Ambiente & Água - An Interdisciplinary Journal of Applied Science
ISSN 1980-993X – doi:10.4136/1980-993X
www.ambi-agua.net
E-mail: ambi.agua@gmail.com

Indirect caffeine modeling in an urban river

ARTICLES doi:10.4136/ambi-agua.2810

Received: 04 Nov. 2021; Accepted: 16 Feb. 2022

Luis Otávio Miranda Peixoto[1*]; Luana Mayumi Takahasi Marques[2]; Alinne Mizukawa[1]; Julio Cesar Rodrigues de Azevedo[2]

[1]Departamento de Hidráulica e Saneamento. Universidade Federal do Paraná (UFPR), Avenida Coronel Francisco H. dos Santos, n° 100, Bloco H, CEP: 81530-000, Curitiba, PR, Brazil. E-mail: alimizu@gmail.com
[2]Departamento Acadêmico de Química e Biologia. Universidade Tecnológica do Paraná (UTFPR), Rua Deputado Heitor de Alencar Guimarães, n° 5000, Bloco EC, CEP: 81280-340, Curitiba, PR, Brazil. E-mail: luanam@alunos.utfpr.edu.br, jcrazevedo@hotmail.com
[*]Corresponding author. E-mail: luisotaviopeixoto@gmail.com

ABSTRACT

Caffeine is used worldwide as a chemical tracer to identify anthropic pressures on urban water resources. Nevertheless, its quantification demands great financial investments. This research created a model that would indirectly determine a range of possible caffeine concentrations along an urban river, without the need for extensive laboratory work. The model is based on Canonical Correlation Analysis (CCA), which can correlate two sets of different-sized independent and dependent variables in order to generate a single empirical equation. This equation takes as input the concentrations of ammonia nitrogen and orthophosphate, as well as the total population and the population inhabiting irregular housing areas. From the model's results, it was possible to elaborate a spectrum of possible concentrations of caffeine along the Atuba River (Curitiba-Brazil). The tendency of water quality degradation of this river was also predicted. This model could become a useful preliminary analysis for water resource managers and researchers alike.

Keywords: caffeine, canonical correlation analysis, water quality modeling.

Modelagem indireta de cafeína em um rio urbano

RESUMO

A cafeína é utilizada mundialmente como um traçador químico para identificar pressões antrópicas em recursos hídricos urbanos. No entanto, a sua quantificação demanda grandes investimentos financeiros. Este estudo tem como objetivo criar um modelo que determinaria, indiretamente, uma banda de possíveis concentrações de cafeína ao longo de um rio urbano, sem a necessidade de esforço laboratorial. O modelo se baseia na análise de correlação canônica, que é capaz de correlacionar dois conjuntos de variáveis, dependentes e independentes, de diferentes dimensões e gerar uma equação que resumiria a relação. Esta equação utiliza como entrada as concentrações de nitrogênio ammoniacal e ortofosfato, e como saída a população total, e habitantes em zonas irregulares, assim como a concentração de cafeína. A partir do modelo, foi possível elaborar uma banda de possíveis concentrações de cafeína ao longo do rio. Este modelo possui a capacidade de ser utilizado como uma ferramenta preliminar para gestores e pesquisadores.

Palavras-chave: análise de correlação canônica, cafeína, qualidade das águas superficiais.

1. INTRODUCTION

The rise in urban population in the last decades has put increasing pressures on urban water bodies. These pressures include increased water demand, wastewater influx and pollution discharge onto these systems. The degradation of the water quality of urban rivers has become a challenge faced by society and academia when attempting to supply the population with water with enough quality and quantity (Woodhouse and Muller, 2017; Han *et al.*, 2018).

Caffeine is a stable compound under different environmental conditions. It is also very soluble and not volatile. It may be used as an indication of the discharge of wastewaters into an urban river. Due to these characteristics, several studies have indicated the viability of caffeine (1,3,7 – trimethylxanthine) as a chemical tracer for the identification of domestic wastewater discharge, compared with other microbiological or chemical indicators (Dafouz *et al.*, 2018; Mizukawa *et al.*, 2019). Also, there exists a possible correlation between high concentrations of caffeine and the presence of viral genome in natural waters has been observed (Gourmelon *et al.*, 2010, Sidhu *et al.*, 2013, Kumar *et al.*, 2019, Rimoldi *et al.*, 2020).

The determination of caffeine concentrations in natural waters can be costly and labor intensive. The machinery, the specialized personnel, the laboratory facilities and glassware, the chemical reagents and consumables are neither easy to acquire nor inexpensive (Colim *et al.*, 2019). Therefore, budgetary limitations constrain repeated caffeine analyses. To circumvent this problem, a model which could indirectly estimate caffeine concentrations in an urban river would be a useful tool for water resource managers and researchers, due to lower costs and faster assessment of the water quality situation. A statistical method that could be used in this context and for which can correlate different sets of multiple independent and dependent variables could be correlated and transformed into an equation is the Canonical Correlation Analysis (CCA) (Hotelling, 1936; Malacarne, 2014; Tiyasha *et al.*, 2020).

CCA has seen an increase in use in environmental sciences for its ability to identify and quantify possible associations among groups of different-sized sets of independent and dependent variables. Some examples of these properties: Gershunov *et al*. (2018) have used CCA to quantitatively assess the effect of the precipitation on water quality of coastal waters in the USA; Wei *et al.* (2018) have observed the relationship among 5 types of heavy metal pollution to 9 different traditional water quality parameters and concluded that for the case of the Dongtinghu Lake orthophosphate, E. Coli and the concentration of dissolved organic presented the highest correlation to the heavy metal pollution; Khalil *et al.* (2011) have implemented a CCA-based neural network to model and forecast water quality parameters of the Nile River, Egypt, using the rainfall in 50 different sub catchments of the river as input.

This study aims to present a model that can predict caffeine concentrations along an urban river, using the Atuba River as the study area. The parameters used as inputs for this model are the concentrations of ammonia nitrogen and orthophosphate and the total population (and the population in irregular housing) in this river's basin. The concentrations for the compounds were set as being dependent on the socio economic parameters. Such a model could improve the understanding of the behavior of contaminants along an urban river and its relationship to social characteristics of the population.

2. MATERIAL AND METHODS

2.1. STUDY AREA

The Atuba River Basin is located entirely within the Curitiba Metropolitan Area (CMA) in the State of Paraná (Southern Brazil). It has a total area of 129.94 km^2. Its main course (the

Atuba River) is 21.57 km long. About 562,700 people inhabited this basin at the time of the sampling. This population is spread through 4 different municipalities (Curitiba, Pinhais, Almirante Tamandare, and Colombo).

The Atuba River was chosen for this study for three reasons: 1) The river's basin is totally urbanized; 2) Though it occupies only 0.8% of the CMA's total area, it is responsible for 15% of the CMA's total gross income; 3) It is socially and physically diverse along its main course.

In early April/2019, 20 surface water samples were collected from the river. The locations of the sampling sites are shown in Figure 1**.** They were designated from P1 to P20, by their proximity to the river's spring. The zone of hydrological influence of each sampling site is also presented. These were used to establish the population of influence for each site.

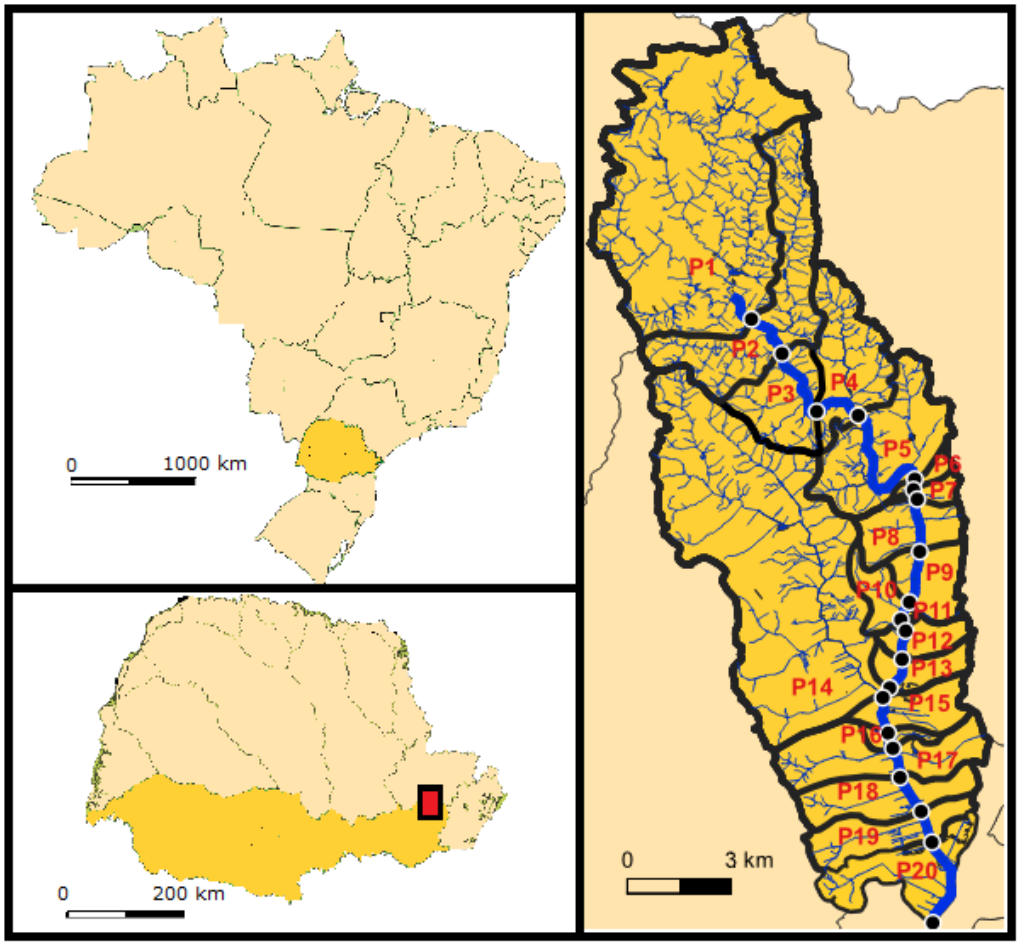

Figure 1. a) Location of the State of Paraná within Brazil; b) Location of the Iguassu River Basin (yellow) and the Atuba River Basin (red) within the State of Parana; c) Location of the sampling sites, and zones of hydrological influence, on the Atuba River Basin.

The socio economic profiles of the inhabitants are unevenly distributed throughout the length of the river. Table 1 presents the distribution of the population for each of the sampling sites, as well as the distribution of the residents that inhabit irregular housing within each zone of hydrological influence.

The socio economic diversities of the profiles within the zones of hydrological influence for the 20 sampling sites may be divided into three specific regions: a) zones for sites P1 to P10 are characterized by sparse regularized horizontal housing; b) zones for sites P10 to P16 are characterized by denser high-income vertical housing; c) zones for sites P17 to P20 are characterized by denser low-income irregular horizontal housing.

Table 1. Distribution of total population and inhabitants is irregular for the zones of hydrological influence of the 20 sampling sites along the Atuba River.

Site	POP	APOP	POPi	APOPi	Site	POP	APOP	POPi	APOPi
P1	58,526	58,526	0	0	**P11**	498	264,423	86	3,780
P2	36,966	95,492	841	841	**P12**	8,768	273,191	263	4,043
P3	21,329	116,821	560	1,401	**P13**	13,068	286,259	194	4,237
P4	35,676	152,497	312	1,713	**P14**	133,831	420,090	2,885	7,122
P5	45,075	197,572	549	2,262	**P15**	15,687	435,777	506	7,628
P6	3,405	200,977	91	2,353	**P16**	8,407	444,184	550	8,178
P7	6,460	207,437	162	2,515	**P17**	43,525	487,709	2,687	10,865
P8	17,912	225,349	716	3,231	**P18**	33,241	520,950	2,943	13,808
P9	19,390	244,739	270	3,501	**P19**	25,320	546,270	4,269	18,077
P10	19,186	263,925	193	3,694	**P20**	16,430	562,700	2,038	20,115

POP = Total inhabitants; APOP = Accumulated Population; POPi = Inhabitants in irregular housing; APOPi = Accumulated population in irregular housing.

Another point of interest for this basin is the existence of the wastewater treatment plant Atuba Sul (WWTP-Atuba Sul). This complex is located between the sampling sites P19 and P20. It is responsible for the treatment of most of the sewage collected in the northern area of the CMA. This facility's design is composed of Upflow Anaerobic Sludge Blanket (UASB) reactors, followed by dissolved air flotation systems. These processes are not designed to efficiently remove ammonia nitrogen.

2.2. SAMPLING

For the determination of the concentration of nutrients (ammonia nitrogen and orthophosphate), twenty samples of surface water of the Atuba River were collected in April 2019. These samples, collected in a 5 L Van Dorn bottle, were stored in twenty 500 mL polyethylene terephthalate (PET) bottles. These bottles had been previously decontaminated by a 5% v/v. HCl solution.

For the caffeine analyses, surface water samples from eight sites were collected (P1, P3, P7, P10, P13, P15, P19, and P20). The samples were stored in 1 L amber bottles until analyzed. These bottles had been decontaminated by a 5% v/v. Extran® detergent solution (Merck Milipore – Darmstadt, Germany).

After collection, all bottles were stored in thermally isolated boxes and immediately conducted to the laboratory.

2.3. CHEMICAL ANALYSES

The analyses for the determination of the concentration of both nutrients (ammonia nitrogen and orthophosphate) were performed in accordance with APHA *et al.* (2012). Orthophosphate concentrations were determined by the molybdate/ascorbic acid colorimetric method. Ammonia nitrogen concentrations were obtained through the nitroprusside/phenol spectrophotometric method.

For the determination of caffeine concentrations, the chromatographic analysis was performed in accordance with the method presented by Ide *et al.* (2013). The surface water samples were filtered through a membrane of cellulose acetate (0.45 µm). These filtered samples proceeded to flow through 6 mL octadecylsilane cartridges (Agilent - Santa Clara, USA). These cartridges had been conditioned by hexane, ethyl acetate and acidified ultra-pure water (pH close to 3) in a velocity that varied from 6 to 8 mL min^{-1}. After being vacuum dried, a solution of acetone and acetonitrile (v/v, 1:1) eluted the cartridges. The extracts were then put in flat-bottomed glass balloons and retro evaporated at 40°C. The contents left in the flat-

bottomed glass balloons were dissolved, yet again, in 1 mL of acetonitrile and passed into a 2 mL vial.

The vials were sent for injection in a liquid chromatograph Agilent (Santa Clara, USA), Model 1260, with a photodiode array. This equipment featured an octadecylsilane column with a length of 250 mm, internal diameter of 4.6 mm, and a particle diameter of 5 µm. Then, 5 µL of sample were injected at a speed of 1.0 $mL.min^{-1}$ on isocratic mode with a composition of 1:1 of ultra-pure water and acetonitrile (HPLC grade purity, ≥ 99.9% – Sigma-Aldrich (St. Louis, USA)).

The monitored wavelength for caffeine was 273 nm and its retention time was 2.4 min. The calibration curves were established with concentrations ranging from 0.02 mg L-1 to 2.0 mg L-1 which resulted in a wide linear range with a regression coefficient (R^2) of 0.9971 with the equivalent equation: 1,742,329x + 67,869. The recovery test was performed with 6 different concentrations of caffeine and resulted in 49.3%±9.3%. Reproducibility and repeatability values of 0.8 and 4.0, respectively, expressed as coefficients of variation, were considered satisfactory (<15%). The repeatability was assessed with the injection of analytes in quintuplicate (three different concentrations) in the same day and reproducibility was assessed with the injection of analytes in quintuplicate (same concentrations as in repeatability) in three different days: 1st day, 7th day and 14th day. The limit of detection (LD) of 12.0 ng.L-1 and limit of quantification (LQ) of 40 ng.L-1, calculated using the standard deviation of at least three blank samples and IC is the inclination of the analytical curve (multiplied by 3 for LD and 10 to LQ).

2.4. MODEL GENERATION

The Pearson coefficient is a tool used to determine the linear correlation between two variables. As the Pearson coefficient is based solely on linear behaviors, non-linearly correlated variables might appear to be uncorrelated. Despite this limitation, the Pearson coefficient is widely used, due to its ease of implementation and its numerically tangible inputs and outputs. The Pearson coefficient is defined as (Equation 1):

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}} \quad (1)$$

The Pearson coefficient is the basis for the Canonical Correlation Analysis method (CCA). The CCA is a statistical method based upon the principles of the Principal Component Analysis. This method aims to explore the dependence between multiple dependent variables (DVs) and independent variables (IVs).

The CCA presents itself as a robust tool to evaluate the relevance of complex relationships between environmental variables (Khalil *et al.*, 2011; Di Felici *et al.*, 2012; Wei *et al.*, 2018; Mangadze *et al.*, 2019).

The method of the CCA was first presented by Hotelling (1936). This process consists of the synthetic creation of two canonical variable matrices (CV). These are composed of two vectors, called $\hat{A}$ and $\hat{B}$. The elements of these vectors, when multiplied by the sample values, would determine the optimal coefficients for the maximal Pearson correlation between the IVs and the DVs. That could be summarized as (Equation 2):

CV_{IV} has the maximum Pearson correlation to CV_{DV}:

$$CV_{IV} = A_1IV_1 + A_2IV_2 \ldots A_nIV_n \quad and \quad CV_{DV} = B_1DV_1 + B_2DV_2 \ldots B_iDV_i \quad (2)$$

CV_{IV} = Coefficients of the canonical variables matrix for the independent variables; CV_{DV} = Coefficients of the canonical variables matrix for the dependent variables; A_n = n^{th} coefficient

of the Â vector; B_i = i^{th} coefficient of the ^B vector; IV_n = n^{th} independent variable; DV_i = i^{th} independent variable.

Therefore, the matrices of coefficients $\hat{A}$ and $\hat{B}$ create a weighted sum of each variable in such a way that the Pearson correlation between the complex CV_{IV} and CV_{DV} is the best possible.

For the determination of these matrices, it is considered the covariance between every single variable, including itself, by means of a covariance matrix (**S**). According to this, the final set of canonical values depend upon the number of the variables (dependent or independent) that have the lowest number of instances (Malacarne, 2014)

After the normalization of the canonical weights to the experimental data, it is possible to approximate the equation $CV_{IV} \approx CV_{DV} + \tau$ (τ is a transposition constant). This approximation enables the creation of an equation capable of interpolating within the experimental data. For the scope of this study, this equation would be capable of determining the quantitative relationship between the socio economic data (total population and population residing in irregular housing) with the concentration of orthophosphate, ammonia nitrogen and caffeine. This quantitative relationship would be turned into an equation.

It is not within the scope of this paper to explain thoroughly the implementation of the CCA. Therefore, the authors point to the didactic works of (Hotelling, 1936; Malacarne, 2014; Rencher, 2002; Ferreira, 2018) for further information on the subject.

2.5. ERROR ANALYSES

To determine the statistical validity of the model, three different procedures were executed, the root mean square error (RMSE), the chi-squared (χ^2) test for association and the determination of the margin of error.

The RMSE is a widely used measure of error of a model. The lower its value, the closer the distribution of the modeled values are to the observed ones. It is given by the following equation (Equation 3).

$$RMSE = \sqrt{\frac{\sum(v_m - v_o)^2}{n}} \quad (3)$$

RMSE = Root Mean Square Error; v_m = value modeled; v_o = value observed; n = nº of samples

One of the ways to test the hypothesis that the correlation/independence between two variables is statistically significant is by the Chi-Squared test for Association. Its statistic is given by the following formula (Equation 4):

$$\chi^2 = \sum \frac{(v_m - v_o)^2}{v_o} \quad (4)$$

χ^2 = **Chi-squared statistic; M = Modeled value; O = observed value.**

The Chi-squared statistic would then be confronted to a curve of the probability density distribution of a Chi-distribution for the same degree of freedom, to discover its equivalent p-value. In the case of the test for association, a p-value lower than 0.05 means that the two variables entered are independent. A p-value higher than 0.05 would then mean that the two variables are associated.

The function that produces a chi-distribution for k degrees of freedom is given by (Equations 5 and 6, where Γ is the gamma distribution, also provided):

$$\begin{cases} for\ x\ \geq 0,\ f(x;k) = \frac{x^{k-1}e^{-\frac{x^2}{2}}}{2^{\frac{k}{2}-1}\Gamma\left(\frac{k}{2}\right)} \\ for\ x < 0,\ f(x;k) = 0 \end{cases} \quad \rightarrow \quad \Gamma\left(\frac{k}{2}\right) = \int_0^\infty x^{\frac{k}{2}}e^{-x}dx \tag{5}$$

To determine the limits of the model, the margin of error was calculated:

$$M = 1.96\frac{\sigma}{\sqrt{n}} \tag{6}$$

M = Margin of error; σ = standard deviation of the modeled values.

3. RESULTS AND DISCUSSION

3.1. Chemical Analyses

The observed concentrations for ammonia nitrogen, orthophosphate, and caffeine for the 20 sampling sites located along the Atuba River are presented on Table 2.

Table 2. Concentrations in the Atuba River for ammonia nitrogen, orthophosphate, and caffeine.

	Chemical Parameters				Chemical Parameters		
Site	**N_A (mg L^{-1})**	**PO_4^{3-} (mg L^{-1})**	**CAF (µg L^{-1})**	**Site**	**N_A (mg L^{-1})**	**PO_4^{3-} (mg L^{-1})**	**CAF (µg L^{-1})**
P1	0.111	0.381	0.639	**P11**	2.959	0.655	-
P2	0.219	0.422	-	**P12**	2.929	0.721	-
P3	0.236	0.452	0.642	**P13**	3.315	0.743	3.124
P4	1.764	0.542	-	**P14**	3.213	0.720	-
P5	2.070	0.589	-	**P15**	3.250	0.724	4.087
P6	2.257	0.590	-	**P16**	3.276	0.682	-
P7	2.162	0.607	1.007	**P17**	3.629	0.828	-
P8	2.569	0.654	-	**P18**	3.298	0.697	-
P9	2.492	0.655	-	**P19**	3.219	0.700	7.168
P10	2.327	0.660	1.186	**P20**	8.873	0.870	8.524

N_A = Ammonia nitrogen; PO_4^{3-} = Orthophosphate; CAF = Caffeine.

The concentrations for all parameters analyzed presented a rising behavior along the course of the river. The smallest concentration was observed on P1, which was closest to the river's spring, while the largest value was observed on the site P20. The site P20 was the one located the furthest from the spring, as well as being the only site which was influenced by the discharge from the WWTP Atuba Sul. The concentration rise found through the sites P19 and P20 may point to an inefficient treatment of the sewage processed on this WWTP regarding the ammonia nitrogen removal

Likewise, the spike in caffeine concentration starting at P13 may be explained by the demographic density found within these zones of hydrological influence. The largest concentration raise was 3.081 µg L^{-1}, between sites P15 and P19. This may be due to the higher density of irregular housing found in this region. As Katukiza *et al.* (2012) and Kelman (2015) proposed, these areas tend to be hotspots for hydric pollution for two reasons: i) lower governmental incentive to connect these residences to the public sewage and solid waste collection systems; ii) the irregular aspect of the domicile creates insecurity for their inhabitants, which usually are not willing to invest in individual wastewater treatment solutions. These factors point to the importance of the study of the social aspects of the residents of a river's

basin to better understand the source and fate of water pollution in urban rivers.

Table 3 presents the concentrations determined for the Atuba River by other researchers that also encompassed the Atuba River in their study areas.

Table 3. Concentrations for caffeine (CAF), ammonia nitrogen (N_A) and orthophosphate (PO_4^{3-}) for other sampling campaigns performed on the Atuba River.

Sampling Date	Reference	Caffeine concentration (µg L^{-1})			N_A (mg L^{-1})	PO_4^{3-} (mg L^{-1})
		Min	Max	Mean		
April/2014	**Mizukawa *et al.* (2019)**	0.65	5.375	2.89	9.77 (±15.6)	n.a.
June/2014		1.58	5.36	3.58	17.5 (±19.4)	n.a.
October/2014		0.53	4.39	3.06	10.6 (±6.7)	n.a.
March/2015		0.3	4.96	3.35	2.0 (±2.0)	n.a.
June/2015		1.46	3.44	2.79	18.4 (±9.9)	n.a.
February/2010	**Ide *et al.* (2017)**	8.97	6.9	7.93	n.a.	n.a.
May/2010		0.39	2.14	1.26	n.a.	n.a.
August/2010		5.14	6.73	5.93	n.a.	n.a.
November/2010		0.5	5.58	3.04	n.a.	n.a.
April/2012 Feb/2013*	**Osawa *et al.* (2015)**	n.a.	n.a.	n.a.	16.2 (±26.8)	0.48 (±0.88)
April/2011-Nov/2011*	**Kramer *et al.* (2015)**	n.a.	n.a.	n.a.	25.91 (±20.72)	1.84 (±1.71)

N_A = Ammonia Nitrogen; PO_4^{-3} = Orthophosphate; * = Four sampling campaigns were performed through this period.

Mizukawa *et al.* (2019) analyzed four sampling sites (named P1, P7, P9 and P20 in this study) and Ide *et al.* (2017) analyzed two (named P19 and P20). Though, the caffeine concentrations found by both groups were lower than the ones determined in this study. This result could be explained by the date of the sampling, which influences the amount of people residing on this basin, as well as the volume of wastewater processed by the WWTP Atuba Sul. According to the Brazilian Institute for Geography and Statistics (IBGE, 2019), the population for the area grew 6.4% from 2010 to 2014, 3.9% from 2014 to 2019 (a total of 10.4% from 2010 to 2019).

The concentrations for ammonia nitrogen determined by Mizukawa *et al.* (2019), Osawa *et al.* (2015) and Kramer *et al.* (2015) were higher than those observed in this study. This might be due to an improvement in ammonia nitrogen removal efficiency by the WWTP Atuba Sul. A reason that might explain a higher concentration of caffeine, yet a lower concentration of ammonia nitrogen, is the dilution of these contaminants by the rainy season, which could indicate that during a drier season the concentration of caffeine might be higher.

The concentration of orthophosphate observed in this study was on par with the ones observed by Osawa *et al.* (2015) and Kramer *et al.* (2015).

The concentrations of caffeine observed by other researchers in Brazil and other countries are presented in Table 4.

Table 4. Concentrations for caffeine observed by other researchers.

Country	Location	Caffeine concentration (µg L^{-1})	Reference
Brazil	Sinos River	3.73 (± 6.76)	Peteffi *et al.* (2019)
	Dourados River	0.14 (± 0.33)	Sposito *et al.* (2018)
	Brilhante River	0.02 (± 0.02)	
Taiwan	Taiwan Strait (Seawater)	0.002 (± 0.002)	Fang *et al.* (2019)
Ukraine	Dnieper River	19.2	Ho *et al*. (2020)
China	Shijing River	(Mass flow) 446.57 g d -1	Yuan *et al.* (2020)

The concentrations observed by the studies shown on Table 4 (Lopez-Doval *et al.*, 2016; Sposito *et al.*, 2018; Peteffi *et al.*, 2019; Fang *et al.*, 2019; Yuan *et al.*, 2020) corroborate the idea that caffeine contamination present on urban water bodies is not a problem confined only to the Atuba River Basin, or even to Brazil, but a worldwide problem. On a smaller scale, caffeine was also found on sea waters across the planet. Thus, the monitoring of caffeine presents itself as a viable way to understand the extent of the anthropic pressures put upon water resources.

Therefore, the existence of a tool (in this case, a model) able to determine indirectly the caffeine concentration on an urban river could improve the understanding of water contamination, across the world.

3.2. MODEL GENERATION

The model created aimed to determine the concentrations of caffeine in an indirect way. The model used as input the following social parameters: total population and population residing in irregular housing within the zones of hydrological influence of each sampling site. The chemical parameters used as inputs were the concentrations of ammonia nitrogen and orthophosphate, found on the same sites.

The first step used to generate the model for the caffeine concentration to determine which of the variables would be independent or dependent. It was decided that the chemical parameters were a portrait of the behavior of the population which resides in the basin, and not the opposite. Therefore, the total population and the population residing in irregular housing were set as the independent variables. Consequently, the chemical parameters were set as the dependent variables.

The second step was to determine the variance and covariance of all variables. The results for this procedure are presented on Table 5.

Table 5. Variance-Covariance matrix for the social and chemical parameters of the Atuba River.

S	IVs		DVs		
	P_T	P_i	CAF	N_A	PO_4
P_T	23188528890	780977474	339923	20833	15643
P_i	780977474	30550886	12846	7724	490.8
CAF	339923	12846	5.629	3.291	0.210
N_A	20833	7724	3.291	3.094	0.184
PO_4	15643	490.8	0.210	0.184	0.015

IVs = Independent Variables; DVs = Dependent Variables; P_T = Total Population; P_i = Population in irregular housing; CAF = Caffeine; N_A = Ammonia nitrogen; PO_4 = Orthophosphate.

Through the **S** matrix, it was possible to determine the vectors with the canonical values for the independent (A) and dependent (B) variables. There are two instances of canonical values for each set of variables, these values are represented by the columns of their respective vectors (Equations 7 and 8).

$$A = \begin{bmatrix} -0{,}00000318 & 0.00001732 \\ -0.00009672 & -0.00047545 \end{bmatrix} \quad (7)$$

$$B = \begin{bmatrix} -0{,}379812 & -0.20066088 \\ 0.05470996 & -0.79124554 \\ -1.88825624 & 16.42867095 \end{bmatrix} \quad (8)$$

The next step was to establish which of the sets of canonical values are the most appropriate; for this purpose, the $R_{U,V}^{-1}$ matrix was used. The $R_{U,V}^{-1}$ matrix determines the value of the linear correlation among the coefficients in both matrix A and B (Equation 9).

$$R_{U,V}^{-1} = \begin{bmatrix} 1.01365326 & 0 \\ 0 & 1.26481073 \end{bmatrix} \quad \begin{cases} r_1 = 0.9865 \\ r_2 = 0.7063 \end{cases} \quad (9)$$

The first column of the matrices A and B presented a correlation of r = 0.9865, which is significantly higher than the second column (r = 0.7063). Therefore, the first set of canonical values was the chosen one. These values, along with the observed variables, were then formed into an equation (Equation 10):

$$-0.00000318P_T - 0.00009672P_i = -0.379812CAF + 0.05470996N_A - 1.88825624PO_4 + 0.2033 \quad (10)$$

To make the equation more easily understandable, the values were transformed to into integers (Equations 11 and 12):

$$\frac{2P_T}{61} + P_i = 3981CAF - \frac{1720N_A}{3} + 19786PO4 - 8409 \quad (11)$$

Solving for CAF:

$$CAF = \frac{P_T}{112270} + \frac{P_i}{3981} - 0.144\,N_A + 4.970\,PO_4 - 2.112 \quad (12)$$

Where: CAF = caffeine concentration (μg L^{-1}); P_T = Total Population; P_i = Population in irregular housing; N_A = concentration of ammonia nitrogen (mg L^{-1}); PO_4 = concentration of orthophosphate (mg L^{-1}).

The modeled values for the caffeine concentrations for the sites P1, P3, P7, P10, P13, P15, P19 and P20 (the ones in which caffeine was analyzed) were determined. The correlation coefficient between the ones modeled was r = 0.984. It presented a total RMSE of 1.18 μg L^{-1}. The culprit for this error statistic is the consistent overestimate of the model (excepting P13), when compared to the observations. The Chi-squared statistic for these two datasets (degrees of freedom = 7) was $\chi^2 = 0.9661$, which provided a p-value of 0.955. This led to the exclusion of the null hypothesis (that these two datasets were independent).

The correlation graph for the observed and modeled values is presented on Figure 2.

The values for the concentrations of caffeine for the twenty sampling sites were then calculated. These are presented in Table 6.

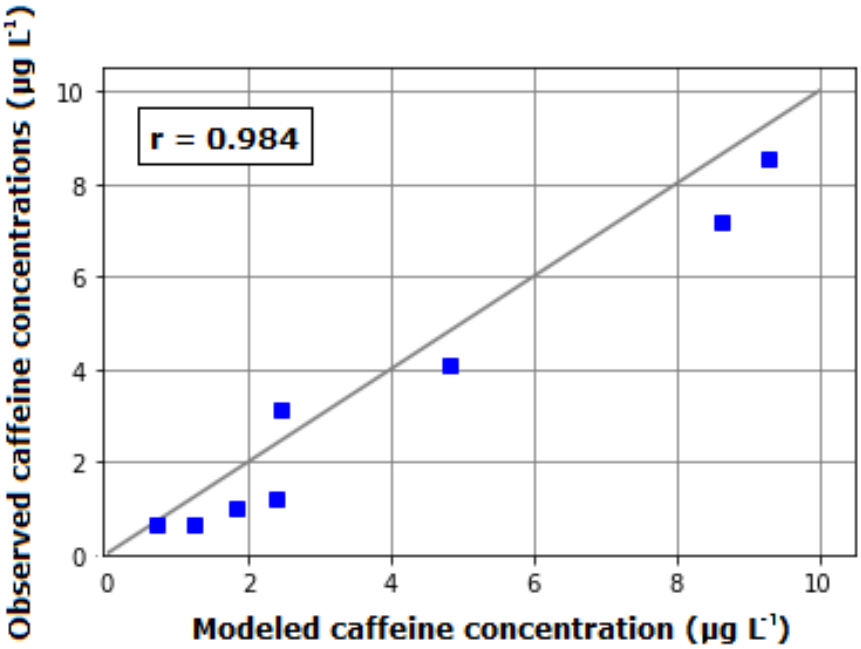

Figure 2. Correlation of the concentration of caffeine observed on the sampling sites P1, P3, P7, P10, P13, P15, P19 and P20 and the values predicted for the same sites using the CCA-generated model.

Table 6. Concentrations of caffeine (µg L^{-1}) found for the 20 sampling sites through the model generated by CCA compared with the values observed at 8 sampling sites across the Atuba River.

Site	Modeled caffeine concentration (mg L^{-1})	Observed Caffeine concentration (µg L^{-1})	Site	Modeled caffeine concentration (mg L^{-1})	Observed Caffeine concentration (µg L^{-1})
P1	0.728	0.639	**P11**	2.539	-
P2	1.069	-	**P12**	2.355	-
P3	1.244	0.642	**P13**	2.472	3.124
P4	1.407	-	**P14**	4.550	-
P5	1.754	-	**P15**	4.815	4.087
P6	1.834	-	**P16**	5.248	-
P7	1.832	1.007	**P17**	5.672	-
P8	2.000	-	**P18**	7.340	-
P9	2.225	-	**P19**	8.640	7.168
P10	2.401	1.186	**P20**	9.283	8.524

The distribution of the modeled values presented a standard deviation of 3.853, which led to a margin of error of 1.6887 µg L^{-1} (probability = 95%, z = 1.96). Figure 3, below, presents the comparison of the model (and its probable spectrum) to the studies that analyzed caffeine concentrations performed on the Atuba River.

The modeled spectrum covers all the values observed for this sampling campaign. This was expected, as these were the concentrations used to generate the model. The values for the other campaigns performed on the Atuba River (by Ide *et al.* (2017) and Mizukawa *et al.* (2019) are not, for the most part, encompassed by the model. This could be explained by the differences that occurred in the study area during this nine-year span. While the concentrations found on sites P19 and P20 were consistently lower than those found on this study, the ones found on

site P7 were consistently higher. This could be explained by the different behavior of the population, as the latter two sampling sites are in a lower income, irregular housing area.

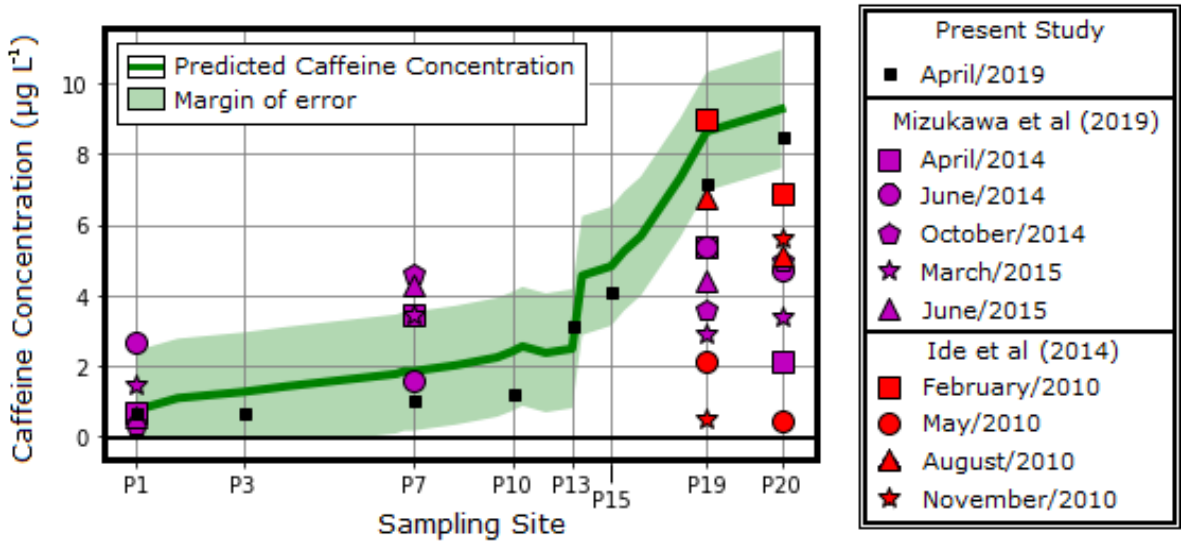

Figure 3. Distribution of three studies performed on the Atuba River compared with the modeled caffeine concentrations.

Another topic that has risen concern recently is the presence of viruses in water resources. This model might be a tool to better understand the behavior of these microorganisms in the aquatic environment. Gourmelon *et al.* (2010) and Kumar *et al.* (2019) have observed concomitantly the concentrations of caffeine and viral genomes in wastewater, rivers and lakes. Their results might indicate somewhat of a correlation between the concentrations of these two parameters. As their results pointed out, higher concentrations of caffeine were usually followed by a higher concentration of viral genome and a higher presence of the bacteria *Escherichia coli*. Their studies analyzed the concentrations of genomes from Norovirus, Hepatitis A, Aichivirus, Pepper Mild Mottle Virus and F-specific RNA bacteriophages. The results observed by Sidhu *et al.* (2013) have also pointed to some correlation between higher concentrations of caffeine and the presence of viral genome in water systems.

Due to the pandemic the world was under in 2021, the analysis of viral presence (especially SARS-CoV-2) in water resources have raised even more concerns. Rimoldi *et al.* (2020) have observed some possible correlation between the presence of this specific virus on waters with high concentrations of caffeine (above 0.44 µg L^{-1}). Interesting future studies could be done to better understand this correlation.

The model was able to interpolate a possible range of concentrations of caffeine that could be observed along the river. It took as input only socio economic data and the concentrations of nutrients, whose determination is cheaper and less labor intensive. The information provided by this model, when applied to other basins, might be useful to managers, researchers, and decision-makers to better understand the pressures a given urban river might be under.

4. CONCLUSIONS

The concentrations for ammonia nitrogen and orthophosphate for 20 sampling sites along the Atuba River were quantified. Also, the population of influence that inhabited each of the zones of hydrological influence was defined (from which the total population and population residing in irregular housing for each sampling site were established). For 8 of these sampling sites, the concentrations of caffeine were determined. Using these observed datasets, a model

was created through the manipulation of the CCA method. The study generated a spectrum of probable ranges of caffeine concentrations for the course of the river, based on 4 inputs: the total population; the population inhabiting irregular housing; then population residing in the zone of hydrological influence; and the concentrations of ammonia nitrogen and orthophosphate at the sampling site. Even though the water resource system of the Atuba River is complex, being home for over half a million people and receiving the discharge from the WWTP Atuba Sul, the model had a satisfactory performance. More studies should be performed to understand the effects of generalization for water resource systems that are under different pressures, as well as different climatic, cultural, and infrastructural situations.

Due to the characteristics of caffeine as a chemical tracer for anthropic pressures on water systems, a model that could predict indirectly its concentration along a river might provide more information at a lower cost. The method for generating the model's equation could be applied to different urban rivers, becoming a useful tool for water-resource managers, researchers, and decision-makers to better understand the environmental impact the population inhabiting a given basin will have on its water resources. This model could be particularly beneficial for regions that do not have systematic sampling campaigns, by being able to determine zones which would demand higher attention from managers and decision-makers, with a small number of sporadic non-uniformly spread data samples.

5. ACKNOWLEDGMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) and the Brazilian National Council for Scientific and Technological Development (CNPq) by Bolsa Productividade (proc. 302736/2016-6) and Call MCTIC/CNPq N° 28/2018 (proc. 407157/2018-2). The authors thank the Graduate Program for Water Resources and Environmental Engineering of the Universidade Federal do Paraná (PPGERHA-UFPR) and the Multidisciplinary Nucleus for Research on Environmental Technologies of the Universidade Tecnológica Federal do Paraná (NIPTA-UTFPR).

6. REFERENCES

APHA; AWWA; WEF. **Standard Methods for the examination of water and wastewater**. 22nd ed. Washington, 2012. 1496 p.

COLIM, A. N.; NASCIMENTO, P. C.; WIETHAN, B. A.; ADOLFO, F. R.; DRESCH, L. R.; CARVALHO, L. M. *et al.* Reversed-Phase High-Performance Liquid Chromatography for the Determination of 15 Rare Earth Elements in Surface Water Sample Collected in a Mining Area from Lavras do Sul/RS, Brazil. **Cromatographia**, v. 82, n. 5, 2019. https://link.springer.com/article/10.1007/s10337-019-03709-w

DAFOUZ, R.; CÁCERES, N.; RODRÍGUEZ-GIL, J. L.; MASTROIANNI, N.; DE ALDA, M. L.; BARCELÓ, D. *et al.* Does the presence of caffeine in the marine environment represent an environmental risk? A regional and global study. **Science of the Total Environment**, v. 615, 632–642, 2018. https://doi.org/10.1016/j.scitotenv.2017.09.155

DI FELICI, V.; MARCINELLI, R.; PROULX, R.; CAMPIGLIA, E. A multivariate analysis for evaluating the environmental and economical aspects of agroecosystem sustainability in central Italy. **Journal of Environmental Management**. v. 98, 119-126, 2012. https://doi.org/10.1016/j.jenvman.2011.12.015

FANG, T.-H.; LIN, C.-W.; KAO, C.-H. Occurrence and distribution of pharmaceutical compounds in the Danshuei River Estuary and the Northern Taiwan Strait. **Marine Pollution Bulletin**, v. 146, p. 509-520, 2019. https://doi.org/10.1016/j.marpolbul.2019.06.069

FERREIRA, D. F. **Estatística Multivariada**. 3. ed. Lavras: Editora UFLA, 2018. p. 624.

GERSHUNOV, A.; BENMARHINA, T.; AGUILERA, R. Human health implications of extreme precipitation events and water quality in California, USA: a canonical correlation analysis. **Lancet Planetary Health**, v. 2, n. 1, p. S9, 2018. https://doi.org/10.1016/S2542-5196(18)30094-9

GOURMELON, M.; CAPRAIS, M. P.; MIESZKIN, S.; MARTI, S.; WÉRY, N.; JARDÉ, E. *et al.* Development of microbial and chemical MST tools to identify the origin of the faecal pollution in bathing and shellfish harvesting waters in France. **Water Research**, v. 44, 4812-4824, 2010. https://doi.org/10.1016/j.watres.2010.07.061

HAN, L.; ZHOU, W.; LI, Y.; QIAN, Y. Urbanization strategy and environmental changes: An insight with relationship between population change and fine particle pollution. **Science of the Total Environment**, v. 642, 789-799, 2018. https://doi.org/10.1016/j.scitotenv.2018.06.094

HO, K. T. *et al*. Contaminants, mutagenicity and toxicity in the surface waters of Kyiv, Ukraine. **Marine pollution bulletin**, v. 155, p. 111153, 2020. https://doi.org/10.1016/j.marpolbul.2020.111153

HOTELLING, H. Relations between two sets of variables. **Biometrika**, v. 28, n. 3/4, 321-327, 1936. https://doi.org/10.1007/978-1-4612-4380-9_14

IDE, A. H.; OSAWA, R. A.; MARCANTE, L. O.; PEREIRA, J. L. G. F. S. C.; AZEVEDO, J. C. R. Occurrence of pharmaceutical products, female sex hormones and caffeine in a subtropical region in Brazil. **CLEAN-Soil Air Water**, v. 45, n. 9, 2017.

IDE, A. H.; OSAWA, R.; MARCANTE, L. O.; PEREIRA, J. L. G. F. S. C.; AZEVEDO, J. C. R. Utilização da cafeína como indicador de contaminação por esgotos domésticos na bacia do Alto Iguaçu. **Brazilian Journal of Water Resources**, v. 18, n. 2, p. 201-211, 2013. https://doi.org/10.1002/clen.201700334

IBGE. **População nos censos demográficos, segundo os municípios das capitais 1872/2010**. 2019. Available on: https://censo2010.ibge.gov.br/sinopse/index.php?dados=6.

KATUKIZA, A. Y.; RONTELTAP, M.; NIWAGABA, C. B.; FOPPEN, J. W. A.; KANSIIME, F.; LENS, P. N. L. Sustainable sanitation technology options for urban slums. **Biotechnology Advances**, v. 30, n. 5, 964–978, 2012. https://doi.org/10.1016/j.biotechadv.2012.02.007

KELMAN, J. Water Supply to the Two Largest Brazilian Metropolitan Regions. **Aquatic Procedia,** v. 5, 13–21, 2015. **https://doi.org/10.1016/j.aqpro.2015.10.004**

KHALIL, B.; OUARDA, T. B. M. J.; SAINT-HILAIRE, A. Estimation of water quality characteristics at ungauged sites using artificial neural networks and canonical correlation analysis. **Journal of Hydrology**, v. 405, 277-287, 2011. https://doi.org/10.1016/j.jhydrol.2011.05.024

KRAMER, R. D.; MIZUKAWA, A.; IDE, A. H.; MARCANTE, L. O.; SANTOS, M. M.; AZEVEDO, J. C. R. Determinação de anti-inflamatórios na água e sedimento e suas relações com a qualidade da água na bacia do Alto Iguaçu, Curitiba-PR. **Brazilian Journal of Water Resources,** v. 20, n. 3, 657–667, 2015.

KUMAR, M.; RAM, B.; HONDA, R.; POOPIPATTANA, C.; CANH, P. D.; CHAMINDA, T. *et al.* Concurrence of antibiotic resistant bacteria (ARB), viruses, pharmaceuticals and personal care products (PPCPs) in ambient waters of Guwahati, India: Urban vulnerability and resilience perspective. **Science of the Total Environment**, v. 693, 133-147, 2019. https://doi.org/10.1016/j.scitotenv.2019.133640

LOPEZ-DOVAL, J. C.; MONTAGNER, C. C.; ALBUQUERQUE, A. F.; MOSCHINI-CARLOS, V.; UMBUZEIRO, G.; POMPEO, M. Nutrients, emerging pollutants and pesticides in a tropical urban reservoir: Spatial distributions and risk assessment. **Science of the Total Environmen***t*, v. 575, 1307-1324, 2016. https://doi.org/10.1016/j.scitotenv.2016.09.210

MALACARNE, R. L. Canonical Correlation Analysis. **The Mathematica Journal**, v. 16, 6-16, 2014.

MANGADZE, T.; TAYLOR, J. C.; FRONEMAN, W. P.; DALU, T. Water quality assessment in a small austral temperate river system (Bloukrans River system, South Africa): Application of multivariate analysis and diatom indices. **South African Journal of Botany**, v. 125, p. 353-359, 2019. https://doi.org/10.1016/j.sajb.2019.08.008

MIZUKAWA, A.; FILIPPE, T. C.; PEIXOTO, L. O. M.; SCIPIONI, B.; LEONARDI, I. R.; AZEVEDO, J. C. R. Caffeine as a chemical tracer for contamination of urban rivers. **Brazilian Journal of Water Resources**, v. 24, 2019. https://doi.org/10.1590/2318-0331.241920180184

OSAWA, R. A.; IDE, A. H.; SAMPAIO, N. M. F. M.; AZEVEDO, J. C. R. Determinação de fármacos anti-hipertensivos em águas superficiais na região metropolitana de Curitiba. **Brazilian Journal of Water Resources**, v. 20, n. 4, p. 1039-1050, 2015.

PETEFFI, G. P.; FLECK, J. D.; KAEL, I. M.; ROSA, D. C.; ANTUNES, M. V.; LINDEN, R. Ecotoxicological risk assessment due to the presence of bisphenol A and caffeine in surface waters in the Sinos River Basin - Rio Grande do Sul – Brazil. **Brazilian Journal of Biology**, v. 79, p. 712-721, 2019. https://doi.org/10.1590/1519-6984.189752

RENCHER, A. C. **Methods of Multivariate Analysis**. 2. ed. Danvers: John Wiley & Sons, 2002. p. 727.

RIMOLDI, S. G.; STEFANI, F.; GIGANTIELLO, A.; POLESELLO, S.; COMANDATORE, F.; MILETO, D. *et al.* Presence and infectivity of SARS-CoV-2 virus in wastewaters and rivers. **Science of the Total Environment**, v. 744, p. 140-148, 2020. https://doi.org/10.1016/j.scitotenv.2020.140911

SIDHU, J. P. S.; AHMED, W.; GERNJAK, W.; ARYAL, R.; MCCARTHY, D.; PALMER, A. *et al.* Sewage pollution in urban stormwater runoff as evident from the widespread presence of multiple microbial and chemical source tracking markers. **Science of the Total Environment**, v. 463-464, p. 488-496, 2013. https://doi.org/10.1016/j.scitotenv.2013.06.020

SPOSITO, J. C. V.; MONTAGNER, C. C.; CASADO, M.; NAVARRO-MARTIN, L.; SOLORZANO, J. C. J.; PINA, B. *et al.* Emerging contaminants in Brazilian rivers: Occurrence and effects on gene expression in zebrafish (*Danio rerio*) embryos. **Chemosphere**, v. 209, p. 696-704, 2018. https://doi.org/10.1016/j.chemosphere.2018.06.046

TIYASHA, T. M.; TUNG, Z. M.; YASEEN. A survey on river water quality modeling using artificial intelligence models: 2000–2020. **Journal of Hydrology**, v. 585, p. 124-186, 2020. https://doi.org/10.1016/j.jhydrol.2020.124670

WEI, H.; YU, H.; ZHANG, G.; PAN, H.; LV, C.; MENG, F. Revealing the correlations between heavy metals and water quality, with insight into the potential factors and variations through canonical correlation analysis in an upstream tributary. **Ecological Indicators,** v. 90, p. 485-493, 2018. https://doi.org/10.1016/j.ecolind.2018.03.037

WOODHOUSE, P.; MULLER, M. Water governance – an historical perspective on current debates. **World development**, v. 92, p. 225-241, 2017. https://doi.org/10.1016/j.worlddev.2016.11.014

YUAN, X.; HU, J.; LI, S.; YU, M. Occurrence, fate, and mass balance of selected pharmaceutical and personal care products (PPCPs) in an urbanized river. **Environmental Pollution**, v. 266, p. 115-124, 2020. https://doi.org/10.1016/j.envpol.2020.115340