Improving Capacity in Forest Resources Assessment in Kenya (IC-FRA)

Technical report on Sampling Design Simulations for National Forest Resources Assessment in Kenya

May 2016

Prepared by:
The Project *Improving Capacity in Forest Resources Assessment in Kenya (IC-FRA)* implemented 2013–2015

Compiled by:
Pekka Hyvönen, Juha Heikkinen, Helena Haakana, András Balázs, Mbae N. Muchiri, Balozi B. Kirongo, Peter N. Nduati, Fredrick Ojuang

Cover caption: Front page photograph by Pekka Hyvönen: Natural forest in Nakuru National Park

Table of contents

List of Figures

List of Tables

Acknowledgement

This Technical Report on Sampling Design Simulations for National Forest Resources Assessment (NFRA) in Kenya is a product of effort by a large number of people and institutions. The compilers of the report would sincerely like to extend their gratitude to institutions that provided staff and allowed them to support this noble initiative.

We most sincerely thank the Government of Finland and Government of Kenya for providing financial support for the IC-FRA Project that has built capacity in terms of staff, equipment and knowledge in Forest Resources Assessment in Kenya. Special thanks are extended to Directors of Luke, DRSRS, KEFRI, KFS, and the Vice Chancellor University of Eldoret for allowing their staff to participate in the Project.

1 Introduction

One target of the IC-FRA project was to study and suggest feasible sampling design for the National Forest Resources Assessment (NFRA) in Kenya. The sampling design can be based on the results of previous inventories or in absence of the results of past inventories, on other relevant methods including sampling design simulations.

The idea of the sampling design simulation is to simulate (imitate) the real inventory. In the simulation, the real forest is replaced with the generated forest. In this study the generated forest means the digital wall-to-wall biomass map which was produced with the *k* nearest neighbour (*k*nn) estimation method utilising measured sample plot data and Landsat satellite images. In this study the sample plot data from the IC-FRA pilot inventory carried out on five test areas in different vegetation types in Kenya was applied.

In simulations different designs can be tested and repeated as many times as needed without extra costs, unlike in the real forest inventory. The best design is then selected with the certain criteria by comparing results from the different simulated sampling designs.

Luke has developed an R-code for the sampling design simulator. The basic simulator can be modified for different purposes and can use different materials like tiff-images. These images can have much information including biomass on a land area and the walking times on different forest types. The results of the basic simulator are the estimates of the forest area, the mean and the total values of the variable in interests (e.g. biomass or volume) and also the error estimates of these variables. As several materials can be utilised in the simulator, it is possible to calculate the time costs for different designs. The basic simulator was used to calculate results for the Nakuru and Gazi test areas.

The latest development of the simulator is “doubly stratified two-phase sampling”. This method was utilised in the sampling design for the whole Kenya. In this method, a very dense systematic grid of clusters is created in the first phase. The grid can vary between 1st-phase strata. The 1st-phase clusters are then stratified into 2nd-phase strata. This can be done e.g. with the help of auxiliary information. There can be different rules for stratifying clusters. In this exercise clusters were stratified based on the number of forested plots in the cluster. The forest type map of the Kenya was used to stratify the 1st-phase strata into 2nd-phase strata. On the next step, the 2nd-phase sample to be measured in the field is selected by random sampling from the 1st-phase sample. This was done separately and independently within each 2nd-phase stratum. As inclusion probabilities can vary between the 2nd-phase strata, more effort can be focused on sampling more interesting clusters, e.g., those containing many forested plots. Allocation of clusters to the different strata was based on optimal allocation minimizing the standard error of the total forest biomass with time cost as a limiting factor (Chapter 4.3)

The two methods (basic simulator and double sampling simulator) were used because with the sample plot data from the IC-FRA pilot inventory, forest biomass map could not be estimated for the whole country (see 3.1). Hence, the biomass map was estimated for two smaller test areas (Nakuru and Gazi) and suitable cluster designs were studied on those areas with the “basic simulator”. Basically, testing of different distances between sample plots on a cluster and cluster shapes would have been enough in this stage because in the sampling method used for the whole Kenya (two-phase sampling) the clusters to be measured in the field are selected from the 1st-phase grid randomly, that is, with varying distances. As the “doubly stratified two-phase sampling” simulator was not yet fully functioning when the study was started, also distances between clusters were tested on Nakuru and Gazi test areas.

The next stage was to decide strata for the whole Kenya and start simulations with the “doubly stratified two-phase sampling” simulator. The idea of dividing Kenya into different spatial strata gives opportunity to utilize different sampling and cluster designs in different parts of Kenya to achieve and guarantee more accurate results. The strata tested were based on the county boundaries and agro- ecological zones. In addition, mangrove forests in the coast formed a stratum (Figure 1 and Chapter 3.4). Dividing Kenya into four strata was agreed in a IC-FRA Project Technical Working Group meeting in late 2014. The strata 1 and 2 were slightly modified following a discussion on the strata homogeneity in a IC-FRA Project in-house workshop in February 2015.

The purpose of this technical report is to describe the procedures and results in developing a sampling design for National Forest Resources Assessment (NFRA) in Kenya. Further, the results and findings of this sampling design study were utilised in the formulation of the "Proposal for Kenya Forest Resources Assessment".

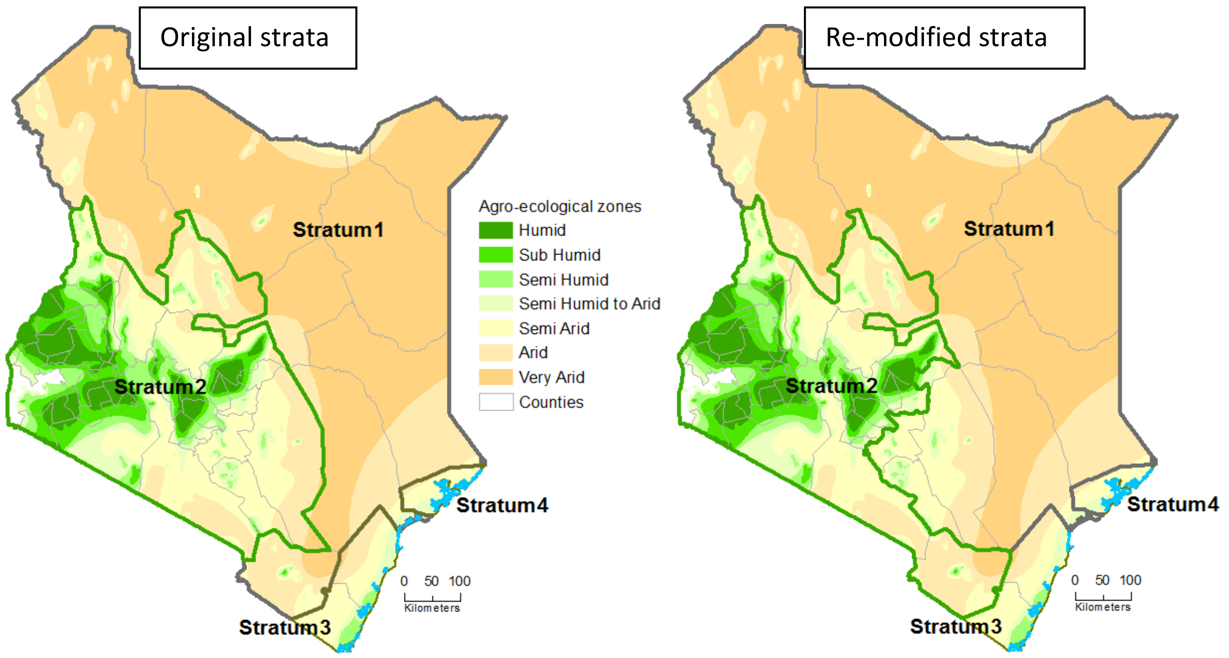

Figure 1. Strata for Kenya.
On the left the original strata and on the right "re-modified" stratum1 and stratum2. (County boundaries with thin line and each stratum with bold line).

2 Data for the Nakuru and Gazi test areas

2.1 General

It was agreed early in 2014 that the test area for sampling design simulations will include major vegetation types in Kenya: forest plantations, indigenous forest including bamboo, woodlands, forests on farms and mangroves. The target was to cover at least the first four areas. The four respective sites chosen were: Kericho, Aberdare, Marigat and Nakuru. The mangrove sites were included later.

2.2 Biomass images

The biomass image for the Nakuru test area was processed in April 2014. The spatial reference of this data was Arc 1960 UTM 37S. The map was produced with the *k*nn method utilising Landsat 8 images and the IC-FRA Pilot inventory sample plot data collected in October 2013 – January 2014. The plot data was processed by Dr Pekka Hyvönen during spring 2014. Two Landsat images were used for the Nakuru test area: LC81690602013278LGN00 (Path 169, Row 60, October 2013) and LC81680602013255LGN00 (Path 169, Row 60, September 2013). The spatial resolution of the estimated biomass image was 30 m x 30 m and it covered the area of 172 km x 133 km (22 876 km^2) (Figure 2, Figure 3).

The biomass image of the Gazi test area was processed in October 2014. Landsat image LC81660632014036LGN00 (Path 166, Row 63, February 2014) was used. The spatial resolution of the estimated biomass image was 30 m x 30 m and it covered the area of 54 km x 33 km of which the land area was about 1 250 km^2. Both images were processed by Luke's remote sensing expert András Balázs.

The vegetation type image for the area of interest was rasterized from the vector data (forest_type_2010.shp and forest_legal_2010.shp) which were received from KFS in February 2013. The spatial reference of this data was Arc1960 UTM 37S. The vegetation type map was merged with the biomass map. The vegetation type codes of this image are shown in Table 1.

Table 1. Vegetation type codes in the biomass image and the estimated walking and measuring times.

CODE	Cnt_ftype	Vegetation type	Walking time min/km	Walking time min/m	Measuring time min/plot*	
1	35	Bamboo	40	0.04	90	80
6	97	Cropland	20	0.02	40	40
5	996	Grassland	20	0.02	40	40
3	1816	NaturalForest	40	0.04	120	90
9	3	Otherland	20	0.02	30	20
4	608	PlantationForest	20	0.02	130	90
8	73	Settlements	15	0.015	40	30
7	71	Wetlands	60	0.06	30	20
2	-	Mangrove	60	0.06	120	90

*Measuring times in the last columns were used for "Kenya ext" simulations and the other times for Nakuru and Gazi test area simulations.

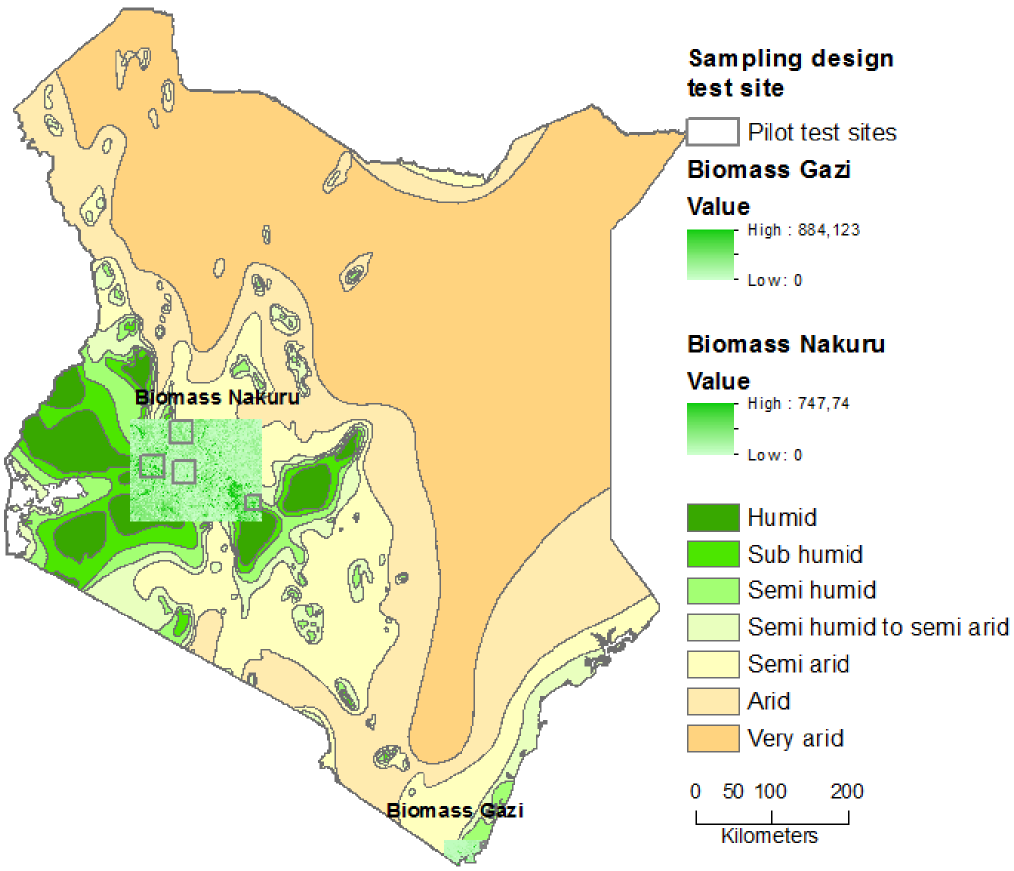

Figure 2. Location of the sampling design test area in the Nakuru area.
(Agro-ecological zones are shown in colours).

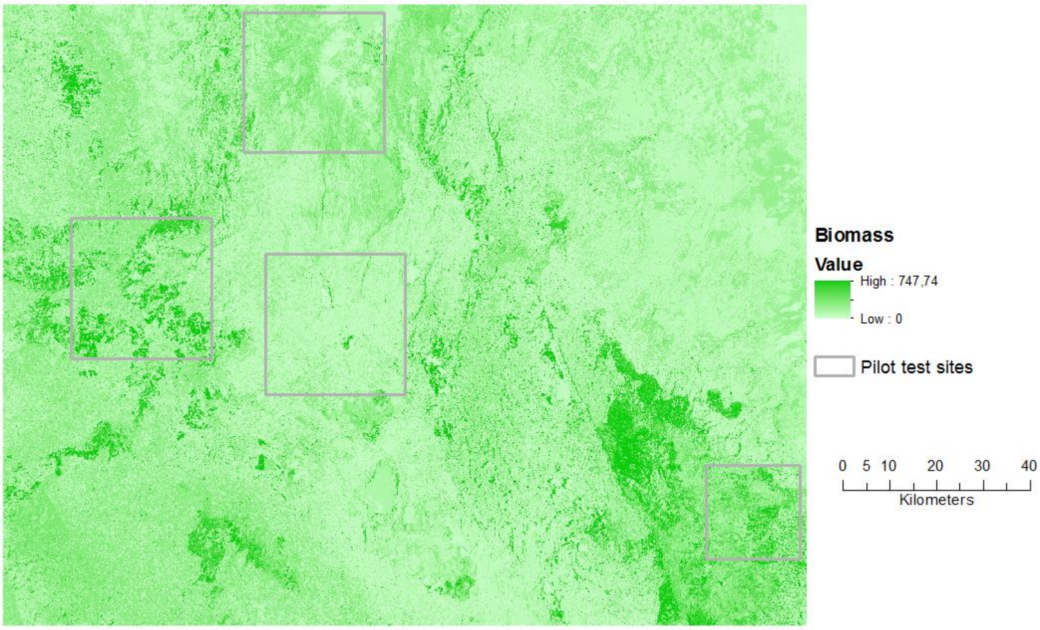

Figure 3. The estimated biomass in the Nakuru area.

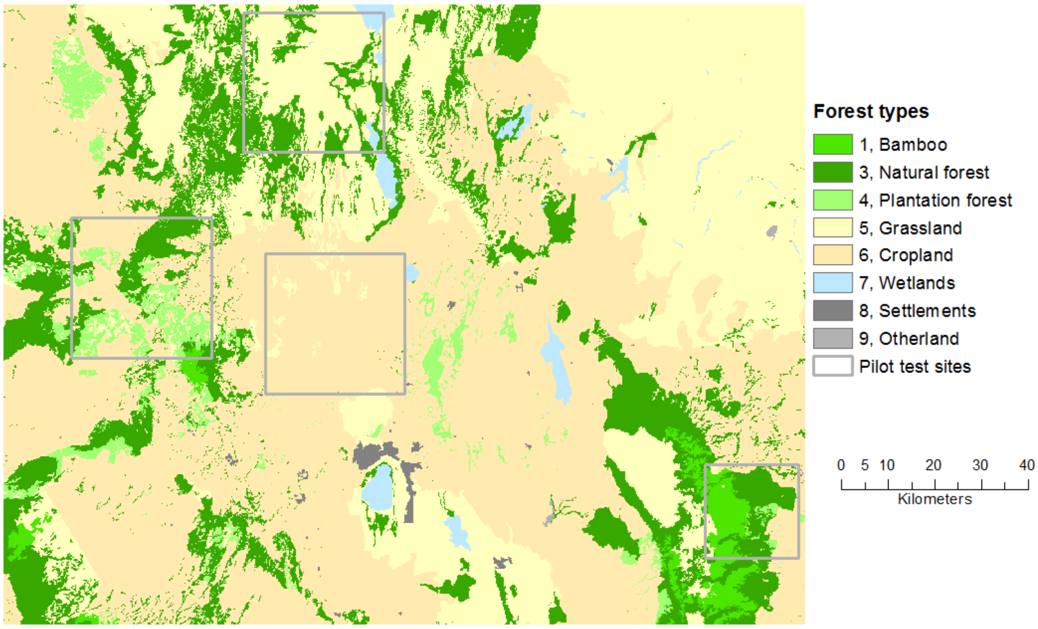

Figure 4. Vegetation types in the Nakuru area.

2.3 Roads

Data of roads for the whole Kenya was received from MMMB Technical Advisor Dr Martin Schweter during spring 2014. The spatial reference of this data was Arc 1960. The roads map was clipped with the area of interest (bm_area_buf5km.shp). A buffer of 5 kilometres around the area of interest was used in order to avoid extra costs in borders. Roads had several classes (Table 2 and Figure 5). The following classes were excluded from the road data: railway, cutline and causeway.

Table 2. Classes of roads.

CODE	Cnt_CODE	First_TYPE
0	3	Railway
1	141	Bound Surface
2	350	Loose Surface
3	765	Dry Weather Road
4	5712	Main Track (Mortorable)
5	12715	Other Tracks and Footpaths
6	1	Causeway
7	340	Cutline
8	53	Railway

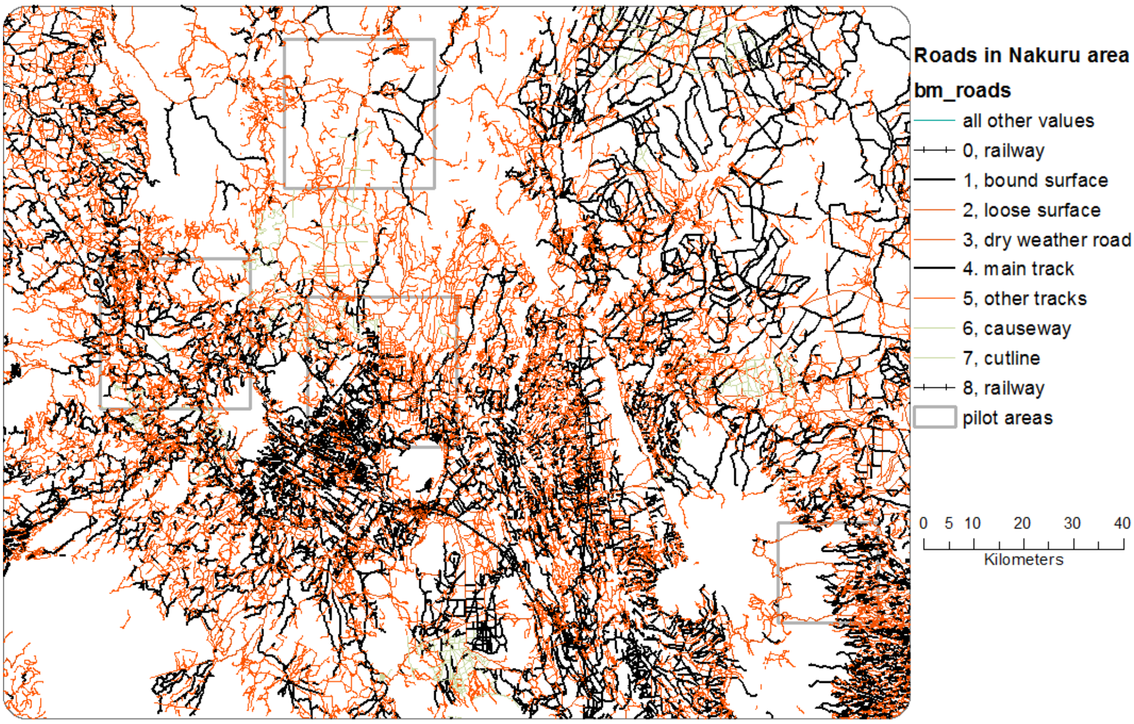

Figure 5. Map of the roads in the Nakuru area. The Pilot inventory test areas are (from the left) Kericho, Marigat, Nakuru and Aberdare.

2.4 Digital elevation model (DEM)

DEM (Aster Global DEM, 30 x 30 m) was downloaded from the USGS web service, http://gdex.cr.usgs.gov/gdex/ (21st August) covering the biomass map area. The spatial reference for this image was WGS84 UTM 37N and it was reprojected to Arc 1960 UTM 37S (as the other data) (Figure 6).

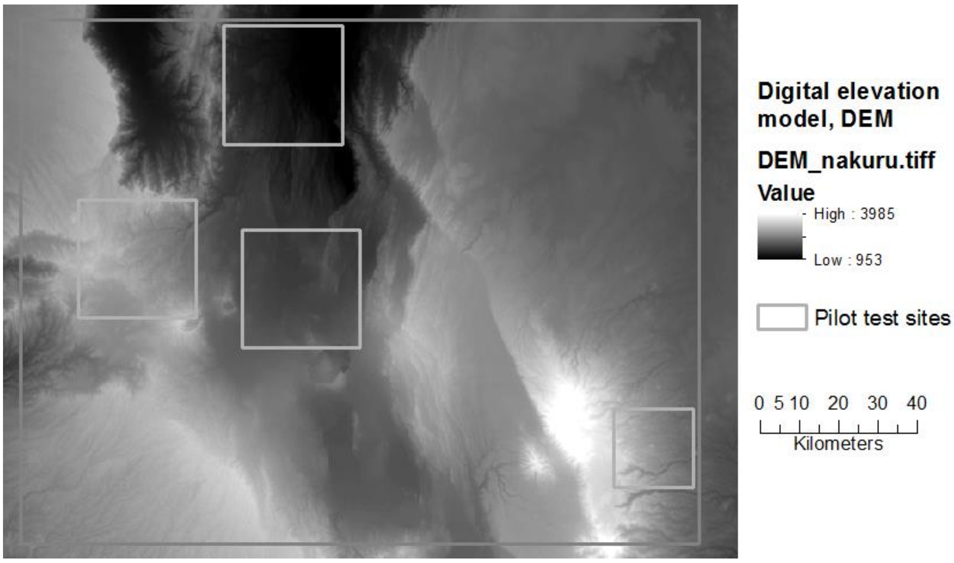

Figure 6. The digital elevation model (DEM) of the Nakuru area.

2.5 Processed images

2.5.1 Distance

The distance from the nearest road taking to account the DEM was calculated for each raster cell (Figure 7). This was done with the Pathdistance function of the ArcGIS software. The Pathdistance function determines the minimum travel cost from the road to the sample plot.

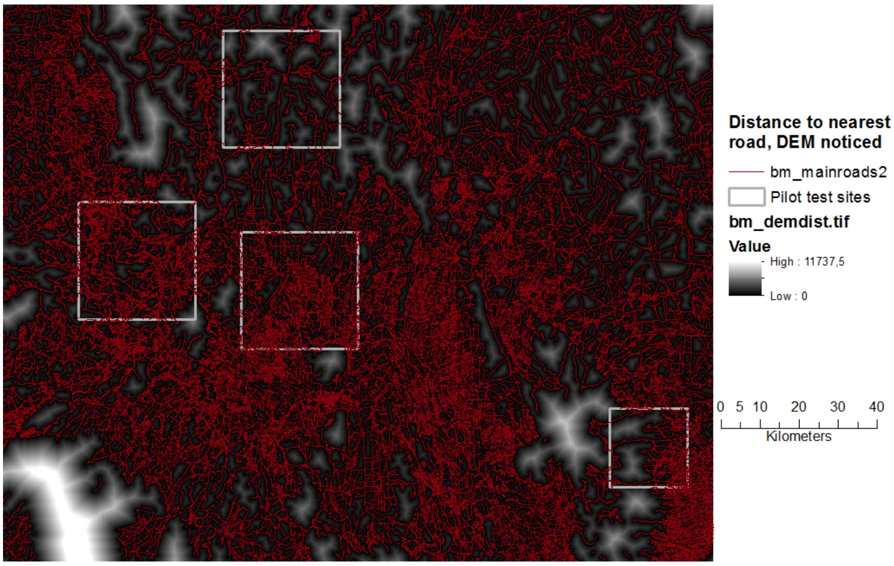

Figure 7. Distance from the road taking into account the DEM. (Roads are shown in red).

2.5.2 Slope

Slope (as percentages) was calculated from the DEM with ArcGIS (Figure 8).

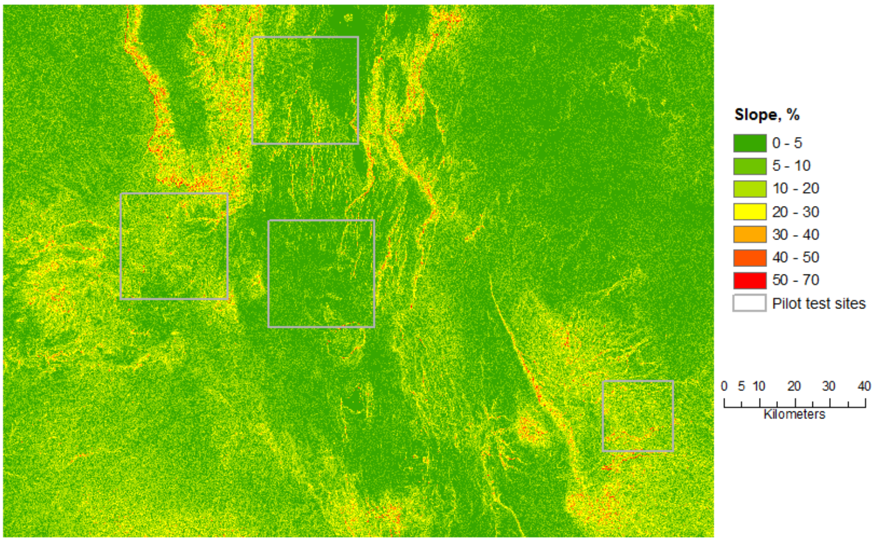

Figure 8. Slope raster in the Nakuru area.

2.5.3 Walking time

Walking time in the different vegetation types was calculated with the help of the time consumption data from the IC-FRA Pilot inventory (Table 1). Appropriate walking times were confirmed also by Fredrick Ojuang who was the overall in charge of the field inventory. First, an image with the walking time in minutes/kilometres was calculated and then, the values were divided by 1 000. Thus, the unit in the walking time image was minutes/metre.

The previously processed slope image was further processed by "the vertical factor correction" (Table 3). This correction takes into account the fact that moving downhill or uphill takes more time than moving on a flat surface. For example, for a slope of 15 degrees the original walking time was multiplied by 1.2. The walking time image was multiplied with this new slope image. As a result, walking-time image which takes into account the forest type and the slope was obtained (Figure 9). This image was used in the sampling simulations to calculate the walking time between the sample plots in a cluster.

Table 3. Vertical factor correction (vf.txt).

Slope from (degrees)	Slope to (degrees)	Vertical factor
0	5	1.0
5	10	1.1
10	15	1.2
15	20	1.4
20	40	2.1
40	50	3.1
50	80	5.0

Another walking time image (time distance) was calculated taking into account the slope, the distance to the nearest road and the walking time as a cost (Figure 10). The same Pathdistance function was used as in calculating the distance image. This image was used in the sampling simulations to estimate the time distance from the sample plot (cluster) to the nearest road.

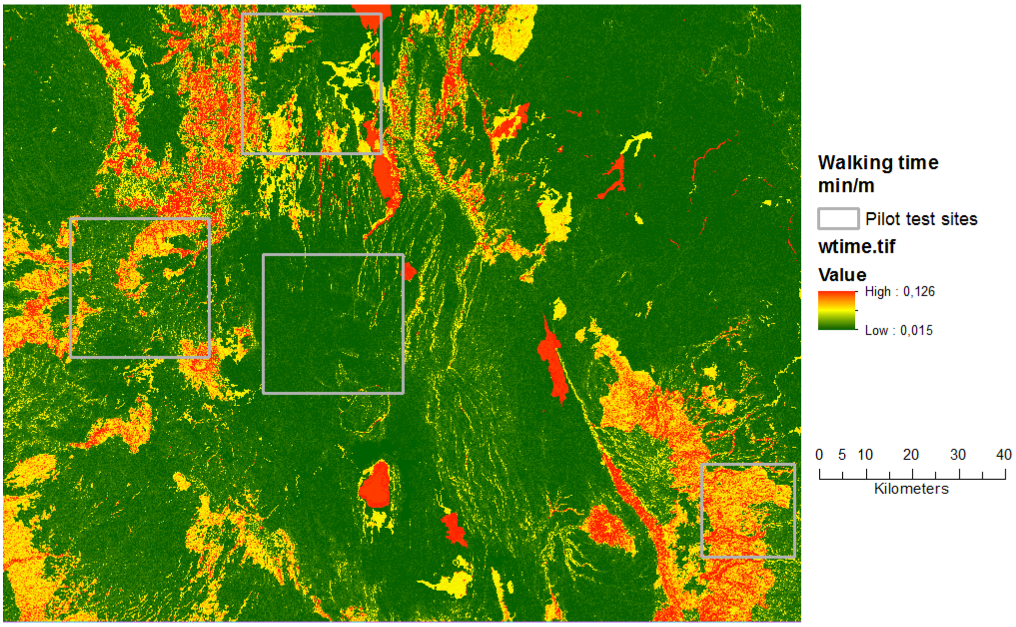

Figure 9. The walking time (min/m) in the Nakuru area taking into account the vegetation type and the slope.

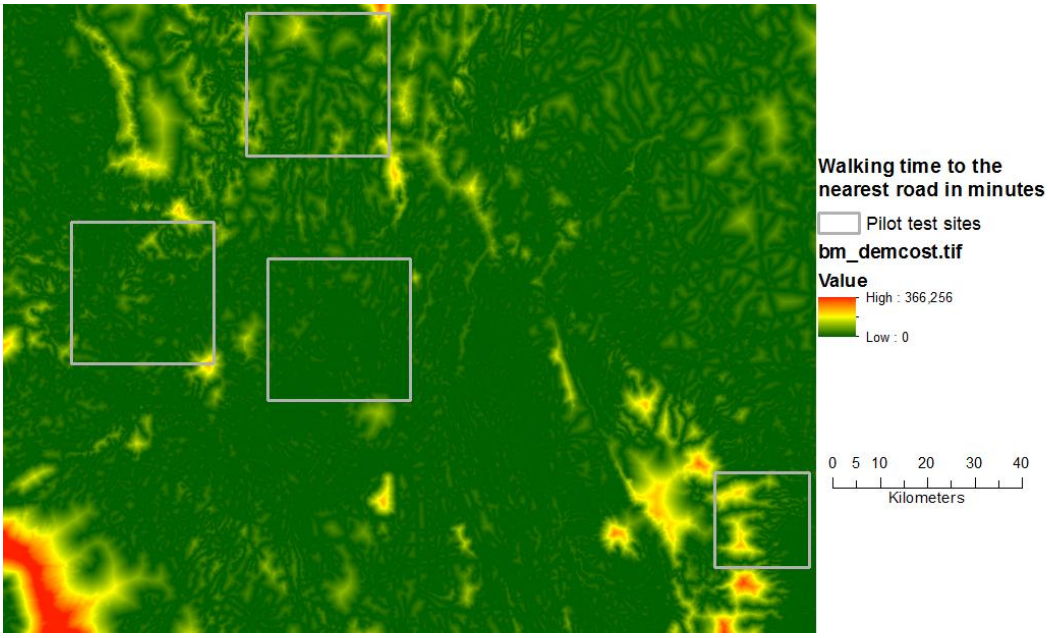

Figure 10. The walking time (minutes) to the nearest road taking into account the DEM and the walking time as a cost.

2.5.4 Measuring time

The measuring time for sample plots in the different vegetation types was analysed from the Pilot inventory data (see IC-FRA 2016, Technical report on the Pilot inventory). Based on these analyses, measuring times were decided as shown in Table 1. In practice, the vegetation type image was reclassified with these values (Figure 11). The cost of measuring time (if done this way) could also be inputted straight to the R script (without image data) as the vegetation type was included in the biomass image.

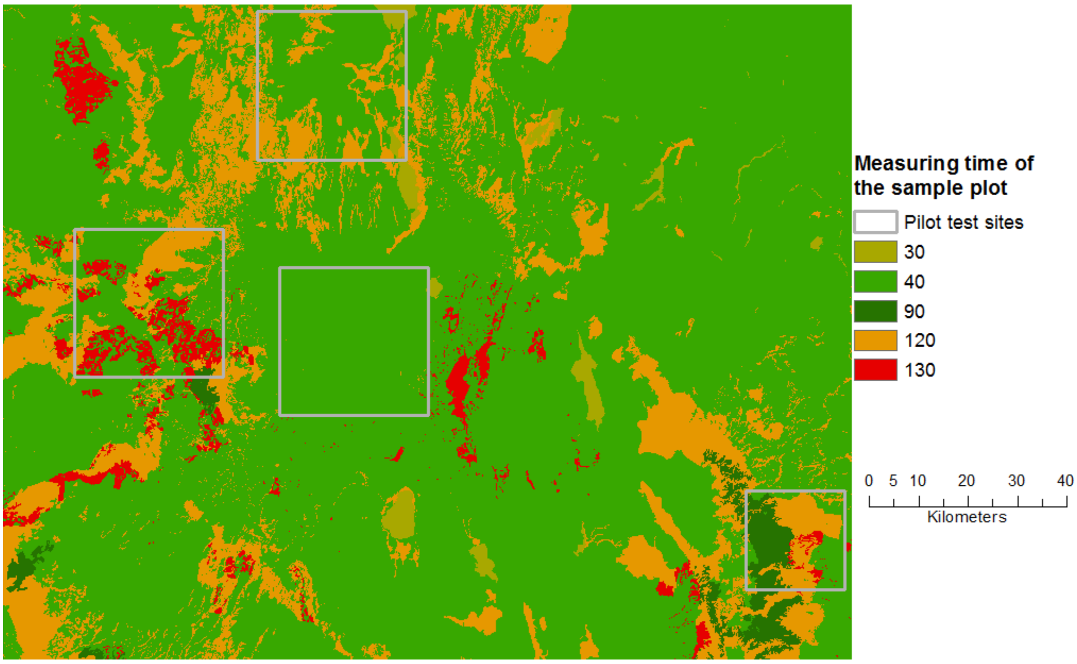

Figure 11. The measuring time of sample plots (min/plot) in the different vegetation types.

3 Data for the whole country

3.1 General

Data for the sampling design simulations for the whole country was prepared basically in a same way as for the Nakuru and Gazi test areas. The difference was the idea of how the biomass image was produced. About 35–40 scenes would be needed to cover the whole country with Landsat images. Processing of that amount of images (e.g. calibration, mosaic) would take a lot of effort. In addition, cloud free (clouds covering less than 5% of the area) images are hard to get, especially in the coastal areas. Therefore, another solution was applied to estimate the biomass image for Kenya, see Chapter 3.3.

3.2 Boundaries of Kenya

As the Pilot inventory data was studied, it was noticed that large areas on the coast were not officially inside Kenya. This is because the Land Survey of Kenya is done when the sea tide is high. This means that, e.g., large areas of mangrove forests are excluded; they are not inside the "official" boundaries. After several inquiries and discussions, some updated maps (boundaries) were found from the databases of KFS. However, also those maps had some incoherence compared to the older map data; e.g. the coastline was drawn with straight lines. In the map data from FAO the coastline seemed to be based on remote sensing material (http://data.fao.org/map?entryId=d028eb8a-00e8-4cf2-81f4-00c21f6f6883) and also accurate (when compared to Google Earth data and downloaded Landsat images). Thus, this data was used as "extended boundaries" for Kenya.

3.3 Biomass image

The map data of forest types in Kenya has been produced in Forest Preservation Programme (FPP, 2013). The data is based on classified remote sensing data and it is validated by field visits. The ALOS AVNIR-2 (189 SCENES of Date 20/3/2009 - 15/2/2011 and DMC (9 SCENES of 6/9/2010 - 13/4/2011) were used in the classification (FPP, 2013). The data of forest types was used for estimating a biomass map for Kenya. The mean biomass and its standard deviation by forest types and agro-ecological zones were derived from the estimated biomass images of Nakuru and Gazi test areas. These values were used to reclassify the forest type map. In other words, each forest type in each agro-ecological zone was given the mean value observed in the test areas and the standard deviation was used to generate variation for the mean values.

3.4 Mangrove data

The FPP had produced also mangrove data based on remote sensing material in the coastal area. Compared to the images on Google Earth and the downloaded Landsat images, the mangrove maps seemed to be somewhat inaccurate. There were, e.g., forests on upland that were classified as mangrove. As this was noticed, other maps of mangroves were inquired from the Kenya Maritime and Fisheries Research Institute (KEMFRI). Very detailed mangrove maps were received at the end of October 2014. Also this data seemed to include some areas that were not mangrove forest. On the other hand, some parts of mangroves seemed to be missing. To be sure that all mangroves (as much as possible with these data sets) are included in further analysis, these two mangrove data sets (from KFS/FPP and KMFRI) were merged and a 100 meter buffer zone around mangrove polygons was created (Figure 12). This buffered mangrove data was not corrected manually, and consequently, some narrow polygons between zones are excluded from and areas in water included in the mangrove data.

Figure 12. Mangroves in the Gazi bay.
With green mangrove data from KFS and with black lines from KEMFRI. Blue line indicates the merged mangrove area buffered with 100 meters.

3.5 Other images

Other image data was processed in a similar way as for the Nakuru and Gazi test areas, see Chapter 2.5.

4 Simulations

4.1 Main principles

As said in the introduction, the idea of sampling design simulation is to simulate (imitate) the real inventory. By varying cluster form (e.g. L-shape), distances between sample plots in a cluster and distances between clusters, different sampling designs can be tested and compared. Simulation steps were as follows:

1. choose design (cluster form, number of plots in a cluster, distances)
2. generate a systematic grid for the area of interest
3. extract values for sample plots (biomass, forest type, time and cost information)
4. calculate results with this data set
5. repeat the steps 2–4 500 times and move the grid randomly (in step 2).

As a final result of each design, the variation of the results (e.g. standard error of the mean biomass) is the measure of “goodness” for that design. Then the steps 1–5 are repeated for another design. The final results from different designs are then evaluated in order to find the most suitable one.

Several cluster forms were tested and evaluated in terms of the forest cover and the mean biomass, i.e. comparing empirical and theoretical (relative) errors they result in. The estimators (the total and the mean forest biomass, proportion of the vegetation type and the errors (empirical and theoretical)) were calculated with the same estimation method (point estimates) as in the pilot inventory (see IC-FRA 2016, Technical Report on the Pilot inventory).

The total time cost was calculated for each design. The total time consisted of the measuring time of the sample plots, the walking time to/from the cluster from/to the road and the walking time on the cluster between the sample plots. Driving time from the accommodation to the cluster (car park) and back was not included. The walking time on a cluster was calculated with the following steps:

- calculate the average walking time between adjacent sample plots: sample plots 1 and 2, sample plots 2 and 3 etc. and sample plot 1 and the last one.
- multiply the average walking time with the real distance (taking account the terrain) between adjacent sample plots
- tally up the walking times to get the total walking time on the cluster
- find the shortest time distance from the road to the cluster
- multiply the time distance by two and tally up with the walking time on the cluster to get the total time on the cluster (cost).

Note; “Total costs hours” (Table 5) is calculated based on an idea of “overnight in the field”. This means that the field teams are supposed to overnight in the cluster in those cases when the sample plots on a cluster are not measured during one working day. “Total cost days” (Table 6) is calculated with the idea that the field teams work in the field 7 hours per day (time from the accommodation to the car park and back is excluded). If the cluster is not measured in one day, the field team will return to the cluster on another day. Time from the car park to the cluster and back is therefore included in time costs as many times as the field team visits the cluster; but not more times than the cluster has sample plots. In other words, a cluster having 5 sample plots is visited in maximum on 5 days and time from/to the car park is added in maximum 5 times to the calculations. In addition, sample plots locating in water are not included in the time calculations (not visited in the field).

The tested distances between sample plots were based on a preliminary study with semivariograms and variograms. Variograms describe the spatial relationship between the variable of interest; in this case how correlated the volumes on sample plots are on different distances. The same was studied also with the estimated biomasses. Spatial correlation was analysed with the help of the sample plot data measured in the IC-FRA pilot inventory (Figures 13 and 14) and with the help of the estimated biomass image (Figure 15).

According to the variograms the optimal distance between sample plots in a cluster might be near 200–250 meters. However, the results with variograms were not clear. Several cluster designs were tested. Basic cluster forms are shown in Figure 16.

.

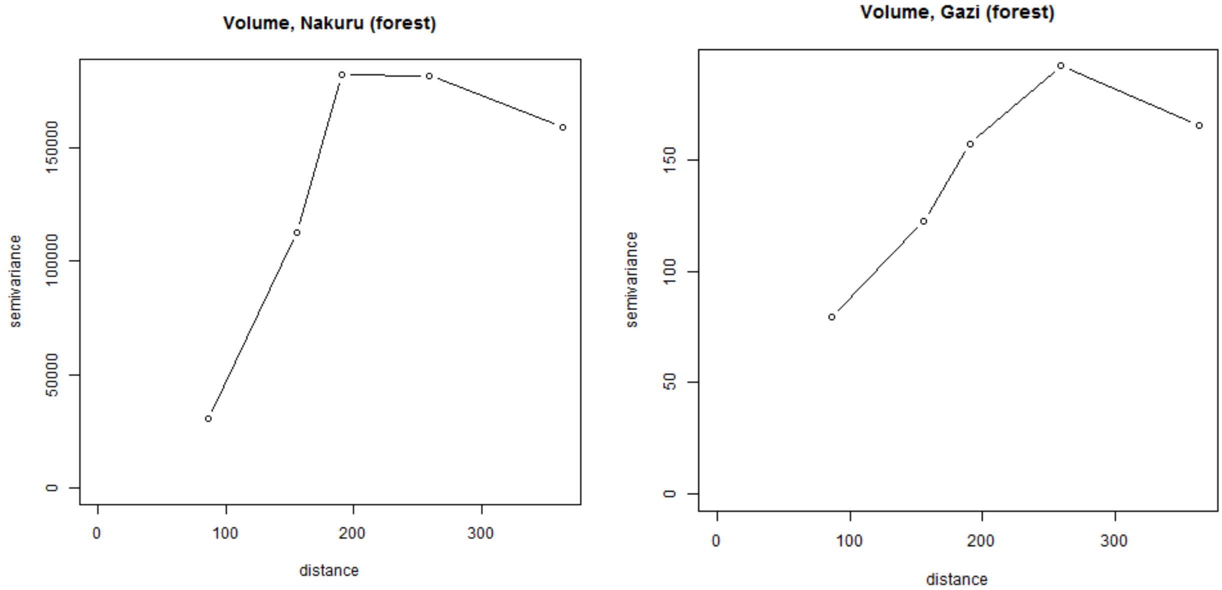

Figure 13. Variograms of the volume based on the sample plot data of the IC-FRA pilot inventory in the upland test areas (Nakuru) and in mangroves (Gazi).

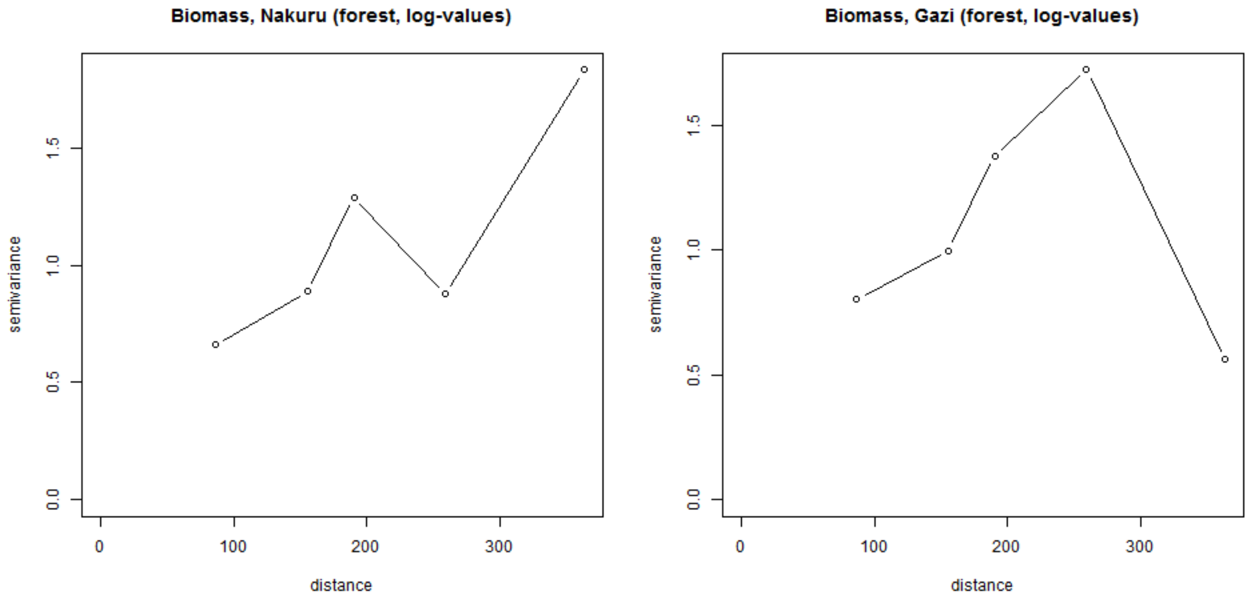

Figure 14. Variograms of the biomass based on the sample plot data of the IC-FRA pilot inventory in the upland test areas (Nakuru) and mangroves (Gazi).

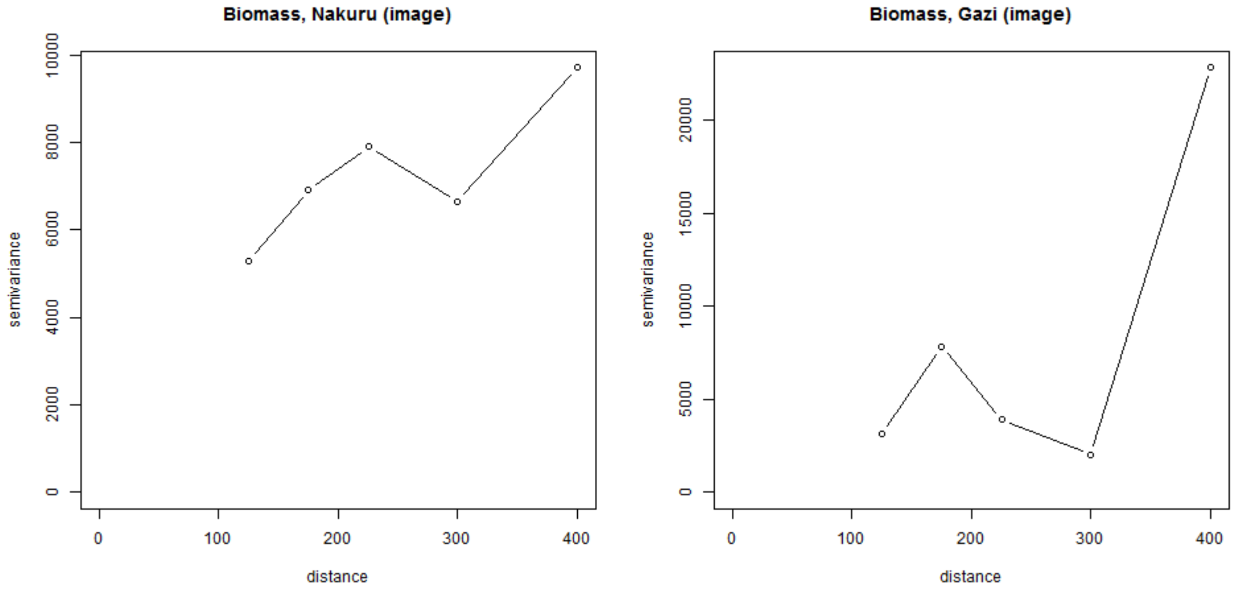

Figure 15. Variograms of the biomass based on the estimated biomass image in the Nakuru test area (left) and in the Gazi test area (right).

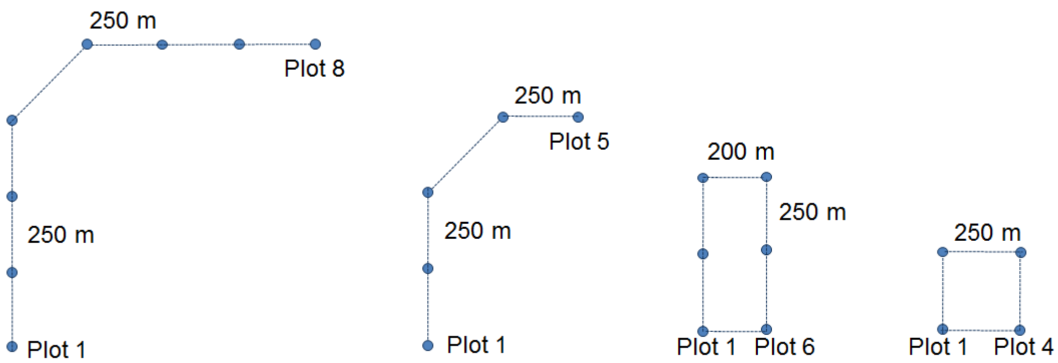

Figure 16. Tested cluster forms. From the left, L-shape with 8 sample plots, L-shape with 5 sample plots, rectangular and square.

4.2 Nakuru and Gazi test areas

Sampling design simulations were carried out with above data sets in the Nakuru and Gazi test areas. The sampling method applied was systematic cluster sampling, with and without stratification. In other words, results were calculated with each design both with and without stratification. The strata in the Nakuru test area were formed roughly with the help of agro-ecological zones, Figure 17.

In the Gazi test area, stratification was done according to mangroves. The mangrove data from KFS and KEMFRI were merged together and buffered with 100 m. This formed stratum 1 (gazi1) and the rest of the area formed stratum 2 (gazi2), Figure 18. The areas of forest types in the strata are shown in Table 4.

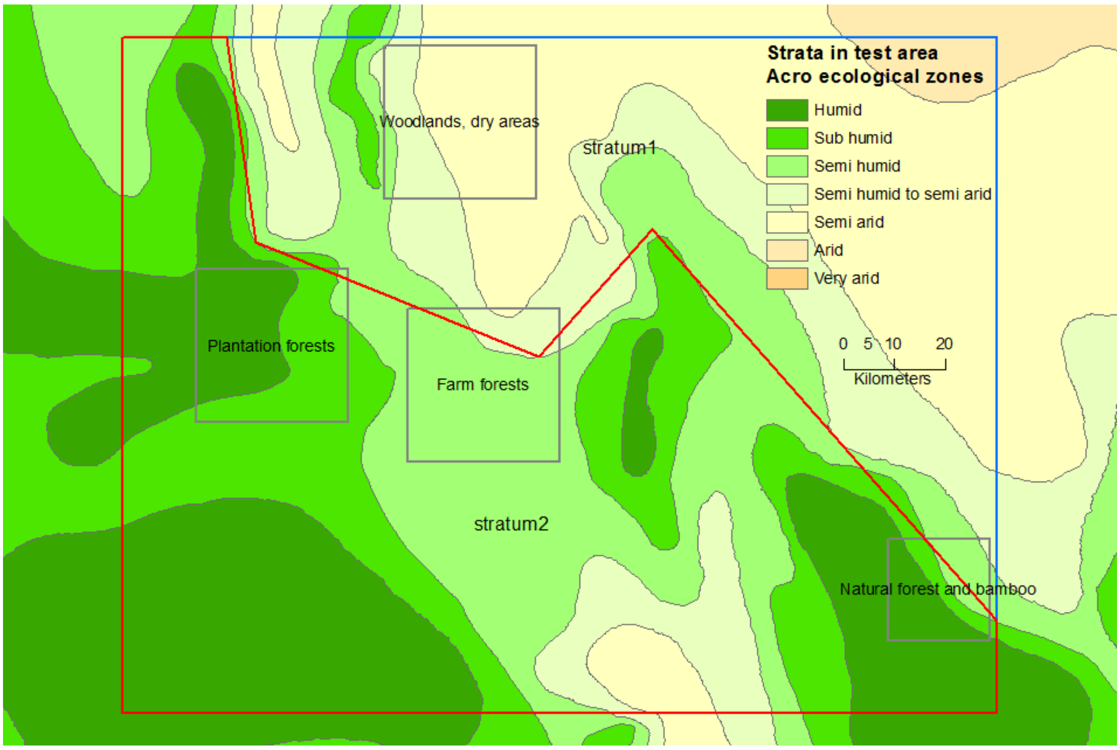

Figure 17. Strata in the Nakuru test area. Stratum 1 with blue line and stratum 2 with red line.

Table 4. Area of different forest types in square kilometres (km^2) in the Nakuru and Gazi test areas.

	Nakuru			Gazi		
	Stratum1	Stratum2	In total	Gazi1	Gazi2	In total
Bamboo	0.00	357.64	357.64	0.00	0.00	0.00
Mangrove	0.00	0.00	0.00	87.14	0.00	87.14
Natural Forest	1,490.46	2,114.18	3,604.64	1.66	89.59	91.25
Plantation	3.16	648.93	652.09	0.00	0.00	0.00
Forest in total	*1,493.6*	*3,120.7*	*4,614.4*	*88.8*	*89.6*	*178.4*
Grassland	5,190.88	1,564.22	6,755.09	10.79	9.86	20.65
Cropland	2,436.53	8,692.74	1,1129.28	9.23	1,007.15	1,016.38
Settlements	3.23	86.15	89.37	0.09	0.19	0.28
Other land	4.63	4.42	9.05	17.55	20.95	38.51
Land in total	*9,128.9*	*1,3468.2*	*2,2597.2*	*126.56*	*1127.8*	*1254.2*
Wetlands	139.58	123.08	262.66	0.88	0.87	1.75
In total	9,268.5	13,591.3	22,859.9	127.4	1,128.8	1,256.0

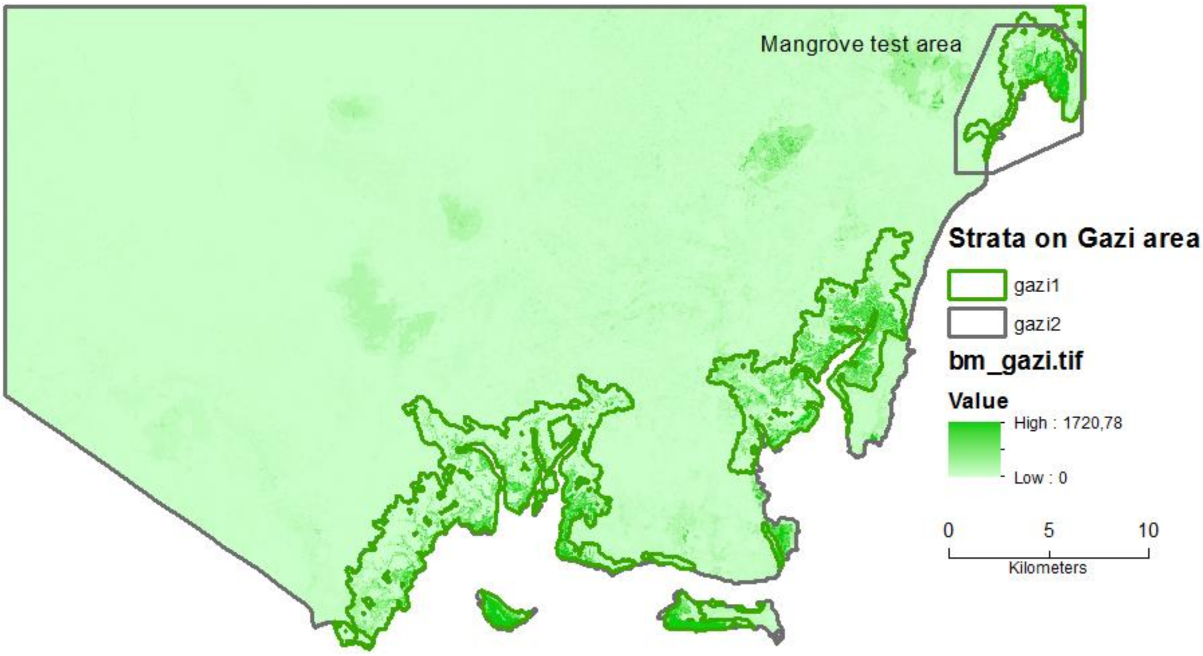

Figure 18. Strata and biomass map in the Gazi test area.

The best combinations (stratified design) according to the relative errors and time consumption are presented in Tables 5 and 6. It was found out that the relative errors decrease as the distance between sample plots increases, Figure 19 – Figure 22. On the other hand, this increases time costs as more time is spent on walking between sample plots, Figure 23 – Figure 25.

Also decreasing the distance between clusters will decrease errors. This is logical as there are more sample plots to measure with shorter distances between clusters than with longer distances.

It seemed that the L-shape and rectangular clusters (with the same cluster distance and with 6 sample plots) gave quite similar results with the forest area, Figure 19 and Figure 21. Errors of mean biomasses were slightly smaller with the L-shape cluster than with the rectangular one. On the other hand, with the L-shape cluster the time cost was clearly higher than with the corresponding rectangular one, Figure 24. The square-shape cluster design resulted in the biggest errors in the forest area and mean biomass estimates in both areas. However, it must be noted that at a given time it might be possible to measure more square-shape clusters than L-shape or rectangular-shape clusters, Figure 26. This originates from the number of sample plots in a cluster; in square 4 but in L or rectangular 5–6. So, adding clusters with the square cluster design might lead to the same accuracy as the other designs. This was tested further with the “doubly stratified two-phase sampling” method for the whole country.

Table 5. Part of the sampling design results in the Nakuru test area.

Cluster	Cluster distance	Plots in cluster	Plot distance	Plots on land	Forest cover		Mean biomass		Total costs
	(km)		(m)		%	s.e. (%)	ton/ha	s.e. (%)	hours
L-shape	6	5	150	3,139	20.5	4.12	76.2	8.4	3,710
			200	3,138	20.4	3.97	75.7	8.3	3,786
			250	3,134	20.4	3.55	76.4	7.8	3,868
	7	5	250	2,300	20.5	5.37	75.9	7.2	2,841
	8-7*	5	200-300	2,087	20.5	5.18	76.1	7.3	2,614
	7	6	250	3,170	20.4	3.78	75.8	6.6	3,889
			300	3,169	20.4	3.68	76.1	6.4	3,998
	8-7	6	300	2,506	20.4	4.81	75.7	6.7	3,251
Rec-tangle	6	6	200-250	3,766	20.4	3.96	76.1	8.0	4,403
			300	3,757	20.5	4.82	76.3	6.9	4,514
			250-300	3,769	20.4	3.99	76.2	7.5	4,458
			250-350	3,766	20.5	3.73	76.5	7.5	4,483
	7	6	250	2,766	20.4	5.92	75.9	7.5	3,289
	8-7	6	200-250	2,507	20.4	5.31	75.9	8.0	2,958
			250-300	2,502	20.4	5.13	76.4	7.8	2,985
Square	7-6	4	250	2,245	20.6	6.11	76.3	8.0	2,797
			350	2,237	20.4	5.72	76.3	7.6	2,868
	7	4	150-200	1,842	20.5	5.50	76.1	8.9	2,204
			200-250	1,840	20.5	5.72	76.3	8.7	2,224
	8-7	4	200	1,673	20.5	6.10	75.3	9.3	2,054
			300	1,668	20.4	5.54	76.4	8.2	2,128

*) Distance between clusters was 8 km in stratum1 and 7 km in stratum 2, the distances between sample plots were 200 m and 300 m, respectively.

Table 6. Part of sampling design results in the Gazi test area.

Cluster	Cluster distance	Plots in cluster	Plot dist.	Plots on land	Forest cover			Mean biomass		Total costs	
	(km)		(m)		%	s.e. (%)*		ton/ha	s.e. (%)	days	hours
L-shape	5	5	150	243	15.5	16.90	14.42	131.1	38.0	56	245
			200	242	15.2	16.48	14.09	132.5	32.6	57	251
			250	242	15.2	16.17	12.47	132.4	32.1	58	258
	3-5**	5	150	281	15.4	16.06	7.62	132.7	24.8	70	323
			200	281	15.5	15.70	7.66	132.0	22.6	72	334
			250	280	15.6	14.90	7.29	132.0	21.1	74	344
	6	5	150	168	15.2	23.26	21.42	131.4	44.4	39	169
			200	168	15.2	24.22	19.68	136.8	40.7	40	174
			250	169	15.4	22.09	16.99	133.1	40.0	40	181
	3-6	5	150	212	15.2	22.69	7.95	134.2	24.8	55	261
			200	213	15.4	20.07	7.65	134.1	24.1	57	269
			250	213	15.5	19.54	7.10	134.0	22.6	59	279
	3-6	6	250	256	15.7	19.19	7.30	133.9	20.4	65	271
	3-5	6	250	336	15.5	14.03	7.36	133.6	19.0	81	417
			350	335	15.3	12.00	7.01	133.4	16.8	85	446
Rec-tangle	3-6	6	300	256	15.6	20.63	7.79	135.1	23.0	62	320
	3-5	6	250	337	15.5	14.93	8.16	134.2	22.2	77	387
			350	336	15.4	14.45	7.74	136.0	20.9	78	398
Square	3-6	4	150	171	15.4	22.31	9.91	136.2	30.3	50	207
			300	172	15.5	21.95	8.75	132.9	26.7	53	224
	3-5	4	150	225	15.6	18.27	9.84	133.1	26.9	64	257
			300	224	15.5	15.46	8.85	134.7	23.9	67	274

*) The relative standard error in the first column is for the whole Gazi test area (gazi1 and gazi2), in the second column for the mangrove stratum (Gazi1).
**) Distance between clusters in Gazi1 (mangrove) was 3 km and in Gazi2 5 km.

Note; codes in the following figure legends are as follows:

- L-86: L-shape cluster (upside/right) with 5 sample plots, the distance between clusters 8 km in stratum1 and 6 km in stratum2.
- L-87-6p: L-shape cluster with 6 sample plots, the distance between clusters 8 km in stratum1 and 7 km in stratum2.
- R-86: rectangle-shape cluster with 6 sample plots, the distance between clusters 8 km in stratum1 and 7 km in stratum2.
- S-77: square-shape cluster, the distance between clusters 7 km in stratum1 and stratum2.
- Note: in rectangle-shape clusters the distance between sample plots in East-West direction was 200 m.

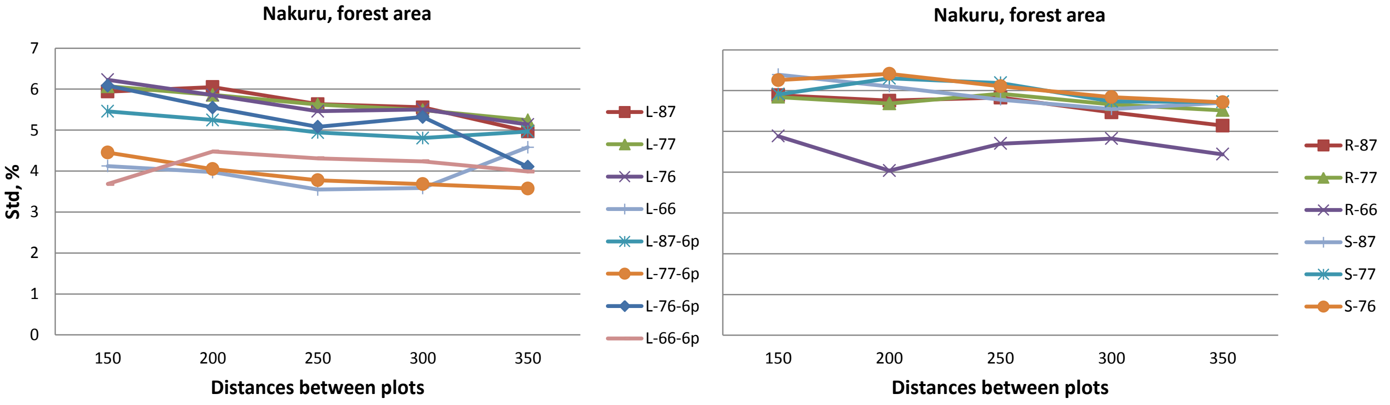

Figure 19. Relative errors of forest area with different cluster designs and sample plot and cluster distances in the Nakuru test area.

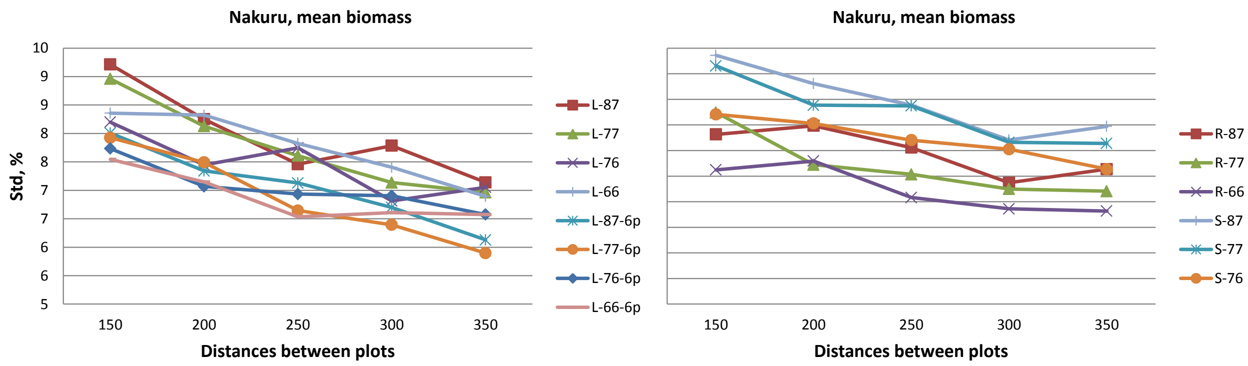

Figure 20. Relative standard errors of mean biomass with different cluster designs and sample plot and cluster distances in the Nakuru test area.

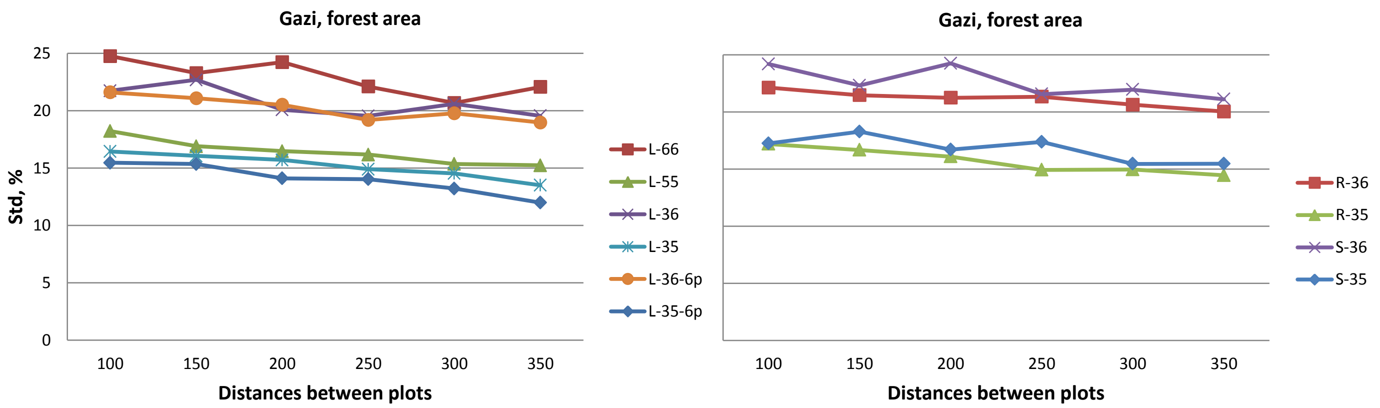

Figure 21. Relative errors of forest area with different cluster designs and sample plot and cluster distances in the Gazi test area.

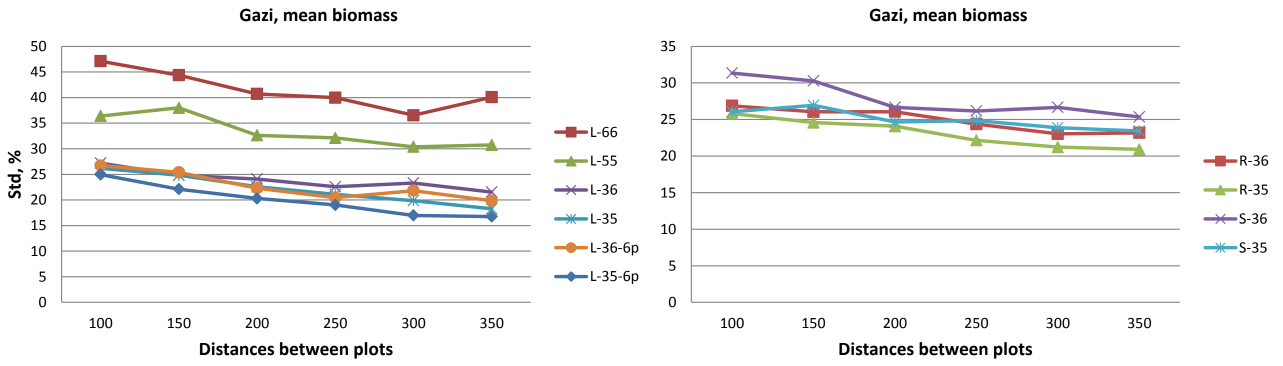

Figure 22. Relative standard errors of mean biomass with different cluster designs and sample plot and cluster distances in Gazi test area.

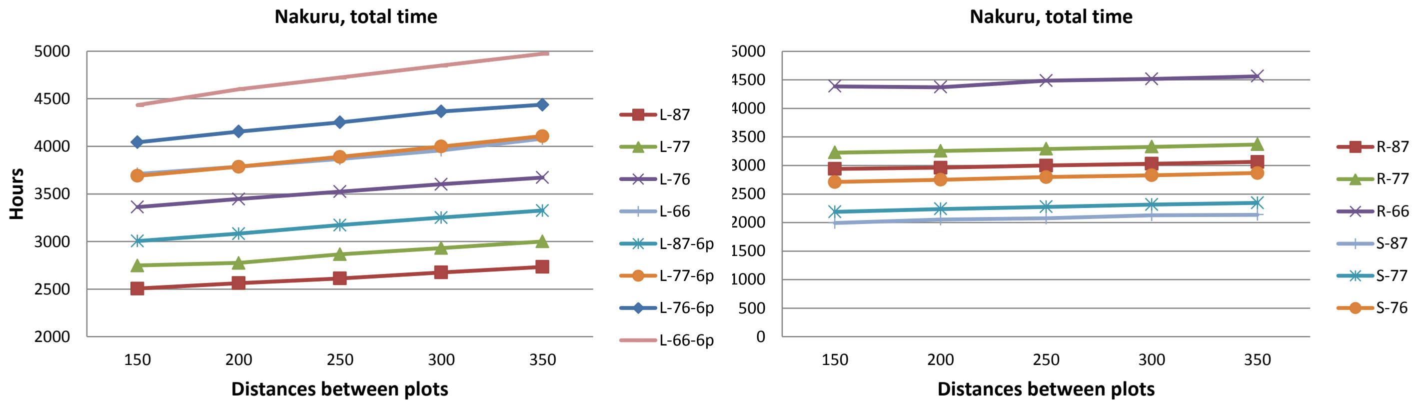

Figure 23. Time consumption with different cluster designs and sample plot and cluster distances in the Nakuru test area.

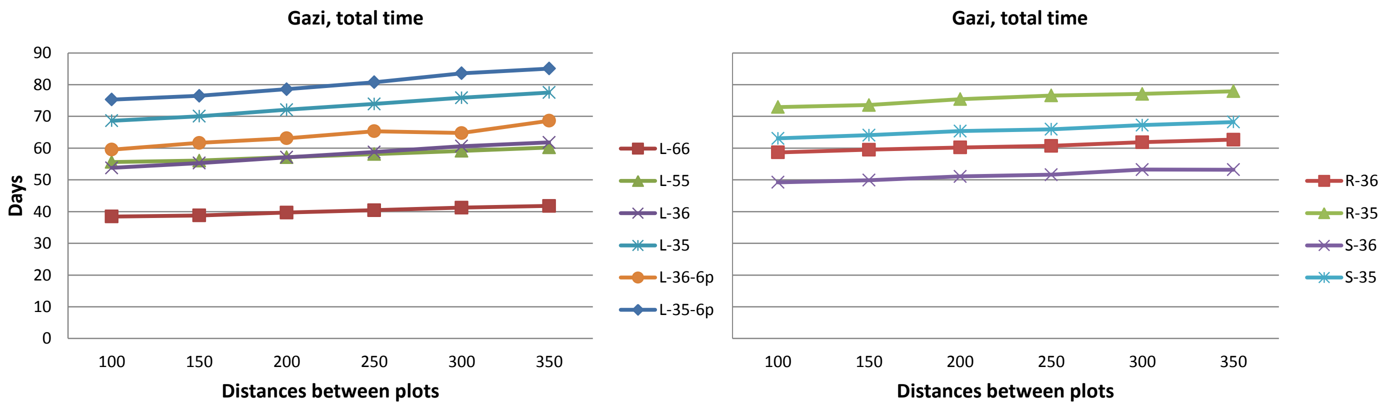

Figure 24. Time consumption with different cluster designs and sample plot and cluster distances in Gazi test area.

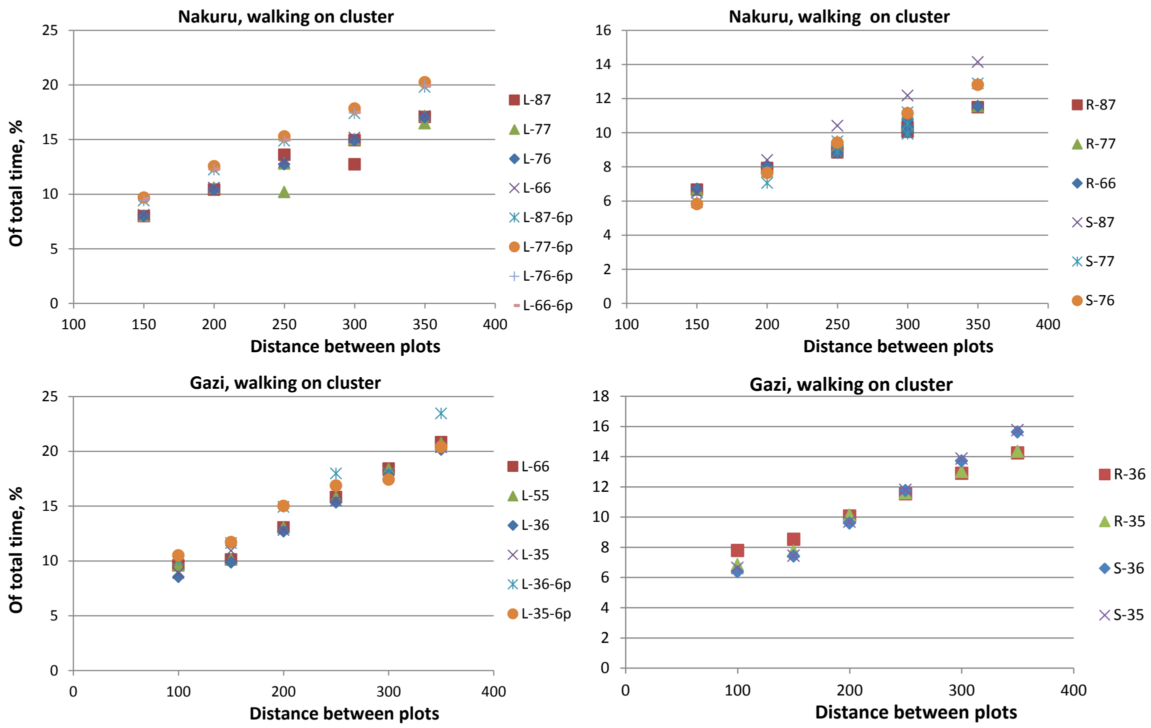

Figure 25. The relative share of the walking time of the total time in a cluster with different cluster designs and distances between sample plots.

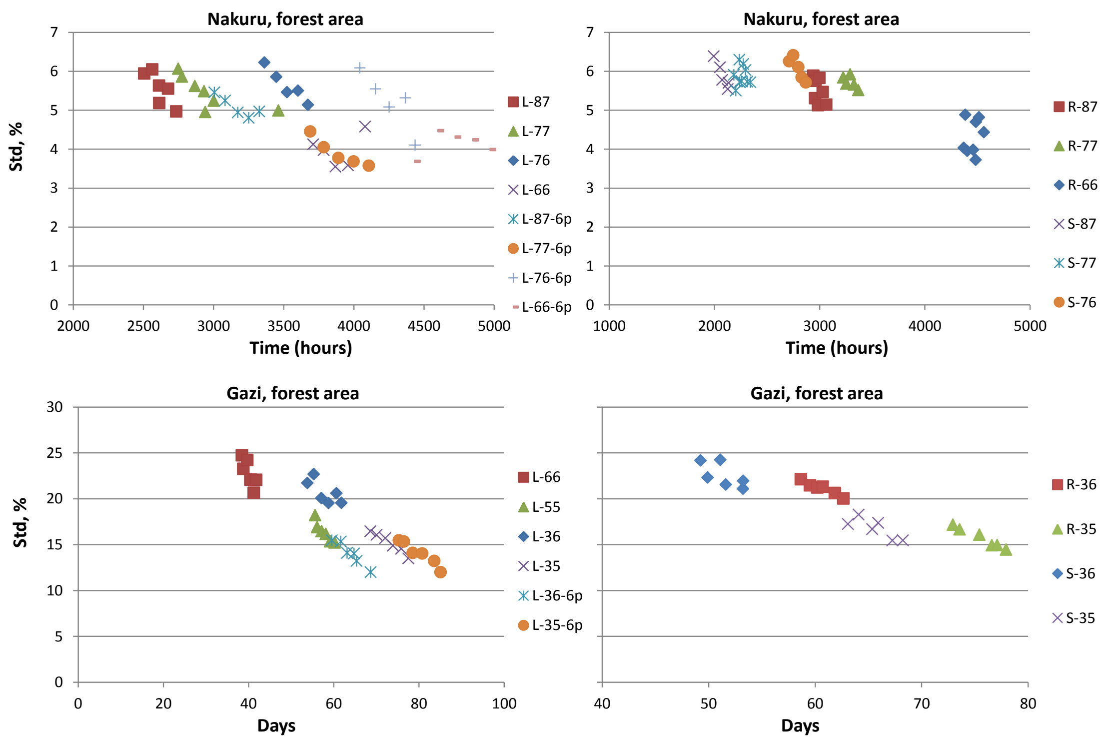

Figure 26. Relationship between the time consumption and the relative standard error of the forest area in test areas.

4.3 Kenya

The results from the Nakuru and Gazi test areas were utilised in the sampling design simulations for Kenya. Those cluster forms and distances between sample plots that gave the best results in the Nakuru and Gazi test areas were first tested for the design of Kenya. Unlike in the Nakuru and Gazi simulations, the distance between clusters was 2 km. As the method of simulations for Kenya was doubly stratified two-phase sampling, different distances between clusters were not tested separately. Kenya was stratified into four spatial strata, Figure 1 and Table 7. These strata were based on the county boundaries and the agro-ecological zones. In addition, in the coast, mangrove forests formed an own stratum as described earlier. Two different boundaries for stratum1 and stratum2 were tested (see Introduction).

Table 7. Area of different forest types in Kenya in square kilometres (km^2) with the original and modified strata (stratum1 and stratum2 modified).

	Stratum1	Stratum2	Stratum3	Stratum4	In total
Bamboo	0.0	856.9	0.0	0.0	856.9
Mangrove	0.0	0.0	0.0	661.1	661.1
Natural Forest	13,390.3	21,647.2	3,429.8	6.5	38,473.9
Plantation	0.0	1922.6	0.0	0.0	1922.6
Forest in total	*13,390.3*	*24,426.7*	*3,429.8*	*667.6*	*41,914.5*
Grassland	322,938.7	87,091.9	13,099.4	143.6	423,273.7
Cropland	1,151.8	89,463.8	10,096.7	51.9	100,764.2
Settlements	119.7	1,026.7	107.7	3.1	1,257.2
Other land	8,005.0	2,405.5	175.7	139.2	10,725.3
Land in total	*345,605.5*	*204,414.6*	*26,909.3*	*1,005.4*	*577,935*
Wetlands	9,096.7	5,649.7	84.8	5.7	14,837.0
In total	354,702	210,064	26,994	1,011	592,772
Re-modified strata					
	Stratum1	**Stratum2**	**Stratum3**	**Stratum4**	**In total**
Bamboo	0.0	856.9	0.0	0.0	856.9
Mangrove	0.0	0.0	0.0	661.1	661.1
Natural Forest	16,082.9	18,954.4	3,429.8	6.5	38,473.9
Plantation	101.4	1,821.3	0.0	0.0	1922.6
Forest in total	*16,184.3*	*21,632.6*	*3,429.8*	*667.6*	*41,914.5*
Grassland	321,259.6	88,771.1	13,099.4	143.6	423,273.7
Cropland	29,287.4	61,329.9	10,096.7	51.9	100,765.9
Settlements	179.3	967.0	107.7	3.1	1,257.2
Other land	8,006.6	2,403.6	175.7	139.2	10,725.1
Land in total	*374,917.3*	*175,104.2*	*26,909.3*	*1,005.4*	*577,936*
Wetlands	9,112.9	5,633.9	84.8	5.7	14,837.3
In total	384,030	180,738	26,994	1,011	592,774

In the simulations, the 1^{st}-phase sample (clusters) was stratified into 2^{nd}-phase strata according the number of forest sample plots in a cluster, Table 8. The used method selects clusters from the 1^{st}-phase grid (2 km by 2 km) by random, that is by varying distances. Allocation to and selecting clusters of different strata was based on optimal allocation having the standard error of the total forest biomass as a target variable and the time cost as a limiting factor:

$$\boldsymbol{n_h = T * \frac{se_h}{\sqrt[2]{c_h}} / \sum_{h=i}^{L}(se_h * \sqrt[2]{c_h})} \qquad \text{(Equation 1)}$$

Where n_h = number of 2^{nd}-phase clusters within stratum h
T = target time (time for the field work)
L = number of strata
se_h = standard error of the total biomass within 2^{nd}-phase stratum h
c_h = the average time cost (walking and measurement) per cluster within 2^{nd}-phase stratum h.

Table 8. Number of forest plots in a cluster in different classes (2nd-phase strata) by strata.

	Number of forest plots in a cluster		
	Class 1	Class 2	Class 3*
Stratum 1, “Grasslands”	0-2 0-1	3-4 2-3	5-n** 4-n
Stratum 2, “Forested areas”	0-1	2-3	4-n
Stratum 3, “Coast”	0-1	2-3	4-n
Stratum 4, “Mangrove”	0-1	2-3	4-n

*) “n” is the maximum number of sample plots in a cluster
**) With the designs having a square-shape cluster, values were 0-1, 2-3 and 4, respectively.

It seemed that the performance of the different cluster designs in the whole country was somewhat different than in the Nakuru and Gazi test sites. The L-shape cluster forms resulted in the biggest relative errors both in forest cover and biomass estimates.

A bit surprisingly, the square-shape cluster form gave about 0.3–0.6 %-unit lower relative errors in forest cover than the L-shape clusters. The rectangle-shape cluster was only a little bit better (0.1–0.2 %-unit) than the L-shape cluster in forest cover estimate. In the biomass estimates differences were even smaller. The cluster forms that combined square- and rectangle-shapes (RS-2 and RS-3) in different stratum gave similar results than the pure square-shape cluster. The relative error of forest cover with these cluster designs was 1.0–1.36%, and that of mean biomass 1.28–1.58% and total biomass 1.21–1.64% (Figure 27, Figure 28 and Table 9). With the modified strata (stratum1 and stratum2) designs RS-1re and RS-3re gave slightly better results than the other designs with the original strata. On the other hand, the differences between all designs were small as compared with the same time consumption. The biggest difference in forest cover was 0.7%-unit, in mean biomass 0.47%-unit and in total biomass 0.36%-unit. With the time cost of 55,000 hours, the differences were 0.6, 0.39 and 0.34, respectively.

The ranking of designs at the stratum level was basically same as at the country level. The square-shape and combination of rectangle-square-shape (RS-1, RS-2, RS-3, RS-1re and RS-3re) cluster designs gave the lowest relative errors in all strata. However, in stratum3 the rectangle-shape cluster (R) gave similar results than the best ones (Figure 30). The L- shape cluster designs gave clearly bigger relative errors in all strata than the other designs, especially in stratum1 and stratum4.

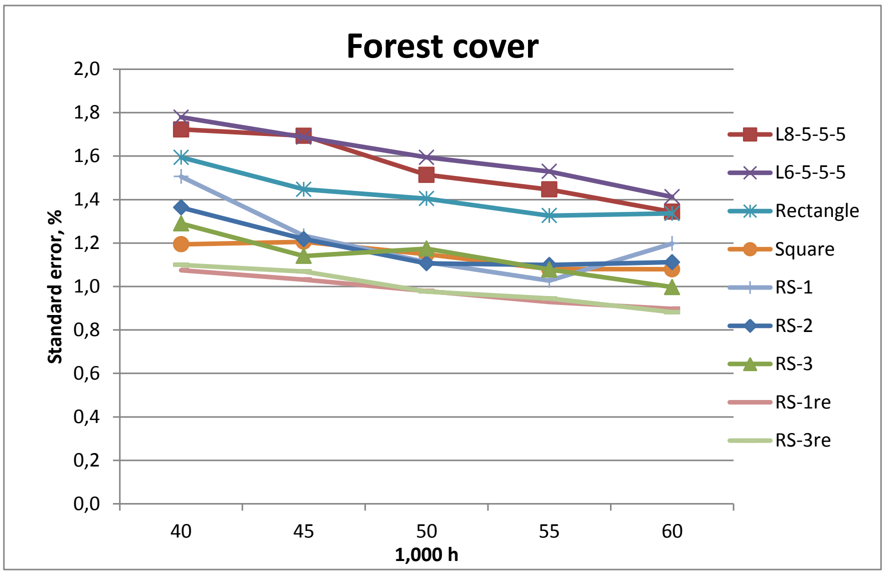

Figure 27. Relative standard errors of the forest cover with different design and time consumption.

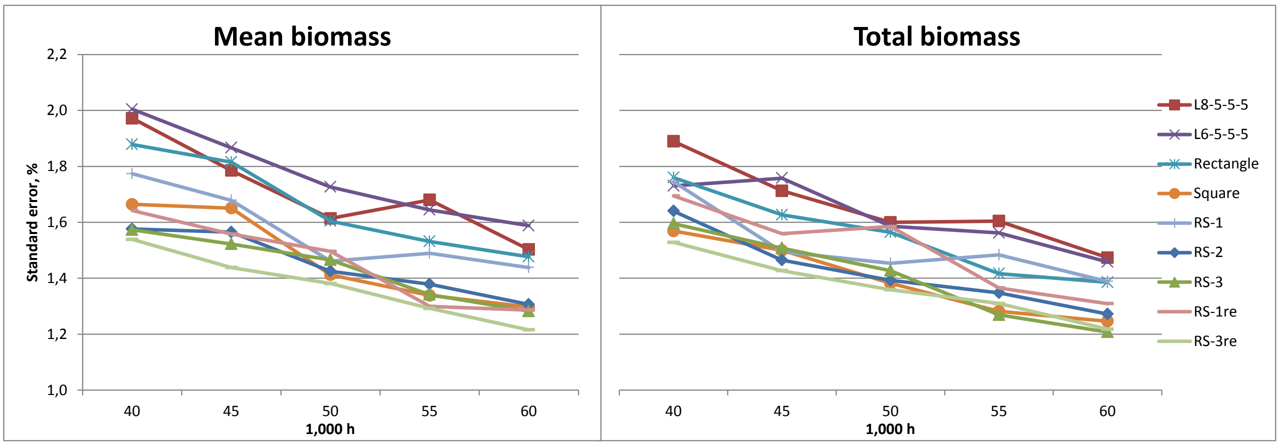

Figure 28. Relative standard errors of the mean and the total biomass with different design and time consumption.

The biggest differences between designs, when comparing relative errors of forest cover and total biomass, were in stratum1 (grasslands) and in stratum4 (mangrove). The combinations of rectangle-square-shape cluster designs performed well in both strata. The square-shape design seemed to be the best one for mangrove, both in forest cover and total biomass. In mangrove the biggest difference between designs in forest cover was 2.18 %-unit, in mean biomass 2.89 %-unit and in total biomass 2.8%-unit. With the time cost of 55,000 hours, the differences were 1.44, 2.03 and 2.51, respectively. In the stratum2 (forested stratum) the differences with tested designs were small. The biggest difference in forest cover was 0.13%-unit, in mean biomass 0.33%-unit and in total biomass 0.39%-unit. With the time cost of 55,000 hours, the differences were 0.09, 0.31 and 0.27, respectively.

There were also great differences between designs when time consumption was studied at the cluster level, Table 10. In L-shape and Rectangle designs there are more clusters that need more than 8 hours to finish the field work (measuring and walking time) than with other designs. The combination of rectangle- and square-shape clusters (RS-2 and RS-3) seemed to be the most efficient regarding the working time. According to the simulations there was a need to overnight in the field about on 25% of clusters with these designs. In Figure 29 there is one realisation of how clusters could be established in RS-3re design.

Table 9. Doubly stratified two-phase sampling simulation results on Kenya.
Notation "8-5-5-5" in Plots in cluster -column means 8 sample plots in each cluster in stratum1 and 5 plots in other strata. Notation in Plot distance -column is similar. Note; 2nd-phase sample optimized with the restriction of 55,000 hours.

Design	Name[1]	Plots in cluster	Plot distance	Clusters		Plots	Forest cover	Mean bio-mass	Total bio-mass
			(m)	1st phase	2nd phase	on land	Sd, %	Sd, %	Sd, %
L-shape	L8-5-5-5	8-5-5-5	200-250-200-150	148,374	5,830	30,347	1.45	1.68	1.60
	L8-6-5-5	8-6-6-5	200-250-250-200	148,476	4,923	30,047	1.36	1.64	1.55
	L6-5-5-5	6-5-5-5	250-250-250-200	148,409	5,984	30,062	1.53	1.64	1.56
Rec-tangle	Rec-tangle	6-6-6-6	200-250-250-150	148,014	5,340	31,445	1.33	1.53	1.42
Square	Square	4-4-4-4	200-250-250-150	147,822	8,182	32,075	1.08	1.34	1.28
Rec-Squ[2]	RS-1	6-6-6-4	250-250-250-150	148,075	5,368	31,509	1.41	1.61	1.55
Rec-Squ[3]	RS-2	4-4-6-4	250-250-250-150	147,969	7,978	31,697	1.10	1.38	1.35
Rec-Squ[4]	RS-3	6-4-6-4	250-250-250-150	148,083	7,781	32,206	1.08	1.34	1.27
Re-modified strata (stratum1 and stratum2)									
Rec-Squ[2]	RS-1re	6-6-6-4	250-250-250-150	148,113	5,300	30,987	0.93	1.30	1.37
Rec-Squ[4]	RS-3re	6-4-6-4	250-250-250-150	147,969	7,428	31,436	0.95	1.29	1.31

(1 Design name in following figures (Figure 27–Figure 31)
(2 Strata 1–3 rectangle- and stratum 4 square-shape clusters
(3 Stratum 3 rectangle- and strata 1, 2 and 4 square-shape clusters
(4 Strata 1 and 3 rectangle- and strata 2 and 4 square-shape clusters

Table 10. Percentages of how clusters are measured in different design by time classes. Note; 2nd-phase sample optimized with the restriction of 55,000 hours.

Design	Time classes						
	< 2h	2-4h	4-6h	6-8h	8-10h	10-12h	>12h
L8-5-5-5	0.2	22.3	49.2	63.8	80.3	88.2	99.8
L8-6-6-5	0.2	1.4	41.2	57.5	69.7	83.6	100
L6-5-5-5	0.3	20.6	49.9	63.7	79.4	88.1	100
Rectangle	0.2	1.2	39.3	53.6	65.7	81.8	100
Square	0.3	41.6	59.7	78.8	88.1	92.9	99.9
RS-1	0.2	1.2	40.1	53.6	65.6	82.5	100
RS-2	0.2	37.1	56.8	75.4	86.4	91.6	99.9
RS-3	0.2	36.7	56.5	73.9	84.4	90.9	99.9
RS-1re	0.2	1.7	41.8	57.9	74.2	85.8	100
RS-3re	0.4	38.6	59.0	77.8	86.5	92.0	100

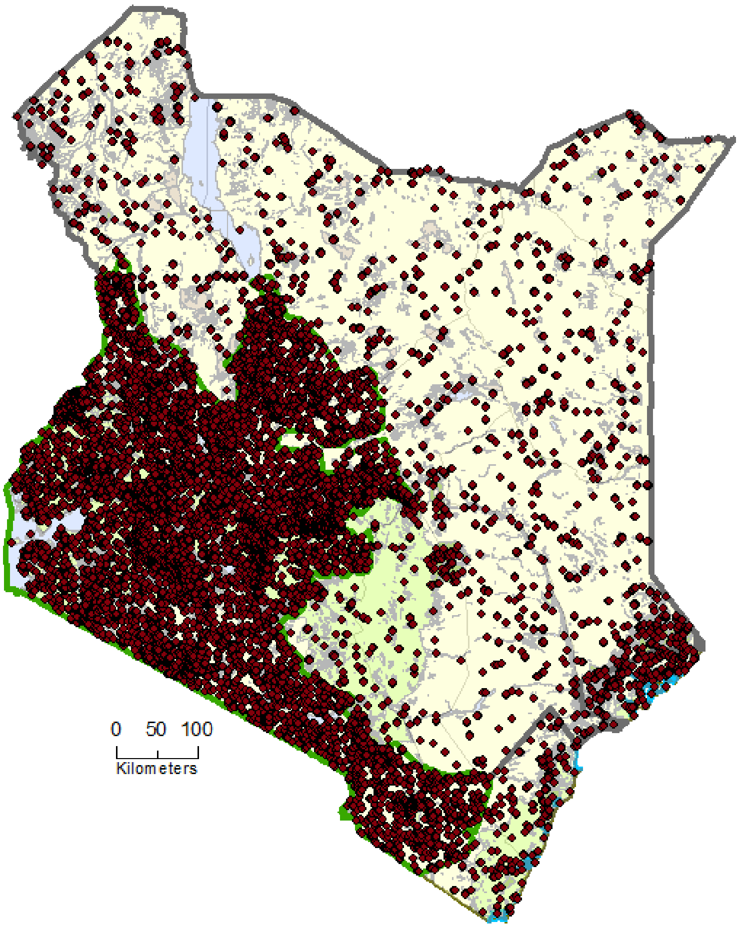

Figure 29. One realisation of cluster locations in Kenya with the RS-1re design.

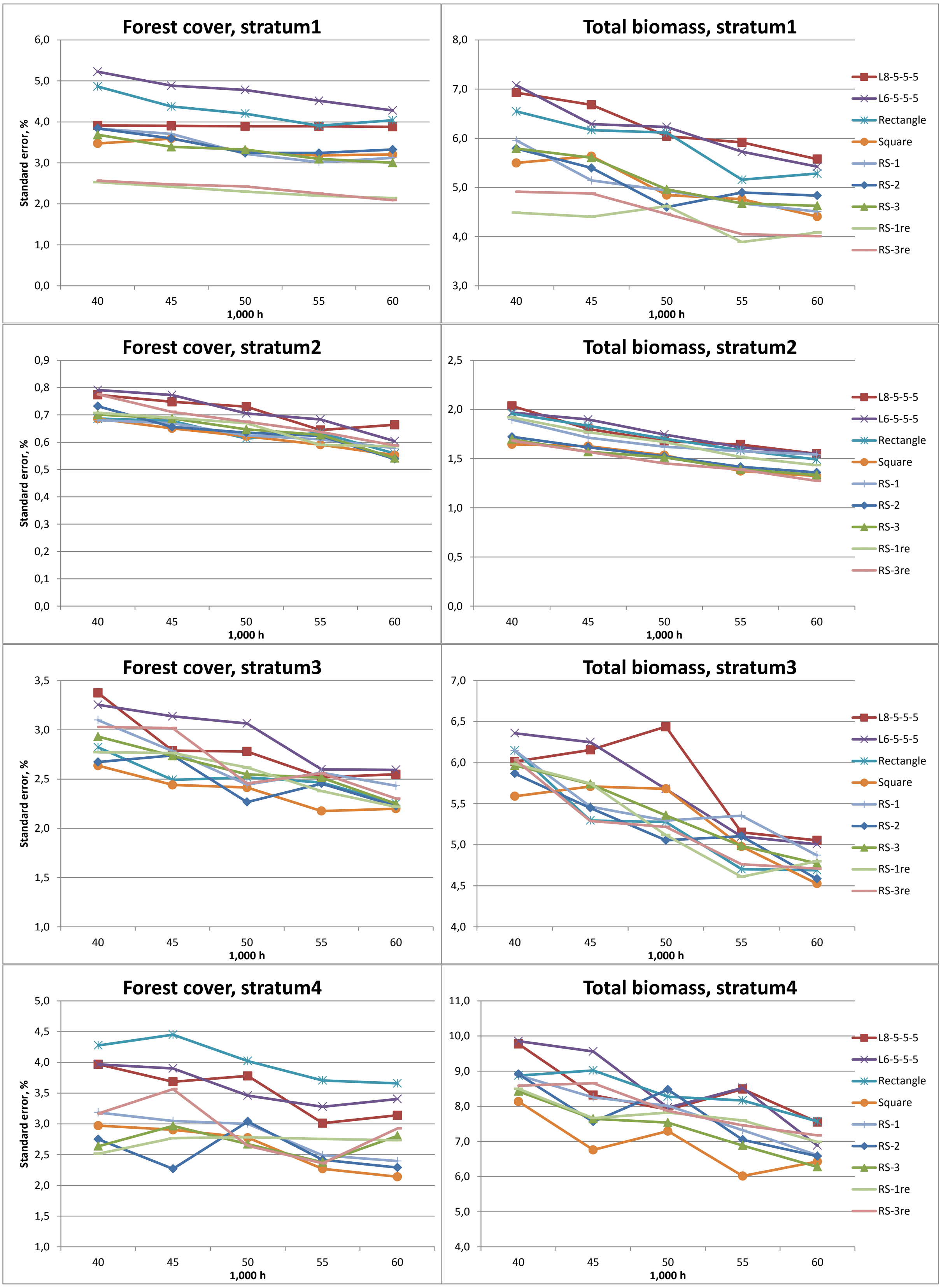

Figure 30. Relative standard errors of forest cover and total biomass by strata with different designs. Note; scale of y-axis varies.

It is obvious, that having more sample plots on land improves accuracies (Figure 31). For example, increasing the number of sample plots with about 50 % (Rectangle) decreases the relative error of forest cover about 0.3 %-unit.

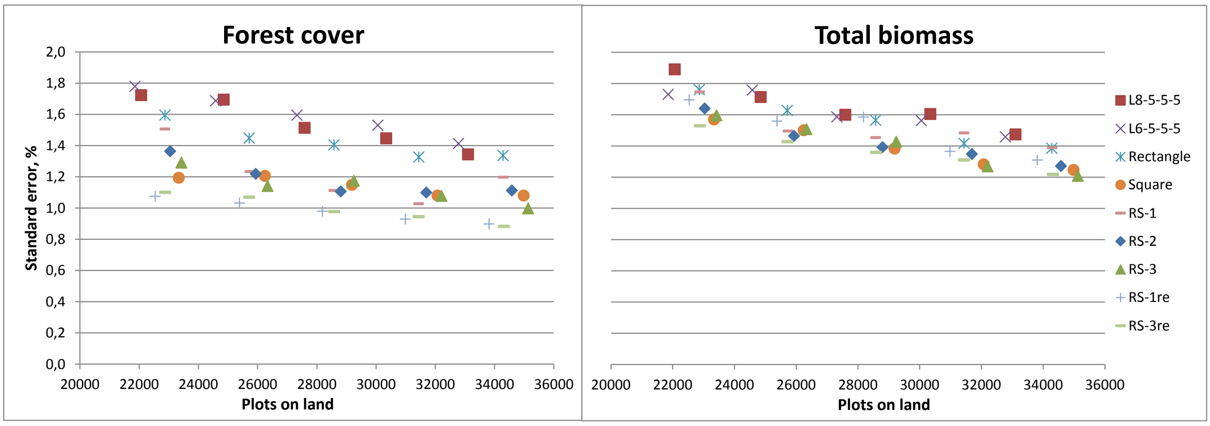

Figure 31. Relationship between the number of sample plots on land and relative errors of the forest cover and total biomass.

Stratified sampling was tested with few designs also for the whole Kenya; the same method than in the Nakuru and Gazi test areas. The strata were the same as in the doubly stratified two-phase sampling: stratum1 – stratum4. The tested designs were L-shape, rectangle-shape and combination of rectangle (stratum1–stratum3) and square (stratum4). The accuracies were not as good as with the comparable design of doubly stratified two-phase sampling, Table 11. E.g. the error of forest cover was, at the lowest, 1.64% (L-shape) and in the two-phase sampling 0.36–1.53%, with total biomass 2.48% (rectangle) and 1.42%, respectively.

Table 11. Results of the stratified sampling for the whole Kenya.

Design	Cluster distance	Plots in cluster	Plot distance	Clusters	Forest cover	Mean biomass	Total biomass	Total costs
			(m)	On land	Sd, %	Sd, %	Sd, %	Hours
L-shape	8-7-7-6	6-5-5-5	200-250-200-150	55,901	1.96	2.10	2.64	87,792
	7-6-6-5	6-5-6-5	200-250-200-200	75,063	1.64	1.92	2.52	118,046
Rectangle	7-6-6-5	6-6-6-6	200-250-200-200	80,779	1.73	1.83	2.48	118,853
Rectangle - square	7-6-6-S3	6-6-6-4	250-250-250-150	81,171	1.76	1.78	2.49	120,517

5 Conclusions

Stratified sampling

The tested designs with stratified sampling in two test areas, Nakuru and Gazi, revealed that the optimal distances between sample plots in a cluster are 200–250 meters. Longer distances will improve the accuracies but also increase the time cost. So, distances near 200 meters can be considered as a trade-off between cost and accuracy.

In mangrove forests the conditions are tough. First, the terrain is usually muddy and slow to walk. Secondly, work in a cluster must be fitted into low tide. So the actual working time in a cluster will be reduced to near 7 hours per day. To be efficient, it might be good to use square-shape clusters and shorter distances, e.g. 150 or 200 metres, between sample plots. With this design one cluster can be measured during a working day.

Doubly stratified two-phase sampling

The results with doubly stratified two-phase sampling in Kenya showed that the best design can be obtained by combining different cluster forms in the different strata. The square-shape cluster design should be used in stratum 4, mangrove area, as this gave the lowest standard error in forest cover. The square-shape design is also cost-efficient with respect to time spent for walking in a cluster.

In stratum3 the rectangle-shape cluster design seemed to be as good as the square-shape and combination of rectangle-square-shape cluster design. The L-shape designs did not work well. When walking times in L-shape and rectangle-shape clusters are compared, the rectangle-shape might be more efficient. This is because going to the rectangle-shape (or square-shape) cluster and back takes less time than with the L-shape cluster.

In Kenya, the amount of 2nd-phase sample plots on land (to be measured in the field) varied from about 30,000 to 32,000 with the doubly stratified two-phase sampling -method (Table 9). With the stratified sampling, the amount of sample plots on land varied between 55,900 and 81,170 (Table 11). As the accuracies between these two methods were clearly different, it is strongly advised to use the doubly stratified two-phase sampling method to avoid extra field work.

In the simulation study, the 2nd-phase clusters were selected randomly from the 1st-phase clusters. However, the selection probabilities varied between clusters so that clusters having more sample plots in forest were more likely to get selected. This method may lead to the result that clusters are not evenly distributed over the area of stratum. In practise, it might be somewhat better system to select 2nd-phase clusters more evenly, e.g., by selecting every nth cluster from the 1st-phase clusters that fulfil the criteria. However, in this simulation study, the used method can be considered to give appropriate results.

Two issues that must be reconsidered before final sampling design are official boundaries of Kenya and mangrove forest. It was discovered that both of these data had some shortcomings. First, official boundary in North-East at the border of South Sudan was somewhat undefined. There were three different delineations of the boundary. The Department of Resource Surveys and Remote Sensing (DRSRS) have upgraded new boundaries for Kenya. However, these were not yet official ones. According to the new boundaries, there are also some new areas at the coastline included to be part of Kenya. Secondly, the mangrove forests have some inaccuracies how they are delineated. As studied visually on Google Earth, there seems to be some areas that should be included in mangroves and some areas to be excluded. However, these issues with boundaries and mangroves are not so remarkable. With updated datasets there will be few more clusters most probably in North-East, in coastline and in mangrove areas compared to this simulation study with used datasets.