D3.2 B

Soils4Africa Sampling Design

Project No. 869200

29 May 2021

Deliverable: Soils4Africa_D3.2B_Sampling_Design_v01

Version 1.0

Contact details

Director of Coordinating Institute – ISRIC: Rik van den Bosch

Project Coordinator: Mary Steverink-Mosugu

Address: Droevendaalsesteeg 3, 6708 PB Wageningen (Building 101), The Netherlands

Postal: PO Box 353, 6700 AJ Wageningen, The Netherlands

Phone: +31 317 48 7634

Email: mary.steverink-mosugu@isric.org

Project details

Project number	862900
Project acronym	Soils4Africa
Project name	Soil Information System for Africa
Starting date	01/06/2020
Duration In months	48
Call (part) identifier	H2020-SFS-2019-2
Topic	SFS-35-2019-2020 Sustainable Intensification in Africa

Document details

Work Package	3
Deliverable number	D3.2B Soils4Africa Sampling design
Version	1
Filename	*Soils4Africa_D3.2B_Sampling_design_v01*
Type of deliverable	Report
Dissemination level	Public
Lead partner	ISRIC
Contributing partners	JRC, ICRAF, IITA, SZIU, ARC
Authors	Bas Kempen, Dick J. Brus, Luís de Sousa
Contributors	Arwyn Jones, Keith Shepherd, Johan Leenaars, Erika Micheli, Adam Csorba, Jeroen Huising, Garry Patterson
Due date	31 May 2021
Submission date	29 May 2021

European Commission | Horizon 2020 European Union funding for Research & Innovation

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

SOILS 4 AFRICA

D3.2 B Table of Contents

European Commission | Horizon 2020 European Union funding for Research & Innovation

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

SOILS 4 AFRICA

European Commission | Horizon 2020 European Union funding for Research & Innovation

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

SOILS 4 AFRICA

Abbreviations

AEZ	Agro-ecological Zones
AfSIS	Africa Soil Information Service
CEC	Cation Exchange Capacity
DEM	Digital Elevation Model
EU	European Union
LDSF	Land Degradation Surveillance Framework
LEAP4NFSSA	Long-term Europe-Africa Research and Innovation Partnership for Food and Nutrition Security and Sustainable Agriculture
LUCAS	Land Use and Coverage Area Frame Survey (EC)
FNSSA	Food, Nutrition, Security and Sustainable Agriculture
GSP	Global Soil Partnership
PPS	Probability Proportional to Size
PSU	Primary Sampling Unit
RSG	Reference Soil Group
SIS	Soil Information System
SOC	Soil Organic Carbon
SSA	Sub-Saharan Africa
SSU	Secondary Sampling Unit
TSU	Tertiary Sampling Unit
WRB	World Reference Base for soil resources (International soil classification system)

European Commission | Horizon 2020 European Union funding for Research & Innovation

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

SOILS 4 AFRICA

Glossary

This section presents an overview of terms frequently used in sampling literature for readers less familiar with these. We use the same terminology here in the report. The definitions are derived from (De Gruijter et al., 2006).

aliquot:	material taken from a sampling unit.
composite:	aliquots bulked together.
domain (of interest):	specification of the part(s) of the sampling universe for which a separate result is required.
monitoring:	collecting information on an object through repeated or continuous observation in order to determine possible changes in the object.
sampling location:	position of a sampling unit in space.
sample size:	number of selected sampling units.
sample support:	geometry of the sampling units in a continuous universe.
sampling frame:	representation of the sampling universe. This can be a map in form of a GIS layer that represents the area to be sampled and from which locations are selected.
sampling pattern:	positions of all sampling units of a sample selected from the sampling frame following a design type.
sampling unit:	single part of the universe that can be selected for sampling.
sampling universe:	precise definition of the universe of interest with boundaries in space (and/or time), possibly a specification of exclusions. Also referred to as 'target universe'.
survey:	collecting information on an object with a spatial extent through observation.
target parameter:	type of statistic for which a result is needed given the target variable and domain(s).
target variable:	precise definition of the variable(s) to be determined for each sampling unit.

European Commission | Horizon 2020 European Union funding for Research & Innovation

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

SOILS 4 AFRICA

1. Introduction

The Soils4Africa project aims to build a soil information system (SIS) that will provide data at trans-continental level that is relevant for a variety of stakeholders involved in sustainable intensification of agricultural production in Africa. These users not only include individuals working in policy support or decision making, private sector, extension services, and (inter)national research institutes, but also international partnerships such as the Global Soil Partnership and inter-governmental platforms such as LEAP4FNSSA, research programs on sustainable intensification such the EU SFS-2019-2020-35 Scope A, and digital agricultural service providers such as the EU H2020 project AfriCultuReS. (Fatunbi and Abishek, 2020) provide an extensive overview of Soils4Africa stakeholders.

An extensive, continent-wide soil dataset should underpin the SIS, as well as the products and services provided by it, This dataset will contain field observations and analytical data on a wide variety of soil properties and soil quality indicators (Moinet et al., 2021) to be collected from about 20,000 sampling locations spread across the agricultural lands of Africa. This report describes the design of the Soils4Africa sampling scheme on basis of which the soil aliquots will be collected in the field. We note that `sampling design' defines the approach for selecting sampling locations, it does not provide guidelines on how soil aliquots are taken in the field. The latter is defined by sampling protocols that will be considered elsewhere (Soils4Africa project deliverable D3.5).

The type of sampling scheme is determined by the aim of the data collection. So before one starts designing a sampling scheme one should determine what information is needed: different information needs typically ask for different sampling schemes. De Gruijter et al. (2006) divide information needs for survey and monitoring into two broad groups. In the first group, the aim is to obtain estimates of *global quantities*, i.e. parameters of the cumulative distribution function of the target variable(s) for the entire sampling universe (the population). These parameters include the mean, median or other quantiles of the distribution (e.g. 5%, 95%) for instance. In the second group, the aim is to obtain some kind of description of the *spatial (and/or temporal) variation* of the target variable(s) *within* the sampling universe. An example is the estimation of quantities for sub-areas of the sampling universe, referred to as *local quantities*. At the extreme the aim is prediction of values at points in the universe, for instance, at the nodes of a discretization grid that covers the universe as is typically done for (digital) soil mapping.

Summarizing, for the first group the aim is to quantify *global* quantities in space (and/or time) and for the second group *local* quantities in space (and/or time). Thus, a first decision when designing a sampling scheme is on the type of quantities that are required: are global quantities required or local quantities, or perhaps both. When local quantities are required, then one should decide for which sub-areas within the sampling universe separate results are required. Are these relatively large, e.g. different land cover classes within a country, or (very) small, e.g. agricultural fields or even points. When the primary interest is in global quantities and local quantities for large sub-areas, the information need typically pertains to estimates of means, totals or areal fractions of the target variables and the main aim of the survey is thus *estimation*. For local quantities for (very) small sub-areas, the main aim of the survey is *mapping* (prediction).

From the information need follows the choice of a sampling approach that is defined as a combination of a mode for sampling unit selection and a mode of statistical inference from sample data. Here we can distinguish two approaches that are fundamentally different: design-based and model-based approaches for sampling and statistical inference. Once a choice is made for a sampling approach, an appropriate design type can be chosen that defines how sampling units will be selected from the sampling universe. Finally, a sampling algorithm is used to select a pre-

European Commission | Horizon 2020 European Union funding for Research & Innovation

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

SOILS4AFRICA

defined number of sampling units given the design type. But it all starts with a clearly defined purpose. In case there are several objectives, a compromise needs to be achieved that can best meet all the objectives.

The main purpose of the Soils4Africa project is to *collect, develop and share **baseline information** on the **prevalence and spatial distribution** of **soil quality indicators and constraints** relevant for sustainable intensification of agriculture in Africa*. In addition, it should lay the foundation of a ***soil monitoring network*** that allows changes in soil conditions and soil quality over time to be quantified, for instance as a result of soil management interventions to permit evaluation of the impact of interventions aiming at sustainable intensification.

A secondary purpose is to develop digital (gridded) soil maps of soil properties and soil quality indicators for the agricultural land in Africa from the collected data.

Chapter 2 is a methodological chapter that further elaborates on sampling design concepts that were already briefly touched upon in this chapter, and major decisions one needs to make when designing a sampling scheme. In chapter 3 we provide a brief review of the LUCAS and AfSIS sampling designs. Chapter 4 translates the purpose of the project to an information need and then describes the design of the Soils4Africa sampling scheme with which this information need can be met. Statistical sampling approaches for monitoring are addressed in Chapter 5.

European Commission | Horizon 2020 European Union funding for Research & Innovation

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

SOILS4AFRICA

2. Designing sampling schemes for spatial survey

In this methodological chapter we provide an overview of some fundamental decisions that need to be made when designing a sampling scheme for spatial survey. These pertain to the basic methods for selecting sampling units and statistical inference and how these relate to the purpose of the sampling campaign and information need (global versus local quantities). This choice of sampling approach defines the basic framework of the sampling scheme and guides further design decisions such as the choice of a design type. In addition, we address three more topics that are relevant in the context of the Soils4Africa project. These are *stratification*, *sampling of sub-areas* and *use of ancillary data*.

Selection and inference methods

De Gruijter et al. (2006) distinguish three possible modes for selecting sampling units: **convenience sampling**, **purposive sampling** and **probability sampling** that we briefly summarize here:

- *Convenience sampling*: sampling units are selected based on convenience such as for instance easily accessible spots such as road sides. The advantage is that it is resource efficient. The disadvantage is that the statistical properties are inferior compared to other selection methods.
- *Purposive sampling*: sampling units are selected in such a way that a given purpose is best served. The *free survey* method traditionally used for soil mapping is an example of purposive sampling. Sampling locations are selected in such a way, subjectively and based on experience, that these are most informative for delineation of soil classes. Another example of purposive sampling is *feature space coverage sampling* that can be used for selecting sampling locations for *digital soil mapping* (Ma et al., 2020; Wadoux et al., 2019). The purpose of this design is to cover feature (covariate) space as good as possible to allow for efficient calibration of prediction models. For geostatistical surveys, sampling units can also be selected purposively, for instance, to minimize prediction uncertainty as quantified by the prediction error variance (kriging variance). The advantage is optimal selection for a given purpose. The disadvantage is that the variance of estimators for means and totals of the study area as a whole and of sub-areas cannot be quantified without making modelling assumptions.
- *Probability sampling*: random selection of sampling units according to a well-defined sampling design so that the probabilities of selecting the sampling units, the **inclusion probabilities**, are **known**; these probabilities need to be **larger than zero** for all population units but **need not to be equal**. The advantage of probability sampling is that it enables model-free, design-based statistical inference, i.e. no modelling assumptions need to be made in estimation of the means (totals) of the study area and of sub-areas. On the other hand probability samples can also be used in model-based statistical inference, for instance to predict the means of small sub-areas from which no data are collected, or for mapping through prediction at the nodes of a fine discretization grid. Hereafter we will give more details about design-based and model-based inference. So probability sampling is more flexible than non-probability sampling (convenience sampling, purposive sampling) as it enables both design-based and model-based inference, where non-probability samples can only be used for model-based inference. The main (operational) disadvantage is that it puts stringent requirements on field work. There is no free choice in deciding which sampling locations are visited. Selected sampling locations must be visited even when these are at inconvenient locations. If a sampling location cannot be reached or when

European Commission | Horizon 2020 European Union funding for Research & Innovation

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

SOILS4AFRICA

permission is denied by the land owner, then the location must either be treated as non-response, or be replaced by a back-up location, selected beforehand. That location could be at considerable distance from the current location. Taking a sample at a convenient location, for instance in a neighboring field, is not allowed.

Besides selection method, one needs to choose a mode for statistical inference from the sample data. De Gruijter et al. (2006) distinguish two that are fundamentally different: **design-based inference** and **model-based inference**. In the design-based approach, inference is based on the inclusion probabilities of the sampling units. Hence, design-based inference requires probability sampling. The inclusion probabilities are known for all selected units. and are used as weights in the statistical inference (estimation). Unbiased point-estimators of the quantities of interest are available, as well as valid confidence intervals of these quantities. The simplest random selection method is Simple Random Sampling. With this sampling design all possible samples of size *n* have the same probability of being selected As a consequence all locations have the same inclusion probability and the sample mean of the observed (or measured) values at the sampling locations is an unbiased estimator of the mean of the target properties

Model-based predictions of the values at points or the means of sub-areas also are weighted averages of the sample data but in this case weights are determined by a stochastic model of the variation in the sampling universe and the sample pattern. Such a model is for instance a model with a constant mean or a mean that is a linear combination of covariates, and spatially correlated residuals used in geostatistics. In that case, the weights of the data are computed from the covariances between the observation points and the covariances between the observation points and a prediction point, that are a function of the spatial coordinates. For model-based inference, purposive sampling is most appropriate, but probability samples can also be used in model-based inference. See Brus (2021) for a more elaborate explanation of design-based and model-based methods and some persistent misconceptions about these.

De Gruijter et al. (2006) define a sampling method or approach as a combination of a method for selecting of sampling units and a method for statistical inference. The design-based approach combines probability sampling with design-based inference and the model-based method combines purposive sampling with model-based inference. The choice of a sampling approach is guided by the purpose of the data collection and the information need. As we have seen in Chapter 1, the information need can be categorized in two broad groups: **global quantities** in space (and or time) and **local quantities** in space (and or time).

Global versus local quantities

De Gruijter et al. (2006) define '*global quantities*' as the spatial cumulative distribution function of the target variable(s) in the sampling universe, or quantities that can be derived of this function such as the mean, median or other quantiles, standard deviation. Global quantities can be defined in space, time or space-time, though here we limit ourselves to global quantities in space for now. Examples of global quantities in space are the total organic carbon stock of the agricultural lands in Africa, the mean topsoil zinc content in a farming system, or the areal fraction of a country that is affected by soil erosion.

Besides interest in the sampling universe as a whole, interest may also be in sub-areas of the sampling universe that are referred to as **domains** (of interest) for which separate estimates are required. De Gruijter et al. (2006) refer to these domain-specific quantities as '*local quantities*'. For instance, for a soil organic carbon stock assessment in a country (the sampling universe) we might not only be interested in the total stock of that country, but we might also be interested to obtain the stocks for each land cover class that occurs in a country. The land cover classes are in this case

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

the domains; domains of interest are also referred to as *reporting units*, i.e., the unit for which we wish to report separate results. Sub-areas can be large such as land cover classes or farming systems, but can also be small or very small. Think for instance about all agricultural fields in the sampling universe.

For both quantities there are design-based as well as model-based sampling approaches available. De Gruijter et al. (2006) provide an elaborate overview of these in Chapters 7 and 8. The suitability of the design-based and model-based methods for spatial survey depends amongst others on the spatial resolution for which estimates are required with at one end of the spatial resolution continuum the entire sampling universe and at the other end of this continuum a single point in the universe. Broadly speaking, as De Gruijter et al. (2006) explain in Chapter 6, design-based methods are most appropriate when estimates of target quantities are required for the entire universe or large domains within the universe. When domains become smaller and more numerous (in a sense that these cannot all be sampled individually anymore, hence estimation for each individual small-area domain is not possible or becomes impractical) model-based methods that rely on prediction, such as geostatistical methods, become more suitable.

Besides the resolution of the result there are more factors that influence the choice for a design (Brus and De Gruijter, 1997; De Gruijter et al., 2006):

- should the estimation of global quantities of the target population be unbiased? Design-based method guarantee against bias in sampling that can arise in convenience or purposive sampling.
- should the accuracy of the estimates be quantified objectively (in the form of a standard error or confidence interval), i.e. without having to make any (model-)assumptions about the spatial variation? If objective assessment of accuracy of the estimates is key, then design-based methods are preferred.
- is random sampling practically feasible? If this is not, then purposive sampling with model-based inference should be considered.
- is a reliable model of the spatial variation available? Only if this is the case then model-based methods can be considered.
- is there substantial spatial autocorrelation of? Only if this is the case then model-based methods can be considered.

Summarizing, when the aim of the spatial survey is *estimation* of global quantities (means and totals) or local quantities for (large) domains of interest then design-based methods based on probability sampling and design-based inference are preferred. If the aim of the survey is *prediction* (mapping) for points, then model-based methods based on purposive (non-probability) sampling and model-based inference are preferred. For estimating quantities of numerous (small) sub-areas both approaches are in principle suitable. An interesting alternative for the latter could also be the *model-assisted* approach that we further describe in section 'Use of ancillary data' hereafter. We further discuss the choice for a basic sample selection and statistical inference method for the Soils4Africa sampling given the overall objective of the project in Chapter 4.

Besides *sampling for estimating means and totals* and *sampling for mapping*, we can distinguish *sampling for monitoring* that includes besides a spatial a temporal dimension. Monitoring soil conditions is not part of the current project. However, because the Soils4Africa sampling design aims to provide a basis for future monitoring, we elaborate on general sampling approaches for monitoring in Chapter 5. Once a choice is made for the sampling approach, a design type must be chosen.

European Commission | Horizon 2020 European Union funding for Research & Innovation

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

SOILS 4 AFRICA

Design types

Once a choice is made for a sampling approach (design-based or model-based), one needs to select a design type. The design type defines how sampling locations are selected. De Gruijter et al. (2006) provide an elaborate overview of a great variety of design types for design- and model-based methods for global quantities in space (Chapter 7) and design- and model-based methods for local quantities in space (Chapter 8), as does Brus (*in prep*). Furthermore, Brus (2019) presents and illustrates with R code various non-probability sampling approaches for mapping. We, therefore, do not repeat such elaborate overview here and limit ourselves to a few examples including those that are relevant for Soils4Africa.

Sampling for estimating means and totals

When the aim of the spatial survey is to estimate means and totals for the sampling universe and domains of interest within that universe then design-based methods based on probability sampling are preferred. There are probability sampling design types that we highlight here: simple random sampling, stratified simple random sampling and two-stage cluster random sampling following De Gruijter et al. (2006). Figure 1 shows notional examples of these three designs.

Simple random sampling. This is the most basic design type. There are no restrictions put on the selection of sampling units besides that the sample size is fixed. All sampling units are selected with equal probability and independently from each other. The advantage of this design is that the implementation is simple as is the statistical inference. Furthermore, because of its simplicity the design is very flexible and sample sizes can be easily adapted. Disadvantages include its efficiency: the sampling variance is usually larger than that of other design types at the same costs. This is because i) the spatial coverage may be poor (especially when sample sizes are small) and ii) visiting the sampling locations may be more time-consuming for logistical reasons because of the irregular distribution across the sampling area, resulting in higher-per sample costs and thus smaller sample size given a fixed budget. A larger sampling variance results in wider confidence intervals of the target quantities. Another important disadvantage is that estimation of local quantities for domains of interest may not be possible since one cannot control the sample size in each domain. It is therefore possible that there are domains without any samples.

Stratified simple random sampling. With stratified simple random sampling the sampling universe is partitioned into sub-areas, called strata. Within every stratum, a fixed and pre-determined number of sampling units are selected with simple random sampling. The number of sampling units per stratum should sum to the total sample size that is also pre-determined. Stratified simple random sampling typically results in a smaller sampling variance (i.e. more accurate estimates) for a given sample size compared to simple random sampling when the strata are chosen in such a way that the sum of squares between strata is larger than the weighted sum of sum of squares within strata (Lohr, 1999). Stratification provides important controls on the selection on sampling units and is often used for spatial surveys and therefore merits more detailed attention in this report. This is provided in the next section of this chapter.

Two-stage cluster random sampling. Like with stratified random sampling, the sampling universe is partitioned in sub-areas. However, the partitioning is not based on the reporting units, neither on expected differences in the mean of the study variable among the sub-areas. The sub-areas are compact polygons serving the selection of spatial clusters of sampling units. The difference with stratified sampling is that not all sub-areas are sampled, but that sampling is restricted to a limited set of randomly selected sub-areas. The selected sub-areas are referred to as *primary sampling units* (PSUs). Within each primary unit a pre-determined number of sampling units is selected that are then visited. These units are referred to as the *secondary sampling units* (SSUs). The most simple selection technique for both primary and secondary units is simple random sampling. Other techniques, however, could also be considered. One could select PUs with stratified random

European Commission | Horizon 2020 European Union funding for Research & Innovation

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

SOILS4AFRICA

sampling and SUs with systematic random sampling to give just an example, resulting in compound design types. A two-stage cluster design can be generalized to multistage cluster random sampling following the same selection principles.

The advantage of two-stage cluster random sampling is its operational efficiency through the clustering of sampling sites within PUs that greatly reduces travel time between locations in the field with a large spatial extent of the study area, as in Soils4Africa. A disadvantage is that spatial clustering of sampling units generally leads to a lower precision of the estimates of the target quantities.

For simple random sampling only one design attribute needs to be chosen (sample size) and for stratified simple random sampling three (definition of the strata, total sample size and allocation of sample sizes to the strata). Two-stage cluster random sampling adds additional complexity to the design. Five design attributes need to be chosen: definition of the PUs, selection mode of the PUs (with or without replacement), selection probabilities of the PUs (equal or proportional to size), the number of PUs to be selected in the first stage and the number of (secondary) sampling units in the PUs.

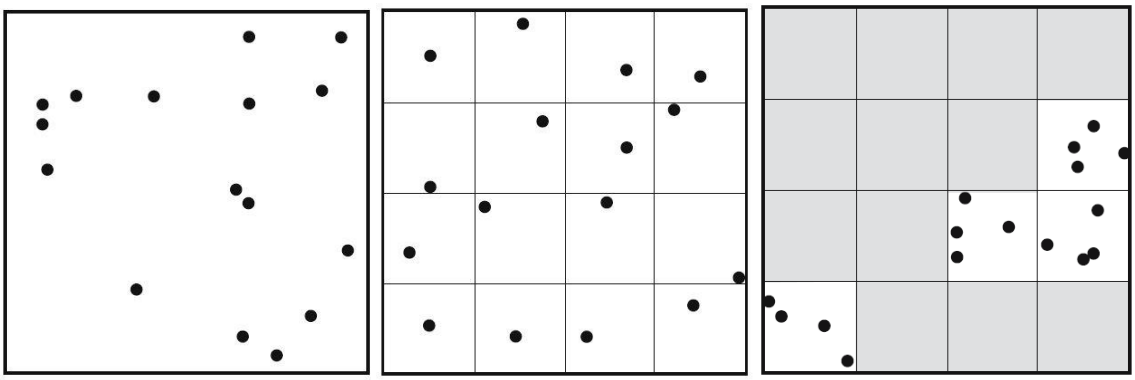

Figure 1. National examples of sampling patterns based on simple random sampling (left), stratified random sampling (middle), two-stage cluster random sampling (right). Adapted from De Gruijter et al. (2006).

Sampling for mapping

When the main purpose of a spatial survey is mapping, i.e. target properties need to be quantified for numerous small domains that cannot be sampled individually, think for instance of agricultural fields or when the aim is to predict the study variable at the nodes of a fine grid, a model-based approach is most advantageous. Probability sampling then is not necessary and could even be suboptimal. For mapping it could be advantageous to have good spatial coverage across the sampling universe or to have good coverage of the covariate space in case ancillary data are used in the statistical models used for mapping such as in regression-kriging (Kempen et al., 2019) or machine learning (Hengl et al., 2017). Design types frequently used for mapping are spatial coverage sampling and conditioned Latin hypercube sampling (cLHS); Figure 2 shows examples. But more design types exist that are elaborately described by De Gruijter et al. (2006) and Brus (2019). Ma et al. (2020) and Wadoux et al. (2019) compare the efficiency of several designs for mapping.

Spatial coverage sampling. This design optimizes sample allocation across geographic space in such a way that the sampling universe is covered as regularly as possible. This can be achieved by clustering prediction locations into geographic sub-areas that are as compact as possible with the number of clusters equal to the sample size. The sampling locations are then formed by the centroids of the geographic clusters. For clustering, algorithms such as k-means can be used.

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

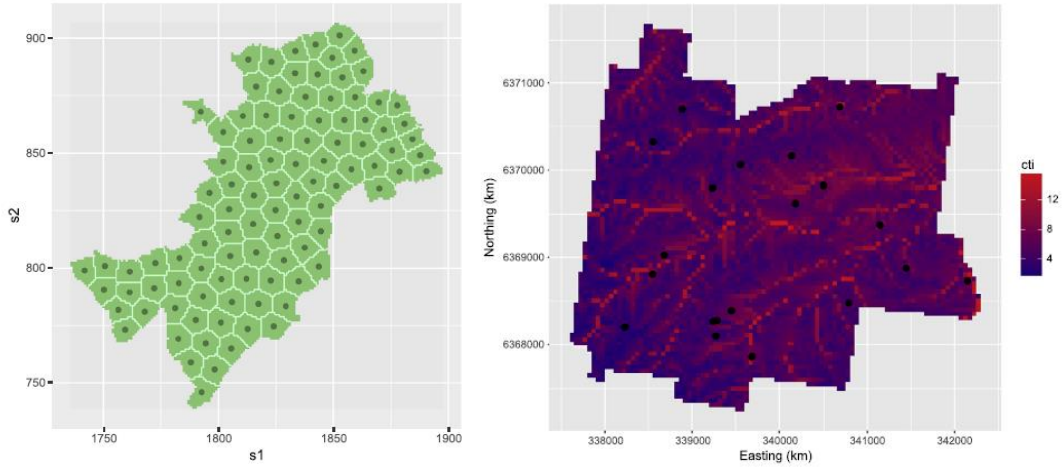

Figure 2. Examples of a spatial coverage (left) and cLHS (right) design. Reproduced from Brus (2019).

Conditioned Latin hypercube sampling. Samples are selected in such a way that the marginal frequency distributions of the quantitative environmental features in the sampling universe (as represented by maps) are reproduced by the sample as closely as possible (Minasny and McBratney, 2006). To this end for each feature, marginal strata are constructed using equally spaced quantiles of the marginal distribution, so that all marginal strata contain an equal number of pixels. Optimization algorithms are subsequently used to select samples across all marginal strata. Though frequently applied for (digital) soil mapping, Ma et al. (2020) and Wadoux et al. (2019) show that in their case studies feature space coverage sampling is more efficient.

Stratification

Rationale and basic principles

Some form of stratification of the sampling universe is often used when designing a sampling scheme. There are two main reasons for stratification (De Gruijter et al., 2006):

- to increase the *efficiency* of the sampling design. One obtains smaller sampling variance at the same costs (i.e. same number of sampling locations) compared to simple random sampling. This means that the estimates of the target quantities (means, totals, areal fractions) become more precise.
- to obtain separate estimates for sub-areas within the sampling universe, the so-called *domains of interest* for which we wish to report separate results. These could for instance be countries, farming systems, agro-ecological zones etc. The domains are then used as strata in the design.

An attractive property of stratification is that one can **control the sample size** of the strata. This is especially important when one is interested in domain estimates. If one does not use a domain as a stratum, then there is a chance that this domain will not be sampled. This is especially a problem when the total sample size is small (low sampling intensity) and when there are domains with a relatively small surface area.

For increasing efficiency of the sampling design, the stratification variables should be associated with the target soil properties, i.e. these should explain some of the spatial variation of these properties. Variation of target properties, expressed in sum of squares, within strata should be smaller than between strata. A soil class map could for instance be a good candidate for stratification when the aim is statistical soil survey since we expect that variation of key soil properties within soil classes is smaller than between soil classes. Using a soil map for stratification

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

could thus serve both purposes: to increase efficiency of the design for estimating global means, as well as to delineate sub-areas of which we wish separate estimates.

De Gruijter et al. (2006) caution that stratification should be done carefully. They explain that inappropriate stratification, that may occur when stratum means differ little, or when sample sizes of the strata that are strongly disproportional to the areas of the strata can result in a loss in efficiency (i.e. estimators become less precise). Disproportional sample sizes can occur if one has many small strata in combination with a relatively small total sample size. The minimum sample size of a stratum is two, to allow unbiased estimation of the sampling variance of the estimated population mean. This minimum sample size may further impair the proportional allocation of the sample size to the strata.

Stratification using cross-classification

It can very well be that one is interested in obtaining estimates of domains of interest that belong to different partitions of the sampling universe. One partitioning could for instance be on basis of soil classes using a soil map with *n* classes (domains), while another partitioning could be on basis of land cover classes using a land cover map with *m* classes (domains). Estimates are required for all domains of both partitions that are *n* + *m* in total.

The standard solution in case of multiple stratification variables is to define strata as *cross-classification classes*. Each stratum is then defined as a unique combination of classes of the stratifying variables. In the above example this will result in *n* x *m* cross-classes (the Cartesian product), though in reality often not all class combinations will exist.

The main drawback of a cross-classification approach is that the number of classes can quickly become very large. Two variables with 15 classes each already yield a maximum of 225 cross-classes, many of which could be small area strata. Having many cross-classes require large total sample sizes that cannot always be permitted. And with a minimum of two sampling locations per stratum there can be a large risk of disproportional allocation when total sample size is relatively small, as described in previous section.

Multi-way stratification

Survey literature provides an attractive alternative for cross-classification in the form of *multi-way stratification* (Falorsi and Righi, 2008). Multi-way stratification allows us to select planned sample sizes for domains of two or more partitions of the sample universe without making use of cross-class strata. We explored multi-way stratification for the Soils4Africa sampling design (see Chapter 4) but finally decided not to implement this form of stratification because estimation of the sampling variance of the estimated means of domains is rather complicated. Nevertheless, we believe it is worthwhile to explain the basic principles here because multi-way stratification is an elegant solution for dealing with multiple stratification variables.

A notional example of two-way stratification is shown in Figure 3. Here two maps are used for stratification. One map has four map units serving as domains of interest (Figure 3, left), the other three (Figure 3, right). The planned sample size for each of the domains of map A is six, and for those of the domains of map B, eight. The actual samples are selected (randomly) in such a way that the planned sample sizes for each of the domains is respected. Figure 3 shows the selected sampling units outlined in black. It also shows that not all cross-class strata are sampled in this case: cross-class strata A1 x B3 is not sampled in this example. From the selected sampling units we can estimate our target quantities for the entire sampling universe as well as for each of the seven individual domains.

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

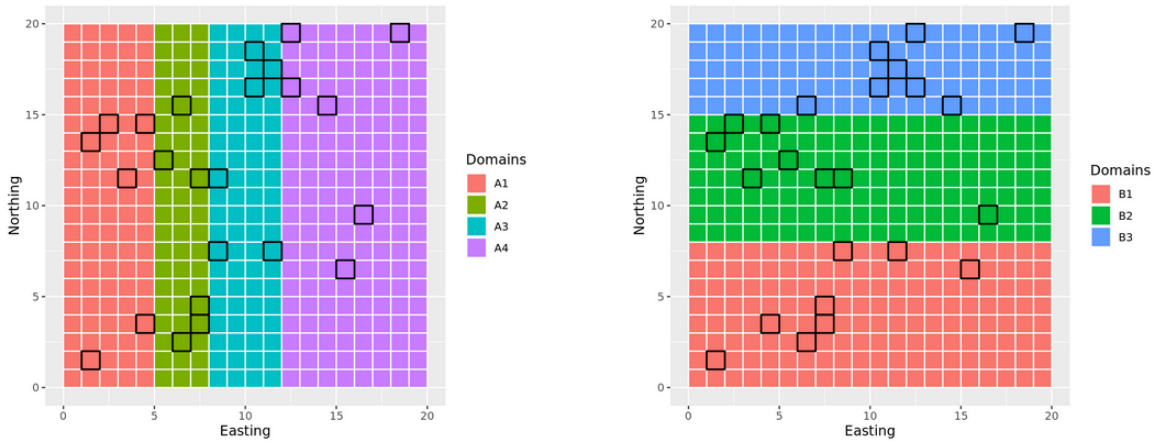

Figure 3. Notional example of a two-way stratified random sample. Reproduced from Brus (in prep).

Multi-way stratification is a constrained (multi-dimensional) optimization problem of the inclusion probabilities of the population units. The total sample size is given, as well as the planned sample sizes for each domain. In case all cross-classification strata are non-empty, i.e. contain at least one population unit as in Figure 3, the optimal inclusion probability of a population unit can be computed as the optimal sample size of the cross-classification stratum of that unit, divided by the total number of units in that cross-classification stratum. This results in equal inclusion probabilities for all units of a cross-classification stratum. The optimal sample size can also apportioned proportional to a covariate related to the soil property of interest, leading to variable inclusion probabilities. The optimal sample size of a cross-classification stratum can be computed as the product of the marginal sample sizes (sample sizes of the individual strata) divided by the total sample size. For instance, suppose that we need to select 100 sampling locations planned as follows: 75 in soil class S1 and 25 in soil class S2 and 50 in land cover class L1 and 50 in land cover class L2. Then the optimal sample sizes for the cross-strata S1xL1 and S1xL2 are 50 x 75 / 100 = 37.5 and for the cross-strata S2xL1 and S2xL2 25/27/100 = 12.5. These optimal sample sizes then must be divided by the number of population units in a cross-classification stratum to obtain the optimal inclusion probabilities. If one or more cross-classification strata are empty, optimization of the inclusion probabilities becomes more complicated.

Sampling unit selection with a multi-way stratification algorithm is accomplished in two main steps. First, the inclusion probabilities of the sampling units are optimized in such a way that these result in sample sizes that are as close as possible to the planned sample sizes of all domains of interest. Simulated annealing, a generic and widely-used optimization algorithm, can be used for this purpose. Second, a sample is selected based on the optimized inclusion probabilities. Sample selection can be done with a **balanced sampling** approach (Falorsi and Righi, 2008; Brus, 2015). Balanced sampling ensures that the samples size of each of the domains of interest are equal to the planned sample sizes.

Sampling of sub-areas and estimation of local quantities

Sampling of sub-areas, or *domains of interest*, of the sampling universe to be able to estimate or predict local quantities in space merits some further attention. As we have seen in the second section of this chapter, the size of the individual sub-areas guides the fundamental choice of a design-based versus a model-based approach to sampling and inference. For large sub-areas, for instance dominant land cover classes in a country, a design-based approach involving probability sampling is preferred (sampling for estimating means, totals and areal fractions), whereas for relatively small sub-areas (that cannot all be sampled), for instance all farmers' fields in a regional

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

model-based approach that relies on a statistical model of the spatial variation of a soil property is preferred (sampling for mapping). Here we focus on design-based estimation of means, totals and areal fractions for sub-areas, following De Gruijter et al. (2006; Section 8.2) who distinguish three situations.

The first is when the domains of interest are known at the beginning of the survey and one can afford a reasonable sample size for each domain. In this situation it is recommended to use the domains as strata and sample each domain separately. This has the great advantage that one can control the sample size in each domain and thus ensure that each domain is sampled.

The second situation is when a map depicting the domains of interest is not used for selecting sampling locations, i.e. for stratification but the number of sampling locations in the domains is large enough to obtain a reliable estimate of the means of the domains from the domain-specific data only. This is for instance the case when one wants to avoid the use of multiple stratification variables resulting in a (too) large number of cross-classification classes. In that case one will only select one map for stratification instead of two or more. It is a misconception that if a domain of interest is not used as a stratum for whatever reason, one cannot obtain (approximately) unbiased and valid estimates of the target quantities for that domain. De Gruijter et al. (2006) present various estimators for this situation. Note that the number of sampling units in a domain of interest not used as a stratum, is uncontrolled. The sample size can be very small or even zero. In case the sample size of a domain is sufficiently large, the mean of that domain can best be estimated by the *ratio estimator*. In this estimator of the domain mean the total of a domain is divided by the *estimated* area of that domain. This ratio estimator is also recommended in case the area of a domain is known because it is more accurate than the usual (Horvitz-Thompson) estimator. Further note that the true area of a land cover type or farming system is actually unknown even when a map of that land cover type is available. A map only gives the mapped area based on a model that contains impurities. If the actual land cover type is recorded for each Soils4Africa sampling location then the mean soil organic carbon content for the observed land cover types can be estimated with the ratio-estimator.

The third situation deals with estimation for small-area domains that are not well represented in the sample. Even in this case, if quantitative and/or qualitative covariates are available it is still possible to estimate the mean of these domains by the generalized regression estimator (De Gruijter et al, 2006; Brus, *in prep.*).

Summarizing, preferably one should use domains of interest as strata. In this way one can control the sample size in each domain. When this is not feasible or impractical, it is still possible to obtain estimates for these domains, even when these are not well represented in the sample.

Use of ancillary data

Ancillary data, in form of GIS layers of environmental properties, can be a great asset when designing a sampling scheme for spatial survey in the design-based approach. Such data cannot only be used in the sample selection stage but also in the estimation stage.

Selection stage

An obvious use of ancillary data in the selection stage is for stratification, as we have seen in the previous sections. Stratification, if done properly, can improve performance of the sampling strategy.

Another use of ancillary data that we have not discussed before is for spreading sampling units in covariate (feature) space. This can be done with a technique referred to as the **local pivotal method** (Grafström et al., 2012; Brus, 2021). With this method we can ensure that a probability

European Commission | Horizon 2020 European Union funding for Research & Innovation

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

SOILS4AFRICA

sample is well spread in the space spanned by the scaled covariates. Think, for instance, of selecting a probability sample covering the space spanned by relevant environmental variables, for instance elevation and slope, and/or climatic gradients, for instance mean annual precipitation and mean annual temperature.

Estimation stage

Besides at the selection stage, ancillary data can also be used a posteriori, after the data are collected, at the estimation stage. The great advantage of using ancillary data that are correlated with the target variables for estimation is that it can increase the accuracy of the estimates. This type of statistical inference is referred to as **model-assisted** inference (Särndal el al, 1992; Brus 2021). Model-assisted inference requires probability sampling and yields approximately unbiased and valid estimates of target quantities.

Model-assisted estimators are for instance the post-stratification estimator which exploits a qualitative (categorical) ancillary variable, and the regression estimator which exploits one or more quantitative ancillary covariates (Brus, 2000, De Gruijter et al. 2006). Newer developments include use of statistical learning techniques, such as the random forest model, in model-assisted estimators (Brus, 2021).

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

3. The LUCAS and AfSIS-LDSF sampling designs

The call under which the Soils4Africa project was proposed asked for the development of a soil database with standard soil properties following a procedure similar to the one used for the EU LUCAS database. For this purpose we reviewed the sampling designs of LUCAS and LUCAS Topsoil to investigate if there are elements in these designs that we could use for a Soils4Africa sampling scheme. A second, Africa-specific, monitoring network that is of interest for this purpose and worth noting here is the Land Degradation Surveillance Framework (LDSF) that was developed in the first phase of the Africa Soil Information Service (AfSIS) project.

LUCAS and LUCAS Soil

Land Use/Cover Area frame Survey (LUCAS)

The Land Use/Cover Area frame Survey (LUCAS) is a network to monitor changes in land use and land cover across the member states of European Union (EU). The first survey was held in 2001, covering 13 of the 15 members of the EU at that time. Over time the LUCAS network expanded covering all 28 member states in 2018. The sampling design was adapted several times after the first survey with the latest major update in 2018 for the 2018 survey. Below we briefly describe the sampling design that was used from 2006 onwards.

The LUCAS sample is two-phase sample. In the first phase a square point grid is randomly selected with a spacing of 2 km covering the member states of the EU territory. This grid is the LUCAS 'master sample' (or sampling frame) and contains about 1.1 million geo-referenced points (EUROSTAT, 2018). In the second phase, a subset of the master is selected containing 25 to 30% of the master sampling units. For the 2018 survey, sample size was 330,000.

Up to 2018, selection of sampling locations in the second phase was done as follows. The points of the master grid are stratified according to seven land cover classes ('STR05'), as identified from aerial photos or satellite imagery. Prior to stratified sub-sampling of the master grid, the grid is divided into square blocks of 9 x 9 points to avoid clustering of sampling sites (Gallego and Delincé, 2010). The 81 points contained in a block are subsequently numbered in such a way that spatial clustering of successive numbers within a block is avoided. The numbering is achieved by implementation of the following algorithm (Brus et al., 2011):

1. Select first point simple randomly (this point receives number 1).
2. Select second point randomly from group of points at maximum distance from first point.
3. Select third until 81th point randomly from group of points that maximizes the minimum distance to points already selected.

In the above algorithm the distance between two points is not simply computed as the Euclidian distance. This would lead to a concentration of small numbers near the edge, and consequently to a concentration of sampling points near the edges of the squares. Instead, distances are defined in such a way as to avoid concentrations along the edges while at the same time ensuring numbering is fully random. The distance definition used is given by Brus et al. (2011). The authors also provide two examples of the random numbering within a block and show, through simulations, that the selection probability is equal for all 81 points in the block. The randomization of the numbers is done only once, meaning that the randomization in all blocks is identical. The set of points with the same number (i.e. position) in each block is referred to as replicate (Gallego and Delincé, 2010).

A stratified sub-sample is selected from the master grid by seven land cover classes and NUTS2 region. Strata are sampled with different densities. The strata 'arable land', 'permanent crops' and

European Commission | Horizon 2020 European Union funding for Research & Innovation

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

SOILS4AFRICA

'grassland' are sub-sampled more intensively than the other land cover types. Sampling units are selected by selecting complete sets of replicates, starting with replicate 1. This means that first all points that fall in stratum *h* with replicate number 1 are selected. Followed by all points in stratum *h* with replicate number 2. This is continued until the remainder of points to be selected to complete the sampling size for stratum *h*, n_h, is smaller than the number of points in the next replicate. The remainder is then selected simple randomly from this replicate.

For LUCAS 2018 the second phase sampling design was redesigned (Ballin et al., 2018). The 'Gallego' design to avoid spatial clustering was abandoned. The land cover stratification of 2005 was updated ('STR18') and in addition two new stratification variables were introduced: the Corine Land Cover class ('CLC') and elevation class ('ELEV'). From these three variables, cross-class (or *atomic*) strata were calculated as the Cartesian product of the three variables (STR18 x CLC x ELEV) for each NUTS2 region. Cross-classification of stratification variables can potentially result in a very large number of cross-classes (equal to the number of unique combinations of the classes of the stratification variables). This was also the case for the LUCAS cross-classification. In a next step, a sophisticated but highly complex optimization algorithm was used to group the atomic strata in such a way that the total sample sizes and allocation were compliant to a set of precision constraints. The optimization procedure is described in detail by Ballin et al. (2018). For LUCAS 2018, this new stratification approach resulted in 19,962 strata, with 55 (Luxembourg) to 2,172 (Spain) strata per country.

LUCAS Soil

The LUCAS Topsoil Survey adds a third phase to the sampling design. LUCAS Soil is a 10% sub-sample of LUCAS, resulting in about 20,000 sampling sites (Figure 4). Though the sub-sampling methodology is described in Tóth et al. (2013a), it remains unclear how this phase was exactly implemented. Four terrain attributes (altitude, slope, curvature and aspect, all derived from SRTM digital elevation model) and land cover (derived from Corine Land Cover 2000) were used for stratification of the LUCAS points. Each terrain attribute was classified into eight classes, each with an equal surface area. Strata were then formed by The four terrain variables and land cover variable were cross-classified, resulting in about 20,000 strata each a unique combination of classes of the five input variables. The gridded strata were converted to vector format giving 30,795 polygons. The polygon number was then associated to each LUCAS point and the number of points per polygon per land use calculated. Points were selected for sub-sampling if the number of points per land use type was larger than three. Next, the selected points per polygon were randomly grouped in triplets.

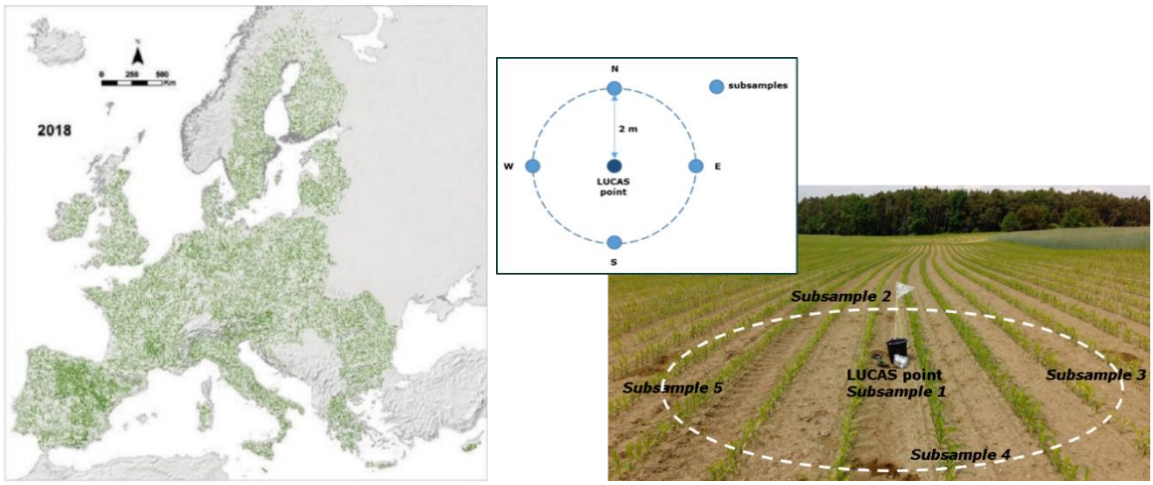

Figure 4. Sampling locations of the 2018 LUCAS Soil survey (Orgiazzi et al., 2018) (left); design of a sampling plot (right).

The LUCAS points within a polygon are used as primary sampling units in two-stage random sampling with the polygons serving as the primary sampling units and the triplets as secondary

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

sampling units. Sample sizes were set for combinations of land use and country and were proportional to the surface areas of the countries and main land use types within each country. From each triplet only one point is selected and sampled for observation. If the first point of a triplet is inaccessible, or for whatever reason unsuitable for soil sampling, another point is sampled et cetera. The soil surveyor is free to determine the order in which the points are visited in the field, which saves time needed for fieldwork, but compromises the computation of inclusion probabilities. If the number of triplets is smaller than the required sample size, more points are selected from the polygons containing more than six LUCAS points. Further details about how this is precisely done are not given by Tóth et al. (2013a). If the number of triplets exceeds the required sample size, those triplets are selected from the polygons of that specific land use that contain the largest number of LUCAS points. Note that the selection of primary sampling units is not random: the selection is targeted at the largest polygons. At each LUCAS Soil sampling site one aliquot is taken at the site centre and four aliquots 2 m in the cardinal directions from the site centre (Figure 4, right). The five soil aliquots are mixed to form a composite sample.

Reflections

LUCAS is a probability sampling design. The inclusion probabilities of the LUCAS points are perfectly known, and as a consequence unbiased design-based estimation of means and totals of reporting units is feasible. However, no unbiased estimator of the variance is available, due to the sampling design used in the first phase (systematic random sampling) and the second phase (cluster random sampling in a way that clusters are separated in geographical space). The variance can only be approximated. LUCAS 2018 redesign, however, is a (highly) complex design especially with respect to the stratification that results in a large number of strata. In the case of LUCAS, this large number of strata (in combination with a requirement of a minimum of two sampling units per stratum) is not an immediate problem as such, because of its large sample size but for smaller sample sizes, such as for Soils4Africa this could be the case. Besides, the complex implementation and statistical inference are making it a less suitable 'blueprint' for the Soils4Africa sampling design.

As for LUCAS Soil, based on the available information, we conclude that the LUCAS Soil is a stratified two-stage (non-random) subsample of LUCAS 2009. We conclude that the inclusion probabilities of the LUCAS Soil points are not traceable, and besides there are points with inclusion probability 0. For that simple reason, design-based statistical inference for the estimation of means ant totals for the sampling universe and parts thereof, qualified as superior by Tóth et al. (2013b) (of which we fully agree), is not feasible. LUCAS Soil therefore does not provide a suitable framework for a similar sampling design in Africa.

In their paper and report Tóth et al. (2013a, 2013b) compute sample means and sample standard deviations for land use/land cover categories within nine climate zones. For instance, mean and standard deviation of SOC concentration for annual croplands and for permanent croplands within each climate zones. The sample means are used as *estimates* of the subpopulation means. Differences in inclusion probabilities of the LUCAS points due to different sampling rates in the LUCAS strata, are not accounted for. In Fig. 6.38 of the report, showing mean C:N ratios for four land cover classes within four climate zones, also error bars are shown, representing the standard error of the estimated means. It is not explained how these standard errors were estimated. Possibly these are computed by the sample standard deviation divided by the square root of the sample size, which is quite a naive approach.

In summary, we conclude that design-based estimation of the means of domains of interest and their variances from LUCAS Soil is not feasible. The only option is model-based inference using, for instance, a geostatistical model (kriging). The estimated means and their estimated standard errors as reported by Tóth et al. (2013a, 2013b) are model-based estimates, that follow from the assumption of independent data (i.e. a pure nugget variogram). The question is how realistic this

European Commission | Horizon 2020 European Union funding for Research & Innovation

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

SOILS4AFRICA

assumption is. In any case such modelling assumptions can best be avoided. It is highly recommendable to select the sampling locations for monitoring the soil by a well-defined probability sampling design, so that means and totals of reporting units can be estimated either model-free by design-based inference, or by model-assisted inference.

AfSIS Land Degradation Surveillance Framework

The sampling design of the AfSIS Land Degradation Surveillance Framework (LDSF) was developed during the first phase of the AfSIS project. The design covers sixty 'sentinel' landscapes across Africa that are representative of the variation in climate, topography and vegetation of Sub-Saharan Africa (SSA) (Vågen et al., 2010).

A sentinel site is 10 x 10 km in size. Each site is divided into sixteen 2.5 x 2.5 km tiles. In each tile, the centroid of a sampling cluster is randomly located. The shape of a cluster is a circle with a 564 m radius from the centroid. This gives the cluster a surface area of 10 ha. Within a cluster, ten sampling plots are randomly selected. Each plot measures 1000 m^2 (0.1 ha) and is made up of four subplots that are 100 m^2 (0.01 ha) in size and configured in a radial arm layout (Figure 5) (Vågen et al., 2010; Vågen and Winowiecki, 2020). A topsoil sample (0-20 cm) is collected from each subplot and a subsoil sample from the center of each subplot. The topsoil aliquots are pooled in a composite sample.

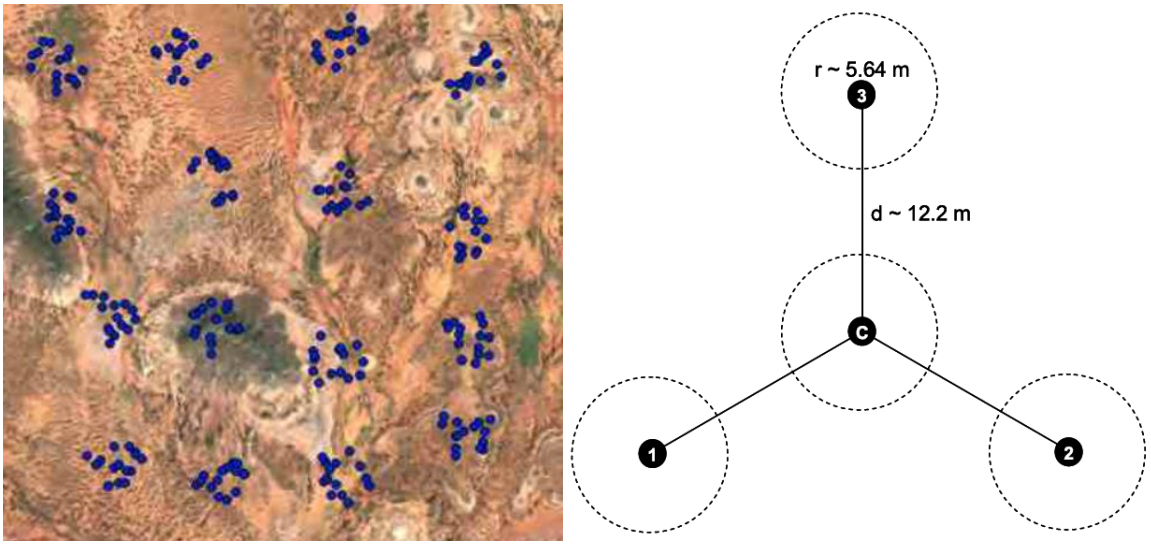

Figure 5. The AfSIS LDSF sampling design. A sentinel site composed of sixteen tiles with each containing a cluster with ten sampling plots (left); a sampling plot (right). From (Vågen et al., 2010).

The cluster centroid and the sampling plots are selected randomly, however, we conclude that the AfSIS LDSF sampling design is not a probability sampling design according to the definition given in section 2. This is because the inclusion (selection) probabilities of the sampling units (in this case subplots) are not known, i.e., cannot be calculated from the design. The design shows that we cannot assume equal inclusion probabilities. Inclusion probabilities of subplots will vary within a tile and will depend on where the cluster centroid will be selected. This is because the locations of possible sampling clusters are not fixed within a tile, i.e., the tile is not discretized into a fixed number of cluster of which one will be selected. Potential clusters, and thus sampling plots, can overlap in varying degrees with higher degrees of overlap in the center of the tile. Locations more towards the center or a tile have larger probability of being selected than locations closer to the tile boundary. This makes that it not possible to develop a sampling framework that lists all potential (discretized) sampling units with an inclusion probability for each of these. Design-based inference

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

is thus not possible. Note, however, that this is merely an observation and that by no means this should be read as criticism on the LDSF sampling design. Whether probability sampling and design-based inference are desirable depends on the purpose of the data collection. For Soils4Africa, we aim to design a sampling scheme based on probability sampling (see Chapter 4) and in that context, the LDSF design as well as LUCAS Soil are less suitable to serve as examples. Though there are some elements of both designs that we can recycle to some extent.

European Commission | Horizon 2020 European Union funding for Research & Innovation

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

SOILS4AFRICA

4. Soils4Africa sampling design

This chapter describes the proposed Soils4Africa sampling design. In Chapter 1 we formulated the overall aim of the Soils4Africa project as "*to collect, develop and share baseline information on the prevalence and spatial distribution of soil indicators and constraints relevant for sustainable intensification of agriculture in Africa"*. In addition, it should lay the foundation of a *soil monitoring network* that should allow to quantify changes in soil conditions and soil quality indicators over time, for instance as a result of soil management interventions. Based on this, we give the rationale for the choice for a sampling approach for sampling unit selection and statistical inference in the first section of this chapter. The second section gives an overview of the ancillary data we collected to support the design of the sampling scheme. The third section presents the sampling design, while the fourth section describes the refinement of the sampling universe considering disputed territories, fieldwork safety and protected areas. This chapter is concluded with a section that shows the implementation of the sampling design with an illustrated example for Kenya.

Sampling approach

The soils information collected in this project should provide a consistent baseline on the *current* status and spatial distribution of soil properties, indicators and constraints across agricultural lands in Africa in a consistent way. With 'consistent' we mean derived from a set of soil aliquots that are collected with uniform protocols, within one relatively short period of time (short in relation to the rate of soil change) and analyzed with one set of methods by one laboratory. The information will serve as a reference against which future observations can be compared to assess changes over time (monitoring).

One prerequisite of the sampling design is that it should provide a framework to allow future monitoring of soil conditions. When monitoring, one is typically not concerned about what happens over time at one particular sampling location because this may not be indicative about what happens with the soil in general (on average) within a larger area in which for instance an intervention took place. One could compare this with running a series of fertilizer field trials for a specific crop, say maize. The outcome of the trial at one individual trial plot is not providing much information about the effectiveness of the fertilizer treatment compared to a control treatment. The observed effect could be a chance result rather than a true effect because of specific (and variable) conditions prevailing at the plots of a trial site at the time of the trial. What we would like to know in this case is whether or not the new treatment produced higher yields compared to the control *on average* in the maize-growing area, thus considering all trial locations. That will give a much better indication about the effectiveness of the new treatment. The same applies here, an individual soil sampling location is not a very relevant monitoring unit. A change that we might observe at a single location may not tell us much about what is happening on average in an area. Monitoring units are typically larger geographic areas that contain multiple sampling locations from which we can infer changes over time. These areas could be countries, regions within countries, land management zones, farming systems etc., or even all agricultural land in the continent.

In Chapter 2 we distinguished two types of spatial survey that require different sampling approaches: *sampling for estimating means and totals* (i.e., assessing global quantities and local quantities for (large) area domains) and *sampling for mapping* (i.e., assessing local quantities for very small area domains and prediction at the nodes of a fine discretisation grid). Given the requirement of providing a statistical framework that allows future monitoring of soil conditions, the primary purpose of the Soils4Africa survey is to provide estimates of means and totals (and other relevant parameters that can be derived from the cumulative distribution function) of

European Commission | Horizon 2020 European Union funding for Research & Innovation

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

SOILS4AFRICA

selected soil properties and indicators for Africa's agricultural land. But apart from providing baseline soil information at continental level, the SIS should also be able to provide information for sub-areas of the sampling universe. Relevant domains of interest for which we might wish to report separate results could for instance include the individual countries to support national and sub-national level policy-making or national reporting requirements, or different agro-ecologies, farming systems or project intervention areas, in the continent or within countries.

The vast expanse of Africa's agricultural land (the sampling universe or *population*) with its great variety in soil conditions will be sampled at 'only' 20,000 locations. Because of the reference function of the information provided by the SIS, it is imperative that the estimates of the population parameters of the target variables are **unbiased**. This means that the values of these parameters are not systematically over- or underestimated. The sampling locations should thus be selected in such a way that the sampling means and totals give an unbiased assessment of the population means and totals.

Besides the unbiasedness requirement, it is important that we can objectively assess the accuracy of the estimated target parameters (e.g. means, totals, areal fractions). We infer target parameters for a set of target soil variables from only a limited number of soil aliquots taken across the sampling universe. We do not know the true value of the parameters, we only have estimates. Hence, we are uncertain about the parameters. Each time we repeat our sampling the estimates will be (slightly) different. We can quantify our uncertainty with the **sampling variance** that provides us with a measure of variation of the estimated target parameters if the sampling would be repeated using the same sampling design for selecting samples. From the sampling variance we can compute confidence intervals. The sampling variance depends on the variation of the target soil variables in the sampling universe, the sampling design and the sample size. The larger the sample size, the more precise our estimates of the target parameters.

The sampling variance is needed for statistical testing. For instance, when one wants to assess if the soil organic carbon content of farming system A is larger than that of farming system B. But also for monitoring it is key that we can quantify sampling variance. If we have measurements from two sampling times and we wish to test, with a defined level of statistical confidence, whether the changes we observe are real changes, i.e. changes in the population mean (total, areal fraction), , we need information about the sampling variances for both time periods. Furthermore, once the sampling variances and components of this variance (e.g., the contribution of the spatial variation of primary unit means to the sampling variance of the estimated population mean) are known after the first round of sampling, these can be used to optimize sample sizes for subsequent sampling rounds leading to more efficient designs. Because of the importance of the sampling variance, its estimation should be unbiased and objective. Ideally we do not want to rely on any (model) assumptions to quantify our uncertainty about the target parameters that might question the validity of the assessment.

Summarizing, primary purpose of the sampling is estimation of means and totals (and other relevant parameters that can be derived from the cumulative distribution function) of a set of target soil properties and soil quality indicators for the sampling universe and parts there-off. Furthermore, the estimates must be (approximately) unbiased and the sampling variance must be quantified objectively. In the **Soils4Africa sampling design** sampling units will be selected by **probability sampling** enabling both design-based estimation of global and local means (totals) and model-based mapping. Once the sample data are collected and suitable ancillary data available, model-assisted estimation of global and local means will be considered (see Chapter 2).

It is important to note here that a choice for a design-based method does not rule out use of the data for mapping. Digital soil mapping models could still be developed from the collected data and

European Commission | Horizon 2020 European Union funding for Research & Innovation

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

SOILS4AFRICA

maps of soil properties and indicators for the agricultural land area produced even though the sampling design is not optimized for this specific purpose.

Ancillary data

A set of ancillary data, with continent-wide coverage, was collected to support the design of the sampling scheme and the statistical inference (see Chapter 2). These datasets include:

- Administrative units level 0 (country) and level 1[1] (FAO, 2014).
- World Reference Base Reference Soil Groups (IUSS WG WRB, 2006; Jones et al., 2013)
- Major farming systems[2] (Dixon et al., 2001)
- Major farming systems level 1 and level 2 – update (Van Velthuizen et al., 2013; Dixon et al., 2019)
- Agro-ecological characteristics (Van Velthuizen et al., 2013)
- Land cover (Buchhorn, 2020)
- Agro-ecological zones (Sebastian, 2009)
- Length of growing season (Fischer et al., 2012; IIASA/FAO, 2012)
- Digital elevation model (DEM; dataset)
- Topographic wetness index (calculated from DEM)
- Accessibility (Nelson, 2008).
- Population density (CIESIN, 2018)
- Protected areas; 3 layers (UNEP-WCMC, 2021)
- Legacy soil point dataset (Batjes and Ribeiro, 2021)

Sampling design

Sampling universe

In space, the sampling universe is broadly defined as the 'agricultural land' in continental Africa (note that for sample selection we will refine this definition somewhat). The agricultural land was mapped by the Soils4Africa project using satellite imagery from the Copernicus Global Land Service. Details of the agricultural land mapping are provided by (Huising et al., 2020). Figure 6 shows a map of the agricultural land. This is a gridded (raster) map with a spatial resolution of 100 m x 100 m. Each of these 1 ha pixels represents a sampling unit and could be selected for sampling the soil. In total there are almost 800 million of these units meaning that the agricultural land area according to this classification is about 8 million km^2. We recognize here that the spatial extent of agricultural land will change with time and so the survey will not give information on areas that will be converted to agriculture in the future. The sampling universe, however, can be updated for future sampling rounds based on up-to-date information on land cover classification.

Clustering of sampling units

Like in the AfSIS-LDSF sampling design, we introduce some form of clustering of the 1ha sampling units. Clustering has great operational advantages in fieldwork. Once a location of a cluster is reached, multiple sampling sites can be visited. This greatly reduces travel time and hereby mitigates some of the risks inherent to a project that aims at continent-wide sampling within a relatively short amount of time. Besides, clustering sampling locations allows assessment of soil

[1] Outdated level 1 administrative boundaries were updated with administrative data from The Humanitarian Data Exchange (https://data.humdata.org/)

[2] To obtain a continental layer the farming systems maps of SSA and Nothern Africa were mosaicked and the legends harmonized.

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

variation at short distances that is to some extent controlled by different farm management practices. If we would not cluster, then this would not be possible given an average sampling density of one sampling locations per 400 km^2.

Clusters are formed by grouping the 1 ha sampling units. Here a cluster is defined as all 1 ha sampling units (100 m x 100 m pixels classified as agricultural land) inside a 2000 m x 2000 m square. All clusters are unique and do not overlap and each 1 ha agricultural land pixel only belongs to one cluster. In this way we can define a total of 3.95 million clusters that cover the sampling universe. Each cluster contains between 1 and 400 1 ha agricultural land pixels (Figure 7). The choice of the cluster size is somewhat arbitrary. It resembles the resolution of the LUCAS master grid and is close to the size of the AfSIS-LDSF sampling plots (2.5 x 2.5 km). Furthermore, the size is small enough that all sampling locations within a cluster are within walking distance which has operational advantages, and large enough that it captures variability at the landscape level.

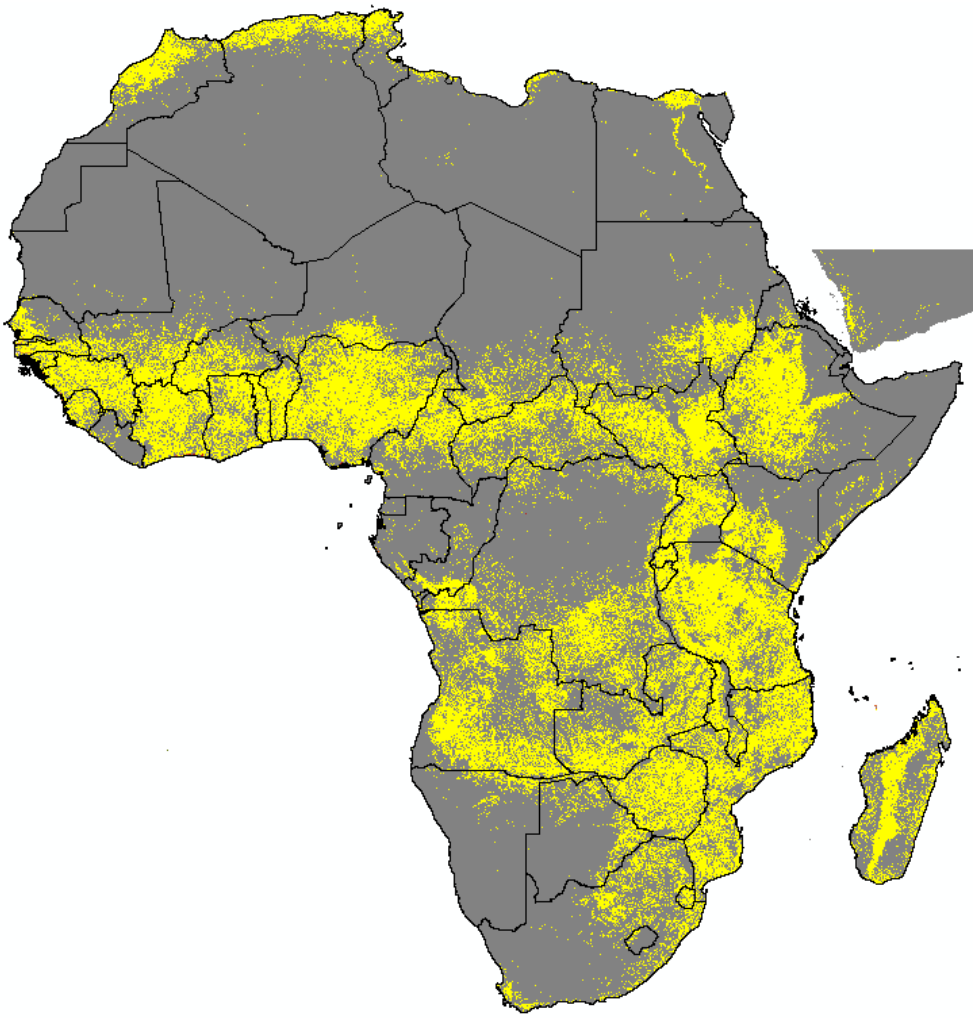

Figure 6.Map of agricultural land, with areas classified as 'agriculture' depicted in yellow (Huising et al, 2020)

Sampling plot

Like in LUCAS Soil and AfSIS-LDSF, composite samples will be taken. A Soils4Africa sampling plot measures 5 m x 5 m. In this sampling plot four soil aliquots are taken in the intercardinal directions from the plot center (i.e. the points where the soil aliquots are collected are at the centre of the four 2.5 m x 2.5 m sub-squares) that are combined in a composite sample. This means that the **sample support** is 5 m x 5 m. The measured values represent an average of the sampling

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

plot. Taking a composite sample has the advantage that some short-range variability which is irrelevant for our purpose is averaged out.

A 5 m x 5 m plot is located within a single 1 ha agricultural land pixel. A 1 ha agricultural land pixel is discretized into 400 non-overlapping 5 m x 5m sampling plots (configured as a 20 x 20 matrix). Of these 400 plots, one is randomly selected for collecting soil aliquots, with equal probabilities for all 400 sampling plots.

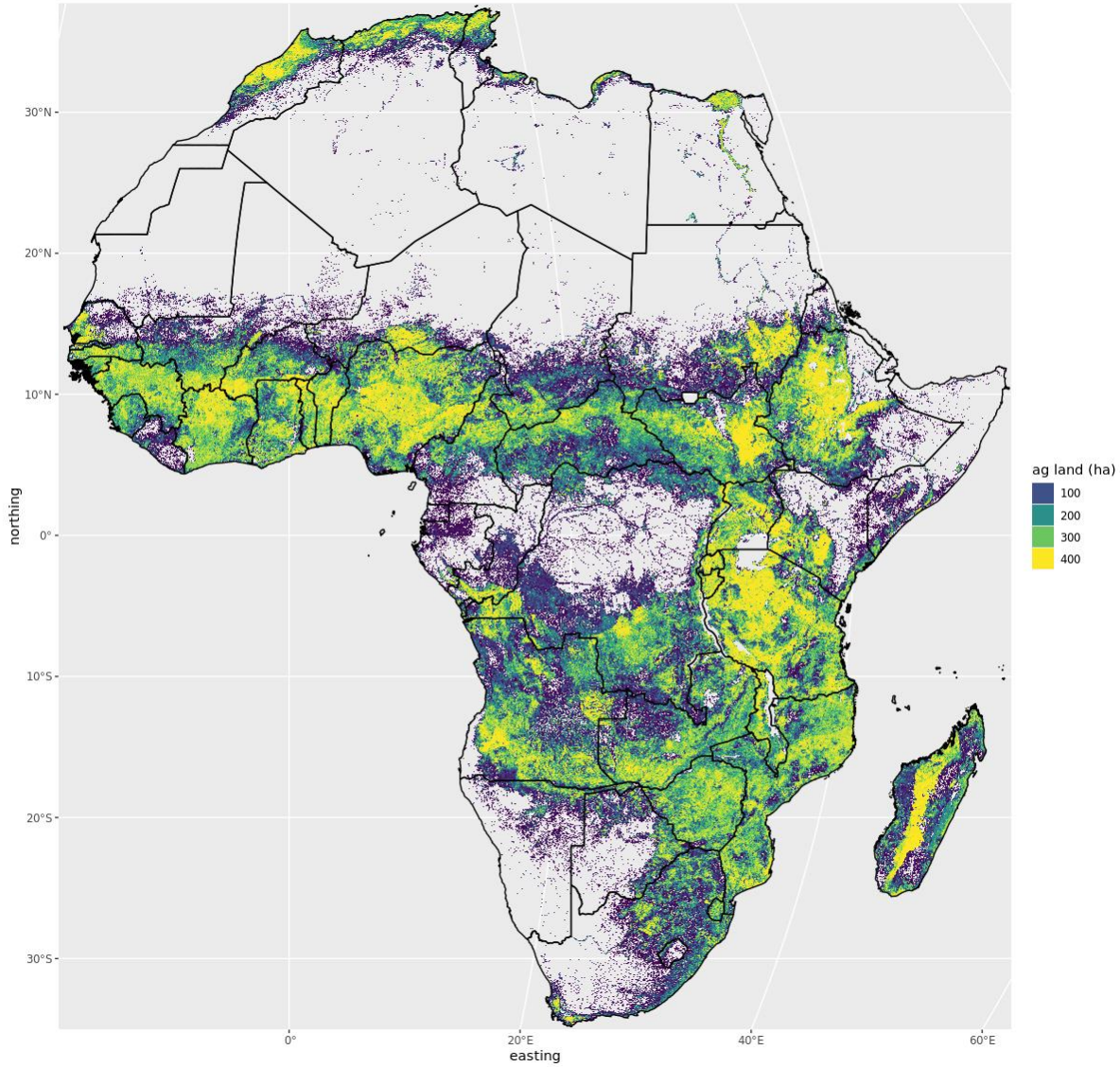

Figure 7.Map of primary sampling units (PSUs), legend indicates the number of 1 ha agricultural land pixels that each PSU contains.

Design type

From the above follows that we have three prerequisites for the sampling design:

1. Sampling unit selection by probability sampling;
2. Clustering of the population units that are represented by 1 ha agricultural land pixels;
3. Composite sampling within a 5 m x 5 m sampling plot (the basic sampling unit) located within a population unit.

From these requirements it becomes clear that we have three types of sampling units that are nested:

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

1. Cluster of agricultural land pixels
2. Agricultural land pixels
3. Sampling plots

Each of these sampling units is a discrete entity in space. An obvious design type for such situation is multistage random sampling, in our case **three-stage random sampling** and this is the basic design type that was selected for the Soils4Africa sampling design. The sampling stages are defined as follows:

- Stage 1: 2 km x 2 km clusters; each containing between 1 and 400 agricultural land pixels. The clusters form the **primary sampling units** (PSUs).
- Stage 2: 100 m x 100 m pixels classified as agricultural land. The agricultural land pixels form the **secondary sampling units** (SSUs). Each SSU contains 400 sampling plots.
- Stage 3: 5 m x 5 m plots; . The plots form the **tertiary sampling units** and are the basic sampling units. The soil in a selected plot is sampled at four points forming a square grid with a spacing of 2.5 m (Figure 8).

Figure 8 gives a schematic representation of the three-stage random sampling design. Note that the 2 km x 2 km squares contain agricultural as well as non-agricultural land and as a consequence the area of agricultural land differs among the PSUs.

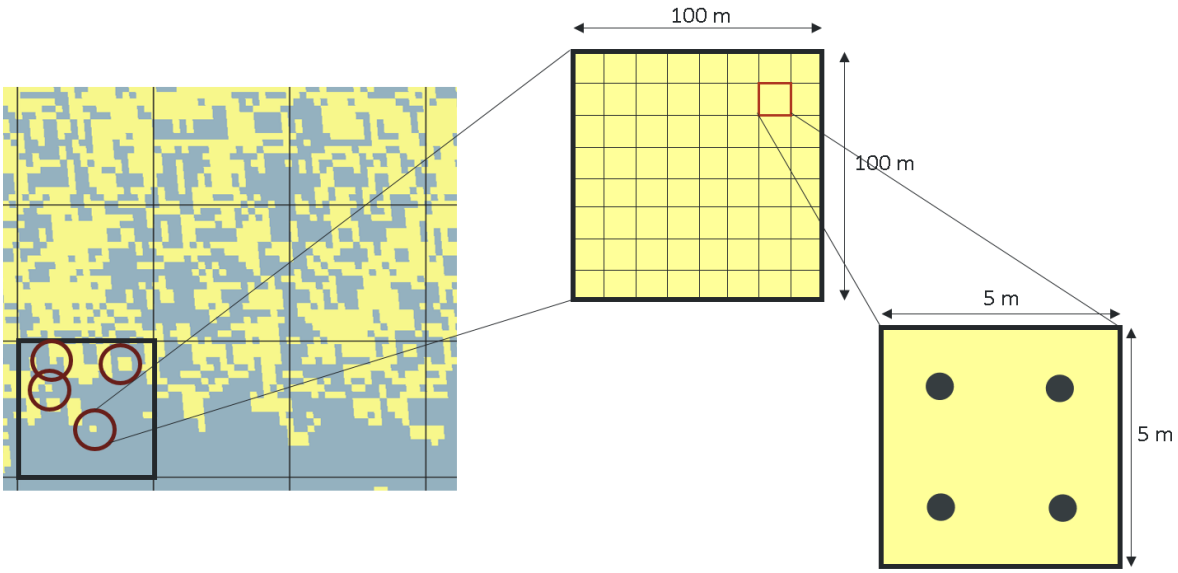

Figure 8.Schemtatic representation of the Soils4Africa sampling design type: three-stage random sampling. The primary sampling unit (PSU) is indicated by the black rectangle in the figure on the left. The PSU contains a number of one ha pixels that are classified as agricultural land, the secondary sampling units (SSUs). A fixed number of SSUs will be selected in each selected PSU. Each SSU is composed of 400 tertiary sampling units (TSUs) that measure 5 x 5 m. In each selected SSU one of these TSUs is selected for field sampling. Soil aliquots are taken at four locations within the TSU and bulked into a composite sample.

Sampling units will be selected in each of the three stages of the design. The sampling locations of the TSU are fixed and therefore do not need to be selected randomly within the TSU. The method for sampling unit selection does not need to be the same for each stage. In other words, each stage can have its own design type.

We have chosen a **stratified random sampling** approach for selecting the primary sampling units (see next section for more information). The probability of selecting a PSU is proportional to its size, i.e. area of agricultural land. As the example in Figure 9 shows, the area classified as agricultural land differs among the PSUs and varies between 1 and 400 ha (with a median area of 214 ha). By selecting PSUs with **probability proportional to size** (pps) we ensure that all SSUs that belong to the same stratum have the same inclusion probability. This will simplify the

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

statistical inference. Furthermore, PSUs will be selected **without replacement (ppswor)**, which means that each PSU cannot be selected more than once. PSUs are selected with the **local pivotal method** (Grafström et al., 2012), using the geographic coordinates of the centers of the PSUs as spreading variables, to improve geographic spreading of the PSUs within strata. We expect that this geographic spreading increases the accuracy of the estimates.

The secondary and tertiary units will be selected with **simple random sampling without replacement (srswor)**. Because the PSUs and SSUs are small there is no need for a more sophisticated selection method. Also, simple random sampling will keep the statistical inference straightforward.

The full design can thus be described as follows: **Three-stage cluster random sampling, with stratified ppswor-sampling in first stage, srswor sampling in the second and third stage.**

In total 5000 PSUs will be selected across the area classified as agricultural land in Africa with a maximum of 4 SSUs in each selected PSU, bringing the total sample size to 20,000. Optimization of the number of PSUs per stratum, number of SSUs per PSU and number of TSUs per SSU *a priori* was not possible because no suitable data were available to assess the various variance components of the design – i.e. the variation within a stratum and within PSUs, SSUs and TSUs – that one would need for such optimization. However, with the data being collected within this project it will be possible to optimize sample sizes for next sampling rounds. This would require though that multiple TSUs are sampled for a small subset of selected SSUs.

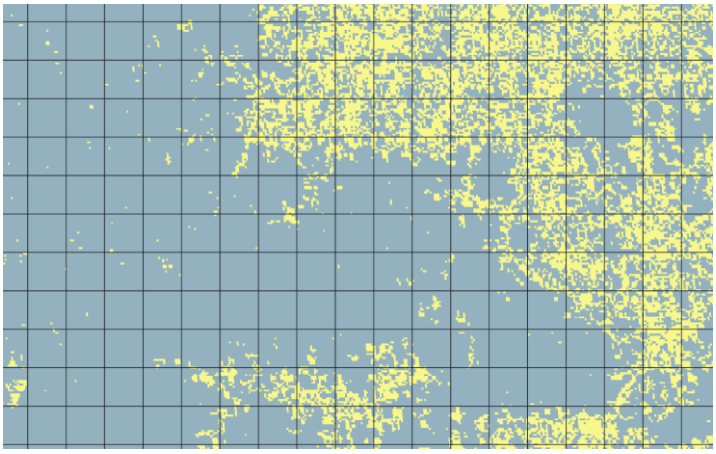

Figure 9. Example of clustering of 1ha pixels classified as agricultural land (depicted in yellow) into primary sampling units (PSUs). A PSU is represented with a transparent square outlined in black. From this example it is obvious that the number of agricultural land pixels can differ greatly between clusters. Some clusters have none, these will not be considered for sampling. Some have only a few, while others have many (up to 400).

Stratification of the primary sampling units

As explained in Chapter 2, stratification of sampling units can be advantageous. It can make the sampling design more efficient and one can control sample sizes in domains of interest (by using these domains as strata) and thus ensure that all domains are sampled which is not guaranteed with simple random sampling. We have collected a number ancillary data layers (see section on 'Ancillary data' in this chapter), some of which are likely candidates for stratification. These include WRB reference soil groups, agro-ecological zones (AEZ), farming systems and land cover classes.

To increase the efficiency of the design the strata and target soil properties and indicators should be associated. The association of four basic soil properties and four candidate stratification variables was studied by fitting one-way ANOVA models, using Africa-wide legacy soil data reported by Batjes and Ribeiro (2021). Results in the form of fraction of explained variation are presented in Table 1 and show that all variables except land cover are reasonable predictors of at least two of the properties tested, with WRB RSGs and farming systems generally outperforming AEZ and land

European Commission | Horizon 2020 European Union funding for Research & Innovation

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

SOILS4AFRICA

cover. The farming system 2001 version performs somewhat better than the 2019 level 1 version. The more detailed 2019 level 2 version, however, outperforms the level 1 layer and all other variables tested.

Table 1. Fraction of explained variation (adjusted R^2 values) of one-way ANOVA models for four basic soil properties (obtained from the WOSIS database: 26,423 data points, 0-30 cm layer) using one out of four categorical variables that are candidates for stratification as a factor: WRB reference soil group, farming system (2001 and 2013 version; with 2 classification levels for the latter), land cover and agro-ecological zone. WRB RSG: 20 classes; FS2001: 16 classes; FS2019lev1: 13 classes, FS2019lev2: 42 classes, Land cover: 5 classes; AEZ: 12 classes.

	WRB RSG	**FS2001**	**FS2019lev1**	**FS2019lev2**	**Land cover**	**AEZ**
Clay	0.133	0.084	0.070	0.132	0.017	0.070
SOC	0.100	0.182	0.137	0.189	0.035	0.204
CEC-pH7	0.246	0.190	0.166	0.313	0.005	0.020
pH-H_2O	0.204	0.232	0.216	0.309	0.099	0.209

PSUs are selected by stratified random sampling. This implies that the stratification is applied at the level of the PSUs so that each PSU belongs to one class of a stratification variable, i.e. can only belong to one soil class, one farming system, one land cover class, etc. It may happen that a class boundary crosses a PSU. An hypothetical example of such situation is given in Figure 10. In that example there are a few PSUs that are partly located in class 'A' and partly in 'B'. For these PSUs the dominant (modal) class is determined and that class is then assigned to the PSU. In this case the dominant class is not the class that covers the largest surface area in the PSU but the class that covers the *largest area classified as agricultural land* within the PSU. See also Figure 10.

So for each of the about 4M PSUs in the sampling universe, we determined the dominant class for each of the categorical ancillary variables and the mean value (calculated only considering the agricultural land pixels) for the each of the continuous ancillary variables (elevation, topographic wetness index, length of growing season, accessibility, population density).

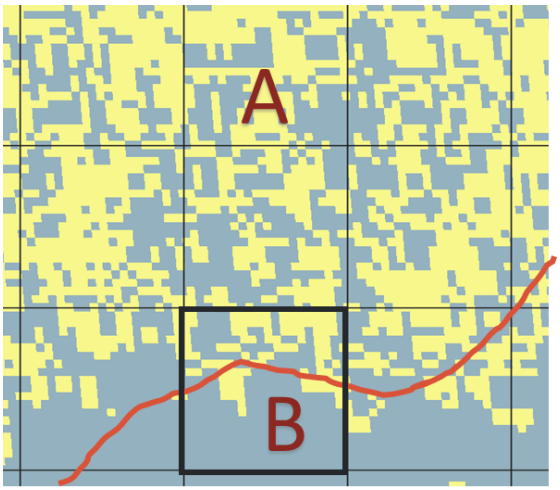

Figure 10. Hypothetical example of stratification of primary sampling units. The curvy line represents the boundary between strata A and B. Some PSUs fall completely within a single stratum and are assigned to that stratum. Others fall on a stratum boundary (e.g. the PSU outlined in bold). The stratum assigned to such PSU is the one that contains the majority of the agricultural land pixels. In this case stratum 'A' despite that 'B' covers a larger surface area.

We have considered using multiple variables for stratification. A cross-classification approach (see Chapter 2, section 'Stratification') was not practically feasible because of the prohibitively large number of cross-class strata (9,000 when using farming system, soil class and land cover class as stratifying variables). We therefore explored and tested the multi-way stratification approach (Chapter 2) that allows to control sample sizes in all classes without making use of cross-classes. Though we succeeded implementing multi-way stratification within the first stage of the three-stage sampling design, we finally abandoned this effort because estimation of the sampling

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

variance of estimated means and totals becomes quite complicated, hampering practicability of the design.

Hence the decision was taken to simplify stratification and use only one thematic variable to stratify the sampling universe. Given the scope of the project (agricultural intensification), farming systems were selected for this purpose since these present a relevant reporting unit and by selecting this variable we ensure that we have soil aliquots collected across all major systems. Besides, table 1 showed that the farming system layers are adequate predictors of four basic soil properties at continental level. We used the update of the 2001 version that was developed and described by Van Velthuizen et al. (2013) and Dixon et al. (2019). This dataset has two levels: 13 major farming systems that are subdivided into 42 sub-systems. Sub-divisions are largely controlled by geography. Figure 11 shows the level 1 farming systems across the sampling universe. Annex 1 the full classification. The level 2 systems were used for stratification.

Even though we do not use other thematic variables for stratification, we point out that we can still exploit information contained in other relevant variables for estimation (see Chapter 2, section 'Sampling of sub-areas and estimation of local quantities', post-stratification). Besides, it is still possible to obtain separate estimates of target quantities for other domains such as soil classes, land cover classes, agro-ecological zones, etc.

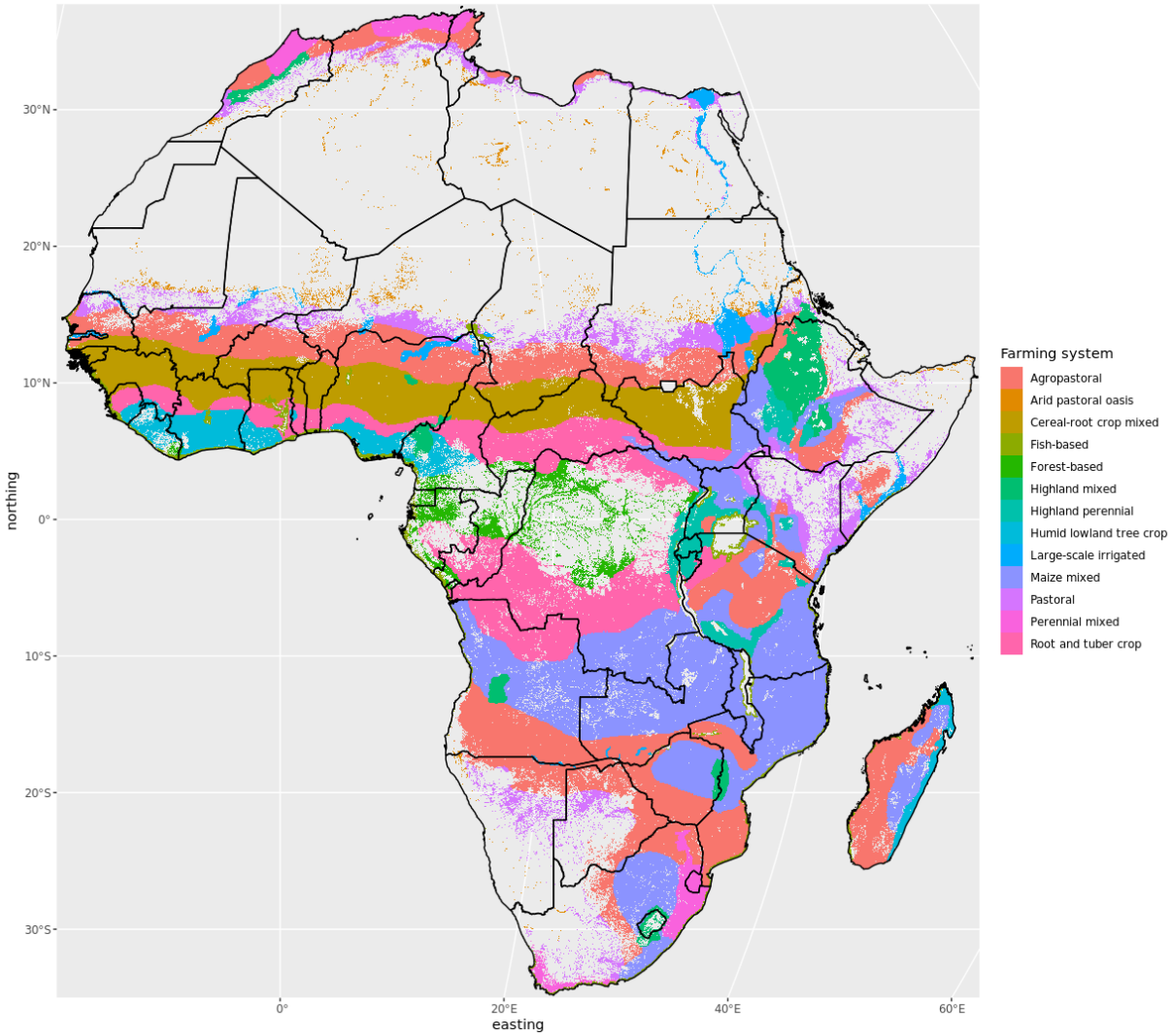

Figure 11. Map of major farming systems in the sampling universe (Van Velthuizen et al., 2013).

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

Refining the sampling universe

Disputed territories

There are several disputed territories in continental Africa. These territories have been excluded from the sampling frame. The excluded are Ilemi triangle (Kenya/South Sudan), Halib triangle (Egypt/Sudan), Bir Tawil (unclaimed by Egypt/Sudan), Western Sahara (Morocco/Sahrawi Arab Democratic Republic). Apart from the Ilemi triangle, the area classified as agricultural land was negligible for these areas.

Elevation

High-elevation areas were excluded from the sampling frame as well. An elevation threshold of 3,000 m was used for this purpose. The excluded areas pertain to relatively small areas mainly in Ethiopia and Kenya covering about 15,000 km^2.

Fieldwork safety

Unfortunately, there are extensive regions in Africa where the current safety situation do not permit fieldwork such as for instance in the parts of the Sahel, horn of Africa or northeast Mozambique. The map in Figure 12 shows areas classified as 'unsafe' for fieldwork (see Annex 2 for an overview). This information was elicited from project partners and other, more general risk assessments[3]. We point out that this overview might not be yet complete and that the overview will be updated once more information about safety becomes available when planning fieldwork. The current overview classifies 9.6% of the area classified as agricultural land as 'unsafe'.

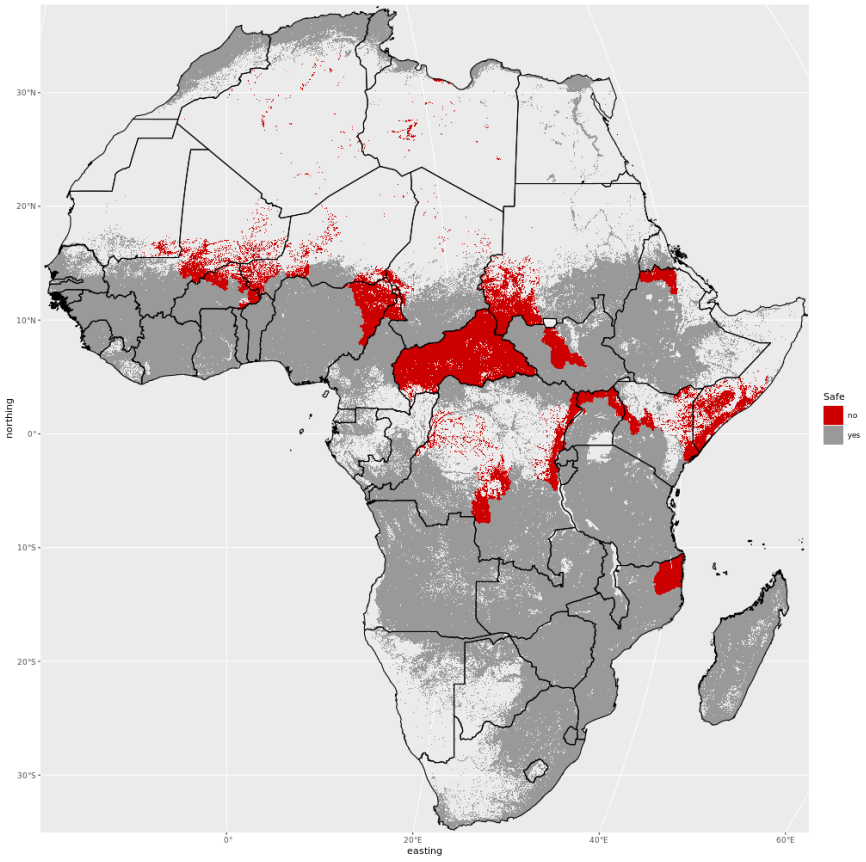

Figure 12. Map indicating unsafe regions in Africa.

[3] https://www.controlrisks.com/-/media/corporate/files/riskmap-2021/riskmap-2021-map-regions-africa-a3.pdf, https://www.controlrisks.com/-/media/corporate/files/riskmap-2021/riskmap-2021-map-regions-mena-a3.pdf

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

Areas classified as 'unsafe' are not excluded from the sampling frame, but the 5000 PSUs were selected from safe areas. Besides, additional units were selected from unsafe areas so that these could be sampled once the safety situation improves. In this way we ensure that each country has at least a complete design.

Protected areas

Africa has vast areas of land that have a protected status, with varying types and degrees of protection. Protected lands are mapped and described by UNEP-WCMC (2021). The data is available in form of three GIS layers. In total, these three layers distinguish 128 unique classes (though there is some overlap between these; e.g. 'Hunting Area', 'Hunting Zone' , 'Hunting Reserve'). About 18.5% of the area classified as 'agricultural land' falls within a protected area.

The 128 protected area classes were standardized into 25 classes depending on type of protected area. A 'sampling class' was associated to each of these based on protected area type (see Annex 3). These classes include "no sampling", "sampling in cultivated land only", "sampling". Main areas excluded for sampling include National Parks, private nature reserves and ranches and (RAMSAR) wetland reserves. Faunal reserves, forest reserves, nature reserves, wildlife reserves and UNESCO Biosphere and World Heritage Sites are sampled only in areas that are under cultivation, under the assumption that non-cultivated land in these reserves is not or only very extensively used for agriculture and therefore of less relevance to this project.

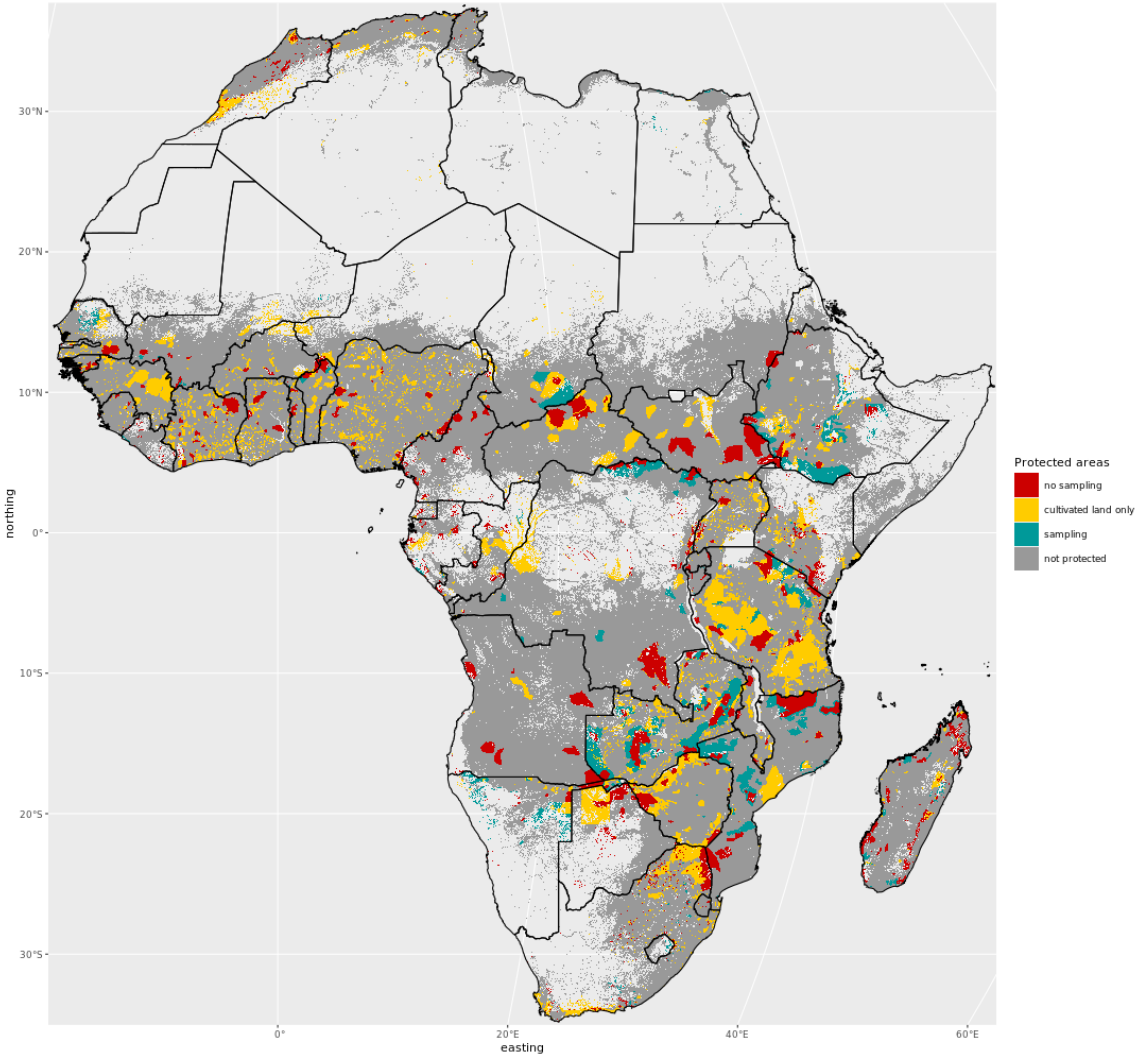

Figure 13. Map of protected areas and classification according to their possibility for soil sampling.

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

The basic 'sampling' classification was refined with information on IUCN protected areas categories[4] (which was only available for a very limited set of protected areas). If a protected area was classified as a Ia, Ib or II category, then 'sampling class' was set to 'no sampling', irrespective of the type of protected area. Further refinement was introduced based on ownership type. Not all privately owned nature reserves were classified as such in the dataset. This was especially true for the nature reserves in South Africa. Because such private nature reserves are typically dedicated to conservation and tourism and do not have a primary agricultural function, we classified all nature reserves for which the government type was 'individual land owner' as private reserve and assigned sampling class 'no sampling' to these.

Of the 18.5% of agricultural land area that falls within a protected area, 5.4% is directly excluded from the sampling frame, in 9.8% we can sample in cultivated areas and in 3.2% we assume we can sample. Figure 13 shows the distribution of 'sample classes' for the protected areas within the sample universe. Implementing this classification results in 13.2% of the area classified as agricultural land to be excluded for sampling (Figure 14). Figure 15 shows an example for Tanzania.

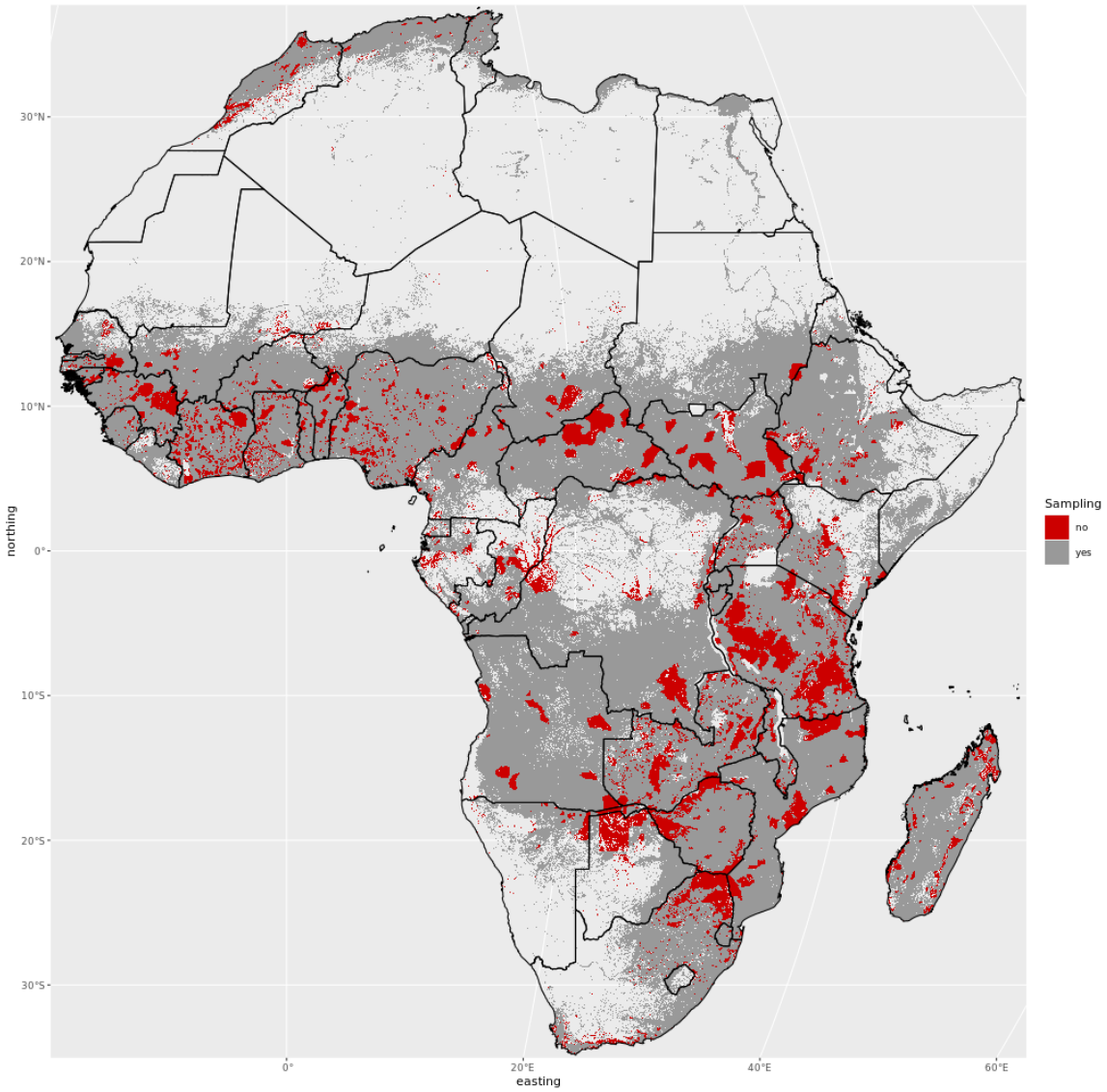

Figure 14. Map of protected areas within area classified as agricultural land where sampling is not possible.

[4] https://www.iucn.org/theme/protected-areas/about/protected-area-categories

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

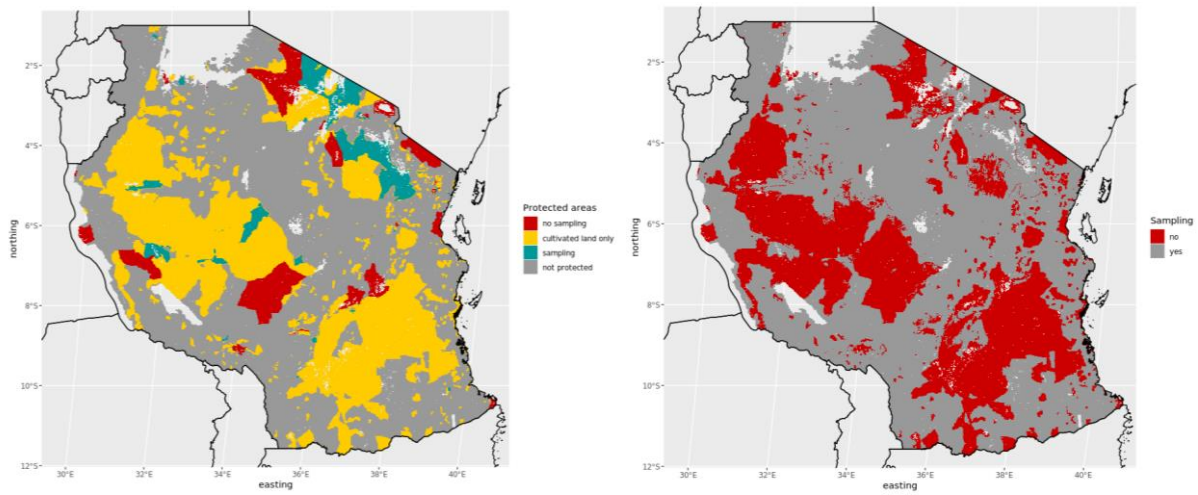

Figure 15. Map of protected areas and classification according to their possibility for soil sampling for Tanzania (left) and map of protected areas within area classified as agricultural land in Tanzania where sampling is not possible.

Sample sizes and sample selection

Sample sizes

The size of the sampling universe after excluding high elevation and protected areas, but including the 'unsafe' areas is 6.7M km^2. The table in Annex 4 lists the area classified as agricultural land per country. We note here that 50% of the countries contain 90% of the area classified as agricultural land.

Countries are important domains for which the project wishes to report separate results. Furthermore, countries will be important drivers of soil change since sustainable agricultural intensification efforts will be determined by national policies. Consequently countries, though not directly related to soil conditions, will form a relevant monitoring unit. We therefore wish to control sample sizes in each country. Country therefore was used as a second stratification variable, so that farming systems within countries were used as strata. Sampling unit selection at country level, with sample sizes that can be selected for each country individually, has the additional advantage that it adds flexibility to the design (see section 'Considerations for designing a monitoring scheme' in the next chapter).

Allocation of sample sizes was done hierarchically: first the total sample sizes per country were determined, then the total sample size of a country was allocated to the farming systems within that country. PSUs are allocated to the countries proportional to the surface area of agricultural land (rounded up). This gives an equal sampling density for all countries. The table in Annex 4 shows the distribution of PSUs across the countries (and main land cover classes and WRB reference soil groups) based on proportional allocation. The table also shows that, as a consequence of having a relatively small area of agricultural land, sample sizes are very small for some countries or even zero for Djibouti. Such small sample sizes are not realistic and for now we assume a minimal sample size of 25 PSUs (100 sampling locations) for each country but this number can be adjusted if necessary. For countries with PSU sample size between 26 and 39, the sample size will be increased to 40; for countries with sample size between 41 and 49 to 50. To stay within the targeted 5,000 PSUs, sample sizes for countries with over 300 PSUs will be reduced by 8% and sample size for countries with 200 to 300 PSUs with 4%. Sampling units of countries excluded from sampling (e.g. Central African Republic) will be proportionally re-allocated to the other countries. Finally, we note that the (proportional to size) sample sizes in Annex 4 are indicative and that these are likely to change when operationalizing the design.

European Commission | Horizon 2020 European Union funding for Research & Innovation

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

SOILS4AFRICA

The sampling frame

Selection of sampling units requires a sampling frame. A sampling frame is a representation of the sampling universe that lists all possible sampling units. Sampling units will be selected from the sampling frame. This typically happens in two steps. First, the inclusion probability of each sampling unit is calculated. The inclusion probabilities are determined by the design. Second, a sampling algorithm is used to select sampling units from the sampling frame, respecting the inclusion probabilities of the units. Construction of one sampling frame with all SSUs in the sampling universe, with an extra column indicating the PSU of each SSU, was not feasible. Therefore we constructed two sampling frames: one containing all PSUs in the sampling universe, and one containing all SSUs for the selected PSUs. Besides the sampling units themselves, the sampling frame contains all information that is required by the design to select sampling units. Annex 5 shows the structure of the sampling frame of the PSUs.

Sampling unit selection

PSUs were selected for each country using the samples sizes listed in Annex 4 with a minimum of 25 samples per country and sample sizes rounded up to 40 and 50 for some countries with relatively few PSUs as described before. Prior to allocation to the strata, strata containing less than 150 PSUs (jointly containing max. 600 km2 of agricultural land) and covering less than 1% of the country's agricultural land area were discarded since these mainly represent polygon slivers resulting from an intersection between country and farming system.

Since we cannot sample in areas classified as 'unsafe' we have to take fieldwork safety into account when selecting sample locations. We therefore cross-classified the farming system strata with a safety indicator (2 classes; Figure 12). In order to reach the minimum requirement of 20,000 sampling locations, the pre-determined PSU sample size was allocated to the area classified as 'safe' and distributed across the farming system strata within this area proportional to stratum size (see previous section) with a minimum of three PSUs per stratum. As a consequence, the country sample sizes are not exactly proportional anymore to the agricultural land area, but this does not impair the requirement of unbiased and valid estimation of target quantities. The sample of 20,000 units was supplemented with a sample for the areas classified as 'unsafe'. The sample size for this area was chosen in such a way that the sampling density equals the sample density in the 'safe' area. Thus giving a uniform sample distribution across the entire, safe and unsafe, agricultural land area in a country. Figure 16 shows the distribution of about 5200 selected PSUs across Africa's agricultural land in areas judged as 'safe'. The total number is somewhat higher than the targeted 5000 because of minimum sample size requirements for countries and strata. In addition, about 800 additional PSUs are selected in area that are judged 'unsafe' to sample. Note that this selection is tentative and small changes can still be expected when planning fieldwork.

Processing of the sampling frame and sample selection was done with the statistical software R (R Core Team, 2020) making it fully reproducible and easily updateable. Selection of sampling units was done with the statistical software R using packages *BalancedSampling* (Grafström and Lisic, 2018) and *sampling* (Tillé and Matei, 2021). All code is stored in the GitLab repository of Wageningen University and Research[5].

An illustrative example for Kenya

In this section we give an example of the selection of PSUs for Kenya. Figure 17 shows the approximately 71,700 PSUs for Kenya and Figure 18 the farming system strata. Figure 19 depicts

[5] https://www.wur.nl/en/Value-Creation-Cooperation/Collaborating-with-WUR-1/Manage-your-source-code-with-GitWUR.htm

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

the fieldwork safety classification and Figure 20 the farming systems x safety cross-strata. The PSU sample size for Kenya is 108. Area classified as 'unsafe' coves 14.7% of the total agricultural land area. The 108 PSU are allocated to area classified as 'safe'. This means that we need to select an additional 19 units for the area classified as 'unsafe' to have the same sampling density in this area bringing the total sample size to 127. These are subsequently allocated to the farming systems x safety cross-strata proportional to strata sizes, with a minimum of three and rounded upwards. Next, a spatially balanced sample is selected within each stratum with the local pivotal method. Actual number of selected sampling locations is 130 because of the minimum requirement and rounding upwards. Figure 21 shows the distribution of the selected PSUs across the farming system strata. The points in red are the sampling locations in 'unsafe' areas.

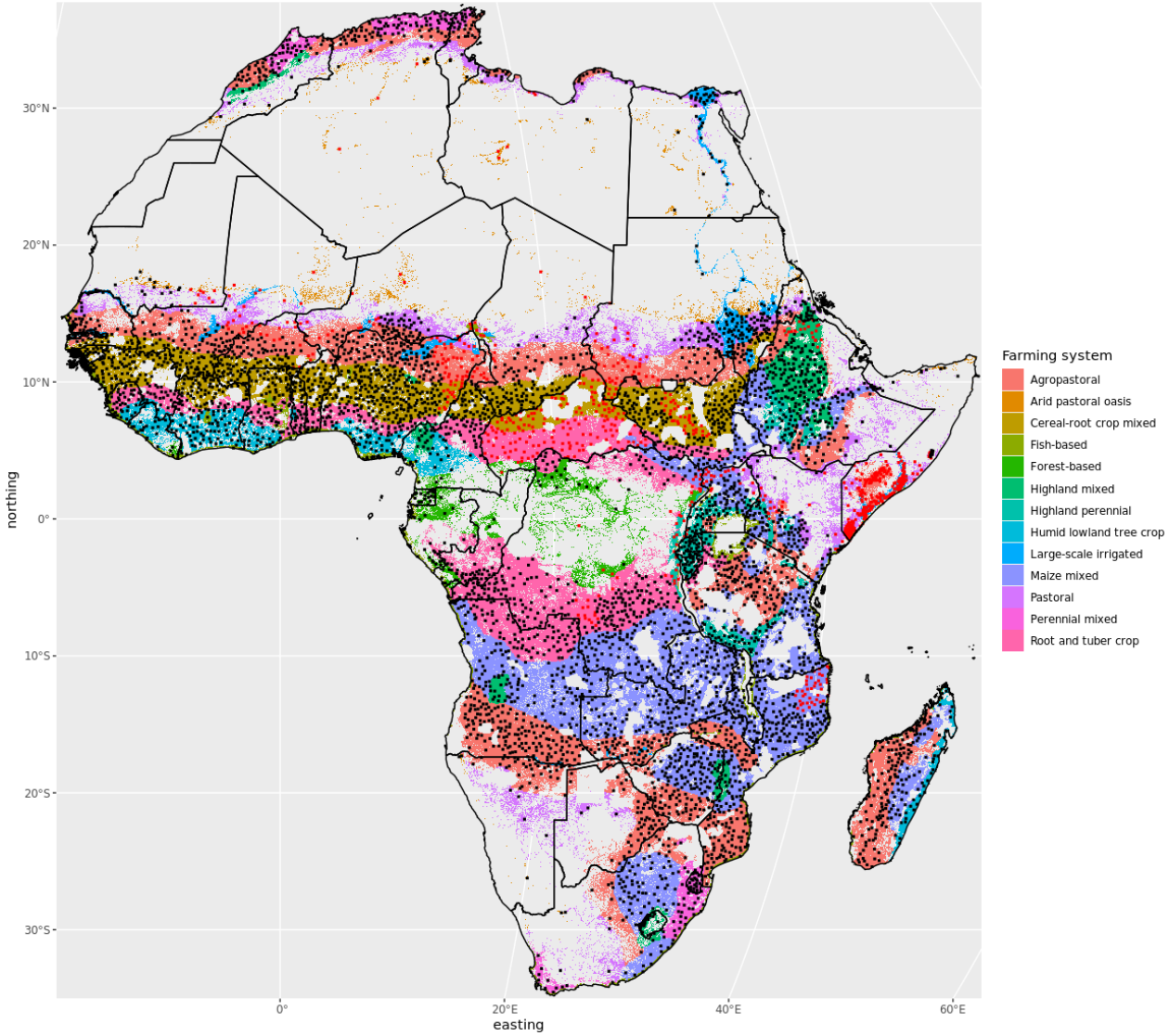

Figure 16. Locations of the selected primary sampling units.

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

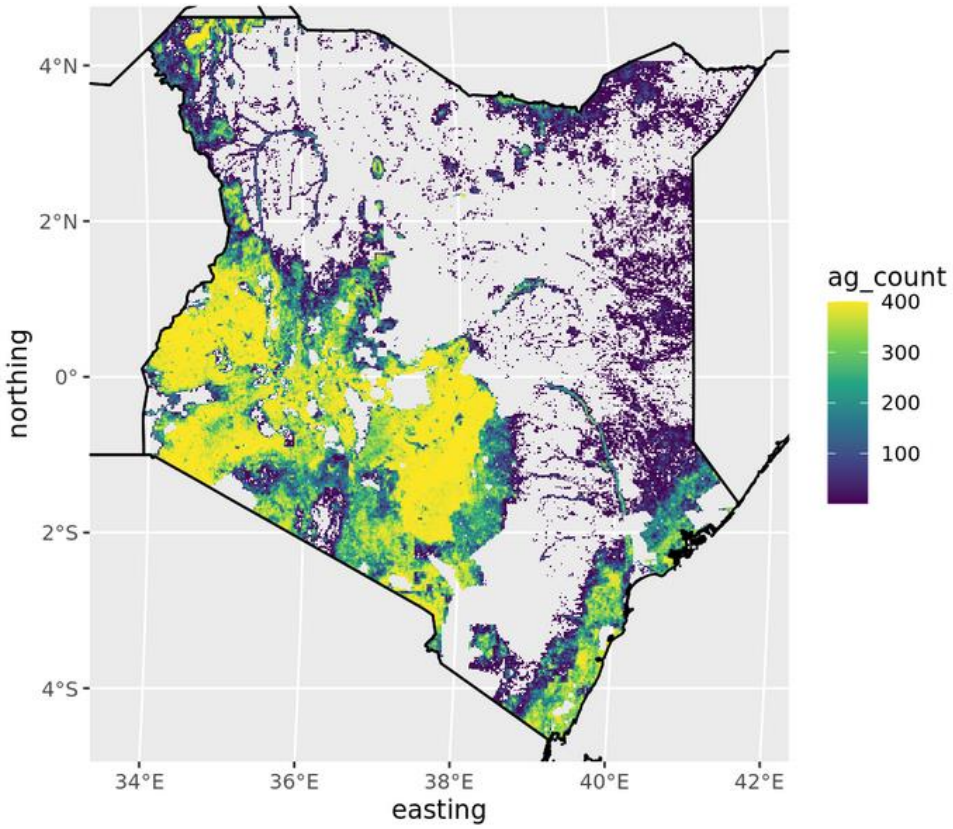

Figure 17. Map of primary sampling units (PSUs) in Kenya that the number of 1ha agricultural land pixels per PSU.

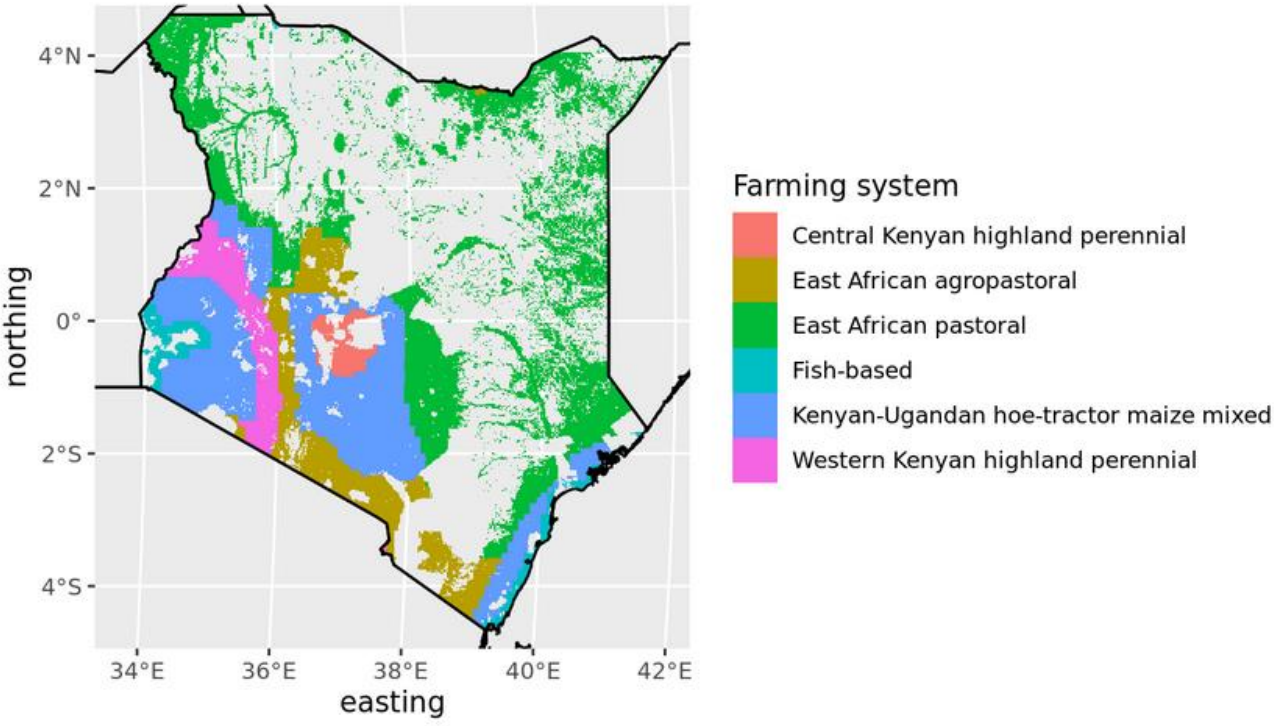

Figure 18. Major farming systems in Kenya following Van Velthuizen et al. (2013).

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

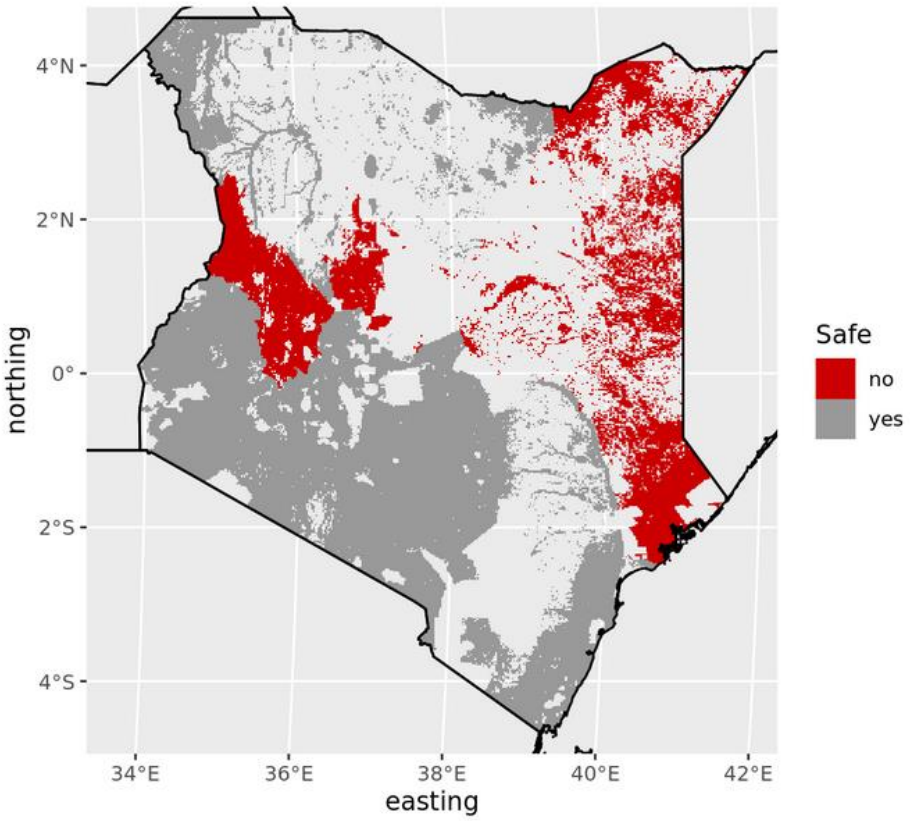

Figure 19. Safety classification. Areas judged unsafe to sample are indicated in red.

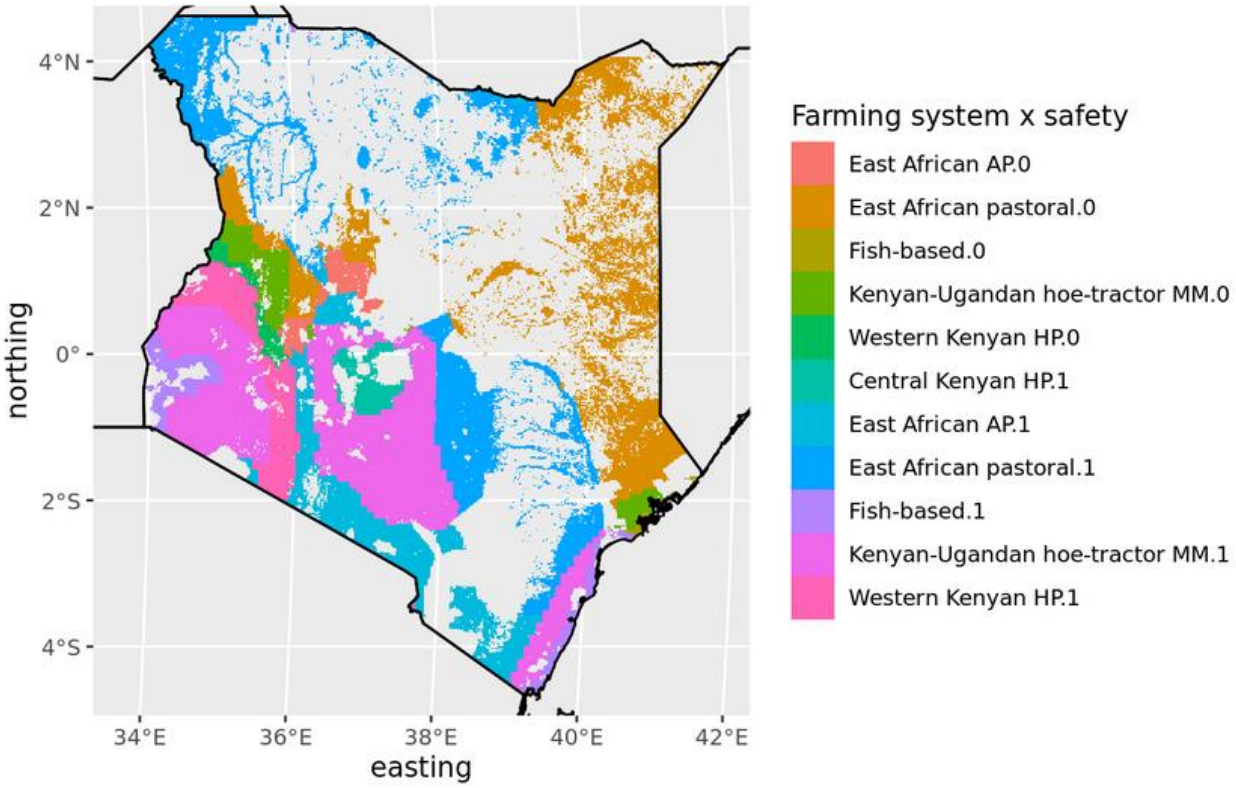

Figure 20. Cross-classification of farming systems and sampling safety classes. The '0' addition in the class name indicates 'unsafe', '1' indicates 'safe'.

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

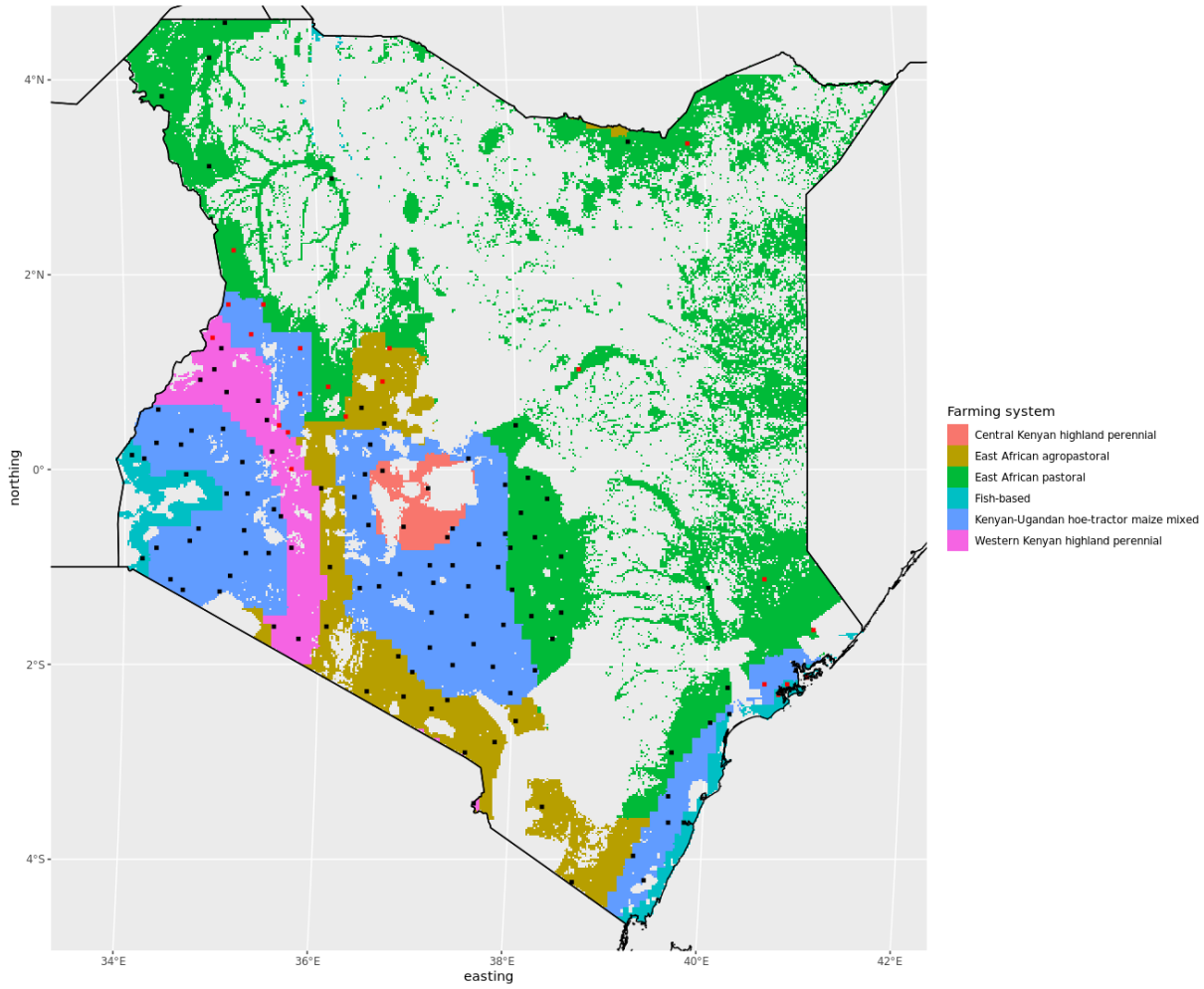

Figure 21. Selected primary sampling units for Kenya (n = 130). Locations indicated in red (n = 22) are in unsafe areas and will not be visited during the project.

Design decisions and trade-offs

Each sampling design type has advantages and disadvantages. Choosing a design type always involves trade-offs, typically between sample sizes that can be afforded for a given cost, accuracy of the estimators and operational efficiency. For the Soils4Africa sampling design we have chosen for a (stratified) three-stage cluster random sampling design type so that spatial clusters of 5 m x 5 m sampling plots are selected, each cluster consisting of four plots. Such design provides clear operational advantages; more sampling units can be observed in a given survey time since less travel time is required (De Gruijter et al., 2006; Brus, *in prep*). Given the scope of the Soils4Africa project, where 20,000 soil aliquots need to be collected within a 8M km2 area with sometimes poor infrastructure and limited accessibility of the terrain, we expect the operational advantages to be huge as compared to other sampling designs spreading out the sampling plots throughout the 8M km^2 study area. Instead of visiting 20,000 individual locations, 5000 clusters of sampling sites are visited which reduces travel time by some 75%. Given the relatively limited time for the sampling campaign, this reduces some of the risks associated to a project of this scope.

In multi-stage random sampling there is a trade-off between operational efficiency and accuracy of the estimates (De Gruijter et al., 2006; Brus, *in prep.*). Given a fixed number of 5 m x 5m sampling plots, estimates obtained from a stratified three-stage cluster random sample are less accurate than those obtained from a stratified simple random sample. This also means that for a

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

fixed sample size, the minimum detectable change in soil conditions is larger. Or the other way around, to be able to detect a given change at a given significance level and with a given statistical power, a larger samples size is needed for stratified three-stage cluster random sampling than for stratified simple random sampling.

It is possible to optimize the sample sizes of the different sampling units of a multi-stage sample design given a maximum budget or minimum precision (maximum variance of estimated mean) (Brus, *in prep*). This requires a cost model and a priori estimates of the variance components of the design (Domburg et al., 1997). In the case of a stratified three-stage cluster random sampling design these variance components are:

1. The variance of the true PSU means (2 x 2 km clusters) within the strata
2. The variance of the true SSU means within the PSUs (100 x 100 m pixels)
3. The variance of the true TSU means within the SSUs (5 x 5 m sampling plots)
4. The variance at points within a TSU (which we can possibly assume to be negligible compared to the other variance components).

As soon as data of the first sampling round are available, these variance components can be estimated as well as the parameters of a costs model. These estimates can then be used to optimize the number of PSUs per stratum, number of SSUs per PSU, number of TSU's per SSU and number of points per TSU. The estimated variance components (for a set of relevant soil properties and soil quality indicators) can then also be used to calculate minimum detectable change for a given sample size and the calculation of sample sizes required for a given minimum detectable change. At this stage though, a priori information about the variance components cannot be derived from available (soil point) data sources. A cost model is not available either. It is therefore not possible to optimize the sample sizes, nor to give an indication of likely minimum detectable changes a priori. Once soil sampling has started, however, such information about the variance components and costs will become available. The Soils4Africa soil dataset will therefore be a great asset that will allow sample size optimization for the current design and sampling efficiency comparisons between alternative designs. To compute the required total sample size for a given minimum detectable change also information is needed of the temporal variation. If sampling location are revisited, an estimate of the temporal correlation of two observations at the same point but at different times is needed.

In Chapter 1 we mentioned that a secondary purpose of the project is to develop gridded soil maps of key soil properties and soil quality indicators. Though probability sampling designs are not always optimal for this purpose, this does not mean that the data collected by the project cannot be used for mapping. Though possibly not optimal for a mapping purpose, the data can still be very well used for mapping, and can even be supplemented with soil point data from other sources to improve the digital soil mapping models. Use of the data for mapping was anticipated with implementation of the local pivotal method that optimizes geographic spreading of the selected samples.

European Commission | Horizon 2020 European Union funding for Research & Innovation

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

SOILS4AFRICA

5. Sampling for monitoring[6]

Monitoring soil conditions over time is not an aim of the Soils4Africa project. Nevertheless, the sampling design and collected data should provide a statistically sound basis from which future monitoring of soil across the African continent or parts there of (e.g. in countries) is possible. The final chapter in this report is therefore dedicated to an overview of sampling approaches for monitoring.

One of the main purposes of monitoring is to estimate the *change* of a (soil) variable between two sampling times (cycles). For instance, we may be interested in the change in the mean soil organic carbon concentration in the topsoil of a country, or the change in areal fraction of degraded soil in a landscape after intervention measures have been put in place. If one has more than two sampling times, then the interest might be in the average change per time unit of the mean, total or fraction, i.e., the temporal trend. There are several (global) space-time quantities that can be estimated (De Gruijter et al., 2006) of which the following are the most relevant in the context of the Soils4Africa project:

- the current mean (i.e., the spatial mean at the most recent sampling time);
- the change of the spatial mean from one sampling time to the other;
- the temporal trend of the spatial mean.

In this report we only consider sampling approaches for monitoring global quantities or local quantities in large sub-areas of the sampling universe. We thus limit ourselves to the estimation of quantities such as means, totals, areal fractions that we believe are most relevant for Soils4Africa. Methods for local (i.e., location-specific) prediction in space-time are not considered here. These methods pertain to spatio-temporal mapping, i.e., mapping a target (soil) variable over the entire space-time universe and updating maps. We refer the reader to De Gruijter et al. (2006) who provide an elaborate overview of (model-based) methods for spatio-temporal mapping such as space-time kriging and Kalman filtering, while more recently Heuvelink et al. (2020) used machine learning for space-time soil organic carbon mapping.

Statistical sampling approaches

In designing a sample for monitoring, not only spatial variation is relevant, but temporal variation must be taken into account as well. This means that sampling times must be selected in addition to sampling locations, leading to four possible sampling approaches, i.e., combinations of probability (P) and non-probability sampling (NP) (Brus, 2014):

- P-sampling in both space and time (P + P),
- NP-sampling in both space and time (NP + NP),
- P-sampling in space and NP-sampling in time (P + NP), and
- NP-sampling in space and P-sampling in time (NP + P).

With P + P the space-time mean (total, fraction) can be estimated with a fully design-based method (Brus, 2014). This means that a model of variation in space and time is not needed for statistical inference that is fully based on the spatial and temporal sampling designs (Brus and De Gruijter, 2012). The P + P approach is advantageous in compliance monitoring of the space-time mean or space-time total (Brus and De Gruijter, 2012), such as for instance the total annual CO2 emission in a country or the change in total SOC stock in a region. The fully design-based approach

[6] This text was originally written by the lead author of this report as contribution to Teuling et al. (2021) for the EJP SOIL project. The text is re-used here in slightly adapted form.

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

avoids using models (and hence model assumptions) which enhances the validity of the result, which is important for compliance monitoring.

In the NP + NP approach a stochastic (statistical) model of the variation in space and time must be postulated to estimate the parameter of interest (Brus, 2014). This means that the statistical inference is fully model-based. De Gruijter et al. (2006) explain that fully design-based methods are generally well suited for estimating space-time means, but that in some cases using fully model-based methods could be advantageous. For instance, this might be the case in a situation where one has prior monitoring data from a purposive (non-probability) sample in space-time that needs to be extended with additional data. Since NP + NP requires models of variation in space-time, these methods are more sophisticated than methods for a fully design-based approach, and require a solid understanding of geostatistics. De Gruijter et al. (2006, Section 15.3) provide an extensive overview of geostatistical methods for a fully model-based sampling approach.

The P + NP approach uses a hybrid estimator. In this hybrid approach sampling locations are selected by probability sampling, whereas sampling times are not. Brus and De Gruijter (2012) provide a theoretical overview of the P + NP approach with a focus on the estimation of the temporal trend of the spatial mean that they illustrate with a case study on forest soil eutrophication and acidification. For this case study sampling locations were selected with probability sampling, sampling times were selected preferentially and a model of the temporal variation of the spatial means was postulated (a time series model). Brus and De Gruijter (2012) show how to do the inference of the P + NP approach with sampling repeated at constant time intervals. According to Brus (2014), the NP + P approach with non-probability sampling in space and probability sampling in time is rarely used in practice and is therefore not considered here.

For Soils4Africa the P + NP is the most relevant approach to consider. Sampling locations are selected by probability sampling. Probability sampling in time as done in the P + P approach is impractical.

Types of sampling pattern

The efficiency of the monitoring design is both determined by the distribution of the sampling events (combination of location and time) in the space-time universe. Monitoring methods are designed to minimize the estimation error due to variation in space and in time and the costs of sampling. Given any of the probability sampling design types, two different kinds of temporal restrictions can be imposed to these designs to increase the efficiency for monitoring purposes. Either sampling locations are visited multiple times (stationarity), or multiple sampling locations are visited simultaneously (synchronicity). By imposing these restrictions in different combinations, four different sampling patterns arise.

Static sampling revolves around the principle of static sampling locations. This means that sampling takes place at a fixed set of locations that are revisited (Figure 22.a). Sampling at the various locations may or may not follow the same pattern in time (De Gruijter et al., 2006).

Synchronous sampling, also known as repeated or dynamic sampling, revolve around the principle that a different set of sampling locations is selected for each sampling time (De Gruijter et al., 2006) (Figure 22.b). Note that synchronous sampling is sometimes referred to as *independent synchronous sampling* (e.g. Brus, 2014).

Combining a spatial sampling design and a temporal sampling design results in a **static-synchronous sampling** design (Figure 23.a). At every selected sampling moment in time, all sampling locations are visited. Such a design is also referred to as a pure panel.

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

The fourth option is **rotational sampling** (Figure 23.b), which is meant as a compromise between the rigid, unbalanced static design and the relatively inefficient synchronous design. The rotational design differs from the static-synchronous design in that the locations of the previous sampling time are partially replaced by new ones (De Gruijter et al., 2006). This is referred to as 'sampling with partial replacement'. In the rotational panel of Figure 23.b every sampling time one-third of the sampling locations of the previous sampling time is replaced by new locations so that the overlap of successive sampling times is two-third.

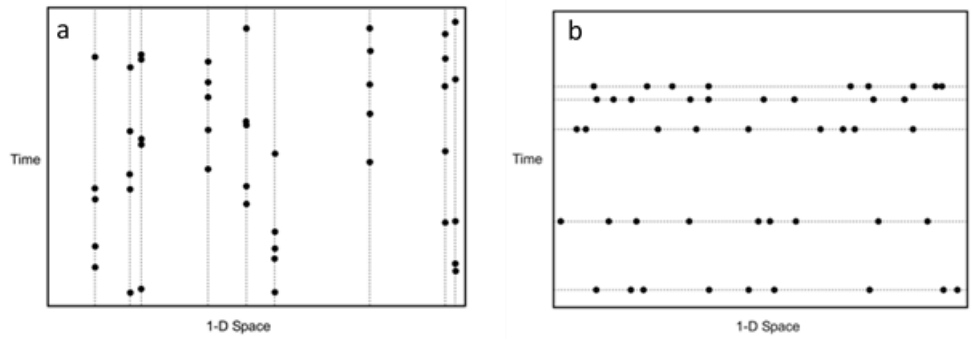

Figure 22. Schematic visualization of a static design with simple random sampling in both space and time (a) and a synchronous design with simple random sampling in both space and time (b). (Reproduced from De Gruijter et al., 2006).

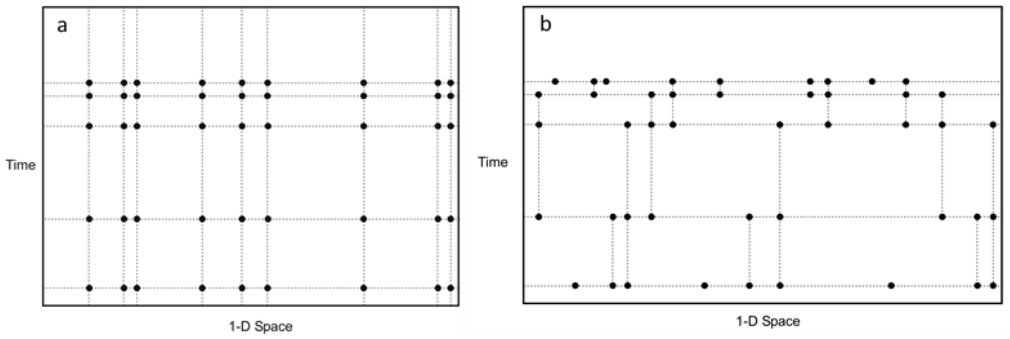

Figure 23. Schematic visualization of a static-synchronous design with simple random sampling in both space and time (a) and a rotational design with simple random sampling in both space and time (b). (Reproduced from De Gruijter et al., 2006).

In addition to these four basic sampling pattern types, two variations on these designs are worth mentioning there. The first is the **serially alternating design** (Brus and De Gruijter, 2011; Brus and De Gruijter, 2013). In serially alternated sampling, disjoint (i.e., non-overlapping) sets of sampling locations are selected and then observed in turn in a cyclic fashion (Brus and De Gruijter, 2011). The second is the **supplemented panel design**. This design is a compromise between a synchronous and static synchronous design. One set of locations (the 'pure panel' part) is maintained through time, supplemented with a different set of locations each time (Brus and De Gruijter, 2011; Brus and De Gruijter, 2013). Figure 24 provides notional examples of five space-time sampling designs that show how the different panels (sets of sampling locations sampled at the same set of moments) are used in each design.

Sampling strategies

Once the space-time sampling pattern has been chosen, one must choose a strategy for selecting the sampling locations as well as the sampling times. As we have seen previously, probability sampling as well as non-probability sampling can be used for this purpose. For monitoring, selecting sampling locations is preferably done with probability sampling. Probability sampling in time is less common than in space (De Gruijter et al. 2006). Typically, sampling times are chosen

European Commission | Horizon 2020 European Union funding for Research & Innovation

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

SOILS 4 AFRICA

purposively. This results in a hybrid P + NP sampling approach. With this approach the change of the mean between two sampling times can be estimated by full design-based inference. The average change per time unit of the spatial mean (total), i.e. the temporal trend of the spatial mean (total) can be estimated either by a hybrid estimator (Brus and De Gruijter, 2013), or fully design-based (Brus and De Gruijter, 2011),. In the latter case the target universe is restricted to the sampling times only; what happens between the sampling times is disregarded.

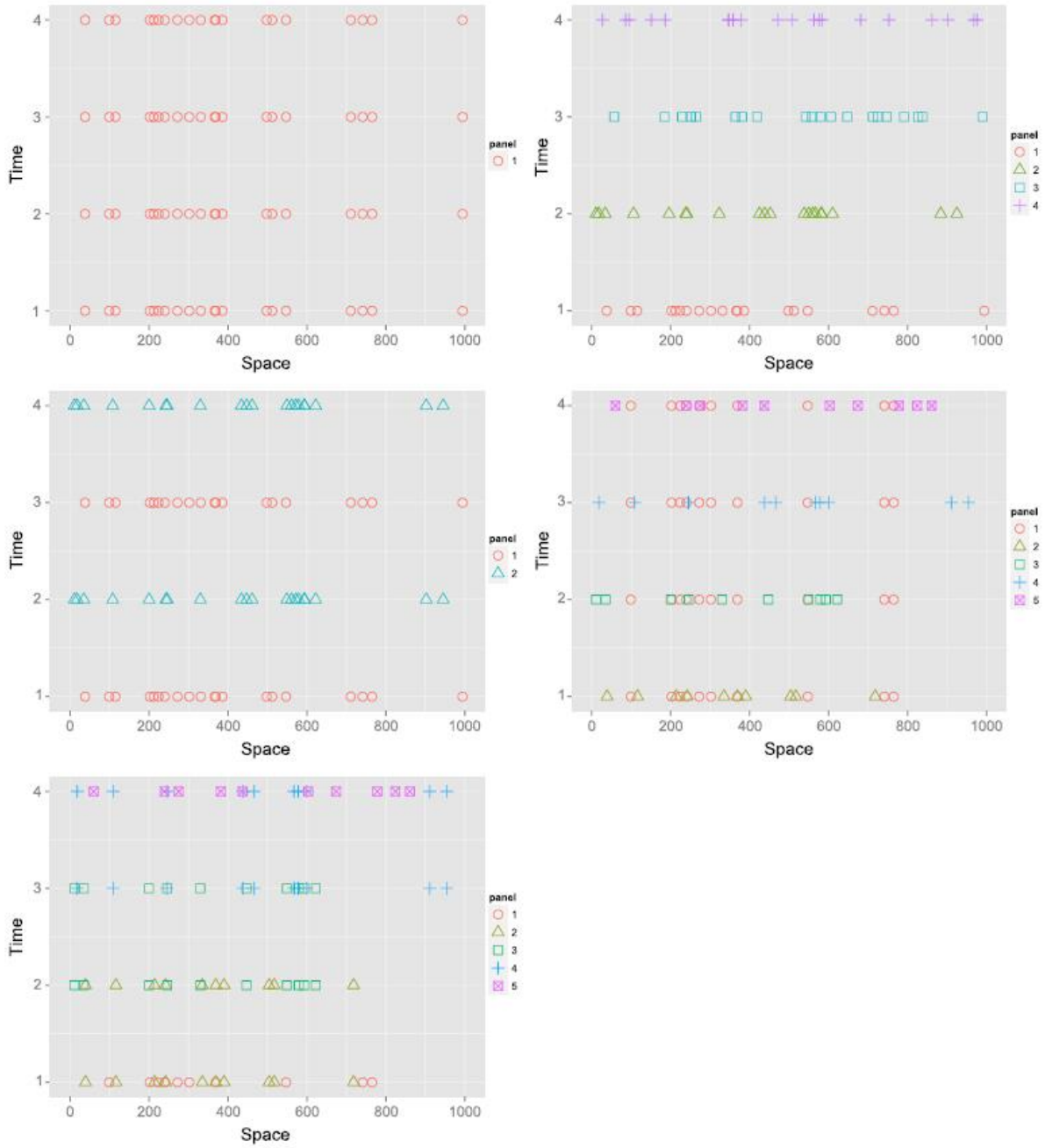

Figure 24. Notional examples of five space-time designs. Static synchronous (top left), (independent) synchronous (top right), serially alternating (middle left), supplemented panel (middle right), rotational (bottom) (Reproduced from Brus and De Gruijter, 2011).

The simplest and most common method of sampling in time is systematic sampling; this means sampling at constant time-intervals (De Gruijter et al., 2006). Systematic sampling requires to choose an appropriate sampling frequency. Sampling at constant time intervals has operational advantages and the methods used to analyze are mathematically relatively simple (De Gruijter et al., 2006). However, De Gruijter et al. (2006) warn that systematic sampling in time is not always

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

the best option. For instance, if temporal variation varies with time then it is more efficient to sample more densely in periods with larger variation, i.e., one might then want to stratify over time. Another problem might be if there is periodicity in time and sampling frequency coincides with the period. For instance, to estimate the monthly average of the topsoil temperature it would be unwise to measure once a day, always at the same time.

Considerations for designing a monitoring scheme

Flexibility of the scheme

De Gruijter et al. (2006) stress an important difference between sampling for estimating global and local quantities and mapping on the one hand and sampling for monitoring on the other hand. The former typically takes place in a relatively short period of time during which the universe of interest as well as the operational and budgetary aspects do not change. With monitoring not only the universe may undergo (large, unexpected) changes but also the sampling conditions, especially in long-term monitoring. This makes that adaptation of a (long-term) monitoring scheme are inevitable or at least desirable. Budgets may vary from year to year and operational constraints that were initially present might disappear or vice versa. Other changes that might occur over time and affect the monitoring scheme are the definition of new objectives (e.g. target variables, domains of interest) or the availability of new measurement techniques. Finally, more and more data become available about the universe of interest that can give good reason to fine-tune or redesign the scheme (De Gruijter et al. 2006). To be able to adapt to (unforeseen) changes in conditions calls for flexibility of the monitoring scheme.

Overton and Stehman (1995) discuss design implications and how the choice of a design is influenced by intended users of the data, whereas Overton and Stehman (1996) elaborately discuss desirable design characteristics for long-term monitoring of ecological variables and ways to adapt the design to changing conditions, including sample restructuring, changing the size of a sample (increase or reduce) and post-stratification. These authors stress the importance of keeping a design as simple as possible. Overton and Stehman (1995) note that complex sampling designs require complex computational formula for statistical inference. By keeping a design simple, standard statistical analyses commonly used by environmental scientists are correct or at least adequate. Furthermore, they note that the primary concern when designing a monitoring scheme should be to develop an adequate design that makes good use of the available resources and not to construct the perfect, optimal design. Practical convenience and simplicity cannot be sacrificed to achieve optimal statistical efficiency (Overton and Stehman, 1995).

It is a misconception that for monitoring global quantities (means, totals, fractions) in space-time one should always return to the same sampling location during each sampling cycle. For estimating the change of the mean, total or areal fraction from one sampling time to another revisiting all sampling locations is the best option. However, as will be explained hereafter, for other space-time quantities it can be more efficient to revisit a subset of the sampling locations only. As we have seen in the previous sections and Figure 24, different sets of sampling locations can be used in different sampling rounds such as in the synchronous design, in which every time all locations are replaced by new locations, or as in the supplemented panel or rotational design in which locations are partially replaced by new locations. The fact that one could select new sampling locations for each sampling cycle gives flexibility to the scheme. Moreover we will not always be able to go back to the exact location, for instance when the site is used for construction or when access permission by the land owner is denied. However, if one does revisit (a selection of) sampling sites in subsequent sampling rounds, as will be the case in most sampling designs, see Figure 24, and a site cannot be accessed or sampled anymore, then this does not mean the statistical integrity of

European Commission | Horizon 2020 European Union funding for Research & Innovation

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

SOILS 4 AFRICA

the monitoring scheme is compromised. Statistical inference still can yield valid and unbiased estimates of global quantities.

Besides 'ad-hoc' changes in sample size as a result of accessibility issues, a more structural reduction in sample size can be required, for instance after a budget cut. Reducing the number of sampling locations to be revisited is straightforward when the existing sample is selected with probability sampling. One can simply take a probability subsample of the original sample since a probability subsample from a probability sample is itself a probability sample of the sampling universe (Overton and Stehman, 1996). Design-based unbiased and valid estimators are still available, but it is evident that these estimators are less precise (have larger variance). Like reduction, increasing the size of a probability sample is in most cases straightforward (Overton and Stehman, 1996).

When designing a scheme for long-term monitoring one must thus anticipate that adaptions will be required over time. Some smart design choices at the beginning can increase the robustness of a design and greatly ease adaptions later on. For instance, choosing a *design type* that allows straightforward adjustment of sample size is of utmost importance. Changing sample size for a systematic random sample is less straightforward than that of a simple, stratified or two-stage cluster random sample (Overton and Stehman, 1996). Another simple measure one can take in case of a stratified design is to ensure that a large enough number of sampling locations is selected from each stratum so that a possible reduction of sampling size in the future is less likely to compromise the sample structure. The minimum sample size for a stratum is two (to be able to estimate the stratum variance). By choosing a sample size per stratum large enough at the beginning, one can ensure that after a (substantial) budget cut at least two (permanent) sample sites remain in each stratum. When designing a sampling scheme this could imply that fewer strata must be defined than initially intended, leading to larger stratum sample sizes that can be reduced when the budget is reduced. Fewer, larger strata may yield less precise results at the original budget but the expected loss of initial precision may be less important than greater adaptability to changing budgets (De Gruijter et al., 2006).

While designing the Soils4Africa sampling scheme, we have attempted to keep the design as simple as possible given the boundary conditions. We have used one thematic stratification variable (farming systems) with a limited number of classes (per country). Statistically valid and unbiased estimates of target parameters can be obtained at the country level that can subsequently be pooled in a combined estimator for the continent. This means that the design can be adjusted at country level without affecting the design and estimation in other countries. Adjustments can include sample sizes – total sample size, strata sample sizes of the PUs and the number of selected SUs per PU – without consequences for the statistical inference. But also the stratification can be refined. If a country can permit a larger sampling size in a next sampling round, then a more refined stratification variable that might be better suited to a country's condition and reporting needs can replace the stratification variable used here. Compatibility with the proposed Soils4Africa design is not compromised as long as the *inclusion probabilities* of each of the newly selected sampling locations in the adapted scheme are known. This will also allow combining the data collected with different probability sampling designs (see for instance Grafström et al., 2019).

Operational aspects

Here we consider some statistical and operational aspects of choosing a space-time design for monitoring as summarized from De Gruijter et al. (2006):

- *Static* and *static-synchronous* patterns are attractive when the costs of repeated sampling at the same location are lower than for sampling at different locations with the same total sample size. This is for instance when semi-permanent measuring equipment is installed at fixed locations. A statistical disadvantage of these two patterns is that only information on

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

temporal variation increases and not that on spatial variation, in contrast to synchronous and rotational patterns (incl. serially alternating and supplemental panel). An advantage of static over static-synchronous is that sampling times may be adapted to local circumstances (e.g. sampling at a time when operational costs are lowest), while static-synchronous patterns reduce the number of sampling times given the sample size compared to static patterns. This means that if costs for an additional sampling time are larger than for sampling additional locations at a given sample time, it can be attractive to reduce the number of sampling times and increase the number of sampling locations per sampling time. This enables more locations to be sampled in total, yielding more accurate estimates of spatio-temporal global quantities. A static pattern is attractive when considerable spatial variation between time series is known to exist. The French soil monitoring network is an example of a static-synchronous pattern.

- *Synchronous* patterns are much more flexible than static and static-synchronous patterns. With each sampling time one is free to choose a spatial or change the existing design (e.g. sample size, possible stratification, clustering) to adapt to the circumstances at the sampling time. Statistical inference from a synchronous sample is much simpler compared to static, static-synchronous and rotational samples since no sampling locations from earlier sampling times are revisited.
- *Rotational* patterns are more flexible and have better spatial coverage than static and static-synchronous patterns. If there is a fair amount of correlations between observations at consecutive sampling times, then rotational patterns are more efficient in estimating current means, totals and fractions as well as temporal trends than synchronous patterns. A disadvantage of rotational patterns is that design-based inference of the sampling variance is somewhat more complicated.
- *Supplemented panel* designs have the same advantages as a rotational design. Besides supplemented panel has the advantage of operational simplicity: a subset of locations is fixed (the pure panel subset), whereas others are swarming. Besides, at the fixed locations a time series of data is obtained.

Suitability of sampling patterns for monitoring

Finally, we briefly consider suitability of the four main sampling patterns for estimating global quantities (see bulleted list in the introduction paragraph of Section 4.2.3). We limit ourselves here to the 'current global quantities' and 'change of the spatial mean', and the 'temporal trend of the spatial mean'. The following summary is adapted from De Gruijter et al. (2006):

- *Estimation of the current means, totals, fractions*: rotational or supplemented panel designs are preferred over synchronous or static-synchronous designs. Because sampling locations partially overlap in these designs, these designs can exploit information from previous sampling times. Synchronous designs do not have such overlap and there is no simple way to use information from a previous sample to estimate a current quantity. Static-synchronous designs have fully overlapping sample locations (use one panel only) so there is no additional information from in the measurements taken in a previous sampling round.
- *Estimation of the change of means, totals and fractions:* Static-synchronous designs are more efficient than synchronous designs (though the latter are suitable for this purpose as well) because one profits most from the correlation of the two sample sets measured at the two times.
- *Temporal trend of the spatial mean*. With more than two sampling times, a supplemented panel design can be more efficient to estimate the average change of a global quantity per time unit than a static-synchronous design. This depends on how persistent the spatial patterns of the soil variables of interest are (Brus and de Gruijter, 2013). When the spatial patterns do not change much over time (strong persistence) a supplemented panel design

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

yields relative precise estimates of a linear time trend of the population mean, whereas for highly dynamic variables resulting in large changes in the spatial pattern over time (e.g., locations with high values at a given time may well have small values at a subsequent time) a synchronous or serially alternating design is the best choice. For moderate persistence the choice is more complicated; supplemented panel or serially alternating designs are good choices in this case (Brus and De Gruijter, 2013).

European Commission | Horizon 2020 European Union funding for Research & Innovation

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

SOILS4AFRICA

Acknowledgements

The authors thank to Julius Buyengo and Eunice Wangui of RCMRD for providing and processing some of the ancillary datasets. The authors are grateful to Harrij van Velthuizen of IIASA and Christopher Auricht of Auricht Projects for their help with obtaining the updated farming systems and agro-ecological characteristics layers. We also express our gratitude to Zhanguo Bai, Maria Ruiperez-Gonzalez and Laura Poggio of ISRIC-World Soil Information for their support with the data processing. Finally, we thank project colleagues from Stellenbosch University, BUNASOLS, KALRO, IRD-Tunisia and IFA-Yangambi for sharing their thoughts and feedback during the various workshops.

European Commission | Horizon 2020 European Union funding for Research & Innovation

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

SOILS 4 AFRICA

References

Ballin, M., Barcaroli, G., Masselli, M., and Scarnó, M. 2018. Redesign sample for Land Use/Cover Area frame Survey (LUCAS) 2018. Statistical Working Papers. Luxembourg: Publications Office of the European Union. Available at: https://ec.europa.eu/eurostat/documents/3888793/9347564/KS-TC-18-006-EN-N.pdf/26cbdb1a-c418-4dfa-bc24-a27d1bc5599c

Batjes, N.H., Ribeiro, E., 2021. Inventory of soil profile data for Africa represented in the ISRIC-WDC holdings., Deliverable 3.2b. Soils4Africa project. European Union's Horizon 2020 research and innovation programme grant agreement No 869200.

Brus, D.J., De Gruijter, J.J., 1997. Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). Geoderma 80(1-2), 1-44.

Brus, D.J., 2000. Using regression models in design-based estimation of spatial means of soil properties. European Journal of Soil Science 51(1), 159-172.

Brus, D.J., Knotters, M., Metzger, M.J., and Walvoort, D.J.J., 2011. Towards a European-wide sampling design for statistical monitoring of common habitats, Wageningen, Alterra, Alterra Report 2213. Available at: https://www.wur.nl/upload_mm/4/0/9/3c60c015-861e-46cc-85fc-b54d969f3f6a_EBONED33TowardsaEuropeanwidesamplingdesign.pdf

Brus, D.J., and de Gruijter, J.J., 2011, Design-based Generalized Least Squares estimation of status and trend of soil properties from monitoring data, Geoderma, 164, 172-180.

Brus, D.J. and de Gruijter, J.J., 2012, A hybrid design-based and model-based sampling approach to estimate the temporal trend of spatial means, Geoderma, 173-174, 241-248.

Brus, D.J., and de Gruijter, J.J., 2013, Effects of spatial pattern persistence on the performance of sampling designs for regional trend monitoring analyzed by simulation of space-time fields, Computers & Geosciences, 61, 175-183.

Brus, D.J., 2014, Statistical sampling approaches for soil monitoring, European Journal of Soil Science, 65, 779-791.

Brus, D.J., 2015. Balanced sampling: A versatile sampling approach for statistical soil surveys. Geoderma 253-254, 111-121. doi: 10.1016/j.geoderma.2015.04.009

Brus, D.J., 2019. Sampling for digital soil mapping: A tutorial supported by R scripts. Geoderma 338, 464-480.

Brus, D.J., 2021. Statistical approaches for spatial sample survey: Persistent misconceptions and new developments. European Journal of Soil Science 72, 686-703. doi: 10.1111/ejss.12988

Brus, D.J., *in prep*. Spatial Sampling with R. CRC Press

Buchhorn, M, Smets, B., Bertels, L., De Roo., B., Lesiv, M., Tsendbazar, N-E., Herold., M., Fritz, S., 2020. Copernicus Global Land Service: Land Cover 100m: collection 3: epoch 2019: Globe (Version V3.0.1) Data set: http://doi.org/10.5281/zenodo.3939050

Center for International Earth Science Information Network - CIESIN - Columbia University. 2018. Gridded Population of the World, Version 4 (GPWv4): Population Density, Revision 11. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). Available at: https://doi.org/10.7927/H49C6VHW.

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

De Gruijter, J.J., Brus, D.J., Bierkens, M.F.P., Knotters, M., 2006. Sampling for natural resource monitoring. Springer.

Dixon, J., Gulliver, A., Gibbon, D., Hall, M., 2001. Farming systems and poverty : improving farmers' livelihoods in a changing world. FAO/World Bank, Rome/Washington, D.C. Available at: http://www.fao.org/3/y1860e/y1860e00.htm. Dataset: http://www.fao.org/geonetwork/srv/en/metadata.show?id=1029&currTab=simple

Dixon, J., Garrity, D. P., Boffa JM., Williams, T.O., Amede, T., Auricht, C., Lott, R., Mburati, G., (Eds.) 2019. Farming systems and Food Security in Africa: Priorities for Science and Policy under Global Change (1st ed.). Routledge. D.C. doi: 10.4324/9781315658841

Domburg, P., de Gruijter, J.J., van Beek, P., 1997. Designing efficient soil survey schemes with a knowledge-based system using dynamic programming. Geoderma 75(3), 183-201.

Eurostat, 2018. LUCAS 2018 (Land Use / Cover Area Frame Survey) Technical reference document S1 Stratification Guidelines. Eurostat Technical Documents. Luxembourg: Eurostat, European Commission. Available at: https://ec.europa.eu/eurostat/documents/205002/7329820/LUCAS2018_S1-StratificationGuidelines_20160523.pdf

Falorsi, P.D., P. Righi, 2008. A balanced sampling approach for multi-way stratification designs for small area estimation. Survey Methodology 34(2), 223-234. Available at: https://www.istat.it/en/files/2016/10/Falorsi-engSURVEY_METH.pdf

FAO, 2014. Global Administrative Unit Layers (GAUL). Available at: http://www.fao.org/geonetwork/srv/en/metadata.show?id=12691

Fatunbi, O.A., Abishek, A., 2020. D2.1: A set of use cases plus supporting soil quality indicators, Deliverable 2.1. Soils4Africa project. European Union's Horizon 2020 research and innovation programme grant agreement No 869200.

Fischer, G., Nachtergaele, F.O., Prieler, S., Teixeira, E., Tóth, G., van Velthuizen, H., Verelst, L., Wiberg, D., 2012. Global Agro-Ecological Zones (GAEZ v3) Model Documentation. IIASA, Laxenburg, Austria and FAO, Rome, Italy, 196 pp. Available at: http://pure.iiasa.ac.at/id/eprint/13290/1/GAEZ_Model_Documentation.pdf

Gallego F.J., and Delincé, J. 2010. The European Land Use and Cover Area-frame statistical Survey (LUCAS). In: Benedetti, R., Bee, M., Espa, G., Piersimoni, F., (Eds.), Agricultural Survey Methods, pp. 151-168. John Wiley & sons, New York. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470665480.ch10

Grafström, A. Lisic, J., 2018. BalancedSampling: Balanced and Spatially Balanced Sampling. R package version 1.5.4. URL http://www.antongrafstrom.se/balancedsampling

Grafström, A., Lundström, N., & Schelin, L., 2012. Spatially Balanced Sampling through the Pivotal Method. Biometrics, 68(2), 514-520. doi: 10.1111/j.1541-0420.2011.01699.x

Grafström, A., Ekström, M., Jonsson, B.G., Esseen, P-A., Ståhl, G., 2019. On combining independent probability samples. Survey Methodology, 45(2), 349-364. Available at: http://umu.diva-portal.org/smash/get/diva2:1338331/FULLTEXT01.pdf

Hengl, T., Mendes de Jesus, J., Heuvelink, G.B.M., Ruiperez-Gonzalez, M., Kilibarda, M., Blagotic, A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B., Guevara, M.A., Vargas, R., MacMillan, R.A., Batjes, N.H., Leenaars, J.G.B., Ribeiro, E., Wheeler, I., Mantel, S., Kempen, B.,

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

2017. SoilGrids250m: Global Gridded Soil Informatin Based on Machine Learning. PLoS ONE 12(2): e0169748. https://doi.org/10.1371/journal.pone.0169748.

Heuvelink, G.B.M., Angelini, M.E., Poggio, L., Bai, Z., Batjes, N.H., van den Bosch, R., Bossio, D., Estella, S., Lehmann, J., Olmedo, G.F., Sanderman, J., 2020. Machine learning in space and time for modelling soil organic carbon change. European Journal of Soil Science, 1-17. doi: 10.1111/ejss.12998

Huising, E.J., Wangui Mwangi, E., Buyengo, J., 2020. D4.1: Map of agricultural land of continental Africa, Soils4Africa project. European Union's Horizon 2020 research and innovation programme grant agreement No 869200.

IIASA/FAO, 2012. Global Agro-Ecological Zones (GAEZ v3.0). IIASA, Laxenburg, Austria and FAO, Rome, Italy. Dataset: https://www.gaez.iiasa.ac.at/

IUSS Working Group WRB, 2006. World Reference Base for Soil Resources 2006. World Soil Resources Reports No. 103. FAO, Rome

Jones, A., Breuning-Madsen, H., Brossard, M., Dampha, A., Deckers, J., Dewitte, O., Gallali, T., Hallett, S., Jones, R., Kilasara, M., Le Roux, P., Michéli, E., Montanarella, L., Spaargaren, O., Thiombiano, L., Van Ranst, E., Yemefack, M., Zougmore, R., (eds.), 2013. Soil Atlas of Africa. European Commission, Publications Office of the European Union, Luxembourg. 176 pp. ISBN 978-92-79-26715-4, doi 10.2788/52319

Kempen, B., Dalsgaard, S., Kaaya, A.K., Chamuya, N., Ruipérez-González, M., Pekkarinen, A., Walsh, M.G., 2019. Mapping topsoil organic carbon concentrations and stocks for Tanzania. Geoderma 337, 164-180.

Lohr, S. L., 1999. Sampling: Design and Analysis. Duxbury Press, Pacific Grove, USA.

Ma, T., Brus, D.J., Zhu, A.X., Zhang, L., Scholten, T., 2020. Comparison of conditioned Latin hypercube and feature space coverage sampling for predicting soil classes using simulation from soil maps. Geoderma 370, 114366.

Minasny, B., McBratney, A.B., 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Computers & Geosciences 32(9), 1378-1388.

Moinet, G., Creamer, R., Leenaars, J.G.B., 2021. Methods for deriving selected soil quality indicators, Deliverable 3.1. Soils4Africa project. European Union's Horizon 2020 research and innovation programme grant agreement No 869200.

Nelson, A., 2008. Estimated travel time to the nearest city of 50,000 or more people in year 2000. Global Environment Monitoring Unit - Joint Research Centre of the European Commission, Ispra Italy. Available at https://forobs.jrc.ec.europa.eu/products/gam/

Orgiazzi, A., Ballabio, C., Panagos, P., Jones, A., and Fernández-Ugalde, O, 2018. LUCAS Soil, the largest expandable soil dataset for Europe: a review. European Journal of Soil Science 69: 140–53. doi: 10.1111/ejss.12499

Overton, W.S. and Stehman, S.V., 1995. Design implications of anticipated data uses for comprehensive environmental monitoring programmes, Environmental and Ecological Statistics, 2 , 287-303.

Overton, W.S. and Stehman, S.V., 1996. Desirable design characteristics for long-term monitoring of ecological variables, Environmental and Ecological Statistics, 3, 349-361.

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

R Core Team, 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Särndal, C. E., Swensson, B., and Wretman, J., 1992. Model Assisted Survey Sampling. Springer, New York.

Sebastian, K., 2009. Agro-ecological Zones of Africa. IFPRI. Harvard Dataverse, V2. Dataset: https://doi.org/10.7910/DVN/HJYYTI,

Teuling, K., Kempen, B., Knotters, M., Saby, N., Brus, D.J., Vašát, R., van Egmond, F. and Bispo, A., 2021. Chapter 4 Sampling theory for mapping and monitoring purposes in 'D6.1 Towards climate-smart sustainable management of agricultural soils' eds. F.M. van Egmond and M. Fantappiè. EJP SOIL project. European Union's Horizon 2020 research and innovation programme grant agreement No 862695 (*under review*).

Tillé, Y. and Matei, A., 2021. sampling: Survey Sampling. R package version 2.9. URL https://cran.r-project.org/web/packages/sampling/index.html

Tóth, G., Jones, A. and Montanarella, L (eds), 2013a. LUCAS Topsoil Survey. Methodology, data and results. Report EUR 26102 EN. Luxembourg: Publications Office of the European Union. Available at: https://esdac.jrc.ec.europa.eu/ESDB_Archive/eusoils_docs/other/EUR26102EN.pdf

Tóth, G., Jones, A. and Montanarella, L, 2013b. The LUCAS topsoil database and derived information on the regional variability of cropland topsoil properties in the European Union. Environmental Monitoring and Assessment 185:7409–7425. doi: 10.1007/s10661-013-3109-3

UNEP-WCMC (2021). Protected Area Profile for Africa from the World Database of Protected Areas. Available at: www.protectedplanet.net

Vågen, T.-G., Sheperd, K.D., Walsh, M.G., Winowiecki, L., Desta, L.T., Tondoh, J.E., 2010. AfSIS Technical Specifications - Soil Health Surveillance, World Agroforestry Centre, Nairobi, Kenya.

Vågen, T.-G., Winowiecki, L., 2020. the Land Degradation Surveillance Framework (LDSF) Field Guide v2020, World Agroforestry Centre (ICRAF), Nairobi, Kenya.

Van Velthuizen, H., Fischer, G., Hizsnyik, E., 2013. The African Farming Systems Update Project. Farming systems and Food Security in Africa: Priorities for Science and Policy under Global Change. IIASA, Laxenburg, Austria. 76 p. Available at: http://pure.iiasa.ac.at/id/eprint/15771/1/The%20African%20Farming%20Systems.pdf

European Commission | Horizon 2020 European Union funding for Research & Innovation

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

SOILS4AFRICA

ANNEX

Annex 1: Farming system classes

Following the updated classification of Van Velthuizen et al., (2013), Dixon et al. (2019).

Level 1	L1 description	Level2	L2 description
1	Maize mixed	11	Ethiopian maize-livestock
1	Maize mixed	12	Kenyan-Ugandan hoe-tractor maize mixed
1	Maize mixed	13	Tanzanian semi-mechanized maize mixed
1	Maize mixed	14	Malawian market-linked maize mixed
1	Maize mixed	15	Central African scattered maize-root crop
1	Maize mixed	16	Mozambiquan extensive maize mixed
1	Maize mixed	17	Southern African dualistic maize mixed
1	Maize mixed	18	Zambian medium population density maize mixed
1	Maize mixed	19	Madagascan maize-rice
2	Agropastoral	21	Sahelian agropastoral
2	Agropastoral	22	East African agropastoral
2	Agropastoral	23	Southern African agropastoral
2	Agropastoral	24	North African dryland mixed
3	Highland perennial	31	Southern Ethiopian highland perennial
3	Highland perennial	32	Albertine Rift highland perennial
3	Highland perennial	33	Northern Tanzanian highland perennial
3	Highland perennial	34	Southern Tanzanian highland perennial
3	Highland perennial	35	Central Kenyan highland perennial
3	Highland perennial	36	Western Kenyan highland perennial
4	Root and tuber crop	41	Yam-cassava
4	Root and tuber crop	42	Cassava-yam-cocoyam
4	Root and tuber crop	43	Cassava-cocoyam
4	Root and tuber crop	44	Cassava-sweet potato-potato
4	Root and tuber crop	45	Cassava
5	Cereal-root crop mixed	51	Cereal-pulse-root crop
5	Cereal-root crop mixed	52	Root crop-cereal
6	Highland mixed	61	Highland livestock-cereal
6	Highland mixed	62	Highland wheat-pulse
6	Highland mixed	63	Highland maize-teff
6	Highland mixed	64	North African highland mixed
7	Humid lowland tree crop	70	Humid lowland tree crop
8	Pastoral	81	Sahelian pastoral
8	Pastoral	82	East African pastoral
8	Pastoral	83	Southern African pastoral
8	Pastoral	84	North African pastoral
9	Fish-based	90	Fish-based
10	Forest-based	100	Forest-based
11	Large-scale irrigated	110	Large-scale irrigated
12	Perennial mixed	121	Grape and deciduous fruit perennial mixed
12	Perennial mixed	122	Sugarcane and production forest perennial mixed
12	Perennial mixed	123	Rainfed perennial mixed
13	Arid pastoral oasis	130	Arid pastoral oasis

European Commission | Horizon 2020 European Union funding for Research & Innovation

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

SOILS4AFRICA

Annex 2: Areas classified as 'unsafe' for soil sampling

Country	Region		Country	Region
Algeria	Adrar		Mozambique	Cabo Delgado
Algeria	Bechar		Niger	Agadez
Algeria	Ghardaia		Niger	Diffa
Algeria	Illizi		Niger	Niamey
Algeria	Ouargla		Niger	Tahoua
Algeria	Tamanrasset		Niger	Tillaberi
Algeria	Tindouf		Nigeria	Borno
Burkina Faso	Yatenga		Nigeria	Yobe
Burkina Faso	Lorum		Nigeria	Adamawa
Burkina Faso	Bam		Somalia	Galgaduud
Burkina Faso	Sanematenga		Somalia	Hiraan
Burkina Faso	Soum		Somalia	Bakool
Burkina Faso	Seno		Somalia	Gedo
Burkina Faso	Oudalan		Somalia	Bay
Burkina Faso	Yagha		Somalia	Shabelle Dhexe
Burkina Faso	Tapoa		Somalia	Shabelle Hoose
Burkina Faso	Kompienga		Somalia	Banadir
Burkina Faso	Komondjari		Somalia	Juba Dhexe
Cameroon	Extreme-Nord		Somalia	Juba Hoose
Central African Republic			South Sudan	Warab
Chad	Tibesti		South Sudan	El Buheyrat
Chad	Borkou		Sudan	Northern Darfur
Chad	Ennedi		Sudan	Western Darfur
Chad	Lac		Sudan	Southern Darfur
Chad	Hadjer Lamis		Uganda	Kaabong
DR Congo	Ituri		Uganda	Lamwo
DR Congo	Equateur		Uganda	Yumbe
DR Congo	Tshuapa		Uganda	Moyo
DR Congo	Nord-Kivu		Uganda	Kitgum
DR Congo	Sud-Kivu		Uganda	Koboko
DR Congo	Sankuru		Uganda	Adjumani
DR Congo	Kasai-Central		Uganda	Amuru
Ethiopia	Tigray		Uganda	Maracha
Kenya	Mandera		Uganda	Arua
Kenya	Wajir		Uganda	Kotido
Kenya	West Pokot		Uganda	Abim
Kenya	Samburu		Uganda	Moroto
Kenya	Isiolo		Uganda	Zombo
Kenya	Baringo		Uganda	Napak
Kenya	Elgeyo-Marakwet		Uganda	Nakapiripirit
Kenya	Garissa		Uganda	Buliisa
Kenya	Lamu		Uganda	Amudat
Mali	Kidal		Uganda	Ntoroko
Mali	Gao		Uganda	Bundibugyo
Mali	Tombouctou		Uganda	Kanungu
Mali	Mopti		Uganda	Kisoro
Mauritania	Hodh Ech Chargi			

European Commission | Horizon 2020 European Union funding for Research & Innovation

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

SOILS4AFRICA

Annex 3: Protected area classification

Type of protected area	Sampling
Ancient Monument	No
Area of Outstanding Natural Beauty	Yes
Biological Reserve	No
Bird Reserve	No
Botanical Reserve	No
Conservation Area	Yes
Ecological Park	Yes
Faunal Reserve	Cultivated land only
Fishery Reserve	No (outside sampling universe)
Forest Plantation	Yes
Forest Reserve	Cultivated land only
Geological Protected Area	No
Hunting Area	Yes
Marine Reserve	No
Mountain Catchment Area	Yes
National Park	No
Nature Reserve	Cultivated land only
Other Protected Area	No
Private Nature Reserve	No
Private Ranch	No
Protected Areas of Mediterranean	Yes
UNESCO-MAB Biosphere Reserve	Cultivated land only
Wetland Reserve	Cultivated land only
Wildlife Reserve	Cultivated land only
World Heritage Site (natural or mixed)	Cultivated land only

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

Annex 4: Area of agricultural land and number of PSUs based on proportional allocation per country, land cover class and soil class

country	area (km2)	area (%)	psu		country	area (km2)	area (%)	psu
Algeria	114,829	1.7	84		Nigeria	563,879	8.3	414
Angola	520,923	7.6	382		Angola	520,923	7.6	382
Benin	75,228	1.1	56		DR Congo	495,072	7.3	363
Botswana	62,118	0.9	46		Ethiopia	443,500	6.5	325
Burkina Faso	128,934	1.9	95		Tanzania	442,520	6.5	325
Burundi	20,909	0.3	16		Mozambique	429,522	6.3	315
Cameroon	138,109	2.0	102		Zambia	320,398	4.7	235
Central African	207,880	3.1	153		South Sudan	300,566	4.4	221
Chad	127,607	1.9	94		Madagascar	241,778	3.6	178
Congo	40,174	0.6	30		Sudan	228,150	3.4	168
Cote d'Ivoire	179,978	2.6	132		Zimbabwe	213,766	3.1	157
DR Congo	495,072	7.3	363		Central African	207,880	3.1	153
Djibouti	50	0.0	0		South Africa	180,904	2.7	133
Egypt	42,488	0.6	31		Cote d'Ivoire	179,978	2.6	132
Equatorial Guinea	536	0.0	1		Mali	149,076	2.2	110
Eritrea	11,054	0.2	8		Kenya	142,361	2.1	105
Eswatini	10,653	0.2	8		Cameroon	138,109	2.0	102
Ethiopia	443,500	6.5	325		Guinea	136,072	2.0	100
Gabon	10,518	0.2	8		Ghana	131,636	1.9	97
Gambia	6,237	0.1	5		Burkina Faso	128,934	1.9	95
Ghana	131,636	1.9	97		Chad	127,607	1.9	94
Guinea	136,072	2.0	100		Uganda	121,533	1.8	89
Guinea-Bissau	16,756	0.3	13		Algeria	114,829	1.7	84
Kenya	142,361	2.1	105		Morocco	105,232	1.5	77
Lesotho	2,913	0.0	2		Niger	88,122	1.3	65
Liberia	8,829	0.1	7		Benin	75,228	1.1	56
Libya	25,629	0.4	19		Senegal	65,644	1.0	48
Madagascar	241,778	3.6	178		Botswana	62,118	0.9	46
Malawi	56,408	0.8	42		Malawi	56,408	0.8	42
Mali	149,076	2.2	110		Namibia	44,363	0.7	33
Mauritania	3,745	0.1	3		Egypt	42,488	0.6	31
Morocco	105,232	1.5	77		Congo	40,174	0.6	30
Mozambique	429,522	6.3	315		Sierra Leone	38,918	0.6	29
Namibia	44,363	0.7	33		Togo	38,620	0.6	29
Niger	88,122	1.3	65		Tunisia	36,597	0.5	27
Nigeria	563,879	8.3	414		Somalia	29,368	0.4	22
Rwanda	19,168	0.3	15		Libya	25,629	0.4	19
Senegal	65,644	1.0	48		Burundi	20,909	0.3	16
Sierra Leone	38,918	0.6	29		Rwanda	19,168	0.3	15
Somalia	29,368	0.4	22		Guinea-Bissau	16,756	0.3	13
South Africa	180,904	2.7	133		Eritrea	11,054	0.2	8
South Sudan	300,566	4.4	221		Eswatini	10,653	0.2	8
Sudan	228,150	3.4	168		Gabon	10,518	0.2	8
Tanzania	442,520	6.5	325		Liberia	8,829	0.1	7
Togo	38,620	0.6	29		Gambia	6,237	0.1	5
Tunisia	36,597	0.5	27		Mauritania	3,745	0.1	3
Uganda	121,533	1.8	89		Lesotho	2,913	0.0	2
Zambia	320,398	4.7	235		Equatorial Guinea	536	0.0	1
Zimbabwe	213,766	3.1	157		Djibouti	50	0.0	0

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

Land cover	**area (km2)**	**area (%)**	**psu**
Open forest	3,054,948	44.8	2240
Cultivated	2,471,727	36.3	1813
Herbaceous vegetation	690,752	10.1	507
Closed forest	374,733	5.5	275
Shrubs	227,109	3.3	167

WRB reference soil group*	**area (km2)**	**area (%)**	**psu**
Ferralsol	875,245	12.8	642
Arenosol/Podzol	846,192	12.4	621
Lixisol	655,362	9.6	481
Plinthosol	652,968	9.6	479
Leptosol	605,582	8.9	444
Luvisol	454,253	6.7	334
Vertisol	448,667	6.6	329
Cambisol	446,870	6.6	328
Acrisol	406,997	6.0	299
Nitisol	336,020	4.9	247
Fluvisol	246,326	3.6	181
Calcisol/Gypsisol/Durisol	180,576	2.7	133
Regosol	130,560	1.9	96
Gleysol	124,647	1.8	92
Solonetz/Solonchack	107,029	1.6	79
Alisol	101,856	1.5	75
Phaezem/Chernozem/Kastanozem	95,605	1.4	70
Stagnosol/Planosol	66,029	1.0	49
Andosol/Umbrisol	29,656	0.4	22
Histosol	8,826	0.1	7

European Commission | Horizon 2020 European Union funding for Research & Innovation

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

SOILS4AFRICA

Annex 5: Sampling frame of the PSUs

The figure below shows an example of the PSU sampling frame.

id	easting	northing	ag_count	admin_0	admin_1	aez	landcover	soil_cl	accsblty	pop_dens	lgp	elevation	twi	agrecol	fs2001	fs16lev1	fs16lev2	pa1	pa2	pa3	protected	sampling	safe
295781	1538612	4156569	31	248	2995	21	126	1	163	163	221	10	112	10	15	12	123	NA	NA	NA	4	1	1
295782	1540612	4156569	100	248	2995	21	126	1	182	163	220	32	108	10	15	12	123	NA	NA	NA	4	1	1
295783	1542612	4156569	104	248	2995	21	126	1	178	163	220	63	105	10	15	12	123	NA	NA	NA	4	1	1
295784	1544612	4156569	98	248	2995	21	126	1	126	163	220	27	110	10	15	12	123	NA	NA	NA	4	1	1
295785	1546612	4156569	220	248	2995	21	126	1	86	1970	220	14	112	10	15	12	123	NA	NA	NA	4	1	1
295786	1548612	4156569	93	248	2995	21	40	1	46	1970	218	6	111	10	15	12	123	NA	NA	NA	4	1	1
295788	1552612	4156569	25	248	2995	21	116	1	74	1970	218	40	107	10	15	12	123	NA	NA	NA	4	1	1
295789	1554612	4156569	64	248	2995	21	126	1	72	1970	218	80	100	10	15	12	123	NA	NA	NA	4	1	1
295790	1556612	4156569	32	248	2995	21	126	1	67	1970	217	24	106	10	15	12	123	NA	NA	NA	4	1	1

The following table gives a description of each variable. References are provided in section 'Ancillary data' in Chapter 4. The numbers of the categorical variables represent a specific class. Class values of each PSU are determined based on a majority vote considering all agricultural land pixels contained within the PSU; values of the continuous variables (excl. easting, northing, ag_count) are means calculated from the agricultural land pixels contained within the PSU (also see section 'Stratification of the primary sampling units' in Chapter 4).

Variable	**Description**	**Type**
id	unique identifier	continuous
easting	longitude (m)	continuous
northing	latitude (m)	continuous
ag_count	number of agricultural land pixels within PSU	continuous
admin_0	country	categorical
admin_1	level 1 administrative unit	categorical
aez	agro-ecological zone	categorical
landcover	land cover class	categorical
soil_cl	soil class	categorical
accsblty	accessibility (travel time; min)	continuous
pop_dens	population density	continuous
lgp	length of growing period	continuous
elevation	elevation	continuous
twi	topographic wetness index	continuous
agrecol	agro-ecological class (soil-climate)	categorical
fs2001	farming system, 2001 classification	categorical
fs16lev1	farming system update, level 1	categorical
fs16lev2	farming system update, level 2	categorical
pa1	protected area class, layer 1	categorical
pa2	protected area class, layer 2	categorical
pa3	protected area class, layer 3	categorical
protected	protected area sampling classification	categorical
sampling	sampling indicator	binary
safe	safety indicator	binary

European Commission | Horizon 2020 European Union funding for Research & Innovation

This project has received funding from the
EU Horizon 2020 Research and Innovation Programme
under Grant Agreement 862900

SOILS4AFRICA