Notes on stage 1 sampling frames for a national Simple Spatial Survey Method survey in Mozambique

Ernest Guevarra

26 August 2020

1 Introduction

This document explores possible stage 1 sampling frames for a national survey in Mozambique using the Simple Spatial Sampling Method.

2 Stage 1 sampling grid

Given a boundary map of Mozambique that includes borders at the *provincias* level (administrative level 1) and point locations of villages throughout the country, various stage 1 sampling grids using `d` values from 10 to 15 kms were tested.

2.1 Estimating sampling points based on area size

The number of sampling points needed for the stage 1 sampling frame can be estimated using the total area size of the survey area and dividing this by the area size of the hexagonal grid based on a specified value of `d`.

Given a specific `d` value, the area size of the hexagonal grid can be calculated as follows:

$$\text{hexagon area size} \ = \ \frac{3 \times \sqrt{3}}{2} \ d^2$$

Mozambique has a total area size of 801590 km^2. The estimated number of sampling points for a national survey in Mozambique using simple spatial sampling method is estimated by:

$$n_{\text{estimated total sampling points}} \ = \ \frac{\text{Total area size of Mozambique}}{\text{Hexagon area size}} \ = \ \frac{801590 \ \text{km}^2}{\frac{3 \times \sqrt{3}}{2} \ d^2}$$

The hexagon area sizes for `d` values from 10 to 15 kms are and the corresponding number of sampling points required are shown in the table below:

Table 1: Hexagonal grid area sizes for d values from 10 to 15 kms

d	Hexagon area size	Number of sampling points
10	259.81	3086
11	314.37	2550
12	374.12	2143
13	439.07	1826
14	509.22	1575
15	584.57	1372

The preceding calculations assume that the villages within Mozambique are spread througout the whole land area of the country. However, as shown in the map below, there are areas within the country with no settlements

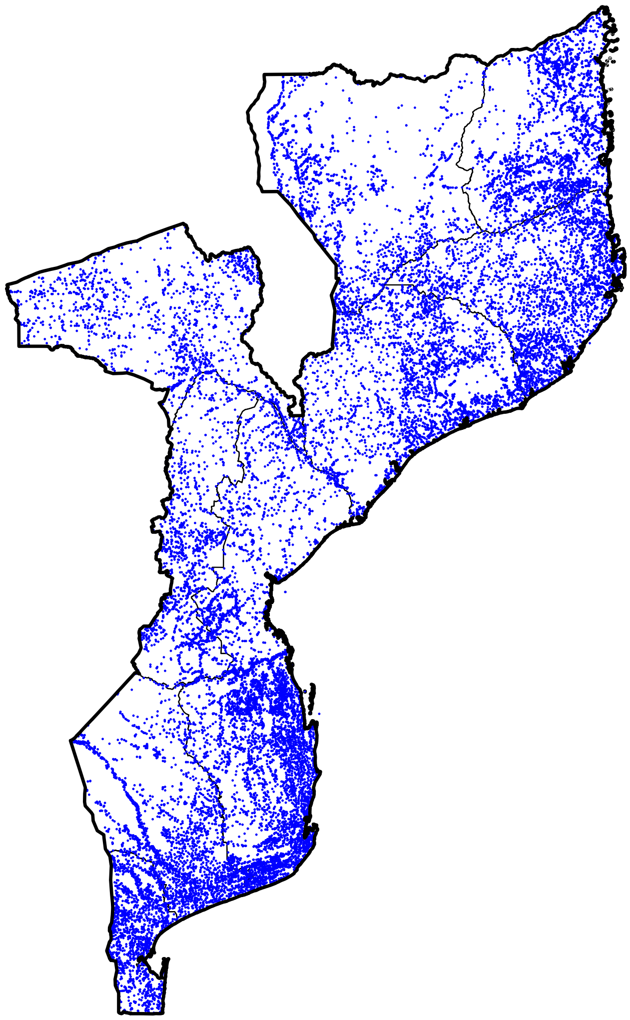

Figure 1: Map of Mozambique provinces with settlements

The implication is that this estimation approach tends to overestimate the total number of sampling points needed. From a planning perspective, this can be acceptable particularly if appropriate maps for use in stage 1 sampling are still being secured. This approach provides a safe estimate to use for planning as resources will be organised based on a higher number of sampling points for the stage 1 sampling frame and eventually can be adjusted down (if needed) once a map-based determination has been performed (as described below).

2.2 Determining number of sampling points using maps

Once maps are available, number of sampling points can be determined based on the simple spatial sampling method approach to stage 1 sampling.

The various stage 1 sampling frames for values of `d` from 10 to 15 kms including respective stage 1 sampling maps are shown below.

Table 2: Stage 1 sample characteristics of various d values

d	Number of sampling points	Number of clusters/ settlements selected
10	3086	2736
11	2550	2334
12	2143	2002
13	1826	1728
14	1575	1520
15	1372	1334

As was expected, the number of sampling points based on actual maps are less than the estimated sampling points. The following maps show the selected settlements for sampling at various values for `d`.

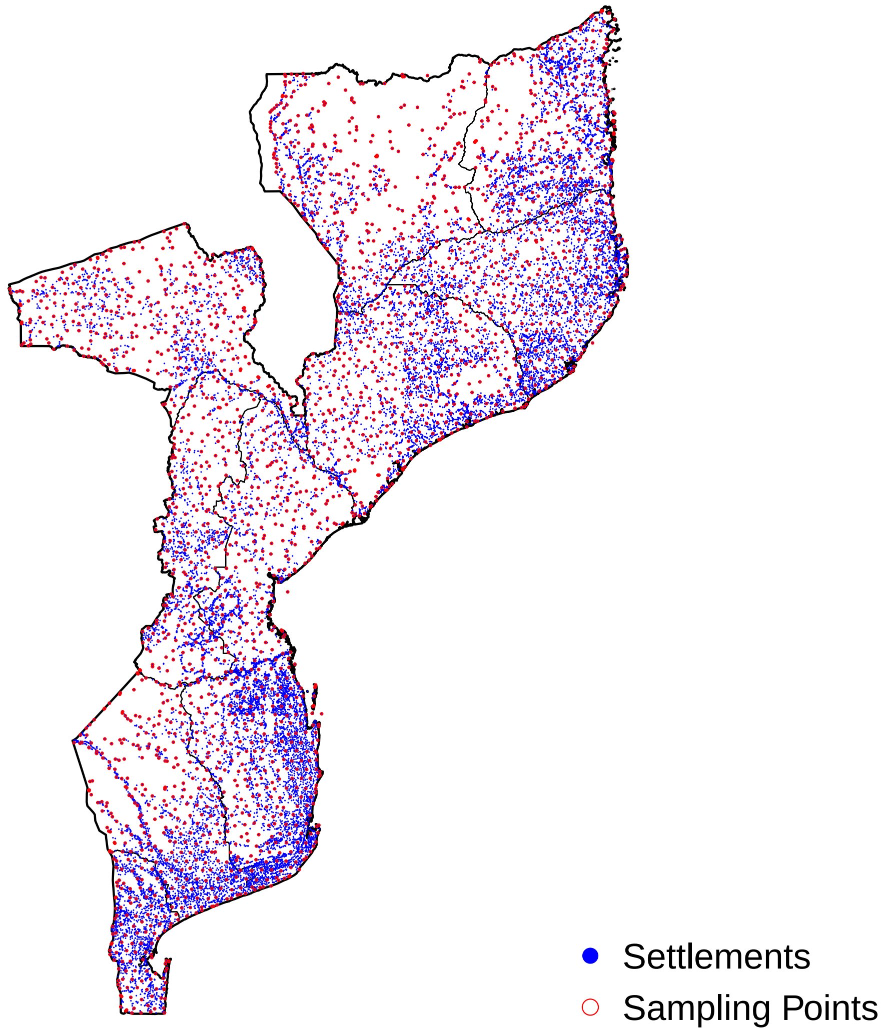

Figure 2: Stage 1 sampling map for d of 10 kms

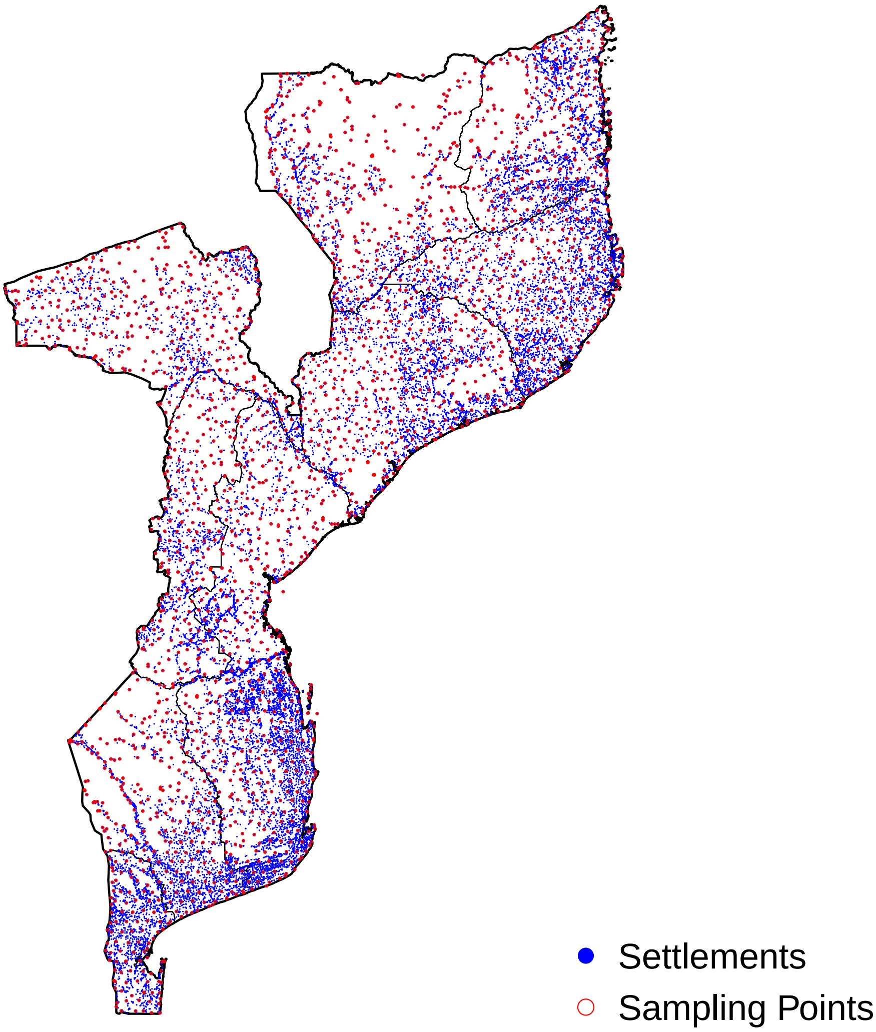

Figure 3: Stage 1 sampling map for d of 11 kms

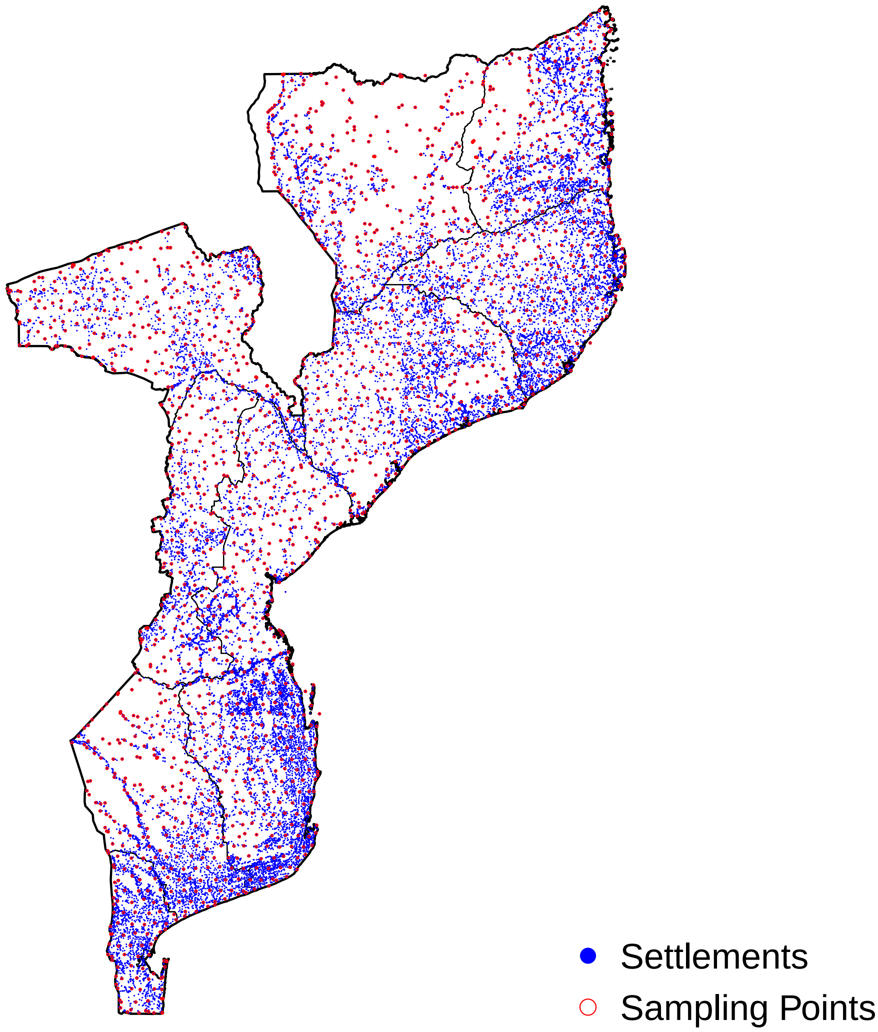

Figure 4: Stage 1 sampling map for d of 12 kms

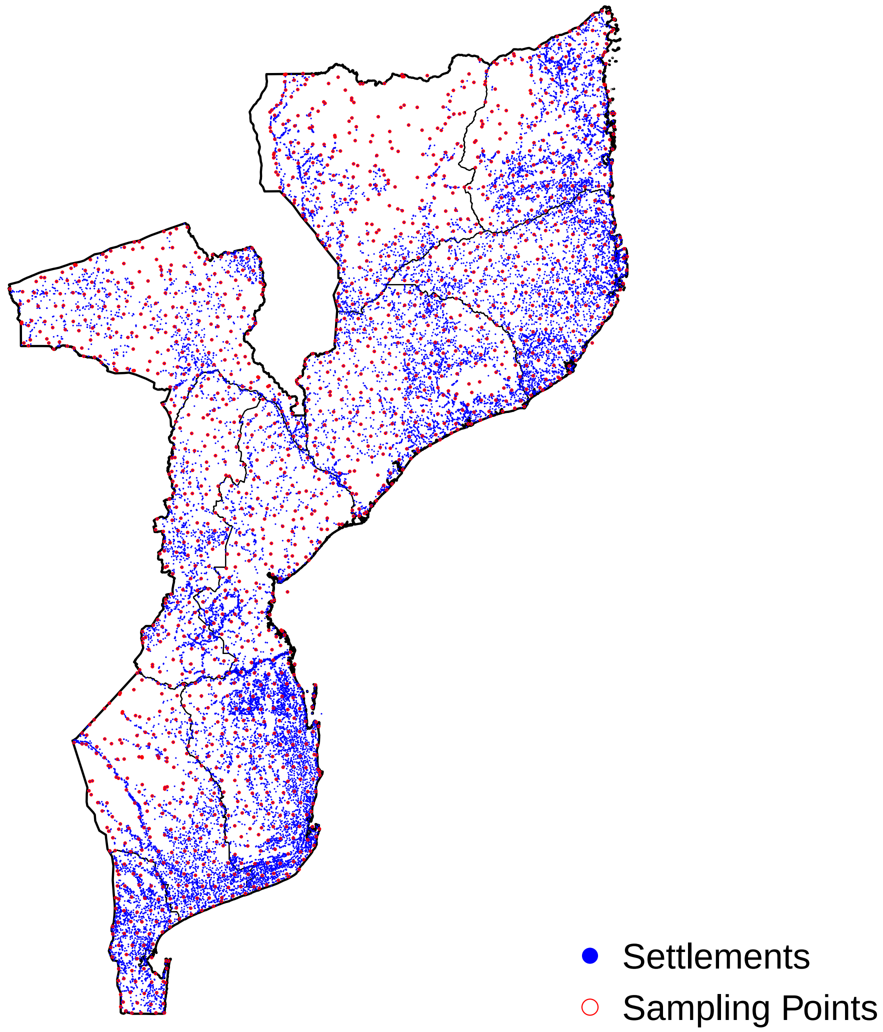

Figure 5: Stage 1 sampling map for d of 13 kms

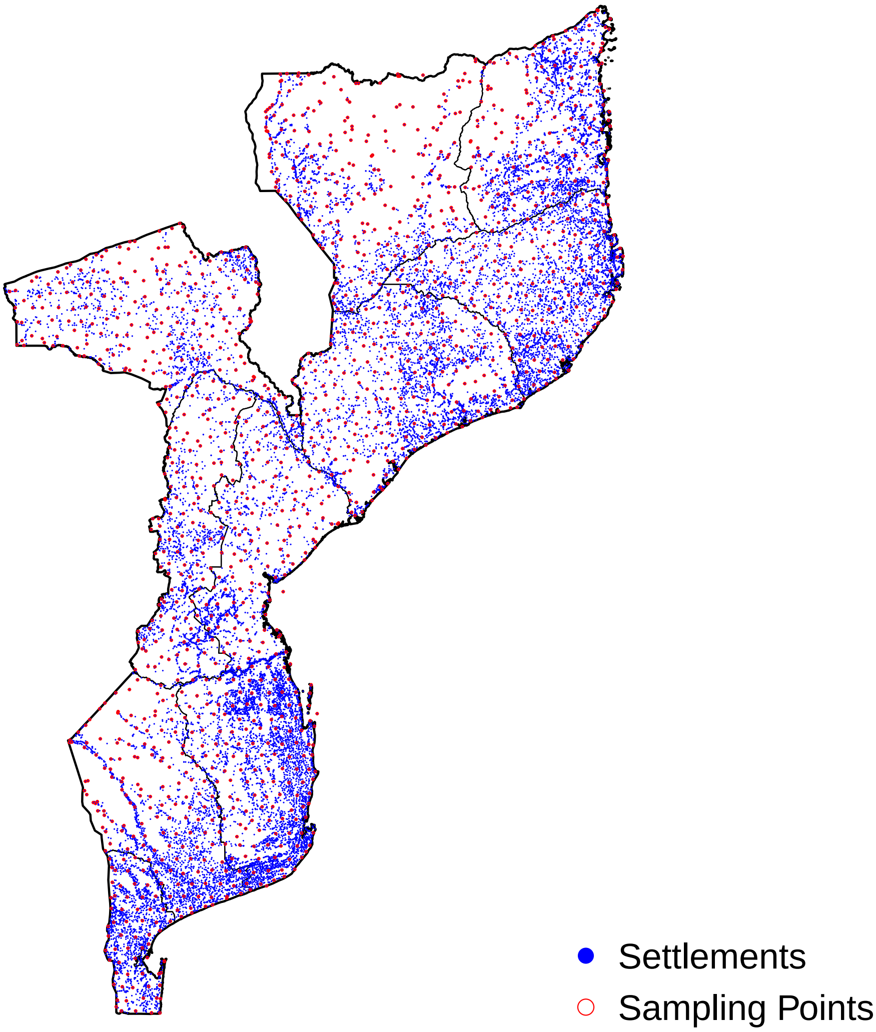

Figure 6: Stage 1 sampling map for d of 14 kms

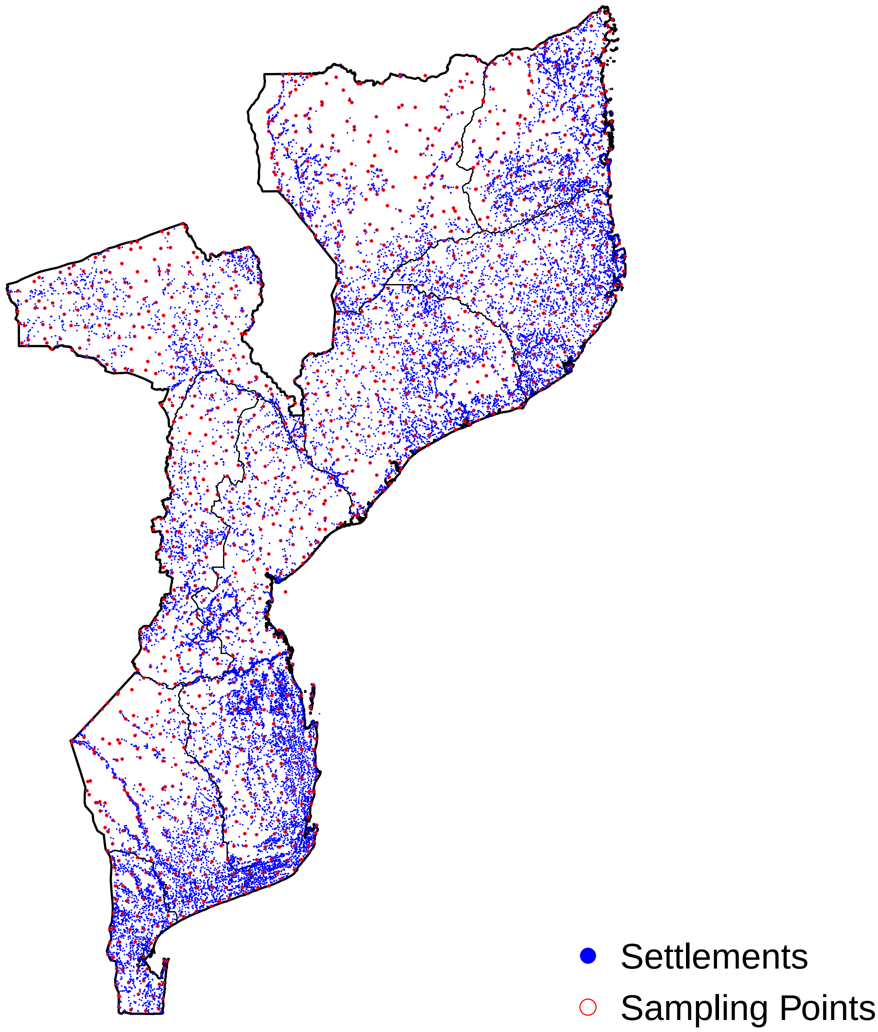

Figure 7: Stage 1 sampling map for d of 15 kms

2.3 Determining d using 'half the distance between markets' approach

Another possible approach to the determination of an appropriate value for `d` is the 'half the distance between markets'. This method is described in the FANTA SQUEAC Technical Reference [Myatt et al., 2012]. In this approach, the value of `d` is approximated by the distance carers are willing to walk/travel to access services. A proxy for this would be the half the average distance of villages to the nearest market.

Using the locations of settlements, and the locations of major towns and cities in Mozambique, half the average distance to markets can be estimated as follows:

```r
library(mozambique)
library(tripack)
library(geosphere)

## location data for major towns and cities
townsCities <- subset(ppl, place %in% c("town", "city"))

## Get distance between towns and cities
distTownsCities <- distm(x = townsCities)

## Craete a triangular irregular network between towns and cities via
## Delaunay triangulation
triTownsCities <- tri.mesh(x = townsCities@coords[ , 1],
                           y = townsCities@coords[ , 2])

## Get network of towns and cities
nbTownsCities <- neighbours(triTownsCities)

## Get vector of distances in network of towns and cities
listDistances <- NULL

for(i in seq_len(length(nbTownsCities))) {
  x <- NULL
  for(j in nbTownsCities[[i]]) {
    x <- c(x, distTownsCities[i, j])
  }
  listDistances <- c(listDistances, x)
}

## Convert distances to kms
listDistances <- listDistances / 1000

## Get half of median distance
median(listDistances) / 2
```

```
## [1] 22.60953
```

```r
## Get half of the mean distance
mean(listDistances) / 2
```

```
## [1] 34.52044
```

In the calculations above, we use both median and mean. The half of the mean distance to markets is about **35 kms** while half the median distance to markets is about **23 kms**.

In this case, median might be the more robust estimator as there are some long distances along the borders of the triangular network that connects extremely distant points on the network.

If we use a `d` of 23 kms, we get the following stage 1 sampling frame:

Table 3: Stage 1 sample characteristics for d values based on half distance between markets

d	Number of sampling points	Number of clusters/ settlements selected
23	584	581
35	252	254

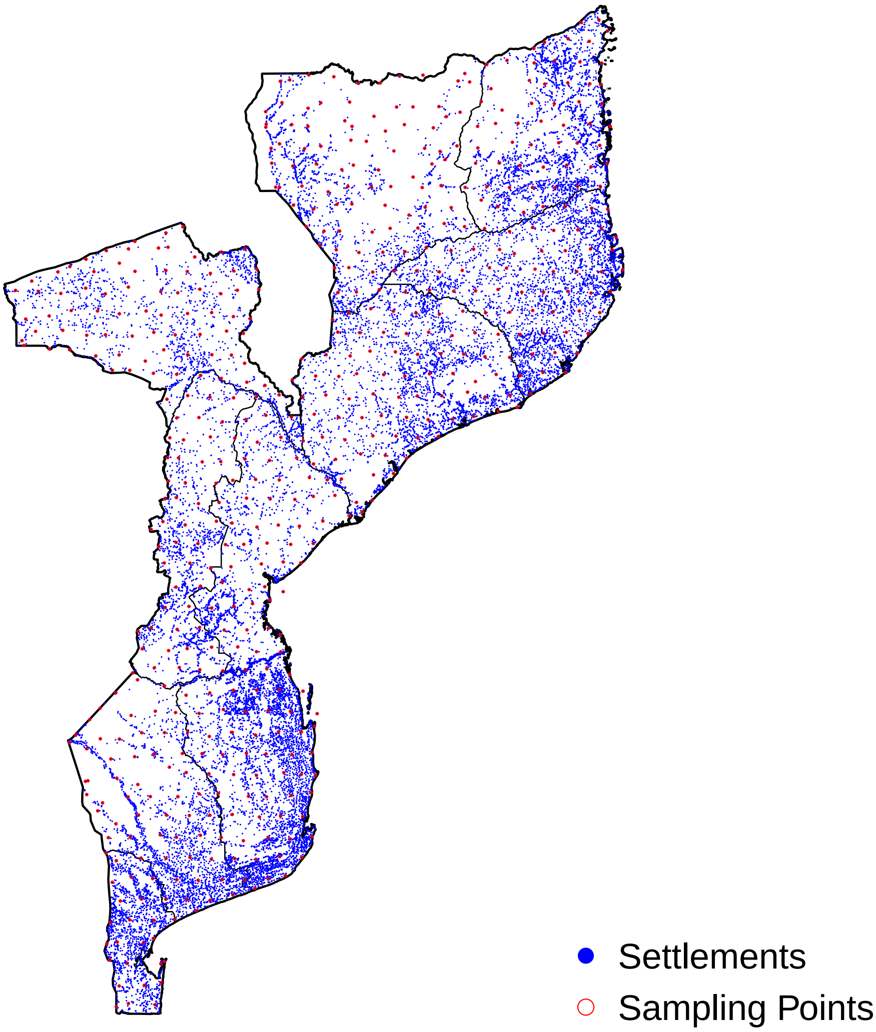

Figure 8: Stage 1 sampling map for d of 23 kms

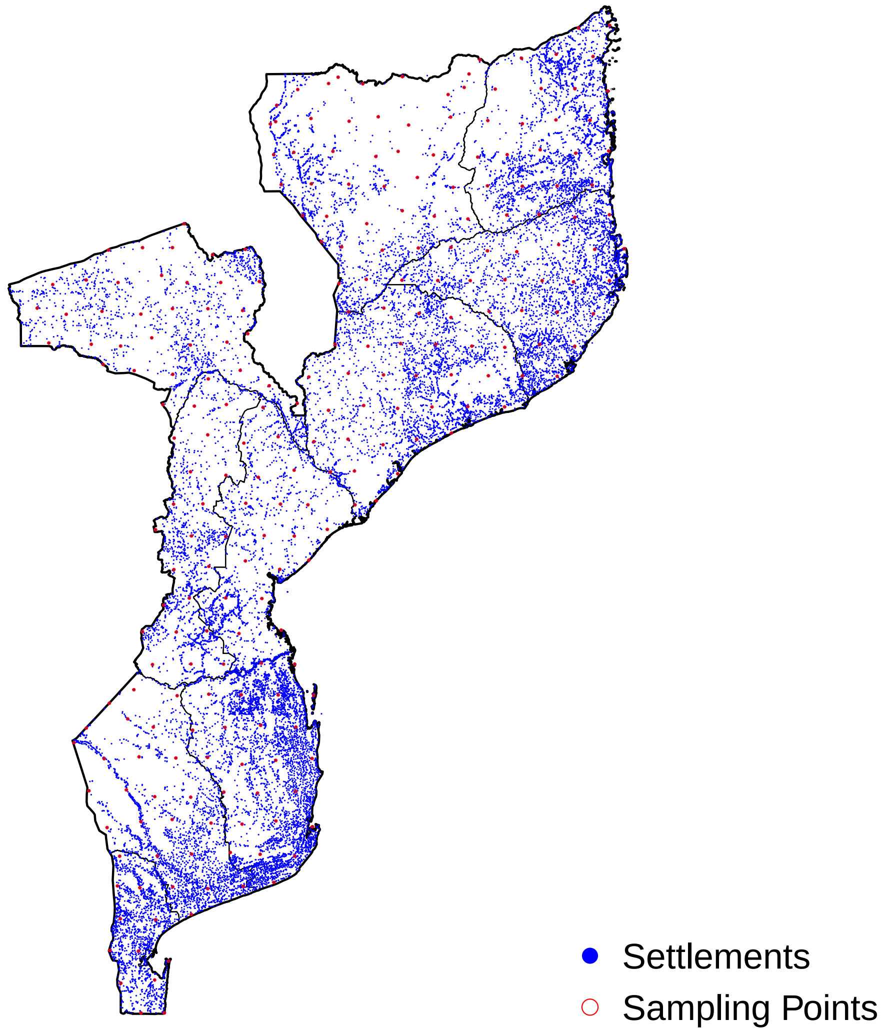

Figure 9: Stage 1 sampling map for d of 35 kms

References

M Myatt, E Guevarra, L Fieschi, A Norris, Saul Guerrero, Lily Schofield, Daniel Jones, Ephrem Emru, and K Sadler. Semi-Quantitative Evaluation of Access and Coverage (SQUEAC)/ Simplified Lot Quality Assurance Sampling Evaluation of Access and Coverage (SLEAC) Technical Reference. FHI 360/FANTA, 2012.