AUTOMATIC ROAD PAVEMENT DETECTION IN MAPUTO, MOZAMBIQUE

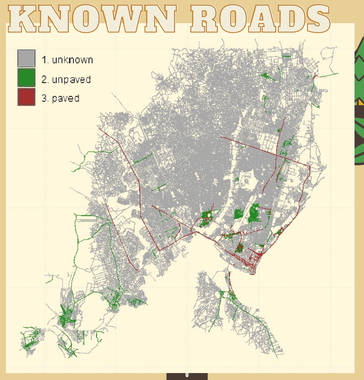
by L. Maci, F. Pagella, M. Poggi, A. Puricelli & G. Venturini

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

Possible stakeholder: Ministry of Transport and Communications of Mozambique.

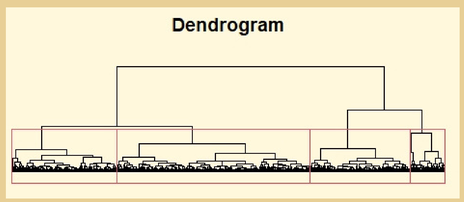
Definition of ARPD: "Development of a machine learning pipeline aimed at the classification of road surfaces based on the color of the pavement and its particular location"

MAP OF KNOWN ROADS



CLUSTERING FOR ROAD CLEANING

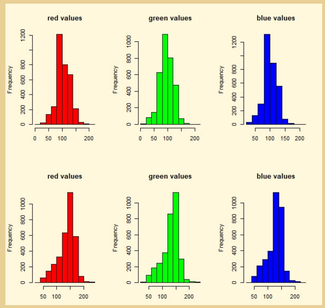
Performing a **hierarchical clustering** with **Euclidean distance**, **complete linkage** and number of clusters **k = 4** , we can identify and remove the pixels a lot darker than average corresponding to the trees.



EXTRACTION OF RGB VALUES

Paved Road

Unpaved Road



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

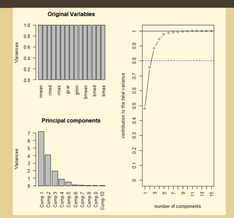
- **Mean** of red, green and blue
- **Median** of red, green and blue
- **Variance** of red, green and blue
- **Maximum** of red, green and blue
- **Minimum** of red green and blue

FINAL DATASET

	osm_typo	osm_surf	rmean	rvar	rmed	rmin	rmax	gmean	gvar
1	unk	paved	131.90279	990.1693	127.0	69	254	120.57800	856.9814
2	residentia	paved	108.40521	456.9249	104.0	66	216	98.91439	449.9038
3	secondary	paved	126.95409	807.2594	121.0	66	233	120.84048	793.7170
4	primary	paved	125.34349	970.1501	122.0	46	227	110.67160	895.4406
5	residentia	paved	94.30961	667.2700	87.0	59	190	91.78025	671.5501
6	residentia	paved	92.90636	419.6536	89.0	58	196	87.86545	393.7671
7	residentia	paved	111.15013	741.5304	105.0	67	239	107.10098	754.4538
8	tertiary	paved	106.12740	837.8268	95.0	71	226	100.13380	812.3197
9	unk	paved	132.22705	1044.2093	123.0	95	251	117.32971	1104.3932
10	residentia	paved	144.72265	868.3482	140.0	90	254	137.21756	971.9552
11	primary	paved	107.53608	575.2520	104.0	70	227	100.12975	550.6538
12	residentia	paved	86.26697	713.2659	80.0	47	223	80.42372	729.9552
13	residentia	paved	117.07309	479.9277	113.0	84	220	111.32737	447.3410
14	residentia	paved	140.70701	1022.8583	129.0	98	240	133.96019	1056.6978
15	secondary	paved	121.15531	713.2773	117.0	80	222	115.51362	719.1615

gmed	gmin	gmax	bmean	bvar	bmed	bmin	bmax
114.0	54	247	128.12165	817.23097	123	71	250
95.0	55	206	103.82556	363.39326	100	64	214
115.0	59	229	123.88756	702.16496	119	59	227
107.0	40	214	116.74203	837.42659	114	48	224
85.0	58	188	99.68416	581.08894	93	62	193
86.0	51	185	94.86727	310.83405	92	63	189
101.0	67	243	113.08132	640.57094	108	76	243
91.0	65	222	106.07362	657.56600	96	74	219
107.0	82	243	125.04245	966.93594	115	88	248
132.0	79	254	134.02099	718.45468	130	82	230
95.0	65	222	104.52215	504.02294	100	73	220
75.0	40	221	87.88849	614.05966	83	51	220
108.0	76	214	115.41213	402.43702	112	83	216
122.0	94	237	134.23461	926.43564	123	94	249
112.0	74	219	120.06948	626.71959	116	83	228

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	-1.81542841	1.29727595	-0.73996580	0.307239867	-0.87559825
5	-4.10657297	-1.48084473	-1.21191012	-1.125126555	-0.51986443
6	-4.22754850	-2.36579861	-1.33190658	-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	-2.65809551	-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 True	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**.
Then we perform kNN classification and get the following results on the test set.

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

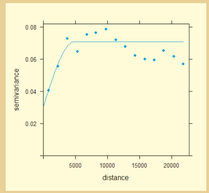
GEOSTATISTICS

Train set: 2300 roads.
Test set: 257 roads.
We perform kNN classification and get the following results on the test set.

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

LOGISTIC REGRESSION

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



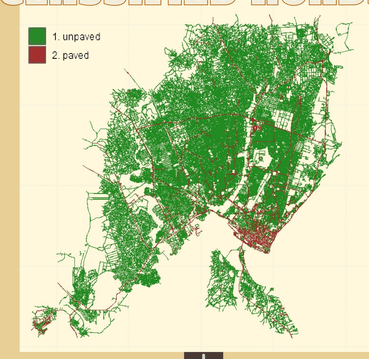
Fitted Variogram:
model psill range
1 Nug 0.03045763 0.000
2 Sph 0.04039150 4834.602

```
Deviance Residuals:
    Min       1Q   Median       3Q      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+00  -6.626 3.44e-11 ***
rmean                 -3.861e-02  8.989e-02  -0.430 0.667539
rvar                   2.711e-04  8.886e-04   0.305 0.760352
rmed                  -8.536e-02  7.906e-02  -1.080 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                  1.554e-02  1.349e-01   0.115 0.908277
gvar                  -1.430e-03  1.736e-03  -0.824 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-03  2.208e-02  -0.057 0.954249
bmean                  1.545e-01  1.266e-01   1.221 0.222159
bvar                  -1.361e-03  1.655e-03  -0.822 0.410968
bmed                  -6.721e-02  1.114e-01  -0.603 0.546224
bmin                   2.864e-02  2.789e-02   1.027 0.304413
bmax                  -9.836e-03  1.879e-02  -0.523 0.600691
osm_typo_footway      -4.532e+00  1.068e+00  -4.245 2.19e-05 ***
osm_typo_primary       2.176e+01  4.685e+02   0.046 0.962953
osm_typo_residentia -1.098e+00  2.889e-01  -3.801 0.000144 ***
osm_typo_secondary     2.119e+01  9.886e+02   0.021 0.982898
osm_typo_tertiary      2.036e+00  3.270e-01   6.227 4.75e-10 ***
osm_typo_unk              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
```

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

MAP OF CLASSIFIED ROADS



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

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9480  -0.3044  -0.1184   0.0000   3.2539

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)           -8.166e+00  1.151e+00 -7.094 1.30e-12 ***
rmed                  -1.068e-01  1.024e-02 -10.428  < 2e-16 ***
rmin                   2.652e-02  1.245e-02   2.130  0.03320 *
rmax                   4.864e-02  1.042e-02   4.668 3.05e-06 ***
gvar                  -2.124e-03  3.942e-04  -5.387 7.18e-08 ***
gmin                  -4.552e-02  1.482e-02  -3.071  0.00213 **
bmean                  1.324e-01  1.450e-02   9.130  < 2e-16 ***
bmax                  -1.390e-02  8.918e-03  -1.558  0.11917
osm_typo_footway      -4.771e+00  1.051e+00  -4.539 5.64e-06 ***
osm_typo_primary       2.170e+01  4.709e+02   0.046  0.96325
osm_typo_residentia -1.093e+00  2.827e-01  -3.868  0.00011 ***
osm_typo_secondary     2.109e+01  9.997e+02   0.021  0.98317
osm_typo_tertiary      2.045e+00  3.203e-01   6.385 1.71e-10 ***
---
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:  790.71  on 2287  degrees of freedom
AIC: 816.71

Number of Fisher Scoring iterations: 18
```

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**.