Big Data et statistiques publiques

Questions de méthodes

Pauline Givord, Benjamin Sakarovitch et Stéphanie Combes (DMCSI, INSEE)

30/11/2016

Au delà des nouvelles données, quelles méthodes utiliser ?

On a vu qu'au sein de ces nouvelles sources de données, les données réellement volumineuses tiennent une place particulière et leur utilisation repose sur un modèle significativement différent (présentation de Benjamin).

Mais la **science des données** (ou **datascience**) c'est aussi : l'ensemble des techniques et technologies visant à :

- acquérir des données,
- les stocker,
- les traiter,
- les visualiser et
- les modéliser

et ce même si elles ne sont pas très volumineuses.

Machine learning

Le machine learning, de quoi parle-t-on ?

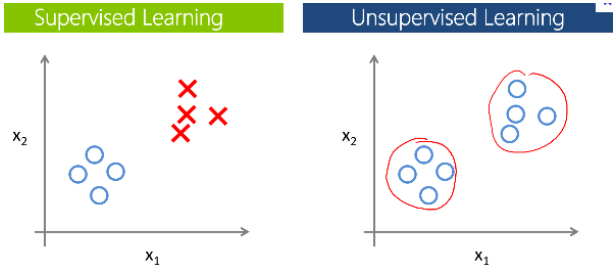
Une discipline qui n'est pourtant pas nouvelle, mais qui revient en avant avec les progrès technologiques permettant de manipuler un grand nombre de données dans un temps raisonnable.

Surtout utilisées en **prédiction**, ces techniques reposent sur **l'automatisation** de la sélection de modèles, ce qui est particulièrement intéressant lorsque l'on dispose d'un **grand nombre d'observations** et/ou de **nombreuses variables**.

- lorsqu'on dispose d'énormément d'observations, on peut **estimer un grand nombre de paramètres** par exemple (poids dans un réseau de neurones)
- lorsqu'on dispose d'un grand nombre de variables, on s'affranchit de la *spécification* au sens économétrique et on **sélectionne automatiquement** les variables pertinentes.

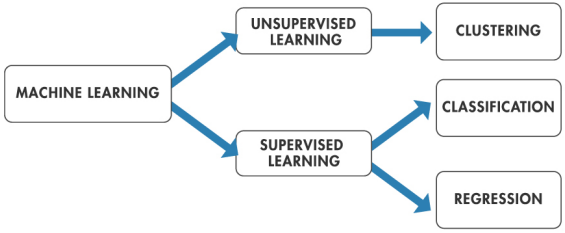
Deux familles de méthodes statistiques

- l'apprentissage **non supervisé** ou **clustering** (classement) qui explore les données, sans *a priori*, pour découvrir des regroupements et chercher à les interpréter.
- l'apprentissage **supervisé** ou **classification**, qui utilise des "exemples" labélisés pour apprendre des règles statistiques réutilisées pour prédire le label de nouvelles observations (détection des fraudes, reconnaissance d'image,..)



Oui mais de quoi parle-t-on concrètement ?

Les approches supervisées peuvent s'appliquer à des variables continues ou qualitatives.



Certaines méthodes peuvent sembler nouvelles pour l'économètre comme les **forêts aléatoires** ou les **k-means**…

…d'autres reposent sur des approches bien connues comme la **régression linéaire ou logistique** ou la **CAH**.

Utiliser la régression en Machine Learning

La différence, c'est qu'on aura tendance à utiliser une régression associée à une procédure de sélection de variables :

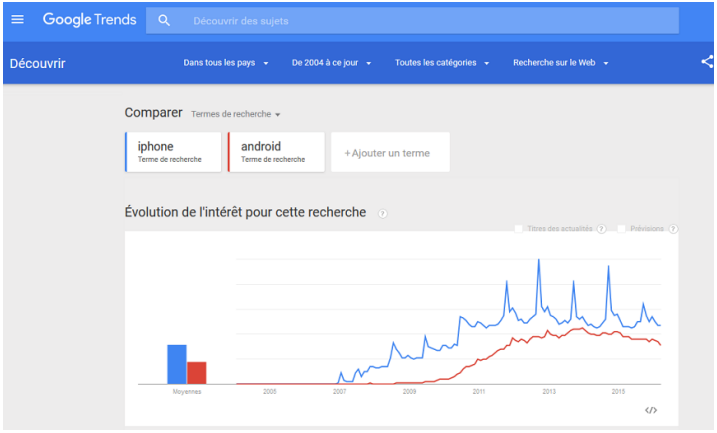
- au lieu de réflechir à une spécification économétrique du modèle, on va choisir un **critère de qualité de la prédiction** : par exemple le MSE (moyenne des erreurs au carré)
- et on va **sélectionner automatiquement les variables** qui aboutissent à la spécification optimisant ce critère.

On parle de stepwise, régression pénalisée (LASSO, Elastic Net)…

Déjà bien utilisées en analyse conjoncturelle

Déjà couramment utilisées en **analyse conjoncturelle** notamment pour construire des étalonnages sur données d'enquête, l'intérêt :

- permet de gérer le **grand nombre de variables** (ici les soldes d'opinion)
- permet d'exploiter des données dont **on n'a pas nécessairement la maîtrise** (expérimentation Google Trends, cf. note de conjoncture mars 2015)



recherches Google et analyse conjoncturelle

Dans cette expérimentation, on avait cherché à voir si les **tendances de recherche des utilisateurs dans Google** permettait de prédire la consommation en biens des ménages (indicateur publié avec un délai de un mois).

Hypothèse : les consommateurs vont se renseigner sur le produit sur Google avant son achat. Les tendances de recherche peuvent donc renseigner sur les **intentions d'achats**.

Beaucoup de variables : une série par terme recherché ou par catégorie de termes, des données dont le **procédé de production n'est pas documenté**.

-> Procédure automatique, résultats mitigés

Machine Learning VS économétrie

Des approches déjà utilisées en séries temporelles mais quid de l'économétrie ?

ML et économétrie se distinguent dans le sens où les approches de ML :

- s'embarrassent assez peu de **formalisation théorique** : peu d'hypothèses en amont, pas de tests sur les résidus, calcul d'intervalles de confiance…
- peuvent s'avérer très performantes en prédiction mais peuvent être **malaisées à interpréter**.
- ne sont pas conçues pour mettre en évidence des **effets causaux**.

Ce n'est pas pour autant une solution de facilité, elles répondent à un **protocole assez précis** que nous illustrerons sur un exemple.

Un regard économétrique sur un problème de machine learning peut être un plus : choix des données, **Google Flu**…

Quel intérêt pour la statistique publique ?

Au delà de l'analyse conjoncturelle, la popularité croissante de ces approches a attiré l'attention des instituts de statistique qui commencent à les considérer pour un certain nombre d'**applications** :

- la codification automatique
- l'identification automatique des anomalies et leur redressement
- l'imputation
- exploration de données

Exemple : détecter les aires urbaines

Dans le cadre d'un partenariat avec Orange (SENSE), Eurostat et l'INSEE (DMAEE, DMRG), nous nous sommes intéressés au potentiel des données de téléphonie mobile pour contribuer à la compréhension de l'**organisation du territoire**.

Les données de téléphonie mobiles (appels, messages, géolocalisés au niveau de l'antenne qui les enregistrent) fournissent en effet des informations relatives à la **densité de population présente** à un endroit à un moment donné et aux **déplacements** des abonnés.

Des travaux sur la **qualification du territoire à partir des profils d'activité des antennes** ont d'ailleurs déjà été menés avec des résultats encourageants (Boston : distinction de zones d'activité, résidentielle, commerciales…). Peut-on généraliser ces études à l'**échelle nationale** ?

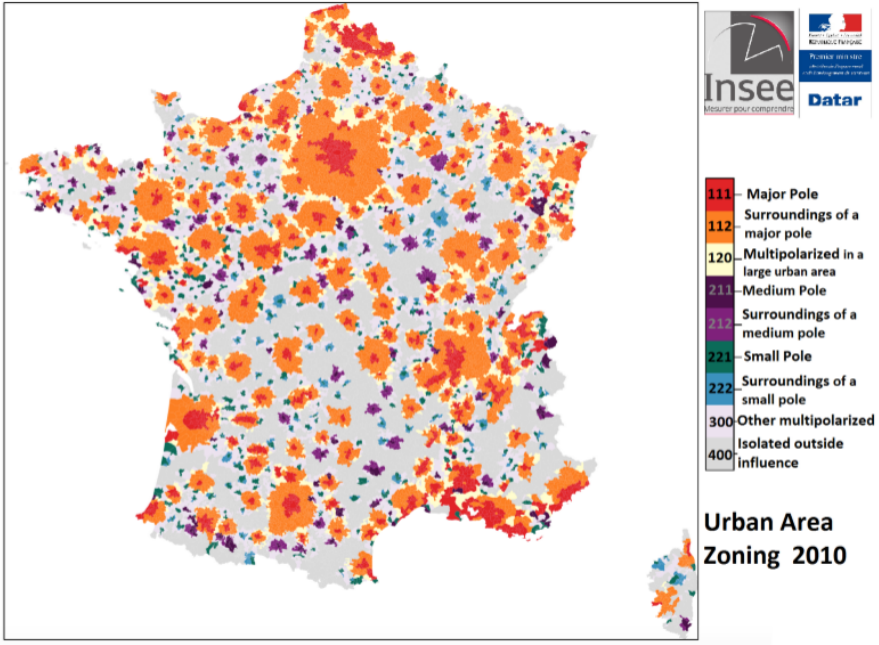
Exemple : détecter les aires urbaines

Plus précisément, nous avons cherché à voir s'il était possible de **reproduire avec les données de téléphonie mobile le zonage du territoire en aires urbaines publié par l'INSEE et la DATAR** (ZAU)

Le zonage permet d'identifier les **poles urbains** (grands, moyens et petits) selon le bâti, la population et l'offre d'emploi et leur **attractivité** sur les communes avoisinnantes (via les flux de déplacements domicile-travail).

Contexte : les données de téléphonie mobile ont été identifiées par Eurostat comme présentant un potentiel important dans divers domaines (population, mobilité, tourisme…), nous cherchons à **évaluer ce potentiel**, même lorsque les données sont agrégées, comme c'est le cas ici

ZAU


Insee
Mesurer pour comprendre
Datar
111 Major Pole
112 Surroundings of a major pole
120 Multipolarized in a large urban area
211 Medium Pole
212 Surroundings of a medium pole
221 Small Pole
222 Surroundings of a small pole
300 Other multipolarized
400 Isolated outside influence
Urban Area Zoning 2010

Données

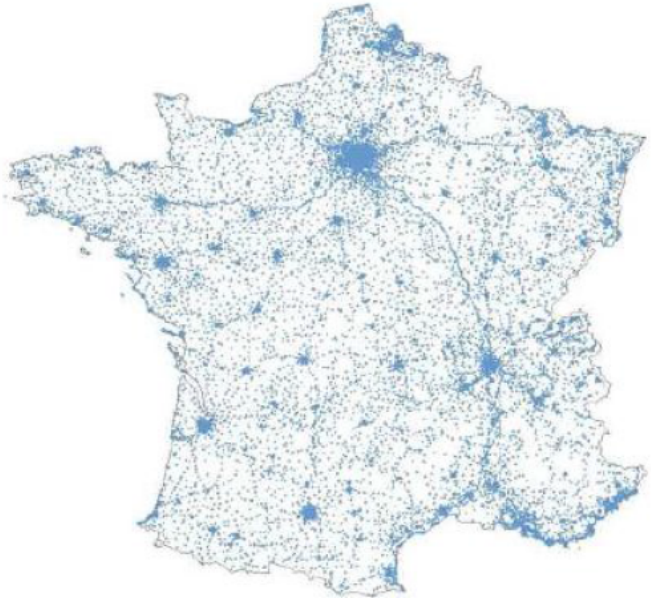
Les données agrégées mobilisées ici sont issues de la base de données **CDR** 2007 d'Orange : des données collectées à des fins de **facturation** essentiellement.

La base originale contient donc un enregistrement par *évènement* (appel reçu ou passé, message reçu ou envoyé), associé à une date et heure, un identifiant d'**abonné** et un identifiant de **tour**.

timestamp	caller	callee	event	duration	area id	tower id
2007/10/01 23:45:00	HJ123423	R482G9342	VO	3656s	1548	53571
2007/13/01 12:10:04	TR234S3	43FG3423	SI	125c	32768	53571

Données

Les données ont ensuite été agrégées au niveau de chaque tour (~18000) pour chaque heure entre **mai et octobre 2007**.



En quoi est-ce un problème de machine learning ?

Les données étant agrégées, il n'était pas possible de les injecter directement dans l'algorithme de production de ZAU originel, dont la calibration est par ailleurs *ad hoc*.

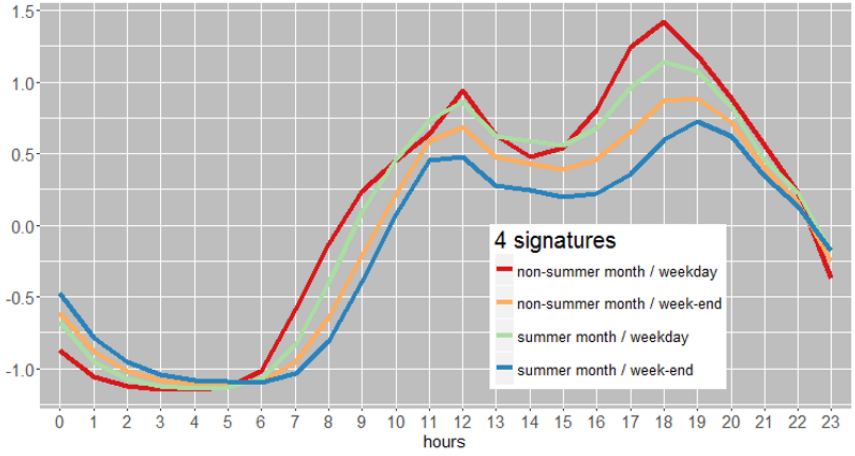
Proposition :

- *Etape 1* : transformer les séries temporelles horaires disponibles pour chaque tour en profil d'activité horaire moyen sur une journée : **features engineering**
- *Etape 2* : **identifier le lien** entre la forme de ces profils et le type d'aire de la commune où se trouve la tour en recourant à des algorithmes de **classification supervisée** (ML)
- *Etape 3* : **évaluer les performances** de la méthode calibrée en 2.

L'ensemble de ces étapes constitue une procédure de ML classique

Etape 1 : construction des variables ou *features engineering* (1/2)

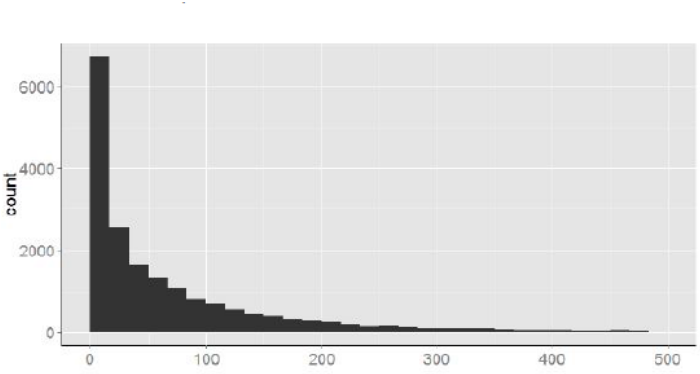
- les nombres d'événements absolus diffèrent pour chaque antenne donc on **standardise** les séries
- puis on **moyennise** par heure pour obtenir un profil horaire typique
- on distingue les jours ouvrés des jours de week-end/fériés, les mois d'été des autres (saisonnalité)

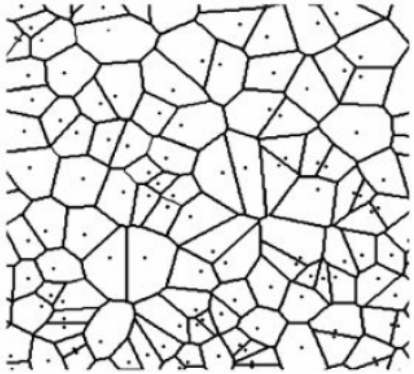


Etape 1 : construction des variables ou *features engineering* (2/2)

Pour prendre en compte l'**intensité** de l'activité globale mesurée au niveau d'une antenne (~proxy de la densité locale de population), on considère également:

- le **nombre moyen d'événements** et/ou la forme des **cellules de voronoï**

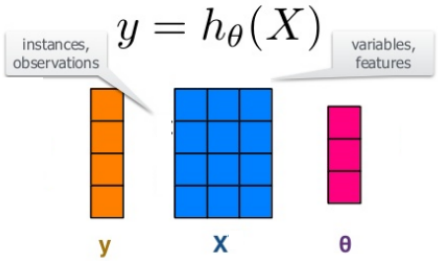




Etape 2 : estimation

L'utilisation de techniques de ML est pertinente ici car on ne sait pas *a priori* quelle est l'**information utile** à la détection du type d'aire urbaine dans les profils d'activité

On rappelle qu'on cherche à estimer le lien entre features (ici les profils horaires et les caractéristiques du voronoï=24x4+2) et type d'aire urbaine



On a testé plusieurs méthodes, dont les **régressions logistiques pénalisées** et les **forêts aléatoires**, qui sont des agrégations d'arbres de classification.

Etape 2 : estimation, exemple de méthode

Un arbre de classification est une approche de ML en soi

- il est construit par **partitionnement récursif** selon une règle de séparation (Gini index par exemple),
- la **classe prédite** (ici le type d'aire urbaine) d'une nouvelle observation (ici une antenne) est la classe majoritaire dans la *feuille* correspondante
- ce classifieur est facile à interpréter mais considéré comme **peu robuste**, car sensible à l'échantillon

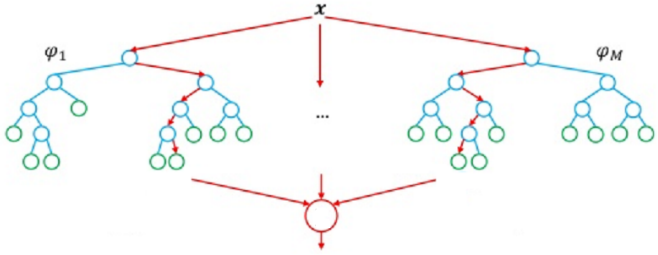
Etape 2 : estimation, exemple de méthode

La forêt consiste à **agréger** des centaines d'arbres :

- construits sur des échantillons **bootstrap**
- l'espace des *features* est réduit à chaque embranchement (tirage aléatoire)

Cette double **randomisation** permet de favoriser la diversification des arbres et d'obtenir un estimateur agrégé plus robuste

- la **classe prédite** (type d'aire ici) est la classe majoritairement prédite par les différents arbres



Etape 3 : évaluation

Pour évaluer la performances d'une prédiction, on compare les classes prédites par l'algorithme avec les classes réelles obervées (les types d'aires de ZAU 2010)

Pour cela on recourt en général à une **matrice de confusion** (table de contingence) :

	Spam (Predicted)	Non-Spam (Predicted)	Accuracy
Spam (Actual)	0	10	0.0
Non-Spam (Actual)	0	990	100.0
Overall Accuracy			99

	Spam (Predicted)	Non-Spam (Predicted)	Accuracy
Spam (Actual)	27	6	81.81
Non-Spam (Actual)	10	57	85.07
Overall Accuracy			83.44

A partir de cette matrice il est possible de construire plusieurs indicateurs :

- la **précision** de la méthode, ou ici le taux de communes dont le type d'aire a été bien prédit par la méthode (diagonale de la matrice)

Etape 3 : évaluation

- on peut aussi considérer les **taux de classification par classe**, d'autant plus important en présence de classes déséquilibrées où la *précision* ne sera pas fiable

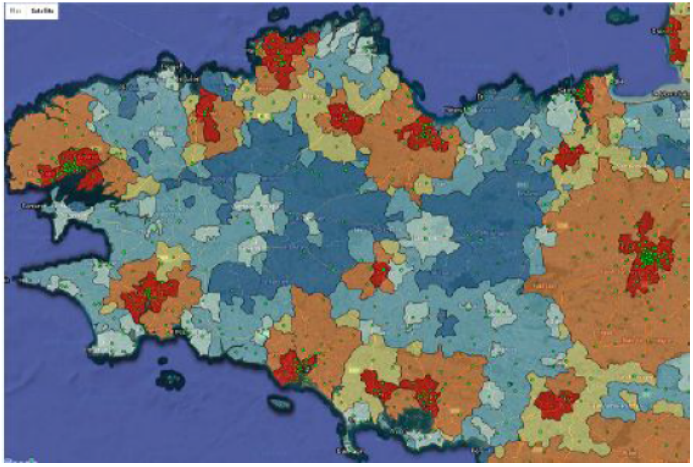
- il existe des indicateurs synthétiques prenant en compte ce biais, par exemple la **G-moyenne** (moyenne géométrique des taux de bonne classification par classe) :

$$G - mean = \left(\prod_{i=1}^{M} R_i \right)^{1/M}$$

Cet indicateur pénalise fortement une méthode qui rate complètement une classe même si celle-ci est faiblement représentée.

Rq : classes désequilibrées

Code	Urban areas	Municipalities	Antennas
111	Major urban center	9%	54%
112	Surroundings of a major urban center	34%	18%
120	Multipolarized in a large urban area	11%	5%
211	Medium urban center	1%	3%
212	Surroundings of a medium urban center	2%	1%
221	Small urban center	2%	3%
222	Surroundings of a small urban center	2%	0.3%
300	Other multipolarized municipality	19%	6%
400	Isolated municipality outside influence	20%	10%



Séparations de l'échantillon d'apprentissage et de test

Au delà des algorithmes et métriques parfois distinctes, une différence fondamentale avec l'économétrie (hors prévisionnistes) réside dans le fait que **l'étape 2 d'estimation et l'étape 3 de validation ne se font pas sur le même échantillon**.

En **étape 0**, on divise les données en un échantillon d'apprentissage (*train*) et de test (*test*) (80/20), l'estimation se fait donc sur le *train* et l'évaluation sur le *test* (on prendra le même *test* si on veut comparer les performances de plusieurs méthodes)

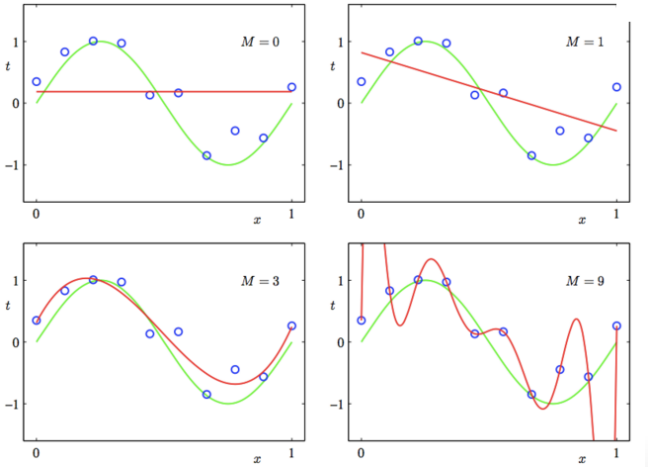
L'objectif étant de vérifier que la méthode calibrée sur un jeu de données fixe, est capable de **généraliser** à d'autres observations, autrement dit de prédire les labels (ici les types d'aires urbaines) **d'observations qu'il n'a jamais vues** avant.

En prévision de séries temporelles (et notamment en analyse conjoncturelle), on parlera de *in sample* et *out of sample*, c'est la même idée.

Séparations de l'échantillon d'apprentissage et de test

Si on évaluait la qualité de l'ajustement sur les données utilisées pour l'estimation, on risquerait de la surestimer, principalement en cas de **surajustement**.

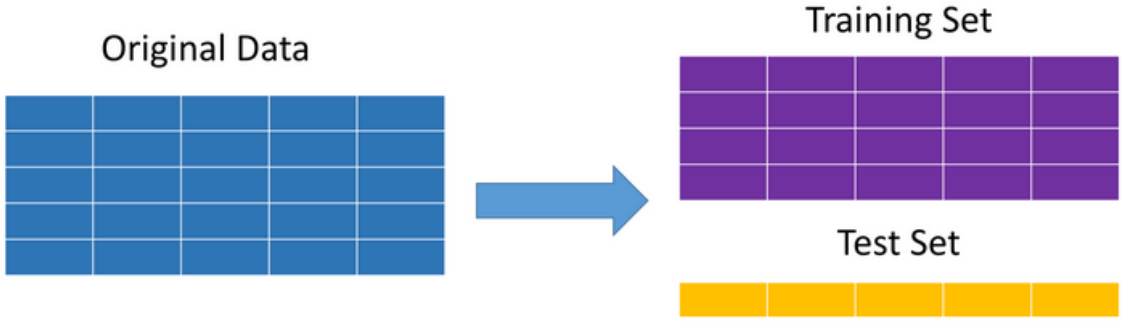
Un surajustement est en effet souvent synonyme d'une mauvaise performance en prédiction, la méthode est **spécialiste** des données utilisées pour l'estimer.



Séparations de l'échantillon d'apprentissage et de test

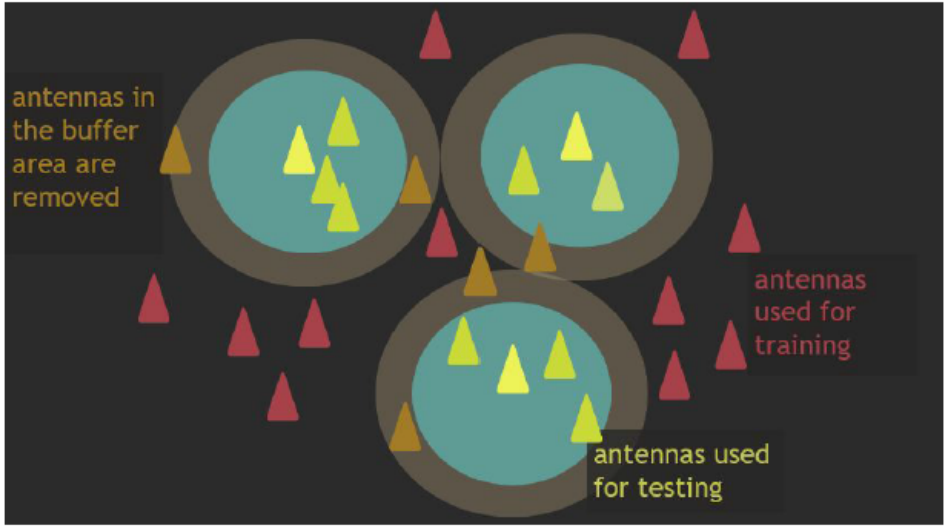
Le problème est classique et les critères de pénalisation type AIC ou BIC servent également à **pénaliser la complexité** des modèles (par exemple quand on cherche à déterminer le nombre optimal de retards dans un AR). Les approches de ML intègrent d'ailleurs souvent un critère de pénalisation.

En résumé, il est donc préférable d'utiliser des **échantillons distincts** (quand le nombre d'observations le permet), pour évaluer les performances **réelles** du modèle en prédiction (voire pour calibrer la méthode : validation croisée cf annexe).



Séparations de l'échantillon d'apprentissage et de test

Ici, on est en présence de données **autocorrélées spatialement**, on doit donc prendre garde à construire des échantillons *d'apprentissage* et *test* indépendants sinon, ça ne sert à rien !



Quelques résultats

Les résultats sont mitigés, les **indicateurs globaux ne sont pas très élevés**, et seule la détection des grands poles urbains est bonne.

Table 1: Global accuracy

Scenario	G-mean	w G-mean	Fuzzy G	Kappa	Fuzzy K	Accuracy	Accuracy 2
Orange	0.52	0.57	0.55	0.49	0.58	0.61	0.78

Table 2: Class detection rates

Scenario	Class1	Class2	Class3	Class4	Class5	Class6
Orange	0.82	0.44	0.39	0.53	0.46	0.59

Quelques résultats

Pour vérifier si les données mobiles sont seules en cause, on réestime les forêts aléatoires en utilisant cette fois des **données Insee** proches de celles utilisées pour produire ZAU, et également des données d'utilisation des sols.

Table 1: Global accuracy

Scenario	G-mean	w G-mean	Fuzzy G	Kappa	Fuzzy K	Accuracy	Accuracy 2
Orange	0.52	0.57	0.55	0.49	0.58	0.61	0.78
INSEE	0.59	0.63	0.61	0.54	0.63	0.65	0.83
INSEE+CORINE	0.63	0.67	0.65	0.61	0.67	0.70	0.86

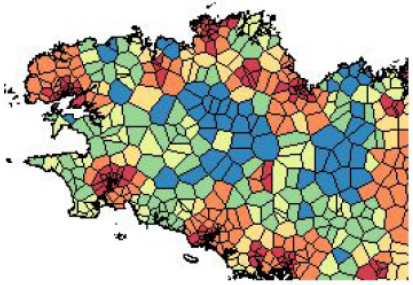
Table 2: Class detection rates

Scenario	Class1	Class2	Class3	Class4	Class5	Class6
Orange	0.82	0.44	0.39	0.53	0.46	0.59
INSEE	0.81	0.53	0.51	0.60	0.45	0.68
INSEE+CORINE	0.87	0.58	0.54	0.62	0.51	0.69

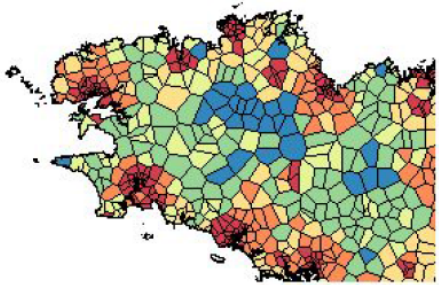
Les performances des approches sur données *officielles* sont meilleures mais restent **modérées**.

Quelques cartes

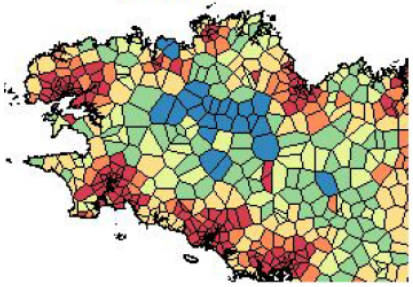
ZAU - official



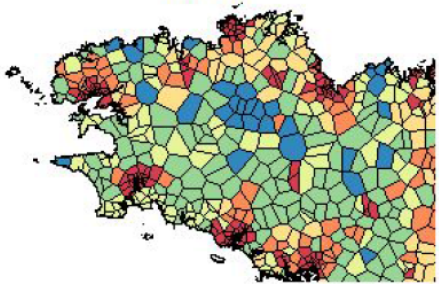
INSEE+CORINE



INSEE



Orange



Interprétabilité

Plus une variable est sélectionnée **proche de la racine** de l'arbre de classification (et dans un grand nombre d'arbres) plus elle contribue à discriminer.

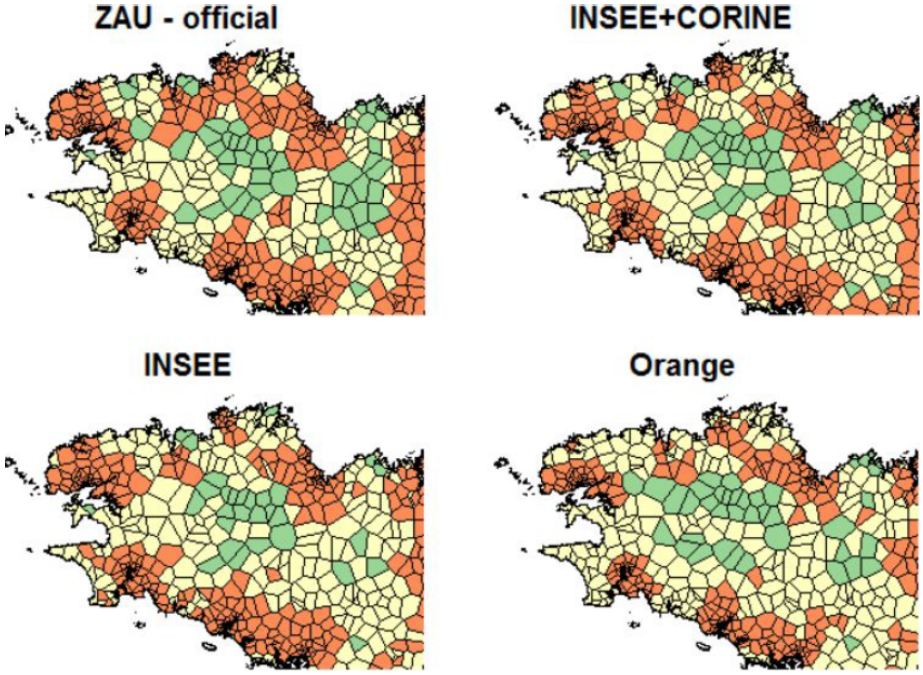
On peut ainsi construire un indicateur d'**importance des variables** dans la prédiction.

On peut voir que la forme du voronoï joue effectivement et ensuite les créneaux horaires correspondant à 9h, 12h, 16-17-18h.

Most important features

Shape Area
Shape Leng
G v8 1 9
event
G v9 0 18
v8 1 12
G v9 0 17
v8 1 17
v8 1 16
G Shape Area

Jouons sur le nombre de classes



Des investigations à poursuivre

Les résultats sont mitigés mais le problème de classification était particulièrement ardu :

- **multiples** types d'aires, sémantiquement proches et géographiquement **floues** (limites de la référence qu'on cherche à reproduire)
- distribution des types d'aires **très déséquilibrée** (même en réechantillonnant ça ne suffit pas)

Toutefois :

- la capacité de la méthode à détecter les grands poles urbains est la même avec des données de téléphonie mobile ou des données officielles,
- mais elles restent une source intéressante pour des **études plus locales** (quelques idées plus tard).

Quid du non supervisé ?

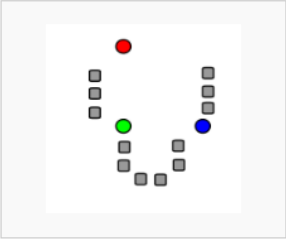
L'idée de base du clustering est de regrouper des **observations similaires**.

L'un des algorithmes de clustering les plus connus chez les statisticiens : la **classification ascendante hiérarchique** (CAH, Manning & Schatze 99)

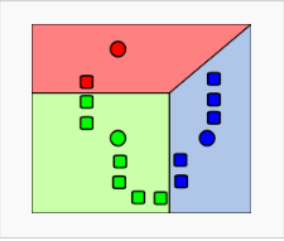
- chaque observation équivaut à un cluster en première étape, puis, à chaque itération, les clusters les plus proches au sens d'une distance choisie sont agglomérés deux à deux.

Plus connu dans la communauté du machine learning : les **k-means**

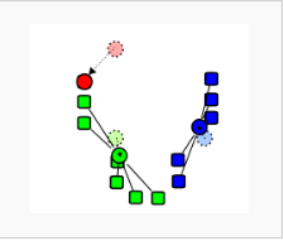
Quid du non supervisé ?



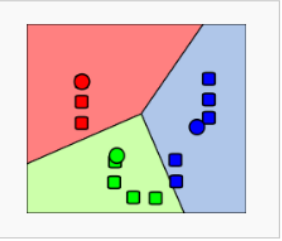
1. *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).



2. *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.



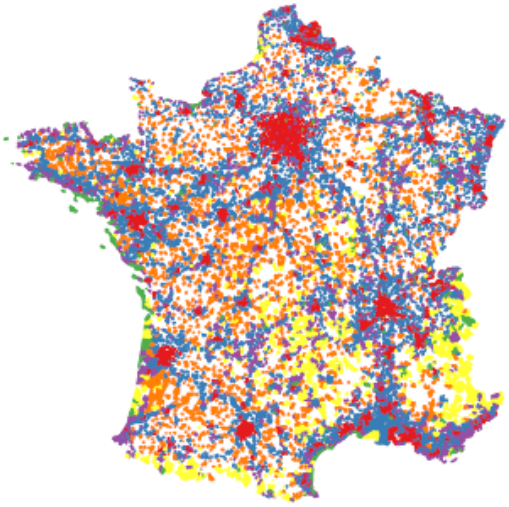
3. The centroid of each of the *k* clusters becomes the new mean.



4. Steps 2 and 3 are repeated until convergence has been reached.

Exemple - carte

- Difficile à interpréter
- Surtout à cette échelle



Conclusions d'un point de vue méthode

Choisir une approche de machine learning n'économise pas un important travail en amont :

- **feature engineering**,
- **séparation apprentissage/test**,
- gestion du **nombre de classes** et définition de l'**unité prédite** (commune/antenne),
- gestion des classes **désequilibrées**,
- choix de la **métrique d'évaluation**

En revanche, la plupart des algorithmes sont accessibles via des **librairies** et ne nécessitent pas d'être recodés.

Exploitation des données textuelles

Concept

Les données textuelles sont **non structurées**, l'information cherchée est perdue au milieu d'une grande quantité d'information ; et **difficiles à manipuler** (bruitées, changeantes, complexes, ambigues, langue-dépendante, de grande dimension).

L'**analyse textuelle** vise à extraire de l'information à partir de données textuelles représentées sous la forme de vecteurs **numériques** (*vector space models*).

Certains aspects de la discipline sont **spécifiques** (prétraitement, analyse sémantique), d'autres sont **classiques** dans le domaine du datamining / machine learning.

Challenge 1 : représenter le texte sous format numérique

Les données textuelles sont constituées de chaînes de caractères qui forment des **documents**, regroupés en **corpus**.

En premier lieu, le texte est décomposé en composants (*tokens*) : termes, ponctuation, nombres, et harmonisé (casse, orthographe : dictionnaire, matching, phonétique).

Une façon classique de décrire les documents est de les représenter dans l'espace de tous les tokens présents dans le corpus : chaque document est décrit par la présence/absence (ou fréquence) des termes qui le composent.

L'empilement de ces vecteurs forme la matrice **documents x termes**, on parle d'approche **sac de mots** (*BoW*) car l'ordre des mots dans les documents n'a pas d'importance.

Challenge 1 : représenter le texte sous format numérique

	call	*time*	*date*	*conference*	*release*	*meeting*	*corporation*	*earnir.*
document 1	2	1	3	2	1	1	1	
document 2	1		2	1	2	1	1	1
document 5		1	2		2	1	1	1
document 6	1	2	1	1	3	1	1	1
document 7	1						1	
document 8			1		1		1	1
document 9	2		1	3	1	1	1	1
document 10	2	1		1	1		1	1
document 13					1			2
document 14							3	
document 15	1			2			1	2

Challenge 1 : représenter le texte sous format numérique

Cette matrice est souvent de **grande dimension** (plusieurs dizaines de milliers de termes) et **très creuse**, en effet la distribution des mots dans le corpus est souvent caractérisée par une loi de puissance (assez peu de mots sont utilisés dans tous les documents).

On va donc chercher à **réduire la dimension** de la matrice pour en faire un objet plus facilement manipulable, par exemple en :

- supprimant les mots non discriminants et notamment les **mots outils** qui peuvent être en très grand nombre mais non informatifs (ex : "au", "avec", "dans", "mais", "elle", "lui", "nos"…)

Challenge 2 : réduire la dimension

- utilisant des classes d'équivalence, on peut considérer que différentes formes d'un même mot (pluriel, singulier, conjugaison) sont équivalentes et les remplacer par une même **forme** :
  - **lemmatisation** qui requiert la connaissance des statuts grammaticaux ou POS tagging, exemple : chevaux -> cheval
  - la **racinisation** plus fruste mais plus rapide, exemple : chevaux -> chev (peut être confondu avec la racine de cheveu)
- **pondérations** de mots, par exemple *TF-IDF* qui pondère plus **les mots qui apparaissent souvent dans peu de documents**, retirer 5% des mots qui ont une pondération trop faible permet donc de filtrer les mots non discriminants

Liens entre textmining et machine learning

La représentation sous forme matricielle facilite l'utilisation d'outils de machine learning (entre autres !)

On peut par exemple chercher à identifier les thèmes abordés dans nos données texte, pour ça naturellement on se tourne vers les outils de **clustering**.

Le fait de représenter les documents sous forme de vecteur numérique facilite le calcul de **distance** entre ceux-ci, on rappelle que le but du clustering est de regrouper des observations **similaires**.

La subtilité sera dans le choix de la distance, ou le choix de l'algorithme.

TM et ML, exemple de clustering : identification des thèmes

Notre précédent exemple était limité par le fait qu'on ne savait pas comment **interpréter les classes**.

On applique ici la CAH à la matrice termesxdocuments issue des réponses à la question "**Pouvez-vous citer les formes de délinquance, les phénomènes qu'il faudrait traiter en priorité dans la société [française] actuelle ?**" de l'enquête CVS.

L'intérêt ici est dans l'extraction d'information contenue dans le corpus sans avoir à dépouiller chaque réponse une à une.

L'interprétation est plus aisée car on n'avait **pas d'a priori** sur les classes.

Rq : il existe des approches plus probabilistes et développées pour le texte (LDA, LSA…)

TM et ML, exemple de clustering : identification des thèmes

Termes spécifiques de la classe 1

	% terme/mod.	% mod.
alcool	27.71	
drogu	20.52	
bagarr	0.97	
jeun	4.94	
chez	1.07	
consomm	0.54	
abus	0.54	
agress	2.36	
physiqu	0.00	
person	0.75	
arrach	0.00	
infract	0.00	
illegal	0.00	
violenc	4.19	
delinqu	1.40	
verbal	0.00	
trafic	1.40	
educ	0.00	
age	0.00	
incivilit	0.86	

Termes spécifiques de la classe 2

	% terme/mod.	% mod./
chomag	0.83	
ecol	0.76	
educ	1.82	
enfant	1.40	
gen	0.86	
insult	0.65	
jeun	3.58	
manqu	3.25	
non	0.98	
nsp	2.51	
parent	0.93	
petit	0.88	
plus	1.19	
proxenet	0.76	
racism	1.02	
racket	0.98	
respect	2.63	
ru	1.00	
securit	0.78	
traffic	0.91	

Termes spécifiques de la clas

	% terme/mod.	%
age	0.90	
agress	6.32	
attaqu	0.46	
bien	0.92	
cambriolag	2.33	
degrad	2.32	
delinqu	3.61	
escroquer	1.80	
financier	1.28	
illegal	1.08	
immigr	1.42	
incivilit	2.62	
infract	1.14	
person	2.83	
physiqu	1.36	
routier	2.02	
verbal	0.99	
manqu	0.85	
respect	0.71	
alcool	0.54	

TM et ML, exemples de classification supervisée

Pourvu que l'on dispose d'un label/classe, on peut également appliquer les algorithmes de classification classique :

- **détecteur de spam**, classification binaire spam / non-spam utilisant les informations textuelles (entêtes, adresses) pour classer les mails.
- **codification automatique**
- identifier les cas de **violence conjugale** à partir du champ textuel des dépots de plainte (travail en cours avec le SSM Intérieur)
- prévoir l'**évolution de l'emploi SMNA à court-terme** à partir des articles de presse (Le Monde), travail en cours avec Thomas Renault et Clément Bortoli
- **analyse de sentiment** (satisfaction, insatisfaction, neutralité) notamment au sujet d'une enquête à partir des commentaires de l'enquête TIC pour les ménages (encadrement d'élèves de 2e année de l'ENSAE)

Liens entre textmining et scraping

On peut scraper toutes sortes de données, **pas nécessairement textuelles**, mais internet regorge de celles-ci…

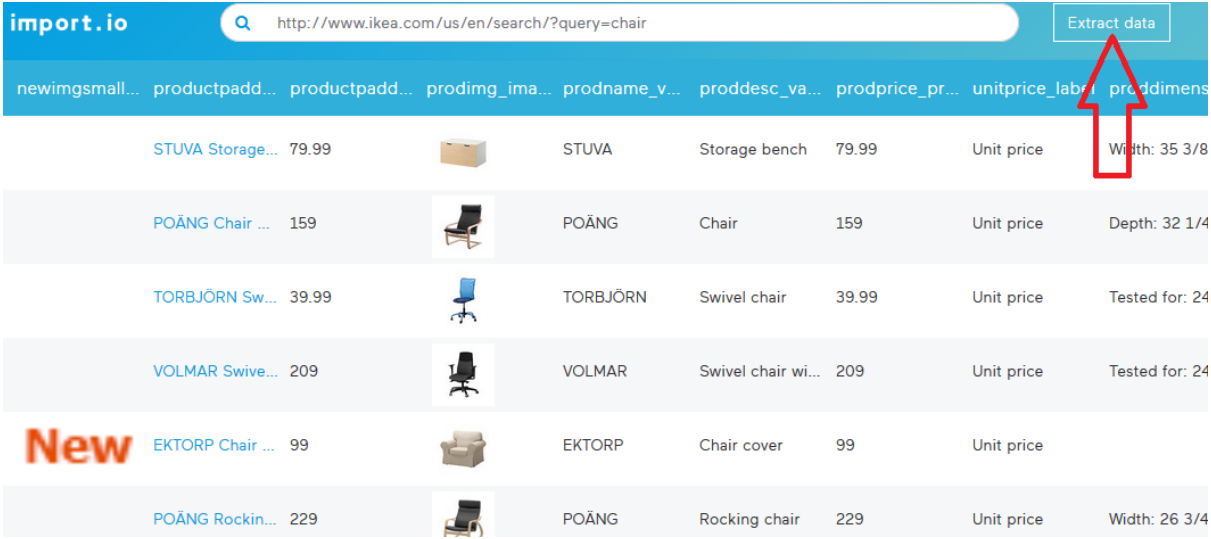
De très nombreux logiciels (**import.io** ou kimonolabs par exemple), gratuits ou payants sont proposés pour scraper internet, mais dans la plupart des cas, ces logiciels n'arrivent à récupérer que des données très bien formatées.

Des **librairies** permettent de coder ses propres robots scraper en R ou en python par exemple.

Attention ! : la pratique du scraping pose des questions éthiques et légales, un scraping de masse pouvant poser des problèmes de trafic pour un site internet, et les conditions générales d'utilisation doivent être respectées.

Remarque : données internet != scraping, il est bien sûr préférable d'établir des partenariats avec les agrégateurs pour obtenir les données directement.

Liens entre textmining et scraping



newimgsmall...	productpadd...	productpadd...	prodimg_ima...	prodname_v...	proddesc_va...	prodprice_pr...	unitprice_labe	proddimens
	STUVA Storage...	79.99		STUVA	Storage bench	79.99	Unit price	Width: 35 3/8
	POÄNG Chair ...	159		POÄNG	Chair	159	Unit price	Depth: 32 1/4
	TORBJÖRN Sw...	39.99		TORBJÖRN	Swivel chair	39.99	Unit price	Tested for: 24
	VOLMAR Swive...	209		VOLMAR	Swivel chair wi...	209	Unit price	Tested for: 24
New	EKTORP Chair ...	99		EKTORP	Chair cover	99	Unit price	
	POÄNG Rockin...	229		POÄNG	Rocking chair	229	Unit price	Width: 26 3/4

Liens entre textmining et scraping, contour de réseaux

L'enquête "Contours de réseaux" 2016 vise à établir un **recensement des réseaux** du commerce alimentaire ainsi que la liste exhaustive des points de vente affiliés.

Un des enjeux majeurs pour garantir la qualité des données de cette opération est de disposer d'une base de sondage précise et exhaustive. Or **il n'existe pas de répertoire des réseaux**. La connaissance des réseaux repose ainsi sur des sources diverses et partielles.

Ces diverses sources permettent d'établir une liste de mots clés correspondant à des enseignes potentielles. Le scraping permet d'**automatiser partiellement les recherches internet** permettant de déterminer qu'un mot clé désigne effectivement un réseau, le cas échéant, on récupère la raison sociale du déposant pour identifier les sites internet portant le nom du réseau.

Cf la prochaine **lettre Big Data** pour plus de détails !

Liens entre textmining et scraping, offres d'emploi

Dans le cas d'usage précédent, l'aspect text-mining est très limité, il sert juste pour un éventuel nettoyage, la recherche de formes, la normalisation des informations scrapées qui peuvent être de format variable.

Un autre projet développé notamment au sein de l'**ESSnet Big Data** (dont DARES), est de scraper des **offres d'emploi** sur des sites d'entreprises ou des agrégateurs dans le but de produire un indicateur.

Le projet est exigeant en termes de **text-mining** et **machine learning**, on va en effet utiliser le texte des offres :

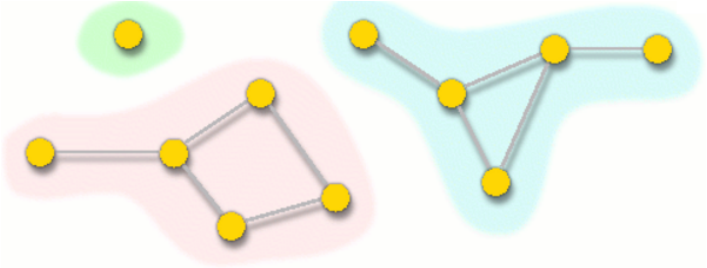
- pour **filtrer** (distinguer vraie offre d'une pub par ex.)
- pour **identifier** et dédoublonner les offres
- pour **catégoriser** les offres ou leur contenu (codification métier, compétence requise etc…)

Exploitation des données en réseau

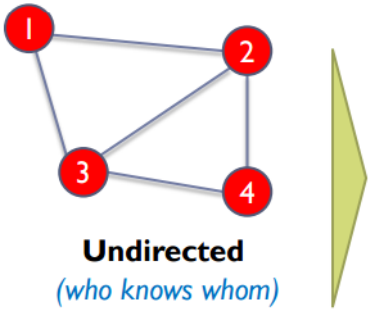
Concept

Pour les données pouvant être représentées sous forme de **graphe** dont les individus constituent les **sommets** (ou noeuds), et leurs relations sont symbolisées par des **arêtes** (ou liens), l'analyse de réseau fournit des indicateurs permettant de décrire :

- le réseau au sens global, caractériser sa **structure** ou son organisation : densité, connexité, communautés…
- les individus au sein du réseau en fonction de leurs intéractions avec les autres noeuds : indicateurs de **centralité** tels que le degré, l'intermédiarité, le pageRank…



Représentation mathématique



Edge list

Vertex	Vertex
1	2
1	3
2	3
2	4
3	4

Adjacency matrix is symmetric

Vertex	1	2	3	4
1	-	1	1	0
2	1	-	1	1
3	1	1	-	1
4	0	1	1	-

Exemple : étude des réseaux de chercheurs

Pour mesurer l'impact sur la visibilité d'un chercheur en Economie de son positionnement au sein de son réseau académique, nous avons :

- Récupéré les données (exhaustives, scrapées) de la base bibliométrique **RePec**, qui nous permet de produire des volumes de citations par année et par article
- Construit des indicateurs décrivant le rôle joué par les auteurs au sein de leur réseau de coauteurs (**analyse des réseaux sociaux**)
- Estimé des **régressions quantiles** pour expliquer le taux de reprise moyen des articles d'un auteur en fonction de ses caractéristiques (distribution à queue épaisse)

Exemple : étude des réseaux de chercheurs

La base RePec contient en effet :

- des relations **auteur-article** (pages auteur), **auteur-auteur** : collaborations (pages article), **article-article** (bibliographies), les dates de publication rendent possible une analyse précise des collaborations et des processus de citations **dans le temps** (< 1960)
- **large couverture** en économie (2 millions de papiers, 2300 journaux, 45000 auteurs) mais un projet collaboratif (volontariat, popularité de l'outil -> effets de bord)
- des caractéristiques individuelles :
  - article : domaine (code JEL), journal, date
  - auteur : affiliation, l'expérience (date de première publication présente dans la base)

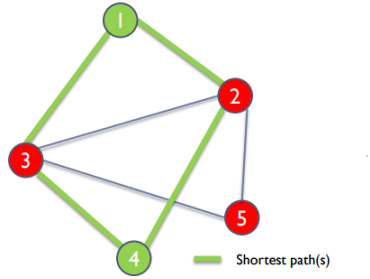
Exemple - description du réseau global des coauteurs

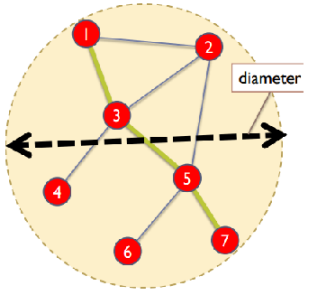
Le graphe des coauteurs correspond au réseau dont les sommets sont les auteurs et les liens les collaborations : il est **non orienté**, **pondéré** et à liens simples.

- Le graphe n'est **pas connexe** : il est constitué de nombreux sommets isolés, quelques dyades, triades ou groupe de quelques dizaines d'auteurs et d'une composante majoritaire qui regroupe de plus en plus d'auteurs

Exemple - le réseau de coauteurs est compact…

- Au sein de la composante majoritaire, les auteurs sont "proches" (en moyenne, il y a **6 degrés de séparation** entre deux auteurs qui ne coécrivent donc pas directement), et le degré de séparation maximal (diamètre) n'est que de 70 (pour plusieurs dizaines de milliers d'auteurs).



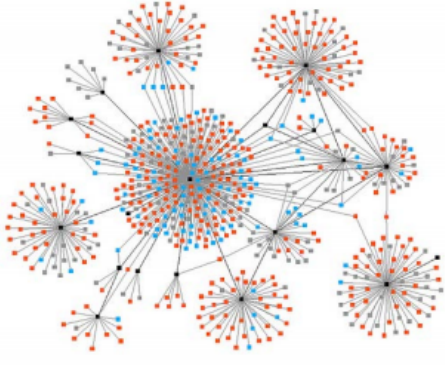


Mais cette apparente **compacité** masque une importante hétérogénéité parmi les auteurs.

Exemple - …mais très peu dense

- dans l'ensemble le nombre moyen de coauteurs par article (<2 en 2010) ou par auteur reste modéré (même s'il est en progression, 4 de moyenne en 2010) : la **densité** du graphe de coauteurs est donc très faible (proportion de liens effectifs sur le nombre de liens possibles)

- finalement, toute la connectivité du graphe s'explique par **quelques auteurs** très productifs avec un réseau de coauteurs très varié, les nouvelles connexions se feront de préférence avec les noeuds bien positionnés.



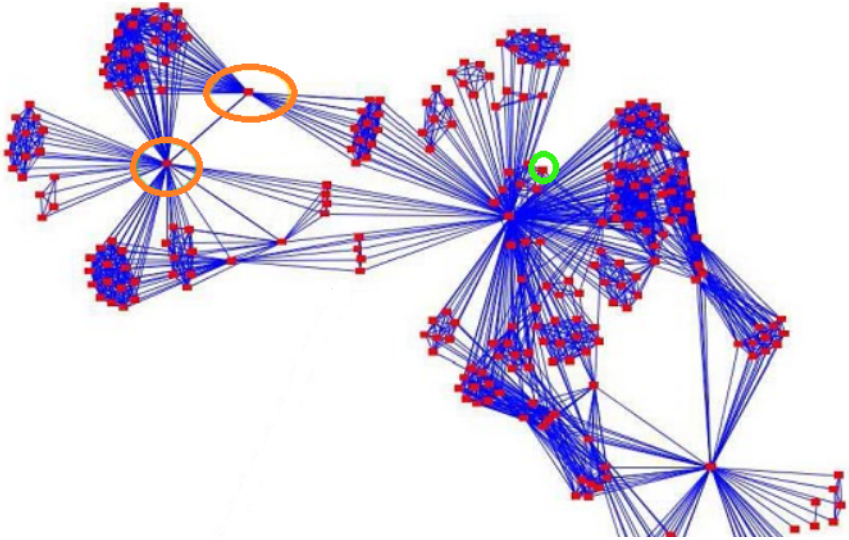
Exemple - description locale du graphe : indicateurs de la centralité

Le **degré** : nombre direct de voisins (ici coauteurs, cf slide précédente)

La centralité de **proximité** prend en compte **tous les sommets** indirectement connectés à un noeud. Elle mesure la moyenne des distances d'un sommet à tous les autres (**vitesse de diffusion** d'une information provenant du sommet dans le reste du réseau).

La centralité d'**intermédiarité** d'un sommet rend compte de son importance dans le maintien de la cohésion du réseau. S'il se trouve sur la plupart des chemins (les plus courts) reliant deux auteurs quelconques du réseau, alors si on le supprimait, on **scinderait** la communauté.

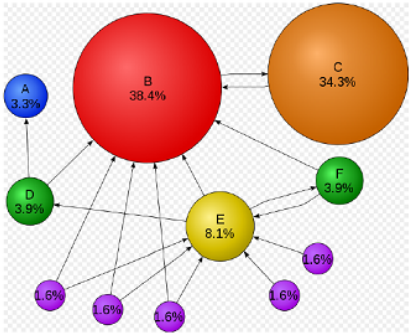
Exemple - d'autres indicateurs de la centralité



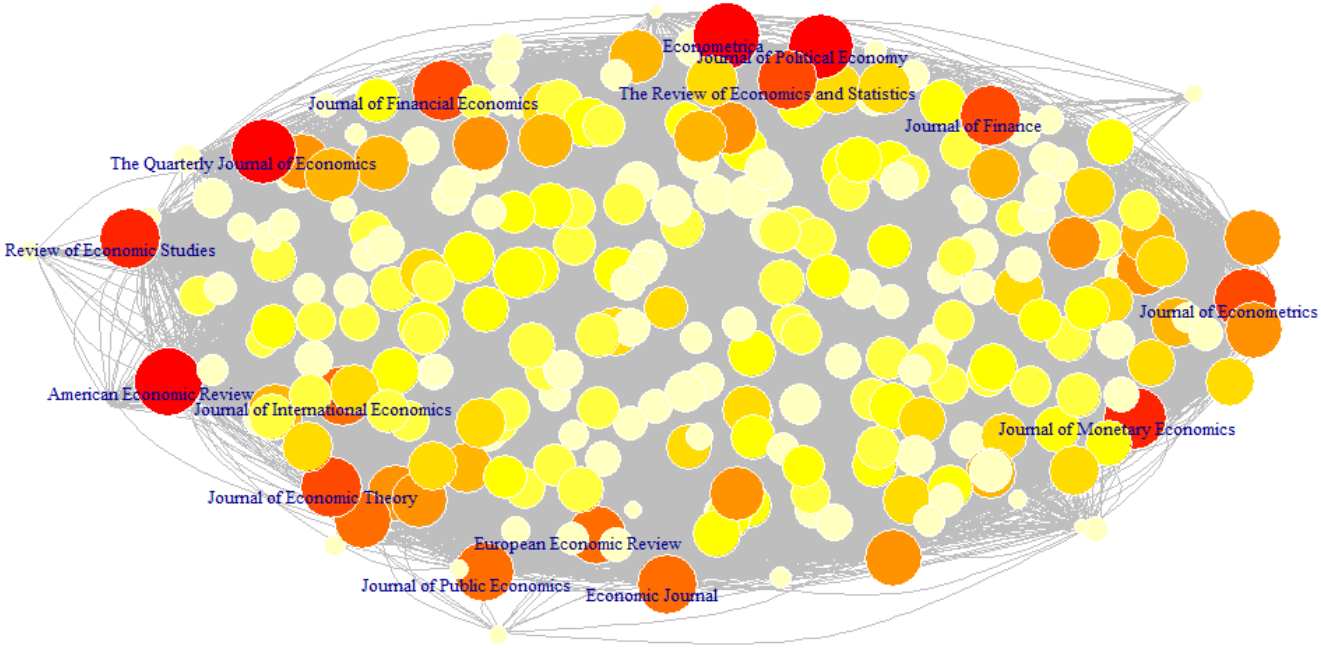
Exemple - le pageRank une mesure de classement

En analyse de réseau, le **PageRank** est classique pour repérer les noeuds influents dans un graphe orienté (en particulier utilisé par Google pour déterminer l'ordre d'apparition des pages web en fonction de leur pertinence suite à une requête).

Calculé sur le réseau orienté des citations, un pageRank élevé signifie qu'un auteur fait d'autant plus autorité qu'il est souvent cité par d'autres chercheurs faisant eux-mêmes autorité.



Exemple - le pageRank une mesure de classement



Exemple - Dynamique du réseau : liens en 1990

Exemple - Dynamique du réseau : liens en 2010

Un exemple typique de projet datascience

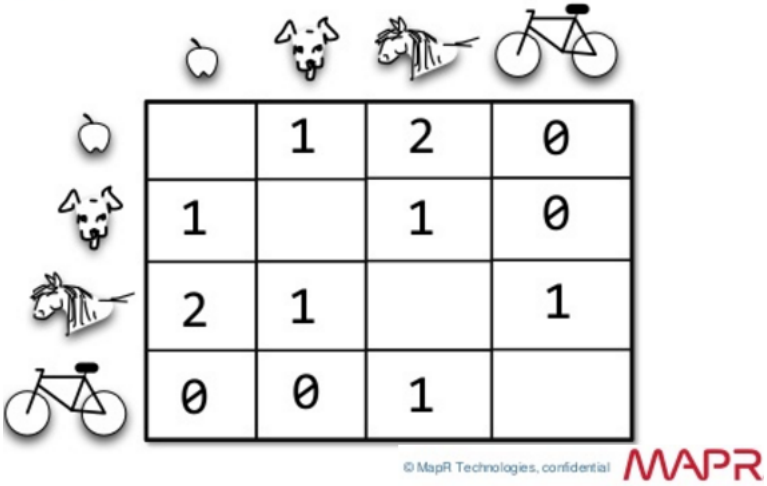
- **scraping** pour la collecte de données
- **Machine Learning** pour l'imputation du genre des auteurs sur les noms prénoms
- **analyse de réseaux** pour l'exploration et l'extraction d'indicateurs
- **visualisation**
- **économétrie**

Les données représentées sous forme de réseaux peuvent également servir d'input à des algorithmes de machine learning (clustering, prediction) :

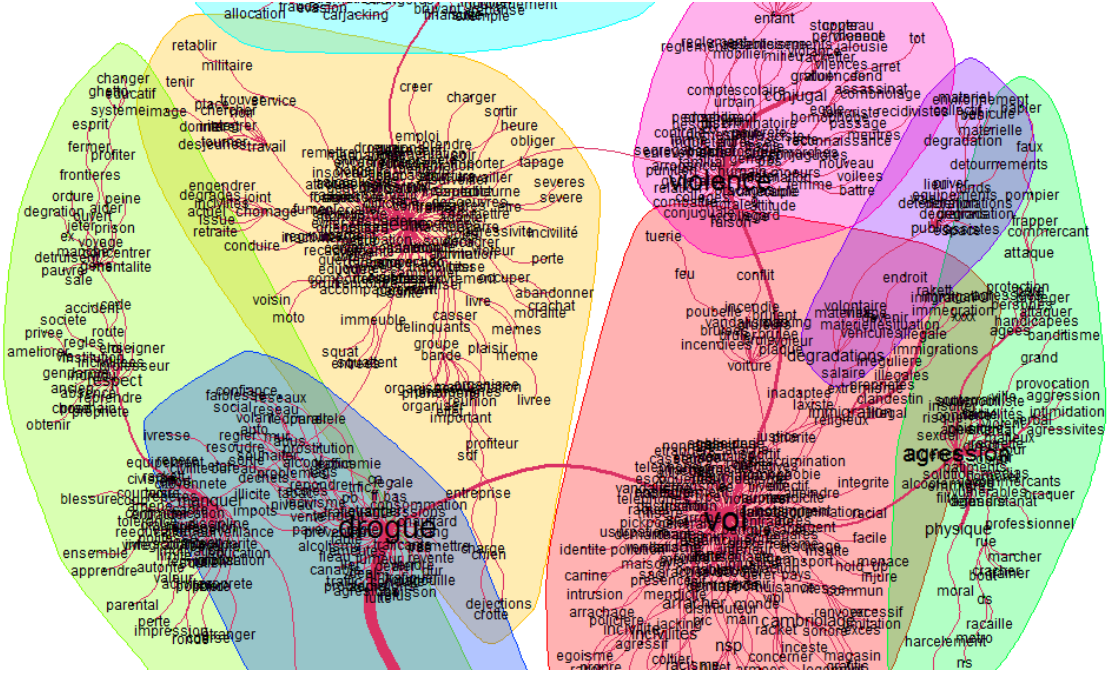
tout est lié !

Lien réseaux et ML : exemple de clustering de graphes

A partir des réponses à la question ouverte : **Pouvez-vous citer les formes de délinquance, les phénomènes qu'il faudrait traiter en priorité dans la société [française] actuelle ?** (CVS), on peut extraire une **matrice de cooccurrences** qui peut être considérée comme une matrice d'adjacence.



Lien réseaux et ML : exemple de clustering de graphes



Visualisation de graphes

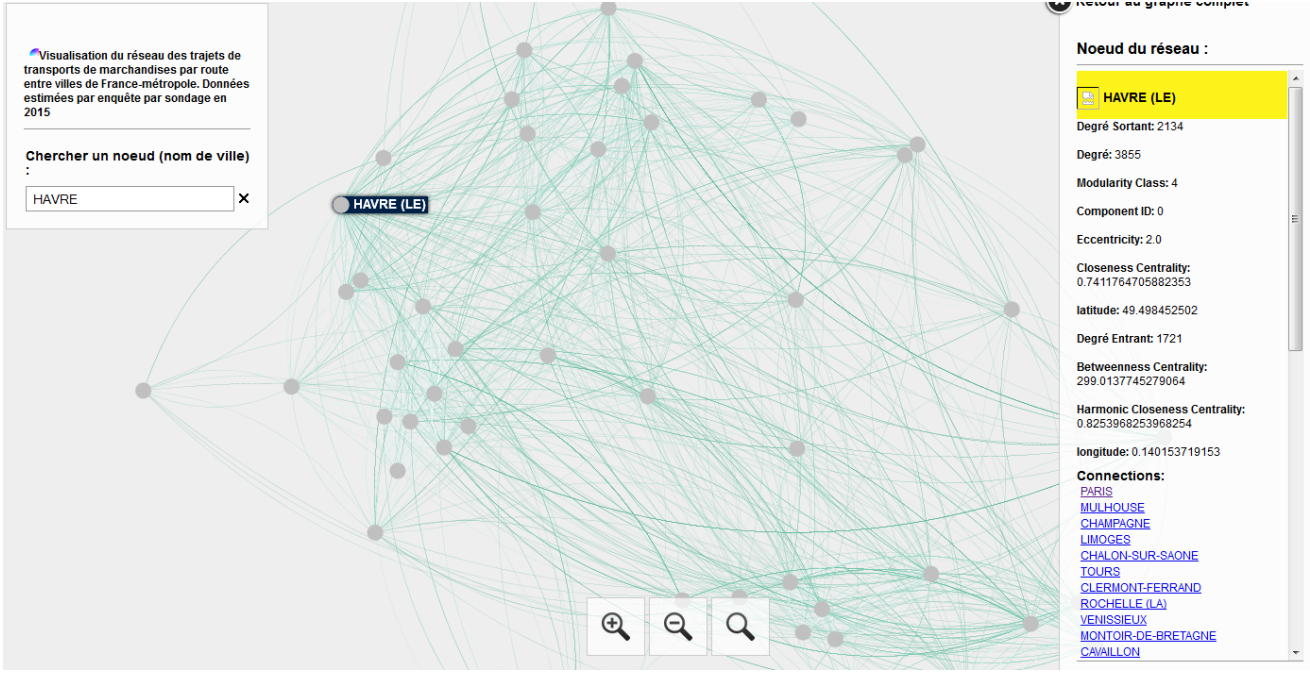
Un des outils les plus mis en avant pour l'exploration de réseaux : **Gephi**, il permet de :

- visualiser un graphe, de facilement filtrer sur les noeuds d'intérêt,
- de calculer les différents indicateurs mentionnés ou de clusteriser.

Tutoriel sur le blog statoscope par Robert Pastorelli sur des données de flux (Enquête transport routier de marchandises (TRM))

		destinations			
		A	B	C	D
origins	A	0	5	7	8
	B	5	0	16	11
	C	7	16	0	20

Visualisation de graphes



Travaux en cours

- dans le prolongement de la **collaboration avec Orange**, on va s'intéresser aux flux de personnes pour analyser la mobilité : **clustering de graphes** pour analyser la **ségrégation urbaine** ?

- encadrement d'un groupe d'élèves de l'ENSAE sur l'**étude des réseaux de professionnels de la santé** (données de la base TRANSPARENCE SANTE), étude des communautés et réseaux de laboratoires-professionnels de santé via les avantages/conventions déclarés.

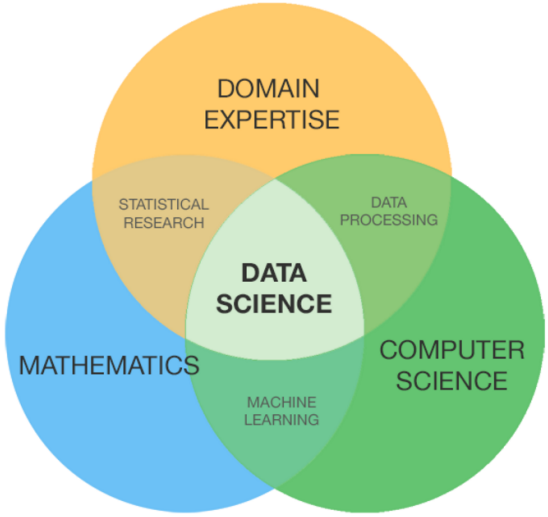
Ressources : **document de travail** sur le partitionnement de graphes (Jean-Michel Floch DAR, Pascal Eusebio, David Levy PSAR-AT), le clustering de graphes appliqué à l'urbanisation

Questions

Des questions ?

Contact : stephanie.combes@insee.fr; pauline.givord@insee.fr; benjamin.sakarovitch@insee.fr

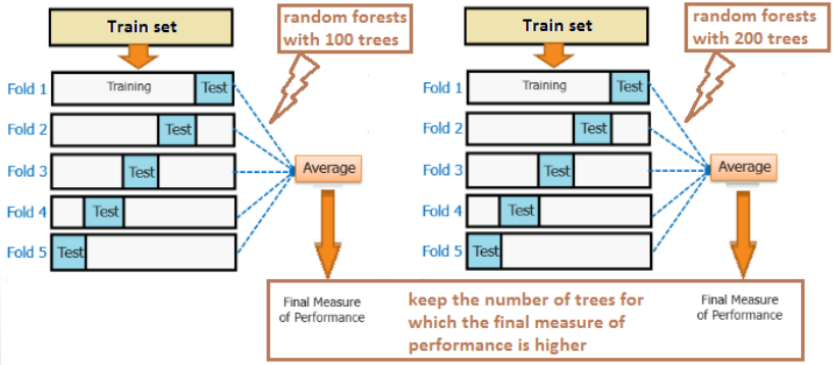
Info : lettre du Big Data et statistiques publiques, yammer, statoscope



Annexe

Validation croisée

It is usually done by **cross validation**: the training set is separated again in a partition of samples (*folds*), each one is used as a test set once to produce a robust measure of the performance for one method with one fixed parameter.



INSEE

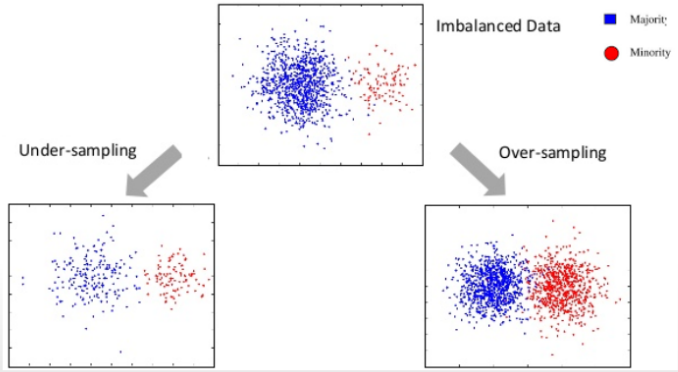
30/

Classes desequilibrées

Our data are characterized by imbalanced distribution (antennas and municipalities): 50% of antennas are in major urban centers.

Two ways to deal with this:

- sampling the data during step 2, ie **artificially balance the dataset** over classes by removing or adding observations.



LASSO

In dimension reduction problems, if we suspect that some variables are more important than others, the emphasis is put on identifying them:

- ***step by step* algorithms** which add (respectively remove) a certain number of variables from an initial empty (respectively full) linear model, on the basis of a significance criterion

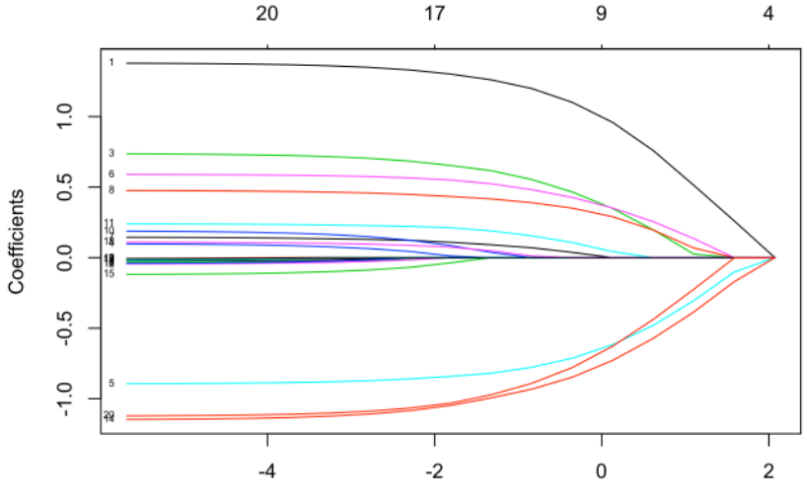
The main risk is to select a model which is far from the best possible model

- More robust approaches in terms of optimization have been developed: such as penalized regressions (LASSO, **Elastic Net**). They incorporate a **penalty term** in the objective function (less sensitive to data used for estimation).

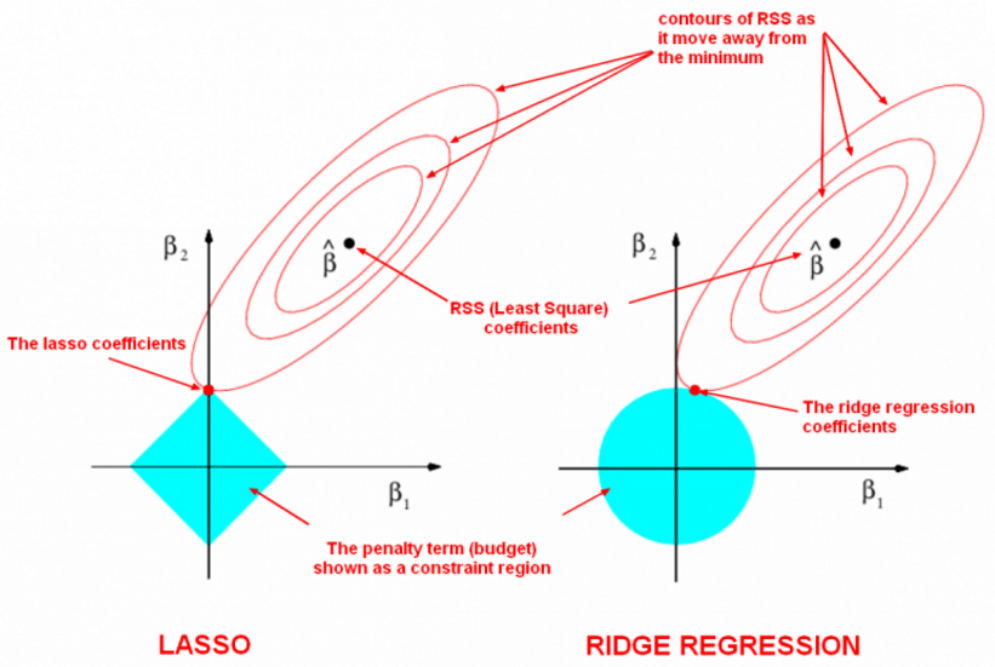
$$\forall \lambda > 0 \ \hat{\beta}^{ElasticNet} = \arg\min_{\beta \in \mathbb{R}^p} ||Y - X\beta||_2^2 + \lambda \left[\alpha |\beta|_1 + (1-\alpha)/2 ||\beta||_2^2 \right]$$

LASSO

Parameters are chosen to optimize the performances of the model. OLS solution is obtained when lambda equals to zero, increasing lambda will allow for sparser and therefore more robust solutions.



LASSO


contours of RSS as
it move away from
the minimum
β_2
$\hat{\beta}$
RSS (Least Square)
coefficients
The lasso coefficients
β_1
The penalty term (budget)
shown as a constraint region
LASSO
β_2
$\hat{\beta}$
The ridge regression
coefficients
β_1
RIDGE REGRESSION