# Chapter 5
# Objective Sampling Estimation of Crop Area Based on Remote Sensing Images

**Alfredo José Barreto Luiz**
*Embrapa, Brazil*

**Antonio Roberto Formaggio**
*INPE, Brazil*

**José Carlos Neves Epiphanio**
*INPE, Brazil*

## ABSTRACT

*Having the ability to estimate crop areas is a necessity ever more pressing for all the stakeholders of productive chains. For many scientists involved in agricultural research it is also important to know the location of crops and the area they occupy. With this information, considered together with data on soil, climate, availability of infrastructure for storage and transportation, among others, it is possible, for example, to build scenarios and fit models to attend multiple demands. In this chapter the authors propose a simple method combining the techniques of statistical sampling with the characteristics of images obtained by remote sensing, to construct estimates of acreage in a micro-region scale. The objective to be achieved is to produce an estimate of crop area in a defined territory, with estimated statistical error associated.*

## INTRODUCTION

Area estimation is one of the most obvious applications of remote sensing because, among other reasons, it has a direct economic impact and very early found a user community with clearly defined accuracy specifications, at least for applications to agriculture statistics (Gallego, 2004). In his chapter, Gallego (2004) cited an abundant literature that made use of different ways to use satellite images for land cover area estimation. However, the linkage with statistical sampling techniques and field work to check or substitute image interpretation are not presented in a simple way yet.

DOI: 10.4018/978-1-61692-871-1.ch005

## Context and History

Proper utilization of statistics within any scope depends primarily on how well the nature of employed data is known, and on how clearly the goals are established. Remote sensing offers a rather particular set of data, which almost predetermines the characteristics that must be taken into consideration when choosing the statistical methods to be used in analysis. Application of those data in agriculture, particularly when the aim is to quantify crop areas, will define quite specific goals that should influence the selection of statistical analysis techniques. An alliance between statistics and remote sensing, based on theory and used in proper ways to estimate crop areas, shall result in a step ahead in the efficient use of data from orbital sensors for agriculture purposes. In this direction, this document presents: 1) a method to prepare and use satellite images in agricultural surveys by sampling; 2) the way to calculate estimates over objective data of crop area and their respective variances; 3) a case study consisting in the estimates of soybean area in a municipality, using remote sensing as auxiliary data; and 4) how to use stratification to improve the quality of estimates.

The availability of reliable information about agricultural production is an increasingly fundamental demand in the decision making process. One of the main variables involved in the assessment of agricultural production is the sowed or planted acreage of important crops (Epiphanio et al, 2009). Methods of survey that take into consideration the increasing availability of remote sensing images, and the use of Global Positioning Systems (GPS) and Geographic Information Systems (GIS) may turn out to be the most practical way for a country produce basic data on major farming commodities (FAO, 1996).

The challenge that arises is the establishment of a method which allows associating the technology provided by orbital remote sensing to the procedures used in agricultural statistics surveys.

## Objectives

The purpose of this chapter is to present a simple and reliable method based on the use of remote sensing images and statistical sampling, which allows the quantification of areas occupied by a particular crop in micro-scale regional or municipal level. This method was developed by Luiz (2003).

To achieve this general goal the following specific objectives are established:

a. Provide a guide for the preparation of images from orbital remote sensing in order to allow their use in agricultural sample surveys;
b. Establish a procedure to extract a random sample of image's pixels from a set of pixels spatially delimited;
c. Present formulas for the calculation of estimated planted area by micro-region or municipality, as well as its variance from data obtained by sampling.

# THEORETICAL FUNDAMENTALS

## Agriculture Estimations and Relationships with Remote Sensing

Despite the decline in the proportional importance of agriculture in global economy, there is a growing need to monitor the agricultural complex (Ryerson et al., 1997). This interest is justified when we consider absolute figures are significant and reach US$ 171 billion annually in the U.S. alone (The World Factbook, 2009).

According to Pino (1999), more countries need to have information and agricultural forecasts that are effective and allow fast perception of change. For Ray et al. (1999), estimates of production in the micro-regional level are essential for management decisions related to the farming-sector economy of any country.

Since the cost of applying traditional techniques over large areas is a limiting factor, the use of remote sensing becomes a practical alternative for some stages of the statistics-obtaining process (Thenkabail et al., 1994). Gurgel et al. (2003) proved that it is possible to use remote sensing data to monitor spatial and temporal dynamics of land covering over large areas.

## Image-Acquiring Systems

Images obtained by remote sensing are indirect ways to capture spatial information. The images obtained by sensors placed in airborne or orbital platforms are stored as matrices. For each picture element (called a "pixel": contraction of "pix" for "picture", and "el" for "element") a value is attributed, which keeps a relationship with the radiance characteristics of an object located on the surface imaged. In digital imaging, each portion of the area imaged by the sensors corresponds to a matrix cell or pixel that must be geographically identified. The pixel is the smallest addressable screen element and each pixel has its own digital address. The address of a pixel corresponds to its coordinates. Pixels are regularly arranged in a 2-dimension grid, and are often represented using dots or squares.

The process of scanning a non-digital image (continuous image) corresponds to a discretization of the scene under observation through a hypothetical mesh overlay and the assignment of an integer value (as represented by digital numbers) to each point of that mesh (in a process called quantization). The spatial resolution of a sensor is defined by the size of its pixels. Sensors commonly used to monitor the characteristics of land surface have spatial resolution ranging from fractions of meters to tens of meters.

The scanned image has a finite number of bits to represent the radiance in each pixel. Radiance is the radiant flux that comes from a source (reflected or emitted) in a given direction, per unit of area. The continuous measurement of the radiance of a scene, represented by digital numbers in the image, is stored in a number of bits per pixel. In general, each object (e.g., a soybean field or a water reservoir) has a typical behavior in relation to the properties of reflectivity or emissivity along the electromagnetic spectrum.

Images obtained by remote sensing are available in the form of "scenes" that correspond to a certain fraction of land surface (Figure 1). Each scene of the Thematic Mapper (TM) sensor installed aboard the Landsat satellite, for example, is provided in the form of 6,000 rows and 6,000 columns (or 36 mega pixels), and corresponds to an approximate area of 180 km x 180 km of land. Thus, each pixel in the TM image corresponds, on the ground, to an area of 30 m x 30 m.

TM images are recorded in seven bands (wavelengths intervals) along the electromagnetic spectrum, numbered from 1 to 7. We can choose any combination of these bands to reproduce the imaged surface into a printed image, by inputting to the digital numbers of each band, one of three colors: red = R, green = G, and blue = B. In general, a composition that represents band 3 by the blue color, band 4 by red, and band 5 by green is named 4R5G3B.

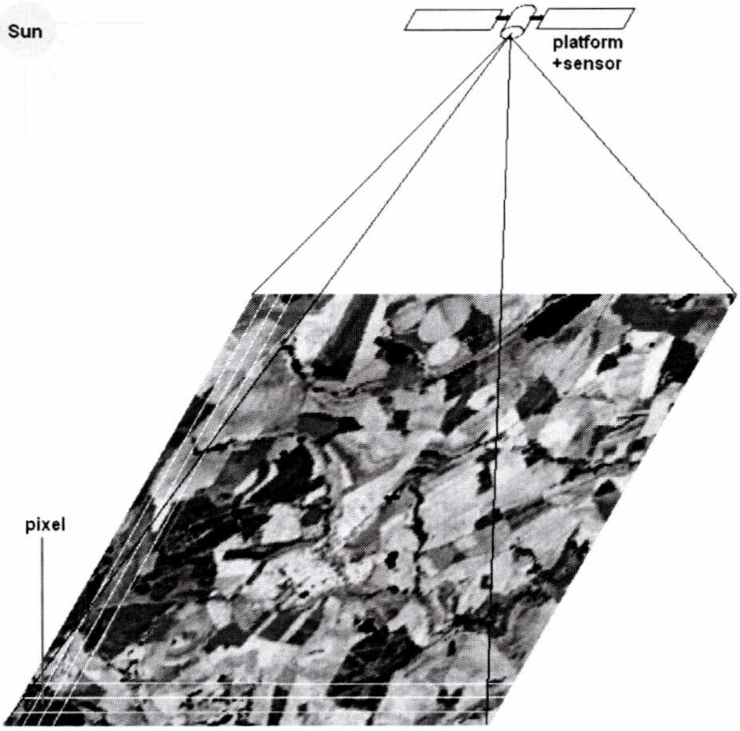
To go deeper into concepts such as swath, geometric resolution, radiometric resolution, spectral resolution, and spectral behavior of objects of land surface, an excellent source of reference is the book of Jensen (2000).

## Crops as Target of Remote Sensing

From the perspective of remote sensing, agricultural activities have certain characteristics that interfere with their identification in digital images obtained by orbital sensors. Unlike natural targets, agricultural targets undergo intense and intentional anthropogenic influence, which helps to give them an aspect of regular geometric figures, with great internal homogeneity. According to Luiz (2003), this is due to some procedures that include:

- plowing and harrowing, which reduce irregularities of the terrain;

*Figure 1. Schematic representation of a remote sensing system, with the source (sun), the platform + sensor, the imaged scene and the pixel*



- liming and fertilization, which subdue differences in soil chemical components;
- mechanized sowing, which provides a geometric character to "plots";
- sowing in a short period of time, and the use of selected seeds, which confer regularity to the green cover across subsequent instants in time; and
- control of pests, diseases and weeds, which warrants regularity and purity for the plant population.

Other factors affecting the homogeneity of agricultural targets are the regional vocation and agricultural calendar, both of which join together to result in a stable and relatively limited number of crops that can occur in a specific image. These features help in the identification of limits of agricultural targets in images obtained by remote sensing satellites (Luiz, 2002).

Moreover, the responses obtained over time, for the same point on the ground occupied by a crop, changes dramatically according to both seasonal factors (climate) and anthropogenic factors (planting, harvesting, soil preparation, etc.).

Agriculture covers a vast number of activities and involves cultivation of several hundred plant species. However, due to environmental and economic constraints, and to agricultural tradition, in a given county only a few species predominate on the cultivated area and constitute a truly significant group for that region (Luiz & Epiphanio, 2001).

An example of this occurs with soybean and corn in Brazil. Speaking of municipalities, the 50 largest producers of these crops, representing 0.91% of the 5,564 Brazilian municipalities, concentrated 14.11% of the area under maize and 36.45% of the area under soybean in 1999 (Tsunechiro & Freitas, 2001).

The fact that the production of a certain crop is concentrated in a few counties, and in these counties it occupies a great proportion of the territory, can mean on the one hand a greater environmental or economic risk, but on the other hand it facilitates the use of remote sensing and statistical sampling.

## Image Preparation

A complete scene obtained by orbital sensors may include many municipalities. To facilitate image handling during fieldwork, it is recommended to extract from the whole scene only the smallest rectangle that will encompass the entire area of the survey's target region. Several types of GIS software can be used to register an image, and then overlap information layers to the limits of the study area, to create a minimal enclosing rectangle. In Brazil there is the SPRING software (Camara et al., 1996), available for free downloading at http://www.dpi.inpe.br/spring/, where the software's operation manual can also be obtained. Image processing procedures described in this chapter were made in SPRING software using Landsat-TM images.

Over the pre-registered image, official municipality boundaries is digitized, which are usually supplied in vector files or in the form of maps by governmental institutions.

When transferring digital material to printed form, an adequate color composition must be chosen to allow for interpretation of the most interesting targets. The R4G5B3 composition is indicated because although it doesn't presents immediate visual identity between colors of targets in real scene and those in the image, it allows a good discrimination between major kinds of land use found in agricultural areas.

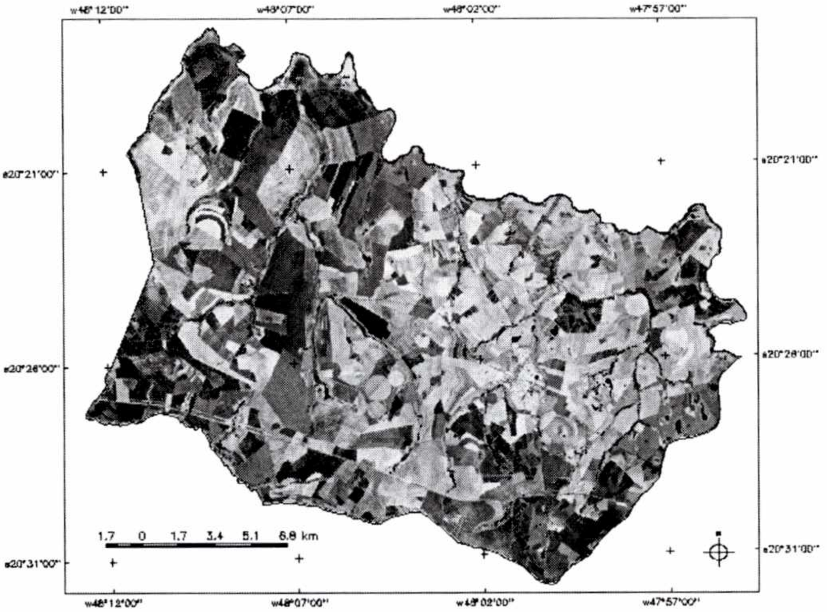
To allow for easy handling of the printed image, while at the same time producing printed material in a scale that renders possible the identification of agricultural plots and the main physiognomic features of the terrain, we suggest the production of image modules in A4-size paper, scale 1:60,000. Prints in this size and scale are fully suitable to a regular clipboard. Each A4 sheet will cover an area of approximately 18,000 ha, which ensures a reasonable synoptic view, without losing the ability to distinguish the main agricultural targets - for example, a squared-shape plot with 9 ha would appear in the print as a square of 5x5 mm.

However, as it is needed several modules in that scale to cover one municipality, it is also necessary to produce a synthesis print in a less detailed scale. This synthesis should present all the modules together in a single A4 sheet, to allow for an overview of the total area of the municipality. Figure 2 shows an image synthesis module of the municipality of Ipuã, State of São Paulo, Brazil, in the summer of 2002.

The SCARTA application, which is part of the SPRING software, can be used to prepare and print A4 modules. The first procedure is the preparation of a printing template, which includes the definition of: paper size (A4), sheet position (landscape), print area (18.7 x 27.0 cm) and scale (1: 60,000). Next, the distribution of modules must be determined to cover the image. The first module should be positioned in the upper left corner of the study area, and the other modules located on the right horizontally and down vertically. It is recommended to leave a small overlap between the modules.

Aiming to not overload the image with texts and lines, and in order to allow overwriting notes during fieldworks, an auxiliary set of modules is printed containing only data such as geographical coordinates and the boundaries of the municipality. These modules are printed on transparent films and then overlaid upon their respective image modules and fixed with adhesive tape, ensuring geographic coincidence through common fiduciary marks that are made for this purpose. Image modules must be printed on high quality paper, specific for images or photographs, using a regular color printer; in other words, it is not necessary to use sophisticated resources such as plotters

*Figure 2. Synthesis module image of the municipality of Ipuã, State of São Paulo, Brazil, Landsat ETM +, WRS 220/74, 4R5G3B, Jan/05/2002*



or photographic techniques in the production of fieldwork material (Luiz et al., 2002).

## Statistical Analysis: Option for Sampling

Obtaining information on planted areas and crop development should be a continuing, periodical, quick, georeferenced, transparent, auditable process with a known accuracy. Such requirements impose certain restrictions on the methods that can be adopted. In Brazil, for example, nationwide coverage in a country of more than 5,500 municipalities and 850,000,000 hectares can hardly be obtained by the census method and, therefore, sampling is more appropriate.

The need for information produced quickly enough to allow intervention still during that harvest cycle requires that time spent gathering, processing and analyzing data must as short as possible, which limits the size of the sample.

For agricultural production seasonality is an important factor, and an effective national program of crop monitoring can only exists through a continuous process of data collection. It requires that sample units be visited and analyzed several times throughout the year, which further restricts the number of sample units that can be observed at a reasonable cost.

Information collected in the sampling units should be georeferenced and stored in a database to allow data be audited at any time, which ensures transparency and the possibility of quality control in the process. This is facilitated by the combined use of remote sensing data in the form of images, global positioning devices and GIS software.

For the survey to produce objective estimates with known precision, it is necessary to define a sampling plan with a probabilistic method, and it is also necessary that data about planted areas be obtained by measurements, not by interviews or questionnaires. The use of images obtained by remote sensing allows the incorporation of the concept of image pixel as the sampling unit, which facilitates achieving an objective measurement (Luiz et al., 2002; Luiz & Gürtler, 2003). The possibility of establishing the accuracy of estimates is ensured by a study of the nature of variables being measured, and by designing an appropriate sampling plan (Epiphanio et al. 2001; Epiphanio et al., 2002; Epiphanio & Luiz, 2001).

## Simple Random Sampling

The Simple Random Sampling (SRS) is a method consisting of selecting a sample of *n* elements from a population total of *N*, such that every possible sample has the same probability of being chosen. To draw a simple random sample of *n* pixels in a digital image, the scene elements that make up the image must be numbered from 1 to *N*, and then using an algorithm, random numbers should be generated to select the *n* elements of the sample.

Because the sampling unit is a pixel that has a known and constant area across the whole region under investigation, the proportion in which pixels with the target crop occur in the sample allows, through expansion, an estimation of the proportion of occurrence of that crop in the municipality. It is therefore necessary and sufficient to measure just one category variable per sampling unit. In other words, in an ordinary case of just one crop ($x_1$), a given element will fall under one of only two categories: *U* if it belongs to a segment occupied by crop $x_1$; or $\bar{U}$ if it belongs to a segment not occupied by crop $x_1$. In order to facilitate exposition of the method, the following notation is adopted:

A number of *U*-class elements in the population (*a* in the sample)

P $= \dfrac{A}{N}$ proportion of *U*-class elements in the population ($p = \dfrac{a}{n}$ in the sample)

Q $= 1 - P$ proportion of $\bar{U}$-class elements in the population ($q = 1 - p$ in the sample)

f $= \dfrac{n}{N}$ sample fraction (or $\dfrac{1}{f} = \dfrac{N}{n}$ = expansion factor)

Considering that the variable is categorical, the process described to obtain a sample, according to Johnson & Kotz (1969), is the classic situation that fits naturally to a discrete distribution of the hypergeometric type. When *N* is big enough, this distribution is approximated to a binomial (Zwillinger & Kokoska, 2000). In this case, according to Cochran (1977), estimation of *P* is given directly by $\hat{P} = p$, while the estimation of *A* is obtained through the application of the expansion factor, namely, $\hat{A} = a \times (\frac{1}{f})$.

Admitting that for each element in the population there is a variable $y_i$ that will take value 1 if that element belongs to class *U*, or value 0 if that element belongs to class $\bar{U}$, it is clear that the total for the population (*Y*) is obtained through Equation 1:

$$Y = \sum_{i=1}^{N} y_i = A \tag{1}$$

and that the averages for population and sample can be calculated through Equations 2 and 3 respectively:

$$\bar{Y} = \frac{\sum_{i=1}^{N} y_i}{N} = \frac{A}{N} = P \tag{2}$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{a}{n} = p \tag{3}$$

Consequently, the problem of obtaining $\hat{A}$ and $\hat{P}$ may be reduced to a simple estimation of the total and the average of a population where every

$y_i$ is either 1 or 0. It is then possible to express population variance ($S^2$) and sample variance ($s^2$) as a function of $P$ and $p$, as follows:

$$S^2 = \frac{N}{N-1} PQ \qquad (4)$$

$$s^2 = \frac{n}{(n-1)} pq \qquad (5)$$

Under such conditions it is possible to affirm that the sample proportion $p = \frac{a}{n}$ is a non-biased estimation of the population proportion $P = \frac{A}{N}$. Also, in practice, in cases where the sample fraction does not exceed 5% ($f$ < 0,05), it is possible to consider that the variance of sample average (that is, $p$ variance) can be estimated in a non-biased way through Equation 6:

$$\hat{s}^2_p = \frac{pq}{(n-1)} \qquad (6)$$

The estimated total of $U$-class elements in the population is represented as $\hat{A} = Np$, and a non-biased estimation of its variance can be obtained through Equation 7:

$$\hat{s}^2_{Np} = \frac{N^2}{(n-1)} pq \qquad (7)$$

With this we can have a good indication of the sample size that is adequate to identify events (agricultural crops, in the present case) with different probabilities of occurrence in the population. In other words, starting from the area of a given municipality and the dimensions of a scene element, we have the $N$ value, and then based on information from previous years, obtained in other surveys, we obtain a value for $p$. It will allow the construction of a table or graph representing the relation between sample size ($n$) and variance of the estimation ($s^2Np$) or its derivatives, such as the coefficient of variation $CV$% that can be calculated with the following expression:

$$CV_{Np} = \frac{100}{p} \times \sqrt{\frac{pq}{n-1}} \qquad (8)$$

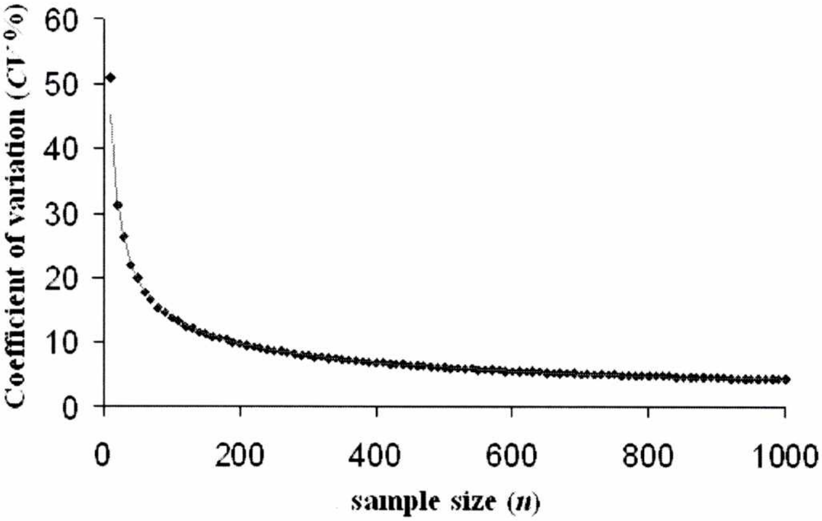
The graph in Figure 3 shows an example where values are the area of the municipality of Ipuã (467,058,750 m$^2$), the area of a scene element corresponding to 25 m × 25 m (625 m$^2$), and the value estimated by IBGE/PAM for the area of harvested soybean in that municipality in 2001, which was 162,000,000 m$^2$ (IBGE, 2003b). In this case, then, $N$ = 747,294 and $P$ = 0.34685144 ≅ 0.35.

## Building a Random Sample

Building a random sample is not a trivial task, and in fact the difficulty to obtain it in agriculture surveys has been a challenge for some decades now, to statisticians involved with the subject. When we consider the problem in a traditional way, basic sampling units consist in harvest plots, farms, units of production or some other type of segment, but always their limits are defined from the perspective of the predominant socioeconomic activity, that is to say, farming. It poses a serious problem for the definition of random samples, because if randomizing is to be warranted, it is necessary to guarantee that all components of a population will have a non-zero and known chance to be part of the sample. Thus, it is necessary to know all elements of a finite population, or the distribution of probabilities in the case of infinite populations.

The case of estimating a planted area characteristically belongs to the first groups, that of finite populations. Therefore, if the basic sampling unit chosen is a farm, for example, in order to

*Figure 3. Curve representing the relation between sample size (n) and the expected coefficient of variation (CV %), given a probability p = 0.35 of occurrence of a crop in the area*



proceed to random sampling it is necessary to have a complete list of all farms existing in the considered population. Among several impediments to such a list, some are: a) the sum of farm areas is not equal to the population's total area, due to measuring errors or mistakes in declaration; b) estate/property registries, depending on when they've been made, are to be found in different bases – public notaries, counties, districts, municipalities, states or territories that may have been created before or after the property was registered; c) plain inexistence of any registry; d) vacant areas; e) areas under litigation – with two or more registries or owners; f) fast modifications in number, size or ownership of properties, due to owner's death, property sale, separation or dissolution of a partnership.

An alternative to bypass such obstacles is the use of sampling panels by areas. Although in theory the panels may contemplate elements with artificially-defined boundaries (Luiz, 2002), this was not the norm until recently, and segments were typically defined by physical boundaries of the terrain (Collares et al., 1993).

Although the advantages that remote sensing images could bring to agriculture sample surveys have been known for a long time, it was only towards the end of the 1990's that utilization of the pixel as a basic sampling unit began to be seen as possible in practice, for agricultural surveys (Luiz et al., 2002). It happened due to the availability of precise GPS equipment at low costs, as well as the uprising of friendly GIS programs – that sometimes are even free of charge, such as SPRING. Starting from that time, the joint utilization of remote sensing images, GIS and GPS devices has been increasing also in agricultural applications (Wilkinson, 1996; Pradhan, 2001; Gao, 2002).

Thus, today it is feasible to use remote sensing images which naturally provide elements with artificial boundaries – the pixels – to obtain random samples for the estimation of the planted area. It renders much easier both the sampling design and the statistical analysis of data collected. Nonetheless, it is necessary to guarantee that the practical procedures to obtain the sample will really result in random process, a fundamental condition for the application of all the remaining statistical concepts to be used in calculation of estimations.

In the past, the use of random sampling per points in statistical surveys to estimate a planted area was nearly impossible, due to the difficulty of point location and the inexistence of an enumerable list of such points. The availability of all three technologies: remote sensing images, GPS equipment, and geographic information systems (GIS) has changed the scenario. The benefit of the use of images when building sampling panels is that the image produces an imaginary division of land surface into scene elements (pixels) of equal dimensions that cover a land area completely and without overlapping. If the area of a municipality is the region of interest, there is a finite number N of pixels that represent in the image the real surface area, where N = (municipality area)/(pixel area). Therefore a numerable list of components of the population is obtained, from which random samples may be drawn. The difficulty of identifying on the actual terrain sample elements with imaginary limits (not real limits such as rivers, woods etc), as is the case with the pixel, is bypassed thanks to the precision and easy operation of GPS equipments that are currently available at low costs. Finally, to integrate benefits from both the use of images and the use of GPS devices, there are the GIS that facilitate the handling of georeferenced information (Luiz et al., 2002).

## Hypotheses

Considering that each digital image obtained by orbital sensors is represented by a $n \times m$ matrix where $n$ is the number of columns and $m$ is the number of lines, and knowing that each cell of the matrix is an element of scene (pixel) that can have its spatial position represented by plane coordinates $X$ and $Y$, it is possible to affirm that:

1. if, through a random number-generating process, $t$ ordered pairs $(x_i, y_j)$ are produced such that determine the positions of $t$ pixels in the image, those pixels shall constitute a Random Sample (RS) that is representative of the image.
2. if, after overlapping an irregular polygon (such as for example the boundaries of a municipality) upon the image, only the $t'$ pixels ($t' \leq k$) are conserved that the centers of which are located within the polygon, the new group of pixels shall continue to constitute a RS that is representative of the region of the image corresponding to the municipality area.

## Demonstration

If $X$ is a random variable (r.v.) with univariate discrete uniform distribution, it means that it may adopt $n$ different values $\{x_1, x_2, ..., x_n\}$ with equal probability. Among the properties of such a distribution, the probability density function (pdf) for $X$ is:

$$f_{(x)} = \begin{cases} \dfrac{1}{n} \quad \forall x_i \quad if \;\; x_i \in \left\{x_1, x_2, \dots, x_n\right\} \\ 0 \qquad\qquad otherwise \end{cases} \tag{9}$$

Let $Y$ be another r.v., independent and identically distributed (i.i.d.) in relation to $X$, that can adopt $m$ different values $\{y_1, y_2, ..., y_m\}$. Then, if $Z$ is defined by the ordered pair $(X, Y)$ in such a way that $z_k = (x_i, y_j)$, where $1 \leq i \leq n$, $1 \leq j \leq m$, and $1 \leq k \leq (n \times m)$, and $i, j, k, n, m$ are integer positive numbers; then we can affirm that $Z$ has a discrete uniform distribution with the following pdf:

$$f_{(y)} = \begin{cases} \dfrac{1}{b-a} \qquad se \quad a < y < b \\ 0 \qquad caso\ contrário\ \left(c.c.\right) \end{cases} \tag{10}$$

This is true if and only if $X$ and $Y$ are i.i.d.

Now, if $W$ is a r.v. that can adopt $t'$ different values, and $t' < n{\times}m$, so that $w_r = z_k$, where $1 \leq r \leq n{\times}m$, $t'$ and $r$ being integer positive numbers; then it follows that $W$ also has a discrete uniform distribution with the following pdf:

$$f_{(y)} = \begin{cases} \dfrac{1}{b-a} & se \quad a < y < b \\ \\ 0 & caso\ contrário\ (c.c.) \end{cases} \qquad (11)$$

In such case, to answer hypothesis a), if the $t$ ordered pairs taken from the population have been produced with any electronic spreadsheet where the cells of two columns, each with $t$ elements, have been filled with integer positive numbers through a random number-generating procedure based on a univariate uniform distribution – in one column numbers vary from 1 to $n$ (number of columns of the image), and in the other column numbers vary from 1 to $m$ (number of lines in the image) – when making the equivalence between those numbers and the pointers of lines and columns in a GIS, the $t$ pixels have been selected that shall constitute a RS representing the image.

In relation to the hypothesis b) exactly the same argumentation is valid, and it is applied twice. In other words, when municipality boundaries are overlapped over the image, a new population is created that is constituted by a subset of the pixels in the original population, in which all elements continue to have identical probabilities of being selected to make the sample; so in parallel, among the pixels selected to make a sample of the original population, the subset of pixels located within the municipality boundaries will still constitute a RS.

It is thus demonstrated that a sample obtained by selecting a set of pixels in an image, and then later taking the subset of pixels contained within the limits of a municipality, is a simple random sample.

## Randomization of Points in the Municipality

After defining a sampling method aided by a remote sensing image, and having selected an area to be studied, it is now necessary to apply the method and obtain estimations originated by the sampling.

The first step is to draw pixels to make the sample. In order to illustrate the method, a sample with size 200 has been drawn.

Considering the irregular characteristics of municipal boundaries, in order to distribute sample points randomly across the municipality, a first draw is made for the enclosing rectangle upon the image stored in SPRING, promoting the equivalence between point and pixel, and then discarding the points that have fallen outside municipal boundaries.

By working upon a matrix representation of the image, the position of each pixel has been defined by the binomial line and column. The name $x$ was given to numbers in the columns, and the name $y$ to line numbers in an image (note that $x$ and $y$ may be multiplied by the pixel dimension, thus transforming their unit in meters, or they may be converted to coordinates of latitude and longitude). The enclosing rectangle will be formed by $r$ lines and $s$ columns, and will be absolutely defined by the ordered pairs of just two scene elements: one in the bottom left corner ($x_1$, $y_1$) and another in the upper right corner ($x_2$, $y_2$). Observe that r = $(|x_2 - x_1| + 1)$, and s = $(|y_2 - y_1| + 1)$, and that the total number of pixels in the rectangle equals the product of $r$ and $s$, which will be called $T$. In order to allow a random selection of the $m$ candidate sample points among all $T$ pixels, it is necessary to generate $m$ ordered pairs ($x_i$, $y_j$). Because the generator used returns two random $z$ numbers $\{0 \leq z < 1\}$ uniformly distributed, it is enough to apply $m$ times the transformations: $x_i = (z_1 \times r)$ and $y_j = (z_2 \times s)$.

These points are considered to be candidates. Since the rectangle is not entirely occupied by the municipality area, a fraction of the $m$ elements will be located outside the limits of the region of interest, and will not be sampled. Only $n$ drawn elements the locations of which are inside the defined limits shall be considered sample points. Because the selection of points is randomized, it is expected that on average the relation between $n$ and $m$ will be the same as exists between the area of the municipality ($A$) and the area of the enclosing rectangle ($R$). From that relation it will be possible to determine the best estimation of $m$ that will produce a certain value of $n$, which is calculated as follows: $m = \{1 + \text{INT}\ [n \times (R/A)]\}$, where INT means the integer part of the expression within brackets.

Considering that the area of the municipality of Ipuã is 46,837.4 ha (IBGE, 2003a), and the area of the enclosing rectangle is 86,265 ha; in order to be able to select a sample of size $n_1 = 200$ it is necessary to draw a total of: $m_2 = \{1 + \text{INT}[200 \times (86{,}265/46{,}837.4)]\} = 369$. That is, if 369 points are generated within the enclosing rectangle, the mathematical expectation is that 200 of them will be located within the municipality boundaries. In order to increase the probability of having the desired number of points inside the area of interest, and because the cost incurred to generate random numbers is computationally very low, we have generated twice that number of points, that is, 738 points to secure 200.

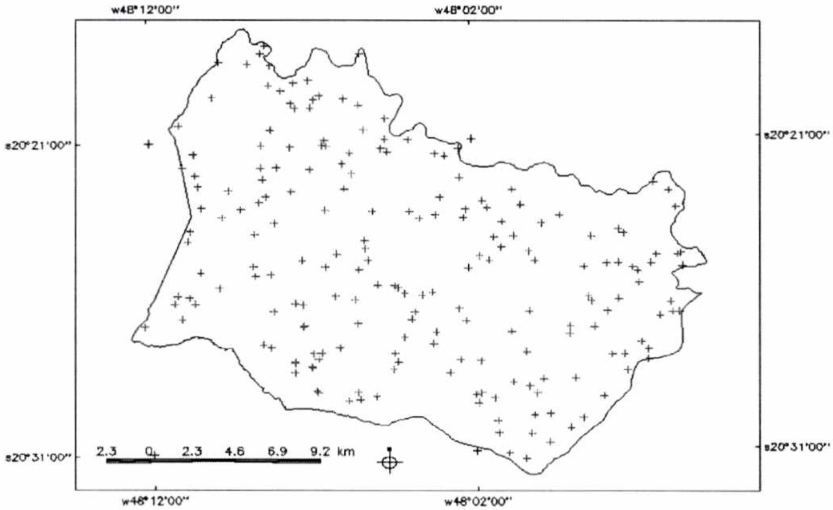
In order to transform the chosen ordered pairs $(x_i, y_j)$ in values compatible with the plane coordinates $(X_i, Y_j)$ of the enclosing rectangle defined in SPRING, in the case of an image with spatial resolution of 25 meters, the following transformation is used: $X_i = X_1 + 12{,}5 + [\text{INT}\ (x_i \times 25)]$ and $Y_j = Y_1 + 12{,}5 + [\text{INT}\ (y_\text{j} \times 25)]$, where the pair $(X_1, Y_1)$ defines the position of the southeastern corner of the enclosing rectangle, in the Southern Hemisphere. The procedure guarantees that each pixel in the image will have the same chance to be picked to make up the sample, and that the selection is made considering the center of the pixel.

The coordinates of points randomly generated as described are inputted into the SPRING GIS software through importation as points' files in the format ASCII-SPRING, which will generate an Information Layer (IL) for each set of points. To eliminate points that are not within the polygon corresponding to Ipuã municipality's boundaries, the IL are cut out using a mask with the boundaries. As the resulting IL contain a number of points that exceeds the desired number, IL are exported to be opened in an Excel spreadsheet. Points are then classified in crescent order (relative to their position in generation process), and from the lists a group is selected with the desired number of points for each file, that is, the 200 first points. The remaining points are deleted from the files that are then imported again into SPRING, finishing this stage. Figure 4 exemplifies the distribution of sample points across Ipuã's territory, for a sample of size 200.

After samples are generated, with IL of the selected points, image modules are printed for fieldwork, with the location of points printed only on plastic films that may be overlay the image, to enable localization of points in the field. Points' coordinates may also be exported directly to the GPS memory, which will further facilitate their localization. It is important to remember that when you can count on a relatively recent image of the area under study, several selected points will be easy to classify without the need of an actual visit, because they will present easily identifiable characteristics such as water bodies, urban areas, forests etc. After filling in the tables with cadastral data and exporting them to an Excel spreadsheet, data are worked on to originate estimations. The procedure consists in a simple count of points in the class of interest.

The area occupied by the class of interest in the municipality can be estimated based on the 200 point sample. You will calculate the area dividing the frequency observed for that class by the size of the sample, and then multiplying the result by the municipality's total area. The ratio between

*Figure 4. Example of distribution of the 200 points selected to make up a random sample in Ipuã, State of São Paulo, Brazil*



the observed frequency and the size of sample is the proportion $p$ of occurrence of that class in the sample, which is a non-biased estimator for the proportion $P$ of that class in the population. Therefore, it suffices to multiply the $p$ of each class by the total area of the municipality $AM$ to obtain the estimated area for that class, that is:

$$\hat{A} = p \times AM \qquad (12)$$

In our example, among the 200 points selected and visited at the field, 68 were classified as soybean. Applying the indicated formulae, we have that the crop of interest occupies a proportion $p = 68/200$ of the territory in the municipality of Ipuã. When we multiply $p$ by the area of the municipality, we estimate as 15,925 ha the area planted with soybean in Ipuã in the Summer 2001/2002.

The procedure to calculate $CV$ is also very simple, and it suffices to apply Equation 8. Getting back to the example of Ipuã in the Summer 2001/2002, with $p$ = 68/200 and $q$ = 1 - $p$, we found that the $CV$ for the estimation of soybean area in the municipality is 9.85%.

## Stratification

Stratification is a regular sampling technique, and there are several reasons for that. Among those reasons we may highlight that when there is the desire to know the degree of data precision for certain subdivisions of the population, it is appropriate to treat each of these divisions as a separate "population", or in other words, each subpopulation is a stratum. Another strong reason would be the existence of administrative advantage; for example, the agency conducting data collection may have different offices, each of them with its own area of scope defined according to criteria that are often non-related to a specific survey. Another reason could be detected in the case of a sampling based on farmers working with a given crop, when one has only a list of large farms, which will be put in a separate stratum, while smaller unlisted properties will be divided across other strata where some type of area sampling could be applied (Cochran, 1977; FAO, 1996; FAO, 1998).

Separation in strata could also produce some gain in precision when estimating characteristics a heterogeneous population into subpopulations that will each be internally homogeneous. This is suggested by the word "stratum", with its implication of a division by layers. If each stratum is homogeneous, so that measures will show little variation from one unit to another, a precise estimation of the average in each stratum can be obtained from a small sample in that stratum. Estimations from all strata can be combined into a single precise estimation of the whole population (Cochran, 1977).

The stratification can be used in two or more stages in a program of statistical surveys of crops. In a first moment, stratification will be made only based on information obtained from the remote sensing image, with the purpose of dividing the municipal territory into areas internally homogeneous, and different among them in relation to the predominant land use. In another way, a group of municipalities – belonging to a State, a Region or even the entire country – could be stratified in order to obtain homogeneous subpopulations as to the variable of interest, which in the present work is the planted area, but such stratification will not be focused in this work.

## Stratification within a Municipality

Stratification inside a municipality is essentially of the spatial type, and it is particularly adequate when the usage of soil varies greatly from one region to another within municipal boundaries. Several reasons may lead to such a situation Farming usage of soil is strongly affected by climate, type of soil and topography; for example, flooded rice in Brazilian conditions of crop handling occurs almost exclusively in hydromorphic plains; while large crops with intense mechanization occupy primarily terrains with little declivity and fertile soils; and perennial and semiperennial cultures may occur in areas of irregular terrain, as the selection of their localization is conditioned mainly by the climate. Another set of factors affecting soil usage are anthropomorphic factors, mainly the agrarian structure and the distribution of infrastructure for transportation and storage of farming products, which makes it relatively easy to understand why within the same municipality a region with difficult access – due to privation of roads ofbridges – shall be neglected in comparison with another region that is better served in those aspects; crops appropriate for regions divided into small rural properties shall be different from those in regions occupied by great landed estates. It is important to observe that although the agrarian structures and the availability of infrastructure are highly correlated with environmental variables such as soil, geography and climate, this is not always valid.

Another factor determining the type of land use in each region in the municipality is the legal or juridical component. The appointment of laws on environment preservation may transform an entire area into a reserve and abolish every form of agricultural use; or else, laws establishing limits for a given farming activity may force a spatial rearrangement of production, as in the case of laws concerning the maximum declivity susceptible to mechanization, or the minimum distance from urban centers where sugar cane can be burnt in preparation for harvest. An excellent source of study to understand the impact that such types of restriction can have on the spatial distribution of crops is the work by Giannotti (2001). The paper deals with the impacts of the environmental law prohibiting the burning of sugar cane in the region of Piracicaba (State of São Paulo), where such prohibition fosters mechanization of crops, and environment variables condition the susceptibility of land to mechanization, which is modifying the spatial formation of the crops.

From the perspective of an estimation of the planted area, some or all the above mentioned factors may act to promote or discourage the spatial concentration of a given crop inside a municipality. It may also vary from one species to another, and there can be in the same territory an example of homogeneous distribution and another example of high spatial concentration. Then, the option for stratification will depend on the study's objective, as the method chosen should favor the estimation of the crop of greatest interest. In other words, depending on the crop of interest, one can choose to use or not use stratification in one municipality; and further still, even when there is an option for stratification, the optimal allocation of sample points across all strata shall also vary according to the crop of interest. For example, if a municipality is occupied mainly with sugar cane and soybean crops; and if spatial distribution of sugar cane crops is strongly influenced by their distance from refineries, thus causing a higher concentration of that culture in one region of the municipality, it could be the case of making a stratification in two areas, one with a higher occurrence of sugar cane, and the other with a higher incidence of soybean. Such stratification will result in a single segmentation of the municipality's territory, regardless whether the main interest is soybean or sugar-cane. But the optimal number of points in each stratum – keeping constant the total number of points – shall vary in accordance with the main interest of the study. In regard to the definition of limits for each stratum, considering the scale of that municipality and admitting that the person in charge of the survey is familiar with the territory under study, the manual method of segmentation is indicated, based on visual interpretation of the remote sensing image (King, 2002).

After the population (area of the municipality split into $N$ pixels) is divided into $l$ strata, and maintaining the notation used in the simple random sampling (SRS), we observe that in the stratified random sampling (StRS), strata have each $N_h$ units, where suffix $h$ defines the $h$-esimal stratum of the population; they do not overlap and constitute together the entire population, that is:

$$N = N_1 + N_2 + \ldots + N_h + \ldots + N_l = \sum_{h=1}^{l} N_h \qquad (13)$$

In order to characterize a StRS it is necessary to take a SRS in an independent way in each stratum. Sample sizes inside strata are named $n_1$, $n_2$,..., $n_h$,..., $n_z$, respectively. The total size of the sample is still called $n$, and it is given by the sum of all $z$ values of $n_h$. To follow with the previous notation, we'll adopt the following:

$A_h$ = number of $U$-class elements in stratum $h$ ($a_h$ in the sample)<br>
$P_h = A_h/N_h$ = proportion of $U$-class elements in stratum $h$ ($p_h = a_h/n_h$ in the sample)<br>
$Q_h$ = 1-$P_h$ = proportion of $\bar{U}$-class elements in stratum $h$ ($q_h$ = 1- $p_h$ in the sample)<br>
$f_h = nh/N_h$ = sample fraction in stratum $h$ ($1/f_h = N_h/n_h$ = expansion factor)<br>
$W_h = N_h/N$ = weight of stratum $h$

In stratified sampling, the number of sample points $n_h$ (or sample size) in each stratum is chosen by the person in charge of sampling. Several criteria may be adopted to determine that amount, and usually an objective is to reduce the estimated variance, keeping a fixed or reduced cost (Cochran, 1977). In a simple case, admitting a linear function of cost, and given the restriction meant by the total number of sample points, one will look for the best allocation of points in each stratum so as to obtain the least variance.

According to Cochran (1977), when StRS is applied to proportions, as is the present case, and one wants to estimate the proportion of units in the population that fall under $U$ class, the ideal stratification is achieved when one puts in the first stratum all units of such class, and in other strata all other classes. But as this usually is not possible in practice, especially when units and strata are spatially distributed, strata should be built in such a way that the proportion in $U$ class will vary as much as possible from one stratum to another. For this situation, the estimation of proportion $P$ that is adequate for StRS is named $\hat{P}_{st}$ and is given by Equation 14:

a heterogeneous population into subpopulations that will each be internally homogeneous. This is suggested by the word "stratum", with its implication of a division by layers. If each stratum is homogeneous, so that measures will show little variation from one unit to another, a precise estimation of the average in each stratum can be obtained from a small sample in that stratum. Estimations from all strata can be combined into a single precise estimation of the whole population (Cochran, 1977).

The stratification can be used in two or more stages in a program of statistical surveys of crops. In a first moment, stratification will be made only based on information obtained from the remote sensing image, with the purpose of dividing the municipal territory into areas internally homogeneous, and different among them in relation to the predominant land use. In another way, a group of municipalities – belonging to a State, a Region or even the entire country – could be stratified in order to obtain homogeneous subpopulations as to the variable of interest, which in the present work is the planted area, but such stratification will not be focused in this work.

## Stratification within a Municipality

Stratification inside a municipality is essentially of the spatial type, and it is particularly adequate when the usage of soil varies greatly from one region to another within municipal boundaries. Several reasons may lead to such a situation Farming usage of soil is strongly affected by climate, type of soil and topography; for example, flooded rice in Brazilian conditions of crop handling occurs almost exclusively in hydromorphic plains; while large crops with intense mechanization occupy primarily terrains with little declivity and fertile soils; and perennial and semiperennial cultures may occur in areas of irregular terrain, as the selection of their localization is conditioned mainly by the climate. Another set of factors affecting soil usage are anthropomorphic factors, mainly the agrarian structure and the distribution of infrastructure for transportation and storage of farming products, which makes it relatively easy to understand why within the same municipality a region with difficult access – due to privation of roads of bridges – shall be neglected in comparison with another region that is better served in those aspects; crops appropriate for regions divided into small rural properties shall be different from those in regions occupied by great landed estates. It is important to observe that although the agrarian structures and the availability of infrastructure are highly correlated with environmental variables such as soil, geography and climate, this is not always valid.

Another factor determining the type of land use in each region in the municipality is the legal or juridical component. The appointment of laws on environment preservation may transform an entire area into a reserve and abolish every form of agricultural use; or else, laws establishing limits for a given farming activity may force a spatial rearrangement of production, as in the case of laws concerning the maximum declivity susceptible to mechanization, or the minimum distance from urban centers where sugar cane can be burnt in preparation for harvest. An excellent source of study to understand the impact that such types of restriction can have on the spatial distribution of crops is the work by Giannotti (2001). The paper deals with the impacts of the environmental law prohibiting the burning of sugar cane in the region of Piracicaba (State of São Paulo), where such prohibition fosters mechanization of crops, and environment variables condition the susceptibility of land to mechanization, which is modifying the spatial formation of the crops.

From the perspective of an estimation of the planted area, some or all the above mentioned factors may act to promote or discourage the spatial concentration of a given crop inside a municipality. It may also vary from one species to another, and there can be in the same territory an example of homogeneous distribution and another example of high spatial concentration. Then, the option for stratification will depend on the study's objective, as the method chosen should favor the estimation of the crop of greatest interest. In other words, depending on the crop of interest, one can choose to use or not use stratification in one municipality; and further still, even when there is an option for stratification, the optimal allocation of sample points across all strata shall also vary according to the crop of interest. For example, if a municipality is occupied mainly with sugar cane and soybean crops; and if spatial distribution of sugar cane crops is strongly influenced by their distance from refineries, thus causing a higher concentration of that culture in one region of the municipality, it could be the case of making a stratification in two areas, one with a higher occurrence of sugar cane, and the other with a higher incidence of soybean. Such stratification will result in a single segmentation of the municipality's territory, regardless whether the main interest is soybean or sugar-cane. But the optimal number of points in each stratum – keeping constant the total number of points – shall vary in accordance with the main interest of the study. In regard to the definition of limits for each stratum, considering the scale of that municipality and admitting that the person in charge of the survey is familiar with the territory under study, the manual method of segmentation is indicated, based on visual interpretation of the remote sensing image (King, 2002).

After the population (area of the municipality split into $N$ pixels) is divided into $l$ strata, and maintaining the notation used in the simple random sampling (SRS), we observe that in the stratified random sampling (StRS), strata have each $N_h$ units, where suffix $h$ defines the $h$-esimal stratum of the population; they do not overlap and constitute together the entire population, that is:

$$N = N_1 + N_2 + \ldots + N_h + \ldots + N_l = \sum_{h=1}^{l} N_h \qquad (13)$$

In order to characterize a StRS it is necessary to take a SRS in an independent way in each stratum. Sample sizes inside strata are named $n_1$, $n_2$,..., $n_h$,..., $n_z$, respectively. The total size of the sample is still called $n$, and it is given by the sum of all $z$ values of $n_h$. To follow with the previous notation, we'll adopt the following:

$A_h$ = number of $U$-class elements in stratum $h$ ($a_h$ in the sample)
$P_h = A_h/N_h$ = proportion of $U$-class elements in stratum $h$ ($p_h = a_h/n_h$ in the sample)
$Q_h$ = 1-$P_h$ = proportion of $\bar{U}$-class elements in stratum $h$ ($q_h$ = 1- $p_h$ in the sample)
$f_h = nh/N_h$ = sample fraction in stratum $h$ ($1/f_h = N_h/n_h$ = expansion factor)
$W_h = N_h/N$ = weight of stratum $h$

In stratified sampling, the number of sample points $n_h$ (or sample size) in each stratum is chosen by the person in charge of sampling. Several criteria may be adopted to determine that amount, and usually an objective is to reduce the estimated variance, keeping a fixed or reduced cost (Cochran, 1977). In a simple case, admitting a linear function of cost, and given the restriction meant by the total number of sample points, one will look for the best allocation of points in each stratum so as to obtain the least variance.

According to Cochran (1977), when StRS is applied to proportions, as is the present case, and one wants to estimate the proportion of units in the population that fall under $U$ class, the ideal stratification is achieved when one puts in the first stratum all units of such class, and in other strata all other classes. But as this usually is not possible in practice, especially when units and strata are spatially distributed, strata should be built in such a way that the proportion in $U$ class will vary as much as possible from one stratum to another. For this situation, the estimation of proportion $P$ that is adequate for StRS is named $\hat{P}_{st}$ and is given by Equation 14:

$$\hat{P}_{st} = \sum_{h=1}^{l} \frac{N_h \times p_h}{N} = \sum_{h=1}^{l} \left( W_h \times p_h \right) \qquad (14)$$

And estimation of variance for $\hat{P}_{st}$ can be calculated with Equation 15:

$$\hat{s}^2_{\hat{P}_{st}} = \sum_{h=1}^{l} W_h^2 \frac{p_h \times q_h}{n_h} \qquad (15)$$

## Applying Stratification in the Municipality

Aiming to evaluate the effect of stratification within a municipality upon the precision of estimations, Ipuã's territory has been divided into two strata. Stratum I was defined so that it covered approximately two thirds of the total area (27,858 ha), being located upon the center and east of the municipality, always north of the road SP 345 that links the municipality of Guaíra to Highway SP 330. Stratum II, in consequence, was located in the western side of the municipality, adjacent to Guaíra, and also included the entire area south of road SP 345, which represented one third of the total area (18,979 ha).

Stratification was made manually, based on visual interpretation of the image, facilitated by familiarity with the field that was acquired along several campaigns conducted in that municipality. We have also tried to use natural limits such as water streams, and anthropomorphic limits such as the road, to draw strata borderlines. The main visual characteristic influencing in the definition of both strata was the presence of sugar cane and soybean crops. Stratum I is predominantly occupied with soybean crops, while Stratum II has a higher frequency of sugar cane.

The same procedure for drawing points has been performed in this stage, and the number of points drawn for the municipality was 200. Two different distributions have been defined among both strata, one considering a greater interest in the estimation of soybean area, and the other admitting a greater interest in sugar cane. In both cases, 86 points were allocated in one stratum, and 114 in the other. These numbers resulted from the calculation of the expected variance that is given by Equation 5. Based on values for soybean and sugar cane in strata I and II, obtained from a mapping effort (Luiz, 2003), we found the value for the proportion $p$ of each crop in both strata. The optimal allocation of sample points is that produces the lowest sum of variances for a given culture in both strata. In the case under study, just as a coincidence, the optimal combination for soybean occurred with 114 points in stratum I and 86 in stratum II; for sugar cane it was just the opposite, that is, 86 points in stratum I and 114 in stratum II.

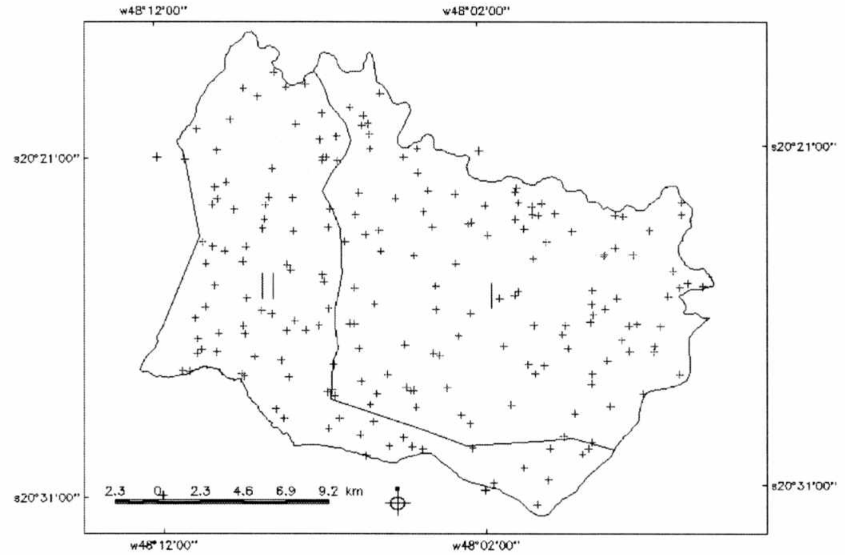
One example of drawing points with 114 points in stratum I and 86 in stratum II is shown in Figure 5. The distribution, in terms of precision, favors the estimation of the area planted with soybean, since the area with greater amount of points is precisely the one with most of the soybean planted in the municipality.

In the other case, opposed to the previous one, 114 points were distributed in stratum II, occupied primarily with sugar cane, and 86 points were distributed in stratum I, that has a higher frequency of soybean crops. This distribution favors the estimation of the area planted with sugar cane, because it intensifies the sampling in the area where cane is predominant.

The procedure employed to calculate the estimation of the planted area in both strata and the municipality, has been at first the same as for SRS. After obtaining the frequency of each class in each stratum, the area of culture estimated in the stratum was obtained simply through multiplication of the proportion found by the area of the stratum, *AH*, that is:

$$\hat{A}_h = p_h \times AH \qquad (16)$$

*Figure 5. Example of distribution of 200 points chosen to make up a stratified random sample, with 114 points in stratum I and 86 points in stratum II, in Ipuã, State of São Paulo, Brazil*



The procedure employed to calculate the coefficient of variation in each stratum, $CV_h$, has also been the same as in the case of SRS, that is:

$$CV_h = 100 \times \sqrt{\frac{q_h}{n_h p_h}} \qquad (17)$$

The procedure has suffered a small change to calculate the estimation for the entire municipality, considering both strata. But even in this case it sufficed to add the areas estimated for each stratum, that is:

$$\hat{A} = \hat{A}_1 + \hat{A}_2 \qquad (18)$$

But to calculate the $CV$ it was necessary to consider first the weight of each stratum, which has been calculated merely dividing the area of each stratum by the municipality area, that is:

$$W_h = \frac{AH}{AM} \qquad (19)$$

Therefore, the $CV$ of the estimation for the municipality that was obtained through StRS, was calculated using the following equation:

$$CV_{\hat{A}} = 100 \times \frac{\sqrt{W_1^2 \frac{p_1 q_1}{n_1} + W_2^2 \frac{p_2 q_2}{n_2}}}{\hat{p}_{est}} \qquad (20)$$

The municipality of Ipuã is relatively homogeneous regarding to its rural occupation, but in order to demonstrate the utility of a spatial stratification within a municipality, it has been split into two strata, one more dedicated to soybean, and another more dedicated to sugar cane. Then, just to verify the effect that a change of objective has upon the precision of estimates, two samplings were performed. In the first case, called case A, admitting that the interest of the survey relates to soybean estimates, a greater amount of points (114) were put in stratum I, where soybean has a higher presence, leaving the remaining 86 points (as the total sample has always been 200) for stratum II, where sugar cane is predominant. In the second case, called case B, the exact opposite actions were taken: admitting a greater interest in the estimation of sugar cane areas, 114 points were put in stratum II, and only 86 points into stratum I. In both cases, estimations of area and *CV* for both sugar cane and soybean were calculated as shown in Table 1.

We can observe in Table 1 that the *CV* for the estimation of soybean is lowest in Case A (8.1%), where sampling has been intentionally in favor of soybean culture, while the lowest *CV* for sugar cane (9.1%) occurred in Case B, for the same reason. It also worth observing that, although the *CV* of the estimation of soybean for the municipality has been lowest in Case A, inside stratum II there has been an increase of *CV* value, as compared with Case B. This is exactly how stratification works: if the interest is directed to a given variable, the method will try to make less mistake where that variable is more frequent (or where it occupies a larger area, in this case), even if it represents to err more where the variable is rare. The important thing is that in the final estimation there is a gain in precision, provided that the objective has been well defined, and stratification has been correctly done.

## FUTURE RESEARCH DIRECTIONS

The estimate of crop areas by objective sampling is of great potential, since it allows predictions provided with quantification of the error. It can be performed at different scales, from the smallest, as in the municipalities, to the larger, such as states, regions or countries.

In addition to its implementation, a verification of costs is recommended for the various phases involved, especially those related to data collection in the field. The costs involved are for equipment (e.g. GPS), daily payment and training for fieldwork data collectors, fuel, etc. One of the comparisons to be made is between these costs for the sampling method proposed here versus the costs for other methods. Also, it would be interesting to perform quantitative comparisons between the accuracy of the crop estimates obtained by this method and by other methods, and between the cost effectiveness of each.

New studies that explore the possibility of reducing the number of points to be visited in the fieldworks, through the use of current remote sensing images, so that points that fall into non-agricultural class can be discarded, should be encouraged. Efforts to improve the method should be directed towards the optimization and reduction of costs and laboriousness, while seeking to increase its optimization, speed and usefulness.

Another goal that can be sought in the near future is the increasing automation of the method, starting with the randomization of sampling points directly into the municipal area or region of interest with no need to go through the phase of the enclosing rectangle. The search and registration

*Table 1. Planted area and CV for sugar cane and soybean, in Ipuã, estimated through StRS*

| | | | Sugar cane | | Soybean | |
|---|---|---|---|---|---|---|
| | Stratum | *nh* | Area (ha) | *CV* (%) | Area (ha) | *CV* (%) |
| Case A | I | 114 | 3,912 | 23.2 | 15,158 | 8.6 |
| | II | 86 | 9,483 | 10.8 | 3,749 | 21.7 |
| | | | 13,395 | 10.4 | 18,907 | 8.1 |
| Case B | I | 86 | 4,861 | 23.5 | 11,991 | 12.4 |
| | II | 114 | 10,980 | 8.0 | 3,494 | 19.7 |
| | | | 15,841 | 9.1 | 15,485 | 10.6 |

of current images of the study area also may be automated so that at any time one wishes to carry out the fieldwork, the most recent image is available.

## CONCLUSION

The agricultural information agile, accurate, inexpensive and safe, with quantified statistical errors, constitutes material of increasing value in a high dynamic scenario within the agricultural national and international markets.

In this chapter we provided the theoretical and practical basis for a method of estimating crop areas, using statistical objective sampling and remote sensing data. The method can also be applied to other targets that occupy large portions of land surface in a delimited area, such as forest, water bodies, urban areas and so on.

The sampling based on the structure of digital images enables a quick survey with known accuracy of the area covered with predominant crops within a municipality or region.

The remote sensing images, for having a comprehensive coverage and are repeatedly obtained, may help to estimate the area occupied by crops in municipal or regional scale.

The stratification approach is an excellent contribution in reducing the sample size or increasing the precision of estimates, provided that there is a strong knowledge base on which we can determine the boundaries of the strata.

The use of objective sampling methods, by region, leads to increased knowledge on the part of local agricultural statistics staff, which will surely bring efficiencies and effectiveness to the process.

Databases, GPS and GIS are tools which increasing efficiency is essential to the objective sampling method. Year after year, the application of the methodology should allow the continued enrichment of the database, in addition to a growth in experience of technical people both in the fieldwork as in the office (those who prepare the material for the field work, receive and organize the data collected and generate final statistics), which will make the method more efficient.

The audits, which can improve the quality of data coming from the field, are of fundamental importance, since the quality of field data are crucial to the quality and accuracy of objective estimates sample.

## REFERENCES

Camara, G., Souza, R. C. M., Freitas, U. M., & Garrido, J. (1996). SPRING: Integrating remote sensing and GIS by object-oriented data modeling. *Computers & Graphics*, *20*(3), 395–403. doi:10.1016/0097-8493(96)00008-8

Cochran, W. G. (1977). *Sampling techniques*. (3ed.). New York: John Wiley & Sons.

Collares, J. E. R., Lauria, C. A., & Carrilho, M. M. (1993). Pesquisa de previsão e acompanhamento de safras baseada em painéis de amostras de áreas. In Instituto Nacional de Pesquisas Espaciais (Ed.), *VII Simpósio Brasileiro de Sensoriamento Remoto,* (Vol. 4, pp. 450-453). São José dos Campos, Brazil: INPE.

Epiphanio, J. C. N., Barros Neto, O. O., Luiz, A. J. B., & Formaggio, A. R. (2001). Sistema de amostragem em imagem como base para estimativa de áreas agrícolas no município de Ipuã-SP. In Instituto Nacional de Pesquisas Espaciais (Ed.), *X Simpósio Brasileiro de Sensoriamento Remoto*, (pp. 59-66). São José dos Campos, Brazil: INPE.

Epiphanio, J. C. N., Luiz, A. J. B., & Formaggio, A. R. (2002). Estimativa de áreas agrícolas municipais, utilizando sistema de amostragem simples sobre imagens de satélite. *Bragantia*, *61*(2), 187–197. doi:10.1590/S0006-87052002000200012

Epiphanio, R. D. V., Formaggio, A. R., Rudorff, B. F. T., Maeda, E. E., & Luiz, A. J. B. (2010). Estimating soybean crop areas using spectral-temporal surfaces derived from MODIS images in Mato Grosso, Brazil. *Pesquisa Agropecuaria Brasileira*, *45*(1), 72–80.

Food and Agriculture Organization of the United Nations – FAO. (1996). *Multiple frame agricultural surveys: volume 1 current survey based on area and list sampling methods*. Rome: FAO. (FAO Statistical Development Series, 7).

Food and Agriculture Organization of the United Nations – FAO. (1999). *Multiple frame agricultural surveys: volume 2 agricultural survey programmes based on area frame or dual frame (area and list) sample design*. Rome: FAO. (FAO Statistical Development Series, 10).

Gallego, F. J. (2004). Remote sensing and land cover area estimation. *International Journal of Remote Sensing*, *25*(15), 3019–3047. doi:10.1080/01431160310001619607

Gao, J. (2002). Integration of GPS with remote sensing and GIS: reality and prospect. *Photogrammetric Engineering and Remote Sensing*, *68*(5), 447–453.

Giannotti, M. A. (2001). *Geotecnologias na análise de impactos sócios-ambientais: o caso da queima da cana-de-açúcar na região de Piracicaba. Unpublished master's theses*. São José dos Campos, Brazil: Instituto Nacional de Pesquisas Espaciais.

Gurgel, H. C., Ferreira, N. J., & Luiz, A. J. B. (2003) Estudo da variabilidade do NDVI sobre o Brasil, utilizando-se a análise de agrupamentos. *Revista brasileira de engenharia agrícola e ambiental,* *7*(1), 85-90. doi: 10.1590/S1415-43662003000100014.

Instituto Brasileiro de Geografia e Estatística – IBGE. (2003a). *CIDADES*. Retrieved May 20, 2003, from http://www.ibge.gov.br/cidadesat

Instituto Brasileiro de Geografia e Estatística – IBGE. (2003b). *SIDRA*. Retrieved April 22, 2003, from http://www.sidra.ibge.gov.br

Jensen, J. R. (2000). *Remote sensing of the environment: an earth resource perspective*. Upper Saddle River: Prentice Hall.

Johnson, N. L., & Kotz, S. (1969). *Discrete Distributions*. New York: John Wiley & Sons.

King, R. B. (2002). Land cover mapping principles: a return to interpretation fundamentals. *International Journal of Remote Sensing*, *23*(18), 3525–3545. doi:10.1080/01431160110109606

Luiz, A. J. B. (2002). *Sensoriamento remoto agrícola*. São José dos Campos, Brazil: INPE.

Luiz, A. J. B. (2003). *Estatísticas agrícolas por amostragem auxiliadas pelo sensoriamento remoto*. Unpublished doctoral dissertation, Instituto Nacional de Pesquisas Espaciais, São José dos Campos, Brazil.

Luiz, A. J. B., & Epiphanio, J. C. N. (2001). Amostragem por pontos em imagens de sensoriamento remoto para estimativa de área plantada por município. In Instituto Nacional de Pesquisas Espaciais (Ed.), *X Simpósio Brasileiro de Sensoriamento Remoto*, (pp. 111-118). São José dos Campos, Brazil: INPE.

Luiz, A. J. B., & Gürtler, S. (2003). Aleatorização de pontos no território de um município, usando o SPRING, para a estimativa de área agrícola por amostragem. In Sociedade Brasileira de Informática Aplicada à Agropecuária e Agroindústria (Ed.), *IV Congresso Brasileiro da Sociedade Brasileira de Informática Aplicada à Agropecuária e Agroindústria SBI-Agro*. (pp. 9-12). Lavras, Brazil: UFLa.

Luiz, A. J. B., Oliveira, J. C., Epiphanio, J. C. N., & Formaggio, A. R. (2002). Auxílio das imagens de satélite aos levantamentos por amostragem em agricultura. *Agricultura em São Paulo*, *49*(1), 41–54.

Pino, F. A. (1999). Estatísticas agrícolas para o século XXI. *Agricultura em São Paulo*, *46*(2), 71–105.

Pradhan, S. (2001). Crop area estimation using GIS, remote sensing and area frame sampling. *JAG: International Journal of Applied Earth Observation and Geoinformation*, *3*(1), 86–92. doi:10.1016/S0303-2434(01)85025-X

Ray, S. S., & Pokharna, S. S., & Ajai. (1999). Cotton yield estimation using agrometeorological model and satellite-derived spectral profile. *International Journal of Remote Sensing*, *20*(14), 2693–2702. doi:10.1080/014311699211741

Ryerson, R. A., Curran, P. J., & Stephens, P. R. (1997). Agriculture. In Philipson, W. R. (Ed.), *Manual of Photographic Interpretation* (pp. 285–397). Bethesda, MD: ASPRS.

*The World Factbook 2009*. (2009). Washington, DC: Central Intelligence Agency. Retrieved February 9, 2010, from https://www.cia.gov/library/publications/the-world-factbook/index.html

Thenkabail, P. S., Ward, A. D., Lyon, J. G., & Merry, C. J. (1994). Thematic mapper vegetation indices for determining soybean and corn growth parameters. *Photogrammetric Engineering and Remote Sensing*, *60*(4), 437–442.

Tsunechiro, A., & de Freitas, B. B. (2001). Os cinqüenta municípios brasileiros maiores produtores de milho e soja. *Informações Econômicas*, *31*(7), 53–58.

Wilkinson, G. G. (1996). A review of current issues in the integration of GIS and remote sensing data. *International Journal of Geographical Information Systems*, *10*(1), 85–101.

Zwillinger, D., & Kokoska, S. (2000). *CRC Standard probability and statistics tables and formulae*. Boca Raton, FL: Chapman & Hall/CRC.

## ADDITIONAL READING

Benedetti, R., Bee, M., Espa, G., & Piersimoni, F. (Eds.). (2010). *Agricultural survey methods*. New York: John Wiley & Sons. doi:10.1002/9780470665480

Bouma, J., Varallyay, G., & Batjes, N. H. (1998). Principal land use changes anticipated in Europe. *Agriculture Ecosystems & Environment*, *67*(2-3), 103–119. doi:10.1016/S0167-8809(97)00109-6

Brink, A. B., & Eva, H. D. (2009). Monitoring 25 years of land cover change dynamics in Africa: A sample based remote sensing approach. *Applied Geography (Sevenoaks, England)*, *29*(4), 501–512. doi:10.1016/j.apgeog.2008.10.004

Carfagna, E., & Gallego, F. J. (2005). Using remote sensing for agricultural statistics. *International Statistical Review*, *73*(3), 389–404. doi:10.1111/j.1751-5823.2005.tb00155.x

Carfagna, E., & Marzialetti, J. (2009). Sequential design in quality control and validation of land cover databases. *Applied Stochastic Models in Business and Industry*, *25*(2), 195–205. doi:10.1002/asmb.742

Eva, H. D., Carboni, S., Achard, F., Stach, N., Durieux, L., Faure, J.-F., & Mollicone, D. (2010). Monitoring forest areas from continental to territorial levels using a sample of medium spatial resolution satellite imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, *65*(2), 191–197. doi:10.1016/j.isprsjprs.2009.10.008

Fjellstad, W. J., & Dramstad, W. E. (1999). Patterns of change in two contrasting Norwegian agricultural landscapes. *Landscape and Urban Planning*, *45*(4), 177–191. doi:10.1016/S0169-2046(99)00055-9

Gonzalez-Alonso, F., Cuevas, J. M., Arbiol, R., & Baulies, X. (1997). Remote sensing and agricultural statistics: crop area estimation in north-eastern Spain through diachronic Landsat TM and ground sample data. *International Journal of Remote Sensing*, *18*(2), 467–470. doi:10.1080/014311697219213

Green, K. (1992). Spatial imagery and GIS: integrated data for natural resource management. *Journal of Forestry*, *90*(11), 32–36.

Griffith, D. A., & Amrhein, C. G. (1997). *Multivariate statistical analysis for geographers*. Upper Saddle River: Prentice Hall.

Hietala-Koivu, R. (1999). Agricultural landscape change: a case study in Yläne, southwest Finland. *Landscape and Urban Planning*, *46*(1-3), 103–108. doi:10.1016/S0169-2046(99)00051-1

Ippoliti-Ramilo, G. A., Epiphanio, J. C. N., & Shimabukuro, Y. E. (2003). Landsat-5 thematic mapper data for pre-planting crop area evaluation in tropical countries. *International Journal of Remote Sensing*, *24*(7), 1521–1534. doi:10.1080/01431160010007105

Klersy, R. (1992). The work and role of the Commission of the European Communities. *International Journal of Remote Sensing*, *13*(6-7), 1035–1058. doi:10.1080/01431169208904177

Lillesand, T. M., & Kiefer, R. W. (1994). *Remote sensing and image interpretation*. 3ed. New York: John Wiley & Sons. 750p.

MacDonald, R. B., & Hall, F. G. (1980, May). Global crop forecasting. *Science*, *208*(4445), 670–679. doi:10.1126/science.208.4445.670

Meyer-Roux, J., & King, C. (1992). Agriculture and Forestry. *International Journal of Remote Sensing*, *13*(6-7), 1329–1341. doi:10.1080/01431169208904194

Monitoring Agriculture with Remote Sensing Techniques (MARS). (1993). *Crop area estimation of annual crops through area frame sampling based on segments: results obtained in Europe in 1992*. Varese: Ispra. 21p.

Mulders, M. A., De Bruin, S., & Schuiling, B. P. (1992). Structured approach to land cover mapping of the Atlantic zone of Costa Rica using single date TM data. *International Journal of Remote Sensing*, *13*(16), 3017–3033. doi:10.1080/01431169208904099

Ortiz, M. J., Formaggio, A. R., & Epiphanio, J. C. N. (1997). Classification of croplands through integration of remote sensing, GIS, and historical database. *International Journal of Remote Sensing*, *18*(1), 95–105. doi:10.1080/014311697219295

Stehman, S. V. (2005). Comparing estimators of gross change derived from complete coverage mapping versus statistical sampling of remotely sensed data. *Remote Sensing of Environment*, *96*(3-4), 466–474. doi:10.1016/j.rse.2005.04.002

Stehman, S. V. (2009). Model-assisted estimation as a unifying framework for estimating the area of land cover and land-cover change from remote sensing. *Remote Sensing of Environment*, *113*(11), 2455–2462. doi:10.1016/j.rse.2009.07.006

Tsiligirides, T. A. (1998). Remote sensing as a tool for agricultural statistics: a case study of area frame sampling methodology in Hellas. *Computers and Electronics in Agriculture*, *20*(1), 45–77. doi:10.1016/S0168-1699(98)00011-8

Villalobos, A. G. (2006). Sample Survey Statistics Teaching: An almost worldwide problem on teaching Agricultural Survey Methods. In *Proceedings of International Conference on the Teaching of Statistics* (ICOTS 7). Salvador, Brazil: IASE - International Association for Statistical Education. http://www.stat.auckland.ac.nz/~iase/publications/17/4f3_gonz.pdf/.

Warner, T. A., Nellis, M. D., & Foody, G. M. (2009). *The SAGE handbook of remote sensing*. London: SAGE Publications.

Xiao, X., Boles, S., Frolking, S., Salas, W., Moore, B., & Li, C. (2002). Landscape-scale characterization of cropland in China using VEGETATION sensor data and Landsat TM imagery. *International Journal of Remote Sensing*, *23*(18), 3579–3594. doi:10.1080/01431160110106069

## KEY TERMS AND DEFINITIONS

**Class of Interest:** It's the kind of land cover that most interests the sample survey. For example, if an agricultural survey that seeks only to estimate the area planted with soybeans, this would be the class of interest and all other types of land use, including other crops, could be grouped in a class defined as non-soy

**Fieldwork:** The fieldwork may or may not be present in an agricultural survey, and is an activity performed at the site where the studied phenomenon occurs naturally and not in the office or lab. Involves all the steps necessary to collect and record data, features and information related to the phenomenon or object of study, including travel, preparation of material and handling equipment.

**Image Processing Procedures:** Image processing procedures are all the procedures used on the raw data of images obtained by remote sensing in order to present them the best possible way for their analysis and interpretation. Procedures can be divided into preprocessing, image enhancement, classification and information extraction.

**Information Layer (IL):** Information layers are digital files of data that can be represented spatially. Each layer can be seen like a map that brings one characteristic represented in space. The geographic information systems (GIS) allow overlapping information layers to cross characteristics and properties located in geographic space as appropriate to the purpose of each investigation.

**Objective Measurement:** Objective measurement is the act of obtaining a value associated with a characteristic of the object observed through the use of devices or instruments, and that does not depend on who makes the measure. It is the opposite of the subjective measure, which is dependent on the bias of who is measuring.

**Sampling Unit:** Each element of the population that can be identified and selected for the sample

**Simple Random Sample:** A simple random sample occurs when the sampling plan ensures that every element of the population has an equal and known probability of being sampled.

**Spatial Information:** Spatial information is any information about an object that is tied to its location in space and its spatial relationship with other objects of the same population.

**Stratified Sampling:** The stratified sampling uses a priori information to divide the target population into subgroups internally homogeneous, called strata. The strata can be defined based on various factors such as topography, political boundaries, roads, rivers, human characteristics, depending on the context of the problem. Once defined the strata, a sample will be extracted in each stratum. Generally, the objective of this procedure is to estimate the true average or the total for a variable in each stratum. If the stratification was correct (relatively homogeneous strata) the estimated average for the population is more accurate than could be obtained by a simple sampling.