



Quadriges² - Référentiel National de gestion des données de la surveillance littorale

Emilie Gauthier

Octobre 2017

Qualification des données Quadriges²

Bilan du processus de qualification automatique et experte

SOMMAIRE

1	Qualification : définition et mise en œuvre	4
1.1	Définition	4
1.2	Historique du processus de qualification	4
1.3	Les thématiques	5
2	Les règles de qualification « automatique » et « expert »	6
2.1	Hydrologie	6
2.1.1	Définition de la thématique	6
2.1.2	Intervenants identifiés	6
2.1.3	Règles de qualification automatique (anomalies recherchées)	8
2.1.4	Règles de contrôle Quadrige	9
2.2	Microbiologie.....	12
2.2.1	Définition de la thématique	12
2.2.2	Intervenants identifiés	12
2.2.3	Règles de qualification automatique.....	13
2.2.4	Règles de contrôle Quadrige	15
2.3	Chimie.....	17
2.3.1	Définition de la thématique	17
2.3.2	Intervenants identifiés	17
2.3.3	Règles de qualification	17
2.3.4	Règles de contrôle Quadrige	21
3	Bilan général de la qualification (septembre 2017)	21
3.1	Volumétrie.....	21
3.2	Niveaux de qualité des données.....	22
3.3	Les intervenants	23
3.4	Les outils informatiques	24
3.5	Hydrologie	24
3.6	Microbiologie.....	26
3.1	Chimie.....	27
3.1.1	Programme RNOPHY (IMPOSEX).....	27
3.1.1	Programme RNOMV.....	27
3.1.2	Programme RNOSED	28
4	Simulations pour les données 2016	30
4.1	Hydrologie	30
4.1.1	Nombre d'anomalies général	30
4.1.2	Simulations détaillées	30
4.1.3	Conclusion	34

4.1	Microbiologie.....	35
4.1.1	Nombre d'anomalies.....	35
4.1.2	Détail des anomalies.....	36
4.1.3	Conclusion.....	38
4.1	Chimie.....	38
4.1.1	Programme RNOPHY.....	38
4.1.2	Programme RNOMV.....	39
4.1.3	Programme RNOSED.....	40
5	Conclusion.....	41
Annexe : spécifications pour la qualification experte des données RNOSED.....		43
Spécification pour qualification RNOSED Experte (V1).....		43

1 Qualification : définition et mise en œuvre

1.1 Définition

La qualification a pour objectif d'attribuer un niveau de qualité aux données : Bon, Douteux ou Faux. Le processus de qualification consiste à rechercher des anomalies plus ou moins complexes dans les données stockées, les anomalies étant définies par les responsables de programme Quadrigé ou responsables thématiques. Ces recherches d'anomalies sont faites en lien avec la Cellule Quadrigé.

On distingue plusieurs types de qualification :

- **Qualification ponctuelle de quelques résultats** : Exemples : on sait qu'un échantillon a été mal conservé et que les résultats de son analyse seront douteux, ou on a détecté un problème analytique et on sait que les résultats sont faux.
- **Qualification d'un jeu de données important** :
 - soit lors d'une reprise de données dont la qualité a été évaluée préalablement à l'intégration dans Quadrigé (dans ce cas les données sont intégrées avec un niveau de qualification déjà défini),
 - soit un expert thématique travaille sur un jeu de données précis, que ce soit pour un rapport d'étude, une publication scientifique, ou toute autre analyse statistique : il peut transmettre les résultats de son expertise à la cellule d'administration Quadrigé, et des niveaux de qualité peuvent être attribués aux données.
- **Qualification dite « automatique »** : mise en place depuis 2009 sur certaines thématiques, elle a pour objectifs de qualifier les données en routine via la recherche d'anomalies prédéfinies par des experts thématiques. Des programmes de recherche de ces anomalies sont lancés périodiquement par la Cellule Quadrigé, générant des fichiers d'anomalies renvoyés aux producteurs de données. Après vérification, les réponses des producteurs sont validées par les experts thématiques et sont intégrées dans la base : les données sont qualifiées après une éventuelle correction.
- **Qualification dite « experte »** : elle consiste à analyser les résultats selon des modèles statistiques définis par des experts thématiques avec l'équipe de biostatisticiens du service VIGIES de l'Ifremer (service auquel appartient la cellule Quadrigé).

1.2 Historique du processus de qualification

L'outil Quadrigé (depuis 1996) puis Quadrigé² (depuis 2008) permet aux utilisateurs de cette base d'attribuer différents niveaux de qualité à leurs données au cours du temps : contrôle, puis validation, puis qualification. Le contrôle et la validation sont effectués en continu par les utilisateurs depuis la mise en service de Quadrigé.

Les travaux de qualification des données de Q² ont été réalisés depuis de nombreuses années sous différentes formes :

- Années 2000 : échanges entre les coordinateurs du réseau REMI (microbiologie) et les laboratoires producteurs de données (LERs) pour corriger / qualifier les données douteuses. Les LERs ont fourni de nombreux retours sous forme de cahiers papier.
- 2004-2005 : travaux de qualification des données de contamination chimique (réseau RNO) : réalisation de programmes informatiques d'édition de graphiques automatisée, et qualification de presque 112 000 résultats.
- 2007 : utilisation de l'outil de qualification de l'application Quadrigé² pour qualifier des données de contamination chimique (plus de 130 000 résultats qualifiés).
- 2009 : préparation d'un processus de qualification automatique des données de toxines phytoplanctoniques (réseau REPHY) : développement des scripts avec plusieurs CDDs successifs
- 2009-2012 : mise en œuvre d'un processus de qualification « automatisé » pour qualifier les données hydrologiques et microbiologiques : plusieurs CDDs se succèdent dans la reprise des scripts existants et leur développement pour les nouvelles thématiques.

- 2010 : audit du processus de qualification « automatisé » par des « Data Manager » en recherche clinique : validation du processus et amélioration de la traçabilité.
- 2013 : qualification en routine des données hydrologiques et microbiologiques : l'avancement est présenté dans le paragraphe suivant.
- 2014 : qualification via le processus « automatisé » de l'ensemble des métadonnées chimie (réseau RNO/ROCCH), et qualification par expertise des résultats d'analyse chimique (même réseau).

Depuis 2014 les processus de qualification doivent être maintenus par le personnel permanent de la Cellule Quadrigé (plus de CDD) et les experts thématiques. Compte tenu de la montée en charge de l'ensemble de l'équipe VIGIES (notamment due à l'arrivée de la DCSMM), les programmes n'ont pas été relancés. L'avancement actuel de la qualification automatique et experte est donc le fruit des travaux menés de 2009 à 2014.

1.3 Les thématiques

Toutes les données de Quadrigé ne font pas l'objet de qualification (même si elles ont toutes vocation à le faire). A ce jour quatre thématiques présentent des données qualifiées (Tableau 1). Pour la qualification « automatique » les thématiques font l'objet de critères de définition qui sont précisés dans les chapitres suivants.

Tableau 1 : Liste des thématiques et des programmes dont des données sont qualifiées.

Thématique	Mode de qualification	Programme	Nombre de résultats de mesure
Chimie	Automatique + experte + ponctuelle	RNOINT	50
		RNOMV	196 449
		RNOPHY	4 717
		RNOSED	33 961
		ROCCEAUCHIMIE	469
	Ponctuelle	REPOMS	887
Hydrologie	Automatique + ponctuelle	ARCHYD	79 503
		REPHY	288 494
		RHLN	12 203
		RNOHYD	698 003
		SRN	23 857
		OMEGA	62
	Reprise de données	REPHY (Réseau des estuaires Bretons)	116 510
		RESLOC_44	80 386
Microbiologie	Automatique + ponctuelle	CMIC	3
		REMIC	15 637
		REMIE1	39
		REMIE2	8
		REMIS	130
		REMI-SURV	83
Zooplankton	Reprise de données + ponctuelle	IGAZOH	17 758

D'autres thématiques sont en attente de mise en place d'un processus automatique et notamment les flores phytoplanctoniques et les toxines phytoplanctoniques. Des travaux ont débuté pour les toxines en 2009 et 2010 (programmes R développés par un sous-traitant sur des spécifications de Catherine Belin) mais sont à reprendre.

2 Les règles de qualification « automatique » et « expert »

2.1 Hydrologie

2.1.1 Définition de la thématique

On entend par « hydrologie » les données des réseaux de surveillance pérennes qui acquièrent des données hydrologiques (Tableau 2).

Tableau 2 : Critères de définition de la thématique Hydrologie pour la qualification « automatique ».

Programmes		Paramètres		
REPHY		TEMP	NH4	
RNOHYD	RHLN	SALI	NO2	<i>La liste est-elle toujours exhaustive?</i>
SRN	ARCHYD	OXYGENE	NO3	
RSLHYD		TURB	NO3+NO2	
		TURB-FNU	PO4	
		CHLOROA	SIOH	PHEO

2.1.2 Intervenants identifiés

Le Tableau 3 présente la liste des personnes identifiées pour chaque laboratoire impliqué dans la qualification comme producteur de données. Les scripts de qualification automatique se basent sur cette liste pour l'envoi automatique des fichiers d'anomalie générés.

Tableau 3 : Liste des intervenants identifiés pour la qualification des données hydrologiques.

Laboratoire	Nom	Rôle	Commentaires
DYNECO PELAGOS	Anne Daniel	Qualificateur	
EMP PHYC	Catherine Belin	Qualificateur	
GUAD	Christelle Batailler	Producteur de données (PARETO)	
GUAD	Anne Daniel	Soutien au Producteur de données	
GUAD	Catherine Belin	Soutien au Producteur de données	
LER AR	Myriam Perrière-Rumèbe		
LER AR	Florence D'Amico		
LER AR	Gilles Trut	Producteur de données	
LER AR	Claire Barbier Méteigner	Producteur de données	
LER BL	Vincent Duquesne	Producteur de données	
LER BL	Alain Lefebvre	Producteur de données	
LER BL	Camille Blondel	Producteur de données	
LER BN	Aurelie Legendre	Producteur de données	

Laboratoire	Nom	Rôle	Commentaires
LER BN	Julien Chevé		
LER BN	Françoise Dagault	Producteur de données	
LER BN	Aurore Lejolviet	Producteur de données	
LER BO	Anne Doner	Producteur de données	
LER BO	Claude le Bec	Chef de station	
LER BO	Audrey Duval	Producteur de données	
LER LR	Eric Abadie	Producteur de données	
LER LR	Claude Chiantella	Producteur de données	
LER LR	Tom Berteaux	Producteur de données	
LER MPL NT	Karine Collin	Producteur de données	
LER MPL NT	Yoann Le Merrer	Producteur de données	
LER MPL NT	Mireille Fortune	Producteur de données	
LER MPL TM	Jacky Chauvin	Producteur de données	
LER MPL TM	Michael Retho	Producteur de données	
LER MPL TM	Raoul Gabellec	Producteur de données	
LER N	Florence Nédelec	Producteur de données	
LER N	Sylvaine Françoise	Producteur de données	
LER N	Emilie Rabiller	Producteur de données	
LER PAC CO	Yoann Baldi	Producteur de données	
LER PAC TL	Françoise Marco-Miralles	Producteur de données	
LER PC LR	Sylvie Genauzeau	Producteur de données	LER PC LR
LER PC LR	Annick Derrien	Producteur de données	LER PC LR
LER PC LR	Anne Schmitt	Producteur de données	LER PC LR
LER PC LT	Sylvie Genauzeau	Producteur de données	LER PC LR
LER PC LT	Annick Derrien	Producteur de données	LER PC LR
LER PC LT	Anne Schmitt	Producteur de données	LER PC LR
MART	Brigitte Ravail		Saisie en sous-traitance des données Martinique
MART	Anne Daniel	Soutien au Producteur de données	
MART	Catherine Belin	Soutien au Producteur de données	
RNOHYD	Didier Claisse	Producteur de données	
RNOHYD	Jean François Chiffolleau	Producteur de données	
RNOHYD	Anne Daniel	Producteur de données	Qualifie les données à la place de Didier Claisse
RSLHYD	Emilie Gauthier	Qualificateur	
SREU	Michel Ropert	Soutien au Producteur de données	
SREU	Jean Turquet	ARVAM	
SREU	Harold Cambert	ARVAM	
SREU	Magali Duval	Soutien au Producteur de données	
AAMP MAY	Magali Duval	Soutien au Producteur de données	

2.1.3 Règles de qualification

2.1.3.1 Qualification automatique (anomalies recherchées)

Tableau 4 : Règles de qualification automatique des données hydrologie.

Pack	Niveau de priorité	N°Anomalie	Description
HPAC	0	HPAC-001	Cohérence des PSFM utilisés dans le cadre de l'hydrologie : est-ce que les paramètres, les supports, les fractions, les méthodes et les PSFM dans leur ensemble sont cohérents par rapport aux analyses effectuées ? Les données incohérentes doivent être identifiées au cas par cas.
HPAC	0	HPAC-002	« Conflits thématiques » : recherche des autres programmes rattachés aux données à qualifier, et recherche des autres résultats rattachés aux données in situ à qualifier.
HPAC	0	HPAC-003	Champs vides : les champs supposés être vides dans le cadre de cette thématique ne le sont pas.
HPAC	1	HPAC-004	1) Vérification de la cohérence des niveaux de saisie des résultats dans les stratégies. 2) Résultats de mesure saisis à un autre niveau que celui prévu dans la stratégie.
HPAC	1	HPAC-005	Doublons de résultats de mesure : plusieurs résultats de mesure pour un même Lieu / Date / Heure / Niveau / Paramètre
HPAC	2	HPAC-006	Cohérence des préleveurs des stratégies
HPAC	2	HPAC-007	Cohérence des analystes des stratégies
HPAC	2	HPAC-008	Données non validées (cela implique de contrôler et valider les résultats, ou de les supprimer ; ainsi que toute leur chaîne de saisie : passage / prélèvement / échantillon).
PASS	1	PASS-001	Heure de passage < 6h ou > 20h
PASS	1	PASS-002	Unité de sonde différente du mètre si la sonde est remplie
PREL	1	PREL-001	Heure de prélèvement différente de celle du passage (si non nulle)
PREL	1	PREL-002	Préleveur saisi différent de celui prévu dans la stratégie applicable (si le préleveur est précisé dans la stratégie)
PREL	1	PREL-003	Unité d'immersion différente du mètre (si immersion renseignée)
PREL	2	PREL-004 ¹	Combinaison sonde - engin - niveau - immersion - taille prélèvement - unité taille prélèvement incohérente
RESU	1	RESU-001	Analyste différent de celui prévu dans la stratégie (si l'analyste est renseigné dans la stratégie)
RESU	2	RESU-002	$1 \leq \text{TEMP} \leq 32$
RESU	2	RESU-003	$0.05 \leq \text{SALI} \leq 41$

¹ Remarque : dans le cadre de la qualification des données microbiologiques, l'anomalie de combinaison engin – niveau – immersion – support est précédée de deux anomalies : engins aberrants et niveaux aberrants, afin d'éliminer les erreurs les plus grossières avant de rechercher les combinaisons douteuses. Dans le cadre de l'hydrologie, les engins et les niveaux de prélèvement aberrants représentent une partie infime des données (<1% des prélèvements au 08/03/2012 sur l'ensemble des données). Il a donc été décidé de rechercher directement les combinaisons sonde – engin – niveau – immersion - taille prélèvement - unité taille prélèvement douteuses.

Pack	Niveau de priorité	N°Anomalie	Description
RESU	2	RESU-004	$0.05 \leq \text{TURB} \leq 70$
RESU	2	RESU-005	$0.05 \leq \text{TURB-FNU} \leq 70$
RESU	2	RESU-006	$0.01 \leq \text{OXYGENE} \leq 20$
RESU	2	RESU-007	$0.05 \leq \text{CHLOROA} \leq 70$
RESU	2	RESU-008	$0.05 \leq \text{PHEO} \leq 5$
RESU	2	RESU-009	$0.05 \leq \text{NH}_4 \leq 40$
RESU	2	RESU-010	$0.01 \leq \text{NO}_3 + \text{NO}_2 \leq 800$
RESU	2	RESU-011	$0.01 \leq \text{NO}_3 \leq 800$
RESU	2	RESU-012	$0.01 \leq \text{NO}_2 \leq 5$
RESU	2	RESU-013	$0.02 \leq \text{PO}_4 \leq 15$
RESU	2	RESU-014	$0.01 \leq \text{SIOH} \leq 400$

2.1.3.2 Qualification experte

Des règles de qualification experte ont été définies par C. Belin et A. Daniel et ont fait l'objet de développements en 2012. Un document présente les règles définies (Daniel A. & Belin C., 2013. Procédures pour la qualification experte des données Quadrigé² - Données hydrologie. Rapport interne Ifremer, 14p.).

2.1.4 Règles de contrôle Quadrigé

Des règles de contrôle à la saisie peuvent être définies dans l'application Quadrigé. Certains LERs avaient défini des règles de contrôle dès 2009. A partir de 2012 des règles nationales s'appliquant à tous les saisisseurs de données des réseaux de surveillance pérennes nationaux et régionaux coordonnées par Ifremer ont été mises en place (liste de règle nommée « Hydrologie » dans les Tableau 5 et Tableau 6).

Tableau 5 : Listes de règles de contrôle en place dans Quadrigé : programmes concernés et laboratoires saisisseurs contrôlés.

Liste de règle de contrôles	Programmes concernés	Laboratoires
controle_LERAR	ARCHYD	PDG-ODE-LITTORAL-LERAR
controle_LERBL	REPHY SRN	PDG-ODE-LITTORAL-LERBL
controle_LERMPL	MORHYD REPHY	PDG-ODE-LITTORAL-LERMPL
controle_REPHY	REPHY	PDG-ODE-VIGIES
Hydrologie	ARCHYD REPHY RHLN RNOHYD RSLHYD SRN	BIOLITT IMPACT-MER PDG-ODE-LER-LERFBN PDG-ODE-LITTORAL-LERAR PDG-ODE-LITTORAL-LERBL PDG-ODE-LITTORAL-LERBN PDG-ODE-LITTORAL-LERBO PDG-ODE-LITTORAL-LERLR PDG-ODE-LITTORAL-LERMPL PDG-ODE-LITTORAL-LERN PDG-ODE-LITTORAL-LERPAC PDG-ODE-LITTORAL-LERPC PDG-ODE-VIGIES

		PDG-RBE-BE PDG-RBE-SREU SPEL2A
LERN	COGEHY REPHY	PDG-ODE-LITTORAL-LERN

Tableau 6 : Liste des règles de contrôle appliquées à la saisie des données hydrologiques.

Attribut contrôlé	Paramètre	Fonction	Contrôle effectué	Liste	Règle
Passage Mnémonique		Not empty		controle_LERAR	controle_LERAR_1
Passage Zone de destination de dragage		Empty		Hydrologie	Hydrologie_1
Passage Heure		Min Max]06:00 ; 20:00[controle_REPHY	controle_REPHY_1
Passage Heure		Min Max]05:59 – 20:01[Hydrologie	Hydrologie_3
Passage Sonde		Min Max]0 ; 101[Hydrologie	Hydrologie_28
Passage Unité de sonde		In	« m »	Hydrologie	Hydrologie_30
Prélèvement Engin		In	Bouteille go-flo Bouteille hydrobios Pompe Seau plastique Bouteille à prélèvement horizontal Flacon plastique sur perche Flacon polypropylène Bouteille type Niskin tous volumes Mesures in situ	Hydrologie	Hydrologie_13
Prélèvement Niveau		In	Surface (0-1m) Mi-profondeur Fond/sonde-1m Colonne d'eau de 3 à 5 mètres de 6 à 10 mètres de 11 à 15 mètres de 16 à 20 mètres de 21 à 25 mètres de 31 à 40 mètres de 41 à 50 mètres 2 mètres Surface-Fond (profondeur <3 m)	Hydrologie	Hydrologie_14
Prélèvement Nombre d'individus		Empty		Hydrologie	Hydrologie_2
Prélèvement Immersion		Min Max]0 ; 51[Hydrologie	Hydrologie_29
Prélèvement Immersion Unité		In	« m »	Hydrologie	Hydrologie_12
Prélèvement Lot		Empty		Hydrologie	Hydrologie_6
Echantillon Nombre d'individus		Empty		Hydrologie	Hydrologie_10
Echantillon Taxon support		Empty		Hydrologie	Hydrologie_7
Echantillon Groupe de taxons support		Empty		Hydrologie	Hydrologie_8
Echantillon Taille de l'échantillon		Empty		Hydrologie	Hydrologie_9
Résultat de mesure Valeur d'incertitude	TEMP	Empty		Hydrologie	Hydrologie_11
Résultat de mesure Valeur numérique	TEMP	Min Max]3 ; 23[controle_LERBL	controle_LERBL_1

Attribut contrôlé	Paramètre	Fonction	Contrôle effectué	Liste	Règle
Résultat de mesure Valeur numérique	TEMP	Min Max]5 ; 25[controle_LERMPL	controle_LERMPL_1
Résultat de mesure Valeur numérique	TEMP	Min Max]0.99 ; 32[Hydrologie	Hydrologie_15
Résultat de mesure Valeur numérique	TEMP	Min Max]2 ; 24[LERN	LERN_3
Résultat de mesure Valeur d'incertitude	SALI	Empty		Hydrologie	Hydrologie_11
Résultat de mesure Valeur numérique	SALI	Min Max]0.049 ; 41[Hydrologie	Hydrologie_16
Résultat de mesure Valeur numérique	SALI	Min Max]30 ; 35.5[controle_LERBL	controle_LERBL_2
Résultat de mesure Valeur numérique	SALI	Min Max]5 ; 35[controle_LERMPL	controle_LERMPL_2
Résultat de mesure Valeur numérique	SALI	Min Max]25 ; 36[LERN	LERN_2
Résultat de mesure Valeur numérique	PH	Min Max]7.9 ; 8.5[controle_LERBL	controle_LERBL_3
Résultat de mesure Valeur d'incertitude	NO2	Empty		Hydrologie	Hydrologie_11
Résultat de mesure Valeur numérique	NO2	Min Max]0.04 ; 1[controle_LERBL	controle_LERBL_6
Résultat de mesure Valeur numérique	NO2	Min Max]0.009 ; 5[Hydrologie	Hydrologie_25
Résultat de mesure Valeur d'incertitude	NO3	Empty		Hydrologie	Hydrologie_11
Résultat de mesure Valeur numérique	NO3	Min Max]0.14 ; 40[controle_LERBL	controle_LERBL_5
Résultat de mesure Valeur numérique	NO3	Min Max]0.099 ; 800[Hydrologie	Hydrologie_24
Résultat de mesure Valeur d'incertitude	NO3+NO2	Empty		Hydrologie	Hydrologie_11
Résultat de mesure Valeur numérique	NO3+NO2	Min Max]0.099 ; 800[Hydrologie	Hydrologie_23
Résultat de mesure Valeur d'incertitude	NH4	Empty		Hydrologie	Hydrologie_11
Résultat de mesure Valeur numérique	NH4	Min Max]0.049 ; 40[Hydrologie	Hydrologie_22
Résultat de mesure Valeur numérique	NH4	Min Max]0.14 ; 4[controle_LERBL	controle_LERBL_4
Résultat de mesure Valeur d'incertitude	PO4	Empty		Hydrologie	Hydrologie_11
Résultat de mesure Valeur numérique	PO4	Min Max]0.04 ; 2[controle_LERBL	controle_LERBL_7
Résultat de mesure Valeur numérique	PO4	Min Max]0.019 ; 15[Hydrologie	Hydrologie_26
Résultat de mesure Valeur d'incertitude	SIOH	Empty		Hydrologie	Hydrologie_11
Résultat de mesure Valeur numérique	SIOH	Min Max]0.09 ; 15[controle_LERBL	controle_LERBL_8
Résultat de mesure Valeur numérique	SIOH	Min Max]0.099 ; 400[Hydrologie	Hydrologie_27
Résultat de mesure Valeur d'incertitude	CHLOROA	Empty		Hydrologie	Hydrologie_11
Résultat de mesure Valeur numérique	CHLOROA	Min Max]0.049 ; 70[Hydrologie	Hydrologie_20
Résultat de mesure Valeur d'incertitude	PHEO	Empty		Hydrologie	Hydrologie_11
Résultat de mesure Valeur numérique	PHEO	Min Max]0.049 ; 5[Hydrologie	Hydrologie_21
Résultat de mesure Valeur d'incertitude	OXYGENE	Empty		Hydrologie	Hydrologie_11
Résultat de mesure Valeur numérique	OXYGENE	Min Max]0.099 ; 20[Hydrologie	Hydrologie_19
Résultat de mesure Valeur d'incertitude	TURB	Empty		Hydrologie	Hydrologie_11
Résultat de mesure Valeur numérique	TURB	Min Max]0.049 ; 70[Hydrologie	Hydrologie_17
Résultat de mesure Valeur d'incertitude	TURB-FNU	Empty		Hydrologie	Hydrologie_11
Résultat de mesure Valeur numérique	TURB-FNU	Min Max]0.049 ; 70[Hydrologie	Hydrologie_18

Ces règles de contrôle doivent limiter les anomalies lors de la qualification automatique. Elles ne peuvent pas en garantir totalement l'absence puisqu'elles ne s'appliquent qu'à la saisie manuelle dans l'application Quadrige². **Tout import de fichier dans les formats standards nationaux (Quadrilabo par exemple) échappe à ces contrôles.**

2.2 Microbiologie

2.2.1 Définition de la thématique

Tableau 7 : Critères de définition de la thématique Microbiologie pour la qualification

Programmes	Paramètres
REMIS	ECOLI (mesure)
REMIC	ENTEROCO (mesure)
REMIE1	SALMOD (mesure)
REMIE2	SALMOI (taxons)
REMI-SURV	STREPTOC (mesure)
CMIC	VIBRIOD (mesure)
	VIBRIOI (taxons)
	VHA
	NoV
	Les résultats sur supports bivalve, gastéropode et échinoderme seront qualifiés les premiers Les résultats sur le support Eau seront qualifiés après (les anomalies concernant ces données ne sont pas encore décrites)

2.2.2 Intervenants identifiés

Tableau 8 : Liste des intervenants identifiés pour la qualification des données microbiologiques

Labo	Noms	Rôle
LER BO	Sylviane Boulben	Producteur de données
LER BN	Julia Penot	Producteur de données
LER BN	Julien Chevé	Producteur de données
LER BN	Aurore Lejolivet	Producteur de données
LER LR	Mathilde Rousselet	Producteur de données
LER LR	Emmanuelle Roque	Chef de station
LER LR	Delphine Granger	
LER LR	Anais Crottier	
LER MPL TM	Jean-Pierre Allenou	Producteur de données
LER MPL TM	Soizick Manach	Producteur de données
LER N	Laure Lamort	Producteur de données
LER PAC TL	Marc Bouchoucha	Producteur de données
LER PAC TL	Francoise Marco Miralles	Producteur de données
LER PAC CO	Yoann Baldi	Producteur de données
LER PC LR	Annick Derrien	Producteur de données
LER BL	Françoise Verin	Producteur de données
LER MPL NT	Gilles Ratiskol	Producteur de données

Labo	Noms	Rôle
LER MPL NT	Sandrine Bonnetot	Producteur de données
LER AR	Isabelle Auby	Producteur de données
LER AR	Claire Barbier	Producteur de données
LER PC LT	Dimitri Morin	Producteur de données
LER PC LT	Cyrielle Montaubin	Producteur de données
LER PC LT	Mathilde Noyer	Producteur de données
EMP	Marie Nedellec	Qualificateur

2.2.3 Règles de qualification

2.2.3.1 Qualification automatique

Le Tableau 9 liste les règles de qualification définies pour les données microbiologiques. Certains passages surlignés sont des remarques ou questions posées par les qualificateurs et auxquelles il faut répondre.

Tableau 9 : Liste des règles de qualification automatique pour la microbiologie.

Pack	Niveau de priorité	N°Anomalie	Description
HPAC	0	HPAC-001	Vérifier les PSFM utilisés dans le cadre du REMI : est-ce que les paramètres, les supports, les fractions, les méthodes et les PSFM dans leur ensemble sont cohérents par rapport aux analyses effectuées dans le cadre du REMI?
HPAC	0	HPAC-002	« Conflits thématiques » : 1/ Recherche des programmes, autres que ceux concernés par le REMI, rattachés aux données (passages, prélèvements, échantillons, résultats) à qualifier. 2/ Recherche des autres résultats rattachés aux passages concernés par le REMI
HPAC	0	HPAC-003	Rechercher les données pour lesquelles les champs normalement vides pour des données microbiologiques ne le sont pas.
HPAC	1	HPAC-004	Rechercher les données pour lesquelles le taxon support est vide ou incohérent par rapport au support de l'échantillon
HPAC	1	HPAC-005	1/ Vérifier la cohérence des niveaux de saisie des résultats dans les stratégies. 2/ Rechercher les résultats de mesure (ECOLI, ENTEROCO, SALMOD, STREPTOC, VIBRIOD, VHA, NoV) et résultats taxons (SALMOI, VIBRIOI) saisis à un autre niveau que celui prévu dans la stratégie.
HPAC	2	HPAC-006	Rechercher les doublons de résultats de mesures : plusieurs résultats de mesure pour un même Lieu / Date / Taxon support / Paramètre / Méthode
HPAC	2	HPAC-007	Rechercher les doublons de résultats taxons : plusieurs résultats taxons pour un même Lieu / Date / Taxon support / Paramètre / Méthode / Taxon
HPAC	2	HPAC-008	Rechercher les doublons de méthodes : plusieurs méthodes pour un même Lieu / Date / Taxon support / Paramètre / Support / Fraction
HPAC	2	HPAC-009	Vérifier la cohérence des combinaisons Lieu de surveillance - Saisisseur passage

Pack	Niveau de priorité	N°Anomalie	Description
HPAC	3	HPAC-010	Les données de surveillance régulière et surveillance en alerte sont différenciées par le type d'évènement qui y est rattaché ou non. Cette anomalie, comportant 4 parties, permet de rechercher uniquement pour les passages portant des résultats ECOLI, les : 1) Evènements non rattachés à un passage et concernant le REMI. 2) Evènements rattachés à un passage non REMIC ni REMI-SURV et concernant le REMI. 3) Evènements rattachés à un passage REMIC ou REMI-SURV et ne concernant pas le REMI. (pas forcément pertinent...) 4) Plusieurs évènements (tous types Q ² confondus) rattachés à un même passage
HPAC	4	HPAC-011	Rechercher les passages portant des résultats ECOLI seulement et dont le saisisseur de l'évènement est différent du saisisseur du passage
HPAC	4	HPAC-012	Vérifier la cohérence des préleveurs des stratégies (vérifier que les préleveurs par défaut sont exacts, et notamment que les préleveurs sont tous remplis (pour REMI-SURV seulement) pourquoi ? c'était surtout pour l'historique que cela avait un intérêt...cad REMIS er REMIC, pour REMI Surv je ne pense pas que ça présente forcément un intérêt et identiques pour toutes les stratégies applicables au même moment sur un même lieu au sein des différents programmes)
HPAC	4	HPAC-013	Vérifier la cohérence des analystes des stratégies (vérifier que les analystes par défaut sont exacts, qu'ils sont tous remplis (pour REMI-SURV seulement) et identiques pour toutes les stratégies applicables au même moment sur un même lieu - PSFM au sein des différents programmes) (je ne sais pas pourquoi il y a ça, je ne suis pas sûre que cela présente un intérêt pour REMI-surv...)
HPAC	4	HPAC-014	Recherche les résultats non validés (cela implique de contrôler et valider les résultats, ou de les supprimer ou de les laisser avec ce statut ; ainsi que toute leur chaîne de saisie : passage / prélèvement / échantillon).
PASS	1	PASS-001	Rechercher les passages dont la date de contrôle ou date de validation des passages est vide (cette anomalie est présente uniquement pour que les passages soient qualifiés. En effet, sans anomalie, pas de qualification).
PREL	1	PREL-001	Rechercher les prélèvements dont le préleveur saisi est différent de celui prévu dans la stratégie
PREL	1	PREL-002	Rechercher les prélèvements dont le support de l'échantillon est « Bivalve », « Gastéropode » ou « Echinoderme » et l'engin est différent de "Main", "Grattoir", "Pelle", "Dragues (non spécifiées)", "Tellinier" (engins définis dans les consignes de saisies)
PREL	2	PREL-003	Rechercher les prélèvements dont le support est « Bivalve », « Gastéropode » ou « Echinoderme » et le niveau est différent de « Emergé », « Mi-profondeur », « Surface(0-1m) », « Fond », "Fond/sonde-1m" (niveaux définis dans les consignes de saisies).
PREL	3	PREL-004	Vérifier que les combinaisons engin / niveau / immersion / support de l'échantillon sont cohérentes
PREL	3	PREL-005	Rechercher les prélèvements dont le support est « Bivalve », « Gastéropode » ou « Echinoderme », et dont l'unité immersion est différente de "m" (c'était aussi pour les données historiques, avec Q ² je pense que cette anomalie n'a plus de sens, elle peut être mise en sommeil pour les données récentes, d'autant que les données avaient fait l'objet d'une correction globale à ce niveau)

Pack	Niveau de priorité	N°Anomalie	Description
ECHA	1	ECH-0001	Rechercher les passages dont la date de contrôle ou date de validation des échantillons est vide (cette anomalie est présente uniquement pour que les échantillons soient qualifiés. En effet, sans anomalie, pas de qualification). Remarque : la seule autre anomalie échantillon est la HPAC-004. (du coup est ce une HPAC ?, + il faudrait rajouter comme anomalie le cas où le taxon support est incomplet (pas l'espèce) : cerastoderma, mytilus)
RESU	1	RES-001	Rechercher les résultats dont l'analyste saisi est différent de celui défini dans la stratégie
RESU	1	RES-002	Rechercher les résultats dont la méthode saisie ne correspond pas à la méthode normalement utilisée à l'époque. Cette anomalie concerne le paramètre ECOLI, les programmes REMIS, REMIC et REMI-SURV et les méthodes d'analyse utilisées dans le cadre de la stratégie régulière.
RESU	2	RES-003	Rechercher les résultats dont la valeur du résultat < à la valeur seuil de détection de la méthode utilisée
RESU	2	RES-004	Rechercher les résultats ayant avec libellé précision alors que la valeur est différente du seuil de détection du PSFM de la méthode
RESU	3	RES-005	Rechercher les résultats dont la valeur du résultat est supérieure 100 000 E.coli/CLI et concernant le paramètre ECOLI et les supports "bivalve" ou "Echinoderme" ou "Gastéropode"
RESU	3	RES-006	Rechercher les résultats ayant paramètre = ECOLI, support = Echinoderme ou Gastéropode, méthode = Colimétrie par conductance-métrie (méthode inadaptée à ces supports : à qualifier « Faux ») oui, mais pourquoi le mettre ici ?
RESU	4	RES-007	Rechercher les résultats ECOLI et STREPTOC dont le support "bivalve" ou "Echinoderme" ou "Gastéropode", la date de passage est postérieure au 01/12/2002, et la valeur du résultat ne respecte pas la règle d'arrondi (que 2 chiffres significatifs)
RESU	4	RES-008	Pour les résultats STREPTOC avec la méthode « Dénombrement MPN Rothe Litsky » car la méthode mal décrite (donc résultats à qualifier douteux)

2.2.3.2 Qualification experte

A l'étape de définition des anomalies à rechercher, deux d'entre elles ont été classées en qualification « experte » dans la mesure où elles consistent à vérifier la cohérence d'une donnée par rapport à d'autres données.

- Résultat SALMOI sur un échantillon portant un résultat SALMOD avec valeur qualitative "Absence"
- Anomalie dans la recherche des salmonelles : échantillon rattaché à la surveillance régulière au lieu d'être rattaché à un événement.

2.2.4 Règles de contrôle Quadrige

La plupart des règles de contrôle appliquées aux données microbiologiques appartiennent à la liste « REMI-SURV » mise en place avec la qualification automatique des données. Les contrôles sur les valeurs du paramètre ECOLI sont complexes : 3 règles de contrôle peuvent s'appliquer simultanément afin de vérifier chacune des plages de valeurs possibles.

Tableau 10 : Liste des règles de contrôle appliquées aux données microbiologiques dans Quadrigé.

Attribut contrôlé	Paramètre	Méthode	Fonction	Contrôle effectué	Bloquante	Liste	Règle
Prélèvement Niveau			In	1,3,2,60000000		REMI-SURV	REMI-SURV_9
Echantillon Support			In	7,11	✓	REMI-SURV	REMI-SURV_6
Echantillon Taxon support			Not empty		✓	REMI-SURV	REMI-SURV_2
Résultat de mesure Précision	ECOLI		In	« < valeur »		contrôle_LERMPL	controle_LERMPL_3
Résultat de mesure Valeur numérique	ECOLI	Colimétrie E. coli NPP 3x5T ou 5x5T + Gélose - nb/hg	Min Max]17 ; 10000000[✓	CMIC	CMIC_1
Résultat de mesure Valeur numérique	ECOLI	Colimétrie E. coli - BacTrac 4300 (tech. validée) - nb/hg	Min Max				
Résultat de mesure Valeur numérique	ECOLI	Colimétrie E. coli NPP 3x5T ou 5x5T + Gélose - nb/hg	Min Max]0 ; 100 000[CMIC	CMIC_3
Résultat de mesure Valeur numérique	ECOLI	Colimétrie E. coli - BacTrac 4300 (tech. validée) - nb/hg	Min Max]66 ; 10 000 000[✓	CMIC	CMIC_4
Résultat de mesure Valeur numérique	ECOLI	Colimétrie E. coli - BacTrac 4300 (tech. validée) - nb/hg	Min Max]130 ; 10 000 000[CMIC	CMIC_5
Résultat de mesure Valeur numérique	ECOLI	Colimétrie E. coli NPP 3x5T ou 5x5T + Gélose - nb/hg	Min Max]18 ; 10 000 000[CMIC	CMIC_11
Résultat de mesure Valeur numérique	ECOLI	Colimétrie E. coli NPP 3x5T ou 5x5T + Gélose - nb/hg	Min Max]17 ; 10 000 000[✓	REMIE1	REMIE1_1
Résultat de mesure Valeur numérique	ECOLI	Colimétrie E. coli NPP 3x5T ou 5x5T + Gélose - nb/hg	Min Max]18 ; 10 000 000[REMIE1	REMIE1_11
Résultat de mesure Valeur numérique	ECOLI	Colimétrie E. coli - BacTrac 4300 (tech. validée) - nb/hg	Min Max				
Résultat de mesure Valeur numérique	ECOLI	Colimétrie E. coli NPP 3x5T ou 5x5T + Gélose - nb/hg	Min Max]0 ; 100 000[REMIE1	REMIE1_3
Résultat de mesure Valeur numérique	ECOLI	Colimétrie E. coli - BacTrac 4300 (tech. validée) - nb/hg	Min Max]66 ; 10 000 000[✓	REMIE1	REMIE1_4
Résultat de mesure Valeur numérique	ECOLI	Colimétrie E. coli - BacTrac 4300 (tech. validée) - nb/hg	Min Max]130 ; 10 000 000[REMIE1	REMIE1_5
Résultat de mesure Valeur numérique	ECOLI	Colimétrie E. coli NPP 3x5T ou 5x5T + Gélose - nb/hg	Min Max]17 ; 10 000 000[✓	REMI-SURV	REMI-SURV_1
Résultat de mesure Valeur numérique	ECOLI	Colimétrie E. coli NPP 3x5T ou 5x5T + Gélose - nb/hg	Min Max]18 ; 10 000 000[REMI-SURV	REMI-SURV_11
Résultat de mesure Valeur numérique	ECOLI	Colimétrie E. coli - BacTrac 4300 (tech. validée) - nb/hg	Min Max				
Résultat de mesure Valeur numérique	ECOLI	Colimétrie E. coli NPP 3x5T ou 5x5T + Gélose - nb/hg	Min Max]0 ; 100 000[REMI-SURV	REMI-SURV_3
Résultat de mesure Valeur numérique	ECOLI	Colimétrie E. coli - BacTrac 4300 (tech. validée) - nb/hg	Min Max]66 ; 10 000 000[✓	REMI-SURV	REMI-SURV_4
Résultat de mesure Valeur numérique	ECOLI	Colimétrie E. coli - BacTrac 4300 (tech. validée) - nb/hg	Min Max]129 ; 10 000 000[REMI-SURV	REMI-SURV_5

2.3 Chimie

2.3.1 Définition de la thématique

Tableau 11 : Critères de définition de la thématique Chimie pour la qualification

Programmes	Paramètres
RNOSED (contaminants chimiques dans les sédiments) : uniquement à partir de 1993 car c'est à cette date que l'Ifremer a pris la maîtrise de cette surveillance RNOMV (contaminants chimiques dans la matière vivante) RNOPHY (suivi de l'IMPOSEX)	Tous les paramètres analysés dans ces programmes <i>La liste est-elle toujours exhaustive?</i>

2.3.2 Intervenants identifiés

Le seul intervenant a été Didier Claisse (RBE/BE), parti à la retraite en 2015. Jean-François Chiffolleau a repris la coordination du réseau ROCCH, puis Anne Grouhel. Aucune qualification n'a été entreprise depuis le départ de D Claisse.

2.3.3 Règles de qualification

2.3.3.1 Qualification automatique

Tableau 12 : Règles de qualification automatique des données Chimie.

Pack	N°anomalie	Niveau priorité	Description
HPAC	HPAC-001	0	Vérification de la liste des PSFMs (état actif / gelé, cohérence paramètre – support – fraction – méthode)
HPAC	HPAC-002	0	Conflits thématiques : recherche des autres programmes liés aux données RNO, et des résultats d'autres thématiques rattachés aux données in situ RNO
HPAC	HPAC-003	0	Champs vides : vérification que les champs qui n'ont pas lieu d'être remplis dans le cadre du RNO sont bien vides. Ont été contrôlés : <ul style="list-style-type: none"> - Passages : observations terrain, habitat observé, zone de destination dragage, nombre d'individus - Prélèvements : Lot, pour RNOSED et RNOPHY : taille de prélèvement et nombre d'individus - Echantillons RNOSED : taxon et groupe de taxon support, taille de l'échantillon et nombre d'individus - Echantillons RNOPHY : groupe de taxon et taille de l'échantillon
HPAC	HPAC-004	1	Support-taxon : vérification de la cohérence des supports d'échantillon avec le taxon support associé. RNOMV et RNOPHY seulement (pas RNOSED)
HPAC	HPAC-005	2	Niveau de saisie : vérification que toutes les PSMFs des stratégies RNO ont le même niveau de saisie (échantillon en l'occurrence), puis vérification que tous les résultats sont saisis au niveau prévu dans la stratégie
HPAC	HPAC-006	3	Doublons de passages : vérification qu'il existe un seul passage par lieu – date.

Pack	N°anomalie	Niveau priorité	Description
HPAC	HPAC-007	3	Doublons de résultats : vérification qu'il existe un seul résultat de mesure par : <ul style="list-style-type: none"> - RNOPHY : lieu – date – taxon – paramètre – méthode – n° d'individu (depuis 2015). - RNOMV : lieu – date – taxon – n° d'individu – paramètre – méthode (1 analyse par individu, donc ce ne sont pas des doublons). - RNOSED : lieu – date – niveau (horizon) – paramètre – fraction – méthode
HPAC	RNOSED-HPAC-007	4	Campagnes RNOSED rattachées à d'autres programmes que le RNOSED, puis cohérence géographique des lieux des passages liés à chaque campagne.
HPAC	RNOPHY-HPAC-008	4	Nombre de résultats saisis : <ol style="list-style-type: none"> 1) Nb résultats par paramètre > nombre d'individus de l'échantillon (depuis 2015) 2) Nb résultats INDVSEX différent du nombre d'individus de l'échantillon
HPAC	HPAC-008	5	Données non validées
PASS	PASS-001	1	Heure de passage = 00:00:00 (à supprimer)
PASS	PASS-002	2	Unité de sonde du passage <> « m »
PASS	RNOSED-PASS-003	2	Sonde nulle ou =0 ou >= 300m
PASS	RNOSED-PASS-004	3	Passages ayant une géométrie réelle (pas habituel pour le RNOSED)
PREL	PREL-001	1	Heure de prélèvement = 00:00:00 (à supprimer)
PREL	PREL-002	1	Unité d'immersion « m » (l'unité est « cm » pour RNOSED puisqu'il s'agit d'horizons sédimentaires).
PREL	PREL-003	2	Combinaison engin – niveau – immersion – support : vérification de la cohérence des 4 valeurs
ECHA	ECHA-001	1	Echantillon non contrôlé / non validé Pas recherché pour RNOSED (aucune anomalie échantillon)
ECHA	RNOPHY-ECHA-004	1	Echantillon dont le nombre d'individus est soit vide, soit supérieur à 40 individus (depuis 2015)
RESU	RNOMV-RES-001	1	Résultat sur le paramètre MS% dont la valeur numérique est > 33.
RESU	RNOMV-RES-002	1	Résultats sur support Poisson : paramètres avec moins de résultats que le nombre d'individus de l'échantillon (résultats manquants)
RESU	RNOPHY-RES-001	1	Résultats VDS et sexe des individus : pour les individus femelle, le résultat VDS doit être présent. Pour les mâles il ne doit pas y avoir de résultat VDS.
RESU	RNOPHY-RES-002	1	Résultats dont le nombre de décimales est >2
RESU	RNOPHY-RES-003	1	Longueur de pénis hors des bornes [0 ; 6,5 cm]

2.3.3.2 Qualification experte

Pour RNOPHY (Imposex) il n'y a pas de qualification experte, D. Claisse ayant considéré que l'expertise avait déjà été réalisée avant saisie des données par Martial Huet (producteur de données et « inventeur » de cette méthode).

Pour RNOMV, la qualification experte consiste en des analyses des séries temporelles via des DLM². Leur programmation a été réalisée par l'équipe de biostatisticiens d'Ifremer/ODE/VIGIES (D. Soudant) selon les spécifications de D. Claisse (Figure 1). Il n'existe pas de document de spécification formel. En revanche il existe un « manuel utilisateur » destiné à l'équipe VIGIES et qui décrit comment faire tourner les programmes de cette qualification.

En plus des graphiques, une liste des données représentées est éditée sous forme de tableur permettant au qualificateur d'indiquer le niveau de qualification final des données. Ce fichier est ensuite utilisé par la cellule Quadrige pour inscrire cette qualification dans la base de données.

Pour RNOSED, il n'y a pas de DLM car pas de séries temporelles longues (au mieux 4 campagnes = série de mesures par façade). De plus, d'une campagne à l'autre, les lieux peuvent être légèrement différents. La représentation graphique choisie est expliquée dans un document de spécifications rédigé par D. Claisse (Annexe). Les programmes R correspondants n'ont pas été développés et D. Claisse est parti à la retraite avant de réaliser cette qualification experte.

² DLM : Dynamic Linear Model

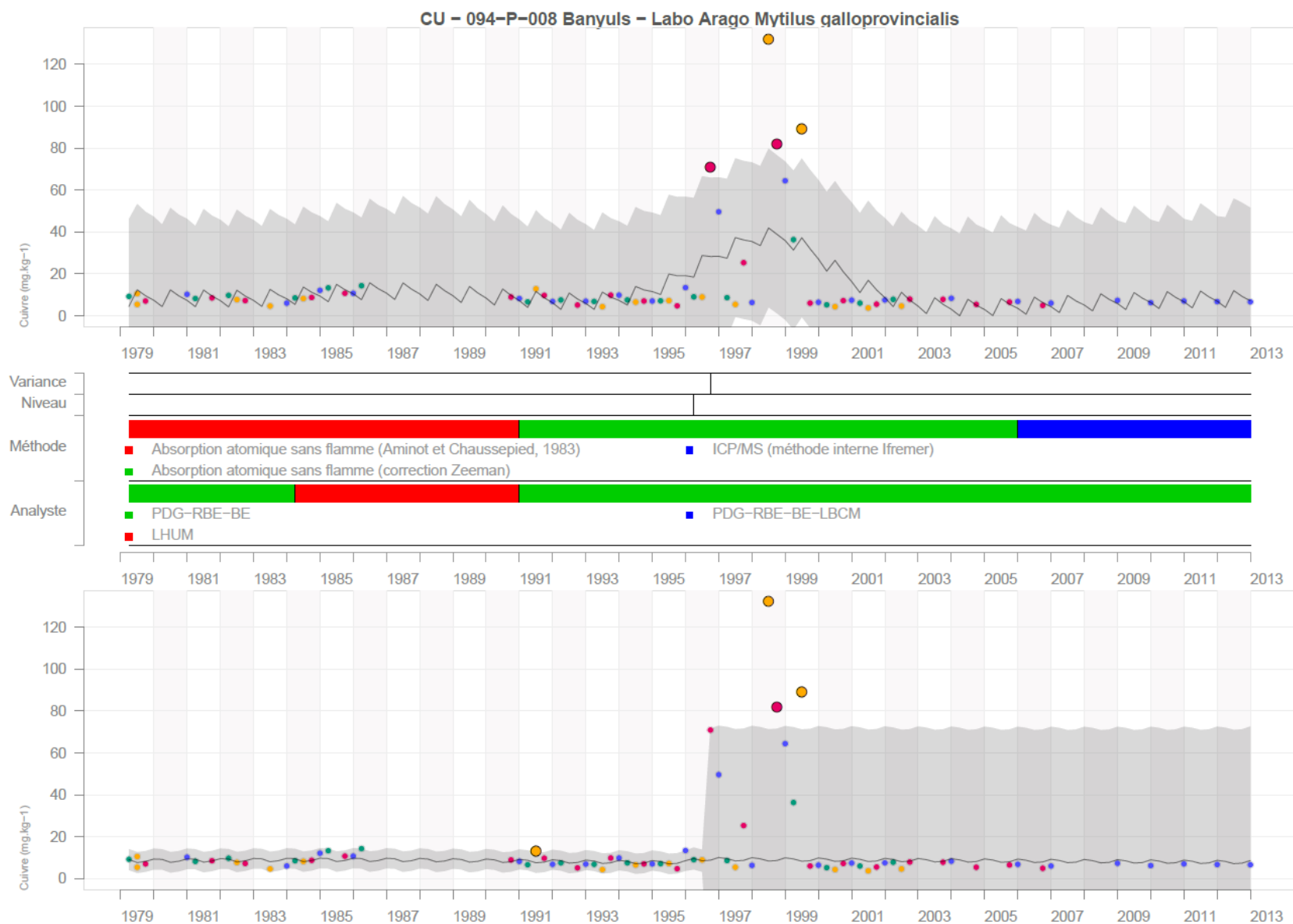


Figure 1 : Exemple de sortie graphique pour la qualification experte RNOMV (DLM avec changement de niveau de variance).

2.3.4 Règles de contrôle Quadrigé

Aucune règle de contrôle nationale n'a été définie dans Quadrigé. Elles permettraient pourtant de s'assurer de l'absence de plusieurs anomalies. Un travail spécifique pourrait être mené avec la coordination du ROCCH et la cellule Quadrigé.

3 Bilan général de la qualification (septembre 2017)

3.1 Volumétrie

Des résultats sont qualifiés pour toutes les années de données de Quadrigé (Figure 2).

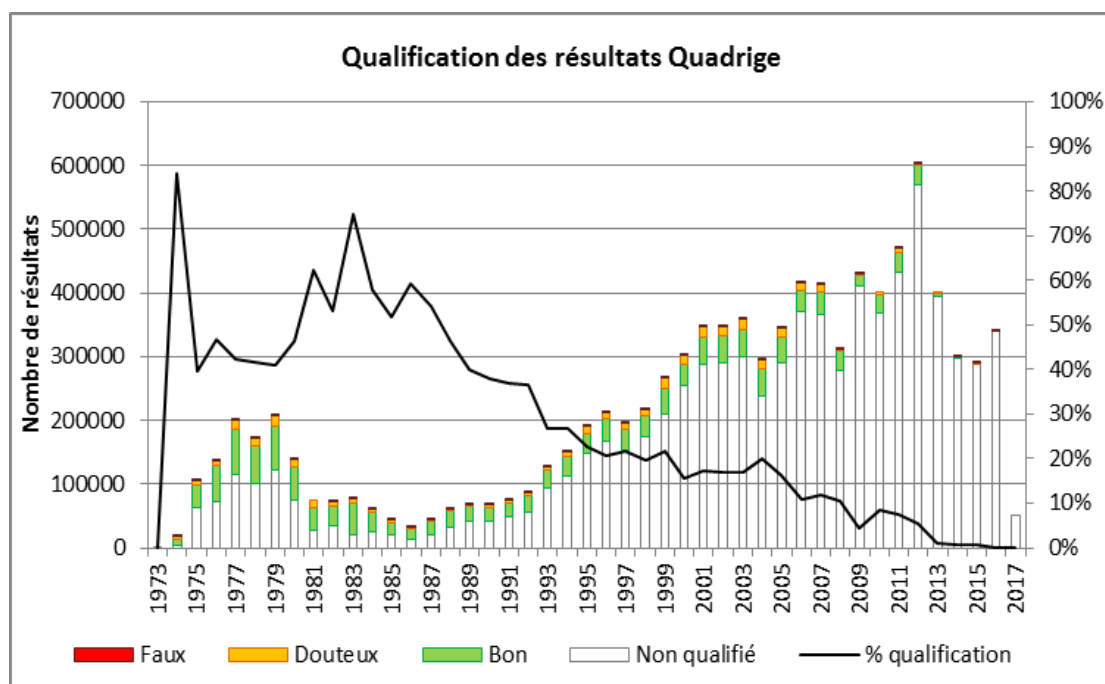


Figure 2 : Avancement de la qualification des résultats de Quadrigé, tous types de résultats confondus (mesures, dénombrements et fichiers).

Le taux de qualification des données dépasse les 40% pour les données anciennes (avant 1990) et diminue jusqu'à moins de 10% pour les résultats des 10 dernières années (Figure 2). Ceci est notamment dû à l'augmentation du volume de données intégrées dans la base Quadrigé (les données produites sont plus nombreuses mais le volume de données qualifiées reste constant jusqu'en 2012).

Par ailleurs la mise en œuvre de la qualification dite « automatique » des données a nécessité un effort particulier : afin de repérer le plus exhaustivement possible les anomalies potentielles, l'examen des jeux de données a nécessité de couvrir tout l'historique.

De plus, les qualificateurs qui ont expertisé ces données ont coordonné ces réseaux de surveillance depuis leur mise en place au début des années 80. Ils avaient donc une parfaite connaissance des données historiques, avaient déjà identifié des anomalies dans les données et savaient comment les corriger et/ou qualifier. Pour pérenniser ces savoirs avant le départ en retraite des experts, un effort particulier de qualification experte a été réalisé sur la thématique Chimie.

Depuis 2014, les ressources humaines disponibles pour qualifier les données se sont raréfiées, tant chez les qualificateurs qu'à la cellule d'exploitation Quadrigé, d'autres projets montant en charge. Le processus de qualification dite « automatique » n'a plus tourné en routine.

3.2 Niveaux de qualité des données

Pour rappel quatre niveaux de qualité sont possibles dans Quadrigé :

- 0 = Non qualifié
- 1 = Bon
- 3 = Douteux
- 4 = Faux.

Ces niveaux sont issus d'une liste internationale³.

Les causes de qualification à Douteux ou Faux sont très diverses. Il est toutefois possible de les classer en 6 types possibles (Figure 3).

Le **manque de métadonnées** représente plus de la moitié des causes. Cela concerne principalement les données historiques (Figure 4). En effet, si dans Quadrigé² certains champs sont obligatoires à la saisie, les fichiers de type tableur utilisés pour saisir des données ne contiennent souvent que les informations sur les paramètres analysés et les valeurs mesurées. Les méthodologies sont souvent décrites dans des rapports à part. Dans le cas des données historiques il est souvent difficile de retrouver ces métadonnées d'autant que les personnes ont pu quitter les laboratoires analystes impliqués. Ce type de données a toutefois été intégré dans Quadrigé soit pour répondre aux besoins de l'appui aux politiques publiques soit pour pérenniser des données qui pourraient être utiles dans le cadre de certaines études scientifiques.

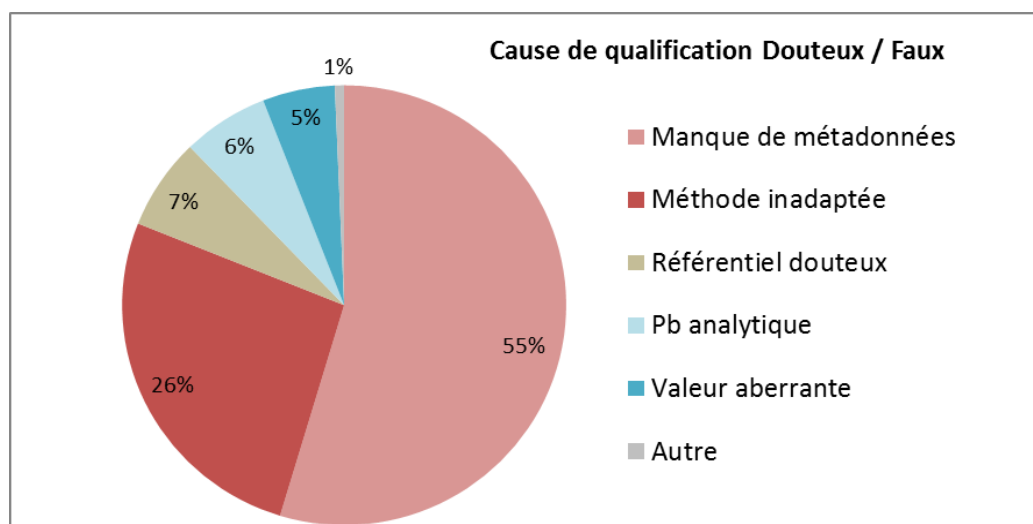


Figure 3 : Causes de qualification des résultats à "Douteux" ou "Faux".

Un quart des données douteuses ou fausses le sont pour cause d'une **méthode inadaptée** soit aux eaux marines, soit au paramètre recherché, soit au support analysé.

Dans 7% des cas c'est le **référentiel qui n'est pas cohérent** : les données ont été saisies sur un PSFMU qui n'a pas de sens et/ou ne correspond aux analyses qui ont été faites. Les travaux de qualification ont permis de corriger le référentiel pour les thématiques « hydrologie », « microbiologie » et « chimie ».

6% des qualifications douteux/faux relèvent d'un **problème au cours du processus d'acquisition de la donnée** : échantillons mal conservés, dérive des instruments de mesure, etc. Seuls 5% des niveaux Douteux et Faux sont dus à des valeurs numériques douteuses voire aberrantes (potentiels « outliers » détectés en qualification experte).

³ Paris. Intergovernmental Oceanographic Commission of UNESCO. 2013. Ocean Data Standards, Vol.3 : Recommendation for a Quality Flag Scheme for the Exchange of Oceanographic and Marine Meteorological Data. (IOC Manuals and Guides, 54, Vol. 3.) 12 pp. (English.) (IOC/2013/MG/54-3).

Si l'on considère uniquement les causes liées aux problèmes analytiques, aux valeurs extrêmes et aux autres problèmes, **le pourcentage de données qualifiées Douteux ou Faux ne dépasse pas les 5%** (Figure 4).

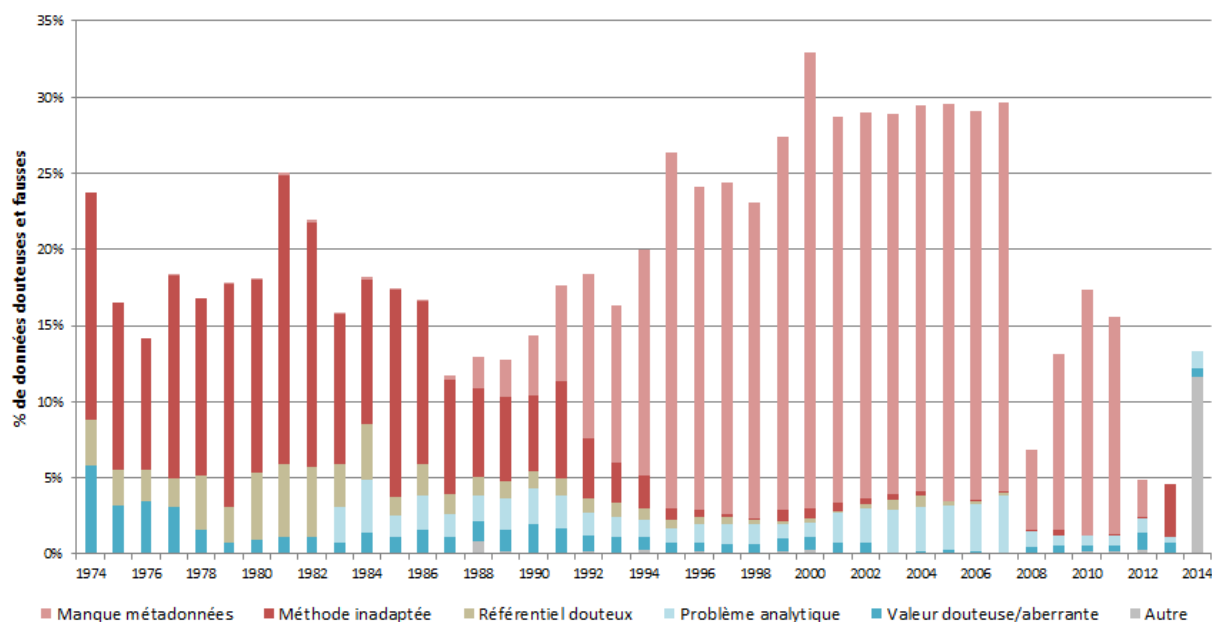


Figure 4 : Pourcentage des données fausses et douteuses par année de données (année des passages).

Remarque sur les commentaires de qualification :

Les commentaires de qualification permettent de préciser l’expertise dont est issu le niveau de qualité indiqué, et, dans le cas des données douteuses et fausses, permet d’expliquer les raisons de ce déclassement.

Or, dans les données qualifiées, certains commentaires ne correspondent pas au niveau indiqué, doivent être précisés ou impliquent une correction des données à réaliser. Par exemple :

BON, commentaire : « Archives non disponibles - Valeur numérique douteuse/fausse - Qualification automatique hydrologie »

➔ si la valeur est douteuse, alors le résultat ne doit pas être qualifié bon.

DOUTEUX, commentaire : « Qualification experte D. Claisse 2014 »

➔ cela n’explique pas ce qui est douteux dans la donnée.

DOUTEUX, commentaire : « Qualification automatique hydrologie. Prélèvement qualifié provisoirement à douteux, en attente de re-saisi de résultats sur échantillon. - Pas d’anomalie détectée lors de la qualification automatique hydrologie »

➔ Le prélèvement est qualifié « Douteux » alors que si les données liées ont été corrigées (re-saisies dans le cas présent) alors il peut être qualifié à « Bon ».

3.3 Les intervenants

Le processus de qualification nécessite des qualificateurs (Tableau 13) et des correspondants chez les producteurs de données (cf. listes dans les chapitres thématiques).

Tableau 13 : Liste des qualificateurs par thématique (pour la qualification automatique déjà opérationnelle).

Thématique	Période	Qualificateur
Hydrologie	Jusqu’en 2015	Catherine BELIN Anne DANIEL-SCUILLER
	A partir de 2016	Nadine NEAUD-MASSON

Thématique	Période	Qualificateur
		Maud LEMOINE Anne DANIEL-SCUILLER
Chimie	Jusqu'en 2015	Didier CLAISSE
	A partir de 2016	Anne GROUHEL
Microbiologie	Jusqu'en 2013	Isabelle AMOUROUX
	Depuis 2014	Jean-Côme PIQUET

La liste des producteurs de données (cf. tableaux par thématique) est à remettre à jour car :

- de nombreux intervenants ont changé de laboratoire (notamment en raison de la fermeture de certaines stations Ifremer entraînant des mouvements de personnel),
- de nouveaux intervenants doivent être identifiés pour les nouvelles données intégrées dans Q² (notamment pour la DCE dans les DOMs).

3.4 Les outils informatiques

Les qualifications dites « ponctuelles » sont effectuées par la cellule Quadrigé via des requêtes SQL exécutées manuellement.

Un outil de qualification existe dans Quadrigé² mais il n'a pas été utilisé depuis plusieurs années pour cause de bugs irrésolus.

La qualification dite « automatique » s'appuie sur des scripts développés sous R par plusieurs CDDs successifs. Les programmes ont été progressivement améliorés et débogués. Ils sont aujourd'hui opérationnels pour les thématiques « Hydrologie » et « Microbiologie ». Leur fonctionnement reste complexe et parfois difficile à déboguer lorsqu'un nouveau cas de données non prévu survient. Une cartographie des programmes indiquant quelle fonction est utilisée par quel programme et donnant les dépendances entre programmes devrait aider à leur maintenance.

Remarque : les programmes de qualification automatique n'ont pas tourné jusqu'au bout pour la microbiologie (la qualification rapide des Bons a été réalisée jusqu'aux résultats, mais la qualification des données en anomalie s'est arrêtée aux échantillons).

La qualification automatique des données Chimie est effectuée par requêtes SQL car elle a été testée sous cette forme et jamais « industrialisée » après le départ en retraite de D. Claisse. Les anomalies sont toutefois très simples à rechercher et les requêtes SQL sont sauvegardées comme les programmes de l'hydrologie et de la microbiologie.

3.5 Hydrologie

Le processus de qualification dite « automatique » est finalisé et peut être lancé en routine. La documentation associée doit être actualisée et finalisée :

- Manuel pour la cellule Quadrigé
- Procédure de qualification Hydrologie : il décrit en détail toutes les étapes de la qualification. Il fait office de document de spécifications.
- Manuel du qualificateur : il décrit comment le qualificateur arbitre les réponses des producteurs de données et guide le qualificateur pour l'attribution des niveaux de qualité.

Concernant les données, l'avancement de la qualification est présenté sur les Figure 5 et Figure 6.

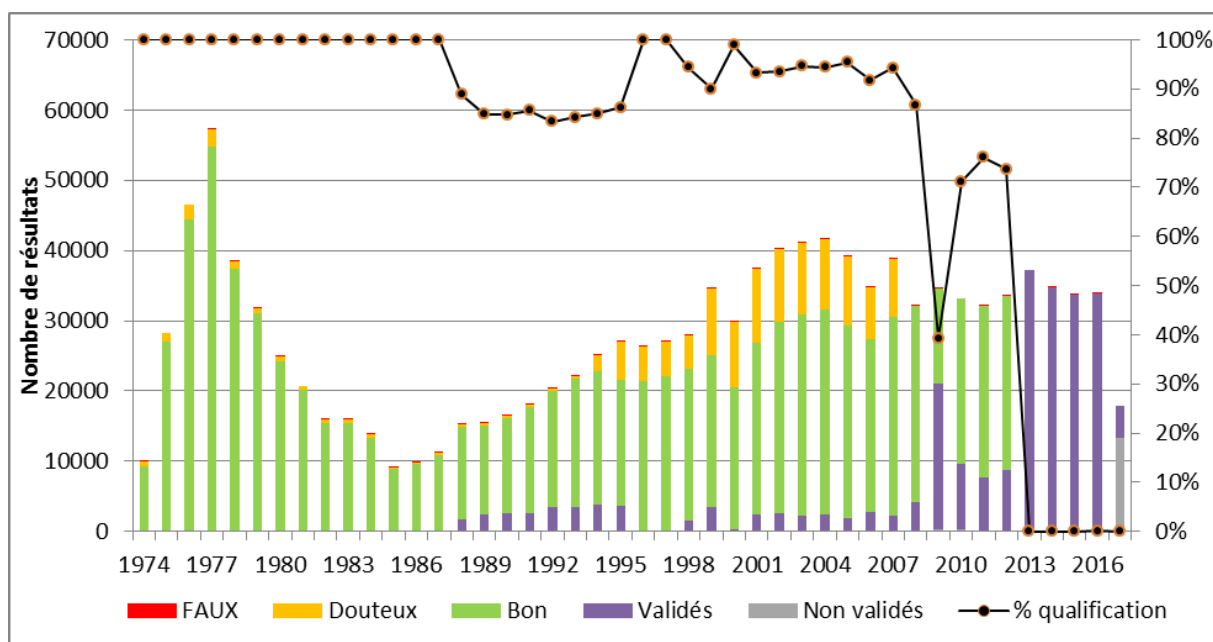


Figure 5 : Avancement de la qualification des résultats hydrologie

Il reste des données à qualifier sur la période 1988-2012. Depuis 2013, quelques résultats ont été qualifiés mais ce sont des qualifications ponctuelles (les données in situ portant ces résultats ne sont pas qualifiées).

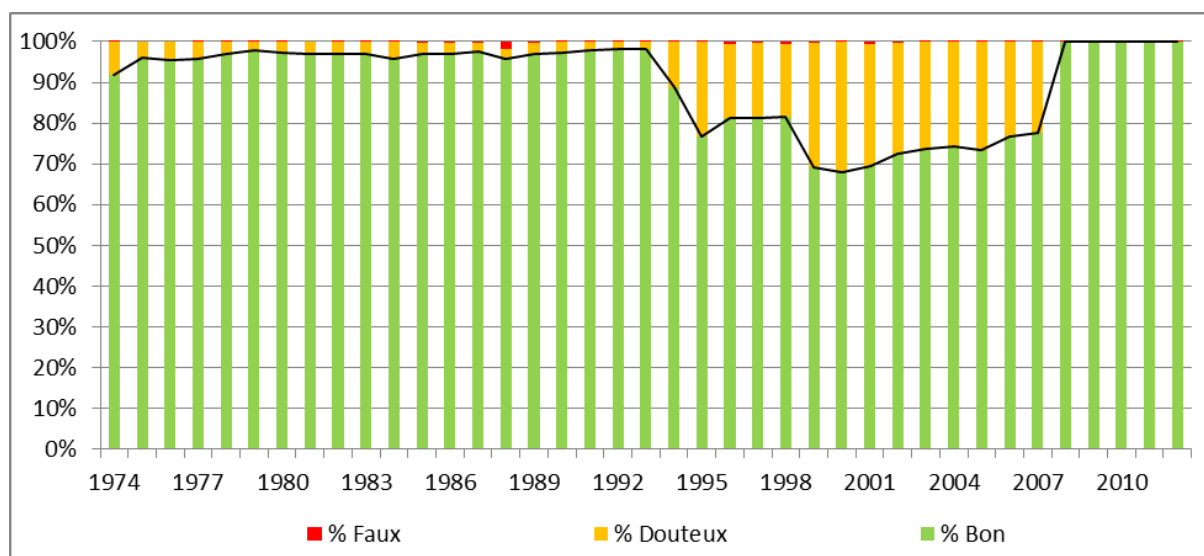


Figure 6 : Pourcentage des trois niveaux de qualité parmi les données qualifiées de 1974 à 2012 (qualification automatique)

Le pourcentage de données douteuses est en moyenne de 3% sur la période 1974-1993. A partir de 1994 ce taux passe à plus de 10%. Il faudrait vérifier les raisons de qualification à « Douteux » de ces résultats pour voir s’il serait possible de lever le doute et requalifier à « Bon » certaines données.

A noter que des commentaires de qualification soulèvent des questions :

- Cas de doublons notés « à supprimer » et qui sont toujours en base,
- Prélèvements qualifiés douteux avec mention « Qualification automatique hydrologie. Prélèvement qualifié provisoirement à douteux, en attente de re-saisi de résultats sur échantillon. - Pas d’anomalie détectée lors de la qualification automatique hydrologie ».

L’ensemble des commentaires de qualification est à reprendre et certaines corrections de données doivent être entreprises en conséquence.

3.6 Microbiologie

Un ordonnancement de la recherche d'anomalies a été défini selon les programmes des données, les paramètres, les supports et les périodes. La priorité a été donnée au **paramètre ECOLI sur supports « Bivalve », « Echinoderme » et « Gastéropode »**. Les données récentes sont prioritaires. Le Tableau 14 montre l'avancement de la qualification des données REMI. Les derniers travaux de qualification datent de 2014.

Seuls des résultats REMIS et REMIC de la période 1998-2002 ont été qualifiés via le processus de qualification automatique (seuls les résultats sans aucune anomalie ont été qualifiés à « Bon ») (Tableau 14, Figure 7). Pour les autres programmes et les autres périodes une partie des anomalies préalables (dites anomalies « Hors Pack » ou « HPAC ») ont été recherchées.

Tableau 14 : Tableau d'avancement de la qualification des données microbiologiques.

Ordre	Jeu de données	Avancement				
		HPAC	PASS	PREL	ECHA	RESU
1	REMI-SURV 2011	HPAC-014 : données non validées				
1	REMI-SURV 2012	HPAC-005 : saisie à un autre niveau qu'échantillon				
2	REMIS / REMIC 2003-2011	HPAC-011 : cohérence saisisseur passage / événement				
3	REMIS / REMIC 1998-2002	X	X	X	X	Qualification rapide des bons effectuée.
4	REMIE1 1987-2011	HPAC-008 : doublons de méthodes				
5	REMIE2 / CMIC	HPAC-006 : doublons de résultats				
6	REMIS / REMIC 1987-2002	HPAC-008 : doublons de méthodes				
7	Support Eau Autres paramètres					

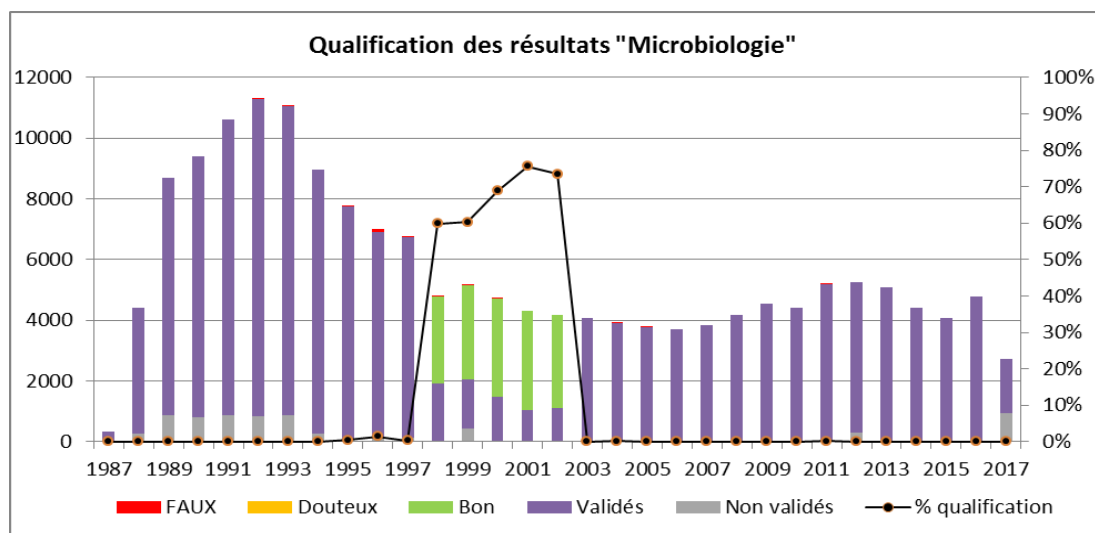


Figure 7: Avancement de la qualification des données microbiologiques.

3.1 Chimie

3.1.1 Programme RNOPHY (IMPOSEX)

L'avancement de la qualification des données du programme RNOPHY est présenté sur les Figure 8 et Figure 9. Il s'agit de qualification « automatique », excepté quelques résultats qualifiés « douteux » sur demande ponctuelle de D. Claisse. Il n'y a pas de qualification experte sur ces données, étant considéré que c'est l'expert lui-même (Martial Huet) qui a saisi, contrôlé, et validé ces données.

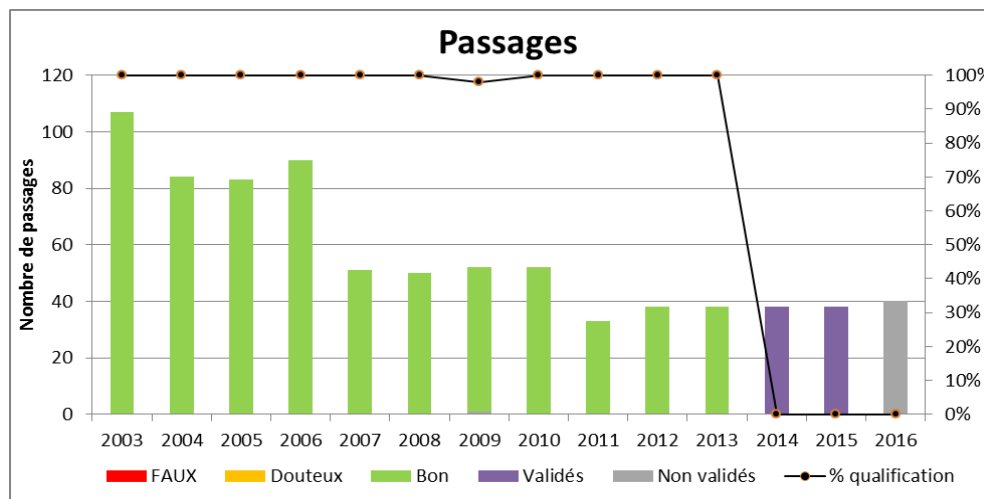


Figure 8 : Avancement de la qualification des passages du programme RNOPHY par année.

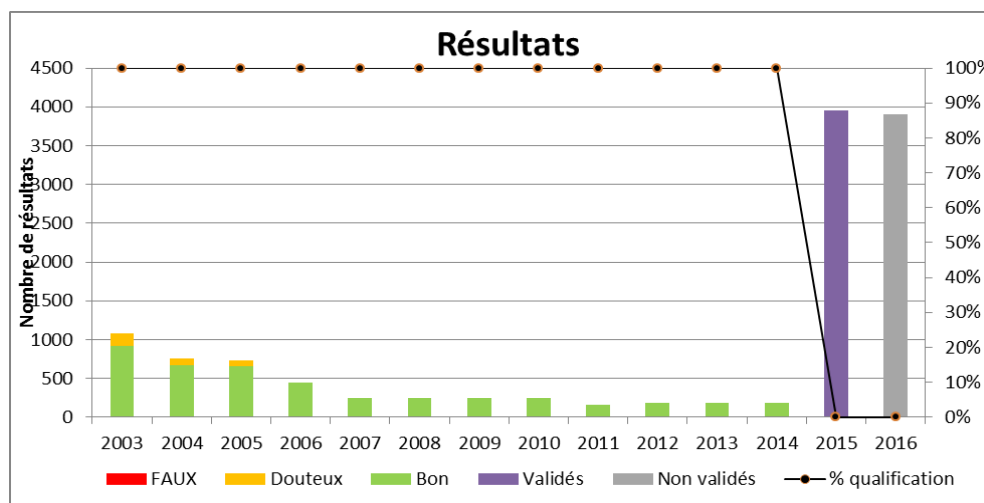


Figure 9 : Avancement de la qualification des résultats du programme RNOPHY par année.

3.1.1 Programme RNOMV

L'ensemble de l'historique des données du programme ont été qualifiées, d'abord avec des anomalies de type qualification « automatique », puis par qualification « experte ».

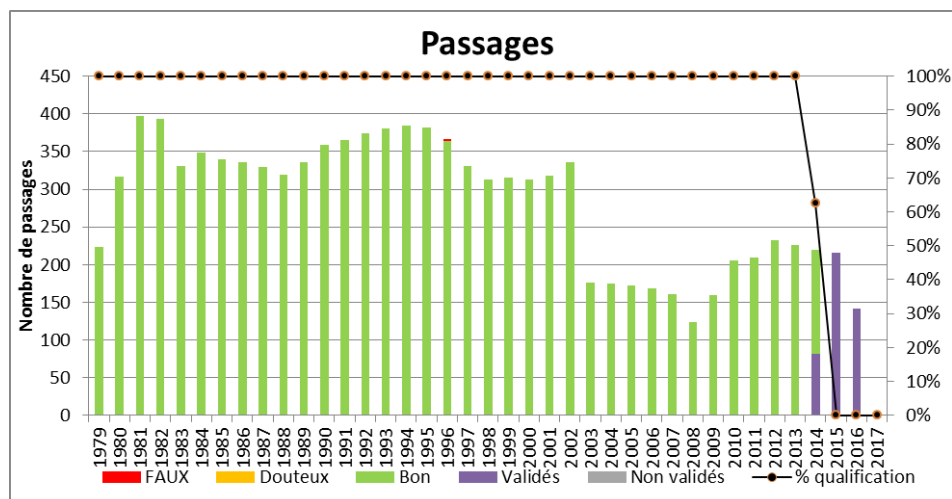


Figure 10 : Avancement de la qualification des passages du programme RNOMV par année.

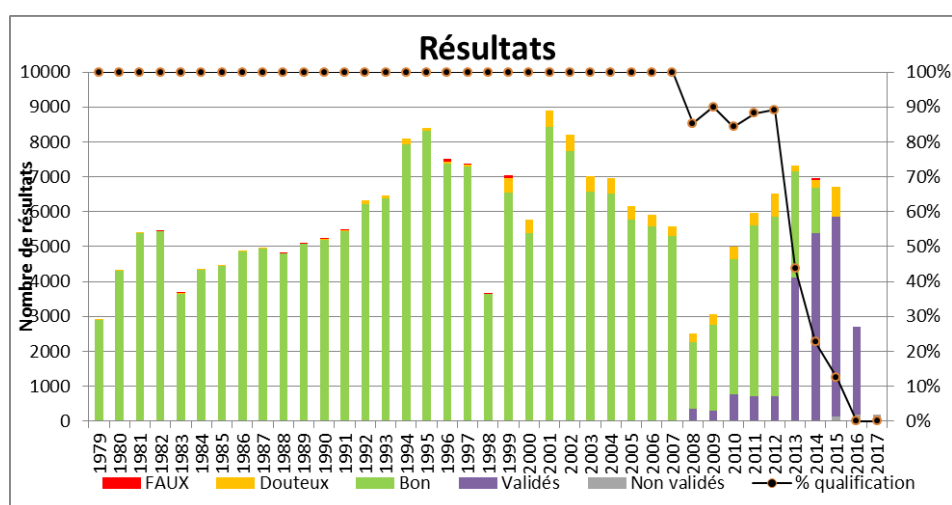


Figure 11 : Avancement de la qualification des résultats du programme RNOMV par année.

3.1.2 Programme RNOSED

Les données « in situ » ont été qualifiées via les programmes de qualification « automatique » (Figure 12). Les données de 2008 n’ont pas été qualifiées car D. Claisse n’avait pas fini de les saisir au moment de la qualification automatique.

Les résultats ont été qualifiés manuellement :

Année de qualification	Période des données	Niveau de qualité	Nb résultats
2005	1993-2001	Bon	22 741
2007	2003-2005	Bon	10 597
2014	1978-1983	Douteux	623

En 2005, la qualification a fait suite aux travaux d’un CDD sur le sujet (C. Fisson). En 2007 D. Claisse s’est replongé dans la qualification des données avant la mise en production du nouvel outil Quadrigé². Enfin, en 2014, suite aux travaux liés à la qualification automatique des données, D. Claisse a identifié des jeux de données douteux qui ont été qualifiés. Mais la qualification experte n’a pas été réalisée.

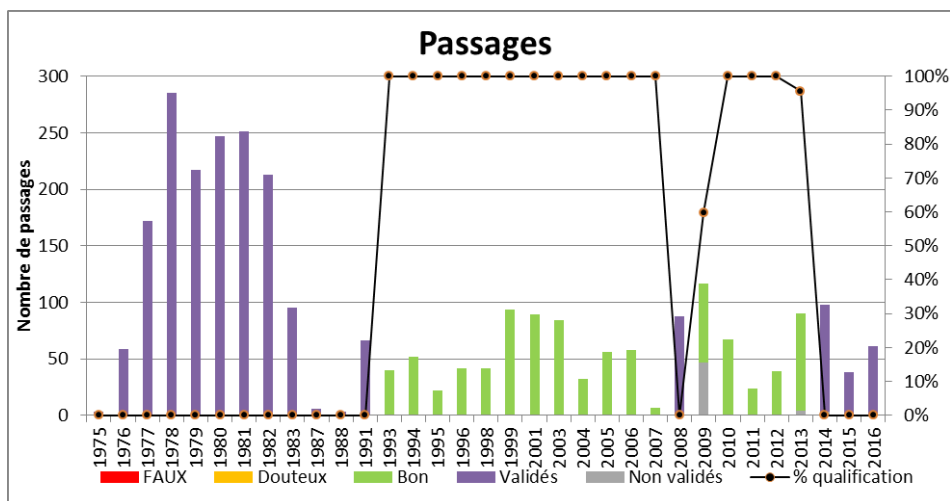


Figure 12 : Avancement de la qualification des passages du programme RNOSED par année.

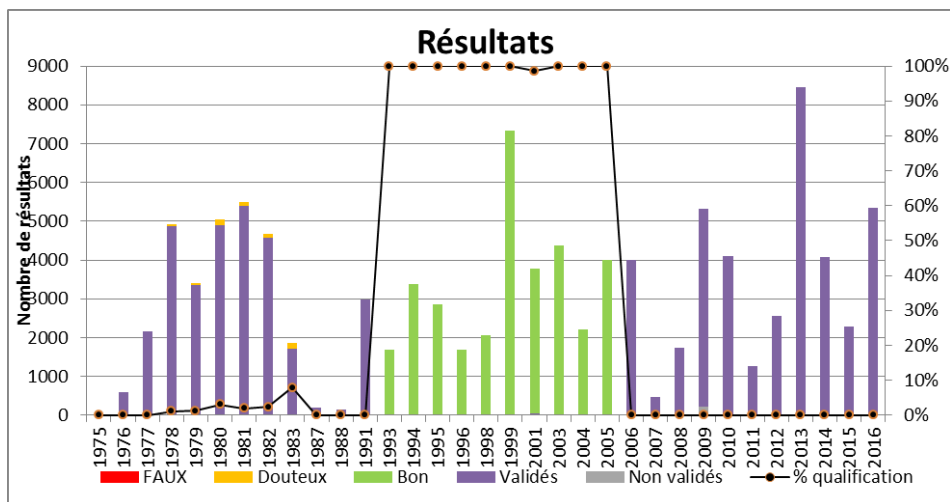


Figure 13 : Avancement de la qualification des résultats du programme RNOSED par année.

4 Simulations pour les données 2016

4.1 Hydrologie

4.1.1 Nombre d'anomalies général

Tranche de dates	HPAC-001	HPAC-002	HPAC-003	HPAC-004	HPAC-005	HPAC-006	HPAC-007	HPAC-008										
	0	0	0	1	1	2	2	2	Cohérence des PSFM	Conflits thématiques	Champs vides	Niveaux de saisie	Doublons	Cohérence préleveurs stratégies	Cohérence analystes stratégies	Validation données		
2016	42	903	476	2	41	0	0	633										
Tranche de dates	PASS-001	PASS-002	PASS-003	PREL-001	PREL-002	PREL-003	PREL-004	PREL-005	ECH-001									
	1	1	2	1	1	1	2	3	1	Heure	Unité sonde	Contrôle + validation passage	Heure	Préleveur	Unité immersion	Combinaisons sonde-engin-niveau-imm-taille	Contrôle + validation prélèvement	Contrôle + validation échantillon
2016	1	0	59	3	749	0	7232	96	95									
Tranche de dates	RES-001	RES-002	RES-003	RES-004	RES-005	RES-006	RES-007	RES-008	RES-009	RES-010	RES-011	RES-012	RES-013	RES-014	RES-015			
	1	2	2	2	2	2	2	2	2	2	2	2	2	2	3			
2016	651	4	47	31	100	2	6	116	37	106	8	28	15	1	383			

4.1.2 Simulations détaillées

Pack	Niveau de priorité	N°Anomalie	Description	Simulation données 2016
HPAC	0	HPAC-001	Cohérence des PSFM	42 PSFMUs dont 40 "nouveaux" en grande partie dus à l'ajout de l'unité dans les libellés de méthode (modification PSFMU par la cellule Quadrigé) 21 PSFMUs vraiment nouveaux

Pack	Niveau de priorité	N°Anomalie	Description	Simulation données 2016
HPAC	0	HPAC-002	« Conflits thématiques »	<p>3 passages CHIMIE_ECHPASS_SBSE 65 passages REMI-SURV 15 passages REPHY-ETUDES 410 passages REPHYTOX</p> <p>44 prélèvements REMI-SURV 38 prélèvements REPHY-ETUDES 175 prélèvements REPHYTOX</p> <p>15 échantillons REPHY-ETUDES 20 échantillons REPHYTOX</p> <p>85 résultats REMI-SURV 33 résultats REPHY-ETUDES</p> <p style="text-align: center;">903 anomalies</p>
HPAC	0	HPAC-003	Champs vides	<p>9 échantillons en anomalie (5 Antifer avec taxon support Mytilus edulis + 4 La Réunion avec un nombre d'individu renseigné) 467 résultats (La Réunion et Mayotte, incertitude remplie)</p>
HPAC	1	HPAC-004	Niveau de saisie	<p>1 PSFMU TURB dont le niveau est « PREL » dans RSLHYD alors que niveau « ECHANT » dans REPHY 1 résultat LER-MPL saisi au mauvais niveau par rapport à sa stratégie</p>
HPAC	1	HPAC-005	Doublons de résultats de mesure	41 doublons (82 résultats)
HPAC	2	HPAC-006	Cohérence des préleveurs des stratégies	Pas d'incohérence entre stratégies mais 70 préleveurs vides (RSLHYD)
HPAC	2	HPAC-007	Cohérence des analystes des stratégies	<p>4 analystes SRN / REPHY 3 analystes RHLN / REPHY</p> <p>➔ Normal mais peut-être vaut-il mieux vérifier les données pour contrôler que le bon analyste a été mis sur la bonne donnée ?</p>
HPAC	2	HPAC-008	Données non validées	<p>59 passages REPHY 96 prélèvements REPHY 95 échantillons REPHY 383 résultats REPHY</p> <p style="text-align: center;">633 données non validées</p>

Pack	Niveau de priorité	N°Anomalie	Description	Simulation données 2016
PASS	1	PASS-001	Heure de passage	1 passage à 05:57 à Mayotte
PASS	1	PASS-002	Unité de sonde <> « m »	0
PASS	2	PASS-003	Passage non contrôlé et/ou non validé	59 passages non validés
PREL	1	PREL-001	Heure de prélèvement	3 prélèvements (liés au passage de l'anomalie PASS-01)
PREL	1	PREL-002	Préleveur <> stratégie	749 prélèvements : <ul style="list-style-type: none"> • 374 cas d'inversion laboratoire Ifremer / bureau d'étude ou professionnel • 168 cas de multi-implantation de LER (ex : dans la stratégie on a « PDG-ODE-LITTORAL-LERPAC » et dans les données on a « LER/PAC/TL ») • 207 cas de changement de prestataire extérieur (ex : CREOCEAN vs IMPACT-MER), principalement dans les DOMs
PREL	1	PREL-003	Unité d'immersion <> « m »	0
PREL	2	PREL-004	Combinaison SENITU	86% des prélèvements sont bloqués à cette étape car : <ul style="list-style-type: none"> - 5455 nouvelles combinaisons (5538 prélèvements concernés) - 1760 combinaisons flaguées « Anormales » dont il faudra contrôler les données (1765 prélèvements concernés)
PREL	3	PREL-005	Prélèvement non contrôlé et/ou non validé	96 prélèvements
ECHA	1	ECHA-001	Echantillon non contrôlé et/ou non validé	95 échantillons
RESU	1	RESU-001	Analyste <> stratégie	651 résultats
RESU	2	RESU-002	TEMP	4 résultats : tous >32°C, dont 3 en métropole qui correspondent peut-être à des inversions de valeur saisie avec la salinité.
RESU	2	RESU-003	SALI	47 résultats tous > 41 et dont 35 résultats dont la valeur est <= 42. Presque tous en lagunes méditerranéennes.
RESU	2	RESU-004	TURB	31 résultats
RESU	2	RESU-005	TURB-FNU	100 résultats
RESU	2	RESU-006	OXYGENE	2 résultats
RESU	2	RESU-007	CHLOROA	6 résultats
RESU	2	RESU-008	PHEO	116 résultats

Pack	Niveau de priorité	N°Anomalie	Description	Simulation données 2016
RESU	2	RESU-009	NH4	37 résultats (4 REPHY + 33 RSLHYD, tous en Méditerranée intégrés via Quadrilabo)
RESU	2	RESU-010	NO3+NO2	106 résultats
RESU	2	RESU-011	NO3	8 résultats (tous à La Réunion)
RESU	2	RESU-012	NO2	28 résultats (tous dans les lagunes méditerranéennes)
RESU	2	RESU-013	PO4	15 résultats (tous dans les lagunes méditerranéennes)
RESU	2	RESU-014	SIOH	1 résultat (lagune méditerranéenne)
RESU	3	RESU-015	Résultat non contrôlé et/ou validé	383 résultats

Les anomalies sur les résultats (valeurs hors des bornes fixées pour chaque paramètre) concernent majoritairement les lagunes méditerranéennes. 57% des anomalies de résultats (bornes des paramètres) concernent des écosystèmes différents de la zone côtière métropolitaine (lagunes et DOMs).

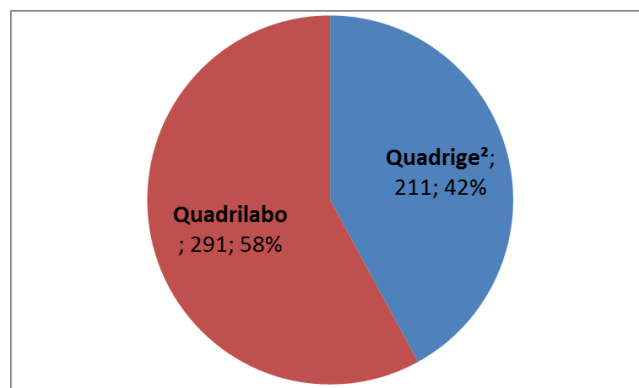


Figure 14 : Répartition des anomalies de valeurs de résultats par mode d'intégration des données.

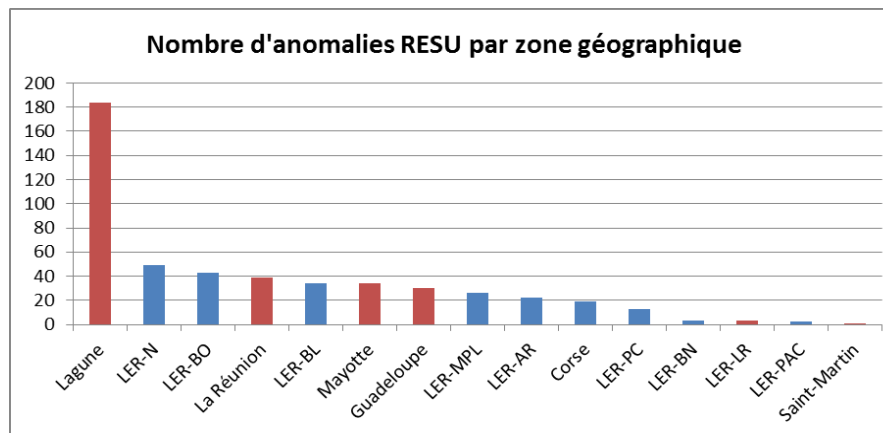


Figure 15 : Nombre d'anomalies de valeurs de résultats par zone géographique. En rouge les intégration par format Quadrilabo, en bleu les saisies par l'application Quadriges².

4.1.3 Conclusion

Le nombre d'anomalies détectées dans les données 2016 est raisonnable, sauf pour trois anomalies :

- HYD-PREL-002 : préleveur saisi différent de la stratégie : il y a encore beaucoup d'anomalies. Elles peuvent être dues :
 - Soit à des stratégies qui ne sont pas à jour,
 - Soit à une duplication d'un ancien prélèvement dont le préleveur était différent de celui de la stratégie actuelle. Remarque : un message d'alerte intervient depuis la v2.3.6 de Quadrigé² (mise en production en juin 2016), mais il n'alerte l'utilisateur qu'en cas de changement de stratégie applicable.
 - Soit à une erreur de préleveur dans des fichiers d'import Quadrilabo (cas des DOMs) : ces imports ne prennent pas en compte les préleveurs des stratégies (seule l'existence d'une stratégie applicable pour le lieu, la date et le programme est contrôlée).
 - Soit aux bilocalisations des LERs Ifremer : ex : dans la stratégie on a « PDG-ODE-LITTORAL-LERPAC » et dans les données on a « LER-PAC/TL ».
- HYD-PREL-004 : combinaisons sonde-engin-niveau-immersion-taille-unité (SENITU) : il y a énormément de nouvelles combinaisons possibles, probablement dues aux immersions très variables qui peuvent être saisies (valeurs numériques « libres », pas de classes de profondeur). La méthode de recherche de ces anomalies pourrait être revue mais cela implique de modifier les programmes R de qualification. D'autre part il y a encore beaucoup de données saisies sur des combinaisons identifiées comme « anormales ». Au total 86% des prélèvements sont bloqués à cette étape (et ne passent pas la qualification « rapide des bons »). L'autre solution est de suspendre cette recherche pour la passer en qualification « experte ».
- HYD-RESU-001 : analyste différent de la stratégie : les anomalies sont là aussi nombreuses. Il s'agit soit d'erreur d'analyste lors de l'import de fichiers Quadrilabo, soit d'une modification volontaire de l'analyste par le saisisseur. En effet, quand le saisisseur initialise les grilles de résultats, l'analyste est pré-rempli avec le laboratoire indiqué dans les stratégies. Pour le modifier le saisisseur doit faire une manipulation volontaire.

Pour les résultats, plus de la moitié des anomalies sont des valeurs à la limite des bornes définies et concernant des habitats spécifiques (33% dans les lagunes et 21% dans les DOMs). Pour rendre plus efficace la recherche d'anomalies il serait possible de définir des anomalies « métropole côtier », « métropole lagunes » et « DOMs ».

4.1 Microbiologie

4.1.1 Nombre d'anomalies

Tranche de dates	HPAC-001	HPAC-002	HPAC-003	HPAC-004	HPAC-005	HPAC-006	HPAC-007	HPAC-008	HPAC-009	HPAC-010	HPAC-011	HPAC-012	HPAC-013	HPAC-014
	0	0	0	1	1	2	2	2	2	3	4	4	4	4
	PSFM's	Conflits thématiques	Champs vides	Support - Taxon	Niveau <> Ech	Doublons résultats mesures	Doublons résultats taxons	Doublons méthodes	Saisisseur passage - lieu	Evénements	Saisisseur passage - év	Préleveurs	Analystes	Données non validées
2016, ECOLI REMI-SURV	2 (OK)	235	0	5 (OK)	0	0	0	0	157	8	0	164	403	32
Tranche de dates	PASS-001	PREL-001	PREL-002	PREL-003	PREL-004	PREL-005								
	1	1	1	2	3	3								
	Passages non contrôlés/validés	Préleveur	Engin	Niveau	Combinaisons Engin-niv-imm-supp	Unité immersion								
2016, ECOLI REMI-SURV	32	1638	0	9	104	0								
Tranche de dates	ECH-001	RES-001	RES-002	RES-003	RES-004	RES-005	RES-006	RES-007	RES-008					
	1	1	1	2	2	3	3	4	4					
	Echantillons non contrôlés/validés	Analyste	Saisisseur - Méthode	Valeur < seuil	Précision - seuil	ECOLI > 100 000	ECOLI support G ou E méthode conductancemétrie	Arrondi	STREPTOC méthode Rothe Litsky					
2016, ECOLI REMI-SURV	32	964	161	0	155	0		0	0					

4.1.2 Détail des anomalies

Pour estimer les travaux à mener pour une reprise de la qualification des données en routine, des simulations de recherche d'anomalies ont été effectuées sur les données 2016 du programme REMI-SURV (paramètre ECOLI seulement, supports Bivalve, Echinoderme et Gastéropode).

N°Anomalie	Description	Simulation 2016
HPAC-001	Vérifier les PSFM utilisés dans le cadre du REMI : est-ce que les paramètres, les supports, les fractions, les méthodes et les PSFM dans leur ensemble sont cohérents par rapport aux analyses effectuées dans le cadre du REMI?	2 PSFMs en tout. Ils semblent nouveaux mais correspondent uniquement à l'ajout de l'unité dans le libellé de la méthode. D'après le fichier support ils sont « normaux ».
HPAC-002	« Conflits thématiques » : 1/ Recherche des programmes, autres que ceux concernés par le REMI, rattachés aux données (passages, prélèvements, échantillons, résultats) à qualifier. 2/ Recherche des autres résultats rattachés aux passages concernés par le REMI	142 passages rattachés à REPHY, REPHYTOX et RSLHYD 35 prélèvements REPHYTOX 22 échantillons REPHYTOX 36 résultats REPHYTOX sur des échantillons REMI
HPAC-003	Rechercher les données pour lesquelles les champs normalement vides pour des données microbiologiques ne le sont pas.	0
HPAC-004	Rechercher les données pour lesquelles le taxon support est vide ou incohérent par rapport au support de l'échantillon	5 nouvelles associations (dus à des évolutions de la taxinomie) : tout OK
HPAC-005	1/ Vérifier la cohérence des niveaux de saisie des résultats dans les stratégies. 2/ Rechercher les résultats de mesure (ECOLI, ENTEROCO, SALMOD, STREPTOC, VIBRIOD, VHA, NoV) et résultats taxons (SALMOI, VIBRIOI) saisis à un autre niveau que celui prévu dans la stratégie.	0
HPAC-006	Rechercher les doublons de résultats de mesures : plusieurs résultats de mesure pour un même Lieu / Date / Taxon support / Paramètre / Méthode	0
HPAC-007	Rechercher les doublons de résultats taxons : plusieurs résultats taxons pour un même Lieu / Date / Taxon support / Paramètre / Méthode / Taxon	0
HPAC-008	Rechercher les doublons de méthodes : plusieurs méthodes pour un même Lieu / Date / Taxon support / Paramètre / Support / Fraction	0
HPAC-009	Vérifier la cohérence des combinaisons Lieu de surveillance - Saisisseur passage	157 nouvelles associations lieux-saisisseurs à évaluer
HPAC-010	Les données de surveillance régulière et surveillance en alerte sont différenciées par le type d'évènement qui y est rattaché ou non. Cette anomalie, comportant 4 parties, permet de rechercher uniquement pour les passages portant des résultats ECOLI, les : 1) Evènements non rattachés à un passage et concernant le REMI. 2) Evènements rattachés à un passage non REMIC ni REMI-SURV et concernant le REMI. 3) Evènements rattachés à un passage REMIC ou REMI-SURV et ne concernant pas le REMI. (pas forcément pertinent...) 4) Plusieurs évènements (tous types Q ² confondus) rattachés à un même passage	8 évènements de type REMI rattachés à aucun passage Aucune autre anomalie détectée.

N°Anomalie	Description	Simulation 2016
HPAC-011	Rechercher les passages portant des résultats ECOLI seulement et dont le saisisseur de l'évènement est différent du saisisseur du passage	0
HPAC-012	Vérifier la cohérence des préleveurs des stratégies (vérifier que les préleveurs par défaut sont exacts, et notamment que les préleveurs sont tous remplis (pour REMI-SURV seulement))	164 nouvelles combinaisons lieu – saisisseur – programme – stratégie à évaluer dont 128 REMI-SURV
HPAC-013	Vérifier la cohérence des analystes des stratégies (vérifier que les analystes par défaut sont exacts, qu'ils sont tous remplis (pour REMI-SURV seulement) et identiques pour toutes les stratégies applicables au même moment sur un même lieu - PSFM au sein des différents programmes)	403 nouvelles combinaisons lieu – PSFM – analyste – programme – stratégie à évaluer (sans doute en grande partie dues à l'ajout des unités dans les libellés de méthode)
HPAC-014	Recherche les résultats non validés (cela implique de contrôler et valider les résultats, ou de les supprimer ou de les laisser avec ce statut; ainsi que toute leur chaîne de saisie : passage / prélèvement / échantillon).	32 résultats dont au moins un des niveaux (résultat, échantillon, prélèvement, passage) n'est pas validé.
PASS-001	Rechercher les passages dont la date de contrôle ou date de validation des passages est vide (cette anomalie est présente uniquement pour que les passages soient qualifiés. En effet, sans anomalie, pas de qualification).	32 passages (cf. HPAC-014)
PREL-001	Rechercher les prélèvements dont le préleveur saisi est différent de celui prévu dans la stratégie	1638 préleveurs saisis différents des stratégies soit 40% des prélèvements ! Dont 60% dus aux codes des LERS (Ex : PDG-ODE-LITTORAL-LERPAC et LER/PAC/TL)
PREL-002	Rechercher les prélèvements dont le support est « Bivalve », « Gastéropode » ou « Echinoderme » et l'engin est différent de "Main", "Grattoir", "Pelle", "Dragues (non spécifiées)", "Tellinier" (engins définis dans les consignes de saisies)	0
PREL-003	Rechercher les prélèvements dont le support est « Bivalve », « Gastéropode » ou « Echinoderme » et le niveau est différent de « Emergé », « Mi-profondeur », « Surface (0-1m) », « Fond », "Fond/sonde-1m" (niveaux définis dans les consignes de saisies).	9 prélèvements (8 dont le niveau est « Haut de pieu » et 1 dont le niveau est « colonne d'eau »)
PREL-004	Vérifier que les combinaisons engin / niveau / immersion / support de l'échantillon sont cohérentes	27 nouvelles combinaisons à évaluer 77 combinaisons anormales dont les données sont à vérifier.
PREL-005	Rechercher les prélèvements dont le support est « Bivalve », « Gastéropode » ou « Echinoderme », et dont l'unité immersion est différente de "m"	0
ECH-0001	Rechercher les passages dont la date de contrôle ou date de validation des échantillons est vide (cette anomalie est présente uniquement pour que les échantillons soient qualifiés. En effet, sans anomalie, pas de qualification).	32 échantillons (cf. HPAC-014)
RES-001	Rechercher les résultats dont l'analyste saisi est différent de celui défini dans la stratégie	964 analystes différentes de celui prévu dans la stratégie
RES-002	Rechercher les résultats dont la méthode saisie ne correspond pas à la méthode normalement utilisée à l'époque. Cette anomalie concerne le paramètre ECOLI, les programmes REMIS, REMIC et REMI-SURV et les méthodes d'analyse utilisées dans le cadre de la stratégie régulière.	161 résultats

N°Anomalie	Description	Simulation 2016
RES-003	Rechercher les résultats dont la valeur du résultat < à la valeur seuil de détection de la méthode utilisée	0
RES-004	Rechercher les résultats ayant avec libellé précision alors que la valeur est différente du seuil de détection du PSFM de la méthode	155 résultats obtenus avec la méthode « Colimétrie E. coli NPP 3x5T ou 5x5T + Gélose - nb/hg » dont le seuil est 18, alors que la valeur saisie pour le résultat est « <20 ».
RES-005	Rechercher les résultats dont la valeur du résultat est supérieure 100 000 E.coli/CLI et concernant le paramètre ECOLI et les supports "bivalve" ou "Echinoderme" ou "Gastéropode"	0
RES-006	Rechercher les résultats ayant paramètre = ECOLI, support = Echinoderme ou Gastéropode, méthode = Colimétrie par conductance-métrie (méthode inadaptée à ces supports : à qualifier « Faux ») oui, mais pourquoi le mettre ici ?	Pas à rechercher : elle est déjà incluse dans la HPAC-001 sur les PSFMs
RES-007	Rechercher les résultats ECOLI et STREPTOC dont le support "bivalve" ou "Echinoderme" ou "Gastéropode", la date de passage est postérieure au 01/12/2002, et la valeur du résultat ne respecte pas la règle d'arrondi (que 2 chiffres significatifs)	0 (règle de recherche d'anomalie à repréciser avec les thématiciens)
RES-008	Pour les résultats STREPTOC avec la méthode « Dénombrement MPN Rothe Litsky » car la méthode mal décrite (donc résultats à qualifier douteux)	0

4.1.3 Conclusion

Les anomalies de préleveur et analyste différents de ceux prévus dans les stratégies génèrent un grand nombre d'anomalie qui risquent de bloquer le processus de qualification et de faire chuter le nombre de données passant dans la boucle de la qualification rapide des bons. Pour les autres anomalies le travail estimé est minime car le nombre d'anomalies détectées reste faible.

4.1 Chimie

4.1.1 Programme RNOPHY

Compte tenu du faible volume de données, les simulations ont été lancées pour 2 périodes restant à qualifier : 2009 d'une part et 2014-2015 d'autre part. Les données 2016 ne sont pas validées (donc non qualifiables).

N°anomalie	Description	Nb « anomalies »
HPAC-001	Vérification de la liste des PSFMs	14 nouveaux PSFMs à examiner
HPAC-002	Conflits thématiques	0
HPAC-003	Champs vides	0
HPAC-004	Support-taxon	0
HPAC-005	Niveau de saisie	0
HPAC-006	Doublons de passages	0
HPAC-007	Doublons de résultats : vérification qu'il existe un seul résultat de mesure par lieu – date – taxon – paramètre – méthode.	0 si le n° d'individu est pris en compte (136 cas sinon, allant de 2 à 40 résultats identiques).
HPAC-008	Données non validées	1 passage en 2009 40 passages en 2016
PASS-001	Heure de passage = 00:00:00	0
PASS-002	Unité de sonde du passage <> « m »	0
PREL-001	Heure de prélèvement = 00:00:00	0

N°anomalie	Description	Nb « anomalies »
PREL-002	Unité d'immersion « m »	0
PREL-003	Combinaison engin – niveau – immersion – support	0
RNOPHY-PREL-004	RNOPHY : prélèvement portant un résultat de TBT : protocole de prélèvement non précisé : à qualifier Douteux	0
ECHA-001	Echantillon non contrôlé / non validé	0
RNOPHY-ECHA-001	Echantillon Gastéropode : à qualifier Bon	Voir si la règle est toujours d'actualité
RNOPHY-ECHA-002	Echantillon d'eau ou sédiment sans résultat TBT : à qualifier Bon	0
RNOPHY-ECHA-003	Echantillon portant un résultat de TBT : protocole de prélèvement non précisé : à qualifier Douteux	0
RNOPHY-RES-001	Résultat sur gastéropode : à qualifier Bon	Voir si la règle est toujours d'actualité
RNOPHY-RES-002	Résultat sur paramètre FRINF250 : à qualifier Bon	0
RNOPHY-RES-003	Résultat TBT : protocole de prélèvement non précisé : à qualifier Douteux	0

Conclusion

La qualification automatique des données RNOPHY ne pose a priori pas de difficulté majeure. Les PSFMUs doivent être examinés mais les erreurs sont peu probables dans la mesure où il n'y a qu'un seul saisisseur qui administre aussi le programme.

En revanche, les règles de contrôle des résultats doivent être revues avec les nouveaux PSFMUs et le changement de protocole (réplicats de mesure sur plusieurs individus). De nouvelles anomalies doivent peut-être être décrites et recherchées.

4.1.2 Programme RNOMV

N°anomalie	Description	Nb « anomalies »
HPAC-001	Vérification de la liste des PSFM	34 nouveaux PSFMUs, mais seulement 5 vrais nouveaux (les autres sont dus à l'ajout de l'unité dans le libellé des méthodes)
HPAC-002	Conflits thématiques	0
HPAC-003	Champs vides	0
HPAC-004	Support-taxon	3 nouvelles combinaisons dont une fautive sûre.
HPAC-005	Niveau de saisie	0
HPAC-006	Doublons de passages	0
HPAC-007	Doublons de résultats : vérification qu'il existe un seul résultat de mesure par lieu – date – taxon – n° d'individu – paramètre – méthode	0
HPAC-008	Données non validées	4 passages, 4 prélèvements, 4 échantillons, 188 résultats
PASS-001	Heure de passage = 00:00:00	0
PASS-002	Unité de sonde du passage <> « m »	0
PREL-001	Heure de prélèvement = 00:00:00	0
PREL-002	Unité d'immersion « m »	0
PREL-003	Combinaison engin – niveau – immersion – support	1 combinaison nouvelle <i>a priori</i> anormale (pas de niveau ni d'immersion) : 1 prélèvement concerné

N°anomalie	Description	Nb « anomalies »
ECHA-001	Echantillon non contrôlé / non validé	4
RNOMV-RES-001	Résultat sur le paramètre MS% dont la valeur numérique est > 33.	0
RNOMV-RES-002	Résultats sur support Poisson : paramètres avec moins de résultats que le nombre d'individus de l'échantillon (résultats manquants)	0

Conclusion

Le volume de données à corriger est faible. Le temps de qualification est extrêmement réduit. Sur le programme RNOMV, la qualification automatique est très facile. Il reste à mettre en place des règles de contrôle pour la rendre quasiment immédiate.

C'est la qualification experte, difficile à simuler, qui est la plus chronophage.

4.1.3 Programme RNOSED

N°anomalie	Description	Nb « anomalies »
HPAC-001	Vérification de la liste des PSFMs	131 nouveaux PSFMUs dont 16 dus à l'ajout des unités dans les libellés de méthodes. 115 vrais nouveaux PSFMUs, dont 50 nouveaux paramètres
HPAC-002	Conflits thématiques	0
HPAC-003	Champs vides	0
HPAC-004	Support-taxon	Pas pour RNOSED
HPAC-005	Niveau de saisie	0
HPAC-006	Doublons de passages	0
HPAC-007	Doublons de résultats : vérification qu'il existe un seul résultat de mesure par lieu – date – niveau (horizon) – paramètre – fraction – méthode	0
RNOSED-HPAC-007	Campagnes RNOSED rattachées à d'autres programmes que le RNOSED, puis cohérence géographique des lieux des passages liés à chaque campagne.	5 anciennes campagnes (2004-2009) à potentiellement vérifier.
HPAC-008	Données non validées	0
PASS-001	Heure de passage = 00:00:00	0
PASS-002	Unité de sonde du passage <> « m »	3 passages
RNOSED-PASS-003	Sonde nulle ou =0 ou >= 300m	0
RNOSED-PASS-004	Passages ayant une géométrie réelle (pas habituel pour le RNOSED)	0
PREL-001	Heure de prélèvement = 00:00:00	0
PREL-002	Unité d'immersion « cm »	0
PREL-003	Combinaison engin – niveau – immersion – support	Pas de nouvelle combinaison, les deux combinaisons utilisées sont normales
ECHA-001	Echantillon non contrôlé / non validé	0

Conclusion

La principale charge est l'examen des PSFMUs et la réflexion autour des règles de contrôle qui pourraient leur être attribuées. Le reste des anomalies ne présente pas de difficulté.

La qualification automatique ne concerne que les données in situ, il n'y a aucune anomalie sur les résultats. L'enjeu consistera donc à mettre en place la qualification experte dont les spécifications sont écrites mais les programmes à développer.

5 Conclusion

Le processus de qualification dite « automatique » peut reprendre rapidement moyennant une mise à jour :

- Des critères de sélection des données (nouveaux programmes, nouveaux paramètres),
- Des intervenants producteurs de données.

Les simulations de recherche d'anomalie sur les données 2016 montrent deux anomalies particulièrement sévères pour les données : les préleveurs et analystes non conformes aux stratégies. Ces anomalies sont très fréquentes dans les données (jusqu'à 40% des prélèvements microbiologiques) et peuvent considérablement freiner la qualification « rapide des bons » (données directement qualifiées bonnes car n'ayant aucune des anomalies recherchées).

Une qualification automatique doit pouvoir se dérouler rapidement si des qualificateurs sont disponibles 5 jours par an : les données 2016 pourraient être qualifiées d'ici février 2018 au plus tard.

Des documents doivent être finalisés pour communiquer sur ce processus : les « procédure de qualification » de chaque thématique doivent être scindées en 2 documents :

- Une présentation thématique de la qualification (quelles données, quelles règles de contrôle) : diffusée dans et hors Ifremer, elle permettrait de communiquer à l'extérieur,
- Une procédure technique détaillant chaque étape du processus : destinée à tous les acteurs de la qualification elle permettrait à chacun de connaître son rôle dans le processus.

Hydrologie

Les qualifications déjà effectuées (données antérieures à 2013) doivent être revues via les commentaires de qualification : certaines actions n'ont pas été finalisées (doublons à supprimer, données re-saisies à qualifier Bon, etc.).

L'anomalie de combinaison sonde-engin-niveau-immersion-taille de prélèvement pose problème car elle implique une expertise manuelle du qualificateur importante (7 232 nouvelles combinaisons à expertiser soit 80% des prélèvements bloqués à cette anomalie). Le mode de recherche de cette anomalie doit être étudié pour l'optimiser.

Par ailleurs la plupart des anomalies sur les résultats concernent des écosystèmes particuliers (lagunes et DOMs). L'application de règles différentes (bornes adaptées) pour ces territoires peut être discutée et pourrait être mise en œuvre rapidement (seul le fichier de paramétrage des anomalies est à distinguer).

Microbiologie

Le processus doit être relancé en totalité selon les priorités de période / programme / paramètre / support donnés par les qualificateurs. Des règles de qualification experte pourraient être spécifiées au fur et à mesure du processus.

Chimie

La qualification automatique ne pose aucun problème spécifique et pourrait même être rapidement réalisée sur les années antérieures à 2016. En effet, la centralisation des saisies sur un seul saisisseur qui n'est autre que le qualificateur accélère considérablement le processus. La qualification des résultats soulève toutefois des questions :

- Imposex : il n'y a pas de règles poussées de contrôle des résultats car jusqu'au départ de D. Claisse c'était l'expert national de la méthode qui fournissait les données. Il faut donc définir des règles de qualification de ces résultats.
- Matière vivante : la qualification experte est la seule à discriminer précisément tous les résultats, la qualification automatique étant limitée à deux paramètres support.
- Sédiments : la qualification automatique ne concerne que les données in situ (métadonnées). Les règles de qualification experte, seules à permettre la qualification des résultats, sont décrites mais les programmes informatiques sont à développer.

Annexe : spécifications pour la qualification experte des données RNOSED

Spécification pour qualification RNOSED Experte (V1)

- **Pas de DLM**, car pas de séries longues. Au mieux 4 campagnes (série de mesures) par façade. Les lieux peuvent être légèrement différents d'une campagne à l'autre.
- **Travailler en données normalisées** (par AI% pour les métaux et par le CORG% pour les substances organiques. Détails de la normalisation en annexe 1.
- **L'unité géographique de regroupement des données** ne sera plus le lieu mais **la Zone Marine (ZM)** de Quadrigé ET les grandes entités géographiques regroupant plusieurs ZM. Je ferai un tableau des regroupements de ZM à faire.
- J'ai cherché quelle était la meilleure façon de représenter les données pour chaque ZM et regroupement de ZM. Voir divers exemples pour mercure et CB153 en annexe 2.
- En fin de compte je pense que le mieux est de faire **3 figures par entité géographique** :
 1. Cohérence de la normalisation
 2. Données par années pour l'entité géographique
 3. Paramètre normalisateur par année (pour vérifier)

Exemple en annexe 3

Sur le graphe 1 il faudrait qu'il y ait l'enveloppe à 90 ou 95% de la droite de régression car tout ce qui est au-dessus ou en dessous est signifiant du point de vue du niveau de contamination.

C'est sur le graphe 2 qu'il faudra faire tourner le même programme que pour RNOMV hors DLM. Le csv devra contenir les lignes correspondantes aux outliers potentiels détectés.

En regardant les données je pense qu'il y aura très peu de douteux sur les paramètres que je pense pouvoir qualifier. C'est en allant voir dans le csv de quel lieu il s'agit que je saurai si c'est douteux ou au contraire représentatif d'une contamination.

Annexe 1

Normalisation

1. le pourquoi du comment

Les contaminants métalliques sont accumulés dans le sédiment par adsorption sur les grains. Le pouvoir d'accumulation d'un sédiment dépend donc de la nature de ce sédiment, en particulier de sa surface spécifique, et donc de la taille des grains (granulométrie, pourcentage de particules fines). Il dépend également fortement de la nature même des grains (pourcentage d'argile). La proportion d'argile est évaluée en mesurant l'aluminium.

Pour les contaminants organiques, l'accumulation dépend aussi en grande partie de la présence ou non de matière organique. L'indicateur de cette présence est le CORG (Carbone organique). De nombreux essais sur de très nombreuses données montrent que la meilleure corrélation est obtenue, non pas avec la granulométrie, mais avec l'aluminium pour les métaux et le CORG pour les contaminants organiques.

On voit donc que pour comparer les niveaux de contamination de lieux différents il est impossible d'utiliser directement des teneurs mesurées dans des sédiments différents. Il est nécessaire de NORMALISER les données brutes pour les ramener à un sédiment théorique commun.

Ceci a conduit OSPAR à définir ses EAC (Ecotoxicological Assessment Criteria) dans le sédiment à un pourcentage de 5% d'aluminium pour les métaux et 2.5% de CORG pour les substances organiques.

2. Méthode de normalisation adoptée pour la qualification experte

La théorie complète de cette normalisation, élaborée par le CIEM, est complexe et nécessite de prélever dans chaque zone un échantillon de sable grossier⁴.

La méthode adoptée pour la qualification sera simplifiée par rapport à la théorie. Elle s'est révélée suffisante dans de nombreuses études pour une gamme de sédiment suffisamment fin. Elle peut diverger plus fortement pour des sédiments présentant des pourcentages de particules fines très faibles. C'est pourquoi, de plus en plus, on évite d'analyser ce type d'échantillon. D'autant plus que les teneurs qui y sont mesurées sont très faibles et souvent inférieures à la LQ.

On se contentera donc d'une règle de 3, après avoir converti les teneurs en Al et CORG en pourcentage car elles sont archivées respectivement en mg/kg et g/kg, poids sec.

mg/kg = PPM (parties par million)

$$\text{Al}\% = \text{Al mg/kg} / 10\,000$$

g/kg = 0/00

$$\text{CORG}\% = \text{CORG g/kg} / 10$$

Ensuite on effectue la règle de 3. Exemples :

$$\text{Hg normalisé} = (\text{Hg brut} / \text{Al}\%) \times 100$$

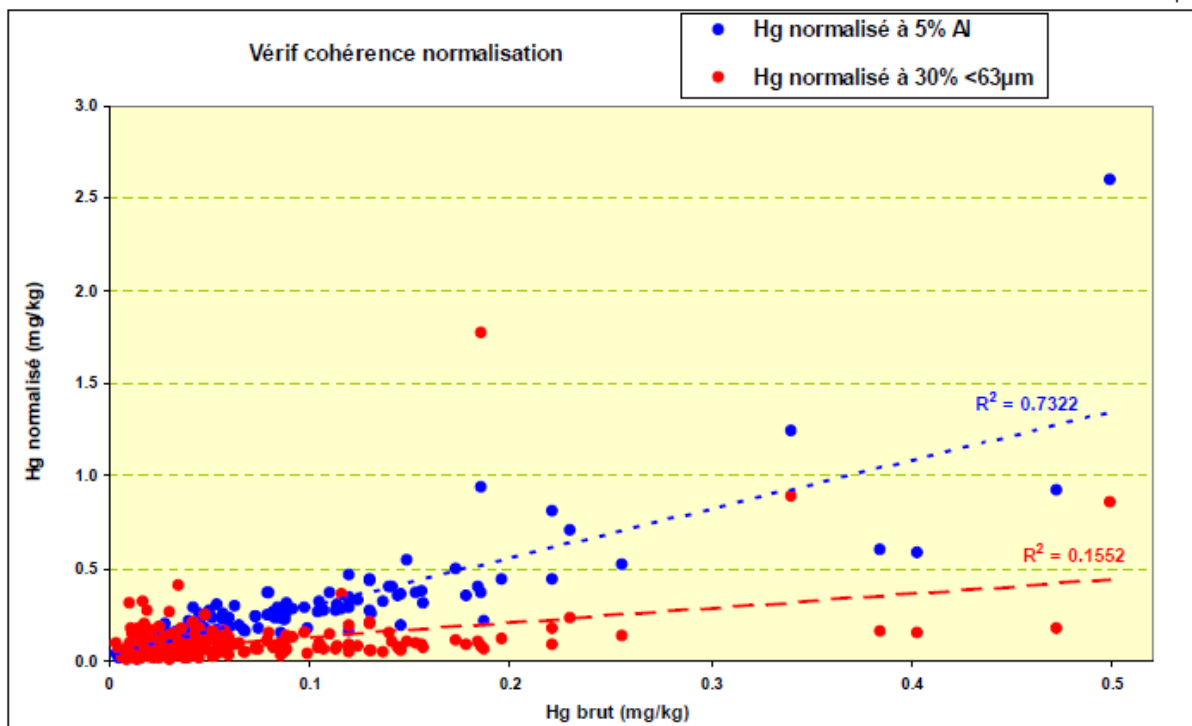
$$\text{CB153 normalisé} = (\text{CB153 brut} / \text{CORG}\%) \times 100$$

⁴ Voir B. Boutier et al. 2005. Les métaux dans les sédiments du Golfe de Gascogne. In Bulletin RNO 2005.

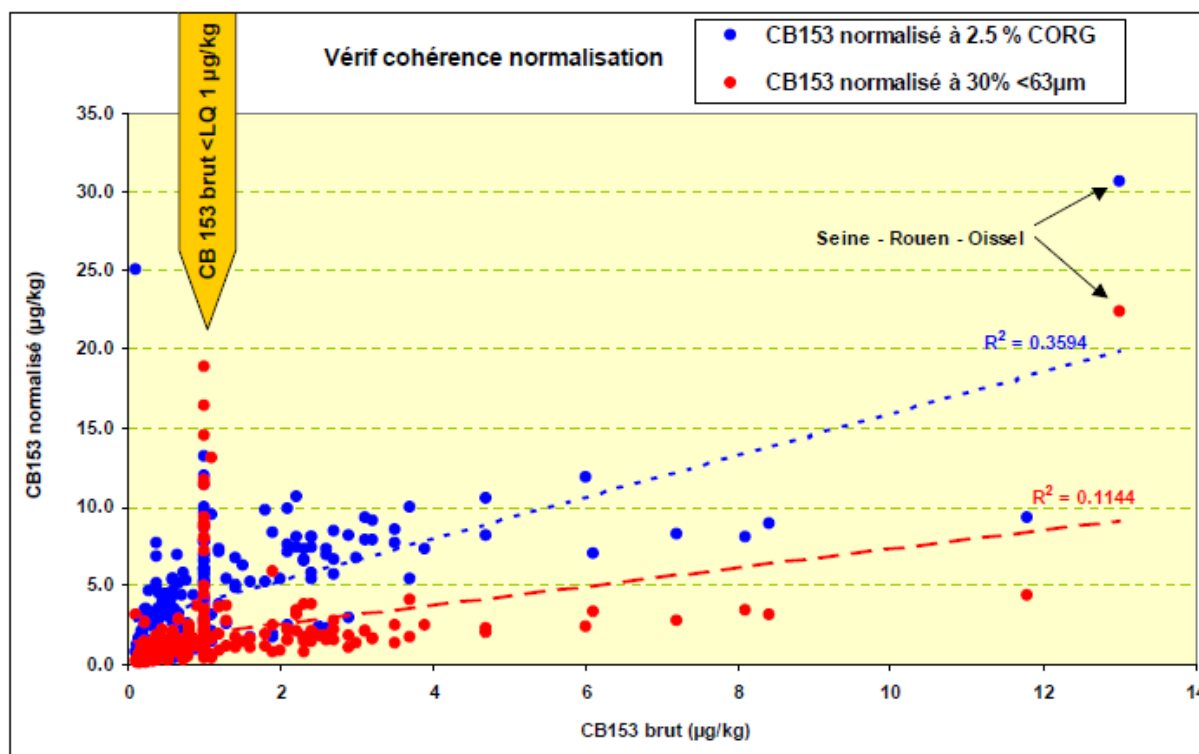
Annexe 2

Recherche du meilleur traitement des données

Comparaison de la normalisation par Al% et par %<63µm (exemple Baie et estuaire de Seine)

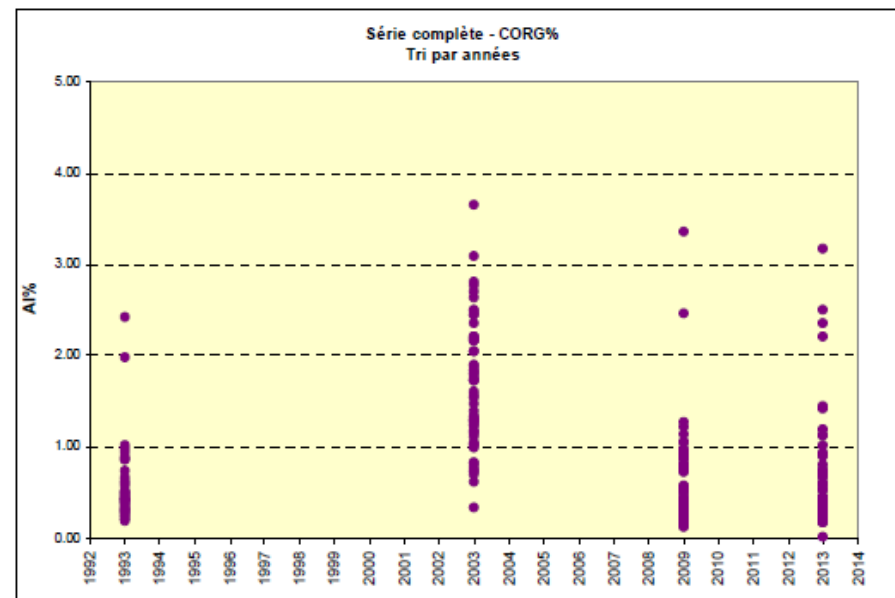
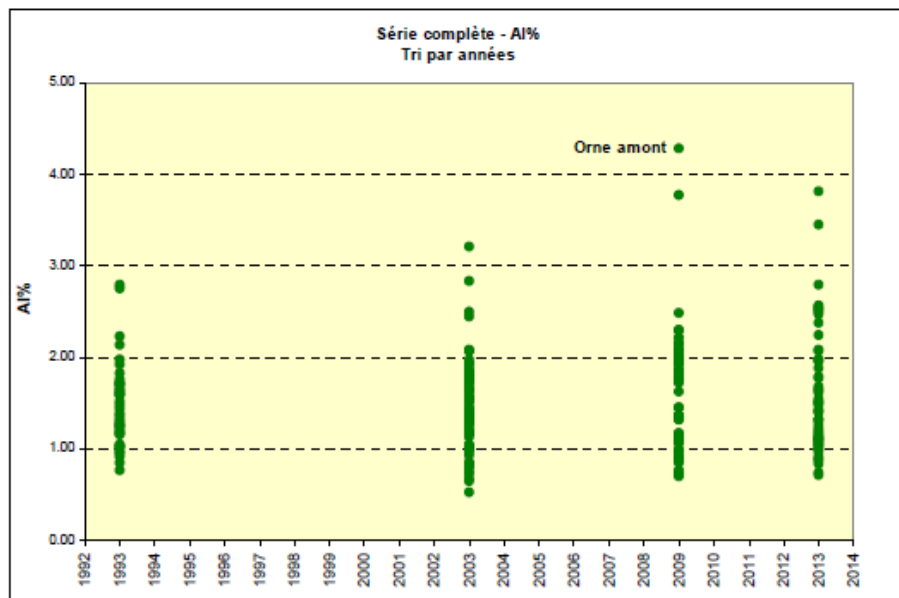
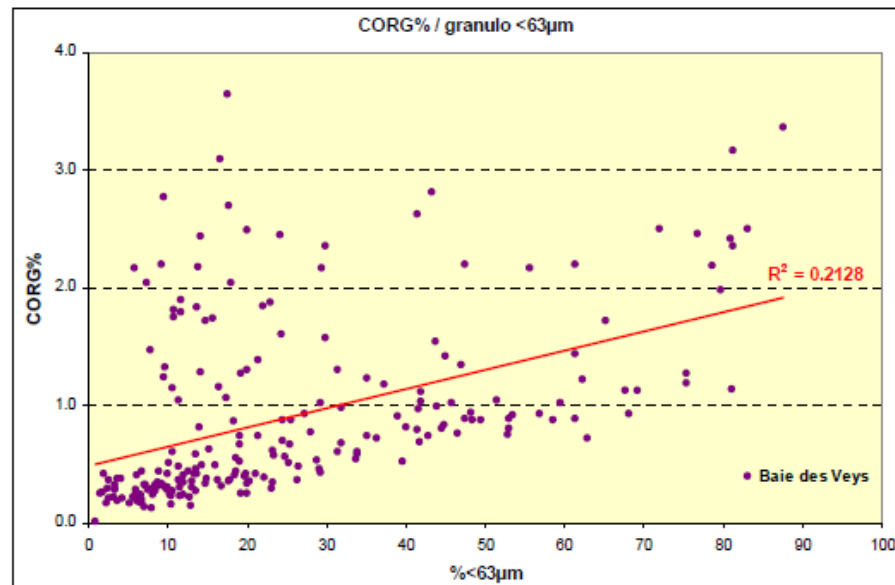
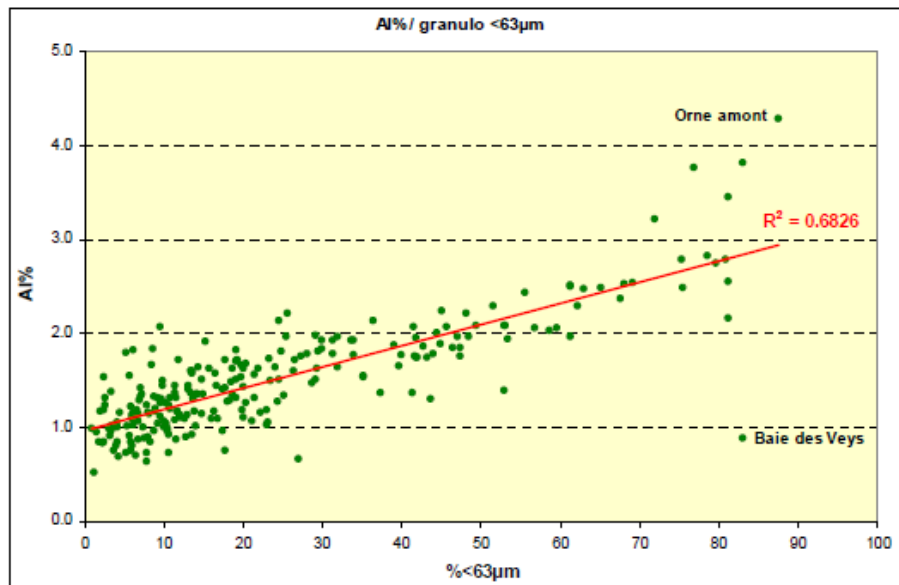


Comparaison de la normalisation par CORG% et par %<63µm

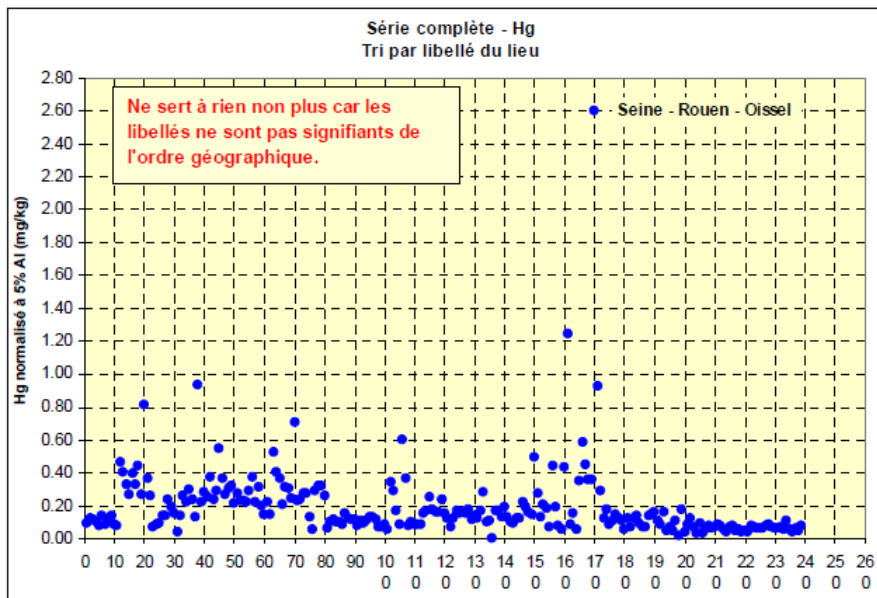
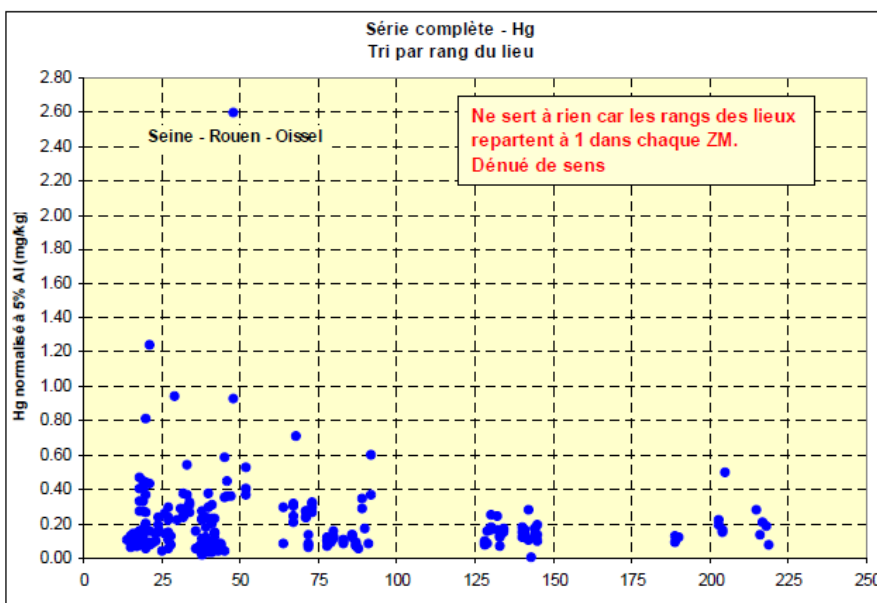
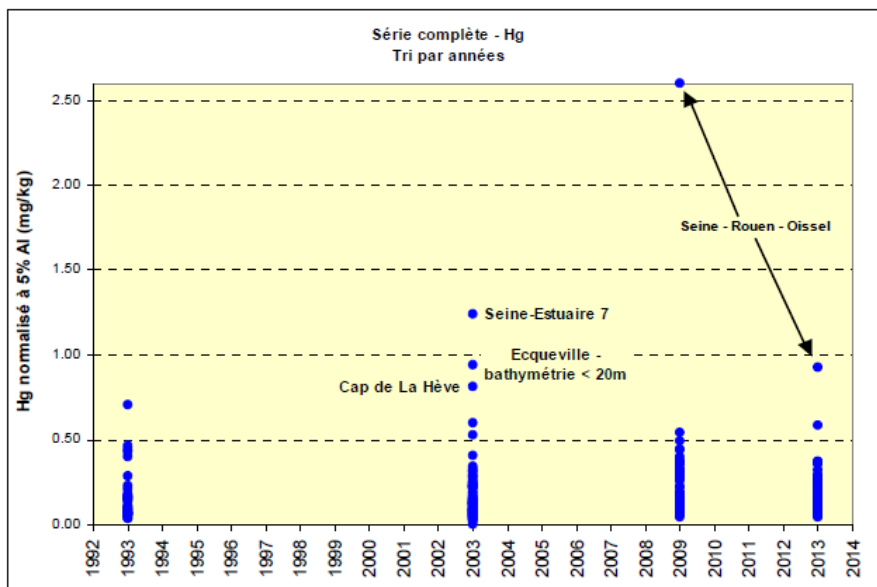


Conclusion : La granulo ne sera pas utilisée. En accord avec OSPAR on utilisera l'aluminium pour les métaux et le CORG pour les organiques.

Cohérence des paramètres normalisateurs :



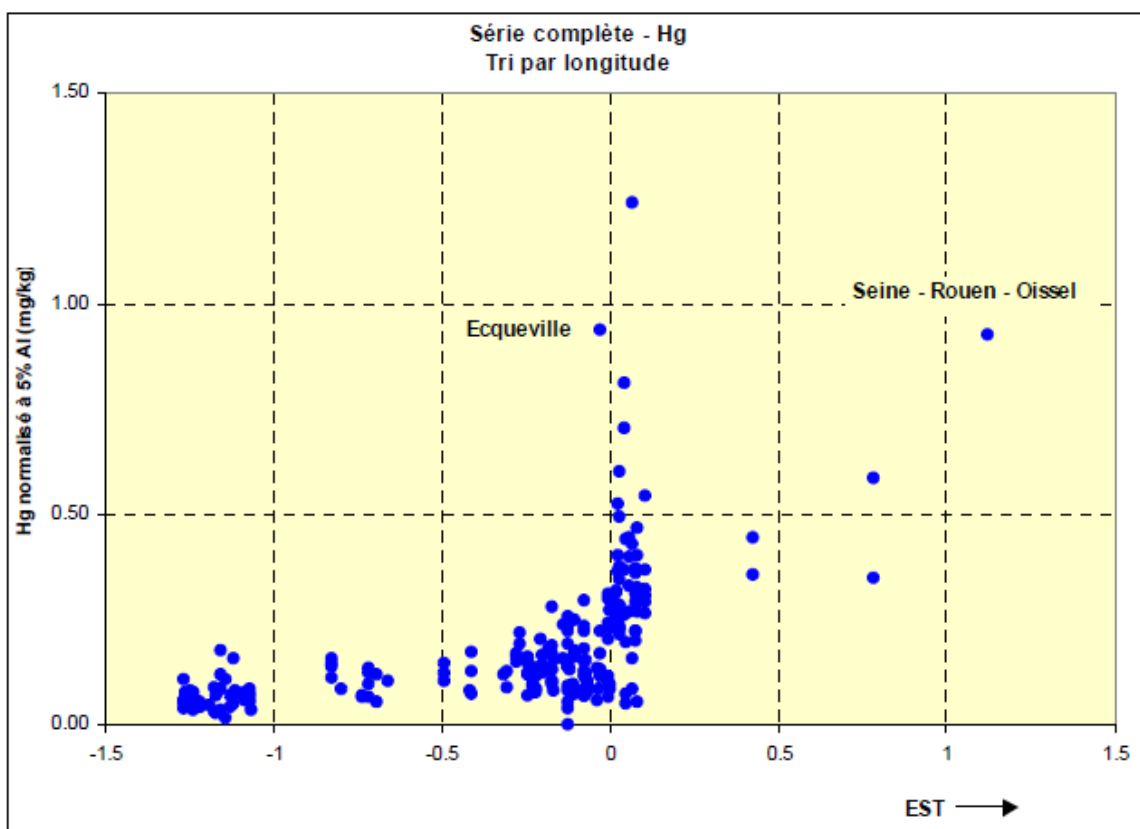
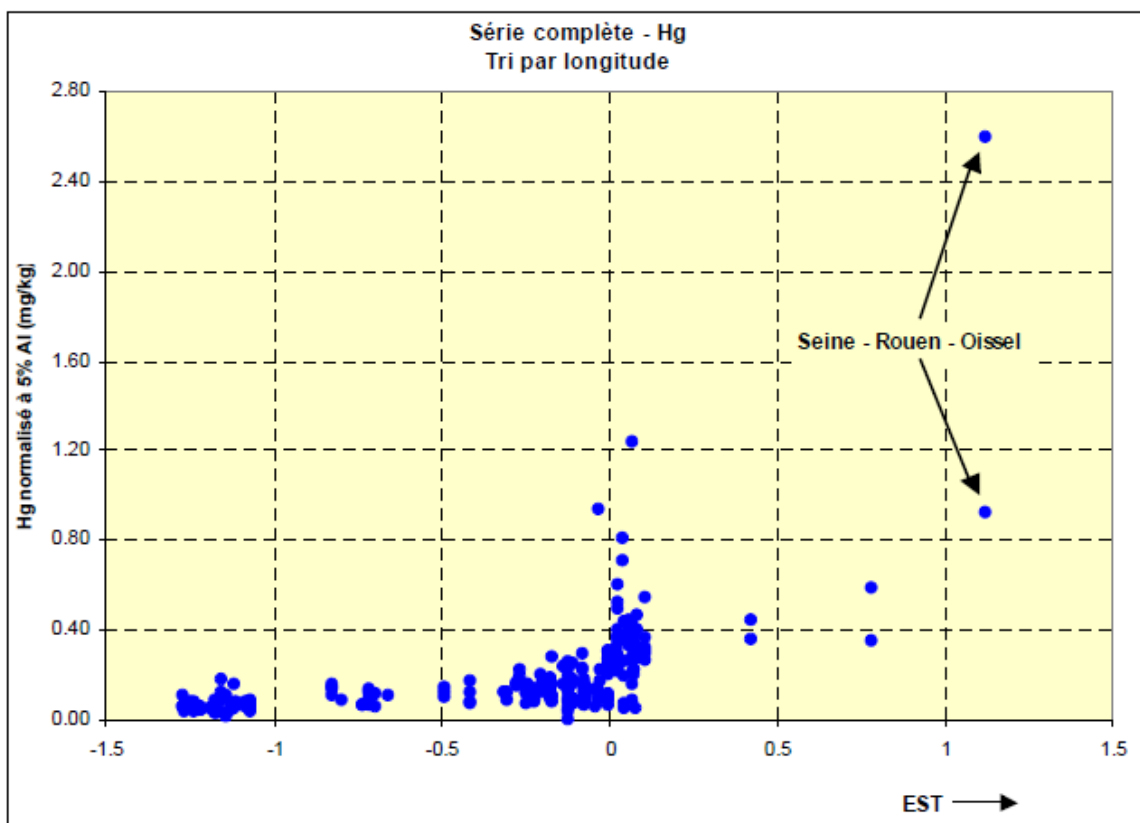
Différents types de représentation des données
exemples avec ZM 03 + ZM 10 + ZM 11 + ZM 12 + ZM 13 + ZM 14 + ZM 15 (Baie de Seine)



C'est le regroupement par campagne (donc par année) qui semble le plus parlant.

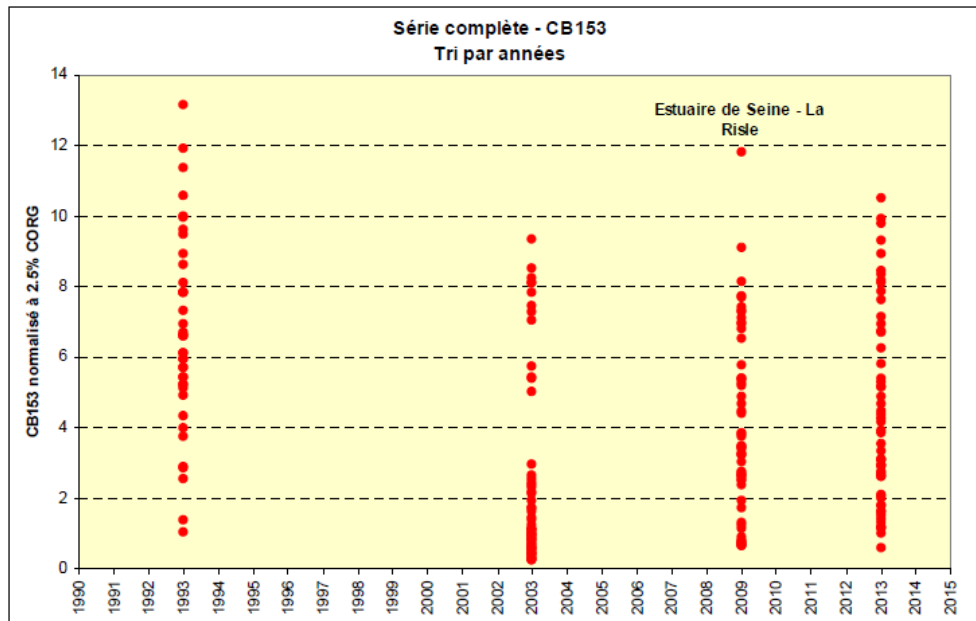
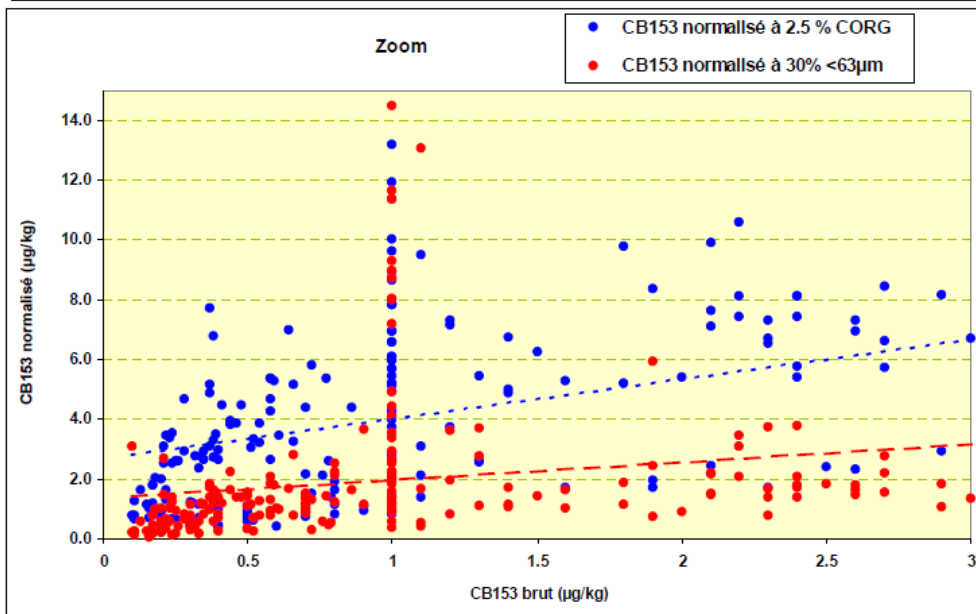
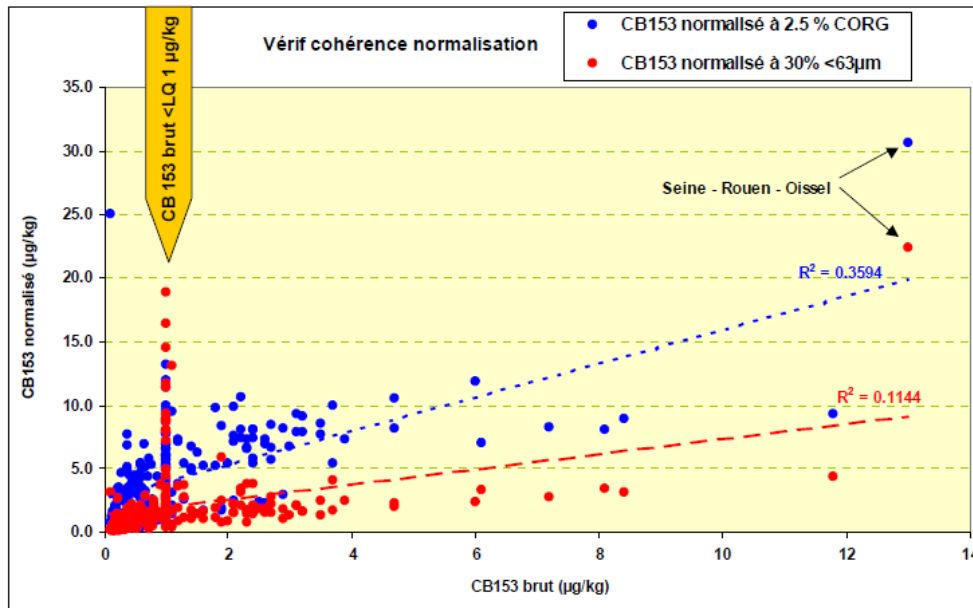
Quadriga ne contient pas de rubrique du genre "ordre géographique".

Essai de classement par la longitude.



Inconvénient : en Manche c'est la latitude qui fait la différence, en Golfe de Gascogne c'est la latitude, dans les zones complexes ça ne marche pas.

Un essai avec un composé organique : CB153



Annexe 3 - Exemple de rendu du traitement souhaité ZM 10 (Baie de Seine et Orne) – Hg

