# GAP ANALYSIS AND PROPOSED METHODOLOGIES FOR ESTIMATION OF CROP AREA AND CROP YIELD UNDER MIXED AND CONTINUOUS CROPPING

***Research on Improving Methods for Estimating Crop Area, Yield and Production under Mixed, Repeated and Continuous Cropping***

**December 2015**

**Working Paper No. 4**

# Global Strategy Working Papers

Global Strategy Working Papers present intermediary research outputs (e.g. literature reviews, gap analyses etc.) that contribute to the development of Technical Reports.

Technical Reports may contain high-level technical content and consolidate intermediary research products. They are reviewed by the Scientific Advisory Committee (SAC) and by peers prior to publication.

As the review process of Technical Reports may take several months, Working Papers are intended to share research results that are in high demand and should be made available at an earlier date and stage. They are reviewed by the Global Office and may undergo additional peer review before or during dedicated expert meetings.

Gap Analysis and Proposed Methodologies for Estimation of Crop Area and Crop Yield under Mixed and Continuous Cropping

Under the project

# Research on Improving Methods for Estimating Crop Area, Yield and Production under Mixed, Repeated and Continuous Cropping

Drafted By

**U.C. Sud**
**Tauqueer Ahmad**
**V.K. Gupta**
**Hukum Chandra**
**Prachi Misra Sahoo**
**Kaustav Aditya**
**Man Singh**
**Ankur Biswas**

**and**

**ICAR-Indian Agricultural Statistics Research Institute**
**New Delhi, India**
**2015**

# Executive Summary

Agriculture is the most important segment of many national economies, particularly those of developing countries. Agricultural statistics are essential for monitoring trends and estimating future prospects for agricultural commodity markets, which can be useful in setting policies involving, for example, price supports, imports and exports and distribution. In particular, reliable and timely availability of crop statistics, namely crop area, yield and production plays an important role in planning and allocating resources for the development of the agriculture sector. The availability and quality of agricultural statistics has been declining in developing and underdeveloped countries. Some countries even lack the capacity to produce a minimum set of data as evidenced by the poor response rates to the Food and Agriculture Organization of the United Nations (FAO) questionnaires (FAO 2010). The Global Strategy to Improve Agricultural and Rural Statistics is a ground breaking effort to strengthen agricultural statistics. The United Nations Statistical Commission endorsed the technical content and strategic directions of the Global Strategy urging the rapid development of an action plan for implementation (hereafter, Global Action Plan) at its forty-first session, which was held in February 2010. One of the issues identified in the Global Strategy under the component data collection methods is estimating crop area, yield and production in the context of mixed, repeated and continuous cropping. Accordingly, a study entitled "Improving Methods for Estimating Crop Area, Yield and Production under Mixed, Repeated and Continuous Cropping" has been awarded to the Indian Agricultural Statistics Research Institute, New Delhi by FAO.

Under this study, several technical reports are being documented. This report is the second in the series which describes the gap analysis that exists in the context of crop area and yield estimation, and in general, focuses more on crop area and yield estimation in the mixed and continuous cropping scenario. It also includes methods of crop area apportioning under the mixed/intercropping scenario and data collection in field testing countries. Based on the gap analysis, an appropriate methodology for estimation of crop area and crop yield under mixed and continuous cropping has been developed. In particular, the domain estimation approach has been proposed in which various crop mixtures are considered as domains. Different measurement methods are proposed for determining crop area and crop yield. In addition, the sample survey-based approach for estimation of crop area and crop yield is suggested.

Selecting an appropriate sample size is an important consideration in any survey. The sample size can be reduced considerably by making use of available relevant auxiliary information in the statistical system. In situations in which the relevant auxiliary information is not readily available, which is often, the double or two-phase sampling approach is used to generate relevant auxiliary information at a nominal cost. This approach has been used in the developed methodology. Using the double sampling approach, the sample size can be reduced substantially by combining data from objective measures such as Global Positioning System (GPS) and rope and compass for crop area estimation, and the auxiliary variable, such as farmer self-reported area or family size, for crop area estimation and the farmers' appraisal/recall of crop production in the case of crop yield. Three approaches are proposed for crop area and yield estimation for mixed and continuous cropping, namely the cadastral map approach, the area frame approach and the household approach.

Apportioning of crop area to the component crops in the crop mixture is a complex problem. Some of the commonly used methods for apportioning the crop mixture area into component crops are amount of seeds used, plant density or row ratio and eye estimate. These methods, though simple, are subjective in nature. Accordingly, an objective method is proposed for apportioning crop area into component crops based on physical observation on plant count for component field crops grown in the same season. Also, methods are proposed for apportioning crop area in cases of mixtures involving annual and seasonal crops, annual and annual crops or annual and perennial or perennial and perennial crops.

To test the developed methodologies, a field test will be carried out in three countries,, namely Indonesia, Rwanda and Jamaica. Data collection is carried out using a traditional paper-based questionnaire and a computer-assisted personal interviewing software for a personal digital assistant (PDA). Therefore, both of these approaches are used in the study.

# Table of Contents

# Abstract

This technical report provides the gap analysis in the context of mixed and continuous cropping with regard to crop area and yield estimation. Based on gap analysis, an appropriate methodology is developed for carrying out estimation of crop area and crop yield in the mixed and continuous cropping scenario. Various concepts and definitions of survey sampling theory related to the developed methodology are given. Several alternatives are considered in this regard depending on the availability of information in the agricultural statistics system. The different methods of area apportionment of crop mixture to various component crops are explained. Questionnaires are designed for data collection on crop area and crop yield for the different alternative methodologies developed. The importance of computer-assisted personal interview software for the efficient collection of survey data is highlighted. The report ends with some conclusions.

# Acknowledgements

The authors express thanks to the Food and Agriculture Organization (FAO) of the United Nations, Rome for awarding this study to ICAR-Indian Agricultural Statistics Research Institute, New Delhi. The authors would also like to thank Naman Keita, Christophe Duhamel and Michael Austin Rahija of FAO for their constant support and help in completing the various activities pertaining to this study efficiently and in a timely manner. The authors are grateful to Consuelo Señoret of FAO for her continuous administrative support. The authors are also grateful to Prof. Raj S. Chhikara for improving the quality of the technical report. The authors gratefully acknowledge the suggestions made by peer reviewers and experts present in the expert meeting held at FAO in Rome, which led to further improvement of the report.

# Acronyms and Abbreviations

| | |
|---|---|
| CAPI | Computer Assisted Telephone Interviewing |
| FAO | Food and Agriculture Organization of the United Nations |
| FSUs | First Stage Units |
| GIS | Geographic Information System |
| GPS | Global Positioning System |
| NASS | National Agricultural Statistics Service |
| NDVI | Normalized Difference Vegetation Index |
| PDA | Personal Digital Assistant |
| PPS | Probability Proportional to Size |
| PSUs | Primary Stage Units |
| SAC | Scientific Advisory Committee |
| SDLC | Software Development Life Cycle |
| SRSWOR | Simple Random Sampling Without Replacement |
| SSUs | Second Stage Units |
| USDA | United States Department of Agriculture |
| USUs | Ultimate Stage Units |

# 1

# Gap Analysis and Proposed Methodologies for Estimation of Crop Area and Crop Yield under Mixed and Continuous Cropping

## 1.1. INTRODUCTION

Information on crop area, yield and production plays a vital role in planning and allocating resources for the development of the agriculture sector. Reliable and timely information on crop area, yield and production acts as a fundamental source of input to planners and policy makers for formulating efficient agricultural policies and making important decisions with respect to procurement, storage, public distribution, import, export and other related issues. The availability of crop area statistics is an essential requirement of an agricultural statistics system for any country as it functions as a key variable for estimating crop production and crop yield. Currently, both subjective and objective methods are being used worldwide to collect crop area statistics. The subjective methods, often used in developing countries, include, field reporting system, eye estimation and interview of farmers. These methods have certain limitations in terms of providing reliable data of crop area. Objective methods of measuring area, such as the polygon method are expected to give reliable crop area estimates, but they tend to be difficult to undertake, costly and time consuming. In addition, under certain unusual and problematic situations, such as fields being of an irregular shape and having irregular boundaries, it is difficult to measure area using these methods. In such situations, modern technology-driven methods, such as GPS, have the potential to provide reliable estimates of crop area. These situations become even more complicated in hilly regions, where crops are grown on slopes, and the measurement of area is not straightforward.

Similarly, both subjective and objective methods are currently being adopted for the collection of yield statistics. Among the subjective methods of estimating crop yield is farmers' assessment, expert opinion and crop card, while the objective methods include, for example, whole plot harvesting and crop-cutting experiments. Similar to crop area statistics, the crop yield statistics can be realistically captured by using the objective methods.

The practice of sowing crops in a mixture in the same parcel of land is prevalent in almost all counties, particularly in countries where landholdings are small. The advantages of growing crops in a mixture are that it provides protection to farmers in adverse weather conditions, such as drought, flooding and pest and disease infestation and it enables farmers to optimize the use of available space, moisture and nutrients in the field. Cultivators usually mix crops that cannot survive in under a particular type of weather conditions with another set of crops that can thrive under the same conditions. The methods employed for sowing such crops vary not only from region to region but also from area to area and even from field to field within the same region. The crops in the mixture are sown either individually in separate rows (intercropping) or mixed together. In the former case, the seeds of constituent crops are kept separate and a certain number of rows of one crop alternate with those of another crop. In the latter case, the seeds of two or more crops are mixed together before sowing and the mixed seeds are either sown in a row or broadcasted. Calculating the area for each crop grown in a mixture is more complex when the number of crops in a mixture increases, proportions of different crops in the mixture vary from field to field, sowing/planting and harvesting time differ, and the growing period (vegetative cycle less than three months to more than a year) of crops have different lengths. The number of crops grown in a mixture in a field may vary according to the growing period of the constituent crops. Accordingly, it is not possible to estimate crop area of constituent crops in a mixture in a single visit.

Sowing and harvesting the same or different crops on the same piece of land one after the other during the agricultural year is referred to as continuous cropping. It is popular because it enables farmers to grow more than one crop on the same piece of land, saving them the need to acquire more land, and tends to be more profitable than traditional cropping systems. Continuous cropping is more common among farmers with small land holdings as it gives them the opportunity to use their land repeatedly, thereby ensuring them with economic support and livelihood. For example, a farmer may grow paddy in different seasons of the agriculture year or wheat in the next season and then maize in the following season. Continuous cropping requires more specialization with regard to crop management, farm implement/equipment and marketing knowledge, among other things.

There are different ways of growing crops on the same piece of land in continuous cropping. Depending on the style of sowing crops, some popular forms of continuous cropping are as follows:

- **Continuous planting/harvesting:** The same or different crop is sown and harvested at a fixed interval on the same piece of land in an agricultural year. This practice is very common in many African countries that grow such crops as cassava.

- **Successive cropping:** One crop is planted after the other crop is harvested. The crops may be the same or different and may be sown and harvested more than once in the same field during the agricultural year. Total (cropped) area is equal to the area sown at a different time under the same or different crop in an agricultural year. It is greater than the physical area (net area). As an example, paddy is being grown during the three different seasons in the eastern part of India. A combination of paddy-wheat-green gram is prevalent in successive cropping in the Northern part of India.

- **Repeated plantings**: In places where crops may be damaged often during the agricultural year because of, among other things, drought, hailstorms and floods, locust and other pest invasions, repeated planting is practiced. This may be in the form of replanting or enlarging the plot gradually.

In some cases, the crop is sown/planted successively as and when the field is ready for sowing/planting such as in wet conditions, the new plants are intercropped with the older ones or planting of new plants after uprooting those plants that have already been harvested. As in case of mixed cropping, multiple field visits to the same farm may be required to realistically capture crop area and yield in the context of continuous cropping.

# 2

# Gap Analysis

Crop production estimates are generally obtained as the product of the crop area and crop yield. Accurate estimations of both the crop area and crop yield are equally important in ensuring accurate determination of their product. Although more emphasis is directed on the methodology related to crop yield estimation, estimation of crop area is equally challenging as there are many complexities associated with carrying it out. First of all, the area planted for harvest of a given crop may change throughout the growing season because of extreme weather, damage or unusual economic conditions. Under such situations, it may be possible to measure first the area intended for planting, followed by measuring the early sown area and then measuring the harvested area at the end of the season. Among the three situations stated above, measurement of the early sown area is the most challenging.

Area estimation in countries beset with problems of drought or floods can be cumbersome. The problem is further compounded when more than one crop is grown on the same piece of land and the planting and harvesting periods of the different crops vary. Landscape factors, such as elevation and/or soil type, also create problems, for example, terraced crops versus those crops planted on a steep slant.

Depending on the availability of resources, such as adequate manpower and funds, crop area estimation is carried out through two different approaches namely, complete enumeration and sample survey. As sampling has evolved, lists of farms began to be sampled, rather than collecting information on all farms, to obtain crop-related information. However, it is expensive to maintain these lists, which may be incomplete or become obsolete, leading to biased estimates. Area sampling has evolved as another approach under which area defined populations are complete and their estimates are unbiased. However, there are also problems with this approach, such as poor lower level area estimates of minor crops. Moreover, the initial cost of creating the area sampling frame is usually very high. Therefore, a combination of area and list frames, known as a multiple frame, has evolved as a better approach. This approach provides a more comprehensive, up-to-date and complete frames and lessens the problems associated with the two approaches when used individually.

An alternative approach for generating crop area statistics is using administrative sources, which can be collected as by-products of administrative data. For example, farm programme sign-ups or mail surveys may ask for changes in the past data leading to ratio estimates based on the previous year's figures. For cotton crop, information may also be provided by the processors of raw crop inputs, such the figures obtained from ginning and pressing mills. Different countries estimate crop area using different approaches; some publish survey data directly while others publish only official estimates based on a panel of experts.

Technological advances, such as remote sensing and geographic information system (GIS) have greatly aided crop area estimation. These tools/techniques have made it much easier to carry out area frame construction, maintenance and sampling. The associated costs have decreased and sampling variation has been reduced through better stratification and later through regression estimates. Improvement in the resolution of sensors and methodology, as well as increased computing power, has significantly improved estimates over time. Currently, LISS-III sensor with 23.5m, and LISS IV with 5.8m resolution are being used. Remote sensing imagery is used for both stratification and estimation. The estimates are not simply sums of local area information, but are also direct expansions of statistically sampled data, crop specific pixel classifications, through error corrected regression and/or calibration estimation. The small area estimation approach is also being used in crop area estimation. It provides timely and reliable area estimates at a smaller level.

## 2.1. GAP ANALYSIS FOR CROP AREA ESTIMATION

Developed countries and some developing countries have established systems for crop area estimation. These systems may be based on censuses or sample surveys. Some countries use area frame and multiple frame-based sample surveys or grid and point based systems or regression estimates based on a combination of survey data and remotely sensed data. Thus, globally, there is still substantial variation in methods and practices used for crop area estimation. It is, therefore, important to examine which method applies to which country and under what conditions. Some of the major issues on which the gaps in these methods can be identified in terms of judging the suitability of each method to be adopted by a particular country are accuracy, cost, complexity, timeliness and availability of existing data.

The degree of accuracy varies from the national level to the small area level. Currently, accuracy measures seem to vary at the national level between 1 percent and5 percent or even less. Statistical measures are rarely available and are generally unknown at the small area level. Over the next five to ten years, there appears to be a great opportunity for developing new measurable accuracy measures of small area crop estimation at the 5 percent level. This seems to be a major gap in the existing methods.

Another major issue is the cost in framing the methodology for crop area estimation. It is recommended that some of the upcoming research be aimed at cost reduction options. These may include use of technologies, such as PDAs/tablets/handsets and GPS, or new sampling methods, such as transect or road network sampling. The use of GPS for measuring an area may require a substantial initial investment, but in the long run it tends to be very accurate and can be used to address a variety of demands in other sectors.

The degree of complexity of a system is also a major factor for a country in determining whether to adopt it. It is always preferable to adopt a simpler system in terms of sampling design, data collection and estimation procedures.

Timeliness in providing the crop area estimates makes a system more desirable to be adopted by any country. The importance of timely information on crop area statistics cannot be ignored by planners and makers for formulating efficient policies and taking decisions with respect to procurement, storage, public distribution, import, export and other related issues at the national level. The other gap analysis opportunities in crop area estimation are: (i) Identification of different typologies from countries/regions regarding landscapes, economical structure, size of agricultural exploitation, the crops and livestock spatial distribution, households and social diversity; (ii) characterization of variables of interest for various typologies identified for constructing the master sampling frame; (iii) determination of the recommended frame for each type of landscape, such as point frame, square segments and segments with physical boundaries.; (iv) analysis of the remote sensing requirements for sampling frame construction; (v) characterization of different sources of information, available or required, for the construction of the master sampling frame (previous sampling frames, cartographic material – paper maps, digital maps, ortophotographs, previous years' satellite images), as well as the appropriate scaling; (vi) design of automated GIS procedures to model master sampling frames; and (vii) design of processes for field data collection, geolocation of elements and surfaces measurement using hand-held computers (PDA) and other equipment.

Another major gap that needs to be addressed is the development of emergency survey infrastructure for major flood and drought situations during the crop growth period.

Newer technologies, such as space technology, including, among others, remote sensing GIS and GPS, has revolutionized the process of estimating crop area statistics, but they have certain limitations. Also, as the technologies are constantly changing, new ones are introduced every year. These technologies must be closely examined.

The topic of crop area estimation is very mature. Multiple methods, including area sampling, list sampling and point sampling with or without a grid, have proven to be successful. Thus, there is less of a gap opportunity in crop area estimation. One area that probably may require continuous research is development of more cost-effective methods for crop area estimation, especially in developing or underdeveloped countries. Thus, any research aimed at reducing the cost of sampling frame construction or data collection would play an important role in developing efficient methods of crop area estimation. Perhaps, even transect sampling or road grid sampling could be studied, as well, with the intent to reduce costs.

## 2.2. GAP ANALYSIS FOR CROP YIELD ESTIMATION

A wide range of methodologies for crop yield estimation are available and being used in developed and developing countries. The two most widely used methods in a number of countries, farmers' estimation and crop-cut methods have their own inherent biases and faults. The farmer estimation method is quicker and cheaper but may result in poor quality data because of intentional over- or underreporting of crop production. The problem is further aggravated because of illiteracy among the farmers and inappropriate timing of the interview before/after the crop harvest. The crop-cut method, on the other hand, has been regarded as a reliable and objective method for estimating crop yield. The method, however, may be accompanied with an inherent upward bias stemming from measurement errors and increased cost and time. These drawbacks, however, can be largely overcome by appropriate training and supervision and through the use of optimum sample sizes and auxiliary data available in the system. A strong advantage of the crop-cut method is that the area of the cut is known and therefore does not introduce an error into the final yield computation.

The choice of these two methods is country specific and varies from one country to another according to such factors as their agricultural statistics system and data availability Therefore, for crop yield estimation, more studies need to conducted indifferent countries to determine which one is more effective.

In addition, other methods are being used to estimate crop yield in different countries. The appropriateness of each method depends on the scale of measurement and available resources. In many countries, the modelling approach is being explored for estimation of crop production. Efforts are being made to combine models that utilize survey and administrative data with remote sensing data. An attempt has also been made to describe a process of incorporating remotely sensed Normalized Difference Vegetation Index (NDVI) data into an empirical Bayes model (Bellow and Lahiri2010). Combining the best available survey data, administrative data, such as precipitation and temperature, and remotely sensed information into modelling efforts could potentially improve estimates of crop production. The basic approach of combining data from a variety of sources in estimating crop yield is being explored by the National Agricultural Statistics Service (NASS) of the United States Department of Agriculture (USDA), in the context of small area estimation.

USDA-NASS has very strict procedures for entering a selected field and placing randomly selected plots for crop counting and measurements throughout a season, such as corn ears, row spacing, corn ear length and circumference. The procedures are expensive and the sample sizes are relatively small but can provide state-level estimates. NASS attempts to keep non-sampling errors to a minimum by imparting training and providing a detailed procedural manuals to the field investigators/enumerators. In addition, it designates a small quality control sample that is checked by supervisors to see if the prescribed procedures are being followed properly. As cost is an issue, the procedures need to be made more efficient in order to be able to transfer this type of methodology, especially to developing or underdeveloped countries.

In most countries, the central statistical offices do not have the structure in place at the subnational levels to conduct objective methods of crop yield estimation. They, instead, rely on extension workers for the crop yield estimation. In most cases, there is a lack of coordination between the central statistics office and extension services for conducting the surveys properly.

Many countries are evaluating the effectiveness of remote sensing to estimate crop yield. It appears likely that remote sensing will become the keystone of agricultural statistics in the future (Zhao et al.2007). Improvement in the resolution of sensors and methodology, as well as increased computing power, has clearly improved the accuracy of the estimates over time. Remote sensing imagery is used for both stratification and estimation. The use of satellite imageries as area frame and crop cutting experiment data from a list frame of villages provides multiple-frame estimators of crop yield that are more precise in comparison to single-frame estimators (Das and Singh 2013). However, considerable research is still needed before remote sensing can be widely applied to estimate crop yield with high accuracy.

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 remote sensed information could also be represented as an auxiliary variable in the procedures used for small area estimation. The remote sensing data can be exploited to estimate the crop production using their link with the yield.

The major gap analysis opportunity in crop yield estimation seems to be conducting more comparative studies in different countries to determine which method is the most effective, carrying out considerable research using remote sensing and GIS technologies in order to estimate crop yield with high accuracy, developing suitable models combining data from different sources in order to forecast the crop production more accurately and control of various non-sampling errors in yield estimation surveys. Also, efficient sampling plans need to be developed to estimate crop production.

## 2.3. GAP ANALYSIS FOR CROP AREA AND YIELD ESTIMATION UNDER MIXED AND CONTINUOUS CROPPING

Mixed cropping is a common phenomenon in many countries. Several approaches have been attempted for apportioning an area under each crop in cases of mixed cropping and estimating average yield. It is a widely used practice by small farmers in many countries, especially in developing countries, as a subsistence need. In a “mixed cropping” scenario, several crops are planted at the same time in the same field with the seeds mixed or planted in rows in some given ratios. Harvesting time of the crops in crop mixture may not always be the same. Hence, the enumerators may be required to make repeated visits in order to estimate yield.

In a “continuous cropping” scenario, the crops are sown and harvested from the same piece of land previously occupied by another crop, or even by the same crop, during the same agricultural year.

Preparation of a sampling frame, sample selection, measurement of area, namely area apportioning, proper sampling plan and suitable estimation procedures as per sampling design, are the major problems in the above situations. To estimate average yield under each of the above situations, different approaches need to be used.

For the estimation of crop areas under continuous cropping, the number of rounds of surveys required depends on the number of different configurations of the crops in the fields. This is a major point of consideration. In a system of regular periodic reporting (monthly, bi-monthly or quarterly, the number of rounds for estimation of crop areas under a system of continuous cropping could be the same. When and how often the data on yield should be collected in the case of continuous cropping are other important points for consideration.

Out of the several methods available for estimation of crop area and crop yield, no matter which method is adopted for crop area or crop yield in a specific area, accurate estimates in the scenario of smallholder agriculture may be complicated by the following: (1) heterogeneous crop performance; (2) continuous planting; (3) mixed cropping; (4) staggered harvesting for crops with an extended harvest period; and (5) planted area not being equal to the harvested area. Two other issues that influence data quality and interpretation are the level at which data are aggregated and the variation of crop yield over time. The gaps in the above context can be discussed as:

**Heterogeneous crop performance**

In any smallholder farming system, crop performance may be heterogeneous both within and between plots, enumeration areas and higher administrative levels. The presence of intercrops, trees, local variability in soil characteristics and non-uniform farm management are some of the factors that contribute to the heterogeneity at the plot level. The crop-cut method that estimates production based on cut plots or subplots in a plot have a larger variance than production estimates that are based on a whole plot observation. However, by increasing the number of subplots sampled, the reliability of production or yield estimates can be improved.

**Continuous planting**

Farmers may plant and harvest crops throughout the year based on the distribution of rainfall. This may complicate the estimation of crop area and production. In order to capture continuous cropping patterns, the crop area for each sample holder can be recorded through multiple visits by the enumerator during the year.

**Intercropping or mixed cropping**

Several approaches have been tried for apportioning an area under each crop and for estimating average yield. Four strategies are commonly used to estimate crop area, production and yield in farming systems that have an important proportion of crops produced under intercropping (Kelly et al. 1995). Each strategy provides resultant estimates with varying degrees of accuracy. Strategy 4 proportionally allocates area to each intercrop and attempts to remove the impact of intercropping on areas and yield as it adjusts all areas and yield to pure stand data.

The crop yield in intercropping systems is affected by a larger range of variables and thus the yield may vary more widely than crop yield in pure stands. Therefore, a larger number of observations are required to obtain yield estimates with an acceptable confidence interval for intercropped fields than for pure stands.

Relay cropping and the mixing of annual with seasonal crops complicates the situation. Sequential cropping or growing cycles that exceed a single season or calendar year complicate the estimation of crop production and crop yield further. No simple solution is available for such situations. No strategies have been developed that take into account more complex intercropping situations, such as relay cropping and the mixing of annual with seasonal crops.

**Crops with an extended harvest period**

There are problems in crop yield estimation for the crops that have an extended harvest period, such as cassava, sweet potato, banana, cotton, coffee and indeterminate legumes. Even crops such as maize and beans, may be harvested at two or more stages (green maize and dry maize). Extended harvest periods may be the result of better in-ground than out-of-ground storability (cassava, sweet potato), continuous planting (all crops), uneven ripening or filling (banana, cassava) and multiple harvest products from one crop (green and dry maize).The crop-cuts and whole plot harvesting methods do not take into account the extended harvest period of the crop under study. These methods cannot be used to estimate banana yield due to the fruit's uneven ripening throughout the year. There have been limited studies that evaluated various methods to estimate yield for crops with an extended harvest period.

**Planted area not being equal to harvested area**

Crop area may be defined in terms of the area planted or the area harvested. The planted area may not always be equal to the harvested area for a variety of reasons, including, among them, poor germination, pest or disease damage, animal grazing, floods and lack of labour or market. Some crops, such as cassava, are grown to meet the emergency needs of a farmer and are only fully harvested during drought or a food shortage. In any of the above circumstances, the definition of crop area has a major influence on area, yield and production estimates.

In addition to the above-mentioned issues, the level at which data are aggregated and the variation of crop yield over time need influence data quality and interpretation and need to be addressed while estimating crop yield. Therefore, it should be noted that among the issues associated with estimating crop area and yield under mixed and continuous cropping scenario, the problem of area apportioning in general and the area apportioning under different situations in particular, such as perennial and annual crops, annual and seasonal crops, perennial and perennial crops and when the component crops are harvested in different seasons, remains a challenge for the researchers and should be addressed judiciously.

# 3

# Methods of crop area apportioning under a mixed/intercropping scenario

When more than one crop is sown simultaneously (crops are sown in mixtures), field-wise recording of an area is a challenging job. The crop mixture area must be apportioned into the component crops in order to estimate the area under various crops. The area can be apportioned by eye estimation or by using any objective method. Apportioning of an area through eye estimation is based on the experience of the enumerator, which may vary. This can lead to an erroneous estimation. Apportioning of area using an objective method is therefore a preferred option to get accurate estimates of a crop area, though it may be expensive and time consuming. The method has the potential to keep the importance of each component crop in a mixture provided there is sufficient plant density (> 10%) in each of the component crop. The area of a field is apportioned between the component crops in order to "adjust" the observed area to pure stand. The area among the various component crops can be apportioned using the quantity of seed sown, plant density, row ratio and physical area occupied by each component crop.

- **Apportioning using seed rates**

  When component crops are sown as under a mixture scenario without any row arrangement, the area under each component crop in the mixture may be apportioned on the basis of adjusted seed rate. In this situation, only the seeds used are considered for apportioning the area; the population of plants in the field is not taken into account. It may be possible that seeds sown have not germinated 100 per cent or the survival of plants is not optimum. The area apportioned, however, to each crop on the basis of seed rate may give biased estimates when all the seeds sown are not germinated.

If *a* kg and *b* kg are the quantities of seeds used for sowing a mixture of two component crops and *A* kg and *B* kg are their normal seed rates when sown as pure, the proportion of the area under each component crop may be estimated as *a/A* :*b/B*.

For illustration, let there be two crops, say, *A* and *B*, in the mixture.

Area under crop mixture = 0.4 ha, quantity of seed used for sowing of crop A in crop mixture A and B = 50 kg, normal seed rate of crop A = 120 kg/ha. Quantity of seed used for sowing of crop B in crop mixture A and B = 1 kg, normal seed rate of crop B = 5 kg/ha. Then the seed rates ratio = (50/120): (1/5) = 0.42: 0.2.

Area of crop A in the mixture of crop A and B = (0.42/(0.42+0.2)) x 0.4 = 0.27 ha

Area of crop B in the mixture of crop A and B = (0.2/(0.42+0.2)) x 0.4 = 0.13 ha

When apportioning of crop area in mixed cropping is carried out on the basis of the seed rate, the reporting of the seed rate by the farmer may be inaccurate. Also, the size of the seed (or test weight) may influence the area of each component crop, which may result in incorrect apportioning of crop area.

- **Apportioning using plant density**

When component crops are sown under a mixture scenario without any row arrangements, the area under each component crop in the mixture may be apportioned on the basis of adjusted plant density. The plants' density per unit area for each component crop in the crop mixture is worked out on the basis of an objective method (the counting the number of plants in a randomly selected small plot) and the area of each component crop can be estimated by calculating the plant density ratio.

Plant density ratio = average number of plants of component crop *A* per unit area/ number of plants per unit area of component crop *A* in pure stand : average number of plants of component crop *B* per unit area/ number of plants per unit area of component crop *B* in pure crop stand.

Let there be three crops in the crop mixture having 0.8 ha area.

Let there be 100 plants of crop A when sown in crop mixture and further let there be 2500 plants when crop A is sown as pure.

Let there be 18 plants of crop B when sown in crop mixture and further let there be 25 plants when crop B is sown as pure.

Let there be 80 plants of crop C when sown in crop mixture and further let there be 200 plants when crop C is sown as pure.
Plant density ratio =100/2500: 18/25: 80/200 = 0.04 : 0.72 : 0.4
Area of crop A in the crop mixture = (0.04/1.16) x 0.8 = 0.0276 ha,
Area of crop B in the crop mixture = (0.72/1.16) x 0.8 = 0.4965 ha,
Area of crop C in the crop mixture = (0.4/1.16) x 0.8 = 0.2759 ha,

- **Apportioning using row ratio** (row intercropping)

When the component crops are sown in separate rows, the area under each component crop may be apportioned on the basis of row ratio of each component crop. The number of rows in a specified length is counted at three places randomly in the selected field to determine the average number of rows.
Example: A and B are intercropped.
Ratio of crop area = (Number of rows of component crop A in a mixture in a specified length / number of rows of crop A as per normal spacing in pure stand in a specified length): (number of rows of component crop B in a mixture in a specified length / number of rows of crop B as per normal spacing in pure stand in a specified length).
Let the intercropped area be equal to 0.8 ha;
Let, rows of component crop A in specified length in intercropping = 32;
Let, rows of crop A in specified length in pure stand = 40;
Let, rows of component crop B in specified length in intercropping = 1;
Let, rows of crop B in specified length in pure stand = 5;
Therefore, crop rows ratio = 32/40 : 1/5 = 0.8 : 0.2.
Area of component crop A in intercropping of crop A and B= (0.8/(0.8+0.2)) x 0.8 = 0.64 ha,
Area of crop component B in intercropping of crop A and B = (0.2/(0.8+0.2)) x 0.8 = 0.16 ha.

- **Apportioning using width ratio** (strip cropping)

For intercropping, where each component crop is sown in a separate distinct group of rows say “strip”, the width of the strip of each component crop is measured randomly at three places. The width ratio between component crops is determined on the basis of average width of the component crop in the strip intercropping. Therefore, the area of component crops can be apportioned using width ratio of component crops.

Example: Let A and B be the component crops.

Ratio of crop area = (Average width of strip of component crop A in the strip intercropping: average width of strip of component crop B in the strip intercropping)

Let the intercropped area be equal to 1.0 ha.

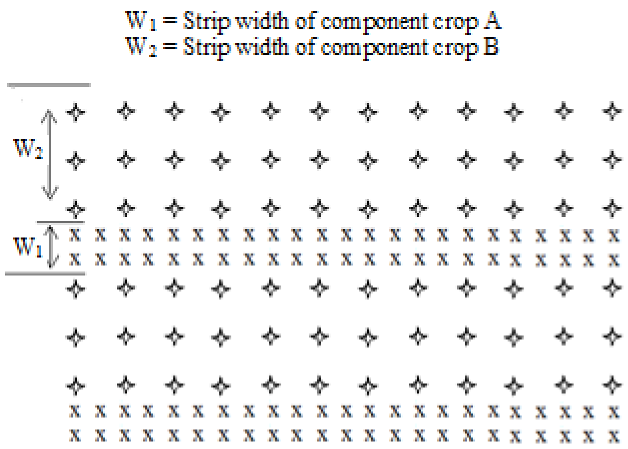
Let average width of strip of component crop A in strip intercropping = 2 m

Let average width of strip of component crop B in strip intercropping = 3 m,

Therefore, width ratio of component crop A and B = 2:3

Area of component crop A in strip intercropping = (1.0/(2+3)) × 2 = 0.4 ha,

Area of crop component B in intercropping = (1.0/(2+3)) x 3 = 0.6 ha.




Apportioning of area on the basis of width ratio

- *Fixed area ratio:* Under complicated crop mixture (large number of crops in the mixture), the area under such a crop mixture is recorded at lower level and aggregated at higher level. Apportionment of an area under each crop may be done at a higher level using a fixed area ratio determined through subjective eye estimates at a periodical interval.

- *Ignore crops occupying a very limited area in a crop mixture:* Crops comprising an extremely small proportion in the mixture of less than 10 percent may be ignored. In cases in which two crops are in the mixture, the entire area may be shown as pure.

- *Total field area sown is divided equally to each component crop:* It is a simple method but it will give over/ underestimates of crop production.

- *Allocate total area sown to each component crop in the mixture*: The method is very crude and may therefore leads to overestimated areas.

- *Allocation of an area when component crops are harvested in different seasons:* When temporary crops (seasonal and annual crops) are sown in a crop mixture at the same time and harvested in different seasons, the entire area of the mixture is treated as double-cropped. The whole area is recorded under each component crop in the respective seasons in which the crop is harvested.

**Proposed method of determination of crop area of component crop in mixture by physical observation**

Generally, when crops are grown as mixed, the recording of an area under each component crop at the field level is difficult. The crop mixture may vary field-wise; even the seed ratio of the component crop may vary in each crop mixture. Therefore, apportioning the area under a crop mixture on the basis of physical observation may provide precise estimates of crop areas of component crops in the crop mixture. Under this situation, it is advisable to observe several data points for each mixture and then work out a fixed ratio for each crop mixture using the averaged value for apportioning the mixture area to each of the component crops. The fixed ratio can be determined on the basis of plant density of component crops of the various crop mixtures. The plant density of each component crop can be determined on the basis of physical observation to be taken in a randomly selected experimental plot. For this purpose, at least five and not more than 10 parcels of land from different villages need to be selected randomly within each stratum of the district for each crop mixture. The apportionment of area to component crops in a crop mixture can be carried out on the basis of plant density using a fixed ratio determined by taking the average value. This exercise can be carried out when crop-cutting experiments are being conducted to determine crop yield. When the crop mixtures are sown in rows, then the method of apportioning using row ratio may be used. For strip intercropping, the method of physical area occupied by each component crop is recommended.

**Sampling design for physical observation for apportioning crop area**

To apportion the crop area through physical observation, it is advisable to use a subsample of the already selected sample for yield estimation. The subsample should be representative of different crop mixtures. Approximately 5 to10 fields need to be selected randomly for each of the crop mixture in the district for physical observation for the purpose of apportioning the crop area.

To demarcate the experimental plot for physical observation, the procedure explained in section 10, field test protocol may be followed. As per the prescribed procedure, the size and shape of the experimental plot for crop mixture may be taken as 5 m × 5 m or 10 m × 10 m in the square shape. The plot size for different food and non-food crops are indicated below:

| Name of the crop | Shape | Length (m) | Breadth (m) | Diagonal (m) |
|---|---|---|---|---|
| Crop mixture as a combination of any of the following two or more crops: paddy, wheat, sorghum, pearl millet, finger millet, maize, groundnut, tobacco, sugarcane, green-gram, chilly, horse-gram, black-gram, sengal-gram, sunflower and similar crops | Square | 5 | 5 | 7.07 |
| Crop mixture involving red-gram, sesamum, caster, cotton and other similar crops | Square | 10 | 10 | 14.14 |

After the experimental plot is demarcated, the plant population of the plot should be counted by the following means:

A well stretched string should be tied around the pegs and lowered gradually to the ground. The position of the string on the ground demarcates the boundary of the experimental plot. The plants on the boundary of the experimental plot are to be counted only if the roots are more than half inside the experimental plot. The plants of each component crop should be counted and recorded carefully.

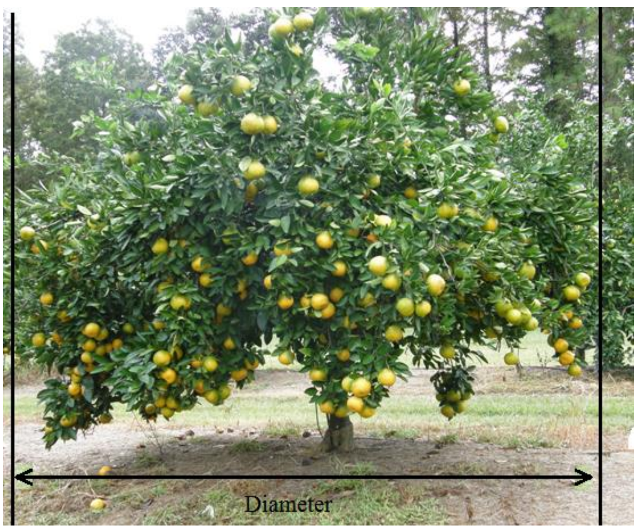
This approach of crop area apportioning can be used in the above mentioned situations. Some of the commonly encountered problems for crop area apportioning involving crop mixtures are discussed below:

Allocation of an area between temporary and permanent crop under mixed cropping is a debatable issue. The production estimates of permanent crops, such as mango, guava and apple, are based on product of per tree yield and number of trees. So, for these crops, the area under trees is not taken into account. Temporary crops grown in the orchards contribute towards production. So, if the area under temporary crops grown in the orchards is not accounted for, an overestimation of the crop yield of temporary crops may occur. For proper estimation of yield of temporary crops, crop areas in the orchards need to be accounted for. However, the area under temporary crops may be ignored if it comprises less than 10 percent of the total area. Under these situations, some of the proposed solutions are as follows:

**Allocation of area when temporary crops are sown with permanent crops:**

Generally, temporary crops are grown in the orchards until the bearing stage. After bearing, they are grown in the space between the permanent crops. Shade-loving crops are also grown in the orchards. The apportioning of crop area between permanent (fruit trees) and temporary (seasonal and annual) crops is a complex problem. The permanent crops are planted once and harvested (picking) when ripened. The crop remains in the field for more than a year occupying the whole area. As the temporary crops occupy a certain area of the orchard, the actual area occupied by temporary crops may be allocated to the temporary crops grown in the orchard. If more than one temporary crop is grown in the mixture, the area between the temporary crops must be apportioned. Under this situation the proposed solution is as follows:

The physical area occupied by permanent crops may be considered under the orchard. The area of the orchard may be estimated using the canopy area of sampled trees. The estimated area may be calculated on the basis of average canopy area ($\pi r^2$) of the three to five randomly selected trees, where *r* is the radius of the canopy. Multiplying the average canopy area per tree with total number of trees gives the total area of the orchard. This area should be deducted from the total area of the field and the remaining area should be allocated to the temporary crop grown in the orchard. In cases in which more than one temporary crop is grown in the orchard, the remaining area under the temporary crops can be apportioned using the method described above.



## Allocation of an area when component crops are permanent

In the case of mixtures of two or more permanent crops, the area may be apportioned on the basis of the number of trees of each component crop so as to adjust the area to pure crop. The number of trees or plants of each permanent crop should be recorded separately. The number of trees in the orchard can be determined by the owner of the orchard or ascertaining the area of component crops in the mixture may be carried out using physical observations as described earlier.

For the purpose of field testing, a comparison of the methods of apportioning using physical observation, seed used and eye estimate is discussed.

# 4

# Measurement methods and data collection in field testing countries

Crop area determined by using GPS and the rope and compass method should be used to determine crop area on a smaller sample to examine the accuracy of the GPS method of recording crop area. In addition, identification of auxiliary variables, such as farmer self-reported area, quantity of seeds used, family size and number of active family members, should be made in consultation with the field-testing countries. The identified auxiliary variables are to be used in the proposed two-phase sampling approach. Crop yield is proposed to be captured by following the crop-cutting experiment approach, expert assessment and sampling of harvested units, crop diary with telephone @ two telephone calls to the farmer per week (for continuous cropping). The accuracy of the proposed methods is to be examined against the whole field harvest method. Auxiliary variables need to be identified in consultation with the field-testing countries. This can be completed through various means, including, among them, farmer appraisal/recall of crop production. The two-phase sampling regression method for estimation of crop area and crop yield should be developed using the auxiliary information.

For the purpose of data collection, appropriate/suitable questionnaires have been designed along with an instruction manual for using /them. These materials are to be fine-tuned in consultation with the funding agency. The questionnaires are to be pre-tested and finalized before the actual data collection. Primary data collection in field testing countries is to be carried out using a traditional paper-based technique. However, computer-assisted personal interviewing (CAPI) software provided by the World Bank using PDAs is also going to be tested in one district in each of the field testing countries. A brief description of CAPI is as follows:

CAPI is a survey data collection technique involving in-person interviewer, or face-to-face interviewing in which a computer is used to administer the questionnaire to the respondent and the responses are captured in the computer. In CAPI, instead of collecting data on paper questionnaires, interviewers use portable computers to enter data directly through a keyboard. This interviewing technique is similar to computer-assisted telephone interviewing, except that the interview takes place in person instead of over the telephone. Although computer-assisted telephonic interview has been used only in the last few years or so, such face-to-face interviews are now being propagated. This method is usually preferred over a telephone interview when the questionnaire is long and complex. The interviewing technique is a relatively a new development in survey research that was made possible by the personal computer revolution that took place in 1980s. The CAPI technique is gaining popularity in survey data collection against the traditional paper-based method because it can be a more rapid way to collect quality data relatively inexpensively.

Some of the important points about CAPI are as follows:

**Data quality:** CAPI enhances the quality of survey data in a number of ways:

- Routing problems within the questionnaire are eliminated;
- Interviewers cannot miss questions or ask the wrong questions;
- Questions are "customized" correctly;
- Mathematical calculations can be carried out within the program;
- The computer checks are possible for inadmissible or inconsistent responses;
- Errors from separate data entry are eliminated.

The way CAPI handles routing is one of its most impressive features. Rather than having to decipher routing instructions during an interview, the computer program takes interviewers automatically to the next appropriate question. This is particularly important when the questionnaire includes complex routing. Similarly, if a set of questions has to be asked a number of times, the computer automatically repeats the questions the correct number of times and then moves on. The routing capabilities of CAPI have two main advantages over paper-based techniques. First, the possibility of error from interviewers failing to follow routing instructions is eliminated; they cannot follow a wrong route and ask inappropriate questions nor can they inadvertently skip over questions. Secondly, the interview flows much more smoothly as the interviewer does not have to keep referring to earlier answers to establish the correct route through the questions.

Interviewing is also made easier by the "customising" of questions. The computer program can recall a piece of data from its memory, such as a name or a date, and insert it at the appropriate place in a question. Another major advantage of CAPI is its ability to spot inadmissible or inconsistent responses that could be the result of either an interviewer or respondent error. CAPI also allows "logic checks"' that can identify inconsistent or contradictory responses. Range and logic checks are powerful features of CAPI, which improve the quality of data at source. Because CAPI interviewers enter data directly into a computer, separate data entry, familiar in paper surveys, is not required. This eliminates one source of error and saves time and money. Responses to open-ended questions can also be typed in directly.

**Timing:** The effect of CAPI on the timing of a survey is twofold. First, the process of converting a paper questionnaire into a computer program is time-consuming. The timetable for a survey using CAPI should, therefore, allow ample time between the design of a questionnaire and the start of fieldwork. This is particularly important if the fieldwork has to be carried out on specified dates. As mentioned earlier, CAPI eliminates a separate process of data entry. As a result, the time between the end of the fieldwork and the supply of a clean data set is reduced. However, there may still be a need for some data cleaning, albeit on a smaller scale than for a paper survey. For example, despite the best intentions of the program designers, some of the possible range or logic checks may not have been included. Hence, some inconsistent or invalid data may result, which will need to be checked, adding additional time to the process.

**Costs:** CAPI, compared with paper surveys, may be costly in some ways, but it can be cost effective in many other ways. In comparing set-up costs only, CAPI is always more expensive because of the time needed to convert paper questionnaires to a computer version. Therefore, the cost is greater for complex questionnaires than for more simple ones. Because each paper questionnaire has an associated data entry cost, the savings generated by CAPI increase as the size of the survey population increases. The costs of cleaning data are also higher for paper surveys as, in a CAPI interview, respondent and interviewer errors are rectified during the interview itself. Administration costs for paper surveys, which include the printing and distribution of questionnaires, also tend to be higher than for CAPI surveys.

This form of interview is substantially cheaper when a large number of respondents are required, because of the following:

- There is no need to transcribe the results into a computer form. The computer program can be constructed so as to place the results directly in a format that can be read by statistical analysis programs.

- The program can be placed on a web site, potentially attracting a world-wide audience.

The survey can miss feedback and clarification/quality control that a personal interviewer could provide. For example, a question that should be interpreted in a particular way, but could also be interpreted differently, can raise questions for respondents.

# 5

# Methodology for estimation of crop area and crop yield under mixed and continuous cropping

The purpose of this section is to describe the methodology for estimation of crop area and crop yield in the context of mixed and continuous cropping. Before explaining the methodology, some useful concepts and definitions are explained explicitly.

## 5.1. CONCEPTS AND DEFINITIONS

**i) Population**

The collection of all units of a specified type in a given region at a particular point or period of time is termed as a population. Thus, for example, a population of parcels of land, households, farmers or orchards in a region depending on the nature of the data required. Populations can be finite or infinite. This study is limited to finite populations.

**ii) Elementary unit**

An elementary unit, or simply a unit, is an element or a group of elements, on which observations can be made or from which the required statistical information can be ascertained based on a well-defined procedure. For example, in a household survey, persons residing in a household are an elementary unit, while for a crop yield estimation survey, a field or a plot may be an elementary unit.

### iii) Sampling unit

Elementary units or a group of such units, which besides being clearly defined, identifiable and observable, are convenient for the purpose of sampling, are called sampling units. For instance, in a family budget enquiry, usually a family is considered to be the sampling unit as it is found to be convenient for sampling and for ascertaining the required information. In a crop survey, a parcel or a group of parcels owned or operated by a household may be considered as the sampling unit. Every elementary unit must belong to one and only one sampling unit.

### iv) Sampling frame

A list of the sampling units belonging to the population to be studied with their identification particulars or a map showing the boundaries of the sampling units is known as a sampling frame. Examples of a frame are a list of parcels or a list of suitable area segments. The frame should be up to date and free from errors of omission and duplication of sampling units.

### v) Sample

A fraction of the population or a proper subset of the population is said to constitute a sample. The number of units (not necessarily distinct) included in the sample is known as the size of the sample and the number of distinct units as the effective size of the sample.

### vi) Random Sample

The sample is considered to be a random or probability sample, if its selection is governed by ascertainable laws of chance. In other words, a random or probability sample is a sample drawn in such a manner that each unit in the population has a predetermined probability of selection.

### vii) Non-random sample

A sample selected by a non-random process is termed a non-random sample. A non-random sample that is drawn using a certain amount of judgment with a view to get a representative sample is termed a judgment or purposive sample. In purposive sampling, units are selected by considering the available auxiliary information more or less subjectively with a view to ensuring a reflection of the population in the sample. This type of sampling is seldom used in large-scale surveys mainly because it is not generally possible to get strictly valid estimates of the population parameters under consideration as there also tends to be sampling errors because of the risk of bias in a subjective selection and the lack of information on the probabilities of selection of the units.

**viii) Population parameter**

Any function of the values of all the population units (or of all the observations constituting a population) is known as a population parameter or simply a parameter. Suppose a finite population consists of the N distinct and identifiable units $U_1, U_2, \ldots, U_N$ and let $Y_i$ be the value of the variable y, the characteristic under study, for the $i^{th}$ unit $U_i$, (i=1,2,…,N). For instance, the unit may be a parcel and the characteristic under study may be the area under a particular crop. Some of the important parameters usually required to be estimated in surveys are population total $Y = \sum_{i=1}^{N} Y_i$ and population mean $\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$ .

**ix) Statistic**

Any function of values of sample observations that is free from unknown population parameters is called a statistic.

**x) Estimator**

An estimator is a statistic obtained by a specified procedure for estimating a population parameter. The estimator is a random variable and its value differs from sample to sample. The samples are selected with specified probabilities.

**xi) Estimate**

A particular value that the estimator takes for a given sample is known as an estimate.

**xii) Sampling error**

An error resulting from drawing inferences about the population parameter on the basis of observations on a part (or sample) of the population is termed sampling error. The sampling error is non-existent in a complete enumeration survey as the whole population is surveyed. The sampling error usually decreases as the sampling size increases. **On-sampling error.**

Errors other than sampling errors, such as those resulting from non-response, incompleteness and inaccuracy of response or measurement are termed non-sampling errors. These errors are likely to be more widespread and important in a complete enumeration survey than in a sample survey. They are the result of various factors, starting at the beginning stage when the survey is planned and designed to the final stage when the data are processed and analysed. Non-sampling errors are likely to increase in line with the size of the sample,

**xiii) Unbiased estimator**

An estimator is said to be unbiased for the population parameter, if the expected value of the estimator is equal to the value of the parameter.

[If $X$ is random variable which assumes values $X_1$, $X_2$, ….,$X_k$ with probabilities $p_1$, $p_2$, …., $p_k$ such that $\sum_{i=1}^{k} p_i = 1,$ then expected value of the variable $X$ is defined as $E(X) = \sum_{i=1}^{k} p_i X_i$ ].

**xiv) Sampling variance**

One of the measures of sampling error is sampling variance. The sampling variance of an estimator $t$ is a measure of the difference of the estimator values from its expected value. Mathematically, sampling variance is defined as the expected value of the squared difference of the estimator t from its expected value, i.e. $Var(t) = E\left[t - E(t)\right]^2$ .

**xv) Standard error**

The positive square root of variance is termed as standard error of the estimator $t$, such as $SE(t) = \sqrt{Var(t)}.$

**xvi) Percentage standard error**

Percentage standard error of the estimator $t$ is defined as $\%SE(t) = \frac{SE(t)}{t} \times 100$

The sampling variance, standard error and percentage standard errors as defined above are based on population values and thus are unknown. The estimated values are obtained from the observed sample.

**xvii) Strata**

Stratification consists of dividing the population into homogeneous subgroups or subpopulations or subsets called strata. Thus, in a district, the various sub districts can be taken as strata.

**xviii) Cluster**

The smallest unit into which a population can be divided is called an elementary unit of the population and a group of such elementary units is known as a cluster. A cluster may be a group of villages in a district or a group of households within a village. Clusters are generally comprised of neighbouring elements.

**xix) Domain**

In many surveys, estimates are required for each of the classes/subpopulations into which the population is subdivided. The term domain refers to these subpopulations. Each unit in the population falls into one of the domains.

**xx) Probability sampling**

If the units in the sample are selected by some probability mechanism then the method of sampling is referred to as a probability sampling.

**xxi) Simple random sampling without replacement**

Simple random sampling without replacement (SRSWOR) is the simplest method of probability sampling. In this method, the units are drawn one by one, assigning equal probability of selection for each unit at each draw and the unit selected at any particular draw is not replaced back to the population before selecting a unit at the next draw. The probability of selecting $i^{th}$ population unit at any draw is the same as probability of selecting that unit at the first draw. This probability is the inverse of the population size.

**xxii) Cluster sampling**

A sampling procedure presupposes division of the population into a finite number of distinct and identifiable units called the sampling units. Many times, a sampling frame of elements is not available, but that of a group of elements known as clusters may be easily obtainable. Cluster sampling is a sampling procedure in which clusters are considered as sampling units and the elements of the selected clusters are enumerated.

For many types of populations, a list of elements is not available and the use of an element as the sampling unit is, therefore, not feasible. The method of cluster sampling is available in such cases. Thus, in a city, a list of all the houses may be available, but one of the persons is rarely available. Also, a list of farms is not available, but those of villages or enumeration districts prepared for the census are. Cluster sampling is, therefore, widely practiced in sample surveys.

For a given number of sampling units, cluster sampling is more convenient and less costly than simple random sampling because of number factors including, among them, the savings associated with travelling, and the identification and contacts. However, cluster sampling is generally less efficient than simple random sampling due to the tendency of the units in a cluster to be similar. In most practical situations, the loss in efficiency may be balanced by the reduction in the cost and the efficiency per unit cost may be more in cluster sampling as compared to simple random sampling.

**xxiii) Stratified sampling.**

Stratified sampling is a commonly used technique of sampling in which the population is divided into non-overlapping homogeneous subgroups or subpopulations called strata that together comprise the entire population and then an independent sample from each stratum is drawn. Stratification is used to increase the precision of population estimates as the technique can provide a representative sample of the population.

**xxiv) Two-stage sampling**

One of the main considerations in adopting cluster sampling is the reduction of travel cost because of the nearness of elements in the clusters. However, this method restricts the spread of the sample over the population, which may result in an increasing of the variance of the estimator. To increase the efficiency of the estimator with the given cost, it is natural to want to conduct further sampling of the clusters and select a greater number of clusters so as to increase the spread of the sample over the population. This type of sampling, which consists of first selecting clusters and then selecting a specified number of elements from each selected cluster, is known as subsampling or two-stage sampling, as the units are selected in two stages. In such sampling designs, clusters are generally termed as first stage units (FSUs) or primary stage units (PSUs) and the elements within the clusters or ultimate observational units are termed as second stage units (SSUs) or ultimate stage units (USUs). Observations are recorded only on the selected second stage units (or ultimate stage units).

**xxv) Two-phase sampling**

A sampling technique that involves sampling in two phases (occasions) is known as two- phase sampling. This technique is also referred to as double sampling. Double sampling is particularly useful in situations in which the enumeration of the character under study (main character) is costly or labour-intensive, whereas an auxiliary character correlated with the main character can be easily observed. Double sampling is also useful in situations in which auxiliary information is not available for all the population units or the population mean or population total of the auxiliary variable is not known. Thus, it may be convenient and economical to take a large sample for the auxiliary variable in the first phase, which would result in precise estimates of the population total or mean of the auxiliary variable. In the second phase, a small sample, usually a subsample is taken wherein both the main character and the auxiliary character may be observed and using the first-phase sampling as supplementary information and utilizing auxiliary information-based estimators, such as the ratio or regression estimators. By the two-phase sampling technique, it is possible to obtain precise estimators for the main character. For example, in estimating the volume of a stand, the diameter or girth of trees or height of tree as auxiliary variables may be used. In two-phase sampling, the observations on auxiliary variables are recorded on both the phases of sampling.

## 5.2. PROPOSED SAMPLING DESIGN FOR CROP AREA ESTIMATION IN MIXED CROPPING

In this section, the recommended sampling methodology for estimation of crop area under mixed cropping is described. Different scenarios are considered in this regard. For example, one scenario is when land cadastral maps are available (category-1) and thus registers exist containing parcel-wise area and crop area season-wise. For example, cadastral maps are readily available in India. Another scenario when an area frame exist (category-2). The area frame approach is suitable for such countries as Jamaica and Rwanda. Another scenario is when none of the above situations prevail. In this context, the household approach (category-3of collecting data on crop area and yield may be adopted. This approach is suitable for country, such as Indonesia.

**CATEGORY - 1: Sampling design for crop area estimation for mixed cropping when cadastral maps are available.**

**Case1: Complete enumeration approach when cadastral maps and crop register with season- wise crop area are available.**

In some countries, land records are maintained by an authorized agency. Each parcel of land has a unique identity, such as a. parcel/survey number, and the activity performed in each parcel is recorded in the crop register at regular intervals. If a parcel is used for cultivation then such information as the area of the parcel, owner of the parcel, the soil type and the crops grown are recorded. Under this situation, the area of each crop/crop mixture is available before harvest in the season. For example, in India, the area of each crop/crop mixture is recorded by the enumerator (*Patwari*) before harvest in the season in the crop register known as *Khasra Register*.

The proposed methodology may be used in situations in which the **land cadastral maps, as well as area records for each parcel are available** for different crop mixtures in each village/enumeration area. In this situation, it is necessary to obtain estimates of the crop area at the district level. As the objective is to develop estimators at the district level, the sub districts are considered as strata. Here, total number of villages (clusters) in each sub district (stratum) is known, but the total number of villages under the constituent crop in the mixture, such as the number of villages having the crop as pure stand, mixture-1, mixture-2 and so on may not be available in general. The number of selected villages within each stratum is fixed, but the number of selected villages within each stratum under the crop as pure stand, mixture-1, mixture-2 and so on may be a random quantity. These different categories under pure stand, mixture-1, mixture-2 and so on should be considered as domains. Therefore, it is proposed to estimate area under the crop and the constituent crops in the mixture using the domain estimation approach (**Särndal et al. 1992**). The domain estimates can be obtained at the district level pooling over the domains for a particular crop provides the estimate under that crop at the district level.

In this situation, the sampling design for estimation of the crop area at the district level is a stratified cluster sampling design in which a probability sample of villages (clusters) is drawn in each sub district (stratum) by a SRSWOR design. The parcels (units) within each selected village are completely enumerated for recording the crop area as available in the crop register. Relevant questionnaires for data collection are given in annex A of the field test protocol document.

**Case 2: Sample survey approach when cadastral maps are available but season-wise crop area is not available in crop register.**

There are situations in which cadastral maps are available, but parcel-wise crop area records may not be available in the crop register. In addition, the size of the village/enumeration area may be very large, for example, there may be a large numbers of parcels in the villages/enumeration area. In this situation, complete enumeration of the villages/enumeration area may be very time consuming exercise. It is, therefore, advisable to select a sample of parcels in the selected villages/enumeration area for recording crop area statistics. Under this condition, stratified a two-stage sampling design is appropriate for the sample selection.

Often, it is possible to observe a variable that is reasonably highly correlated with the study variable at a relatively low cost. /Thus, for estimation of a crop area, the variable "seed used" may be correlated with the apportioned area under the component crops. The information on "seed used" can be collected relatively inexpensively and on a larger sample, while the information on the main variable, the apportioned area under the component crops, can be collected on a smaller sample. By using this technique, which is known as two-phase/double sampling, it is possible to obtain a precise estimator of the population parameter by observing the study variable (study variable is assumed to be observed/measured expensively) on a reduced sample size by incurring a little extra expenditure on a variable that is observable for a relatively low cost on a large sample. This technique can be easily extended to two-stage/multi-stage sampling designs. This approach may be useful even for smaller villages.

Thus, in this case, under the stratified two-stage sampling design framework, a two-phase sampling approach is used for estimation of the crop area at the district level with two phases at each stage of sampling. First, a preliminary SRSWOR sample of villages is selected in each district and then, in the second stage, parcels are selected by SRSWOR for collecting auxiliary information regarding the selected parcel, such as seed used and, farmers' assessment of crop area. Again, a subsample of villages is selected from the initial sample of villages and in each of these selected villages a subsample of parcels is selected for area measurement by GPS or by the polygon method using a chain and compass. In this situation, questionnaires given in annex A of the field test protocol document are the best option to use for data collection of the different attributes. Then, the area for component crops can be obtained by apportioning using the fixed ratio approach based on physical observations.

**CATEGORY 2 (Rwanda and Jamaica): Sampling design for crop area estimation for mixed cropping when area frame is available.**

Countries where an area frame is available, a stratified two-stage cluster sampling design with two phases at each stage can be employed using the available area frame. The sub districts within the districts are the strata. First, a preliminary probability proportional to size (PPS) with a replacement sample of enumeration areas need to be selected in each sub district. Total agricultural land can be considered as the auxiliary information for the selection of the enumerations areas. Each enumeration area can be divided into a numbers of segments. In the second stage, segments need to be selected by SRSWOR design and completely enumerated with regard to, among other things, the variable seed used, family size, number of active member in the family and, self-reporting by the farmer.

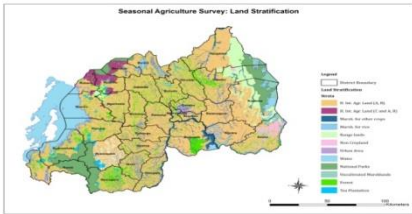
From the already selected enumeration area, a subsample of the area is selected and from the already selected segments, a subsample of the segments is selected. The parcels within the selected segments should be completely enumerated for area measurement by GPS and the rope and compass method. The parcels crops are grown in different forms of mixture, such as a pure stand, mixture-1and mixture-2. In this case, the different crop mixtures are taken as the domains of the study.

For this situation, questionnaires given in annex B of the field test protocol document should be used for data collection of the different attributes. Then, the area for component crops is obtained by apportioning using the fixed ratio approach based on physical observations.

**Sampling design for crop area estimation for mixed cropping when an area frame is available.**

A stratified two-stage cluster sampling design with two phases at each stage may be employed if an area frame is available.

Strata (subdistricts within the districts).




**PPSWR**

Total agricultural land as the size variable for PPS sampling.

Phase I: Select a larger sample.

SRSWOR

Phase II: Select Smaller Sample of enumeration areas (PSU).

**SRSWOR**

Phase I: Select a larger sample of a segments (SSU) within each selected enumeration area and complete enumeration for seed used, family size, farmers' self-reporting etc.

SRSWOR

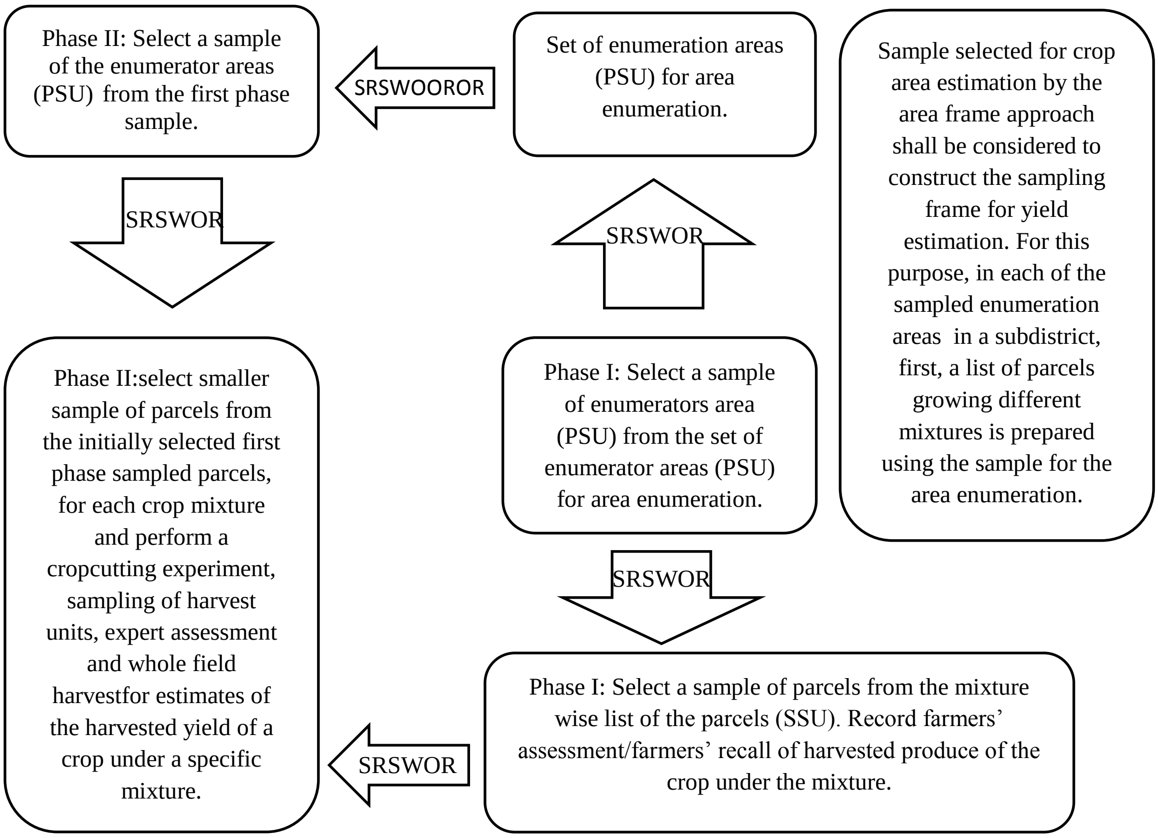
Phase II: Select Smaller Sample of Segments (SSU) within already selected enumeration areas and complete enumeration of the parcels using GPS and a rope and compass.

**CATEGORY 3: Sampling design for crop area estimation for mixed cropping using the household approach.**

Sometimes, identification of parcels in the selected villages/enumeration area may not be available because land cadastral maps and parcel-wise crop registers are not available. In this situation it is advisable to adopt the household approach. Therefore, village/enumeration area /census block may be considered as PSU and a sample of households should be selected from each selected village/enumeration area/census block. The fields belonging to a selected household should be enumerated for crop area estimation.

Sampling design in this case needs to be a stratified two-stage cluster sampling design with two phases within each stage for crop area estimation. The sub districts within the districts are to be treated as strata. First, a preliminary probability proportional to size (PPS) with replacement sample of PSUs, namely census block/village/enumeration area needs to be selected in each sub district/stratum. For this purpose, total agricultural land is to be considered as the auxiliary information for selection of PSUs. At the second stage the households are selected by SRSWOR. The parcels belonging to the selected households within the selected PSUs should be completely enumerated for recording such information as the seed used and family size of the owner of the parcel. From the already selected PSUs, a subsample of PSUs needs to be selected and from the already selected households, a subsample of households is selected. All the parcels in selected households must be completely enumerated. The areas of each of the parcels growing crop mixtures of these sampled households are measured by GPS and a compass and rope. In the selected parcels crops are grown in different forms of mixture, such as pure stand, mixture-1 andmixture-2. The area for component crops can be obtained by apportioning using the fixed ratio approach based on physical observations, seed used and eye estimate. Data collection should be made using the questionnaires given in annex C of the field test protocol document.

**Sampling design for crop yield estimation for mixed cropping when area frame is available**



## 5.3. PROPOSED SAMPLING DESIGN FOR CROPP YIELD ESTIMATION UNDER MIXED CROPPING

The sampling design and estimation procedure for average crop yield under a specific mixture as per a specified sampling design is as follows:

**CATEGORY 1: Sampling design for crop yield estimation for mixed cropping when cadastral maps are available.**

**Case 1: Cadastral maps and crop register with season-wise crop area are available.**

In these situations, there may be two ways to sample parcels in the selected villages/enumerations Areas. These are as follows:

Option1: Crop mixture-wise sampling

At the time of "estimation of crop area" under a stratified cluster sampling design framework as described in the previous section, in each sub district (stratum) from the set of villages (PSUs), a sample of villages is selected by SRSWOR and within each selected village the parcels (SSUs) are completely enumerated for area enumeration by GPS or farmers' enquiry. A crop mixture-wise sampling frame is to be prepared during area enumeration and used for sample selection for yield estimation. The sampling design under consideration for crop yield estimation needs to be stratified in a two-phase two stage sampling design.

In the first phase of sampling, from the set of villages selected for area enumeration, a large sample of villages should be selected by SRSWOR design and a sample of parcels (SSUs) should be chosen within the selected villages by SRSWOR from the total number of parcels growing a specific crop mixture in the village. From the selected parcels farmer's eye estimates of harvested yield of the crop under mixture needs to be recorded. In the second phase, a subsample of villages should be selected by SRSWOR design from the initially selected villages. Within each selected village, for each mixture, a subsample of parcels needs to be selected from the initial first phase second stage units in the selected village of the sub district.

From these parcels, estimates of the harvested yield of a crop under a specific mixture can be obtained by conducting crop-cutting experiments. The harvested produce of each component crop should be recorded separately. The crop yield is determined on the basis of the apportioned area as set by a fixed ratio of each component crop in the crop-cut plot. See section 4.2 of Technical Report I for a detailed procedure for conducting crop cutting data collection on farmers' eye estimates of crop yields and crop cutting experiment questionnaires needs to carried out using questionnaires (option 1) given in annex 1 of the field test protocol document.

Option 2: Overall sampling

In previous option, the sample at the second stage is selected from each of the second stage strata (crop mixtures). At the time of area enumeration, when the number of crop mixtures observed in villages is large, the selection of samples from each second stage stratum can be a very laborious exercise.

In this case, rather than selecting mixture by a mixture sample, an overall sample of a large number of villages is selected in the first phase from the already selected sample of villages for area enumeration. In the second stage of the first phase, a random sample of parcels is selected within the selected village for observations on auxiliary variable, such as farmers' eye estimates of harvested yield of the crop under mixture.

In the second phase, a subsample of villages are drawn from the first phase villages and a subsample of parcels are selected from the parcels selected in the first phase by SRSWOR. Estimates of the harvested yield of a crop under a specific mixture are obtained by conducting a crop cutting experiment for crop mixtures from all these parcels.

Questionnaires (option 2) given in annex A of the field test protocol document should be used for data collection on farmers' eye estimates of crop yield as well as for crop cutting experiments. See section 4.2 of Technical Report I for a detailed procedure for conducting a crop cutting experiment.

Case 2: Cadastral maps are available but season-wise crop area is not recorded in the crop register

In these situations, there may be two ways for sampling of parcels in the selected villages/enumerations areas. They are as follows:

At the time of "estimation of crop area" under a stratified two- stage two-phase sampling design framework, within each sub district (stratum) from the set of all villages (PSUs), a sample of villages and within each selected village, a sample of parcels, are to be selected in two phases for area enumeration under a specific mixture of a crop. While surveying the parcels for area enumeration, it is possible to post-stratify the sampled parcels into different mixtures. The sample observed for area enumeration is used as a sampling frame for the selection of a sample for yield estimation. Therefore, at the first phase of sampling, a sample of villages is selected and within the selected villages, mixture by mixture samples of parcels are selected by SRSWOR for collecting data on farmers' eye estimates of harvested yield of the crop under mixture. In the second phase of sampling, a subsample of villages is selected by SRSWOR design from the first phase selected villages and within each selected villages, mixture by mixture sub-sample of parcels are selected from the first phase second-stage parcels in the selected village of the sub district for data collection by crop cutting experiments. Data collection on farmers' eye estimates of crop yields and for conducting crop cutting experiments, questionnaires (option 1) should be used as given in annex A of the field test protocol document. Procedures for conducting crop cutting experiments are given in section 4.2 of Technical Report I.

Option 2: Overall sampling

In the previous option, the sample at the second stage was selected from each of the second stage strata, namely crop mixtures. When the number of crop mixtures observed in villages is large, the selection of a sample from each second stage stratum may be a very laborious. In this case, rather than selecting a mixture-by-mixture sample, an overall sample of a large number of villages is selected in the first phase from the already-selected sample of villages for area enumeration. In the second stage, an overall sample of parcels are drawn for observations on auxiliary information, such as farmers' eye estimates of harvested yield of the crop under the mixture scenario, are taken from these parcels. In the second phase of sampling, a subsample of villages is drawn from first phase villages and a subsample of parcels is selected from the second-phase parcels by SRSWOR. From these parcels, estimates of the harvested yield of a crop mixture are obtained by conducting a crop cutting experiment for mixed cropping. See section 4.2 of Technical Report I for a detailed procedure for conducting a crop cutting experiment. Data collection on crop cutting experiments and farmers' eye estimates of crop yields should be carried out using questionnaires (option 2) given in annex A of the field test protocol document.

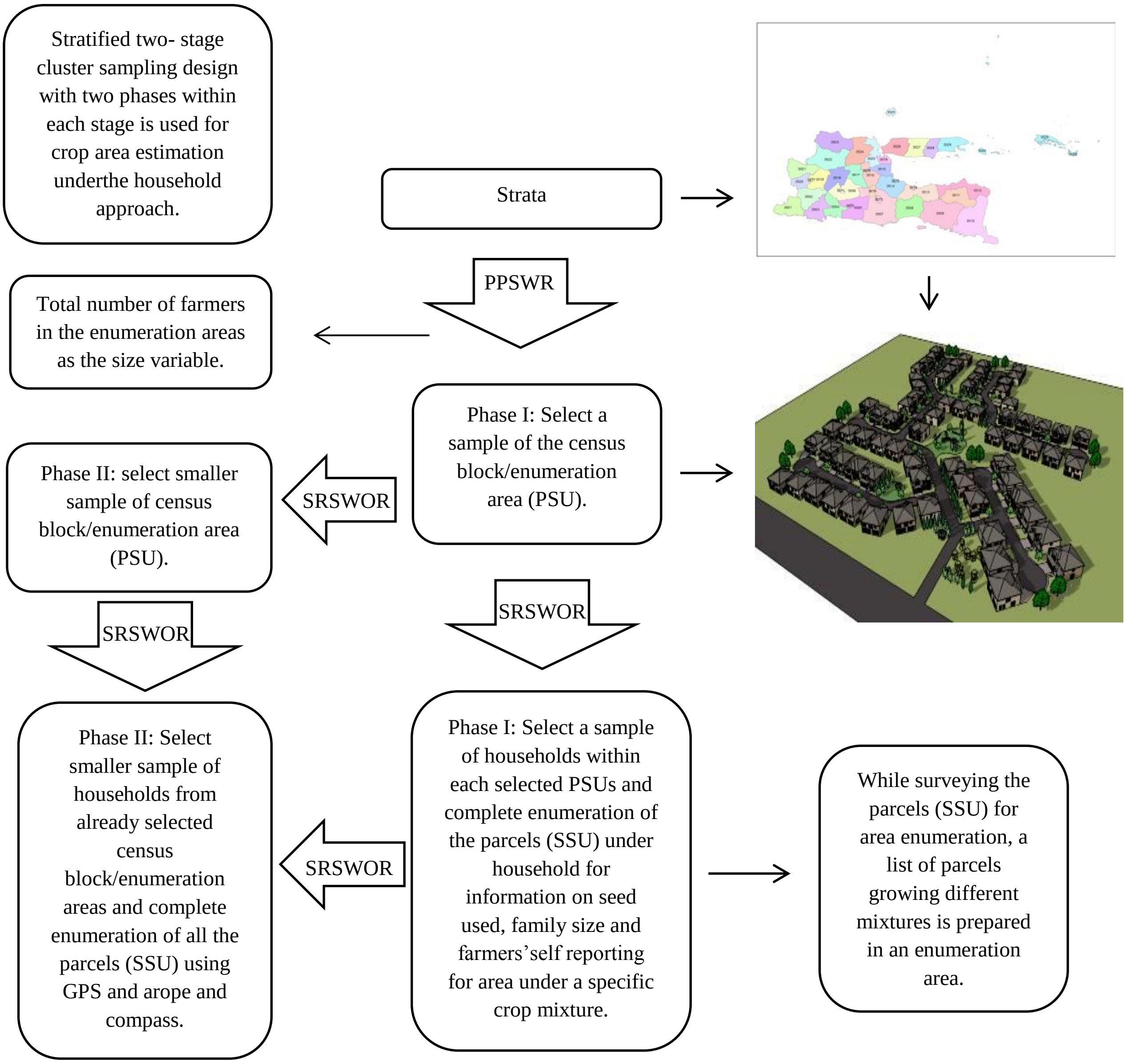
**CATEGORY-2: Sampling design for crop yield estimation for mixed cropping using the area frame approach.**

In situations in which area frames are available, using the area frame approach for crop yield estimation is advisable, wherein the sample selected should be considered to construct the sampling frame for yield estimation. For this purpose, in each of the sampled enumeration areas in a sub district, a list of parcels growing different mixtures is prepared using the sample for area enumeration. It should be noted that that segment- wise listing is not necessary in this case.

In the first phase of sampling, from the set of enumeration areas (PSU) for area enumeration, a sample of enumeration areas are selected by SRSWOR design. Within the selected enumeration areas, for each of the crop mixtures, a sample of parcels (SSUs) is selected by SRSWOR design from the set of parcels used for recording crop area. From the selected parcels, farmers' assessment/farmers' recall of harvested produce of the crop under the mixture has to be recorded.

In the second phase of sampling, a sample of enumeration areas is selected by SRSWOR design from the first phase of enumeration areas and within each selected enumeration area, for each crop mixture, a subsample of parcels is selected from initially selected first phase sampled parcels in each sub district. From the parcels in the final sample, estimates of the harvested yield of a crop under a specific mixture are obtained by conducting a crop cutting experiment for mixed cropping. See section 4.2 of Technical Report I for detailed procedures for conducting a crop cutting experiment. Data collection on farmers' assessment/farmers' recall of harvested produce should be carried out using questionnaire-5(C-2) in the Annex of the field test protocol document and for data collection on crop yield, crop-cutting experiment questionnaires given in the same annex should be used.

**Sampling design for crop estimation for mixed cropping using the household approach**



**CATEGORY-3: Sampling design for crop yield estimation for mixed cropping using the household approach**

In situations in which information of parcels in the selected villages/enumeration areas/census blocks may not be available in the records, it is advisable to adopt the household approach for crop yield estimation. Under this approach, the sample selected for crop area estimation by should be taken as the sampling frame. For this purpose, a list of parcels growing different mixtures is prepared using the results from the area enumeration in each of the sampled villages/enumeration areas/census blocks in a sub district.

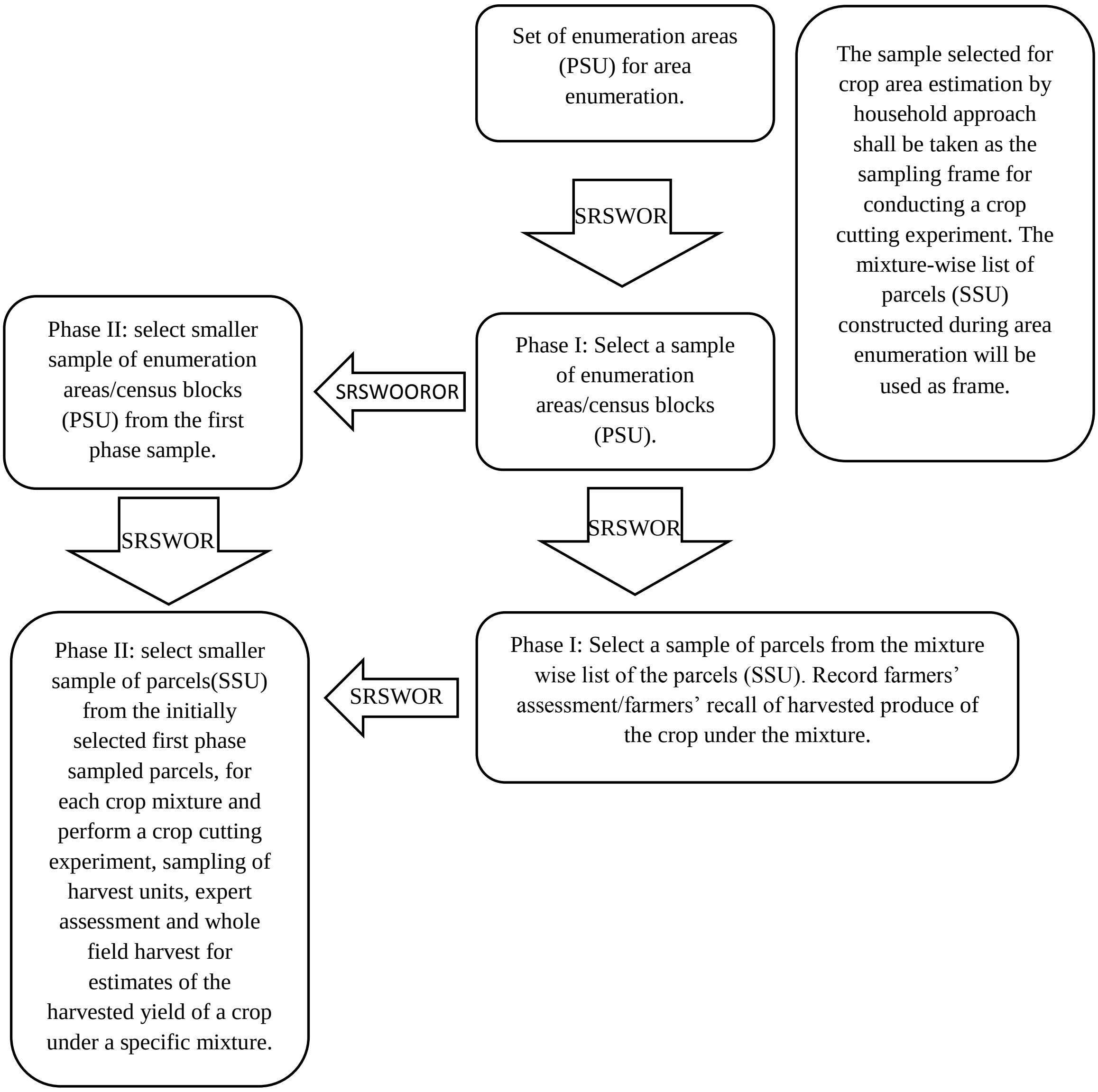
Suppose, for "estimation of crop area" by the household approach under a stratified two-phase two-stage cluster sampling design framework, within each sub district (stratum) from the set of total villages/enumeration areas/census blocks (PSUs), a sample of villages was selected. In the second stage, a subsample of households was selected in each selected village and the parcels of the selected households were completely enumerated for an area under a specific crop mixture. While surveying the parcels for area enumeration, a list of parcels growing different mixtures is prepared in a selected village. It should be noted that in the above-mentioned situation, there is no need for household-wise listing of parcels.

In the first phase of sampling, from the set of villages/enumeration areas/census board (PSU) for area enumeration, a sample of villages is elected by SRSWOR. Within the selected villages, for each of the mixtures, a sample of parcels (SSUs) is selected by SRSWOR from the parcels used for recording the area. From the selected farmers' assessment/ farmers' recall of harvested produce of the crop under the mixture has to be recorded.

In the second phase of sampling, a sample of villages/enumeration areas/census boards is selected by SRSWOR design from the initially selected -phase sampled villages/enumeration areas/census boards and within each selected villages, for each mixture, a sample of parcels is selected from the initially selected first phase sampled parcels in each sub district. From those sampled parcels, estimates of the harvested yield of a crop under a specific mixture are obtained by conducting a crop cutting experiment for mixed cropping. See section 4.2 of Technical Report I for detailed procedures for conducting a crop cutting experiment.

Data collection on farmers' assessment/farmers' recall of harvested produce should be carried out using questionnaire-6(C-3) in annex C of the field test protocol document and for data collection on crop yield, crop cutting experiment questionnaires given in the same annex be used.

## SAMPLING DESIGN FOR CROP YIELD ESTIMATION FOR MIXED CROPPING USING HOUSEHOLD APPROACH



## 5.4. APPLICATION OF SMALL AREA APPROACH FOR CROP AREA AND YIELD ESTIMATION IN MIXED CROPPING

As the domain estimation approach is being proposed for crop area and crop yield estimation, the sample sizes for some of the domains may be very small or in the worst case, there may not be any sample data point for some of the domains. This may particularly be the case when there is large number of crop mixtures. In such situations, the small area estimation approach should be used for developing estimators for domains with a very small sample size or no sample data point. Depending on the level of availability of auxiliary data, an area level model or unit level model should be used for small area estimation. See section 7 of Technical Report 1.

## 5.5. CROP AREA AND YIELD ESTIMATION UNDER CONTINUOUS CROPPING

In some cases, the crop is sown and planted successively as and when the field is ready for sowing/planting, such as a wet condition of field/receding water, the new plants are intercropped with the older ones or planting of new plants after uprooting plants that have already been harvested. As in case of mixed cropping, multiple field visits may be required to realistically capture the crop area and yield in the context of continuous cropping.

**Area estimation:**

1. One crop is harvested and another is planted: The crop area may be recorded once in each crop growing season before harvesting. In each season, if a crop is grown as pure, the entire area may be recorded under the crop. Under a crop mixture scenario, the crop area may be apportioned on the basis of already-available methods, such as physical observation and seed used.

2. When a new crop is sown between a standing crop nearing maturity, the crop area may be recorded in the season in which the crop is harvested. When the crops are sown under a mixed approach, the area of each component crop may be worked out through apportioning.

3. For cases in which the crop area under a field is gradually enlarged, a multi-round survey is most appropriate for realistically capturing the area of each crop. The frequency of each round is determined on the basis of the vegetative cycle of the crop (less than three months, preferably two months). In this situation, the area to be harvested before

the next visit may be recorded in each round. The apportioning of the area may be done as in the case of mixed cropping.

For pure and mixed crops, the crop area may be estimated using the domain estimation approach under a stratified cluster sampling design when using the cadastral map approach and a stratified two-phase two-stage sampling design when using an area frame approach or a household approach. The crop area may be determined using a compass and rope and GPS as the main variable and self-reporting by farmer, family size and number of active family members as the auxiliary variable.

**Yield estimation:**

In the absence of an enumerator, the farmer should record harvested produce in a separate pro forma provided by the enumerator or by using the method of crop diary/pro forma along with telephone @ two telephone calls per week as follow-up. Another alternative recording the produce is for the enumerator to count the total plant population in a small plot (separate for each crop in a mixture situation) and record the produce harvested plant-wise on the day of his/her visit. On the basis of produce obtained from a sufficient number of plants from the plot, the total produce of the plot may be worked out.

For both pure and mixed crops, the crop yield may be estimated using the domain estimation approach under a stratified two-phase two-stage sampling design in the case of the cadastral map approach and a stratified two-phase two-stage sampling design when using the area frame and household approach. The crop yield is determined through a crop cutting experiment, expert assessment, sampling of the harvest unit, method of produce harvested plant-wise and crop diary pro forma along with telephone calls as the main variable and farmers' prediction and recall of the harvested produce as the auxiliary variable. For the purpose of accuracy assessment of proposed methods, the method of whole field harvest shall be used on a smaller sample.

## 5.6. CRITERION FOR SAMPLE SIZE

The sample size needs to be suitably fixed by examining the degree of variability in the data. The guiding principle in determining the sample size is by fixing the total cost of the survey for a given level of precision or maximizing the precision of the estimator for a given cost. Other important criteria in determining the sample size are fixing of the coefficient of variation of the estimator or using the known value of the design effect. The criterion of standard error should be taken into account for examining the variability. More specifically, the estimated percentage standard errors at the district level should be worked out. The available guidelines with respect agricultural surveys suggest that at the district level, the percentage standard errors should be between 5and 10percent. Accordingly, the sample size at the district level should be fixed so that the percentage standard errors at that level do not exceed 10 percent.

## 5.7. CONTROL OF NON – SAMPLING ERRORS

The data quality aspect is an important consideration in all surveys, including agricultural surveys. Data quality may be seriously affected if the crop area and crop yield surveys are not carried out properly. Non-sampling errors mainly are the result of numerous factors and can occur at every stage of the survey, including survey design, field work, tabulation or analysis of data. Sometimes, the non-sampling errors are more significant than the sampling errors and thus may affect the results substantially. There are different types of non-sampling errors such as non-response errors, measurement errors and data processing errors.

Non-response errors occur at all stages of the survey, such as the survey design, planning and execution. However, most of the non-response errors are mainly because of: (i) not-at-home, the respondent may not be at home when the enumerators call on them, in such situations frequent visits to the farmer may reduce the level of non-response; and, (ii) refusal, the respondent may refuse to provide information to the enumerators. The reliability of estimates depends to a great extent on the response level when conducting crop cutting experiments. Sometimes the selected farmer may not be very cooperative, which may lead to non-response. Another form of non-response is that from time to time, the farmer informs the enumerator of a particular date of harvest but when the enumerator reaches the field, the crop is already harvested. Better rapport between the enumerator and the farmer may reduce the level of non-response.

Measurement errors or response errors arise in data collection or while taking observations and are mainly contributed by the respondent or the enumerator or both. Faulty instruments may also lead to measurement errors. Thus, if the weights and balances used for recording crop produce are defective, there could be an erroneous recording of crop produce. The findings of crop estimation surveys in India reveal the following sources of error while conducting crop-cutting experiments:

i) Error in selection of survey/sub survey numbers;,
ii) Error in selection of field within the survey/subsurvey number;
iii) Error in the measurement of the field;
iv) Error in selection of random numbers, location and marking of plots;

v) Error in weighing of produce;
vi) Error in recording ancillary information;
vii) Inadequate arrangements for storing produce for driage and incorrect reporting of constituents in mixture.

Such errors may seriously affect the data quality. The enumerator has the tendency to harvest plants which are on the border and thus may lead to an over estimation of the production.

Response errors refer to the differences between the individual true value and the corresponding observed sampling value, irrespective of the reasons for discrepancies. The errors occurring during data entry and data processing are also important and need to be controlled.

To control non-sampling errors, the developed questionnaires should not be lengthy. The questionnaires should be simple and easy to understand. In designing the questionnaire and schedule, special care needs to be taken to include certain items of the information that serve as a check on the quality of the data to be collected. These additional items of information may be canvassed on a small sample. For example, inclusion of items leading to stable ratios, such as the sex ratio, may be useful in assessing the quality of census and survey data. The enumerators should be imparted proper training and supervision of data collection should be done at regular intervals. The questionnaires should be designed in such a way that cross checking of data can be done in the filled-in questionnaires. The validation checks should also be in place while entering the data.

One way to assess and control non-sampling errors in censuses and surveys is to independently duplicate the work at the different stages of operation with a view to facilitating the detection and rectification of errors.

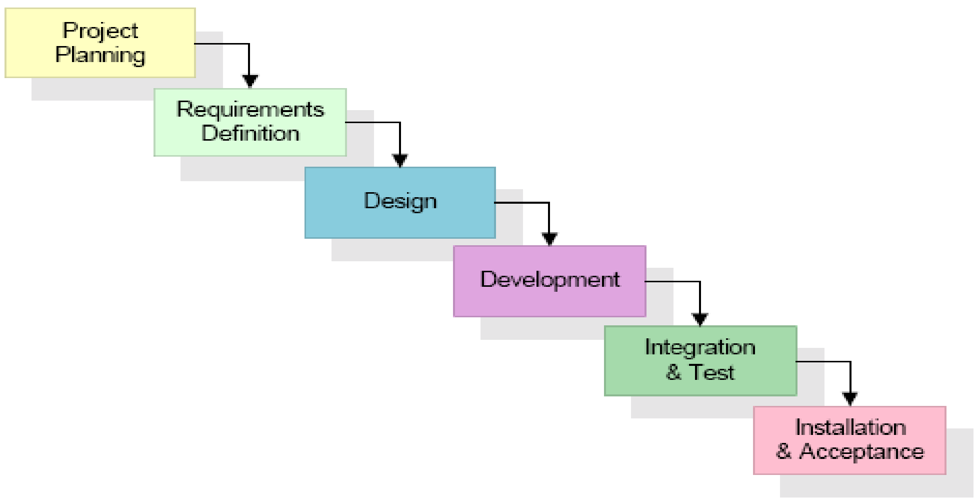
An important type of sample check, commonly used to assess non-sampling errors, consists in selecting a small sample of units covered in the census or survey and in resurveying it while using better trained and more experienced survey staff.

Another method, for assessing non-sampling errors, especially in census work, is to take a sample of relevant units from a different source, if available, and check whether the units have been enumerated in the main investigation and whether there are discrepancies between the values when matched.

# 6

# Software for data entry and data analysis

Under this study, software for data entry as well as analysis of data is being developed using the Software Development Life Cycle (SDLC). Small to medium database software projects are generally broken down into six stages:




The relationship of each stage to the others can be roughly described as a waterfall, in which the outputs from a specific stage serve as the initial inputs for the following stage.

To follow the waterfall model, it is necessary to proceed from one phase to the next in a purely sequential manner. For example:

- After completing the “project planning” phase, the "requirements definitions" phase must be completed.

- When and only when the requirements are fully completed, the design can be completed. The design should be a plan for implementing the requirements given.
- When and only when the design is fully completed, implementation of that design is made by coders. Towards the later stages of this implementation phase, disparate software components produced by different teams are integrated.
- After the implementation and integration phases are complete, the software product is tested and debugged; any faults introduced in earlier phases are removed here.
- Then, the software product is installed, and later maintained to introduce new functionality and remove bugs.

Thus, under the waterfall model, moving to a phase should only occur when the proceeding phase is completed and perfected. Phases of development in the waterfall model are thus discrete, and there is no jumping back and forth or overlap between them.

The central idea behind the waterfall model time spent in the early part of development to ensure that requirements and design are absolutely correct is very useful in economic terms (it will save much time and effort later). It is also necessary to ensure that each phase is 100 percent complete and absolutely correct before proceeding to the next phase of program creation.

Many believe that the waterfall model in general are well suited to software projects that are stable (especially with unchanging requirements) and when possible and likely that designers are able to fully predict problem areas of the system and produce a correct design before implementation is started. The waterfall model also requires that implementers follow the well-made, complete design accurately to ensure that the integration of the system proceeds smoothly.

The waterfall model, however, is argued by many to be a bad idea in practice, mainly based on the view that it is impossible to get one phase of a software product's lifecycle "perfected" before moving on to the next phases and learning from them (or at least, the belief that this is impossible for any non-trivial program). These models may address some or all of the criticism of the "pure" waterfall model. Once the primary data is collected using designed questionnaire for the study, the data entry should be done by the field testing countries using data entry software. The developed data entry software should take care of validation of data and further data analysis of the entered data shall be done using data analysis software.

# 7

# Conclusion

In this report, an assessment of various gaps that exist in the context of crop area and crop yield estimation in general and in crop area and yield estimation in the mixed/ intercropping scenario, in particular has been made. Some of the notable gaps that exist in the estimation of crop area and crop yield in this scenario are: a) choice of an appropriate sampling frame or development of workable sampling frame for the choice of a suitable sampling design for selection of a representative sample; b) development of an appropriate estimation procedure; c) determination of optimum sample size; and d) development of safeguards for the control of various types of non-sampling errors. These are issues of consideration.

On the basis of gap analysis, an attempt has been made to find a solution to the problems of estimation of crop area and crop yield in the context of mixed and continuous cropping. Accordingly, appropriate methodology has been developed for the estimation of crop area and crop yield in the context of mixed/intercropping. The domain estimation approach has been proposed for the estimation of crop area of components crops in the crop mixtures. The various crop mixtures are considered as domains. Generally, measurement-based methods are recommended for determining the crop area and crop yield. These methods, though capable of giving accurate figures, are time-consuming, are costlier and require adequate trained manpower. Therefore, the sample survey approach for estimation of crop area and crop yield is most appropriate.

Choice of an appropriate sample size is very critical in agricultural surveys involving measurements. The sample size in any survey can be reduced considerably by making use of available relevant auxiliary information in the statistical system. When the relevant auxiliary information is not readily available, the double or two-phase sampling approach can be used to generate relevant auxiliary information by incurring a nominal cost. The double sampling approach has been used extensively to generate suitable auxiliary information for the development of appropriate estimators of crop area and crop yield so that sample sizes are kept at a bare minimum. The advantages of the proposed approach are twofold. First there is a reduction in cost of a survey by keeping the sample sizes under check using auxiliary information. Second, the non-sampling errors can also be minimized through proper training and supervision of field work. An added benefit of using the proposed approach is that the data processing and analysis can be performed with relative ease.

An important consideration in estimation involving crop mixtures is apportioning of crop area to the component crops in the crop mixture. While standard procedures for apportioning of crop areas in mixture to component crops, such as use of seed rate, plant density or number of rows, can be used for field crops, no such procedures are readily available for apportioning in case of mixtures involving annual and seasonal crops, annual and annual crops or for that matter in the case of mixtures of annual and perennial or perennial and perennial crops. Some methods of apportioning in such cases are given in the report.

The objective methods for determination of crop area and crop yield are being proposed. The use of GPS is recommended when the plot sizes are not available. In view of limitation of GPS instruments in providing accurate estimates for small plot sizes and for hilly areas, correction factors for rectification of crop area figures those areas should be developed. The correction factors can be generated by determining plot areas in respect of 20-25 small plots using GPS and the same number of plots in hilly areas through an objective method, such as the polygon method/rectangulation/triangulation method. The difference in the two figures can be used for generating a correction factor. Correction factors should be generated separately for plains and hilly regions. The exercise of updating plot area figures through use of GPS instruments can be carried out after an interval of five years or even longer in countries where plot sizes change infrequently. It is well known that the harvested area is generally less than the area planted as crops may be damaged because of, among other things, flood, drought and disease or pest attack. To the extent possible, the crop area data should be collected not more than two months before harvest.

The crop cutting experiment technique is proposed for crop yield estimation. This technique is tedious and time consuming, but, if carried out properly, has the potential to objectively capture the crop yield data. To significantly reduce the sample size for crop cutting experiments, crop cutting experiment data and the farmers' appraisal based crop production data are combined. The homestead data related to crop area and crop yield are to be collected by inquiry from the members of households. The double sampling approach is being proposed for capturing yield data of crops involving multiple pickings. The developed methodology can capture the area and yield figures from stray trees.

Crop area and crop yield data in the context of continuous cropping, especially when cultivation is gradually enlarged, is proposed to be captured through multiple rounds. However, the same sample would be retained for data capture in various rounds. The optimum number of rounds should be suggested after the experience gained in the field survey.

The data collection is sought to be carried out through the use of computer-assisted personal interviewing software. User- friendly software should be developed for data entry and the analysis of data.

## ANNEX A1

**Category -1 - When cadastral maps are available.**

**Case1: When season-wise crop area is available in the crop register.**

**Estimation procedure for crop area estimation for mixed cropping using the complete enumeration approach.**

Our objective is to obtain estimates of the crop area at the district level. Let a finite population $U=\{1,\ldots,j,\ldots, N\}$ be divided into $H$ strata $U_1,\ldots, U_h,\ldots,U_H$ such that $h^{th}$ stratum consists of $N_h$ clusters (village). Let there be $M_{hi}$ parcels in the $i^{th}$ cluster (village), $i= 1,\ldots,N_h$. In this situation, we assume that there are $D$ domains $U_{h1},\ldots,U_{hd},\ldots,U_{hD}(d=1,2,\ldots,D)$ (domains in this case are different crop mixtures). Let $N_{hd}$ clusters among $N_h$ clusters contain units belonging to $d^{th}$ domain. In addition, let $M_{hid}$ units out of $M_{hi}$ units of $i^{th}$ cluster fall in the $d^{th}$ domain.

The partitioning equation for domains are given by

$$U_h = \bigcup_{d=1}^{D} U_{hd} \; ;h=1,2,\ldots,H. \qquad ...(1)$$

Total population size is

$$N = \sum_{h=1}^{H}\sum_{i=1}^{N_h} M_{hi} = \sum_{d=1}^{D}\sum_{h=1}^{H}\sum_{i=1}^{N_{hd}} M_{hid} \; .$$

Let $y_{hij}$ be the crop area of $j^{th}$ sample parcel (unit) in the $i^{th}$ village (cluster) in the $h^{th}$subdistrict (stratum).

The total crop area for the $d^{th}$ mixture (domain) is given as

$$Y_d = \sum_{h=1}^{H}\sum_{i=1}^{N_{hd}}\sum_{j=1}^{M_{hid}} y_{hij} = \sum_{h=1}^{H}\sum_{i=1}^{N_{hd}} Y_{hid}, \qquad ...(2)$$

Where $Y_{hid} = \sum_{j=1}^{M_{hid}} y_{hij}$ =total of the units within $i^{th}$ cluster belonging to $d^{th}$ domain in the $h^{th}$ stratum.

The total crop area in a district based on the domains is given as

$$Y = \sum_{d=1}^{D}\sum_{h=1}^{H}\sum_{i=1}^{N_{hd}}\sum_{j=1}^{M_{hid}} y_{hij} = \sum_{d=1}^{D}\sum_{h=1}^{H}\sum_{i=1}^{N_{hd}} Y_{hid} \; . \qquad ...(3)$$

The sampling design for estimation of a crop area at the district level is a stratified cluster sampling design. Let a probability sample $s_h$ of size $n_{hA}$ clusters be drawn from the $h^{th}$ stratum, $U_h$, of $N_h$ clusters by SRSWOR design; also, the units within each selected cluster are completely enumerated for recording the crop area as available in the crop register. Relevant questionnaires for data collection are given in subsequent sections.

Let $n_{hAd}$ clusters, out of $n_{hA}$ selected clusters, contain units belonging to the $d^{th}$ domain and $M_{hid}$ units of $i^{th}$ cluster fall in the $d^{th}$ domain. In this situation, $M_{hid}$ could be very small if the number of domains is large. For such cases, a large sample of clusters may be required to be observed from each of the stratum. To develop an estimator of overall domain total, we consider the case of domain size assumed unknown which is the most common case in domain estimation (Särndal et al., 1992).

In this situation, the objective is to estimate the total crop area ($Y$), as well, under different forms of crop mixtures, namely domains ($Y_d$), $d$=1,2,...,$D$. The proposed estimator of the total area under the $d^{th}$ mixture (domain) in a district is given by

$$\hat{Y}_d = \sum_{h=1}^{H} \frac{N_h}{n_{hA}} \sum_{i=1}^{n_{hAd}} Y_{hid} . \qquad ...(4)$$

The approximate variance of the estimator $\hat{Y}_d$ is given by

$$V\left(\hat{Y}_d\right) = \sum_{h=1}^{H} N_h^2 \left(\frac{1-f_h}{n_{hA}}\right)\left(P_{hd} S_{byhd}^2 + P_{hd} Q_{hd} \bar{Y}_{hd}^2\right), \qquad ...(5)$$

where

$$f_h = \frac{n_{hA}}{N_h},\ P_{hd} = \frac{N_{hd}}{N_h},\ Q_{hd} = 1 - P_{hd},\ \bar{Y}_{hd} = \frac{1}{N_{hd}} \sum_{i=1}^{N_{hd}} Y_{hid},\ S_{byhd}^2 = \frac{1}{N_{hd}-1} \sum_{i=1}^{N_{hd}} \left(Y_{hid} - \bar{Y}_{hd}\right)^2 .$$

An approximate estimator of the variance of $\hat{Y}_d$ is given by

$$\hat{V}\left(\hat{Y}_d\right) = \sum_{h=1}^{H} N_h^2 \left(\frac{1-f_h}{n_{hA}}\right)\left(p_{hd} s_{byhd}^2 + p_{hd} q_{hd} \bar{y}_{hd}^2\right), \qquad ...(6)$$

where

$$p_{hd} = \frac{n_{hd}}{n_{hA}},\ q_{hd} = 1 - p_{hd},\ \bar{y}_{hd} = \frac{1}{n_{hAd}} \sum_{i=1}^{n_{hAd}} Y_{hid},\ s_{byhd}^2 = \frac{1}{n_{hAd}-1} \sum_{i=1}^{n_{hAd}} \left(Y_{hid} - \bar{y}_{hd}\right)^2 .$$

Then, the estimator of percentage standard error of the proposed estimator of the total crop area under dth mixture, $Y_d$, is given by

$$\% \widehat{SE}(\hat{Y}_d) = \frac{\sqrt{\hat{V}\left(\hat{Y}_d\right)}}{\hat{Y}_d} \times 100. \qquad ...(7)$$

The estimator of the total crop area based on the domains at the district level for a specific crop $c$is given by

$$\hat{Y}_{c1} = \sum_{d=1}^{D*} \sum_{h=1}^{H} \frac{N_h}{n_{hA}} \sum_{i=1}^{n_{hd}} Y_{hid} \ . \qquad ...(8)$$

where the first sum is over those domains containing a particular crop $c$,$d$=1,2,...,$D*$.

An approximate estimator of the variance of $\hat{Y}_{c1}$ is given by

$$\hat{V}\left(\hat{Y}_{c1}\right) = \sum_{d=1}^{D*} \sum_{h=1}^{H} N_h^2 \left(\frac{1-f_h}{n_{hA}}\right)\left(p_{hd} s_{byhd}^2 + p_{hd} q_{hd} \bar{y}_{hd}^2\right). \qquad ...(9)$$

Then, the estimator of percentage standard error of the proposed estimator of population total, $\hat{Y}_{c1}$, is given by

$$\% \widehat{SE}(\hat{Y}_{c1}) = \frac{\sqrt{\hat{V}\left(\hat{Y}_{c1}\right)}}{\hat{Y}_{c1}} \times 100. \qquad ...(10)$$

**Estimation of crop yield for mixed cropping using a stratified two-phase two-stage sampling design framework.**

**Option1: Crop mixture-wise sampling.**

At the time of "estimation of crop area" under a stratified cluster sampling design framework under category 1, in the $h^{th}$ sub district (stratum) from the set of $N_h$ villages (PSUs),$n_{hA}$ villages were selected by SRSWOR and within each selected village, the $M_{hi}$ parcels (SSUs) were completely enumerated for area enumeration by GPS or farmers' enquiry. Suppose, while surveying the parcels for area enumeration in the $i^{th}$ village of the $h^{th}$ sub districts, $D$ different forms of mixture of the specific crop had been observed in the total $M_{hi}$ parcels. Let $D$

different mixtures be followed in $\{M_{h1}, M_{h2}, ..., M_{hd}, ..., M_{hD}\}$ parcels of the $i^{th}$ village of the $h^{th}$ sub district.

For the purpose of average yield estimation of the different forms of mixture, as well as the overall yield, the following sampling design needs to be followed:

Here, the different forms of crop mixtures are considered as different post strata obtained during area enumerations. Now, from the set of $n_{hA}$ villages (PSU) $n_h^{'}$ villages are selected by SRSWOR design and a sample of $m_{hid}^{'}$ parcels (SSUs) are selected by SRSWOR from the $M_{hid}$ parcels growing $d^{th}$ mixture, $\forall\ d = 1, 2, ..., D$. From the selected parcels, farmer's eye estimates of the harvested yield of the crop under the mixture scenario need to be recorded.

Let $x_{hijd}$ define the eye-estimated harvested yield of the crop in the $j^{th}$ parcel growing the $d^{th}$ mixture in the $i^{th}$ village of the $h^{th}$ sub district.

In the second phase, let a sample of $n_h$ villages be selected by SRSWOR design from the $n_h^{'}$ selected villages and within each selected village, for each mixture $d = 1, 2, ....D$, a sample of $m_{hid}$ parcels are selected from the $m_{hid}^{'}$ first-phase second stage-units in the $i^{th}$ village of the $h^{th}$ sub district.

It can be observed that, out of $n_h$ sample villages, $n_{hd}$ villages fall under the $d^{th}$ mixture, $\forall d = 1, 2, ....D$. Now, from all those $m_{hid}$ parcels, estimates of the harvested yield of a crop under a specific mixture are obtained by conducting crop cutting experiments. The harvested produce of each component crop is recorded separately. The crop yield is determined on the basis of apportioned area as determined by a fixed ratio of each component crop in the crop- cut plot. See section 4.2 of Technical Report I for detailed procedures for conducting a crop cutting experiment. Data collection on farmers' eye estimates of crop yields should be carried out using questionnaires (option 1) and for conducting a crop cutting experiment questionnaires given in at the end of annex A of the field test protocol document should be used.

Now, let $y_{hijd}$ define the harvested yield of the crop under a specific $d^{th}$ mixture from the $j^{th}$ parcel of the $i^{th}$ village of the $h^{th}$ sub district.

From the harvested yield of the sampled parcels, an estimate of the average crop yield under a specific crop mixture for a district can be given by

$$\bar{y}_d = \frac{1}{H}\sum_{h=1}^{H}\frac{1}{n_h}\sum_{i=1}^{n_{hd}}\frac{1}{m_{hid}}\sum_{j=1}^{m_{hid}} y_{hijd}. \quad ...(11)$$

Using the information from the eye-estimated yields under a specific mixture (domain) and using a double sampling approach, a linear regression estimate can be formed as

$$\bar{y}_{1lrd} = \bar{y}_d + b_{1d}(\bar{x}'_d - \bar{x}_d), \qquad ...(12)$$

where $\bar{x}_d = \frac{1}{H}\sum_{h=1}^{H}\frac{1}{n_h}\sum_{i=1}^{n_{hd}}\frac{1}{m_{hid}}\sum_{j=1}^{m_{hid}} x_{hijd}$ and $\bar{x}'_d = \frac{1}{H}\sum_{h=1}^{H}\frac{1}{n'_h}\sum_{i=1}^{n'_{hd}}\frac{1}{m'_{hid}}\sum_{j=1}^{m'_{hid}} x_{hijd}$.

By minimizing the variance of $\bar{y}_{lrd}$ with respect to $b_{1d}$ and ignoring the variation in $b_{1d}$, the value of $b_{1d}$ is given by

$$b_{1d} = \frac{c_{1xyd}}{c_{1xxd}}, \qquad ...(13)$$

where,

$$c_{1xyd} = \frac{1}{H^2}\sum_{h=1}^{H}\left[\left(\frac{1}{n_h} - \frac{1}{n'_h}\right)(p_{hd}s^2_{bxyhd} + p_{hd}q_{hd}\bar{x}_{hd}\bar{y}_{hd}) + \frac{1}{n_h n'_h}\sum_{i}^{n_{hd}}\left(\frac{1}{m_{hid}} - \frac{1}{m'_{hid}}\right)s_{xyhid}\right],$$

$$c_{1xxd} = \frac{1}{H^2}\sum_{h=1}^{H}\left[\left(\frac{1}{n_h} - \frac{1}{n'_h}\right)(p_{hd}s^2_{bxhd} + p_{hd}q_{hd}\bar{x}^2_{hd}) + \frac{1}{n_h n'_h}\sum_{i}^{n_{hd}}\left(\frac{1}{m_{hid}} - \frac{1}{m'_{hid}}\right)s^2_{xhid}\right],$$

$$p_{hd} = \frac{n_{hd}}{n_h}, \quad q_{hd} = 1 - p_{hd},$$

$$s^2_{bxhd} = \frac{1}{n_{hd}-1}\sum_{i=1}^{n_{hd}}(\bar{x}_{hid} - \bar{x}_{hd})^2, \; s_{bxyhd} = \frac{1}{n_{hd}-1}\sum_{i=1}^{n_{hd}}(\bar{x}_{hid} - \bar{x}_{hd})(\bar{y}_{hid} - \bar{y}_{hd}),$$

$$s^2_{xhid} = \frac{1}{m_{hid}-1}\sum_{j=1}^{m_{hid}}\left(x_{hijd} - \bar{x}_{hid}\right)^2,$$

$$s_{xyhid} = \frac{1}{m_{hid}-1}\sum_{j=1}^{m_{hid}}\left(x_{hijd} - \bar{x}_{hid}\right)\left(y_{hijd} - \bar{y}_{hid}\right),$$

$$\bar{x}_{hid} = \frac{1}{m_{hid}}\sum_{j}^{m_{hid}} x_{hijd}, \quad \bar{y}_{hid} = \frac{1}{m_{hid}}\sum_{j}^{m_{hid}} y_{hijd}, \; \bar{x}_{hd} = \frac{1}{n_{hd}}\sum_{i=}^{n_{hd}}\bar{x}_{hid},$$

$$\bar{y}_{hd} = \frac{1}{n_{hd}}\sum_{i=1}^{n_{hd}}\bar{y}_{hid}.$$

An approximate estimate of variance of the linear regression estimator $\bar{y}_{1lrd}$ is given by

$$\hat{V}\left(\bar{y}_{1lrd}\right)=\frac{1}{H^2}\sum_{h=1}^{H}\left[\left\{\left(\frac{1}{n_h}-\frac{1}{N_h}\right)\left(p_{hd}s_{byhd}^2+p_{hd}q_{hd}\bar{y}_{hd}^2\right)+\frac{1}{n_hN_h}\sum_{i=1}^{n_{hd}}\left(\frac{1}{m_{hid}}-\frac{1}{M_{hid}}\right)s_{yhid}^2\right\}\right. \quad \ldots(14)$$
$$\left.-r_{1d}^2\left\{\left(\frac{1}{n_h}-\frac{1}{n_h^{'}}\right)\left(p_{hd}s_{byhd}^2+p_{hd}q_{hd}\bar{y}_{hd}^2\right)+\frac{1}{n_hn_h^{'}}\sum_{i=1}^{n_{hd}}\left(\frac{1}{m_{hid}}-\frac{1}{m_{hid}^{'}}\right)s_{yhid}^2\right\}\right],$$

where

$$s_{byhd}^2=\frac{1}{n_{hd}-1}\sum_{i=1}^{n_{hd}}\left(\bar{y}_{hid}-\bar{y}_{hd}\right)^2,\ s_{yhid}^2=\frac{1}{m_{hid}-1}\sum_{j=1}^{m_{hid}}\left(y_{hijd}-\bar{y}_{hid}\right)^2,$$

$$r_{1d}^2=\frac{c_{1xyd}^2}{c_{1xxd}\ c_{1yyd}},$$

and $c_{1yyd}$ is in same functional form as $c_{1xxd}$ as defined earlier.

Then, the estimator of percentage standard error of the proposed linear regression estimator of average yield under of a crop under $d^{\text{th}}$ mixture, $\bar{y}_{1lrd}$, is given by

$$\%\,\widehat{SE}(\bar{y}_{1lrd})=\frac{\sqrt{\hat{V}\left(\bar{y}_{1lrd}\right)}}{\bar{y}_{1lrd}}\times 100. \quad \text{...(15)}$$

The estimator of yield for a specific crop *c* at the district level is given by

$$\hat{\bar{Y}}_{c1}=\frac{1}{D*}\sum_{d=1}^{D*}\bar{y}_{1lrd}\ . \quad \text{...(16)}$$

where the sum is over those domains containing a particular crop *c* in the different mixtures, $d$=1,2,...,$D$*.

An approximate estimator of the variance of $\hat{\bar{Y}}_{c1}$ is given by

$$\hat{V}\left(\hat{\bar{Y}}_{c1}\right)=\frac{1}{D^{*2}}\sum_{d=1}^{D*}\hat{V}\left(\bar{y}_{1lrd}\right). \quad \text{...(17)}$$

Then, the estimator of percentage standard error of the proposed estimator of population total, $\hat{\bar{Y}}_{c1}$, is given by

$$\% \widehat{SE}\left(\hat{\bar{Y}}_{c1}\right)=\frac{\sqrt{\hat{V}\left(\hat{\bar{Y}}_{c1}\right)}}{\hat{\bar{Y}}_{c1}}\times 100. \qquad ...(18)$$

**Option 2: Overall sampling**

In the previous section, the sample at the second stage was selected from each of the second-stage strata (crop mixtures). At the time of area enumeration, when the number of crop mixtures observed in villages is large, the selection of a sample from each second-stage stratum may be a very laborious exercise.

In this case, instead of going with stratification for different crop mixtures, a random sample of $n_h'$ villages is selected from $n_{hA}$ villages used for area enumeration and in the second stage of the first phase, a random sample of $m_{hi}'$ parcels are drawn out of $M_{hi}$ parcels in the village and observations on an auxiliary variable, such as farmers' eye estimates of harvested yield of the crop under a mixture scenario are taken. In the second phase, a subsample of size $n_h$ villages are drawn from $n_h'$ first phase villages and $m_{hi}$ parcels from $m_{hi}'$ parcels by SRSWOR. Estimates of the harvested yield of a crop under a specific mixture are obtained by conducting a crop cutting experiment for crop mixtures from these $m_{hi}$ parcels. Questionnaires (option 2) given in annex A of the field test protocol document should be used for the data collection on farmers' eye estimates of crop yield and for recording crop yield by crop cutting experiments. See section 4.2 of Technical Report I for a detailed procedure for conducting a crop cutting experiment.

It can be observed that out of these $n_h$ selected villages, let $n_{hd}$ villages follow a specific $d^{th}$ mixture in $m_{hid}$ sampled parcels (out of $m_{hi}$ parcels), $d=1,2,...,D$. This is a random event. We will take a sufficient sample size so that each domain has adequate sample for design -based estimation of the average crop yield at the district level.

Let $x_{hij}$ define the eye-estimated harvested yield of the crop in the $j^{th}$ parcel grown in the $i^{th}$ village of the $h^{th}$ sub district. Again, let $y_{hij}$ define the harvested yield of the crop from the $j^{th}$ parcel of the $i^{th}$ village of the $h^{th}$ sub district.

Using the information from the eye-estimated yields $(x_{hij})$ as well as the harvested yield by a crop cutting experiment $(y_{hij})$ under a specific mixture (domain) under a stratified two-phase two-stage sampling design, a linear regression estimate can be formed as

$$\bar{y}^*_{1lrd} = \bar{y}^*_d + b^*_{1d}(\bar{x}_d{}^{*\prime} - \bar{x}_d{}^*) \qquad ...(19)$$

where

$$\bar{y}^*_d = \frac{1}{H}\sum_{h=1}^{H}\frac{1}{n_h}\sum_{i=1}^{n_{hd}}\frac{1}{m_{hi}}\sum_{j=1}^{m_{hid}} y_{hij}$$

$$\bar{x}_d{}^{*\prime} = \frac{1}{H}\sum_{h=1}^{H}\frac{1}{n'_h}\sum_{i=1}^{n'_{hd}}\frac{1}{m'_{hi}}\sum_{j=1}^{m'_{hid}} x_{hij}$$

$$\bar{x}_d{}^* = \frac{1}{H}\sum_{h=1}^{H}\frac{1}{n_h}\sum_{i=1}^{n_{hd}}\frac{1}{m_{hi}}\sum_{j=1}^{m_{hid}} x_{hij}$$

By minimizing the variance of the regression estimator with respect to $b^*_{1d}$, the value of $b^*_{1d}$ is given by

$$b^*_{1d} = \frac{c^*_{1xyd}}{c^*_{1xxd}} \qquad ...(20)$$

where,

$$c^*_{1xyd} = \frac{1}{H^2}\sum_{h=1}^{H}\left[\left(\frac{1}{n_h} - \frac{1}{n_h{}'}\right)\left(p_{hd}s_{bxyhd} + p_{hd}q_{hd}\bar{x}_{hd}\bar{y}_{hd}\right) + \frac{1}{n_h n_h{}'}\sum_{i=1}^{n_{hd}}\left(\frac{1}{m_{hi}} - \frac{1}{m_{hi}{}'}\right)\left(p_{hid}\, s_{xyhid} + p_{hid}\, q_{hid}\, \bar{x}_{hid}\bar{y}_{hid}\right)\right]$$

$$c^*_{1xxd} = \frac{1}{H^2}\sum_{h=1}^{H}\left[\left(\frac{1}{n_h} - \frac{1}{n_h{}'}\right)\left(p_{hd}\, s^2_{bxhd} + p_{hd}\, q_{hd}\, \bar{x}^2_{hd}\right) + \frac{1}{n_h n_h{}'}\sum_{i=1}^{n_{hd}}\left(\frac{1}{m_{hi}} - \frac{1}{m_{hi}{}'}\right)\left(p_{hid}\, s^2_{xhid} + p_{hid}\, q_{hid}\, \bar{x}^2_{hid}\right)\right]$$

$$p_{hd} = \frac{n_{hd}}{n_h}, \quad q_{hd} = 1 - p_{hd}$$

$$p_{hid} = \frac{m_{hid}}{m_{hi}}, \quad q_{hid} = 1 - p_{hid}$$

$$s^2_{bxhd} = \frac{1}{n_{hd} - 1} \sum_{i=1}^{n_{hd}} (p_{hid} \bar{x}_{hid} - \bar{x}_{hd})^2$$

$$s_{bxyhd} = \frac{1}{n_{hd} - 1} \sum_{i=1}^{n_{hd}} (p_{hid} \bar{x}_{hid} - \bar{x}_{hd})(p_{hid} \bar{y}_{hid} - \bar{y}_{hd})$$

$$\bar{x}_{hid} = \frac{1}{m_{hid}} \sum_{j=1}^{m_{hid}} x_{hijd}, \; \bar{y}_{hid} = \frac{1}{m_{hid}} \sum_{j=1}^{m_{hid}} y_{hijd}$$

$$\bar{x}_{hd} = \frac{1}{n_{hd}} \sum_{i=1}^{n_{hd}} p_{hid} \bar{x}_{hid}, \; \bar{y}_{hd} = \frac{1}{n_{hd}} \sum_{i=1}^{n_{hd}} p_{hid} \bar{y}_{hid}$$

$$s^2_{xhid} = \frac{1}{m_{hid} - 1} \sum_{j=1}^{m_{hid}} \left( x_{hijd} - \bar{x}_{hid} \right)^2$$

$$s_{xyhid} = \frac{1}{m_{hid} - 1} \sum_{j=1}^{m_{hid}} \left( x_{hijd} - \bar{x}_{hid} \right) \left( y_{hijd} - \bar{y}_{hid} \right)$$

An approximate estimate of variance of the regression estimator is given by

$$\begin{aligned} \hat{V}\left(\bar{y}^*_{1lrd}\right) = \frac{1}{H^2} \sum_{h=1}^{H} \Bigg[ \Bigg\{ & \left( \frac{1}{n_h} - \frac{1}{N_h} \right) \left( p_{hd} s^2_{byhd} + p_{hd} q_{hd} \bar{y}^2_{hd} \right) \\ & + \frac{1}{n_h N_h} \sum_{i=1}^{n_{hd}} \left( \frac{1}{m_{hi}} - \frac{1}{M_{hi}} \right) \left( p_{hid} s^2_{yhid} + p_{hid} q_{hid} \bar{y}^2_{hid} \right) \Bigg\} \\ & - r^{*2}_{1d} \Bigg\{ \left( \frac{1}{n_h} - \frac{1}{n'_h} \right) \left( p_{hd} s^2_{byhd} + p_{hd} q_{hd} \bar{y}^2_{hd} \right) + \\ & \frac{1}{n_h n'_h} \sum_{i=1}^{n_{hd}} \left( \frac{1}{m_{hi}} - \frac{1}{m'_{hi}} \right) \left( p_{hid} s^2_{yhid} + p_{hid} q_{hid} \bar{y}^2_{hid} \right) \Bigg\} \Bigg] \end{aligned} \quad \ldots(21)$$

where

$$s^2_{byhd} = \frac{1}{n_{hd} - 1} \sum_{i=1}^{n_{hd}} (p_{hid} \bar{y}_{hid} - \bar{y}_{hd})^2, \; s^2_{yhid} = \frac{1}{m_{hid} - 1} \sum_{j=1}^{m_{hid}} \left( y_{hijd} - \bar{y}_{hid} \right)^2$$

$$r_{1d}^{*2} = \frac{c_{1xyd}^{*2}}{c_{1xxd}^{*}\, c_{1yyd}^{*}},$$

and $c_{1yyd}^{*}$ is in same functional form as $c_{1xxd}^{*}$ as defined earlier.

Then, the estimator of percentage standard error of the proposed linear regression estimator of average yield under a crop under $d^{\text{th}}$ mixture, $\bar{y}_{1lrd}^{*}$, is given by

$$\% \widehat{SE}(\bar{y}_{1lrd}^{*}) = \frac{\sqrt{\hat{V}\left(\bar{y}_{1lrd}^{*}\right)}}{\bar{y}_{1lrd}^{*}} \times 100. \qquad ...(22)$$

The estimator of yield for a specific crop *c* based on the domains at the district level is given by

$$\hat{\bar{Y}}_{c1}^{*} = \frac{1}{D*}\sum_{d=1}^{D*} \bar{y}_{1lrd}^{*} . \qquad ...(23)$$

where the sum is over all those domains containing a particular crop *c* in the form of different mixtures, *d*=1,2,...,*D**.

An approximate estimator of the variance of $\hat{\bar{Y}}_{c1}^{*}$ is given by

$$\hat{V}\left(\hat{\bar{Y}}_{c1}^{*}\right) = \frac{1}{D^2}\sum_{d=1}^{D} \hat{V}\left(\bar{y}_{1lrd}^{*}\right).$$

...(24)

Then, the estimator of percentage standard error of the proposed estimator of population total, $\hat{\bar{Y}}_{c1}^{*}$, is given by

$$\% \widehat{SE}\left(\hat{\bar{Y}}_{c1}^{*}\right) = \frac{\sqrt{\hat{V}\left(\hat{\bar{Y}}_{c1}^{*}\right)}}{\hat{\bar{Y}}_{c1}^{*}} \times 100. \qquad ...(25)$$

## ANNEX A2

**Category -1 - When cadastral maps are available.**

**Case1: When season-wise crop area is available in the crop register.**

**Estimation procedure for crop area estimation for mixed cropping using the sample survey approach.**

Let a finite population $U=\{1,\ldots, j,\ldots,N\}$ be divided into $H$ strata $\{U_1,\ldots,U_h,\ldots,U_H\}$ and the $h^{th}$ stratum contains $N_h$ primary stage units (PSU). Let, there be $M_{hi}$ second stage units (SSU) in the $i^{th}$ PSU, $i= 1,\ldots,N_h$. As for example, suppose, in a district there are $H$ sub districts that can be considered as $H$ strata. Let, $N_h$ villages/Enumeration Areas in the $h^{th}$ sub district (stratum)be considered as the PSUs, $h=1,\ldots,H$. In general, the total number of villages in each sub district, $N_h$, is known. Suppose, in each village there are $M_{hi}$ number of parcels (SSU) growing crops in different forms of mixtures, such as a pure stand, mixture-1 or mixture-2.

Thus, the different crop mixtures are taken as domains of the study. It is assumed that in each $h^{th}$ sub district, $D$ different crop mixtures are being followed as pure stand, mixture-1, mixture-2 and so on. Here, there would be $\{U_{h1}, \ldots,U_{hd},\ldots,U_{hD}\}$ domains in the $h^{th}$ stratum. Let $N_{hd}$ villages among $N_h$ villages contain parcels (SSU) belonging to $d^{th}$ mixture (domain),$d=1,2,...,D$. Also, let out of $M_{hi}$ parcels (SSU) in the $i^{th}$ cluster of $h^{th}$ strata, $M_{hid}$ parcels (SSU) belong to $d^{th}$ mixture (domain),$d=1,2,...,D$. Although the number of villages ($N_h$) in each sub district is known, the number of villages ($N_{hd}$) following a specific mixture form is generally unknown. The same applies to the number of parcels in a village ($M_{hi}$) as well as specific mixture growing parcels in a village ($M_{hid}$).

The partitioning equation for $D$ domains is given as

$$U_h = \bigcup_{d=1}^{D} U_{hd} \;;h=1,2,\ldots,H. \qquad ...(26)$$

Let $y_{hij}$ be the crop area of $j^{th}$ sample parcel (SSU) in the $i^{th}$ villages (PSU) in the $h^{th}$ sub district (stratum).The total area under $d^{th}$ crop mixture (domain) in a district is given by

$$Y_d = \sum_{h=1}^{H}\sum_{i=1}^{N_{hd}}\sum_{j=1}^{M_{hid}} y_{hij} = \sum_{h=1}^{H}\sum_{i=1}^{N_{hd}} Y_{hid},$$

where, $Y_{hid} = \sum_{j=1}^{M_{hid}} y_{hij}$ =total area of the units within $i^{th}$ PSU belonging to $d^{th}$ domain in the $h^{th}$ stratum.

The population total based on the domains is given as

$$Y = \sum_{d=1}^{D} Y_d = \sum_{d=1}^{D}\sum_{h=1}^{H}\sum_{i=1}^{N_{hd}}\sum_{j=1}^{M_{hid}} y_{hij}.$$

Another sampling design for estimation of crop area at the district level is a stratified two-stage sampling design with two phases at each stage of sampling. Now, suppose, a SRSWOR sample of size $n_h'$ is drawn from $N_h$ PSU (villages). In the second stage, $m_{hi}'$ parcels are selected by SRSWOR for collecting auxiliary information regarding the parcel, such as seed used and farmers' assessment. Let, $n_{hd}'$ villages out of selected $n_h'$ villages follow the specific $d^{th}$ mixture. In the same way, let, $d^{th}$ mixture be grown in $m_{hid}'$ parcels in the $i^{th}$ village out of total selected parcels, $m_{hi}'$ .Again, a sample of $n_h$ villages is selected from $n_h'$ initially selected villages by SRSWOR and in each of those selected villages, a subsample of $m_{hi}$ parcels is selected by SRSWOR. The areas of these sampled parcels are measured by GPS or by the polygon method using a chain and compass. In this situation, questionnaires given in annex A of the field test protocol document should be used for the data collection of the different attributes. Then, the area for component crops is obtained by apportioning using a fixed-ratio approach based on physical observations. Suppose, out of these $n_h$ villages, $n_{hd}$ villages follow a specific $d^{th}$ mixture in $m_{hid}$ sampled parcels.

In the current scenario, the aim is to estimate the total crop area under a specific crop($Y$) and under different mixtures ($Y_d$), $d$=1,2,...,$D$. Here, the information on auxiliary variable needs to be used to obtain more precise estimators.

Let $x_{hij}$ be the auxiliary information, such as seed used and farmers' assessment corresponding to $j^{th}$ sample parcel (SSU) in the $i^{th}$ village (PSU) in the $h^{th}$ sub district (stratum).

A linear regression estimator of the total area under a $d^{th}$ mixture under the stratified two- phase two-stage design can be formed as

$$\hat{Y}_{lr2d} = \hat{Y}_d + b_{Ad}(\hat{X}_d' - \hat{X}_d) \qquad ...(27)$$

where,

$$\hat{Y}_d = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i=1}^{n_{hd}} \frac{M_{hi}}{m_{hi}} \sum_{j=1}^{m_{hid}} y_{hij}, \; \hat{X}_d = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i=1}^{n_{hd}} \frac{M_{hi}}{m_{hi}} \sum_{j=1}^{m_{hid}} y_{hij},$$

$$\hat{X}_d' = \sum_{h=1}^{H} \frac{N_h}{n_h'} \sum_{i=1}^{n_{hd}'} \frac{M_{hi}}{m_{hi}'} \sum_{j=1}^{m_{hid}'} y_{hij}.$$

By minimizing the variance of the linear regression estimator $\hat{Y}_{lr2d}$ with respect to $b_{Ad}$ and ignoring the variation in $b_{Ad}$, the value of $b_{Ad}$ can be shown to be

$$b_{Ad} = c_{Axyd} / c_{Axxd}, \qquad ...(28)$$

where,

$$c_{Axyd} = \sum_{h=1}^{H} N_h^2 \left[ \left( \frac{1}{n_h} - \frac{1}{n_h'} \right) \left( p_{hd}\, s_{bxyhd} + p_{hd}\, q_{hd}\, \bar{x}_{hd} \bar{y}_{hd} \right) + \frac{1}{n_h n_h'} \sum_{i=1}^{n_{hd}} M_{hi}^2 \left( \frac{1}{m_{hi}} - \frac{1}{m_{hi}'} \right) \left( p_{hid}\, s_{xyhid} + p_{hid}\, q_{hid}\, \bar{x}_{hid} \bar{y}_{hid} \right) \right],$$

$$c_{Axxd} = \sum_{h=1}^{H} N_h^2 \left[ \left( \frac{1}{n_h} - \frac{1}{n_h'} \right) \left( p_{hd}\, s_{bxhd}^2 + p_{hd}\, q_{hd}\, \bar{x}_{hd}^2 \right) + \frac{1}{n_h n_h'} \sum_{i=1}^{n_{hd}} M_{hi}^2 \left( \frac{1}{m_{hi}} - \frac{1}{m_{hi}'} \right) \left( p_{hid}\, s_{xhid}^2 + p_{hid}\, q_{hid}\, \bar{x}_{hid}^2 \right) \right],$$

$$p_{hd} = \frac{n_{hd}}{n_h}, \quad q_{hd} = 1 - p_{hd}, \; p_{hid} = \frac{m_{hid}}{m_{hi}}, \quad q_{hid} = 1 - p_{hid},$$

$$s_{bxhd}^2 = \frac{1}{n_{hd} - 1} \sum_{i=1}^{n_{hd}} (M_i\, p_{hid} \bar{x}_{hid} - \bar{x}_{hd})^2,$$

$$s_{bxyhd} = \frac{1}{n_{hd} - 1} \sum_{i=1}^{n_{hd}} (M_i\, p_{hid}\, \bar{x}_{hid} - \bar{x}_{hd})(M_i\, p_{hid} \bar{y}_{hid} - \bar{y}_{hd}),$$

$$s_{xhid}^2 = \frac{1}{m_{hid} - 1} \sum_{j=1}^{m_{hid}} \left( x_{hij} - \bar{x}_{hid} \right)^2, \; s_{xyhid} = \frac{1}{m_{hid} - 1} \sum_{j=1}^{m_{hid}} \left( x_{hij} - \bar{x}_{hid} \right) \left( y_{hij} - \bar{y}_{hid} \right),$$

$$\bar{x}_{hid} = \frac{1}{m_{hid}} \sum_{j}^{m_{hid}} x_{hij}, \quad \bar{y}_{hid} = \frac{1}{m_{hid}} \sum_{j}^{m_{hid}} y_{hij}, \ \bar{x}_{hd} = \frac{1}{n_{hd}} \sum_{i=}^{n_{hd}} M_{hi} p_{hid} \bar{x}_{hid},$$

$$\bar{y}_{hd} = \frac{1}{n_{hd}} \sum_{i=1}^{n_{hd}} M_{hi} p_{hid} \bar{y}_{hid}.$$

An approximate estimate of variance of the linear regression estimator $\hat{Y}_{lr2d}$ is given by

$$\hat{V}\left(\hat{Y}_{lr2d}\right) = \sum_{h=1}^{H} N_h^2 \left[ \left\{ \left( \frac{1}{n_h} - \frac{1}{N_h} \right) \left( p_{hd} s_{byhd}^2 + p_{hd} q_{hd} \bar{y}_{hd}^2 \right) + \frac{1}{n_h N_h} \sum_{i=1}^{n_{hd}} M_{hi}^2 \left( \frac{1}{m_{hi}} - \frac{1}{M_{hi}} \right) \left( p_{hid} s_{yhid}^2 + p_{hid} q_{hid} \bar{y}_{hid}^2 \right) \right\} - r_{Ad}^2 \left\{ \left( \frac{1}{n_h} - \frac{1}{n_h^{'}} \right) \left( p_{hd} s_{byhd}^2 + p_{hd} q_{hd} \bar{y}_{hd}^2 \right) + \frac{1}{n_h n_h^{'}} \sum_{i=1}^{n_{hd}} M_{hi}^2 \left( \frac{1}{m_{hi}} - \frac{1}{m_{hi}^{'}} \right) \left( p_{hid} s_{yhid}^2 + p_{hid} q_{hid} \bar{y}_{hid}^2 \right) \right\} \right],$$

where

$$s_{byhd}^2 = \frac{1}{n_{hd} - 1} \sum_{i=1}^{n_{hd}} \left( M_{hi} p_{hid} \bar{y}_{hid} - \bar{y}_{hd} \right)^2,$$

$$s_{yhid}^2 = \frac{1}{m_{hid} - 1} \sum_{j=1}^{m_{hd}} \left( y_{hij} - \bar{y}_{hid} \right)^2, \ r_{Ad}^2 = \frac{c_{Axyd}^2}{c_{Axxd} \, c_{Ayyd}},$$

and $c_{Ayyd}$ is in same functional form as $c_{Axxd}$ as defined earlier.

Then, the estimator of percentage standard error of the proposed linear regression estimator of population total area under $d^{\text{th}}$ mixture, $\hat{Y}_{lr2d}$, is given by

$$\% \widehat{SE}(\hat{Y}_{lr2d}) = \frac{\sqrt{\hat{V}\left(\hat{Y}_{lr2d}\right)}}{\hat{Y}_{lr2d}} \times 100. \qquad \text{...(30)}$$

The estimator of total area for a specific crop *c* based on all the domains at district level is given by

$$\hat{Y}_{c2} = \sum_{d=1}^{D*} \hat{Y}_{lr2d} \ . \qquad \text{...(31)}$$

where the sum is over all those domains containing a particular crop *c*, *d*=1,2,...,*D**.

An approximate estimate of variance and of the proposed linear regression estimator of population total $\hat{Y}_{c2}$ is given by

$$\hat{V}\left(\hat{Y}_{c2}\right)=\sum_{d=1}^{D*}\hat{V}(\hat{Y}_{lr2d}) \qquad ...(32)$$

The percentage standard error of the proposed linear regression estimator of population total $\hat{Y}_{c2}$ is given by

$$\% \widehat{SE}(\hat{Y}_{c2})=\frac{\sqrt{\hat{V}\left(\hat{Y}_{c2}\right)}}{\hat{Y}_{c2}}\times 100. \qquad ...(33)$$

**Estimation of crop yield for mixed cropping using stratified two-stage two-phase sampling design framework.**

**Option1: Crop mixture-wise sampling.**

At the time of estimation of crop area" under a stratified two-stage two-phase sampling design framework, in $h^{th}$ sub-district (stratum) from the set of $N_h$ villages (PSUs) containing $M_{hi}$ parcels (SSUs) each, a sample of $n_h$ villages with $m_{hi}$ parcels was selected in two phases for area enumeration under a specific mixture of a crop. While surveying the parcels for area enumeration, it was possible to know that out of $m_{hi}$ sampled parcels of the $i^{th}$ village of the $h^{th}$ sub districts, $m_{hid}$ parcels were being grown as the $d^{th}$ mixture of the crop, $\forall\ d=1,2,...,D$.

To come up with average yield estimation of the different forms of mixture, the following sampling design needs to be followed:

In the first phase of sampling, from the set of $n_h$ villages (PSU) $n_h^{"}$ villages are selected by SRSWOR design and within the selected villages, for each of the $d^{th}$ mixture, $d=1,2,...,D$, samples of $m_{hid}^{"}$ parcels (SSUs) are selected by SRSWOR from the $m_{hid}$ parcels used for area enumeration. From the selected parcels, farmers' eye estimates of harvested yield of the crop under mixture have to be recorded.

Let, $x_{hijd}$ define the eye-estimated harvested yield of the crop in the $j^{th}$ parcel grown in the $d^{th}$ mixture in the $i^{th}$ village of the $h^{th}$ sub district.

Again in the second phase of sampling, let a sample of $n_h^{'''}$ villages be selected by SRSWOR design from the $n_h^{''}$ selected villages and within each of the selected villages, for each mixture $d = 1, 2, ....D$, sample of $m_{hid}^{'''}$ parcels are selected from the $m_{hid}^{''}$ third-phase second-stage units in the $i^{th}$ village of the $h^{th}$ sub district. Data collection on farmers' eye estimates of crop yields should be carried out using questionnaires (option 1) given in annex A of the field test protocol document and for conducting a crop-cutting experiment, questionnaires given in same annex should be used. See section 4.2 of Technical Report I for detailed procedures for conducting a crop cutting experiment.

In the final sample, it can be observed that out of $n_h^{'''}$ sample villages, $n_{hd}^{'''}$ villages are following the $d^{th}$ mixture, $\forall d = 1, 2, ....D$. From these $m_{hid}^{'''}$ parcels in the final sample, estimates of the harvested yields of a crop under a specific mixture are obtained by conducting a crop cutting experiment for mixed cropping.

Let $y_{hijd}$ define the harvested yield of the crop under $d^{th}$ mixture from the $j^{th}$ parcel in the $i^{th}$ village of the $h^{th}$ sub district. From the harvested yield of the sampled parcels, an estimate of the average crop yield under a specific crop mixture for a district can be given by

$$\bar{y}_{2d} = \frac{1}{H}\sum_{h=1}^{H}\frac{1}{n_h^{'''}}\sum_{i=1}^{n_{hd}^{'''}}\frac{1}{m_{hid}^{'''}}\sum_{j=1}^{m_{hid}^{'''}} y_{hijd}. \qquad \ldots(34)$$

Using the information from the eye-estimate yields under a specific mixture (domain) and using double sampling approach, a linear regression estimator can be formed as

$$\bar{y}_{2lrd} = \bar{y}_{2d} + b_{2d}(\bar{x}'_{2d} - \bar{x}_{2d}) \qquad \ldots(35)$$

Where $\bar{x}_{2d} = \frac{1}{H}\sum_{h=1}^{H}\frac{1}{n_h^{'''}}\sum_{i=1}^{n_{hd}^{'''}}\frac{1}{m_{hid}^{'''}}\sum_{j=1}^{m_{hid}^{'''}} x_{hijd}$ and $\bar{x}'_{2d} = \frac{1}{H}\sum_{h=1}^{H}\frac{1}{n_h^{''}}\sum_{i=1}^{n_{hd}^{''}}\frac{1}{m_{hid}^{''}}\sum_{j=1}^{m_{hid}^{''}} x_{hijd}$.

By minimizing the variance of $\bar{y}_{2lrd}$ with respect to $b_{2d}$ and ignoring the variation in $b_{2d}$, the value of $b_{2d}$ is given by

$$b_{2d} = \frac{c_{2xyd}}{c_{2xxd}} \qquad \ldots(36)$$

where,

$$c_{2xyd}=\frac{1}{H^2}\sum_{h=1}^{H}\left[\left(\frac{1}{n_h'''}-\frac{1}{n_h''}\right)(p_{hd}s_{bxyhd}+p_{hd}q_{hd}\bar{x}_{hd}\bar{y}_{hd})+\frac{1}{n_h'''n_h''}\sum_{i}^{n_{hd}'''}\left(\frac{1}{m_{hid}'''}-\frac{1}{m_{hid}''}\right)s_{xyhid}\right],$$

$$c_{2xxd}=\frac{1}{H^2}\sum_{h=1}^{H}\left[\left(\frac{1}{n_h'''}-\frac{1}{n_h''}\right)(p_{hd}s_{bxhd}^2+p_{hd}q_{hd}\bar{x}_{hd}^2)+\frac{1}{n_h'''n_h''}\sum_{i}^{n_{hd}'''}\left(\frac{1}{m_{hid}'''}-\frac{1}{m_{hid}''}\right)s_{xhid}^2\right],$$

$$s_{bxhd}^2=\frac{1}{n_{hd}'''-1}\sum_{i=1}^{n_{hd}'''}(\bar{x}_{hid}-\bar{x}_{hd})^2,\ s_{bxyhd}=\frac{1}{n_{hd}'''-1}\sum_{i=1}^{n_{hd}'''}(\bar{x}_{hid}-\bar{x}_{hd})(\bar{y}_{hid}-\bar{y}_{hd}),$$

$$s_{xhid}^2=\frac{1}{m_{hid}'''-1}\sum_{j=1}^{m_{hid}'''}\left(x_{hijd}-\bar{x}_{hid}\right)^2,$$

$$s_{xyhid}=\frac{1}{m_{hid}'''-1}\sum_{j=1}^{m_{hid}'''}\left(x_{hijd}-\bar{x}_{hid}\right)\left(y_{hijd}-\bar{y}_{hid}\right),$$

$$p_{hd}=\frac{n_{hd}'''}{n_h'''},\quad q_{hd}=1-p_{hd},\ p_{hid}=\frac{m_{hid}'''}{m_{hi}'''},\quad q_{hid}=1-p_{hid},$$

$$\bar{x}_{hid}=\frac{1}{m_{hid}'''}\sum_{j}^{m_{hid}'''}x_{hijd},\quad \bar{y}_{hid}=\frac{1}{m_{hid}'''}\sum_{j}^{m_{hid}'''}y_{hijd},$$

$$\bar{x}_{hd}=\frac{1}{n_{hd}'''}\sum_{i=}^{n_{hd}'''}\bar{x}_{hid},\ \bar{y}_{hd}=\frac{1}{n_{hd}'''}\sum_{i=1}^{n_{hd}'''}\bar{y}_{hid}.$$

An approximate estimate of variance of the linear regression estimator $\bar{y}_{2lrd}$ is given by

$$\hat{V}(\bar{y}_{2lrd})=\frac{1}{H^2}\sum_{h=1}^{H}\left[\left\{\left(\frac{1}{n_h''}-\frac{1}{N_h}\right)\left(p_{hd}s_{byhd}^2+p_{hd}q_{hd}\bar{y}_{hd}^2\right)+\frac{1}{n_h''N_h}\sum_{i=1}^{n_{hd}''}\left(\frac{1}{m_{hid}''}-\frac{1}{M_{hid}}\right)s_{yhid}^2\right\}\right. \quad \ldots(37)$$
$$\left.-r_{2d}^2\left\{\left(\frac{1}{n_h'''}-\frac{1}{n_h''}\right)\left(p_{hd}s_{byhd}^2+p_{hd}q_{hd}\bar{y}_{hd}^2\right)+\frac{1}{n_h'''n_h''}\sum_{i=1}^{n_{hd}'''}\left(\frac{1}{m_{hid}'''}-\frac{1}{m_{hid}''}\right)s_{yhid}^2\right\}\right]$$

where

$$s_{byhd}^2=\frac{1}{n_{hd}'''-1}\sum_{i=1}^{n_{hd}'''}\left(\bar{y}_{hid}-\bar{y}_{hd}\right)^2,\ s_{yhid}^2=\frac{1}{m_{hid}'''-1}\sum_{j=1}^{m_{hid}'''}\left(y_{hijd}-\bar{y}_{hid}\right)^2,$$

$$r_{2d}^2 = \frac{c_{2xyd}^2}{c_{2xxd}\, c_{2yyd}},$$

and $c_{2yyd}$ is in same functional form as $c_{2xxd}$ as defined earlier.

Then, the estimator of percentage standard error of the proposed linear regression estimator of average yield under a crop under $d^{th}$ mixture, $\bar{y}_{2lrd}$, is given by

$$\%\,\widehat{SE}(\bar{y}_{2lrd}) = \frac{\sqrt{\hat{V}\left(\bar{y}_{2lrd}\right)}}{\bar{y}_{2lrd}} \times 100. \qquad \ldots(38)$$

The estimator of yield for a specific crop *c* based on all the domains at the district level is given by

$$\hat{\bar{Y}}_{c2} = \frac{1}{D^*}\sum_{d=1}^{D^*} \bar{y}_{2lrd}\ . \qquad \ldots(39)$$

where the sum is over the domains containing a particular crop *c* in the different mixtures, *d*=1,2,...,*D**.

An approximate estimator of the variance of $\hat{\bar{Y}}_{c2}$ is given by

$$\hat{V}\left(\hat{\bar{Y}}_{c2}\right) = \frac{1}{D^{*2}}\sum_{d=1}^{D^*} \hat{V}\left(\bar{y}_{2lrd}\right). \qquad \ldots(40)$$

Then, the estimator of percentage standard error of the proposed estimator of population total, $\hat{\bar{Y}}_{c2}$, is given by

$$\%\,\widehat{SE}\left(\hat{\bar{Y}}_{c2}\right) = \frac{\sqrt{\hat{V}\left(\hat{\bar{Y}}_{c2}\right)}}{\hat{\bar{Y}}_{c2}} \times 100. \qquad \ldots(41)$$

**Option2: Overall sampling.**

In the previous option, the sample at the second stage was selected from each of the second-stage strata (crop mixtures). At the time of area enumeration, when the number of crop mixtures observed in the villages is large, the selection of sample from each second stage stratum could be a very laborious exercise. In this case, instead of going with a second stage stratification for different crop mixtures, a random sample of $n_h''$ villages is drawn from $n_{hA}$ area enumeration sampled villages, and in the second stage of the first phase, an overall sample of $m_{hi}''$ parcels is drawn out of $M_{hi}$ parcels in the village and observations on auxiliary information, such as farmers' eye estimates of the harvested yield of the crop under mixture, are taken from these parcels. In the second phase of sampling, a subsample of size $n_h'''$ villages is drawn from $n_h''$ villages and $m_{hi}'''$ parcels from $m_{hi}''$ second phase parcels by SRSWOR. From these $m_{hi}$ parcels, estimates of the harvested yield of a crop mixture can be obtained by conducting a crop cutting experiment for mixed cropping. Refer to section 4.2 of Technical Report I for detailed procedures for conducting a crop cutting experiment. Data collection on farmers' eye estimates of crop yields should be carried out using questionnaires (option 2) and for conducting crop cutting experiment, questionnaires given in annex A of field test protocol should be used.

In the final sample, it can be observed that out of $n_h'''$ sample villages, $n_{hd}'''$ villages are following the $d^{th}$ mixture and in the $i^{th}$ village $m_{hid}'''$ parcels $d^{th}$ mixture is grown, $\forall d = 1, 2, ....D$.

Let, $x_{hij}$ defines the eye-estimated harvested yield of the crop in the $j^{th}$ parcel grown in the $i^{th}$ village of the $h^{th}$ sub-district. Again, let, $y_{hij}$ defines the harvested yield of the crop from the $j^{th}$ parcel the $i^{th}$ village of the $h^{th}$ sub district.

Using the information from the eye-estimated yields , $(x_{hij})$ as well as the harvested yield by crop cutting experiment $(y_{hij})$ under a specific mixture (domain) under a stratified cluster two-phase two-stage sampling design, a linear regression estimator can be formed as

$$\bar{y}^*_{2lrd} = \bar{y}^*_{2d} + b^*_{2d}(\bar{x}_{2d}{}^{*\prime} - \bar{x}_{2d}{}^{*}) \qquad ...(42)$$

where

$$\bar{y}_{2d}{}^{*} = \frac{1}{H}\sum_{h=1}^{H}\frac{1}{n_h^{'''}}\sum_{i=1}^{n_{hd}^{'''}}\frac{1}{m_{hi}^{'''}}\sum_{j=1}^{m_{hid}^{'''}} y_{hij},\ \bar{x}_{2d}{}^{*} = \frac{1}{H}\sum_{h=1}^{H}\frac{1}{n_h^{'''}}\sum_{i=1}^{n_{hd}^{'''}}\frac{1}{m_{hi}^{'''}}\sum_{j=1}^{m_{hid}^{'''}} x_{hijd}$$

and

$$\bar{x}_{2d}{}^{*\prime} = \frac{1}{H}\sum_{h=1}^{H}\frac{1}{n_h^{''}}\sum_{i=1}^{n_{hd}^{''}}\frac{1}{m_{hi}^{''}}\sum_{j=1}^{m_{hid}^{''}} x_{hij}.$$

By minimizing the variance of the regression estimator with respect to $b_{2d}^{*}$, the value of $b_{2d}^{*}$ is given by

$$b_{2d}^{*} = \frac{c_{2xyd}^{*}}{c_{2xxd}^{*}} \qquad ...(43)$$

where,

$$c_{2xyd}^{*} = \frac{1}{H^2}\sum_{h=1}^{H}\left[\left(\frac{1}{n_h^{'''}} - \frac{1}{n_h^{''}}\right)(p_{hd}s_{bxyhd}^{2} + p_{hd}q_{hd}\bar{x}_{hd}\bar{y}_{hd}) + \frac{1}{n_h^{'''}n_h^{''}}\sum_{i}^{n_{hd}^{'''}}\left(\frac{1}{m_{hid}^{'''}} - \frac{1}{m_{hid}^{''}}\right)\left(p_{hid}\, s_{xyhid} + p_{hid}\, q_{hid}\, \bar{x}_{hid}\bar{y}_{hid}\right)\right],$$

$$c_{2xxd}^{*} = \frac{1}{H^2}\sum_{h=1}^{H}\left[\left(\frac{1}{n_h^{'''}} - \frac{1}{n_h^{''}}\right)(p_{hd}s_{bxhd}^{2} + p_{hd}q_{hd}\bar{x}_{hd}^{2}) + \frac{1}{n_h^{'''}n_h^{''}}\sum_{i}^{n_{hd}^{'''}}\left(\frac{1}{m_{hid}^{'''}} - \frac{1}{m_{hid}^{''}}\right)\left(p_{hid}\, s_{xhid}^{2} + p_{hid}\, q_{hid}\, \bar{x}_{hid}^{2}\right)\right],$$

$$p_{hd} = \frac{n_{hd}^{'''}}{n_h^{'''}}, \quad q_{hd} = 1 - p_{hd},\ p_{hid} = \frac{m_{hid}^{'''}}{m_{hi}^{'''}}, \quad q_{hid} = 1 - p_{hid},$$

$$s_{bxhd}^{2} = \frac{1}{n_{hd}^{'''} - 1}\sum_{i=1}^{n_{hd}^{'''}}(p_{hid}\bar{x}_{hid} - \bar{x}_{hd})^2,$$

$$s_{bxyhd} = \frac{1}{n_{hd}^{'''} - 1}\sum_{i=1}^{n_{hd}^{'''}}(p_{hid}\bar{x}_{hid} - \bar{x}_{hd})(p_{hid}\bar{y}_{hid} - \bar{y}_{hd}),$$

$$\bar{x}_{hid} = \frac{1}{m_{hid}^{'''}}\sum_{j}^{m_{hid}^{'''}} x_{hij}, \quad \bar{y}_{hid} = \frac{1}{m_{hid}^{'''}}\sum_{j}^{m_{hid}^{'''}} y_{hij},$$

$$\bar{x}_{hd} = \frac{1}{n'''_{hd}} \sum_{i=}^{n'''_{hd}} \bar{x}_{hid}, \ \bar{y}_{hd} = \frac{1}{n'''_{hd}} \sum_{i=1}^{n'''_{hd}} \bar{y}_{hid}.$$

$$s^2_{xhid} = \frac{1}{m'''_{hid} - 1} \sum_{j=1}^{m'''_{hid}} \left( x_{hij} - \bar{x}_{hid} \right)^2, \ s_{xyhid} = \frac{1}{m'''_{hid} - 1} \sum_{j=1}^{m'''_{hid}} \left( x_{hij} - \bar{x}_{hid} \right)\left( y_{hij} - \bar{y}_{hid} \right),$$

An approximate estimate of variance of the regression estimator is given by

$$\begin{aligned} \hat{V}\left(\bar{y}_{2lrd}\right) = \frac{1}{H^2} \sum_{h=1}^{H} \Bigg[ \Bigg\{ \left( \frac{1}{n'''_h} - \frac{1}{N_h} \right) \left( p_{hd} s^2_{byhd} + p_{hd} q_{hd} \bar{y}^2_{hd} \right) \\ + \frac{1}{n'''_h N_h} \sum_{i=1}^{n'''_{hd}} \left( \frac{1}{m'''_{hid}} - \frac{1}{M_{hid}} \right) \left( p_{hid} s^2_{yhid} + p_{hid} q_{hid} \bar{y}^2_{hid} \right) \Bigg\} \\ - r^{*2}_{2d} \Bigg\{ \left( \frac{1}{n'''_h} - \frac{1}{n''_h} \right) \left( p_{hd} s^2_{byhd} + p_{hd} q_{hd} \bar{y}^2_{hd} \right) \\ + \frac{1}{n'''_h n''_h} \sum_{i=1}^{n'''_{hd}} \left( \frac{1}{m'''_{hid}} - \frac{1}{m''_{hid}} \right) \left( p_{hid} s^2_{yhid} + p_{hid} q_{hid} \bar{y}^2_{hid} \right) \Bigg\} \Bigg] \end{aligned} \quad ...(44)$$

where,

$$s^2_{byhd} = \frac{1}{n'''_{hd} - 1} \sum_{i=1}^{n'''_{hd}} \left( p_{hid} \bar{y}_{hid} - \bar{y}_{hd} \right)^2, \ s^2_{yhid} = \frac{1}{m'''_{hid} - 1} \sum_{j=1}^{m'''_{hid}} \left( y_{hij} - \bar{y}_{hid} \right)^2,$$

$$r^{*2}_{2d} = \frac{c^{*2}_{2xyd}}{c^{*}_{2xxd} \, c^{*}_{2yyd}},$$

and $c^*_{2yyd}$ is in same functional form as $c^*_{2xxd}$ as defined earlier.

Then, the estimator of percentage standard error of the proposed linear regression estimator of average yield under of a crop under $d^{\text{th}}$ mixture, $\bar{y}^*_{2lrd}$ , is given by

$$\% \widehat{SE}(\bar{y}^*_{2lrd}) = \frac{\sqrt{\hat{V}\left(\bar{y}^*_{2lrd}\right)}}{\bar{y}^*_{2lrd}} \times 100. \quad ...(45)$$

The estimator of yield for a specific crop *c* based on the domains at the district level is given by

$$\hat{\bar{Y}}^*_{c2} = \frac{1}{D^*}\sum_{d=1}^{D^*} \bar{y}^*_{2lrd} \ . \qquad \text{...(46)}$$

where the sum is over all those domains containing a particular crop *c* in the different mixtures, *d*=1,2,...,*D**.

An approximate estimator of the variance of $\hat{\bar{Y}}^*_{c2}$ is given by

$$\hat{V}\left(\hat{\bar{Y}}^*_{c2}\right) = \frac{1}{D^{*2}}\sum_{d=1}^{D^*} \hat{V}\left(\bar{y}^*_{2lrd}\right). \qquad \text{...(47)}$$

Then, the estimator of percentage standard error of the proposed estimator of population total, $\hat{\bar{Y}}^*_{c2}$, is given by

$$\% \widehat{SE}\left(\hat{\bar{Y}}^*_{c2}\right) = \frac{\sqrt{\hat{V}\left(\hat{\bar{Y}}^*_{c2}\right)}}{\hat{\bar{Y}}^*_{c2}} \times 100. \qquad \text{...(48)}$$

## ANNEX B

**Category 2- Area frame approach**

**Estimation procedure for crop area under mixed cropping when area frame is available.**

Let, in a district, there are $H$ sub districts that can be considered as $H$ strata. Let there be $N_h$ enumeration areas in the $h^{th}$ sub district (stratum) that are considered as PSUs, $h$=1,…,H. In general, the total number of enumeration areas in each sub district, $N_h$, are known. Suppose, each enumeration area can be divided into $M_{hi}$ segments (SSU), each containing $T_{hij}$ parcels (USU).In this, parcels crops are grown in different forms of mixture, such as a pure stand, mixture-1 and mixture-2.Therefore, in this case, the different crop mixtures are taken as domains of the study. It is assumed that in each $h^{th}$ sub district, $D$ different crop mixtures are being followed as, for example, pure stand, mixture-1 and mixture-2. Thus, there would be $\{U_{h1}, \ldots, U_{hd}, \ldots, U_{hD}\}$ domains in the $h^{th}$ stratum.

Let $y_{hijk}$ be the crop area of $k^{th}$ parcel (USU) in the $j^{th}$ segment (SSU) of the $i^{th}$ EA (PSU) in the $h^{th}$ sub district (stratum).

The total area under $d^{th}$ crop mixture (domain) in a district is given by

$$Y_d = \sum_{h=1}^{H} \sum_{i=1}^{N_{hd}} \sum_{j=1}^{M_{hid}} \sum_{k=1}^{T_{hijd}} y_{hijk}, \qquad \text{...(49)}$$

The population total based on the domains is given as

$$Y = \sum_{d=1}^{D} Y_d = \sum_{d=1}^{D} \sum_{h=1}^{H} \sum_{i=1}^{N_{hd}} \sum_{j=1}^{M_{hid}} \sum_{k=1}^{T_{hijd}} y_{hijk}.$$

In these situations, a sampling design for estimation of the crop area at the district level is stratified in a two-stage cluster sampling design with two phases in each stage of sampling using the area frame approach. Suppose, in the first phase, a sample of $n_h'$ enumeration areas are selected from $N_h$ enumeration area as (PSU) by probability proportional to the size with replacement (PPSWR) design from the enumeration areas in the $h^{th}$ stratum. The probability of selecting $i^{th}$ EA in the $h^{th}$ stratum is taken $z_{hi}' = X_{hi}/X_h$, where $X$ may be taken as total agricultural land. In the second stage of first phase, within each of the selected enumeration areas, from the set of $M_{hi}$ segments, a subsample of $m_{hi}'$ segments are selected by SRSWOR design. Then, each of the $T_{hij}$ parcels growing crop mixtures are completely enumerated for collecting auxiliary information regarding the parcel, such as seed used and farmers' assessment, j=1,… $m_{hi}'$.Let, $n_{hd}'$ enumeration areas out of the selected $n_h'$ Enumeration areas follow the specific $d^{th}$ mixture. In the same way, let, $d^{th}$ mixture be grown in $m_{hid}'$ selected segments in their $T_{hijd}$ parcels.

In the second phase of sampling, a sample of $n_h$ enumeration areas are selected from $n_h'$ initially selected enumeration areas (PSU) by SRSWOR. In each of the selected enumeration areas, a subsample of $m_{hi}$ segments are selected by SRSWOR design and the $T_{hij}$ parcels in the selected household are completely enumerated. The areas of each of the parcels growing crop mixtures of those sampled segments are measured by GPS or by the polygon method using a chain and compass. Data collection should be made using the questionnaires given in annex B of the field test protocol document. Then, the area for component crops are obtained by apportioning using a fixed ratio approach based on physical observations. Suppose, out of those parcels growing crop mixtures by $m_{hi}$ segments of $n_h$ enumeration areas, $n_{hd}$ enumeration areas are following a specific $d^{th}$ mixture in $T_{hijd}$ parcels of $m_{hid}$ sampled segments.

In the current scenario, the objective is to estimate the total crop area under a specific crop($Y$) as well as under different mixtures ($Y_d$), $d$=1,2,...,$D$.

Let, $x_{hijk}$ be the auxiliary information, such as seed used and farmers' assessment, corresponding to $k^{th}$ parcel (SSU) of the $j^{th}$ selected segment in the $i^{th}$ enumeration area(PSU) in the $h^{th}$ sub-district (stratum), whereas, $y_{hijk}$ is the crop area of the corresponding parcel measured by GPS and the rope and compass method.

A linear regression estimator of the total area under a $d^{th}$ mixture under the stratified two-phase two-stage cluster sampling design can be formed as

$$\hat{Y}_{lr2d} = \hat{Y}_d + b_{A2d}(\hat{X}_d' - \hat{X}_d) \qquad ...(50)$$

where,

$$\hat{Y}_d = \sum_{h=1}^{H}\frac{1}{n_h}\sum_{i=1}^{n_{hd}}\frac{1}{z_{hi}^{'}}\frac{M_{hi}}{m_{hi}}\sum_{j=1}^{m_{hid}}\sum_{k=1}^{T_{hijd}} y_{hijk} = \sum_{h=1}^{H}\frac{1}{n_h}\sum_{i=1}^{n_{hd}}\frac{1}{z_{hi}^{'}}\frac{M_{hi}}{m_{hi}}\sum_{j=1}^{m_{hid}} y_{hij.d} = \sum_{h=1}^{H}\frac{1}{n_h}\sum_{i=1}^{n_{hd}}\frac{\hat{Y}_{hid}}{z_{hi}^{'}},$$

$$\hat{X}_d = \sum_{h=1}^{H}\frac{1}{n_h}\sum_{i=1}^{n_{hd}}\frac{\hat{X}_{hid}}{z_{hi}^{'}},\quad \hat{X}_d^{'} = \sum_{h=1}^{H}\frac{1}{n_h^{'}}\sum_{i=1}^{n_{hd}}\frac{\hat{X}_{hid}^{'}}{z_{hi}^{'}}$$

$$\hat{Y}_{hid} = \frac{M_{hi}}{m_{hi}}\sum_{j=1}^{m_{hid}}\sum_{k=1}^{T_{hijd}} y_{hijk},\ \hat{X}_{hid} = \frac{M_{hi}}{m_{hi}}\sum_{j=1}^{m_{hid}}\sum_{k=1}^{T_{hijd}} x_{hijk},\ \hat{X}_{hid}^{'} = \frac{M_{hi}}{m_{hi}^{'}}\sum_{j=1}^{m_{hid}^{'}}\sum_{k=1}^{T_{hijd}} x_{hijk}\ .$$

By minimizing the variance of the linear regression estimator $\hat{Y}_{lr2d}$ with respect to $b_{A2d}$ and ignoring the variation in $b_{A2d}$, the value of $b_{A2d}$can be shown as

$$b_{A2d} = c_{A2xyd} / c_{A2xxd}\ , \qquad \ldots(51)$$

where,

$$c_{A2xyd} = \sum_{h=1}^{H}\left(\frac{1}{n_h} - \frac{1}{n_h^{'}}\right)\left(p_{hd}\ s_{bxyhd} + p_{hd}\ q_{hd}\ \hat{X}_{hd}\hat{Y}_{hd}\right),$$

$$c_{A2xxd} = \sum_{h=1}^{H}\left(\frac{1}{n_h} - \frac{1}{n_h^{'}}\right)\left(p_{hd}\ s_{bxhd}^2 + p_{hd}\ q_{hd}\ \hat{X}_{hd}^2\right),$$

$$p_{hd} = \frac{n_{hd}}{n_h},\quad q_{hd} = 1 - p_{hd},$$

$$s_{bxhd}^2 = \frac{1}{n_{hd}-1}\sum_{i=1}^{n_{hd}}\left(\frac{M_{hi}\,p_{hid}\,\overline{x}_{hid}}{z_{hi}^{'}} - \hat{X}_{hd}\right)^2,$$

$$s_{bxyhd} = \frac{1}{n_{hd}-1}\sum_{i=1}^{n_{hd}}\left(\frac{M_{hi}\,p_{hid}\,\overline{x}_{hid}}{z_{hi}^{'}} - \hat{X}_{hd}\right)\left(\frac{M_{hi}\,p_{hid}\,\overline{y}_{hid}}{z_{hi}^{'}} - \hat{Y}_{hd}\right),$$

$$\overline{x}_{hid} = \frac{1}{m_{hid}}\sum_{j=1}^{m_{hid}} x_{hij.d},\quad \overline{y}_{hid} = \frac{1}{m_{hid}}\sum_{j}^{m_{hid}} y_{hij.d},$$

$$\hat{X}_{hd} = \frac{1}{n_{hd}}\sum_{i=}^{n_{hd}}\frac{M_{hi}\,p_{hid}\,\overline{x}_{hid}}{z_{hi}^{'}}, \hat{Y}_{hd} = \frac{1}{n_{hd}}\sum_{i=1}^{n_{hd}}\frac{M_{hi}\,p_{hid}\,\overline{y}_{hid}}{z_{hi}^{'}}.$$

An approximate estimate of variance of the linear regression estimator $\hat{Y}_{lr2d}$ is given by

$$\hat{V}\left(\hat{Y}_{lr2d}\right)=\sum_{h=1}^{H}\left(\frac{1}{n_h}-r_{A2d}^{2}\left(\frac{1}{n_h}-\frac{1}{n_h^{'}}\right)\right)\left(p_{hd}\,s_{byhd}^{2}+p_{hd}\,q_{hd}\,\hat{Y}_{hd}^{2}\right), \qquad ...(52)$$

where

$$s_{byhd}^{2}=\frac{1}{n_{hd}-1}\sum_{i=1}^{n_{hd}}\left(\frac{M_{hi}\,p_{hid}\,\bar{y}_{hid}}{z_{hi}^{'}}-\hat{Y}_{hd}\right)^{2}, \quad r_{A2d}^{2}=\frac{c_{A2xyd}^{2}}{c_{A2xxd}\,c_{A2yyd}},$$

and $c_{A2yyd}$ is in same functional form as $c_{A2xxd}$ as defined earlier.

Then, the estimator of percentage standard error of the proposed linear regression estimator of total area under $d^{\text{th}}$ mixture, $\hat{Y}_{lr2d}$, is given by

$$\%\,\widehat{SE}(\hat{Y}_{lr2d})=\frac{\sqrt{\hat{V}\left(\hat{Y}_{lr2d}\right)}}{\hat{Y}_{lr2d}}\times 100. \qquad ...(53)$$

The estimator of the total area for a specific crop c based on the domains at the district level is given by

$$\hat{Y}_{c2}=\sum_{d=1}^{D*}\hat{Y}_{lr2d}\ . \qquad ...(54)$$

where the sum is over the domains containing a particular crop $c$,$d$=1,2,...,$D*$.

An approximate estimate of variance of the linear regression estimator $\hat{Y}_{c2}$ is given by

$$\hat{V}\left(\hat{Y}_{c2}\right)=\sum_{d=1}^{D*}\hat{V}(\hat{Y}_{lr2d})$$

Then, the estimator of percentage standard error of the proposed estimator of population total, $\hat{Y}_{c2}$, is given by

$$\%\,\widehat{SE}(\hat{Y}_{c2})=\frac{\sqrt{\hat{V}\left(\hat{Y}_{c2}\right)}}{\hat{Y}_{c2}}\times 100. \qquad ...(55)$$

**Estimation of crop yield under mixed cropping using a stratified two-stage two-phase sampling design framework under an area frame approach.**

Crop yield estimation in the context of the area frame approach, is proposed, wherein the sample selected for the crop area estimation by area frame approach should be considered to construct the sampling frame for yield estimation. For this purpose, in each of the sampled enumeration areas in a sub district, first, a list of parcels growing different mixtures is prepared using the sample or area enumeration.

Suppose, for the "estimation of crop area" by the area frame approach under a stratified two- stage cluster sampling design framework (as discussed earlier), in $h^{th}$ sub district (stratum) from the set of $N_h$ enumeration areas (PSUs), a sample of $n_h$ enumeration areas was selected. Then, a subsample of segments was selected in each selected enumeration area and the parcels of selected segments were completely enumerated for area under a specific crop mixture. While surveying the parcels for area enumeration, a list of parcels growing different crop mixtures should be prepared in an enumeration area. Let, in total, there are $m_{hid}$ parcels where the $d^{th}$ crop mixture is grown out of the total $m_{hi}$ sampled parcels in the $i^{th}$ enumeration areas of the $h^{th}$ sub districts, $\forall\ i=1,...,n_{hd};\ d=1,...,D; h=1,...,H.$ Here, it should be noted that, a segment-wise crop mixture list need not be prepared.

In the first phase of sampling, from the set of $n_h$ EAs (PSU) $n_h^{''}$ enumeration areas are selected by SRSWOR design. Within the selected enumeration areas, for each of the $d^{th}$ mixture, $d=1,2,...,D$, samples of $m_{hid}^{''}$ parcels (SSUs) are selected by SRSWOR from the $m_{hid}$ parcels used for recording the area. From the selected parcels farmers' eye estimates of harvested yield of the crop under the mixture has to be recorded. Let, $x_{hijd}$ defines the eye-estimated harvested yield of the crop in the $j^{th}$ parcel grown in the $d^{th}$ mixture in the $i^{th}$ enumeration areas of the $h^{th}$ sub district.

In the second phase of sampling, let a sample of $n_h^{'''}$ enumeration areas be selected by SRSWOR design from the $n_h^{''}$ selected enumeration areas and within each selected enumeration area, for each mixture $d=1,2,....D$, a sample of $m_{hid}^{'''}$ parcels are selected from the $m_{hid}^{''}$ units in the $i^{th}$ enumeration areas of the $h^{th}$ sub district.

In the final sample, it can be observed that out of $n_h^{'''}$ sample villages, $n_{hd}^{'''}$ villages are following the $d^{th}$ mixture, $\forall\, d=1,2,....D$. From those $m_{hid}^{'''}$ parcels in the final sample, estimates of the harvested yields of a crop under a specific mixture are obtained by conducting a crop cutting experiment for mixed cropping. See section 4.2 of Technical Report I for a detailed procedure for cutting experiment. Data collection on farmers' eye estimates of crop yields should be carried out using the questionnaire in annex B of the field test protocol document, and for conducting a crop cutting experiment, the questionnaires given in the same annex should be used.

Let, $y_{hijd}$ defines the harvested yield of the crop under $d^{\text{th}}$ mixture from the $j^{\text{th}}$ parcel in the $i^{\text{th}}$ enumeration areas of the $h^{\text{th}}$ sub district.

By following a double sampling approach using the information from the eye-estimated yields and the harvested yield of the sampled parcels under a specific mixture (domain), a linear regression estimator of the average crop yield under a specific crop mixture for a district can be given by

$$\bar{y}_{2lrd} = \bar{y}_{2d} + b_{2d}(\bar{x}'_{2d} - \bar{x}_{2d}) \qquad \ldots(56)$$

where, $\bar{y}_{2d} = \frac{1}{H}\sum_{h=1}^{H}\frac{1}{n_h'''}\sum_{i=1}^{n_{hd}'''}\frac{1}{m_{hid}'''}\sum_{j=1}^{m_{hid}'''} y_{hijd}$,

$$\bar{x}_{2d} = \frac{1}{H}\sum_{h=1}^{H}\frac{1}{n_h'''}\sum_{i=1}^{n_{hd}'''}\frac{1}{m_{hid}'''}\sum_{j=1}^{m_{hid}'''} x_{hijd},$$

$$\bar{x}'_{2d} = \frac{1}{H}\sum_{h=1}^{H}\frac{1}{n_h''}\sum_{i=1}^{n_{hd}''}\frac{1}{m_{hid}''}\sum_{j=1}^{m_{hid}''} x_{hijd},$$

$$b_{2d} = \frac{c_{2xyd}}{c_{2xxd}}$$

$$c_{2xyd} = \frac{1}{H^2}\sum_{h=1}^{H}\left[\left(\frac{1}{n_h'''} - \frac{1}{n_h''}\right)(p_{hd}s_{bxyhd} + p_{hd}q_{hd}\bar{x}_{hd}\bar{y}_{hd}) + \frac{1}{n_h'''n_h''}\sum_{i}^{n_{hd}'''}\left(\frac{1}{m_{hid}'''} - \frac{1}{m_{hid}''}\right)s_{xyhid}\right],$$

$$c_{2xxd} = \frac{1}{H^2}\sum_{h=1}^{H}\left[\left(\frac{1}{n_h'''} - \frac{1}{n_h''}\right)(p_{hd}s^2_{bxhd} + p_{hd}q_{hd}\bar{x}^2_{hd}) + \frac{1}{n_h'''n_h''}\sum_{i}^{n_{hd}'''}\left(\frac{1}{m_{hid}'''} - \frac{1}{m_{hid}''}\right)s^2_{xhid}\right],$$

$$s^2_{bxhd} = \frac{1}{n_{hd}''' - 1}\sum_{i=1}^{n_{hd}'''}(\bar{x}_{hid} - \bar{x}_{hd})^2,\; s_{bxyhd} = \frac{1}{n_{hd}''' - 1}\sum_{i=1}^{n_{hd}'''}(\bar{x}_{hid} - \bar{x}_{hd})(\bar{y}_{hid} - \bar{y}_{hd}),$$

$$s_{xhid}^{2} = \frac{1}{m_{hid}''' - 1} \sum_{j=1}^{m_{hid}'''} \left( x_{hijd} - \bar{x}_{hid} \right)^{2},$$

$$s_{xyhid} = \frac{1}{m_{hid}''' - 1} \sum_{j=1}^{m_{hid}'''} \left( x_{hijd} - \bar{x}_{hid} \right)\left( y_{hijd} - \bar{y}_{hid} \right),$$

$$p_{hd} = \frac{n_{hd}'''}{n_{h}'''}, \quad q_{hd} = 1 - p_{hd}, \; p_{hid} = \frac{m_{hid}'''}{m_{hi}'''}, \quad q_{hid} = 1 - p_{hid},$$

$$\bar{x}_{hid} = \frac{1}{m_{hid}'''} \sum_{j}^{m_{hid}'''} x_{hijd}, \quad \bar{y}_{hid} = \frac{1}{m_{hid}'''} \sum_{j}^{m_{hid}'''} y_{hijd},$$

$$\bar{x}_{hd} = \frac{1}{n_{hd}'''} \sum_{i=}^{n_{hd}'''} \bar{x}_{hid}, \; \bar{y}_{hd} = \frac{1}{n_{hd}'''} \sum_{i=1}^{n_{hd}'''} \bar{y}_{hid}.$$

An approximate estimate of the variance and percentage standard error of the proposed linear regression estimator of average yield under a crop under $d^{\text{th}}$ mixture, $\bar{y}_{2lrd}$, is given by

$$\hat{V}\left(\bar{y}_{2lrd}\right) = \frac{1}{H^{2}} \sum_{h=1}^{H} \left[ \left\{ \left( \frac{1}{n_{h}'''} - \frac{1}{N_{h}} \right) \left( p_{hd} s_{byhd}^{2} + p_{hd} q_{hd} \bar{y}_{hd}^{2} \right) + \frac{1}{n_{h}''' N_{h}} \sum_{i=1}^{n_{hd}'''} \left( \frac{1}{m_{hid}'''} - \frac{1}{M_{hid}} \right) s_{yhid}^{2} \right\} \right.$$
$$\left. - r_{2d}^{2} \left\{ \left( \frac{1}{n_{h}'''} - \frac{1}{n_{h}''} \right) \left( p_{hd} s_{byhd}^{2} + p_{hd} q_{hd} \bar{y}_{hd}^{2} \right) + \frac{1}{n_{h}''' n_{h}''} \sum_{i=1}^{n_{hd}'''} \left( \frac{1}{m_{hid}'''} - \frac{1}{m_{hid}''} \right) s_{yhid}^{2} \right\} \right], \quad \ldots(56)$$

and

$$\% \widehat{SE}(\bar{y}_{2lrd}) = \frac{\sqrt{\hat{V}\left(\bar{y}_{2lrd}\right)}}{\bar{y}_{2lrd}} \times 100. \quad \ldots(57)$$

where

$$s_{byhd}^{2} = \frac{1}{n_{hd}''' - 1} \sum_{i=1}^{n_{hd}'''} \left( \bar{y}_{hid} - \bar{y}_{hd} \right)^{2}, \; s_{yhid}^{2} = \frac{1}{m_{hid}''' - 1} \sum_{j=1}^{m_{hid}'''} \left( y_{hijd} - \bar{y}_{hid} \right)^{2},$$

$$r_{2d}^{2} = \frac{c_{2xyd}^{2}}{c_{2xxd}\, c_{2yyd}},$$

and $c_{2yyd}$ is in the same functional form as $c_{2xxd}$ , as defined earlier.

The estimator of yield for a specific crop *c* based on all the domains at the district level along with its approximate estimator of the variance is given by

$$\hat{\bar{Y}}_{c2} = \frac{1}{D^*}\sum_{d=1}^{D^*}\bar{y}_{2lrd} \text{ and} \qquad ...(58)$$

$$\hat{V}\left(\hat{\bar{Y}}_{c2}\right) = \frac{1}{D^{*2}}\sum_{d=1}^{D^*}\hat{V}\left(\bar{y}_{2lrd}\right). \qquad ...(59)$$

where the sum is over the domains containing a particular crop *c* in the different mixtures, *d*=1,2,...,*D**.

Then, the estimator of percentage standard error of the proposed estimator of population total, $\hat{\bar{Y}}_{c2}$, is given by

$$\% \widehat{SE}\left(\hat{\bar{Y}}_{c2}\right) = \frac{\sqrt{\hat{V}\left(\hat{\bar{Y}}_{c2}\right)}}{\hat{\bar{Y}}_{c2}} \times 100. \qquad ...(60)$$

## ANNEX C

**Category 3- Household approach**

**Estimation procedure for crop area under mixed cropping using the household approach**

Let, in a district, there are $H$ sub districts, which can be considered as $H$ strata. Let, $N_h$ villages/enumeration areas/census block be in the $h^{th}$ sub district (stratum), which are considered as the PSUs, $h$=1,…,H. In general, the total number of villages/enumeration areas/census blocks in each sub district, $N_h$, are known. Suppose, in each village/enumeration area, there are $M_{hi}$ number of households (SSU).Let j$^{th}$ household in the $i^{th}$ village/enumeration area/census block has $T_{hij}$ parcels (USU) where different forms of crop mixture, such as a pure stand, mixture-1 and mixture-2, are grown.

Consequently, the different crop mixtures are taken as domains of the study. It is assumed that in each $h^{th}$ sub district, $D$ different crop mixtures are being followed as, for example, pure stand, mixture-1 and, mixture-2.Here there would be $\{U_{h1}, \ldots, U_{hd}, \ldots, U_{hD}\}$ domains in the $h^{th}$ stratum, $U_h$. Although, the number of villages/enumeration areas/census blocks ($N_h$), number of households ($M_{hi}$) and number of parcels ($T_{hij}$) in each sub district are known, for a specific $d^{th}$ form of mixture, the number of villages/enumeration areas/census blocks ($N_{hd}$), number of households ($M_{hid}$) and number of parcels ($T_{hijd}$) are generally unknown.

Let $y_{hijk}$ be the crop area of $k^{th}$ parcel (USU) in the $j^{th}$ household (SSU) of the $i^{th}$ villages/enumeration areas/census blocks (PSU) in the $h^{th}$ sub district (stratum) of a district.

The total area under $d^{th}$ crop mixture (domain)in a district is given by

$$Y_d = \sum_{h=1}^{H} \sum_{i=1}^{N_{hd}} \sum_{j=1}^{M_{hid}} \sum_{k=1}^{T_{hijd}} y_{hijk}, \qquad \text{...(61)}$$

The population total based on the domains is given as

$$Y = \sum_{d=1}^{D} Y_d = \sum_{d=1}^{D} \sum_{h=1}^{H} \sum_{i=1}^{N_{hd}} \sum_{j=1}^{M_{hid}} \sum_{k=1}^{T_{hijd}} y_{hijk}.$$

The sampling design for the estimation of crop area at the district level using the household approach is a stratified two-stage cluster sampling design with two phases in each of the stages of sampling. Suppose, in the first phase, a PPSWR sample of size $n_h'$ is drawn from $N_h$ villages/enumeration areas/census blocks (PSU). The probability of selecting $i^{th}$ village/enumeration area/census block in the $h^{th}$ stratum is taken $z'_{hi} = X_{hi}/X_h$, where $X$ may be taken as the total agricultural household. In the second stage of first phase, $m_{hi}'$ households are selected by SRSWOR design and the $T_{hij}$ parcels in selected household are completely enumerated for collecting auxiliary information regarding the parcel, such as seed used and farmers' assessment. Let, $n_{hd}'$ villages out of the selected $n_h'$ villages are following the specific $d^{\text{th}}$ mixture. In the same way, let, $d^{\text{th}}$ mixture be grown by $m_{hid}'$ selected households in their $T_{hijd}$ total parcels. Again in the second phase, a sample of $n_h$ villages is selected from $n_h'$ initially selected villages/enumeration areas (PSU) by SRSWOR. In each of those selected villages, a subsample of $m_{hi}$ households are selected by SRSWOR design and the $T_{hij}$ parcels in the selected household are completely enumerated. The areas of each of the parcels growing crop mixtures of those sampled households are measured by GPS or by the polygon method using a chain and compass or farmers' enquiry. Data collection should be made using the questionnaires given in annex C of the field test protocol document. Then, the area for component crops are obtained by apportioning using a fixed ratio approach based on physical observations. Suppose, out of the parcels growing crop mixtures by $m_{hi}$ households of $n_h$ villages, $n_{hd}$ villages are following a specific $d^{\text{th}}$ mixture in $T_{hijd}$ parcels of $m_{hid}$ sampled households.

In the current scenario, the objective is to estimate the total crop area under a specific crop($Y$) as well as under different mixtures($Y_d$), $d$=1,2,...,$D$.

Let $x_{hijk}$ be the auxiliary information, such as seed used and farmers' assessment, corresponding to $k^{\text{th}}$ parcel (SSU) of the $j^{th}$ selected household in the $i^{\text{th}}$ village (PSU) in the $h^{\text{th}}$ sub district (stratum).

Whereas, $y_{hijk}$ is the crop area measured by GPS and the rope and compass method of $k^{\text{th}}$ parcel (USU) in the $j^{\text{th}}$ household (SSU) of the $i^{\text{th}}$ villages/EA/CB (PSU) in the $h^{\text{th}}$ sub district (stratum) of a district.

A linear regression estimator of the total area under a $d^{\text{th}}$ mixture under the stratified two-phase two-stage cluster sampling design can be formed as

$$\hat{Y}_{lr3d} = \hat{Y}_d + b_{A3d}(\hat{X}_d' - \hat{X}_d) \quad ...(62)$$

where,

$$\hat{Y}_d = \sum_{h=1}^{H} \frac{1}{n_h} \sum_{i=1}^{n_{hd}} \frac{1}{z_{hi}^{'}} \frac{M_{hi}}{m_{hi}} \sum_{j=1}^{m_{hid}} \sum_{k=1}^{T_{hijd}} y_{hijk} = \sum_{h=1}^{H} \frac{1}{n_h} \sum_{i=1}^{n_{hd}} \frac{1}{z_{hi}^{'}} \frac{M_{hi}}{m_{hi}} \sum_{j=1}^{m_{hid}} y_{hij.} = \sum_{h=1}^{H} \frac{1}{n_h} \sum_{i=1}^{n_{hd}} \frac{\hat{Y}_{hid}}{z_{hi}^{'}},$$

$$\hat{X}_d = \sum_{h=1}^{H} \frac{1}{n_h} \sum_{i=1}^{n_{hd}} \frac{\hat{X}_{hid}}{z_{hi}^{'}}, \quad \hat{X}_d^{'} = \sum_{h=1}^{H} \frac{1}{n_h^{'}} \sum_{i=1}^{n_{hd}} \frac{\hat{X}_{hid}^{'}}{z_{hi}^{'}},$$

$$\hat{Y}_{hid} = \frac{M_{hi}}{m_{hi}} \sum_{j=1}^{m_{hid}} \sum_{k=1}^{T_{hijd}} y_{hijk}, \ \hat{X}_{hid} = \frac{M_{hi}}{m_{hi}} \sum_{j=1}^{m_{hid}} \sum_{k=1}^{T_{hijd}} x_{hijk}, \ \hat{X}_{hid}^{'} = \frac{M_{hi}}{m_{hi}^{'}} \sum_{j=1}^{m_{hid}^{'}} \sum_{k=1}^{T_{hijd}} x_{hijk}.$$

By minimizing the variance of the linear regression estimator $\hat{Y}_{lr3d}$ with respect to $b_{A3d}$ and ignoring the variation in $b_{A3d}$, the value of $b_{A3d}$can be shown as

$$b_{A3d} = c_{A3xyd} / c_{A3xxd}, \quad ...(63)$$

where,

$$c_{A3xyd} = \sum_{h=1}^{H} \left( \frac{1}{n_h} - \frac{1}{n_h^{'}} \right) \left( p_{hd}\, s_{bxyhd} + p_{hd}\, q_{hd}\, \hat{X}_{hd} \hat{Y}_{hd} \right),$$

$$c_{A3xxd} = \sum_{h=1}^{H} \left( \frac{1}{n_h} - \frac{1}{n_h^{'}} \right) \left( p_{hd}\, s_{bxhd}^2 + p_{hd}\, q_{hd}\, \hat{X}_{hd}^2 \right),$$

$$p_{hd} = \frac{n_{hd}}{n_h}, \quad q_{hd} = 1 - p_{hd},$$

$$s_{bxhd}^2 = \frac{1}{n_{hd} - 1} \sum_{i=1}^{n_{hd}} \left( \frac{M_{hi}\, p_{hid}\, \overline{x}_{hid}}{z_{hi}^{'}} - \hat{X}_{hd} \right)^2,$$

$$s_{bxyhd} = \frac{1}{n_{hd} - 1} \sum_{i=1}^{n_{hd}} \left( \frac{M_{hi}\, p_{hid}\, \overline{x}_{hid}}{z_{hi}^{'}} - \hat{X}_{hd} \right) \left( \frac{M_{hi}\, p_{hid}\, \overline{y}_{hid}}{z_{hi}^{'}} - \hat{Y}_{hd} \right),$$

$$\overline{x}_{hid} = \frac{1}{m_{hid}} \sum_{j=1}^{m_{hid}} x_{hij.d}, \quad \overline{y}_{hid} = \frac{1}{m_{hid}} \sum_{j}^{m_{hid}} y_{hij.d},$$

$$\hat{X}_{hd} = \frac{1}{n_{hd}} \sum_{i=}^{n_{hd}} \frac{M_{hi}\, p_{hid}\, \overline{x}_{hid}}{z_{hi}^{'}}, \quad \overline{y}_{hd} = \frac{1}{n_{hd}} \sum_{i=1}^{n_{hd}} \frac{M_{hi}\, p_{hid}\, \overline{y}_{hid}}{z_{hi}^{'}}.$$

An approximate estimate of variance of the linear regression estimator $\hat{Y}_{lr3d}$ is given by

$$\hat{V}\left(\hat{Y}_{lr2d}\right) = \sum_{h=1}^{H} \left( \frac{1}{n_h} - r_{A2d}^2 \left( \frac{1}{n_h} - \frac{1}{n_h^{'}} \right) \right) \left( p_{hd}\, s_{byhd}^2 + p_{hd}\, q_{hd}\, \hat{Y}_{hd}^2 \right), \quad ...(64)$$

where

$$s_{byhd}^2 = \frac{1}{n_{hd}-1}\sum_{i=1}^{n_{hd}}\left(\frac{M_{hi}\,p_{hid}\,\overline{y}_{hid}}{z_{hi}^{'}} - \hat{Y}_{hd}\right)^2, \quad r_{A3d}^2 = \frac{c_{A3xyd}^2}{c_{A3xxd}\,c_{A3yyd}},$$

and $c_{A3yyd}$ is in the same functional form as $c_{A3xxd}$ as defined earlier.

Then, the estimator of percentage standard error of the proposed linear regression estimator of total area under $d^{\text{th}}$ mixture, $\hat{Y}_{lr3d}$, is given by

$$\%\,\widehat{SE}(\hat{Y}_{lr3d}) = \frac{\sqrt{\hat{V}\left(\hat{Y}_{lr3d}\right)}}{\hat{Y}_{lr3d}} \times 100. \qquad ...(65)$$

The estimator of total area for a specific crop *c* based on all the domains at district level is given by

$$\hat{Y}_{c3} = \sum_{d=1}^{D^*}\hat{Y}_{lr3d}\,. \qquad ...(66)$$

where the sum is over the domains containing a particular crop $c$,$d$=1,2,...,$D^*$.

An approximate estimate of variance of the linear regression estimator $\hat{Y}_{c3}$ is given by

$$\hat{V}\left(\hat{Y}_{c3}\right) = \sum_{d=1}^{D^*}\hat{V}(\hat{Y}_{lr3d}) \qquad ...(67)$$

Then, the estimator of percentage standard error of the proposed estimator of population total, $\hat{Y}_{c3}$, is given by

$$\%\,\widehat{SE}(\hat{Y}_{c3}) = \frac{\sqrt{\hat{V}\left(\hat{Y}_{c3}\right)}}{\hat{Y}_{c3}} \times 100. \qquad ...(68)$$

**Estimation of crop yield for mixed cropping using a stratified two-stage two-phase sampling design framework.**

In a situation in which information on parcels in the selected villages/enumeration area is available in the records, it is advisable to adopt the **household approach** for crop yield estimation. Under this approach, the sample selected for crop area estimation should be taken as the sampling frame. For this purpose, first, a list of parcels growing different mixtures is prepared using the results from area enumeration in each of the sampled villages/enumeration areas in a sub district.

Suppose, for "estimation of crop area" by the household approach under the stratified two- phase two-stage cluster sampling design framework, in $h^{th}$ sub district (stratum) from the set of $N_h$ villages/EAs/CBs (PSUs), a sample of $n_h$ villages were selected. Then, a subsample of households was selected in each selected village/enumeration area/census block and the parcels of selected household were completely enumerated for the area under a specific crop mixture. While surveying the parcels for area enumeration, a list of parcels growing different mixtures is prepared in a village/enumeration area/census block. Let, $m_{hid}$ parcels are being grown as the $d^{th}$ mixture of the crop out of $m_{hi}$ sampled parcels in the $n_{hd}$ villages/enumeration areas/census blocks of the $h^{th}$ sub districts, $\forall\ d = 1, 2, ..., D$.

In the first phase of sampling, from the set of $n_h$ villages/enumeration areas/census block (PSU) $n_h^{''}$ villages are selected by SRSWOR design. Within the selected villages, for each of the $d^{th}$ mixture, $d = 1, 2, ..., D$, samples of $m_{hid}^{''}$ parcels (SSUs) are to be selected by SRSWOR from the $m_{hid}$ parcels used as the recording area. From the selected parcels, farmers' eye estimates of harvested yield of the crop under mixture have to be recorded.

Let, $x_{hijd}$ defines the eye-estimated harvested yield of the crop in the $j^{th}$ parcel grown in the $d^{th}$ mixture in the $i^{th}$ village of the $h^{th}$ sub district.

Again, in the second phase of sampling, let a sample of $n_h^{'''}$ villages/enumeration areas/census blocks be selected by SRSWOR design from the $n_h^{''}$ selected villages and within each selected villages/enumeration areas/census blocks, for each mixture, $d = 1, 2, ....D$, sample of $m_{hid}^{'''}$ parcels be selected from the $m_{hid}^{''}$ first phase second stage units in the $i^{th}$ village of the $h^{th}$ sub district.

In the final sample, it can be observed that out of $n_h'''$ sample villages/enumeration areas/census blocks, $n_{hd}'''$ villages/enumeration areas/census blocks are following the $d^{th}$ mixture, $\forall\, d = 1, 2, ....D$. Now, from all those $m_{hid}'''$ parcels in the final sample, estimates of the harvested yields of a crop under a specific mixture can be obtained by conducting a crop cutting experiment for mixed cropping. See section 4.2 of Technical Report I for the detailed procedure for conducting a crop cutting experiment. Data collection on farmers' eye estimates of crop yields should be carried out using questionnaire-6(C-3) in annex C of the field test protocol document and for data collection on crop yield, the crop cutting questionnaires given in the same annex should be used.

Let, $y_{hijd}$ defines the harvested yield of the crop under $d^{\text{th}}$ mixture from the $j^{\text{th}}$ parcel in the $i^{\text{th}}$ village/enumeration area of the $h^{\text{th}}$ sub district.

Using the information from the eye-estimated yields and the harvested yield of the sampled parcels under a specific mixture (domain) and using a double sampling approach, a linear regression estimator of the average crop yield under a specific crop mixture for a district can be given by

$$\bar{y}_{lr3d} = \bar{y}_{3d} + b_{3d}(\bar{x}'_{3d} - \bar{x}_{3d}) \qquad \ldots(69)$$

where, $$\bar{y}_{3d} = \frac{1}{H}\sum_{h=1}^{H}\frac{1}{n_h'''}\sum_{i=1}^{n_{hd}'''}\frac{1}{m_{hid}'''}\sum_{j=1}^{m_{hid}'''} y_{hijd},$$

$$\bar{x}_{3d} = \frac{1}{H}\sum_{h=1}^{H}\frac{1}{n_h'''}\sum_{i=1}^{n_{hd}'''}\frac{1}{m_{hid}'''}\sum_{j=1}^{m_{hid}'''} x_{hijd},$$

$$\bar{x}'_{3d} = \frac{1}{H}\sum_{h=1}^{H}\frac{1}{n_h''}\sum_{i=1}^{n_{hd}''}\frac{1}{m_{hid}''}\sum_{j=1}^{m_{hid}''} x_{hijd},$$

$$b_{3d} = \frac{c_{3xyd}}{c_{3xxd}}$$

$$c_{3xyd} = \frac{1}{H^2}\sum_{h=1}^{H}\left[\left(\frac{1}{n_h'''} - \frac{1}{n_h''}\right)(p_{hd}s_{bxyhd} + p_{hd}q_{hd}\bar{x}_{hd}\bar{y}_{hd}) + \frac{1}{n_h'''n_h''}\sum_{i}^{n_{hd}'''}\left(\frac{1}{m_{hid}'''} - \frac{1}{m_{hid}''}\right)s_{xyhid}\right],$$

$$c_{3xxd} = \frac{1}{H^2}\sum_{h=1}^{H}\left[\left(\frac{1}{n_h^{'''}} - \frac{1}{n_h^{''}}\right)(p_{hd}s_{bxhd}^2 + p_{hd}q_{hd}\overline{x}_{hd}^2) + \frac{1}{n_h^{'''}n_h^{''}}\sum_{i}^{n_{hd}^{'''}}\left(\frac{1}{m_{hid}^{'''}} - \frac{1}{m_{hid}^{''}}\right)s_{xhid}^2\right],$$

$$s_{bxhd}^2 = \frac{1}{n_{hd}^{'''} - 1}\sum_{i=1}^{n_{hd}^{'''}}(\overline{x}_{hid} - \overline{x}_{hd})^2,\ s_{bxyhd} = \frac{1}{n_{hd}^{'''} - 1}\sum_{i=1}^{n_{hd}^{'''}}(\overline{x}_{hid} - \overline{x}_{hd})(\overline{y}_{hid} - \overline{y}_{hd}),$$

$$s_{xhid}^2 = \frac{1}{m_{hid}^{'''} - 1}\sum_{j=1}^{m_{hid}^{'''}}\left(x_{hijd} - \overline{x}_{hid}\right)^2,$$

$$s_{xyhid} = \frac{1}{m_{hid}^{'''} - 1}\sum_{j=1}^{m_{hid}^{'''}}\left(x_{hijd} - \overline{x}_{hid}\right)\left(y_{hijd} - \overline{y}_{hid}\right),$$

$$p_{hd} = \frac{n_{hd}^{'''}}{n_h^{'''}},\quad q_{hd} = 1 - p_{hd},\ p_{hid} = \frac{m_{hid}^{'''}}{m_{hi}^{'''}},\quad q_{hid} = 1 - p_{hid},$$

$$\overline{x}_{hid} = \frac{1}{m_{hid}^{'''}}\sum_{j}^{m_{hid}^{'''}}x_{hijd},\quad \overline{y}_{hid} = \frac{1}{m_{hid}^{'''}}\sum_{j}^{m_{hid}^{'''}}y_{hijd},$$

$$\overline{x}_{hd} = \frac{1}{n_{hd}^{'''}}\sum_{i=}^{n_{hd}^{'''}}\overline{x}_{hid},\ \overline{y}_{hd} = \frac{1}{n_{hd}^{'''}}\sum_{i=1}^{n_{hd}^{'''}}\overline{y}_{hid}.$$

An approximate estimate of variance and percentage standard error of the proposed linear regression estimator of average yield under a crop under $d^{\text{th}}$ mixture, $\overline{y}_{lr3d}$, is given by

$$\begin{aligned}\hat{V}\left(\overline{y}_{lr3d}\right) = \frac{1}{H^2}\sum_{h=1}^{H}\Bigg[&\left\{\left(\frac{1}{n_h^{'''}} - \frac{1}{N_h}\right)\left(p_{hd}s_{byhd}^2 + p_{hd}q_{hd}\overline{y}_{hd}^2\right) + \frac{1}{n_h^{'''}N_h}\sum_{i=1}^{n_{hd}^{'''}}\left(\frac{1}{m_{hid}^{'''}} - \frac{1}{M_{hid}}\right)s_{yhid}^2\right\}\\ &- r_{3d}^2\left\{\left(\frac{1}{n_h^{'''}} - \frac{1}{n_h^{''}}\right)\left(p_{hd}s_{byhd}^2 + p_{hd}q_{hd}\overline{y}_{hd}^2\right) + \frac{1}{n_h^{'''}n_h^{''}}\sum_{i=1}^{n_{hd}^{'''}}\left(\frac{1}{m_{hid}^{'''}} - \frac{1}{m_{hid}^{''}}\right)s_{yhid}^2\right\}\Bigg],\end{aligned} \quad \ldots(70)$$

and

$$\% \widehat{SE}(\overline{y}_{lr3d}) = \frac{\sqrt{\hat{V}\left(\overline{y}_{lr3d}\right)}}{\overline{y}_{lr3d}} \times 100. \quad \ldots(71)$$

where

$$s_{byhd}^2 = \frac{1}{n_{hd}^{'''}-1}\sum_{i=1}^{n_{hd}^{'''}}\left(\bar{y}_{hid}-\bar{y}_{hd}\right)^2,\ s_{yhid}^2 = \frac{1}{m_{hid}^{'''}-1}\sum_{j=1}^{m_{hid}^{'''}}\left(y_{hijd}-\bar{y}_{hid}\right)^2,$$

$$r_{3d}^2 = \frac{c_{3xyd}^2}{c_{3xxd}\,c_{3yyd}},$$

and $c_{3yyd}$ is in same functional form as $c_{3xxd}$ as defined earlier.

The estimator of yield for a specific crop *c* based on all the domains at district level along with its approximate estimator of the variance is given by

$$\hat{\bar{Y}}_{c3} = \frac{1}{D^*}\sum_{d=1}^{D^*}\bar{y}_{lr3d} \qquad ...(72)$$

and

$$\hat{V}\left(\hat{\bar{Y}}_{c3}\right) = \frac{1}{D^{*2}}\sum_{d=1}^{D^*}\hat{V}\left(\bar{y}_{lr3d}\right). \qquad ...(73)$$

where the sum is over all those domains containing a particular crop *c* in the different mixtures, $d$=1,2,...,$D^*$.

Then, the estimator of percentage standard error of the proposed estimator of population total, $\hat{\bar{Y}}_{c3}$, is given by

$$\%\,\widehat{SE}\left(\hat{\bar{Y}}_{c3}\right) = \frac{\sqrt{\hat{V}\left(\hat{\bar{Y}}_{c3}\right)}}{\hat{\bar{Y}}_{c3}}\times 100. \qquad ...(74)$$

# References

**Bellow, M. & Lahiri, P.**2010. Empirical Bayes methodology for the NASS county estimation program. Proceedings of the section on Survey Research Methods.

**Das S.K. & Singh R.** 2013. A multiple-frame approach to crop yield estimation from satellite- remotely sensed data. *International Journal of Remote Sensing*, 34(11), 3803-3819.

**Food and Agriculture Organization of the United Nations (FAO)**.2010. Global strategy to improve agricultural and rural statistics. Report No. 56719-GLB.

**Kelly, V., Hopkins, J., Reardon, T. & Crawford, E. 1995**. Improving the measurement and analysis of African agricultural productivity: Promoting complementarities between micro and macro data. International Development Paper No. 16, East Lansing, Michigan, USA, Michigan State University.

**Särndal, C.E., Swensson, B., & Wretman, J.**1992. *Model Assisted Survey Sampling.* New York:Springer-Verlagk.

**Zhao, J., Shi, K. & Wei, F**. 2007. Research and application of remote sensing techniques in Chinese agricultural statistics. Paper presented at the *Fourth International Conference on Agricultural Statistic*s, 22-24 October 2007, Beijing.