

AUTOMATIC ROAD PAVEMENT DETECTION IN MAPUTO, MOZAMBIQUE

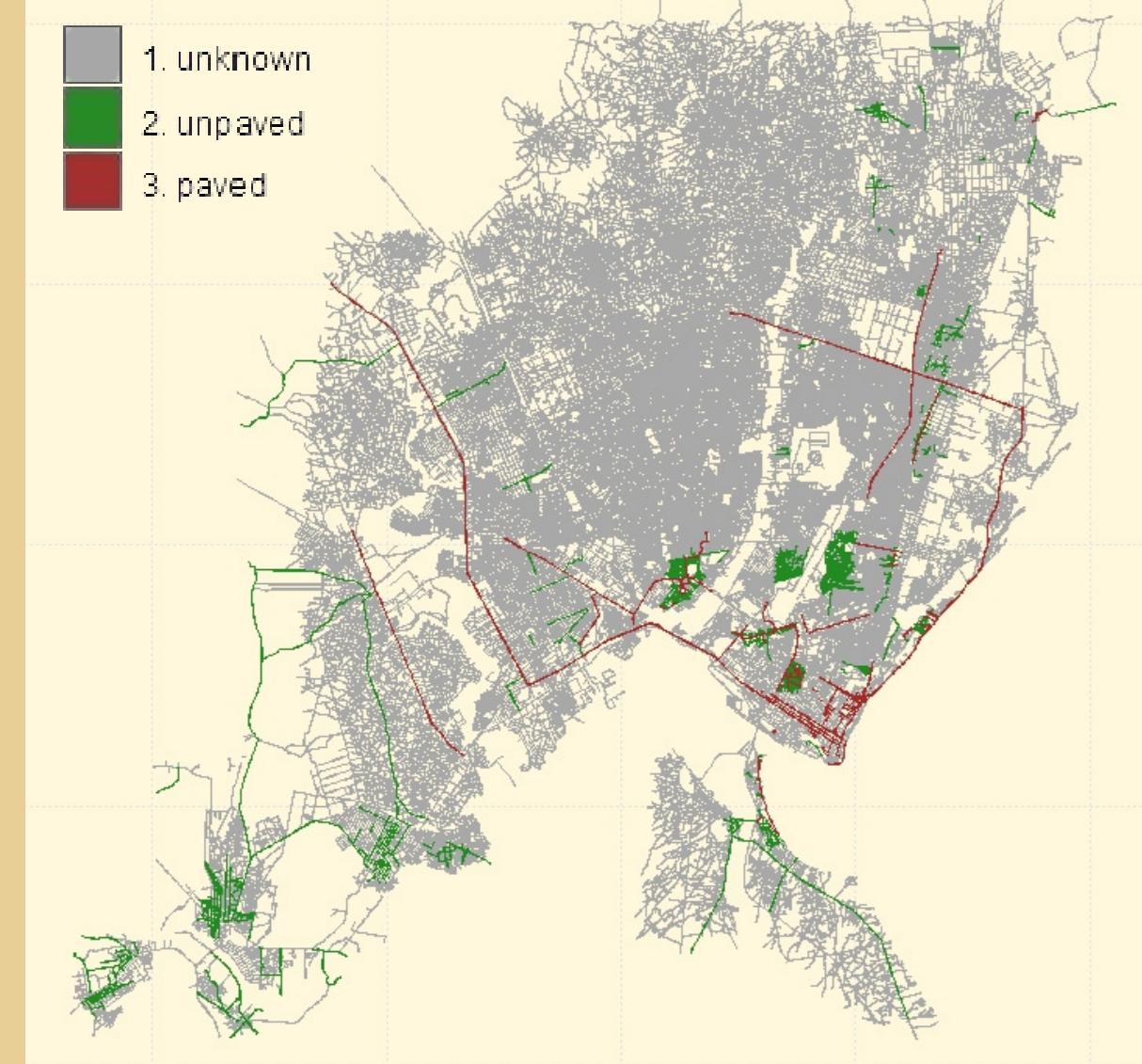
by L. Maci, F. Pagella, M. Poggi, A. Puricelli & G. Venturini

Goal: Road Pavement Classification in **paved** and **unpaved**.

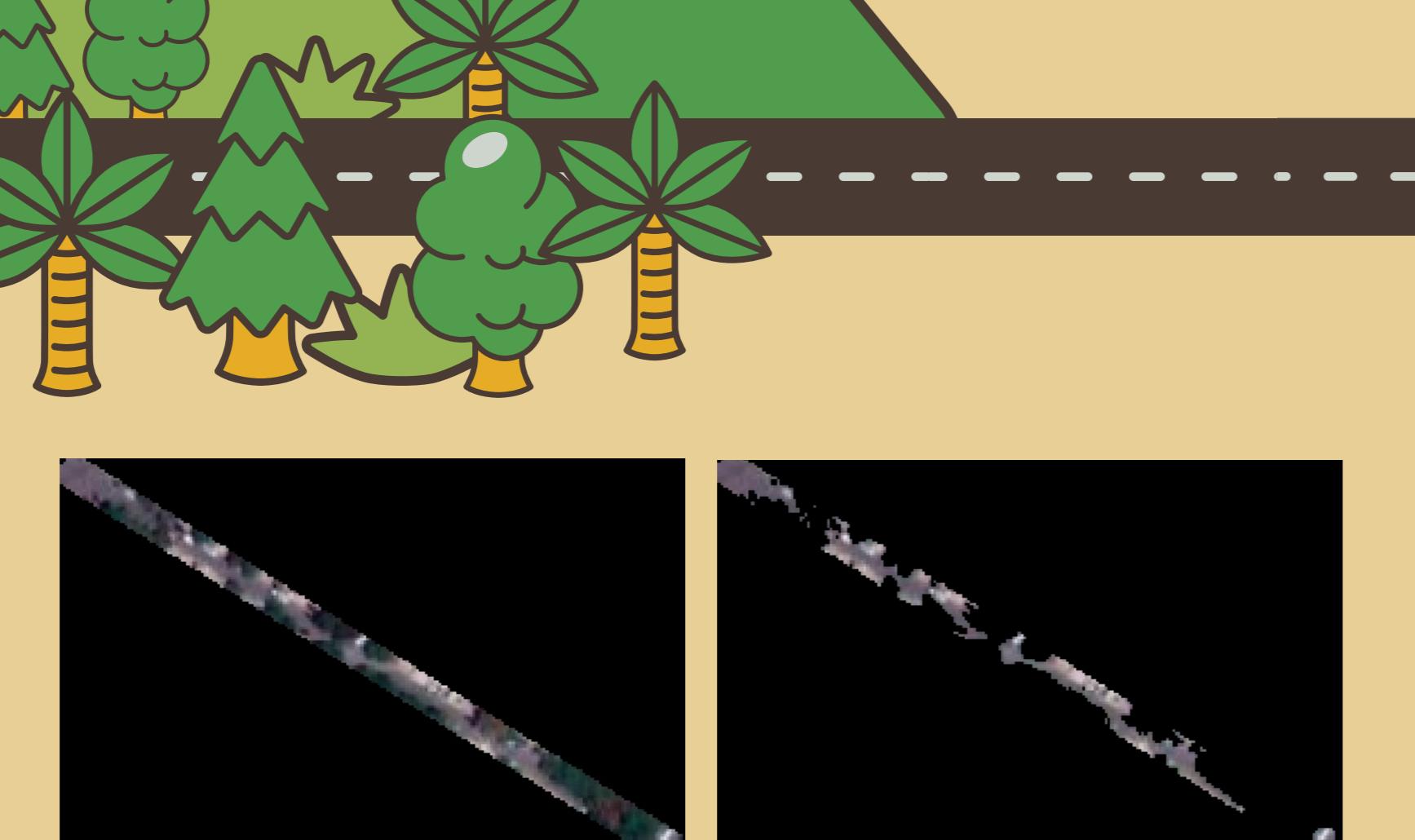
Use known pavement data (5%) and pixel color of satellite images to detect the road pavement of unknown data (95%) in the road network.

We work on a **reduced dataset** using 2558 known roads and 2558 to be classified.

MAP OF KNOWN ROADS

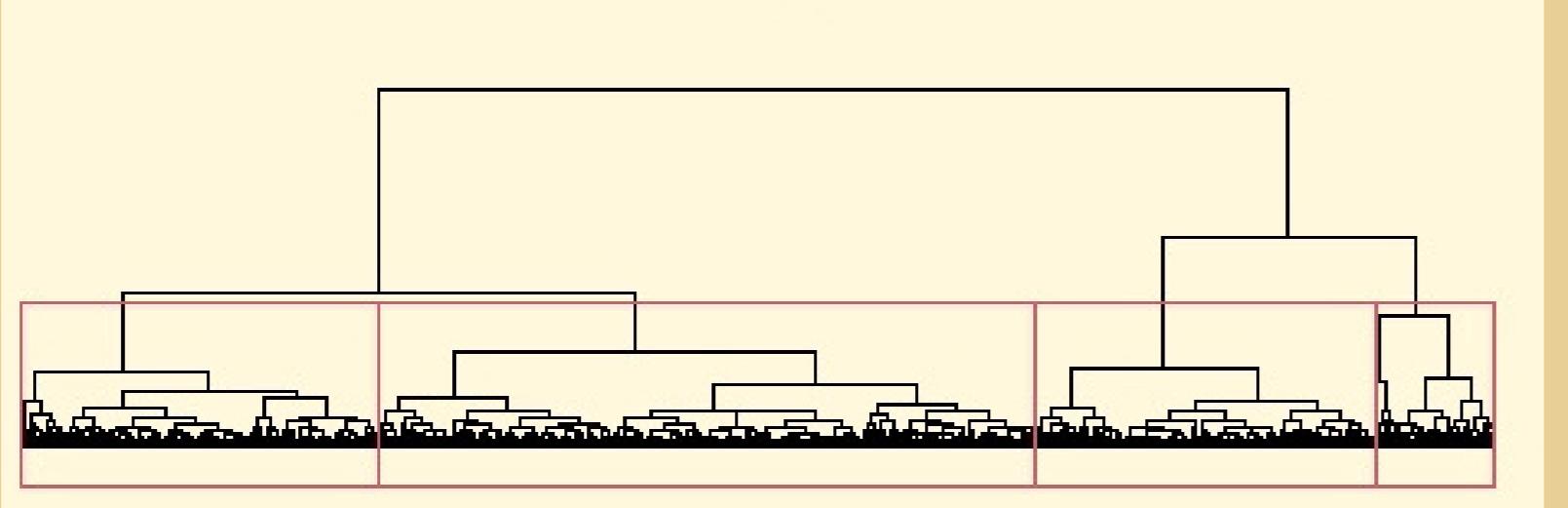


CLUSTERING FOR ROAD CLEANING

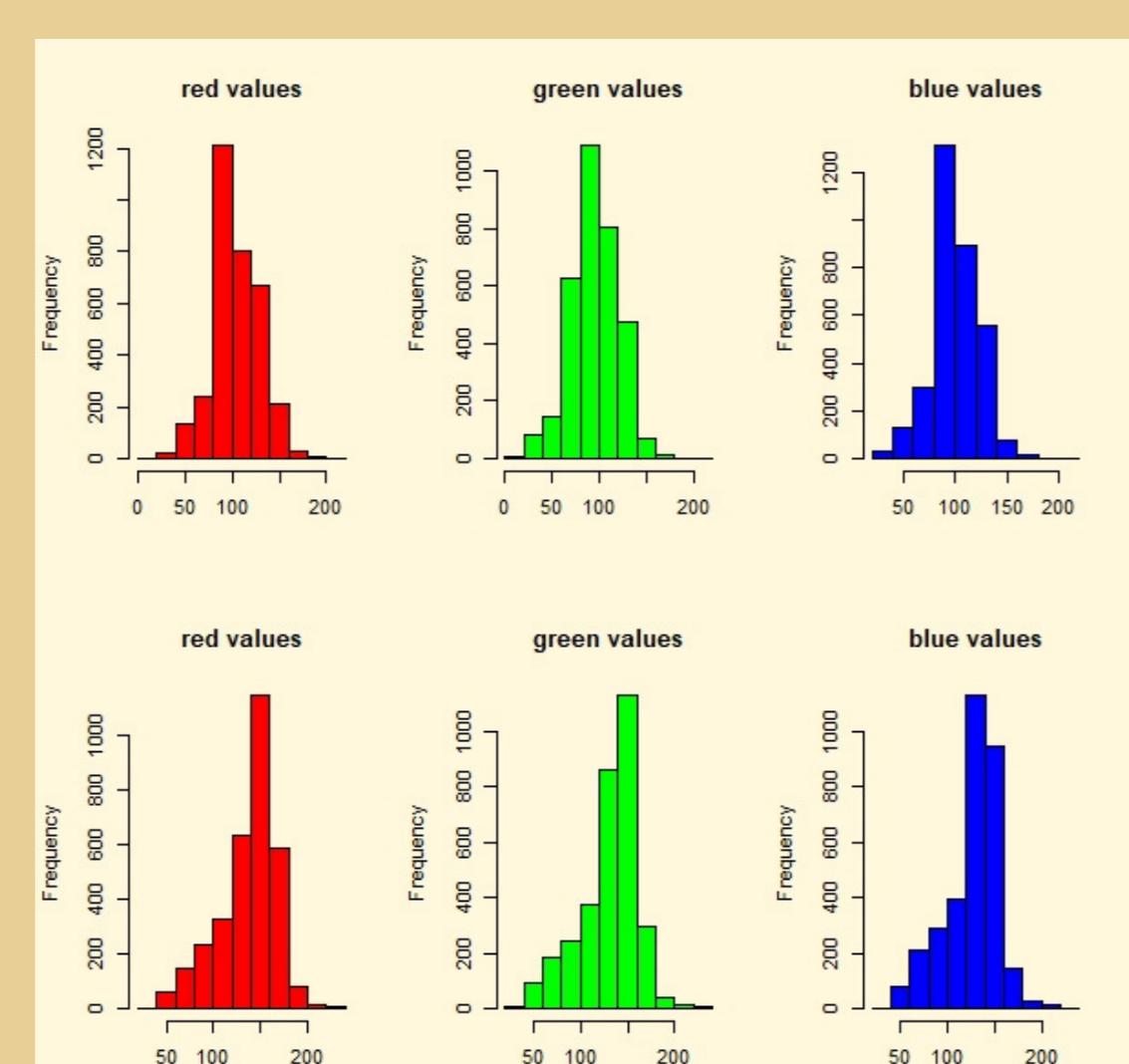
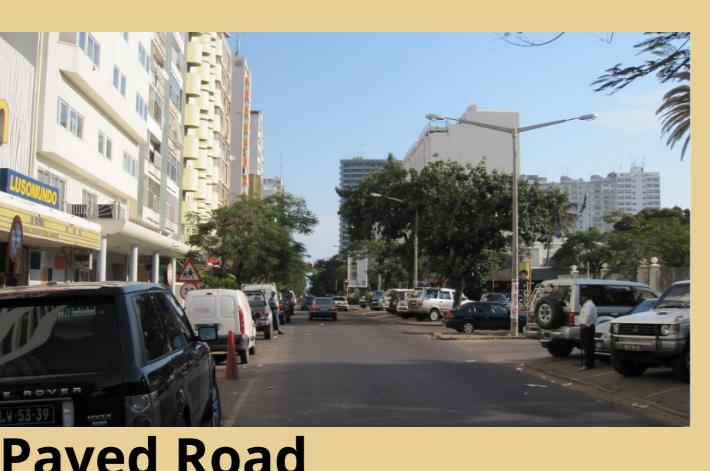
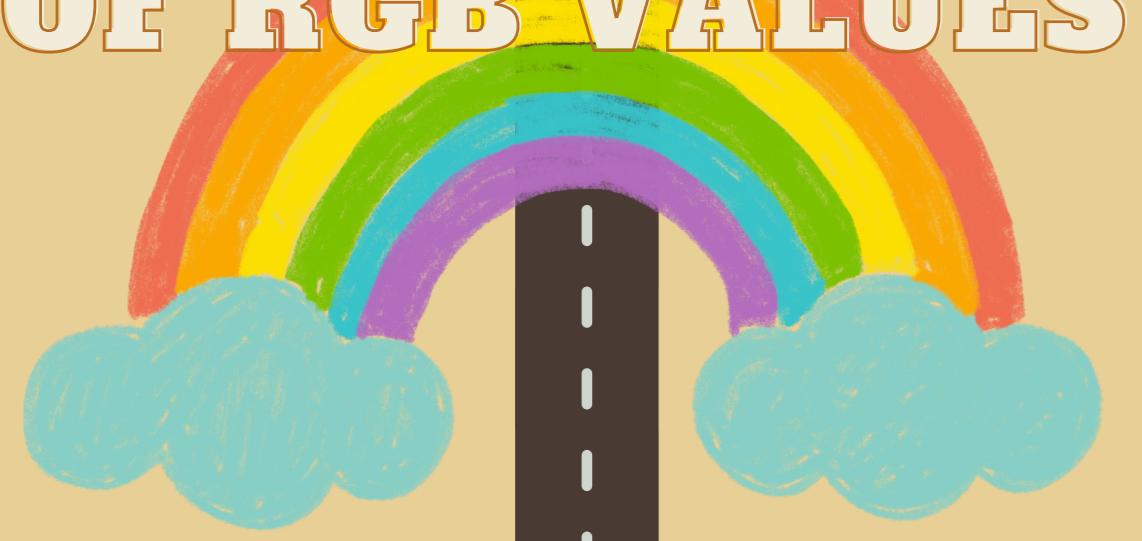


Performing a **hierarchical clustering** with **Euclidean distance**, **complete linkage** and number of clusters **k = 4**, we can identify and remove the pixels a lot darker than average corresponding to the trees.

Dendrogram



EXTRACTION OF RGB VALUES



We remove black pixels and then, using **library raster**, we can extract from the pictures the **RGB values**. Every road is than summarized by the following attributes:

- **Mean** of red, green and blue
- **Median** of red, green and blue
- **Variance** of red, green and blue
- **Maximum** of red, green and blue
- **Minimum** of red green and blue

```
Deviance Residuals:
    Min      Q1     Median      Q3      Max 
-1.9939 -0.2962 -0.1182  0.0000  3.3217 

Coefficients: (1 not defined because of singularities)
Estimate Std. Error z value Pr(>|z|) 
(Intercept) -7.949e+00 1.200e+11 -6.626 3.44e-11 *** 
rmean -3.861e-02 8.989e-02 -0.430 0.665759  
rvar 2.236e-02 8.906e-02 0.252 0.802621  
med 8.815e-02 1.280e-02 -0.680 0.280261  
rmin 2.998e-02 1.940e-02 1.546 0.122123  
rmax 4.850e-02 1.382e-02 3.509 0.000450 *** 
gmean 1.354e-02 1.380e-02 0.101 0.962777  
gvar 1.427e-02 1.766e-02 0.824 0.609919  
gmed 4.297e-02 1.194e-01 0.360 0.719041  
gmin -7.826e-02 2.774e-02 -2.821 0.004781 ** 
gmax 1.267e-02 2.226e-02 -0.572 0.952485  
bvar 1.1361e-03 1.655e-03 -0.822 0.410968  
bmed 2.846e-02 2.729e-02 1.028 0.380413  
bmin 8.913e-03 1.779e-02 -0.523 0.600023  
osm_type_footway 4.532e+00 1.068e+01 -4.245 2.19e-05 *** 
osm_type_primary 2.176e+01 4.685e+02 0.046 0.962953  
osm_type_residential -1.098e+00 2.889e-01 -3.685 0.000144 *** 
osm_type_secondary 2.036e+00 9.886e-01 0.221 0.982803  
osm_type_unk NA NA NA NA NA NA 
--- 
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Dispersion parameter for binomial family taken to be 1 

Null deviance: 2760.89 on 2299 degrees of freedom 
Residual deviance: 786.23 on 2279 degrees of freedom 
AIC: 828.23 

Number of Fisher Scoring iterations: 18
```

```
Deviance Residuals:
    Min      Q1     Median      Q3      Max 
-1.9480 -0.3044 -0.1184  0.0000  3.2399 

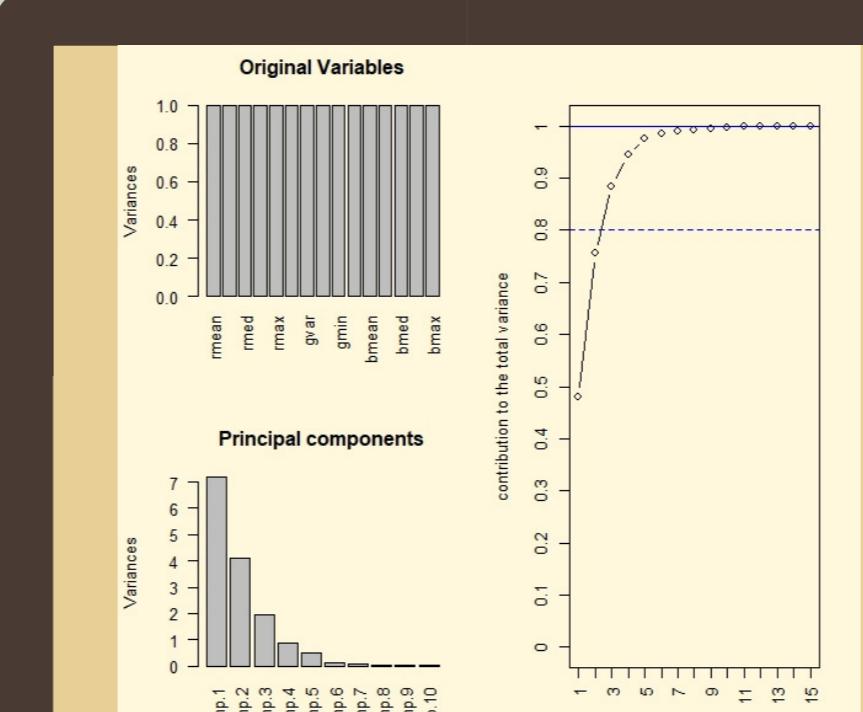
Coefficients: 
Estimate Std. Error z value Pr(>|z|) 
(Intercept) -8.186e-01 1.151e+00 -7.094 1.39e-12 *** 
rmean -2.000e-02 8.989e-02 -0.222 0.838120  
rvar 2.652e-02 1.245e-02 2.130 0.03320 
med 4.864e-02 1.042e-02 4.668 1.01e-06 *** 
rmin 1.132e-03 1.779e-02 -0.632 0.500023  
rmax 1.452e-02 2.729e-02 1.028 0.380413  
gmean 5.512e-02 1.482e-02 -3.071 0.00213 ** 
gvar 1.354e-02 1.766e-02 0.782 0.409919  
gmed 4.297e-02 1.194e-01 0.360 0.719041  
gmin -7.826e-02 2.774e-02 -2.821 0.004781 ** 
gmax 1.267e-02 2.226e-02 -0.572 0.952485  
bvar 1.1361e-03 1.655e-03 -0.822 0.410968  
bmed 2.846e-02 2.729e-02 1.028 0.380413  
bmin 8.913e-03 1.779e-02 -0.523 0.600023  
osm_type_footway 4.532e+00 1.068e+01 -4.245 2.19e-05 *** 
osm_type_primary 2.176e+01 4.685e+02 0.046 0.962953  
osm_type_residential -1.098e+00 2.889e-01 -3.685 0.000144 *** 
osm_type_secondary 2.036e+00 9.886e-01 0.221 0.982803  
osm_type_unk NA NA NA NA NA NA 
--- 
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Dispersion parameter for binomial family taken to be 1 

Null deviance: 2760.89 on 2299 degrees of freedom 
Residual deviance: 786.23 on 2279 degrees of freedom 
AIC: 828.23 

Number of Fisher Scoring iterations: 18
```

PCA



We perform **PCA** and find that the **first 5 PCs** explain 97.66% of the total variance. For the following points we are going to use this **reduced dataset**:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
1	0.35938084	1.62278094	-2.23307676	0.482444906	-0.25574999
2	-2.55854965	-1.49283674	-1.78274060	0.355884062	-0.63203258
3	-0.45000622	0.62346240	-1.36905465	0.194727782	-0.54328501
4	-1.81542441	1.29727595	-0.73996580	0.307239867	-0.87559825
5	-4.10657297	-1.48084473	-1.21191012	-1.125126555	-0.51986443
6	-4.22754850	-2.36579861	-1.33190568	-0.254601665	-0.86182319
7	-1.08811250	0.30401706	-3.05276235	0.067767377	-0.10495380
8	-2.34750901	-0.25609776	-2.43018571	-0.520313985	0.35337121
9	0.67228982	1.20028232	-2.65805951	-0.785531662	1.25502764
10	2.11729584	0.96519000	-1.66884522	0.031016113	0.46692916
11	-2.03373794	-0.95631223	-2.41396635	0.135137241	-0.06051613

QDA

Train set: 2300 roads.
Test set: 257 roads.
Even if normality assumptions are not met we perform QDA on the first 5 PCs getting the following results:

QDA		Class Assigned	
		paved	unpaved
Class	paved	46	23
	unpaved	25	163
Accuracy: 0.8132296			

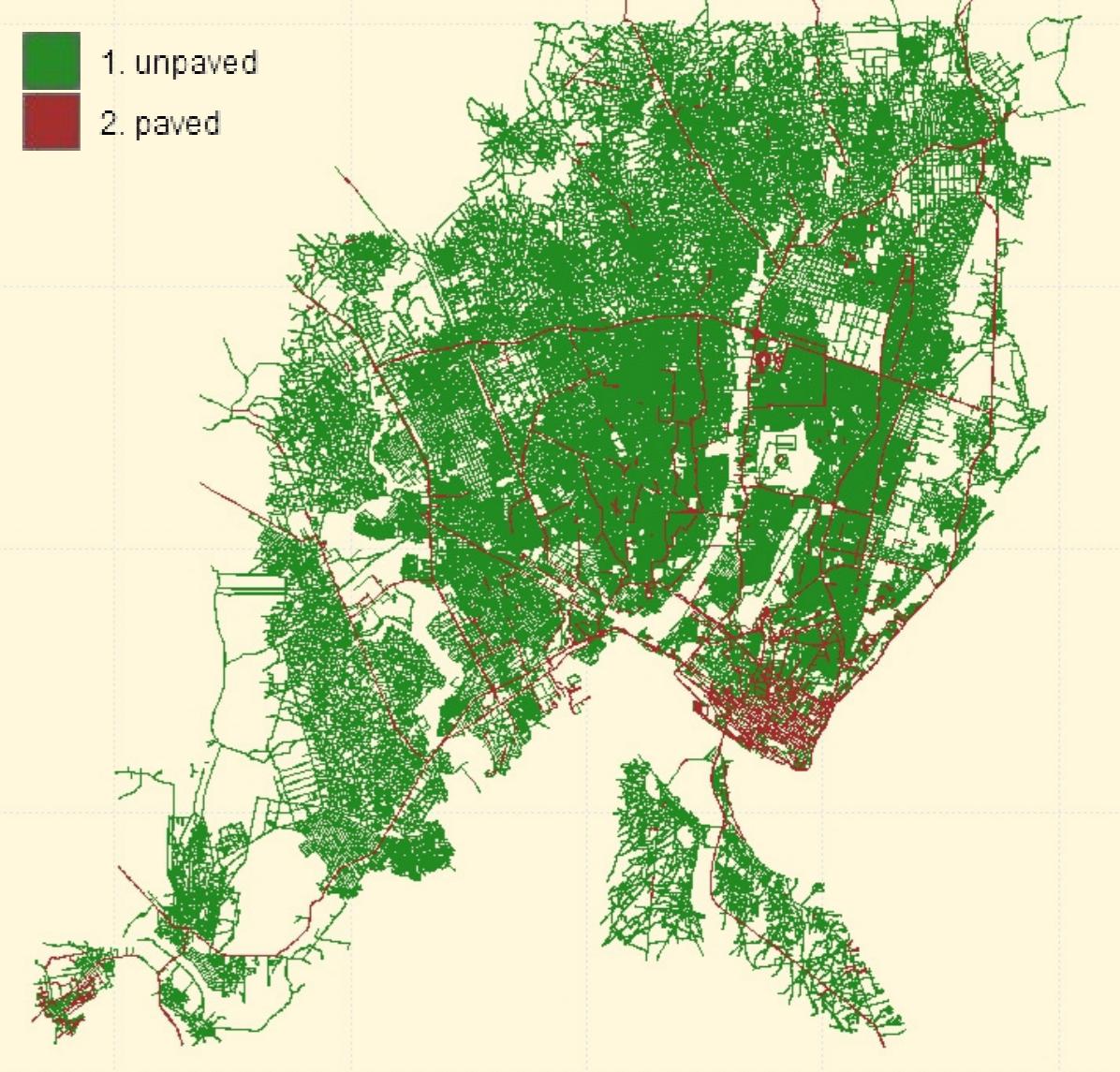
KNN

Train set: 2300 roads.
Test set: 257 roads.
Via **Leave-One-Out cross-validation** we select the optimal number of neighbours: **K = 12**.

KNN		Class Assigned	
		paved	unpaved
Class	paved	43	26
	unpaved	12	176
Accuracy: 0.8521401			

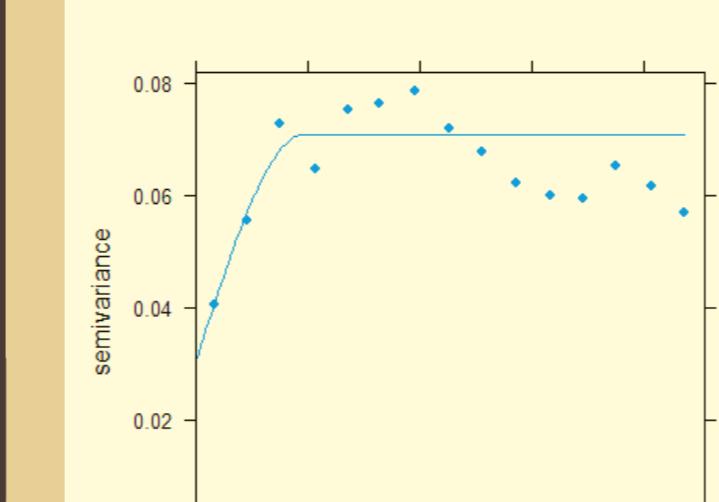
Geostatistics		Class Assigned	
		paved	unpaved
Class	paved	52	17
	unpaved	2	186
Accuracy: 0.92607			

MAP OF CLASSIFIED ROADS



LOGISTIC REGRESSION

Model: **osm_surf ~ dist + osm_typo + rmean + rvar**, where dist is the Euclidean distance from the City Hall.



Fitted Variogram:
model psill range
1 Nug 0.03045763 0.000
2 Sph 0.04039150 4834.602

Train set: 2300 roads.
Test set: 257 roads.
We select the best threshold in terms of accuracy and find **threshold = 0.5**.

- Predicted osm_surf > 0.5:
paved
- Predicted osm_surf <= 0.5:
unpaved

Logistic Reg		Class Assigned	
		paved	unpaved
Class	paved	59	10
	unpaved	5	183
Accuracy: 0.9416342			

Train set: 2300 roads.
Test set: 257 roads.
We predict if a street is **paved (1)** or **unpaved (0)** based on the distribution of the colors and the type of the street using adequate dummy variables.
Using the **backward stepwise selection** we obtain the parameters of the final model. We start with a model containing 21 variables and AIC = 828.23, to end up with the reduced final model with **AIC = 816.71**.
To select the threshold parameter we use **10-fold cross-validation** and find that the best one is **threshold = 0.35**.