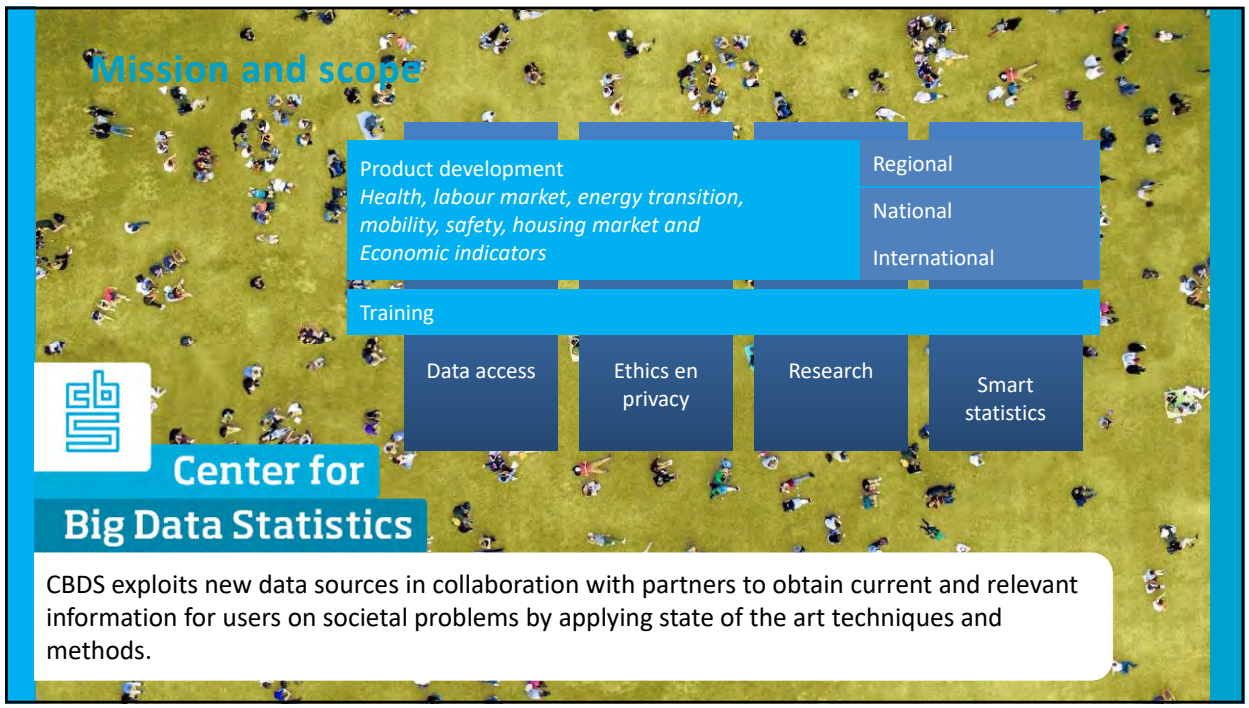
To what extent can CBS benefit from Earth observation to monitor and report on local to national level

Synergistic use of EO and Register data to evaluate build-up area in the Netherlands

Curier R. Lyana,

1



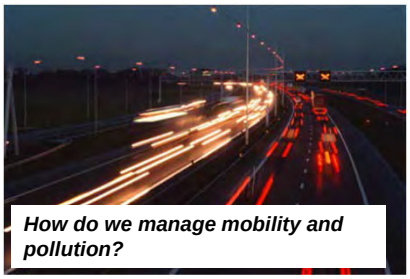
2

Answering complex policy questions

How do we tackle the energy transition?

Can we make early warning systems for refugee streams?

How do we match offer and demand in the labour market?



How do we manage mobility and pollution?

Center for Big Data Statistics

3

Earth Observation (EO) datasets

- Earth Observation (EO) datasets offers many opportunities to improve the monitoring of these SDGs for both reaching the SDG targets and reporting on progress.
- EO allows for tracking of global change at high resolution and in real time.
- Geospatial information provided by EO data will allow for implementation at local to national levels.

  - MODIS
  - LANDSAT 8
  - SENTINEL-1/2
  - SUPERVIEW

Center for Big Data Statistics

4

4

Delimit the built-up area of the urban agglomeration:

Context

❑ESSnet Big Data is a project within the European statistical system (ESS) integration of big data in the regular production of official statistics, through pilots exploring the potential of selected big data sources, and through building and implementing concrete applications

Why

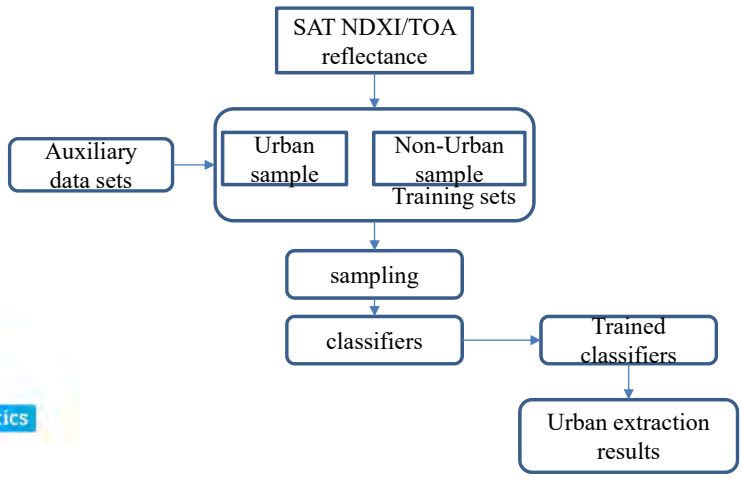
- ❑ Evaluate Urban Sprawl
- ❑ SDG 11.3.1 : Ratio of land consumption rate to population growth rate
- ❑ SDG 11.7.1 : Average share of the built-up area of cities that is open space for public use for all, by sex, age and persons with disabilities

Center for Big Data Statistics

5

Delimit the built-up area of the urban agglomeration

First step: assess the robustness of state-of-the art classifier to distinguish between artificial surfaces and others surfaces type using high spectral, spatial and temporal resolution at national scale.



Center for Big Data Statistics

6

Area of Interest

Random LandSat scene for April 2015 was resample to the Training domain. 30 m pixels.

Training and Testing

Random subset of 5 000 000 pixels
70% Trainning
30% Testing

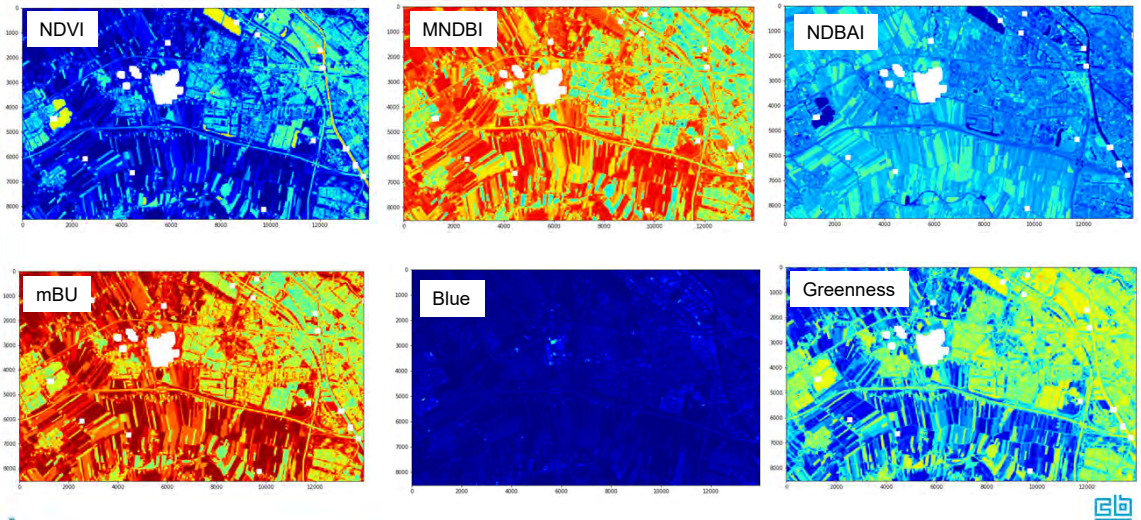
Blind Test

Center for Big Data Statistics

7

7

Datasets: Landsat data



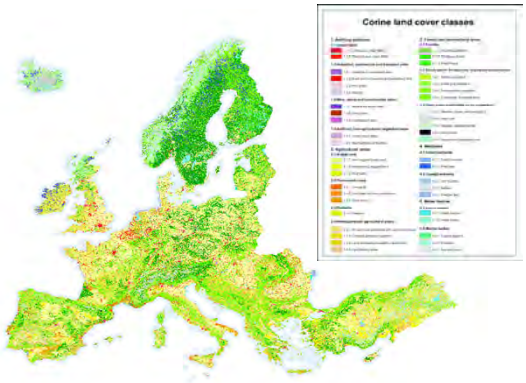
Center for Big Data Statistics

8

8

Datasets: Administrative sources

- BAG (Building Polygons, Address, Status, Function)
- Land cover register (national and European)



Center for Big Data Statistics

These sources are used for training and validation purposes.

9

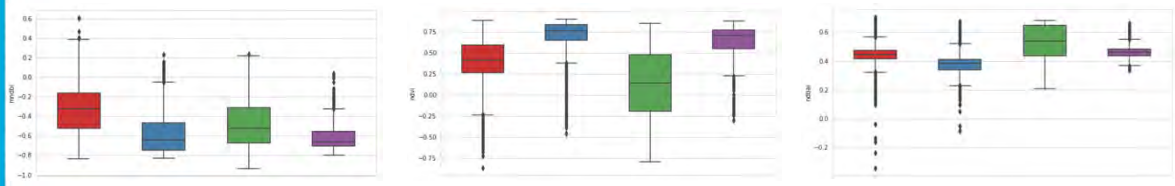
9

Datasets: NDXI distribution per class

Only main categories are kept

Agricultural areas Forest semi natural areas Artificial Surfaces Water bodies Wetlands

MNDBI, NDVI and NDBAI for Training area



Agricultural areas Artificial Surfaces Forest semi natural areas Water bodies

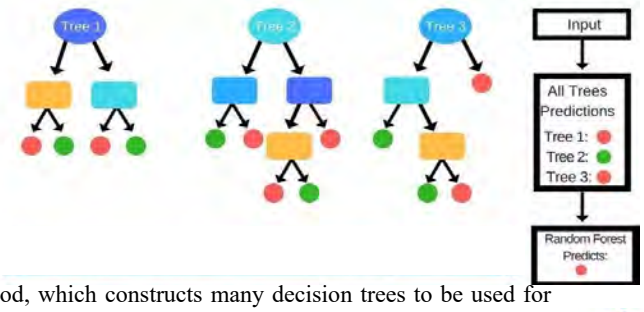
Center for Big Data Statistics

10

10

Methods

1. Three traditional classification algorithms were considered:
   - ❑ k-nearest-neighbors (k-NN),
   - ❑ support vector machine (SVM)
   - ❑ random forests (RF).
2. Deep Learning



ensemble method, which constructs many decision trees to be used for classifying a new instance by the majority vote. Each decision tree node uses a subset of attributes randomly selected from the original set of attributes.

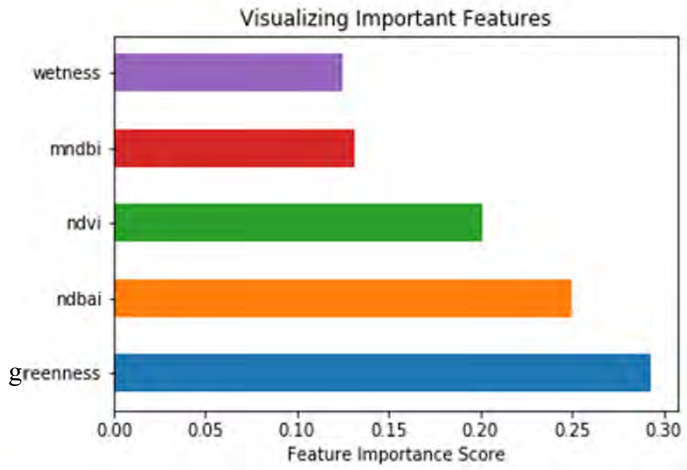
Additionally, each tree uses a different bootstrap sample data

Center for Big Data Statistics

11

11

Results: Training and Testing



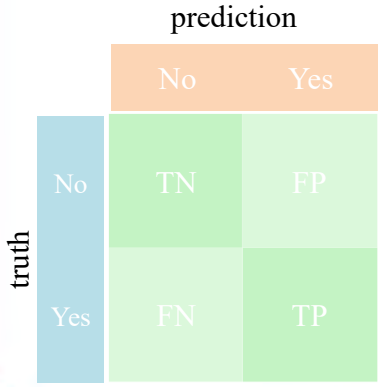
Center for Big Data Statistics

12

12

Results: Evaluation

The metrics of accuracy, precision, recall and F1-score were chosen.



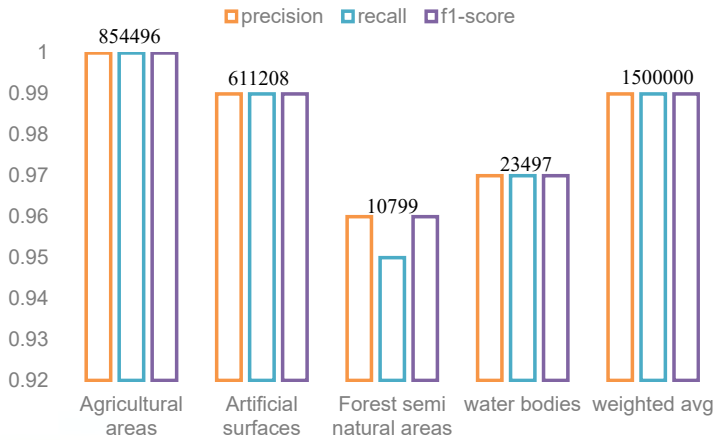
accuracy = TN + TP / TN + FP + FN + TP
precision = TP / TP + FP
recall = TP / FN + TP
F1-score is the harmonic mean between the recall and precision, it reaches its best value at 1 and worst at 0.

Center for Big Data Statistics

13

13

Results: Training and Testing



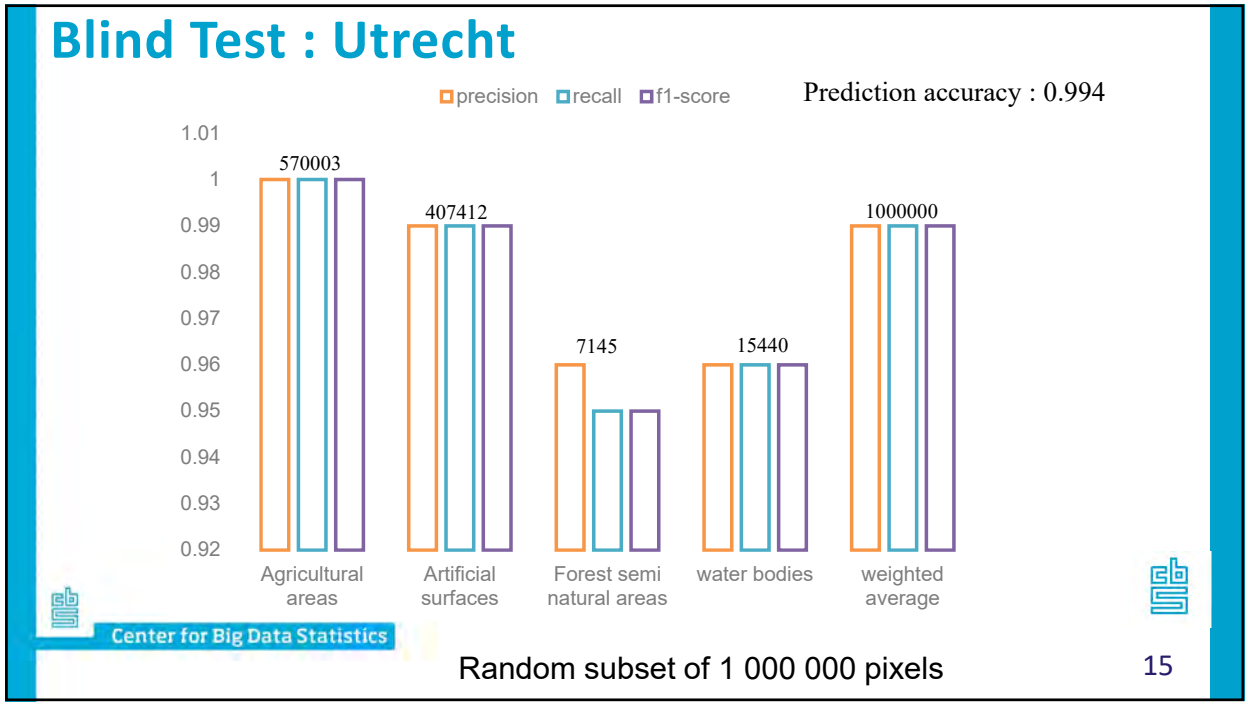
Overall testing accuracy : 0.994

Overfitting?
K-Fold cross-validation says, no…

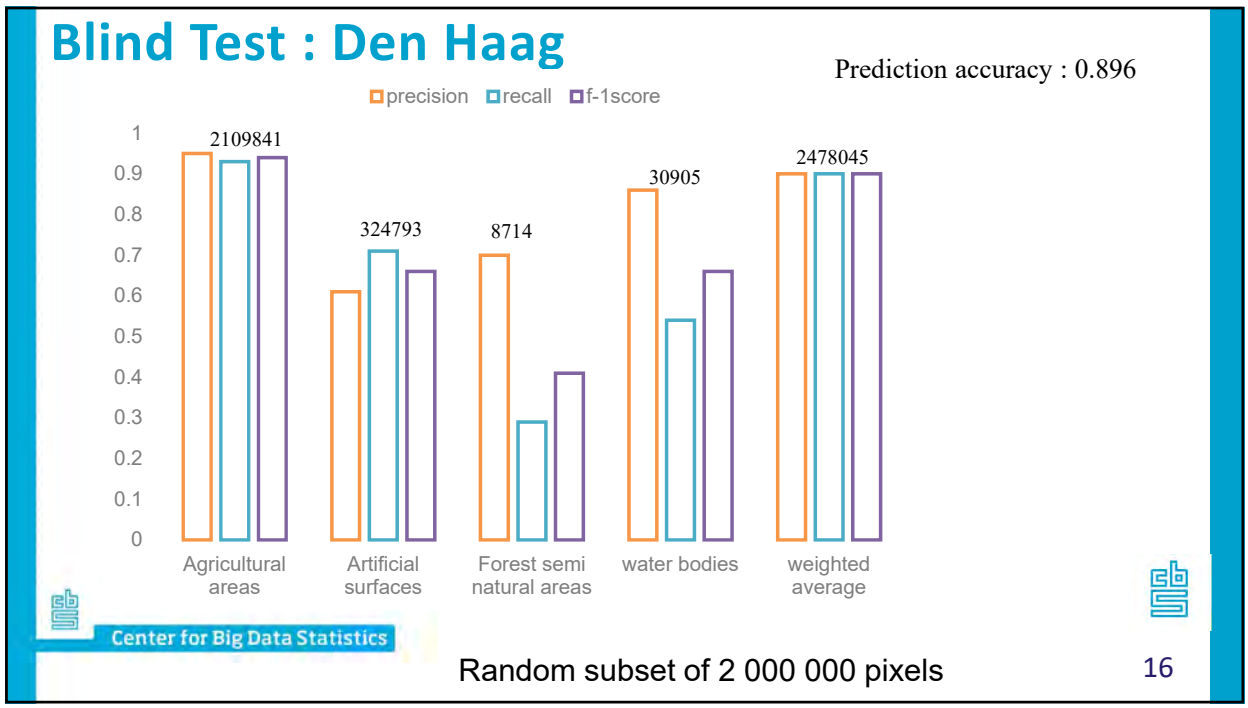
Center for Big Data Statistics

14

14


Blind Test : Utrecht
precision
recall
f1-score
Prediction accuracy : 0.994
1.01
1
0.99
0.98
0.97
0.96
0.95
0.94
0.93
0.92
570003
407412
7145
15440
1000000
Agricultural areas
Artificial surfaces
Forest semi natural areas
water bodies
weighted average
Center for Big Data Statistics
Random subset of 1 000 000 pixels
15

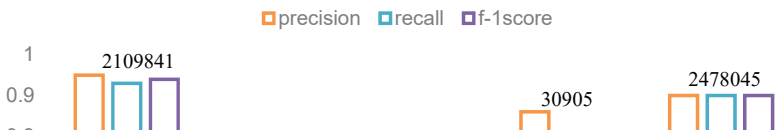
15


Blind Test : Den Haag
precision
recall
f-1score
Prediction accuracy : 0.896
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
2109841
324793
8714
30905
2478045
Agricultural areas
Artificial surfaces
Forest semi natural areas
water bodies
weighted average
Center for Big Data Statistics
Random subset of 2 000 000 pixels
16

16

Blind Test : Den Haag

Prediction accuracy : 0.896



First Findings

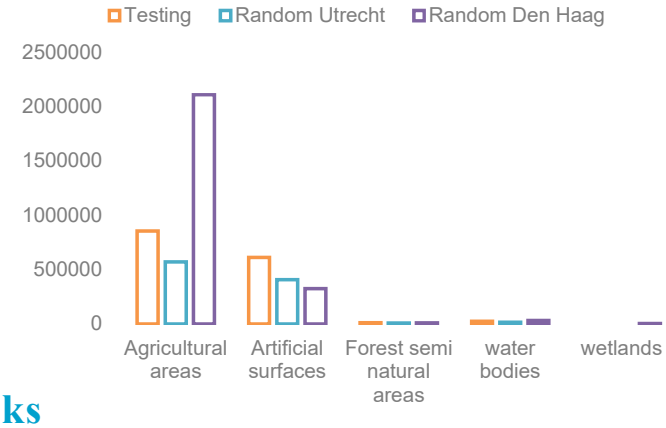
- A random forest classifier for land cover mapping using LandSat NDXI was trained for the Netherlands.
- Although some overfitting was observed, the blind test experiment shows that agricultural areas and artificial surfaces could be confidently identify and correctly classified.

Center for Big Data Statistics

Random subset of 2 000 000 pixels

17

17



Further tasks

- Retrain the classifier while imposing a cost penalty on the minority class misclassification can be useful. For example, a class weight from the ratio between the number of dataset in agricultural areas and the number of the dataset in the other classes.
- Delimiter the build-up area to evaluate year-to-year changes

Center for Big Data Statistics

18

19