# SAMPLING SEGMENTS IN AN AREA FRAME WITH A DISTANCE THRESHOLD

M. Fuentes

F. J. Gallego

# SAMPLING SEGMENTS IN AN AREA FRAME WITH A DISTANCE THRESHOLD

M. Fuentes: Department of Statistics, University of Chicago, 5734 University Av., Chicago Il 60637.1514.
F.J. Gallego: JRC. 21020 Ispra (Varese), Italy

## ABSTRACT

A simple random sample in an area frame usually gives a number of pairs of elements that are close to each other. These elements give redundant information since there is usually a high spatial autocorrelation at short distances. The efficiency of sampling is generally improved if we impose that the distance between two elements of the sample cannot be less than a certain threshold. However applying this restriction can introduce a significant perturbation of the sampling probability. Elements near the borders of the region are more likely to be selected. In the case of aligned sampling by repetition of a pattern in a square block, a distance threshold does not modify the marginal probability of each element of the population, although crossed probabilities are slightly changed.

Keywords: *Area frame sampling, distance threshold, systematic sampling.*

## 1. Introduction

Most variables observed in a geographic area have a spatial autocorrelation that is rather high at short distances and tends to decrease when the distance becomes larger. For example when segments (pieces of land) are sampled in an agricultural landscape, if a segment has high proportion of wheat, it is likely that another segment close to it has a high proportion of wheat.

A number of sampling procedures have been developed to optimize the precision of the estimates in the presence of spatial autocorrelation (Bellhouse 1977, Iachan 1985, Haining 1990, Arbia 1993, Cressie 1993 Benedetti, 1994). In general a knowledge is needed of the spatial autocorrelation function, since there is no optimal design plan for a general family of correlation functions (Bellhouse 1977).

When a sampling plan is set up, frequently no information is available on the autocorrelation function. In this case setting a distance threshold gives a simple alternative to more sophisticated sampling algorithms. Imposing that the distance between sample elements cannot be less than a certain threshold $r$ improves the sampling efficiency if it can be accepted that the autocorrelation is higher at distances shorter than $r$ than the one at distances longer than $r$.

## 2. Modifying a random sampling with a distance threshold.

Let us discuss an example of area frame in a small region with a square grid of 1 km represented in Figure 1 (the graphics actually correspond to the province of Varese, Italy). The area frame is defined on the basis of the region limits in the form of cartographic coordinates (UTM in this particular case). The approximately 1200 segments of the frame correspond to 100 ha except those on the border. This kind of area frame is being used for estimation of crop areas.

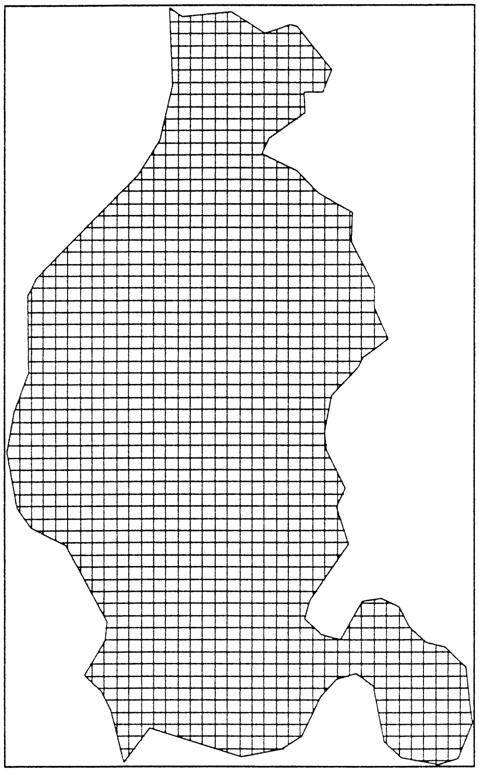


**Figure 1: Square grid on a small region**

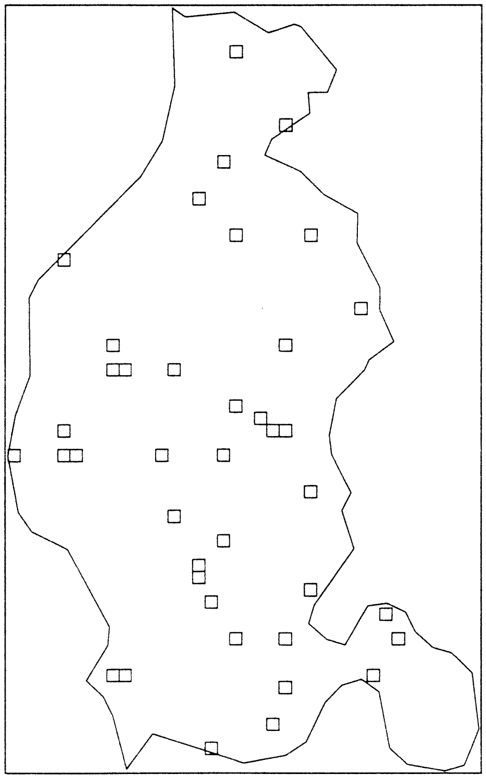


**Figure 2: Simple random sample**

Figure 2 shows the result of a simple random sampling of 40 segments in the grid. The quality of the estimates depends rather on the number of segments in the sample than on the sampling rate. It depends as well on how important the crop is in the region (Gallego, 95); minor crops are often poorly estimated. The results are also improved if the variability of the region is correctly represented by the sample. The geographic distribution of the sample is disturbing: sample segments are concentrated in some areas while other areas are missed by the sample. Some pairs of segments are contiguous and will presumably give redundant information.

## *2.1 Stepwise and Global Application of the Threshold.*

The easiest way to ensure that a distance threshold is respected is drawing a sample at random (with or without replacement), checking if there is a pair of elements at a distance less than $r$ and rejecting the sample in this case.

We can apply the threshold in two different ways:

- Stepwise application of threshold: In the $i^{th}$ step, the sample element $x_i$ is selected. If it does not respect the condition $(d(x_i ,x_j) < r \;\; for\; j<i),$ we select a new $x_i$ until the threshold condition is respected.
- Global application of the threshold: If the distance threshold is not respected for one pair, we repeat the complete sampling procedure.

It is obvious that the amount of computation needed is smaller with the stepwise application of the threshold.

The probability of possible samples depends on the way the threshold is applied.

## *2.2 Perturbations of the Probability that each Element is in the Sample.*

We draw a sample of $n$ segments out of $N$ using random sampling. The units can be points or aerial units (for example square segments of 1 km × 1 km for which distances are defined with the usual Euclidean distance between their centers). We impose that the distance between two points in the sample has to be always greater or equal than $s$ . The probability that a particular unit is in the sample depends on where that sampling unit is situated. Without the request of the distance threshold among sampling units, all of the points would have the same probability.

### 2.2.1 A fictitious example

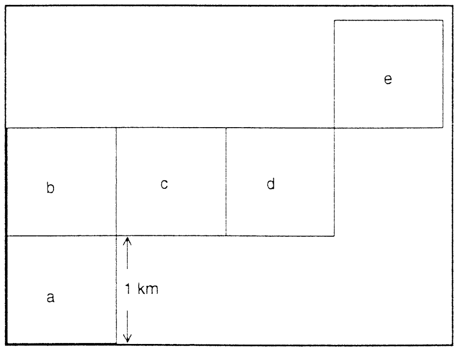


**Figure 3: Example of population of 5 elements**

The computation of exact probabilities for each sample in most real situations requires an unaffordable computing time, since the probability of each possible sample must be separately calculated; we see here an extremely small example to illustrate the modification of sampling probabilities when a distance threshold is introduced. We consider a population of 5 squares with unity side length as in Figure 3. We draw a sample of 3 elements out of these 5 with a distance threshold $r = 1.1$ km.

There are only 3 samples that meet the distance threshold; the probabilities we get applying the global threshold procedure are the same, while the stepwise procedure gives unequal probabilities.

To compute for example the probability of the sample {a, c, e}, we separately consider each of the 6 possible ways of getting the sample if order is considered: p($x_1$=a, $x_2$=c, $x_3$=e)= p($x_1$=a) × p($x_2$=c/$x_1$=a) × p($x_3$=e/$x_1$=a,$x_2$=c) = 1/5 × 1/3 × 1 =1/15, p(a,e,c)=1/30, p(c,a,e)=1/10, p(c,e,a)=1/10, p(e,a,c)=1/40, p(e,c,a)=1/20.

The probability of each sample is:

| Possible samples | Stepwise threshold | Global threshold |
|---|---|---|
| {a, c, e} | 3/8 | 1/3 |
| {a, d, e} | 2/8 | 1/3 |
| {b, d, e} | 3/8 | 1/3 |

The probability that each element is in the sample is:

| Elements | Stepwise threshold | Global threshold |
|---|---|---|
| a | 5/8 | 2/3 |
| b | 3/8 | 1/3 |
| c | 3/8 | 1/3 |
| d | 5/8 | 2/3 |
| e | 1 | 1 |

A quick analysis of this small example suggests that the probability $p(\underline{x})$ that a point $\underline{x}$ is in the sample is strongly correlated with the number of points inside the open ball $m(\underline{x}) = \#B(\underline{x},r) = \#\{\underline{z} / d(\underline{z},\underline{x})<r\}$: the more elements inside the ball, the more often $\underline{x}$ is not allowed to belong to the sample because of the incompatibility with another element, and the lower the probability that $\underline{x}$ belongs to the sample.

### 2.2.2 Examples in real regions.

We can modify the sample of Figure 2 with a distance threshold of 2.5 km. In this case we will obtain the sample represented in Figure 4. The geographic distribution is improved: contiguous segments are avoided but again some large areas are missed. A higher threshold would be necessary to avoid missing such large areas. Of course this can be achieved by selecting a larger sample, but ground survey costs would be higher.

When a region is stratified, the distance threshold can be separately applied to each stratum.

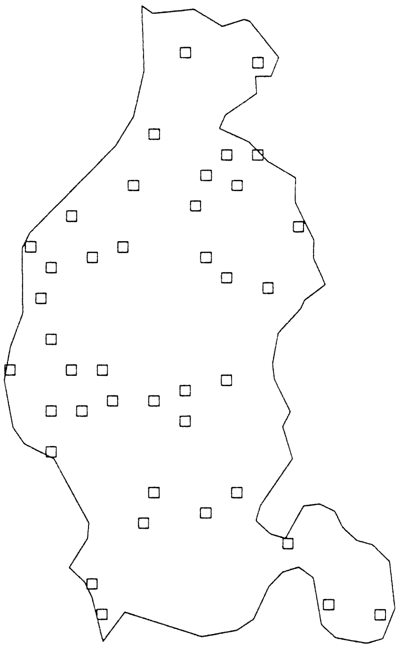


**Figure 4: Random Sample with a Distance Threshold**

We could see in the previous example that the probability that an element belongs to the sample depends on its position in a region, but in real cases it is not easy to compute this probability, even with a perfect knowledge of the region. Here we give the estimates of such probabilities by repeating the sampling in a small region: the island of Evvoia (Greece): 935 segments of 2 km × 2 km , out of which we draw a sample of 15 segments.

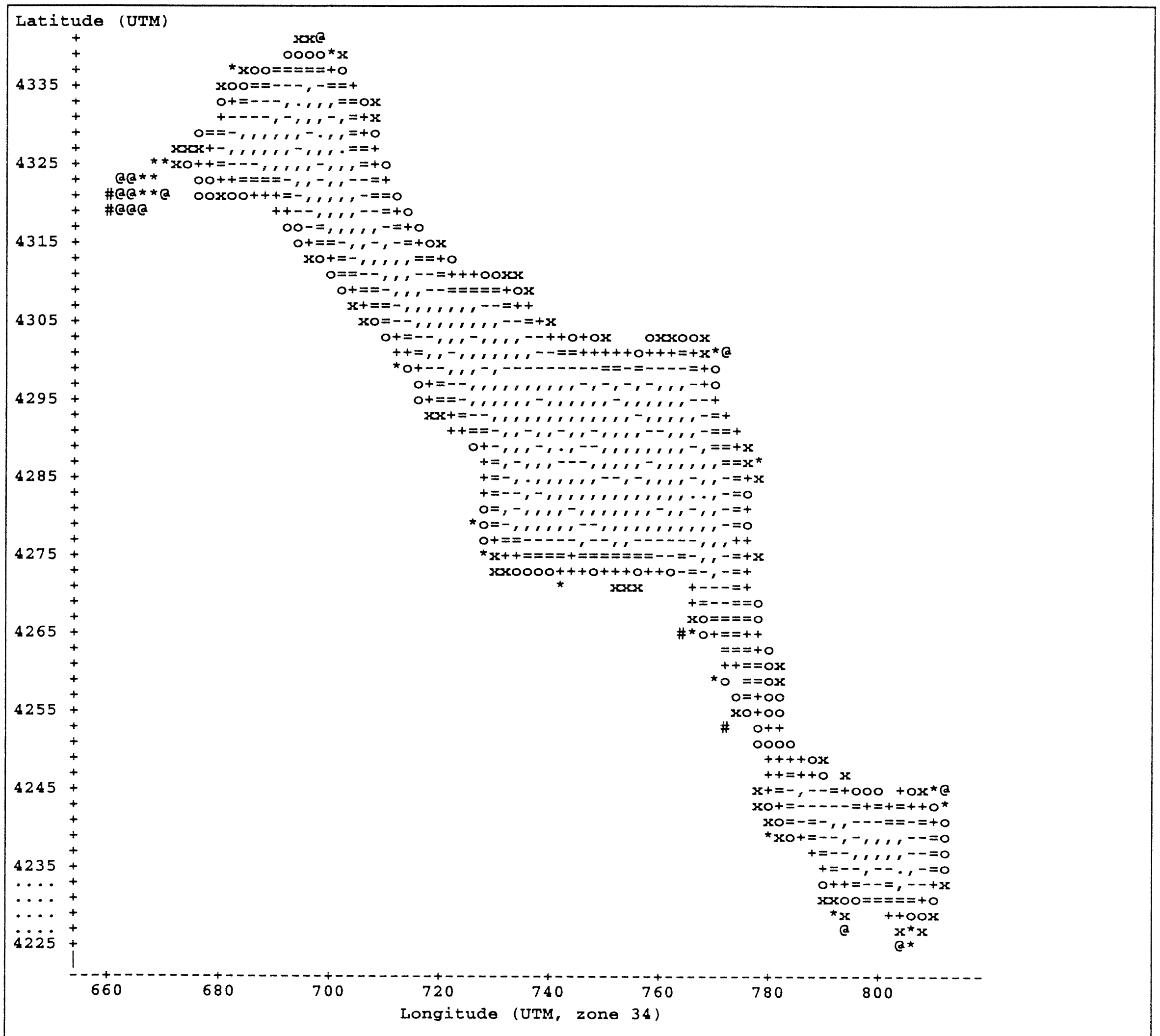
With a stepwise application of a threshold of 8 km between centres of segments, the probability is significantly higher for segments in the fringe of the region (Figure 5). The result is similar for a global application of the threshold, excepting that the modification of the probability is stronger: the range is 1.28 % - 2.48 % for the global application of the threshold instead of 1.39 % - 2.07 % for the stepwise application.

Some suggestions come out from this simulation:

- the stepwise application of a threshold alters the probability less than the global application.
- bias can be avoided by reducing the "a priori probability of elements near the border" in the basic sampling, or alternatively by using a Horvitz-Thomson estimator, that compensates the unequal sampling probabilities (Cochran, 1977). The Horvitz-Thomson estimator is unbiased if the sampling probabilities are known. However in real cases, sampling probabilities are to be estimated and the Horvitz-Thomson estimator becomes biased.

### 2.2.3 The choice of the threshold *r*

This sampling method clearly depends on the $r$ . We have not conducted specific studies on the stability of the results when $r$ is modified. So far we have selected a value of r with a subjective compromise: relatively large, but not so large that nearly all the region is covered by the "forbidden" areas $\{x/(d(x_i,x_j) < r\}$ at the end of the sample selection. A value $r \cong \sqrt{\frac{D}{3n}}$ , where D is the area of the region and n is the targeted sample size comfortably ensures this condition.



| | | | | |
|---|---|---|---|---|
| ... 13.50 - 14.25 | --- 15.00 - 15.75 | +++ 16.50 - 17.25 | xxx 18.00 - 18.75 | @@ 19.50 - 20.25 |
| ,,, 14.25 - 15.00 | === 15.75 - 16.50 | ooo 17.25 - 18.00 | *** 18.75 - 19.50 | ### 20.25 - 21.00 |

**Figure 5: sampling probability alteration after sequential application of a distance threshold in Euboea island (Greece)**

## 3. Systematic aligned sampling with a distance threshold.

Systematic geographic sampling is often used to improve the efficiency of sampling when there is a decreasing auto-correlation (Iachan, 83). The MARS Project of the E.C. has adopted this strategy to sample segments for area and production crop estimates at regional or national level (Gallego *et al* 94).

### *3.1 Applying a distance threshold to a sample of segments by square blocks.*

A distance threshold can be applied for systematic sampling on square blocks. Some care is necessary if we want the distance threshold to be met by segments in contiguous blocks, as well as inside each block.

### 3.1.1 Sampling square segments by repeating a pattern in a block

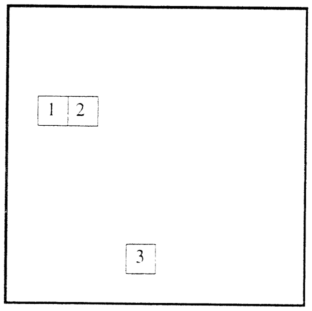


**Figure 6: Random pattern in a block**

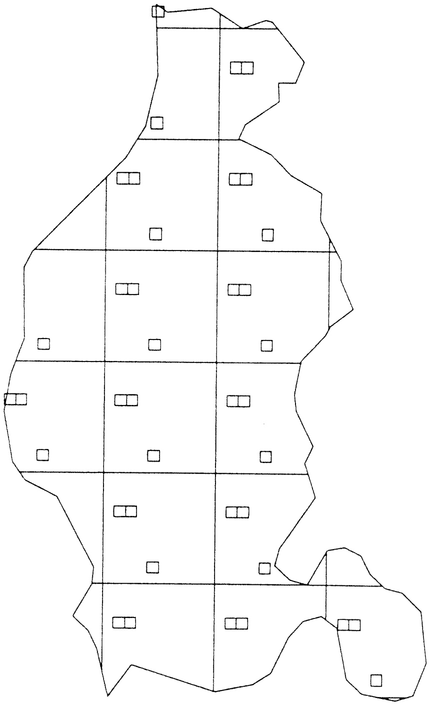
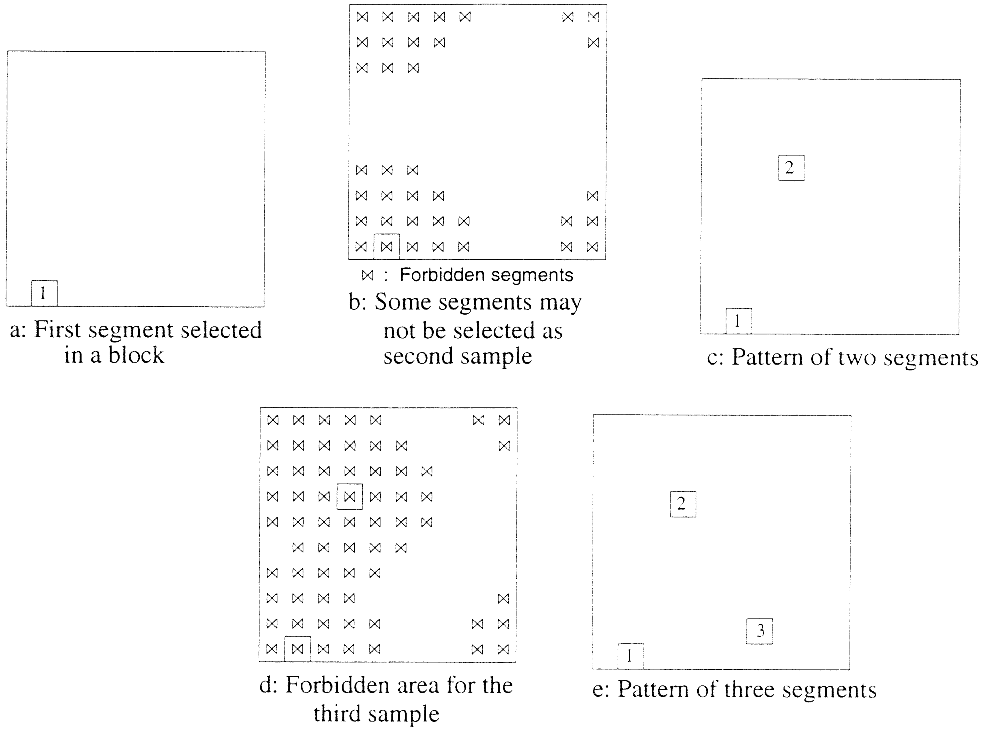


**Figure 7: Aligned sample by repeating a pattern.**

A sample by square blocks can be drawn by selecting at random a pattern that is repeated across the region as the example of Figure 7. In this example sampling is made using blocks of 10 km × 10 km . Several segments are chosen at random without replacement in a block.

The pattern is repeated for all the blocks in the region. This way of sampling is often known as systematic aligned sampling. The set of all the segments with the same relative position in all the blocks is called a replicate. Here we take a sample of 3 replicates.

Compared with a purely random sample, systematic sampling usually gives better precision (Cochran, 77). The main risk of systematic sampling is that a serious perturbation can appear if there is a periodical phenomenon with an interval that happens to coincide with the size of the block. This risk is negligible here: it is very unlikely that crops have a periodic behavior with a cycle of 10 km.



a: First segment selected in a block

b: Some segments may not be selected as second sample

c: Pattern of two segments

d: Forbidden area for the third sample

e: Pattern of three segments

**Figure 8: Sampling a pattern in a block with a distance threshold**

### 3.1.2 Systematic aligned sampling with a distance threshold.

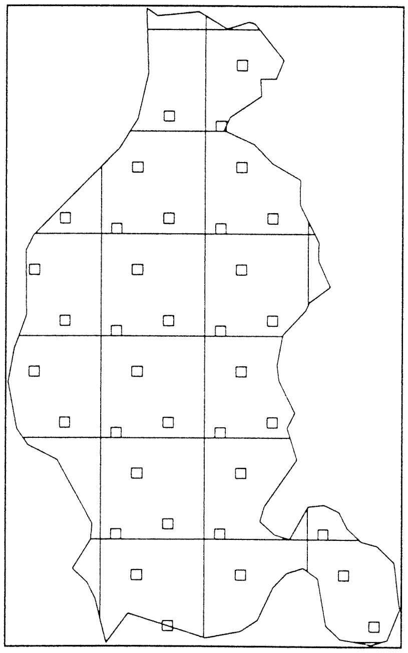


**Figure 9: Aligned sample with a distance threshold.**

In the random selection of the pattern in a block, it can happen that two or more elements are geographically close to each other. Figure 8 illustrates a way to get a pattern that will generate an aligned sample with a distance threshold.

We first draw a segment at random (Figure 8a); in the example it happens to fall close to the SW corner. Figure 8b shows the segments of the block at a distance less than 3.5 km, that are forbidden for the second replicate of the sample -second segment in the block-; notice that a number of segments around the other corners of the block are equally forbidden because they would be too close to the first replicate in another contiguous block. We surround the second replicate with a new forbidden area and so on.

Figure 9 represents the result of this operation after matching with the administrative limits of a region.

We draw a sample of *n* segments out of *N* using systematic sampling with square blocks, each block has $N_1 * N_2 = N$ segments. The sample units can be written as *(b,j)* from replica *j (j=1,r)* in block *b (b=1,B)*. A cluster $C_j$ is made up of all the segments *{(b,j), b=1,...,B}* that have the same relative position in each block. We define the next distance between segments $\underline{a} = (a_1, a_2)$, and $\underline{b} = (b_1, b_2)$:

$$d(\underline{a},\underline{b}) = \min_{\substack{\alpha=-1,0,1\\ \beta=-1,0,1}}\left(\sqrt{(b_1 - a_1 - \alpha N_1)^2 + (b_2 - a_2 - \beta N_2)^2}\right) \qquad (1)$$

In random sampling we have seen that the probability of each segment $\underline{a}$ of being in the sample is different; in systematic sampling is the same for all of them:

$$p(\underline{a}) = \sum_{\substack{d\left(x_{i_j},x_{i_k}\right)\geq r \ \forall j,k=1,\dots,n \\ d\left(x_{i_j},\underline{a}\right)\geq r \ \forall j=1,\dots,n}} p\left(x_{i_1}, x_{i_2}, \dots, \underline{a}, \dots, x_{i_n}\right). \qquad (2)$$

At first sight this probability seems to depend on the point $\underline{a}$ because for each point the segments that are at distance more than $s$ are different. However it can be easily proved that $p(\underline{a})$ is the same for any $\underline{a}$ . The proof is based on the invariance with regard to the translation modulo $(N_1, N_2)$:

$$(a_1, a_2) \oplus (z_1, z_2) = (a_1 + z_1 \pm N_1, a_2 + z_2 \pm N_2) \qquad (3)$$

The sampling probability remains constant because the borders of the region are not considered when the block pattern is sampled, and the probability of each point (or segment) depends only of its position in the block.

However there is a slight change of the joint probability that a pair $(\underline{a},\underline{b})$ is in the sample. The joint probability depends on the number of elements common to the balls $B(\underline{a},s)$ , $B(\underline{b},s)$, $B(\underline{a}, 2s)$, and $B(\underline{b},2s)$. For example, drawing 2000000 times a pattern of 4 points in a grid of 1 km step with a block size of 12 km × 12 km and a threshold of 2.9 km gave joint probabilities for compatible pairs of points ranging from 0.0631 % to 0.0665 % . This alteration of joint probabilities introduces a slight bias in the estimation of the variance.

## REFERENCES:

ARBIA G., 1993. The use of GIS in spatial statistical surveys. International Statistical review, vol. 63, n. 2, pp. 339-359.

BELLHOUSE D.R., 1977. Some optimal designs for sampling in two dimensions, Biometrika 64, 605-611.

BENEDETTI R., PALMA D., 1994, Optimal sampling designs for dependent spatial units. Environmetrics

COCHRAN W., 1977, Sampling Techniques. New York: Wiley

CRESSIE N., 1993, Statistics for spatial data. Revised edition. New York. Wiley.

GALLEGO, F.J., 1995, Sampling Frames of Square Segments, Report EUR 16317, Office for Publ. of the E.C. Luxembourg. 68 pp. ISBN 92-827-5106-6

GALLEGO, F.J., DELINCÉ, J., CARFAGNA E., 1994, Two-Stage Area Frame Sampling On Square Segments For Farm Surveys. Survey Method. vol 20, No. 2 , pp. 107-115.

HAINING R.P., 1990, Spatial data analysis in the social and environmental sciences. Cambridge University Press.

IACHAN R., 1983, Asymptotic theory of systematic sampling, Annals of Statistics, 11, 959-969.

IACHAN R., 1985, Plane sampling. Statistics and probability letters. 3, 151-159.