Verification of excluding small sample from area framework

Shoji KIMURA

ASEAN Food Security Information System, email: wood_v@yahoo.co.jp

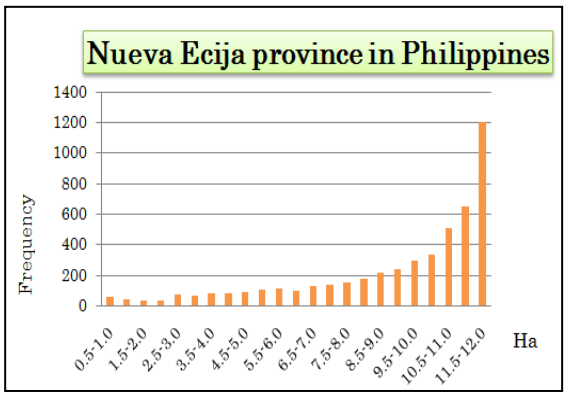
This study report is a complementary report of "The consideration for making an area framework in ALIS" which is written for a working level statistician who responses on conducting an area sample survey by ALIS. Therefore this practical method excluding a small area sample is confined to the utilization of an area sample survey which a sample of a framework has a limited area information. In addition, the working level statistician has to make consideration for a total statistical survey works like an environment of field survey not only a theoretical data accuracy.

1. Preface

ALIS operator selects an area meshes including a cultivated land by his or her visual operation. These meshes become a statistics framework of an area sample survey. In fact, this visual operation gives a big influence for the number of framework. The rule of ALIS operation is defined that *a narrow cultivated land in the area mesh does not assume as a cultivated land*. And I indicate a definition of the small area sample in ALIS as it would be less than 10% cultivated land area; it would be about less than 1.0ha in case of 300m mesh. This study report shows the verification on this consideration.

2. Reason 1 (Verification)

When estimating a total area, even if it excludes the small area samples from the area framework, it does not give the big influence to an estimation result.



Graph 1

Graph 1 shows the cultivated land area condition of the area meshes in the first samples on Nueva Ecija province in Philippines. Bureau of Agricultural Statistics (BAS) in Philippines took 4,928 samples as the first sample. This sample size itself is unnecessarily big[note]. However, these many samples data give many suggestions for us. See this graph, x-axis take the cultivated area in the area mesh and y-axis takes the appearance frequency of the area mesh. We can image this province locates a plain area where corrected the cultivated land around proper area for crop production.

[note] It is shown by study report "Verification on number of first sample" edited by Shoji Kimura.

Exclusion of small area mesh

----- Nueva Ecija province, Philippines -----

	Number of First Sample	Number of excluded meshes $(n-n')$	Rate of exclusion	Estimated Number of Framework $N'=\frac{n'}{n}N$	Area Average $\bar{y}=\frac{\sum y}{n}$	Estimated area $Y=\frac{\sum y}{n}N$	Standard deviation $\sigma=\sqrt{\frac{\sum(y-\bar{y})^2}{n-1}}$	Standard Error $SE=\sqrt{\frac{n-n'}{n-1}}\frac{\sigma}{\sqrt{n}}$	Standard Error Rate $SER=\frac{SE}{\bar{y}}\times100$
Feasibility study	n = 4,928	-	-	N= 24,021	922.29a	221,543ha	286.02a	3.63a	0.39%
Exclusion Less 1 ha mesh	n' = 4,869	59	1.2%	N'= 23,733	932.76a	221,372 ha	271.36a	3.47a	0.37%
Exclusion Less 2 ha mesh	n' = 4,793	135	2.7%	N'= 23,363	945.21a	220,829ha	254.64a	3.28a	0.35%
Exclusion Less 3ha mesh	n' = 4,686	242	4.9%	N'= 23,118	960.93a	222,148ha	234.97a	3.07a	0.32%

Table 1

Table 1 show the result of estimated agricultural land area by excluding the small area meshes from a sample group. On the feasibility study in Nueva Ecija province, the number of the statistics framework consisted of 24,021 area meshes and it was extracted 4,928 area meshes as the first sample. An average of the first sample was 922.29a. We can estimate the total agricultural land area by the simple estimation. And the estimated total area was 221,543ha. At the time, a standard deviation is 286.02a, a standard error is 3.63a, and a standard error rate is 0.39%.

For example, in case it excludes less 1ha meshes, concerned meshes are 59. In this case, the number of sample is 4,869; we can estimate the number of the statistics framework will be 23,733. At the time, the average of the first sample is 932.76a and the estimated total area is 221,372ha. Continuously, it calculates in case of exclusion less 2ha mesh and less 3ha mesh as below.

It will be understood that the number of area mesh in framework degrease with excluding the small area sample. On the other hand, the area averages of area mesh increase with excluding the small area sample. As the result, the estimated areas in each case are almost the same relative to the number of framework and area average. In addition, theoretical accuracy rate increase by excluding the small area sample based on a value of the standard deviation.

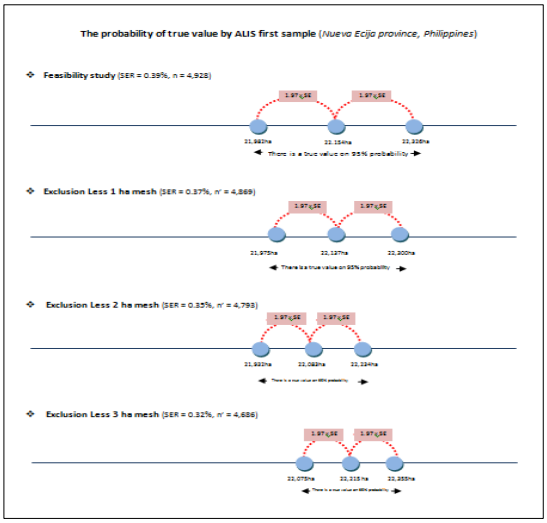
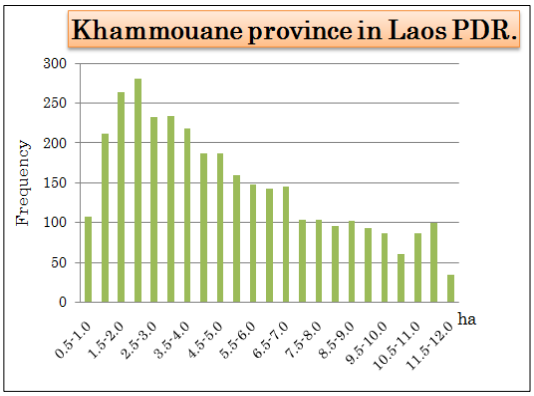


Figure 1

Figure 1 shows a provability range that there is a true value on 95% provability in case of the feasibility study and each case of exclusion of the small sample. This figure means, in case of feasibility study, with a central focus on the estimated area 221,154ha, there is a true area value between ±2σ, from 219,826ha to 223,261ha on 95% provability.

And it probably means that the more narrowed of a range, consequently increase of accuracy. As long as checking this diagram, an exclusion of the small area sample not seems to give a big influence to the total estimated data. Because the provability rang of all cases of exclusion of small sample enter to the provability range of feasibility which took the most samples.



Graph 2

Graph 2 shows the cultivated land area condition of the area meshes in the first samples on Khammouane province in Lao PDR. Department of Planning in Laos took 3,383 samples as the first sample. As I said, this sample size itself is unnecessarily big. I consider everybody trend to take many samples for a data estimation.

On this graph, x-axis takes the cultivated area in the area mesh and y-axis takes the appearance frequency of the area mesh, we can image this province locates on a mountain area where cultivated lands are dispersed geographically.

Exclusion of small area mesh

----- Khammoane province, Laos PDR-----

	Number of First Sample	Number of excluded meshes (n−n')	Rate of exclusion	Estimated Number of Framework $N' = \frac{n'}{n}N$	Area Average $\bar{y} = \frac{\sum yi}{n}$	Estimated area $\hat{Y} = \frac{\sum yi}{n}N$	Standard deviation $\sigma = \sqrt{\frac{\sum(yi-\bar{y})^2}{n-1}}$	Standard Error $SE = \sqrt{\frac{N-n}{N-1}}\frac{\sigma}{\sqrt{n}}$	Standard Error Rate $SER = \frac{SE}{\bar{y}} \times 100$
Feasibility study	n = 3,383	-	-	N = 18,311	499.81a	91,520ha	299.09a	4.64a	0.93%
Exclusion Less 1 ha mesh	n' = 3,275	108	3.2%	N' = 17,726	513.68a	91,055ha	293.88a	4.64a	0.90%
Exclusion Less 2 ha mesh	n' = 2,791	592	17.5%	N' = 15,107	575.76a	86,980ha	274.07a	4.68a	0.81%
Exclusion Less 3ha mesh	n' = 2,286	1,097	32.4%	N' = 12,373	648.07a	80,186ha	250.23a	4.87a	0.75%

Table 2

Once time let explain about this table, Table2 shows the result of the estimated agricultural land area by excluding the small area meshes from the sample group. On the feasibility study in Khammoane province, the number of the statistics framework consisted of 18,311 area meshes and it was extracted 3,383 area meshes as the first sample. The average of the first sample is 499.81a. We can estimate the total agricultural land area by the simple estimation. And the estimated total area is 91,520 ha. At the time, the standard deviation is 299.09a, the standard error is 4.64a, and the standard error rate is 0.93%.

For example, in case it excludes less 1 ha meshes, concerned meshes are 108. In this case, the number of sample is 3,275; we can estimate the number of the frame work will be 17,726. At the time, the average of the first sample is 513.68a and the estimated total area is 91,055ha.

Continuously, it calculates in case of exclusion less 2 ha mesh and less 3 ha mesh as below. It will be understood that the number of the area mesh in the framework degrease with excluding the small area sample. On the other hand, the area averages of the area mesh increase with excluding the small area sample.

It differs from Nueva Ecija province in Philippines in that, the estimated areas in case of exclusion less 2ha mesh and less 3ha mesh decrease obviously.

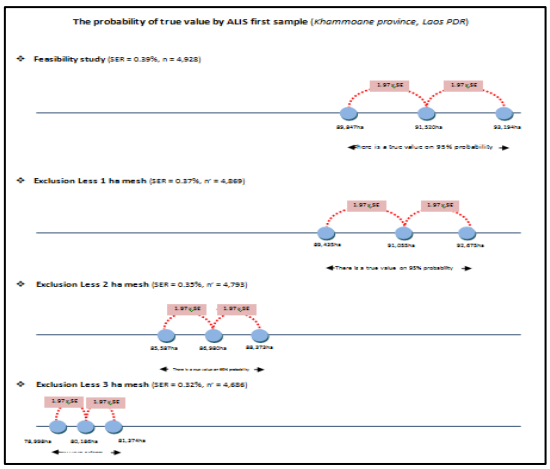


Figure 2

Checking this provability range figure, in case of Khammoane province, an exclusion of over 1 ha mesh gives an influence to the total area estimation and the data reliability.

From this verification using actual data, we can consider that *if one sample is smaller than a certain area level, the exclusion of this sample from the framework does not give an influence to estimate the total area and the data reliability*.

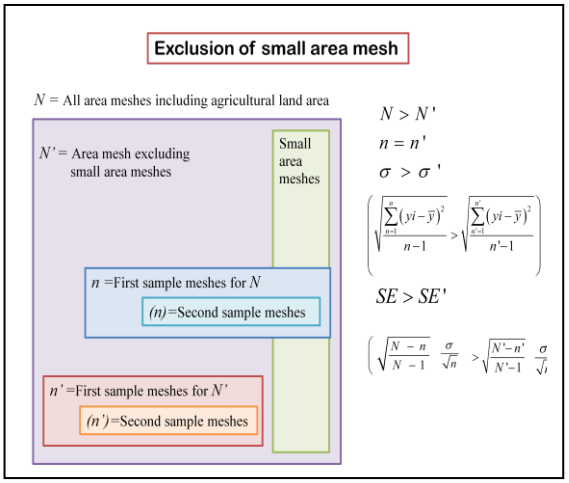


Figure 3

Figure 3 is a conceptual figure on exclusion of the small area mesh in ALIS.

N becomes all area meshes including the agricultural land area. N' becomes area mesh excluding the small agricultural land area mesh. In fact, N´becomes less than N. The number of n and n' sample is same.

In this case, the first sample meshes and the second sample meshes automatically exclude the small area meshes by excluding the small area meshes from the framework. It means that the standard deviation of n´ becomes less than n. In fact, the data variability becomes small. And the statistical accuracy increase.

3. Reason 2

As reason 2, it is necessary to consider about the field survey impossible risk with difficulty road to enter the survey field and safety risk for a researcher. Please remember figure 3; the second sample meshes automatically exclude the small area meshes by excluding the small area meshes from the framework.

4. Reason 3

As reason 3, it is important to increase the field survey efficiency for the second area sample for more cultivated land borderline and planted crops. Needless to say, the area mesh including a wide cultivated land has possibility to get many samples data than the small area sample. And to increase the field survey efficiency lead to increase the reliability of the estimated data.