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

Sibusisiwe Khuluse-Makhanya[a,b,*], Alfred Stein[b], André Breytenbach[a], Athi Gxumisa[a], Nontembeko Dudeni-Tlhone[a], Pravesh Debba[a,c]

[a] *CSIR Built Environment, PO Box 395, Pretoria, South Africa, 0001*
[b] *Faculty of Geo-Information and Earth Sciences (ITC), University of Twente, PO Box 217, 7500 AE Enschede, Netherlands*
[c] *School of Statistics and Actuarial Sciences, University of Witwatersrand, Johannesburg, South Africa*

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

*Corresponding author
Email address: smakhanya@csir.co.za (Sibusisiwe Khuluse-Makhanya)

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 [1, 2, 3, 4]. Chemical processes refer to the formation of particles through condensation of gases produced by incomplete combustion as industrial, vehicle and biomass burning fumes [5]. 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. [6, 7]. 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 [8]. 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 [9, 10, 8]. 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 [11]. There are various approaches for extracting land cover classes from imagery and a common technique is pixel-based supervised maximum likelihood (ML) classification [12, 13]. [13] 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 [14]. In a recent study by [15], 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 [16].

Land-use regression (LUR) models are a popular tool in air pollution exposure assessments [17, 18, 3]. 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 [1, 3]. 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 [19]. Given no specific grouping factor, clusters can be identified by applying a k-means clustering technique [20] prior to regression modelling. The strengths of the k-means technique include the capability of handling large data sets [21, 22] and ease of adaptation into advanced clustering procedures aimed at superior data handling efficiency [23]. 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 [24] and in creating a multi-scale trans-border oriented landscape typology for Belgium in [21]. 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 paper is presented as follows: Section 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; the results are in Section 4; the discussion in Section 5 and the concluding remarks in Section 6.

2. Materials

2.1. The study area

Air quality monitoring stations (Figure 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 [25, 26]. 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 [27]. 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 [28]. 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 [29, 30], 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.

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 [31], 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 [6]. Therefore, we focus on an area that is limited to a 4 km radius from the location of an air quality station (Figure 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 [6, 32]. [32] 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 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 2, Table 1).

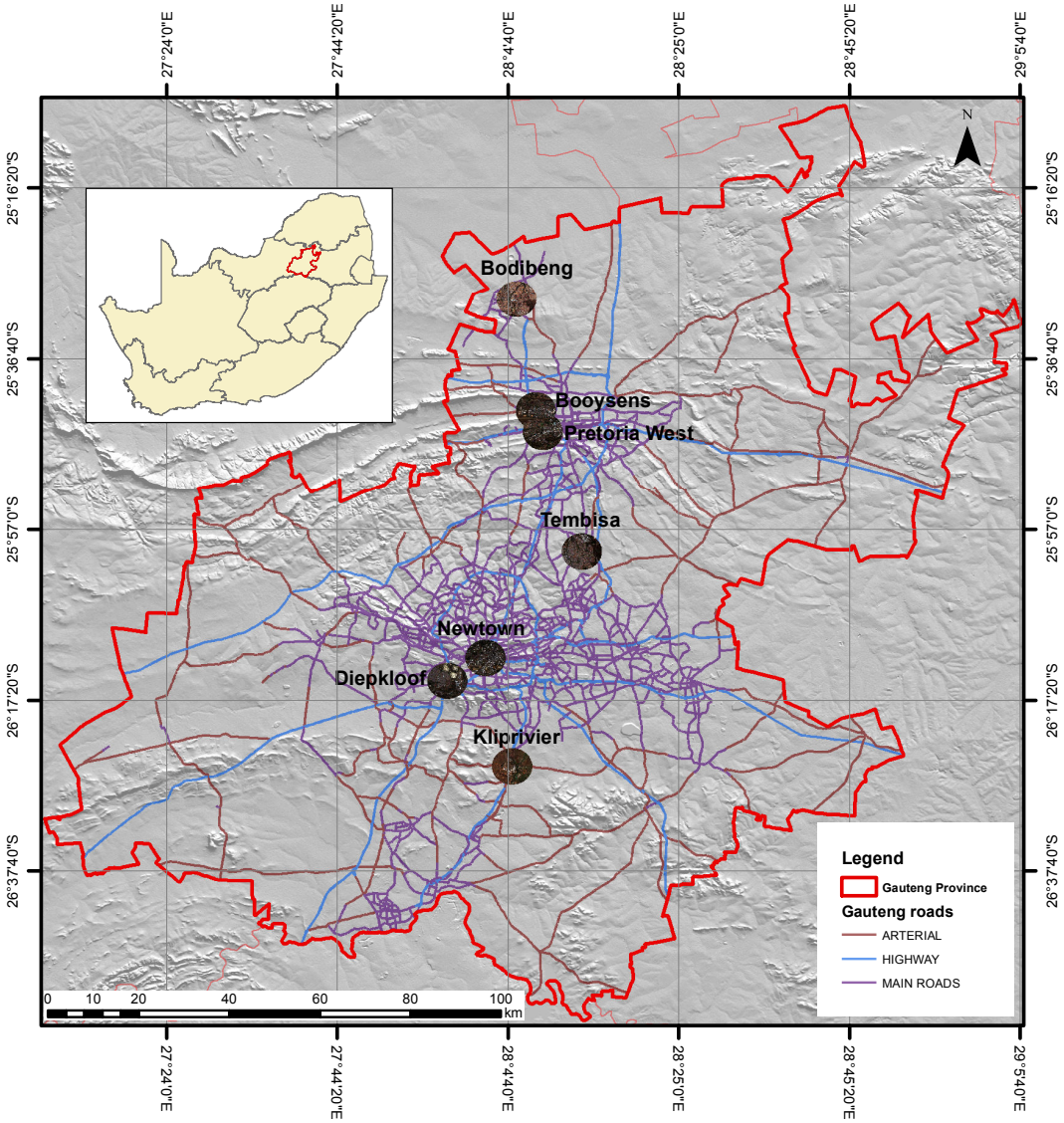
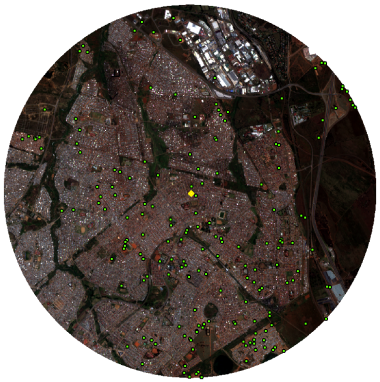
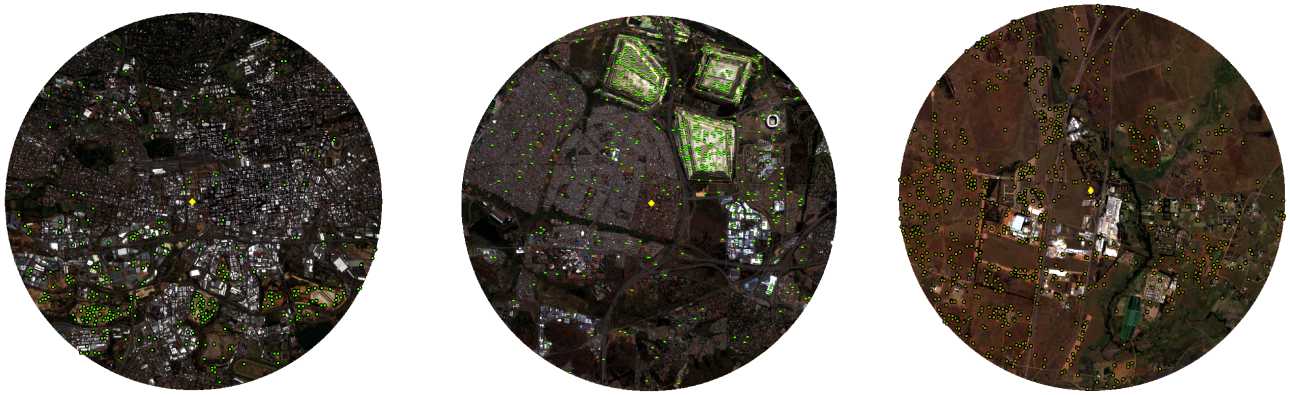


Figure 1: All 23 air quality monitoring stations are located within Gauteng's provincial boundaries, but for classifier development, the seven circular areas are shown with the Bodibeng, Booysens, Pretoria West, Tembisa, Newtown, Diepkloof and Kliprivier air quality stations located at their respective centres

(a) Bodibeng (b) Booysens (c) PTA West



(d) Tembisa



(e) Newtown (f) Diepkloof (g) Kliprivier

Figure 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

3. Methods

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 [12, 33]. 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:

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

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 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

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

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. [34] 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, 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 [14]. 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

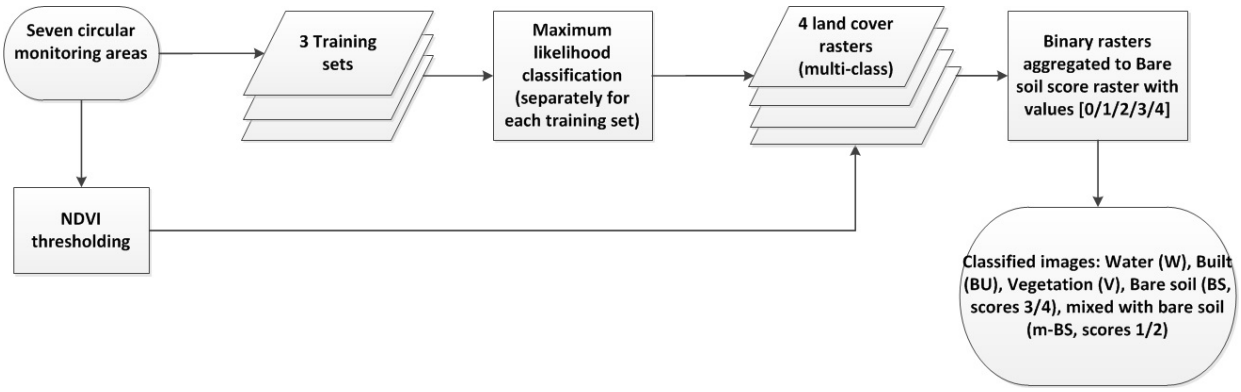


Figure 3: Ensemble maximum likelihood classification with focus on improved accuracy for the bare soil class

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, 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 [34]. 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. [35] 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.

3.2. 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 [36]. An important consideration is the determination of sample size. For a multinomial population [37], the marginal distribution for each class $i = 1, 2, \ldots, m$ is binomial with parameter p_i. Sample size is n such that,

$$n = B \times \frac{p(1-p)}{e^2} \tag{3}$$

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

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

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 3.1 focuses 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 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 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 [38].

Table 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.)** $N_{BS(i)}/N_{BS(all)}$	**BS validation sample size**
Bodibeng (Dom, Tshwane)	Townships, unpaved roads, yards and sidewalks, bare fields East of station	93 537	0.15	398
Booysens (Bg, Tshwane)	Pretoria CBD south of station, mining, agric fields north of station	56 129	0.09	238
PTA West (Ind, Tshwane)	Pretoria CBD east of station, mining south of station	34 466	0.05	147
Diepkloof (Traf, Vaal Triangle)	Mine tailing storage facilities N-NE of station, township	130 050	0.20	553
Kliprivier (Ind, Vaal Triangle)	Extensive agric fields, low density residential, townships NE of station	204 724	0.32	872
Newtown (Traf, JHB)	JHB CBD, mining south of station, suburbs N of station	73 534	0.12	313
Tembisa (Dom, Ekurhuleni)	Township, sport fields with bare areas, unpaved sidewalks and roads	42 188	0.07	180
Total		634 628	1	2 701

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 [39, 40]. 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 [40]. For the Witwatersrand area which is part of our study area, [41] 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 [19] is:

$$y_i = \alpha_{j[i]} + \beta x_i + \epsilon_i \tag{5}$$

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 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 [40] as:

$$\begin{aligned} \mathrm{E_{dust}}(10^9\ \mathrm{g\,m^{-2}}) \quad = \quad & \{\text{land area}, \mathrm{m^2}\} \times \\ & \{[\text{spike emission rate, g m}^{-2}]+ \\ & [(\text{duration of erosion event, h}) \times (\text{emission factor, g m}^{-2}\ \mathrm{h^{-1}})]\} \end{aligned} \tag{6}$$

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 [40] 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 2.1. Values by [40] 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 [32]. It is expressed as:

$$\beta = \log\left(1 + \frac{\sum_i a_i \times n_{CL(i)}}{\sum_i n_{CL(i)}}\right) \tag{7}$$

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 [32]. 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 [42]. 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,

$$O = \sum_{i=1}^{k} \sum_{x_j \in V_i} \parallel x_j - m_i \parallel^2 \tag{8}$$

with our choice of distance function $\| x_j - m_i \|$ 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 9 describes the balance between land cover classes in each circular neighbourhood [43]. That is,

$$-\frac{\sum_i p_i \ln p_i}{\ln N} \tag{9}$$

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.

4. Results

4.1. Ensemble land cover classification results

Figure 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 2 and Figure 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 bare soil classification accuracy, classification quality is also assessed based on an analysis of confidence. Classification confidence in Figure 5 is "a measure of confidence that quantifies how closely a classified observation matches the exemplars of the training set" [44]. Fourteen classification confidence levels are defined relative to predefined discrete points

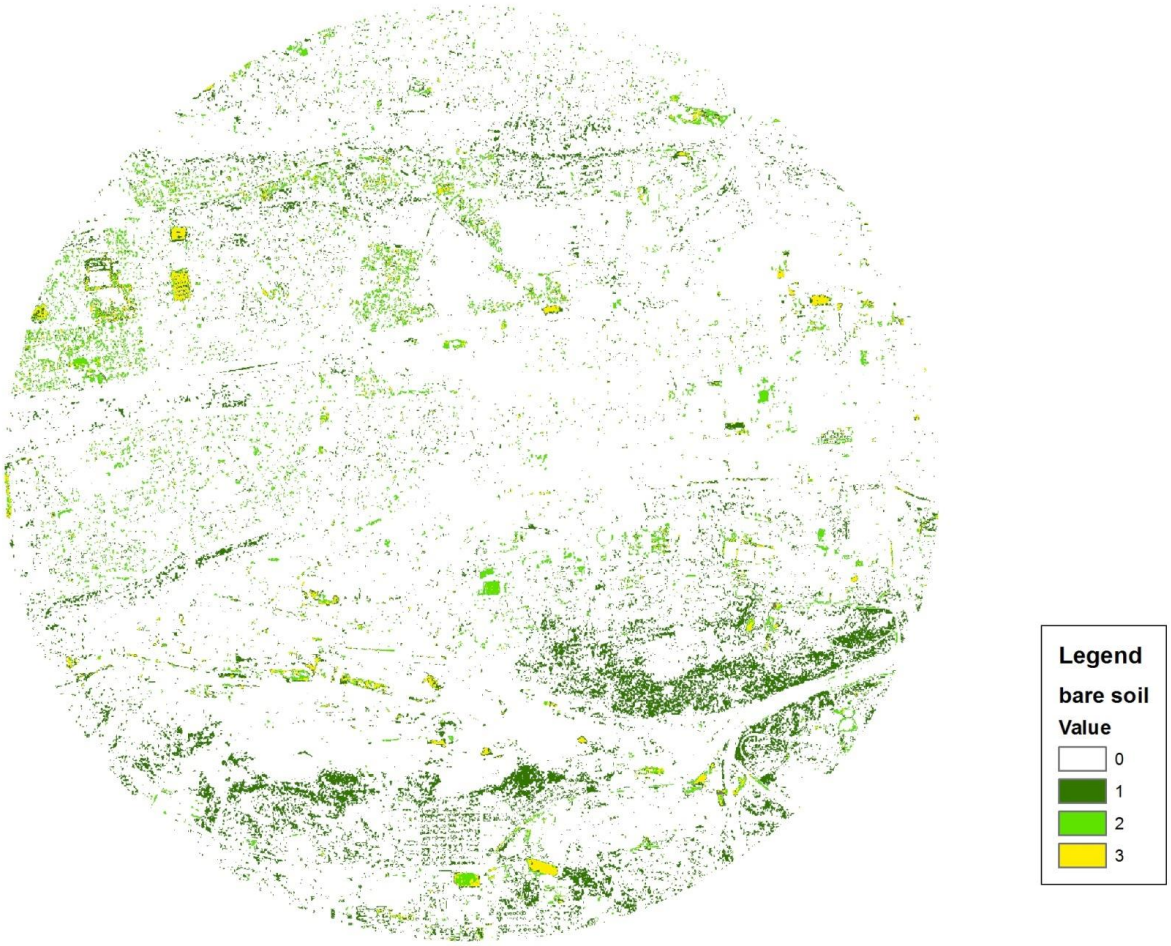


Figure 4: Map for bare soil derived by adding binary rasters obtained from each ML classification run (Pretoria West AOI)

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 pixels that are furthest from the mean vector of the input signature and therefore 99.5% (0.995 rejection fraction) of such pixels are at risk of misclassification. Apart from Diepkloof in Figure 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 6 show high prevalence of the built class in Tembisa, PTA West and

Table 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 $\hat{p}$

Newtown. This is confirmed by coverage statistics in Table 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 AOIs with the highest proportions of vegetation pixels are Kliprivier and Diepkloof. The least built-up AOI is Kliprivier. When the other 16 neighbourhoods were classified using the ensemble classifier, there were areas that had higher density of buildings, the least (< 2 km^2) bare soil coverage and very high average PM_{10} concentrations corresponding to days with wind speeds in excess of 4 m s^{-1}. Diepsloot noted in Figure 7(a) was one of those neighbourhoods.

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 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 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 4 is challenging for the classifier because of a higher risk of confusion with either

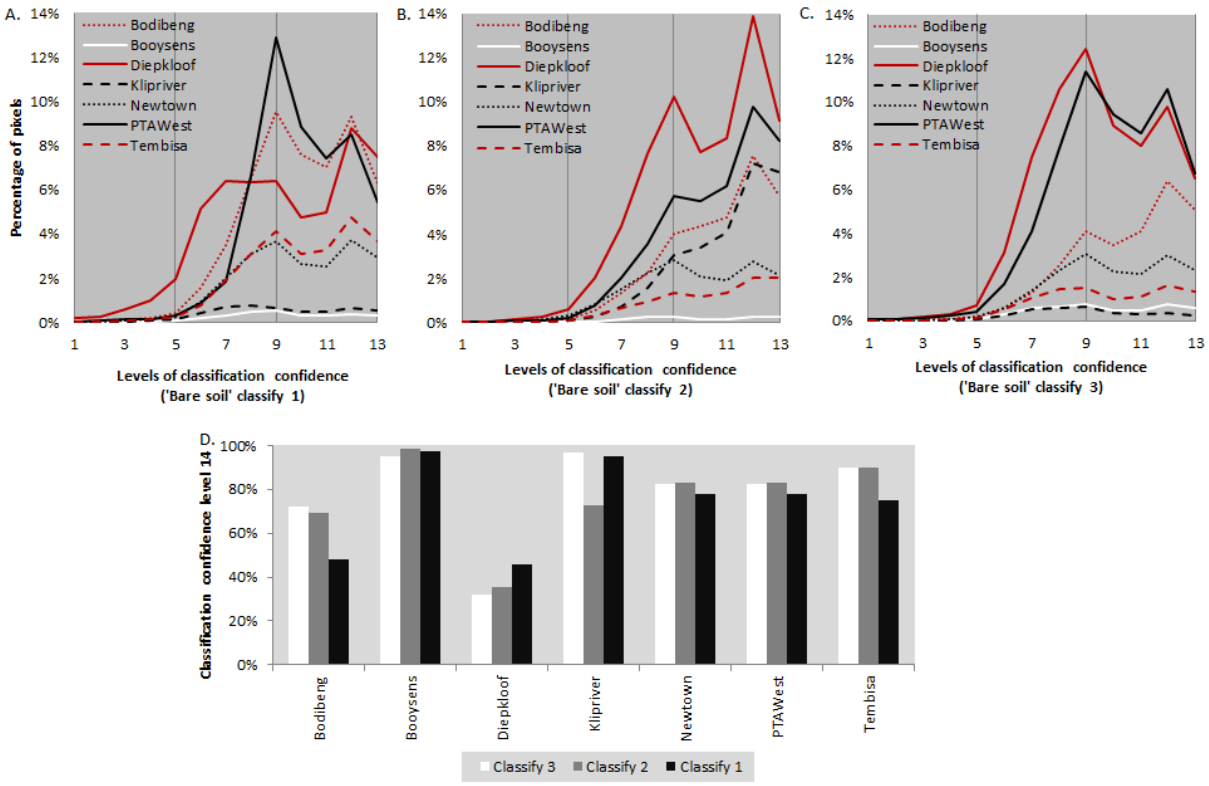


Figure 5: Levels of confidence for the three ML classification iterations for seven AOIs used in training the classifier. Graphics A-C show coverage (percentage of pixels) for each AOI corresponding to confidence levels 1-13 for each iteration; D. Shows the percentage of pixels per AOI for which there is least confidence of correct classification for the three iterations.

bare soil or mixed bare soil classes.

4.2. Relating land cover with PM_{10} concentrations

Figure 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 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 7(b–c), vegetation and built-up coverage when considered individually are not predictive of ambient PM_{10} concentrations.

Table 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 of the seven circular areas

AOI	Cluster	Land cover (in km^2)[a]					PM_{10}[b]	Wind sp[c]	Wind dir[d]	Shannon[e] index	Janssen[e] β index	E_{dust} index[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	0.72	0.98	8.4
Booysens	4	28.3	13.6	2.0	20.1	0.0	35	6.1	NNW	0.72	0.84	10.9
Diepkloof	3	12.7	27.2	4.7	19.3	0.2	56	8.4	NNE	0.78	0.88	64.9
Kliprivier	3	2.6	42.1	7.4	11.9	0.2	55	8.1	NNW	0.61	0.41	35.6
Newtown	4	41.0	8.1	2.0	12.7	0.2	26	4.6	NNE	0.62	0.98	4.5
PTA West	4	41.7	7.0	1.2	14.0	0.1	49	4.7	ENE	0.58	0.94	14.8
Tembisa	4	42.5	3.7	1.5	16.3	0.0	84	4.7	WNW	0.55	0.98	14.8

[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

[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 [32]; E_{dust} is a proxy for wind blown dust emissions of PM_{10} based on bare soil coverage

From Figure 7 the statistical relationship between observed average 'windy day' PM_{10} concentrations and areal coverage of the different land cover types, especially bare soil, is not straight-forward linear. Close examination of the scatter plots and further exploratory analysis revealed some grouping effects. Therefore, instead of directly relating PM_{10} with bare soil coverage in circular areas around air quality stations, cluster analysis was performed to formally identify homogenous groups of air quality stations based on the multivariate space consisting of wind-related average PM_{10} and different areal coverage combinations for bare soil, mixed bare soil, built-up, vegetation and water classes. Land cover composition clusters can therefore be related to average 'windy day' PM_{10} concentrations. Figure 8(a) 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 8(a) 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 8(d)).

We observe two main outlier clusters in Figure 8(b), namely Cluster 2 and 6, with the latter represent-

Table 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%

ing 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 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 8(b) and 8(d)). The Shannon evenness indicator describes the diversity of land cover types within each circular neighbourhood and Figure 8(c) 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 8(c), Table 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 8(b)).

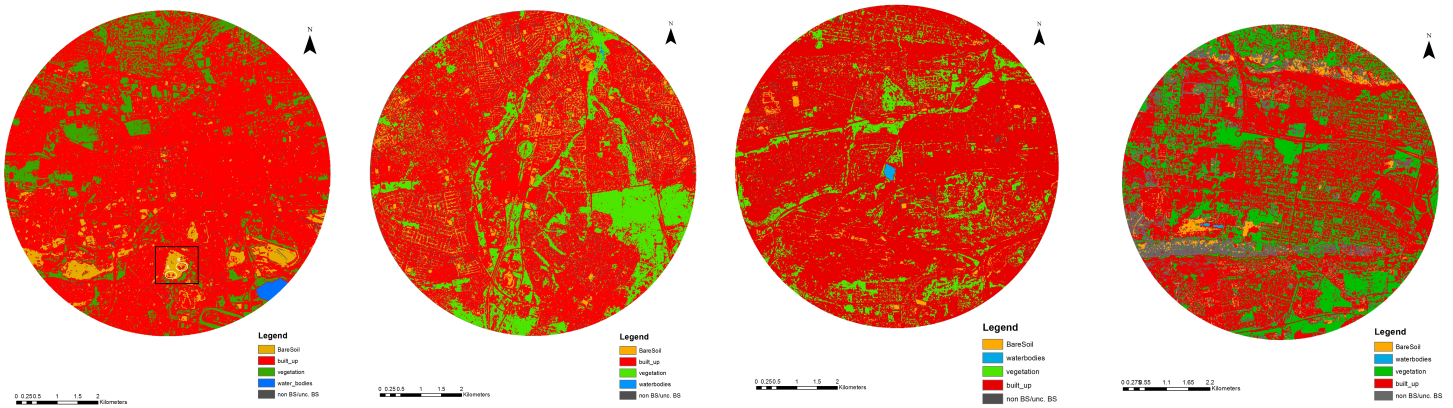
Land cover characteristics as represented by the six clusters significantly (p-value $< 1e-3$) explain more than 70% of the variability in average PM_{10} concentrations associated with wind speeds in excess of 4 m s^{-1} (Table 5). Further, at 10% level of significance, all six clusters are significant 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 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

Table 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

Figure 9(b), 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 8(b). Similarly the low levels of PM_{10} emission from bare soil indicated by E_{dust} for Clusters 1 and 6 in Figure 9(a) are in contrast to the high observed PM_{10} values for these clusters in Figure 8(b).



(a) Newtown (b) Bodibeng (c) Pretoria West (d) Booysens

(e) Tembisa (f) Diepkloof (g) Kliprivier

Figure 6: Preliminary ensemble ML land cover classification output for the seven AOIs

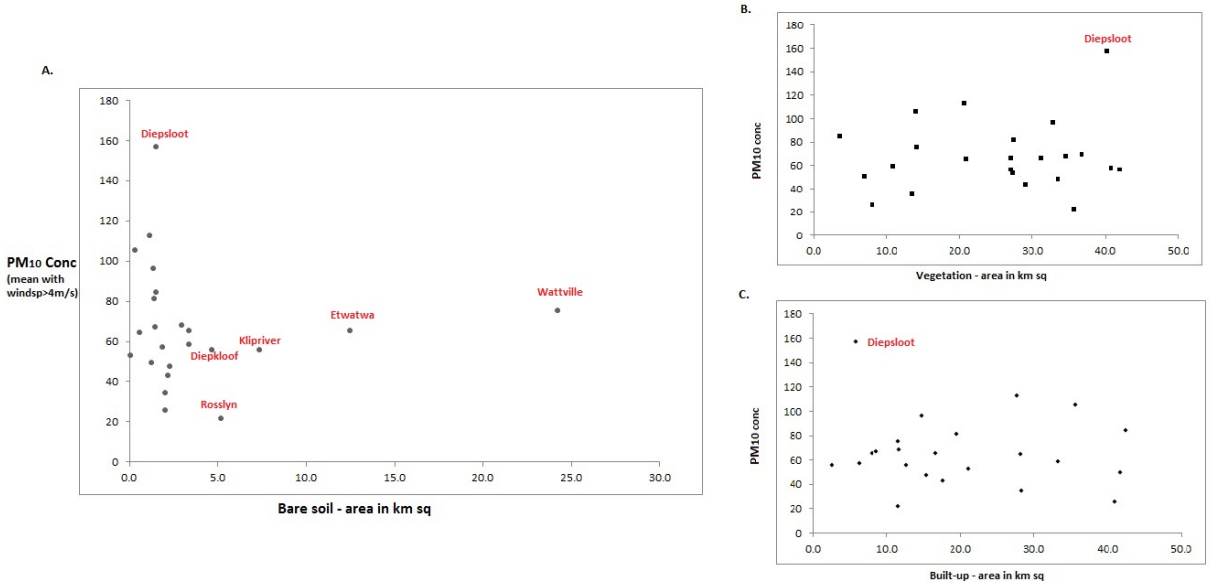
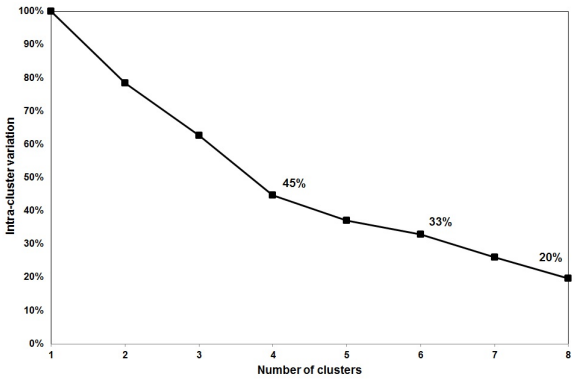
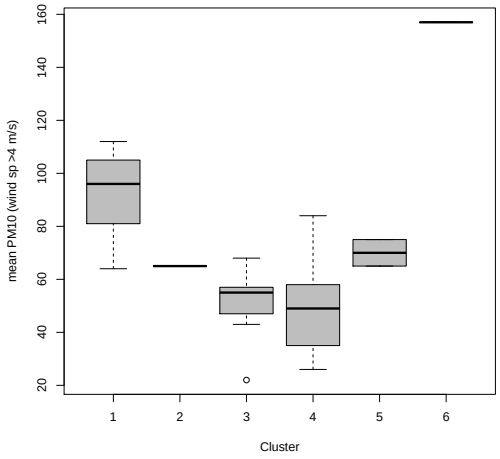


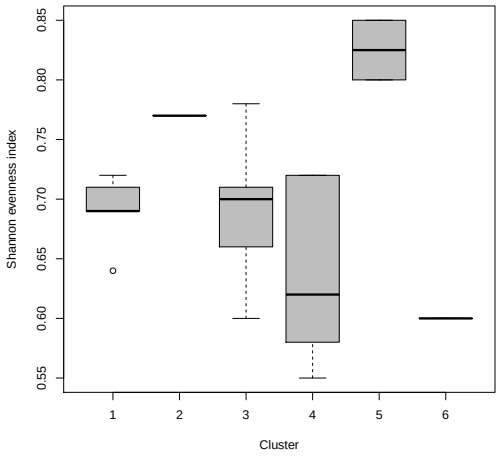
Figure 7: Exploratory assessment of the statistical relationship between vegetation, built-up and bare soil coverage and ambient PM_{10}



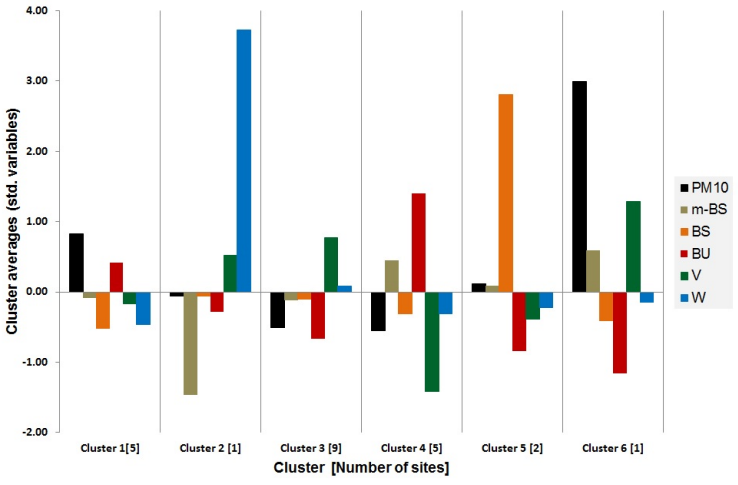
(a) Elbow criterion for determining the number of clusters



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

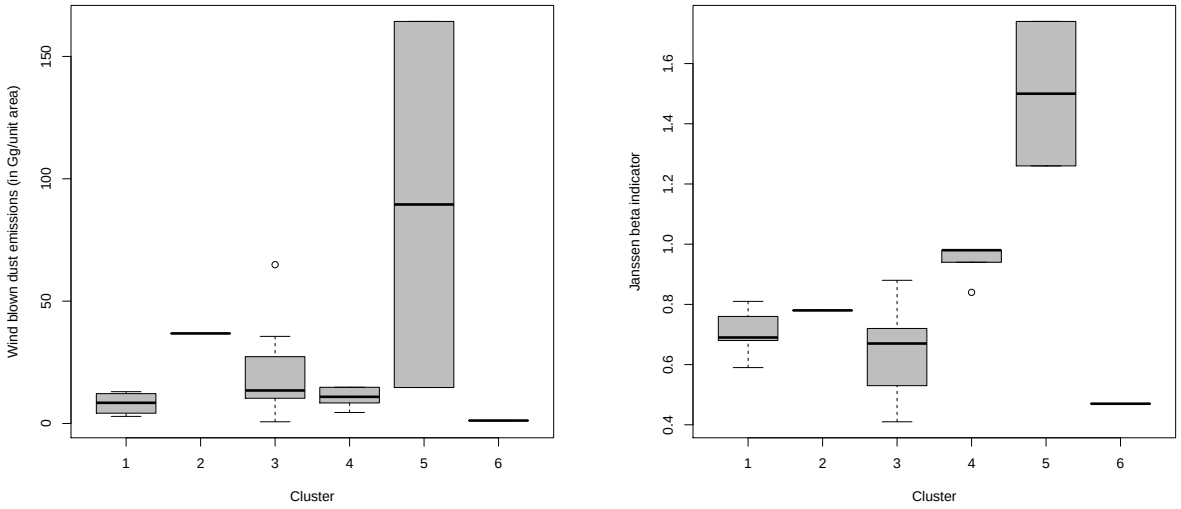


(c) The distribution of the Shannon evenness index for the six land cover clusters



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

Figure 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 9: Proxy variables considered as fixed effects in the varying intercepts model that relates land cover characteristics to observed PM_{10} values

5. Discussion

(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 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)

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} [32]. 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 [11]. 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 10(a)) and bare soil areas in a waste treatment plant (Figure 10(c)). 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 [8]. 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 [8]. 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 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 [41]. 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 [40, 10, 41]. 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 [40]. [32] 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} would enable 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) [45]. 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 [46]. 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.

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

This research was funded by the NUFFIC Netherlands Fellowship Program (Grant number CF7517/2011) and the Council for Scientific and Industrial Research of South Africa (CSIR). The South African National Space Agency is acknowledged for providing SPOT 6 images and the South African Weather Service for providing air quality data from the South African Air Quality Information System.

References

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

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

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

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

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

[6] J. G. Watson, J. C. Chow, 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 (2000).

[7] E. Athanasopoulou, M. Tombrou, A. G. Russell, A. Karanasiou, K. Eleftheriadis, A. Dandou, 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) (2010) 1–21.

[8] M. A. Kneen, M. E. Ojelede, H. J. Annegarn, Housing and population sprawl near tailings storage facilities in the Witwatersrand: 1952 to current, South African Journal of Science 111 (11) (2015) 1–9. doi:10.17159/sajs.2015/20140186.

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

[10] M. E. Ojelede, H. J. Annegarn, M. A. Kneen, Evaluation of aeolian emissions from gold mine tailings on the Witwatersrand, Aeolian Research 3 (2012) 477–4869. doi:10.1016/j.aeolia.2011.03.010.

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

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

[13] R. Khatami, G. Mountrakis, S. V. Stehman, 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 (2016) 89–100. doi:{10.1016/j.rse.2016.02.028}.

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

[15] B. Banerjee, F. Bovolo, A. Bhattacharya, L. Bruzzone, S. Chaudhuri, B. Mohan, A new self-training-based unsupervised satellite image classification technique using cluster ensemble strategy, IEEE Geoscience and Remote Sensing Letters 12 (4) (2015) 741–745. doi:{10.1109/LGRS.2014.2360833}.

[16] C. Li, J. Wang, L. Wang, L. Hu, P. Gong, Comparison of classification algorithms and training sample sizes in urban land classification with Landsat Thematic Mapper imagery, Remote Sensing 6 (2) (2014) 964–983. doi:{10.3390/Rs6020964}.

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

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

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

[20] J. MacQueen, 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, 1967, pp. 281–297.
URL http://projecteuclid.org/euclid.bsmsp/1200512992

[21] V. Van Eetvelde, M. Antrop, 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) (2009) 160–170.

[22] N. Dudeni-Tlhone, J. Holloway, S. Khuluse-Makhanya, R. Koen, 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, 2013, pp. 23–30.

[23] Y. Chen, P. Gong, Clustering based on eigenspace transformation– CBEST for efficient classification, Journal of Photogrammetry and Remote Sensing 83 (2013) 64–80. doi:{10.1016/j.isprsjprs.2013.06.003}.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[40] M. Korcz, J. Fudala, C. Kliś, Estimation of wind blown dust emissions in Europe and its vicinity, Atmospheric Environment 43 (2009) 1410–1420. doi:{10.1016/j.atmosenv.2008.05.027}.

[41] O. Oguntoke, M. E. Ojelede, H. J. Annegarn, Frequency of mine dust episodes and the influence of meteorological parameters on the Witwatersrand area, South Africa, International Journal of Atmospheric Sciences (2013) 1–10doi:10.1155/2013/128463.

[42] EEA, CORINE land cover project, Report Part II - Nomenclature, European Environment Agency (1995).

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

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

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

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