



# Assessment of the risk criteria: summary of survey on MS strategies. Example of LT

## outline

- Regulatory context
- Results of survey on assessment of risk factors
- Principle of CART methods
- Example on LT data from 2007

## Regulatory context (art 27 of Reg 796/2004)

- The effectiveness of the risk analysis shall be assessed
- on an annual basis (i.e. before selecting the new OTS check sample)
- by establishing the **relevance of each risk factor**

By comparing the results of the risk based & random samples -> may indicate that risk factors have to be revised

## Survey: how are risk factors assessed by MS?

- “we follow art 27”
- “we use GIS”
- “we use previous experience”
- “in 2007, we used the list of factors in art 27. No change for 2008”
- “this is the 1<sup>st</sup> year, we are starting to think about it”
- “Our sample is too small”
- “we use factors A, B, C and D” (most common reply)
- BE Wa, CZ, EE, ES, SE: for each risk factor, look at % of dossiers (concerned by the factor) with payment reduction, % of CwRS rejects, % of parcels outside tolerance, mean area not found
- IT: CART

## Survey: how are risk factors assessed by MS?

- BE Wa method: for each risk factor, look at % of dossiers (concerned by the factor) with payment reduction

	Factor 1	Factor 2	Factor 3
Region 1	6% (1)	22% (6)	13% (5)
Region 2	7% (1)	0% (0)	7% (4)
overall	7%	11%	9%

5 applications with penalty = 13% of applications checked OTS, with factor 3 & falling in region 1

- Sampling rate for each risk factor based on % of anomalies, taking account of representativity of sample (# of dossiers checked per factor)

## Decision Tree methods

Objective: identify the criteria / factors that will classify the population of claims in **classes** as **homogeneous** as possible with respect to a **target variable** (e.g. at crop group level: payment reduction (Y/N), % of area not found, amount of aid not paid...) => goal: identify classes that best predict the target variable

- *Advantage of decision tree methods*: automatic analysis of all possible combinations of factors
- *2 main algorithms*: **CHAID** and **CART**

## Chi square Automatic Interactive Detection (CHAID)

- restricted to categorical variables (continuous variables have to be recoded into categories prior to analysis)
- partitions the sample into groups that maximize the between-group differences measured on the target variable
- uses chi square tests to create multi-way splits
  - $\chi^2$  statistic measures dependence between factor X and target variable Y based on the distance between observed (O) and expected (E) frequencies of Y values; with E based on hypothesis that X does not help predicting Y (i.e. same number of cases for each value of Y)

## Example of Chi<sup>2</sup> with candidate risk factor “amount of aid”

Number of applications with

OTS results (O)	No penalty	Penalty
aid ≤ 5000€	30	70
Aid > 5000€	60	20

Expected freq (E)	No penalty	Penalty
aid ≤ 5000€	50	50
Aid > 5000€	40	40

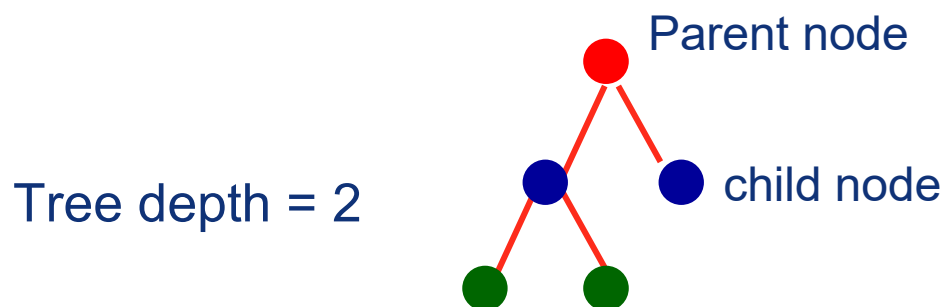
$$\chi^2 = \sum (O-E)^2/E = (30-50)^2/50 + (70-50)^2/50 + (60-40)^2/40 + (20-40)^2/40$$

Algorithm tests all possible factors and selects the one giving the largest  $\chi^2$  statistic (for a given node)



# Classification And Regression Tree (CART)

- Uses categorical & continuous variables
- Can process cases with missing values for predictors (i.e. risk factors)
- Splits a dataset into 2 groups that maximize the homogeneity of child nodes with respect to the value of the target variable
- For both methods, need to indicate max tree depth, minimum cases in parent node, in child node



## Measures of homogeneity in CART

- For categorical variable: Gini impurity =  $1 - \sum_i f_i^2$  where  $f_i$  is frequency of cases where target variable = value  $i$

The lower Gini, the better (purer) the node -> best factor

Ex: target variable takes 0 (no penalty) or 1 (penalty)

Factor aid: Aid>5000

% of 0	% of 1
0	100%

$$\text{Gini} = 1 - (0^2 + 1^2) = 0$$

Good risk factor

Factor aid: Aid>5000

% of 0	% of 1
50%	50%

$$\text{Gini} = 1 - (0.5^2 + 0.5^2) = 0.5$$

Poor risk factor

- For continuous variable: within-node variance

## Data format

For each application or rather crop group:

- candidate risk factors: factors characterizing the application before the campaign
- target variables: result of OTS check (presence of anomaly, area not found, % of area not found, amount of aid not paid)

## Case of LT data, Ex of LT **2007** OTS check sample (1/4)

Random sample: **5 738** crop groups

Attributes of claims (crop groups) selected as

Factors in 2007:

- 9 Territorial unit (district)
- 9 Municipality
- 9 Farmer's Age, Gender
- 9 New application
- 9 Claim lodged with delay
- 9 Crop group

## Case of LT data, Ex of LT **2007** OTS check sample (2/4)

- 9 Checked before (2004 - 2006)
- 9 Irregularities in previous check
- 9 Area not found in previous check
- 9 % of area not found in previous check
- 9 Changes in area from previous year
- 9 Holding area close to 1 ha (<1.5ha)
- 9 Area declared for crop group (ha)

## Case of LT data, Ex of LT **2007** OTS check sample (3/4)

- 9 Number of parcels declared
- 9 Payment claimed
- 9 Smallest parcel size
- 9 Largest parcel size
- 9 Average parcel size
- 9 number of parcels in [x, y ha[
- 9 Number of LPIS blocks claimed < 90 %

## Case of LT data, Ex of LT **2007** OTS check sample (4/4)

Type of crop group:

9 **SAPS**

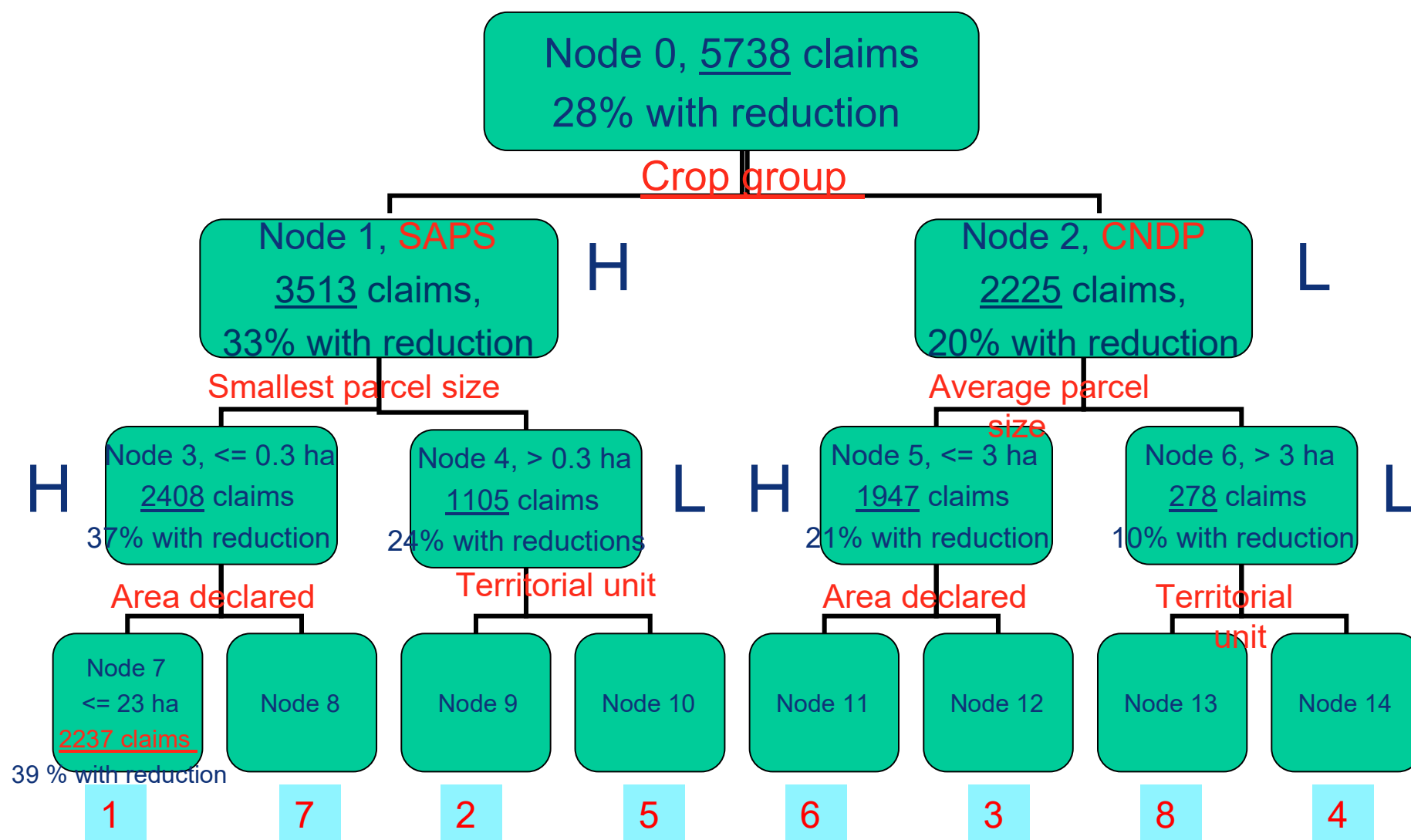
9 **CNDP** arable crops, protein crops, energy crops for bio fuel, potatoes for starch, certificated seed and flax

Target variable: **% of area not found**

% of area not found  $\leq 3\%$  (0) or  $> 3\%$  (1)

CART method applied

## Tree based on maximization of % of area not found (CART)

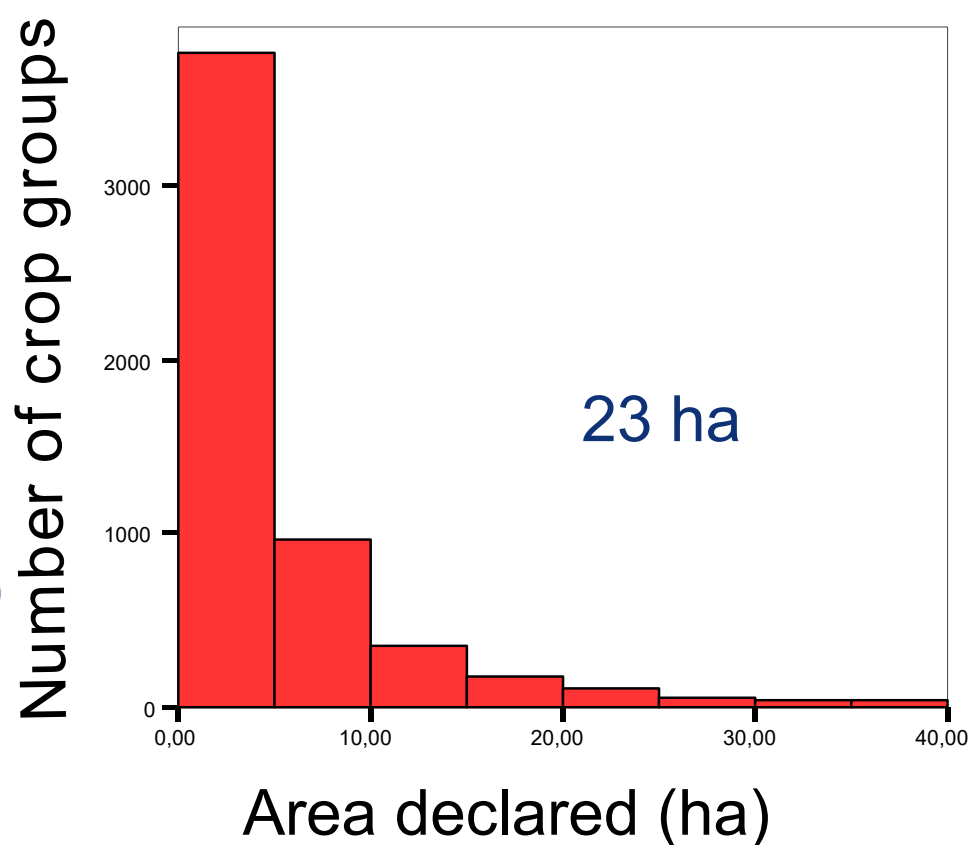
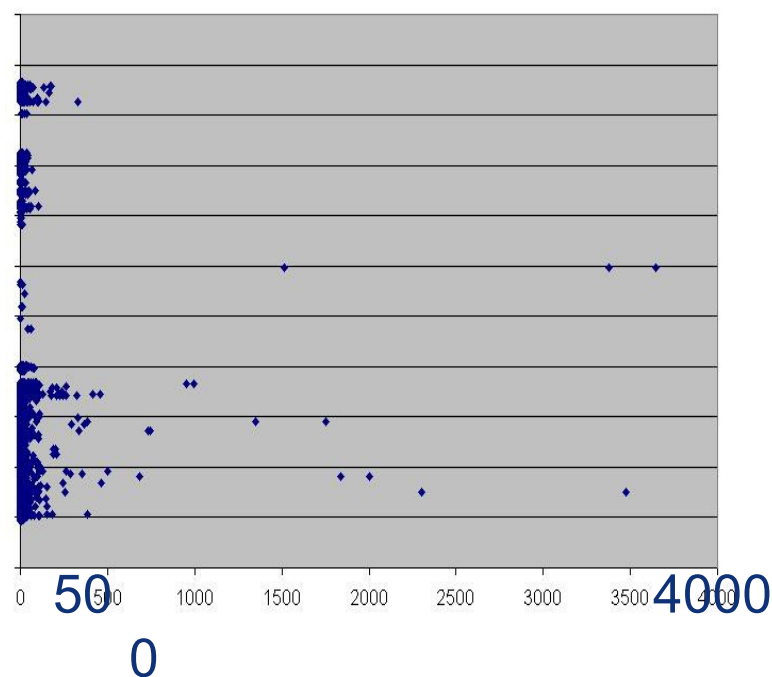




## Results: classes found with CART

5 738 crop groups classified in 8 classes (terminal nodes)	% application with reduction	# claims in class
1. SAPS & smallest parcel $\leq 0.3$ ha & area declared $\leq 23$ ha	39%	2237
2. SAPS & smallest parcel $> 0.3$ ha & districts 1,4,5,7,8,9	29%	588
3. CNDP for arable crops, protein crops, energy crops, starch potatoes, certified seed & Average parcel area $\leq 3$ ha & area declared $\leq 0.7$ ha	25%	1327
4. CNDPs & Average parcel area $> 3$ ha & districts 3,5,8,10	18.5%	98
5. SAPS & smallest parcel $> 0.3$ ha & districts $\neq$ in class 2	18%	517
6. CNDPs & Average parcel area $\leq 3$ ha & area decl $> 0.7$ ha	14%	620
7. SAPS & smallest parcel $\leq 0.3$ ha & area declared $> 23$ ha	12%	171
8. CNDPs & Average parcel area $> 3$ ha & districts $\neq$ in class 4	5%	180

# Distribution of crop groups areas in LT



## Verification based on simple statistics 1/2

<i>Crop groups First level factor</i>	<i>Total # of applications</i>	<i>% of applications with area not found &gt; 3 %</i>
<i>1. SAPS</i>	<i>3 513</i>	<i>33 %</i>
<i>2. CNDP for arable crops</i>	<i>2 116</i>	<i>20 %</i>
<i>3. Protein crops</i>	<i>64</i>	<i>12 %</i>
<i>4. Energy crops</i>	<i>38</i>	<i>8 %</i>
<i>5. Potatoes for starch</i>	<i>3</i>	<i>0</i>
<i>6. Certified seed</i>	<i>4</i>	<i>0</i>
<i>Total</i>	<i>5 738</i>	<i>28 %</i>

## Verification based on simple statistics 2/2

<i>Second level factors</i>	<i>Total # of applications</i>	<i>% of applications with area not found &gt; 3 %</i>
<i>1. Smallest parcel size <math>\leq 0.3</math> ha</i>	<i>3 335</i>	<i>32 %</i>
<i>2. Smallest parcel size <math>&gt; 0.3</math> ha</i>	<i>2 403</i>	<i>22 %</i>
<i>3. Average parcel size <math>\leq 3</math> ha</i>	<i>5 104</i>	<i>29 %</i>
<i>4. Average parcel size <math>&gt; 3</math> ha</i>	<i>634</i>	<i>16 %</i>

## Summary of analysis of LT data

- CART method allows analyzing many potential risk factors simultaneously (10s with SPSS on a PC); in the present case 26 **qualitative** and **quantitative** potential risk factors were assessed with respect to the **% of claims with reduction**
- dataset: **5 738** claimed crop groups from the 2007 OTS check **random sample** – assumed representative of population
- **Crop group, Parcel size, Area declared, District** were identified as the factors with the highest risk for target variable **“% of area not found”** with **CART** (CHAID identified other factors such as the number of declared parcels, holdings with declared area close to 1 ha, changes from previous year, age of applicant, crop group, district)
- Softwares dedicated to decision tree analysis can be found around 500€

Thank you very much for your attention!