Food and Agriculture Organization
of the United Nations

Technical Report on **Developing More Efficient and Accurate Methods for the Use of Remote Sensing in Agricultural Statistics**

**Publication prepared in the framework of
the Global Strategy to improve Agricultural and Rural Statistics**

September 2014

Technical Report on
Developing More Efficient and Accurate Methods for the Use of Remote Sensing in Agricultural Statistics

Table of contents

Preface

This Technical Paper on **Developing More Efficient and Accurate Methods for the use of Remote Sensing** was prepared in the framework of the *Global Strategy to Improve Agricultural and Rural Statistics*. The Global Strategy is an initiative that was endorsed in 2010 by the United Nations Statistical Commission. It provides a framework and a blueprint to meet current and emerging data requirements and the needs of policy makers and other data users. Its goal is to contribute to achieving greater food security, reduced food price volatility, higher incomes and greater well-being for rural populations, through evidence-based policies. The Global Strategy consists of 3 pillars: (1) establishing a minimum set of core data; (2) integrating agriculture in National Statistical Systems (NSS); and (3) fostering statistical systems' sustainability, through governance and statistical capacity building.

The Action Plan to Implement the Global Strategy includes an important Research programme, to address methodological issues for improving the quality of agricultural and rural statistics. The envisaged outcome of the Research Programme consists in scientifically sound and cost-effective methods that will be used as inputs in preparing practical guidelines for use by country statisticians, training institutions, consultants, etc.

To enable countries and partners to benefit, at an early stage, from results of the Research activities, it was decided to establish a **Technical Reports Series**, to allow widespread dissemination of the technical reports and advanced draft guidelines and handbooks available. This would also enable countries to provide earlier and more feedback on the papers.

The Technical Reports and draft guidelines and handbooks published in this Technical Report Series are prepared by Senior Consultants and Experts, and reviewed by the Scientific Advisory Committee (SAC)1 of the Global Strategy, the Research Coordinator at the Global Office and other independent Senior Experts. For some of the research topics, field tests are organized before final results can be included in the relevant guidelines and handbooks.

This technical report on **Developing More Efficient and Accurate Methods for the Use of Remote Sensing** is the result of a comprehensive literature review on the subject, followed by a gap analysis and, finally, the development of innovative methodological proposals for addressing the various issues arising.

[1] The SAC is composed of 10 well-known Senior Experts in various fields relevant to the Global Strategy's Research Programme. These Experts are appointed for a 2-year term. . The current membership is composed of Vijay Bhatia, Seghir Bouzaffour, Ray Chambers, Jacques Delincé, Cristiano Ferraz, Miguel Galmes, Ben Kiregyera, Sarah Nusser, Fred Vogel, and Anders Walgreen.

The main purpose of this report is to summarize certain topics concerning the use of remotely sensed data for agricultural statistics. In particular, the report seeks to describe the new technologies of remote sensing, and their influence on any improvements that can be made to the methods proposed.

The Report first provides a comprehensive review of the literature on this subject, dividing it into six main topics: *New technologies of remote sensing, Methods for using remote sensing data at the design level, Extension of the regression or calibration estimators, Robustness of the estimators adopted for producing agricultural and rural statistics, Comparison of regression and calibration estimators with small area estimators, Statistical methods for quality assessment of land use/land cover databases.*

The Report also contains a critical analysis of the papers in these six topics, and proposes methodological and operational solutions to fill the gaps and to analyse certain methodological issues identified and examined subsequently to the literature review. The work will constitute an input for several Guidelines, including the Guidelines on developing and maintaining a Master Sampling Frame for integrated agricultural surveys, which are currently under preparation.

The technical papers are updated as required with the results of in-country field tests and with country feedback and experiences.

Acknowledgments

The Technical Paper on Developing More Efficient and Accurate Methods for the Use of Remote Sensing was prepared by Roberto Benedetti (Professor), University “G. d’Annunzio” of Chieti-Pescara, with the assistance of Paolo Postiglione (Professor), University “G. d’Annunzio” of Chieti-Pescara, Domenica Panzera (PhD), University “G. d’Annunzio” of Chieti-Pescara, and Danila Filipponi (PhD, researcher), Italian National Statistical Institute. The Report was prepared with the guidance and supervision of Elisabetta Carfagna, FAO, Javier Galego, JRC, and Naman Keita, FAO.

Valuable inputs and comments were provided at various stages by the SAC members and by Loredana Di Consiglio, ISTAT.

This publication was prepared with the support of the Trust Fund of the Global Strategy, funded by the UK’s Department for International Development (DFID) and the Bill and Melinda Gates Foundation (BMGF).

1

Introduction

Remote sensing is an important tool in the study of natural resources and environment. The possible applications of remotely sensed data are manifold: the identification of potential archaeological sites, the assessment of drought and flood damage, the monitoring and management of land use, the compilation of crop inventories and forecasts, etc. Today, remotely sensed images constitute a basic instrument for monitoring natural resources, the environment, and agriculture. Remote sensing has also become crucial for protecting the global environment, reducing disaster losses, and achieving sustainable development. Furthermore, they provide invaluable information on the state of art of the agricultural sector for both developed and developing countries.

Remote sensing is defined as the technique of deriving information on the features of the earth's surface, and estimating their geo-bio-physical properties using electromagnetic radiation as a medium of interaction (Canada Centre for Remote Sensing, 2003). This acquisition occurs without physical contact with the earth. The process involves registering observations using sensors (i.e. cameras, scanners, radiometer, radar, etc.) mounted on platforms (i.e. aircraft and satellites), which are at a considerable height from the earth's surface, and recording the observations on a suitable medium (i.e. images on photographic films, and videotapes or digital data on magnetic tapes). Then, the data obtained is usually stored and manipulated using computers.

Remote sensing data, also combined with *in situ* observations, are widely used in agriculture and agronomy (Dorigo *et al.*, 2007). They ensure a good spatial coverage of the field under investigation. This is particularly evident for many developing countries, where satellite data represent the only source of information on the weather, the earth's surface, and, of course, agriculture.

Remote sensing has been increasingly considered for its capacity to provide a standardized, faster, and possibly cheaper methodology for agricultural statistics. Several developing and developed countries have remote sensing projects to support the official agricultural statistics programs (see Becker-Reshef *et al.*, 2010).

The wavelengths used in most agricultural remote sensing applications cover only a small region of the electromagnetic spectrum. Wavelengths are measured in micrometers (μm) or nanometers (nm). One μm is equivalent to 1,000 nm. In remote sensing, we deal with ultraviolet (UV) radiation, which has wavelengths between 10 nm to 400 nm, to radar wavelengths. The visible region of the electromagnetic spectrum is from approximately 400 nm to approximately 700 nm. The green color associated with plant vigor has a wavelength around 500 nm. The major part of the electromagnetic spectrum that is used for sensing earth resources consists of the visible/infrared and the microwave range.

The reflectance measured by a sensor can be considered as a proxy variable of some biophysical phenomena, such as: the geographical coordinates (x,y) of an object, the temperature, the colour, and the moisture content of the soil and of the vegetation, among others. These covariates are often called direct variables (Jensen, 2004). On the other hand, it is possible to derive certain hybrid variables, defined through the simultaneous analysis of several biophysical variables. For example, considering the absorption characteristics of a plant, its temperature, and its moisture content, it is possible to determine the stress undergone by a plant, which represents a hybrid variable.

The remote sensing system usually contains a platform, a device for navigation, one or more sensors, and a module for data processing and interpretation. The platform is usually a satellite or an aircraft. The device determines the location of the navigation platform and the land area to be investigated; the interpreter can be human or an automated system that supervises the entire operation and the platform.

The remote sensing systems can be active or passive. Active systems such as radars and lasers emit their own electromagnetic radiation and then, later, analyse the characteristics of signals reflected from the illuminated objects. Therefore, images can be acquired both by day and by night, completely independently of solar illumination; this is particularly important at high latitudes (polar night). The microwaves emitted and received are at a much longer wavelength than optical or infrared waves. Microwaves can therefore easily penetrate clouds, and images of the surface can be acquired irrespective of local weather conditions.

On the other hand, passive systems are based on electromagnetic waves that are not generated by themselves, but from external sources of energy like the sun. Note that the joint analysis of optical and radar data can provide unique information that is not visible in the separate images.

When electromagnetic energy from the sun goes down to the plants, three different possibilities may materialize. Depending on the energy’s wavelength and on the plants’ features, the energy will be reflected, absorbed, or transmitted. Reflected energy bounces off leaves, and is recognized by the human eye as the green colour of plants. Sunlight that is not reflected or absorbed is transmitted through the leaves to the ground. Interactions between reflected, absorbed, and transmitted energy can be detected by remote sensing. The differences in leaf colours, textures, and shapes determine how much energy will be reflected, absorbed or transmitted. The relationship between reflected, absorbed and transmitted energy is used to determine the spectral signatures of individual plants. Spectral signatures are unique to plant species.

However, as well-stated by Carfagna and Gallego (2005), the spectral response and the identification of crops are not in one-to-one correspondence. Indeed, the radiometric response of the same crop in different conditions can vary across the pixels of an image. A more appropriate approach is to consider the spectral response of a crop as a function of the probability distribution of its spectral reflectance. To solve this problem, the literature has explored some methods for the correct identification of the crops. We refer to this group of techniques as "classification methods".

There are two broad classes of classification procedure. The first is referred to as unsupervised classification, while the second is defined as supervised classification. In unsupervised classification, an image is segmented into unknown classes. The researcher's aim is to label these classes at a subsequent stage. Unsupervised classifications seek to group pixels having similar spectral reflective characteristics into distinct clusters. These spectral clusters are then labeled with a certain class name. Supervised classification uses a set of user-defined spectral signatures to classify an image. The spectral signatures are derived from training areas (or sets) that are created by depicting features of interest on an image. The main difference between the two approaches is that in unsupervised classification, the classes need not be defined *a priori*. Richards and Jia (2006) provide further detail on this topic.

In recent decades, satellite remote sensing technology, in combination with *in situ* observations, has become an important factor in the enhancement of the current systems for acquiring and generating agricultural data. To reap the benefits of remotely sensed data, managers, consultants, and technicians must be able to understand and interpret the images.

Remote sensing techniques are widely used in agriculture and agronomy (Dorigo *et al.*, 2007). Indeed, remotely sensed images provide spatial coverage of a field, and can be used as proxies to measure crop and soil attributes (Fitzgerald *et al.*, 2006). In several developing countries, and over much of the oceans, satellite data is the only source of quantitative information on the state of the atmosphere and of the earth's surface, and it is an invaluable source of real-time information on severe weather, which is critical for safety in these areas.

It is necessary to use remote sensing, since the monitoring of agriculture raises special problems, which are not common to other economic sectors (World Bank, 2011). Indeed, agricultural production depends heavily upon seasonal patterns related to the life cycle of crops.

In addition, production varies according to the physical landscape (i.e. soil type), as well as to climatic conditions and the agricultural management practices adopted. Finally, all agricultural variables differ greatly in space and time. For these reasons, agricultural monitoring systems must be capable of operating on a timely basis. Remote sensing can significantly contribute towards addressing these needs, and is highly appropriate for the collection of information over large areas with high revisit frequency.

As mentioned above, remote sensing has been increasingly considered for the development of a standardized, faster, and possibly cheaper methodology for agricultural statistics. Many countries have remote sensing programs in support of the official agricultural statistics programs, including the EU countries, China, India, and some less developed countries in Africa, Southeast Asia, and Latin America. Today, agricultural knowledge is required to address various social requirements. For example, national and international agricultural policies, and global agricultural organizations dealing with food security issues, depend to a great extent on reliable and timely crop production information (Becker-Reshef *et al.*, 2010).

Carfagna and Gallego (2005) provide a first exhaustive description of the different possible uses of remote sensing for agricultural statistics. In particular, remote sensing techniques may represent a suitable tool for particular problems in agricultural survey, such as: the reliability of data, incomplete sample frames and sample sizes, the methods to select units, measurement of areas, non-sampling errors, gaps in geographical coverage, and the unavailability of statistics at a disaggregated level.

Remote sensing can be properly used at the design level. Remotely sensed images provide a synopsis of the area under investigation, and are useful for the construction of the spatial reference frame. Furthermore, classified satellite images can be used as auxiliary variables to improve the precision of ground survey estimates, generally with a regression or a calibration estimator. The remotely sensed information could also represent an auxiliary variable in the process of small area estimation. Finally, remote sensing data have been exploited to estimate the production of crops, using their link with the yield. The most common indicators are based on the Normalized Difference Vegetation Index (NDVI, Benedetti and Rossini 1993, Benedetti et al., 1994) that can be computed through a remotely sensed image. However, as highlighted by Carfagna and Gallego (2005), the link between the NDVI and crop yield is high only for some crops, under certain conditions.

The cases described above are only some of the possible examples of the application of remote sensing to agricultural data. The purpose of this Report is to review the main contributions in the literature on this issue. To pursue this objective, we have identified six main topics that will be discussed in detail in this Report:

1. New technologies of remote sensing
2. Methods for using remote sensing data at the design level;
3. Extension of the regression or calibration estimators;
4. Robustness of the estimators adopted for producing agricultural and rural statistics;
5. Comparison of regression and calibration estimators with small area estimators;
6. Statistical methods for quality assessment of land use/land cover databases;

The following sections of this document will be devoted to the critical analysis of the papers grouped according to the six topics listed above.

This Report also seeks to propose possible methodological and operational solutions to fill the existing gaps in this regard, and to analyse certain methodological issues identified and examined after the literature review.

In particular, we aim to describe the new technologies of remote sensing, and their influence on possible improvements to the methods proposed. Following these lines, Section 2 will describe the impact of the new remote sensing technologies, devoting some attention to free-of-charge images. Section 3 will present some suggestions for the development of *ad hoc* methodologies for employing remote sensed data, during both the design and the estimation phases. Section 4 will outline some extensions of the regression and calibration estimators, to address the particular typology of sampling units as, for example, points. Section 5 will analyse the robustness of the estimators currently adopted for producing agricultural and rural statistics. Section 6 will compare regression and calibration estimators with small area estimators, with particular attention to the latter class of estimators. Section 7 will discuss some issues concerning improved statistical methods for the quality assessment of land use/land cover databases, and of methods for detecting change in land covers. Section 8 will highlight some issues related to the applicability of the proposed methods in developing countries. Finally, Section 9 will provide some conclusions.

This Report follows a six-part scheme. A summary table is included at the end which summarizes the main research issues and their characteristics in terms of applicability, recommendations and gaps to address.

Below, we summarize the topics that require further methodological development, highlighting the gaps encountered and the recommendations submitted. A literature review on the topic and possible solutions to the limitations are presented in each respective section.

A - Methods for using remote sensing data at the design level

1. Sample selection with probability proportional to size, when a size measure is multivariate (see Section 3.1).
   -*πps* designs are known to be not robust for outliers. A single sampling weight that is too high may change the estimates to a remarkable degree.
   -It is necessary to adopt a method to evaluate the inclusion probabilities as a linear combination of a multivariate set of auxiliaries, which should represent advancement with respect to the maximal Brewer selection used by USDA.

2. Optimal stratification with multivariate continuous auxiliary variables (see Section 3.2).
   -The literature is mainly univariate, while X is usually multivariate.
   -An algorithm must be developed to optimally stratify with irregularly shaped strata.

3. Spatially balanced samples (see Section 3.3).
   -The algorithms are very slow; their application to very large populations may be impossible.
   -A flexible selection method, with probability proportional to the within sample distance, must be developed.

4. Models linking survey variables, with some auxiliary variables to design the sample (sample size, sample allocation etc.) (see Section 3.4).
   -As is often the case when a data model is suggested, the model should be adjusted for each application.
   -It is necessary to develop the theory of anticipated moments for "zero-inflated" models (some zeros in the data that alter the estimated parameters), and their estimation and test on remotely sensed data.

B - Extension of the regression or calibration estimators

1. Non-response (see Section 4).
   -How to cope with the prescence of missing values in the auxiliary vector and estimate variance in the prescence of inputed data.

2. Models for space varying coefficients (see Section 4).
   -A model allowing the coefficients to vary as smooth functions of the geographic coordinates should be introduced.

-Methods should be devised for the identification of local stationarity zones, i.e. post strata. These could greatly increase the efficiency of the model-assisted estimators.
-A method is needed to identify the best partition and the estimation and test of models on remotely sensed data.

3. Zero-inflated data (see Section 4).
   -Crops data may present excess zeros. Zero-inflated count models provide a powerful way to model this type of situation.
   -A method for the estimation and test of "zero-inflated" models on remotely sensed data.

C - Robustness of the estimators adopted for producing agricultural and rural statistics

1. Direct Estimation (see Section 5).
   -The sample design, which allows for increasing the sample size in small areas allowing for direct estimates, could be developed (linked to Point A4).

2. Model-based Small Area Estimation (see Section 5).
   -Most of the proposed benchmarking approaches only adjust for the overall bias, irrespective of the bias at the small-area level. Further investigations on the benchmarking procedures could be developed.

D - Comparison of regression and calibration estimators with small area estimators

1. Models for space varying coefficients (see Section 6).
   -Local stationarity zones should be identified.
   -A geographically-weighted regression model, which allows the coefficients to vary across the area of interest, should be formulated.
   -The estimation and the test for partitioning or smoothing algorithms on remotely sensed data should be defined.

2. Models for non-Gaussian data (see Section 6).
   -Non-parametric methods incorporating spatial information into the M-quantile approach (Chambers et al., 2013) should be used.

3. Missing values in the auxiliary variable (see Section 6).
   -Multiple imputation methods should be adopted, to reflect the uncertainty due to the imputed values.

E - Statistical methods for quality assessment of land use/land cover databases

1. Sampling Design – linked to Point A3 (see Section 7).
   -Since reference data might serve different accuracy analyses, more efforts should be directed towards the definition of sampling designs that are amenable to multiple uses, which fulfill both precision and cost/efficiency criteria.

2. Response Design (see Section 7).
   -A protocol for assessing the accuracy of the response design should be developed.

3. Analysis (see Section 7).
   -The accuracy assessment for soft classifications must be investigated further.

2

New technologies of remote sensing

The main aim of an agricultural monitoring system should be to provide timely information on crop production, status, and yield. This information is needed for policy makers, in a standardized and regular manner, at national as well as regional levels. Estimates should be delivered as early as possible during the growing seasons, and updated periodically through the season until harvest. Naturally, stakeholders need these data to present statistically valid precision and accuracy. In obtaining such information, remote sensing represents a very important tool. This instrument should be combined with other sophisticated modeling systems, to provide agricultural information in a timely manner, across large areas, with high spatial detail and with reasonable costs.

Today, approximately 37% of the earth's land is employed for agricultural purposes, with approximately 11% used for crops and the rest for pasture. Satellites are applied to agriculture in several ways, initially as a means of estimating crop yields. Optical and radar sensors can provide an accurate picture of the acreage being cultivated, while also differentiating between crop types and determining their health and maturity.

Satellites are also used to identify detailed characteristics of individual farmers' fields; they often used in combination with Geographical Information Systems (GIS), to enable more intensive and efficient cultivation practices.

Researchers collect information on the earth's features, to formulate models and to validate hypotheses. Analysts may then use this information and directly examine the phenomenon under investigation by means of special sensors, with remote sensing methods.

In the past, aerial photography was the principal tool used for earth remote sensing based on analogical devices; however, with technological advances, this method was integrated by other special spatial sensors. Currently, the two different technologies can often be considered similar.

Multispectral detection can provide new information that cannot be obtained with the vision-based methods. For example, infrared sensors measure the thermal emission of an object. The temperatures thus obtained may constitute important parameters to study. From a statistical point of view, this information is a typical example of a multivariate data set.

Several types of remote sensing systems are used in agriculture. However, the most common is a passive system that uses the electromagnetic energy reflected from plants. The sun is obviously the most common source of energy for these systems. Passive system sensors can be mounted on satellites, aircraft, or directly on farm equipment.

Before describing how new remote sensing technology has influenced agricultural monitoring, it is necessary to introduce some definitions that will be commonly used in the following pages. First, we discuss different concepts of resolution.

The resolution of a sensor is defined as the measurement of the optical system that is capable of recognizing signals that are spatially close or spectrally similar. We consider four types of resolution: spectral, spatial, temporal, and radiometric (Jensen 2004).

Spectral resolution refers to the size and number of specific ranges of wavelengths to which a sensor is sensitive. Different materials respond to electromagnetic radiation in different ways. Thus, the bands are usually chosen to improve the contrast between the object under investigation and its borders. According to the number of spectral bands used in data acquirement, the satellite images can be classified as: mono-spectral or panchromatic (i.e. with a single wavelength band), multispectral (i.e. with several spectral bands), superspectral (i.e. with tens of several spectral bands), and finally hyperspectral (i.e. with hundreds of spectral bands). The spatial resolution defines the level of spatial precision denoted in the image, and it is a measure of the smallest linear or angular separation between two objects that can be detected by the sensor. The higher the spatial resolution, the more accurately will the sensor detect the phenomenon. In terms of spatial resolution, the images can be classified as: low-resolution images (approximately 1 km or more), medium-resolution images (approximately, from 100 m to 1 km), high-resolution images (approximately, from 5 m to 100 m), and very high-resolution images (approximately, 5 m or less). The radiometric resolution outlines the differences in the sensitivity of a sensor signal to the radiation emitted or reflected from the earth. The radiometric range is the maximum number of quantization levels that can be recorded by a given sensing system. Most sensors record data in 8 bits, with values ranging from 0 to 255 (i.e. 256 levels of gray). Finally, the temporal resolution (or revisit period) concerns the frequency with which a sensor receives images in a specific area. The ability to collect images of the same area of the earth's surface during different periods of time is one of the most important elements for applying remote sensing data. For instance, with multiple analyses of data received at different points in time, it is possible to study how a phenomenon evolves.

Another important concept is the swath of a satellite. This can be defined as the width of the strip observed by each satellite pass. Indeed, as a satellite orbits the Earth, the sensor *sees* a certain portion of the Earth's surface. Image swaths for sensors generally vary between tens and hundreds of kilometres wide. As the satellite circles around the Earth from pole to pole, its east-west position would not change if the Earth did not rotate. However, because the Earth rotates from west to east, the satellite appears to be shifting. This allows the satellite swath to cover a **new area with each consecutive pass**. Finally, some sensors can be only directed straight downwards (i.e. nadir viewing). If the device can point laterally, the sensor has an off-nadir pointing capability.

Several satellite missions have been held to acquire remotely sensed images. At first, these missions were launched principally for gathering weather information; only later was the observation of earth resources included among their main objectives (see Atzberger, 2013, and the references cited therein). In the following pages, we only describe the main satellites and instruments used for agricultural monitoring.

Important weather satellites are now commonly used by the National Oceanic and Atmospheric Administration (NOAA, see http://www.noaa.gov/). NOAA and the National Aeronautics and Space Administration (NASA) have jointly developed a valuable series of Polar-orbiting Operational Environmental Satellites (POES). These spacecrafts have been operative since 1978. NOAA-19, denoted as NOAA-N' (NOAA-N Prime) is the last of the NOAA series of weather satellites. NOAA-19 was launched on 6 February 2009. For our purposes, the principal sensor of interest is the NOAA AVHRR (Richards and Jia, 2006 and Atzberger, 2013). The AVHRR (Advanced Very High Resolution Radiometer) is a radiation-detection imager that can be used to remotely determine cloud cover and the Earth's surface temperature. This scanning radiometer uses 6 sensors, that collect different bands of radiation wavelengths as shown below. The instrument measures reflected solar (visible and near-IR) energy and radiated thermal energy from land, sea, clouds, and the intervening atmosphere. The first AVHRR was a 4-channel radiometer. The latest version is AVHRR/3, with 6 channels. It was first carried on NOAA-15, which was launched in May 1998, although only five are transmitted to the ground at any time.

The use of satellites is usually based on low-resolution images (Rembold *et al.*, 2013). The term "low-resolution satellite images" essentially refers to optical sensors in the reflective domain (i.e. from the visible to the short-wave infrared) and with a spatial resolution between 250 m and several kilometers. The early studies in this field are often related to the use of different sensors of the NOAA AVHRR series. These types of images were typically available at the national and multinational levels with a 1-km resolution and, at the continental and global levels, with a resolution of 4.6 km or lower.

A classical application of this data is to derive vegetation growth profiles that integrate a data sets relating to agricultural land-use; namely, survey-based production statistics. Because the technique reallocates aggregated production statistics consistent with low-resolution imagery and with limited ground control information, it is a particularly cost-effective method of extensive land-use mapping (Walker and Mallawaarachchi,1998; Benedetti and Rossini, 1993; Benedetti *et al.*, 1994).

The French–Belgian–Swedish satellite SPOT was equipped with a 1-km resolution sensor for vegetation monitoring on a global scale only at the end of the 1990s. In addition, several so-called medium-resolution sensors (250 m maximum) have become operational since the year 2000; among the best known is MODIS, belonging to the TERRA/AQUA platforms. All low- and medium-resolution sensors have found effective application in agricultural studies.

The main advantages of the low- and medium-resolution images are represented by the large spatial coverage and high temporal revisit frequency. For these reasons, they are particularly useful for near real-time information collection on a regional scale. Several stakeholders specifically request this information. For example, national and international agricultural agencies, insurance agencies, and international agricultural organizations need maps of crop types, to compile inventories on crop growth in certain areas. The rising availability of low- and medium-resolution satellite images is particularly important for several countries, especially developing countries with arid and semi-arid climates. This crop monitoring system, based on low- and medium-resolution satellite images, is very simple, but timely and accurate; furthermore, it is particularly significant in these countries, where temporal and geographic rainfall variability can cause high inter-annual fluctuations in production and pose a high risk of famines. In these areas, this source of information is often the only possibility for monitoring agriculture. Indeed, the Earth's entire surface is scanned daily, and the specific costs per ground area unit are very low. These systems are typically used in many food-insecure countries by FAO, FEWSNET (Famine Early Warning System) of the USAID (United States Agency for International Development), and the MARS project of the European Commission.

These sensors' intrinsic drawback is, of course, related to their low spatial resolution, with pixel sizes of about 1 km^2, i.e. far above typical field sizes.

The relationship between crops' spectral properties and their biomass/yield was recognized since the very first spectrometric field experiments. As highlighted by Rembold *et al.*, (2013), there are generally three main groups of techniques that are widely used for coarse scale crop monitoring and yield estimation. These three groups are:

- qualitative crop monitoring;
- quantitative crop yield predictions by regression modeling;
- quantitative yield forecasts using crop growth models.

Crop monitoring methods that are based on the qualitative interpretation of remote sensing-derived indicators (for example, the NDVI index) are considered qualitative crop monitoring. These methods are based on the comparison of the actual crop status to previous seasons or to what can be assumed to be the standard situation. The results obtained tend to identify anomalies of the agricultural system, and to depict possible yield limitations. As opposed to the qualitative approaches, the regression approaches must necessarily be calibrated using appropriate covariate information. In most cases, agricultural statistics and, specifically, crop yields, are used as reference information. This prerequisite limits their applicability in many regions of the world. Finally, crop growth modeling involves the use of mathematical simulation models including the knowledge previously obtained by plant physiologists. The models describe the primary physiological mechanisms of crop growth and their interactions with the underlying environmental driving variables (e.g. air temperature, soil moisture, nutrient availability) by using mechanistic equations (Delécolle *et al.* 1992).

MetOp (see http://www.esa.int/Our_Activities/Observing_the_Earth/The_Living_Planet_Programme/ Meteorological_missions/MetOp) is a series of three polar orbiting meteorological satellites managed by Eumetsat. The satellites use a payload containing 11 scientific instruments. To provide data continuity between MetOp and NOAA POES, both series of satellites carry several instruments. MetOp-A was launched on 19 October 2006, and is Europe’s first polar orbiting satellite used for operational meteorology. The launching of subsequent satellites in the series is provisionally planned at 5-year intervals, to match the satellites’ 5-year design life and to provide continuity of service.

The **Agricultural Stress Index System** (ASIS) by FAO (http://www.fao.org/climatechange/asis/en/) is based on 10-day satellite data of vegetation and land surface temperature from the METOP-AVHRR sensor, at a resolution of 1 km.

The early French SPOT (Système pour l’Observation de la Terre, http://www.cnes.fr/web/CNES-en/1415-spot.php) satellites had two imaging sensors, referred to as High Resolution Visible (HRV). These instruments utilize two different images modes: one is a multispectral mode, and the other is panchromatic. The following SPOT missions (i.e. SPOT 4 and SPOT 5) mounted sensors with similar characteristics, as well as the Vegetation instrument. SPOT

5 was launched on 4 May 2002. SPOT 5 has two High-Resolution Geometrical (HRG) instruments. They provide a higher resolution of 2.5 to 5 m, in panchromatic mode and 10 meters in multispectral mode.

The vegetation program (http://www.spot-vegetation.com/index.html) is the result of the space collaboration between various European partners: Belgium, France, Italy, Sweden, and the European Commission. The program consists of two observation instruments: VEGETATION 1, aboard the SPOT 4 satellite, and VEGETATION 2 aboard SPOT 5. These deliver measurements that were specifically designed to monitor land surfaces' parameters with a frequency of approximately once per day, on a global basis (some gaps remain near the Equator), and an average spatial resolution of one kilometer. The mission is now nearing the end of its life cycle. From the summer of 2013 onwards, the role of SPOT-VEGETATION was taken over by the European Space Agency (ESA)'s technologically advanced PROBA-V mission (see below).

The SPOT-6 satellite (http://www.astrium-geo.com/en/147-spot-6-7) built by Astrium was successfully launched on 9 September 2012. SPOT-6 is an optical imaging satellite that is capable of imaging the Earth with a resolution of 1.5 m in panchromatic mode, and 6 m in multispectral mode (i.e. Blue, Green, Red, Near-IR), and that can produce images useful for the purposes of defence, agriculture, and environmental monitoring. SPOT-6 and SPOT-7 (that will probably be launched in 2014) will provide a daily revisit for each location that can be detected on Earth, with a total coverage of 6 million km^2 per day.

The SPOT program is now complemented by Pléiades satellites (see http://www.astrium-geo.com/pleiades/). Pléiades 1A and Pléiades 1B work as a constellation in the same orbit, phased 180° apart. These identical twin satellites provide very-high-resolution optical images in record time, and offer a daily revisit capability for any point on the globe. The Pléiades satellites are designed to obtain data and to acquire imagery from anywhere in the world in less than 24 hours, in response to a crisis or natural disaster, and to ensure a regular monitoring even on a daily basis, if required.

Envisat (Environmental Satellite) is an ESA satellite that is still in orbit (http://www.esa.int/Our_Activities/Observing_the_Earth/Envisat_overview). It was the largest Earth Observation spacecraft ever built. It transported ten sophisticated optical and radar instruments to provide continuous observation and monitoring of Earth's land, atmosphere, oceans and ice caps. Its largest payload was the Advanced Synthetic Aperture Radar (ASAR). Operating at C-band, it ensured continuity of data after ERS-2. The Medium Resolution Imaging Spectrometer (MERIS) was an imaging spectrometer with a ground spatial resolution of 300 m, 15 spectral bands. MERIS enabled global coverage of the Earth every three days. In April 2012, contact with Envisat was suddenly lost, and ESA declared the mission finished.

Some satellite missions are particularly noteworthy, especially for the high spatial resolution images that provide. IKONOS (http://www.digitalglobe.com/about-us/content-collection#ikonos) is a commercial satellite that was launched on 24 September 1999. It is a sun-synchronous satellite, with a 3-days revisit capacity with off-nadir pointing capability (the frequency depends on the latitude). It provides multispectral and panchromatic images, and it was the first to collect publicly available high-resolution imagery at 0.82 (i.e. in panchromatic band) and 3.2-meter resolution (i.e. multispectral mode) at nadir.

QuickBird is DigitalGlobe's primary commercial satellite (http://www.digitalglobe.com/about-us/content-collection#quickbird) that offers sub-metrer resolution images, high geo-locational accuracy, and large on-board data storage. It delivers both panchromatic and multispectral images, and it is designed to support a wide range of geospatial applications.

WorldView-1 (http://www.digitalglobe.com/about-us/content-collection#worldview-1), launched in September 2007, is the first of DigitalGlobe's new-generation satellites. It operates at an altitude of 496 kilometres. WorldView-1 has an average revisit time of 1.7 days (depending on the latitude), and is capable of collecting over 1 million km^2 per day of half-metre images. This satellite is also equipped with state-of-the-art geo-location accuracy instruments.

WorldView-2 (http://www.digitalglobe.com/about-us/content-collection#worldview-2) was launched in October 2009, and is the first high-resolution 8-bands multispectral commercial satellite. It operates at an altitude of 770 km, and it provides images with 46-cm resolution for the panchromatic sensor and with 1.85 m resolution for the multispectral device. WorldView-2 has an average revisit time of 1.1 days (depending on the latitude), and is capable of collecting up to 1 million km^2 of 8-band imagery per day.

WorldView-3 (http://www.digitalglobe.com/about-us/content-collection#worldview-3) will probably be launched in 2014. WorldView-3 will provide images with 31 cm panchromatic resolution and 1.24 m multispectral resolution. WorldView-3 will have an average revisit time of less than one day, and will collect data for up to 680,000 km^2.

Among the missions having the monitoring of earth resources as a main objective, the Landsat expeditions are highly remarkable (see Richards and Jia, 2006). The first three Landsats (see http://landsat.usgs.gov/) had identical orbit features. All satellites nominally obtained images at 9:30 a.m. local time on a descending (i.e. north-to-south) path. The complete coverage of the earth's surface is ensured, with 251 revolutions in 18 days. The characteristics of the orbits of second-generation Landsats (from Landsat 4 onward) are different from those of previous generations. Again, images are acquired nominally at 9:30 a.m.

local time, but the earth's surface is covered with a total of 233 revolutions in 16 days. The current version (Landsat 7) is a similar satellite in all respects. Three different sensors have been used on the Landsat satellites: the Return Beam Vidicon (RBV), the Multispectral Scanner (MSS), and the Thematic Mapper (TM). The primary sensor on board Landsats 1, 2, and 3 was the MSS, with an image resolution of approximately 80 m in four spectral bands, ranging from the visible green to the near-infrared (IR) wavelengths. The MSS was not used after Landsat 5.

With the launch of Landsat 7, Enhanced Thematic Mapper + (i.e. ETM+) was added. The Thematic Mapper has improved spectral, spatial, and radiometric characteristics. Seven wavelength bands are used. The spatial resolution is of 30 m for the visible, near-IR, and shortwave infrared (SWIR) bands, and the addition of a 120-m thermal-IR band. Besides, the ETM+, mounted on Landsat 7, also includes a panchromatic band. On 30 May 2013, data from the Landsat 8 satellite (launched on 11 February 2013) became available. This project, known as the Landsat Data Continuity Mission (LDCM), endures the acquisition of high-quality data that meet the scientific and operational requirements of both NASA and the United States Geological Survey (USGS) for observing land use and land change. Landsat 8 operates in the visible, near-infrared, short wave infrared, and thermal infrared spectrums. The Operational Land Imager (OLI) and the Thermal InfraRed Sensor (TIRS) sensors are used. The OLI collects data in nine shortwave bands, eight spectral bands at a 30 m resolution and one panchromatic band at 15 m. The TIRS captures data in two long wave thermal bands with 100 m resolution, and is registered to and delivered with the OLI data as a single product. The USGS currently distributes Landsat data at no charge to users, via the Internet.

For further information on NASA's next missions to complement and/or substitute the Landsat program, readers are referred to http://www.nasa.gov/missions/schedule/index.html#.UqsWAI0lvx4.

The Moderate Resolution Imaging Spectroradiometer (MODIS, see http://modis.gsfc.nasa.gov/), included in NASA's Earth Observing Systems (EOS) project, is crucial for monitoring agriculture resources. MODIS is a scientific instrument launched by NASA in 1999, on board the Terra satellite, and in 2002 on board the Aqua satellite. MODIS Terra's orbit passes from north to south across the Equator in the morning, while Aqua passes south to north over the Equator in the afternoon. Terra MODIS and Aqua MODIS observe the Earth's entire surface every 1 to 2 days, acquiring data in 36 spectral bands that range in wavelength from 0.4 microns to 14.4 microns, and with different spatial resolutions (bands 1-2 at 250 m, bands 3-7 at 500 m, and bands 8-36 at 1 km). The measurements seek to improve understanding of global dynamics, including changes in the Earth's cloud cover, the radiation budget and the processes occurring in the oceans, on land, and in the lower atmosphere. MODIS plays a vital role to support policy makers in making appropriate decisions concerning environmental protection (see also Roy *et al.*, 2002).

The availability of information will increase with Proba-V sensors (http://www.esa.int/Our_Activities/Technology/Proba_Missions), and with the launch of ESA's new Sentinel mission (see http://www.esa.int/Our_Activities/Observing_the_Earth/GMES/Overview4).

Proba-V, where V stands for vegetation, is a small satellite that uses a redesigned version of the Vegetation imaging instruments previously aboard France's SPOT-4 and SPOT-5 satellites, which have observed Earth since 1998. The Proba-V project was initiated by the Space and Aeronautics Department of the Belgian Science Policy Office. Today, it is operated by ESA. It was launched very recently (on 7 May 2013) to fill the gap between the end of SPOT's missions and the upcoming Sentinel project (see below). However, because of the change of Sentinel's satellites, Proba-V will ensure the continuation of the Vegetation program. Proba-V will support applications such as land use, worldwide vegetation classification, crop monitoring, famine prediction, food security, disaster monitoring and biosphere studies. Proba-V data will be available at a spatial resolution of 100 m.

The Sentinel's project started in 2014, and is specifically formulated to meet the operational needs of the Global Monitoring for Environment and Security (GMES) program. GMES seeks to provide, accurate, timely, and easily accessible information to improve the management of the environment. The Sentinel project consists of 5 missions, the last of which is scheduled for 2020. Copernicus is GMES' new name, as was announced on 11 December 2012, by the European Commission's Vice-President Antonio Tajani, during the Competitiveness Council.

Sentinel-1 (Torres et al., 2012) is a near polar sun-synchronous, day-and-night radar imaging mission for land and ocean services, and seeks to continue SAR's operational applications. Sentinel-1 satellites are being built by an industrial consortium led by Thales Alenia Space (Italy) as the Prime Contractor, while Astrium (Germany) is responsible for the C-band Synthetic Aperture Radar (CSAR) payload, which incorporates the central radar electronics subsystem developed by Astrium. Sentinel-1's revisit frequency and coverage are dramatically better than those of the European Remote Sensing satellites (ERS-1 and 2) SAR, and the Envisat ASAR. Compared to its predecessors, the Sentinel-1 mission represents a significant increase in capability. Sentinel-1 satellites are expected to provide coverage over Europe, Canada and main shipping routes in 1–3 days, regardless of weather conditions. Sentinel-1 was designed to address mainly medium- to high-resolution applications, through a main mode of operation that features both a wide swath (250 km) and high spatial (5x20 m) and radiometric resolution. Based on the mission's requirements, the following main modes of operational measurement are implemented: Interferometric Wide-swath mode (IW), Wave mode (WV), and in the interests of continuity and to meet emerging user requirements, Strip Map mode (SM), and Extra Wide-swath mode (EW). In particular, Sentinel-1 provides SAR imaging for monitoring sea-ice zones and the polar environment, mapping in support of humanitarian aid in crisis situations, surveillance of marine environments, monitoring land surface motion risks, and mapping of land surfaces (forest, water and soil, and agriculture). Except for the WV mode, which is a single-polarization mode, the CSAR instrument supports operation in dual polarization.

Sentinel-2 satellites (Drusch *et al*., 2012) operate simultaneously with a sun-synchronous orbit, at an altitude of 786 km. The two satellites will work on opposite sides of the orbit. They provide multispectral high-resolution images for land cover, land usage and land-use-change detection maps, inland waterways and coastal areas, geophysical variable maps (i.e. leaf chlorophyll content, leaf water content, leaf area index), risk mapping, and fast images for disaster relief efforts. Sentinel-2 can also deliver information for emergency services. The two satellites were designed as a reliable multispectral Earth observation system that will ensure the continuity of Landsat and SPOT observations and improve the availability of data for users. Compared to the SPOT and Landsat, the Sentinel-2 mission offers an unique combination of systematic global coverage of land surfaces from 56°S to 84°N, including coastal waters, the Mediterranean and selected calibration sites, a high revisit frequency (every five days at the Equator under the same viewing conditions), a high spatial resolution (10 m, 20 m, and 60 m), multispectral information with 13 bands in the VNIR and SWIR parts of the spectrum, and a wide field of view (290 km). The launch of the first Sentinel-2 satellite was scheduled for 2014.

Sentinel-3 (Donlon *et al*., 2012) is multi-instrument mission with the objective of measuring variables such as sea-surface topography, and sea and land-surface temperature. The mission expects to launch a series of satellites, each having a 7-year lifespan, over a 20-year period, starting with the launch of Sentinel-3A in late 2013 and of Sentinel-3B in late 2014. When the mission is fully operative, two identical satellites will be maintained in the same orbit, with a phase delay of 180°. Sentinel-3 directly follows the path outlined by ERS-2 and Envisat. Its innovative instrument package includes several devices. The first is the Sea and Land Surface Temperature Radiometer (SLSTR), which is based on Envisat's Advanced Along Track Scanning Radiometer (AATSR). SLSTR measures in nine spectral channels, plus two additional bands optimized for fire monitoring, and has a dual view (i.e. near-nadir and inclined). SLSTR has a spatial resolution in the visible and shortwave infrared channels of 500 m, and of 1 km in the thermal infrared channels. An Ocean and Land Colour Instrument (OLCI) is based on Envisat's Medium Resolution Imaging Spectrometer (MERIS). The OLCI has 21 bands, compared to the 15 on MERIS, and a spatial resolution of 300 m over all surfaces. The OLCI swath is not centred at nadir (as in the MERIS design), but is tilted 12.6° westwards. A dual-frequency (Ku and C band) advanced Synthetic Aperture Radar Altimetre (SRAL) provides measurements at a spatial resolution of about 300m in SAR mode. SRAL is supported by a dual frequency passive microwave radiometer (MWR) for wet-tropospheric correction, and a DORIS receiver for orbit positioning. This combined topography package will provide exact measurements of sea-surface height, which are essential for ocean forecasting systems and climate monitoring. The pair of Sentinel-3 satellites will enable a short revisit time of less than two days for OLCI, and of less than one day for SLSTR at the Equator.

The Sentinel-4 and Sentinel-5 missions will be devoted to monitoring the composition of the atmosphere (Ingmann *et al.*, 2012). Both missions will be carried out on meteorological satellites operated by Eumetsat. The Sentinel-4 mission includes an Ultraviolet Visible Near-infrared (UVN) spectrometer and data from Eumetsat's thermal InfraRed Sounder (IRS), both boarded on the MTG-Sounder (MTG-S) satellite. Once the MTG-S satellite is in orbit, the Sentinel-4 mission will also include data from Eumetsat's Flexible Combined Imager (FCI), embarked on the MTG-Imager (MTG-I) satellite. The first MTG-S satellite will be probably launched in 2019, and the first MTG-I in 2017. For a tentative description of the requirements of Sentinel-4 and Sentinel-5, see Ingmann *et al.* (2012).

Other important sensors that will be launched in the near future are VENµS (see http://smsc.cnes.fr/VENUS/index.htm) and the hyperspectral HyspIRI (see http://hyspiri.jpl.nasa.gov/).

The Vegetation and Environment monitoring on a New Micro-Satellite (VENµS) project is designed to enable the combination of high spatial and temporal resolution, and will be launched in 2014 (VM1 is its first mission). Indeed, the VENµS microsatellite was created by the French *Centre National d'Etudes Spatiales* (CNES) and the Israeli Space Agency (ISA) for this very aim. This satellite operates in the visible to the near infrared range, and the camera will cover a total of 12 spectral bands. Besides, VENµS will observe 100 sites that are representative of the main terrestrial and coastal ecosystems in the world, every two days. VENµS's main objective is the monitoring of vegetation growth. This data will also be useful for improving estimates of the evolution of water resources.

NASA's Hyperspectral Infrared Imager (HyspIRI) aims to detect the responses of ecosystems and climate change. The HyspIRI mission includes two instruments mounted on a satellite in Low Earth Orbit. There is an imaging spectrometer that measures from the visible to shortwave infrared (VSWIR: 380 nm – 2500 nm) in 10 nm contiguous bands, and a multispectral imager measuring from 3 to 12 um in the mid- and thermal infrared (TIR). The spatial resolution is of 60 m at nadir. The VSWIR will have a revisiting time of 19 days and the TIR will have a revisit of 5 days. The mission is currently at the study stage; its website serves as a principal information centre on the mission.

With the goal of capturing frequent pictures that show the planet's changes in real time, the Planet Labs company (http://www.planet-labs.com/) launched two demonstration satellites, to test technologies and operations on a space-based platform. The satellites are known as *Dove 1* and *Dove 2*. In early 2014, Planet Labs launched 28 small satellites, which constitute the world's largest constellation of Earth-observing satellites. Each Dove image will consist of relatively affordable 10-centimetre-wide building blocks called CubeSats.

The RapidEye constellation (http://www.satimagingcorp.com/satellite-sensors/rapideye.html) of five Earth Observation satellites has been in operation since February 2009. The system images a swath that is 77 kilometres wide, which enables over five million square kilometres of Earth to be depicted every day for its archive, and over one billion km^2 every year per day, at a nominal spatial resolution of 6.5 m.

It is clear that the availability of data has increased over the last few decades. In particular, it is noteworthy that the satellite sensors provide images with very different spectral, spatial, temporal, and radiometric characteristics. Therefore, depending on the purpose of the analysis, it is possible to choose an appropriate data type. These characteristics are a synopsis of each satellite's main advantages and drawbacks. The different features of the data collected by the operational and future satellite payloads described above are summarized in Tables 2.1 and 2.2 below.

Agricultural monitoring is not a recent practice. The first monitoring system can be traced back to the ancient Egyptians, who assessed the cultivated areas affected by the water level fluctuations of the River Nile, for the purposes of taxation and preventing famine (Atzberger, 2013). As highlighted by Becker-Reshef *et al.* (2010), the Landsat system was the first system designed to provide near-global coverage of the earth's surface on a regular and predictable basis. NASA and the US Department of Agriculture (USDA) have been working together to monitor global agriculture from space since the 1970s. Landsat 1 was launched by NASA on July 1972. To improve domestic and international crop forecasting methods, in 1974 the USDA, NASA and NOAA initiated the Large Area Crop Inventory Experiment (LACIE).

The NOAA AVHRR sensor, enabling daily global monitoring, led to the definition of the AgRISTARS (Agriculture and Resource Inventory Surveys Through Aerospace Remote Sensing) program, launched in the early 1980s.

More recently, NASA and the USDA Foreign Agricultural Service (FAS) have introduced the Global Agricultural Monitoring (GLAM) Project (see Becker-Reshef *et al*., 2010), focused on applying data from NASA's MODIS instrument.

Currently, there are several other operational agricultural monitoring systems that make use of remote sensing information, and that provide critical agricultural information.

The Famine Early Warning Systems Network (FEWSNET; see http://www.fews.net/Pages/default.aspx) is the USAID-funded activity that has the objective of delivering timely warnings and vulnerability information on emerging and evolving food security issues. The project provides monthly food security updates for 25 countries, as well as regular food security outlooks.

The FAO Global Information and Early Warning System (GIEWS; see http://fao.org/gviews) seeks to keep the world food supply/demand situation under continuous review, and to provide early warnings of impending food crises in individual countries.

The Joint Research Centres's (JRC) Monitoring Agricultural ResourceS action of the European Commission in Ispra, Italy (MARS; see http://mars.jrc.ec.europa.eu/), focuses on crop production, agricultural activities and rural development. MARS offers timely forecasts and early assessments, thus enabling efficient monitoring and control systems.

TABLE 2.1. Main characteristics of certain operational satellite sensors.

	Spatial resolution	Channels	Swath at nadir (Km)	Revisit days at nadir	Off nadir pointing
NOAA –	1.09 Km	6	2900	1	No

AVHRR/3					
Landsat 7 ETM + (multispectral)	30 m	6	185	16	No
Landsat 7 ETM + ETM (thermal)	60 m	1			
Landsat 7 ETM + (panchromatic)	15 m	1			
Landsat 8 OLI + (multispectral)	30 m	8	185	16	Yes
Landsat 8 OLI + (panchromatic)	15 m	1			
Landsat TIRS (thermal)	100	2	185	16	Yes
SPOT 5 (multispectral)	10-20 m	4	60	26	Yes
SPOT 5 (panchromatic)	2.5 m	1	60	(2-3 off –nadir)	
SPOT 6 (multispectral)	8 m	4	60	26	Yes
SPOT 6 (panchromatic)	1.5 m	1		(1-3 off nadir)	
IKONOS (multispectral)	3.2 m	4	11.3	≃141	Yes
IKONOS (panchromatic)	0.82 m	1			
QuickBird (multispectral)	2.44 m	4	16.8	≃2.4	Yes
QuickBird (panchromatic)	61 cm	1		(40°N Lat)	Yes
WorldView – 1 (panchromatic)	50 cm	1	17.7	≃1.7 (40°N Lat)	Yes
WorldView – 2 (multispectral)	1.85 m	8	16.4	≃ 1.1	Yes
WorldView – 2 (panchromatic)	46 cm	1		(40° N Lat)	Yes
MODIS	250 m (bands 1-2) 500 m (bands 3-7) 1 Km (bands 8-36)		2330	1	No
Proba – V	100 m	4	2250	1-2	Yes

TABLE 2.2. Main characteristics of certain satellite sensors, soon to become operational.

	Spatial resolution	**Channels**	**Swath (Km)**	**Revisit days at nadir**	**Off nadir pointing**
WorldView-3	1.24 m	8	13.1	<1	Yes

(multisp.) [2014]				(40º N Lat)	
WorldView-3 (panchromatic)	31 m	1			
WorldView-3 (SWIR)	3.70 m	8			
WorldView-3 (CAVIS)	30 m	12			
Sentinel -1 (IW) [2013]	5x20 m	1 mode	250	12(with 1 satellites)	Yes
Sentinel -1 (WV)	5x5 m	1 mode	20		
Sentinel -1 (SM)	5x5 m	1 mode	80	6 (with 2 satellites)	
Sentinel -1 (EW)	20x40 m	1 mode	400		
Sentinel-2 [2014]	10 m	4	290	<5 at equator	Yes
	20 m	6			
	60 m	3			
Sentinel -3 (SLSTR) [2014]	500 m – 1 Km	9+2 for fire mon.	1420	<1 at equator	Yes
Sentinel-3 (OLCI)	300 m		1270	<2 at equator	Yes
Sentinel-3 (SRAL)	300 m		>2	27	No
VENµS -VM1 [2014]	5.3 m	12	27.5	2	Yes
HyspIRI (VSWIR)	60 m	220	145	19	Yes
HyspIRI (TIR)	60m	8	600	5	No

The China CropWatch System (CCWS; see http://www.cropwatch.com.cn/en/) was introduced in 1998 by the Institute of Remote Sensing Application (IRSA) of the Chinese Academy of Sciences (CAS), and has been operational ever since. This project covers the entire country of China, as well as 46 major grain-growing countries of the world. The system monitors the conditions of growing crops, crop production, drought, and of the cropping index.

However, the GLAM system is currently the only source of regular, timely, and objective crop production forecasts on a global scale. This result, as highlighted by Atzberger (2013), is ascribable to the close cooperation between USDA and NASA, and is based on the use of MODIS. Consequently, monitoring systems must rely heavily upon the time series provided by these sensors.

ESA's new missions (i.e. Proba-V and Sentinel) will, of course, represent a drastic improvement in addressing the specific needs of stakeholders dealing with agricultural data.

The data acquired from large satellite sensors, such as those described above, can be employed in a wide range of agricultural applications (Atzberger, 2013). The main agricultural applications of remote sensing occur in relation to crop type classification, crop condition assessment (e.g. crop monitoring and damage assessment), crop yield estimation, mapping of soil characteristics and type, and soil erosion.

As already mentioned, the main characteristics that a satellite system should possess to adequately monitor agriculture concern spatial coverage, and a high revisit frequency. Currently, the Earth's land surface can only be covered by coarse/medium resolution sensors, such as MODIS (Atzberger, 2013). Consequently, a monitoring system must rely heavily upon the time series provided by these sensors. The situation will improve considerably with the upcoming Sentinel 2 and 3 and Proba-V sensors. For example, Proba-V will provide images with a spatial resolution of 100 m, Sentinel-2 will provide 10–30 m data at five-day revisit intervals and the Venus sensor will be launched in April 2015, with 12 spectral bands (5 m ground resolution) and a two-day revisiting time (see Tables 2.1 and 2.2 above).

This dramatic improvement in the spatial resolution of satellite images will lead to a positive impact on the prediction of agriculture variables, especially with regard to the precision of the auxiliary information (represented by the satellite images) that is fundamental to the phase of sample design and estimation in several methodologies.

Remotely sensed images could be used to produce maps of crop types. This information is added to the traditional methods of census and ground surveying. The use of satellites is valuable, as it can systematically cover of a large area, and provide data on the health of the vegetation. Satellite data are used by agricultural agencies to prepare an inventory of the crops grown in certain areas, and when they were grown. See e.g. Gallego (1999) for some examples relating to the MARS project.

A document on best practices for crop area estimation with remote sensing was prepared by the Global Earth Observation System of Systems (GEOSS, 2009); the text focuses on the use of remote sensing data as an auxiliary variable for improving the precision of estimates for specific crops.

Remote sensing features several attributes that contribute towards monitoring crop health. Remote sensing can assist in identifying crops affected by conditions that are too dry or wet, insects, weed or fungal infestations, or weather-related damage.

Besides, monitoring agricultural crop conditions during the growing season and estimating the potential crop yields are very important for the determination of seasonal production. For yield prediction and estimation of crops it is necessary to achieve a very high accuracy and reliability. See Ferencz et al (2004) for an interesting application of estimate of the yield of different crops in Hungary from satellite remote sensing.

The disturbance of soil by land use impacts the quality of the environment, entailing, for example, salinity, soil acidification and erosion. Remote sensing is a good method for mapping and predicting soil degradation.

Finally, it is important to highlight that Euroconsult periodically writes excellent reports on the satellite market. For further details, readers may refer to: http://www.euroconsult-ec.com/research-reports-28.html.

In the following sections, we present a comprehensive review on how remote sensing information can be a valuable instrument in statistical theory applied to agricultural data.

2.1 A note on the use of high- resolution images

The monitoring of agriculture is not a recent issue (Becker-Reshef *et al*., 2010). In this Report, we will explore some issues of satellite missions relevant to the system for monitoring agriculture, with particular reference to the introduction of new high-resolution images and to the use of information from Google Maps.

Image technology has recently made considerable progress. For example, the introduction and widespread use of Google Maps and Google Earth could constitute a valid support for a country's agricultural monitoring system. Google Maps is an open technology that offers maps (Choimeun et al., 2010). Operators can use web browsers to examine the maps. The technology contains several convenient built-in tools, such as zooming in and out, marking, and view of satellite data. While Google Maps provides map information, Google Earth is the software for viewing satellite data in high resolution. Several alternatives to these products exist on the market, such as Yahoo! Maps, Microsoft Bing, Nokia Maps, Cyclomedia and WikiMapia; however, all of these share similar characteristics, inherited from the precursor Google Maps.

The visual localization of images is an important objective. For example, the images that can be extracted from Google Maps Street View are a remarkable illustration of such a data set. Google Maps Street View is a highly comprehensive dataset, consisting of 360° panoramic views of almost all main streets and roads in a number of countries, with a distance of about 12 metres between locations.

According to Zamir and Shah (2010), the main advantages of Google Maps Street View images are:

- query independence: the localization task is independent of the popularity of the objects in the query image and the location.

- accuracy: as the images in the data set are spherical 360° views taken approximately every 12 metres, it is possible to correctly localize an image with a higher degree of accuracy.
- secondary applications: using a structured database enables us to derive additional information, without the need for further in-depth computation. Localization and orientation determination are issues that even hand-held GPS devices cannot reach without motion information.

However, the main benefit from the use of this type of images is, obviously, the fact that they are provided free of charge.
Unfortunately, the use of the Google Maps Street View dataset also has some drawbacks. For example, the huge number of images can hinder fast localization. This need to capture a large number of images always adds some noise and geometric distortions to the images. Storage limitations make it impossible to save very high quality images; therefore, the information available consists in a distorted, low-quality, large-scale image data set. For further details, see Zamir and Shah (2010).

Our main recommendation relating to the use of Google Maps images in an agricultural monitoring system regards the reason for which this product was introduced. The objective of these images is to provide users with some information on a territory, city or other entities. In other words, the main objective is to provide generic visual evidence about a location; it is a powerful tool, for example, for viewing a zone that a person wishes to visit or become acquainted with. Therefore, some caution is needed when applying Google Maps as the *main source of information* in agriculture monitoring resources. On the other hand, it is an effective source of auxiliary information that can support the use of *classical* satellite images and *in situ* surveys, in monitoring agricultural systems.

The main problem is that Google Earth and Street View can be considered as two cartographic products; however, they cannot be considered as remotely sensed data that can be used as auxiliary information (in particular, Street View images are entirely unrelated to remote sensing). Google Earth image resolution varies between 20 cm, 1 m and 20 m; the period in which individual images were recorded is unknown. Absolute geometry is not guaranteed (even if local geometry is good). Image dates vary abruptly. Thus, their potential use in agricultural statistics is necessarily limited to being field documents or, at most, to assist photo-interpreters in partitioning a territory (although in this case, it is particularly important that the reference date and period is unknown).

3

Methods for using remote sensing data at the design level

3.1 Introduction

Surveys are routinely used to gather primary data in agricultural research. The units to be observed are often randomly selected from a finite population whose main feature is its geo-referenced nature. Thus, in designing the sample, the spatial distrubition of the sampled units is crucial.

In business surveys in general, and in multipurpose agricultural surveys in particular, the problem of designing a sample from a frame usually consists of three different aspects. The first is the choice of a rule for stratifying the population when several size variables are available; the second is the definition of the selection probabilities for each unit in the frame; and the third is devoted to sample size determination and sample allocation to a given set of strata. The main property required of the sample design is that it should provide a specified level of precision for a set of variables of interest, using as few sampling units as possible.

Stratification is introduced into sampling designs for a number of different reasons: for example, to select the sample from a given frame, to obtain reliable estimators for sub-populations (i.e. domains), or to improve the estimators' efficiency of global population parameters.

In most cases, populations are either naturally stratified or can be stratified easily, on the basis of practical considerations such as administrative subdivisions. In other circumstances, strata are established to satisfy the interest in identifying characteristics of sub-populations. When such straightforward definitions are not possible, then a decision should be taken on the number of strata and their respective boundaries.

These typical design issues should be considered together with that of the importance of bearing geographical position in mind when selecting samples of statistical units. Today, this issue is recognized more than ever when measuring many phenomena, due to several reasons. First, because evidence exists that statistical units are defined by using purely spatial criteria, as occurs in most agricultural and environmental studies. Second, in several countries, it is common practice for the National Statistical Institute (NSI) to geo-reference the typical sampling frames of physical or administrative bodies not only according to the codes of a geographical nomenclature, but also by adding information on the exact or estimated, position of each record.

Often, spatial units are also artificially defined, and made available over a domain partitioned into a number of predetermined regularly- or irregularly- shaped sets of spatial objects. This may occur, for example, when the original data lie over a continuous spatial domain and, to simplify the problem, researchers decide to observe them only in a selection of fixed points, potentially made at random or averaged over a selection of predefined polygons.

Although the monitoring and estimation of infinite populations covers an important part of sampling problems in the context of natural resources, agricultural surveys deal mainly with finite populations.
Indeed, in this latter context, the spatial distribution of the frame is a constraint.

For this reason, it is suspected that it could have considerable impact on the performance of a random sampling method. For example, the traditional solution of extending the systematic sampling to multidimensional data by simply overlaying a grid of points to a spatial domain may not be feasible, if the population cannot be considered as distributed on a regular grid because it is clustered or displays different intensities of units across the domain.

Assume that it is sought to estimate a parameter of a set $\mathbf{Y}=\{\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_v\}$ of v variables of interest, generally denoted as survey variables. Let $U=\{1,2,\ldots,N\}$ be a finite population recorded on a frame together with a set of k auxiliary variables $\mathbf{X}=\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_k\}$ and a set of h (usually h=2) coordinates $\mathbf{C}=\{\mathbf{c}_1,\mathbf{c}_2,\ldots,\mathbf{c}_h\}$, obtained by geo-coding each unit, where $\mathbf{x}_l=\{x_{1l},x_{2l},\ldots,x_{il},\ldots,x_{Nl}\}$ is the generic l-th auxiliary and $\mathbf{c}_l=\{c_{1l},c_{2l},\ldots,c_{il},\ldots,c_{Nl}\}$ is the generic l-th coordinate. From $\mathbf{C}$ we can always derive, for any distance definition, a matrix $\mathbf{D}_U=\{d_{ij};i=1,\ldots,N,j=1,\ldots,N\}$, which specifies how far apart are all the pairs of units in the population.

In several agricultural surveys, the geographical position is an intrinsic characteristic of the unit and, given the particular nature of this information, its efficient use in sample design often requires methods that cannot be adapted from those used when dealing with classical auxiliary variables.

This is not only a consequence of its multivariate nature and of traditional design solutions, as the *πps* (i.e. inclusion probability proportional to size) can handle only one auxiliary (Bee *et al.*, 2010). To use certain covariates, we always assume that there is, at least approximately, a certain degree of correlation between a survey variable y and the set **X**. With regard to the use of set **C**, the commonly used distance matrix as a synthesis of the spatial information emphasizes the importance of the sample's spread over the region under study. This feature may be related to this dependence, but also to some form of similarity between adjacent units.

Usually, **X** and **C** in agricultural surveys play different roles according to the definition of the statistical unit:

1. When U is a list of agricultural households, **C** is rarely obtainable, depending on the availability of accurate cadastral maps, and should be constituted by a map of polygons representing the parcels of land used by each holding. **X** is usually filled with administrative data sources, previous census data and, only if **C** is available, remotely sensed data obtained through the overlay of the polygon map with a classified image;
2. If U is a list of regularly- or irregularly-shaped polygons defined *ad hoc* for the agricultural survey, **C** is always available, since it represents the very definition of each statistical unit and **X**, unless an overlay of **C** with a cadaster is possible, can be constituted only by some geographical coding and by summarizing a classification arising from remotely sensed data within each polygon;
3. Another possible choice often made in agricultural surveys is for U to represent a list of points, usually the corners of the regular grid, overlaid over the survey geographical domain. This, therefore, represents a non-exhaustive population of the study area and only a first stage of sampling. In this case, **X** can be only represented with a geographical nomenclature and with a design matrix of codes for land use classification obtained, with previous land use maps or with a classification of remotely sensed data, while **C** represents the coordinates of each point.

In the first type of survey, the relevant structural characteristic to be controlled is that the population under investigation is highly skewed, as the concentration of farms and agricultural households' sizes is very high. Most of these units have a small size, and are not important in economic terms, although they may be interesting for a rural development analysis. On the other hand, a limited number of large units represents a significant part of the population, and must, therefore, always be included in any sample survey. This is a typical situation arising in any business survey in which the population of interest is extremely positively skewed, due to the presence of a few *large* units and several *small* units. Thus, when estimating an unknown total of the population, many small observations make a negligible contribution, whereas a few large observations have a dramatic impact on the estimates.

In sampling theory, a large concentration of the population with respect to surveyed variables constitutes a problem that is difficult to handle, without using selection probabilities that are proportional to a size measure or a stratification or partition tool. These issues will be examined in Sections 3.2 and 3.3, respectively. As for the efficient use of the spatial information in C, the interest is focused on probability samples that are well-spread over the population in every dimension; recent literature defines these as spatially balanced samples. We discuss this topic in Section 3.4 below. Finally, Section 3.5 describes the problem of auxiliary and survey variables.

3.2 Multivariate auxiliaries in πps sampling

One of the methods for the *ex ante* use of auxiliary information is to use a sampling scheme with inclusion probabilities that are proportional to given size measures, a so-called πps scheme (Rosén, 1997a; Rosén, 1997b; Foreman and Brewer, 1971). This sampling scheme presents desirable properties, but cannot be applied in practical situations where the frame contains a multivariate **X,** because it is seriously limited by the drawback that the method can use only one auxiliary variable (Benedetti *et al.*, 2010).

The design of a πps random sample from a finite population, when multivariate auxiliary variables are available, deals with two main issues: the definition of a selection probability for each unit in the population, as a function of the whole set of the auxiliary variables; and the determination of the sample size required to achieve a constrained precision level for each auxiliary variable. These precisions are usually expressed as a set of upper limits on the coefficients of variation of the estimates.

Define $p_i^j = x_{i,j} \Big/ \sum_{i=1}^{N} x_{i,j}, i = 1,\ldots,N, j = 1,\ldots,k$ and suppose that the k vectors of these size measures are available, one for each auxiliary, and that the k first order inclusion probabilities $\pi_i^j = np_i^j$, where n is the sample size. Without loss of generality, for every j we will assume that $0 \leq \pi_i^j \leq 1$ and that, for at least one j, $\pi_i^j > 0$ (otherwise, the unit i is outside the frame) and $\pi_i^j < 1$ (otherwise, the problem is trivial, because the unit i is surely included in the sample).

Deville and Tillé (1998) suggested some interesting solutions to the problem of selecting a sample by using a πps scheme; Chauvet and Tillé (2006) review the application of several πps algorithms. However, their focus was mainly on how to observe the defined probabilities and the performance of each selection procedure, measured with reference to relationships between one outcome variable and a unique covariate. These classical methods deal with the univariate case and cannot be easily extended to cover the case – which often arises in real circumstances, particularly in agricultural surveys – in which researchers must deal with a multipurpose survey and exploit multiple covariates in the sampling design, such as land use classes arising from a remotely sensed data classification.

Besides, the coefficient of variation constraints is related to the auxiliary variables **X**, rather than to the survey variables **Y**. The typical hypothesis underlying this assumption is that they can be considered equal for the purposes of determining the sample size required to reach a certain degree of precision. However, if there are considerable differences between the auxiliary variables and the survey variables, then the solution will be sub-optimal, because it is well-known that in practice, this hypothesis is only an approximation of the true situation, and that using the auxiliary variables to design the sample may underestimate the sample size required to reach a predetermined level of precision. An alternative sample size determination could use a model for an unknown **Y** in terms of the known **X**. Such models can be derived from past surveys, or by using remotely sensed data (see Section 3.5).

This approach should be followed if the two sets **X** and **Y** are correlated. No major problems arise when we deal with the design of a survey repeated in time, in which the two sets have the same size and definition but are recorded from different sources (surveys and remote sensing). This is the case, for example, of most business surveys carried out by National Statistical Institutes (NSIs) (Bee *et al.*, 2010; Hidiroglou and Srinath, 1993).

We propose a solution for a πps scheme that is capable of considering more auxiliaries in the sample selection process; we refer to this approach as a multivariate πps. As stated above, a general methodological framework to deal with this situation is absent from existing literature, although several practical efforts have been already made in this direction: in some NASS-USDA (National Agricultural Statistical Service – U.S. Department of Agriculture) surveys, it was suggested to use the maximum probability $\pi_i^0 = \max_{j=1,2,\ldots,k} \{\pi_i^j\}$ (Kott and Bailey, 2000; Bee *et al*., 2010). In a previous work, sound results in terms of Root Mean Square Error (RMSE) were obtained by simply defining the vectors of first order inclusion probabilities as the averages of these probabilities for each auxiliary variable $\pi_i^0 = \sum_{j=1}^{k} \pi_i^j \Big/ k$ (Bee *et al*., 2010).

An interesting approach is based on the use of a vector of selection probabilities that should limit the variation coefficients of the estimates of the totals of a given set of auxiliary variables. This outcome can also be achieved when carrying out the survey estimates by means of certain well-known estimators, such as calibration weighting (see Section 4). However, an optimal, or at least well-designed, sample selection of the statistical units should be considered as being complementary to an appropriate estimator, and certainly not as an alternative.

Moreover, it is important to mention that there are other ways to deal with multiple auxiliary variables in the sample selection procedure (Bee *et al*., 2010). In particular, the Cube method for balanced sampling (Chauvet, 2009; Chauvet and Tillé, 2006; Deville and Tillé, 2004; Tillé, 2011; Tillé and Favre, 2005; Falorsi and Righi, 2008), with constant or varying inclusion probabilities, can be used to select a sample that satisfies a given vector of selection probabilities, and that is, at the same time, balanced on a set of auxiliary variables. Therefore, the Horvitz-Thompson (HT) estimators of these variables are exactly equal to the known population totals, and thus have zero variance. Without the balancing constraint, this property can be satisfied by a *πps* selection procedure only if the vector of selection probabilities is strictly proportional to all auxiliary variables. This implies that they are linearly dependent.

Following these considerations, some recent studies focused on the computation of optimal inclusion probabilities for balanced sampling on given auxiliary variables (Tillé and Favre, 2005; Chauvet *et al*., 2011). The basis of this approach lies in the minimization of the residuals arising from a linear regression between a set of variables of interest and the balancing variables. Within this framework, procedures to compute the selection probabilities should not be considered an alternative to the Cube method, but rather, can be jointly used.

From the sampling theory, we know that a design is uniquely defined by the list of all the possible samples that can be selected from a population U and by the stochastic distribution $p(s)$, which assign a selection probability to each of them.

The selection of sampling units is almost always performed using some scheme that, by means of a proper use of some pseudo-random numbers routine, is proved to generate samples from $p(s)$.

When planning a specific sampling design, the random selection procedure should satisfy several requirements, among which the sampling error; that this should be as low as possible is not necessarily the most important concern. Often, organizational matters such as the availability of the frame or the cost of the data collection activity may play a significant role in the choices related to the sample selection. However, good survey organization stems from the ability to translate operational needs into methodologically correct decisions that will help planning a reasonable sampling scheme.

Instead, it is truly important to gain maximum knowledge of the structure of the population; the capacity of a sampling design to exploit the structure of U is always a desirable feature.

It is well-known that the adopted selection algorithm determines the statistical properties of the Horvitz-Thompson (HT) estimator, in particular its variance, known as the sampling error. What is less clear is that its impact is also substantial in the non-sampling errors, as the rate of non-response, the measurement errors, etc.

In agricultural surveys, statistical units do not necessarily have the same size; in particular, if we are dealing with legal bodies such as households or farms, they tend to have a highly skewed size distribution. The size measure involved should not be, inevitably, an economic or business indicator, as the same results will be obtained if we are dealing with the total surface of the farm, its arable land or the size of the livestock. Within spatial units, the use of points and regular polygons will ensure that this situation is avoided, but the widespread technique of partitioning the study area into irregular polygons will surely define a set of units, which are probably not skewed but, certainly, will not have the same size, although this requirement is considered essential during frame setup. Moreover, in multi-stage sampling, the aggregates that are used as primary sampling units are traditionally defined by administrative boundaries, and will surely not include the same number of secondary sampling units. Thus, if an equal probability sample is performed in the first stage, in the second stage it will be impossible for all the secondary sampling units to have the same selection probability.

Thus, the challenge lies in the fact that size has a considerable impact on the precision of survey estimates. Failure to notice that units should be selected using this attribute will probably introduce serious inefficiencies into the estimation of the population characteristics. On the other hand, when the distribution of the survey variable is concentrated in a few large units, an appropriate random selection plan exploiting this feature would provide investigators with a sample of smaller size and of higher efficiency (Holmberg, 2007; Falorsi and Righi, 2008).

It is widely recognized among statisticians that the evaluation of the sample size is a crucial element in the design of the sampling strategy – not only because of its impact on the overall cost of the survey and on the efficiency of producing reliable statistics for a given phenomenon, but also due to the typically predictive nature of the problem. Evaluation of sample size is one aspect of sample design, and a common intuitive ambiguity is that the sample size required is a function of the size of the target population. Indeed, we often enquire upon the sampling rate necessary to achieve a given precision of the estimates, because our expectation is that the sample for an entire country should be larger than that for a region. Generally, this is not true as in the simplest situation, we can consider it as inversely related to the variance, or standard error of survey estimates and, through the finite population correction, only slightly dependent on *N*.

A traditional approach to dealing with multivariate auxiliary variables in designing the sample is to employ a stratification scheme, such that the population units are classified in a stratum according to the values of their auxiliary variables (Kott and Bailey, 2000; Benedetti et al., 2008; Benedetti and Piersimoni, 2012; Vogel, 1995). Thus, a simple random sampling without replacement, or a *πps*, is selected within each stratum.

The most common solution suggested for decision-making problems regarding multiple criteria is not to summarize the multivariate nature of the problem through a simple statistic (average, maximum and so on), but to define a unique criterion, which combines the original functions by using a set of weights, which could exploit the relevance and uncertainty of each criterion.

Our problem can be formalized as follows. The aim is to find a vector π_i^0 of inclusion probabilities, which, through a convex linear combination, represents a mixture of the original univariate inclusion probabilities π_i^j:

$$\pi_i^0 = \sum_{j=1}^{k} \pi_i^j \alpha_j \tag{3.1}$$

with an unknown vector **α** of weights such that $\sum_{j=1}^{k} \alpha_j = 1$ and $\alpha_j \in [0;1]$, in such a way that $\sum_{i=1}^{N} \pi_i^0 = n$.

When the population is highly skewed, some of the inclusion probabilities π_i^j may be greater than one, because $n\left(x_{i,j} \middle/ \sum_{i=1}^{N} x_{i,j} \right)$ is an approximation of the inclusion probabilities. While, in the univariate case, these units are usually selected with certainty, and revised selection probabilities are calculated by excluding these units, in this multivariate framework, only the units with a π_i^0 greater than one should certainly be selected. As a consequence, the inclusion probabilities, and their weights α_j, must be calculated iteratively, by excluding this kind of unit from each iteration, until all probabilities are lower than 1.

Several sample selection algorithms have been proposed in the literature (for a review, see Brewer and Hanif, 1983; Tillé, 2006). Among these, a very simple way to draw a without-replacement *πps* sample is the use of the Poisson sampling design. For every unit i in the frame, an independent random number r_i is generated, uniformly distributed in the interval [0,1]. Each unit of the population is thus included in the sample if $r_i \leq n\, p_i$.

The major limitation of this selection procedure is that the sample size thus obtained is a random variable $n(s)$ with possible values 0, 1, ..., N, and is a sum of independent Bernoulli trials, which is usually called a Poisson-Binomial distribution (Tillé, 2006). As this distribution can approximately be considered as a Poisson with variance n (Ohlsson, 1998), the deviations from the desired size n may be considerable. Thus, it is often important to use its counterpart, which yields a fixed sample size n: the sequential Poisson sample selection procedure (Ohlsson, 1998). The adjustment in the selection criterion consists in selecting the n units having the smallest r_i / p_i ratio.

While the Poisson design yields exact prescribed inclusion probabilities, the sequential Poisson design does not. As a result, although it is asymptotically unbiased, the HT estimator is slightly biased for the sequential Poisson design. Poisson sampling and sequential Poisson sampling offer two easy ways – random and fixed sample size, respectively – of drawing probability proportional to size samples from a finite population.

The extension of the results obtained to other *πps* designs will depend only on the possibility of inverting the variance of the estimator arising from these designs, to derive the requested sample size for a fixed precision of the estimates. This could be even more necessary in cases using the sequential Poisson sampling. Indeed, there are many other *πps* designs that, although more difficult to implement, could be a better choice, since they have a fixed sample size and give exact inclusion probabilities (Tillé, 2006).

As usual, the variances of the HT estimators arising from these two sampling designs will be a function of the sample size n, of the population variance σ_x^2 of a variable X, of its total t_x , and of the inclusion probabilities π_i^0 and, as consequence, of the set of weights of the linear convex combination (3.1).

The variances of the HT estimators of these two sampling designs are reported below. For Poisson sampling, it is (Särndal *et al.*, 1992, p. 86):

$$V_{Poisson}\left(\hat{t}_{X_j}\right)=\frac{1}{n_j}\sum_{i=1}^{N}\left(1-n_j p_i^0\right)\left(\frac{x_{i,j}}{p_i^0}\right)^2 p_i^0,$$

where n_j is the sample size necessary to achieve a given precision of the j-th variable. On the other hand, for sequential Poisson sampling, it is (Ohlsson, 1998):

$$V_{SPS}\left(\hat{t}_{X_j}\right)=\frac{1}{n_j}\frac{N}{N-1}\sum_{i=1}^{N}\left(1-n_j p_i^0\right)\left(\frac{x_{i,j}}{p_i^0}-t_{X_j}\right)^2 p_i^0,$$

in which the correction factor $N/(N-1)$ can be usually considered as approximately equal to 1. Thus, the sequential Poisson sampling presents the same approximate variance of the ratio model that is used to adjust the HT estimator of a Poisson Sampling, to take into account the potential difference between the planned and observed numbers of sampling units (Särndal *et al.*, 1992, p. 248).

Inverting the constraint $CV_j=\frac{\sqrt{Var\left(\hat{t}_{X_j}\right)}}{t_{X_j}}\leq c_j$ to derive the sample size for fixed sampling errors c_j for each variable j, when Poisson sampling is used, we obtain:

$$n_j=\frac{\sum_{i=1}^{N}\frac{x_{i,j}^2}{p_i^0}}{c_j^2 t_{X_j}^2+\sum_{i=1}^{N}x_{i,j}^2}, \tag{3.2}$$

while for sequential Poisson sampling, we have:

$$n_j=\frac{\sum_{i=1}^{N}\left(\frac{x_{i,j}}{p_i^0}-t_{X_j}\right)^2 p_i^0}{c_j^2 t_{X_j}^2+\sum_{i=1}^{N}\left(\frac{x_{i,j}}{p_i^0}-t_{X_j}\right)^2\left(p_i^0\right)^2} \tag{3.3}$$

Where, for Poisson sampling, n is the expected sample size.

In (3.2) and (3.3), parameter c is chosen *a priori* by researchers, and the X_j are assumed to be known for the whole population. Thus, if we denote with $\boldsymbol{\alpha}=\{\alpha_1,\alpha_2,\ldots,\alpha_k\}$ the generic vector of the set of all possible weighting systems, it is possible to conclude that n_j is a function only of the unknown $\boldsymbol{\alpha}$: $n_j=f_j(\boldsymbol{\alpha})$.

At this point, it is clear that the problem lies in identifying the vector $\boldsymbol{\alpha}$ that can minimize (3.2) or (3.3), given the desired level of precision c_j. It is noteworthy that these constraints are assumed to be fixed at the same level that the πps sampling is performed. In other words, under the adopted framework, it is not possible to treat higher- or lower-level constraints.

In particular, since it is sought to estimate the totals for each of the J variables by means of the same number of sampling units, the optimal sample size can be defined as follows:

$$\bar{n} = \min_{\boldsymbol{\alpha}}\left\{ \max_{j=1,2,\ldots,k} \left\{ n_j = f_j\left(\boldsymbol{\alpha}\right) \right\} \right\}. \tag{3.4}$$

The term $\max\limits_{j=1,2,\ldots,k}\{\boldsymbol{n}_j\}$ in (3.4) means that the optimization concerns the largest of the sample size n_j corresponding to each auxiliary variable j. This is a classical cautionary solution adopted in sample design.

To solve the minimax problem (3.4), it is necessary to analyse the features of the function $f_j(\boldsymbol{\alpha})$. In particular, the following must be introduced:

Assumption 1: Every $\mathrm{f_j}(\boldsymbol{\alpha})$ is a strictly monotone decreasing function of α_j, i.e. the partial derivatives are always lower than zero: $\dfrac{\partial f_j\left(\boldsymbol{\alpha}\right)}{\partial \alpha_j} < 0$.

The above assumption means that the sample size required to reach a given precision on a variable j is expected to decrease, when the inclusion probability π_i^0 approaches proportionality to X_j, since the weight α_j increases. This requirement appears rather reasonable, and from an empirical point of view it is usually verified in practical survey designs on real populations.

This feature of the function $f_j(\boldsymbol{\alpha})$ plays an important part in solving (3.4). For $k = 2$, this monotonic behaviour implies that the two functions $f_1(\alpha_1)$ and $f_2(\alpha_1)$ are strictly monotone decreasing and increasing, respectively, in the range $\alpha_1 \in [0;1]$, with extreme values $f_1(1)=0$, $f_1(0)>0$, $f_2(1)>0$, $f_2(0)=0$. This is because of the exact proportionality of the inclusion probabilities to X_1 or to X_2 respectively, when $\alpha_1=1$ or $\alpha_1=0$. Due to these distinctive characteristics, and considering the two functions $f_1(\alpha_1)$ and $f_1(\alpha_1)$ as continuous, there is, at least for k=2, certainly one – and only one – intersection between the two functions, e.g. $\overline{\alpha}_1$, and when $\alpha_1 < \overline{\alpha}_1$ the maximum of the two functions is represented by $f_1(\alpha_1)$, while when $\alpha_1 > \overline{\alpha}_1$ the maximum of the two functions is represented by $f_2(\alpha_1)$. It noteworthy noticing that, in this situation, the maximum of the two functions is minimized when $\alpha_1 = \overline{\alpha}_1$. In other words, under *Assumption* 1, for k=2, the minimax problem (3.4) consists in finding the value of α_1 such that n_1=n_2, thus solving a system of two nonlinear equations in two variables (α_1 and n), the solution to which exists in the admissible range of **α**, and is unique.

In *k* dimensions, we can thus define the following system of *k* non-linear equations in *k* unknowns, i.e. *k*-1 weights (their sum is fixed), and *n*:

$$F(\boldsymbol{\alpha}) = \begin{cases} f_1(\boldsymbol{\alpha}) - n = 0 \\ \quad \cdots\cdots\cdots \\ f_j(\boldsymbol{\alpha}) - n = 0 \\ \quad \cdots\cdots\cdots \\ f_k(\boldsymbol{\alpha}) - n = 0 \end{cases} \tag{3.5}$$

There is no formal proof that the uniqueness of the solution found in two dimensions will also hold for $k > 2$. However, in all practical applications of this method, it was possible to find a unique solution in the admissible range of **α**. Thus, to extend the results to a general multivariate *k*-dimensional space, we should suppose that:

Assumption 2: Under **Assumption 1**, the minimax problem (3.4) is reduced to find the vector of coefficients $\overline{\alpha}$, such that $n_1 = n_2 = \ldots\ldots = n_k$, by solving the system of non-linear equations (3.5) whose solution in the admissible range of α exists, and is unique.

Under **Assumption 2**, the vector of coefficient **α** can be obtained by finding a solution to the system (3.5); this problem can be solved numerically, with the classical Newton algorithm:

$$s^{h+1} = s^h - \left[J\left(s^h\right)\right]^{-1} F\left(s^h\right), \tag{3.6}$$

where, for each iteration *h*, s^h is a vector of the values of the *k* variables $\{\alpha_1, \alpha_2, .., \alpha_{k-1}, n\}$, $F(s^h)$ is a vector of the values of the *k* equations, and $J(s^h)$ is the Jacobian of the first-order partial derivatives.

However, this well-known algorithm has proved to be unsuitable for this application, because it has two drawbacks. The first is that it requires the evaluation of the Jacobian, which is not simple, from an analytical point of view, as it must be derived for (3.2) and (3.3). However, this limitation can be overcome by using numerical evaluations of the partial derivatives. The second, much more serious limitation, is that in many cases, it proposes solutions that are not in the admissible range $\alpha_j \in [0;1]$ for some j. Thus, the Newton algorithm should be modified to introduce the necessary constraints, which will ensure its observance of the requirement $\alpha_j \in [0;1]$.

To find a solution, we propose an alternative algorithm, that is much simpler, since it does not require the first-order derivatives to be evaluated. It usually requires a few more iterations of the Newton algorithm, but with less computational burdens within each iteration and, more importantly, always within the constraint $\alpha_j \in [0;1]$.

Given a starting value $\alpha_{0,j}$, within each iteration *h*, the algorithm iteratively updates the coefficients according to the rule:

$$\alpha_{h+1,j} = \frac{\alpha_{h,j} \frac{n_{h,j}}{\sum_{j=1}^{k} n_{h,j}}}{\sum_{j=1}^{k} \alpha_{h,j} \frac{n_{h,j}}{\sum_{j=1}^{k} n_{h,j}}}, \tag{3.7}$$

where $n_{h,j}$ is evaluated as a function of $\alpha_{h,j}$ by using (3.1) or (3.2). This rule will generate a sequence of $n_{h,j}$ for each variable whose distance from the optimal sample size $\bar{n}_j$ is necessarily decreasing. This procedure is formulated such that in each iteration *h*, all variables j with a sample size $n_{h,j}$ lower than its average among the other variables, will receive a lower weight $\alpha_{h+1,j} < \alpha_{h,j}$, which will necessarily increase the sample size $n_{h+1,j} > n_{h,j}$ (due to the monotone functions in **Assumption 1**).

Thus, being of a descending type, and given the existence and uniqueness of the solution stated by **Assumption 2**, it will surely converge to the solution of System (3.5).

Concerning the convergence criterion, there is no notable gain in the algorithm speed, whether due to changes in the vector of weights, in the vector of sample sizes or in the multivariate inclusion probabilities between the algorithm's various iterations. We decide to use the most restrictive of these suitable criteria: the maximum absolute variation in the multivariate inclusion probabilities fixed at 1×10^{-10}.

Good results in terms of speed of convergence, and accuracy of the proposed solution, have been obtained by using, as starting value for the vector $\boldsymbol{\alpha}$:

$$\alpha_{0,j} = \frac{\frac{1}{c_j}}{\sum_{j=1}^{k} \frac{1}{c_j}}. \tag{3.8}$$

Although there is no proof of the algorithm's convergence, in empirical applications on real data, the convergence was always reached in a very small number of iterations, usually less than 10 and never more than 20, for a number of auxiliaries ranging between 4 and 15.

3.3 Optimal stratification

A traditional approach to dealing with multivariate auxiliary variables when designing the sample is to employ a stratification scheme, such that the population units are classified in a stratum according to the values of their auxiliary variables (Benedetti *et al.*, 2008; Vogel, 1995). Thus a simple random sampling without replacement or a πps is selected within each stratum.

In many agricultural surveys, the main use of X consists of actions that are not related to the sample design, but are performed after the sample selection. The most common context for the production of sample estimates consists in a standard design; only after the data collection and editing phase, are the auxiliary information used. It is at this stage that NSIs make the greatest efforts in using and developing highly complex estimators that can lead to greater efficiency (see Section 6). In sample design, the common procedure is to perform stratification by size, obtained by defining a set of threshold levels for each auxiliary variable included in the sampling frame. After the survey has been performed, the initial direct estimates are corrected through the use of calibration estimators (see Section 4), in which the external consistency constraints to known totals are assigned, usually with reference to a previous Census.

Most of the literature on optimal stratification relies on the early works of Dalenius and Hodges, written in the 1950s (Dalenius and Hodges, 1959; see Horgan, 2006 for a review). The solutions they propose, usually based on linear programming, are still widely popular in applied survey sampling (Khan et al., 2008). This strategy can be addressed by the introduction of a take-all (censused) stratum and of one or more take-some (sampled) strata. NSIs commonly use this procedure to select samples, although it is not easy to provide a unique definition of the boundaries of such strata when they must be based on a multivariate set of size measures.

This approach is not new, and has been widely employed by survey practitioners, often using a heuristic rule to determine the part of the population to be censused (for example, households having over ten hectares). This procedure, typically undertaken pursuant to the desire to match administrative criteria, usually ignores the statistical implications for estimate precision.

The univariate methodological framework for this problem was suggested by Hidiroglou (1986), who proposed an algorithm for the determination of the optimal boundary between the two strata: census and sample. The literature has proposed several formal extensions to the univariate optimal determination of the boundaries between more than two strata (Kozak, 2004; Horgan, 2006; Verma and Rizvi, 2007), through the use of algorithms that usually derive, simultaneously, the sample size required to guarantee a given accuracy level for the resulting estimates, and the sample allocation to the strata. A generalization of these algorithms, extended in Baillargeon and Rivest (2009 and 2011), is used when the survey variable and the stratification variable differ. However, these classical methods only deal with the univariate case, and cannot be extended easily when there are multiple covariates for stratification (Briggs *et al.*, 2000).

This approach produces optimal stratification, in the sense of a minimum variance for the stratified estimator of the population mean, under the assumption that the (univariate) character of interest **Y** is known for each population unit. Since **Y** is unknown before sampling, it is suggested to adopt a linearly approximated solution, based on a *highly correlated* auxiliary variable (or set of variables) **X**, known for the entire population. The optimality properties of these methods rely on distributional assumptions regarding the target variable **Y** in the population, the assumed linear relationship between **Y** and **X**, the type of allocation of sample units to the strata, and the sampling design within strata (typically, simple random sampling).

In this context, stratification trees (Benedetti *et al.*, 2008) display several advantages over classical univariate Dalenius-type methods. First, stratification trees do not require either distributional assumptions on the target variable, or any hypotheses on the functional form of the relationship between this variable and the covariates. Moreover, when many auxiliary variables are available, the stratification tree algorithm can automatically select the most powerful variables for the construction of strata. Identified strata are easier to interpret than those based on linear methods. Finally, they do not require any particular sample allocations to the strata, as it simultaneously allocates the sampling units, using the Bethel or the Cromy algorithm in each iteration (Benedetti *et al.*, 2008).

However, such an approach is equivalent to partitioning the population into strata that have *box-shaped* boundaries, or that are approximated through the union of several such boxes. This constraint prevents irregularly-shaped strata boundaries to be identified, unless a grid constituted by several rectangles of different sizes are used to approximate the solution required.

Optimal data partitioning is a classical problem in statistical literature, following

Fisher's early work on linear discriminant analysis. However, our problem is more directly related to the use of unsupervised classification methods to cluster a set of units (in this case, a population frame). The main difference between the two problems lies in the fact that the underlying objective functions are different: in sampling design, the aim is usually to minimize the sample size; in clustering, it is a common practice to minimize the variance within a cluster. There is an intuitive connection between these two concepts, although the definition of sample size depends not only on the variance within each stratum, but also on other parameters (sample size, population size, unknown total, etc.).

Assume that the population U can be partitioned into H groups according to a given criterion, and let $\{U_1,U_2,\ldots,U_h,\ldots,U_H\}$ be this set of groups such that $\bigcup_{h=1}^{H} U_h = U$ and $U_h \bigcap U_r = \varnothing, \forall h \neq r$; in other words, that this set of groups called strata be exhaustive and non-overlapping. Let $\theta \in \Theta = \{1,\ldots,H\}^N$ be the unknown stratification codes, and $\{N_1, N_2,\ldots, N_h,\ldots, N_H\}$ the number of units of the population belonging to each stratum, evidently $\sum_{h=1}^{H} N_h = N$. Such a partition implies that certain basic choices have been or should be established, i.e.:

1. the selection of the set of the stratifying covariates (usually, the crop area derived from a land use map, obtained by classifying remotely sensed data);
2. how to determine the required number of strata H;
3. the definition of the criterion adopted to stratify the population. If discrete auxiliaries are used, it is necessary to define the corresponding list of codes or code combinations to be used, while for continuous variables, strata borders or limits should be carefully evaluated;
4. how to allocate the sample units to the strata.

In a sampling strategy, the population is stratified for three main reasons: administrative purposes, definition of the planned domains of analysis, and improvements in the efficiency of the estimates.

Also, it is important to consider that stratification on many auxiliaries often risks being carried out through an excessively-fine partition that is not always applicable in practice (Benedetti *et al.*, 2008).

Before describing the third topic, it is noted that the set of stratifying covariates is usually selected to meet the requirements of the first two topics. The first issue essentially concerns the organization of the data collection process, and legal and administrative aspects, such as the availability of the frame at the local level exclusively. Besides, the second aspect concerns the dissemination of the survey's data. In particular, this is connected to the estimation of unplanned domains that usually creates several difficulties due to the fact that sample size within each domain is not defined. The best way to avoid these difficulties is to establish the sample size in each estimation domain, by introducing an auxiliary variable into the set of stratifying covariates. The codes of this auxiliary variable identify the estimation domains such that it becomes planned, even though it may cause too fine a partition.

The third and fourth topic can be solved simultaneously by using an optimizing approach, which does not assume a given shape for the optimal partition, such as Simulated Annealing (SA). This is a stochastic optimization method for finding a function's global minimum (Kirkpatrick *et al.*, 1983). The method is a generalization of the Metropolis-Hastings algorithm (Metropolis *et al.*, 1953), and is one of the most popular optimization strategies for solving complex combinatorial problems.

Let $\theta \in \Theta = \{1, \ldots, H\}^N$ be a vector of size *N* whose *i*-th element may assume *H* possible values: the codes of the stratifying variable. In the approach advanced by Geman and Geman (1984), the optimization problem can be viewed as a stochastic process described through a family of distributions:

$$\pi_T(\theta) = \frac{\exp\{-f(\theta)/T\}}{\sum_{\theta\in\Theta} \exp\{-f(\theta)/T\}}, \tag{3.9}$$

where f(*θ*) is the energy function, T is a positive parameter called temperature, and the denominator of Equation (3.9) is called the model's normalization constant. It can be shown (Brémaund, 1999) that $\lim_{T\to 0} \pi_T(\theta) = \pi_{\lim}(\theta)$, where $\pi_{\lim}(\theta)$ takes the same positive value at any configuration *θ* corresponding to a global minimum of the energy function, and $\pi_{\lim}(\theta) = 0$ otherwise.

The SA algorithm is an iterative random search procedure, where at each step a non-homogeneous Markov-Chain with a reduced value for temperature *T* is generated. Specifically, for *k* auxiliary variables, the energy function for each iteration *h* and for each visited unit *i* can be defined as $f(\theta_{h,i}) = \max(n_{x_1,h,i}, n_{x_2,h,i}, \ldots, n_{x_k,h,i})$, where each sample size is evaluated using the Bethel (1989) multivariate allocation algorithm under configuration *θ*. Given a configuration $\theta_{h,i}$ at the *h*-th iteration, another configuration, say $\theta_{h+1,i}$, is then chosen on the basis of a visiting schedule for each unit *i*. Here, we visit units sequentially, following an initial random ordering of the list frame. The new configuration of the *i*-th element of the vector *θ*, following such a visit, is defined by first exchanging the code $\theta_{h,i}$ of the unit *i* with an alternative code selected at random from the H-1 remaining codes, and then accepting this change if it leads to a reduction of the energy function defined by the sample size.

For any suitable choice of stopping criterion, a final configuration is therefore obtained. Generally, this corresponds to a local minimum. To avoid convergence to local minima, a stochastic decision rule is used, which allocates a positive probability to the exchange of configuration even when an increase in the energy function is obtained. In particular, the algorithm replaces the solution obtained at the h-th iteration $\theta_{h,i}$ with a new solution $\theta_{h+1,i}$, according to an acceptance rule known as the Metropolis criterion, which enables hill-climbing of the objective function.

More formally, the algorithm can be summarized as follows. The procedure starts at the first iteration h=1, with an initial value T_1, and randomly selects the initial configuration θ_1 from $\{1,\ldots,H\}^N$. At step h, the elements of $\theta_{h,i}$ are updated as follows:

- a unit in the population is selected in accordance with the visiting schedule, and its status is exchanged (from $\theta_{h,i}$ to any other possible stratum code);
- the new proposed vector of codes is denoted with $\theta_h{}^*$; and $f(\theta_h^*)$ denotes the sample size, evaluated as the Bethel solution with k auxiliary variables. Whether or not to adopt θ_h^* according to a Boltzmann distribution is randomly determined:

$$\theta_h = \begin{cases} \theta_h^* \text{ with probability } p = \min\left(1, \exp\left\{\dfrac{[f(\theta_h) - f(\theta_h^*)]}{T_h}\right\}\right) \\ \theta_h \text{ otherwise} \end{cases} \tag{3.10}$$

- these steps must be repeated, say m times, for all the units of the population, and the temperature is updated according to a simple rule: $T_{h+1} = \rho T_h = \rho^h T_1$, with $0<\rho<1$ representing a control parameter on the algorithm's cooling speed. Values close to 1 increase the number of iterations required to reach the optimum, but also prevent convergence to local minima;
- the procedure must be stopped when $\left|f(\theta_h) - f(\theta_{h+1})\right| / f(\theta_h) \leq \varepsilon$, or when the number of iterations exceeds a fixed maximum.

The algorithm's most serious drawback is its computational burden. Partitioning populations with a huge number of units may not be feasible, since the procedure is applied to all units in the population. While this may not be a major drawback for relatively small populations in the case of two strata, it could be a serious shortcoming for very large populations and, potentially, a severe drawback in the case of over two strata. In the latter case, it would be necessary to determine optimal multivariate sample allocations for each unit in the population at each iteration, leading to a huge computational burden.

A way to accelerate the annealing process is to use a genetic algorithm (Ballin and Barcaroli, 2013; Barcaroli, 2014) or Generalized Simulated Annealing (GSA; Tsallis and Stariolo, 1996), in which the updating criterion (3.10) is again random, but with an acceptance probability given by:

$$\theta_h = \begin{cases} \theta_h^* \text{ with } p = \min\left(1, \dfrac{1}{\left[1 + (q_A - 1)\left(\dfrac{f(\theta_h^*) - f(\theta_h)}{T_h}\right)\right]^{\frac{1}{q_A - 1}}}\right), \\ \theta_h \text{ otherwise} \end{cases} \tag{3.11}$$

and with a temperature that decreases according to the rule:

$$T_h = T_1 \frac{2^{q_V - 1} - 1}{(1 + h)^{q_V - 1} - 1} \tag{3.12}$$

The two constants q_A and q_V regulate the acceptance probability distribution and the rate of temperature decrease. When $q_A = q_V = 1$, distribution (3.11) corresponds to (3.10), and the temperature decreases logarithmically with the increase of the iteration number, while when q_A =2 and $q_V = 1$ the algorithm is equivalent to the so-called Fast Simulated Annealing or Cauchy Machine (Tsallis and Stariolo, 1996).

If the size N of the list frame is such that even the GSA is unfeasible, then it is possible to further accelerate the convergence of the annealing process using Iterated Conditional Modes (ICM; Besag, 1986). This replaces the stochastic rule (3.10) by the deterministic rule:

$$\theta_h = \begin{cases} \theta_h^* & \text{if } f(\theta_h) > f(\theta_h^*) \\ \theta_h & \text{otherwise.} \end{cases} \tag{3.13}$$

Note that this rule will lead to convergence to a local, but not necessarily a global, minimum. However, this theoretical disadvantage could be compensated by a dramatic decrease in the computational burden. In our empirical experience, ICM has always prompted solutions that are rarely equal, but usually very close, to the global optimum, and with a number of iterations that is negligible compared to that required by random search procedures as SA and GSA.

3.4 Spatially balanced samples

In recent decades, the spatial balancing of samples has become so peculiar that several researchers and survey practitioners have suggested sampling algorithms to achieve it (Wang et al., 2012). Surprisingly, this balancing is mainly based on intuitive considerations, and it is not so clear when and to what extent it could have an impact on estimate efficiency. Moreover, it is also useful to note that this feature was not defined with sufficient adequacy; as a consequence, a range of possible interpretations exists that makes it unfeasible to perform any comparison between different methods, only because they are most likely intended to obtain selected samples with different formal requirements.

In design-based sampling theory, if it is assumed that there is no measurement error, the potential observations concerning each unit of the population cannot be considered as dependent. However, an inherent and fully recognized feature of spatial data is that it is dependent; this much was succinctly expressed in Tobler's *First Law of Geography*, according to which *everything is related to everything else, but near things are more related than distant things*. Thus, it is clear that sampling schemes for spatial units can be given reasonable treatment by introducing a suitable model of spatial dependence into a model-based or at least a model-assisted framework. In literature (Benedetti and Palma, 1995; Dunn and Harrison, 1993; Rogerson and Delmelle, 2004), this approach proved to be helpful in rationalizing the intuitive procedure to spread the selected units over the space, because closer observations will provide overlapping information as an immediate consequence of dependence. However, on this assumption, the concern is, necessarily, how to find the sample configuration that can best represent the entire population; this leads us to define our selection as a problem of combinatorial optimization. Indeed, if the sample size is fixed, the aim is to minimize an objective function defined over the whole set of possible samples, which is a measure of the loss of information due to dependence.

An optimal sample selected with certainty is of course not acceptable, if we assume the randomization hypothesis that is the background for design-based inference. Thus, it is necessary to depart from the concept of dependence in favour of that of spatial homogeneity, measured in terms of the local variance of the observable variables, where all the units of the population within a given distance could be defined as local units.

An intuitive method of producing samples that are well spread across the population, and that is widely used by practitioners, is to stratify the population units on the basis of their location. The problems arising from the adoption of this strategy lie in the evidence that it fails to have direct and substantial impact on the second-order inclusion probabilities, certainly not within a given stratum; and that the way to obtain a good partition of the study area is often unclear. These drawbacks are related, to a certain extent; for this reason, they are usually approached in combination by defining a maximal stratification, i.e. partitioning the study in as many strata as possible, and selecting one or two units per stratum. However, this straightforward and fast scheme to ensure that the sample is well-spread across the population, is somewhat arbitrary, because it greatly depends upon the stratification criterion, which should be general and efficient.

The basic principle is to extend the use of systematic sampling to two or more dimensions, a notion on which the Generalized Random Tessellation Stratified (GRTS) design is based (Stevens and Olsen 2004): to select the units systematically, the design maps the two-dimensional population into one dimension, while attempting to preserve some multi-dimensional order.

This approach is essentially based on the use of Voronoi polygons, which are used to define an index of *spatial balance*.

Let denote with S the set of all possible random samples of fixed size n, which can be selected from U, where its generic element is $s = \{s_1, s_2, \ldots, s_N\}$, and s_i is equal to 1 if the unit with label i is in the sample, and 0 otherwise. For any unit i, let π_i $(i \in U)$ be the first-order inclusion probability, and for any couple $\{i,j\}$, let $\pi_{i,j}$ $(i,j \in U)$ be the second order-inclusion probability.

For a generic sample s, the Voronoi polygon for the sample unit s_i=1 includes all population units closer to s_i than to any other sample unit s_j=1. If we let v_i be the sum of the inclusion probabilities of all units in the *i-th* Voronoi polygon, for any sample unit u_i, we have E(v_i)=1; and for a spatially balanced sample, all v_i should be close to 1. Thus, the index V(v_i) (i.e. the variance of the v_i) can be used as a measure of spatial balance for a sample.

Note that this concept is rather distant from that of balanced sampling, introduced in model-based sampling (Deville and Tillé, 2004) and that is reasonably accepted even in the design-based approach, through the introduction of the cube method (Chauvet and Tillé 2006), as a restriction of the support S of samples that can be selected by imposing a set of linear constraints on the covariates. These restrictions represent the intuitive requirement that the sample estimates of the total, or of the average, of a covariate should be equal to the known parameter of the population. In a spatial context, this plan could be applied by imposing the requirement that any selected sample should respect the first p moments for each coordinate, assuming implicitly that the survey variable y follows a polynomial spatial trend of order p (Breidt and Chauvet, 2012).

However, these selection strategies do not use the concept of distance, which is a basic tool to describe the spatial distribution of the sample units, which leads to the intuitive criterion that units that are close, seldom appear simultaneously in the sample. This condition can be considered reasonable under the assumption that, if the distance between two units i and j increases, the difference |yi - yj| between the values of the survey variable always increases. In these situations it is clear that the HT variance of estimates will necessarily decrease, if high joint inclusion probabilities are assigned to couples with very different y values due to their great distance from one other, to the detriment of couples that are expected to have similar y values due to their closeness.

In this connection, Arbia (1993) – inspired by purely model-based assumptions on the dependence of the stochastic process generating the data, according to the algorithm typologies identified by Tillé (2006) – suggested a draw-by-draw scheme: the dependent areal units sequential technique (DUST), which began with a unit selected at random, say i, and in any step $t<n$ updates the selection probabilities according to the rule $\pi_j^{(t)} = \pi_j^{(t-1)}\left(1 - e^{-\lambda d_{i,j}}\right)$, where λ is a tuning parameter that is useful in controlling the sample's distribution over the region under study. This algorithm, or at least the sampling design that it implies, can be interpreted and analysed easily in a design-based perspective, especially referring to a careful estimation and analysis of its first- and second-order inclusion probabilities.

Recently, some advances have been proposed for list sequential algorithms whose updating rules present the crucial property of preserving the fixed first-order inclusion probabilities (Grafström, 2012; Grafström *et al.*, 2012; Grafström and Tillé, 2013). In particular, Grafström (2012) suggested a list sequential algorithm that, for any unit i, in any step t, updates the inclusion probabilities according to a $\pi_i^{(t)}= \pi_i^{(t-1)}-w_t^{(i)}(\mathrm{I}_t - \pi_t^{(t-1)})$ rule, where $w_t^{(i)}$ are the weights given by unit t to the units $i=t+1$, $t+2,\ldots, N$ and I_t is an indicator function set equal to 1, if the unit t is included in the sample and otherwise equal to 0.

The weight determines how the inclusion probabilities for the unit i should be affected by the outcome of unit t. These are defined such that, if they satisfy upper and lower bounds, the initial π_i are not modified. The suggested *maximal weights* criterion places as much weight as possible upon the closest unit, then to the second closest unit, and so on.

Two alternative procedures to select samples having a fixed π_i and correlated inclusion probabilities were derived (Grafström *et al.*, 2012), as an extension of the Pivotal method introduced to select πps (Deville and Tillé, 1998). These are essentially based on an updating rule of the π_i and π_j probabilities, that should, at each step, locally keep the sum of the updated probabilities as constant as possible, and differ from each other so that the two nearby units *i* and *j* can be chosen. These two methods are known as the Local Pivotal Method 1 (LPM1), which, according to the authors' suggestion, is the most balanced of the two; and the Local Pivotal Method 2 (LPM2), which is simpler and faster.

To understand the circumstances in which it is efficient to spread the selected units over the population, it must be supposed that the distance matrix summarizes all the features of the spatial distribution of the population and, as a consequence, of the sample. Within a model-based perspective, this general hypothesis is equivalent to assuming that the data-generating process is stationary and isotropic (i.e. its distribution does not change the coordinates' space is we shifted or rotated). Focusing on set **C**, without using any other information from **X**, this assumption implies that the problem in selecting *spatially balanced samples* lies in the definition of a design $p(s)$, with a probability proportional to some synthetic index $M(d_s)$ of the within sample distance matrix d_s, when it is observed within each possible sample *s* by using an MCMC algorithm to select the sample (Traat *et al.*, 2004).

There are several circumstances in which due care should be taken when selecting samples, so that they are spatially well-distributed:

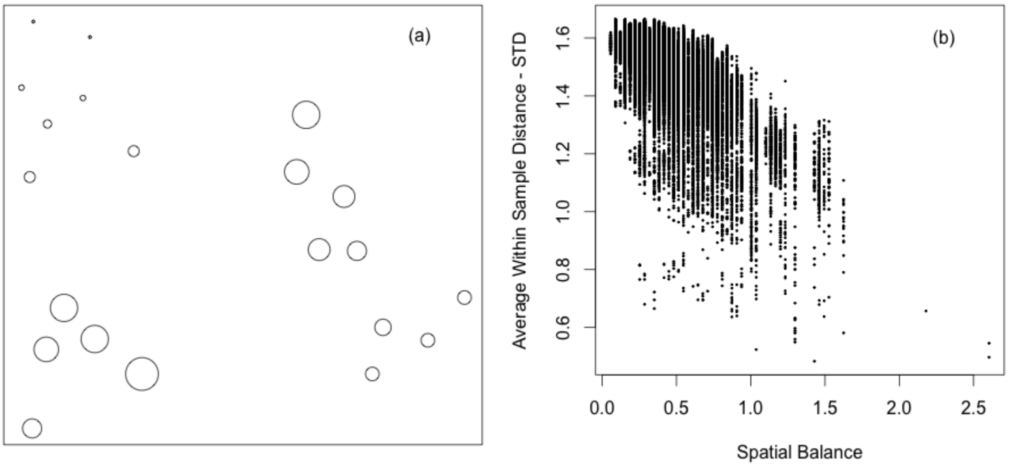
1. *y* has a linear or monotone spatial trend;
2. there is spatial autocorrelation (i.e. the data relating to close units is more similar than those relating to distant units);
3. *y* appears to follow zones of local stationarity of the mean and/or of the variance; in other words, the observed phenomenon displays spatial stratification;
4. the population's units present a spatial pattern, which can be clustered; in other words, the units' intensity varies across the region under study.

It is noteworthy that, while the distance between a pair is a basic concept in all these features of the phenomenon, the index $V(v_i)$ of *spatial balance* appears to be directly related to the third aspect, but only indirectly with the other three. This consideration and the practical impossibility of using the index $V(v_i)$, because it involves the π_i, prompt the use of a rule that sets the probability $p(s)$ of selecting a sample *s* proportionally, or more than proportionally, to a synthesis of the distance matrix within the sample d_s. Below is an example that can clarify this argument.

Example: A small population of size $N = 21$ is used (Grafström *et al.*, 2012; see Figure 3.1). The population has three natural strata, and it is used to verify whether the index of spatial balance $\mathrm{V}(v_i \in s)$, evaluated in all the possible samples of size n=6, is related to the design's efficiency and to an alternative spatial balance index, defined as the average of the distance matrix (3.15) within the sample d_s.

The correlation coefficient between the two indexes is equal to -0.504, but its value is influenced by the evidence that the π_i are assumed to be constant, to evaluate $\mathrm{V}(v_i)$. On the other hand, they are variable if we use a design with a $p(s)$ proportional to the average of d_s. To obtain a similar design, but with constant first-order inclusion probabilities, it is enough to standardize the distance matrix such that its row and column totals are constant. After this standardization, the correlation coefficient becomes -0.715, thus increasing our confidence that the use of a summary index of the distance matrix, in the sample design, will generate samples that are spatially well-balanced according to the index $\mathrm{V}(v_i \in s)$.

Figure 3.1: (a) Spatial distribution of the population of size N=21; the radius of each circle is proportional to the target variable y. (b) Scatterplot of the spatial balance index and of the average of the standardized distance matrix d_s, within any possible sample of size n=6.

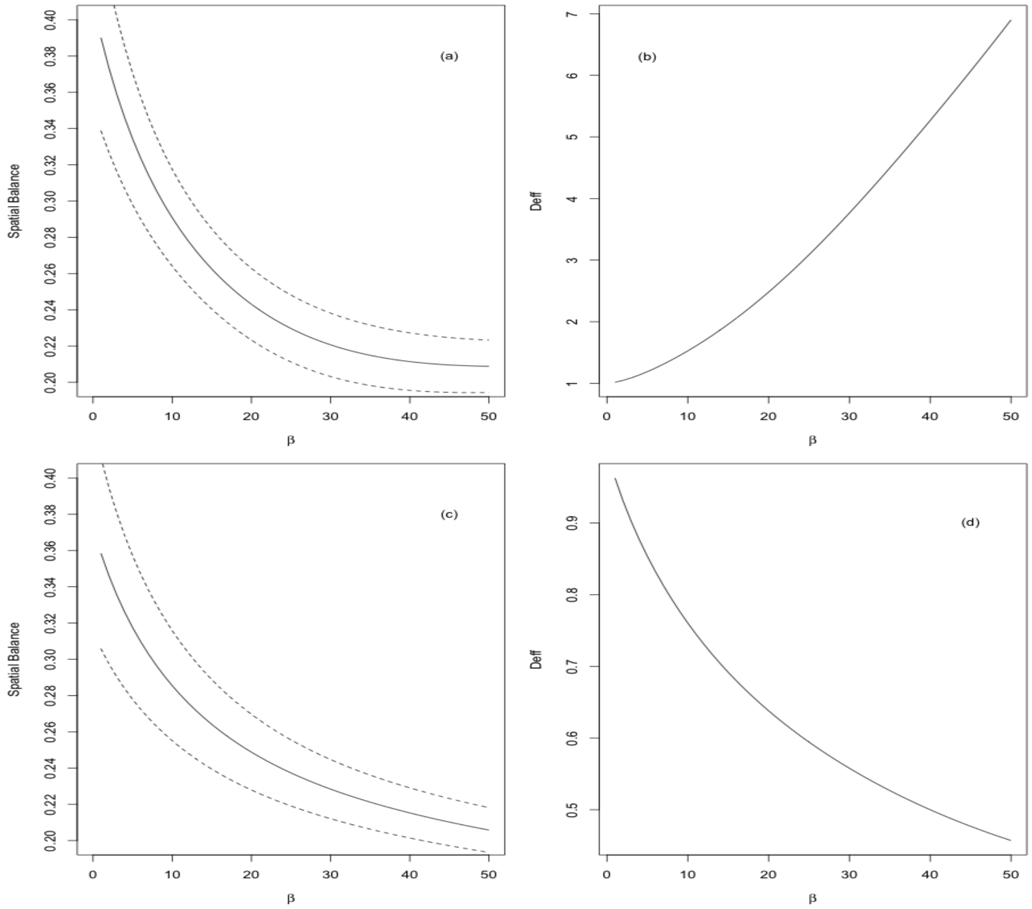


In particular, it can be conjectured that if each sample is selected with a probability that is proportional, with a constant greater than one, to its average within distance, the expected value $\mathrm{E}(\mathrm{V}(v_i \in s))=\Sigma_s p(s)\mathrm{V}(v_i \in s)$ should necessarily decrease; in other words, we should obtain samples that are better balanced in spatial terms. One possible way to acquire such a design is to set the $p(s)$ proportional to the power β of the average distance. Increasing this parameter, samples with high within distance will have a greater probability of being selected; thus, we will obtain designs that are more spatially balanced.

However, it will be interesting to see whether this easily achieved property implies that this design is more efficient than Simple Random Sampling (SRS) design. The usual efficiency measure of a design is $Deff = V\left(\hat{Y}\right)/V_{SRS}\left(\hat{Y}\right)$, that can be evaluated for several values of the parameter β (see Figure 3.2). Because of the variability of the first-order inclusion probabilities, which is unrelated to the variability of the target variable y, but rather to problems such as the edge effect or the isolation of a point, the variance of the estimates obtained by means of the distance-based design is greater than that of the SRS; its efficiency decreases as β increases, and the uncontrolled effects of the spatial distribution of the population amplifies the non-required variability of the π_i.

Spatial balance can also be negatively related to the efficiency of a design, especially if it is used to compare different designs having a different set of first-order inclusion probabilities. When the π_i are kept constant, again by using a standardized distance matrix, the efficiency of the design increases with β, like the spatial balance. It must be noted that, when the product of the standardized within sample distance matrix is used instead of the average, the expected spatial balance index is equal to 0.227 and the *Deff* is equal to 0.449; these values are very similar to those obtained with the design having the highest value of β.

Figure 3.2: (a) Expected value of the spatial balance index for different parameters β, when the distance matrix is not standardized (dashed lines represent the expected index, plus or minus its standard deviation); (b) Deff for different parameters β when the distance matrix is not standardized; (c) Expected value of the index of spatial balance for different parameters β, when the distance matrix is standardized (dashed lines represent the expected index, plus or minus its standard deviation); (d) Deff for different parameters β, when the distance matrix is standardized.



If it is assumed that the phenomenon survey may be affected by one or more of the spatial effects listed above, the problem is to ensure use of a design that will give higher probabilities to samples with higher within distance. Such a design $p(s)$ can be obtained by setting each $p(s)=\mathrm{M}(D_s)/\Sigma_s \mathrm{M}(D_s)$ as proportional to some synthetic index $\mathrm{M}(D_s)$ of the matrix D_s, observed within each possible sample s.

The most common sample selection algorithms (for a review, see Tillé, 2006) usually do not attempt to make a suitable choice for the sampling design's probability $p(s)$; however, its observance is at most verified only *a posteriori*. Traat *et al*. (2004) review the sampling designs and the sampling selection issues from the perpective of distribution. They begin from the assumption that the probability function $p(s)$ of the sampling design is known. Thus, drawing a sample $s\in\{0,1\}^N$ from a population U according to some sampling design, means generating an outcome from the multivariate design distribution $p(s)=P(s=S)$, with $\Sigma p(s)=1$ (the sampling design is of fixed size, thus $p(s) = 0$ when $\Sigma s_i \neq n$). Each element of the design vector is a Bernoulli random variable, and the joint distribution of the vector is a multivariate Bernoulli distribution whose moments of the first order are the π_is.

Traat *et al*. (2004) list different functional forms of the multivariate Bernoulli distribution, and develop a general list-sequential method for drawing a sample from any sampling design.

Markov Chain Monte Carlo (MCMC) methods and Gibbs-sampling, especially, can be used to generate samples from any high-dimensional distribution, if the probability function is known (Robert and Casella, 1999, Chapters 6 and 7). For example, Gibbs-sampling is an efficient algorithm for drawing a fixed-size sample from a multivariate Bernoulli design (Traat *et al*. 2004).
This algorithm is an iterative procedure, in which each step consists in running a Markov-Chain in which, given a configuration $s^{(t)}$ at the t-th iteration, another configuration – say $s^{(t+1)}$ – is chosen according to an acceptance rule known as the Metropolis criterion.

The proposed algorithm can be summarized as follows. The procedure starts at iteration t=0, with an initial point $s^{(0)}$, randomly selected from $\{0,1\}^N$ according to an SRS with constant inclusion probabilities. In a generic iteration t, the elements of $s^{(t)}$ are updated in these steps:

1. select two units at random, the first $i\in s$ included in the sample in the previous iteration and the second $j\notin s$ not included in the sample in the previous iteration. Formally, one must be from the units within the sample for which $s_i^{(t)}=1$, and another from the units outside the sample for which

$s_j^{(t)} = 0$;

2. denote with ${}^*s_i^{(t)}$ the sample in which the units in positions *i* and *j* exchange their status. Randomly decide whether or not to adopt ${}^*s_i^{(t)}$, that is:

$$s^{(t+1)} = \begin{cases} {}^*s^{(t)} & \text{with probability } p = \min\left\{1, \left(\frac{M\left(D_{{}^*s^{(t+1)}}\right)}{M\left(D_{s^{(t+1)}}\right)}\right)^{\beta}\right\} \\ s^{(t)} & \text{otherwise} \end{cases} \tag{3.14}$$

3. repeat steps (1) and (2) above $m \times q$ times (our computational experience indicates that in all the populations and sample sizes analysed, the choice $10 \times N$ produces samples with a frequency that closely approximates the proportionality to the index M(D_s)).

It has been shown (Robert and Casella, 2010, p. 182) that for a suitable choice of the parameters *m* and *q*, this iterative procedure will generate a random outcome from a multivariate probability, with *p*(*s*) proportional to the particular index used in (3.14).

Moreover, it is possible to coordinate the sample selection. The algorithm is easily implemented and extremely flexible, by simply modifying the criteria used in the first step to select the two candidate units *i* and *j*, or by changing the index M(D_s).

For example, it is possible to select a stratified sample proportional to M(Ds) in all the sample units, and only within each stratum h by simply starting with a stratified SRS with fixed sizes nh, and then selecting the candidate j from within the same stratum of the candidate I, rather than from among all the population units.

Regarding the index M(D_s), it is possible to modify both the definition of the distance matrix and how it is summarized within the sample. Concerning the first topic, there are no limits on the use of any distance or similarity definition. Assume, for example, to adopt the Euclidean distance, that we wish to emphasize that there are no limits on the use of powers in constructing this distance matrix, such as D_s^{γ}, to increase its effects on the spread of the sample over the study region. The parameter β plays an important rule in the design, since it controls our requirements on the within distance of the selected samples, setting the *p*(*s*) as more than proportional to the distance than we require. In addition, we also found it very useful to standardize the distance matrix to fixed row totals and, for the symmetry, column totals. To achieve this property of the matrix, we iteratively constrained the matrix's row (or column) sums, which were then again symmetric again by performing an average with its transpose. This is a very simple and accurate method for obtaining a distance matrix that is not affected by the problems that may arise from the use units with different inclusion probabilities, due to the non-required features of the phenomenon, such as edge effects and isolated points.

When seeking a class of candidates for the index M(D_s), the choice to adopt the power means based on their known properties and flexibilities appears natural. In particular, remarkable results have been obtained by using:

$$M_1(D_s) = \sum_{i;s_i=1} \sum_{j;s_j=1} d_{ij} \tag{3.15}$$

$$M_0(D_s) = \prod_{i;s_i=1} \prod_{j \neq i;s_j=1} d_{ij} \tag{3.16}$$

$$M_{-\infty}(D_s) = \min_{i;s_i=1, j \neq i;s_j=1} \{d_{ij}\} \tag{3.17}$$

Notice that while the division by n in (3.15) does not entail any alteration with respect to the average, because of the proportionality of $p(s)$ to the used index, the lack of the power $2/n(n$-1) in (3.16) is rather different from the $p(s)$ that could be obtained by using the geometric mean. In particular, it can be obtained by using the geometric mean, but with $\beta = n(n$-1)/2. We preferred to use it without the root for empirical reasons only (i.e. better results were obtained), although it could be interesting to obtain some theoretical results to better comprehend the circumstances in which the use of the products would be more efficient, instead of the sums or the minimum. Moreover, Function (3.16) can be given an appealing interpretation, since it is compatible with the underlying assumption that the distance between two units d_{ij} and their second-order inclusion probabilities π_{ij} is proportional. If this situation holds, Function (3.16) could be viewed as an attempt to approximate the entire sample's joint probability through the product of the probabilities of each couple $\{i,j\}$. In other words, we believe that the additional hypothesis of independence of the probabilities of order higher than two could be realistic. Setting a $p(s)$ in this way is compatible with the conventional practice of attempting to control the π_is and the π_{ij}s as they directly influence the estimates and their variance without considering higher order probabilities.

The option of using the entire distance matrix, or only its upper or lower part, depends only on computational reasons as in (3.15). (3.17) does not entail any variation to the obtained $p(s)$, while in (3.16) there should be a square root; however, in our experiments, this did not lead to any substantial difference in the results.

The performance of the proposed algorithm strictly depends on the degree to which the ratio in (3.14) can discriminate between different candidates to be included in the sample. If two candidate units display a very different pattern of distance with respect to the other (n-1) units included in the sample, this ratio should be very far from 1, which represents the situation of indifference. With respect to this property, the Index (3.16) displays good behavior, while (3.15) and (3.17) greatly depend on the distance between all the n units of the sample, and the ratio (3.14) can be equal, or very close, to 1 even though the two candidates present very different distances compared to the other sampling units. This could prevent the algorithm from generating reasonable outcomes in an acceptable number of iterations; thus, when using the average and the minimum, we chose the following indexes:

$$M_{1,i}(D_s) = \sum_{j;s_j=1} d_{ij} \tag{3.18}$$

$$M_{-\infty,i}(D_s) = \min_{j \neq i; s_j=1} \{d_{ij}\} \tag{3.19}$$

Although they no longer represent the average and the minimum distance within all sampling units, the distributions of the within sample distances obtained appear to confirm that their use accelerates the algorithm, without violating its convergence properties.

Estimation, specifically variance estimation, may be problematic for this sampling design: unfortunately, explicit derivations of π_i and π_{ij} for each unit and pairs of units in the population may be prohibitive for most summary indexes of distance. Consequently, the use of the HT estimator can be precluded. Since a frame population is under study, and the sampling design does not depend on unknown characteristics of the population, it is possible to generate as many independent replicates from the selection algorithm as necessary, and the π_is and π_{ij}s can be estimated on the basis of the proportion of times in which the units or the pairs of units are selected. These estimated inclusion probabilities could be adopted in the estimation process, instead of their theoretical counterparts (Fattorini, 2006 and 2009). Nevertheless, an evident property of the suggested selection procedure is that, unless d_{ij}=0 for one or more couples $\{i,j\}$, every π_{ij} is greater than 0 because any (or at least one) sample s with s_i= s_j=1 will have $p(s)$>0. This will always make it possible to compute an HT estimation of the variance that avoids a typical problem of spatially balanced sampling designs, which, for this reason, must usually propose some *ad hoc* variance estimation procedures (Stevens and Olsen, 2003 and 2004).

Furthermore, the use of this estimation procedure immediately drew our interest to useful evidence concerning the results of the empirical frequencies of units in selected samples. We observed that they are approximately constant, if the same index M(Ds) used in p(s) is constant when applied to each row or column of the population distance matrix.

3.5 Auxiliary and survey variables

Until now, the coefficients of variation constraints have been imposed on the auxiliary variables **X** rather than on the survey variables. The typical assumption is that optimal sample design (stratified or *πps*), based on specifying target levels of precision for a set of auxiliary variables, will lead to a design that achieves the required target precision k_j for each survey variable j.

The assumption is that these variables can be considered equal or similar, to determine the sample size required to reach a target precision; in other terms, the optimal sample design based on specifying target levels of precision for a set of auxiliary variables will lead to a design that achieves the required target precision, for each survey variable i. However, if considerable differences exist between the auxiliary variables and the survey variables, the solution will be sub-optimal, because it is well known that in practice, this hypothesis is only an approximation of the true situation, and that using the auxiliary variables to design the sample might underestimate the sample size required to reach a predetermined level of precision.

This is particularly true if administrative or remotely sensed data are used; if the survey variables and the auxiliary variables are simply the same variables recorded in two different periods, this hypothesis may also be considered acceptable.

In particular, as for the use of remotely sensed data, the survey variables and the auxiliary variables are not simply the same variables recorded in two different moments; therefore, considerable differences between them could exist. In these cases, the solutions that could be suggested in the above sections may be sub-optimal; it is well known that in practice, the previous hypothesis only approximates the true situation, and using the auxiliary variables to design the sample could thus underestimate the sample size required to reach a predetermined level of precision.

A standard alternative is to use a model for any of the unknown q survey variables **Y**, in terms of the known matrix of auxiliaries **X**. The solution that underpins the approach adopted by Baillargeon and Rivest (2009 and 2011) is to derive, from past surveys, a model that relates each $\boldsymbol{y}_l$ with its counterpart $\boldsymbol{x}_j$ observed in previous years. The sample allocation to each stratum is then made on the basis of the anticipated moments of **Y**, given **X**. It is important to emphasize that there is considerable advantage to designing a survey that is repeated at two time periods, in which the variables collected at each period have the same definition and the phenomenon being investigated is known to be highly dependent on its past values.

An important issue in this approach relates to the implicit use of a linear model linking the auxiliaries and the variable of interest **Y**. Clearly, a simple linear regression can be used if each variable has its own counterpart within the auxiliaries; or multiple regressions, if they represent a set of completely different information that is only related to the set of covariates. In these simple models, a log-scale relationship should help to reduce the effects of heteroschedastic errors and skewness of the population data.

The basis of this approach lies in the assumption that our prior knowledge is such that at the design stage, the finite population may be considered as if it were a sample from an infinite super-population, whose characteristics may be described by the model *M* (Isaki and Fuller, 1982).

Thus, to design a survey, it is necessary to search for the optimal anticipated variance (AV) of the estimator $\hat{\boldsymbol{t}}$ of total *t*, which can defined as the variance of the random variable $(\hat{t}-t)$ under both the design and the super-population models:

$$AV(\hat{t}-t)=E_M\left\{E_s\left[(\hat{t}-t)^2\right]\right\}+\left[E_M\left\{E_s(\hat{t}-t)\right\}\right]^2, \tag{3.20}$$

where E_M and E_S denote the expectation with regard to the model and to the sample design, respectively. Clearly, if the estimator is design-unbiased, as the HT estimator, the second part of (3.20) will be equal to 0.

The linear model to be used when a sample is designed on an estimated **y**, and not on the auxiliary **X** is (Särndal *et al.*, 1992, p. 449):

$$\begin{cases} y_k = \mathbf{x}_k^t \beta + \varepsilon_k \\ E_\xi(\varepsilon_k) = 0 \\ V_\xi(\varepsilon_k) = \sigma_k^2 = \sigma x_k^\gamma \\ E_\xi(\varepsilon_k \varepsilon_l) = 0 \end{cases} \tag{3.21}$$

where β is a vector of regression coefficients and ε_k are random variables.

Heteroschedastic variance (Särndal *et al.*, 1992, 12.2.6) is important to obtain a solution that is still a function of the auxiliary **X**; otherwise it will disappear from the Anticipated Variance (AV), introduced by Isaki and Fuller (1982), that under Model (3.21) will be (Särndal *et al.*, 1992, 12.2.13):

$$AV\left(\hat{t}_r\right)=\sum_{k=1}^{N}\left(\frac{1}{\pi_k}-1\right)\sigma x_k^{\gamma} . \tag{3.22}$$

If it is assumed that $Var_M(\varepsilon_k)=\sigma_l^2$ and $Cov_M(\varepsilon_k,\varepsilon_l)=\sigma_k\sigma_l\rho_{kl}$, under Model (3.21), the AV of the HT estimator of the total of a variable **y**, given **X**, is (Grafström and Tillé, 2013):

$$AV\left(\hat{t}_{HT}-t\right)=E_s\left[\left(\sum_{k\in s}\frac{x_k}{\pi_k}-\sum_{k\in U}x_k\right)^{t}\beta\right]^2+\sum_{k\in U}\sum_{l\in U}\sigma_k\sigma_l\rho_{kl}\frac{\pi_{kl}-\pi_k\pi_l}{\pi_k\pi_l} . \tag{3.23}$$

It can be seen that the first part of (3.23) is the error made by the HT estimator in estimating the true and known totals of the covariate, weighted through the regression coefficients; while the second term represents the typical expanded covariance of the indicator random variables used in the HT variance estimator, but weighted by the correlation of the model residuals. The introduction of such a super-population model could assist comprehension of the impact of space in the design of a sample of geo-coded units. Indeed, from (3.22), it is clear that the optimal design would be balanced on the set of auxiliaries, such that the first term is equal to 0, and is spatially balanced with sampling units so far apart that we can assume that ρ_{kl}=0 for each pair $k \neq l$ within the sample.

In practice, especially in stratified and πps sampling, the solutions proposed in literature are not explicitly finalized at minimizing the AV, but simply substitute the moments of the variable of interest with the Anticipated Moments (AM) of **y** given **X** (Baillargeon and Rivest, 2009 and 2011). Several alternatives to the linear regression model have been proposed, such as the log-scale relationship, which should contribute to reducing the effects of heteroschedastic errors and of skewness of the population, or a zero-inflated linear model (Fletcher *et al.*, 2005, Karlberg, 2000), that is useful when dealing with household surveys; indeed, a unit may go out of business in the time between the collection of the **X** variables and the date of the survey (Baillargeon and Rivest, 2009 and 2011; Benedetti *et al.*, 2008; Benedetti and Piersimoni, 2012).

The probability of being zero, i.e. of going out of business, or of suspending or postponing the activity of interest, typically decreases as the size of the establishments increases. The proposed model can then be based on a log-scale mixture model, with survival probabilities that are assumed to be constant for each unit i belonging to the same stratum *h:*

$$y_{i,j} = \begin{cases} \exp(\alpha_j + \beta_j \log(x_{i,j}) + \varepsilon_{i,j}) & \text{with probability } p_h \\ 0 & \text{with probability } 1 - p_h \end{cases}, \tag{3.24}$$

where $\varepsilon_{i,j} \sim N(0, \sigma_j^2)$. Such models, the parameters of which can be estimated by using maximum likelihood (Liu and Chan, 2010), are widely used for ecological count data, and have recently been extended to the analysis of economic microdata (Cameron and Trivedi, 2005). The anticipated moments under (3.24) can be derived from Baillargeon and Rivest (2009):

$$\mu_{y_j, S_{r,h,j}} = p_h e^{\alpha_j + \frac{\sigma_j^2}{2}} \left(\frac{\sum_{i \in h} x_{i,j}^{\beta_j}}{N_h} \right) \tag{3.25}$$

$$V_{y_j, S_{r,h,j}}^2 = p_h e^{2\alpha_j + 2\sigma_j^2} \left(\frac{\sum_{i \in h} x_{i,j}^{2\beta_j}}{N_h} \right) - p_h^2 e^{2\alpha_j + \sigma_j^2} \left(\frac{\sum_{i \in h} x_{i,j}^{\beta_j}}{N_h} \right)^2 \tag{3.26}$$

$$t_{y_j}^2 = e^{\alpha_j + \frac{\sigma_j^2}{2}} \sum_{h \in \{S,C\}} p_h \left(\sum_{i \in h} x_{i,j}^{\beta_j} \right) \tag{3.27}$$

It is important to highlight that the AM approach, although based on the assumption of a super-population model *M*, does not necessarily presume that the estimates are not design-based or that any inference made on the sample does not respect the randomization principles that are widely accepted by any statistician. Here, the model used is only a tool to better forecast the parameters of the investigated population and, thus, to better calibrate the sample size and its allocation.

The main advantages and drawbacks of the methods described in this topic are summarized in Table 3.1 below.

TABLE 3.1. Summary information for each sub-topic.

	πps sample selection	**Optimal Stratification**	**Spatially Balanced Samples**	**Models for anticipated moments**
Assessment of applicability in developing countries	These methods are widely used for frames of households; in developing countries, polygon and point frames are more commonly used.		The reliability of the data sets of reference should be assessed.	The reliability of the reference data sets should be assessed.
Recommendations on methods proposed in literature	These designs are usually not robust for outliers in X/Y, and are only univariate.	These practices are robust for outliers in X, and are usually multivariate.	The algorithms are very slow, and their application may be impossible for very large populations.	These models should be tuned for each single application. Their generalization is difficult.
Outline of research gaps and recommendations on areas for further research	A method is required to evaluate the π_i as a linear combination of a multivariate set of auxiliaries.	An algorithm to optimally stratify with irregularly shaped strata must be developed.	A flexible selection method with a probability proportional to the within sample distance.	Estimation and test of linear and nonlinear models that are “zero inflated” on remotely sensed data.

4

Extension of the regression or calibration estimators

4.1 Introduction

Survey statisticians place considerable efforts in the design of their surveys, to be able to use the auxiliary information for producing precise and reliable estimates. The class of calibration estimators is an example of a very general and practical approach to incorporating auxiliary information into the estimation. These estimators are used in most surveys performed by the major NSIs.

Agricultural surveys are highly specialized with respect to other surveys. These surveys are conducted to gather information on the crop area, crop yield, livestock, and other agricultural resources. Apart from the difficulties typical of business data, such as the quantitative nature of several variables and their high concentration, agricultural surveys are indeed characterized by certain additional peculiarities. In the case of auxiliary information, two specific issues must be discussed.

First, the definition of the statistical units is not unique. The list of possible statistical units is extensive, and its choice depends not only on the phenomenon for which the data is being collecting, but also on the availability of a frame of units (unless Indirect Sampling is used; see Lavallée, 2007). Second, a rich set of auxiliary variables, other than dimensional variables, is available: consider, for example, the information provided by airplane or satellite remote sensing.

Concerning the first issue, agricultural surveys can be conducted using a list frame or a spatial reference frame. Generally, a list frame is based on an agricultural census, a farm register or administrative data. A spatial reference frame is defined by a cartographic representation of the territory and by a rule that defines how it is divided into units. According to the available frame, different statistical units are available.

Agricultural holdings are the statistical units of a list frame. Surveys based on agricultural holdings are generally cheaper, since it is possible to collect a significant amount of information in a single interview. However, these presume that the list is recent and is of good quality: conditions that are not always satisfied.

Points are an example of statistical units of a spatial reference frame. Surveys based on points are often called point frame surveys. In theory, points are dimensionless, but they can be defined as having a certain size, for the sake of coherence with the observation rules, or the location accuracy that can be achieved. Segments are a second type of statistical units of a spatial reference frame. The choice of segment size depends on the landscape. Also, segments can be delimited by physical elements.

Two main differences among the statistical units must be highlighted. Information on the positioning (i.e. geo-referencing) of agricultural holdings is not always available, while it is instead always obtainable for points and segments. Geo-reference is seen as an important source of data, to be complemented with spatial agricultural information such as satellite images, land cover maps or other geo-referenced information layers. As is usual with business surveys, the population of agricultural holdings is markedly asymmetrical. Usually, asymmetry is positive, as small family-owned holdings coexist with large industrial companies.

With regard to the second issue, the rich set of auxiliary variables in agricultural surveys is mainly available by means of remote sensing data. Remote sensing can significantly contribute to provide a timely and accurate picture of the agricultural sector, because it is extremely suitable for gathering information over large areas with high revisit frequency. Indeed, a large range of satellite sensors regularly provides us with data that covers a broad spectral range. To derive the information sought, a large number of spectral analysis tools have been developed. For a review of the remote sensing applications devoted to the agricultural sector, see Section 2 above and the references cited therein.

An auxiliary variable that is commonly used for crop area estimates is the Land Use/Cover (LULC) data. LULC refers to data that is a result of raw satellite data classification into categories based on the return value of the satellite image. LULC data are most commonly presented in a raster or grid data structure, with each cell having a value that corresponds to a certain classification. LULC have been widely applied in estimating crop area. Hung and Fuller (1987) combine data collected by satellite with data collected by means of area survey, to estimate crop areas. Basic survey regression estimation is compared with two methods of transforming the satellite information, prior to regression estimation. González and Cuevas (1993) used thematic maps to estimate crop areas. The estimates were made using regression methods. Pradhan (2001) presents an approach to develop a Geographic Information System (GIS) for crop area estimation that supports a crop forecasting system at a regional level. The overall system combines spatial reference frame sampling and remote sensing.

Remote sensing data also provide information on different factors that influence crop yield. The most popular indicator for studying vegetation health and crop production is the Normalized Difference Vegetation Index (NDVI; see Section 2 above). This is a normalized arithmetic combination of vegetation reflectance in the red and near infrared. Studies have shown that NDVI values present a significant correlation with crop yields (see Section 2 above). Doraiswamy *et al.* (2005) evaluate the quality of the MODIS 250 m resolution data to retrieve crop biophysical parameters that could be integrated into crop yield simulation models. For a comprehensive review of the different ways to use remote sensing for agricultural statistics, see also Gallego (2004) and some references cited in Section 2.

The availability of remote sensing data does not eliminate the need for ground data, since satellite data do not always present the accuracy required. However, this information can be used as auxiliary data to improve the precision of the direct estimates. In this framework, the calibration estimator can improve the efficiency of crop areas and yield estimates for a large geographical area, when classified satellite images and NDVI can, respectively, be used as auxiliary information.

Section 4.2 outlines the calibration approach. Some possible extensions are presented in Section 4.3. In Section 4.4, we review the issue of model calibration. When complex auxiliary information is available, calibration methods assume special features; these are presented in Section 4.5. Section 4.6 describes a calibration approach for non-response adjustment, while Section 4.7 deals with the problem of missing data in the covariates. Finally, Section 4.8 features some remarks on computational issues.

4.2 The calibration approach to estimation

The technique of estimation by calibration was introduced by Deville and Särndal (1992). The basic idea is to use auxiliary information to obtain new sampling weights, called calibration weights that lead the estimates to agree with known totals. The estimates are generally design-consistent, and with a smaller variance than the HT estimator.

Consider a probability sample s selected from a finite population $U = \{1,2,\ldots,k,\ldots,N\}$, using a probability sampling $p(.)$. The first- and second-order inclusion probabilities, $\pi_k = P(k \in s)$ and $\pi_{lk} = P(k \,\&\, l \in S)$ respectively, are assumed to be strictly positive. Let $y_k, k = 1,...,N$ be the study variable. Suppose that we are interested in estimating the population total $t_y = \sum_U y_k$. An HT estimator of t_y is $\hat{t}_{y\pi} = \sum_s d_k y_k$, where $d_k = 1/\pi_k$ is the sampling weight for unit *k*. The HT estimator is guaranteed to be unbiased, regardless of the sampling design *p*(.). Its variance under *p*(.) is given as $V(\hat{t}_y) = \sum_U \sum_U (\pi_{kl} - \pi_k \pi_l) \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$

.

Let us assume that *J* auxiliary variables are available. Let $\mathbf{x}_k = (x_{k1},...,x_{kj},...,x_{kJ})^t$, *k*=1,...,*N* be a *J*-dimensional vector of auxiliary variables associated with unit *k*. The totals of the *J* auxiliary variables $\mathbf{t}_x = (t_{x_{k1}},...,t_{x_{kj}},...,t_{x_{kJ}}) = (\sum_U x_{k1},..., \sum_U x_{kj},..., \sum_U x_{kJ})$ are known.

The link between the variables of interest and the auxiliary information is crucial, if the latter is to be successfullyused. In agricultural surveys, there are differences between the statistical units regarding the use of the auxiliary variables available.

When the statistical units are agricultural holdings, the use of auxiliary information depends on the availability of information on the positioning of agricultural holdings. If the agricultural holdings are geo-referenced, the vector of auxiliary information for crop area estimates related to the farm *k* is given by $\mathbf{x}_k = (x_{k1},...,x_{kj},...,x_{kJ})^t$, with $\boldsymbol{x}_{kj}$ containing the number of pixels classified in crop type *j* according to the satellite data in farm *k*. When the statistical units are points, the vector of auxiliary information for crop area estimates related to point *k* is given by $\boldsymbol{\delta}_k = (\delta_{k1},...,\delta_{kj},...,\delta_{kJ})^t$, *k*=1,...,*N*, where δ_{kj} is an indicator variable with value $\delta_{kj} = 1$ if the point *k* is classified in crop type *j*, and $\delta_{kj} = 0$ otherwise.

The location accuracy between the ground survey and satellite images, and the difficulties inherent in improving this accuracy through geometrical correction, are considered one of the main problems in relating remote sensing satellite data to crop areas or yields, mainly in point frame sample surveys where the sampled point represents a very small portion of the territory.

When the statistical units are regular or irregular polygons, similarly to agricultural holdings, the vector of the auxiliary information for the crop area related to point *k* is given by $\mathbf{x}_k = (x_{k1},...,x_{kj},...,x_{kJ})^t$, *k*=1,...,*N*, with $\boldsymbol{x}_{kj}$ containing the number of pixels classified in crop type *j* according to the satellite data in point *k*.

Ideally, $\sum_s d_k \mathbf{x}_k = \mathbf{t}_x$, but often this is not true. In rough terms, the methodology proposed by Deville and Särndal (1992) identifies weights by means of a distance measure, and a system of calibration equations. The procedures can be summarized as follows:

1. Compute the initial design weight $d_k = 1/\pi_k$, directly obtained from the sampling design.
2. Compute the quantities γ_k to correct the initial weights as little as possible, for consistency with the auxiliary variables.
3. Compute the final weight as $w_k = d_k \gamma_k$.

Formally, the class of calibration estimators, calibrated to $\boldsymbol{t_x}$, is the class of estimators of the form:

$$\hat{t}_{y_w} = \sum\nolimits_s w_k y_k \,, \tag{4.1}$$

where $\boldsymbol{w_k}$ satisfies:

$$\sum\nolimits_s w_k \mathbf{x}_k = t_x \,. \tag{4.2}$$

The set of final weights *wk* is found by solving an optimization problem as follows:

$$\begin{cases} \min \sum\nolimits_s G(w_k, d_k) \\ \sum\nolimits_s w_k \mathbf{x}_k = \mathbf{t}_x \end{cases}, \tag{4.3}$$

where $G(w_k, d_k)$ is a function that measures the distance between the original weight *dk* and the new weight *wk*. To define a finite and unique solution, the function should satisfy precise conditions (Deville and Särndal 1992). To find the solution *wk* of System (4.3), it is necessary to define the Lagrangian as:

$$\min \sum\nolimits_s d_k G(w_k, d_k) - \boldsymbol{\lambda}\left(\sum\nolimits_s w_k \mathbf{x}_k - \mathbf{t}_x\right), \tag{4.4}$$

where the vector $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_j, ..., \lambda_J)^t$ are Lagrange multipliers. Differentiating (4.4) with respect to *wk*, we obtain:

$$g_k(w_k, d_k) - \mathbf{x}_k^t \boldsymbol{\lambda} = 0, \tag{4.5}$$

where $g(x) = dG(x)/dx$. Finally, we solve for *wk*, to obtain:

$$w_k = d_k F(\mathbf{x}_k^t \boldsymbol{\lambda}), \tag{4.6}$$

where $F(u) = g^{-1}(u)$ denotes the inverse function of *g*(.). To determinate the values of **λ**, we must solve the calibration equations as:

$$\varphi_s(\boldsymbol{\lambda}) = \sum_s d_k \left(F_k(\mathbf{x}_k^t \boldsymbol{\lambda}) - \mathbf{1}\right)\mathbf{x}_k = \mathbf{t}_x - \hat{\mathbf{t}}_{x\pi}, \tag{4.7}$$

where **λ** is the only unknown. Once **λ** is defined, the resulting calibration estimator is:

$$\hat{\mathbf{t}}_{yw} = \sum_s d_k F_k(\mathbf{x}_k^t \boldsymbol{\lambda}) y_k . \tag{4.8}$$

We can therefore summarize the procedure proposed by Deville and Särndal (1992) as follows:

1. Definition of a distance function $G(w_k, d_k)$.
2. Given a sample *s* and the function *F*(.) chosen in the previous step, solution, with respect to **λ**, of the calibration equations in (4.7), where the quantity on the right-hand side is known.
3. Computation of the calibration estimator of t_y, i.e. $\mathbf{t}_{yw} = \sum_s w_k y_k = \sum_s d_k F_k(\mathbf{x}_k^t \boldsymbol{\lambda}) y_k$.

This estimator will give closer estimates of t_y as the relationship between *x* and *y* becomes stronger. Examples of distance functions *G* are presented in Deville and Särndal (1992):

1. Chi-squared distance: $(w_k - d_k)^2 / 2d_k q_k$
2. Logarithm distance: $q_k^{-1}\left(w_k \log(w_k / d_k) - w_k + d_k\right)$
3. Hellinger distance: $2(\sqrt{w_k} - \sqrt{d_k})^2 / q_k$
4. Minimum entropy distance: $q_k^{-1}\left(-d_k \log(w_k / d_k) + w_k - d_k\right)$
5. Modified chi-squared distance: $(w_k - d_k)^2 / 2w_k q_k$
6. Truncated (L,U) logarithm distance or Logit:
$$\frac{1}{A}\left[\left(\frac{w_k}{d_k} - L\right)\log\left(\frac{(w_k / d_k) - L)}{(1-L)}\right) + \left(U - \frac{w_k}{d_k}\right)\log\left(\left(\frac{U - (w_k / d_k))}{U - 1}\right)\right)\right] \quad L < \frac{w_k}{d_k} < U$$

7. Truncated (L, U) chi-square distance

$$\begin{cases} (w_k - d_k)^2 / 2d_k q_k & L < \dfrac{w_k}{d_k} < U \\ \infty & \text{otherwise} \end{cases},$$

where q_k is a tuning parameter that can be manipulated to achieve the optimal minimum, and L and U are two constants such that $L<1<U$ and $A=(U\text{-}L)/((1\text{-}L)(U\text{-}1))$. The choice of distance function depends on the statistician and the problem.

It can be shown that most traditional estimators are a special case of the calibration estimator.

For example, the GREG estimator is a special case of the calibration estimator, when the chosen distance function is the Chi-square distance.

Consider the chi-square distance function $G(w_k, d_k) = (w_k - d_k)^2 / 2d_k q_k$; then, $\boldsymbol{F(u) = (1 + q_k u)}$ leads to the calibration weight $w_k = d_k(1 + q_k \mathbf{x}_k^t \boldsymbol{\lambda})$, where the vector of Lagrange multipliers λ is determined from Equation (4.7) as $\boldsymbol{\lambda} = \mathbf{T}_s^{-1}(\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi})$, and where $\mathbf{T}_s = \sum_s d_k q_k \mathbf{x}_k \mathbf{x}_k^t$, assuming that the inverse exists and that $\hat{\mathbf{t}}_{x\pi}$ is the Horvitz-Thompson estimator for **x**. The resulting calibration estimator is:

$$\hat{\mathbf{t}}_{yw} = \sum_s w_k y_k = \hat{\mathbf{t}}_{y\pi} + (\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi})^t \hat{\mathbf{B}}_s \qquad (4.9)$$

where $\hat{\mathbf{B}}_s = \mathbf{T}^{-1} \sum_s d_k q_k \mathbf{x}_k y_k$. Written in this form, it can be seen that the calibration estimator is the same as the GREG estimator.

Thus, when **x** are quantitative, it is nothing more than a multivariate extension of the classical regression estimator used with segments by NASS and MARS. However, when both **x** and **y** are land use/cover codes, it will become very similar to the confusion matrix-based estimator proposed by Gallego and Delince (1994), which can be useful when using a sample selected from a point frame.

If we take $\mathbf{x}_k = x_k$ and consider the Chi-square distance function, with $q_k = 1/x_k$, then $\lambda = t_x / \hat{t}_{x\pi} - 1$, $w_k = d_k(1 + q_k x_k^t \lambda) = d_k(1 + \lambda) = d_k t_x / \hat{t}_{x\pi}$. From Equation (4.9), the calibration estimator $\hat{t}_{yw} = \hat{t}_{y\pi} t_x / \hat{t}_{x\pi}$ is the ratio estimator.

Deville *et al.* (1993), Zhang (2000), and Breidt and Opsomer (2008) explain the post-stratified estimator and the raking as a special case of calibration estimation, when the information available consists of, respectively, known cell counts or known marginal counts in a contingency table. For the sake of simplicity, let us consider a two-way consistency table with R tows and C columns, and thus RxC=J cells. The cell (r,c), r=1,...,R; c=1,...,C contains N_{rc} elements. Then, $N=\sum_{r=1}^{R}\sum_{c=1}^{C}N_{rc}$. In the case of complete post-stratification, the vector of auxiliary information $\mathbf{x}_k=(x_{k1},...,x_{kj},...,x_{kJ})^t=(\delta_{k1},...,\delta_{kj},...,\delta_{kJ})^t$ is composed of J elements indicating the cell to which the unit k belongs, i.e. $\delta_{kj}=1$, if the unit k belongs to the cell j, and $\delta_{kj}=0$ otherwise.

Then, $\mathbf{t}_k=\left(\sum_U\delta_{k1},...,\sum_U\delta_{kj},...,\sum_U\delta_{kJ}\right)=(N_{11},...,N_{rc},...,N_{RC})$ is the vector of known population cell counts. Regardless of the F function, from the calibration equations (4.7) $F(\mathbf{x}_k^t\boldsymbol{\lambda})=F(\lambda_{ij})=N_{rc}/\hat{N}_{rc}$, where $\hat{N}_{rc}=\sum_{s_{rc}}d_k$ and s_{rc} denote the sample in cell (r,c). The resulting calibration estimator is $\hat{\mathbf{t}}_{ypos}=\sum_{r=1}^{R}\sum_{c=1}^{C}N_{rc}\sum_{s_{rc}}y_kd_k/\hat{N}_{rc}$; in other words, the calibration estimator is the same as the post-stratified estimator. The post-stratified estimator is the calibration estimator when the statistical units are points, and the vector of auxiliary information is given by crop type.

When the marginal cell count $N_{r.}$ and $N_{.c}$, r=1,...,R; c=1,...,C are known, but the cell count N_{rc} are not, we denote the estimate procedure as the cell count raking ratio procedure. Deville *et al.* (1993) obtained the raking ratio weights by minimizing the distance function 2, the logarithm distance.

Among all these distance functions, Andersson and Thorburn (2005) consider the issue of the determination of the optimal estimator, and found that it is based on the distance closely related to (but not identical to) that generating the GREG estimator, i.e. the chi-square distance.

A limitation of the calibration estimator with the chi-square distance function consists in the fact that the weights can assume negative or extremely large values. Deville and Särndal (1992) recognized this issue, and demonstrated how to restrict the weights so that they would fall within a certain range. The distance functions 2, 3, 4 and 5 guarantee positive weight. However, in each of the aforementioned cases, the weights can be unacceptably large with respect to the initial weights. They therefore consider the two additional functions 6 and 7, which have the attractive property of yielding weights that are restricted to an interval that statisticians can specify in advance.

It is important to note that, depending on the distance function chosen, there may not be a closed form solution to (4.7). Indeed, when the model for the correcting factors $F\left(\mathbf{x}_k^t\boldsymbol{\lambda}\right)$ is a linear function of **x**, it is possible to rewrite (4.7) in the form $\varphi_s(\boldsymbol{\lambda}) = \mathbf{T}_s\boldsymbol{\lambda}$, where $\mathbf{T}_s$ is a symmetric positive definite (J x J) matrix. The solution is therefore given by $\boldsymbol{\lambda} = \mathbf{T}_s^{-1}\left(\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi}\right)$. When the function $F\left(\mathbf{x}_k^t\boldsymbol{\lambda}\right)$ is non-linear, the solution can be found by using iterative techniques, usually based on the Newton–Raphson algorithm.

Deville and Särndal (1992) state that for any function $F_k(u)$ satisfying certain conditions, the calibration estimator is asymptotically equivalent to the regression estimator given in (4.9). Then, the two estimators have the same asymptotic variance (AsV), namely:

$$_{asy}V(\hat{\mathbf{t}}_{yw})\sum\nolimits_U \sum\nolimits_U \left(\pi_{kl} - \pi_k\pi_l\right)\left(d_kE_k\right)\left(d_lE_l\right),$$

where $E_k = y_k - \mathbf{x}_k^t\mathbf{B}$, with **B** solution of the equation

$$\left(\sum\nolimits_U q_k \mathbf{x}_k^t \mathbf{x}_k^t\right)\mathbf{B} = \sum\nolimits_U q_k \mathbf{x}_k y_k \ .$$

The asymptotic variance of $\hat{\mathbf{t}}_{yw}$ can be estimated as:

$$_{asy}\hat{V}(\hat{\mathbf{t}}_{yw})\sum\nolimits_s \sum\nolimits_s \left(\frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}}\right)\left(w_ke_k\right)\left(w_le_l\right),$$

where the e_k are the sampling residuals, $e_k = y_k - \mathbf{x}_k^t\hat{\mathbf{B}}$ with $\hat{\mathbf{B}} = \mathbf{x}^t\Gamma\mathbf{x}\left(\mathbf{x}^t\Gamma\mathbf{x}\right)^{-1}$ and $\Gamma = diag(\mathbf{d})$ is the nxn diagonal matrix of the direct weights.

4.3 Extension of calibration estimator

To obtain calibration weights, an alternative to distance minimization is the instrumental vector method. Estevao and Särndal (2000 and 2006) removed the requirement of minimizing a distance function, to introduce the functional form of calibration weights $w_k = d_kF\left(\boldsymbol{\lambda}^t\mathbf{z}_k\right)$, where $\mathbf{z}_k$ is an instrumental vector sharing the dimension of the specified auxiliary vector, and the vector λ is determined from the calibration equation (4.2). Several choices of the function $F(.)$ can be carried out, where the function $F(.)$ plays the same role as in the distance minimization method; for example, the linear function $F(u)$=1+n corresponds to the chi-square distance, and the exponential function $F(u)$=exp(u) corresponds to the logarithm distance. If we choose a linear function, and $\mathbf{z}_k$=$q_k\mathbf{x}_k$, the resulting calibration estimator is given by (4.9).

For a fixed set of auxiliary variables and a sampling design, Estevao and Särndal (2004) find an asymptotically optimal z vector:

$$\mathbf{z}_k = d_k^{-1} \sum_{l \in s} \left(d_l d_k - d_{kl} \right) \mathbf{x}_l ,$$

where d_{kl} is the inverse of the second-order inclusion probability, assumed to be strictly positive.

The resulting calibration estimator,

$$\hat{\mathbf{t}}_{y\mathrm{cal}} = \sum_{k} d_k \left(1 - \lambda^t \mathbf{z}_k \right),$$

is essentially the *randomization optimal estimator.*

4.4 Model calibration

The calibration estimator has many practical advantages. However, a limitation is that it relies on an implicit linear relationship between the variable **y** under study and the auxiliary variables **X**. Thus, a non-linear relationship between **y** and **X** exists, the calibration estimator does not perform as well as the HT estimator, i.e. if the auxiliary variable is ignored altogether (Wu and Sitter, 2001).

Särndal (2007) discusses the theoretical motivation and practical advantages and disadvantages of nonlinear GREG estimators (for logistic GREG, see Lehtonen and Veijanen, 1998), with respect to the standard GREG that is assisted by a linear fixed-effects model. He concludes that more research is needed to find better reasons in favour of using the nonlinear GREG estimators for practical purposes, e.g. in the production of routine official statistics.

The purpose of this Section is to propose efficient calibration and regression estimators for agricultural data (Gallego and Delincé, 1994). In particular, the introduction of complex models and flexible techniques, that make use of complete auxiliary information, is important in agricultural surveys where:

- the relationship between x and y may take on several forms and is highly dependent on the statistical units;
- complete auxiliary information, such as satellite images, land cover maps or other geo-referenced information layers, is available for all population units (Gallego and Delincé, 1994).

The calibration approach is a method for computing weights that reproduce the specified auxiliary totals, without an explicit assisting model. The calibration weights are justified principally by their consistence with the auxiliary variables. However, statisticians are trained to think in terms of models, and they often feel obligated to always have a statistical procedure that states the associated relationship of y to $\mathbf{x}$.

The notion behind the model calibration estimator proposed by Wu and Sitter (2001), Wu (2003), and Montanari and Ranalli (2005), is to extend the calibration estimator (Deville and Särndal, 1992) by assuming more general models to describe the relationship between the auxiliary variables and the survey variable. The reason is that, when the auxiliary information $\mathbf{x}_k$ is known for all population units, this should be used in a more effective way than is possible in model-free calibration, in which a known total is enough. More specifically, the method proposed by Wu and Sitter (2001) can be included in a model-assisted approach to inference, for a finite population. Model-assisted survey estimation (Särndal *et al.*, 1992) is an approach for incorporating auxiliary information in design-based survey estimation, assuming the existence of a working model ξ that describes the relationship between the auxiliary variables and the variable to be sampled. The population quantities are then estimated so that desirable design properties are present, such as asymptotic design unbiasedness and design consistency, regardless of the working model used and efficiency, if the model employed is correct. in other words, model-assisted estimation may lead to a considerably reduced variance, when appropriate auxiliary information is available and the correct model is used.

Consider a probability sample s selected from a finite population $U = \{1,2,...,k,...,N\}$ using a probability sampling $p(.)$. The first- and second-order inclusion probabilities, $\pi_k = P(k \in s)$ and $\pi_{lk} = P(k \,\&\, l \in S)$ respectively, are assumed to be strictly positive. Let $[y_k]_{k\in U}$ be the study variable, and $[\mathbf{x}_k]_{k\in U} = \left[(x_{k1},...,x_{kj},...,x_{kJ})^t\right]_{k\in U}$ J auxiliary variables available for all population units.

The model calibration estimator proposed by Wu and Sitter (2001) is a flexible idea. They consider an assisting model of the following type:

$$E_\xi(y_k \mid \mathbf{x}_k) = \mu(\mathbf{x}_k, \boldsymbol{\theta}), V_\xi(y_k \mid \mathbf{x}_k) = v_k^2\sigma^2, \quad k = 1,...,N, \tag{4.10}$$

where $\boldsymbol{\theta} = (\theta_1,...,\theta_p)^t$ and σ^2 are unknown assisting model parameters, $\mu(\mathbf{x}_k, \theta)$ is a specified function of $\mathbf{x}_k$ and $\boldsymbol{\theta}$, and v_k^2 is a known function of $\mathbf{x}_k$. E_ξ and V_ξ which denote, respectively, the mean and the variance with respect to the model specified.

A variety of estimators become possible via assisting models (4.10), where explicit formulations for the model mean and variance are given. One application of (4.10) is when $\mu(\mathbf{x}_k,\theta)$ is a linear or nonlinear function in $\mathbf{x}_k$, and $v_k = v(\mathbf{x}_k)$ is a strictly positive function of $\mathbf{x}_k$. Then, we have the linear or nonlinear regression model $y_k = \mu(\mathbf{x}_k,\boldsymbol{\theta}) + v_k \varepsilon_k, k = 1,...,N$, with ε_ks iid random variables with a mean of zero and variance σ^2. Other applications of (4.10) include generalized linear models, such that $g(\mu_k) = \mathbf{x}_k^t \boldsymbol{\theta}$ and $V_\xi(y_k \mid \mathbf{x}_k) = v(\mu_i)$, for a specified link function g(.) and an appropriate variance structure v(.).

We estimate the unknown parameter θ by $\hat{\theta}$, leading to values $\hat{y}_k = \boldsymbol{\mu}(x_k,\hat{\theta})$ that can be computed for all $k \in U$. Then, the weights are required to be consistent with the population total obtained with the predicted values $\hat{y}_k$. The weight system is not necessary consistent with the known population total of the auxiliary variable. If minimum chi-square distance is used, we may find the weights of the model calibration estimator $\hat{\mathbf{t}}_{y\mathrm{MCAL}} = \sum_s w_k y_k$ by minimizing the distance function $\sum_s (w_k - d_k)^2 / 2d_k q_k$, and using the calibration equation $\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$, where $\mathbf{x}_k = (1, \hat{y}_k)^t$. It follows that the population size N is known and plays an important part in the calibration. Then, the model calibration estimator is:

$$\hat{\mathbf{t}}_{y\mathrm{MCAL}} = N\left(\bar{y}_{s,d} + \left(\bar{\hat{y}}_U - \bar{\hat{y}}_{s,d}\right)\bar{\beta}_{s,d}\right), \tag{4.11}$$

where $\bar{y}_{s,d} = \sum_s d_k y_k / \sum_s d_k$, $\bar{\hat{y}}_{s,d} = \sum_s d_k \hat{y}_k / \sum_s d_k$ and

$$\bar{\beta}_{s,d} = \frac{\left[\sum_s d_k \left(\hat{y}_k - \bar{\hat{y}}_{s,d}\right) y_k\right]}{\left[\sum_s d_k \left(\hat{y}_k - \bar{\hat{y}}_{s,d}\right)^2\right]}.$$

That is, $\hat{\mathbf{t}}_{y\mathrm{MCAL}}$ can be considered a regression estimator that uses the predicted y-values as the auxiliary variable.

To estimate the model parameters $\boldsymbol{\theta}$ under a design-based framework, Wu and Sitter (2001) considered the following approach. The first step is obtaining an estimate $\boldsymbol{\theta}_N$ of $\boldsymbol{\theta}$ based on the entire population. Then, they obtain a design-based estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}_N$ from the sample data; for example, consider a generalized linear model. The estimate $\boldsymbol{\theta}_N$ of the population parameters $\boldsymbol{\theta}$ is defined as the solution of the estimating equation:

$$\sum_{k\in U} \mathbf{X}_k \, [dg(u)/du]^{-1}\{\mu(\mathbf{x}_k,\boldsymbol{\theta})\}v^{-1}\{\mu(\mathbf{x}_k,\boldsymbol{\theta})\}[y_k-\mu(\mathbf{x}_k,\boldsymbol{\theta})]=0, \qquad (4.12)$$

where $\mathbf{X}_k=(1,\mathbf{x}_k)$. The estimate $\hat{\boldsymbol{\theta}}$ is defined as the solution of the design-based sample version of (4.12), which is the solution of the following equation:

$$\sum_{k\in s} d_k \mathbf{X}_k \, [dg(u)/du]^{-1}\{\mu(\mathbf{x}_k,\boldsymbol{\theta})\}v^{-1}\{\mu(\mathbf{x}_k,\boldsymbol{\theta})\}[y_k-\mu(\mathbf{x}_k,\boldsymbol{\theta})]=0,$$

where $d_k=1/\pi_k$ are the design weights. It can be shown (Wu, 1999) that $\hat{\boldsymbol{\theta}}=\boldsymbol{\theta}_N+O_p(n^{1/2})$.

Once the unknown parameter $\boldsymbol{\theta}$ is estimated by $\hat{\boldsymbol{\theta}}$, the available auxiliary information can be included into the fitted function, leading to values $\hat{y}_k=\mu(\mathbf{x}_k,\hat{\boldsymbol{\theta}})$ that can be computed for all $k\in U$. The model calibration estimator of the mean can be defined as $\hat{\overline{Y}}_{MC}=N^{-1}\sum_{k\in s} w_k y_k$, where the calibration weights w_k s are obtained by minimizing the distance function $\sum_{k\in s}(w_k-d_k)^2/2d_k q_k$, under the calibration equations $\sum_{k\in s} w_k = N\sum_{k\in s} w_k\mu(\mathbf{x}_k,\hat{\boldsymbol{\theta}})=\sum_{k\in U}\mu(\mathbf{x}_k,\hat{\boldsymbol{\theta}})$.

It follows that the population size *N* is known and has an important role in the calibration. Then, the model calibration estimator (Wu and Sitter, 2001) is:

$$\hat{\overline{Y}}_{MC}=\hat{\overline{Y}}_{HT}+\left\{N^{-1}\sum\nolimits_U \mu(\mathbf{x}_k,\hat{\boldsymbol{\theta}})-N^{-1}\sum\nolimits_s d_k\mu(\mathbf{x}_k,\hat{\boldsymbol{\theta}})\right\}\hat{\beta}, \qquad (4.13)$$

with $\hat{\beta}=\sum_{k\in s} d_k q_k(\hat{\mu}_k-\overline{\mu})(y_k-\overline{y})/\sum_{k\in s} d_k q_k(\hat{\mu}_k-\overline{\mu})^2$,

$\overline{y}=\sum_{k\in s} d_k q_k y_k/\sum_{k\in s} d_k q_k$,

$\hat{\mu}_k=\mu(\mathbf{x}_k,\hat{\boldsymbol{\theta}})$ and $\overline{\mu}=\sum_{k\in s} d_k q_k\mu(\mathbf{x}_k,\hat{\boldsymbol{\theta}})/\sum_{k\in s} d_k q_k$, where q_k are a set of tuning parameters.

The properties of the model calibration estimator (4.13) were assessed in Wu and Sitter (2001), where a set of regularity conditional upon the behaviour of parameters $\hat{\boldsymbol{\theta}}$ and of the model mean $\mu(\mathbf{x}_k,\boldsymbol{\theta})$ must be assumed.

The important properties of (4.13) can be summarized as follows:

1. $\hat{\bar{Y}}_{MC}$ is asymptotically design-unbiased, regardless of the working model used.
2. The asymptotic variance of $\hat{\bar{Y}}_{MC}$ is given by:
   $${}_{asy}V(\hat{\bar{Y}}_{MC}) = N^{-2}\sum_U\sum_U(\pi_{kl}-\pi_k\pi_l)\left(\frac{E_k}{\pi_k}-\frac{E_i}{\pi_i}\right)^2,$$
3. where $E_k = y_k - \mu_k\beta$, with
   $\beta = \sum_U q_k(\mu_k-\bar{\mu}_N)(y_k-\bar{y})/\sum_s d_k(\mu_k-\bar{\mu}_N)^2$, $\mu_k = \mu(\mathbf{x}_k,\boldsymbol{\theta}_N)$ and $\bar{\mu}_N = N^{-1}\sum_U\mu_k$.
4. The asymptotic variance of $\hat{\bar{Y}}_{MC}$ can be estimated as:
   $$v\left(\hat{\bar{Y}}_{MC}\right) = N^2\sum_s\sum_s(\pi_{kl}-\pi_k\pi_l)\left(\frac{e_k}{\pi_k}-\frac{e_l}{\pi_l}\right)^2,$$
   where $e_k = y_k - \hat{\mu}_k\hat{\beta}$.
5. Using a linear working model, the model calibration estimator is reduced to the usual generalized regression estimator.

Wu and Sitter (2001) compare the linear calibration estimator ($\hat{\bar{Y}}_{MC}$) with the non-linear GREG ($\hat{\bar{Y}}_{GR}$), for the same non-linear assisting model. The study shows that the non-linear GREG is generally less efficient than the model calibration estimator. The reduction of the design variance when $\hat{\bar{Y}}_{GR}$ is used instead of $\hat{\bar{Y}}_{HT}$ is related to a sound prediction of the values taken by the survey variable in non-sampled units.

If the relationship between y and x is not sufficiently strong, it might be possible that $V(\hat{\bar{Y}}_{GR}) > V(\hat{\bar{Y}}_{HT})$. The idea of a model calibration estimator is different. $\hat{\bar{Y}}_{MC}$ involves fitting a general working model, and then calibrating on the resulting fitted values, while keeping as close to $\hat{\bar{Y}}_{HT}$ as possible. In other words, $\hat{\bar{Y}}_{MC}$ may be considered as a regression estimator that uses $\mu(\mathbf{x}_k, \hat{\boldsymbol{\theta}})$ as the auxiliary variable. The reduction of the design variance when using $\hat{\bar{Y}}_{MC}$ instead of $\hat{\bar{Y}}_{HT}$ is then related to a positive association between the variable y_k under study and the fitted values $\mu(\mathbf{x}_k, \hat{\boldsymbol{\theta}})$. Even when the model is misspecified, the gain from using a regression estimator on these variables remains.

Wu and Sitter (2001) show that the estimator, regardless of the model, is almost design-unbiased under minor conditions on the assisting model and on the sampling design. They also compare the linear calibration estimator $\hat{\mathbf{t}}_{y\mathrm{w}}$ with the model calibration estimator $\hat{\mathbf{t}}_{\mathbf{yMC}}$. The linear calibration estimator is less efficient than the model calibration, but $\hat{\mathbf{t}}_{y\mathrm{w}}$ presents several practical advantages over the model calibration estimator $\hat{\mathbf{t}}_{y\mathrm{MC}}$. Indeed, in this case, the auxiliary information is not required for all population units; the population total $\sum_U \mathbf{x}_k$ is enough. The same weights can be applied to all *y*-variables, because they do not depend on *y*; the estimator is identical to the linear GREG estimator. Moreover, in an empirical study, Wu and Sitter (2001) compare $\hat{\mathbf{t}}_{y\mathrm{MCAL}}$ with the non-linear GREG $\hat{\mathbf{t}}_{y\mathrm{GREG}}$ for the same non-linear assisting model. The study shows that the non-linear GREG is generally less efficient than the model calibration estimator. Demnati and Rao (2010) analyse the estimator of the total variance of non-linear population parameters when model calibration is used for estimation. Some comparison of alternative variance estimators of the GREG can be found in Stukel *et al.*, 1996.

Montanari and Ranalli (2005) provide further evidence. In an empirical study, they compare $\hat{\mathbf{t}}_{y\mathrm{MCAL}}$ with the non-linear $\hat{\mathbf{t}}_{y\mathrm{GREG}}$, where the assisting model is fitted via non-parametric regression (local polynomial smoothing). The model calibration estimator achieves only marginal improvement over the non-linear GREG. In model calibration, auxiliary information is required for all population units. When such information is not available, Wu and Luan (2003) propose a two-phase sample, where a large first-phase sample measures over the auxiliary variables.

In agricultural surveys, where complete auxiliary information is available from satellite data (thus available for the population), the non-linear assisting model may give a considerably reduced variance. The relationship between *x* and *y* can have several forms, thus yielding a great variety of possible assisting models that generate a wide family of model calibration estimators taking the form of (4.10). Cicchitelli and Montanari (2012) address the estimation of a spatial population's mean, using a model-assisted approach that considers semi-parametric methods. The idea is to assume a spline regression model that uses spatial coordinates as auxiliary information. With a simulation study, they show that significant gains in efficiency are made compared to the HT estimator, under a spatially stratified design. They also suggest using quantitative and qualitative covariates when available, other than the spatial coordinates, to increase the estimator's precision and to capture the target variables' spatial dependence.

The model calibration estimators proposed by Wu and Sitter (2001) rely on a class of working models that can be productively enlarged to account for more complex data structures. Some specific concerns arise when dealing with agricultural surveys; these must be addressed if a working model describing the relationship between the survey variable and the auxiliary variables is to be specified.

As is usual in the case of business surveys, the population of agricultural holdings is markedly asymmetric. Usually, asymmetry is positive, as small family-owned holdings coexist with large industrial companies. In this case, a log-scale relationship model should contribute towards reducing the effects of heteroschedastic errors and of the skewness of the population data. In addition, non-parametric modeling can be used.

As with all spatial populations, nearby locations tend to be influenced by the same set of factors. Spatial autocorrelation statistics measure and analyse the degree of dependence between observations in a geographic space. Positive spatial autocorrelation indicates the clustering of similar values across geographic space, while negative spatial autocorrelation indicates that neighbouring values are dissimilar. Thus, it is reasonable to model the response variable assuming the spatial autocorrelation among neighboring units.

An alternative approach to the introduction of spatial information into the model is to allow parameters to vary in space. That is, the effect of a covariate on the response variable changes depending on the area location. Thus, the areas' spatial pattern is not associated with random effects, but with the explanatory variables' influence on the response variable. In this modeling approach, a covariate's coefficient is not uniquely defined, but varies across areas. In this case, we can assume a local-stationarity model for the regression parameters that implies that the model is a finite mixture of e.g. C stationary models.

Another complex issue that often arises, when statistical units are usually farms or portions of land, is that the observed phenomenon can also be equal to zero with a non-null probability. Such a zero-inflated situation, where $\mathbf{X} > 0$ and $\mathbf{Y} = 0$, may arise because a unit has gone out of business between the collection of the $\mathbf{X}$ variables and the date of the survey.

Sub-section 4.4.1 below introduces some extensions of model calibration estimators that can be efficient when dealing with agricultural survey data.

4.4.1 Nonparametric model-assisted estimators

Breidt and Opsomer (2000) considered a nonparametric, model-assisted regression estimator using local polynomial regression. They showed that the use of nonparametric methods provide significant improvements in predicting the value of the variable of interest in non-sampled units. This feature increases the efficiency of the resulting estimators when compared with classical parametric estimators, especially when the underlying functional relationship is rather complex. Montanari and Ranalli (2005) combined model calibration estimation and nonparametric methods, and proposed nonparametric model-calibration estimators for a finite population mean.

Breidt and Opsomer (2000) assume that the working model ξ can describe the relationship between the survey variable and one auxiliary variable:

$$y_k = m(x_k) + \varepsilon_k, \quad k = 1,\ldots,N,$$

where ε_k s are independent random variables with mean of zero and a variance of $v(x_k)$. $m(x_k) = E_\xi(y_k \mid x_k)$ is a smooth function of x and is called the regression function, while $v(x_k) = V_\xi(y_k \mid x_k)$ is a smooth and strictly positive function of *x*, and is called the variance function.

Let $K_h(u) = h^{-1}K(u/h)$, where K denotes a continuous kernel function and *h* is the bandwidth. A local polynomial kernel estimator of degree *p*, of the regression function at x_k, based on the entire population, is given by:

$$m_k = \mathbf{e}_1^t(\mathbf{X}_{Uk}^t \mathbf{W}_{Uk} \mathbf{X}_{Uk})^{-1} \mathbf{X}_{Uk}^t \mathbf{W}_{Uk} \mathbf{y}_U = \mathbf{w}_{Uk}^t \mathbf{y}_U, \tag{4.14}$$

which is defined as long as $\mathbf{X}_{Uk}^t \mathbf{W}_{Uk} \mathbf{X}_{Uk}$ is invertible. Here, $\mathbf{e}_r$ is a column vector of length *p*+1, with a 1 in the *r*-th position and 0 elsewhere, $\mathbf{y}_u = [y_k]_{k\in U}$ is the *N*-vector of y_ks, $\mathbf{W}_{Uk} = diag\left[\frac{1}{h}K\left(\frac{x_j - x_k}{h}\right)\right]_{j\in U}$ is a $N\times N$ matrix and $\mathbf{X}_{Uk} = [1 x_j - x_k \ldots (x_j - x_k)p]_{j\in U}$ is an *N*×(*p*+1) matrix. In a sample context, m_k as defined in (4.14) cannot be calculated, since only $\mathbf{y}_u \in s \subset U$ are known. Breidt and Opsomer (2000) then obtain a design-based estimate $\hat{\boldsymbol{m}}_{\boldsymbol{k}}$ of m_k from the sample data:

$$\hat{m}_k = \mathbf{e}_1^t(\mathbf{X}_{sk}^t \mathbf{W}_{sk} \mathbf{X}_{sk})^{-1} \mathbf{X}_{sk}^t \mathbf{W}_{sk} \mathbf{y}_s = \mathbf{w}_{sk}^t \mathbf{y}_s, \qquad (4.15)$$

as long as $\mathbf{X}_{sk}^t \mathbf{W}_{sk} \mathbf{X}_{sk}$ is invertible, where $\mathbf{y}_s = [y_k]_{k \in s}$ is the n-vector of y_ks, $\mathbf{W}_{sk} = diag\left[\frac{1}{\pi_k h} K\left(\frac{x_j - x_k}{h}\right)\right]_{j \in s}$ is a $n \times n$ matrix, and $\mathbf{X}_{sk} = [1 x_j - x_k ... (x_j - x_k) p]_{j \in s}$ is an $n \times (p+1)$ matrix. The sample estimator $\hat{\boldsymbol{m}}_{\boldsymbol{k}}$ defined in (4.15) is an asymptotically design-unbiased and design-consistent estimator of the finite population smooth m_k based on a bandwidth h, because of the probability weights included in the smoothing weights.

Once the regression function $m(x_k)$ is estimated by $\hat{m}_k$ all $k \in U$, the local polynomial model calibration estimator of the mean can be defined as $\hat{\overline{Y}}_{MC}^{Lp} = N^{-1} \sum_s w_k y_k$, where the calibration weights wk are obtained by minimizing the distance function $\sum_s (w_k - d_k)^2 / 2 d_k q_k$ under the calibration equations $\sum_s w_k = N,\ \sum_s w_k \hat{m}_k = \sum_U \hat{m}_k$.

The minimization problem is solved as in Deville and Särndal (1992), and provides the following estimator:

$$\hat{\overline{Y}}_{MC}^{lp} = \hat{\overline{Y}}_{HT} + \left\{N^{-1} \sum_U \hat{m}_k - N^{-1} \sum_s d_k \hat{m}_k \right\} \hat{\beta}^{lp},$$
(4.16)

with $\overline{y} = \sum_s d_k q_k y_k / \sum_s d_k q_k$ $\overline{y} = \sum_s d_k q_k y_k / \sum_s d_k q_k$
and $\overline{m} = \sum_s d_k q_k \hat{m}_k / \sum_s d_k q_k$.

The properties of the local polynomial model calibration were assessed in Montanari and Ranalli (2005), under regularity conditions considered in Breidt and Opsomer (2000). The properties can be summarized as follows:

1. $\hat{\overline{Y}}_{MC}^{lp}$ is asymptotically design-unbiased and design-consistent;

   the asymptotic variance of $\hat{\overline{Y}}_{MC}^{lp}$ is given by:

   $${}_{asy}V(\hat{\overline{Y}}_{MC}) = N^{-2} \sum_U \sum_U (\pi_{kl} - \pi_k \pi_l) \left(\frac{E_k}{\pi_k} - \frac{E_i}{\pi_i}\right)^2,$$

2. where $E_k = y_k - m_k \beta^{lp}$, with

$$\beta^{lp} = \sum\nolimits_U q_k \left(m_k - \overline{m}_N\right)\left(y_k - \overline{y}\right) / \sum\nolimits_s d_k \left(m_k - \overline{m}_N\right)^2 \text{ and } \overline{m}_N = N^{-1}\sum\nolimits_U m_k$$

3. the asymptotic variance of $\overset{\wedge}{\overline{Y}}{}^{lp}_{MC}$ can be estimated as:

$$v(\overset{\wedge}{\overline{Y}}_{MC}) = N^{-2}\sum\nolimits_s\sum\nolimits_s\left(\pi_{kl} - \pi_k\pi_l\right)\left(\frac{e_k}{\pi_k} - \frac{e_i}{\pi_i}\right)^2 ,$$

   where $e_k = y_k - \hat{\mu}_k\hat{\beta}^{lp}$.

In principle, other nonparametric methods described in literature can be employed, within a model-assisted framework, to estimate values of the survey variable on non-sampled units. Breidt and Opsomer (2000) first considered a local polynomial regression estimator as a generalization of the ordinary regression estimator. However, such a technique is not easily extendable to multivariate auxiliary information.

Further attempts to handle multivariate auxiliary information, within a model-assisted framework, make use of generalized additive models, as in Opsomer *et al*. (2001). Besides, Breidt *et al*. (2005) have proposed the use of penalized splines. Finally, Montanari and Ranalli (2005) consider neural networks. However, treatment here is limited to local polynomials methods as a method that is easy to use in practice, and with commonly-available software.

4.4.2 Model-assisted estimator of spatially autocorrelated data

In agricultural statistics, the statistical units are points or areas. Probably, it is thought that the nearby location can be influenced by the same factors. Response variables, such as crop or crop yield that are close together may exhibit some degree of spatial dependence. Thus, in a design-based framework to survey spatial populations, the auxiliary information, provided by the spatial coordinates, should be used. Methods for achieving more efficient designs using the spatial pattern in the response variable at the design stage have received a great deal of attention. Less attention has been devoted to techniques aimed at improving efficiency at the estimation stage, using working models to capture the spatial pattern under the model-assisted framework.

Cicchitelli and Montanari (2012) deal with the estimation of the mean of a spatial population, using a model-assisted approach that considers semi-parametric methods. The idea is to assume a spline regression model that uses the spatial coordinates as auxiliary information, and then to employ the resulting population-fitted values in a difference estimator. The idea is to simulate what happens in the kriging predictor where a covariance function is estimated from the data, and the values of the variable are predicted as an average of the sample observations, with weights depending on the estimated correlation function.

Let us consider as spatial sampling frame U a regular grid of points ($\mathbf{x}_1,...,\mathbf{x}_N$), where $\mathbf{x}_k = (x_{k1}, x_{k2})$ are the spatial coordinates of the point k, for complete coverage of the area of interest. Cicchitelli and Montanari (2012) assume that the population values $\mathbf{y} = [y(\mathbf{x}_k)]^t{}_{k\in U}$ are realizations of the following linear mixed model ξ:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathrm{u} + \varepsilon, \tag{4.17}$$

where **X** and **Z** are a known full rank matrix, **β** a vector of parameters, $E_\xi(\mathbf{u})=\mathbf{0}$, $E_\xi(\boldsymbol{\varepsilon})=\mathbf{0}$, $Cov_\xi(\mathbf{u}, \boldsymbol{\varepsilon}) = \mathbf{0}$, $Var_\xi(\mathbf{u}) = \mathbf{G}$, and $Var_\xi(\boldsymbol{\varepsilon}) = \mathbf{R}$.

Let $\boldsymbol{\kappa}_1,...,\boldsymbol{\kappa}_l$ be a representative subset of locations $L \subset U$, such that the $l\times l$ matrix $\boldsymbol{\Omega} = \left\{(\|\kappa_r - \kappa_l\|)^2 \log(\|\kappa_r - \kappa_l\|)\right\}_{r,l=1,...,l}$ is non-singular. Then, model (4.17) is a low-rank radial smoother if:

- **X** is a $N \times 3$-matrix of spatial coordinates defined as $\mathbf{X} = [1, x_{k1}, x_{k2}]_{k\in U}$;
- **Z** is the $N\times l$ matrix of pseudo-covariate values defined as $\mathbf{Z} = [z_1(\mathbf{x}_k),..., z_l(\mathbf{x}_k)]_{k\in U}$, with $[z_1(\mathbf{x}_k),..., z_l(\mathbf{x}_k)] = [\tilde{z}_1(\mathbf{x}_k),...,\tilde{z}_l(\mathbf{x}_k)]\boldsymbol{\Omega}^{-1/2}$ and $\tilde{z}_l(\mathbf{x}_k) = (\|\boldsymbol{\kappa}_k - \boldsymbol{\kappa}_l\|)^2 \log(\|\boldsymbol{\kappa}_k - \boldsymbol{\kappa}_l\|)$;
- $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^t, \mathbf{u} = (u_1 ..., u_l)^t$ and σ_ε^2 are unknown assisting model parameters;
- ***G*** is an $l\times l$ matrix defined as $\mathbf{G} = \sigma_u^2 I_l$, with $\boldsymbol{I_l}$ being a *lxl* identity matrix;
- **R** is an $N\times N$ matrix defined as $\mathbf{R} = \sigma_\varepsilon^2 I_N$, with I_N being an $N\times N$ identity matrix.

The low-rank radial smoother belongs to the class of thin-plate splines. If we specify $l=N$ in the definition of matrix **Z,** the model (4.17) represents the full-rank radial smoother that would be obtained if we considered the entire set of locations $\mathbf{x}_1,...,\mathbf{x}_N$ as knots. Moreover, if $\mathbf{Z} = \boldsymbol{\Sigma}^{-1/2}$, where Σ is the covariance matrix of $\mathbf{y} = [y(\mathbf{x}_k)]_{k\in U}{}^t$ divided by σ^2, the model (4.17) represents the kriging smoother. Kriging assumes that $\mathbf{y} = [y(\mathbf{x}_k)]_{k\in U}{}^t$ derives from the performance of a stochastic process with a spatial dependence structure that is estimated from the data, while smoothing splines use a particular generalized covariance function (see Cressie, 1993). Cicchitelli and Montanari (2012) assume a low-rank smoother because it enables them to obtain the design-based consistent estimator.

If it is assumed that the values of $\mathbf{y}$ are known for the entire population, the parameters of the spline regression model (4.17) can be estimated by minimizing, with respect to $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^t$ and $\mathbf{u} = (u_1 ..., u_l)^t$, the function:

$$\frac{1}{\sigma_\varepsilon^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})^t(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) + \frac{1}{\sigma_{\mathrm{u}}^2}\mathbf{u}^t\mathbf{u}.$$

The estimator of $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^t$ and $\mathbf{u} = (u_1 ..., u_l)^t$, for fixed values of σ_u^2 and σ_ε^2, are given by:

$$\begin{bmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\mathbf{u}} \end{bmatrix} = \left[\begin{bmatrix} \mathbf{X}^t\mathbf{X} & \mathbf{X}^t\mathbf{Z} \\ \mathbf{Z}^t\mathbf{X} & \mathbf{Z}^t\mathbf{Z} \end{bmatrix} + \frac{\sigma_\varepsilon^2}{\sigma_u^2}\mathbf{D} \right]^{-1} \begin{bmatrix} \mathbf{X}^t\mathbf{y} \\ \mathbf{Z}^t\mathbf{y} \end{bmatrix},$$

where $\mathbf{D} = blockdiag[\mathbf{0}_{3\times 3}, \mathbf{I}_l]$. The fitted population variables $\tilde{\mathbf{y}} = [\tilde{y}(\mathbf{x}_k)]^t{}_{k\in U}$ are then given by $\tilde{\mathbf{y}} = \tilde{\mathbf{S}}_\sigma \mathbf{y}$, where $\tilde{\mathbf{S}}_\sigma$ is the smoothing matrix defined as:

$$\tilde{\mathbf{S}}_\sigma = [\mathbf{X},\mathbf{Z}] \left[\begin{bmatrix} \mathbf{X}^t\mathbf{X} & \mathbf{X}^t\mathbf{Z} \\ \mathbf{Z}^t\mathbf{X} & \mathbf{Z}^t\mathbf{Z} \end{bmatrix} + \frac{\sigma_\varepsilon^2}{\sigma_u^2}\mathbf{D} \right]^{-1} \begin{bmatrix} \mathbf{X}^t \\ \mathbf{Z}^t \end{bmatrix}. \qquad (4.18)$$

In a sample context, the smoothing matrix as defined in (4.8) cannot be calculated, since only the $\mathbf{y} = [y(\mathbf{x}_k)]_{k\in s}{}^t$ $s \subset U$ are known. Cicchitelli and Montanari (2012) defined a design-based estimate $\tilde{S}_{\sigma s}$ of $\tilde{S}_\sigma$ from the sample data as:

$$\tilde{\mathbf{S}}_{\sigma s} = [\mathbf{X},\mathbf{Z}] \left[\begin{bmatrix} \mathbf{X}_{\mathrm{s}}{}^t\Pi_{\mathrm{s}}\mathbf{X}_{\mathrm{s}} & \mathbf{X}_{\mathrm{s}}{}^t\Pi_{\mathrm{s}}\mathbf{Z}_{\mathrm{s}} \\ \mathbf{Z}^t{}_{\mathrm{s}}\Pi_{\mathrm{s}}\mathbf{X}_{\mathrm{s}} & \mathbf{Z}^t{}_{\mathrm{s}}\Pi_{\mathrm{s}}\mathbf{Z}_{\mathrm{s}} \end{bmatrix} + \frac{\sigma_\varepsilon^2}{\sigma_u^2}\mathbf{D} \right]^{-1} \begin{bmatrix} \mathbf{X}_{\mathrm{s}}\Pi_{\mathrm{ss}}{}^t \\ \mathbf{Z}^t{}_{\mathrm{s}}\Pi_{\mathrm{ss}} \end{bmatrix},$$

where X_s and $\mathbf{Z}_s$ are the sub-matrices of $\mathbf{X}$ and $\mathbf{Z}$, containing only the rows for the units $k \in s$, and $\Pi_{\mathrm{s}} = diag[(1/\pi_k)]_{k\in s}$. A design-based consistent estimator of $\tilde{\mathbf{y}} = [\tilde{y}(\mathbf{x}_k)]_{k\in U}{}^t$ is provided by $\hat{\tilde{\mathbf{y}}} = \tilde{\mathbf{S}}_{\sigma s}\mathbf{y}$.

Once the variable y is fitted by $\tilde{y}(\mathbf{x}_k)$ for all $k \in U$, Cicchitelli and Montanari (2012) define a model-assisted difference estimator of the population mean as:

$$\hat{\bar{Y}}_{GR}^{Spl} = N^{-1}\sum_{U}\hat{\tilde{y}}(\mathbf{x}_k) + N^{-1}\sum_{s}\frac{y(\mathbf{x}_k) - \hat{\tilde{y}}(\mathbf{x}_k)}{\pi_k} \ . \tag{4.19}$$

The properties of the proposed estimator are assessed in Cicchitelli and Montanari (2012); they introduce an appropriate asymptotic framework and regularity conditions. These properties can be summarized as follows:

1. $\hat{\bar{Y}}_{GR}^{Spl}$ is asymptotically design-unbiased and design-consistent;
2. The asymptotic variance of $\hat{\bar{Y}}_{GR}^{Spl}$ can be estimated as:

   $$v(\hat{\bar{Y}}_{GR}^{Spl}) = N^{-2}\sum_{s}\sum_{s}\left(\frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}}\right)\left(\frac{e(\mathbf{x}_k)}{\pi_k}\frac{e(\mathbf{x}_i)}{\pi_i}\right),$$

   where $e(\mathbf{x}_k) = y(\mathbf{x}_k) - \hat{\tilde{y}}(\mathbf{x}_k)$.

The estimator (4.19) is not calibrated with respect to the fitted values $\hat{\tilde{y}}(\mathbf{x}_k)$. As discussed in Section 4.2 above, a GREG estimator is generally less efficient than model calibration estimators that use the same working-model. A spline regression model-calibrated estimator could be derived, to obtain a more efficient estimator.

4.4.3 Model-assisted estimator for geographical local-stationary data

An alternative method to the introduction of spatial information into the working model is to allow the effect of a covariate to change across the area location, i.e. to allow some predictor variables to have differing impacts on the response variable, depending on their location. The data's spatial pattern is then described not with random effects, but with the impact of the explanatory variables on the response variable; the coefficient of a covariate is not uniquely defined, but varies from area to area.

In this modeling approach, spatial heterogeneity can be taken into account by considering the statistical units as belonging to different clusters, where the clusters can be viewed as groups of areas where the local parameters are stationary. Another approach, based on multilevel models, was suggested by Moura and Holt (1999).

Suppose that the statistical units are areas. Let $\mathbf{y}=[y_k]_{k\in U}{}^t$ be the population values measured at $k\in U$. To take the heterogeneity of the spatial parameters into account, the population values are assumed to be a realization of the following model ξ:

$$E_\xi(y_k \mid \mathbf{x}_k)=\mu(\mathbf{x}_k,\boldsymbol{\beta}_{c_k}), V_\xi(y_k \mid \mathbf{x}_k)=\sigma^2_{c_k}, \quad k=1,\dots,N,$$

where $\mathbf{x}_k$ is the p dimensional vector of predictors observed at $k\in U$, $\boldsymbol{\beta}_{c_k}$ is a p dimensional vector of locally stationary parameters, $\sigma^2_{c_k}$ is the variance of the group of k-th unit, and $\mathbf{c}=(c_1,c_2,\dots,c_k,\dots,c_N)\in C^N$ is the label of vector representing the cluster related to each single unit k belonging to the cluster $c_k\in\{1,2,,\dots,C\}$, with C number of clusters or heterogeneous zones.

From a theoretical point of view, the estimate of the heterogeneous regression parameters can be computed through the resolution of a combinatorial optimization problem. If the number C of local stationary zones is fixed, the aim is to minimize an objective function, defined in terms of best fit over all possible partitions of the study area. Theoretically, this spatial combinatorial problem could be solved through any search algorithm; unfortunately, this does not always enable us to avoid a local optimum. Postiglione *et al.* (2013) propose two algorithms for the identification of heterogeneous zones based on Simulated Annealing (SA) and Iterated Conditional Modes (ICM).

SA is a stochastic optimization method proposed by Kirkpatrick *et al.* (1983) for finding a function's global minimum. The method is a generalization of the Metropolis-Hastings algorithm (Metropolis *et al.*, 1953) in the optimization context, and is one of the most popular and commonly-used optimization strategies for the solution of complex combinatorial problems.

Generally, in the approach of Geman *et al.* (1990), the spatial combinatorial optimization problem can be considered a Markov Random Field (MRF), described by reference to a family of distributions (see also Section 3.2 above for the unconstrained, unpenalized, version of annealing):

$$\pi(\mathbf{x},\mathbf{y},\mathbf{c})=\frac{\exp\{-U(\mathbf{x},\mathbf{y},\mathbf{c})/T\}}{\sum_{\mathbf{c}\in C^N}\exp\{-U(\mathbf{x},\mathbf{y},\mathbf{c})/T\}}, \tag{4.20}$$

where $U(\mathbf{x},\mathbf{y},\mathbf{c})$ is the energy function, and plays a fundamental role in the optimization procedures, with $\mathbf{y}$ and $\mathbf{x}$ observed data, T is a positive parameter called temperature and the denominator of Equation (4.20) is the model's normalization constant.

The energy function is the procedure's objective function, and should be specified to take into account the model's goodness of fit to our data, and an adjacency constraint that considers the possibility of obtaining more or less aggregated spatial configurations. For these reasons, the energy function $U(\mathbf{x},\mathbf{y},\mathbf{c})$ can be defined as the sum of two terms: an interaction term $I(\mathbf{y},\mathbf{x},\mathbf{c})$, which depends on the data observed and on the labels, and a penalty term $V(\mathbf{c})$, which depends only on **c**. At the *j*-th iteration, the energy function can be defined as:

$$U_j(\mathbf{y}, \mathbf{x}, \mathbf{c}) = I_j(\mathbf{y}, \mathbf{x}, \mathbf{c})+V_j(\mathbf{c}). \tag{4.21}$$

The interaction term *I*(.) is given by a distance (usually Euclidean) function between the observations and the estimated regression function at the *j*-th iteration:

$$I_j(\mathbf{x},\mathbf{y},\mathbf{c}) = \sum_{k=1}^{N} e_{k.j}^2(\boldsymbol{\beta}), \tag{4.22}$$

where $e_{k.j}^2(\boldsymbol{\beta})$ is a distance function at *j*-th iteration. The adjacency constraints are formalized through the penalty function; for this, we adopt a Potts model (Sebastiani 2003):

$$V_j(\mathbf{c}) = -\chi \sum_{s=1}^{N}\sum_{v=1}^{N} p_{s,v} \mathbf{1}_{(c_s^j = c_v^j)}, \tag{4.23}$$

where $p_{s,v}$ is the element (*s*,*v*) of a binary contiguity matrix, $\mathbf{1}_{(c_s^j = c_v^j)}$ is the indicator function of the event $(c_s^j = c_v^j)$, and χ is a parameter that discourages configurations that do not have contiguous zones.

Now let us define as $\mathbf{c}^j \in C^N$ a configuration at the *j*-th iteration, and as $U(.,\mathbf{c}^j)$ the objective function observed for the configuration $\mathbf{c}^j$. The algorithm's classical version uses a logic according to which, given a configuration $\mathbf{c}^j$ at the *j*-th iteration, another configuration at the (*j*+1)-th iteration, say $\mathbf{c}^{j+1}$, is chosen according to a *visiting schedule.* The status can be exchanged if $U(.,\mathbf{c}^{j+1}) < U(.,\mathbf{c}^j)$. The final configuration is therefore obtained for any suitable choice of the stopping criterion.

In our algorithm, at the (*j*+1)-th iteration, for each geographical unit $k = 1,...,N$, a new candidate label ${c_k}^{j+1} \in \{1,2,...,C\}/c_k^j$ is randomly selected. The energy function $U(.,\mathbf{c}^{j+1})$ is computed and compared with the current energy $U(.,\mathbf{c}^j)$. If $U(.,\mathbf{c}^{j+1}) < U(.,\mathbf{c}^j)$, and the label ${c_k}^j$ is replaced with ${c_k}^{j+1}$. The algorithm will stop if $\mathbf{c}^{j+1} = \mathbf{c}^{j}$ hold true.

Generally, this result implies that a local, and not necessarily a global, minimum is reached. To avoid entrapments in local minima, it is necessary to provide a non-deterministically descending rule and, thus, to define the positive probability for the change of configuration, even when there is an increase in the energy function. Therefore, the algorithm replaces the solution obtained at the *j*-th iteration $\mathbf{c}^{j}$ with a new solution c^{j+1}, according to an acceptance rule known as the Metropolis criterion:

$$p = \begin{cases} 1 & \text{if} U(.,\mathbf{c}^{j+1}) < U(.,\mathbf{c}^{j}) \\ \exp\{-(U(.,\mathbf{c}^{j+1}) - U(.,\mathbf{c}^{j}) / T(j)\} & \text{otherwise} \end{cases}$$

This means that a move from a configuration $\mathbf{c}^{j}$ to a worse configuration $\mathbf{c}^{j+1}$ is allowed with a probability *p*, which depends on the value of the parameter *T(j)*, which indicates the temperature at the *j*-th step of the procedure and slowly decreases to 0, according to a schedule (see Section 3.2 above). It is clear that larger values of the parameter *T(j)* are expected to move the solution away from a local optimum, while smaller values of *T(j)* are expected to decreasing very slowly to zero.

This version of SA enables us to identify zones of local stationarity, and to classify each site in one, and only one, of the *C* groups.

In practice, since the value of *T* decreases very slowly to zero, the computational burden may become prohibitive. Thus, deterministic versions of the SA called ICM, proposed by Besag (1986), is a possible alternative to SA in solving complex combinatorial optimization problems. ICM is exactly equivalent to *instantaneous freezing* in SA (Besag, 1986); in other words, the temperature is set equal to zero, thereby ensuring a rapid convergence. See Besag (1986) for further details. Thus, the algorithm does not avoid entrapments in local minima. However, ICM is computationally faster than SA, and guarantees good solutions in practical applications.

Using the modified SA or ICM, it is possible to estimate the heterogeneous regression parameters $\tilde{\boldsymbol{\beta}}_{c_k}$. Then, the prediction of the values of the variable of interest is obtained as:

$$\hat{\tilde{y}}_k = \mu\left(\mathbf{x}_k, \hat{\tilde{\boldsymbol{\beta}}}_{c_k}\right). \tag{4.24}$$

These methods are usually formulated in a statistical framework that is not specific to sample surveys. Here, we observe only $\mathbf{y} = [y(\mathbf{x}_k)]_{k \in s}{}^{t}$, where $s \subset U$ and then, the methods should explicitly consider the sample nature of the data.

The simplest general way to consider the selection method's effect is to attach sampling weights to the regression estimates. The prediction $\hat{\tilde{y}}_k$ of the variable of interest is therefore defined as the solution of the design-based sample version of (4.24), which is the solution of the following equation:

$$\hat{\tilde{y}}_k = \mu\left(\mathbf{x}_k, \hat{\tilde{\boldsymbol{\beta}}}_{c_k}\right),$$

where $\hat{\tilde{\boldsymbol{\beta}}}_{c_k}$ is the estimate of $\tilde{\boldsymbol{\beta}}_{c_k}$. The points that are not included in the sample were predicted by classifying them into one of the classes identified with the proposed approach, according to the minimum-distance criterion.

Once the variable y is fitted by $\hat{\tilde{y}}_k$ for all $k \in U$, the model calibration estimator of the mean for non-stationary data can be defined as:

$$\hat{\bar{Y}}^{NS}_{MC} = \hat{\bar{Y}}_{HT} + \left\{N^{-1}\sum\nolimits_U \hat{\tilde{y}}_k - N^{-1}\sum\nolimits_s d_k \hat{\tilde{y}}_k\right\}\hat{\beta}^{NS}, \tag{4.25}$$

with $\hat{\beta}^{NS} = \sum_s d_k q_k \left(\hat{\tilde{y}}_k - \dddot{y}\right)\left(y_k - \bar{y}\right) / \sum_s d_k q_k \left(\hat{\tilde{y}}_k - \bar{y}\right)^2$,
$\dddot{y} = \sum_s d_k q_k y_k / \sum_s d_k q_k$

and $\bar{y} = \sum_s d_k q_k \hat{\tilde{y}}_k / \sum_s d_k q_k$.

The properties of this model calibration estimator must be assessed.

4.4.4 Model-assisted estimator for zero-inflated data

In several agricultural surveys, it is very common for the response variable to be positively skewed and to contain a substantial proportion of zeros. In literature, Zero-Inflated (ZI) models provide a way to analyse the situation when the variable of interest contains more zeros than expected. In ZI models, the response variable is modeled as a mixture of distributions. In particular, for each observation, there are two possible data generation processes; the result of a Bernoulli trial determines which process is used. For each observation *k*, Process 1 is chosen with a probability 1-p_k , and Process 2 with probability p_k. Process 1 generates only zero counts, whereas Process 2 generates counts from either a Poisson or a negative binomial model. In general:

$$\begin{cases} 0 & \text{withprobability1 - } p_k \\ f(y_k / \mathbf{x}_k) & \text{withprobability} \, p_k \end{cases} \tag{4.26}$$

The model (4.16) includes the Zero-Inflated Poisson (ZIP) model and the Zero-Inflated Negative Binomial (ZINB) model.

An alternative approach that has been suggested for the analysis of this type of data (Fletcher *et al.*, 2005; Karlberg, 2000) is given by the lognormal-logistic model. Following Fletcher *et al.* (2005), this approach works in three steps. It separately models (i) the occurrence of a zero value (as a Bernoulli random variable) and (ii) the positive abundances. We can model these two aspects of the data separately, also using different covariates. Then, the two models are combined in estimation (iii). The analysis is simpler than with the mixture-model approach, as the two models' parameters can be independently estimated.

More formally, according to the lognormal-logistic model ξ, the response variable $y_k = \delta_k y_k^*$ is the product of a lognormal component y_k^* and a logistic component δ_k. For the lognormal component, we assume that $z_k^* = \log(y_k^*), k \in U$ has a normal distribution, with $E_\xi(z_k^* \mid \mathbf{x}_k) = \mu(\mathbf{x}_k) = \alpha\mathbf{x}_k$ and $V_\xi(z_k^* \mid \mathbf{x}_k) = \sigma^2 v(\mathbf{x}_k)$, where α and σ^2 are unknown parameters and $v(\mathbf{x}_k)$ is a known function of $\mathbf{x}_k$. For the logistic component, we assume that δ_ks, $k\in U$ are distributed as independent Bernoulli (p_k):

$$P(y_k > 0 \mid \mathbf{x}_k) = P(\delta^* = 1 \mid \mathbf{x}_k) = p_k = \frac{\exp(\boldsymbol{\beta}\mathbf{x}_k)}{1+\exp(\boldsymbol{\beta}\mathbf{x}_k)},$$

where β is a vector of unknown parameters.

If we define with $s_+ = \{k \in s : y_k > 0\}$ the sub-sample for which the survey variable is positive, and with $\mathbf{Z}_{s_+}$, $\mathbf{X}_{s_+}$, and $\mathbf{V}_{s_+}$ the vectors of the logarithmic values, the matrix of auxiliary variables and the inverse of the diagonal matrix containing the variance coefficients, respectively, and defined for the strictly positive survey variable, an unbiased estimator $\hat{z}_k$ of z_k^* is given by:

$\hat{z}_k = \hat{\alpha}\mathbf{x}_k$,

where $\hat{\alpha} = (\mathbf{Z}_{s_+}\mathbf{V}_{s_+}\mathbf{X}^t{}_{s_+})(\mathbf{X}_{s_+}\mathbf{V}_{s_+}\mathbf{X}^t{}_{s_+})^{-1}$ is the ML estimation of α. Under a set of conditions of regularity, Karlberg (2000) shows that a ξ unbiased estimator $\hat{y}_k^*$ of y_k^* is given by:

$$\hat{y}_k^* = \exp(\hat{z}_k)\exp\left(\frac{\hat{\sigma}^2}{2}(v_k - \alpha_{kk}) - \frac{\hat{\sigma}^4}{4n^+}\right),$$

where $\hat{\sigma}^2 = \dfrac{\mathbf{Z}_{s_+}\mathbf{V}_{s_+}\mathbf{Z}^t_{s_+} - \hat{\alpha}(\mathbf{X}_{s_+}\mathbf{V}_{s_+}\mathbf{X}^t_{s_+})\hat{\alpha}^t}{n^+ - l - 1}$ is a bias estimator of σ^2, l is the number of auxiliary variables, $n^+ = |s_+|$ is the number of positive sample units and $\alpha_{kk} = \mathbf{X}^t_k(\mathbf{X}_{s_+}\mathbf{V}_{s_+}\mathbf{X}^t_{s_+})^{-1}\mathbf{X}_k$.

Regarding the logistic component, an asymptotically-unbiased estimator of p_k is given by $\hat{p}_k$. An approximate unbiased estimator of the survey variable is given by $\hat{y}_k = \hat{p}_k \hat{y}^*_k$.

Once the variable y is fitted by $\hat{y}_k$ for all $k \in U$, it is possible to propose a calibration estimator of the mean for zero-inflated data, whose properties must be assessed.

4.4.5 Endogenous post-stratification

Post-stratification is a method to improve the accuracy of survey estimates, both by reducing bias and by increasing precision. Post-stratification combines data collected in the survey with aggregate data on the population from other sources. In agricultural surveys, auxiliary information can be obtained from remote sensing data, classified into land cover maps. However, these maps may be based on classification models that are fitted to the sample data, in violation of the standard post-stratification assumptions that observations are classified without error into post-strata, and post-stratum population counts are known.

Breidt and Opsomer (2008) considered the post-stratification from a modeling perspective, in which population variables are treated as random variables under a model, and population quantities are based on predicted values obtained from the assumed model. Then, post-strata are constructed by dividing the range of the model predictions into predetermined intervals. They define this post-stratification of the sample data, based on categories derived from a working-model, as endogenous post-stratification.

Consider a finite population $U = \{1,2,...,k,...,N\}$. For each unit $k \in U$, an auxiliary vector $\mathbf{x}_k$ is observed. Given a scalar-valued function $\mu(.)$ and stratum boundaries $-\infty \leq \tau_0 < ... < \tau_h < ... < \tau_H \leq \infty$, we can construct the scalar index $\mu(\mathbf{x}_k)_{k \in U}$, and use it to partition the population U into H strata, where unit k is in stratum h if and only if $\tau_{h-1} < \mu(\mathbf{x}_k) < \tau_h$. If $\mu(\mathbf{x}_k)$, where for all $k \in U$, we could compute:

$$A_{Nhl}(\mu) = N^{-1} \sum_{k \in U} y_k^l I_{\{\tau_{h-1} < \mu(\mathbf{x}_k) < \tau_h\}}$$

and

$$A^*_{Nhl}(\mu) = N^{-1} \sum_{k \in s} \frac{y_k^l}{\pi_k} I_{\{\tau_{h-1} < \mu(\mathbf{x}_k) < \tau_h\}} ,$$

where *l=0,* 1, 2, the stratum index $h = 1,...,H$, and $I_{\{C\}}=1$ if the event *C* occurs, and zero otherwise. In this notation, A_{Nh0} is the population stratum proportion for the strata *h*, A^*_{Nh0} is the design-weighted sample post-stratum proportion for the strata *h*, and A^*_{Nh1} is the design-weighted sample post-stratum total of the *y* variable. The classical design-weighted Post Stratification Estimator (PSE) for the population mean is defined as:

$$\begin{aligned} \hat{\bar{y}}_{PSE} &= \sum_{h=1}^{H} A_{NH0}(\mu) \frac{A^*_{Nh1}(\mu)}{A^*_{Nh0}(\mu)} \\ &= \sum_{k \in s} \left(\sum_{h=1}^{H} A_{NH0}(\mu) \frac{N^{-1} \pi_k^{-1} I_{I_{\{\tau_{h-1} < \mu(\mathbf{x}_k) < \tau_h\}}}}{A^*_{Nh0}(\mu)} \right) y_k = \sum_{k \in s} w_k(\mu) y_k \end{aligned} \quad . \qquad (4.27)$$

The estimator (4.27) is the PSE if the function $\mu(.)$ is known for all the population units. However, the function $\mu(.)$ is unknown, and must be estimated from the sample units. Breidt and Opsomer (2008) assume a parametric model of the following type:

$$E(z_k \mid \mathbf{x}_k) = \mu(\mathbf{x}_k, \boldsymbol{\theta}), V(z_k \mid \mathbf{x}_k) = v(\mathbf{x}_k), \ k = 1,...,N$$

,

where z_k is the post-stratification variable, $\boldsymbol{\theta}^t$ are unknown model parameters, $\mu(\mathbf{x}_k, \boldsymbol{\theta})$ is a specified function of $\mathbf{x}_k$ and $\boldsymbol{\theta}$, and $v()$ is a known function of $\mathbf{x}_k$. E and V denote, respectively, the mean and the variance with respect to the specified model. Breidt and Opsomer (2008) obtain the estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}^t$ from the sample data, using standard methods.

The parametric Endogenous Post Stratification Estimator (EPSE) for the population mean is then defined as:

$$\hat{\bar{y}}_{EPSE} = \sum_{h=1}^{H} A_{NH0}\left(\mu(\mathbf{x}_k, \hat{\boldsymbol{\theta}})\right) \frac{A^*_{Nh1}\left(\mu(\mathbf{x}_k, \hat{\boldsymbol{\theta}})\right)}{A^*_{Nh0}\left(\mu(\mathbf{x}_k, \hat{\boldsymbol{\theta}})\right)} = \sum_{k \in s} w_k\left(\mu(\mathbf{x}_k, \hat{\boldsymbol{\theta}})\right) y_k ,$$

where the weights are defined as in (4.27), and they can be applied to any variable y under study. Breidt and Opsomer (2008) showed that if $\hat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}^{t}$, the parametric EPSE:

- is design-consistent under general unequal probability sampling designs and mild further assumptions;
- under a working model, is consistent and asymptotically normal;
- has the same asymptotic variance as the traditional post-stratified estimator with fixed strata.

Simulation experiments have demonstrated that the effect of first fitting a model to the survey data before post-stratifying is limited, even for relatively small sample sizes.

4.5 Calibration on complex auxiliary information

In many situations, the auxiliary information has a more complex structure than that described in previous Sections, featuring a single-phase sampling of elements and without any non-response. The complexity of the information increases with the complexity of the sampling design. In designs having two or more phases, or two or more stages, the auxiliary information may be composed of more than one variable, depending upon the structure of the design. For example, in two-phase sampling, some variables may be available in the first phase and other information in the second phase. Thus, estimation by calibration must consider the information's composite structure, to ensure the estimates' best possible accuracy.

Two-phase sampling is very frequent in agricultural surveys. Generally, in the first phase, a systematic sample is selected. Each point is then classified, using orthophotos or satellite images, into land use categories. In the second phase, a subsample is selected for the ground survey. The auxiliary information has a composite structure that the calibration estimator must take into account.

A two-phase sampling design, in its simplest form, is as follows. First, a sample s_1 is selected, and then a sample s_2 is chosen from the members of the population selected in sample s_1. The design weights are $d_{1k} = 1/\pi_{1k}$, for the sample s_1, and $d_{2k} = 1/\pi_{2k}$, for the sample s_2. The basic unbiased estimator is given by $\hat{\mathbf{t}}_y = \sum_s d_k y_k$, with $d_k = d_{1k} d_{2k}$. The auxiliary information may be available for the entire population unit and for the units belonging to the first phase sample.

That is, two different kinds of auxiliary variables may be available:

- *Population level*. The variables $\mathbf{x}_{1k}$ are known for $k \in U$; thus, the total $\sum_U \mathbf{x}_{1k}$ is known.
- *Sample level*. The variables $\mathbf{x}_{2k}$ are known only for the units in the sample s_1. The total $\sum_U \mathbf{x}_{2k}$ is estimated by $\sum_{s_1} d_{1k}\mathbf{x}_{2k}$.

Alternative formulations of the calibration problem are possible. Estevao and Särndal (2006 and 2009) illustrate some possible uses of the composite information: one-step or two-step calibration options. In the single-step option, we determine the calibration weights w_k that satisfy the condition $\sum_{s_2} w_k \mathbf{x}_k = \mathbf{t}_x$, where $\mathbf{t}_x = \begin{pmatrix} \sum_U \mathbf{x}_{1k} \\ \sum_{s_1} d_{1k}\mathbf{x}_{1k} \end{pmatrix}$ and $\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_{1k} \\ \mathbf{x}_{2k} \end{pmatrix}$. In the two-step option, we first find the weights w_{1k} such that $\sum_{s_1} w_{1k}\mathbf{x}_{1k} = \sum_U \mathbf{x}_{1k}$; we then compute the final calibration weights w_k that satisfy $\sum_{s_2} w_k \mathbf{x}_k = \mathbf{t}_x$, where $\mathbf{t}_x = \begin{pmatrix} \sum_U \mathbf{x}_{1k} \\ \sum_{s_1} w_{1k}\mathbf{x}_{2k} \end{pmatrix}$. The efficiency of the different options depends on the pattern of correlation between y_k, $\mathbf{x}_{1k}$, and $\mathbf{x}_{2k}$.

4.6 Calibration for non-response adjustment

Total non-response is an increasingly important issue affecting sample surveys. It is generally due to non-contact, refusal or inability to respond to the survey, on part of the sample units. If it is not treated, unit non-response is a source of bias, if non-respondents are systematically different from respondents with respect to the survey's characteristics of interest. Survey sampling theory has increasing needs, to address the consequences of non-response. In particular, a main issue is to examine the bias and to attempt to reduce it as much as possible.

Like all surveys, agricultural surveys too deal with non-response, regarding which the reasons are different depending upon the statistical units. Indeed, if the statistical units are farms, the total non-response is generally due to non-contact or refusal from agricultural holdings. If the statistical units are points or areas, the total non-response is due to the inability to observe the selected point or area. Consider a probability sample s selected from a finite population $U = \{1,2,...,k,...,N\}$; the known inclusion probability of the unit k is $\pi_k = P(k \in s)$, and the design weight is $d_k = 1/\pi_k$. If non-response occurs, the response set $r \subseteq s$ and the study variable yk are observed only for $k \in r$. The classical procedures for dealing with non-response consist in adjusting the design weight for non-response, based on non-response modeling.

If the unknown response probability of element *k* is defined as $P(k \in r/s) = \theta_k$, the unbiased estimator $\hat{\mathbf{t}}_{y\mathrm{NR}} = \sum_s d_k / \theta_k y_k$. Standard statistical techniques such as logistic modelling or response homogeneous groups are often used to estimate response propensities, on the basis of auxiliary covariates that are available both for respondents and non-respondents.

Calibration can also be used to construct adjusted weights for unit non-response (Särndal and Lundström, 2005 and 2008). The calibration approach for non-response consists of a reweighting scheme, which distinguishes between two different types of auxiliary variables:

- sample-level variables, which aim to remove non-response bias in survey estimates. The variables $\mathbf{x}_k^{\circ}$ must be known only for the units in the sample *s*. Contrary to simple calibration, their control totals are not required. The total $\sum_U \mathbf{x}_k^{\circ}$ is an estimate without bias by $\sum_s d_k \mathbf{x}_k^{\circ}$.
- Population level variables, which aim to reduce sampling variance. Like any usual calibration variable, the benchmark totals must be known from other sources. The variables $\mathbf{x}_k^{*}$ must be known for $k \in U$; thus the total $\sum_U \mathbf{x}_k^{*}$ is known.

The calibration can be performed by considering the combined auxiliary vectors and total information:

$$\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_k^{*} \\ \mathbf{x}_k^{\circ} \end{pmatrix}; \mathbf{t}_x = \begin{pmatrix} \sum_U \mathbf{x}_k^{*} \\ \sum_s d_k \mathbf{x}_k^{\circ} \end{pmatrix}.$$

Using the functional form, the calibration weights are $w_k = d_k F(\boldsymbol{\lambda}' \mathbf{z}_k)$. $\upsilon_k = F(\boldsymbol{\lambda}' \mathbf{z}_k)$ is the non-response adjustment factor, with the vector $\boldsymbol{\lambda}$ determined through the calibration equation $\sum_r w_k \mathbf{x}_k = \mathbf{t}_x$.

Here, $F(\boldsymbol{\lambda}' \mathbf{z}_k)$ estimates the inverse response probability $\varphi_k = 1/\theta_k$.

In agricultural surveys, there are several potential auxiliary variables. A decision must then be made as to which of these variables should be selected for inclusion in the auxiliary vector to make it as effective as possible, especially for bias reduction. Särndal and Lundström (2010) develop a bias indicator that is useful to select auxiliary variables that are effective to reduce the non-response bias. The main advantage of using calibration to deal with unit non-response is that auxiliary variables need no longer be available for the population. In addition, as there is no need for explicit response modelling, the calibration approach is simple and flexible.

4.7 Missing values in auxiliary spatial variables

So far, we have assumed that complete auxiliary information, like satellite images, land cover maps or other geo-referenced information layers, is available for all population units. However, outliers and missing data are often present in satellite information. These are mainly due to cloudy weather, that does not enable crop areas to be identified or correctly recognised from the digital images acquired. The availability of sound auxiliary data is fundamental to the success of any model-based method. Therefore, attention must be paid to its control and imputation.

Generally, missing values represent an obstacle for data analysts in drawing efficient inferences. Unless the incomplete data can be considered representative of the entire population, any statistical analysis that takes into account only the data observed risks being seriously biased. Moreover, even if the observed data and the complete data do not differ, discarding the units which present missing information is still inefficient, because of the reduction of the sample size, especially in multivariate analyses involving several variables. In general, when values are missing, three main concerns arise for survey methodologists:

- bias, because the complete case cannot be representative of the entire population;
- inefficiency, because a portion of the sample has not been observed;
- complications for the data analysis.

In the single-imputation approach, the units for which data is missing are not discarded; the incomplete records are filled in, in an attempt to preserve the distribution of the variables of interest and their relationships with other variables. However, although single imputation may save the marginal and joint distributions of the variables of interest, it does not generally reflect the uncertainty arising from the fact that the missing values have been entered, by imputed values. In other words, single imputation does not take into account the fact that the imputed values are only a guess. This may have very serious implications, especially if several observations are missing. The problems relating to single-imputation methods are well documented in Rubin (2004). The main problem of all single-imputation methods is the undercoverage of confidence intervals. The low performance in terms of coverage is due to the fact that single-imputation methods tend to understate the level of uncertainty. However, there is a vast literature on how to account for imputation in variance estimation (Rao, 1996; Kim *et al*., 2011).

An interesting alternative to single imputation are multiple-imputation methods. Rubin (2004) introduces the idea of Multiple Imputation (MI), which he describes as a three-step process:

1. for each missing datum, *m*>1 likely values are drawn from the predictive distribution $P(Y_{mis} / Y_{obs})$, where Y_{obs} is the observed part and Y_{mis} is the missing part of the data matrix;
2. *m* possible alternative versions of the complete data are produced by substituting the *i*-th simulated value, *i* = 1,…,*m*, for the corresponding missing data. The *m* imputed data sets are then analysed, using standard procedures for complete data;
3. the results are combined in such a way as to produce statistical inferences that properly reflect the uncertainty arising from missing values.

Rubin (2004) presents a procedure that combines the results of the analysis and generates valid statistical inferences. Let *Q* be a scalar parameter, such as a mean or a total. Let $\hat{Q} = \hat{Q}(Y_{obs}, Y_{mis})$ be the statistics to be used in estimating *Q* if no data were missing, and $U = U(Y_{obs}, Y_{mis})$ the squared standard error. If there are missing values, then for each unit with missing values *m*, independent simulated versions $Y_{mis}^{(1)},...,Y_{mis}^{(m)}$ are generated and *m* different data sets are created and analysed, as if they were complete. From these, we calculate *m* estimates for Q and U , that is, $\hat{Q}^{(i)} = \hat{Q}^{(i)}(Y_{obs}, Y_{mis})$ and $\hat{U}^{(i)} = \hat{U}^{(i)}(Y_{obs}, Y_{mis})$, *i* = 1,…,*m*. The combined point estimate for Q, using a multiple imputation approach, is the average of the *m* complete-data estimates:

$$\overline{Q} = m^{-1} \sum_{i=1}^{m} \tilde{Q}^{(i)} ,$$

and the estimate of the uncertainty associated with $\overline{Q}$ is given by:

$$T = \overline{U} + (1 + m^{-1})B ,$$

where $\overline{U} = m^{-1} \sum_{i=1}^{m} \hat{U}^{(i)}$ is the within-imputation variance, which is the average of the *m* complete-data estimates, and $B = m^{-1} \sum_{i=1}^{m} \left(\tilde{Q}^{(i)} - \overline{Q} \right)^2$ is the between-imputation variance, i.e. the variance among the *m* complete data estimates. Confidence intervals when Q is a scalar are based on $\overline{Q} \pm kT^{1/2}$, where the number *k* is a quantile of Student’s t-distribution, with degrees of freedom

$$\upsilon = (1-m)\left[1+\frac{\overline{U}}{(1+m^{-1})B}\right].$$

The MI method can be used to impute missing values in satellite image data. However, the method's strength depends on how the imputations are performed. To generate imputations for the missing values, it is necessary to impose a probability model upon the complete data (observed and missing values). On average, the imputation model should give reasonable predictions for the missing data, while also reflecting the uncertainty on the true value to impute. Rubin (2004) recommends that imputations can be created through a Bayesian process: a parametric model for the complete data must be specified, a prior distribution must be applied to the unknown model parameters, and m independent draws from the conditional distribution of the missing data must be simulated, given the observed data pursuant to Bayes' Theorem. In simple problems, the computations necessary for creating MIs can be performed explicitly, through formulas. In non-trivial applications, special computational techniques such as MCMC must be applied.

Once the incomplete information has been filled in, the complete data can be used for various analyses, and different working models can be assumed. The question is then if the imputation model should be related with the working model. As in the single-imputation approach, the incomplete records must be entered while trying to preserve the distribution of the variables of interest and their relationships with other variables. In this case, the statistical model used to complete the data sets should preserve the relationships between variables measured on a subject, i.e. the joint distribution in the imputed values. For this reason, it is not necessary to distinguish between dependent and independent variables; however, the variables can be treated as a multivariate response. For example, if we deal with crop area or crop yield, with satellite data as auxiliary variables, we can assume that their joint distribution is multivariate normal, i.e. $(y_k, \mathbf{x}_k) \sim N(\boldsymbol{\mu}, \Sigma)$, where $(\boldsymbol{\mu}, \Sigma)$ are unknown parameters, which may be estimated through an appropriate imputation model (Benedetti and Filipponi, 2010).

4.8 Computational issues

A great number of software packages are available for computing calibrated weights, such as the SAS macro CALMAR (Deville *et al.*, 1993), the SPSS program G-CALIB, and the functions offered by the R packages "Survey" and "regenesees"; these are R-based, but are applicable to large data sets, such as those in statistical offices, and have a user-friendly GUI (https://joinup.ec.europa.eu/software/regenesees/description).

In different ways, these packages seek to resolve computational issues such as: excluding negative weights that satisfy given calibration equations, maintaining computed weights within desirable bounds, dropping certain *x* variables to remove near-linear dependencies, reduce the weight of outlying values in auxiliary variables that may cause extreme weights. In particular, calibration in the presence of outliers is discussed in Beaumont and Alavi (2004), who present a practical way to implement M-estimators for multipurpose surveys, where the weights of influential units are modified and a calibration approach is used to obtain a single set of robust estimation weights.

Readers may find excellent reviews on the calibration estimator in Zhang (2000), Estevao and Särndal (2006), Särndal (2007), and Kim and Park (2010). Zhang (2000) presents a synthesis of the relations between post-stratification and calibration. Estevao and Särndal (2006) describe some recent progress in the field, and offer new perspectives in several non-standard set-ups, including estimation for domains in one-phase sampling, and estimation for two-phase sampling. Särndal (2007) reviews the calibration approach, with an emphasis on the progress achieved over the past decade. Kim and Park (2010) present a review of the class of calibration estimator, considering the functional form of the calibration weight.

The main advantages and drawbacks of the methods described in this topic are summarized in the following Table 4.1.

TABLE 4.1. Summary information for each sub-topic.

	Non-response	**Models for space-varying coefficients**	**Zero-inflated data**
Assessment of applicability in developing countries	The reliability of the reference data sets should be assessed.	The reliability of the reference data sets should be assessed.	These methods are particularly useful when the statistical units are points or farm units – the latter case is not very frequent in developing countries (frame not available)
Recommendations on methods proposed in literature	Incorporating relevant auxiliary variables into the $\mathbf{x}_k$-vector, to reduce the bias in the calibration estimator	Model allowing the coefficients to vary as smooth functions of an area's geographical location. Methods for the identification of local stationarity zones, i.e. post-strata.	Crops data are frequently characterized by excess zeros. Zero-inflated count models provide powerful ways to model this type of situation.
Outline of research gaps and recommendations on areas for further research	Missing value in the auxiliary vector. Variance estimation in the presence of imputation.	Estimation and test of models on remotely-sensed data	Estimation and test "zero-inflated" models on remotely sensed data

5

Benchmarking the estimators adopted for the production of agricultural and rural statistics

Agricultural and rural statistics are essential to inform policies and decisions regarding a variety of important issues, including economic development, food security and environmental sustainability.

Statistical information on land use and rural development can be derived with reference to different observation units, such as households, agricultural holdings, and parcels of land or points. Furthermore, agricultural and rural statistics can be derived through sampling and non-sampling methods. Non-sampling methods mainly include agricultural censuses and the use of administrative data collected for different purposes. Sampling methods can be based upon a list frame or an area frame, or rely on the combined use of different sampling frames. The use of administrative sources in producing agricultural and rural statistics was discussed by Carfagna and Carfagna (2010). The main advantages and drawbacks deriving from using list and area frames in agricultural sample surveys were analysed by Carfagna and Carfagna (2010), and Cotter *et al*. (2010).

The collection of data for producing agricultural and rural statistics can be based upon sampling and non-sampling methods. Non-sampling methods include agricultural censuses and the use of administrative data collected for different purposes. Sources of error in agricultural censuses have been investigated by House (2010). When a sampling survey is carried out, there are several decisions to be made, which all affect its statistical and cost efficiencies. These decisions concern the definition of land use strata, the size and the spatial scale of the sampling units, as well as the sampling strategy to be adopted (see e.g, Cotter *et al*., 2010).

Data collected from sample surveys can be used to derive reliable direct estimates for large areas, making use of auxiliary information from agricultural censuses, administrative sources, or remotely sensed data. Auxiliary information can be used before the sample selection, in designing the survey, as well as after the sample selection, during the estimation procedure (Bee *et al.*, 2010).

Several different estimators may be applied in different practical circumstances. A very important research question is the choice of the most appropriate estimator for the particular case under investigation. In addition, these different estimators should be compared. This issue is not extensively analysed in the specialist literature. Below, we provide only some of the ideas and contributions that can be applied to this context.

Sample surveys are usually designed to provide reliable estimates of finite population parameters for large areas. Design-unbiased, or approximately design-unbiased, direct estimates can be derived for these areas, when the sample size is sufficiently large. Auxiliary information from agricultural censuses, administrative sources, or remotely sensed data, can be used before the sample selection – in designing the survey – as well as after the sample selection, during the estimation procedure (Bee *et al.*, 2010). The *ex-ante* use of auxiliary information mainly concerns the construction of optimal sample stratification, and the definition of balanced sampling designs. The auxiliary information can be introduced into the estimation procedure by means of generalized regression estimators (see e.g. Cassel *et al.*, 1976; or Beaumont and Alavi, 2004) or by following the calibration approach (Deville and Särndal, 1992; Särndal, 2007). The use of calibration and regression estimators to combine information from ground data and remotely sensed data in agricultural surveys has been discussed by Gallego *et al.* (2010).

The increasing demand for statistical information on land use and rural development on a small scale is a reason for using small area estimation methods (see Section 6, below). When dealing with small areas or domains, the area-specific sample size may not be sufficiently large to support reliable direct estimates. By contrast, indirect small-area estimation enables reliable estimates of characteristics of interest to be produced, for areas or domains for which only small samples or no samples are available. While direct estimators only use data from the area of interest, small-area indirect estimators are based on either implicit or explicit models that relate small areas, such that information from other areas contribute to the estimation in a particular small area. A review of the small-area methods used in the context of agricultural surveys is available in Rao (2003). Indirect estimation, based on implicit models, includes synthetic and composite estimators. Recent developments in small-area estimation include Empirical Best Linear Unbiased Predictors (EBLUP), Empirical Bayes (EB) and Hierarchical Bayes (HB) estimation. These approaches use explicit models, categorized mainly as area-level and unit-level models, to delineate the relationships between small areas, and display advantages over traditional indirect estimators (see e.g. Rao, 2003; Pfeffermann, 2013).
Small-area estimation required the implicit or explicit use of models that connect

related areas (see e.g. Rao, 2010). Indirect estimates based on implicit models include composite and synthetic estimators.

The use of model-based small area estimators raises the question of the robustness of the inference in light of possible model misspecifications. Furthermore, when a reliable direct estimate is available for an aggregate of small areas, the model-based small area estimates must be consistent with the direct estimate for the larger area. This condition is crucial when the direct estimate for the larger area is official endorsed.

A potential difficulty with these model-based estimates is that when they are aggregated, the overall estimate for a larger geographical area may be different from the corresponding direct estimate, which is often assumed to be rather reliable. One way to avoid this problem is the so-called benchmarking approach, which consists in modifying these model-based estimates so that, when aggregated, they match the direct estimate for the larger geographical area.

A number of benchmarking procedures, intended to ensure consistency between model-based small-area estimates and direct estimates for large areas, have been developed (see e.g. Wang *et al.*, 2008; Pfeffermann, 2013). The benchmarking procedures make robust the inference forcing the model-based small-area predictors to agree with the design-based estimator for an aggregate of the areas (Pfeffermann, 2013). Denoting by θ_i a parameter of interest in area i, for i=1,…,m, and assuming that the larger area contains all the m small areas under investigation, a general formulation for the benchmarking equation is given by:

$$\sum_{i=1}^{m} w_i \hat{\theta}_{i,\text{model}} = \sum_{i=1}^{m} w_i \hat{\theta}_{i,\text{design}} , \tag{5.1}$$

where w_i, for i=1,…,m, are sampling weights, such that $\sum_{i=1}^{m} w_i \hat{\theta}_{i,\text{design}}$ is a design-consistent estimator for the total. For the condition in (5.1), the model-based predictors $\hat{\theta}_{i,\text{model}}$, i=1,…,m, provide a stable total, as a reliable direct estimator in the larger area covering all the small areas.

For the condition in (5.1), the model-based predictors $\hat{\theta}_{i,\text{model}}$, for i=1,…,m, provide a stable total (or mean) as a reliable direct estimator in the larger area covering all the small areas. Any set of estimators $\{\hat{\theta}_{i,\text{model}}\}$ that satisfies the restriction in (5.1) possesses the self-benchmarking property. Self-benchmarked small-area estimators were derived by You and Rao (2002), and Wang *et al.* (2008).

Several procedures to derive externally-benchmarked estimators have been developed. Given a small-area estimator, which does not satisfy the benchmarking condition in (5.1), a common way to achieve benchmarking is by means of a ratio-type adjustment (see e.g. Pfeffermann, 2013), that is:

$$\hat{\theta}^{\text{bench}}{}_{i,\text{ratio}} = \left(\frac{\sum_{i=1}^{m} w_i \hat{\theta}_{i,\text{design}}}{\sum_{i=1}^{m} w_i \hat{\theta}_{i,\text{model}}} \right) \cdot \hat{\theta}_{i,\text{model}} \cdot \qquad (5.2)$$

By setting $\hat{\theta}_{i,\text{model}} = \hat{\theta}^{HB}_{i,\text{model}}$ in (5.2), with $\hat{\theta}^{HB}_{i,\text{model}}$ denoting the HB estimator, a ratio benchmarked HB estimator, $\hat{\theta}^{RBHB}_{i,\text{model}}$, has been derived by You *et al.* (2004).

A general criterion for deriving benchmarked predictors for small-area parameters has been developed by Wang *et al.* (2008). According to the proposed criterion, any estimator $\hat{\theta}_{i,\text{model}}$ of θ_i, for i=1,…,m, can be adjusted as follows:

$$\hat{\theta}^{a}_{i,\text{model}} = \hat{\theta}_{i,\text{model}} + a_i \left(\sum_{i=1}^{m} b_i \hat{\theta}_{i,\text{design}} - \sum_{i=1}^{m} b_i \hat{\theta}_{i,\text{model}} \right),$$

where $\sum_{i=1}^{m} a_i b_i = 1$. The adjusted predictor $\hat{\theta}^{a}_{i,\text{model}}$ is forced to satisfy the benchmarking condition operating through the coefficient a_i, for i=1,…,m (see, also, You *et al.*, 2013). When $\hat{\theta}_{i,\text{model}} = \hat{\theta}^{BLUP}_{i,\text{model}}$, with $\hat{\theta}^{BLUP}_{i,\text{model}}$ denoting the best linear unbiased predictor, the benchmarked BLUP predictor, $\hat{\theta}^{BBLUP}_{i,\text{model}}$, is derived as follows (Wang *et al.* 2008):

$$\hat{\theta}^{BBLUP}_{i,\text{model}} = \hat{\theta}^{BLUP}_{i,\text{model}} + a_i \left(\sum_{i=1}^{m} b_i \hat{\theta}_{i,\text{design}} - \sum_{i=1}^{m} b_i \hat{\theta}^{BLUP}_{i,\text{model}} \right),$$

where $a_i = \left(\sum_{i=1}^{m} \varphi_i^{-1} b_i^2 \right)^{-1} \varphi_i^{-1} b_i$, and φ_i are chosen positive weights. A similar procedure can be applied to derive the adjusted EBLUP estimator (see You *et al.*, 2013).

The predictors described above are internally benchmarked. Externally benchmarked predictors can be also derived through an *a-posteriori* adjustment of model-based predictors. A recent review of these approaches can be found in Wang *et al.* (2008). Additional benchmarking procedures, developed in both a Bayesian and a frequentist framework, are described in Pfeffermann (2013).

Imposing the benchmarking restriction implies the possibility that the small-area model is misspecified and that the predictors are biased (Wang *et al.*, 2008). Also, self-benchmarked estimators, for which $\sum_{i=1}^{m} b_i \hat{\theta}_{i,\text{design}} - \sum_{i=1}^{m} b_i \hat{\theta}_{i,\text{model}} = 0$, are of the form $\hat{\theta}^{a}_{i,\text{model}}$.

A criterion that could be adopted to derive the set of best unbiased predictors for the small-area parameters $\boldsymbol{\theta}=(\theta_1,\ldots,\theta_m)^t$ has been provided by Wang *et al.* (2008), and can be expressed as follows:

$$Q\left(\hat{\boldsymbol{\theta}}^{a}_{\text{model}}\right) = \sum_{i=1}^{m} \varphi_i \left(\hat{\theta}^{a}_{i,\text{model}} - \hat{\theta}^{a}_{i,\text{design}}\right)^2, \tag{5.3}$$

where φ_i, i=1,…, m, denote positive weights. The criterion in (5.3) corresponds to a loss function to be minimized. The choice of weights $\boldsymbol{\phi}_i$ depends on the problem under investigation. Weights can be chosen to be a function of the variance components, or be chosen such that the derived predictors possess certain desirable properties (Wang *et al.*, 2008).

By forcing the model-based predictors to agree with the design-based estimator for the larger area, the benchmarking procedures make the inference on the small-area parameters more robust, and enable the overall bias to be reduced due to model misspecification. The reduction of bias at the small-area level was addressed by Wang *et al.* (2008). This objective was achieved by proposing an augmented model for the small areas, which results in a predictor that satisfies the self-benchmarked property. This suggests that when bias is a concern, the self-benchmarked augmented model is preferable to external benchmarking.

The use of small-area estimation is essential to help estimating detailed information, while improving reliability; benchmarking could reduce the bias and, as a consequence, improve the accuracy of the estimates of agricultural and rural statistics.

The main advantages and drawbacks of the methods described above are summarized in the following Table 5.1.

TABLE 5.1. Summary information for each sub-topic.

	Direct Estimation	Model-based Small Area Estimation
Assessment of applicability in developing countries	Estimation procedures that make use of auxiliary information must rely on reliable data from censuses, administrative sources and remote sensing.	
Recommendations on methods proposed in literature	Direct estimation requires a sufficiently large domain-specific sample. These techniques may not provide enough statistical precisión, because of inadequate sample size in small domains.	These methods use data from similar domains to estimate quantities of interest in a particular small area, assuming explicit or implicit models. They provide reliable estimates for small domains in which small samples or no samples are available. Benchmarking procedures are needed to derive small-area predictors, which agree with design-consistent direct estimates in an aggregate of the small areas.
Outline of research gaps and recommendations on areas for further research	A sample design could be developed that enables the sample size to be increased in small areas, thereby allowing for direct estimates.	Most of the benchmarking approaches proposed only adjust for the overall bias, irrespective of the bias at the small-area level. Further investigations on the benchmarking procedures could be developed.

6

Comparison of regression and calibration estimators with small-area estimators

6.1 Introduction

Regression and calibration estimator are techniques used to improve the precision of a sample estimator. However, these estimators are not sufficiently precise to produce Small-Area Estimates (SAE) of surveyed variables, due to the small sample sizes in the small area considered. The literature features several contributions seeking to increase the precision of the SAE.
To compare regression and calibration estimators, described in Section 4 above, with small-area estimators, in this Section we briefly review the problem of Small Area Estimation (SAE).

The term "small area" generally refers to a small geographical area or a spatial population unit for which reliable statistics of interest cannot be produced, due to certain limitations of the available data. For example, small areas include small geographical regions such as counties, municipalities or administrative divisions; domains or subpopulations, such as a particular economic activity or a sub-group of individuals obtained by cross-classifying demographic characteristics, are called small areas if the domain-specific sample size is small. SAE is a research topic of great importance, due to the rising demand for reliable small-area statistics even when only very small samples are available for these areas. The problem affecting SAE is twofold. The first issue is how to produce reliable estimates of characteristics of interest for small areas or domains, based on the very small samples that can taken from these areas. The second issue is how to assess the estimation error of these estimates.

In the context of agriculture, "small area" usually refers to crop areas and crop yield estimates at the level of the small geographical area. Agricultural statistics are generally obtained through sample surveys, where the sample sizes are chosen to provide reliable estimators for large areas. A limitation of the available data in the target small areas severely affects the precision of estimates obtained from area-specific direct estimators.

When auxiliary information is available, the design-based regression estimator is a classical technique used to improve the precision of a direct estimator. This technique has been widely applied to improve the efficiency of crop area estimates (Flores and Martinez, 2000), where the auxiliary information used is given by satellite image data. Unfortunately, direct area-specific estimates may not be able to provide adequate precision at the small area (SA) level; in other words, they are expected to return undesirable large standard errors due to the small, or even zero, size of the sample in question. Furthermore, when there are no sample observations in some of the relevant small domains, the direct estimators cannot be calculated.

To increase the precision of area-specific direct estimators, various types of estimators have been developed that combine both the survey data for the target small areas, and auxiliary information from sources outside the survey, such as data from a recent agricultural census, remote sensing satellite data and administrative records. Such estimators, referred to as indirect estimators, are based on (implicit or explicit) models that provide a link to related small areas by means of auxiliary data, to borrow information from the related small areas and thus increase the effective sample size. Torabi and Rao (2008) derived the model mean squared error of a GREG estimator and two-level model-assisted new GREG estimator of a small area mean. They show that, due to the borrowing of strength from related small areas, estimators based on explicit model exhibit significantly better performance compared to the GREG and the new GREG estimators.

As mentioned above, the literature features many contributions on the topic of SAE. In particular, Ghosh and Rao (1994), Rao (2002, 2003), and Pfeffermann (2002, 2013) have highlighted the main theories upon which the practical use of small area estimator is based.

To compare the performance of small-area estimators with the calibration estimators described in Section 4, we must identify the appropriate SAE, considering the different types of agricultural data. SAEs based on linear mixed models may be inefficient when dealing with agricultural data. Indeed, all the issues discussed in Section 4.4 above also apply here. In this Report, two different models will be considered:

- in the presence of variables with a high portion of values equal to zero and a continuous skewed distribution for the remaining values, we propose using zero-inflated models, as suggested by Chandra and Chambers (2008) and Chandra and Sud (2012);
- small domains are often geographical areas. An adequate use of geographic information and spatial modelling can provide more accurate estimates for small area parameters. Several attempts to generalize the Fay–Herriot model considering the correlated random area effects

between neighboring areas have been performed using a Simultaneously Autoregressive (SAR) process (Petrucci and Salvati, 2006; Pratesi and Salvati, 2008 and 2009). However, it is common in agricultural and environmental studies for the population to be divided into latent heterogeneous spatially-dependent subgroups of areas, in which the effect of covariates on the variable being studied is stationary. Here, we suggest using a local stationarity approach to estimate the parameters of the Fay–Herriot model (Benedetti et al., 2013).

The problem of outliers and missing data, often present in satellite information, is another issue that will be examined. In this connection, all the considerations concerning the appropriateness of imputing the missing covariate and of using a multiple imputation approach have been already presented in Section 4.7 and also apply also here.
Section 6.2 describes the models for small-area estimation. Sections 6.3 and 6.4 will contain a review of the main SA approaches, namely the area-level model and the unit-level model. In Section 6.5, we describe the spatial approach to SAE. The theory on zero-inflated and local stationarity in small-area models is discussed in Sections 6.6 and 6.7 respectively.

6.2 Models for small-area estimation

Indirect small-area estimates that make use of explicit models for considering specific variations between different areas have received a great deal of attention, for several reasons:

1. The explicit models used are a special case of the linear mixed model, and are thus very flexible for handling complex problems in SAE (Fay and Herriot, 1979; Battese et al., 1988).
2. Models can be validated from sample data.
3. The MSE of the prediction is defined and estimated with respect to the model.

Let us consider a partition of the population $U = \{1,2,...,k,...,N\}$ into D small sub-domains $U_1,...,U_d,...,U_D$, with N_d the size of U_d. Thus, we have:

$$U = \bigcup_{d=1}^{D} U_d \ ; \ N = \sum_{d=1}^{D} N_d \ . \tag{6.1}$$

Let y_{dk} be the study variable for area d and unit k, for $d = 1, 2,..., D$ and $k = 1,2,..., N_d$, and $\mathbf{x}_{dk} = \{x_{dk1}, x_{dk2},..., x_{dkq}\}^t$ a p-dimensional vector of auxiliary variables associated with unit k in the area d, where q is the number of the auxiliary variables. We are interested in estimating the total at the small-domain level, defined as $t_d = \sum_{U_d} y_k, d = 1,..., D$.

Small-area estimates that make use of explicit models are generally referred to as Small Area Models (SAM), and they can be broadly classified into two types: area-level models and unit-level models (Rao 2003).

6.3 Area-level models

This approach is used when area-level auxiliary data are available. Let $\mathbf{x}_d = \{x_{d1}, x_{d2},..., x_{dq}\}^t$ be the auxiliary vector at d area level, and let $\theta_d = g(t_d)$ be the parameter of interest, for some function $g(.)$.

The area-level model consists of two components: the linking model and the sampling model. In the linking model, we assume that θ_d are related to $\mathbf{x}_d$ through a linear model such as:

$$\theta_d = \mathbf{x}_d^t \boldsymbol{\beta} + b_d \upsilon_d, d = 1,2,..., D \qquad (6.2)$$

where $\boldsymbol{\beta}$ is the $q\times1$ regression parameters vector, the b_d 's are known positive coefficients and $\upsilon_d \overset{iid}{\sim} N(0, \sigma_\upsilon^2)$. The υ_d 's are area-specific random effects that represent a measure of homogeneity of the areas, after accounting for the covariates $\mathbf{x}_d$.

In the sampling model, we suppose that the direct estimator $\hat{t}_d$ or its transformation $\hat{\theta}_d = \theta(\hat{t}_d)$ is available and defined as:

$$\hat{\theta}_d = \theta_d + e_d, d = 1,2,...D, \qquad (6.3)$$

where $e_d|\theta_d \overset{ind}{\sim} N(0, \psi_d)$ are the known sampling errors. This assumption implies that the estimators $\hat{\theta}_d$ are not biased with respect to the design. Besides, the samples variances ψ_d are supposed to be known.
Combining Equations (6.2) and (6.3), we obtain the Fay-Herriott model (1979):

$$\hat{\theta}_d = \mathbf{x}_d^t \boldsymbol{\beta} + b_d \upsilon_d + e_d, d = 1,2,..., D \qquad (6.4)$$

Equation (6.4) is known as the Fay-Herriott model (Fay and Herriot, 1979). The model is a mixed linear model with two random components: the first (e_d) caused by the design, and the second (υ_d) due to the model.

To predict the random effects under the assumed mixed model (6.4), the Best Linear Unbiased Prediction (BLUP) approach is commonly used. The BLUP estimator for θ_d under Model (6.4) is (Ghosh and Rao 1994):

$$\tilde{\theta}_d = \mathbf{x}_d^t \tilde{\boldsymbol{\beta}} + \gamma_d (\hat{\theta}_d - \mathbf{x}_d^t \tilde{\boldsymbol{\beta}}) = \gamma_d \hat{\theta}_d + (1-\gamma_d)\mathbf{x}_d^t \tilde{\boldsymbol{\beta}}, \tag{6.5}$$

where $\gamma_d = \frac{\sigma_\upsilon^2 b_d^2}{(\sigma_\upsilon^2 b_d^2 + \psi_d)}, 0 \le \gamma_d \le 1$, and $\tilde{\boldsymbol{\beta}}(\sigma_\upsilon^2)$ is the is the weighted least square estimator of $\boldsymbol{\beta}$, which is defined as:

$$\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\sigma_\upsilon^2) = \left(\sum_d \mathbf{x}_d \mathbf{x}_d^t / (\psi_d + \sigma_\upsilon^2 b_d^2) \right)^{-1} \left(\sum_d \mathbf{x}_d \hat{\theta}_d / (\psi_d + \sigma_\upsilon^2 b_d^2) \right) \tag{6.6}$$

Estimator (6.5) is the Best Linear Unbiased predictor (BLUP), and is a weighted combination of the direct estimator $\hat{\theta}_d$ and the regression synthetic estimator $\mathbf{x}_d^t \tilde{\boldsymbol{\beta}}$. More weight is given to the direct estimator when the sampling variance is small compared to the total variance and more weight to the synthetic estimator when sampling variance is large or model variance is small. In practice, the BLUP estimator depends on the variance component σ_υ^2, which is generally unknown in practical applications. The most common methods for estimating model parameters are moment methods, MM (Fay and Herriot, 1979), ML and REML (Cressie 1992). Replacing σ_υ^2 with $\hat{\sigma}_\upsilon^2$, we obtain an Empirical BLUP estimator, which is known in the literature as an EBLUP estimator. The EBLUP estimator can be written as:

$$\tilde{\theta}_d^{EBLUP} = \hat{\gamma}_d \hat{\theta}_d + (1-\hat{\gamma}_d)\mathbf{x}_d^t \tilde{\boldsymbol{\beta}} \ ,$$

where $\hat{\gamma}_d = \frac{\hat{\sigma}_\upsilon^2 b_d^2}{(\hat{\sigma}_\upsilon^2 b_d^2 + \psi_d)}$.

Under the EBLUP, we use an estimate of $MSE(\tilde{\theta}_d^{EBLUP}) = E(\tilde{\theta}_d^{EBLUP} - \theta_d)^2$ as a measure of the variability of $\tilde{\theta}_d^{EBLUP}$, where the expectation is with respect to Model (6.4). A great deal of attention has been given to the estimation of the

MSE. Unfortunately, closed forms of $MSE\left(\tilde{\theta}_d^{EBLUP}\right)$ exist only in particular cases. Therefore, several scholars decided to place importance upon the identification of accurate MSE approximations.

A valid approximation for $MSE\left(\tilde{\theta}_d^{EBLUP}\right)$, if D is large, and under the assumption of the normality of the errors υ and e is (Rao, 2003):

$$MSE\left(\tilde{\theta}_d^{EBLUP}\right) \approx g_{1d}\left(\sigma_\upsilon^2\right) + g_{2d}\left(\sigma_\upsilon^2\right) + g_{3d}\left(\sigma_\upsilon^2\right), \tag{6.7}$$

where $g_{1d}\left(\sigma_\upsilon^2\right) = \gamma_d \psi_{di} < \psi_d$, $g_{2d}\left(\sigma_\upsilon^2\right) = (1-\gamma_d)^2 \mathbf{x}_d^T \left(\sum_d \mathbf{x}_d \mathbf{x}_d^t / (\psi_d + \sigma_\upsilon^2 b_d^2)\right)^{-1} \mathbf{x}_d$ and $g_{3d}\left(\sigma_\upsilon^2\right) = \left\lfloor \psi_d^2 b_d^4 / \left(\sigma_\upsilon^2 b_d^2 + \psi_d\right)^{-3} \right\rfloor \overline{V}\left(\hat{\sigma}_\upsilon^2\right)$, with $\overline{V}\left(\hat{\sigma}_\upsilon^2\right)$ as the asymptotic variance of an estimator of σ_υ^2.

Note that the main term $g_{1i}\left(\sigma_\upsilon^2\right)$ in (6.7) shows that $MSE\left(\tilde{\theta}_i^{EBLUP}\right)$ may be considerably smaller than $MSE\left(\hat{\theta}_d\right)$, if the weight γ_d is small or if σ_υ^2 is small compared to ψ_d. This means that the process of SAE depends largely upon the availability of good auxiliary information that contributes to reducing the model variance σ_υ^2 with respect to ψ_d.

The assumption of known sampling variances ψ_d can be problematic. You and Chapman (2006) consider the situation in which the sampling error variances are individually estimated by direct estimators. A full hierarchical Bayes (HB) model is constructed for the direct survey estimators and for the sampling error variances estimators.

Various methods other than EBLUP have been introduced in literature to estimate θ_d under Model (6.4). The most common are Empirical Bayes (EB) and Hierarchical Bayes (HB). These methods have been reviewed by Rao (2003), Ghosh and Rao (1994), and Pfeffermann (2002 and 2013).

Area-level models such as the Fay–Herriot model are widely used to obtain efficient model-based estimators for small areas in agricultural statistics, where a rich set of auxiliary variables is mainly available from remote sensing data. Benedetti and Filipponi (2010) addressed the problem of improving the land cover estimate at a small-area level, relating to the quality of the auxiliary information. They considered two different aspects associated with the quality of remote sensing satellite data:

1. the location accuracy of the ground survey and the satellite images;
2. outliers and missing data in the satellite information.

The first problem is addressed by using the area-level model; the small-area direct estimator is related to area-specific auxiliary variables, i.e. the number of pixels classified in each crop type according to the satellite data for each small area. The problem of missing data is addressed by using a multiple imputation.

Furthermore, various extensions have been proposed. Datta *et al.* (1991) proposed a multivariate version of the Fay-Herriot model that lead to more efficient estimators. Rao and Yu (1994) suggested an extension of Equation (6.4) for the analysis of time series and cross-sectional data.

6.4 Unit-level models

This approach is used when unit-level auxiliary data are available. The model assumes that the values of a study variable are related to unit-specific auxiliary data. More formally, if y is a continuous response variable, a basic unit-level model relates the y_{dk} to the $\mathbf{x}_{dk}$ by means of a one-fold/single nested error regression model:

$$y_{dk} = \mathbf{x}_{dk}^T \boldsymbol{\beta} + \upsilon_d + e_{dk}, d = 1,2,\ldots,D, k = 1,2\ldots,N_d, \tag{6.8}$$

where $\boldsymbol{\beta}$ is a fixed set of regression parameters, $\upsilon_d \overset{iid}{\sim} N\left(0,\sigma_\upsilon^2\right)$ are random sample area effects, and $e_{dk} \overset{iid}{\sim} N\left(0,\sigma_e^2\right)$ are the residual errors. Furthermore, the υ_ds are independent from the residual errors e_{dk}s (Battese *et al.*, 1988).

If $\overline{Y}_d$ and $\overline{\mathbf{X}}_d = \{\overline{X}_{d1},\ldots,\overline{X}_{dp}\}$ are, respectively, the population mean of a study variable and the population mean of the auxiliary variables for the area *d*, we assume that $\overline{Y}_d = \overline{\mathbf{X}}_d^T \boldsymbol{\beta} + \upsilon_d, d = 1,2,..,D$. Then, the EBLUP estimate of $\overline{Y}_d$ is:

$$\widetilde{y}_d^{EBLUP} = \hat{\gamma}_d[\overline{y}_d + (\overline{\mathbf{X}}_d - \overline{\mathbf{x}}_d)^t \widetilde{\boldsymbol{\beta}}] + \left(1 - \hat{\gamma}_d\right)\overline{\mathbf{X}}_d^t \widetilde{\boldsymbol{\beta}}, d = 1,2,..,D, \tag{6.9}$$

where $\widehat{\gamma}_d = \dfrac{\widehat{\sigma}_\upsilon^2}{\left(\widehat{\sigma}_\upsilon^2 + \widehat{\sigma}_e^2 n_d^{-1}\right)}$ and $\widetilde{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}\left(\hat{\sigma}_\upsilon^2, \hat{\sigma}_e^2\right)$ is the weighted least squares of $\boldsymbol{\beta}$; $\hat{\sigma}_\upsilon^2$ and $\hat{\sigma}_e^2$ are the estimated variance components obtained with the method of fitting constants (Battese et al., 1998) or of restricted maximum likelihood. As the small-area sample size increases, the EBLUP estimator approaches the survey regression estimator.

On the other hand, for small sample sizes and small $\hat{\sigma}_{\upsilon}^{2}/\hat{\sigma}_{e}^{2}$, the EBLUP tends towards the regression synthetic estimator.

Also, in agricultural statistics, unit-level models have been often used to obtain efficient model-based estimators for small areas. Battese *et al*. (1988) first used the unit-level model to predict areas planted with corn and soybeans for 12 counties in north-central Iowa. The areas of corn and soybeans in the 37 segments (PSUs) of the 12 counties were determined by interviewing farm operators. Each segment represents approximately 250 hectares. The sample information was integrated with auxiliary data derived from satellite imagery readings. Crop areas for each segment were estimated from satellite images, by counting the number of individual pixels in the satellite photographs. The model used assumed that a linear relationship exists between the survey data and the satellite data, with county-specific random effects.

In agricultural statistics, y_{dk} are often not continuous variables. For example, if the statistical units are sampled points, crop area estimates related to the point *k in the small area d* are given by $\boldsymbol{\delta}_{kd} = (\delta_{kd1},...,\delta_{dkj}), d = 1,...,D, k = 1,...,N_d$, where δ_{kdj} is an indicator variable with value $\delta_{dkj} = 1$ if the point *k* is classified in crop type *j*, and $\delta_{dkj} = 0$ otherwise. In this case, the SA quantities of interest are usually proportions or counts, and the mixed linear models described above are no longer applicable. MacGibbon and Tomberlin (1989) defined a Generalized Linear Mixed Model (GLMM) for SAE that is widely used for this type of problem.

Rashid and Nandram (1998) use a rank-based method to estimate the mean of county- level crop production data when the data are not normally distributed. They use the nested error regression model to borrow strength from other areas. Then, the estimates of the model parameters are used to construct a predictor of the population mean of a small area, and the mean squared error of the predictor. They apply the methodology using satellite and survey data obtained from 12 counties to estimate crop area.

Datta *et al*. (1998) consider multivariate HB prediction of small-area means using a multivariate nested error regression model. The advantages of using a multivariate approach instead of a univariate approach were demonstrated through simulations. Moreover, they analyse the corn and soybean data provided by Battese *et al*. (1988) using both multivariate and univariate models.

6.5 Extension of area-level model for the analysis of spatially autocorrelated data

Spatial autocorrelation statistics measure and analyse the degree of dependence between observations in a given geographic space. Positive spatial autocorrelation indicates the clustering of similar values across geographic space, while negative spatial autocorrelation indicates that neighbouring values are dissimilar. In the case of agricultural statistics, the statistical units are points or areas. The attribute data (crop and crop yield) is likely to exhibit some degree of spatial dependence in the form of positive spatial autocorrelation.

The spatial autocorrelation between neighbouring areas or units can be introduced into small-area estimation. A possible improvement in the EBLUP method can be achieved by including spatial structure in the random area effects (Cressie, 1991).

An area-level model with conditional spatial dependence between random effects can be considered as an extension of the Fay-Herriot model (6.4), where area-specific random effect υ_d takes into account the interaction between neighbouring areas. There are two different approaches to describing the spatial information: the conditional autoregressive model (CAR), and the simultaneous autoregressive model (SAR).

Denote with $N(d)$ the set of neighbourhoods of the small area d. For the random effect $b_d\upsilon_d$, it is possible to define the CAR spatial model:

$$b_d\upsilon_d \left| \{\upsilon_l, l \neq d\} \sim N\left(\sum_{l \neq d} c_{ld} b_d \upsilon_l, \sigma_\upsilon^2\right)\right., \tag{6.10}$$

where c_{ld} denotes spatial dependence parameters that are non-zero only if $l \in N(d)$. Cressie (1991) used a CAR model in an SAE framework in the context of US census undercount.

Now, define the Fay-Herriott model in matrix notation. Model (6.4) can be written as:

$$\hat{\boldsymbol{\theta}} = \mathbf{x}\boldsymbol{\beta} + \mathbf{B}\boldsymbol{\upsilon} + \mathbf{e}. \tag{6.11}$$

The area-specific random effects $\boldsymbol{\upsilon}$ can be defined by means of a SAR process having spatial autoregressive coefficient ρ and a d×d proximity matrix **W.** In this case, $\boldsymbol{\upsilon}$ has a covariance matrix **G** defined as:

$$\mathbf{G} = \sigma_u^2\left[(\mathbf{I}-\rho\mathbf{W})^{-1}(\mathbf{I}-\rho\mathbf{W}^t)^{-1}\right],$$

with $(\mathbf{I}-\rho\mathbf{W})$ non-singular and **e** defined as previously described.

The Spatial BLUP estimator of θ_d is obtained from Equation (6.5) as (Pratesi and Salvati, 2008):

$$\begin{aligned}\tilde{\theta}_d &= \mathbf{x}_d\hat{\boldsymbol{\beta}} + \mathbf{z}_d^t\left\{\sigma_u^2\left[(\mathbf{I}-\rho\mathbf{W})^{-1}(\mathbf{I}-\rho\mathbf{W}^t)^{-1}\right]\right\}\mathbf{B}^t \\ &\text{x}\left\{diag(\psi_i)+\sigma_u^2\mathbf{B}\left[(\mathbf{I}-\rho\mathbf{W})^{-1}(\mathbf{I}-\rho\mathbf{W}^t)^{-1}\right]\mathbf{B}^t\right\}^{-1}\left(\hat{\boldsymbol{\theta}}-\mathbf{x}\hat{\boldsymbol{\beta}}\right)'\end{aligned} \qquad (6.12)$$

where $\hat{\boldsymbol{\beta}} = \left(\mathbf{x}^t\mathbf{V}^{-1}\mathbf{x}\right)^{-1}\mathbf{x}^t\mathbf{V}^{-1}\hat{\boldsymbol{\theta}}$ and $\mathbf{z}_d^t$ is the 1x*d* vector (0,0,...,0,1,0,...,0) with 1 in the *d*-th position. The spatial BLUP is reduced to the traditional BLUP when *ρ*=0.

The spatial BLUP depends on the unknown variance σ_u^2 and ρ. Replacing these parameters with the corresponding estimators, we can define a two-stage estimator, denoted as Spatial EBLUP (SEBLUP):

$$\begin{aligned}\tilde{\theta}_d^{SEBLUP} &= \mathbf{x}_d\hat{\boldsymbol{\beta}} + \mathbf{z}_d^t\left\{\hat{\sigma}_u^2\left[(\mathbf{I}-\hat{\rho}\mathbf{W})^{-1}(\mathbf{I}-\hat{\rho}\mathbf{W}^t)^{-1}\right]\right\}\mathbf{B}^t \\ &\text{x}\left\{diag(\psi_i)+\hat{\sigma}_u^2\mathbf{B}\left[(\mathbf{I}-\hat{\rho}\mathbf{W})^{-1}(\mathbf{I}-\hat{\rho}\mathbf{W}^t)^{-1}\right]\mathbf{B}^t\right\}^{-1}\left(\hat{\boldsymbol{\theta}}-\mathbf{x}\hat{\boldsymbol{\beta}}\right)\end{aligned}$$

Assuming the normality of the random effects, σ_u^2 and *ρ* can both be estimated using ML and REML procedures. For further details on the estimation process, see Pratesi and Salvati (2008).

It is noteworthy that different estimation methods such as calibration, regression and SAE estimators, can lead to different results both in terms of the magnitude of coefficient estimates, and of the values of the relevant estimated standard error. The appropriate use of estimators depends especially upon the data available and on the objective of the analysis. For example, depending upon the availability of auxiliary information (i.e. remotely sensed images), one approach may be preferred instead of another. Furthermore, researchers should pay attention to the definition of the methods used, to compare statistical properties.
Indeed, calibration and regression estimators are model-assisted methods, and the properties must be assessed in terms of design. These estimators are design-unbiased. On the other hand, SAEs are model-based techniques, which means that the statistical properties should be analysed with reference to the model. Therefore, analysts should interpret the comparisons with caution.

However, for possible ideas on comparing SAE techniques, readers may refer to Section 5.

6.6 Small-area estimation under a zero-inflated model

Assume that a survey variable y is zero-inflated. Chandra and Chambers (2008) define small-area estimators of a survey variable y, assuming a lognormal-logistic model (Fletcher *et al.*, 2005) to accommodate for the excess of zeros in the data. The lognormal-logistic model is described in Section 4.3.4 above.

In the lognormal-logistic model, the response variable $y_k = \delta_k y_k^*$ is the product of a lognormal component y_k^* and a logistic component δ_k. Let $z_d^* = \log(y_d^*)$ be a $N_d \times 1$ vector, referred to as the lognormal component for the units within the small area d, with $\boldsymbol{N_d}$ being the number of units within the small area d, $\mathbf{x}_d$ as an N_d x m matrix of the auxiliary variables. They assume that the lognormal component z_d^* follows a linear mixed model:

$$z_d^* = \mathbf{x}_d \beta + G_d u_d + e_d ,$$

where β is the $m \times 1$ vector of fixed effects, G_d is the $N_d \times q$ matrix of known covariates, u_d is the area-specific random effect associated with area d, and e_d is a vector of individual level random errors. The two random effects are assumed to be independently and normally distributed, with zero means and variances $\mathrm{var}(u_d) = \Sigma(\theta)$ and $\mathrm{var}(e_d) = \sigma_e^2 I_{N_d}$, respectively. Here, $\mathrm{var}(z_d^*) = V_d = \sigma_e^2 I_{N_d} + G_d \Sigma(\theta) G_d^t$.

For the logistic component δ_k, the authors assume that δ_ks, $k \in U$ are distributed as an independent Bernoulli (p_k). To estimate the probability p_k in the context of SAE, they assume a generalized linear mixed model with a logit link function. The generalized linear mixed model for the small area d is:

$$logit(p_d) = \ln(p_d/(1-p_d)) = \eta_d = \mathbf{x}_d \alpha + g_d b_d ,$$

where α is a vector of fixed effects, g_d is a matrix of known covariates, and b_d is the area-specific random effect associated with area d.
The estimation of the parameters is described in Chandra and Chambers (2008) and in Chandra and Sud (2012). For the lognormal component, a second-order bias-corrected predictor for the unit $k \in d$ is:

$$\hat{\hat{y}}_k = k_k^{-1}\exp(x_d\widehat{\beta}+\frac{\hat{v}_{kk}}{2}), \tag{6.13}$$

where k_k is the bias correction factor (see Chandra and Chambers, 2008):

$$\widehat{\beta} = (\sum_{d=1}^{D}\mathbf{X}^t_{ds_+}\hat{\mathbf{V}}^{-1}_{ds_+}\mathbf{X}_{ds_+})^{-1}(\sum_{d=1}^{D}\mathbf{X}^t_{ds_+}\hat{\mathbf{V}}^{-1}_{ds_+}\mathbf{Y}_{ds_+}),$$

where $ds_+ = \{k \in s_d : y_k > 0\}$ is the sub-sample belonging to the small area *d*, for which the survey variable is positive and $\mathbf{Y}_{ds_+}$, $\mathbf{X}_{ds_+}$, and $\mathbf{V}_{ds_+}$ are, respectively, the vectors of the logarithmic values, the matrix of auxiliary variables and the inverse of the diagonal matrix containing the variance coefficients. Finally, $v_{kk} = \sigma_e^2 + G_i\Sigma(\hat{\theta})G_i$. For the logistic component, the probabilities predicted are:

$$\hat{p}_k = \exp(\mathbf{x}_k\hat{\alpha} + g_k\hat{b}_k)\{1+\exp(\mathbf{x}_k\hat{\alpha} + g_k\hat{b}_k)\}^{-1}. \tag{6.14}$$

The estimation of the unknown parameters α and b_k are described in Manteinga *et al.* (2007). Using the results in (6.13) and (6.14) at the area level, and assuming that $\hat{\hat{y}}_d$ and $\hat{p}_d$ are uncorrelated, an approximately model-unbiased predictor of the survey variable y_d is: $\hat{y}_d = \hat{p}_d\left(k_d^{-1}\exp\left(x_d\widehat{\beta}+\frac{\hat{v}_d}{2}\right)\right)$.

Further details are available in Chandra and Chambers (2008) and Chandra and Sud (2012).

6.7 Local stationarity in small area estimation models

Small-area estimators are based on the assumption that the relationships between the direct estimator and the auxiliary covariates are stationary across the area of interest. This hypothesis is clearly inadequate, however, when the population is divided into heterogeneous latent subgroups. To take into account local stationarity in the classical area-level model, the Fay–Herriot model can be modified, for the *d*-th area *d*=1,…,*D*, as (Benedetti *et al*., 2013):

$$\hat{\theta}_d = \mathbf{x}_d^t\boldsymbol{\beta}_{k_d} + u_d + \varepsilon_d, \tag{6.15}$$

where $\hat{\theta}_d$ is the direct estimator of the parameter of interest; $\mathbf{x}_d$ is the *p* × 1 vector of the area-specific auxiliary covariates; $\boldsymbol{\beta}_{k_d}$ is the local regression parameters vector *p* × 1 where $k_d \in \{1, 2, \dots, t\}$ is the generic element of $\mathbf{k} = (k_1, k_2, \dots, k_d, \dots, k_D) \subset \mathrm{d}^t$; *t* is the number of local stationarity zones; **k** is the unknown label vector representing *t* local stationarity zones related to each single area *d*; u_d is the area effect with zero mean and variance σ_u^2; and ε_d is the error term with zero mean and variance ψ_d . If the label vector ***k*** is known, the Fay–Herriot model could be fitted to each of the *t* local zones, and the BLUP estimator of θ_d would be:

$$\widetilde{\theta}_{L.d}(\sigma^2_{u.L}) = \mathbf{x}^t_d\widetilde{\boldsymbol{\beta}}_{k_d} + \sigma^2_{u.L}\mathbf{b}^t_{L.d}\mathbf{Z}^t_L\{diag(\psi_d)_L + \sigma^2_{u.L}\mathbf{Z}_L\mathbf{Z}^t_L\}^{-1}(\hat{\theta}_L - \mathbf{x}^t_L\widetilde{\boldsymbol{\beta}}_{k_d}), \quad (6.16)$$

where $\sigma^2_{u.L}$, $\mathbf{b}_{L.d}$, $\mathbf{Z}_L$, $diag(\psi_i)_L$, $\hat{\theta}_L$ and $\mathbf{x}_L$ refer to local stationarity zones related to area *d*. If q_{kd} is the number of small areas in sub-group k_d, then $\mathbf{Z}_L$ is a matrix of known positive constants of dimensions q_{kd}×q_{kd}, and $\mathbf{b}^t_{L.d}$ is a q_{kd} dimension vector of zeros and one in the *d*-th position.

From a theoretical point of view, the estimate of the heterogeneous parameters can be computed by resolving a combinatorial optimization problem. If the number *t* of local stationary zones is fixed, the aim is to define a function that must be minimized over the whole set of possible partitions of the study area into *t* groups. Since the objective of small-area methods is to obtain estimates of the study variable with small standard errors, a reasonable choice of the function is the sum of the mean squared errors of the small-area estimates. Benedetti *et al.* (2013) propose an algorithm to identify heterogeneous zones based on SA, which is described in Section 4.3.3 above, as a stochastic optimization method.

Here, the interaction term in the energy function at *j*-th iteration becomes:

$$I_j(\mathbf{x},\boldsymbol{\theta},\mathbf{k}) = \sum_{d=1}^{D} MSE_j[\hat{\theta}_d(\hat{\sigma}^2_u;\widetilde{\boldsymbol{\beta}}_{k(j)_d})],$$

where $MSE_j[\hat{\theta}_d(\hat{\sigma}^2_u;\widetilde{\boldsymbol{\beta}}_{k(j)_d})]$ is MSE at *j*-th iteration. Benedetti *et al.* (2013) estimated the MSE_j through standard formulas (see Rao, 2003). The adjacency constraints are formalized through the penalty function, for which they adopt a Potts model (Sebastiani, 2003):

$$V_j(\mathbf{k}) = -\lambda\sum_{s=1}^{D}\sum_{v=1}^{D} p_{s,v}\mathbf{1}_{(k^j_s=k^j_v)} ,$$

where $p_{s,v}$ is the element (*s*, *v*) of a binary contiguity matrix, $\mathbf{1}_{(k^j_s=k^j_v)}$ is the indicator function of the event $(k^j_s = k^j_v)$, and λ is a parameter that discourages configurations with non-contiguous zones. Then, given a starting configuration S_0, the algorithm moves from S_j to S_{j+1} with probability:

$$p = \begin{cases} 1 & \text{if} U(\hat{\theta}(\sigma_u^2), \mathbf{X}, k_{ij+1}) < U(\hat{\theta}(\sigma_u^2), \mathbf{X}, k_{ij}) \\ \exp\{-(U(\hat{\theta}(\sigma_u^2), \mathbf{X}, k_{ij+1}) - U(\hat{\theta}(\sigma_u^2), \mathbf{X}, k_{ij+1}) / T(j)\} & \text{otherwise} \end{cases}$$

.

The results of the local Fay–Herriot model, as described in Benedetti *et al.* (2013), show that introducing local stationarity into a small-area model may lead to significant improvements in the performance of the estimators.

The main advantages and drawbacks of the methods described are summarized in Table 6.1 below.

TABLE 6.1: Summary information for each sub-topic.

	Models for space-varying coefficients	**Models for non-Gaussian data**	**Missing values in auxiliary variable**
Assessment of applicability in developing countries	The reliability of the data sets referred to should be assessed.	The reliability of the data sets referred to should be assessed.	The reliability of the data sets referred to should be assessed.
Recommendations on the methods proposed in the literature	Local stationarity zones. Geographically weighted regression model allowing coefficients to vary across the area of interest.	M quantile methods. Generalized linear mixed models addressing binary and count data.	Multiple imputation methods to reflect the uncertainty arising from the imputed values.
Outline of research gaps and recommendations on areas for further research	Estimation and test of partitioning or smoothing algorithms on remotely sensed data.	Use of non-parametric methods incorporating spatial information into the M-quantile approach.	Test of the behaviour of small-area predictors under a different model used to impute the data.

7

Statistical methods for quality assessment of land use/land cover databases

Information on land cover and land use is critical when addressing a range of ecological, socioeconomic, and policy issues. Land cover refers to the composition and characteristics of land surface, whereas land use concerns the human activities that are directly related to the land. Over the past few decades, the scientific community has been committed in producing land cover and land use information at several spatial scales.

Land cover and land use information are extracted mainly from remotely sensed data, such as aerial photography or satellite imagery, by applying photo-interpretation or semi-automatic classification methods. Remotely sensed data are photo-interpreted according to a land cover legend in which the classes, or labels, correspond to land cover types. The land cover classes can be identified according to an existing classification (see e.g. Anderson *et al*., 1976), or derived from the objectives of the classification project (Congalton, 1991).

Thus, remote sensing plays a crucial role in deriving land cover and land use data. Image processing techniques, such as photo-interpretation or semi-automatic classification, when applied to remotely sensed data, enable the extraction of land cover and land use information that is usually available in a digital map format. Thematic maps depict categorical outputs, corresponding to land cover\land use patterns that may result from a crisp (i.e. traditional) classification or a fuzzy classification. In the crisp classification, each map unit is labeled as exactly one class. Fuzzy classification allows each mapped unit to have multiple or partial memberships.

In computer processing, basic classification methods involve supervised classification, in which some prior knowledge of the cover types to be mapped is assumed, and unsupervised classification, in which no prior information is required. Several alternatives to these basic approaches are available, including e.g. maximum likelihood classification (Foody *et al*., 1992), decision trees (Hansen *et al*., 1996), neural networks (Foody, 1995), and fuzzy classification (Foody 1996 and 1998). The result of photo-interpretation or semi-automatic classification is a land cover/land use database, usually in a digital map format, the basic units of which are pixels or polygons, according to a raster or a vector approach, respectively.

Recent advances in remote sensing technology have determined the availability of a large volume of data (see Section 2 for further details), and assessing the reliability of the resulting maps has become a required component of any classification project (Congalton, 1991). Information on the accuracy and reliability associated with land cover/land use maps is necessary for the map users to evaluate whether the map quality meets their specific needs. Furthermore, map accuracy assessment is crucial for map producers to detect errors and any weaknesses of a particular classification strategy (Liu *et al*., 2007).

An accuracy assessment should be performed during the data production process, to examine the quality of the classification as well as the validation of the resulting map, i.e. the assessment of the degree to which the map agrees with reality (Lunetta *et al*., 1991; Carfagna and Marzialetti, 2009a; Gallego *et al*., 2010). Often**,** the accuracy assessment is performed only after the completion of the land cover/land use map (Lunetta *et al*., 1991), and is based on the comparison of the map attributes with some reference data, for a sample of units. The reference data are gathered mainly via ground visits or aerial photography, and are assumed to be more accurate than mapped information.

When collecting reference data, ground visits tend to be preferable, compared to photo-interpretation of images, because the latter may lead to questionable results (Congalton, 1991; Nusser and Klaas, 2003). The sampling units, which represent the basis for comparing the map and the reference classification, usually consist of areal units, as pixels, polygons or fixed-area plots (Stehman and Czaplewski, 1998). Pixels and polygons are areal units that are directly associated with the map representation. Fixed-area plots correspond to some predetermined areal extent, and are usually regular in shape. There is no prevailing opinion as to the observation units that must be considered in assessing land cover/land use map accuracy. A list of sampling units employed in different classification projects is available in Stehman and Czaplewski (1998).

Stehman and Czaplewski (1998) identified three major components of land cover/land use map accuracy assessments: the sampling design, the response – or measurement – design, and the estimation and analysis protocol.

The sampling design identifies the protocol according to which the reference units (i.e. the units upon which the map accuracy assessment is based) are selected. A sampling design requires the definition of a sampling frame, which corresponds to any tools or devices used to gain access to the elements of the target population (Särndal et al., 1992, p.9). Two types of sampling frames, namely list and spatial reference, respectively, are possible. The list frame is a list of all the sampling units. The spatial reference frame consists of a list of spatial locations that provide indirect access to the assessment units. To ensure a rigorous statistical foundation for inference, the sampling design should be a probability sampling design (Stehman and Czaplewski, 1998; Stehman, 2001; Strahler et al., 2006). A probability sampling design is defined in terms of inclusion probabilities. A probability sampling design requires the inclusion probabilities to be non-zero for all the units in the population, and to be known for the units included in the sample (see e.g. Särndal et al., 1992). Requiring the inclusion probabilities to be known ensures that consistent estimates for the accuracy measures are derived. Basic probability sampling designs proposed in map accuracy assessment include simple random, systematic, stratified random, and cluster sampling (Stehman and Czaplewski, 1998).

Defining a sample design first entails deciding whether the assessment units are to be regarded as individual entities or should be grouped into clusters. The assessment units or clusters could also be grouped into strata, specified according to the classes mapped or defined by geography and space.

Furthermore, the selection protocol could be a simple SRS, or systematic sampling. Combining these different choices yields different sampling schemes. A simple random sample is obtained by treating the assessment units as individual entities, and by applying a simple random selection protocol. In systematic sampling design, the assessment units are selected according to a random starting point and to fixed intervals. Within the class of stratified sampling, a common choice in map accuracy assessment is stratified random sampling, in which the assessment units are grouped into strata, and then selected via the simple random sampling within each stratum. Usually, the stratification is based on land-cover types or geographic locations. In cluster sampling designs, clusters of pixels or polygons may be selected via simple random or systematic sampling. A two-stage cluster sampling can be also implemented, by selecting the individual assessment units within each sampled cluster. The literature presents several applications of probability sampling schemes.

Congalton (1988) verifies the performance of simple random and stratified sampling, in a simulation study. A stratified random sampling is applied to assess the accuracy of a global land cover map (Scepan, 1999), and an application of a two-stage cluster sampling can be found in Nusser and Klaas (2003). Besides these basic approaches, alternative sampling schemes have also been developed. For example, Carfagna and Marzialetti (2009a) propose a sequential sampling design for both quality control and validation of land cover databases. Different strata are identified, according to the land cover type and the size of polygons, and sampling units are selected within each stratum according to the permanent random number method.

Further issues concern the sample size, which should be selected with caution and be sufficient to provide a representative basis for assessing map accuracy (Foody, 2002). Equations for choosing the appropriate sample size have been proposed in literature (Congalton 1991). These equations are mainly based on the binomial distribution or on the normal approximation to the binomial distribution (see e.g. Fitzpatrick-Lins, 1981).

The number of the sampling units selected may be not predetermined, unlike in sequential acceptance sampling. The use of this approach in performing the quality control of land cover databases is discussed in Carfagna and Marzialetti (2009b). According to the sequential acceptance sampling, a sequence of samples is selected, and at each stage of the quality control process, the decision between terminating the inspection or selecting a further sample depends on the results obtained in the previous stage.

Besides statistical rigor, the choice of the sample size should be also driven by practical considerations directed towards develop a cost-effective accuracy assessment (Congalton, 1991; Foody, 2002).

General requirements in choosing a sample design concern its statistical rigor and cost-effectiveness. To establish a statistically rigorous basis for inference on the accuracy parameters, a basic recommendation is that the sampling design should be a probability sampling design (see e.g. Stehman and Czaplewski, 1998; Stehman, 2001; Strahler *et al.*, 2006). Probability sampling requires the inclusion probabilities to be known for the units selected in the sample. The inclusion probabilities represent the weights that must be attached to the sample units in computing the accuracy estimates, according to the principle of consistent estimation. A further requirement is that the adopted sampling strategy should produce accuracy estimators with adequate precision (Strahler *et al.*, 2006).

Strata and clusters are often employed in accuracy assessment sampling designs. Stratification targets the objective of increasing the precision of the accuracy estimates, enabling an adequate sample size to be obtained for rare land cover classes. Cluster sampling is chosen mainly due to cost-effectiveness criteria, as it reduces the time and costs associated with the acquisition of the assessment units. Combining the advantage of stratification with cluster sampling would be desirable.

A sampling design aimed at selecting the smallest sample that enables a pre-assigned precision of the accuracy estimates to be reached is the adaptive sequential sampling, developed by Carfagna and Marzialetti (2009a). In the procedure proposed, the reference units (i.e. polygons) are stratified according to land cover types and to the size of the polygons. A first stratified random sample is selected with probability proportional to the stratum size, using the permanent random numbers method (Ohlsson, 1995). The sample size for the entire area, denoted by n, is chosen to be small, and this first sample's principal objective is to produce an estimate of the standard errors of an accuracy parameter in each stratum. The standard errors estimated in the different strata are used in computing the Neyman allocation with sample size n+1. One sample unit is thus selected from the stratum presenting the maximum difference between actual allocation and the Neyman allocation. The accuracy parameter and its precision are then estimated. If the precision can be considered acceptable with respect to a pre-assigned estimate precision, the process can stop; otherwise, the procedure is repeated.

An alternative sampling design aimed at improving the precision of accuracy estimates is adaptive cluster sampling (Thompson, 1990). In the adaptive cluster sampling design, an initial probability-based sample is selected. When a variable of interest for a unit in the sample satisfies a pre-specified condition, additional units in the neighbourhood of the original unit are added to the sample. If any of these additional units still meet the pre-specified criterion, further units are included in the sample. The process stops when no units satisfying the condition of interest are found in the neighbourhoods. The final sample consists of clusters of units selected around the initial sample units. The set of all units meeting the pre-specified criterion in the neighbourhood of another defines a network. The units that were adaptively sampled, and do not meet the criterion, are called edge units. Implementing adaptive cluster sampling thus entails selecting an initial sample, specifying a criterion for performing additional sampling, and defining the neighbourhood of a sample unit. Unlike conventional sampling designs, in adaptive cluster sampling, the selection procedure depends on the observed value of a variable of interest.

This sampling design is suitable for sampling rare and spatially clustered land use/land cover types. Furthermore, adaptive cluster sampling may be usefully implemented in assessing change detection accuracy (see e.g. Stehman, 2001).
The response design refers to the protocol according to which the reference classification for the sample units is determined; (the response design can be subdivided into two parts, e.g. the evaluation protocol and the labeling protocol (Stehman and Czaplewski, 1998). The evaluation protocol consists of identifying the support region in which the reference information will be collected. The support region identifies the size, geometry, and orientation of the space in which an observation is collected (Stehman and Czaplewski, 1998).
The support region for sampling areal units does not necessarily coincide with the areal unit itself. Whatever the support region defined, the reference classification is applied only to the sampling units. Identifying the areal units (i.e. pixels, polygons or fixed-area plot) for which the reference classification will be determined is part of the response design. The labeling protocol assigns labels (classes) to the sample units. Commonly, the labels are assigned so that each sample belongs to a single land cover type (hard classification). The reference classification should be exhaustive and mutually exclusive, and capable of ensuring a direct comparison with the classification depicted in the map to be assessed (Nusser and Klaas, 2003). A hierarchical classification scheme could be conveniently applied to the reference units, so that more detailed categories can be collapsed into general categories which could be compared with the categories depicted in the map (Congalton, 1991).

The assignment of the selected units to the land cover/land use classes is part of the response design. The response design is identified by two components, the evaluation and the labeling protocol (Stehman and Czaplewski, 1998). The evaluation protocol consists of the procedures used to collect information, which contributes to determining the reference classification. The labeling protocol assigns, to the sampling units, the land cover/land use classification defined through the information collected in the evaluation protocol.

The reference classification is compared with the mapped classification in the analysis and estimation protocol. The validation of the map classification involves analysing positional as well as thematic accuracy (Lunetta *et al*., 1991; Foody, 2002). Positional accuracy, which refers to the accuracy of a unit's location in the map relating to its location in the reference data, is commonly measured in terms of root mean squared error (Lunetta *et al*., 1991). Thematic accuracy refers to the degree at which the land cover types depicted in the map agree with the land cover types in the reference data. Two main types of thematic error, the omission and the commission errors, can be identified. The omission error occurs when a case belonging to a class is not allocated to that class. The commission error occurs when a case belonging to a class is erroneously allocated to another class. Thematic accuracy is commonly measured by accuracy metrics, derived from a confusion or error matrix. The confusion matrix summarizes the correct classifications and the misclassifications for the sample units, and is constructed as a square matrix whose rows and columns indicate the land use/land cover classes of interest in the map and in the reference classification, respectively. An example of confusion matrix is reported in Table 7.1 below.

Potential sources of error in the reference data are discussed by Congalton and Green (1999). The effects of errors in reference data on the accuracy assessment estimates are investigated by Verbyla and Hammond (1995) and Hammond and Verbyla (1996). Given the inherently interpretive nature of the reference land cover/land use classification, an accuracy assessment of the response design could be also required (Stehman and Czaplewski, 1998).

As mentioned above, two main types of thematic errors, the omission and the commission errors, can be identified. Thematic accuracy is typically assessed through a confusion or error matrix. The confusion matrix summarizes the correct classifications and the misclassifications in a contingency table format. Usually, the rows of the confusion matrix represent the map labels, and its columns identify the reference labels.

Congalton (1991) identifies four major historical stages of map accuracy assessment. In the first stage, accuracy assessment is simply based upon a visual analysis of the derived map. The second step is characterized by a non-site-specific accuracy assessment in which the areal extent of the land cover classes depicted in the derived map is compared with the areal extent of the same classes in the reference data. In the third stage, accuracy assessment is based on the calculation of accuracy metrics, derived mainly from the comparison between the classes depicted in the map and the reference data at specific locations. The fourth step focuses on the confusion matrix, which is still widely used in land cover/land use map accuracy assessment, being the starting point of a series of descriptive and analytical statistical techniques (Congalton, 1991; Foody, 2002).

An example of confusion matrix is reported in Table 7.1 below. The entry of the confusion matrix, p_{ij}, denotes the proportion of the area in the mapped land-cover class i and the reference land-cover class j, for i, j=1,…,m. The row total p_{i+} identifies the proportion of the area mapped as land-cover class i, and the column total p_{+i} represents the proportion of the area classified as land cover class i in the reference data, for i=1,…,m.

TABLE 7.1. Confusion matrix.

Map	Reference				
	1	2	…	k	Total
1	p_{11}	p_{12}	…	p_{1k}	**p_{1+}**
2	p_{21}	p_{22}	…	p_{2k}	**p_{2+}**
…	…	…	…	…	**…**
k	p_{k1}	p_{k2}		p_{kk}	**p_{k+}**
Total	**p_{+1}**	**p_{+2}**	**…**	**p_{+k}**	

These proportions can be derived from pixels or polygon counts, or by measurement (Stehman and Czaplewski, 1998). The entries of the confusion matrix can be also reported in terms of counts rather than proportions (see e.g. Congalton, 1991; Foody, 2002).

The entries of the confusion matrix must be estimated on the sampled units, in order to obtain estimates of accuracy parameters. Several accuracy measures can be derived from the confusion matrix. There is no single standard approach to land cover/land use maps accuracy assessment; each accuracy measure, being suitable for a particular purpose, incorporates specific information on the confusion matrix (Liu *et al.*, 2007).

The entries of the confusion matrix can also be reported in terms of counts rather than proportions (see e.g. Congalton, 1991; Foody, 2002). The confusion matrix is assumed to be the starting point for estimating the accuracy of individual classes and the overall map accuracy.

The overall accuracy P_o expresses the probability that a randomly selected unit is correctly classified by the map, and is defined by the sum of the diagonal entries of the confusion matrix, that is:

$$P_o = \sum_{i=1}^{k} p_{ii} . \tag{7.1}$$

The overall accuracy expresses the probability that a randomly selected unit is correctly classified by the map, and provides a measure of the quality of the map as a whole.

The accuracy of individual land cover/land use classes may be also assessed. Story and Congalton (1986) distinguished between *producer's accuracy* (P_{Ai}) and *user's accuracy (P_{Ui}),* which are respectively computed as follows:

$$P_{Ai} = \frac{p_{ii}}{p_{+i}} \qquad P_{Ui} = \frac{p_{ii}}{p_{i+}},$$

for i=1,…,m. The producer's accuracy for land cover/land use class i expresses the conditional probability that a randomly selected unit classified as category i by the reference data is classified as category i by the map. It is referred to as producer's accuracy, because the producer of a land cover/land use map is interested in how well a reference category is depicted in the map. The user's accuracy for land cover/land use class i expresses the conditional probability that a randomly selected unit classified as category i in the map, is classified as category i by the reference data. The row and column totals can be also used to quantify the probabilities of omission and commission errors, which are respectively given by p_{ij}/ p_{+j} and p_{ij}/ p_{i+}, for i, j=1,…,m.

Further category-level and map-level accuracy measures have been proposed in the literature. A comprehensive review of these measures is available in Liu *et al.*, (2007). To compute the accuracy measures from the confusion matrix reported in Table 7.1 above, the proportions p_{ij} are estimated from the sample units. The inclusion probabilities, which define the adopted sampling design, must be included in the proportion estimates $\hat{p}_{ij}$, for i, j=1,...,m. If the sample units have been selected according to the simple random sampling, the proportion estimates can be computed as $\hat{p}_{ij} = n_{ij} / n$, where n_{ij} is the number of units classified as class i by the map and as class j by the reference data, for i,j=1,…,m, and n is the total number of sample units in the confusion matrix. If a stratified

random sampling has been employed, the proportion estimates are given by $\hat{p}_{ij} = (n_{ij} / n_{i+})(N_{i+} / N)$, where n_{i+} and N_{i+} represent the sample and population sizes in stratum *i*, respectively, and *N* denotes the population size.
Other accuracy measures can be computed by replacing pij with $\hat{p}_{ij}$ in the corresponding formulas. These estimation approaches lead to consistent estimators of the parameters of interest (Stehman and Czaplewski, 1998). Furthermore, since most accuracy measures are expressed as totals, they could be estimated with the HT estimator (Stehman, 2001). Specific guidelines to implement consistent estimators for accuracy parameters are also given by Strahler et al., (2006).

Given the known inclusion probabilities of the probability sampling designs, it is possible to construct consistent estimators for the accuracy metrics.

To derive consistent estimators, the accuracy parameters could be expressed as a function of population totals, and each total could be estimated using the HT estimator (Stehman, 2001). Accuracy estimators can be also derived by first estimating the entries p_{ij} of the confusion matrix (Strahler *et al.*, 2006).

The accuracy estimators derived for the stratified random sampling can be used in the adaptive sequential procedure developed by Carfagna and Marzialetti (2009a). The consistency of estimators poses certain problems when adaptive cluster sampling is used. In this type of sampling, the inclusion probabilities cannot be determined for those units that do not satisfy the pre-established condition and that are not included in the initial sample. Therefore, the inclusion probabilities generally cannot be determined for all the units in the final sample; it is for this reason that the modified version of the HT and the Hansen-Hurwitz estimators are derived (Thompson, 1990).

The definition of variance estimates associated with the estimated accuracy measures is extensively analysed by Czaplewski (1994). An approach to variance estimation is discussed by Stehman (1995), and a general formula for the variance estimator is provided by Strahler *et al.*, (2006).

Other than the computation of the described accuracy metrics, the confusion matrix represents the appropriate starting point for the performance of analytical statistical techniques (Congalton, 1991). Discrete multivariate techniques have been proposed, appropriate for remotely sensed data (Congalton, 1991). An analytical technique, which relies on the normalization of the confusion matrix, can be used for comparing different matrices (Congalton, 1991). Normalizing the confusion matrix is implemented by iterative proportional fitting, and results in a matrix in which rows and columns add up to a common value. Normalizing the error matrix eliminates differences in sample size, thus making the entries of the matrix comparable, and enabling direct comparison of entries of different matrices. This approach is criticized by Stehman and Czaplewski (1998) and

Foody (2002). A normalized confusion matrix could lead to accuracy parameter estimates that violate the consistence criterion, and tend to equalize accuracy measures, such as the user's and the producer's accuracies, which may instead differ significantly.

A different multivariate statistical technique used in map accuracy assessment is the Kappa analysis (Congalton, 1991), which focuses on the computation of the KHAT statistic, i.e. the maximum likelihood estimator of the Kappa coefficient of agreement (Cohen, 1960). Formulas for the KATH estimator and its standard error have been provided by Bishop *et al.*, (1975), under the assumption of multinomial sampling. The KHAT value enables determination of whether the results in the error matrix are significantly better than a random result (Congalton, 1991). This accuracy measure incorporates more information than does overall accuracy, also involving the off-diagonal elements of the error matrix in its computation. An estimator for the Kappa coefficient under stratified random sampling has been derived by Stehman (1996).

Therefore, the quality of the map as a whole is represented by the Kappa coefficient of agreement (Cohen, 1960). The Kappa coefficient is based upon the comparison between the proportion of cases in agreement (i.e. those correctly allocated) and the proportion of agreement expected to have arisen by chance. It is expressed by:

$$Kappa = \frac{P_o - P_c}{1 - P_c}, \tag{7.2}$$

where P_o is defined as in Equation (7.1), and $P_c = \sum_{i=1}^{k} p_{i+} p_{+i}$.

Accuracy measures computed from the confusion matrix rely on a hard classification, in which each unit is assigned to a single land cover/land use category. By contrast, pixels, and other areal units, may exhibit membership of multiple categories.

Some attempts to solve the mixed pixels problems have been made. Fuzzy classification, as opposite to hard classification, allows each pixel to have multiple or partial memberships. The fuzzy set theory (Zadeh, 1965) was introduced into the map accuracy assessment process by Gopal and Woodcock (1994), who define a linguistic measurement scale to evaluate map products relative to reference data. Starting from five linguistic categories, ranging from "absolutely wrong" to "absolutely right", the authors derived a number of fuzzy measures (see also Woodcock and Gopal, 2000). A review of the different methods used for assessing maps based on fuzzy classification, and the techniques for assessing a hard classification scheme using fuzzy-class reference data, is provided by Foody (2002).

Although there exist well-established methods for assessing the quality of land cover/land use maps quality, there are several research areas to be further developed. Some research priorities are highlighted by Strahler *et al.* (2006). These mainly concern the standardization of land cover map legends, to promote comparisons, the integration of errors related to the reference data in map accuracy assessment, and the improvement of the existing validation methods to better meet the specific needs of map users.

Finally, it should be highlighted that a more informed use of land cover/land use maps requires a detailed description of the accuracy assessment approach adopted (Foody, 2000). This means that it becomes necessary to report, in addition to quantitative metrics, information on the sampling design and on the reliability of reference data.

7.1 Land cover change analysis

The confusion matrix, traditionally used in map accuracy assessment, could be usefully implemented in land cover change analysis.

Several techniques are available to detect change occurring in land cover classes. Basic approaches include a comparative analysis of independently produced land cover classifications on two different dates (post-classification comparison), and simultaneous analysis of multitemporal data (Serra *et al.*, 2003). Regarding the latter approach, several procedures, such as image differencing, principal component analysis and change vector analysis, have been developed (Serra *et al.*, 2003; Lu *et al.*, 2004).

Post-classification comparison is the most commonly used technique to detect change, because of its applicability to available single-date classifications (Foody, 2002). It enables identification of land cover changes between two time periods, and requires high accuracy of the classifications to be compared (see e.g. Foody, 2002).

Classifications at two different time periods can be compared through a confusion matrix modified for the purpose. The matrix that summarizes land cover changes occurring between the time periods t_1 and t_2 is reported in Table 7.2 below.

TABLE 7.2. Change detection matrix.

t t_1	t_2				
	1	2	…	k	Total
1	n_{11}	n_{12}	…	n_{1k}	n_{1+}
2	n_{21}	n_{22}	…	n_{2k}	**n_{2+}**
…	…	…	…	…	**…**
k	n_{k1}	n_{k2}		n_{kk}	**n_{k+}**
Total	**n_{+1}**	**n_{+2}**	**…**	**n_{+k}**	**n**

The matrix rows report the land cover classes at time t_1, and the matrix columns refer to the classes identified at time t_2; the diagonal entries of the matrix, n_{ii}, identify the number of areas classified as land cover class i in both the time periods under investigation, for i=1,…,k. The off-diagonal elements of the matrix, n_{ij}, express the number of units classified as land cover class i at time t_1, and as land cover class j at time t_2, for $i{\neq}j$, i, j=1,…,k.

The row totals, $n_{i+} = \sum_{j=1}^{k} n_{ij}$, express the number of areas classified as land cover class i at time t_1, for i=1,...,k. The column total, $n_{+j} = \sum_{i=1}^{k} n_{ij}$, expresses the total number of areas classified as land cover class j at time t_2, for j=1,...,k. Land cover change indices may be computed starting from the conditional distributions derived from the change detection matrix.

The conditional distribution for land cover class i at time t_1 is derived as shown in Table 7.3 below.

TABLE 7.3. Conditional distribution of the *i*-th row of the change detection matrix.

t_1	1	2	…	i	…	k	Total
i	n_{i1}/n_{i+}	n_{i2}/n_{i+}	…	n_{ii}/n_{i+}	…	n_{ik}/n_{i+}	1

For each class i, a land cover change index can be derived as follows:

$I_{VAR}(i)$= 1 - n_{ii}/n_{i+}, for i=1,…,k. (7.3)

The land cover change index in (7.3) ranges from 0 to 1. Specifically, $I_{VAR}(i)$=0 if no area, classified as land cover class i at time t_1, is differently classified at time t_2, and $I_{VAR}(i)$=1 otherwise. This index provides information on the number of areal units affected by a change from the land cover type j at time t_1 to the land cover type $i \neq j$ at time t_2, for i, j=1,…,k.

The conditional distribution for land cover class j at time t_2 is derived as reported in Table 7.4 below.

TABLE 7.4. Conditional distribution of the j-th column of the confusion matrix.

t_2	1	2	…	i	…	k	Total
j	n_{1j}/n_{+j}	n_{2j}/n_{+j}	…	n_{ij}/ n_{+j}	…	n_{kj}/ n_{+j}	1

Starting from this conditional distribution, the land cover change index can be derived as follows:

$I_{VAR}(i)$= 1 - n_{jj}/n_{+j} for i=1,…,k. (7.4)

The land cover change index in (7.4) provides information on the number of areal units interested by a change from land cover class i at time t_1, to land cover class $j \neq i$ at time t_2, for i, j=1,…,k.

The indices in (7.3) and (7.4) enable the detection of changes for a single land cover class. A global index, which considers all the k land cover classes, can be defined as:

$$I_{VAR} = 1 - \sum_{i=1}^{k} \frac{n_{ii}}{n}.$$

Additional information on land cover change could be acquired by using the Kappa coefficient as defined in (7.2), where P_o and P_c are estimated starting from the change detection matrix given in Table 7.2 above.

Additional issues arise from assessing the accuracy of change detection techniques. The accuracy of the maps involved in post-classification comparison must be assessed. Results from the accuracy assessment performed on the classifications on individual dates can be summarized in the binary change detection error matrix (see e.g. van Oort, 2007) reported in Table 7.5 below.

TABLE 7.5. Binary change detection error matrix.

	Correct at t_2	Incorrect at t_2	*Total*
Correct at t_1	u_{11}	u_{12}	$\boldsymbol{u_{1+}}$
Incorrect at t_2	u_{21}	u_{22}	$\boldsymbol{u_{2+}}$
Total	$\boldsymbol{u_{+1}}$	$\boldsymbol{u_{+2}}$	**1**

Three accuracy metrics are included in the binary change detection error matrix: the overall accuracy at time t_1, denoted by u_{1+}; the overall accuracy at time t_2, expressed by u_{+1}; and the overall detection accuracy, given by u_{11}.

An alternative specification of the change detection error matrix is given by Macleod and Congalton (1998). This matrix, also known as full transition error matrix (see van Oort, 2007) has the same characteristics as the single date confusion matrix reported in Table 7.1 above, but also assesses errors in changes between the two time periods under investigation (Macleod and Congalton, 1998).

An example of full transition error matrix constructed for two land cover categories (A and B) is reported in Table 7.6 below.

TABLE 7.6. Full transition error matrix.

Test Data		***Reference Data***			
		No Change		Change	
		AA	BB	AB	BA
No Change	AA	n_{AAAA}	n_{AABB}	n_{AAAB}	n_{AABA}
	BB	n_{BBAA}	n_{BBBB}	n_{BBAB}	n_{BBBA}
Change	AB	n_{ABAA}	n_{ABBB}	n_{ABAB}	n_{BAAB}
	BA	n_{BAAA}	n_{BABB}	n_{BAAB}	n_{BABA}

The matrix's diagonal entries indicate the correct classifications. The off-diagonal elements of the matrix provide information on the different types of confusion in the classification. The change detection error matrix enables determination of the accuracy of change detection techniques and, as its main advantage, enables application of standard accuracy techniques that are already available for single-date accuracy assessment.

The full change detection error matrix can be collapsed into the change/no change error matrix reported in Table 7.7 below. The elements of this matrix are derived by adding the cells in the appropriate sections of the change detection error matrix.

TABLE 7.7. Change/No Change error matrix.

Test data	**Reference data**		
	No Change	Change	Total
No Change	x_{11}	x_{12}	x_{1+}
Change	x_{21}	x_{22}	x_{2+}
Total	x_{+1}	x_{+2}	1

The diagonal entries of the change/no change error matrix indicate the proportion of area for which change (no change) in land cover classes is reported in both the test data and the reference data. The sum of these diagonal entries provides information on the detection accuracy (van Oort, 2007). The off-diagonal elements of the matrix indicates misclassifications, i.e. the number of areal units for which land cover classes are classified as unchanged (changed) in the test data, and as changed (unchanged) in the reference data.

The change/no change error matrix enables quantification of how accurately the change was distinguished by no change in the test data. To further analyse whether the errors are due to the classification or to the change detection technique, a condensed change detection error matrix can be constructed (van Oort, 2007). An example is given in Table 7.8 below.

TABLE 7.8. Condensed transition error matrix.

Test Data	***Reference Data***				
	No Change		**Change**		
	Correct class	**Incorrect class**	**Correct class**	**Incorrect class**	**Total**
No Change	y_{11}	y_{12}	y_{13}		$y_{1+}= x_{1+}$
Change	y_{21}		y_{22}	y_{23}	$y_{2+}= x_{2+}$

While in the change/no change error matrix, both the elements nAAAA and nAABB of the full transition error matrix are assigned to x11 (true change), in the condensed transition matrix, they are partitioned into y11 (true change, correct class) and y12 (true change, incorrect class). On the basis of the condensed change detection error matrix, the change detection accuracy can be computedas y11+y22.

When standard accuracy metrics are computed from the change detection error matrix, it is implicitly assumed that errors at individual date are independent. The introduction of temporal correlation in change detection error matrices is discussed by van Oort (2007).

8

Assessment of the applicability of the methods developed in developing countries

Remote sensing includes techniques that serve to register data on the condition of the Earth's surface, for example by using sensors carried on board of aircraft or satellites. Remote sensing is applied extensively, for example, in resource management in agriculture, forestry and fishery, exploration of raw material deposits, obtaining environmental information, monitoring urban development, and catastrophe management.

The technical development of satellite sensors as well as GIS will surely further extend the applicability of remote sensing in the field of agriculture and, in general, of earth monitoring in the near future. In developing countries, the importance of the use of remote sensing in monitoring social, environmental, and economic issues is even more important, due to the presence of increasing social (urban development) and ecological problems (desertification, erosion, etc.).

The advantages of remote sensing are, in particular, the ability to record data on inaccessible areas, and comparatively lower costs per unit of recorded area. These characteristics make it highly attractive for developing countries.

Indeed, remote sensing and GIS have great potential in the land use and land cover mapping of developing countries. Remotely sensed data can be used for effective planning and decision making at local, regional, national, and international levels. In particular, high-resolution satellite imagery provides cost-effective information with extensive spatial coverage and spectral information, and with a high repetitive cycle to investigate temporal changes of land use and land cover. Aerial photographs are also a valid platform for up-to-date surveying, but such photography is usually more expensive, and requires more resources for wide area mapping. Furthermore, remote sensing may be an appropriate approach in developing countries, to handle, store, and use different typologies of spatial data.

The main problem is that remote sensing-based mapping implemented with aerial photography as well as with satellite imagery and ground inventory often requires great and distinctive resources in terms of skills, hardware, and software that are often insufficient in developing countries. In addition, the management and distribution of data can be problematic and delayed by ineffective computational capacity and the relatively sparse distribution of Internet services in developing countries.

As highlighted by this brief discussion, the use of remote sensing as auxiliary information in the phases of design, estimation, and control is particularly recommended in developing countries. This powerful tool does not present any counterintuitive and general drawbacks to its use in specific countries and developing regions. The problem is the availability of such information, which is sometimes unsatisfactory in developing countries. The cost of images and expenditure for the training of local individuals may limit the chances to use such techniques. However, the launch of new satellites (see Section 2 of this Report, above) that, in the near future, will provide images with good spatial resolution and high visit frequency should dramatically improve the situation in developing countries.

Thus, the applicability of the methods proposed in this Report depends on the availability of satellite information. We maintain that this, at least for the near future, will also hold true for the developing countries. Moreover, qualified personnel and software to apply the methods proposed are needed. Below, we outline some of the main particular issues concerning the applicability of the methods proposed to developing countries.

As for the use of remote sensing data at the design level, it must be noted that the sample selection with probability proportional to size, when a size measure is multivariate, is widely used for households frames, while in developing countries, polygon and point frames are more common (see Section 3.1 of this Report). In the case of the extension of regression or calibration estimators, when dealing with zero inflated data, these methods are particularly useful when the statistical units are points or farm units (the latter case is not very frequent in developing countries; see Section 4.3.4 of this Report). Besides, the robustness of the estimators adopted for producing agricultural and rural statistics, which make use of auxiliary information, must rely on reliable data from censuses, administrative sources, and remote sensing. This occurrence is not always verified in developing countries (see Section 5 of this Report).

Finally, the sampling designs used for the quality assessment of land use/land cover databases should achieve cost-efficiency criteria (see Section 7 of this Report).

9

Concluding remarks

In this Report, we have presented a further step in assessing the possibility of using remotely sensed images in sampling design and estimation. Satellite and/or aerial remote sensing technology, in combination with in-situ observations, are a very important tool in enhancing the monitoring system of the earth and, in particular, of agriculture. Remote sensing provides information that is available for several countries at a certain spatial and temporal resolution. We have sought to identify possible solutions to the gaps outlined in Section 1. Our proposal can be easily applied to the developing countries that do not present unusual characteristics. Applicability depends essentially on the availability of images, which may not always be satisfied in developing countries. We consider that new satellite technology will provide images that will solve this problem in the very near future. A synopsis of the recommendations and the suggestions on the methods proposed in this Report is given below, in Tables 9.1, 9.2, 9.3 and 9.4.

TABLE 9.1. Applicability in developing countries (XXX – Suggested; XX – Can be used; X – Undesirable use).

	Data Frame			
Research Activity / Methods	**List (Farms)**	**Irregular Polygons**	**Regular Polygons**	**Points**
Methods for using remote sensing data at the design level				
Sample selection with probability proportional to size when a size measure is multivariate	XXX	XX	XX	X
Optimal stratification with multivariate continuous auxiliary variables	XXX	XX	XX	X
Spatially balanced samples	X	XX	XXX	XXX
Models linking survey variables with some auxiliary variables to design the sample	XXX	XXX	XXX	XX
Extension of the regression or calibration estimators				
Non-response	XXX	XXX	XXX	XXX
Models for space-varying coefficients	X	XXX	XXX	XXX
Zero-inflated data	XXX	XX	XX	XXX
Robustness of the estimators adopted for producing agric. and rural statistics				
Direct Estimation	XXX	XXX	XXX	XXX
Model-based Small Area Estimation	XX	XX	XX	XXX
Comparison of regression and calibration estimators with small area estimators				
Models for space-varying coefficients	XX	XX	XX	XXX
Models for non-Gaussian data	X	X	X	XXX
Missing values in the auxiliary variable	X	XX	XX	XX
Statistical methods for quality assessment of land use/land cover databases				
Sampling Design	X	XXX	XXX	XXX
Response Design	X	XXX	XXX	XXX
Analysis	X	XXX	XXX	XXX

TABLE 9.2. Summary of Activities and Benefits.

Research Activity / Methods	Currently used	Proposal	Advantages
Methods at the design level			
Multivariate πps	Maximal Brewer	Optimal Linear Combination	Theoretical properties: Flexibility More reasonable distribution of weights
Optimal stratification	Sequence of thresholds chosen by eye or by trial and error	Optimal Algorithm given H	Considerable reduction of variance (over 60%, depending on the auxiliary)
Spatially balanced samples	Systematic when possible, or rules that rarely allow an efficient use of HT (one unit per stratum)	Spatially Balanced Sampling, in particular samples with prob. prop. to the within distance	Theoretical properties: Flexibility Considerable reduction of variance (0-100% on simulated data, and over 50% on real data depending on the homogeneity of the survey variable)
Calibration estimators			
Non-response	RHG on original strata	Calibration	Theoretical properties Flexibility Considerable reduction of variance depending on the aux (Remote Sensing)
Models for space-varying coefficients	Theory exists but not applied in this field	Partition (post-stratification) on Calibration models	Considerable reduction of variance
Zero-inflated data	Theory exists but not applied in this field	Model Calibration	Better fit of the models to the auxiliary data
Small Area			
Area Level Models	The theory exists but the operational use is very limited in spatial agricultural surveys	Simply an FH model	Promising results when used on a point frame (correlation at an area level, avoid geom. errors)

The main issues concerning the topic of statistical methods for quality assessment of land use/land cover databases are summarized in Table 7.9 below.

TABLE 7.9. Summary information for each sub-topic.	**Sampling Design**	**Response Design**	**Analysis**
Assessment of applicability in developing countries	Sampling designs should meet cost-efficiency criteria.	The reliability of the reference data sets should be assessed.	
Recommendations on methods proposed in literature	The sampling design identifies the protocol to be followed when selecting the locations at which the reference data are obtained. Probability sampling designs display advantages over non-probability sampling designs, in terms of the statistical rigor of the inference.	The response design refers to the protocol according to which the reference classification for the sample units is determined. The reference classification should be capable of ensuring a direct comparison with the classification depicted in the land cover/land use map to be assessed.	The analysis protocol consists of constructing estimators for accuracy metrics at map level and category level. Accuracy metrics have been mainly derived for hard classifications, relying on the use of confusion matrices. The inclusion probabilities, which define the sampling design used, must be included in the accuracy measure estimates.
Outline of research gaps and recommendations on areas for further research	Since reference data might be useful for multiple accuracy analyses, greater efforts should be directed towards the definition of sampling designs amenable to multiple uses, which meet both precision and cost efficiency criteria.	A protocol for the accuracy assessment of the response design should be developed.	The accuracy assessment for soft classifications must be investigated further.

TABLE 9.3. Research requirements.

Research Activity / Methods	**Theoretical development**	**IT development**	**Empirical comparison**
Methods at the design level			
Multivariate π ps	XXX	XXX	XXX (vs MaxBrewer & Balanced - Cube)
Optimal stratification	XXX	XXX	XXX
Spatially balanced samples	XXX	XXX	XXX
Calibration estimators	X	X	XXX (vs Regression Estimator)
Non-response	X	X	XXX (vs Regression Estimator)
Models for space-varying coefficients	XXX	XXX	XX (particular data sets are needed)
Zero-inflated data	XXX	X	XX (particular data sets are needed)
Small Area			
Area Level Models	X	X	XXX (vs Regression Estimator)
Benchmarking	X	XX	XXX (vs Regression Estimator)

XXX – High effort
XX – Medium effort
X – Low effort

TABLE 9.4. Research contribution, training and field test.

	Contribute to:				
Research Activity / Methods	**Country's ability to provide estimates**	**Country's development of a master sampling frame and its implementation**	**Guidelines for technical assistance and training**	**Training and/or technical assistance**	**Field tested**
Methods at the design level					
Multivariate π ps	XXX	XXX	Prepare guidelines based on good practices and findings of research for development of master sampling frames, integration of surveys, improved estimation practices, and use of administrative data Prepare guidelines based on good practices and findings of research for use of remote sensing, global positioning systems, statistical software and portable data entry devices Prepare guidelines based on good practices and findings of research for sample design, data collection, estimation and analysis.	X	XXX (vs MaxBrewer & Balanced - Cube)
Optimal stratification	XXX	XXX		X	XXX
Spatially balanced samples	XXX	X		XX	XXX
Calibration estimators	XXX	X		XXX	XXX (vs Regression Estimator)
Non-response	XXX	X		XXX	XXX (vs Regression Estimator)
Models for space-varying coefficients	XX	XX		XX	XX (particular data sets are needed)
Zero-inflated data	XX	X		XX	XX (particular data sets are needed)
Small Area					
Area Level Models	XXX	X		XXX	XXX (vs Regression Estimator)
Benchmarking	XXX	X		XXX	XXX (vs Regression Estimator)

XXX – High
XX – Medium

X – Low

References

Sections 1-2

Atzberger, C. 2013. Advances in remote sensing of agriculture: Context description, existing operational monitoring systems and major information needs. *Remote Sensing*, 5: 949-981.

Becker-Reshef, I., Justice, C., Sullivan, M., Vermote, E., Tucker, C., Anyamba, A., Small, J., Pak, E., Masuoka, E., Schmaltz, J., Hansen, M., Pittman, K., Birkett, C., Williams, D., Reynolds, C., and Doorn, B. 2010. Monitoring global croplands with coarse resolution earth observations: The Global Agriculture Monitoring (GLAM) Project. *Remote Sensing*, 2: 1589-1609.

Benedetti, R. and Rossini, P. 1993. On the use of NDVI profiles as a tool for agricultural statistics: the case study of wheat yield estimate and forecast in Emilia Romagna. *Remote Sensing of Environment*, 326(1): 311–326.

Benedetti, R., Rossini, P., and Taddei, R. 1994. Vegetation classification in the Middle Mediterranean area by satellite data. *International Journal of Remote Sensing*, 15(3):583–596.

Canada Centre for Remote Sensing. (2003). *Principles of remote sensing*. Publication of the Centre for Remote Imaging, Sensing and Processing (CRISP), National University of Singapore. Available at: http://www.crisp.nus.edu.sg/~research/tutorial/rsmain.htm.

Carfagna, E. and Gallego, F.J. 2005. Using remote sensing for agricultural statistics. *International Statistical Review*, 73: 389-404.

Chambers, R.L., Chandra, H., Salvati, N., and Tzavidis, N. 2013. Outlier robust small area estimation. *Journal of the Royal Statistical Society: Series B*, 176(4):1–23.

Choimeun, S., Phumejaya, N., Pomnakchim, S., and Chantrapornchai, C. 2010. Tool for collecting spatial data with Google Maps API. In Kim T. H., Ma, J., Fang, W.C., Park, B., Kang, B.H., Slezak, D (eds), *U- and E-Service, Science and Technology, Communications in Computer and Information Science*, vol. 124, pp. 107–113, Springer-Verlag, Berlin.

Delécolle, R., Maas, S.J., Guérif, M., and Baret, F. 1992. Remote sensing and crop production models: Present trends. *ISPRS Journal of Photogrammetry and Remote Sensing*, 47: 145–161.

Donlon, C., Berruti, B., Buongiorno, A., Ferreira, M.H., Féménias, P.,

Frerick, J., Goryl, P., Klein, U., Laur, H., Mavrocordatos, C., Nieke, J., Rebhan, H., Seitz, B., Stroede, J., and Sciarra, R. 2012. The Global Monitoring for Environment and Security (GMES) Sentinel-3 mission. *Remote Sensing of Environment*, 120: 37–57.

Dorigo, W.A., Zurita-Milla, R., de Wit, A.J.W., Brazile, J., Singh, R., and Schaepman, M.E. 2007. A review on reflective remote sensing and data assimilation techniques for enhanced agroecosystem modeling. *International Journal of Applied Earth Observation and Geoinformation*, 9: 165-193.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, A., Spoto, F., Sy, O., Marchese, F., and Bargellini, P. 2012. Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote Sensing of Environment*, 120: 25–36

Ferencz, C., Bognar, P., Lichtenberger, J., Hamar, D., Tarcsai, G., Timar, G., Molnar, G., Pasztor, S., Steinbach, P., Szekely, B., Ferencz, O.E., and Ferencz-arkos, I. 2004. Crop yield estimation by satellite remote sensing. *International Journal of Remote Sensing*, 25: 4113-4149.

Fitzgerald, G.J., Lesch, S.M., Barnes, E.M., and Luckett, W.E. 2006. Directed sampling using remote sensing with a response surface sampling design for site-specific agriculture. *Computers and Electronics in Agriculture.* 53: 98-112.

Gallego, F.J. 1999. Crop area estimation in the MARS project. Conference on ten years of the MARS project, Brussels, April 1999.

GEOSS (Global Earth Observation System of Systems). 2009. *Best Practices for Crop Area Estimation with Remote Sensing*. Available at: http://www.earthobservations.org/documents/cop/ag_gams/GEOSS best practices area estimation final.pdf.

Ingmann, P., Veihelmann, B., Langen, J., Lamarre, D., Stark, H., and Courrèges-Lacoste, G.B. 2012. Requirements for the GMES atmosphere service and ESA's implementation concept: Sentinels-4/-5 and -5p. *Remote Sensing of Environment*, 120: 58–69.

Jensen, J.R. 2004. *Introductory digital image processing: a remote sensing perspective*. Prentice Hall: New Jersey, USA.

Rembold, F., Atzberger, C., Savin, I., and Rojas, O. 2013. Using low resolution satellite imagery for yield prediction and yield anomaly detection. *Remote Sensing*, 5: 1704-1733.

Richards, J.A. and Jia, X. 2006. *Remote sensing digital image analysis. An*

Introduction. Springer Verlag: Berlin Heidelberg.

Roy, D.P., Borak, J.S., Devadiga, S., Wolfe, R.E., Zheng, M., and Descloitres, J. 2002. The MODIS Land product quality assessment approach. *Remote Sensing of Environment*. 83: 62–76.

Torres, R., Snoeij, P., Geudtner, D., Bibby, D., Davidson, M., Attema, E., Potin, P., Rommen, B., Floury, N., Brown, M., Navas Traver, I., Deghaye, P., Duesmann, B., Rosich, B., Miranda, N., Bruno, C., L'Abbate, M., Croci, R., Pietropaolo, A., Huchler, and M., Rostan, F. 2012. GMES Sentinel-1 mission. Remote Sensing of Environment, 120: 9–24.

Walker, P.A. and Mallawaarachchi, T. (1998). Disaggregating Agricultural Statistics Using NOAA-AVHRR NDVI. *Remote Sensing of Environment*, 63:112–125.
World Bank (2011). Global strategy to improve agricultural and rural statistics. Report No. 56719-GLB, Washington, DC, USA. **Zamir, R. and Mubarak, S.** 2010. Accurate image localization based on Google Maps Street view. In Danilidis, K., Maragos, P., and Paragios, N. (eds), *Computer Vision- ECCV-2010*. Springer-Verlag: Berlin, pp. 255–268.

Web References

Agriculture Stress Index System (ASIS). Available at:
http://www.fao.org/climatechange/asis/en/.

China CropWatch System (CCWS). Available at:
http://www.cropwatch.com.cn/en/.

Euroconsult. Available at: http://www.euroconsult-ec.com/research-reports-28.html.

Envisat. Available at :
http://www.esa.int/Our_Activities/Observing_the_Earth/Envisat_overview.

Famine Early Warning Systems Network (FEWSNET). Available at:
http://www.fews.net/Pages/default.aspx.

Global Information and Early Warning System (GIEWS). Available at:
http://fao.org/gviews.

Hyperspectral Infrared Imager (HyspIRI). Available at:
http://hyspiri.jpl.nasa.gov/.

IKONOS. Available at: http://www.digitalglobe.com/about-us/content-collection#ikonos.

Landsat. Available at: http://landsat.usgs.gov/.
MetOp. Available at:
http://www.esa.int/Our_Activities/Observing_the_Earth/The_Living_Planet_Programme/Meteorological_missions/MetOp.

Monitoring Agricultural ResourceS (MARS). Available at:
http://mars.jrc.ec.europa.eu/.

Moderate Resolution Imaging Spectroradiometer (MODIS). Available at:
http://modis.gsfc.nasa.gov/.

NASA missions. Available at :
http://www.nasa.gov/missions/schedule/index.html#.UqsWAI0lvx4.

National Oceanic and Atmospheric Administration (NOAA). Available at:
http://www.noaa.gov/.

Planet Labs. Available at: http://www.planet-labs.com/.

Pleiades. Available at : http://www.astrium-geo.com/pleiades/.

Proba-V. Available at:
http://www.esa.int/Our_Activities/Technology/Proba_Missions.

QuickBird. Available at: http://www.digitalglobe.com/about-us/content-collection#quickbird.

Rapideye. Available at : http://www.satimagingcorp.com/satellite-sensors/rapideye.html.

Sentinel. Available at :
http://www.esa.int/Our_Activities/Observing_the_Earth/GMES/Overview4.

SPOT (Système pour d'Observation de la Terre). Available at:
http://www.cnes.fr/web/CNES-en/1415-spot.php.

SPOT-6/7 satellites. Available at: http://www.astrium-geo.com/en/147-spot-6-7.

SPOT vegetation program. Available at: http://www.spot-vegetation.com/index.html.

VENµS. Available at: http://smsc.cnes.fr/VENUS/index.htm.
WorldView-1. Available at: http://www.digitalglobe.com/about-us/content-

collection-worldview-1.

WorldView-2. Available at: http://www.digitalglobe.com/about-us/content-collection-worldview-2.

WorldView-3. Available at: http://www.digitalglobe.com/about-us/content-collection-worldview-3.

Section 3

Arbia, G. 1993. The use of GIS in spatial statistical surveys. *International Statistical Review*, 61: 339–359.

Baillargeon, S. and Rivest, L.-P. 2009. A general algorithm for univariate stratification. *International Statistical Review*, 77: 331-344.

Baillargeon, S. and Rivest, L.-P. 2011. The construction of stratified designs in R with the package stratification. *Survey Methodology*, 37: 53-65.

Ballin, M. and Barcaroli, G. 2013. Joint determination of optimal stratification and sample allocation using genetic algorithm. *Survey Methodology*, 39(2): 369–393.

Barcaroli, G. 2014. Samplingstrata: an r package for the optimization of stratified sampling. *Journal of Statistical Software*, 1(1):1–20.

Bee, M., Benedetti, R., Espa, G., and Piersimoni, F. 2010. On the use of auxiliary variables in agricultural surveys design. In Benedetti, R., Bee, M., Espa, G., and Piersimoni, F. (eds.), *Agricultural Survey Methods*, pp. 107-132, John Wiley & Sons: Chichester, UK.

Benedetti, R., Bee, M., Espa, G., and Piersimoni, F. 2010. *Agricultural survey methods*. John Wiley & Sons: Chichester, UK.

Benedetti, R., Espa, G., and Lafratta, G. 2008. A tree-based approach to forming strata in multipurpose business surveys. *Survey Methodology*, 34: 195-203.

Benedetti, R. and Palma, D. 1995. Optimal sampling designs for dependent spatial units. *Environmetrics*. 6: 101-114.

Benedetti, R. and Piersimoni, F. 2012. Multivariate boundaries of a self representing stratum of large units in agricultural survey design. *Survey Research Methods*, 6, 3, 125–135.
Besag, J. 1986. On the statistical analysis of dirty pictures. *Journal of the Royal*

Statistical Society, Series B, 48: 259–302.

Bethel, J. 1989. Sample allocation in multivariate surveys, *Survey Methodology*, 15: 47–57.

Breidt, F.J., Chauvet, G. 2012. Penalized balanced sampling. *Biometrika*. 99: 945-958.

Brémaund, P. 1999. *Markov Chain: Gibbs Fields, Monte Carlo Simulation and Queues*. Spinger-Verlag: New York, USA.

Brewer, K.R.W. and Hanif, M. 1983. *Sampling with Unequal Probabilities*. Springer-Verlag: New York, USA.

Briggs, J., Duoba, V., Zealand, S.N. 2000. *STRAT2D: Optimal bi-variate stratification system. Proceedings of the 2nd International Conference on Establishment Surveys*, pp. 1589-1594.

Cameron, A.C., Trivedi, P.K. 2005. *Microeconometrics: methods and applications*. Cambridge University Press: New York, USA.

Chauvet, G. 2009. Stratified balanced sampling. *Survey Methodology*. 35: 115-119.

Chauvet, G., Bonnéry, D., and Deville, J.C. 2011. Optimal inclusion probabilities for balanced sampling. *Journal of Statistical Planning and Inference*. 141: 984-994.

Chauvet, G. and Tillé, Y. 2006. A fast algorithm of balanced sampling. *Computational Statistics*. 21: 53-62.

Dalenius, T. and Hodges, J.L. Jr. 1959. Minimum Variance Stratification. *Journal of the American Statistical Association*, 54, 285, 88-101.

Deville, J.C. and Tillé, Y. 1998. Unequal probability sampling without replacement through a splitting method. *Biometrika*, 85: 89-101.

Deville, J.C. and Tillé, Y. 2004. Efficient balanced sampling: The cube method. *Biometrika*. 91: 893–912.

Dunn, R. and Harrison, A.R. 1993. Two-dimensional systematic sampling of land use. *Journal of the Royal Statistical Society, Series C (Applied Statistics).* 42: 585-601.

Falorsi, P.D. and Righi, P. (2008). A balanced sampling approach for multi-way stratification designs for small area estimation. *Survey Methodology*, 34(2):223–234.

Fattorini, L. 2006. Applying the Horvitz–Thompson criterion in complex designs: A computer-intensive perspective for estimating inclusion probabilities. *Biometrika*. 93: 269–278.

Fattorini, L. 2009. An adaptive algorithm for estimating inclusion probabilities and performing the Horvitz–Thompson criterion in complex designs. *Computational Statistics*, 24: 623–639.

Fletcher, D., Mackenzie, D., Villouta, E. 2005. Modelling skewed data with many zeros: a simple approach combining ordinary and logistic regression. *Journal of Environmental and Ecological Statistics*, 12: 45–54.

Foreman, E.K. and Brewer, K.R.W. 1971. The Efficient use of supplementary information in standard sampling procedures. *Journal of the Royal Statistical Society, Series B (Methodological)*, 33: 391-400.

Geman, S. and Geman, D. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. PAMI-7, 6: 721-741.

Grafström, A. 2012. Spatially correlated Poisson sampling. *Journal of Statistical Planning and Inference*. 142: 139–147.

Grafström, A., Lundström, N.L.P. and Schelin, L. 2012. Spatially balanced sampling through the Pivotal method. *Biometrics*, 68: 514-520.

Grafström, A. and Tillé, Y. 2013. Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. *Environmetrics*, 24: 120-131.

Hidiroglou, M.A. 1986. The construction of a self-representing stratum of large units in survey design. *The American Statistician*, 40: 27-31.

Hidiroglou, M.A. and Srinath, K.P. 1993. Problems associated with designing subannual business surveys. *Journal of Business & Economic Statistics*. 11: 397-405.

Holmberg, A. 2007. *Using unequal probability sampling in business surveys to limit anticipated variances of regression estimators*. In International Conference on Establishment Surveys, III, 550–556.

Horgan, J.M. 2006. Stratification of skewed populations: A review. *International Statistical Review*, 74: 67-76.

Isaki, C.T. and Fuller, W.A. 1982. Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77: 89–96.

Karlberg, F. 2000. Survey estimation for highly skewed populations in the presence of zeroes. *Journal of Official Statistics*, 16: 229–241.

Khan, M.G.M., Nand, N. and Ahma, N. 2008. Determining the optimum strata boundary points using dynamic programming. *Survey Methodology*, 34: 205-214

Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P. 1983. Optimization by simulated annealing. *Science*, 220: 671-680.

Kott, P.S. and Bailey, J.T. 2000. *The theory and practice of maximal Brewer selection with Poisson PRN sampling*. In International Conference on Establishment Surveys, II, 1–12.

Kozak, M. 2004. Optimal Stratification Using Random Search Method in Agricultural Surveys. *Statistics in Transition*, 5: 797-806.

Liu, H. and Chan, K.S. 2010. Introducing COZIGAM: An R Package for unconstrained and constrained zero-inflated generalized additive model analysis. *Journal of Statistical Software*, 35: 1-26.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, N.M., Teller, A.H., and Teller, E. 1953. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21: 1087-1092.

Ohlsson, E. (1998). Sequential Poisson Sampling. *Journal of Official Statistics*, 14: 149-162.

Robert, C.P. and Casella, G. 1999. *Monte Carlo statistical methods*. Springer: New York, USA.

Robert, C.P. and Casella, G. 2010. *Introducing Monte Carlo methods with R. Use R*. Springer: New York, USA.

Rogerson, P.A. and Delmelle, E. 2004. Optimal sampling design for variables with varying spatial importance. *Geographical Analysis*, 36: 177-194.
Rosén, B. 1997a. Asymptotic theory for order sampling. *Journal of Statistical Planning and Inference*, 62: 135-158.

Rosén, B. 1997b. On sampling with probability proportional to size. *Journal of Statistical Planning and Inference*, 62: 159-191.

Särndal, C.E., Swensson, B., Wretman, J. 1992. *Model assisted survey sampling*. Springer Verlag: New York, USA.

Stevens, D.L. Jr, Olsen, A.R. 2003. Variance estimation for spatially balanced samples of environmental resources. *Environmetrics*, 14: 593–610.

Stevens, D.L. Jr, Olsen, A.R. 2004. Spatially balanced sampling of natural resources. *Journal of the American Statistical Association*, 99: 262–278.

Tillé, Y. 2006. *Sampling algorithms. Springer series in statistics*. Springer: New York, USA.

Tillé, Y. 2011. Ten years of balanced sampling with the cube method: An appraisal. *Survey Methodology*, 2: 215-226.

Tillé, Y. and Favre, A.-C. 2005. Optimal allocation in balanced sampling. *Statistics & Probability Letters*, 1: 31-37.

Tsallis, C. and Stariolo, D.A. 1996. Generalized simulated annealing. *Physica A*, 233: 395-406.

Traat, I., Bondesson, L., Meister, K. 2004. Sampling design and sample selection through distribution theory. *Journal of Statistical Planning and Inference*, 123: 395-413.

Verma, M.R. and Rizvi, S.E.H. 2007. Optimum stratification for PPS sampling using auxiliary information. *Journal of the Indian Society of Agricultural Statistics*, 61: 66-76.

Vogel, F.A. 1995. The evolution and development of agricultural statistics at the United States Department of Agriculture. *Journal of Official Statistics*, 11: 161-180.

Wang, J.-F., Stein, A., Gao, B.-B. and Ge, Y. 2012. A review of spatial sampling. *Spatial Statistics*, 2: 1–14.

Section 4

Andersson, P.G. and Thorburn, D. 2005. An optimal calibration distance leading to the optimal regression estimator. *Survey Methodology*, 31: 95-99.

Beaumont, J.-F. and Alavi, A. 2004. Robust generalized regression estimation. *Survey Methodology*, 30: 195-208.

Besag, J. 1986. The statistical analysis of dirty pictures. *Journal of the Royal Statistical Society, Series B*, 48: 259–302.

Benedetti, R. and Filipponi, D. 2010. Estimation of land cover parameters when some covariates are missing. In: Benedetti, R., Bee, M., Espa, G., Piersimoni, F.

(eds), *Agricultural survey methods.* John Wiley & Sons: Chichester, UK, pp. 213–230.

Breidt, F.J., Claekens, G., Opsomer, J.D. 2005. Model-assisted estimation for complex surveys using penalized splines. *Biometrika*, 92: 831-846.

Breidt, F.J. and Opsomer, J.D. 2000. Local polynomial regression estimators in survey sampling. *The Annals of Statistics*, 28: 1026-1053.

Breidt, F.J. and Opsomer, J.D. 2008. Endogenous post-stratification in surveys: classifying with a sample-fitted model. *The Annals of Statistics*, 36: 403-427.

Cicchitelli, G. and Montanari, G.E. 2012. Model-assisted estimation of a spatial population mean. *International Statistical Review*, 80: 111-126.
Cressie, N. 1993. *Statistics for Spatial Data*. Wiley: New York, USA.

Demnati, A. and Rao, J.N.K. 2010. Linearization variance estimators for model parameters from complex survey data. *Survey Methodology*, 36: 193-201.

Deville, J.C. and Särndal, C.E. 1992. Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87: 376-382.

Deville, J.C., Särndal, C.-E., and Sautory, O. 1993. Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88: 1013-1020.

Fletcher, D., Mackenzie, D., and Villouta, E. 2005. Modelling skewed data with many zeros: a simple approach combining ordinary and logistic regression. *Journal of Environmental and Ecological Statistics*, 12: 45–54.

Doraiswamy, P.C., Sinclair, T.R., Hollinger, S., Akhmedov, B., Stern, A., and Prueger, J. 2005. Application of MODIS derived parameters for regional crop yield assessment. *Remote Sensing of Environment*, 97: 192-202.

Estevao, V.M. and Särndal, C.E. 2000. A functional form approach to calibration. Journal of Official Statistics, 16: 379-399.

Estevao, V.M. and Särndal, C.E. 2004. Borrowing strength is not the best technique within a wide class of design-consistent domain estimators. *Journal of Official Statistics*, 20: 645-669.

Estevao, V.M. and Särndal, C.E. 2006. Survey estimates by calibration on complex auxiliary information. *International Statistical Review*, 74: 127-147.

Estevao, V.M. and Särndal, C.E. 2009. A new face on two-phase sampling with calibration estimators. *Survey Methodology*, 35: 3-14.

Gallego, F.J. 2004. Remote sensing and land cover area estimation. *International Journal of Remote Sensing*, 25: 3019-3047.

Gallego, F.J. and Delincé, J. (1994). Using a confusion Matrix for Area Estimation with Remote Sensing. In *Atti Convegno Associazione Italiana Telerilevamento (Roma: AIT)*, pp. 99–102.

Geman, D., Geman, S., Graffigne, C. and Dong, P. 1990. Boundary detection by constrained optimization. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 12: 609–628.

González, F. and Cuevas, M. 1993. Remote sensing and agricultural statistics: crop area estimation through regression estimators and confusion matrices. *International Journal of Remote Sensing*, 14: 1215-1219.

Hung, H.M. and Fuller, W.A. 1987. Regression estimation of crop acreages with transformed landsat data as auxiliary variables. *Journal of Business and Economic Statistics*, 5: 475-482.

Karlberg, F. 2000. Survey estimation for highly skewed populations in the presence of zeroes. *Journal of Official Statistics*, 16: 229–241.

Kim, J.K., Fuller, W.A. and Bell, W.R. 2011. Variance estimation for nearest neighbor imputation for US Census long form data. *Annals of Applied Statistics*, 5, 2Am 824-842.

Kim, J.K. and Park, M. 2010. Calibration estimation in survey sampling. *International Statistical Review*, 78: 21-39.

Kirkpatrik, S., Gelatt, C.D. and Vecchi, M.P. 1983. Optimization by simulated annealing. *Science*, 220: 671–680.

Lavallée, P. 2007. *Indirect sampling*. Springer: New York, USA.

Lehtonen, R. and Veijanen, A. 1998. Logistic generalized regression estimators. *Survey Methodology* 24, 51-55.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. 1953. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21: 1087–1092.

Montanari, G.E. and Ranalli, G. 2005. Nonparametric model calibration estimation in survey sampling. *Journal of the American Statistical Association*, 100: 1429-1442.

Moura, F.A.S. and Holt, D. 1999 Small area estimation using multilevel models. *Survey Methodology*, 25(1), p.73-80.

Opsomer, J.D., Moisen, G.G. and Kim, J.Y. 2001. Model-assisted estimation of forest resources with generalized additive models. *Proceedings of the Section on Survey Research Methods*, American Statistical Association. Alexandria, VA, USA.

Postiglione, P., Andreano, S. and Benedetti, R. 2013. Using constrained optimization for the identification of convergence clubs. *Computational Economics*, 42: 151–174

Pradhan, S. 2001. Crop area estimation using GIS, remote sensing and area frame sampling. *International Journal of Applied Earth Observation and Geoinformation*, 3: 86-92.

Rao, J.N.K. 1996. On variance estimation with imputed survey data. *Journal of the American Statistical Association*, 91(434):499–506.

Rubin, D.B. 2004. *Multiple imputation for nonresponse in surveys*. Wiley-Interscience: USA.

Särndal, C.E. 2007. The calibration approach in survey theory and practice. *Survey Methodology*, 33: 99-119.

Särndal, C.E. and Lundström, S. 2005. *Estimation in surveys with nonresponse*. John Wiley and Sons.

Särndal, C.E. and Lundström, S. 2008. Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator. *Journal of Official Statistics*, 24: 167-191.

Särndal, C.E. and Lundström, S. 2010. Design for estimation: Identifying auxiliary vectors to reduce nonresponse bias. *Survey Methodology*, 36: 131-144.

Särndal, C.E., Swensson, B. and Wretman, J. 1992. *Model-assisted survey sampling*. Springer-Verlag: New York, USA.

Sebastiani, M.R. 2003. Markov random-field models for estimating local labour markets. *Journal of the Royal Statistical Society, Series C, Applied Statistics*, 52: 201–211.

Stukel, D., Hidiroglou, M.A. and Särndal, C.E. 1996. Variance estimation for calibration estimators: a comparison of jackknifing versus Taylor linearization. *Survey Methodology*, 22, 117-125.

Wu, C. 1999. *The effective use of complete auxiliary information from survey data*. Simon Fraser University, Canada. (Ph.D. Thesis, unpublished).

Wu, C. 2003. Optimal calibration estimators in survey sampling. *Biometrika*, 90: 937-951.

Wu, C. and Luan, Y. 2003. Optimal calibration estimators under two-phase sampling. *Journal of Official Statistics*, 19:119–131.

Wu, C. and Sitter, R.R. 2001. A model calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96: 185-193.

Zhang LC. 2000. Post-stratification and calibration - a synthesis. *The American Statistician*, 54: 178-184.

Section 5

Beaumont, J.F. and Alavi, A. 2004. Robust generalized regression estimation. *Survey Methodology*, 30: 195-208.

Bee, M., Benedetti, R., Espa, G., and Piersimoni, F. 2010. On the use of auxiliary variables in agricultural survey design. In Benedetti, R., Bee, M., Espa, G., Piersimoni, F. (eds). Agricultural survey methods, pp. 107-132. John Wiley & Sons, Ltd, Chichester, UK.

Carfagna, E. and Carfagna, A. 2010. Alternative sampling frames and administrative data. What is the best data source for agricultural statistics? In Benedetti, R., Bee, M., Espa, G., Piersimoni, F. (eds). Agricultural survey methods, pp. 45-61. John Wiley & Sons: Chichester, UK,.

Cassel, C.M., Särndal, C.E. and Wretman, J.H. 1976. Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63(3): 615-20.

Cotter, J., Davies, C., Nealon, J. and Roberts, R. 2010. Area frame design for agricultural surveys. In Benedetti, R., Bee, M., Espa, G., Piersimoni, F. (eds). *Agricultural survey methods*, pp. 169-192. John Wiley & Sons: Chichester, UK.

Deville, J.C. and Särndal, C.E. 1992. Calibration estimators in sampling survey. *Journal of the American Statistical Association*, 87: 376-382.

Gallego, J., Carfagna, E., Baruth, B. 2010. Accuracy, objectivity and efficiency of remote sensing for agricultural statistics. In Benedetti, R., Bee, M., Espa, G., Piersimoni, F. (eds). *Agricultural survey methods*,pp. 193-211. , John Wiley & Sons: Chichester, UK.

House, C.C. 2010. Statistical aspects of a census. In: Benedetti, R., Bee, M., Espa, G., Piersimoni, F. (eds). *Agricultural survey methods*, pp. 63-72. John Wiley & Sons: Chichester, UK.

Pfeffermann, D. 2013. New important developments in small area estimation. *Statistical Science*, 28: 40-68.

Rao, J.N.K. 2003. *Small Area Estimation*. John Wiley & Sons: Hoboken, USA.

Rao, J.N.K. 2010. Small-area estimation with applications to agriculture. In Benedetti, R., Bee, M., Espa, G., Piersimoni, F. (eds). *Agricultural survey methods*, pp. 139-147. John Wiley & Sons: Chichester, UK..

Särndal, C.E. 2007. The calibration approach in survey theory and practice. *Survey Methodology*, 33: 99-119.

You, Y. and Rao, J.N.K. 2002. A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *The Canadian Journal of Statistics*, 30: 431-439.

You, Y., Rao, J.N.K., and Dick, P. 2004. Benchmarking hierarchical Bayes small area estimators in the Canadian census undercoverage estimation. *Statistics in Transition*, 6: 631-640.

You, Y., Rao, J.N.K. and Hidiroglou, M. 2013. On the performance of self benchmarked small area estimators under the Fay-Herriot area level model. *Survey Methodology*, 39: 217-229.

Wang, J., Fuller, W.A. and Qu, Y. 2008. Small area estimation under a restriction. *Survey Methodology*, 34: 29-36.

Section 6

Battese, G., Harter, R.M., and Fuller, W. 1988. An error-components model for prediction of county crop areas using survey and satellite Data. *Journal of the American Statistical Association*, 83: 28-36.

Benedetti, R. and Filipponi, D. 2010. Estimation of land cover parameters when some covariates are missing. In Benedetti, R., Bee, M., Espa, G., and Piersimoni, F. (eds.), *Agricultural Survey Methods*, pp. 213–230. John Wiley & Sons: Chichester, UK,.

Benedetti, R., Pratesi, M., and Salvati, N. 2013. Local stationarity in small area estimation models. *Statistical Methods and Applications*, 22:81–95.

Chandra, H. and Chambers, R. 2008. *Estimation of small domain means for* *zero contaminated skewed data*. Centre for Statistical and Survey Methodology, University of Wollongong, Working Paper 07-08.

Chandra, H. and Sud, U.C. 2012. Small area estimation for zero-inflated data. *Communications in Statistics - Simulation and Computation*, 41: 632–643.

Cressie, N. 1991. *Small area prediction of undercount using the general linear model*. Proceedings of Statistics Symposium 90: Measurement and Improvement of Data Quality, Statistics Canada, Ottawa, pp. 93–105.

Cressie, N. 1992. REML Estimation in Empirical Bayes Smoothing of Census Undercount, *Survey Methodology*, 18: 75-94.

Datta, G.S., Day, B., and Maiti, T. 1998. Multivariate Bayesian small area estimation: An application to survey and satellite data. *Sankhya: The Indian Journal of Statistics, Series A*, 60: 344–362.

Datta, G.S., Fay, R.E., and Ghosh, M. 1991. *Hierarchical and empirical Bayes multivariate analysis in small area estimation*. Proceedings of Bureau of the Census 1991 Annual Research Conference, US Bureau of the Census, Washington DC, pp. 63-79.

Fay, R.E. and Herriot, R.A. 1979. Estimates of income for small places: An application of James–Stein procedures to census data. *Journal of the American Statistical Association*, 74: 269–277.

Fletcher, D., Mackenzie, D., and Villouta, E. 2005. Modelling skewed data with many zeros: a simple approach combining ordinary and logistic regression. *Journal of Environmental and Ecological Statistics*, 12: 45–54.

Flores, L.A. and Martinez, L.I. 2000. Land cover estimation in small areas using ground survey and remote rensing. *Remote Sensing of Environment*, 74: 240–248.

Ghosh, M. and Rao, J.N.K. 1994. Small area estimation: an appraisal. *Statistical Science*, 9: 55-93.

MacGibbon, B. and Tomberlin, T.J. 1989. Small area estimates of proportions via empirical Bayes techniques. *Survey Methodology*, 15: 237–252.

Petrucci, A. and Salvati, N. 2006. Small area estimation for spatial correlation in watershed erosion assessment. *Journal of Agricultural Biological and Environmental Statistics*, 11: 169–182.

Pfeffermann, D. 2002. Small area estimation: New developments and directions. *International Statistical Review*, 70: 125–143.

Pfeffermann, D. 2013. New important developments in small area estimation. *Statistical Science*, 28: 40–68.

Pratesi, M. and Salvati, N. 2008. Small area estimation: the eblup estimator based on spatially correlated random area effects. *Statistical Methods and Application*, 17:114–131.

Pratesi, M and Salvati, N. 2009. Small area estimation in the presence of correlated random area effects. *Journal of Official Statistics*, 25: 37–53.

Rao, J.N.K. 2002. Small area estimation: update with appraisal. In Balakrishnan, N. (ed.) *Advances on methodological and applied aspects of probability and statistics*, pp. 113-139. Taylor and Francis: New York, USA.

Rao, J.N.K. 2003. *Small Area Estimation*. John Wiley & Sons: Hoboken, USA.

Rao, J.N.K. and Yu, M. 1994. Small area estimation by combining time series and cross- sectional data. *Canadian Journal of Statistics*, 22: 511-528.

Rashid, M.M. and Nandram, B. 1998. A rank-based predictor for the finite population mean of a small area: An application to crop production. *Journal of Agricultural, Biological, and Environmental Statistics*, 3: 201–222.

Sebastiani, M.R. 2003. Markov random-field models for estimating local labour markets. *Journal of the Royal Statistical Society, Series C, Applied Statistics*, 52: 201–211.

Torabi, M. and Rao, J.N.K. 2008. Small area estimation under a two-level model. *Survey Methodology*, 34: 11–17.

You, Y. and Chapman, B. 2006. Small area estimation using area level models and estimated sampling variances. *Survey Methodology*, 32: 97–103.

Section 7

Anderson, J.R., Hardy, E.E., Roach, J.T., and Witmer, R.E. 1976. *A Land Use and Land Cover Classification System for Use with Remote Sensor Data*. USGS Professional Paper 964, U.S. Geological Survey, U.S. Government Printing Office, Washington, D.C.

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. 1975. *Discrete multivariate analysis. Theory and practice*. MIT Press, Cambridge, USA.

Carfagna, E. and Marzialetti, J. 2009a. Sequential design in quality control and validation of land cover databases. *Applied Stochastic Models in Business and*

Industry, 25: 195-205.

Carfagna, E. and Marzialetti, J. 2009b. Continuous innovation of the quality control of remote sensing data for territory management. In: Erto, P. (ed.). *Statistics for Innovation, Statistical Design of "Continuous" Product Innovation*, pp. 172-188. Springer Verlag, Milan.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20: 37-46.

Congalton, R.G. 1988. A comparison of sampling schemes used in generating error matrices for assessing the accuracy of map generated from remotely sensed data. *Photogrammetric Engineering & Remote Sensing*, 54: 593-600.

Congalton, R.G. 1991. A Review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment*, 37: 35-46.

Congalton, R.G. and Green, K. 1999. *Assessing the accuracy of remotely sensed data: principles and practices*. Lewis Publishers: Boca Raton, USA.

Czaplewski, R.L. 1994. *Variance approximations for assessments of classification accuracy*. Research Paper RM-316, USDA, Forest Service, Rocky Mountain Forest and Range Experiment Station, Fort Collins, CO, USA.

Fitzpatrick-Lins, K. 1981. Comparison of sampling procedures and data analysis for a land use land cover map. *Photogrammetric Engineering & Remote Sensing*, 47: 343-351.

Foody, G.M. 1995. Land-cover classification by an artificial neural-network with ancillary information. *International Journal of Geographical Information Science*, 9: 527-542.

1996. Approaches for the production and evaluation of fuzzy land cover classifications from remotely-sensed data. *International Journal of Remote Sensing*, 17: 1317–1340.

1998. Sharpening fuzzy classification output to refine the representation of sub-pixel land cover distribution. *International Journal of Remote Sensing*, 19: 2593–2599.

2000. Estimation of sub-pixel land cover composition in the presence of untrained classes. *Computers & Geosciences*, 26: 469-478.
2002. Status of land cover classification accuracy. *Remote Sensing of Environment*, 80: 185-201.

Foody, G.M., Campbell, N.A., Trodd, N.M., and Wood, T.F. 1992. Derivation and applications of probabilistic measures of class membership from the maximum likelihood classification. *Photogrammetric Engineering and Remote Sensing*, 58: 1335-1341.

Gallego, F.J., Carfagna, E., and Baruth, B. 2010. Accuracy, objectivity and efficiency of remote sensing for agricultural statistics. In Benedetti, R., Bee, M., Espa, G., Piersimoni, F. (eds.), *Agricultural Survey Methods*, pp. 193-211. John Wiley & Sons: Chichester, UK.

Gopal, S. and Woodcock, C.E. 1994. Accuracy assessment of thematic maps using fuzzy sets I: theory and methods. *Photogrammetric Engineering & Remote Sensing*, 60: 181-188.

Hammond, T.O. and Verbyla, D.L. 1996. Optimistic bias in classification accuracy assessment. *International Journal of Remote Sensing*, 7: 1261-1266.

Hansen, M.C., Dubayah, R., and DeFries, R. 1996. Classification trees: an alternative to traditional land cover classifiers. *International Journal of Remote Sensing*, 17: 1075-1081.

Liu, C., Frazier, P., and Kumar, L. 2007. Comparative assessment of the measures of thematic classification accuracy. *Remote Sensing of Environment*, 107: 606-616.

Lu, D., Mausel, P., Brondízio, E., and Moran, E. 2004. Change detection techniques. *International Journal of Remote Sensing*, 25: 2365-2407.

Lunetta, R.S., Congalton, R.G., Fenstemarker, L.K., Jensen, J.R., McGwire, K.C., and Tinney, L.R. 1991. Remote sensing and geographic information system data integration: Error sources and research issues. *Photogrammetric Engineering & Remote Sensing*, 57: 677-687.

Macleod, R.D. and Congalton, R.G. 1998. A quantitative comparison of change detection algorithms for monitoring eelgrass from remotely sensed data. *Photogrammetric Engineering & Remote Sensing*, 64: 207-216.

Ohlsson, E. 1995. Coordination of samples using permanent random numbers, In: Cox, B.G., Binder, D.A., Nanjamma Chinnappa, B., Christianson, A., Colledge, M.J., and Kott, P.S. (eds.), *Business Survey Methods*, pp. 153–169. John Wiley & Sons: New York, USA.

Nusser, S.M. and Klaas, E.E. 2003. Survey methods for assessing land cover map accuracy. *Environmental and Ecological Statistics*: 10, 309-331.

Särndal, C.E., Swensson, B., and Wretman, J. 1992. *Model-assisted survey sampling*. Springer-Verlag: New York, USA.

Scepan, J. 1999. Thematic validation of high-resolution global land-cover data sets. *Photogrammetric Engineering & Remote Sensing*, 65: 1051-1060.

Serra, P., Pons, X., and Saurí, D. 2003. Post-classification change detection with data from different sensors: some accuracy considerations. *International Journal of Remote Sensing*, 24: 3311-3340.

Stehman, S.V. 1995. Thematic map accuracy assessment from the perspective of finite population sampling. *International Journal of Remote Sensing*, 16: 589-593.

1996. Estimating the Kappa coefficient and its variance under stratified random sampling. *Photogrammetric Engineering & Remote Sensing*, 62: 401-407.

2001. Statistical rigor and practical utility in thematic map accuracy assessment. *Photogrammetric Engineering & Remote Sensing*, 67: 727-734.

Stehman, S.V. and Czaplewski, R.L. 1998. Design and analysis for thematic map accuracy assessment: fundamental principles. *Remote Sensing of Environment*, 64: 331–344.

Story, M. and Congalton, R. 1986. Accuracy assessment: a user's perspective. *Photogrammetric Engineering & Remote Sensing*, 52: 397-399.

Strahler, A.H., Boschetti, L., Foody, G.M., Friedl, M.A., Hansen, M.C., Herold, M., Mayaux, P., Morisette, J.T., Stehman, S.V. and Woodcock, C.E. 2006. *Global land cover validation: recommendations for evaluation and accuracy assessment of global land cover maps*. GOFC-GOLT Report No 25, Office for Official Publication of the European Communities, Luxembourg.

Thompson, S.K. 1990. Adaptive cluster sampling. *Journal of the AmericanStatistical Association*, 85: 1050-1059.

van Oort, P.A.J. 2007. Interpreting the change detection error matrix. *Remote Sensing of Environment*, 108: 1-8.

Verbyla, D.L. and Hammond, T.O. 1995. Conservative bias in classification accuracy assessment due to pixel-by-pixel comparison of classified images with reference grids. *International Journal of Remote Sensing*, 16: 581-587.

Woodcock, C.E. and Gopal, S. 2000. Fuzzy set theory and thematic maps: accuracy assessment and area estimation. *International Journal of Geographical Information Science*, 14: 153-172.

Zadeh, L.A. 1965. Fuzzy Sets. *Information and Control*, 8: 338-353.

Global Strategy to Improve Agricultural and Rural Statistics