Crop Area Estimates Using LANDSAT
and Point Sampling Survey Data [1]

by JEAN MEYER-ROUX
1981

Statistical Reporting Service, Research Division
U.S. Department of Agriculture, Washington, D.C.

The Statistical Reporting Service of USDA has been successful combining LANDSAT and ground survey data to improve its crop area estimates in the Midwest for major crops mainly wheat, corn, and soybeans.

This research attempts to find a similar method which could be automated or semi-automated and which could be applied to the operational program in France where:

a) soils and land cover are much more diversified,
b) fields are much smaller, and
c) basic survey is a point sampling survey.

The region chosen for this study is the "department" of Indre et Loire (A "department" is a political unit which can be compared to a large county.) This region has small fields with diversified agriculture, and LANDSAT products were readily available.

1. French agriculture statistics system

The French agricultural statistics system is based on two frames. The most important one is a list frame kept up to date by an agricultural census every 10 years and large interview surveys in intervening years.

Since 1967 a point sample of the total land area has been used for specific surveys in addition to providing land cover or crop area estimates.

The list frame is more efficient for economic data and cattle, whereas the point sample tends to be more useful for crop related surveys or surveys where a farmer/operator interview is not required.

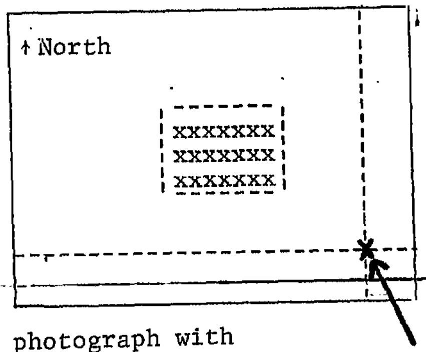
If national estimates have a high priority, almost every estimate has to be given at the "department" level.

2 Point sampling survey

Aerial photography coverage at a scale of approximately 1/10,000 forms the basis for the area frame in France. For our purposes, the most recent are often eight or ten years old.

The point sampling survey designed in France is a two stage sample. Prior to 1981 a national grid of 8000 points generated a first-stage sample of 8000 photographs. In 1981 the size of the first sample was 15,000 photographs. The second stage is a gridof points inside the photograph.

[1] Presented at the International Symposium on Remote Sensing of Environment, First Thematic Conference: Remote Sensing of Arid and Semi-Arid Lands, Cairo, Egypt, January, 1982.



The "Institute National Geographiave" provides these photographs the distance between points. In the developing process they superimpose a grid with 72 evenly spaced points on each photograph.

We enumerated the 36 points of the photograph (northern half or southern half) which are closest to the point of the national grid which generated the photograph.

Each year in June, July, enumerators visit the points and note
-the physical cover type (forage, wheat, woods, house), and
-the function of the cover type (farmland, recreational area, military use, etc.).

The point is in fact an area of basically 10 square meters, but specific instructions are given for linear cover types such as line of trees or empty areas inside a forest.

An estimate of total hectares of a cover type h, denoted y_h^1, within a target area is:

$$y_h^1 = \frac{A}{m} \sum_i \frac{n_{hi}}{36}$$

Where n_{hi} = number of points of the cover type h in the photograph i,
m = number of photographs in the target area, and
A = hectarage of the target area

An estimate of the variance of y_h^1 (using random sample formula) will be:

$$v\,(y_h^1) = \frac{A^2}{m(m-1)} \quad {}_{i} \quad \left(\frac{n_{hi}}{36} - \frac{n_{h.}}{36}\right)^2$$

where $n_{h.}$ = mean number of points of cover type h per photograph.

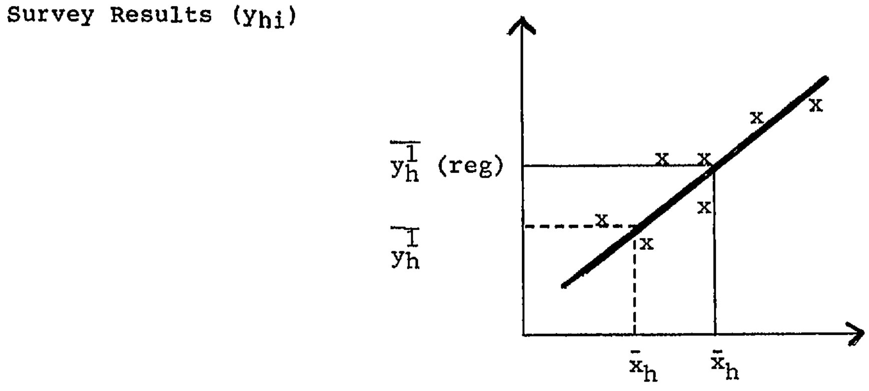
3. General Methodology for Combining ground and LANDSAT Data

We consider that the ground data from 36 points give a crop-area estimate for the part of the photo containing the points. This part will be called a segment.

If y_{hi} is the hectarage of segment i for crop h, an unbiased estimate of this hectarage is $y_{hi}^1 = A_i \frac{n_{hi}}{36}$ where A_i is the hectarage of the segment.

The classification of the pixels for sigment i gives another result x_{hi}, which is the number of pixels classified as crop h within segment i.

There is a correlation (r) between y_{hi}^1 and x_{hi} as illustrated in the following figure:



Other quantities defined in the figure are:

y^1_h = estimate of the average hectarage of crop h in the segments from the ground survey only,

b^1_h = slope (estimate) of the least square line,

$\bar{x}_h$ = average hectarage of $crop_h$ in the studied segments given by the classification of LANDSAT pixels, and

$\bar{X}_h$ = average hectarage of $crop_h$ by segment given by the classification of the complete area A.

Then Y^1_h, the estimator of total crop area used only on ground data is given by $Y^1_h = N \bar{y}^1_h$. Where N = number of A_i in the total area A or total number of segments.

An estimator which uses both ground data and LANDSAT data, called a regression estimator, to estimate total crop area is given by Y^1_h(reg) - N $\bar{y}^1_h$(reg) where $\bar{y}^1_h$(reg) = $y^1_h + b^1_h (\bar{X}_h - \bar{x}_h)$

The variance of these estimators are:

$$\text{Var } \bar{y}^1_h(\text{reg}) = (1 - r^2) \text{ Var } \bar{y}^1_h$$

$$\text{Var } y^1_h(\text{reg}) = (1 - r^2) \text{ Var } y^1_h$$

The regression estimator using both LANDSAT and ground survey data is a better estimate that the estimate based only on ground data since its variance is lower. Note that the value of r indicates the movement or efficiency of the method.

4 Methodology used in classifying pixels

The different steps of the study are the following:

-Pixels of the segments are read from the LANDSAT tape and classified using an unsupervised algorithm

-The ground truth allows us to have a correspondance between labeled points and unlabled but classifed pixels. We use this correspondence to label the groups resulting from the unsupervised classification. Some groups are ignored since they are so confused we cannot label them.

-Different statistics resulting from combining the groups and/or using a prior probability are tried. For each of the approaches a classification of segments (called small-scale classification) is performed to obtain the correlation coefficient r. The highest r gives us the best combination of the groups and prior probabilities. It gives also b_h^1 and x_h in the regression estimate formula.

-A large scale classification is then performed to obtain $\overline{X}_h$. All the elements are then available to calculate the regression estimate. The relative efficiency of the regression estimator is given by:

$$Ef = \frac{\mathrm{Var}\ y_h^1}{\mathrm{Var}\ y_h^1(\mathrm{reg})} = \frac{1}{1 - r^2}$$

A relative efficiency of two would mean that we would have to double the sample to achieve the same accuracy of the estimate without using LANDSAT.

If we include the practical work involved we have the following chronology:

-Digitization of the segments and their calibration to a map base.

-Unsupervised clustering on a small scale (segments only).

-Labeling of cluster groups. This step has been very time consuming but could be automated since the points have a defnite location within the segment.

-Creating statistics files for the identified groups. We assume that each cluster group corresponds to a multivariate normal distribution. Here we try the different approaches to creating statistics such as combining groups and addition of priors.

-Using a maximum likelihood classification, we obtain a new small-scale classification but this time with labeled pixels.

(At this level, using the ground truth, we could get an idea of the omission and commission errors in the classification, but we will choose the type of grouping with the higher r.)

-Digitization of the boundaries of the area studied, Indre et Loire.

-Large-scale classification with the best statistics file.

5 Results of the study

The northern part of the Indre et Loire appeared cloudy on the LANDSAT image. Consequently, the northern and southern parts were studied separately.

Twenty-two photographs or segments were used in the study, 8 in the north, and 14 in the south. In the south, most of the training groups were present in each of the different segments; whereas in the north the groups were really clustered in some segments. In the north, the classification accuracy seemed to depend on the location of the segment, so 8 additional segments have been digitized, classified, and the correlation coefficient calculated.

a) Unsupervised classification

In the unsupervised clustering using EDITOR we can enter as a parameter the number of clusters we want. This parameter was set at 40.

	unsupervised clusters	identified clusters	categories	labels
Northern part: 8 segments	40	24	7	7
Southern part: 14 segments	40	18	12	10

From these 40 groups approximately half of the clusters could be labeled. (See Appendices 1 and 2) In the northern part, cluster group labeling was easier than in the southern part, but having fewer points in the north made it difficult to develop statisics for as many crops (seven crops in the north instead of ten in the south).

For the 10 crops in the south, 12 groups were used --2 groups for wheat and 2 groups for forage since these crops would overlap if the different groups had been pooled and 1 group each for the remaining 8 crops.

b) Correlation coefficients

Since the correlation coefficients were very low in the northern part using the independent segments, the southeren statistics file was used there. The results were better and the categories more homogeneous over the entire area.

	Correlation coeffi: R			Efficiency		
	Wheat	Corn	Woods	Wheat	Corn	Woods
Northern part 8 trained segments Northern stat.file	.91	.89	.94	5.9	4.8	8.3
Northern part 8 independent seg. Northern stat.file	.23	.07	.60	1.1	1.0	1.6
Northern part 8 independent seg. Southern stat.file	.43	.48	.61	1.2	1.3	1.6
Southern part 14 segments	.51	.76	.94	1.4	2.4	8.3
Total 22 segments (14+8 indep)	.44	.63	.80	1.23	1.67	2.78

For the 10 land cover categories the results are the following:

land cover	Wheat	Barley	Corn	Forage	Other ag.	Hardw.	Conifers	Other woods	Urban	Water
R	.44	.45	.63	.25	.57	.42	.81	.38	.34	.93
Efficiency	1.2	1.2	1.7	1.1	1.6	1.2	2.9	1.2	1.1	8.1

c) Correlation estimates for wheat, corn and woods

$$\bar{y}^{1}_{h(reg)} = \bar{y}^{1}_{h} + b^{1}_{h} \ (\bar{X}_{h} - x_{h})$$

The classification results and ground truth in Appendix 3 are given in percent of the segment.

h	b_h^1	$\bar{y}_h^1$	$\bar{x}_h$	$\bar{X}_h$	$y_{h(reg)}^1$
wheat	.50	14.4	21.8	23.4	15.2
corn	.72	11.1	7.4	8.9	12.2
woods	.67	27.7	30.8	27.7	25.6

d) Observations

The relative efficiencies were very low, especially for wheat, which was one of the main items to study. The differences of the results between the north and the south can be attributed mainly to clouds in the north. The least squares line shows that the classification overestimates a land cover when its area proportion is high, underestimates when it is low.

6. Characteristics of the method

In order to determine the reason for the relative failure of the method we will examine the difficulties met during the study.

- Small number of training pixels. If we want to use the pixels located at the 36 observation points only 36 pixels are labeled. By comparison the segment which contains the points has about 650 pixels. This can be especially limiting if we want to observe minor crops or land cover categories instead of working only with major crops.

It is not a problem if we want to use the method to increase the accuracy obtained with something like a hundred photographs, but is was a problem in this study where we wanted to achieve the same quality as the actual survey with a much more limited number of photographs.

We have to keep in mind that to have satisfactory and stable statistics---means, variances and covariances--we need at least 100 pixels (as a rule of thumb).

- If the pixels of the point are used for training then we must have a very precise registration. This is true too if we want to have a precise idea of the commission and ommission error as shown by the study of M. Lointier in Moselle who describes very accurately these steps using an image analysis system. For ten hours spent on each segment, three were needed for an accurate registration and seven for analysis.

In order to avoid these two difficulties, unsupervised clustering was performed using all the pixels of the segments and no training pixels were used. The pixels related to the ground truth points were used only for labeling. Consequently, good registration of the segments is not necessary for the first step. However, to label the groups it is important. The registration errors will be mixed with the ommissin and commission errors even with well defined groups. It is thus not necessary to have a perfect registration since there is an uncertainity of 10% to 30% in the labeling of the groups due to percent incorrect classification.

The unsupervised clustering gives a solution which solves two main difficulties of the point sampling method, but becomes highly sensitive to another problem the mixed pixels.

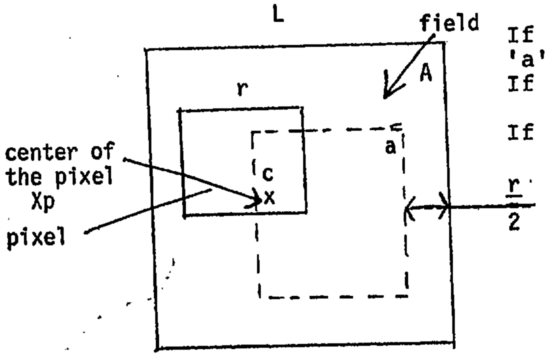
By isolating pixels that should be wheat (even if mixed together) we will get with certainly, groups of almost pure wheat. But if we let the pixels group each other we get normal distribution of pixels where the peaks are mixed pixels and the tails pure pixel of different categories. In unsupervised clustering using the NASA-developed Classy program we don't insert the number of clusters, rather it is one of the results, which should be generally an advantage. For this type of study it was not, since only a very small number of large clusters were obtained. These were mixed and could not be labelled.

Consequently, this method becomes very sensitive to mixed pixels, which is a problem related to the resolution of the satellite sensor and the size of the fields.

7. Probability of getting pure pixels.

With a sensor of resolution r what is the probability of obtaining a pure pixel as a function of field size?

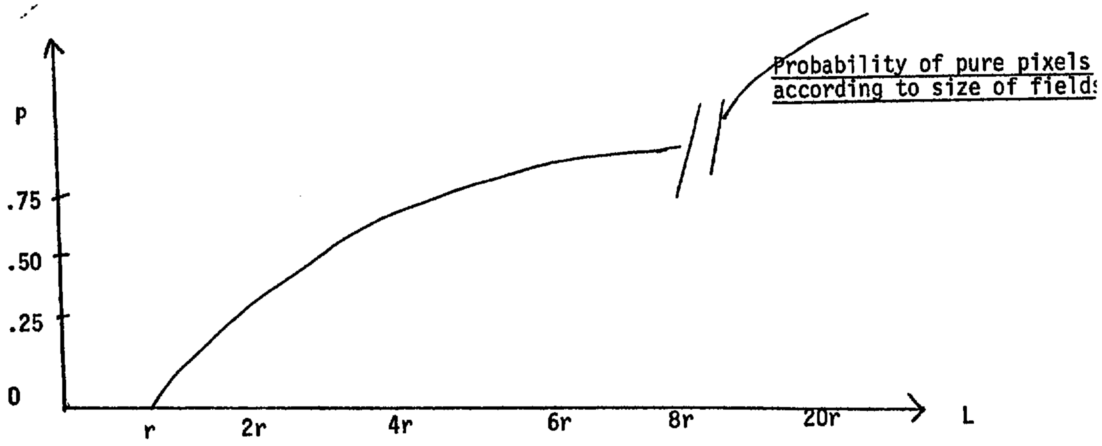
- A pixel will be defined pure if it is completely within the field and mixed otherwise.
- To simplify calculations, fields are considered square, of the same averae size (length = L), and the sensor has a well defined way of scanning the field as shown in the following figure.



If the center C of the pixel is inside 'a' then the pixel is pure: event X_p

If L : r $P(X_p/C \text{ in } A) = 0$

If L ; r $P(X_p/C \text{ in } A) = \frac{a}{A} = \frac{(L-r)^2}{L^2}$



Probability of pure pixels according to size of fields

Probab.	0	.25	.50	.75	.90
r					
80 m	.64	2.56	7.40	40.96	2.56
30 m	.09	.36	1.04	5.76	36
20 m	.04	.16	.64	2.56	16
10 m	.01	.04	.16	.64	4

Related area(hectares) of fields to accuracy of and probability of pure pixels.

The average field size in Indre et Loire appears to be between three and four hectares which means that we can expect at least two out of three pixels to be mixed. This problem would impair any kind of training method, with unsupervised clustering being much more sensitive than other methods.

8. Limit of the method in reducing the variance of estimators.

The total variance of an estimator in a two-stage, point sampling survey as described can be divided as such: variance between photographs, V_B, and variance within or inside photographs or segments, V_I.

In the methodology used, a perfect classification can nullify V_B but it cannot reduce V_I.

In the methodological study "Statistique agricole no. 104," written by M. Fournier these two types of variance are given for different areas. V_I is only 10% to 25% of the total variance, but it can vary a great deal depending on the land cover types and the aggregation (national, "department", etc).

Another way of showing the limit of the method in reducing the variance is to show the relationships between R , the LANDSAT correlation for a complete survey of the segment; with R_{36}, the LANDSAT correlation for the survey of 36 points (determining Y_{36}, the segment estimate for the given crop).

The classification results X are not affected.

$$R^2 = - \frac{\mathrm{cov}^2(Y\ ,X)}{\mathrm{Var}(Y\ )\ \mathrm{Var}(X)} \qquad R^2_{36} = \frac{\mathrm{cov}^2(Y_{36},X)}{\mathrm{Var}(Y_{36})\ \mathrm{Var}(X)}$$

The large changes are in Var (Y) and Var (Y_{36}). The covariance should be rather similar.

If we assume $\mathrm{cov}^2\ (Y\ ,X) = \mathrm{cov}^2(Y_{36},X)$, then

$$R^2 = R^2_{36}\ \frac{\mathrm{Var}(Y_{36})\ \times\ \mathrm{Var}(X)}{\mathrm{Var}(Y\ )\ \times\ \mathrm{Var}(X)}$$

In fact V is V between photographs V_B and $V36 = V_T = V_B + V_I$

If we take $V_B = 0.80\ V_T$, a frequent case as suggested by the document already mentioned, then

$$R^2 = R^2_{36} \quad \frac{1}{.80} \quad = \quad 1.25 \quad R^2_{36}$$

The best R^2_{36} we can get is .80 and since in other studies where only R^2 is used, the results are frequently around .80 and the R_{36} would be .64.

R^2	R^2_{36}	Ef	Ef36
1	0.80		5
0.80	0.64	5	2.8
0.50	0.40	2	1.7
0.25	0.20	1.33	1.25

The variances between and within segments were not computed in this study, but because of the low efficiency values this type of difficulty was not the main one. However, it would be a serious limitation of the method in a better efficiency.

Conclusions

1. For Indre et Loire and similar regions, this kind of methodology cannot be applied directly but it could perhaps be applied in regions like the Parisian Bassin where the fields are much larger. In such regions this methodology could possibly reduce the number of first stage photographs required for the survey.
2. We cannot conclude from the study that LANDSAT data cannot be applied to regions like Indre et Loire where a large number of mixed pixels are generated. A method less sensitive to that problem must be developed for this type of region. The unsupervised classification plus labeling must be eliminated by using training pixels from a different type of survey from that of strictly point sampling.
3. Some qualities of the method, like no additional ground work required and digitization process required only once would be lost but the use of a point sampling survey would limit this extra work to a minimum.
4. The new sensor planned will have better resolution. Consequently, this method can possibly be applied directly in regions like Indre et Loire in a few years.

References

1. Fournier, M. Ph. "Etude sur l'Utilisation du Territoire - Methodologie". Statistique Agricole Supplement Serie Etudes no. 104, 1972.
2. Hanuschak, G., Sigman, R., Craig, M., Ozga, M., Luebbe, R., Cook, P., Kleweno, D., Miller, C., "Crop-area Estimates from LANDSAT: Transition from Research and Development to Timely Resusts", Statistical Reporting Service, U.S. Department of Agriculture. 1979
3. Lointier, "Utilisation des Donnes d'Enquete Terut Comme Verite Terrain", Document OPIT. 1979.
4. Wigton, W.H., "Use of LANDSAT Technology by Statistical Reporting Service", Proc. Symposium on Machine Processing of Remotely Sensed Data. 1976.

Appendix 1A _ Relationship between ground truth and clusters on the point sample and labeling of the clusters (1 - 20) (Northern part)

Clusters		Wheat	Barley	Other sm grains	Corn	Annual Forrage	Perennial Forrage	Vineyards	Orchards	Other Crops	Hardwood	Conifers	Mixed Woods	Water	Urban Land	Others	Labels
1							1							2			Water
2														2	1		Water
3							1				4	1	2				Hardwood
4														1	4		Urban
5							2				3	3	6				Mixed Woods
6											1		2				"
7										1					3		Urban
8					1										5		"
9							1				2			1	1		
10	A						1				2		2				Mixed Woods
11	B	1	1		1												
12	C														4		Urban
13	D					1	2				9						Hardwoods
14	E						2				6		2				"
15	F						1				4		1				"
16	G		1												1		
17	H	1					3				5		1				
18	I				1						8		1				Hardwoods
19	J														2		Urban
20	K	6									7						

Appendix 2A **Relationship between ground truth and clusters on the point sample and labeling of the clusters (1 -20) (Southern part)**

Clusters		Wheat	Barley	Other sm grains	Corn	Annual Forrage	Perennial Forrage	Vineyards	Orchards	Other Crops	Hardwood	Conifers	Mixed Woods	Water	Urban Land	Others	Labels
1													1	8			Water
2							1					18	2	2			Conifers
3															1		Urban
4											4	8	2	3		1	Other Woods
5						1					2		1				Other Woods
6		1					1				5		3				
7		1				1					8	2	2	1			Hardwoods
8		1	1		2						12	1	1	1			Hardwoods
9				4	1					4	1				1		
10	A					1	1				8		1				Hardwoods
11	B		1		1				1	1	1		1				
12	C	4	1		1		2			3	3		2				
13	D	1	1	1	2		1			1	1		1			1	
14	E	2				2	4			1	4		4		1	1	
15	F	3	3	1	1		2				7		1				
16	G	1	1		3		1						1				
17	H	1	1	1	6					1			1				Corn
18	I	7	1		1					3	3		2		1		Wheat
19	J	1	1	1	3		1			1							
20	K	3	1		3	1	1			7				1	1		Other Ag.

Appendix 1B . **Relationship between ground truth and clusters on the point sample and labeling of the clusters (21- 40) (Northern part)**

Clusters		Wheat	Barley	Other sm grains	Corn	Annual Forrage	Perennial Forrage	Vineyards	Orchards	Other Crops	Hardwood	Conifers	Mixed Woods	Water	Urban Land	Others	Labels
21	L		1				2			2	3						
22	M	2					1				5		1			1	
23	N	1									1				1		
24	O		1	1	3	1		1		1	1				2		
25	P	3			2						4						
26	Q	1					2			1	2						
27	R	3	2					1							1		
28	S	5	1		1												Wheat
29	T				1	1					4				1		Hardwoods
30	U	5					1			1	1						Wheat
31	V	1	1	1	1	1	1			2							Other Crops
32	W	3		1		1			1		1						
33	X	2			1		2			1							
34	Y	1	1		2												
35	Z	3		1	1												Wheat
36	a	2			4				1	1							Corn
37	b									1							Other Crops
38	c				2												Corn
39	d					1	1										
40	e	2	1	1	4	2											Corn

Appendix 2B **Relationship between ground truth and clusters on the point sample and labeling of the clusters (21- 40) (Southern part)**

Clusters		Wheat	Barley	Other sm grains	Corn	Annual Forrage	Perennial Forrage	Vineyards	Orchards	Other Crops	Hardwood	Conifers	Mixed Woods	Water	Urban Land	Others	Labels
21	L	4	1			2	5			1	2		1				Forage
22	M	1		1	3		1			3							
23	N	7	3		1	1	3				5		4				
24	O	6	2	1	2	4	2		2	3							
25	P	3			1	6	3			4					1		Forage
26	Q	1	2	2	4			1		3							
27	R	1	1	4				4		1							
28	S	4	1	1	2		7				1		4				Forage
29	T	3	4		3	1	2			3	1						
30	U	1	8		1	1											
31	V	6	2	2	1	1	1			5							Wheat
32	W	2	1		1	2	2			2	1				1		
33	X	1	2			3	4			2					1		Forage
34	Y	2	3			2				1							
35	Z	6	1			4	1				1						Wheat
36a		1		1		1	1		1								
37	b		1			3							1				
38	c		1	3	2	2				4							Other Ag.
39	d					3											
40	e		2							1							