Data acquisition strategies for developing digital soil maps of Ethiopia

Markus Walsh, Jiehua Chen and Sonya Ahamed

Africa Soil Information Service (AfSIS), POB. 2704, Arusha, Tanzania

Contents

1 Introduction

People depend on soil for a wide range of essential ecosystem services. Soil is a key resource in the production of food, forage, fuel and fiber. Soil stores water from rainfall and irrigation and filters toxic substances through clay sorption and precipitation processes that determine water quality. Soil organisms decompose organic materials, cycle nutrients, and regulate gas fluxes to and from the atmosphere. However, as human populations have grown rapidly, there has been a strong tendency to trade off increases in the demand for

provisioning ecosystem services (e.g., for food and other commodities), for regulating (e.g., nutrient, greenhouse gas, and hydrological cycling) and supporting services (e.g., biodiversity), worldwide.

1.1 Rationale

Inherently low soil fertility, nutrient imbalances and potentially accelerating soil physical, chemical and biological degradation, are thought to constitute threats to increasing agricultural productivity and other ecosystem services in Ethiopia. However, at present the geographical extent of existing soil problems, their location specific trends and opportunities for managing these over time are not known. Some of the key information gaps include:

- The area of arable land and the proportion of arable land that is being used for cultivation have not been identified with sufficient accuracy to allow spatial targeting of appropriate management practices. Rapidly growing populations in Ethiopia (currently ~87.1M, and ~173.8M projected by 2050) will require substantial increases in either agricultural productivity per unit of currently cultivated land area and/or the expansion of cultivated lands to meet food demands.

- The rate of yield increases in staple crops (e.g. maize, millet, sorghum, tef, rice wheat, barley and pulses) may be slowing and various soil constraints appear to be one important factor influencing this trend. However, the spatial distribution of these constraints is not known and it is therefore difficult to devise location specific nutrient input and crop management recommendations.

- Economically viable and ecologically sustainable nutrient input/import options have not been previously established at sub-national scales nor for specific agro-ecologies and landuse systems. As a consequence, blanket fertilizer formulations and input recommendations are being applied country-wide.

- Soil degradation reduces agricultural productivity and other ecosystem service deliveries to people. However, the geographical distribution, rates and impacts of important soil degradation processes (e.g. wind and water erosion, organic matter and nutrient depletion, acidification, salinization, alkalization, compaction, loss of below ground biodiversity) are poorly documented and understood.

- Water-use efficiency in crop and rangeland production systems is determined in part by soil management. Again, the actual and potential water use efficiencies of agricultural systems in Ethiopia are poorly characterized and should be assessed to evaluate e.g. the potential tradeoff's between needs for agricultural production and the availability of potable water for rapidly growing human populations.

- Current land use systems may be increasingly disrupted by changing climatic conditions. The sustainability of these systems will depend on interactions between climate, soil, and land management as well as by the policies that enable or undermine them. An ability to track these potential changes and anticipate the consequences requires reliable information on climate change and the characteristics and distribution of soils.

These information gaps are not unique to Ethiopia, and many other countries in Africa (and beyond) face similar problems and soil resource related information needs. In response to this situation, the Africa Soil Information Service (AfSIS, see: `http://www.africasoils.net`) is developing a freely accessible, open source soil information system that will be generally applicable to mapping and monitoring soil resources at any spatial scale and at any given location in sub Saharan Africa.

The implementation of this system is also closely aligned to similar efforts that are being undertaken for the entire planet (see, `http://www.GlobalSoilMap.net`, and `http://www.fao.org/nr/water/landandwater_gsp.html`) and will be interoperable with global earth observation data exchange standards (see e.g. `http://www.earthobservations.org/geoss.shtml`).

1.2 Digital soil mapping

The general proposal that our natural environment should be mapped and monitored is widely supported by agencies responsible for managing natural resources, industry groups and community organizations. This information provides an essential basis for devising, implementing and monitoring land management activities with the intent of improving returns on investment to agriculture and land management and reducing the risks to communities and the environment.

Digital soil mapping has been defined as: "the creation of spatial soil information systems using field and laboratory methods coupled with spatial and non-spatial soil inference systems" [Lagacherie et al., 2006]. A *digital soil map* (DSM) is essentially a spatial database of soil properties that is based on a statistical sample of landscapes, which permits functional interpretation, spatial prediction, mapping and monitoring of soil properties that are relevant to soil management and policy decisions. DSM's can provide:

- Space-time information on the soil's capacity to deliver important ecosystem services (e.g. the ability to infiltrate water, produce crops, store carbon and cycle nutrients) in a given region of interest (ROI),
- Space-time predictions of significant soil constraints (e.g. nutrient status and potential depletion problems, aluminum toxicity, organic matter (carbon) deficits, plant rooting depth and crop water use restrictions, among others) with known confidence,
- Spatial targeting of soil management recommendations,
- Spatially explicit baselines for soil change detection and management impact assessments, and
- Quantitative assessments of the uncertainty in the predictions of soil properties and any recommendations that are derived from these.

Some key advantages of DSM versus conventional soil mapping approaches are: DSM's are based on "statistical inference", or the process of drawing conclusions from data that are subject to random variations (i.e. uncertainty), which can also include quantitatively stated expert opinion. DSM's can also be readily updated (spatially and over time),

as new data become available, and/or as soil information needs change. Many of the associated processes and workflows can now be automated using computer programs that provide previously unprecedented possibilities for the development of "evidence based" soil management recommendations [McBratney et al., 2003].

2 Main goals of this report

The main goals of this report are to provide advice on collecting and recording georeferenced soil data for developing an initial DSM of Ethiopia by the end of 2013. A critical part of obtaining high quality results, in this regard, involves large-scale soil survey activities that will be conducted by Ministry of Agriculture staff and other stakeholders e.g., Crop Nutrition Services (CNS) in Nairobi (`http://www.cropnuts.com`), and the CASCAPE project (`http://www.cascape.info`) that is being coordinated by Wageningen University and Research Center (WUR).

Coordinating these activities to achieve the best possible results necessitates the application of quality control measures on the resulting data and the information products and services that are generated from these. This report provides some initial recommendations in this regard and proposes several follow up actions (see Section 5).

To clarify the various procedures involved, we initially review key terminology and principles for developing DSM survey designs, including processes for data acquisition and recording (in Section 3). We then discuss a worked example including R-code (in Section 4) that can be used by collaborators to develop, explore and document their own DSM survey design options. A subsequent report (*Data analysis strategies for developing digital soil maps of Ethiopia*), has been targeted for the end of September, 2012 and will focus on DSM workflows for statistical estimation and prediction of soil properties.

3 Soil survey specifications

This section illustrates some of the decisions that need to be made in soil survey planning for digital soil mapping (as well as for other types of soil surveys), prior to engaging in field activities. The main reason for being quite explicit about these decisions from the start, is that they will overwhelmingly affect both the costs and the quality of any resulting DSM's.

We follow a slightly modified version of the extremely helpful scheme provided by de Gruijter et al. [2006, p. 29]. Our version of this contains a list of 15 items, which when completed would constitute the essential meta-data for a given soil survey activity. We are also attaching an accompanying spreadsheet to this report that collaborators can use to record the relevant meta-data for their specific sampling designs. Thoroughly documenting how and for what purpose data were acquired in a large, collaborative project like EthioSIS will be important, if the results from different surveys are to be usefully combined at a later stage.

3.1 DSM survey checklist

The examples provided in the checklist (below) were developed from preliminary consultations with ATA and MoA staff and cover the agricultural areas located within 113 Woredas of Ethiopia that have been identified by the ATA and the MoA as priority areas for topsoil fertility surveys.

The intent here is not to prescribe a specific survey design and data acquisition plan, but simply to describe the key decisions that need to be made for developing a plan that is suitable for DSM applications. An actual plan would need to be discussed and formulated in greater detail and in close consultation with all of the relevant stakeholders <*see follow up action in Section 5*>.

1. State the survey goals, including:
   (a) *Short title*: e.g., Survey for mapping topsoil fertility indicators in priority AGP Woredas of Ethiopia.
   (b) *Main purpose*: e.g., Soil fertility mapping to advise fertilizer formulations and application rates.
   (c) *Responsible*: e.g., CNS (`http://www.cropnuts.com`)
   (d) *Timeline for results*: ~? number of months, to be completed by date ?
   (e) Provisional budget: ~? Birr, assuming sample collection, processing and relevant lab analyses at ~4000 locations.
2. Provide design specifications of the survey including:
   (a) *Target universe*: or the overall Region of Interest (ROI), e.g the 0-20 cm topsoil of cultivated areas falling within 83 priority Woredas of Ethiopia (also see Fig. 4.1 and Section 4.1).
   (b) *Domain*: a specification of the parts of the ROI for which separate results are needed, if any. In this example, results would be needed for every 1 ha grid cell within the entire ROI.
   (c) *Target variable(s)*: a precise definition of all of the variables to be determined for each sampling unit, e.g. N, P, S, K and micronutrient concentrations (g kg^{-1} or mg kg^{-1}), soil organic carbon stocks (kg m^{-2}), bulk density (kg m^{-3}), % clay, silt and sand etc. Alternatively, variables indicating whether soil properties fall above or below a certain threshold value, e.g. plant available P < 5 ppm.
   (d) *Target parameter*: the type of statistic for which a result is needed e.g., the mean and variance of P concentration for every 1 ha grid cell within the domain. Alternatively, the proportion of with plant available P < 5 ppm for every 1 ha grid cell within the domain.
   (e) *Target quantity*: the combination of a domain, target variable and target parameter, e.g. the mean P concentration (parameter) between 0-20 cm of soil depth (target variable) for every 1 ha grid cell within the ROI (domain). In this example, there would be 4,832,100 such values to predict for each target variable by target parameter combination, at potentially multiple levels of soil depth (see Section 4.1 for further details).

(f) *Type of result*: i.e., quantitative or qualitative, which will determine the mode of inference that can be applied. In this example, the type of result is quantitative and would therefore be predicted.

3. *Define a quality measure*: meaning define the quantity that will be used to numerically express the statistical quality of the survey result. Examples include: 95% prediction intervals in estimation, mean errors of prediction, and classification error rates. In this example the quality measure that would be used is the mean standard error of prediction (SEP) of an independent model validation sample.

4. *Identify constraints*: this includes the allocated budget and/or minimum quality of the result as well as factors such site accessibility and/or travel time restrictions between sample locations. Note that for large-area soil surveys (such as this one), travel times between sampling locations are likely to impose major time and cost constraints on field work.

5. *Assess the availability of prior and supporting information*: e.g.,

   (a) Legacy soil profile data: A GIS compatible file containing previously collected soil profile data contained in the Africa.profiles database (v.1 description and download at: `http://www.africasoils.net/data/legacyprofile`).

   (b) Previous AfSIS surveys: currently 44, 10×10 km sentinel sites on which detailed soil observations and measurements. Four of these sites are located in Ethiopia. This data set is in progress; however, some of the data available on request.

   (c) Remote sensing and GIS data: A wide range of remote sensing, digital terrain derivatives at 100 m - 1 km resolution and GIS layers including a 1:1,000,000 soil map and a map of the distribution of agricultural land in priority Woredas in the year 2000 are available on request.

   (d) Model of the spatial variation: not available currently for all needed target variables, under development on the basis of AfSIS sentinel site surveys.

6. *Specify the sample support*: referring to the shape size a position and orientation of a sample, e.g., a standard auger core from 0-20 cm soil depth. If subsoil properties are of interest, additional samples from 30-50 cm soil depth may be obtained.

7. *Specify assessment method*s: this refers to standard field and/or laboratory measurement procedures; see e.g. documentation of standard methods and operating procedures at: `http://www.africasoils.net` and `http://www.cropnuts.com/`.

8. *Will composite sampling be used?* e.g, yes, a well mixed composite sample consisting of 4 auger cores from 0-20 cm depth collected at 4 locations within a 1000 m^2 area would be used as per the AfSIS standard operating procedures.

9. *Choose design* or *model based inference*? See de Gruijter et al. [2006, Chapter 8]. In this example, design-based inference would be used (also see Section 4).

10. For design based inference: *select a random sampling design type and the attributes of the chosen design*. In this case, because of significant travel time restrictions between sampling locations, we recommend a *Multistage Sampling* pattern on a 100 m resolution equal area grid (UTM zone 37N, WGS84 datum, also see Section 4.2).

11. For model based inference: *choice of sampling pattern type and optimization algorithm*: not applicable in this case.

12. *Method of statistical inference*: calculation of 95% prediction intervals based on Hierarchical Models [Banerjee et al., 2004] using remote sensing and other spatial covariates and/or means and associated variances from 3-D Regression Kriging [Cressie and Wikle, 2011].

13. *Identification of the actually selected sample*: including generating GIS maps and GPS compatible files of sampling locations that have been selected (for example see Section 4).

14. *Protocols for field and laboratory data recording*: Electronic field and laboratory data recording procedures and database schemas are available. These have been designed and tested by AfSIS over the last 3 years and conform with international soil data and meta-data standards.

15. *Conduct an ex-ante assessment of operational costs and quality of results*: document under development.

3.2 Statistical inference and use of spatial covariates

The statistical inference of DSM models generally consists of three parts, as expressed by the following equation:

$$y(s) = \mu(s) + \omega(s) + \varepsilon(s) \tag{3.1}$$

for which, $y(s)$ is an observed or measured soil variable at a 3-dimensional location (s, described in $x, y, depth$ coordinate space). $\mu(s)$ refers to a mean estimate at a given location, that can be conditioned on environmental data (e.g., remote sensing grids and/or existing soil maps). The covariates (item 5c in the checklist) are usually selected to fill the entire domain (item 2b), meaning these data values are available for every pixel. However, more local covariates such as soil spectra and other proximal sensing measurements (among many others) at the location could also be used as covariates. In those instances we would refer to the resulting models as "pedo-transfer models".

The main benefit of using covariates, is that their use improves the precision and accuracy of predictions of soil properties at unsampled locations in the domain. In this particlar case, a wide range of covariates are available for all of Ethiopia, via the Africa.grids database (server access to this database at: `sftp://afsisdata.ciesin.columbia.edu`, can be provided on request).

The residuals of these types DSM models are then typically further partitioned into two pieces: one which is spatial, $\omega(s)$ and one which represents the remaining uncorrelated "noise" or model uncertainty, $\varepsilon(s)$. The spatial part, $\omega(s)$ invokes Waldo Tobler's 1^{rst} Law of Geography, which states that "Everything is related to everything else, but near things

are more related than distant things". The spatial correlation that is implied in Tobler's Law can often be exploited for improving spatial predictions of DSM models and forms the theoretical basis for most spatial inference procedures.

One way of solving equation 3.1 is via 3D Regression Kriging [Cressie and Wikle, 2011], but newer approaches such as Bayesian hierarchical methods exist and have been successfully used in other Earth and Life Science disciplines [Banerjee et al., 2004]. The latter methods offer the advantage of more realistic representations of the underlying soil processes and of the uncertainty of the model based predictions that are generated.

We have been developing and testing code e.g., by updating the `geoR` and `gstat` libraries in R and working with Bayesian models in the `R2Jags` package in R (`http://mcmc-jags.sourceforge.net/`) and WinBugs (`http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml`). This represents foundation code development for various soil mapping and monitoring workflows, but also the basis for calibration and Bayesian updating procedures of soil spectral libraries [Shepherd and Walsh, 2002, 2007]. Evaluation of these statistical inference methods is an ongoing activity in the AfSIS project. We are also working on automating and packaging these procedures in R+GRASS. A subsequent report, targeted for the end of September 2012, will provide additional information and advice about data analysis, mapping options and code <*see follow up action in Section 5*>.

3.3 Generating a sampling plan

To generate an actual sampling plan (i.e, item 13 of the checklist), one would have to implement the overall design in some form of computer code. While there are a number of basic facilities available for this purpose in free and open software such as R (`http://www.R-project.org`) and GRASS GIS (`http://grass.fbk.eu/`), additional scripting is typically required. Section 4 illustrates this based on the assumptions that have been made currently, and we will create a R+GRASS package over the coming 6 months to largely automate these workflows.

At present we restrict ourselves to the application of probability based sampling procedures for Ethiopia, primarily because we lack reliable models of the spatial variability of the various soil variables being considered. Once such information becomes available, it will be possible to compare *probability vs purposive* sampling approaches for predicting local quantities in space [de Gruijter et al., 2006, Chapter 8], and to offer additional recommendations in this regard <*see follow up action in Section 5*>. We will not pursue *convenience sampling* (e.g., road-side sampling) on the grounds that it generally does not lead to statistically interpretable results. In the mean time, the probability based sampling approaches we are suggesting in this report are robust, fairly simple to implement and will produce data that is useful for DSM applications.

The code presented under section 4.3 is quite general in that it can cover a wide variety of probability based sampling procedures, including e.g., simple random sampling, stratified random sampling, cluster random sampling, regression estimator (double) sampling, two-stage and multistage sampling and various combinations of these. The code can typically be adjusted to reflect specific user requirements and design criteria by changing just two to three of the main design parameters and the ROI.

3.4 Data collection protocols

Similarly, protocols for recording field and laboratory data (i.e., item 14 on the checklist) need to be translated into functioning data collection toolkits and database schemas prior to initiating field work. Because many Earth and Life Science disciplines use, or would like to use, soil data, it is important that the resulting databases are interoperable with other databases, meaning that they can interact with other products or systems, present or future, without restrictions to data access or implementation of new methods. The requires adherence to international data models and standards, see e.g. the Open Geospatial Consortium (OGC) at: `http://www.opengeospatial.org/` and the Global Earth Observation System of Systems (GEOSS) at: `http://www.earthobservations.org/geoss.shtml`.

The protocols that have been developed by AfSIS over the last three years fit these requirements and are sufficiently flexible to meet specific user needs with regard to new target variables (item 2c in the checklist), or new assessment methods (item 7), for example. These can now be accommodated easily in OGC standards compliant databases.

Field data collection protocols have been implemented in Open Data Kit, which is "a free and open-source set of tools which help organizations author, field, and manage mobile data collection solutions" on Android mobile devices (see `http://opendatakit.org`). Access to the AfSIS field forms for soil and socioeconomic surveys can be obtained at: `http://formhub.org`, on request. Lab database schemas that can be used to record both wet chemistry and spectral data are also available and can be customized to meet user needs and access requirements.

Remote sensing, field and laboratory data can be linked to one another through universally unique soil sample ID's, and the associated bar and/or Quick Response (QR) codes that will allow tracking of samples through the entire processing chain from the field, through the lab and to the analysis stage. Collectively, these tools are expected to enable the rapid acquisition of scientifically justifiable and high quality soil data.

Though we do strongly encourage free and open access to all soil and remote sensing data, we understand that such data may be considered to be sensitive information in some countries. Our database protocols can therefore be maintained and updated in secure locations, in this case under e.g. the full control of EthioSIS staff, who can determine user restrictions and access levels should this be desirable or needed <*see follow up action in Section 5*>.

3.5 Ex ante and ex post evaluations of sampling plans

Evaluation of a proposed sampling plan should ideally consider the *ex ante* costs and timelines of implementing the proposed plan, as well as the expected quality of the result. This can be done in different ways, but fundamentally an explicit ex ante evaluation of a sampling plan provides a final check on whether the stated goals of the survey are likely to be achieved within budget and on time. Depending on the outcome of such an evaluation, the plan, the budget and/or the timeline may need to be revised to achieve the goals and specifications set out under items 1 & 2 on the checklist.

In this context, obtaining rough estimates of the sample size that is affordable within a given budget and a realistic timeline, are generally more easily obtained than predictions of the quality of the results. While in some instances it may be possible to use legacy

data for this <*see follow up action in Section 5*>, pre-surveys may need to be carried out to evaluate the spatial (co)variance of target variables (item 2c in the checklist), within a given domain. Also the use of covariates in this context can significantly reduce sampling efforts and costs required at given levels of precision and accuracy.

In our current example, the quality measure of the survey result would be expressed as the mean standard error of prediction (SEP, item 3 on the checklist) as a function of the target quantity (item 2e), which would be established on the basis of an independent model validation sample. Subsequently, comparisons of the *ex ante* and the *ex post* SEP would provide opportunities for improving future sampling and should really be part of the monitoring and evaluation component of EthioSIS, as it reflects both the costs and quality of the DSM products that are obtained.

4 Worked example

This section of the report derives a worked example of the main steps we have implemented in deriving a sample for ~4000 point locations within the ROI (item 2a in the checklist). We would like to emphasize again that this is simply an example to illustrate the flexibility and transparency of the underlying methods and the R-code for the implementing the approach. It does not constitute a sampling plan proposal.

4.1 Delineating the ROI

In this particular case we were provided with a GIS file of the ROI dating from 2000 that delineates the agricultural land area in 82 of the priority Woredas of Ethiopia (see Fig. 4.1). This area covers ~48,000 km^2 of Ethiopia. A 200 m resolution rasterized, geoTif version of this file (`Priority_Ag_areas.tif`) as well as a Google Earth file (`Priority_Ag_areas.kml`) are available in the AfSIS-Ethiopia dropbox at the following link: `https://www.dropbox.com/sh/pe1pu7gzwdfiyp3/RqSvy_KCGd`. The kml file can be used to obtain a visual impression of how well the ROI boundary matches the approximately current extent of agricultural land in the 83 Woredas that it covers.

Unfortunately, we know very little about the provenance and subsequent processing of the original data that were used to construct this map, and are thus unable to comment on specifically how this area was delineated or what classification error rates were obtained in differentiating between agricultural and non-agricultural land areas in the year 2000. It is also likely that the actual area of agricultural land has changed over the intervening 12 years. The extent of those changes is currently not known. Thus, we strongly recommend that a new land cover classification be undertaken on the basis of georeferenced field data collections currently underway and current remote sensing data (i.e., MODIS and SRTM from the Africa.grids data base), <*see follow up action in Section 5*> .

The second GIS file that was made available to us contains the outlines of 113 (AGP as well as non-AGP) Woredas of Ethiopia that have been designated as soil survey priority areas for 2012/2013 (also shown in Fig.). Both shp and kml versions of this file (`Priority_woredas.*`) are also available in the AfSIS-Ethiopia dropbox. The file also contains information about which team (CNS, or CASCAPE) would be responsible for sampling which of the Woredas. We are uncertain as to wether this listing is definitive,

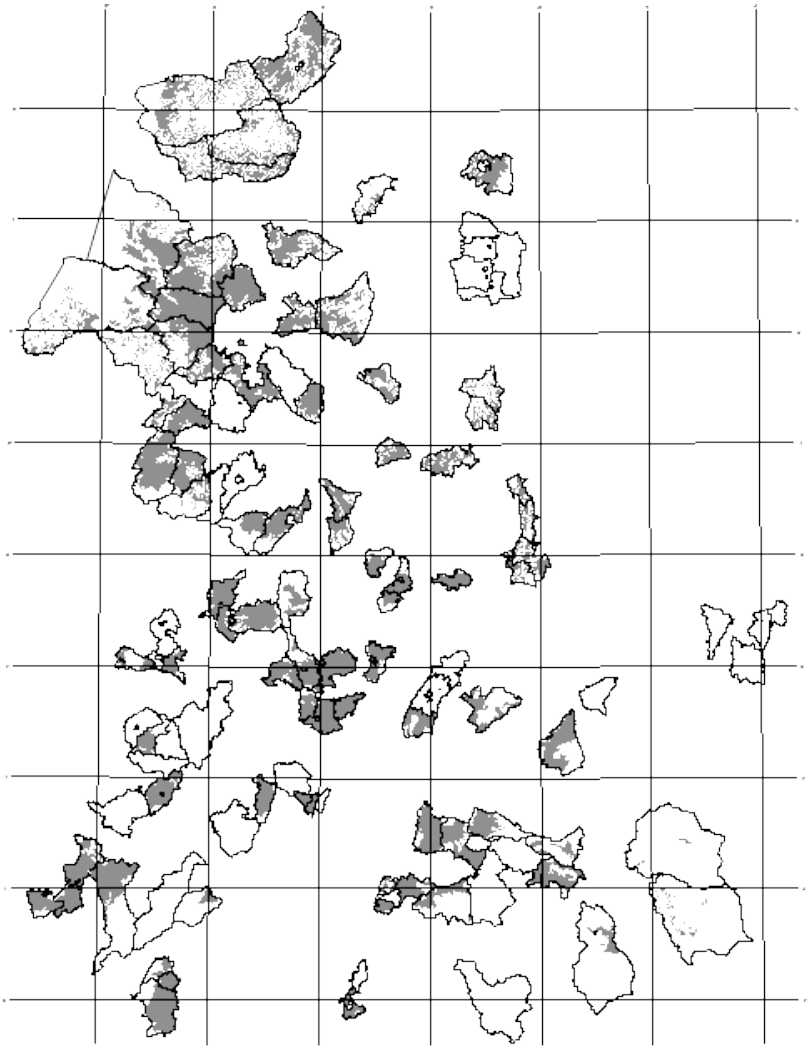

Figure 4.1: Agricultural areas (in grey) located within priority Woredas of Ethiopia.

but again the specifics of this can be changed easily once the relevant arrangements have been finalized.

The third file (`Priority_Ag.csv`) in the AfSIS-Ethiopia dropbox contains an overlay of the previous two files on a 1×1 km grid. This file can be imported and manipulated as a spreadsheet and serves as the main input to the R-script described under Section 4.3. To avoid confusion the 1×1 km is simply a small trick that we currently use to provide starting points for the randomization algorithm that we have used. The actual sampling locations occur on a 100 × 100 m grid (see item 2b on the checklist), which provides for exactly 4,832,100 possible sampling locations within the current ROI. These are geolocated on the Ethiopian National UTM zone 37, WGS84 coordinate reference system. We could of course also zoom in even further, say to a 25 m grid spacing, and this would provide us with a total of $4,832,100 \times 16 = 77,313,600$ possible sampling locations within the ROI.

The question now is: how to select among these possibilities to derive a sampling plan consisting of e.g., ~4000 (substitute any number >4,000 here) sampling locations

that is useful for DSM applications, yet which is also practical in terms of field logistics. It is likely that several iterations and significant additional feedback will be needed to complete this step. The next section describes what we think the main options are given our current understanding of the overall goals of the EthioSIS project.

4.2 Main options for sampling the ROI

For large area soil surveys, simple random sampling (SRS) is relatively inefficient in terms of the amount of time spent on traveling between locations versus the time that can be spent on actual sampling activities. Additionally, the sampling variance that is obtained is usually larger than for other designs at the same level of cost. SRS can also result in poor spatial coverage, leaving large empty spaces between sampling locations.

One way of dealing with empty spaces and high variance is to add some form of *stratification* to the design. By stratification we mean dividing the ROI into smaller sub-areas, or *strata*. This can be done, using either: a set of supporting variables e.g., existing soil maps, terrain models and/or remote sensing data, or alternatively using spatial coordinates. The two main advantages of stratification are that efficiency is increased compared to SRS, and that separate estimates for the different sub-areas are supported.

In this particular example we shall use the Woredas as geographical strata and allocate sampling effort in proportion to the size of the agricultural area in each of the Woreda's under consideration. We are also working on a biophysical stratification scheme for all of Ethiopia that will be presented at a later stage <*see follow up action in Section 5*>. The simple probability proportional to size allocation formula under stratification is:

$$n_i = round\left(\frac{a_i}{A} \cdot N\right) \tag{4.1}$$

for which n_i is the number of sampling locations to be allocated to the i^{th} stratum (Woreda in this case), a_i is the agricultural land area within the i^{th} stratum, A is the total agricultural land area within the ROI, and N is the target sample size (e.g., 4000). This step solves the between stratum allocation problem by spreading sampling locations across the entire ROI and weighting the number of locations within Woredas in proportion to their agricultural land areas. It also has the logistical advantage that the Woredas could be sampled separately, in whatever priority ordering, with field teams being deployed accordingly.

While one could now use an SRS randomization for sampling every stratum, this would not solve the problem of travel time restrictions within large Woredas. A good alternative in this regard is to use a multistage strategy. In *multistage sampling* the area is again subdivided into sub-areas, usually on a regular square, triangular or hexagonal grid that can be nested at different levels of resolution, i.e., at 1 km, 500 m and 100 m grid resolution as in this particular case.

Sampling is then restricted to selected sub-areas. Whereas, in stratified random sampling all strata are sampled, in multistage sampling only parts of the sub-areas are sampled. Figure 4.2 provides a graphical illustration of this. The main advantage of this approach is in reducing the travel times between sampling locations. The main disadvantage is that the resulting spatial clustering (see Fig. 4.2) can lead to reduced precision. In general however, the operational/logistical advantages that are gained, allow for a larger sample size at the same overall budget and typically outweigh any losses in precision.

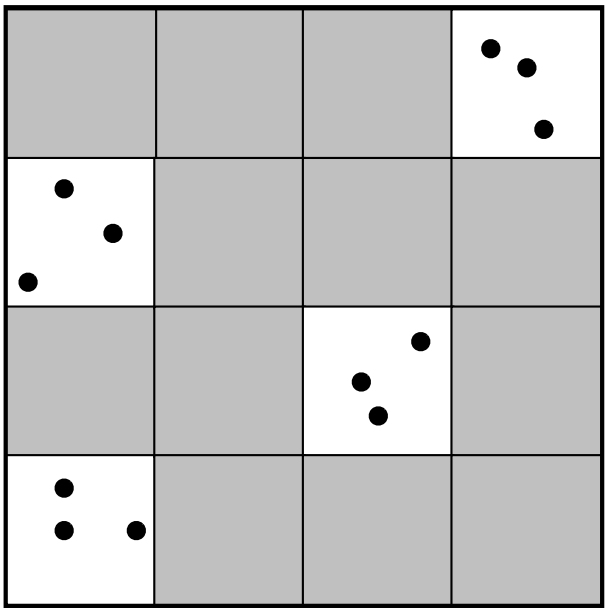

Figure 4.2: Notional example of a two stage sampling process in which four pixels are selected in the first stage, and three sampling locations in the second stage. Note the spatial clustering that occurs.

4.3 Selection procedure and R-code

The procedure that is implemented in the attached R-script `https://www.dropbox.com/sh/pe1pu7gzwdfiyp3/RqSvy_KCGd` combines stratified with multistage sampling approaches by initially stratifying by Woredas, and subsequently drawing a multistage sample from nested grids. We refer to this as *stratified multistage sampling*. The site selection procedure used in this particular example consists of the following steps:

1. Read a file (e.g,, `Priority_Ag.csv`) into R that contains the center locations of a 1×1 km grid of areas that which are fully included within the ROI boundary and that include a listing of all the relevant strata i.e., Woredas in this particular case. This file can be generated in free and open GIS software, such as GRASS (`http://grass.fbk.eu`) or QGIS (`http://www.qgis.org`).

2. Randomly select 1×1 km grid centers for each Woreda, in proportion to the area of agricultural land that is contained within within each Woreda, or more generally, contained within each stratum of the ROI.

3. Divide each randomly selected 1×1 km grid into cell into e.g., four 500×500 m grid cells, and then further divide each 500×500 grid cell into twenty five 100×100 grid cells.

4. Randomly select e.g., two 100×100 m grid cells within each of the four 500×500 m grids. In this case, this leads to a sample of eight sampling locations per selected 1×1 km grid cell.

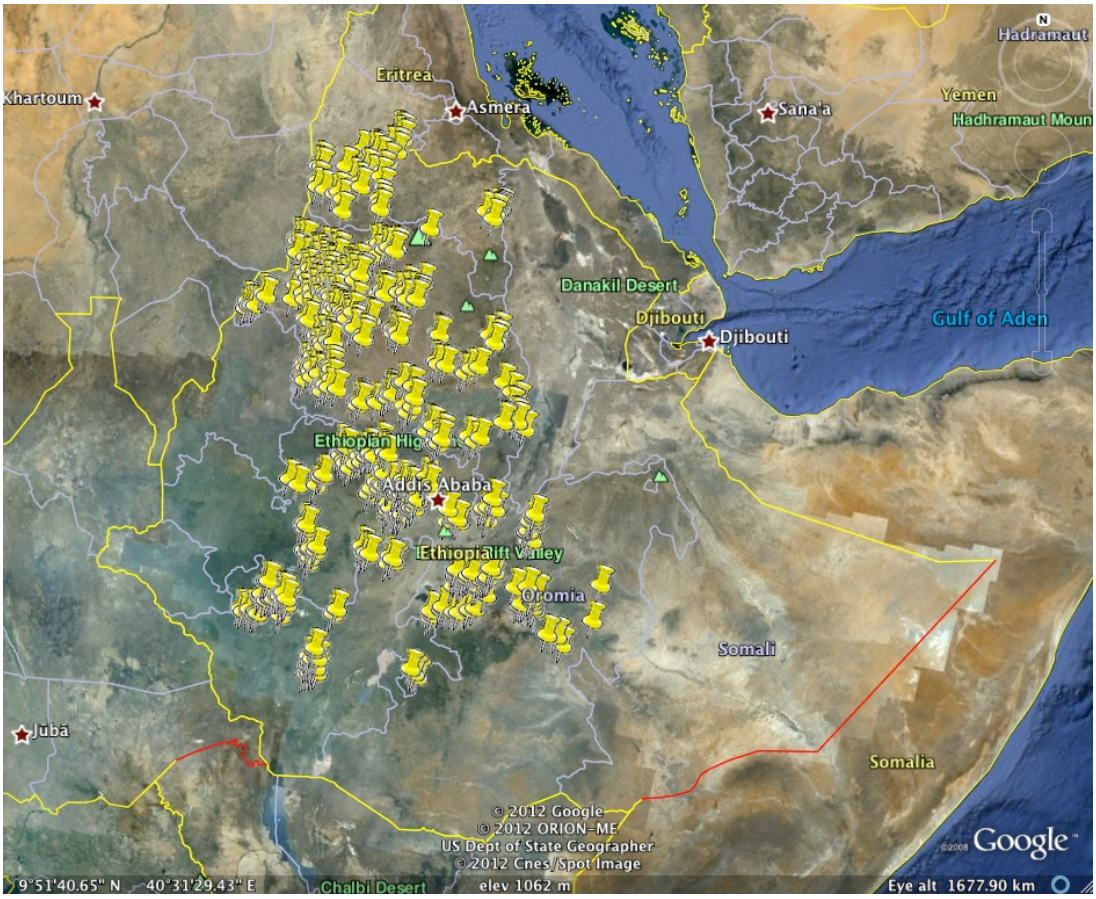

Figure 4.3: Sample realization in Google Earth.

5. Randomly select one location point within each selected 100 × 100 m grid cell. In this particular case case this provides for areal support of the sample, under the assumption that samples would be physically composited over a 1000 m area (item 8 on the checklist).

6. Reproject the sampling coordinates to the Geographic projection (Lat/Lon, WGS84) that is compatible with Google Earth, and export the a file for visualization in e.g., Google Earth (see `samplelocations.kml` in the Dropbox, also see Fig. 4.3).

The procedure is quite flexible, since the sizes of grids and the allocations of samples in each level of grids can be adjusted in accordance with varying field situations and prior information. It can also include different stratification schemes.

Note that you will need to install the the following packages in R prior to running the script, for Windows and Linux the installation from within R is as follows:

```
> install.packages(sampling, rgdal, proj4, lattice)
```

Under Mac OS X use:

```
> install.packages(“sampling”, “proj4”, “lattice”)
> setRepositories(ind=1:2)
> install.packages("rgdal")
```

Also note, that in both instances you would have to install the GDAL (Geospatial Data Abstraction Library) on your computer. You can download this either at: `http://www.gdal.org/`, for Windows and Linux, or at: `http://www.kyngchaos.com`, for Mac OS X.

4.4 Sample frame visualization

One realization of the sampling procedure is presented in kml format, and the kml file in in the sampling_code / sampled_locations folder in the AfSIS-Ethiopia dropbox at the following link: `https://www.dropbox.com/sh/pe1pu7gzwdfiyp3/RqSvy_KCGd`). This can be used to check factors such as site accessibility, and can also be used for prioritization and planning of field activities. Finally most GPS receivers, navigation software and tablet computers will readily read kml formatted files. More generally, the script will support any GDAL compatible formats e.g., for inclusion in GIS databases.

5 Main followup actions to this report

References

S. Banerjee, B.P. Carlin, and A.E. Gelfand. *Hierarchical Modelling and Analysis for Spatial Data*. Chapman and Hall, Boca Raton, 2004.

N. Cressie and C.K. Wikle. *Statistics for Spatio-Temporal Data*. John Wiley and Sons, 2011.

J.J. de Gruijter, D. Brus, M. Bierkens, and M. Knotters. *Sampling for Natural Resource Monitoring*. Springer Verlag, Berlin; New York, 2006.

P. Lagacherie, A.B. McBratney, and M Voltz. *Digital Soil Mapping: An Introductory Perspective*. Elsevier, 2006.

A.B. McBratney, M.L. Mendonça Santos, and B. Minasny. On digital soil mapping. *Geoderma*, 117:3–52, 2003.

K.D. Shepherd and M.G. Walsh. Development of reflectance libraries for characterization of soil properties. *Soil Science Society of America Journal*, 66:988–998, 2002.

K.D. Shepherd and M.G. Walsh. Infrared spectroscopy-enabling an evidence-based diagnostic surveillance approach to agricultural and environmental management in developing countries. *Journal of Near Infrared Spectroscopy*, 15:1–19, 2007.