Using remote sensing for crop area estimation

Javier.gallego@jrc.it

Joint Research Centre (JRC)

IPSC - Institute for the Protection and Security of the Citizen

Ispra - Italy

http://ipsc.jrc.ec.europa.eu/

http://www.jrc.ec.europa.eu/

Satellite images (possibly classified) can be used as

- Basic information for area estimation
- Covariates for a posteriori accuracy improvement
- Graphical support for ground work
- Tool to improve the sampling design of a ground survey (stratification)
- Indication for quality control of a ground survey

Area is estimated by counting pixels in a classified image

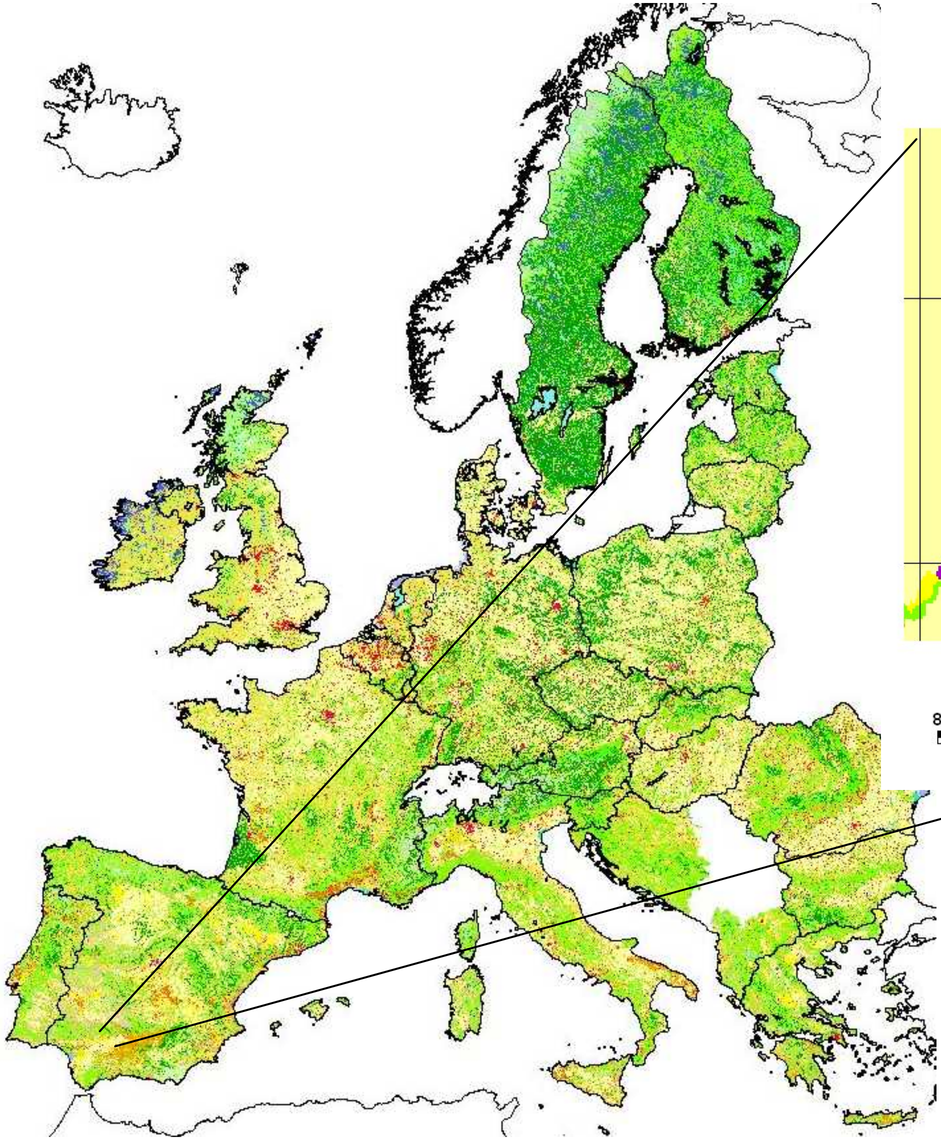
Sources of area estimation error:

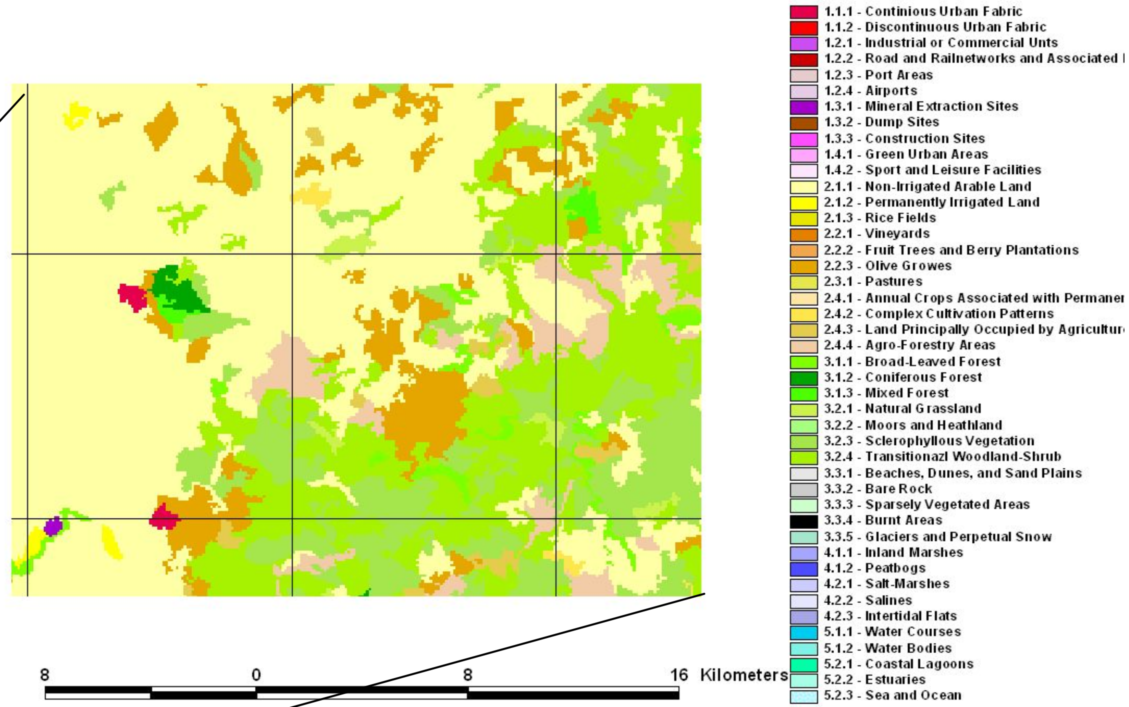
- Mixed pixels (boundary). Error depends on
  - Resolution, geometry (% of mixed pixels)
  - Relative radiometry of different classes
  - "suitable resolution": most pixels should be pure
- Misclassification of pure pixels

Direct area estimation by photo-interpretation (polygon area measurement)

- **Example: CORINE Land Cover**
- **By photo-interpretation of TM images**
- **Nearly homogeneous rules in most European Countries**
- **Nomenclature of 44 classes**
- **Minimum polygon size: 25 ha**
- **Some mixed classes such as agro-forestry, complex agricultural patterns, etc.**
- **In the early times of CLC (90's), it was often presented as a source of direct land cover area estimators**
  - But further analysis has shown that this is only acceptable if there is no alternative

CORINE Land Cover 2000





Partial view (rasterised 100m)

confusion matrix with "pure LUCAS points" (excluding points too close to boundaries)

CLC \ LUCAS	Artificial	agriculture	permanent grass	Broadleaved forest	Coniferous forest	Mixed forest	Woddland-shrub-heath	other	Total
artificial	1602	83	288	44	26	20	94	51	**2208**
agriculture	815	13892	2692	341	90	65	455	268	**18618**
Pasture	306	579	5678	194	107	44	544	409	**7861**
Broadleaved forest	98	85	276	3411	411	597	477	149	**5504**
Coniferous	162	74	216	541	7992	1675	354	597	**11611**
Mixed forest	74	36	101	596	1478	1056	156	262	**3759**
Woddland-shrub-heath	80	158	1330	603	1402	422	2192	795	**6982**
heterogeneous	387	2357	2087	705	167	127	593	147	**6570**
other	24	30	447	20	178	48	287	3753	**4787**
Total	**3548**	**17294**	**13115**	**6455**	**11851**	**4054**	**5152**	**6431**	**67900**

LUCAS / CLC2000	Buildings	Other artificial	Arable land	Vineyard	other permanent	Olive trees	permanent grass	Broadleaved forest	Coniferous forest	Mixed forest	Woddland-shrub-heath	Bare land	Water
Urban	**25,1**	23,4	7,2	0,4	1,7	0,4	26,6	2,3	1,3	0,9	8,1	1,7	0,8
Other artificial	12,8	**27,1**	5,4	0,3	0,4	0,2	28,3	4,6	2,4	1,4	7,7	6,3	3,1
Arable	1,0	3,2	**73,7**	0,7	1,0	0,5	10,7	2,5	0,9	0,7	3,2	0,9	1,0
Vineyard	1,5	3,6	14,6	**53,3**	4,0	2,8	6,5	2,4	1,1	0,4	8,3	0,9	0,7
Fruits	1,5	3,0	13,7	1,8	**54,2**	3,1	5,9	3,5	2,0	0,2	7,6	1,7	1,8
Olive	0,9	2,4	10,1	2,2	3,2	**64,5**	4,3	2,4	0,8	0,7	6,6	1,0	0,7
Pasture	0,8	2,4	11,8	0,1	0,3	0,1	**62,0**	3,8	2,1	0,9	10,1	4,0	1,5
Broadleaved forest	0,3	1,3	2,7	0,2	0,3	0,2	5,9	**56,6**	7,9	10,9	10,4	0,8	2,5
Coniferous	0,1	1,6	1,3	0,1	0,1	0,1	2,2	5,1	**65,4**	14,8	3,5	0,8	5,1
Mixed forest	0,3	1,5	1,8	0,1	0,1	0,1	2,9	15,4	38,8	**27,2**	4,5	0,8	6,6
Woddland-shrub-heath	0,1	1,2	1,7	0,1	0,3	0,6	11,1	9,0	23,0	7,0	**34,0**	5,8	6,2
Bare land	0,1	0,2	0,7	0,1	0,4	0,3	11,2	0,9	2,7	0,6	21,4	**58,5**	2,9
Water	0,3	0,7	0,6	0,0	0,0	0,0	5,1	1,0	7,3	2,0	5,0	1,1	**76,9**
Heterogeneous	1,8	3,6	29,6	2,9	3,4	3,8	24,8	10,6	3,5	2,6	10,7	1,1	1,5
Burnt	0,0	0,0	3,6	0,0	0,0	0,0	7,1	10,7	39,3	10,7	28,6	0,0	0,0

Fine scale profiles of aggregated CLC2000 classes, based on LUCAS 2001

Land cover change: Example of straight estimation

- Consider CORINE Land Cover (CLC90) and CLC2000
- Both layers have the same geometry in an area of 3.5 Mkm2
- Direct overlay gives an “estimate”of ~20% of change in land cover type
- Remaking the photo-interpretation of both layers gives <5% change in land cover type.
- Probably closer to reality
- No sampling error, but
- Bias due to
  - Photo-interpretation errors,
  - Scale effect.
- For this period these figures are acceptable
  - Because we have no alternative
  - We should have better figures for 2006-2009

Errors from misclassification of pure pixels

No sampling error if complete image

Possible large bias

Λ = confusion matrix for the population

$$\hat{Z}_c = \frac{\lambda_{+c}}{\lambda_{++}} D = \frac{\text{pixels classified as } c}{\text{total pixels}} \text{ area of the region}$$

$$Commission\ \ error\ \ \varphi_c = 1 - \frac{\lambda_{cc}}{\lambda_{+c}} \qquad Omission\ \ error\ \ \psi_c = 1 - \frac{\lambda_{cc}}{\lambda_{c+}}$$

$$relative\ \ bias \quad b_c = \frac{\lambda_{+c} - \lambda_{c+}}{\lambda_{c+}} = \varphi_c \frac{\lambda_{+c}}{\lambda_{c+}} - \psi_c$$

Pixel counting as area estimator

Rule of thumb: do not use pixel counting if your expected commission/omission error is more than twice the targeted accuracy.

Example: if you want an accuracy of ± 5% (semi-confidence interval?), do not use pixel counting unless you are confident that your classification accuracy is >90%.

Gaussian distribution does not protect against bias or subjectivity

Pixel counting as area estimator

Example with maximum likelihood supervised classification (discriminant analysis)

- **Region of ~ 100,000 km^2**
- **Area of cereals ~ 2 Mha**
- **Accuracy of classification ~ 70%**
- **Tuning the parameters (a priori prob.), we can easily get an area of pixels classified as cereals between 1.5 and 2.5 Mha.**
  - If we think the area is 2.3 Mha, we will tune the classification to get that figure.
  - It may be right, but we are using RS as a “sexy dress” to make our belief more attractive.
  - There may be a tendency to underestimate changes if we use historical statistical data as a reference

Pixel counting as area estimator (2)

- We can tune the parameters to balance commission and omission errors on a test sample
- This gives a good protection against bias if the sample is statistically valid (random, systematic, etc…)
  - Random sample ≠ hap-hazard set
- We are implicitly using a calibration estimator.
  - We better use a calibration estimator explicitely.

Bias ≈ Commission error – omission error

If we have a confusion matrix, we can correct the bias.
Cannot we?

Ex: Photo-interpretation made for the EU LUCAS survey
Raw confusion matrix (simplified nomenclature):

Strata \ Ground	Arable	Perm. Crops	Perm. Grass	Forest Wood	Other	*Total*	Comm. Error %	Omis. Error %
Arable land	67313	1751	17597	2035	2760	*91456*	**32.9**	**8.3**
Perm. Crops	651	9516	546	573	287	*11573*	**16.9**	**21.7**
Perm. Grass	4940	658	26969	3693	4244	*40504*	**28.6**	**43.1**
Forest & Wood	308	185	1962	16248	1277	*19980*	**16.4**	**28.5**
Other	195	47	299	186	2925	*3652*	**6.3**	**74.5**
Total	*73407*	*12157*	*47373*	*22735*	*11493*	**167165**		

Let us look at the class "forest and wood"

Commission < Omission ⇒ We should <u>increase</u> the estimates by ca. 12%
Right?

Bias and confusion matrix

But in LUCAS the sampling rate of the non-agricultural strata is 5 times lower
the corresponding rows of the confusion matrix should be multiplied by 5

Weighted confusion matrix

Ground / Strata	Arable	Perm. Crops	Perm. Grass	Forest Wood	Other	*Total*	Comm. Error %	Omis. Error %
Arable land	67313	1751	17597	2035	2760	*91456*	**32.0**	**10.7**
Perm. Crops	651	9516	546	573	287	*11573*	**15.7**	**27.3**
Perm. Grass	4940	658	26969	3693	4244	*40504*	**24.0**	**52.2**
Forest & Wood	1540	925	9810	81240	6385	*99900*	**21.1**	**8.2**
Other	975	235	1495	930	14625	*18260*	**12.8**	**48.3**
	75419	*13085*	*56417*	*88471*	*28301*	*261693*		

Commission > Omission ⇒ We should <u>reduce</u> the estimates by ca. 13%

Confusion matrices for calibration

Confusion matrices should be computed on a proper sample of test pixels

Correctly extrapolated

Independent of the training pixels

- (everybody knows, but…)
- Spatially uncorrelated (this is sometimes forgotten)
- Not very important for robust classifiers

The proper way to use a confusion matrix for area estimation is the calibration estimator (extensive bibliography)

- The calibration estimator inherits bias from ground data, not from image classification
- It has a sampling error that depends on the size of the test set.

Combining ground survey and satellite images to improve the accuracy of estimates

Main approaches: calibration and regression estimators.

Common features:

- combine accurate information on a sample (ground survey) with less accurate information in the whole area, or most of it.
- Unbiasedness is provided by the ground survey.
- The more accurate the ground survey, the higher the added value of RS.

Variant if ground data are too difficult/expensive (e.g: forest in very large areas):

- Accurate information from high or medium resolution on a sample of images
- Less accurate information from coarse resolution (AVHRR, VEGETATION, MODIS, MERIS)

A : **Confusion matrix on a sample of test pixels**

A_g : **ground truth totals**

A_c : **pixels classified by class**

Λ : **Confusion matrix on the population**

Λ_g : **ground truth totals (unknown to be estimated)**

Λ_c : **pixels classified by class**

Error matrices:

$$\Pi_c(g,c) = \frac{\lambda(g,c)}{\lambda(g,+)} \qquad \Pi_g(g,c) = \frac{\lambda(g,c)}{\lambda(+,c)}$$

$$P_c(g,c) = \frac{a(g,c)}{a(g,+)} \qquad P_g(g,c) = \frac{a(g,c)}{a(+,c)}$$

Calibration estimators with confusion matrices

Straightforward identities:

$$\Lambda_g = \Pi_g \Lambda_c \qquad \Lambda_g = \Pi_g \Lambda_c$$

$$A_g = P_g A_c \qquad A_c = P_c A_g$$

Estimators:

$$\hat{\lambda}_{dir}(g) = P_g \Lambda_c \qquad \lambda_{inv} = P_c^{-1} \Lambda_c$$

Relative efficiency of the same order of regression estimator.

Y: Ground data (% of wheat)

X: Classified satellite image (% od pixels classified as wheat)

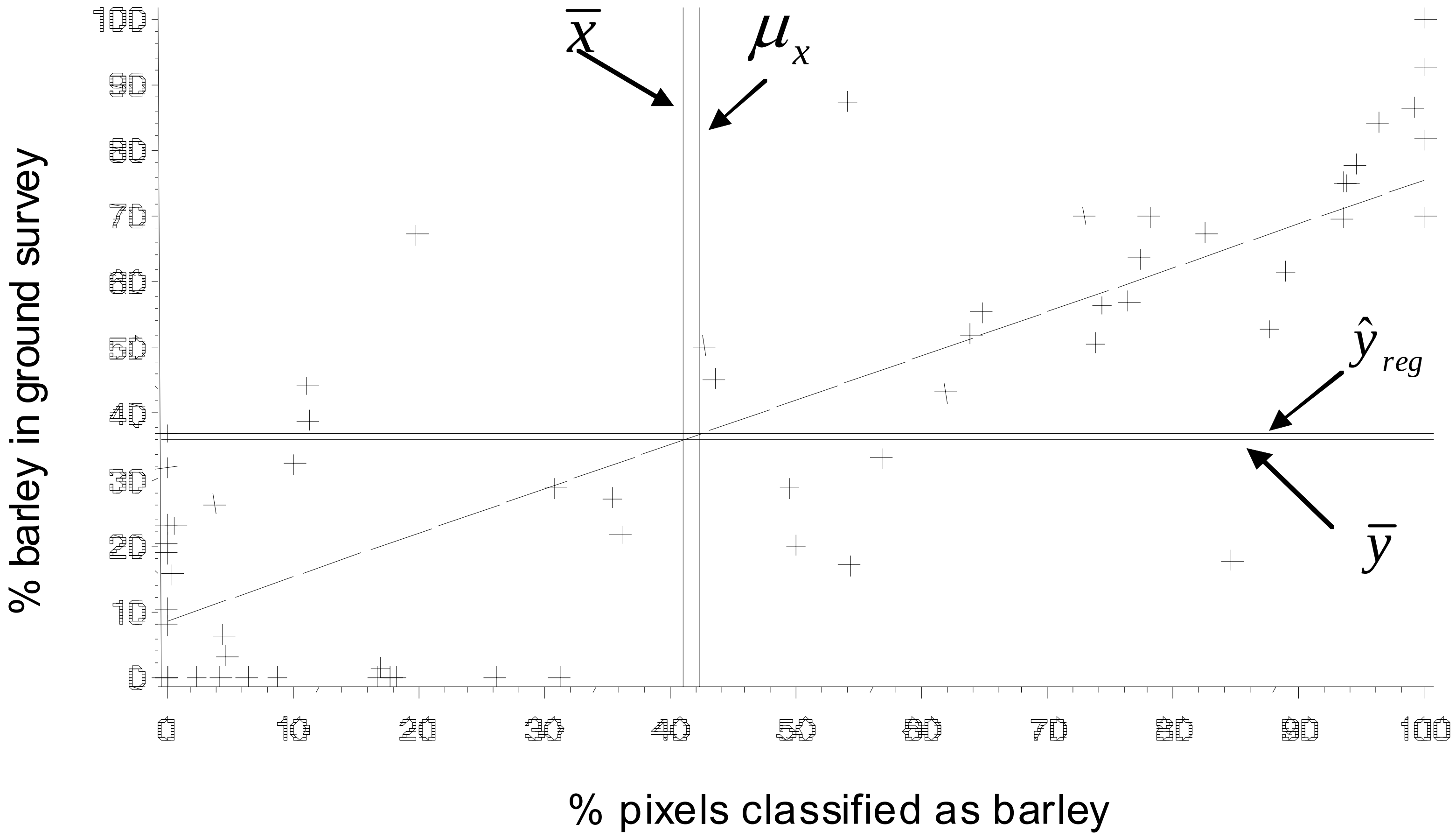
$$Y = a + bX + \varepsilon$$

Regression estimator $\hat{y}_{reg} = \overline{y} + b\left(\mu_x - \overline{x}\right)$

Difference estimator if slope b **pre-defined: less efficient, but more robust.**

Ratio estimator if $a = 0$

Regression estimator



Regression estimator

Relative efficiency (coarse approximation)

$$rel\ eff \sim \frac{1}{1-r_{xy}^2}$$

better approximation:

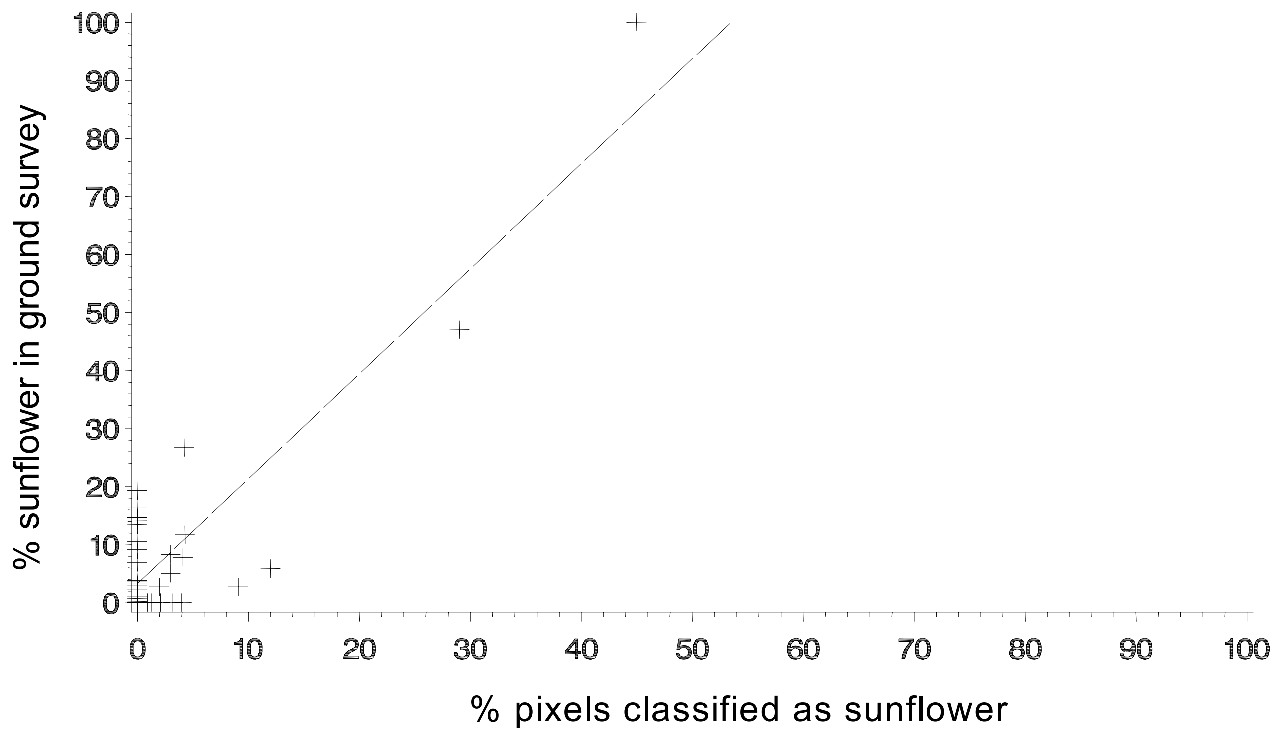
$$V(\hat{y}_{reg}) = \frac{N-n}{N \times n}\left(1 + \frac{1}{n-3} + \frac{2G_x^2}{n^2}\right)\sigma_y^2\left(1-\rho^2\right) \qquad G_x = \frac{k_{3x}}{\sigma_x^3}$$

An efficiency = 2 means that :
n segments + regression ~ 2n segments (only ground survey)
Criterion to assess cost-efficiency

Relative efficiency of the same order of calibration estimator.

Regression is not very suitable for point sampling: only 4 points in the regression plot: (0,0), (0,1), (1,0), (1,1)

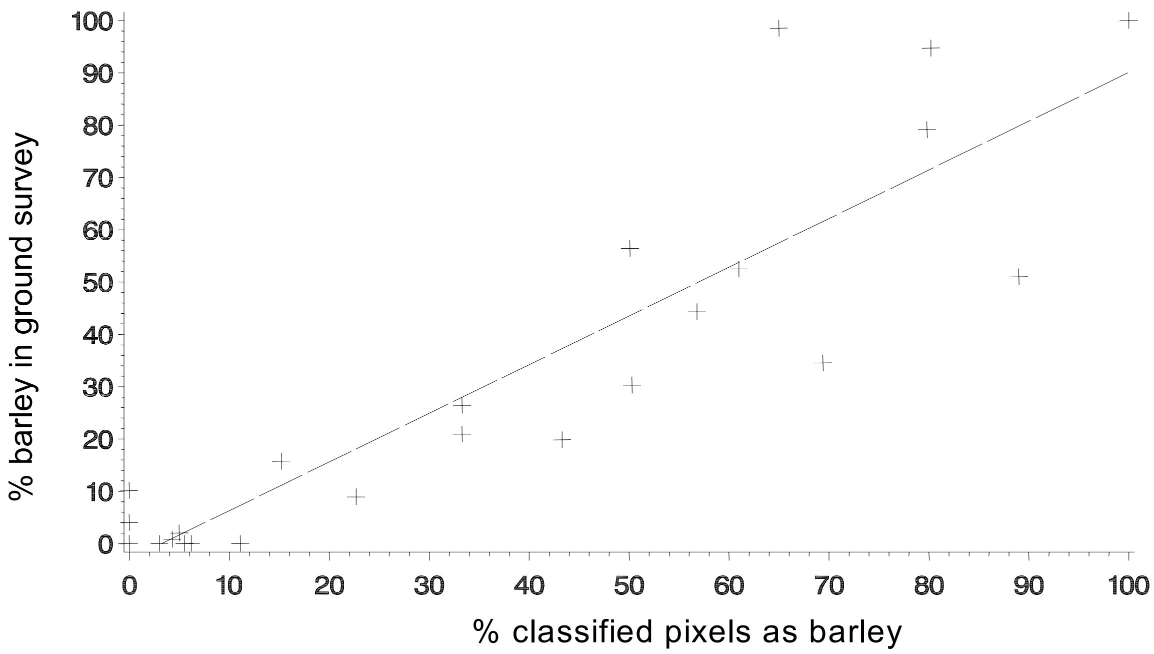
Regression estimator is not always reliable



n = 39 but unreliable regression (maximum Belsley's β = 4.7)

⇒ use tools to detect influential observations

Regression Estimator



n = 24 but reliable regression

(maximum Belsley's β = 0.8)

Caution!!!!

X must be the same variable in the sample and outside the sample

- Use all pixels (including mixed pixels) to compute X on the sample
- Do not use the same sample for training pixels and for regression,
  - or at least use a classification with a similar behaviour for training and test pixels (few parameters to estimate)

If this is not respected, regression estimator can degrade the ground survey estimates

Practical obstacles for operational use of remote sensing

In the 80's-early 90's: cost efficiency was insufficient

- Cost of images
- Cost/time of image processing.
- In the late 90's RS area estimation became nearly cost-efficient with Landsat TM, but…. no continuity of the mission.

Timeliness: 1-2 months after ground survey estimates

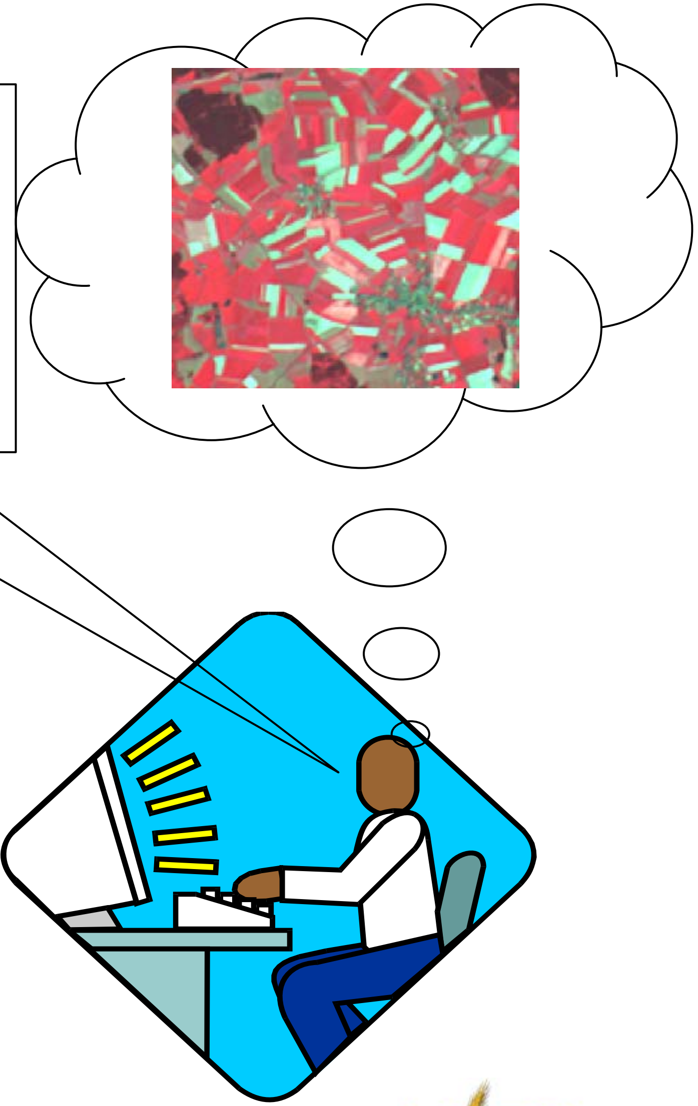
Autonomy of official organisations.

Currently new image types need to be better assessed (e.g: DMCII)

Remote sensing over-marketing

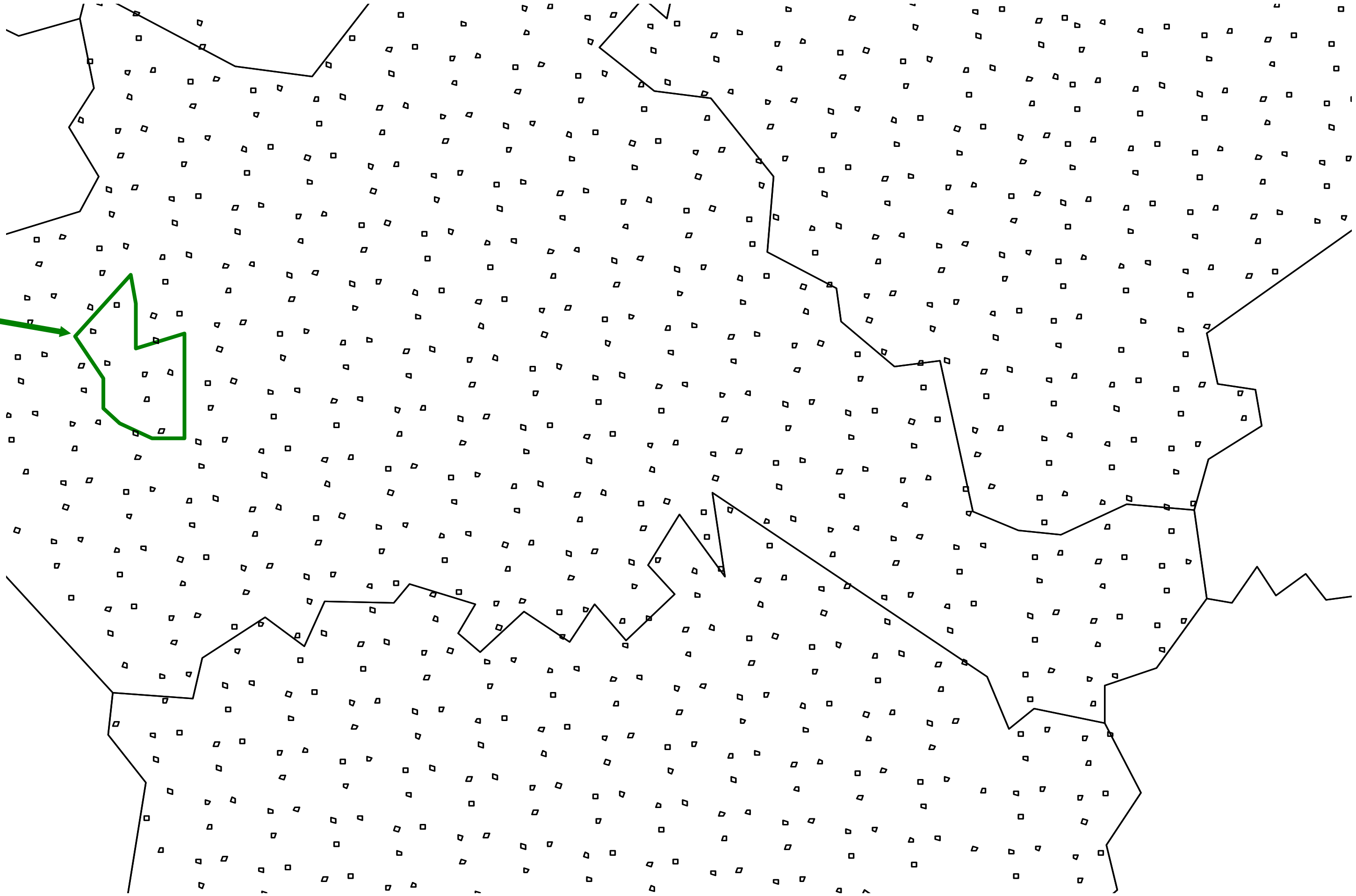
We have the solution.

Which is your problem?



Small area estimators

Small area



Small area estimators use

- The sample inside the area (possibly n=0)
- A covariable inside the area (classified satellite image)
- The link between variable and covariable outside the area.

Small area estimators are model-dependent

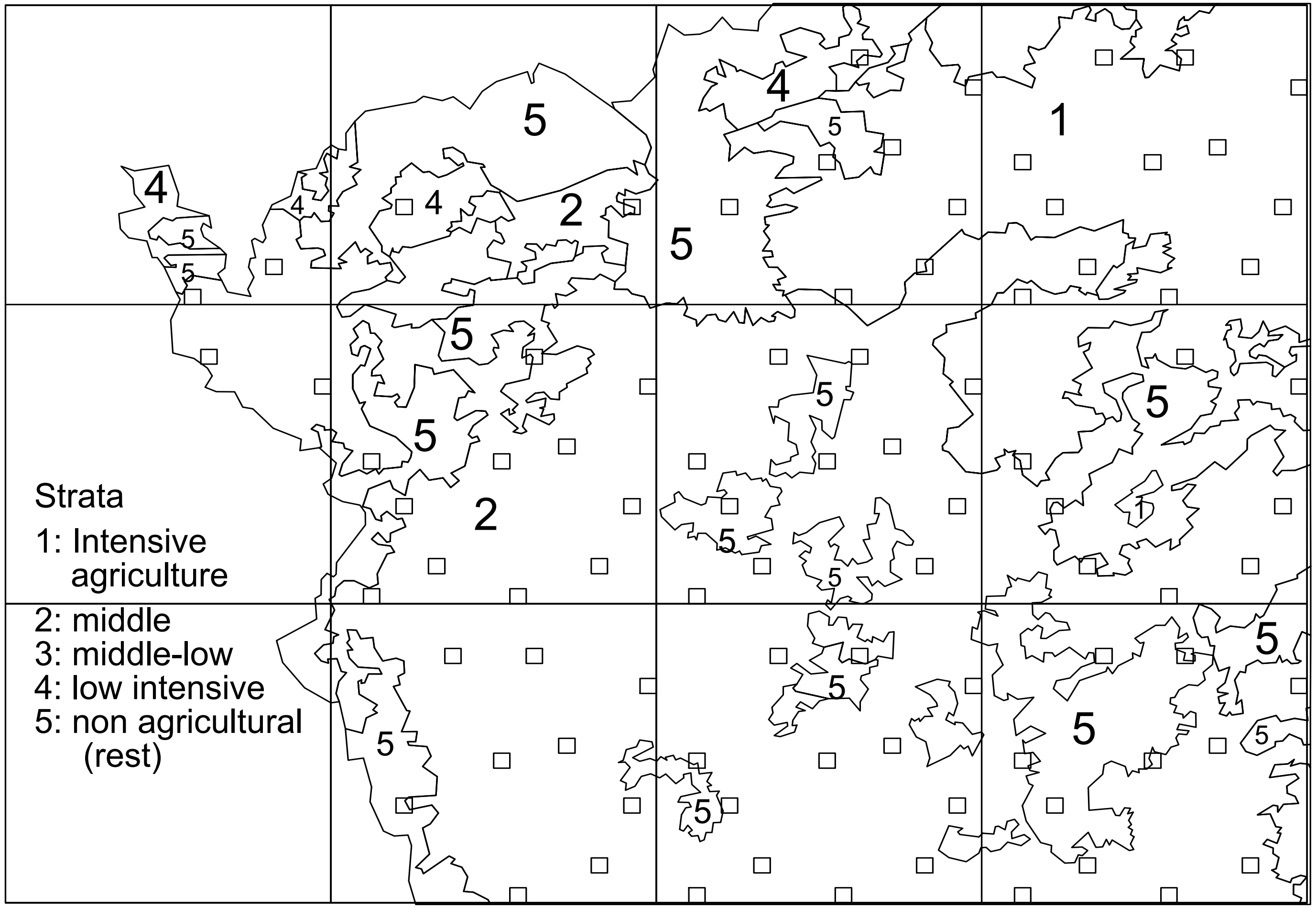
Improving an area sampling frame with satellite images

Stratification: strata defined by an indicative land cover pattern

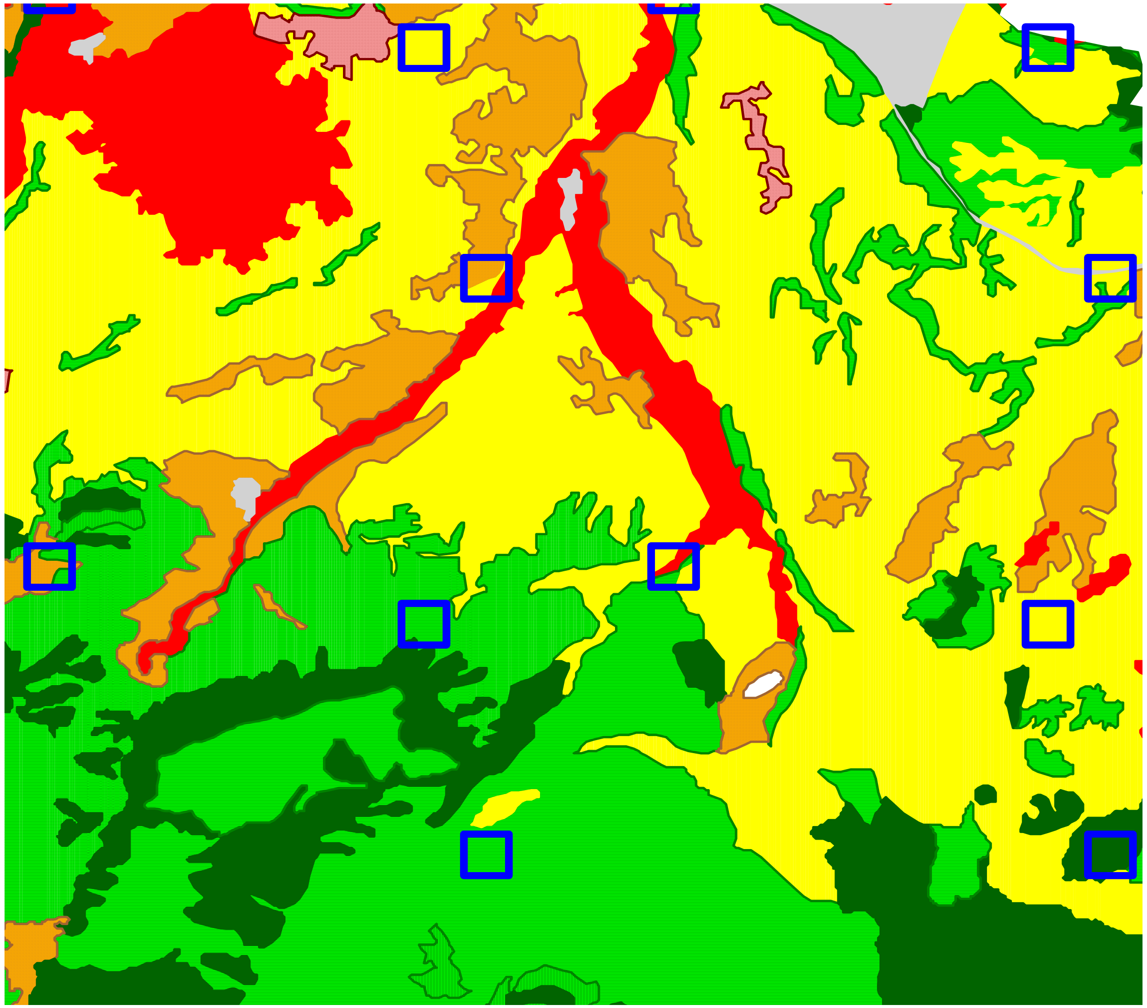
Two-phase sampling: large random or systematic pre-sample and subsampling with unequal probability.

Stratification and two-phase (double) sampling efficiency is generally moderate (often between 1.5 and 2) but the operation is not too expensive and is valid for several years.

Square segments and crop-intensity stratification



CORINE Land Cover as stratification

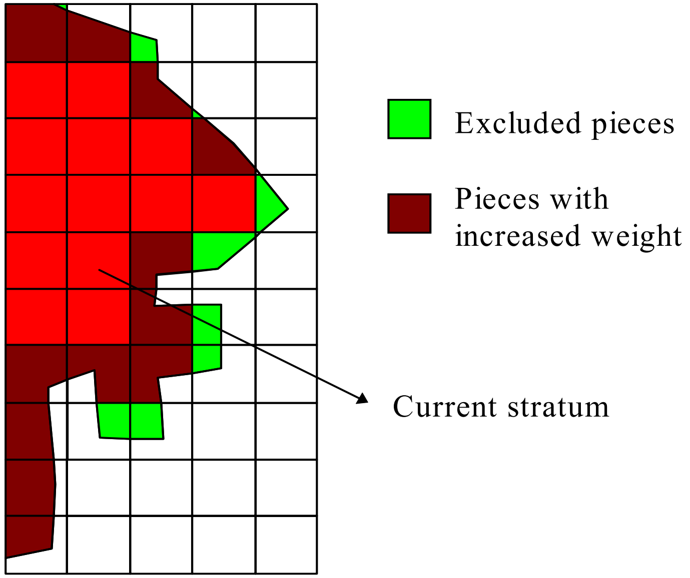


Area sampling frame of square segments

Segments can straddle strata limits

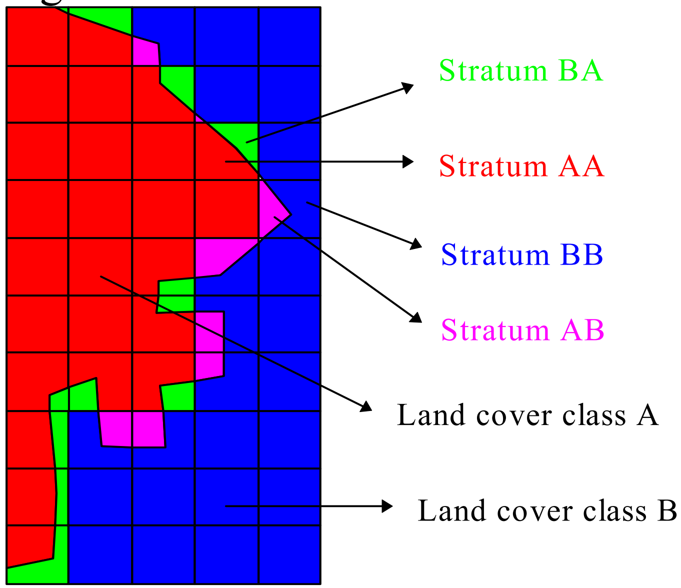
Largest piece strategy

Small pieces are excluded from estimation. Large pieces receive the same weight as a full square grid cell, and compensate in some way.



Splitting strategy

Squares sampled with the probability corresponding to the largest piece.
Stratum BA={pieces in B sampled with the probability corresponding to A}
➔ large number of strata for estimation



Other options, e.g: Attribute each square to a stratum with an “agricultural abundance indicator

Not recommended

Efficiency of stratification

How much did we gain with the stratification?

$$Eff_{str} = \frac{V_{nostr}}{V_{str}}$$

V_{nostr} Variance that we would have got with the same sample size without stratification.

But we do not have such a sample….

For stratified random sampling:

$$V_{nostr}(\bar{y}) = \frac{N-n}{n(N-1)}\left\{Var(\hat{y}_{st}) + \left(\sum_h \frac{N_h}{n_h}\sum_{hi=1}^{n_h} y_{hi}^2\right) - \hat{y}_{st}^2\right\}$$

Do not use:

$$V_{nostr} \neq V_0 = \left(1-\frac{n}{N}\right)\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$$

Using images for incomplete stratification in a two-phase sampling

- **Reminder of LUCAS (Land Use/Cover Area-frame Survey)**
- **Area frame of points (each point is a circel of 3 m.)**
- **Points are unclustered: single stage sampling.**
- **Two-phase sampling:**
  - First phase: systematic sample with 2 km step
  - Stratification by photo-interpretation of the pre-sample
  - Subsampling with different rates for each stratum
- **Observation of the points on the ground (GPS monitoring)**
- **Digital pictures from each point (landscape database)**

Substituting ground data with remote sensing data

- **When a proper ground survey is not possible**
- **Principles remain the same, with**
  - A sample of HR-VHR images instead of the ground data (<10 m?)
  - A wall-to-wall (complete as much as possible) cover of medium resolution images (TM for example)
- **Differences:**
  - The sampling plan (size of PSUs) has to take into account the size of HR/VHR images.
  - The main non-sampling error (commission/omission errors) needs to be assessed:
    - Some ground observations, approximately balanced, are better than no ground data at all
    - If no ground data at all can be collected, assess commission/omission errors in an area with similar landscape

Square segment and farm sampling by points