Canada - Crop Yield Modelling Using Remote Sensing, Agroclimatic Data and Statistical Survey Data

Gordon Reichert[1]*, Frédéric Bédard[1], Chris Mohl[2], Wesley Benjamin[2], Valéry Dongmo Jiongo[2],Aston Chipanshi[3], Yinsuo Zhang[4]
DOI: 10.1481/icasVII.2016.g43d

[1]Statistics Canada Agriculture Division 150 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6

[2]Statistics Canada Business Survey and Methods Division 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6

[3]Agriculture and Agri-Food Canada, AgroClimate, Geomatics, and Earth Observations Division2010 12th Avenue, Regina, Saskatchewan, Canada, S4P 0M3

[4]Agriculture and Agri-Food Canada, AgroClimate, Geomatics, and Earth Observations Division 960 Carling Avenue, Ottawa, Ontario, Canada, K1A 0C6

[1]Gordon.Reichert@Canada.ca
[1]Frederic.Bedard@Canada.ca
[2]Chris.Mohl@Canada.ca
[2]Wesley.Benjamin@Canada.ca
[2]Valery.DongmoJiongo@Canada.ca
[3]Chipanshi.Aston@agr.gc.ca
[4]Yinsuo.Zhang@agr.gc.ca

ABSTRACT

Statistics Canada's goal for modelled yield estimates was to produce a midseason estimate of crop yield and production based on information received as of the end of August, similar to what has been traditionally done with the September Farm Survey. The November Farm Survey estimates are considered the most accurate estimates of yield for a given year, due to the fact that the data are collected after the majority of harvesting has been completed and the sample size is the largest of the Field Crop Reporting Series. The modelled yield estimates and September Farm Survey estimates were both compared to the November Farm Survey estimates to verify the accuracy of the yield model results compared to the survey. Nineteen crops were introduced to the modelling process but published results were restricted to 15 when rules based on data availability and quality were implemented.

In 2015, the model-based yield estimates were disseminated for the first time by Statistics Canada as a supplemental publication 3 weeks in advance of the September Farm Survey and 11 weeks in advance of the November Farm Survey results.The modelled yield estimates had less deviation from the November Farm Survey than the September Farm Survey for canola, corn for grain,mixed grains, oats, rye, soybeans, and canary seed. Conversely, the September Farm Survey had less deviation from the November Farm Survey than the model for barley, flaxseed, dry peas, spring wheat, winter wheat, lentils, and mustard seed. Equal deviation was noted for durum wheat yield.

Feedback through government and industry consultation has been very positive and commencing in 2016, Statistics Canada replaced the September Farm Survey with the Model-based Principal Field Crop Estimates.

Key words: Remote Sensing, Normalized Difference Vegetation Index, Yield Model, Agriculture, Crop Statistics

1. Introduction

Innovative approaches in estimating crop yields are continuously being sought with the objective of reducing respondent burden while producing accurate, timely and reliable estimates. Statistics Canada, in cooperation with Agriculture and Agri-Food Canada, has developed a crop yield modelling approach as a non-intrusive method of producing yield forecasts that incorporates the 1 km resolution Normalized Difference Vegetation Index (NDVI) data used as part of Statistics Canada's Crop Condition Assessment Program, statistical survey data from Statistics Canada's Field Crop Reporting Series, and agroclimatic data for the agricultural regions of Canada. Although both the agroclimate and crop yield data had a longer time series, the study period was chosen according to the availability of the satellite data; a 29-year time series from 1987 to 2015.

Each year, Statistics Canada has traditionally conducted six farm surveys, in part, for estimating seeded area, harvested area,expected yield and production as part of the Field Crop Reporting Series. Like many other national statistical agencies, it is under increasing pressure to reduce response burden and cost of the traditional surveys while maintaining relevance, accuracy, timeliness, accessibility, interpretability and coherence.

Statistics Canada has therefore been researching and evaluating alternate methods of incorporatingadministrative data into its program to produce non-intrusive estimates of field crop yields and production.Agriculture and Agri-Food Canada (AAFC), has also been investigating the use of yield modelsfor the same purpose.To ensure no duplication of effort, a yield model that was being developed by AAFC using R statistical language software was transferred to Statistics Canada. The two organizations workedtogether on developing a robust yield model. Within Statistics Canada, the yield model was ported to a SAS platform.

The two departments modified the model with the goal of producing principal field crop yield estimates as of August 31.The 2015 modelled results were deemed of acceptable quality and were published by Statistics Canada3 weeks in advance of the September Farm Surveyresults and 11 weeks in advance of the November Farm Survey results.

This paper provides an overview of the background and general methods used to model reliable crop yield estimates as a preliminary estimate ofthe November Farm Survey estimates.

2. Methodology

A methodology for modelling crop yield was developed and tested in fiveCanadian provinces(Alberta, Saskatchewan, Manitoba, Ontario, and Quebec)forcrops that are typically published at the provincial and national levels by the September Farm Survey. These five provinces account for about 98% of the agricultural land in Canada and for the purpose of this paperare referred to as the national level when the yield model resultsare discussed.

2.1 Data sources used in the model

The modelling methodology used three data sources: 1) NDVI derived from coarse resolution satellite data (Latifovic et al., 2005)an integral component of Statistics Canada's Crop Condition Assessment Program (Bédard, 2010); 2) area and yield data collected through Statistics Canada's Field Crop Reporting Series, and 3) agroclimatic data for the agricultural regions of Canada.

2.1.1 Normalized Difference Vegetation Index (NDVI)

Since 1987, Statistics Canada has monitored crop conditions across Canada using the 1-km resolution, Advanced Very High Resolution Radiometer (AVHRR) sensor aboard the National Oceanic and Atmospheric Administration (NOAA) series of satellites. The NDVI was processed on a weekly basis throughout the growing season and used within Statistics Canada'syield model as a standardized index of vegetation health. These weekly NDVI values are available for download fromStatistics Canada's Canadian Socio-Economic Information Management System (CANSIM), Table 001-0100.

2.1.2 Survey area and yield data

Survey estimatesfrom Statistics Canada's Field Crop Reporting Series provided accurate and timely estimates of the seeded area, harvested area, yield and production of the principal field crops in Canada at the provincial level (Statistics Canada, Table 001-0010; Table 001-0017).

Results from the surveyswere only utilized in modelling activities when the crop wasrelatively abundant. If the crop wasabundant in a province, the yield estimates wereavailable at sub-provincial geographicunits.This finer level of geography usually correspondedto the Census Agriculture Regions (CAR) of which there are 82 across the agriculture region of the country (Statistics Canada, 2011). If the crop wasnot abundant, then yield estimateswereavailable at the provinciallevel only.

For abundant crops, CAR level crop yield estimates from the July and November Farm Surveys from 1987 to present were used as input variables for developing the model while yield estimates from the September Farm Survey and the November Farm Survey were used to validate the yield model results. For less abundant crops, the survey data and model results wereanalyzedat the provincelevel.

Area data from the June Farm Survey were used to aggregate yield estimates to larger geographic regions as described in Section 3.2. This area data along with yield data from the July and November Farm Surveys were used as part of the publication rules to determine which of the modelled yields were of acceptable quality for publication. The publication rules are described in Section 4.1.

2.1.3 Agroclimatic data

Climate data from 416 climate stations throughout the agriculture region of the five provinces was the third data source used as part of the crop yield modelling process. The station-based daily temperature and precipitation data provided by Environment and Climate Change Canada and other partner institutions were re-analyzedby AAFC to generate the climate-based predictors which amongst others included crop moisture stress, cumulative precipitation and growing degree days(Newlands et al. 2014, Chipanshi et. al. 2015). These data were provided to Statistics Canada by AAFC.

3. Modelling survey yields

3.1 Development of Statistics Canada'syield model

AAFC has an extensive history in developing field crop yield models. The most recent were documented in Newlands et al. (2014), and Chipanshi et al. (2015). These models incorporated non-Bayesian and Bayesian methods at different steps. The variable selection step used a non-Bayesian approach by the least-angle robust regression algorithm. Yields were then estimated using a Bayesian approach.

Statistics Canada had different modelling needs than AAFC. The AAFC model used Bayesian methods in order to estimate yields throughout the growing season at monthly intervals. Early season estimates were produced when data for the current year were not available. Unavailable data for the rest of the growing season were generated using a random forest method (Liaw and Wiener, 2002) which allowed crop yield results to be displayed as a probability.The Statistics Canada model was to be used in the middle of the growing season when the majority of thedata forthe current season werealready available, thereforethe Bayesian approach was not required. Statistics Canada also required that the model run on a SAS platform which is the standard programming tool used at the Agency. The AAFC models were programmed using Rstatistical language software.

Statistics Canada'smodelling goalwas to predict the final crop yield, therefore, the dependent variable of the model was the crop yield estimate from the November Farm Survey. There were 80 potential explanatory variables derived from the three data sources described in Section 2. Thus it was necessary to implement an appropriate method of selecting the model's explanatory variables. Bédard and Reichert (2013), established that the optimal number of explanatory variables to be selected for modelling was five. Khan et al., (2007), emphasized the importance of using robust modelling methods for selecting the explanatory variables for the model and estimating the yields. As there was no robust variable selection procedure in the SAS software it was necessary to use non-robust algorithms as an alternative at the selection step and then to estimate the model in a robust way. The Least Absolute Shrinkage and Selection Operator (LASSO) method was selected from the five variable selection algorithms available in SAS. The MM method (Yohai, 1987) was chosen from the robust regression methods available in SAS due to its ability to effectively treat outliers (Copt et al. 2006).

Preliminary evaluations were conducted using the data from 1987 to 2014. The median absolute differences in yield at the national level between the AAFC and Statistics Canada models for the seven largest crops in Canada were all between 0.9% and 2.4% (barley, 0.9%; canola, 1.0%;

corn for grain, 1.4%; durum wheat, 1.3%; oats, 0.9%; soybeans, 2.4%; and spring wheat, 0.9%)(Statistics Canada 2015). Since the two methods produced similar results, Statistics Canada made the decision to adopt a model using LASSO variable selection and the MM robust regression estimation in SAS.Throughout the remainder of the paper, results will only be discussed for this model used by Statistics Canada and will be referred to as the "yield model".

3.2 Aggregating modelled yield estimates

For the majority of the crops, modelling was done at the CAR level, the smallest geographic unitfor which historical survey data were available, or, for less abundant crops, the provincial level. The CAR level yield estimates are weighted based on seeded area estimates from the June Farm Survey and aggregated to produce a provincial estimate. For crops that are less common in a province, the model estimates werebuiltat only the provincial level. A similar weighting approach was used to aggregate provincial and the national yield estimates.

3.3 Model evaluation method

The November Farm Survey estimatesare considered the most accurate estimate of yield for a given year, due to the fact that the data are collected after the majority of harvesting is completed and the sample size is the largest of all six of the survey occasions. The results of the September Farm Survey can be considered a preliminary estimate of the November results. Therefore, Statistics Canada'sgoal for the yield model wasnot to replicate the results of the September Farm Survey but rather to obtain a sufficiently accurate yield estimate in advance of the November Farm Survey results.

The relative difference (presented as a percentage) between the yield estimate of a given method (i.e., September Farm Survey or the yieldmodel) and the November Farm Survey yield estimate was the measure of accuracy.A negative relative difference indicated that the given yield estimate was smaller than the November Farm Survey estimate, while a positive relative difference indicated that the given yield estimate was larger than the November Farm Survey estimate.

$$Relative\ difference = 100 * \frac{Given\ yield\ estimate - November\ Farm\ Survey\ yield\ estimate}{November\ Farm\ Survey\ yield\ estimate}$$

4. Publishing the yield estimates

Modelled crop yield estimates wereproduced at the CARlevel whenever possible and then rolled-upto the provincial and national levels.Statistics Canada has established three criteria based on data availability and quality that must be met to ensure the statistical integrity of the estimates and to determine which of the modelled crop yields were of acceptable quality for publication. Each year, the yield model estimates for individualcropsmust be evaluated to determine whether there is sufficient quality to warrant publication.

4.1 Publication rules for modelled yields

A minimum of 12 years of historical survey yield data for both theJuly and November Farm Surveys must be available as well as area and yield estimates for the current year from the June and July Farm Surveys, respectively. If these conditions are not met, then a modelled yield estimate will not be produced for that CAR or province.

The provincial yield estimate for a crop will not be published if the total cultivated area estimated by the June Farm Survey from suppressed regions (based on the previous set of conditions) exceeds 10% of the provincial area for the crop. Similarly, if provincial yield estimates for a crop were not published, the national level estimate will not be published if the total cultivated area for the suppressed provinces exceeds 10% of the national area.

Finally, if the coefficient of variation (CV) of the provincial or national estimate from the model was greater than 10%, the estimate was not published at that level. Model based CVs are calculated differently than those for survey estimates. Different CV thresholds are used to determine which estimates are suitable for publication than those used in the Field Crop Reporting Series. The 10% CV threshold for the model is the approximate equivalent to allowing a maximum absolute relative difference of 25% between the modelled yield and the November Farm Survey yield estimate.

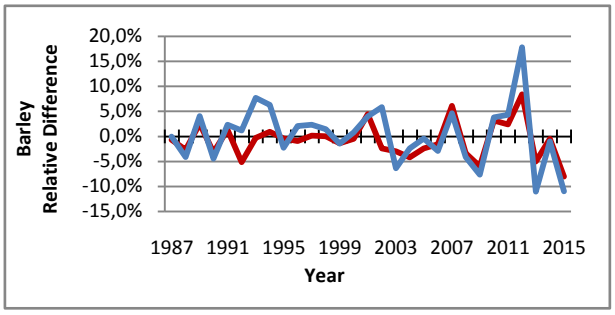
In cases where the estimates for some provinces were suppressed due to quality, but an estimate for the national level was still produced, only provincial estimates that were of an acceptable level of quality were used.

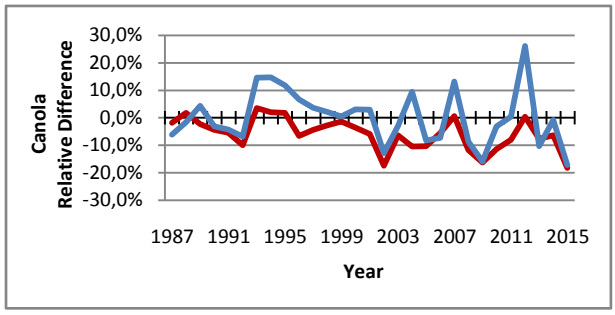
5. Results

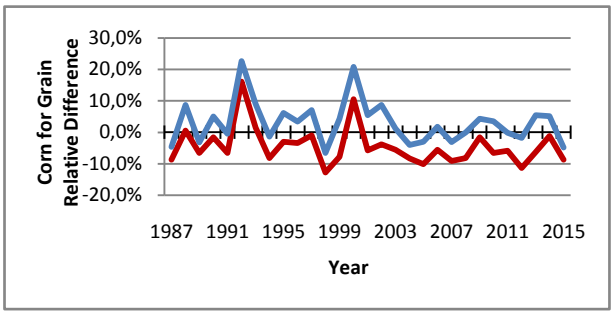
5.1 Comparisons of the modelled and survey yields

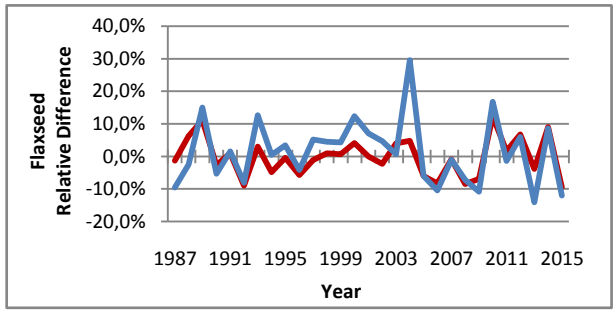
Nineteen crops were introduced to the modelling process at Statistics Canada but published results in 2015 were restricted to 15 when rules on data availability and quality were implemented. The four crops suppressed were chick peas, coloured beans, sunflower seed, and white beans.

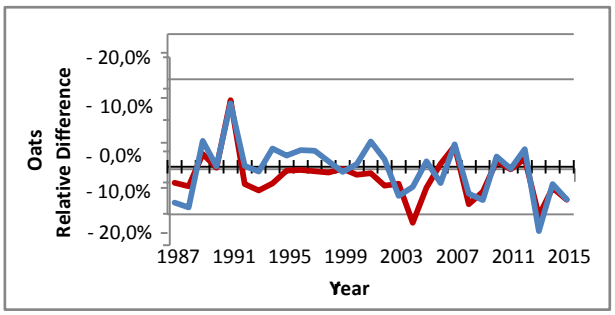
To verify the accuracy of the yield model,the relative difference of its yield estimates relative to those from the November Farm Survey were computed from 1987-2015. The September Farm Survey yield estimates were also compared to November Farm Survey yield results to provide a comparison of the accuracy of both methods.Figure 1 presents the comparison graphs for the 1987-2015 time series for the 15 crops for which modelled yield results were released in 2015.

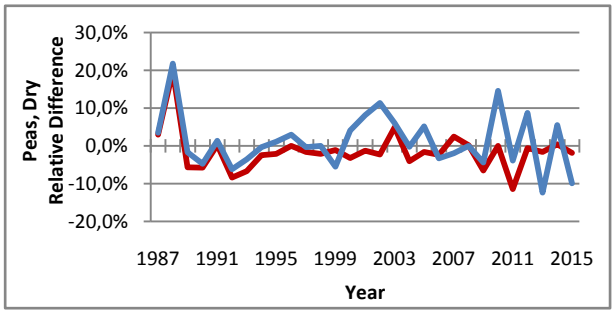


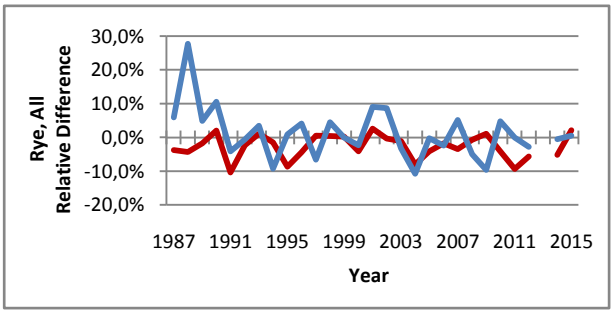












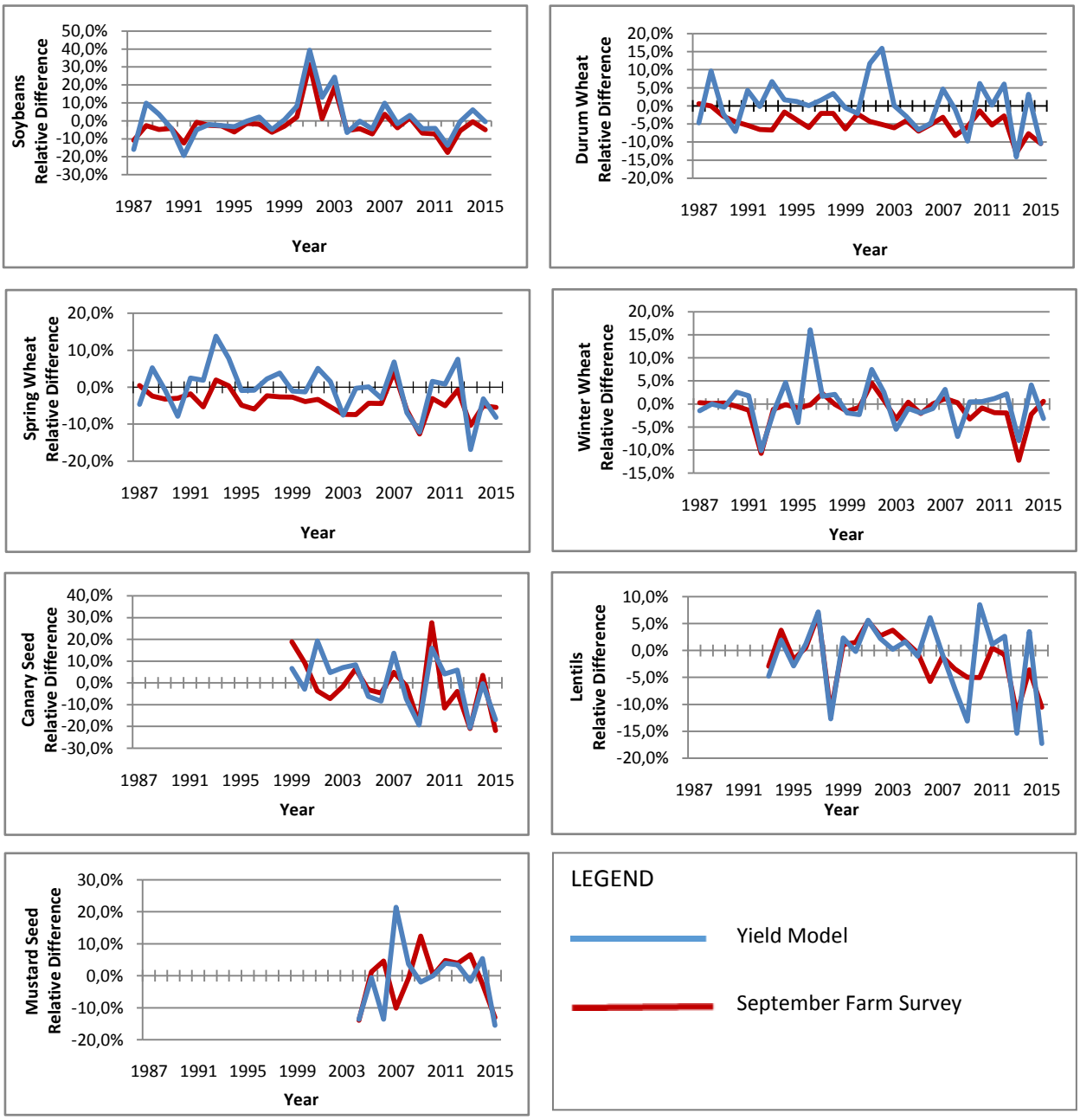


Figure 1. Relative difference of the yield model and the September Farm Survey from the November Farm Survey yields at the national level, 1987 to 2015.

The analysis shows that there is no consistent pattern when the yield model estimates and the September Farm Survey yield estimates are compared to the November Farm Survey for the 1987-2015 time series (Figure 1). Neither method is consistently closer to the November Farm Survey estimates for any crop. For soybeans and corn for grain, the two methods follow a similar pattern of estimates for the 29 years with regard to how the estimates change from year to year. However, this pattern is not present for the other crops. Additionally, for any given year, one method does not consistently perform better for all crops. In general, the yield model and the September Farm Survey yield estimates have comparable relative differences from the November Farm Survey estimates. However, the modelled estimates tend to have larger relative differences in

cases where an extreme relative difference is observed (e.g., the maximum and minimum relative differences are larger).

One pattern that can be seen is that the September Farm Survey results tend to be low when compared with the November Farm Survey results (below the x-axis) more often than the model results. This is particularly evident with canola, corn for grain, durum wheat, spring wheat, and rye. For more details on the comparative analysis refer to Statistics Canada, 2015.

On September 17, 2015, Statistics Canada disseminated the Model-based Principal Field Crop Estimates for the first time as a supplement publication 3 weeks in advance of the September Farm Survey estimates and 11 weeks in advance of the November Farm Survey estimates (Statistics Canada, Table 001-0075). Feedback to the modelled estimates through government and industry consultation has been very positive because of reduced response burden and reduced survey cost all while maintaining relevance, accuracy, timeliness, accessibility, interpretability and coherence.

Table 1 contains the 2015 summary comparison for yield and relative difference between the yield model and the November Farm Survey and between the September Farm Survey and the November Farm Survey.

Table 1. Summary comparison at the national level of Statistics Canada's yield model, the September and November Farm Survey, 2015.

Crop	Yield Model August 31, 2015	September Farm Survey	November Farm Survey	Yield Model compared to November Farm Survey	September Farm Survey compared to November Farm Survey
	Yield (bushels per acre)			Difference (%)	
Barley	**57.8**	**59.8**	**65.0**	**-11.0**	**-8.0**
Canola	**32.6**	**32.2**	**39.4**	**-17.3**	**-18.3**
Corn for grain	**158.9**	**150.8**	**165.5**	**-4.0**	**-8.9**
Flaxseed	**20.5**	**21.1**	**23.3**	**-11.9**	**-9.3**
Mixed Grain	**65.6**	**67.5**	**65.4**	**0.3**	**3.2**
Oats	**79.6**	**79.4**	**85.7**	**-7.1**	**-7.4**
Peas, dry	**29.1**	**31.7**	**32.3**	**-10.0**	**-2.1**
Rye, all	**38.2**	**38.8**	**38.0**	**0.5**	**2.2**
Soybeans	**43.3**	**41.3**	**43.5**	**-0.5**	**-5.1**

Wheat, durum	**30.9**	**30.9**	**34.5**	**-10.3**	**-10.3**
Wheat, spring	**40.5**	**41.6**	**44.0**	**-8.0**	**-5.5**
Wheat, winter	**62.8**	**65.2**	**64.8**	**-3.1**	**0.6**
	Yield (pounds per acre)			**Difference (%)**	
Canary seed	**865**	**813**	**1,040**	**-16.8**	**-21.8**
Lentils	**1,151**	**1,246**	**1,392**	**-17.3**	**-10.5**
Mustard seed	**711**	**731**	**841**	**-15.4**	**-13.1**

The September and November yield estimates listed in Table 1 have been adjusted to take into account any suppression that was applied during the yield modelling process as described earlier thereby providing a normalized comparison of the results between the three occasions.

The yield model had less deviation for canola, corn for grain, mixed grain, oats, rye, soybeans, and canary seed. Conversely, the September Farm Survey had less deviation than the model from the November Farm Survey for yield for barley, flaxseed, dry peas, spring wheat, winter wheat, lentils, and mustard seed. The two methods had equal deviation for durum wheat yield.

In general, the results from 2015 yield model and the September Farm Survey estimates had deviations from the November Farm Survey estimates of varying degrees. For certain crops the yield model estimates had less deviation while for others the September Farm Survey estimates had less deviation. Both methods produce estimates that can be both very similar to the November estimates for some crops while having more significant deviation for other crops.

8. Summary

The estimates produced by the yield model were comparable to those produced by the September Farm Survey in terms of relative difference from the November Farm Survey estimates for the 15 crops modelled.

In 2015, modelled yield estimates for field crops deemed to have a sufficient level of quality were published as a preliminary estimate 3 weeks in advance of the September Farm Survey estimates and 11 weeks in advance of the November Farm Survey results. Statistics Canada consulted with provincial and federal government counterparts, members of the grain industry, and academia regarding the yield model. Based on a proven, non-intrusive, scientific method and the

strong outreach support coupled with the federal government's desire to reduce respondent burden and survey cost, it was decided that, commencing in 2016, Statistics Canada would replace the September Farm Survey with the Model-based Principal Field Crop Estimates. The replacement of a statistical field crop survey with a remote sensing model-based administrative data approach is a first for any statistical agency worldwide. Moving forward, Statistics Canada and Agriculture and Agri-Food Canada are evaluating methods of using other administrative data sources (such as crop insurance and additional satellite crop classification data) to derive crop area estimates which can be used in conjunction with the modelled yield estimates to create reliable estimates of crop production.

9. REFERENCES

Bédard, F., 2010. *Methodology document: Satellite image data processing at Statistics Canada for the Crop Condition Assessment Program (CCAP)*. http://www23.statcan.gc.ca/imdb-bmdi/pub/document/5177_D1_T9_V1-eng.htm

Bédard, F., and Reichert, G, 2013. *Integrated Crop Yield and Production Forecasting using Remote Sensing and Agri-Climatic data*. Analytical Projects Initiatives final report. Remote Sensing and Geospatial Analysis, Agriculture Division, Statistics Canada

Chipanshi, A., Zhang, Y., Kouadio, L., Newlands, N., Davidson, A., Hill, H., Warren, R., Qian, B., Daneshfar, B., Bedard, F. and Reichert, G, 2015. *Evaluation of the Integrated Canadian Crop Yield Forecaster (ICCYF) Model for In-season Prediction of Crop Yield across the Canadian Agricultural Landscape.* Agricultural and Forest Meteorology, vol. 206, pp 137-150. I: http://dx.doi.org/10.1016/j.agrformet.2015.03.007

Copt, S., and Heritier, S, 2006. *Robust MM-Estimation and Inference in Mixed Linear Models*. Cahiers du département d'économétrie, Faculté des sciences économiques et sociales, Université de Genève. http://www.unige.ch/ses/metri/cahiers/2006_01.pdf

Khan, J. A., Aelst, S. V., and Zamar, R. H., 2007. *Robust Model Selection Based on Least Angle Regression*. Journal of the American Statistical Association, Vol. 102, No 480, pp. 1289-1299. http://dx.doi.org/10.1198/016214507000000950

Latifovic, R., Trishchenko, A. P., Chen J., Park W.B., Khlopenkov, K. V., Fernandes, R., Pouliot, D., Ungureanu, C., Luo, Y., Wang, S., Davidson, A., Cihlar, J., 2005. *Generating historical AVHRR 1 km baseline satellite data records over Canada suitable for climate change studies.* Canadian Journal of Remote Sensing, vol. 31, no 5, pp 324-346.http://pubs.casi.ca/doi/abs/10.5589/m05-024

Liaw, A. and Wiener, M., 2002. Classification and Regression by Random Forest. R news, 2(3): 18-22. https://cran.r-project.org/doc/Rnews/Rnews_2002-3.pdf

Newlands, N.K., Zamar, D., Kouadio, L., Zhang, Y., Chipanshi, A., Potgieter, A., Toure, S., Hill, H.S.J., 2014. *An integrated model for improved seasonal forecasting of agricultural crop*

yield under environmental uncertainty. Front. Environ. Sci. 2, 17. Doi: http://dx.doi.org/10.3389/fenvs.2014.00017

Statistics Canada. 2011. *Reference Maps and Thematic Maps, Reference Guide.*Census year 2011. Catalogue no. 92-143-G. http://www.statcan.gc.ca/pub/92-143-g/92-143-g2011001-eng.pdf

Statistics Canada, 2015. *Integrated Crop Yield Modelling Using Remote Sensing, Agroclimatic Data and Survey Data.*http://www23.statcan.gc.ca/imdb-bmdi/document/5225_D1_T9_V1-eng.htm

Statistics Canada. 2016. *Crop Condition Assessment Program*.http://www26.statcan.ca/ccap-peec/start-debut-eng.jsp

Statistics Canada. *Table 001-0010 - Estimated areas, yield, production and average farm price of principal field crops, in metric units, annual,* CANSIM: http://www5.statcan.gc.ca/cansim/a26?lang=eng&retrLang=eng&id=0010010&&pattern=&stByVal=1&p1=1&p2=50&tabMode=dataTable&csid

Statistics Canada. *Table 001-0100 - Normalized difference vegetation indices at one kilometre resolution by land use type for agricultural areas of Canada, weekly (index),* CANSIM: http://www5.statcan.gc.ca/cansim/a26?lang=eng&retrLang=eng&id=0010100&&pattern=&stByVal=1&p1=1&p2=50&tabMode=dataTable&csid

Statistics Canada. *Table 001-0017 - Estimated areas, yield, production, average farm price and total farm value of principal field crops, in imperial units, annual,* CANSIM: http://www5.statcan.gc.ca/cansim/a26?lang=eng&retrLang=eng&id=0010017&&pattern=&stByVal=1&p1=1&p2=50&tabMode=dataTable&csid

Statistics Canada. *Table 001-0075 - Model-based Principal Field Crop Estimates, in metric and imperial units, annual,* CANSIM:http://www5.statcan.gc.ca/cansim/a26?lang=eng&retrLang=eng&id=0010075&&pattern=&stByVal=1&p1=1&p2=50&tabMode=dataTable&csid

Yohai, V.J, 1987. *High breakdown-point and high efficiency robust estimates for regression*, The Annals of Statistics, Vol. 15, pp. 642-656. http://www.jstor.org/stable/2241331