**Supplementary material for:**

**Spatial variation in the detection rates of frequently studied pharmaceuticals in Asian, European and North American rivers**

*Yusuf Jameel1, a\*,**Denis R Valle1 and Paul Kay2*

1 School of Forest Resources and Conservation, University of Florida, Gainesville, Florida, USA.

2 School of Geography, University of Leeds, Leeds, West Yorkshire LS2 9JT, U.K.

a Now at Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA.

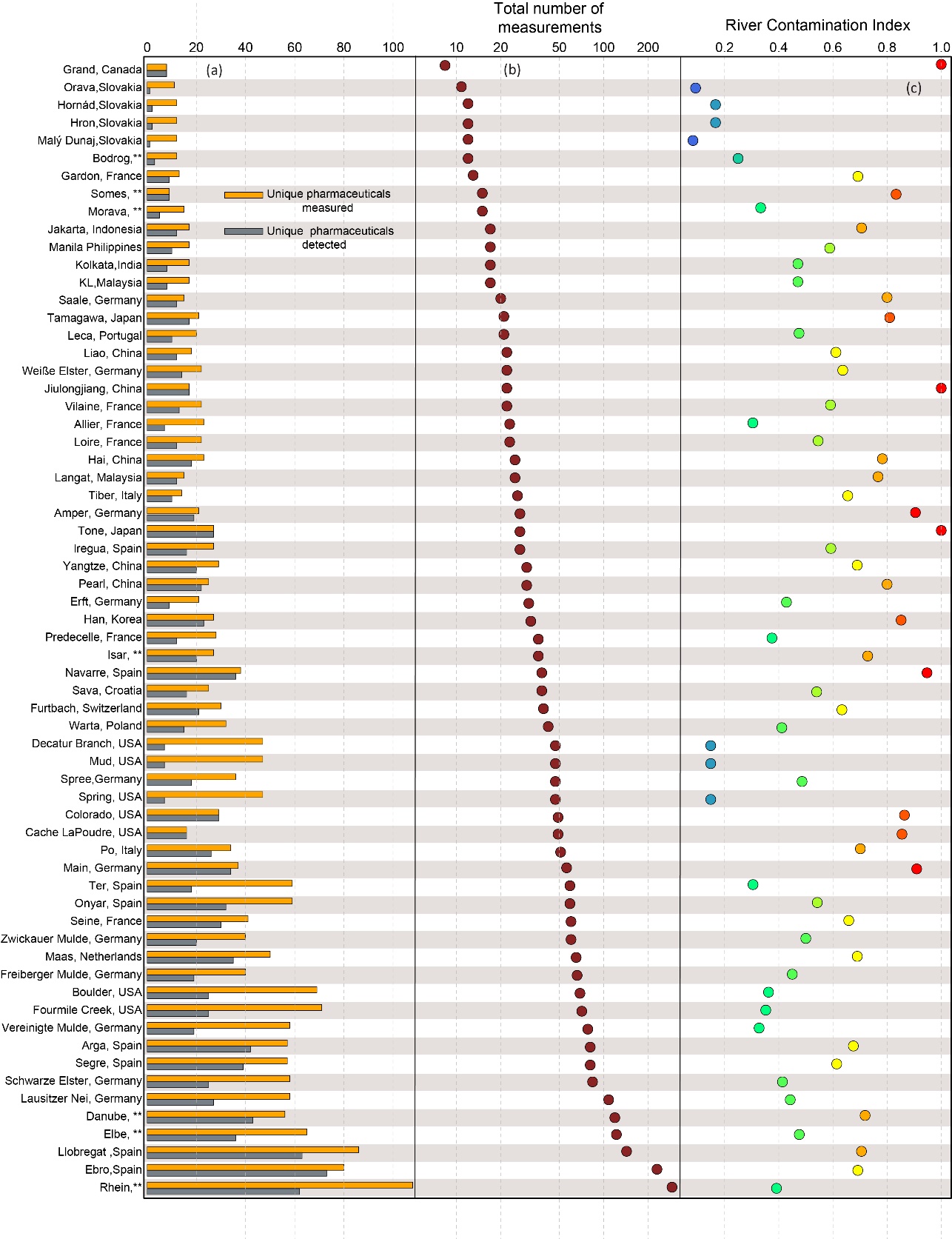
\*Corresponding author: yjameel@mit.edu

Summary:

22 pages

5 Figures

5 Tables

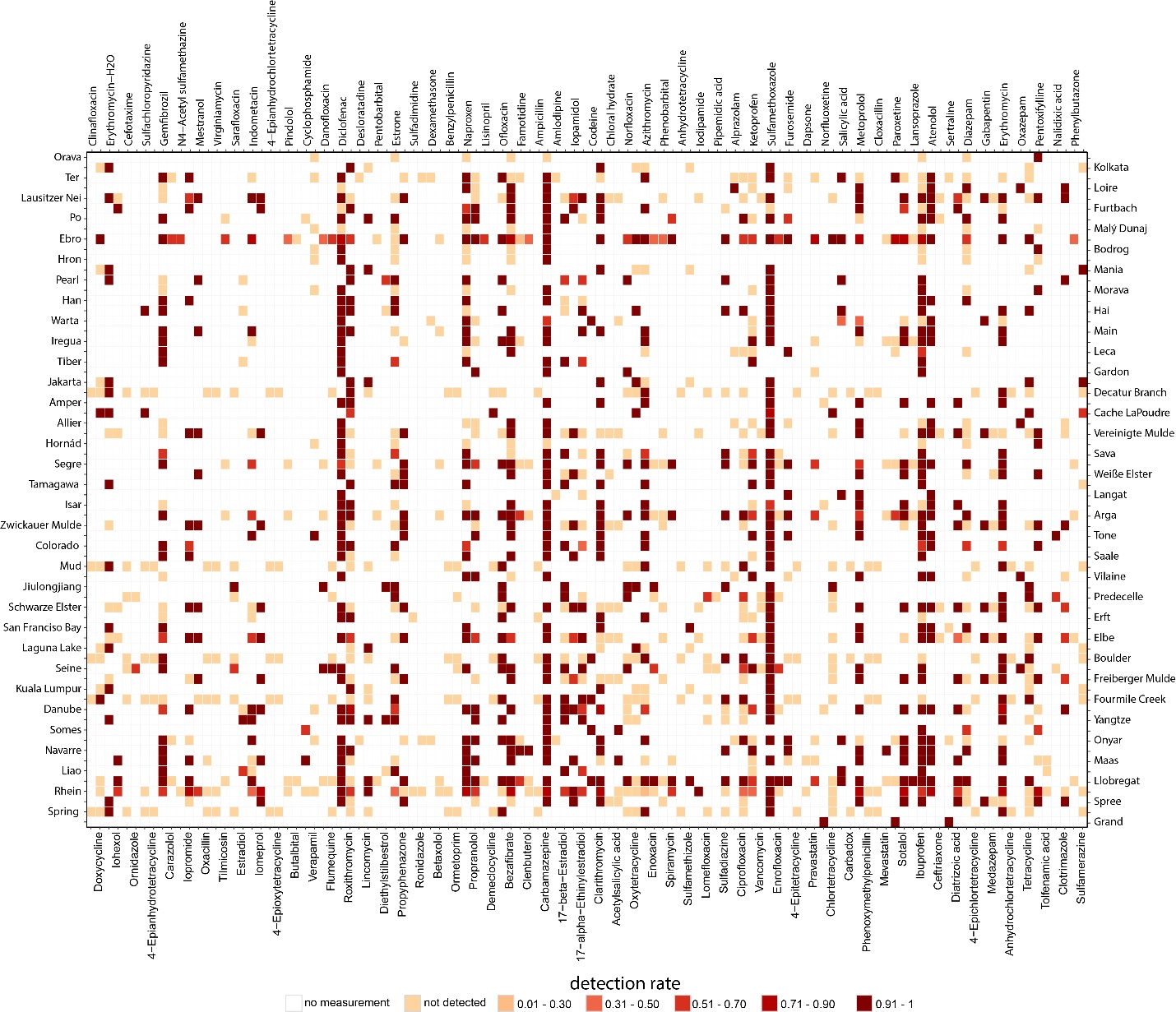
**

*Figure S2. (a): Number of unique pharmaceutical measured and detected at the 64 rivers included in this analysis. (b): Total number of pharmaceuticals measured at each river. As some pharmaceuticals were measured multiple times, this value is greater than the total number of unique pharmaceuticals. (c): River contamination index of each river (higher value means greater contamination). \*\* represents rivers flowing through multiple countries.*

A screenshot of a cell phone

Description automatically generated

*Figure S2. Total number of measurements of the 112 pharmaceuticals (columns) studied across the 64 rivers (rows). White squares represent pharmaceuticals that were not studied at that river. The matrix has been ordered according to the river-pharmaceutical block. Rows 1-9, 10-21, 22-35, 36-60 and 61-64 represents river groups 1,2,3,4, and 5 (partitioned by magenta lines). Columns 1-67, 68-81, 82-85, 86-91, 92-103 and 104-112 represents pharmaceutical clusters A, B, C, D, E and F (partitioned by blue lines). Each rectangle enclosed by the magenta and blue lines is a pharmaceutical-river block. For example, the left most bottom rectangle is block “A–1”.*

**

*Figure S3. Detection rate of the 112 pharmaceuticals (columns) studied across the 64 rivers (rows). White squares represent pharmaceuticals that were not studied at that river. This matrix illustrates the apparent lack of pattern in the detection rates of the pharmaceuticals across the rivers before grouping the data using the stochastic block model (SBM).*

**

*Figure S4. Removal rate of 30 pharmaceuticals in clusters A to C and 20 pharmaceuticals in clusters D to F. The data represents the conventional active sludge (CAS) removal rate obtained from (Verlicchi et al., 2012). Removal rate of the remaining pharmaceuticals in clusters A to F were not provided in Verlicchi et al (2012).*

*Table S1: Therapeutic groups of pharmaceuticals in clusters D to F.*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Pharmaceutical** | **Therapeutic group** | **Pharmaceutical cluster** |  | **Pharmaceutical** | **Therapeutic group** | **Pharmaceutical cluster** |
| Azithromycin | Antibiotics | D |  | Naproxen | Analgesics | E |
| Ibuprofen | Analgesics | D |  | Ofloxacin | Antibiotics | E |
| Mestranol | Estrogen | D |  | Propranolol | Beta blockers | E |
| Metoprolol | Beta blockers | D |  | Sotalol | Beta blockers | E |
| Pentoxifylline | Beta blockers | D |  | Atenolol | Beta blockers | F |
| Propyphenazone | Analgesics | D |  | Carbamazepine | Antiepileptic drugs | F |
| Bezafibrate | Lipid-lowering drugs | E |  | Codeine | Morphine derivates | F |
| Clarithromycin | Antibiotics | E |  | Diclofenac | Analgesics | F |
| Diatrizoic acid | Radiocontrast agents | E |  | Gabapentin | Anticonvulsants | F |
| Erythromycin | Antibiotics | E |  | Iomeprol | Radiocontrast agents | F |
| Erythromycin-H2O | Antibiotics | E |  | Oxazepam | Anxiolytics | F |
| Gemfibrozil | Lipid-lowering drugs | E |  | Salicylic acid | Natural product | F |
| Iopamidol | Radiocontrast agents | E |  | Sulfamethoxazole | Antibiotics | F |
| Iopromide | Radiocontrast agents | E |  |  |  |  |

*Table S2: Therapeutic groups of pharmaceuticals in clusters A to C.*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Pharmaceutical** | **Therapeutic group** | **Pharmaceutical cluster** |  | **Pharmaceutical** | **Therapeutic group** | **Pharmaceutical cluster** |
| 4-Epianhydrochlortetracycline | Antibiotics | A |  | Ormetoprim | Folic acid antagonist | A |
| 4-Epianhydrotetracycline | Antibiotics | A |  | Ornidazole | Antibiotics | A |
| 4-Epichlortetracycline | Antibiotics | A |  | Oxacillin | Antibiotics | A |
| 4-Epioxytetracycline | Antibiotics | A |  | Paroxetine | Antidepressant | A |
| 4-Epitetracycline | Antibiotics | A |  | Pentobarbital | Barbiturates | A |
| Alprazolam | Antidepressant | A |  | Phenobarbital | Anticonvulsants | A |
| Amlodipine | Calcium Channel Blocker | A |  | Phenoxymethylpenicillin | Antibiotics | A |
| Ampicillin | Antibiotics | A |  | Phenylbutazone | NSAID | A |
| Anhydrochlortetracycline | Antibiotics | A |  | Pindolol | Beta blockers | A |
| Anhydrotetracycline | Antibiotics | A |  | Pipemidic acid | Antibiotics | A |
| Benzylpenicillin | Antibiotics | A |  | Pravastatin | Lipid-lowering drugs | A |
| Betaxolol | Beta blockers | A |  | Ronidazole | Antibiotics | A |
| Butalbital | Analgesics | A |  | Sarafloxacin | Antibiotics | A |
| Carazolol | Beta blockers | A |  | Sertraline | Antidepressant | A |
| Carbadox | Antibiotics | A |  | Sulfachloropyridazine | Antibiotics | A |
| Cefotaxime | Antibiotics | A |  | Sulfadimidine | Antibiotics | A |
| Ceftriaxone | Antibiotics | A |  | Sulfamerazine | Antibiotics | A |
| Chloral hydrate | Hypnotics | A |  | Sulfamethizole | Antibiotics | A |
| Chlortetracycline | Antibiotics | A |  | Tilmicosin | Antibiotics | A |
| Clenbuterol | Anti-asthmatics | A |  | Tolfenamic acid | Anti-inflammatory drugs | A |
| Clinafloxacin | Antibiotics | A |  | Vancomycin | Antibiotics | A |
| Cloxacillin | Antibiotics | A |  | Verapamil | Antihypertensives | A |
| Cyclophosphamide | Chemotherapeutic agents | A |  | Virginiamycin | Antibiotics | A |
| Danofloxacin | Antibiotics | A |  | 17-beta-Estradiol | Estrogen | B |
| Dapsone | Antibiotics | A |  | Acetylsalicylic acid | Analgesics | B |
| Demeclocycline | Antibiotics | A |  | Ciprofloxacin | Antibiotics | B |
| Desloratadine | Antihistamine | A |  | Diazepam | Psychiatric medication | B |
| Dexamethasone | Anti-inflammatory drugs | A |  | Estrone | Estrogen | B |
| Diethylstilbestrol | Estrogen | A |  | Furosemide | Diuretics | B |
| Doxycycline | Antibiotics | A |  | Iodipamide | Radiocontrast agents | B |
| Enoxacin | Antibiotics | A |  | Iohexol | Radiocontrast agents | B |
| Enrofloxacin | Antibiotics | A |  | Lincomycin | Antibiotics | B |
| Estradiol | Estrogen | A |  | Norfloxacin | Antibiotics | B |
| Famotidine | Antihistamine | A |  | Oxytetracycline | Antibiotics | B |
| Flumequine | Antibiotics | A |  | Roxithromycin | Antibiotics | B |
| Lansoprazole | Antacids | A |  | Spiramycin | Antibiotics | B |
| Lisinopril | Antihypertensives | A |  | Sulfadiazine | Antibiotics | B |
| Lomefloxacin | Antibiotics | A |  | Tetracycline | Antibiotics | B |
| Medazepam | Sedatives | A |  | 17-alpha-Ethinylestradiol | Estrogen | C |
| Mevastatin | Lipid-lowering drugs | A |  | Clotrimazole | Antifungal medication | C |
| N4-Acetyl sulfamethazine | Antibiotics | A |  | Indometacin | Analgesics | C |
| Nalidixic acid | Antibiotics | A |  | Ketoprofen | Analgesics | C |
| Norfluoxetine | Antidepressant | A |  |  |  |  |

*Table S3: Median value of detection limit (ng/l) for the 38 pharmaceuticals with more than 20 unique measurements. N is the number of measurements for each pharmaceutical used to calculate the median value of the detection limit. R is the ratio between the mean observed concentration and mean detection limit for each pharmaceutical respectively.*

|  |  |  |  |
| --- | --- | --- | --- |
| Name of the pharmaceutical | Median | N | R |
| 17-alpha-Ethinylestradiol | 0.19 | 37 | 32.6 |
| 17-beta-Estradiol | 0.20 | 33 | 71.8 |
| Atenolol | 0.02 | 37 | 113.2 |
| Azithromycin | 0.14 | 35 | 63.5 |
| Bezafibrate | 0.08 | 58 | 53.2 |
| Carbamazepine | 0.10 | 79 | 47.9 |
| Ciprofloxacin | 0.10 | 26 | 6.6 |
| Clarithromycin | 0.10 | 37 | 38.7 |
| Clofibric acid | 0.03 | 41 | 0.6 |
| Clotrimazole | 0.01 | 36 | 0.01 |
| Diatrizoic acid | 0.02 | 25 | 2.8 |
| Diazepam | 0.03 | 37 | 19.8 |
| Diclofenac | 0.05 | 98 | 105.3 |
| Erythromycin | 0.01 | 35 | 94.5 |
| Erythromycin-H2O | 0.10 | 25 | 112.0 |
| Estriol | 0.10 | 35 | 9.4 |
| Estrone | 0.17 | 54 | 1.7 |
| Fluoxetine | 0.02 | 32 | 4.2 |
| Furosemide | 0.01 | 21 | 79.7 |
| Gemfibrozil | 0.02 | 51 | 83.8 |
| Ibuprofen | 0.01 | 104 | 76.6 |
| Indometacin | 0.02 | 21 | 38.5 |
| Ketoprofen | 0.10 | 49 | 41.9 |
| Metoprolol | 0.03 | 44 | 52.5 |
| Naproxen | 0.03 | 63 | 54.3 |
| Ofloxacin | 0.02 | 25 | 90.7 |
| Paracetamol | 0.30 | 36 | 188.8 |
| Phenazone | 0.01 | 26 | 38.5 |
| Propranolol | 0.01 | 37 | 53.9 |
| Roxithromycin | 0.01 | 37 | 81.7 |
| Sotalol | 0.01 | 28 | 239.9 |
| Sulfadimethoxine | 0.10 | 27 | 10.6 |
| Sulfamethazine | 0.10 | 35 | 281.5 |
| Sulfamethoxazole | 0.30 | 86 | 76.1 |
| Sulfathiazole | 0.80 | 21 | 16.0 |
| Tetracycline | 1.00 | 30 | 13.1 |
| Trimethoprim | 0.28 | 58 | 38.1 |
| Tylosin | 0.25 | 23 | 1.1 |

**Data sub-setting and aggregation**

Below we describe the procedure/steps followed for obtaining the data used in our analysis. To obtain the surface water pharmaceutical data, the first step consisted of sub-setting justsurface water data by only including observations for which the “environmental matrix” was described as "Surface Water - River/Stream. We subsequently chose just the observations for which “Sampling Locations” was not blank (i.e., this field included the name of the river from where the sample was collected). This was done because for many surface water data included in the database, the sampling location was not provided. Finally, we aggregate these data by calculating the maximum value stored in the “MEC original” column for each river, study (column “Literature Citation”), and pharmaceutical (column “Name of Analyte”). The maximum value was 0 if the pharmaceutical under study is “below detection”.

Our algorithm resulted in a dataset with 3770 unique entries, where each row contained information on the literature from which the entry was obtained, the name of the surface water system, the pharmaceutical measured at the site and its maximum concentration. We removed pharmaceuticals that were measured only in 3 or less rivers. Similarly, rivers where there were very few pharmaceutical measurements (< 10) were also removed. This was done to remove pharmaceuticals that were rarely measured and exclude rivers with very limited measurements. Following this, the dataset consisted of 214 unique pharmaceutical measurements measured across 96 rivers. We summarized the dataset in two large matrices (matrix S1 and matrix S2, included in supplementary material) with the rows containing information on each river and the columns containing information on each pharmaceutical. Each cell *(i, j)* of matrix S1 contained information on the total number of times pharmaceutical *j* was measured in river *i*. If a given pharmaceutical *j* was not measured for river *i*, the corresponding cell value was 0. Similarly, each cell *(i, j)* of matrix S2 contained information on the number of times pharmaceutical *j* was positively observed for river *i*. The rationale behind presenting the results for only 64 rivers and 112 pharmaceuticals is discussed in the next section.

To obtain the WWTP pharmaceutical data, we followed a similar procedure as described above and only include information for which the “environmental matrix” was described as "WWTP”. We only included those countries from where the rivers included in the surface water dataset originated.

By extracting the maximum value for of a pharmaceutical measured on a river for a given study, we lose the temporal component of the study. However, in our analysis the number of studies with multiple measurements for a pharmaceutical on a river were limited. Indeed 42% of the measurements included in this analysis were one-time and single-site measurement, around 25% of the measurements included 2 measurements, majority of the samples in these cases were upstream and downstream sites around a WWTP but were collected concurrently. Another 23% of the samples included between 3 to 10 measurements and in these cases also temporal resolution was limited as the multiple measurements were made at sites along the river. Only, 10% of the samples had more than 10 measurements for a pharmaceutical measured on a river and most of these sample were collected by a network of reference laboratories, research centers and related organizations for monitoring of emerging environmental substances (NORMAN, https://www.norman-network.net/?q=node/125) that measured pharmaceuticals in rivers across Europe from 2002 to 2010. By extracting the “maximum concentration”, we do lose the temporal resolution and also likely risk obtaining information from base flow conditions as the highest concentrations are expected during base flow. If we were using measured concentration in our analysis, it would have indeed been problematic as we would have been using concentration values in our analysis that we most likely an overestimating of the mean concentration. This or similar work would benefit from a clearer understanding of the relationship between flow and pharms but at present this doesn’t exist. In terms of model output, since the NORMAN dataset constituted less than 10% of our subsetted data, their impact on our interpretations is most likely negligible.

Another major challenge by combining existing datasets arises due to variable reporting limits reported by different studies for a given pharmaceutical. It is possible that a pharmaceutical was undetected in a river because the method employed had a higher detection limit whereas the same pharmaceutical with similar concentration in another river was detected because the method used had a lower detection limit. To understand effect of differences in detection limit across studies, we calculated the range of detection limit for all the pharmaceuticals with greater than 20 measurements. We observed that the median detection limit was low for most of the pharmaceuticals (Table S3). Further, the ratio between mean observed concentration and mean detection limit for each pharmaceutical was big suggesting that the observed concentration for a pharmaceutical were many times higher than the method detection limits.

**Description of the stochastic block model**

Let be the number of measurements on river *i* for pharmaceutical *j* and be the subset of these measurements for which pharmaceutical *j* was positively detected. The stochastic block model (SBM) assumes that follows a binomial distribution with total number of trials given by and with success probability given by . The subscripts for this parameter depend on the group membership of river *i* and pharmaceutical *j*. More specifically, we assume that:

where is a latent variable indicating the group membership of river *i*, is a latent variable indicating the group membership of pharmaceutical *j*, and is the probability associated with ‘river group *q*’ and ‘pharmaceutical group *r*’. To avoid confusion with river groups, we refer pharmaceutical groups as ‘pharmaceutical clusters’ from hereon.

We assume that the latent river group membership is given by:

where is a vector of probabilities that sum to one of the assignment of river *i* to each group. Similarly, we assume that the latent pharmaceutical cluster membership is given by:

where is a vector of probabilities that sum to one of the assignment of pharmaceutical j to each cluster.

In traditional clustering algorithms, the appropriate number of groups/clusters has to be a priori specified. As a result, the standard approach is to systematically vary the number of groups and chose the optimal number of groups using a performance metric (e.g., AIC, BIC, gap statistic, integrated classification likelihood, minimum message length - MML) (Charrad et al., 2015; Tibshirani et al., 2001) Instead of adopting this approach, we relied on Bayesian non-parametric ‘truncated stick-breaking approach’ following (Valle et al., 2018) to determine the optimal number of river groups and pharmaceutical clusters. Following this approach, the priors for and are given by:

for k = 1,…,K−1, > 0 and .

for k = 1,…,K−1, > 0 and .

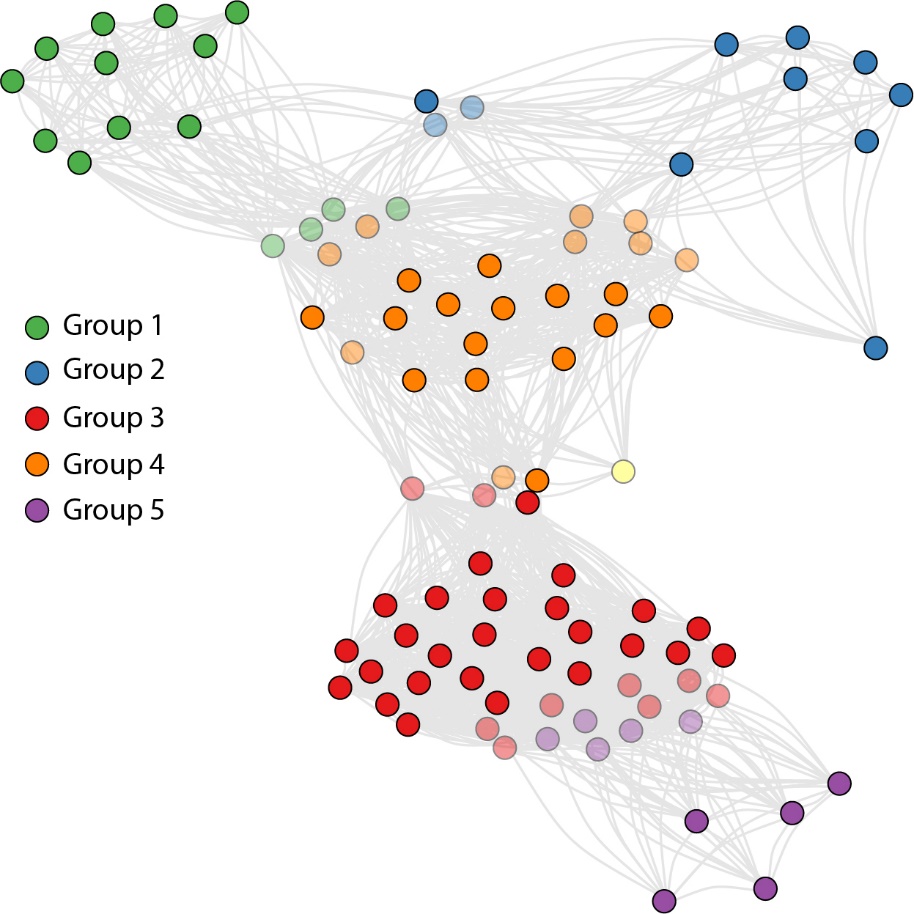
We finish the specification of our model by specifying an uninformative uniform prior for the parameter :

The full conditional distribution of all the parameters have been derived below. We fit the Stochastic Block Model using a Gibbs sampler and ran the model for 7,000 iterations after discarding the first 3,000 iterations as burn-in.

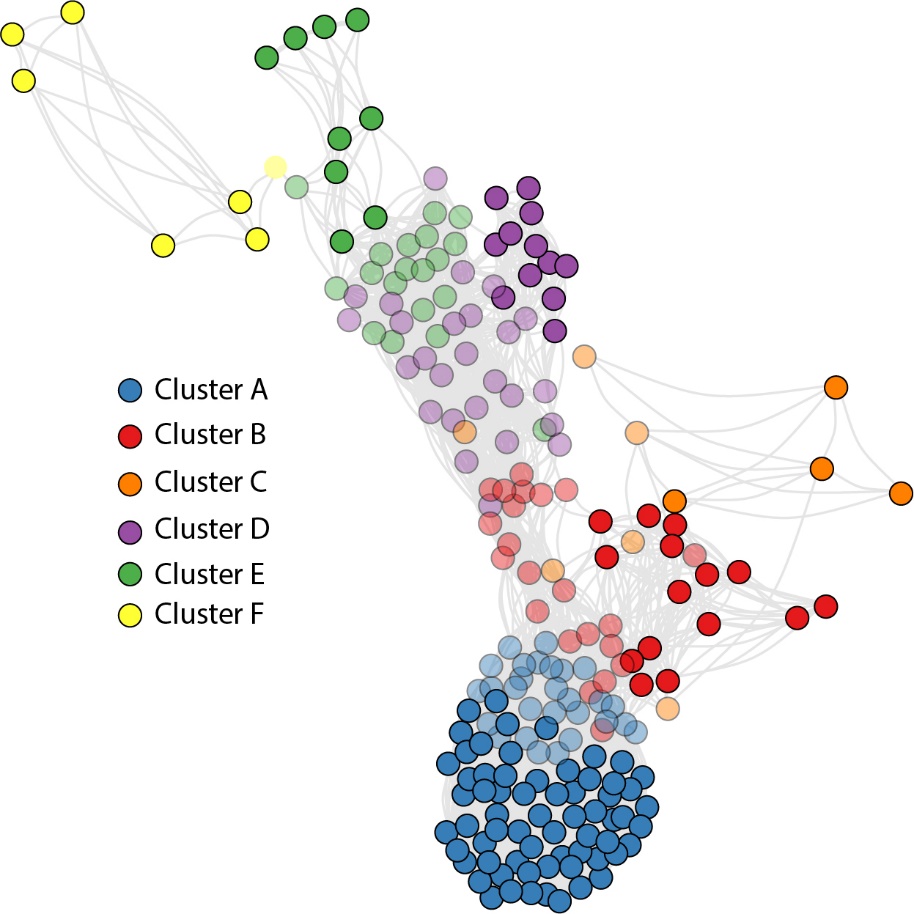
The model outputs the proportion of times each river and each pharmaceutical were assigned to the different groups and clusters, respectively. A typical group assignment for river *i* would look like [0.2, 0.18, 0.55, 0.07, 0, 0] (if there were 6 river groups) suggesting that river *i* was assigned to river groups 1 and 3, 20% and 55% of the times respectively. Similarly, the river was assigned 0% of the times to groups 5 and 6 in all Gibbs iterations. The vector also suggests that this river belongs dominantly to ‘river group 3’. The model also outputs the posterior distribution for , which is the probability of positively detecting a pharmaceutical in ‘cluster *r*’ at a river in ‘group *q*’. If the group assignment for river *j* is [0.2, 0.3, 0.3, 0.15, 0.0.5, 0] then this suggests that river *j,* was assigned to groups 2 and 3 the maximum number of times, however no one group is dominantly assigned to the river. The non-assignment of pharmaceuticals and rivers to a specific pharmaceutical cluster and river group even after multiple Gibbs iteration suggests that they do not share the characteristics of a single pharmaceutical cluster and river group but rather reflect characteristics of multiple groups and clusters. This is possible for rivers and pharmaceuticals that have mean detection rate intermediary between two or more groups (for rivers) and clusters (for pharmaceuticals). Indeed, almost all the river and pharmaceutics that were not assigned specifically to a single group and cluster reflected detection rate that were transitional between two groups and clusters respectively (see Figures S5 and S6).

Here we present results only for rivers (and pharmaceuticals) that were assigned dominantly to a single group. Accordingly, we defined a river to belong to a group exclusively only if the assigned proportion associated with that group is greater than 0.6 (60%) and the assignment proportion associated with the second largest group is less than 0.3 (30%). The same procedure was followed for pharmaceutical clusters. We did this to include only those pharmaceuticals in a group that are very similar to each other in their detection rates and exclude pharmaceuticals that did not belong to that group dominantly. The same was done for rivers where only those rivers were included that were very similar to each other in their detection rates. This allows us to be very confident in our prediction about positively observing a non-measured pharmaceutical for a given river in a group.

Following the criteria described above, only 64 rivers were assigned to a single group and the remaining 30 rivers were not assigned to any single group. Similarly, out of 214 pharmaceuticals in the dataset, only 112 of them were repeatedly assigned to a specific group and the remaining 102 were not assigned to any specific group and we present the result for those rivers. It is worth mentioning that if the groups assigned to a river and cluster assigned to a pharmaceutical were based upon the maximum number of assignments, our result would not have changed as all the rivers/pharmaceuticals included in our analysis already belong to a group/cluster from which they have the highest assignments (greater than 60%). However, if we were allocating groups/clusters based upon maximum number of allocations, we would have also included additional rivers/pharmaceuticals in the group/cluster that might exhibit mixed membership. For example, if the probability of river *i* belonging to groups 1 to 6 were [0.40, 0.38, 0.06, 0.10, 0.04, 0.02], then the river *i* would be assigned to group1 (following the selection criteria of highest assignment to a group), however, the likelihood of it belonging to group 2 is similar and just fractionally lower (0.38 for group 2 and 0.40 for group 1). This suggests that this river exhibits a detection rate in-between of the mean of group 1 and group 2 (see the lighter shade circles in figure S5) and hence was excluded.



*Figure S5. Similarity between the 96 rivers included in the original analysis plotted following the Fruchterman-Reingold algorithm that plots similar river closer to each other and dissimilar rivers further apart. Here, similar rivers are defined as rivers with comparable detection rates. The 64 rivers included in and 32 rivers excluded from the final analysis are shown by darker and lighter shades respectively. Clearly, rivers that were excluded from the final analysis (lighter shades) exhibited detection rates in-between of two groups. The distance between two groups shows the similarity/dissimilarity between them. It is also evident from the plot that the largest distance is between group 5 (most contaminated river group) and group 1 (least contaminated river group) suggesting that these are the most dissimilar groups which is indeed the case.*



*Figure S6. Similarity between the 214 pharmaceuticals included in the original analysis plotted following the Fruchterman-Reingold algorithm that plots similar pharmaceuticals closer to each other and dissimilar pharmaceuticals further apart. Here, similar pharmaceuticals are defined as pharmaceuticals with comparable detection rates. The 112 pharmaceuticals included in and 102 pharmaceuticals excluded from the final analysis are shown by darker and lighter shades respectively. Clearly, pharmaceuticals that were excluded from the final analysis (lighter shades) exhibited detection rates in-between of two clusters. The distance between two clusters shows the similarity/dissimilarity between them. It is evident from the plot that the largest distance is between the most contaminated pharmaceutical cluster (cluster F) and the least contaminated pharmaceutical cluster (cluster A) suggesting that these are the most dissimilar clusters.*

**Full conditional distributions**

To create our Gibbs sampler, we need to derive the full conditional distributions for all the parameters and latent variable in our model.

* For

Because there is a finite number of clusters, this implies that:

We draw this variable from a multinomial distribution with probability vector given by the above expression.

* For

Because there is a finite number of clusters, this implies that:

We draw this variable from a multinomial distribution with probability vector given by the above expression.

* For

where and

This implies that can be sampled from the following beta distribution:

* For

Where and

This implies that can be sampled from the following beta distribution:

* For

where and . This implies that can be sampled from the following beta distribution:

**Model implementation**

We used R statistical software to run the model. The code can be downloaded from https://github.com/drvalle1/git\_pharma\_sbm. The matrix S1 and S2 containing the information on the total number of times pharmaceutical *j* was measured in river *i* and the number of times pharmaceutical *j* was positively observed for river *i* is included in the supporting information*.*  The rivers and pharmaceuticals are sorted alphabetically in matrices S1 and S2. Tables S4 and S5 provides the actual name of each river and pharmaceutical used in the matrices.We urge the users to install all the dependencies and packages before running the code. Please contact Denis Valle (drvalle@ufl.edu) or Yusuf Jameel (m.jameel@ufl.edu) regarding questions on the model and application of the model to similar datasets.

*Table S4: Numeric code of the rivers in matrices S1 and S2 and their corresponding names.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| River code | Name of the river | | River code | Name of the river |
| River2 | Allier |  | River64 | Main |
| River3 | Amper |  | River65 | Malý Dunaj |
| River4 | Arga |  | River67 | Mania |
| River5 | Atibaia |  | River68 | Mekong |
| River6 | Aura |  | River69 | Mess |
| River8 | Bodrog |  | River70 | Moine |
| River9 | Boulder |  | River71 | Morava |
| River10 | Cache LaPoudre |  | River72 | Mud |
| River13 | Clain |  | River73 | Navarre |
| River14 | Colorado |  | River74 | Nitra |
| River16 | Danube |  | River75 | Onyar |
| River17 | Dead Horse |  | River76 | Orava |
| River18 | Decatur Branch |  | River77 | Paraiba do Sul |
| River20 | Douro |  | River78 | Pearl |
| River21 | Ebrach |  | River79 | Petrusse |
| River22 | Ebro |  | River81 | Po |
| River23 | Elbe |  | River82 | Predecelle |
| River24 | Ely |  | River84 | Rhein |
| River25 | Erft |  | River86 | Ruhr |
| River26 | Fourmile Creek |  | River87 | Saale |
| River27 | Freiberger Mulde |  | River88 | San Franciso Bay |
| River28 | Furtbach |  | River89 | Sava |
| River29 | Gardon |  | River90 | Schelde |
| River30 | Grand |  | River91 | Schwarzach |
| River32 | Hai |  | River92 | Schwarze Elster |
| River33 | Halifax |  | River95 | Segre |
| River34 | Han |  | River96 | Seine |
| River35 | Hau |  | River97 | Somes |
| River36 | Hoje |  | River98 | Spree |
| River37 | Holtemme |  | River99 | Spring |
| River38 | Hornád |  | River100 | Strengbach |
| River39 | Hron |  | River101 | Svratka |
| River41 | Iregua |  | River102 | Taff |
| River42 | Isar |  | River103 | Tamagawa |
| River43 | Jakarta |  | River104 | Ter |
| River44 | Jiulongjiang |  | River105 | Tiber |
| River46 | Kolkata |  | River106 | Tisza |
| River47 | Kuala Lumpur |  | River107 | Tone |
| River49 | Laguna Lake |  | River108 | Vah |
| River51 | Lake Ontario |  | River109 | Vereinigte Mulde |
| River52 | Langat |  | River110 | Vienne |
| River53 | Lausitzer Nei |  | River111 | Vilaine |
| River54 | Leca |  | River112 | Vistula |
| River55 | Liao |  | River113 | Warta |
| River57 | Llobregat |  | River115 | Weiße Elster |
| River59 | Loire |  | River117 | Würm |
| River62 | Maas |  | River119 | Yangtze |
| River63 | Mackinaw |  | River120 | Zwickauer Mulde |

*Table S5: Numeric code of the pharmaceuticals in matrices S1 and S2 and their corresponding names.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Pharma code | Name of the pharmaceutical |  | Pharma code | Name of the pharmaceutical |
| spp2 | 10,11-Dihydro-10,11-Dihydroxy Carbamazepine |  | spp184 | Iotaminic acid |
| spp3 | 10,11-Dihydro-10,11-Epoxycarbamazepine |  | spp186 | Ioxitalminic acid |
| spp4 | 17-alpha-Estradiol |  | spp188 | Isochlortetracycline |
| spp6 | 17-alpha-Ethinylestradiol |  | spp190 | Josamycin |
| spp8 | 17-beta-Estradiol |  | spp192 | Ketoprofen |
| spp10 | 2-Ethyl-2-phenylmalonamide |  | spp193 | Lansoprazole |
| spp12 | 4-Acetaminoantipyrine |  | spp196 | Levonorgestrel |
| spp14 | 4-Epianhydrochlortetracycline |  | spp197 | Lidocaine |
| spp15 | 4-Epianhydrotetracycline |  | spp198 | Lincomycin |
| aspp16 | 4-Epichlortetracycline |  | spp199 | Lisinopril |
| spp17 | 4-Epioxytetracycline |  | spp200 | Lomefloxacin |
| spp18 | 4-Epitetracycline |  | spp201 | Loratadine |
| spp20 | 4-Hydroxydiclofenac |  | spp202 | Lorazepam |
| spp27 | Acetyl-sulfamethoxazole |  | spp206 | Meclocycline |
| spp28 | Acetylsalicylic acid |  | spp208 | Medazepam |
| spp31 | Acyclovir |  | spp209 | Mefenamic acid |
| spp33 | Alpha-apo-oxytetracycline |  | spp212 | Meprobamate |
| spp35 | Alprazolam |  | spp213 | Mestranol |
| spp37 | Amitriptyline |  | spp215 | Metformin |
| spp38 | Amlodipine |  | spp219 | Metoprolol |
| spp39 | Amoxicillin |  | spp220 | Metronidazole |
| spp40 | Amphetamine |  | spp221 | Mevastatin |
| spp41 | Ampicillin |  | spp222 | Miconazole |
| spp43 | Anhydrochlortetracycline |  | spp223 | Minocycline |
| spp44 | Anhydroerythromycin |  | spp224 | Monensin |
| spp45 | Anhydrotetracycline |  | spp228 | N-formyl-4-aminoantipyrine |
| spp46 | Atenolol |  | spp229 | N4-Acetyl sulfamethazine |
| spp47 | Atorvastatin |  | spp234 | Nadolol |
| spp50 | Azithromycin |  | spp236 | Nalidixic acid |
| spp54 | Benzylpenicillin |  | spp237 | Naproxen |
| spp56 | Betaxolol |  | spp245 | Norfloxacin |
| spp57 | Bezafibrate |  | spp247 | Norfluoxetine |
| spp58 | Bisoprolol |  | spp249 | Nystatin |
| spp59 | Bromazepam |  | spp252 | Ofloxacin |
| spp62 | Butalbital |  | spp253 | Olanzapine |
| spp64 | Carazolol |  | spp257 | Ormetoprim |
| spp65 | Carbadox |  | spp258 | Ornidazole |
| spp66 | Carbamazepine |  | spp259 | Oseltamivir |
| spp67 | Carbamazepine-2OH |  | spp260 | Oseltamivir carboxylic acid |
| spp69 | Carboxyibuprofen |  | spp261 | Oxacillin |
| spp71 | Cefalexin |  | spp262 | Oxazepam |
| spp72 | Cefotaxime |  | spp264 | Oxcarbazepine |
| spp73 | Ceftriaxone |  | spp265 | Oxolinic acid |
| spp74 | Cefuroxime |  | spp267 | Oxytetracycline |
| spp76 | Chloral hydrate |  | spp269 | Paracetamol |
| spp77 | Chloramphenicol |  | spp270 | Paroxetine |
| spp81 | Chlortetracycline |  | spp273 | Pentobarbital |
| spp82 | Cimetidine |  | spp274 | Pentoxifylline |
| spp83 | Ciprofloxacin |  | spp277 | Phenazone |
| spp85 | Citalopram |  | spp278 | Phenobarbital |
| spp86 | Clarithromycin |  | spp279 | Phenoxymethylpenicillin |
| spp87 | Clenbuterol |  | spp280 | Phenylbutazone |
| spp88 | Clinafloxacin |  | spp281 | Phenytoin |
| spp89 | Clindamycin |  | spp282 | Pindolol |
| spp90 | Clofibrate ethyl |  | spp284 | Pipemidic acid |
| spp91 | Clofibric acid |  | spp285 | Piperacillin |
| spp93 | Clotrimazole |  | spp288 | Pravastatin |
| spp94 | Cloxacillin |  | spp290 | Primidone |
| spp96 | Codeine |  | spp292 | Progesterone |
| spp98 | Cyclophosphamide |  | spp293 | Propranolol |
| spp101 | Danofloxacin |  | spp295 | Propyphenazone |
| spp102 | Dapsone |  | spp298 | Ranitidine |
| spp103 | Dehydronifedipine |  | spp301 | Ronidazole |
| spp104 | Demeclocycline |  | spp302 | Roxithromycin |
| spp105 | Desloratadine |  | spp303 | Salbutamol |
| spp110 | Dexamethasone |  | spp304 | Salicylic acid |
| spp112 | Diatrizoic acid |  | spp307 | Sarafloxacin |
| spp113 | Diazepam |  | spp308 | Sertraline |
| spp114 | Diazinon |  | spp310 | Simvastatin |
| spp115 | Diclofenac |  | spp311 | Sotalol |
| spp117 | Diethylstilbestrol |  | spp312 | Spiramycin |
| spp120 | Diltiazem |  | spp317 | Sulfachloropyridazine |
| spp122 | Diphenhydramine |  | spp319 | Sulfadiazine |
| spp125 | Doxycycline |  | spp321 | Sulfadimethoxine |
| spp128 | Enalapril |  | spp322 | Sulfadimidine |
| spp130 | Enoxacin |  | spp327 | Sulfamerazine |
| spp131 | Enrofloxacin |  | spp328 | sulfamethazine |
| spp132 | Epi-iso-chlorotetracycline |  | spp329 | Sulfamethazine |
| spp135 | Erythromycin |  | spp330 | Sulfamethazine-n4-acetyl |
| spp136 | Erythromycin-H2O |  | spp332 | Sulfamethizole |
| spp137 | Estradiol |  | spp333 | Sulfamethoxazole |
| spp140 | Estriol |  | spp334 | Sulfamethoxypyridazine |
| spp142 | Estrone |  | spp341 | Sulfapyridine |
| spp148 | Famotidine |  | spp346 | Sulfathiazole |
| spp149 | Fenofibrate |  | spp350 | Sulfisoxazole |
| spp150 | Fenofibric acid |  | spp352 | Tamoxifen |
| spp151 | Fenoprofen |  | spp356 | Testosterone |
| spp156 | Flumequine |  | spp357 | Tetracycline |
| spp157 | Fluoxetine |  | spp359 | Thiabendazole |
| spp160 | Furosemide |  | spp360 | Tiamulin |
| spp161 | Gabapentin |  | spp361 | Tilmicosin |
| spp162 | Gemfibrozil |  | spp362 | Timolol |
| spp163 | Gentamicin |  | spp363 | Tolfenamic acid |
| spp164 | Glibenclamide |  | spp365 | Tramadol |
| spp167 | Guanylurea |  | spp367 | Triamterene |
| spp169 | Hydrochlorothiazide |  | spp368 | Triclocarban |
| spp170 | Hydrocodone |  | spp369 | Triclosan |
| spp172 | Hydroxyibuprofen |  | spp370 | Trimethoprim |
| spp173 | Ibuprofen |  | spp371 | Tylosin |
| spp175 | Ifosfamide |  | spp372 | Valsartan |
| spp176 | Imipenem |  | spp373 | Vancomycin |
| spp177 | Indometacin |  | spp374 | Venlafaxine |
| spp178 | Iodipamide |  | spp375 | Verapamil |
| spp179 | Iohexol |  | spp376 | Virginiamycin |
| spp180 | Iomeprol |  | spp377 | Warfarin |
| spp181 | Iopamidol |  | spp380 | Zidovudine |
| spp183 | Iopromide |  | spp381 | Zolpidem |

**References**

Charrad, M., Ghazzali, N., Boiteau, V., Maintainer, A.N., 2015. Package “NbClust” Title Determining the Best Number of Clusters in a Data Set Depends R (>= 3.1.0).

Tibshirani, R., Walther, G., Hastie, T., 2001. Estimating the number of clusters in a data set via the gap statistic. J. R. Stat. Soc. Ser. B (Statistical Methodol. 63, 411–423. https://doi.org/10.1111/1467-9868.00293

Valle, D., Albuquerque, P., Zhao, Q., Barberan, A., Fletcher, R.J., 2018. Extending the Latent Dirichlet Allocation model to presence/absence data: A case study on North American breeding birds and biogeographical shifts expected from climate change. Glob. Chang. Biol. 24, 5560–5572. https://doi.org/10.1111/gcb.14412

Verlicchi, P., Al Aukidy, M., Zambello, E., 2012. Occurrence of pharmaceutical compounds in urban wastewater: Removal, mass load and environmental risk after a secondary treatment-A review. Sci. Total Environ. https://doi.org/10.1016/j.scitotenv.2012.04.028