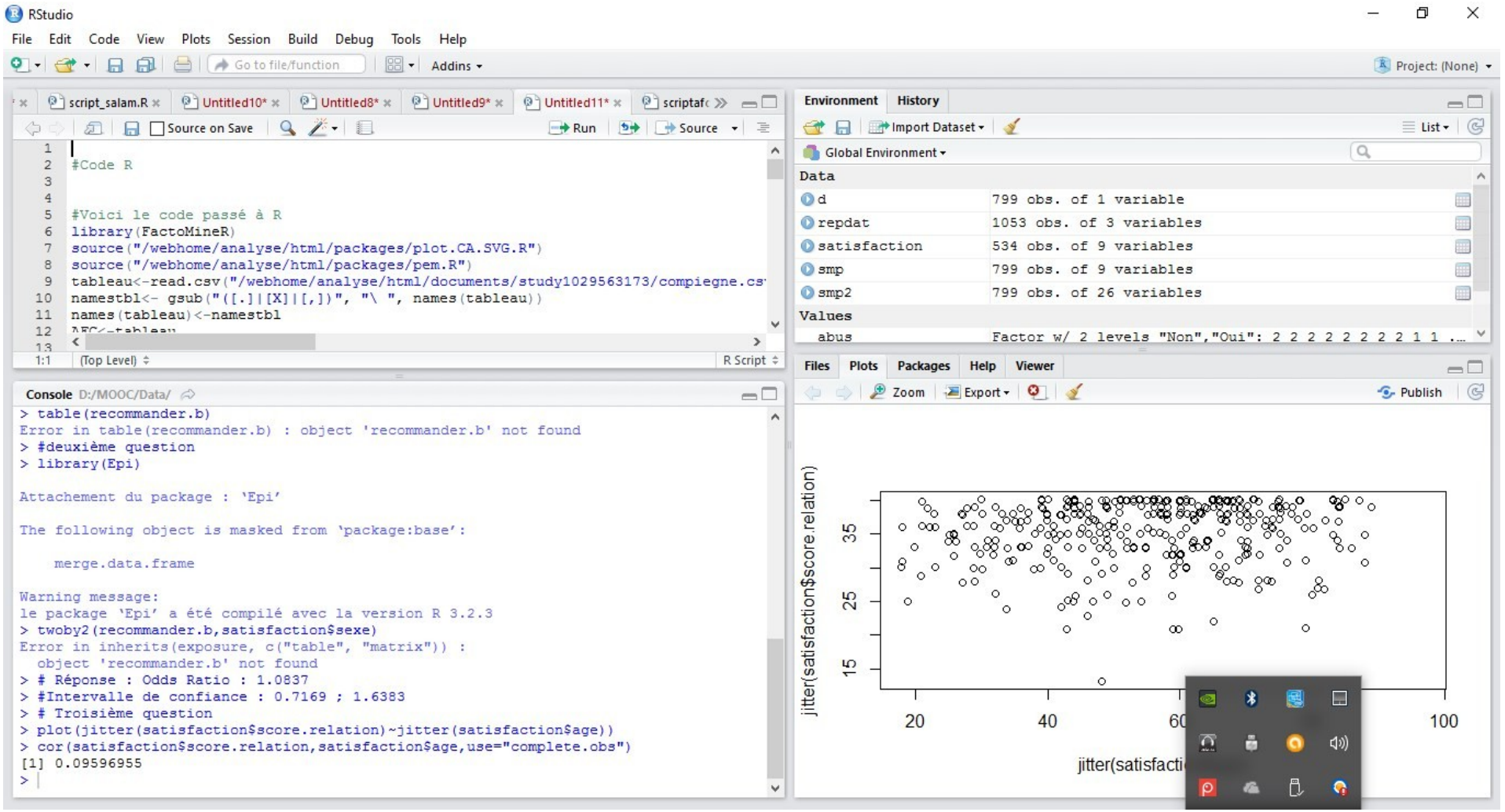
Quelques petites réflexions sur le traitement des données



Luc Sanson, 22 Octobre 2016

- Un peu d'intro : au sein de l'inrap, beaucoup d'opérations archéologiques : Diagnostics, fouilles...

- Elles génèrent, dans le cadre du déploiement des SIG des données spatiales de plusieurs natures

- 3 poupées russes principales : emprises, ouvertures, unités d'observation

- Réseau des référents SIG : on est convaincus de la nécessité de rendre publiques ces données (et toutes les données liées à l'archéo au sens large). C'est une mission de service public !

- Parmi les collègues certains sont du coté obscur (« ce sont mes données, elles sont à moi »), mais confiance dans l'avenir.

- Donc on a beaucoup de données liées à l'opérationnel qu'on souhaiterait rendre publique.

- Après, il y aura d'inévitables débats sur quoi rendre public (question des détectoristes, du pillage des sites, de la situation de concurrence suite à la loi de 2003...).

- L'inrap développe aussi des programmes de recherches (PAS, PCR). Comme le PCR de la Plaine de Troyes, ou le PAS sur les fosses mésolithiques.

- L'articulation entre *l'opérationnel* et *la recherche* est une problématique assez ardue en soi (et je ne parle pas des problématiques logistiques, ou des susceptibilités).
- C'est que, en soi, c'est fondamentalement différent de mener à bien une opération ou de mener à bien un programme de recherche (j'ai presque envie de dire que ce sont deux zones du cerveau différentes).

- Cela implique un changement dans l'enregistrement, dans le choix des géométries sur sig...

- Dans le monde de la recherche, et sur certains logiciels : logique de reproductibilité
- Par exemple : les scripts R (j'en utilise à foison!) ou le service GIThub

[BILLET] Cartographie reproductible

PAR NICOLAS LAMBERT · PUBLICATION 04/07/2016 · MIS À JOUR 30/08/2016

Camarades cartographes, partant du constat que les géographes/cartographes ont beaucoup à apprendre des informaticiens en terme de méthode de travail, je vous parle aujourd'hui de cartographie reproductible. Ce terme – *la cartographie reproductible* – que je pose pour la première fois dans ce billet, est directement inspiré de ce qu'on appelle la recherche reproductible (voir). Réfléchissant depuis longtemps à un *process* de travail permettant d'organiser, d'archiver et de partager mon travail de conception cartographique, je vous propose ici un dispositif qui me semble efficace. Il repose sur deux piliers : **R** et **github**.

R c'est quoi ?

R est un langage de programmation open source dédié à l'analyse statistique. Outre les fonctionnalités statistiques présentes de base dans R, certains packages additionnels permettent d'effectuer en sus des opérations SIG (rgeos), d'analyse spatiale (SpatialPosition) ou de cartographie thématique (cartography). Avec R, une carte se conçoit donc comme un processus traçable permettant de couvrir toute la chaîne collecte-traitement-analyse-représentation, et cela dans un seul et unique environnement logiciel (A l'exception de l'étape de mise en scène cartographique). Dans R, une carte se conçoit donc en écrivant des lignes de code et non via des clics sur une interface graphique. Si ce procédé peut paraître rebutant, il peut aussi faire gagner beaucoup de temps une fois qu'un programme est bien écrit, puisqu'il est alors facilement modifiable et réutilisable (avec d'autres données, d'autres hypothèses, d'autres paramètres, etc.)

Github c'est quoi ?

Github est un service web d'hébergement et de gestion de développement de logiciels lancé en 2008, basé sur le système de gestion de versions git. S'il permet l'hébergement de projets payants, l'avantage de github est surtout de proposer des comptes gratuits pour les projets de logiciels libres. Github dispose de nombreuses fonctionnalités utiles : wiki, issue tracking system (suivi des problèmes), gestion de versions, etc. Un tel système vous permettra donc de stocker et partager vos productions cartographiques mais également les différents fichiers et programmes qui vous ont servi à les construire.

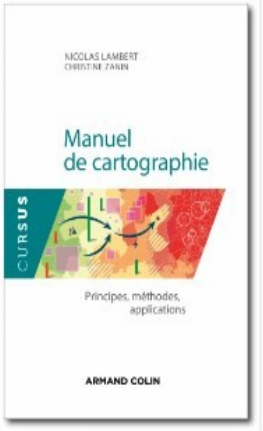
Éléments clefs pour une cartographie reproductible

–

HISTOIRE SUIVANTE
Manuel de cartographie : recension pour la Cliothèque

HISTOIRE PRÉCÉDENTE
[BILLET] Des cartes dans vos têtes

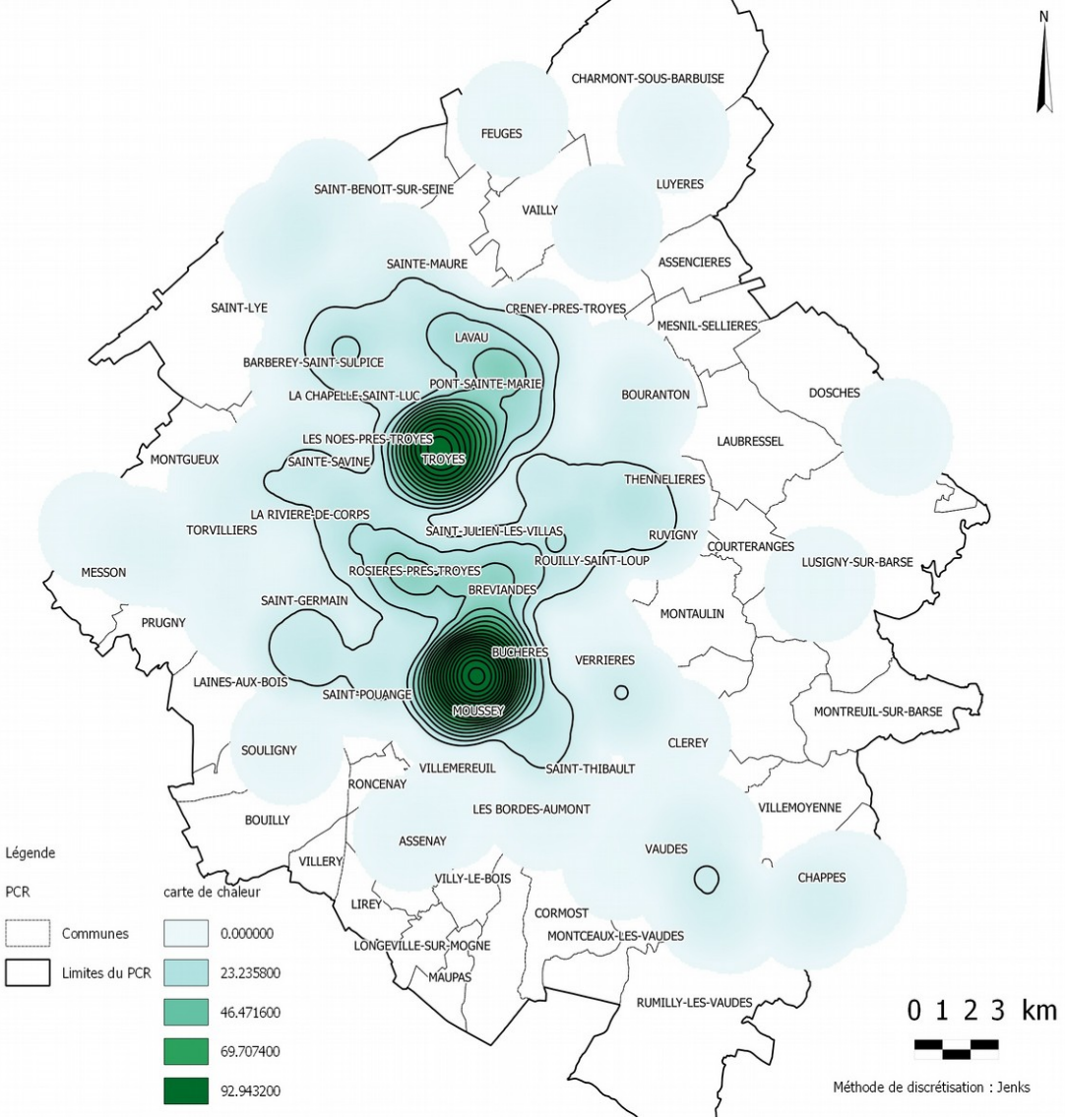
VIENT DE PARAÎTRE



Serial Mapper

- Pour la plaine de Troyes, on est 2 principaux « tripatouilleurs » de données (V. Riquier et moi).
- La formation statistiques niveau 1 proposée par Sylvain Badey permettra aux collègues de tripatouiller également.
- On fait de calculs des valeurs d'indépendance, des écarts, des afc pour ensuite aller voir les spécialistes et tenter de lier ce qu'on perçoit à un discours synthètique.
- J'ai déjà un problème sur ce principe...

- La question, c'est où s'arrêter ? Quelle est la valeur de ce qu'on fait ?
- Dans une logique de reproductibilité, on va fournir nos scripts R et sur une plateforme (github ou autre), nos tableaux de données (et tableaux de contingence).
- Nos données spatiales : idem (serveur PostGIS)
- Du coup d'autres chercheurs pourront vérifier ce qu'on a fait, mais pourront également mettre en évidence des phénomènes d'après ces mêmes tableaux...



Méthodo

7 - structuration données - geol.xls - OpenOffice Calc

Fichier Édition Affichage Insertion Format Outils Données Fenêtre Aide

B17 = 4330

	A	B	C	D	E	F
1	occupation	surface	annee	COMMUNE	LIEU_DIT	LAMBERT93
2	VM079	57	2007	BREVIANDR	ZAC de Cou	781117
3	VM079	7203	2007	BREVIANDR	ZAC de Cou	781117
4	LS004	6057	2015	SAINT-LYE	rue du mouli	774139
5	VM004	1163	2010	RUVIGNY	Demeures d	787436
6	VM004	5797	2010	RUVIGNY	Demeures d	787436
7	VM005	6916	2010	RUVIGNY	Demeures d	787434
8	VM005	7226	2010	RUVIGNY	Demeures d	787434
9	VM006	6424	2010	RUVIGNY	Demeures d	787435
10	VM006	9924	2010	RUVIGNY	Demeures d	787435
11	VM062	3632	2004	BOURANTO	La LouviSre	787070
12	VM064	943	2004	BOURANTO	La LouviSre	787035
13	VM063	3835	2004	BOURANTO	La LouviSre	787034
14	VM053	77	2004	SAINT-AND	Echenilly	777228
15	VM052	77	2004	SAINT-AND	Echenilly	777232
16	VM067	1364	2004	BARBEREY	Les Vingt-Ar	776950
17	VM076	4330	2001	SAINT-POU	Rue du Pont	776922
18	VM085	1295	2007	SAINT-GER	DerriSre le P	775400
19	VM087	15390	2009	SAINT-GER	Le Pont-de-l	776888
20	VM087	2634	2009	SAINT-GER	Le Pont-de-l	776888
21	VM093	5767	1996	BARBEREY	A rodrome	775053
22	VM096	11599	1976	BARBEREY	Les Vingt-Ar	776309
23	VM100	322	1987	BARBEREY	Rue du prof.	776437

22	6803405	686	0	Fouille	1976	Tomasson_P	Plaine craye	cra					
23	6804437	1565	0	D,couverte fo	1987	Lenoble_Mic	Plaine craye	lim					

Arrière-plan de cellule :

dure de cellule :

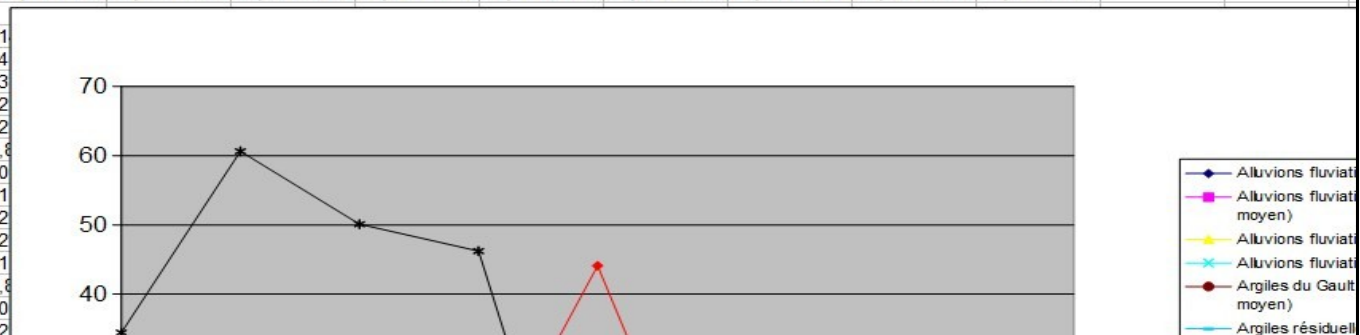
	A
3	Somme de surface
4	DESCR
5	Alluvions fluviatiles actuelles et r€centes (Quaternaire-Holoc€ne)
6	Alluvions fluviatiles anciennes (Quaternaire-Pleistoc€ne inf€rieur ◆ moyen)
7	Alluvions fluviatiles anciennes (Quaternaire-Pleistoc€ne moyen)
8	Alluvions fluviatiles anciennes (Quaternaire-Pleistoc€ne sup€rieur)
9	Colluvions indiff€renci€es (Quaternaire)
10	Craie grise ◆ blanche argileuse (Cr€tac€ sup€rieur-Turonien moyen)
11	Craie noduleuse, craie marneuse (Cr€tac€ sup€rieur-Turonien inf€rieur)
12	Craies ◆ Acanthoceras rhotomagense (Cr€tac€ sup€rieur-C€nomanien moyen)
13	Craies ◆ Actinocamax plenus (Cr€tac€ sup€rieur-C€nomanien sup€rieur)
14	Marnes, argiles, craies, indiff€renci€es (Cr€tac€ sup€rieur-C€nomanien inf€rieur ◆ moyen)
15	Marnes, marnes glauconieuses, marnes crayeuses, argiles sableuses (Cr€tac€ sup€rieur-C€nomanien inf€rieur)
16	Total

	B	C
3		
4	Total	
5	15692	
6	2962	
7	65003	
8	37074	
9	124196	
10	12877	
11	10542	
12	8122	
13	6275	
14	1674	
15	6129	
16	290546	

E1 = différence phase 8

	A	B	C	D	E
1	DESCR	pourc_surf	pourc_ope	pourc_ph8	différence
2	Alluvions fluv	12,28044431	12,99137848	2,057380451	-10,22306
3	Alluvions fluv	1,284768891	0,043311033	1,205054557	-0,079714
4	Alluvions fluv	8,874149753	21,77343019	18,92529638	10,051146
5	Alluvions fluv	5,687461923	10,34259524	16,040814	10,353352
6	Alluvions fluv	0,002241978	0	0	-0,002241
7	Argiles du G	4,428973672	0,239538709	0	-4,428973
8	Argiles et ca	0,860198945	0	0	-0,860198
9	Argiles grise	1,473055755	0	0	-1,473055
10	Argiles résid	0,345823579	0	0	-0,345823
11	Argiles, fer o	1,140223163	0,012711092	0	-1,140223
12	Argiles, sabl	1,870186851	0	0	-1,870186
13	Calcaires à *	0,036386445	0	0	-0,036386
14	Calcaires po	0,01334365	0	0	-0,013343
15	Colluvions in	11,40378984	21,988658	39,41488539	28,011095
16	Craie à Micr	2,458490003	0,091524917	0	-2,458
17	Craie à Micr	15,34586599	5,480410245	0	-15,3458
18	Craie grise à	5,042547692	12,59946836	3,963823953	-1,078723
19	Craie nodule	4,749398691	5,536456988	7,654659517	2,9052608
20	Craies à Aca	2,803968979	4,795521922	6,613560729	3,8095917
21	Craies à Act	2,761977995	1,596251102	0	-2,761977
22	Formations ◆	1,534395289	0,994820748	0	-1,534395
23	Grézes ou g	0,275475893	0	0	-0,275475
24	hydrographie	1,495348175	0,171222073	0	-1,495348
25	Limons des	2,050346407	0	0	-2,050346
26	Marnes de B	5,927571578	0,272944436	0,949967046	-4,977604
27	Marnes, argi	1,858685183	0,611831406	0,681047039	-1,177638
28	Marnes, mar	1,143675144	0,312283271	2,493510932	1,3498357
29	Sables des I	0,737755847	0	0	-0,737755
30	Sables verts	2,113448377	0,145641788	0	-2,113448
31	Total	100	100	100	

	A	B	C	D	E	F	G	H	I	J	K	L
1	DESCR	phase 1	phase 2	phase 3	phase 4	phase 5	phase 6	phase 7	phase 8	phase 9		
2	Alluvions fluv	-1,00743402	-7,17699953	-5,26575192	-7,02825741	-2,30977348	7,935726951	0,892737658	-10,2230639	-6,87957835		
3	Alluvions fluv	23,38679416	11,97886154	-1,28476889	-1,28476889	-1,28476889	-0,90336832	-0,99298159	-0,07971433	-0,26530898		
4	Alluvions fluv	-8,71996788	-8,67569302	10,34719096	4,107913207	20,50744151	6,419267128	6,890540831	10,05114663	13,49855543		
5	Alluvions fluv	-5,68746192	-5,68746192	-5,51956419	-5,35289959	-5,03847596	-2,58813533	1,025616196	10,35335208	7,072651794		
6	Alluvions fluv	-0,00224198	-0,00224198	-0,00224198	-0,00224198	-0,00224198	-0,00224198	-0,00224198	-0,00224198	-0,00224198		
7	Argiles du G	-4,42897367	-4,42897367	-4,42897367	-4,42897367	-4,42897367	-4,42897367	-4,42897367	-4,42897367	-4,42897367		
8	Argiles et ca	-0,86019894	-0,86019894	-0,86019894	-0,86019894	-0,86019894	-0,86019894	-0,86019894	-0,86019894	-0,86019894		
9	Argiles grise	-1,47305576	-1,47305576	-1,47305576	-1,47305576	-1,47305576	-1,47305576	-1,47305576	-1,47305576	-1,47305576		
10	Argiles résid	-0,34582358	-0,34582358	-0,34582358	-0,34582358	-0,34582358	-0,34582358	-0,34582358	-0,34582358	-0,34582358		
11	Argiles, fer o	-1,14022316	-1,14022316	-1,14022316	-1,14022316	-1,14022316	-1,14022316	-0,78489676	-1,14022316	-1,14022316		
12	Argiles, sabl	-1,87018685	-1,87018685	-1,87018685	-1,87018685	-1,87018685	-1,87018685	-1,87018685	-1,87018685	-1,87018685		
13	Calcaires à *	-0,03638645	-0,03638645	-0,03638645	-0,03638645	-0,03638645	-0,03638645	-0,03638645	-0,03638645	-0,03638645		
14	Calcaires po	-0,01334365	-0,01334365	-0,01334365	-0,01334365	-0,01334365	-0,01334365	-0,01334365	-0,01334365	-0,01334365		
15	Colluvions in	34,4429051	60,62839439	50,10364218	46,22005445	-8,33358703	22,78239044	22,86919974	28,01109555	31,34193717		
16	Craie à Micr											
17	Craie à Micr	-1										
18	Craie grise à	-4										
19	Craie nodule	-3										
20	Craies à Aca	-2										
21	Craies à Act	-2										
22	Formations ◆	0,8										
23	Grézes ou g	-0										
24	hydrographie	-1										
25	Limons des	-2										
26	Marnes de B	-2										
27	Marnes, argi	-1										
28	Marnes, mar	7,8										
29	Sables des I	-0										
30	Sables verts	-2										



Méthodo

- Ce qui est drôle, d'un point de vue méthodologique, c'est que les interrogations naissent de par le jeu avec les tableaux, les TCD...
- C'est de ce triturage qui me fait poser des questions.
- Une journée, je peux me dire « tiens je vais regarder si... ». Le choix du questionnement est complètement arbitraire.
- Donc on ne peut pas trop prévoir ni planifier les orientations que vont prendre les problématiques au préalable...

Script R

```
library(FactoMineR)
source("/webhome/analyse/html/packages/plot.CA.SVG.R")
source("/webhome/analyse/html/packages/pem.R")
tableau<-read.csv("/webhome/analyse/html/documents/study1029563173/compiegne.csv", header=TRUE, row.names=1, sep="\t",  na.strings="NA",  dec=".", strip.white=TRUE, blank.lines.skip=TRUE, encoding="UTF-8")
namestbl<- gsub("([.]|[X]|[,])", "\ ", names(tableau))
names(tableau)<-namestbl
AFC<-tableau
AFC<-apply(AFC, 2, function(x) ifelse(is.na(x), 0, x))
AFC<- t(AFC)
tbl<-AFC
tbl[,margin.table(t(tbl),1)!=0]
tbl[margin.table(t(tbl),2)!=0,]
AFC<-tbl
AFC.ca <- CA(AFC, ncp=5, graph=FALSE)
```

- Ok pour fournir les scripts, mais à quoi bon refaire exactement les mêmes opérations ?
- A partir des mêmes données, on retombe sur les mêmes résultats…

Quelques questionnements

- Ou bien les scripts peuvent servir d'inspiration pour d'autres travaux.
- *Effet bénéfique pour d'autres collègues.*

- Ou bien, la démarche de recherche n'est jamais finie. Et les scripts peuvent tourner en routine sur des tableaux de données sans cesse alimentés.
- *Mais encéphalogramme plat en revanche*

Quelques questionnements

- Comment publier ? Avec quelles données ? Les tableaux et les scripts en plus des résultats ? Problème avec la taille des tableaux !
- Une publi ne serait qu'un pêchage de résultats possibles parmi les nombreuses possibilités et nombreux calculs autres.

Perspectives de la recherche

- Différentes possibilités depuis l'arbitraire total du choix du chercheur.
- Ou bien on s'inspire de ce qu'on a vu ailleurs (on reprend un bout de script). Mais on utilise d'autres tableaux de données.
- Ou bien on ne fait plus que de la saisie de données, et les scripts ne bougent pas et tournent sans cesse sur des tableaux sans arrêt complétés.
- Ou bien les données sont explorées dans de toutes autres directions par d'autres personnes.