Influence functions for CART

Jean-Michel Poggi

Orsay University

JOINT WORK WITH AVNER BAR-HEN AND SERVANE GEY

March 2011

Plan

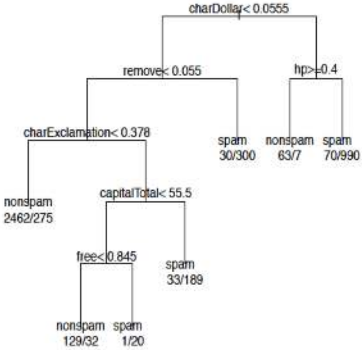
CART Classification And Regression Trees, Breiman et al. (1984)

- Learning set $L = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, n i.i.d. observations of a random vector (X, Y)
- Vector $X = (X^1, ..., X^p)$ of explanatory variables, $X \in \mathbb{R}^p$, and $Y \in \mathcal{Y}$ where $\mathcal{Y}$ is either a class label or a numerical response
- For classification problems, a classifier t is a mapping $t : \mathbb{R}^p \to \mathcal{Y}$ and the Bayes classifier is to estimate
- For regression problems, we suppose that

$$Y = f(X) + \varepsilon$$

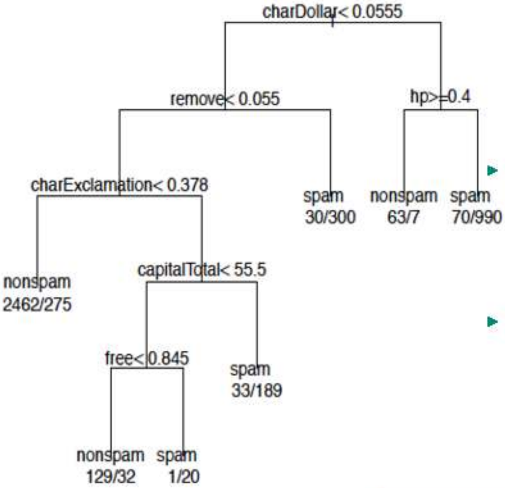
and f is the regression function to estimate

spam dataset



- Design an automatic spam detector (supervised learning problem)
- ftp.ics.uci.edu
- n=4601 email messages (1813 spams, 40%)
- p=57 predictors:
  - 54 are the % of words in the email matching a given word or character like "$", "!", "remove", "free"
  - 3 related to the lengths of uninterrupted sequences of capital letters (average length, length of the longest one, sum of the lengths of such sequences)

CART tree on *spam* dataset



- CART tree structure: 7 leaves, splits involve *charDollar*, *remove*, *hp*, *free*, *charExclamation*, and *capitalTotal*
- CART tree prediction: leaf labels: prediction of Y (spam or nonspam) and distribution of Y
- Interpretation: path root - 3rd leaf = an email containing many "$", "!", "remove", capital letters and "free" is almost always a spam

CART Classification And Regression Trees, Breiman et al. (1984)

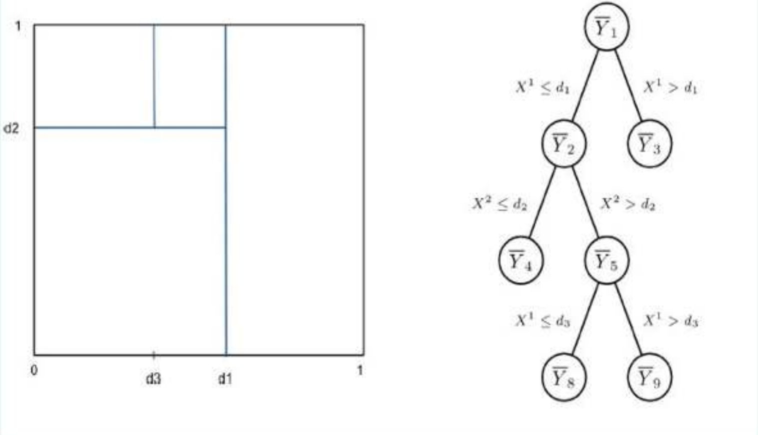
- nonparametric model + data partitioning
- numerical + categorical predictors
- easy to interpret models
- non linear modelling

- base rule for: bagging, boosting, random forests
- single framework for: regression, binary or multiclass classification

- see Zhang, Singer (2010) and Hastie, Tibshirani, Friedman (2009)

CART tree

CART tree as a piecewise constant function



Growing step, stopping rule:

- recursive partitioning by maximizing local decreasing heterogeneity
- do not split a pure node or a node containing a few data

Pruning step:

- the maximal tree overfits the data
- an optimal tree is pruned subtree by penalizing the prediction error by the model complexity

Penalized criterion

$$crit_\alpha(T) = R_n(f, \hat{f}_{|T}, \mathcal{L}_n) + \alpha \frac{|\tilde{T}|}{n}$$

$R_n(f, \hat{f}_{|T}, \mathcal{L}_n)$ the error term (MSE for regression or misclassification rate)
$|\tilde{T}|$ the number of leaves of T

A typical result for regression problems (Gey, Nedelec 2005)

There exist C_1, C_2, C_3 positive constants such that:

$$\mathbb{E}\left[\|\tilde{f} - f\|^2 | \mathcal{L}_1\right] \leqslant C_1 \inf_{T \preceq T_{max}} \left[\inf_{u \in S_T} \|u - f\|^2 + \sigma^2 \frac{|\tilde{T}|}{n_1}\right] + \frac{C_2}{n_1} + C_3 \frac{\ln n_1}{n_2}$$

where S_T is the set of piecewise constant functions defined on the partition $\tilde{T}$

In the sequel, CART trees obtained using

- *R* package *rpart*
- the default parameters (Gini heterogeneity function to grow the maximal tree and pruning with 10-fold CV)

A boosting-based outlier detection

- Cheze, Poggi (2006)
- Outliers = observations not "consistent" with most of data
  Rousseeuw, Leroy 1987

- PCA and related robust methods Jolliffe (2002)
- Methods supported by robustness ideas and based on linear modeling:
  - Minimum Covariance Determinant (MCD) Rousseeuw, Van Driessen (1999)
  - Least Trimmed Squares (LTS) Rousseeuw, Leroy (1987)
  - Least Median Squares (LMS) Rousseeuw (1984)

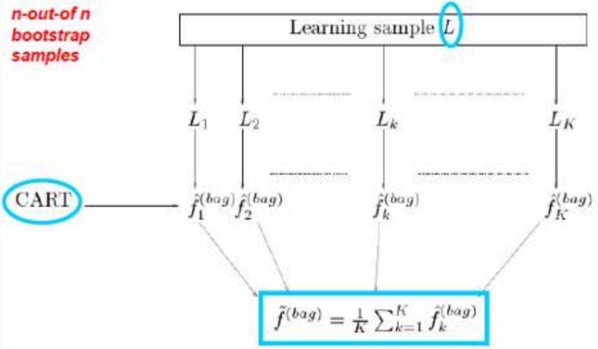
- General nonparametric regression design method
  CART trees to ensure flexible estimation and weakly structured model
- Automatic data-driven outlier detection procedure
  Boosting by adaptive resampling to explore different features of the data

Bagging

Bootstrap and Aggregating

Algorithm 1: Bagging

Breiman (1996)



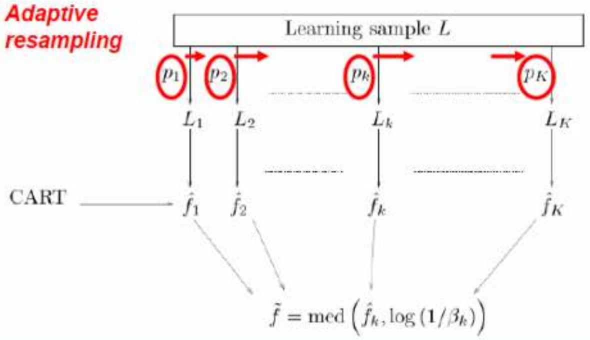
Boosting

Adaptive Bootstrap and Aggregating

Algorithm 2: Boosting

Drucker (1997)

AdaBoost (*Freund et Schapire*, 1997)



Boosting procedure $[M, i_0] = boost(L, K)$

Boosting procedure

Boosting procedure $[M, i_0] = boost(L, K)$

Input: sample L size n and K the number of iterations

Initialization: Set $p_1 = D$ the uniform distribution on $\{1, \ldots, n\}$

Loop: for $k = 1$ to K do

step 1 - generate L_k of size n, from L using p_k

step 2 - using CART, construct $\hat{f}_k$ of f from L_k

step 3 - set from L: $i = 1, \ldots, n$

$$p_{k+1}(i) = \beta_k^{1-d_k(i)} p_k(i)$$

step 4 - compute $I_{i,k} = Card\{i \in L_k\}^{\#}$

Output: $M = max_{i \in L}\ S_i$ $\quad i_0 = argmax_{i \in L}\ S_i$

where $S_i = \frac{1}{K} \sum_{k=1}^{K} I_{i,k}$

Maximum of the average number of appearances in Lk

Index of the most frequently resampled observation

Outlier detection procedure by iterated boosting

Outlier detection procedure

Outlier detection procedure

Input: J the number of applications of boosting
L the initial sample, α the significance level
K the number of iterations of each boosting

Initialization: Set $L^1 = L$

Stage 1: for $j = 1$ to J do
$[M_j, i(j)] = \textit{boost}(L^j, K)$
$L^{j+1} = L^j \setminus i(j)$
$H = L \setminus L^J$

Stage 2: Outliers are defined as the observations of index
$i(j) \in H$ such that $(M_j > C_\alpha)$

CART and stability

- CART instability
- Briand *et al.* (2009) sensitivity using a similarity measure between trees
- Bousquet, Elisseeff (2002) stability through jackknife
- Classically, robustness deals with model stability, considered globally

- Focus on individual observations diagnosis issues rather than model properties or variable selection problems
- Huber (1981) influence curve theory to define
  - different classes of robust estimators
  - measure of sensitivity for usual estimators

- We use decision trees to perform diagnosis on observations
- We use influence function, a classical diagnostic method to measure the perturbation induced by a single observation: stability issue through jackknife

Influence function

- $X_1, \ldots, X_n$ r.v. of common distribution function (df) F on $\mathbb{R}^d$ ($d \geq 1$)
  Statistic $T(F)$ naturally estimated by $T(F_n)$ where F_n is the empirical df

$$F_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$$

- The influence of an infinitesimal perturbation along δ_x (Hampel (1988))

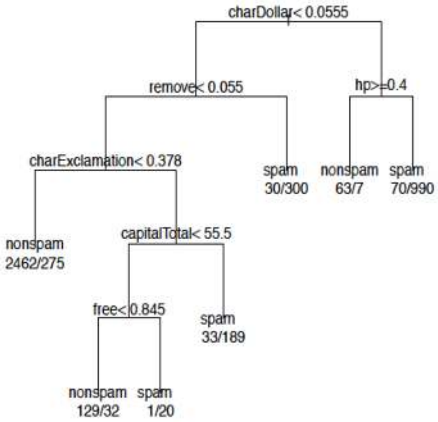
$$IC_{T,F}(x) = \lim_{\epsilon \to 0} \frac{T\big((1-\epsilon)F + \epsilon\, \delta_x\big) - T(F)}{\epsilon}$$

  is used to evaluate the importance of an observation $x \in \mathbb{R}^d$

- Connection between influence function and jackknife (Miller (1974)):
  let $F_{n-1}^{(i)} = \frac{1}{n-1} \sum_{j \neq i} \delta_{x_j}$, then $F_n = \frac{n-1}{n} F_{n-1}^{(i)} + \frac{1}{n} \delta_{x_i}$. If $\epsilon = -\frac{1}{n-1}$, we have:

$$\begin{aligned} IC_{T,F_n}(x_i) &= \frac{T\big((1-\epsilon)F_n + \epsilon\, \delta x_i\big) - T(F_n)}{\epsilon} \\ &\approx (n-1)(T(F_n) - T(F_{n-1}^{(i)})) \\ &\approx T_{n,i}^* - T(F_n) \end{aligned}$$

spam dataset: jackknife trees



- Influence of observation *i* is based on $T^{(-i)}$ the jackknife tree based on all observations except *i*
- collection of 4601 jackknife trees
  - retained variables:
  - *charDollar*, *charExclamation*, *hp* and *remove* are always present
  - *capitalLong* is present in 88 trees, *capitalTotal* in 4513 and *free* in 4441 trees
  - the 4441 trees containing *free* also contain *capitalTotal*
  - number of leaves:
  - *differences explained by the presence (or not) of the variable free: removing one observation is enough to remove the split leading to misclassification*

Plan

Influence functions for CART

- Quantifying the differences between
  - reference tree T obtained from the complete sample $\mathcal{L}_n$
  - jackknife trees $\left(T^{(-i)}\right)_{1 \leqslant i \leqslant n}$ obtained from $(\mathcal{L}_n \setminus \{(X_i, Y_i)\})_{1 \leqslant i \leqslant n}$

Three kinds of IF for CART

- we derive three kinds of IF based on jackknife trees
  - influence on predictions focusing on predictive performance
  - influence on partitions highlighting the tree structure

  following a classical distinction, see Miglio and Soffritti (2004)

  $+$

  - CART specific influence derived from the pruned sequences of trees

Influence on predictions

I_1 and I_2 are based only on the predictions

Definition I_1 and I_2

- I_1, closely related to the resubstitution estimate of the prediction error, evaluates the impact of a single change on all the predictions

$$I_1(x_i) = \sum_{k=1}^{n} \mathbb{1}_{T(x_k) \neq T^{(-i)}(x_k)}$$

- I_2, closely related to the leave-one-out estimate of the prediction error

$$I_2(x_i) = \mathbb{1}_{T(x_i) \neq T^{(-i)}(x_i)}$$

Influence on predictions

I_3 is based on the distribution of the labels in each leaf

Definition I_3

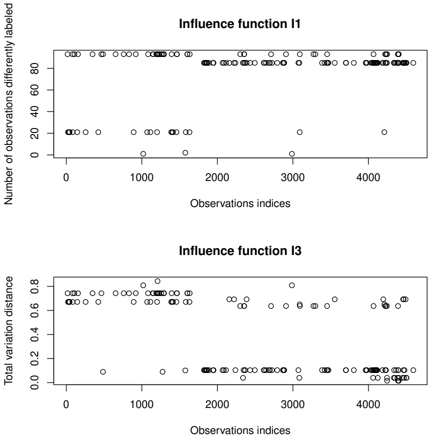
- I_3 measures the distance between the distribution of the label in the nodes where x_i falls

$$I_3(x_i) = d\left(\mathrm{p}_{x_i,T}, \mathrm{p}_{x_i,T^{(-i)}}\right)$$

where d is the total variation distance

$$d(p,q) = \max_{A \subset \{1;\ldots;J\}} |p(A) - q(A)| = 2^{-1} \sum_{j=1}^{J} |p(j) - q(j)|$$

Influence on predictions: *spam* dataset



- 163 jackknife trees lead to a nonzero number of observations for which the predicted label changes
- I_1 and I_3 for the 163 observations for which I_1 is nonzero
- 77 observations for which $I_2 = 1$ (i.e. $T(x_i) \neq T^{(-i)}(x_i)$) lead to a distance between $\mathrm{p}_{x_i,T}$ and $\mathrm{p}_{x_i,T^{(-i)}}$ larger than 0.6
- the others lead to a distance smaller than 0.1

Influence on partitions

Definition

- I_4 measures the variations on the number of clusters in each partition

$$I_4(x_i) = |T^{(-i)}| - |T|$$

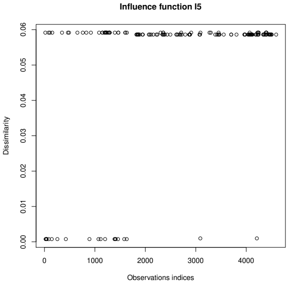
- I_5 is based on the dissimilarity difference between the two partitions

$$I_5(x_i) = 1 - J\left(\widetilde{T}, \widetilde{T}^{(-i)}\right)$$

where $J\left(\widetilde{T}, \widetilde{T}^{(-i)}\right)$ is the Jaccard dissimilarity between the partitions of $\mathcal{L}$ defined by $\widetilde{T}^{(-i)}$ and $\widetilde{T}$ (the sets of the leaves of the trees)

- Jaccard coefficient $J(C_1, C_2) = \frac{a}{a+b+c}$
  a = number of pairwise points of $\mathcal{L}$ in the same cluster in both partitions C_1 and C_2
  b (resp. c)= number of pairwise points in the same cluster in C_1, but not in C_2 (resp. in C_2, but not in C_1)

Influence on partitions: *spam* dataset



- 160 jackknife trees $\widetilde{T}^{(-i)}$ having one leaf less than T and 163 leading to a partition different from $\widetilde{T}$
- I_5 of 160 observations for which the jackknife tree has one leaf less than T
  - 139 lead to a partition $\widetilde{T}^{(-i)}$ of dissimilarity larger than 0.05
  - 21 trees with one less leaf than T, but leading to a similar partition
- all jackknife trees partitions are of dissimilarity smaller than 0.06 from $\widetilde{T}$ (very local perturbations around x_i)
- 163 trees leading to nearby partitions $\widetilde{T}^{(-i)}$ different from $\widetilde{T}$
  = the 163 trees leading to a nonzero number of mails classified differently

CART specific influence

Focus on the *cp* complexity cost constant

- consider the $N_{cp} \leqslant K_T + \sum_{1 \leqslant i \leqslant n} K_{T(-i)}$ distinct values $\{cp_1; \ldots; cp_{N_{cp}}\}$ where K_T is the length of the sequence leading to tree T
- usually $N_{cp} << K_T + \sum_{1 \leqslant i \leqslant n} K_{T(-i)}$, since the jackknife sequences are the same for many observations

Definition I_6

- I_6 is the number of complexities for which these predicted labels differ

$$I_6(x_i) = \sum_{j=1}^{N_{cp}} \mathbb{1}_{T_{cp_j}(x_i) \neq T_{cp_j}^{(-i)}(x_i)}$$

$\mathbb{1}_{T_{cp_j}(x_i) \neq T_{cp_j}^{(-i)}(x_i)}$ indicates if the reference and jackknife subtrees corresponding to the same complexity cp_j provide different predicted labels for x_i

CART tree: pruning sequence

Penalized criterion

$$crit_{\alpha}(T) = R_n(f, \hat{f}_{|T}, \mathcal{L}_n) + \alpha \frac{|\tilde{T}|}{n}$$

$R_n(f, \hat{f}_{|T}, \mathcal{L}_n)$ the error term and $|\tilde{T}|$ the number of leaves

Pruning procedure: how to find T_{α} minimizing $crit_{\alpha}(T)$ for any given α

- a finite decreasing (nested) sequence of subtrees pruned from T_{max}

$$T_K = \{t_1\} \prec T_{K-1} \prec ... \prec T_1$$

corresponding to critical complexities

$$0 = \alpha_1 < \alpha_2 < ... < \alpha_{K-1} < \alpha_K$$

such that if $\alpha_k < \beta \in< \alpha_{k+1}$ then $T_{\beta} = T_{\alpha_k} = T_k$

- Remark: this sequence is a subsequence of the best trees of m leaves

CART specific influence: *spam* dataset

Pruned subtrees sequences contain around 6 elements and lead to $N_{cp} = 27$ distinct values of the *cp* parameter (from 0.01 to 0.48)

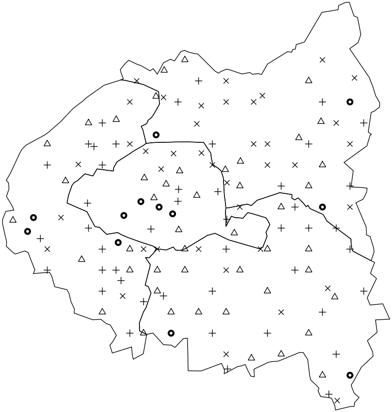
I_6	0	1	2	3	4	7	12	*13*	*14*	*17*	*18*	*21*
Nb. Obs.	2768	208	1359	79	62	1	1	66	30	2	23	2

Table: Frequency distribution of I_6 over the 4601 emails

- 123 observations leading to different predictions for at least half of the pruned subtrees
- the 2 most influential mails for I_6 lead to 78% of the complexity values for which predicting labels change
- these 2 mails are also the most influential for I_2 and I_4. They are a spam and a nonspam mails defining the second split of the reference tree: the threshold on *remove* is the middle of their corresponding values

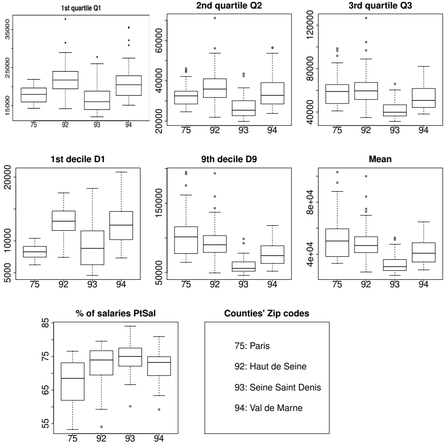
Plan

PATARE dataset



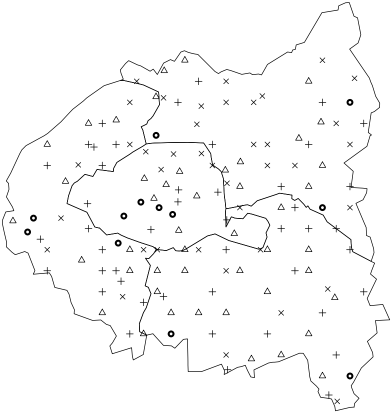
- Tax revenues of households in 2007 from the 143 cities surrounding Paris
- Cities are grouped into four counties ("département" in french)
  - Paris: 20 "arrondissements" (districts)
  - Seine-Saint-Denis (north of Paris): 40 cities
  - Hauts-de-Seine (west of Paris): 36 cities
  - Val-de-Marne (south of Paris): 48 cities
- Data freely available on `http://www.data-publica.com/data`

PATARE dataset: the variables



- Variables = characteristics of the distribution of the tax revenues per city
- For each city:
  - first and 9th deciles (D1, D9)
  - quartiles (Q1, Q2 and Q3)
  - mean, and % of the tax revenues coming from the salaries and treatments (PtSal)

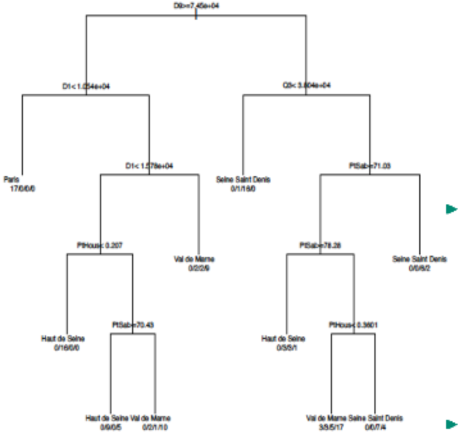
PATARE dataset: the classification problem



- supervised classification problem (quaternary explained variable): *to predict the county of the city with the characteristics of the tax revenues distribution*
- it cannot be easily retrieved from the explanatory variables considered without the county information

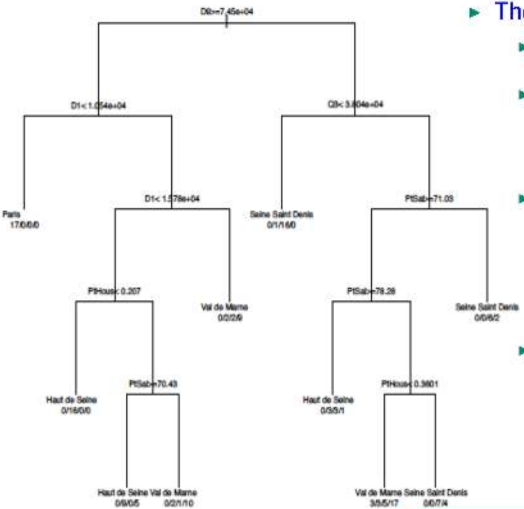
  poor recovery of counties through clusters: map of the cities drawn according to a k-means (k=4) clustering superimposed with the borders of the counties

PATARE dataset: CART reference tree



- Terminal nodes:
  - each leaf: the predicted county and the distribution of the county 75/92/93/95
  - on the left subtree, homogeneous
  - half the nodes of the right subtree are highly heterogeneous
- Labels distinguish
  - Paris and Hauts-de-Seine on the left from Seine-Saint-Denis on the right
  - while Val-de-Marne appears in both sides
- The splits

PATARE dataset: CART reference tree



- The splits
  - extreme quantiles separate richest from poorest counties
  - global predictors are useful to further discriminate between intermediate cities
  - splits on the left part mainly based on deciles D1, D9 while PtSal is only used to separate Hauts-de-Seine from Val-de-Marne
  - splits on the right part are based on all the variables but involve PtSal and mean variables to separate Seine-Saint-Denis from Val-de-Marne

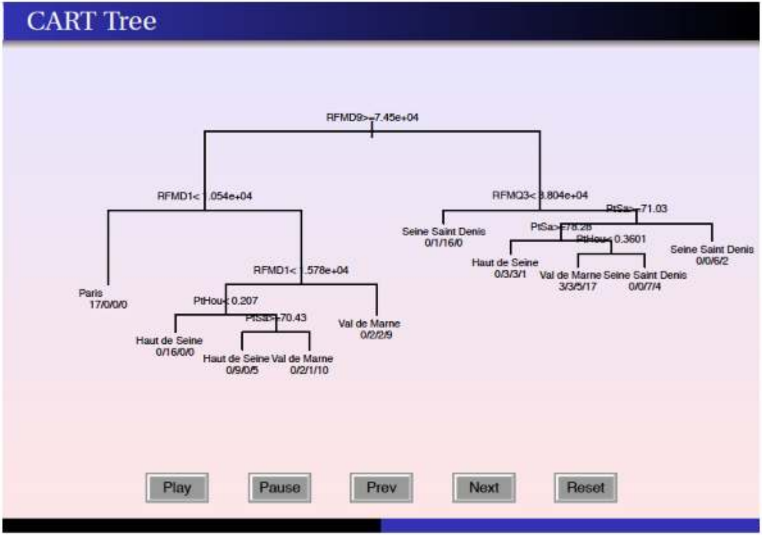
PATARE dataset: reference tree performance

- Surprisingly, the predictions are generally correct: resubstitution misclassification rate = 24.3%

Actual \ **Predicted**	75	92	93	94
Paris (75)	20	0	0	0
Haut de Seine (92)	0	30	1	5
Seine Saint Denis (93)	1	4	28	7
Val de Marne (94)	3	9	5	30

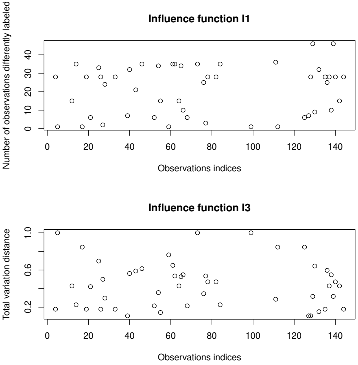
- Since the cities within each county are very heterogeneous, we look for the cities which perturb the reference tree
- *the 143 jackknife trees*

PATARE dataset: the 143 jackknife trees



PATARE dataset: influential observations

- I_1 and I_3 computed on the 75 cities (over 143) for which I_1 is nonzero

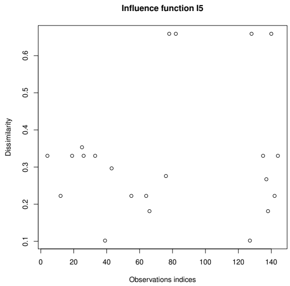


PATARE dataset: influential observations

- 45 cities (over 143) are classified differently by T and $T^{(-i)}$
- 44 jackknife trees have a different number of leaves than T, i.e. $I_4 \neq 0$

I_4	-3	-2	-1	0	1
Nb. Obs.	1	8	25	99	10

- I_5 on the 45 observations of the PATARE dataset for which I_4 is nonzero:



PATARE dataset: I_6-influential observations

Pruned subtrees sequences lead to $N_{cp} = 29$ distinct values of the *cp* complexity parameter

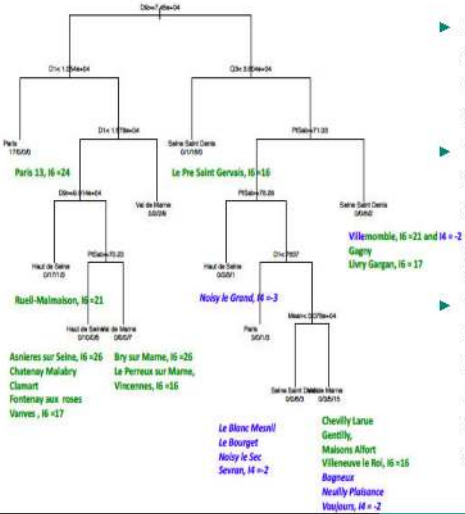
I_6	**0-2**	3	4	6	7	9	10	12	13	14	16	17	*21*	*24*	*26*
Nb	**61**	17	9	2	14	5	1	3	3	10	7	6	*2*	*1*	*2*

Table: Frequency distribution of I_6 over the 143 cities

- 2 cities change prediction labels of trees for 26 complexities: *Asnieres-sur-Seine* and *Villemomble*
- one city changes labels of trees for 24 complexities, and 2 cities for 21 complexities: *Paris 13eme*, and *Bry-sur-Marne* (from “Val-de-Marne”), *Rueil-Malmaison* (from “Hauts-de-Seine”)
- these *5 cities change labels for more than* 72% *of the complexities*

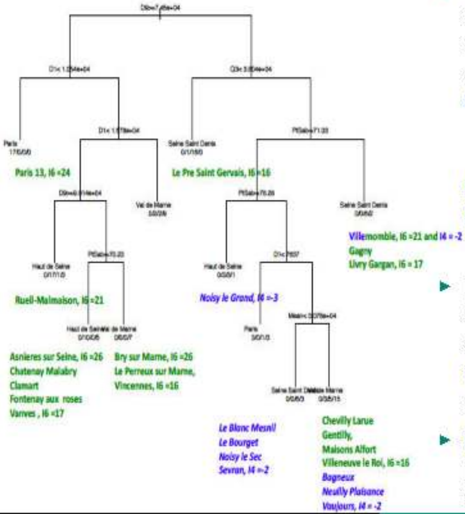
- *61 cities change labels of trees for less than* 7% *of the complexities*

PATARE dataset: influential cities - interpretation



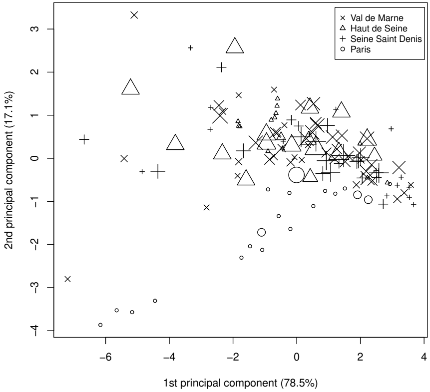
- **non trivial detection:** only 3 cities among the 26 influential cities magnified by I_6 or I_4 are misclassified by the reference tree
- index I_6 highlights cities for which **two parts of the city can be distinguished**: a popular one with a low social level and a rich one of high social level
- index I_4 highlights **cities far from Paris and of middle or low social level**. Cities of index of -3 or -2 are located in nodes of the right part of the tree whereas the rich cities are concentrated on the left part

PATARE dataset: non influential cities - interpretation



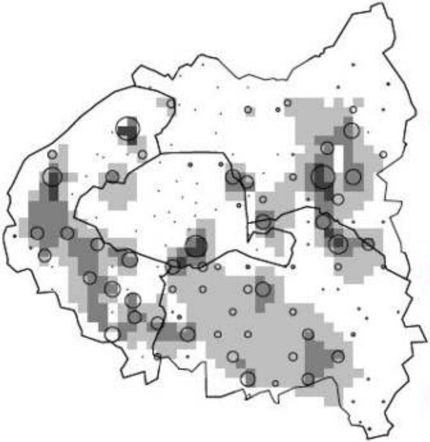
- Exploring the converse: the 51 cities of lowest values of I_6 (0 or 1) the less influential, the more stable, correspond to the 16 rich district of Paris downtown (*Paris 1er* to *12eme* and *Paris 14eme* to *16eme*) and mainly cities near Paris or directly connected by the RER line
- Influence indices **cannot** be easily explained neither by central descriptors like the mean or the median nor by dispersion descriptors as Q3-Q1 and D9-D1
- Bimodality seems the **key property** to explain high values of the influence indices

PATARE dataset: back to unsupervised analysis



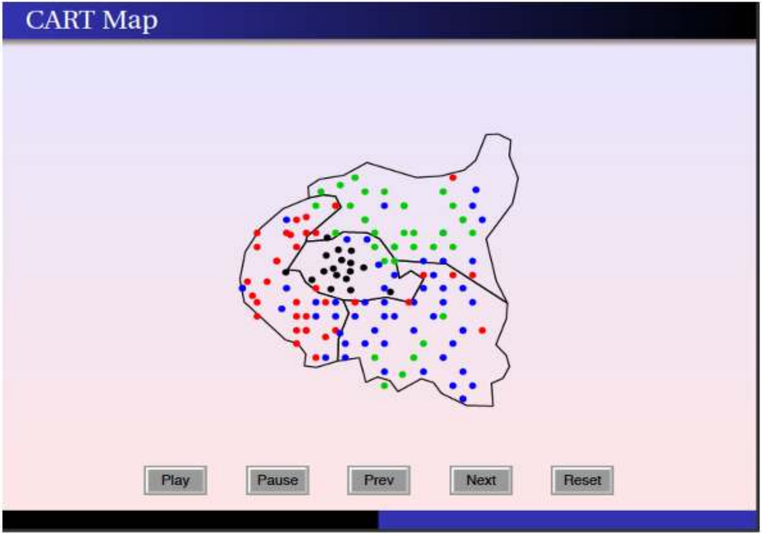
- influential observations for PCA are not related to influential cities detected using I_6 index
- in the plane of two first principal components capturing more than 95% of the total variance, each city = a symbol of size proportional to its I_6 index
- the influential points for PCA (those far from the origin) are generally of small influence for influence index I_6

PATARE dataset: influential cities - spatial interpretation



- Map useful to capture the spatial interpretation and complement the previous comments based on prior knowledge about the Paris area sociology
- Each of the 143 cities = a circle proportional to its index I_6
  + a spatial interpolation performed using 4 gray levels
- Paris is stable, and that each surrounding county contains a stable area: the richest or the poorest cities
- Remarkable fact: the white as well as the gray areas are clustered

Back to the data: spatial visualization of jackknife trees



References

- Breiman, Friedman, Olshen, Stone (1984) *Chapman & Hall*
- Briand, Ducharme, Parache, Mercat-Rommens (2009) *CSDA*
- Bousquet, Elisseeff (2002) *JMLR*
- Cheze, Poggi (2006) *JSRI*
- Gey, Poggi (2005) *CSDA*
- Gey, Nedelec (2005) *IEEE PAMI*
- Hampel (1988) *JASA*
- Huber (1981) *Wiley*
- Jolliffe (2002) *Springer*
- Miller (1974) *Biometrika*
- Miglio, Soffritti (2004) *CSDA*
- Rousseeuw (1984) *JASA*
- Rousseeuw, Leroy (1987) *Wiley*
- Rousseeuw, Van Driessen (1999) *Technometrics*
- Verboven, Hubert (2005) *Chemometrics and Int. Lab. Syst.*
- Zhang, Singer (2010) *Springer*
- Bar-Hen, Gey, Poggi (2010) *hal.archives-ouvertes.fr/docs/00/56/20/39/PDF/cart.influence.pdf*

Thanks and greetings to Anestis, undoubtedly a highly ”influential” statistician ...