Modélisation de la formation des trihalométhanes dans les réseaux de distribution d'eau destinés à la consommation humaine en France

C. GALEY[1], A. ZEGHNOUN[1], O. BOUDOUCH[2], P. BEAUDEAU[1], C. ROSIN[2]

Mots-clés : exposition, sous-produits de chloration, trihalométhanes, modèle prédictif, réseau de distribution
Keywords: exposure, disinfection by-products, trihalomethanes, predictive model, drinking water system

Introduction

La chloration de l'eau potable est largement utilisée dans le monde pour prévenir et limiter le risque infectieux véhiculé par l'eau du robinet. En France, son utilisation date de plus d'un siècle dans plusieurs grandes villes. Depuis 2003, les autorités françaises ont recommandé d'étendre son utilisation à l'ensemble des réseaux d'eau quelle que soit la taille de la population desservie. En 2007, plus de 99 % des débits produits sont associés à une désinfection [1].

De par ses propriétés oxydantes, le chlore réagit avec la matière organique de l'eau, les ions bromure et iodure, pour former des sous-produits de chloration (SPC). Près de 600 SPC sont identifiés à ce jour [2]. Parmi les familles majoritaires, les trihalométhanes (THM) et les acides haloacétiques (HAA) représentent à elles deux entre 20 % et 30 % de la masse totale des SPC [3]. En France, les THM sont réglementés par le Code de la santé publique [4] et font l'objet de contrôles réguliers dans l'eau distribuée. Les prélèvements d'eau s'effectuent en sortie des stations de traitement disposant d'une étape de chloration, et en réseau si la teneur en chlore dans le réseau dépasse 0,5 mg/L ou s'il y a une rechloration. La formation des SPC dépend de la nature de l'eau brute, des traitements mis en place pour éliminer la matière organique et de la stratégie de désinfection (points d'injection, doses appliquées, temps de contact).

La présence de SPC pose un problème de santé publique en raison des risques sanitaires associés et de la taille de la population exposée. Les études épidémiologiques indiquent une association entre l'exposition aux SPC, évaluée généralement par les mesures de THM effectuées dans le cadre des contrôles réglementaires, et l'apparition de cancers de la vessie, chez les hommes uniquement [5]. Une association entre THM et cancer colorectal est par ailleurs évoquée, sans consensus sur le sujet à ce jour [6]. Les effets suspectés sur la reproduction et le développement, bien que largement étudiés, sont toujours controversés [7-11]. L'estimation de l'exposition reste de façon générale le point faible des études épidémiologiques.

La formation des THM évolue dans le réseau de distribution de l'eau. Plusieurs études ont montré une augmentation des concentrations des THM d'un facteur 2 à 6 entre la sortie de l'usine et la périphérie du réseau de distribution d'eau potable [12].

1. Objectifs

Afin de mieux cerner l'exposition de la population française aux THM, un premier modèle de régression a été construit sur trois sites de production et de distribution d'eau destinée à la consommation humaine en 2009 par l'Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail (Anses) et l'Institut de veille sanitaire (InVS) [12, 13] pour prédire les concentrations des THM

[1] Institut de veille sanitaire – 12, rue du Val-d'Osne – 94415 Saint-Maurice cedex. Courriel : c.galey@invs.sante.fr
[2] Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail (Anses) – Direction de l'évaluation des risques, Unité évaluation des risques liés à l'eau – 27-31, avenue du Général-Leclerc – BP19 – 94701 Maisons-Alfort cedex.
Courriel : christophe.rosin@anses.fr

dans les réseaux d'eau à partir de données mesurées à la sortie des usines de traitement. Les données provenant de cinq autres sites de production et de distribution ont été ensuite utilisées à des fins de validation externe. La comparaison des prédictions fournies par ce modèle (dit « 2009 ») aux données mesurées sur ces cinq sites n'a toutefois pas permis d'établir la validité du modèle au-delà des trois sites considérés pour sa conception.

L'objectif de l'étude est donc de proposer deux variantes d'un nouveau modèle de régression fondé sur l'analyse de l'ensemble des données (données issues des trois sites ayant servi à l'établissement du modèle « 2009 » et des cinq sites ayant servi à sa validation externe), afin de disposer d'un modèle présentant un domaine d'application plus étendu que le précédent. Un modèle complet utilisant toutes les variables fournies par les exploitants des différents sites a été construit, ainsi qu'un modèle simplifié retenant un sous-ensemble minimal de variables, réduit à celles qui sont indispensables, ou facilement accessibles et produites en routine en France.

2. Matériel et méthode

2.1. La base de données du contrôle sanitaire de l'eau

Comme pour l'ensemble des paramètres du contrôle sanitaire de l'eau, les résultats concernant les niveaux de THM sont enregistrés dans la base de données nationale SISE-Eaux gérée par le ministère de la Santé. Cette base, qui rassemble à ce jour plus de 70 millions de résultats d'analyses depuis les années 1995, est régulièrement utilisée pour estimer l'exposition de la population et contribue à la surveillance des risques d'origine hydrique en France. Néanmoins, dans le cas des THM, les données contenues dans cette base ne permettent pas d'estimer précisément les niveaux d'exposition au robinet du consommateur, car la variabilité spatiale des concentrations en THM n'est généralement pas prise en compte dans le programme de contrôle. En revanche, les informations contenues dans la base SISE-Eaux permettent d'avoir une bonne connaissance des concentrations à la sortie de toutes les usines de traitement d'eau disposant d'une étape de chloration.

2.2. Sites d'étude

Huit sites ont été utilisés pour la construction et la validation des modèles. Tous ces sites sont alimentés par une eau de surface ou de retenue et comportent une filière de traitement complète avec une étape d'affinage par filtration sur charbon actif en grain ou filtration bicouche, et une étape d'ozonation. Ces sites ne présentent pas d'étape de préchloration. La désinfection finale par du chlore est effectuée avant la mise en distribution de l'eau dans le réseau.

Les données proviennent de différentes campagnes de prélèvements d'échantillons et d'analyses, effectuées à des saisons différentes. Lors de chaque campagne, un prélèvement d'eau a été systématiquement effectué en sortie d'usine de traitement à l'aval du point de chloration en usine (point de référence 0), et un à plusieurs prélèvements ont été effectués en différents points du réseau, avant ou après une éventuelle étape de rechloration dans ce dernier.

En conséquence, les données complètes utilisées se répartissent entre les différents sites de la manière rapportée dans le *tableau I*.

2.3. Domaine de variation des paramètres étudiés

Le *tableau II* présente la description des variables de qualité d'eau et des variables d'exploitation qui peuvent influencer la formation des THM. Certaines sont facilement disponibles, car elles sont réglementées et sont enregistrées dans la base du contrôle sanitaire de l'eau SISE-Eaux.

2.4. Modélisation

La méthode utilisée pour ajuster ces modèles repose sur le découpage aléatoire des données en deux sous-échantillons. Le premier, appelé échantillon d'apprentissage, est constitué de 75 % des données disponibles tirées aléatoirement et est utilisé pour construire le modèle. Le second, appelé échantillon de validation, est constitué des données restantes (25 %) et est utilisé pour mesurer la capacité de généralisation du modèle en comparant ses prédictions aux valeurs observées.

Les covariables sont introduites sous forme de fonctions polynomiales de degré 1 à 3, pour prendre en compte la possible non linéarité de la relation entre les niveaux de THM présents dans le réseau et les

covariables. Différents modèles de régression ont ensuite été testés avec les variables ainsi retenues, en introduisant d'éventuelles interactions. Ces modèles ont été appréciés en considérant :

• le coefficient de détermination R^2 qui permet de connaître la contribution des variables testées dans l'explication de la variabilité de la réponse, et l'erreur standard résiduelle (*root-mean-square*

Site	Nombre de campagnes	Nombre total de valeurs de trihalométhanes en réseau	Temps de séjour hydraulique (min-max, en heures)
Sites 2009 -1	3	48	(11-27)
Sites 2009 -2	3	62	(26-160)
Sites 2009 -3	3	55	(30-210)
Site 4	4	16	(64-160)
Site 5	7	48	(5-280)
Site 6	7	14	(19-57)
Site 7	7	16	(5-53)
Site 8	2	3	(15-53)
Total	–	**262**	–

Tableau I. Synthèse sur les échantillonnages des sites d'étude : nombre de campagnes, de prélèvements et gamme des temps de séjour hydraulique explorés par site d'étude

Variable	Intitulé	Unité	Min	Max
Variables du contrôle sanitaire, disponibles dans SISE-Eaux				
THM_0	[THM] en sortie d'usine	µg/L	1,3	68
		µmol/L	0,01	0,5
$Cl2_0$	Résiduel de chlore en sortie usine	mg/L	0,05	1,3
$Temp_0$	Température de l'eau	°C	7	23
COT_0	Carbone organique total en sortie usine	mg/L	1,1	4
pH_0	pH en sortie usine		7,2	8,5
Variables d'exploitation, dépendant du réseau, ou non réglementées				
$Cl2_{inj}$	Dose de chlore injectée dans le réservoir de chloration de la filière de traitement	mg/L	1,2	6
TCtp	Temps de contact dans le réservoir de chloration de la filière de traitement	Heures	0,5	6,9
TSH_i	Temps de séjour hydraulique entre le point i du réseau et la sortie usine	Heures	4,5	280
RCP_i	Présence d'une rechloration en amont du point i considéré.	Variable qualitative (Oui/Non)		
Br^-_0	Ions bromures sortie usine	mg/L	0,003	0,97
$Absuv_0$	Absorbance UV à 254 nm, en sortie usine	m^{-1}	0,003	0,08

Tableau II. Variables utilisées pour construire les modèles et domaines de variation (données provenant des huit sites d'étude. Le point i est le point en réseau pour lequel la teneur en trihalométhanes [THM] doit être prédite)

error, RMSE) qui correspond à l'erreur faite sur la prédiction ;

• la qualité d'ajustement du modèle aux données utilisées pour le construire (échantillon d'apprentissage), en regardant graphiquement la normalité de la distribution des résidus (histogramme et QQ-plot des résidus), les résidus en fonction des valeurs prédites et la concordance entre les valeurs prédites et celles observées ;

• une validation sur les données non utilisées pour sa construction (échantillon de validation), appréciée sur la base de :

– la racine carrée d'erreur quadratique moyenne (RMSE) ;

– N_{25} qui représente le pourcentage des prédictions ayant une erreur relative inférieure à 25 % en valeur absolue ;

– et N_{5inc} qui représente le pourcentage des prédictions ayant une erreur relative liée à l'incertitude, calculée sur les intervalles de confiance[3], inférieure à 5 % en valeur absolue.

Plus ces deux dernières valeurs sont élevées, plus la capacité de généralisation du modèle est importante.

La stabilité des deux modèles retenus a été vérifiée par validation croisée sur huit sous-échantillons constitués aléatoirement à partir de l'échantillon de départ. Le travail a été effectué avec le logiciel R (V2.14.2).

3. Résultats

3.1. Modèle simplifié

La recherche d'un modèle simplifié vise à disposer d'un outil prédictif, utilisant un sous-ensemble minimal de variables explicatives facilement accessibles

[3] La valeur de l'erreur relative liée à l'incertitude représente la plus petite différence entre les extrémités des deux intervalles de confiance à 95 %.

Variables du modèle simplifié	Coefficient	Écart type	*Pr*(>|t|)
Constante	**145,00**	40,10	0,000 4
THM_0	**1,25**	0,12	< 0,000 1
$THM_0 \times THM_0$	**– 1,24**	0,27	< 0,000 1
$Cl2_0$	**0,08**	0,02	< 0,000 1
TSH_i	**0,001 2**	0,000 4	0,002 5
$TSH_i \times TSH_i$	**– 0,000 009**	0,000 003	0,005 5
$TSH_i \times TSH_i \times TSH_i$	**0,000 000 03**	0,000 000 01	0,001 1
pH_0	**– 55,00**	15,40	0,000 4
$pH_0 \times pH_0$	**6,97**	1,96	0,000 5
$pH_0 \times pH_0 \times pH_0$	**– 0,29**	0,08	0,000 5
COT_0	**0,11**	0,03	0,000 4
$COT_0 \times COT_0$	**– 0,02**	0,01	0,000 7
RCP_i [0 si non, 1 si oui]	**0,33**	0,08	< 0,000 1
Prise en compte de l'interaction			
Si $RCP_i = 0$ (sans rechloration dans le réseau avant le point de prélèvement)			
$Temp_0$	**– 0,01**	0,01	0,287 0
$Temp_0 \times Temp_0$	**0,000 5**	0,000 3	0,075 8
Si $RCP_i = 1$ (rechloration dans le réseau avant le point de prélèvement)			
$Temp_0$	**– 0,05**	0,01	< 0,000 1
$Temp_0 \times Temp_0$	**0,001 8**	0,000 3	< 0,000 1

Tableau III. Variables du modèle simplifié, obtenues en utilisant l'échantillon d'apprentissage : coefficients avec leur erreur standard et leur degré de signification

(présentes dans la base SISE-Eaux) ou indispensables pour permettre à l'outil de prédire correctement.

Après exploration de la relation entre les THM_i et les variables, la forme du modèle simplifié retenu est une forme polynomiale, de degré 1 à 3 selon les variables, utilisant quatre variables du contrôle sanitaire et présentes dans SISE-Eaux – THM, carbone organique total (COT), pH, résiduel de chlore et température de l'eau, en sortie d'usine –, et deux variables liées à l'exploitation du réseau, indispensables pour disposer d'un modèle ayant de bonnes performances (temps de séjour de l'eau dans le réseau et présence ou non d'une rechloration dans le réseau avant le point de prélèvement). Il présente un terme d'interaction entre une rechloration en réseau et la température de l'eau *(tableau III)*.

La qualité d'ajustement et les performances prédictives de ce modèle sont les suivantes :

– Construction sur l'échantillon d'apprentissage (N = 197) :

R^2 = 87,15 %

RMSE = 0,048 4

p < 2,2E-16

– Validation sur l'échantillon de validation (N = 65) :

RMSE = 0,062 5

N_{25} = 67,7 %

N_{5inc} = 81,5 %

Le modèle simplifié ajuste bien les données observées. En effet, l'histogramme et le QQ-plot des résidus montrent que la distribution des résidus est proche d'une distribution normale. De plus, les valeurs résiduelles ne présentent pas de tendance particulière *(figure 1)*.

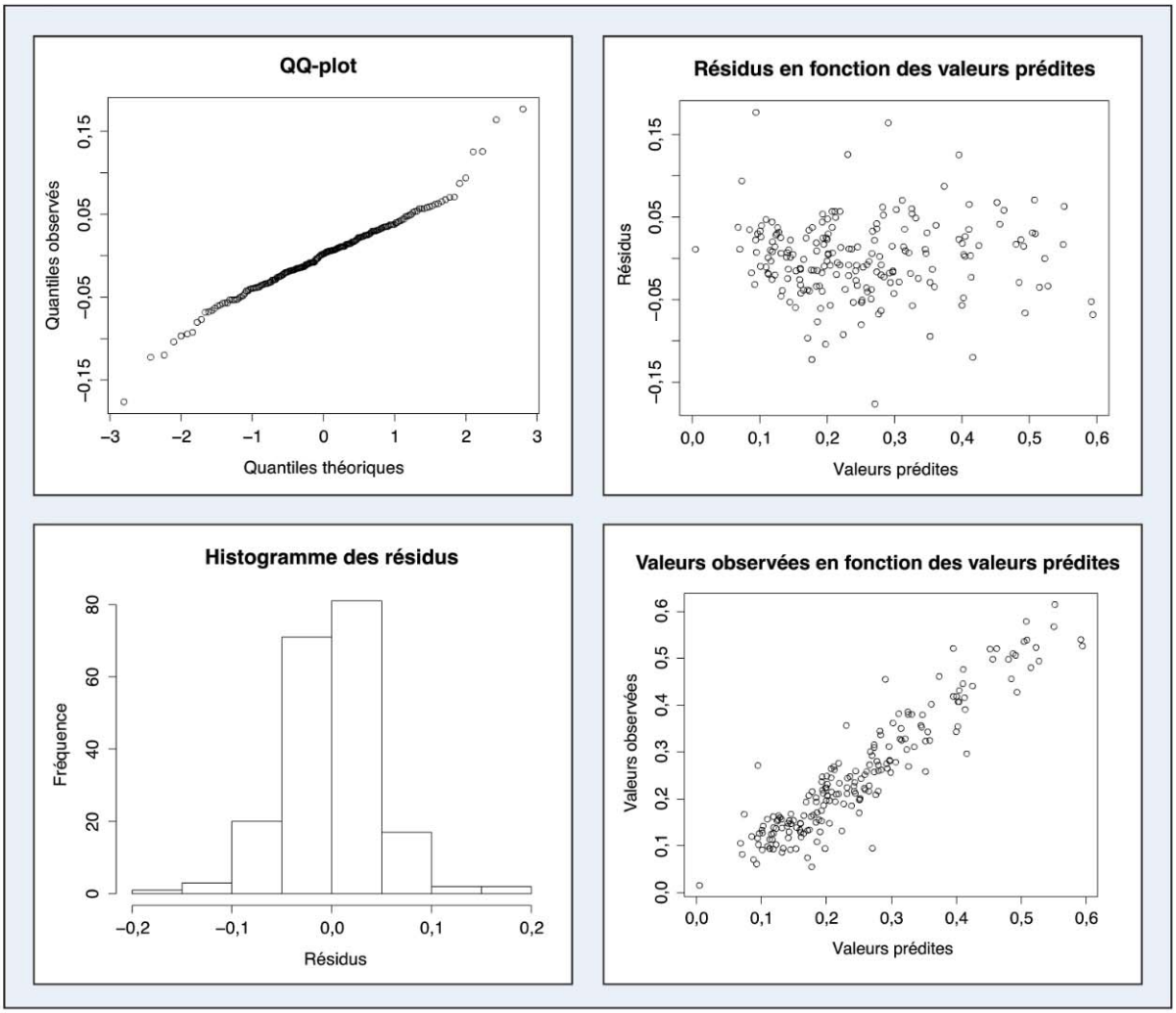


Figure 1. Qualité d'ajustement du modèle simplifié (échantillon d'apprentissage) : histogramme et QQ-plot des résidus, résidus en fonction des valeurs prédites et concordance entre les valeurs prédites et observées

Il a également de bonnes performances prédictives, car la grande majorité des prédictions de l'échantillon de validation sont proches des valeurs observées (*figure 2*, N_{25} proche de 70 % et N_{5inc} supérieur à 80 %).

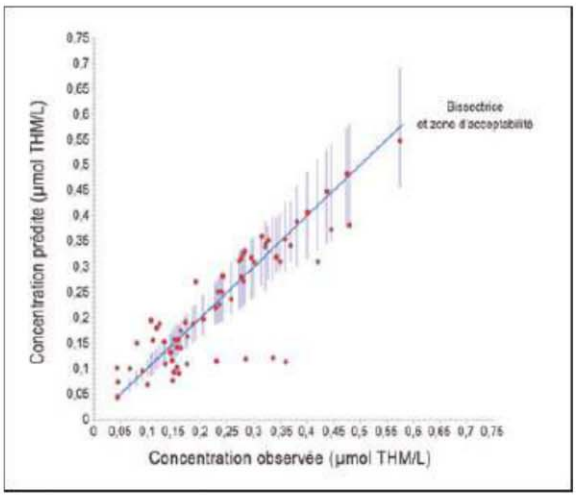
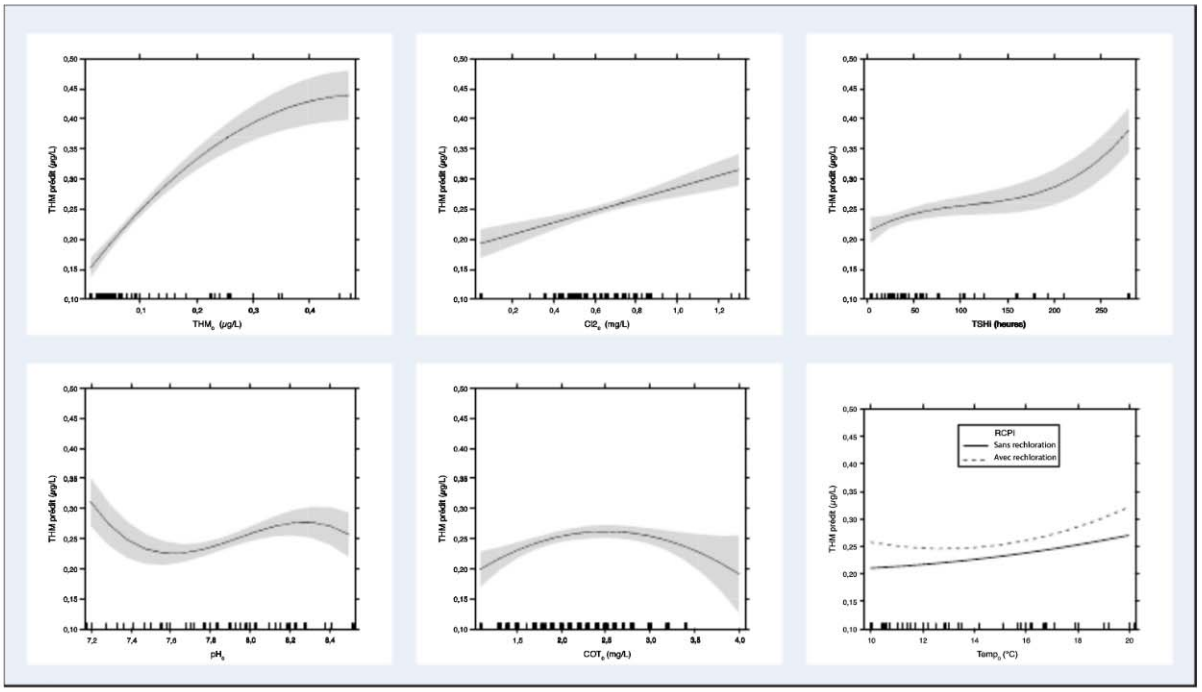


Figure 2. Validation du modèle simplifié (échantillon de validation) : graphe des concentrations prédites selon les concentrations observées. Bissectrice et zone d'acceptabilité

Les quatre concentrations observées atypiques comprises entre 0,23 et 0,37 µmol/L, pour lesquelles le modèle simplifié prédit une valeur autour de 0,1 µmol/L, appartiennent au même site (le site 7). Elles ont toutes été mesurées au printemps, trois durant la même campagne. Les quatre points de prélèvements sont différents, mais présentent une double chloration en réseau (pour trois sur quatre) ou un temps de contact mal connu et qui a pu être l'objet d'une forte sous-estimation (information > 40 heures ou > 48 heures, pour trois sur quatre. Les valeurs des temps de contacts entrés dans la base ont respectivement été 45 et 53 heures et sous-estiment sans doute fortement les temps réels qui pourraient dépasser les 200 heures).

La forme des relations entre les niveaux de THM présents dans le réseau et chaque variable explicative du modèle simplifié permet d'apprécier la cohérence des relations avec les mécanismes mis en jeu, même si le modèle construit repose sur une approche prédictive, et en aucun cas explicative *(figure 3)*.

Une relation croissante est observée entre la formation des THM_i en réseau et les variables THM_0 (THM en sortie d'usine), $Cl2_0$ (résiduel de chlore en sortie d'usine), TSH_i (temps de séjour de l'eau au point i de prélèvement), $Temp_0$ (température de l'eau) en absence de rechloration. Ces résultats sont conformes aux attentes.



Les relevés des différentes campagnes sont signalés en abscisse.

Figure 3. Relations entre les concentrations en trihalométhanes (THM) prédites dans le réseau et chaque variable du modèle simplifié (courbes en noire), avec l'intervalle de confiance (zone en grisé). Deux relations sont représentées pour la variable température, selon l'existence ou non d'une rechloration en réseau (interaction entre $Temp_0$ et RPC_i).