EVALUATION OF THE CANADIAN POTATO AREA ESTIMATION PROGRAM

Patricia J. Whitridge, Claude Poirier, Statistics Canada
Business Survey Methods Division, 11-I R.H.Coats Bldg, 120 Parkdale Ave, Ottawa, Ontario, K1A 0T6

Key Words: Regression Estimates, Remote Sensing

1 INTRODUCTION

The Potato Area Estimation Program (PAEP) produces estimates of potato acreage and potato yield in several Canadian provinces. It includes several surveys whose design may differ between the provinces. The surveys that are conducted by Statistics Canada are: the Potato Survey in Manitoba, Saskatchewan and British Columbia, the Potato Objective Yield Survey (POYS) in Prince Edward Island (P.E.I.) and New Brunswick (N.B.), the Remote Sensing Program held also in Prince Edward Island and New Brunswick, the January Livestock Survey and the June Crops Survey. This report evaluates POYS and the Remote Sensing Program in P.E.I. and N.B.

2 POTATO OBJECTIVE YIELD SURVEY (POYS)

2.1 Sample Design

The POYS estimates the potato acreage and yields in P.E.I. and N.B. This includes estimates of the total area planted and the average harvested yield of potatoes. The 1993 target population consisted of all of the farms in both provinces that planted potatoes in 1993 but the survey frame included only the farms that reported potato acreage in the 1991 Census of Agriculture. Adjustments were done to the estimates to compensate for the geographical areas excluded. The frame, containing 906 farms, was divided into a specified stratum for large farms, then five other strata according to size. The size was defined by the farm potato acreage. The sigma-gap method was used to determine the boundary for the specified farms, which were selected in the sample with certainty. A sample of about 150 farms for each province was allocated optimally between the first five strata. The selected farms were contacted in June and September to produce respectively preliminary and final area estimates.

A sub-sample was drawn for yield data collection in September. To collect the yield data, fields were selected within each sampled farm and an Agriculture Canada inspector went to the farm to dig a part of the selected fields and to weigh the potatoes found. Yield estimates were produced based on these data.

2.2 Non-Response

The non-response for POYS included farmers who refused to respond and those that were not contacted because they were not available at the collection time. Whenever possible, the non-responding farms were replaced by other farms selected in the same stratum. A list of replacement farms for each stratum was prepared at the time of sampling. In 1993, a total of 30 non-respondents, representing 10% of the sample, were replaced by 28 farms in the two provinces together.

2.3 Frame Deficiencies

The frame was built using the known potato producers present on the Farm Register, which is a list of all farmers identified by the 1991 Census of Agriculture. Potato data from the Census were edited and then imputed as necessary using the nearest neighbour donor imputation method. As a result, zero potato acreage may have been erroneously assigned to some producers or, conversely, positive acreage could be assigned to non-producers. Also, some farms have gone out of business since the last census and some new ones have started production. This represents frame overcoverage and undercoverage respectively.

The frame overcoverage was estimated using the out-of-business farms and the farms that no longer produced potatoes, both identified at the collection stage. The resulting estimates showed that 87 farms (i.e. 10% of frame) were not in business or no longer produced potatoes.

The undercoverage is more difficult to evaluate than overcoverage. By matching the POYS frame with the potato producers identified by the Farm Financial Survey, we estimated the P.E.I. undercoverage to be 105 farms, representing 4,100 acres. The N.B. undercoverage was estimated to be 110 farms totalizing 6,600 acres. Since the undercoverage was not considered in the estimation, the survey figures may underestimate by about 10%.

For future occasions of the survey, we should make sure the frame uses up-to-date potato information, available from other agriculture surveys. Other surveys identifying potato production are the January Livestock Survey, the Farm Financial Survey, the June Crops Survey and the Area Farm Survey. The farms which do not produce potatoes should be included in a zero stratum and be sampled with a small sampling fraction. This would provide an estimate of the farms which have begun potato production since the Census and would make up for the frame undercoverage. This could be done by increasing the overall sample size rather than by allocating a part of the current sample to this stratum.

2.4 Sampling Errors

When the frame was built, the imputed potato acreage available on the 1991 Census was used to identify the potato producers. This acreage, called the frame value, represented a proxy for the survey value and was used to define the farm size in the stratification process. The frame values could be summed to obtain a frame total. They could also be used when the sample was selected to produce a frame estimate with a coefficient of variation (CV) from the selected farms. By comparing the frame estimate with the frame total, we can get an idea of how good the sample might be. Table 1 shows the frame total acreage, the frame estimate and the final POYS survey estimate for both P.E.I. and N.B.

From this table we can evaluate the relative difference between the frame totals and the frame estimates. This difference should be less than 2*CV, 19 times out of 20. The relative difference observed are 0.31% for P.E.I. and 1.66% for N.B. Since both are lower than twice the CVs, there is no evidence to conclude that the sample is unrepresentative. The CVs increased in the final estimates because the survey data, which were unknown at the stratification stage, were different from the frame data. Compared to the CVs of the frame estimates, the final POYS CVs were respectively 2.1 and 2.7 times bigger for P.E.I. and N.B.

3 REMOTE SENSING PROGRAM

3.1 Sample Design

The Remote Sensing estimates were based on a sample of geographical cells. First, both P.E.I. and N.B. were divided into cells of 2 by 3 km for the 1992 survey year. Values representing potato acreage were assigned to each cell based on the analysis of 1991 satellite images, with some information taken from 1990 images where there were clouds in 1991. Cells with at least one acre of potatoes were kept on the frame. In N.B., as for the POYS survey, cells outside the potato belt were removed. Adjustments were done to the estimates to compensate for all such frame exclusions. The frame was stratified into five strata based on size, defined here by the cell potato acreage. A sample of 60 cells for each province was allocated optimally to the strata. The selected cells were flown over in July to identify their potato acreage using air photos. Estimates based only on the aerial surveillance were produced.

Auxiliary data, in the form of satellite imagery, were available that corresponded to the aerial data. Satellite imagery was acquired from either LANDSAT or SPOT satellites depending on which one covered the studied regions within a clear weather period. The data were received through the Gatineau, Quebec receiving station, between July 27 and August 10, 1993. A SPOT image was used to cover the P.E.I. central zone whereas the images covering the two P.E.I. ends and N.B. came from LANDSAT. Note that both satellites do not have the same resolution. The most accurate, the SPOT satellite, has a resolution unit of 400 m^2 compared to 900 m^2 for the LANDSAT. Spectral analyses were performed on the satellite data to provide an auxiliary potato value for almost all frame units. The data from the satellite images, by themselves, cannot be used to produce a reliable estimate of potato acreage. From the satellites, the number of pixels in potatoes can be obtained, but the "ground truthing" using the aerial surveillance data is required to train the computer to recognize these pixels as potatoes. The direct sum of the potato pixels does not give a representative indication of potato acreage, since a pixel is not 100% certainly composed of potatoes. There is always some amount of "noise" or error associated with the classification. This noise changes from one year to the next, depending upon the degree of maturity of the potato plants, the weather when the image was taken and other variables. Using the sampled cells, the relationship between the satellite images and the air photos is established. This relationship is used to bring the sampled data from the aerial surveillance closer to the population data from the satellite images. The CVs are then reduced and the estimates are adjusted for any lack of representativeness in the sample. By using the satellite data in conjunction with the air photo data through a regression estimator, a

Table 1: 1993 Design and POYS Estimates with their CVs

	P.E.I.	N.B.
Frame Total	76,038	49,568
Frame Estimate CV (%)	75,802 1.61	50,418 0.90
POYS Estimate CV (%)	82,805 3.39	49,833 2.43

second set of estimates can be produced. The relationship is subject to the same noise due to the conditions surrounding the satellite image, so the parameters must be recalculated each year.

It was felt that data based on more than one year would provide a stronger frame for the sample design, since there can be a number of factors affecting the quality of data for any given year. It was expected that the frame would be recreated each year using data from more than one year to improve the quality. The sample would then be rotated by around 25%, providing good quality estimates of both levels and trends. For 1993, no updates were made to the P.E.I. frame or sample due to cost and time constraints. The frame for N.B. was updated using the average of the values derived from the satellite image analyses from 1991 and 1992. A replacement of 12 cells, corresponding to 20%, was done when the 1993 N.B. sample was redrawn. This difference between the treatment from 1992 to 1993 for P.E.I. and N.B. would provide an opportunity to examine the impact of frame recreation and rotation.

There are many different ways in which data from multiple years could be combined to create the sampling frame. A straight average was used in this case for N.B. However, in the future, other combinations should be considered. For example, the sum of the potato pixels from the satellite data is available for each year, as is the published estimates of potato acreage. The published estimates could be used to benchmark the potato pixels for each year. Several years of this combined data could then be averaged. Such a procedure would provide stable frame data, strengthened by taking into account possible crop rotations, different weather conditions, and different stages of potato growth over time.

3.2 Non-Response

The quality of the satellite images depends on many factors such as the weather, the ability to differentiate crops, the time of year, and the crop conditions. Clouds can make the analysis of affected regions very difficult , if not impossible. The auxiliary data may then be unavailable for some cells, constituting non-response. An additional factor that generated non-response in 1993 was that the N.B. satellite image did not correspond to the one requested. The image that was received was shifted south by about 15 km and did not cover the northern region of the province.

Because of its size, P.E.I. required 3 images to be covered this year, and even then a gap was observed between the images, causing non-response. Any non-response related to the satellite data was accounted for by an adjustment factor applied to the total of the available data at estimation. The problem became more important when the satellite data did not exist for some of the cells of the aerial sample, because the regression parameters were based only on this sample. In 1993, satellite data were imputed for 9 sampled cells to allow estimation of regression parameters. In these cases, the frame values were used for imputation. Table 2 details the non-response observed in 1993.

3.3 Quality of the Satellite Data

The quality of the satellite data is very difficult to evaluate. The final estimate from the satellite data depends partly on which fields were used to train the computer to recognize potatoes. From different training fields, different results can be observed andthere is no way to identify the best ones. The choice of the training fields seems somewhat

Table 2: Non-response Related to Satellite Data

	P.E.I.	N.B.
Cells on frame	830	537
Cells in cloudy regions	47	9
Cells not covered by satellite	3	33
Total number of cells without data (% of the potato acreage affected)	50 (9.1%)	42 (1.3%)
Sampled cells without data	5	4

subjective. The possibility of selecting these fields using stronger probabilistic concepts should be considered. It would be useful to study the contribution of the actual field selection process to the regression estimation.

From this process, we understand that the computer was trained using part of the aerial sample which may have introduced a bias into the estimated CVs. This is because the computer probably classified the training fields with more success than any other fields since they were used to model the pixel recognition. Since the correlation was measured on the aerial sample, including the training fields, it may have looked better than what it should have been, underestimating the regression CV for the same. This means that the regression estimates looked more precise than they really were. The bias consequently depends on the number of fields used for training and may be considered low in our case. Ideally, training should have been done on cells which were not used in the estimation.

The potato area identified by spectral analysis must be correlated to the air photo data to provide reliable regression estimates. This is measured by the correlation coefficient ρ which varies between 0 and 1 for positive dependence, 0 indicating no correlation at all and 1 indicating a perfect model. In 1993, regression seemed good in N.B. with a coefficient $\rho=0.93$, but in P.E.I., with $\rho=0.43$, we did not expect reliable estimates. Note, that the P.E.I. coefficient is provided here as one composite number even though 3 regressions were applied in practice. The coefficients measured in the three zones were respectively 0.61, 0.31 and 0.50. It is difficult to explain why the central zone presented the worst correlation when it was covered by the SPOT satellite with the best resolution. It seems that the bad weather observed during the summer of 1993 did upset the farmers' schedules and, the potato growing period became different between farms, causing a lack of consistency in the spectral analyses. By looking at historical estimates in Table 3, we can see that such a poor correlation is unusual.

Table 3 also gives the total acreage derived from the satellite data. As mentioned before, this total is not considered in the Potato Area Estimation Program because it is not very reliable. By comparing the satellite totals with the current estimates, we can easily observe that they are not stable. This does not represent a problem in the regression estimator because the only property required from the data is a good correlation with the aerial data, not to represent a good indicator of the potato acreage.

3.4 Sampling Errors

The sampling error in the Remote Sensing Program was measured using the coefficient of variation (CV). As explained in 3.1, two sets of estimates were produced: one using only the air photo data and a second using both air photo and satellite data combined through a regression. The 1993 CVs associated with the air photo estimates were relatively high but this was expected since the survey was not designed specifically for these estimates. The CVs were to be reduced by the use of regression. In N.B., it was reduced to a level comparable to POYS. On the other hand, the regression CV observed in P.E.I. did not really decrease due to the poor correlation mentioned in 3.3. To reduce the P.E.I. CV to something similar to POYS, given the 1993 data quality, the P.E.I. sample used for the aerial surveillance would have be doubled to 115 cells. Table 3 gives the 1993 estimates in addition to the ones obtained for the two previous years.

Table 3: Historical Acreage Estimates

	NEW BRUNSWICK						PRINCE EDWARD ISLAND					
	Current estimate	Estimates & CVs (%)				ρ	Current estimate	Estimates & CVs (%)				ρ
		POYS	Aerial	Regr.	Sat.			POYS	Aerial	Regr.	Sat.	
1993	51000	49833 2.43	52358 7.40	56679 2.48	32142	.93	87000	82805 3.39	92846 8.15	87840 7.35	79973	.43
1992	53000	52627 2.49	54102 8.18	57362 5.62	16913	.73	85000	77300 2.73	89649 11.04	84192 6.13	70305	.85
1991	50600	50106 2.70	47465 10.70	49872 4.02	20739	.92	77800	75268 3.90	79115 12.70	77777 2.68	81487	.97

In Table 3, we can observe that the aerial and the regression estimates are generally higher than the POYS estimates. This may signify that the POYS estimates are negatively biased as mentioned in 2.3. A sample drawn from the non-producers of potatoes would be a possible solution.

For any estimates, a confidence interval may be built centred on the estimate with a relative length of ± 2CV. It can be observed that the POYS and the Regression intervals do not overlap in New Brunswick. This puts emphasis on the fact that either the POYS or the Regression estimate may be biased.

3.5 Use of the Satellite Data

The satellite data are used for more the regression estimators. The satellite data represent the best option for frame values. The frame was built and updated based on the satellite data. Census values could also be used but they are not up-to-date and they are available at the EA level only. To derive segment values we would have to make many assumptions, especially regarding farms or EAs overlapping more than one segment.

The satellite data are also used in the aerial program when the potato fields are digitized. In that process, the potato fields outlined in the photos must be captured in the computer. To do this, an up-to-date version of the segment area must be loaded as background on the computer screen. This allows the operator to identify the fields of interest and to digitize them. The computer image used is provided by the satellite data.

4. CONCLUSIONS

The Potato Area Estimation Program produces three estimates of potato acreage from POYS and the Remote Sensing Program. Each of these estimates can be improved by changing some aspects of the sampling design and/or the estimation procedure.

First, for POYS, we recommend that the survey frame be created using the Farm Register data as well as any potato information collected by agricultural surveys like the January Livestock Survey, the Farm Financial Survey, the June Crops Survey and the Area Farm Survey. We also recommend that a zero stratum be created and a small sample be selected from it. This would reduce the frame undercoverage and would eliminate a potential bias in the estimates. The farms selected in the zero stratum would be selected only for the acreage data collection, which is conducted by phone. The overall sample size would be increased rather than allocating a part of the existing sample size to the zero stratum.

For the Remote Sensing Program, we recommend that the 1994 survey frame be updated using a three year average benchmarked each year to the published potato estimate, based on the 1991, 1992 and 1993 satellite data. This would stabilize the frame values by reducing the impact of the crop rotation. A simple two year average was used for N.B. in 1993, but a benchmarked average should be done in future to take into account the fact that the satellite data are not stable. The frame would be restratified and the 1994 sample would be independently redrawn by maximizing its overlap with the 1993 sample, for both P.E.I. and N.B. Additional sample replacement should be done to reach a total replacement rate of 25%. This would result in more reliable estimates on a multi-year basis. The impact of rotation and frame recreation for 1993 in N.B. should be examined.

The selection of the training fields should use stronger probabilistic concepts. This would eliminate any subjective decisions in the computer training process.

The use of satellite data in conjunction with the aerial data is supposed to reduce the CVs. In 1993, this was not observed in P.E.I. because the data were poorly correlated with each other. If the regression estimates are not as reliable as expected, the possibility of increasing the aerial sample size should be considered to reduce the aerial CV itself, without having to do any regression. Historically, the aerial estimate was considered to be closer to the truth than the regression estimate.

Note, that any decisions concerning the needs of the aerial and/or satellite projects should consider the fact that they are both related. For example, the digitizing of the potato fields of the aerial project and the survey frame updates require satellite data collected on a regular basis.

Sources of information and expertise about survey methodology and remote sensing technology need to be combined to see if there are not better uses that could be made of data collected through remote sensing means.