Multidimensional Sampling of Farms within R: A Successful Kazakh-German Cooperation

Sven SCHMIEDEL, PhD, (sven.schmiedel@destatis.de)
Federal Statistical Office of Germany
Meyram SEYDAZIM, (me.seydazim@stat.kz)
Committee on Statistics of Kazakhstan
Zhandos KOZBANOV, (zh.kozbanov@stat.kz)
Committee on Statistics of Kazakhstan

ABSTRACT

Within the project "strengthening the national statistical system of Kazakhstan", the Federal Statistical Office of Germany (DESTATIS) and the Committee on Statistics in Kazakhstan (CSK) intensively collaborated in the field of sampling methodology and the related technical implementation. One part of the project was to implement the multidimensional sampling methodology for agricultural surveys applied by the CSK in the statistical software "R" and thereby automatize the work process at CSK.

The frame of the sample is the agricultural register. Different types of crops are sampled together in a multidimensional approach. Key variable for the sampling design is the seed area for each crop, which is used to calculate the inclusion probability. Hence, the sample is proportional to size. The inclusion probability is used to select farms randomly. The method includes additionally an exponent, which reduces the probability of very large farms to be included.

The method was implemented successfully in "R" and thereby reduced massively the work load of the personnel of the division of sampling surveys in Kazakhstan.

Keywords: *Sampling, Agriculture, International Cooperation, Asia, Kazakhstan, Europe, Germany, Statistical Software R*

JEL Classification: *C83, Q12*

INTRODUCTION

The law "On ratification of Agreement on Loan (KAZSTAT: Project for Strengthening the National Statistical System of the Republic of Kazakhstan) between The Republic of Kazakhstan and International Bank for Reconstruction and Development" was signed by the head of state on December 27, 2011.

The project is implemented in the partnership with a consortium of statistical offices of Germany, Finland, Czech Republic, Slovakia, South Korea and Russia. The head of the consortium is the Federal Statistical Office of Germany.

The main goal of the KAZSTAT project is to provide users with qualitative statistical information and to promote the efficiency at the Committee on Statistics (CS) of Kazakhstan in accordance with international methodology and best practice.

Goal is to bring the national statistical system to the level of European countries, to increase users' satisfaction with qualitative statistical data to 80 % and to reduce the time spent by enterprises for completion and submission of statistical accounting forms to 40 %.

One component of this project named "Improving the quality and methods of conducting sample surveys" included theoretical and practical missions in Kazakhstan and Germany.

During these missions the methods of sampling in agricultural statistics were reviewed. Even though the method of sampling was well and clearly documented, the workers in Kazakhstan used an old version of the statistical software Statistica to set up their sampling frame and the final selection of units was realized in the spreadsheet program Excel of Microsoft Office.

This work was not very handy, made the sampling of units cumbersome and time consuming. Hence, the workers at CS needed several weeks to finalize the sample.

The German experts on statistic recommended the use of the statistical software R in order to streamline the process and produce the sample in a single step in one software package.

In this paper the method of sampling is explained. The R code is found in the Appendix.

MATERIAL AND METHODS

Material

Basis for agricultural surveys is the Kazakh Statistical Agricultural Registry (SAR). It was created in the first program of the reform of official statistics in Kazakhstan in 1993-1996 under methodological and technical assistance of specialists of the National Agricultural Statistics Service (NASS) of the United Stated of America. The registry is the tool for statistical reporting of economic entities, allows statistical surveillance in the agricultural sector and provides information about agricultural producers.

The SAR resolves the following main tasks:

- Define the general population;
- Set up a sampling frame;
- Evaluate the results of surveys;
- Receive information about the availability of agricultural commodities, available farmland, livestock and poultry by region.

Every agricultural unit in Kazakhstan (about 180 000) is identified by a unique registration code, the Business Identification Number. Agricultural units are

- agricultural enterprises;
- country (farmer) farms;
- households;
- gardening and country cooperatives.

An agricultural census is done every decade, twice a year the registry is updated from administrative sources and every month from statistical forms. Furthermore, the SAR is updated annually, by three samples, in spring for livestock, in summer for crops and in winter for livestock and crops. Additionally data of the tax authorities are used to keep the registry up-to-date.

The important regional unit in Kazakhstan is the "oblast", comparable to the NUTS-1 level, and inside the oblasts the raions, comparable to NUTS-2 level.

Main variables of the registry are the National Classifier of Administrative-Territorial Objects (KATO) which has nine characters, e.g. "111032100" first two digits "11" stand for the oblast level. The first four digits "1110" stand for the raion level. In Kazakhstan there are 16 oblasts of which two are cities (Astana and Almaty) and 210 raions. Multidimensional sampling is formed at raion level.

The registry includes about 300 variables of which about 55 are frequently used like the types of livestock and crops like cows, horses, sheep, goats, camels, pigs, poultry, chicken, rabbits, arable land, hayfields, pastures, crop, wheat, rye, rice, corn, beans, tobacco and cotton.

Methods

Sample Size

When planning a sample of the statistical units in agriculture, it is important to calculate the sample size. A sample that is too large leads to a high consumption of resources. A sample that is too small reduces the quality of the survey results. An optimal sample size estimates the general population parameter with the specified accuracy. The sample size is computed for each raion. The following formula based on simple random sampling with replacement is used for sample size estimation

$$n = \frac{Ns^2 z_{1-\alpha}^2}{s^2 z_{1-\alpha}^2 + e^2 N},$$

where, $z_{1-\alpha}$ is the quantile of the normal distribution at α, the "confidence level"

N– size of general population;
s^2– variance;
e– expected error, expressed as half length of the confidence interval.

If the error is expressed as a coefficient of variation (*CV*), the following formula can be used:

$$n = \frac{N\ CV^2 z_{1-\alpha}^2}{CV^2 z_{1-\alpha}^2 + e^2 N}.$$

If adjustment for finite population is ignored (simple random sampling without replacement), the formula will be as follows

$$n = \frac{z_{1-\alpha}^2\ CV^2}{e^2}.$$

In the agricultural statistics, the following sample sizes have been estimated and adopted after the statistical units have been surveyed over several years:

- For peasant and farming enterprises – 30 %;
- For subsistence farms – 5 %;
- For horticultural, market-gardening and dacha associations/ cooperatives – 5 %.

Sample selection

The method explained in the following was introduced by the NASS to the CS. Goal of this project was to transfer this method into an R code (see Appendix).

The method of selecting units is of great importance for a high-quality sample. Two methods are used to generate a sample for agricultural surveys:

1. One-dimensional sampling with probabilities proportional to size;
2. Multidimensional sampling with probabilities proportional to size. It is sampled separately in every raion.

One-dimensional sampling

One-dimensional sampling which is based on one indicator as selection criteria is also called continuous stratification. Here an approach following the method of systematic proportional to size sampling (Cochran 1977, Särndal 1991) is used. In contrast to the stratification method, which implies that the probability of inclusion is determined for each stratum, in one-dimensional sampling the probability is estimated for each element of the general population in accordance with the selection criteria (e.g. size of a sown area of peasant or farming enterprises).

One-dimensional sampling is used for surveying the yield of crops. This method is based on selecting certain segments (e.g. villages and different categories of farms) as primary sampling units (PSUs) proportional to their size, as well as on selecting the fixed number of secondary sampling units (SSUs), such as fields within each selected PSU.

The probability of inclusion in the sample for peasant or farming enterprises, when the criterion is one indicator (e.g., sown area from the SAR), is estimated by the following formula:

$$p_i = max\left(1, n\frac{X_i^f}{\sum_{i=1}^{N} X_i^f}\right), \text{where}$$

- p_i is the probability of inclusion in a sample for farm i;
- n is the specified size of sample;
- X_i^f is the value of the selection criterion (sown area) for farm i with expansion factor f, where $f \in [0.5; 1.0]$ (usually $f = 0.75$);
- $\sum_{i=1}^{N} X_i$ is the sum of the selection criterion (sown area) for the general population.

In order to raise the probability of inclusion for small farms and to reduce the probability of inclusion for large farms to be sampled the selection criterion is adjusted by the expansion factor *f*. Among the recommended values of the exponent (from 0.50 to 1.00), value 0.75 is usually used.

The sample is generated so that each unit can only be selected once. Hence, the probability of inclusion is reduced to 1 if $p_i > 1$. If latter happens the number units chosen reduces automatically and has to be adopted.

The sampling process is a systematic selection of agricultural producers (Table 1):

1 Accumulate the amount of the selection criterion

$$X_1/p_1, X_1/p_1 + X_2/p_2, X_1/p_1 + X_2/p_2 + X_3/p_3, \ldots$$

2. Calculate the selection interval $I = \sum_{i=1}^{N}(X_i/p_i)/n$

3 Select a random value $r_0 \in [0,1]$;

4. Calculating a random start $r = r_0 I$

5 Calculating a series of $r, r + I, r + 2I, r + 3I, \ldots$

One-dimensional sampling

Table 1

Farm i	Sown area in hectare x_i	Trans-formed area $x_i^{0.75}$	Proba-bility of inclusion p_i	Converted data of hectare x_i / p_i	Accumulative amount $\sum_{j=1}^{i} x_j / p_j$	Sample number	Selected expanded hectare
1	105	32.80	1.000	105	105		
2	147	42.22	1.000	147	252		
3	69	23.94	1.000	69	321		
4	95	30.43	1.000	95	416		
5	142	41.14	1.000	142	558	1	529
6	400	89.44	1.356	295	853		
7	160	44.99	1.000	160	1 013		
8	45	17.37	1.000	45	1 058		
9	84	27.75	1.000	84	1 142	2	1 074
10	120	36.26	1.000	120	1 262		
11	190	51.18	1.000	190	1 452		
12	170	47.08	1.000	170	1 622	3	1 619
13	400	89.44	1.356	295	1 917		
14	1 887	286.30	4.339	435	2 352	4	2 165
15	380	86.07	1.304	291	2 643		
16	123	36.93	1.000	123	2 766	5	2 710
…	…	…	…	…	…	…	…
141	100	31.62	1.000	100	24 542	45	24 525

Multidimensional sampling

In the case of a survey with multiple indicators, one criterion of sampling is not enough to obtain a representative sample. Here multidimensional sampling with probabilities proportional to size is used.

Probability of inclusion of farm i in multidimensional sampling is determined the following formula

$$p_i = min\left\{1, max\left\{n_1 \frac{X_{1,i}^f}{\sum_{i=1}^{N} xX_{1,i}^f}, \ldots, n_k \frac{X_{k,i}^f}{\sum_{i=1}^{N} X_{k,i}^f}\right\}\right\}, \text{where}$$

- k is the number of indicators taken from the register as selection criteria;
- $X_{1,i}^f - X_{k,i}^f$ are the values of selection criterion k for farm i;

- n_1 - n_k are the sample sizes assumed for each selection criterion;
- $f \in [0.5; 1.0]$ (usually $f = 0.75$).

Inclusion probabilities higher than 1 are set to 1. The highest inclusion probability is selected from the estimations of each criterion. The upper limit for the inclusion probability is set to 1 with the aim to prevent the inclusion probability of large farms into the sample during the systematic selection repeatedly (Table 2). This approach of limiting the inclusion probability to 1 may lead to reduction of the predefined sample size, which has then to be adopted.

For the selection, the inclusion probability is cumulated. Thereafter a single random start value between 0 and 1 is chosen and added to the cumulated probability. The farms where the cumulated probability exceeds the first time a natural number are sampled.

Multidimensional sampling

Table 2

Farm i	Crops			Probability of inclusion with $n_1 = 8$, $n_2 = 5$, $n_3 = 4$					Cumulative	Inclusion in sample
	$x^f_{1,i}$	$x^f_{2,i}$	$x^f_{3,i}$	p_1	p_2	p_3	$max\{p_i\}$	$min\{1, max\{p_i\}\}$	$\sum_i min\{1, max\{p_i\}\}$	
1	118.8	3.6	0.0	0.77	0.09	0.00	0.77	0.77	0.77	0
2	348.2	8.3	0.0	2.27	0.20	0.00	2.27	1.00	1.77	1
3	3.9	3.9	3.9	0.03	0.09	0.18	0.18	0.18	1.95	0
4	10.0	10.0	6.0	0.07	0.24	0.27	0.27	0.27	2.23	1
5	7.0	7.0	7.0	0.05	0.17	0.32	0.32	0.32	2.54	0
6	5.5	5.5	5.0	0.04	0.13	0.23	0.23	0.23	2.77	0
7	71.0	3.4	3.4	0.46	0.08	0.15	0.46	0.46	3.23	1
8	10.0	10.0	2.0	0.07	0.24	0.09	0.24	0.24	3.47	0
9	121.0	12.0	0.0	0.79	0.29	0.00	0.79	0.79	4.26	1
10	4.9	4.9	4.9	0.03	0.12	0.22	0.22	0.22	4.49	0
11	16.3	5.7	2.7	0.11	0.14	0.12	0.14	0.14	4.62	0
12	20.0	20.0	8.0	0.13	0.48	0.36	0.48	0.48	5.10	1
13	200.0	50.0	0.0	1.30	1.19	0.00	1.30	1.00	6.10	1
14	135.8	7.9	3.5	0.89	0.19	0.16	0.89	0.89	6.98	0
15	11.8	4.7	4.7	0.08	0.11	0.21	0.21	0.21	7.20	1
16	20.0	20.0	14.0	0.13	0.48	0.64	0.64	0.64	7.84	0
17	2.3	2.3	2.3	0.01	0.05	0.10	0.10	0.10	7.94	0
18	95.8	5.6	4.0	0.62	0.13	0.18	0.62	0.62	8.57	1
19	16.5	16.5	16.5	0.11	0.39	0.75	0.75	0.75	9.32	1
20	8.0	8.0	0.0	0.05	0.19	0.00	0.19	0.19	9.51	0
Total	1 226.8	209.3	87.9							9

Sample judgment

Following values can be derived for each crop from the sample and are used for assessment. The index *k* is neglected, because for each crop the following values are calculated.

1) *N, n* the population size and the predefined sample sizes.
2) *n/N* - predefined sample size as a percentage of population.
3) Obtained – the sample size of the selection
4) Extrapolated sum $\hat{Y}$ – the sum of the register values of the sample, weighted by their inclusion probability

$$\hat{Y} = \sum_{i=1}^{n} \frac{1}{p_i} X_i.$$

5) "ratio adjustment"–ratio of the extrapolated sum $\hat{Y}$ to the sum of the crop Y in the SAR "ratio adjustment"–must been between 0.7 and 1.3
6) Coefficient of variation in percent

$$\widehat{CV} = \frac{s}{\bar{x}} * 100$$

The coefficient of variation estimates deviation of the results of the sample from the true value in the population in percent.

7) $\sum_{i=1}^{n} X_i , \sum_{j=1}^{N} X_j$ - sum of (unweighted) register values for the general population and the sample.
8) $\sum_{i=1}^{n} X_i / \sum_{j=1}^{N} X_j$ - the sum of the register values of the sample as a percentage of the sum of the registry population.
9) Mean - the mean (unweighted) of the value for the registry population and the sample is calculated using the following formula:

$$\bar{X} = \frac{1}{N} \sum_{j=1}^{N} X_j , \bar{x} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

10) Minimum - the minimum value of the registry population and the sample.
11) Maximum - the maximum value of the registry population and the sample.

Especially the "ratio adjustment" was very important for the assessment of the sample. The workers at CS rejected a sample if the "ratio adjustment" was greater than 1.3 or smaller than 0.7.

Summary

The described method was computed in R (see Appendix). This led to a remarkable reduction of work load for the workers of the CS. Especially the repeated search for a sample that meets the criteria like "ratio adjustment" is now automatically done in R. Additionally, the fact that for the one-dimensional and the multidimensional sampling it is not clear beforehand how many units are sampled, which comes from the approach of reducing the probability of large units to be sampled, was time consuming. Both problems are now transferred in nested loops in R.

R is now one of the standard software at CS.

References

1. **Cochran, WG**, 1977, *Sampling Techniques*, John Wiley & Sons, New York.
2. **Särndal CE, Swensson B, Wretman J**, 1992, *Model Assisted Survey Sampling*, ***Springer Series in Statistics***, Springer-Verlag, New York.

APPENDIX – THE R CODE FOR MULTIDIMENSIONAL SAMPLING

```
#######################################
# Multidimensional selection of farms  #
# Study visit at DESTATIS, Wiesbaden   #
# 18-22 May 2015                       #
# Author: Sven Schmiedel               #
#######################################


#The program selects farms by weighted number of livestock in a systematic
#manner. It is called multidimensional as it is possible to include
#a multiple number of different cattle
#Large farms are reduced by the factor exponent

#library foreign is needed when data has a dbf format
library(foreign)

#here the folder and the data name to be read have to be stated
readfile <- "..."

#the directory where the sample results should be written to is stated
writedir <- ".....//"

#the first and last column of the cattle data in the dataframe are declared
first_col    <-    9
last_col     <-   44

#states the minimum number of farms with a certain cattle
min.farm     <-    3

#if this or a smaller number of farms is found in a raion all farms are selected
min.raion    <-   10

#the variable "exponent" is between 0.5 and 1.0
#the standard that was used is 0.75. It weighs down the probability
#for farms with a large number of one or more cattle.
exponent     <- 0.75
```

```
#interval for the accepted ratio adjustment is declared
ratio_adj_min <- 0.7
ratio_adj_max <- 1.3

#the fraction of farms which should be included is declared
farm_fraction <- 0.3

#the data of one oblast is read
oblast <- read.csv2(file=readfile)

#sorting the dataframe by raion
oblast <- oblast[order(oblast$RAION),]

#for checking the program, a seed for the random number generator can be set
#set.seed(1)

#The directory, where the results are saved, is created
dir.create(path=writedir)

#before the loop starts, variables for data collection are declared
sample_oblast <- NULL
N_oblast <- rep(x=0, times=last_col-first_col+1)
n_oblast <- rep(x=0, times=last_col-first_col+1)

#start of the loop. The loop variable is the RAION
for (raion_sample in unique(x=oblast$RAION)){

  #the data of the raion of the current loop is extracted
  raion <- oblast[oblast$RAION==raion_sample,]

  #the last two columns are not needed
  raion <- raion[,1:(ncol(x=raion)-2)]

  #starting values for the loop
  ratio.adj <- 2
  j <- 0

  #population sum for each cattle
  pop.sum <- apply(X=raion[,(first_col:last_col)],MARGIN=2,FUN=sum)


 #number of farms which keep a certain cattle
 N <- apply(X=raion[,(first_col:last_col)]>0,MARGIN=2,FUN=sum)

 #the loop ends when the ratio adjustment (the ratio between the
 #extrapolated and the population value) lies in the interval
 #[ratio_adj_min;ratio_adj_max]
 while(!(max(ratio.adj,na.rm=TRUE)<ratio_adj_max &
           min(ratio.adj,na.rm=TRUE)>ratio_adj_min))
 {
   #counter for the loop
   j <- j + 1

   #if after 100 loops no result is found the counter is reset to 1 and
   #to the minimum of farms is added 1
   if (j==101)
   {
     min.farm <- min.farm + 1
     j <- 1
   }

   #When the loop is here the first time with a new "min.farm" p is set
   #to zero because a new p has to be found for the new "min.farm".
   if (j==1) p <- 0 else p <- p.memo - 0.01

   #the data with the randomly chosen farms is set up here
   prop.inc <- data.frame(inc=0)

   #the loop repeats as long as less than the fraction given by the variable
   #farm_fraction of the farms in the raion is chosen
```

```
 while(round(x=nrow(raion)*farm_fraction)-1>=sum(prop.inc$inc))
 {
   #p is the the fraction farms with a certain cattle
   p <- p+0.01

   if (nrow(x=raion)<=min.raion) n <- N
   else {
     #number of farms which should be included in the sample with
     #certain cattle
     n <- round(x=N*p, digits = 0)

      #where n is lower than the accepted minimum, the minimum is set
      n[n < min.farm] <- pmin(min.farm,N[n < min.farm])
    }
    #the exponent is applied
    dataexp <- raion[,(first_col:last_col)]^exponent

    #the relative frequency of each cattle is calculated
    prop.dataexp <- prop.table(x=dataexp, margin=2)

    #n*relative frequency of each cattle is calculated
    prop.inc <- data.frame(t(x=t(x=prop.dataexp)*n))

    #to the names of the data.frame the "_freq" is added
    names(x=prop.inc) <- paste0(names(x=prop.inc),"_freq")

    #the maximum of the probability of selection for each farm
    prop.inc <- cbind(prop.inc,Pmax=apply(X=prop.inc,MARGIN=1,FUN=max,
                                          na.rm= TRUE))

    #if the probability exceeds 1 this is reduced to 1
    prop.inc <- cbind(prop.inc,Plast=pmin(prop.inc$Pmax,1))

    #the cumulative sum of probability of selection is calculated
    prop.inc <- cbind(prop.inc,cum.amount=ave(x=prop.inc$Plast,FUN=cumsum))

    #random factor for inclusion is set and added to the cumulative sum
    random_number <- runif(n=1,min=0,max=1)
    prop.inc$cum.amount <- prop.inc$cum.amount + random_number

    #the variable which states the inclusion to the sample is set to 0
    prop.inc$inc <- 0

    #the included farms are found
    for (i in 1:max(prop.inc$cum.amount))
      if (sum(prop.inc$cum.amount<=i)==i) prop.inc$inc[i] <-1 else
        prop.inc$inc[sum(prop.inc$cum.amount<=i)+1] <-1
 }

  #the used p is put in a memo
  p.memo <- p

#the result of the sampling is bind to the raion dataset
raion <- cbind(raion[,1:last_col],prop.inc)

#calculation of the extrapolation factor
raion$weight <- 1/raion$Plast

#the sample is put in a own dataset
sample<-raion[prop.inc$inc==1,]

#a dataframe of all cattle multiplied by the extrapolation factor
#of the respective farm is created
expanded <- sample[,(first_col:last_col)]*sample$weight

#the extrapolated sum of each cattle is calculated
exp.sum <- apply(X=expanded,MARGIN=2,FUN=sum)

#ratio adjustment is calculated
ratio.adj <- pop.sum/exp.sum
```

```
    #the results of the current loop are printed on screen
    print(x=paste("j = ",j," minimum ratio adj = ",
                  round(min(ratio.adj,na.rm=TRUE),2)," maximum ratio adj = ",
                  round(max(ratio.adj,na.rm=TRUE),2)," p = ",p," min.farm = ",
                  min.farm,cat(names(sort(ratio.adj,decreasing =TRUE))[1:3])))
  }

  #the sample of the oblast is memorized
  sample_oblast <- rbind(sample_oblast,raion)
  #N and n of the raion is added to the sum of the oblast
  N_oblast <- N_oblast+N
  n_oblast <- n_oblast+n

  #the number of farms with a certain cattle is summarized
  obtained <- apply(X=sample[,(first_col:last_col)]>0,MARGIN=2,FUN=sum)

  #the sampling error is calculated
  smpl.err <- sqrt(apply(X=sample[,(first_col:last_col)],MARGIN=2,var)/n)*(1-n/N)


  #the coefficient of variation is calculated
  CV <- smpl.err/apply(X=sample[,(first_col:last_col)],MARGIN=2,mean)


  sample_sort <- apply(X=sample[,(first_col:last_col)],MARGIN=2,FUN=rank,ties.method = "min")

  #the warnings are set off as the minimum produces warnings, because for certain
  #cattle all farms have a value of 0
  options(warn=-1)

  #the descriptive table for the raion is calculated
  descriptive <- cbind(N,n,pop.perc=n/N*100,obtained,exp.sum,pop.sum,
                       ratio.adj,TwoCVperc=2*CV*100,
                       SampleSum=apply(X=sample[,(first_col:last_col)],MARGIN=2,sum),
                       PopSum=apply(X=raion[,(first_col:last_col)],MARGIN=2,sum),
PercPop=apply(X=sample[,(first_col:last_col)],MARGIN=2,sum)/apply(X=raion[,(first_col:last_col)],MARGIN=2,sum)*100,
                       mean_n=apply(X=sample[,(first_col:last_col)],MARGIN=2,sum)/n,
                       mean_N=pop.sum/N,
                       #!!! the minimum is actually not a minimum, it is the
                       #second smallest value
                       min_n=apply(X=sample[,(first_col:last_col)],MARGIN=2, function(x) return(min(x[x!=0]))),
                       min_N=apply(X=raion[,(first_col:last_col)],MARGIN=2, function(x) return(min(x[x!=0]))),
                       max_n=apply(X=sample[,(first_col:last_col)],MARGIN=2,max),
                       max_N=apply(X=raion[,(first_col:last_col)],MARGIN=2,max))

  #warning option set to on
  options(warn=0)

  #when for all farms 0 is found for a certain cattle this produces a value of
  #infinity, this is now set back to 0
  descriptive[descriptive==Inf] <- 0

  #the descriptive table for the current raion is saved
  write.csv2(x=descriptive,file=paste0(writedir,"descriptive",raion_sample,".csv"))
}

#the sampled farms are extracted
sampled_farms <- sample_oblast[sample_oblast$inc==1,]



  #the fraction of the number of farms per raion is calculated
  fraction <- data.frame(farms_inc=tapply(X=sample_oblast$inc,INDEX=sample_oblast$RAION,FUN=sum),
                farms=tapply(X=sample_oblast$inc,INDEX=sample_oblast$RAION,FUN=length),
                fraction=tapply(X=sample_oblast$inc,INDEX=sample_oblast$RAION,FUN=mean))

  #here the raionnumber is extracted from the rownames as numeric variable
  fraction$RAION <- as.numeric(x=rownames(fraction))

  #the fraction is merged to the dataframe of the sampled farms
  sampled_farms <- merge(x=sampled_farms,y=fraction,by.x="RAION")

  #the sample of the oblast and the dataframe containing only the sampled
  #farms are saved
  write.csv2(x=sample_oblast,file=paste0(writedir,"sample_oblast.csv"))
  write.csv2(x=sampled_farms,file=paste0(writedir,"sampled_farms.csv"))


  #the descriptive table like in software Statistica is calculated
  obtained <- apply(X=sampled_farms[,(first_col:last_col)]>0,MARGIN=2,FUN=sum)

  #a dataframe of all cattle multiplied by the extrapolation factor
  #of the respective farm is created
  expanded <- sampled_farms[,(first_col:last_col)]*sampled_farms$weight

  #the extrapolated sum of each cattle is calculated
  exp.sum <- apply(X=expanded,MARGIN=2,FUN=sum)
```

```
    #ratio adjustment is calculated
    ratio.adj <- pop.sum/exp.sum

    #the sampling error is calculated
    smpl.err <- sqrt(apply(X=sampled_farms[,(first_col:last_col)],MARGIN=2,var)/n_oblast)*(1-n_oblast/N_oblast)

    #the coefficient of variation is calculated
    CV <- smpl.err/apply(X=sampled_farms[,(first_col:last_col)],MARGIN=2,mean)
#the warnings are set off as the minimum produces warnings, because for certain
#cattle all farms have a value of 0
options(warn=-1)
#the descriptive table for the whole oblast is calculated
descriptive <- cbind(N_oblast,n_oblast,pop.perc=n_oblast/N_oblast*100,obtained,
                exp.sum,pop.sum,
                ratio.adj,TwoCVperc=2*CV*100,
                SampleSum=apply(X=sampled_farms[,(first_col:last_col)],MARGIN=2,sum),
                PopSum=apply(X=oblast[,(first_col:last_col)],MARGIN=2,sum),

PercPop=apply(X=sampled_farms[,(first_col:last_col)],MARGIN=2,sum)/apply(X=oblast[,(first_col:last_col)],MARGIN=2,su
m)*100,
                mean_n=apply(X=sampled_farms[,(first_col:last_col)],MARGIN=2,sum)/n_oblast,
                mean_N=pop.sum/N_oblast,
                #!!! the minimum is actually not a minimum, it is the
                #second smallest value
                min_n=apply(X=sampled_farms[,(first_col:last_col)],2, function(x) return(min(x[x!=0]))),
                min_N=apply(X=oblast[,(first_col:last_col)],2, function(x) return(min(x[x!=0]))),
                max_n=apply(X=sampled_farms[,(first_col:last_col)],MARGIN=2,max),
                max_N=apply(X=oblast[,(first_col:last_col)],MARGIN=2,max))
#warning option set to on
options(warn=0)

#when for all farms 0 is found for a certain cattle this produces a value of
#infinity, this is now set back to 0
descriptive[descriptive==Inf] <- 0

#the descriptive table of the oblast is saved
write.csv2(x=descriptive,file=paste0(writedir,"descriptive_oblast.csv"))
```