# Mapping impervious urban surfaces using regression analysis on synthetic hyperspectral EnMAP data

*Geography Department, Humboldt-Universität, Berlin, Germany*

[redacted]

# Mapping impervious urban surfaces using regression analysis on synthetic hyperspectral EnMAP data

With the envisaged launch of the EnMAP satellite in 2018 high accuracy hyperspectral image data of the earth's surface will be provided creating new opportunities for many research fields. Especially for the exploration of challenging environments like urban areas with their heterogeneous composition this data is expected to be of great value. With regard to increasing urbanization all over the world the monitoring and screening of impervious materials in urban areas are of particularly high importance. In this study we aim at quantifying fractions of impervious surfaces using a simulated EnMAP scene of Berlin, Germany, with a spatial resolution of 30 m. To extract as much information as possible from every pixel we used a multiple linear regression approach to identify possible fractions of impervious surface materials on each pixel. The basis for training our linear regression models are synthetically mixed spectra from 75 impervious and non-impervious surfaces. A semi-manual principal component analysis was performed on training and image data providing respectively three comparable principal components that explain over 99% of the variances of both datasets. Our multiple linear regression aims at the development on one model for each spectral training mixture and the application of that model on each image pixel. It turns out that there are numerous models that predict a pixel in an adequate way. On an area clip, each pixel has been correctly modeled (imperviousness of 0 to 100%) from 213 to 1495 different models. Maximum and minimum fractions modeled do not allow conclusions about actual rates of imperviousness neither. A result validation method would be designed after the enhancement of our central analysis method which could include a different mixture or modeling approach.

Keywords: EnMAP, hyperspectral data, imperviousness, principle component analysis, PCA, multiple linear regression

## 1. Introduction

According to the UN more than half of the world's population lives in urban areas today. Following current rates of urbanization cities are expected to be a home to more than two thirds of mankind in 2050 [14]. Therefore a more detailed analysis of urban structures and surface characteristics is needed for a deeper understanding of the functioning of this special kind of human-environment-interaction. As the impervious urban surface is emerging more and more as a major indicator of the environmental quality [15] a special emphasis is laid on its mapping. However, research on this topic is often connected to field surveys that are both time consuming and too expensive for essential regular updates. Even if the growing field of remote sensing techniques offers a chance to provide information on the properties of urban surfaces instead [7] urban areas are highly challenging for remote sensing data analysis. Compared to natural environments, the high spectral diversity of different materials in combination with restricted spatial resolution of sensors leads to the need to explore sub-pixel information [8].

In this study we aimed at quantifying fractions of impervious surfaces using multiple linear regression analyses on synthetically generated EnMAP training data. A principle component analysis (PCA) is used to reduce data complexity as a PCA is legitimately the most popular dimension reduction technique for hyperspectral data [4,5].

In contrast to us using a combination of a linear regression and a PCA there are different methods of spectral mixture analysis (SMA) for subpixel information. Besides regular linear SMAs based on linear equation systems and alternatively fixed or variable amounts of endmembers and image or pixelwise approaches, the multiple endmember SMA (MESMA) is a very common technique. It is a pixelwise linear unmixing approach based on a pool of known endmembers and is, e.g., used in [11] and [7]. Also appealing are support vector machines (SVM) and support vector regressions (SVR) due to their capability to work even with limited training data [6,7,8]. As these methods are broadly used many researchers have published enhancements on SVM especially focusing on using SVM on simulated EnMAP data [1].

## 2. Study area and data

Our study region consists of an approximate area of 6 km in east-west and 23 km in north-south direction comprising the south-western part of the German capital, Berlin, and adjacent rural areas of Brandenburg (Fig. 1). Our project is based on three data sets resulting from [7].

Resulting maps for imperviousness will be deduced from synthetic satellite data of the German hyperspectral Environmental Mapping and Analysis Program (EnMAP) presumably launched in 2018. Source imagery for data simulation originates from hyperspectral airborne iamgery taken 2009 by the HyMap sensor that offers 128 spectral bands from 440 nm to 2500 nm [2]. The simulated EnMAP image features 111 bands and a spatial resolution 30 m x 30 m.

Model training is based on a set of training data including 2106 artificial mixtures of 75 (39 impervious and 54 non-impervious surfaces including intra-class mixtures) different surface spectra, respectively mixed in steps of 20%. Table 1 presents a list of all materials predefined in [9] and identified after an adapted classification scheme by [12]. That list of materials underlines the broad range of urban surface types, which is, in combination with small-scale structures, a challenge for remote sensing.

Validation data consists in a block wise high resolution reference map by the Berlin Urban and Environmental Information System and the results of [7] and will find its way into the study depending on our model outcomes.

## 3. Methods

Our project is realized using open source software provided by [13] and complementary packages for statistical computing specified in the code annex.

The first step of our study consists in the reduction of data complexity using a principal component analysis (PCA). Since the number of spectral bands corresponds to the number of predictor variables in our linear regression models, we will reduce the image's 111 bands to a manageable number of bands that still explain a considerable part of the image variance (Fig. 2).

The central idea of the following multiple linear regression approach is that one respective multiple linear model can be established for any of the 2106 (39 * 54) spectral mixtures of surfaces. Any of those models is able to predict surface fractions for any pixel of our imagery. The challenge is to find out which of the 2106 models predicts the surface material fraction best on a pixelwise basis. Knowing the index of the respective best model we can deduce information about each pixel's surface features.

### *3.1 Reduction of data complexity (Principal Component Analysis)*

Hyperspectral sensors offer a high dimensionality in data. This can be advantageous for analyzing complex processes on the earth's surface allowing to benefit from a high level of spectral detail. On the other hand, that level of detail can massively increase computation time considering that in addition, there is a redundancy in data through high band correlations. A principle component analysis transforms correlated data into a new set of uncorrelated synthetic components explaining the same information. That transformation is based on a transformation matrix built from the eigenvectors of a correlation matrix of all input variables in order to minimize correlations between the different variables. The transformation matrix contains rotation values for each variable and each component. Components are calculated by the sum of multiplying each rotation factor with each original data [10].

In our case, the original data consists of 111 bands resulting in 111 new components. We first transform our training data assuming that it is the perfect data for our analysis, because it has been synthetically created. We then extract the rotation matrix of that principal component analysis in order to transform our simulated EnMAP image. The goal is that both datasets have comparable components. This approach is possible because the endmember spectra of our training data originate from the EnMAP image [7].

### *3.2 Multiple linear regression development*

A multiple linear regression model is a regression model with more than one predictor variable. It is characterized by an intercept and coefficients for each predictor variable. In contrast to a simple linear regression, the choice of predictor variables and their interactions is crucial. Higher-order terms might also be included. Linear regressions underlay several conditions, such as constant variances over time or constant error probability during measurements. The challenge is to find the best model fit with the least possible complexity. Performing a multiple linear regression model in R, parameters are estimated by the method of least squares [3,10].

Since our study is based on remote sensing data, the distribution of variance and measurement error probability is negligible as a factor or constraint. Our approach aims at determining the best model for each pixel amongst all models of spectral training data mixtures, whereas our selection of first components is our set of predictor variables and the fraction of the first material in our training data is our response variable. Creating models out of our training data results in perfect linear models with no residuals, since our training data is synthetic and linearly mixed. The best model is the one that is able to predict fractions as our response variable between 0% and 100%, i.e. that all predictor variables are able to locate the fraction variable on a more-dimensional hyperplane with a plausible fraction of impervious surface.

### *3.3 Fraction analysis*

In a final step, we will analyze how many models can predict fractions of imperviousness and if there is one best model. We will map the number of matching models as well as fraction statistics of that number of models. In case that there is one model that predicts the fractions best on a pixel basis we can make a statement about the materials existing within that pixel.

## 4. Methods

### *4.1 Reduction of data complexity (Principal Component Analysis)*

We performed an uncentered and unscaled principal component analysis on our training data set, since all spectral data are already scaled and comparable. The first three of our 111 principal components explained 99.62% of the variance in our data (Fig. 3). As expected, correlations between principal components are negligible (Table 2).

Conducting a new automated principal component analysis on our image data would have led to the identification of different principal components, since eigenvectors and loadings were different, too. We use the automatically derived rotation matrix of the principal component analysis performed on our training data in order to manually create the same principal components on our image data. Our first image component is, thus:

```
pca.berlin.enmap.pc1 = sum(berlin.enmap * as.vector(pca.training.rotation[,1]))
```

Our assumption is that correlations between the three principal components of our image are similarly low and that they explain a similar proportion of variance. However, correlations between the image's components are not ignorable (Table 3).

The presentation of cumulative proportions of variance shows, though, that the same three components explain again more than 99% of our data variance (Fig. 4).

### *4.2 Multiple linear regression development*

The algorithm for our multiple linear regression development is based on a loop that performs every possible model on every pixel iteratively. It fills a result raster image containing information on the number of models per pixel that are adequate models as well as the minimum and the maximum fraction predicted. Its idea is presented using commented pseudo code (Table 4).

Our three score layers that indicate how many models were adequate for this pixel and both minimum and maximum fractions are used for analysis. The calculated fractions can be interpreted as fractions of imperviousness, because they give information on the fractions of the underlying classes of our training data. If a pixel's count information is 1, than there is one single and best model for that pixel.

### *4.3 Fraction analysis*

The following figures display different aspects of a selected clip within our study area near Bundesplatz, Berlin. Fig 5a shows our study area as a Google Earth aerial image, whereas Fig. 5b is the same area as a simulated EnMAP data image. The area features a Berlin living environment with a green belt including a lake in the western part of the image (west of sports grounds). The central road in east-west direction is part of an urban motorway. Rooftops consist of red clay and dark shale roofs. Most structures remain identifiable on the simulated EnMAP image despite of its low spatial resolution.

The result image on valid model counts (Fig. 5c) shows that each pixel is far away from having a single and best model. However, some structures of the aerial image are recognizable. Pixels representing streets and water seem to have a lower number of matching models (starting at 213), other structures offer up to 1495 models that predict imperviousness between 0 and 100%.

The model for the minimum value modeled per pixel always predicts an imperviousness rate of 0, whereas maximum values vary between 86% and 100%. Though, there is no shape of the aerial image recognizable, even if maxima might be lower in the green belt and the garden plots in the south-west of the clipped area (Fig. 5d and 5e).

Since none of the pixels offer a model count that is next to manageable, we give up the step of validating our results. We would first need to enhance our approach since until now, results are close to being meaningless with regard to our research question.

## 5. Discussion

Our study contributed to researching sub-pixel information on hyperspectral EnMAP imagery in heterogeneous urban areas. We use principal component analysis on synthetically generated training spectra for data reduction followed by a multiple linear regression analysis modeling fractions of imperviousness.

### *5.1 Principal component analysis*

As shown in chapter 3.1 the automatic reduction of training data complexity leads to suitable results featuring low correlations. Data variance is explained by only three principal components. Manual application of the PCA’s rotation matrix on the EnMAP scene, however, results in higher correlations. Nevertheless, only three components explain most of the data variance again.

In order to reduce processing complexity, data reduction is, in general, a useful and necessary tool. Still, high correlation values of our manipulated PCA (as we use the rotation matrix from a different PCA) might have a negative impact on the following steps.

### *5.2 Multiple linear regression analysis*

Our approach of a simple and comprehensible multiple linear regression analysis in order to achieve subpixel information might be useful in general and for a more limited set of data. However, despite the reduced data complexity the computing time for a large amount of multiple linear regression models is a non-negligible factor. Furthermore, it is uncertain whether this approach is able to find one best model for each pixel at all. Spectral profiles of synthetic mixtures of fractions are maybe too similar and, as a result, many models are capable of predicting reasonable fraction values. A positive aspect of our method is the possibility to intervene as necessary into the algorithm in contrast to a black-box-approach with predefined functions.

We identify three possible ways to improve the study's results. The use of fewer surface material classifications (e.g. by combining different classes in groups) would imply less models and thus less model matches. As our study reveals too many classes may produces redundancies and increase the uncertainty of the approach.

A second way could consist of considering all 111 spectral bands instead of a selected number of principal components. The expected effect is that spectral similarities within training data are reduced and fewer models are assumed to predict adequate pixel fractions.

The problem of high computing time is due to the fact that our algorithm iterates 2106 models over about 180,000 pixels. This could be approached by a model pre-selection. In a first step the residuals between a pixel's spectrum and both pure surface spectra (0% and 100%) of each training data set are summed up. Then only those models with the lowest residual values are tested on the respective pixel.

### *5.3 Further propositions*

Processing our analysis with Rstudio seems to be one reason for high computing times as this program does not originally support the use of multi-core processors. Though, installing further packages for Rstudio or switching to another programming language could possibly solve this problem.

One approach that would offer another possibility to validate the outcomes of this study includes further model classification and combination. Adjoining the linear regression all models adequate for one pixel are evaluated and divided into groups regarding the surface materials involved. Presumably many models of the same surface type (e.g. Tree Nr. 4) would fit the same pixel thus giving additional information about its texture.

The use of support vector regression (SVR) or support vector machine (SVM) is also possible for this kind of survey following [7] and [8]. These methods would be especially valuable to use the full potential of hyperspectral data.

## Acknowledgement

This work is part of the university course ‘Quantitative Methods’ at the Department of Geography at Humboldt University of Berlin. The authors wish to thank all course lecturers for their support, especially Cornelius Senf for his important advices in R programming and Akpona Okujeni for providing data sets and literature from his doctoral thesis.

**References**

[1] A.C. Braun, *Classification in high-dimensional Feature Spaces – Assessment using SVM, IVM and RVM with focus on simulated EnMAP Data*, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 18 (2012), pp. 436–443.

[2] T. Cocks, R. Jenssen, and A. Stewart, *The HyMap™ airborne hyperspectral sensor: The System, Calibration and Performance*, 1st Earsel Workshop on Imaging Spectroscopy, Zurich, Switzerland, 1998.

[3] M. Crawley, *The R Book*, John Wiley & Sons, Chichester, 2007.

[4] M.A. Farrell Jr., and R.M. Mersereau, *On the Impact of PCA Dimension Reduction for Hyperspectral Detection of difficult Targets*, IEEE Geoscience and Remote Sensing Letters 2 (2005), pp. 192–195.

[5] K. Koonsanit, C. Jaruskulchai, and A. Eiumnoh, *Band Selection for Dimension Reduction in Hyper Spectral Image using Integrated Information Gain and Principal Components Analysis Technique*, International Journal of Machine Learning and Computing 2 (2012), pp. 248–251.

[6] G. Mountrakis, J. Im, and C. Ogole, *Support Vector Machines in Remote Rensing: A review*, ISPRS Journal of Photogrammetry and Remote Sensing 66 (2011), pp. 247–259.

[7] A. Okujeni, S. van der Linden, and L. Tits, *Support Vector Regression and synthetically mixed Training Data for Quantifying Urban Land Cover*, Remote Sensing of Environment 137 (2013), pp. 184–197.

[8] A. Okujeni, S. van der Linden, and B. Jakimow, *A Comparison of Advanced Regression Algorithms for Quantifying Urban Land Cover*, Remote Sensing 6 (2014), pp. 6324–6346.

[9] A. Okujeni, S. van der Linden, and P. Hostert, *Extending the Vegetation-Impervious-Soil Model using simulated EnMAP Data and Machine Learning*, Remote Sensing of Environment 158 (2015), pp. 69–80.

[10] G. Quinn, and M. Keough, *Experimental Design and Data Analysis for Biologists*, Cambridge University Press, 2002.

[11] D.A. Roberts, M. Gardner, and R. Church, *Mapping Chaparral in the Santa Monica Mountains using Multiple Endmember Spectral Mixture Models*, Remote Sensing of Environment 65 (1998), pp. 267–279.

[12] S. Roessner, K. Segl, and U. Heiden, *Automated Differentiation of Urban Surfaces based on Airborne Hyperspectral Imagery,* IEEE Transactions on Geoscience and Remote Sensing, 39 (2001), pp. 1525–1532.

[13] *The R Project*, The R Foundation for Statistical Computing, Vienna, Austria, 2015, software available at http://cran.r-project.org/mirrors.html.

[14] United Nations Department of Economic and Social Affairs (UNDESA) *World Urbanization Prospects: The 2014 Revision*, UNDESA – Population Division, 2014.

[15] Q. Weng, *Remote Sensing of impervious Surfaces in the Urban Areas: Requirements, Methods, and Trends*, Remote Sensing of Environment 117 (2011), pp. 34–49.

Table 1. List of considered materials for analyses [9]

| Group | Class | Nr of spectra |
|---|---|---|
| Roof | Red clay tile | 4 |
| | Red cement tile | 3 |
| | Bitumen | 5 |
| | Brown roof tile | 1 |
| | Brown roof shingle | 1 |
| | White roof material (polyethylene) | 1 |
| | White roof material (unknown) | 1 |
| | Zinc roof material | 1 |
| Pavement | Asphalt | 4 |
| | Concrete | 2 |
| Grass | Grass (intensively manicured) | 2 |
| | Grass (extensively manicured) | 1 |
| | Grass (dry) | 2 |
| Tree | Deciduous tree | 7 |
| Other | Tartan | 1 |
| | Railtrack (concrete sleepers) | 1 |
| | Railtrack (wooden sleepers) | 1 |
| | Sand (playground) | 1 |
| | Soil | 1 |
| | Water | 1 |

Table 2. Correlation matrix of the training data

| | PC 1 | PC2 | PC3 |
|---|---|---|---|
| PC1 | 1.00 | 0.17 | 0.02 |
| PC2 | 0.17 | 1.00 | 0.00 |
| PC3 | 0.02 | 0.00 | 1.00 |

Table 3. Correlation matrix of the EnMAP data

| | PC 1 | PC2 | PC3 |
|---|---|---|---|
| PC1 | 1.00 | -0.37 | -0.46 |
| PC2 | -0.37 | 1.00 | 0.63 |
| PC3 | -0.46 | 0.63 | 1.00 |

Table 4. Idea of our multiple linear regression development

| | |
|---|---|
| for (any pixel from 1 to 12636 in steps of 6) | Our training data consists of 2106 models represented in blocks of 6 artificial mixtures. |
| { | |
| subs = subset(rows i to i+5) | |
| fit = fit(predict fractions from PC1 + PC2 + PC3 in subs) | One model is based on one subset.<br>Train model |
| for (any pixel in the enmap image) | Apply model on every pixel of our image |
| { | |
| a = prediction(fit) | Predict temporary fraction value |
| if (a is between 0 and 100) | If the prediction turns out that the model is adequate |
| { | |
| Count++ | Increase model count |
| Update minFractions | And update min and max fractions if |
| Update maxFractions | necessary! |
| } | |
| } | |
| } | |

Figure 1. Study area in western Berlin, Germany [9]

Figure 2. Process chart of this study’s methods

Figure 3. Variance of the training data explained by the PCA

Figure 4. Variance of the EnMAP data explained by the PCA

Figure 5a. Aerial imagery by Google Earth

Figure 5b. Simulated EnMAP data

Figure 5c. Maximum fractions of imperviousness

Figure 5d. Minimum fractions of imperviousness

Figure 5e. Count of valid models