Combining satellite imagery and machine learning to predict atmospheric heavy metal contamination

Alexander Uzhinskiy[1], Gennady Ososkov[1], Pavel Goncharov[2], Marina Frontsyeva[1]

1. Joint Institute for Nuclear Research
2. Sukhoi State Technical University of Gomel, Belarus.

The 8th International Conference "Distributed Computing and Grid-technologies in Science and Education", Dubna 2018

Introduction

Air pollution has a significant negative impact on the various components of ecosystems, human health, and ultimately, cause significant economic damage.

More than nine out of 10 of the world's population – 92% – lives in places where air pollution exceeds safe limits, according to research from the World Health Organization (WHO).

Air pollution is the fourth-largest threat to human health, behind high blood pressure, dietary risks and smoking.
The health risks of breathing dirty air include respiratory infections and cardiovascular diseases, stroke, chronic lung disease and lung cancer.

The study by the World Bank and the Institute for Health Metrics and Evaluation (IHME) calculated the economic cost of air pollution. It found that air pollution led to one in 10 deaths in 2013, which cost the global economy about $225 billion in lost labour income.

Introduction

There are a lot of regional and international environment control programs. They uses different techniques and tools but as result they all want to understand what is the current situation and how it will evolve. Generally to get some indexes researches taking samples and analyses them. For the natural reasons sampling is carried out rarely and the dimension of the sampling grid could be very big. In such a situation modeling can be the right choice.

In our research we try to predict atmospheric heavy metal contamination by combining satellite imagery and machine learning.

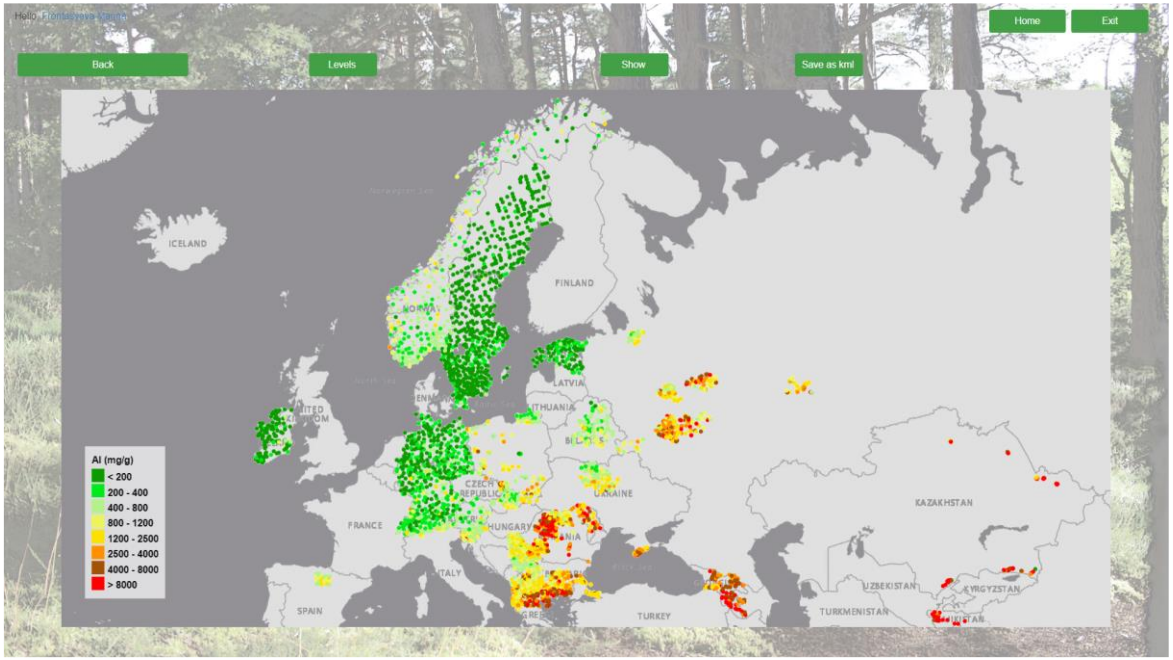
The idea is to use real-life information about heavy metals concentration and indexes taken from an satellite images to train a special statistical model. After that the model together with a new satellite indexes can be used to predict contamination for an area with better dimension at any time.

It is clear that getting the indexes from satellite images is much easer process then field sampling. Sampling and analysis for the area like Moscow Region can take 4-6 month. Gathering indexes for the same area can be done for a few days.

ICP Vegetation (Where to get real data from?)

The aim of the **UNECE International Cooperative Program (ICP)** Vegetation in the framework of the United Nations Convention on Long-Range Transboundary Air Pollution (CLRTAP) is to identify the main polluted areas of Europe, produce regional maps and further develop the understanding of the long-range transboundary pollution. Atmospheric deposition study of heavy metals, nitrogen, persistent organic compounds (POPs) and radionuclides is based on the **analysis of naturally growing mosses** through moss surveys carried out **every 5 years.**

Specialists of the Joint Institute of Nuclear Research (JINR) developed cloud platform (ICP Vegetation Data Management System, DMS, dms.jinr.ru) consists of a set of interconnected services. The platform intended to provide ICP Vegetation participants with modern unified system of collecting, analyzing and processing of biological monitoring data.



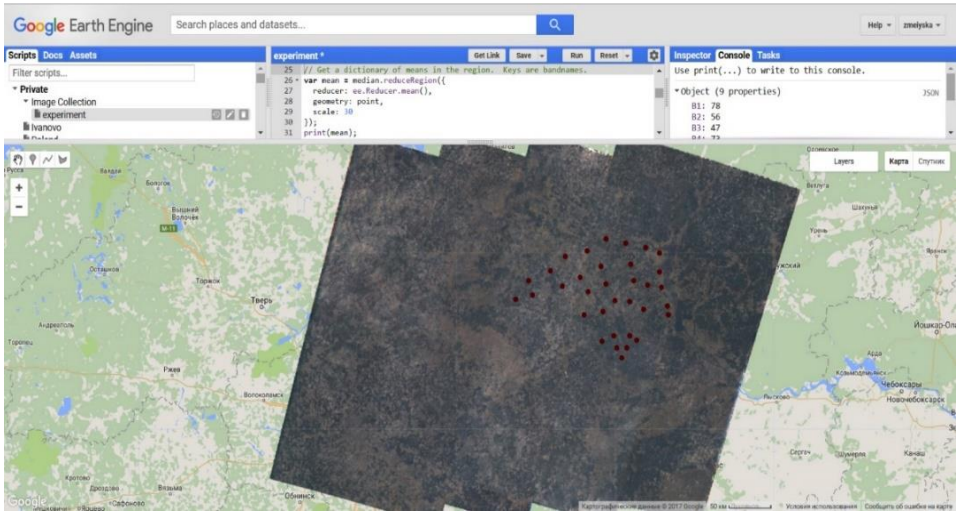
Example of the map in DMS

More than **6000** sampling sites from **40** regions of different countries are presented at the DMS now.

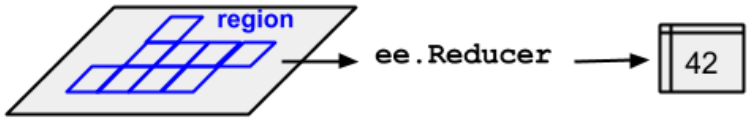
Google Earth Engine (Additional data for model)

Google Earth Engine is a cloud-based platform for planetary-scale environmental data analysis. The purpose of Earth Engine is to:

- Perform highly-interactive algorithm development at global scale
- Push the edge of the envelope for big data in remote sensing
- Enable high-impact, data-driven science
- Make substantive progress on global challenges that involve large geospatial datasets



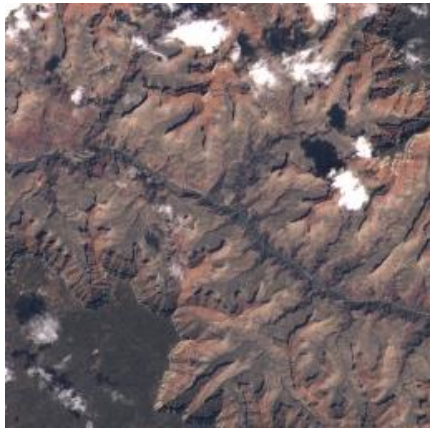
Google Earth Engine JavaScript online editor




```
var region = ee.Geometry.Rectangle(20.661, 44, 28, 48.5);

var collection = ee.ImageCollection('MODIS/006/MOD09A1')
   .filterDate('2015-01-01', '2016-12-31')
   .filterBounds(region)
   .sort('CLOUD_COVER', true);

var median = collection.median();

var area =ee.Geometry. Rectangle(21.00, 42.00, 61.01, 42.01);

// Get a dictionary of means in the region.  Keys are bandnames.
var mean = median.reduceRegion({
 reducer: ee.Reducer.mean(),
 geometry: area,
 scale: 30
});
```

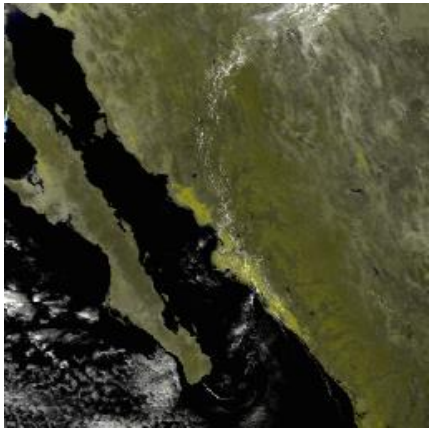
There are more than **100 satellite programs** and modeled datasets. Google Earth Engine has **JavaScript online editor** to create and verify code and **python API** to communicate with users applications.

Google Earth Engine (Programs)

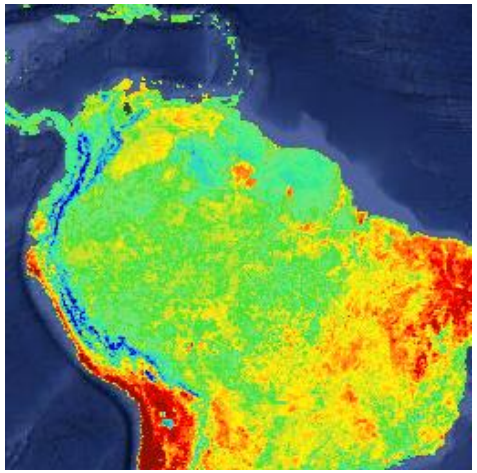


Landsat (15-30m Resolution)

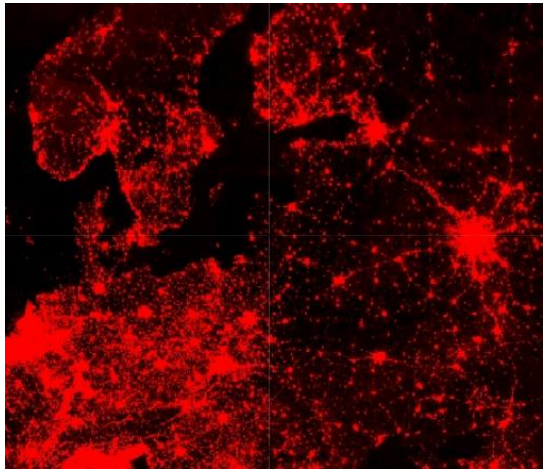
Modis (250-500m Resolution)



Sentinel (250-500m Resolution)



The MOD11A2 V6 product provides an average 8-day land surface temperature (LST) in a 1200 x 1200 kilometer grid.



Monthly average radiance composite images using nighttime data from the Visible Infrared Imaging Radiometer Suite (VIIRS) Day/Night Band (DNB).

The MOD13A2 V6 product provides two Vegetation Indices (VI): the Normalized Difference Vegetation Index (NDVI) and the Enhanced Vegetation Index (EVI).

Correlation of the contamination and satellite images

We have create a piece of software that takes coordinate of the sampling site from DMS and calculate indexes from different programs satellite images for them and then calculate the correlation between the contamination and indexes.

7 Countries: Norway, France, Germany, Sweden, Rumania, Serbia, Iceland

>20 programs: "IDAHO_EPSCOR/TERRACLIMATE","TRMM/3B42","NOAA/PERSIANN-CDR","NOAA/CDR/AVHRR/NDVI/V4","NOAA/CDR/AVHRR/LAI_FAPAR/V4","NCEP_RE/surface_temp","NASA_USDA/HSL/soil_moisture","MODIS/006/MOD17A2H","MODIS/006/MOD16A2","MODIS/006/MYD13Q1","MODIS/006/MOD13A1","MODIS/006/MOD11A1","MODIS/006/MOD09A1","MODIS/006/MCD15A3H","NOAA/CDR/AVHRR/SR/V4","MODIS/006/MYD09GA","MODIS/006/MCD43A4","LANDSAT/LE07/C01/T1_RT","COPERNICUS/S3/OLCI","MODIS/006/MOD11A2","COPERNICUS/S2","LANDSAT/LC8","VITO/PROBAV/S1_TOC_100M","VITO/PROBAV/C1/S1_TOC_333M","WHBU/NBAR_1YEAR","MODIS/MCD43A1","ASTER/AST_L1T_003","NOAA/VIIRS/DNB/MONTHLY_V1/VCMCFG","NOAA/VIIRS/DNB/MONTHLY_V1/VCMSLCFG","TOMS/MERGED", ..

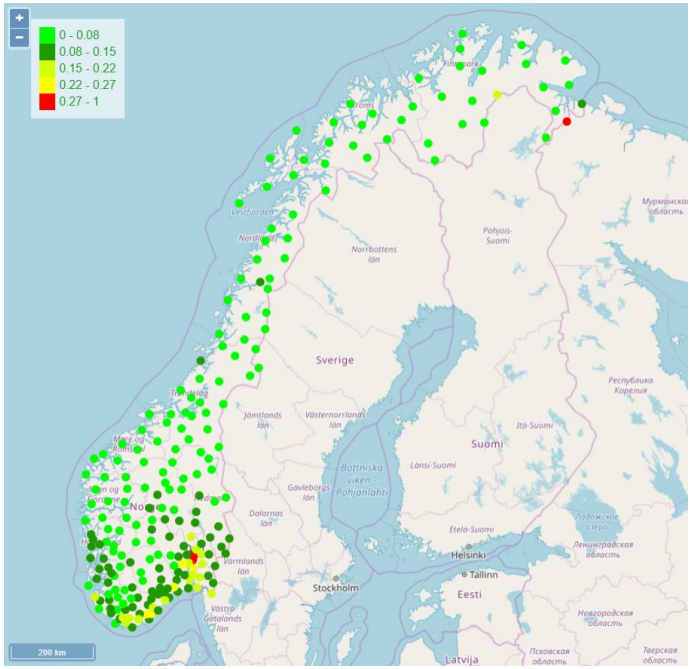
We have found connection of some elements at some countries with satellite images indexes.

Examples:

France
al SZA **0.55429233715691** :VITO/PROBAV/C1/S1_TOC_333M max 0.005
Norway
na aet **0.60229223715898** :IDAHO_EPSCOR/TERRACLIMATE min 0.001
sb LST_Day_1km **0.60905240558684** :MODIS/006/MOD11A2 max 0.005
mn PsnNet **0.5663583915194** :MODIS/006/MOD17A2H sum 0.005
Serbia
mn def **-0.59954259961496** :IDAHO_EPSCOR/TERRACLIMATE sum 0.001
na srad **0.60440015641321** :IDAHO_EPSCOR/TERRACLIMATE mean 0.01
Romania
u sur_refl_b04 **0.73149618161662** :MODIS/006/MOD09A1 mean 0.001
Iceland
sb avg_rad **0.82128040586616** :NOAA/VIIRS/DNB/MONTHLY_V1/VCMSLCFG max 0.005

But to train a model we should find 8 or more indexes with weak **cross correlation**.
And we have found it.

Experiment 1 (Sb in Norway)



We have information about Sb concentration at 228 sampling sites at Norway.

ELEMENT	RANGE	MEAN	MEDIAN	± ST.DEV.
Sb	0.0065 - 0.376	0.08	0.06695	0.06

Mostly Sb sources are anthropogenic (traffic and industrial) so we have few programs at the list that represents temperature, radiance and other signs of human population.

Here is 8 indexes that we choose to train model:

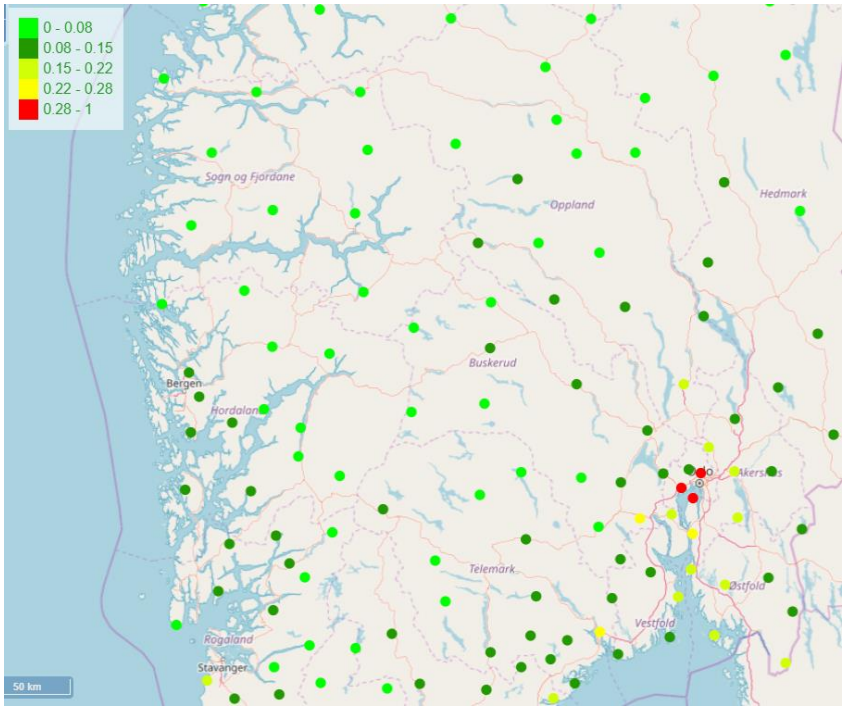
Program	Index	Area	Correlation
PROBA-V Co Level 3 Top Of Canopy Daily Synthesis at 100m resolution [PROBA-V 100m resolution]	sum(NDVI)	~ 36km2	0,636
MOD11A2.006 Land Surface Temperature and Emissivity 8-Day Global 1km [MOD11A2.006]	median(LST_Day_1km)	~ 16km2	0,628
PROBA-V C1 Level 3 Top Of Canopy Daily Synthesis at 333m resolution [PROBA-V 333m resolution]	median(SZA)	~ 6,25km2	-0,605
Sentinel-3 OLCI EFR: Ocean and Land Color Instrument Earth Observation Full Resolution [Sentinel-3 OLCI EFR]	max(Oa03_radiance)	~ 25km2	-0,57
VIIRS Nighttime Day/Night Band Composites Version 1 [VIIRS Nighttime]	max(avg_rad)	~ 16km2	0,587
USGS Landsat 7 Raw Scenes [Landsat 7]	max(B6_VCID_2)	~ 20,25km2	0,593
ASTER L1T Radiance [ASTER L1T Radiance]	max(B13)	~ 16km2	0,587
MODIS Nadir BRDF-Adjusted Reflectance, daily 500m [MCD43A4.006]	max(Nadir_Reflectance_Band5)	~ 49km2	-0,571

Models and results

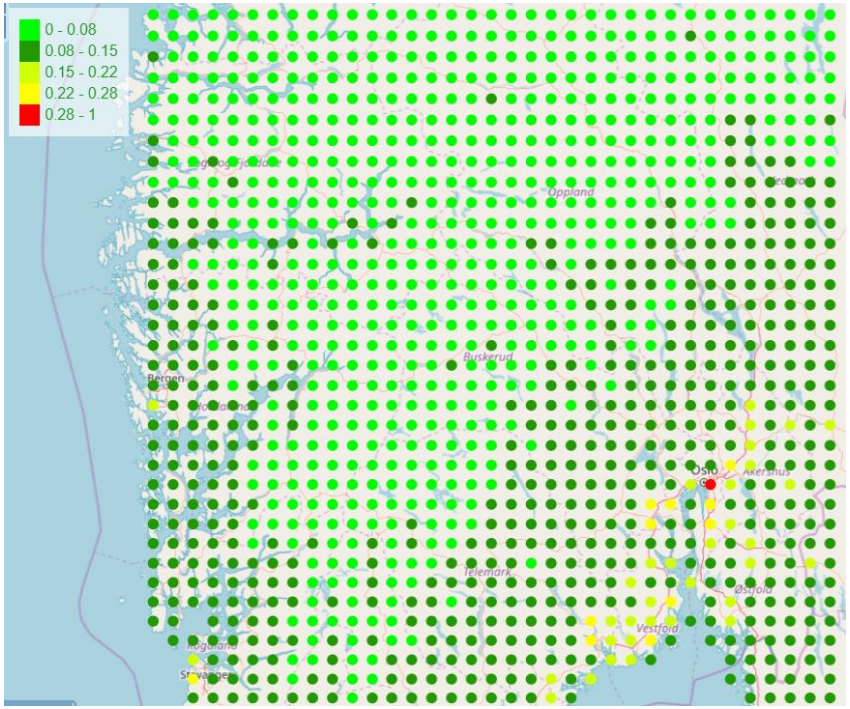
We have used 2 types of model: regression and classification. And 2 approaches for each class: neural (multilayer perceptron) and tree-based (gradient boosting, decision tree, random forest, bagging).

To find optimal parameters for tree-based model we use special parameter selection algorithm. One can get more information about models and their parameters at (**Компьютерные исследования и моделирование, 2018, том 4, Перспективы использования данных с космоснимков для прогнозирования загрязнения воздуха тяжелыми металлами**)

Visually best result shows gradient boosting.

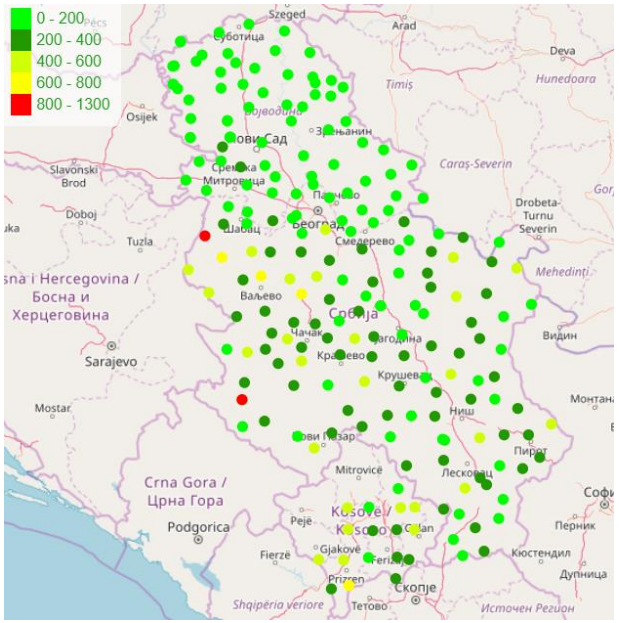


Real situation



Modeled situation

Experiment 2 (Mn in Serbia)



We have information about Mn concentration at 216 sampling sites at Serbia.

ELEMENT	RANGE	MEAN	MEDIAN	± ST.DEV.
Mn	17.5933 - 1136.1082	226	188.83565	190.31

Sources of Mn is not too clear it could have nature, agriculture or industrial origin. As one can see in the experiment analyzed areas was much smaller then for an Sb at Norway.

Here is 9 indexes that we choose to train model:

Program	Index	Area	Correlation
Monthly Climate and Climatic Water Balance for Global Terrestrial Surfaces, University of Idaho [IDAHO_EPSCOR/TERRACLIMATE]	min(def) min(pr) max(soil)	~ 0,3 km2 ~ 1,5 km2 ~ 0,3 km2	-0.601 0.552 0.564
Sentinel-3 OLCI EFR: Ocean and Land Color Instrument Earth Observation Full Resolution [COPERNICUS/S3/OLCI]	mean(Oa21_radiance)	~ 0,9km2	0,556
NASA-USDA Global Soil Moisture Data [NASA_USDA/HSL/soil_moisture]	min(ssm)	~ 0,9km2	-0,54
MOD17A2H.006: Terra Gross Primary Productivity 8-Day Global 500m [MODIS/006/MOD17A2H]	sum(PsnNet)	~ 3km2	0,585
PROBA-V C1 Top Of Canopy Daily Synthesis 333m [VITO/PROBAV/C1/S1_TOC_333M]	max(SAA) max(VNIRVZA)	~ 0,9km2 ~ 0,9km2	-0,547 0,53
NOAA CDR AVHRR LAI FAPAR: Leaf Area Index and Fraction of Absorbed Photosynthetically Active Radiation[NOAA/CDR/AVHRR/LAI_FAPAR/V4]	max(FAPAR)	~ 3km2	0,563

We removed 2 indexes with correlation ~0.52 from experiment to get better results

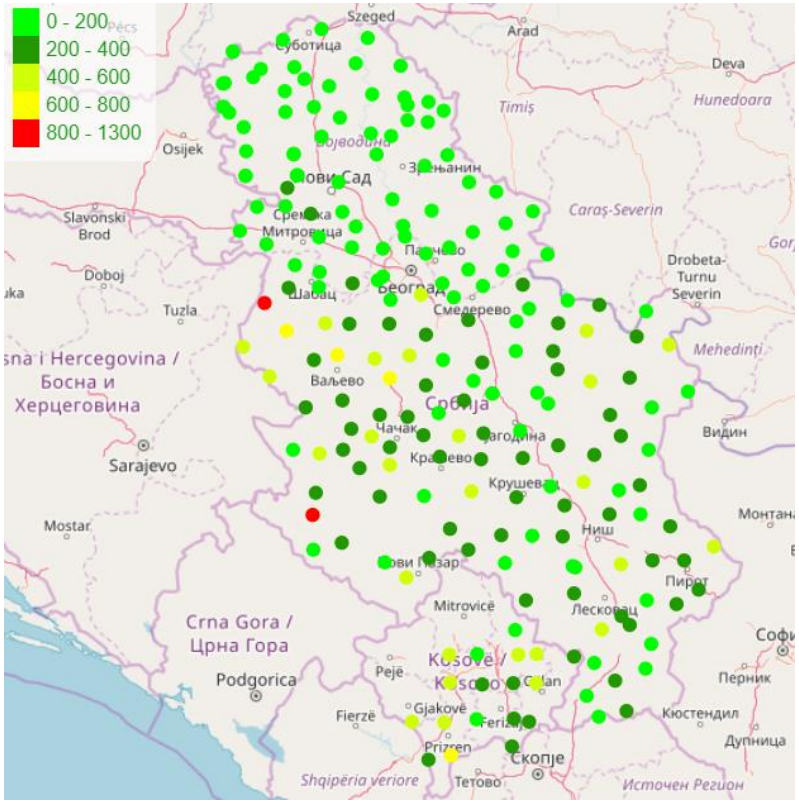
Models and results

We used our old models and gradient boosting was best again.

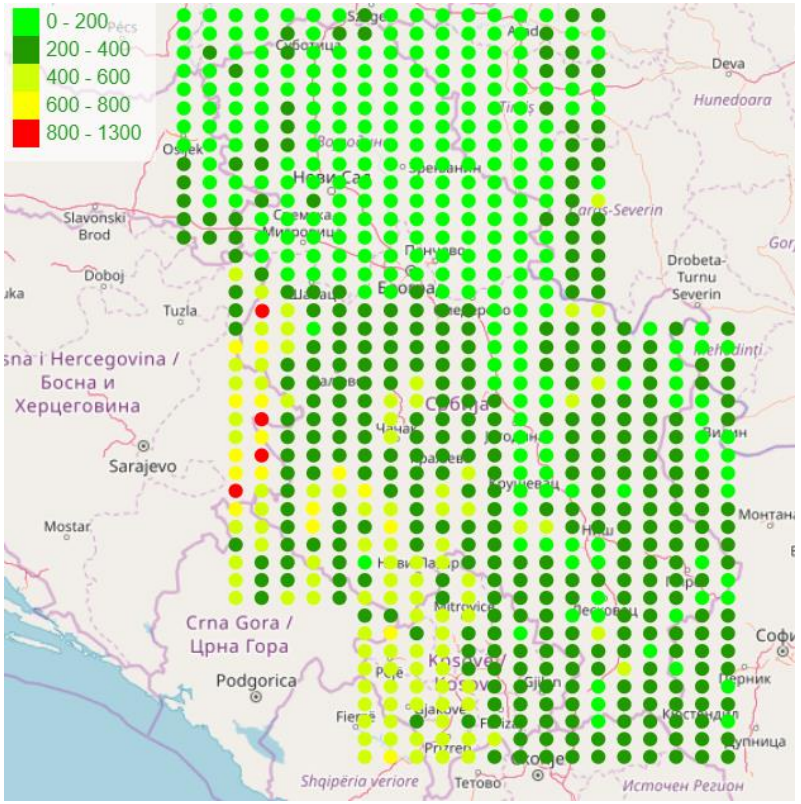
Than we have try few new models and improvements. We have applied minmax and robust normalization to the indexes and to the concentration. Surprisingly neural regression model with minmax normalization shows nice results but more experiments needed.

Regression models was better then classification one.

We also try seamese network for classification task but predictions was really bad.



Real situation



Modeled situation

Conclusion and plans

- Indexes from satellite images combined with special statistical models can be used to predict atmospheric contamination of some heavy metal at some regions.
- The connection between elements and indexes varies from region to region so there is no unified solution or unified model.
- Modeling can open new horizons for contamination analysis. Researchers will be able to:
  - monitor the evaluation of situation when it needed,
  - get detailed information about areas of interests,
  - check the situation in the cross border areas,
  - partly automate environment control process (automatically run the model and get notification when contamination level is higher then critical level) .

We are working on mechanisms to verify and select best model for an region automatically.

We will keep on searching connection of the contamination and satellite indexes and testing new models and approaches.

We will try to find new sources of the satellite images indexes to verify Google Earth Engine but mostly to improve scalability of our solutions.

Thank you for your attention