The Evolution of the Use of Satellite and Administrative Data in Estimating Mid-season Field Crop Yields at Statistics Canada

Herbert Nkwimi Tchahou

Chris Mohl / Gordon Reichert / Frédéric Bédard

May 16, 2022

Delivering insight through data for a better Canada

Background and Objective

Traditionally Statistics Canada conducted six field crop survey annually. In three occasions farmers were asked about crop yields

- July and September – mid-season estimated crop yields
- November – actual yields - the final values

Mid-season estimates were usually underestimates of the final values

The challenge

Can we make use of alternative data sources to reduce cost and response burden and produce a mid-season set of estimates of equal or better quality than the mid-season surveys provide?

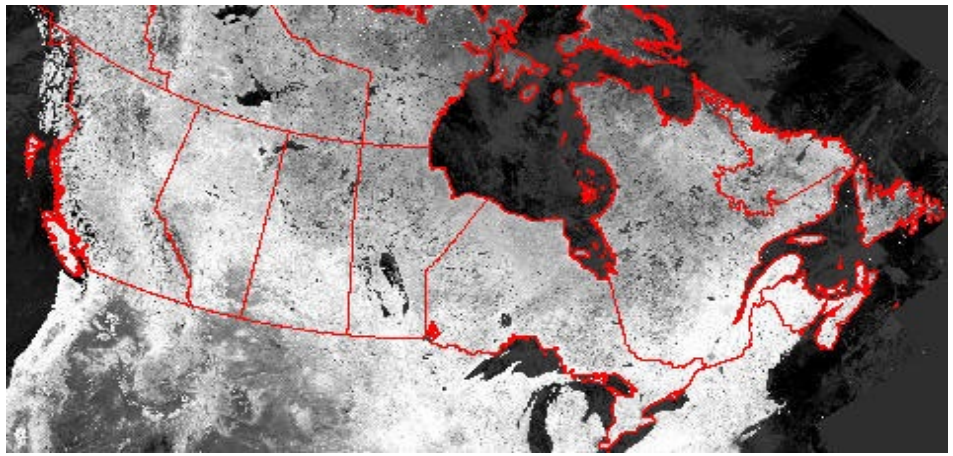
Alternative data sources – satellite imagery

From satellite imagery a weekly Normalized Difference Vegetation Index (NDVI) can be computed

- NDVI is used to determine the relative health of the vegetation using the red and near infra-red satellite channels
- Generally, the higher the NDVI value, the higher the crop yield potential

Two sources of images are available

- AVHRR (1 km resolution) 1987 to present
- MODIS (250 m) 2000 - present



Alternative data sources

Agroclimatic information

- Raw data provided by Environment and Climate Change Canada with value-added agroclimatic variables added by Agriculture and Agri-Food Canada (AAFC)
- Data associated with temperature, precipitation, soil moisture, sunshine hours etc.

Crop insurance

- Provided by provincial crop insurance corporations (requires data sharing agreements)
- Contains the location of the land parcel, what crop is being grown, the acreage of each crop sown and, after the growing season, the resulting yield

Challenges – satellite imagery

Atmospheric conditions affect quality

- Clouds and other contaminants
- Create a weekly portrait from the best of daily images

Intensive IT processes to process them (data transfer, disk space, computing capacity)

Comparability through time when satellites change (different resolutions)

- Can be controlled through calibration

Satellite Failure

- Rare occurrence

Challenges - agroclimatic data

Number of weather stations varies significantly from region to region

- Distance between a farm and the weather station can be quite far in some cases and not always representative

Location of weather stations can change over time

- Information from nearby stations is combined to create a complete data series over the required timeline (in some models, more than 30 years' of data are used)

Challenges – crop insurance data

Data acquisition: negotiations with data providers for access to their information

- StatCan provides its in-season field-level yield estimates to the crop insurance providers as an incentive to continue sharing the data
- Data not available for all provinces to use in the model

Uneven data structure from province to province

- Some provinces have more information than others

Not all crops are insured: Varies by region and crop

May be multiple crops growing within a land parcel (ex. quarter section)

- Cannot tell where each crop type is grown and associate its NDVI

Initial crop yield model

The goal was to replace the September survey collection with modelled yield estimates

The initial model relied solely on satellite imagery and agroclimatic data (no crop insurance).

Models estimated the yield at the Census Agriculture Region (CAR) level and combined them to the provincial level

- Between 500 and 5000 farms per CAR in most significant crop regions

Initial model

Starting point :

- multiple variable regression model developed at AAFC

Statistics Canada made a number of adjustments

- Use of LASSO (Least Absolute Shrinkage and Selection Operator) for variable selection (maximum of five variables)
- Adopted MM robust regression to account for outliers

Dependent variable:

- CAR-level November survey crop yield estimate

Predictors included:

- July survey crop yield estimate
- Weekly composite CAR-level NDVI values by crop
- Agroclimatic information from a single weather station within the CAR

Measuring quality and publication rules

Absolute relative differences between the model and the November survey yields were compared with those from the September survey over a 28-year period

- Results were comparable between the model and the September survey

Coefficients of variation (CVs) were calculated for each CAR-level crop yield estimate

- A CV value of 10% was considered to be of acceptable quality

A provincial crop yield was not published if

- The CARs with CVs over 10% accounted for at least 10% of the entire area of that crop in the province

Initial Publication

A parallel run was implemented in 2015 (national results below)

The September survey was cancelled in 2016 and replaced with modelled estimates

Earlier publication and a savings of approximately $150,000

	Model relative to November (%)	September survey relative to November (%)		Model relative to November (%)	September survey relative to November (%)
Barley	-11.0	-8.0	Soybeans	-0.5	-5.1
Canola	-17.3	-18.3	Durum Wheat	-10.3	-10.3
Corn for Grain	-4.0	-8.9	Spring Wheat	-8.0	-5.5
Flaxseed	-11.9	-9.3	Winter Wheat	-3.1	0.6
Mixed Grain	0.3	3.2	Canary Seed	-16.8	-21.8
Oats	-7.1	-7.4	Lentils	-17.3	-10.5
Dry Peas	-10.0	-2.1	Mustard Seed	-15.4	-13.1
Rye	0.5	2.2			

Crop insurance-based model

Goal: to eliminate both the July and September surveys. Start with the same general model but integrate crop insurance data as well

- Unit of observation is now the individual parcel, not CAR
- Dependent variable: yield reported to the crop insurance company for the parcel at the end of the year
- Drop the July yield as an explanatory variable
- Permit more than five explanatory variables to be selected
- Develop the model at the ecoregion level rather than CAR (ecoregion has more homogenous characteristics)

Model the yield estimates at the parcel level and combine the estimates up to the CAR, watershed, ecoregion, or provincial level

Crop insurance-based model

What to do with the parcels with more than one crop in them?

- Dropped from the model
  - Only parcels with 90% of the land covered by a single crop are retained.
  - In Manitoba, this eliminates about 2/3 of the parcels and ½ of the sown acreage
- Based on analyses of the yields reported to the insurance company, these mixed parcels had generally lower yields.
  - An adjustment was made to the estimate to compensate for this.

What to do with uninsured crops?

- Nothing. We assume they have a similar yield as insured crops

Crop insurance-based model

First studied in Manitoba. Nine years of estimates produced

- July modelled estimates are much closer to November values than those from the July survey
- September estimates are comparable with the survey
- In both cases, we were able to publish estimates for less common crops which we could not publish through the survey

This methodology is currently in use in Manitoba and Saskatchewan

- The initial model is used in other provinces until a sufficient number of years of crop insurance information is obtained.

Future work

Research on the use of Machine Learning methods

- Tests using XGBoost appear promising

Use of additional variables in the model

- Example: in one crop insurance dataset we have a field-level irrigation indicator

Transition to a protected cloud computing platform

- Enhance confidentiality and speed of image processing

Lessons learned

The methodology has shown to be a good replacement for mid-season surveys estimates

- Cost and response burden savings
- They still generally underestimate the values from the end of season survey

In extreme cases (2021 was a year of severe drought in the Prairies), the model struggles

Farm-level estimates can be less reliable

Relying on non-publicly available information can have risks

- Continuous negotiations with crop insurance corporations for access to their data

References

Documentation on the initial model

https://www.statcan.gc.ca/en/statistical-programs/document/5225_D1_T9_V1

Documentation on the crop insurance model

https://www.statcan.gc.ca/en/statistical-programs/document/3401_D2_V1

Thank you / Merci

herbert.nkwimitchahou@statcan.gc.ca

Cette information est aussi disponible en français