ESTIMATING POTATO AREA USING AERIAL SURVEILLANCE

Chantal Belley, Wesley Yung and Richard Dobbins, Statistics Canada
Chantal Belley, 3-A R.H. Coats Building, Tunney's Pasture, Ottawa, Ont., Canada, K1A 0T6

Key Words: Aerial surveillance, PPS sampling, Remote sensing

Abstract:

For the 2001 crop season, the Spatial Analysis and Geomatics Applications unit and Business Survey Methods Division of Statistics Canada developed a methodology to generate an experimental estimate of potato area of Prince Edward Island. This approach involved using a combination of statistical sampling, geographic information system technology and aerial surveillance techniques. The potato area estimate and data quality indicators produced under this project are comparable to other estimates produced in 2001 by Statistics Canada using traditional statistical survey methods.

1. Introduction

Prior to 2001, the Prince Edward Island Land-Cover Classification project (PLCCP) was undertaken by the Spatial Analysis and Geomatics Applications (SAGA) unit of Statistics Canada to provide annual land-cover classification (i.e., major crops and forests) of Prince Edward Island, to the Prince Edward Island Department of Agriculture and Forestry (DAF). In addition to providing land cover data, experimental crop area estimates were also generated for the major crops, which include potatoes, hay, pasture and grains.

For 2001, funding for the PLCCP by DAF was not available. However, due to factors such as speculation that potato area in PEI would decrease substantially in 2001, (because of an existing trade embargo imposed by the United States), there was interest at Statistics Canada in obtaining an aerial surveillance-based estimate for the province. This estimate would also be used to corroborate Statistics Canada's Potato Survey estimates and also to validate the 2001 Canadian Census of Agriculture (CEAG) potato data. Thus, a proposal to produce potato area estimates using a statistical sampling and estimation approach in conjunction with aerial surveillance was made and approved for the 2001 crop season. This project, referred to as the Potato Area Estimation Project (PAEP), was carried out for the 2001 crop season.

In this paper, we describe the proposed project methodology and report on the results. In Section 2, we describe the PLCCP project as implemented in the past. Section 3 describes the sample design employed by the 2001 PAEP, as well as, the data collection methods and results. Finally, conclusions and recommendations are presented in Section 4.

2. The Prince Edward Island Land-Cover Classification Project

In general, the PLCCP combined the processing of multi-date Landsat digital satellite imagery with ground "truth" data, which was acquired on a sample basis by aerial surveillance, to produce land-cover data (i.e., crops and forest lands) throughout PEI. Using digital image processing software, it was possible to distinguish between different crops. Generally, this separation is made as each land-cover type reflects a distinctive wavelength of light providing a unique spectral signature. Therefore, by associating the crop type with each spectral signature on the Landsat image data, it is possible to identify, separate and map land-cover classes throughout each satellite image and the entire province.

Using Geographic Information System (GIS) software, a grid of 1,217 cells, with each cell being 2 km by 3 km in dimension, was constructed for the entire province. A random sample of 75 cells was selected for ground "truthing", which consisted of identifying all crops visually within each of the 75 cells, either by means of aerial surveillance, or in case of questionable data and as required, supplemented by field verification. This information was then used to "train" the digital image analysis system to separate crops and forestlands and to classify the digital satellite images. Once the crop and forest classifications were generated, provincial crop area estimates were easily generated using the GIS software.

3. The Potato Area Estimation Project Survey Design

To minimise the cost of the PAEP, a combined statistical sampling and estimation method using aerial surveillance was proposed. In this section, the survey design of the PAEP is described.

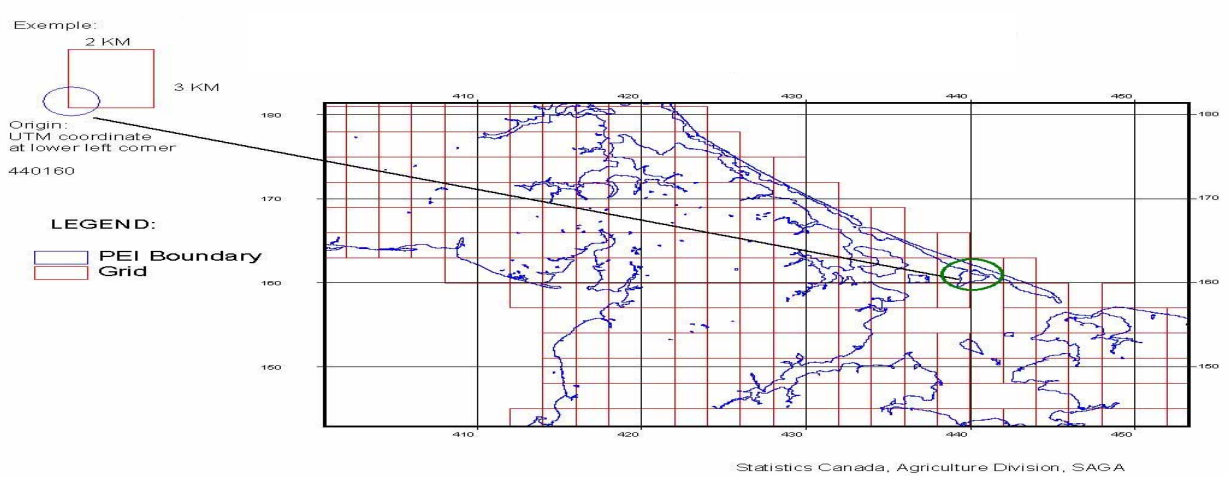

Figure 1: Prince Edward Island – Grid Coverage - UTMID Description

3.1 Frame Creation

The geographic frame consisted of the system of 1,217 cells covering the entire island of PEI. For each cell, information on major crops (potatoes, hay, grain) from previous occasions of the PLCCP was available in GIS format. In addition, each grid cell was assigned a unique identifier, referred to as the UTMID (Universal Transverse Mercator Identifier) that was composed of 6 numbers: the first three were from the *x* (horizontal) UTM co-ordinate, and the last three were from the *y* (vertical) UTM co-ordinate. The origin of the cell is it's lower left corner. UTMID values increase from West to East and South to North.

For example, suppose a cell has a UTM *x* value of 440 000 and a *y* value of 160 000. The UTMID for the cell will be 440 160. The first three numbers of the *x* co-ordinate (440) and the last three will come from the *y* co-ordinate (160). Figure 1 illustrates this.

The first step of the sample design involved dividing the province into three strata, loosely based on the concentration of agricultural activity. The concentration of activity was derived from the geographic frame as an average of the area of potatoes, grains, hay/pasture and other crops between 1998 and 2000. This delineation was undertaken to isolate the area of high concentration of activity so as to ensure an adequate representation of the entire island when selecting our sample. Figure 2 is a map of Prince Edward Island, showing the delineation of strata boundaries based on concentration of agricultural activity.

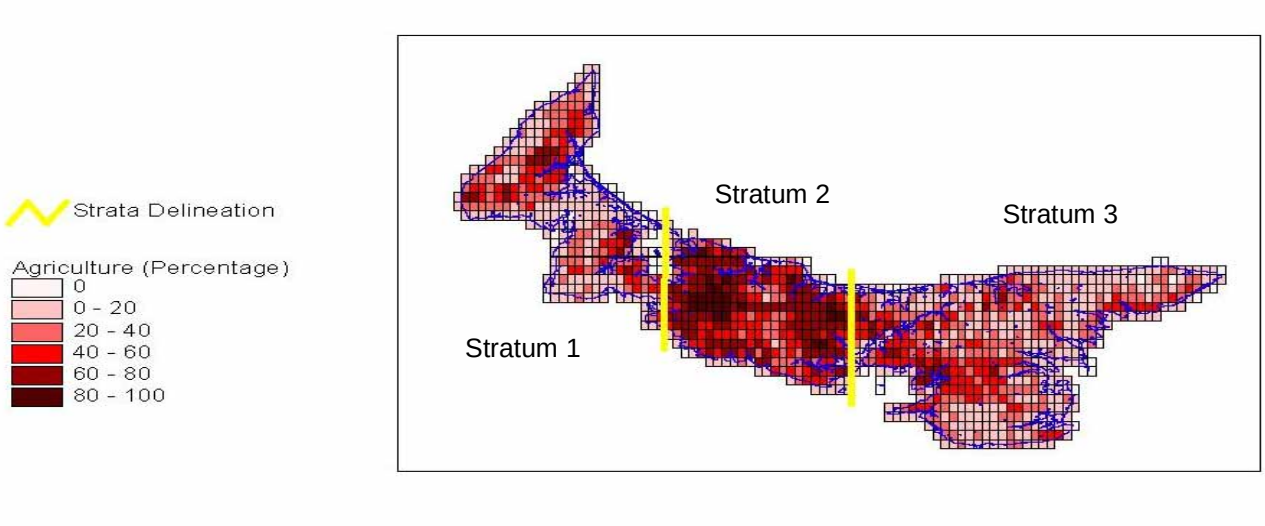

Figure 2: Strata Delineation based on Agricultural Activity (1998-2000)

3.2 Sample Size Determination and Allocation

Several limitations were placed on the sample design due to budgetary constraints. The first of these limitations was the sample size. The budget allowed for a sample of 75 cells only. Would it be possible to obtain reliable estimates with such a small sample? The second limitation, also due to budgetary constraints, was the desire to maximize the overlap between the current sample and that of the 2001 PLCCP. A maximum overlap was desired due to the time and effort involved in preparation of the cells for aerial surveillance. Could this be done without compromising the efficiency of our estimates? The final limitation was related to the sample selection method. As described in Section 2, the PLCCP selected a simple random selection of cells and included some cells in which there was no agricultural activity. This was inefficient from both a sampling point of view and a cost effectiveness perspective, as some of the project funding was spent to survey expected non-agricultural land (i.e. forests).

It was obvious that a more complex survey design would be necessary if we wanted to improve our chances of producing a reliable estimate. As potato area was being measured, it was necessary to find a selection method that would increase the chances of sampling cells that contained potatoes. Probability Proportional to Size (PPS) sampling is commonly used when there exists auxiliary data that is correlated with the variable of interest to increase the efficiency of the estimators. The more closely correlated the auxiliary variable is to the variable of interest, the more efficient the estimate. For more on PPS sampling, see Cochran (1977, pg 250). Because of the decision to use PPS sampling, maximising the overlap with the past sample was not possible. For this project, the sampling unit consisted of a 2 km by 3 km cell, or plot of land. As mentioned previously, historical data on the area of crops and forestlands were available for every cell from previous PLCCP work. Note that the data was obtained by means of the automated crop recognition process and not direct observation or surveying. The historical data that seemed the most relevant to our survey were investigated. These were:

- Average potato acreage (1998, 1999, 2000)
- Average agricultural acreage (1998, 1999, 2000)
- Agricultural acreage for the year 2000

In order to decide which auxiliary variable would perform the best for the PAEP, two comparison studies were performed.

First, a sample size determination study was performed in order to compare the effectiveness of these three variables as possible measures of size. For each of the possible auxiliary variables, the sample size necessary to achieve a desired level of precision, as expressed by the coefficient of variation (CV), was calculated using potato area from 2000. These results are presented in Table 1. Of note, when using PPS sampling, all cells must have a probability of selection greater than zero. That is, the measure of size variable must be greater than zero. Table 1 includes the number of cells where the auxiliary variable is greater than zero. In Table 1, the variable names are defined as follows; AVPOT is the average potato acreage over the years 1998, 1998 and 2000; AVAGR is the average agricultural acreage over the years 1998, 1999 and 2000; and AGR00 is the agricultural acreage for the year 2000.

Table 1. Sample Size Determination Study

Population Size	Auxiliary Variable	Desired CV	Sample Size
1168	AVPOT	5%	70
		6%	49
		7%	36
		8%	28
1168	AVAGR	5%	187
		6%	130
		7%	96
		8%	73
1076	AGR00	5%	197
		6%	137
		7%	101
		8%	77

From Table 1 it is clear that the most efficient auxiliary variable is the average potato acreage over 1998-2000. With a sample of 70 cells, the expected CV is only 5%. For comparison, a sample of 197 cells would be necessary to produce a similarly precise estimate if the area of all agricultural activity during the year 2000 was used. However, it was questioned whether the use of this auxiliary variable would be wise given the anticipated fluctuation in potato production.

Also, one has to note that at the time of the estimation of the CV, the potato area for all farms was available. However, the true estimate of the CV would be based on a sample of cells and would not be as efficient. Thus, the actual observed CV would likely turn out to be larger than the one presented in Table 1. The final decision on which of the three variables to use as an auxiliary variable was made after further evaluation.

In addition to the sample size determination study, a simulation study was completed, whereby, potato estimates for the year 2000 were produced using each of the auxiliary variables and compared to the known total (i.e., sum of Potato 2000 on the historical database). The first step was to allocate the sample of 75 cells to the three strata proportionally to the sum of the auxiliary variable. That is, the sample sizes within each of the 3 strata were given as

$$n_h = round\left(n \times {X_h}/{X} \right)$$

where n_h is the sample size for stratum h, n is the total sample size (75), X_h is the sum of the auxiliary variable in stratum h and X is the sum of the auxiliary variable across the three strata. Characteristics of these samples are given in Table 2.

Table 2. Simulation Study of Sample Distribution

Auxiliary variable	Stratum	n_h	X_h	X
AVPOT	1	20	27,891	105,415
	2	33	45,736	
	3	22	31,788	
AVAGR	1	15	109,515	532,435
	2	32	226,257	
	3	28	196,663	
AGR00	1	14	103,353	551,584
	2	32	237,404	
	3	29	210,827	

The second step involved selecting n_h cells within each stratum using PPS systematic sampling without replacement. Weighting and estimation were performed and estimates were produced. Table 3 presents the potato area estimates for the year 2000 along with the associated levels of precision based on the samples using the three different auxiliary variables.

Table 3. Simulation Study Results

Auxiliary variable	Estimate	CV(%)
AVPOT	111,129	4.273
AVAGR	102,980	7.602
AGR00	110,056	8.718

The total potato area for the year 2000 as obtained from our historical database was 105,617. As can be seen from the results in Table 3, all three auxiliary variables produced comparable area estimates. Interestingly, the CV for the average potato area variable was noticeably lower than the other two, however, this variable is most likely to be severely affected by the anticipated changes in potato acreage for 2001. Thus, it was deemed preferable to go with the more conservative choice of using the average agricultural activity over 1998, 1999 and 2000 as the auxiliary variable for estimating the 2001 potato area. While this choice could possibly yield results that would not be as efficient as if average potato area was used, average agricultural activity would likely be more robust to changes in potato area during the last few years.

3.3 Sample Selection and Preparation

As previously noted, when using PPS sampling, all cells on the frame must have a probability of selection greater than zero, that is, the auxiliary variable AVAGR must have a value greater than zero. Sample allocation of the 75 cells to each of the three strata was done proportionally to total AVAGR and then PPS sampling was used to draw the sample within each stratum.

Once the sampled cell UTMIDs were chosen, large-scale black and white aerial photographs were downloaded from the Prince Edward Island provincial government website. Hard-copy paper prints for each of the 75 sampled cells were made, and the cell boundaries were manually drawn on the photos. These photos were then processed in order to integrate the gound data, which would be collected by aerial surveillance (see section 3.4) into the GIS. This processing was done using digital image processing software (e.g., EASI/PACE, PCI Geomatics V6.1). It was also necessary to build an Arcview 3.1 GIS project file, which contained various digital boundary and image files, including roads, hydrology, a 1997 agricultural field boundary file (provided by DAF) and the PLCCP grid overlay (i.e., UTM PEI grid consisting of 1,217 cells).

3.4 Data collection

The ground "truth" (i.e., crop field data) was collected by means of aerial surveillance. For each of the 75 sampled cells, all crops were visually identified from an airplane, and field locations were recorded directly on the aerial photographs. For questionable fields, which were not easily identifiable from the air, ground field checking was undertaken by car to confirm the exact crop type. Figure 3 provides an example of one of the completed aerial photographs.

Figure 3: Aerial Photograph with Crop Field Data

Potatoes (•) Grains (G) Hay (H) Soybean (S)

The processing of the "ground-truthed" photos was done using GIS software. Basically, the digitized aerial photographs were geo-corrected to align precisely with a master PLCCP satellite image mosaic. Using the GIS, all potato fields located within the 75 sampled cells were then digitized (i.e., draw/trace boundaries of potato fields) using existing high-resolution satellite imagery as a backdrop image. The area of each potato field within each sampled cell was then calculated using the GIS and was used in the estimation process to follow.

3.5 Estimation and Results

Once the total potato area for each of the sampled cells was obtained from the GIS system, Statistics Canada's Generalized Estimation System (GES) was used to produce the estimate of total potato area and a corresponding measure of precision. GES was developed at Statistics Canada to allow users to calculate point estimates and corresponding measures of precision that take into account the survey design. While the system can handle PPS sampling, in order to simplify the calculations at the stage of variance estimation it was decided to treat the units as if they were sampled with replacement. The overestimation of the variance caused by this assumption is believed to be negligible. Table 4 presents the total potato area estimate obtained by the PAEP as well as the potato figures estimated from Statistics Canada's 2001 Potato Survey (Statistics Canada, 2001) and the 2001 Canadian Census of Agriculture (Statistics Canada, 2002).

Table 4. Prince Edward Island Potato Area Estimates

Indicator	Estimate	CV
PAEP	103,447	7.76%
Potato Survey	101,756	4.5%
2001 Census	106,888	n/a

As seen from Table 4, the Census figure is well within the sampling variability of the estimates obtained from the PAEP and the Potato Survey. The Potato Survey is a semi-annual survey based on approximately 500

farms in five provinces across Canada. Selected farms are contacted by telephone and potato area information is collected using Computer Assisted Telephone Interviewing (CATI) techniques. For the year 2001, approximately 180 farms in PEI were contacted by the Potato Survey. While the PAEP estimate is closer to the Census figure, it has a CV that is considerably larger than that obtained from the Potato Survey. This can probably be attributed to the fact that the Potato Survey consists of more than twice as many sampling units as the PAEP.

4. Conclusions

The combination of statistical sampling, aerial surveillance and the use of GIS have allowed the calculation of accurate and timely estimates of potato acreage for PEI. By using PPS sampling, cells with agricultural activity in the past can be selected with higher probabilities, thus avoiding costly aerial surveillance of non-agricultural land. As indicated by the results, estimates obtained in this fashion can be very accurate and precise for even a small number of sampled cells. In addition, the growing problem of respondent burden is avoided, as farms do not need to be contacted to obtain acreage information.

The main limitation associated with this survey design is budget because aerial surveillance is obviously quite costly. This method is perhaps better suited for surveys where relatively small areas need to be covered.

REFERENCES

Cochran, W. G. (1977). *Sampling Techniques (Third Edition)*, New York: John Wiley and Sons, Inc.

Statistics Canada (2001). *Canadian Potato Production*. Publication number 22-008-UIB. Statistics Canada.

Statistics Canada (2002). *Farm Data for the 2001 Census of Agriculture.* Publication number 95F0301XIE. Statistics Canada.