Proceedings of Statistics Canada Symposium 2014
Beyond traditional survey taking: adapting to a changing world

The Potential Use of Remote Sensing to Produce Field Crop Statistics at Statistics Canada

James Brisbane and Chris Mohl[1]

Abstract

In an effort to reduce response burden on farm operators, Statistics Canada is studying alternative approaches to telephone surveys for producing field crop estimates. One option is to publish harvested area and yield estimates in September as is currently done, but to calculate them using models based on satellite and weather data, and data from the July telephone survey. However before adopting such an approach, a method must be found which produces estimates with a sufficient level of accuracy. Research is taking place to investigate different possibilities. Initial research results and issues to consider are discussed in this paper.

Key Words: Remote Sensing; Agriculture; Crop Monitoring; Crop Statistics.

1. Introduction

Each year, Statistics Canada carries out six surveys of field crops, estimating areas and yields. Each survey is done by telephone interviews with farm operators. The estimates are used by the crop industry and by government policy-makers. An important statistic for the crop industry is the in-season prediction of the level of production of individual crops. They are required to make good predictions of price, which in turn are required for business decisions.

Like many other national statistical organizations, Statistics Canada is under increasing pressure to find alternative approaches to traditional surveys for generating statistics in order to reduce response burden and cost. The use of remote sensing and administrative data is being examined as an alternative to replace one occasion of the current field crop surveys. In the context of this paper, remote sensing refers to the use of satellite information and images as input to models which are used to estimate the parameters in question. Statistics Canada has done some small-scale trials on individual crops with this technology in the past twenty years, however it has not been used in production, nor been used to cover a wide range of crops at one time.

This paper discusses some of the goals of the research on the use of remote sensing and its associated challenges. In section 2, the current strategy for collecting field crop information is presented. Section 3 discusses some of the potential data sources that can be used for producing field crop statistics. In sections 4 and 5, the use of remote sensing and administrative data for modelling estimates of crop yield and areas respectively are discussed. Some results from an initial yield model are presented. Section 6 describes some of the future work that will be taking place and is followed by conclusions.

2. Current methods for producing field crop statistics at Statistics Canada

The crop industry benefits from the best possible estimates of crop production. These estimates play an important role as predictions of the supply of crops from the upcoming harvest, and help to set the price that farm operators can expect to receive for their crops. For these estimates, two components are required: the area harvested and the

[1]Statistics Canada, Business Survey Methods Division, 150 Tunney's Pasture Driveway, Ottawa, Ontario, Canada K1A 0T6.

yield. Within a season, as crops grow, new weather events occur, new information becomes available, and so it is important to update the estimates over time. Consequently, Statistics Canada conducts six occasions of its field crop telephone survey per year, asking farm operators about fifty-two different crops. Table 2-1 shows the timing of the six surveys, the approximate sample size, and the information collected in each survey.

Table 2-1
Statistics Canada's telephone surveys for field crop statistics

Month of data collection	Approximate sample size (in 2014)	Quantities estimated
March	11,500	• Planned area to be seeded for each crop • Amount of each crop in storage
June	24,500	• Actual area that has been seeded for each crop
July	13,100	• Actual area that has been seeded for each crop • Expected area to be harvested for each crop • Expected yield for each crop • Amount of each crop in storage
September	9,300	The same quantities as for July, except crop in storage
November	26,800	The same quantities as for September
December	8,700	Only the amount of each crop in storage.
Total	93,900	

Each survey collects information independently of other occasions. Farm operations are surveyed at most twice during the year. The most accurate estimates of harvested area and yield are produced by the November survey, after the harvest is complete, but these estimates are not available until early December, and it's important to produce accurate estimates before that time. These are referred to as in-season estimates. During the summer, Statistics Canada could potentially make use of other data sources to estimate expected harvested areas and yields. One course of action under consideration is to cancel the September telephone survey and publish modelled estimates in September based on data from remote sensing, the July telephone survey, and other available sources such as weather stations.

3. Potential data sources for producing crop estimates.

Statistics Canada is evaluating how different sources of data can be used to produce field crop statistics. Possible sources include

1. Conventional telephone surveying of farm operators
2. Direct observation of fields (by staff using cars or planes)
3. Crop insurance data
4. Satellite data
5. Weather station data.

Each of these five data sources comes with challenges.

Conventional telephone surveying of farm operators. Statistics Canada's agriculture survey program places a significant burden on farm operators. Managing response burden has been identified as a priority for the agency**.**

Direct observation of fields. This is expensive due to travel costs and requires knowledgeable staff to carry it out. It also requires an area frame of fields which Statistics Canada no longer maintains.

Crop insurance data. This indicates only how much crop has been insured. Estimates of the uninsured areas are still needed from other sources.

Satellite data. When using satellite images to estimate crop areas, each image requires the development of a classification algorithm specifically for that image. The accuracy of this classification must be tested, by comparing it against small samples of land, known as "ground-truth", where the identity of the crop is assumed known without error.

Weather station data. This can be used in conjunction with other sources to estimate yield. Its usefulness depends on how well it represents the weather conditions in the fields where the crops are grown.

Statistics Canada has obligations to produce reliable statistics, to be cost-effective in all its undertakings, and to minimize response burden. The present research aims to design a program with lower response burden and cost by increasing the use of data from crop insurance sources, satellites, and weather stations, in conjunction with conventional surveys, while maintaining an acceptable level of data quality. The cancellation of the September survey occasion and conducting only five telephone surveys per year would be a way to reduce the burden. Under such a plan, Statistics Canada would still publish harvested yield estimates in September, but they would be calculated using satellite and weather data, and estimates from the July telephone survey. Over the short term the area estimates would be carried forward from the July survey. In the longer term, remote sensing approaches for area estimates could be adopted.

4. Estimating crop yield from satellite, weather, and telephone survey data

Statistics Canada is examining the use of regression models to estimate in advance what the yield estimate from the November survey will be. Models and estimates are developed separately and independently for each census agricultural region (CAR) of the country and the results are aggregated to the provincial level. A CAR is a sub-provincial area of land composed (usually) of Census Divisions with similar agricultural properties such as climate and soil type.

One of the predictor variables in the model is the yield estimate from the July telephone survey. The other predictors come from various types of satellite and weather station measurements. The satellite measurements consist of a calculated Normalized Difference Vegetation Index (NDVI) measured at different times during the growing season and averaged over the crop area of the CAR. NDVI is a common satellite-based measure of the amount of green vegetation, measured by the light sensor on the satellite. The other variables are measurements made at weather stations including total precipitation, growing degree days, water stress index, and soil water content. The first two are cumulated over different time periods during the growing season, and the second two are averaged over different time periods. Then in all four cases the numbers are averaged over all stations in the CAR to produce the predictor variable. It is important to consider different time periods because the sensitivity of crop yield to weather varies over time.

In the results presented in this paper, around sixty predictor variables were produced using data from the start of the growing season to early September and the best five were selected using the stepwise selection method for multiple linear regression. Twenty-seven years of data (1987-2013) is available to generate the models. The regression model was then used to estimate the November survey yield estimate. Graphs are presented below, comparing this estimate with the November survey yield estimate from the July and September telephone surveys.

Figure 4-1 shows the spring wheat yield in Saskatchewan over the last twenty-seven years, as estimated by the November survey occasion. Saskatchewan is the largest province in terms of field crop production, and spring wheat is one of its principal crops. Average yield per acre is about 30 bushels. While the general trend is upwards, yields can vary significantly between years. For example, the yield was about 15 bu below average in 1988 and about 18 bu above average in 2013. These outlying data points make modelling the yield more challenging.

Figure 4-2 shows the relative differences between the November survey estimate of spring wheat yield and the three predictions over the last twenty-seven years. A positive value means that the survey estimate is higher than the prediction, and a negative value means that it is lower than the prediction. This initial multiple linear regression model produced estimates which were visibly different than those from either the July or September survey

occasions. Whereas the survey data tended to underestimate the yield in comparison to the November estimates, the model overestimated them slightly more than half of the time. The magnitude of the differences is also of note.

Figure 4-1
Yield estimated by the November telephone survey for 27 years of spring wheat in Saskatchewan

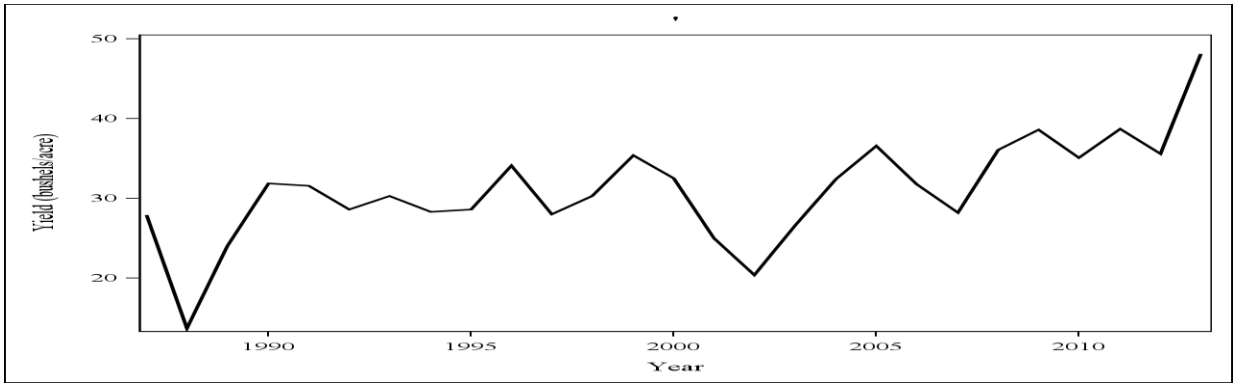
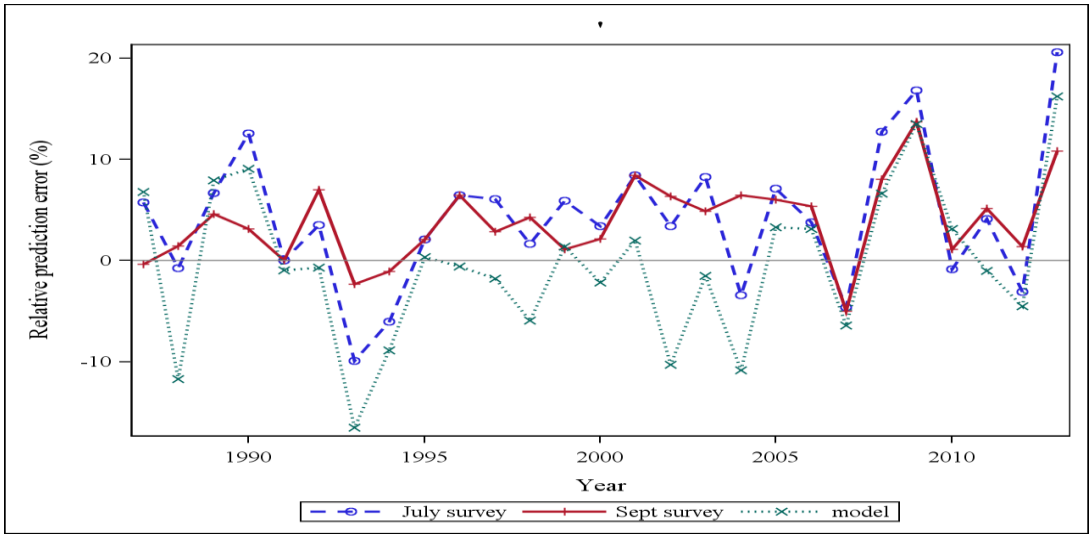


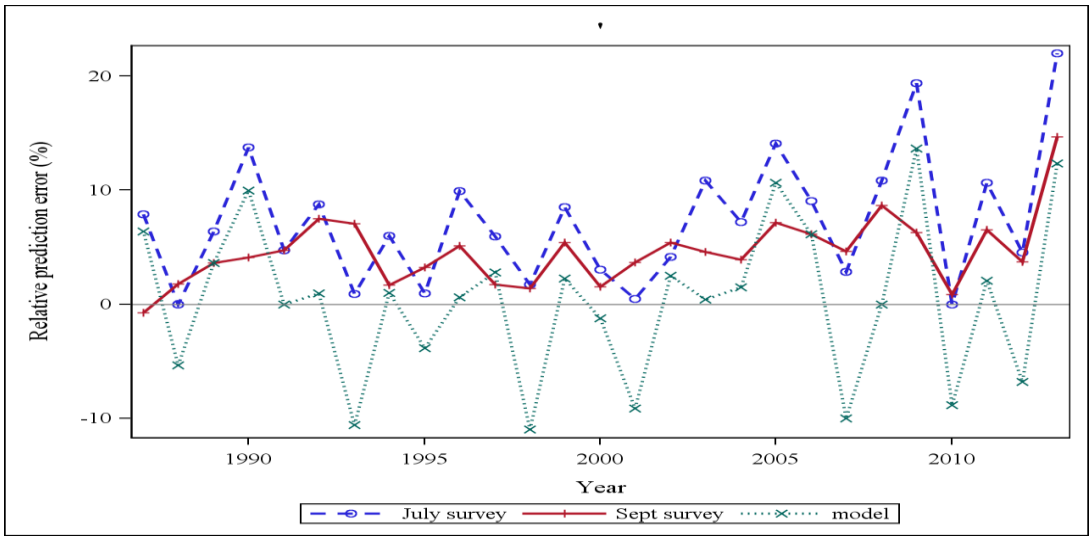
Figure 4-2
Prediction errors relative to the November survey estimates for 27 years of spring wheat in Saskatchewan



From an absolute point of view the modelled numbers deviated from the November estimated yields more than the survey numbers. Not surprisingly, the model had some difficulty predicting the yield in outlying years when the yield was much lower (as in 1988 or 2002) or higher (as in 2013) than usual.

Figure 4-3 shows a similar graph for Saskatchewan durum wheat. While still an important crop in Saskatchewan, it typically has less than half the acreage of spring wheat. The resulting graph shows similar patterns to spring wheat. The July and September survey occasions tended to underestimate the production whereas the model showed a more even distribution of over and underestimation. However the spikes indicate that the model had much more instability from one year to the next and tended to more poorly predict the November numbers compared to the September survey. This is the typical pattern that was observed among the five provinces and seven important crops that were examined with the initial model. It indicates that some investigation into improved modelling is necessary to address some of the current model's weaknesses. A standard multiple linear regression model may be insufficient to properly address some of the outliers in the historical datasets.

Figure 4-3
Prediction errors relative to the November survey estimates for 27 years of durum wheat in Saskatchewan



5. Estimating crop area from satellite image data

This involves the satellite capturing an image of the ground. The ground is sampled using an area frame and the crops growing on the sampled areas are recorded. This sample data is called ground-truth because it's assumed free from errors in the same way that respondents are assumed to give the correct answers in a telephone survey. Two ground samples are selected, and all crops whose areas are to be estimated must exist in both. With the first sample, ground data is compared to the data of the satellite image, and that comparison is used to develop a classifier algorithm, referred to as simply "the classifier".

The classifier is a set of decision rules mapping the light reflectance values at different wavelengths to a crop. Different algorithms exist for developing the classifier. Statistics Canada uses the See5 algorithm provided by RuleQuest software (www.rulequest.com) with free images provided by the Landsat 8 satellite of the US Geological Survey. Landsat 8 provides reflectance measured over six wavebands, which can improve classifier accuracy over cases where fewer wavebands are measured. Using several images taken on different days can also improve accuracy. The output data containing the predicted identity of the crop grown in each area of land is called "classification data". The strength of the method is that the classifier produces classification data for the whole population area covered by the satellite image, even though the ground-truth covers a relatively small area.

If the method is implemented well, the classification can be quite accurate. However the population total area classified to a given crop is always a biased estimate, to some extent, because of the so-called errors of omission and commission (see Congalton and Green (1999)). The bias is removed by using the second ground sample as the conventional survey sample, and using the classification data as auxiliary data for the estimation. Calibration or model-based estimation can be used to remove bias and produce estimates with known sampling variance. For this, and more information, see the European Commission, Joint Research Centre, MARS (1999).

Statistics Canada has identified two major challenges in using satellite image data. Firstly, the classifier has difficulty distinguishing crops that look similar from a distance. For example, wheat and barley can't be distinguished, and in some cases it's difficult to distinguish between soybeans and potatoes. Since acreages by crop type are required, this will need to be addressed. A simple solution might be to split the estimated acreage of the indistinguishable crops into individual crops using the proportions observed in the July survey. Secondly, it is difficult to estimate areas of rare crops with good relative accuracy without a large ground sample. If the rare crop does not exist in the ground sample, the area estimate is zero. An estimate of zero has a small absolute error, but a

large relative error, because the true area is small and non-zero. A solution could be to get data on rare crop areas from crop insurance or from a ground survey collecting GPS-mapped data from known growers of rare crops. This non-randomly sampled data could then be used to develop a classifier for satellite image data.

Crop insurance plays a potential role in the estimation of crop areas, but its use has limits. As previously mentioned, crop insurance data only provides information about the insured crop area. It contains no information about the uninsured area. It can be used as the first ground sample used to develop the satellite classifier, but it cannot be used as the main survey sample to estimate the uninsured area. However, it can be put to good use in the estimation as auxiliary data. It can be used to correct errors in the satellite classification data, before it is used in the estimation.

6. Future work at Statistics Canada

A decision on whether to cancel the September 2015 survey occasion and replace it with modelled estimates will be made near the end of 2014. If yes, survey yield estimates will be replaced with yield estimates from a regression model. In the meantime additional research will take place to improve the accuracy of the initial models and address some of the weaknesses that were observed. The research to this point has focused solely on seven major crops. The model will also need to be robust enough to generate sufficient estimates for some rarer crops. Agriculture and Agri-Food Canada (AAFC) has also been investigating how best to make use of satellite information for crop prediction (Newlands et. al, 2014). Statistics Canada will continue to work with AAFC with the common goal of being able to derive accurate and timely estimates using remote sensing technology.

The second phase of the project is to more deeply investigate the potential use of remote sensing to predict crop acreage. Work up to this point has already indicated that there will be some challenges in distinguishing individual field crops. This is not an issue which was encountered in some of the small, specialized investigations that have taken place at Statistics Canada in the past. As part of this study, in the summer of 2014 Statistics Canada collected a ground-truth sample from three sub-provincial areas in Canada. This data was gathered by trained Statistics Canada personnel visiting the individual fields and noting the crop growing in them. This information can be used to better understand the potential limits of remote sensing as well as the extent to which crop insurance data might be integrated into the acreage estimation process.

Over the longer term, Statistics Canada will need to be vigilant in examining additional alternatives for producing its field crop statistics. Farm operations are continually adopting more smart technology into their daily work. One potentially exciting source of information is the data that some farm equipment now generates which can tabulate how much cropland is being farmed and later, what the production from this cropland is.

7. Conclusion

The use of remote sensing and satellite technology has the potential to play a role in the production of field crop statistics at Statistics Canada. The fact that they can be used with little or no burden on the farm operator makes them appealing as Statistics Canada searches for ways to produce statistics in manners other than traditional survey taking. On the other hand, a more thorough investigation of the advantages and disadvantages of these approaches must be done before they can be used as part of the official data collection for the survey. The quality of estimates resulting from these approaches must be sufficient to meet the needs of the data users. Crop insurance information is another source of data which, on its own, is insufficient for producing complete crop statistics, but which can play an important complementary role in the process.

8. Acknowledgements

The crop yield predictions described in section 4 of this paper were produced by Frédéric Bédard, from the Remote Sensing and Geospatial Analysis Section of Statistics Canada's Agriculture Division. The authors thank Frédéric Bédard, Gordon Reichert and Martin Renaud, for their helpful comments on this paper.

References

Congalton, R.G. and K. Green (1999). *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*, Boca Raton, FL: Lewis.

European Commission, Joint Research Centre, MARS (1999), Best Practices for Crop Area Estimation with Remote Sensing. URL: http://mars.jrc.ec.europa.eu/mars/Bulletins-Publications (accessed December 2014).

Newlands, N.K., Zamar, D., Kouadio, L.A., Zhang, Y., Chipanshi, A., Potgeiter, A., Toure, S., and H.S.J. Hill (2014), "An Integrated Probabilistic Model for Improved Seasonal Forecasting of Agricultural Crop Yield Under Environmental Uncertainty", *Frontiers in Environmental Science*, 2(17).

.