EXTRAPOLATING INTRA-CLUSTER CORRELATION TO OPTIMIZE THE SIZE OF SEGMENTS IN AN AREA FRAME.

E. Carfagna

F. J. Gallego

EXTRAPOLATING INTRA-CLUSTER CORRELATION TO OPTIMIZE THE SIZE OF SEGMENTS IN AN AREA FRAME.

E. Carfagna. Univ. Bologna. Dep. of Statistics, V. Belle Arti 41, 40126 Bologna, Italy
F.J. Gallego: JRC, 21020 Ispra (Varese), Italy

ABSTRACT

In the frame of the "Rapid Crop Area Estimates in the European Community" we use a sample of squared segments (pieces of land) of 49 ha. each; estimates are made for difference of crop areas between years. The optimum size seems to be larger than the current one, and much larger if ground survey data are obtained by photo-interpretation.
The main purpose of this paper is to assess a method, based on splitting the segments into pieces of 1/4 ha., to forecast the precision of estimates with larger segments. The tests made for France confirm the belief that better estimates can be obtained with a fixed cost by enlarging the size of the segments.
We consider as well the problem of assessing smaller segments, for which an easier technique can be used.

Keywords: *Area frame sampling, segment size, intra-cluster correlation*

1. THE MARS PROJECT OF THE EUROPEAN COMMUNITY.

The MARS project (Monitoring Agriculture with Remote Sensing) of the European Community (E.C.) was launched to assess and develop applications of Remote Sensing to Agricultural Statistics. It is carried out by the Institute of Remote Sensing Applications (IRSA) of the Joint Research Centre (JRC) of the E.C. We refer here to two activities of this project, "Regional Crop Inventories" and "Rapid Estimates at the E.C. level", based on area frame sampling and high resolution satellite images.

1.1. Regional crop inventories by segment sampling and remote sensing.

This activity has to do with estimation methods for crop area and production based on area frame sampling and satellite images (Gallego, 93, Fuentes 94). In 1988 annual crops had an absolute priority: soft and durum wheat, barley, rapeseed, dried pulses, sunflower, maize, cotton, tobacco, sugar beet, potatoes, rice, soya and fallow. Attention is now being given more and more to permanent crops, pastures, and non-agricultural land. Squared segments with a size of 49 ha. were selected for most regions in the E.C. The same technique has been used in some countries of eastern Europe (Czech republic, Rumania), where the size of the segments was larger to take into account the size of the fields.

1.2 **Rapid Estimates at the E.C. Level.**

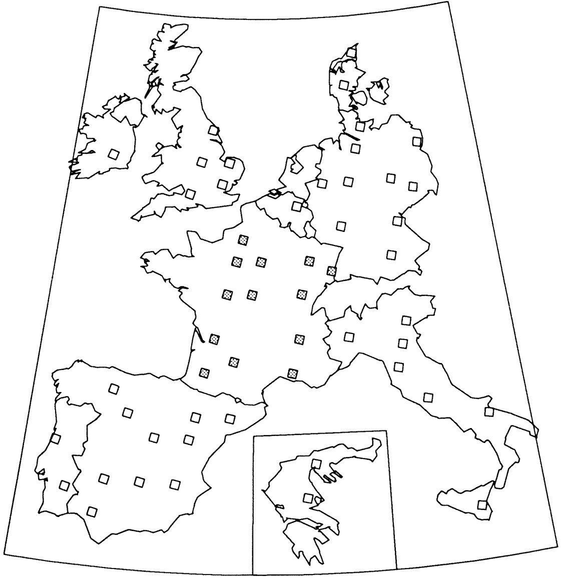

Figure 1: Sample of 53 sites used for rapid estimates in the E.C.

The main goal is giving rapid estimates of area and yield change of annual crops compared with the previous year based on a two-stage sampling scheme: 53 sites (figure 1) of 40 Km × 40 Km. with a sample of 16 squared segments of 700 m. × 700 m. in each of the sites. Individual data are acquired by photo-interpretation of SPOT-XS or Landsat-TM images. An average of three images is analysed for each site with a minimum of ground information, (a general knowledge of the dominant crops in each area). A ground survey is made for an a posteriori validation of the photo-interpretation. A regular report (4-8 issues a year) is produced with an update of the estimates.

2. **ASSESSING SMALLER SEGMENTS (CZECH REPUBLIC)**

Since 1992 the IRSA is giving support to the Administration of the Czech Republic to use area frame surveys based on segments. In 1992, a sample of 417 segments of 400 ha. (2000m. × 2000m.) was drawn (fig. 2). Standard formulae for stratified sampling (Cochran, 77, Allen, 90, Gallego, 93) are used to compute area estimates.

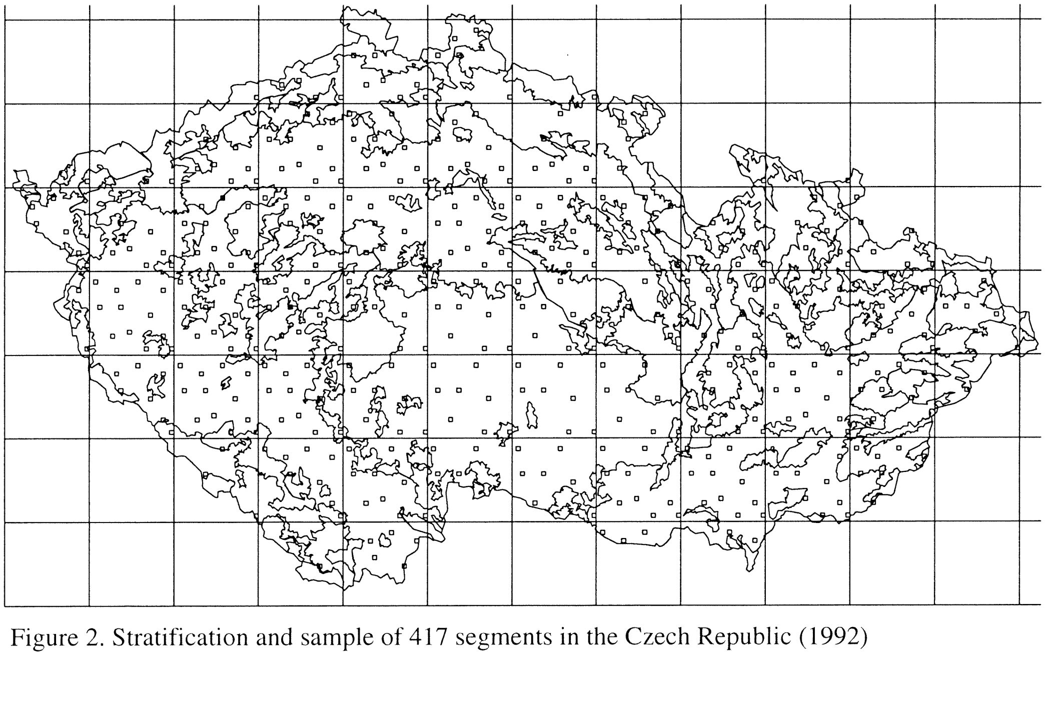

Figure 2. Stratification and sample of 417 segments in the Czech Republic (1992)

2.1. **Simulating smaller segments**.

In order to study the adequacy of this segment size, we cut each of the segments to get squared segments with sides of 1800m., 1600m, and so on down to 200m. (figures 3a, 3b). The new smaller segments have the same centre of the original segment.

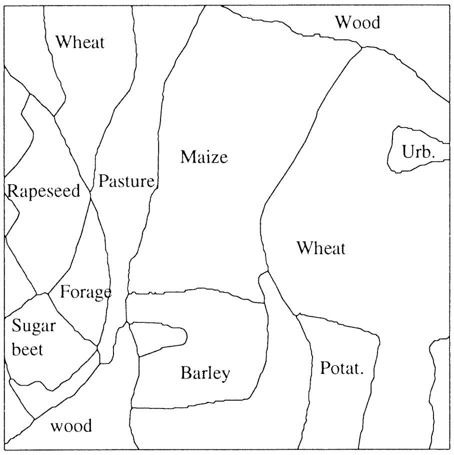

Fig. 3a : Example of squared segment of 400 ha.

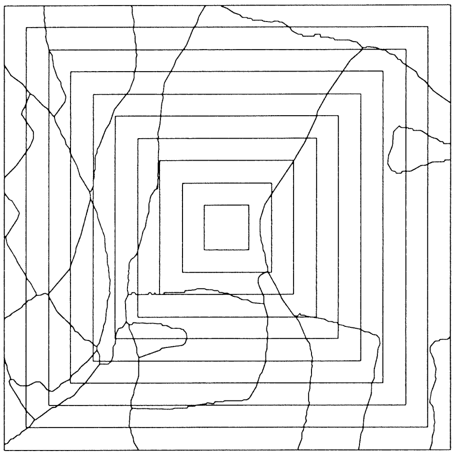

Fig. 3b: Cutting the segment into smaller sizes

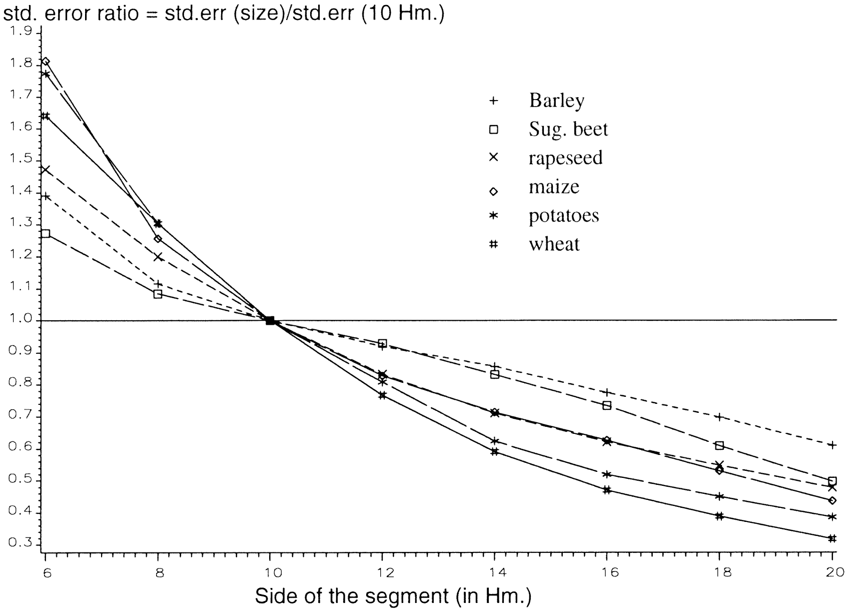

Figure 4: Standard error ratio for different segment sizes compared with the standard error for segments of 100 ha. (side= 10 Hm.)

Area estimates have been computed for each segment size. Since the sample size (number of segments) remains the same, the standard errors are larger when the segments are smaller, as it could be expected (fig. 4). The standard error depends strongly on the segment size for wheat and potatoes, and much less for barley and sugar beet. Now we have to compare standard errors to the survey cost as a function of the segment size.

2.2. **Cost Function**.

The cost function has been determined in a rather rough way. The enumerators were asked in a meeting about the average time they would need to visit a segment of a different size (100 ha., 225 ha.) compared to the actual segments of 400 ha., including location, walking across the segment and drawing fields. As they did not have much experience, they could not answer at first. They were asked later if the time needed for a segment of 100 ha. would be more or less than half the time needed for a segment of 400 ha.. They found that "about half the time" was a reasonable answer. Following this opinion, a linear cost function was used with value 600 at 400 ha. and 300 at 100 ha.:

$$Cost(S) = 200 + S \qquad (1)$$

where S is the area of the segment in hectares. Cost units do not correspond to any particular currency.

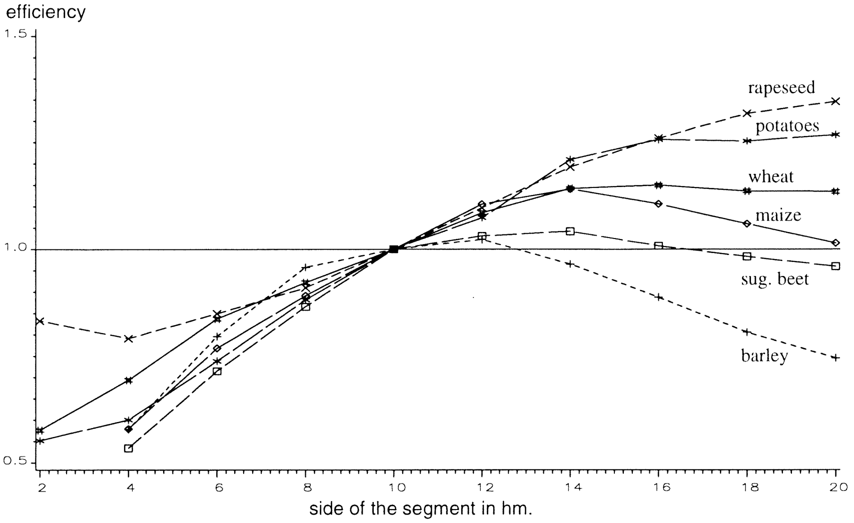

Figure 5: Relative efficiency of the segment size

This cost function can be discussed to take into account different strata. Intensive agriculture and marginal areas have different cost structure. General costs, such as digitizing segments should be included, though they are small compared to the ground work cost.

2.3. **Optimum size of the segment.**

For a particular crop and stratum, the optimum size is the one that minimises $Var(S) \times Cost(S)$, where *Var(S)* is the variance of the area estimate. Figure 5 gives such functions for the stratum "intensive agriculture". We have plotted the cost efficiency using the segment of 100 ha. as a reference:

$$Eff(S) = \frac{Var(100\,ha.) \times Cost(100\,ha.)}{Var(S) \times Cost(S)} \qquad (2)$$

For this stratum a segment size of about 200 ha. (1400m. × 1400m.) is a reasonable compromise although larger segments are more efficient for some crops, as rapeseed.

3. FORECASTING THE BEHAVIOUR OF LARGER SEGMENTS.

In the previous section the optimum size was smaller than the available, and no extrapolation was needed. In other situations we need to forecast the variance for larger segments. This is more likely if segments are not visited on the ground, but photo-interpreted. In this case general costs (not related to the size of the segment) are high, including image acquisition and processing and training photo-interpreters, while marginal costs (cost of photo-interpreting one additional hectare in a segment) are lower, and hence the optimum size of the segment is likely to be larger. The actual size of the segments in this case is 49 ha. The results below are obtained on a sample of 206 segments in 13 sites in France (fig. 1). The problem can be estimating areas in a single year or estimating the evolution between two years. We shall override the fact that it is a two-stage sampling and focus on a method to extrapolate the variance of the estimators.

3.1. Splitting the segment into elementary units. Variance of the estimates and intra-cluster correlation..

The approach we have followed is considering the segment as a cluster of elementary units. Fig. 6 shows an example of segment with an overlaid grid of 100m. producing elementary units of 1 ha.

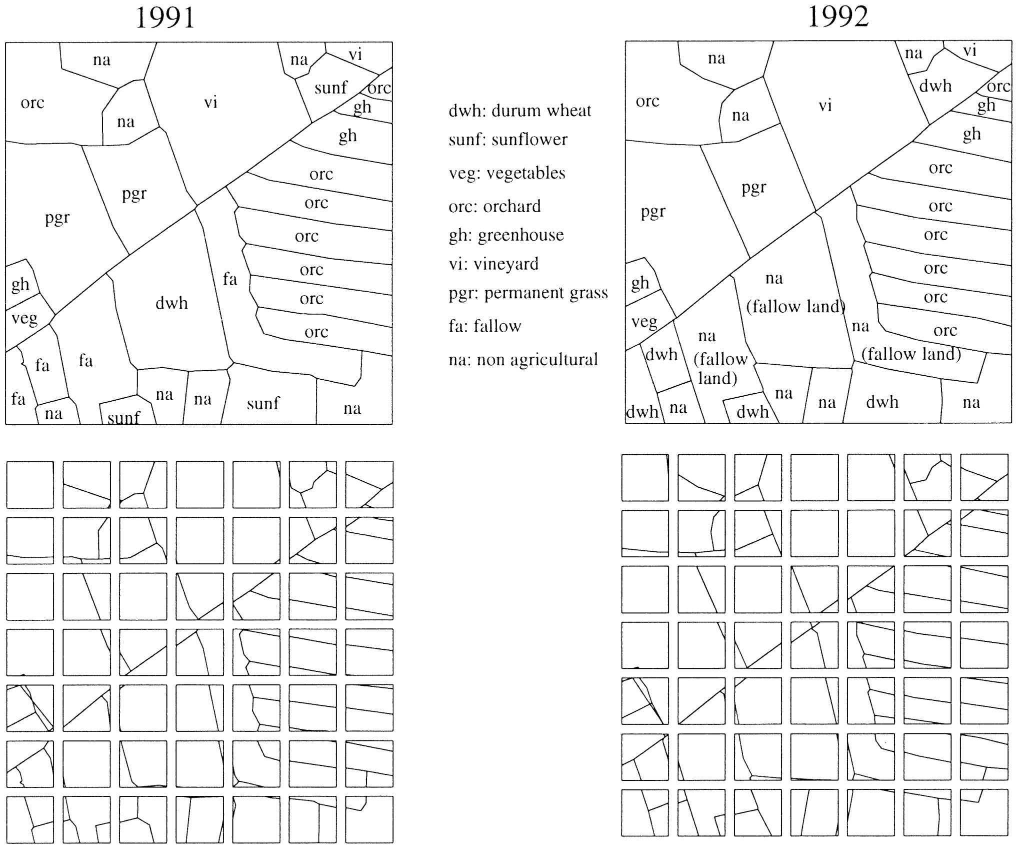

Fig. 6: Example of segment split into elementary units of 1 ha.

The results given below correspond to splitting each segment into 196 elementary units of 50m. × 50m.

We have a sample of n clusters (segments) out of a population of *N*. The number of elements per cluster is *M*. The variance among elements is

$$S^2 = \frac{1}{NM-1}\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij} - \bar{\bar{Y}})^2 \qquad (3)$$

estimated by s^2 computed on the sample (substituting *N* by *n* and $\bar{\bar{Y}}$ by $\bar{\bar{y}}$).
The intra-cluster correlation is estimated by:

$$\rho_M = \frac{1}{(M-1)(nM-1)s^2}\sum_{i}\sum_{j\neq k}(y_{ij} - \bar{\bar{y}})(y_{ik} - \bar{\bar{y}}) \qquad (4).$$

In the example we have computed ρ for squared segments of M=4, 9, 16, 25, 36,...196 elementary units of 50m. × 50m . The variance of the estimate can be written as:

$$\hat{V}(\bar{\bar{y}}) = \frac{1-f}{nM}s^2(1+(M-1)\rho_M) \qquad (5)$$

where $$\rho_M = \frac{1}{(M-1)(nM-1)S^2}\sum_{i}\sum_{j\neq k}(y_{ij} - \bar{\bar{y}})(y_{ik} - \bar{\bar{y}}) \qquad (6)$$

Forecasting ρ_M will give a criterion to optimize the variance of the estimate for larger segment sizes. We will consider here a simple functional extrapolation, rather than time series methods. We focus instead on the possibility of decomposing ρ_M in order to improve the extrapolation.

We can write the estimated intra-cluster correlation as a sum of terms indexed by the distance between the pairs of elements inside a cluster:

$$\rho_M = A_M\sum_{d}\sum_{i}\sum_{jk/d(jk)=d}(y_{ij} - \bar{\bar{y}})(y_{ik} - \bar{\bar{y}}) = A_M n\sum_{d} M_d\varphi_d \qquad (7)$$

For simplicity we work with the L_1 distance between elements within a cluster: $d(j,k) = |j_1 - k_1| + |j_2 - k_2|$ (*j* and *k* are rows and columns of elements in the clusters).
$A_M = \frac{1}{(M-1)(nM-1)S^2}$ is a term that depends on the cluster size M and M_d is the number of pairs of elements at a distance *d* in a cluster of size M and

$$\varphi_d = \frac{1}{nM_d}\sum_{i}\sum_{jk/d(jk)=d}(y_{ij} - \bar{\bar{y}})(y_{ik} - \bar{\bar{y}}) \qquad (8)$$

is basically a covariance at distance d.

3.2. **Extrapolating intra-cluster correlations.**

If we accept that *y* is second order stationary, and hence φ_d is a valid estimate inside clusters of any size M , extrapolating φ_d leads to an extrapolation of ρ_M .The intra-cluster correlations can be directly extrapolated using a simple functional adjustment, such as $r = a \cdot M^b$, as suggested by Cochran (1977). Figure 7 shows an example of

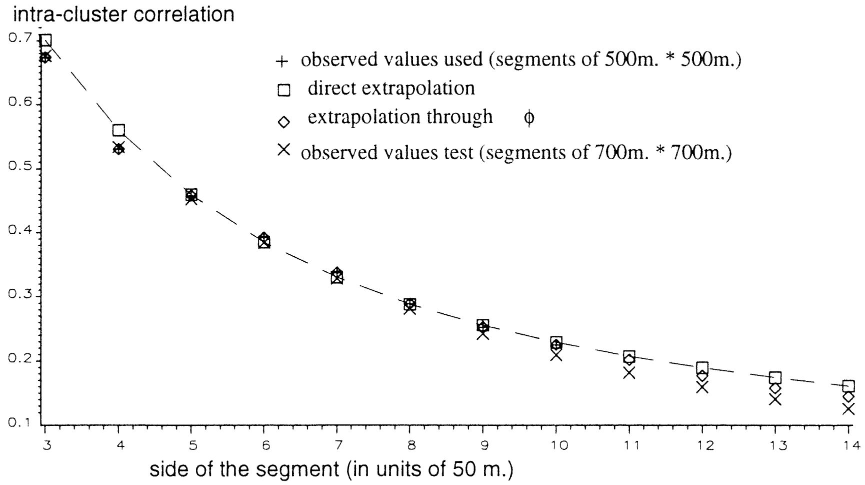

Figure 7: Extrapolation of intra-cluster correlation for soft wheat

extrapolation using $r = a \cdot (M + c)^b$ for common wheat, which is the most important crop in the area.

If a similar extrapolation is made on φ_d, a more reliable extrapolation of the intra-cluster correlation can be obtained by substituting in (1) the missing φ_d by their extrapolated values $\overset{*}{\varphi}_d$:

$$\overset{*}{r}_M = A_M n \left[\sum_{d=1}^{d0} M_d \varphi_d + \sum_{d0+1}^{d1} M_d \overset{*}{\varphi}_d \right] \qquad (9)$$

where $d0$ is the maximum distance between single elementary units for which φ_d can be estimated on the available data and $d1$ is the maximum distance for segments of size M, unavailable, for which we wish to extrapolate the intra-cluster correlation.

We have taken the example of soft wheat to compare extrapolation of intra-cluster correlation in both ways: directly and through φ_d, we have made as if we had only had segments of 25 ha. (side=10, M=100, *d*0=18), and used the same kind of functional extrapolations.

None of the extrapolations is very good (they may be improved for example by using time series techniques). Both extrapolations are compared (table 1) with the actual values computed on the segments of 49 ha. The reason of this improvement is that, in formula (9), the weights M_d corresponding to the extrapolated $\overset{*}{\varphi}_d$ are small compared to the rest.

side	M	observed (test)	extrapolated directly	extrapolated through φ_d
11	121	0.182	0.207	0.202
12	144	0.159	0.189	0.177
13	169	0.141	0.174	0.158
14	196	0.126	0.161	0.145

Table 1: extrapolation of intra-cluster correlation by two methods

3.3. **Comparison of both extrapolation methods in terms of optimum segment size**.

Another kind of comparison can be done between both ways of extrapolating intra-cluster correlation, directly and through φ_d . We could check in the example of the previous section that the extrapolated correlations are rather different; the question now is whether the conclusions on the most efficient size of the segment are substantially different.

By optimum size of the segment for an area frame sampling we mean the size that minimises the estimation variance under the constraint of a fixed budget. In the case of rapid estimates the variable is the difference between the surfaces of main crops in successive years. Obviously, the optimum segment size is determined for each crop, due to the different behaviour of variances. A compromise can be achieved at a second step, on the basis of the importance of different crops.

Our aim now is to analyse the effect of differences of the two kinds of extrapolation on the behaviour of variances; thus we disregard here the compromise size of a multipurpose survey and focus on one single widespread crop: soft wheat.

A main element of the methodology to identify the optimum segment size is the link between estimate precision and segment size, that allows the identification of the size that minimises standard errors of the estimates. This link is the cost function that gives the number of segments of each size that can be observed with fixed budget. The difference of cost between a large segment and a small one decreases if photointerpretation is performed instead of ground survey.

Fixed costs do not influence the optimum size, being of the same amount for different sizes and number of segments, thus they have not been taken into account. Variable costs for ground survey have been formalised by the expression: $C = n(35 + 0.7M)$, where C is the total variable cost (total cost minus fixed cost), M is the segment size and n is the number of segments.

Using the expression (5) the variance of the estimate can be described as a function of the segment size. Figure 8 shows the standard error for y = "area of soft wheat in 1992 - area of soft wheat in 1991" .

It can be seen that the minimum value for estimate standard error corresponds to a segment size of 196 hectares for standard error estimated adopting the extrapolation of ρ_M and to segment size 100 adopting the extrapolation of ϕ_d. The two estimates corresponding to optimum segment sizes are: 96.2 extrapolating ρ_M and 97.9 extrapolating ϕ_d. The main difference between the two behaviours concerns the higher increase of estimate standard error for segment sizes larger than 196 hectares. The standard error estimated through ϕ_d extrapolation shows more clearly the optimum size.

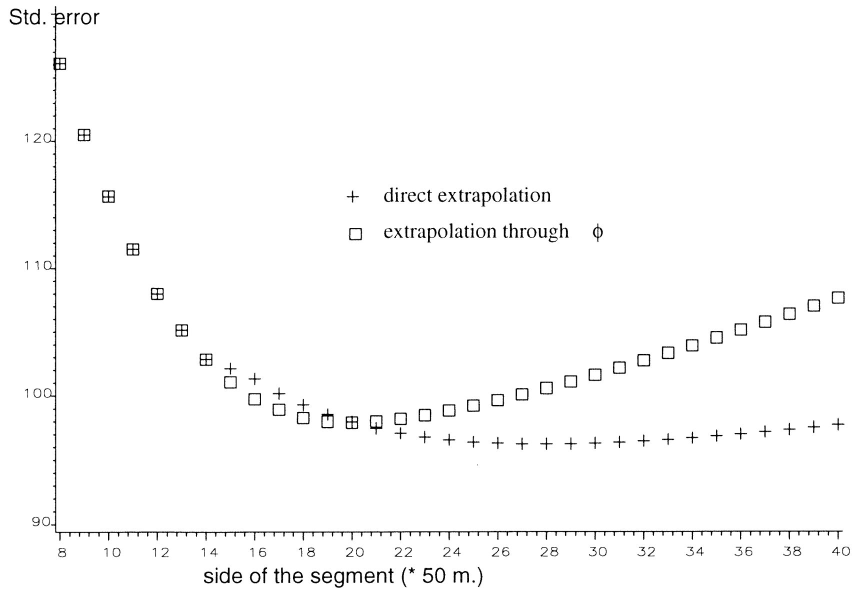

Figure 8: Standard error as a function of the segment size with fixed cost (ground survey).

Figure 9 shows the estimate standard error with the cost function $C = n\ (68.75 + 0.125\ M)$, better adapted to a photo-interpretation of segments. It can be noticed that the optimum segment size for photo-interpretation is higher than 400 hectares for this variable. The behaviour of the two curves is rather similar, although standard error tends to be higher when extrapolated through φ_d .

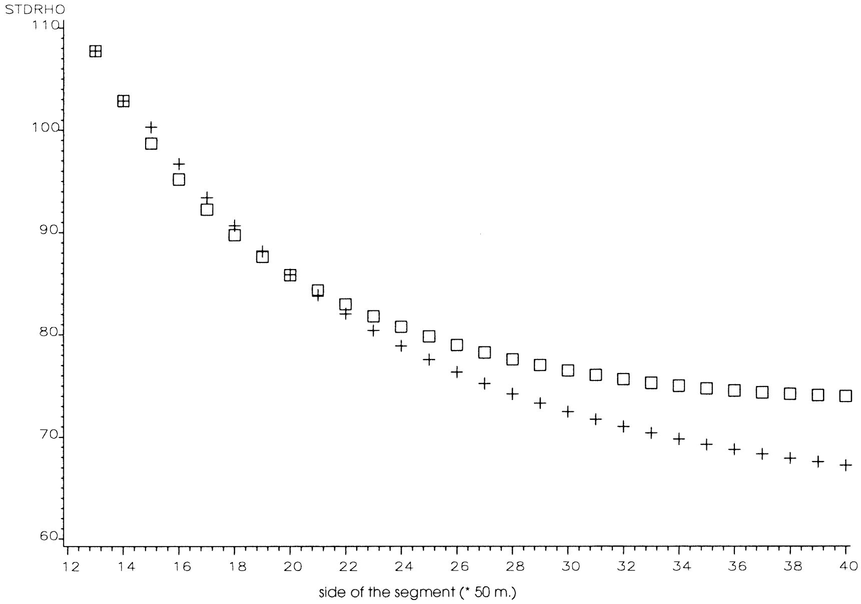

Figure 9: Standard error as a funcion of the segment size with fixed cost (photo-interpretation)

SUMMARY

In an area frame of squared segments on squared blocks, where clusters are built with segments that have the same relative position in each block, if there is no stratification and blocks are relatively small, cluster estimator gives lower variances than standard formulae, although there is the problem of its unstable variance. Making independent permutations of the sampling segments in each block, if blocks are complete, gives the same estimation as using cluster estimator, but improves the stability of the variance. When there is a stratification, cluster estimator is not a good choice due to the correlation among segments in each stratum: The cluster estimator works well when segments in each cluster are quite heterogeneous. Area frame in Spain is without stratification; then cluster estimator with permutations is a good option to improve the variance.

ACKNOWLEDGEMENTS

We are grateful to the Spanish Ministry of Agriculture, in particular to P. Sánchez, J.M. Fernández del Pozo, and M.J. Postigo that have kindly provided the data. L. Ambrosio made a number of useful comments.

REFERENCES:

Allen, J.D., 1990, A Look at the Remote Sensing Applications Program of the National Agricultural Statistics Service. Jou. of Official Stat. Vol 6, n. 4, pp. 393-409.

Ambrosio L., Alonso R., Villa A., 1993, Estimación de superficies cultivadas por muestreo de áreas y teledetección. Precisión relativa. Estadística Española, Vol 35, pp. 91-103.

Cochran W., 1977, Sampling Techniques. New York: John Wiley and Sons

Cotter, J. Nealon J. (1987), Area Frame design for Agricultural Surveys. U.S. Dept. of Agriculture. Nat. Agr. Stat. Serv.

Gallego F.J., Delincé J., Rueda C., 1993, Crop area estimates through remote sensing: stability of the regression correction. Int. J. Remote Sensing. Vol 14, n.18, pp 3433-3445.

Gallego, F.J., Delincé, J.,Carfagne E., 1994 Two-Staged Area Frame Sampling On Squared Segments For Farm Surveys. Survey Methodology (in print).

González F., López S., Cuevas J.M., 1991, Comparing Two Methodologies for Crop Area Estimation in Spain Using Landsat TM Images and Ground Gathered Data. Remote Sens. Environ. no 32, pp. 29-36.

Meyer-Roux J., 1990. Présentation du Projet Pilote de Télédétection Appliquée aux Statistiques Agricoles. Conference on the Appl. of Remote Sensing to Agricultural Statistics. Office for Publications of the E.C. Luxembourg.