PhD : Crop performance prediction with satellites and environmental data

Johann Desloires
Dates : 15/10/2020-14/10/2023
Academic supervisor : Dino Ienco
Industrial supervisor : Antoine Botrel

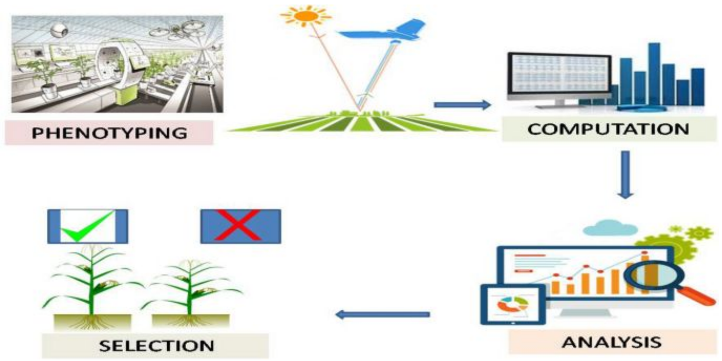
Plan

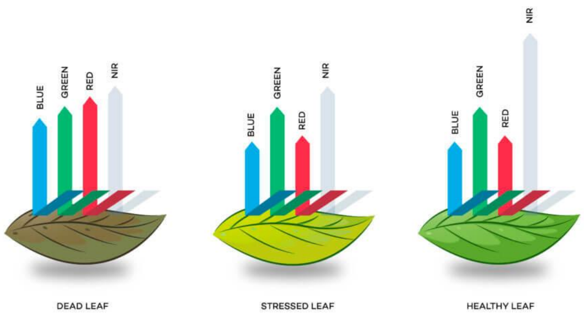
syngenta

Project Introduction

- **Syngenta** is a leading science-based agtech company
  - Data analytics expertise in genetics, biostatistics, system modelling and computer vision for **varieties selection**.
  - Need to develop skills on **satellites data** and deep learning.
- Why predict yield in season?
  - Help farmers to decide on what to grow and when to grow
  - Stock management
  - Optimize human intervention in the fields

□ Remote sensing-based crop yield prediction demonstrated in papers (You et al., 2017)

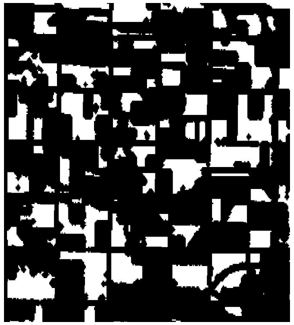




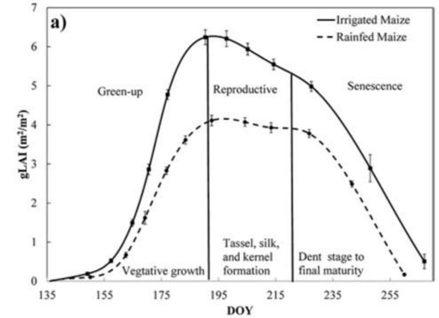
syngenta

Research plan PhD

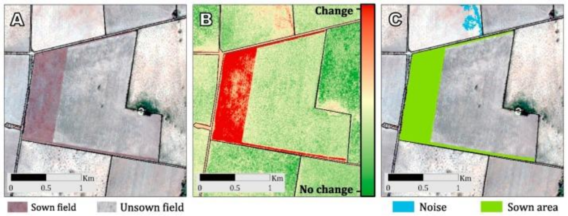
- 1^{st} year : Methodology to detect a particular land cover class with Positive Unlabeled Learning settings.
  □ Objective : identify pixels from a given crop to deploy yield prediction models



- 2^{nd} : Yield prediction of maize varieties in seed production fields from satellites observations
  □ Objective : Identify predictors for vegetation status using Satellite Images Time Series data



- 3^{rd} : TBD ~ Sowing date detection at field scale using unsupervised change detection (PlanetScope data)
  □ Objective : Sowing date is a required input for yield prediction models



Yield prediction using satellites and environmental data

PhD overview : Research plan 2nd year

- ***Objective*** :
  - Yield prediction at the **field level** using environmental and multi-source satellites data **for a new year**
     Very few papers in such setting
  - Focus on recent advances on machine learning for EO instead of crop modelling
- ***Experimental settings*** :
  - Sentinel-2 (S2) time series data on **calendar time**, while being robust to temporal shifts of the growing seasons …
  - S2 time series data on **thermal time** from the sowing date to improve generalization
  - Multi-source satellites and environnemental data

syngenta

Data

- Corn production fields:
  - Parent lines to **form hybrids** between a variety A (role of male) and B (role of female)
  - In-situ data available (irrigated fields, varieties, sowing and flowering dates, …)
  - Harvested yield per female acre

- Sentinel-2 tiles:
  - 250 fields per year in average (1200 in total) and distributed over 11 S2 tiles since 2017

- Environmental:
  - Agro-Meteorological using European Remote Sensing 5 (ERA5)

  ⇒ 0.25 * 0.25 degrees spatial resolution

Fig.1 : Corn production field with female and male rows for breeding pipeline

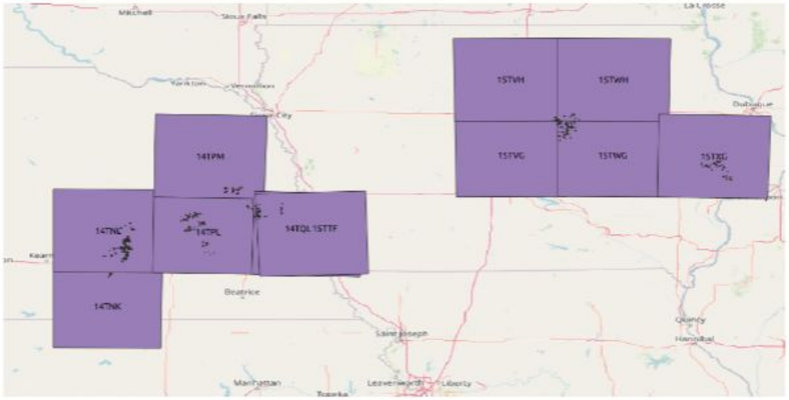


Fig.2 : Spatial distribution over S2 tiles from production fields

syngenta

Methodology : Sentinel-2 (optical) data

- Biophysical parameter estimates from PROSAIL RTM (Weiss and Baret, 2016)
  □ Improve yield prediction (Segarra *et al*., 2022)
  - ***Leaf Chlorophyll Content*** (Cab, in mg) ~ red-edge & swir
    - Proxy relationship between chlorophyll concentration and leaf nitrogen content (Dordas, 2017)
  - ***Fraction of Absorbed Photosynthetically Active Radiation*** (fAPAR) ~ red-edge
    - Radiometric quantity (radiation interception) : Fraction of incident solar radiation that is absorbed by land vegetation for photosynthesis
    - Estimation of primary production / photosynthetic activity, especially for **accumulated values** (Duveiller et al., 2013)

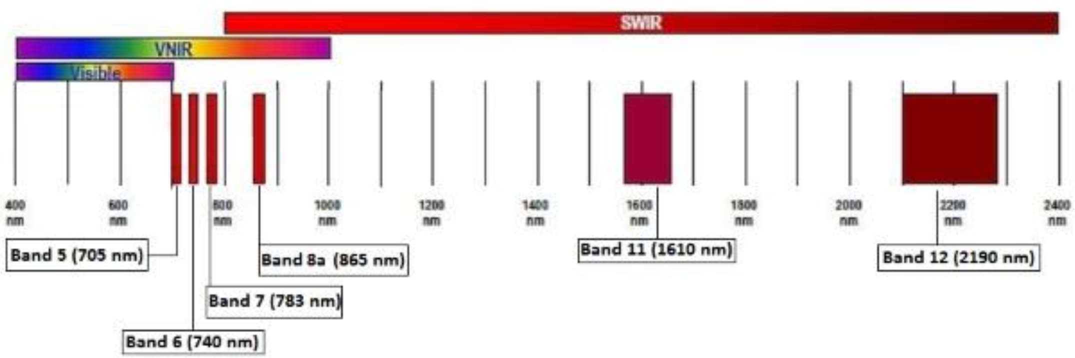


Fig.3 : Sentinel-2 bands in the VIS and IR regions of the electromagnetic spectrum

Methodology : Agro-Meteorological data

- European Remote Sensing 5 (ERA5) satellite-based air temperature data (0.25° resolution)

  ⇒ **Temperature** is the **primary climatic driver** of US agricultural yields (Ortiz-Bobea et al., 2019)

  - Accumulated mean **daily** air temperatures at 2 m ag.l above a **crop-specific threshold** (McMaster and Wilhelm, 1997) :
    - Good proxy for the crop development stage (Duveiller et al., 2013) ~ Growing Degree Days (GDD)

$$GDD = \left[\frac{(\text{Max Temp} + \text{Min Temp})}{2}\right] - \text{Base Temp}$$

  - **Descriptive statistics** (mean, minimum and maximum) daily temperature at 2 m a.g
  - **Number of days** where temperature lower and higher than crop-specific thresholds ~ stress index

Methodology : time series processing and validation

- Resampling over thermal time ⇒ capture temporal anomalies
  - **Calendar time** : temporal anomalies could be related to the shift in a vegetation season
  - **Thermal time** : derivation to a multiannual average calculated for the same thermal time, i.e. **same development** stage

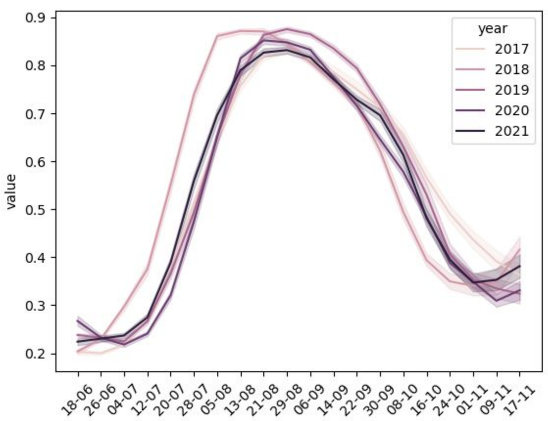


*Fig.4 : NDVI time series profile with **weekly periods***

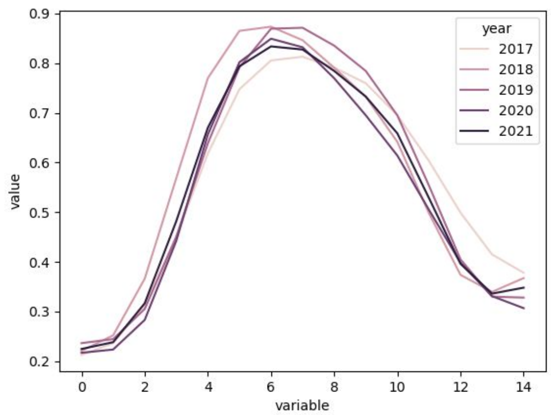


*Fig.5 : NDVI time series profile with 10-day periods **from planting date***

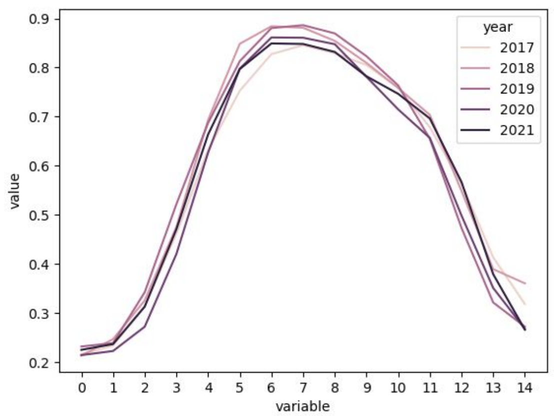


*Fig.6 : NDVI time series profile with 140-**GDD** periods from planting date*

 Thermal time resampled values ensure **year-to-year comparability** of vegetation conditions

syngenta

Methodology : Features

- Sentinel-2 (S2) time series :
  - Biophysical parameters and accumulated values
  - Standard deviation at the vegetation peak (i.e. NDVI is maximum) ⇒ field variability
- Agro-Meteorological (AM) data:
  - Averaged values between the vegetation peak and 5 periods before ~ **stress at vegetative phase**
- In-situ data:
  - Relative Maturity (RM)
    □ early maturities **require less heat** units to reach **physiological maturity**
  - Irrigated fields (dummy)
  - Geographical location ~ agricultural practices

⇒ Total : **69 predictors**

- S2 : 3 time series with 13 timestamps
- AM : 5 time series with 5 timestamps
- In-situ : 5 features

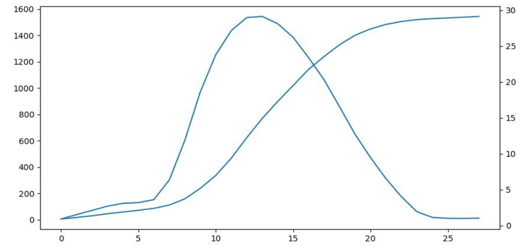


Fig.7 : Time series profile and accumulated values

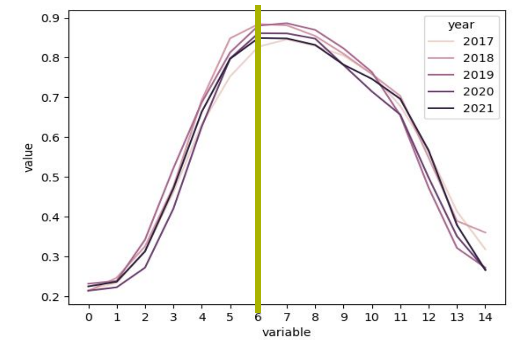


Fig.8 : Vegetation peak

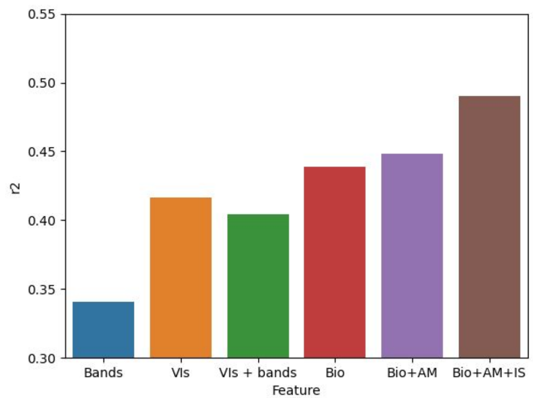
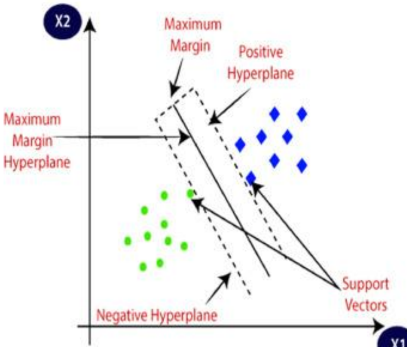


Fig.9 : R^2 w.r.t each feature group by fitting a RF with 10-folds random CV

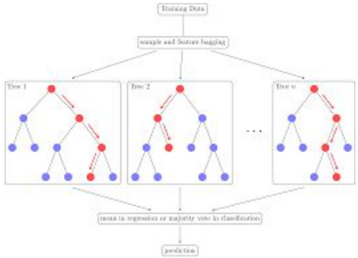
syngenta

Methodology : machine learning models

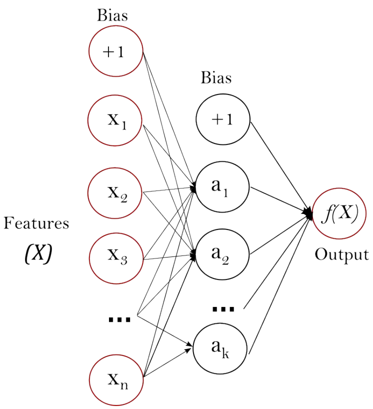
- **Support Vector Regression** (Cortes & Vapnik, 1995)
  - Finding hyperplane that has the maximum number of points.
  - Map the original feature space to some higher-dimensional space using kernel tricks.



- **Random Forest** (Breiman, 2001)
  - Combination of multiple individual decision trees to act as an ensemble
  - Random sub-samples of our dataset with replacement and calculate average prediction from each model.

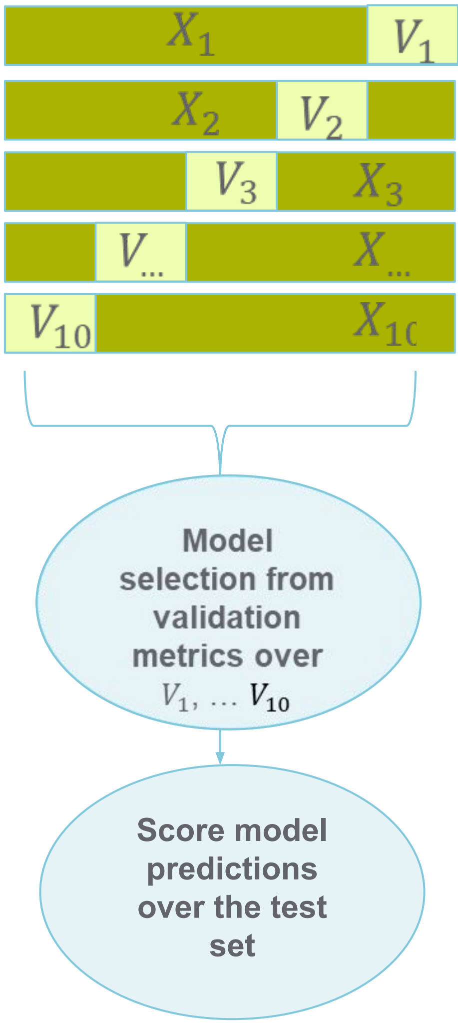


- **Multilayer Perceptron** (Haykin, 1994)
  - Learn a **non-linear function** approximation between the input and the output layer, with one or more non-linear layers ("hidden layers")



Experimental settings

- **<u>Training:</u>**
  - 90/10% training/validation (X_i, V_i)
  - 4 years as training/validation and 1 year of testing
- **<u>Model selection</u>**:
  - Choice of a model evaluation metric (R-square)
  - Model hyperparameters tuning:
    - Average the metric from each model configuration on the validation sets
    - Average the validation metrics over all testing years
  - Score the model selected prediction over each testing year
- **<u>Evaluation</u>**:
  - Compare daily and thermal time (e.g. RF_{daily} and RF_{GDD})
  - Define an ensemble ~ averaged model predictions



$$\hat{y}_{ens} = \sum_{i=1}^{10} \hat{y}_i$$

Results

Year	SVM_{daily}	SVM_{GDD}	RF_{daily}	RF_{GDD}	MLP_{daily}	MLP_{GDD}
2017	0.28 ± 0.01	0.27 ± 0.01	0.25 ± 0.01	**0.30 ± 0.01**	0.18 ± 0.08	0.30 ± 0.02
2018	0.28 ± 0.01	**0.31 ± 0.01**	0.25 ± 0.02	0.31 ± 0.01	0.29 ± 0.01	0.29 ± 0.03
2019	0.27 ± 0.02	0.31 ± 0.03	0.21 ± 0.02	0.32 ± 0.01	0.21 ± 0.05	**0.36 ± 0.02**
2020	0.37 ± 0.01	0.42 ± 0.02	0.40 ± 0.02	0.44 ± 0.01	0.38 ± 0.03	**0.50 ± 0.02**
2021	0.12 ± 0.02	0.37 ± 0.01	0.10 ± 0.02	0.30 ± 0.01	0.10 ± 0.07	**0.38 ± 0.03**
Average	0.26 ± 0.02	0.34 ± 0.02	0.24 ± 0.02	0.33 ± 0.01	0.23 ± 0.05	**0.37 ± 0.02**

*Table.1 : Results (R-squared) using **calendar time vs thermal time** for an unseen new year*

Year	SVM_{GDD}	$\text{SVM}_{GDD_{ens}}$	RF_{GDD}	$\text{RF}_{GDD_{ens}}$	MLP_{GDD}	$\text{MLP}_{GDD_{ens}}$
2017	0.27 ± 0.01	0.28	0.30 ± 0.01	0.30	0.30 ± 0.02	**0.32**
2018	0.31 ± 0.01	**0.32**	0.31 ± 0.01	0.32	0.29 ± 0.03	0.31
2019	0.31 ± 0.03	0.32	0.32 ± 0.01	0.33	0.36 ± 0.02	**0.38**
2020	0.42 ± 0.02	0.43	0.44 ± 0.01	0.44	0.50 ± 0.02	**0.51**
2021	0.37 ± 0.01	0.38	0.30 ± 0.01	0.31	0.38 ± 0.03	**0.40**
Average	0.34 ± 0.02	0.34	0.33 ± 0.01	0.34	0.37 ± 0.02	**0.39**

Table.2 : Ensemble the predictions from the 10-folds CV boost Multilayer Perceptron R-squared

syngenta

Conclusion

- **Conclusions** :
  - **Thermal time (GDD)** significantly improved results with a simpler model
  - Sentinel-2 ~ estimated of **Leaf Chlorophyll Content** is the best yield predictor
  - Environmental data ~ refining periods w.r.t the periods **before vegetation peak** from S2 time series improved results
  - Pipeline automatized at the field level in **python module** https://github.com/j-desloires/eo-crops

- **Perspectives :**
  - Tackle **domain shift** using domain adaptation techniques
  - **Article submission** in “Computers and Electronics in Agriculture” in September
  - Prepare **3rd year** PhD subject
    - Proposal ongoing : sowing date detection at field scale using unsupervised change detection

References

- Joel Segarra, Jose Luis Araus, Shawn C. Kefauver, Farming and Earth Observation: Sentinel-2 data to estimate within-field wheat grain yield, International Journal of Applied Earth Observation and Geoinformation, Volume 107, 2022, 102697, ISSN 1569-8432, https://doi.org/10.1016/j.jag.2022.102697.
- C. Dordas, “Nitrogen nutrition index and leaf chlorophyll concentration and its relationship with nitrogen use efficiency in barley (Hordeum vulgare L.),” J. Plant Nutrition, vol. 40, no. 8, pp. 1190–1203, 2017.
- Grégory Duveiller, Raul Lopez-Lozano, Bettina Baruth. Enhanced Processing of 1-km Spatial Resolution fAPAR Time Series for Sugarcane Yield Forecasting and Monitoring. Remote Sensing, MDPI, 2013, 5 (3), pp.1091-1116.
- Duveiller, G.; Baret, F.; Defourny, P. Using Thermal Time and Pixel Purity for Enhancing Biophysical Variable Time Series: An Interproduct Comparison. IEEE Trans. Geosci. Remote Sens. 2013, 51, 2119–2127
- Weiss, M.; Baret S2ToolBox Level 2 products: LAI, FAPAR, FCOVER; INRA: Paris, 2016.
- Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In Advances in neural information processing systems. 6402–6413
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297.
- Breiman, L. (2001) Random Forests. Machine Learning, 45, 5-32.
- Haykin S. Neural networks: a comprehensive foundation. Prentice Hall PTR; 1994.
- Amoukou et al., The Shapley Value of coalition of variables provides better explanations, 2021.