Plant disease forecasting using data mining and machine learning: a case study on Fusarium head blight and deoxynivalenol in winter wheat

S. Landschoot [1] G. Haesaert[1]

[1]Department of Plants and Crops,
Faculty of Bioscience Engineering
Ghent University Ghent

8^{th} Online Workshop on the status of Mycotoxins predictive models in Africa, 26th of October 2020

Overview

1. PART 1: Plant disease modelling
2. PART 2: A case study on FHB

Overview

1. PART 1: Plant disease modelling
2. PART 2: A case study on FHB

Three main steps in model building

Step 1: Data collection and data mining

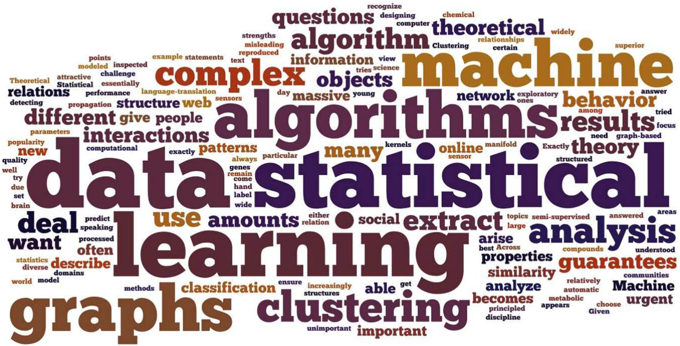
Discovering patterns in the dataset, finding the most important variables for modelling

- Summarising data (e.g. calculating frequencies, . . .)
- Presenting data (e.g. histograms, boxplots, . . .)
- Statistical inference (e.g. parametric and non-parametric test, correlation analysis, . . .)

Three main steps in model building

Step 2: Machine learning

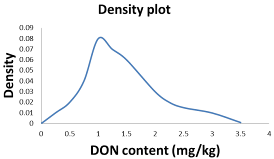
Using algorithms to predict one (or more) variable(s) based on properties in the input data



Three main steps in model building

Step 2: Machine learning

- **Continuous output variable**: the output variable can have an infinite number of values e.g. toxin content, ...



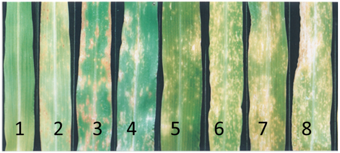
$\Rightarrow$ metric regression techniques

- Multiple linear and ridge regression
- Linear and non-linear mixed models
- Regression trees
- Support vector regression

Three main steps in model building

Step 2: Machine learning

- **Ordinal output variable**: the output variable is discrete, but there exist an ordening e.g. disease classes,. . .



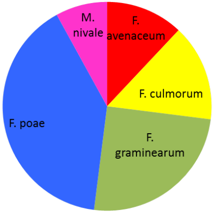
⇒ Ordinal regression techniques

- Proportional odds models
- Support vector ordinal regression

Three main steps in model building

Step 2: Machine learning

- **Categorical output variable**: output variable is discrete and there exist no ordening e.g. distribution of *Fusarium* species.



⇒ Classification techniques

- Logistic regression
- Support vector machines
- Classification trees

Three main steps in model building

Step 3: Model validation

Model validation techniques address two fundamental problems:

- Model selection and parameter estimation
- Performance estimation: How will the model perform in the future?

Cross-year cross-location validation:
observations for a given year and location never participate in the model construction and the model evaluation process simultaneously.

	Location 1	Location 2	Location 3	Location 4	Location 5	Location 6
Year 1	Test	Omitted	Omitted	Omitted	Omitted	Omitted
Year 2	Omitted	Train	Train	Train	Train	Train
Year 3	Omitted	Train	Train	Train	Train	Train
Year 4	Omitted	Train	Train	Train	Train	Train
Year 5	Omitted	Train	Train	Train	Train	Train

Figure: fold 1

Performance of the model for new years and new locations

Three main steps in model building

Step 3: Model validation

Model validation techniques address two fundamental problems:

- Model selection and parameter estimation
- Performance estimation: How will the model perform in the future?

Cross-year cross-location validation:
observations for a given year and location never participate in the model construction and the model evaluation process simultaneously.

	Location 1	Location 2	Location 3	Location 4	Location 5	Location 6
Year 1	Omitted	Test	Omitted	Omitted	Omitted	Omitted
Year 2	Train	Omitted	Train	Train	Train	Train
Year 3	Train	Omitted	Train	Train	Train	Train
Year 4	Train	Omitted	Train	Train	Train	Train
Year 5	Train	Omitted	Train	Train	Train	Train

Figure: fold 2

Performance of the model for new years and new locations

Three main steps in model building

Step 3: Model validation

Model validation techniques address two fundamental problems:

- Model selection and parameter estimation
- Performance estimation: How will the model perform in the future?

Cross-year cross-location validation:
observations for a given year and location never participate in the model construction and the model evaluation process simultaneously.

	Location 1	Location 2	Location 3	Location 4	Location 5	Location 6
Year 1	Omitted	Omitted	Test	Omitted	Omitted	Omitted
Year 2	Train	Train	Omitted	Train	Train	Train
Year 3	Train	Train	Omitted	Train	Train	Train
Year 4	Train	Train	Omitted	Train	Train	Train
Year 5	Train	Train	Omitted	Train	Train	Train

Figure: fold 3

Performance of the model for new years and new locations

Three main steps in model building

Step 3: Model validation

Model validation techniques address two fundamental problems:

- Model selection and parameter estimation
- Performance estimation: How will the model perform in the future?

Cross-year cross-location validation:
observations for a given year and location never participate in the model construction and the model evaluation process simultaneously.

	Location 1	Location 2	Location 3	Location 4	Location 5	Location 6
Year 1	Omitted	Omitted	Omitted	Test	Omitted	Omitted
Year 2	Train	Train	Train	Omitted	Train	Train
Year 3	Train	Train	Train	Omitted	Train	Train
Year 4	Train	Train	Train	Omitted	Train	Train
Year 5	Train	Train	Train	Omitted	Train	Train

Figure: fold 4

Performance of the model for new years and new locations

Three main steps in model building

Step 3: Model validation

Model validation techniques address two fundamental problems:

- Model selection and parameter estimation
- Performance estimation: How will the model perform in the future?

Cross-year cross-location validation:
observations for a given year and location never participate in the model construction and the model evaluation process simultaneously.

	Location 1	Location 2	Location 3	Location 4	Location 5	Location 6
Year 1	Omitted	Omitted	Omitted	Omitted	Test	Omitted
Year 2	Train	Train	Train	Train	Omitted	Train
Year 3	Train	Train	Train	Train	Omitted	Train
Year 4	Train	Train	Train	Train	Omitted	Train
Year 5	Train	Train	Train	Train	Omitted	Train

Figure: fold 5

Performance of the model for new years and new locations

Three main steps in model building

Step 3: Model validation

Model validation techniques address two fundamental problems:

- Model selection and parameter estimation
- Performance estimation: How will the model perform in the future?

Cross-year cross-location validation:
observations for a given year and location never participate in the model construction and the model evaluation process simultaneously.

	Location 1	Location 2	Location 3	Location 4	Location 5	Location 6
Year 1	Omitted	Omitted	Omitted	Omitted	Omitted	Test
Year 2	Train	Train	Train	Train	Train	Omitted
Year 3	Train	Train	Train	Train	Train	Omitted
Year 4	Train	Train	Train	Train	Train	Omitted
Year 5	Train	Train	Train	Train	Train	Omitted

Figure: fold 6

Performance of the model for new years and new locations

Three main steps in model building

Step 3: Model validation

Model validation techniques address two fundamental problems:

- Model selection and parameter estimation
- Performance estimation: How will the model perform in the future?

Cross-year cross-location validation:
observations for a given year and location never participate in the model construction and the model evaluation process simultaneously.

	Location 1	Location 2	Location 3	Location 4	Location 5	Location 6
Year 1	Train	Train	Train	Train	Train	Omitted
Year 2	Train	Train	Train	Train	Train	Omitted
Year 3	Train	Train	Train	Train	Train	Omitted
Year 4	Train	Train	Train	Train	Train	Omitted
Year 5	Omitted	Omitted	Omitted	Omitted	Omitted	Test

Figure: fold 30

Performance of the model for new years and new locations

Overview

1. PART 1: Plant disease modelling
2. PART 2: A case study on FHB

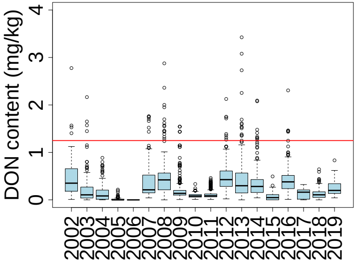
Fusarium head blight

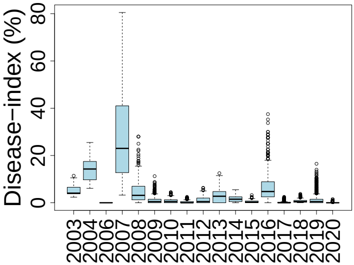
- Winter wheat is one of the most important cereal crops in Belgium (200,000 ha)
- FHB is a fungal disease caused by a complex of *Fusarium* species
- It is a world-wide problem in wheat growing areas due to
  - yield losses
  - quality losses (mycotoxins)
- The structure and distribution of FHB population is determined by
  - weather conditions (temperature, moisture, rainfall, . . .)
  - agricultural factors (crop rotation, weed management, tillage, . . .)




Visual symptoms and DON content

These figures show the yearly fluctuations in disease index and DON content. It is clear that the incidence clearly differs between years.





Weather conditions

Weather conditions are most important factors influencing the FHB incidence.
A correlation analysis was performed to study the influence of the weather variables on disease index and DON content.

- Early in the season: positive correlation with temperature

Mild temperatures favour build-up of inoculum and mycelium growth

- At anthesis: positive correlation with humidity and rainfall

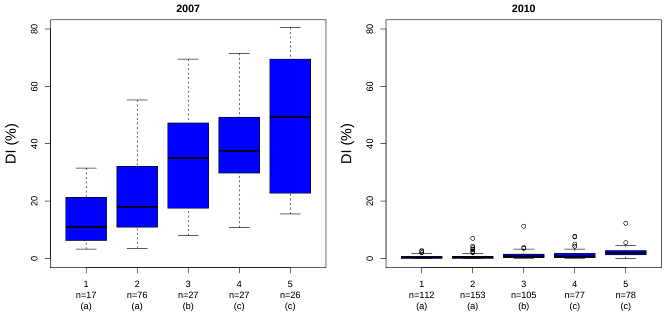
Rainfall and moisture promote production and dispersal of spores

- The correlations between the disease index and weather variables were higher than between DON content and weather variables

Toxin production is complex and influenced by many factors including host resistance, chemotype of isolates, interactions between species, ...

Agronomic factors

Agronomic factors such as crop rotation and wheat variety susceptibility also influence the FHB incidence



Influence of agronomic factors is more pronounced during years with a high disease pressure

Ordinal regression: class boundaries

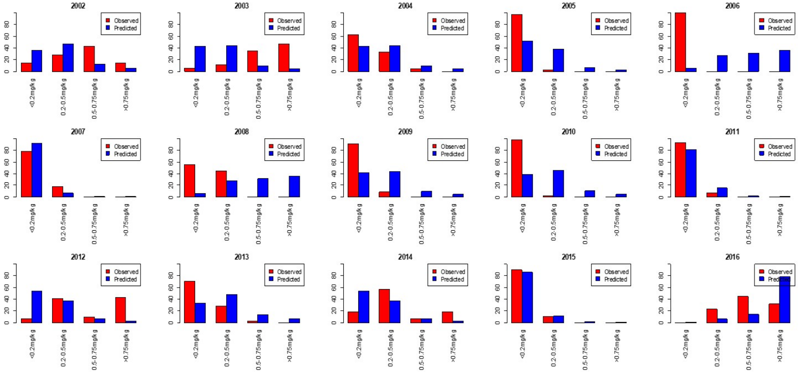
To predict the DON content, a 4-class ordinal regression approach was selected. This model predicts between which boundaries the DON content will be situated.

Class boundaries for DON content model are based on the European legislation:

Limit	Product
0.20 $\mathrm{mg/kg}$	processed cereal-based food for infants and babies
0.50 $\mathrm{mg/kg}$	bread, pastries, etc
0.75 $\mathrm{mg/kg}$	cereals intended for direct human consumption
1.25 $\mathrm{mg/kg}$	unprocessed cereals

The threshold of 1.25 $\mathrm{mg/kg}$ was not considered, as our dataset contains only a very small number of observations exceeding this threshold, but the model can be extended with a fifth class.

Evaluation of the DON content model



Predictions were similar to the measured DON contents