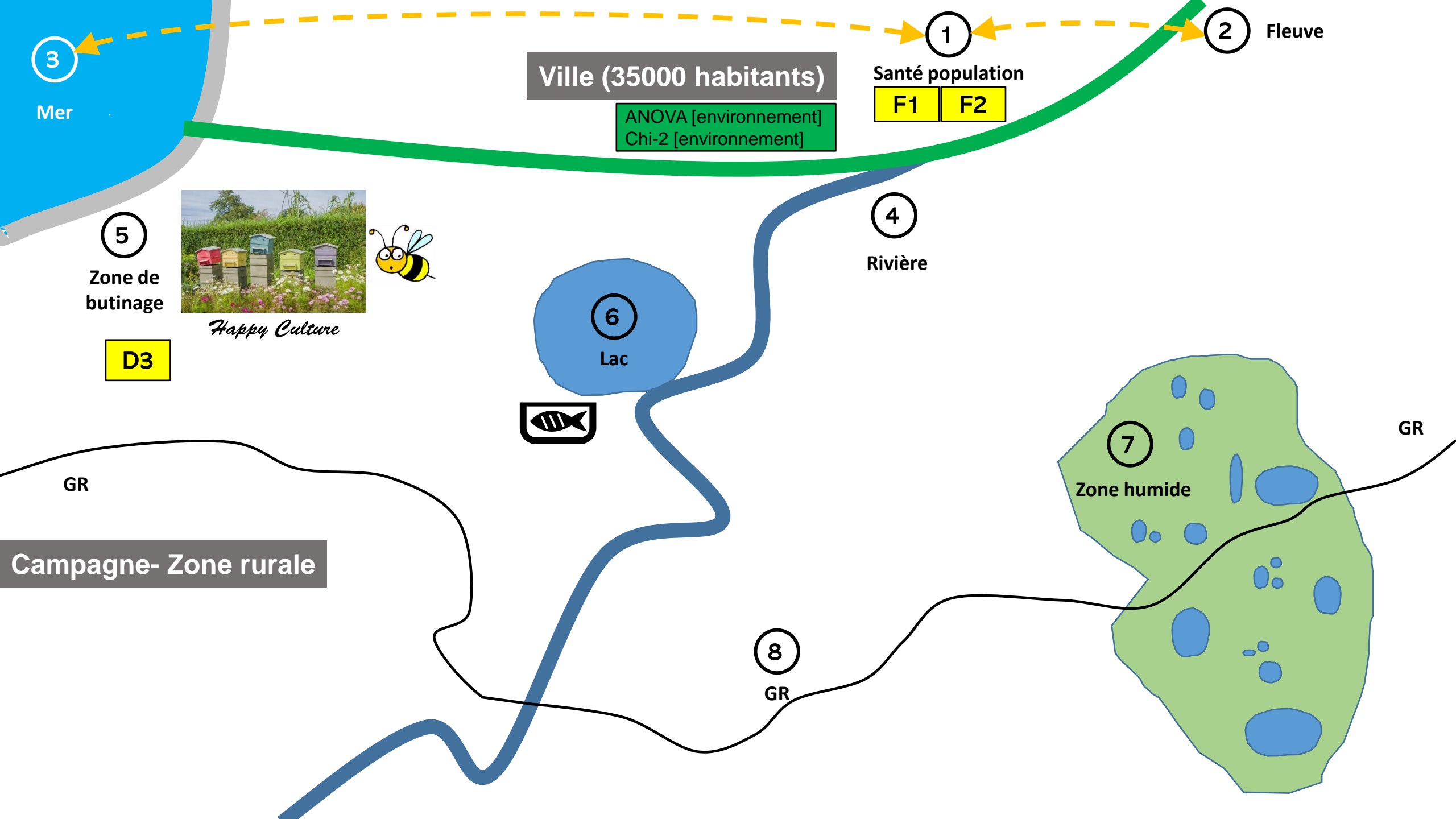


Data sciences : Graphique contextuel

(Nouvelles données)



A quoi pouvons-nous nous intéresser ? A vos idées !

- ▶ Revêtez votre casquette de *data scientist*
- ▶ Travaillez par groupe de 3

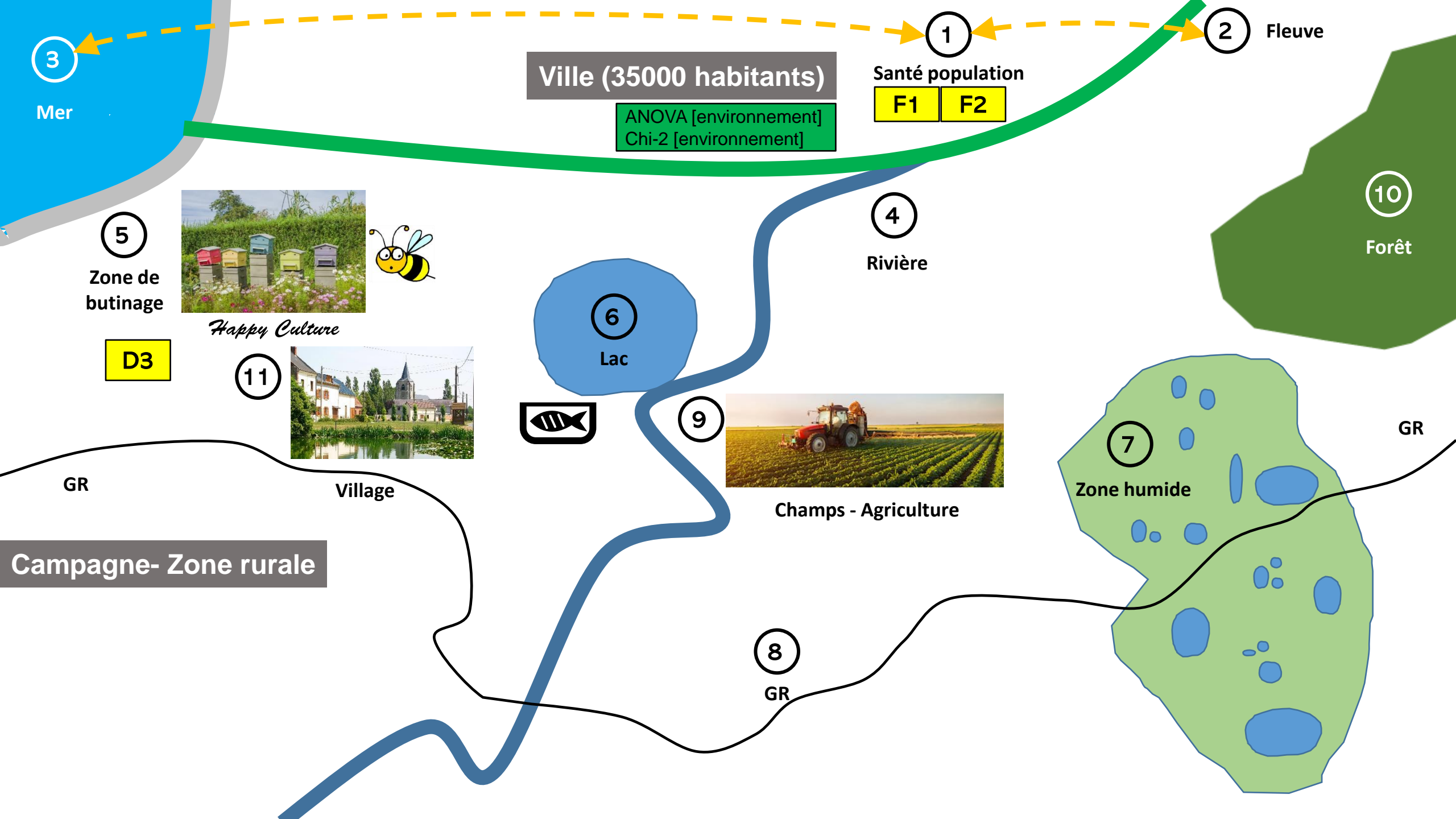


- Questionnements ?
- Variables ?
- Lois ?
- Stats descriptives ?
- Stats inférentielles ?
- Machine learning?

(No limit!)

Que pouvons nous en tirer ? A vos idées !

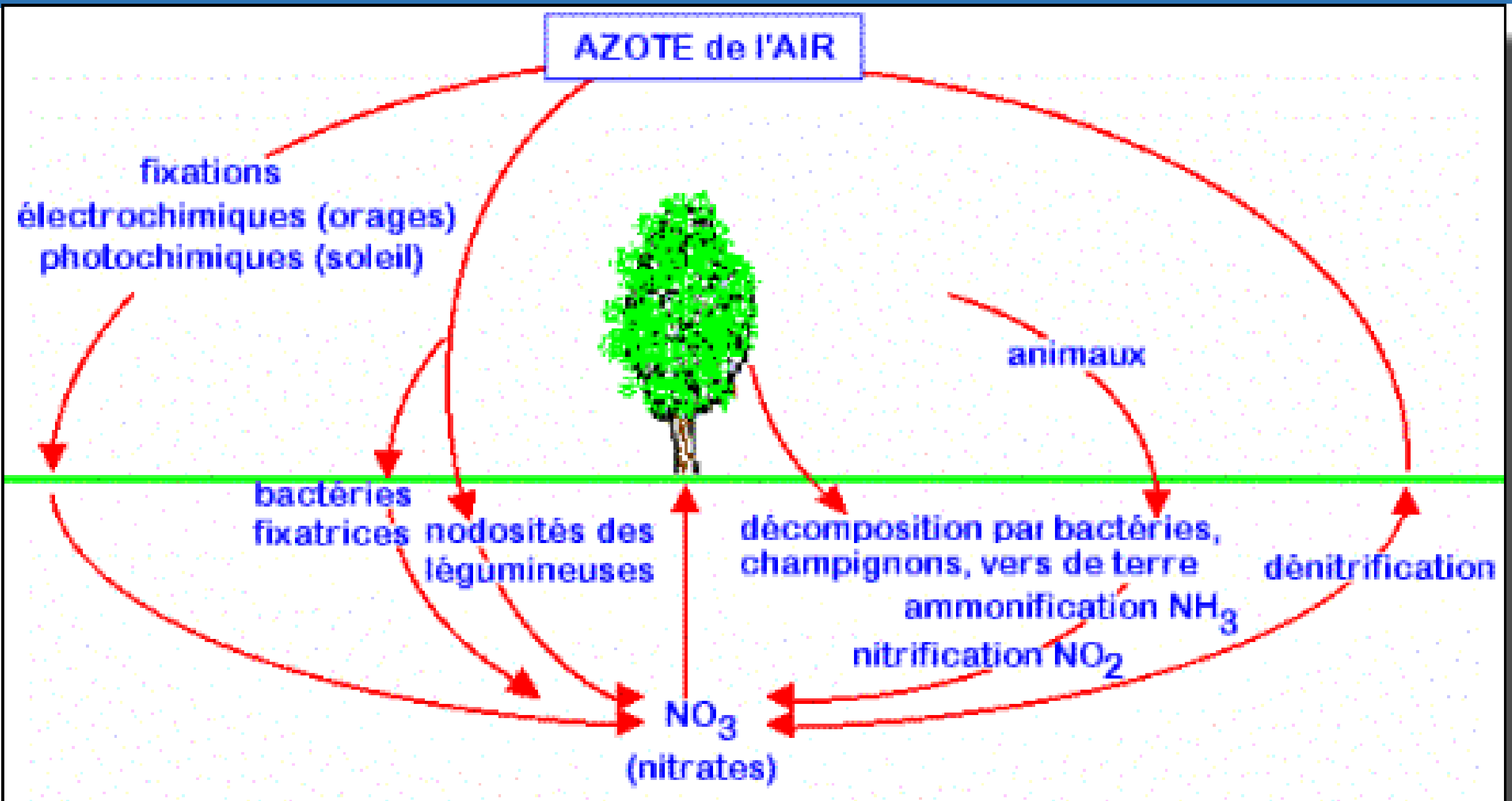




Qu'en dites-vous?



Qu'en dites-vous?



Qu'en dites-vous?



Qu'en dites-vous?

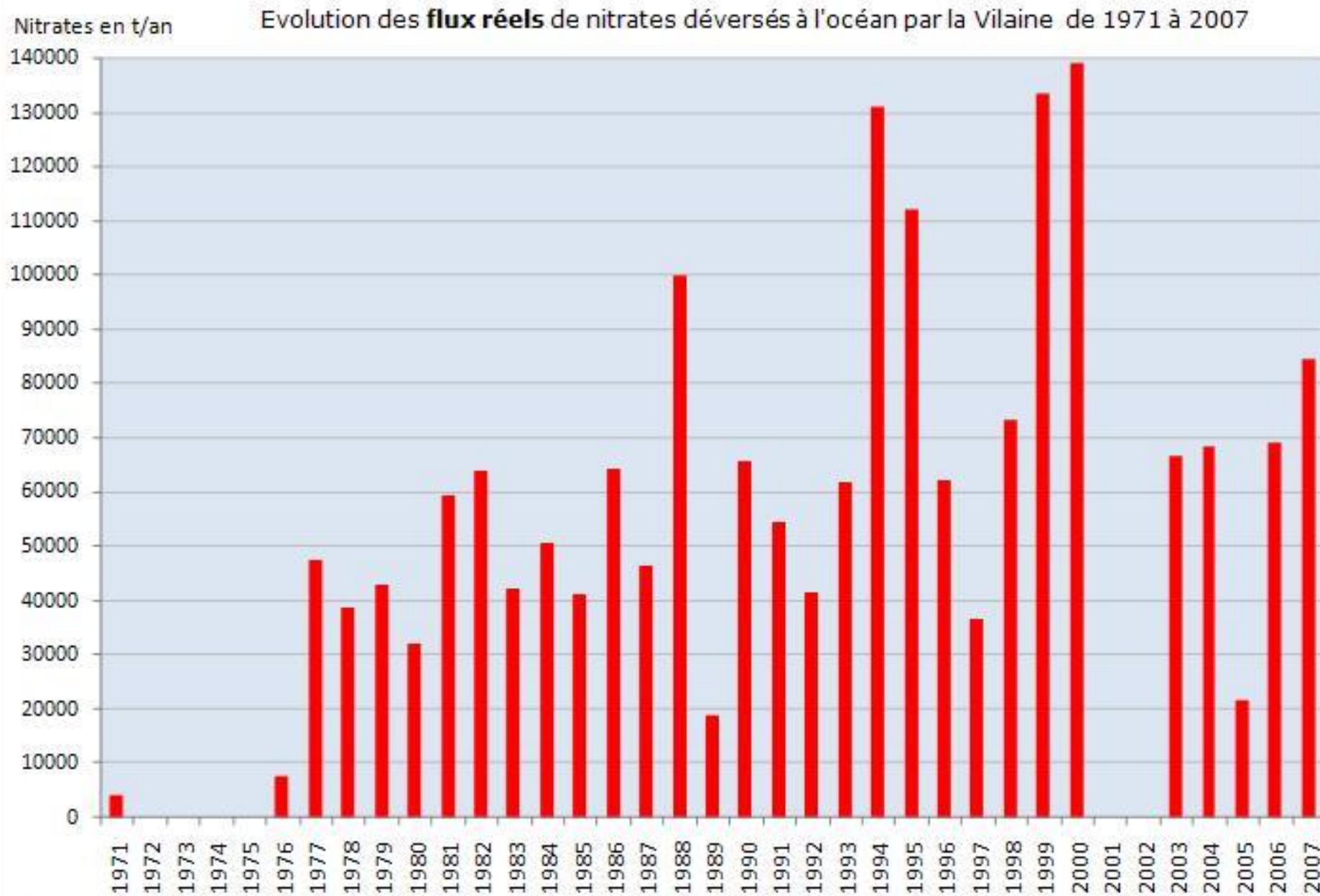
Nitrates dans les cours d'eau bretons

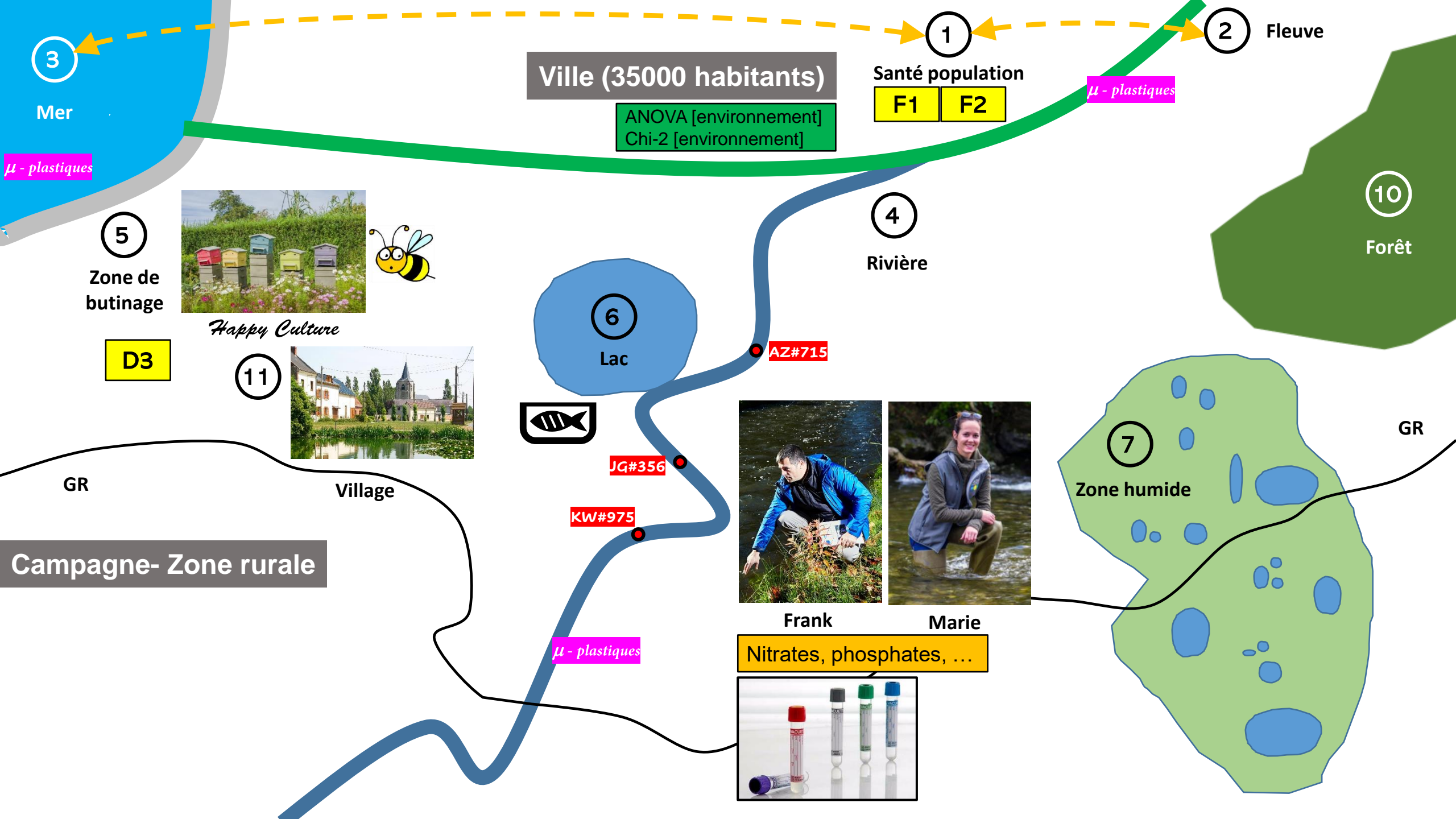
Source : Observatoire de l'environnement en Bretagne

Bon état ($\leq 10\text{mg/l}$)	1%
Etat moyen ($\leq 25\text{mg/l}$)	17%
Etat médiocre ($\leq 50\text{mg/l}$)	59%
Mauvais état ($> 50\text{mg/l}$)	17%
Non évalué	6%



Qu'en dites-vous?



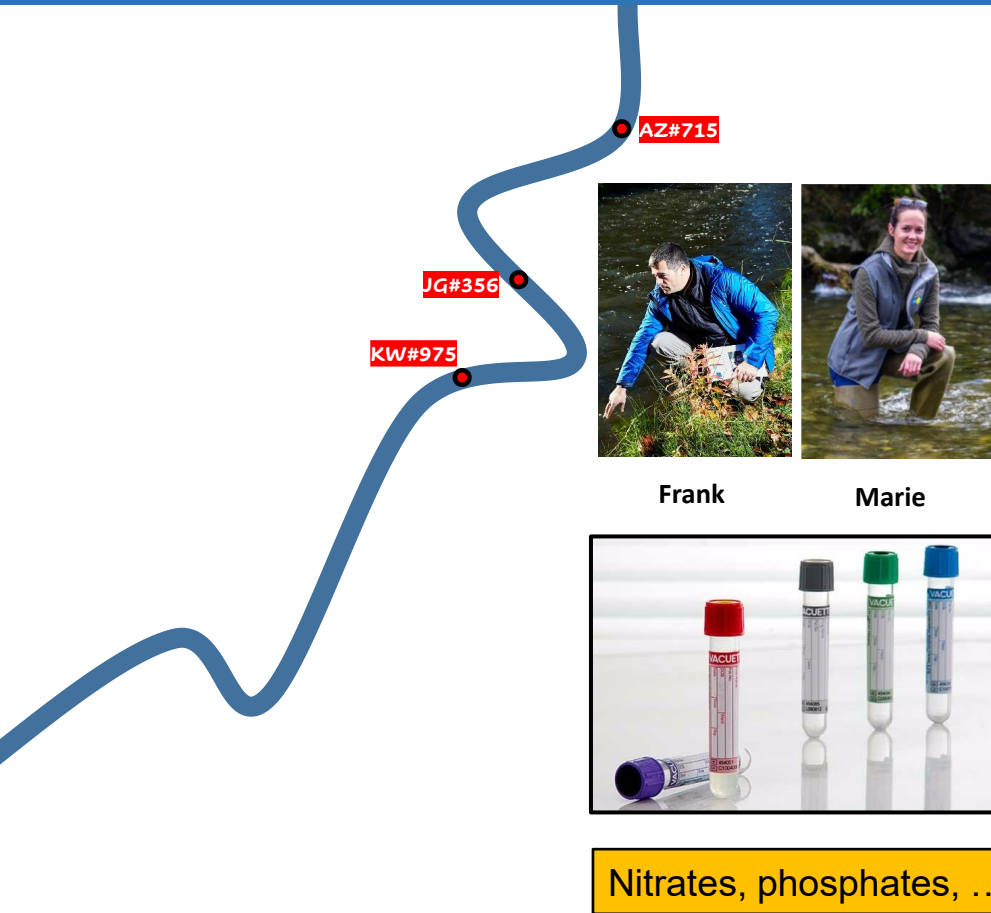



Que doit contenir le fichier rassemblant les mesures (données, format...) ?

The image shows a Microsoft Excel spreadsheet with a blue line graph plotted across several rows. The graph starts at row 32 and ends at row 11. Three data points are highlighted with red dots and labeled with codes in red boxes: KW#975 at row 19, JG#356 at row 16, and AZ#715 at row 11. Below the graph, there are two photographs of people: Frank (a man in a blue jacket) and Marie (a woman in a grey jacket). Below these photos is a photograph of four test tubes, each labeled 'Nitrates, phosphates, ...'. The spreadsheet interface includes a formula bar at the top, a grid of columns (A-R) and rows (1-38), and a status bar at the bottom showing 'Prêt' and 'Feuil1'.

- Utilisation ?
- Données ?
- Variables ?
- Individus ?
- Lignes et colonnes ?
- Population ?
- Exploitation des données ?
- Ce qu'il faudra surveiller ?

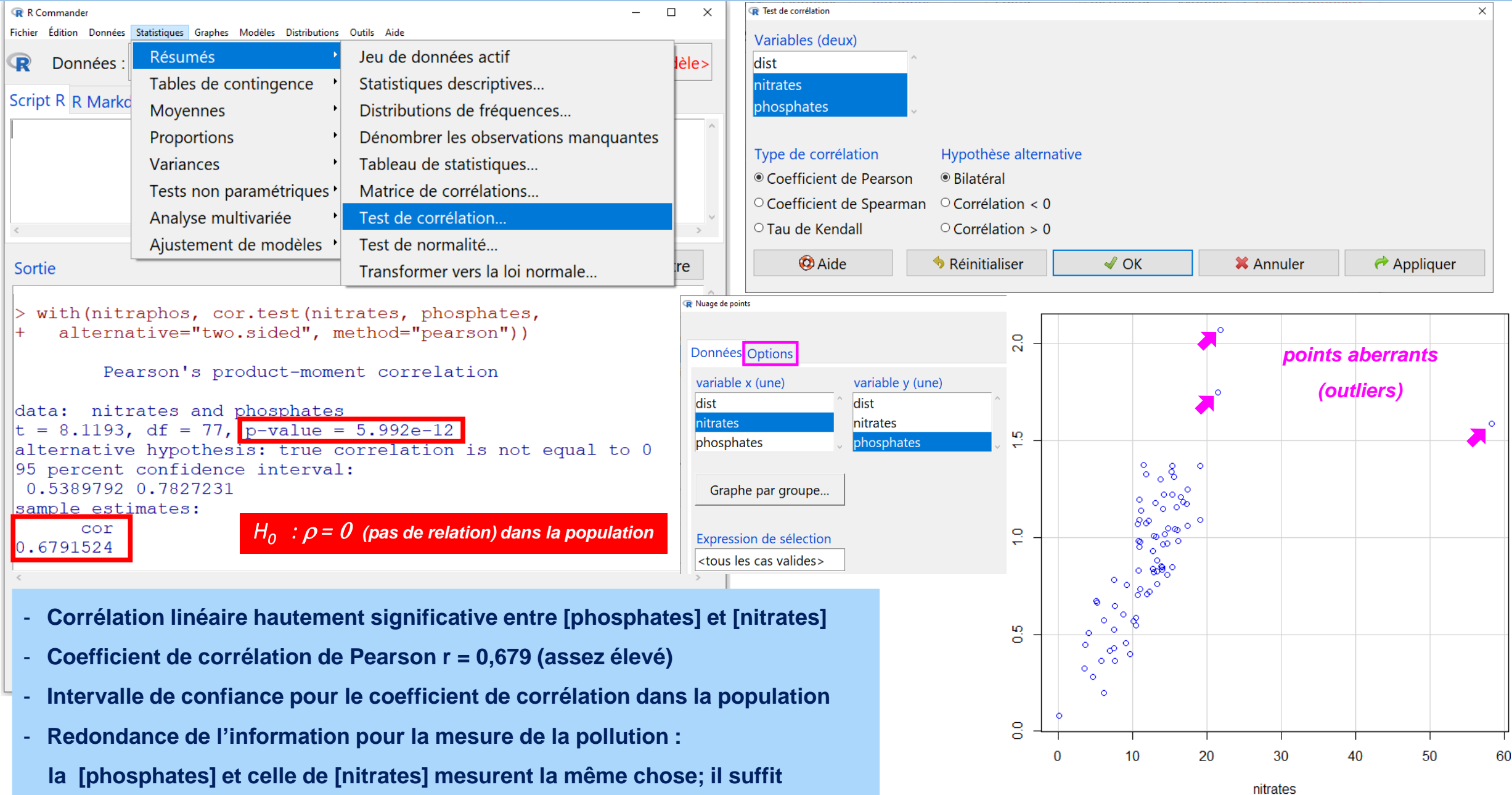
Que doit contenir le fichier rassemblant les mesures (données, format...) ?



ID tube	site 	nitrates	phosphates	ingénieur	dist
1	AZ#715	3,48	0,32	Frank	600
2	AZ#715	8,71	0,60	Marie	600
3	AZ#715	7,54	0,78	Frank	600
4	AZ#715	7,55	0,36	Marie	600
5	AZ#715	0,06	0,08	Marie	600
6	AZ#715	7,43	0,43	Frank	600
7	AZ#715	10,78	0,98	Marie	600
8	AZ#715	7,53	0,52	Marie	600
9	AZ#715	9,07	0,45	Frank	600
10	AZ#715	6,95	0,42	Marie	600
11	AZ#715	7,55	0,65	Frank	600
12	AZ#715	10,75	0,83	Frank	600
13	AZ#715	4,08	0,51	Frank	600
14	AZ#715	9,21	0,75	Marie	600
15	AZ#715	4,67	0,28	Marie	600
16	AZ#715	10,70	0,70	Frank	600
17	AZ#715	6,14	0,57	Frank	600

- Colonnes \Rightarrow variables (VD : mesures; VI : variables fixées)
- Lignes \Rightarrow individus : prélèvements effectués/tubes/ « échantillons »
- Population \blacktriangleright populations au pluriel = les différentes zones explorées/étudiées de la rivière
- Exploitation des données \Rightarrow utilisation de Rcmdr (ou autre logiciel) pour réaliser stats descriptives et inférentielles
- Ce qu'il faudra surveiller \Rightarrow Entre autres, les valeurs manquantes et les valeurs aberrantes.

Corrélation : jusqu'à quel point l'information est-elle redondante?



- Corrélation linéaire hautement significative entre [phosphates] et [nitrates]
- Coefficient de corrélation de Pearson $r = 0,679$ (assez élevé)
- Intervalle de confiance pour le coefficient de corrélation dans la population
- Redondance de l'information pour la mesure de la pollution :
la [phosphates] et celle de [nitrates] mesurent la même chose; il suffit de considérer l'une ou l'autre pour avoir une évaluation de la pollution