

**RECHERCHE DES SUBSTANCES ÉMERGENTES DANS
LES EAUX ET INTÉRESSANT LA SANTÉ PUBLIQUE ET
L'ENVIRONNEMENT**

–

PROGRAMME DE RECHERCHE IMHOTEP

Inventaire des Matières Hormonales et Organiques en
Traces dans les Eaux Patrimoniales et Potabilisables

**ANNEXE 2 DU RAPPORT FINAL : RAPPORT D'EPHESIA CONSULT –
COMPARAISON DES DONNÉES DE STEPs**

JUIN 2018

Comparaison des données de STEP

Dimitri D'Or, Ephesia Consult

22 mars 2017

Table des matières

1	Le jeu de données	1
2	Objectif	2
3	Préambule sur les tests statistiques	2
4	Identification de différences de concentration à l'intérieur des séries	3
5	Identification de différences entre les deux STEPs	3
6	Identification de différences entre la moyenne sur un jour et la moyenne sur une semaine	4
A	Tableaux des moyennes sur une journée et sur une semaine	6
B	Matrices de corrélation	6
C	Graphiques des séries temporelles	11

1 Le jeu de données

Le jeu de données est fourni dans le fichier Excel *STEP_pour RAMSES.xlsx*, produit par Vincent Brahy début février 2017. Il contient des mesures de concentration pour 54 molécules (médicaments et certains pesticides "traceurs") réalisées dans 2 stations d'épuration (STEP) (Basse-Wavre, grosse STEP accueillant plutôt des effluents d'un environnement urbain et de petites entreprises et Basècles, petite STEP accueillant plutôt des effluents d'un environnement rural). A chaque STEP, les échantillonnages ont été réalisés selon deux fréquences :

- toutes les 2 heures pendant 1 journée ;
- tous les jours pendant 1 semaine.

L'échantillon composite du jour 1 de Basse-Wavre a été prélevé durant toute la journée durant laquelle les échantillons horaires ont été prélevés. Il s'agissait du mercredi. Aucune information n'est disponible sur la constitution de l'échantillon composite de Basècles.

Le fichier original a subi les modifications suivantes avant d'être utilisé pour l'analyse statistique :

1. Remplacement des "," par des "." comme séparateur de décimales

2. Correction de la valeur de Isoproturon_pharma pour le jour J7 de 1.7 à 17 ng/l (probablement une erreur de dilution)
3. Remplacement des caractères accentués dans les noms de variables
4. Suppression des deux premières lignes et de la ligne 4 (unités)
5. Remplacement des valeurs de l'échantillon horaire 9 de Basse-Wavre par des Non Valeurs car cet échantillon n'a pas été analysé (d'après les renseignements fournis par Vincent Brahy et confirmés par Francis Delloye). Les valeurs présentes dans le fichier original sont celles de l'échantillon composite.

2 Objectif

Sur base des données fournies, il est demandé d'apporter une réponse aux deux questions suivantes :

1. Est-il possible d'identifier des différences de concentrations selon l'heure ou le jour dans les séquences fournies ?
2. Y a-t-il des différences de concentrations entre les deux STEPs ?
3. Y a-t-il des différences entre les concentrations moyennes mesurées sur les deux séries de données dans chaque STEP ?

3 Préambule sur les tests statistiques

Les tests statistiques utilisés dans ce rapport doivent être interprétés avec précaution pour trois raisons :

1. Ils font tous l'hypothèse d'un échantillon iid (indépendant et identiquement distribué). En outre, ils sont optimaux pour des distributions normales. Or, dans le cas présent, même s'il est difficile de vérifier l'hypothèse de normalité sur des jeux de données de si petite taille, il est généralement admis que les données de type concentration suivent plutôt des distributions lognormales avec beaucoup de valeurs faibles et peu de valeurs élevées.
2. Ensuite, on peut remettre également en question l'hypothèse d'indépendance entre les mesures à l'intérieur d'une série. En effet, il est hautement probable que deux données successives dans le temps montrent des valeurs plus proches que des données plus éloignées dans le temps. Par manque de données, il est toutefois impossible d'estimer cette corrélation temporelle avec une précision suffisante. Dans le cas où les données ne sont pas indépendantes, on remplace le nombre de données effectivement mesurées par un nombre plus faible (parfois appelé nombre équivalent de données indépendantes - NEDI) pour tenir compte de la corrélation. Si les données sont parfaitement corrélées, ce NEDI vaut 1. L'échantillon se réduit donc à une seule valeur. Ce nombre plus faible a pour effet de rendre le test moins puissant. Dans les tests réalisés ici, nous n'avons pas corrigé le nombre de données pour tenir compte de la corrélation. Si un test donne un résultat non significatif, il ne sera pas significatif non plus avec un nombre corrigé. A l'inverse, une différence significative avec le nombre de données effectif pourrait disparaître après correction pour tenir compte de la corrélation entre données. Il faut donc rester prudent dans l'interprétation des résultats des tests.

- Enfin, un résultat non significatif ne signifie pas une absence de différence, mais l'impossibilité de mettre en évidence une différence éventuelle. Cette impossibilité peut être soit due au fait que la différence n'existe pas, soit que le nombre de données disponibles ne permet pas de la mettre en évidence.

4 Identification de différences de concentration à l'intérieur des séries

Le jeu de données disponible ne permet pas de mettre en évidence des différences de concentrations selon l'heure ou le jour de mesure. En effet, comme il n'y a pas de répétition, il n'est pas possible de calculer des moyennes et de comparer celles-ci.

Tout au plus est-il possible d'afficher graphiquement les évolutions sur la journée ou la semaine mesurées. Il est cependant hasardeux de tirer des conclusions générales à partir de l'observation d'une seule journée ou d'une seule semaine.

A titre d'information, les tableaux de moyennes sur une journée et sur une semaine sont disponibles à l'Annexe A et les matrices de corrélation entre composés pour les fréquences horaires et journalières sont données à l'Annexe B.

Les graphiques des séries temporelles sont montrés à l'Annexe C.

5 Identification de différences entre les deux STEPs

Pour comparer les deux STEPs, un test de student pour données paires a été mis en oeuvre. Les données ont été considérées comme paires parce que, dans les deux séries, elles sont prélevées au même moment. Si les variances apparaissent statistiquement différentes, la méthode de Welch est utilisée pour approximer les degrés de liberté.

Pour la fréquence horaire, les composés suivants montrent une différence hautement significative ($\alpha = 0.01$) :

[1]	"Isoproturon_pharma"	"MCPA_pharma"	"Bentazone_pharma"
[4]	"Diclorobenzamide pha"	"Lincomycine"	"Sulfamethazine"
[7]	"Sulfamethoxazole"	"Trimethoprim"	"Diclofenac"
[10]	"Ibuprofene"	"Hydroxyibuprofene"	"Ketoprofene"
[13]	"Naproxene"	"Tramadol"	"Levamisole"
[16]	"Atenolol"	"Metoprolol"	"Rosuvastatine"
[19]	"Acide.fenofibrique"	"Losartan"	"Irbesartan"
[22]	"Ramipril"	"Ramiprilate"	"Carbamazepine"
[25]	"Carbamazepine.10.11"	"Oxazepam"	"Citalopram"
[28]	"Venlafaxine"	"Furosemide"	"Hydrochlorothiazide"
[31]	"Cafeine"	"Cotinine"	"Iomeprol"

Aucun composé ne montre de différence significative pour $\alpha = 0.05$.

Pour la fréquence journalière, les composés suivants montrent une différence hautement significative ($\alpha = 0.01$) :

[1]	"Bentazone_pharma"	"Diclorobenzamide_pha"	"Lincomycine"
[4]	"Sulfamethazine"	"Sulfamethoxazole"	"Diclofenac"
[7]	"Ibuprofene"	"Hydroxyibuprofene"	"Ketoprofene"
[10]	"Naproxene"	"Tramadol"	"Levamisole"
[13]	"Atenolol"	"Sotalol"	"Metoprolol"
[16]	"Losartan"	"Irbesartan"	"Ramipril"
[19]	"Ramiprilate"	"Citalopram"	"Venlafaxine"
[22]	"Furosemide"	"Hydrochlorothiazide"	"Cotinine"
[25]	"Iomeprol"		

Les composés suivants montrent une différence significative ($\alpha = 0.05$) :

[1]	"MCPA_pharma"	"Trimethoprim"	"Rosuvastatine"
[4]	"Acide.fenofibrique"	"Carbamazepine"	"Carbamazepine.10.11"
[7]	"Oxazepam"		

En conclusion, les différences significatives, voire hautement significatives, pour une grande majorité de composés montrent que les deux STEPs sont différents du point de vue des concentrations mesurées.

6 Identification de différences entre la moyenne sur un jour et la moyenne sur une semaine

Pour comparer les concentrations horaires avec les concentrations journalières, on utilise un test de Student pour données non paires.

Pour Basse-Wavre, aucun composé ne montre de différence significative pour $\alpha = 0.01$.

Les composés suivants montrent une différence significative ($\alpha = 0.05$) :

[1]	"Irbesartan"	"Iomeprol"
-----	--------------	------------

Pour Basècles, les composés suivants montrent une différence hautement significative ($\alpha = 0.01$) :

[1]	"Isoproturon_pharma"	"Lincomycine"
-----	----------------------	---------------

Les composés suivants montrent une différence significative ($\alpha = 0.05$) :

[1]	"Sulfamethoxazole"	"Trimethoprim"	"Ketoprofene"	"Irbesartan"
[5]	"Cafeine"			

En conclusion, les concentrations mesurées sont significativement différentes entre la moyenne horaire et la moyenne hebdomadaire seulement pour un petit nombre de données. Sur base des données disponibles, il n'est pas possible de mettre en évidence une différence entre les deux types de moyenne avec confiance. Cela peut être dû à un manque de données ou à l'absence réelle d'une telle différence. En l'état des choses, il ne semble donc pas y avoir de raison de privilégier des mesures à un moment plutôt qu'un autre dans la journée ou dans la semaine.

TABLE 1 – Moyennes horaires et journalières.

Composé	Basse-Wavre		Basècles	
	Moyenne horaire (ng/l)	Moyenne journalière (ng/l)	Moyenne horaire (ng/l)	Moyenne journalière (ng/l)
Isoproturon pharma	17.08	17.66	31.52	19.49
MCPA pharma	74.06	50.54	25.68	22.84
Bentazone pharma	87.61	91.37	4.58	3.84
Diclorobenzamide pha	69.1	72.89	8	8
Estrone	0	0	5.31	9.14
Clarithromycine	490.18	558.6	453.56	450.79
Lincomycine	12.34	10.09	65.63	45.81
Sulfamethazine	2.48	3.47	0	0
Sulfadiazine	3.64	0	0	0
Sulfamethoxazole	125.81	142.01	15.8	34.02
Trimethoprine	57.76	49.87	20.26	28.36
Diclofenac	1043.73	1199.74	593.26	631.97
Ibuprofene	414.23	509.29	24.37	56.19
Hydroxyibuprofene	839.27	905.95	127.03	158.08
Ketoprofene	172.58	177.81	50.71	72.54
Naproxene	345.14	392.91	63.4	101.25
Paracetamol	8.65	21.83	10.36	12.51
Tramadol	1120.85	1268.41	683.25	690.13
Levamisole	21.28	22.98	9.17	10
Metrifonate	0	0	0	1.33
Atenolol	93.43	119.48	57.8	51.69
Sotalol	973.2	1063.82	910.65	846.51
Metoprolol	363.92	387.57	70.12	69.42
Rosuvastatine	76.53	93.73	48	52.58
Acide.fenofibrique	1207.94	1233.25	535.98	635.53
Losartan	234.63	276.06	70.31	73.43
Irbesartan	1394.12	1625.65	757.56	891.68
Ramipril	7	7.26	0	0.83
Ramiprilate	93.25	98.56	0	0
Carbamazepine	653.09	708.6	532.1	513.26
Carbamazepine.10.11	59.93	64.28	46.14	44.4
Oxazepam	120.58	121.05	92.32	91.04
Citalopram	114.16	124.47	76.13	82.18
Venlafaxine	797.55	824.82	434.1	397.38
Furosemide	658.01	683.44	73.8	112.81
Hydrochlorothiazide	1456.23	1566.72	781.52	835.55
Ranitidine	347.5	377.64	279.42	317.83
Cafeine	490.67	695.4	187.1	502.99
Cotinine	31.62	38.87	15.43	15.59
Iomeprol	2523.59	1756.94	449.33	385.14

A Tableaux des moyennes sur une journée et sur une semaine

Le Tableau 1 fournit les moyennes sur une journée et sur une semaine pour tous les composés des deux STEPs dont au moins une concentration mesurée est supérieure à 0. Les composés non montrés dans ce tableau ont toutes leurs valeurs à zéro.

B Matrices de corrélation

Les coefficients de corrélation sont calculés entre composés pour chaque STEP et pour chaque fréquence de mesure. Les graphiques sont donnés aux Figures 1 à 4. Les composés n'apparaissant pas dans les graphiques ont des valeurs manquantes ou toutes identiques empêchant le calcul du coefficient de corrélation. Sur chaque graphique, les composés sont ordonnés selon les coefficients de corrélation. Les valeurs bleues indiquent des valeurs proches de 1 (corrélation parfaite) et les valeurs rouges, des valeurs proches de -1 (corrélation inverse parfaite). Les cases blanches indiquent des coefficients de corrélations non significatifs au niveau $\alpha = 0.05$. L'ellipse représente la forme du nuage. Plus elle est ronde, plus le coefficient de corrélation est proche de 0 ; plus elle tend vers une droite, plus le coefficient tend vers 1 (pente positive) ou -1 (pente négative).

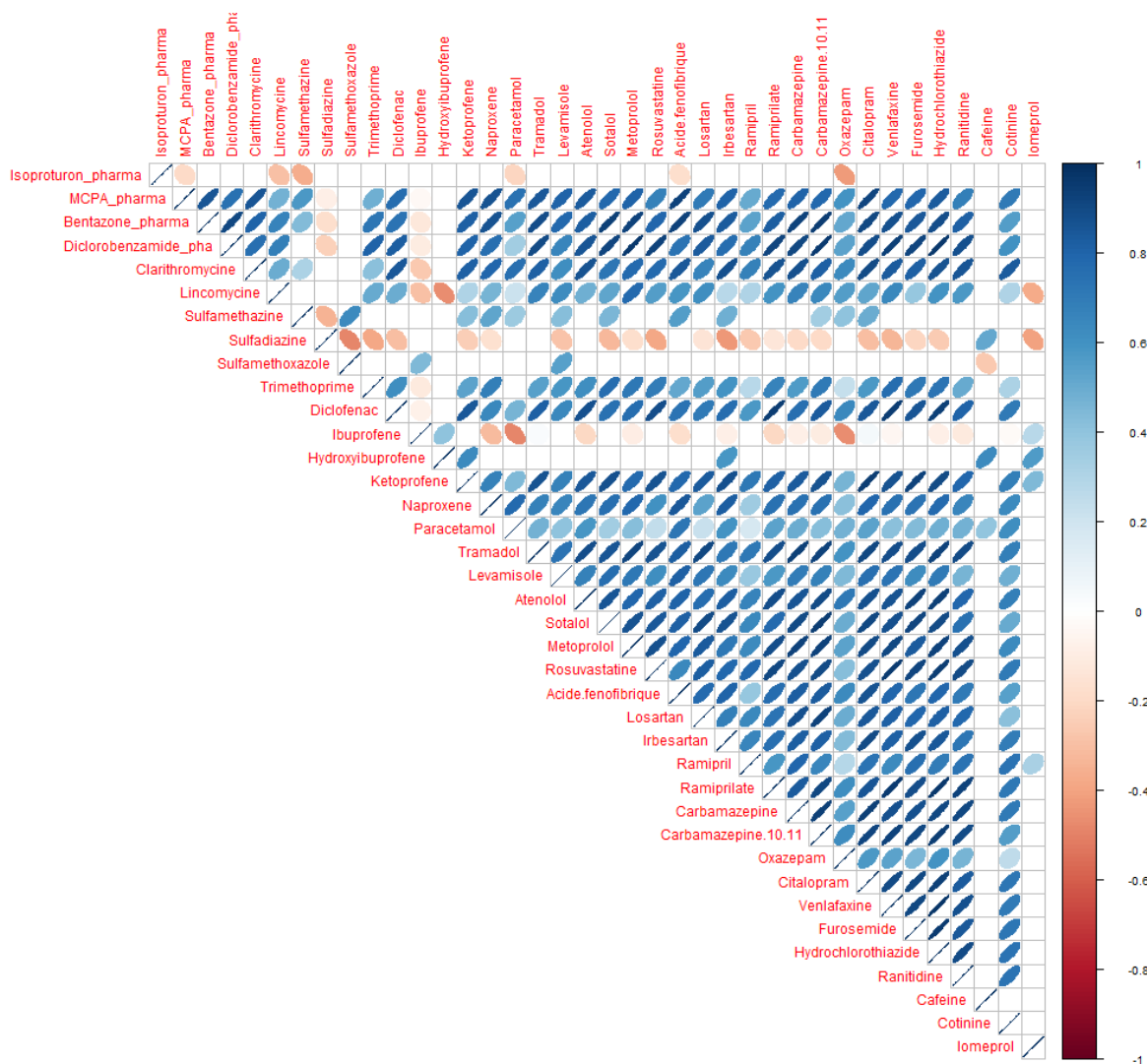


FIGURE 1 – Matrice de corrélation entre composés pour les concentrations mesurées à une fréquence horaire de Basse-Wavre.

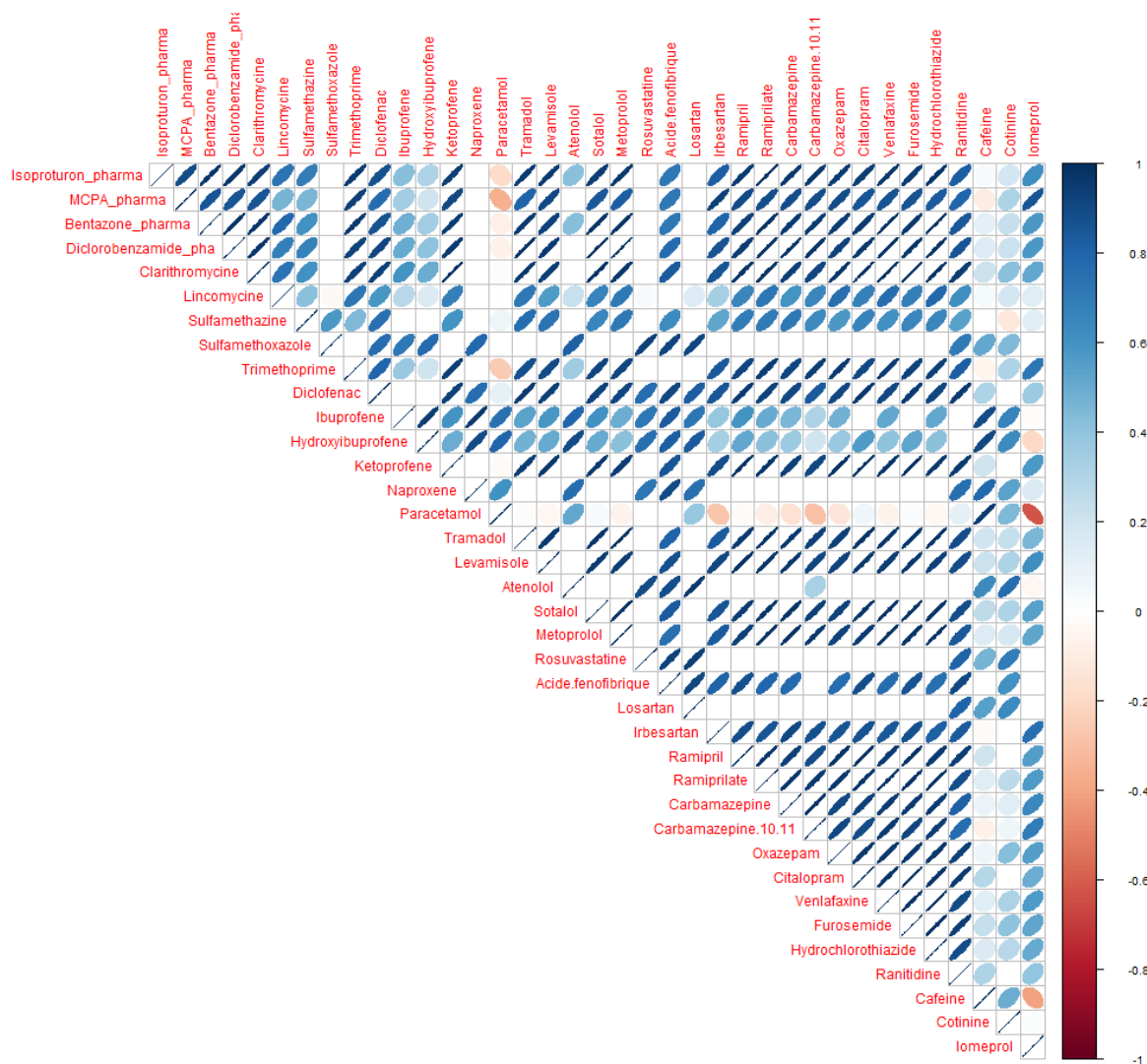


FIGURE 2 – Matrice de corrélation entre composés pour les concentrations mesurées à une fréquence journalière de Basse-Wavre

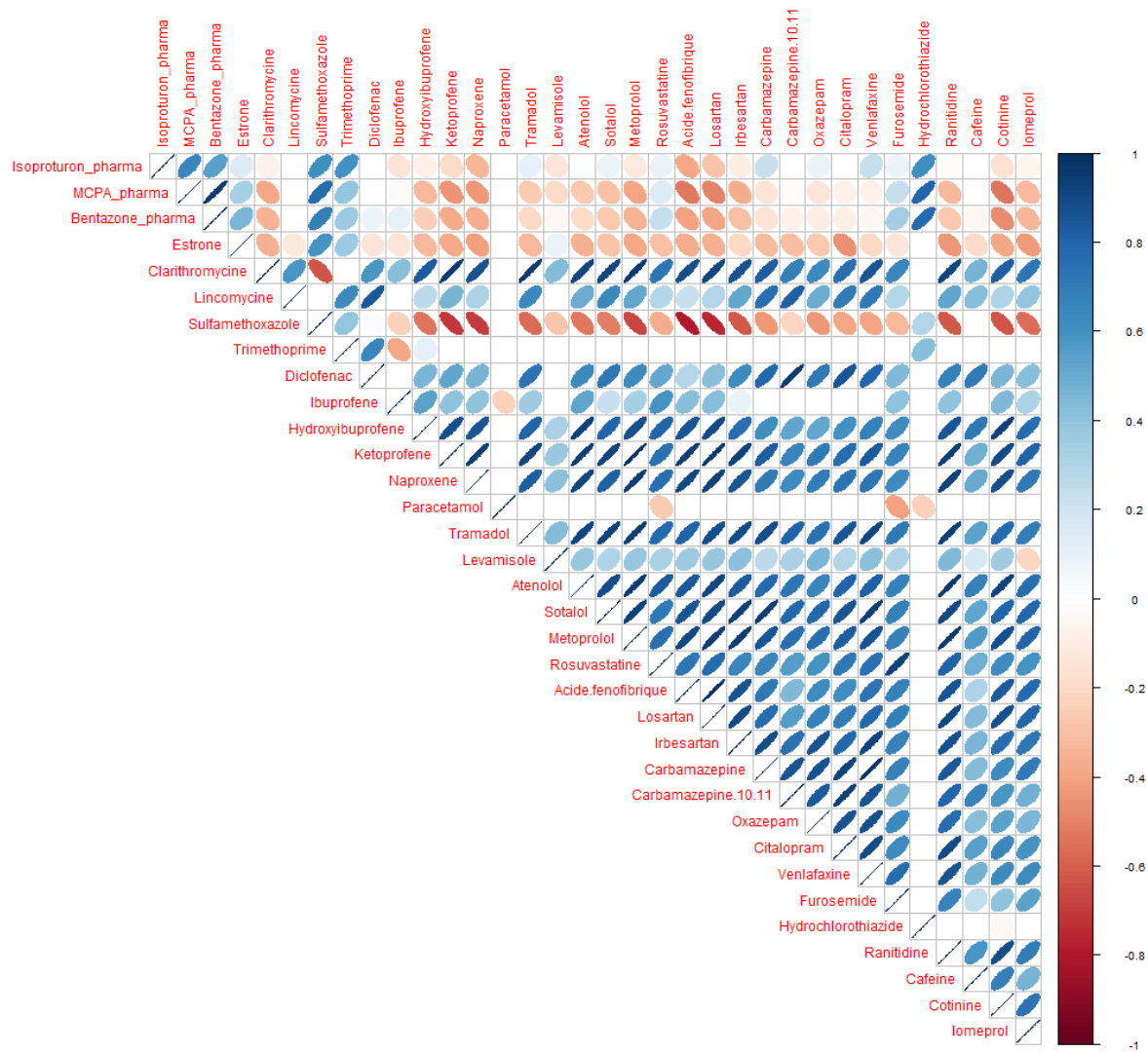


FIGURE 3 – Matrice de corrélation entre composés pour les concentrations mesurées à une fréquence horaire de Basècles

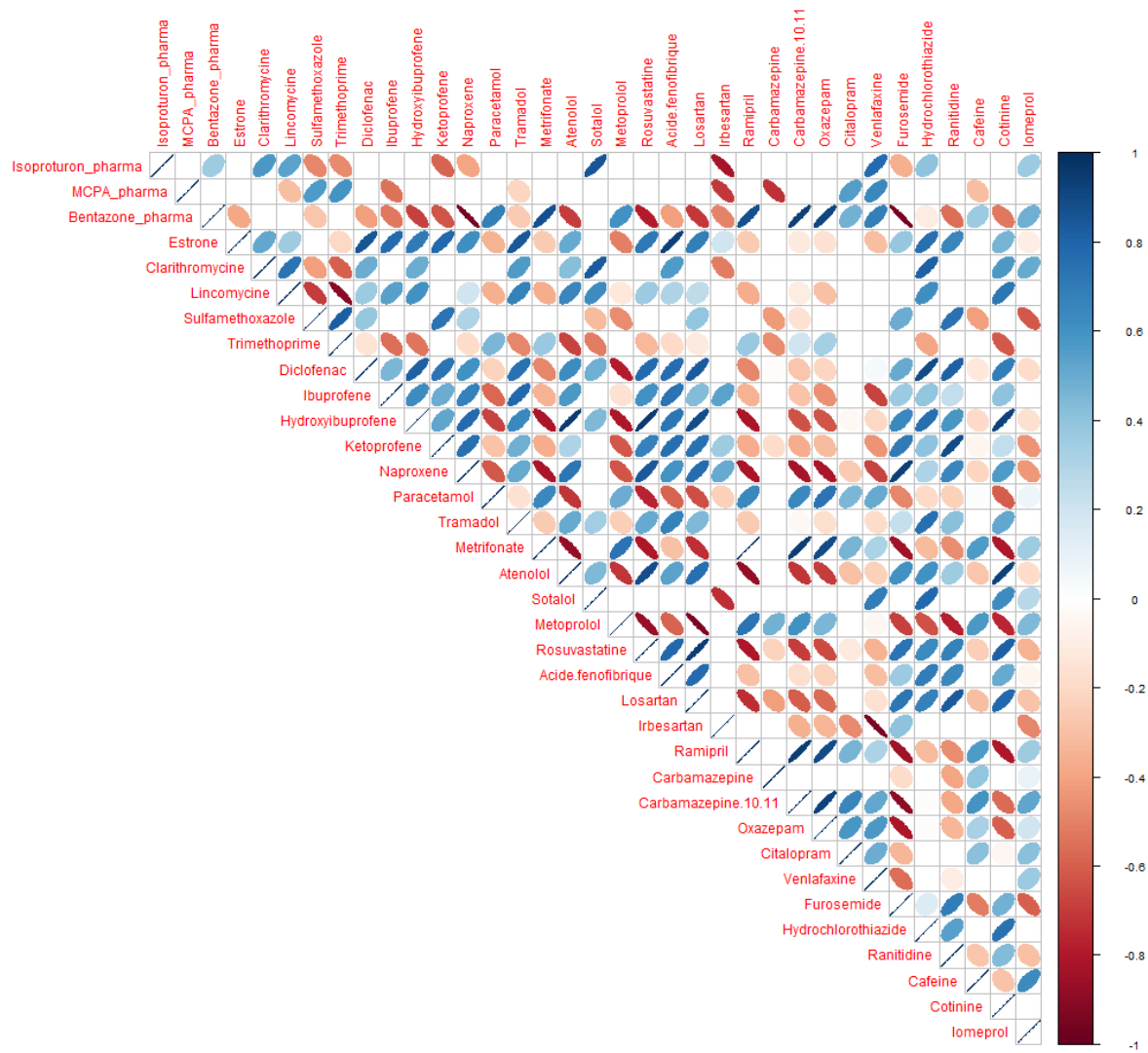


FIGURE 4 – Matrice de corrélation entre composés pour les concentrations mesurées à une fréquence journalière de Basècles

C Graphiques des séries temporelles

