NIBIO
NORWEGIAN INSTITUTE OF
BIOECONOMY RESEARCH

Spatial Big Data
Application examples from NIBIO

NIBIO REPORT | VOL. 7 | NO. 156 | 2021

Jonathan Rizzi[1], Misganu Debella-Gilo[1], Åsmund Ertshus Mathisen[2], Tor-Einar Skog[3]
[Div. Survey and Statistics, [1]Dept. Geomatics - [2]Dept. Land Inventory] [[3]Div. Biotechnology and Plant Health, Dept. Fungal Plant Pathology in Forestry, Agriculture and Horticulture]

TITTEL/TITLE

Spatial Big Data. Application examples from NIBIO

FORFATTER(E)/AUTHOR(S)

Jonathan Rizzi, Misganu Debella-Gilo, Åsmund Ertshus Mathisen, Tor-Einar Skog

DATO/DATE:	RAPPORT NR./ REPORT NO.:	TILGJENGELIGHET/AVAILABILITY:	PROSJEKTNR./PROJECT NO.:	SAKSNR./ARCHIVE NO.:
03.10.2021	7/156/2021	Open	11028	19/00827

ISBN:	ISSN:	ANTALL SIDER/ NO. OF PAGES:	ANTALL VEDLEGG/ NO. OF APPENDICES:
978-82-17-02917-5	2464-1162	23	

OPPDRAGSGIVER/EMPLOYER:	KONTAKTPERSON/CONTACT PERSON:
NIBIO	Jonathan Rizzi

STIKKORD/KEYWORDS:	FAGOMRÅDE/FIELD OF WORK:
Stordata	Geomatikk, IKT
Big Data	Geomatics, ICT

SAMMENDRAG/SUMMARY:

Rapporten dokumenterer utvalgte eksempler på bruk av *stordata* (engelsk: *big data*) teknologi og metode i NIBIO. Det første eksemplet er knyttet til oppdatering av arealressurskartet AR5, hvor det undersøkes om stordata-tilnærming kan benyttes for å identifisere lokaliteter der kartet må oppdateres. De neste eksemplene er hentet fra fagområdet plantehelse og tar for seg mulighetene for å bruke stordata-metode for å bedre prediksjonsmodeller og gjenkjenning av for skadegjørere.

LAND/COUNTRY: Norway

FYLKE/COUNTY:

KOMMUNE/MUNICIPALITY:

STED/LOKALITET:

GODKJENT /APPROVED	PROSJEKTLEDER /PROJECT LEADER
Hildegunn Norheim	Jonathan Rizzi
NAVN/NAME	NAVN/NAME

NIBIO
NORWEGIAN INSTITUTE OF BIOECONOMY RESEARCH

Preface

Most activities of NIBIO require that geospatial information of various type is available or generate new spatial information as output. The organization handles large amounts of data, and new information is generated as output leading to a rapid increase of the volume of data. It is therefore vital for NIBIO to have the most advanced competences to deal with geospatial big data. In 2018 NIBIO started an internal competence project to build knowledge in this sector across all the divisions of the institute.

NIBIO is already using big data technologies and methodologies in several departments, and particularly in the centres for precision agriculture and the centre for precision forestry. These centres also have researchers with competences on the subject. However, until 2018 there was no coordination and little communication across divisions, which is leading to limited exploitation of the huge potential already available in NIBIO as well as a loss of efficiency due to replication of activities, data, or processes. The "*Stordata*" project, hosted by the Geomatics department under the Survey and Statistics Division is aiming to better link people and activities about big data for the benefit of the entire NIBIO organization.

The aim of this document is to present selected examples of big data applications in NIBIO.

In addition to the authors of the report, other authors contributed with Chapter 3:

- Therese W. Berge (Division of Biotechnology and Plant Health, Department of Fungal Plant Pathology in Forestry, Agriculture and Horticulture)
- Håvard Eikemo (Division of Biotechnology and Plant Health, Department of Fungal Plant Pathology in Forestry, Agriculture and Horticulture)
- Anne-Grete Roer Hjelkrem (Division of Food Production and Society, Department of Agricultural Technology and System Analysis)
- Ingerd Skow Hofgaard (Division of Biotechnology and Plant Health, Department of Fungal Plant Pathology in Forestry, Agriculture and Horticulture)
- Brita Linnestad (Division of Biotechnology and Plant Health, Department of Fungal Plant Pathology in Forestry, Agriculture and Horticulture)
- Jiangsan Zhao (Division of Environment and Natural Resources, Department of Biogeochemistry and Soil Quality)

Ås, 03.10.21

Hildegunn Norheim

Content

1 Introduction

The importance of a well-functioning bioeconomy is increasingly recognised in addressing challenges such as food safety, natural resource scarcity, climate change, unsustainable development, and consumption patterns. Defined as an economy in which food, materials and energy are derived from renewable biological resources involving the land and the sea (Commission, 2012), bioeconomy is seen as a central component of sustainable development. Availability of data and information is crucial to be able to take correct decisions at all levels and in many other sectors, from healthcare to transport and energy, including bioeconomy, big data have become fundamental.

The implementation of big data technologies and methodologies in NIBIO is examined in the report Spatial *Big Data tools and methods within NIBIO. A guide to new opportunities and possible ideas and challenges for NIBIO*. During the development of that report, the authors also collected materials describing examples of *big data* usage inside NIBIO. The materials are published in this report.

2 How Big Data can support the AR5 workflow

AR5 is the authoritative land resource map at a scale of 1: 5000. AR5 is a detailed, nationally comprehensive dataset and the best source of information on the country's land resources. The map classifies the land by area type, forest site quality, tree species and soil conditions.

A process which is particularly relevant for NIBIO is the AR5 update, which is done in cooperation with municipalities. While municipalities update AR5 continuously, NIBIO contribute by performing periodic updates every few years. Inaccuracies and area changes neglected in the continuous updating made by the municipalities are captured using orthophotos. In addition, AR5 is harmonized with other updated data sets, such as road and water as needed.

The periodic updating of AR5 is done in cooperation with Geovekst, and AR5 is one of the ordinary Geovekst projects. This means that the municipalities will submit the data for updating in the same way as other FKB data sets. A periodic update must be adapted to aerial photography and management, operation, and maintenance agreements (FDV) under the auspices of Geovekst. Where there are no comprehensive local photography projects for the municipality, orthophoto from Geovekst can be combined with other photos.

The update process of AR5 is done manually. It is not realistic to expect to completely automatize the update process, but it may be possible to take advantage of methods such as machine learning to improve the efficiency of the process and introduce some automated steps.

Periodic updates made by NIBIO use the latest version of FKB-AR5 as a starting point. This contains changes from the municipality's continuous updating. AR5 is placed over the latest orthophoto so that it is possible to visually see if the map matches what can be seen in the image. The map is updated where orthophotos show that there are changes in the landscape. In areas that need to be mapped, information from other archives, such as economic maps or soil data, is also used. Changes must not be projected, it is the current state that must be registered in AR5. In the case of periodic updating, the area condition at the time of aerial photography is registered. If the municipality has made registrations of more recent date than the orthophotos, the municipality's registrations are retained.

2.1 AR5 Classification

The classification of AR5 is shown in Figure 1, including colours and symbols used in the maps. Some of the themes are difficult to identify from an aerial image since they are not related only to the land cover but also information not directly visible from aerial images such as site quality. The presence of an operator performing the classification is therefore fundamental and, in these case, even big data methods cannot provide good result; those layers are elaborated in section 10. However, most of the layers do not have this problem and it is therefore important to exploit the potential coming from big data methods for automated or semi-automated mapping.

Most problematic layers are fully cultivated, surface cultivated, infield grazing and open permanent land that can potentially be classified as the same area type in machine learning. Bogs (Mire?) can also fall under the same class. Built-up area with garden can also be misinterpreted. Further, grass-covered area classification is particularly challenging for a machine.

Arealtype (ARTYPE)	Treslag (ARTRESLAG)	Skogbonitet (ARSKOGBON)	Grunnforhold (ARGRUNNF)
Fulldyrka jord (21)	Barskog (31)	Særs høy (15)	Jorddekt (44)
Overflatedyrka jord (22)	Lauvskog (32)	Høy (14)	Organiske jordlag (45)
Innmarksbeite (23)	Blandingsskog (33)	Middels (13)	Grunnlendt (43)
Skog (30)	Ikke tresatt (39)	Lav (12)	Fjell i dagen (42)
Åpen fastmark (50)	Ikke relevant (98)	Impediment (11)	Blokkmark (41)
Myr (60)	Ikke registret (99)	Ikke relevant (98)	Konstruert (46)
Snøisbre (70)		Ikke registrert (99)	Ikke relevant (98)
Ferskvann (81)			Ikke registrert (99)
Hav (82)			
Samferdsel (12)			
Bebygd (11)			
Ikke kartlagt (99)			

2.2 AR5 workflow

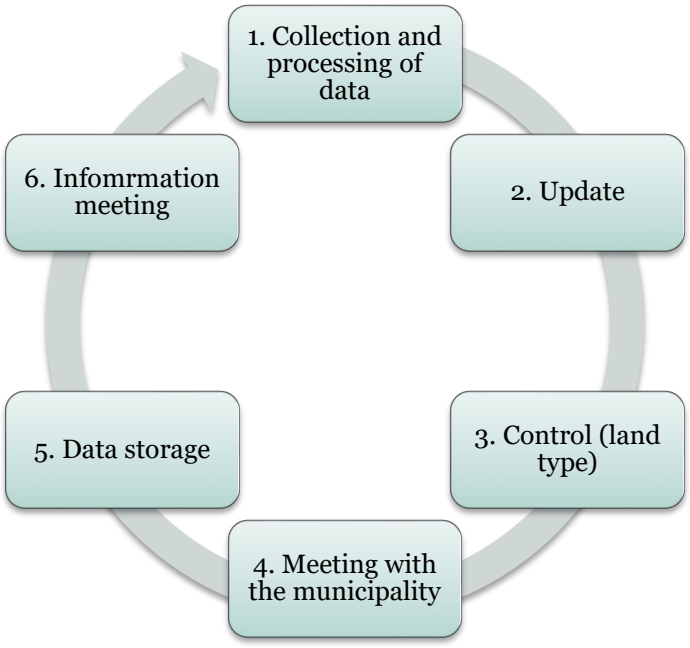


Figure 2. Current AR5 workflow

Collection and processing of data

The first step is retrieving AR5 from the central base (SFBK). FKB-Water, road and railway are also collected. All road and water data contained in AR5 is replaced and updated with the latest data.

A technical inspection is taken of the continuous updating carried out by the municipality. The technical control checks that curves, surfaces, signatures, and data capture date are correct. Depending on how much and how good a job the municipality has done, the latter task can take from less than an hour to several days of work.

Furthermore, AR5 is adapted for the operators who will update AR5. Areas where the municipality has changed ground conditions and ARTYPE are highlighted for the operators. Areas that may have incorrect soil condition based on information from soil data will also be highlighted.

Before updating there is also a control of what the municipality has done in terms of updates to the map, such as changing an area from forest to cultivated area. By doing this we get a good overview of the changes the municipality has made, as well as an assessment of the quality of the municipality's work.

If huge discrepancies/errors are found in the municipality's continuous work, the municipality will be contacted for clarifications.

Update

All stages of the AR5 flow are important, and good updating is the very foundation of AR5.

The whole municipality is checked manually using orthophoto, in addition other aids used are:

- Economic data
- Soil maps
- Google Streetview

During the update, operators will make changes where they believe it is appropriate, but there are two exceptions:

- The municipality has registrations newer than the available orthophoto
- Infield pastures - With regard to strict requirements for defining an area as infield pastures, this is an area that must be checked in the field.

It will in some cases not be possible to determine the condition of an area using orthophoto; in such cases, the operator will have the opportunity to mark the area with a question to show they are uncertain. These questions can then be asked to the municipality in the clarification meeting.

Technical control

The technical control checks that the entire municipality has been updated by the operators. It also includes correction of technical errors such as surfaces, junctions etc.

Finally, this step also includes the production of change analysis, a file that shows what we have changed during periodic update and is used in the next control.

Ground control

The fourth step is the ground-based final inspection, which is a complete review of the entire municipality before a clarification meeting is carried out. Both the municipalities and NIBIO's work is controlled.

In particular, the focus is the following activities:

- Identify possible areas that have been forgotten to be updated;

- Areas that are not correctly updated according to current practices and rules;

Clarification meeting

The clarification meeting, marking the end of the periodic update, is carried out for all municipalities and is the last thing done before the new AR5 file is sent back to the Mapping Authority. There is a control that checks that there are no technical errors in the file before it is sent back. A file that shows changes done during the periodic update is also made.

Information meeting

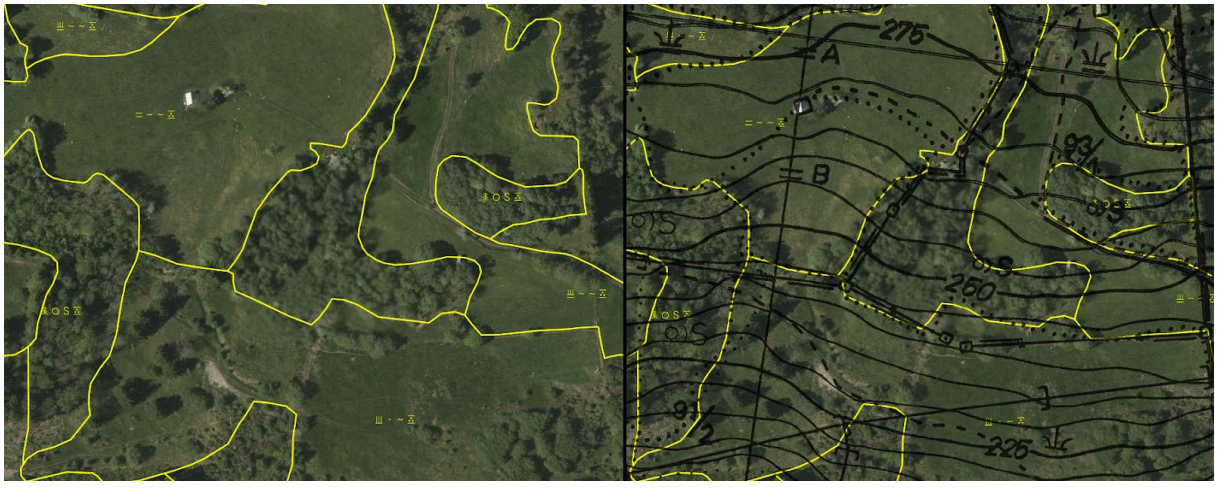
An information meeting with the municipality is held after all the data has been updated in our map services. This meeting aims to give the municipality information on what they should focus on in the future and give them information about how they can look at the changes we have done. There is also information on how they can update AR5 based on the GIS-software the municipality is using.

2.3 Big data Methods to improve the process of update of AR5

A full automation of the updating process is not foreseen due to the characteristics of the maps (it is a combination of land cover and (potential) land use), but several steps can benefit from it.

Identification of areas requiring update

Periodic updates of the AR5 datasets are dependent on the availability of high-resolution aerial images, which are taken with intervals of several years. However, satellite images, with a much higher temporal frequency, can help identify areas where changes have occurred. The limitation of these images is the resolution, which is much lower than the precision needed by the AR5. Therefore, satellite images cannot be directly used to update the geometry of the AR5 polygons.



A competence project made in cooperation with external data scientists, allowed to evaluate the possible results that can be obtained using machine learning or deep learning algorithms on aerial images. Several solutions have been tested. The first step towards a more automated update of AR5 is represented by the possibility of dividing the area under investigation into tiles and for each tile calculate the probability that changes have occurred after the previous updates. This can be achieved through the following steps.

1. Extract AR5 and images for last update and new update to perform
2. Divide the maps and the images into tiles

3. Compare the possible maps obtained by machine learning applied to both images
4. Quantify the probability of changes in that tile

This procedure potentially represents an improvement compared to the actual process since operators do not need to check all images thoroughly but can focus mainly on areas where the described procedure reports a higher probability of changes (and thus the need of an update). Advanced support for the identification of new AR5 areas

Further methods have been tested in the aforementioned cooperation project, aiming to find an algorithm able to propose an updated AR5 map. It is important to remind that some AR5 classes cannot be identified from an automated procedure and require the interaction with a human operator.

Although an automated procedure provides good results, the results might have an accuracy too low compared to that required by AR5.

Integration of images from multiple sources at different scales

A further development towards a more automated update of AR5 can be provided using different data sources at different temporal and spatial resolution. This can allow to have more frequent updates on areas where there have been more changes. The following Figure show the possible data sources that can be used in a sequence, from higher temporal resolution and lower spatial resolution, to on demand images with very high resolution.

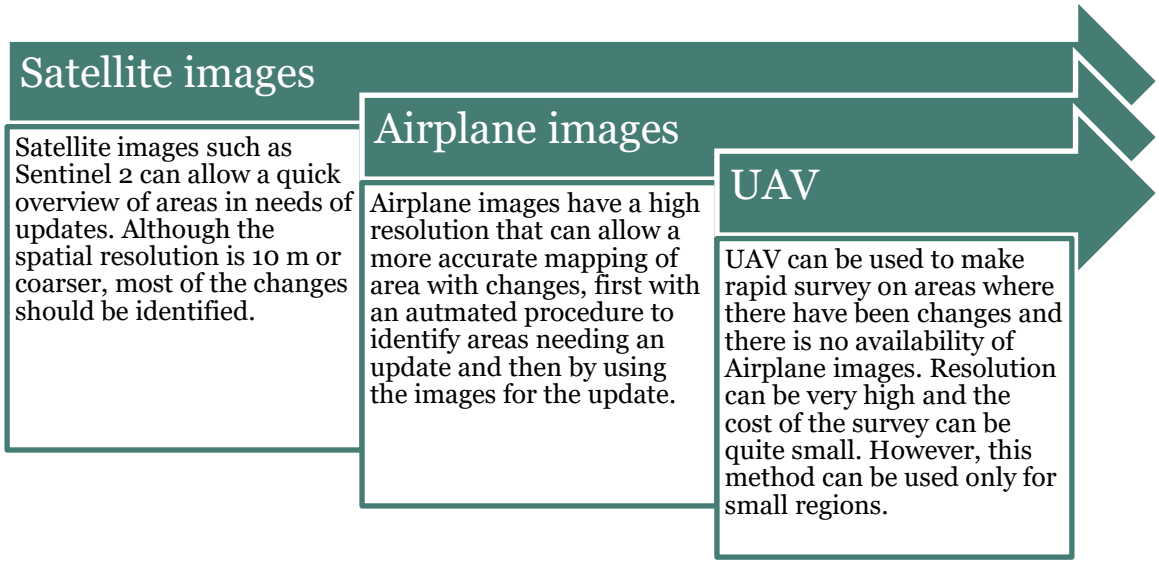


Figure 4. Possible data source to be used for the update of AR5.

3 Applications in the Plant Health division

3.1 Use of artificial intelligence to improve existing disease prediction models in VIPS

Weather factors have different and varying effects on potato late blight (PLB) in different stages of disease development which are challenging to model. Various conditions are favourable for spore production, spore spread, survival of the spores and infection, and if one of the stages does not have good enough conditions, it will not be a successful infection. NIBIO has nine years of data (weather data, spore release and infection data from trap plants). Based on these data, the Nærstad model [1] was developed, and it is in use in VIPS 2 today. It is a process-based model that consists of several parameters. The use of artificial intelligence would be ideal to use on such an advanced data set, both due to the complicated biology and all the factors affecting the various stages of the disease. This will give a good indication of whether machine learning is a suitable method for improving existing and developing new disease infection models.

The aim of the pilot application was to test the use of machine learning on PLB data. NIBIO's PLB data are from the seasons 2006 – 2015 and contains a total of 286 daily values. The PLB data was combined with weather data for the given hours. Everything was collected in a csv file. The data has been run in a program developed in Python, using the Jupyter Notebook. The libraries that have been used are Pandas, NumPy and Keras.

Although in this case the available dataset is rather small compared to a traditional big data application, it was possible to use the PLB dataset to look at different machine learning models, with different parameters as input. For all parameter combinations, the accuracy was estimated, describing the percentage of hits when the test data was used in a prediction calculation. Furthermore, a confusion matrix was defined for each combination of parameters, showing the proportion of false positives and false negatives.

Three different algorithms have been tested on the dry rot data set:

- Logistisc Regression (LR); is a good starting point for binary classification,
- Support Vector Machine (SVM); is also a good starting point for binary classification,
- Random Forest (RS).

The first two algorithms are strong on binary classifications, while the last one performs well with multiclass classifications.

The purpose of the application is to predict risk of PLB. In the observed data, a threshold value of zero lesions per plant was used to distinguish between infection/non-infection. The number of lesions per plant is a daily value. New weather parameters were calculated based on the hourly values. The day was defined as being between from 15:00 and 15:00, and included the following parameters:

- TM: average temperature;
- RR: total precipitation;
- UM: number of hours with humidity over 80%;

[1] https://www.vips-landbruk.no/forecasts/models/NAERSTADMO/

[2] https://www.vips-landbruk.no/

- BT: number of hours with leaf moisture;
- MIN: minimum temperature;
- MAX: maximum temperature;
- #hours with TM > 10°C: Number of hours with temperatures above 10 ° C.

The data set was divided into a training set, containing 80% of the data, and a validation set, containing the remaining 20% of the data. Data weas randomly split into the two datasets.

Table 1. Overview of the results obtained by the different combinations of parameters using the three algorithms.

Parameter	LR	SVR	RS
TM, RR	0,810	0,778	0,655
TM, RR, UM	0,810	0,778	0,328
TM, RR, UM, #hours with TM > 10°C	0,810	0,759	0,810
TM, RR, UM, #hours with TM > 10°C, BT	0,845	0,759	0,810
TM, RR, UM, #hours with TM > 10°C, BT, MIN, MAX	0,897	0,810	0,828

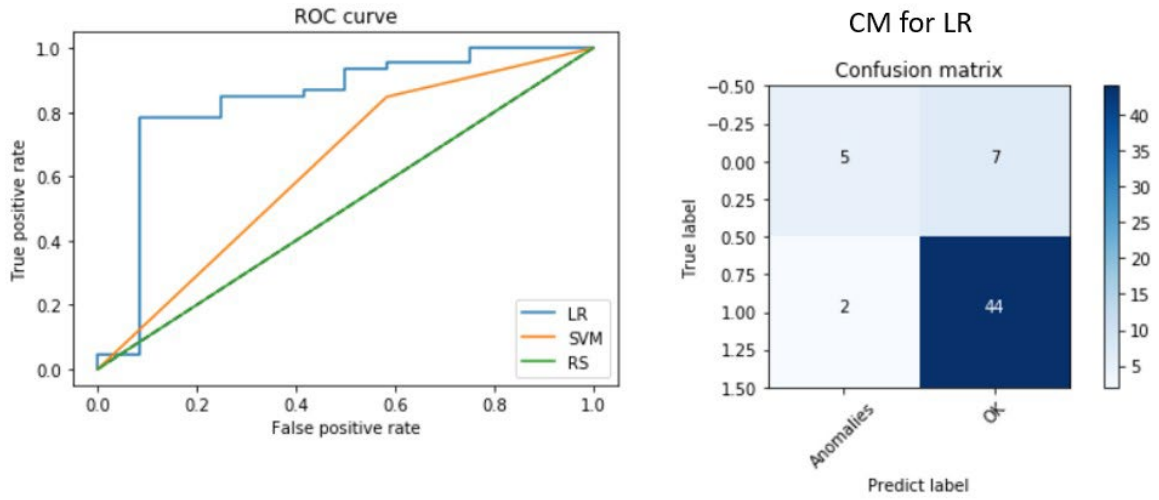


Figure 5. Accuracy and confusion matrix using the logistic regression with the following combination of parameters: TM, RR, UM, #hours with TM > 10°C, BT.

Results show that all models get better accuracy with an increasing number of input parameters. It would certainly be possible to determine more parameters based on existing data and get even better results.

Although it is not possible to directly compare these results with the model created based on the same dataset, it seems that the method can find good correlations with high accuracy between climatic factors and disease on the plants even based on few parameters. Further developments should focus on the use of neural networks and deep learning. However, it is important to have a larger dataset to have a more robust training.

3.2 Use of artificial intelligence to identify associations between environmental factors and mycotoxins in oats.

The aim of this pilot is to test machine learning algorithms (specifically unsupervised learning) on existing data sets to investigate the association between cultivation conditions and the development of mycotoxins in Norwegian oats. Since the aim is to develop warning models with known groupings for risk/non-risk, it was early decided to switch to supervised learning. When using machine learning algorithms for classifying data, most of the time (~ 90%) is spent on preparing the data sets. In this case, by using existing datasets, simple pilot models were developed for warning of the risk of the mycotoxin deoxynivalenol (DON). Implemented algorithms are the following ones:

- Classification trees
- K-nearest neighbour
- Linear discriminant analysis.

An example of a classification tree is given in the following figure.

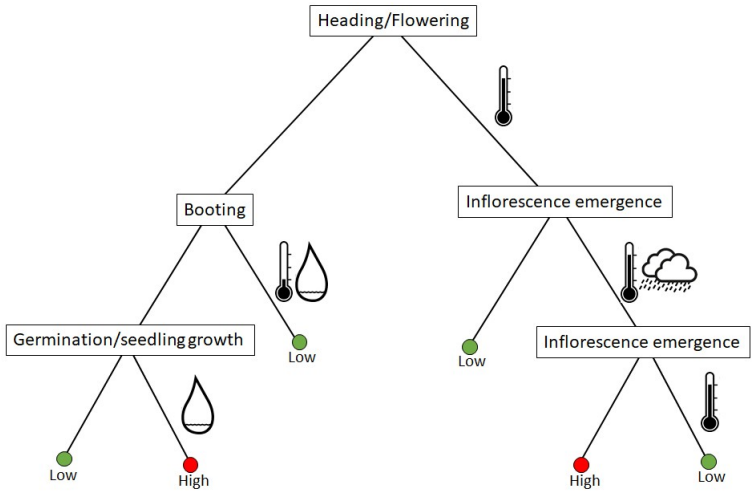


Figure 6. An example of a tree model for warning of the risk of disease in cereals based on weather data.

The pilot models gave good results, but since the data set was relatively small for a robust machine learning application, it could not completely take the advantages of the benefits that lie in machine learning and big data methods.

Fusarium species that produce DON need warm and humid conditions for infection and disease development. Therefore, relative humidity (RH) is an important weather factor that is often included in our warning models. Unfortunately, RH is difficult to measure and there is also considerable uncertainty associated with these measurements. We have lost large parts of our data sets as a number of weather stations have delivered too low values over several years. This is a general weakness of instruments that measure RH. NIBIO / LMT has a lot of historical data on this variable, and we envisage that it should be possible to use these data sets in machine learning / deep learning to estimate relative humidity or to be able to detect changes / needs for service at weather stations.

3.3 Automatic identification and quantification of plant pests

Artificial intelligence has proven to be an effective approach to image analysis and classification, far more so than traditional digital image analysis. Neural networks have in several attempts been trained to analyse images more efficiently and accurately than humans. For example, machines are now better than specialists at determining whether a mole is benign or malignant. To date, NIBIO has used traditional, rule-based algorithms for image analysis. Using existing datasets (images of pests), we can train artificial neural networks to recognize plant pests automatically. This can be used in, for example, precision agriculture, automatic detection of pests in the field, and reduction of manual labour in connection with analyses under a microscope. It is possible to compare artificial intelligence methodologies with classical methods, both machine vision algorithms and manual counting. In this pilot there have been two applications related to: i) fungal spores/potato rot and ii) weeds.

3.3.1 Automatic identification of plant pests (fungal spores / potato rot) from images taken under a microscope

Sporangia of Phytophthora infestants which cause late blight in potatoes (PLB), have a characteristic shape that makes them well suited for exploring image recognition as a detection method. By monitoring the presence of spores in the air, one can estimate the risk of attack of the disease at a given time and compare with weather data and other observations. This is an important tool in developing models for many pests. Today, spores of this disease are counted manually using a microscope, and the efficiency potential of automation is significant. NIBIO has previously used traditional image analysis algorithms to detect fungal spores of an insect pathogen. The experiences from these two different methodologies can be compared.

Preparation of slides

Pictures were taken from microscope slides with a tape containing spores of Phytophthora infestants (PI). This is the way spores are caught in the field using volumetric spore traps. To use as training data, slides prepared in the lab containing only spores of PI (no contamination like soil, pollen, or other spores) were used to identify the spores. At a later stage, real spore trap slides were included to provide a more challenging environment.

Image annotation

Images are labelled with graphical image annotation tool (as shown in the following figure), Labelme, to mark the locations of spores on these images. Then a csv-file of the coordinates of these spores highlighted is prepared together with the original images for training.

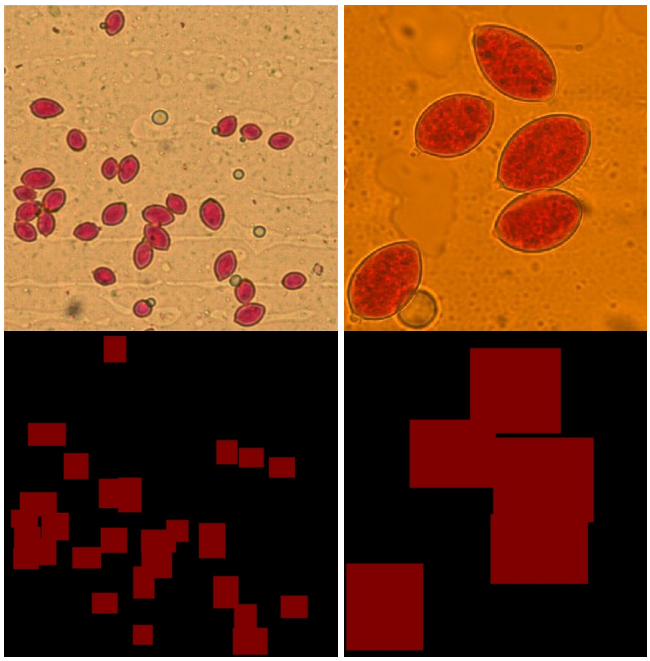


Figure 7. Original image of the spores (top) and mask of annotated spores (bottom).

Model training

Due to the small size of the dataset, transfer learning using pretrained model, ResNet101 architecture trained on Coco data set, was used to train for spore detection. The model was trained for 10 epochs in total with 3000 iterations in each epoch; the learning rate was set at 1e-5 and kept constant during training.

The free of charge cloud service, Google Colaboratory (Colab) together with Python 3 runtime, was the major platform for all the model training. Colab is equipped with a 2.3 GHz and 12.6 GB RAM Intel Xeon processor with two cores and a NVIDIA Tesla K80 GPU with 12 GB RAM. Data preparation was done in python 3.

Results and Discussion

Through transfer learning, the model is already good enough to detect spores of potato late blight on relatively clean slides with high accuracy after 10 epochs training (see figure), which enables quick spore quantification, number, and size in high throughput manner. However, the images collected from the field can be much noisier, more data from field situations should be used for training and more extensive data augmentation methods will be applied to enhance the model robustness in the future.

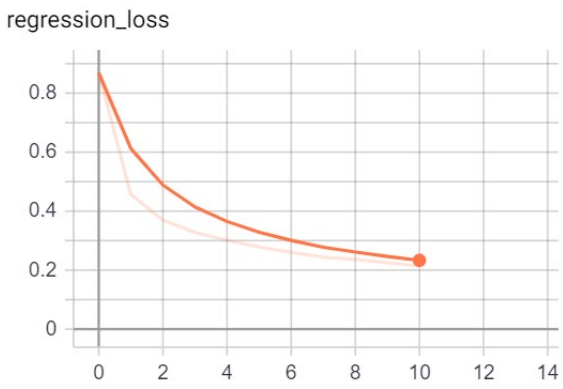


Figure 8. Graph of the regression loss of the training, with constant decrease to 0.21 at epoch 10.

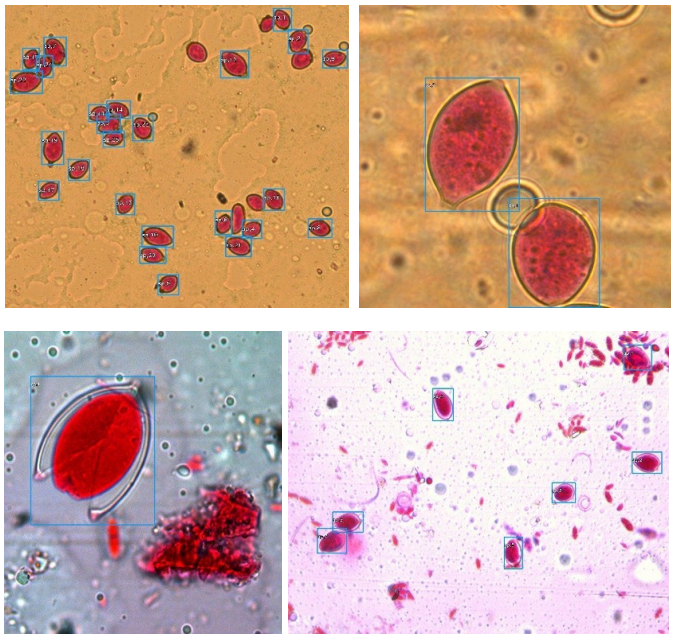


Figure 9. Detection results of trained images (above) and spores detected on new test images (below).

An attempt to compare automatic counting with the traditional, manual counting, unfortunately failed due to low performing in-house equipment (too long time to generate images with a quality that is insufficient). More advanced equipment can now be found on the marked to produce higher resolution pictures, having additional possibilities like creating a multi- focus image by adjusting focus while the picture is created (e.g., LAS X "Live Z-Image Builder" in Leica Application Suite) or creating low magnification picture with high depth of field. Further investments of time and money in this direction, given the promising preliminary results, are encouraged for NIBIO. In this direction, the cooperation with external partner organizations such as NMBU and its Imaging Centre can allow to have access to advanced equipment without large investments.

These results show that the potential for automatic recognition and counting of spores is high when you have spores with characteristics making them easy to distinguish from other spores or contamination. But this requires that a single, high quality picture can be generated from one slide. Further, the resources spent to create this image should be lower than what is needed to do the spore counting manually. In addition, including a multi-focus image can further increase the precision of this method compared to the manual, because “dirty” slides with spores in several focal planes are challenging to count accurately in the manual way.

3.3.2 Automatic identification of weeds

The aim of this activity was to complete a scientific manuscript that deals with automatic identification and quantification of young seed-propagated weeds (dicots and monocots) vs. spring barley in proximate (ground-based) RGB images. The result is a comparison between a new algorithm based on neural networks and a traditional rule-based algorithm. The new algorithm was generally more precise than the traditional one but coped poorly for images with extremely high weediness. The algorithms based on neural networks gave better results than traditional algorithms, but representativity in training data is important. The usefulness of robust AI–based algorithms is large because they are basis for innovative precision weeding methods that can reduce the herbicides usage and other weeding measures in conventional and organic production, while maintaining or even increase weeding efficacy, and thus secure high crop yields.

The results are described in detail in a submitted manuscript "Precision weeding in cereals: Evaluation of a novel machine vision algorithm based on deep learning" (by Therese W. Berge et al.). The results, presented also in an oral speech with the title "Precision weed harrowing in spring cereals" at a conference in late 2019, represents a building block for future innovations in precision agriculture/precision weeding technologies and integrated pest management (IPM, in particular IPN principle No. 6). Further use of this result may be variable rate weed harrowing in spring cereals, a precision agriculture approach to reduce herbicide use and risk of herbicide resistance development in weeds (conventional/IPM production) or diesel use during mechanical weeding (organic production).

4 Processing satellite data using open-source libraries for distributed computing

In this chapter, we present an example application where an integrated method based on a cluster of personal computers and a set of open-source software packages have been used to build-up a local solution for distributed processing of satellite data (in this case, Sentinel-2 satellite images are used, but the same could be done with any satellite or other remote sensing data). The infrastructure has been tested under different configurations by computing the Normalized Difference Vegetation Index (NDVI) and the monthly median values of time series of Sentinel-2 images.

4.1 Hardware and software

For this experiment, we created a network of 10 workstations with different computational and storage capacities, as detailed in Table 1. All devices were connected to the local network using high speed ethernet cables with data transfer speed of about 1 Gb/s.

Table 2. Specifications of the computers used in the cluster. The last computer (J) is used as a client and it is not involved in the computation, therefore its RAM and cores are not part of the total RAM and cores.

ID	Operating system	RAM (GB)	# of cores/threads	Processor speed (GHz)
A	Linux (Centos)	64	6/12	3.7
B	Windows	32	4/8	3.4
C	Linux (Ubuntu)	16	4/8	3.4
D	Linux (Centos)	16	6/6	3
E	Linux (Centos)	8	4/4	3.3
F	Linux (Centos)	8	4/4	3.3
G	Linux (Centos)	4	2/4	1.9
H	Linux (Centos)	4	1/1	2.5
I	Linux (Centos)	2	2/2	2.2
J	Windows	8*	4*	2.7
Total		154	33/49	

The integrated system is tested both under the Anaconda Spyder IDE and the Jupyter lab with the main software and their versions listed in Table 2.

Table 3. List of the open-source software used in this experiment.

SW Component	Version	SW Component	Version
Python	3.7.5	**Dask**	2.30.0
Anaconda	4.9.2	**Spyder IDE**	3.3.6
Xarray	0.15.1	**Jupyter**	1.0.0

4.2 Analysis and results

The selected analysis for this experiment are two indicators: NDVI (Normalized Vegetation Difference Index,) and the monthly median value calculated for each band. Although the calculations are not complex, they can be demanding from a computational point of view. In particular, the median requires sorting all the values of the considered dataset, which is an operation that require large memory when computed over large datasets.

The NDVI allow to understand if an observed pixel contains living vegetation and allows the evaluation of the health status of the observed vegetation, such as forest, grass, or any crop. The NDVI is calculated as follows:

NDVI = (NIR – Red) / (NIR + Red)

where:

- NIR: value of reflectance from the Near InfraRed band (Band 8 of Sentinel-2),
- Red: value of reflectance from the Red band (visible, Band 4 of Sentinel-2).

The median of a series of values (in this case a time series of the values for each pixel) is the value separating the lower and higher half values of a given dataset (not to be confused with the mean or average, which is a calculated value). The median allows to understand if the distribution of the values is or is not skewed (although need to be used together with other indexes). For a series of values, the median is found by ordering increasingly all values and selecting the middle one (in case of odd number of observations) or the mean of the two middle values (in case of even number of values). The median value can be useful in optical satellite image time series (SITS) data to extract an image that is not influenced by clouds and shadows, that have high and low values respectively.

The first two experiments have been done to check the performances using a single machine at a time and doing the computation splitting the data into chunks of different size. As it was expected, computers with higher RAM and multiple cores are faster in calculating both the NDVI and the median. It was noted that the chunk size can have a great influence on the performance of the infrastructure. Splitting the data into small chunks, increases the overall computation time: this is due to the higher time used for the transmission of data. Increased chunk size, instead, reduces the computation time if chunks have a size which is well under the available memory of the used machines, otherwise the elaborations can be slower, or some machine could even be excluded from the network if RAM is not enough to handle the chunks. In our case, chunk size of 2745 pixels resulted in the fastest computation (a chunk of 2745 pixels is one-fourth of a Sentinel-2 frame, which is 10980 by 10980 pixels).

Calculations have been successively performed in different ways to evaluate how different configurations of the cluster perform. When the NDVI or median are calculated using the Dask cluster starting from the less powerful machine and then adding a machine at a time in order of increasing performance, the computation time is decreasing as shown in Figures 10A and 10C, with the fastest computation obtained when the cluster includes all the computers. However, when the calculation of the NDVI or median are done starting from the best performing machine and adding all others in decreasing order of performances, the computation time keeps on decreasing up to a point where it starts to increase, as shown in Figure 10A and Figure 10B. Anyway, the computation time achieved by using all the computers in the cluster is the same in this case to the previous case with increasing order of performances as the final cluster is the same. The lowest computation time is registered when 4 (for NDVI) or 3 (for median) best performing computers are used.

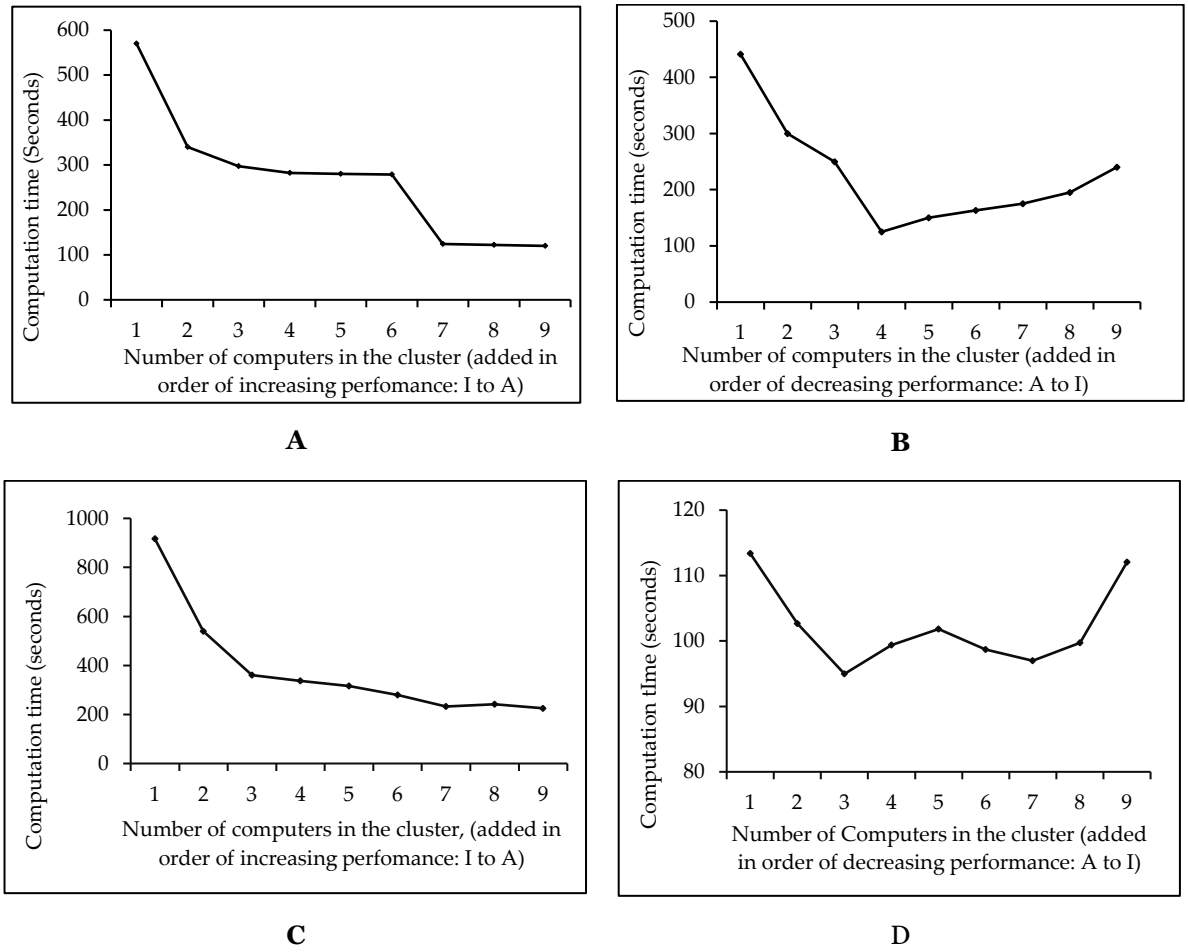


Figure 10. Time for computing the NDVI (A and C) and the median (B and D) by adding new machines to the framework in order of increasing (from I to A) and decreasing (from A to I) performances.

Last experiment focused on the use of different source of data, comparing local data vs. remotely stored data. Results showed a big difference between the computation time when using local network disk compared to the use of a remote server. Data acquisition from a remote server can increase the execution time due to the low bandwidth compared to data acquisition from a local source. In the performed experiment the execution time decuplicated due the much higher time spent on data transfer.

4.3 Conclusions

The open-source infrastructure briefly presented and tested for the calculation of some indices, showed that doing analysis of big data that goes over the capacity of a single personal computer, can be possible in an efficient way also using a local infrastructure, before scaling up with very large analysis that require High Performance Computers (HPC) or Cloud services. Further, the tested infrastructure can be used for testing of processing routines before implementing on HPC's or Cloud services. The implementation is simple, and it is based on Python, which is now widely used for big data analysis. Additionally, the proposed solution does not require a dedicated hardware, but can be built up on personal computers that are often idle during the daytime and almost always during the night in medium/large organizations.

The performed test highlighted some aspects to be considered when setting up an infrastructure similar to the proposed one.

- Machines with low performances (e.g., little RAM or single core) can slow down the overall processing speed and therefore should not be included.
- The connection of the devices used within the cluster and between the cluster and the data storage (if external) should be fast to avoid the creation of a bottleneck.
- In the executed script, it is important to optimize data access (or eventually scattering), dimension of chunks used to split data, and computational algorithms. It is recommended to break down the entire job into smaller tasks.

References

Cattell, R. (2011). Scalable SQL and NoSQL data stores. *ACM SIGMOD Record, 39*(4), 12-27. doi:10.1145/1978915.1978919

Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile networks and applications, 19*(2), 171-209.

Commission, E. (2012). Innovating for sustainable growth: a bioeconomy for Europe. *Communication from the Commission to the European Parliament, the Council, the European economic and Social Committee and the Committee of the regions, Brussels, 13*, 2013.

Dean, J., & Ghemawat, S. (2004). *MapReduce: simplified data processing on large clusters*. Paper presented at the Proceedings of the 6th conference on Symposium on Opearting Systems Design & Implementation - Volume 6, San Francisco, CA.

Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. Communications of the ACM, 51(1), 107-113.

Fan, J., Han, F., Liu, H. (2014). Challenges of Big Data analysis, National Science Review, Volume 1, Issue 2, 293–314, https://doi.org/10.1093/nsr/nwt032

Ghemawat, S., Gobioff, H., & Leung, S.-T. (2003). The Google file system. *SIGOPS Oper. Syst. Rev., 37*(5), 29-43. doi:10.1145/1165389.945450

Hilbert, M., & Lopez, P. (2011). The world's technological capacity to store, communicate, and compute information. *Science, 332*(6025), 60-65. doi:10.1126/science.1200970

Hoyer, S., & Hamman, J. (2017). xarray: ND labeled arrays and datasets in Python. *Journal of Open Research Software, 5*(1).

Iorga, M., Feldman, L., Barton, R., Martin, M. J., Goren, N. S., & Mahmoudi, C. (2018). *Fog computing conceptual model*. Retrieved from

Li, J., Liao, W.-k., Choudhary, A., Ross, R., Thakur, R., Gropp, W., . . . Zingale, M. (2003). *Parallel netCDF: A high-performance scientific I/O interface*. Paper presented at the Supercomputing, 2003 Acm/Ieee Conference.

Li, Y., & Manoharan, S. (2013). *A performance comparison of SQL and NoSQL databases*. Paper presented at the Communications, computers and signal processing (PACRIM), 2013 IEEE pacific rim conference on.

Rew, R., & Davis, G. (1990). NetCDF: an interface for scientific data access. *IEEE computer graphics and applications, 10*(4), 76-82.

Ryan, A., kevin, p., joe, h., matthew, r., chiara, l., michael, t., . . . Davide, D. V. (2017). *Pangeo NSF Earthcube Proposal*.

Shvachko, K., Kuang, H., Radia, S., & Chansler, R. (2010). *The Hadoop Distributed File System*. Paper presented at the Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST).

Uehara, M. (2017). *Mist computing: Linking cloudlet to fogs*. Paper presented at the International Conference on Computational Science/Intelligence & Applied Informatics.

NIBIO
NORWEGIAN INSTITUTE OF
BIOECONOMY RESEARCH

NIBIO - Norwegian Institute of Bioeconomy Research was established July 1 2015 as a merger between the Norwegian Institute for Agricultural and Environmental Research, the Norwegian Agricultural Economics Research Institute and Norwegian Forest and Landscape Institute.

The basis of bioeconomics is the utilisation and management of fresh photosynthesis, rather than a fossil economy based on preserved photosynthesis (oil). NIBIO is to become the leading national centre for development of knowledge in bioeconomics. The goal of the Institute is to contribute to food security, sustainable resource management, innovation and value creation through research and knowledge production within food, forestry and other biobased industries. The Institute will deliver research, managerial support and knowledge for use in national preparedness, as well as for businesses and the society at large.
NIBIO is owned by the Ministry of Agriculture and Food as an administrative agency with special authorization and its own board. The main office is located at Ås. The Institute has several regional divisions and a branch office in Oslo.

Front cover: image by Gerd Altmann from Pixabay.
Back cover: image by Elchinator from Pixabay.

nibio.no