landate :	word		Landsat scene date in seconds since January 1, 1970 (in PEDITOR DATETIME format)
nparts :	word	(1..9)	Number of parts in a multi-part mask file
frame :	byte	[12]	ASCII Landsat frame name, left-justified, blank filled
	word	(0)	not used
styr :	byte	[4]	ASCII state and year for segment.Specifically, styr[1..2] is the two letter state code (i.e. postal) and styr[3..4] is the last two digits of the year. This may be all zero or all 'XXXX' if no state and year was specified.(This is propagated from other programs.)
nflds :	word	(1..255)	Number of fields

Each field has a two word descriptor of the form:
fe record

<u>IF File type is 17</u>

	byte	(0)	
cover :	byte	(0..255)	Field cover. See 'area' below
area :	half	(1..)	Area in acres. If the area exceeds 65535, the cover information will be overwritten by the ex'ension of the area acreage
tract :	byte	[2]	ASCII tract names with a range from 'A ' to 'ZZ' in the first 14 bits of the two bytes with the first character in the leftmost 7 bits and the second character in the second 7 bits.The remaining two bits are zeroes.
field :	half	(100..9999)	Field number and subfield number encoded as (field # * 100 + subfield #). Thus a value of 801 is field 8.1 as displayed.

<u>ELSE</u>

area :	word	(1..)	Area in acres.
strata number	bytes	[2]	Strata number in 2 bytes, in binary format.
count unit :	bytes	[2]	Count unit in binary format within 2 bytes and NOT factored by any value.

<u>ENDIF</u>

end		[1..nflds]	field entries

Body

The data consists of 2 byte chunks which list the number of pixels (length) and the field number for a piece of the mask file. It is stored beginning at the northernmost row of the image to the southernmost row. Each row always starts at the first column of the containing window as specified in the header. If an entire data entry is zero, the end of the row is indicated. The maximum length for any chunk is 127 pixels (7 bits).

Name	Type	Value [length]	Description
bits :	record		
blen :	byte		blen = (field length) + (boundary flag) * 128.

			Field length is 0..127, and boundary flag is 0 or 1, 1 if it is a boundary.
field :	byte	(0..255)	The field number. 0 field number is the background field. Note: if both blen and field are zero, then that bits item is taken to be the "end of a row" of bits information, and the next bits will be for the next image row. Background field will normally not have boundary flag on.
end		[until n-s + 1 "end of rows" have been found]	bits entries

STANDARD NAMING CONVENTIONS

Segment or Strata mask file (type F17)
name:{segment number}.MASK_{first five of scene-id}${second five of scene- id}_{ssyy}

SEGMENT / STRATA NETWORK (type 47 / 63)

Segment Network File is File Type 47.
Strata Network File is File Type 63.

These files are created by the SEGED module during digitization of segments or strata in counties. They contains all of the digitization information needed to create segment masks which are needed to link the JES ground truth data with the Landsat data.

FILE CONTENTS

Header

The header of the segment network file contains the following information:

Segment #
Calibration coefficients
State and year
Current and total number of segment parts
A table mapping tract and field #'s to the digitized field information

Body

The body of the file contains the following digitization information:

A table identifying the X & Y coordinates for each vertices.
A table identifying the top and bottom vertices, and the right
and left polygons for each edge.
Number of fields
Number of edges
Number of vertices
Latitude and longitude of segment center

NOTE

The strata network FILE CONTENTS are the same as for the segment network file except that segment actually refers to county, tract refers to strata, and fields to the different parts of a particular strata. The segment number actually represents the three digit FIPS county code. The information is used to create strata masks which are needed in estimation to provide average pixel counts per sample unit within strata.
In estimation, it is necessary that the tract code assigned in digitization be correctly mapped to the corresponding strata code to be used for estimation.

FILE FORMAT

Header

Name	Type	Value [length]	Description
	word	(0)	
	word	(47 or 63)	File type

	word	(0)	This word may have some use in strata files created in an image display system.
segnum	word	(0..9999)	Segment number
nlabels		(0..)	Unused for type=47 number of labels for type=63.
nflds:	word	(1..)	Number of fields
nedges:	word	(3..4000)	Number of edges.
nverts:	word	(3..4000)	Number of vertices.
cal	real	[6]	Calibration coefficients. If cal[1] > = 10**9 ,then no calibration is available. Otherwise, the xv[] and yv[] values below can be mapped to UTM values via UTMx = cal[1] * xv + cal[2] * yv + cal[3] UTMy = cal[4] * xv + cal[5] * yv + cal[6] The segment scale factor is also calculated from the cal[] values; see the code for details.
curprt:	word	(1..9)	Current part. Segments that cannot be calibrated by one cal[] (e.g., they are on different photos) are broken up into multiple parts, which are then concatenated. Specifically, a multi-part segment file is identical to the concatenation of the single part files. All fields, including the file type, are present. curprt always starts at 1, but not all parts may be in a multi-part file. Specifically, curprt is always < = maxprt, but may be < maxprt for the last part in a file. A segment multi-part file ends when curprt=maxprt,or on EOF.
maxprt:	word	(1..9)	Maximum # of parts. All parts of a multi-part file must have the same maxprt number
clat:	real	(-90..90)	Central latitude in degrees.
clon:	real	(-180..180)	Central longitude in degrees of the segment. These numbers are used to perform the inversion of the UTM coordinates when needed. Negative values mean either south of the equator or east of Greenwich.
date:	word		Date of file creation. Set to zero if not used. This entry, when used, is a standard PEDITOR DATETIME value.
digtim:	word		Time to digitize segment,in seconds. Set to zero if not used.
styr:	byte	[4]	State & year for segment, in ASCII. Specifically, styr[1..2] is the two letter state code (i.e., the postal two letter code), and styr[3..4] is the last two digits of the year.
	word		Not used.
diag1,:diag2	word		Diagonal values. Used as a check that the same corner points are used when the segment or county is put up later.

Body

The nflds entries below are attributes for each digitized field or count unit. These entries are different for file types 47 and 63.

IF filetype = 47 THEN

tract:	byte	[4][nflds] ('A' to 'ZZ')	nflds entries of 4 bytes. ASCII tract names, blank filled. Tract names are one or two alphabetic letters.
field:	word	[nflds] (100..9999)	Field number and subfield number encoded as field # * 100 + subfield #. Thus, a value of 204 is field 2.4 as displayed.
crop:	word	[nflds] (0..255)	Crop code for that field. These entries are obsolete and set to zero.

ELSE if ftype = 63 then

stata	word	[nflds] (1..99)	the strata number.
count unit	word	[nflds] (1.999)	the count unit (PSU) number.
special ndicator	word	[nflds] (1.99)	a special indicator which may be used with some count units, set to zero if not used.

ENDIF

area:	real	[nflds]	Acreage for the field or count unit. Acreages are calculated based on either a user-specified scale or the scale value determined from the calibration. Acreages can be negative if the field was digitized as a hole. This is normally a user error, but the SEGED program permits such files to be created

The entries below describe the data structure of connecting vertices, edges, and fields.

v:	word	[nedges] (1..nverts)	Top vertex of edge.

Each digitized edge is stored once, in the following form:

```
        (TV)
(LFLD)        (RFLD)
        (BV)
```

We have either YV(TV) > YV(BV) or YV(TV) = YV(BV) & (XV(TV) > XV(BV).

bv:	word	[nedges] (1..nverts)	Bottom vertex of edge.
lfld:	word	[nedges] (0..nflds)	Left field or count unit, an index into the array of fields or count units. Zero indicates background, that is outside the county or segment.
rfld:	word	[nedges] (0..nflds)	Right field or count unit,an index into the array of fields or count units. Zero indicates background, that is, outside the segment or county.
xv:	real	[nverts]	X coordinate of vertex, in inches. The range of values is usually between -100 to 100, but that is not guaranteed.
yv:	real	[nverts]	Y coordinate of vertex, in inches.

The following label data is used only for file type=63 and occurs only if nlabels is greater than zero. This label data is only necessary to use a plotting program at the Fairfax office and is not used by any MARS-PED program.

label field:	word	[nlabels] (1..)	The count unit for the label, an index into the array of count units.
label X:	real	[nlabels]	The X-coordinate of the label.
label Y:	real	[nlabels]	The Y-coordinate of the label.

STANDARD NAMING CONVENTIONS

Segment:

name:{segnum}.SEG_{ssyy}

SEGMENT CATALOG FILE (type 49)

The segment catalog file contains information on characteristics of agricultural segments so that groups of segments having common characteristics may be easily accessed by programs. The information contained for a segment consists of the county, land use strata, Landsat frame ids, map type, quad, expansion and adjustment factors, and analysis district.
When a segment is entered in the catalog, some (or all) of the information may be omitted, then added or updated later.
The file contains information about sample characteristics of area frame segments included in the current survey. The file is used extensively during many phases of the project to group segments by county, analysis districts, or Landsat Frame. The basic sample information is obtained from the area record, the map information is generated and the Landsat related information is entered as analysis decisions are made.
Routines to read file type 49 are contained in the library rdscf.prc. See the documentation in rdscf.ld.

FILE CONTENTS

The file consists of words (integers), reals, and bytes. The byte values are ascii characters. The constant CPW is a machine dependent value giving the number of characters per word. All ascii character strings are packed CPW chars per word, with garbage chars added if necessary to get to a full word boundary.

Header

# of segments in the file

Body

Segment number
County name
Landuse strata
Landsat frame ID
Quad map name
Map type
expansion factor
Problem segment adjustment factor
Analysis district name

FILE FORMAT

Header

A segment catalog file has a four word header followed by one or more segment entries.

Name	Type	Value [length]	Description
	word	(0)	
	word	(49)	file type code
nseg:	word	(>0)	no. of segment entries in the file
	word	(>=6)	no. of words in the file, including the header

A segment entry has a two word intro, followed by up to eight subentries. There may be no subentries, so a segment entry can be as small as two words. Some of the subentries are character strings, these are always packed four chars to a word, with garbage characters added if needed to reach a full word boundary.

Body

Name	Type	Value [length]	Description
segent	record		
	word	(>=2)	no. of words in the segment entry
	word	(0..9999)	segment i.d. number
	record		
	word	(1)	subentry type 1 = county name
nb:	word	(1..80)	no. of bytes (ascii chars) in the name
	word	[nb\CPW]	The county name. nb\CPW is integer divide rounded up, not truncated. There may be some garbage bytes following the last char of the county name
	end	[zero or one]	
	record		
	word	(2)	subentry 2 = strata number
	word	(1)	
	word	(0..99)	strata number
	end	[zero or one]	
	record		
	word	(3)	subentry 3 = Frame ids.
nb:	word	(10..80)	no of bytes in the frame id str
	word	[nb\CPW]	The frame id str. There can be several frame is in the str, seperated by commas
	end	[zero or one]	
	record		
	word	(4)	subentry 4 = quad name
nb:	word	(1..80)	no. of bytes in the quad name
	word	[nb\CPW]	The ascii chars of the name, garbage filled to full word
	end	[zero or one]	
	record		
	word	(5)	subentry 5 = map type name
nb:	word	(1..80)	no. of ascii chars in map type name
	word	[nb\CPW]	ap type name, garbage filled to a full word boundary
	end	[zero or one]	
	record		
	word	(6)	subentry 6 = expansion factor value
	word	(1)	
	real	(0..10)	expansion factor value
	end	[zero or one]	
	record		

	word	(7)	subentry 7 = adjustment factor value
	word	(1)	
	real	(0..10)	adjustment factor value
	end	[zero or one]	
	record		
	word	(8)	subentry 8 = analysis district name
nb:	word	(1..80)	no. of bytes in a.d. name
	word	[nb\CPW]	a.d. name, ascii, filled with garbage chars to full word boundary
	end	[zero or one]	
	end	[nseg segment entries]	There is no set limit to the number of segment entries.

STANDARD NAMING CONVENTIONS

name:{ssyy}.CATLG

SEGMENT SHIFTS FILE

The segment shifts file contains the raw and column shifts determined necessary after local calibration (segment shifting). If the automatic segments shifting program (ASMA) is used, this file is created automatically, otherwise it must be created manually using a editor.

FILE CONTENTS

This is an ASCII file containing the following information for each segment needing to be shifted locally.

Segment number
Row and column shifts (R.R., C.C)

STANDARD NAMING CONVENTIONS

name:{anything}.SFT

STATISTICS FILE (type 54)

Statistics files are created in clustering and contain the mean values, variance-covariance matrices, apriori probabilities, class names, crop indices, and number of points for multi-channel data for all categories within a given cover. Statistics files for each cover may be merged to create a total statistics file including all covers in an analysis district. The generated statistics may be displayed and edited using the STATED module.

FILE CONTENTS

The file consists of 3 parts: a 120 byte header, an array of statistics records with one record for each of the classes represented, and an array of words which are the crop indices for the crop names specified by the statistics records. The record size of statistics records is variable and is dependent on the number of channels of data.The record size is 2n**2+6n+32 bytes, where n is the number of channels. Real values (means, variances, covariances, and prior probabilities) are in IEEE floating point format.

Header

File Type
Number of crop indices
Number of channels of Landsat data
Number of classes or (categories)
ASCII ID information

Body

Statistics for each class
Class name
Class sequence number
Number of points in class
Number of crop names in class
Apriori probability
Mean value for each channel
Variance-covariance matrix for all classes within a cover

FILE FORMAT

Name	Type	Value[length]		Description
	word	(0)		Zero
	word	(54)		File type 54
	word	(0..)		Number of crop name indices in file (0 indicates crop indices not present)
n:	word	(1..)		Number of channels
nc:	word	(1..)		Number of classes
nb:	word	(40..)		Record size of statistics records (2n(n+3)+32 bytes)
id:	byte	[96]	(0..127)	ASCII id information (a zero byte terminates the string, actual info is limited to 95 characters)
class:	record			For each class
	byte	[8]	(0..127)	Class name, 8 characters ASCII (if less than 8 characters, then left adjusted blank filled)

	word	(0..)	Sequence number from original clustering or 0 if "created" or "pooled"
np:	word	(1..)	Number of points in class
ncr:	word	(0..)	Number of crop names in class
	word	(0..)	Pointer to first crop name index in the crop name index list.
	word		Not used
	real	(0.0..)	Apriori probability.
	real	[n] 0.0..)	Mean values for n channels
	real	[n(n+1)/2] (..)	Variance/covariance matrix stored lower triangle row-wise. See note.
	end	[1..nc]	
names:	word	[sum of ncr for all classes] (0..)	Index to list of crop names stored in the crops file. One word for each crop name.

STANDARD NAMING CONVENTIONS

name:{cover}.CFS

NOTE

The variance-covariance matrix is diagonally symmetric and only the lower half is stored. For example, the following matrix

```
C11 C21 C31 . . .
C12 C22 C32 . . .
C13 C23 C33 . . .
. . . . . . . . .
```

is stored as C11,C12,C22,C13,C23,C33, . .

STRATA NETWORK FILE (type 63)

For the FILE CONTENTS and format see File Type 47.
The strata network file is created as a segment network file using the SEGED module in EDITOR during digitization. The segment number actually represents the three digit FIPS county code. The FILE CONTENTS are the same as for the segment network file except that segment actually refers to county, tract refers to strata, and fields to the different parts of a particular strata. The information is used to create strata masks which are needed in estimation to provide average pixel counts per sample unit within strata.
In estimation, it is necessary that the tract code assigned in digitization be correctly mapped to the corresponding strata code to be used for estimation.
The FILE FORMAT is like F47
The FILE CONTENTS are exactly the same as for segment network files as described for segment network files except as noted in File Type 47.

STANDARD NAMING CONVENTIONS

name:{ssyy}.STRN

TABLE FILE (type 56)

The table file is used to store a table of pixel counts by category and cover type. Input is a packed file of raw or categorized data. If the input is uncategorized data then there is only one category and the pixel contents are not reflected in the table. The table file may be organized as one large table for all segments by using the option 'tabulation by all' or as a separate table for each segment with the option 'tabulation by segment.' A table file is created by using the Tabulate function within the Pack program. A table file may be displayed using the Print function within the Pack program.

FILE CONTENTS

Header

file type
number of segments
number of categories
type of table 0=by all, 1=by segment
frame numbers in ASCII
packing information

Body

For each segment
segment number
number of covers
For each cover within a segment
cover number
array of categories and associated pixel counts

FILE FORMAT

Header

Type	Value [length]	Description
word	(0)	file type, first word.
word	(56)	file type, second word.
word		number of segments, =1 for tabulation by all.
word		number of categories, HICAT, =1 for raw data table
word	(0..1)	type of table, 0=by all and 1=by segment.
array [1..5] of word	(0)	reserved for expansion.
packed array [1..100] of byte		frame numbers, in ASCII, separated by a space and terminated by '#'. The remainder of the area is undefined, not necessarily set to zero.
packed array [1..5980] of byte		packing information.

The first two bytes contain the number of segments, followed by that many segments, each occupying two bytes.Next comes the length in bytes of the packing code stored in two bytes. Finally comes the packing code. This is a condensed, Polish Postfix representation of the statement entered under SELECT OPTIONS.

Body

The body of the table file is divided up into entries by segment. Although the entries will typically be in numerical order, this is not required by the format of the file. The entry for each segment has a two word header:

Type	Value [length]	Description
word		segment number, set to -1 for tabulation by all.
word		number of covers for this segment. This entry may be different for different segments.

within each segment, there is one entry for each cover ,as follows

word		cover, the cover number from PEDITOR.CROPS.
array	[1..HICAT] word	each entry is the number of pixels for that category and cover.

STANDARD NAMING CONVENTIONS

name:$NB.TABLE

TOTALS FILE (type 58)

The totals file is used to store size values for segments. It contains the total acreage, by segment, for the crops of interest in the state being analyzed. The file is generated by reading all of the ground truth files one at a time and totaling the acreages in the fields. There is one list of segments for the entire file. The size values are stored in the body of the file in structures called option blocks. Each option block has the SELECT OPTIONS code used, the default use, the size type, the units, and the size for each segment.

Entries in the totals file for an analysis district must be present before small scale estimation for that analysis district can be run.

If any changes are made to the ground truth files in an analysis district after the totals files has been updated for that analysis district the totals file must be updated again.

FILE CONTENTS

The totals file consists of three parts, a small ten-word header, the list of segments, and the option blocks.

Header

file type
number of segments
number of option blocks
state and year

Body

list of segments
size type
units (acres)
default use
number of covers
cover list
sizes

FILE FORMAT

Header

The format of the header is:

Type	Value [length]	Description
word	(0)	file type, first word.
word	(58)	file type, second word.
word		unused, reserved for expansion, set to zero.
word		number of segments, NSEG.
word		number of option blocks.
array [1..4] of byte		state and year, the two-letter postal state abbreviation followed by the two low-order digits of the year.
array [1..4] of word		reserved for expansion.

The list of segments occurs once in the file and is:

array [1..NSEG] of word   the list of segments, one segment per word.

Body

The remainder of the file consists of the option blocks for as many option blocks as are specified in the header. Each option block is as follows:

Type	Value [length]	Description
word	(0..4)	size type, 0=field, 1=planted, 2=harvested, 3=abandoned, 4=digitized. Abandoned is the difference between planted and harvested. Digitized is not implemented yet.
word	(0..1)	units, 0=hectares and 1=acres.
word	(1..4)	default use, used in the SELECT OPTIONS statement where a use is required but not explicitly specified, also used to obtain the cover for a field.
word	(1..10)	number of covers stored for this option block, NCOV.
word	(0..1)	overflow indicator, = 1 if during size computation more than ten covers were found.
array [1..10] of word		cover list, first NCOV words each contain a cover index into the standard PEDITOR crops file. The contents of any remaining words is undefined, not necessarily zero.
array [1..NSEG] of real		sizes, in the same order as the list of segments.
word		length of the code in bytes, CODELEN.
packedarray [1..CODELEN] of byte		the internal representation of the SELECT OPTIONS code.

STANDARD NAMING CONVENTIONS

name:{anything}.TOT

WINDOW COORDINATES FILE

The window coordinates file contains Landsat row column coordinates for all segments contained in the Landsat scene. It can be created as soon as the PCAL-3 file is available and segment containment has been run. In video dig states, the window coordinates file is created on the PDP-11 at the same time that the B&P masks are created. For manual dig states, the window coordinates file is created by running extent on MMDS. This process is discussed in documentation 110.

FILE CONTENTS

This is an ASCII file containing the following information for each segment to be included in analysis of the scene.
Landsat row and column coordinates for NW & SE corners of the window containing segment
Segment number

STANDARD NAMING CONVENTIONS

name:{anything}.WINCOORD

WINDOW OR CATEGORIZED FILE

MULTIWINDOW FILE

Multiwindow files contain Landsat pixel data for up to 300 segment windows. The file is originally created by the TAPWIN program which extracts data windows from the original input tape of Landsat data. The size of the window can vary from one pixel to the entire Landsat scene. They may be distinct, or overlap to any extent and contain some or all of the channels of data from the original tape. Operationally windows are extracted with 15 pixel borders around the JES segment to allow for local calibration to occur. The SUBWIN program can read this file and create a new multiwindow file to reduce the border around each segment window.

CATEGORIZED FILE

Categorized files contain the category in which each pixel was classified during small scale classification. A categorized file is just a window file with one channel and category numbers for values instead of sensor reflectance values.
The output from small scale classification will always be written to the file VECTOR.BUFSSO(XXXXXXXX) where ss is the state abbreviation and xxxxxxx is assigned by the program..

FILE CONTENTS

Header

File type
FILE FORMAT, information including
Format for storing pixel data
Row sampling increment
Column sampling increment

Number of channels selected
Total # of pixels
Window type
Number of windows
North, west, south and east coordinates for each segment window in the file

Body

Multiwindow file: Pixel data by channel for each segment window in the file
Categorized file: Category numbers for each pixel classified.

FILE FORMAT

The window file has 2 sections. A 6400-byte [1600-word] header contains information about the windows contained in the file. This information applies to all windows in the file (ie, all windows have the same format, sampling increments, etc).

Header

The header is organized as follows:

Type	Value [length]	Description
word	(0)	all peditor files start with 0 in the first word
word	(51)	file type code; identifies file as a window file
word	(0..3)	pixel type: 0 = byte; 1 = half; 2 = word; 3 = real; indicates size andtype of image data
word	(0..2)	window format id: 0 = PIL (pixel-interleaved) 1 = BIL (band-interleaved) 2 = BSQ (band sequential); indicates how image pixels are arranged in file
word	(1..)	row sampling increment
word	(1..)	column sampling increment
word	(1..maxchan)	number of channels selected; maxchan is a constant defined in winlib.
word	(1..)	total number of pixels in file
word	[10]	reserved space for future expansion
word	(0..1)	window type: 0=image data;1=packed data

The next entries describe individual windows in the file

case window type of image data:

word	(1..300)	number of windows
record		
word	[4]	north,west,south and east coordinates for each window
word	[1]	segment number
end		one record for each window in file
end		image case

case window file of packed data:

word	[1501]	contents undefined at present
end packed case		
word	[20]	header information; this is descriptive

information either read from the input tape header, as with edips tapes, or supplied by the user, as with bare tape

word	[60]	left for future expansion

The header is followed by the data for each window in the file.

Body

These data are a sequential series of pixels. SUBWIN and other programs using window files must interpret the window data accoding to the information in the window file header.
The file image data are arranged as follows:

case window format of

```
PIL:
record
          record
                    pixel:item          [number of channels]
          end       (number of rows) * (number of columns)
end       # windows

BIL:
record
          record
                    record
                                   pixel:item          [number of columns]
                    end            [number of channels]
          end        number of rows
end       number of windows

BSQ:
record
          record
                    record
                                   pixel:item          [number of columns]
                    end            [number of rows]
          end       [number of channels]
end       [number of windows]

end of file
```

STANDARD NAMING CONVENTIONS

Multiwindow file:

name:{anything}.MWN

Categorized file

name:{anything}.CAT

VOCABULARY

Area frame: frame made out of surface elements.

Clustering (unsupervised classification): partition into a set of classes that are defined by the algorithm itself.

Confusion matrix: contingency table giving, for a particular set of pixels, the number of pixels, the number of pixels of class c (ground truth) classified in to clss c'.

Discriminant analysis (supervised classification): partition into a previously defined set of labelled classes.

Frame: list of units of the investigated population. Each unit belongs usually to one stratum.

Segment: each of the units of an area frame. Often referred only to the units in the surveyed sample. Squared segments are sometimes called "quadrats".

Spectral signature of a land use: probability of frequency distribution of the radiometric values of pixels corresponding to this land use.

ANNEX 1

CROP AREA ESTIMATION THROUGH AREA FRAME SAMPLING AND REMOTE SENSING

J. Gallego, J. Delincé
MARS Project, Institute for Remote Sensing Applications
JRC, 21020 Ispra (Varese) Italy

0. REMOTE SENSING AND AREA ESTIMATION

There are at least two main steps in which remote sensing can play a role for area estimation: stratification and pixel classification to compute the estimates themselves. The cartographic representation of land use classification can be an important product of this kind of study, but our purpose now is to present the way of using remote sensing to improve the estimates obtained through an area frame sampling. We shall consider here an unstratified population. Stratification problems are presented in another chapter of this course.

1. AREA FRAME SAMPLING

The standard way to compute unbiased area estimates is a sampling survey on a frame (list frame or area frame). We shall consider here area frames (their units are called here "segments") rather than frames made out of farms or other units (Meyer-Roux, 1990).

The classical expansion formulae gives an unbiased estimate for the area Z_c of a land use c or for its proportion $Y_c=Z_c/D$, where D is the total area of the region. The formulae:

$$\bar{y}_c = \frac{1}{n} \sum_{i=1}^{n} y_{ic} \qquad\qquad \dot{Z}_c = D\,\bar{y}_c \tag{1}$$

give generally better results than

$$\ddot{Z}_c = N\,\bar{z}_c = \frac{N}{n} \sum_{i=1}^{n} z_{ic} \tag{2}$$

Both are equivalent when the size of all units is exactly the same, but (1) behaves better with unequal sizes, that are often the result of inaccuracies of ground survey material (e.g. if squared units are not exactly squared since common aerial photographs are used because orthophotographies are not available). The variance of the estimates are

$$Var(\bar{y}_c) = \left(1-\frac{n}{N}\right)\frac{1}{n(n-1)}\sum_{i=1}^{n}(y_{ic}-\bar{y}_c)^2 \qquad Var(\dot{Z}_c) = D^2\, Var(\bar{y}_c) \qquad (3)$$

For crop acreage estimates in most of the regions of the EC, a appropriate size of the area frame units (segments) can be about 25 to 100 Ha., depending on the average field size and homogeneity of the land use. The trials carried out up to now suggest that, for crop area estimation, smaller segments are more convenient in difficult, mountainous areas.

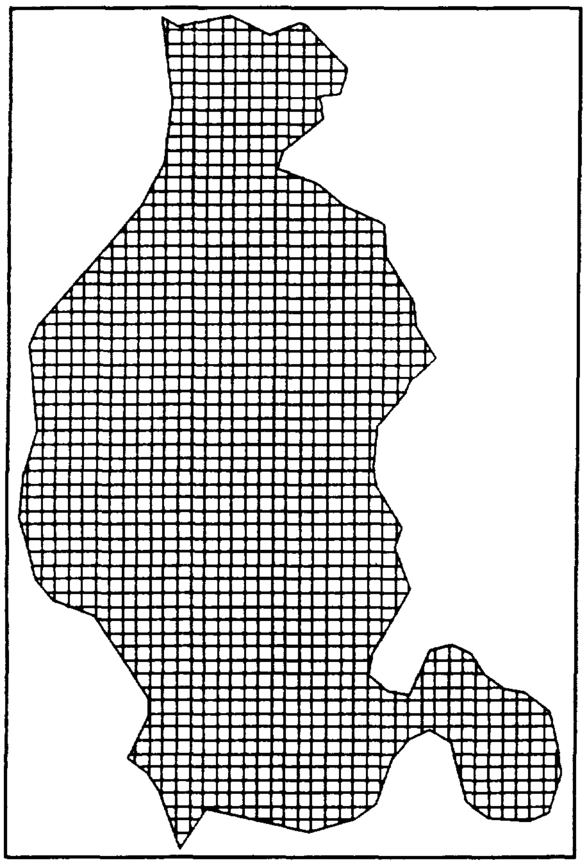

Figure 1: 1 Km. squared grid upon a region.

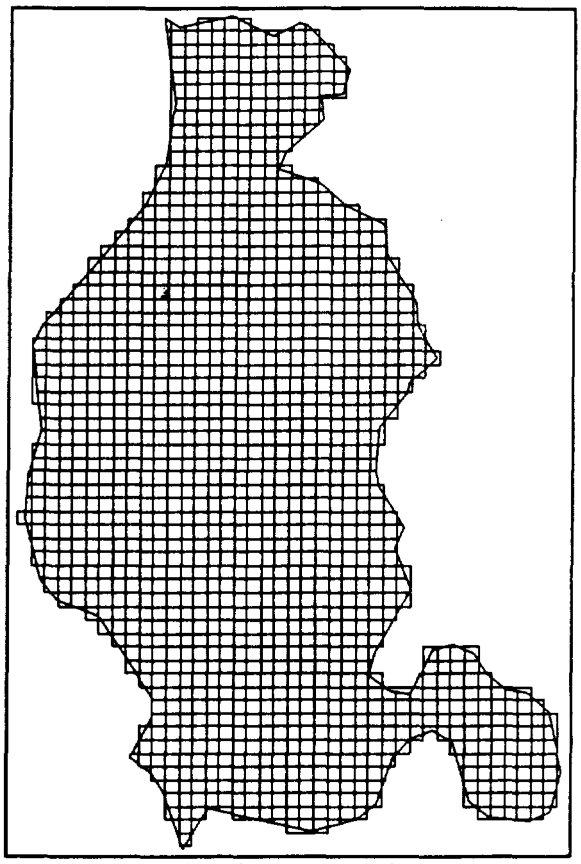

Figure 2: Making full squares in the borders.

Let us see a hypothetical example of area frame in a small region (the graphics correspond actually to the province of Varese) represented in figure 1 with a squared grid of 1 Km. segments of the frame have 100 Ha. excepting those in the border. To simplify ground survey work and computing, it can be decided that squares in the border are completely included if they have more than 50 Ha. inside the region and completely excluded otherwise (fig. 2). Hence the region studied in fact is approximated following the squared grid (fig. 3).

A random sampling is performed using blocks of 10 Km * 10 Km. In each block several segments (sampling replicates) are chosen at random (fig. 4). In our example the pattern is repeated for all the blocks, which is an easier solution, but different random choices for each block can be better for further data analysis.

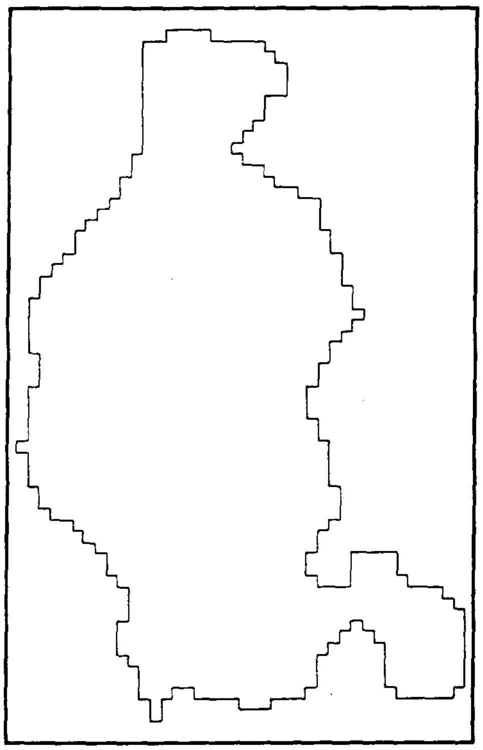

Figure 3: Surveyed region (approximation to the grid).

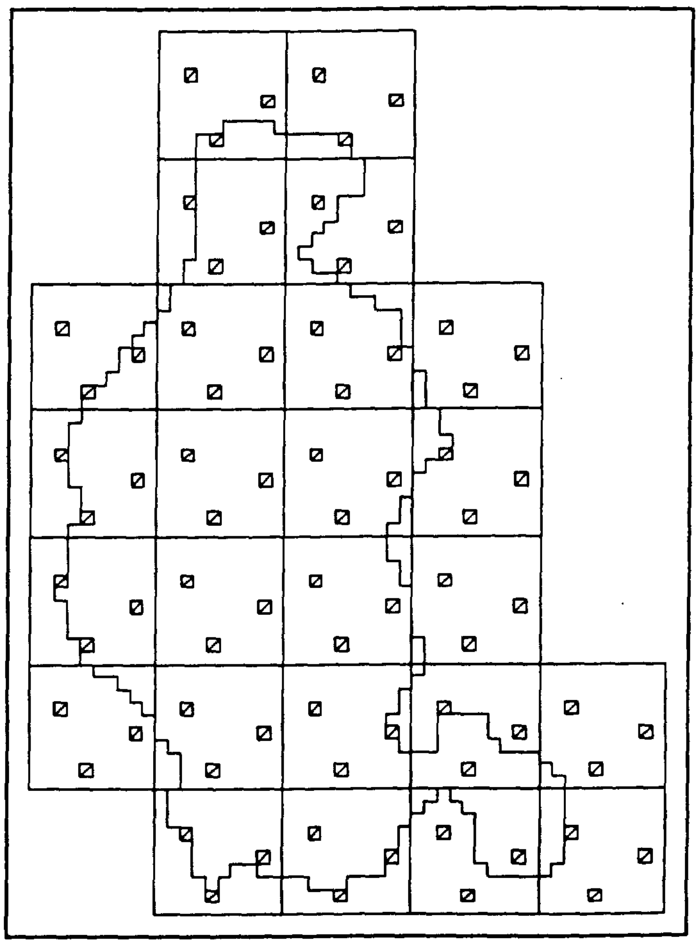

Figure 4: Systematic random sampling with three replicates in blocks of 10 Km.

2. AREA ESTIMATION USING CLASSIFIED SATELLITE IMAGES.

Satellite images give useful information for area estimation, in particular after a supervised classification of the images into the categories of a suitable nomenclature. This classification, performed most often by a maximum likelihood discriminant analysis (Anderson, 1984), leads to a thematic map. In general the nomenclature for image classification is less detailed than the nomenclature for ground survey.

Several approaches are possible to estimate areas from a classified image:

2.1. Direct Estimate from a Classified Image.

Supervised classification is sometimes used directly to estimate the area S_c covered by a land use c in a region of total area D. The estimator can be written as

$$\bar{Z}_c = \frac{x_c}{x} D \tag{4}$$

where x_c is the number of pixels classified into the land use c, and x is the total number of pixels. This estimator is very bad in general. Its properties depend very strongly on the (unknown) distribution of the radiometric values of the pixels for each land use (spectral signature). The estimates are reasonably good only if spectral signatures are very clearly discriminated. This may happen estimating the area of irrigated crops on a summer image in a very dry region. Unfortunately there is usually an important confusion between spectral signatures, and this direct estimator is strongly biased.

2.2. Global Estimate using Confusion Matrices.

If the confusion matrix A, giving the number A_{ij} of pixels of land cover LC_i in a test set classified into the land use LC_j, is an acceptable estimator of the confusion matrix Λ for the whole population (this is true if the test set is a random sample of the population without geographic relationship with the training set), A can be used for the so called global estimators (Hay, 1988, 1989; Jupp, 1989, Delincé, 1990(b)), that balance the stronger or weaker tendency of pixels to be classified into a land use.

If we call r the column vector whose elements are the unknown $r_i=A_{i.}$: number of pixels in the test set that correspond to land use LC_i (ground truth), C is the column vector with elements $C_j=A_{.j}$: number of pixels classified into class j, Pr and Pc the error matrices with the proportions $Pr_{ij}=A_{ij}/A_{i.}$ and $Pc_{ij}=A_{ij}/A_{.j}$, the following identities are straightforward:

$$r = Pc\,C \qquad\qquad C = Pr\,r \tag{5}$$

(beware that this does not mean $Pc=Pr^{-1}$).

In a parallel way, if the confusion matrix Λ for the whole population were known, we would get

$$\Theta = \Pi c\ \mu \qquad\qquad \mu = \Pi r\ \Theta \tag{6}$$

where $\mu_j=\Lambda_{.j}$ are known, $\Pi r_{ji}=\Lambda_{ij}/\Lambda_{i.}$, $\Pi c_{ij}=\Lambda_{ij}/\Lambda_{.j}$ are estimated by Pr and Pc, and $\theta_i=\Lambda_{i.}$: pixels in the whole region corresponding to a ground truth LC_i is what we want to estimate. The direct and inverse estimators are:

$$\tilde{\theta}_{dir} = Pc\ \mu \qquad and \qquad \tilde{\theta}_{inv} = Pr^{-1}\ \mu \tag{7}$$

that lead to the area estimates

$$\tilde{Z}_{dir} = \frac{\tilde{\theta}_{dir}}{Npix}\,D \qquad and \qquad \tilde{Z}_{inv} = \frac{\tilde{\theta}_{inv}}{Npix}\,D \tag{8}$$

where Npix is the total number of pixels in the region.

This technique has not yet been sufficiently tested, but may give excellent results in the near future thanks to the deeper use made of the information contained in the confusion matrix.

2.3. Regression Correction of Ground Survey Estimates.

The most usual technique to integrate a ground survey and satellite images is the Regression Estimator, (Ozga, 1977, Hanushak, 1982, Chhikara, 1986, Consorzio ITA, 1987, Allen 1988, Porchier 1990). The regression correction of survey estimates is a classic technique (Cochran 1977) to estimate the mean μ_y of a variable Y known for the n units of a sample, by correction of using an auxiliary variable X known for the N elements of the whole population and correlated with Y.

2.3.1 Single Regression.

The area estimates are obtained separately for each land use c; hence we shall drop the index c unless specification is necessary. y_i is the percentage of land use c in segment i (Ground Survey); the auxiliary variable x_i is most often the percentage of pixels classified into c, although other choices are possible.

The linear regression estimate of μ_y is:

$$\hat{y} = \bar{y} + b(\mu_x - \bar{x}) \tag{9}$$

where b is the estimated change of Y if X is increased by unity. If $b=b_0$ is fixed (independent of the actual sample in the current stratum), the regression estimator is unbiased with variance

$$V(\tilde{y}) = \frac{N-n}{N*n}\,(S_y^2 - 2b_0 S_{xy} + b_0^2 S_x^2) \qquad \textbf{(10)}$$

where S^2, and S_{xy} , variances and covariance of Y and X, can be estimated from the sample. If b is a least squares estimate on the sample, there is a bias of order 1/n and its approximate variance for large samples is:

$$V(\tilde{y}) \approx \frac{1}{n}\, S_y^2\left(1 - r_{xy}^2\right) \qquad \textbf{(11)}$$

Some points should be made concerning the choice of X:

-X must be the same variable for the segments in the sample and out of it. This means in particular that if X="% of pixels in each segment classified into class c", mixed pixels must be included, even if they have been excluded from the training set for the supervised classification.

-As it has been said, X can be another variable. For instance, if spectral discrimination between classes c_1 and c_2 is very bad, it can be decided to aggregate them into c_{12} for the supervised classification; the % of pixels in this new class can act as X in the regression corrections for Y_{c1} and Y_{c2}. Hence the regression correction can be obtained separately for both land uses c_1 and c_2.

2.3.2. Multiple Regression.

More than one variable can be used as regressor. For two regressors, the estimator would be similar to that of single regression:

$$\tilde{y} = \bar{y} + b_1\,(\mu_{x1} - \bar{x}_1) + b_2\,(\mu_{x2} - \bar{x}_2) \qquad \textbf{(12)}$$

as well as the expressions of variance and method to estimate the parameters b_1 and b_2 (Cárdenas, 78, Konijn, 74).

Multiple regression must be used very carefully, since an important bias can appear if the regressors X_1 and X_2 are strongly correlated. Such a correlation is likely to be found if both regressors are proportions of pixels in the same image classified as land use c_1 and as land use c_2.

3. THE PILOT PROJECT OF REMOTE SENSING APPLIED TO AGRICULTURAL STATISTICS OF THE EUROPEAN COMMUNITIES: REGIONAL INVENTORIES.

The main aim of this action is the economical evaluation and technical improvement of a methodology able to be used by national or regional organisms. Operational estimates are produced for a number of important annual crops in some regions of the EEC.

When this action was first implemented in 1988, an absolute priority had been given to the estimation of areas and yields at a regional level for annual crops: soft and durum wheat, barley, rapeseed and dried pulses (winter and spring crops); and sunflower, maize, cotton, tobacco, sugar beet, potatoes, rice and soya (summer crops). Attention is being shared more and more by permanent crops and the possible adaptations of the methodology to support the control of frauds concerning the subsidies given by the EC to each hectare cultivated with a particular crop, such as durum wheat.

The method can be applied as well to environmental problems for area estimates of different types of forest, area concerned by a particular plant illness etc. The main conditions for such a direct adaptation are:

- The nomenclature must be clear enough to avoid interpretation differences from one enumerator to another.
- The fields covered by each of the land uses must not be so small that most pixels are mixed (shared by several land uses).
- The size of the region must not be so small that the analysis of high resolution images is too expensive for the aimed purpose.

Topographic maps at 1:50000 or 1:25000 and a coverage of aerial photographs are important tools, but if they are not available they can be substituted with satellite images of the preceding year.

If a very large region is to be studied and the analysis of a complete coverage of high resolution images is too expensive the methodology can be adapted using a sample of images selected in a way similar to the one used in action 4 presented hereinafter (section 4).

3.1 Pilot Regions.

The study started on five pilot regions of about 20.000 Km^2 each. The selection criteria included the % of area covered by crops of interest and probability of image acquisition. Some regions have been enlarged with the collaboration of local authorities, the region Oberpfalz-Niederbayern has been dropped, and new regions have been added. The activities have been shifted to the south of the EC because of the need of more objective agricultural statistics.

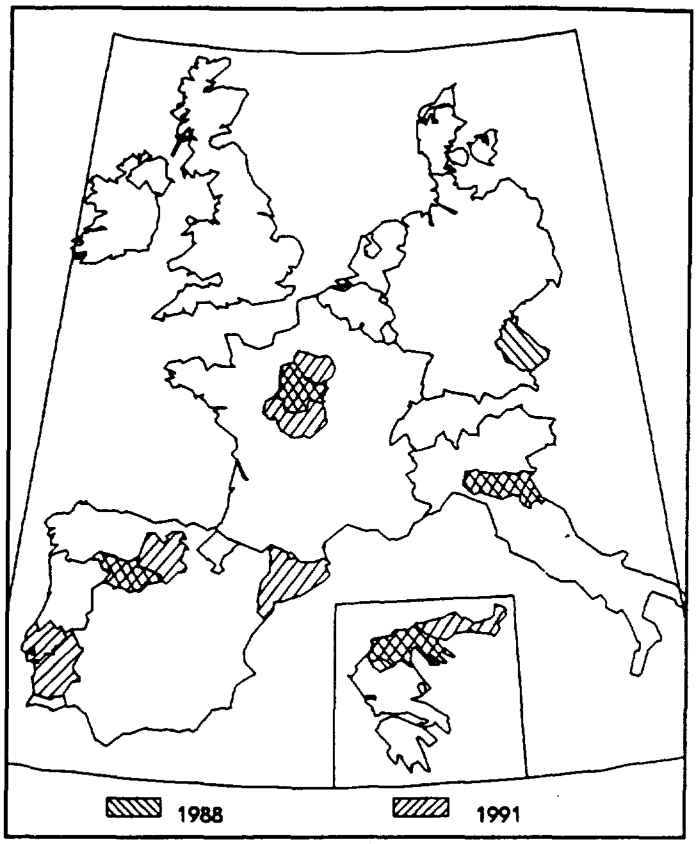

Figure 5: Regional inventories: Pilot regions in 1988 and 1991

In 1991 the next regions are being studied:
Makedonia, Région Centre and Ile de France, Emilia Romagna, Castilla y León (Valladolid, Zamora, Palencia and Burgos), Cataluña, Alentejo and Ribatejo-Oeste (Fig. 5).

3.2 Ground Survey

3.2.1 Squared and Cadastral Segments.

The ground survey is based on an Area Frame where the basic units are in most cases squared segments of 700m*700m (1Km*1Km in Alentejo, 500m*500m in Ribatejo).

In Emilia Romagna the so called "cadastral segments", with physical elements as borders, are being used. When "cadastral segments" are used, sampling is much heavier than with squared segments, since the segments should be drawn and digitized before sampling. This task is simplified defining "primary units", containing about 8-10 segments. Segments are drawn and sampled only in the primary units selected in a presampling. Some comparisons between both methods suggest that square segments give better precision (González, 1990), though ground survey errors can decrease with cadastral segments.

3.2.2. Stratification.

In most cases the number of strata ranged between 6 and 10; Each stratum is partitioned into segments. The procedure used up to now requires that each sampling unit belongs completely to one, and only one, stratum.

3.2.3. Sampling.

A random sampling was performed using blocks of segments (16*16 segments for instance). The number of segments sampled in each block (repetitions) depends on the crop intensity, for example 4 repetitions for intensively cultivated strata and 1 repetition for non-agricultural strata. The average sampling rate ranges generally between 0.7% and 1.5%.

3.2.4. Ground Survey Documents

Two kinds of basic documents are used for the ground survey: topographic maps at a scale 1:50000 or 1:25000, and aerial photographs at 1:30000 (or 1:15000) to help the enumerator to find the right position of the segment in the field, and a enlargement of aerial photograph at a scale 1:5000 for each segment of the sample over which the segment will be drawn on an overhead. Important improvements can be achieved if orthophotographic documents at 1:5000 can be used instead. Different scales can be better if very large or very small segments are used.

Other actions of the project (action 6: "Area Frame Sampling") show that these documents can be complemented or even replaced by satellite images (for example SPOT-Panchromatic with a resolution of 10m. or SPOT-XS with a resolution of 20m.).

3.2.5. Collecting Ground Observations

The segments of the sample are located and drawn on topographic maps and aerial photographs (orthophotographs if available), then visited (eventually twice). The fields in each segment are drawn on a transparency on an aerial photograph at a suitable scale (1:5000) to fit in a Din-A4 sheet, and their land use recorded (type of crop, forest, etc). About 5%-10% of the segments were visited again by supervisors to assess the quality of the ground survey data and to make some pressure on the field workers.

The area of each land use in each segment is calculated by digitizing. Digitizing can be substituted with a moderate loss of accuracy by putting a regular grid on the transparency and counting the number of points in each field.

3.2.6 Ground Survey estimates

The area of each land use in each stratum is calculated from the surveyed segments as in section 1.

An idea about the level of precision that can be reached with a ground survey on squared segments of about 50 Ha. and a sampling rate of 1% is given by the following regression lines that have been fit on the estimated areas and standard errors (S and SE in 1000 Ha) and coefficients of variation (CV in percentage) obtained in 1989 on regions about 20,000 Km^2 (Delincé, 1990(a)):

$$\begin{aligned} Log(SE) &= -0.2 + 0.56\log(S) \qquad (r^2=0.91) \\ Log(CV) &= 1.22 - 0.44\log(\%) \qquad (r^2=0.87) \end{aligned} \tag{13}$$

where logarithms are decimal and % means the percentage of the current land cover in the region. This means for example that for a crop with about 100.000 Ha. in the regio ι, a standard error of about 8.300 Ha. is to be expected.

3.3. Image Analysis and Regression Estimates.

The Regression procedure presented in section 2.3 is used to get more precise estimates than those of the Ground Survey. This estimate is nearly unbiased and reliable if it is used with some care (Gallego 1991).

3.3.1. Image acquisition

For each region a coverage with high resolution satellite images is acquired. After some tests (specially in 1988) LANDSAT-TM has been preferred to SPOT-XS for regions of this size, though the critical point is the possibility of getting cloud and haze-free images in a previously defined window of dates in which the discrimination of crops is expected to be optimal.

Each image or set of images from the same date and the same sensor is treated separately; this means that an additional stratification (neostratification) must be defined from the ground survey strata and the limits of the acquired images.

3.3.2. Geometric Correction and Resampling.

Geometric correction of images is performed using ground control points located in topographic maps and images. At the time being there is no general agreement about the best resampling procedure in this context (nearest neighbor, bicubic convolution), or the convenience of filtering images.

3.3.3. Training Pixels.

The nomenclature can be simplified grouping land uses before the supervised classification in each neostratum. The optimal way to choose training pixels is still an open question. In any case the radiometric distribution of the training pixels should be close to the distribution for the set of pixels in the whole sample of segments. Pure pixels in a random subset of segments give good results as training pixels.

Results are very bad if many training pixels have a wrong label. This can happen in particular if there is a geometric distortion in the digitized segments because they are in a hilly land or far from the center of the aerial photograph used by the enumerator. In this case pure pixels can be understood in a more restrictive way as being at some distance from any field border (considering thick borders).

3.3.4 Preclustering

Final estimates are in general significantly better if a clustering (Everitt, 1980) is performed on the set of pixels of each crop in the surveyed segments. Before the Supervised Classification a land use c is split into clusters $c_1 \ldots . c_m$. Each cluster is considered as a different land use by the maximum likelihood algorithm. The proportion X_c of pixels classified into c is computed by aggregation of $X_{c1} \ldots . X_{cm}$.

3.3.5. Regression Correction of the Ground Survey Estimates.

After supervised classification we have, for a given land use c, two different magnitudes for each segment: the proportion X_c of pixels classified into c and the proportion of area Yc observed in the ground survey. Hence the ground survey estimate can be corrected with the regression procedure presented in section 3.2.5. X_c must be computed with pure and mixed pixels since its meaning must be the same for the segments in and out of the sample, and we ignore which pixels are pure out of the sample segments. If in a given stratum Ah there are too few segments with non-zero values or some strongly influential segments with high values of Y_c or X_c, the ground survey estimate should be kept rather than accepting unreliable estimates apparently more accurate.

The ratio between the variance of the ground survey area estimate and the variance after correcting by regression is approximately $1/(1-r^2)$ in each neostratum, r being the correlation between Y_c and X_c. It is called relative efficiency of the regression. An efficiency 2 means that the same precision would have been obtained multiplying by 2 the number of segments in the sample and not using remote sensing.

3.4. Economic Evaluation: relative efficiency.

Relative efficiency of remote sensing gives a criterion for an economical evaluation of the procedure. It will be economical if the relative efficiency reaches a threshold that can be computed comparing the marginal cost of a supplementary segment in the sample and the total cost of image analysis:The value of this threshold is very different from one country to another, but ranges in general between 2 and 3 for a monotemporal study.

Relative efficiency of remote sensing is closely bound to stratification. We shall come back to this question in the chapter concerning stratification.

This efficiency threshold is not yet reached in an absolutely regular way (for all the crops of interest), but there are several reasons in favor of remote sensing even if the short-term economic evaluation gives doubtful results:

- The efficiency threshold will decrease since the image analysis procedures are becoming more automatic and computer costs are smaller.
- Relative efficiencies are being progressively improved. Improvements should be substantial with new technologies, such as microwave radar.
- Other products, (thematic maps), are given.

3.5 Mono or Multitemporal Image Coverage.

There is no doubt that multitemporal image coverage gives better crop discrimination than a single coverage; this option is to be chosen when cartographic accuracy is the central scope. If the purpose is improving the statistical accuracy of area estimations, the relative efficiency is the main criterion and it does not seem to raise enough to make up for the increase of costs, excepting in regions with very high rates of both winter and summer crops.

In Portugal and Greece trials are being made to assess a working scheme in which a multitemporal remote sensing correction is performed only every second or third year.

3.6. The choice of the satellite.

LANDSAT-TM is superior to SPOT-XS in that each scene covers a larger area (180Km.*180Km. instead of 60Km.*60Km.). Neostratification is easier and both acquiring images and computing are less expensive. TM is also superior in that it has a higher number of usable channels (6 instead of 3). A handicap of SPOT is the absence of a medium infrared channel. This superiority is partly compensated by the spatial resolution of SPOT.

The systematic comparisons carried out in 1988 suggest that the quality of the regression correction depends more on the suitability of the date of image acquisition than on the satellite. Priority is now given to TM for the regional inventories in medium size regions.

3.7. Yield and Production Survey.

In 1989 a yield survey was done in several sub-regions based on a sample of fields selected in a subsample of segments (random selections in any case). The scope was getting data on at least 100 fields for each targeted crops. The farmers were located and interviewed both before and after harvest. The test concluded that the procedure was feasible and the coefficients of variation of yields were most often significantly lower than the coefficients of variation of the area estimates (Anonymous, 1990). There are still some doubts about the possible bias in the answer given by the farmer; a comparison with objective measures is scheduled for 1990.

An exploratory study was made as well in 1989 on the correlation between the yield and the NDVI (Normalized Difference Vegetation Index):

$$NDVI = \frac{Near\ Infrared - Red}{Near\ Infrared + Red}$$

computed from the same images that had been used for area estimation; the result were rather deceiving, with values of r^2 under 0.3. The correlations were so low probably because the image dates were not adapted to this purpose. In any case it seems unlikely that we will be able to significantly improve estimates until we can integrate other information, such as meteorological data.

A different test was performed in 1990. Its aim is assessing the possibility of undertaking farm-oriented surveys with a sampling based on an area frame. The probability of selection for each farm is proportional to is agricultural land acreage excepting for the very big farms, for which a list frame approach can be chosen. This method allows other magnitudes to be estimated on the basis of an area frame, concerning cattle, manpower, financial situation, or others. The precision of the estimators achieved with this approach seems to be highly dependent on the particular farm structure in each region.

3.8 Results for area estimates

Tables 1 to 6 give some area estimates and their standard erros obtained in the pilot regions between 1988 and 1990.

France	Centre		Ile de France	
* 1000 Ha.	area	stderr	area	stderr
Soft Wheat	836.6	34.9	255.5	17.1
Durum Wheat	176.7	17.4	6.4	2.7
Barley	197.2	13.7	50.2	6.4
Dried Pulses	102.3	11.0	54.9	7.6
Rapeseed	110.2	13.9	26.9	5.1
Maize	192.3	14.7	62.2	7.4
Sugar Beets	27.7	5.6	40.7	6.8
Sunflower	241.7	18.2	31.4	5.5
Fallow	40.4	6.0	2.4	1.7

Tab. 1: Estimates in France in 1990.

Région Centre	1988		1989		1990	
(3 Dépts)	area	stderr	area	stderr	area	stderr
Wheat	537.7	11.7	566.2	15.4	553.3	15.6
Barley	99.8	6.0	87.2	7.4	110.2	8.4
Dried Pulses	55.0	3.2	72.0	6.0	66.3	3.5
Colza	57.0	5.2	42.7	4.1	27.0	2.6
Maize	148.5	6.6	143.3	8.0	125.7	7.8
Sugar Beets	22.0	3.0	24.5	4.8	29.4	5.0
Sunflower	82.9	5.3	70.2	5.6	79.7	7.1

Tab 2: Estimates for the sub-region studied en 1988-89 (Loir et Cher, Eure et Loir, Loiret)

Emilia Romagna	1988		1989		1990	
	area	stderr	area	stderr	area	stderr
Soft Wheat	203.6	9.5	209.3	9.9	212.8	12.2
Durum Wheat	63.3	6.7	41.9	5.8	46.2	6.9
Barley	42.8	3.9	37.0	3.3	43.6	4.9
Maize	92.6	8.4	69.7&	6.0	85.6	7.3
Rice	13.1	6.0	16.3	4.2	2.2	1.4
Sugar Beets	102.0	6.4	119.5	5.4	111.0	7.9
Soya	81.3	6.6	53.6	5.3	77.0	4.6
Vineyard	80.1	7.9	79.0	8.0	78.3	10.4

Tab. 3: Estimates in Emilia Romagna.

Castilla-León	4 provinces		2 provinces (Valladolid-Zamora)					
	1990		1988		1989		1990	
* 1000 Ha.	area	stderr	area	stderr	area	stderr	area	stderr
Wheat	324.1	13.4	120.0	9.2	106.9	6.9	106.3	7.1
Barley	1037.5	15.7	552.8	11.6	548.3	13.0	502.5	10.2
Dried Pulses	37.6	5.1	16.4	3.0	9.0	1.8	21.2	3.9
Maize	13.0	2.6	4.9	1.6	9.4	1.3	9.2	1.9
Potatoes	15.9	3.3	7.4	2.0	9.1	2.1	4.1	0.8
Sugar Beets	48.7	5.1	33.7	3.7	29.2	2.0	34.3	4.2
Sunflower	57.7	6.2	5.2	2.1	20.2	3.0	32.5	4.6
Fallow	393.4	12.6						

Tab 4: Estimates in Castilla-León (Valladolid-Zamora-Palencia-Burgos).

Kentriki Dytiki	1988		1989		1990	
	area	stderr	area	stderr	area	stderr
Soft Wheat	158.5	11.7	163.6	11.0	152.1	11.4
Durum Wheat	175.4	12.3	161.1	10.0	230.2	12.6
Barley	60.4	6.4	65.5	5.5	51.4	5.1
Dried Pulses	0.6	0.3	2.2	1.0	0.8	0.3
Maize	27.5	3.3	22.5	2.0	23.7	1.9
Rice	14.1	1.1	9.9	0.5	10.9	1.5
Potatoes	6.0	3.0	0.9	0.7	2.9	1.1
Sugar Beets	15.9	3.8	13.9	1.2	15.3	1.4
Sunflower	16.6	3.5	4.0	1.2	4.5	1.6
Tobacco	31.7	4.7	17.7	2.2	18.3	3.0
Cotton	73.9	6.2	60.5	4.0	46.4	6.1
Fallow			53.5	8.2	36.1	6.0

Tab. 5: Estimates in Makedonia (Kentriki-Dytiki) in thousands of Ha.

Oberpflaz	1988		1989		1990	
Niederbayern	area	stderr	area	stderr	area	stderr
Winter Wheat	153.3	4.6	151.2	4.7	153.2	3.0
Winter Barley	86.8	3.5	85.3	3.4	88.7	5.3
Spring Wheat	7.6	0.9	10.0	2.0	3.5	0.8
Spring Barley	71.0	4.5	72.3	4.6	71.9	4.0
Rapeseed	24.6	2.0	25.8	1.3	45.5	4.5
Dried Pulses	2.8	0.9	5.4	1.2	2.0	0.5
Maize	155.9	6.5	156.2	8.2	153.9	3.3
Potatoes	21.0	2.5	21.3	3.1	19.5	1.9
Sugar Beets	29.0	2.9	29.9	3.0	32.8	1.0

Tab 6: Estimates in Oberpfalz-Nierderbayern.

3.9. Some Conclusions.

The method combining a ground survey on an area frame with a regression correction using a classified high resolution coverage as auxiliary variable is operational in medium-size regions if topographic maps at 1:50000 or 1:25000 and a coverage of aerial photographs are available.

The method allows to get area estimates and their precision with a homogeneous criterion in different countries. Nevertheless some care is needed on agricultural aspects: the ground survey may have been done before some crops emerge, associated crops in the same fields require specific computing criteria, and so on.

The procedure has two separate parts, ground survey and remote sensing correction, and the former can be used without the latter. Some methodological hints can be given for applications wider than the regression correction:

- After a global geometric correction of a satellite image, visual shifting of each digitized segments is important to make ground data fit better with the image. Geometric correction through ground control points of each aerial photography is advised unless orthophotographic documents have been used.
- Using all the 6 TM channels with 30m. resolution improves the results; no channel should be dropped even if it has a high correlation with other channels. If computing time must be saved, it is preferable to subsample images (each second or third line and column), keeping full resolution only for operations on the segments of the sample.

3.10. Software for Regional Inventories: the package MARS-PED.

For the method to become operational at a large scale, an important feature is the availability of a user friendly software that can be used by an operator without a highly specialized background.

The MARS project is trying to provide such a package, written on the basis of "Peditor" by the NASS-USDA (National Agricultural Statistical Service of the U.S. Department of Agriculture). The basic modules are actually running and in the next months a user manual should be available, so that the system can be used by the national or regional organisms that wish to use the regression estimator.

MARS-PED does not need, excepting for a few modules, a specific equipment for image analysis, so that it can be used on a workstation or even on a powerful PC.

4. ACTION 4: RAPID ESTIMATES OF AREAS AND YIELDS AT THE EC LEVEL.

The main goal of this action is also estimating areas and yields of annual crops, however there are some important differences with regard to the "regional inventories":

- A regular report (4-8 issues a year) must be produced with an update of the estimates. This means that the method must be multitemporal.
- The region under study is much larger.
- The requirements about delays for estimates to be available are much harder than in the case of action 1. The final scope is that each report should use all the images acquired more than 15 days before.
- The estimates are produced on the relative changes with regard to the last year rather than on the magnitudes themselves.
- The precision requirements are weaker than in Regional Inventories. A qualitative indication such as "high increase in area" is considered satisfactory.
- The method should be adaptable to be used in any country: ground data must be used as little as possible. In particular, no ground data from the current year should be required.

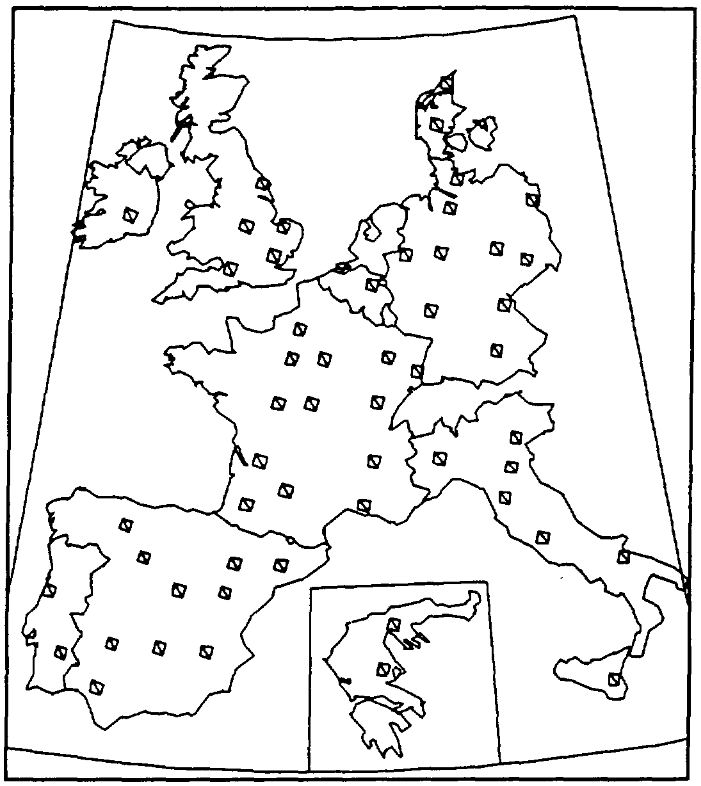

Figure 6: Rapid estimates at EEC level: sample of 53 sites

4.1. Sampling Sites in a Large Region.

When the studied region has several millions of squared kilometers, acquiring and working out a complete coverage of high resolution is extremely heavy and expensive. The alternative approach tested in this action is studying a sample of representative sites in the important agricultural regions of the EC (only regions with more than 10% of arable land have been selected).

The sample had initially 50 sites; 3 more sites have been added to include East Germany. Each site is a square of

40Km.*40Km. Their location (fig. 6) has been chosen taking into account some constraints:

- No site should straddle a national border (an exception is made in the Benelux).- Each site is wished to be included in as many nominal SPOT scenes and TM quarters of scene as possible, so that the probability of acquiring cloud free images in the wished dates is maximized. This means that the estimates from the sample of sites cannot be computed as if the sample was random, so that specific estimators are being developed using a stratification of the EC built up on the basis of NOAA-AVHRR images.

4.2. Ground Survey.

In each site a sample of about 16 squared segments of 700m*700m is drawn (eventually after a stratification within the site). A ground survey is performed in these segments, including location and interviews with farmers.

4.3. Image Analysis.

On each site 3 to 5 images are analyzed by photointerpretation of the sample of segments. No ground data from the current year are available to the photointerpreter. This will give the possibility to assess the method for its use on other countries.

ANNEX 2

STRATIFICATION FOR ACREAGE REGRESSION ESTIMATORS WITH REMOTE SENSING

J. Gallego, J. Delincé
MARS Project, Institute for Remote Sensing Applications
JRC, 21020 Ispra (Varese) Italy

1. **Stratified Ground Survey to Improve the Accuracy of Estimators.**

In a previous chapter (Gallego, Delincé, 91) we presented briefly some basic ideas on Area Frame Sampling, with an example in the case of squared units (segments), but assuming that no stratification is performed.

A stratification is a division of a population Ω of size N into non-overlapping subpopulations Ω_h of size N_h. The closer is the behaviour of the N_h elements within each stratum, the more efficient is the stratification.

Classically, the strata are defined so that no segment straddles the border between two strata. Each segment of the population belongs to one, and only one of the H strata.

1.1 Estimators and their Precision

The expansion formulae given for an estimator without stratifiaction need a slight adaptation for stratified sampling. If Z_h is the area of a land use c in the stratum Ω_h of total area D_h, and $Y_h = Z_h / D_h$, we get the estimators:

$$\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_i \qquad \dot{Z}_h = D_h \, \bar{y}_h \qquad (1)$$

where i are the sample segments in Ω_h. The variance is:

$$Var(\bar{y}_h) = \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h (n_h - 1)} \sum_{i=1}^{n_h} (y_i - \bar{y}_h)^2 \qquad (2)$$

$$Var(\dot{Z}_h) = D_h^2 \, Var(\bar{y}_h) \qquad (3)$$

that give for the whole region

$$\dot{Z} = \sum_{h=1}^{H} \dot{Z}_h \qquad\qquad \hat{y}_{st} = \frac{1}{N}\sum_{h=1}^{H} N_h \bar{y}_h \qquad (4)$$

with variance

$$Var(\dot{Z}) = \sum_{h=1}^{H} Var(\dot{Z}_h) \qquad\qquad V_{st} = Var(\hat{y}_{st}) = \frac{1}{N^2}\sum_{h=1}^{H} N_h^2 \, Var(\bar{y}_h) \qquad (5)$$

(we are assuming here equal size segments).

We can see from (2) that the variances of the estimators do not depend on the variability of the proportion Y in the whole population, but on its variability within each stratum.

1.2 Poststratification

The formulae above assume that n_h is known before sampling. In the so called poststratification, the sample is selected for the whole population with sample size n fixed, but the size n_h for each stratum is random. The formula to compute the variance of the estimator must be corrected (Cochran, 1977):

$$V(\bar{y}_{post}) = V_{st} + \frac{1}{n^2}\sum_{h=1}^{H}\left(1-\frac{N_h}{N}\right) V_h(Y) \qquad (6)$$

where $V_h(Y)$ is the variance of Y (not of $\bar{y}$) in Ω_h.

Poststratification gives a slightly higher variance, but has the advantage of allowing a change of stratification for further years with the same sample of segments. This means that aerial photographs are already available, field workers know the location of segments, and better estimates can be obtained for the % of variation from one year to the next. The post-stratification correction in the calculation of variance is important in the case of small strata.

1.3 Classical Stratification Tools.

The most common stratification tools are topographic or thematic maps (including land use maps, geological, and pedological maps). Each stratum obtained is in general formed by one or a few relatively large polygons (continuous areas). This is the kind of stratification used in the pilot region of Valladolid-Zamora (Fig.4).

If statistical data are available for small geographical units, such as municipalities, a clustering procedure can lead to strata with a large number of small scattered pieces. The stratification of Emilia Romagna (Fig. 2) is an example of this case.

1.4 Efficiency of the Stratification

The criterion to assess the quality of a stratification is the decrease of the variances of the estimates. The efficiency of stratification is computed as a ratio

$$EFFST = \frac{V_{ran}}{V_{st}} \qquad (7)$$

where V_{ran} is the estimate of the variance that would have been obtained without stratification:

$$V_{ran}(\bar{y}) = \frac{(N-n)}{n(N-1)}\left(\frac{1}{N}\sum_{h}\frac{N_h}{n_h}\sum_{hi=1}^{n_h} y_{hi}^2 - \hat{y}_{st}^2 + Var(\hat{y}_{st})\right) \qquad (8)$$

It should be noticed that V_{ran} is not the variance that would have been computed through a straightforward estimation if the actual sample had been obtained through a random sampling:

$$V_{ran} \neq V_0 = \left(1-\frac{n}{N}\right)\frac{1}{n(n-1)}\sum_{i=1}^{n}(y_i-\bar{y})^2 \qquad (9)$$

This variance V_0 is close to V_{ran} only in the case of constant intensity of sampling for all the strata, but can be very different if the allocation of the sample by strata has been done to optimise the global variance.

Tables 1.1 to 1.5 give the relative efficiencies of the ground survey stratification in five pilot regions in 1988-90. They are in general somewhat lower than expected. This means that the stratification procedures and tools used should be improved. We will refer later to some possible ways of improvement. The forthcoming procedures will be likely to use Geographic Information Systems and different kinds of satellite imagery as main tools.

Let us take the example of common wheat in 1989 in Obepfalz-Niederbayern. With a sample of 449 segments, we estimated 152118 Ha. with a standard error of 6850 Ha. (coefficient of variation: 4.5%). To get the same precision without stratification, the sample size should have been multiplied by the relative efficiency: 449 * 2.37= 1064 segments.

The efficiency of stratification can be lower than 1, this happens for instance in the same region for spring wheat. In fact this is a minor crop, concentrated in a non-agricultural stratum. The result for this crop is worse than the one we would have obtained with a simple random sample because of the low sampling intensity (0.4%) in this stratum. In general the efficiencies of stratification are less than 1 for land uses that appear mainly in strata with low sampling intensity (e.g. forest or urban areas)

Castilla-León	Stratification			Remote Sensing		
	1988	1989	1990	1988	1989	1990
Wheat	1.17	1.17	1.10	1.16	1.83	1.36
Barley	1.63	1.63	1.95	2.32	2.11	2.10
Dried Pulses	1.18	1.16	1.10	*	*	*
Maize	1.88	1.88	1.22	*	1.60	*
Potatoes	1.18	1.18	0.72	*	*	*
Sugar Beets	1.23	1.23	1.17	*	3.33	*
Sunflower	1.00	1.00	1.16	*	1.17	*
Fallow			1.18	*	*	1.76

Tab. 1.1: Relative efficiency of stratification and remote sensing in Castilla-León

Makedonia	Stratification			Remote Sensing		
	1988	1989	1990	1988	1989	1990
Soft Wheat	1.25	1.39	1.43	1.84	1.99	1.72
Durum Wheat	1.17	1.49	1.57	2.17	2.13	1.94
Barley	1.12	1.29	1.24	*	1.67	1.93
Dried Pulses	1.44	1.72	1.58	3.56	*	*
Maize	1.70	1.99	2.37	1.67	3.11	3.06
Rice	7.04	6.27	5.66	6.31	18.86	2.76
Potatoes	0.67	1.10	1.28	0.99	7.61	2.26
Sugar Beets	1.36	1.99	1.80	1.23	7.16	6.73
Sunflower	1.16	1.09	1.10	1.38	1.41	*
Soya			1.76			*
Tobacco	1.30	1.60	1.66	1.06	2.21	1.17
Cotton	2.26	2.47	2.50	1.54	3.19	1.58

Tab. 1.2: Relative efficiencies in Makedonia

France	Stratification		Remote Sensing			
	1988	1989	1988	1989	1990 RC	1990 IF
Wheat	1.68	1.74	1.75	1.80	2.56	4.46
Barley	1.07	1.10	1.17	1.03	1.10	1.40
Dried Pulses	1.20	1.21	1.77	1.53	5.14	3.47
Rapeseed	1.04	1.08	2.90	1.98	2.96	5.83
Maize	1.04	1.04	1.65	2.17	1.58	1.76
Sugar Beet	1.19	1.17	1.42	*	1.56	2.79
Sunflower	1.06	1.20	1.24	2.26	1.77	3.04

Tab 1.3: Relative efficiencies in France (RC: Région Centre. IF: Ile de France)

Bayern	Stratification			Remote Sensing		
	1988	1989	1990	1988	1989	1990
Winter Wheat	1.86	2.37	2.18	2.15	2.10	5.32
Spring Wheat	1.05	0.60	1.08	1.42	*	*
Winter Barley	1.25	1.41	1.20	1.76	2.07	*
Sprin Barley	1.16	0.89	0.73	1.10	1.73	2.31
Rapeseed	1.04	1.28	1.03	2.33	7.50	*
Dried Pulses	1.01	1.31	1.27	1.35	*	*
Maize	1.32	1.17	1.16	*	*	5.71
Potatoes	1.16	1.35	1.27	*	*	3.70
Sugar Beet	1.83	2.77	2.88	*	*	10.62

Tab 1.4: Relative efficiencies in Bayern

	Stratification			Remote Sensing		
	1988	1989	1990	1988	1989	1990
Soft Wheat	1.32	1.37	1.23	1.24	*	*
Durum Wheat	1.20	1.20	1.39	1.19	*	*
Barley	1.13	1.13	1.31	1.02	*	*
Maize	1.22	1.21	1.13	1.06	1.45	1.69
Rice	1.01	1.01	1.55	1.03	1.67	*
Sugar Beet	1.59	1.58	1.20	1.16	2.46	1.46
Soya	1.36	1.18	1.48	1.24	2.14	2.98
Vineyard	1.12	1.11	0.60	*	*	1.23

Tab 1.5: Relative efficiencies in Emilia Romagna.

1.5 Optimum Allocation of the Sample

Let us assume that we have decided that the sample will have n segments. We shall improve the estimates with a good distribution of the sample in the strata.

If we accept that the cost of surveying each segment is approximately the same in different strata, the variance V_{st} is minimised if the size n_h of the sample in each stratum Ω_h is proportional to $N_h V_h(Y)$. This rule is known as Neymann allocation.

The main problem arising to apply this rule is that we want to estimate from the same sample several crop areas rather than a single variable Y. The optimum allocation is different for each crop, and a compromise is necessary.

Another problem is found if we want to keep the same sample of segments for several years having the possibility of changing the stratification: if a part of the stratum Ω_h is to be moved to Ω'_h, both strata must have the same sampling intensity.

This is a powerful reason to keep few different values for the sampling intensity n_h/N_h rather than looking for a refined optimisation in the allocation.

The variance $V_h(Y)$ for the proportion of the area of a crop is usually higher in the strata in which this crop covers a larger area. This means that, in most agricultural regions, a reasonably good choice of allocation can have three or four levels of sampling rate: around 1.5% for the most intensively agricultural strata, around 0.5% for the little agricultural strata. For mountains or urban areas, one possible solution is to choose a very low rate (0.1%), or using the same rate as in the little agricultural strata, but surveying every fourth or fifth year. In some regions, two different sampling rates can be selected for intensively agricultural strata, making a difference between irrigated areas.

1.6 How Large Should Be The Ground Survey Strata?

The smaller the strata, the easier is to have low intra-stratum variances, and the better precision we can reach, according to formula (2), as long as the sizes of the strata are not so small that the correction (6) for poststratification becomes important or some variances cannot be computed (this happens if $n_h=1$).

The advantage of small strata can be illustrated by a trial performed on a 1988 ground survey for Emilia Romagna (sample size: 422 segments), comparing the efficiencies of 5 stratifications:

a) 8 strata based in localised production statistics and existing topographic and thematic maps.

b) 2 strata, agricultural and non agricultural, obtained by union of strata in a) with the same sampling intensity.

c) 13 strata obtained splitting b) by a square grid of 70*70 km. and grouping of too small strata.

d) 32 strata. Like c) with a 35*35 km. grid.

e) 138 strata. Like c) with a 11.2*11.2 Km. grid.

Stratifications c), d), and e) are certainly not refined, but

d) and e) give the best relative efficiencies for most of the dominant crops in the region.

	Ha.	Relative efficiency of Stratification				
		8 strata	2 strata	13 strata	32 strata	138 strata
Durum Wheat	61789	1.48	1.29	1.45	1.54	1.58*
Soft Wheat	206706	1.41	1.27	1.46*	1.45	1.39
Barley	43300	1.02*	0.95	1.01	0.97	0.94
Rice	11796	1.30	1.29	1.43	1.49*	1.40
Maize	83005	1.53	1.38	1.65	1.76	1.95*
Sugar beets	98632	2.02*	1.48	1.86	1.91	1.91
Soya	73448	1.73	1.39	2.03	2.29	2.44*
Vineyard	80708	1.39	1.34	1.51	1.51	2.23*

Table 2 : Estimated acreages and relative efficiency of 5 stratifications in Emilia Romagna

The interest of smaller strata is still to be assessed with an efficient stratification method, instead of the "blind" method used for this example. In practice the number of strata is limited by the amount of work required to deal with too many of them. This limitation might disappear or be reduced with more automatised procedures.

The point made in about the interest of a small number of different sampling intensities is equivalent to saying that pre-stratification (for sampling) should have few large strata, even if they are split to compute the estimates.

2. Stratification for Analysis of High Resolution Satellite Images

For the time being, we have worked in the Regional Inventories with the same stratification for all the tasks concerning remote sensing application for regression correction, i.e. image classification, estimation of regression parameters, and application of these parameters to compute the corrected estimates. We shall first keep this assumption and discuss later the possibility of using different stratifications for the different tasks.

2.1 Stratification and Image Coverage: Neostratification.

Image classification is performed separately in zones in which each land use has a radiometric behaviour as homogeneous as possible. That means in general for each image or set of images of the same date and same instrument, and for an agronomically homogeneous stratum.

If a ground survey stratum is not covered by a single image, it must be split (Fig.3) following a line that depends mainly on the quality of images in the overlapping area. If some pieces are too small, they can be grouped within the same image. This stratification is a poststratification, but we shall call it a neo-stratification to distinguish it from the general concept given in section 2.2.

The neostratification requires a new computation of the ground survey results: we shall have an average $\bar{y}_{neo}$ that will be slightly different from $\bar{y}$ and a new variance V_{neo} and relative efficiency $EFFST_{neo}=V_{ran}/V_{neo}$ that may be quite different to the values obtained for the original ground survey stratification.

If some areas are cloudy in the images or out of the successfully acquired scenes, a neostratum Ω_0 must be made with these areas. It is also possible to deal with cloudy areas creating a class "clouds" in the supervised classification, but this can induce a bias in the final estimates, unless it can be accepted that the distribution of clouds is uncorrelated with the land cover.

Haze and shades of clouds can be considered to build neostrata if they affect large areas, but this can be avoided if a pre-clustering is performed before the supervised classification. One or several classes LC_m that will correspond roughly to all land covers under haze or shade.

2.2 Different Levels of Neostratification.

As it has been said above, three different stratifications can be used for different tasks related with remote sensing. In the MARS Project, a single stratification is being used because it is operationally easier. The variance of the final regression estimators would be lower by using the following three levels. Research is being made to evaluate if the decrease of the variance makes up for the bigger complexity.

2.2.1 Strata for the Supervised Classification.

For the supervised classification, the criteria about the stratum size for optimal results have not been sufficiently tested, but the procedure is easier and the computation is faster with rather large strata, and can give as good results as small strata if a pre-clustering is performed.

2.2.2 Strata to Estimate Regression Parameters.

For the estimation of the regression parameter b, large strata are strongly recommended. There is some tradition in statistics of taking n=30 as the limit for a "large sample". This is reasonable for very dominant land uses, but is not enough in a situation in which a few segments have a very strong influence on the results of the regression.

Figure 1 gives some examples of situations in a stratum with a sample of 43 segments. In this example, the training pixels for image classification had been taken in 15 segments selected at random. Some hint can be got about the reliability of regression deleting the segments that contain training pixels (this should be done in general to get unbiased estimates). The figures given in the upper left corner of each scattergram are the estimates of the slope b and the squared correlation r^2 if we include or exclude the segments with training pixels.

In the cases of sugar beet and maize, one segment, lying far from the others, has a very strong influence in the estimates of regression slope b. On the other hand these parameters are heavily modified if the segments with training pixels are not used for the parameter estimation. In these cases, larger strata are necessary for reliable estimates.

The regression for wood is much more stable, a relatively important number of segments having a large proportion of wood. In the case of soya, one segment might be considered as outlier, and hence suspected as influencial, but the estimate of b is rather stable when segments containing training points are deleted in the regression. In these two cases, 43 segments seem to be enough for a reliable estimation of parameters.

2.2.3 Strata to Correct Acreage Estimates.

For the regression correction and its variance we have again similar formulae when used stratum by stratum:

$$\tilde{y}_h = \overline{y}_h + b(\mu_{xh} - \overline{x}_h) \tag{10}$$

$$V(\tilde{y}_h) = \frac{N_h - n_h}{N_h * n_h} (S_{yh}^2 - 2b_0 S_{xyh} + b_0^2 S_{xh}^2) \tag{11}$$

If $b=b_0$ is fixed, and apporximately

$$v(\tilde{y}_h) \approx \frac{1}{n_h} S_{yh}^2 \left(1 - r_{xyh}^2\right) \tag{12}$$

if b is a least squares estimate on the sample.

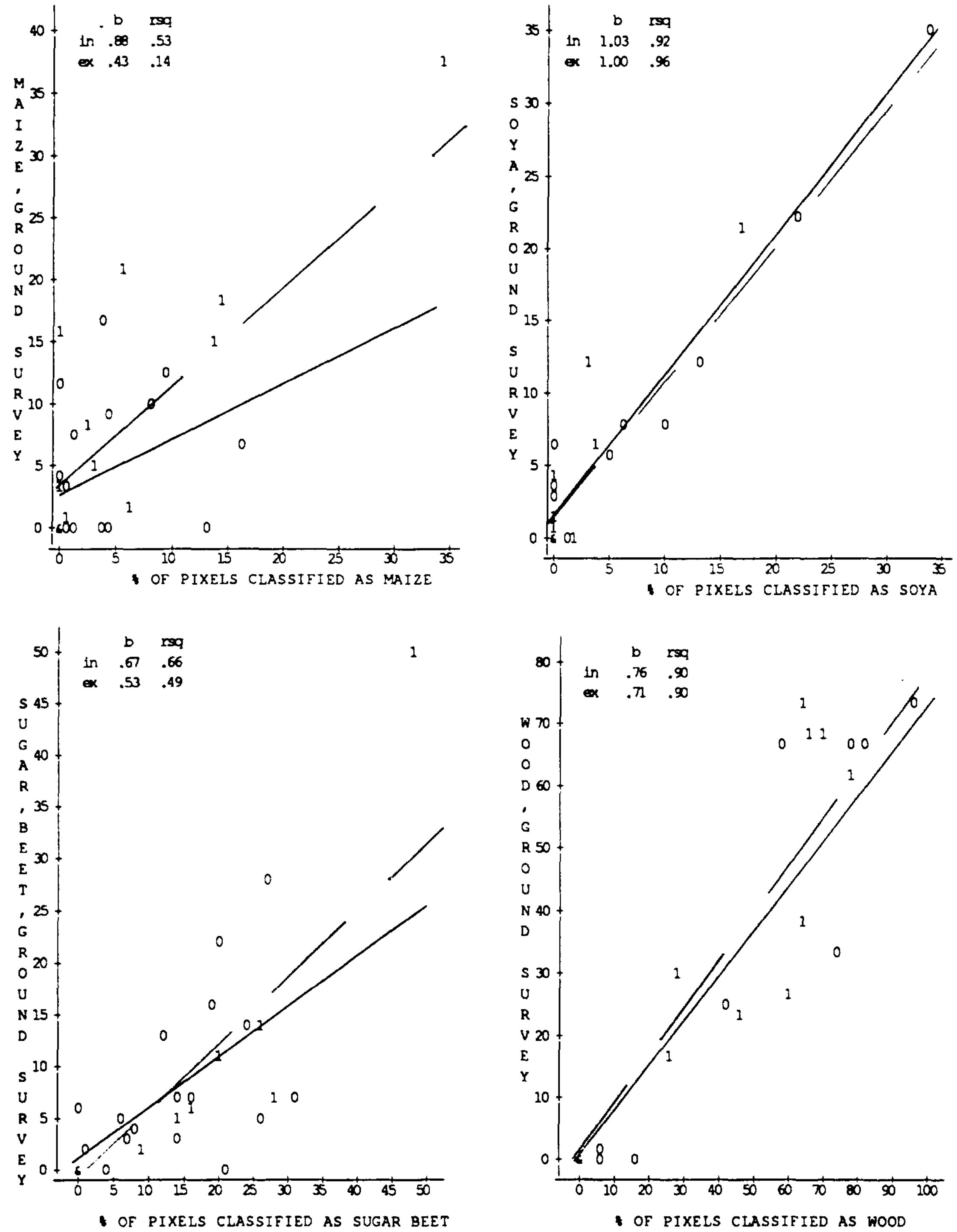

Fig. 1 : Some examples of regressions in a stratum with a sample of 43 segments. Segments with training pixels are indicated with "1". They are excluded in the "ex" version of the parameter estimate of the regression.

the variance will be lower in general if small strata are used, but in this case there is a risk of getting unreliable estimates. A reasonable solution to avoid these risk is the estimation of b by least squares in the union of several strata, and the computation of estimates separately in each of the individual strata. In this case, formula (11) seems better than formula (12), at least if the current stratum is not a very large part of the set in which b was estimated. However it is not completely correct, since b is not independent of the current stratum.

If the estimates of regression parameters are not very reliable, even after aggregation of several strata, the ground survey estimates ant their corresponding variances are to be kept for the computation of the final results. This can happen for minoritary crops concentrated in small areas. An attempt to improve the results can be done in this case through a new specific stratification, rather than using a risky regression correction.

To summarise, the neostratification should be rather detailed, but some strata should be grouped to estimate the regression slope b, and eventually for image classification.

3. Examples of Stratification in Different Countries

We give here an overview of the stratification procedures used in the five Pilot Regions of the MARS Project and the method used for an agricultural stratification in the USA. In general, the strata are drawn up by manual amalgamation of available maps, some statistical data for small administrative units, and, in some cases, satellite images.

3.1 Région Centre (France)

Three "Départements" were studied in 1988 and 1989: Eure et Loir et Cher, et Loiret. 10 strata were defined; each of them was contained in a Département, and was built up joining some of the 24 Agricultural Regions in the area of study. The aggregation was made using topographic and soil criteria, as well as average field size and percentage covered by agricultural land. Each stratum is contained in one "Département".

The area of study has been enlarged in 1990 to 51.000 Km^2, and no specific stratification has been used, the "Departements" acting as strata.

3.2 Emilia Romagna (Italy)

Three kinds of information have been used: altitude, land use map, and statistical data at the level of the Agricultural Region (48 Agricultural Regions in Emilia Romagna).

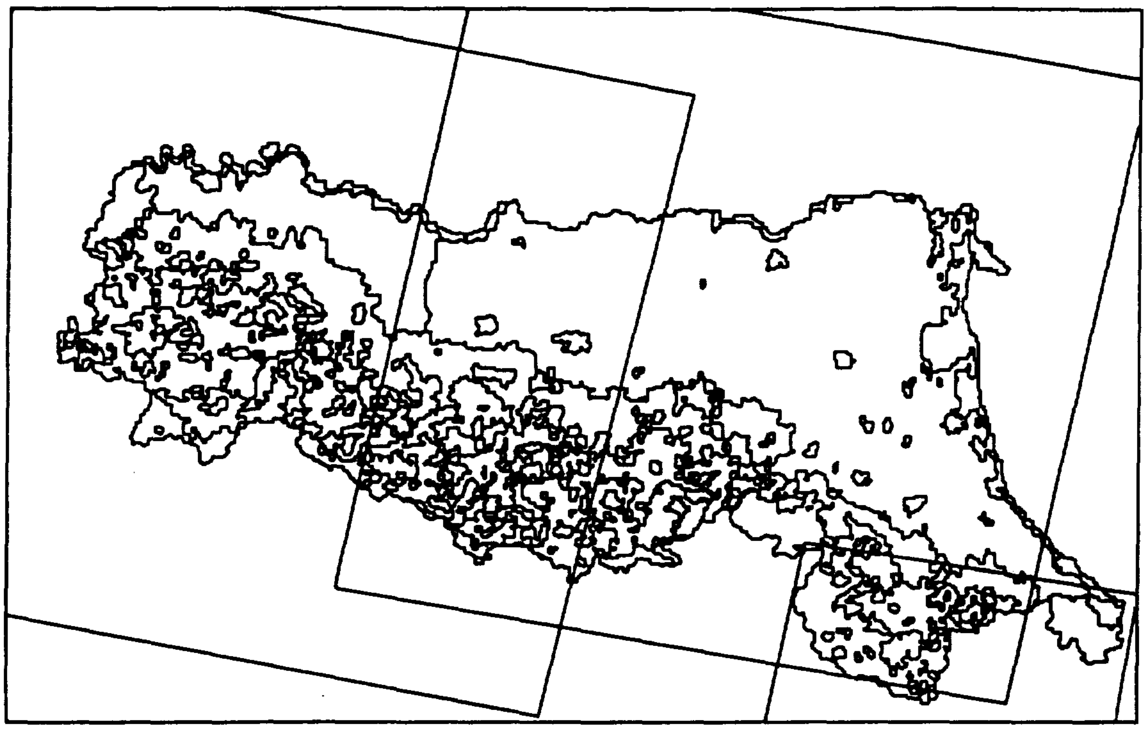

Figure 2: Emilia Romagna. Ground survey stratification for summer crops and coverage of Landsat-TM scenes.

The altitude is classified into zones of plane, hills, and mountain. The land use map had a nomenclature of 20 classes, that have been aggregated into two: agricultural land (excluding pastures) and other land uses. Statistical acreage data of common wheat, durum wheat and barley at the "Agricultural Region" level have been used for the stratification concerning winter-spring crops; maize and soya acreage data have been used for the summer crops stratification. The 341 "Comune" in Emilia Romagna are characterised by these three variables, and clustered to get 8 strata for winter-spring crops and 7 strata for summer crops (Fig.2). After splitting and regrouping, a neostratification with 8 strata is obtained for summer crops (Fig.3).

The stratification method was relatively sophisticated, however the relative efficiencies obtained were rather low.

One of the reasons can be the fact that in the same "Comune", there can be several very different areas, and this increases the within-stratum variance.

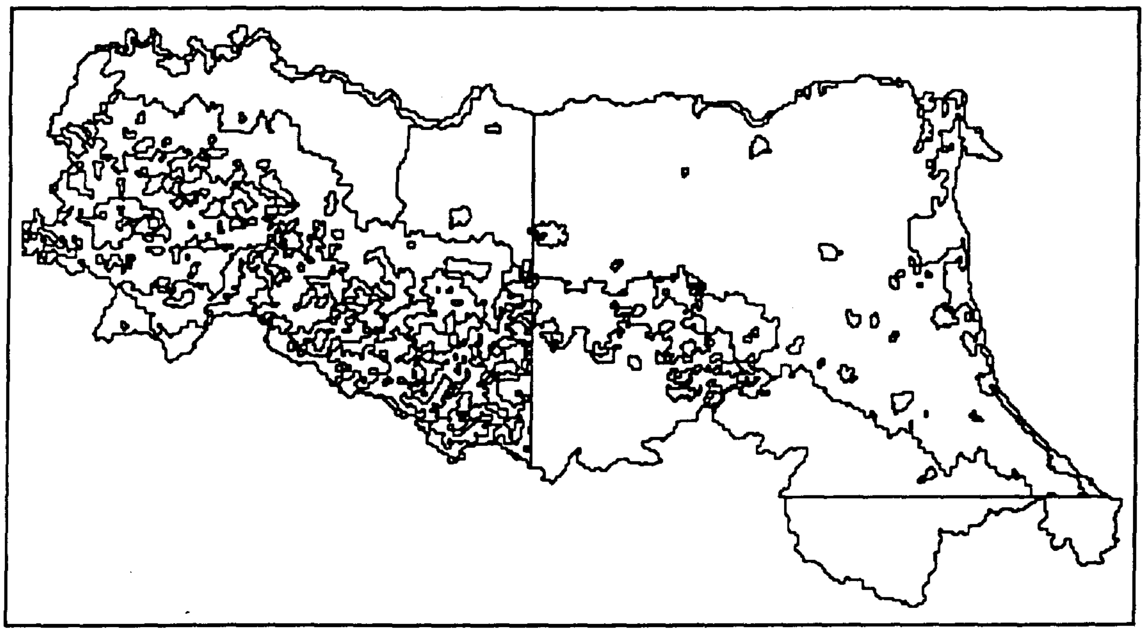

Figure 3: Emilia Romagna. Neostratification for summer crops.

3.3. Bayern (Germany)

The "Bezirke" Niederbayern and Oberpfalz have been studied. The stratification for 1988 was based on a previously existing ecological classification, based on climatic, phenologic, geologic, and soil criteria. The 121 units of this classification were first aggregated to 68 using mean annual temperature and rainfall, and then to 10 strata by land use, phenological and soil criteria. Slight modifications were performed in 1989, the most important of them being the aggregation of two very similar strata.

3.4. Castilla y León (Spain)

The area studied was made up of the provinces of Valladolid and Zamora in 1988 and 1989, and enlarged to another two provinces in 1990. The stratification is mainly based on topographic, geologic, and slope maps, as well as statistical data of proportion of arable land per "municipio". A 1987 LANDSAT TM image covering most of the region was used to revise strata and solve some doubts.

A manual approach gave 6 strata corresponding to: Mayor river valleys, limestone uplands, arable plains, mixed arable with vineyards, hilly land, and mountains. An additional problem appears in this case: the region is shared by the zones 29 and 30 of the UTM coordinates (Fig.4). Each zone is treated separately for segment sampling and strata digitisation.

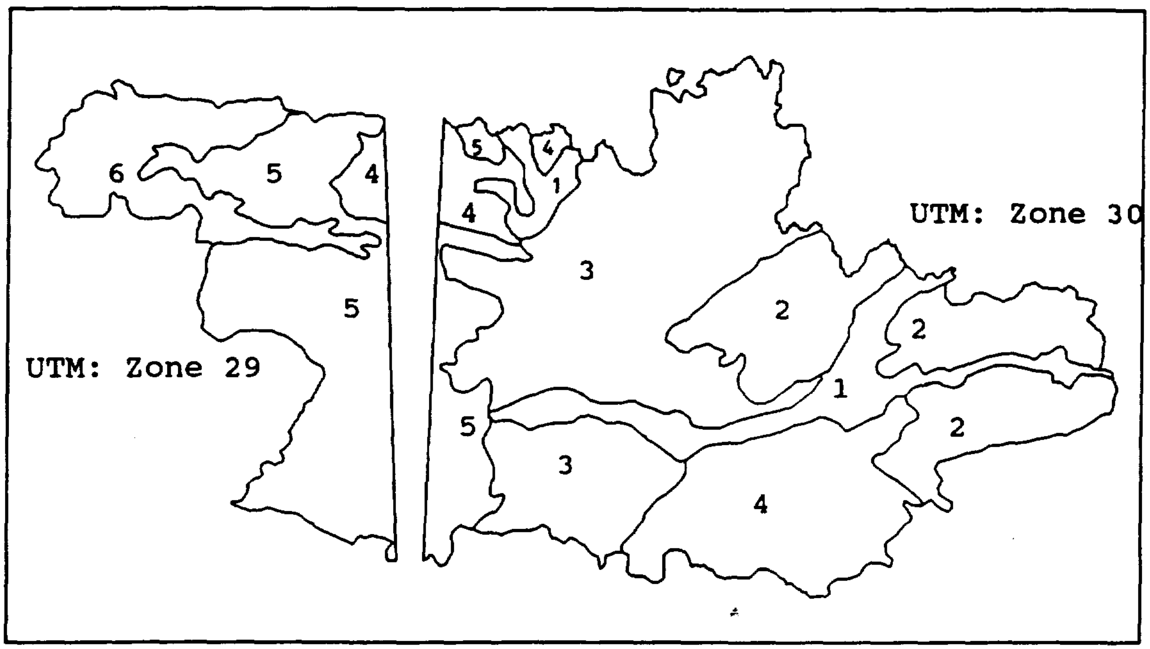

Figure 4: Valladolid-Zamora. Ground survey strata in UTM zones 29 and 30.

3.5 Makedonia (Greece)

The subregions Kentriki and Dytiki have been studied. The basic informations for stratification in 1988 were :
-map of administrative limits
-official classification of municipalities into three categories: plain, hill, and mountain.
-topographic and geologic maps,

Seven strata have been defined after studying this material: High mountain, mountains and hills with crops, hills and plains of Kalkidiki, high plains and basins, non irrigated plains and hills, irrigated plains, and Axios delta.

In 1989 the limits of the strata were redefined by photointerpretation of TM images; about 20% of the region changed from one stratum to another. The result has been an important improvement in the efficiencies of stratification.

3.6 The NASS-USDA stratification (U.S.A.)

The National Agricultural Statistical Service of the U.S. Department of Agriculture (NASS-USDA) performs a detailed stratification that improves very much the quality of ground survey results. Here is a brief description of its main features as described by Cotter (1987). The stratification is used for a period of 15-20 years.

Some strata are defined a priori:
- Crop land: 15-50%, 50-74%, and 75% or more cultivated.
- agricultural-urban (<15% cultivated and >20 dwellings per sq.mile)
- Residential-commercial
- range and pasture (<15% cultivated)
- non agricultural
- Water
- Indian reservations (western states)

The materials used include satellite imagery, aerial Photography, several types of maps, lists of National parks, wildlife or recreation areas, military installations, airports, number and size of farms, and acreages by county.

The stratification of each county is performed by one person (stratifier). The procedure is essentially manual. The boundaries between strata are roads, railways, rivers, etc.. In each stratum, primary sample units are defined. The minimum size is one segment and the maximum is 12 sq. miles.

A substratification is performed by some multivariate procedure of cluster analysis in two main steps: The clustering units are counties in one of them and primary sampling units in the other. The kind of algorithm used is not explained in the document, except that it seems to use an ordering of cluster units trying as far as possible that two consecutive units should be adjacent. The number of substrata for each of the most intensively cultivated stratum seems to be high: 15 for the first stratum in the example of Nebraska.

The cost of the stratification of one State is between 50000 and 250000 $, depending on size and complexity.

4. Some Tools for an Improved Stratification

More refined stratifications can be obtained in the near future. They should use multivariate algorithms (Cluster Analysis) and be managed in a way as automatic as possible, specially if the use of several stratification levels gives a significant improvement of the estimators precision.

4.1 Geographic Information System (GIS).

The basic units of a GIS to manage stratifications should be the segments defined for the survey, some layers being fed by the elements used for traditional stratifications. Some would summarise information coming from satellite images, and some would contain the different stratifications.

4.2 High Resolution Satellite Images (Landsat TM or SPOT).

The aggregation of pixels to the segment level can be done simply by computing averages of each channel, vegetation indexes, brightness, or other indexes. The information of classified images can be represented as percentages for each land use in a simplified nomenclature in which crops usually associated in the region by rotation practices should be grouped (in this sense, fallow is understood as a crop).

4.3 Low resolution Satellite Images (NOAA-AVHRR)

NOAA-AVHRR images have the advantage of a high temporal repetitiveness allowing for building up cloud-free mosaics, but the sharp problems arising for a good geometric location are a serious drawback. Still some information can be obtained about the area surrounding a particular segment by smoothing images.

If a stratification is drawn up using standard information and high resolution images, some segments can have missing data (clouds in many cases) and be unusable for clustering. Low resolution images can be used by a Discriminant Analysis to add the segments with missing data to the existing classes.

4.4 Clustering Segments

A Clustering scheme for getting strata can be as follows:

a) defining a dissimilarity index between segments. This index must cope with the mixture of categorical and continuous variables. A combination of the chi-2 distance (Lebart, 1984) for categorical data and an euclidean distance for continuous variables is a solution that has proved to give good results.

b) A quick clustering algorithm (k-means for example), eventually with restrictions of geographic contiguity, to get a relatively large number of classes (200-1000).

c) Hierarchical clustering without restrictions of geographic contiguity to get approximately the desired number of strata. Uninteresting or small strata can be aggregated by hand after photointerpretation.

5. Sampling Units Shared by Different Strata.

As it has been said, classic area sampling frames sre such that a sampling unit (segment) belongs to one, and only to one stratum. Most of them are based on a squared grid or made out of irregular units following physical limits. Comparisons carried out up to now suggest that estimations based on squared units have lower variances. Defining an area frame and preparing the ground survey documents is much heavier with irregular units.

Irregular units are superior in that they can manage detailed stratifications obtained by photointerpretation in which each stratum is made up of many polygons that can be as small as the sampling units. If a sampling unit must belong to a single stratum, much of the information provided by a detailed stratification is lost if strata borders must follow a fixed squared grid.

A procedure that is being tested in Spain and Portugal that keeps borders with an irregular shape even if it uses squared sampling units.

The main features of this procedure are:

- A squared sampling unit (quadrat) is split into two or more estimation units (segments) if it straddles a stratum border.

- A new stratification is made for estimation if sampling strata have different sampling rates.

- Area estimation is made by a ratio estimator (Cochran, 1977) using the total area as an auxiliary variable.

- The estimates in different strata are not independent. Covariances between the estimates in different strata must be added to compute the variance of the total estimate.

- Information from classified satellite images (SPOT or TM) can be used in a multiple regression estimator (Cárdenas, where the regressors are the segment size and the % of pixels classified as some land use.

A program in C has been written to compute the estimates with this method. The user does not need to care about substrata created in the different steps, excepting the options to aggregate some of them if their sample sizes are too small.

Some problems are still open, in particular the computation of the variance components due to post-stratification.

Figure 5 shows an example of 20Km * 20Km block with nine sampled repetitions. Quadrats n. 2, 6, and 9 are mainly in stratum C and belong to the actual sample as well as quadrats 3, 4, and 5, in stratum F and quadrat n.1 in stratum A; quadrats 7 and 8 do not belong to the sample since they are mainly in F where the sampling rate is 5/400.

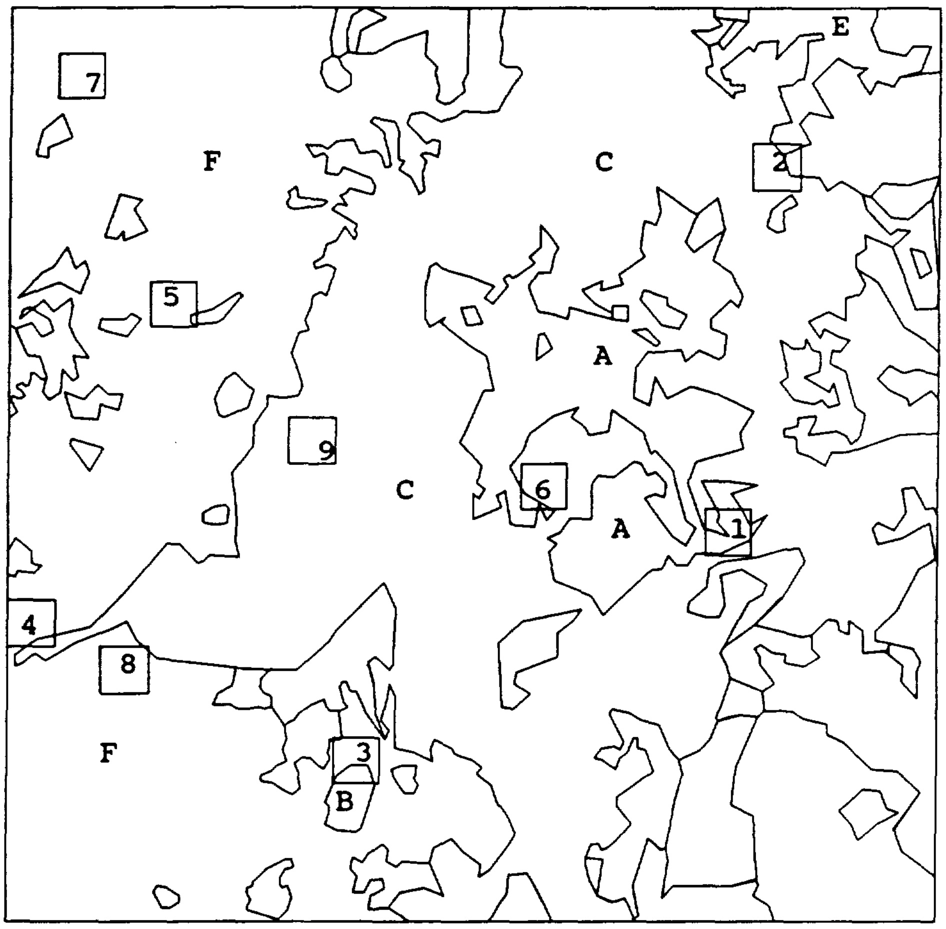

Figure 5: Block of 20Km*20Km with a presample of 9 quadrats in a detailed stratification.

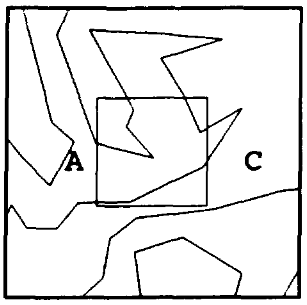

Fig. 6: Quadrat n. 1

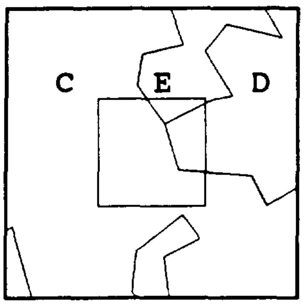

Fig.7: Quadrat n.2

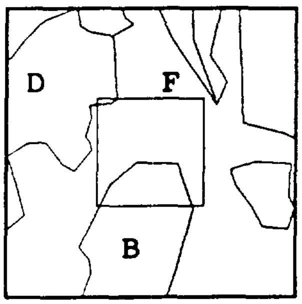

Fig. 8: Quadrat n.3

BIBLIOGRAPHY

Allen J.D., Hanuschak G.A. (1988) - The remote sensing applications programme of the National Agricultural Statistics Service. Report n. SRB 88-08. NASS-USDA. Washington.

Anderson T.W. (1984) - An introduction to multivariate statistical analysis. Wiley. New York.

Angelici G., R. Slye, M.Ozga, P.Ritter (1986), "PEDITOR - a portable image processing system" - IGARSS 86 Symposium, Zurich.

Anonymous (1990) - Tele-Agri-News n. 2 and 3. IRSA. JRC Ispra.

Cárdenas M., Hanuschak G.A. (1978) - Multiple Regression Estimation using classified Landsat data. report NASS-USDA . Washington.

Chhikara R. (1986) - Crop acreage estimation using a Landsat-based estimators as an auxiliary variable. IEEE Trans. Geosc. & R.S.

Cochran W. (1977). Sampling Techniques. Wiley, New York

Consorzio ITA. (1990) - Telerilevamento in agricoltura. Previsione delle produzioni di frumento in tempo reale. Min. Agric. Roma.

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

Delincé J. (1990) - Premier bilan de l'Action 1 "Inventaires Règionaux". Conf. Appl. of Rem. Sens. to agr. Stat. (Varese). Office Publ. EC.

Everitt B.S. (1980) - Cluster analysis. Heineman. London.

Gallego J., Rueda C., Delincé J. (1991) - Stability of regression correction through remote sensing for crop acrage estimation. Submitted do Int.J. of Remote Sensing.

Gonzáles F. et al. (1990) - Comparison of JRC and USDA segments. Conf. Appli. Rem. Sens. to Agr. Stat. (Varese). Office Publ. EC:

Hay A.M. (1988) - The derivation of global estimates form a confusion matrix. int. J. of Remote Sensing. 9, 1395-98.

Hay A.M. (1989) - Global estimate from confusion matrices: a reply to Jupp. Int. J. of Remote Sensing . 10. 1571-73.

Institute for Remote Sensing Applications (1988), "Television Tracking System, Digitization of boundaries with a Video Camera" - Joint Research Centre of EEC, Ispra.

Jupp D. (1989) - The stability of global estimates from confusion matrices. Int. J. of Remote Sensing. 10. 1563-69.

Konijn H.S. (1974) - Statistical theory of sample survey design and analysis. Eslevier-North Holland.

Lebart L., Morineau A., Tabard N.(1977).Techniques de la description statistique, Dunod, Paris.

Meyer-Roux J. (1990) - Agricultural statistical systems. Ispra course on application of Remote Sensing to agricultural statistics. IRSA, JRC, Ispra.

Ozga Martin, Walter E. Donovan, Chapman P. Gleason (1977), "An interactive system for agricultural acreage estimates using Landsat data" - 1977 Machine Processing of Remotely Sensed Data Symposium.

Ozga Martin (1985), "USDA/SRS software for Landsat MSS-based crop-acreage estimation - IGARSS '85 Symposium, Amherst, Massachusetts.

Porchier J.C. (1990) - Le tèlèdètection dans le programme d'enquetas du SCEES. Conf. Appl. of Rem. Sens. to Agr. Stat. (Varese). Office Publ. EC. Luxembourg.