Landsat-8 satellite images for human density prediction

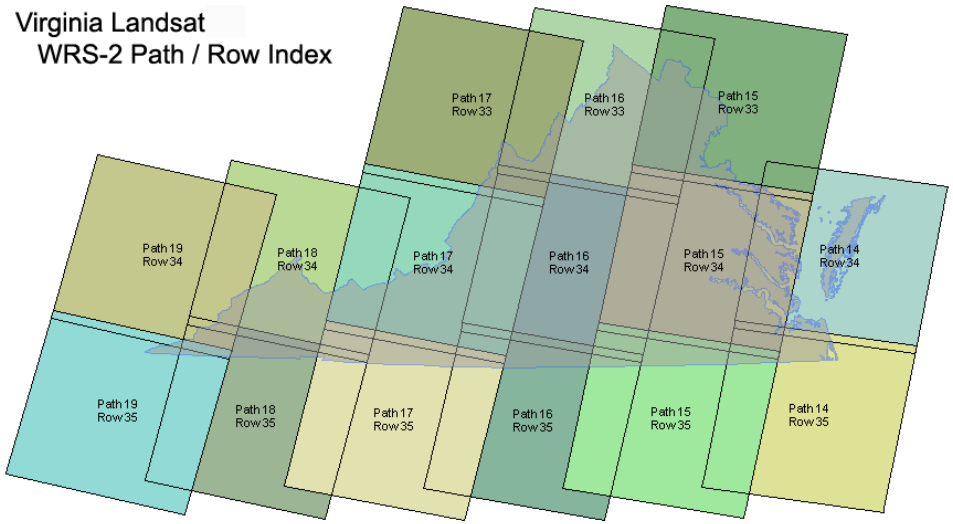
Youcef Kacer

www.github.com/ykacer

2017,January 25th

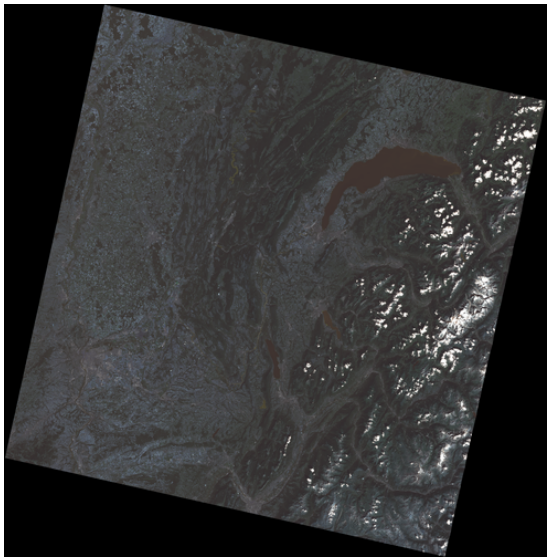
- Population census is expensive using conventional methods
  - 180 millions euros in France (officials,1999)
- How Satellites images could explain human density ?
- How to transform satellite images to best explain human density ?

- Total earth covering defined by (path,row) grid pattern and achieved every 16 days



Landsat-8 grid covering (path,row) for Virginia (*USA*)

- Landsat-8 images are georeferenced which means each pixel has (x,y) meter coordinates in a certain Projection Coordinates System (*ex : UTM, Lambert 93, Web Mercator,...*)

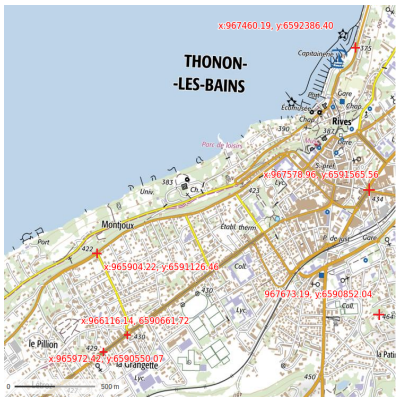


Landsat-8 Eastern-France image
path=196,row=028
georeferenced in *UTM* system
(image containing city
Thonon-les-Bains)

UTM corner coordinates (meters)

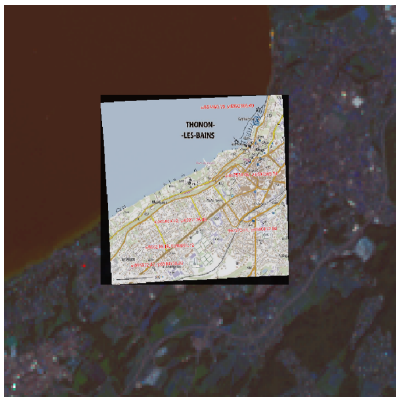
- upper-left (486202.777, 5958940.580)
- lower-left (486202.777, 5610657.403)
- upper-right (828135.467, 5958940.580)
- lower-right (828135.467, 5610657.403)
- center (657169.122, 5784798.992)

- Landsat-8 georeferencement can be checked comparing with another georeferenced source like *IGN* using a *SIG* (open source *QGIS*).



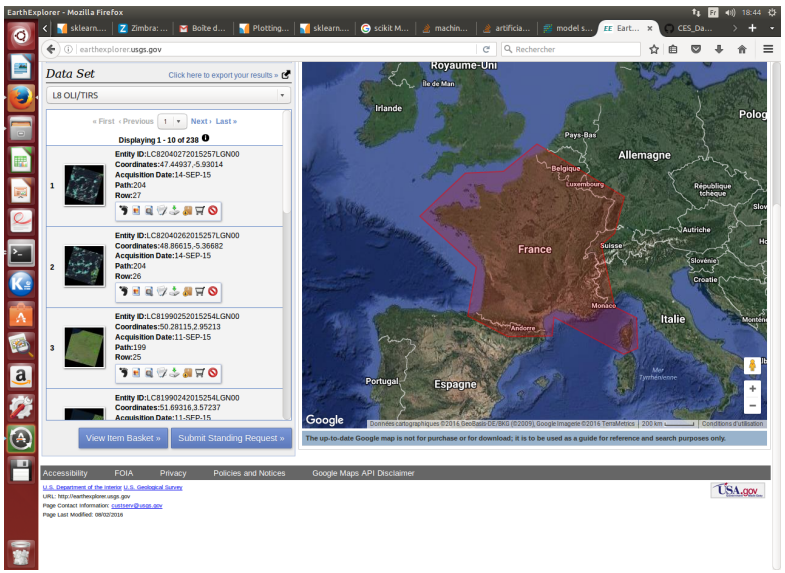
IGN map georeferenced in *Lambert* 93 system

Then IGN image is tranformed to be georeferenced in *UTM* system

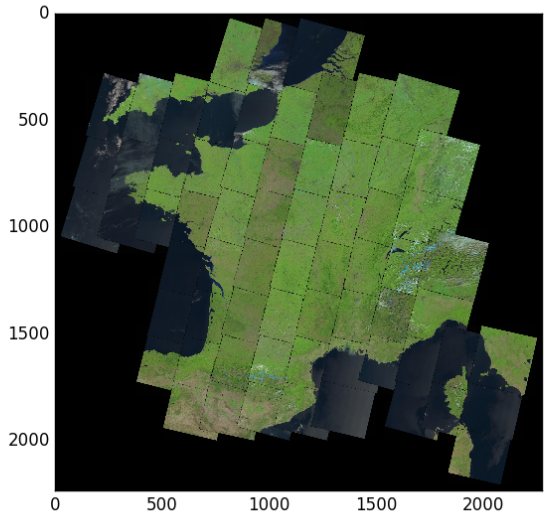


Superposition of IGN map and Landsat-8 image both georeferenced in *UTM* system

- Query images from *U.S geological Survey* website with criterias :
  - cloud covering $\leq 20\%$
  - day acquisition
  - between May, 2013 and September, 2013

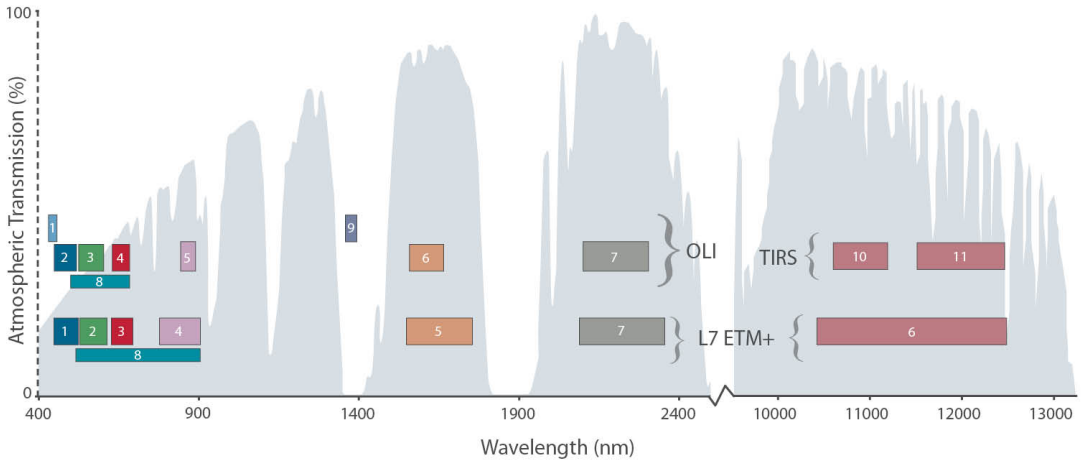


Polygon selection on *USGS* website



70 resulting datasets georeferenced in *Web Mercator* system

- Landsat-8 dataset is composed of
  - 11 bands (OLI/TRS sensors) + 1 quality band (cloudyness of each pixel)
  - Possible combinaison of bands to extract information (bands 4 and 5 for vegetation presence)



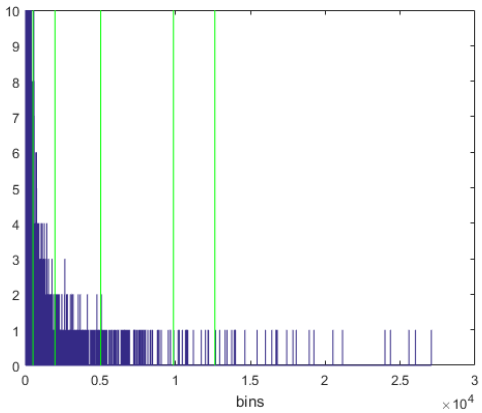
Landsat-8 bands (OLI/TIRS sensors)

- Take cities surfaces and densities from 2013 official census (INSEE)
- Take cities latitude and longitude from Google geolocator (Python API Geopy)

name	latitude (degrees)	longitude (degrees)	surface (km^2)	density ($habs/km^2$)
Ozan	46.391534	4.915265	6.6	98.3
Cormoranche-sur-Saône	46.240532	4.830863	9	118.9
Paris	48.856614	2.352222	105.4	21153.9
Lyon	45.764043	4.835659	47.87	10117.0
Tours	47.394144	0.68484	34.67	3888.2
Besancon	47.237829	6.024054	65.05	1797.9
...	...	...	...	...

34190 cities (instances)

- Categorize densities applying clustering (Otsu mutli-thresholding)



Density histogram (habs/km^2)

catégory 1 : density between 0 and 500 habs/km^2

catégory 2 : density between 500 and 2000 habs/km^2

catégory 3 : density between 2000 and 5000 habs/km^2

catégory 4 : density between 5000 and 10000 habs/km^2

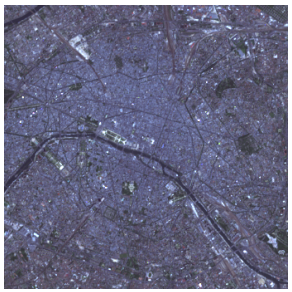
catégory 5 : density between 10000 and 13000 habs/km^2

catégory 6 : density greater than 13000 habs/km^2

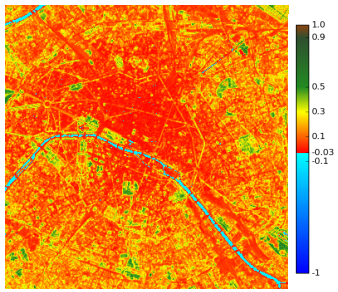
city	latitude (degrees)	longitude (degrees)	surface (km^2)	density (habs/km^2)	density (category)
Ozan	46.391534	4.915265	6.6	98.3	1
Cormoranche-sur-saone	46.240532	4.830863	9	118.9	1
Paris	48.856614	2.352222	105.4	21153.9	6
Lyon	45.764043	4.835659	47.87	10117.0	5
Tours	47.394144	0.68484	34.67	3888.2	3
Besancon	47.237829	6.024054	65.05	1797.9	1
...	...	...	...	...	...

- Compute Normalized Difference Vegetation Indice using bands 4 (Red) and 5 (Near-Infra-Red) for each dataset :
- Values between -1 and 1
- ≤ 0 for water, snow and cloud
- ≤ 0.2 for ground without vegetation
- ≥ 0.2 for ground vegetation

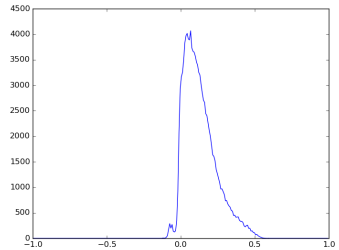
$$NDVI = \frac{NIR - R}{NIR + R}$$



RGB (Paris,May 2013)

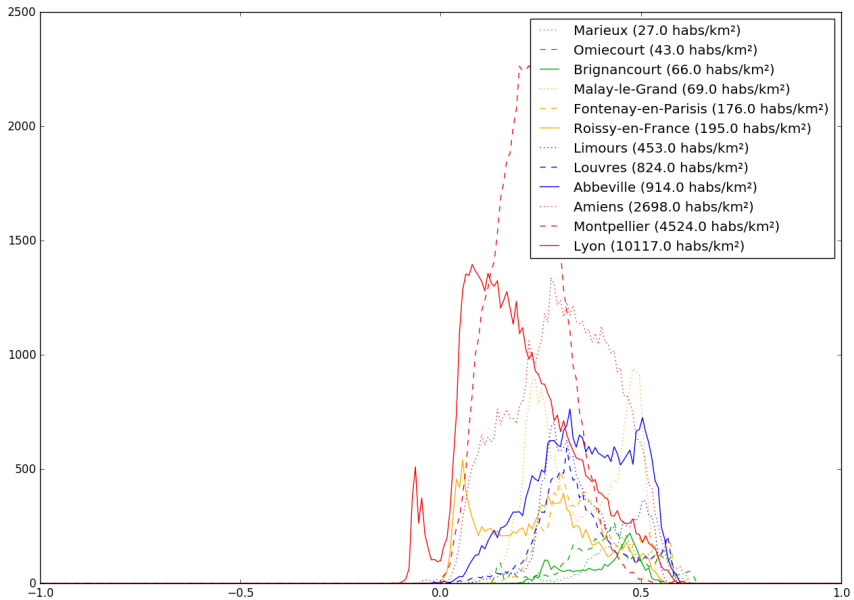


NDVI (Paris,May 2013)



NDVI (1024 bins)-histogram (Paris,May 2013)

- NDVI histogram could explain human density
  - low ndvi mode for high density (poor vegetation)
  - high ndvi mode for low density (rich vegetation)

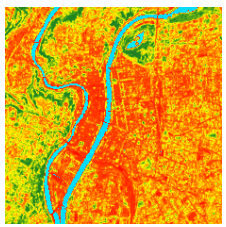


NDVI (1024 bins)-histograms (French cities, July 2013)

NDVI - Brignancourt (66 habs/km^2) - July 2013)

NDVI - Louvres (824 habs/km^2) - July 2013)

NDVI - Abbeville (914 habs/km^2) - July 2013



NDVI - Lyon (10117 habs/km^2) - July 2013)

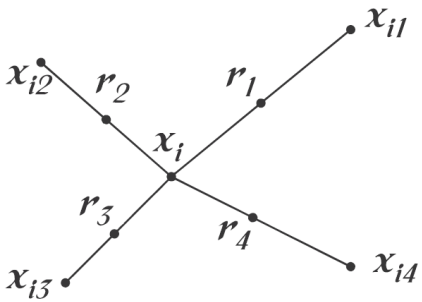
- Explanatory variables : NDVI (1024 bins)-histograms of the 34190 cities
- Predictable variable : density of the 34190 cities for regresion
- Predictable variable : density category of the 34190 cities for classification
- Principal Component Analysis over 600 components (100% of variance explained)

city	bin-1	bin-2	...	bin-511	bin-512	...	bin-1023	bin-1024	density (habs/km^2)	density (category)
Ozan	0	0	...	1	5	...	0	0	93.8	1
Cormoranche-sur-Saône	0	0	...	1	4	...	0	0	118.9	1
Paris	0	0	...	1953	1815	...	0	0	21153.9	6
Lyon	0	0	...	1099	1032	...	0	0	10117.0	5
Tours	0	0	...	268	238	...	0	0	3888.2	3
Besancon	0	0	...	97	122	...	0	0	1797.9	1
...	...	...	...	...	...	...	...	...	...	...

Very imbalanced data for classification !

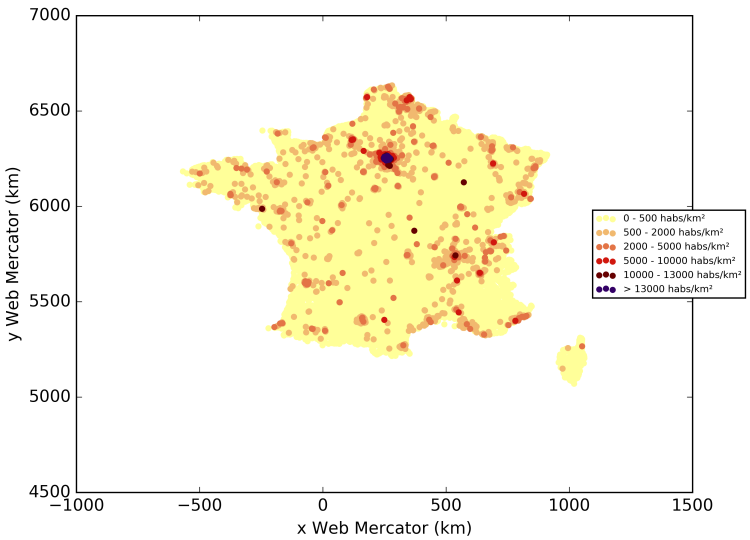
category	number of samples
1	32533
2	1252
3	288
4	78
5	15
6	24

- Overcome imbalanced data by oversampling minority classe C_i, using SMOTE technique (Synthetic Minority Oversampling TEchnique) with a factor n_1/n_i

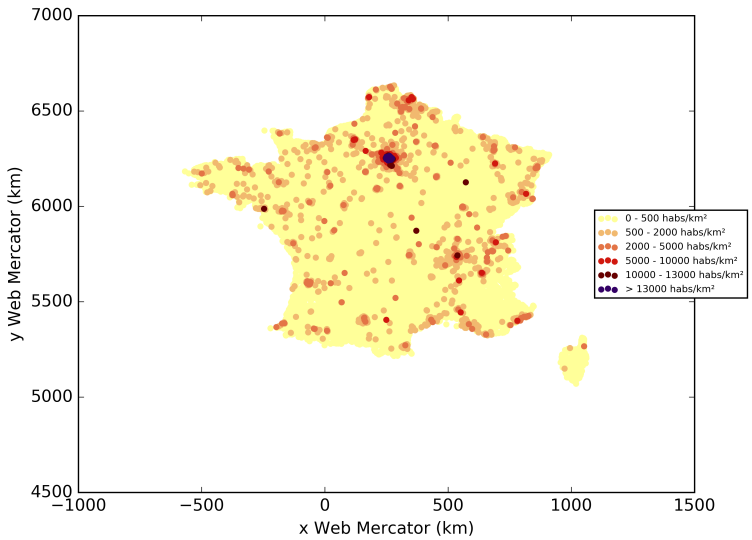


Let's x_i,x_{i1}, x_{i2}, x_{i3} and x_{i4} be points of minority class i and y_{jk} the points of majority class j , choose each new sample r_l for class i that maximizes $\sum_{k=1}^{n_j}(y_{jk}-r_l)^2$

- Best cross-validation (stratified 3-folds) of 94.74% obtained for number of neighboors $k = 5$.

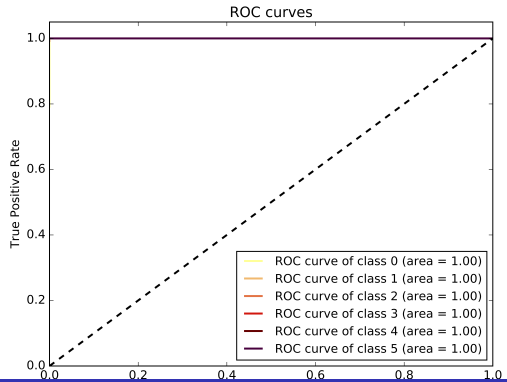


France density category - ground truth

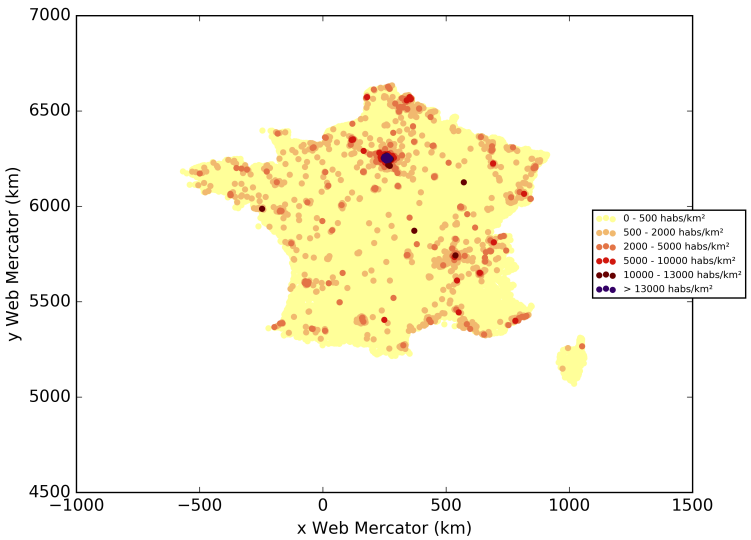


France density category - prediction

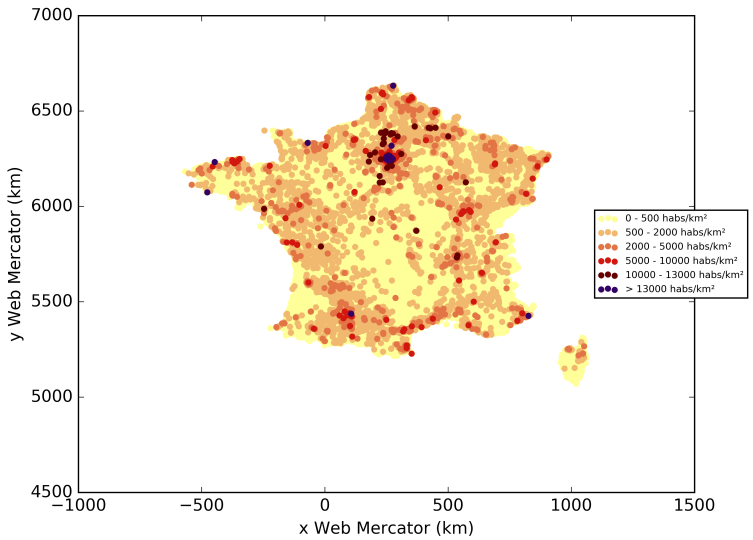
Confusion matrix after refitting						
	1	**2**	**3**	**4**	**5**	**6**
1	32533	0	0	0	0	0
2	0	1252	0	0	0	0
3	0	0	288	0	0	0
4	0	0	0	78	0	0
5	0	0	0	0	15	0
6	0	0	0	0	0	24



- Overcome imbalanced data by increasing regularization parameter C_i for minority classes i : $C_i = n/n_i$
- Best cross-validation (stratified 3-folds) of 84.07% obtained for $\gamma = 0.01$.

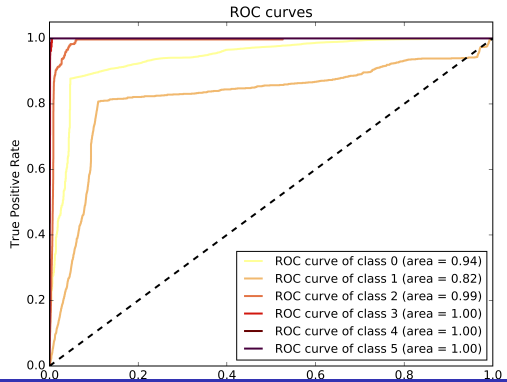


France density category - ground truth

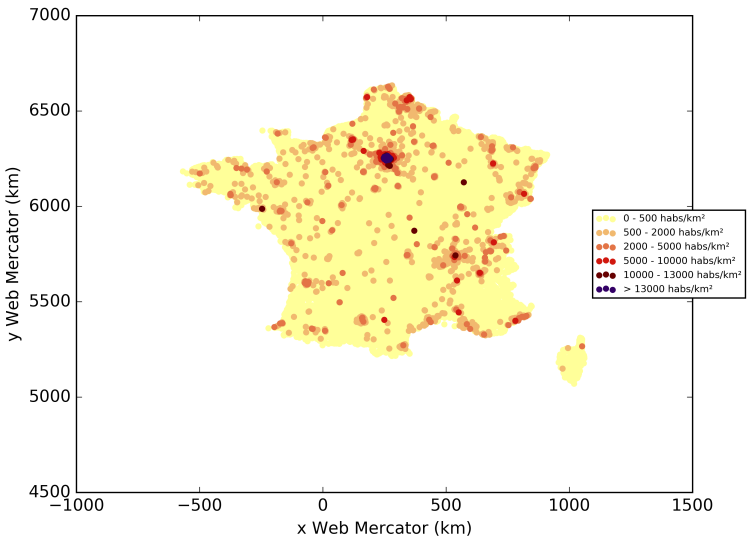


France density category - prediction

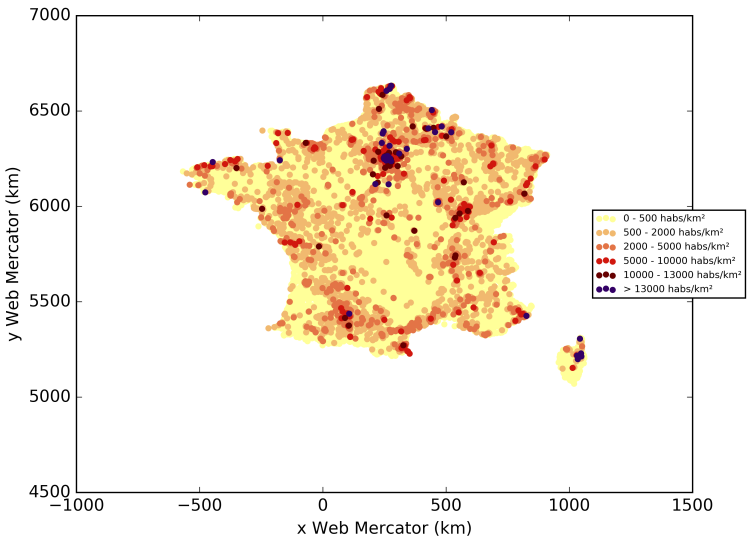
Confusion matrix after refitting						
	1	**2**	**3**	**4**	**5**	**6**
1	28569	3632	251	51	27	3
2	77	1021	139	11	1	3
3	1	25	254	7	0	1
4	0	0	3	75	0	0
5	0	0	0	0	15	0
6	0	0	0	0	0	24



- Overcome imbalanced data by majoring error by weights to amplify error on minority class : $w_i = \frac{n_1}{n_i}$.
- Best cross-validation (stratified 3-folds) of 90.46% obtained for :
  - One layer of 1200 neurons
  - Stochastic Gradient Descent learning rate of 0.001
  - Penalization of 0.001.
- Early stopping activation to avoid overfitting (i.e. stop iterative training when no more validation error decrease)

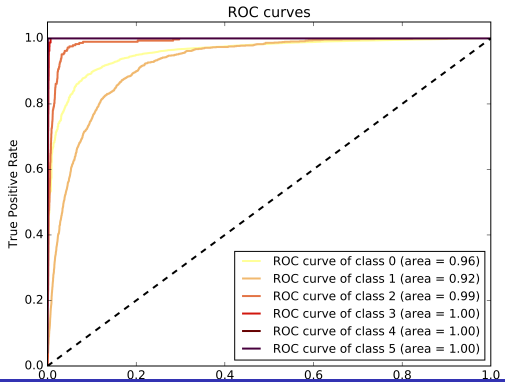


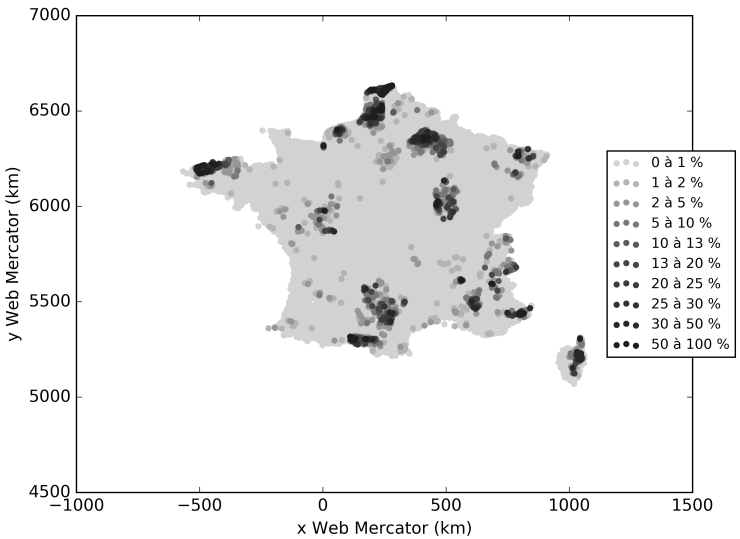
France density category - ground truth



France density category - prediction

Confusion matrix after refitting						
	1	**2**	**3**	**4**	**5**	**6**
1	29743	2281	367	95	24	23
2	207	853	164	23	2	3
3	2	36	236	9	2	3
4	0	0	2	76	0	0
5	0	0	0	0	15	0
6	0	0	0	0	0	24





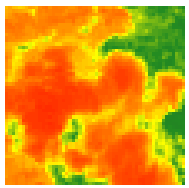
France density classification error ($y_{pred} - y_{true}$)

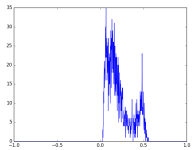
France cloudyness

- Green points shows that a lot of low densities (category 1) are predicted as high densities (categorie 5/6) (see Corsica).
- This can be related to cloudyness that provides zero values for NDVI like high densities does :

Silvareccio
cloudyness : 49.5%
true : 1
pred : 6
error : 5

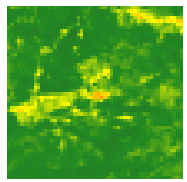

rgb

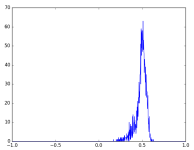

ndvi


histogram

Porri
cloudyness : 0.0%
true : 1
pred : 1
error : 0


rgb


ndvi


histogram

- Take ground truth densities and surfaces for Switzerland (2013), Belgium (2015) and Netherlands (2014).

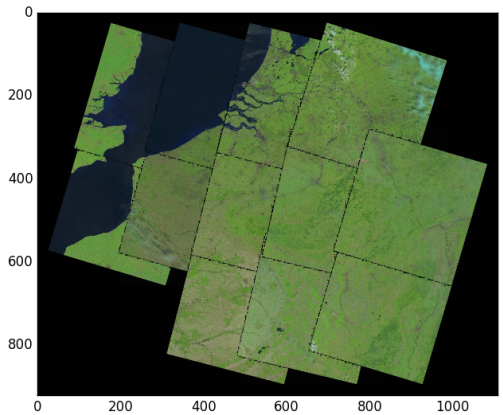
Switzerland (2013)	
category	number of samples
1	1857
2	428
3	57
4	9
5	1
6	0
total	2352

Belgium (2015)	
category	number of samples
1	406
2	154
3	14
4	7
5	1
6	7
total	589

Netherlands (2014)	
category	number of samples
1	234
2	121
3	31
4	1
5	0
6	0
total	388

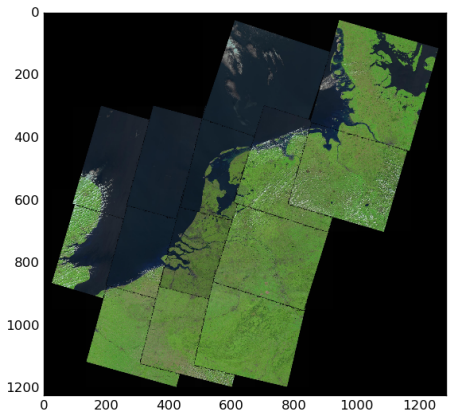
- Take longitudes and latitudes of each city using Google geolocator
- Take corresponding Landsat-8 datasets for each country in the corresponding year (between May and September, day acquisition, cloud covering $\leq 20\%$)

```
i../../data/Suiss
```

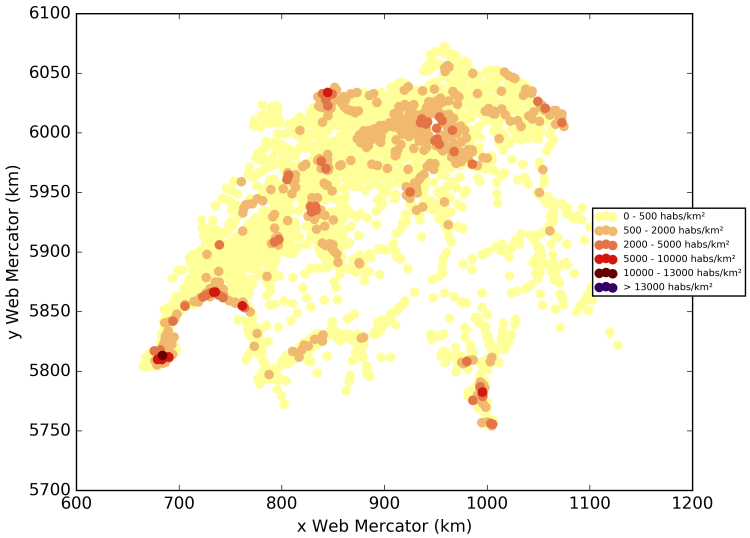


Switzerland datasets georeferenced in *Web Mercator* system

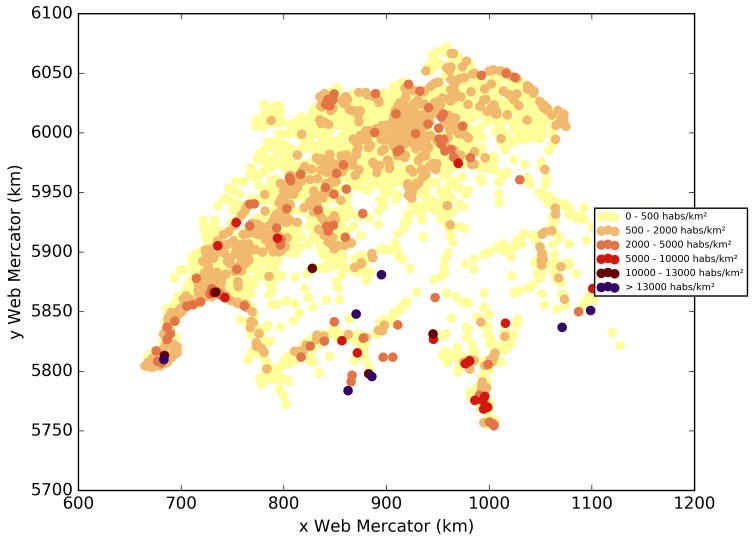
Belgium datasets georeferenced in *Web Mercator* system



Netherlands datasets georeferenced in *Web Mercator* system

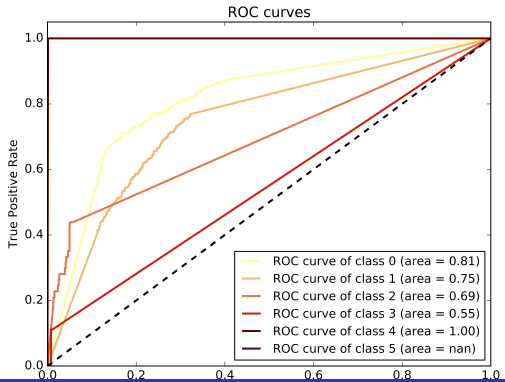


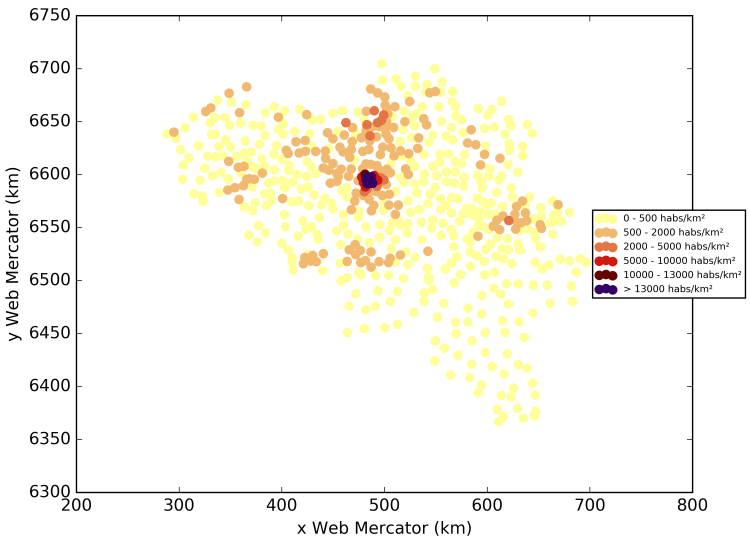
Switzerland density category - ground truth



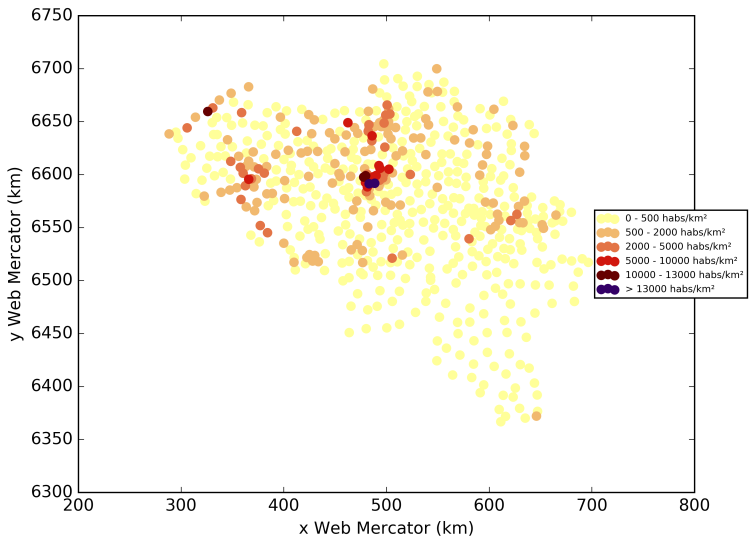
Switzerland density category - prediction

Confusion matrix						
	1	**2**	**3**	**4**	**5**	**6**
1	1411	396	33	8	3	6
2	114	268	40	6	0	0
3	1	35	16	5	0	0
4	1	4	2	0	1	1
5	0	0	0	0	1	0
6	0	0	0	0	0	0



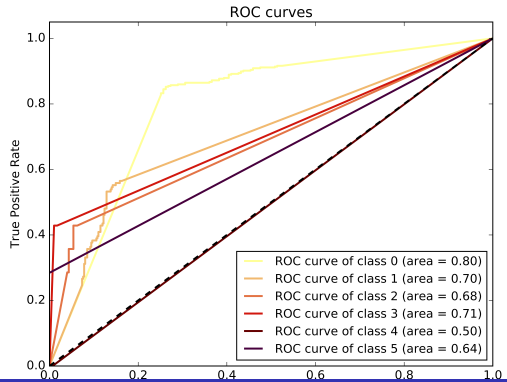


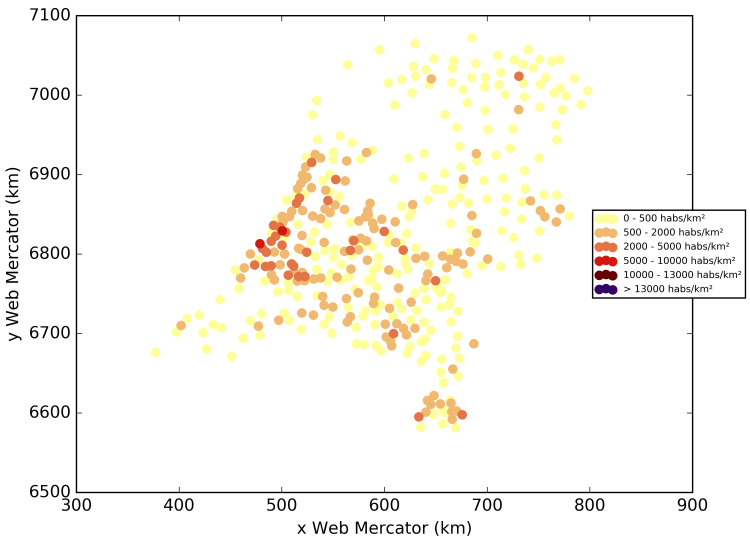
Belgium density category - ground truth



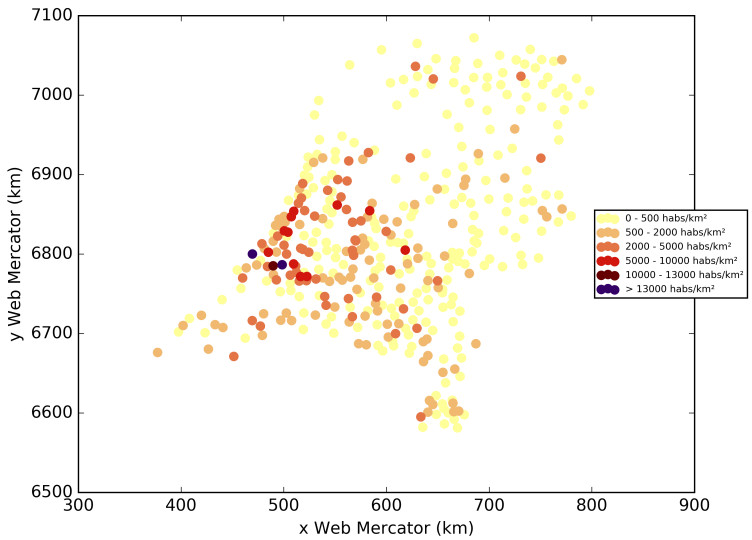
Belgium density category - prediction

Confusion matrix						
	1	**2**	**3**	**4**	**5**	**6**
1	351	48	7	0	0	0
2	59	71	19	4	1	0
3	1	6	5	2	0	0
4	0	0	2	3	2	0
5	0	0	1	0	0	0
6	0	0	2	3	0	2



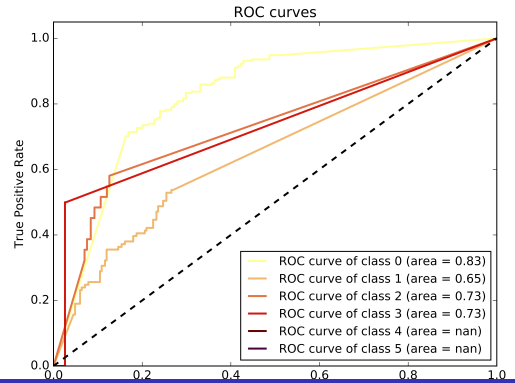


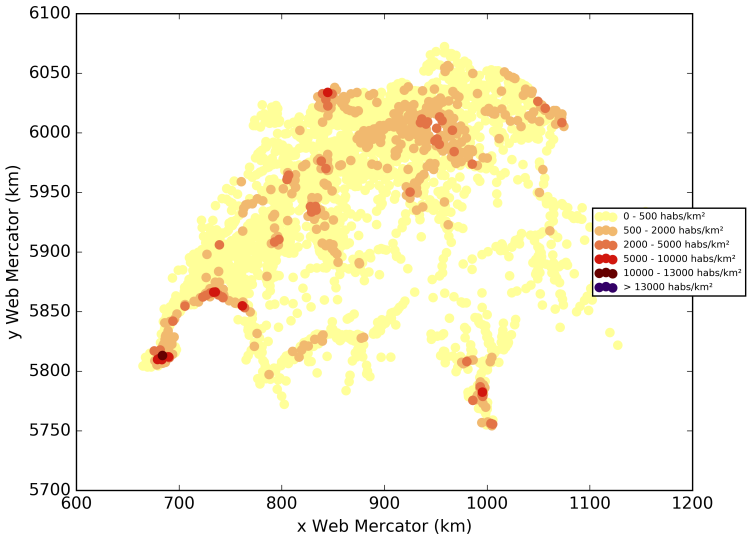
Netherlands density category - ground truth



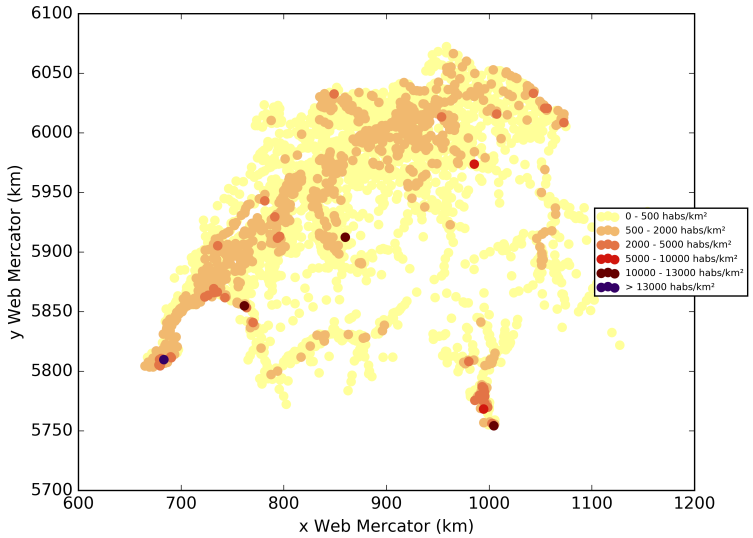
Netherlands density category - prediction

Confusion matrix						
	1	**2**	**3**	**4**	**5**	**6**
1	187	38	9	0	0	0
2	41	46	28	4	0	2
3	2	6	16	6	1	0
4	0	0	1	1	0	0
5	0	0	0	0	0	0
6	0	0	0	0	0	0



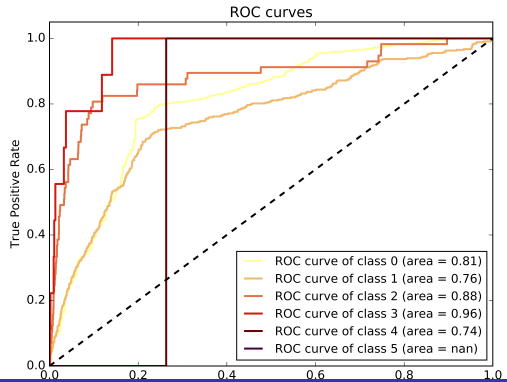


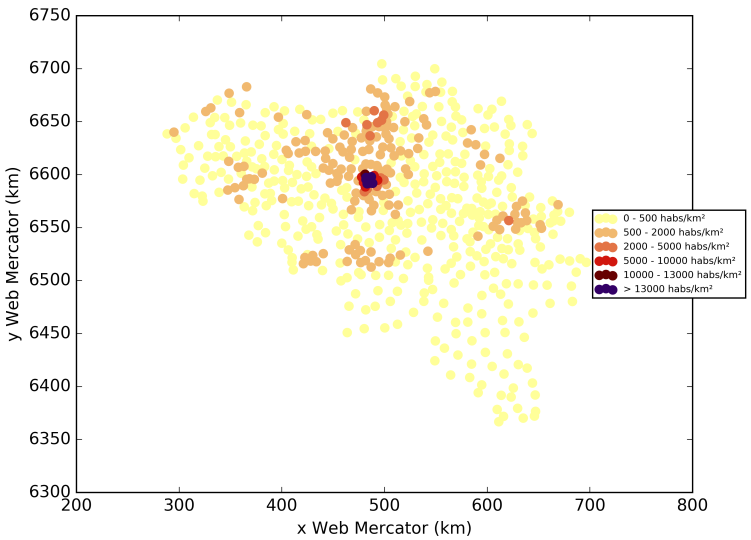
Switzerland density category - ground truth



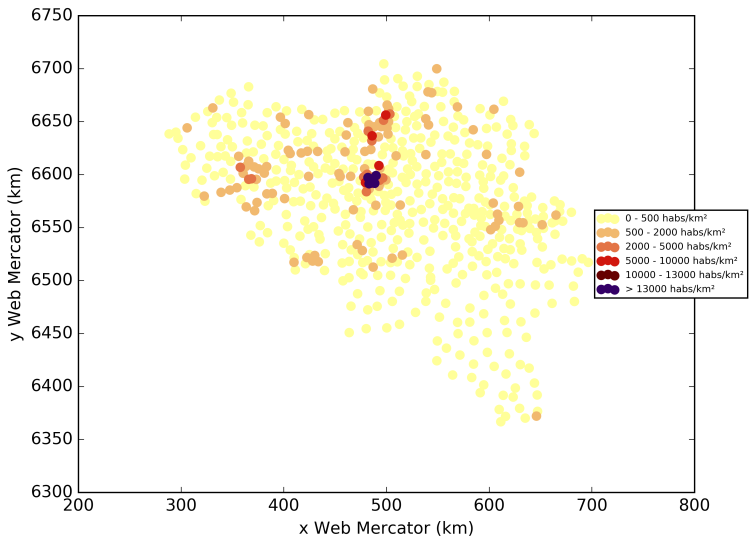
Switzerland density category - prediction

Confusion matrix						
	1	**2**	**3**	**4**	**5**	**6**
1	1471	381	3	1	1	0
2	114	296	17	0	1	0
3	5	37	14	1	0	0
4	1	0	6	0	1	1
5	1	0	0	0	0	0
6	0	0	0	0	0	0



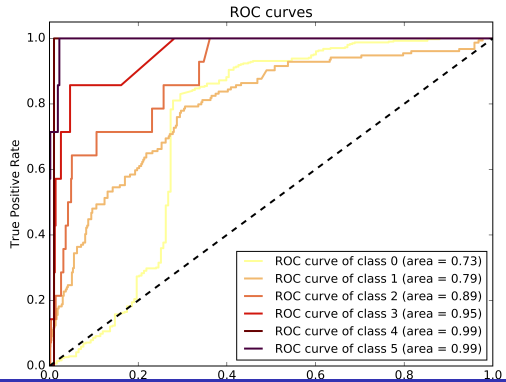


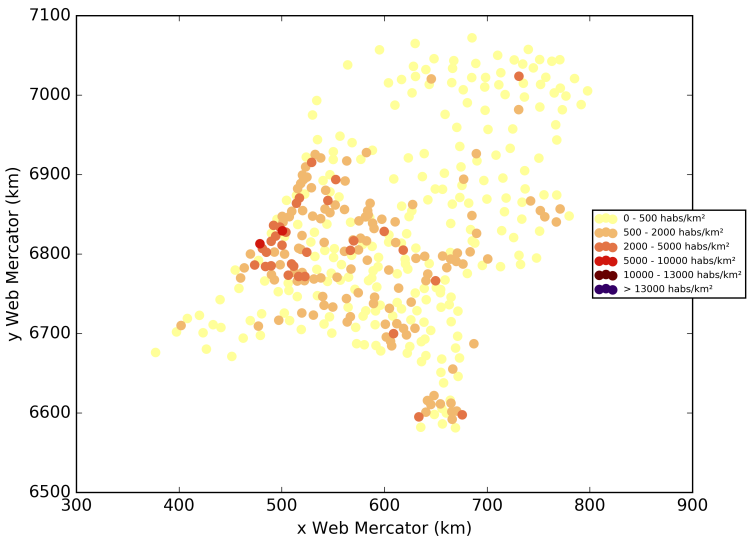
Belgium density category - ground truth



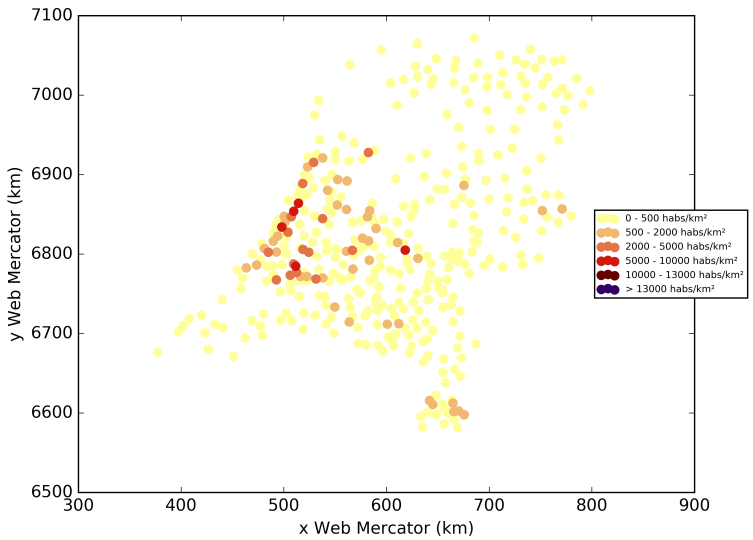
Belgium density category - prediction

Confusion matrix						
	1	**2**	**3**	**4**	**5**	**6**
1	378	28	0	0	0	0
2	88	59	6	1	0	0
3	2	7	3	2	0	0
4	1	1	3	1	0	1
5	0	0	1	0	0	0
6	1	1	0	1	0	4



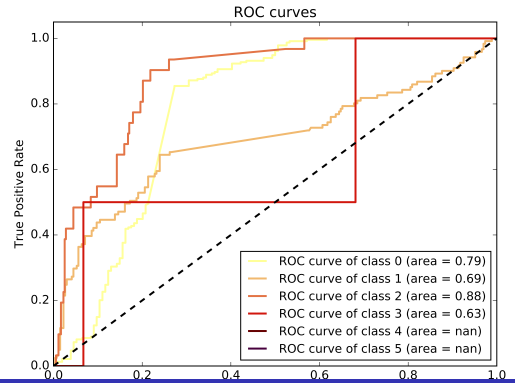


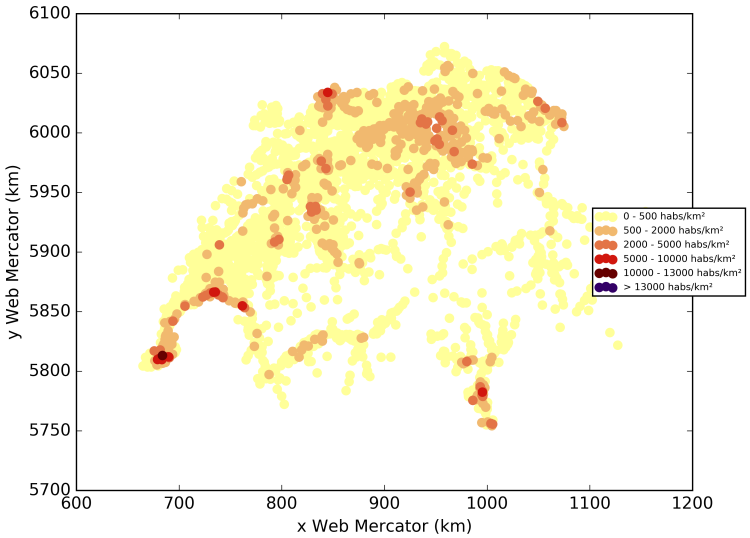
Netherlands density category - ground truth



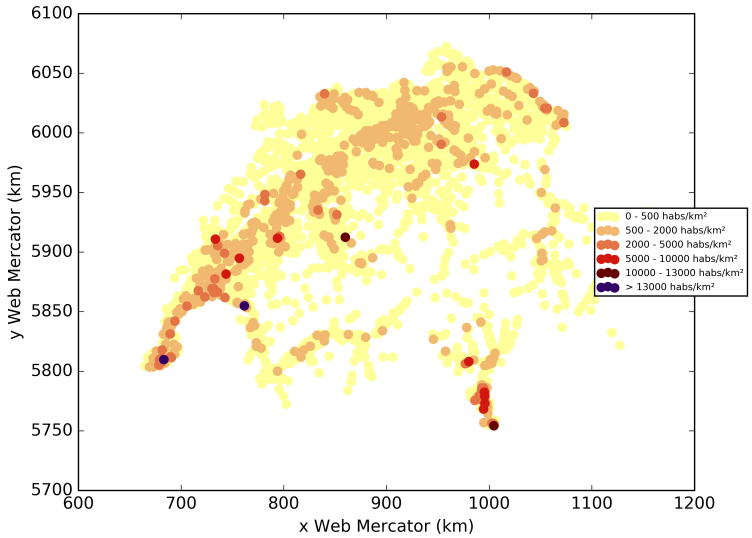
Netherlands density category - prediction

Confusion matrix						
	1	**2**	**3**	**4**	**5**	**6**
1	233	1	0	0	0	0
2	80	32	8	1	0	0
3	12	8	7	4	0	0
4	2	0	0	0	0	0
5	0	0	0	0	0	0
6	0	0	0	0	0	0



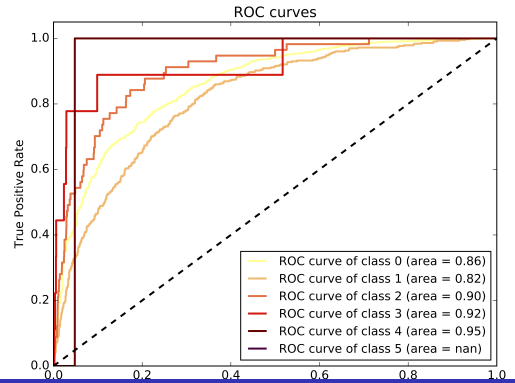


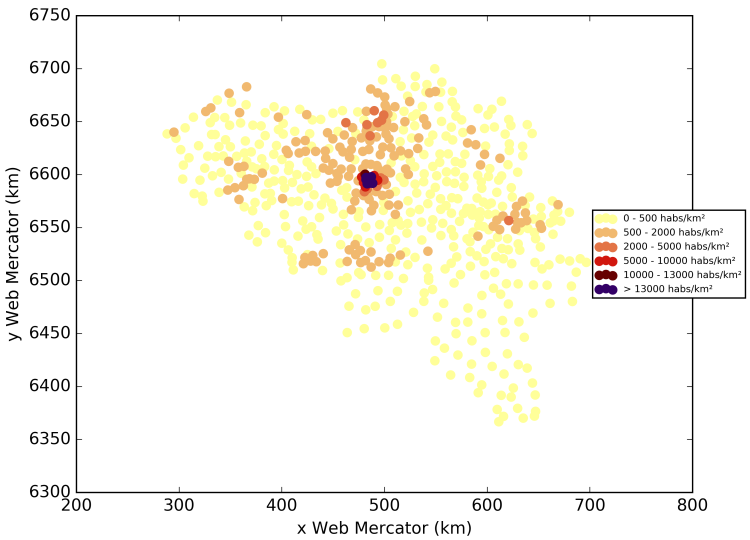
Switzerland density category - ground truth



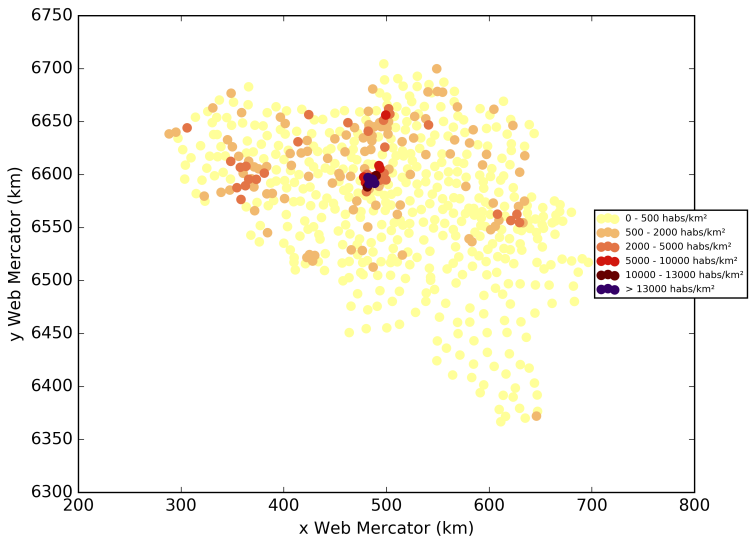
Switzerland density category - prediction

Confusion matrix						
	1	**2**	**3**	**4**	**5**	**6**
1	1601	244	7	4	1	0
2	158	249	18	2	1	0
3	2	36	16	3	0	0
4	1	0	5	1	0	2
5	1	0	0	0	0	0
6	0	0	0	0	0	0



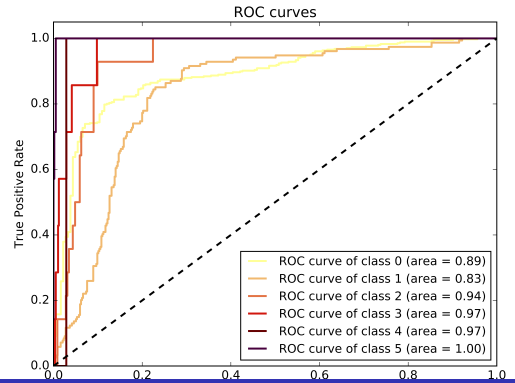


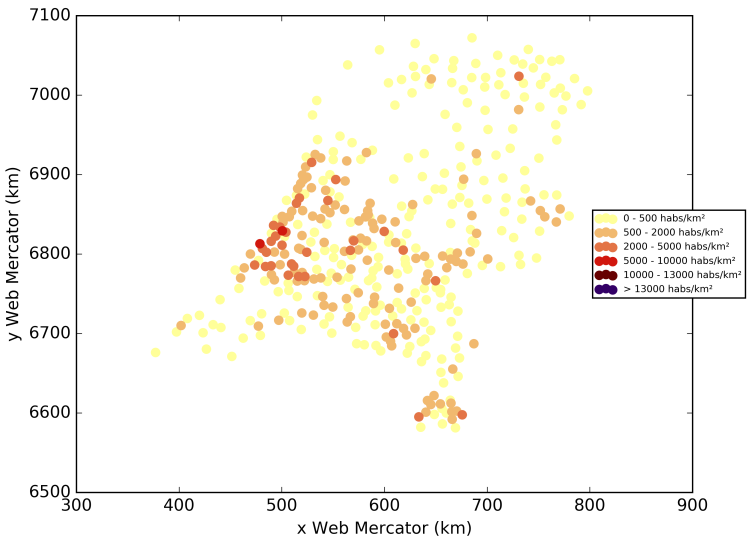
Belgium density category - ground truth



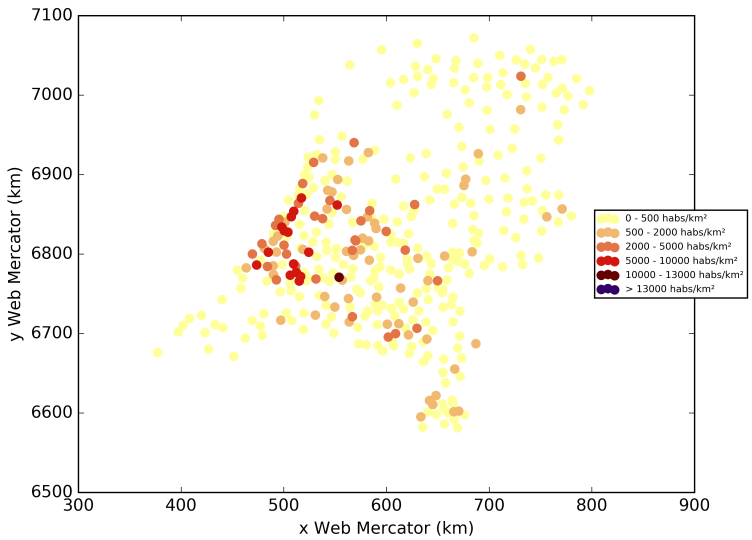
Belgium density category - prediction

Confusion matrix						
	1	**2**	**3**	**4**	**5**	**6**
1	363	41	2	0	0	0
2	73	60	19	2	0	0
3	1	6	6	1	0	0
4	1	0	2	2	2	0
5	0	0	1	0	0	0
6	0	2	0	1	0	4



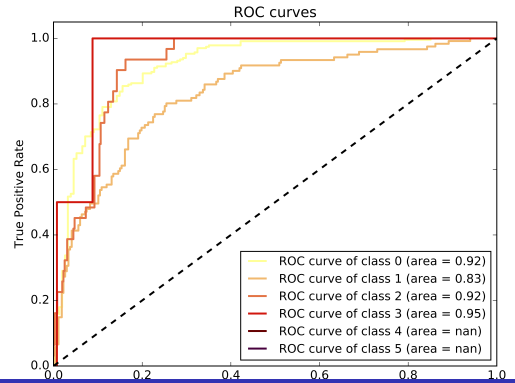


Netherlands density category - ground truth



Netherlands density category - prediction

Confusion matrix						
	1	**2**	**3**	**4**	**5**	**6**
1	277	5	2	0	0	0
2	54	47	14	5	1	0
3	2	6	14	9	0	0
4	0	0	1	1	0	0
5	0	0	0	0	0	0
6	0	0	0	0	0	0



- Training Area Under Curve

Mean AUC	
Classifier	**France**
Nearest Neighbors	1.00
Support Vector Machine (Gaussian)	0.96
Neural Network	0.98

- Testing Area Under Curve

Mean AUC			
Classifier	**Switzerland**	**Belgium**	**The Netherlands**
Nearest Neighbors	0.89	0.87	0.85
Support Vector Machine (Gaussian)	0.94	0.95	0.94
Neural Network	0.97	0.97	0.96

- Nearest Neighbors clearly overfitted (important gap between training and testing)
- Best Generalization for Neural Network

- practical improvment : Get higher dimensional features and train it using cluster-distributed features and Spark MLib or Hadoop XGBoost.
- theoretical improvment : Overcome local cloudyness with a finer NDVI. For example, indice GEMI.

$$GEMI = \eta(1 - 0.25\eta) - \frac{R - 0.25}{1 - R}$$

$$\eta = \frac{2(NIR^2 - R^2) + 1.5NIR + 0.5R}{NIR + R}$$