Faculteit Bio-ingenieurswetenschappen

Academiejaar 2015-2016

Het ontwerpen van een zwarte koolstof kaart van stadsregio Gent voor gebruik bij routing problemen

Annelies Van den Hove
Promotors: prof. dr. Bernard De Baets en dr. ir. Jan Verwaeren
Tutor: ir. Joris Van den Bossche

Masterproef voorgedragen tot het behalen van de graad van
Master in de bio-ingenieurswetenschappen: Milieutechnologie

Gent, juni 2016

De promotors, De tutor,

prof. dr. Bernard De Baets dr. ir. Jan Verwaeren ir. Joris Van den Bossche

De auteur,

Annelies Van den Hove

Woord vooraf

Met deze masterproef sluit ik een verrijkende periode van vijf jaar af, een periode die mij als student, maar zeker ook als mens gevormd heeft. Zonder de hulp en raad van bepaalde personen zou deze masterproef nooit tot stand gekomen zijn. Dit is dan ook een uitgelezen moment om een aantal personen te bedanken.

Allereerst wil ik prof. dr. Bernard De Baets bedanken om mij de kans te geven aan deze masterproef te werken, voor zijn advies en ideeën. Daarnaast wil ik dr. ir. Jan Verwaeren bedanken voor de vele tijd die hij heeft vrijgemaakt, de hulp en deskundige uitleg. Verder wil ik ir. Joris Van den Bossche bedanken voor de uitstekende begeleiding en tips gedurende het realiseren van deze masterproef. Alsook een dankwoord aan de mensen van de vakgroep.

Met speciale aandacht wil ik ook mijn ouders bedanken voor de kans die ze mij gegeven hebben om bio-ingenieurswetenschappen te studeren en voor de vele steun. Ook wil ik mijn zus Elien bedanken voor de vele steunberichtjes en om in mij te geloven. Mijn vriend Stef krijgt in dit dankwoord een bijzondere vermelding. Hem wil ik bedanken voor de laatste lezing en om er altijd voor mij te zijn. Tevens een dankwoord aan alle vrienden en vriendinnen die ik de voorbije vijf jaar heb leren kennen en voor de fijne momenten.

Samenvatting

Deze masterproef beoogt het ontwerpen van een zwarte koolstof kaart van de stadsregio Gent, en het gebruik van deze kaart om een routeplanner te ontwerpen die de blootstelling van een fietser aan zwarte koolstof minimaliseert. Zwarte koolstof (Eng.: *black carbon*) is een lichtabsorberende en koolstofbevattende component aanwezig in fijn stof. Enkele bronnen van zwarte koolstof zijn: mobiele bronnen, huishoudelijke verwarming en open biomassa verbranding. Op Europees of Vlaams niveau is er op dit moment geen emissiegrenswaarde voor zwarte koolstof. Toch kan zwarte koolstof een waardevolle additionele luchtkwaliteitsindicator zijn van schadelijke partikels afkomstig van verbrandingsprocessen.

Voor het ontwerpen van deze kaart werd gebruik gemaakt van pollutiedata (i.e. mobiele metingen die toelaten de ruimtelijke en temporele variabiliteit in kaart te brengen). In deze masterproef werden enerzijds nieuwe pollutiedata verzameld met het instrument MicroAeth® en anderzijds gebruik gemaakt van mobiele data die verzameld werden door het Gents MilieuFront (GMF) voor de Vlaamse Instelling voor Technologisch Onderzoek (VITO). Voor locaties waar geen pollutiedata beschikbaar waren, werden landgebruiksregressiemodellen (Eng.: *land-use regression models*) ontwikkeld. Voor het ontwikkelen van deze landgebruiksregressiemodellen werden in een eerste stap data verzameld over omgevingseigenschappen die konden gebruikt worden om de concentratie aan zwarte koolstof te voorspellen. Deze verzamelde data bestonden in hoofdzaak uit een aantal zorgvuldig gekozen Geografisch InformatieSysteem (GIS)-lagen. Zo werd gebruik gemaakt van OpenStreetMap (OSM), meerbepaald de *highway-key*, Urban Atlas, het Centraal ReferentieAdressenBestand (CRAB), het verkeersmodel referentiejaar 2014 en het hoogtemodel van Vlaanderen om daaruit een *sky view factor* kaart te berekenen. Deze GIS-lagen werden gebruikt om een uitgebreide feature-extractie te kunnen uitvoeren. Bij het extraheren werd rekening gehouden met de mogelijke correlatie tussen de verschillende omgevingsvariabelen en zwarte koolstof. Uiteindelijk werden veertig features geëxtraheerd.

Deze features konden dan gebruikt worden als inputvariabelen voor een landgebruiksregressiemodel. In deze masterproef werd een selectie gemaakt van regressietechnieken die cou-

rant gebruikt worden in hedendaagse voorspellingsmodellen: lineaire regressie, *forward* en *backward stepwise selection*, ridge regressie, lasso, *support vector* regressie, regressiebomen, *random forests* en *K-Nearest Neighbors* regressie. Van deze verschillende regressietechnieken werd hun performantie geanalyseerd door middel van ruimtelijk gestratificeerde 4-fold cross-validatie. Uit deze analyse werd besloten om de zwarte koolstof kaart van de stadsregio Gent te ontwerpen met behulp van de regressietechniek lasso. Dit landgebruiksregressiemodel kon 43.57% van alle variabiliteit in de data aanwezig verklaren.

Deze zwarte koolstof kaart werd vervolgens gebruikt bij het ontwerpen van een routeplanner die de blootstelling van een fietser aan zwarte koolstof minimaliseerde. Hiervoor werd met elke straat een kost geassocieerd. Deze kost was de zwarte koolstof hoeveelheid. Voor het berekenen van het optimale pad werd gebruik gemaakt van een implementatie van het Dijkstra algoritme. Tot slot werden traditionele kortste pad trajecten vergeleken met trajecten die bekomen werden uit de ontworpen routeplanner. Voor een aantal start- en eindlocaties traden er geen verschillen op in de zwarte koolstof blootstelling en in de lengte van het traject. Het kortste pad stemde in die gevallen overeen met de laagste blootstelling aan zwarte koolstof. Echter in meer dan de helft van de berekende trajecten werd een afname van de blootstelling aan zwarte koolstof vastgesteld wanneer gebruik werd gemaakt van de ontworpen routeplanner. De procentuele afname aan zwarte koolstof hoeveelheid liep voor de berekende negentien routes op tot maximaal 15.96% (0.527 μg). Dit ging telkens gepaard met een toename in de lengte van het traject.

Inhoudsopgave

Lijst van afkortingen

AGIV	Agentschap voor Geografische Informatie Vlaanderen
BC	Black Carbon
CRAB	Centraal ReferentieAdressenBestand
DHMV	Digitaal HoogteModel Vlaanderen
EEA	European Environment Agency
ESCAPE	European Study of Cohorts for Air Pollution Effects
GIS	Geografisch InformatieSysteem
GMF	Gents MilieuFront
IRCEL	Intergewestelijke Cel voor het Leefmilieu
KNN	K-Nearest Neighbors
LUR	Land-Use Regression
MAAP	Multi-Angle Absorption Photometer
MSE	Mean Squared Error
ONA	Optimized Noise-reduction Averaging
OSM	OpenStreetMap
PM	Particulate Matter
POI	Point Of Interest
RMSE	Root Mean Squared Error
RSS	Residual Sum of Squares

RV	Relatieve luchtvochtigheid
SVR	Support Vector Regressie
UFP	UltraFine Particles
VITO	Vlaamse Instelling voor Technologisch Onderzoek
VMM	Vlaamse MilieuMaatschappij

HOOFDSTUK 1

Inleiding

1.1 Situering en motivatie

Hoewel de luchtkwaliteit in Europa de afgelopen decennia aanzienlijk verbeterd is, blijft ze de belangrijkste milieufactor die in verband wordt gebracht met vermijdbare ziektes en voortijdige sterfte in de Europese Unie. Tevens heeft de luchtkwaliteit een grote impact op een groot deel van de natuurlijke omgeving [28]. De luchtkwaliteit in stedelijke omgeving wordt grotendeels bepaald door fijn stof, een wijdverspreide polluent die bestaat uit een mengsel van vaste en vloeibare partikels gesuspendeerd in de atmosfeer. Vaakgebruikte indicatoren die fijn stof beschrijven zijn PM_{10} en $PM_{2.5}$, die refereren naar respectievelijk de massaconcentratie van partikels met een aerodynamische diameter van minder dan 10 μm en minder dan 2.5 μm [81]. De Europese doelstellingen voor PM_{10} werden in 2013 voor Vlaanderen slechts deels behaald. Zo was er aan de Europese jaargrenswaarde van 40 $\frac{\mu g}{m^3}$ voldaan, maar was er bij 3 van de 36 meetstations een overschrijding van de Europese daggrenswaarde. Dit betekent dat er op jaarbasis meer dan 35 dagen waren waarbij de concentratie hoger was dan 50 $\frac{\mu g}{m^3}$. Alle Vlaamse meetstations voldeden in 2013 aan de toen toekomstige jaargrenswaarde van $PM_{2.5}$, namelijk 25 $\frac{\mu g}{m^3}$ tegen 2015 (VMM [77]). De bovenstaande gegevens illustreren dat fijn stof nog steeds een probleem vormt voor de luchtkwaliteit in België. In het kader van deze masterproef wordt daarom de ruimtelijke variabiliteit in kaart gebracht en mitigerende maatregelen voorgesteld voor de nadelige effecten.

In deze masterproef wordt een specifieke fractie van fijn stof bestudeerd, namelijk zwarte koolstof. Zwarte koolstof (Eng.: *black carbon*) is een licht-absorberende en koolstofbevattende component aanwezig in fijn stof [25]. Een uitgebreidere definitie is terug te vinden in Sectie 2.1. Voor zwarte koolstof is er echter op dit moment nog geen grenswaarde op Europees of Vlaams niveau (VMM [77]). Deze masterproef beoogt het ontwerpen van een luchtvervuilingskaart van de stadsregio Gent die de verschillende concentratieniveaus van zwarte koolstof

weerspiegelt om zo de ruimtelijke variabiliteit van deze component in kaart te brengen. Voor het ontwerpen van deze zwarte koolstof kaart wordt gebruik gemaakt van pollutiedata; i.e. mobiele metingen die toelaten de ruimtelijke en temporele variabiliteit van zwarte koolstof in kaart te brengen. Voor locaties waar geen data beschikbaar zijn, worden landgebruiksregressiemodellen[1] ontwikkeld. Deze modellen maken gebruik van verklarende variabelen, afgeleid uit Geografisch InformatieSysteem (GIS)-lagen, en kunnen gebruikt worden om concentraties aan zwarte koolstof te voorspellen. Vervolgens wordt deze kaart gebruikt om aan routing te doen; i.e. het plannen van een (fiets)traject dat de blootstelling van de reiziger aan zwarte koolstof minimaliseert.

1.2 Doelstellingen

De doelstelling van deze masterproef is tweeledig: (i) het ontwerpen van een zwarte koolstof kaart van de stadsregio Gent, en (ii) het gebruik van deze kaart bij routing problemen. Er zal getracht worden om volgende doelstellingen te bereiken:

1. Het opstellen van een zwarte koolstof kaart op basis van mobiele meetdata.
2. Onderzoeken welke omgevingseigenschappen het meest geschikt zijn om zwarte koolstof te voorspellen.
3. Evalueren van de performantie die men kan behalen bij het voorspellen van zwarte koolstof.
4. Het ontwerpen van een routeplanner die de blootstelling van een fietser aan zwarte koolstof minimaliseert.
5. Het vergelijken van traditionele kortste pad trajecten met trajecten die bekomen worden uit Doelstelling 4 met betrekking tot de blootstelling aan zwarte koolstof.

1.3 Opbouw van de masterproef

Het ontwerpen van een zwarte koolstof kaart kan opgesplitst worden in verschillende deeltaken, in overeenstemming met de eerste hoofdstukken in deze masterproef. Vooreerst wordt meer duiding gegeven over zwarte koolstof in Hoofdstuk 2. In Hoofdstuk 3 wordt de verzameling en verwerking van pollutiedata besproken. Daarnaast werden ook data verzameld over omgevingseigenschappen die kunnen gebruikt worden om de concentratie van zwarte koolstof te voorspellen op plaatsen waarvoor geen pollutiedata beschikbaar zijn. De verzamelde data bestaan in hoofdzaak uit een aantal zorgvuldig gekozen GIS-lagen. In Hoofdstuk 4 wordt

[1] Eng.: *Land-Use Regression* (LUR) *models*.

beschreven hoe uit deze GIS-lagen een uitgebreide feature-extractie wordt uitgevoerd. Deze features kunnen dan gebruikt worden als inputvariabelen voor een regressiemodel. In Hoofdstuk 5 worden verschillende regressietechnieken gebruikt om een landgebruiksregressiemodel te ontwikkelen dat op basis van de omgevingsvariabelen de locale concentratie van zwarte koolstof kan voorspellen. Verschillende regressietechnieken werden daarbij onderling vergeleken. Op basis van het beste regressiemodel wordt vervolgens een zwarte koolstof kaart van de stadsregio Gent ontworpen. Deze kaart wordt tenslotte gebruikt bij het ontwerpen van een routeplanner die de blootstelling van een fietser aan zwarte koolstof minimaliseert. In Hoofdstuk 6 wordt meer duiding gegeven over deze routeplanner en worden enkele routing problemen uitgewerkt.

1.4 Gerelateerd onderzoek

Deze sectie geeft een overzicht van gerelateerd onderzoek waarbij mobiele monitoringsmethoden werden gebruikt voor het monitoren van polluenten voor diverse doeleinden. Aangezien men bij het ontwerpen van de luchtvervuilingskaart de spatiale variabiliteit in kaart wil brengen, dient men de spatio-temporele data te aggregeren over de tijd. Mogelijke aggregratiemethoden uit de literatuur worden daarom vervolgens toegelicht. Daarnaast wordt toegelicht welke regressietechnieken en features frequent worden gebruikt bij het opbouwen van landgebruiksregressiemodellen. Tot slot worden enkele routeplanners op basis van luchtverontreiniging besproken.

1.4.1 Mobiele monitoringsmethoden

Voor het monitoren van de concentraties van polluenten in de atmosfeer wordt steeds vaker gebruik gemaakt van mobiele monitoringsmethoden. Verschillende publicaties vermelden het gebruik van mobiele monitoringsmethoden voor het bestuderen van verschillende fijn stof fracties, zoals PM_{10}, $PM_{2.5}$, PM_1 en ultrafijne partikels (UFP, i.e. fijn stof met een aerodynamische diameter kleiner dan 0.1 μm) [10, 14, 17, 19, 34, 38, 43, 56, 57, 80, 82], zwarte koolstof [22, 43, 57, 72, 80], stikstofoxiden (NO, NO_2, NO_x) [9, 19, 33, 43, 80], partikelgebonden polycyclische aromatische koolwaterstoffen [19, 80, 82] en koolstofmonoxide (CO) [34, 80]. Ook de doeleinden van mobiel monitoren kunnen verschillen: het onderzoeken van de spatiale en seizoenale variatie [17], het onderzoeken van de variabiliteit van de pluim van polluenten onder stabiele weerscondities [19], het evalueren van persoonlijke blootstelling aan een polluent [22], het maken van een luchtvervuilingskaart met een hoge spatiale resolutie [34, 38]. Hierdoor kan ook de manier waarop mobiele data worden bekomen eveneens verschillen. Elen et al. [24] beschrijven de ontwikkeling van de Aeroflex, een fiets uitgerust met meettoestellen die de luchtkwaliteit meten door onder andere fijn stof en zwarte koolstof te monitoren. Hierbij werd elke meting gelinkt aan zijn corresponderende geografische locatie en tijdstip. Op

die manier bekwam men metingen met een hoge spatiale en temporele resolutie. Naast het inwinnen van mobiele data met een fiets [14, 57], worden ook mobiele data ingewonnen door voetgangers [82], door meettoestellen te plaatsen op elektrische voertuigen [19, 34, 43, 80], op trams [33, 38] of door meettoestellen te plaatsen in voertuigen [14].

1.4.2 Aggregatiemethoden

Berghmans et al. [10] brachten de variabiliteit van PM_{10} en UFP in kaart. Het opstellen van deze kaarten gebeurde telkens op basis van één passage. Mobiele meetdata verzameld gedurende één passage kan vergeleken worden met een *snapshot* dat informatie bevat over de spatiale variabiliteit. Om tevens de temporele variabiliteit in rekening te brengen zal men gebruik moeten maken van meerdere passages gespreid in de tijd. Door deze temporele variabiliteit kan het tijdstip waarop de metingen worden genomen een sterk effect hebben op de gemeten concentratie van bijvoorbeeld UFP. Zo kan de concentratie lager zijn in een drukke straat dan in een rustige straat, wanneer de meting werd uitgevoerd respectievelijk gedurende de nacht en gedurende de spits. Daarnaast kunnen ook occasionele gebeurtenissen een rol spelen in de gemeten concentraties, wat niet noodzakelijk representatief is voor deze locatie over een langere periode. Mogelijke strategieën om de voornoemde twee impacten te reduceren, zijn het toepassen van een achtergrondcorrectie of het uitvoeren van een meetcampagne. Bij een achtergrondcorrectie worden de data gecorrigeerd door de achtergrondconcentratie af te trekken van de gemeten concentraties. Op die manier geven de gecorrigeerde data enkel de lokale bijdrage van aanwezige bronnen weer. Een achtergrondconcentratie kan afkomstig zijn van een stationair meetstation. Bij een meetcampagne wordt meermaals dezelfde route gereden op verschillende tijdstippen, waarbij vervolgens op een aantal gekozen locaties op de route de spatio-temporele data worden geaggregeerd over de tijd [24, 56]. Een mogelijke aggregatiemethode is het opdelen van het studiegebied volgens een grid in verschillende cellen van gelijke grootte gevolgd door het berekenen van de gemiddelde concentraties van de polluent per cel [38]. Naast het gemiddelde kan men de mediaan berekenen van de concentraties van de polluent en bovendien gebruik maken van straatsegmenten in plaats van een grid. De keuze voor de mediaan en niet het gemiddelde wordt gemotiveerd door het feit dat luchtkwaliteitsdata niet symmetrisch verdeeld zijn [56]. Peters et al. [57] maakten gebruik van een spatiale database die vaste punten bevatte om de 10 m gelegen op de route. Aan elk vast punt werd vervolgens een waarde toegekend op basis van het gewogen gemiddelde met als wegingscoëfficiënt de afstand. Een andere aggregatiemethode is het α-getrimd gemiddelde om zo de spatiale variabiliteit in kaart te brengen [72]. Het voordeel van deze methode is dat het ruimtelijk geaggregeerd resultaat minder gevoelig wordt voor extreme waarden. Hagler et al. [34] maakten gebruik van een *bias-removal* algoritme om occasionele gebeurtenissen te verwijderen. Hierbij beschouwde men het verdubbelen van de gemeten CO concentratie binnen een tijdspanne van 1 seconde als een indicator van een occasionele gebeurtenis en

werden deze data daaropvolgend verwijderd vanaf het verdubbelen van de concentratie tot de concentratie minder dan 75% was van de initiële piek. Een nadeel van deze methode is dat het algoritme berust op subjectieve criteria. Vervolgens werden alle datapunten geaggregeerd die in een cirkel met diameter van 15 m liggen met als middelpunt een specifieke geografische locatie, waarna het gemiddelde werd genomen van de CO concentraties.

1.4.3 Landgebruiksregressiemodellen

Algemeen kunnen landgebruiksregressiemodellen (LUR modellen) als volgt worden voorgesteld [47]:

$$Y = f(X) + \epsilon, \tag{1.1}$$

met Y een kwantitatieve outputvariabele en $X = (X_1, X_2, \ldots, X_p)$, een vector van p verschillende features. Hierbij wordt aangenomen dat er een relatie bestaat tussen Y en X, waarbij f een ongekende functie is. ϵ is een foutterm die onafhankelijk is van X en waarvan de verwachtingswaarde nul is. Volgens Hoek et al. [42] zijn veelgebruikte features verkeersvariabelen, populatiedensiteit, landgebruiksdata, hoogte-informatie en topografie, meteorologie en geografische coördinaten. Bovendien ontwikkelt men in de meeste studies LUR modellen met behulp van *forward stepwise selection*, *backward stepwise selection* of *best subset selection*. Deze regressietechnieken worden in Sectie 5.1 toegelicht.

Het ESCAPE (*European Study of Cohorts for Air Pollution Effects*) project is een toonaangevend project[2] dat de invloed bestudeert van de blootstelling aan luchtverontreiniging op de gezondheid van de mens. LUR modellen worden in het kader van het ESCAPE project ontwikkeld op een gestandaardiseerde manier. Voor het extraheren van features werd gebruik gemaakt van onder andere volgende types databronnen: het wegennetwerk, landgebruiksdata, data over populatiedensiteit en hoogtegegevens, aangevuld met gegevens over verkeersintensiteit. Op die manier extraheerde men features, zoals oppervlakte van hoog- en laagstedelijk gebied (m^2), havenoppervlakte (m^2), oppervlakte aan stedelijk groengebied (m^2), aantal huishoudens (-), totale lengte van alle straten in een buffer (m), verkeersintensiteit op de dichtstbijzijnde weg (voertuigen/dag), totale verkeersintensiteit van alle straten in een bepaalde buffer (som van (verkeersintensiteit·lengte van alle straatsegmenten)) (voertuigen/dag·m), enz. De ontwikkeling van het LUR model gebeurde met behulp van een *supervised forward stepwise* procedure [9, 23]. Hasenfratz et al. [38, 39] extraheerden twaalf features: populatie (aantal inwoners/ha), hoogte gebouwen (aantal verdiepingen/ha), hoogte terreinen (m/ha), afstand tot dichtstbijzijnde weg (m), terreinhelling (graden/ha), verkeersvolume (voertuigen/(dag·ha)), industrie (aantal industriegebouwen/ha), verwarming (aantal olie- en gasverwarmingen/ha), wegtype (i.e. residentiële weg, tertiaire weg, secundaire weg, primaire weg en snelweg) (drukste wegtype/ha), afstand tot de dichtstbijzijnde grote weg

[2]Een project binnen het zevende kaderprogramma van de EU.

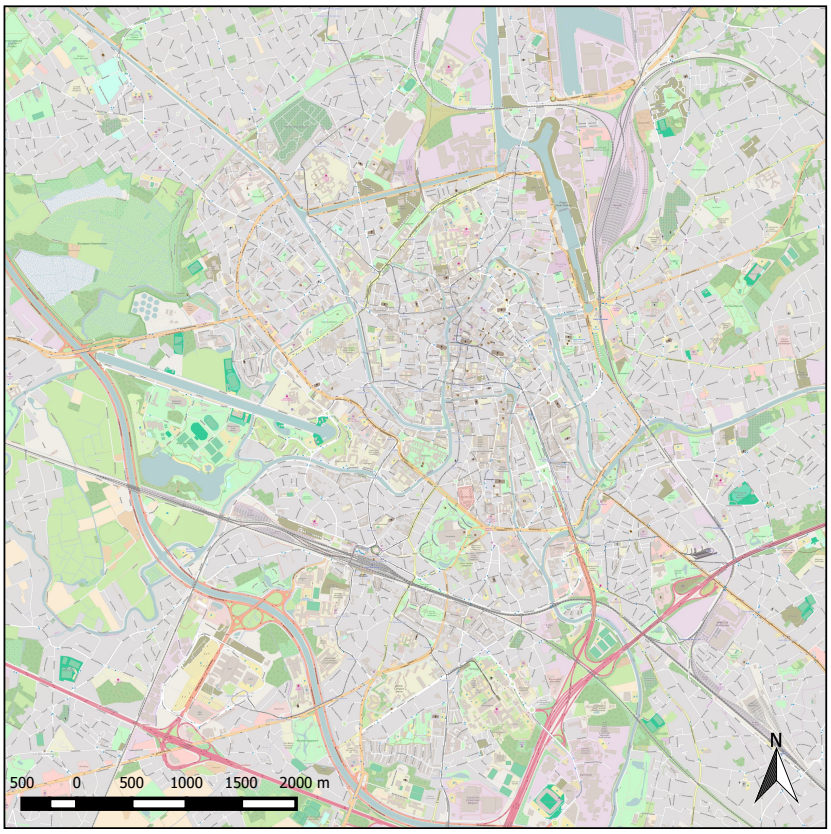
(i.e. snelweg, primaire weg, secundaire weg) (m), terreinaspect (graden/ha) en afstand tot het dichtstbijzijnde verkeerslicht (m). Voor de ontwikkeling van het LUR model werd gebruik gemaakt van *Generalized Additive Models* (GAMs).

1.4.4 Routeplanner op basis van luchtverontreiniging

Vaak wordt met een routeplanner de kortste afstand tussen twee locaties berekend. Soms kan men als voetganger of als fietser niet alleen geïnteresseerd zijn in het kortste pad, maar ook in het afleggen van een traject dat rekening houdt met luchtverontreiniging. De website `https://walkit.com/` [79] is hiervan een toepassing. Dit is een routeplanner, waarbij voetgangers kunnen kiezen tussen het kortste pad, een pad dat drukke straten vermijdt, of een pad waarbij de blootstelling aan NO_2 wordt geminimaliseerd. Hertel et al. [41] vonden dat een goede routekeuze kan leiden tot een significante reductie in de blootstelling aan luchtverontreiniging. In deze studie maakte men voor het berekenen van de routes echter gebruik van een verkeersmodel als benadering van de blootstelling aan luchtverontreiniging. Bovenstaande luchtkwaliteitsmodellen (landgebruiksregressiemodellen) kunnen ook gebruikt worden om aan routing te doen. Su et al. [67] gebruikten een landgebruiksregressiemodel om een routeplanner voor fietsers te ontwikkelen die routes berekende met onder andere de laagste gemiddelde NO_2 concentratie. Wil men de optimale route vinden, dan moeten kosten met straatsegmenten worden geassocieerd. Sharker et al. [66] onderzochten een kostfunctie die kosten berekende op basis van de blootstelling aan luchtverontreiniging. Hiervoor gebruikte men data afkomstig van stationaire meetstations. De kostfunctie werd bekomen door de gemiddelde luchtkwaliteitsindex per straatsegment te vermenigvuldigen met de tijd nodig om dit straatsegment te nemen. Daarnaast werd in een studie van Hasenfratz et al. [38] een nieuwe kostfunctie bekomen voor alle straatsegmenten door de voorspelde UFP concentratie voor een bepaald straatsegment te vermenigvuldigen met de lengte van dit straatsegment. De voorspelde concentraties waren afkomstig van een landgebruiksregressiemodel. Het pad dat rekening hield met UFP in plaats van het kortste pad, leidde tot een gemiddelde blootstellingsreductie van 7.1% ($6.8 \cdot 10^6$ $\frac{\text{partikels}}{\text{cm}^3\text{m}}$). De lengte van dit pad was wel gemiddeld 6.4% (548 m) langer dan het kortste pad.

1.5 Keuze studiegebied

In deze masterproef werd als studiegebied de stadsregio Gent gekozen (Figuur 1.1). Rekening houdend met het bereik van de ontvangen pollutiedata werd een gebied van 7.5 x 7.5 km geselecteerd. Dit gebied omvat onder meer het centrum van Gent, Gentbrugge, het stedelijk natuurreservaat Bourgoyen-Ossemeersen en sport- en recreatiepark Blaarmeersen.



Figuur 1.1: Studiegebied. Kaartgegevens afkomstig van OpenStreetMapdatabase gevisualiseerd met QGIS [62].

HOOFDSTUK 2

Zwarte koolstof

Dit hoofdstuk behandelt de volgende aspecten: (i) de definitie van zwarte koolstof samen met de belangrijkste bronnen, (ii) een bespreking van de voornaamste technieken die gebruikt worden om zwarte koolstof in een stadsomgeving te meten, (iii) een studie van zwarte koolstof concentraties in België (deze bevindingen zijn het resultaat van een analyse van data afkomstig van de Intergewestelijke Cel voor het Leefmilieu (IRCEL) die in het kader van deze masterproef werd uitgevoerd), (iv) de invloed van zwarte koolstof op mens en milieu.

2.1 Definitie van zwarte koolstof

Een definitie van zwarte koolstof is: 'Zwarte koolstof is een licht-absorberende en koolstofbevattende component aanwezig in fijn stof (EEA [25]).' Het is een primaire polluent, dus het wordt rechtstreeks geëmitteerd in de lucht vanaf de bron en wordt niet gevormd in de atmosfeer vanuit precursoren. Het is een indicator voor de roetconcentratie in de omgevingslucht. Hierbij bevindt zwarte koolstof zich voornamelijk in de ultrafijne fractie van fijn stof en is het afkomstig van onvolledige verbrandingsprocessen. Zwarte koolstof absorbeert sterk het zichtbare licht en is vuurvast met een verdampingstemperatuur nabij 4000 K. Tevens is het onoplosbaar in water en in gebruikelijke organische solventen. Het heeft een korte levensduur in de atmosfeer van enkele dagen tot weken, waarna het verdwijnt uit de atmosfeer via depositie of via contact met oppervlaktes. Hierdoor is de concentratie aan zwarte koolstof nabij de bron hoger in vergelijking met meer afgelegen regio's. Enkele bronnen van zwarte koolstof zijn [13, 25, 48, 77, 78]:

1. Mobiele bronnen, voornamelijk voertuigen, machines gebruikt in de bosbouw en landbouw, locomotieven, schepen, enz.

2. Huishoudelijke verwarming, voornamelijk het verbranden van biomassa, hout en kolen.

3. Open biomassa verbranding, zoals bosbranden of het verbranden van landbouwafval.

Op Europees of Vlaams niveau is er op dit moment geen emissiegrenswaarde voor zwarte koolstof. Wel bestaan er grenswaarden voor fijn stof, die op die manier bijdragen aan een indirecte regulering van zwarte koolstof emissies [25, 77].

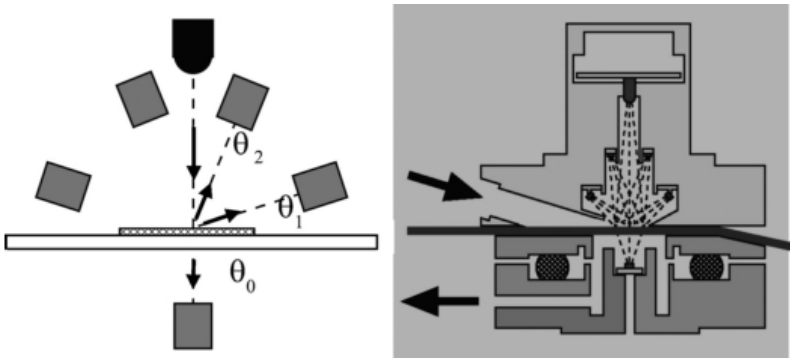
Roet, wat het product is van onvolledige verbranding, kan echter op verschillende manieren geanalyseerd worden. Zo kan roet gemeten worden op basis van zijn licht-absorberende eigenschappen of op basis van zijn thermo-optische eigenschappen. Als men de licht-absorberende eigenschappen meet, dan beschrijft men roet als zwarte koolstof. Meet men echter de thermo-optische eigenschappen, dan beschrijft men roet als elementair koolstof [25]. Zwarte koolstof is dus niet hetzelfde als elementair koolstof [48]. Deze masterproef beperkt zich tot zwarte koolstof.

2.2 Meten van zwarte koolstof

Voor de bepaling van zwarte koolstof bestaan er nog geen gestandaardiseerde meetmethoden. Het Europees comité voor standaardisatie inzake luchtkwaliteit (CEN/ TC 264) richtte een werkgroep (WG35) op. Deze groep werkt sinds eind 2013 aan de ontwikkeling van een referentiemethode om zwarte koolstof te meten. In deze sectie worden de twee meest gebruikte filterabsorptie fotometers besproken: de *Multi-Angle Absorption Photometer* (MAAP) en de aethalometer [25, 77]. Daarnaast wordt de MicroAeth® besproken, die werkt volgens hetzelfde meetprincipe als de aethalometer, maar is licht en in zakformaat beschikbaar.

2.2.1 Multi-Angle Absorption Photometer

De *Multi-Angle Absorption Photometer* (MAAP) is een instrument om aerosol lichtabsorptie en zwarte koolstof te bepalen door simultaan de straling te meten die passeert doorheen en teruggekaatst wordt door een partikelgeladen glasvezelfilter. Hierbij benadert men de partikelgeladen glasvezelfilter als twee lagen: een aerosol-filterlaag en een partikelvrije filterlaag. De aerosol-filterlaag bevat de partikels afgezet op de filter. De meting van aerosol lichtabsorptie in het zichtbare spectrum is sterk gecorreleerd met de meting van zwarte koolstof. De meting wordt uitgevoerd op drie detectiehoeken ($\theta_0 = 0°$, $\theta_1 = 130°$ en $\theta_2 = 165°$) om de invloed van lichtverstrooiende effecten veroorzaakt door de aerosol en door de filtermatrix te verminderen (Figuur 2.1) [58, 59]. Dit instrument wordt gebruikt door de Vlaamse MilieuMaatschappij (VMM) om zwarte koolstof te monitoren en meet op halfuurbasis. VMM meet en rapporteert de verontreiniging van de omgevingslucht. In Gent heeft VMM twee meetstations die zwarte koolstof meten. Hiervan bevindt zich één in het Baudelopark (i.e. stedelijk gebied) en één in de Gustaaf Callierlaan (i.e. verkeersgericht gebied) [45, 77].



(a) (b)

Figuur 2.1: Optische sensor van MAAP. (a) Positie van de fotodetectors op de detectiehoeken $\theta_0 = 0^\circ$, $\theta_1 = 130^\circ$ en $\theta_2 = 165^\circ$ ten opzichte van de invallende lichtstraal. (b) MAAP sensor. Pijlen duiden de stroming van lucht doorheen de sensor aan [58].

2.2.2 Aethalometer

De aethalometer is een instrument dat de *real-time* concentratie van zwarte koolstof meet door middel van absorptie op een filter [36]. Hierbij wordt continu de attenuatie van een lichtstraal gemeten die doorheen een filter wordt gestuurd samen met de partikelhoudende lucht. Bij een constante snelheid van lucht doorheen de filter is de depositie van zwarte koolstof op de filter evenredig met de concentratie in de atmosfeer en geeft het een corresponderende toename in optische attenuatie. Het meten van deze optische attenuatie is het basisprincipe van deze methode. In volgende paragraaf wordt het werkingsprincipe uitgelegd.

De aethalometer maakt gebruik van een filter die uniform wordt verlicht door een lamp. Deze filter wordt afgedekt met een transparant masker, vaak een quartsvezel, die een opening bevat van 5 mm diameter. Op die manier gaat slechts een deel van de luchtstroom doorheen de filter waarop de partikels worden verzameld. Het overige deel van de filter dient als een referentie voor optische metingen of als een blanco voor chemische analyses. Onder het collecterend gedeelte en het referentiegedeelte van de filter bevinden zich twee optische vezels die het doorgelaten licht overdragen aan een paar aangepaste fotodetectoren. De attenuatie A van de intensiteit I, die doorheen het collecterend gedeelte van de filter gaat, relatief ten opzichte van de intensiteit I_0, die doorheen het referentiegedeelte gaat, is $A = 100 \cdot \ln(I_0/I)$. Dit is evenredig met de concentratie van zwarte koolstof. De output is een spanning die proportioneel is met de optische attenuatie A. De toename in spanning is evenredig met de snelheid van depositie van zwarte koolstof op de filter. Na verloop van tijd moet de filter vervangen worden om effecten van optische saturatie te vermijden [36].

2.2.3 MicroAeth®

Voor *real-time* en persoonlijke metingen kan men beroep doen op de MicroAeth®. Dit instrument werkt volgens hetzelfde principe als de aethalometer, maar is licht en beschikbaar in zakformaat. Een nadeel van deze methode is dat de relatieve luchtvochtigheid (RV) en temperatuur de datakwaliteit beïnvloeden. Zo kunnen snelle veranderingen in RV aanleiding geven tot onder- en overschattingen in zwarte koolstof concentraties. Een snelle toename in RV kan resulteren in condensatie van vocht op het collecterend gedeelte van de filter. Dit vocht kan vervolgens het licht verstrooiien en bijdragen aan lichtattenuatie. Bijgevolg kan het aanleiding geven tot overschattingen. Onderschattingen worden bekomen door evaporatie van het vastgehouden vocht. Sterke onderschattingen resulteren in berekende zwarte koolstof concentraties die negatief zijn, wat niet mogelijk is. Dit kan doordat de berekening van zwarte koolstof gebaseerd is op opeenvolgende verschillen in attenuatie. Bij de berekening van attenuatie wordt het collecterend gedeelte vergeleken met het referentiegedeelte van de filter. Gezien er door het referentiegedeelte geen lucht wordt doorgelaten, zal dit niet zo snel beïnvloed worden door veranderingen in RV of in temperatuur in vergelijking met het collecterend gedeelte. Daardoor kunnen deze veranderingen daar een snellere impact hebben [18].

Een mogelijke oplossing is het gebruik van een droger, bijvoorbeeld een Nafion® tube, waardoor de impact van het binnenkomend vochtgehalte wordt gereduceerd. Dit is niet nodig wanneer de data uitgemiddeld wordt over een lange periode, bijvoorbeeld > 1 uur, gezien de onder- en overschattingen dan elkaar opheffen en verwaarloosbaar worden. De temperatuur kan gebufferd worden door de MicroAeth® te dragen in nabijheid van het lichaam, bijvoorbeeld in een jaszak [18]. Voorts kan de data sterk variabel zijn en ruis bevatten, zeker wanneer de tijdstap tussen de verschillende metingen klein is of als de concentraties laag zijn. Teneinde het echte signaal van de ruis te onderscheiden, is naverwerking van de data noodzakelijk (Sectie 3.1.1) [18, 35].

2.3 Studie van zwarte koolstof concentraties in België

De Intergewestelijke Cel voor het Leefmilieu (IRCEL) is een samenwerkingsakkoord tussen de Belgische Staat, het Vlaamse Gewest, het Waalse Gewest en het Brusselse Hoofdstedelijk Gewest inzake het toezicht op emissies in de lucht en op de structurering van gegevens hierover. Het doel van deze overeenkomst is een permanente samenwerking tussen de gewesten te verzekeren en dit op het vlak van het behoud van een gemeenschappelijke wetenschappelijke basis inzake de registratie, de interpretatie van de gegevens en de totstandkoming van rapporten met betrekking tot luchtvervuiling. Verder is er overleg inzake het beheer van de netwerken voor luchtmeting en staan ze ook in voor de ontwikkeling en het beheer van een permanente structuur voor het inzamelen van de gewestelijke gegevens. Voorts verzorgen ze

ook de werking van het Belgisch knooppunt van het Agentschap [45].

In 2014 waren er in België 23 meetstations voor het monitoren van zwarte koolstof concentraties [46]. Naargelang hun locatie worden deze meetstations ingedeeld in één van de volgende landgebruiksklassen: 'landelijk', 'voorstedelijk', 'stedelijk', 'industrieel' of 'verkeer' [45]. De gebruikte data[1] bevatten uurlijkse metingen van zwarte koolstof voor de periode van 15 oktober 2014 tot en met 14 oktober 2015. In het kader van deze masterproef werd voor elke landgebruiksklasse het dagelijkse gemiddelde en bijhorende standaarddeviatie berekend samen met de minimum- en maximumwaarde. De gekozen meetstations en locaties samen met de berekende waarden zijn terug te vinden in Tabel 2.1. Het rekenkundig gemiddelde per landgebruiksklasse werd vervolgens uitgezet in Figuur 2.2. Uit Tabel 2.1 kan men het volgende opmerken:

1. Uit de gemiddelde waarden blijkt dat de concentratie aan zwarte koolstof het laagst is in de landelijke gebieden. In het verkeersgericht gebied observeert men een concentratie die drie maal zo hoog is als deze geobserveerd in de landelijke gebieden (Figuur 2.2). Dit strookt met de bevindingen in Sectie 2.1, waaruit blijkt dat het verkeer een belangrijke bron is.

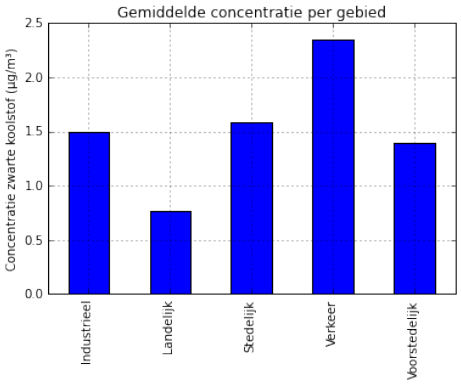
2. De gemiddelde waarden (Tabel 2.1, samengevat in Figuur 2.2) verschillen niet substantieel tussen de landgebruiksklassen 'voorstedelijk', 'stedelijk', 'industrieel' en schommelen rondom 1.5 $\frac{\mu g}{m^3}$.

Figuur 2.3 toont het verloop van de concentratie van zwarte koolstof in België uitgezet in functie van de tijd (berekend op basis van de data aangeleverd door IRCEL). Opmerkelijk in Figuur 2.3 is dat het vervuilingsniveau hoger is in de herfst en de winter dan in de lente en de zomer. Dit hoger vervuilingsniveau is vermoedelijk een gevolg van het frequenter voorkomen van temperatuursinversie[2] van oktober tot maart, waardoor de polluenten verhinderd worden de onderste luchtlagen te verlaten en zich daar gaan ophopen. Deze bevindingen stemmen overeen met een studie uitgevoerd door Hasenfratz et al. in Zwitserland [38]. Temperatuursinversie en een stabiele grenslaag zijn dan ook de slechtste condities met betrekking tot dispersie van polluenten in de atmosfeer. Hierbij wordt de turbulentie onderdrukt en de opwaarste beweging geëlimineerd [52]. Daarnaast wordt gedurende de herst en de winter meer gebruik gemaakt van huishoudelijke verwarming, waaronder het verbranden van hout, wat tevens een bron is van zwarte koolstof.

[1] Data gedownload via `http://sos.irceline.be/` op 15 oktober 2015 [1].
[2] Warme luchtlaag op een koudere luchtlaag [52].

Tabel 2.1: Code van meetstations en respectievelijke gemeente opgedeeld volgens de classificatie van de meetstations. Gemiddelde, minimum en maximum concentratie in $\frac{\mu g}{m^3}$ gemeten door de verschillende meetstations voor de periode van 15 oktober 2014 tot en met 14 oktober 2015.

Code	Gemeente	Classificatie	Gemiddelde	Standaardafwijking	Minimum	Maximum
44N029	Houtem (Veurne)	Landelijk	0.707	0.620	0.076	3.831
42N016	Dessel	Landelijk	1.142	0.742	0.189	4.369
43N085	Vielsalm	Landelijk	0.463	0.240	0.068	1.609
40SZ01	Steenokkerzeel	Voorstedelijk	1.439	0.884	0.242	5.810
40AL01	Antwerpen-Linkeroever	Voorstedelijk	1.289	0.886	0.167	6.255
42N045	Hasselt	Voorstedelijk	1.442	0.937	0.265	7.239
44R701	Gent (Baudelopark)	Stedelijk	1.528	0.919	0.333	5.386
41R001	Sint-Jans-Molenbeek	Stedelijk	1.838	0.956	0.539	6.575
43R221	Herstal	Stedelijk	1.374	1.106	0.180	9.382
44M705	Roeselare (Haven)	Industrieel	1.538	1.118	0.192	5.625
44R750	Zelzate	Industrieel	1.542	0.965	0.220	5.265
42R815	Zwijndrecht	Industrieel	1.431	0.994	0.185	6.919
42M802	Antwerpen (Luchtbal)	Industrieel	1.795	1.143	0.363	7.392
40AB01	Antwerpen (Boudewijnsluis)	Industrieel	1.478	0.910	0.310	5.504
40SA04	Hoevenen	Industrieel	1.270	0.911	0.210	5.492
45R512	Marchienne-Au-Pont	Industrieel	1.431	0.853	0.287	4.858
44R702	Gent (Gustaaf Callierlaan)	Verkeer	2.045	1.124	0.413	6.276
41R002	Elsene	Verkeer	2.205	0.866	0.538	5.073
42R802	Borgerhout (straatkant)	Verkeer	2.799	1.549	0.526	8.144



Figuur 2.2: Gemiddelde concentratie zwarte koolstof voor de verschillende landgebruiksklassen.

2.4 Invloed van zwarte koolstof op mens en milieu

2.4.1 Gezondheidseffecten

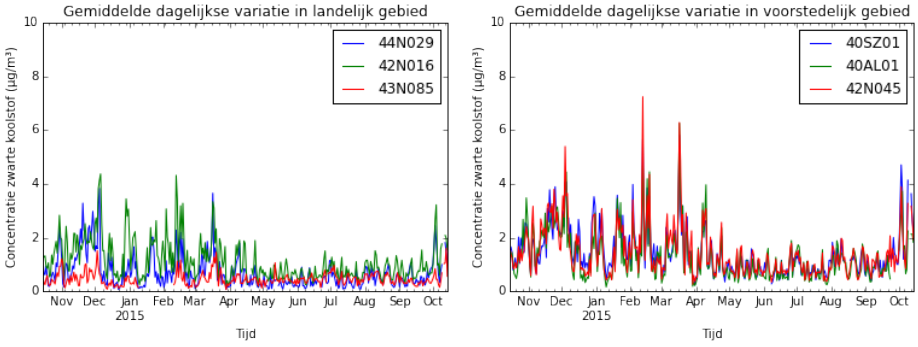
Blootstelling aan zwarte koolstof kan in verband worden gebracht met hart- en vaatziekten. De gezondheidseffecten geassocieerd met zwarte koolstof kunnen niet kwantitatief op dezelfde manier vergeleken worden met de fijn stof concentratie. Daarom is zwarte koolstof een waardevolle additionele luchtkwaliteitsindicator van schadelijke partikels afkomstig van verbrandingsprocessen (WHO [48]).

2.4.2 Toxiciteit

Zwarte koolstof is op zichzelf niet toxisch, maar kan wel een sleutelrol spelen als universele drager van toxische componenten. Deze toxische componenten kunnen componenten zijn die vrijgesteld werden bij het verbrandingsproces zelf of componenten zijn die eraan adsorberen gedurende het transport doorheen de atmosfeer (WHO [48]).

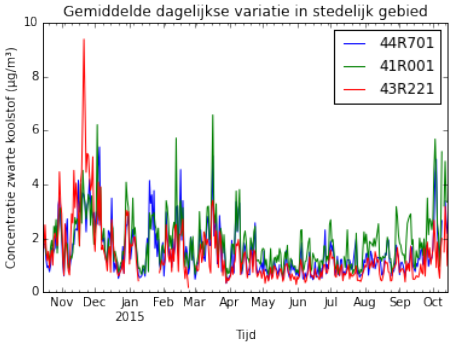
2.4.3 Invloed op het klimaat

Naast gezondheids- en toxiciteitseffecten zou zwarte koolstof ook bijdragen tot de klimaatverandering. Recent werd het beschouwd als een belangrijke bijdrager van *short-lived climate forcing* (SLCF). Zo zou zwarte koolstof niet alleen de invallende zonnestralen verstrooien, maar ook absorberen. Zwarte koolstof acteert over een kortere periode dan de klassieke broeikasgassen (bijvoorbeeld koolstofdioxide (CO_2)), omdat zijn levensduur in de atmosfeer ook korter is gezien het snel verwijderd wordt via depositie. Een reductie van zwarte koolstof emissies heeft dan ook een sneller effect dan reducties van andere broeikasgassen. Toch is

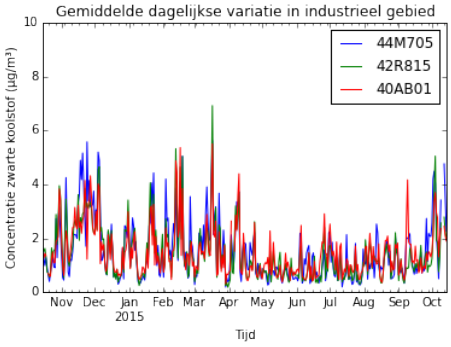


(a) Landelijk

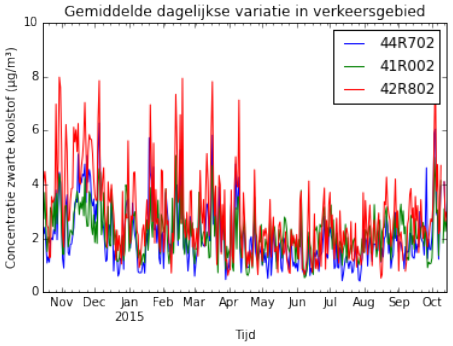
(b) Voorstedelijk



(c) Stedelijk



(d) Industrieel



(e) Verkeer

Figuur 2.3: Verloop van de concentratie van zwarte koolstof uitgezet in functie van de tijd en opgedeeld volgens de landgebruiksklassen: (a) 'landelijk', (b) 'voorstedelijk', (c) 'stedelijk', (d) 'industrieel' en (e) 'verkeer'.

er een substantiële onzekerheid over de bijdrage van de zwarte koolstofrijke bronnen aan de netto klimaatopwarming. Dit is voornamelijk te wijten aan het gebrek aan kennis over de interactie tussen wolken met zwarte koolstof en co-geëmitteerde organische koolstof [13, 25].

HOOFDSTUK 3

Dataverzameling en -verwerking

Uit voorgaand hoofdstuk blijkt dat verschillende omgevingsvariabelen gecorreleerd zijn met zwarte koolstof. Bij de ontwikkeling van een stratennetwerkdekkende kaart van zwarte koolstof concentraties voor de stadsregio Gent is het daarom nodig om geografische data te verzamelen waaruit relevante omgevingsvariabelen kunnen afgeleid worden. Op volgende databronnen werd daarom beroep gedaan: (i) pollutiedata, (ii) OpenStreetMap (OSM), meerbepaald de *highway-key*, (iii) Urban Atlas, (iv) Centraal ReferentieAdressenBestand (CRAB), (v) het verkeersmodel referentiejaar 2014 en (vi) het hoogtemodel van Vlaanderen om daaruit een *sky view factor* kaart te berekenen. In wat volgt wordt meer informatie verschaft over deze databronnen, alsook de keuze beargumenteerd. Daarnaast wordt ook de verwerking ervan besproken.

3.1 Databronnen

3.1.1 Pollutiedata

In deze masterproef werden enerzijds nieuwe pollutiedata verzameld (via mobiele monitoring) en anderzijds werd gebruik gemaakt van mobiele data die verzameld werden door het Gents MilieuFront (GMF) voor de Vlaamse Instelling voor Technologisch Onderzoek (VITO). GMF is een regionale milieuvereniging uit Gent die actief is op het vlak van het leefmilieu en duurzame mobiliteit [31]. VITO is een Europese onafhankelijke onderzoeksorganisatie. Het biedt objectief onderzoek, studies en adviezen op het gebied van *clean technology* en duurzame ontwikkeling [75]. Als pollutiedata wordt gebruik gemaakt van mobiele data (1-seconde data) verzameld met behulp van het instrument MicroAeth® (Sectie 2.2.3). Het gebruikte MicroAeth® model AE51 heeft een meetnauwkeurigheid van 0.1 $\frac{\mu g}{m^3}$ zwarte koolstof bij een debiet van 150 $\frac{mL}{minuut}$ en een temporele resolutie van 1 minuut [2]. Aangezien de nauwkeurigheden zijn gemeten onder ideale omstandigheden, zal men in de praktijk mogelijks metingen

bekomen met een lagere meetnauwkeurigheid. Gelet op meetfouten die frequent voorkomen bij het gebruik van het instrument MicroAeth® (Sectie 2.2.3) bekomt men tijdreeksen die een vrij groot aantal negatieve concentraties bevatten. Het *Optimized Noise-reduction Averaging* (ONA) algoritme corrigeert het voorkomen van deze negatieve concentraties zonder een sterke afbreuk te doen aan tijdsresolutie en dynamische trends [35]. Deze pollutiedata werden tot slot aangevuld met bijkomende meetdata verzameld in het kader van deze masterproef. Gezien het doel van deze masterproef enkel berust op het in kaart brengen van de spatiale variabiliteit van de component zwarte koolstof, werd bij het verzamelen van de mobiele meetdata een meetcampagne opgestart. Hierbij werd op verschillende tijdstippen overdag dezelfde route meermaals gereden. De meetcampagne van het GMF werd uitgevoerd in de periode van 26 september 2012 tot en met 12 november 2012. De bijkomende mobiele meetdata werden ingewonnen van 7 december 2015 tot en met 15 december 2015. Om de extreme waarden op het ruimtelijk geaggregeerd resultaat te verminderen, werd gebruik gemaakt van het α-getrimd gemiddelde met α gelijk aan 0.5%. Het α-getrimd gemiddelde verwijdert de α grootste en α kleinste waarden aan zwarte koolstof concentraties en berekent vervolgens het gemiddelde van de overgebleven waarden [72].

3.1.2 OpenStreetMap

OpenStreetMap (OSM) werd opgericht in 2004. Het is een internationaal project met als doel de wereld op een open manier in kaart te brengen. Vrijwilligers verzamelen gegevens over wegen, rivieren, bossen, spoorwegen, enz. met behulp van luchtfoto's, GPS-apparaten en low-tech veldkaarten. Deze gegevens worden vervolgens opgeslagen in een database. OSM is open data en dus vrij toegankelijk [53, 54]. In deze masterproef werd gebruik gemaakt van het wegennetwerk in OSM (*highway-key*) om zwarte koolstof concentraties te linken aan verschillende types wegen. Hasenfratz et al. [38] maakten eveneens gebruik van het wegennetwerk in OSM voor het opstellen van een UFP kaart met een hoge spatiale resolutie.

3.1.3 Urban Atlas

Urban Atlas voorziet data over het landgebruik en bodembedekking op basis van satellietbeelden. Het is een gemeenschappelijk initiatief van de Europese Commissie Directoraat-generaal Regionaal Beleid en Stadsontwikkeling en Directoraat-generaal voor Ondernemingen en Industrie in samenspraak met de Europese Ruimtevaartorganisatie en het Europees Milieuagentschap [26, 27]. Op basis van Urban Atlas kan het verschillend landgebruik gelinkt worden aan zwarte koolstof. Zo wordt verwacht dat in een stedelijk gebied, waar meer mobiele bronnen en huishoudelijke verwarming aanwezig zijn, de concentratie aan zwarte koolstof hoger is dan in een groengebied, waar minder bronnen van zwarte koolstof aanwezig zijn.

3.1.4 Centraal ReferentieAdressenBestand

Het Centraal ReferentieAdressenBestand (CRAB) bevat alle officiële adressen en hun corresponderende geografische positie. Deze authentieke bron voor adressen in Vlaanderen werd tot 1 juni 2011 beheerd door het Agentschap voor Geografische Informatie Vlaanderen (AGIV), maar nu, mits een overgangsfase van vier jaar, door de Vlaamse steden en gemeenten. Deze data zijn eveneens open data en vrij toegankelijk [3, 5]. CRAB is in deze masterproef nuttig voor het inschatten van zwarte koolstof concentraties afkomstig van huishoudelijke verwarming.

3.1.5 Verkeersmodel

Voor het verkeersmodel werd beroep gedaan op het verkeersmodel referentiejaar 2014 afkomstig van het Mobiliteitsbedrijf Stad Gent [71]. Dit heeft als output de verkeersintensiteit (i.e. het aantal voertuigen gepasseerd over een bepaalde afstand van het wegsegment gedurende een bepaalde tijdsperiode) weer tijdens de ochtend (8 – 9 uur)- en avondspits (17 – 18 uur). Het verkeersmodel maakt verder een onderscheid tussen het aantal vrachtwagens en het aantal personenwagens per uur op een bepaald wegsegment. Hierbij zijn de verkeersintensiteiten modelmatig gegenereerd en wordt verondersteld dat deze voorspellingen in grootte-orde overeenstemmen met de absolute verkeersintensiteiten. Uit Figuur 2.2 bleek dat verkeer een grote invloed heeft op de concentratie aan zwarte koolstof. Daarom is een model dat de verkeersintensiteit weergeeft per wegsegment een essentieel element bij het opstellen van de luchtvervuilingskaart.

In Vlaanderen maakt men gebruik van provinciale verkeersmodellen die worden opgesteld met behulp van het model BASMAT en het Multimodaal Model (MM). In BASMAT berekent men de herkomst-bestemmingsmatrices of verplaatsingsmatrices. Deze verplaatsingsmatrices worden bekomen door eerst voor een beschouwde tijdsperiode te berekenen hoeveel verplaatsingen er in iedere zone vertrekken en aankomen (tripgeneratie) en vervolgens deze globale verplaatsingen per zone te verdelen over alle herkomsten en bestemmingen (tripdistributie). Men berekent deze matrices per provincie en per motief (werk, school, winkel, recreatief/sociaal bezoek en overig). Deze resulterende verplaatsingsmatrices dienen als input van MM. In MM worden de verplaatsingsmatrices opgedeeld in verplaatsingsmatrices per vervoersmodus (auto, fiets, te voet of openbaar vervoer) en vindt een kalibratie plaats van auto en openbaar vervoer in functie van de beschikbare tellingen. Als laatste worden de resulterende verplaatsingsmatrices toegedeeld voor de vervoerswijzekeuzes auto en openbaar vervoer [50, 60].

3.1.6 Sky view factor

De *sky view factor* is de ratio van de hoeveelheid hemel zichtbaar vanaf een gegeven punt op het aardoppervlak tot datgene dat potentieel zou moeten aanwezig zijn (i.e. de proportie van het hemelhalffront ingesloten door een horizontaal oppervlak) [52]. De berekening van de *sky view factor* heeft vooreerst nood aan hoogtegegevens van Gent. Hiervoor wordt gebruik gemaakt van het Digitaal HoogteModel Vlaanderen II, DSM, raster, 1 m.

Hoogtemodel Vlaanderen

Voor het bepalen van de hoogtegegevens in Gent wordt gebruik gemaakt van het Digitaal HoogteModel Vlaanderen (DHMV). Dit is een verzamelnaam voor alle gebiedsdekkende hoogtegegevens van Vlaanderen waarover het AGIV beschikt. Hierbij wordt een onderscheid gemaakt tussen DHMV I en DHMV II. DHMV I werd reeds aangemaakt in de periode van 2001 tot 2004 en DHMV II vormt hiervan een actualisatie en wordt gezien als een eerste stap naar Vlaanderen in 3D. DHMV II werd aangemaakt in de periode van 2013 tot 2015 en wordt gradueel vrijgegeven. Van DHMV II bestaan vier standaard afgeleide producten die als gratis open data ter beschikking staan [6]:

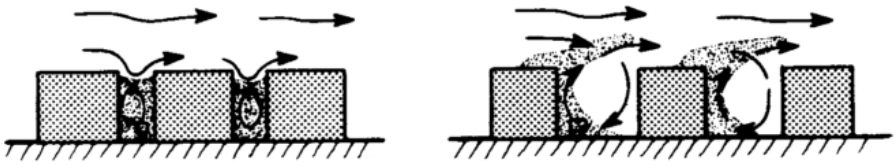
1. Digitaal terreinmodel (DTM) van het maaiveld in rasterformaat met een grondresolutie van 1 m.
2. Digitaal terreinmodel van het maaiveld in rasterformaat met een grondresolutie van 5 m.
3. Digitaal oppervlaktemodel (DSM) van het maaiveld inclusief objecten in rasterformaat met een grondresolutie van 1 m.
4. Digitaal oppervlaktemodel van het maaiveld inclusief objecten in rasterformaat met een grondresolutie van 5 m.

Deze producten zijn beschikbaar per 1/1 NGI-kaartblad; dit is een topografisch kaartblad van het Nationaal Geografisch Instituut op toepassingsschaal 1 : 50 000. De verschillende kaartbladen sluiten exact op elkaar aan en overlappen niet [6]. Hier werd geopteerd voor het derde standaard afgeleide product, namelijk DSM van het maaiveld inclusief objecten in rasterformaat met een grondresolutie van 1 m. Er werd gekozen voor DSM, omdat in dit geval naast hoogte-informatie van het terrein ook hoogte-informatie van objecten zoals gebouwen of vegetatie belangrijk zijn. Verder werd gekozen voor het model met de grootste resolutie met een pixelgrootte van 1 m [4]. DSM, nog steeds in ontwikkelingsfase, is een standaard afgeleid product van de brondata 'LiDAR Digitaal HoogteModel Vlaanderen II - ruwe remote sensing data'. De hoogtegegevens werden ingewonnen met behulp van LiDAR-technologie vanuit een vliegtuig. Het coördinatenreferentiesysteem is Belge 1972/ Belgian

Lambert 72, EPSG: 31370 en de hoogtewaarden, uitgedrukt in meter met centimeterprecisie, zijn gerefereerd aan de Tweede Algemene Waterpassing (TAW) [6].

Gebruik van sky view factor

De atmosfeer kan polluenten transporteren, verdunnen, transformeren en verwijderen. Na vrijstelling van de polluenten wordt de dispersie ervan gecontroleerd door atmosferische bewegingen (wind en turbulentie). Wind en turbulentie zijn dus cruciaal bij de dispersie van de polluenten in de atmosfeer. De uitwisseling tussen straatniveau, waar de polluenten worden geëmitteerd, en hoger dan dakniveau, zal afhangen van de ratio gemiddelde hoogte van de gebouwen en breedte van de straat. In nauwere straten (*street canyons*) is de atmosferische uitwisseling gelimiteerd in vergelijking met een meer open gebied, waar de vortexcirculatie de atmosferische uitwisseling tussen straatniveau en hoger niveau bevordert (Figuur 3.1). Is de ventilatie zwak, dan kunnen de polluenten zich daar gaan ophopen [52]. Nauwere straten hebben een lagere *sky view factor* dan bredere straten, waardoor de sky view factor kan gelinkt worden aan de atmosferische uitwisseling van polluenten.



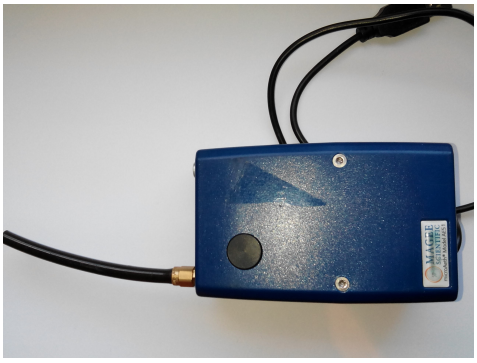
Figuur 3.1: Invloed van gebouwen op de dispersie van polluenten. Links: Nauwere straat waarbij de dispersie van polluenten beperkt is. Rechts: Bredere straat waarbij de dispersie van polluenten reeds groter is [52].

3.2 Verwerking van de verzamelde data

3.2.1 Pollutiedata

De pollutiedata verzameld door het GMF werden aangevuld met bijkomende meetdata verzameld in het kader van deze masterproef. De bijkomende meetdata betreft geen nieuwe meetcampagne, gezien dezelfde route werd gereden. De bijkomende meetdata zijn eveneens mobiele data bekomen met behulp van het instrument MicroAeth® model AE51 (Magee Scientific) (Figuren 3.2 en 3.3) en een GPS-toestel. Bij gebrek aan meer professionele GPS-toestellen werd gebruik gemaakt van een smartphone-app. Zwarte koolstof werd gemeten met een temporele resolutie van 1 seconde met een debiet van 150 $\frac{\text{mL}}{\text{minuut}}$. Op de ontvangen pollutiedata werd reeds naverwerking toegepast. Op de bijkomende meetdata werd het ONA algoritme toegepast met een attenuatiethreshold van 0.05 met behulp van MATLAB [68], waardoor de negatieve concentraties sterk reduceerden in voorkomen van 19.91% tot 0.69%.

Tot slot werden de bijkomende meetdata geanalyseerd op basis van de geografische locatie van het meetpunt. Meetpunten, voornamelijk deze bij de start van een nieuwe route, vertoonden afwijkende geografische locaties overeenkomstig met de gereden route. Deze meetpunten werden verwijderd uit de bijkomende meetdata, evenals meetdata waarvan de geografische locatie ontbrak.



Figuur 3.2: MicroAeth® model AE51 (Magee Scientific).

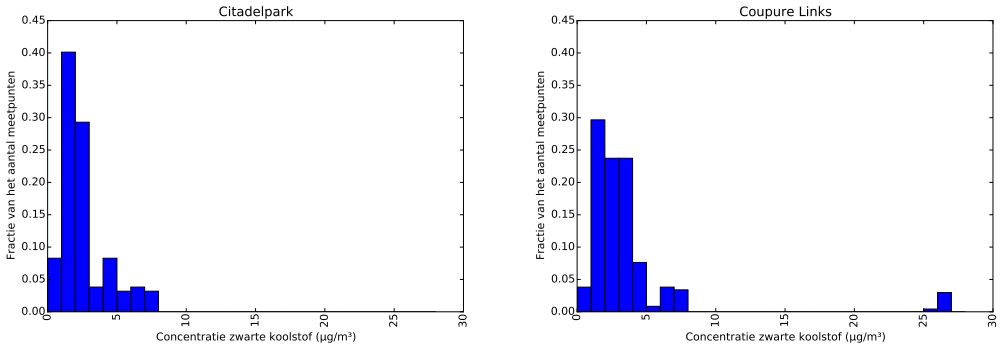
Figuur 3.3: Filter afkomstig van MicroAeth® model AE51 (Magee Scientific). Links: Nieuwe filter. Rechts: Gesatureerde filter. Zwarte vlek wijst op depositie van zwarte koolstof.

Vervolgens werden deze twee datasets (mobiele meetdata ontvangen van het GMF en bijkomende mobiele meetdata) samengevoegd. Daarnaast werd de gereden route geëxtraheerd als een GIS-laag uit OSM. Op basis van deze laag werden 370 *points of interest* (POIs) gekozen, die gelijkmatig verspreid lagen (bij benadering 50 m van elkaar verwijderd) op de gereden route. Op basis van deze POIs was het mogelijk om de meetpunten met bijhorende zwarte koolstof concentraties uit de samengevoegde dataset te aggregeren. Een eerste stap in deze aggregatie was elk meetpunt toekennen aan een POI. Het criterium om een meetpunt toe te kennen aan een POI is gebaseerd op de afstand van het meetpunt tot het POI. Meetpunten werden telkens toegekend aan het dichtstbijzijnde POI. Zo bekomt men een nieuwe dataset

met een resolutie van 50 m. Uiteindelijk werd één zwarte koolstof concentratie bekomen per POI door het α-getrimd gemiddelde (α gelijk aan 0.5%) te berekenen van de verschillende gemeten zwarte koolstof concentraties toegewezen aan dit POI. Deze werkmethode is gebaseerd op een studie uitgevoerd door Van den Bossche et al. in Antwerpen [72]. In de volgende paragraaf wordt de distributie van de geobserveerde concentraties op vier locaties besproken.

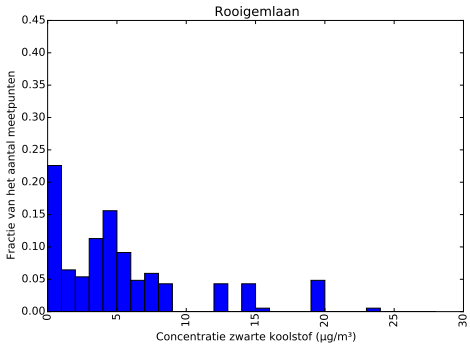
Figuur 3.4 geeft de histogrammen van geobserveerde concentraties op vier verschillende locaties (POIs) weer. Deze locaties duiden op één punt in de desbetreffende straat of park en dus niet op alle punten gelegen in die straat of park. Uit de vier histogrammen valt op te merken dat luchtkwaliteitsdata niet symmetrisch verdeeld zijn, zoals eerder besproken in Sectie 1.4.2. In Figuur 3.4 (a) was de maximale gemeten concentratie 8 $\frac{\mu g}{m^3}$, wat minder dan de helft is in vergelijking met de drie andere histogrammen. De lagere geobserveerde concentraties in het Citadelpark kunnen het gevolg zijn van minder aanwezige bronnen van zwarte koolstof, zoals het ontbreken van gemotoriseerd verkeer. Verder is enkel in Figuur 3.4 (d) geen concentratie gemeten tussen 0 $\frac{\mu g}{m^3}$ en 1 $\frac{\mu g}{m^3}$. Een mogelijke oorzaak kan de lagere *sky view factor* zijn, waardoor zwarte koolstof minder sterk gedispergeerd wordt in de atmosfeer en/of de afwezigheid van een gescheiden fietspad zijn, waardoor de fietser zich dichter bij verschillende verkeersvormen bevindt. Opvallend in Figuur 3.4 (b) zijn de hogere concentraties boven 25 $\frac{\mu g}{m^3}$. Deze concentraties kunnen afkomstig zijn van occassionele gebeurtenissen, zoals het rijden achter een brommer op het fietspad.

Het resultaat van de aggregatie wordt weergegeven in Figuur 3.5, waarbij de kleurencode wijst op de concentratie aan zwarte koolstof in $\frac{\mu g}{m^3}$. Gezien er op dit moment nog geen emissiegrenswaarde voor zwarte koolstof is (Sectie 2.1), werd de kleurenschaal gebaseerd op de kleurenschaal van airQmap (VITO NV [76]). AirQmap brengt de blootstelling aan zwarte koolstof voor fietsers en voetgangers in kaart. Uit het resultaat valt op te merken dat zwarte koolstof spatiaal variabel is en dat er een verband is met de verkeersintensiteit en met de topologie van straten (conform *sky view factor*). Lage zwarte koolstof concentraties werden teruggevonden in het Citadelpark en in de Langemunt, waar de verkeersintensiteit voor voertuigen nagenoeg nul is. Hoge zwarte koolstof concentraties werden teruggevonden in de Charles de Kerchovelaan, Godshuizenlaan en Sint-Lievenslaan, die getypeerd worden door een hoge verkeersintensiteit. Hoewel de verkeersintensiteit in de Sint-Jacobsnieuwstraat circa 50% lager is relatief ten opzichte van de laatst genoemde straten volgens het verkeersmodel observeert men daar eveneens een hoge zwarte koolstof concentratie. Deze hogere concentratie kan het gevolg zijn van de lagere *sky view factor* in vergelijking met de laatst genoemde straten en/of geen aanwezigheid van een fietspad, zoals reeds besproken in voorgaande paragraaf. Deze bevindingen zijn in overeenstemming met een studie uitgevoerd door Van den Bossche et al. in Antwerpen [72].

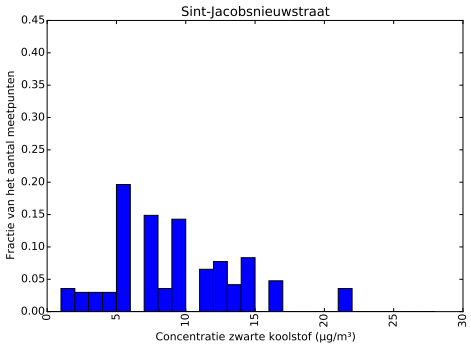


(a) Citadelpark

(b) Coupure Links

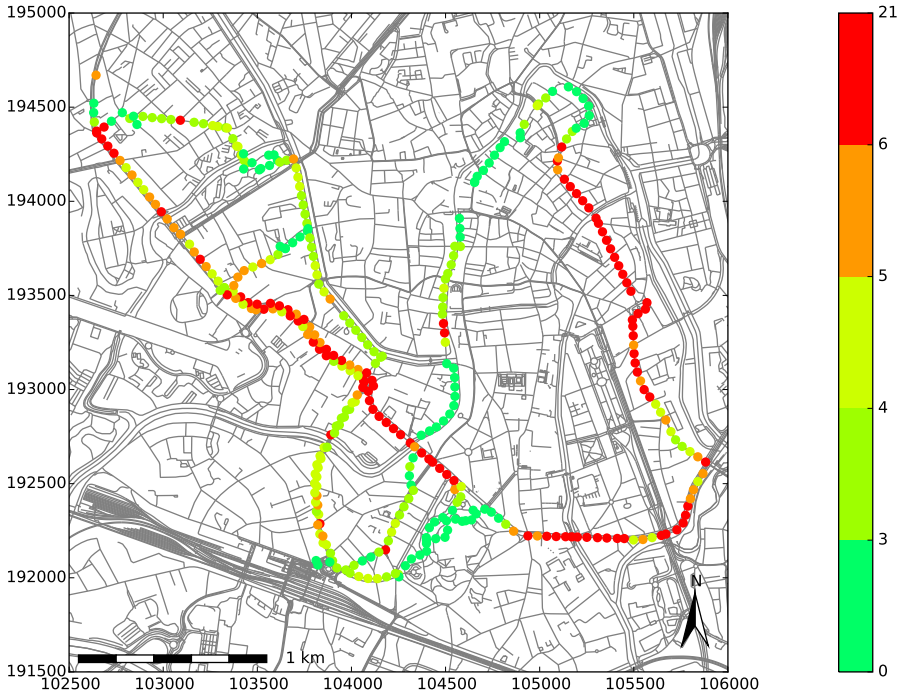


(c) Rooigemlaan



(d) Sint-Jacobsnieuwstraat

Figuur 3.4: Histogram van distributie van geobserveerde concentraties op de locaties (a) Citadelpark, (b) Coupure Links, (c) Rooigemlaan en (d) Sint-Jacobsnieuwstraat.



Figuur 3.5: Aggregatie van mobiele meetdata ontvangen van het GMF en bijkomende meetdata verzameld in het kader van deze masterproef. De kleurencode weerspiegelt de concentratie aan zwarte koolstof in $\frac{\mu g}{m^3}$.

3.2.2 OpenStreetMap

In deze masterproef werd beroep gedaan op de *highway-key* van OSM. Hiervoor werd OSM ingelezen in QGIS en werd vervolgens gezocht naar de regio Gent. Aangezien hier gebruik wordt gemaakt van het wegennetwerk, wordt het objecttype 'polygonen (*open ways*)' geselecteerd. De gekozen geëxporteerde tags waren '*highway*' en '*name*'. Tot slot werd het studiegebied uitgesneden en werd Belge 1972/ Belgian Lambert 72, EPSG: 31370 gebruikt als coördinatenreferentiesysteem. Het uiteindelijke resultaat leverde een onderscheid op van de verschillende types wegen (*primary, secondary, tertiary, motorway*, enz.)[1]. Deze kaart kan gebruikt worden om de dichtstbijzijnde types straten te linken aan een bepaalde zwarte koolstof concentratie.

[1] `http://wiki.openstreetmap.org/wiki/Map_Features`

3.2.3 Urban Atlas

Het gebruikte downloadpakket omvatte enkel de regio Gent [26]. Uit deze regio werd het studiegebied uitgeknipt. Voorts diende het coördinatenreferentiesysteem nog omgezet te worden naar Belge 1972/ Belgian Lambert 72, EPSG: 31370. De gegevens dateren van 2009 en zijn afkomstig van *Earth Observation data*, topografische kaarten en *Commercial Off-The-Shelf* (COTS) navigatiedata. De toepassingsschaal is 1 : 10 000 en de pixelgrootte is ongeveer 5 m. De GIS-laag maakt een onderscheid tussen de stedelijke weefsels, industriële, commerciële, militaire, publieke en private eenheden, havens, bossen, enz. [27]. Het verschillend landgebruik en bodembedekking kan vervolgens gelinkt worden aan de concentratie aan zwarte koolstof.

3.2.4 Centraal ReferentieAdressenBestand

Voor gegevens over de huisnummers in Gent en hun positie werd gebruik gemaakt van CRAB adresposities. Dit downloadpakket bevat een lijst met 2.7 miljoen huisnummers en 600 000 bus- en appartementnummers in Vlaanderen met hun positie op de kaart. De posities van kadastrale en administratieve percelen en gebouwen waaraan een huisnummer is gekoppeld worden weergegeven als centroïdes, waarbij de positie het middelpunt van het terreinobject aanduidt waarnaar het verwijst. De toepassingsschaal is 1 : 50 en 1 : 1 000 000. In CRAB worden de historische gegevens bijgehouden, waarbij de geldigheidsperiode van het ontstaan en eventuele afschaffing van een object wordt weergegeven door de velden 'BEGINDATUM' en 'EINDDATUM' [8]. Uit de beschikbare GIS-laag werd het studiegebied geselecteerd en uitgesneden. Gezien enkel de niet-afgeschafte objecten hier van belang zijn, worden enkel die gegevens geselecteerd waarvan de einddatum nog niet verstreken is. Tevens werd Belge 1972/ Belgian Lambert 72, EPSG: 31370 gebruikt als coördinatenreferentiesysteem.

3.2.5 Verkeersmodel

Uit het verkeersmodel werd eveneens het studiegebied geselecteerd en gebruik gemaakt van het coördinatenreferentiesysteem Belge 1972/ Belgian Lambert 72, EPSG: 31370. Dit werd zowel toegepast op het model dat de ochtendspits simuleert als het model dat de avondspits simuleert.

3.2.6 Sky view factor

Vooraleer men de *sky view factor* kan berekenen voor het studiegebied, werd eerst beroep gedaan op het DHMV II, DSM, raster, 1 m. Nadat hieruit het studiegebied werd geselecteerd, meerbepaald uit het kaartblad nr. 22 [7], werd de *sky view factor* kaart berekend. Deze kaart werd bekomen met QGIS [62], waarbij gebruik werd gemaakt van acht sectoren of kijkrichtingen.

HOOFDSTUK 4

Feature-extractie

In een volgende stap is het noodzakelijk om features te extraheren uit de verzamelde GIS-lagen besproken in het voorgaande hoofdstuk. Bij het extraheren wordt rekening gehouden met de mogelijke correlatie tussen de verschillende omgevingsvariabelen en zwarte koolstof. In de volgende secties wordt besproken hoe deze features geëxtraheerd werden, gevolgd door een overzicht van de verschillende geëxtraheerde features. In het voorgaande hoofdstuk bepaalde men op de POIs de zwarte koolstof concentraties. In dit hoofdstuk zijn POIs deze geografische posities ($> 24\,000$) waarop men de waarden bepaalt van een aantal omgevingsvariabelen. Deze waarden moeten bepaald worden om LUR modellen te kunnen opbouwen en voorspellingen te kunnen maken. Gelet op de complexiteit en uitgebreidheid van de data werden berekeningen niet uitgevoerd met een traditionele ruimtelijke query, maar werd geopteerd om zelf broncode te schrijven. Deze broncode werd geschreven in Python [61]. Hierbij werd gebruik gemaakt van de bibliotheken 'geopandas', 'matplotlib', 'numpy', 'pandas' en 'shapely'. De *sky view factor* werd wel berekend in QGIS [62].

4.1 Aantal puntbronnen in nabijheid van het POI

Uit Sectie 2.1 blijkt dat zwarte koolstof afkomstig kan zijn van puntbronnen, zoals huishoudelijke verwarming. De CRAB dataset (Sectie 3.1.4) bevat alle adressen en hun corresponderende geografische positie. Huizen kunnen beschouwd worden als puntbronnen van zwarte koolstof. Om het aantal huizen in een bepaalde buffer (i.e. een cirkelschijf met als middelpunt het POI) te berekenen, bepaalt men de afstand tussen het POI en het aantal huizen. Het aantal huizen gelegen in de buffer wordt bekomen door het aantal huizen te tellen gelegen in de buffer. De buffergroottes zijn gebaseerd op het ESCAPE project. In dat project werd voor het aantal huizen gebruik gemaakt van buffers met een straal van 100 m, 300 m, 500 m, 1 000 m en 5 000 m [9, 23]. Gezien het bereik van de mobiele data in deze masterproef werd geopteerd geen features te berekenen waarvan de straal van de buffer groter is dan 500 m. Op

die manier bekomt men volgende features: het aantal huizen gelegen in een buffer met straal 100 m, 300 m en 500 m.

4.2 Afstand tot dichtstbijzijnde lijnbronnen

Naast puntbronnen zijn ook lijnbronnen, zoals wegen, een belangrijke bron van zwarte koolstof (Sectie 2.1). Daarom werd de afstand tot de dichtstbijzijnde types wegen ten opzichte van de verschillende POIs geëxtraheerd uit OSM: afstand tot dichtstbijzijnde snelweg, rijksweg, primaire weg, secundaire weg, tertiaire weg, dienstweg, residentiële weg, leefstraat, voetgangersstraat en pad.

4.3 Afstand tot dichtstbijzijnde kruispunt

Gezien stationair draaiende en optrekkende voertuigen nabij een kruispunt eveneens een bron kunnen zijn van zwarte koolstof, werd de afstand tot het dichtstbijzijnde kruispunt ten opzichte van de verschillende POIs bepaald. De GIS-data die hiervoor werd gebruikt is afkomstig uit OSM.

4.4 Afstand tot dichtstbijzijnde park

Men kan verwachten dat in een park, waar minder bronnen van zwarte koolstof aanwezig zijn, de concentratie aan zwarte koolstof ook lager is dan in een stedelijk gebied, waar meer punten lijnbronnen aanwezig zijn. Vanuit dit opzicht werd ook de afstand tot het dichtstbijzijnde park/stedelijk groengebied ten opzichte van de verschillende POIs bepaald. De GIS-data die hiervoor werd gebruikt is afkomstig uit Urban Atlas.

4.5 Oppervlaktebronnen in nabijheid van het POI

Zoals vermeld in voorgaande sectie kan men verwachten dat in stedelijke gebieden de concentratie aan zwarte koolstof hoger is dan in groengebieden. Om die reden werden ook de volgende features geëxtraheerd uit Urban Atlas: oppervlakte aan park, aan hoogstedelijk en laagstedelijk gebied gelegen in een buffer met straal 100 m, 300 m en 500 m. De keuze van buffergroottes is analoog als in Sectie 4.1. Voor hoogstedelijk, respectievelijk laagstedelijk gebied, werden alle stedelijke gebieden met een densiteit van meer dan 50%, respectievelijk minder dan 50% samengenomen. De pseudocode voor het bepalen van deze oppervlaktebronnen in een vastgelegde buffer is weergegeven in Tabel 4.1. Een eerste stap is het specifiëren van de buffer aan de hand van zijn straal. Na het nemen van de doorsnede van de buffer

met de oppervlaktebronnen kan men als resultaat de totale oppervlakte verkrijgen door te sommeren over de oppervlaktebronnen gelegen in de buffer.

Tabel 4.1: Pseudocode: Oppervlaktebronnen binnen vastgelegde buffer.

Input: Coördinaten van het POI, straal, coördinaten van de oppervlaktebronnen.
Output: Oppervlaktebronnen binnen vastgelegde buffer.

1. Bepalen van buffer met behulp van opgegeven straal rond het POI.
2. Nemen van doorsnede buffer met de oppervlaktebronnen.
3. Bepalen en sommeren van de oppervlaktebronnen binnen vastgelegde buffer.

4.6 Verkeersintensiteit in nabijheid van het POI

De verkeersintensiteit in de nabijheid van het POI is een andere bron van zwarte koolstof. Het berekenen van features omtrent de verkeersintensiteit gebeurt vrij analoog aan Sectie 4.5 (pseudocode: zie Tabel 4.2). Het grote verschil is dat hier een doorsnede wordt genomen met wegen en niet met oppervlaktebronnen en tevens dat de lengte van de wegen wordt vermenigvuldigd met de bijhorende verkeersintensiteit van die weg gelegen in de buffer. De reden dat niet enkel rekening wordt gehouden met de verkeersintensiteit gelegen in de buffer, maar ook met de lengte van de wegen, is om te voorkomen dat een kortere weg met een hoge verkeersintensiteit zwaarder zou gaan doorwegen. Deze methode voor het berekenen van verkeersintensiteit is analoog aan de methode toegepast in het ESCAPE project [9, 23]. De bekomen features zijn: aantal voertuigen, aantal vrachtwagens en aantal personenwagens voorkomend in een buffer met straal 25 m, 50 m, 100 m, 300 m en 500 m. De keuze van buffergroottes is opnieuw gebaseerd op het ESCAPE project (Sectie 4.1) [9, 23]. Bij het berekenen van deze features werd enkel gebruik gemaakt van het ochtendmodel van het verkeersmodel. Dit gezien er in het kader van deze masterproef geen rekening wordt gehouden met de temporele component, omdat daarvoor te weinig data voorhanden is en waarbij werd aangenomen dat het ochtendmodel en het avondmodel niet sterk van elkaar verschillen.

4.7 Sky view factor op het POI

Om de *sky view factor* te kunnen berekenen op het POI werd eerst een GIS-laag aangemaakt die louter de geografische positie van het POI omvatte. Op die manier kon de *sky view factor*

Tabel 4.2: Pseudocode: Verkeersintensiteit binnen vastgelegde buffer.

Input: Coördinaten van het POI, straal, coördinaten van de wegen met bijhorende verkeersintensiteit.
Output: Verkeersintensiteit binnen vastgelegde buffer.

1. Bepalen van buffer met behulp van opgegeven straal rond het POI.
2. Nemen van doorsnede buffer met wegen.
3. Vermenigvuldigen van lengte wegen gelegen in de buffer met de bijhorende verkeersintensiteit van de weg.
4. Sommeren over het product berekend in Stap 3.

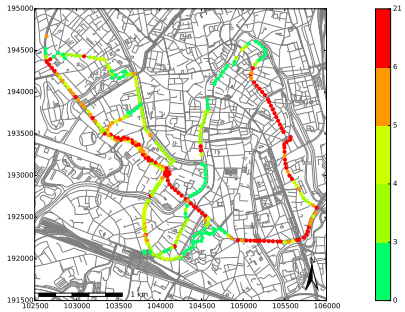
op die geografische posities bekomen worden door gebruik te maken van de *Point sampling tool* in QGIS die als input de voorgenoemde GIS-laag en de *sky view factor* kaart vereist.

4.8 Bespreking features

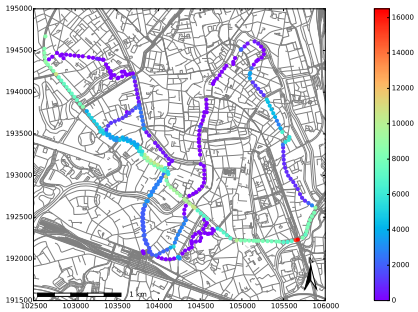
Een overzicht van de verschillende berekende features met gebruikte databron en methode is terug te vinden in Tabel 4.3. Verder wordt in deze tabel de determinatiecoëfficiënt R^2 weergegeven van de verschillende features met zwarte koolstof. Hieruit blijkt dat verkeersgerelateerde features (aantal voertuigen, aantal vrachtwagens en aantal personenwagens) en de features 'afstand tot dichtstbijzijnde tertiaire weg' en 'afstand tot het dichtstbijzijnde kruispunt' een relatief hogere correlatie vertonen met zwarte koolstof in vergelijking met de overige features. In Figuur 4.1 worden twee features ('aantal vrachtwagens in een buffer met straal 25 m' en '*sky view factor*') in meer detail weergegeven. Bij het feature 'aantal vrachtwagens in een buffer met straal 25 m' is de correlatie duidelijk waarneembaar: op de R40 observeert men een hoog aantal vrachtwagens en hogere zwarte koolstof concentraties en in het Citadelpark, waar het aantal vrachtwagens nul is, observeert men lage zwarte koolstof concentraties. Bij het feature '*sky view factor*' is dit verband minder waarneembaar, waardoor de voorspellende kracht waarschijnlijk niet zo groot is. Toch kan het zijn dat dit feature in een model nog belangrijk wordt door potentiële interactie-effecten.

Tabel 4.3: Overzicht geëxtraheerde features met gebruikte databron, methode van extractie en keuze van straal voor buffer. In de laatste kolom wordt de correlatie van het feature met zwarte koolstof weergegeven. De afstanden zijn telkens tot het dichtstbijzijnde type weg, kruispunt of park. Opp staat voor oppervlakte.

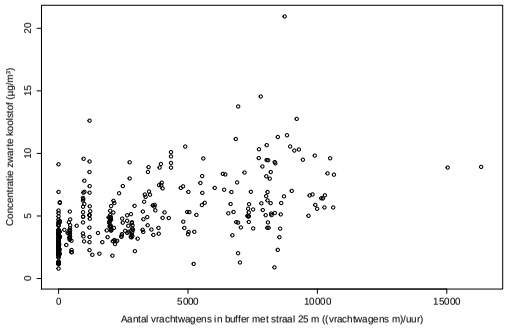
Feature	Databron	Methode extractie	Straal buffer (m)	Determinatiecoëfficiënt (%)
`Aantal huizen`	CRAB	Sectie 4.1	100, 300, 500	0.37, 0.98, 2.29
`Afstand snelweg`	OSM	Sectie 4.2	Niet toepasbaar	1.69
`Afstand rijksweg`	OSM	Sectie 4.2	Niet toepasbaar	5.56
`Afstand primaire weg`	OSM	Sectie 4.2	Niet toepasbaar	9.35
`Afstand secundaire weg`	OSM	Sectie 4.2	Niet toepasbaar	0.48
`Afstand tertiaire weg`	OSM	Sectie 4.2	Niet toepasbaar	16.44
`Afstand dienstweg`	OSM	Sectie 4.2	Niet toepasbaar	0.24
`Afstand residentiële weg`	OSM	Sectie 4.2	Niet toepasbaar	0.61
`Afstand leefstraat`	OSM	Sectie 4.2	Niet toepasbaar	0.86
`Afstand voetgangersstraat`	OSM	Sectie 4.2	Niet toepasbaar	4.58
`Afstand pad`	OSM	Sectie 4.2	Niet toepasbaar	0.21
`Afstand kruispunt`	OSM	Sectie 4.3	Niet toepasbaar	15.80
`Afstand park`	Urban Atlas	Sectie 4.4	Niet toepasbaar	2.07
`Opp park`	Urban Atlas	Sectie 4.5	100, 300, 500	10.94, 9.29, 7.43
`Opp hoogstedelijk gebied`	Urban Atlas	Sectie 4.5	100, 300, 500	0.30, 1.42, 4.63
`Opp laagstedelijk gebied`	Urban Atlas	Sectie 4.5	100, 300, 500	3.79, 2.07, 1.76
`Aantal voertuigen`	Ochtendverkeersmodel	Sectie 4.6	25, 50, 100, 300, 500	33.94, 33.14, 27.89, 13.94, 6.80
`Aantal vrachtwagens`	Ochtendverkeersmodel	Sectie 4.6	25, 50, 100, 300, 500	34.91, 34.39, 32.11, 20.55, 11.18
`Aantal personenwagens`	Ochtendverkeersmodel	Sectie 4.6	25, 50, 100, 300, 500	33.49, 32.70, 27.34, 13.46, 6.55
`Sky view factor`	Sky view factor	Sectie 4.7	Niet toepasbaar	2.82



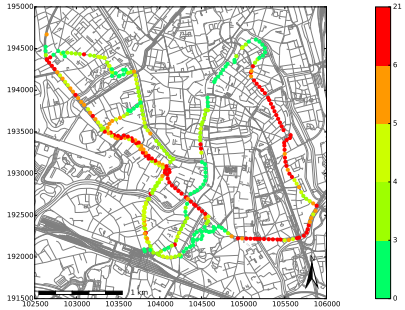
(a) Kleurencode: Concentratie zwarte koolstof ($\frac{\mu g}{m^3}$).



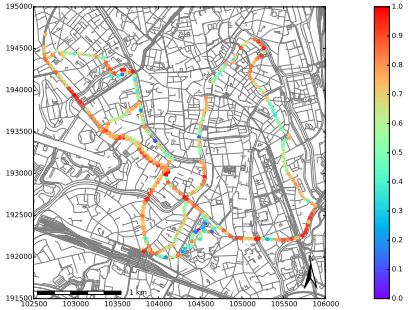
(b) Kleurencode: Aantal vrachtwagens in een straal van 25 m ($\frac{\text{vrachtwagens}}{\text{uur}}$ m).



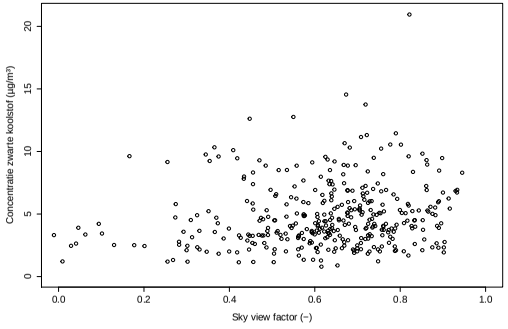
(c)



(d) Kleurencode: Concentratie zwarte koolstof ($\frac{\mu g}{m^3}$).



(e) Kleurencode: *Sky view factor* (-).



(f)

Figuur 4.1: (a) Aggregatie van mobiele meetdata. (b) Aantal vrachtwagens in een buffer met straal 25 m. (c) Scatterplot van concentratie aan zwarte koolstof in functie van het aantal vrachtwagens in een buffer met straal 25 m. (d) Aggregatie van mobiele meetdata. (e) *Sky view factor*. (f) Scatterplot van concentratie aan zwarte koolstof in functie van de *sky view factor*.

HOOFDSTUK 5

Opbouw regressiemodellen

In dit hoofdstuk worden verschillende regressietechnieken beschreven, toegepast en wordt hun performantie geëvalueerd. Er werd een selectie gemaakt van regressietechnieken die courant gebruikt worden in hedendaagse voorspellingsmodellen [47]. De implementaties beschikbaar in het statistisch softwarepakket R [63] werden gebruikt. Vervolgens werd de zwarte koolstof kaart opgesteld met behulp van de regressietechniek lasso en geëvalueerd. Tot slot worden enkele kritische bemerkingen toegelicht.

5.1 Regressietechnieken

In deze sectie worden de verschillende toegepaste regressietechnieken toegelicht, namelijk lineaire regressie, *forward* en *backward stepwise selection*, ridge regressie, lasso, *support vector* regressie, regressiebomen, *random forests* en *K-Nearest Neighbors* regressie.

5.1.1 Lineaire regressie

Het meervoudige lineaire regressiemodel met p verschillende features wordt voorgesteld door volgend statistisch model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon, \tag{5.1}$$

met X_j de j^{de} feature, β_j de regressiecoëfficiënt bij de variabele X_j en ϵ een foutterm die onafhankelijk is van X_j en waarvan de verwachtingswaarde nul is. β_j kan beschouwd worden als het gemiddelde effect op Y door toename van X_j met één eenheid, waarbij alle overige features constant worden gehouden. Lineaire regressie is een parametrische benadering, omdat het aanneemt dat $\mathrm{E}(Y)$ een lineaire functie is van X_j.

Het fitten van het meervoudige lineaire regressiemodel komt neer op het schatten van de regressiecoëfficiënten β_0, β_1, ..., β_p. Deze regressiecoëfficiënten worden geschat met behulp van de *least squares* benadering. Hierbij wordt geopteerd voor een minimalisering van de *sum of squared residuals* (RSS):

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2, \tag{5.2}$$

waarbij y_i en $\hat{y}_i$ de werkelijke respons, respectievelijk de voorspelling weergeeft voor de i^{de} observatie. De waarden β_0, β_1, ..., β_p die Vgl. (5.2) minimaliseren zijn de meervoudige *least squares* regressiecoëfficiëntschattingen.

5.1.2 Forward stepwise selection

Lineaire regressie heeft een aantal beperkingen. Wanneer er een lineair verband is tussen X_j en Y en het aantal observaties veel groter is dan het aantal features, dan zullen de meervoudige *least squares* regressiecoëfficiëntschattingen heel stabiel zijn (lage variantie) en leiden tot een regressiemodel met een hoge performantie. Is echter het aantal observaties niet veel groter dan het aantal features, dan neemt de variabiliteit toe, wat kan resulteren in overfitting van de data en een lagere performantie van het uiteindelijke regressiemodel. Wanneer het aantal features groter is dan het aantal observaties bestaat er bovendien geen unieke kleinste kwadratenschatter. Het reduceren van het aantal features kan in dergelijke gevallen een positief effect hebben op de performantie van het regressiemodel, doordat de variantie van de schatters gereduceerd wordt en maar een beperkte stijging in bias (i.e. de systematische afwijking van een model ten opzichte van de werkelijkheid) tot gevolg heeft. Daarnaast kan het reduceren van het aantal features ook leiden tot een model dat minder complex en beter interpreteerbaar is. Vanuit dit opzicht wordt geopteerd voor *forward stepwise selection*. Bij *forward stepwise selection* start men met een nulmodel dat geen features bevat. Vervolgens voegt men één feature toe per stap tot alle features aanwezig zijn in het model. Zo wordt bij elke stap een feature toegevoegd die leidt tot de beste verbetering van het model. Het algoritme van *forward stepwise selection* is weergegeven in Tabel 5.1. Om dit regressiemodel op te stellen werd beroep gedaan op het pakket 'leaps' in R [69].

5.1.3 Backward stepwise selection

In tegenstelling tot *forward stepwise selection* begint *backward stepwise selection* met het volledige model dat alle features bevat en verwijdert het één feature per stap. Het algoritme van *backward stepwise selection* is weergegeven in Tabel 5.2. Om dit regressiemodel op te stellen werd net als bij *forward stepwise selection* beroep gedaan op het pakket 'leaps' in R [69].

Tabel 5.1: Algoritme: *Forward stepwise selection.*

1. M_0 is het nulmodel dat geen features bevat.
2. Voor $k = 0, \ldots, p-1$:
   a Beschouw alle $p-k$ modellen die de features in M_k verhogen met één extra feature.
   b Kies het beste model tussen de voorgaande $p-k$ modellen en noem het M_{k+1}. Het beste model is hier het model met de laagste RSS of de hoogste R^2.
3. Het enige beste model wordt gekozen tussen $M_0, \ldots, M_p$ door gebruik te maken van crossvalidatie. Het beste model is hier het model met de laagste gemiddelde kwadratische fout.

Tabel 5.2: Algoritme: *Backward stepwise selection.*

1. M_p is het volledige model dat alle features bevat.
2. Voor $k = p, p-1, \ldots, 1$:
   a Beschouw alle k modellen die alle features behalve één feature van M_k bevatten voor een totaal van $k-1$ features.
   b Kies het beste model tussen de voorgaande k modellen en noem het M_{k-1}. Het beste model is hier het model met de laagste RSS of de hoogste R^2.
3. Het enige beste model wordt gekozen tussen $M_0, \ldots, M_p$ door gebruik te maken van crossvalidatie. Het beste model is hier het model met de laagste gemiddelde kwadratische fout.

5.1.4 Ridge regressie

Naast subset selectietechnieken zoals *forward* en *backward stepwise selection* kan men ook gebruik maken van regularisatie (*shrinkage*). Bij deze techniek worden alle p features gebruikt om een model te fitten, maar worden de coëfficiëntschattingen door middel van een *shrinkage penalty* beperkt in grootte. Deze regularisatie kan de variantie reduceren. Een regressietechniek die gebruik maakt van regularisatie is ridge regressie. Ridge regressie is gelijkaardig aan *least squares* (Vgl. (5.2)), behalve dat bij ridge regressie een tweede term is toegevoegd, namelijk een *shrinkage penalty*:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = RSS + \lambda\sum_{j=1}^{p}\beta_j^2, \tag{5.3}$$

met λ een tuningsparameter. De eerste term zoekt naar coëfficiëntschattingen die de data goed benaderen door te zoeken naar een kleine RSS, terwijl de tweede term klein is als de coëfficiëntschattingen nul naderen. Op die manier zorgt de tweede term voor een krimpend effect van de coëfficiëntschattingen naar nul toe. De mate van relatieve impact van de twee termen op de coëfficiëntschattingen wordt bepaald door λ. Bij $\lambda = 0$ heeft de *shrinkage penalty* geen effect en levert ridge regressie niets anders op dan de coëfficiëntschattingen bekomen met *least squares*. Wanneer $\lambda \to \infty$ zal de impact van de *shrinkage penalty* groter worden, waardoor de coëfficiëntschattingen nul benaderen. Het opstellen van dit statistisch model werd uitgevoerd in R met behulp van het pakket 'glmnet' [29].

5.1.5 Lasso

Een andere regressietechniek die gebruik maakt van regularisatie is lasso [70]. Het voordeel van lasso ten opzichte van ridge regressie is dat hier sommige coëfficiëntschattingen effectief nul kunnen worden. Hierdoor kan lasso eveneens aan feature-selectie doen, wat ervoor zorgt dat het model minder complex wordt en beter te interpreteren is. Het verschil tussen ridge regressie en lasso is dat de term β_j^2 in ridge regressie is vervangen door $|\beta_j|$ in lasso:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j|. \tag{5.4}$$

Door het veranderen van deze term worden sommige coëfficiëntschattingen effectief nul wanneer λ voldoende groot wordt. Voor het opstellen van dit statistisch model werd ook beroep gedaan op het pakket 'glmnet' in R [29].

5.1.6 Support vector regressie

Bij *support vector* regressie (SVR) maakt men gebruik van een 'ϵ-*insensitive*' verliesfunctie $V_\epsilon(r)$ om de regressiecoëfficiënten te schatten [40, 73]:

$$H = C\sum_{i=1}^{n} V_\epsilon(y_i - f(x_i)) + \frac{\beta_j^2}{2}, \tag{5.5}$$

met

$$V_\epsilon(r) = \begin{cases} 0 & \text{, als } |r| < \epsilon, \\ |r| - \epsilon & \text{, anders.} \end{cases} \tag{5.6}$$

Deze manier om regressiecoëfficiënten te schatten is verschillend van de voorgaande regressietechnieken die gebruik maken van een kwadratische verliesfunctie. Het voordeel van deze benadering is dat het model robuuster wordt, doordat outliers een minder sterke invloed en kleine fouten geen invloed hebben op de modelperformantie. De parameter C is een regularisatieparameter, met een functie gelijkaardig aan λ in ridge regressie en lasso, die de *trade-off* bepaalt tussen de complexiteit van het model en zijn fit aan de data. Er kan aangetoond worden dat de lineaire functie die H minimaliseert kan geschreven worden als volgt:

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i K(x_i, x), \tag{5.7}$$

met n parameters α_i, $i = 1, \ldots, n$, één per trainingsobservatie en $K(x_i, x)$ het inwendig product van x_i en x. Een inwendig product van twee observaties x_i en $x_{i'}$ wordt gedefinieerd als: $\langle x_i, x_{i'} \rangle = \sum_{j=1}^{p} x_{ij} x_{i'j}$. Het schatten van de parameters $\alpha_1, \ldots, \alpha_n$ en β_0 gebeurt aan de hand van $\binom{n}{2}$ inwendige producten $\langle x_i, x_{i'} \rangle$ tussen alle paren van trainingsobservaties. Interessant hierbij is het feit dat veel α's gelijk zullen zijn aan nul. Optimalisatie-algoritmen kunnen deze eigenschap uitbuiten zodat het fitten van dergelijke modellen computationeel zeer efficiënt verloopt. Bovendien kan aangetoond worden dat, indien men de ruimte van alle lineaire functies uitbreidt tot alle polynomen van graad d, opnieuw geldt dat de functie die H minimaliseert kan geschreven worden als Vgl. (5.7), maar met:

$$K(x_i, x_{i'}) = (c_1 + c_2 \sum_{j=1}^{p} x_{ij} x_{i'j})^d, \tag{5.8}$$

waarbij x_i en $x_{i'}$ twee vectoren van features zijn en c_1 en c_2 twee parameters. In deze masterproef werden c_1 en c_2 gelijkgesteld aan respectievelijk 0 en $\frac{1}{\text{aantal features}}$. Dit zijn standaardwaarden die vaak gebruikt worden in de literatuur [51]. Algemeen noemt men K een kernelfunctie, die vaak gezien wordt als een functie die de gelijkenis tussen twee observaties kwantificeert. In R werd SVR opgesteld met een polynomiale kernel aan de hand van het pakket 'e1071' [51].

5.1.7 Regressiebomen

Voorgaande regressietechnieken veronderstellen dat er een lineaire relatie is tussen de concentratie aan zwarte koolstof en de verschillende features (met uitzondering van SVR met een polynomiale kernel). Deze regressietechnieken zullen een hoge performantie opleveren indien aan deze onderliggende relatie voldaan is. Is de relatie tussen de concentratie aan zwarte koolstof en de verschillende features niet lineair, dan zullen niet-lineaire regressietechnieken, zoals regressiebomen, een hogere performantie opleveren dan voorgaande regressietechnieken. Het opstellen van regressiebomen gebeurt in twee stappen:

1. Het opsplitsen van de featureruimte (i.e. alle mogelijke waarden voor X_1, X_2, ..., X_p) in J verschillende en niet-overlappende regio's R_1, R_2, ..., R_J.

2. Voor elke observatie die in de regio R_J valt, maakt men een voorspelling dat het gemiddelde is van de responswaarden voor de trainingsobservaties in R_J.

Het bekomen van de verschillende regio's in Stap 1 gebeurt via *recursive binary splitting*. Deze *top-down* benadering start aan de top van de regressieboom, waarbij alle observaties nog tot één regio behoren. De featureruimte splitst opeenvolgend, waarbij bij elke split twee nieuwe takken ontstaan. Bij *recursive binary splitting* wordt een feature X_j en een *cutpoint* s geselecteerd, zodat de featureruimte wordt gesplitst in de regio's $\{X|X_j < s\}$ en $\{X|X_j \geq s\}$ die leiden tot de grootst mogelijke reductie in RSS, met

$$RSS = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2. \tag{5.9}$$

Hierbij is $\hat{y}_{R_j}$ de gemiddelde respons voor de trainingsobservaties in de j^{de} regio. Men beschouwt alle features X_1, X_2, ..., X_p en alle mogelijke waarden van *cutpoint* s voor elk van de features en kiest vervolgens het feature en *cutpoint* zodanig dat de resulterende boom de laagste RSS heeft. Het opsplitsen van de featureruimte in regio's blijft doorgaan tot een bepaald stopcriterium bereikt wordt (bijvoorbeeld: geen enkele regio mag meer dan vijf observaties bevatten). Om overfitting van de data en complexiteit tegen te gaan, wordt vervolgens de regressieboom gesnoeid. Het snoeien van regressiebomen kan gebeuren via *cost complexity pruning*. Hierbij beschouwt men die regressiebomen die zijn aangeduid door een niet-negatieve parameter α, waarbij α de *trade-off* bepaalt tussen de complexiteit van een regressieboom en zijn fit aan de trainingsdata. Met elke waarde van α correspondeert een subregressieboom $T \subset T_0$ zodat

$$\sum_{m=1}^{|T|} \sum_{i:x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha|T| \tag{5.10}$$

zo klein mogelijk is. Hierbij is $|T|$ het aantal eindknopen van de regressieboom, R_m de subset van de featureruimte die correspondeert met de m^{de} eindknoop en $\hat{y}_{R_m}$ is het gemiddelde van

de trainingsobservaties in R_m. Het algoritme om regressiebomen op te stellen is weergegeven in Tabel 5.3. Het gebruikte pakket in R was 'tree' [65].

Tabel 5.3: Algoritme: Opbouw regressieboom.

1. Gebruik *recursive binary splitting* teneinde de grootst mogelijke regressieboom op te stellen met behulp van de trainingsdata tot het stopcriterium is bereikt.
2. Toepassen van *cost complexity pruning* op de grootst mogelijke regressieboom teneinde de sequentie van beste subregressiebomen te bekomen in functie van α.
3. Gebruik van ruimtelijke gestratificeerde 4-fold crossvalidatie voor het bepalen van α. Voor elke k=1, 2, 3, 4:
   a Herhaal Stappen 1 en 2 op alle trainingsdata met uitzondering van de k^{de} fold.
   b Evalueer de *mean squared prediction error* op de k^{de} fold in functie van α.

   Neem het gemiddelde voor elke waarde van α en selecteer vervolgens die waarde van α die de laagste gemiddelde fout oplevert.
4. Neem de subregressieboom van Stap 2 die overeenstemt met de gekozen waarde voor α.

5.1.8 Random forests

Het nadeel van regressiebomen is dat de voorspellende kracht in de praktijk vaak beperkter is dan die van andere regressietechnieken. Deze voorspellende kracht kan substantieel verhoogd worden door het aggregeren van verschillende regressiebomen, wat wordt toegepast in *random forests* [16]. Vooraleer deze regressietechniek wordt toegelicht, is het noodzakelijk om eerst het principe *bagging* toe te lichten. Bij *bagging* worden B niet-gesnoeide regressiebomen opgesteld door gebruik te maken van B *bootstrapped* trainingsdatasets. Vervolgens wordt het gemiddelde genomen van deze regressiebomen, volgens volgende vergelijking:

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x), \tag{5.11}$$

met $\hat{f}^{*b}(x)$ de voorspelling van de b^{de} *bootstrapped* trainingsdataset en $\hat{f}_{bag}(x)$ de finaal voorspelde regressieboom. Een individuele regressieboom heeft een hoge variantie, maar een lage bias. Door het gemiddelde te nemen van een groot aantal regressiebomen wordt deze variantie

gereduceerd. Dit regressiemodel laat toe om de testfout te berekenen zonder crossvalidatie. Elke regressieboom maakt slechts gebruik van ongeveer twee derden van de observaties, waardoor één derde van de observaties niet gebruikt wordt voor het fitten van het model aan de data. Deze observaties, *out-of-bag* (OOB) observaties genoemd, kunnen gebruikt worden om de testfout te berekenen door het gemiddelde te nemen van de verschillende OOB-voorspellingen. *Random forests* verschilt van *bagging* in de manier waarop een regressieboom wordt gesplitst. Bij *random forests* worden telkens willekeurig m features gekozen als splitkandidaten, terwijl bij *bagging* gekozen wordt uit de volledige set van p features. De split mag slechts één feature gebruiken van de m features. Voor regressiemodellen wordt vaak geopteerd voor $m \approx \frac{p}{3}$. Doordat bij *random forests* niet alle features worden beschouwd bij het splitten van de regressieboom, zullen de verschillende regressiebomen minder met elkaar zijn gecorreleerd. De reden hiervoor is dat door het willekeurig nemen van m features een sterke feature in de dataset niet telkens het opstellen van de regressiebomen domineert, waardoor de verschillende regressiebomen minder op elkaar lijken. Het gemiddelde van niet-gecorreleerde regressiebomen leidt tot een grotere reductie in variantie in vergelijking met het gemiddelde van gecorreleerde regressiebomen. Daardoor kan *random forests* beschouwd worden als een verbetering van *bagging*. Het opstellen van *random forests* gebeurde met behulp van het pakket 'randomForest' in R [49].

5.1.9 K-Nearest Neighbors regressie

Lineaire regresie is een parametrische methode (Sectie 5.1.1). Dit heeft als voordeel dat men slechts een klein aantal regressiecoëfficiënten dient te schatten, maar als nadeel dat er een assumptie over de functionele vorm van $f(x)$ wordt gemaakt. Indien echter niet voldaan is aan deze assumptie, dan zal het resulterend model slecht fitten aan de data en bestaat het risico dat er verkeerde conclusies worden genomen. Een niet-parametrische methode, zoals *K-Nearest Neighbors* (KNN) regressie, is hiervoor een alternatief gezien het geen functionele vorm van $f(x)$ aanneemt. Doordat KNN regressie geen veronderstelling maakt van de onderliggende data is het een meer flexibele techniek. KNN regressie identificeert de K dichtstbijzijnde observaties ten opzichte van een voorspellingspunt x_0, die worden voorgesteld door N_0. Vervolgens schat KNN regressie $f(x_0)$ door het gemiddelde te nemen van alle responsen aanwezig in N_0, voorgesteld door volgende vergelijking:

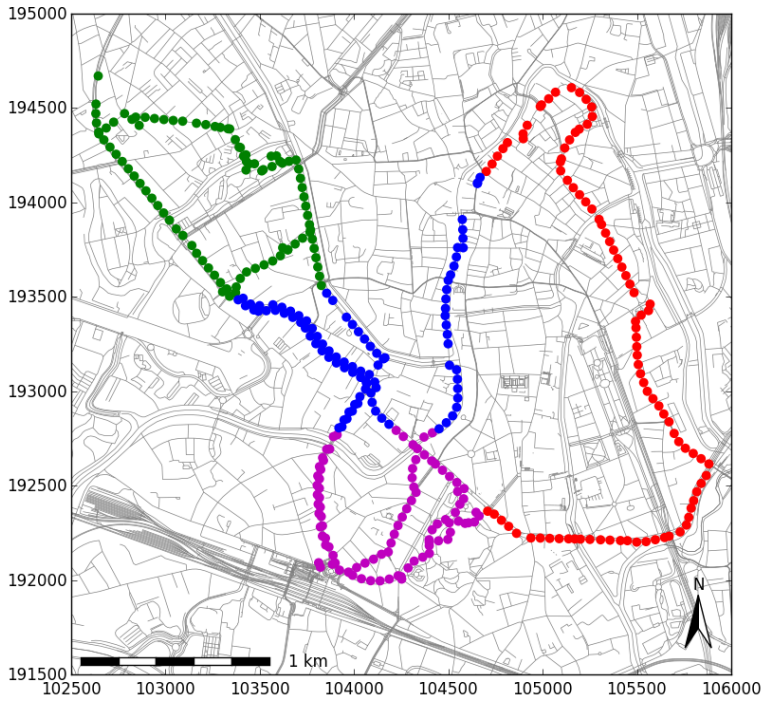
$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \epsilon N_0} y_i. \tag{5.12}$$

De optimale waarde voor K hangt af van de *bias-variance trade-off*. Een kleine waarde voor K produceert een meer flexibele fit, wat overeenstemt met een lage bias, maar een hoge variantie. De hoge variantie wordt veroorzaakt door het feit dat een voorspelling in een bepaalde regio volledig afhangt van een beperkt aantal observaties (hoe kleiner K, hoe minder observaties).

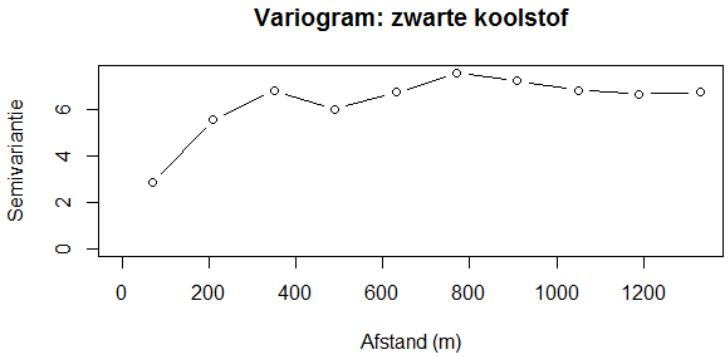
Grotere waarden voor K produceren daardoor een minder variabele fit, omdat de voorspelling in een regio dan het gemiddelde is van verschillende observaties, waarbij het veranderen van één enkele observatie minder effect heeft. Een grotere K kan een hogere bias veroorzaken, wat dan weer een nadeel is. Het uitvoeren van KNN regressie gebeurde met behulp van het pakket 'FNN' in R [11].

5.2 Performantie-analyse

Eén van de doelstellingen was het evalueren van de performantie die men kan behalen bij het voorspellen van zwarte koolstof (Sectie 1.2). Om die doelstelling te kunnen bereiken, dient het model getest te worden op data die niet werden gebruikt voor het trainen van het model. Bij crossvalidatie worden data opgesplitst in test- en trainingsdatasets. Het evalueren van de performantie van modellen gebeurde in het ESCAPE project [9, 23] door middel van *leave-one-out* crossvalidatie en in Hasenfratz et al. [38] door middel van 10-fold crossvalidatie. In deze masterproef werd echter gebruik gemaakt van ruimtelijk gestratificeerde 4-fold crossvalidatie. Hiertoe werden de beschikbare data met bijhorende features ruimtelijk opgedeeld in vier folds (Figuur 5.1). Bij de keuze voor deze folds werd rekening gehouden met de ruimtelijke verdeling van de geëxtraheerde features, waarbij werd getracht in elke fold een gelijk aantal datapunten te bekomen die bovendien een ruime range van waarden bevatten voor elk feature. Ruimtelijk gestratificeerde 4-fold crossvalidatie werd verkozen boven de eerder vernoemde crossvalidatietechnieken, omdat zwarte koolstof concentraties op POIs die dicht bij elkaar liggen te sterk afhankelijk van elkaar zijn. Als gevolg daarvan zijn test- en trainingsdatasets niet onafhankelijk van elkaar, wat een voorwaarde is om modelperformanties correct te kunnen inschatten. Uit het variogram van zwarte koolstof (Figuur 5.2) blijkt dat een afstand van minimum 375 m tussen twee POIs vereist is zodat de zwarte koolstof concentraties niet langer afhankelijk van elkaar zijn. Dit variogram werd opgesteld met behulp van de pakketten 'geoR' [64] en 'rgdal' [12] in R. De optimale waarde voor λ in ridge regressie en lasso en de optimale waarde voor α in regressiebomen werd bekomen door gebruik te maken van ruimtelijk gestratificeerde 3-fold crossvalidatie. Hierbij werden de drie folds gebruikt die zich bij het evalueren van de performantie van de regressietechniek in de trainingsdataset bevonden. Ruimtelijk gestratificeerde 3-fold crossvalidatie werd eveneens gebruikt om te onderzoeken welke graad d in de polynomiale kernel leidde tot de laagste gemiddelde kwadratische fout (voor elke waarde van graad d werd gelijktijdig de kost C geassocieerd met de verliesfunctie getuned). Voorts werd het aantal B *bootstrapped* trainingdatasets voldoende groot gekozen bij regressietechniek *random forests*.



Figuur 5.1: Ruimtelijke opdeling van de dataset in vier delen (fold 1: groen, fold 2: paars, fold 3: rood, fold 4: blauw).



Figuur 5.2: Variogram van zwarte koolstof.

5.3 Bespreking modelresultaten

5.3.1 Vergelijking performantie van verschillende regressietechnieken

In Tabel 5.4 wordt de gemiddelde kwadratische fout (MSE) van de individuele folds en de totale MSE weergegeven. Hieruit blijkt dat regressietechnieken die gebruik maken van een beperkt aantal features beter scoren dan regressietechnieken die alle veertig features bevatten. Zo is de gemiddelde kwadratische fout bij lineaire regressie minstens tien maal groter in vergelijking met *forward* en *backward stepwise selection* en lasso. Ridge regressie en lasso leveren respectievelijk de laagste en tweede laagste gemiddelde kwadratische fout op. Hieruit blijkt dat een betere fit wordt bekomen met regressietechnieken die gebruik maken van regularisatie. Daarnaast blijkt dat de niet-lineaire regressietechnieken regressiebomen en *random forests* en de niet-parametrische regressietechniek KNN regressie hogere gemiddelde kwadratische fouten opleveren dan lineaire regressietechnieken, op lineaire regressie na. Dit besluit kan eveneens getrokken worden uit de performanties van SVR, waar gebruik werd gemaakt van een polynomiale kernel. De polynomiale kernel met graad één, dus een lineaire kernel, levert voor alle vier de folds de laagste gemiddelde kwadratische fout op. Bij het analyseren van de gemiddelde kwadratische fout valt op te merken dat fold 3 telkens met de grootste fout wordt voorspeld (met uitzondering van lineaire regressie). Mogelijke oorzaken hiervoor kunnen een niet-optimale ruimtelijke verdeling van de verschillende features zijn en het feit dat de route gelegen in fold 3 voornamelijk werd gereden gedurende de vroege namiddag, terwijl het tijdstip waarop de overige routes werden gereden zich meer situeerde in de ochtend- of avondspits. De voorspellingen van fold 3 kunnen daardoor leiden tot een systematische overschatting van de gemeten waarden.

Naast het vergelijken van de verschillende gemiddelde kwadratische fouten werd ook een Friedman test [15, 20] uitgevoerd. Een Friedman test is een test die aantoont of één van de regressietechnieken consequent een hogere of lagere performantie heeft dan de overige regressietechnieken. De nulhypothese (H_0) en alternatieve hypothese (H_a) luiden als volgt:
H_0: 'De performantie van alle regressietechnieken is gelijk.'
H_a: 'Er is minstens één regressietechniek met een afwijkende performantie.'
De p-waarde van de Friedman test is 0.0022. Gezien deze p-waarde kleiner is dan 0.05 kan men besluiten dat niet alle regressietechnieken equivalent zijn en dat minstens één van de regressietechnieken significant afwijkt van de andere regressietechnieken. Bij dit besluit is enige voorzichtigheid geboden, gezien de Friedman test veronderstelt dat de resultaten voor de verschillende folds onafhankelijk van elkaar zijn. Aan deze veronderstelling is echter niet voldaan, aangezien dezelfde observaties deels gebruikt worden bij het bouwen van de verschillende modellen. Hierdoor is de bekomen p-waarde slechts een benadering van de 'exacte' p-waarde, die waarschijnlijk iets groter zal zijn dan degene hier bekomen. Doordat

Tabel 5.4: Gemiddelde kwadratische fout van de individuele fold en totale gemiddelde kwadratische fout per regressietechniek. MSE uitgedrukt in $(\frac{\mu g}{m^3})^2$. Bij SVR was de gebruikte kernel lineair.

	MSE Fold 1	MSE Fold 2	MSE Fold 3	MSE Fold 4	MSE Totaal
Lineaire regressie	163.764	12.368	22.748	10.597	52.369
Forward stepwise selection	3.311	2.595	9.395	4.329	4.908
Backward stepwise selection	3.499	3.002	9.620	4.473	5.148
Ridge regressie	2.964	3.359	7.565	5.161	4.762
Lasso	2.898	3.578	7.960	4.826	4.815
SVR	8.034	2.849	8.386	5.214	6.121
Regressiebomen	3.986	9.426	10.214	4.880	7.127
Random forests	3.938	3.287	9.997	3.963	5.296
KNN regressie (K=1)	6.364	10.208	20.633	6.875	11.020
KNN regressie (K=5)	5.859	6.641	18.860	5.363	9.181

de nulhypothese werd verworpen, werd vervolgens een Conover post-hoc test (Tabel 5.5) uitgevoerd, waarbij een p-waarde kleiner dan 0.05 wijst op een significante afwijking van de gelijkheid, en dus dat de ene regressietechniek een hogere performantie oplevert dan de andere. Uit deze test kan men besluiten dat lineaire regressie en KNN regressie (K=1) duidelijk een lagere performantie vertonen en significant verschillen van *forward stepwise selection*, ridge regressie en lasso. Verder kan men uit deze test niet besluiten dat ridge regressie significant afwijkt van *forward* en *backward stepwise selection* en lasso. Het gebruik van ridge regressie vergt echter een grotere rekenkracht in vergelijking met de andere eerder genoemde regressietechnieken, aangezien bij ridge regressie alle features dienen te worden berekend. Vanuit dit opzicht wordt gekozen om de zwarte koolstof kaart op te stellen met behulp van lasso, aangezien deze regressietechniek de tweede laagste totale gemiddelde kwadratische fout oplevert (Tabel 5.4) en er niet kan besloten worden dat lasso en ridge regressie significant van elkaar verschillen in performantie.

Tabel 5.5: p-waarden Conover post-hoc test. Significante p-waarden zijn aangeduid in het vet.

	Lineaire regressie	KNN regressie (K=1)	KNN regressie (K=5)	Forward stepwise selection	Backward stepwise selection	Ridge regressie	Lasso	Random forests	Regressie-bomen	SVR
Lineaire regressie	–	–	–	–	–	–	–	–	–	–
KNN regressie (K=1)	1	–	–	–	–	–	–	–	–	–
KNN regressie (K=5)	1	1	–	–	–	–	–	–	–	–
Forward stepwise selection	**0.0009**	**0.0083**	0.0703	–	–	–	–	–	–	–
Backward stepwise selection	**0.0083**	0.0703	0.4682	1	–	–	–	–	–	–
Ridge regressie	**0.0054**	**0.0463**	0.3337	1	1	–	–	–	–	–
Lasso	**0.0035**	**0.0304**	0.2277	1	1	1	–	–	–	–
Random forests	**0.0126**	0.1029	0.6698	1	1	1	1	–	–	–
Regressiebomen	0.6698	1	1	0.3337	1	1	0.8776	1	–	–
SVR	0.1029	0.6698	1	1	1	1	1	1	1	–

5.3.2 Bespreking relevante features

Het aantal overgebleven features als omgevingsvariabelen aanwezig in de regressietechnieken *forward* en *backward stepwise selection* en lasso varieerden van één tot negen (Tabel A.1). Dit aantal verschilt sterk van het totale aantal features. Bij Beelen et al. [9] varieerde dit aantal tussen twee en zeven omgevingsvariabelen en bij Eeftens et al. [23] tussen twee en vijf omgevingsvariabelen. Opmerkelijk is dat in elk van deze regressietechnieken minstens één feature terugkomt gebaseerd op de verkeersintensiteit. Acht van de twaalf regressiemodellen bevatten het feature 'aantal voertuigen in een straal van 25 m'. Eveneens komt in acht van de twaalf regressiemodellen het feature 'afstand tot dichtstbijzijnde tertiaire weg' voor. Verkeersgerelateerde features bleken eveneens een belangrijke rol te spelen in het opbouwen van LUR modellen in het ESCAPE project [9, 23]. Verder valt op te merken dat in geen enkel regressiemodel noch het aantal huizen, noch de *sky view factor* als relevante feature terugkomt. Een gelijkaardig besluit kan genomen worden voor *random forests* (Figuren A.1, A.2, A.3 en A.4). Uit een analyse van de belangrijkste omgevingsvariabelen blijkt dat verkeersgerelateerde features opnieuw een belangrijke rol spelen. De feature 'aantal voertuigen in een straal van 25 m' staat bij elk van de vier regressiemodellen in de top drie van belangrijkste features. Slechts bij één regressiemodel staat het aantal huizen in een bepaalde buffer in de top tien van belangrijkste features. Wanneer men kijkt naar de correlatie tussen de features en zwarte koolstof (Tabel 4.3), dan valt op te merken dat net deze features die de hoogste correlatie vertonen met zwarte koolstof frequent voorkomen, terwijl features die een lage correlatie vertonen met zwarte koolstof niet voorkomen als relevante feature. Zo is de correlatie met zwarte koolstof voor 'aantal voertuigen in een straal van 25 m' en 'afstand tot dichtstbijzijnde tertiaire weg' respectievelijk 33.94% en 16.44%, terwijl de correlatie met zwarte koolstof voor het aantal huizen in een bepaalde buffer maximaal 2.29% bedraagt.

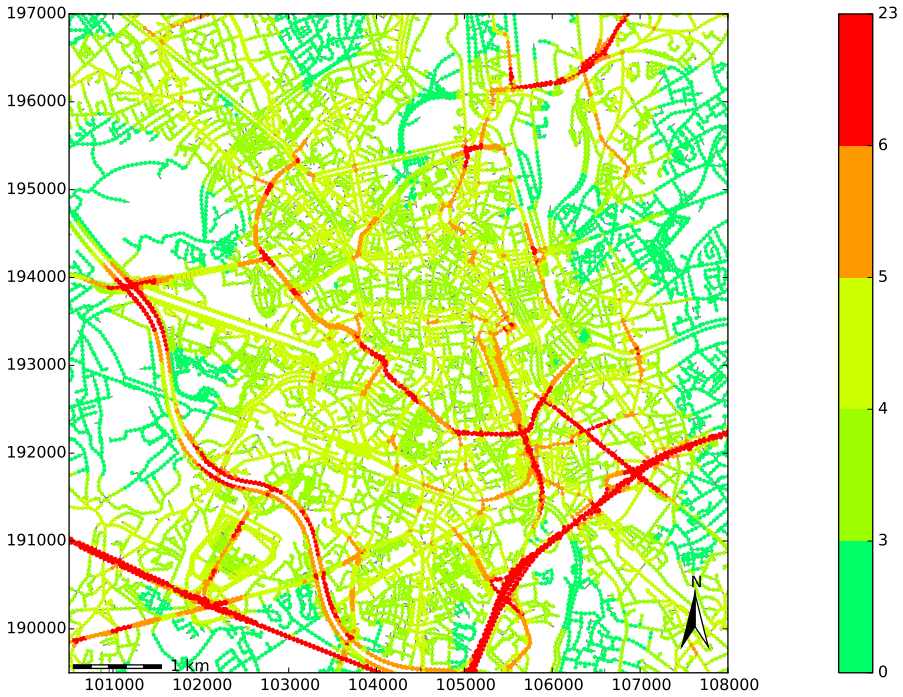
5.3.3 Analyse van het opstellen van de zwarte koolstof kaart met lasso

Het opstellen van de zwarte koolstof kaart van de stadsregio Gent gebeurde met behulp van de regressietechniek lasso. De waarde van de tuningsparameter λ werd bekomen door middel van ruimtelijk gestratificeerde 4-fold crossvalidatie en bedroeg 0.552. De inputvariabelen van lasso waren de features: 'afstand tot dichtstbijzijnde tertiaire weg', 'aantal voertuigen in een straal van 25 m', 'aantal vrachtwagens in een straal van 25 m', 'aantal vrachtwagens in een straal van 50 m' en 'aantal vrachtwagens in een straal van 100 m'. Op die manier bekomt men volgend model:

$$y = 4.302 - 3.213 \cdot 10^{-3} x_1 + 3.937 \cdot 10^{-6} x_2 + 1.825 \cdot 10^{-4} x_3 + 1.969 \cdot 10^{-7} x_4 + 6.847 \cdot 10^{-6} x_5, \quad (5.13)$$

met y de zwarte koolstof concentratie ($\frac{\mu g}{m^3}$), x_1 de afstand tot de dichtstbijzijnde tertiaire weg (m), x_2 het aantal voertuigen in een straal van 25 m ($\frac{\text{voertuigen}}{\text{uur}}$ m), x_3 het aantal vrachtwagens in een straal van 25 m ($\frac{\text{vrachtwagens}}{\text{uur}}$ m), x_4 het aantal vrachtwagens in een straal van

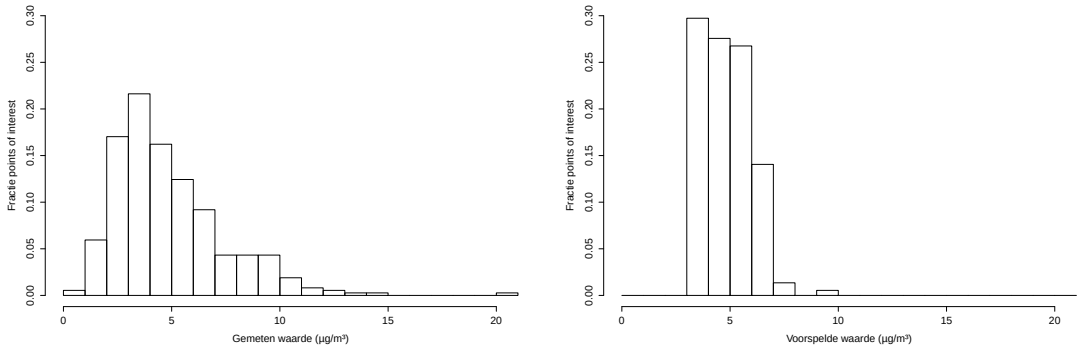
50 m ($\frac{\text{vrachtwagens}}{\text{uur}}$ m), x_5 het aantal vrachtwagens in een straal van 100 m ($\frac{\text{vrachtwagens}}{\text{uur}}$ m). Uiteindelijk bekomt men de zwarte koolstof kaart van de stadsregio Gent (Figuur 5.3) door het toepassen van dit model en het gebruik van de omgevingsvariabelen van de stadsregio Gent als inputvariabelen. Eventuele negatieve berekende concentraties werden gelijk gesteld aan 0 $\frac{\mu g}{m^3}$. De kleurenschaal is analoog aan Figuur 3.5. Figuur 5.3 toont duidelijk aan dat zwarte koolstof spatiaal variabel is. In het rood zijn voornamelijk de E40 en E17 herkenbaar. Andere drukke wegen, zoals R4, B401 en R40, zijn terug te vinden door hun oranje-rode kleur. Het stedelijk natuurreservaat Bourgoyen-Ossemeersen wordt gekenmerkt door een groene kleur. De spatiale variabiliteit van zwarte koolstof is in overeenstemming met de bronnen van zwarte koolstof.



Figuur 5.3: Zwarte koolstof kaart van de stadsregio Gent. De kleurencode weerspiegelt de concentratie aan zwarte koolstof in $\frac{\mu g}{m^3}$.

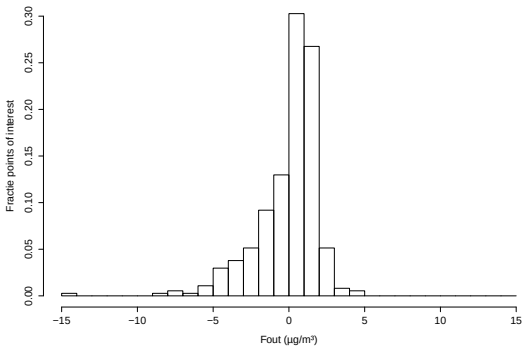
Door regularisatie worden de voorspelde waarden minder extreem dan de gemeten waarden (Figuur 5.4). Uit het histogram van de fouten (Figuur 5.5) valt op te merken dat deze fouten niet normaal verdeeld zijn. Hierbij bedroeg de gemiddelde fout (RMSE) 2.07 $\frac{\mu g}{m^3}$. Wanneer men de gemiddelde fout deelt door het gemiddelde van de gemeten waarden van zwarte koolstof (4.86 $\frac{\mu g}{m^3}$), dan bekomt men een waarde van 42.60%. Dit is nog een behoorlijk grote fout,

maar het is wel mogelijk om locaties met een hoge concentratie (ongeveer 8 $\frac{\mu g}{m^3}$) te onderscheiden van locaties met een lage concentratie (ongeveer 0–2 $\frac{\mu g}{m^3}$). De gemiddelde gekwadrateerde fout van de voorspelde waarden (MSE) bedroeg dus 4.27 $(\frac{\mu g}{m^3})^2$ en de gemiddelde gekwadrateerde afwijking van de gemeten waarden ten opzichte van de gemiddelde gemeten waarde bedroeg 6.76 $(\frac{\mu g}{m^3})^2$. Uit de gepaarde Wilcoxon rank test blijkt dat het model significant beter is dan een strategie die op elke locatie de gemiddelde (achtergrond)concentratie voorspelt (p-waarde < 0.0001) op basis van de mediaan van de gekwadrateerde fouten (1.46 $(\frac{\mu g}{m^3})^2$) en de mediaan van de gekwadrateerde afwijkingen van de gemeten waarden ten opzichte van de gemiddelde gemeten waarde (2.61 $(\frac{\mu g}{m^3})^2$). De determinatiecoëfficiënt R^2 bedroeg voor testfold 1, 2, 3 en 4 respectievelijk 31.44%, 55.73%, 16.62% en 36.42%. Globaal kan het vooropgestelde model 43.57% van alle variabiliteit in de data aanwezig verklaren. Uit de scatterplot van de voorspelde en gemeten waarden (Figuur 5.6) valt voornamelijk de hogere gemeten waarde van 20.94 $\frac{\mu g}{m^3}$ op. Deze waarde is gelegen in fold 4. De extreme waarde kan mogelijks te wijten zijn aan het feit dat er niet werd rekening gehouden met het aantal passages bij het opstellen van de zwarte koolstof kaart. Bij het verder analyseren van de fouten valt op te merken dat fold 1 relatief gezien de grootste bijdrage ($> 40\%$) levert aan gemiddelde fouten > 5 $\frac{\mu g}{m^3}$. Gemiddelde fouten tussen 3 $\frac{\mu g}{m^3}$ en 5 $\frac{\mu g}{m^3}$ worden vooral veroorzaakt door fold 3 (bijdrage van 60%). Folds 1, 2 en 4 dragen relatief gezien ongeveer evenveel bij tot gemiddelde fouten < 1 $\frac{\mu g}{m^3}$, terwijl fold 3 dit voor slechts 15% doet.

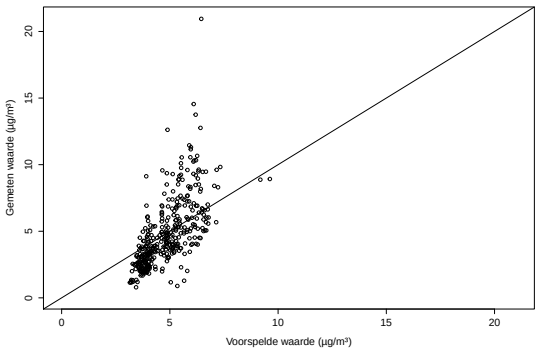


(a) (b)

Figuur 5.4: Histogrammen van (a) gemeten en (b) voorspelde waarden bekomen met regressietechniek lasso.



Figuur 5.5: Histogram van de fouten.



Figuur 5.6: Scatterplot van de voorspelde waarden bekomen met regressietechniek lasso en gemeten waarden. Lijn is de eerste bissectrice.

5.4 Kritische bemerkingen

Gedurende de meetcampagne werden enkel metingen uitgevoerd in het stedelijke gebied. Gezien er geen metingen werden uitgevoerd in het natuurreservaat Bourgoyen-Ossemeersen of langsheen de E17 of E40 kan men niet met zekerheid zeggen of het model ook extrapoleerbaar is naar deze gebieden. Verder werd reeds eerder aangehaald dat er in deze masterproef geen rekening werd gehouden met het aantal passages. Een mogelijke verbetering is de dataset aanvullen met nieuwe meetdata. Volgens Peters et al [56] zijn twintig tot vierentwintig passages uitgevoerd op verschillende dagen en op verschillende tijdstippen van de dag over een periode van twee tot drie weken noodzakelijk om straten met een hogere en lagere concentratie aan PM_{10} en UFP significant van elkaar te kunnen onderscheiden. Naast het aanvullen van de huidige dataset zou het gebruik van gewogen regressie eveneens een verbetering kunnen zijn. In een studie van Hasenfratz et al. [38] leidde het gebruik van het aantal metingen per cel als gewicht echter niet tot een verbetering van de modelperformantie. Het instrument MicroAeth® is in deze masterproef niet onderworpen geweest aan een datakwaliteitscontrole.

Een mogelijke datakwaliteitscontrole is het verzamelen van additionele stationaire metingen met MicroAeth®'s nabij de twee stationaire meetstations gelegen in Gent (Sectie 2.2.1). Men verwacht dat data afkomstig van MAAP's minder beïnvloed worden door de ladingseffecten in vergelijking met data afkomstig van MicroAeth®'s [58, 72]. Uit een studie van Van den Bossche et al. [72] uitgevoerd in Antwerpen blijkt dat MicroAeth®'s een gemiddelde absolute fout van 0.3 $\frac{\mu g}{m^3}$ vertonen in vergelijking met MAAP's gebaseerd op 30 minuutsgemiddelden. Verschillende studies houden eveneens rekening met meteorologische factoren teneinde de variabiliteit nog aanwezig in de data louter te kunnen toeschrijven aan de locatie. Hagler et al. [34] kwamen tot de conclusie dat verschillende meteorologische condities (in deze studie in het bijzonder windsnelheid en windrichting) leidden tot significante verschillen in concentratieniveaus nabij wegen en verschillen in spatiale patronen. Bukowiecki et al. [17] raadde aan om niet enkel gezondheidseffecten toe te schrijven aan een bepaald type locatie, maar ook rekening te houden met de invloed van weersomstandigheden op die specifieke locatie. In deze masterproef werden enkel data verzameld gedurende de seizoenen herfst en winter. Bij eventueel verder onderzoek zou men eveneens meetcampagnes kunnen uitvoeren in andere seizoenen, zoals werd gedaan in het ESCAPE project [23] of door Hasenfratz et al. [38]. Op die manier kan men seizoenale luchtvervuilingskaarten bekomen. In het finale model gebruikt voor het opstellen van de zwarte koolstof kaart waren onder andere volgende drie omgevingsvariabelen aanwezig: het aantal vrachtwagens in een straal van 25 m, 50 m en 100 m. Aangezien deze variabelen met verschillende buffergrootte elkaar overlappen, kunnen deze ook herschreven worden door gebruik te maken van aangrenzende concentrische ringen. In dat geval zouden de drie laatstgenoemde omgevingsvariabelen kunnen herschreven worden tot: 'het aantal vrachtwagens in een straal van 25 m', 'het aantal vrachtwagens in een straal tussen 25 m en 50 m' en 'het aantal vrachtwagens in een straal tussen 50 m en 100 m'. Dit herschrijven van variabelen heeft als voordeel dat het finale model beter te interpreteren is [9].

HOOFDSTUK 6

Routeplanner met minimale blootstelling aan zwarte koolstof

In dit hoofdstuk wordt besproken hoe een routeplanner, die de blootstelling van een fietser aan zwarte koolstof minimaliseert, werd ontworpen. Als startpunt worden twee algoritmes besproken, namelijk het Dijkstra algoritme en het A^* algoritme, die frequent gebruikt worden bij het oplossen van kortste pad problemen. Verder worden enkele routing problemen besproken om tot slot te eindigen met enkele kritische bemerkingen.

6.1 Gewogen grafen en het kortste pad probleem: definities en notatie

Een graaf G is een wiskundige manier om een verzameling (K) van objecten en de verbindingen (B) tussen deze objecten voor te stellen. De objecten noemt men de *knopen* van een graaf en de verbindingen tussen de objecten de *bogen* van de graaf. Beschouw daartoe een verzameling van k objecten $K = \{n_1,\ \ldots,\ n_k\}$, hierna knopen genoemd. Deze knopen zijn bijvoorbeeld alle kruispunten in het studiegebied. Beschouw tevens de verzameling $B \subseteq K \times K$. De elementen van B noemt men de bogen van de graaf. Een boog is bijgevolg een koppel van knopen. Indien een koppel $(n_i,\ n_j) \in B$, dan zegt men dat de knopen n_i en n_j verbonden zijn met elkaar. In deze masterproef zullen twee knopen (kruispunten) n_i en n_j verbonden zijn door een boog indien men zich van n_i naar n_j kan verplaatsen via het wegennetwerk zonder dat men een derde kruispunt moet passeren. Een rij van ℓ knopen $n_{i_1},\ n_{i_2},\ \ldots,\ n_{i\ell}$ waarvoor geldt dat elk koppel $(n_{i_j},\ n_{i_{j+1}})$ van opeenvolgende knopen behoort tot B (en dus verbonden is door een boog) noemt men een *pad*. De koppels van opeenvolgende knopen van het pad noemt men de bogen van een pad. Een *gewogen graaf* is een graaf waarbij men bovendien aan elke boog een positief reëel getal toekent. Men noemt dit getal het gewicht of de kost van de boog. Men noteert de kost die wordt toegekend aan de boog $(n_i,\ n_j)$ als $c_{i,j}$. De kost van een pad is de som van de kosten van alle bogen van dat pad. Voor twee gegeven knopen n_a en n_b

kan men trachten het volgende vraagstuk op te lossen: 'Zoek het pad met de laagste kost dat start in n_a en eindigt in n_b'. Dit vraagstuk noemt men *het kortste pad probleem* [30, 37, 55].

6.2 Kortste pad algoritmes: Dijkstra en A^*

Het Dijkstra algoritme [21] en het A^* algoritme [37] zijn algoritmen die men kan gebruiken om een pad met een minimale kost te zoeken tussen een gegeven beginknoop n_o en een gegeven eindknoop n_d van een gewogen graaf. Het pad met een minimale kost wordt vaak het kortste pad genoemd (vandaar de naamgeving kortste pad algoritmes). Hierna wordt het Dijkstra algoritme kort besproken.

Beschouw vooreerst een gewogen graaf G, met verzameling van knopen K en verzameling van bogen B. Bemerk dat $c_{i,j}$ de kost voorstelt geassocieerd met de boog $(n_i,\ n_j)$ en kan gezien worden als een eigenschap van deze boog. Op dezelfde manier kan men aan elke knoop $n \in K$ de volgende eigenschappen toekennen:

1. **afstand(n)**: een bovengrens voor de kost van het kortste pad tussen beginknoop n_o en n, deze wordt gelijk gesteld aan $+\infty$ tijdens de initialisatiestap van het algoritme.
2. **ouder(n)**: de voorlaatste knoop in het (tot dan toe gevonden) kortste pad dat n_o verbindt met n, deze wordt 'None' (i.e. niet gekend) gekozen tijdens de initialisatiestap van het algoritme.

Het Dijkstra algoritme zal vervolgens de volgende stappen doorlopen:

1. Voeg alle knopen uit K toe aan een nieuwe verzameling Q, deze verzameling stelt de tot dan toe onbezochte knopen voor.
2. Update afstand(n_o) als volgt: `afstand(`n_o`) <- 0`.
3. Kies van alle knopen in Q de knoop n_i met de kleinste waarde voor afstand(n_i).
4. Verwijder n_i uit Q.
5. Beschouw de verzameling K_{n_i} van alle buren van n_i (i.e. de knopen die verbonden zijn door een boog met n_i), die tevens tot Q behoren en voer voor elke buur $n_j \in K_{n_i}$ de volgende update uit:
   `if afstand(`n_j`) >` `afstand(`n_i`) +` $c_{i,j}$ `then:`

   `afstand(`n_j`) <- afstand(`n_i`) +` $c_{i,j}$

   `ouder(`n_j`) <-` n_i

   (Indien de *if*-voorwaarde vals is, wordt geen update uitgevoerd.)

6. Indien de geselecteerde knoop n_d is, dan beëindigt men het iteratieproces, anders gaat men opnieuw naar Stap 3.

Wanneer de iteratie beëindigd wordt, dan kan men eenvoudig het kortste pad bekomen door de eindknoop n_d te selecteren en vervolgens de ouder van deze knoop te bepalen. Van deze ouder bepaalt men opnieuw de ouder, enz., tot men de beginknoop bereikt. De knopen die men op die manier selecteert, vormen samen het kortste pad [55].

Doorheen de iteraties van het Dijkstra algoritme kan men steeds drie verzamelingen van knopen onderscheiden: X is de verzameling van knopen die reeds bezocht zijn (i.e. de knopen waarvoor het kortste pad tot n_o reeds gekend is) en die niet meer in Q zitten; Y is de verzameling van knopen waarvoor een tentatieve afstand gekend is (de bovengrens is niet meer gelijk aan $+\infty$, maar is mogelijks nog niet optimaal); en de verzameling Z bevat de overige knopen. Grafisch zullen de knopen die tot Y behoren een schil vormen, die de knopen die tot X behoren, scheidt van de knopen die tot Z behoren. Initieel bevat X enkel n_o, maar naarmate X groeit, zal deze schil zich verwijderen van n_o. Wanneer deze schil n_d bereikt, zal de beëindiging van het iteratieproces ingeleid worden [55].

Een nadeel van het Dijkstra algoritme is dat het computationeel intensief kan zijn. De oorzaak hiervan is dat het algoritme niet gebruik maakt van *a priori* kennis. Stel dat de knoop n_o gelegen is in het centrum van de stad en dat de knoop n_d gelegen is in het zuiden. Het Dijkstra algoritme zou net zo waarschijnlijk het korste pad zoeken ten noorden van n_o als dat het zou zoeken ten zuiden van n_o. De efficiëntie kan dus verbeterd worden door het incorporeren van *a priori* kennis in het zoekproces. Dit wordt toegepast in het A^* algoritme, waarin wordt gebruik gemaakt van een heuristische evaluatiefunctie. Hierdoor wordt de kost om zich te verplaatsen van knoop n_o tot n_i vermeerderd met een ondergrens voor de kost van het kortste pad van n_i naar n_d. Deze ondergrens is bijvoorbeeld de afstand van n_i naar n_d in vogelvlucht. Het gebruik van het A^* algoritme leidt in de praktijk vaak tot een reductie van 50% in computationele tijd in vergelijking met het Dijkstra algoritme [30, 37, 74].

6.3 Ontwerpen routeplanner

Eén van de doelstellingen was het ontwerpen van een routeplanner die de blootstelling van een fietser aan zwarte koolstof minimaliseert (Sectie 1.2). Om deze doelstelling te bereiken werd als eerste het wegennetwerk aangepast. Hierbij werden wegen zoals snelwegen of rijkswegen verwijderd teneinde het wegennetwerk van fietsers beter te benaderen. Voorts werd aan elke boog een zwarte koolstof hoeveelheid (μg) toegekend als kost volgens volgende formule:

$$F(\text{concentratie}, \text{afstand}) = \text{concentratie} \cdot \text{minuutventilatie} \cdot 60 \cdot \frac{\text{afstand}}{\text{snelheid}}, \qquad (6.1)$$

met 'concentratie' de gemiddelde concentratie aan zwarte koolstof ($\frac{\mu g}{m^3}$), 'minuutventilatie'[1] gelijk aan $52.65 \cdot 10^{-3}$ $\frac{m^3}{\text{minuut}}$, 'afstand' de booglengte (km) en 'snelheid' gelijk aan 17.5 $\frac{\text{km}}{\text{uur}}$. De waarden voor 'minuutventilatie' en 'snelheid' werden bekomen uit een studie van Int Panis [44], waarbij het gemiddelde werd genomen van de waarden bekomen voor vrouwen en mannen. Aangezien het berekenen van het optimale pad geen relatief grote computationele tijd vergt (grootte-orde: seconden), mogelijks door de beperkte grootte van het studiegebied, werd gebruik gemaakt van een implementatie van het Dijkstra algoritme. De routeplanner werd geïmplementeerd in Python [61]. Hierbij werd gebruik gemaakt van volgende bibliotheken:

1. **geopandas**: Werd gebruikt voor het inlezen en manipuleren van GIS-data (ondermeer bewerken van *shapefiles* en het uitvoeren van ruimtelijke *querries*).
2. **matplotlib**: Werd gebruikt om berekende routes te visualiseren.
3. **networkx**: Werd gebruikt voor het aanmaken van een graaf van het geïmplementeerde wegennetwerk voor fietsers die gebruikt wordt bij het berekenen van het kortste pad (de implementatie van het Dijkstra algoritme maakt gebruik van deze graaf).
4. **numpy**: Werd gebruikt voor het werken met matrices (gegevenstype *array*).
5. **pandas**: Werd gebruikt voor het werken met *dataframes*.
6. **pyproj**: Werd gebruikt voor het transformeren van coördinaten naar een ander coördinatenreferentiesysteem.
7. **shelve**: Werd gebruikt om grote Python objecten (in deze masterproef de gewogen graaf die de zwarte koolstof concentraties linkt aan het wegennetwerk) weg te schrijven naar en in te lezen uit een bestand op de harde schijf.
8. **smopy**: Werd gebruikt voor het inladen van OSM-afbeeldingen in Python, waarop vervolgens de routes worden geplot met behulp van matplotlib.

6.4 Vergelijking routes bekomen met criterium BC en criterium afstand

In wat volgt duiden de termen 'criterium BC' en 'criterium afstand' op het berekenen van het optimale pad met als kost respectievelijk de zwarte koolstof hoeveelheid en de booglengte. In Tabel 6.1 worden een aantal karakteristieken weergegeven van paden die gevonden werden op basis van beide criteria. Als start- en eindlocatie voor deze paden werden acht verschillende

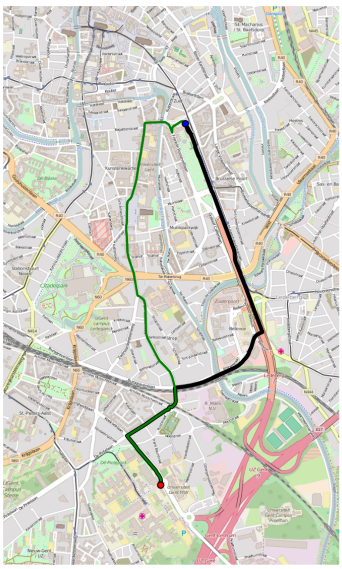
[1] Minuutventilatie is de ademfrequentie vermenigvuldigd met het ademvolume [44].

drukbezochte locaties in het studiegebied gekozen: Vrijdagmarkt, Veldstraat, Shoppingcenter Gent Zuid, Dok (Koopvaardijlaan), ingang voetgangers en fietsers van de Blaarmeersen, Gent Sint-Pietersstation ter hoogte van het Koningin Maria Hendrikaplein, hoofdingang van het Universitair Ziekenhuis (UZ), ingang Faculteit Bio-ingenieurswetenschappen gelegen aan Coupure Links. Hieruit valt op te merken dat voor een aantal start- en eindlocaties de berekende paden niet beïnvloed werden door het gekozen criterium (BC of afstand). Een logisch gevolg is dat hierbij dan ook geen verschillen optreden in de zwarte koolstof blootstelling en in de lengte van het traject. Het korste pad stemt in die gevallen overeen met de laagste blootstelling aan zwarte koolstof. Uit Vgl. (6.1) blijkt dat in het criterium BC indirect wordt rekening gehouden met afstand, waardoor de afstand bekomen met het criterium BC niet sterk zal afwijken van de afstand bekomen met het criterium afstand. Het maximale verschil in deze twee afstanden is 227 m (Vrijdagmarkt - Shoppingcenter Gent Zuid) in Tabel 6.1. Voor bepaalde start- en eindlocaties verschilt echter wel de route afhankelijk van het gekozen criterium. Hierdoor kan de procentuele afname aan zwarte koolstof hoeveelheid oplopen tot 15.96%. De routes die aanleiding geven tot de drie grootste waarden in procentuele afname aan zwarte koolstof hoeveelheid worden in wat volgt in meer detail besproken: Shoppingcenter Gent Zuid - Universitair Ziekenhuis, Dok - Universitair Ziekenhuis, Vrijdagmarkt - Shoppingcenter Gent Zuid.

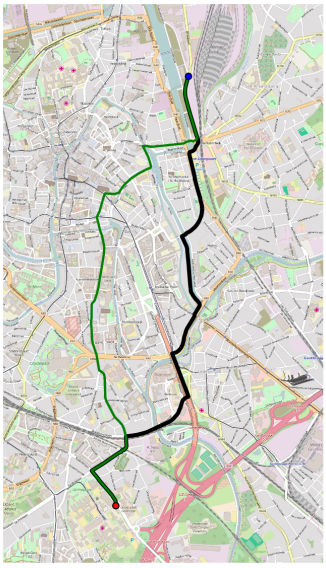
De berekende routes van het Shoppingcenter Gent Zuid naar het Universitair Ziekenhuis op basis van de twee criteria verschillen duidelijk van elkaar (Figuur 6.1). De route berekend op basis van het criterium BC mijdt de meer verkeersgerichte gebieden in vergelijking met de route berekend op basis van het criterium afstand. Dezelfde conclusie kan worden getrokken uit Figuur 6.2 die de route weergeeft van het Dok naar het Universitair Ziekenhuis. De route berekend op basis van het criterium afstand maakt gebruik van de R40, terwijl de R40 wordt vermeden met het andere criterium. Uit Figuur 3.5 bleek al dat de Sint-Lievenslaan en het Keizerviaduct, gelegen op de R40, gepaard gaan met relatief hogere concentraties aan zwarte koolstof. Een opmerkelijk verschil in de berekende routes van de Vrijdagmarkt naar het Shoppingcenter Gent Zuid in Figuur 6.3 is het vermijden van de Sint-Jacobsnieuwstraat wanneer de route berekend wordt met behulp van het criterium BC. Zoals reeds eerder vermeld in Sectie 3.2.1 werden er in deze straat hoge concentraties aan zwarte koolstof waargenomen, terwijl er in de Langemunt net lage concentraties aan zwarte koolstof werden waargenomen. De route berekend met het criterium BC loopt dan ook doorheen de Langemunt in plaats van doorheen de Sint-Jacobsnieuwstraat. Dit in tegenstelling tot de route berekend met het criterium afstand die wel doorheen de Sint-Jacobsnieuwstraat loopt.

Tabel 6.1: Totale blootstelling aan zwarte koolstof en lengte van de reisweg voor optimale paden. Criterium BC en criterium afstand wijzen op waarden bekomen door het optimale pad te berekenen met als kost respectievelijk de zwarte koolstof hoeveelheid en de booglengte. BC staat voor zwarte koolstof.

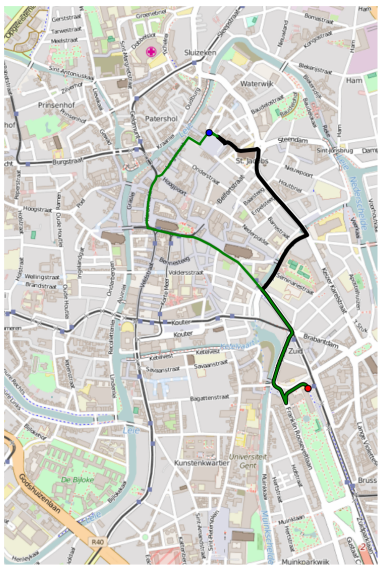
Route		BC blootstelling ($\frac{\mu g}{rit}$)		Afstand ($\frac{km}{rit}$)		Procentuele	Procentuele
Start	Einde	Criterium BC	Criterium afstand	Criterium BC	Criterium afstand	afname BC (%)	toename afstand (%)
Blaarmeersen	Veldstraat	2.202	2.246	2.791	2.667	1.99	4.65
Blaarmeersen	Vrijdagmarkt	2.712	2.757	3.573	3.449	1.62	3.60
Coupure Links	Gent Sint-Pietersstation	2.275	2.275	2.568	2.568	0.00	0.00
Coupure Links	Shoppingcenter Gent Zuid	1.587	1.611	2.215	2.167	1.51	2.25
Dok	Blaarmeersen	4.063	4.108	5.066	4.942	1.09	2.51
Dok	Universitair Ziekenhuis	4.449	4.913	5.338	5.113	9.44	4.39
Dok	Veldstraat	1.861	1.861	2.275	2.275	0.00	0.00
Gent Sint-Pietersstation	Shoppingcenter Gent Zuid	2.217	2.217	2.600	2.600	0.00	0.00
Gent Sint-Pietersstation	Vrijdagmarkt	2.399	2.399	3.031	3.031	0.00	0.00
Shoppingcenter Gent Zuid	Dok	2.230	2.254	2.701	2.640	1.07	2.30
Shoppingcenter Gent Zuid	Universitair Ziekenhuis	2.776	3.303	3.246	3.236	15.96	0.31
Universitair Ziekenhuis	Coupure Links	3.467	3.467	4.064	4.064	0.00	0.00
Universitair Ziekenhuis	Vrijdagmarkt	3.466	3.466	4.211	4.211	0.00	0.00
Veldstraat	Gent Sint-Pietersstation	1.889	1.889	2.249	2.249	0.00	0.00
Veldstraat	Shoppingcenter Gent Zuid	0.928	0.952	1.210	1.162	2.55	4.19
Veldstraat	Universitair Ziekenhuis	3.048	3.092	3.694	3.567	1.44	3.57
Vrijdagmarkt	Coupure Links	1.110	1.110	1.710	1.710	0.00	0.00
Vrijdagmarkt	Shoppingcenter Gent Zuid	1.371	1.411	1.837	1.610	2.84	14.06
Vrijdagmarkt	Veldstraat	0.593	0.593	0.907	0.907	0.00	0.00



Figuur 6.1: Route van het Shoppingcenter Gent Zuid (blauwe stip) naar het Universitair Ziekenhuis (rode stip). De groene lijn toont de route op basis van het criterium BC en de zwarte lijn toont de route op basis van het criterium afstand.



Figuur 6.2: Route van het Dok (blauwe stip) naar het Universitair Ziekenhuis (rode stip). De groene lijn toont de route op basis van het criterium BC en de zwarte lijn toont de route op basis van het criterium afstand.



Figuur 6.3: Route van de Vrijdagmarkt (blauwe stip) naar het Shoppingcenter Gent Zuid (rode stip). De groene lijn toont de route op basis van het criterium BC en de zwarte lijn toont de route op basis van het criterium afstand.

6.5 Kritische bemerkingen

De routeplanner berekende soms routes op basis van het criterium afstand die afweken van deze bekomen met Google Maps [32]. Afstanden bekomen in Tabel 6.1 volgens het criterium afstand vertoonden een overschatting van maximaal 29.07% en een onderschatting van maximaal 15.88% ten opzichte van de kortste afstand berekend met Google Maps voor fietsers. Een mogelijke oorzaak hiervan kan het gebruikte benaderde fietsennetwerk zijn. Mogelijks zijn hierin wegen aanwezig waar geen fietsers op toegestaan zijn of zijn er net wegen afwezig waar wel fietsers op toegestaan zijn. Het aanmaken van een wegennetwerk louter voor fietsers zou hier een oplossing voor kunnen bieden. Een andere mogelijke oorzaak is dat bepaalde wegen niet toegankelijk zijn voor de ontworpen routeplanner. Dit is mogelijk doordat de knopen en bogen werden aangemaakt met behulp van de bibliotheek 'networkx'. Bij het ontwerpen van de routeplanner werd de *shapefile*, die het benaderde wegennetwerk voor fietsers voorstelt, ingelezen, waarbij de bibliotheek 'networkx' vervolgens de graaf aanmaakt. Mogelijks ontbreken hierdoor bepaalde knopen en bogen, waardoor de routeplanner omwegen genereert. Indien het gewenst is om in de toekomst sneller de routes te berekenen, kan men opteren het A^* algoritme te implementeren.

HOOFDSTUK 7

Besluit

7.1 Algemene conclusies

Een eerste doelstelling was het ontwerpen van een zwarte koolstof kaart op basis van mobiele meetdata. Deze zwarte koolstof kaart (Figuur 5.3) werd opgesteld met behulp van de regressietechniek lasso en toont duidelijk aan dat zwarte koolstof spatiaal variabel is. Drukke wegen, zoals de E17, E40 en R40, gekenmerkt door een oranje-rode kleur, kan men onderscheiden van het stedelijk natuurreservaat Bourgoyen-Ossemeersen en sport- en recreatiepark Blaarmeersen, gekenmerkt door een lichtgroene-donkergroene kleur. Aangezien er voor het opstellen van de zwarte koolstof kaart gebruik werd gemaakt van mobiele meetdata die werden verzameld in enkel het stedelijke gebied, kan men echter niet met zekerheid zeggen of het model ook extrapoleerbaar is naar de E17, E40, het natuurreservaat en het sport- en recreatiepark.

Uit onderzoek van de omgevingseigenschappen bleek dat voornamelijk verkeersgerelateerde features het meest geschikt zijn om zwarte koolstof te voorspellen. Deze features bleken eveneens een belangrijke rol te spelen in het opbouwen van landgebruiksregressiemodellen in het ESCAPE project [9, 23]. Zo kwam er in de regressietechnieken *forward* en *backward stepwise selection* en lasso minstens één feature terug die gebaseerd is op verkeersintensiteit. Daarnaast kwam ook het feature 'afstand tot dichtstbijzijnde tertiaire weg' frequent voor in de regressiemodellen. Dezelfde conclusie kon worden getrokken uit de *variable importance plot* bepaald door *random forests*. Deze grafiek geeft de belangrijkheid van de verschillende features weer. Features gerelateerd aan de verkeersintensiteit hadden eveneens een hoge correlatie met zwarte koolstof concentraties. Hoe kleiner de straal van buffer werd gekozen voor deze features, hoe groter de determinatiecoëfficiënt R^2. De determinatiecoëfficiënt R^2 bedroeg eveneens meer dan 10% voor de features 'afstand tot dichtstbijzijnde tertiaire weg', 'afstand

tot dichtstbijzijnde kruispunt' en 'oppervlakte park in buffer met straal 100 m'.

Uit een analyse van de performantie van de verschillende regressietechnieken bekomen door middel van ruimtelijk gestratificeerde 4-fold crossvalidatie bleek dat regressietechnieken, zoals *forward* en *backward stepwise selection* en lasso, die gebruik maakten van een beperkt aantal features beter scoorden dan regressietechnieken die alle veertig features bevatten. Bovendien leverden de niet-lineaire regressietechnieken regressiebomen en *random forests* en de niet-parametrische regressietechniek *K-Nearest Neighbors* regressie hogere gemiddelde kwadratische fouten op. Voorts was de gemiddelde kwadratische fout het laagst wanneer er gebruik werd gemaakt van een polynomiale kernel met graad één bij *support vector* regressie in vergelijking met een polynomiale kernel met een graad groter dan één. Uit deze resultaten kan men besluiten dat, alvast in deze masterproef, enkel lineaire patronen in de data kunnen worden gebruikt om regressiemodellen te bouwen. Daarnaast hadden de regressietechnieken die gebruik maken van regularisatie (ridge regressie en lasso) de beste performantie. Gezien ridge regressie en lasso niet significant verschilden in performantie en het gebruik van ridge regressie een grotere rekenkracht vergde in vergelijking met lasso, werd geopteerd om de zwarte koolstof kaart op te stellen met behulp van regressietechniek lasso. De gemiddelde fout bedroeg dan finaal 2.07 $\frac{\mu g}{m^3}$. Dit is nog een behoorlijke grote fout, maar het is wel mogelijk om locaties met een hoge concentratie (ongeveer 8 $\frac{\mu g}{m^3}$) te onderscheiden van locaties met een lage concentratie (ongeveer 0–2 $\frac{\mu g}{m^3}$). Bovendien bleek dat het model signifant beter is dan een strategie die op elke locatie de gemiddelde (achtergrond)concentratie voorspelt. De determinatiecoëfficiënt R^2 bedroeg voor testfold 1, 2, 3 en 4 respectievelijk 31.44%, 55.73%, 16.62% en 36.42%. Globaal kon het vooropgestelde model 43.57% van alle variabiliteit in de data aanwezig verklaren.

In deze masterproef werd een routeplanner ontworpen die de blootstelling van een fietser aan zwarte koolstof minimaliseerde. Dit was mogelijk door aan elk straatsegment een zwarte koolstof hoeveelheid (μg) toe te kennen als kost en vervolgens gebruik te maken van een implementatie van het Dijkstra algoritme.

Tot slot werden traditionele kortste pad trajecten vergeleken met trajecten die bekomen werden met de ontworpen routeplanner. Voor een aantal start- en eindlocaties traden er geen verschillen op in de zwarte koolstof blootstelling en in de lengte van het traject. Het kortste pad stemde in die gevallen overeen met de laagste blootstelling aan zwarte koolstof. Echter in meer dan de helft van de berekende trajecten werd een afname van de blootstelling aan zwarte koolstof vastgesteld wanneer gebruik werd gemaakt van de ontworpen routeplanner. De procentuele afname aan zwarte koolstof hoeveelheid liep voor de berekende negentien routes op tot maximaal 15.96% (0.527 μg). Dit ging telkens gepaard met een toename in lengte van het traject.

7.2 Suggesties voor verder onderzoek

In deze masterproef werd één zwarte koolstof kaart bekomen. Men zou in verder onderzoek kunnen meetcampagnes uitvoeren in alle seizoenen om op die manier seizoenale luchtvervuilingskaarten te bekomen. Bij eventueel verder onderzoek zou men ook kunnen rekening houden met meteorologische condities, zoals windsnelheid en windrichting, bij het ontwikkelen van de landgebruiksregressiemodellen. Voorts zou men kunnen onderzoeken of deze modellen extrapoleerbaar zijn naar gebieden waar geen mobiele metingen werden uitgevoerd. Verder kan men van de ontworpen routeplanner een *grafische user interface* (GUI) ontwikkelen, waardoor de gebruiker op een eenvoudige manier zijn (fiets)traject kan plannen dat rekening houdt met de blootstelling aan zwarte koolstof. Tot slot kan men niet alleen zwarte koolstof in kaart brengen, maar eveneens ook andere polluenten die schadelijk zijn voor de mens en spatiaal variabel zijn.

Bibliografie

[1] 52°North (2015). SOS TestClient Version 2. http://sos.irceline.be/. Bezocht op 15/10/2015.

[2] AethLabs (2015). microAeth Model AE51: operating manual. Technical report.

[3] Agentschap voor Geografische Informatie Vlaanderen (2013a). Centraal Referentieadressenbestand. https://www.agiv.be/producten/crab. Bezocht op 30/9/2015.

[4] Agentschap voor Geografische Informatie Vlaanderen (2013b). Ondersteuning. https://www.agiv.be/producten/digitaal-hoogtemodel-vlaanderen/meer-over-dhm-v/ondersteuning-faq#vraag7. Bezocht op 8/10/2015.

[5] Agentschap voor Geografische Informatie Vlaanderen (2013c). Wat is het CRAB-project? https://www.agiv.be/producten/crab/meer-info-over-crab/algemeen/wat-is-crab. Bezocht op 30/9/2015.

[6] Agentschap voor Geografische Informatie Vlaanderen (2014a). Digitaal Hoogtemodel Vlaanderen II, DSM, raster, 1 m. https://download.agiv.be/Producten/Detail?id=937&title=Digitaal_Hoogtemodel_Vlaanderen_II_DSM_raster_1_m. Bezocht op 8/10/2015.

[7] Agentschap voor Geografische Informatie Vlaanderen (2014b). Kaartbladversnijdingen NGI, numerieke reeks. https://download.agiv.be/Producten/Detail/40. Bezocht op 8/10/2015.

[8] Agentschap voor Geografische Informatie Vlaanderen (2015). CRAB adresposities. https://download.agiv.be/Producten/Detail?id=102&title=CRAB_adresposities. Bezocht op 30/9/2015.

[9] Beelen, R., Hoek, G., Vienneau, D., Eeftens, M., Dimakopoulou, K., Pedeli, X., Tsai, M.-Y., Künzli, N., Schikowski, T., Marcon, A., Eriksen, K. T., Raaschou-Nielsen, O., Stephanou, E., Patelarou, E., Lanki, T., Yli-Tuomi, T., Declercq, C., Falq, G., Stempfelet, M., Birk, M., Cyrys, J., von Klot, S., Nádor, G., Varró, M. J., Dedele, A., Grazuleviciene, R.,

Mölter, A., Lindley, S., Madsen, C., Cesaroni, G., Ranzi, A., Badaloni, C., Hoffmann, B., Nonnemacher, M., Krämer, U., Kuhlbusch, T., Cirach, M., de Nazelle, A., Nieuwenhuijsen, M., Bellander, T., Korek, M., Olsson, D., Strömgren, M., Dons, E., Jerrett, M., Fischer, P., Wang, M., Brunekreef, B., and de Hoogh, K. (2013). Development of NO_2 and NO_x land use regression models for estimating air pollution exposure in 36 study areas in Europe - the ESCAPE project. *Atmospheric Environment*, 72:10–23.

[10] Berghmans, P., Bleux, N., Int Panis, L., Mishra, V. K., Torfs, R., and Van Poppel, M. (2009). Exposure assessment of a cyclist to PM_{10} and ultrafine particles. *Science of the Total Environment*, 407:1286–1298.

[11] Beygelzimer, A., Kakadet, S., Langford, J., Arya, S., Mount, D., and Li, S. (2013). FNN: fast nearest neighbor search algorithms and applications. R package version 1.1. `https://cran.r-project.org/package=FNN`.

[12] Bivand, R., Keitt, T., and Rowlingson, B. (2016). rgdal: bindings for the geospatial data abstraction library. R package version 1.1-7. `https://cran.r-project.org/package=rgdal`.

[13] Bond, T. C., Doherty, S. J., Fahey, D. W., Forster, P. M., Berntsen, T., DeAngelo, B. J., Flanner, M. G., Ghan, S., Kärcher, B., Koch, D., Kinne, S., Kondo, Y., Quinn, P. K., Sarofim, M. C., Schultz, M. G., Schulz, M., Venkataraman, C., Zhang, H., Zhang, S., Bellouin, N., Guttikunda, S. K., Hopke, P. K., Jacobson, M. Z., Kaiser, J. W., Klimont, Z., Lohmann, U., Schwarz, J. P., Shindell, D., Storelvmo, T., Warren, S. G., and Zender, C. S. (2013). Bounding the role of black carbon in the climate system: a scientific assessment. *Journal of Geophysical Research: Atmospheres*, 118:5380–5552.

[14] Boogaard, H., Borgman, F., Kamminga, J., and Hoek, G. (2009). Exposure to ultrafine and fine particles and noise during cycling and driving in 11 Dutch cities. *Atmospheric Environment*, 43:4234–4242.

[15] Bouckaert, R. R. (2003). Choosing between two learning algorithms based on calibrated tests. *Proceedings of the Twentieth International Conference on Machine Learning*, 1:51–58.

[16] Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

[17] Bukowiecki, N., Dommen, J., Prévôt, A. S. H., Weingartner, E., and Baltensperger, U. (2003). Fine and ultrafine particles in the Zürich (Switzerland) area measured with a mobile laboratory: an assessment of the seasonal and regional variation throughout a year. *Atmospheric Chemistry and Physics*, 3:1477–1494.

[18] Cai, J., Yan, B., Ross, J., Zhang, D., Kinney, P. L., Perzanowski, M. S., Jung, K., Miller, R., and Chillrud, S. N. (2014). Validation of MicroAeth® as a black carbon monitor for

fixed-site measurement and optimization for personal exposure characterization. *Aerosol and Air Quality Research*, 14:1–9.

[19] Choi, W., He, M., Barbesant, V., Kozawa, K. H., Mara, S., Winer, A. M., and Paulson, S. E. (2012). Prevalence of wide area impacts downwind of freeways under pre-sunrise stable atmospheric conditions. *Atmospheric Environment*, 62:318–327.

[20] Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30.

[21] Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik*, 1:269–271.

[22] Dons, E., Int Panis, L., Van Poppel, M., Theunis, J., and Wets, G. (2012). Personal exposure to black carbon in transport microenvironments. *Atmospheric Environment*, 55:392–398.

[23] Eeftens, M., Beelen, R., de Hoogh, K., Bellander, T., Cesaroni, G., Cirach, M., Declercq, C., Dedele, A., Dons, E., de Nazelle, A., Dimakopoulou, K., Eriksen, K., Falq, G., Fischer, P., Galassi, C., Grazuleviciene, R., Heinrich, J., Hoffmann, B., Jerrett, M., Keidel, D., Korek, M., Lanki, T., Lindley, S., Madsen, C., Mölter, A., Nádor, G., Nieuwenhuijsen, M., Nonnemacher, M., Pedeli, X., Raaschou-Nielsen, O., Patelarou, E., Quass, U., Ranzi, A., Schindler, C., Stempfelet, M., Stephanou, E., Sugiri, D., Tsai, M.-Y., Yli-Tuomi, T., Varró, M. J., Vienneau, D., von Klot, S., Wolf, K., Brunekreef, B., and Hoek, G. (2012). Development of land use regression models for $PM_{2.5}$, $PM_{2.5}$ absorbance, PM_{10} and PM_{coarse} in 20 European study areas; results of the ESCAPE project. *Environmental Science & Technology*, 46:11195–11205.

[24] Elen, B., Peters, J., Van Poppel, M., Bleux, N., Theunis, J., Reggente, M., and Standaert, A. (2013). The aeroflex: a bicycle for mobile air quality measurements. *Sensors*, 13:221–240.

[25] European Environment Agency (2013). Status of black carbon monitoring in ambient air in Europe. Technical report.

[26] European Environment Agency (2014). Urban Atlas. `http://www.eea.europa.eu/data-and-maps/data/urban-atlas`. Bezocht op 27/10/2015.

[27] European Union (2011). Mapping guide for a European Urban Altas. Technical report.

[28] Europese Commissie (2013). Mededeling van de Commissie aan het Europees Parlement, de Raad, het Europees Economisch en Sociaal Comité en het Comité van de Regio's - Programma 'Schone lucht voor Europa'. Technical report.

[29] Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33:1–22.

[30] Fu, L., Sun, D., and Rilett, L. R. (2006). Heuristic shortest path algorithms for transportation applications: state of the art. *Computers & Operations Research*, 33:3324–3343.

[31] Gents MilieuFront (2015). gentsmilieufront. `http://www.gentsmilieufront.be/`. Bezocht op 25/2/2016.

[32] GoogleMaps (2016). Routeplanner. `https://www.google.be/maps`. Bezocht op 16/05/2016.

[33] Hagemann, R., Corsmeier, U., Kottmeier, C., Rinke, R., Wieser, A., and Vogel, B. (2014). Spatial variability of particle number concentrations and NO_x in the Karlsruhe (Germany) area obtained with the mobile laboratory 'AERO-TRAM'. *Atmospheric Environment*, 94:341–352.

[34] Hagler, G. S. W., Thoma, E. D., and Baldauf, R. W. (2010). High-resolution mobile monitoring of carbon monoxide and ultrafine particle concentrations in a near-road environment. *Journal of the Air & Waste Management Association*, 60:328–336.

[35] Hagler, G. S. W., Yelverton, T. L. B., Vedantham, R., Hansen, A. D. A., and Turner, J. R. (2011). Post-processing method to reduce noise while preserving high time resolution in aethalometer real-time black carbon data. *Aerosol and Air Quality Research*, 11:539–546.

[36] Hansen, A. D. A., Rosen, H., and Novakov, T. (1984). The aethalometer - an instrument for the real-time measurement of optical absorption by aerosol particles. *Science of the Total Environment*, 36:191–196.

[37] Hart, P. E., Nilsson, N. J., and Raphael, B. (1968). A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions of Systems Science and Cybernetics*, 4:100–107.

[38] Hasenfratz, D., Saukh, O., Walser, C., Hueglin, C., Fierz, M., Arn, T., Beutel, J., and Thiele, L. (2015). Deriving high-resolution urban air pollution maps using mobile sensor nodes. *Pervasive and Mobile Computing*, 16:268–285.

[39] Hasenfratz, D., Saukh, O., Walser, C., Hueglin, C., Fierz, M., and Thiele, L. (2014). Pushing the spatio-temporal resolution limit of urban air pollution maps. *IEEE International Conference on Pervasive Computing and Communications*, 14:69–77.

[40] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning*. Springer Series in Statistics, Second edition.

[41] Hertel, O., Hvidberg, M., Ketzel, M., Storm, L., and Stausgaard, L. (2008). A proper choice of route significantly reduces air pollution exposure - a study on bicycle and bus trips in urban streets. *Science of the Total Environment*, 389:58–70.

[42] Hoek, G., Beelen, R., de Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P., and Briggs, D. (2008). A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric Environment*, 42:7561–7578.

[43] Hu, S., Paulson, S. E., Fruin, S., Kozawa, K., Mara, S., and Winer, A. M. (2012). Observation of elevated air pollutant concentrations in a residential neighborhood of Los Angeles California using a mobile platform. *Atmospheric Environment*, 51:311–319.

[44] Int Panis, L., de Geus, B., Vandenbulcke, G., Willems, H., Degraeuwe, B., Bleux, N., Mishra, V., Thomas, I., and Meeusen, R. (2010). Exposure to particulate matter in traffic: a comparison of cyclists and car passengers. *Atmospheric Environment*, 44:2263–2270.

[45] IRCEL (2015). Intergewestelijke Cel voor het Leefmilieu (IRCEL). `http://www.irceline.be/nl`. Bezocht op 18/10/2015.

[46] IRCEL-CELINE Lucht (2015). Jaarrapport luchtkwaliteit in België 2014. Technical report.

[47] James, G., Witten, D., Hastie, T., and Tibishirani, R. (2013). *An introduction to statistical learning with applications in R*. Springer Texts in Statistics, First edition.

[48] Janssen, N. A. H., Gerlofs-Nijland, M. E., Lanki, T., Salonen, R. O., Cassee, F., Hoek, G., Fischer, P., Brunekreef, B., and Krzyzanowski, M. (2012). Health effects of black carbon. Technical report, World Health Organization.

[49] Liaw, A. and Wiener, M. (2002). Classification and regression by randomForest. *R News*, 2:18–22.

[50] Liebens, J. (2013). Opbouw van het Multimodaal Model (MM) versie 3.6.1 - gedetailleerde beschrijving modelprocessen. Technical report.

[51] Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2015). e1071: misc functions of the department of statistics, probability theory group (formerly: E1071), TU Wien. R package version 1.6-7. `https://CRAN.R-project.org/package=e1071`.

[52] Oke, T. R. (1987). *Boundary layer climates*. Taylor & Francis Group, Second edition.

[53] OpenStreetMap-auteurs (2015a). OpenStreetMap. `http://www.openstreetmap.org`. Bezocht op 29/9/2015.

[54] OpenStreetMap-auteurs (2015b). OpenStreetMap België. `http://osm.be/`. Bezocht op 29/9/2015.

[55] Papadimitriou, C. H. and Steiglitz, K. (1998). *Combinatorial optimization: algorithms and complexity*. Dover Publications, Inc., Second edition.

[56] Peters, J., Theunis, J., Van Poppel, M., and Berghmans, P. (2013). Monitoring PM_{10} and ultrafine particles in urban environments using mobile measurements. *Aerosol and Air Quality Research*, 13:509–522.

[57] Peters, J., Van den Bossche, J., Reggente, M., Van Poppel, M., De Baets, B., and Theunis, J. (2014). Cyclist exposure to UFP and BC on urban routes in Antwerp, Belgium. *Atmospheric Environment*, 92:31–43.

[58] Petzold, A., Schloesser, H., Sheridan, P. J., Arnott, W. P., Ogren, J. A., and Virkkula, A. (2005). Evaluation of multiangle absorption photometry for measuring aerosol light absorption. *Aerosol Science and Technology*, 39:40–51.

[59] Petzold, A. and Schönlinner, M. (2004). Multi-angle absorption photometry - a new method for the measurement of aerosol light absorption and atmospheric black carbon. *Journal of Aerosol Science*, 35:421–441.

[60] Puttemans, C. (2013). Beschrijving BASMAT versie 3.6. Technical report.

[61] Python Software Foundation (2015). Python language reference, version 3.4.3. `http://www.python.org`.

[62] QGIS Development Team and Open Source Geospatial Foundation (2015). QGIS geographic information system. `http://qgis.osgeo.org`.

[63] R Core Team and R Foundation for Statistical Computing (2015). R: a language and environment for statistical computing. `https://www.R-project.org/`.

[64] Ribeiro Jr, P. J. and Diggle, P. J. (2015). geoR: analysis of geostatistical data. R package version 1.7-5.1. `https://CRAN.R-project.org/package=geoR`.

[65] Ripley, B. (2016). tree: classification and regression trees. R package version 1.0-37. `https://CRAN.R-project.org/package=tree`.

[66] Sharker, M. H. and Karimi, H. A. (2014). Computing least air pollution exposure routes. *International Journal of Geographical Information Science*, 28:343–362.

[67] Su, J. G., Winters, M., Nunes, M., and Brauer, M. (2010). Designing a route planner to facilitate and promote cycling in Metro Vancouver, Canada. *Transportation Research Part A*, 44:495–505.

[68] The MathWorks and Inc. (2015). MATLAB and Statistics Toolbox Release 2015b.

[69] Thomas Lumley using Fortran code by Alan Miller (2009). leaps: regression subset selection. R package version 2.9. `https://CRAN.R-project.org/package=leaps`.

[70] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58:267–288.

[71] Van Cauwenberge, B. (2015). Persoonlijke communicatie.

[72] Van den Bossche, J., Peters, J., Verwaeren, J., Botteldooren, D., Theunis, J., and De Baets, B. (2015). Mobile monitoring for mapping spatial variation in urban air quality: development and validation of a methodology based on an extensive dataset. *Atmospheric Environment*, 105:148–161.

[73] Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer-Verlag, Second edition.

[74] Vernieuwe, H., Ducheyne, E., Hendrickx, G., and De Baets, B. (2010). Efficient management of transportation logistics related to animal disease outbreaks. *Computers and Electronics in Agriculture*, 71:148–157.

[75] VITO NV (2015). Over VITO. `https://vito.be/nl/over-vito`. Bezocht op 25/2/2016.

[76] VITO NV (2016). airQmap additional information. `http://airqmap.com/info.html`. Bezocht op 18/04/2016.

[77] Vlaamse Milieumaatschappij (2014). Luchtkwaliteit in het Vlaamse Gewest - jaarverslag immissiemeetnetten - 2013. Technical report.

[78] Vlaamse Milieumaatschappij (2015). Wat is fijn stof? `https://www.vmm.be/lucht/fijn-stof/wat-is-fijn-stof#section-2`. Bezocht op 15/10/2015.

[79] Walkit (2016). The urban walking route planner. `https://walkit.com/`. Bezocht op 18/05/2016.

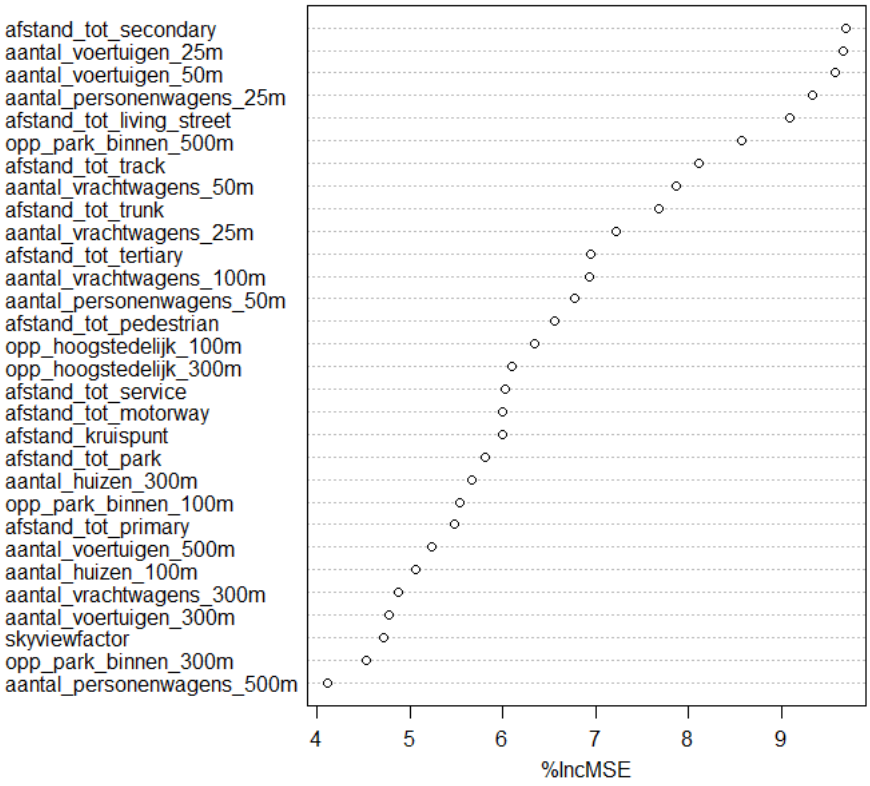
[80] Westerdahl, D., Fruin, S., Sax, T., Fine, P. M., and Sioutas, C. (2005). Mobile platform measurements of ultrafine particles and associated pollutant concentrations on freeways and residential streets in Los Angeles. *Atmospheric Environment*, 39:3597–3610.

[81] World Health Organization (2013). Health effects of particulate matter: policy implications for countries in eastern Europe, Caucasus and central Asia. Technical report.

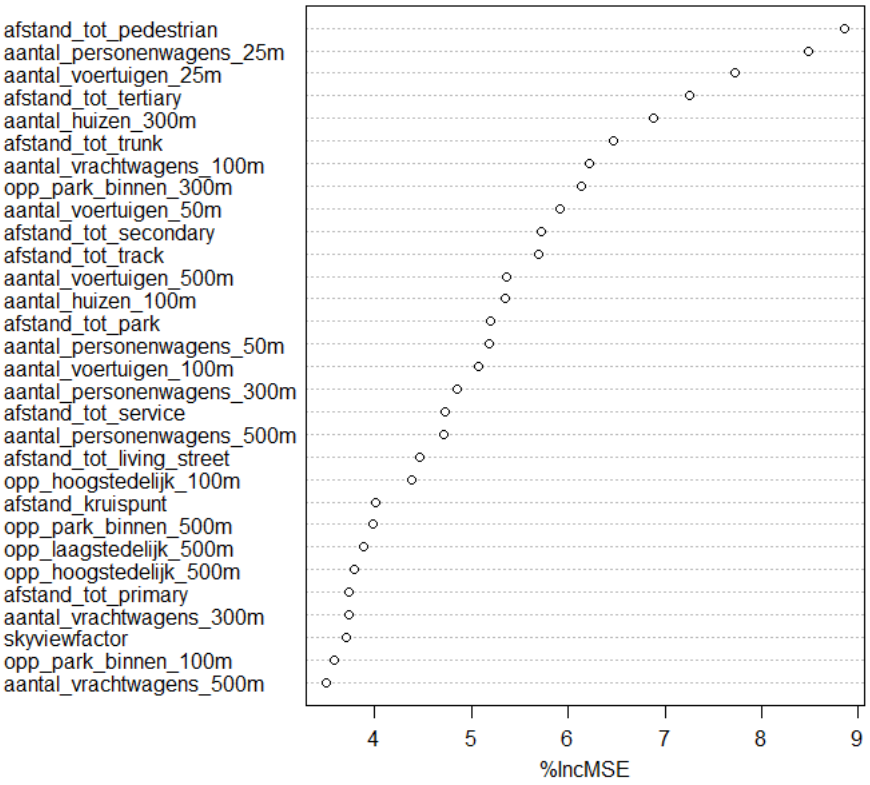
[82] Zwack, L. M., Paciorek, C. J., Spengler, J. D., and Levy, J. I. (2011). Modeling spatial patterns of traffic-related air pollutants in complex urban terrain. *Environmental Health Perspectives*, 119:852–859.

APPENDIX A

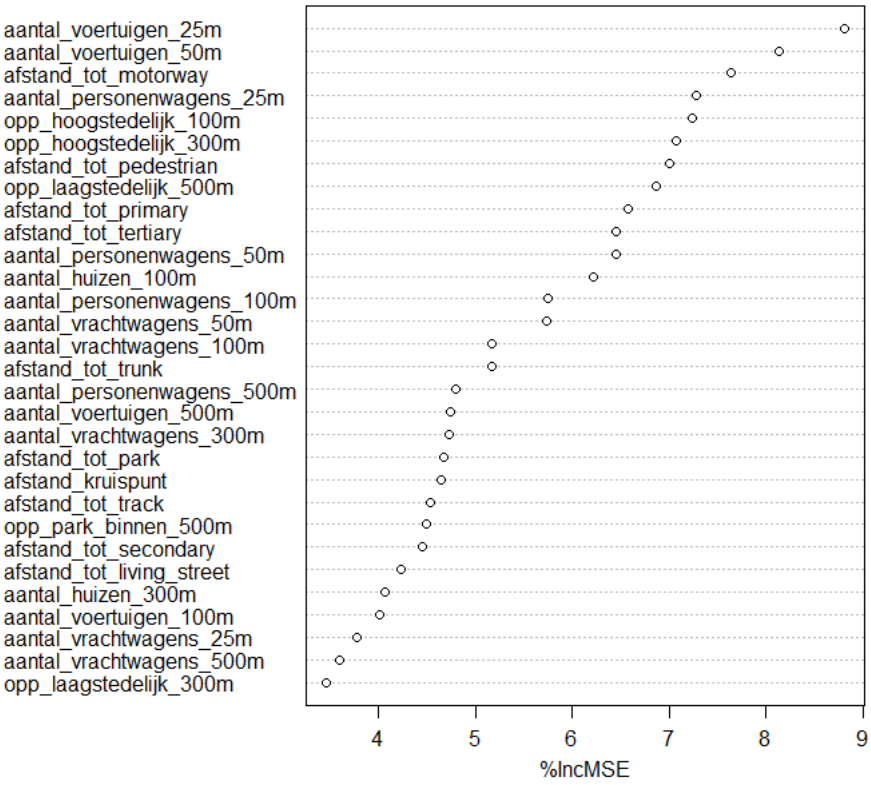
Relevante features



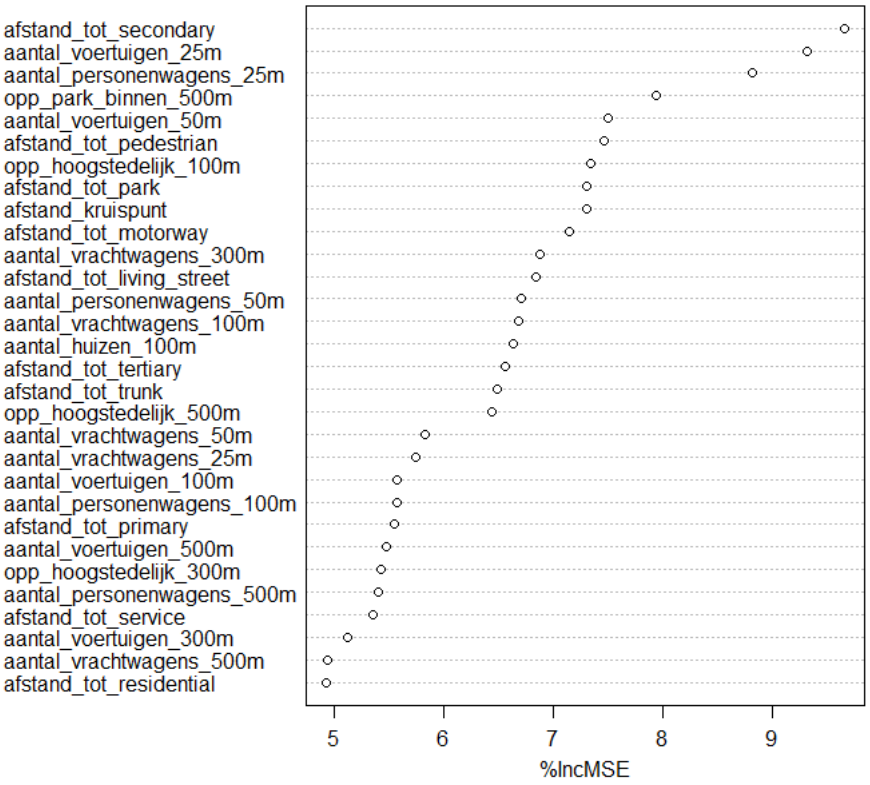
Figuur A.1: *Variable importance plot* bepaald met regressietechniek *random forests* voor testfold 1. Op de x-as wordt de procentuele toename in gemiddelde kwadratische fout weergegeven indien deze feature niet als omgevingsvariabele in het model aanwezig is. Op de y-as wordt de belangrijkheid van de features weergegeven. Hoe hoger de feature zich bevindt op de y-as, hoe belangrijker de feature is.



Figuur A.2: *Variable importance plot* bepaald met regressietechniek *random forests* voor testfold 2. Op de x-as wordt de procentuele toename in gemiddelde kwadratische fout weergegeven indien deze feature niet als omgevingsvariabele in het model aanwezig is. Op de y-as wordt de belangrijkheid van de features weergegeven. Hoe hoger de feature zich bevindt op de y-as, hoe belangrijker de feature is.



Figuur A.3: *Variable importance plot* bepaald met regressietechniek *random forests* voor testfold 3. Op de x-as wordt de procentuele toename in gemiddelde kwadratische fout weergegeven indien deze feature niet als omgevingsvariabele in het model aanwezig is. Op de y-as wordt de belangrijkheid van de features weergegeven. Hoe hoger de feature zich bevindt op de y-as, hoe belangrijker de feature is.



Figuur A.4: *Variable importance plot* bepaald met regressietechniek *random forests* voor testfold 4. Op de x-as wordt de procentuele toename in gemiddelde kwadratische fout weergegeven indien deze feature niet als omgevingsvariabele in het model aanwezig is. Op de y-as wordt de belangrijkheid van de features weergegeven. Hoe hoger de feature zich bevindt op de y-as, hoe belangrijker de feature is.

Tabel A.1: Geselecteerde features per regressietechniek en per testfold 1, 2, 3 of 4 aangeduid met 'x'. In deze tabel worden enkel deze features weergegeven die in één van deze regressietechnieken voorkwamen.

Feature (Straal buffer)	Forward stepwise selection				Backward stepwise selection				Lasso			
	1	2	3	4	1	2	3	4	1	2	3	4
Afstand tot dichtstbijzijnde tertiaire weg	x	x		x		x			x	x	x	x
Afstand tot dichtstbijzijnde dienstweg		x										x
Afstand tot dichtstbijzijnde residentiële weg		x										
Afstand tot dichtstbijzijnde kruispunt		x										
Afstand tot dichtstbijzijnde park							x					
Oppervlakte park (100 m)												x
Oppervlakte park (500 m)			x								x	x
Oppervlakte laagstedelijk gebied (100 m)												x
Aantal voertuigen (25 m)			x			x	x	x	x	x	x	x
Aantal voertuigen (50 m)											x	
Aantal vrachtwagens (25 m)	x	x		x					x	x		x
Aantal vrachtwagens (50 m)					x				x			
Aantal vrachtwagens (100 m)										x	x	x
Aantal vrachtwagens (300 m)												x
Aantal personenwagens (50 m)		x										
Aantal personenwagens (500 m)		x										