Estimating the worlds number of buildings, swimming pools, silos, windmills, intersections and the worlds proportion of ocean.

Maria Simonsen, Martin Jensen, Svend Nielsen and Simon Simonsen
Aarhus University

Abstract

This papers purpose is to examine the total number of buildings, swimming pools, intersections, silos and windmills in the world and the proportion of ocean. For achieving this rather ambitious goal, we will estimate these numbers by using a simple random sample.That is, we choose 1000 coordinates uniformly around the globe and construct a quadrant of 360000 square meters, where we by air photos manually determine the number of wanted objects. We conclude that the estimates we find for these objects are rather imprecise and have large variances but they are still fair estimates.

I. Introduction

This paper is written as an introductory assignment in the course Survey Sampling at Aarhus University. We will in this paper be using the notation from (Sampling by S.K. Thompson, 2012) in which we mainly will be using methods from Chapter II & III. For the estimation part we uniformly draw 1000 coordinates around the globe. From these coordinates we construct a 600 by 600 meter large quadrant, in where we count the numbers of seen objects. This will be an representative and simple random sample in the notation used in [Sampling, 2012] if we assume that we will have no overlaps. In our setup we have

$$P(\text{Observing more than 1 point in a quadrant}) = 0.00079.$$

So the assumption of no overlap is not completely unfair since the true probability is close to zero. In Figure 1, we see the first 200 coordinates we have been looking at. In Figure 2; one example of a 600m by 600m quadrant in which we have been counting the desired objects.

Figure 1: *Sampled coordinates*

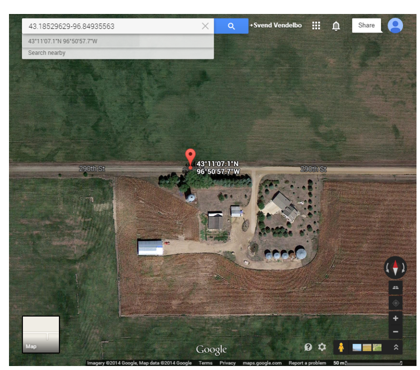

Figure 2: *quadrant with 6 silos, 10 buildings.*

II. Sampling

Sampling a uniformly point on a sphere can be carried out using polar coordinates. In order to use the coordinates on Google Maps we had to convert between polar and longitude/latitude coordinates. Both simulation and conversion was carried out by online software[Altitude and longitude sampler]. All coordinates were typed into google maps search bar and then we manually zoomed in to fixed distance. This arises a few problems, because the fixed distance is not fixed(in case of mountains etc.). Therefore there is some variation to the plot size. In some uncivilized areas the resolution is too poor to see anything, thus we can have missed a shed or two. In better areas with normal resolution we might also have missed things due to the angle of the photograph. In Africa people use huts as silos which are nondistinguishable from other huts. Those, we have not counted.

III. Data

A sample of the data is neatly being represented in the following Table 1.

Table 1: *The data*

	# of objects		
Quadrant:	1	$\cdots$	1000
Swimming Pools	0	$\cdots$	0
Buildings	4	$\cdots$	0
Silos	6	$\cdots$	0
windmills	0	$\cdots$	0
intersections	2	$\cdots$	0
1 if ocean 0 otherwise			
Ocean	0	$\cdots$	1

If we sum the rows we get the values represented in Table 2

Table 2: *SumRow data*

Variable	sum	$\sqrt{s^2}$
Swimming Pools	11	0.2
Buildings	1202	10.6
Silos	6	0.2
windmills	0	n/a
intersections	203	1.5
Ocean	712	

We can now go ahead and use the methods represented in (Sampling, 2012)

IV. Method

By using the notation from [Sampling 2012] we have that $N = 1416866666$ and $n = 1000$. We have that S is our sample containing the values from the 1000 different quadrants. We estimate the sample means and variances by (1) and (2)

$$\bar{y} = \frac{1}{n}\sum_{i\in S} y_i \tag{1}$$

$$\widehat{\mathbf{VAR}}(\bar{y}) = \frac{(N-n)s^2}{Nn}, \tag{2}$$

where $s^2 = (n-1)^{-1}\sum_{i\in S}(y_i - \bar{y})^2$. We can now estimate the population totals by using (Sampling, 2012) equations (2.7) and (2.10), that is

$$\hat{\tau} = N\bar{y} \tag{3}$$

$$\widehat{\mathbf{VAR}}(\hat{\tau}) = N(N-n)\frac{s^2}{n}. \tag{4}$$

Since n is large we can use normal approximations for confidence intervals with the estimate for the variances , that is

$$a_{\pm} = \bar{y} \pm z_{0.95}\sqrt{\frac{(N-n)s^2}{Nn}}. \tag{5}$$

We obtain the confidence intervals for the Ocean variable by using the hypergeometicdistribution as is described on page 59 in [Sampling, 2012].

V. Conclusion

We summed the results up in this very neat Table 3

Table 3: *summary table in millions*

Variable	a_-	$\hat{\tau}$	a_+	$\sqrt{\widehat{\mathbf{VAR}}(\hat{\tau})}$
SP	-0	16	31	9
BU	919	1703	2486	476
SI	-5	9	22	8
WM	n/a	0	n/a	n/a
IC	179	288	397	66
OC	0.69	0.712	0.74	0.014

The variance is big and we would want to make it smaller next time with less work. To make better estimates of number of land objects most pictures were irrelevant, because they were water. To improve this we could have used cluster analysis to group the pictures into water and into non-water. Another thing we could have done, was to gather our coordinates uniformly over land areas, to improve the effective sample size. If that sounds too hard, one can go ask the the HR department at Aarhus university - since they most surely wont have anything better to do :p - to look through more air photos such that the estimates variances is reduced and the accuracy is improved.

References

[Altitude and longtitude sampler] http://www.geomidpoint.com/random/

[Sampling, 2012] Sampling, Third Edition, Steven K. Thompson, SFU. Wiley publication.