Livrable Final Phase 3

Restitution des traitements statistiques pour les 4 cas d'étude

Projet Aerm

Application de techniques statistiques pour l'analyse exploratoire des données de suivi de la qualité des rivières

Détails Document	
Nom	**BPM_AERM_Livrable_Final_Phase3.pdf**
Auteur	BPM-Conseil
Date	11/09/2017

Livrable Final Phase 3

Restitution des traitements statistiques pour les 4 cas d'étude

Projet Aerm

Application de techniques statistiques pour l'analyse exploratoire des données de suivi de la qualité des rivières

Détails Document	
Nom	**BPM_AERM_Livrable_Final_Phase3.pdf**
Auteur	BPM-Conseil
Coordonnées	Siège social : 9 rue Pierre Blanc CP 1160 69203 LYON Etablissement : 39 rue de la Cité 69003 LYON aerm@bpm-conseil.com 06 09 20 47 59 479 400 129 000 17
Date	11/09/2017

Date de révision	Auteur	Commentaires
26/06/2017	BPM-Conseil	Création initiale du document
24/08/2017	BPM-Conseil	Livraison AERM
11/09/2017	BPM-Conseil	MAJ document aux clauses d'édition AERM

RESUME

Dans le cadre de son étude « Application de Techniques Statistiques pour l'Analyse Exploratoire des Données de Suivi de la Qualité des Rivières », l'Agence de l'Eau Rhin-Meuse a mandaté la société BPM Conseil, et son expertise statistique, dans le but de faciliter l'aide à la décision des experts métiers en charge de la validation des données.

En effet, pour assurer ses missions de connaissances de la qualité des eaux des milieux, l'Aerm collecte des volumes croissants de données. Il devient donc complexe de les exploiter convenablement d'une part, et d'en extraire de l'information porteuse et éloquente d'une autre part.

Afin de répondre aux différents cas d'étude, BPM Conseil a fourni une série de méthodologies statistiques, détaillées dans ce document, permettant de dégrossir certains sujets métiers mais surtout d'apporter des éléments de réponses concrets.

Dans un second temps, cette étude a permis à l'Agence de l'Eau, d'améliorer drastiquement, sa connaissance structurelle des données stockées, mais également de repérer les informations peu ou mal relevées.

Pour cela, l'étude menée et composée de quatre cas particuliers, a permis :

- D'identifier les données anormales par un algorithme de spécifications des valeurs fortes, en fonctions des corrélations 2 à 2 des paramètres polluants.
- En utilisant la méthode d'Analyse en Composantes Principale couplée d'un k-means, deux cartes d'identité ont été proposées ; l'une par groupe de station aux comportements de pollution similaires, et une autre par station à l'unité pour déceler des prélèvements dits « hors-norme ».
- De déterminer des cycles et des pics, haut ou bas, sur chaque paramètre, en utilisant la méthode d'autocorrélation totale et partielle en fonction de la sporadicité avérée des mesures de polluants.
- Une méthode ANOVA a également démontré l'impact des phases hydrologiques sur les valeurs des polluants mesurées dans les eaux des milieux.
- De caractériser, à partir d'une régression Stepwise (pas à pas), l'impact de la composition et l'occupation des sols des bassins versants et corridors des cours d'eau sur les valeurs mesurées des polluants.
- Un modèle de risque de présence de valeurs fortes de paramètres polluants dans les cours d'eau non surveillés a ensuite été élaboré en utilisant l'algorithme de construction des arbres de décision.

TABLE DES MATIERES

INTRODUCTION

Ce document constitue le dernier livrable du marché **intitulé « Application de techniques statistiques pour l'analyse exploratoire des données de suivi de la qualité des rivières »** pour l'Agence de l'Eau Rhin-Meuse.

Conformément à ce qui a été approuvé en début de marché, ce document marque le terme de la troisième et dernière phase du marché, au cours de laquelle il est prévu la capitalisation des deux premières phases.

Ainsi, ce document retrace toute la réflexion établie dans le choix des méthodes de traitement statistique et synthétise l'ensemble des résultats obtenus pour répondre aux problématiques métier qui cernaient les quatre cas d'étude suivants :

- **Cas 1** – Identification des données anormales
- **Cas 2** – Variabilité temporelle des détections de polluants dans les eaux et identification des phénomènes cycliques et de leur période
- **Cas 3** – Origine des HAP, métaux et pesticides
- **Cas 4** – Analyse exploratoire et modèle de risque

Toutes les analyses menées sont établies sur la base des besoins métier précisément décrits dans le document *Spécifications des analyses de cas.pdf* fourni par l'AERM en début de marché.

Les résultats proposés sont issues de la seconde phase de développement et restitués dans le document *BPM_AERM_Livrable_Final_Phase2.pdf*. Ils sont par conséquence le fruit des échanges nombreux et réguliers entre BPM et l'AERM tout au long de la période de réalisation des quatre cas.

Il est important de noter que certaines méthodes proposées au cours de la phase 1 se sont ainsi affinées au cours de la phase 2, et de nouvelles techniques que nous croyons pertinentes dans le cadre de l'analyse sont nées d'une réflexion plus approfondie.

Le second objectif de ce livrable est de détailler les modalités de réutilisation des résultats des cas d'études pour la réponse de cas similaires. En effet tout le processus d'extraction de connaissances métiers à partir des données réalisés pour répondre aux quatre cas d'étude doit permettre de fournir une base solide de savoirs métiers capables de diriger d'autres problématiques statistiques.

NE SERONT PAS EXPLICITEES DANS CE DOCUMENT LES MODALITES DE SELECTION DES DONNEES NI LES METHODOLOGIES DE CALCUL SUR LES DONNEES.

1 CAS 1

IDENTIFICATION DES DONNEES ANORMALES

1.1 Rappel des objectifs

L'objectif de ce cas d'étude est de fournir une analyse statistique permettant de dégager une carte d'identité des points de surveillance. Les résultats obtenus ont pour vocation d'apporter une aide à l'expert métier dans sa prise de décision quant à la validité d'une mesure de prélèvement, et notamment sur la distinction aberration / atypisme.

Pour répondre efficacement aux besoins de l'agence de l'eau une signature par point de surveillance a été proposée sur la base des corrélations entre paramètres, aussi appelées cartes d'identité. La réflexion statistique permettant de répondre efficacement aux différents objectifs s'est déroulée selon les étapes suivantes :

- Identification des covariances entre paramètres
- Détection et spécification des valeurs fortes
  - Détection des valeurs fortes (à la hausse comme à la baisse) dans les séries de valeurs mesurées pour chaque paramètre
  - Spécification métier de la nature d'une valeur forte : aberrante ou atypique
- Détermination d'une carte d'identité des points de surveillance

Les sections suivantes détaillerons la méthodologie sélectionnée au regard de l'objectif ciblé, puis le résultat final obtenu et présenté pour répondre au sujet et enfin une ouverture sur les modalités de réutilisation des procédés statistiques proposés.

1.2 Méthodologie appliquée

La réponse au premier cas d'étude s'est articulée, comme précisé à la section précédente, autour de trois grands sous objectifs.

En effet afin d'identifier les valeurs anormales au sein de la masse de données, il paraissait important de pouvoir déterminer ce qu'est une donnée anormale par rapport à une donnée atypique, pour ensuite définir des groupes de paramètres polluants amenés à fluctuer de façon analogue.

Ainsi ce premier volet s'attache d'une part à identifier les liaisons qui existent entre les différents couples de paramètres pour un même point de surveillance, puis d'une autre part à décider de la validité d'un prélèvement sur une date donnée ou par rapport à la carte d'identité d'un point de surveillance.

1.2.1 Identification des covariances entre paramètres

Partant du principe que si un paramètre propose une valeur de prélèvement anormale sur une date donnée, alors tous les autres paramètres, ou du moins la majorité, sur lesquels nous avons décelés une corrélation forte (coefficient supérieur à 0.5), doivent proposer également une variation significative sur ce même prélèvement.

L'objectif ici est de construire des groupes de paramètres liés, positivement ou négativement, sur lesquels se baser pour le calcul et la spécification des valeurs fortes dans le second objectif de ce cas d'études.

Pour élaborer ces groupes une matrice de corrélation est d'abord proposée pour chaque point de surveillance afin d'en conserver les corrélations les plus fréquentes.

En effet une récurrence signifiera ici que sur un point de surveillance, en fonction d'une certaine probabilité, si l'un des paramètres polluants observe un phénomène anormal, alors ses autres paramètres, parmi ceux les plus souvent corrélés avec le premier, ont également de fortes chances eux aussi de posséder cet écart anormal sur son prélèvement.

Outre cette étape de production du socle d'analyse indispensable dans la suite de notre construction d'algorithme permettant la détection de valeur anormales, toutes les matrices de corrélations calculées, une fois synthétisées constituent une première et immense source d'extraction de connaissances pour l'Agence Eau Rhin-Meuse.

1.2.2 Détection et spécification des valeurs fortes

Disposant de groupes bien définis et d'une matrice synthétique de l'ensemble des couples de paramètres variant de façon analogue, il s'agit maintenant de pouvoir spécifier si une valeur forte observable sur un prélèvement précis d'un paramètre est à rapprocher à une erreur de prélèvement ou à un phénomène plausible car également observable sur un bon nombre de paramètres liés.

Ainsi pour assurer une bonne détection il est tout d'abord nécessaire de sélectionner un bon indicateur de seuil permettant de décider à partir de quel ordre de grandeur une valeur est considérée comme forte.

Telles que nous les avons définies au cours de la phase 2, les valeurs fortes sont les enregistrements d'une série temporelle qui s'éloignent de la moyenne de cette dernière à plus de deux écarts-types, à la hausse comme à la baisse.

Pour assurer cette distinction de valeur forte et pour la bonne comparaison des valeurs entre paramètres, toutes les séries sont normalisées, c'est-à-dire centrées et réduites sur une moyenne et un écart type de référence. Par ce procédé les fluctuations sont certes lissées mais elles sont surtout conservées pour comparer les paramètres entre eux malgré des différences d'unité ou d'échelle.

Une fois décelées, les valeurs fortes sont ensuite extraites et conservées au sein d'un jeu de données.

L'idée ici est de confronter toutes les valeurs fortes observées sur une même date données et sur des paramètres conjointement liés. En d'autres termes des paramètres fortement corrélés pour un point de surveillance doivent nécessairement enregistrer des variations similaires.

De ce principe directeur sont nées deux règles complémentaires de qualification d'une valeur forte :

- Si pour deux paramètres covariants, les dates (mois et année) d'anomalie coïncident, les deux valeurs mesurées sont considérées comme **atypiques**.
- Dans le cas contraire (i.e une seule des deux séries corrélées enregistre une valeur forte), la valeur mesurée est considérée comme **aberrante**.

En fin de chaîne un jeu de données comptabilise, pour une valeur forte relevée sur un paramètre, le nombre de valeur également forte à la même période sur les autres paramètres qui lui sont liés.

De cette façon il est facile de proposer une probabilité permettant aux experts métiers de l'AERM de décider si, sur ce prélèvement, les valeurs sont atypiques ou aberrantes.

1.2.3 Détermination d'une carte d'identité des points de surveillance

Maintenant que nous sommes capables de définir quels types de paramètres fluctuent ensemble et de donner la spécification d'une valeur forte en valeur aberrante ou atypique, le dernier objectif de ce cas d'étude est de proposer une carte d'identité par station.

Pour cela nous avons proposé deux méthodologies distinctes dans leur approche mais complémentaires dans leur résultat.

Dans un premier temps nous avons déployé une analyse factorielle (ACP) basé sur la matrice synthétisée des corrélations des paramètres 2 à 2.

L'idée ici est de pouvoir projeter toutes les stations, en fonction de leurs corrélations duales de paramètres sur un plan orthonormé, puis de procéder à un algorithme de regroupement K-means. Ce dernier cherche à rassembler les points en groupe homogène en fonction des distances existantes entre eux.

En effet lorsque plusieurs points de surveillance observent des corrélations entre paramètres identiques il parait cohérents de pouvoir les rassembler au sein d'une seule et même entité.

La détermination des groupes a pour finalité d'apporter une première réponse dans la validation d'un prélèvement. Dans l'idée de vérifier si un prélèvement n'a pas été échangé, il est possible de comparer la moyenne du prélèvement à une date donnée aux moyennes observées pour chaque groupe.

Nous pouvons donc approximer à quel groupe le prélèvement extrait peut appartenir.

Dans un second temps une analyse plus micro a été menée pour comparer les différents prélèvements station par station en fonction de 3 critères de dispersion : la moyenne, la variance et le Kurtosis (coefficient d'aplatissement)

Cette méthode utilise également une analyse factorielle (ACP) sur laquelle l'éloignement d'un point au centre d'inertie du nuage, indique sur le plan factoriel un écart de ce prélèvement par rapport à la 'norme' du point de surveillance étudié.

Ces indicateurs ont été calculés de sorte à agréger sur une date de prélèvement donnée toutes les valeurs de paramètres mesurées, point de surveillance par point de surveillance. En d'autres termes nous retrouvons pour chaque date de mesure une moyenne, une variance et un Kurtosis indiquant la dispersion de façon générale des valeurs prélevées.

Avec un jeu de données par station, nous avons construit pour chacun, une analyse factorielle (ACP). Ce procédé nous permet d'une part de maximiser l'information contenue dans la dispersion des données sur un repère orthonormé, mais aussi de projeter sur les 2 axes factoriels les différentes dates de prélèvements.

C'est donc sur ce plan que les prélèvements s'écartant de la norme seront positionnés loin du centre d'inertie du nouveau nuage.

1.3 Synthèse des résultats du cas

Afin d'illustrer la méthodologie avancée à la section précédente, cette synthèse des résultats développe et explique pour chaque objectif les réponses que nous avons apportées.

Pour aller plus loin, l'entièreté des étapes de calculs et restitutions produites sont trouvables dans le document livré pour la fin de la phase de développement : *BPM_AERM_Livrable_Final_Phase2_20170609*

1.3.1 Matrice de corrélation des paramètres par points de surveillance

Le premier objectif de ce cas d'étude requérait d'extraire pour chaque point de surveillance les couples de paramètres les plus fortement corrélés. Pour cela la construction d'une matrice de corrélation par point de surveillance nous a permis de projeter un seul et même objet tous les coefficients de corrélation des paramètres 2 à 2.

A la fin du traitement, chaque point de surveillance s'est vu pourvu de sa propre matrice de corrélation, dont nous fournissons ci-dessous quelques exemples d'illustration.

		pH	MES	Couleur Mesurée	Oxygène Dissous	DBO5	DCO	Azote Kjeldahl	Ammonium	Nitrites	Nitrates	Phosphore Total	Cote à l'échelle	Orthophosphates (PO4)	Carbone Organique
		1302	1305	1309	1311	1313	1314	1319	1335	1339	1340	1350	1429	1433	1841
pH	1302														
MES	1305	-0,352													
Couleur Mesurée	1309	-0,188	0,254												
Oxygène Dissous	1311	0,247	-0,406	0,109											
DBO5	1313	-0,300	0,489	0,003	-0,278										
DCO	1314	-0,279	0,874	0,314	-0,396	0,325									
Azote Kjeldahl	1319	-0,270	0,711	0,092	-0,544	0,640	0,651								
Ammonium	1335	-0,264	-0,044	-0,049	-0,204	0,207	0,048	0,165							
Nitrites	1339	-0,470	0,515	0,156	-0,325	0,222	0,507	0,402	0,303						
Nitrates	1340	0,061	-0,148	0,510	0,637	-0,190	-0,053	-0,233	-0,088	-0,046					
Phosphore Total	1350	-0,391	0,404	0,157	-0,414	0,166	0,365	0,428	0,088	0,424	-0,216				
Cote à l'échelle	1429	0,118	0,213	0,508	0,470	0,081	0,213	0,025	-0,165	-0,047	0,728	-0,244			
Orthophosphates (PO4)	1433	-0,105	-0,050	0,019	-0,211	-0,410	0,125	-0,031	0,134	0,249	-0,075	0,676	-0,270		
Carbone Organique	1841	-0,143	0,341	0,805	0,017	0,188	0,423	0,332	0,079	0,173	0,502	0,101	0,575	-0,023	

Extrait de la matrice de corrélation du PS 02041850 – Falkensteinbach à Gundershoffen

		pH	MES	Couleur Mesurée	Oxygène Dissous	DBO5	Hydrogénocarbonates	Ammonium	Chlorures	Sulfates	Nitrites	Nitrates	Silicates	TAC	Phosphore total	Uranium	Bore	Potassium	Arsenic	Magnésium	Titane	Calcium	Sodium	Cobalt
		1302	1305	1309	1311	1313	1327	1335	1337	1338	1339	1340	1342	1347	1350	1361	1362	1367	1369	1372	1373	1374	1375	1379
pH	1302																							
MES	1305	-0,378																						
Couleur Mesurée	1309	-0,086	0,560																					
Oxygène Dissous	1311	0,352	-0,017	-0,030																				
DBO5	1313	0,348	0,288	0,249	0,508																			
Hydrogénocarbonates	1327	0,519	-0,218	-0,026	0,262	0,146																		
Ammonium	1335	-0,300	0,121	0,012	0,468	0,082	0,058																	
Chlorures	1337	-0,043	-0,292	-0,330	-0,318	-0,202	0,320	-0,013																
Sulfates	1338	0,057	-0,480	-0,423	-0,279	-0,316	0,457	-0,104	0,858															
Nitrites	1339	-0,178	-0,090	0,176	-0,446	-0,444	0,076	-0,116	0,200	0,255														
Nitrates	1340	0,025	0,502	0,514	0,365	0,363	0,218	0,195	-0,321	-0,440	0,147													
Silicates	1342	-0,318	0,162	0,175	-0,003	-0,357	-0,047	0,264	-0,251	-0,242	0,439	0,449												
TAC	1347	0,531	-0,219	-0,015	0,281	0,173	0,996	0,048	0,313	0,443	0,056	0,219	-0,077											
Phosphore total	1350	-0,451	0,102	-0,050	-0,675	-0,553	-0,080	-0,152	0,473	0,479	0,494	-0,277	0,214	-0,102										
Uranium	1361	-0,007	-0,042	0,179	-0,057	-0,109	0,240	-0,079	0,099	0,074	0,105	0,311	0,228	0,236	-0,009									
Bore	1362	-0,102	-0,380	-0,155	-0,481	-0,523	0,201	-0,149	0,629	0,629	0,447	-0,388	-0,106	0,200	0,471	0,288								
Potassium	1367	-0,481	-0,162	-0,239	-0,317	-0,338	0,102	0,265	0,670	0,588	0,405	-0,135	0,182	0,076	0,549	0,103	0,542							
Arsenic	1369	-0,399	-0,219	-0,192	-0,666	-0,608	-0,170	-0,109	0,443	0,534	0,414	-0,482	0,148	-0,195	0,774	0,009	0,521	0,519						
Magnésium	1372	0,278	-0,421	-0,299	-0,009	-0,088	0,697	-0,054	0,692	0,889	0,090	-0,322	-0,361	0,692	0,234	0,058	0,471	0,356	0,257					
Titane	1373	-0,289	0,892	0,627	0,119	0,269	-0,254	0,118	-0,514	-0,602	-0,044	0,613	0,300	-0,261	-0,002	-0,007	-0,467	-0,226	-0,270	-0,522				
Calcium	1374	0,337	-0,266	-0,130	0,130	0,030	0,885	0,029	0,619	0,726	0,194	0,138	-0,094	0,880	0,142	0,235	0,364	0,374	0,055	0,833	-0,337			
Sodium	1375	-0,075	-0,315	-0,308	-0,425	-0,315	0,260	-0,101	0,952	0,876	0,282	-0,354	-0,155	0,245	0,592	0,168	0,660	0,693	0,586	0,673	-0,483	0,568		
Cobalt	1379	-0,115	0,031	0,233	0,004	0,016	0,025	0,215	0,282	0,055	0,076	0,076	-0,013	0,047	0,033	0,161	0,420	0,173	0,022	-0,002	-0,019	0,129	0,185	

Extrait de la matrice de corrélation du PS 020450000 – La Moder à Drusenheim

Pour une meilleure lisibilité nous avons ciblé les corrélations les plus fortes, soit supérieures ou égales à 0.5, symbolisées par un code couleur, rouge pour les corrélations inverses, vert pour les corrélations positives.

Disposant de toutes ces matrices de corrélation, c'est-à-dire une matrice par point de surveillance, il nous a fallu résumer leur information au sein d'un seul et même tableau pour identifier les couples de paramètres les plus fréquents et donc les plus utiles.

	2124000	2118000	2116000	2113000	2109000	2106600	2106410	2100150	2098300	2097000	2096900	2096480	2095600	...
1305_1302	-0,1431	-0,0172	-0,1907	-0,1841	-0,1811	-0,1773	-0,3373	-0,3385	-0,3175	0,0567	-0,3777	-0,1656	0,1094	...
1309_1302	-0,2525	-0,0649	-0,3166	0	0	-0,3699	-0,1405	-0,1848	-0,3764	-0,3408	-0,0862	-0,1419	0,0961	...
1311_1302	0,2067	0,6574	0,2348	0,4822	0,5116	0,5934	0,5161	0,5308	0,0616	0,3044	0,3521	-0,1013	0,389	...
1313_1302	0,1374	0,1702	-0,1464	0,0839	-0,063	0,09	0,4297	0,0939	-0,336	0,2069	0,348	0	0,1785	...
1327_1302	0,2642	0,4834	0,2276	0,5036	0,5297	0,0601	0,2176	0,5938	0	0	0,5185	0	0,1737	...
1337_1302	0,0237	-0,3713	0,0235	-0,2581	-0,2958	-0,2182	-0,0803	-0,0103	0	0	-0,0433	0	-0,3523	...
1338_1302	-0,0246	-0,3923	-0,0133	0,0481	-0,193	-0,1564	-0,2018	0,2042	0	0	0,0574	0	-0,2542	...
1339_1302	-0,0699	-0,2388	-0,1492	-0,3378	-0,3125	-0,1144	-0,2196	-0,4705	-0,4343	-0,187	-0,1782	-0,1045	-0,2426	...
1340_1302	0,3093	0,7169	0,2543	0,5242	0,3734	0,0641	0,0924	0,1701	0,1653	0,003	0,0247	0,018	0,2775	...
1342_1302	-0,2833	0,0256	-0,1252	0,17	0,0449	0,0744	-0,4751	-0,1668	0	-0,3235	-0,3183	0	0,0295	...
1347_1302	0,2905	0,4955	0,2487	0,5759	0,5608	0,0864	0,2802	0,5946	0	0	0,5313	0	0,1742	...
1350_1302	-0,2151	-0,3493	-0,1262	-0,224	0,0684	-0,3024	-0,6723	-0,657	-0,2374	-0,2291	-0,4506	-0,0507	-0,4268	...
1361_1302	0,0196	-0,0371	0,1811	0,0832	0,148	0,1577	0,1305	0,1112	0	0	-0,0072	0	0,1125	...
1362_1302	0,403	0,1718	0,5243	0,3081	-0,0735	-0,0839	-0,1287	-0,1599	0	0	-0,1019	0	-0,2189	...
1367_1302	-0,1823	-0,2457	-0,2022	-0,3299	-0,4015	-0,3378	-0,4689	-0,2545	0	0	-0,4813	0	-0,3686	...
1372_1302	0,2005	0,2407	0,2398	0,222	0,0144	0,0262	-0,1043	0,5447	0	0	0,2782	0	-0,0611	...
1373_1302	-0,1327	-0,0615	-0,2047	-0,2542	0	-0,2671	-0,3679	-0,2872	0	0	-0,2891	0	0,1862	...
1374_1302	0,3551	0,4803	0,3712	0,5767	0,4545	0,5184	-0,17	0,5926	0	0	0,3365	0	0,1572	...
1375_1302	-0,0169	-0,4608	-0,0134	-0,2297	-0,3079	-0,2647	-0,1274	-0,1148	0	0	-0,0753	0	-0,2697	...
1379_1302	0,2924	0,3933	0,4405	0,1636	0,1649	-0,2496	0,2029	0,0039	0	0	-0,1152	0	0	...
1383_1302	0,1067	0	0,0237	0	0	0	0,0995	-0,1075	0	0	6,00E-04	0	0,2367	...
1384_1302	-0,1591	-0,592	0,1315	-0,2933	-0,2984	-0,3179	-0,4962	-0,5813	0	0	-0,6328	0	-0,1822	...
1386_1302	0,1005	0	0,1686	0	0,0308	-0,0733	0,2625	-0,008	0	0	0,0721	0	0,1448	...
1392_1302	-0,0992	0,0907	-0,0329	-0,1599	-0,1553	-0,0458	-0,0871	-0,2689	0	0	-0,346	0	0,0635	...
1396_1302	0,2608	-0,0577	0,2828	0,2717	0,0499	-0,0508	-0,1418	0,1774	0	0	0,005	0	-0,1468	...
1433_1302	-0,3729	-0,5012	-0,1127	-0,2028	-0,002	-0,2953	-0,6117	-0,5729	0,0158	-0,2076	-0,3703	0,0187	-0,4243	...
1436_1302	0,1355	0,037	-0,0716	-0,125	-0,3257	-0,1256	0,141	-0,0388	0	0,181	0,1782	0	-0,0655	...
1439_1302	0,0707	0,1496	-0,0627	0,0672	-0,067	-0,172	0,5323	0,2983	0	0,0595	0,3231	0	-0,0585	...
1841_1302	-0,282	-0,1151	-0,2342	-0,1547	-0,2532	-0,4717	-0,2199	-0,2496	-0,5504	-0,3205	-0,3094	-0,1605	-0,2148	...
1907_1302	-0,2292	-0,5212	-0,0938	0	0,0869	-0,3711	-0,3924	-0,3831	0	0	-0,2351	0	-0,3275	...
7073_1302	0,04	-0,142	-0,2332	-0,2528	-0,0743	0,1287	0,3141	0,1631	0	0	-0,0518	0	-0,3146	...
1309_1305	0,3347	0,5958	0,4193	0	0	0,6582	0,4427	0,54	0,7513	0,1952	0,5596	0,7057	0,533	...
1311_1305	0,1026	0,2945	0,173	-0,0803	0,0977	0,0379	-0,1038	0,006	0,2357	0,2524	-0,0167	0,2086	0,0015	...
1313_1305	0,1288	0,1941	0,3182	0,2395	0,3797	0,2824	0,0879	0,2331	0,1412	0,1969	0,2875	0	0,3237	...
1327_1305	-0,42	-0,2868	-0,556	-0,2909	-0,4809	-0,616	-0,611	-0,4437	0	0	-0,2183	0	0,0065	...
1337_1305	-0,5238	-0,6099	-0,6048	-0,4667	-0,4924	-0,3531	-0,5143	-0,4188	0	0	-0,2921	0	-0,2841	...
1338_1305	-0,4703	-0,513	-0,6341	-0,1737	-0,3423	-0,7067	-0,5361	-0,5626	0	0	-0,4799	0	-0,3043	...
1339_1305	0,3501	0,3836	0,2909	0,4621	0,5009	-0,1631	0,1636	-0,0653	0,263	-0,0212	-0,0901	0,2531	0,0388	...
1340_1305	0,1241	0,1313	0,1884	-0,0072	0,0713	0,2085	0,4837	0,1202	0,1401	0,2154	0,5021	0,3212	0,4242	...
1342_1305	0,3525	0,5018	0,4071	0,4221	0,4419	0,374	0,0619	-0,0333	0	-0,0648	0,1623	0	0,1341	...
1347_1305	-0,4588	-0,3078	-0,6031	-0,3182	-0,4814	-0,626	-0,6115	-0,4436	0	0	-0,2185	0	0,0112	...
1350_1305	0,3379	0,7086	-0,0101	0,2497	0,5603	-0,1421	0,1662	0,3658	0,4893	-0,2048	0,1015	-0,1371	0,0256	...

Extrait de la matrice des corrélations obtenue

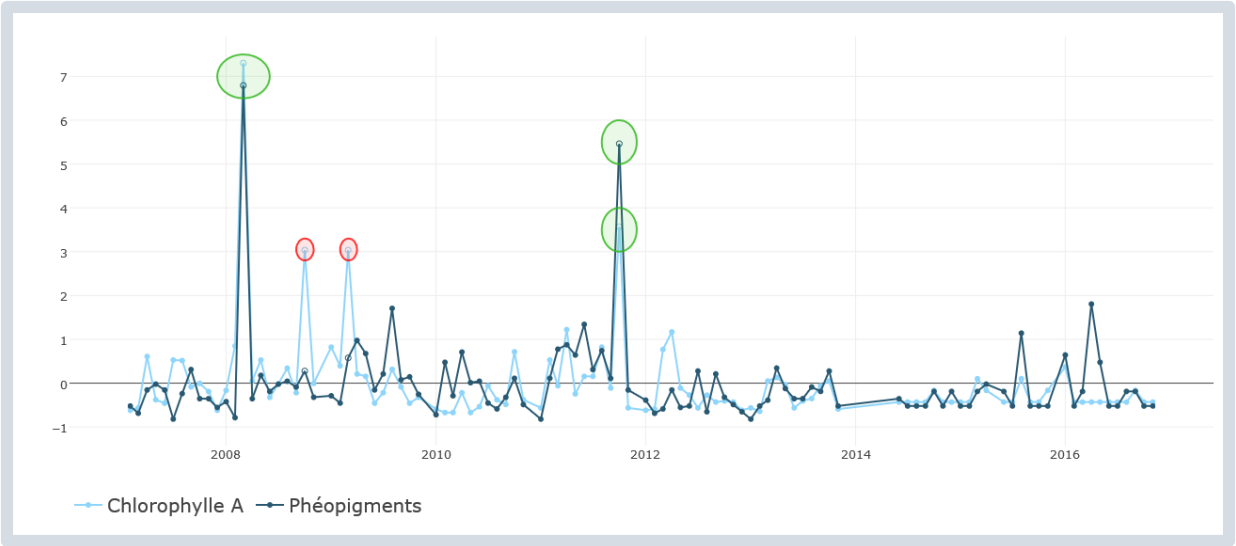
Ce tableau de synthèse décrit en effet pour chaque couple de paramètres distinct repéré le niveau de corrélation retrouvé par station.

Ainsi sur une ligne, i.e. pour un couple de paramètre, plus les coefficients sont forts et récurrents sur les points de surveillance identifiés en colonnes, plus cette combinaison de paramètres devient intéressante à étudier par l'Agence de l'Eau Rhin-Meuse.

1.3.2 Spécifications des pics de concentrations

La seconde partie de ce cas d'études réside dans la détermination des valeurs fortes relevées en valeur atypique ou en valeur aberrante. En effet un pic de concentration peut, par définition métier, être expliqué si ce phénomène est également observable sur les autres paramètres.

Cette étape nécessite la construction de la matrice des corrélations agrégée car l'analyse des phénomènes anormaux et identiques ne s'établira maintenant que sur les paramètres fortement liés et non plus sur la masse entière des paramètres.



Qualification métier des valeurs fortes pour la Chlorophylle A (1433) et les MES (1305)
Sur le PS 02020000

Dans l'exemple ci-dessus nous avons décelé une corrélation forte entre les paramètres 1433 et 1305. En comparant les 2 courbes normalisées nous pouvons valider que 2 points anormaux peuvent être potentiellement spécifiés comme atypique comme les 2 courbes fluctuent conjointement (cerclés en vert) et que 2 autres sont potentiellement aberrants car seul l'un des 2 paramètres observe un pic particulier.

Fort de cette détection de points forts et de cette spécification du type de points anormaux en atypique ou aberrant nous pouvons dorénavant pour une date donnée, agréger les différents éléments pour offrir une probabilité de classification du point en atypisme réel ou aberrance confirmée.

En d'autres termes pour un même point de surveillance, un même paramètre peut être corrélé à plusieurs autres, ce qui va donner lieu à autant de comparaison graphique pour déterminer la nature d'une valeur forte ; certains prélèvements peuvent alors apparaître tantôt atypiques, tantôt aberrants, selon le paramètre corrélé mis en comparaison.

Dans ce cas, un **bon indicateur de la significativité du diagnostic** serait de calculer le ratio classement atypique / aberrant, comme nous le présentons sur l'exemple ci-dessous.

PS	Code Paramètre	Catégorie	Date	Quantité
02001050	1340	**Aberrant**	11/2009	8 **(80%)**
02001050	1340	**Atypique**	11/2009	2 **(20%)**

Résultat du diagnostic pour le PS 02001050 avec les nitrates (1340) au mois de septembre 2009

D'après l'exemple ci-dessus, on peut **diagnostiquer une valeur aberrante en septembre 2009 pour le paramètre 1340 et le PS 02001050 à un seuil de 80%.**

1.3.3 Détermination d'une carte d'identité des points de surveillance

Conformément à la méthodologie annoncée en section précédente la construction des cartes d'identité a été formulée en combinant 2 méthodes. Le but principal ici est de proposer aux agents de l'AERM des éléments de réponses sur des cas probables d'inversion de prélèvements entre 2 stations par un laboratoire

La première se destine à rassembler les stations par profils types afin de rapprocher les stations entre elles en fonctions des corrélations de couples de paramètres identiques.

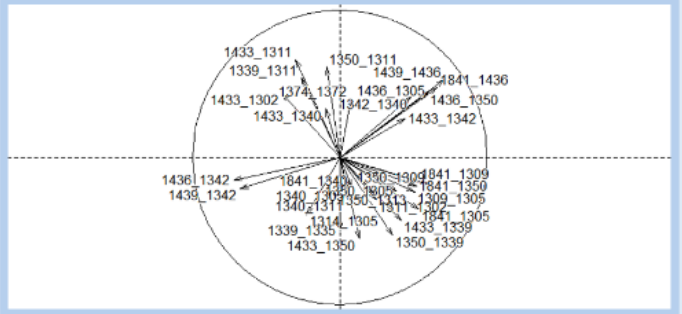
La seconde propose une analyse individuelle par station, capable de définir une norme des niveaux des concentrations permettant de détecter les prélèvements déviants.

1.3.3.1 Utilisation de la carte par groupes

Cette construction de groupes de stations, réalisée à partir de la combinaison d'une analyse factorielle et d'un algorithme de type k-means, a pour but de réduire le champ des possibles lors de la recherche d'inversion de prélèvements des stations.

En effet cette méthodologie permet comme énoncé plus haut de rassembler des stations en fonctions de leurs caractéristiques communes, ici leurs couples de paramètres fortement corrélés. Ces regroupements permettent donc de fournir un premier élément de réponse dans la confirmation d'inversion de prélèvement.

Cercle des corrélation au seuil de récurrence de 10% (50 paires de paramètres retenues)



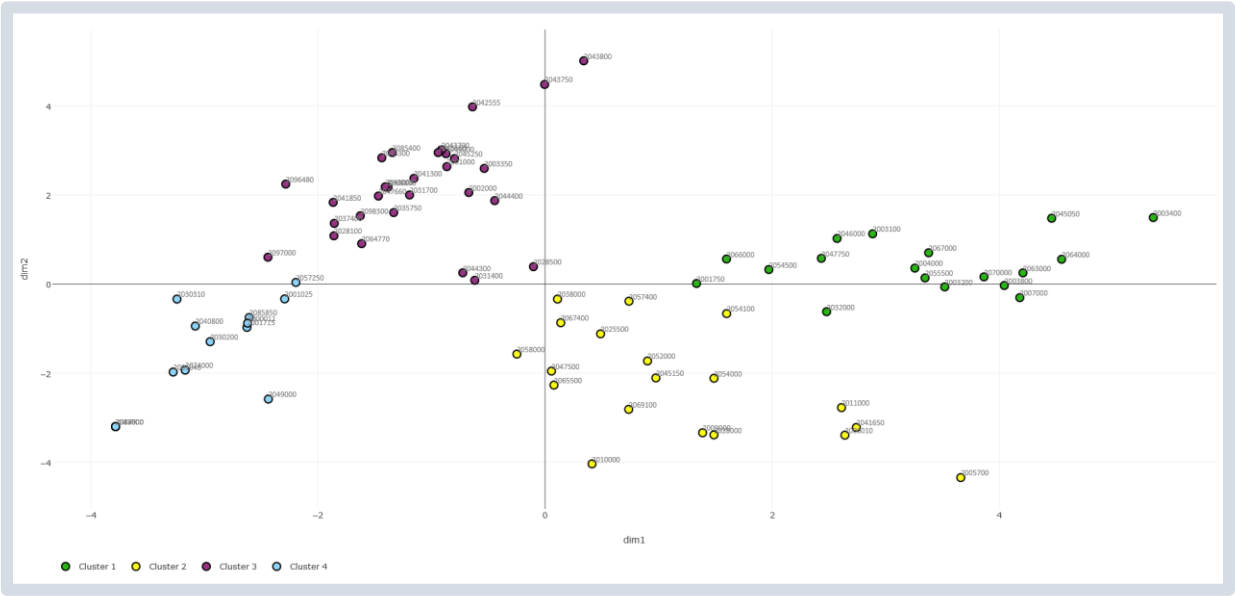
	Axe 1			Axe 2		
Inertie retenue	18,44%			15,31%		
	Codes paramètres	Familles	Noms	Codes paramètres	Familles	Noms
Couples représentatifs	1439_1436	9_9	Chlorophylle A_Phéopigments	1339_1335	3_NA	Nitrites_Ammonium
	1841_1436	1_9	Carbone Organique_Phéopigments	1350_1339	1_12	Phosphore total_Nitrites
	1436_1305	9_12	Phéopigments_MES	1433_1350	NA_12	Orthophosphates_Phosphore total
	1436_1350	9_3	Phéopigments_Phosphore total	1314_1305	1_NA	DCO_MES
	1350_1309	3_NA	Phosphore total_Couleur mesurée	1433_1339	3_1	Orthophosphates_Nitrites

Résultat de l'ACP des différentes couples de paramètres

Sur cette analyse factorielle nous observons **33% d'inertie du nuage**, cependant même s'il n'est pas possible de déterminer une signification métier aux axes, l'objectif principal de l'ACP est rempli, c'est-à-dire **résumer l'information** sur repère orthonormé afin de pouvoir projeter les stations de surveillance et les classer en groupe aux caractéristiques communes.

La mise en place de l'algorithme va en effet pouvoir regrouper les différentes stations projetées afin de les regrouper en fonction de leur distance sur le plan factoriel résumant au mieux l'information contenu dans la matrice agrégée des corrélations.

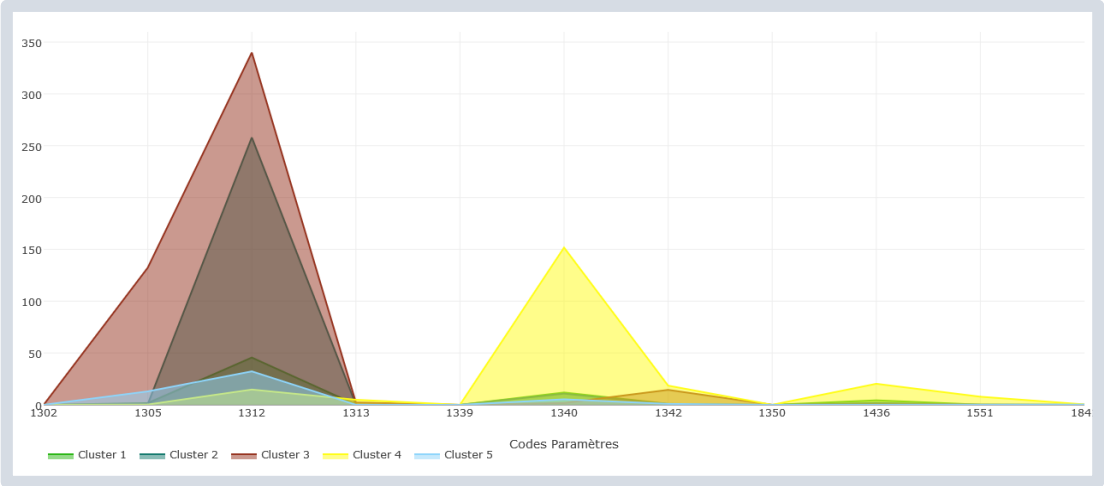
Le plus de cette méthode de regroupement k-means c'est que le nombre de groupe à retrouver est déterminé en amont par l'utilisateur. De cette façon il est possible de tester plusieurs paramétrages sur le nombre de groupe, mais aussi sur la distance entre points utilisée par exemple.



Visualisation des groupes calculés par le k-means (k = 4)

Connaissant les différents groupes de stations, nous pouvons extraire une date de prélèvement suspecte afin de vérifier son rapprochement de tel ou tel groupe de station.

Pour cela nous comparons les carrés des écarts à la moyenne des groupes de stations et les valeurs observées sur le prélèvement extrait paramètre par paramètre. Plus les valeurs du prélèvement sont proches d'un groupe plus il a de chance d'appartenir à ce groupe.



La ligne d'abscisse correspond au prélèvement du PS 0207000 du 20/02/2007

Les valeurs du cluster 1 (verte) semblent se rapprocher le plus de ce dernier

En effet si le prélèvement extrait se rapproche du groupe de stations auquel appartient sa station de base alors cela implique deux possibilités :

- Le prélèvement a de forte chance d'être conforme
- Le prélèvement peut cependant appartenir à une autre station de son groupe.

Si à l'inverse le prélèvement extrait se rapproche d'un autre groupe de stations que celui d'appartenance de sa station de base alors cela implique deux possibilités :

- Le prélèvement a de forte chance d'appartenir à une autre station
- Le prélèvement peut cependant aussi être un prélèvement sur une journée atypique

1.3.3.2 Projection individuelle

Nous savons dorénavant rapprocher un prélèvement à un ensemble de stations précis, mais dans le cas où le prélèvement suspect appartient bien au groupe de la station dont il a été extrait, il est nécessaire de pouvoir l'étudier comparativement aux autres prélèvements de sa station d'origine.

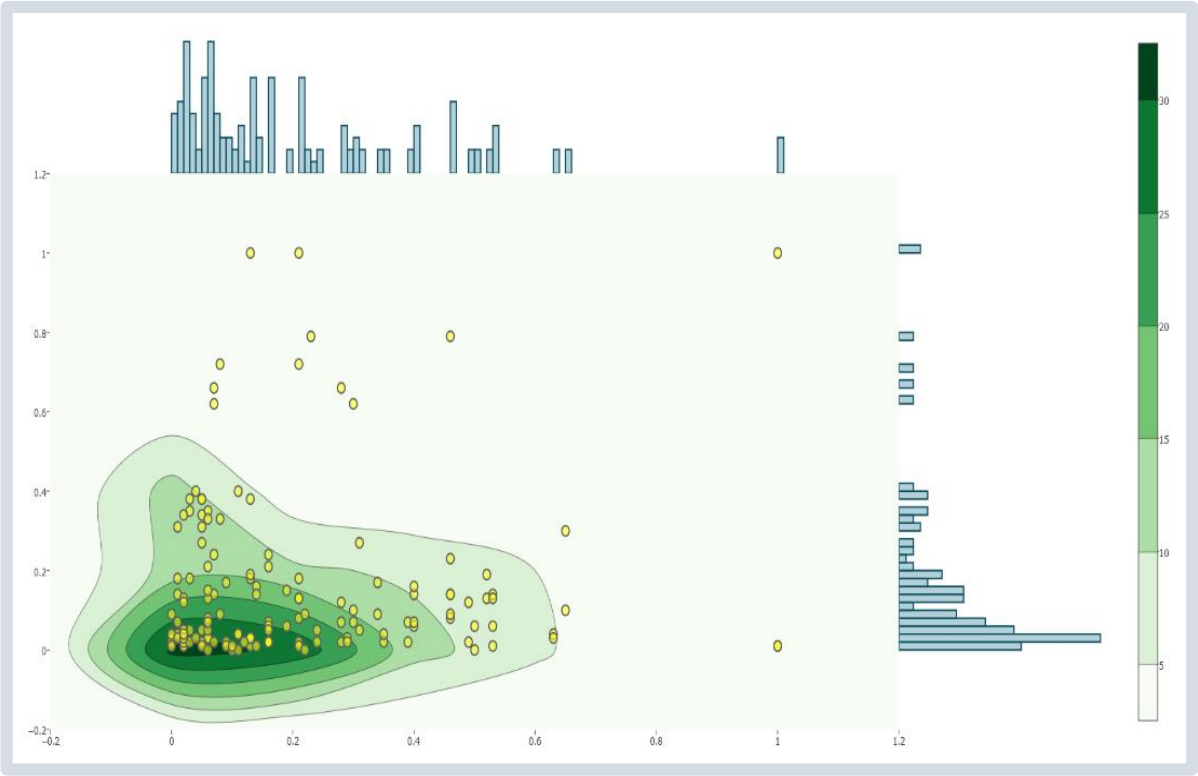
Pour réaliser cela nous agrégeons sur chaque date de prélèvement les différentes valeurs de paramètres observées sous 3 critères de dispersion distincts : la moyenne, la variance et le Kurtosis (coefficient d'aplatissement)

Date	Moyenne	Variance	Kurtose
16/01/2012	0,65	0,99	-1,41
13/02/2012	0,76	1,10	-1,47
13/03/2012	1,12	2,27	-0,94
10/04/2012	0,20	0,75	-0,46
09/05/2012	-0,18	0,25	-0,74
04/06/2012	-0,31	0,25	-1,26
02/07/2012	-0,53	0,21	-1,52
27/08/2012	-0,75	0,62	-0,69
24/09/2012	-0,31	0,60	-0,02
22/10/2012	-0,61	0,29	0,71
19/11/2012	-0,31	0,20	-1,44
17/12/2012	0,73	0,99	-0,34
07/01/2013	0,39	1,11	-0,54
04/02/2013	0,34	0,58	-1,44
04/03/2013	0,55	1,12	-0,65
02/04/2013	0,76	0,54	-0,73
27/05/2013	0,05	0,25	-0,80
...	...	...	...

Extrait des indicateurs de dispersion agrégés sur une station

Ces données sont ensuite analysées au sein d'une nouvelle ACP dans le but identique de résumer le maximum d'informations de la dispersion représentée par les 3 indicateurs sélectionnés sur un plan factoriel.

Ensuite en projetant les différentes dates de prélèvement sur le nouveau repère calculé alors nous pouvons aisément calculer les densités d'apparition des données et repérer les prélèvements qui s'éloignent de la 'norme' calculée.



Carte d'identité du point de surveillance 0200010

Sur la représentation graphique ci-dessus nous observons grâce au dégradé des patatoïdes verts les différents niveaux de densité d'apparition des prélèvements.

En d'autres termes le centre d'inertie du nuage de point maximisé par le calcul du plan factoriel se retrouve dans la zone la plus sombre. C'est à cet endroit que la plupart des prélèvements se regroupent, on peut logiquement appeler cette zone l'identité du point de surveillance.

Les points les plus éloignées représentent des dates de prélèvements à priori hors norme, qui mériteraient d'être extraites du jeu de données pour être étudiées de façon plus détaillée.

1.4 Modalités de réutilisation

Composé de plusieurs méthodologies, ce premier cas d'étude propose à la fois la mise en place d'analyses factorielles seules ou couplées à un algorithme de regroupement, la normalisation des données ou encore la recherche de valeurs fortes.

Nous avons en effet vu la puissance délivrée par la mise place d'une ACP en utilisation les corrélations entre paramètres. Classées parmi les méthodes descriptives, les analyses factorielles transforment un nuage de données projetables sur un plan à N dimensions, en un nuage de points visualisable sur un plan orthonormé capable de faire ressortir les meilleures interactions entre variables.

Comme nous l'avons vu, les matrices de corrélation ont été une source immense de réponses métiers et sont à l'origine de la construction des cartes d'identité des stations de surveillance.

En effet nous avons démontré que réutiliser les corrélations entre paramètres nous a permis de spécifier les valeurs comme atypiques ou aberrantes en fonction de la récurrence de phénomène anormaux.

Il va de soi que les règles de décision fournit par la récurrence d'apparition d'un phénomène anormal nécessite bien entendu la validation métier que l'AERM pourra enrichir ou agrémenter car connaissant généralement l'origine de pics fort sur certains jours de prélèvements particulier.

L'utilisation du modèle de regroupement (k-means) proposé pour ce cas d'étude résulte d'un cas classique de combinaison d'analyse factorielle et d'algorithme de classement.

En d'autres termes à partir d'un immense jeu de données, nous recherchons d'une part les interactions et d'une autre part la position des individus les uns par rapport aux autres.

En conclusion l'ACP permet de réduire les différences entre les individus et la méthode des k-means cherche à les rassembler en groupe distinct. La définition des groupes est proposée grâce aux interactions obtenues dans l'ACP.

Pour aller plus loin, la combinaison des deux méthodes permet d'obtenir une variable discriminante de classification réutilisable dans méthodes de classification automatique comme les arbres de décision utilisés par exemple dans le cas 4.

2 CAS 2

VARIABILITE TEMPORELLE DES DETECTIONS DE POLLUANTS DANS LES EAUX
IDENTIFICATION DES PHENOMENES CYCLIQUES ET DE LEURS PERIODES

2.1 Rappel des objectifs

L'objet d'étude central du Cas 1 était le point de surveillance, dont il nous a fallu dresser la signature à travers l'identification des covariances inter-paramètres caractéristiques et des sites présentant des relevés anormaux comparativement à leur comportement standard.

Pour ce Cas 2, nous nous intéressons plus particulièrement aux paramètres qui deviennent l'objet central de l'étude. Il s'agit de fournir aux experts métier de l'AERM des clés d'analyse métier pour permettre de mieux déchiffrer leur comportement dans le milieu, à travers :

- La détection et l'analyse de phénomènes cycliques par paramètres
- L'identification et l'analyse d'impact des phases hydrologiques sur le comportement individuel des paramètres dans le milieu.

  Par extension, ce second objectif vise à identifier l'élasticité des concentrations de paramètres face à d'importants épisodes pluvieux. En effet, le calcul des phases hydrologiques sur la base des données débits peut être considéré comme un substitut des données pluviométriques qui ne sont pas disponibles en base.

Les sections suivantes détaillerons la méthodologie sélectionnée au regard de l'objectif ciblé, puis le résultat final obtenu et présenté pour répondre au sujet et enfin une ouverture sur les modalités de réutilisation des procédés statistiques proposés.

2.2 Méthodologie appliquée

La première partie de ce cas demande de fournir une méthodologie capable d'analyser les courbes de concentrations et de charges des différents paramètres afin d'en extraire des cycles puis de les caractériser dans le temps et la durée.

Le second volet s'attache quant à lui à l'analyse d'un possible impact des phases hydrologiques sur les valeurs mesurées de paramètres dans les cours d'eau. Les phases hydrologiques définies par l'AERM se basent sur les débits relevés sur les cours d'eau.

2.2.1 Détection et caractérisation des phénomènes cycliques

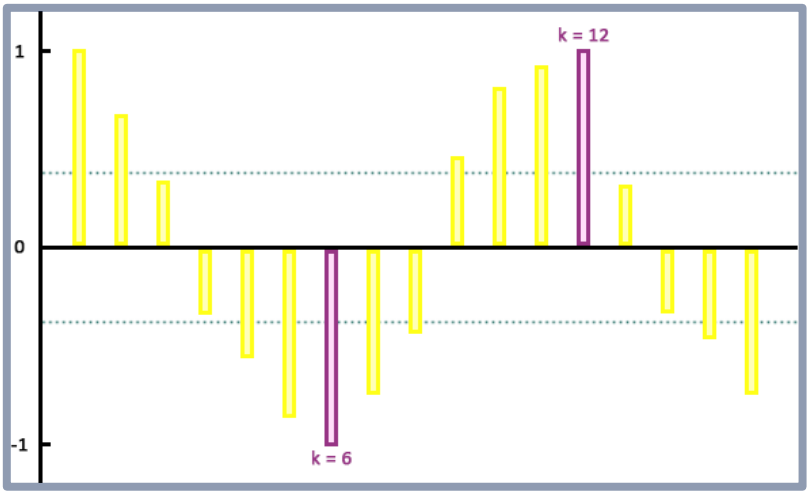
Afin d'extraire le cycle des différents paramètres nous avons sélectionné l'algorithme d'autocorrélation totale, ainsi que sa variante partielle pour le cas des paramètres peu mesurés.

En effet afin d'étudier en masse tous les cycles possibles, de tous les paramètres, sur tous les points de surveillance, cet algorithme d'autocorrélation est apparu comme légitime car automatisable, rapide d'exécution et proposant un résultat facilement interprétable.

L'autocorrélation d'ordre k mesure la corrélation d'une série avec elle-même, mais retardée de k périodes. Si cette dernière est significativement différente de 0, cela signifie qu'il existe une relation linéaire entre les observations en t et t-k, ce qui traduit l'existence d'un cycle de k périodes.

Nous avons donc calculé pour chaque paramètre la corrélation qui lie une observation en t à son observation retardée de k périodes pour k=1,….,24. Il n'est en effet pas forcément cohérent de chercher des cycles dont la période est supérieure à 2 ans : cette dernière peut notamment concerner des pesticides dont l'utilisation par les exploitants agricoles est très sporadique.

En sortie d'algorithme un autocorrélogramme peut être tracé afin de fournir un aperçu visuel des cycles potentiels d'une série. En effet il restitue graphiquement toutes les corrélations calculées entre t et t-k, ainsi que le seuil de significativité associé, qui représente le seuil pour lequel le coefficient de corrélation calculé est significativement différent de 0.



Autocorrélogramme d'une série périodique de 1 an sinusoïdale

Dans le cadre de séries composées de données de moindre qualité, l'autocorrélation totale délivre peu de résultats robustes, sa variante l'autocorrélation partielle nous a permis d'extraire des cycles cohérents en termes métiers et significatifs statistiquement.

L'autocorrélation partielle a l'avantage de négliger l'influence des valeurs entre t et t-k, si bien qu'on retient uniquement la liaison linéaire entre ces deux valeurs, qui peut dès lors s'interpréter comme le coefficient de la régression linéaire (par MCO) de de X_t sur X_{t-k}.

Un même paramètre étant testé sur plusieurs sites de surveillance distincts, il est important de pouvoir agréger les résultats pour en déterminer le cycle le plus récurrent et donc traduisant le comportement caractéristique de ce paramètre.

Pour répondre totalement à l'objectif, il est nécessaire, après avoir détecté les cycles, de pouvoir déterminer leur période de fortes valeurs. En effet en retirant la phase de pic, haut ou bas, l'AERM détient un nouvel élément pour comprendre et analyser le comportement des paramètres.

En regroupant les prélèvements par mois et en calculant leur moyenne mensuelle nous pouvons déterminer le mois avec le pic le plus haut et le mois du pic le plus bas, c'est la méthode du monthplot.

Tous comme pour la détection des cycles, un même paramètre est testé plusieurs fois et peut faire apparaître des périodes de pics différentes, même si elles sont proches, l'agrégation de tous les résultats par paramètre permet de délivrer une probabilité d'existence réelle d'un cycle et de ses périodes de fortes et/ou faibles valeurs.

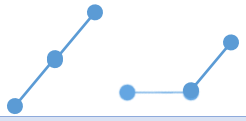
2.2.2 Analyse d'impact des phases hydrologiques sur la détection des polluants

Comprendre l'impact hydrologique sur les valeurs mesurées des paramètres polluants, c'est aider l'AERM à comprendre de quelles manières le ruissellement des pluies intervient dans la pollution des cours d'eau.

Afin d'analyser cet impact, l'Agence de l'Eau Rhin-Meuse nous a fourni une série de règles permettant de caractériser les phases hydrologiques d'un cours d'eau. La notion d'épisode pluvieux est par extension masquée dans la plupart des phases de crues qui, elles, vont influencer les valeurs de concentrations mesurées par paramètre.

En période de forte pluie les paramètres seront à priori plus dilués dans l'eau et, à l'inverse en période de sécheresse, les paramètres seront très concentrés dans un faible niveau d'eau.

Ainsi chaque phase hydrologique sera calculée en considérant les débits observés sur 3 jours consécutifs (j-1, j et j+1). De cette façon la mesure du débit prélevée au moment j sera ainsi typée et recodée selon les cas suivants :

Numéro de phase	Description schématique	Définition
1		**Stabilité hydrologique**, pas de modification nette des débits
2		**Montée de crue**, le débit augmente après le prélèvement et augmente ou est stable avant le prélèvement
3		**Pic de crue**, le débit augmente avant le prélèvement et redescend ou reste stable après ou le débit est stable avant le prélèvement et diminue après.
4		**Descente de crue**, le débit diminue avant ou avant et après le prélèvement
5		**Entre deux crues**, le débit diminue puis augmente après le prélèvement

Note : les phases de stabilité et d'entre deux crues ont été regroupées au cours de la phase de développement car trop identiques

Fort de cette nouvelle variable qualitative, il nous sera possible, une fois rapprochée aux valeurs de concentration de quantifier l'impact d'une phase sur le comportement de chaque paramètre. En effet chaque mesure de concentration est positionnée sur une date sur laquelle nous connaissons maintenant sa phase hydrologique

Pour calculer cet impact nous avons sélectionné la méthode ANOVA. Ce modèle statistique permet dans ce cas de tester si les échantillons sont significativement identiques, qu'ils proviennent de la même population.

En d'autres termes le modèle vérifie pour chaque phase hydrologique si les valeurs de concentrations sont significativement identiques. Et, si le test de l'ANOVA est rejeté alors on peut en déduire que la phase hydrologique à l'origine du rejet a eu un impact significatif sur les valeurs de concentrations du paramètre étudié.

2.3 Synthèse des résultats du cas

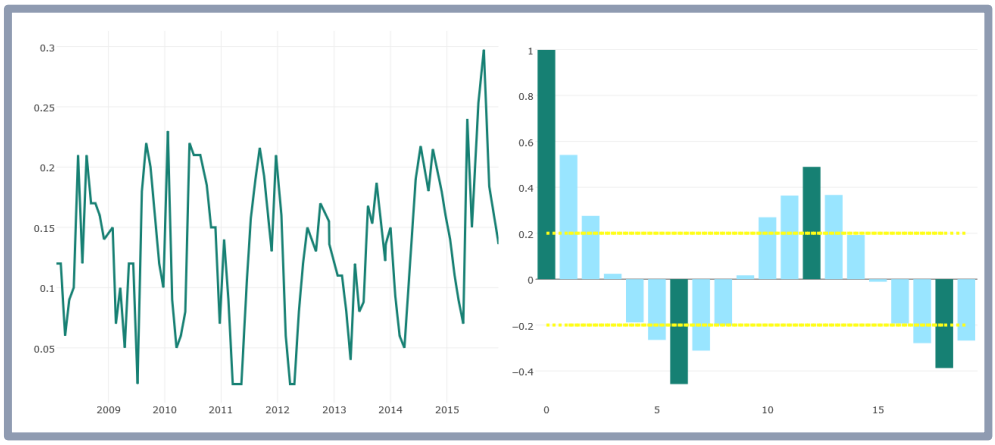
Afin d'illustrer la méthodologie avancée à la section précédente, cette synthèse des résultats développe et explique pour chaque objectif les réponses que nous avons apportées.

Pour aller plus loin, l'entièreté des étapes de calculs et restitutions produites sont trouvables dans le document livré pour la fin de la phase de développement : *BPM_AERM_Livrable_Final_Phase2_20170609*

2.3.1 Analyse et spécification du cycle détecté par paramètre

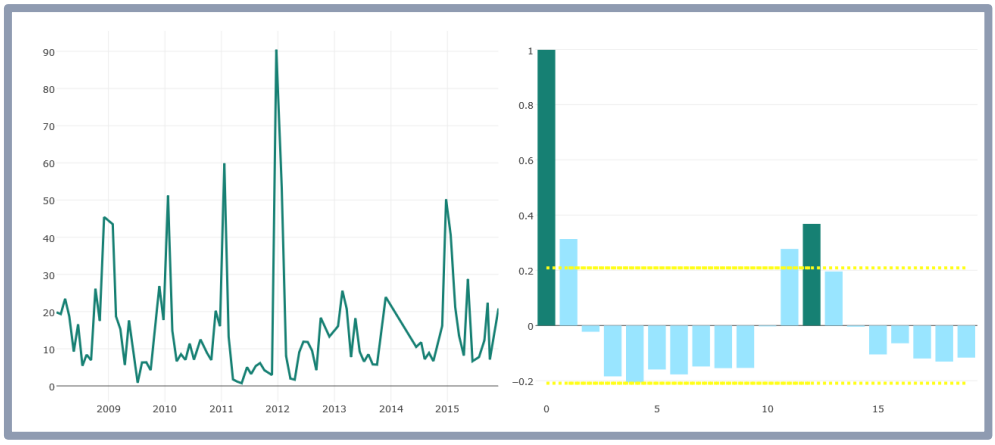
Comme spécifié plus haut, l'algorithme d'autocorrélation totale, et partielle, nous a permis de détecter et de spécifier les cycles des paramètres. En synthèse la majorité des paramètres étudiés ont pu être rapprochés à un cycle annuel. En effet les paramètres écartés représentent essentiellement des paramètres pour lesquels trop de valeurs seuils sont enregistrées dans les données que même l'autocorrélation partielle n'apporte aucun résultat.

En sortie d'algorithme nous avons préparé, pour chaque paramètre, une comparaison entre la série réelle et les résultats de l'algorithme d'autocorrélation sous forme de lagplot, où chaque barre indique le niveau du coefficient de corrélation entre k et les termes retardés. Cette analyse a été développée sur les valeurs de concentration et de charges pour comparaison des résultats et validation des tendances.



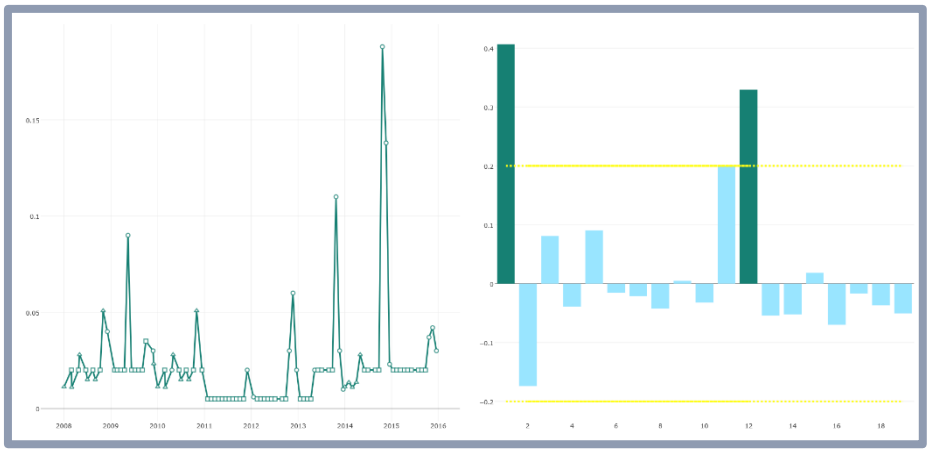
Chronogramme et autocorrélogramme – Orthophosphates (1433) – Ham-Sur-Meuse (02123000)

***CONCENTRATION - Durée du cycle :** 1 an sinusoïdal sur 6 mois*



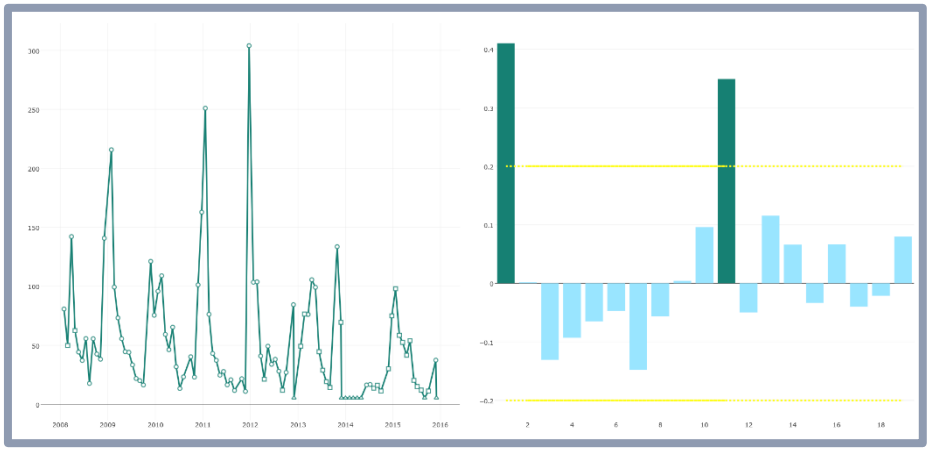
Chronogramme et autocorrélogramme – Orthophosphates (1433) – Ham-Sur-Meuse (02123000)

***CHARGE - Durée du cycle :** 1 an*



Chronogramme et autocorrélogramme – Isoproturon (1208) – Meuse à Givet (02124000)

***CONCENTRATION - Durée du cycle :** 1 an*



Chronogramme et autocorrélogramme – Azote (1319) – Meuse à Donchéry (02117000)

***CHARGE - Durée du cycle :** 1 an*

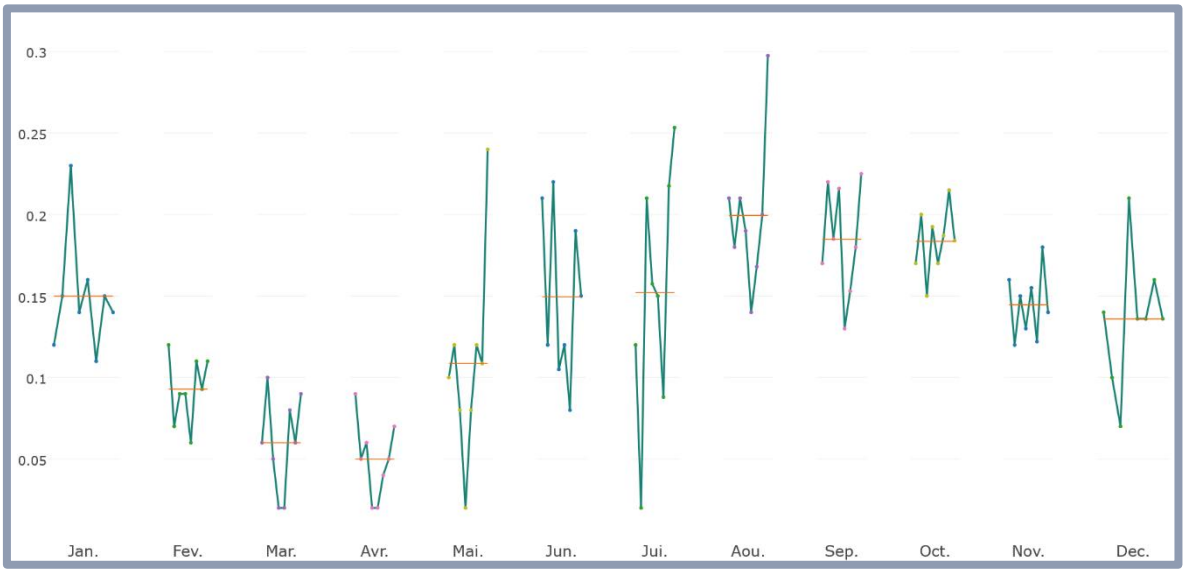
Pour la caractérisation des cycles et de leur période il suffit donc de ne conserver que les coefficients supérieurs au seuil de décision. Ainsi en fonction de la valeur k du terme retardé le plus significatif nous obtenons la durée du cycle.

Note : la plupart des cycles significatifs sur 6 mois le sont également sur 12 mois, c'est donc au métier de trancher paramètre par paramètre sur la durée réelle du cycle.

2.3.2 Spécification du cycle détecté

Connaissant le cycle des paramètres le second objectif de ce cas était de déterminer leurs périodes de pics. Par période nous entendons à partir de quels mois de l'année le cycle atteint son niveau le plus fort et le niveau le faible.

La méthode du monthplot est visualisable sur la restitution suivante, sur laquelle nous pouvons, grâce aux droites rouges, repérer les moyennes mensuelles les plus hautes et les plus faibles. Ainsi malgré les fluctuations d'une année sur l'autre il est possible d'agréger pour chaque mois sa valeur de concentration moyenne.



Moyennes mensuelles pour la détection des périodes de fortes valeurs pour l'orthophosphate (1433)

Une fois extraites, les valeurs les plus fortes sont comptabilisées sur l'ensemble des stations pour un paramètre donné.

La récurrence du cycle donnera la probabilité de validation de ce dernier sur son paramètre

La récurrence des pics de valeurs hautes et faibles indiquera la probabilité de validation des phases hautes et basses du paramètre ciblé.

Toutes ces informations sont résumées au sein d'un unique tableau d'agrégation pour offrir aux agents de l'AERM une meilleure visibilité par paramètre.

CODE_PARAMETRE	NOM_PARAMETRE	NOMBRE DE PS ANALYSES					CYCLE PROBABLE				
		CYCLE_DETECTE	(%)	CYCLE_NON_DETECTE	(%)	TOTAL_PS	PERIODE	NB_PS	(% CYCLE_DETECTE)	PIC_HAUT	PIC_BAS
1302	Potentiel en Hydrogène (pH)	209	0,80	53	0,20	262	12	66	0,32	Mar	Aug
1305	Matières en suspension	147	0,64	84	0,36	231	12	50	0,34	Dec	Jul
1309	Couleur mesurée	115	0,68	53	0,32	168	12	28	0,24	Dec	Feb
1311	Oxygène dissous	262	1,00	1	0,00	263	12	199	0,76	Feb	Aug
1313	Demande Biochimique en oxygène en 5 jours (D.B.O.5)	145	0,87	22	0,13	167	12	44	0,30	Mar	Dec
1314	Demande Chimique en Oxygène (DCO)	40	0,62	25	0,38	65	12	9	0,23	Dec	Apr
1319	Azote Kjeldahl	86	0,91	8	0,09	94	2	19	0,22	Jan	Dec
1327	Hydrogénocarbonates	21	0,95	1	0,05	22	12	9	0,43	Feb	Aug
1335	Ammonium	93	0,68	43	0,32	136	12	25	0,27	Jan	Aug
1337	Chlorures	19	0,83	4	0,17	23	12	9	0,47	Dec	Aug
1338	Sulfates	17	0,77	5	0,23	22	12	10	0,59	Oct	Jan
1339	Nitrites	190	0,84	37	0,16	227	12	81	0,43	Jun	Sep
1340	Nitrates	240	0,94	16	0,06	256	12	147	0,61	Feb	Aug
1342	Silicates	180	0,96	8	0,04	188	12	141	0,78	Dec	Apr
1347	Titre alcalimétrique complet (T.A.C.)	20	0,91	2	0,09	22	12	9	0,45	Feb	Aug
1350	Phosphore total	179	0,72	68	0,28	247	12	76	0,42	Jul	Mar
1362	Bore	14	0,93	1	0,07	15	12	5	0,36	Jun	Jan
1367	Potassium	20	0,91	2	0,09	22	12	9	0,45	Jul	Jan
1369	Arsenic	12	0,92	1	0,08	13	12	8	0,67	Jul	Mar
1369	Arsenic	12	0,92	1	0,08	13	12	8	0,67	Aug	Mar
1372	Magnésium	27	0,73	10	0,27	37	12	12	0,44	Feb	Jan
1372	Magnésium	27	0,73	10	0,27	37	12	12	0,44	Aug	Jan
1372	Magnésium	27	0,73	10	0,27	37	12	12	0,44	Jul	Jan
1372	Magnésium	27	0,73	10	0,27	37	12	12	0,44	Feb	Jan
1373	Titane	4	0,57	3	0,43	7	12	2	0,50	Dec	Aug
1373	Titane	4	0,57	3	0,43	7	12	2	0,50	Dec	Sep
1374	Calcium	27	0,73	10	0,27	37	12	9	0,33	Feb	Aug
1375	Sodium	20	0,91	2	0,09	22	12	9	0,45	Dec	Jul
1375	Sodium	20	0,91	2	0,09	22	12	9	0,45	Dec	Aug
1377	Béryllium	1	0,33	2	0,67	3	12	1	1,00	Dec	Jul
1379	Cobalt	1	1,00	0	0,00	1	4	1	1,00	Apr	Jul
1383	Zinc	14	0,74	5	0,26	19	2	8	0,57	Jan	Dec
1383	Zinc	14	0,74	5	0,26	19	2	8	0,57	Feb	Dec
1384	Vanadium	7	0,70	3	0,30	10	12	3	0,43	Sep	Mar
1386	Nickel	7	1,00	0	0,00	7	5	2	0,29	Apr	Mar
1392	Cuivre	13	0,72	5	0,28	18	2	7	0,54	Jan	Aug
1392	Cuivre	13	0,72	5	0,28	18	2	7	0,54	Jan	Sep
1392	Cuivre	13	0,72	5	0,28	18	2	7	0,54	Apr	Sep
1392	Cuivre	13	0,72	5	0,28	18	2	7	0,54	Apr	Aug
1395	Molybdène	2	0,67	1	0,33	3	13	1	0,50	Aug	Jan
1395	Molybdène	2	0,67	1	0,33	3	13	1	0,50	Sep	Jan
1395	Molybdène	2	0,67	1	0,33	3	14	1	0,50	Aug	Feb
1395	Molybdène	2	0,67	1	0,33	3	14	1	0,50	Sep	Feb

Extrait du résultat de l'agrégation des cycles par paramètres, en concentration

2.3.3 Impact des phases hydrologique sur les concentrations

L'idée pour cet objectif est de rapprocher les valeurs de concentration mesurées pour un paramètre aux phases hydrologiques détectées sur les débits sur chaque station. L'agrégation des informations tous points de surveillance confondus permet de construire pour chaque paramètre un jeu de données sur lequel les hypothèses de l'ANOVA seront testées.

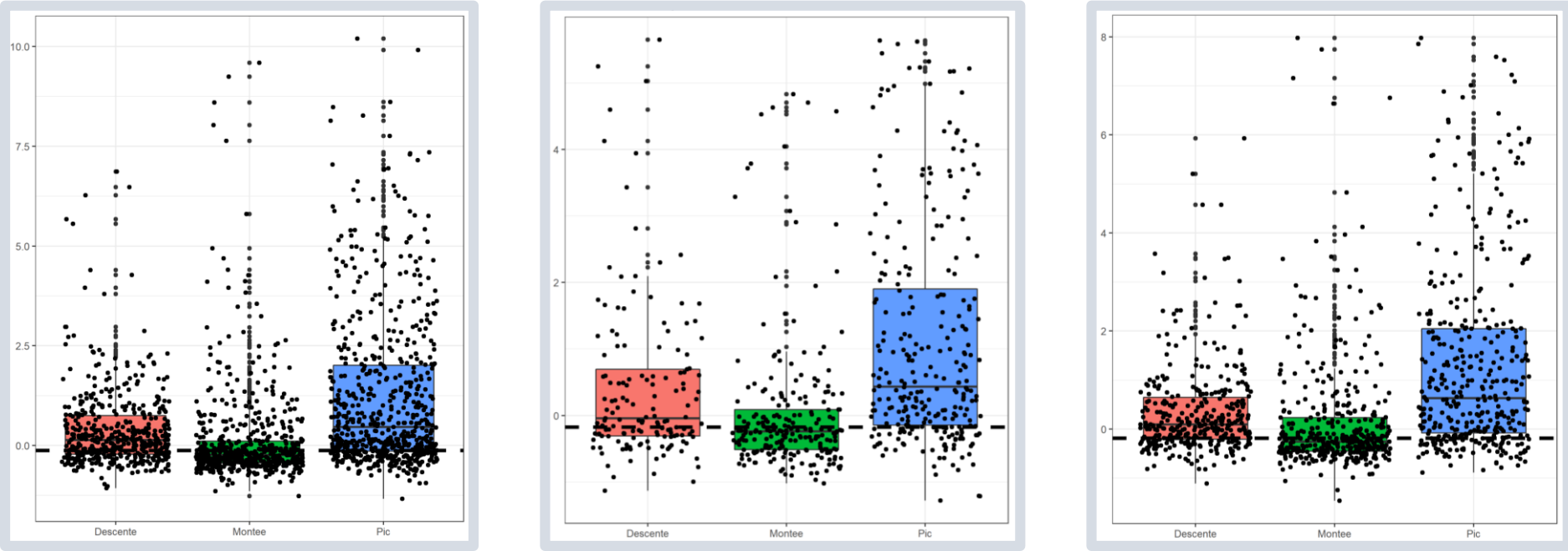
Paramètre 1308 - Atrazine		
NUMNAT	PHASE	CONCENTRATION
2000990	Entre deux	-1,79
2000990	Stabilite	-0,92
2000990	Stabilite	1,68
2000990	Montee	1,55
2000990	Stabilite	0,31
2000990	Stabilite	-0,3
2000990	Descente	-0,3
2000990	Montee	-0,3
2000990	Stabilite	-0,12
2000990	Descente	0,5
2057250	Stabilite	-0,64
2057250	Stabilite	-0,1
2057250	Stabilite	0,99
2057250	Stabilite	-1,73
2057250	Stabilite	1,54
2106410	Stabilite	-0,37
2106410	Stabilite	0,37
2106410	Stabilite	-1,11
2106410	Pic	-0,37
2106410	Montee	2,58
2117650	Stabilite	-0,58
2117650	Stabilite	-0,91
2117650	Stabilite	0,07
2117650	Descente	-0,26
2117650	Stabilite	0,72

Structure d'un jeu de données d'alimentation de l'Anova

Pour un paramètre

Sous forme de tableau il est en effet impossible de déterminer si une phase en particulier impacte les valeurs de concentration relevées. Avant toutes analyses ANOVA il est toujours intéressant de visualiser les données sous la forme de boxplots, ou boîtes à moustache.

Ces graphiques donnent en effet un aperçu rapide des distributions par modalité et permet de valider visuellement si une des modalités notamment, ici une phase hydrologique en particulier, est significativement différente des autres.



Boxplot du paramètre Matières en Suspension (1305) – Total / Eté / Hiver

A la lecture des boxplots, produits pour un même paramètre mais sur des périodes de temps différents, on observe clairement des dispersions assez marquées en fonction des phases hydrologiques par rapport à la moyenne globale (ligne pointillée).

L'ANOVA va justement chercher à valider les écarts à la moyenne des différentes modalités représentée en couleur ci-dessus.

Encore une fois, ici, si le test d'homogénéité des moyennes des modalités à la moyenne globale est rejeté, alors cela indique qu'au moins une phase impacte la mesure des concentrations du paramètre étudié.

Afin de résumer au mieux les résultats de l'ANOVA obtenus sur l'ensemble des paramètres nous avons produit un tableau récapitulant l'ensemble des tests statistiques validant la robustesse de l'analyse, mais aussi des indicateurs visuels permettant de donner le sens et l'intensité de l'impact d'une phase hydrologique sur les concentrations.

CODE_PARAMETRE	NOM_PARAMETRE	TESTS PROBABILISTES		ANALYSE IMPACT GLOBAL		IMPACT IMPACT INDIVIDUEL			EFFECTIFS DES PHASES				
		NORMALITE (Shapiro-Wilk)	HOMOGENEITE (Levene)	KRUSKAL-WALLIS	ANOVA	Descente	Montee	Pic	Global	Descente	Montée	Pic	Stabilité Entre Deux
1108	Atrazine déséthyl	Non Normal	Homogènes	Identique	Identique				266	12	30	7	217
1113	Bentazone	Non Normal	Non Homogènes	Non Identique	Non Identique	0,6733		0,8208	226	18	19	15	174
1115	Benzo(a)pyrène	Non Normal	Non Homogènes	Non Identique	Non Identique	0,4937	0,59	0,9933	1277	102	120	173	882
1116	Benzo(b)fluoranthène	Non Normal	Non Homogènes	Non Identique	Non Identique	0,371	0,578	1,0302	1314	108	121	171	914
1118	Benzo(g,h,i)pérylène	Non Normal	Non Homogènes	Non Identique	Non Identique	0,4601	0,6174	1,0165	1278	101	107	160	910
1191	Fluoranthène	Non Normal	Non Homogènes	Non Identique	Non Identique		0,2886	0,5575	1023	98	118	174	633
1204	Indéno(1,2,3-cd)pyrène	Non Normal	Non Homogènes	Non Identique	Non Identique	0,3049	0,5858	1,0002	1302	98	116	166	922
1302	Potentiel en Hydrogène (pH)	Non Normal	Non Homogènes	Non Identique	Non Identique	-0,3676	-0,2474	-0,4408	3222	398	489	516	1819
1305	Matières en suspension	Non Normal	Non Homogènes	Non Identique	Non Identique	0,5217	0,2372	1,3519	5380	471	599	651	3659
1309	Couleur mesurée	Non Normal	Non Homogènes	Non Identique	Non Identique	0,5524	-0,1729	0,6828	3439	428	499	568	1944
1311	Oxygène dissous	Non Normal	Non Homogènes	Non Identique	Non Identique		-0,2041	-0,2574	6253	513	679	666	4395
1313	Demande Biochimique en oxygène en 5 jours (D.B.O.5)	Non Normal	Non Homogènes	Non Identique	Non Identique			0,5378	3765	395	481	566	2323
1314	Demande Chimique en Oxygène (DCO)	Non Normal	Non Homogènes	Non Identique	Non Identique	0,3167		0,9562	2874	354	434	522	1564
1319	Azote Kjeldahl	Non Normal	Non Homogènes	Non Identique	Non Identique	0,1404		0,6899	3153	366	448	527	1812
1327	Hydrogénocarbonates	Non Normal	Non Homogènes	Non Identique	Non Identique	-0,5709		-0,5543	3574	243	351	343	2637
1328	Carbonates	Non Normal	Homogènes	Identique	Identique				283	13	21	11	238
1335	Ammonium	Non Normal	Non Homogènes	Non Identique	Non Identique	-0,4275		0,1368	3326	351	443	514	2018
1337	Chlorures	Non Normal	Non Homogènes	Non Identique	Non Identique	-0,7568		-0,3524	3349	244	348	324	2433
1338	Sulfates	Non Normal	Homogènes	Non Identique	Non Identique	-0,6847		-0,5522	3478	242	344	339	2553
1339	Nitrites	Non Normal	Non Homogènes	Non Identique	Non Identique	-0,2624	-0,2528		2853	363	468	485	1537
1340	Nitrates	Non Normal	Non Homogènes	Non Identique	Non Identique	0,2872			5814	495	638	646	4035
1342	Silicates	Non Normal	Non Homogènes	Non Identique	Non Identique	0,4689	0,1024		4786	420	568	576	3222
1347	Titre alcalimétrique complet (T.A.C.)	Non Normal	Non Homogènes	Non Identique	Non Identique	-0,6306	-0,1409	-0,6375	3216	228	322	313	2353
1350	Phosphore total	Non Normal	Non Homogènes	Non Identique	Non Identique		0,1321	0,7819	4424	445	593	609	2777
1367	Potassium	Non Normal	Non Homogènes	Non Identique	Non Identique	-0,4212			2549	227	310	313	1699
1369	Arsenic	Non Normal	Homogènes	Identique	Identique				1139	97	121	135	786
1372	Magnésium	Non Normal	Non Homogènes	Non Identique	Non Identique	-0,7168		-0,5622	3409	275	371	366	2397
1374	Calcium	Non Normal	Non Homogènes	Non Identique	Non Identique	-0,6219		-0,5654	3271	270	371	364	2266
1375	Sodium	Non Normal	Non Homogènes	Non Identique	Non Identique	-0,6825		-0,3447	3132	234	333	334	2231
1383	Zinc	Non Normal	Non Homogènes	Non Identique	Non Identique		0,2332	0,5052	1229	89	143	161	836
1386	Nickel	Non Normal	Homogènes	Identique	Identique				980	119	114	142	605
1389	Chrome	Non Normal	Homogènes	Identique	Identique				475	37	44	67	327
1392	Cuivre	Non Normal	Non Homogènes	Non Identique	Non Identique	0,2692	0,2925	0,7804	1128	96	125	147	760
1429	Cote à l'échelle	Non Normal	Non Homogènes	Non Identique	Non Identique	0,4655		0,7483	3110	292	391	386	2041
1433	Orthophosphates (PO4)	Non Normal	Non Homogènes	Non Identique	Non Identique	-0,2278	0,1055	0,2235	4626	457	582	600	2987
1439	Chlorophylle a	Non Normal	Non Homogènes	Non Identique	Non Identique	-0,3309	-0,2654		4140	379	529	546	2686
1312	Taux de saturation en oxygène	Non Normal	Non Homogènes	Non Identique	Non Identique		-0,2522	-0,2923	4933	460	621	640	3212
1345	Dureté totale	Non Normal	Non Homogènes	Non Identique	Non Identique	-0,7082		-0,6189	3971	284	385	381	2921
1436	Phéopigments	Non Normal	Non Homogènes	Non Identique	Non Identique	-0,1649		0,6147	5151	434	588	606	3523
1537	Pyrène	Non Normal	Non Homogènes	Non Identique	Non Identique			0,5848	992	97	113	166	616
1551	Azote global (N.GL.)	Non Normal	Non Homogènes	Non Identique	Non Identique	0,4245	0,1213	0,5169	4676	481	603	626	2966
1841	Carbone Organique	Non Normal	Non Homogènes	Non Identique	Non Identique	0,7823		1,0082	6466	782	929	1041	3714
5535	Somme de Benzo(b)fluoranthène et Benzo(k)fluoranthène	Non Normal	Non Homogènes	Non Identique	Non Identique	0,3668	0,5023	0,9842	1311	107	113	160	931
5536	Somme de Benzo(g,h,i)pérylène et Indéno(1,2,3-cd)pyrène	Non Normal	Non Homogènes	Non Identique	Non Identique	0,4409	0,509	0,9851	1385	99	112	165	1009
5932	Azote organique	Non Normal	Non Homogènes	Non Identique	Non Identique	0,3392	0,1078	0,6997	5179	410	529	606	3634
6276	Somme des pesticides totaux	Non Normal	Non Homogènes	Non Identique	Non Identique	0,1715		0,3064	3793	226	334	328	2905
1370	Aluminium	Non Normal	Non Homogènes	Non Identique	Non Identique	0,9534	0,1466	1,2532	2806	200	332	313	1961
1141	2,4-D	Non Normal	Homogènes	Identique	Identique				1663	102	188	172	1201
1361	Uranium	Non Normal	Non Homogènes	Non Identique	Non Identique			-0,5935	498	49	70	74	305
1362	Bore	Non Normal	Homogènes	Non Identique	Non Identique	-0,3918		-0,2197	1330	69	120	121	1020
1368	Argent	Non Normal	Homogènes	Identique	Identique				145	5	7	19	114
1373	Titane	Non Normal	Homogènes	Non Identique	Non Identique	0,3973			618	59	71	100	388
1384	Vanadium	Non Normal	Homogènes	Non Identique	Non Identique		0,3233	0,4599	580	49	72	85	374
1393	Fer	Non Normal	Non Homogènes	Non Identique	Non Identique	0,7034		1,2966	1912	127	247	219	1319
1394	Manganèse	Non Normal	Non Homogènes	Non Identique	Non Identique	0,1949	0,1773	1,0306	1721	124	239	213	1145
1396	Baryum	Non Normal	Homogènes	Non Identique	Non Identique	-0,6119		-0,2434	1000	59	116	119	706
1957	NONYLPHENOLS	Non Normal	Non Homogènes	Non Identique	Non Identique		-0,4249		562	29	50	51	432
6616	Di(2-ethylhexyl)phtalate	Non Normal	Homogènes	Identique	Identique				1559	96	166	156	1141
7073	Fluorure anion	Non Normal	Non Homogènes	Identique	Identique				990	96	150	157	587
1385	Sélénium	Non Normal	Homogènes	Identique	Identique				133	13	8	15	97
1517	Naphtalène	Non Normal	Homogènes	Identique	Identique				818	51	72	51	644
1032	Polychlorobiphényles totaux	Non Normal	Homogènes	Identique	Identique				93	4	14	5	70
1177	Diuron	Non Normal	Homogènes	Non Identique	Non Identique		0,3866		522	38	69	71	344

Extrait du fichier final des résultats relatifs à l'analyse d'impact des phases hydrologiques

2.4 Modalités de réutilisation

Au cours de ce cas d'étude nous avons démontré que l'analyse des séries temporelles nécessite d'une part un historique assez grand pour la détection de cycle ou de saisonnalité, mais également des données de bonnes qualités pour assurer de meilleurs résultats.

En effet contrairement à d'autres domaines statistiques, les séries chronologiques se basent sur une expertise passée, donc connue et donc valide et validée.

C'est pourquoi l'algorithme d'autocorrélation a été préféré aux méthodes plus classiques de décomposition des courbes et autres modèles d'analyses de tendances de séries temporelles.

La méthode d'Anova quant à elle peut être utilisable sur plusieurs domaines. En effet elle répond toujours à la problématique : Mes échantillons d'individus sont-ils issus de la même population de référence ?

Les cas d'école de cette méthode tentent par exemple de comparer le système de notation de deux professeurs sur les mêmes copies des élèves pour valider leur impartialité et leur sévérité.

Plus proche de la chimie, il est également courant de valider l'impact d'un sérum sur des niveaux de globules blancs de deux populations distinctes de patients, l'une ayant reçu bien évidemment un placebo. Si le test est rejeté alors le sérum est à priori efficace.

Une attention est portée aux résultats des tests statistiques préliminaires pour la mise en place de l'ANOVA. En effet malgré le rejet de la plupart de ces tests sur des cas réels d'utilisation, il n'est pas nécessaire d'abandonner l'analyse. En effet la robustesse de l'ANOVA en elle-même suffit pour valider les résultats attendus. Le métier est en effet toujours maître de la validation des résultats.

3 CAS 3

ORIGINE DES HYDROCARBURES AROMATIQUES POLYCYCLIQUES (HAP) METAUX ET PESTICIDES

3.1 Rappel des objectifs

Au cas précédent, nous avons mis en œuvre une analyse d'impact pour étudier la réponse comportementale des paramètres face à différentes phases hydrologiques : Montée de Crue, Pic de Crue et Descente de Crue. Plus particulièrement, nous avons mis en exergue la réactivité de leur concentration respective face à des épisodes pluvieux importants.

Pour ce cas, nous tâchons de déterminer l'impact non plus de phases hydrologiques, mais de conditions environnementales qui pourraient expliquer la détection plus ou moins importante de certains paramètres ciblés par l'Aerm, appartenant à la famille des HAP (25), métaux (19) ou pesticides (21).

Par-là, notre étude va tenter de résoudre la problématique métier suivante : au sein d'un bassin versant, la détection dans les eaux de substances nocives peut-elle être expliquée par les caractéristiques environnementales alentours ? Si oui, avec quelle intensité ?

L'analyse pour le Cas 3 va donc consister à dépister, identifier, puis quantifier d'éventuelles corrélations significatives entre :

- Variables environnementales caractéristiques d'un bassin versant ;
- Détections dans les eaux de paramètres appartenant aux familles suivantes : HAP (25), Métaux ou Micropolluants Minéraux (19) et Pesticides (21)

Les sections suivantes détaillerons la méthodologie sélectionnée au regard de l'objectif ciblé, puis le résultat final obtenu et présenté pour répondre au sujet et enfin une ouverture sur les modalités de réutilisation des procédés statistiques proposés.

3.2 Méthodologie appliquée

Afin de déterminer l'impact des différentes variables environnementales, comme l'occupation ou la structure des sols des bassins versants et des corridors, sur les valeurs d'une série de paramètres polluants nous avions proposé la mise en place d'une analyse factorielle (ACP).

Cependant la complexité de la structure du tableau de données en entrée, c'est-à-dire autant de colonnes que de modalités existantes dans les variables environnementales, a forcé le basculement de la méthode ACP vers une méthode de régression parcimonieuse, aussi appelée régression pas-à-pas.

NUMNAT	DONNEES QUALITE	DONNEES ENVIRONNEMENTALES											
	P90	Variables Mères						Variables Filles					
		M_1	M_2	M_3	M_4	M_5	...	F_1	F_2	F_3	F_4	F_5	...
2000990	0,145	0,908	0,316	0,208	0,996	0,349	...	0,670	0,178	0,668	0,747	0,410	...
2001000	0,873	0,517	0,936	0,767	0,210	0,104	...	0,284	0,144	0,339	0,308	0,486	...
2001010	0,307	0,754	0,532	0,353	0,109	0,586	...	0,923	0,626	0,669	0,610	0,600	...
2001025	0,441	0,606	0,409	0,693	0,514	0,794	...	0,088	0,155	0,671	0,672	0,149	...
2001030	0,781	0,889	0,822	0,072	0,492	0,934	...	0,839	0,311	0,727	0,054	0,738	...
2001500	0,343	0,517	0,661	0,793	0,382	0,422	...	0,974	0,858	0,988	0,417	0,505	...
2001725	0,038	0,439	0,554	0,119	0,301	0,571	...	0,159	0,437	0,753	0,844	0,501	...
2001750	0,427	0,134	0,705	0,810	0,253	0,676	...	0,813	0,212	0,295	0,377	0,286	...
2001915	0,072	0,061	0,999	0,588	0,711	0,423	...	0,355	0,499	0,758	0,687	0,966	...
2001955	0,946	0,672	0,368	0,937	0,045	0,338	...	0,811	0,440	0,874	0,711	0,390	...
2001990	0,553	0,982	0,059	0,597	0,145	0,114	...	0,283	0,255	0,617	0,782	0,157	...
2002000	0,080	0,446	0,050	0,038	0,972	0,996	...	0,053	0,484	0,396	0,536	0,013	...
2003100	0,920	0,378	0,246	0,664	0,562	0,718	...	0,607	0,817	0,338	0,742	0,114	...
2003350	0,774	0,811	0,508	0,413	0,847	0,345	...	0,181	0,686	0,231	0,478	0,898	...
2003620	0,809	0,958	0,044	0,100	0,271	0,290	...	0,216	0,011	0,551	0,533	0,347	...
2003670	0,015	0,884	0,951	0,126	0,076	0,837	...	0,385	0,469	0,182	0,901	0,046	...
2006500	0,361	0,662	0,730	0,104	0,872	0,143	...	0,396	0,858	0,518	0,190	0,356	...
2007250	0,279	0,182	0,436	0,079	0,595	0,901	...	0,041	0,302	0,722	0,639	0,888	...
2007380	0,467	0,071	0,808	0,778	0,938	0,304	...	0,252	0,906	0,559	0,091	0,458	...
2009000	0,233	0,186	0,568	0,344	0,046	0,910	...	0,977	0,612	0,907	0,827	0,626	...

Structure d'un jeu de données en entrée pour un paramètre (données non conformes)

Cette méthode de régression multiple permet par itération de recalculer le coefficient de détermination à chaque ajout de variable, ou suppression dans le cas de la méthode descendante, dans l'équation. Lorsque le modèle n'a pas amélioré son critère sur une itération il s'arrête.

Le critère de qualité sélectionné est donc le critère d'Akaiké (ou critère AIC). La régression pas-à-pas va donc produire toutes les combinaisons possibles de variables exogènes, et choisir la régression qui optimise (minimise si régression descendante) à chaque étape le critère AIC.

De cette façon nous avons construit des modèles de régression sur un nombre significatif et impactant de variables environnementales pour déterminer les valeurs P90 des paramètres polluants.

D'autre part, l'objectif se basant sur le seuil d'apparition flagrante et exceptionnelle de concentrations fortes des paramètres polluants, seul l'indicateur P90, soit les 90% des valeurs les plus fortes, n'a été conservé.

3.3 Synthèse des résultats du cas

Afin d'illustrer la méthodologie avancée à la section précédente, cette synthèse des résultats développe et explique pour chaque objectif les réponses que nous avons apportées.

Pour aller plus loin, l'entièreté des étapes de calculs et restitutions produites sont trouvables dans le document livré pour la fin de la phase de développement : *BPM_AERM_Livrable_Final_Phase2_20170609*

Grâce à la méthode de régression parcimonieuse nous obtenons donc par chaque paramètre une liste de variables environnementales représentées chacune par un coefficient de régression. De cette façon il devient simple de déterminer le sens et la teneur de l'impact d'une variable dans son équation de régression et donc dans la construction de la valeur P90 du paramètre.

En sortie d'algorithme un fichier agrégé des influences des variables par paramètres a été construit, il contient notamment :

- Les variables environnementales (en bassins versants et corridors) qui impactent significativement chacun des paramètres
- Le sens de l'impact, donné par le signe du coefficient :
    - Les coefficients mentionnés en rouge (négatifs) signifient que le paramètre et la variable environnementale en question évoluent dans un sens contraire
    - Les coefficients mentionnés en vert (positifs) signifient que le paramètre et la variable environnementale en question évoluent dans le même sens.
- L'intensité de l'impact, donnée par la valeur absolue du coefficient, et soulignée par la symbolique suivante :

SIGNE	SENS	INTENSITE
↑	Positif	> 0,5
↗	Positif	> 0 et < 0,5
↘	Négatif	> -0,5 et < 0
↓	Négatif	< -1 et > -0,5

Note : Tous les résultats sont significatifs au risque de 5%.

VARIABLES MERES		Benzo(a)pyrène	Fluoranthène	Métolachlore	Arsenic	Zinc	Mercure	Cuivre	Glyphosate	Naphtalène	Métazachlore	Diflufenicanil	Nicosulfuron	AMPA	Metolachlor OXA	Metolachlor ESA	Métazachlore OXA	Métazachlore ESA	meta	meta	amp	Indice pesticides
		1115	1191	1221	1369	1383	1387	1392	1506	1517	1670	1814	1882	1907	6853	6854	6894	6895	11221	11670	11907	12000
bv_clc	111 - Tissu urbain continu							0,227							0,616						0,511	
bv_clc	112 - Tissu urbain discontinu		0,504	0,537				-0,216	0,177	-0,419	0,399	0,208			-0,805						-0,798	
bv_clc	121 - Zones industrielles et commerciales					0,629					0,726					0,251	0,399					
bv_clc	122 - Réseaux routier et ferroviaire et espaces associés		-0,278			-0,203			0,310		-0,587											0,230
bv_clc	123 - Zones portuaires			-0,189																		
bv_clc	124 - Aéroports		0,537								-0,710		0,238			0,300						
bv_clc	131 - Extraction de matériaux		-0,389			0,319					-0,331					-0,311	0,581					
bv_clc	132 - Décharges		-0,233	0,391		0,087					0,328											
bv_clc	133 - Chantiers																		0,436			
bv_clc	141 - Espaces verts urbains			-0,129		-0,242					-0,162											
bv_clc	142 - Equipements sportifs et de loisirs			-0,256											0,277							
bv_clc	211 - Terres arables hors périmètres d'irrigation				0,666									0,545		-0,214	0,293		-0,245			0,303
bv_clc	221 - Vignobles								0,262													
bv_clc	222 - Vergers et petits fruits		-0,174								-0,153				0,586			0,512	-0,173		0,593	
bv_clc	231 - Prairies		0,223				-0,215		0,172										0,219			
bv_clc	242 - Systèmes culturaux et parcellaires complexes		-0,328	-0,150	-0,312				-0,200		0,297	0,402				0,200	-0,287					
bv_clc	243 - Territoires principalement occupés par l'agriculture,avec présence de végétation naturelle importante				-0,185											0,214	0,678	0,545	0,133			
bv_clc	311 - Forêts de feuillus				-0,152				0,345		-0,154											0,264
bv_clc	312 - Forêts de conifères							-0,193							-0,309						-0,294	
bv_clc	313 - Forêts mélangées																					
bv_clc	321 - Pelouses et pâturages naturels																					
bv_clc	322 - Landes et broussailles					-0,263																
bv_clc	324 - Forêt et végétation arbustive en mutation														-0,240						-0,242	
bv_clc	332 - Roches nues																					
bv_clc	411 - Marais intérieurs					-0,102																
bv_clc	412 - Tourbières																					
bv_clc	511 - Cours et voies d'eau		0,262			0,075													-0,269			
bv_clc	512 - Plans d'eau																					
bv_pente	t_faible																					
bv_pente	faible		0,177							-0,237		-0,197	0,306									
bv_pente	moyenne											-0,145	0,243									
bv_pente	forte		-0,145																			
bv_pente	t_forte														-0,189						-0,191	
bv_pente	tt_forte														0,201						0,218	
bv_rhf	aletrique_poreuse								0,256						-0,300						-0,308	
bv_rhf	mediocre																					
bv_rhf	carb_karst																					
bv_rhf	inconnu																					
bv_rhf	metamorphique																					
bv_rhf	carb_fissuree																					
bv_rhf	tres_peu_permeable																					
bv_pcpracc	Oui																					
bv_pcpracc	Non																					

VARIABLES FILLES		Benzo(a)pyrène	Fluoranthène	Métolachlore	Arsenic	Zinc	Mercure	Cuivre	Glyphosate	Naphtalène	Métazachlore	Diflufenicanil	Nicosulfuron	AMPA	Metolachlor OXA	Metolachlor ESA	Métazachlore OXA	Métazachlore ESA	meta	meta	amp	Indice pesticides
		1115	1191	1221	1369	1383	1387	1392	1506	1517	1670	1814	1882	1907	6853	6854	6894	6895	11221	11670	11907	12000
bv_cultagri	prairie								0,180		0,220						0,259	0,258				
bv_cultagri	maïs			0,540							-0,130		0,460		0,359	0,620			0,550			
bv_cultagri	blé				-0,258			-0,170	0,380		0,404	0,308	0,269				0,378	0,403		0,455	0,076	0,330
bv_cultagri	orge			-0,230	-0,233			-0,240	0,270		0,405	0,262		-0,240		-0,370	0,443	0,480	-0,256	0,470		0,220
bv_cultagri	colza			-0,240	-0,220			-0,190	0,320		0,519	0,345		-0,260		-0,400	0,505	0,538	-0,278	0,502		0,230
bv_cultagri	autres oléagineux																					
bv_cultagri	divers										0,266						0,343	0,228		0,190		-0,180
bv_cultagri	autres céréales																					
bv_cultagri	Légumes Fleurs																					
bv_cultagri	Verger																					
bv_cultagri	Autres cultures industrielles																					
bv_cultagri	Vignes																					
bv_cultagri	Protéagineux																					
bv_cultagri	Fruits à Coques																					
bv_cultagri	Legumineuses																					
bv_cultagri	Tournesol																					
bv_cultagri	Semences																					
bv_cultagri	Estives Landes																					
bv_cultagri	Inconnu																					
bv_cultagri	Canne à Sucre																					
bv_cultagri	Plantes à Fibres																					
bv_sgh	Densité			0,190							0,183		0,289									
bv_basol	O																					
bv_basol	F																					
bv_basol	K																					
bv_basol	T																					
bv_basol	C																					
bv_basol	D																					
bv_basol	H																					
bv_basol	J																					
bv_basol	Estives Landes																					
bv_basol	L																					
bv_basol	G																					
bv_basol	I																					
bv_routes	nb_chausse_1	0,350	0,340			0,256		0,230	0,250	0,260	-0,180			0,230						-0,196	0,250	
bv_routes	nb_chausse_2																					
bv_idgr	ratio	0,250	0,250		-0,190					0,180												

Extrait de la table de résultat du Cas 3 – Variables environnementales qui impactent significativement le P90 des paramètres du Jeu 1

3.4 Modalités de réutilisation

En pratique la construction d'une équation de régression, et ce quelle que soit la méthode utilisée (multiple, logistique, polynomiale, parcimonieuse, etc…) permet de répondre à plusieurs objectifs dont on peut citer les principaux :

- Evaluer l'impact des variables explicatives sur les valeurs de la variable à expliquer,
- Prédire le niveau de la variable à expliquer en fonction de nouvelles valeurs estimées des variables explicatives,
- Sélectionner les variables les plus déterminantes dans la construction de l'équation,
- Détecter les observations qui ne suivent pas le modèle.

Pour répondre à ce cas d'étude nous avons vu comment construire pour chaque paramètre, une équation composée des variables les plus impactantes, significativement, les unes par rapport aux autres, dans la définition des valeurs des paramètres.

En effet ce choix de la régression parcimonieuse s'est effectué de par le nombre conséquent de variables environnementales à disposition et notamment à cause de la faible importance de certaines variables par rapport à d'autres au regard de l'apriori métier que l'Agence de l'Eau nous a démontré.

Chercher les origines, les caractéristiques ou encore les paramètres qui influent le plus sur une notion métier peut en fonction du format des données être réalisé comme ici à partir d'une régression. Mais il est également possible de réutiliser la méthode des analyses factorielles utilisées pour le premier cas d'étude.

En effet comprendre les interactions (i.e. corrélations) entre les différentes variables sur une variable de référence à expliquer permet de dégager plusieurs groupes d'observations déterminables par des règles métiers

4 CAS 4

ANALYSE EXPLORATOIRE ET MODELE DE RISQUE

4.1 Rappel des objectifs

La surveillance ininterrompue des substances polluantes pour la préservation du milieu est onéreuse : pour de nombreux bassins versants de taille réduite (faible ordre de Strahler), l'Aerm ne dispose pas de relevés QUALITE qui lui permettrait pourtant d'accroître sa connaissance du milieu.

En revanche, les données environnementales et géomorphologiques dont elle dispose couvrent la quasi-totalité de son territoire d'exercice.

L'analyse du cas précédent a permis d'établir une liste de variables environnementales qui impacte significativement les concentrations de paramètres : à partir de là, peut-on prédire le risque de présence de pesticides à partir de la seule connaissance du contexte environnemental lié au bassin versant ?

Le Cas 4 va donc consister en l'élaboration d'un modèle prédictif capable de déterminer le risque de détections élevées de pesticide au sein d'un bassin versant, à partir de l'information disponible livrée à travers la connaissance de sa nature environnementale uniquement.

Les sections suivantes détaillerons la méthodologie sélectionnée au regard de l'objectif ciblé, puis le résultat final obtenu et présenté pour répondre au sujet et enfin une ouverture sur les modalités de réutilisation des procédés statistiques proposés.

4.2 Méthodologie appliquée

Pour construire le modèle de risque de pollution des cours d'eau en fonction des variables environnementales il nous a fallu construire une variable de risque.

En effet grâce à l'algorithme de régression parcimonieuse, nous avons obtenu une équation capable de calculer la valeur P90 des paramètres pour tous les points de surveillance. Et ne sont conservées que les variables environnementales qui ont révélées un impact significatif dans le niveau du P90 des paramètres/

Ainsi, nous possédons pour chaque point de surveillance autant d'équations de régression que de paramètres. Il est néanmoins nécessaire de pouvoir recoder ces valeurs calculées du P90 par paramètre sous une valeur qualitative et binaire de risque ou non risque. Pour cela nous avons sélectionné les règles de calculs suivantes :

- **Recodage 1** : Comparaison par rapport à la règle des deux écarts-types du paramètre (Rappel du Cas 1 : $valeur > 2\sigma$ ➔ valeur forte, donc risquée)
- **Recodage 2** : Comparaison à 2σ par rapport à la moyenne de la série des P90
- **Recodage 3** : Comparaison à $1{,}5\sigma$ par rapport à la moyenne de la série des P90

Ainsi, pour chaque paramètre nous obtenons un tableau de données contenant une notion de risque d'apparition grâce aux différentes valeurs des variables environnementales pour tous les points de surveillance utilisés dans le cas 3.

NUMNAT	DONNEES QUALITE	DONNEES ENVIRONNEMENTALES SIGNIFICATIVES CAS 3											
	P90	Variables Mères						Variables Filles					
		M_1	M_2	M_3	M_4	M_5	...	F_1	F_2	F_3	F_4	F_5	...
2000990	0	0,908	0,316	0,208	0,996	0,349	...	0,670	0,178	0,668	0,747	0,410	...
2001000	0	0,517	0,936	0,767	0,210	0,104	...	0,284	0,144	0,339	0,308	0,486	...
2001010	1	0,754	0,532	0,353	0,109	0,586	...	0,923	0,626	0,669	0,610	0,600	...
2001025	1	0,606	0,409	0,693	0,514	0,794	...	0,088	0,155	0,671	0,672	0,149	...
2001030	1	0,889	0,822	0,072	0,492	0,934	...	0,839	0,311	0,727	0,054	0,738	...
2001500	0	0,517	0,661	0,793	0,382	0,422	...	0,974	0,858	0,988	0,417	0,505	...
2001725	1	0,439	0,554	0,119	0,301	0,571	...	0,159	0,437	0,753	0,844	0,501	...
2001750	0	0,134	0,705	0,810	0,253	0,676	...	0,813	0,212	0,295	0,377	0,286	...
2001915	1	0,061	0,999	0,588	0,711	0,423	...	0,355	0,499	0,758	0,687	0,966	...
2001955	0	0,672	0,368	0,937	0,045	0,338	...	0,811	0,440	0,874	0,711	0,390	...
2001990	0	0,982	0,059	0,597	0,145	0,114	...	0,283	0,255	0,617	0,782	0,157	...
2002000	1	0,446	0,050	0,038	0,972	0,996	...	0,053	0,484	0,396	0,536	0,013	...
2003100	0	0,378	0,246	0,664	0,562	0,718	...	0,607	0,817	0,338	0,742	0,114	...
2003350	1	0,811	0,508	0,413	0,847	0,345	...	0,181	0,686	0,231	0,478	0,898	...
2003620	0	0,958	0,044	0,100	0,271	0,290	...	0,216	0,011	0,551	0,533	0,347	...
2003670	1	0,884	0,951	0,126	0,076	0,837	...	0,385	0,469	0,182	0,901	0,046	...
2006500	0	0,662	0,730	0,104	0,872	0,143	...	0,396	0,858	0,518	0,190	0,356	...
2007250	1	0,182	0,436	0,079	0,595	0,901	...	0,041	0,302	0,722	0,639	0,888	...
2007380	1	0,071	0,808	0,778	0,938	0,304	...	0,252	0,906	0,559	0,091	0,458	...
2009000	0	0,186	0,568	0,344	0,046	0,910	...	0,977	0,612	0,907	0,827	0,626	...

Exemple de tableau de données pour un paramètre avec sa variable de risque calculée

Avec un tel tableau de données en entrée il est possible de chercher à comprendre dans quelles mesures la valeur '1' de risque d'apparition de risque de pollution est présente ? Ou encore à partir de quel niveau de valeurs des variables environnementales il est possible d'observer un risque d'apparition d'une valeur forte du paramètre ciblé ?

Pour répondre à ce genre de problématique, nous proposons l'utilisation d'un algorithme d'apprentissage automatique. En effet ce type d'algorithme va, en parcourant les données, essayer de comprendre quelles variables environnementales ressortent le plus souvent dans les lignes des points de surveillance à risque.

Ainsi nous avons testé et sélectionné plusieurs méthodes de classification automatique pour ne conserver in fine que la méthode des arbres de décision.

Ce choix fut déterminé grâce aux bons résultats et à la compréhension aisée par des personnes peu formées à la Statistique, qu'offre cette méthode par rapport aux autres. Il est important de noter que la problématique du cas d'étude et la composition des jeux de données en entrée répondent également parfaitement aux prérequis des arbres de décision.

Cette technique propose en effet une série successive de nœuds organisés de façon descendante permettant de suivre la construction de l'arbre selon un raisonnement logique. En parcourant les différents nœuds de la racine jusqu'à la fin de l'arbre nous obtenons des règles de classement, appelées règles de décision.

Pour être valide et validé, un modèle d'apprentissage automatique a besoin de 2 jeux de données. 1 premier jeu pour apprendre justement dans quelles mesures la modalité de risque apparaît, et un second pour tester si ses règles de classement fonctionnement correctement ou si le modèle n'est pas en sur-apprentissage.

Il est ensuite facile de tester et de vérifier la robustesse d'un modèle en calculant la sensibilité et la précision du classement grâce à la construction d'une matrice de confusion. En effet sur cette matrice nous retrouvons aisément le nombre d'erreur de classement au sein des deux modalités 0 et 1 de notre variable de risque.

	Grp1	Grp2
Grp1	28	5
Grp2	4	22

Les cases vertes indiqueront par exemple le nombre de point de surveillance correctement classés dans leur groupe d'appartenance révélé en sortie du cas 3.

4.3 Synthèse des résultats du cas

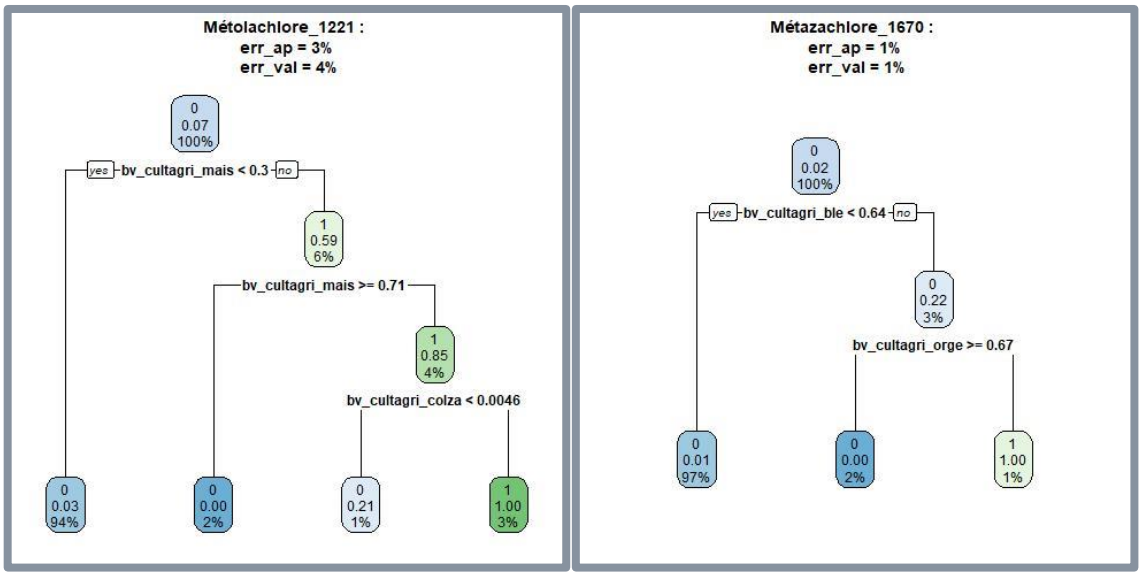
Afin d'illustrer la méthodologie avancée à la section précédente, cette synthèse des résultats développe et explique pour chaque objectif les réponses que nous avons apportées.

Pour aller plus loin, l'entièreté des étapes de calculs et restitutions produites sont trouvables dans le document livré pour la fin de la phase de développement : *BPM_AERM_Livrable_Final_Phase2_20170609*

Comme précisé dans la section précédente, un algorithme d'apprentissage automatique nécessite la construction de 2 jeux de données. Le premier, le jeu d'entraînement, soit (70%) des points de surveillance issus du cas 3, va permettre de calculer les différentes règles de classement des points de surveillance en risque ou non risque d'apparition de paramètres polluants. Le second, le jeu de

validation, soit (30%) des points de surveillance issus du cas 3 agira quant à lui comme un base de vérification des différentes règles.

En effet comme nous connaissons le risque de pollution, 0 ou 1, de tous les points de surveillance issus du cas 3, nous pourrons valider que les 30% des stations utilisées comme base de test sont correctement classées.



Résultat des règles de classement dans les modalités de la variable risque par l'arbre de décision Pour le métolachlore (a) et le métazachlore (b)

Bassins Versants

En reprenant l'exemple ci-dessus, nous pouvons proposer plusieurs règles de risque d'apparition du métolachlore et du métazachlore.

En effet pour construire la règle d'apparition du risque du 1er paramètre, il faut par exemple qu'un point de surveillance combine :

- Un ratio de culture agricole de maïs supérieur ou égale à 0.71
  -> successivement supérieur à 0.3 puis supérieur ou égale à 0.71
- Un ratio de culture de colza de 0.0046

Afin d'aller plus loin nous avons également préparé un jeu de données de test 'réel'. Réel car contenant 72 observations supplémentaires pour lesquelles le P90 n'était pas disponible mais sur lesquelles nous connaissions les valeurs des variables environnementales.

Les règles de décision comme celle citées ci-dessus pour le métolachlore, seront donc testées sur ces nouvelles données afin de donner la possibilité d'apparition par exemple du pesticide sur des cours d'eau non surveillés/prélevés.

En d'autre termes si risque il y a, alors il faudra valider que le cours d'eau en question ne requiert pas la mise en place d'un point de mesure de pollution pour valider le classement et démarrer des campagnes de traitement des eaux et de prévention de pollution de l'environnement au sein du bassin versant.

		Benzo(a)pyrène	Fluoranthène	Métolachlore	Arsenic	Zinc	Mercure	Cuivre	Glyphosate	Napthalène	Métazachlore	Diflufénicanil	Nicosulfuron	AMPA	Métolachlore OXA	Métolachlore ESA	Métazachlore OXA	Métazachlore ESA	amp	méta	méto	Indice Pesticide
		1115	1191	1221	1369	1383	1387	1392	1506	1517	1670	1814	1882	1907	6853	6854	6894	6895	11221	11670	11907	12000
	RISQUE ERREUR PREDICTION	6%	6%	8%	36%	29%	25%	38%	18%	19%	8%	3%	5%	11%	13%	20%	10%	3%	27%	14%	-	5%
2001020	LE SAURENTZ À SIERENTZ (AMONT)	Risque	Risque	Risque	Non Risque	Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	-	Non Risque
2001040	LE RUISSEAU DE MUHLBACH À ESCHENTZWILLER	Risque	Risque	Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	-	Non Risque
2003400	LA LARGUE À FRIESEN	Risque	Risque	Risque	Non Risque	Risque	Risque	Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	-	Non Risque
2004300	LA DOLLER À SEWEN (LERCHENMATT)	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	-	Non Risque
2022800	LA BLIND À BALDENHEIM	Risque	Risque	Risque	Risque	Risque	Risque	Risque	Risque	Risque	Non Risque	Non Risque	Risque	Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	-	Non Risque
2022825	LE DAECHERTSGRABEN À SELESTAT	Risque	Risque	Risque	Risque	Risque	Risque	Risque	Risque	Risque	Non Risque	Non Risque	Risque	Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	-	Non Risque
2023000	LE GIESSEN À THANVILLE	Risque	Risque	Non Risque	Risque	Risque	Risque	Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	-	Non Risque
2025700	LA LUTTER À BENFELD	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Risque	Non Risque	Risque	Non Risque	Non Risque	Non Risque	Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	-	Non Risque
2026500	LA ZEMBS À KRAFFT	Risque	Risque	Risque	Risque	Risque	Risque	Risque	Risque	Risque	Non Risque	Non Risque	Non Risque	Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	-	Non Risque
2028075	LA KIRNECK À BARR	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	-	Non Risque
2028200	L'ANDLAU À SCHAEFFERSHEIM	Risque	Risque	Risque	Risque	Risque	Risque	Risque	Risque	Risque	Non Risque	Non Risque	Risque	Risque	Non Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	-	Non Risque
2029160	L'EHN À BOERSCH	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	-	Non Risque
2029200	L'EHN À OTTROTT	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	-	Non Risque
2030450	LE VIEIL ERGELSENBACH À GEISPOLSHEIM	Risque	Risque	Risque	Risque	Risque	Risque	Risque	Risque	Risque	Non Risque	Non Risque	Risque	Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	-	Non Risque
2031540	LE RUISSEAU D'ALBET À LA BROQUE (LA CLAQUETTE)	Risque	Risque	Non Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	-	Non Risque
2031560	LE RUISSEAU DE FRAMONT À SCHIRMECK	Risque	Risque	Non Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	-	Non Risque
2031580	LE BARENBACH À BAREMBACH	Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	-	Non Risque
2031595	LE BASS DE RUSS À RUSS	Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	-	Non Risque
2031600	LA BRUCHE À WISCHES	Risque	Risque	Non Risque	Non Risque	Non Risque	Risque	Non Risque	Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	-	Non Risque
2031650	LE NETZENBACH À LUTZELHOUSE	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	-	Non Risque
2031800	LA MAGEL À MOLLKIRCH	Risque	Risque	Non Risque	Non Risque	Non Risque	Risque	Non Risque	Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	-	Non Risque
2032800	LA MOSSIG À ROMANSWILLER	Risque	Risque	Non Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	-	Non Risque
2034000	LA MOSSIG À WANGEN	Risque	Risque	Non Risque	Risque	Risque	Risque	Risque	Risque	Risque	Non Risque	Non Risque	Risque	Risque	Non Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	-	Non Risque
2036270	LE MUHLBACH À ACHENHEIM	Risque	Risque	Risque	Risque	Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	-	Non Risque
2037300	LA SOUFFEL À QUATZENHEIM	Risque	Risque	Risque	Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	-	Non Risque
2037350	LA SOUFFEL À GRIESHEIM-SUR-SOUFFEL	Risque	Risque	Risque	Risque	Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	-	Non Risque
2037450	LE LEISBACH À PFULGRIESHEIM	Risque	Risque	Risque	Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Non Risque	Risque	Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	-	Non Risque
2041000	LA MODER À INGWILLER	Risque	Risque	Non Risque	Risque	Non Risque	Risque	Risque	Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	-	Non Risque
2041100	LA MODER À MENCHHOFFEN	Risque	Risque	Non Risque	Risque	Risque	Risque	Risque	Risque	Risque	Non Risque	Non Risque	Risque	Risque	Non Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	-	Non Risque
2041700	LE FALKENSTEINBACH À NIEDERBRONN-LES-BAINS	Risque	Risque	Non Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	-	Non Risque
2043725	LE ROHRBACH À LANDERSHEIM	Risque	Risque	Risque	Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Non Risque	Risque	Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	-	Non Risque
2045220	LA SAUER À SURBOURG	Risque	Risque	Risque	Risque	Risque	Risque	Risque	Risque	Non Risque	Non Risque	Non Risque	Risque	Risque	Non Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	-	Non Risque
2045275	LE HALBMÜHLBACH À WALBOURG	Risque	Risque	Risque	Risque	Risque	Risque	Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Risque	Non Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	-	Non Risque
2045315	L'ASCHBACHGRABEN À RITTERSHOFFEN	Non Risque	Risque	Risque	Risque	Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	-	Non Risque
2045425	LE BRUMBACH À HAGUENAU	Non Risque	Non Risque	Non Risque	Non Risque	Risque	Risque	Risque	Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	-	Non Risque
2045500	L'EBERBACH À LEUTENHEIM	Risque	Risque	Risque	Risque	Risque	Risque	Risque	Risque	Risque	Non Risque	Non Risque	Risque	Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	-	Non Risque
2046400	LE SELTZBACH À SOULTZ-SOUS-FORÊTS	Risque	Risque	Risque	Risque	Risque	Risque	Risque	Risque	Risque	Non Risque	Non Risque	Risque	Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	-	Non Risque
2046500	LE WINTZENBACH À HOFFEN	Risque	Risque	Risque	Risque	Risque	Risque	Non Risque	Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	-	Non Risque
2046600	LE HAUSAUERBACH À HUNSPACH	Risque	Risque	Risque	Risque	Risque	Risque	Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	-	Non Risque
2048960	LE RUISSEAU DE LA COLLINE DE FRESSE À FRESSE-SUR-MOSELLE	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	-	Non Risque
2049500	LA MOSELOTTE À SAULXURES-SUR-MOSELOTTE	Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	-	Non Risque
2051820	LE BARBA À DOCELLES	Risque	Risque	Non Risque	Non Risque	Risque	Risque	Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	-	Non Risque
2066000	LA MEURTHE À THIAVILLE	Risque	Risque	Non Risque	Risque	Risque	Risque	Risque	Risque	Risque	Non Risque	Non Risque	Risque	Risque	Non Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	-	Non Risque
2067400	LE VACON À BARBAS	Risque	Risque	Risque	Risque	Non Risque	Risque	Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Risque	Risque	Risque	Risque	Non Risque	Risque	Non Risque	-	Non Risque
2085500	LE GRIOLOT À CONFLANS-EN-JARNISY	Risque	Risque	Non Risque	Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Non Risque	Risque	Risque	Risque	Risque	Risque	Risque	Non Risque	Non Risque	-	Non Risque
2085600	L'YRON À SAINT-BENOÎT-EN-WOEVRE	Risque	Risque	Non Risque	Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Risque	Risque	Non Risque	Risque	Risque	Non Risque	Non Risque	-	Non Risque
2086390	LE WOIGOT À MAIRY-MAINVILLE	Risque	Risque	Non Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Risque	Risque	Non Risque	Risque	Risque	Non Risque	Non Risque	-	Non Risque
2086400	LE WOIGOT À TUCQUEGNIEUX	Risque	Risque	Non Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Risque	Risque	Non Risque	Risque	Risque	Non Risque	Non Risque	-	Non Risque
2086460	LE RUISSEAU DE LA VALLÉE À BETTAINVILLERS	Risque	Risque	Non Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Risque	Risque	Risque	Risque	Risque	Risque	Non Risque	Non Risque	-	Non Risque
2088200	LE CONROY À FONTOY	Risque	Risque	Non Risque	Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Risque	Risque	Risque	Risque	Risque	Risque	Non Risque	Non Risque	-	Non Risque
2088600	LE CONROY À MOYEUVRE-GRANDE	Risque	Risque	Non Risque	Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Risque	Risque	Risque	Risque	Risque	Risque	Non Risque	Non Risque	-	Non Risque
2091700	LA FENSCH À FONTOY (AVAL)	Risque	Risque	Non Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Risque	Risque	Non Risque	Risque	Risque	Non Risque	Non Risque	-	Non Risque
2091800	LA FENSCH À KNUTANGE	Risque	Risque	Risque	Risque	Non Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	Risque	Risque	Non Risque	Risque	Risque	Non Risque	Non Risque	-	Non Risque
2091900	LA FENSCH À SEREMANGE	Risque	Risque	Risque	Risque	Risque	Non Risque	Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	Risque	Risque	Non Risque	Risque	Risque	Non Risque	Non Risque	-	Non Risque
2092400	LE RUISSEAU DE VEYMERANGE À THIONVILLE (VEYMERANGE)	Risque	Risque	Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Risque	Risque	Non Risque	Non Risque	-	Non Risque
2093600	LA CANNER À HOMBOURG-BUDANGE	Risque	Risque	Risque	Risque	Non Risque	Risque	Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Risque	Non Risque	Risque	Risque	Risque	Non Risque	Non Risque	-	Non Risque
2094973	LA SARRE BLANCHE À TURQUESTEIN-BLANCRUPT (AMONT STORINDAL)	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	-	Non Risque
2094978	LA SARRE ROUGE À SAINT-QUIRIN (LES DEUX RIVIERES)	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	-	Non Risque
2096400	L'ISCH À HIRSCHLAND	Risque	Risque	Risque	Risque	Risque	Risque	Risque	Risque	Risque	Non Risque	Non Risque	Risque	Risque	Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	-	Non Risque
2098450	LE BUTTENBACH À LORENTZEN	Risque	Risque	Risque	Risque	Risque	Risque	Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Risque	Risque	Risque	Risque	Risque	Non Risque	Non Risque	-	Non Risque
2101800	LA MERLE À L'HÔPITAL	Risque	Risque	Non Risque	Non Risque	Non Risque	Risque	Non Risque	Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	-	Non Risque
2103500	LA ROSSELLE À MORSBACH	Risque	Risque	Risque	Risque	Risque	Risque	Risque	Risque	Risque	Non Risque	Non Risque	Risque	Risque	Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	-	Non Risque
2103830	LA BISTEN À DIESEN	Risque	Risque	Risque	Risque	Non Risque	Risque	Risque	Risque	Risque	Non Risque	Non Risque	Risque	Risque	Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	-	Non Risque
2103840	LE RUISSEAU DE DIESEN À DIESEN	Risque	Risque	Risque	Risque	Non Risque	Risque	Risque	Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	-	Non Risque
2103910	LE GROSSBACH À MERTEN	Risque	Risque	Risque	Risque	Risque	Risque	Risque	Risque	Risque	Non Risque	Non Risque	Risque	Risque	Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	-	Non Risque
2103920	LE WEISBACH À BERVILLER	Risque	Risque	Risque	Risque	Non Risque	Risque	Non Risque	Non Risque	Risque	Non Risque	Non Risque	Non Risque	Risque	Non Risque	Risque	Risque	Risque	Non Risque	Non Risque	-	Non Risque
2112150	LE RUISSEAU DE FORGES À BETHINCOURT	Non Risque	Risque	Non Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Risque	Risque	Risque	Risque	Risque	Risque	Non Risque	Non Risque	-	Non Risque
2115653	LA CRUSNES À SERROUVILLE	Risque	Risque	Non Risque	Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Non Risque	Risque	Risque	Risque	Risque	Risque	Risque	Non Risque	Risque	-	Non Risque

Extrait du fichier récapitulatif des prédictions de risque de valeurs élevées pour les 21 substances ciblées par l'Aerm

4.4 Modalités de réutilisation

Les cas de réutilisation des algorithmes d'apprentissage automatique sont nombreux et divers, cependant pour être déployés il est nécessaire qui valident tous les prérequis suivants :

1. Une variable à expliquer

De préférence à 2 modalités même si certains algorithmes sont capables d'en gérer plusieurs. Dans notre cas d'étude la variable à expliquer était le risque d'apparition de valeurs fortes d'un ou plusieurs pesticides.

Nous avons vu dans ce cas que la variable de risque n'était pas donnée par les données sources mais qu'il nous a fallu la construire, notamment grâce aux résultats des régressions parcimonieuses du cas3.

2. Plusieurs variables explicatives

Le nombre de variables explicatives n'a pas, ou peu d'importance, mais il faut néanmoins posséder plus d'individus que de variables pour assurer la robustesse du modèle.

Il n'est également pas nécessaire que les données soient toujours quantitatives pour être utilisées dans un algorithme de classification automatique, cependant il est obligatoire qu'elles partagent toujours le même type, sauf pour les rares cas d'algorithmes exotiques et prototypiques.

Certains algorithmes sont en effet utilisables essentiellement sur des variables qualitatives et certains autres, à l'inverse, utilisables essentiellement sur des variables numériques.

3. 'Assez de données' de bonnes qualités

Cette notion assez abstraite dépend en fait essentiellement du domaine d'application de la méthode statistique. En effet plus on a de données plus le modèle a de la ressource pour construire ses règles et améliorer sa robustesse. Cependant parfois le nombre d'individus peut rendre flou la détermination de règles de décision parlante pour une application métier.

C'est à l'expert statistique de juger et de jauger du paramétrage de la méthode au regard de la problématique métier.

Enfin pour être réutilisée, notre méthode statistique proposée pour ce cas, pourra d'une part reprendre une notion de risque calculée à partir d'une véritable discussion entre experts métiers et expert de la donnée, ou encore être améliorée au fil du temps à partir d'autre règles de codification que celles que nous avons proposées.

En effet la notion de risque ne doit pas nécessairement être produite à partir d'un calcul mais peut par exemple combiner de la recherche métiers, des exceptions métiers et du recodage techniques pour traduire au mieux la décision métier.

5 CONCLUSION

Ces différents traitements analytiques et statistiques ont permis d'apporter des réponses à l'AERM pour la totalité des objectifs fixés en début de marché.

Même si toutes les méthodes proposées n'offrent pas de résultats satisfaisants par rapport aux aprioris métiers, fructueux ou alors directement exploitables, ces dernières ont toujours permis d'extraire des clés d'analyse supplémentaires pour les experts métiers, au regard des trois grandes entités métier de référence :

- Les sites de surveillance
- Les paramètres
- Le contexte environnemental

En effet, comme pour bon nombre de projets qui nécessitent l'application de méthodes exploratoires, la fouille de données soulève de nouvelles questions, qui engagent l'amélioration des techniques statistiques proposées ou des filtres de données sélectionnées.

Toutefois, cela remplit également l'un des objectifs sous-jacent de ce projet, qui consistait en l'amélioration de la connaissance structurelle des données. La mise en place d'une méthodologie de traitements a montré :

- Qu'il est possible d'employer la volumétrie de données enregistrées pour répondre à des problématiques métier clairement identifiées
- Que certaines données qui présentent un intérêt statistique mériteraient d'être mieux renseignées, à l'exemple de certaines variables environnementales pour la construction des arbres
- Que certaines données sont trop cloisonnées, et mériteraient d'intégrer une base métier pour être accessible à tous, à l'exemple des durées d'ensoleillement qui auraient pu être exploitées dans le cadre de l'étude d'impact des variables environnementales sur les concentrations

Au-delà de l'amélioration des techniques statistiques, ou du paramétrage des méthodes proposées, un projet d'analyse exploratoire des données trouve un réel aboutissement dans l'exécution de sa traduction fonctionnelle.

Idéalement, des spécifications devront être réalisées pour automatiser – sur la base des résultats produits dans le cadre de ce projet – la production de graphiques analytiques et pertinents, organisés par plans thématiques, filtrables et paramétrables par substances ou site de surveillance par exemple.

L'utilisation d'une plateforme de visualisation de données permettrait à terme de capitaliser sur l'entièreté des résultats obtenus durant le projet. Effectivement ce type de plateforme garanti la profitabilité en rendant les résultats accessibles à tous, que ce soit en administration (exécuter la production des restitutions), ou en simple consultation (via un déploiement des résultats exécutés sur un portail intranet local par exemple).

6 ANNEXES

6.1 Bibliographie

6.1.1 Détection de valeurs aberrantes

- *V.Planchon, Traitement des valeurs aberrantes: concepts actuels et tendances générales, Biotechnol. Agron. Soc. Environ. 2005.*
- *V.J.Hodge, A.Austin, survey of outlier detection methodologies, Artificial Intelligence Review, pp. 85-126, 2004.*

6.1.2 Combinaison ACP et K-means :

- *P.Prabhu et al, Improving the Performance of K-Means Clustering For High Dimensional Data Set, International Journal on Computer Science and Engineering (IJCSE) , june 2011.*
- *S.Momon et al, Identification de la signature acoustique des différents mécanismes sources lors d'essais de fatigue sur CMC : Application de classificateur supervisé et non supervisé, 2009.*
- *C.Ding and X.H.Lawrence, Principal Component Analysis and Effective K-means Clustering, Berkeley National Laboratory University of California, 2004.*
- *C.Ding and X.H.Lawrence, K-means Clustering via Principal Component Analysis, Berkeley National Laboratory University of California, 2004.*

6.1.3 Méthode d'autocorrélation

- M Gajić-Čapka, Periodicity of annual precipitation in different climate regions of Croatia, Theoretical and applied climatology, 1994.
- M.Small, K.Judd, Detecting periodicity in experimental data using linear modeling techniques, Centre for Applied Dynamics and Optimization Department of Mathematics University of Western Australia, 2008.

6.1.4 Régression pas à pas

- P-C.Telmo, J.Lousada, N.Moreira, Proximate analysis, backwards stepwise regression between gross calorific value, ultimate and chemical analysis of wood, Bioresource Technology, 2010.
- G.Hegyi, L.Z.Garamszegi, Using information theory as a substitute for stepwise regression in ecology and behavior, Behav Ecol Sociobiol, 2011.

6.1.5 Arbre de décision

- Q.Chen, A.E.Mynett, Predicting Phaeocystis globosa bloom in Dutch coastal waters by decision trees and nonlinear piecewise regression, 2004.
- S.P. Sevionovic, Risk in sustainable water resources management, Sustainability of Water Resources under Increasing Uncertainty, 1997.

6.2 Notes méthodologiques

6.2.1 Notes méthodologiques Cas 1

Dans cette section, nous présentons les notes méthodologiques relatives au Cas 1.

6.2.1.1 Calcul du coefficient de corrélation

6.2.1.1.1 Coefficient de corrélation – définition

Le coefficient de corrélation permet de donner une mesure synthétique de l'intensité de la liaison qui unit deux variables quantitatives ou numériques. En d'autres termes, il quantifie la relation linéaire entre deux variables.

La valeur de la corrélation entre deux séries de valeurs mesurées de paramètre (notées X et Y) a été calculée par le **coefficient de corrélation de Bravais-Pearson**, qui correspond à la covariance de X et Y divisée par le produit de leur écarts-types respectifs.

Ainsi, si la **covariance** entre deux séries de valeurs mesurées de paramètre X et Y est donnée par :

$$Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})\,(Y_i - \bar{Y})$$

Où :

- X et Y sont deux variables aléatoires de valeurs mesurées
- $\bar{X}$ et $\bar{Y}$ les moyennes respectives des variables aléatoires X et Y
- X_i et Y_i les réalisations des variables aléatoires X et Y en i
- N le nombre total d'observations

Alors le coefficient de corrélation de Bravais-Pearson est donné par :

$$r(X,Y) = r_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

Où σ_X et σ_Y sont les écarts-types respectifs des variables X et Y (cf. note méthodologique sur l'écart-type).

Le coefficient de corrélation de Bravais-Pearson varie entre -1 et 1. Son interprétation est la suivante :

- Si $r_{X,Y}$ **est proche de 0**, il n'y a pas de relation linéaire entre X et Y
- Si $r_{X,Y}$ **est proche de 1**, il existe une forte relation linéaire positive entre X et Y
- Si $r_{X,Y}$ **est proche de -**1, il existe une forte relation linéaire négative entre X et Y

Ainsi, le signe de $r_{X,Y}$ indique le sens de la corrélation, tandis que sa valeur absolue indique l'intensité de cette dernière, i.e la capacité à prédire les valeurs de Y en fonction de celles de X.

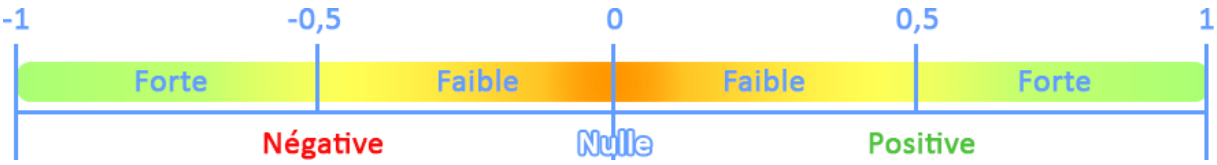


Schéma récapitulatif de l'interprétation du coefficient de corrélation de Bravais-Pearson

Note : Le coefficient de corrélation n'indique en aucun cas une relation de causalité d'une variable sur l'autre ou inversement.

Pour aller plus loin, on pourrait calculer les corrélations par le **coefficient de Spearman**, capable de détecter les relations non-linéaires entre deux variables. En d'autres termes, il permet de mettre en exergue l'existence de relations monotones (croissante ou décroissante), quelle que soit leur forme précise (linéaire, exponentielle, puissance, …)

6.2.1.1.2 Matrice des corrélations

Les matrices de corrélations correspondent simplement à une représentation matricielle des coefficients de corrélations entre deux variables deux à deux.

Sur la diagonale sont toujours représentés les coefficients de corrélations entre une variable et elle-même. Par définition, **la corrélation entre une variable et elle-même vaut toujours 1**.

Par construction, les parties diagonales supérieures et inférieures sont toujours symétriques, c'est pourquoi on ne représente généralement que l'une d'entres elles.

6.2.1.1.3 Coefficient de corrélation avec R

Dans R, le coefficient de corrélation entre deux variables X et Y est obtenu par la fonction `cor()` :

```
cor(X, Y, method = "pearson")
```

6.2.1.2 Normalisation

6.2.1.2.1 Normalisation - définition

La normalisation permet de simplifier la comparaison des variations entre séries de données, notamment lorsque :

- Elles ne présentent pas le même ordre de grandeur
- Elles ne sont pas enregistrées dans la même unité de mesure

L'avantage de cette technique statistique et qu'elle ne **crée pas d'incidence sur les profils de variation des deux séries** : les coefficients de corrélation restent identiques avant et après la normalisation.

Généralement, la normalisation est obtenue par centrage puis réduction de la variable en question. Ainsi, pour toute observation en i de la variable à centrer X, la transformation est la suivante :

$$\widehat{X}_i = \frac{X_i - \bar{X}}{\sigma_X}$$

Où :

- $\widehat{X}_i$ est la valeur centrée-réduite de la série en i
- $\bar{X}$ est la moyenne de la variable aléatoire X
- σ_X est l'écart-type de la variable aléatoire X

Note : On utilise généralement la normalisation par standardisation (centrage et réduction de variable) lorsque les distributions sont normales. Lorsque ce n'est pas le cas, on pourra préférer une **normalisation entre 0 et 1**. Dans ce cas, on peut appliquer la transformation suivante :

$$X_{min} \leq X_i \leq X_{max}$$
$$\Leftrightarrow\ 0 \leq\ X_i - X_{min} \leq X_{max} - X_{min}$$
$$\Leftrightarrow 0 \leq \frac{X_i - X_{min}}{X_{max} - X_{min}} \leq 1$$

6.2.1.2.2 Normalisation avec R

Dans R, la normalisation d'une variable par centrage, puis réduction s'obtient grâce à la fonction `scale()`:

```
scale(X, center = TRUE, scale = TRUE)
```

6.2.1.3 Calcul de l'écart-type

6.2.1.3.1 Ecart-type – définition

L'écart-type est une **mesure de dispersion des observations d'une variable aléatoire**, i.e qu'il donne une indication sur la dispersion des données enregistrées pour une série de valeurs mesurées autour de leur moyenne.

Techniquement, l'écart-type correspond à la racine carrée de la variance d'une variable, ou en d'autres termes, la **moyenne quadratique des écarts des observations par rapport à la moyenne**. Il est donné par :

$$\sigma_X = \sqrt{E[(X - E[X])^2]} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(X_t - \bar{X})^2}$$

Où :

- X est une variable aléatoire ou une série de valeurs mesurées
- X_t correspond à une réalisation de la variable aléatoire X
- $\bar{X}$ correspond à la moyenne de la variable aléatoire X
- n correspond au nombre total d'observations de la variable aléatoire X

6.2.1.3.2 Ecart-type avec R

Dans R, l'écart-type d'une variable aléatoire X est obtenu par la fonction `sd()` :

```
sd(X, na.rm = TRUE)
```

6.2.1.4 Analyse en Composantes Principales (ACP)

6.2.1.4.1 ACP – Principes Généraux

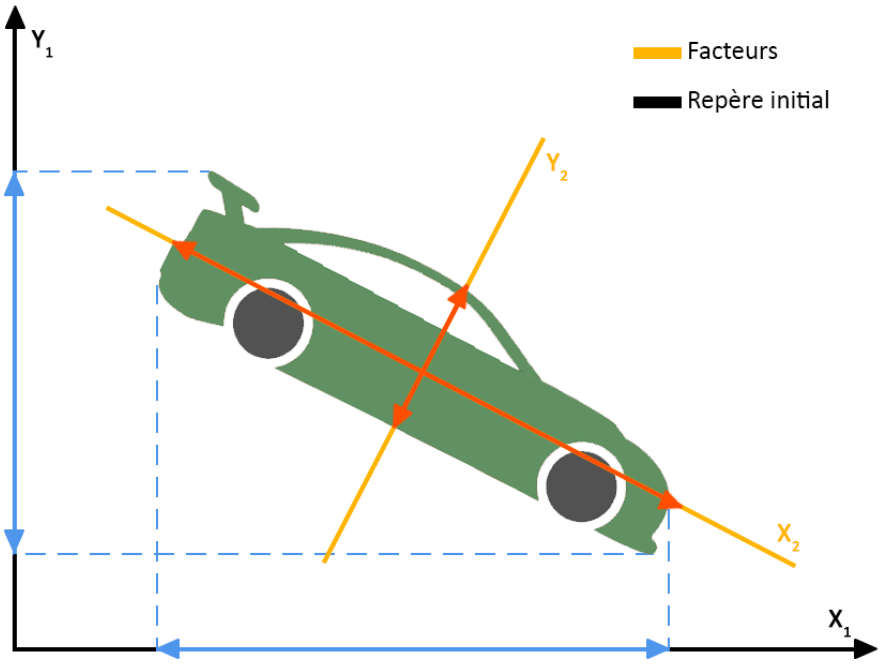
L'ACP est une méthode fondamentale en statistique descriptive multidimensionnelle. Elle permet de traiter simultanément un grand nombre de variables, toutes quantitatives, telles que dans le présent cas traité.

Cette méthode est purement descriptive : elle ne suppose, a priori, **aucun modèle sous-jacent de type probabiliste**. Il n'est donc pas nécessaire que les variables du modèle soient distribuées selon une loi normale ou respectent des hypothèses précises. On raisonne uniquement sur un calcul spatial basé sur des distances normées.

Parmi les méthodes de statistique descriptive multidimensionnelle, l'ACP est une méthode factorielle : on cherche, parmi toutes les variables initiales à priori liées entre elles, à déterminer des variables latentes (ou facteurs) *via* combinaison linéaire des variables observées. Ces axes factoriels sont indépendants, ou décorrellées, afin de restituer au mieux la variabilité des données, i.e l'information qu'elles contiennent.

L'ACP est donc à la fois :

- **Une approche géométrique** : on passe d'une représentation dans la base canonique des variables initiales à une représentation dans la base des axes factoriels (définis par les vecteurs propres de la matrice des corrélations). Le but étant de représenter les variables initiales dans un nouvel espace dimensionnel résumant au mieux toute l'information disponible.
- **Une approche statistique** : On sélectionne des axes qui ont une signification statistique importante, sur la base de la variance des données.

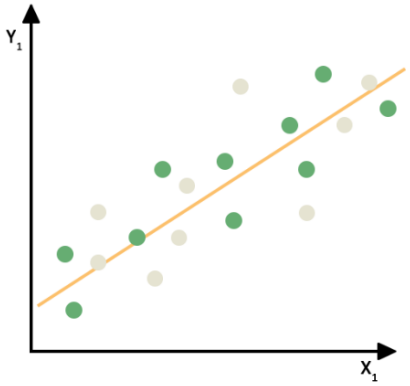


Exemple d'axe factoriel

Sur la figure ci-dessus, les axes factoriels X_2 et Y_2 sont obtenus par combinaison linéaire des axes du repère initial X_1 et Y_1. Le nouveau plan factoriel traduit bien mieux la dispersion de l'objet, interprété comme un nuage de points, car il est orthogonal à celui-ci. X_2 (resp. Y_2) donne directement la dispersion maximale horizontale (resp. verticale) de l'objet, alors que les axes initiaux mélangent les deux informations.

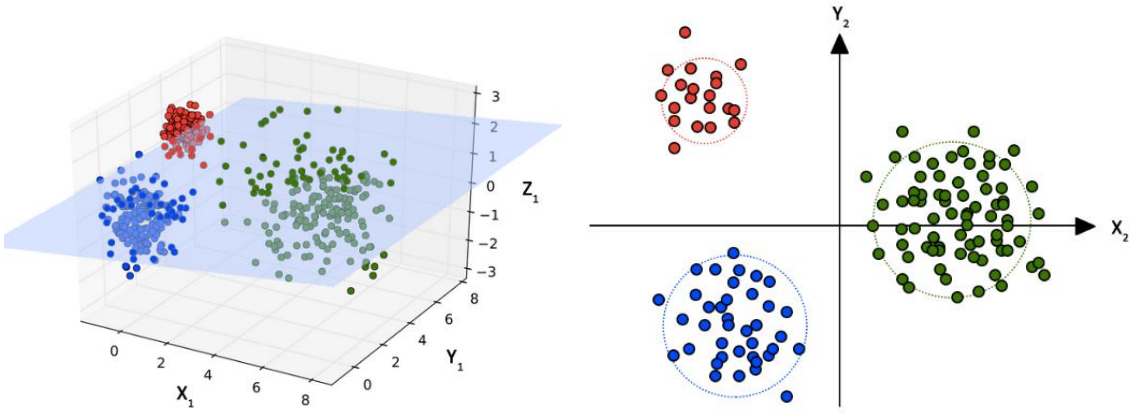
L'objectif de ce cas est de détecter des liens entre variables correspondantes à des couples de paramètres (colonnes de la matrice agrégée des corrélations). Si les corrélations entre deux variables peuvent aisément être rendues au sein d'un repère bidimensionnel voire tridimensionnel, comment visualiser les liaisons simultanées qui pourraient exister dans un espace multidimensionnel, ou chaque variable considérée représente une dimension ?

En d'autres termes, quel graphique permettrait de généraliser le nuage de points tracé dans le cas de deux variables permettant d'aborder la structure de corrélation présente entre plus de deux variables ? Nous proposons ci-dessous une illustration graphique.



Sur le graphique ci-contre, la projection des observations dans le repère normé par les variables X_1 et Y_1 permet d'identifier aisément la liaison positive qui unit les deux variables.

Dans une ACP, puisque le nombre de dimensions est très élevé et empêche l'observation directe, on va chercher à le réduire pour projeter les points dans un espace de dimension réduite afin de pouvoir observer les corrélations. Ce nouvel espace est formé par des axes qui forment une combinaison linéaire de toutes les variables initiales.



Sur la figure ci-dessus, on voit que l'espace normé par (X_1, Y_1, Z_1) a été réduit pour former l'espace normé par (X_2, Y_2), dans lequel les relations entre les observations sont plus aisément identifiables. L'expérience est encore plus flagrante sur un espace initial de plus de 3 dimensions.

6.2.1.4.2 ACP avec R

Dans R, lorsque l'on souhaite réaliser une ACP sur un tableau de données, on utilise la fonction `PCA()` de la librairie `FactoMiner` :

```
PCA(tableau_de_données, scale.unit = TRUE, ncp = nb_de_cp, graph = T)
```

6.2.1.5 Regroupement d'observation en classe par k means

La méthode k-means est une méthode itérative de classification supervisée qui consiste à partitionner les individus en classes homogènes, et dont le nombre « k » est défini au préalable.

Par exemple, si l'on souhaite partitionner une nuée initiale de n observations en **k groupes distincts**, l'algorithme effectue les étapes suivantes :

1. Il sélectionne k observations (tirées au hasard ou non), et considère qu'il s'agit des k barycentres (i.e le centre de chaque groupe). C'est le point de départ de la première itération.
2. Il calcule la distance entre les n observations et les k centres, et les affecte aux centres dont elles sont le plus proches.

   Lors de la première itération et à la fin de cette étape, on remarque que k groupes sont formés, mais ils ne sont pas forcément homogènes.
3. L'algorithme recalcule et redéfinit le barycentre de chaque groupe sur la base des observations qu'il leur a affecté à l'étape précédente
4. Il réaffecte les observations aux nouveaux barycentres calculés à l'étape précédente
5. Il réitère les phases 3 et 4 jusqu'à ce que la convergence soit atteinte, i.e lorsque la différence entre le barycentre calculé à l'étape p et p-1 n'est plus suffisamment significative.

A noter que l'inconvénient de cette méthode est qu'elle ne permet pas de découvrir quel peut être le nombre cohérent de classes, à l'instar de la classification hiérarchique ascendante.

6.2.1.6 Lecture d'un arbre de décision

Les arbres de décision sont des modèles qui permettent d'obtenir des modèles au pouvoir à la fois **explicatif** (détermination et validation des règles de classification en langage naturel ou métier sur un jeu d'apprentissage et un jeu de validation), et **prédictif** (utilisation des règles déterminées pour effectuer des prévisions sur le jeu de test).

Leur simplicité d'interprétation, du fait de leur visualisation didactique sous forme d'arbre, ainsi que leurs bonnes performances, les rend extrêmement populaires.

En principe, la méthode vérifie récursivement pour chaque nœud si une séparation est possible sur la base de la mesure choisie (indice de Gini), à partir de l'effectif total initial.

6.2.2 Notes méthodologiques Cas 2

Dans cette section, nous présentons les notes méthodologiques relatives au Cas 2.

6.2.2.1 Fonctions d'autocorrélations

La fonction d'autocorrélation permet d'étudier le lien linéaire qui existe entre les termes d'une série mesurée en t et ses termes retardés en $t-k$. Il existe deux types de fonctions d'autocorrélations.

6.2.2.1.1 Autocorrélation totale

L'autocorrélation totale mesure la corrélation d'un processus par rapport à une version décalée dans le temps de lui-même.

Ce signal est calculé exactement de la même façon que le coefficient de Pearson, i.e par la covariance des deux termes en t et $t-k$ divisée par le produit de leur écart-type respectif.

6.2.2.1.2 Autocorrélation partielle

L'autocorrélation partielle entre le terme en t et $t-k$ correspond à la régression linéaire de la série en t sur ses termes retardés, i.e à l'estimation par MCO de l'équation suivante :

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_k X_{t-k} + \varepsilon_t$$

Le coefficient ϕ_k estimé représente l'autocorrélation partielle d'ordre k, i.e l'apport d'explication de X_{t-k} à X_t, *toute chose égale par ailleurs*, i.e étant donné qu'on régresse également sur X_{t-1}, X_{t-2},…, X_{t-k+1}.

6.2.2.2 Analyse de la variance (ANOVA)

6.2.2.2.1 Tests préliminaires

Test de Shapiro-Wilk

Le test de Shapiro-Wil teste l'hypothèse nulle selon laquelle un échantillon est issu d'une population normalement distribuée.

Sachant que l'hypothèse nulle est que la population est normalement distribuée, si la p-value est inférieure au niveau alpha choisi, alors l'hypothèse nulle est rejetée (i.e. on conclut que les données ne sont pas issues d'une population normalement distribuée).

Si la p-value est supérieure au niveau alpha choisi, alors on ne peut pas rejeter l'hypothèse nulle selon laquelle les données sont issues d'une population normalement distribuée.

Par exemple, pour un niveau alpha de 0.05, un jeu de données avec une p-value de 0.32 n'entraîne pas le rejet de l'hypothèse nulle selon laquelle les données sont issues d'une population normalement distribuée. Donc si la p-value est grande, la distribution tend vers une distribution normale.

Test de Lévène

Le test de Lévène est un test de l'égalité des variances qui permet de vérifier l'égalité des variances entre des des niveaux de facteurs.

Sachant que l'hypothèse nulle est que toutes les variances sont égales, si la p-value est supérieure au niveau alpha choisi, alors l'hypothèse nulle est rejetée (i.e. on conclut que les effectifs ne sont pas de variance homogène).

Si la p-value est supérieure au niveau alpha choisi, alors on ne peut pas rejeter l'hypothèse nulle : toutes les variances ne sont pas égales.

6.2.2.2.2 Test Anova

L'analyse de la variance permet de comparer des moyennes sur plusieurs échantillons, afin de vérifier que ces échantillons sont bien issus d'une même population.

Par extension, il s'agit d'étudier l'influence des modalités (ou niveau) d'une variable qualitative (ou facteur), sur la distribution d'une variable quantitative à expliquer.

Remarque : Une comparaison de moyennes sur deux échantillons est possible grâce au test de Student ou d'un test z utilisant la loi normale. En revanche, une analyse sur trois échantillons indépendants ou plus nécessite un test Anova.

L'hypothèse à vérifier (H0) est que tous les échantillons ont la même moyenne. L'hypothèse alternative est qu'au moins une moyenne est significativement différente des autres.

La règle de décision du test est la suivante :

- Si le **F de Fisher** est supérieur à sa valeur critique, on rejette H0 et inversement.
- **Règle équivalente :** On accepte H si la **p-value** du test est supérieure au risque de première espèce α

6.2.2.3 Régression linéaire sur variables qualitatives (Moindres Carrés Ordinaires)

L'équation de régression de notre variable de concentration sur les modalités de la variable PHASE_HYDROLOGIQUE s'écrit :

$$Y = \mu\mathbb{I} + A_C\alpha + \varepsilon$$

Où :

- Y est le vecteur des concentrations mesurées
- $A_C = (\mathbb{I}_{Pic}, \mathbb{I}_{Montée}, \mathbb{I}_{Descente}, \mathbb{I}_{Stabilité})$ la matrice indiquant l'appartenance aux modalités Pic, Montée, Descente et Crue
- $\mathbb{I}_{Pic} + \mathbb{I}_{Montée} + \mathbb{I}_{Descente} + \mathbb{I}_{Stabilité} = \mathbb{I}$
- α la matrice de coefficients à estimer par MCO

Afin de pouvoir identifier un α unique, i.e d'avoir un modèle identifiable, la méthode la plus classique consiste à donner des contraintes.

Nous choisissons la contrainte suivante : $\alpha_{Stabilité} = 0$, ce qui revient à dire que la modalité « Stabilité » va servir de référence.

Dans ce cas, les estimateurs des moindres carrés (MCO) des paramètres liés à chaque phase hydrologique sont :

$$\hat{\mu} = \overline{y_{Pic}}$$

$$\widehat{\alpha_i} = \overline{y_i} - \overline{y_{Pic}}$$

La modalité « Pic » sert de référence. Le coefficient $\hat{\mu}$ (i.e la constante du modèle) est donc égal à la moyenne empirique de la cellule de référence « Pic ».

Les coefficients individuels estimés de chaque modalité $\widehat{\alpha_i}$ correspondent à l'effet différentiel entre la moyenne de la cellule i et la moyenne de la cellule « Pic ».

6.2.3 Notes méthodologiques Cas 3

Dans cette section, nous présentons les notes méthodologiques relatives au Cas 3.

6.2.3.1 Coefficient de détermination (R^2) d'une régression

Le coefficient de détermination mesure l'adéquation entre un modèle issu d'une régression linéaire (simple ou multiple) et les données observées qui ont permis de l'établir.

Il se situe entre 0 (le modèle ne vaut rien) et 1 (il est parfait).

Il se définit par le rapport **entre la variance expliquée par le modèle et la variance totale.**

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}$$

Le coefficient de détermination ne tient donc pas compte du nombre de variables. Son principal défaut est par ailleurs de croître avec le nombre de variables explicatives. Or, un excès de variables produit des modèles peu robustes.

C'est pourquoi l'on utilise le critère AIC dans la régression pas-à-pas comme critère d'optimisation.

6.2.3.2 Critère d'Akkaike (AIC)

L'AIC est également un critère d'évaluation et d'adéquation de modèles. Il permet notamment de les comparer entre eux, contrairement au coefficient de détermination.

L'AIC utilise le maximum de vraisemblance, mais en pénalisant les modèles comportant trop de variables, ce qui en fait un candidat de choix pour le critère d'optimisation de la régression parcimonieuse.

Sa formulation est la suivante :

$$AIC = -2\ln L(\theta) + 2k$$

Où $L(\theta)$ représente la log-vraissemblance du modèle et k le nombre de paramètre.