Keven Bosa, Kenneth Chu
Section des méthodes et de la qualité, DScD
Methods and Quality Section, DScD

HLG-MOS, UNECE
ML Virtual Sessions
October 15, 2020

Deploying Machine Learning Techniques for Crop Yield Prediction

Background – Field Crop Reporting Series (FCRS)

- Publishes final annual crop yield **estimates** towards **end** of each reference year.

- Also publishes full-year crop yield **predictions** a few times **during** reference year.
- In particular, contact farms in early July, ask them for their own full-year crop yield predictions. Publishes resulting yield predictions in August.

Yield prediction question was phased out from July data collection for **Manitoba** in 2019 (to reduce cost/response burden).

- A model-based method ("**baseline**") was used instead to generate the Manitoba/July crop yield predictions.
- July prediction $\rightsquigarrow$ early season prediction, deemed difficult.

Crop Yield Prediction Project

Question :

Can ML improve upon Baseline?

Approach :

Try and compare a (large) number of combinations of ML techniques and hyperparameter configurations

Main contribution :

Introduction of *rolling window forward validation*,[†] which mimics FCRS production setting, as validation protocol

[†] Schnaubelt, Matthias (2019) : A comparison of machine learning model validation schemes for non-stationary time series data, FAU Discussion Papers in Economics, No. 11/2019, Friedrich-Alexander-Universität Erlangen-Nürnberg, Institute for Economics, Nürnberg. http ://hdl.handle.net/10419/209136

Background – Data

- Availability : (2000, . . . , 2017) + (2018, 2019)
- Parcel-level†
  - **yield**‡ := (crop production) / (harvested area)
  - satellite (weekly, wks 16 – 31) : NDVI (normalized difference vegetation index)
  - crop insurance : insured crop type
  - geographical : Census Agricultural Region (CAR), eco-region, etc.
  - operational : seeded area, harvested area, etc.
- CAR-level
  - weather (weekly, wks 18 – 31) : total precipitation, average soil water content, etc.
- Derived variables of NDVI and weather time series
  - totals, maxima, rolling averages, etc.

† insured parcels only ; 1 parcel = 160 acres

‡ measured in (number of bushels) / acre

Underlying prediction/regression technique

Phase 1

XGBoost (Linear) | parcel-level within (eco-region × crop)

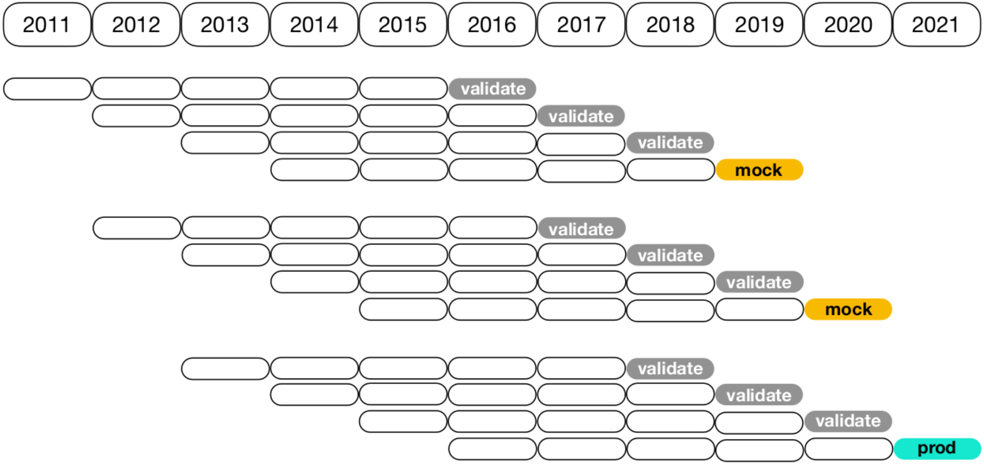
Phase 2

XGBoost (Linear) | parcel-level within (crop)

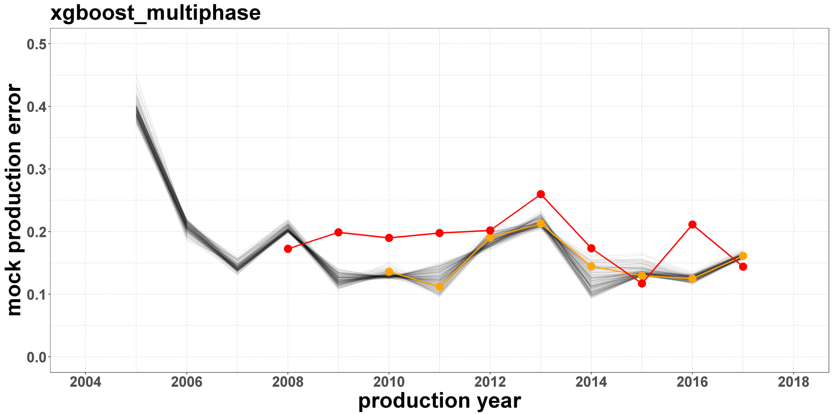
Question :

How to tune hyperparameters ?

Rolling Window Forward Validation – schematic



Preliminary results



- Each point : (year, h.config.)
- Red : Baseline mock production errors
- Orange : XGBoost/rwFV mock production errors
- Light gray : XGBoost(Linear) with 196 $(\alpha, \lambda_{\text{weights}})$'s
- Training window : five years
- Validation window : five years

Next Steps

`stcCropYield`

- R package
  - two-phase XGBoost(Linear)
  - rolling window forward validation
  - persisted trained model for use in production
  - documentation + sample code
- Near completion

Extend mock production to :

RY2018, RY2019 RY2020

Compare against baseline model

Personne-ressource

Pour plus d'information,
veuillez contacter :

For more information,
please contact:

Keven Bosa

keven.bosa@canada.ca

613-863-8964

Kenneth Chu

kenneth.chu@canada.ca

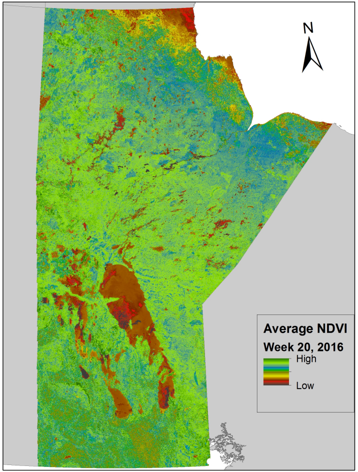
613-852-7361

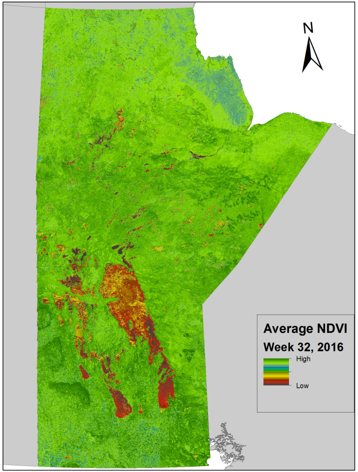
Cette présentation décrit des approches théoriques et ne présente pas des méthodes mises en œuvre présentement à Statistique Canada.
This presentation describes theoretical approaches and does not reflect currently implemented methods at Statistics Canada.

Statistics Canada Statistique Canada

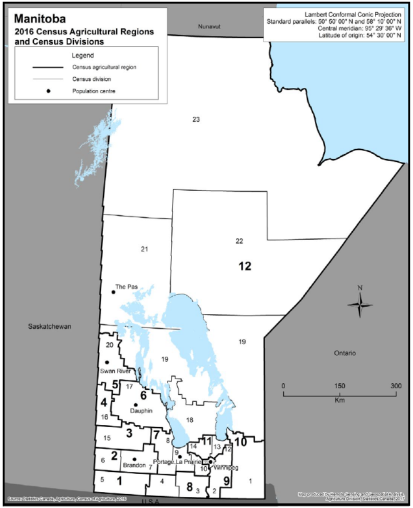
Canada

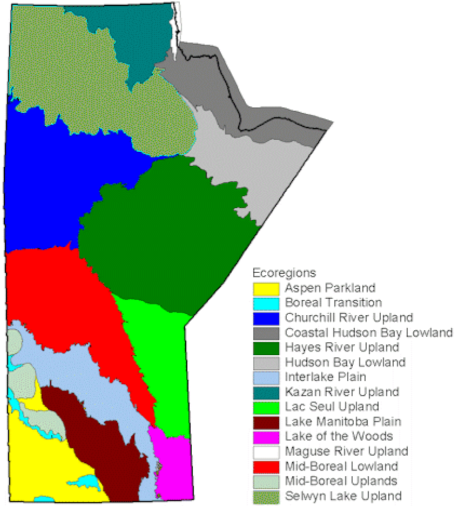
Background – NDVI





Background – CARS, eco-regions





Background – Baseline model

(deployed for Manitoba/July 2019)

$$\begin{pmatrix}\text{variable}\\ \text{selection}\\ \text{via Lasso}\end{pmatrix} + \begin{pmatrix}\text{robust}\\ \text{linear}\\ \text{regression}\end{pmatrix} \Bigg| \begin{array}{c}\text{parcel-level}\\ \text{within}\\ (\,\text{eco-region} \times \text{crop}\,)\end{array}$$

Rolling Window Forward Validation

- Take advantage of long history of available data (2000 – 2017).

- Mimic **multi-year production runs** :

  To generate **sequence(s) of yearly prediction errors** that would have been obtained for each candidate strategy had it been deployed in production in the past.

- Key design features : For each (ML method, hyperparameter configuration),
  - perform separately training/validation for consecutive reference years,
  - for each validation year, train a model based on data from **strictly preceding years**,
  - compute prediction errors for the trained model based on data from the validation year.

  Compare the (ML method, hyperparameter configurations)'s based on prediction errors.

Tuning objective function

$$\text{Across-validation-year average of}$$

$$\left(\begin{array}{c}\text{harvested-area-weighted} \\ \color{red}{\text{average}} \text{ of the } \mathscr{E}\text{'s}\end{array}\right)$$

where

$$\mathscr{E}\text{'s} \;:=\; \left(\begin{array}{c}\text{within-year (ecoregion, crop)-level} \\ \text{relative errors of crop production}\end{array}\right)$$

Reminder : crop production $=$ yield $\times$ harvested area

Tentative performance metrics

$(y, r, c) = (\text{year, eco-region, crop})$ and $(m, h) = (\text{ML method, hyperparameter configuration})$:

Crop production for (y, r, c) and predicted crop production for (y, r, c) and (m, h) :

$$P_{r,c}^{(y)} := \sum_{l \in (y,r,c)} \begin{pmatrix} \text{crop} \\ \text{yield} \end{pmatrix}_l \times \begin{pmatrix} \text{harvested} \\ \text{area} \end{pmatrix}_l , \qquad \widehat{P}_{r,c}^{(y,m,h)} := \sum_{l \in (y,r,c)} \begin{pmatrix} (m,h)\text{-predicted} \\ \text{crop yield} \end{pmatrix}_l \times \begin{pmatrix} \text{harvested} \\ \text{area} \end{pmatrix}_l$$

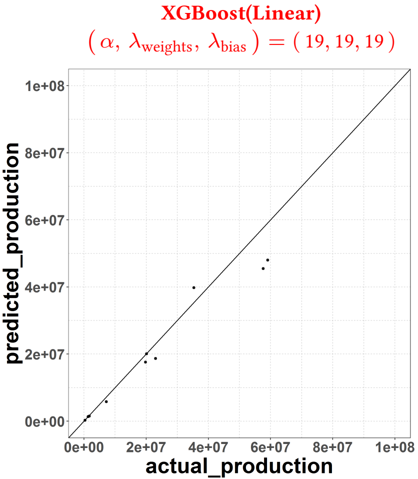
Production-induced relative error $\varepsilon_{r,c}^{(y,m,h)}$ and weight $w_{r,c}^{(y)}$ for (y, r, c), and number $N^{(y)}$ of nonzero weights for y :

$$\varepsilon_{r,c}^{(y,m,h)} := \left| \widehat{P}_{r,c}^{(y,m,h)} - P_{r,c}^{(y)} \right| \Big/ P_{r,c}^{(y)} , \qquad w_{r,c}^{(y)} := P_{r,c}^{(y)} \Big/ \sum_{(\xi,\zeta)} P_{\xi,\zeta}^{(y)} , \qquad N^{(y)} := \sum_{(\xi,\zeta)} 1_{\left\{ w_{\xi,\zeta}^{(y)} > 0 \right\}}$$

Production-weighted relative error and standard deviation for (y, m, h) :

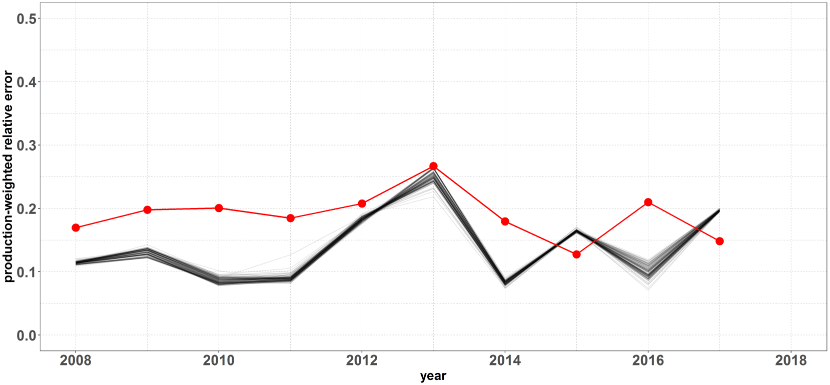
$$\mathrm{wErr}(y, m, h) := \sum_{(\xi,\zeta)} w_{\xi,\zeta}^{(y)} \cdot \varepsilon_{\xi,\zeta}^{(y,m,h)} , \qquad \mathrm{wSd}(y, m, h) := \sqrt{ \frac{N^{(y)}}{N^{(y)} - 1} \cdot \sum_{(\xi,\zeta)} w_{\xi,\zeta}^{(y)} \cdot \left(\varepsilon_{\xi,\zeta}^{(y,m,h)} - \mathrm{wErr}(y, m, h) \right)^2 }$$

Prototype results



- Validation Year : **2017**
- Training data : 2012, . . . , 2016.
- Each point : crop
- Absolute value of relative error :
  - Canola : 18.77%
  - Hard red spring wheat : 21.09%

Prototype results



- Each point : (year, method, h.config.)
- Red : Baseline
- Light gray :

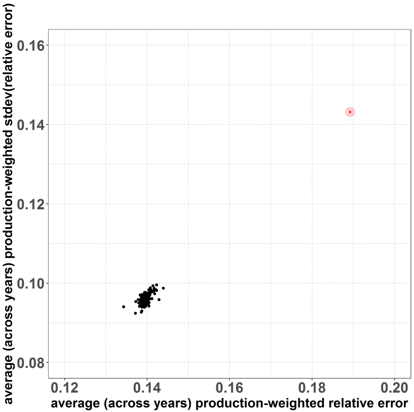
  XGBoost(Linear) with 125 $(\alpha, \lambda_{\text{weights}}, \lambda_{\text{bias}})$'s
- Included :

  Top 7 crops (by parcel count)
- Training window :

  five years

Prototype results

XGBoost(Linear) with
125 $(\alpha, \lambda_{\text{weights}}, \lambda_{\text{bias}})$'s



- Each point : (method, h. config.)
- Red : Baseline
- Light gray :

  XGBoost(Linear)
  with 125 $(\alpha, \lambda_{\text{weights}}, \lambda_{\text{bias}})$'s
- Included :

  Top 7 crops (by parcel count)
- Training window : five years