Maps in R". In: *Journal of Statistical Software* 84.6 (2018), p. 1–39. DOI: 10.18637/jss.v084.i06.

[24] University of West Georgia. *Scatterplots and Correlation*. Geraadpleegd op 30 mei 2025. Onbekend. URL: https://www.westga.edu/academics/research/vrc/assets/docs/scatterplots_and_correlation_notes.pdf.

[25] Zhuosen Wang. *Black Marble User Guide (Version 1.3)*. NASA. 2022.

[26] Zach Bobbitt. *How to Identify Influential Data Points Using Cook's Distance*. Geraadpleegd op 26 juni 2025. 2020. URL: https://www.statology.org/how-to-identify-influential-data-points-using-cooks-distance/.

[27] Zach Bobbitt. *How to Interpret Diagnostic Plots in R*. Geraadpleegd op 30 mei 2025. 2021. URL: https://www.statology.org/diagnostic-plots-in-r/.

[28] Zach Bobbitt. *The Complete Guide: How to Interpret Q-Q Plots*. Geraadpleegd op 30 mei 2025. 2024. URL: https://www.statology.org/qq-plot-interpretation/.

[29] Zach Bobbitt. *What is Backward Selection? (Definition & Example)*. Geraadpleegd op 17 juni 2025. 2022. URL: https://www.statology.org/backward-selection/.

[30] Zach Bobbitt. *What is Forward Selection? (Definition & Example)*. Geraadpleegd op 17 juni 2025. 2022. URL: https://www.statology.org/forward-selection/.

11 Appendix: R-code

11.1 Packages

```
#Nasa data
library(ggplot2)
library(dplyr)
library(rhdf5)

#Functie plekken in regio's
library(sp)
library(tmap)
library(secr)
library(sf)

#Jaren data
library (readr)

#Opmaak
library(modelsummary)
library(tidyr)
```

11.2 Data laden

11.2.1 NASA

```
h5ls('C:/Users/zoebe/OneDrive/Documenten/Leiden/scriptie/
VNP46A2.A2024361.h18v03.001.2025003112013.h5')
data_2025003 <-
h5read("C:/Users/zoebe/OneDrive/Documenten/Leiden/scriptie/
VNP46A2.A2024361.h18v03.001.2025003112013.h5",
"/HDFEOS/GRIDS")

#Juiste data uit h5 halen en data manipuleren om het te kunnen gebruiken
data_2025003A <- data_2025003[[1]][[1]][[3]]
data_2025003A[data_2025003A==65535] <- NA
data_2025003A[data_2025003A == 0] <- NA
data_2025003B <- log(data_2025003A)

#Data maniupeleren naar waar interesse in is
B_rotated <- data_2025003B[, ncol(data_2025003B):1]
B_cropped <- B_rotated[-(1:800),-((ncol(B_rotated)-1440):ncol(B_rotated))]
B_cropped2 <- B_cropped[-((nrow(B_cropped)-640):nrow(B_cropped)),]
```

11.2.2 Lichtvervuiling data

```
load("C:/Users/zoebe/OneDrive/Documenten/Leiden/scriptie/data/Datalight/DataLight2012.Rdata")
load("C:/Users/zoebe/OneDrive/Documenten/Leiden/scriptie/data/Datalight/DataLight2013.Rdata")
load("C:/Users/zoebe/OneDrive/Documenten/Leiden/scriptie/data/Datalight/DataLight2014.Rdata")
load("C:/Users/zoebe/OneDrive/Documenten/Leiden/scriptie/data/Datalight/DataLight2015.Rdata")
load("C:/Users/zoebe/OneDrive/Documenten/Leiden/scriptie/data/Datalight/DataLight2016.Rdata")
load("C:/Users/zoebe/OneDrive/Documenten/Leiden/scriptie/data/Datalight/DataLight2017.Rdata")
load("C:/Users/zoebe/OneDrive/Documenten/Leiden/scriptie/data/Datalight/DataLight2018.Rdata")
load("C:/Users/zoebe/OneDrive/Documenten/Leiden/scriptie/data/Datalight/DataLight2019.Rdata")
load("C:/Users/zoebe/OneDrive/Documenten/Leiden/scriptie/data/Datalight/DataLight2020.Rdata")
load("C:/Users/zoebe/OneDrive/Documenten/Leiden/scriptie/data/Datalight/DataLight2021.Rdata")
load("C:/Users/zoebe/OneDrive/Documenten/Leiden/scriptie/data/Datalight/DataLight2022.Rdata")
load("C:/Users/zoebe/OneDrive/Documenten/Leiden/scriptie/data/Datalight/DataLight2023.Rdata")
load("C:/Users/zoebe/OneDrive/Documenten/Leiden/scriptie/data/Datalight/DataLight2024.Rdata")
```

11.3 Functies voor coördinaten

```
#' Translate coordinates in rijksdriehoek (Dutch National Grid) system to WGS84
#'
#' @param x vector of x-coordinates in rijksdriehoek system; or without y a
#'   vector of length 2 containing the x and y coordinates.
#' @param y vector of y-coordinates in rijksdriehoek system
#'
#' @returns
#' A list with lamba (decimal degrees; longitude) and phi (decimal degrees;
#' latitude), or if just x was given a numeric vector of length 2 with lambda
#' and phi.
#'
#' @export
rd_to_wgs84 <- function(x, y) {
  if (!is.numeric(x)) stop("x needs to be numeric.")
  if (nargs() == 1  && length(x) == 2) {
    y <- x[2]
    x <- x[1]
  }
  if (!is.numeric(y)) stop("y needs to be numeric.")
  if (length(x) != length(y)) stop("x and y need to have same length.")

  x0 <- 155000.00
  y0 <- 463000.00
  phi0 <- 52.15517440
  lambda0 <- 5.38720621

  k <- list(
    list(p=0, q=1, k=3235.65389),
    list(p=2, q=0, k= -32.58297),
    list(p=0, q=2, k=  -0.24750),
    list(p=2, q=1, k=  -0.84978),
    list(p=0, q=3, k=  -0.06550),
    list(p=2, q=2, k=  -0.01709),
    list(p=1, q=0, k=  -0.00738),
    list(p=4, q=0, k=   0.00530),
    list(p=2, q=3, k=  -0.00039),
    list(p=4, q=1, k=   0.00033),
    list(p=1, q=1, k=  -0.00012))
  l <- list(
    list(p=1, q=0, l=5260.52916),
    list(p=1, q=1, l= 105.94684),
    list(p=1, q=2, l=   2.45656),
    list(p=3, q=0, l=  -0.81885),
    list(p=1, q=3, l=   0.05594),
    list(p=3, q=1, l=  -0.05607),
    list(p=0, q=1, l=   0.01199),
    list(p=3, q=2, l=  -0.00256),
    list(p=1, q=4, l=   0.00128))
  dx <- (x - x0)/1E5
  dy <- (y - y0)/1E5
  phi <- rep(phi0, length(x))
  lambda <- rep(lambda0, length(x))

  for (i in seq_along(k)) {
    p <- k[[i]][["p"]]
    q <- k[[i]][["q"]]
    ki <- k[[i]][["k"]]
    phi <- phi + (ki * (dx^p) * (dy^q))/3600
  }
```

```
  for (i in seq_along(l)) {
    p <- l[[i]][["p"]]
    q <- l[[i]][["q"]]
    li <- l[[i]][["l"]]
    lambda <- lambda + (li * (dx^p) * (dy^q))/3600
  }
  if (nargs() == 1) {
    c(lambda[1], phi[1])
  } else list(lambda=lambda, phi=phi)
}


#' Translate coordinates in WGS84 to rijksdriehoek (Dutch National Grid)
#'
#' @param lambda numeric vector with longitutes (in decimal degrees), or
#'   without phi a vector of length 2 containing both phi and lambda.
#' @param phi numeric vector with latitudes (in decimal degrees).
#'
#' @returns
#' A list with x and y coordinates in Rijksdriehoek system coordinates, of if
#' just lambda was given a numeric vector of length 2 with both coordinates.
#'
#' @export
wgs84_to_rd <- function(lambda, phi) {
  if (!is.numeric(lambda)) stop("lambda needs to be numeric.")
  if (nargs() == 1 && length(lambda) == 2) {
    phi <- lambda[2]
    lambda <- lambda[1]
  }
  if (!is.numeric(phi)) stop("phi needs to be numeric.")
  if (length(lambda) != length(phi))
    stop("lambda and phi need to have same length.")
  x0 <- 155000.0;
  y0 <- 463000.0;
  phi0 <- 52.15517440;
  lambda0 <-  5.38720621;
  r <- list(
    list(p=0, q=1, r= 190094.945),
    list(p=1, q=1, r= -11832.228),
    list(p=2, q=1, r=   -114.221),
    list(p=0, q=3, r=    -32.391),
    list(p=1, q=0, r=     -0.705),
    list(p=3, q=1, r=     -2.340),
    list(p=1, q=3, r=     -0.608),
    list(p=0, q=2, r=     -0.008),
    list(p=2, q=3, r=      0.148))
  s <- list(
    list(p=1, q=0, s= 309056.544),
    list(p=0, q=2, s=   3638.893),
    list(p=2, q=0, s=     73.077),
    list(p=1, q=2, s=   -157.984),
    list(p=3, q=0, s=     59.788),
    list(p=0, q=1, s=      0.433),
    list(p=2, q=2, s=     -6.439),
    list(p=1, q=1, s=     -0.032),
    list(p=0, q=4, s=      0.092),
    list(p=1, q=4, s=     -0.054))
  dphi = 0.36*(phi - phi0)
  dlambda = 0.36*(lambda - lambda0)
  x <- rep(x0, length(phi))
  y <- rep(y0, length(phi))
```

```
  for (i in seq_along(r)) {
    p <- r[[i]][["p"]]
    q <- r[[i]][["q"]]
    ri <- r[[i]][["r"]]
    x <- x +  ri * (dphi^p) * (dlambda^q)
  }
  for (i in seq_along(s)) {
    p <- s[[i]][["p"]]
    q <- s[[i]][["q"]]
    si <- s[[i]][["s"]]
    y <- y +  si * (dphi^p) * (dlambda^q)
  }
  if (nargs() == 1) {
    c(x[1], y[1])
  } else list(x=x, y=y)
}


#' Reproject a geoJSON file to an other coordinate system
#'
#' @param filein the input file or connection.
#' @param fileout the output file or connection.
#' @param projection a function that projects coordinates from the original
#'   coordinate system to the new system. The function has to accept numeric
#'   vectors of length two as its input and return a vector of length two.
#' @param inencoding when filein is a character string this encoding is used
#'   when reading from the file.
#'
#' @returns
#' A list containing the converted geoJSON. This output is usually not needed.
#'
#' @export
reproject_geojson <-
    function(filein, fileout, projection, inencoding=getOption("encoding")) {
  library(rjson)
  trans <- function(coor) {
    if (is.numeric(coor)) {
      return(projection(coor))
    } else {
      res <- vector(length=length(coor), mode="list")
      for (i in seq_along(coor)) {
        res[[i]] <- trans(coor[[i]])
      }
      return(res)
    }
  }
  if (is.character(filein)) filein <- file(filein, "rt", encoding=inencoding)
  map <- fromJSON(file=filein)
  nfeatures <- length(map$features)
  for (i in seq_len(nfeatures)) {
    coor <- map$features[[i]]$geometry$coordinates
    map$features[[i]]$geometry$coordinates <- trans(coor)
  }
  json <- toJSON(map)
  writeLines(json, con=fileout)
  invisible(map)
}
```

11.4 Functie die checkt of punten in een regio zitten

```
data_2025003B <- data_2025003[[1]][[1]][[3]]
data_2025003B[data_2025003B==65535] <- NA
data_2025003B[data_2025003B==0] <- 1/2
data_2025003B <- log(data_2025003B)

data_2025003B_rot <- data_2025003B[, ncol(data_2025003B):1]

data_2025003B_cropped <- data_2025003B_rot[-(1:790),-(1:150)]
data_2025003B_cropped2 <- data_2025003B_cropped
    [-((nrow(data_2025003B_cropped)-650):nrow(data_2025003B_cropped)),
    -((ncol(data_2025003B_cropped)-1500):ncol(data_2025003B_cropped))]
```

11.4.1 Districtenmatrix

```
distrmat <- matrix(ncol=2400,nrow=2400)
averages <- t(sapply(1:3340, function(i) {
  colMeans(as.matrix(NLD_dist$geometry[[i]][[1]][[1]]))
}))

# Maak een matrix van alle combinaties van i en j coördinaten
coords <- expand.grid(i = 1:2400, j = 1:2400)
coords$long <- coords$i / 240
coords$lat <- 60 - coords$j / 240

# Pas de coordinaat transformatie toe op alle punten tegelijk
newcors <- mapply(wgs84_to_rd, lambda = coords$long,
    phi = coords$lat, SIMPLIFY = FALSE)

# Zet de punten om naar een sf-object
# WGS 84 naar het nl stelsel heeft die 4326 nodig
points_sf <- st_as_sf(data.frame(long = coords$long, lat = coords$lat),
    coords = c("long", "lat"), crs = 4326)

# Zet de polygons om naar een sf-object
polygons_sf <- st_as_sf(NLD_dist, crs = 4326)
points_sf <- st_transform(points_sf, st_crs(polygons_sf))

# Gebruik een ruimtelijke intersectie in plaats van een dubbele loop
matches <- st_intersects(points_sf, polygons_sf)

# Opslaan in een matrix
distrmat <- matrix(NA, nrow = 2400, ncol = 2400)
for (idx in seq_along(matches)) {
  if (length(matches[[idx]]) > 0) {
    i <- coords$i[idx]
    j <- coords$j[idx]
    distrmat[i, j] <- matches[[idx]][1]
  }
}

#Bijsnijden
distrmat_rot <- distrmat[, ncol(distrmat):1]
distrmat_cr <- distrmat_rot[-(1:790),-(1:150)]
distrmat_cr2 <- distrmat_cr[-((nrow(distrmat_cr)-650):nrow(distrmat_cr)),
    -((ncol(distrmat_cr)-1500):ncol(distrmat_cr))]
```

11.4.2 Gemiddelde per district

```
Districtenmatrix = distrmat_cr2
licht_matrix <- as.matrix(data_2025003B_cropped2)
```

```
# Data frame definiëren
df_licht <- data.frame(
  regio = as.vector(Districtenmatrix),
  licht = as.vector(licht_matrix)
)

# Hier opniew starten bij apparte dagen
gem_licht_verv <- df_licht %>%
  group_by(regio) %>%
  summarise(gem_licht = ifelse(all(is.nan(df_licht$licht)),
    NA, mean(licht, na.rm = FALSE)))

print(gem_licht_verv)
gem_licht_dist <- rep(NA,3340)
for(i in 1:3340){
  index <- i==gem_licht_verv$regio
  if(sum(index,na.rm=TRUE)>0) {
    gem_licht_dist[i] <- gem_licht_verv$gem_licht[which(index)]
  }
 print(i)
}
NLD_dist$gem_lichtvervuiling <- gem_licht_dist
```

11.5 Model opstellen

```
selectie <- function(data, doel, pred) {
  resultaat <- data.frame(Model = character(),
  AIC = numeric(), stringsAsFactors = FALSE)

  if(length(pred)> 1){
  # Loop mogelijkheden in model
  for (k in 1:length(pred)) {
    combos <- combn(pred, k, simplify = FALSE)

    for (combo in combos) {
      subdata <- data[, c(doel, combo)]
        # Verwijder rijen met NA's
      subdata <- na.omit(subdata)
      valid <- valid <- all(sapply(combo, function(var)
        length(unique(subdata[[var]])) > 1))

      if(valid){
      formula <- as.formula(paste(doel, "~",
        paste(combo, collapse = "+")))
      model <- glm(formula, data = data)
      # Opslaan van resultaten
      resultaat <- rbind(resultaat, data.frame(
        Model = paste(combo, collapse = "+"),
        AIC = AIC(model),
        stringsAsFactors = FALSE
      ))
      print(AIC(model))
      }
    }
  }

  # Sorteer
  resultaat <- resultaat[order(resultaat$AIC), ]
  return(resultaat)
```

```
}
}

# Invullen
data <- NLD_dist
doel <- "gem_lichtvervuiling"
pred <- c("area", "province", "urbanity", "population", "dwelling_total",
    "employment_rate", "edu_appl_sci", "pop_0_14", "pop_15_24", "pop_25_44",
    "pop_45_64", "pop_65plus", "dwelling_value", "dwelling_ownership",
    "income_low", "income_high")

# Run code
resultaat <- selectie(data, doel, pred)
print(resultaat)

# Invullen
data <- Year2024
doel <- "gem_lichtvervuiling"
pred <- c("gm_naam", "a_1p_hh", "a_hh_z_k", "a_hh_m_k", "g_hhgro", "a_woning",
    "a_inkont", "a_pau", "a_bst_b", "a_bst_nb", "g_pau_hh", "g_pau_km",
    "a_m2w", "a_opp_ha", "a_lan_ha", "a_wat_ha", "ste_oad")

# Run code
resultaat <- selectie(data, doel, pred)
print(resultaat)

toevoeging<- c("a_inw","a_00_14","a_15_24","a_25_44","a_45_64","a_65_oo","a_hh",
    "bev_dich","g_wozbag", "a_bedv","a_bed_a","a_bed_bf","a_bed_gi","a_bed_hj",
    "a_bed_kl","a_bed_mn","a_bed_oq","a_bed_ru")
```

11.5.1 Model aanpassen

```
model1 <- glm(gem_licht_dist~log(area) + province +
    urbanity + population + dwelling_total +
    employment_rate + edu_appl_sci + pop_15_24 +
    pop_25_44 + pop_65plus + dwelling_value + dwelling_ownership + income_high,
    data = NLD_dist)

AIC(model1)

plot.ecdf(residuals(model1))
plot(density(residuals(model1)))
qqnorm(residuals(model1))
plot(NLD_dist$urbanity,NLD_dist$gem_lichtvervuiling)
plot(NLD_dist$population,NLD_dist$gem_lichtvervuiling)

model2 <- glm(gem_licht_dist ~ log(area) + province +
    urbanity + log(population) + dwelling_total +
    employment_rate + edu_appl_sci + pop_15_24 + pop_25_44 +
    pop_65plus + dwelling_value + dwelling_ownership + income_high,
    data = NLD_dist)
AIC(model2)

plot(NLD_dist$dwelling_total,NLD_dist$gem_lichtvervuiling)
model3 <- glm(gem_licht_dist~log(area) + province +
    urbanity + log(population) + log(dwelling_total) +
    employment_rate + edu_appl_sci + pop_15_24 + pop_25_44 +
    pop_65plus + dwelling_value + dwelling_ownership + income_high       ,
    data = NLD_dist)
AIC(model3) #minimaal verschil, maar wel beter
```

```
plot(NLD_dist$employment_rate,NLD_dist$gem_lichtvervuiling)
NLD_dist$emp2<-NLD_dist$employment_rate * NLD_dist$employment_rate
model4 <- glm(gem_licht_dist~log(area) + province +
    urbanity + log(population) + log(dwelling_total) + emp2 +
    edu_appl_sci + pop_15_24 + pop_25_44 + pop_65plus +
    dwelling_value + dwelling_ownership + income_high,
    data = NLD_dist)
AIC(model4) # minimaal verschil

plot(NLD_dist$pop_15_24,NLD_dist$gem_lichtvervuiling)
model5 <- glm(gem_licht_dist~log(area) + province +
    urbanity + log(population) + log(dwelling_total) + emp2 +
    edu_appl_sci + log(pop_15_24) + pop_25_44 + pop_65plus +
    dwelling_value + dwelling_ownership + income_high,
    data = NLD_dist)
AIC(model5)

plot(NLD_dist$pop_25_44,NLD_dist$gem_lichtvervuiling)
model5 <- glm(gem_licht_dist~log(area) + province +
    urbanity + log(population) + log(dwelling_total) + emp2 +
    edu_appl_sci + log(pop_15_24) + log(pop_25_44) +
    pop_65plus + dwelling_value + dwelling_ownership + income_high,
    data = NLD_dist)
AIC(model5)

plot(NLD_dist$income_high,NLD_dist$gem_lichtvervuiling)
model6 <- glm(gem_licht_dist~log(area) + province +
    urbanity + log(population) + log(dwelling_total) + emp2 +
    edu_appl_sci + log(pop_15_24) + log(pop_25_44) + pop_65plus +
    dwelling_value + dwelling_ownership + income_high,
    data = NLD_dist)
AIC(model6)

model7 <- glm(gem_licht_dist~log(area) + province +
    urbanity + log(population) + log(dwelling_total) + emp2 +
    edu_appl_sci + log(pop_15_24) + log(pop_25_44) + pop_65plus +
    dwelling_value + dwelling_ownership + income_high + population/area ,
    data = NLD_dist)
AIC(model7)

NLD_dist$valown <- NLD_dist$dwelling_value * NLD_dist$dwelling_ownership

model8 <- glm(gem_licht_dist~log(area) + province +
    urbanity + log(population) + log(dwelling_total) + emp2 +
    edu_appl_sci + log(pop_15_24) + log(pop_25_44) + pop_65plus +
    dwelling_value + dwelling_ownership + income_high + population/area + valown,
    data = NLD_dist)
AIC(model8)
```

11.5.2 Outliers

```
print(NLD_dist$name[2351])
print(NLD_dist$name[2751])
print(NLD_dist$name[1072])
print(NLD_dist$name[100])
print(NLD_dist$name[2052])
print(NLD_dist$name[3293])
```

11.5.3 Model bij CBS data

```
model1a <- glm(gem_lichtvervuiling ~ gm_naam + a_hh_z_k +
    a_hh_m_k + g_hhgro + a_pau + g_pau_hh + g_pau_km + a_opp_ha + ste_oad,
    data=Year2024)
AIC(model1a)
plot(model1a)
model1b <- glm(gem_lichtvervuiling ~ a_inw + gm_naam +
    a_hh_z_k + a_hh_m_k + g_hhgro + a_pau +
    g_pau_hh + g_pau_km + a_opp_ha + ste_oad + a_bed_a +
    a_bed_bf + a_bed_gi + a_bed_hj + a_bed_kl + a_bed_mn +
    a_bed_oq + a_bed_ru + g_wozbag + bev_dich + a_hh +
    a_00_14 + a_15_24 + a_25_44 + a_65_oo,
    data=Year2024)
AIC(model1b)
plot(model1b)
```

11.5.4 Opname variabelen

```
plot(Year2024$a_65_oo,Year2024$gem_lichtvervuiling)

Year2024$bev_dich2 <- Year2024$bev_dich*Year2024$bev_dich

model2b <- glm(gem_lichtvervuiling ~ log(a_inw) + gm_naam +
    log(a_hh_z_k) + a_hh_m_k + g_hhgro + log(a_pau) +
    g_pau_hh + log(g_pau_km) + a_opp_ha +  log(ste_oad) +
    a_bed_a + a_bed_bf + a_bed_gi + a_bed_hj + a_bed_kl +
    a_bed_mn + a_bed_oq + a_bed_ru + g_wozbag + bev_dich2 +
    log(a_hh) + log(a_00_14) + a_15_24 + log(a_25_44) + a_65_oo,
    data=Year2024)
AIC(model2d)
plot(model2d)
```

11.5.5 Urbanity model

```
model_urb <- glm(gem_licht_dist ~ urbanity, data=NLD_dist)

stargazer(model_urb,
          type = "latex",
          title = "Samenvatting van Linear Regression Model voor urbanity",
          label = "tab:model_summary",
          out = "model_summary.tex")
```

11.6 Meerdere jaren

```
Stats <- function(mat){
  min_waarde <- apply(mat, 1, function(x) min(x, na.rm = TRUE))
  max_waarde <- apply(mat, 1, function(x) max(x, na.rm = TRUE))
  mean_waarde <- apply(mat, 1, function(x) mean(x, na.rm = TRUE))
   Verschil_min_max <- max_waarde - min_waarde

   positie_min <- apply(mat, 1, function(x) {
     which(x == min(x))
   })

   positie_max <- apply(mat, 1, function(x) {
     which(x == max(x))
   })
```

```
  result <- cbind(min = min_waarde, max = max_waarde, mean = mean_waarde,
    Verschil_min_max, positie_min, positie_max)

  return(result)
  print(result)
}
```

11.6.1 Data frame maken over gemiddelden van jaren

```
Y2012light_df <- as.data.frame(Y2012light)
Y2012light_df$gem <- rowMeans(Y2012light_df, na.rm=TRUE)

Y2013light_df <- as.data.frame(Y2013light)
Y2013light_df$gem <- rowMeans(Y2013light_df, na.rm=TRUE)

Y2014light_df <- as.data.frame(Y2014light)
Y2014light_df$gem <- rowMeans(Y2014light_df, na.rm=TRUE)

Y2015light_df <- as.data.frame(Y2015light)
Y2015light_df$gem <- rowMeans(Y2015light_df, na.rm=TRUE)

Y2016light_df <- as.data.frame(Y2016light)
Y2016light_df$gem <- rowMeans(Y2016light_df, na.rm=TRUE)

Y2017light_df <- as.data.frame(Y2017light)
Y2017light_df$gem <- rowMeans(Y2017light_df, na.rm=TRUE)

Y2018light_df <- as.data.frame(Y2018light)
Y2018light_df$gem <- rowMeans(Y2018light_df, na.rm=TRUE)

Y2019light_df <- as.data.frame(Y2019light)
Y2019light_df$gem <- rowMeans(Y2019light_df, na.rm=TRUE)

Y2020light_df <- as.data.frame(Y2020light)
Y2020light_df$gem <- rowMeans(Y2020light_df, na.rm=TRUE)

Y2021light_df <- as.data.frame(Y2021light)
Y2021light_df$gem <- rowMeans(Y2021light_df, na.rm=TRUE)

Y2022light_df <- as.data.frame(Y2022light)
Y2022light_df$gem <- rowMeans(Y2022light_df, na.rm=TRUE)

Y2023light_df <- as.data.frame(Y2023light)
Y2023light_df$gem <- rowMeans(Y2023light_df, na.rm=TRUE)

Y2024light_df <- as.data.frame(Y2024light)
Y2024light_df$gem <- rowMeans(Y2024light_df, na.rm=TRUE)

df_gem_jaren <- data.frame(gem2012=Y2012light_df$gem, gem2013=Y2013light_df$gem,
    gem2014=Y2014light_df$gem, gem2015=Y2015light_df$gem,
    gem2016=Y2016light_df$gem, gem2017=Y2017light_df$gem, gem2018=Y2018light_df$gem,
    gem2019=Y2019light_df$gem, gem2020=Y2020light_df$gem, gem2021=Y2021light_df$gem,
    gem2022=Y2022light_df$gem, gem2023=Y2023light_df$gem, gem2024=Y2024light_df$gem)

df_gem_jaren$locatie_id <- 1:nrow(df_gem_jaren)
df_totaal <- df_gem_jaren %>%
  pivot_longer(
    cols = starts_with("gem"),
    names_to = "jaar",
    names_prefix = "gem",
```

```
    values_to = "gem"
  )
df_totaal$jaar_factor <- factor(df_totaal$jaar)
model_jaren <- lm(gem ~ jaar_factor, data = df_totaal)
```

11.7 Meerdere dagen

```
Stats_metplot <- function(mat){
  min_waarde <- apply(mat, 1, function(x) min(x, na.rm = TRUE))
  max_waarde <- apply(mat, 1, function(x) max(x, na.rm = TRUE))
  mean_waarde <- apply(mat, 1, function(x) mean(x, na.rm = TRUE))
  Verschil_min_max <- max_waarde - min_waarde

  positie_min <- apply(mat, 1, function(x) {
    geldige_min_x <- ifelse(is.nan(x), Inf, x)
    which(geldige_min_x == min(geldige_min_x))
  })

  positie_max <- apply(mat, 1, function(x) {
    geldige_max_x <- ifelse(is.nan(x), -Inf, x)
    which(geldige_max_x == max(geldige_max_x))
  })

  positie_min_str <- sapply(positie_min, function(x) paste(x, collapse = ","))
  positie_max_str <- sapply(positie_max, function(x) paste(x, collapse = ","))

  result <- cbind(min = min_waarde, max = max_waarde, mean = mean_waarde,
    Verschil_min_max, positie_min = positie_min_str, positie_max=positie_max_str)

  #Minimale posities
  positie_min_list <- strsplit(result[,5], ",")
  alle_min_posities <- as.integer(unlist(positie_min_list))
  hist(alle_min_posities,
     breaks = 60,
     main = "Aantal keer laagste waarde per kolom",
     xlab = "Dagen van het jaar",
     col = "lightgreen")

  #Maximale posities
  positie_max_list <- strsplit(result[,6], ",")
  alle_max_posities <- as.integer(unlist(positie_max_list))
  hist(alle_max_posities,
     breaks = 60,
     main = "Aantal keer hoogtse waarde per kolom",
     xlab = "Dagen van het jaar",
     col = "lightblue")

  return(result)
  print(result)
}
```

12 Appendix: Tabellen

12.1 Invloed van variabelen in het model op lichtvervuiling

Tabel 8: Samenvatting van Linear Regression Model

	Dependent variable:
	gem_lichtvervuiling
log(a_inw)	0.181
	(0.252)
gm_naamAalsmeer	1.461***
	(0.308)
gm_naamAalten	0.512*
	(0.270)
gm_naamAchtkarspelen	0.253
	(0.304)
gm_naamAlblasserdam	1.169***
	(0.251)
gm_naamAlbrandswaard	1.351***
	(0.233)
gm_naamAlkmaar	1.363***
	(0.191)
gm_naamAlmelo	1.329***
	(0.184)
gm_naamAlmere	0.638***
	(0.145)
gm_naamAlphen-Chaam	0.732***
	(0.246)
gm_naamAlphen aan den Rijn	0.861***
	(0.175)
gm_naamAltena	0.681***
	(0.164)
gm_naamAmeland	0.516
	(0.501)
gm_naamAmersfoort	1.558***
	(0.157)
gm_naamAmstelveen	0.494**
	(0.197)
gm_naamApeldoorn	1.016***
	(0.171)
gm_naamArnhem	1.238***
	(0.164)

Vervolg op volgende pagina

Tabel 8 – vervolg

	Dependent variable:
	gem_lichtvervuiling
gm_naamAssen	0.857***
	(0.194)
gm_naamAsten	1.278***
	(0.203)
gm_naamBaarle-Nassau	0.538*
	(0.303)
gm_naamBaarn	0.913***
	(0.274)
gm_naamBarendrecht	1.421***
	(0.167)
gm_naamBarneveld	0.728***
	(0.195)
gm_naamBeek	2.134***
	(0.362)
gm_naamBeekdaelen	1.203***
	(0.217)
gm_naamBeesel	1.291***
	(0.362)
gm_naamBerg en Dal	0.488***
	(0.187)
gm_naamBergeijk	0.765***
	(0.221)
gm_naamBergen (L.)	0.513**
	(0.216)
gm_naamBergen (NH.)	0.048
	(0.222)
gm_naamBergen op Zoom	0.920***
	(0.247)
gm_naamBerkelland	0.221
	(0.252)
gm_naamBernheze	0.937***
	(0.232)
gm_naamBest	0.601
	(0.503)
gm_naamBeuningen	0.978***
	(0.269)
gm_naamBeverwijk	1.560***

Vervolg op volgende pagina

Tabel 8 – vervolg

	Dependent variable:
	gem_lichtvervuiling
	(0.198)
gm_naamBladel	0.012
	(0.247)
gm_naamBlaricum	0.832
	(0.510)
gm_naamBloemendaal	0.530**
	(0.259)
gm_naamBodegraven-Reeuwijk	0.795***
	(0.219)
gm_naamBoekel	−0.101
	(0.364)
gm_naamBorger-Odoorn	0.058
	(0.159)
gm_naamBorne	0.939***
	(0.363)
gm_naamBorsele	0.608***
	(0.170)
gm_naamBoxtel	0.210
	(0.247)
gm_naamBreda	1.340***
	(0.193)
gm_naamBronckhorst	0.595**
	(0.251)
gm_naamBrummen	0.382*
	(0.229)
gm_naamBrunssum	1.611***
	(0.248)
gm_naamBunnik	0.963***
	(0.306)
gm_naamBunschoten	0.182
	(0.511)
gm_naamBuren	0.527**
	(0.218)
gm_naamCapelle aan den IJssel	1.190***
	(0.204)
gm_naamCastricum	0.877***
	(0.210)

Vervolg op volgende pagina

Tabel 8 – vervolg

	Dependent variable:
	gem_lichtvervuiling
gm_naamCoevorden	0.543**
	(0.218)
gm_naamCranendonck	0.559**
	(0.229)
gm_naamCulemborg	0.172
	(0.364)
gm_naamDalfsen	0.372
	(0.309)
gm_naamDantumadiel	0.058
	(0.269)
gm_naamDe Bilt	0.214
	(0.199)
gm_naamDe Fryske Marren	0.169
	(0.140)
gm_naamDe Ronde Venen	0.705***
	(0.210)
gm_naamDe Wolden	0.154
	(0.177)
gm_naamDelft	1.545***
	(0.198)
gm_naamDen Helder	0.674***
	(0.208)
gm_naamDeurne	−0.797***
	(0.272)
gm_naamDeventer	0.686***
	(0.173)
gm_naamDiemen	0.812***
	(0.239)
gm_naamDijk en Waard	1.531***
	(0.158)
gm_naamDinkelland	0.017
	(0.184)
gm_naamDoesburg	−0.424
	(0.499)
gm_naamDoetinchem	0.870***
	(0.189)
gm_naamDongen	0.960***
	(0.363)

Vervolg op volgende pagina

Tabel 8 – vervolg

	Dependent variable:
	gem_lichtvervuiling
gm_naamDordrecht	0.881***
	(0.192)
gm_naamDrechterland	0.462**
	(0.218)
gm_naamDrimmelen	0.961***
	(0.247)
gm_naamDronten	1.016***
	(0.223)
gm_naamDruten	0.117
	(0.269)
gm_naamDuiven	1.175***
	(0.247)
gm_naamEcht-Susteren	0.874***
	(0.187)
gm_naamEdam-Volendam	0.505**
	(0.201)
gm_naamEde	1.229***
	(0.173)
gm_naamEemnes	0.561
	(0.500)
gm_naamEemsdelta	0.606***
	(0.168)
gm_naamEersel	0.430*
	(0.231)
gm_naamEijsden-Margraten	0.975***
	(0.230)
gm_naamEindhoven	2.026***
	(0.169)
gm_naamElburg	0.260
	(0.304)
gm_naamEmmen	0.496***
	(0.150)
gm_naamEnkhuizen	1.644***
	(0.271)
gm_naamEnschede	0.880***
	(0.205)
gm_naamEpe	−0.085

Vervolg op volgende pagina

Tabel 8 – vervolg

	Dependent variable:
	gem_lichtvervuiling
	(0.271)
gm_naamErmelo	0.687***
	(0.248)
gm_naamEtten-Leur	1.823***
	(0.231)
gm_naamGeertruidenberg	1.176***
	(0.304)
gm_naamGeldrop-Mierlo	1.576***
	(0.365)
gm_naamGemert-Bakel	−0.217
	(0.218)
gm_naamGennep	0.738***
	(0.246)
gm_naamGilze en Rijen	1.196***
	(0.304)
gm_naamGoeree-Overflakkee	0.250
	(0.192)
gm_naamGoes	0.191
	(0.199)
gm_naamGoirle	1.399***
	(0.183)
gm_naamGooise Meren	0.607**
	(0.242)
gm_naamGorinchem	1.602***
	(0.197)
gm_naamGouda	0.789***
	(0.204)
gm_naamGroningen	0.782***
	(0.167)
gm_naamGulpen-Wittem	0.216
	(0.216)
gm_naamHaaksbergen	0.693***
	(0.217)
gm_naamHaarlem	1.066***
	(0.171)
gm_naamHaarlemmermeer	1.685***
	(0.173)

Vervolg op volgende pagina

Tabel 8 – vervolg

	Dependent variable:
	gem_lichtvervuiling
gm_naamHalderberge	1.160***
	(0.246)
gm_naamHardenberg	0.370**
	(0.151)
gm_naamHarderwijk	0.434**
	(0.179)
gm_naamHardinxveld-Giessendam	0.479*
	(0.272)
gm_naamHarlingen	1.014***
	(0.361)
gm_naamHattem	0.554
	(0.363)
gm_naamHeemskerk	1.447***
	(0.198)
gm_naamHeemstede	1.233***
	(0.313)
gm_naamHeerde	0.342
	(0.363)
gm_naamHeerenveen	0.550**
	(0.217)
gm_naamHeerlen	1.397***
	(0.162)
gm_naamHeeze-Leende	0.423
	(0.270)
gm_naamHeiloo	0.839***
	(0.222)
gm_naamHellendoorn	0.390*
	(0.203)
gm_naamHelmond	1.412***
	(0.190)
gm_naamHendrik-Ido-Ambacht	1.697***
	(0.253)
gm_naamHengelo	0.841***
	(0.195)
gm_naamHet Hogeland	0.028
	(0.183)
gm_naamHeumen	0.744**
	(0.362)

Vervolg op volgende pagina

Tabel 8 – vervolg

	Dependent variable:
	gem_lichtvervuiling
gm_naamHeusden	0.574***
	(0.194)
gm_naamHillegom	1.604***
	(0.250)
gm_naamHilvarenbeek	0.291
	(0.247)
gm_naamHilversum	1.230***
	(0.212)
gm_naamHoeksche Waard	0.643***
	(0.175)
gm_naamHof van Twente	0.822***
	(0.193)
gm_naamHollands Kroon	0.094
	(0.165)
gm_naamHoogeveen	0.300*
	(0.182)
gm_naamHoorn	1.885***
	(0.186)
gm_naamHorst aan de Maas	0.664***
	(0.168)
gm_naamHouten	1.287***
	(0.204)
gm_naamHuizen	0.554***
	(0.188)
gm_naamHulst	0.626***
	(0.170)
gm_naamIJsselstein	0.381
	(0.505)
gm_naamKaag en Braassem	1.079***
	(0.248)
gm_naamKampen	0.801***
	(0.232)
gm_naamKapelle	0.735**
	(0.303)
gm_naamKatwijk	0.847***
	(0.252)
gm_naamKerkrade	1.939***

Vervolg op volgende pagina

Tabel 8 – vervolg

	Dependent variable:
	gem_lichtvervuiling
	(0.306)
gm_naamKoggenland	0.546***
	(0.193)
gm_naamKrimpen aan den IJssel	1.364***
	(0.508)
gm_naamKrimpenerwaard	0.557***
	(0.184)
gm_naamLaarbeek	0.576**
	(0.270)
gm_naamLand van Cuijk	0.510***
	(0.145)
gm_naamLandgraaf	1.301***
	(0.305)
gm_naamLandsmeer	0.239
	(0.307)
gm_naamLansingerland	1.879***
	(0.179)
gm_naamLaren	1.232**
	(0.520)
gm_naamLeeuwarden	0.878***
	(0.158)
gm_naamLeiden	0.639***
	(0.206)
gm_naamLeiderdorp	0.857***
	(0.307)
gm_naamLeidschendam-Voorburg	1.349***
	(0.184)
gm_naamLelystad	1.047***
	(0.203)
gm_naamLeudal	0.476***
	(0.167)
gm_naamLeusden	0.906***
	(0.232)
gm_naamLingewaard	1.011***
	(0.230)
gm_naamLisse	1.307***
	(0.271)

Vervolg op volgende pagina

Tabel 8 – vervolg

	Dependent variable:
	gem_lichtvervuiling
gm_naamLochem	0.699***
	(0.231)
gm_naamLoon op Zand	1.161***
	(0.269)
gm_naamLopik	1.003**
	(0.507)
gm_naamLosser	0.326
	(0.246)
gm_naamMaasdriel	0.647***
	(0.247)
gm_naamMaasgouw	0.821***
	(0.208)
gm_naamMaashorst	0.761***
	(0.231)
gm_naamMaassluis	1.676***
	(0.212)
gm_naamMaastricht	1.757***
	(0.226)
gm_naamMedemblik	0.666***
	(0.168)
gm_naamMeerssen	1.428***
	(0.303)
gm_naamMeierijstad	0.521*
	(0.279)
gm_naamMeppel	0.721***
	(0.191)
gm_naamMiddelburg	0.742***
	(0.171)
gm_naamMidden-Delfland	3.070***
	(0.365)
gm_naamMidden-Drenthe	0.114
	(0.167)
gm_naamMidden-Groningen	0.610***
	(0.156)
gm_naamMoerdijk	1.251***
	(0.192)
gm_naamMolenlanden	0.548***
	(0.162)

Vervolg op volgende pagina

Tabel 8 – vervolg

	Dependent variable:
	gem_lichtvervuiling
gm_naamMontferland	0.747**
	(0.305)
gm_naamMontfoort	0.931***
	(0.304)
gm_naamMook en Middelaar	0.643*
	(0.362)
gm_naamNeder-Betuwe	1.041***
	(0.306)
gm_naamNederweert	0.308
	(0.246)
gm_naamNieuwegein	1.829***
	(0.172)
gm_naamNieuwkoop	1.055***
	(0.364)
gm_naamNijkerk	1.215***
	(0.249)
gm_naamNijmegen	1.203***
	(0.220)
gm_naamNissewaard	1.102***
	(0.163)
gm_naamNoardeast-Fryslan	0.247
	(0.174)
gm_naamNoord-Beveland	0.056
	(0.229)
gm_naamNoordenveld	0.561**
	(0.247)
gm_naamNoordoostpolder	0.925***
	(0.190)
gm_naamNoordwijk	0.613**
	(0.251)
gm_naamNuenen, Gerwen en Nederwetten	1.073***
	(0.304)
gm_naamNunspeet	0.411
	(0.369)
gm_naamOegstgeest	0.863***
	(0.255)
gm_naamOirschot	0.331

Vervolg op volgende pagina

Tabel 8 – vervolg

	Dependent variable:
	gem_lichtvervuiling
	(0.270)
gm_naamOisterwijk	1.177***
	(0.175)
gm_naamOldambt	0.310
	(0.192)
gm_naamOldebroek	0.404
	(0.305)
gm_naamOldenzaal	0.734***
	(0.191)
gm_naamOlst-Wijhe	0.221
	(0.187)
gm_naamOmmen	0.460**
	(0.187)
gm_naamOost Gelre	0.364*
	(0.217)
gm_naamOosterhout	1.811***
	(0.184)
gm_naamOoststellingwerf	−0.151
	(0.176)
gm_naamOostzaan	0.484
	(0.500)
gm_naamOpmeer	−0.082
	(0.362)
gm_naamOpsterland	0.144
	(0.177)
gm_naamOss	0.960***
	(0.159)
gm_naamOude IJsselstreek	0.548**
	(0.271)
gm_naamOuder-Amstel	0.828
	(0.510)
gm_naamOudewater	0.321
	(0.270)
gm_naamOverbetuwe	1.103***
	(0.167)
gm_naamPapendrecht	1.012***
	(0.221)

Vervolg op volgende pagina

Tabel 8 – vervolg

	Dependent variable:
	gem_lichtvervuiling
gm_naamPeel en Maas	0.711***
	(0.247)
gm_naamPekela	0.345
	(0.362)
gm_naamPijnacker-Nootdorp	2.211***
	(0.311)
gm_naamPurmerend	1.464***
	(0.212)
gm_naamPutten	0.396
	(0.367)
gm_naamRaalte	0.085
	(0.199)
gm_naamReimerswaal	0.453*
	(0.232)
gm_naamRenkum	0.504***
	(0.190)
gm_naamRenswoude	0.228
	(0.500)
gm_naamReusel-De Mierden	−0.051
	(0.270)
gm_naamRheden	−0.071
	(0.307)
gm_naamRhenen	0.369*
	(0.204)
gm_naamRidderkerk	1.723***
	(0.202)
gm_naamRijssen-Holten	0.906***
	(0.270)
gm_naamRijswijk	0.928***
	(0.199)
gm_naamRoerdalen	0.605***
	(0.229)
gm_naamRoermond	1.139***
	(0.193)
gm_naamRoosendaal	1.309***
	(0.179)
gm_naamRotterdam	0.987***
	(0.217)

Vervolg op volgende pagina

Tabel 8 – vervolg

	Dependent variable:
	gem_lichtvervuiling
gm_naamRozendaal	−0.091
	(0.502)
gm_naamRucphen	1.501***
	(0.247)
gm_naams-Gravenhage	0.794***
	(0.160)
gm_naams-Hertogenbosch	0.858***
	(0.180)
gm_naamSchagen	0.305*
	(0.166)
gm_naamScherpenzeel	0.091
	(0.500)
gm_naamSchiedam	1.337***
	(0.214)
gm_naamSchouwen-Duiveland	0.325*
	(0.177)
gm_naamSimpelveld	1.140***
	(0.362)
gm_naamSint-Michielsgestel	0.531**
	(0.248)
gm_naamSittard-Geleen	1.509***
	(0.209)
gm_naamSliedrecht	0.909***
	(0.306)
gm_naamSluis	−0.072
	(0.171)
gm_naamSmallingerland	−2.051***
	(0.501)
gm_naamSoest	0.934***
	(0.211)
gm_naamSomeren	−0.114
	(0.272)
gm_naamSon en Breugel	0.947***
	(0.363)
gm_naamStadskanaal	0.010
	(0.246)
gm_naamStaphorst	0.530**

Vervolg op volgende pagina

Tabel 8 – vervolg

	Dependent variable:
	gem_lichtvervuiling
	(0.220)
gm_naamStede Broec	1.145***
	(0.305)
gm_naamSteenbergen	0.712***
	(0.246)
gm_naamSteenwijkerland	0.338**
	(0.147)
gm_naamStein	1.322***
	(0.304)
gm_naamStichtse Vecht	0.712***
	(0.185)
gm_naamSudwest-Fryslan	0.290*
	(0.161)
gm_naamTerneuzen	0.743***
	(0.155)
gm_naamTerschelling	−0.007
	(0.270)
gm_naamTexel	0.196
	(0.500)
gm_naamTeylingen	1.318***
	(0.204)
gm_naamTholen	0.175
	(0.217)
gm_naamTiel	0.525*
	(0.271)
gm_naamTilburg	1.557***
	(0.144)
gm_naamTubbergen	0.115
	(0.209)
gm_naamTwenterand	0.418*
	(0.231)
gm_naamTynaarlo	0.124
	(0.170)
gm_naamTytsjerksteradiel	0.358*
	(0.192)
gm_naamUitgeest	0.895*
	(0.500)

Vervolg op volgende pagina

Tabel 8 – vervolg

	Dependent variable:
	gem_lichtvervuiling
gm_naamUithoorn	1.277***
	(0.207)
gm_naamUrk	0.360
	(0.510)
gm_naamUtrecht	1.311***
	(0.259)
gm_naamUtrechtse Heuvelrug	0.549**
	(0.251)
gm_naamVaals	0.933***
	(0.362)
gm_naamValkenburg aan de Geul	1.076***
	(0.229)
gm_naamValkenswaard	0.575*
	(0.305)
gm_naamVeendam	0.768**
	(0.303)
gm_naamVeenendaal	1.657***
	(0.234)
gm_naamVeere	0.147
	(0.182)
gm_naamVeldhoven	1.334***
	(0.307)
gm_naamVelsen	0.904***
	(0.196)
gm_naamVenlo	1.368***
	(0.161)
gm_naamVenray	1.019***
	(0.155)
gm_naamVijfheerenlanden	0.627***
	(0.147)
gm_naamVlaardingen	1.426***
	(0.221)
gm_naamVlieland	−0.477
	(0.503)
gm_naamVlissingen	0.856***
	(0.248)
gm_naamVoerendaal	1.338***
	(0.230)

Vervolg op volgende pagina

Tabel 8 – vervolg

	Dependent variable:
	gem_lichtvervuiling
gm_naamVoorschoten	0.670
	(0.503)
gm_naamVoorst	0.948***
	(0.270)
gm_naamVught	0.470*
	(0.272)
gm_naamWaadhoeke	0.209
	(0.144)
gm_naamWaalre	1.568***
	(0.503)
gm_naamWaalwijk	0.867***
	(0.308)
gm_naamWaddinxveen	1.954***
	(0.272)
gm_naamWageningen	1.177***
	(0.195)
gm_naamWassenaar	0.499
	(0.373)
gm_naamWaterland	0.140
	(0.232)
gm_naamWeert	1.185***
	(0.170)
gm_naamWest Betuwe	0.834***
	(0.183)
gm_naamWest Maas en Waal	0.498**
	(0.207)
gm_naamWesterkwartier	−0.034
	(0.162)
gm_naamWesterveld	0.044
	(0.164)
gm_naamWestervoort	1.259**
	(0.499)
gm_naamWesterwolde	0.058
	(0.192)
gm_naamWestland	3.132***
	(0.207)
gm_naamWeststellingwerf	0.334**

Vervolg op volgende pagina

Tabel 8 – vervolg

	Dependent variable:
	gem_lichtvervuiling
	(0.167)
gm_naamWierden	0.961***
	(0.232)
gm_naamWijchen	0.750***
	(0.199)
gm_naamWijdemeren	−0.019
	(0.224)
gm_naamWijk bij Duurstede	0.579**
	(0.270)
gm_naamWinterswijk	1.291***
	(0.370)
gm_naamWoensdrecht	0.935***
	(0.269)
gm_naamWoerden	0.977***
	(0.209)
gm_naamWormerland	0.539*
	(0.304)
gm_naamWoudenberg	0.280
	(0.500)
gm_naamZaanstad	1.023***
	(0.168)
gm_naamZaltbommel	1.444***
	(0.270)
gm_naamZandvoort	0.512*
	(0.309)
gm_naamZeewolde	1.821***
	(0.590)
gm_naamZeist	0.929***
	(0.255)
gm_naamZevenaar	0.654***
	(0.186)
gm_naamZoetermeer	0.714***
	(0.214)
gm_naamZoeterwoude	0.738***
	(0.272)
gm_naamZuidplas	1.891***
	(0.272)

Vervolg op volgende pagina

Tabel 8 – vervolg

	Dependent variable:
	gem_lichtvervuiling
gm_naamZundert	0.637**
	(0.248)
gm_naamZutphen	0.468**
	(0.233)
gm_naamZwartewaterland	0.764***
	(0.231)
gm_naamZwijndrecht	1.136***
	(0.211)
gm_naamZwolle	0.944***
	(0.176)
log(a_hh_z_k)	−0.472***
	(0.078)
a_hh_m_k	−0.00001
	(0.00004)
g_hhgro	−0.199*
	(0.119)
log(a_pau)	−0.345*
	(0.179)
g_pau_hh	0.347**
	(0.150)
log(g_pau_km)	0.316***
	(0.021)
a_opp_ha	−0.0001***
	(0.00002)
log(ste_oad)	0.293***
	(0.021)
a_bed_a	0.001
	(0.0005)
a_bed_bf	−0.0001
	(0.0002)
a_bed_gi	0.001***
	(0.0002)
a_bed_hj	0.001*
	(0.001)
a_bed_kl	0.001
	(0.001)
a_bed_mn	−0.001***
	(0.0003)

Vervolg op volgende pagina

Tabel 8 – vervolg

	Dependent variable:
	gem_lichtvervuiling
a_bed_oq	0.001**
	(0.0003)
a_bed_ru	−0.0004
	(0.0004)
g_wozbag	0.001***
	(0.0001)
bev_dich2	−0.000**
	(0.000)
log(a_hh)	0.366
	(0.244)
log(a_00_14)	−0.270***
	(0.072)
a_15_24	−0.00005
	(0.00003)
log(a_25_44)	0.369***
	(0.075)
a_65_oo	0.00001
	(0.00003)
Constant	−1.140***
	(0.372)
Observations	2,920
Log Likelihood	−1,831.148
Akaike Inf. Crit.	4,388.295
Note:	*p<0.1; **p<0.05; ***p<0.01