

Investigating Gene- and Pathway-environment Interaction analysis approaches

Camilo Broc¹ Marina Evangelou² Thérèse Truong³ Benoit
Liquet¹

¹LMAP (UPPA)

²Imperial College

³INSERM

June 28, 2017

Investigating Gene- and Pathway-environment Interaction analysis approaches

Camilo Broc¹ Marina Evangelou² Thérèse Truong³ Benoit
Liquet¹

¹LMAP (UPPA)

²Imperial College

³INSERM

June 28, 2017

Introduction

Objective : Adapt and compare 2 existing methods Fisher and ARTP to interaction test

- Statistic content

- Interaction test :

$$\text{logit}[P(Y = 1|SNP_\ell, E)] = \alpha_\ell + \beta_\ell SNP_\ell + \beta_{E,\ell} E + \gamma_\ell E \times SNP_\ell$$

- Resamplings
 - p-value combination :

- Data structure constraints

- SNPs (Single Nucleotide polymorphism)
 - Gene : set of SNPs
 - pathway : set of pathway
 - environment data

- Respect previous ARTP package structure

Introduction

Objective : Adapt and compare 2 existing methods Fisher and ARTP to interaction test

- Statistic content

- Interaction test :

$$\text{logit}[P(Y = 1|SNP_\ell, E) = \alpha_\ell + \beta_\ell SNP_\ell + \beta_{E,\ell} E + \gamma_\ell E \times SNP_\ell]$$

- Resamplings
 - p-value combination :

- Data structure constraints

- "Omics" data
 - SNPs (Single Nucleotide polymorphism)
 - Gene : set of SNPs
 - pathway : set of pathway
 - environment data

- Respect previous ARTP package structure

Procedure

For SNP_ℓ

- Fit the model under the null hypothesis $H_{0,\ell}$ on original data → statistic test τ_0
- Perform resamplings according to null hypothesis (permutation or bootstrap)
- Fit model on resamplings → statistic test τ_1, \dots, τ_N
- Compute p-value comparing original data and resampled data statistic tests → $\frac{\sum_{n=1}^N (\tau_n \geq \tau_0)}{B+1}$

Type of statistical test

- $\text{logit}[P(Y = 1|SNP_\ell, E)] = \alpha_\ell + \beta_\ell SNP_\ell + \beta_{E,\ell} E + \gamma_\ell E \times SNP_\ell$
- For binary outcome : Logistic model and Likelihood Ratio Test
example : disease diagnostic
- For continuous outcome : Linear regression and Wald test
Example : survival data

P-value combination

- For N resamplings and k independent variables $N \times k$ pvalues performed
- Fischer method :

$$FM = -2 \sum_{\ell=1}^L \log(p_\ell)$$

- ARTP (Adaptive Rank Truncating Product) :

$$W_K = \sum_{k=1}^K \log(p_k) = \log \left(\prod_{k=1}^K p_{(k)} \right)$$

$$\hat{s}_j^{(b)} = \frac{\sum_{b^*=0}^B I\left(W_j^{(b^*)} \leq W_j^{(b)}\right)}{B+1}$$

Resampling

- Permutation
 - Permute outcome among all observations
 - Permute outcome within levels of environment variable.
 - Permute simultaneously outcome and an environment term
- Parametric bootstrap

Discussion : Permutation not always feasible

Syntax - Data structure

SNP-Gene-pathway structure

```
> list.gene.snp  
$gene1  
[1] "rs10904382" "rs11252861" "rs2904802"  
[4] "rs2904803" "rs3930966" "rs4445550"  
[7] "rs7908994" "rs11252885" "rs11816204"  
  
$gene2  
[1] "rs11252885" "rs11816204" "rs12414884"  
[4] "rs17134158" "rs1937855" "rs1937863"  
[7] "rs1937868" "rs1937888" "rs1937889"  
[10] "rs28488494" "rs2854466" "rs2854482"  
[13] "rs2854494" "rs4143630" "rs7082231"  
[16] "rs7083583" "rs7099721" "rs7915365"  
  
$gene3  
[1] "rs10904416" "rs12242350" "rs17396032"  
[4] "rs2275928" "rs2298305" "rs2518049"  
[7] "rs4242785" "rs4559587"
```

```
> data.pathway  
      path1 path2  
gene1      1     0  
gene2      1     0  
gene3      1     0  
gene4      1     0  
gene5      1     0  
gene6      0     1  
gene7      0     1  
gene8      0     1  
gene9      0     1  
gene10     1     0  
gene11     1     0  
gene12     1     0
```

Syntax - Resampling

Resampling notable arguments

- Outcome.model : binary, continuous
- var.inter : interaction to test
- class.inter : type of test
- method : type of permutation
- nbcpu : for parallelism
- Npermut : Number of resamplings

ARTP notable arguments

- SNP selection options
- Gene selection options

Application

- General information
 - 1126 breast cancer cases and 1174 controls
 - Interaction with Night work/Cancer
- Data structure
 - Circadian Pathway
 - 23 Genes
 - 577 SNPs
- Objective : estimate the interaction between night work and circadian genes in breast cancer risk

Application

Circadian Pathway		FM		ARTP	
Gene	Size	Perm.	Bootstrap	Perm.	Bootstrap
ARNTL	24	0.030	0.001	0.047	0.001
PER1	5	0.094	0.039	0.051	0.069
Pathway		0.517	0.014	0.720	0.038

Perspectives

- Improve manual
- Apply to data sets with several pathways
- Compare performance with existing methods

Thank you for your attention!



FactoInvestigate

Description automatique de résultats d'analyse factorielle

Simon THULEAU
Eulidia



François HUSSON
Agrocampus Ouest



FactoInvestigate

Description automatique de résultats d'analyse factorielle

Simon THULEAU
Eulidia



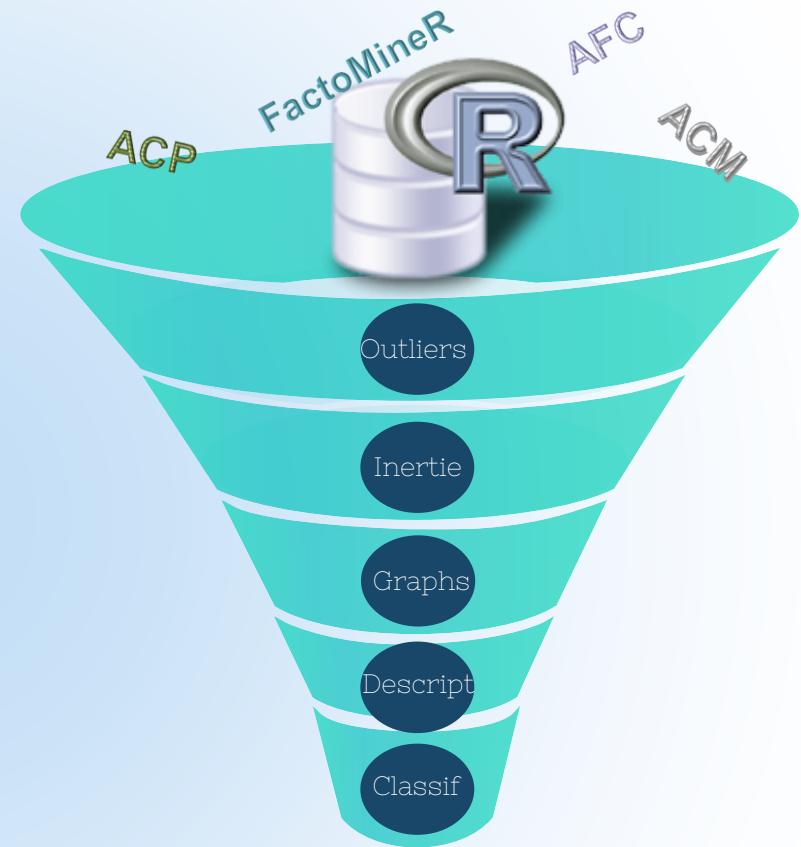
François HUSSON
Agrocampus Ouest



Investigate()

APERÇU :

- Fonction centrale, synthétique
- Sous-fonctions indépendantes
- Continuité de FactoMineR
- Aide à l'interprétation

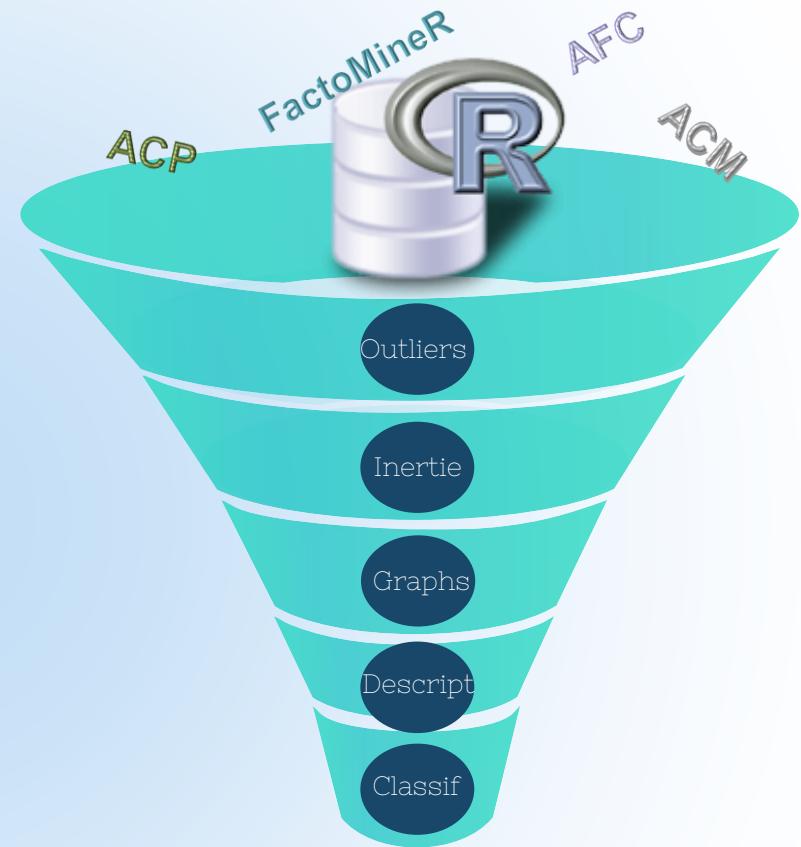




Investigate()

APERÇU :

- Fonction centrale, synthétique
- Sous-fonctions indépendantes
- Continuité de FactoMineR
- Aide à l'interprétation

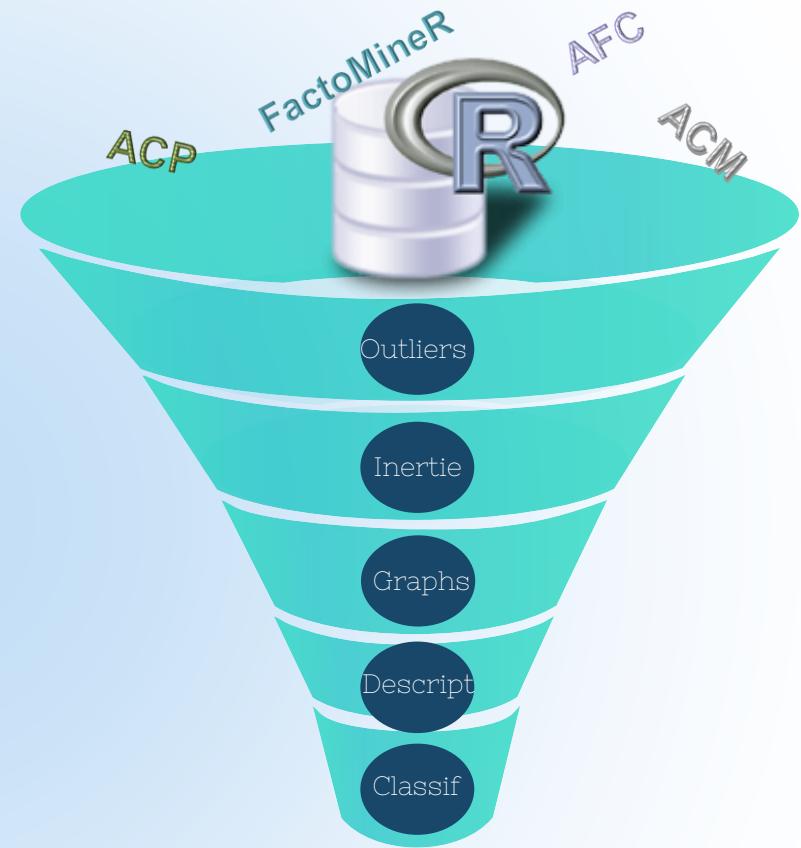




Investigate()

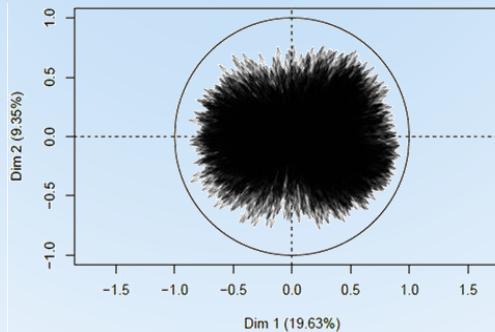
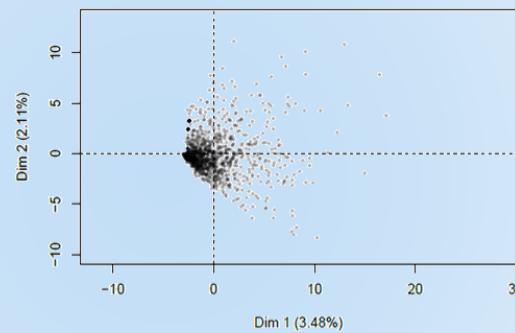
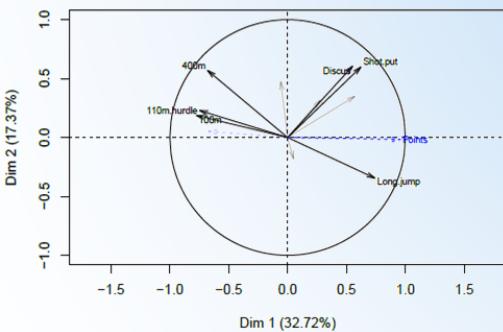
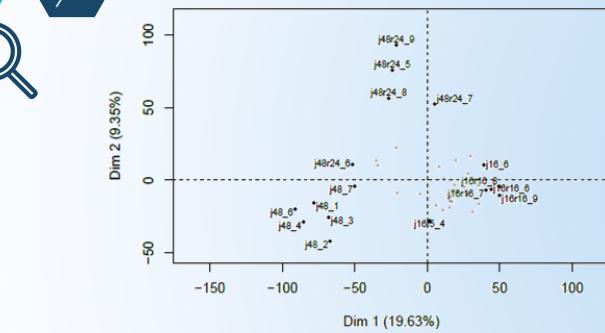
APERÇU :

- Fonction centrale, synthétique
- Sous-fonctions indépendantes
- Continuité de FactoMineR
- Aide à l'interprétation





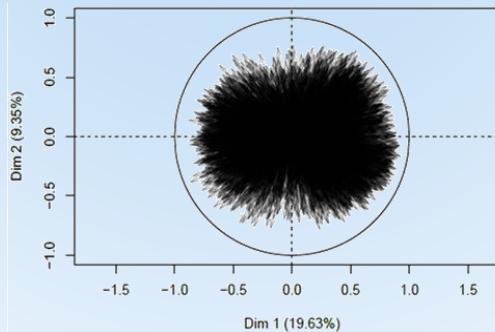
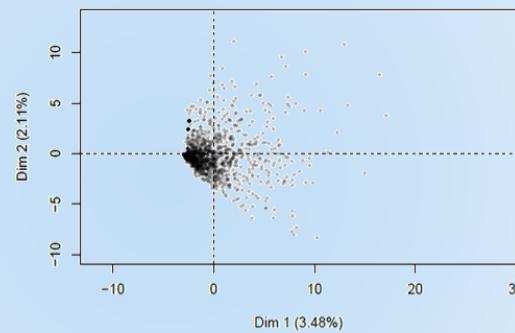
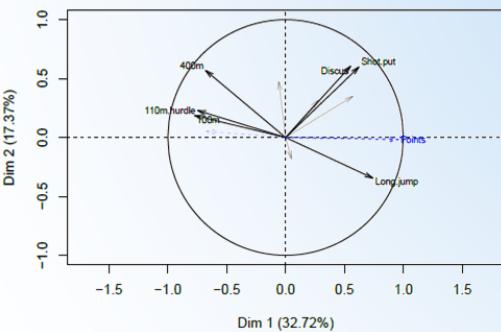
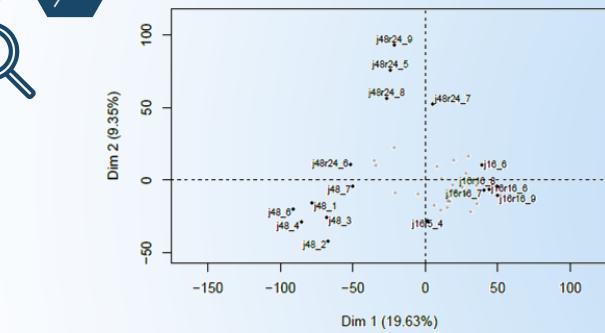
Description des nuages



- Adapté aux jeux de données de toutes tailles
- Clustering des individus les mieux représentés (révèle les groupes caractéristiques des dimensions)
- Pour chaque groupe révélé, on explicite :
 - sa composition
 - son positionnement par rapport aux dimensions
 - ses principales variables caractéristiques



Description des nuages

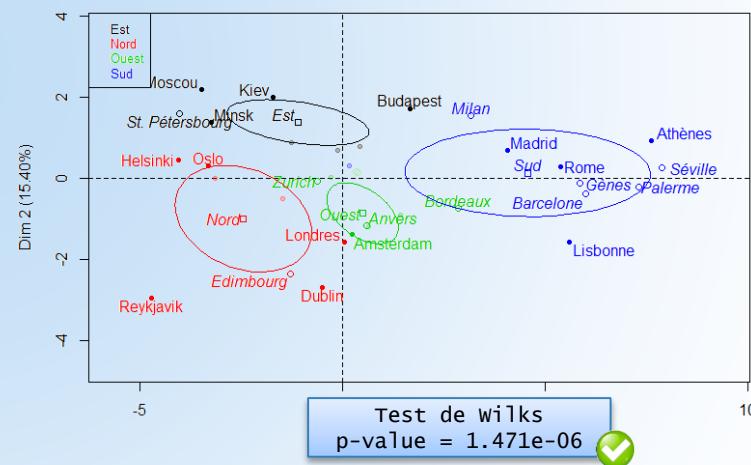
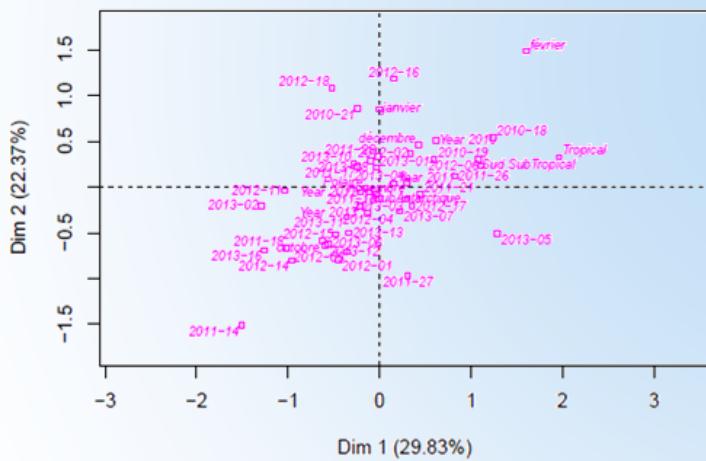


- Adapté aux jeux de données de toutes tailles
- Clustering des individus les mieux représentés (révèle les groupes caractéristiques des dimensions)
- Pour chaque groupe révélé, on explicite :
 - sa composition
 - son positionnement par rapport aux dimensions
 - ses principales variables caractéristiques



Variables illustratives

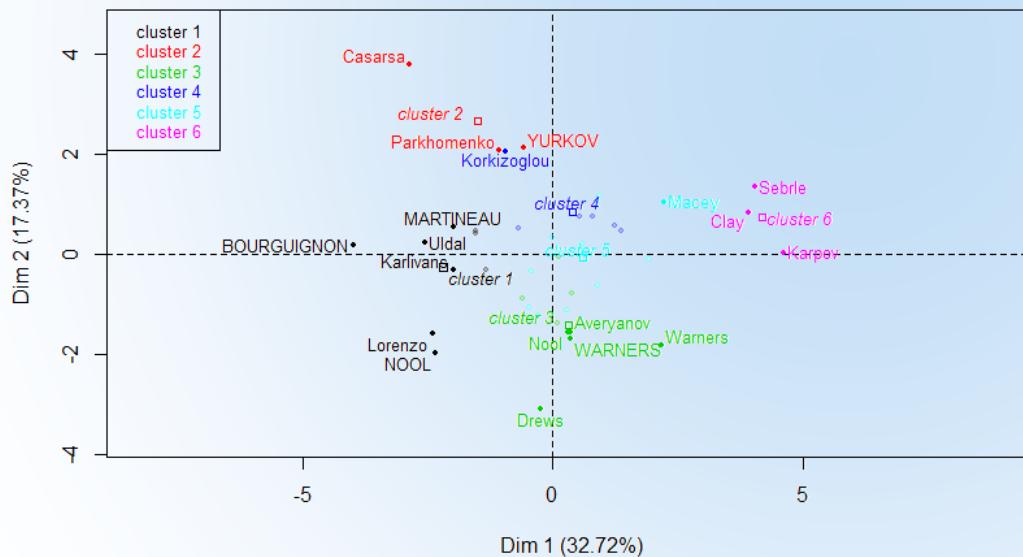
- Variables qualitatives : habillage du nuage des individus



- Variables quantitative : projetées sur le cercle des corrélations



Classification et Annexes



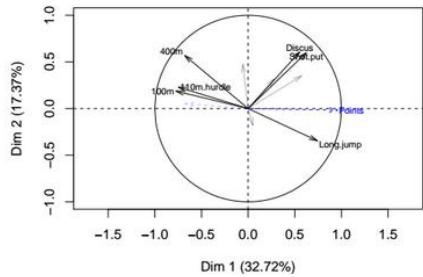
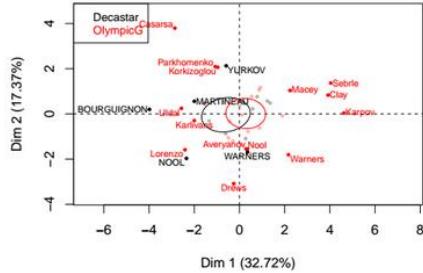
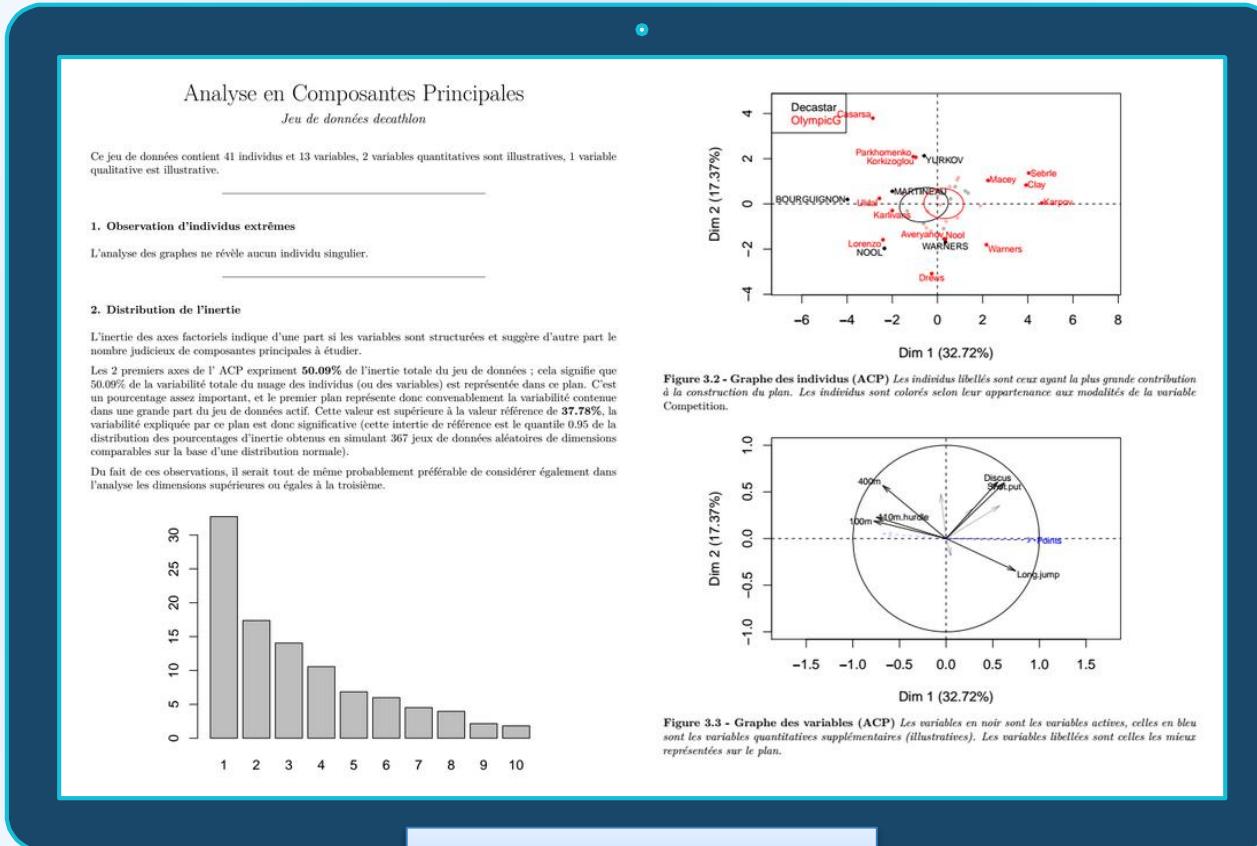
```
dimdesc(res, axes = 1:3)
```

```
$Dim.1
$Dim.1$quantitative
correlation p.value
Points      0.9561543 2.099191e-22
Long.jump   0.7418997 2.849886e-08
Shot.put    0.6225026 1.388321e-05
High.jump   0.5719453 9.362285e-05
Discus     0.5524665 1.802220e-04
Rank        -0.6705104 1.616348e-06
400m       -0.6796099 1.028175e-06
110m.hurdle -0.7462453 2.136962e-08
100m      -0.7747198 2.778467e-09
```

```
$Dim.2
$Dim.2$quantitative
correlation p.value
Discus      0.6063134 2.650745e-05
Shot.put    0.5983033 3.603567e-05
400m        0.5694378 1.020941e-04
1500m       0.4742238 1.734405e-03
High.jump   0.3502936 2.475025e-02
Javeline    0.3169891 4.344974e-02
Long.jump  -0.3454213 2.696969e-02
```

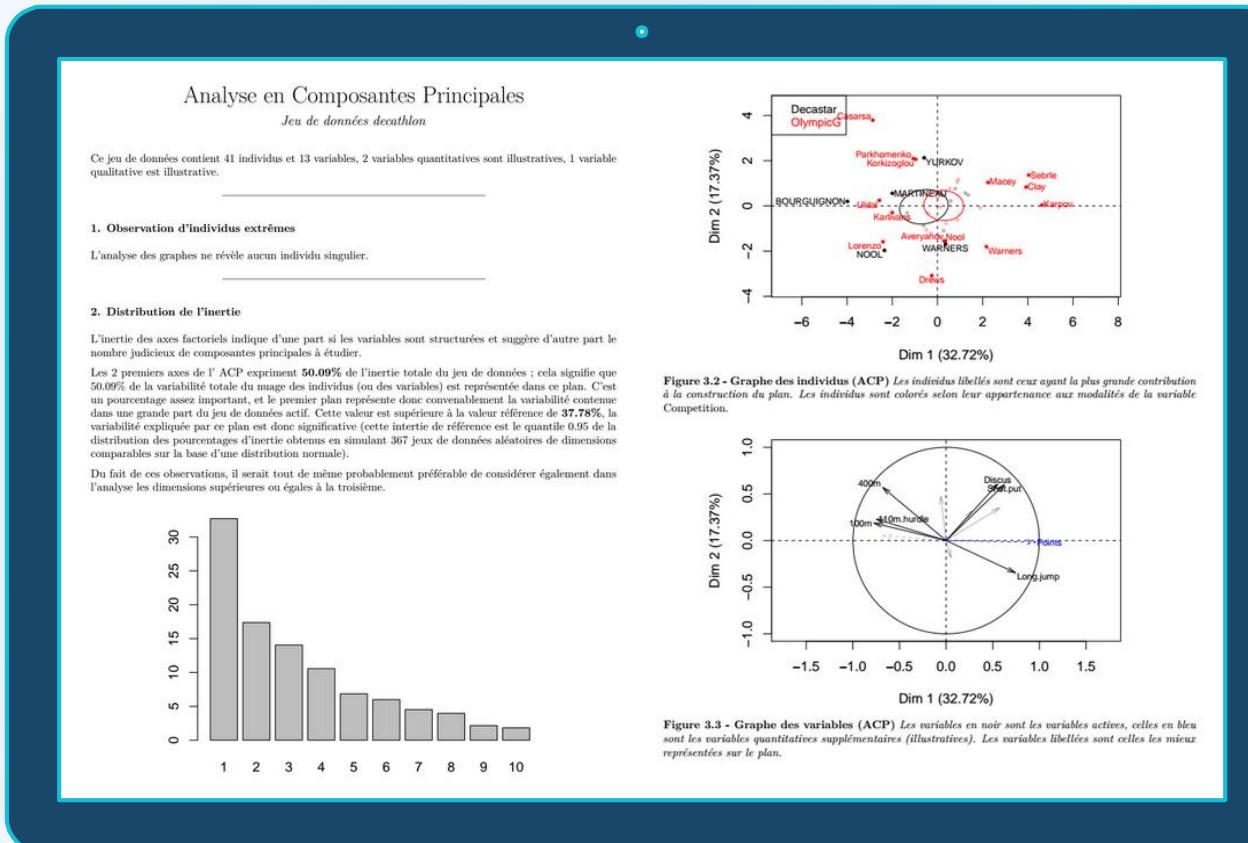
```
$Dim.3
$Dim.3$quantitative
correlation p.value
1500m      0.7821428 1.554450e-09
Pole.vault  0.6917567 5.480172e-07
Javeline   -0.3896554 1.179331e-02
```

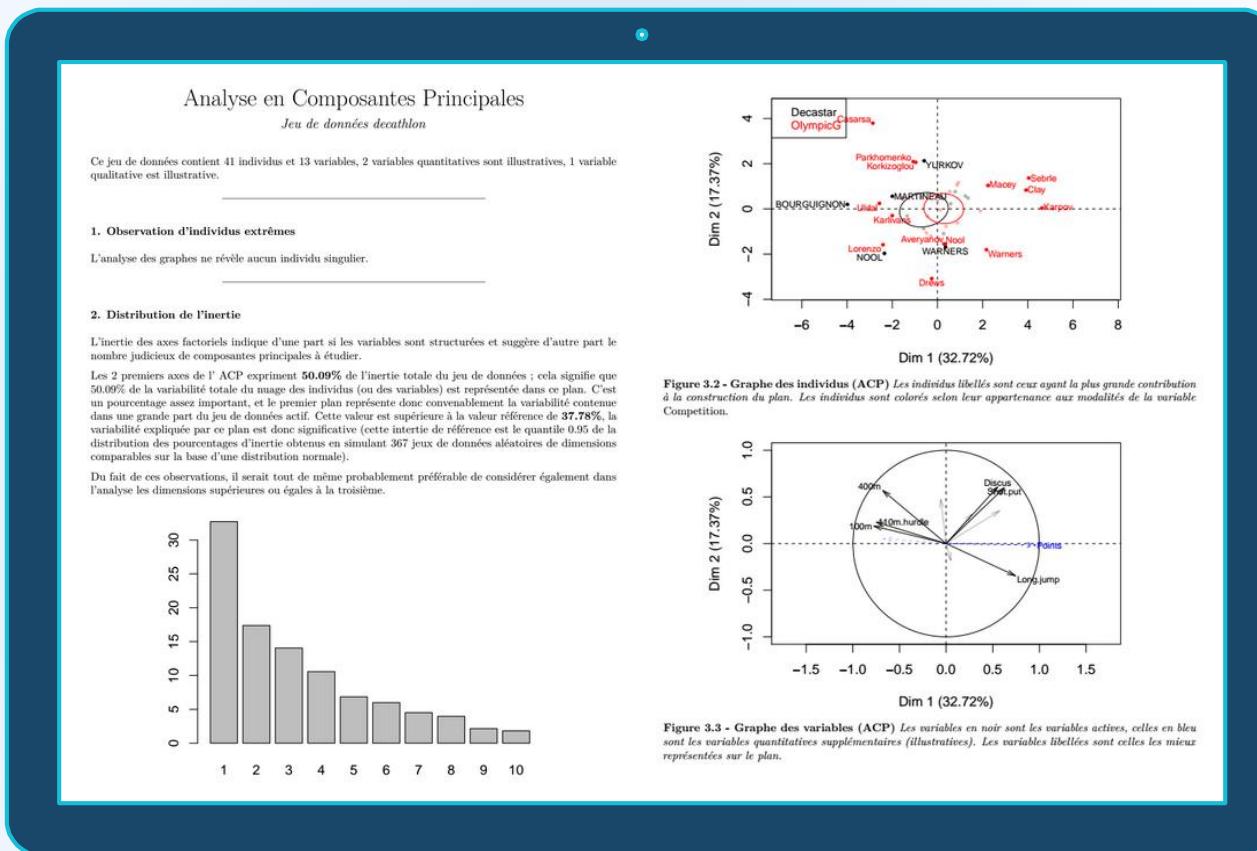
Figure 6 - Liste des variables caractéristiques des dimensions de l'analyse.



Investigate(res.pca)







Analyse en Composantes Principales

Jeu de données decathlon

Ce jeu de données contient 41 individus et 13 variables, 2 variables quantitatives sont illustratives, 1 variable qualitative est illustrative.

1. Observation d'individus extrêmes

L'analyse des graphes ne révèle aucun individu singulier.

2. Distribution de l'inertie

L'inertie des axes factoriels indique d'une part si les variables sont structurées et suggère d'autre part le nombre judicieux de composantes principales à étudier.

Les 2 premiers axes de l'ACP expriment **50,09%** de l'inertie totale du jeu de données ; cela signifie que 50,09% de la variabilité totale du nuage des individus (ou des variables) est représentée dans ce plan. C'est un pourcentage assez important, et le premier plan représente donc convenablement la variabilité contenue dans une grande partie du jeu de données actif. Cette valeur est supérieure à la valeur référence de **37,78%**, la variabilité expliquée par ce plan est donc significative (cette inertie de référence est le quantile 0,95 de la distribution des pourcentages d'inertie obtenus en simulant 367 jeux de données aléatoires de dimensions comparables sur la base d'une distribution normale).

Du fait de ces observations, il serait tout de même probablement préférable de considérer également dans l'analyse les dimensions supérieures ou égales à la troisième.

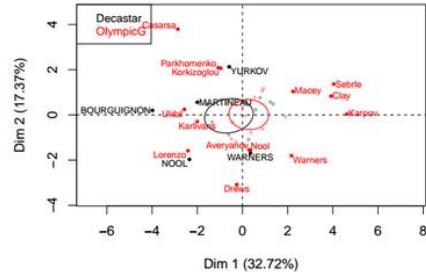
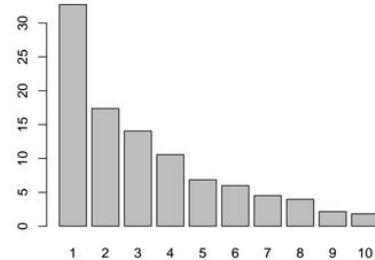


Figure 3.2 - Graphe des individus (ACP) Les individus libellés sont ceux ayant la plus grande contribution à la construction du plan. Les individus sont colorés selon leur appartenance aux modalités de la variable Competition.

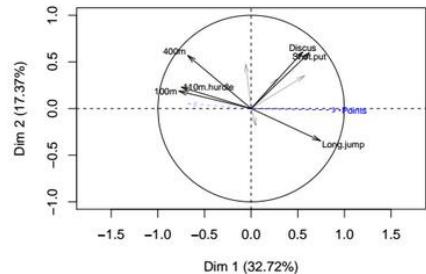


Figure 3.3 - Graphe des variables (ACP) Les variables en noir sont les variables actives, celles en bleu sont les variables quantitatives supplémentaires (illustratives). Les variables libellées sont celles les mieux représentées sur le plan.





Thanks!

Any questions?

factoinvestigate@gmail.com

Simon THULEAU
Eulidia

François HUSSON
Agrocampus Ouest



VALORISATION DES DONNÉES DE LOGS VIA R

SOMFY - « Connected Solutions »

Alison PATOU, Data Scientist

Alison.Patou@keyrus.com

28/06/2017



somfy.[®]

VALORISATION DES DONNÉES DE LOGS VIA R

SOMFY - « Connected Solutions »

Alison PATOU, Data Scientist

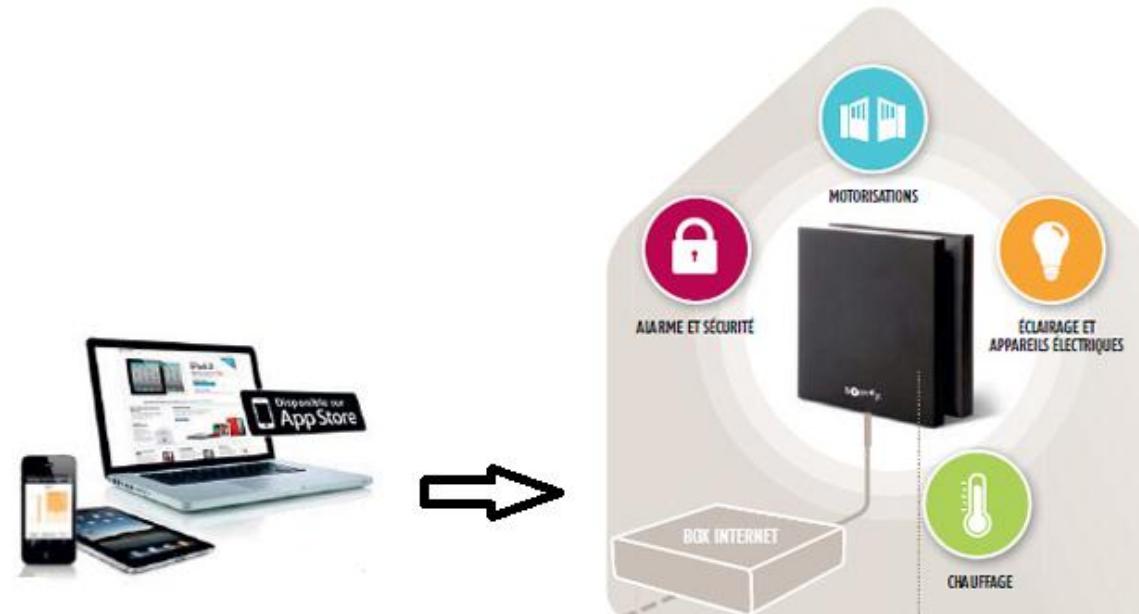
Alison.Patou@keyrus.com

28/06/2017



somfy.[®]

somfy® spécialisé dans la domotique



Un ordre est donné via le mobile par exemple, l'action est effectuée et l'information est enregistrée et stockée en log

CONTEXTE

- **Contexte :**

Exploiter ses données d'usages (provenant de logs) et impliquer les métiers.

A ce jour, très faible connaissance de leurs données, juste des « intuitions »

- **Attendus :**

Statistiques simples et valorisation des données

Outil permettant de visualiser les données

PLAN

I. Préparation des logs

II. Segmentation des usages

III. Analyse sémantique

IV. R Shiny

V. Next Step



I. Préparation des logs

A PERIMETRE

Sources de données (via Overkiz) :

- Overkiz décrypte et simplifie les logs pour nous fournir en sortie des fichiers Excel
- 3 fichiers donnant des informations sur le type de produits/exécutions/pays/dates/ ...

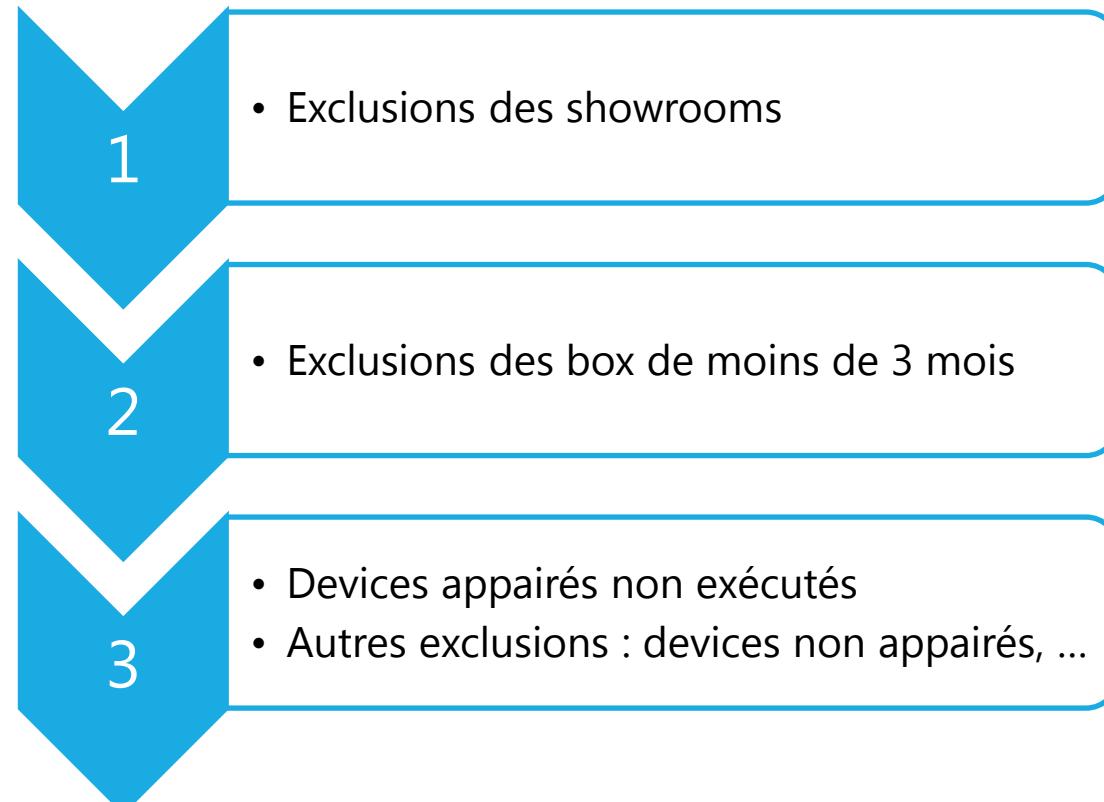
La CNIL autorise seulement à garder 3 mois d'historique et les données sont évidemment anonymisées.

Volume : 5 Go de données // Plusieurs millions de lignes

B FILTRAGES

L'étude portait sur une population de box homogènes. Aussi, nous avons du réaliser des premiers filtres pour exclure les box spécifiques.

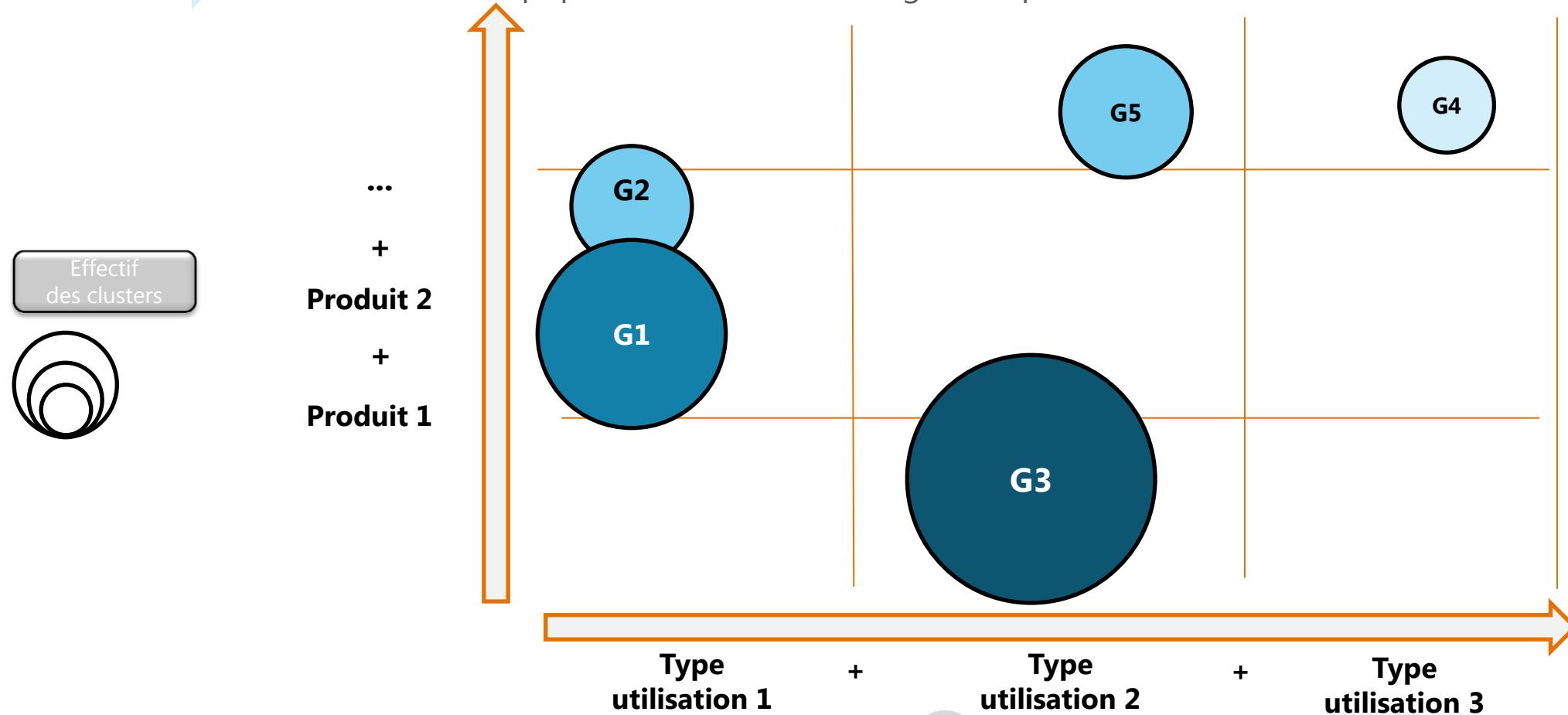
Les différentes étapes d'exclusions ont été faites en amont via un **script R**:



II. Classification des usages

Allocation des box dans le cluster le plus pertinent

Classification de la population selon leurs usages des produits



Allocation des box dans le cluster le plus pertinent

Classification de la population selon leurs usages des produits

Démarche analytique :

a) Détection d'axes discriminants

- Création de d'axes avec analyse discriminante sur l'échantillon (*linear discriminant analysis*)

b) Sélection de l'algorithme de réallocation des box à partir des axes discriminants

- Analyse discriminante
- Arbre de décision : *package tree*
- Forêts aléatoires : *package randomForest*
- K plus Proches Voisins : *package stats*

III. Analyse sémantique

Analyse sémantique des personnalisations

→ Préparation des données

T
R
A
I
T
E
M
E
N
T

- Suppression des chiffres
- Uniformisation en minuscule
- Enlever la ponctuation
- Suppression espaces superflus
- Remplacer/supprimer les caractères spéciaux
- ...



Packages :

- ✓ Tm
- ✓ Stringr (word())

III. Analyse sémantique

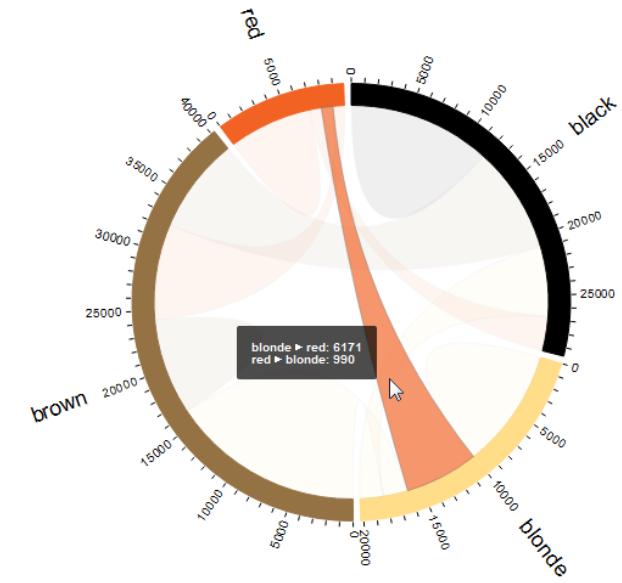
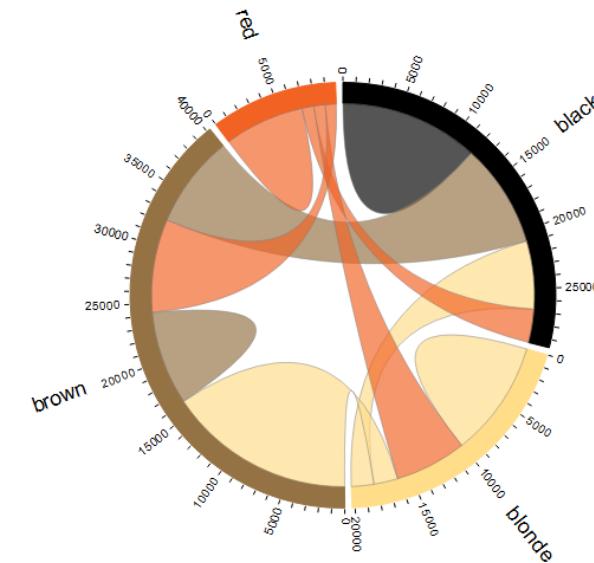
Analyse sémantique des personnalisations

→ Représentations et visualisations des résultats

Package ‘wordcloud’



Package 'chorddiag'



A CONTEXTE CLIENT

Problématiques

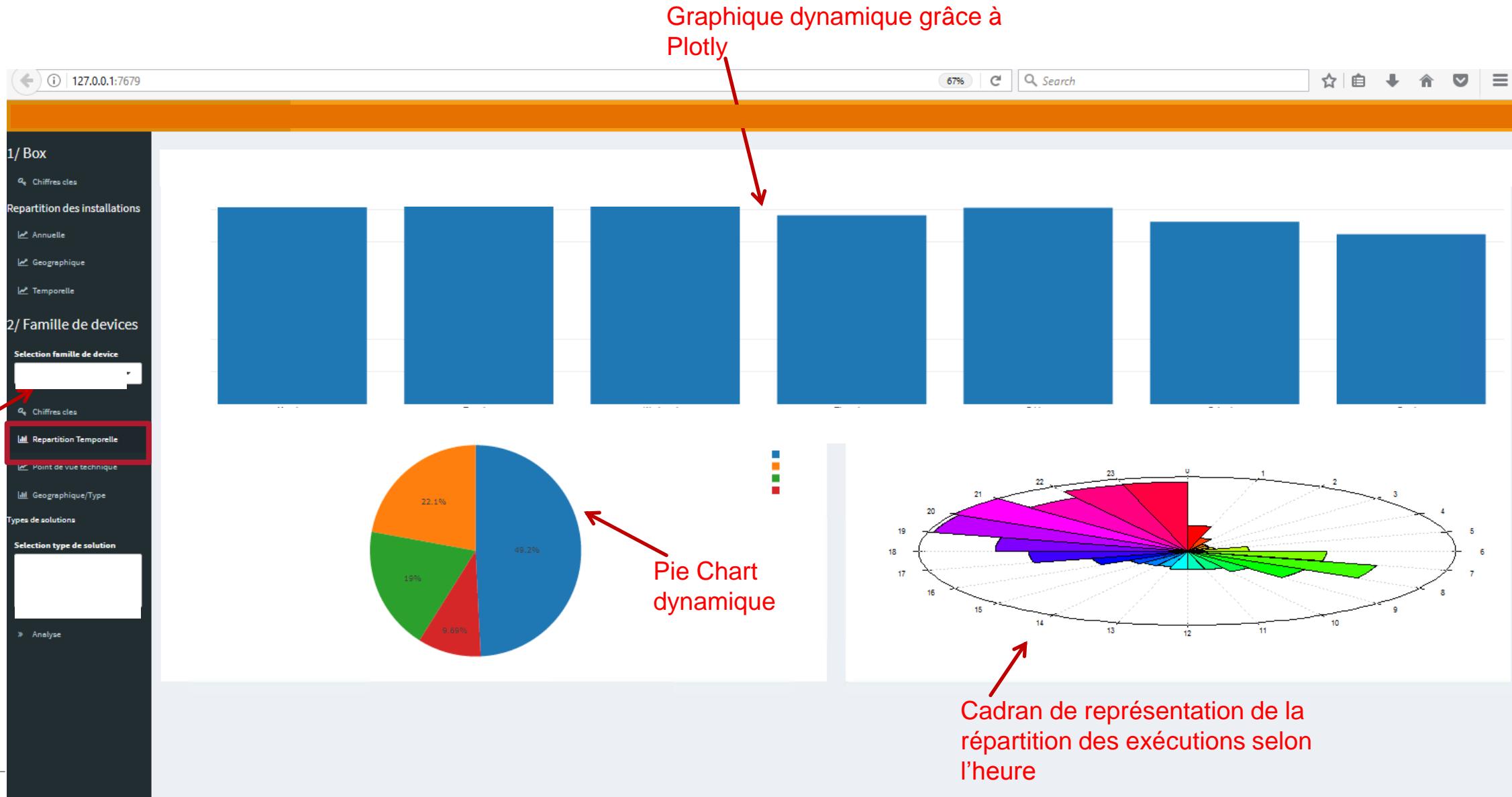
- Visualiser des chiffres-clés
- Permettre la sélection selon le type de famille
- Faire un dashboard simple, complet et user-friendly
- Possibilité de récupérer les graphiques

Solutions apportées

- Restitution sous-forme d'interface simple & accessible -> **Shiny**
- Traitement, nettoyage, calcul des indicateurs -> **RStudio**
- Génération de graphiques dynamiques -> **plotly**

IV. R SHINY

B LIVRABLE



V. Next Step

- Déploiement de l'application R Shiny sur un serveur afin que chacun puisse l'utiliser et manipuler les données qui les intéressent
- Volonté de récupérer des logs machines plus seulement via les applications mobiles (ipad/iphone/...) mais également via toutes les logs machines pilotées même en direct
- Faire de la maintenance prédictive
- Lever des alertes lors d'usages intempestifs



Identification des bassins d'habitat des Pyrénées-Atlantiques

Utilisation de R en agence d'urbanisme

Anglet, Parc Montaury
30 juin 2017



www.audap.org

Identification des bassins d'habitat des Pyrénées-Atlantiques

Utilisation de R en agence d'urbanisme

Anglet, Parc Montaury
30 juin 2017



www.audap.org



L'idée du projet



Projet expérimental pour l'Agence et l'INSEE qui consiste à définir des bassins d'habitat dans le département des Pyrénées-Atlantiques :

- Définir ce qu'est un bassin d'habitat et ce que ça traduit*
- Créer la base de données grâce aux nombreuses sources d'informations
- Utiliser R
- Définir les bassins d'habitat sur notre territoire.

L'utilisation de R dans les agences d'urbanisme n'est pas usuelle. Dans cet exercice nous n'avons pas effectué de programmation mais nous avons combiné différentes méthodes statistiques et travaillé sur la représentation de l'information sous forme cartographique.

* DRIANT,J-C, (2016). Bassin d'habitat, Politique du logement, *Analyses et débats*, 1 p.



Création de la base de données



26 Variables quantitatives

-
- 4** Thèmes:
- Démographie
 - Economie
 - Fiscalité
 - Foncier
-
- 7** Sources :
- INSEE
 - Etat Civil
 - DGFiP (fiscalité)
 - SIRS (foncier)
 - SOES (Sit@del2: logement)
 - SAFER (foncier)
 - L'observatoire des territoires
-

547 Individus : communes des Pyrénées-Atlantiques



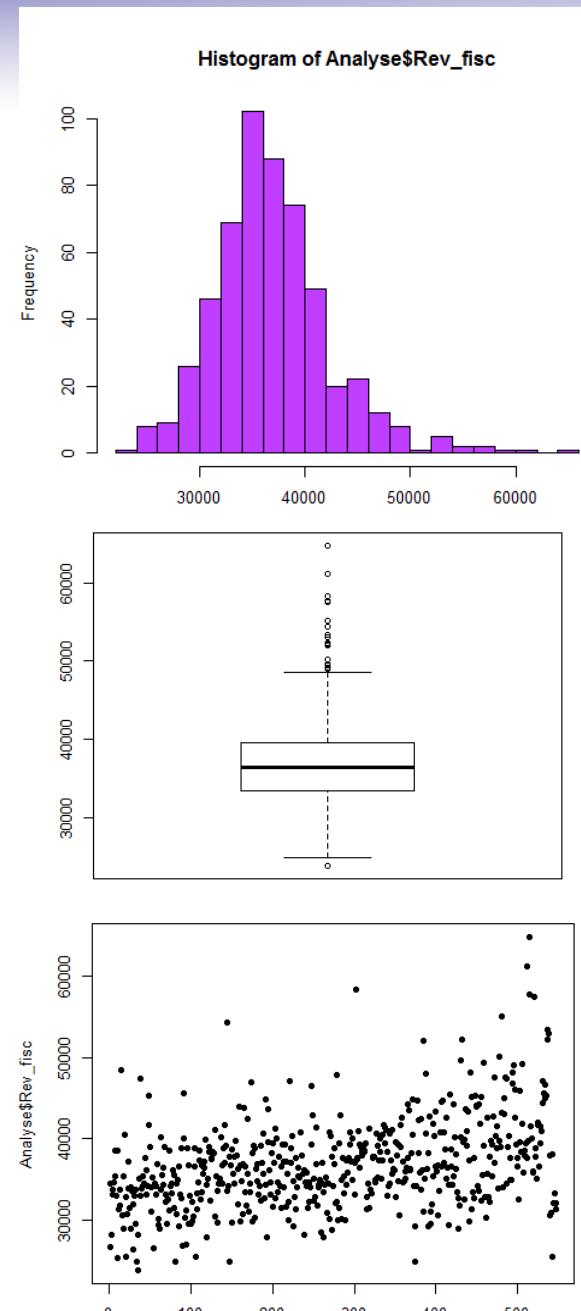
Traitements statistiques



1er travail : Analyse Univariée

Exemple de traitement pour la population : observe les outputs, la distribution, etc.

INSEE	Commune	Pop_2013	TCAM_08_13
Min. : 64001	Length: 547	Min. : 31.0	Min. :-0.098280
1st Qu. : 64141	Class : character	1st Qu. : 170.5	1st Qu. :-0.002755
Median : 64282	Mode : character	Median : 316.0	Median : 0.007668
Mean : 64281		Mean : 1214.0	Mean : 0.008579
3rd Qu. : 64422		3rd Qu. : 755.0	3rd Qu. : 0.017555
Max. : 64560		Max. : 77575.0	Max. : 0.095932
part_ER_13	P_ActOcc_13	Ind_EMPL	Rev_fisc
Min. : 0.00000	Min. : 0.2469	Min. : 0.04615	Min. : 23831
1st Qu. : 0.06196	1st Qu. : 0.4057	1st Qu. : 0.28314	1st Qu. : 33388
Median : 0.08679	Median : 0.4353	Median : 0.41539	Median : 36333
Mean : 0.08893	Mean : 0.4332	Mean : 0.58997	Mean : 36866
3rd Qu. : 0.11321	3rd Qu. : 0.4630	3rd Qu. : 0.66247	3rd Qu. : 39490
Max. : 0.25424	Max. : 0.6077	Max. : 7.50432	Max. : 64784
Int_Const	Typo_log	PxMoy_M	PavecEnf13
Min. : 0.00000	Min. : 0.00000	Min. : 60000	Min. : 0.00000
1st Qu. : 0.04309	1st Qu. : 0.02360	1st Qu. : 143083	1st Qu. : 0.2724
Median : 0.07171	Median : 0.05193	Median : 170304	Median : 0.3333
Mean : 0.08406	Mean : 0.19396	Mean : 179675	Mean : 0.3313
3rd Qu. : 0.10986	3rd Qu. : 0.16667	3rd Qu. : 202838	3rd Qu. : 0.4000
Max. : 0.90164	Max. : 5.86628	Max. : 501726	Max. : 0.6250
TMoyMen_13	Ind_jeun	Ind_vieilli	P_Ag_13
Min. : 1.667	Min. : 0.0000	Min. : 0.000	Min. : 0.0000000
1st Qu. : 2.286	1st Qu. : 0.5178	1st Qu. : 1.287	1st Qu. : 0.009399
Median : 2.438	Median : 0.7551	Median : 2.200	Median : 0.037440
Mean : 2.426	Mean : 0.8862	Mean : 4.962	Mean : 0.055911
3rd Qu. : 2.586	3rd Qu. : 1.0000	3rd Qu. : 4.918	3rd Qu. : 0.083333
Max. : 3.111	Max. : 13.0000	Max. : 59.256	Max. : 0.454545
P_Cadre_13	P_Retr_13	p_RS_13	p_LV_13
Min. : 0.00000	Min. : 0.0000	Min. : 0.00000	Min. : 0.00000
1st Qu. : 0.02243	1st Qu. : 0.2532	1st Qu. : 0.02858	1st Qu. : 0.04852
Median : 0.04938	Median : 0.3142	Median : 0.06800	Median : 0.06836
Mean : 0.05370	Mean : 0.3155	Mean : 0.10440	Mean : 0.07504
3rd Qu. : 0.07860	3rd Qu. : 0.3732	3rd Qu. : 0.13405	3rd Qu. : 0.09610
Max. : 0.23810	Max. : 0.6923	Max. : 0.92062	Max. : 0.25435
ARTI	p_Proprio	p_LOCHLM	vol_Prox
Min. : 0.003596	Min. : 0.3578	Min. : 0.00000	Min. : 0.00
1st Qu. : 0.040305	1st Qu. : 0.7540	1st Qu. : 0.00000	1st Qu. : 3.00
Median : 0.060213	Median : 0.8209	Median : 0.00000	Median : 7.00
Mean : 0.096166	Mean : 0.8004	Mean : 0.01611	Mean : 34.36
3rd Qu. : 0.101727	3rd Qu. : 0.8694	3rd Qu. : 0.01868	3rd Qu. : 19.00
Max. : 0.909929	Max. : 0.9706	Max. : 0.53085	Max. : 1964.00
Immo_0914	Foncier_0914	Px_terr_0914	SAFER_px
Min. : 0.00	Min. : 0.00	Min. : 0.05418	Min. : 0.04653
1st Qu. : 4.00	1st Qu. : 11.00	1st Qu. : 1.40120	1st Qu. : 0.64076
Median : 9.00	Median : 18.00	Median : 3.47089	Median : 1.08031
Mean : 70.76	Mean : 30.84	Mean : 11.35473	Mean : 2.24948
3rd Qu. : 23.00	3rd Qu. : 35.00	3rd Qu. : 10.97125	3rd Qu. : 1.90035
Max. : 7612.00	Max. : 359.00	Max. : 305.04098	Max. : 47.00000





Méthodes d'analyse statistique

Package -> FactoMiner*
Factoextra**

Analyse en Composantes Principales

Dé-correlées
centrer réduire les données

+

1^{ère} observation du territoire

+

Choix de 10 axes :
74 % de restitution de l'information



Classification Ascendante Hiérarchique

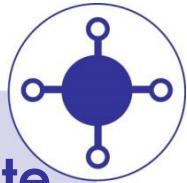
2 choix de classes :
5 ou 9 classes

+

Lecture du territoire

+

Choix de 9 classes
pour les bassins d'habitat

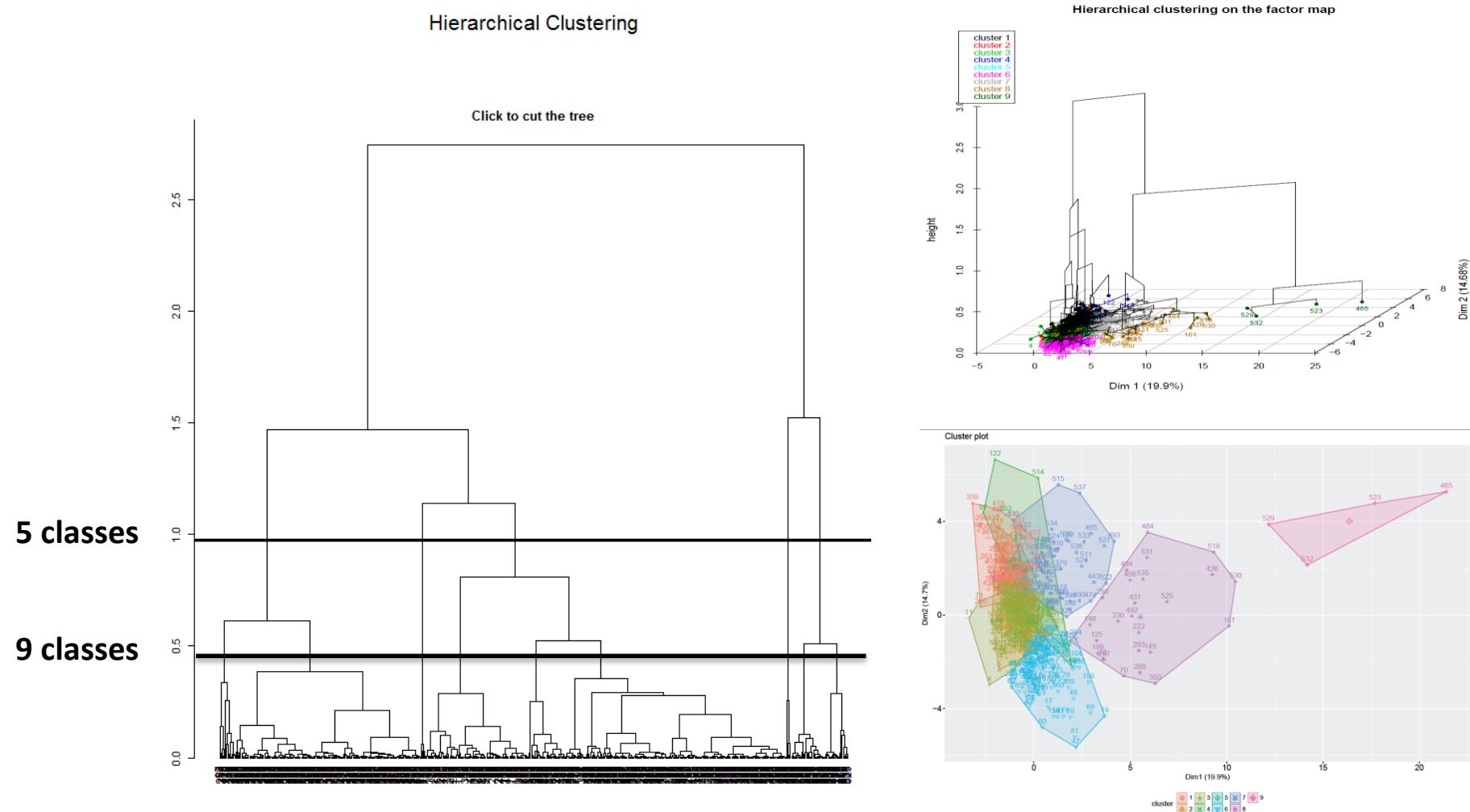


*François Husson, Julie Josse, Sébastien Le, Jeremy Mazet [aut] (2017). Multivariate Exploratory Data Analysis and Data Mining

**Alboukadel Kassambara [aut, cre], Fabian Mundt [aut] (2017). Extract and Visualize the Results of Multivariate Data Analyses

a

La classification Ascendante Hiérarchique





La 9 classes de communes

Clust

1	2	3	4	5	6	7	8	9
87	39	206	14	4	114	53	26	4

Territoires dynamiques

Une population jeune grâce à la présence de familles avec enfants et d'une sous représentation des retraités. Les actifs sont présents et les cadres sont ici en surreprésentation

Territoires ruraux vieillissants

Une population active agricole mais avec un vieillissement attendu important

Territoires « campagnards »

Une population ancrée dans le territoire, familles d'agriculteurs propriétaires avec des enfants

Territoires en transition #1

Marché foncier cher et un nombre de ventes de terrains élevé

Territoires en transition #2

Marché immobilier dynamique, sur du foncier cher

Territoires de résidences secondaires avec « mamie qui vit à côté »

Une population de retraités, de personnes isolées à faible revenus et des résidences secondaires

Territoires « IN » comme Brooklyn

Une population avec des revenus importants. Construction, artificialisation, l'immobilier devient cher

Territoires dynamiques et attractifs

Une population locataire, sans enfants. Communes avec des emplois, des appartements et artificialisées

La VILLE

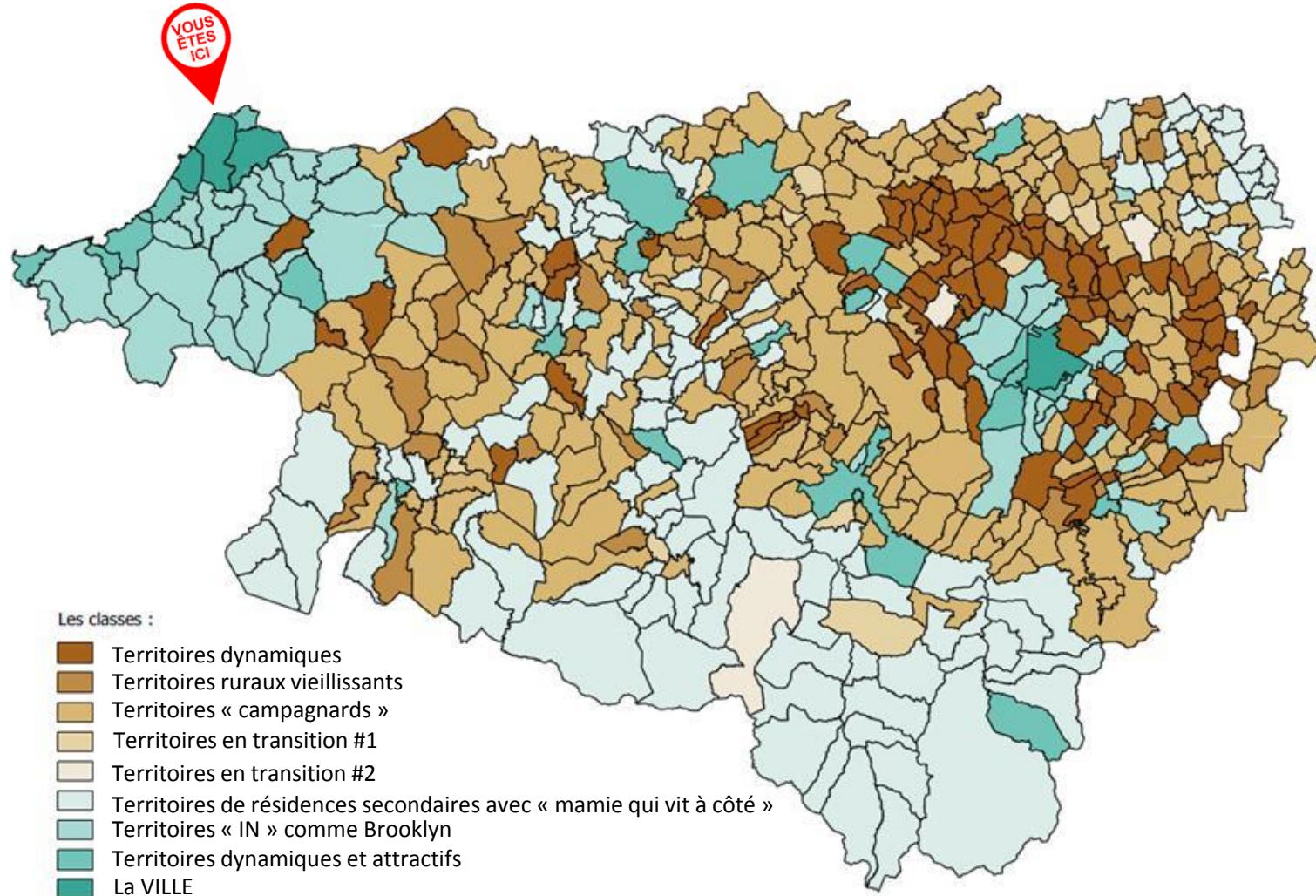
Une population très importante, avec beaucoup d'équipements de proximité.



a| Représentation des classes sous cartographie

Cartographie des 9 classes pour le département des Pyrénées-Atlantiques

Package -> cartography*



sources : INSEE, SAFER, DVF, SIRS

• * Timothée Giraud [cre, aut], Nicolas Lambert [aut] (2016). Thematic Cartography.



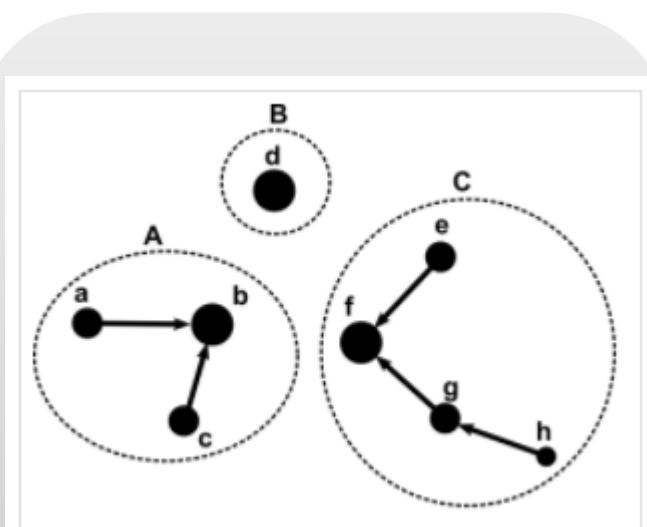
Flux dominants-dominés : théorie développée en 1961 par deux géographes, John D. Nystuen et Michael F. Dacey

- Analyse de réseau pour hiérarchiser les flux
- Ordonner et grouper les villes en fonction de l'intensité et de la direction des flux

Utilisation de la méthode des flux dominants/dominés avec les données « Domicile/Travail » de l'INSEE.

Illustrer les dynamiques entre communes du département

Observer des relations qui existent selon les types de classifications.



a est dominé par b si : a envoie son flux le plus important vers b et si la somme des arrivées de b est plus importante que la somme des arrivées de a.

Source : "A graph theory interpretation of nodal regions", de John D.Nystuen et Michael F.Dacey

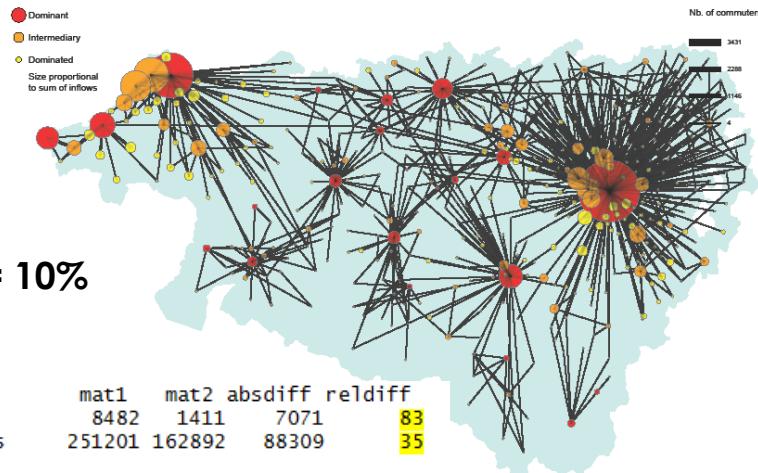
*Timothée Giraud [cre, aut], Laurent Beauguitte [aut], Marianne Guérois [ctb] (2016). Selections on flow matrices, statistics on selected flows, map and graph visualisations, Package R.

** Roger Bivand [cre, aut], Tim Keitt [aut], Barry Rowlingson [aut, ctb], Edzer Pebesma [ctb], Michael Sumner [ctb], Robert Hijmans [ctb], Even Rouault [ctb] (2017). Bindings for the Geospatial Data Abstraction Library.

a

Projection des Flux dominants-dominés

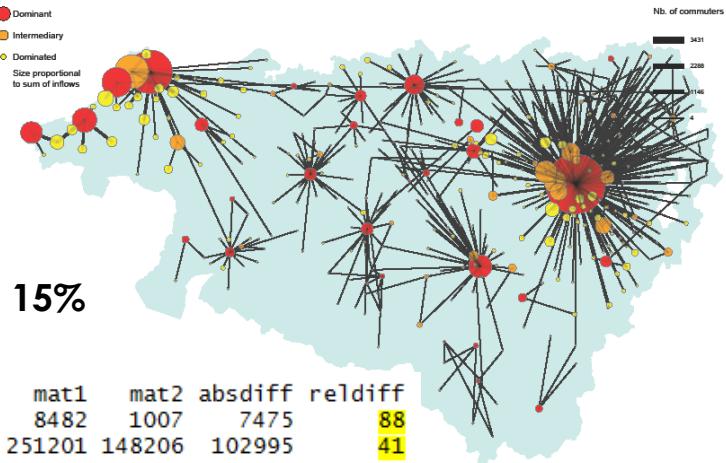
Les flux Domicile-Travail entre les communes pour K=10 %



17 % des flux restants

65 % des informations restantes

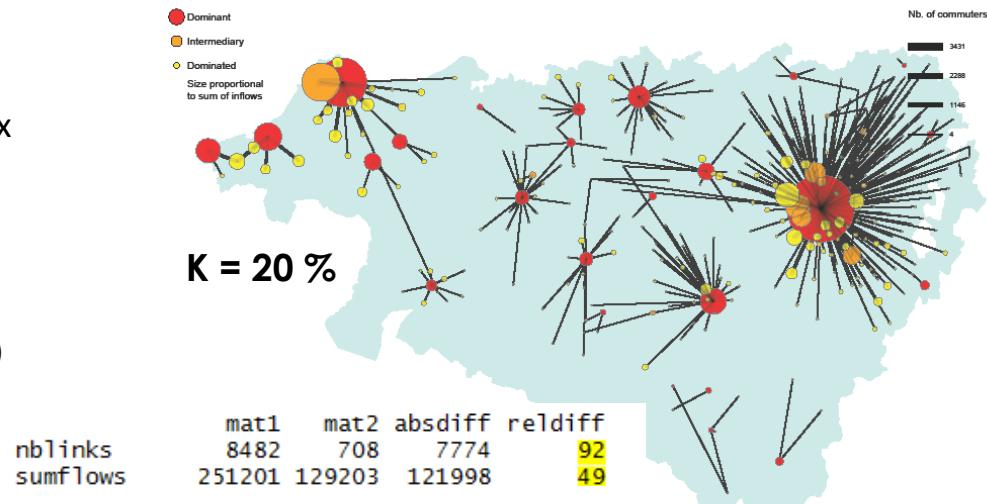
Les flux Domicile-Travail entre les communes pour k=15 %



12 % des flux restants

59 % des informations restantes

Les flux Domicile-Travail entre les communes pour K=20%



8 % des flux restants

51 % des informations restantes

Utilisation de la méthode :

xfirts (selects flows greater than k) pour les flux domicile/travail.

2 valeurs pour K :

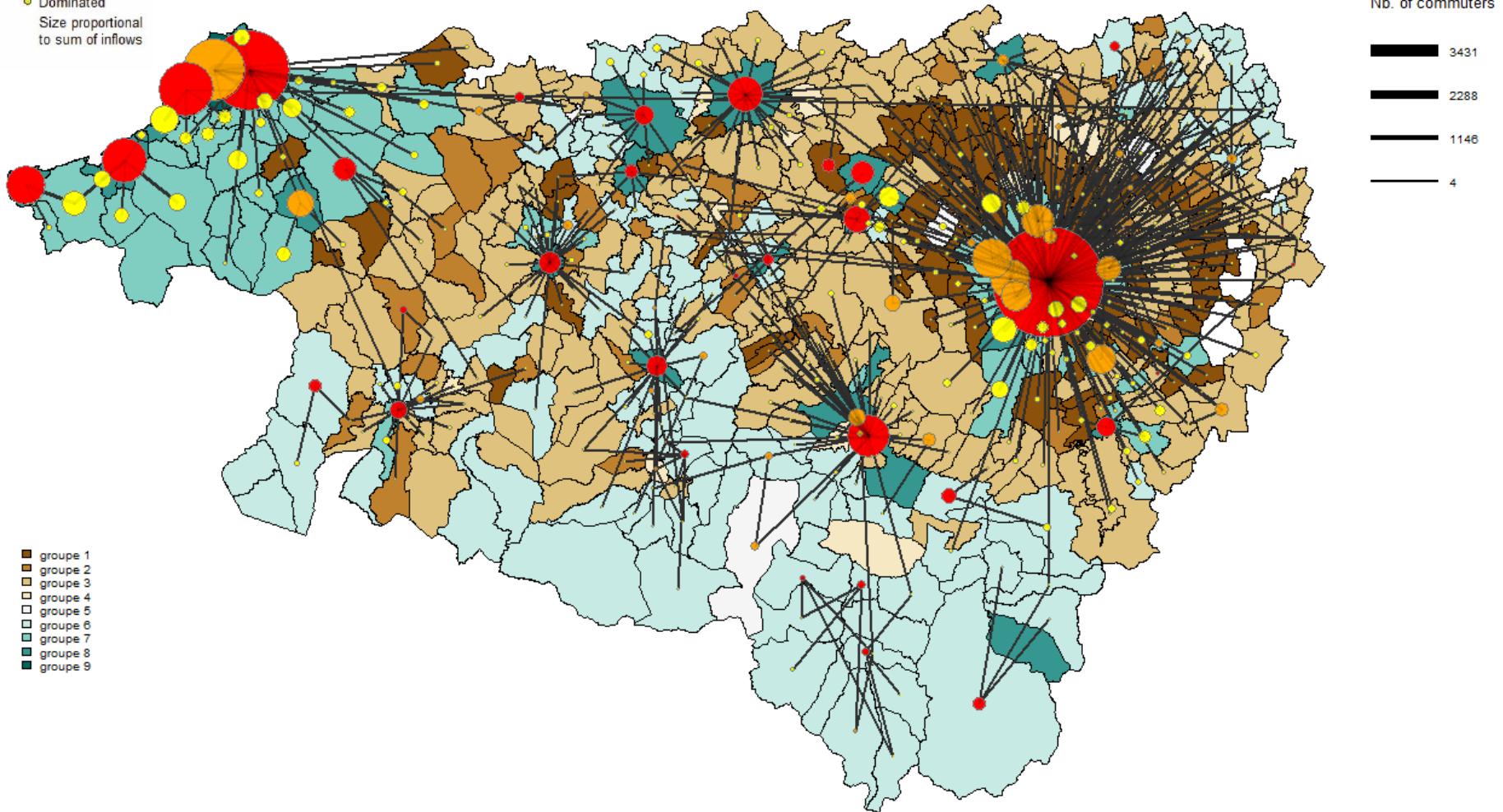
- Les flux sélectionnés pour l'analyse, ex si k = 15 %
- Le rapport dominant / dominé : k = 1 (ne change pas)

a

Flux dominants-dominés sur la CAH

- Dominant
 - Intermediary
 - Dominated
- Size proportional to sum of inflows

Les flux Domicile-Travail entre les communes pour k=15 %

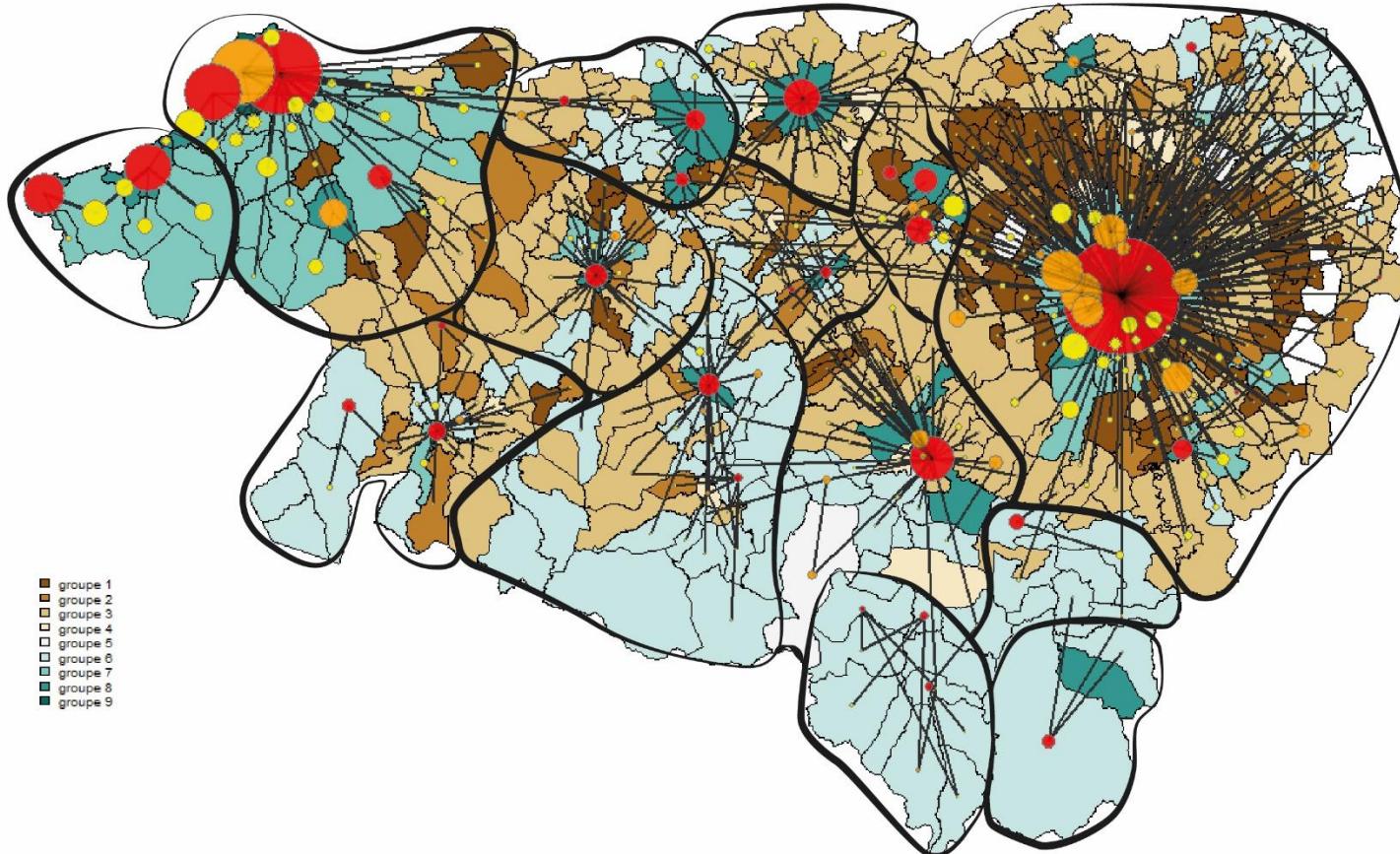
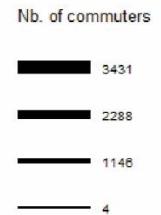


a|

Les 14 bassins d'habitat

Les flux Domicile-Travail entre les communes pour k=15 %

- Dominant
- Intermediary
- Dominated
- Size proportional to sum of inflows



Emma DAMITIO

Pôle Évolutions Spatiales
Connaissance & Évaluation

Agence d'Urbanisme Atlantique et Pyrénées
Ingénierie d'intérêt public pour la promotion d'un urbanisme durable dans les Pyrénées Atlantiques et le sud des Landes



Petite caserne
2 allée des platanes - BP 628
64 406 Bayonne Cedex
Tél. 05 59 46 50 10

4 rue Henri IV
Porte J
64 000 Pau
Tél. 05 33 64 00 30

➤ www.audap.org

05 33 64 00 40



Les membres de droit de l'Agence d'urbanisme Atlantique & Pyrénées

Comment je suis devenue crolute

Maëlle Salmon (@ma_salmon, @RLadiesBCN)
Scott A. Chamberlain (@sckottie, @rOpenSci)

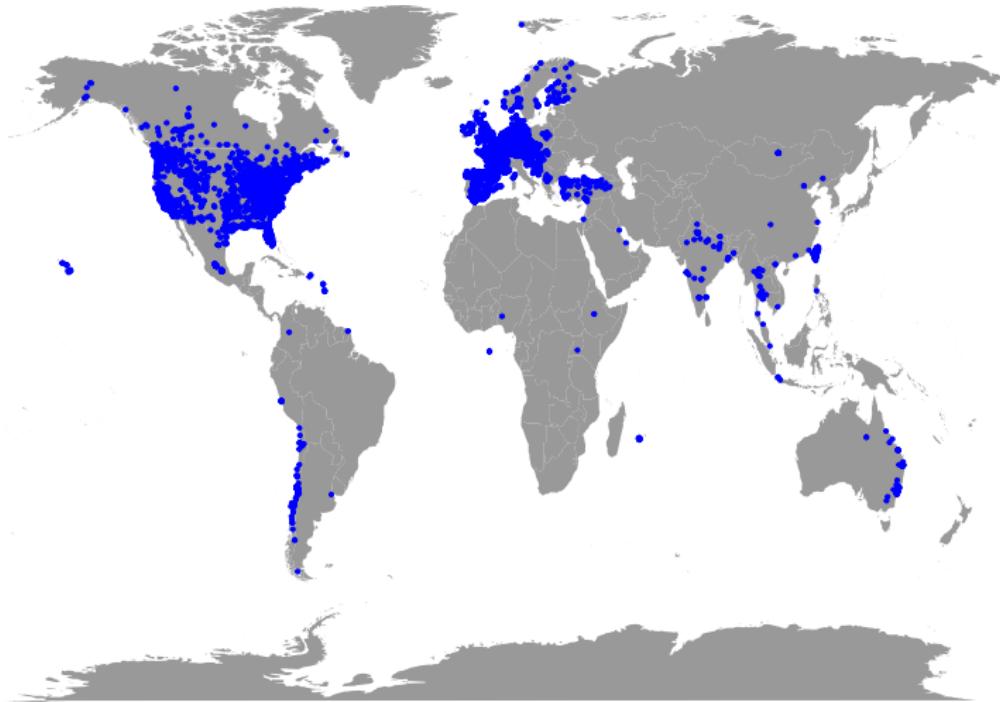
Vendredi 30 juin 2017

Comment je suis devenue crolute

Maëlle Salmon (@ma_salmon, @RLadiesBCN)
Scott A. Chamberlain (@sckottie, @rOpenSci)

Vendredi 30 juin 2017

OpenAQ, accès à des données libres de qualité de l'air



Des données de l'air dans R via `ropenaq`

Paquet créé en décembre 2015.

Accès aux 5 services de l'API d'OpenAQ:

- `measures` (`aq_measures`),
- `latest` (`aq_latest`),
- `countries` (`aq_countries`),
- `cities` (`aq_cities`),
- `locations` (`aq_locations`).

La nouvelle galaxie de mon paquet, rOpenSci

Organisation pour la science ouverte, la reproductibilité.

Développement de paquets et d'une communauté.

Système de revue par les pairs de paquets. `ropenaq` a été revu par et amélioré grâce à Andy Teucher et Andrew MacDonald.

<https://github.com/ropensci/ropenaq>



Mais au fait comment parler avec une API ?

Envoi et réception d'information. Outil de base = libcurl interfacé par les paquets RCurl et curl. http et crul sont des paquets dépendants de curl.

On fait des *requêtes* à l'API avec des *paramètres*.

L'API peut avoir un système de *pagination*: pas toute la commande d'un coup !

Utiliser crul pour ropenaq

ropenaq a d'abord dépendu de httr.

Découverte de crul, paquet de rOpenSci développé par Scott Chamberlain... pourquoi pas changer ?

Crul = une planète de Star Wars. Crolutes et Gilliands.

Argument 1 : la vitesse... (1/3)

Imaginons une demande de données pour laquelle l'API donnerait 10 pages = 10 requêtes à faire.

2 solutions:

- requêtes les unes après les autres: *requêtes synchrones*
- requêtes toutes à la fois ou au moins en paquets: *requêtes asynchrones*

crul permet les requêtes synchrones ou asynchrones !

Argument 1 : la vitesse... (2/3)

Petit expérience de comparaison avec microbenchmark.

Objectif: obtention de 20 538 mesures de 14 jours à Delhi, 1000 mesures par page.

100 répétitions.

Argument 1 : la vitesse... (3/3)

De 37 à 6 secondes.

```
aq_measurements(city = "Delhi",
                 date_from = "2017-01-01",
                 date_to = "2017-01-15",
                 limit = 1000)
```

Argument 2 : classes R6 (1/2)

```
# URL
url <- "https://httpbin.org/get"

# code httr
response <- httr::GET(url)
parsed_content <- httr::content(response, as = "text")
status <- httr::status_code(response)
```

Argument 2 : classes R6 (2/2)

```
# URL
url <- "https://httpbin.org/get"

# code crul
client <- crul::HttpClient$new(url = url)
res_get <- client$get()
parsed_content <- res_get$parse()
status <- res_get$status_code
```

Pourquoi utiliser curl?

- Certes pas encore toutes les options offertes par httr
- Requêtes asynchrones: vitesse!
- Classes R6
- Paquet bien maintenu

Y a-t-il d'autres trésors dans la collection de paquets rOpenSci?

Oui ! Allez sur <https://ropensci.org/> !

- tabulizer
- magick
- rdefra
- Votre paquet ? Soumettez-le !
<https://github.com/ropensci/onboarding/>