Spatial statistical modelling of urban particulate matter

Sibusisiwe A. Khuluse

Graduation committee

Chair and Secretary		
	Prof.dr.ir. A. Veldkamp	University of Twente
Supervisor		
	Prof.dr.ir. A. Stein	University of Twente
Co-supervisor		
	Prof.dr.ir. P. Debba	University of the Witwatersrand
Members		
	Prof.dr.ir. M.F.A.M. van Maarseveen	University of Twente
	Prof.dr.ir. R.J. Boucherie	University of Twente
	Prof.dr. E. Pebesma	Universität Münster
	Prof.dr.ir. P.J.M. Cluitmans	Eindhoven University

ITC dissertation number 305
ITC, P.O. Box 217, 7500 AE Enschede, The Netherlands

ISBN: 978-90-365-4370-5
DOI: 10.3990/1.9789036543705
Printed by: ITC Printing Department, Enschede, The Netherlands

SPATIAL STATISTICAL MODELLING OF URBAN PARTICULATE MATTER

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
prof.dr. T.T.M. Palstra,
on account of the decision of the graduation committee,
to be publicly defended
on Friday, July 21st, 2017 at 14.45 hrs

by

Sibusisiwe Audrey Khuluse
born on August 19th, 1985 in
Durban, South Africa

This dissertation is approved by:

Prof.dr.ir. A. Stein (supervisor)
Prof.dr.ir. P. Debba (co-supervisor)

To my late parents and my family

Summary

Chronic exposure to poor air quality poses a risk to respiratory health. The spatial distribution of air quality and socioeconomic vulnerability, however, is not equitable, as those most vulnerable often reside in areas with poorer air quality. In the Highveld region in South Africa, like in other rapidly developing urban regions in developing countries, air quality mapping for the purpose of investigating population exposure is important given the need for interventions to reduce the environmental impact on health. In this context statistical air quality mapping is challenging because of the sparsity of air quality monitoring network in space and time. The aim of this thesis was to assess the risk of exposure to poor air quality indicated by excessive ambient concentrations of PM_{10} and $PM_{2.5}$. This hinged on the development of methods to overcome data constraints in the form of high proportions of missing data per air quality station and the limited number of stations in the study area.

When the target variate is measured at few locations, a suitable and spatially extensive covariate can improve the reliability of predictions at unmeasured locations. The first objective was to compare ordinary kriging and model-based geostatistical methods, and to assess the significance of housing related factors as proxies for domestic emissions for spatial prediction of the annual PM_{10} exceedance rate at unmeasured locations. The exceedance threshold is the PM_{10} South African national air quality standard (NAQS) of 120 μg m^{-3} for average daily ambient concentrations. Four geostatistical methods were explored, two based on kriging and the others on the model-based geostatistical methods. A Poisson genaralized linear geostatistical model was considered because of the type of data, namely PM_{10} yearly exceedance counts from 36 air quality stations in the South African Highveld region for the period 2008 to 2012. The other models were the log-Gaussian geostatistical model, ordinary and external drift kriging. The spatial patterns of the PM_{10} NAQS exceedance rate, namely the location of hot-spots, were similar for kriging and the generalized linear geostatistical models. All four models were biased upwards. The relative accuracy of predictions to the actual data was highest for ordinary kriging as compared to the log-Gaussian and Poison models without covariates. External drift kriging predictions were more precise as compared to the model-based alternatives with covariates. Predictions from models with covariates were higher in areas where the density of informal

dwellings was higher. The Poisson model performed better than the log-Gaussian model in terms of prediction accuracy at test sites if the covariate was considered, otherwise they were similar. Kriging was superior in terms of prediction accuracy at test sites. From the three covariates considered, namely household biofuel use for cooking, heating and housing informality, it was housing informality that was statistically significant. Housing informality coincides with household use of biofuels, especially for heating, and being located close to industrial areas.

The deterioration of air quality in urban areas as a result of fugitive dust was explored because of the presence of mines, mine residue deposits, unpaved roads and agricultural fields in the study region. The second objective was to determine if land cover could be statistically related to observed PM_{10} and be used as a covariate to improve the reliability of PM_{10} predictions at locations without air quality stations. In the absence of readily available land cover data, high resolution SPOT 6 images were obtained for land cover classification. An ensemble maximum likelihood pixel-based land cover classifier was developed with five primary classes, namely water, bare soil, vegetation, built-up and a mixed class for pixels where there was difficulty in separating bare soil and degraded grass. The ensemble classifier which was based on iterative training enabled inclusion of information on known sources of variability which contribute to difficulties in classifying bare soil in the study region. These sources of variability were mainly soil colour tones due to variation in soil types. Overall accuracy of the classifier in terms of the Kappa index was 0.78. Various landscape features affect the dispersion and sedimentation of dust particles and as such a statistical relationship between ambient concentrations of PM_{10} and a factor for land cover composition was sought. Firstly, a k-means cluster analysis was used to derive homogenous land cover groups in neighbourhoods (within 4 km radius) of air quality stations that could be related to observed PM_{10} concentrations. Secondly, average PM_{10} calculated for days where wind speeds were conducive for dust emissions were related with a factor for land cover composition group in a varying intercepts regression model, where the factor was found to be a significant covariate. Therefore, land cover data can be processed into a suitable covariate for improved prediction of PM_{10} at locations without air quality monitoring stations in spatially sparse networks.

High quality monitoring data are important, but data from air quality stations often suffer from substantial incompleteness. In sparse monitoring networks imputation of missing pollutant data is favourable compared to discarding a station's record or analyzing data of low coverage. With multiple imputation missing values are imputed and the uncertainty associated with the imputations is quantifiable. The third objective was to develop a bootstrap regression multiple method to multiply impute missing meteorological and pollutant values for each air quality station. The method leverages on the availability of better quality meteorological data from nearby weather stations to impute multiple values for each missing relative humidity, temperature, wind speed and direction value. Subsequently, NO_2, SO_2 and

eventually PM_{10} and $PM_{2.5}$ were imputed based on the completed meteorological datasets. Regression imputation models were customized for each variable, such as circular regression for wind direction and log-transformation with inclusion of a wind intensity indicator for wind speed. Inference was based on generalized least squares to incorporate first order autoregressive residual structure to account for temporal autocorrelation. To avoid overstating the precision, regressions were performed sequentially and at each stage imputations were drawn from the Gaussian predictive distribution parameterized by the deterministic predicted value and the prediction standard error, thus incorporating uncertainty into the imputed values including errors propagating from meteorological imputations to imputed pollutant values. Using meteorological, gaseous pollutant and seasonal factor variables as covariates resulted in the preservation of seasonal patterns in imputed data. When the bootstrap regression imputation method was compared with the approximate Bayesian bootstrap (ABB) method, ABB imputations reverted to the mean, had reduced variability and lacked seasonal structure. Overall, the bootstrap regression multiple imputation method resulted in improved imputation quality of pollutants and meteorological variables compared to the ABB.

The fourth objective was to map the risk of exposure to poor air quality, integrating hazard probabilities with indicators of population at risk and inability of exposed communities to cope with the adverse effects of poor air quality. Hazard probabilities were defined as annual average concentrations of $PM_{2.5}$ and PM_{10} exceeding specific regulatory standards, namely 25 μg m^{-3} for $PM_{2.5}$ and 50 μg m^{-3} for PM_{10}. They were obtained using conditional simulations based on spatiotemporal kriging with external drift. Covariates and a joint spatiotemporal covariance function solved the problem of spatiotemporal sparsity of air quality data. Covariates included land cover composition and population counts to account location specific properties of pollutant emissions and dispersion. Exceedance probabilities for PM_{10} and $PM_{2.5}$ were high in central and southern parts of Gauteng and the main towns in the Highveld priority air-shed in Mpumalanga. A composite spatial indicator for social vulnerability was developed using geographically weighted principal components analysis. High social vulnerability was indicated for the south-eastern parts of Mpumalanga, characterized by the prevalence of child-headed households, insufficiency of access to basic services such as piped water for residential use and routine waste collection. Areas marked by moderate to high social vulnerability in Gauteng were characterized by the prevalence of housing informality, female household leadership (mostly single income households) and immigration. Combining the three risk dimensions resulted in high risk of exposure to excessive ambient PM_{10} concentration throughout Gauteng and Mpumalanga. For $PM_{2.5}$, small areas with low to medium risk of exposure to excessive ambient concentrations occurred away from the major towns of Mpumalanga and in protected areas towards the periphery of Gauteng. In Gauteng, $PM_{2.5}$ risk was highest in the city region. These retrospective risk maps can be used to initiate detailed investigations into the human and housing conditions in high risk areas for confirmation to

inform mitigation efforts.

To summarize, this dissertation provides a framework for air quality risk mapping, contributing specific methods to improve the quality of the data including the integration of ancillary data from disparate sources.

Samenvatting

Chronische blootstelling aan slechte luchtkwaliteit is een bedreiging voor de gezondheid van de luchtwegen. De ruimtelijke spreiding van luchtkwaliteit en sociaal-economische kwetsbaarheid, is echter onrechtvaardig, omdat de meest kwetsbaren vaak in gebieden met de slechtste luchtkwaliteit wonen. In de Highveld regio in Zuid-Afrika, net als in andere zich snel ontwikkelende stedelijke gebieden in ontwikkelingslanden, is het karteren van luchtkwaliteit belangrijk voor het onderzoek naar blootstelling van de bevolking gezien de behoefte aan interventies voor de terugdringing van de milieu-efecten op de gezondheid. In dit verband is het statistisch karteren van luchtkwaliteit een uitdaging vanwege de beperktheid van het luchtkwaliteitmeetnet in ruimte en tijd. Het doel van dit onderzoek was om het risico van blootstelling aan slechte luchtkwaliteit te schatten, met name naar buitensporige omgevingsconcentraties van PM_{10} en $PM_{2.5}$. Het proefschrift richt zich op de ontwikkeling van methoden om de beperkingen van de gegevens te overwinnen die bestaan uit hoge percentages ontbrekende gegevens voor ieder luchtkwaliteitmeetpunt en het beperkte aantal meetpunten in het studiegebied.

Wanneer de doelvariabele wordt gemeten op enkele locaties, dan kan een geschikte en ruimtelijk uitgebreide covariabele de betrouwbaarheid van de voorspellingen meten op locaties zonder meetpunt. De eerste doelstelling was om ordinary kriging en model-based geostatististische methoden te vergelijken en de significantie te bepalen van factoren die gerelateerd zijn aan huisvesting als proxies voor nationale emissies voor de ruimtelijke voorspelling van de jaarlijkse mate van overschrijding van PM_{10} op ongemeten locaties. De overschrijdingsdrempel is de Zuid Afrikaanse nationale luchtkwaliteit standaard national (NAQS) voor PM_{10}, gelijk aan 120 μg m^{-3} voor de gemiddelde dagelijkse omgevingsconcentratie. Vier geostatistische methoden zijn verkend, twee zijn gebaseerd op kriging en de andere twee zijn model-based geostatistische methoden. Een Poisson gegeneraliseerd lineair geostatistisch model is beschouwd vanwege het type gegevens, namelijk het jaarlijkse aantal PM_{10} overschrijdingen van 36 luchtkwaliteitsmeetpunten in de Highveld regio gedurende de periode van 2008 tot 2012. De andere modellen waren het log-Gaussische geostatistische model, ordinary kriging en externe drift kriging. De ruimtelijke patronen van de mate van overschrijding van de NAQS PM_{10} standaard, namelijk de plaats van hot-spots, waren hetzelfde voor kriging en de generaliseerde lineaire geostatistische

modellen. Alle vier modellen hadden een positieve onzuiverheid. De relatieve nauwkeurigheid van de voorspellingen in vergelijking met de meetwaarden was het hoogst voor ordinary kriging in vergelijking met de log-Gaussische en de Poison modellen zonder covariabelen. Voorspellingen met externe drift kriging waren preciezer in vergelijking met de model-based alternatieven met covariabelen. Voorspellingen met modellen met covariabelen waren hoger in gebieden waar de dichtheid van informele onderkomens hoger was. Het Poisson model gaf betere resultaten dan het log-Gaussische model in termen van de nauwkeurigheid in voorspellingen op test locaties als de covariabele hierbij betrokken was, anders waren ze vergelijkbaar. Kriging was beter in termen van de de nauwkeurigheid van de voorspellingen op test locaties. Wat betreft de drie beschouwde covariabelen, namelijk het gebruik in huishoudens van biobrandstof voor koken, verwarming en informaliteit van wonen, was de laatste statistisch significant. Informaliteit van wonen valt samen met het gebruik in huishoudens van biobrandstof, in het bijzonder voor verwarming, en met een positie dicht bij een industriegebied.

De achteruitgang van luchtkwaliteit in stedelijke gebieden als gevolg van stofwinden is verkend vanwege de aanwezigheid van mijnen, opslagplaatsen van mijnafval, ongeplaveide wegen en landbouwgronden in het studiegebied. De tweede doelstelling was om te bepalen of landbedekking statistisch gekoppeld kon worden aan waargenomen PM_{10} en gebruikt kon worden als een covariabele om de betrouwbaarheid te verhogen van PM_{10} voorspellingen op plaatsen zonder luchtkwaliteitmeetpunten. In de afwezigheid van gemakkelijk beschikbare landbedekkinggegevens zijn hoge resolutie SPOT 6 beelden gebruikt voor een landbedekkingclassificatie. Een ensemble maximum likelihood landbedekkingclassificatie op pixel basis is ontwikkeld met vijf primaire klassen, te weten water, kale grond, vegetatie, bebouwing en een gemengde klasse voor pixels waar het moeilijk was om kale grond te scheiden van gedegradeerd gras. De ensemble classificatie die gebaseerd is op een iteratieve verbetering bood de mogelijkheid om informatie op te nemen van bekende bronnen van variabiliteit die bijdragen aan moeilijkheden om kale grond te classificeren in het studiegebied. Deze bronnen van variabiliteit waren hoofdzakelijk tonen van bodemkleur op grond van verschillende bodemtypes. De totale nauwkeurigheid van de classificatie in termen van de Kappa index was 0.78. Verschillende landschappelijke kenmerken hebben een effect op de dispersie en sedimentatie van stofdeeltjes en om die reden is een statistische relatie gezocht tussen omgevingsconcentraties van PM_{10} en een factor voor de samenstelling van de landbedekking. Eerst is een k-gemiddelden clusteranalyse gebruik om homogene landbedekkingsgroepen in de omgeving (binnen een straal van 4 km) van luchtkwaliteitsmeetpunten die gerelateerd kunnen worden aan waargenomen PM_{10} concentraties. Als tweede is de gemiddelde PM_{10} waarde, berekend op dagen dat windsnelheden aanleiding gaven voor stof emissies, gerelateerd aan een factor voor de landbedekkings samenstelling groep in een regressie model met variabele intercepten. De factor bleek een significante covariabele te zijn. Om die reden zijn de landbedekkingsgegevens vertaald in een geschikte covariabele voor een verbeterde voorspelling van PM_{10} op plaatsen zonder luchtkwaliteitsmeetpunten in een

ruimelijk dun netwerk.

Het monitoren van gegevens van hoge kwaliteit is belangrijk. Gegevens van luchtkwaliteitsmeetpunten lijden vaak aan een substantiele onvolledigheid. In dunne monitoring netwerken is het toevoegen van missende verontreinigingsgegevens gunstig in vergelijking met het weglaten van die waarnemingen of het werken met de beperkte gegevens. Met meervoudige methoden zijn de ontbrekende waarden toegevoegd en is de onzekerheid gekwantificeerd die hieraan verbonden is. De derde doelstelling was het ontwikkelen van een meervoudige bootstrap regressie methode om op een meervoudige basis ontbrekende meteorologische en verontreingingswaarden toe te voegen voor ieder luchtkwaliteitsmeetpunt. De methode is gebaseerd op de beschikbaarheid van meteorologische gegevens bij nabij gelegen weerstations van een betere kwaliteit, om meervoudige waarden toe te voegen voor iedere ontbrekende waarde aan relatieve vochtigheid, temperatuur, windsnelheid en windrichting. Vervolgens zijn NO_2, SO_2 en tenslotte PM_{10} en $PM_{2.5}$ toegevoegd op basis van de aangevulde meteorologische bestanden. Regressie modellen voor toevoeging zijn toegesneden op iedere variable, zoals een circulair regressie model voor windrichting en een log-transformatie met toevoeging van een indicator voor windintensiteit voor windsnelheid. Het schatten van parameters is gebaseerd op gegeneraliserde kleinste kwadraten om de eerste orde autoregressieve residu structuur op te nemen die compenseert voor autocorrelatie in de tijd. Om te voorkomen dat de nauwkeurigeheid te hoog wordt ingeschat, zijn regressies sequentieel uitgevoerd. Op ieder moment zijn de toevoegingen getrokken uit de Gaussische voorspellende verdeling, met parameters die afgeleid zijn van de deterministische voorspelde waarde en de standaardfout van de voorspellingen. Op deze manier is de onzekerheid meegenomen in de toegevoegde waarden, alsmede de voortplantingsfouten van de meteorologische toevoegingen in de ingebrachte verontreinigingswaarden. Het gebruik van meteorologische variabelen, verontreinigende gassen en seizoensgebonden variabelen als covariabelen resulteerde in seizoensgebonden patronen in de ingebrachte gegevens. Bij het vergelijken van de toevoegmethode gebaseerd op bootstrap regressie met de benaderde Bayesiaanse bootstrap (ABB) methode, ontdekten we dat de ABB toevoegingen een negatief effect hadden op het gemiddelde, een lagere variabiliteit hadden en dat het ontbrak aan seizoensstructuur. Alles bij elkaar genomen resulteerde de meervoudige toevoegmethode gebaseerd op bootstrap regressie in een verbeterde kwaliteit van de zowel de verontreinigings als de methodologische variabelen in vergelijking met ABB.

De vierde doelstelling was het karteren van het risico voor blootstelling aan slechte luchtkwaliteit waarbij risico kansen werden geintegreerd met indicatoren voor de populatie die hieraan blootstond en met de onmacht van de blootgestelde gemeenschappen om om te gaan met de ongunstige effecten van slechte luchtkwaliteit. Risicokansen zijn gedefinieerd als de jaarlijkse gemiddelde concentraties van $PM_{2.5}$ en PM_{10} die de specifieke drempelwaarden overschrijden, namelijk 25 μg m^{-3} voor $PM_{2.5}$ en 50 μg m^{-3} voor PM_{10}. Deze zijn verkregen via conditionele simulaties gebaseerd op

spatiotemporele externe drift kriging. Covariabelen en een multivariate spatiotemporele covariantie functie losten het probleem op van het ontbreken van voldoende spatiotemporele luchtkwaliteitsgegevens. Covariabelen waren o.a. landbedekkingssamenstelling en bevolkingsaantallen om rekening te houden met locatie specifieke eigenschappen van de emissie en dispersie van verontreinigende stoffen. Overschrijdingskansen voor PM_{10} en $PM_{2.5}$ waren hoog in centraal en zuid Gauteng en in de belangrijkste steden in de Highveld priority air-shed in Mpumalanga. Een samengestelde ruimtelijke indicator voor sociale kwetsbaarheid is ontwikkeld die is gebaseerd op geografisch gewogen hoofdcomponenten analyse. Hoge sociale kwetsbaarheid komt voor in de zuid-oostelijke delen van Mpumalanga, die gekarakteriseerd worden door de aanwezigheid van huishoudens die door kinderen worden geleid en onvoldoende toegang hebben tot basale diensten zoals waterleidingen voor residentieel gebruik en het regelmatig ophalen van afval. Gebieden die gemiddeld tot hoog sociaal kwetsbaar zijn in Gauteng worden gekarakteriseerd door het voorkomen van informele bewoning, huishoudens die geleid werden door een vrouw (vooral huishoudens met een enkel inkomen) en immigratie. Het combineren van de drie risico dimensies resulteerde in een hoog risico op blootstelling aan buitensporige PM_{10} concentraties in Gauteng en Mpumalanga. Voor $PM_{2.5}$ vomd overschrijding plaats in kleinere gebieden met lage tot gemiddelde risico's op blootstelling aan exessieve omgevingsconcentraties, op grote afstand van de belangrijke steden in Mpumalanga en in beschermde gebieden richting de grens met Gauteng. In Gauteng was het $PM_{2.5}$ risico het hoogste in het stedelijk gebied. Deze retrospectieve risico kaarten kunnen gebruikt worden om meer gedetailleerd onderzoek te starten naar de menselijk en bewoningscondities in gebieden met een hoog risico om verlichtingsmaatregelen sturen.

Samenvattend biedt dit proefschrift een kader voor het karteren van het risico van luchtkwaliteit. Het geeft specifieke methoden om de kwaliteit van de gegevens te verbeteren, door met name ondersteunende gegevens op te nemen van wezenlijk verschillende bronnen.

Acknowledgments

> "After climbing a great hill, one only finds that there are many more hills to climb." –Nelson Mandela

From the start I knew that the climb to the summit of my PhD would be challenging, but I did not anticipate just how fraught with difficulties it would be. However, the wells dug by difficulty and pain can be filled with deep appreciation, love and renewed strength. Faith, perseverance and the support of others brought me to this point.

Words are not enough to express my gratitude to Prof Alfred Stein for believing in my abilities, patience and for pushing me to the threshold of my mind. I have deep admiration for your creativity, tenacity, efficiency and the passion that you have for your work. I believe that the lessons I learned during the period I was under your guidance will continue to shape my career in statistical research and practice. I hope to continue to learn from you beyond the PhD.

I am deeply grateful to Prof Pravesh Debba who has supported my development as a statistician from the time I was an undergraduate student until this point. Thank you for introducing me to the area of spatial statistics. I am deeply grateful for the opportunities that you alerted and supported me through, and especially for introducing to ITC and Prof Stein. As my manager at the CSIR, you did your best to ensure that I had some time to work on my PhD. I admire your focus and dedication to personal development as well as the development of others.

As a sandwich program student, I am grateful for all the administrative assistance I received at ITC, especially the assistance of Teresa Brefeld and Loes Colenbrander that made my student life easier. I am grateful for the financial support from the Nuffic Netherlands Fellowship Program and the CSIR which made my studies possible. I am deeply grateful for the management and administrative support I received from the CSIR, especially from my research group leaders Reneé Koen and previously Dr Chris Elphinstone. Thank you to my CSIR colleagues Jenny Holloway, Nontembeko Dudeni-Tlhone, Paul Mokilane, Dr Abel Ramoelo, Dr Moses Cho, Nosizo Sebake, Dr Sonali Das and Thandulwazi Magadla for your encouragement throughout

my studies.

To all my friends, thank you for believing in me and for being that warm, familiar place I could retreat to when I needed replenishment. Special thank you to my girls Xoli, Nokuthula, Nolwazi, Nontembeko, Thabile, Nombuso and Nombali for your support through personal traumas and for lightening my heart with your presence and laughter. I am grateful for the comradeship of friends that I made at ITC, especially Dr Nthabiseng Mohlakoana for her constant presence and support.

To my brother, thank you for inspiring me to work harder and for being a reminder of who I am in essence. To my extended family thank you for your support through difficult times which encouraged me to carry on with my studies. To Zama, thank you for caring for my son, whenever I had to be in the office late and on weekends to work on my thesis. To my parents, on my first day at university I had imagined seeing you beaming with pride on the day that I would eventually graduate with a PhD. It was not to be, but I am faithful that you are proud of me wherever your souls are. Thank you for your love and for teaching me to always pursue my dreams.

To my husband, thank you for being my pillar of strength. You always listened patiently to my ideas, doubts and disappointments. You remained supportive and loving even during my periods of absence from home. Finally, to my son, Olwethu thank you for teaching me to appreciate the smallest pleasures of life and invigorating me to work harder.

"The important thing is to keep the fire in your heart and be strong to overcome hard moments"
–Paulo Coelho

Contents

List of Figures

List of Tables

Introduction 1

1.1 Air quality – background and data issues

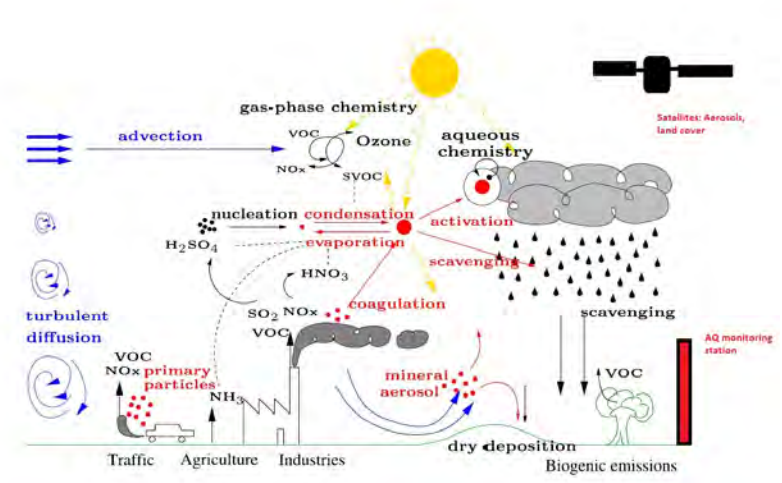


Figure 1.1: Sketch of atmospheric processes and pollutants on a regional landscape (Adapted from Sportisse (2009))

Airborne particulate matter is a mixture of solid and liquid matter suspended in air. Particulate matter (PM) can be either emitted directly into ambient air (primary) or be formed from nucleation and condensation of gas molecules (secondary) such as nitrogen oxides, sulfur dioxide and volatile organic compounds as shown in Figure 1.1. Primary PM from mechanical processes usually forms the coarse fraction (> 1 μm in aerodynamic diameter), while that from combustion processes (secondary PM) is usually smaller (Sportisse, 2009). Small particles can collide with other particles especially larger ones (coagulation), resulting in an increase in coarse mode PM. Coagulation is dependent on temperature and relative humidity (Sportisse, 2009). Temperature influences the stochastic motion of particles in the air, with higher temperatures increasing the likelihood of coagulation. Temperature gradients lead to dispersion of particles from high to low concentration areas, the flow is from high to low temperatures. Lower temperatures have been noted to be one of the determinants for the high concentrations observed in winter. Relative humidity affects particle growth/size especially for inorganic particulate matter species which are soluble (Sportisse, 2009). For such species, high relative humidity leads to larger liquid aerosols while these particles revert to the solid particle form when relative humidity is low (Sportisse, 2009).

Removal processes of PM from the atmosphere include *dry deposition* (both sedimentation and impaction) and *wet scavenging*. Sedimentation is particle removal when gravitational force acting on it overcomes air resistance, which is proportional to the ratio of the mass to the aerodynamic diameter. Therefore larger particles (≥ 10 μm) have higher sedimentation rates, corresponding to

short atmospheric residence time and transport distances. Impaction refers to removal by collision with a surface such as a wall, a feature that is dominant in urban landscapes. Both surface roughness and wind play an influential role in this regard. Wet scavenging can be described as washing out of PM by precipitation. Therefore residence time for particulate matter is a function of both particle diameter and wet scavenging. Residence time decreases with increasing precipitation intensity. Particle size is also associated to residence time. Coarse particles also have short residence time, between an hour to a few days (much less than a week) due to sedimentation and impaction. The fine particle mode is more stable, with residence times of up to 10 days (Sportisse, 2009).

It can be deduced that inaccuracy in PM concentration estimation over urban areas is attributable to the complex mixture of emission sources as well as differences in spatial and temporal scales over which the various components of PM and influential factors have impact (Jerrett et al., 2005; Beelen et al., 2009; Zwack et al., 2011; Sportisse, 2009). Urban areas, especially in developing countries are heavily affected by exposure to dust produced locally as well as transport of dust particles from eroded agricultural landscapes on the periphery (Ozer, 2006; Goudie, 2009; De Longueville et al., 2010; Chervenkov and Jakobs, 2011). Crustal particles aerially suspended as a result of vehicle-road surface interaction are considered as dust and are typically dispersed over shorter distances < 5 km (Sportisse, 2009).

Issues with in situ air quality monitoring data

Air quality monitoring stations provide time series data of mass concentrations of particulate matter and co-pollutants. However, there are usually only a few stations for monitoring within large geographic areas (Elliott et al., 2000; Millar et al., 2010a). This means that in urban environments PM concentrations close to a monitoring station can be different to concentration levels at targeted micro-environments like homes and schools where individuals are exposed (Elliott et al., 2000; Millar et al., 2010a,b). Therefore, common practice is to spatially interpolate PM data from all monitoring stations in that region instead of relying on PM concentrations from a station that may be too distant to reliably make inferences about health effects of exposure to PM. Another issue is that some stations do not have measurements for $PM_{2.5}$ (fine particles with aerodynamic diameter that is less than 2.5 μm) because monitoring of $PM_{2.5}$ came into effect later than coarse PM and other pollutants (Christakos et al., 2005). To overcome this, a popular technique is to estimate $PM_{2.5}$ from PM_{10} (particles with aerodynamic diameter that is less than 10 μm) using the ratio $PM_{2.5}/PM_{10}$ calculated from stations which monitor both quantities (Norman et al., 2007; Christakos et al., 2005). The ratio method takes advantage of the property that $PM_{2.5}$ is a component of PM_{10} and as such the two are expected to be strongly correlated. A problem with this ratio approach is it's determinism, where a constant value is used for large regions, despite both PM components being variable in time and space (Christakos et al., 2005). An alternative is to take advantage of the

fact that $PM_{2.5}$ consists of mostly fine particles produced from combustion processes. Gaseous pollutants such as CO, SO_2 and NO_2 are also products of incomplete combustion and may be used in predicting $PM_{2.5}$ at sites where they are measured (Christakos et al., 2005; Sportisse, 2009).

1.2 Consequences and socioeconomic determinants of exposure to poor air quality

The effects of chronic exposure to particulate matter pollution on common chronic respiratory diseases (CCRD) mainly asthma and chronic obstructive pulmonary diseases (COPD) is of interest in air pollution epidemiology due to the high global burden of these diseases (Pope III, 2000; Pope III and Dockery, 2006; WHO, 2007; Millar et al., 2010a). Exposure to poor air quality is associated with the risk to respiratory health because of the involuntary nature of ambient air inhalation. Toxicological findings indicate that $PM_{10-2.5}$ deposits mainly in the nose, pharynx, trachea and bronchiolar regions in Figure 1.2 and that the rates for males and females are similar, while infants have higher deposition rates in the nasal region. Generally, deposition of coarse particles is much lower in the bronchial region as most are deposited in the extra-thoracic tract, while fine particles have higher likelihood of deposition in the alveolar region (Pope III and Dockery, 2006; Lazaridis, 2011). Particles deposited in the extra-thoracic tract are cleared through the production of mucus with observable symptoms being coughing and sneezing. Chronic particle deposition in the bronchi and alveoli can lead to a sustained state of oxidative stress and inflammation resulting in the death of endothelial cells that line them (Li et al., 2003; Churg et al., 2003; Schlesinger et al., 2006; Chirino et al., 2010). In the bronchi this inflammation induces symptoms such as chest tightness and wheezing, while damage to alveolar walls leads to increased lung volume and forced expiration (coughing) that is characteristic of COPD (Churg et al., 2003).

The higher likelihood of fine particle deposition in the lower respiratory tract in early childhood and more so in women compared to men concurs with the observation of high rates of non-infectious lower respiratory tract illnesses in early childhood and higher rates of adulthood asthma and COPD in women than in men (WHO, 2007). These gender and age gradients in asthma and COPD prevalence are more pronounced in communities of lower socioeconomic status in low and middle income countries (WHO, 2007; Ehrlich and Jithoo, 2006). This has been attributed to exposure to indoor particles from combustion of fossil fuels for cooking and heating, under-diagnosis, under-treatment and inadequate access to emergency medical care (WHO, 2007; Ehrlich and Jithoo, 2006). The contribution of residency in hazardous areas in terms of air pollution such as mining and other industrial waste sites to the burden of respiratory diseases needs more attention. Case in point is lack of published research on community effects of wind-blown dust from gold mine residue stockpiles which are rich in uranium and other toxic

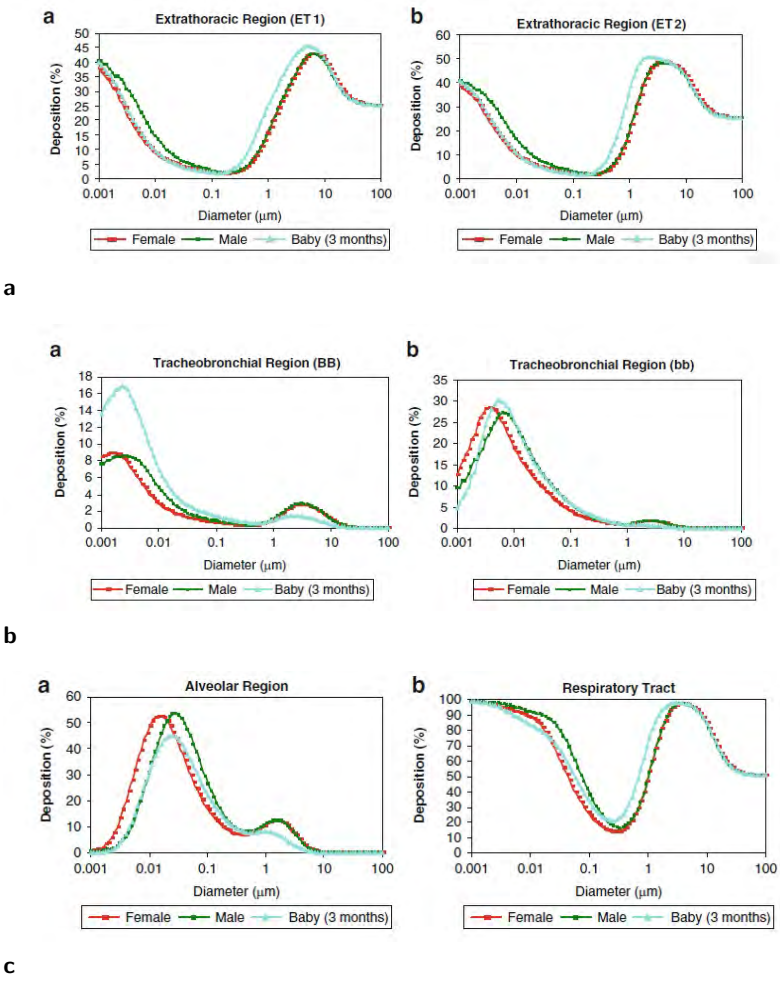


Figure 1.2: Differential deposition rates of particles in respiratory systems of males, females and infants (Source: Lazaridis (2011))

elements (Kneen et al., 2015). This is important in light of findings from long-term occupational health studies of miners, where apart from pneumoconiosis and silica tuberculosis, exposure to dust during mining activities is associated with development of chronic airway obstruction (Hnizdo and Vallyathan, 2003; Ehrlich and Jithoo, 2006; Rees and Murray, 2007).

Another aspect to consider when assessing exposure to PM is that particles from outdoors can enter indoor environments as a result of poor housing infrastructure, a prevalent challenge in developing countries. Settlement types and socioeconomic information at neighborhood scale is important for this purpose. Insight from what is known about the biological effects and gradients relative to the characteristics of the population exposed provide motivation

for integration of official statistics on socioeconomic conditions at small area spatial scales with neighbourhood/settlement information extracted from satellite imagery to determine both elements at risk as well as differentials in vulnerability amongst those elements at risk (Ebert et al., 2009; Millar et al., 2010a; Schwartz et al., 2011). Incorporating stochastic vulnerability of the population at risk remains a challenge in air quality exposure and risk assessment, given that until recently in statistical approaches to risk, the vulnerability component was not considered (Schwartz et al., 2011).

1.3 Urban air quality in a developing country context

Public health impacts of poor air quality have been under-recognized in the South African policy sphere as the focus has been primarily on reducing the burden of HIV, AIDS and Tuberculosis on society. Six metropolitan areas were considered previously where PM_{10} and $PM_{2.5}$ were estimated to have caused 3.7% of national mortality caused by cardiopulmonary diseases in adults over the age of 30 years and 1.1% of mortality from acute respiratory infections in children aged five years and younger (Norman et al., 2007).

The study area considered is approximately $67\,000\,\mathrm{km}^2$ and lies between latitudes $(-25.00; -27.00)$ and longitudes $(27.00; 30.00)$. It consists of the central part of a high plateau region of South Africa, that encapsulates the Gauteng province which covers an area of approximately $17\,000\,\mathrm{km}^2$. According to the 2016 Community Survey results by Statistics South Africa the population in Gauteng was estimated at 13.4 million, an increase of nearly 9% from the census 2011 population count of 12.3 million. A part of the Mpumalanga province made of the Gert Sibande and Nkangala district municipalities beyond the eastern boundary of Gauteng is also part of the study area shown in Figure 3.1.

The city region of Gauteng is an agglomeration of three of the country's six metropolitan municipal areas. The city of Johannesburg (JHB) municipality covers an area of approximately $1\,645\ \mathrm{km}^2$, while the Tshwane municipality north of JHB has an approximated area of $6\,368\ \mathrm{km}^2$. East of JHB is the Ekurhuleni municipality, with an area of approximately $1\,925\ \mathrm{km}^2$. Gauteng is an industrial hub of South Africa, with a high density of industries, including mining and chronic traffic congestion along major routes within the province. There are 31 airports in the province, including the OR Tambo international airport. Gauteng has three operational coal-based power stations, while neighbouring Mpumalanga has eleven, accounting for approximately 80% of the country's coal-produced electric power. Key regional anthropogenic sources of PM pollution, therefore, include industrial, domestic and vehicular combustion of fossil fuels. Domestic combustion of coal, wood and paraffin has in previous studies been associated with locations of poorer socioeconomic conditions in less affluent areas in South Africa (Norman et al., 2007; Wright et al., 2011). The severity of haze episodes due to domestic fuel burning has declined since electrification started in 2001. Domestic combustion of

Figure 1.3: A Google Earth image that has been adapted to show the study area with air quality monitoring stations from Tshwane, Johannesburg and Ekurhuleni municipalities as well as the Vaal Triangle and Highveld Priority airsheds from the Department of Environmental Affairs

biofuels however has not been completely eradicated in informal settlements due to being unelectrified in some cases and lack of finance to purchase enough for household demand in other cases. Housing infrastructure is an important factor in studying air pollution in this region not only with regards to domestic use of fossil fuels, but also because formal low-cost and informal housing tends to be located in areas close to high pollution industries and mine waste sites. Further, there is marked heterogeneity in housing infrastructure in the region with residential buildings varying from formal up-market suburban dwellings and flats, four-roomed and other low cost township houses, to backyard shacks in townships and informal settlements.

The ground-level air quality monitoring network for this study consists of 37 stations, of which 17 have $PM_{2.5}$ measurements. In the JHB Metropolitan area there are 8 monitoring stations, and only one station (Newtown), located in the city centre, has $PM_{2.5}$ measurements. The overall observation period is from 2006 to 2015, with two stations being operational from 2004 and some were temporarily decommissioned in 2010. Various meteorological and atmospheric chemistry parameters are measured at each site; those chosen for this study include: temperature, wind speed and direction, relative humidity, $PM_{2.5}$, PM_{10}, NO_2 and SO_2. For each station the data are available at a daily level of aggregation. Meteorological data are also available from local weather stations. There are 12 such weather stations within the study area.

There is no publicly available emissions inventory for this area. Dust from

mine dumps is a local source of coarse particulate matter around Johannesburg and the Ekurhuleni metropolitan areas. Incidents of dust 'storms' are recorded and there are dust monitoring gauges along the Witwatersrand mining belt, but access to these data was not granted by the custodians. Therefore, satellite imagery (SPOT 6) will be used in this study to identify mine dumps and other sources of fugitive dust. Unpaved roads which are highly prevalent in informal settlements in the area also contribute to dusty conditions.

1.4 Research statement and objectives

The aim of this research is to statistically map PM_{10} and $PM_{2.5}$ for the purpose of population exposure and risk assessment (Figure 1.4). Core to this is the search for solutions to problems of sparseness of air quality data. These solutions are methods that are developed for the integration of data from disparate sources and of differing quality to yield valid pollutant maps and make inference about the populations at risk of exposure to poor air quality.

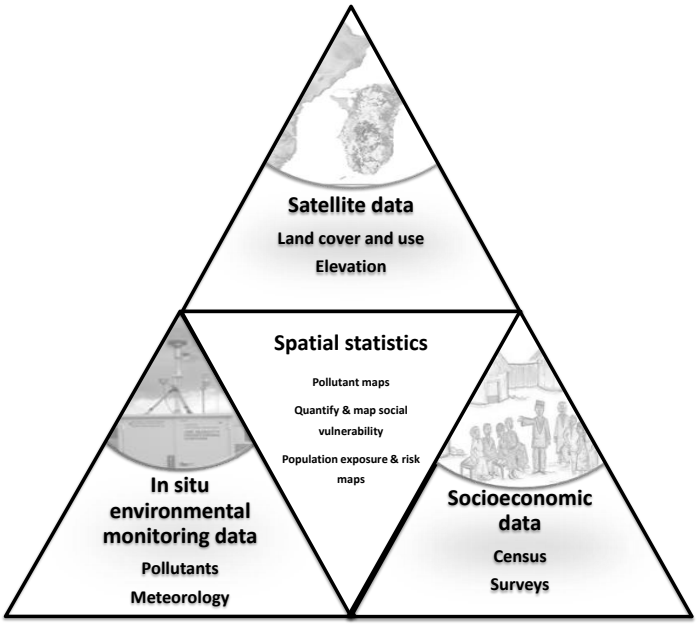


Figure 1.4: An overview of various types of data considered important for air quality exposure and risk assessment and how they will be integrated in this research study using spatial statistical methods

1.4.1 First objective: Modelling considerations for statistical mapping of particulate matter in a large densely populated area with few monitoring stations

A map of PM_{10} showing locations where air quality thresholds are frequently exceeded is useful for mitigation and identification of populations whose health is at risk. Knowing factors that contribute to the formation of these hot spots serves a purpose in mitigation planning, but more importantly variables representative of these factors can be used to improve statistical mapping of PM_{10} when the sample of monitoring stations is small. The objective is to evaluate the performance of the generalized linear geostatistical models (GLGM) and kriging for mapping the threshold exceedance rate of PM_{10}. Additionally, the applicability of household census data as proxies of domestic sources of PM_{10} emissions for predicting PM_{10} at unmeasured locations is assessed. These methods are implemented on daily PM_{10} data from 36 air quality monitoring sites in the Highveld in South Africa for the 48 month period from September 2008 to August 2012. Domestic emission proxies of interest for this objective are extracted from the South African census 2011 small area dataset. These are percentages per small area of informal dwellings, household using solid biofuel energy for heating and for cooking.

1.4.2 Second objective: Use satellite imagery to identify fugitive dust sources of $\mathbf{PM_{10}}$

Bare ground including unpaved roads and mine residue deposits (or dumps) are dominant features and sources of health nuisance in the study region. Fugitive dust is defined as solid particulate matter suspended into the atmosphere through mechanical processes caused by wind and human activities. Typically natural sources of fugitive dust inland are bare soil and vegetation (spores and pollen) while there are many various anthropogenic sources of fugitive dust especially in urban areas of developing countries. Image classification of sources of fugitive dust particles is complicated by partial vegetation cover, mixture of bare soil with gravel and similarities between soil and building materials such as pavement bricks when the spectral resolution of the image is limited to few bands in the visible part of the spectrum. The objective here is to extend the maximum likelihood classifier for improved classification of bare soil by accounting for the variation in soil types. A sub-objective is to develop a method to process land cover class output into variables that can be related to average PM_{10} concentrations on days when local wind speeds are favourable to fugitive dust emissions. SPOT 6 images taken in March and April 2013 are used as well as wind-related averages calculated from daily PM_{10} data from 23 air quality monitoring stations within the Gauteng province in South Africa for period from September 2011 to February 2015.

1.4.3 Third objective: Imputation of missing air quality data

Data is available from 37 stations of the air quality monitoring network in the South African Highveld region. The challenge with this data is the high proportion of missing observations per station. Ignoring missing data and using annual air quality statistics as inputs in other models can lead to biased and unreliable parameter estimates. In this objective relationships between PM_{10}, NO_2, SO_2 and meteorological variables (relative humidity, temperature, wind speed and wind direction) are exploited using a bootstrap regression multiple imputation method to account for temporal variation and enable evaluation of uncertainty associated with imputing missing values. Part of this objective is to determine how data collected at nearby locations on common variables can be of use. Daily data from five air quality stations and their weather stations are considered for methodological development.

1.4.4 Fourth objective: Neighbourhood-level risk of exposure to high ambient concentrations of particulate matter

Exposure to even moderate concentrations of particulate matter leads to increased relative risks of cardiopulmonary morbidity and mortality. Those living in poor settlement conditions typically encounter various other socioeconomic issues which render them vulnerable. In this objective $PM_{2.5}$ and PM_{10} data will be integrated with data on population size and social vulnerability for inference on risks posed by exposure to poor urban air quality. In the first and second objective focus was on identifying covariates that would improve mapping of PM_{10} for a spatially sparse air quality network. Those covariates will be used together with annual averages for $PM_{2.5}$ and PM_{10} derived from completed data (missing values imputed) from the 37 air quality stations for the years 2008 until 2014. Population and social vulnerability data are extracted from the South African census 2011 small area dataset. Landscape characteristics for mapping $PM_{2.5}$ and PM_{10} are derived from the 2013–2014 South African national land cover dataset.

Statistically mapping the PM_{10} annual exceedance rate for a sparse air quality network

2

This chapter is based on papers: Khuluse-Makhanya S., Dudeni-Tlhone N., Holloway J., Schmitz P., Waldeck L., Stein A., Debba P., Stylianides T., Du Plessis P., Cooper A., Baloyi E., 2016. The applicability of the South African Census 2011 data for evidence-based urban planning. *Southern African Journal of Demography* 17(1), 67–132.

Khuluse-Makhanya S., Stein A., Debba P. Exploring housing informality and domestic solid biomass fuel use as predictors of the PM_{10} exceedance rate through kriging and generalized linear geostatistical models. Submitted to *South African Geography Journal*.

Abstract

An important objective in air quality mapping is to determine high concentration areas and to identify factors that contribute to their formation. This paper focuses on the annual average exceedance frequency of the South African PM_{10} standard (NAQS) of $120\,\mu g\,m^{-3}$. It compares classical kriging and model-based geostatistical methods, and assesses whether selected housing related factors are significant spatial predictors of the annual PM_{10} exceedance rate. Methods were implemented using PM_{10} yearly exceedance count data from 36 air quality stations in the South African Highveld region for the period 2008 to 2012. Selected predictors were percentage of households per small area that were living in informal dwellings (shacks) and percentage of households per small area using biofuels for heating and for cooking. The basis for this selection were exploratory analysis findings where higher concentrations of PM_{10} were found in high density residential areas marked by a high prevalence of informal dwellings. Informal dwellings percentage was found to be statistically significant as a predictor of the PM_{10} annual exceedance rate. All four models were biased upwards. Without covariates, the relative accuracy of predictions to the actuals was highest for ordinary kriging. Higher prediction accuracy was achieved with external drift kriging compared to the model-based alternatives with covariates. Predictions with models with covariates were higher in areas where the density of informal dwellings was higher. Overall, maps of the PM_{10} annual exceedance rate from all considered methods were similar in terms of location of high concentration areas and areas of lower exceedance rates, but kriging methods were better in terms of prediction accuracy.

Keywords: Urban planning, spatial data quality, exceedances, kriging, generalized linear geostatistical models

2.1 Introduction

The South African PM_{10} air quality standard stipulates that the daily average PM_{10} concentrations in an area should be below 120 μg m^{-3} and that this threshold should not be exceeded more than four times per year (RSA Government, 2009). Air quality monitors are installed with the purpose of monitoring compliance with these standards. Monitors, however, are stationed in specific locations. It is thus a challenge to estimate pollutant concentration surfaces over the whole region of interest. Geostatistical methods are useful for that purpose as they enable estimation of regional pollution surfaces from in situ air quality monitoring data.

The fundamental hypothesis in geostatistics is that observed values $Y(s_i)$ at locations s_i are a realization of an unknown spatial process $S(s_i)$. Inference therefore aims at estimation of parameters describing this hidden process, while prediction is aimed at obtaining values at unmeasured sites (Diggle and Ribeiro Jr, 2007). In kriging the unknown process is described by a regional mean trend $\mu(\mathbf{s}) = X(\mathbf{s})'\beta$ which is a deterministic function of covariates $X(\mathbf{s})$ and the residual spatial variation is described by a deterministic semivariance function (Bivand et al., 2008). In model-based geostatistics distributional assumptions are made about the unknown process, with inference focused on the conditional expectation $E[Y_i \mid S(s_i)] = \mu(s_i) + S(s_i)$ where the semivariance parameters are estimated with their corresponding uncertainty. The incorporation of knowledge about physical structure of the process being modelled gives geostatistics a key advantage over techniques such as inverse distance weighting that simply interpolates observed values (Diggle and Ribeiro Jr, 2007). In model-based geostatistics the main advantage is that likelihood and Bayesian inference can be pursued, unifying geostatistics with general and generalized linear statistical models and their extensions which are applied extensively in many fields of study (Banerjee et al., 2004; Diggle and Ribeiro Jr, 2007; Cressie and Wikle, 2011; Blangiardo and Cameletti, 2015).

In the case of pollutant concentrations where interest is in the spatial distribution of average concentrations, counts of exceeding regulatory thresholds, amongst other statistics, the log-Gaussian geostatistical model is commonly applied (van de Kassteele, 2006; Cocchi et al., 2007; Hamm et al., 2015). This is a special case of the generalized linear geostatistical model (GLGM) with an identity link function (Diggle and Ribeiro Jr, 2007). More generally for the GLGM, the conditional expectation is a non-linear function of the conditional mean $E[Y_i \mid S(s_i)] = h^{-1}(\mu(s_i) + S(s_i))$ where h is a link function. This accommodates a wider range of models such as Poisson log-linear models for count data. Alternatives such as Negative-Binomial models are also plausible when the spatial variation in exceedance counts is not adequately captured by the Poisson model because of over-dispersion (van de Kassteele, 2006). Similar efforts to account for the distributional assumptions in kriging can be done through regression kriging which involves multiple steps. In the first step, a generalized linear model is applied to the response variable and the covariates.

The second step subjects residuals from the first analysis to semivariogram modelling. As the third step, the predicted residual surface is added to the predicted mean surface to get the response surface. Poisson kriging for count data is based on this framework and was found to perform better in terms of prediction accuracy than ordinary kriging (Monestiez et al., 2005).

In this study we compared ordinary kriging (OK) and kriging with external drift (KED) as well as the log-Gaussian and Poisson generalized linear geostatistical models for mapping the average number of days per year that the South African PM_{10} national air quality standard (NAQS) of $120\,\mu g\,m^{-3}$ is exceeded. We assess whether housing related factors can be used as spatial predictors to improve mapping PM_{10} statistics given that the air quality monitoring network over the study area is sparse. We look specifically at variables that are characteristic of domestic fuel burning as an emission source motivated by exploratory finding of elevated PM_{10} concentrations on monitors located in areas with high prevalence of informal settlements. In Section 2.2 the study region is discussed with reference to sources of airborne particles, data used for the study and findings from preliminary analysis. Kriging and model-based geostatistical methods for mapping PM_{10} NAQS exceedance counts are described in Section 2.3. This is followed by results in Section 2.4 and a reflection on the findings of this study in Sections 2.5 and 2.6.

2.2 Materials

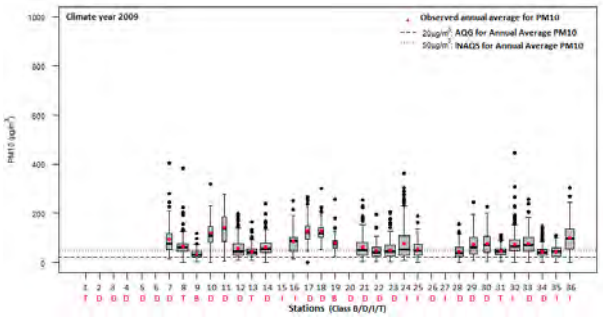
The study area as shown in Figure 3.1 consists of the Highveld region of South Africa with 36 air quality monitors from the Gauteng city region and parts of Mpumalanga province. Stations in Mpumalanga are typically close to or down-wind from industrial areas. Parts of Mpumalanga that are in the study area may be classified as rural, however the monitoring stations are located in towns and are therefore relevant for this study. Further, polluted air masses transported from the power generation regions may contribute to pollution levels observed within the Gauteng province. A common feature between the small towns of Mpumalanga and the large metropolitan centers of Gauteng is the incidence of informal and mixed township type settlements in close proximity to industrial areas. This is of interest for air quality because it has been observed in South Africa that domestic combustion of solid biofuels and coal is more prevalent in such areas (Balmer, 2007). The size of the monitoring network being 36 stations is small relative to the size and heterogeneity of the study area. Further, the limitation of monitoring air quality to specific sites results in what is defined as a preferential sample in geostatistics (Diggle and Ribeiro Jr, 2007). Without supplementary information, the resulting map can be expected to be informative in close proximity of the monitors which translates to vast areas particularly in Mpumalanga with low predicted values and high prediction uncertainty for the statistic of interest. For this reason an improvement in statistical air quality mapping is sought.

2.2.1 Data

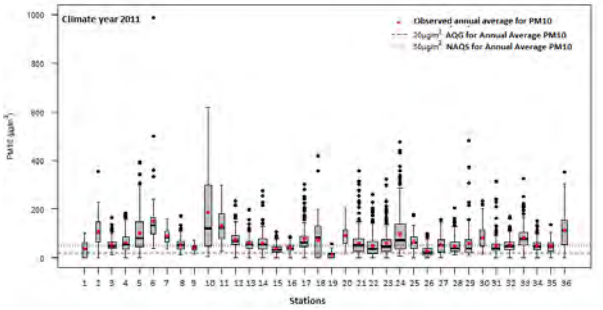
The focus here is on PM_{10} mass concentrations collected at 36 air quality stations between 1 September 2008 and 31 August 2012. We define this period as the 2009 to 2012 climatic years to denote annual cycles in which all four seasons are covered. We note that from the three stations classified as urban background stations, two are in residential areas, namely the Booysens (19) and Olievenhoutbosch (17) in the Tshwane municipal area. The other is the Delta monitor (9) located in a park in Johannesburg. Booysens is an older suburb, with lower housing density, paved roads and trees, whereas Olievenhoutbosch is an actively developing location with higher density of housing including informal backyard dwellings, some unpaved roads and a mine within 4 km South East of the monitor. The distinction of Olievenhoutbosch in terms of higher than typical observed levels of PM_{10} for background stations is observed from Figure 2.1. Therefore Olievenhoutbosch is considered in this study to be one of the 22 residential stations. There are four and eight stations classified as dominated by traffic and industrial emissions respectively, indicated by label T and I in Figure 2.1a. Due to the complexity of the study area in terms of air pollution sources, the classification of stations is not pure. Hence emissions from traffic or industrial sources can be substantial in areas classified as being dominated by emissions from domestic sources.

The summary of observed PM_{10} values per climate year in Figure 2.1 shows the best spatial coverage in 2011, while temporal coverage as expressed by the width of the box-plots was best in 2012 and worst in 2009. The observed annual mean PM_{10} level as represented by the red points is consistently the highest in areas categorized as dominated by domestic fuel emissions. It breaks even for locations dominated by traffic emissions and is moderately high in industrial areas. At each station, including background stations, more than 75% of PM_{10} values are above the World Health Organization's air quality guideline (AQG) which is based on minimizing public health risks. The NAQS are higher because they are country specific and take public health risks as well as activities necessary for economic growth and development into account.

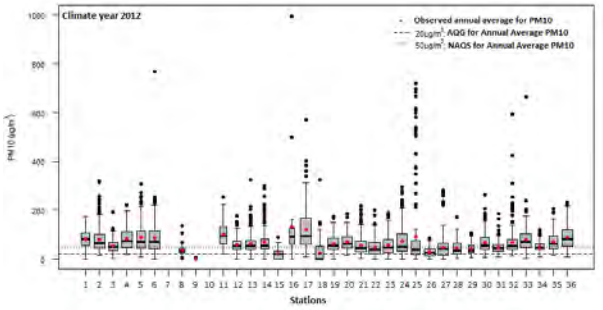
For this study, settlement-related spatial predictors of interest are extracted from the South African census 2011 small area dataset (Statistics South Africa, 2012b,a), with 24 584 small areas for the study area. Selected predictors as shown for Gauteng in Figure 2.2 are the percentage of households per small area. They include residing in informal dwellings; using solid biofuel energy for heating; and using solid biofuel energy for cooking. White spaces in Figure 2.2 are non-residential small areas, typically industrial areas including mines and quarries. Informal settlements are observed as slivers of red throughout the province, typically on the periphery of industrial or mining areas. On the southern boundary of Ekurhuleni, a moderate to high proportion of informality is observed, attributable to a mixture of informal settlements and formal residences with backyard shack dwellings in the townships of Vosloorus, Thokoza, Katlehong and Tsakane. Figure 2.2 shows



a PM_{10} observations for 2009



b PM_{10} observations for 2011



c PM_{10} observations for 2012

Figure 2.1: Empirical distribution summaries of PM_{10} observations at each station for the years 2009, 2011 and 2012

that for the three metropolitan municipalities, household use of biofuels for heating is most prevalent in areas with high proportions of informality. This similarity pattern is weaker for biofuel usage for cooking. In fact it is nearly absent in Ekurhuleni, Johannesburg and the core of Tshwane, except for areas on the periphery of the province.

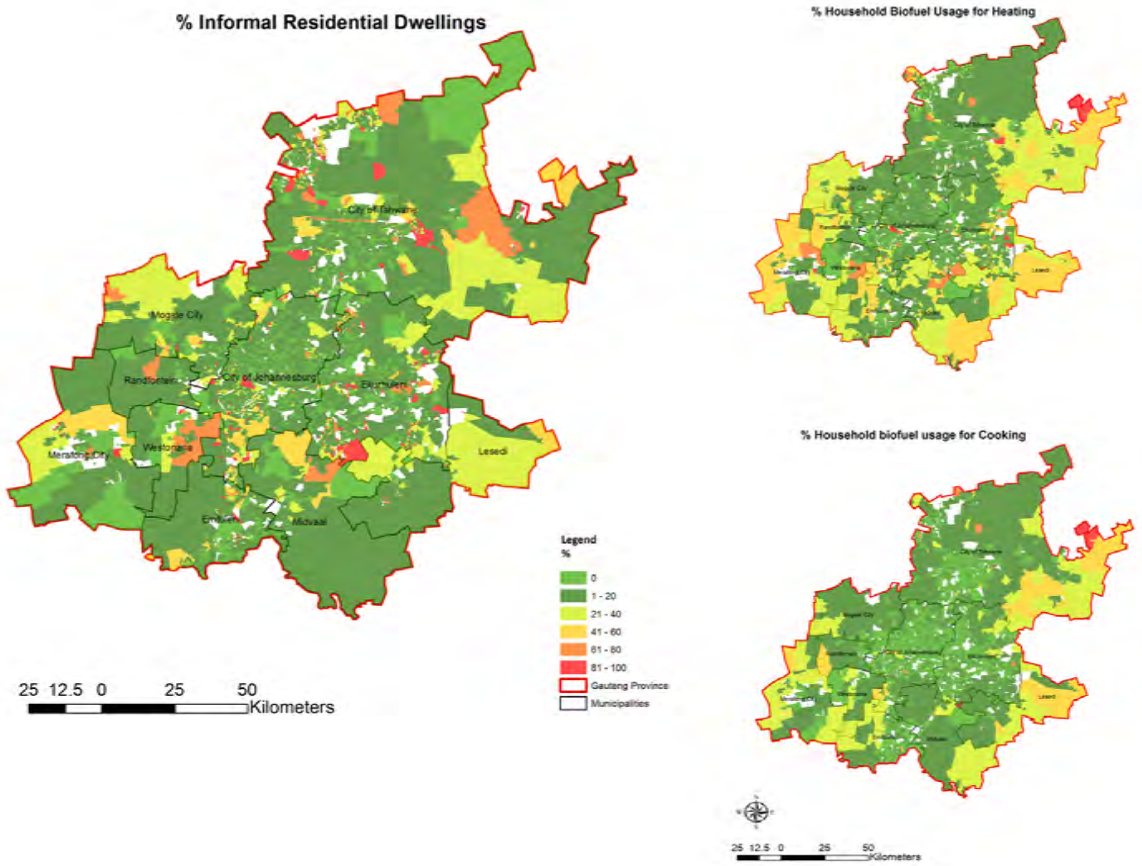


Figure 2.2: Variables from the South African census 2011 small area dataset that are assessed for plausibility in spatially predicting the number of exceedances of the PM_{10} NAQS

2.2.2 Pre-processing

The initial pre-processing step involves calculating PM_{10} NAQS exceedance counts for each station. The divisor in calculating the annual average number of exceedances is reflective of the number of observations available for that station in years. That is, for a station with 50% of the observations missing, the divisor is 2 (years) rather than 4. A plot of these average exceedance counts is shown in Figure 2.3. A general observation is that high counts are prevalent in locations where domestic sources of pollution are thought to be more dominant. Even for industrial source locations, only those sources interwoven with townships have substantial emissions from domestic sources that are high. These townships are: Embalenhle in Secunda, Zamdela and Sharpeville. All are categorized as industrial due to their location in the Vaal Triangle, an intensively industrial area.

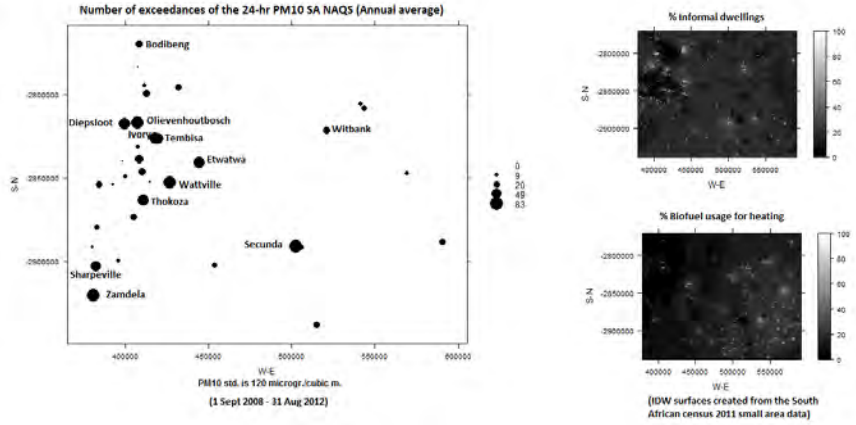


Figure 2.3: Average number of days per year where $PM_{10} > 120\,\mu g\,m^{-3}$ visually compared with percent informality of settlement and percent biofuel usage for household heating

Small area covariates are translated into 5 km resolution grids by means of inverse distance weighting. Grids for the percentage of households per small area in informal dwellings and usage of biofuels for heating are shown in Figure 2.3. The high concentration areas in the informal dwellings grid are similar to the pattern of high values on the exceedance count plot. This similarity signal however is weaker with the biofuel grid. Selection of covariate is followed by regressing the log-transformed exceedance rate with firstly all three covariates and eliminating until only covariates with a significant effect on the response variable remain. From this analysis the large scale spatial trend in number of days above the NAQS emerged as having a multiplicative association only with percentage informality with less than 0.1 probability that such an association is due to randomness. Although the association is significant, only 6.4% percent of the variability in the exceedance rate is

explained by this large scale trend alone. Therefore methods that account for both large scale and local spatial variability in the number of days above the PM_{10} NAQS are used as discussed in Section 2.3.

2.3 Methods

There are several ways to achieve the objective of mapping an air quality statistic. In this study we consider kriging and generalized linear geostatistical models (GLGM). In kriging the main decision regards the choice of function for spatial dependence, while in GLGM the main decision concerns plausible distributions for the latent spatial process assumed to have generated the observed values. The semivariogram as a representation of the underlying spatial dependence is at the core of both methods. Consider a spatial observation $y(s_i)$ being the annual average number of days where observed concentrations of PM_{10} in a 24-hour are above the regulatory level of $120\,\mu g\,m^{-3}$ at location s_i. Intrinsic stationarity and isotropy is assumed, hence the spatial correlation between observation pairs that are distance $h = \|s_i - s_j\|$ depends only on the separation distance h rather than on the actual location and direction. The exponential semivariogram is chosen and expressed as

$$\begin{aligned} \gamma(h) &= \frac{1}{2}\mathrm{E}\left(Y(s) - Y(s+h)\right)^2 \\ &= \sigma^2\left(1 - \exp(-h/\phi)\right) \end{aligned} \tag{2.1}$$

where σ^2 and ϕ represent the partial sill and range parameters (Bivand et al., 2008). The following subsections present kriging and the GLGM methods implemented in this study.

2.3.1 Kriging

Kriging is a spatial interpolation technique, a distance-weighted average, where the weights are given by the spatial correlation function in Equation 2.1. In this study two variants are considered, namely ordinary and external drift kriging. In ordinary kriging only the local scale spatial variation of the response variable as described by Equation 2.1 is modelled, while the mean of the underlying spatial process producing the observed values is assumed to be constant. In kriging with external drift, the assumption is that there are two components to spatial variability, the non-constant spatial trend which can be captured by known spatial regressors $X_j(s)$ and the semivariance captures the residual spatial variation after removal of the trend (Bivand et al., 2008). That is, a linear trend is assumed between the response surface and predictors,

$$Y(\mathbf{s}) = \sum_{j=0}^{p} X_j(\mathbf{s})\beta_j + e(\mathbf{s}) \tag{2.2}$$

where the residual spatial variation refers to the fluctuation in $e(\mathbf{s})$. Weighted least squares is implemented, minimizing the deviation of the exponential variogram from the sample variogram to obtain the sill and range parameters.

2.3.2 Log-Gaussian and Poisson geostatistical models

In kriging the uncertainty in the variogram parameters is not considered, the function is deterministic. In contrast, the GLGM method treats the semivariance stochastically, where the parameters are inferred from likelihood based functions and distributional assumptions are made for the latent spatial process. For this study, two distributional assumptions are considered. Initially the log-transformed annual exceedance frequency, $\log(Y(\mathbf{s}))$, is assumed to be Gaussian with the mean $\mu(\mathbf{s})$ being a linear function of the predictors, in this case percentage informality per small area and covariance matrix $\mathbf{V} = \sigma^2\mathbf{R}$. Elements of matrix $\mathbf{R}$ describe the correlation $\rho(h) = \gamma(h)/\gamma(0)$ with the semivariance function described in Equation 2.1 and the variance σ^2 are diagonal elements of $\mathbf{V}$. Note that the mean, variance and correlation of $Y(\mathbf{s})$ are as follows,

$$\mu_Y(\mathbf{s}) = \exp\left(\mu + \frac{\sigma^2}{2}\right) \tag{2.3}$$

$$\sigma^2_Y = \exp\left(2\mu + \sigma^2\right)\left(\exp(\sigma^2) - 1\right) \tag{2.4}$$

$$\rho_Y(h) = \frac{\exp\left(\sigma^2\rho(h)\right) - 1}{\exp(\sigma^2) - 1} \tag{2.5}$$

The log-Gaussian geostatistical model is commonly applied if areas of small values are in close proximity to large values. Given that the observations are counts, these can be considered realizations of a spatially correlated Poisson process. Spatial correlation is invoked through the Poisson intensity $\lambda(\mathbf{s})$, a function of an unobserved spatial process with spatial trend $\mu(\mathbf{s})$ captured by the covariates and residual spatial variation through a latent zero-mean Gaussian process $W(\mathbf{s})$ with covariance matrix $\mathbf{V} = \sigma^2\mathbf{R}$ similar to the log-Gaussian model. This means that the Poisson log-linear geostatistical model can be expressed as,

$$Y(\mathbf{s})|W(\mathbf{s}) \sim \text{Poisson}(\lambda(\mathbf{s})) \tag{2.6}$$

$$\lambda(\mathbf{s}) = \exp(\mu(\mathbf{s}) + W(\mathbf{s})) \tag{2.7}$$

Given the hierarchical nature of the GLGM, spatial prediction follows by Markov Chain Monte Carlo (Diggle and Ribeiro Jr, 2007). In its generic form, the parameters of the latent zero-mean Gaussian process are estimated and these are used in the next level to calibrate parameters for the Poisson distribution or for the distribution of the actual rather than the log-transformed variable in the case of the log-Gaussian models. Realizations at unobserved sites can then be obtained from these predictive densities.

2.4 Results

An assessment of spatial correlation through the semivariogram is the first step in kriging and GLGM approaches to statistical mapping. Table 2.1 shows parameter estimates obtained with weighted least squares and maximum likelihood inference for an exponential semivariogram with and without percentage informality as covariate are shown as WLS_{null}, WLS_{trend}, ML_{null} and ML_{trend} respectively. Minor positive changes were observed in goodness of fit statistics, namely WLS, the log-likelihood and the Bayesian Information Criterion, when percentage informality was added as a covariate. During pre-processing the small proportion of spatial variation captured by informal settlement prevalence as predictor of the mean spatial trend was noted in Section 2.2.2. This finding is confirmed by the small effect size for percentage informality (β_1) in comparison with the intercept β_0 in Table 2.1. There is a marked difference in spatial variation accounted for by informal settlement prevalence between WLS and ML, with the former attributing 20% variability to this regional trend and 4% is attributed by the latter. The larger effect size for β_0 results in similarities between partial sill and range estimates for the ML and WLS_{trend} estimators. From these results, we observe that there is a strong spatial dependence in the NAQS exceedance frequency of PM_{10} at distances below 8 km.

Table 2.1: Parameter estimates for the exponential semivariogram model from weighted least square and maximum likelihood methods, with the nugget (σ_0^2) fixed at 0.1

Par	WLS_{null}	WLS_{trend}	ML_{null}	ML_{trend}
$\beta_0^{(s.e)}$		$2.67^{(0.23)}$	$3.02^{(0.24)}$	$2.88^{(0.26)}$
$\beta_1^{(s.e)}$		$0.02^{(0.01)}$		$0.01^{(0.01)}$
σ^2	1.43	1.15	1.14	1.19
ϕ in km	9.14	7.71	8.00	8.00

The first purpose of mapping the PM_{10} exceedances is to determine areas that chronically have poor air quality. A second purpose is to determine whether there is an association with settlement-related drivers when exploratory analysis results in Figure 2.1 show that stations monitoring emissions from domestic sources have higher PM_{10} concentrations. Therefore maps in Figure 2.4–2.6 are organized to show differences between using and not using the chosen covariate. The advantage of using a spatially extensive covariate if the response variable is spatially sparse is that the map can be improved especially at unobserved locations. A caveat is that the degree of improvement depends upon the strength of the correlation between the response variable and the covariate.

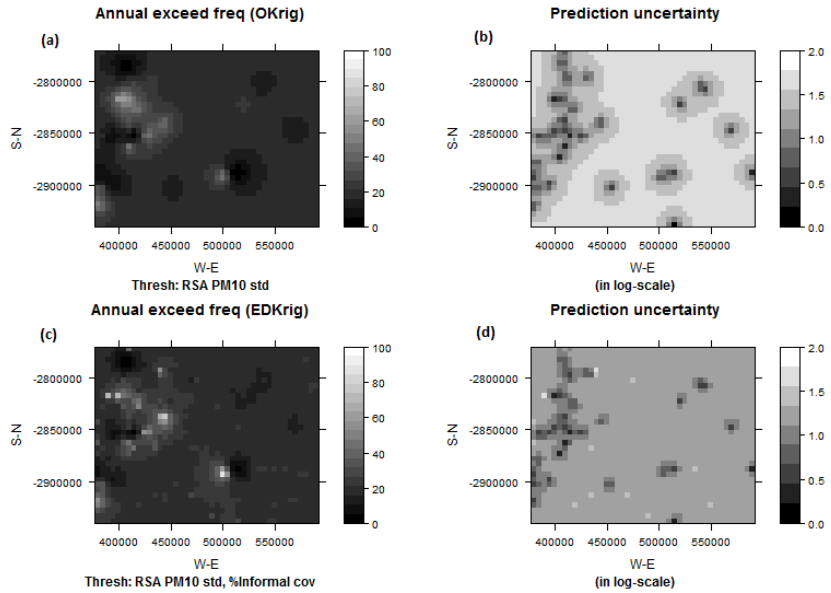


Figure 2.4: Maps of the average number of days per year where $PM_{10} > 120\,\mu g\,m^{-3}$ obtained through ordinary and external drift kriging

The results from implementing OK and KED as shown in Figure 2.4 indicate exceedance hot-spots to be in Ekurhuleni, the Vaal triangle and Secunda (refer to Figure 3.1). KED with the percentage of informal dwellings per small area as the explanatory variable, resulted in a more detailed spatial pattern with hot-spots extending further. Figure 2.4(b) and (d) are prediction error variance maps, showing that as expected, precision in spatial prediction is lowest (high prediction error variance) in areas without air quality stations. With the inclusion of a proxy for informal settlement prevalence, prediction uncertainty is reduced especially in areas without air quality monitors between Gauteng and Mpumalanga as well as areas surrounding each monitoring station.

The GLGM maps in Figure 2.5 are similar to kriging maps with regards to the location of exceedance hot-spots and the detailed spatial pattern observed when the covariate is considered. Maps from the Poisson log-linear geostatistical model have similar patterns to the log-Gaussian maps and are therefore not included. A thousand predicted surfaces were simulated from the predictive density of the log-Gaussian model from which the 2.5% and 97.5% quantile surfaces are extracted and shown in Figure 2.5 (b),(c),(e) and (f), expressing the model's prediction uncertainty. Prediction uncertainty is pronounced for high values and in areas without observations. In Figure 2.6, it is observed that including the covariate leads to higher predicted values in kriging and both GLGMs for locations identified as hot-spots and lower

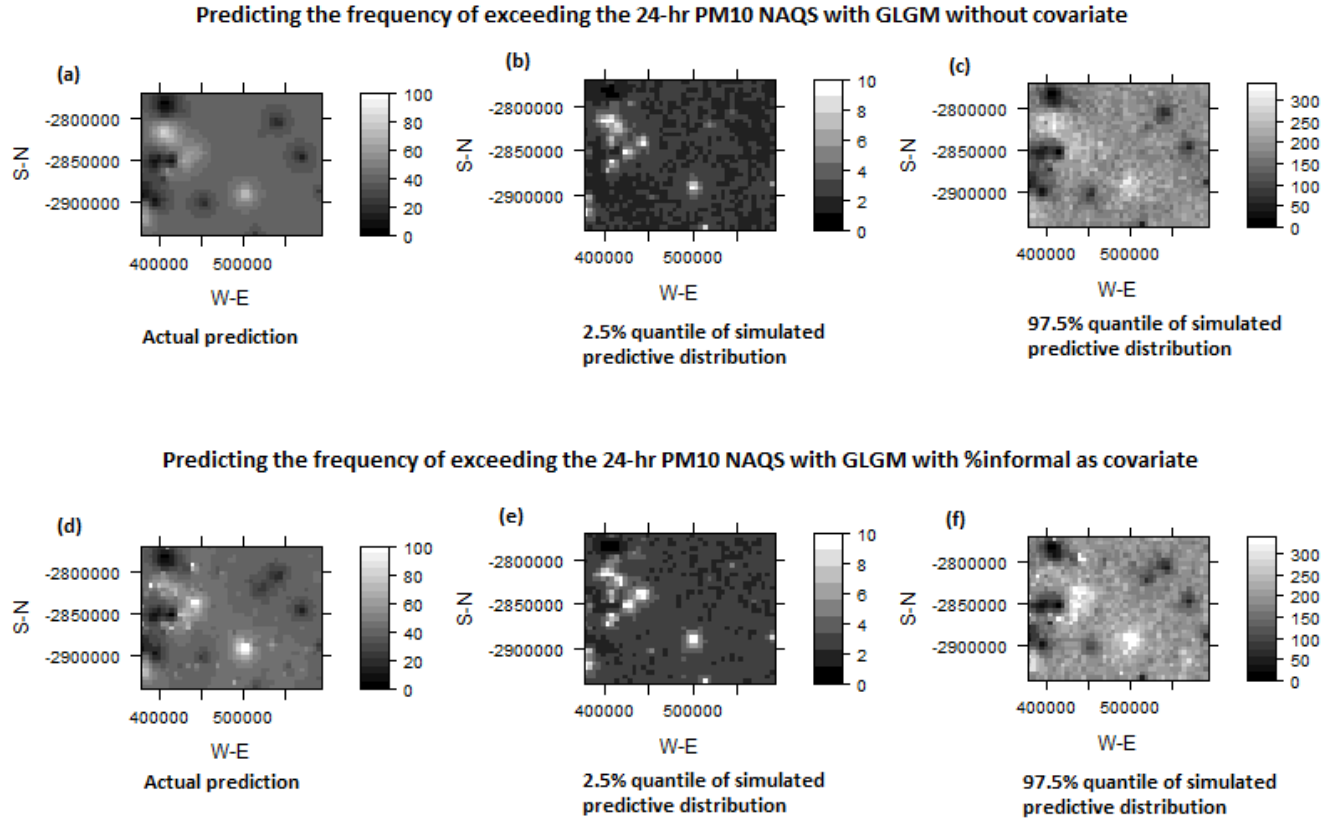


Figure 2.5: Maps of the average number of days per year where $PM_{10} > 120\,\mu g\,m^{-3}$ obtained from the log-Gaussian geostatistical model with and without a covariate, with the corresponding predictive simulation quantile maps showing prediction uncertainty

predictions for locations where exceedance counts are low. This pattern is in particular pronounced for GLGMs. The predictions are closer at unsampled locations. Further, quantile maps show that hot-spots are less likely to meet the regulatory standard of no more than four days per year in which the 24-hour PM_{10} threshold may be exceeded. This is investigated further by calculating the probability of exceeding this threshold of no more than four exceedances per year. The finding is that for 75% of the study area, the probability of exceeding this annual count threshold is less than 0.1 for both the covariate and no covariate log-Gaussian models. For Bodibeng in Tshwane, Wattville and Etwatwa in Ekurhuleni as well as Sharpeville in the Vaal Triangle, the exceedance probability ranges between 0.4 and 0.7.

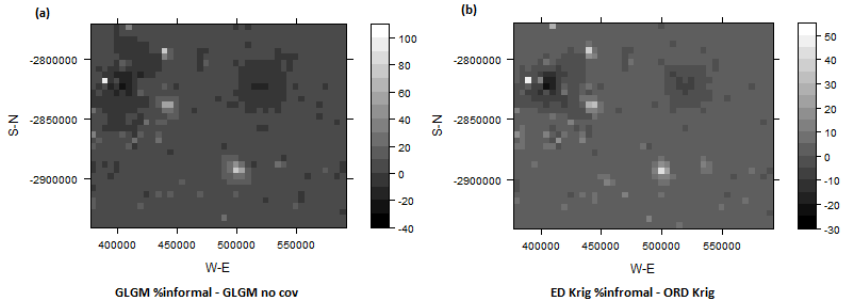


Figure 2.6: Differences resulting from adding percentage settlement informality in mapping exceedance frquency using (a) the log-Gaussian model and (b) kriging

Exceedance counts from four monitoring stations were excluded in building models that produced results shown in this section. These stations are: Delta which is an urban background station; Buccleuch which monitors traffic emissions as a source; Tembisa which monitors domestic emission sources and Langverwacht which monitors industrial emissions. The observed average number of exceedance days per year for these stations are 1, 8, 51 and 11 respectively. The values observed at modelling stations closest to these validation sites are worth noting. These are 35 for Alexandra which is close to both Delta and Buccleuch stations; 57 for Ivory which is in the same precinct as Tembisa and 83 for Secunda which is in the same precinct as Langverwacht. With an exception of Tembisa in Figure 2.7, all models poorly predict the observed values at test locations. This may be the result of the counts at the validation sites being substantially lower than the values at their nearest stations that are used in building the models. Figure 2.7 also shows that both kriging and GLGM methods are biased upwards. Further, for outlying observations, that is where differences between values for test and nearest modelling locations are large, incorporation of percentage informality as a predictor for large scale spatial trend leads to less accurate predictions. The GLGM is more affected in this regard.

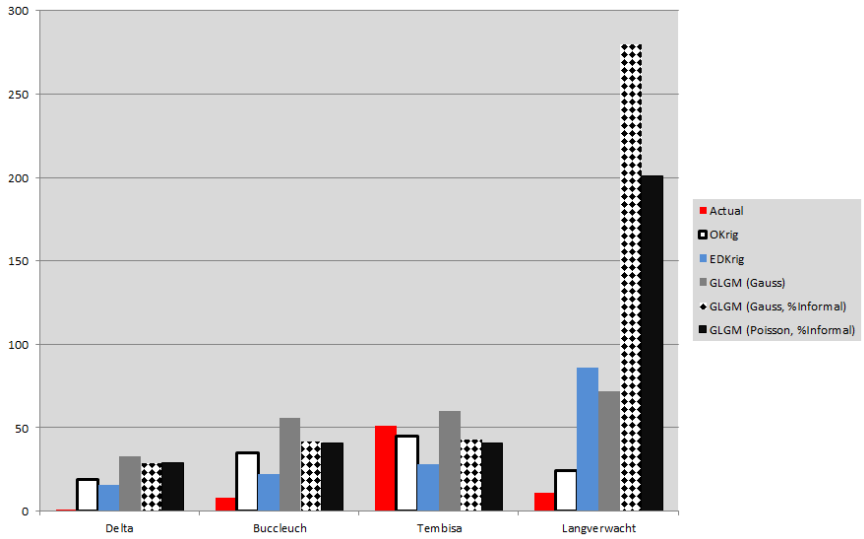


Figure 2.7: Predictions of the average number of days per year above the NAQS for PM_{10} for test locations

Results of further evaluation of spatial prediction performance for the different methods are shown in Table 2.2. Accuracy measures summarizing the discrepancies in prediction are the unbiased root mean squared error (URMSE) and the sum of squared log accuracy ratio (SSLAR). The URMSE is a bias corrected standard deviation of the predictions from actuals and should ideally be zero (van de Kassteele, 2006), whereas SSLAR is the relative accuracy of predictions to the actuals and should also ideally be zero (Tofallis, 2015). According to Tofallis (2015) the SSLAR is more suitable for this study given the positive and heteroscedastic nature of the data and the corresponding multiplicative error modelling methods. Assessment of accuracy using URMSE is included for comparison and its sensitivity to large errors is expected. Mathematically these are expressed as,

$$URMSE = \sqrt{\frac{1}{4}\sum_{i=1}^{4}(\hat{y}_i - y_i)^2 - ME^2} \tag{2.8}$$

$$SSLAR = \sum_{i=1}^{4}\left(\ln\left(\frac{\hat{y}_i}{y_i}\right)\right)^2 \tag{2.9}$$

where the mean error, $ME = \sum_{i=1}^{4}(\hat{y}_i - y_i)/4$, is the bias correction term. There is consensus between kriging and the GLGM techniques of a deterioration in prediction accuracy when percentage informality is included as a covariate.

Kriging prediction accuracy is the highest, whilst the Poisson log-linear geostatistical model performed better than the log-Gaussian model.

Table 2.2: Evaluation of spatial prediction methods for the 24-hour PM_{10} NAQS exceedance counts

Method	URMSE	SSLAR
Ordinary Kriging	12	11
External Drift Kriging	35	13
GLGM (Gaussian)	19	20
GLGM (Gaussian, covariate)	110	25
GLGM (Poisson)	19	19
GLGM (Poisson, covariate)	77	22

2.5 Discussion

A major difficulty in geostatistical mapping of PM_{10} indicators is caused by the multiple ways in which particles may be formed or removed from the atmosphere. If this complex process is broken down into sources of variability, variability due to the temporal evolution of the process involving factors such as residence times and accumulation of particles in ambient air would be one component. In terms of exceedance of NAQS this relates to persistence, where exceedances occur on successive days. Then there is a spatial variability component which concerns the effects of landscape heterogeneity and locations of emission sources (Janssen et al., 2008). Spatiotemporal variability due to meteorology is another component (Velders and Matthijsen, 2009; Barmpadimos et al., 2011). Cocchi et al. (2007) found that temporal persistence was the greatest source of variability (68%), while spatial random effects and meteorology accounted for 18% and 5% respectively. We focussed on spatial variability in exceedance frequency explained by settlement related variables as proxies for domestic emissions of PM_{10}. Therefore, small proportions of explained variability were anticipated.

Prevalence of informal dwellings emerged as a statistically significant predictor of the number of days the daily PM_{10} standard was exceeded. Domestic biofuel usage variables however were not significant, given that previous environmental health studies identified domestic fuel burning as a significant contributor to poor air quality in the study area (Wright et al., 2011). A reason for their insignificance is collinearity with the informal dwelling percentage variable. Percentage informal dwellings per small area is a better predictor because informal settlements are by definition constrained in terms of provision of basic services including electricity. Financial constraints are another reason that some residents in electrified informal settlements, low-cost township housing and backyard shacks choose to use biofuel energy sources for energy intensive activities such as cooking and heating.

Figure 2.8 shows housing conditions in Ekurhuleni municipality, where according to results presented in Section 2.4 and Figure 2.3, stations in this area (Tembisa, Ivory, Etwatwa, Wattville and Thokoza) record high levels of ambient PM_{10} often exceeding the NAQS. From Figure 2.2 small areas surrounding Etwatwa, Wattville and Thokoza show higher prevalence of biofuel use for domestic heating. 21% of small areas in Ekurhuleni are dominantly formal '4-room' township (pre-1994) and RDP [1] houses mixed with informal dwellings in the form of backyard shacks and settlements with stand-alone shacks (Khuluse-Makhanya et al., 2016). Socioeconomic characteristics include over-crowding given average household size of six or more in small houses, as well as low income, education and employment levels with approximately 27% of these households being headed by children aged 15 years and younger (Dudeni-Tlhone et al., 2013). Studies on determinants of household energy choices and transition from domestic biomass and coal combustion to gas or electricity in other developing countries, also identified these socioeconomic factors including energy costs as being of significant influence (Mensah and Adu, 2015; Chen et al., 2016).

Secunda is also identified as a hot-spot in Section 2.4. A chemical and fuel plant in the area is a known air pollution source. Figure 2.3, however shows that energy poverty and a general insufficiency of basic services characteristic of informal settlements also results in a contribution and vulnerability to air pollution specifically in the township of Embalenhle which is in Secunda (John and Das, 2012).

Two statistical mapping methods considering informal dwelling prevalence as a covariate as well as the Poisson and Gaussian distributional assumptions were implemented in this paper. The best linear unbiased predictor, kriging, out-performed the Poisson log-linear and log-Gaussian models in predicting the number of exceedances of the PM_{10} NAQS at four test locations. van de Kassteele (2006) found that neither the Poisson or log-Gaussian models fitted well to their annual Ozone exceedance count data. However they noted desirable properties of these models including better fit of the log-Gaussian density to the long-tailed empirical distribution of their data and the Poisson as an appropriate distribution for count data, eliminating the need for transformation. Similarly, we found that prediction uncertainty is larger for larger predicted values and that the accuracy of Poisson model was marginally better than the log-Gaussian model. With the inclusion of a covariate, as explored in this study, the Poisson model's accuracy was still better than the log-Gaussian model, but kriging with external drift remained the better of the three models.

[1] Reconstruction and Development Programme (RDP) low-cost housing provided after 1994 to help alleviate housing backlogs and improve living conditions in townships and rural areas.

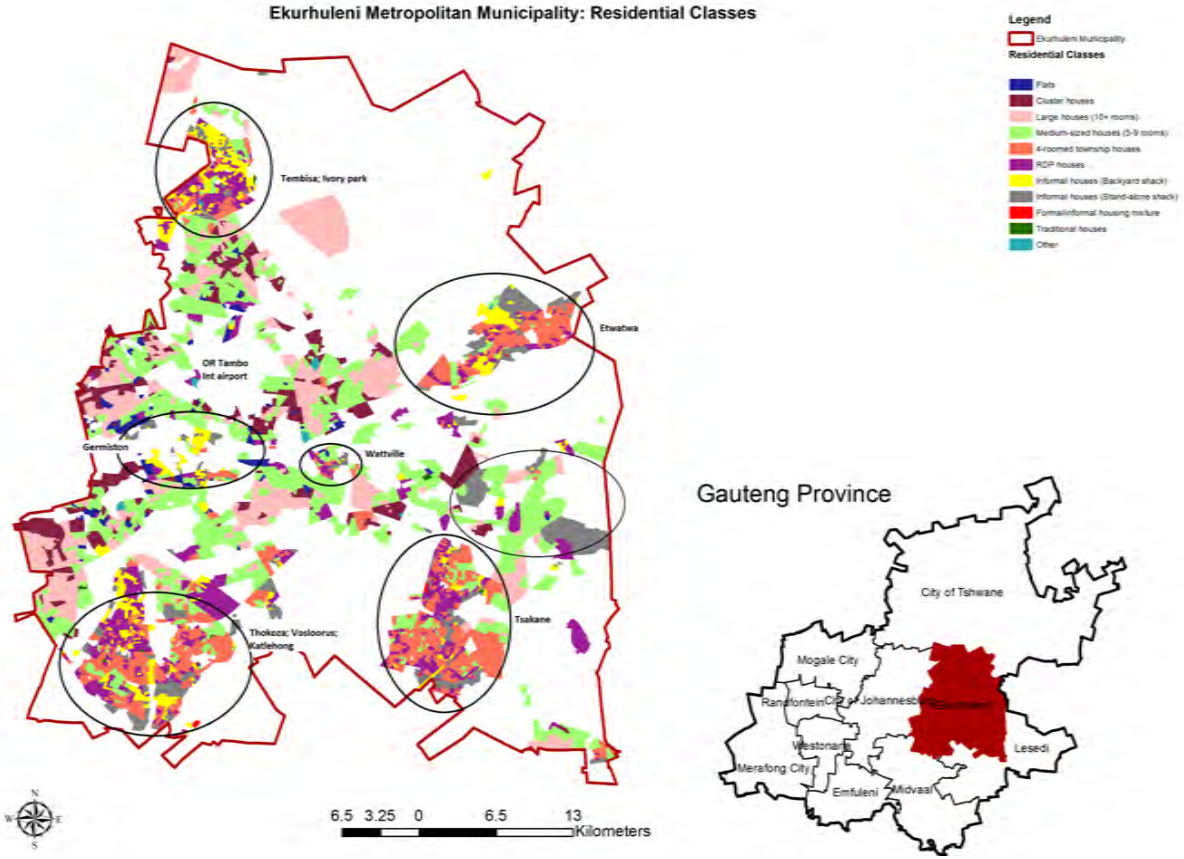


Figure 2.8: Ekurhuleni metropolitan municipality housing conditions extracted from the first two stages of the multi-stage k-means geo-demographic clustering (Khuluse-Makhanya et al., 2016)

2.6 Conclusion

The objective of this study is to statistically map the average number of times per year that the South African air quality standard of 120 $\mu g\,m^{-3}$ for PM_{10} is exceeded in the Highveld region based on four years of data between 2008 and 2012. Further, due to the study region being extensively urban with informal dwellings being a substantial part of the housing stock in this area, housing characteristics are assessed as potential spatial predictors of elevated ambient concentrations of PM_{10}. The percentage of dwellings classified as informal per small area from the Census 2011 is the only significant explanatory variable, hence dwelling total per small area and household use of biofuels (wood, coal and dung) for cooking and heating are not used. The selected predictor is satisfactory given the statistical requirement to avoid collinearity amongst predictors and energy poverty being a prominent issue in informal settlements. The spatial pattern of the PM_{10} NAQS exceedance rate in terms of location of hot-spots is consistent for kriging, the Poisson and log-Gaussian generalized linear methods. In this study prediction accuracy of kriging methods is superior to the model-based geostatistical models. The advantage of using spatial predictors to improve mapping PM_{10} statistics given a spatially sparse monitoring network can only materialize if the predictor and response surfaces are correlated, with the degree of improvement dependent on the strength of this correlation. Prediction uncertainty at unsampled areas including the validation sites is improved by incorporating the explanatory information in this study.

3

Ensemble classification for identifying neighbourhood sources of fugitive dust and associations with observed PM_{10}

This chapter is based on the paper (under revision): Khuluse-Makhanya S., Stein A., Breytenbach A., Gxumisa A., Dudeni-Tlhone N., Debba P. Ensemble classification for identifying neighbourhood sources of fugitive dust and associations with observed PM_{10}. *Atmospheric Environment*.

Abstract

In urban areas the deterioration of air quality as a result of fugitive dust receives less attention than the more prominent traffic and industrial emissions. We assessed whether fugitive dust emission sources in the neighbourhood of an air quality monitor are predictors of ambient PM_{10} concentrations on days characterized by strong local winds. An ensemble maximum likelihood method is developed for land cover mapping in the vicinity of an air quality station using SPOT 6 multi-spectral images. The ensemble maximum likelihood classifier is developed through multiple training iterations for improved accuracy of the bare soil class. Five primary land cover classes are considered, namely built-up areas, vegetation, bare soil, water and 'mixed bare soil' which denotes areas where soil is mixed with either vegetation or synthetic materials. Preliminary validation of the ensemble classifier for the bare soil class results in an accuracy range of 65–98%. Final validation of all classes results in an overall accuracy of 78%. Next, cluster analysis and a varying intercepts regression model are used to assess the statistical association between land cover, a fugitive dust emissions proxy and observed PM_{10}. We found that land cover patterns in the neighbourhood of an air quality station are significant predictors of observed average PM_{10} concentrations on days when wind speeds are conducive for dust emissions. This study concludes that in the absence of an emissions inventory for ambient particulate matter, PM_{10} emitted from dust reservoirs can be statistically accounted for by land cover characteristics. This supports the use of land cover data for improved prediction of PM_{10} at locations without air quality monitoring stations.

Keywords: Particulate matter, fugitive dust, land cover, ensemble classifier, k-means clustering, varying intercepts regression model

3.1 Introduction

Particulate matter (PM) is a highly erratic pollutant in urban landscapes due to its formation from both mechanical and chemical processes, sensitivity to meteorological conditions, volatile residence times as a result of sedimentation and increased likelihood of impaction given the larger built-up footprint in urban areas (Beelen et al., 2009; Velders and Matthijsen, 2009; Zwack et al., 2011; Barmpadimos et al., 2011). Chemical processes refer to the formation of particles through condensation of gases produced by incomplete combustion as industrial, vehicle and biomass burning fumes (Sportisse, 2009). Mechanical formation of PM refers to direct emission of particles as dust from: agricultural fields, construction sites, unpaved roads and yards, mining and quarrying sites including mine residue deposits (MRDs) or "mine dumps", re-suspension from vehicle tyre and road surface interactions, etc. (Watson and Chow, 2000; Athanasopoulou et al., 2010). This is especially relevant in urban landscapes of developing countries where areas of bare soil in the form of active surface mining areas and residue deposits, unpaved roads and yards in informal settlements as well as natural bare ground, are intertwined with impervious surfaces (Kneen et al., 2015). Dust in this context is often reported as a "health nuisance" especially in communities close to mine residue deposits. With human settlement growth near such sources, quantitative studies of the association between dust emissions, monitored particulate matter concentrations and health are necessary (Chikusa, 1994; Ojelede et al., 2012; Kneen et al., 2015). Apart from aeolian emissions from bare ground, there are process-based fugitive dust emission sources such as material-altering industrial operations, agricultural tilling and disturbances on roads and parking lots (paved and unpaved).

High spatial resolution satellite images are an important source of urban land cover data (Myint et al., 2011). There are various approaches for extracting land cover classes from imagery and a common technique is pixel-based supervised maximum likelihood (ML) classification (Besag, 1986; Khatami et al., 2016). Khatami et al. (2016) concluded that additional dimensions in the form of texture, ancillary, multi-time or multi-angle data led to the greatest improvements in overall pixel-based classification accuracy. Another way to introduce additional information is by training multiple classifiers on the same problem with the purpose of combining the outputs to achieve greater accuracy than the individual classification result. This method is known as ensemble classification (Maimon and Rokach, 2010). In a recent study by Banerjee et al. (2015), an ensemble clustering routine was developed for improving class means and covariances used in initializing an expectation-maximization classification algorithm. However, this self-training cluster ensemble prior to ML classifier was outperformed by a supervised ML classifier, with the latter achieving 97.2% overall accuracy. Such superior performances of the ML classifier, even in cases of less representative training samples, justify its popularity in land cover mapping, in terms of both direct application and development of advanced classification procedures (Li et al., 2014).

Land-use regression (LUR) models are a popular tool in air pollution exposure assessments (Jerrett et al., 2005; Millar et al., 2010b; Zwack et al., 2011). LUR models achieve the objective of mapping air quality by using the correlation between pollutants of interest and land-use indicators such as population density, traffic intensity and distances to known pollution sources (Beelen et al., 2009; Zwack et al., 2011). A varying intercept model is a regression model that is suitable when the data are clustered. It is common in applications of this model that grouping factors are known (Gelman and Hill, 2007). Given no specific grouping factors, clusters can be identified by applying a k-means clustering technique (MacQueen, 1967) prior to regression modelling. The strengths of the k-means technique include the capability of handling large data sets (Van Eetvelde and Antrop, 2009; Dudeni-Tlhone et al., 2013) and ease of adaption into advanced clustering procedures aimed at superior data handling efficiency (Chen and Gong, 2013). A k-means cluster analysis was used to transform land cover classification outputs derived from satellite imagery with ancillary spatial data into landscape metrics for inference about determinants of rural land cover change in Reger et al. (2007) and in creating a multi-scale trans-border oriented landscape typology for Belgium in Van Eetvelde and Antrop (2009). A well-known difficulty with clustering is the interpretation of the clusters, therefore other information can be useful for describing the clusters.

The objective in this study is to assess whether there is a statistical association between land cover and observed PM_{10} concentrations as a basis for using spatially extensive land cover data to predict PM_{10} in locations without air quality monitoring stations. For this purpose, we propose a varying intercept model using land cover and observed PM_{10} data. In the absence of recent high spatial resolution land cover data, we propose an ensemble maximum likelihood classification method characterized by the use of multiple training sets to capture the heterogeneity in local sources of fugitive dust emissions.

The chapter is presented as follows: Section 3.2 consists of details on the study area and data; the ensemble classification and the k-means prior to varying intercepts modelling methods are presented in Section 3.3; the results are in Section 3.4; the discussion in Section 3.5 and the concluding remarks in Section 3.6.

3.2 Materials

3.2.1 The study area

Air quality monitoring stations (Figure 3.1) are located in the Gauteng province of South Africa. From the 2011 census, nearly one million households were living in informal dwellings in Gauteng, where solid biofuels, gas or paraffin provided energy for cooking (75%), heating (58%) and lighting (73%) due to lack of electricity (Balmer, 2007; Housing Development Agency,

2013). Gauteng's traffic corridors carry large volumes of passenger and freight vehicles because half of South Africa's main freight corridors which transport on average 246 metric tonnes of freight annually are located in the province (CSIR, 2014). Prevailing industrial activities include mining, mineral and metal processing. Gauteng is in a grassland biome and is dominated by soils (acrisols, leptosols and lixisols) with low nutrient retention capacity which are susceptible to wind erosion (Jones et al., 2013). The colour tone of soils found predominantly in the province range from reddish brown (northern Gauteng) to light yellow at locations with high anthropogenic disturbances as well as fine particle 'black-clay' soils (vertisols) south of the province. Air pollution sources in the study area include vehicles, industries and domestic fuel burning activities (Alade, 2010; Ncipha, 2011), however our interest is on relating observed ambient PM_{10} concentration to land cover characteristics particularly as they relate to fugitive dust emissions. Therefore, variability in surface properties of soils becomes an important consideration in identifying dust emission reservoirs from optical satellite images.

3.2.2 Land cover imagery and air quality data

Daily PM_{10} and wind data from 23 air quality stations in Gauteng for the period starting March 2011 until February 2015 were used. The dominant pollution source classification of these stations by the custodians of the data is as follows: urban background (3 stations), industry (6 stations), domestic (13 stations) and traffic (4 stations). High spatial resolution multi-spectral images without panchromatic sharpening were obtained from the South African National Space Agency (SANSA) and used for land cover mapping. These are SPOT 6 multi-spectral images with 6 m ground sampling distance. The images had been orthorectified using ground control points, 50 cm aerial reference imagery and a 2 m digital elevation model. They were radiometrically calibrated implying that sensor-received light radiances were recomputed to the top-of-atmosphere (TOA) normalized reflectance. The effect of panchromatic sharpening would have been an increase in the spatial resolution of the images to 1.5 m, but the unsharpened images were used. Wavelengths for SPOT 6 bands on the electromagnetic spectrum: Blue ($0.455 - 0.525$ μm); Green ($0.530 - 0.590$ μm); Red ($0.625 - 0.695$ μm); NIR ($0.760 - 0.890$ μm). The images are dated 17 March 2013 and 17 April 2013, where the latter image covers four stations located south of the province.

The purpose of land cover mapping in this case is to provide inputs for statistically assessing the association between measured PM_{10} concentrations and sources of fugitive dust emissions. According to Kok et al. (2012), wind-blown dust can reduce visibility to less than 200 m during severe dust storms to 10 km during episodes of local strong winds. Further, the travel distance associated with a 10 μm aerodynamic diameter particle given 3 m s^{-1} wind speed is 1 km which rapidly increases to 4 km for smaller particles of diameter 5 μm (Watson and Chow, 2000). Therefore, we focus on an area that is limited to a 4 km radius from the location of an air quality station (Figure 3.1). This

corresponds to the upper bound of the neighbourhood scale (500 m − 4 km) commonly applied in intra-urban air pollution modelling and emission source apportionment assessments (Watson and Chow, 2000; Janssen et al., 2008). Janssen et al. (2008) considered the radius to be a free parameter and they found 2 km radii to be optimal for maintaining site-specific character of the CORINE land cover class distribution and discriminating between urban areas of different sizes. We selected seven monitoring areas (Figure 3.1) for developing the classifier. These areas are representative of variations in dominant emission sources of PM_{10}, land cover and soil types in the province (Figure 3.2, Table 3.1).

3.3 Methods

3.3.1 A land cover classification procedure for increased bare soil class precision

Land cover classification is undertaken with interest in basic land cover types, namely vegetation (V), bare soil (BS), built-up (BU) and water bodies (W). A fifth class for mixed bare soil areas (m-BS) which are defined by mixture of bare ground and grass (typically degraded) or synthetic materials is derived in the aggregation step of the ensemble classifier. The ensemble classification output consists of aggregating output from three iterations of maximum likelihood classification and output from one iteration of classification using NDVI thresholds. Details about the ensemble classification method appear later in this section.

Given the heterogeneity of the landscape being studied, supervised classification was preferred because knowledge about the area could be incorporated during training of the classifier. Specifically the maximum likelihood classification method is chosen (Besag, 1986; Dean and Smith, 2003). Maximum likelihood classification starts with an initial set of class means $\{\mu_1, \mu_2, \ldots, \mu_{\mathrm{K}}\}$ that are derived from a training set, allocating to each pixel p with feature vector $\mathbf{x}_p$, the class of highest probability as follows:

$$P(C_i|\mathbf{x}_p) = \frac{P(\mathbf{x}_p|C_i)}{\sum_{j=1}^{\mathrm{K}} P(\mathbf{x}_p|C_j)} \tag{3.1}$$

where $P(C_i|\mathbf{x}_p)$ is the *a posteriori* probability of class $\{i : i = 1, 2, \ldots, K\}$ given the pixel feature vector $\{\mathbf{x}_p : p = 1, 2, \ldots, q\}$ where q is the total number of pixels for the imaged area. Equation 3.1 results from assuming equal *a priori* probabilities for all classes. The conditional probability that a pixel belongs to class i is Gaussian with variance-covariance matrix $M_i = (\mathbf{x}_p - \mu_i)^{\mathrm{T}} \mathbf{V}_i^{-1} (\mathbf{x}_p - \mu_i)$, expressed as

$$P(\mathbf{x}_p|C_i) = \frac{1}{(2\pi)^{D/2} |\mathbf{V}_i|^{1/2}} \exp\left(-\frac{M_i}{2}\right) \tag{3.2}$$

where dimensionality parameter D represents the number of spectral bands. The $D \times D$ variance-covariance matrix of class C_i is denoted by $\mathbf{V}_i$. $P(\mathbf{x}_p|C_i)$

need not be limited to the Gaussian density. Gorte and Stein (1998) showed that higher overall accuracy can be achieved by localized k-nearest neighbour estimation of the probability density $P(\mathbf{x}_p|C_i)$.

In Figure 3.3, an ensemble classifier based on maximum likelihood (ML) classification and on thresholds of the normalized difference vegetation index (NDVI) is presented. Ensemble classification entails aggregating results from multiple individual classification runs (Maimon and Rokach, 2010). The differences between the individual outputs can either be in terms of application of different classifiers or differences in input parameters for the same classifier. The objective of ensemble classification is to improve the accuracy achievable through a single classification effort. The basis for choosing the ensemble method in this study is the need for improved accuracy given the higher risk of inaccurate classification of bare soil in urban areas due to imagery of limited spectral resolution relative to the heterogeneity of the urban landscape. With reference to the workflow illustrated in Figure 3.3, the ensemble classification method consists of the following:

1. *Signature development* – Seven circular air quality neighbourhoods or areas of interest (AOIs) were selected from which three non-overlapping training regions were selected. In each training region 14 areas were selected representatively in terms of land cover classes and variations in soil types. From each training area, 20–30 pixels were selected resulting in a training sample size that was within the recommended range for maximum likelihood classification (Gorte and Stein, 1998). The training samples were the basis for signature development for the ML classifier and for determining NDVI thresholds for the bare ground, water and vegetation classes. In this step two classes for vegetation (grass and shrubs, trees) and four classes for bare soil corresponding to dominant colours ranging from reddish brown (plinthosols, leptosols and nitisols), black (vertisols), yellow to white (characteristic of acrisols and soils heavily contaminated by chemicals (technosols)) are considered. For the built-up (BU) and water classes (W) no sub-divisions were considered and these features were also represented in the training samples. For the BU class, during training there was emphasis on including pixels for different colours and types of roofs and pavements.

2. *Classification* – Three iterations of ML classification were performed for each of the training sets, resulting in the image being classified into four classes, namely V, W, BU and BS during each iteration. A this stage reference is made to four classes rather than five because the "mixed with bare soil class" is only derived in the aggregation stage of the ensemble of classification outputs. The last part of the ensemble classifier uses NDVI thresholds for different natural land cover types. Reference values for water bodies, bare ground and vegetation are accepted as: $< 0, 0 - 0.2, 0.2 - 0.9$ respectively. The applicability of these threshold ranges for our scenes was verified during training where the different land cover types were known. Montandon and Small (2008)

also recommended verification of thresholds using local scene NDVI values especially in regions dominated by grass, shrubs and variation in soil properties. Bare ground includes bare soil and built up areas.

3. *Dichotomizing land cover classes* – The final outputs from the ML and NDVI classifiers were reclassified into binary rasters. Each of the three land cover classification outputs from the ML classifier were dichotomized by assigning one to BS pixels and zero to W, V and BU cells. The NDVI classification output was dichotomized by assigning one to bare ground pixels and zero to W and V cells. The final BS class label was determined by aggregating the four binary rasters. For the other classes the aggregate score was assessed in combination with multi-class output discussed in the second step.

4. *Bare soil total score raster* – The four binary rasters were aggregated into a bare soil score raster where each cell had a value that equals either 0, 1, 2, 3 or 4. A value of 4 meant that the class assigned by the ML and NDVI classifiers was bare soil, whereas a value of zero indicated that none of the classifiers identified that pixel as bare soil. Pixels with zero score were assigned W or V as final class labels depending on class majority from the four multi-class classification output layers. A value of 3 indicates a strong likelihood (75% agreement) that the pixel can be assigned to the bare soil class, whereas for pixels with a score of 2 there is 50% agreement with a bare soil class assignment. Therefore, pixels with bare soil scores of 3 and 4 were classified as bare soil (BS). An aggregate score of 1 emanating from the NDVI classification raster combined with a majority of BU class assignment from the three iterations of ML classification, resulted in BU being the final class labels for those pixels. The remaining pixels with scores of 1 and 2, without BU class majority, were defined by areas where soil was mixed with natural or synthetic features. The "mixed with bare soil" (m-BS) class label was given to these pixels. Therefore, the ensemble classifier resulted in five land cover classes, namely V, W, BS, BU and m-BS.

3.3.2 Sampling for performance evaluation of the classification routine

Assessment of classification accuracy is based upon visual inspection of the true-colour composite pan-sharpened 1.5 m resolution SPOT 6 image of the AOIs in conjunction with Google Earth imagery according to a simple random sampling design (Brus and Gruijter, 1997). An important consideration is the determination of sample size. For a multinomial population (Tortora, 1978), the marginal distribution for each class $i = 1, 2, \ldots, m$ is binomial with parameter p_i. Sample size is n such that,

$$n = B \times \frac{p(1-p)}{e^2} \tag{3.3}$$

where B is the $(\alpha/k) \times 100$th percentile of the χ^2_1 distribution for the proportion parameter. In our case we have no prior knowledge of standard

deviation of each class. Therefore, using $p = 0.5$, sample size per AOI is equal to

$$n = \frac{B}{4e^2} \tag{3.4}$$

Assuming the probability of Type I error is $\alpha = 0.05$ and 80% precision ($e = 0.2$), results in a sample size of 384 pixels.

Our ensemble classification method as discussed in Section 3.3.1 focusses on obtaining the bare soil class as accurate as possible. Therefore, validation is performed twice. The binomial test for accuracy was chosen for preliminary validation of how frequently bare soil was correctly classified at each of the seven AOIs during the classifier development stage. Table 3.1 shows that validation pixels were proportionally allocated to each AOI based on the unequal number of bare soil pixels results and sample size calculated in Equation 3.4. The purpose of preliminary validation is to identify common instances of bare soil misclassification as a basis for improving our classifier through further training. Thereafter, all 23 AOIs are classified. The conditional or intra-class Kappa assessment are chosen for final validation or testing at a randomly selected AOI where all five land cover classes are evaluated (Banerjee et al., 1999).

Table 3.1: Description of the seven areas of interest (AOIs) with respect to pollution sources and sample size chosen for the preliminary validation of the bare soil (BS) class

AOI (Source, Region)	**Description of AOI**	**BS pixel count**	**Prop. alloc.**	**BS test sample size**
Bodibeng (Dom, Tshwane)	Township, unpaved roads & yards, bare fields E of station	93 537	0.15	398
Booysens (Bg, Tshwane)	Pretoria CBD south of station, mining, agric fields N of station	56 129	0.09	238
PTA West (Ind, Tshwane)	Pretoria CBD E of station, mining south of station	34 466	0.05	147
Diepkloof (Traf, Vaal)	Mine residue deposits N-NE of station, township	130 050	0.20	553
Kliprivier (Ind, Vaal)	Extensive agric fields, low density residential, townships NE of station	204 724	0.32	872
Newtown (Traf, JHB)	JHB CBD, mining S of station, suburbs N of station	73 534	0.12	313
Tembisa (Dom, Ekurhuleni)	Township, bare sport fields, unpaved sidewalks & roads	42 188	0.07	180
Total		634 628	1	2 701

3.3.3 Assessing the variability of observed PM_{10} attributed to land cover

Natural and anthropogenic sources of fugitive dust emissions in urban areas are a challenge in terms of separability (Mansell et al., 2007; Korcz et al., 2009). This is complicated by the reduction effect that vegetation and built structures have on ambient dust particles. Wind activity is an important factor in the dust emission process, with the strength of emissions being a function of wind speed and landscape characteristics including soil type, vegetation and built surfaces. According to previous studies, wind speeds in excess of 6 m s^{-1} have been observed to induce emission fluxes in desert and arable areas (Korcz et al., 2009). For the Witwatersrand area which is part of our study area, Oguntoke et al. (2013) found that wind speeds during dust episodes were at least 4 m s^{-1}. Therefore, this wind threshold velocity is chosen for segmenting the PM_{10} data, focussing only on daily observations corresponding to wind speeds in excess of 4 m s^{-1}.

A varying intercept regression model is chosen to assess how much of the variability in average PM_{10} is explained by land cover characteristics. All land cover classes and a proxy for wind-blown dust emissions are considered as explanatory variables. The emission of dust is a non-continuous spatial process because of the intermittency of dispersion mechanisms and the heterogeneity of land cover, especially in urban areas. Our choice of model is motivated by the latter, the assumption being that neighbourhoods with similar land cover characteristics tend to have similar ambient PM_{10} concentrations. The expression for a varying intercept model by Gelman and Hill (2007) is:

$$y_i = \alpha_{j[i]} + \beta x_i + \epsilon_i \tag{3.5}$$

The observed average PM_{10} at air quality station i for days when wind speeds were in excess of 4 m s^{-1} is represented by y_i. The total number of distinct clusters (or groups) in terms of land cover is J. Given a baseline cluster, the intercepts α_j for $j = 1, 2, \ldots, J-1$ represent the changes in mean PM_{10} level in as a result of land cover characteristics represented by each cluster in comparison to the characteristics of the baseline. The final term is assumed at this stage to be a zero-mean constant variance Gaussian error term.

Two proxy variables for particulate matter emissions adapted from previous studies are considered for the slope term. The first is a proxy for wind-blown dust emitted from bare ground. It describes the horizontal flux of dust and is expressed by Korcz et al. (2009) as:

$$\begin{aligned} \mathrm{E}_{\mathrm{dust}}(10^9\,\mathrm{g\,m^{-2}}) \quad = \quad & \{\text{land area, m}^2\} \times \\ & \{[\text{spike emission rate, g m}^{-2}] + \\ & [(\text{duration of erosion event, h}) \times \\ & (\text{emission factor, g m}^{-2}\ \text{h}^{-1})]\} \end{aligned} \tag{3.6}$$

It is applied specifically to areas covered by bare soil and mixed bare soil pixels. The values for spike emission rate (0.318), duration of wind erosion

event (being 10 times the number of days with wind speed in excess of 4 m s^{-1}) and emission factor (1.73) are taken from Korcz et al. (2009) for unstable fine textured soils. The assumption of unstable soils is based on the region being dominated by erosion-susceptible soils and the high prevalence of anthropogenic disturbances as described in Section 3.2.1. Values by Korcz et al. (2009) correspond to wind threshold velocities in the range $8.9 - 11.1$ m s^{-1}, which are more than double our threshold of 4 m s^{-1}.

The other proxy, which we refer to as the Janssen's beta indicator, "is a single value indicator that correlates local land use characteristics to the local air pollution levels" according to Janssen et al. (2008). It is expressed as:

$$\beta = \log\left(1 + \frac{\sum_i a_i \times n_{CL(i)}}{\sum_i n_{CL(i)}}\right) \tag{3.7}$$

where $n_{CL(i)}$ is the number of pixels of class i in the neighbourhood and a_i weights the importance of land cover class i on ambient pollutant concentration levels. We use the optimized set of weights for PM_{10} published by Janssen et al. (2008). Zero weights are applied to vegetation and water pixels because they correspond to semi-natural areas and water bodies. We adapted the weighting coefficients for the built and bare soil classes in this study because these are at a lower level of detail when compared to the CORINE land cover classes. According to the nomenclature for the CORINE land cover data, the 'continuous urban fabric' is applicable when artificially surfaced areas cover more than 80% of the surface (EEA, 1995). None of our neighbourhoods have that level of impervious surface coverage. Neighbourhoods with the highest built-up coverage expressed as a percentage are Newton (64%), Pretoria West (65.1%) and Tembisa (66.3%). Therefore these neighbourhoods are considered to be continuous urban areas in this study. The others are considered to be areas with discontinuous urban fabric. In all our neighbourhoods there is road infrastructure of significant width and length as well as industries. Due to this heterogeneity, a geometrically averaged weight is used for built-up pixels, averaging over the road (2.23), industrial (2.07) and discontinuous urban fabric (1.00) weights. The three continuous urban fabric neighbourhoods are exempt from this tripartite weight. The 'mine, dump and construction sites' (10.99) weight is applied to bare soil pixels, with the exception of two neighbourhoods where geometrically averaging 'agricultural areas and arable land' (0.64) and mining weights seemed appropriate.

A k-means clustering procedure is used to identify homogeneous groups of land cover characteristics referred to in our varying-intercept model. A starting point in a k-means cluster algorithm, is the specification of the number of clusters k. The objective of the algorithm is to find k cluster centroids m_j where $j = 1, 2, \ldots, k$, for a data set $\mathbf{X}_{n\times p}$ in p-dimensional space thereby obtaining partition sets $\mathbf{V} = \{V_1, V_2, \ldots, V_k\}$. The objective function that ensures minimum discrepancy between data points and the

cluster centroid in each partition is expressed as,

$$O = \sum_{i=1}^{k} \sum_{x_j \in V_i} \parallel x_j - m_i \parallel^2 \tag{3.8}$$

with our choice of distance function $\parallel x_j - m_i \parallel$ being Euclidean. The k-means cluster algorithm is implemented iteratively with the initial cluster centroids assigned randomly and convergence being reached when all data points have been assigned and the within cluster distances are minimum. Multiple randomly selected initial cluster centroids are considered to avoid convergence of the algorithm to a local minimum. Further, the specification of the number of clusters k follows the use of an elbow criterion, a graphical tool where the profile of the intra-cluster to inter-cluster variance ratio is assessed for 'kinks' or change points, especially those below 50% as candidates for the number of clusters.

One of the challenges of k-means cluster analysis is the difficulty of interpreting clusters, therefore we use a landscape diversity metric to interpret our clusters. The Shannon evenness index which is expressed in Equation 3.9 describes the balance between land cover classes in each circular neighbourhood (Leitão and Ahern, 2002). That is,

$$-\frac{\sum_i p_i \ln p_i}{\ln N} \tag{3.9}$$

where the maximum number of classes considered is N and the proportion of pixels assigned to class i is p_i. This index ranges from zero to one, where low values (< 0.5) indicate lack of variety or evenness in land cover composition whereas high values (> 0.7) indicate evenness.

3.4 Results

3.4.1 Ensemble land cover classification results

Figure 3.4 shows the bare soil output from the three ML iterations for the PTA West AOI. The white areas are pixels attributed to the other classes, namely the built, water and vegetation classes. The yellow areas correspond to an ensemble BS score of 3, showing pixels that are attributed to the bare soil class in all three iterations of the ML classifier. Pixels with an ensemble score of 2 occur in close proximity to the yellow areas on the map and these pixels are areas where bare soil is mixed man-made features in some cases and with degraded grass in others. Dark green pixels have a BS score of 1, being made of mostly areas of degraded grass mixed with bare soil.

The results for interim accuracy assessment which are specific to the bare soil class are presented in Table 3.2 and Figure 3.5. The overall preliminary accuracy for bare soil is 88%. For the seven AOIs, accuracy is lowest for Newtown at 65% and highest for Kliprivier at 98%. In addition to assessing

Table 3.2: Binomial assessment of accuracy for the bare soil (BS) class

AOI	$\hat{p}$[a]	90% Conf. int.[b] lower bound	90% Conf. int.[b] upper bound
Bodibeng	0.95	0.93	0.97
Booysens	0.83	0.79	0.87
Diepkloof	0.84	0.81	0.87
Kliprivier	0.98	0.98	0.99
Newtown	0.65	0.61	0.70
PTA West	0.79	0.73	0.85
Tembisa	0.87	0.82	0.91
Overall	0.88	0.87	0.89

[a] Estimated probability of success, where success is defined by the number of pixels from the validation sample that were correctly classified as bare soil
[b] Lower and upper bounds respectively for the 90% confidence interval for $\hat{p}$

bare soil classification accuracy, classification quality is also assessed based on an analysis of confidence. Classification confidence in Figure 3.5 is "a measure of confidence that quantifies how closely a classified observation matches the exemplars of the training set" (Strahler et al., 2006). Fourteen classification confidence levels are defined relative to predefined discrete points on the cumulative distribution of the rejection fraction from 0.0 to 0.995. A confidence level of 1 corresponds to a rejection fraction of 0 meaning that every cell can be correctly classified, whereas a confidence level of 14 is indicative of cells that are furthest from the mean vector of the input signature and therefore 99.5% (0.995 rejection fraction) of such cells are at risk of misclassification. Apart from Diepkloof in Figure 3.5, more than 70% of pixels attributed to bare soil in the other AOIs are likely to be misclassified. Classifier uncertainty is lower for Diepkloof relative to the other seven AOIs, with only 30–45% of pixels with level 14 and 17–23% of pixels with at least 0.5 probability that incorrect BS class assignment will be rejected (confidence level ≤ 8).

Land cover classification maps in Figure 3.6 show high prevalence of the built class in Tembisa, PTA West and Newtown. This is confirmed by coverage statistics in Table 3.3 which presents final classification results after the classifier was improved by additional training on features where bare soil was prevalently misclassified. The mixed bare soil class corresponds to large areas with degraded grass in Diepkloof and Booysens. From Table 3.3 AOIs with the highest proportions of vegetation pixels are Kliprivier, Olievenhoutbosch and Diepkloof. The least built-up AOI is Kliprivier. Table 3.3 also shows results for the other 16 neighbourhoods which were classified using the improved ensemble classifier. Wattville has the highest proportion of

pixels classified as bare soil, whereas Sharpeville has the largest area at approximately 1.4 km^2 that is classified as water. Further, areas with highest average PM_{10} concentration corresponding to days with wind speeds in excess of 4 m s^{-1} have the least (< 2 km^2) bare soil coverage.

Etwatwa is randomly selected for the final classification accuracy assessment. The test sample consists of 384 randomly selected pixels. The percentage of pixels assigned to the water class is 0.03%, therefore only four pixels are in the validation sample. From the intra-class kappa assessment results in Table 3.4, all four pixels are incorrectly classified as water. They are bright rooftops of non-residential buildings. The overall accuracy is $\kappa \approx 0.78$ with 12% standard deviation. Table 3.4 shows that the performance of the classifier is superior for vegetation pixels. Reproducibility is also highest for this class. Apart from water, the built-up class according to Table 3.4 is challenging for the classifier because of a higher risk of confusion with either bare soil or mixed bare soil classes.

3.4.2 Relating land cover with PM_{10} concentrations

Figure 3.7(a) is an explorative assessment of whether areal coverage of bare soil as a source of fugitive dust emissions in a neighbourhood is correlated with average 'windy day' PM_{10} concentrations. The scatter-plot reveals three broad groups and Diepsloot as an outlying observation. The first group includes the three AOIs namely Wattville, Etwatwa and Kliprivier, with the largest bare soil coverage, which appears to be positively correlated with PM_{10}. The other two groups identified from Figure 3.7(a) show that if bare soil coverage is small (an area less than 10 km^2), there seem to be a negative correlation with PM_{10}. The two groups differ in their intercepts. From Figure 3.7(b–c), vegetation and built-up coverage when considered individually are not predictive of ambient PM_{10} concentrations.

Cluster analysis is performed to formally identify homogenous groups of observations (air quality stations) based on the multivariate space consisting of wind-related average PM_{10} and areal coverage of the bare soil, mixed bare soil, built-up, vegetation and water classes. Figure 3.8a shows changes in intra-cluster variance as a result of changes in cluster size, which is the basis of the elbow criterion for selecting a plausible number of clusters (k) for the k-means cluster algorithm. Cluster sizes four and six are identified from Figure 3.8a as change points in the intra-cluster variation curve. A six cluster solution is chosen because it corresponds to a lower intra-cluster variance and better separation of the groups with respect to land cover composition and PM_{10} levels is observed (Figure 3.8d).

We observe two main outlier clusters in Figure 3.8b, namely Cluster 2 and 6, with the latter representing high ambient PM_{10} levels (more than two standard deviations from the mean) for a neighbourhood with the smallest built-up footprint and the largest coverage of open fields with vegetation

Table 3.3: PM_{10} and wind summary statistics from air quality observations from the period March 2011 – February 2015 and land cover estimates from ensemble classification of SPOT 6 images taken 17 March and 17 April 2013

AOI	Cluster	Land cover (in km^2)[a]					PM_{10}[b]	Wind sp[c]	Wind dir (Spr)[d]	Shannon[e] index	Janssen[e] β index	E_{dust} indicator[e] (10^9 g m^{-2})
		BU	V	BS	mBS	W						
Bodibeng	4	33.2	10.9	3.4	16.4	0.1	58	4.5	ESE (ENE)	0.72	0.98	8.4
Booysens	4	28.3	13.6	2.0	20.1	0.0	35	6.1	NNW (NNW)	0.72	0.84	10.9
Diepkloof	3	12.7	27.2	4.7	19.3	0.2	56	8.4	NNE (NNE)	0.78	0.88	64.9
Kliprivier	3	2.6	42.1	7.4	11.9	0.2	55	8.1	NNW (NNE)	0.61	0.41	35.6
Newtown	4	41.0	8.1	2.0	12.7	0.2	26	4.6	NNE (NNE)	0.62	0.98	4.5
PTA West	4	41.7	7.0	1.2	14.0	0.1	49	4.7	ENE (ENE)	0.58	0.94	14.8
Tembisa	4	42.5	3.7	1.5	16.3	0.0	84	4.7	WNW (ENE)	0.55	0.98	14.8
Bedfordview	1	14.8	32.9	1.3	14.9	0.1	96	5.0	SSW (NNW)	0.69	0.59	13.0
Buccleuch	3	8.7	34.8	1.5	19.0	0.2	67	4.3	ENE (E)	0.66	0.57	0.7
Diepsloot	6	5.8	40.3	1.5	16.3	0.1	157	5.1	NNE (NNE)	0.60	0.47	1.2
Etwatwa	5	8.0	27.2	12.5	16.3	0.0	65	4.6	SSW (NNW)	0.80	1.26	14.7
Germiston	3	17.6	29.2	2.2	14.5	0.6	43	4.4	WSW (WSW)	0.75	0.72	1.4
Ivory	1	27.7	20.7	1.2	14.5	0.0	112	4.6	NNE (E)	0.71	0.76	2.9
Jabavu	1	28.2	21.0	0.6	14.3	0.0	64	-	-	0.69	0.68	-
Mamelodi	1	19.5	27.5	1.4	15.6	0.0	81	4.5	SSW (WNW)	0.72	0.69	5.5
Olievenhoutbosch	3	6.3	40.9	1.9	14.9	0.0	57	4.5	ESE (ENE)	0.60	0.53	10.3
Orangefarm	3	11.7	36.9	2.9	12.5	0.0	68	6.1	ENE (ENE)	0.68	0.68	20.8
Rosslyn	3	11.5	35.8	5.2	11.5	0.0	22	4.4	ENE (ENE)	0.71	0.86	11.1
Sebokeng	3	15.4	33.6	2.3	12.7	0.0	47	4.6	NNE (NE)	0.70	0.67	27.3
Sharpeville	2	16.6	31.3	3.4	11.4	1.4	65	4.6	WNW (NNW)	0.77	0.78	36.8
Thokoza	1	35.6	14.0	0.3	14.0	0.1	105	5.4	WSW (NNE)	0.64	0.81	11.4
Three Rivers	3	21.1	27.4	0.1	14.7	0.7	53	4.6	NE (NE)	0.70	0.52	13.5
Wattville	5	11.6	14.2	24.3	13.8	0.2	75	5.4	NNW (NNE)	0.85	1.74	164.3

[a] Land cover classes: BU– Built-up, V– Vegetation, BS– Bare Soil, m-BS– Bare soil mixed with man-made features or degraded vegetation, W– Water
[b] The average PM_{10} concentrations in $\mu g\, m^{-3}$ for days when wind speeds exceeded 4 m s^{-1}
[c] Average wind speed exceeding 4 m s^{-1}
[d] The most prevalent wind direction over all seasons, with Spring seasonal prevalence in brackets
[e] Landscape metrics based on land cover: Shannon's evenness index for diversity; Janssen's β is an indicator of expected pollutant response to land cover and use (Janssen et al., 2008); E_{dust} is a proxy for wind blown dust emissions of PM_{10} based on bare soil coverage

Table 3.4: Evaluating the performance of the ensemble land cover classifier through an assessment of the intra-class Kappa coefficient

Class	User's accuracy				Producer's reliability			
	Naïve agreement	κ_{i+}	$\sigma(\kappa_{i+})$	CV(κ_{i+})	Naïve reliability	κ_{+j}	$\sigma(\kappa_{+j})$	CV(κ_{+j})
W	0	0	0	-	0	0	0	-
V	0.99	0.98	0.02	2%	0.97	0.96	0.02	2%
BU	0.71	0.62	0.05	9%	0.79	0.71	0.05	8%
BS	0.76	0.70	0.05	8%	0.89	0.86	0.05	5%
m-BS	0.90	0.88	0.06	6%	0.57	0.52	0.06	12%

and degraded grass mixed with bare soil. Clusters 1, 5 and 6 have higher PM_{10} concentrations than the average of 67 $\mu g\,m^{-3}$, whereas the other three clusters have PM_{10} concentrations that are lower. Cluster 5 consist of Wattville and Etwatwa which have the largest bare soil coverage and PM_{10} values that exceed the cluster average. With the exception of Cluster 4, neighbourhoods with lower than average PM_{10} values are characterized by higher than average vegetation cover and water bodies. Clusters 1 and 4 are the only clusters consisting of neighbourhoods with higher than average proportion of built-up area, however they have conflicting PM_{10} responses (Figures 3.8b and 3.8d). The Shannon evenness indicator describes the diversity of land cover types within each circular neighbourhood and Figure 3.8c illustrates how this varies between the six clusters of neighbourhoods. Land cover composition is more evenly distributed across the five classes for neighbourhoods in Cluster 1, 2, 3 and 5. The latter cluster consists of two neighbourhoods with the biggest areal coverage of bare soil compared to neighbourhoods in other clusters. Neighbourhoods in Clusters 4 and 6 exhibit lower levels of evenness due to dominance of the built-up class in Cluster 4 and vegetation in Cluster 6 (Figure 3.8c, Table 3.3). Variation in land cover composition is the highest amongst neighbourhoods in Cluster 4 which matches with the high variation in average PM_{10} for this cluster (Figure 3.8b).

Table 3.5: Varying intercept model results relating land cover patterns to ambient PM_{10}

Coefficients	Estimate	Std. Error	Pr($>|t|$)
Intercept	97.78	8.27	$< 1e-3$
Cluster 2	-36.03	18.74	0.074
Cluster 3	-47.60	10.00	$< 1e-3$
Cluster 4	-48.32	11.00	$< 1e-3$
Cluster 5	-35.68	17.99	0.066
Cluster 6	59.12	18.36	$< 1e-2$
E_{dust}	0.09	0.14	0.525

p-value: $< 1e-3$; R^2: 0.79 and adjusted R^2: 0.71

Land cover characteristics as represented by the six clusters significantly (p-value $< 1e-3$) explain more than 70% of the variability in average PM_{10}

concentrations associated with wind speeds in excess of 4 m s^{-1} (Table 3.5). Further, at 10% level of significance, all six clusters are significant predictors of observed PM_{10} levels. From pairwise analysis of variance comparisons, we found Cluster 6 to be significantly different from all other clusters and Cluster 1 to be significantly different from Clusters 3 and 4. Our proxy for wind-blown dust emissions E_{dust} in Table 3.5 is not statistically significant as a predictor for observed wind-related average PM_{10}. A model with the adapted Janssen's β-indicator as the slope term was also considered. The results indicated that it is also not a statistically significant predictor of observed wind-related average PM_{10}. From Figure 3.9b, high ambient PM_{10} levels are expected for neighbourhoods in Clusters 4 and 5, whereas lower concentrations are expected for Cluster 6. This is in contrast to observations where the highest PM_{10} value corresponds to Cluster 6 and the median value for Cluster 4 is the lowest in Figure 3.8b. Similarly the low levels of PM_{10} emission from bare soil indicated by E_{dust} for Clusters 1 and 6 in Figure 3.9a are in contrast to the high observed PM_{10} values for these clusters in Figure 3.8b.

3.5 Discussion

PM_{10} is an erratic pollutant influenced by numerous local sources of emissions, hence finding covariates that capture local variation of concentration is an important step for statistical mapping of PM_{10} (Janssen et al., 2008). Spatially extensive covariates, like those derived from land cover and use, are valuable for this, especially in regions where the air quality network is minimal and the data cannot realistically support region-wide regulatory decisions. However, usefulness of a spatially extensive predictor depends on the strength of its correlation with the target variable. Our interest was on land cover data with one caveat being the lack of high spatial resolution data for our study period when the research was undertaken. We therefore considered land cover classification using available SPOT 6 images taken during our study period as a starting point towards our objective of assessing the proportion of variability of PM_{10} concentrations attributable to land cover characteristics. We considered the classification of four major land cover classes, namely bare soil, vegetation, built up and water bodies. However, separability from a satellite image with limited spectral resolution is a challenge in complex landscapes like in urban areas if land cover consists of bare soil, rocks, degraded grass and soil aggregates which contain synthetic materials (Myint et al., 2011). To overcome this, we focussed on developing an ensemble classifier, by training iteratively over areas of bare soil from locations which differed in soil types. The particular focus of our classifier on bare soil aligned with our interest on dust emission reservoirs because of the mining heritage of our study region.

The ensemble classifier is able to discriminate some synthetic materials from bare soil, such as waste and leachate on mine residue deposits (Figure 3.10a) and bare soil areas in a waste treatment plant (Figure 3.10c).

However, mixtures of vegetation and chemically treated soils that contain synthetic materials used to cover the three mine residue deposits are a source of confusion. This is partly attributable to erosion of vegetation cover due to the time that has lapsed since last rehabilitation and deposition of fine dust from reprocessing of nearby smaller MRDs for residual gold (Kneen et al., 2015). Human-induced degradation is another source of confusion here, given that during the period from 1952 until 2011 more than 700% growth in housing has been realised within 3 km of the three MRDs (Kneen et al., 2015). Built-up areas tend to be misclassified as bare soil, especially clay-tile, thatched and canvas rooftops in areas with unpaved roads. Therefore, as shown in Table 3.4, the performance of the classifier is lower for the bare soil and built-up classes. Higher accuracy is achieved for neighbourhoods with lower building density and low landscape heterogeneity, while in high intensity built-up areas and heterogeneous landscapes like Newtown, classification accuracy is lower. Further work will investigate using additional information in the form of height and texture to improve discrimination between bare soil and built-up pixels. The misclassification of bright (white on the image) rooftops as water is a concern with no further action taken in this study because the proportion of land covered by water bodies is less than 0.1. In future work this will be considered during the training stage.

With respect to describing local variation in PM_{10}, monitoring stations were grouped into six clusters revealing a statistically significant relationship between land cover patterns and average PM_{10} concentrations observed when daily average wind speeds exceeded a threshold of 4 m s^{-1} which correlates with known dust episodes within the study area (Oguntoke et al., 2013). An interesting result was that Clusters 1, 5 and 6 consisting of monitoring stations with either larger built-up or bare soil areas, had PM_{10} concentrations higher than the average of 67 $\mu g\, m^{-3}$. Differences in the nature of the built-up areas and bare ground in neighbourhoods represented Cluster 1 and 4 could be contributing to the conflicting land cover and PM_{10} patterns. Neighbourhoods in Cluster 1 include dense informal settlements with unpaved roads, pavements and yards which are more prone to wind-blown dust. Neighbourhoods in Cluster 4 are also densely built-up, however these are mostly high-rise buildings and there is no settlement informality. High-rise buildings are suspected to have a shielding effect on particles emitted from dust reservoirs found on the periphery of these neighbourhoods. This could be contributing to the lower PM_{10} values in comparison to the average of 67 $\mu g\, m^{-3}$.

Based on Shannon index, we observed that the clusters varied in respect of land cover composition, with most clusters being evenly distributed rather than having land cover classes that prevalently dominate. We observed that more diverse air quality neighbourhoods had higher concentrations of ambient PM_{10} concentrations. The fugitive dust emissions proxy was not a statistically significant predictor of average PM_{10} values associated with strong winds. However, this does not imply that wind-blown dust emissions do not have an effect on ambient PM_{10} concentrations (Korcz et al., 2009; Ojelede et al., 2012; Oguntoke et al., 2013). Lack of significance for E_{dust} can

be attributed to additional assumptions considered in calculating our proxy values due to data limitations which would have widened the 20-50% variability in dust horizontal emission (E_{dust}) attributable to uncertainty (Korcz et al., 2009). Janssen et al. (2008) captured local variation in PM_{10} through a land cover and use indicator which was optimized for Europe using the CORINE land cover data. Adapting this indicator for our study area did not work because the indicator was found to be insignificant as a predictor of observed wind-related average PM_{10}. In the absence of emission inventories on which indicators that link local land cover and use patterns to local air pollution levels can be optimized, our method of identifying homogenous land cover groups based on their statistical relation to observed PM_{10} enables improved prediction of PM_{10} where there are no air quality stations.

In our previous work, the South African census 2011 small area layers on percentage housing informality and domestic energy fuel usage were explored as potential spatially extensive covariates for mapping the annual exceedance frequency of the PM_{10} national air quality standard (NAQS) (Khuluse-Makhanya et al., 2016). Three geostatistical models were compared, namely kriging with external drift, a Log-Gaussian and a Poisson generalized linear geostatistical model. Housing informality was found to be a statistically significant predictor, accounting for approximately 20% of the variability of the PM_{10} NAQS annual exceedance rate. The PM_{10} response variable of interest in this study is arguably different from the previous study, however both responses are obtained from the same daily PM_{10} observations which are realizations of the same underlying unknown complex process. Therefore, for future work, it is plausible that including proportion housing informality as a predictor could account for a portion of the remaining 30% variation in average PM_{10} values associated with strong winds that could not be attributable to land cover characteristics in this study. Another portion of the remaining variability may be attributable to spatial correlation from the underlying particulate matter emission processes. Therefore, the varying intercept model will be extended into a spatially varying intercept geostatistical model (Hamm et al., 2015). This will enable quantification of the remaining variability that is due to the spatial covariance of PM_{10} values and mapping of key PM_{10} statistics.

3.6 Conclusions

This study showed that land cover patterns in the neighbourhood of an air quality station are significant predictors of average PM_{10} concentrations, in particular on days when wind speeds have locally been observed to cause significant dust emission episodes. This justifies the use of a land cover data set in mapping air quality, especially given a spatially sparse air quality monitoring network and lack of a regional emissions inventory. An ensemble maximum likelihood pixel-based land cover classifier enabled inclusion of information on known sources of variability that contribute to difficulties in classifying bare soil through iterative training. For urban areas, this

learning can be extended to the built-up class which is also susceptible to misclassification and for which improved accuracy is important for the use of land cover as a covariate in mapping air quality. A k-means cluster analysis is effective in separating air quality stations into homogenous groups with respect to land cover characteristics in the vicinity of the air quality station that can be related to observed PM_{10} concentration.

Acknowledgment

The South African National Space Agency is acknowledged for providing imagery and the South African Weather Service for provision of air quality data from the South African Air Quality Information System.

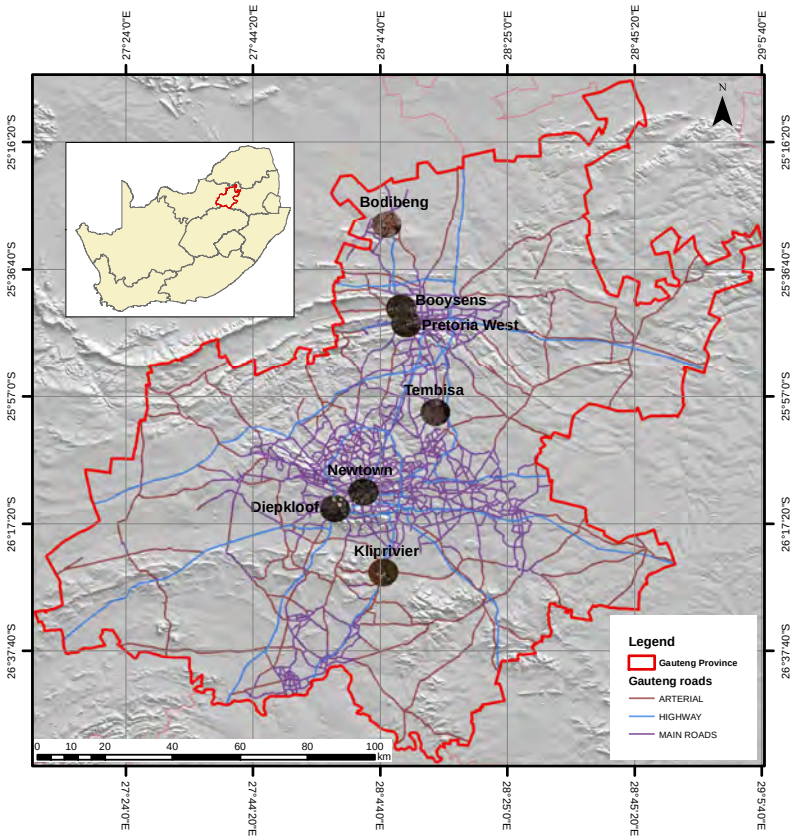


Figure 3.1: All 23 air quality monitoring stations are located within Gauteng's provincial boundaries, but for classifier development, the seven circular areas are shown with the Bodibeng, Booysens, Pretoria West, Tembisa, Newtown, Diepkloof and Kliprivier air quality stations located at their respective centres

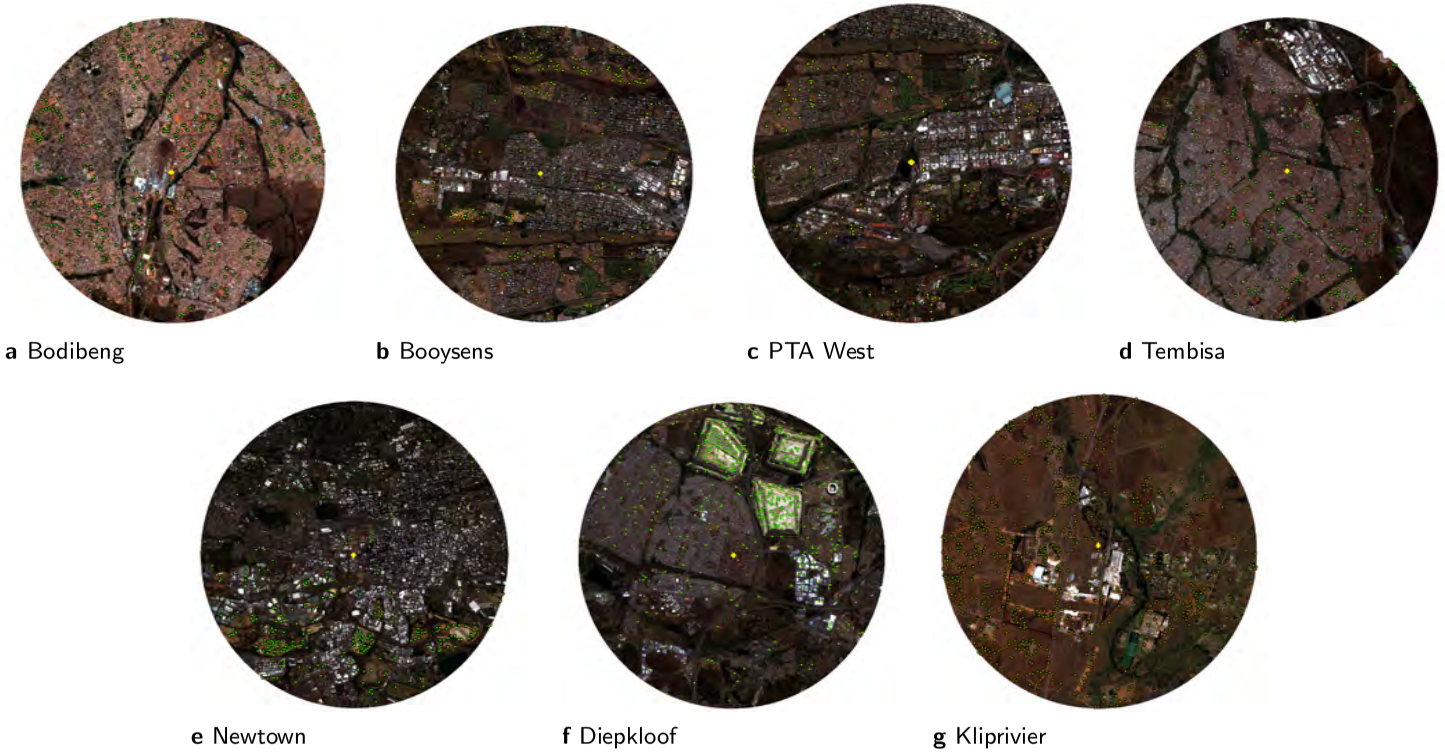


Figure 3.2: Circular areas with an air quality monitoring station located at the centre (yellow points) of each circular area. Point locations for pixels chosen for validating bare soil class are expressed in bright green

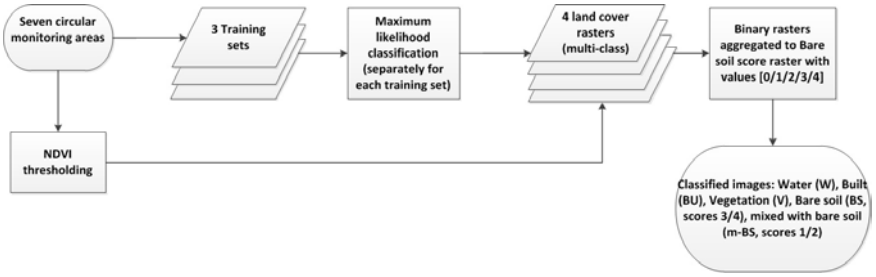


Figure 3.3: Ensemble maximum likelihood classification with focus on improved accuracy for the bare soil class

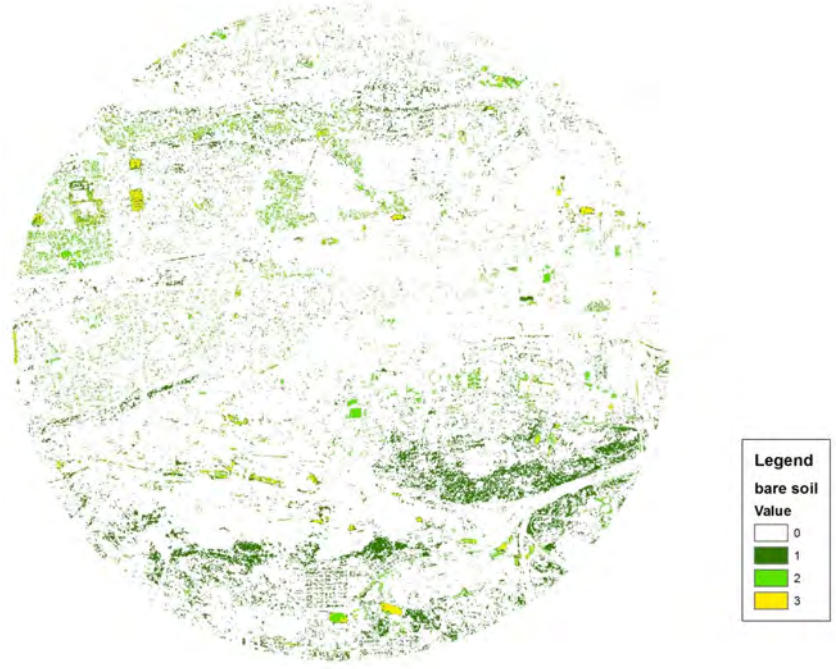


Figure 3.4: Map for bare soil derived by adding binary rasters obtained from each ML classification run (Pretoria West AOI)

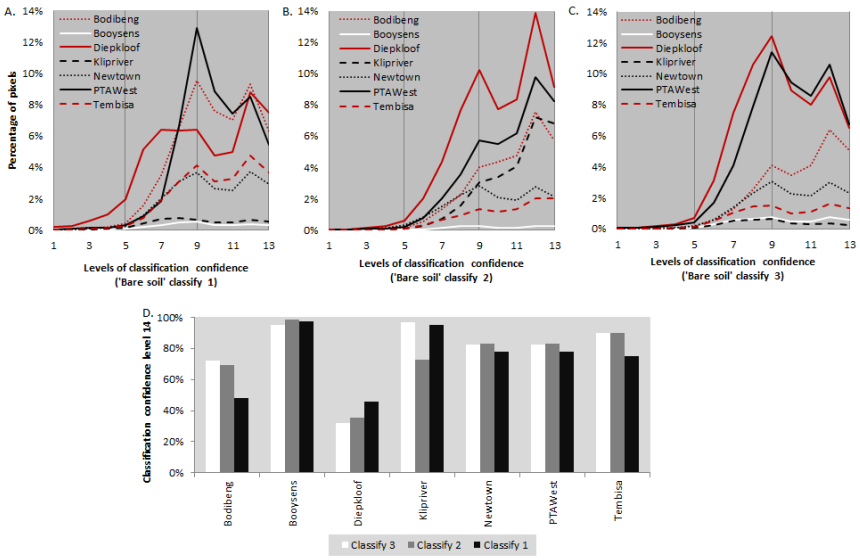


Figure 3.5: Levels of confidence for the three ML classification iterations for seven AOIs used in training the classifier. Graphics A-C show coverage (percentage of pixels) for each AOI corresponding to confidence levels 1-13 for each iteration; D. Shows the percentage of pixels per AOI for which there is least confidence of correct classification for the three iterations.

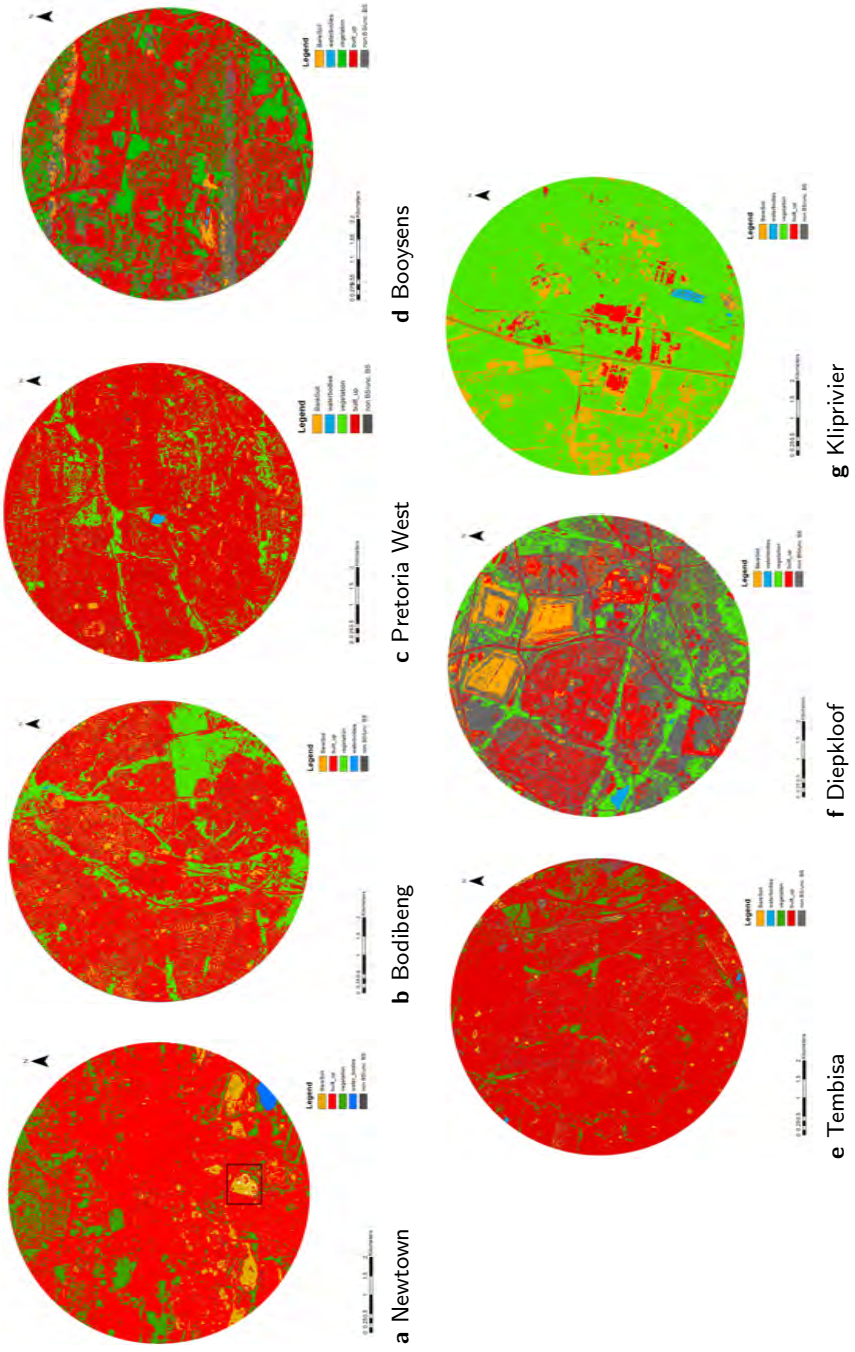


Figure 3.6: Preliminary ensemble ML land cover classification output for the seven AOIs

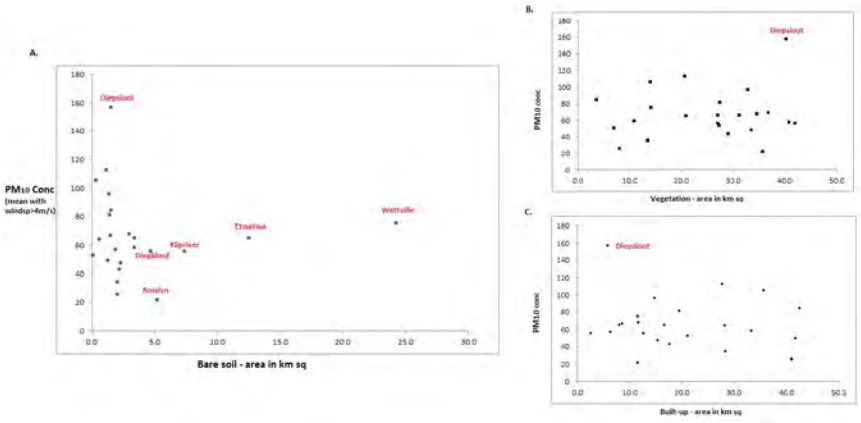
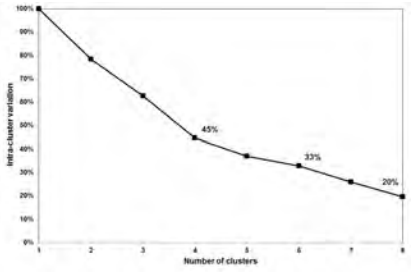
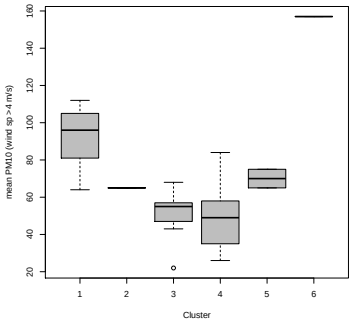


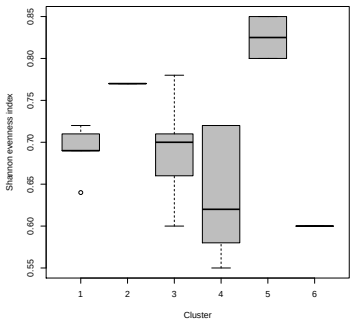
Figure 3.7: Exploratory assessment of the statistical relationship between vegetation, built-up and bare soil coverage and ambient PM_{10}



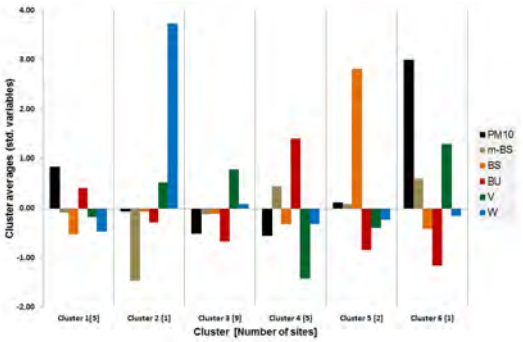
a Elbow criterion for determining the number of clusters



b The distribution of mean PM_{10} for the six land cover clusters

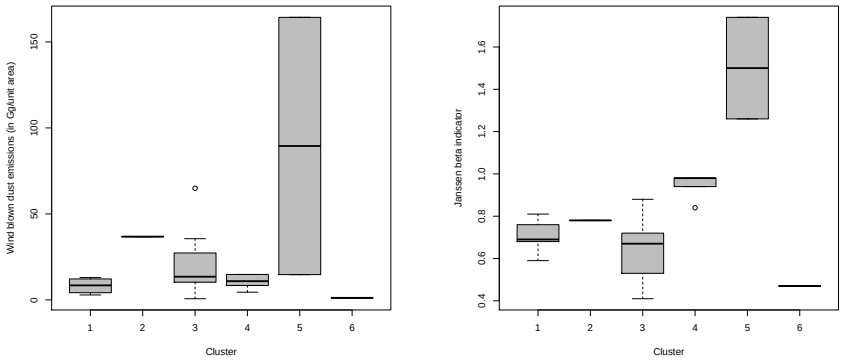


c The distribution of the Shannon evenness index for the six land cover clusters



d Deviation of land cover coverage and mean PM_{10} from the cluster mean

Figure 3.8: A graphic summary of the characteristics of the six land cover clusters



a Illustrating how the proxy for PM_{10} emissions from bare soil is distributed across the six land cover clusters

b The adapted Janssen's β-indicator, illustrating expected pollutant levels associated with land cover characteristics of the six clusters

Figure 3.9: Proxy variables considered as fixed effects in the varying intercepts model that relates land cover characteristics to observed PM_{10} values

a Man-made objects (waste) identified on top of the MRD and assigned to the built-up class

b Leachate identified as a synthetic feature and classified as built-up

c Dry soiled areas in a waste treatment plant classified as bare soil and the rest as built-up

Figure 3.10: Unique features within Newtown AOI for which the ensemble classifier successfully identified bare soil from synthetic materials (Source: Google Earth imagery, 28 April 2013)

Multiply imputing missing air quality data using bootstrap methods

4

This chapter is based on the paper: Khuluse-Makhanya S., Stein A., Debba P. Multiple imputation of missing air quality data using bootstrap methods. Submitted to *Environmental and Ecological Statistics* journal.

Abstract

High quality monitoring data are important for developing air quality regulations. Commonly, data from air quality stations, however, suffer from incompleteness. In sparse monitoring networks statistically sound imputation of missing pollutant data is a better solution than simply discarding a station's record or relying only on observed data for a secondary analysis. Assuming air quality data to be missing at random (MAR), this study explored relationships between coarse particulate matter (PM_{10}), atmospheric NO_2 and SO_2 and meteorological variables namely relative humidity, temperature, wind speed and wind direction. The aim was to develop a bootstrap based regression method to multiply impute the missing pollutant values. Challenges posed by pollutant data included non-constant variance and serial correlation. Therefore, the normal linear model for log-transformed dependent variates with the error distribution assumed to be Gaussian was applied with inference based on generalized least squares to incorporate a first order autoregressive structure for the residuals to account for temporal autocorrelation. The approximate Bayesian bootstrap (ABB) method was implemented as a benchmark against which results achieved by the bootstrap regression multiple imputation method were assessed. The performance of the bootstrap regression imputation method was also assessed by means of prediction of hold-out samples of actual data for PM_{10}. In using additional variables, namely seasonal factors and meteorological data from neighbouring weather stations, fluctuations observed in non-missing values of PM_{10}, NO_2, SO_2, relative humidity, temperature, wind speed and wind direction were preserved in the imputations. For meteorological variables, bootstrap regression imputation using data on the same variables from neighbouring weather stations was superior to direct substitution and ABB. There was an improvement in the quality of the imputations obtained through the bootstrap regression method for pollutants NO_2, SO_2 and PM_{10}. When the bias and precision of imputed values were evaluated against actuals for PM_{10}, better accuracy was achieved for the multiple imputation method based on bootstrap regressions compared to the ABB method.

Keywords: Missing data, Multiple imputation, Bootstrap regression, Circular regression, Approximate Bayesian bootstrap imputation method, Air quality data

4.1 Introduction

An air quality monitoring network is composed of point locations where stations with recording equipment are situated. These locations are often where the custodians of the network expect air pollution to be a problem. Resource constraints can result in large proportions of missing data per monitoring station. These irregularities can pose a challenge for calculating metrics such as the number of times air quality standards are exceeded per year because ignoring missing values can result in biased and imprecise estimates. An alternative to ignoring missing values is statistically valid imputation. There are various methods for imputing missing values. Advanced methods account for the mechanism or random process that leads to data being missing (von Maltitz, 2015). These methods are valuable when the imputer is the data owner, but in many applications the imputer is a secondary user of the data and the missing data mechanism is unknown. Therefore, in such cases it is common to assume that the data is missing at random (MAR). MAR implies that the probability that a datum is missing is independent of the missing data in the analysis and it depends only on the available information (Gelman and Hill, 2007; von Maltitz, 2015).

Air quality monitoring data are amenable to regression-based imputation methods because pollutants and meteorology are physically interrelated. Regression-based imputation is based on building a model on the complete portion of the data for prediction of missing values. They are referred to as conditional mean or conditional distribution imputation approaches based on whether imputed values are deterministic predicted values $\hat{\mathbf{y}} = \mathbf{x}\hat{\beta}$ or whether they are simulated values from the posterior predictive distribution $P(\tilde{\mathbf{y}}|\mathbf{y}, \theta)$ where the regression parameters θ are drawn from their joint posterior distribution. Single conditional mean imputation is not ideal when the proportion of data that is missing is substantial because of the risk of obtaining incorrectly low variance among imputed values and the introduction of bias in the completed data set (von Maltitz, 2015). When the proportion of missing data is low, single imputation based on simulating from the posterior predictive distribution is considered sufficient and has been noted to avoid the problem of artificially low variance for data completed through imputation. The main limitation with single imputation is that parameter uncertainty due to variation between imputed values cannot be assessed.

Multiple imputation (MI) methods are an alternative to single imputation methods where each missing value is replaced with several ($M > 1$) plausible values resulting in M complete data sets (Rubin, 1996). Combining rules are applied to obtain overall estimates and to assess the uncertainty attributable to imputing the missing data within and between the imputed data sets. Core to multiple imputation is the selection of a statistically valid method that will be implemented M times to yield multiple values for each missing datum. Statistical validity refers to unbiasedness of point estimates and coverage of estimated confidence intervals that do not exceed actual intervals (if data were not missing). The approximate Bayesian bootstrap (ABB) is a simple

non-parametric multiple imputation method, where for a sample of size N with $N - n$ missing observations, M imputations $Y_{\text{miss}}^{(1)}, Y_{\text{miss}}^{(2)}, \ldots, Y_{\text{miss}}^{(M)}$ are generated by repeating a two-step resampling procedure (Rubin and Schenker, 1986; Schafer, 1999). The two-step procedure starts with sampling with replacement n observations from the observed data (Y_{obs}) and is followed by sampling with replacement the $N - n$ missing values from the new set of observed data (Y_{obs}^{*}). Even though ABB is a valid imputation method (Efron, 1994), the variance estimator for multiple imputations was found to be biased for moderate sample sizes and ways to reduce this bias were proposed (Kim, 2002; Parzen et al., 2005).

Multivariate incompleteness is a common feature of air quality data and therefore a property that is desirable for such data is an imputation method that handles multivariate missing data (Heitjan and Little, 1991). Proper imputation methods are required to include all variables involved in the definition of estimands and response mechanism because omitting variables can lead to correlation biased towards zero for variables that should be correlated (Rubin, 1996). Imputation methods based on sampling from the full joint multivariate distribution are applicable for multivariate missing data, however differences in the variation structures of targeted variables often makes specification of appropriate joint multivariate distributions difficult in practice (von Maltitz, 2015). In the simple ABB imputation method, missing values can be imputed separately for each variable. Accounting for multivariate missing data in ABB can be achieved by stratifying the response-predictor variable space and implementing the imputation technique per stratum. A concern for using ABB with stratification in situations where the response and predictor variable are known to be correlated is that it does not overcome the problem of correlation being biased toward zero because the method is not effective in preserving the joint multivariate distribution of the data.

Sequential regression multivariate imputation (SRMI), also known as the fully conditional specification is a multiple imputation method that circumvents the challenge of joint multivariate distribution specification by approximating the multivariate distribution as a product of univariate conditional distribution functions whose parameter are obtained from individual imputation models for each variable that has missing values (Raghunathan et al., 2001; White et al., 2011). The univariate conditional distribution functions must be built and sequenced (or chained) in such a way that the response-covariate correlation structure is preserved (Schenker and Taylor, 1996; Schafer, 1999). Denoting variables with missing values as $\{\mathbf{Y}_j : j = 1, 2, \ldots, p\}$ and those without as $\mathbf{X}$. The $\mathbf{Y}_j$ are ordered in increasing order of incompleteness, where for each, the regression model is denoted as $P(\mathbf{Y}_j|\mathbf{Y}_{-j}, \mathbf{X}, \theta_j)$. $\mathbf{Y}_{-j}$ is the set of all variables with missing values already imputed, excluding the j^{th} variable. The joint multivariate conditional density function for the

incomplete variables can be factored as

$$f(\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_p \mid \mathbf{X}, \theta_1, \theta_2, \ldots, \theta_p) \quad = \\ f_1(\mathbf{Y}_1 \mid \mathbf{X}, \theta_1) f_2(\mathbf{Y}_2 \mid \mathbf{X}, \mathbf{Y}_1, \theta_2) \times \cdots \times f_p(\mathbf{Y}_p \mid \mathbf{X}, \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_{p-1}, \theta_p)$$

where $f_1(\mathbf{Y}_1|\mathbf{X}, \theta_1)$ denotes the conditional density function for $\mathbf{Y}_1$ as the variable with the least number of missing values which is imputed first, $f_2(\mathbf{Y}_2|\mathbf{X}, \mathbf{Y}_1, \theta_2)$ as the conditional density function for $\mathbf{Y}_2$ which supersedes $\mathbf{Y}_1$ in terms of the number of missing values and $f_p(\mathbf{Y}_p \mid \mathbf{X}, \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_{p-1}, \theta_p)$ as the conditional density function of $\mathbf{Y}_p$, the variable with the most number of missing values which is imputed last during one cycle of the SRMI (Raghunathan et al., 2001).

In practice, each cycle of SRMI is initiated by imputing all missing values by simple random sampling with replacement from the observed values for each variable prior to implementing the sequential regressions and it has been reported that between 10 and 20 cycles may be needed for the regression coefficient denoted θ_j and its standard error to reach stability for variables $\mathbf{Y}_j$ conditional on all the other variables $(\mathbf{Y}_{-j}, \mathbf{X})$ (Raghunathan et al., 2001; White et al., 2011). Coefficients from the last cycle are used to specify the posterior distributions from which the missing values are singly imputed for each variable. Multiple imputations follow by M repetitions of initializing all incomplete variables with simple random imputation and iteratively fitting sequential regression models until the parameters are stable. The other option is to select imputations corresponding to M instances within a single run of iterative fitting of sequential regression models, given that the number iterations is sufficiently large (Raghunathan et al., 2001). During model building, the compatibility of the conditional regression imputation models with the outcome or substantive model needs to be considered (White et al., 2011; Bartlett et al., 2015).

We present a multiple imputation method that is based on interrelations between PM_{10}, NO_2, SO_2, relative humidity, temperature, wind speed and direction. Our method is based on sequentially implementing bootstrap regression models to impute the missing values for meteorological variables, NO_2, SO_2 and subsequently PM_{10}. Model-based multivariate imputation requires careful consideration of model assumptions and therefore we assess the suitability of the Gamma generalized linear model with log-link function and the normal linear model for a log-transformed response variable because of the positive skewness of pollutant data. We also consider incorporating an autoregressive model structure to account for serial correlation of pollutant data. We use Rubin's combining rules for multiple imputation to assess overall uncertainty due to missing information and hold-out samples of 50 observations from each station to assess the discrepancy between our imputed values and the actual observations. The approximate Bayesian bootstrap is used as a simple non-parametric method against which to gauge the performance of our method.

The rest of this chapter is structured as follows. Section 4.2 introduces the pollutant and meteorological data from five air quality monitoring stations as well as a discussion on how the data was screened. The imputation method is introduced in Section 4.3 while Section 4.4 contains details on how the individual regression models were developed and checking of model assumptions. Results are presented in Section 4.5 and Section 4.6 is a reflection on what was achieved with reference to other regression-based multiple imputation methods applied to air quality data. Section 4.7 is the conclusion.

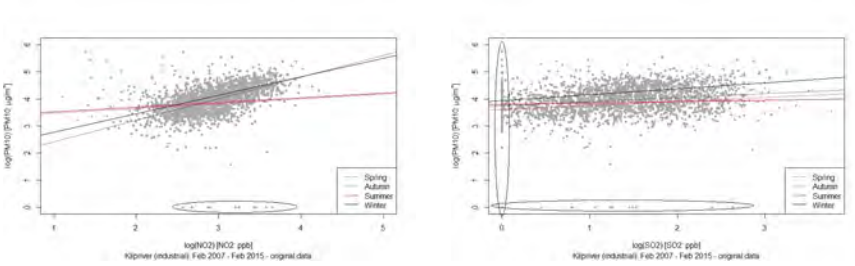
4.2 Data and the quality screening

Daily records from five air quality stations located in Gauteng and Mpumalanga provinces (Table 4.1) are considered. Dominant emission source classification for each station vary from urban background (Booysens), traffic (Buccleuch), residential (Tembisa and Ermelo) to industrial (Kliprivier). Two residential stations are considered, one in a formal settlement with a power station and coal mining nearby and the other is a township settlement with a mixture of formal and informal housing. The study period is September 2006 to February 2015. We note in Table 4.1 that observation periods differ for each station with the extent of missing data per variable determined relative to each station's observation period. Tembisa has the highest degree of incompleteness across all variables, averaging at 66% of the data set being missing, whereas Kliprivier has the least, averaging at 12% of data being missing. Pollutant data have a higher rate of incompleteness than meteorological data.

Table 4.1: Percentage incompleteness for pollutant and meteorological variables for the five selected air quality stations

Air quality station (AQS) (Obs. period)	PM_{10}	NO_2	SO_2	Meteorological variables			
				RH	Temp	Wind Dir	Wind Sp
Tembisa (Jan 2011:Feb 2015)	59	61	58	90	73	67	59
Buccleuch (Sep 2006:Jun 2014)	43	53	54	44	44	44	44
Ermelo (Jan 2008:Feb 2015)	15	30	45	13	16	12	13
Booysens (Jul 2009:Nov 2014)	35	56	57	38	32	45	46
Kliprivier (Feb 2007:Feb 2015)	24	15	15	9	11	8	8

Missing values were identified in two separate instances. In the first instance missing data were identified by network custodians or officers for known periods of instrument failure, scheduled maintenance, etc. As secondary users, the data that was received already had those values removed. In the second instance, quality screening based on guidelines for air quality data validation was used to identify suspicious observations (Ministry for the Environment, NZ, 2009; Liu et al., 2016). As part of this, practical limit values for pollutant concentrations and meteorological variables were used to detect values that were out of bounds. Practical ranges are as follows: 0–1000 parts per billion or μg m^{-3} for $NO_2, SO_2, PM_{2.5}$ and PM_{10}; 0–100% for relative humidity; 0–50 m s^{-1} for wind speed; 0–360° for wind direction and -10–50°C



a Assessing the relationship between log-transformed PM_{10} and NO_2 and the presence of seasonal differences

b Assessing the relationship between log-transformed PM_{10} and SO_2 and the presence of seasonal differences

Figure 4.1: An example of multivariate scatter-plots used to assess the correlation between PM_{10} and gaseous pollutants and to identify suspicious observations

for ambient temperature in the region. Pollutant and meteorological values outside of these practical ranges were removed.

Apart from removal of observations outside of practical limits, an assessment of patterns on time plots such as Figure 4.2 and multivariate scatter plots such as Figure 4.1 also helped in identifying suspicious values. These included consecutive sequences of zeros and other small values (close to zero) which are seen as points along the axes in Figure 4.1, spikes and $PM_{2.5}$ values which exceeded PM_{10} values. The latter required the removal of corresponding values from both particulate matter series because $PM_{2.5}$ is a component of PM_{10} and should therefore be less in quantity than PM_{10}. The reason for removing repetitive zeros (or other small values) was that they can be recorded by an instrument when atmospheric concentrations of pollutants fall below the instrument's detection limits as well as when an instrument malfunctions or is recalibrated. Suspicious values surrounded by missing values were removed, while those whose neighbouring values were higher or lower by a clear scaling factor (or drift) were corrected through multiplication by that fixed number.

Figure 4.2 shows time periods where data were missing simultaneously for all variables. In Figure 4.2 (a) attention is drawn to $PM_{2.5}$ values that are higher than PM_{10} values. These were removed for both particulate matter series. For Buccleuch this amounted to the removal of 7.4% of the $PM_{2.5}$ and PM_{10} series. Figure 4.2 (d) shows the temperature and relative humidity series, highlighting temperature observations that are within practical limits for temperature but are not plausible for South Africa. The co-occurrence of erroneous values in the temperature and relative humidity series could be the result of relative humidity being a quantity that is estimated from observed temperature rather than being directly measured. Similarities in the cyclical patterns of the temperature and relative humidity series is a

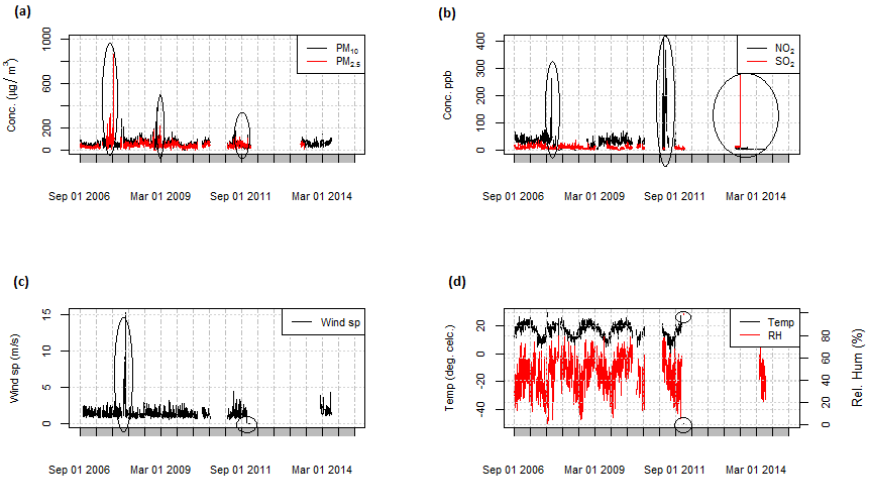


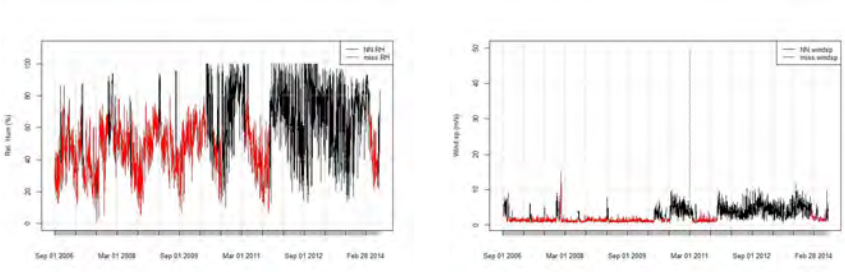
Figure 4.2: Time series plots of pollutant and meteorological variables recorded at the Buccleuch AQS showing suspicious observations in the form of spikes, zeros and recurring small values before removal during the quality screening process

reflection of that functional relationship. For Buccleuch 8.9% of the relative humidity data and less than a percent of the temperature data were removed. Figure 4.2 (b) shows consistently higher NO_2 concentrations in comparison to the SO_2 series. This is expected for Buccleuch because of its location near a large and busy highway intersection and NO_2 being an indicator of vehicular exhaust pollutant emissions.

Subsequent to quality screening, pollutant data were pre-processed to circumvent problems caused by differences in scale. In particular, NO_2 and SO_2 concentrations which were recorded and archived in volumetric units (parts per billion), were changed to units of mass concentration ($\mu g\,m^{-3}$) corresponding to units used for particulate matter. The conversion was based on factors for each contaminant's molecular weight at 0°C which are 2.85 for SO_2 and 2.05 for NO_2 (Ministry for the Environment, NZ, 2009).

The dispersion of pollutants is dependant on weather, which makes meteorological data indispensable in imputation models for pollutant data. When meteorological variables are incomplete, it is accepted practice in air quality monitoring to substitute missing meteorological observations with data from the nearest weather office (WO) (Ministry for the Environment, NZ, 2009). This was the approach considered initially for relative humidity, wind speed and wind direction. We limited the distance to the nearest WO to 20 km which is within the meso-scale of horizontal motion in the atmosphere

characterized by micro- to medium sized terrain-modulated meteorologic phenomena (Wallace and Hobbs, 2006). Irene WO is 9.5 km from the Tembisa AQS, Johannesburg WO is 17.5 km from the Buccleuch AQS, Ermelo WO is 1.4 km from the AQS, Pretoria WO is 6 km from the Booysens AQS and the Vereeniging automatic weather station is 21 km from the Kliprivier AQS. The weather office data also had missing values, but they could be used because of the small percentages of missing values which was generally less than 10%.



a Missing relative humidity values from Buccleuch AQS substituted with values from the JHB WO

b Missing wind speed values from Buccleuch AQS substituted with values from the JHB WO

Figure 4.3: An illustration of the differences between the original series with missing values for the Buccleuch AQS with values nearest weather office (Johannesburg WO)

The practice of substituting missing meteorological values with those from the nearest weather station proved problematic because of the large proportion of data that are missing. This is illustrated in Figure 4.3 where the means and variances of the incomplete relative humidity and wind speed series observed at the Buccleuch air quality station differ substantially from values that were substituted from the Johannesburg weather office. From this we can deduce that using meteorological data with missing values that were substituted with those from the nearest station would negatively affect substantive modelling that use these data as inputs. In our case the substantive model is that of PM_{10} where the meteorological and gaseous pollutant variables are predictors.

4.3 The imputation method

The inadequacy of the substitution from the nearest weather station approach for missing meteorological data prompted the formulation of the bootstrap based multiple imputation method that starts with the imputation of the missing meteorological data, followed by the missing gaseous pollutant data and ultimately the missing PM_{10} data. Consider a linear regression model relating PM_{10} to gaseous pollutants and meteorological variables $\mathbf{X} = (\mathbf{1}, \mathbf{NO_2}, \mathbf{SO_2}, \mathbf{T_{dp}}, \mathbf{W_s}, \mathbf{W_d})$, expressed as:

$$\mathbf{y} \equiv \log \mathbf{PM_{10}} = \mathbf{X}\beta + \epsilon \tag{4.1}$$

where $\epsilon \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{V})$. The covariance term $\sigma^2\mathbf{V}$ implies there is no restriction to the assumption of homoscedasticity of errors. If none of the covariates are missing, then imputing missing values based on this model would be simple. However, given that none of the covariates are complete, a regression based imputation method is developed that uses all available data recorded at air quality monitoring stations as well as meteorological data from neighbouring weather monitoring offices. Figure 4.4 is a conceptual depiction of this method which starts with multiply imputing missing temperature, relative humidity, wind speed and wind direction data using data from weather offices that are nearest to each air quality station. This multiple imputation method for the meteorological data based on the bootstrap as presented in Section 4.3.1, requires careful development regression models for each meteorological variable which is discussed in Section 4.4.1. The multiply completed meteorological data are subsequently used in predicting missing gaseous pollutant values. Finally, multiply completed meteorological and gaseous pollutant data sets are used in the prediction of missing PM_{10} values. The multiple imputation combining rules as presented in Section 4.3.2 are used at each stage to assess uncertainty within and between multiply imputed data sets, but brevity we show these results for NO_2, SO_2 and PM_{10} in Section 4.5.

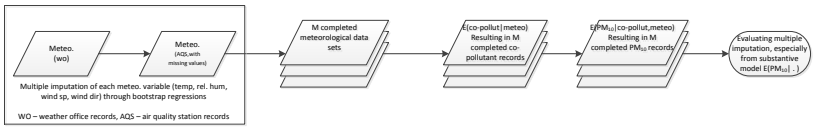


Figure 4.4: Conceptual framework for imputing missing meteorological and air pollutant observations at each air quality station

4.3.1 Bootstrap multiple imputation of meteorological variables using data from nearest weather office

Bootstrapping is a statistical method where the observed data are repetitively sampled with replacement, producing $b = 1, 2, \ldots, B$ bootstrap samples which are each subject to model fitting, with final parameter estimates obtained through averaging over all estimates obtained from fitting the model B times (Efron, 1979). In regression modelling, the sampling with replacement of $i = 1, 2, \ldots, N$ cases is either in terms of response-covariate pairs $(y_i^*; x_{i1}^*, x_{i2}^*, \ldots, x_{ip}^*)$ or the resampling of residuals. For each bootstrap sample $(\mathbf{y}_{N\times 1}^{*b}, \mathbf{x}_{N\times p}^{*b})$, a regression model is fit yielding $\hat{\beta}_{p+1}^{*b}$ through minimization of the least squares criterion similar to $\hat{\beta}_{p+1}$ the unbiased estimator of the unknown true regression coefficients β_{p+1}. Therefore B bootstrap samples result in a random sample $\left(\hat{\beta}_{p+1}^{*1}, \hat{\beta}_{p+1}^{*2}, \ldots, \hat{\beta}_{p+1}^{*B}\right)$ which is an approximate distribution for $\hat{\beta}_{p+1}^{*} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}^*$ with mean $\hat{\beta}$ and covariance

$\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$. An advantage with bootstrapping is that data-driven estimates of uncertainty are derived which is useful when there are violations of model assumptions which make approximations based on asymptotic theory inappropriate (Efron, 1979; Freedman, 1981). Resampling pairs is appropriate in this study because the predictors are stochastic rather than fixed and may be related to the errors (Freedman, 1981).

Assuming that the missing data mechanism is ignorable and data are missing at random, we develop a regression-based multiple imputation method using the resampling concepts of the approximate Bayesian bootstrap (ABB) as follows:

1. Independently obtain B bootstrap samples, where at each iteration observed pairs $\{(y_i, x_i) : i = 1, 2, \ldots, n\}$ are sampled with replacement.
2. For each first level bootstrap sample a linear model is applied, obtaining upon completion of all iterations, the set of parameter vectors $\left(\hat{\beta}^{*1}_{p+1}, \hat{\beta}^{*2}_{p+1}, \ldots, \hat{\beta}^{*B}_{p+1}\right)$ and the mean $\hat{\beta}^* = \dfrac{1}{B}\sum_{b=1}^{B}\hat{\beta}^{*b}$ and its standard error are calculated.
3. In the second level bootstrap, $m = 1, 2, \ldots, M$ regression coefficient vectors are sampled with replacement from $\left(\hat{\beta}^{*1}_{p+1}, \hat{\beta}^{*2}_{p+1}, \ldots, \hat{\beta}^{*B}_{p+1}\right)$ and used independently to predict missing values for the response variable. This results in M completed daily records for each meteorological variable.

According to Schafer (1999) for proportions of missing data observed in our air quality data set (typically between 30% and 60%), the relative efficiency of an estimate based on $5 \leq m \leq 10$ imputations is nearly equivalent to that based on an unlimited number of imputations. There are other propositions on what m should be based on arguments of efficiency and reproducibility such as the proposition that m be at least equal to the percentage of incomplete cases (White et al., 2011). Considering percentage completeness of our pollutant data against the need for efficient and practically beneficial imputations, $m = 1, 2, \ldots, 30$ is chosen for this study. The resampling for the second level bootstrap applies to the whole set B of coefficient vectors to capture the uncertainty due imputing missing values using our chosen models. The alternative would be to resample within a local neighbourhood of $\hat{\beta}^*$ which has an advantage of reducing computing time and incorporating prior information on $\hat{\beta}$ where it is available as weights in an importance sampling set-up (Efron, 2012, 2015).

The approximate Bayesian bootstrap method implemented in this study was proposed by Rubin and Schenker (1986), with the exception that $d < n$ values are sampled with replacement from the observed $\{y_1, \ldots, y_n\}$. Once d observed values have been randomly sampled with replacement, $(N - n)$ missing values are sampled from the pool of d values. These two steps are repeated independently M times. N denotes the number of observations there would be if the data were complete. The overall mean and its variance

for the completed data are obtained using the combining rules discussed in Section 4.3.2. The smaller sample size

$$d = \frac{(n-1)(N-n-1)(N-2)}{(N-1)(N-n+1)+N+n-1} < n \tag{4.2}$$

has the effect of reducing the bias of the imputation variance estimator, which occurs for moderately sized data sets and number of imputations (Kim, 2002; Parzen et al., 2005). The ABB gives an asymptotically unbiased estimator of the mean μ_y and its variance as sample size and number of imputations approach infinity, which can be violated for small sample sizes as typically observed when the proportion of missing data is substantial (Rubin and Schenker, 1986; Parzen et al., 2005). With the reduced sample size, the bias of the variance estimator is reduced by means of a correction factor f for V in Equation 4.6 resulting in the total variance estimator $V^* = f \times V$, where

$$f = \frac{\left(\frac{N^2}{n} + \frac{N-n}{M}\left(\frac{N-1}{n} - \frac{N}{n^2}\right)\right)}{\left(\frac{N^2}{n} - \frac{(N-n)N}{n}\left(\frac{3}{N} + \frac{1}{n}\right) + \frac{N-n}{M}\left(\frac{N-1}{n} - \frac{N}{n^2}\right)\right)} \tag{4.3}$$

4.3.2 Subsequent multiple imputation of pollutants and imputation uncertainty assessment

In the previous section a two-level bootstrap regression method was described for multiply imputing missing meteorological data from each air quality station using meteorological data from neighbouring weather stations. The target variable is PM_{10} which in our view is dependent on ambient concentrations of gaseous pollutants and meteorology. As a next step the multiple sets of completed meteorological data are used as covariates in multiple predictions of missing values for NO_2 and SO_2. NO_2 is predicted independently from SO_2 to accommodate the applicability of our sequential method to cases where air quality stations collect information on either one of the gaseous pollutants but not both. The final model in the sequence of regressions is the model expressed by Equation 4.1, where the covariates include NO_2, SO_2 and meteorological data. In this imputation model set-up, the error associated with imputed PM_{10} propagates from the initial step in the sequence of regressions which pertains imputation of missing meteorological data. To enable the propagation of uncertainty the imputed values starting with meteorological data, then gaseous pollutants and eventually PM_{10} are random regression predictions. This means that values to be imputed are drawn from the predictive distribution whose deviation is given by the residual stand error and is centered on the deterministic predicted value estimated from the observed-cases regression model. The prediction standard errors account for the variation in the predictive distribution, thus incorporating uncertainty into the imputed values (Gelman and Hill, 2007). Assuming a Gaussian error distribution, the missing values are randomly drawn from $\mathrm{N}(\hat{\alpha} + \hat{\beta}\mathbf{x}, \hat{\sigma}^2\mathbf{I})$, where $\mathbf{x}$ denotes the covariates.

The evaluation of imputation quality is based on the ‘combining rules in

multiple imputation' defined by Rubin (1996). Denoting observed data as $\mathbf{Y}_{\text{obs}}$, an estimand of interest is denoted as $Q = Q(\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}})$, this being the estimator of the mean μ in this instance. For a completed data set, $\hat{Q}^{(m)}$ denotes the mean estimate obtained for the m^{th} iteration, with it's variance estimate $V_w^{(m)}$ such that the multiple imputation estimator of Q that would be obtained if the data were complete is:

$$\bar{Q} = \frac{1}{M} \sum_{m=1}^{M} \widehat{Q}^{(m)} \tag{4.4}$$

The corresponding total variance estimator is an average of the within and between imputation variance

$$V = \frac{1}{M} \sum_{m=1}^{M} V_w^{(m)} + \frac{M+1}{M} V_b \tag{4.5}$$

The between imputation variance is V_b, given by

$$V_b = \frac{1}{M-1} \sum_{m=1}^{M} (\widehat{Q}^{(m)} - \bar{Q})(\widehat{Q}^{(m)} - \bar{Q})^T \tag{4.6}$$

Given that a finite number of imputations are considered ($m < \infty$), tests and confidence intervals based on a Student's t approximation imply that

$$\frac{\bar{Q} - Q}{\sqrt{T}} \sim t_\nu \quad \text{with} \quad \nu = (m-1)\left[1 + \frac{\overline{V}_w}{(1+m^{-1})V_b}\right]^2 \quad \text{degrees of freedom} \tag{4.7}$$

This is important when looking at imputation quality in terms of the degree of influence missing information has on the quantity being estimated. A greater degree of information loss results in a reduced quality of the imputation. According to Schafer (1999), the rate of information loss due to incompleteness is

$$\lambda = \frac{r + \frac{2}{\nu+3}}{1+r} \tag{4.8}$$

where r is the relative increase in uncertainty due to missing data denoted by

$$r = \left(1 + m^{-1}\right) \frac{V_b}{\overline{V}_w} \tag{4.9}$$

Increase uncertainty about imputations, namely between imputation variance V_b relative to the within imputation variance result in an increase r, an increase in the rate of missing information λ and consequently a decrease in imputation quality. Therefore the quality of missing data imputation for smaller datasets will be lower because the between imputation variance tends to be higher. Imputations for NO_2, SO_2 and PM_{10} that are obtained using our method are evaluated using these estimates of imputation quality and compared to results obtained for the approximate Bayesian bootstrap imputation method. An additional assessment of prediction accuracy for both methods on hold-out samples of observed PM_{10} is also considered.

4.4 Preprocessing

4.4.1 Building the regression models for missing meteorological data

Temperature data obtained from the weather office were the daily minimum and maximum values which differed from daily averages obtained from air quality stations. It was therefore obvious for temperature that direct substitution from the nearest weather station's record would be inappropriate and that a model based on the nearest station's maximum temperature record would be needed for predicting the missing average daily temperatures. This initiated the idea of imputing meteorological data from air quality stations based on regression models using weather station data as input variables. In this section regression models developed for each meteorological variable are discussed with graphics of singly imputed values used to show improvement that can be achieved even with single regression based imputation instead of directly substituting missing meteorological values. These regression models in the context of our bootstrap based multiple imputation method discussed in Section 4.3.1 are implemented multiple times.

Varying intercept regression models were set-up for each air quality station i with average daily temperature as the response variable $\mathbf{y}_i$ such that for stations $i = 1, 2, \ldots, 5$ and seasons $j = 1, 2, 3$ we have

$$\mathbf{y}_i = \alpha_{j[i]} + \beta \mathbf{x}_i + \epsilon_i \tag{4.10}$$

The intercepts α_j represent the changes in mean daily average temperatures recorded by the air quality monitors in response to seasonal changes, while the relation to the nearest weather station's maximum temperature is incorporated through the slope β. A Gaussian error distribution is assumed. The seasonal factor and the maximum temperature variables were statistically significant for the five stations. The adjusted R^2 was above 0.8 for Tembisa, Buccleuch and Kliprivier which indicates that the chosen predictors captured most of the variability and were therefore good predictors of average daily temperatures at those air quality monitoring locations. The smaller proportions (0.51 and 0.66) of variation explained for average temperatures in Ermelo and Booysens were unexpected because these AQSs were the closest (maximum distance 6 km) to a weather office. Figure 4.7a shows the average daily temperature series for the Buccleuch air quality station after single imputation using the random regression method on the maximum temperature data from the Johannesburg WO. We observe that the imputed temperature values are consistent with the observed values in terms of seasonality, trend and overall variation. Similar models were set-up for RH and we observe from Figure 4.7b that singly imputed RH values are consistent with the observed values, which is in stark contrast with the difference in variability observed in Figure 4.3a.

The dependence between relative humidity and temperature is a concern for collinearity which can affect their significance as predictors in the substantive pollutant models. However, both variables are important in terms of second-

ary particles formed from gaseous pollutants (Sportisse, 2009). Therefore to incorporate both variables, dew point temperature (T_{dp} in °C) which is a function of both variables is derived after missing values for both variables have been imputed. At a given constant pressure, T_{dp} is defined as the temperature at which air must be cooled for dew or frost to form, with the latter resulting from temperature below freezing point. Amongst the various expressions for the derivation of T_{dp} from RH and T, the Magnus formula (Bolton, 1980; Lawrence, 2005) was chosen for this study and is expressed as follows:

$$T_{dp} = \frac{c\ f(T, RH)}{b\ -\ f(T, RH)} \quad \text{where} \quad f(T, RH) = \ln\left(\frac{RH}{100}\right) + \frac{b\ T}{c\ +\ T} \tag{4.11}$$

Various approximations for b and c have been proposed (Bolton, 1980; Alduchov and Eskridge, 1996; Lawrence, 2005), but we chose constants $b = 17.62$ and $c = 243.12$°C that are applicable when $-45°C \leq T \leq 60°C$ and $1\% < RH < 100\%$ with maximum error of 0.1% (Alduchov and Eskridge, 1996).

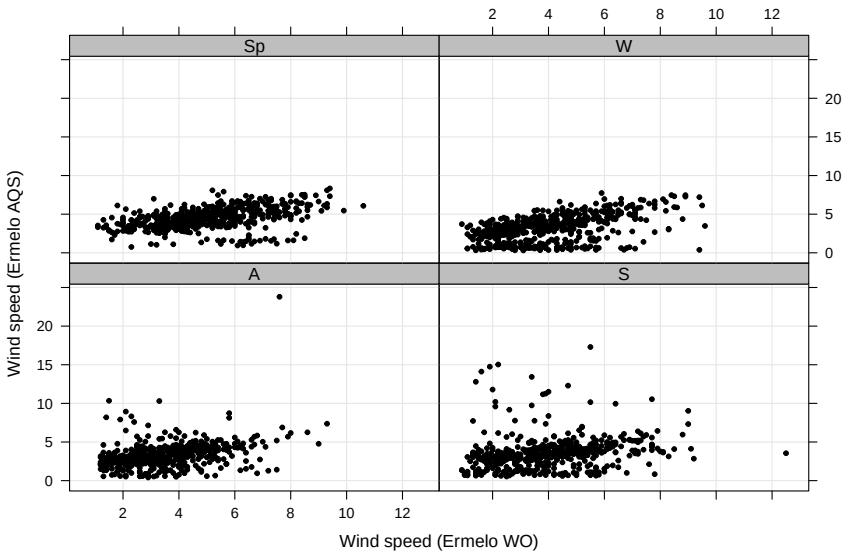


Figure 4.5: Wind speed data from the Ermelo air quality station compared with data from the nearest weather office, showing similar weak linear trend and non-constant variance per season

Wind speed data posed a challenge in terms of the fit of the varying intercepts linear regression model used for relative humidity and temperature. Wind speeds from the air quality and weather stations were weakly correlated, with the Spearman correlation coefficient ranging from 0.21 for AQSs furthest from a WO to 0.4 for Ermelo and Booysens which are within 6 km of a WO.

This was contrary to the temperature and RH data where AQS and WO observations were strongly correlated with Pearson correlation coefficients being mostly above 0.60. Scatter plots similar to Figure 4.5 indicated plausibility of the linear model for all seasons and the presence of non-constant variance. For the spring (Sp) and winter (W) plots, two groups of points can be seen, that is distinct linear trends one on lower AQS wind speed values and another on higher values.

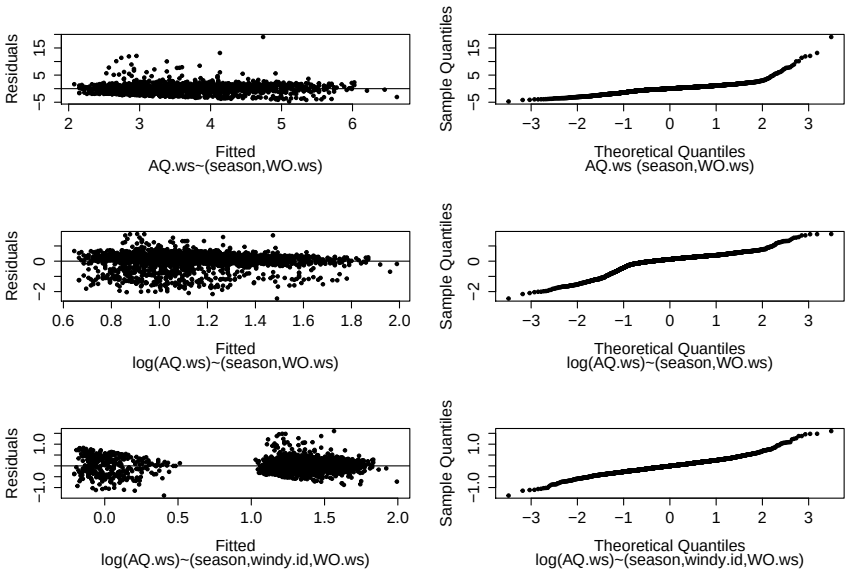


Figure 4.6: Residual diagnostics from three linear regression models fitted to wind speed data from the Ermelo air quality station which differ in covaraites used, starting with seasonal factor and WO wind speeds as predictors (top), log-transformed response variable with seasonal factor and WO wind speeds as predictors (middle) and log-transformed response variable with seasonal factor, wind intensity factor and WO wind speeds as predictors (bottom)

The starting point of finding an appropriate model was implementing the seasonal random intercept linear model presented in Equation 4.10. For this model, the proportion of variation of the AQS's wind speed values explained by seasonal changes and the observed wind speeds at the nearest WO was generally small (less 0.20) for all stations and residual diagnostic plots similar to the first two plots in Figure 4.6 for Ermelo indicated error distributions with heavier upper tails than the Gaussian distribution. There were also indications of non-constant error variance from the plots of residuals against fitted values. Remedying these deficiencies with log transformation of the response variable was not effective because that resulted in reduced adjusted R^2 and persistence of non-constant error variance as shown in the second set of diagnostic plots in Figure 4.6. An effective modification was to incorporate

an indicator term $\gamma_{[i]}$ for calm ($< 2\ \mathrm{m\,s^{-1}}$) versus windy conditions for station i and to keep the response variable log-transformed (Figure 4.6).

$$\log\left(\mathbf{y}_{AQ.ws[i]}\right) = \alpha_{j[i]} + \gamma_{[i]} + \beta\mathbf{x}_{WO.ws[i]} + \epsilon_i \quad \epsilon_i \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I}) \tag{4.12}$$

The inclusion of an indicator variable for wind conditions for Ermelo (bottom of Figure 4.6) results in the grouped residual pattern observed on the residual scatter plot and a normal q-q plot that shows no significant violations of the assumption of Gaussian distributed residuals. The model (Eq. 4.12) resulted in increases in adjusted R^2 for all five air quality stations from less than 0.2 to above 0.7. The inconsistency in mean level and variation between wind speed data imputed from the nearest weather station and data observed at the Buccleuch AQS is reduced when Figure 4.3b is compared to Figure 4.7d. However, it is not completely eradicated due to the weak correlation of wind speed data observed between the two locations.

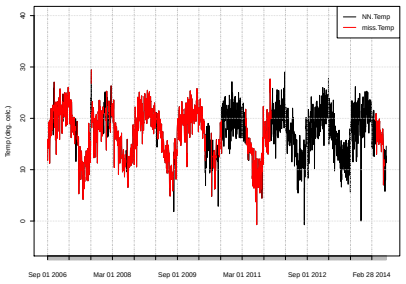
Wind direction data are circular, making the normal linear regression model inappropriate for the restricted range and periodicity associated with this type of data. Circular regression models are better suited to this type of data. Firstly, an assessment of whether there was circular correlation between AQS and WO wind direction data was performed on observed data. Significant (0.05 significance level) positive and negative associations between wind directions collected by AQSs and those recorded at the nearest WO were found. Positive angular correlations were 0.44 and 0.63 for Kliprivier and Ermelo, while negative correlations were -0.11, -0.46 and -0.55 for Tembisa, Buccleuch and Booysens. Consider wind directions converted from degrees to radians such that $(\mathrm{rad}(X_{WO.wd}), \mathrm{rad}(Y_{AQS.wd})) = (\eta, \theta)$ where $\eta \geq 0$ and $\theta < 2\pi$. AQS wind directions θ are predicted from the conditional expectation of $e^{i\theta}$ given η (the conditional characteristic function):

$$\begin{aligned} \mathrm{E}\left(\exp[i\theta] \mid \eta\right) &= \mathrm{g}_1(\eta) + i\mathrm{g}_2(\eta) && (4.13)\\ &= \mathrm{E}(\cos\theta \mid \eta) + i\mathrm{E}(\sin\theta \mid \eta) && (4.14)\\ &= \rho(\eta)\exp[i\mu(\eta)] && (4.15) \end{aligned}$$

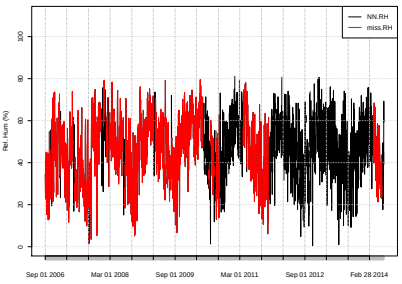
where $\mu(\eta)$ is the conditional mean direction of θ given η and $0 \leq \rho(\eta) \leq 1$ the conditional concentration towards that direction (Sarma and Jammalamadaka, 1993). The unknown structure of $(\mathrm{g}_1(\eta), \mathrm{g}_2(\eta))$ is approximated by periodic functions expressed by m-degree trigonometric polynomials such that one has the following model terms:

$$\begin{aligned} \cos\theta &= \sum_{k=0}^{m}(A_k\cos k\eta + B_k\sin k\eta) + \epsilon_1 && (4.16)\\ \sin\theta &= \sum_{k=0}^{m}(C_k\cos k\eta + D_k\sin k\eta) + \epsilon_2 && (4.17) \end{aligned}$$

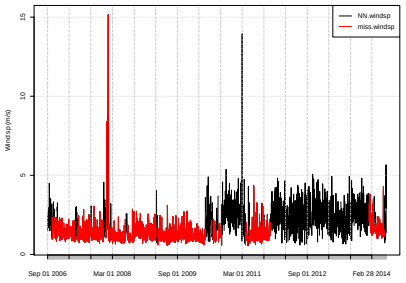
where the errors $\epsilon = (\epsilon_1, \epsilon_2)$ are zero mean random variables with unknown covariance matrix $\mathbf{\Sigma}$. In parametric inference the von Mises and the wrapped



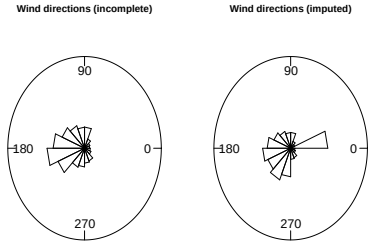
a Daily average temperature series before and after random regression imputation from the JHB WO's daily maximum temperature data



b Relative humidity values from Buccleuch AQS before and after random regression imputation from the JHB WO's observations



c Wind speed values from Buccleuch AQS before and after random regression imputation from the JHB WO's observations



d Rose diagram for wind directions for Buccleuch AQS before and after circular regression imputation from the JHB WO's observations

Figure 4.7: Showing the original meteorological data with missing values for the Buccleuch AQS and the data after missing values were singly imputed with predicted values from variable-specific regression models that used meteorological data from the weather office in Johannesburg

Cauchy distribution are commonly assumed for the error distribution.

We assume the von Mises distribution which closely approximates the wrapped normal distribution (a Gaussian distribution wrapped around the unit circle) and therefore circular regression based on this assumption is similar in terms of statistical inference to the Gaussian linear regression model considered for temperature, relative humidity and wind speed (Fisher and Lee, 1992). The von Mises density function for the conditional distribution of θ given η is

$$f(\theta \mid \eta) = [2\pi I_0(\kappa)]^{-1} \exp[\kappa \cos(\theta - \mu(\eta))] \tag{4.18}$$

where $I_0(k)$ is the modified Bessel function of order 0 and the concentration parameter is $\kappa = \rho(\eta)$ which may depend on η. The location μ and dispersion

κ^{-1} parameters are equivalent to the mean and variance σ^2 of the Gaussian distribution and it is especially for large concentration parameters ($\kappa \geq 2$) that Gaussian approximation is most applicable. For small and non-zero κ, f is unimodal and symmetrical, that is close to the circular uniform distribution which is attained when $\kappa = 0$. Irrespective of the distributional assumptions, the predicted wind directions would be given by:

$$\hat{\theta} = \mu(\eta) = \begin{cases} \arctan \frac{g_2(\eta)}{g_1(\eta)}, & \text{if} \quad g_1(\eta) \geq 0 \\ \pi + \arctan \frac{g_2(\eta)}{g_1(\eta)}, & \text{if} \quad g_1(\eta) \leq 0 \\ \text{undefined}, & \text{if} \quad g_1(\eta) = g_2(\eta) = 0 \end{cases} \tag{4.19}$$

When the θ given η is an assumed von Mises variate, the periodic functions used in Equation 4.19 are:

$$g_1(\eta) = \frac{I_1(\kappa)}{I_0(\kappa)} \cos \mu(\eta) \quad \text{and} \quad g_2(\eta) = \frac{I_1(\kappa)}{I_0(\kappa)} \sin \mu(\eta)$$

Important parameters to be estimated include the order of the trigonometric polynomial, the location μ and concentration parameters (Sarma and Jammalamadaka, 1993). Figure 4.7d shows the wind rose diagram for Buccleuch incomplete and then singly imputed wind directions data. The dominance of southerly winds (from east-south-east to west-south-west) is similar for both series. The singly imputed series differs from the incomplete series only in terms of the high frequency of winds coming from the north as shown in Figure 4.7d.

Figure 4.8(a)–(d) shows the fit of regression models discussed in this section for meteorological variables from the Ermelo air quality station. Temperatures in Figure 4.8(a) show strong linear pattern with summer and autumn temperatures being higher as would be expected. Similarly in Figure 4.8(b) RH is higher and summer and autumn due to the rainy season in this region which runs from late spring to early autumn, peaking in summer. There is more variability in RH values and this is more pronounced for values of RH above 80%. We observe in Figure 4.8(c) that for calm wind conditions, wind speeds from Ermelo AQS are weakly related to those from the weather office, but for strong wind speeds the regression model appears to fit the data well. Two dominant quadrants are observed for wind directions between the two locations in Figure 4.8(d) and the circular regression model seem to fit the data well in those two quadrants.

4.4.2 Checking of distributional assumptions for the PM_{10} model

The choice of error distribution for regression based imputation models is important, but it is common in practice that errors are assumed to be independent and identically distributed (IID) Gaussian variables without diagnostic assessment of the suitability of such assumptions. For our imputation method, a log-Gaussian linear model is proposed for imputing missing PM_{10} data (Equation 4.1) which is appropriate given a non-negative continuous response variable whose distribution is positively skewed and evidence that

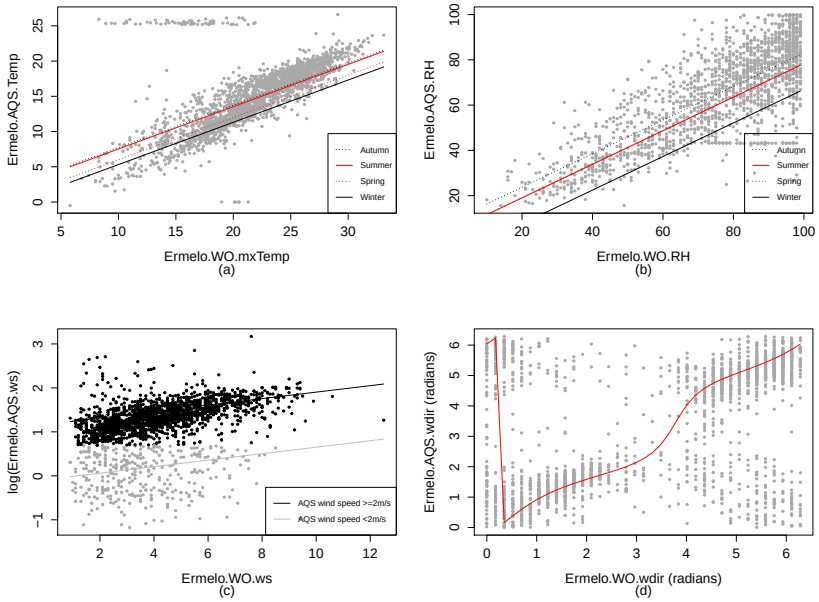


Figure 4.8: Regression model fits for Ermelo air quality station using data from Ermelo weather office: (a) Seasonal intercepts with daily maximum temperatures (WO) linearly related to daily average temperatures (AQS); (b) Seasonal intercepts with AQS and WO daily relative humidity linearly related; (c) AQS and WO wind speeds are linearly related, with shifts depending on whether AQS wind speeds considered are indicative of calm wind conditions or not; (d) Circular linear relation between AQS and WO wind directions

the error variance is non-constant. An assumption in the log-Gaussian model is that the predictors have a multiplicative effect on the response variable. In this section an analysis of the plausibility of the log-Gaussian model for PM_{10} is tested and compared with an appropriate alternative model. Regression based imputation methods such as SRMI were shown to be robust against violations of the IID Gaussian assumption when the true error distribution is moderately heavy-tailed. Despite this robustness for data from moderately heavy-tailed distributions, strong deviations from normality are a concern because of the potentially poor fit of the Gaussian distribution to extreme values (He and Raghunathan, 2009). An alternative model we consider is a generalized linear model assuming a Gamma(ν, λ) distribution with a logarithm link function which relates the expected value of the response variable to explanatory variable $\mathbf{X}$ such that,

$$\log(\mathrm{E}(\mathbf{PM}_{10})) = \alpha_j + \beta_1 \mathbf{I}_{\mathrm{pub}.h} + \beta_2 \mathbf{T}_{dp} + \beta_3 \mathbf{W}_s^{-1} + \beta_4 \sin(\mathbf{W}_d) + \beta_5 \cos(\mathbf{W}_d) \tag{4.20}$$

The parameters ν and λ are the shape and scale of the Gamma distribution. An assumption in this model is that the predictors have a multiplicative effect on the mean of PM_{10} rather than on PM_{10} itself as is the case for the log-Gaussian model. Denoting $\mathbf{PM}_{10}$ as $\mathbf{Y}$, we have $\mathrm{E}(\mathbf{Y}) = \frac{\nu}{\lambda} = \mu$ and $\mathrm{Var}(\mathbf{Y}) = \frac{\nu}{\lambda^2} = \mu^2\sigma^2$ with $\sigma^2 = \nu^{-1}$ (Faraway, 2006).

Gaseous pollutants as predictors are not considered for this preliminary analysis, but they are considered in the substantive imputation model. Dew point temperature, a function of humidity and temperature as discussed in Section 4.4.1 is used as a predictor ($\mathbf{T}_{dp}$) in Equation 4.20. Meteorological data used for this analysis includes values that have been singly imputed through random regression of values from the nearest weather office as discussed in Section 4.4.1. Varying intercepts ($\alpha_j, j = 1, 2, 3$) accommodate rate of change in mean PM_{10} due to seasonal change and the indicator variable $\mathbf{I}_{\mathrm{pub}.h}$ captures rate of change in mean PM_{10} attributed to whether it is a working day or not. Non-working days include Saturday, Sunday, public holidays and the annual Christmas closure period from 25 December to 1 January.

Table 4.2: Diagnostic assessment of the suitability of the Gamma-GLM and the Gaussian linear model with log-transformed response variable

Station	Model	Null dev. D_0	Resid. dev. D_1	Prop dev. explained	Resid. s.e.
Tembisa	Gauss	92.04	25.57	0.72	0.37
	Gamma	83.67	24.21	0.71	0.32
Buccleuch	Gauss	336.12	274.72	0.18	0.42
	Gamma	321.36	264.13	0.18	0.41
Ermelo	Gauss	1528.65	851.38	0.44	0.63
	Gamma	1301.61	800.66	0.38	0.59
Booysens	Gauss	435.32	274.48	0.37	0.54
	Gamma	401.52	258.10	0.36	0.49
Kliprivier	Gauss	586.62	414.10	0.29	0.42
	Gamma	574.34	419.61	0.27	0.43

To assess the goodness-of-fit of the log-Gamma and log-Gaussian models, we use the proportion of deviance explained which is a generalization of R^2 such that $R^2 \equiv 1 - (\mathrm{D}_1/\mathrm{D}_0)$. D_0 is the deviance for a model with just a constant term and the residual deviance, D_1, is for the fitted model. The square root of the log-Gamma's dispersion parameter estimate is comparable to the residual standard error for the log-Gaussian model. As anticipated from theory and practice (David, 1988; Faraway, 2006; Moran et al., 2007), results in Table 4.2 show that the differences in fit between the two models is not substantial. The log-Gaussian model consistently shows a slight advantage in terms of proportion of variation explained. Tembisa stands out with over 70% of the variation in $\log(PM_{10})$ being accounted for by ambient moisture and wind conditions, seasonal changes and emissions mainly from

industrial and transport sources which are related to normal working day activities. For the other stations the percentage is low at less than 40%. The residual standard error is slightly lower for the log-Gamma model which can be interpreted as the Gamma distribution being a better fit to the residuals than the Gaussian, the main difference in fit being in the tail region. Consequent to this preliminary analysis, the log-Gaussian model is chosen for imputing missing PM_{10} data because of its plausibility for the type of data, comparability to the log-Gamma model in terms of model fit and consistency with models considered for imputing missing meteorological data.

4.5 Results

For the ABB multiple imputation method each variable is considered independently and correlations between pollutants and meteorological variables are not considered. From the results in Table 4.4 there is barely any difference between the multiple imputation estimate of the mean and the sample mean calculated from the incomplete data. Looking at the means, Tembisa has the highest average daily ambient concentration of PM_{10}, whilst the highest for NO_2 corresponds to the Buccleuch AQS situated at one of the busiest highway intersection in the study region. Ermelo has the highest average daily SO_2 concentrations among the five stations considered which was anticipated due to the presence of power generation plants in the vicinity of this air quality station. Ambient concentrations of PM_{10} are highly variable when compared to gaseous pollutants. For all three pollutants considered, the station with the highest mean concentrations also has the largest variance for that pollutant.

Uncertainty within each completed pollutant series remains high for both the ABB and the bootstrap regression imputations and is similar to the sample variances of the incomplete pollutant series. Tembisa had the highest percentage ($\approx 60\%$) of missing data (Table 4.1) and we observe from Table 4.4 that imputation quality statistics V_b, r and λ corresponding to this station are larger, implying lower imputation quality in comparison to the other four stations. When comparing r and λ statistics for imputations from the ABB and bootstrap regression methods for Tembisa, these are lower for the bootstrap regression method, indicating higher imputation quality from this method. Imputation quality results for Kliprivier for NO_2 and SO_2 are interesting because the percentage of missing data is small and sampling with replacement from the same dataset results in reduced variability between the 30 imputed datasets. This leads to reduced uncertainty due to missing data. Another consequence of the reduced variability in pollutant data for Kliprivier is larger λ values which imply that multiple imputations from the ABB method for this station are less informative when compared to the other stations with a larger percentage of missing data. V_b and r from the bootstrap regression method for Kliprivier are smaller than for the ABB method, implying better quality of imputations compared to the ABB method. Imputations for Buccleuch were the closest to the actuals for the

Table 4.3: Summary statistics for PM_{10}, NO_2 and SO_2 after implementing the approximate Bayesian bootstrap and bootstrap regression multiple imputation methods for missing pollutant values

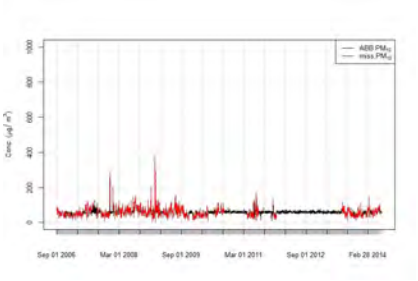
Pollut.	Station	Before Imputation[a]		After ABB Imputation[b]						After Bootstrap Regression Imputation[c]					
		Mean	s^2	Mean	$\bar{V}_w$	V_b	V^*	r	λ	Mean	$\bar{V}_w$	V_b	$\widehat{V}$	r	λ
PM_{10}	Tembisa	83	1787.06	83	1765.45	2.01	1771.51	1.2e-3	1.2e-3	80	1868.68	0.40	1869.09	2.2e-4	2.2e-4
	Buccleuch	58	791.51	59	801.50	0.27	802.39	3.5e-4	3.5e-4	56	749.42	0.14	749.56	1.9e-4	2.0e-4
	Ermelo	52	1703.22	52	1711.69	0.13	1712.29	7.9e-5	8.0e-5	54	1954.91	0.16	1955.07	8.2e-5	8.3e-5
	Booysens	57	1259.86	57	1277.15	0.63	1278.98	5.1e-4	5.1e-4	58	1397.84	0.26	1398.11	1.9e-4	1.9e-4
	Kliprivier	64	1341.00	64	1330.28	0.18	1330.96	1.4e-4	1.4e-4	62	1235.98	0.10	1236.09	8.8e-5	9.2e-5
NO_2	Tembisa	35	314.65	36	309.13	0.22	310.03	7.4e-4	7.5e-4	35	298.51	0.11	298.62	3.9e-4	4.5e-4
	Buccleuch	52	1217.50	52	1214.51	0.47	1216.15	4.0e-4	4.0e-4	50	1147.17	0.18	1147.35	1.6e-4	1.6e-4
	Ermelo	28	628.41	28	625.43	0.11	625.85	1.8e-4	2.0e-4	28	625.92	0.09	626.02	1.5e-4	1.7e-4
	Booysens	30	316.85	30	316.45	0.28	317.20	9.1e-4	9.2e-4	31	297.64	0.05	297.69	1.7e-4	4.8e-4
	Kliprivier	41	287.98	41	287.84	0.02	287.93	7.2e-5	1.7e-3	41	293.06	10e-3	293.07	3.4e-5	5.0e-3
SO_2	Tembisa	22	211.44	22	211.47	0.16	212.07	7.8e-4	8.4e-4	22	203.91	0.04	203.96	2.2e-4	9.7e-4
	Buccleuch	25	256.23	25	259.66	0.11	260.02	4.4e-4	5.2e-4	26	244.28	0.04	244.33	1.9e-4	7.1e-4
	Ermelo	37	1129.34	37	1130.27	0.34	1131.54	3.1e-4	3.1e-4	38	1127.66	0.22	1127.88	2.0e-4	2.0e-4
	Booysens	14	134.72	14	135.71	0.09	136.01	6.9e-4	1.1e-3	13	133.17	0.04	133.21	2.9e-4	2.3e-3
	Kliprivier	14	112.75	14	113.23	0.01	113.26	9.1e-5	1.5e-2	14	111.33	6e-3	111.33	5.3e-5	2.6e-2

[a] Sample means (in $\mu g\,m^{-3}$) and variances (s^2) for the incomplete pollutant data

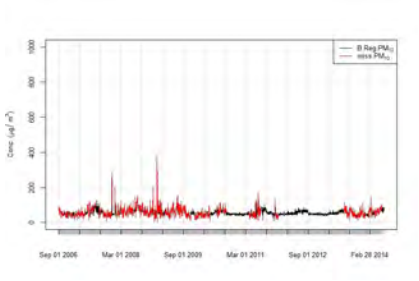
[b] Approximate Bayesian bootstrap multiple imputation summary statistics include the average within imputation variance ($\bar{V}_w$), the between imputation variance (V_b), the overall variance estimate (V^*), the rate of missing information in the multiple imputation λ and the relative increase in variance due to missing data r.

[c] Bootstrap regression-based multiple imputation summary statistics similar to summaries for ABB except the overall variance estimate is $\widehat{V}$ which implies this is not adjusted by the bias reduction correction factor as in ABB's overall variance estimate V^*.

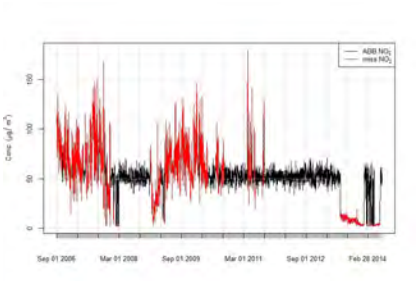
ABB method according to statistics in Table 4.4 which is consistent with Figure 4.9a showing closeness in mean level of the ABB imputed and the incomplete series for PM_{10}. The mean obtained after bootstrap regression imputation is smaller than the mean of the incomplete data for PM_{10} and NO_2, but the quality of the imputations is better for the regression method when compared to ABB imputations.



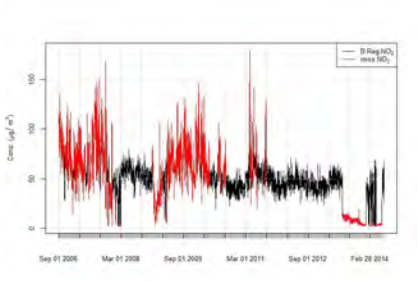
a PM_{10} ABB imputed series superimposed on the initial incomplete series



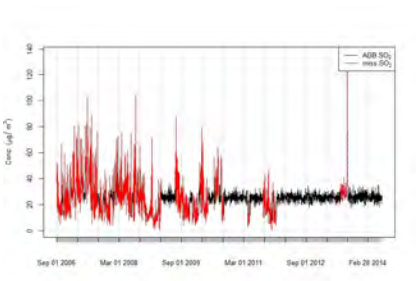
b PM_{10} Bootstrap regression imputed series superimposed on the initial incomplete series



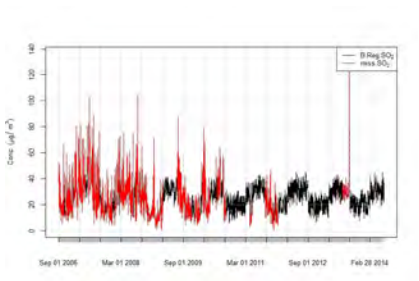
c NO_2 ABB imputed series superimposed on the initial incomplete series



d NO_2 Bootstrap regression imputed series superimposed on the initial incomplete series



e SO_2 ABB imputed series superimposed on the initial incomplete series



f SO_2 Bootstrap regression imputed series superimposed on the initial incomplete series

Figure 4.9: Each missing pollutant value for the Buccleuch AQS is filled in with the mean of the 30 plausible imputation values

Table 4.4: Accuracy of the multiple imputations from the two methods for a hold-out samples of 50 PM_{10} observations for each air quality station

Station	ABB		Bootstrap Regression	
	ME	URMSE	ME	URMSE
Tembisa	12.4	44.3	5.6	24.8
Buccleuch	-1.6	22.6	-1.4	21.6
Ermelo	5.3	32.6	0.5	22.6
Booysens	1.9	33.3	3.7	29.9
Kliprivier	-5.1	32.1	-2.4	26.0

From Figure 4.9, the NO_2, SO_2 and PM_{10} daily series completed using the ABB method, lack the seasonal pattern present in the observed data. Further, there is mean reversion of imputed values resulting in reduced variability in the completed dataset than it would be if there were no missing values. The pattern observed for the mean of the multiple imputations from the bootstrap regression method, consists of seasonal cycles similar to those observed in the incomplete series and imputed values are very close to observed values for periods where few daily values are missing. Prediction performance was also evaluated using a test sample y_i and the mean of multiply imputed values $\hat{y}_i$ and summarized using the mean error (ME) for the bias and the unbiased root mean square error (URMSE) for the precision:

$$\text{ME} = (1/50)\sum_{i=1}^{50}(y_i - \hat{y}_i) \tag{4.21}$$

$$\text{URMSE} = \sqrt{(1/50)\sum_{i=1}^{50}(\hat{y}_i - y_i)^2 - \text{ME}^2} \tag{4.22}$$

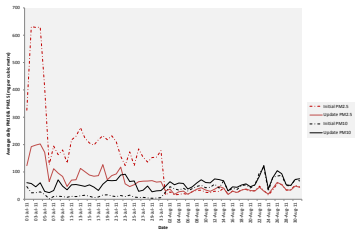
$$\tag{4.23}$$

The predictive accuracy is higher for the bootstrap regression method in comparison to the ABB method.

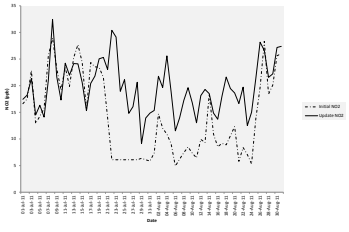
4.6 Discussion

Missing data is a common problem in air quality monitoring. This is a concern for environmental health studies which consider these data as inputs, hence we considered ways in which missing pollutant and collocated meteorological values can be imputed using available data from the air quality station and neighbouring weather stations. We also considered in detail the quality screening of data obtained from network custodians. Our quality screening method helped us identify values that were potentially errors. These suspect values should ideally be checked for and corrected by the custodian's air quality officers prior to release of the data to secondary users. Quality screening is therefore recommended as a first step for secondary users of air quality data (Le et al., 2007; Liu et al., 2016). In our case, we found that

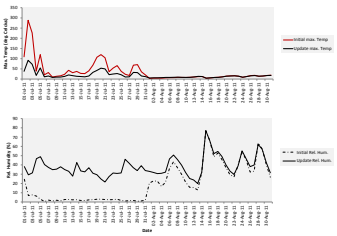
switching of personnel and custodians also contributed to the imperfections we found in our data. For an example in Figure 4.10 for the Kliprivier station we observe differences within a purposefully chosen overlapping period 1 July 2011 until 31 August 2011 between data that was acquired in 2011 and updates received in 2012. Differences are attributed to drifts that occurred in the instrument during the two-weekly zero and calibration checking period which coincided with changes in custodianship of the network. They were later corrected when the data was re-examined, hence the data acquired in 2012 was updated. With the exception of the NO_2 series in Figure 4.10b, the discrepancies between the initial and updated data sets are pronounced during the month of July. We noted that even with the updated data sets in Figure 4.10a, $PM_{2.5}$ values were still erroneously larger than PM_{10} for the month of July and therefore these were removed during quality screening.



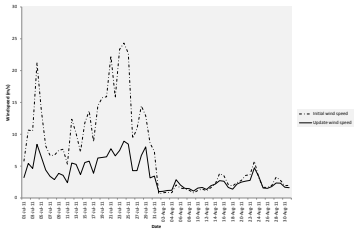
a $PM_{2.5}$ concentrations exceed PM_{10} in both versions of the data before 1 August 2011



b NO_2 values are different throughout the period of overlap



c Peaks in the average daily temperature erroneously exceed the upper limit of $50°C$ and trailing minuscule numbers are observed for the initial relative humidity data



d Wind speeds seem to be adjusted downwards for July in 2011

Figure 4.10: Differences in air quality data from the Kliprivier station for the period (1 Jul 2011 to 31 Aug 2011) where data sets acquired on two separate dates overlap

Incompleteness of covariates is a problem for regression based imputation of

missing response values because ignoring missing covariate data results in a small number of observed data that can be used to build the imputation model as well as a smaller proportion of values for the response variable that can be imputed. Our novel approach to this problem was to build a sequence of regression models starting with regressing meteorological data from air quality stations with data from neighbouring weather stations, then imputing missing gaseous pollutant values using the completed meteorological data and eventually imputing missing particulate matter values. The closest in concept to our approach is the Imputation Robot system (I-BOT) of the California Air Resources Board which considers variable-by-variable single imputation through linear regression of missing values at a target air quality station using values from a donor station which is chosen based on correlation strength and maximum distance of 100 km (Larsen and Shah, 2016). Our approach differs from I-BOT in being a multiple imputation through bootstrap method where multivariate regression models are developed instead of univariate regressions. Another alternative would be SRMI using data only from that individual station, starting with random imputation of the least missing meteorological variable ending with PM_{10}. SRMI does not take advantage of the availability of additional data from nearby weather stations like we have done in our method. Further, in SRMI there can be issues of incompatibility of the regression sequences for the covariates with the substantive model for PM_{10} which, if not remedied, can lead to systematic bias in substantive model parameter estimation and consequently compromised validity of imputed values (Bartlett et al., 2015).

Defining an imputed value as an expectation given a particular context has the consequence that imputed values revert to the mean and that precision is overstated. Our solution to these challenges was to consider covariates for pollutants and to implement the bootstrap on random regression models assuming the Gaussian distribution. Extreme value patterns were not considered and this can be pursued in future especially where the accuracy of missing values that may be much higher than regulatory thresholds is important. This would entail consideration of other models like the generalized additive and extreme value models (De Jong et al., 2016). Air quality datasets for each station are large and implementing our multiple imputation method for several stations simultaneously presented a challenge in terms of memory in R. Therefore further work is required on improving computational efficiency if a system like I-BOT is to be built to assist air quality officers with imputations using our method (Larsen and Shah, 2016). Another area of further research concerns visual diagnostics for multivariate imputations an efficient way to check whether the imputed values are sensible when one has several stations to consider at a time (Abayomi et al., 2008).

4.7 Concluding remarks

There are various reasons for missing air quality data and for secondary users of these data, multiple imputation is an effective way to impute missing values

and to quantify the uncertainty associated with the imputations. Quality screening using a combination of known physically possible ranges for the different variables and plots to identify values that are potentially invalid was performed. We found this to be an important first step as users of the data, but we conclude that it would be more advantageous if the data is quality screened by the custodians prior to dispatch so that the inclusion-exclusion criteria for suspicious values is based on expert knowledge of the station's location, instrumentation and operational conditions. Consideration of all observed data is encouraged in multiple imputation and can include data available from other types of monitoring stations as illustrated in this study with meteorological data from nearby weather stations. Covariates in the form of meteorological variables, gaseous pollutant and seasonal factor variables lead to imputations with similar structural patterns to observed values. The bootstrap regression multiple imputation method developed in this study exploits correlations between meteorological and pollutant variables, leading to improved imputation quality when compared to the approximate Bayesian bootstrap method.

5

Statistically mapping the risk of exposure to high ambient concentrations of PM_{10} and $\mathrm{PM}_{2.5}$

This chapter is based on the paper: Khuluse-Makhanya S., Stein A., Debba P., Dudeni-Tlhone N., Ngidi M. A statistical approach to air quality mapping and the risk of exposure to excessive particulate matter pollution. Submitted to the *Spatial Statistics* journal.

Abstract

Exposure to even moderate concentrations of particulate matter leads to increased risks of cardiopulmonary morbidity and mortality. Poor housing conditions increase the susceptibility to exposure. A methodology is developed to quantify population risks related to exposure to PM pollution by combining information on $PM_{2.5}$ and PM_{10} exceedance probabilities with information on population and housing characteristics. $PM_{2.5}$ and PM_{10} data for the years 2008 until 2014 was obtained from 37 air quality stations. Exceedance probabilities result from conditional simulations based on kriging with external drift models with joint spatiotemporal covariance functions which pool data across space and time. Covariates, namely land cover clusters and population counts represent location characteristics that either promote or inhibit the emissions and dispersion of particulate matter. The probability of exceeding PM_{10} and $PM_{2.5}$ regulatory thresholds is high (> 0.6) for most areas in the study region especially in central Gauteng. A composite spatial indicator is developed to quantify vulnerability to poor air quality relevant within the study area using geographically weighted PCA. The spatial patterns of social vulnerability differ from the pure spatial distribution of population counts in our study region. Small areas associated with high (> 0.7) social vulnerability are located south-east of Mpumalanga where lack of basic services and low education levels are indicated. PM_{10} exceedance probabilities and the size of the population are also high in this area, resulting in more small areas with higher (> 0.7) PM_{10} risk compared to other parts of Mpumalanga. Combining particulate matter exceedances and social vulnerability information enhances understanding of social conditions in areas where thresholds are regularly exceeded.

Keywords: Spatiotemporal kriging, conditional simulation, geographically weighted principal components analysis, particulate matter, social vulnerability, risk assessment

5.1 Introduction

Exposure to even moderate ambient concentrations of particulate matter (PM) pollution was associated with 2.9 million deaths and 69.7 million DALYs (disability adjusted life years) in 2013 (GBD 2013 Risk Factors Collaborators, 2015). Health risks associated with exposure to high concentrations of PM are exacerbated by poor living conditions including inadequate housing, living near sources, poverty and lack of immediate access to health care. The risk and vulnerability of communities that is associated with exposure to ambient PM pollution can be evaluated by integrating maps of areas with high long-term PM averages and/or areas with greater frequency of exceedance of limit values with indicators of socioeconomic vulnerability (Wright and Diab, 2011). This can inform decisions regarding improvement of air quality thereby reducing the incidence of air quality related morbidity and mortality.

Quantitative assessment of risk in socioecological settings, as is the case in associating particulate matter exposure to health, need to account for differing levels of susceptibility of exposed individuals (Schwartz et al., 2011). Data are obtained from various sources and often not for the purpose of such analyses. While incompleteness, errors in variables and preferential network design are problems common to air quality data (Chapter 4), socioeconomic data are aggregated into administrative units and are therefore misaligned with the point located air quality stations (Szpiro and Paciorek, 2013; Shaddick et al., 2016). In combining environmental and socioeconomic data spatial misalignment is an issue caused by the disparateness of data sources (Young and Gotway, 2007; Gryparis et al., 2009) and it needs to be considered when evaluating population risk to PM pollution.

Vulnerability is considered an important component in risk assessment. A social vulnerability index (SVI) based on generally accepted variables that represent drivers of social vulnerability in South African was created by le Roux et al. (2015). The purpose of the social vulnerability index was to assist South African government agencies to prioritise socially vulnerable communities in spatial integrated planning and in disaster management. The index was created statistically using principal components analysis (PCA) of the census 2011 data at ward level which is the minimum spatial unit for implementing policy plans. Three components were extracted and used to build the SVI. The first component was representative of the rural, female-headed poor households. The second component represented the shack dwelling urban communities, of low socioeconomic status, with some education and means of living due to access to formal and informal employment opportunities. The third component also represented rural households who lacked in basic services, had a higher proportion of immigrants and were not living below the poverty line. The vulnerability of shack dwellers was associated with settlements located in hazard prone areas and with limited access to basic social services. Wright and Diab (2011) developed an air pollution population exposure and vulnerability prioritization framework, where vulnerability considered demographic variables, employment status, education, income,

disease prevalence, access to health care, access to basic services (energy, water and waste removal) and life expectancy. They found that the variables they envisioned for their framework were difficult to obtain, especially at the required spatial resolution.

The objective of this chapter is to integrate spatial information on exceedances of PM_{10} and $PM_{2.5}$ annual thresholds with data on population and housing characteristics for inference about risks posed by exposure to excessive levels of particulate matter pollution. This requires development of a methodology consisting of spatiotemporally interpolating PM_{10} and $PM_{2.5}$ using suitable covariates given the sparsity of the air quality network and techniques for reducing the high dimensionality of data on population characteristics in a meaningful way for air quality risk assessment. For spatiotemporal interpolation, we consider kriging with external drift with a spatiotemporal metric covariance function. For information on population characteristics, we augment the SVI of le Roux et al. (2015) to reflect socioeconomic vulnerability related to exposure to poor air quality in urban areas and to account for spatial dependence by implementing a geographically weighted PCA.

The following sections include a description of the study area, air quality, land cover and population data in Section 5.2. This includes details on data preparation. In the methodology, Section 5.3, the spatiotemporal kriging and the geographically weighted PCA methods are presented. Results emanating from implementation of these methods follow in Section 5.4, followed by our reflections on this study in Section 5.5. Finally, we conclude in Section 5.6.

5.2 Data and pre-processing

5.2.1 Air quality data

Thirty seven air quality stations located in the Gauteng and Mpumalanga provinces (Figure 5.1) were considered. Within the Mpumalanga province, the stations are concentrated in the Gert Sibande and Nkangala district municipal areas, hence our study area does not extend to the Ehlanzeni district municipality. For each station daily observations of PM_{10}, $PM_{2.5}$, NO_2, SO_2, relative humidity, temperature, wind speed and wind direction were quality screened for treatment of erroneous values. Missing values in all variables were imputed using the bootstrap regression based multiple imputation method developed in Chapter 4. From the summary statistics of PM_{10} in Table 5.1, the overall average daily ambient PM_{10} concentration was 62 μg m^{-3}. If Olievenhoutbosch is considered an urban background station, the average for background stations was higher than the regional average at 65 μg m^{-3}, but this drops to 48 μg m^{-3} otherwise. Ambient PM_{10} concentrations were highest where domestic sources of pollution dominate, averaging 73 μg m^{-3} and lower (54 μg m^{-3}) for stations classified to be monitoring industrial air pollution sources.

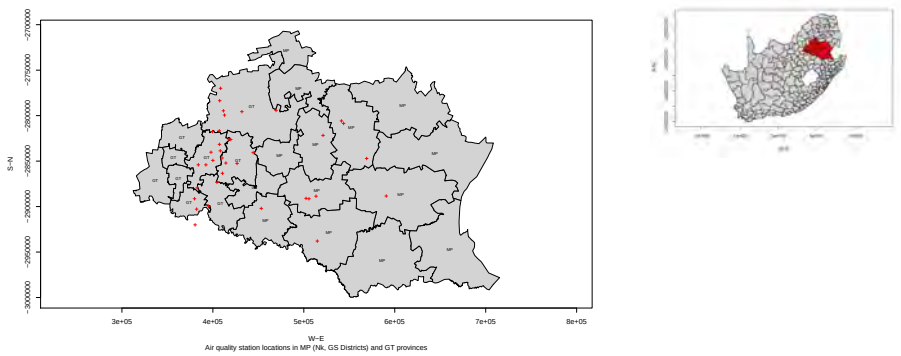


Figure 5.1: Study region which includes Gauteng province and the Nkangala and Gert Sibande districts in Mpumalanga

From Table 5.1, imputation quality is adversely affected by the proportion of valid data during each station's observed period. Missing values were imputed sequentially, starting with meteorological variables, ending with PM_{10} series and subsequently $PM_{2.5}$ for those stations where it was measured. Stations whose meteorological, gaseous pollutant and PM_{10} data were extensively missing had poorer imputation quality statistics for PM_{10} due to the accumulation of uncertainty caused by missing values. In cases where either relative humidity or temperature data were not collected (Germiston, Jabavu, Newtown, Orange farm, Club, Langverwacht and Pretoria West), dew-point temperature being a function of both variables could not be used as a regressor in pollutant models, hence either humidity or temperature was substituted for dew-point temperature. For stations where data for only one of NO_2 or SO_2 were collected, namely the Delta, Diepsloot, Ivory, Jabavu, Newtown and Orange farm stations, the PM_{10} model was adjusted to include only that available pollutant variable. The impact of these structural adaptations on imputation uncertainty was only negative for stations with higher proportions of missing data in all variables. Another adaptation to the imputation method concerned the 20 km maximum distance criterion for matching an air quality station with the nearest weather station (Chapter 4). For some stations this was relaxed to 40 km maximum because the nearest weather office was further. Only four stations, namely Balfour, Standerton, Club and Langverwacht, considered data from weather offices that were between 60 and 70 kilometers away. For Club and Langverwacht stations, this was only for wind speed imputation because their nearest weather station in Bethal ($\sim$ 35 km) had more than 50% of wind speed data missing for the period required for imputation. The impact of this on imputation quality depends upon data coverage, with poorer quality where data coverage is low.

We focus on mapping the annual average ambient PM_{10} and $PM_{2.5}$, consid-

Table 5.1: Summary statistics for PM_{10} after implementing the bootstrap regression multiple imputation method for missing values. Percentage of incompleteness for the $PM_{2.5}$ series is indicated only for those stations where it is measured.

Custodian	Station	Obs. period	Source[a]	% missing		PM_{10} pre-imputation[b]		PM_{10} post-imputation[c]				
				PM_{10}	$PM_{2.5}$	Mean	s^2	Mean	$\bar{V}_w$	V_b	$\widehat{V}$	λ
Highveld	Ermelo	Jan 2008:Feb 2015	Domestic	15	15	51.9	1703.2	53.7	1954.9	0.16	1955.1	8.3e-05
	Hendrina	Aug 2008:Feb 2015	Domestic	36	36	43.6	1259.8	41.8	1614.9	0.40	1615.3	2.6e-04
	Middelburg	Jan 2008:Feb 2015	Domestic	19	19	43.7	1265.0	43.6	1356.3	0.10	1356.4	7.8e-05
	Secunda	Jan 2008:Feb 2015	Industry	18	18	74.6	5008.6	76.9	5573.1	0.60	5573.7	1.1e-04
Mpumalanga	Club	Jan 2011:Aug 2012	Industry	13	49	29.3	391.3	31.4	1502.3	1.89	1504.2	1.3e-03
	Langverwacht	Jan 2011:Aug 2012	Industry	42	62	57.6	2158.6	60.3	7031.3	6.43	7037.9	9.4e-04
	Balfour	Nov 2008:Jul 2013	Domestic	51	41	65.5	7182.0	58.2	4422.8	0.39	4423.2	9.1e-05
	Middelburg	Nov 2008:Jul 2013	Domestic	44	47	50.9	1013.9	49.5	1008.6	0.16	1008.8	1.7e-04
	Standerton	Nov 2008:Jul 2013	Domestic	57	54	62.5	3560.7	59.9	2792.5	0.47	2793.0	1.8e-04
	Witbank	Nov 2008:Feb 2015	Domestic	48	48	51.6	1639.1	49.4	1843.9	0.16	1844.1	9.3e-05
Vaaltriangle	Diepkloof	Feb 2007:Feb 2015	Traffic	25	46	46.2	593.9	45.9	578.0	0.03	578.0	3.1e-04
	Kliprivier	Feb 2007:Feb 2015	Industry	24	32	63.7	1341.0	62.1	1236.0	0.10	1236.1	9.2e-05
	Sebokeng	Feb 2007:Feb 2015	Domestic	38	52	51.3	689.1	53.0	948.0	0.12	948.1	1.4e-04
	Sharpeville	Feb 2007:Feb 2015	Domestic	16	22	75.7	2016.9	73.5	1927.8	0.04	1927.8	3.4e-05
	Three Rivers	Feb 2007:Feb 2015	Industry	32	31	53.4	731.5	53.0	768.1	0.06	768.1	1.1e-04
	Zamdela	Feb 2007:Feb 2015	Industry	35	10	80.4	2548.6	76.0	2501.7	0.18	2501.9	7.4e-05
Johannesburg	Buccleuch	Sep 2006:Jun 2014	Traffic	43	56	58.5	791.5	56.2	749.4	0.14	749.6	1.9e-04
	Alexandra	Sep 2006:Oct 2010	Domestic	39		122.9	11021.9	124.0	12151.0	3.05	12154.2	2.6e-04
	Delta	Sep 2006:Jan 2012	Background	32		37.8	313.3	37.4	302.2	0.05	302.3	4.4e-04
	Diepsloot	Mar 2009:Aug 2011	Domestic	60		146.8	15981.8	146.4	18747.5	9.58	18757.4	5.3e-04
	Ivory	Mar 2009:Apr 2012	Domestic	57		119.8	3431.8	114.5	4000.3	0.91	4001.2	2.3e-04
	Jabavu	Sep 2006:Aug 2012	Domestic	27		60.3	1216.5	58.2	1347.8	0.16	1348.0	1.2e-04
	Newtown	Sep 2006:June 2014	Traffic	35		50.7	722.6	48.6	680.3	0.06	680.3	1.3e-04
	Orangefarm	Sep 2006:June 2014	Domestic	37		61.5	1267.0	63.0	1382.2	0.20	1382.4	1.5e-04
Tshwane	Bodibeng	Jul 2011:Jul 2014	Domestic	8		65.3	1270.8	65.9	1464.9	0.12	1465.1	8.6e-05
	Booysens	Jul 2009:Nov 2014	Background	35		56.7	1259.9	57.5	1397.8	0.26	1398.1	1.9e-04
	Ekandustria	Sep 2012:Nov 2014	Industry	24		42.6	563.9	40.8	581.1	0.18	581.3	3.3e-04
	Mamelodi	Jul 2009:Nov 2014	Domestic	71		83.2	1989.2	89.8	3893.3	1.08	3894.4	2.9e-04
	Olievenhoutbosch	Jul 2009:Nov 2014	Background	31		97.3	4540.9	98.5	5043.4	0.91	5044.4	1.9e-04
	Pretoria West	Feb 2009:Nov 2014	Industry	51		66.8	1626.9	60.1	1726.9	0.32	1727.2	1.9e-04
	Rosslyn	Jul 2009:Nov 2014	Industry	25		28.4	332.8	28.3	330.8	0.03	330.8	6.9e-04
Ekurhuleni	Tembisa	Jan 2011:Feb 2015	Domestic	59		82.6	1787.1	79.9	1868.7	0.40	1869.1	2.2e-04
	Bedfordview	Jan 2011:Feb 2015	Traffic	70		85.0	2196.8	79.1	2103.4	1.47	2104.9	7.5e-04
	Etwatwa	Jul 2011:Feb 2015	Domestic	60		87.2	2811.0	80.5	2723.7	1.00	2724.8	3.8e-04
	Germiston	Jan 2011:Feb 2015	Domestic	60		53.4	634.0	52.1	718.0	0.43	718.5	6.2e-04
	Thokoza	Jan 2011:Feb 2015	Domestic	72		96.7	4204.3	92.5	6402.4	4.26	6406.8	7.0e-04
	Wattville	Jan 2011:Feb 2015	Domestic	63		87.8	3155.0	82.7	3235.3	1.27	3236.6	4.1e-04

[a] Classification of stations by custodians in terms of dominant pollution sources, where 'background' refers to urban background stations
[b] Sample means (in μg m^{-3}) and variances (s^2) for the incomplete PM_{10} data
[c] Average within imputation variance ($\bar{V}_w$), between imputation variance (V_b), overall variance estimate $\widehat{V}$ and rate of information loss due to incompleteness λ

ering areas where the annual thresholds are exceeded. For brevity, predicted annual means are shown for the years 2009, 2011 and 2014. Twenty seven stations were operational in 2009, 36 in 2011 and 25 stations in 2014 were operational for 12 months while two operated for six months. Further the South African population census was done on 10 October 2011 and the national ambient air quality standards (NAQS) for $PM_{2.5}$ were officially proposed in August 2011. NAQS for the other priority pollutants were promulgated in December 2009. The year 2014 was the last year prior to the NAQS for PM_{10} changing to lower values in January 2015.

5.2.2 Land cover data

In Chapters 2 and 3 the prevalence of housing informality and land cover composition in a neighbourhood were significant predictors for PM_{10}. Previously, lack of an updated high spatial resolution land cover map led to land cover classification using SPOT 6 imagery in Chapter 3. Subsequently, the South African national land cover dataset 2013-2014 based on Landsat 8 at 30 m spatial resolution became available (GTI, 2015). These data are used in mapping PM_{10} and $PM_{2.5}$ in this chapter. The gridded land cover data with 72 classes was first converted into polygons, namely the census 2011 small area polygons to match the units used in constructing the social vulnerability indicator. Each datum in the gridded dataset is a land cover class assigned to that pixel. By converting to the small area polygons, land cover class proportions (area coverage) in each small area were calculated using the geometric intersection tool in a geographic information system. Further processing on the small area land cover proportions data included collapsing the variables corresponding to the 72 classes into 27 variables using the reduced verified land cover and use class hierarchy (GTI, 2015). The full 2013-2014 South African national land cover legend is in GTI (2015) as Appendix A and Table 5.2 is a list of the 27 variables. The vegetation descriptions appended to all the built-up classes accommodate the mixing of features within a pixel. There are four classes for housing informality in the land cover data. These include urban informal bare (built-up and no vegetation), urban informal with low vegetation or grass, urban informal with open trees or bush and urban informal with dense vegetation or bush. These data will account for housing informality in mapping particulate matter instead of data on percentages of households per small area classified as living in informal dwellings from the census used in Chapter 2. The urban residential classes refer to areas with formal housing.

The final step in processing the land cover data involves applying the k-means cluster algorithm to find homogenous groups of land cover features to which each small area will be assigned. These land cover feature clusters will be covariates in the spatial model to map PM_{10} and $PM_{2.5}$. Four clusters were identified and illustrated in Figure 5.2. Cluster 4 is associated strongly with urban informal residences, while Cluster 3 is associated with formal residential areas. The formal residential areas are also characterized by dense bushes and trees, whereas in informal and township areas grass and bare areas are more common. A cluster that is strongly associated with

Table 5.2: Twenty seven land cover variables defined as proportion of coverage in a small area corresponding to that particular land cover type

Variable[a]	Class[b]	Variable[a]	Class[b]
Water	seasonal, permanent water	Mine water	mine water seasonal, permanent
Grassland	grassland	Low shrubland	low shrubland
Thicket dense bush	thicket or dense bush	Wetlands	wetlands
Woodlands open bush	woodlands or open bush	Indigenous forest	indigenous forest
Cultivated fields	cultivated comm fields high, cultivated comm fields med, cultivated comm fields low	Cultivated pivots	cultivated comm pivots high, cultivated comm pivots med, cultivated comm pivots low
Cultivated orchards	cultivated orchards high, cultivated orchards med, cultivated orchards low	Cultivated subsistence	cultivated subsistence high, cultivated subsistence med, cultivated subsistence low
Plantation	plantation/woodlot mature, plantation/woodlot young, plantation/woodlot clear-felled	Urban sport golf	sport & golf (dense trees/bush), sport & golf (open trees/bush), sport & golf (low veg/grass), sport & golf bare
Bare none veg	bare none vegetated	Erosion donga	erosion (dongas & gullies)
Urban commercial	urban commercial	Urban industrial	urban industrial
Mine buildings	mine buildings	Mines bare	mines bare, mines semi-bare
Urban school sports ground	urban school & sports ground	Village	urban village (dense trees/bush), village (open trees/bush) village (low veg/grass), village bare
Smallhold	urban smallhold (dense trees/bush), smallholding (open trees/bush), smallhold (low veg/grass), smallhold bare	Residential	urban residential (dense trees/bush), residential (open trees/bush), residential (low veg/grass), residential bare
Township	urban township (dense trees/bush), township (open trees/bush), township (low veg/grass), township bare	Informal	urban informal (dense trees/bush) informal (open trees/bush), informal (low veg/grass), informal bare
Built other[c]	urban built-up (dense trees/bush), built-up (open trees/bush), built-up (low veg/grass), built-up bare		

[a] Variable names for land cover class coverage in each small area
[b] Classes aggregated to form that variable which are taken from the legend of the 72-class 2013-2014 South African national land cover dataset
[c] The urban built-up class in the SA-NLC data consists of areas that are not clearly identifiable as one of the other built-up classes including runways, major infrastructure development sites, roads, holiday chalets, car parks, cemeteries, etc (GTI, 2015).

townships is Cluster 2. Cluster 1 is village or traditional areas outside of the city regions which can be associated with open areas and grasslands, farms and the surrounds of commercial complexes, industrial areas and other non-residential built-up features such as roads, airports and cemeteries.

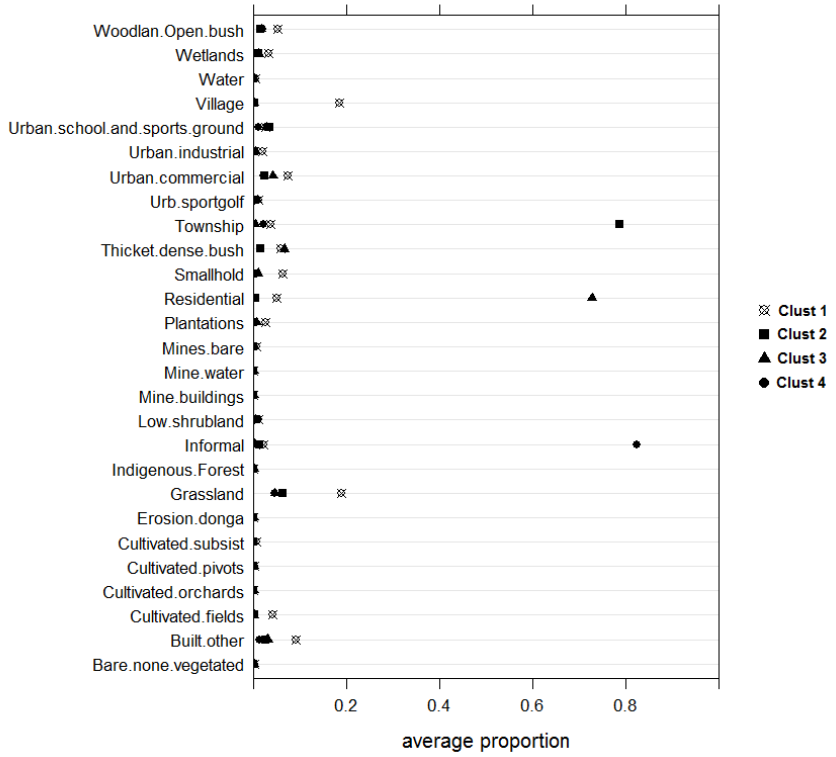


Figure 5.2: Profiles of the average proportion of land cover types characterizing the small areas that form each of the four clusters

5.2.3 Socioeconomic data

The availability of data at the correct spatial scale from credible sources was one of the criteria used for variable selection for the South African SVI (le Roux et al., 2015). This resulted in South African census 2011 being the data source for the index (Statistics South Africa, 2012b). The electoral wards are larger than the small areas (neighbourhood scale) which are minimum spatial units for disseminating official statistics. The SVI was revised by using small area census 2011 data and adding variables such as long-term residency (`resid_since2001`), thus resulting in 17 variables instead of 14 in the ward level index. The age dependent population was defined as those younger than 15 and older than 69. The percentage of child headed households

was included. The percentage of the population that is employed instead of unemployed was included because of stronger correlation with the other variables. The percentage of the population living below the poverty line was redefined using recent updates of the upper and lower bound poverty lines. The percentage of the population aged 25 with no education was changed to percentage of population aged 25 with low educational levels described by highest level of education that is lower than primary school. The percentage of the population that is disabled was restricted to those with difficulty or complete lack of capacity to self-care. The percentage of households using non-electric sources of energy for cooking excluded cleaner non-electric energy sources, namely solar and gas. The percentage households without telephone lines was swapped for percentage of households without sanitary toilet and waste removal services because of the importance of hygiene in health.

Table 5.3: Seventeen variables selected for inclusion in the social vulnerability index

Variable id	Variable name	Description
`agedepend`	Dependent	% dependent population i.e. persons younger than 15 or older than 69 years
`not_citizen`	Non-citizen	% of the population without South African citizenship
`resid_since2001`	Changed residence	% of the population that did not move between 2001 & 2011, and those born after 2001
`pct_25low_edu`	Uneducated adults	% of the population aged 25 years or older with primary school as highest level of education
`employ`	Employed	% of the population that is employed (includes formal, informal and domestic work)
`slfcare_impair`	Self-care disability	% of the population with difficulty with self-care from some difficulty to total incapacity
`avghhsize`	Household size	Average household[a] size for that small area
`noinc_grant_dep`	Living in poverty	% of households with annual household income[b] range of R 0 − R 38 200
`hh_female_pct`	Female household head	% of households headed by females
`hh_child_pct`	Child household head	% of households headed by children
`ruralpct`	Rural	% of rural[c] households
`informaldwell`	Informal dwelling	% of households living in informal dwellings[d]
`nocar`	No car	% of households without car ownership
`lowateraccess`	Piped water access	% of households without adequate[e] access to piped water
`energypov_cook`	Energy poverty	% of households using other energy[f] sources for cooking instead of electricity, gas or solar
`unsanitary_toilet`	Toilet system	% of households without access to sanitary toilets[g]
`unsanitary_waste`	Refuse removal	% of households without regular waste collection services

[a] Household according to Statistics South Africa is a group of persons who live together and provide themselves jointly with food and/or other essentials for living (on average four nights a week), or a single person who lives alone.

[b] Household income is defined as "all receipts by all members of a household, in cash and in kind, in exchange for employment, or in return for capital investment, or receipts obtained from other sources such as social grants, pension, etc" (Statistics South Africa, 2012c).

[c] These are farms and communally owned land under jurisdiction of traditional leader, characterized by low levels of population density, economic activities and infrastructure (Statistics South Africa, 2012c).

[d] A dwelling unit is a structure mean't to be occupied by at least one household and informal dwellings are shacks (makeshift structures) including those in backyards in formal residences that have not been approved by local authority and not intended for permanent residence.

[e] Inadequate access to piped water includes household without any access and those that have to walk further than 1 km for water.

[f] Use of wood, coal, dung, paraffin and other unspecified energy sources for cooking.

[g] Households without sanitary toilets may either have no toilet or use the bucket system or pit toilets without ventilation

The age dependency lower boundary is based on Section 43 of the Basic Conditions of Employment Act which specifies the employment of children younger than 15 years as criminal with the exception of children in the arts and that children between 15 and 18 may be employed provided that work activities are appropriate (Department of Labour, 2016). There is no specific threshold for retirement age, but 65 is a common retirement age specified contractually by employers. People may continue to work privately beyond 65 and therefore 69 years was chosen. Child headed households are defined as those headed by children younger than 18 years in age. For the social vulnerability index by le Roux et al. (2015), poverty was described as percentage of the population earning less than R 400 a month based on

the food poverty line of R 321 per capita, per month. The food poverty line (FPL) captures the extremely poor households that cannot afford enough food to satisfy their daily caloric needs, but it misses those that are poor in terms of not having adequate resources to consume adequate food and non-food items without having to sacrifice non-food items for food. Those who can afford both adequate food and non-food items are at or above the upper bound poverty line (UBPL) (Statistics South Africa, 2015). Rebasing of poverty lines based on the South African income and expenditure survey of 2010 and 2011 led to the FPL being adjusted from R 321 to R 335 per capita per month and the UBPL adjusted to R 779 per capita per month. The revision of the poverty indicator consisted of considering an annual aggregation of per capita monthly incomes equal to the UBPL for the average South African household size of four (Statistics South Africa, 2012c). This resulted in focussing on the percentage of households per small area with annual household incomes between zero and R 38 200 which is the closest income bracket in the data to the annual per capita UBPL of R 37 392.

5.3 Methods

Population health risk to air pollution is assumed to be a function of vulnerability and exposure to health hazards, namely, $PM_{2.5}$ and PM_{10} in excess of regulatory limits. The spatial scale of assessment is the neighbourhood (or small area) level which is appropriate for public health interventions. Previous chapters discussed innovations to improve maps of particulate matter derived from data of limited spatial and temporal (missing daily data) coverage. To improve temporal coverage at each station a bootstrap regression imputation method was implemented leading to the completed $PM_{2.5}$ and PM_{10} daily time series that will be used in the chapter (Figure 5.3). To improve spatial coverage land cover data and an indicator of housing informality were identified as suitable covariates. These will be used in a kriging with external drift (universal kriging) model to map PM_{10} and $PM_{2.5}$ respectively. Existing thresholds that will be considered are the national air quality standards for $PM_{2.5}$ and PM_{10}. The pollutant maps show the hazard component of risk.

5.3.1 Space-time kriging implemented to map $\mathbf{PM}_{10}$ and $\mathbf{PM}_{2.5}$

Let $\{\mathbf{Y}(\mathbf{s},t) : \mathbf{s} \in D_s, t \in D_t\}$ denote a spatiotemporal random field for atmospheric concentrations of either $\log(PM_{10})$, or $\log(PM_{2.5})$. We have observations $\{\mathbf{z}(\mathbf{s},t) : \mathbf{s} = (s_1, s_2, \ldots, s_d) \times t = (t_1, t_2, \ldots, t_T) \in D_s, t \times D_t \subseteq \mathbb{R}^2 \times \mathbb{R}\}$ at air quality station locations (Gräler et al., 2016). The aim is to model $\mathbf{Y}$ using observations $\mathbf{z}$ and predict particulate matter levels for all small areas in Gauteng and Mpumalanga provinces. Predictions of PM_{10} and $PM_{2.5}$ at the small area level (centroids) are aimed at alignment with the social vulnerability and population exposed indicators derived from the census 2011 small area data. Land cover clusters and population counts which are considered as q spatial covariates $\{\mathbf{x}(\mathbf{s}) = (x_1(\mathbf{s}), \ldots, x_q(\mathbf{s}))'\}$ in

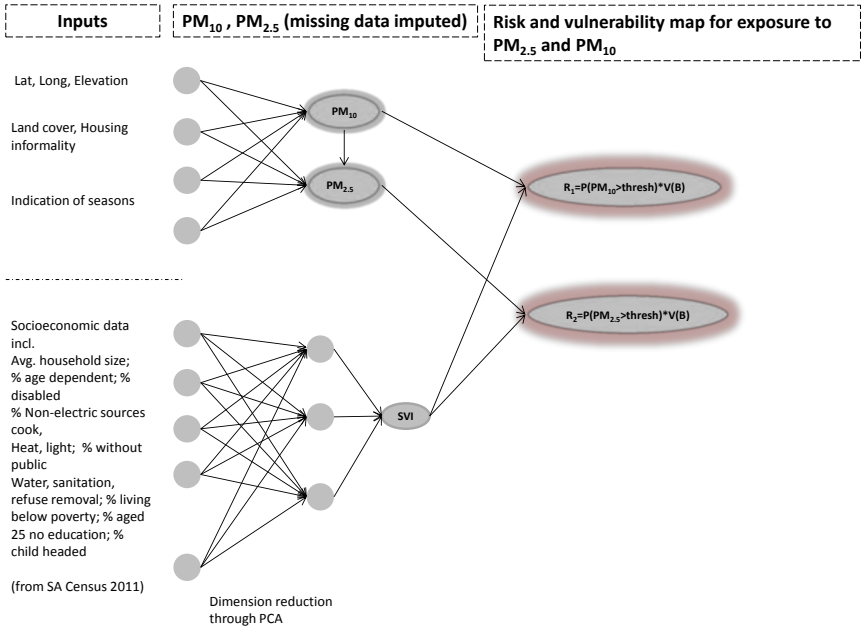


Figure 5.3: Conceptual framework of the methodology to assess population risk of exposure to excessive concentrations of ambient $PM_{2.5}$ and PM_{10}

modelling $\mathbf{Y}$ are also in small area polygons. The years are denoted as t, matching the exceedance probability standards for the annual average concentrations of PM_{10} and $PM_{2.5}$. The data model and the true underlying spatiotemporal field model according to Cressie and Wikle (2011) can be expressed as,

$$\begin{aligned} \mathbf{Z}(\mathbf{s},t) &= \mathbf{Y}(\mathbf{s},t) + \epsilon(\mathbf{s},t), \quad \text{where} \qquad \epsilon(\mathbf{s},t) \sim \mathrm{N}\left(\mathbf{0}, \sigma^2_\epsilon \mathbf{I}_d\right) \qquad (5.1) \\ &= \mathbf{x}(\mathbf{s})'\beta + \mathbf{w}(\mathbf{s},t) + \epsilon(\mathbf{s},t) \qquad (5.2) \end{aligned}$$

where $\mathbf{w}(\mathbf{s},t)$ is a zero-mean stationary Gaussian process with covariance function $C_{st}(\mathbf{h},\tau) = \mathrm{Cov}\left(\mathbf{Z}(\mathbf{s},t), \mathbf{Z}(\tilde{\mathbf{s}},\tilde{t})\right)$ dependent on separation distances $\mathbf{h} = \mid \mathbf{s} - \tilde{\mathbf{s}} \mid$ between locations and $\tau = \mid t - \tilde{t} \mid$ between years for any pair of spatiotemporal points $(\mathbf{s},t), (\tilde{\mathbf{s}},\tilde{t}) \in D_s, t \times D_t$. Covariance functions will not be estimated directly, but the variogram $\gamma_{st}(\mathbf{h},\tau) = C_{s,t}(\mathbf{0},0) - C_{s,t}(\mathbf{h},\tau)$ will be estimated.

The annual PM_{10} and $PM_{2.5}$ data are neither spatially nor temporally extensive at (37×7) and (17×7) resolutions, respectively. Pooling data across space and time with an adjustment for temporal dependence is one way to increase the reliability of variogram estimation when the sample size in space and time is small (Sterk and Stein, 1997). The spatiotemporal metric covariance model is based on the same principle of pooling by jointly modelling covariance in space and time, treating spatial, temporal and spatiotemporal distances

equally after aligning space and time by an anisotropy correction parameter κ. The anisotropy correction parameter scales time to an equivalent spatial distance, such that κ is given as units of $\mathbf{h}$ per unit of τ (Gräler et al., 2016). Given a purely spatial covariance function $C_s(\cdot)$ and this anisotropy parameter,

$$\begin{aligned} C_{st}(\mathbf{h}, \tau) &\equiv C_s(\mathbf{h} - \kappa\tau) \\ &\equiv C\left(\sqrt{\mathbf{h}^2 + (\kappa \cdot \tau)^2}\right) \end{aligned}$$

is a non-separable, geometrically anisotropic covariance function in metric $\mathbb{R}^3$ space. This corresponds to a spatial field which temporally evolves with constant velocity, also known as Taylor's frozen turbulence hypothesis (Cressie and Wikle, 2011). Higgins et al. (2012) used a field experiment on humidity to study the effectiveness of this hypothesis and concluded that the hypothesis is best applicable to atmospheric fields exhibiting long range dependence (temporal persistence). Therefore, we assume the metric covariance function is appropriate for our annual data. We will estimate the nugget, range, partial sill and spatiotemporal anisotropy parameters for the variogram model

$$\gamma_{st}(\mathbf{h}, \tau) = \gamma\left(\sqrt{\mathbf{h}^2 + (\kappa \cdot \tau)^2}\right)$$

Our goal is to predict the expected value of our hidden spatiotemporal field at $(\mathbf{s}_0, t_0)$ given the observed data $\mathbf{z}$ through external drift kriging:

$$\begin{aligned} E(Y(\mathbf{s}_0, t_0) \mid \mathbf{Z}) &= \mu(\mathbf{s}_0, t_0) + \mathbf{c}_0' \mathbf{C}_\mathbf{Z}^{-1}(\mathbf{Z} - \mu) \\ &= \mathbf{x}(\mathbf{s}_0)\hat{\beta} + \mathbf{c}_0' \mathbf{C}_\mathbf{Z}^{-1}(\mathbf{Z} - \mathbf{X}\hat{\beta}) \end{aligned} \tag{5.3}$$

with the generalized least squares estimates of the large scale spatial trend coefficients $\hat{\beta} = \left(\mathbf{X}'\mathbf{C}_\mathbf{Z}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{C}_\mathbf{Z}^{-1}\mathbf{Z}(\mathbf{s})$. The second term of the prediction equation (Eq. 5.3) consists of the simple kriging weights $\mathbf{c}_0'\mathbf{C}_\mathbf{Z}^{-1}$, weighting the average of the residuals that contribute to the mean function. The minimized mean squared prediction error, namely the external drift kriging variance is

$$\sigma^2(\mathbf{s}_0, t_0) = \mathbf{C}(\mathbf{0}, 0) - \mathbf{c}_0'\mathbf{C}_\mathbf{Z}^{-1}\mathbf{c}_0 + (\mathbf{x}(\mathbf{s}_0) - \mathbf{c}_0'\mathbf{C}_\mathbf{Z}^{-1}\mathbf{X})(\mathbf{X}'\mathbf{C}_\mathbf{Z}^{-1}\mathbf{X})^{-1}(\mathbf{x}(\mathbf{s}_0) - \mathbf{c}_0'\mathbf{C}_\mathbf{Z}^{-1}\mathbf{X})' \tag{5.4}$$

The last term in Equation 5.4 is the contribution of the trend estimation error to the prediction error. The accuracy of prediction will be evaluated using 10-fold cross-validation (Bivand et al., 2008).

For mapping population risk of exposure to particulate matter pollution, we target the probability of exceedances of the national standards for $PM_{2.5}$ and PM_{10} as hazard quantities. The national standard for the annual average $PM_{2.5}$ concentrations is 25 μg m^{-3} and 50 μg m^{-3} for annual average PM_{10} (RSA Government, 2009; Department of Environmental Affairs and Tourism, 2012). According to the World Health Organization (WHO), air quality guidelines (AQG) of 10 μg m^{-3} and 20 μg m^{-3} for annual average

$PM_{2.5}$ and PM_{10} are the lowest levels at which total, cardiopulmonary and lung cancer mortality have been shown to increase with long-term exposure (WHO, 2006; Krzyzanowski and Cohen, 2008). Our national standards are equivalent to the Interim Target-2 thresholds of the WHO associated with 2-11% higher long-term mortality risk relative to the AQG level (Krzyzanowski and Cohen, 2008).

We denote the probability of $\log(PM_{2.5})$ or $\log(PM_{10})$ exceeding the prescribed threshold y_c at location $\mathbf{s}$ as

$$\begin{aligned} \mathrm{P}(\mathbf{Y}(\mathbf{s}) > y_c \mid \mathbf{z}) &= 1 - \mathrm{P}(\mathbf{Y}(\mathbf{s}) \leq y_c \mid \mathbf{z}, \mu, \mathbf{C_Z}) \\ &= 1 - \mathrm{F}(\mathbf{s}, y_c \mid \mathbf{z}, \mu, \mathbf{C_Z}) \end{aligned} \tag{5.5}$$

where $\mathrm{F}(\cdot \mid \cdot)$ is the conditional cumulative probability which can be defined in terms of the conditional expectation of the exceedance indicator $I(\mathbf{s}, y_c \mid \mathbf{z})$, namely

$$\mathrm{F}(\mathbf{s}, y_c \mid \mathbf{z}, \mu, \mathbf{C_Z}) = E\{I(\mathbf{s}, y_c) \mid \mathbf{z}, \mu, \mathbf{C_Z}\}$$

The exceedance indicator equals one for values below the threshold and zero otherwise (Goovaerts et al., 1997). Once the parameters of the log-transformed $PM_{2.5}$ and PM_{10} pollutant surfaces have been estimated, other possible realizations of random fields can be simulated conditional on the same mean, covariance structure and observed values. From the simulated surfaces of log-transformed $PM_{2.5}$ and PM_{10}, $\mathrm{F}(\cdot \mid \cdot)$ is empirically estimated as counts of non-exceedances of $\log(25)$ and $\log(50)$ respectively. Subtracting $\hat{\mathrm{F}}(\cdot \mid \cdot)$ from one yields an estimate of the probability of exceedance.

5.3.2 Social vulnerability indicator development

"Social vulnerability is the exposure of groups or individuals to stress as a result of social and environmental change" (Zhou et al., 2014). Our interest is differential risk related to ambient levels of particulate matter and heterogeneity in social vulnerability of the exposed population. Complexity in quantifying social vulnerability can be attributable to the multitude of factors such as social inequality, social capital differences, susceptibility to disease, access to basic and social services or quality of the built environment. Some factors can be quantified through routinely collected data, such as demographic data from census, while others are constructs which are difficult to measure. There is criticism against the ad-hoc manner in which social vulnerability indicators are developed, the difficulty of evaluating their accuracy and concerns regarding weighting and aggregation in indicator development in general (Paruolo et al., 2013). Despite these methodological challenges, social vulnerability indicators do provide a mechanism for determining locations of vulnerable communities and identifying a combination of factors that render those communities more sensitive to stresses within their environment. Factor analysis based on principal components extracted from the data is used here to quantify social vulnerability.

In PCA interrelationships among a large number of variables are used to extract common factors or components which explain the underlying structure in a dataset. The components explain the data in fewer dimensions than the original variables. Extraction of components that are meaningful for creating a social vulnerability indicator, require care in obtaining reliable, unique and relevant data as discussed previously. We denote the J variables derived from the census as $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_J\}$ where each is of length n corresponding to the number of census small areas considered. As preparation for PCA, variables are standardized using the means $\{\mu_1, \mu_2, \ldots, \mu_J\}$ and standard deviations $\{\sigma_1, \sigma_2, \ldots, \sigma_J\}$. For the standardized dataset $\mathbf{X}_{n\times J}$ with correlation matrix $\mathbf{\Sigma}$, positive latent roots $\{\lambda_1, \lambda_2, \ldots, \lambda_J\}$ and an orthogonal matrix $\mathbf{H}_{J\times J}$ can be found through a singular value decomposition such that each column $\mathbf{h}_j$ of $\mathbf{H}$ is an eigenvector with corresponding eigenvalue λ_j. The j^{th} principal component of $\mathbf{X}$ is defined as $\mathbf{U}_j = \mathbf{h}_j^{'}\mathbf{X}$, the normalized linear combination of the components $\mathbf{X}$ which has maximum variance λ_j out of all normalized linear combinations which are uncorrelated with $\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_{j-1}$ with $j = 1, 2, \ldots, J$. The number of components chosen is normally less than the number of variables in $\mathbf{X}$. For standardized data, the Guttman-Kaiser criterion can be used to provide a lower bound on the number of components (Guttman, 1954; Kaiser, 1961). This criterion specifies the lower bound on the number of components as the number of eigenvalues that are no less than unity (Yoemans and Golder, 1982). With the Guttman-Kaiser criterion as a first consideration, the cumulative percentage variability explained was also considered, with 75% chosen as suitable threshold. The stability of the number of components chosen was tested using leave-one-out cross validation. Alternatively bootstrapping methods could be used.

The PCA as described above is global in that the covariance structure as represented by the eigen-decomposition model is assumed constant over the study area. Therefore we further develop the social vulnerability index by le Roux et al. (2015) by considering the spatial heterogeneity of the covariance structure (Harris et al., 2011). Theoretically, this means for an observed variable $\mathbf{x}_{ij}$ at spatial location s_i with geographic coordinates (u, v), we can assume that for a geographically weighted mean vector and variance-covariance matrix, the $\{\mathbf{x}_{ij} : i = 1, 2, \ldots, n; j = 1, 2, \ldots, J\}$ are Gaussian, namely $\mathbf{x}_{ij} \mid (u, v) \sim \mathrm{N}(\mu(u, v), \mathbf{\Sigma}(u, v))$. The geographically weighted eigenvalues and eigenvectors are obtained by decomposing $\mathbf{\Sigma}(u, v)$ such that,

$$\mathbf{\Sigma}(u, v) = \mathbf{X}^T\mathbf{W}(u, v)\mathbf{X} \tag{5.6}$$

$$\mathbf{\Sigma}(u_i, v_i) = \mathbf{H}\mathbf{V}\mathbf{H}^T \mid (u_i, v_i), \quad \text{for } i = 1, 2, \ldots, n \tag{5.7}$$

where $\mathbf{\Sigma}(u_i, v_i)$ is the location specific geographically weighted variance-covariance matrix and $\mathbf{V}$ is the diagonal matrix of eigenvalues specific to the location. The matrix of geographic weights $\mathbf{W}(u, v)$ is generated using a kernel function with either a user specified or an estimated bandwidth (geographic distance r). There are various kernel functions including Gaussian and exponential functions, but we chose the bi-square function which for a

pair of small areas that are distance d_{ik} apart, the weights are

$$w_{ik} = (1 - (d_{ik}/r)^2)^2 \text{ if } d_{ik} \leq r; \quad w_{ik} = 0 \text{ otherwise} \tag{5.8}$$

Our choice of bandwidth, namely 30 km, is within the urban scale of $5-50$ km applicable in urban social studies and $J = 17$, the variables derived from the census data and described in Table 5.3.

5.3.3 Quantifying risk by aggregating exceedance probability, population exposure and socioeconomic vulnerability

To make statements about population health risk associated with ambient particulate matter pollution we consider estimating the probability that $PM_{2.5}$ and PM_{10} exceed the regulatory levels associated with increased long-term mortality risk. Another important component in population health risk is a measure of population sensitivity to the effects of exposure to poor air quality and their capacity to cope with adverse outcomes. If $I_{\text{Pop}}(\mathbf{s})$ is the population exposed indicator at location $\mathbf{s}$, the social vulnerability indicator $I_{SVI}(\mathbf{s})$ captures the different levels of susceptibility of the exposed population at location $\mathbf{s}$. Therefore, we define a risk index by the geometric mean of pollutant exceedance probability, the population exposed and the vulnerability of the population exposed (JRC European Commission, 2008):

$$\text{Risk}(\mathbf{s}) = (\text{P}(\mathbf{Y}(\mathbf{s}) > y_c \mid \mathbf{z}) \times I_{\text{Pop}}(\mathbf{s}) \times I_{\text{SVI}}(\mathbf{s}))^{1/3} \tag{5.9}$$

The indicator of population exposed is derived using the min-max normalization on the log-transformed population counts, where the transformation is necessitated by the positive skewness of the population count distribution. That is, for a particular location $\mathbf{s}_0$ and population count variable Pop,

$$I_{\text{Pop}}(\mathbf{s}_0) = \frac{\log(\text{Pop}(\mathbf{s}_0)) - \log(\min(\text{Pop}))}{\log(\max(\text{Pop})) - \log(\min(\text{Pop}))}$$

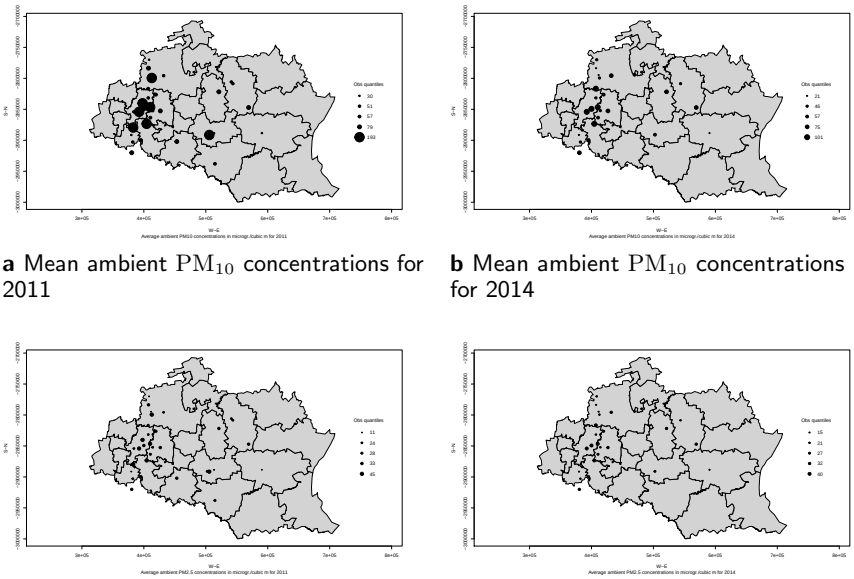
has a range $[0, 1]$, indicating the size of the population exposed relative to the highest (11 720) and lowest (10) population counts in small areas forming the study region. Pollutant hazard, exposure and vulnerability contribute equally to risk defined in Equation 5.9. The Risk($\mathbf{s}$) index ranges from zero to one.

5.4 Results

5.4.1 Spatiotemporal kriging prediction of $\mathbf{PM}_{2.5}$ and $\mathbf{PM}_{10}$ at locations without stations

The basis of our kriging models are the annual averages from 2008 to 2014 of PM_{10} and $PM_{2.5}$ calculated from daily observations after imputation of missing values at each air quality station. PM_{10} data were available for 35 stations in 2011 compared to only 27 stations in 2014 as shown by the reduced number of points in Figure 5.4b compared to Figure 5.4a. The

maximum annual means were higher in 2011 compared to 2014 for both types of particulate matter (Figure 5.4c and 5.4d). Observation quantiles refer to the minimum and the four quartiles of the observed frequency distributions. The overall medians and third quartiles are similar for the two years but we observed that three stations in the Johannesburg metropolitan municipality, namely Diepkloof, Newtown and Bedfordview consistently have higher annual average PM_{10} and $PM_{2.5}$ values relative to other stations nearby.



a Mean ambient PM_{10} concentrations for 2011

b Mean ambient PM_{10} concentrations for 2014

c Mean ambient $PM_{2.5}$ concentrations for 2011

d Mean ambient $PM_{2.5}$ concentrations for 2014

Figure 5.4: Annual average concentrations for PM_{10} and $PM_{2.5}$ at air quality station locations

The first step toward spatiotemporal prediction through kriging is variogram modelling. A metric variogram was chosen for our data. In exploratory linear regression modelling, a varying intercepts model was fitted, where the varying intercepts were the four land cover composition clusters introduced in Section 5.2.2 and the slope variable was population counts. This model was significant when $\log(PM_{10})$ and $\log(PM_{2.5})$ were individually considered as dependent variable and the adjusted coefficients of determination (adj-R^2) in both instances were larger than 0.95. This was the justification for considering kriging with external drift and Table 5.4 gives the generalized least squares estimates of the large scale spatial trend coefficients obtained during variogram modelling. An interesting finding was that a difference of 1 000 in population counts corresponds to an expected increase of 17% in annual average PM_{10} and a 77% decrease in annual average $PM_{2.5}$.

Table 5.4: Spatial trend coefficients capturing the relation between land cover composition clusters, population counts and ambient concentration of PM_{10} and $PM_{2.5}$

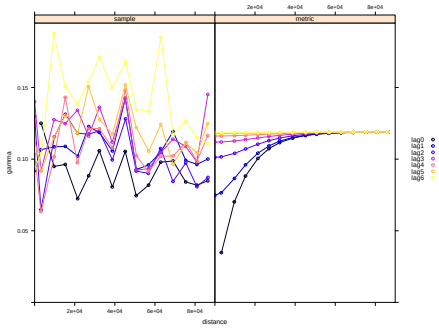
	$\log(PM_{10})$		$\log(PM_{2.5})$	
Coefficients	**Estimate**[a]	**Std. Error**	**Estimate**	**Std. Error**
$\hat{\beta_1}$: Cluster 1	4.10	0.11	3.66	0.08
$\hat{\beta_2}$: Cluster 2	3.87	0.12	3.74	0.17
$\hat{\beta_3}$: Cluster 3	3.91	0.12	3.80	0.15
$\hat{\beta_4}$: Cluster 4	3.66	0.12	3.99	0.14
$\hat{\beta_5}$: Pop[b]	0.17	0.10	-0.77	0.21

[a] Generalized least squares estimates of trend coefficients β
[b] Scaled population count variable $I(Pop/1000)$, namely population in thousands

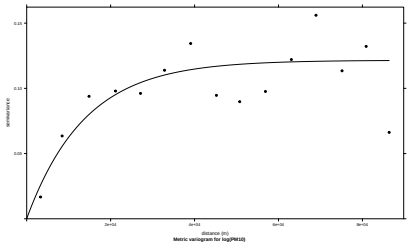
Once large scale spatial trends in $\log(PM_{10})$ and $\log(PM_{2.5})$ had been accounted for, parameters were estimated for the Exponential form of the metric spatiotemporal variogram model for residual spatial variation. Spatial dependence in observations further than two years apart diminishes towards zero as shown in Figure 5.5a. An estimate of the spatiotemporal anisotropy parameter κ was $2.8\,\mathrm{km\,yr^{-1}}$ for $\log(PM_{10})$, therefore a close value of $3\,\mathrm{km\,yr^{-1}}$ was used for both $\log(PM_{10})$ and $\log(PM_{2.5})$. The estimated range for $\log(PM_{10})$ is 13.9 km which is twice the estimated range for $\log(PM_{2.5})$. The means of the weighted squared deviations (wMSE) between the samples and fitted variogram surfaces are close to zero, indicating good fits (Figure 5.5b and 5.5c). Prediction accuracy was also evaluated using 10-fold cross validation, with summary statistics of bias and precision, namely mean error (cvME) and unbiased root mean square error (cvURMSE) for kriging PM_{10} and $PM_{2.5}$ listed in the captions Figures 5.5b and 5.5c. Bias of $2.9e-04$ for PM_{10} and $9.7e-04$ for $PM_{2.5}$ is close to zero which is ideal.

Figures 5.6 and 5.7 are the back-transformed predicted annual mean concentrations and kriging standard deviations for PM_{10} and PM_{10} at point locations corresponding to the small area centroids. The maximum predicted value for PM_{10} in Figure 5.6 is $745\,\mu\mathrm{g\,m^{-3}}$ which is more than five times larger than the 0.99 quantile, namely $134\,\mu\mathrm{g\,m^{-3}}$. These very high annual average PM_{10} predictions are in central Gauteng province, in the south east direction where Jabavu, Diepkloof and Thokoza stations lie. This extreme positive skewness is not observed in predicted values for $PM_{2.5}$ in Figure 5.7. Predicted PM_{10} annual means for the three years are smaller than $75\,\mu\mathrm{g\,m^{-3}}$ for most areas in Mpumalanga. Prediction standard deviations are higher for locations further away from station locations.

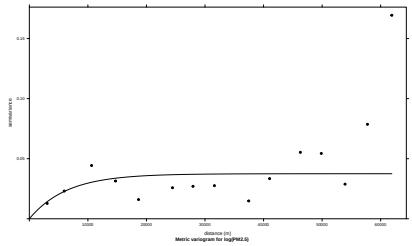
Differences in predictions between the years 2009, 2011 and 2014 in Figure 5.6 and Figure 5.7 are difficult to see, hence differences in the predictions of 2014



a Sample variogram and the fit of the exponential joint/metric variogram for the seven annual lags



b Metric variogram fit for $\log(PM_{10})$: nug=1e-04, psill=0.12, range=13.9 km, wMSE=4e-09, cvME=2.9e-04, cvURMSE=5.5e-02



c Metric variogram fit for $\log(PM_{2.5})$: nug=1e-03, psill=0.04, range=6.3 km, wMSE=4.6e-10, cvME=9.7e-04, cvURMSE=4.4e-02

Figure 5.5: The spatiotemporal metric variograms for the log-transformed annual average $\log(PM_{10})$ and $\log(PM_{2.5})$

and 2011 as well as 2009 and 2011 are mapped in Figure 5.8. These show that for PM_{10} predicted annual averages are higher for the western parts of Gauteng and for most of the two districts of Mpumalanga for 2009 compared to 2011, but when 2014 predictions are compared to those of 2011 we notice that they are lower for these same areas. Similarly for $PM_{2.5}$, predicted means for 2009 are higher than those of 2014 when both are compared to 2011.

5.4.2 A spatially explicit social vulnerability indicator

The initial step in adapting the social vulnerability index to account for spatial heterogeneity in correlation was to derive the global principal components. With our 17 standardized variables as inputs, keeping four to six components seems to be appropriate according to the scree plot (Figure 5.9). The percentage of variation explained in Table 5.5 by a component starts falling below 5% after the sixth component. The cumulative percentage of

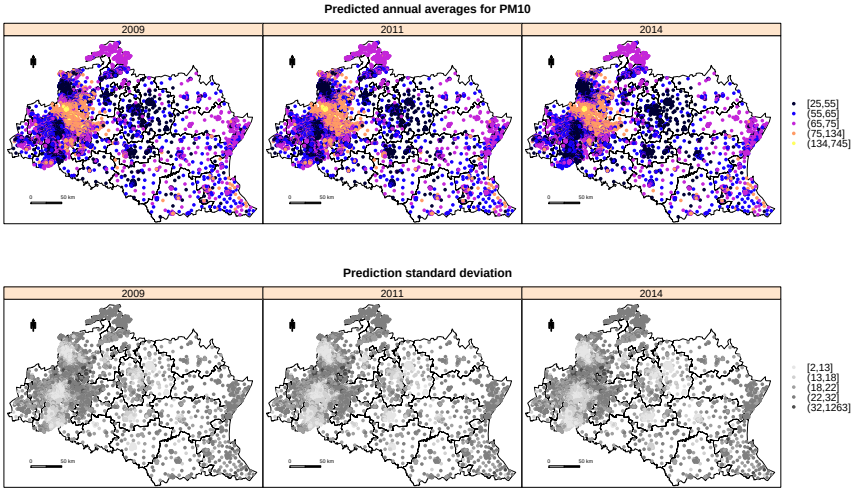


Figure 5.6: Kriging maps for PM_{10} for the years 2009, 2011 and 2014, where the point locations are the small area centroids

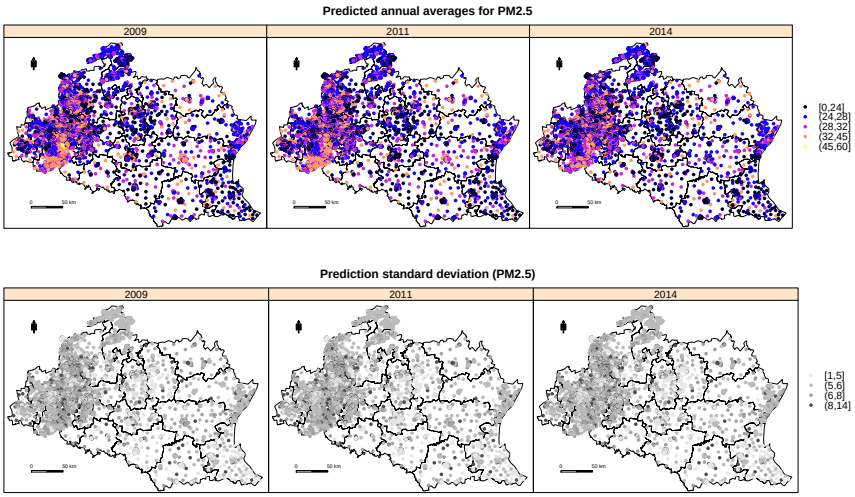


Figure 5.7: Kriging maps for $PM_{2.5}$ for the years 2009, 2011 and 2014

variation accounted for by the first six components is 72%, which grows to 80% when the eighth component is considered. High cumulative variance percentage, that is above 90%, is only achieved when the first 11 components are considered. According to the Guttman-Kaizer criterion, the minimum number of components is three. With only the first six component loadings shown in Table 5.5, the first component loading seems representative of small areas with urban male headed households that are affluent, mobile and have

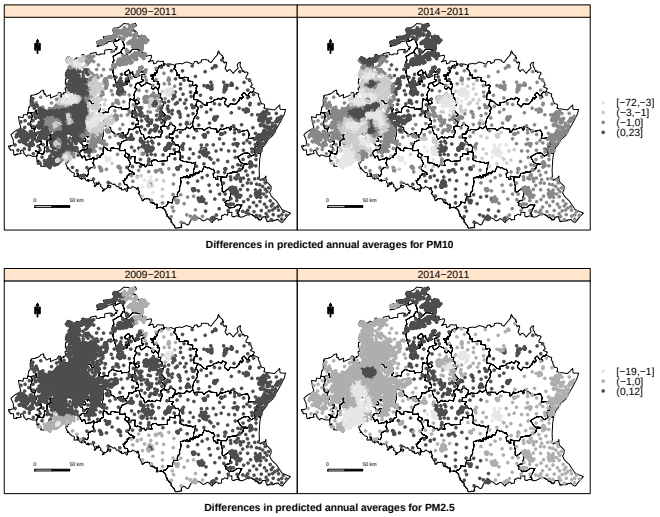


Figure 5.8: Differences predict means in PM_{10} and $PM_{2.5}$ for 2009 and 2014 compared to 2011

Table 5.5: Summary of global PCA results

	PC1	PC2	PC3	PC4	PC5	PC6
Eigenvalues	2.23	1.67	1.20	1.04	0.94	0.88
Proportion of variance	0.31	0.16	0.09	0.06	0.05	0.05
Component loadings:						
Dependent	-0.26	0.29	-0.08	0.02	-0.06	0.17
Non-citizen	0.09	-0.34	0.09	-0.47	0.08	-0.30
Changed residence	-0.26	0.25	0.15	0.29	0.12	-0.18
Uneducated adults	-0.37	-0.05	-0.01	-0.07	0.01	-0.26
Employed	0.34	-0.19	-0.19	-0.05	-0.07	-0.03
Self-care disability	-0.17	0.07	-0.03	-0.03	-0.95	0.12
Household size	-0.17	0.39	-0.10	0.21	0.06	-0.13
Living in poverty	-0.37	-0.13	0.29	-0.08	0.04	-0.13
Female household head	-0.12	0.29	0.08	-0.40	0.10	0.50
Child household head	-0.17	0.06	-0.10	-0.56	0.18	0.26
Rural	-0.22	0.08	-0.46	-0.22	0.02	-0.42
Informal dwelling	-0.19	-0.39	0.22	0.18	-0.01	0.22
No car	-0.34	-0.14	0.36	-0.10	0.02	-0.16
Piped water access	-0.19	-0.12	-0.48	-0.03	-0.03	-0.10
Energy poverty	-0.26	-0.35	-0.10	0.09	0.01	0.17
Toilets	-0.16	-0.30	-0.22	0.22	0.07	0.28
Refuse removal	-0.17	-0.17	-0.36	0.13	0.14	0.20

lower levels of dependency in the form of children or the elderly and the functionally disabled individuals. Component three seems representative of urban poor households, living in informal dwellings with access to basic services and higher levels of unemployment. Component four is mostly female headed households without citizenship in traditional and farm areas, living below the upper bound poverty line, limited possession of private vehicles but with access to basic services such as cleaner forms of energy for cooking and solid waste removal mechanisms.

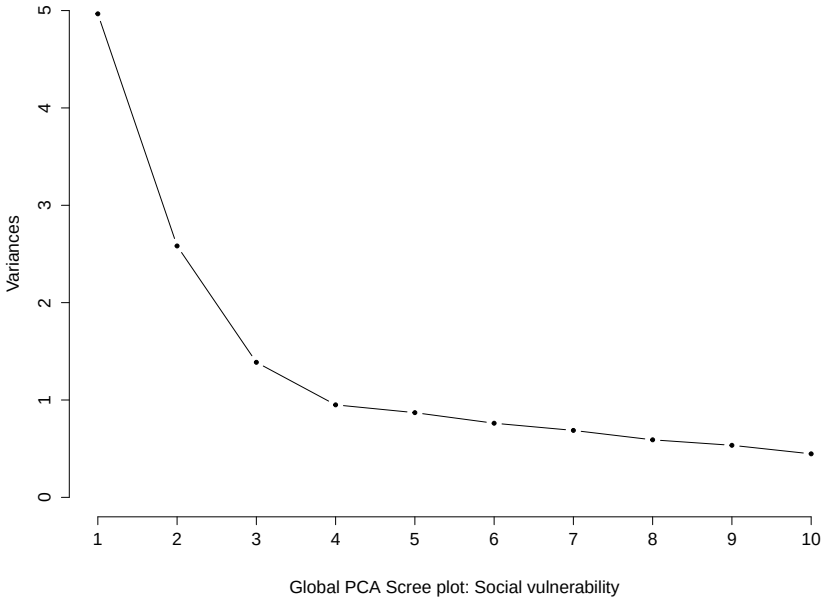


Figure 5.9: Scree plot for determining the number of global principal components for the vulnerability index

Geographically weighted PCA is implemented because there is variability across small areas of the cumulative proportion of variance explained for the six components (Table 5.6). When spatial heterogeneity is accounted for through localized components, the original variables are reduced to fewer factors which capture higher proportion of variability than global PCA. Choosing the first four components for all small areas is in line with our criteria for the number of components. Visual assessment of the 'winning' variables as defined by Harris et al. (2011) in Figure 5.10 reveals a pattern in northern Gauteng that is similar to rural parts of Mpumalanga. That is, the variables with highest loads are access to basic services for water and waste, dependent population prevalence and low levels of education among the adult population. In the extensively urban areas, centrally in Gauteng and in the main towns in Mpumalanga, dominating loads are on average household size, employment, household leadership, informality of dwellings and poverty. Small variable loads (less than 0.01) are not mapped. This is the case with average annual household incomes below the upper bound poverty line (`noinc_grant_dep`) which is omitted in Figure 5.10b.

A single index of vulnerability was obtained by weighted additive aggregation of geographically weighted factor scores. Factor scores are determined per small area as the sum of the product of small area loadings for that factor and the standardized data. The weights for adding the factor scores are the

Table 5.6: Spatial variation of the cumulative proportion of variance explained for the chosen four factors and that of the vulnerability index

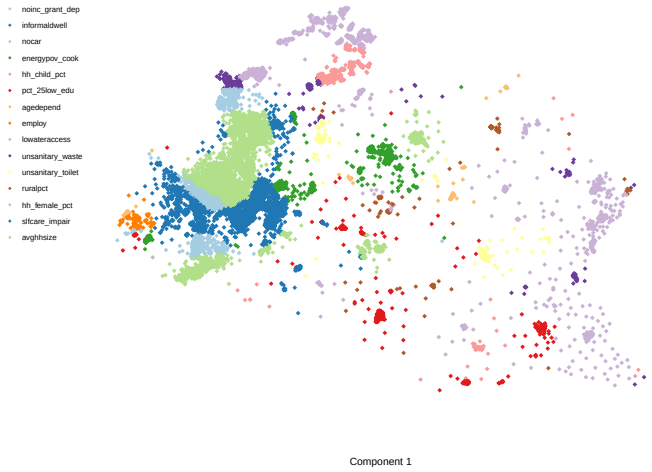
	Exploratory GWPCs					Vulnerability index	
	2 PCs	3 PCs	4 PCs	5 PCs	6 PCs	VI	VI (normalized)
Minimum	0.45	0.63	0.77	0.90	1.00	-7.29	0.00
1st quartile	0.63	0.76	0.85	0.93	1.00	-0.52	0.54
Median	0.65	0.78	0.87	0.94	1.00	0.24	0.58
3rd quartile	0.67	0.79	0.87	0.94	1.00	1.00	0.62
Maximum	0.94	0.97	0.99	1.00	1.00	7.45	1.00

proportion of total variance in each small area explained by the factor. As a result of the data being standardized, the index (VI in Table 5.6) is a real number with negative and positive values. For ease of interpretation, it is normalized to range between zero and one through the min-max normalization function. The overall summary of social vulnerability for the study region in terms of quartiles in Table 5.6 indicates that for most neighbourhoods, vulnerability with respect to socioeconomic factors is moderate. It becomes high for 25% of all small areas considered. Very high vulnerability levels (> 0.8) are indicated spatially in Figure 5.11 for more rural areas characterized by limitations in provision of basic services namely piped water access, routine waste collection and sanitary toilet systems (Figure 5.10a and 5.10b).

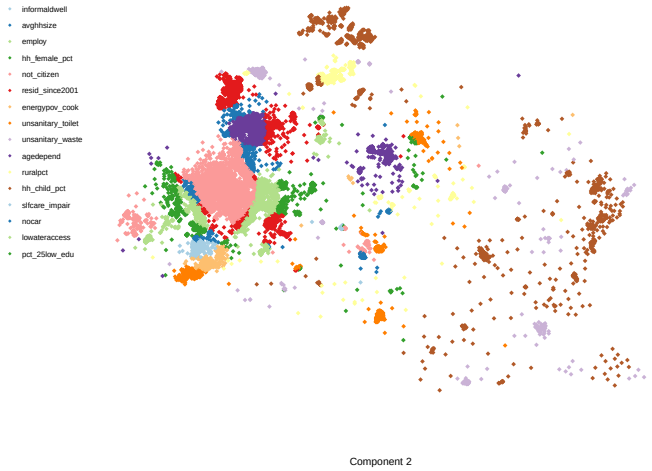
5.4.3 Mapping the risk of exposure to $PM_{2.5}$ and PM_{10}

Exceedance probabilities for PM_{10} and $PM_{2.5}$ were derived using conditional simulation and the resulting non-exceedance indicators formulated in Section 5.3.1. From Figure 5.12 high exceedance probability (> 0.6) of the PM_{10} threshold is observed in central Gauteng, along the north-west to south-east gradient, as well as close to towns in Mpumalanga. High exceedance probabilities are also indicated for the two municipalities on the north- and south-eastern border of Mpumalanga. For $PM_{2.5}$ the pattern in exceedance probabilities across Gauteng is more diffuse compared to the pattern for PM_{10}, but it is mostly high, namely more than 0.6 for most small area centroids. Similarly, $PM_{2.5}$ clusters of high exceedance probability occur around towns in Mpumalanga and on the north-eastern border.

We observe medium to high population density (0.4–0.8) throughout the study region as depicted in Figure 5.13. Small areas with the highest population densities (> 0.8) are in Gauteng. Risk maps for PM_{10} and $PM_{2.5}$ (Figure 5.14a and 5.14b) were obtained by geometric averaging of exceedance probabilities and the indicators of population density socioeconomic vulnerability. Figure 5.14a shows that population risk associated with exposure to excessive levels of PM_{10} is high (> 0.6) throughout the study area, with clusters of low risk areas (0.21–0.30) found west of the City of Tshwane, around Secunda and Hendriena. From Figure 5.14b the risk associated with $PM_{2.5}$ is more variable, but areas of high risk are still dominant, especially in Gauteng from south of the City of Tshwane to Vaal Marina.



a Variables with highest absolute local loading in first component



b Variables with highest absolute local loading in second component

Figure 5.10: Spatial distribution of variables with highest absolute local loadings

In Mpumalanga high risk areas include those around Balfour, Embalenhle, Ermelo and Kranskop.

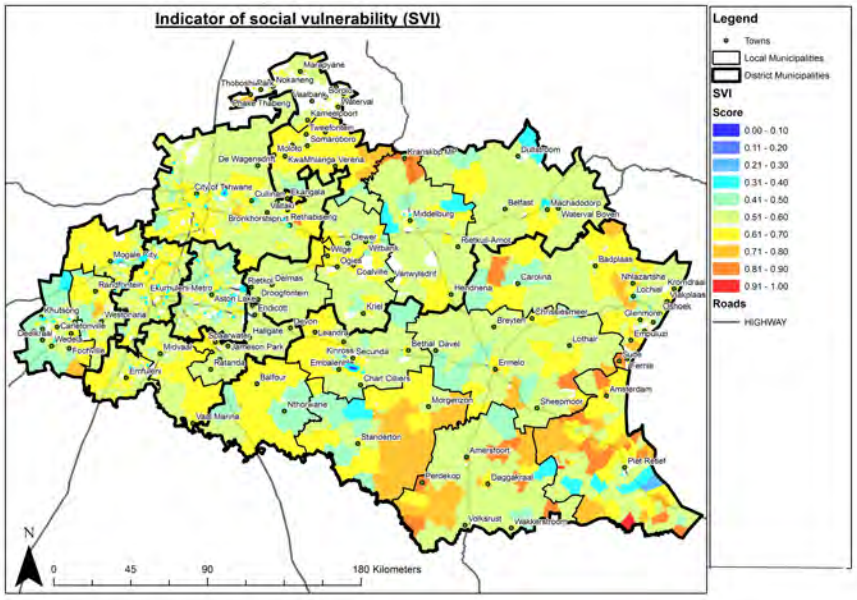


Figure 5.11: An indicator of social vulnerability mapped for census 2011 small areas in Gauteng and Mpumalanga (Nkangala and Gert Sibande districts)

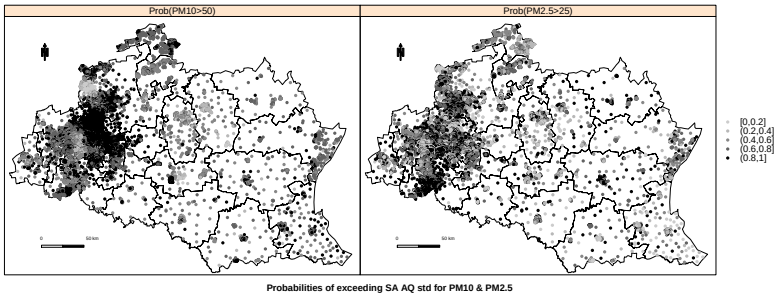


Figure 5.12: Probability of the annual averages $PM_{2.5}$ and PM_{10} concentrations being above 25 μg m^{-3} and 50 μg m^{-3} respectively

5.5 Discussion

Our objective was to integrate information on spatial exceedances of air quality standards for the annual average ambient concentrations of PM_{10}

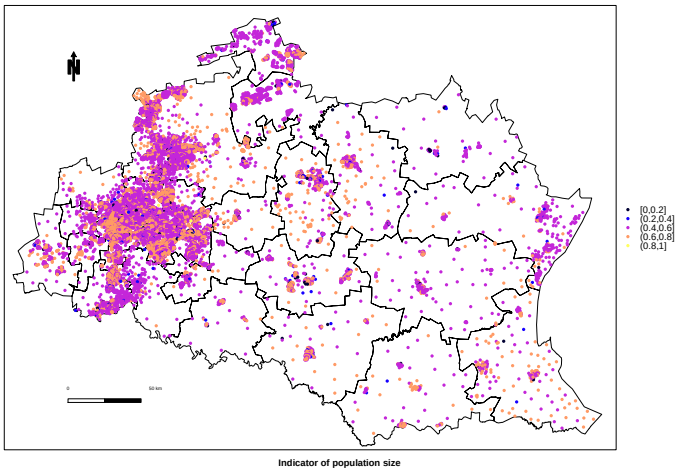
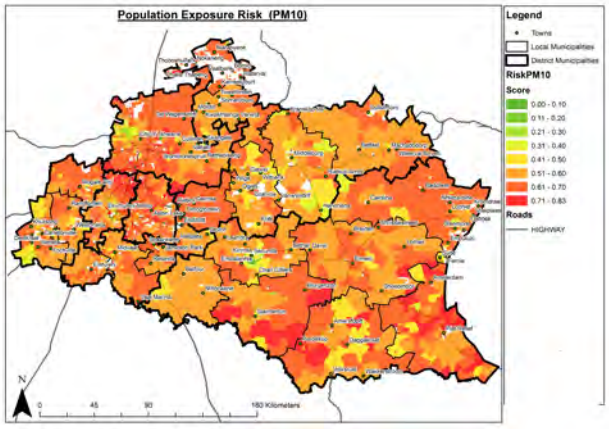


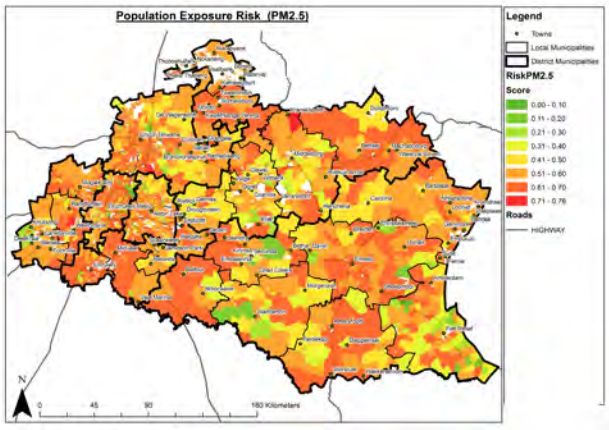
Figure 5.13: An indicator of relative population size at census 2011 small area centroids in Gauteng and Mpumalanga (Nkangala and Gert Sibande districts)

and $PM_{2.5}$ with information on population and housing characteristics for inference about the risk of exposure to excessive levels of particulate matter pollution. Excessive levels of PM_{10} and $PM_{2.5}$ were defined as hazardous and a probabilistic approach to hazard quantification was chosen. Specifically, spatiotemporal kriging with external drift based on a spatiotemporal metric covariance function was chosen for predicting PM_{10} and $PM_{2.5}$ at point locations without an air quality station. These prediction locations were purposefully chosen to be centroids of small areas for the South African census 2011 data for alignment with the population indicator for exposure and the social vulnerability index. These small areas can be linked to electoral wards, municipalities and provinces, making it possible to easily aggregate results to larger scale geopolitical units if required. The type of kriging method was chosen based on its applicability to data with irregular space-time pattern as a result of stations differing in operational times. Another selection factor was the availability of spatially extensive land cover and population count data for modelling non-stationarity of the random fields for PM_{10} and $PM_{2.5}$. This increases the reliability of PM_{10} and $PM_{2.5}$ predictions at unmeasured locations given our small spatial sample size.

The probability that the annual mean concentrations for PM_{10} will be above 50 μg m^{-3} is greater than 0.4 for the majority of small areas, indicating poor air quality for Gauteng and the Highveld in Mpumalanga. This is consistent with the Vaal region that is located south of Gauteng, eastern Gauteng and then western parts of Nkangala and Gert Sibande districts in Mpumalanga being declared priority air-sheds in 2006 and 2007 (Department of Environmental Affairs and Tourism, 2007). One of the reasons for declaring priority air-sheds is for implementing remedial actions. Comparing kriged



a Population risk associated with excessive annual average concentrations of PM_{10}



b Population risk associated with excessive annual average concentrations of $PM_{2.5}$

Figure 5.14: Spatial distribution (in small areas) of the population risk associated with exposure to particulate matter pollution

values for PM_{10} and $PM_{2.5}$ for 2009 and 2014 against those of 2011 revealed that annual average concentrations for 2009 were higher and that 2014 values were lower compared to 2011 concentrations. An enquiry into whether there were interventions implemented after the declarations in these areas and an evaluation of their effectiveness remains a subject of further research.

Differences in susceptibility to air pollution hazards were captured by using official data on population characteristics (Schwartz et al., 2011; Zhou et al., 2014). Previously, population counts were identified as a good proxy for pollutant emissions and the type of area, whether rural or urban for

mapping of ambient pollutant concentration (Paciorek and Liu, 2009; Kloog et al., 2011; Finazzi et al., 2013). Indicators of population size are also commonly used to quantify exposure or the number of elements at risk when quantifying environmental health risk. In our method for quantifying risk we also considered information on population vulnerability, going beyond population size as was considered in a previous air quality risk evaluation study (Finazzi et al., 2013). The risk estimates for PM_{10} and $PM_{2.5}$ were obtained by geometrically averaging exceedance probabilities, population at risk and the social vulnerability indicators, where the three components were weighted equally. Other ways of aggregation that reflect each components importance in quantifying risk and quality assessment of the resulting risk index will be explored in further work (Saisana et al., 2005; Paruolo et al., 2013).

The social vulnerability index of le Roux et al. (2015) was augmented by redefining input variables, using small areas rather than wards and accounting for spatial heterogeneity by applying a geographically weighted PCA (Saib et al., 2015). Through the social vulnerability index, socioeconomic disparities which are commonly known to be associated with disparities in health were accounted for. Further, spatial inequity in population risk related to air pollution which is important to consider in a developing country context (Wright et al., 2011) was also accounted for. Inequalities in socioeconomic conditions and lack of control of housing developments can result in communities whose standards of living are lower to be located close to pollution sources or for households to contribute to the pollution problem through burning of biofuels for cooking and heating (Saib et al., 2015). We concede that by using census data these and other situations concerning the states of the built, natural and social environments are only partially accounted for in quantifying the SVI. Census data are limited in terms of temporal efficiency and the amount of information that can be collected and therefore other spatially extensive and frequently collected data such as satellite imagery would need to be considered to enrich the current social vulnerability index. Ebert et al. (2009) showed that proxies of social vulnerability could be derived from earth observation imagery and GIS data, explaining almost 60% of the variance of an index derived from census small area data. The use of similar proxies and other community survey data for extending the SVI will be pursued in further research.

Maps of non-infectious chronic disease incidence and prevalence such as asthma do not exist for South Africa. This makes it difficult to assess whether there similarities between the PM_{10} and $PM_{2.5}$ population exposure risk maps in Figure 5.14 and spatial patterns in asthma prevalence over the study area. However, mortality statistics, namely the 10 leading causes of death are published at district municipal levels (Statistics South Africa, 2014). We focussed on deaths caused by chronic lower respiratory diseases (includes asthma) and found that from the deaths recorded at each district, Nkangala district municipality had higher percentages attributed to chronic lower respiratory diseases (Figure 5.15). For Tshwane, Gert Sibande and Ekurhuleni municipalities, chronic lower respiratory diseases were not among

the 10 leading causes of death. Upon visual inspection of the $PM_{2.5}$ population exposure risk map (Figure 5.14b) these three municipalities have higher proportions of small areas with lower risk scores (< 0.5) compared to the West Rand, Sedibeng and Johannesburg municipalities. There are published self-reported asthma prevalence estimates from cross-sectional public health surveys conducted within specific communities in our study area and location of these is indicated in red in Figure 5.15. Bertrams which is located in Johannesburg has the highest self-reported prevalence of asthma, the highest $PM_{2.5}$ risk estimate and high social vulnerability. Embalenhle (in Gert Sibande district) and Hillbrow have the lowest self-reported prevalence of asthma and the lowest $PM_{2.5}$ risk estimates. Apart from these locations where $PM_{2.5}$ exposure risk seems to be positively correlated with self-reported asthma prevalence, the pattern for the other locations is not clear. Comparison of our results with asthma prevalence based on medical diagnosis from a larger scale, epidemiologic study with would be more appropriate for validation.

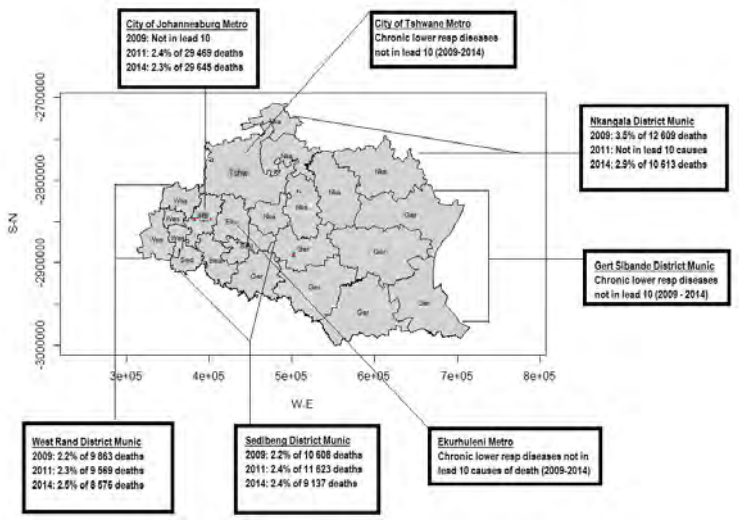


Figure 5.15: Mortality statistics attributed to chronic lower respiratory diseases at district municipality level with the district that each local municipality belongs to indicated. In red are the locations where cross-sectional public health studies were conducted between 2006 and 2010 and for which asthma prevalence estimates are presented in Table 5.7

$PM_{2.5}$ and PM_{10} are composites and as such it can be assumed that there is some degree of information loss resulting from interpolating them independently. In an effort to minimize such information loss, the same covariates were used to spatiotemporally interpolate both modes of particulate matter. Further work will include investigating the applicability of a multivariate spatiotemporal cokriging method developed by Liu and Koike (2007) for non-collocated network configurations and irregular absences of data. Interest

Table 5.7: Self-reported asthma prevalence from public health survey studies, with corresponding mean estimates of SVI, PM_{10} and $PM_{2.5}$ exposure risk for those specific communities surveyed

Sub-place name	SVI (std. dev.)	PM_{10} risk (std. dev.)	$PM_{2.5}$ risk (std. dev.)	Asthma Prevalence
Bertrams	0.62 (0.02)	0.64 (0.07)	0.61 (0.07)	20.0%[a]
Bram Fischerville	0.56 (0.03)	0.62 (0.05)	0.50 (0.11)	7.0%[a]
Hillbrow	0.64 (0.02)	0.66 (0.10)	0.46 (0.21)	2.0%[a]
Hospital Hill	0.59 (0.04)	0.58 (0.09)	0.58 (0.09)	3.0%[a]
Riverlea	0.62 (0.04)	0.54 (0.10)	0.52 (0.18)	19.0%[a]
Embalenhle	0.55 (0.04)	0.67 (0.03)	0.45 (0.10)	2.0%[b]
KwaGuqa	0.58 (0.04)	0.57 (0.07)	0.47 (0.13)	13.2%[c]
Vosman	0.64 (0.08)	0.49 (0.09)	0.53 (0.16)	9.6%[c]
eMpumelelweni	0.61 (0.05)	0.58 (0.05)	0.52 (0.08)	8.3%[c]

[a] Estimates for five areas surveyed in the Johannesburg Health, Environment and Development study that were published in Mathee et al. (2009)
[b] Estimates from a 2006 study of vulnerability to air pollution in eMbalenhle township near Secunda in the Gert Sibande district municipality (Matooane et al., 2011)
[c] Estimates from a cross-sectional public health survey in three communities at risk of exposure to poor air and water quality in the Nkangala district municipality (John et al., 2014)

will be on an extension that accommodates the presence of covariates. This has been explored in terms of an expectation-maximization algorithm for a linear coregionalization model with dynamic components to calibrate aerosol optical depth measurements from a satellite with PM_{10} measurements from a network of ground monitoring stations (Fassò and Finazzi, 2011).

Our exceedance probabilities for particulate matter were derived using geostatistical conditional simulations where the Gaussian distribution is assumed for the pollutant random fields. A semi-parametric method for deriving exceedance probabilities and associated confidence intervals based on bootstrapping can also be applied in further work (Schelin and Sjöstedt-de Luna, 2010; Cameletti et al., 2013a). This will be an opportunity to evaluate differences in results from the two methods. An alternative parametric method, based on Bayesian inference would be to implement a hierarchical spatiotemporal model for deriving exceedance probabilities and associated uncertainty (Cocchi et al., 2007; Cameletti et al., 2013b; Hamm et al., 2015).

An issue with our air quality monitoring network might be that the stations are preferentially located close to high pollution sources (Diggle et al., 2010). The effect of this would be upwardly biased predictions of $PM_{2.5}$ and PM_{10} at unmeasured locations. Our assumption was that effects of preferential sampling were mitigated by the use of land cover composition and population as covariates, such that preferential sampling would have negligible effect in variogram estimation for the residual spatial dependence (Finazzi et al., 2013). Further work would be to assess the effects of preferential sampling and implement methods for adjusting our resulting estimates to account for any biases that might exist (Zidek et al., 2014; Shaddick et al.,

2016).

5.6 Conclusion

A risk and vulnerability framework for evaluating population health risks related to exposure to excessive levels of ambient particulate matter pollution requires integration of quantities that describe pollutant hazard, exposure and gradients in susceptibility to the effects of exposure. We developed a methodology consisting of spatiotemporally interpolating PM_{10} and $PM_{2.5}$ using land cover composition and population counts covariates; and a joint spatiotemporal covariance function to overcome the challenge of the spatiotemporal sparsity of our air quality data. We found that the probability of exceeding PM_{10} and $PM_{2.5}$ regulatory thresholds is high for most areas in Gauteng and within the Highveld priority air-shed in Mpumalanga. However, annual average PM_{10} and $PM_{2.5}$ concentrations across the study area might be declining because our kriging results for 2014 showed that concentrations in most areas were lower in comparison to 2011 and 2009. We incorporated social vulnerability in quantifying risk to account for differences in susceptibility. Census data was used to quantify socioeconomic vulnerability related to exposure to poor air quality with the spatial dependence in small area population and housing characteristics accommodated by a geographically weighted PCA. The socioeconomic vulnerability map provides valuable spatial information in starting to understand the human and housing conditions in areas where the regulatory air quality limits are regularly exceeded.

Synthesis

6

Elucidating areas and times for which ambient particulate matter concentrations are chronically high is important for determining associated health effects and health risk assessment. The link with health requires reliable health data for deriving relative risk estimates that form the basis of concentration-response functions (WHO, 2016). In the absence of specific health data, understanding the spatial and temporal dynamics of particulate matter (PM) remains important for more general population exposure and risk assessments. Typical steps in population exposure and risk assessment include identifying key emission sources in an area, determining the spatial distribution of PM and assessing exposure and risk at a neighbourhood scale by linking the estimated PM surface with variables of population density and vulnerability. Spatial statistical models are to be considered throughout this process. They can be used to reduce the complexity of pollutant dispersion in urban air by identifying influential factors. Spatial statistical methods can also be used to solve problems of spatial and temporal sparsity of air quality monitoring network data. Spatial sparsity refers to the number of air quality stations being small and irregularly placed in a densely populated conurbation, and temporal sparsity refers to data coverage, defined by Brown and Woods (2014) as, "the proportion of a relevant reference period for which valid measurements are available". Clearly, the presence of missing values reduces data coverage.

6.1 Research findings

The aim of this research was to statistically map PM_{10} and $PM_{2.5}$ for the purpose of population exposure and risk assessment. Specific challenges with air quality data were the presence of missing values and the limited number of stations in the study region. To achieve the aim, four specific objectives were pursued and in this section research findings and conclusions for each objective are discussed.

6.1.1 First objective: Modelling considerations for statistical mapping of particulate matter in a large densely populated area with few monitoring stations

The first objective was to evaluate the performance of the generalized linear geostatistical models (GLGM) and kriging for mapping the threshold exceedance rate of PM_{10}. Four models, namely ordinary kriging, external drift kriging, the log-Gaussian and the Poisson generalized linear geostatistical models were implemented. As part of this the applicability of household census data chosen as representative of domestic sources of PM_{10} emissions as covariates in the GLGM and kriging were assessed. Daily PM_{10} data from 36 air quality monitoring sites in the Highveld in South Africa for the 48 month period from September 2008 to August 2012 were aggregated into yearly exceedance counts. Missing values at this stage were not imputed. Domestic emissions proxies were extracted from the 2011 small area dataset of the South African census. These were percentages per small area of informal dwellings and the household use of solid biofuels (wood, coal and dung) for heating and for cooking. Exploratory analysis results indicated that higher concentrations of PM_{10} coincided with high density residential areas, especially in those with informal dwellings. Informal dwellings percentage was a statistically significant predictor of the PM_{10} annual exceedance rate. Household use of biofuels for heating was a significant covariate on its own, but became insignificant when informal dwelling percentage was included indicating collinearity. To this end dwelling informality was viewed as being a stronger proxy of domestic emissions resulting from energy poverty, the higher likelihood of being located in pollution prone areas and the state of the built environment in these areas, namely the prevalence of dust from unpaved roads and yards.

Mapping of the annual average exceedance rate of the South African PM_{10} standard of 120 μg m^{-3} was considered important for determination of hot spots and factors that contribute to their formation for the purpose of air quality management. Kriging and the GLGM predictions were systematically higher than actual observations at the validation stations. The relative accuracy of predictions to the actual observations was highest for ordinary kriging as compared to the log-Gaussian and Poison models without covariates. A higher relative prediction accuracy was achieved with external drift kriging as compared to the model-based alternatives with covariates. Predictions from models with covariates were higher in areas where the density of informal dwellings was higher. Overall, maps of the PM_{10} annual exceedance rate from all four methods were similar in terms of location of hot-spots and areas of lower exceedance rates, but kriging methods were better in terms of prediction accuracy. The advantage of using spatial predictors to improve the reliability of PM_{10} maps given a spatially sparse monitoring network can only materialize if the covariate and response surfaces are correlated. The degree of improvement depends upon the strength of this correlation. In our case prediction uncertainty in areas without air quality stations improved if the covariate was included.

6.1.2 Second objective: Use satellite imagery to identify fugitive dust sources of PM_{10}

The research question being answered in this objective was whether information on fugitive dust emission sources in the neighbourhood of an air quality monitor could be used in predicting ambient PM_{10} concentrations on days characterized by strong winds. The first part of the objective was to identify sources of fugitive dust from satellite imagery by extending the maximum likelihood classifier to improve classification of bare soil with the variation in soil types in the study region. The second part was to develop a method to postprocess land cover class output such that it can be statistically associated with average PM_{10} concentrations on days when local wind speeds were favourable to fugitive dust emissions.

An ensemble maximum likelihood method was developed for pixel-based land cover classification of SPOT 6 multi-spectral images for circular areas of 4 km from an air quality station. The ensemble maximum likelihood classifier was based on iterative training in targeted areas that represent the variety in soil types that are prevalent in the study region for the purpose of improving the accuracy of bare soil classification. Other primary land cover classes that were considered were built-up areas, vegetation, water and 'mixed bare soil'. The mixed bare soil class was defined as areas where soil was mixed with either vegetation (mainly degraded grass) or synthetic materials. Preliminary validation of the ensemble classifier for the bare soil class resulted in accuracies ranging from 65% to 98% for seven considered neighbourhoods. It was lower for densely built-up areas, specifically in the central business districts with high rise buildings, commercial and industrial properties, whereas areas that were less heterogeneous in terms of the types of buildings had high classification accuracies for bare soil. Final validation of all classes on test neighbourhoods resulted in an overall classification accuracy of 78%. Vegetation classification was most accurate, whereas there was some confusion between the built-up and mixed bare soil classes. It would therefore be worthwhile to find ways to improve the ensemble learning for the built-up class because it is important in urban air quality mapping.

Post-processing of land cover data involved applying k-means cluster analysis and a varying intercepts regression model to regress land cover clusters and a fugitive dust emissions proxy with average PM_{10} for days where wind speeds were in excess of $4\,\mathrm{m\,s^{-1}}$. Land cover clusters in the neighbourhood of an air quality station were significant predictors of wind-related average PM_{10} concentrations, whereas the fugitive dust emissions proxy was not a significant predictor. The proxy or indicator was developed in Europe using land cover data, soil texture map, gridded meteorological data and emissions factors. Similarly, detailed information was not available for our study area and this could be the reason for the indicator's insignificance as a covaraite. In the absence of an emissions inventory for PM_{10} from which dust emissions could be accurately quantified, land cover clusters derived to characterize dust emission reservoirs, were significantly correlated with wind-related average

PM_{10}. Therefore, land cover data were identified as a source in a search for suitable covariates to improve the prediction of PM_{10} at locations without air quality monitoring stations.

6.1.3 Third objective: Imputation of missing air quality data

High quality monitoring data are important for developing air quality regulations, but data from air quality stations commonly suffer from incompleteness. The third objective was to develop a bootstrap based regression method to multiply impute the missing pollutant values. This method leveraged on correlations between PM_{10}, NO_2, SO_2 and meteorological variables namely, relative humidity, temperature, wind speed and wind direction. In a multiple imputation method, the uncertainty associated with imputing missing values can be evaluated. The aim was to impute pollutant data using the air quality station's meteorological variables as covariates, but the covariates were themselves substantially incomplete. Therefore, the method was developed to start with the multiple imputation of meteorological variables using data from a weather station that was closest to the air quality station. Next, NO_2 and SO_2 would be regressed on the completed meteorological variables for the prediction of missing values. Multiple imputations of NO_2 and SO_2 resulted from multiple completed meteorological datasets. Finally, PM_{10} was multiply imputed conditional on multiply completed NO_2, SO_2 and meteorological variables.

Other data challenges included non-constant variance and serial correlation of the variates. Non-constant variance has implications for the suitability of the Gaussian distributional assumption. Therefore, suitable parametric models for the bootstrap regressions were assessed, namely the generalized linear model assuming a Gamma error distribution with a log-link function and the Gaussian linear model for log-transformed dependent variates. The log-Gaussian model was chosen because of the higher proportion of variance explained by the model in comparison to the log-Gamma generalized linear model. To account for serial correlation, generalized least squares inference was implemented incorporating a first order autoregressive structure for the residuals. Prediction accuracy of the bootstrap regression imputation method was compared to the approximate Bayesian bootstrap (ABB) method using a hold-out sample of observed PM_{10} values. Better imputation quality for pollutants NO_2, SO_2 and PM_{10} was achieved with bootstrap regression method compared to the ABB method. The inclusion of covariates in the form of meteorological variables, gaseous pollutant and seasonal factor variables in the bootstrap regression method resulted in imputations with seasonal structure and variability similar to observed values, whereas ABB imputations had no seasonal structure and minimum variation. Instrument failure, shortages in human and financial resources are some of the reasons for missing air quality data. For air quality officers and secondary users of these data, multiple imputation is an effective way to impute missing values and to quantify the uncertainty associated with the imputations. Prior to releasing the data to users, the custodians should ensure that they are quality screened based

on their expert knowledge of the station's location, instrumentation and operational conditions. We demonstrated that users can screen the quality of the data by using published guidelines on practical ranges for the values of the different atmospheric variables and exploratory plots to identify values that may be invalid.

6.1.4 Fourth objective: Neighbourhood-level risk of exposure to high ambient concentrations of particulate matter

Exposure to even moderate concentrations of particulate matter leads to increased risks of cardiopulmonary morbidity and mortality. Those living in poor housing conditions are more susceptible to the effects of being exposed to poor air quality. The objective was to develop a methodology to quantify population risks related to exposure to particulate matter pollution by combining information on $PM_{2.5}$ and PM_{10} exceedance probabilities with data on population size and social vulnerability. Data that were used included the annual average $PM_{2.5}$ and PM_{10} concentrations from 37 air quality monitoring stations for the years 2008–2014, the 2013–2014 South African land cover data and the 2011 small area census dataset. Exceedance probabilities of $PM_{2.5}$ and PM_{10} were derived using conditional simulations based on kriging with external drift models with joint spatiotemporal covariance functions. With external drift kriging PM_{10} and $PM_{2.5}$ were spatiotemporally interpolated using land cover composition and population counts covariates to account for location specific particle emission and dispersion characteristics. The probability of exceeding PM_{10} and $PM_{2.5}$ regulatory thresholds was high (> 0.6) especially in central Gauteng and within the priority air-shed in Mpumalanga. There was an indication that annual average PM_{10} and $PM_{2.5}$ concentrations were declining because concentrations in most areas were lower in 2014 compared to 2011 and 2009.

Social vulnerability was incorporated in quantifying risk to account for differences in susceptibility of the exposed population. A composite spatial indicator of social vulnerability was developed using geographically weighted PCA. The spatial patterns of social vulnerability differed from the pure spatial distribution of population counts in our study region. Small areas associated with higher (> 0.7) social vulnerability were located south-east of Mpumalanga where low education levels and lack of some basic services such as piped water at residences were indicated. PM_{10} exceedance probabilities were also high for these small areas, resulting in higher (> 0.7) PM_{10} risk compared to other parts of Mpumalanga. The addition of social vulnerability provides valuable spatial information for understanding social and settlement conditions in areas where regulatory thresholds are regularly exceeded.

6.2 Reflections and outlook

With the rapid pace of urbanization, there is an increased pressure on governments to ensure that cities thrive in terms of economic development, quality

of life and sustenance of the environment. This is the ethos of Goal 11 of the United Nations sustainable development agenda. The sixth target of this goal is to "reduce the adverse per capita environmental impact of cities by 2030, including by paying special attention to air quality and municipal and other waste management". Therefore, one of the indicators for monitoring progress against this target is the population weighted annual mean level of $PM_{2.5}$ and PM_{10}. Mapping of air quality has therefore become increasingly important in all countries. This presents a challenge for developing countries because historically the monitoring of air quality was restricted to locations with heavy polluting industries. Further, the limitation of resources and politics have had an impact on the number of stations commissioned and the quality of the collected data.

When air quality data are substantially imperfect, mapping the targeted indicators is a challenge. In this thesis I presented solutions to the problems of (i) the poor quality data due to errors and missing values, and (ii) the limitation having few observation locations to reliably map the targeted indicators for a given urban region. Methods for imputing missing data on a station by station basis are relevant if the network is owned in parts by different entities (municipalities, environment department or industrial companies) and each entity owns a handful of stations. Ideally such data quality assessments and corrections would be done online to ensure that data coverage is close to 100%. This would require that current monitoring infrastructure be upgraded to smart networks of sensors and information systems for online monitoring supported by a sustainable pool of skilled human resources for data quality control and management. In financially constrained environments, such upgrades are not be immediately possible, but an alternative can be the environmental department's issuance of stringent requirements for 100% data coverage at each station and shorter reporting periods. This would encourage frequent data quality checks throughout the year and early detection and treatment of invalid or missing data.

In this thesis I used spatially extensive covariates to improve the reliability of $PM_{2.5}$ and PM_{10} maps from a limited number of air quality stations that were in the study area. For accurate monitoring and reporting, the monitoring network needs to be upgraded with current and mid-term population growth and spatial development prospects considered. Previously air quality monitoring network design considered locations of heavy pollution sources. However, with the current global and national policy environment focussing strongly on sustainability, land use and transportation planning should be considered when networks are designed. Further, advocacy for improved quality of life and inclusive communities also imply that population exposure and vulnerability need to be considered. Population exposure and vulnerability should include insights on locating schools and communal residential facilities for the aged, the physically impaired and the frail. Community health care facilities are less of a concern in developed nations where there are strict regulations regarding proximity to pollution sources and the quality of indoor environments. This is not always the case in developing

countries, especially in less affluent areas where building standards are not adhered to. If, however, network design decisions are left to be governed by policy, economic, geographical and human factors, then consistent and comprehensive monitoring would not be achieved. Therefore, to avoid suboptimal network designs, geostatistical methods are used in practice to delineate optimal sample locations for spatial networks based on typically minimizing some objective function of the variance. Ways of incorporating data on land use, specific emission source information as well as population exposure and vulnerability would need to be considered in future research.

Suitable and spatially extensive ancillary data improves the reliability of air quality maps when there are few ground monitoring stations. The area studied in this thesis is extensively industrial, especially mining with much of the landscape having features that are evident of the mining heritage. Gold mining residue deposits are found in central Gauteng whilst fly-ash deposits are found mostly close to power generation facilities which are more abundant in Mpumalanga. Housing informality is also prevalent in the study area. Land cover and household-related census datasets were therefore ancillary data sources used in this thesis to extract variables on informal dwellings, land cover composition and domestic burning of biofuels for mapping PM_{10} and $PM_{2.5}$. Significant covariates were dwelling informality and land cover composition (clusters) and these were successfully used to map annual average PM_{10} and $PM_{2.5}$ as well as exceedance probabilities of annual standards set for these pollutants. Continued efforts for rehabilitation of mine dumps, including use of waste materials to manufacture construction materials as well as rapid development of the built environment imply that land cover data would need to be regularly updated for use in mapping air quality. For example, the period between the national land cover dataset used in this study and the previous one is 12 years and it is inappropriate to use them for monitoring applications as discussed in this thesis.

Land cover change studies report significant encroachment of settlements around mine residue deposits. The residents also regularly complain about the dust from these structures. The abatement of dust through fertile soil and vegetation cover fails if there is no continuous monitoring and maintenance of the vegetation cover sealing the dumps. Another rehabilitation strategy has been to convert dumps into other land uses including turning them into leisure facilities. This could work if the dumps are permanently sealed with durable synthetic material that also prevents leaching of chemicals that are trapped in the residue. Fly ash is used to enhance the quality of concrete and in manufacturing bricks and pavers. This is found to be a cost-effective and environmentally sustainable way to eradicate ash dumps and develop skills in communities close to these sites. More resources should be mobilized towards finding similar technology solutions for transforming other types of mining waste material into products, thus eradicating these sources of hazardous dust from the landscape.

Vehicular and industrial combustion-related emissions are a concern in the

study area. Further work in mapping air quality should incorporate data on vehicular and industrial emissions using available traffic count and industrial activities data. For example, locations of power generation plants and some large industries are known. These could be combined with published production statistics and specific industry related emission factors to account for the industrial contribution to measured ambient levels of $PM_{2.5}$ and PM_{10}. In this thesis, through the use of land cover data, I restricted myself to only the industrial proportion of the small area. Another air quality mapping consideration, especially for daily spatiotemporal modelling, is inclusion of the effect of local variation meteorological conditions in combination with surface roughness and other landscape variables. In this thesis the importance of meteorology in accounting for temporal variability of pollutants was shown. For daily spatiotemporal pollutant models with meteorological covariates, however, gridded meteorological data would be required at spatial resolutions appropriate for pollutant dispersion in urban air. This can be a challenge.

The final result that the thesis aimed for was neighbourhood level mapping of population risk due to exposure to poor air quality. Such information can be useful in directing epidemiological research towards specific communities to assess whether areas where current regulatory thresholds are being exceeded continuously also show increased relative risks for particular health outcomes. Community health clinic aggregated admissions count data can also be assessed against PM_{10} and $PM_{2.5}$ risk maps produced in this thesis to check whether clinics in high risk areas have higher counts for specific conditions such as asthma. This could be further investigated by checking for correlations between daily pollutant and admissions data. In further work, the insights on annual average PM_{10} and $PM_{2.5}$, the exceedance probabilities, social vulnerability and risk will be used for further analysis of an existing cross-sectional public health study, namely the Johannesburg Health, Environment and Development (HEAD) study (Mathee et al., 2009). The HEAD study sampled households from five communities of lower socioeconomic standards, annually from 2006 until 2010 and included self-reported prevalence of asthma and other respiratory illness in the survey. They found that Riverlea, a community surrounded by mine dumps near Diepkloof in Soweto had higher prevalence of asthma, but there was no further analysis to determine whether outdoor air pollution levels contributed to the higher relative risks of asthma in this community. This will clearly be of interest for further work.

Information similar to this thesis can be used in spatial planning including transport planning and urban design as a basis to advocate for certain types of developments that promote for example walking, cycling, use of public transport and the availability of green spaces as a way to improve urban air quality. Remote sensing data are already being used to study carbon sequestration potential of trees, therefore promoting the planting of trees along street canyons and in urban forests to reduce the footprint of a city in terms of green house gases. This area of research has also expanded into investigations of the role of vegetation and buildings in the removal of

particulate matter. In the future, smart sensors could be used to regulate traffic in cities to minimize congestion and associated emissions. Further, the use of mobile data to track diurnal population movements for better understanding of exposure patterns could become common as will be the demand for online information on the state of the air around cities. Therefore air quality mapping and population risk assessment will continue to grow as an area of research, especially as solutions are sought for challenges of spatial data quality and the need to integrate different types of data from disparate sources. Some of the decisions for creating healthy liveable cities will depend on insights gained from such research.

Bibliography

Abayomi, K., Gelman, A., Levy, M., 2008. Diagnostics for multivariate imputations. Journal of the Royal Statistical Society 57 (3), 273–291.

Alade, O. L., 2010. Characteristics of particulate matter over the South African industrialized Highveld. MSc research report, University of the Witwatersrand, South Africa.

Alduchov, O. A., Eskridge, R. E., 1996. Improved magnus form approximation of saturation vapor pressure. Applied Meteorology 35, 601–609.

Athanasopoulou, E., Tombrou, M., Russell, A. G., Karanasiou, A., Eleftheriadis, K., Dandou, A., 2010. Implementation of road and soil dust emission parameterizations in the aerosol model CAMx: Applications over the greater Athens urban area affected by natural sources. Journal of Geophysical Research 115 (D17301), 1–21.

Balmer, M., 2007. Household coal use in an urban township in South Africa. Journal of Energy in Southern Africa 18 (3), 27–32.

Banerjee, B., Bovolo, F., Bhattacharya, A., Bruzzone, L., Chaudhuri, S., Mohan, B., 2015. A new self-training-based unsupervised satellite image classification technique using cluster ensemble strategy. IEEE Geoscience and Remote Sensing Letters 12 (4), 741–745.

Banerjee, M., Capozzoli, M., McSweeney, L., Sinha, D., 1999. Beyond Kappa: A review of interrater agreement measures. Canadian Journal of Statistics 27 (1), 3–23.

Banerjee, S., Carlin, B. P., Gelfand, A. E., 2004. Hierarchical modeling and analysis for spatial data. Monographs on statistics and applied probability. Chapman and Hall CRC, USA.

Barmpadimos, I., Hueglin, C., Keller, J., Henne, S., Prévôt, A. S. H., 2011. Influence of meteorology on PM_{10} trends and variability in Switzerland from 1991 to 2008. Atmospheric Chemistry and Physics 11, 1813–1835.

Bartlett, J. W., Seaman, S. R., White, I. R., Carpenter, J. R., 2015. Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. Statistical methods in medical research 24 (4), 462–487.

Beelen, R., Hoek, G., Pebesma, E., Vienneau, D., de Hoogh, K., Briggs, D. J., 2009. Mapping of background air pollution at a fine spatial scale across the European Union. Science of The Total Environment 407 (6), 1852–1867.

Besag, J., 1986. On the statistical analysis of dirty pictures. Journal of the Royal Statistics Society 48 (3), 259–302.

Bivand, R. S., Pebesma, E. J., Gómez-Rubio, V., 2008. Applied spatial data analysis with R, 1st Edition. Use R. Springer, USA.

Blangiardo, M., Cameletti, M., 2015. Spatial and spatio-temporal Bayesian models with R-INLA, 1st Edition. Wiley, United Kingdom.

Bolton, D., 1980. The computation of equivalent potential temperature. Monthly Weather Review 108, 1046–1053.

Brown, R. J. C., Woods, P. T., 2014. Proposals for new data quality objectives to underpin ambient air quality monitoring networks. Accreditation and Quality Assurance 19 (6), 465–471.

Brus, D. J., Gruijter, J. J., 1997. Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). Geoderma 80, 1–44.

Cameletti, M., Ignaccolo, R., Sylvan, D., 2013a. Assessment and visualization of threshold exceedance probabilities in complex space-time settings: A case study of air quality in Northern Italy. Spatial Statistics 5, 57–68.

Cameletti, M., Lindgren, F., Simpson, D., Rue, H., 2013b. Spatio-temporal modeling of particulate matter concentration through the SPDE approach. Advanced Statistical Analysis 97, 109–131.

Chen, Q., Yang, H., Liu, T., Zhang, L., 2016. Household biomass energy choice and its policy implications on improving rural livelihoods in Sichuan, China. Energy Policy 93, 291–302.

Chen, Y., Gong, P., 2013. Clustering based on eigenspace transformation–CBEST for efficient classification. Journal of Photogrammetry and Remote Sensing 83, 64–80.

Chervenkov, H., Jakobs, H., 2011. Dust storm simulation with regional air quality model: Problems and results. Atmospheric Environment 45 (24), 3965–3976.

Chikusa, C. M., January 1994. Pollution cased by mine dumps and its control. MSc dissertation, Rhodes University, Grahamstown, South Africa.

Chirino, Y. I., Sánchez-Pérez, Y., Osornio-Vargas, A. R., Morales-Bárcenas, R., Gutirrez-Ruíz, M. C., Segura-García, Y., Rosas, I., Pedraza-Chaverri, J., García-Cuellar, C. M., 2010. PM10 impairs the antioxidant defense system and exacerbates oxidative stress driven cell death. Toxicology Letters 193 (3), 209–216.

Christakos, G., Olea, R. A., Serre, M. L., Yu, H.-L., Wang, L., 2005. Interdisciplinary public health reasoning and epidemic modelling: The case of black death. Springer-Verlag, New York.

Churg, A., Brauer, M., Avila-Casado, M. C., Fortoul, T. I., Wright, J. L., 2003. Chronic exposure to high levels of particulate air pollution and small airway remodeling. Environmental Health Perspectives 111 (5), 714–718.

Cocchi, D., Greco, F., Trivisano, C., 2007. Hierachical space-time modelling of PM_{10} pollution. Atmospheric Environment 41, 532–542.

Cressie, N., Wikle, C. K., 2011. Statistics for spatio-temporal data. Probability and Statistics. Wiley, USA.

CSIR, 2014. 10^{th} Annual State of Logistics Survey for South Africa, 2013. Tech. rep., CSIR, Pretoria, South Africa.

David, F., 1988. Multiplicative errors: Log-Normal or Gamma? Journal of the Royal Statistical Society. Series B (Methodological) 50 (2), 266–268.

De Jong, R., Van Buuren, S., Spiess, M., 2016. Multiple imputation of predictor variables using generalized additive models. Communications in Statistics-Simulation and Computation 45 (3), 968–985.

De Longueville, F., Hountondji, Y. C., Henry, S., Ozer, P., 2010. What do we know about effects of desert dust on air quality and human health in West Africa compared to other regions? Science of The Total Environment 409 (1), 1–8.

Dean, A. M., Smith, G. M., 2003. An evaluation of per-parcel land cover mapping using maximum likelihood class probabilities. International Journal of Remote Sensing 24 (14), 2905–2920.

Department of Environmental Affairs and Tourism, November 2007. Declaration of the Highveld as Priority Area in terms of Section 18(1) of the National Environmental Management: Air Quality Act, 2004 (Act No. 39 of 2004). Gazette 30518, Government of the Republic of South Africa, Pretoria, South Africa.

Department of Environmental Affairs and Tourism, June 2012. National environmental management air quality Act 2004: National ambient air quality standard for particulate matter with aerodynamic diameter less than 2.5 micron metres ($PM_{2.5}$). Gazette 35463, Government of the Republic of South Africa, Pretoria, South Africa.

Department of Labour, August 2016. Basic guide to child labour. Online, accessed 10 January 2017.

Diggle, P. J., Menezes, R., Su, T., 2010. Geostatistical inference under preferential sampling. Journal of the Royal Statistical Society 59 (2), 191–232.

Diggle, P. J., Ribeiro Jr, P. J., 2007. Model-based geostatistics. Statistics. Springer, USA.

Dudeni-Tlhone, N., Holloway, J., Khuluse-Makhanya, S., Koen, R., November 2013. Clustering of housing and household patterns using 2011 population census. In: Annual proceedings of the South African Statistical Association Conference. South African Statistical Association, South Africa, pp. 23–30.

Ebert, A., Kerle, N., Stein, A., 2009. Urban social vulnerability assessment with physical proxies and spatial metrics derived from air- and spaceborne imagery and GIS. Natural Hazards 48, 275–294.

EEA, 1995. CORINE land cover project. Report Part II - Nomenclature, European Environment Agency.

Efron, B., 1979. Bootstrap methods: Another look at the jackknife. The Annals of Statistics 7 (1), 1–26.

Efron, B., 1994. Missing data, imputation and the bootstrap. Journal of the American Statistical Association 89 (426), 463–475.

Efron, B., 2012. Bayesian inference and the parametric bootstrap. The Annals of Applied Statistics 6 (4), 1971–1997.

Efron, B., 2015. Frequentist accuracy of Bayesian estimates. Journal of the Royal Statistical Society 77 (3), 617–646.

Ehrlich, R., Jithoo, A., 2006. Chronic diseases of lifestyle in South Africa: 1995–2005. Technical report, South African Medical Research Council, Cape Town, South Africa.

Elliott, P., Wakefield, J., Best, N., Briggs, D. (Eds.), 2000. Spatial epidemiology: Methods and applications. Oxford University Press, USA.

Faraway, J. J., 2006. Extending the linear model with R: Generalized linear, mixed effects and nonparametric regression models. Texts in Statistical Science. Chapman & Hall/CRC.

Fassò, A., Finazzi, F., 2011. Maximum likelihood estimation of the dynamic coregionalization model with heterotopic data. Environmetrics 22 (6), 735–748.

Finazzi, F., Scott, E. M., Fassò, A., 2013. A model-based framework for air quality indices and population risk evaluation, with an application to the analysis of Scottish air quality data. Journal of the Royal Statistical Society 62 (2), 287–308.

Fisher, N. I., Lee, A. J., 1992. Regression models for an angular response. Biometrics 48 (3), 665–677.

Freedman, D. A., 1981. Bootstrapping regression models. The Annals of Statistics 9 (6), 1218–1228.

GBD 2013 Risk Factors Collaborators, 2015. Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks in 188 countries, 1990–2013: A systematic analysis for the Global Burden of Disease Study 2013. Lancet 386, 2287–2323.

Gelman, A., Hill, J., 2007. Data analysis using regression and multilevel/hierachical models. Analytical methods for social research. Cambridge University Press, USA.

Goovaerts, P., Webster, R., Dubois, J.-P., 1997. Assessing the risk of soil contamination in the Swiss Jura using indicator geostatistics. Environmental and Ecological Statistics 4 (1), 49–64.

Gorte, B., Stein, A., 1998. Bayesian classification and class area estimation of satellite images using stratification. IEEE Transactions in geoscience and remote sensing 36 (3), 803–812.

Goudie, A. S., 2009. Dust storms: Recent developments. Journal of Environmental Management 90 (1), 89–94.

Gräler, B., Pebesma, E., Heuvelink, G., 2016. Spatio-temporal interpolation using gstat. R Journal 8 (1), 204–218.

Gryparis, A., Paciorek, C. J., Zeka, A., Schwartz, J., Coull, B. A., 2009. Measurement error caused by spatial misalignment in environmental epidemiology. Biostatistics 10 (2), 258–274.

GTI, 2015. 2013-2014 South African national land-cover dataset. Data User Report and Metadata Version 05, GEOTERRAIMAGE, Pretoria, South Africa.

Guttman, L., 1954. Some necessary conditions for common-factor analysis. Psychometrika 19 (2), 149–161.

Hamm, N. A. S., Finley, A. O., Schaap, M., Stein, A., 2015. A spatially varying coefficient model for mapping PM_{10} air quality at the European scale. Atmospheric Environment 102, 393–405.

Harris, P., Brunsdon, C., Charlton, M., 2011. Geographically weighted principal components analysis. International Journal of Geographical Information Science 25 (10), 1717–1736.

He, Y., Raghunathan, T. E., 2009. On the performance of sequential regression multiple imputation methods with non Normal error distributions. Communications in Statistics (Simulation and Computation) 38, 856–883.

Heitjan, D. F., Little, R. J. A., 1991. Multiple imputation for the fatal accident reporting system. Journal of the Royal Statistics Society 40 (1), 13–29.

Higgins, C. W., Froidevaux, M., Simeonov, V., Vercauteren, N., Barry, C., Parlange, M. B., 2012. The effect of scale on the applicability of Taylor's Frozen Turbulence Hypothesis in the atmospheric boundary layer. Boundary-Layer Meteorology 143 (2), 379–391.

Hnizdo, E., Vallyathan, V., 2003. Chronic obstructive pulmonary disease due to occupational exposure to silica dust: A review of epidemiological and pathological evidence. Occupational and Environmental Medicine 60 (4), 237–243.

Housing Development Agency, 2013. Gauteng informal settlement status (2013). Technical research report, Housing Development Agency, Johannesburg, South Africa.

Janssen, S., Dumont, G., Fierens, F., Mensink, C., 2008. Spatial interpolation of air pollution measurements using CORINE land cover data. Atmospheric Environment 42, 4884–4903.

Jerrett, M., Arain, A., Kanaroglou, P., Beckerman, B., Potoglou, D., Sahsuvaroglu, T., Morrison, J., Giovis, C., 2005. A review and evaluation of intraurban air pollution exposure models. Exposure Analysis and Environmental Epidemiology 15, 185–204.

John, J., Das, S., 2012. Vulnerability of a low-income community in South Africa to air pollution: Exploring the use of structural equations modelling to identify appropriate interventions. Journal of Integrative Environmental Sciences.

John, J., Wright, C. Y., Oosthuizen, M. A., Steyn, M., Genthe, B., le Roux, W., Albers, P., Oberholster, P., Pauw, C., 2014. Environmental health

outcomes and exposure risks among at-risk communities living in the Upper Olifants River Catchment, South Africa. International Journal of Environmental Health Research 24 (3), 195–214.

Jones, A., Breuning-Madsen, H., Brossard, M., Dampha, A., Deckers, J., Dewitte, O., Gallali, T., Hallett, S., Jones, R., Kilasara, M., Le Roux, P., Micheli, E., Montanarella, L., Spaargaren, O., Thiombiano, L., Van Ranst, E., Yemefack, M., Zougmoré, R., E., 2013. Soil atlas of Africa. Tech. rep., European Commission, Publications Office of the European Union, Luxembourg.

JRC European Commission, 2008. Handbook on constructing composite indicators: Methodology and User guide. OECD publishing.

Kaiser, H. F., 1961. A note on Guttman's lower bound for the number of common factors. British Journal of Statistical Psychology 14 (1), 1–2.

Khatami, R., Mountrakis, G., Stehman, S. V., 2016. A meta-analysis of remote sensing research on supervised pixel-based land-cover image classification processes: General guidelines for practitioners and future research. Remote Sensing of Environment 177, 89–100.

Khuluse-Makhanya, S., Dudeni-Tlhone, N., Holloway, J., Schmitz, P., Waldeck, L., Stein, A., Debba, P., Stylianides, T., Du Plessis, P., Cooper, A., Baloyi, E., 2016. The applicability of the South African Census 2011 data for evidence-based urban planning. Southern African Journal of Demography 17 (1), 67–132.

Kim, J. K., 2002. A note on approximate Bayesian bootstrap imputation. Biometrika 89 (2), 470–477.

Kloog, I., Koutrakis, P., Coull, B. A., Lee, H. J., Schwartz, J., 2011. Assessing temporally and spatially resolved $PM_{2.5}$ exposures for epidemiological studies using satellite aerosol optical depth measurements. Atmospheric Environment 45 (35), 6267–6275.

Kneen, M. A., Ojelede, M. E., Annegarn, H. J., 2015. Housing and population sprawl near tailings storage facilities in the Witwatersrand: 1952 to current. South African Journal of Science 111 (11), 1–9.

Kok, J. F., Parteli, E. J. R., Michaels, T. I., Bou Karam, D., 2012. The physics of wind-blown sand and dust. Reports on Progress in Physics 75 (10), 106901–106973.

Korcz, M., Fudala, J., Kliś, C., 2009. Estimation of wind blown dust emissions in Europe and its vicinity. Atmospheric Environment 43, 1410–1420.

Krzyzanowski, M., Cohen, A., 2008. Update of WHO air quality guidelines. Air Quality, Atmosphere & Health 1 (1), 7–13.

Larsen, L. C., Shah, M., 2016. A context-intensive approach to imputation of missing values in data sets from networks of environmental monitors. Journal of the Air & Waste Management Association 66 (1), 38–52.

Lawrence, M. G., 2005. The relationship between relative humidity and the dew point temperature in moist air. Bulletin of the American Meteorological Society 86 (2), 225–233.

Lazaridis, M., 2011. Human exposure and health risk from air pollutants. In: First principles of meteorology and air pollution. Vol. 19 of Environmental Pollution. Springer Netherlands, pp. 305–354.

Le, H. Q., Batterman, S. A., Wahl, R. L., 2007. Reproducibility and imputation of air toxics data. Journal of Environmental Monitoring 9 (12), 1358–1372.

le Roux, A., Khuluse, S., Naude, A. J. S., 2015. Creating a high resolution social vulnerability map in support of national decision makers in South Africa. In: Sluter, R., Madureira Cruz, C., Bernadete, C., de Menezes, L., Márcio, P. (Eds.), Cartography– Maps connecting the world: 27th International Cartographic Conference. Springer, Ch. 19, pp. 283–294.

Leitão, A. B., Ahern, J., 2002. Applying landscape ecological concepts and metrics in sustainable landscape planning. Landscape and Urban Planning 59, 65–93.

Li, C., Wang, J., Wang, L., Hu, L., Gong, P., 2014. Comparison of classification algorithms and training sample sizes in urban land classification with Landsat Thematic Mapper imagery. Remote Sensing 6 (2), 964–983.

Li, N., Hao, M., Phalen, R. F., Hinds, W. C., Nel, A. E., 2003. Particulate air pollutants and asthma: A paradigm for the role of oxidative stress in PM-induced adverse health effects. Clinical Immunology 109 (3), 250–265.

Liu, C., Koike, K., 2007. Extending multivariate space-time geostatistics for environmental data analysis. Mathematical Geology 39 (3), 289–305.

Liu, J., Li, W., Li, J., 2016. Quality screening for air quality monitoring data in China. Environmental Pollution, 1–4.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Vol. 1. University of California Press, Berkeley, California, pp. 281–297.
URL http://projecteuclid.org/euclid.bsmsp/1200512992

Maimon, O., Rokach, M. (Eds.), 2010. Data mining and knowledge discovery handbook, 2nd Edition. Springer.

Mansell, G., Lester, J., Pollack, A., September 2007. Studies of emissions from anthropogenic and natural dust sources in the Western United States. Air & Waste Management Association, 23–26.

Mathee, A., Harpham, T., Barnes, B., Swart, A., Naidoo, S., De Wet, T., Becker, P., 2009. Inequity in poverty: the emerging public health challenge in Johannesburg. Development Southern Africa 26 (5), 721–732.

Matooane, M., Oosthuizen, R., John, J., 2011. Self-reported hypertension in eMbalenhle, Mpumalanga, South Africa: findings from a vulnerability to air pollution assessment. Southern African Journal of Epidemiology and Infection 26 (4), 280–284.

Mensah, J. T., Adu, G., 2015. An empirical analysis of household energy choice in Ghana. Renewable and Sustainable Energy Reviews 51, 1402–1411.

Millar, G., Abel, T., Allen, J., Barn, P., Noullett, M., Spagnol, J., Jackson, P. L., 2010a. Evaluating human exposure to fine particulate matter (Part I): Measurements. Geography Compass 4 (4), 281–302.

Millar, G., Abel, T., Allen, J., Barn, P., Noullett, M., Spagnol, J., Jackson, P. L., 2010b. Evaluating human exposure to fine particulate matter (Part II): Modeling. Geography Compass 4 (7), 731–749.

Ministry for the Environment, NZ, 2009. Good practice guide for air quality monitoring and data management. Guidelines, New Zealand Ministry for the Environment, Wellington, New Zealand.

Monestiez, P., Dubroca, L., Bonnin, E., Durbec, J.-P., Guinet, C., 2005. Geostatistics Banff 2004. Springer Netherlands, Dordrecht, Ch. Comparison of model based geostatistical methods in ecology: Application to fin whale spatial distribution in Northwestern Mediterranean sea, pp. 777–786.

Montandon, L. M., Small, E. E., 2008. The impact of soil reflectance on the quantification of the green vegetation fraction from NDVI. Remote Sensing of Environment 112, 1835–1845.

Moran, J. L., Solomon, P. J., Peisach, A. R., Martin, J., 2007. New models for old questions: Generalized linear models for cost prediction. Journal of Evaluation in Clinical Practice 13 (3), 381–389.

Myint, S. W., Gober, P., Brazel, A., Grossman-Clarke, S., Weng, Q., 2011. Per-pixel vs. object-based classification of urban land cover extraction using high spatial resolution imagery. Remote Sensing of Environment 115 (5), 1145–1161.

Ncipha, X. G., 2011. Comparison of air pollution hotspots in the Highveld using airborne data. MSc dissertation, University of the Witwatersrand, South Africa.

Norman, R., Cairncross, E., Witi, J., Bradshaw, D., Group, S. A. C. R. A. C., 2007. Estimating the burden of diesease attributable to urban outdoor air pollution in South Africa in 2000. South African Medical Journal 97 (7), 782–790.

Oguntoke, O., Ojelede, M. E., Annegarn, H. J., 2013. Frequency of mine dust episodes and the influence of meteorological parameters on the Witwatersrand area, South Africa. International Journal of Atmospheric Sciences, 1–10.

Ojelede, M. E., Annegarn, H. J., Kneen, M. A., 2012. Evaluation of aeolian emissions from gold mine tailings on the Witwatersrand. Aeolian Research 3, 477–4869.

Ozer, P., September 2006. Dust in the wind and public health: Example from Mauritania. In: International Conference on Desertification, Migration, Health, Remediation and Local Governance. Royal Academy for Overseas Sciences, United Nations, Brussels, pp. 55–74.

Paciorek, C. J., Liu, Y., 2009. Limitations of remotely sensed aerosol as a spatial proxy for fine particulate matter. Environmental Health Perspectives 117 (6), 904–909.

Paruolo, P., Saisana, M., Saltelli, A., 2013. Ratings and rankings: voodoo or science? Journal of the Royal Statistical Society 176 (3), 609–634.

Parzen, M., Lipsitz, S. R., Fitzmaurice, G. M., 2005. A note on reducing the bias of the approximate Bayesian bootstrap imputation variance estimator. Biometrika 92 (4), 971–974.

Pope III, C. A., 2000. Epidemiology of fine particulate air pollution and human health: Biologic mechanisms and who's at risk? Environmental Health Perspectives 108 (4), 713–723.

Pope III, C. A., Dockery, D. W., 2006. Health effects of fine particulate air pollution: Lines that connect. Air and Waste Management Association 56, 709–742.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., Solenberger, P., 2001. A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Methodology 27 (1), 85–95.

Rees, D., Murray, J., 2007. Silica, silicosis and tuberculosis. In: Chan-Yeung, M. (Ed.), State of the Art Series: Occupational lung disease in high- and low-income countries. Vol. 11. International Union Against Tuberculosis and Lung Disease, pp. 474–484.

Reger, B., Otte, A., Waldhardt, R., 2007. Identifying patterns of land-cover change and their physical attributes in a marginal European landscape. Landscape and Urban Planning 81 (1–2), 104–113.

RSA Government, December 2009. National environmental management air quality Act 2004: National ambient air quality standards. Gazette 32816, Government of the Republic of South Africa, Pretoria, South Africa.

Rubin, D. B., 1996. Multiple imputation after 18+ years. Journal of the American Statistical Association 91 (434), 473–489.

Rubin, D. B., Schenker, N., 1986. Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. Journal of the American Statistical Association 81 (394), 366–374.

Saib, M.-S., Caudeville, J., Beauchamp, M., Carré, F., Ganry, O., Trugeon, A., Cicolella, A., 2015. Building spatial composite indicators to analyze environmental health inequalities on a regional scale. Environmental Health 14 (1), 68.

Saisana, M., Saltelli, A., Tarantola, S., 2005. Uncertainty and sensitivity analysis techniques as tools for the quality assessment of composite indicators. Journal of the Royal Statistical Society 168 (2), 307–323.

Sarma, Y., Jammalamadaka, S., 1993. Circular regression. Statistical Science and Data Analysis, 109–128.

Schafer, J. L., 1999. Multiple imputation: A primer. Statistical methods in medical research 8, 3–15.

Schelin, L., Sjöstedt-de Luna, S., 2010. Kriging prediction intervals based on semiparametric bootstrap. Mathematical Geosciences 42 (8), 985–1000.

Schenker, N., Taylor, J. M. G., 1996. Partially parametric techniques for multiple imputation. Computational Statistics and Data Analysis 22, 425–446.

Schlesinger, R. B., Kunzli, N., Hidy, G. M., Gotschi, T., Jerrett, M., 2006. The health relevance of ambient particulate matter characteristics: Coherence of toxicological and epidemiological inferences. Inhalation Toxicology 18 (2), 95–125.

Schwartz, J., Bellinger, D., Glass, T., 2011. Expanding the scope of environmental risk assessment to better include differential vulnerability and susceptibility. American Journal of Public Health 101 (S1), S88–S93.

Shaddick, G., Zidek, J. V., Liu, Y., 2016. Mitigating the effects of preferentially selected monitoring sites for environmental policy and health risk analysis. Spatial and Spatio-temporal Epidemiology 18, 44–52.

Sportisse, B., 2009. Fundamentals in air pollution: From processes to modelling. First Edition. Springer.

Statistics South Africa, 2012a. Census 2011: Metadata. Official statistics publication, Statistics South Africa, Pretoria, South Africa.

Statistics South Africa, 2012b. Census 2011: Statistical release. Official statistics publication, Statistics South Africa, Pretoria, South Africa.

Statistics South Africa, 2012c. Income and expenditure of households 2010/2011. Official Statistics Publication P0100, Statistics South Africa, Pretoria, South Africa.

Statistics South Africa, 2014. Mortality and causes of death in South Africa, 2014: Findings from death notifications. Official Statistics Publication P0309.3, Statistics South Africa, Pretoria, South Africa.

Statistics South Africa, 2015. Methodological report on rebasing of national poverty lines and development on pilot provincial poverty lines. Technical report 03-10-11, Statistics South Africa, Pretoria, South Africa.

Sterk, G., Stein, A., 1997. Mapping wind-blown mass transport by modeling variability in space and time. Soil Science Society of America Journal 61 (1), 232–239.

Strahler, A. H., Boschetti, L., Foody, G. M., Freidl, M. A., Hansen, M. C., Herold, M., Mayaux, P., Morisette, J. T., Stehman, S. V., Woodcock, C. E., 2006. Global land cover validation: Recommendations for evaluation and accuracy assessment of global land cover maps. Research report EUR 22156 EN, European Commission, Luxembourg.

Szpiro, A. A., Paciorek, C. J., 2013. Measurement error in two-stage analyses, with application to air pollution epidemiology. Environmentrics 24 (8), 501–517.

Tofallis, C., 2015. A better measure of relative prediction accuracy for model selection and model estimation. Journal of the Operational Research Society 66, 1352–1362.

Tortora, R. D., 1978. A note on sample size estimation for multinomial populations. The American Statistician 32 (3), 100–102.

van de Kassteele, J., 2006. Statistical air quality mapping. Doctoral Thesis, Wageningen University, The Netherlands.

Van Eetvelde, V., Antrop, M., 2009. A stepwise multi-scaled landscape typology and characterisation for trans-regional integration, applied on the federal state of Belgium. Landscape and Urban Planning 91 (3), 160–170.

Velders, G. J. M., Matthijsen, J., 2009. Meteorological variability in NO_2 and PM_{10} concentrations in the Netherlands and its relation with EU limit values. Atmospheric Environment 43, 3858–3866.

von Maltitz, M. J., June 2015. Extending the reach of sequential regression multiple imputation. Doctoral thesis, University of the Free State, South Africa.

Wallace, J. M., Hobbs, P. V., 2006. Atmospheric Science: An introductory survey, 2nd Edition. International Geophysics. Elsevier, USA.

Watson, J. G., Chow, J. C., 2000. Reconciling urban fugitive dust emissions inventory and ambient source contribution estimates: Summary of current knowledge and needed research. Technical research report, Desert Research Institute, NV, USA.

White, I. R., Royston, P., Wood, A. M., 2011. Multiple imputation using chained equations: Issues and guidance for practice. Statistics in Medicine 30 (4), 377–399.

WHO, 2006. Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide. Summary report of risk assessment, World Health Organization, Geneva, Switzerland.

WHO, 2007. Global surveillance, prevention and control of chronic respiratory diseases.

WHO, 2016. Health risk assessment of air pollution - general principles. Tech. rep., World Health Organization, Regional Office for Europe, Copenhagen, Denmark.

Wright, C., Diab, R., 2011. Air pollution and vulnerability: solving the puzzle of prioritization. Journal of Environmental Health 73 (6), 56–64.

Wright, C., Garland, R., Thambiran, T., Diab, R., May 2011. Air quality: A South African perspective. Environmental Scientist, 25–27.

Yoemans, K., Golder, P. A., 1982. The Guttman-Kaiser criterion as a predictor of the number of common factors. Journal of the Royal Statistical Society 31 (3), 221–229.

Young, L., Gotway, C., 2007. Linking spatial data from different sources: The effects of change of support. Stochastic Environmental Research and Risk Assessment 21 (5), 589–600.

Zhou, Y., Li, N., Wu, W., Wu, J., Shi, P., 2014. Local spatial and temporal factors influencing population and societal vulnerability to natural disasters. Risk analysis 34 (4), 614–639.

Zidek, J. V., Shaddick, G., Taylor, C. G., 2014. Reducing estimation bias in adaptively changing monitoring networks with preferential site selection. The Annals of Applied Statistics 8 (3), 1640–1670.

Zwack, L. M., Paciorek, C. J., Spengler, J. D., Levy, J. I., 2011. Modeling spatial patterns of traffic-related air pollutants in complex urban terrain. Environmental Health Perspectives 119 (6), 852–859.

Biography

Sibusisiwe Khuluse was born on 19 August 1985 in Durban, South Africa. She obtained a BSc degree in Applied Mathematics and Statistics and a BSc Honours degree in Statistics at the University of KwaZulu-Natal, Durban, South Africa in 2007. In 2007 she was employed as a Candidate Researcher in Statistics at the Council for Scientific and Industrial Research (CSIR), in Pretoria, South Africa. In 2009 she spent three months at the Faculty of Geo-Information Science and Earth Observation (ITC) at the University of Twente for her MSc research supported by the Tata Africa Scholarship, ITC and CSIR. She graduated with a MSc in Mathematical Statistics degree from the University of Witwatersrand, Johannesburg, in 2011 whilst in full-time employment at the CSIR. She was awarded the Harvard South Africa Fellowship in 2009 and went to Harvard University Graduate School of the Arts and Sciences as a visiting student from August 2010 until May 2011. In July 2011 she returned to ITC for her PhD research supported by the Nuffic Netherlands Fellowship Programme and the CSIR. Her PhD research focussed on the development of a statistical framework for air quality risk mapping, contributing specific methods to improve the quality of the data including the integration of ancillary data from disparate sources. This thesis is the output of her research.

Author's publication

Submitted ISI journal articles

Khuluse-Makhanya S., Stein A., Debba P. Exploring housing informality and domestic solid biomass fuel use as predictors of the PM_{10} exceedance rate through kriging and generalized linear geostatistical models. *South African Geography Journal*.

Khuluse-Makhanya S., Stein A., Debba P. Multiple imputation of missing air quality data using bootstrap methods. *Environmental and Ecological Statistics*.

Khuluse-Makhanya S., Stein A., Debba P., Dudeni-Tlhone N., Ngidi M. A statistical approach to air quality mapping and the risk of exposure to excessive particulate matter pollution. *Spatial Statistics*.

ISI journal article under review

Khuluse-Makhanya S., Stein A., Breytenbach A., Gxumisa A., Dudeni-Tlhone N., Debba P. Ensemble classification for identifying neighbourhood sources of fugitive dust and associations with observed PM_{10}. *Atmospheric Environment* (major revision).

Published journal articles

Khuluse-Makhanya, S., Dudeni-Tlhone N., Holloway J., Schmitz P., Waldeck L., Stein A., Debba P., Stylianides T., Du Plessis P., Cooper A., Baloyi E., 2016. The applicability of the South African Census 2011 data for evidence-based urban planning. *Southern African Journal of Demography* 17(1), 67–132.

Article published in conference book

le Roux A., *Khuluse S.*, Naude A.J.S., 2015. Creating a high resolution social vulnerability map in support of national decision makers in South Africa. Cartography– Maps Connecting the World: 27th International Cartographic Conference 2015. Edited by Sluter R., Madureira Cruz C., Bernadete C., de Menezes L., Mrcio P. Springer International Publishing. ISBN 9783319177373, Chapter 19, pp. 283–294.

Conference presentations

Khuluse-Makhanya S., Stein A., Debba P., 2016. Identifying local emission sources of particulate matter to improve geostatistical mapping of PM_{10}. Oral Presentation at the South African Statistical Association Conference, Cape Town.

Khuluse-Makhanya S., Stein A., Debba P., 2015. Sequential regression imputation of air quality data. Oral presentation at the South African Statistical Association Conference, Pretoria.

Khuluse S., Stein A., Debba P., 2014. A multiple imputation approach for missing air quality measurements. Oral presentation at the South African Statistical Association Conference, Grahamstown.

Khuluse S., Stein A., 2013. Mapping the annual exceedance frequencies of the PM_{10} air quality standard: Comparing kriging to a generalized linear spatial model. Poster presentation at the South African Statistical Association Conference, Polokwane.

ITC Dissertation List

http://www.itc.nl/Pub/research_programme/Research-review-and-output/PhD-Graduates