Research Article

Máté Krisztián Kardos* and Adrienne Clement

Predicting small water courses' physico-chemical status from watershed characteristics with two multivariate statistical methods

https://doi.org/10.1515/geo-2020-0006
Received Oct 07, 2019; accepted Dec 23, 2019

Abstract: Watershed area and a bunch of relief, land use, and wastewater characteristics for 32 upland and 33 lowland small river courses are generated. Based on these characteristics, logistic binary regression models are trained to predict if the river achieves the good physico-chemical status, and discriminant analysis models are trained to predict the physico-chemical status class on a five-class scale.

Univariate models revealed that elevation (for upland rivers), the share of artificial surfaces (for lowland rivers) along with forests, and wastewater quality variables such as biochemical oxygen demand, chemical oxygen demand, and phosphorus are the most significant predictors. Discriminant analysis models performed better on upland than on lowland rivers. Achievement of good status could be predicted with an accuracy of ~90% (with 2 to 4 variable logit models), whereas the status class with an accuracy of 63/48% (with 2 to 4 variable discriminant analysis models) for upland and lowland rivers, respectively. This contribution uses Hungary as a case study.

Keywords: binary logistic regression, diffuse pollution, land use, linear discriminant analysis, point source pollution, water quality monitoring, Water Framework Directive, Central Europe

1 Introduction

Starting in the early 20th century, the pollution of waters, and, in particular, rivers has received growing attention worldwide. Protecting them to fulfill human needs turned out not to be sustainable in the long term. Starting in the late 20th century, water protection measures began to focus on the ecosystems of the water. Two examples are the Clean Water Act in the US [1] or the Water Framework Directive (WFD) in Europe [2]. The goal of the latter is to achieve good ecological status / potential of all surface and subsurface waters by (at the latest) 2027 from a biological point of view. Groundwater, rivers, lakes, transitional and coastal waters as well as estuaries are all in the scope of the WFD.

The primary tool of the WFD is the river basin management plans to be prepared by each member state in a 6-year cycle. Basic unit of the river basin management plans are the water bodies comprising one or more stretches/parts of the waters mentioned above. Categorizing water bodies into a few types facilitates their management. For surface freshwaters, the typology is based on altitude, slope (in case of rivers), geology, sediment, and catchment size of the water body (Table 1) [2–4].

After delimiting the water bodies and their watershed, river basin management plans require to list all pressures (*i.e.*, natural and human effects influencing the water quality) and to assess the status of each water body. The status evaluation results in each water body assigned to one of the classes *high*, *good*, *moderate*, *poor* or *bad*. Assessing the reliability of the classification (*high*, *medium* or *low*) is also part of the status evaluation. The primary base of the classification is water quality monitoring data. Monitoring here means hydro-morphological, biological and physico-chemical monitoring of each water body. For more details on the physico-chemical status assessment of Hungarian river water bodies, the reader is referred to [5, 6].

For most countries, implementation of the monitoring required by the WFD is a considerable challenge. Traditional monitoring of all water bodies with the necessary reliability would require unreasonably high efforts [7, 8]. This statement is particularly true for Hungary. With a few exemptions, smaller rivers and lakes have not been monitored before the year 2007. The status of only a tiny part (170 out of 1078) of all surface water bodies could be assessed with high reliability in the 2nd river basin management plan, while 145 out of 1078 surface water bodies

***Corresponding Author: Máté Krisztián Kardos:** Budapest University of Technology and Economics Budapest, Hungary; Email: kardos.mate@epito.bme.hu
Adrienne Clement: Budapest University of Technology and Economics Budapest, Hungary

Table 1: Hungarian river water body types, and number of the particular water bodies. nWB = number of water bodies in the particular type category. High = number of water bodies classified with high reliability.

type #	altitude	slope	geology	sediment	catchment size	nWB	high
1	upland	high	siliceous	coarse	small	20	9
2	upland	high	calcareous	coarse	small to medium	31	7
3	upland	medium	calcareous	any	small to medium	359	32
4	upland	medium	calcareous	coarse	large to very large	19	12
5	lowland	low	calcareous	coarse	small to medium	23	8
6	lowland	low	calcareous	medium to fine	small to medium	376	33
7	lowland	low	calcareous	medium to fine	large	33	20
8	lowland	low	calcareous	medium to fine	very large	18	15
9	lowland	m – low	calcareous	coarse	Danube-size	9	9
10	lowland	low	calcareous	medium to fine	Danube-size	1	1

stayed "grey" (meaning unknown status) [4, 9]. Emerging methods like citizen science, remote sensing, and big data, along with machine learning algorithms, are to be considered as a solution to the monitoring dilemma. However, these methods are not widely known and elaborated yet [10, 11]. On the other hand, monitoring of large rivers is in some cases excessive [12–14].

A plethora of studies reveal that there is an apparent link between a watershed's characteristics and it's water quality [15–17]. In the hypothetical case of having all the knowledge on the background factors and the processes, no monitoring would be needed. In reality, the relationships are complex, and it is due to this fact that deterministic models (based on the physical processes and referred to as water quality or watershed models) usually do not perform better than statistical ones [18–21].

The statistical modeling task is usually data-driven: the watershed properties taken into account as predictor variables are those the data is available for. The most important predictor variables are geology, land use, and point sources pollutions [22–24]. The established relationships, however, are strongly site-specific, since agricultural and industrial as well as wastewater treatment technologies have a substantial variation around the globe. The statistical modeling method has widely been applied in the Americas [25–27] or Asia [28–30] but less frequently in Europe. A few European examples are [31, 32].

Logistic binary regression (in short: logit) is widely used in finances (*e.g.*, credit assessment) or medicine (*e.g.*, in predicting disease from the way of life or deoxyribonucleic acid) [33, 34]. It is somewhat less known in predicting the probability of a water pollution event [35, 36]. Linear discriminant analysis is widely used for dimensionality reduction of large datasets, among other things of water quality monitoring [37–39]. It can also be used to predict class assignation based on continuous or categorical covariates. Since water quality is more mapped on a continuous scale rather than as nominal categories, the use of linear discriminant analysis for predicting water quality is rare. The authors are not aware of any of these methods used to establish a direct link between the watershed's characteristics, and it's physico-chemical status class on a regional scale.

1.1 Objectives of the study

This study aims at defining relationships between the physico-chemical status of small watercourses and the physical characteristics of their watershed that can be calculated from databases covering large areas. Also, we aim at defining the accuracy and reliability of such relationships. In particular, we will

- determine linear discriminant analysis and logit models with watershed properties as the predictor and the water body's physico-chemical status class as the predicted variable;
- describe the reliability / accuracy of these models.

2 Material & Methods

2.1 Study site

Hungary is a landlocked country situated in the Carpathian basin (Figure 1). About two-thirds of the country's area is flat, and the rest is hilly. The highest point is 1014 and the lowest 76 meters above sea level. Mountainous regions are typically calcareous whereas lowland regions are loamy or sandy.

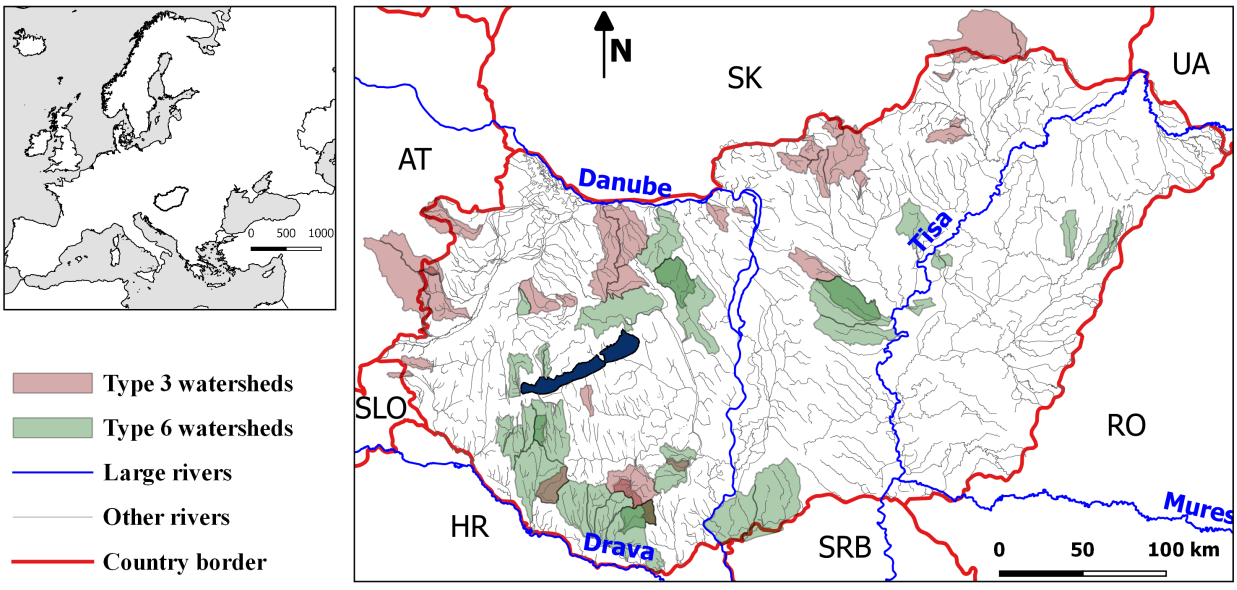


Figure 1: Cumulated watershed of study water bodies. Brown: type 3, green: type 6 watersheds. Darker colors show nested catchments.

The climate is continental; yearly mean precipitation is between 500 mm on central lowland regions and 850 mm on the southwestern and hilly areas. Mean monthly temperatures range from –2°C in January to 20°C in July. Larger rivers (except for the Danube) originate in direct neighboring countries (it is a typical "downstream country").

The average population density is 105 capita km^{-2}. Seventy percent of the inhabitants live in towns covering 3.6% of the country's surface; more than one-fourth of the population lives in or around the capital (Budapest). The most important economic activities include agriculture, industry, and tourism.

2.2 Material

The study presented in this paper is based on the physico-chemical classification of the 2nd river basin management plan of Hungary [40]. The classification was based on water quality measurements from years 2009 – 2012 for following water quality variables: pH, electric conductivity, chloride ion concentration, dissolved oxygen, oxygen saturation, biochemical oxygen demand, chemical oxygen demand, total organic carbon, ammonium ion concentration, total inorganic nitrogen, total nitrogen, orthophosphate ion concentration, total phosphorus [17, 41]. Only river water bodies classified with high reliability were included in the present study.

The cumulated watershed belonging to the outflow point of each water body was generated by summing up the immediate catchment of all upstream water bodies [42]. Basins extending to neighboring countries were created based on the topography (EU-DEM v1.1 [43]) with the TauDEM algorithm [44]. Only type 3 and type 6 water bodies classified with high reliability were included in the study (Table 1). In both types, all five classes were represented; in both of them, status *good* was the most frequent (Table 2).

Table 2: Frequency of physico-chemical classes in the studied types of water bodies.

	type 3	type 6
high	3	5
good	13	11
moderate	9	6
poor	6	7
bad	1	4
total	32	33

The EU-DEM also served for calculating mean elevation and slope for each watershed. Based on the Corine Land Cover database [45], land use share for four categories was generated for each basin (Table 3). Point source pollution values were based on two databases. The European Environment Agency's Urban Wastewater Directive

Table 3: Characteristics of study watersheds. Mean (minimum – maximum) values; LMQ = long term mean flow of the water body; masl = meters above sea level; PE = population equivalent. WB = Water body, WWTP = wastewater treatment plant.

	Type 3	Type 6
Relief & hydrology		
Watershed area [km^2]	240 (25 - 1000)	360 (4 - 1000)
Percent inland [%]	86 (6.7 – 100)	100 (89 – 100)
Mean elevation [masl]	280 (150 - 540)	150 (85 - 270)
Mean slope [%]	8.8 (1.7 - 16)	3.1 (0.32 - 8.8)
Long term specific runoff [mm]	97 (31 - 170)	67 (18 - 140)
Long term mean flow of the WB (LMQ) [$m^3 s^{-1}$]	0.75 (0.053 – 4.0)	0.75 (0.008 - 3.5)
Land cover and land use		
Artificial surfaces [%]	7.0 (2.4 - 33)	6.6 (1.6 - 22)
Agricultural areas [%]	47 (5.6 - 88)	62 (26 - 90)
Forest and semi natural areas [%]	46 (4.3 - 86)	30 (1.1 - 71)
Wetlands and water bodies [%]	0.27 (0 - 2)	1.4 (0 - 12)
EU urban wastewater database [46]		
Number of WWTP-s [-]	2.7 (0 - 12)	2.6 (0 - 11)
Load relative to LMQ [PE $m^{-3} s^1$] (eq. (1))	620 (0 - 5900)	2900 (0 - 54000)
HU-RWBM WWTP database [48]		
Number of WWTP-s [-]	5.1 (0 - 23)	6.1 (0 - 28)
BOD relative to WB LMQ [mg l^{-1}]	0.61 (0 - 5)	3.3 (0 - 53)
COD relative to WB LMQ [mg l^{-1}]	3.3 (0 - 56)	14 (0 - 230)
TN relative to WB LMQ [mg l^{-1}]	0.85 (0 - 5)	4.7 (0 - 63)
TP relative to WB LMQ [mg l^{-1}]	0.074 (0 - 0.63)	0.74 (0 - 11)

Treatment Plants database [46] contains a list of all European Union wastewater plants along with their effluent load values (in population equivalent) and the treatment technology applied, classified into one of seven categories (no treatment / primary / secondary / secondary + nitrogen removal / secondary + phosphorus removal / secondary + nitrogen and phosphorus removal / secondary + other). From this database, the load entering each water body was calculated with the following formula.

$$L = L_0 + 0.65L_1 + 0.15L_2 + 0.02L_3 \quad (1)$$

where L_0 means load from plants with no treatment, L_1 means load from plants with primary treatment, L_2 means load from plants with secondary treatment, and L_3 means plants with secondary + optionally any other treatment. The numbers 0.65, 0.15 and 0.02 are intended to represent mean removal efficiencies [47].

The second source of point sources was a cadaster of the Hungarian wastewater treatment plants enclosed to the 2nd river basin management plan [48]. This database comprises the self-control reports of wastewater treatment plants from the years 2010-2012. It contains yearly mean discharge (in m^3/s) and yearly mean effluent biochemical oxygen demand, chemical oxygen demand, total nitrogen, and total phosphorus concentrations for each plant. Annual load values were calculated, summed up for each watershed, and divided by the long-term mean flow of the respective water body. The values can be interpreted as yearly mean concentrations originating from point sources, with the hypothesis of no in-stream retention and degradation. Tables 3-4 list the watershed properties along with their statistical values.

2.3 Methods

Logistic binary regression is a special case among the generalized linear models. Instead of predicting a continuous variable (as does a linear regression model), the probability of falling into one of two classes is predicted. In our case, the probability of achieving good status is predicted as the function of one or more watershed properties.

While being in their use very similar to regression models, discriminant analysis models have very different underlying mathematics. Instead of defining a predictor function, they aim at describing the discriminant functions that are the best in differentiating among the categories

Table 4: Matrix of Pearson's linear correlation coefficients of the predictor variables. Upper right part: Type 3, lower left part: Type 6 dataset. Elev = elevation; artif = artificial surfaces; agric = agricultural surfaces. BOD = Biochemical oxygen demand; COD = chemical oxygen demand; TN = total nitrogen; TP = total phosphorus; PCC = physico-chemical status class. Significance codes: 0 < *** < 0.001 < ** < 0.01 < * < 0.05 < ^{x} < 0.1 < ' ' < 1.

	area	elev	slope	artif.	agric	forest	lake	wwQ	load	BOD	COD	TN	TP	PCC
area		0.15	−0.02	−0.15	0.21	−0.17	0.05	−0.07	−0.06	−0.12	−0.13	0	−0.05	0.03
elev	0.16		0.77***	0.01	−0.63***	0.63***	−0.42*	−0.19	−0.27	−0.18	−0.16	−0.15	−0.15	−0.50**
slope	0.00	0.83***		0.07	−0.72***	0.70***	−0.28	−0.01	−0.09	0.06	0.07	−0.03	0.05	−0.25
artif.	−0.19	0.12	0.08		−0.14	−0.16	−0.18	0.21	0.15	0.12	0.06	0.11	0.14	0.31^{x}
agric.	0.09	−0.67***	−0.52**	0.00		−0.96***	0.35*	−0.04	0.14	−0.02	−0.08	0.03	−0.03	0.33^{x}
forest	−0.05	0.60***	0.45**	−0.27	−0.95***		−0.32^{x}	−0.01	−0.18	−0.01	0.06	−0.06	−0.01	−0.42*
lake	0.08	0.26	0.31^{x}	−0.12	−0.27	0.17		−0.18	−0.11	−0.14	−0.14	−0.19	−0.18	−0.04
wwQ	−0.17	0.00	−0.13	0.76***	−0.07	−0.14	−0.08		0.86***	0.81***	0.96***	0.79***	0.88***	0.63***
load	−0.11	−0.09	−0.14	0.64***	−0.11	−0.06	−0.14	0.92***		0.81***	0.88***	0.77***	0.90***	0.59***
BOD	−0.15	−0.10	−0.19	0.68***	−0.04	−0.14	−0.13	0.98***	0.96***		0.82***	0.91***	0.88***	0.67***
COD	−0.14	−0.08	−0.20	0.68***	−0.03	−0.15	−0.14	0.98***	0.95***	1.00***		0.72***	0.85***	0.54**
TN	−0.17	0.00	−0.15	0.71***	−0.06	−0.14	−0.07	0.99***	0.89***	0.97***	0.98***		0.91***	0.64***
TP	−0.19	−0.08	−0.21	0.69***	0.03	−0.20	−0.17	0.96***	0.86***	0.96***	0.98***	0.98***		0.64***
PCC	0.30^{x}	0.05	0.05	0.41*	0.30^{x}	−0.39*	−0.23	0.39*	0.36*	0.37*	0.39*	0.38*	0.37*	

Table 5: List of studied models. All indicated combinations of covariates were applied with both of the logit and the linear discriminant analysis methods. WW = wastewater; COD = chemical oxygen demand; TP = total phosphorus.

Model #	Elevation	Artificial	Forest	WW COD	WW TP	Dataset
1.3					x	type 3
1.6					x	type 6
2.3			x		x	type 3
2.6			x		x	type 6
3.3	x		x		x	type 3
3.6		x	x		x	type 6
4.3	x		x	x	x	type 3
4.6		x	x	x	x	type 6

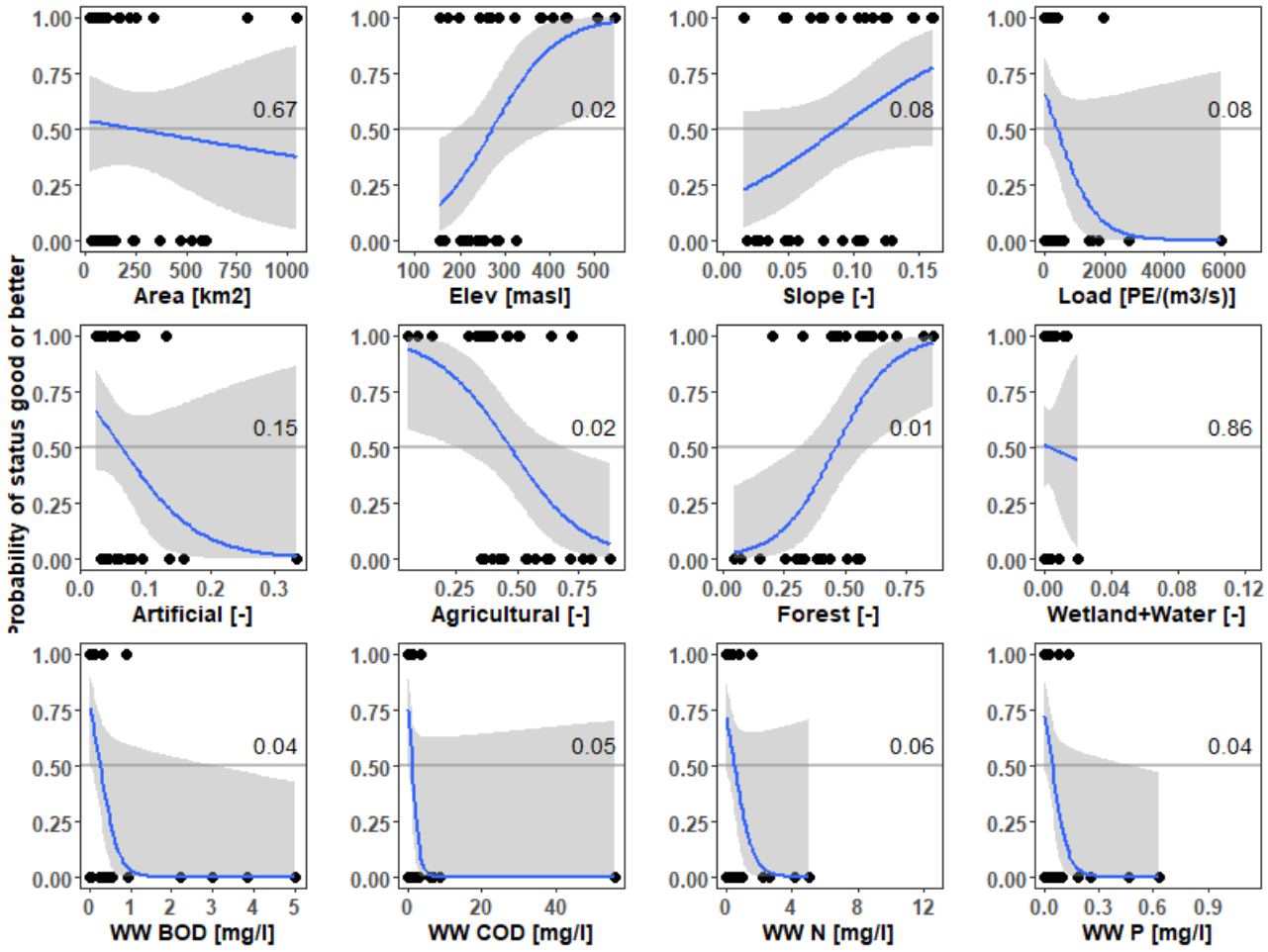


Figure 2: Relative probability of status good or high as function of unique watershed properties and 95% confidence intervals. Type 3 water bodies. Black dots show observations. X-axis ranges are calculated as the union of type 3 and type 6 watersheds, but two type 6 watersheds with extremely high point source pollution were excluded. Significance indicated (for the whole dataset).

of the function variable. As a result, they still do predict a categorical variable (of two or more categories).

As a first step of the present study, univariate logit models are established and visualized. Based on these models, and predictor variable's correlation table (Table 4), variables to be included in multivariable models are determined. Two times four multivariable models are investigated on both datasets and with both multivariable methods (Table 5).

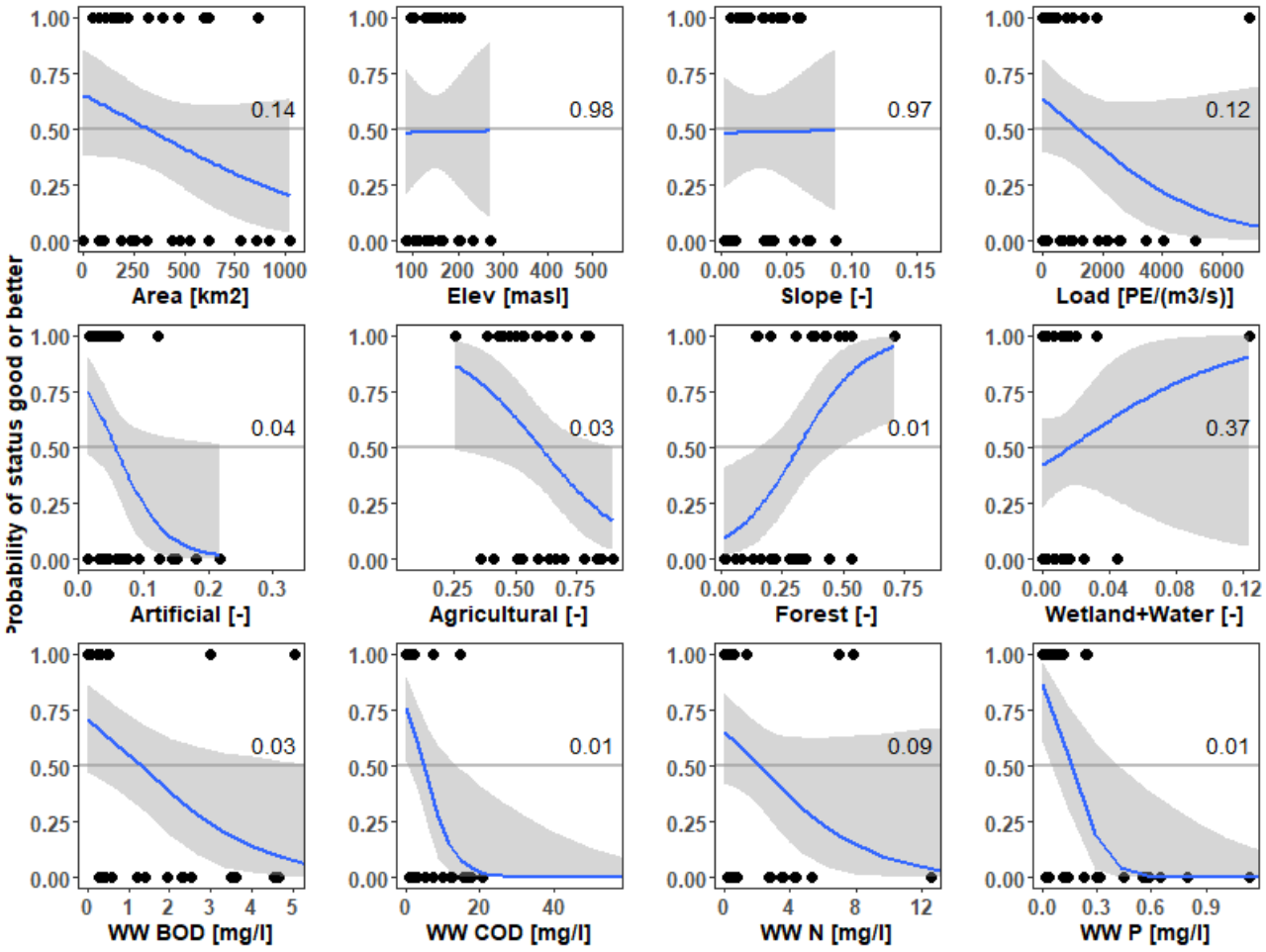


Figure 3: Relative probability of status good or high as function of unique watershed properties, and 95% confidence intervals. Type 6 water bodies. Black dots show observations. X-axis ranges are calculated as the union of type 3 and type 6 watersheds, but two type 6 watersheds with extremely high point source pollution were excluded. Significance indicated (for the whole dataset).

The present study had the aim to predict physico-chemical water quality based on the simplest possible watershed properties. Due to the relatively low number of training cases (32/33 watersheds for type 3 / type 6, respectively), the number of predictors could not exceed ~5 (5-6 training cases per predictors, see *e.g.* [34]). Relatively simple parameters were chosen as predictor variables: relief properties (area, slope, elevation), main land use categories (Table 3), and a few wastewater indicators. The chosen variables are supposed to have the most substantial influence on physico-chemical water quality. Other possible variables would include catchment shape indicators, drainage network (river network) density in the catchment, the slope of the channel network, fragmentation indicators of individual land-use types, land slope within individual land-use classes. These are subjects of future studies.

Variables were added sequentially. Models 1.3 and 1.6 consisted of one and the most significant predictor variable: total phosphorus emitted from point sources. Predictor variables of Models 2.3 and 2.6 were the share of forests (indicating absence of diffuse pollution) on the watershed and the phosphorus from point sources. The results of this model can be visualized on the 2D plane and thus help the reader to understand it's functioning.

The third predictor variable was elevation (for type 3) and artificial surfaces (type 6). These are still quite significant and at the same time, possibly uncorrelated predictors (Figures 2-3 and Table 4). As a fourth predictor, chemical oxygen demand emitted from point sources was added, representing another aspect of point sources pollution, but is highly correlated with total phosphorus.

The same combination of covariates was applied when running the logit and the linear discriminant analysis models. In the case of the logit models, the significance level

of the predictor variables, the accuracy of the model, and the "area under the curve" are used as model performance indicators. In the case of linear discriminant analysis models, accuracy and false negative predictions, as well as the difference in the predicted classes, are studied.

Calculations in this study were conducted using the R programming language [49]. For logit models, the stats package, for linear discriminant analysis models, the MASS package [50] was used. ggplot2 package was used to prepare the figures [51].

Table 6: Performance indicators of logit models [%]. AUC = "area under the curve".

Model #	accuracy	AUC	Model #	accuracy	AUC
1.3	75	81	**1.6**	79	91
2.3	78	93	**2.6**	88	94
3.3	88	96	**3.6**	91	94
4.3	91	97	**4.6**	91	94

3 Results

3.1 Logit models

Figures 2 and 3 show fit for one-variable logit models and their 95% confidence interval. Elevation (for type 3 only), artificial surfaces (type 6 only), agricultural and forested areas as well as three wastewater concentrations (except for nitrogen) show high significance levels. The significance of wastewater nitrogen is somewhat weaker (0.06 and 0.09 for type 3 and type 6 waters, respectively). The significance of the aggregated wastewater load (in population equivalent) is around 0.1 (0.08 for type 3 and 0.12 for type 6). As an overall tendency, significance levels are comparable for type 3 and type 6 watersheds. In addition to the already mentioned ones, the most important differences are: area – higher significance for type 6 watersheds; slope – much more significant for type 3 watersheds; wetland and water surfaces – higher significance for type 6.

Concerning the multivariate models, the performance indicators generally increase with the number of variables, although already the one-variable models 1.3 and 1.6 perform quite well: accuracy of 75 – 79% (Table 6). While in the case of Type 3, the elevation, in case of Type 6, the area of artificial surfaces is a better predictor. Adding the second wastewater indicator (COD) hardly adds anything to Type 3 models and adds nothing to type 6 models. The best accuracy is 91% for both types, and the best AUC is 97/94% for type 3 and type 6, respectively.

Tables 7 and 8 show confusion matrices for 2×2 selected models. These tables show the amount and type of errors. The terms "positive" and "negative" are used considering the water management point of view: a case (a specific water body) is regarded as positive when interventions/measures are needed to ensure it's good status (so it's status is in fact not good). Type II errors (false negatives) are the more severe errors: the waters where the need for a measure is not predicted although needed. In the presented models, Type II errors amount to 6 – 13% of the cases (red numbers in Tables 7 and 8).

3.2 Linear discriminant analysis models

Just as with the logit models, we start with a graphical investigation. This step can not be done if the number of predictor variables is higher than two, thus only models 2.3 and 2.6 are graphically investigated. As a first step, each water body is depicted on a 2D plane as a function of the model variables, with the color representing the status. Secondly, prediction areas of the models are filled up with color for the respective class. At the same time, an uncertainty analysis is conducted: reliability (confidence) of the models is tested with the bootstrapping method [34]: the model is fitted on a random 90% subsample of the training dataset ("submodels"). This step is repeated many times. Those points of the prediction area that belonged to the same class in 95% of the submodels are enclosed with a black line; those belonging to the same class in 80% of the models with a grey line on Figures 4 left and right.

The above figures – along with the multivariate models – also help us in concluding the role of the single variables. Considering the models,

- only water bodies with a forest share above 80% have a chance to be in high status;
- only water bodies with forest share above ~30% have an opportunity to achieve good status;
- water bodies where wastewater total phosphorus relative to long term mean flow is higher than 0.5 (type3) or 4 mg/l (type 6) are unlikely to achieve good status.

These numbers will be slightly different with the inclusion of other variables or with a different training dataset; what is essential now is that they can be defined.

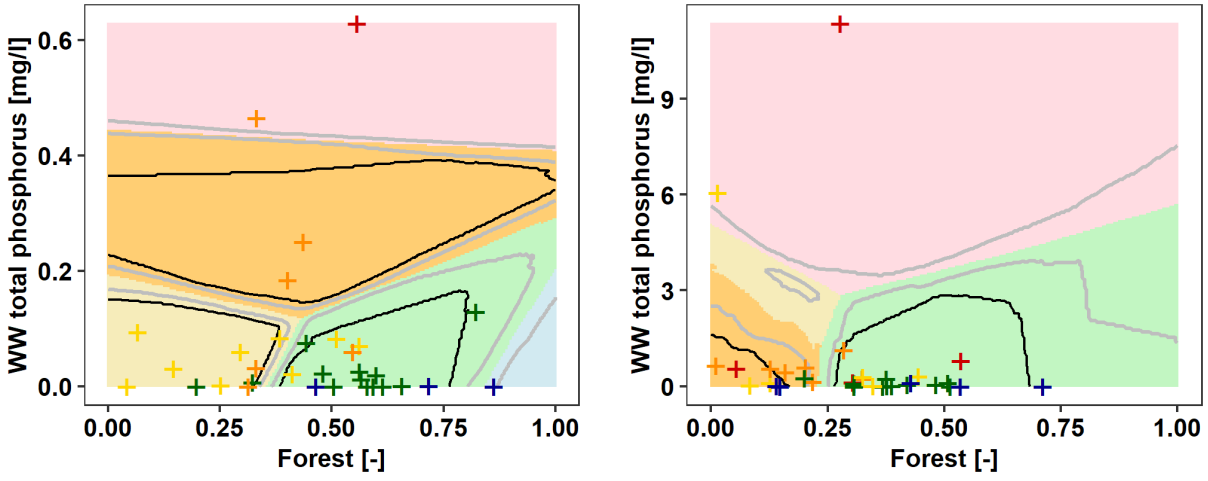
As already mentioned, rather than predicting if a water body achieves or not the good status, linear discriminant analysis models aim at predicting its status on a five-class scale. It is easy to understand: without any knowl-

Table 7: Confusion matrices for models 2.3 and 4.3. Red numbers indicate type II errors. Accur. = accuracy.

measured		model 2.3			model 4.3		
	total	good	not good	accur.	good	not good	accur.
good	16	13	3	81%	15	1	94%
not good	16	4	12	75%	2	14	88%
total	32	17	15	78%	17	15	91%

Table 8: Confusion matrices for models 2.6, and 4.6. Red numbers indicate type II errors. Accur. = accuracy.

measured		model 2.6			model 4.6		
	total	good	not good	accur.	good	not good	accur.
good	16	14	2	88%	15	1	94%
not good	17	2	15	88%	2	15	88%
total	33	16	17	88%	17	16	91%



"+": observations. Background color: prediction area of different status classes.
Grey and black lines: 80 and 95% confidence. Colors: high, good, moderate, poor, bad.

Figure 4: Composite graphs of models 2.3 (left) and 2.6 (right).

Table 9: Linear discriminant analysis models' performance indicators. Increment = increment in accuracy compared to the naïve model.

Model #	accuracy	increment	Model #	accuracy	increment
1.3	47	06	**1.6**	36	03
2.3	66	25	**2.6**	48	15
3.3	63	22	**3.6**	48	15
4.3	63	22	**4.6**	48	15

edge, the class (on a five-class scale) will be met with a 20% probability (blind model). Knowing the frequencies of the unique classes and presuming all cases in the most frequent class, a higher probability can be achieved. In our study, the accuracy of the so-called naïve model will be 13/32 = 41% and 11/33 = 33% for type 3 and type 6 models, respectively (Table 2). To quantify the performance difference between type 3 and type 6 models, the accuracy increments (compared to the naïve model) are also represented in Table 9.

Tables 10 and 11 contain confusion matrices for two- and four-variable linear discriminant analysis models.

Table 10: Confusion matrix for type 3 models. Red: false negative predictions.

measured		model 2.3						model 4.3					
	total	high	good	mod.	poor	bad	accur.	high	good	mod.	poor	bad	accur.
high	3	**1**	2	0	0	0	33%	**1**	2	0	0	0	33%
good	13	0	**11**	2	0	0	85%	1	**10**	2	0	0	77%
mod.	9	0	3	**6**	0	0	67%	0	4	**5**	0	0	56%
poor	6	0	1^x	2	**2**	1	33%	0	1^x	2	**3**	0	50%
bad	1	0	0	0	0	**1**	100%	0	0	0	0	**1**	100%
total	32	1	17	10	2	2	66%	2	17	9	3	1	63%

^{x}prediction two classes aside. *prediction three classes aside. Mod. = moderate; accur. = accuracy. Zeros not indicated.

Table 11: Confusion matrix for type 6 models. Red: false negative predictions.

measured		model 2.6						model 4.6					
	total	high	good	mod.	poor	bad	accur.	high	good	mod.	poor	bad	accur.
high	5	0	3	0	2*	0	0%	**0**	3	0	2*	0	0%
good	11	0	**10**	0	1^x	0	91%	0	**9**	0	2^x	0	82%
mod.	6	0	3	0	2	1^x	0%	0	3	**1**	2	0	17%
poor	7	0	2^x	0	**5**	0	71%	0	2^x	0	**5**	0	71%
bad	4	0	2*	0	1	**1**	25%	0	2*	0	1	**1**	25%
total	33	0	20	0	11	2	48%	0	19	1	12	1	48%

^{x}prediction two classes aside. *prediction three classes aside. Mod. = moderate; accur. = accuracy. Zeros not indicated.

Table 12: Water bodies misclassified by two or more classes by any of models 2.3 to 4.3 or 2.6 to 4.6.

Type	ID	wbname	measured	model 1.3 / 2.6	model 3.3 / 3.6	model 4.3 / 4.6
3	31	Dobroda-creek and tributaries	poor	good^x	good^x	good^x
6	61	Ferenc-channel	high	poor*	poor*	poor*
6	62	Hunyor-creek	high	poor*	poor*	poor*
6	63	Lanka-channel	good	poor^x	poor^x	poor^x
6	64	Tapolca-creek	good	good	poor^x	poor^x
6	65	Répce-spillway	mod	bad^x	bad^x	mod
6	66	Kõrös-ér	poor	good^x	good^x	good^x
6	67	Mirhó-Gyolcsi-channel	poor	poor	mod	good^x
6	68	Pécsi-víz middle	poor	good^x	poor	poor
6	69	Nádor-channel (Sárvíz) upper	bad	good*	good*	good*
6	70	Völgységi-creek to Rák-creek	bad	good*	good*	good*

^{x}prediction two classes aside. *prediction three classes aside. Mod. = moderate.

Two kinds of errors are investigated: first, false negatives (type II errors) and second, misclassification by two or more classes (the formers are marked with red, the latter with ^{x} and * in Tables 10 – 11). The number of false negative cases is 4 and 5 for models 2.3 and 4.3, respectively, and 7 for both of the models 2.6 and 4.6.

"Very big" mistakes (misclassification by three classes) only happen with type 6 models, however, in both directions. Their number is 2×2 with both models (2× underestimation by three classes, 2x overestimation by three classes. The number of "big" mistakes (over- or underestimation by two classes) is just the same for type 6 models. One type 3 case is overestimated by both of the models 2.3 and 4.3.

4 Discussion

Two times four logit models plus two times four linear discriminant analysis models were studied. All of them yielded consistent results, which indicates the suitability of the training data – modeling methods combinations for the required purpose.

Concerning the basic watershed properties (area, elevation, slope), only elevation is significant and only with the type 3 dataset. The cause of this fact might be that elevation of the watershed is in a strong correlation with land use: both settlements and agricultural activities tend to concentrate on lower areas. The next two most significant covariates are slope with type 3 watersheds (significance = 0.08) and area with type 6 (significance = 0.14). The slope of type 6 basins is not so indicative because it covers only a narrower range (Table 3). Relatively higher importance of catchment area on lowlands can be understood, taking into account that agricultural surfaces have a higher (62 versus 47) whereas forests a lower (30 versus 46) percentage on type 6 versus type 3 watersheds.

From the four land use covariates, agricultural and forested surfaces along with artificial surfaces turned out to be significant, this latter only with type 6. Non-significance of wetlands and waters might be explained by the fact that the extent of their effect very much depends on their location (near to outflow point versus close to the origin) [52] which was not accounted for in the models.

The aggregated wastewater indicator (load in population equivalent) turned out not to be significant. From the unique components of wastewater, both biochemical and chemical oxygen demand are significant with both types; however, biochemical oxygen demand more with type 3 and chemical oxygen demand more with type 6. From the nutrient indicators, only total phosphorus is significant, which emphasizes the role of point sources in phosphorus contamination of rivers [53].

Univariate logit models quantify the role of each watershed characteristic in water bodies' status. Concerning type 3 watersheds, those with a mean elevation above 400 meters above sea level or with an agricultural share < 30% or with a forest share > 60% will achieve good status (confidence > 95%). On the contrary, type 3 watersheds with elevation < 200m, agricultural share > 70% or forest share < 30% will not achieve good status. As for type 6 watersheds, forest share > 45% or phosphorus load below 0.1 mg/l are guarantees for achieving; whereas forest share < 15%, COD load above 15 mg/l or phosphorus above 0.5 mg/l for not achieving good status. Kändler *et al.* [31] also concluded that forest proportions bigger than 70% lead to low concentrations of contaminants; however, concerning arable land, their threshold was somewhat lower (40%).

Regarding the multivariate logit models, already the two-variable (forest + total phosphorus) models perform quite well: they show an accuracy of 78/88% and an "area under the curve" value of 93/94% for type 3 and type 6 models, respectively. Three- and four-variable models supersede these values.

Considering multivariate linear discriminant analysis models, two- or more variable models perform better than univariate ones. Concerning type 3 models, there is no significant difference between models 2.3, 3.3, 4.3; the two-variable model even performs slightly better. Concerning type 6 models, there is no difference at all, between models 2.6, 3.6, and 4.6.

The more marked difference (concerning discriminant analysis models) is between type 3 and type 6 models: the former perform much better. A possible cause for this is the presence of waters loaded with extremely high point source pollution in this data set (Figure 4 right).

A comparison of logit models with linear discriminant analysis models, in general, is not straightforward since they had a different objective. Counting the amount / proportion of water bodies misclassified to achieve good status, logit models perform better. The fact that linear discriminant analysis models treat classes as nominal variables serves as a reason for this. The finding is in line with Avila *et al.* [36], who concluded that multinomial regression models performed slightly better than linear discriminant analysis models in terms of cross-validation error rates.

The most frequent cause for status overprediction (by discriminant analysis models) is that water quality at the monitoring location is influenced by a near wastewater inlet (IDs 31, 67-70, Table 12) or industrial wastewater (ID 69) or a fishing pond (ID 66). Industrial wastewater was not accounted for in the models due to a lack of data. In many cases, a more recent study [54] assessed a status closer to the one predicted by the models (IDs 31, 63, 65-70).

Two of the under-predicted waters are water diversion channels (IDs 61 and 65, Table 12); water quality here is determined not by the own watershed, instead by the source water (Duna and Répce). Hunyor-creek (ID 62) is a small creek, with mainly of agricultural land use on the watershed. However, the monitoring point far upstream. The proportion of forests on the basin of the monitoring point is much higher than on the watershed draining to the water body outflow point. Tapolca-creek (ID 64) has a high wastewater share; however, effluent wastewater thresholds in this region are stricter than in the other areas of

the country due to the vulnerability of the Lake Balaton (receptor of the Tapolca-creek).

Bearing the false predictions in mind, future developments should be the exclusion of water diversion channels (where their watershed does not determine water quality). Also, the next models should account for the distance between the monitoring location and the source of wastewaters (including industrial plants and fishing ponds).

5 Conclusions

Both logit and linear discriminant analysis models are useful in predicting if a water body achieves good status and / or it's status class. The most significant covariates for both upland and lowland rivers, with both of the logit and linear discriminant analysis methods were the share of agricultural land and the share of forests, the organic wastewater indicators and wastewater TP load. Models perform better on upland than on lowland rivers.

Achievement of good status could be predicted with an accuracy of ~90% (with 5-variable logit models), whereas the status class with an accuracy of 72/55% (with 5-variable linear discriminant analysis models) for upland and lowland rivers, respectively.

Acknowledgement: The Higher Education Excellence Program of the Ministry of Human Capacities, Hungary supported the research presented in this paper in the frame of the Water sciences and Disaster Prevention research area of the Budapest University of Technology and Economics (BME FIKP-VÍZ).

References

[1] Congress, U.S., 1972: *Federal water pollution control act*, 33 U.S.C. 1251 et seq. USA.

[2] *Directive 2000/60/EC of the European Parliament and of the Council*, 2000. European Comission, Bruxelles.

[3] Boda, P., Móra, A., Deák, C., Krasznai, E., Csercsa, A., Zagyva, A., & Várbíró, G., 2014: Testing the adequacy of the Hungarian typological system on the watercourses of the Ipoly basin, based on the macroinvertebrate communities. *Acta Biologica Debrecina, Suppl. Oecol. Hung.*, *32*, 9–18.

[4] Borics, G., Ács, É., Boda, P., Boros, E., Erős, T., Grigorszky, I., Kiss, K.T., & Lengyel, S., 2016: Water bodies in Hungary – an overview of their management and present state. *Hungarian Journal of Hydrology*, *86*, 57–67.

[5] Clement, A., Szilágyi, F., & Kardos, M.K., 2015: Classification of surface waters based on physico-chemical characteristics supporting ecology - lessons learned during status assessment and the planning of interventions In: *Proceedings of the XXXIII. National Meeting of the Hungarian Hydrological Society* (In Hungarian: Felszíni vizek minősítése az ökológiát támogató fizikai-kémiai jellemzők szerint - az állapotértékelés tanulságai az intézkedési programok tervezése szempontjából, In: *A Magyar Hidrológiai Társaság XXXIII. Vándorgyűlése*). 1-3 July 2015, Szombathely, Hungary (ed. Szlávik, L., Gampel, T. & Szigeti, E.). Hungarian Hydrological Society, pp 1–11.

[6] Clement, A., & Szilágyi, F., 2015: Physico-chemical status evaluation of surface water bodies – River Basin Management Plan background document no 6-2. (In Hungarian: Felszíni víztestek fizikai kémiai állapotértékelési rendszere. OVGT 6-2 háttéranyag). Budapest, 1–15 p. Downloadable from http://www.vizugy.hu/vizstrategia/documents/988BF7DB-B869-46C6-9463-E9E4BFC81D2A/6_2_hatteranyag_Fizikokemiai_minosites.pdf. Accessed 01/Nov/2019.

[7] Dworak, T., Gonzalez, C., Laaser, C., & Interwies, E., 2005: The need for new monitoring tools to implement the WFD. *Environmental Science and Policy*, *8*, 301–306. doi:10.1016/j.envsci.2005.03.007

[8] Hering, D., Borja, Á., Carstensen, J., Carvalho, L., Elliott, M., Feld, C.K., Heiskanen, A.S., Johnson, R.K., Moe, J., Pont, D., Solheim, A.L., & de Bund, W. van, 2010: The European Water Framework Directive at the age of 10: A critical review of the achievements with recommendations for the future. *Science of the Total Environment*, *408*, 4007–4019. doi:10.1016/j.scitotenv.2010.05.031

[9] Kerekes-Steindl, Z., 2016: Water quality protection in Hungary - policy and status. *Hungarian Journal of Hydrology*, *96*, 43–56.

[10] Tyler, A.N., Hunter, P.D., Spyrakos, E., Groom, S., Constantinescu, A.M., & Kitchen, J., 2016: Developments in Earth observation for the assessment and monitoring of inland, transitional, coastal and shelf-sea waters. *Science of the Total Environment*, *572*, 1307–1321. doi:10.1016/j.scitotenv.2016.01.020

[11] Carvalho, L., Mackay, E.B., Cardoso, A.C., Baattrup-Pedersen, A., Birk, S., Blackstock, K.L., Borics, G., Borja, Á., Feld, C.K., Ferreira, M.T., Globevnik, L., Grizzetti, B., Hendry, S., Hering, D., Kelly, M., Langaas, S., Meissner, K., Panagopoulos, Y., Penning, E., Rouillard, J., Sabater, S., Schmedtje, U., Spears, B.M., Venohr, M., van de Bund, W., & Solheim, A.L., 2019: Protecting and restoring Europe's waters: An analysis of the future development needs of the Water Framework Directive. *Science of The Total Environment*, *658*, 1228–1238. doi:10.1016/j.scitotenv.2018.12.255

[12] Chapman, D.V., Bradley, C., Gettel, G.M., Hatvani, I.G., Hein, T., Kovács, J., Liska, I., Oliver, D.M., Tanos, P. & Trásy, B., 2016: Developments in water quality monitoring and management in large river catchments using the Danube River as an example. *Environmental Science & Policy* 64, pp. 141–154. doi:10.1016/j.envsci.2016.06.015

[13] Kovács, J., Kovács, S., Hatvani, I.G., Magyar, N., Tanos, P., Korponai, J. & Blaschke, A.P., 2015: Spatial Optimization of Monitoring Networks on the Examples of a River, a Lake-Wetland System and a Sub-Surface Water System Water Resources Management 29:14 pp. 5275-5294. doi:10.1007/s11269-015-1117-5

[14] Tanos, P., Kovács, J., Kovács, S., Anda, A. & Hatvani, I.G., 2015: Optimization of the monitoring network on the River Tisza (Central Europe, Hungary) using combined cluster and discriminant analysis, taking seasonality into account. *Environmental Monitoring & Assessment*, 187, pp. 1-14. doi: 10.1007/s10661-015-4777-y

[15] Singh, K.P., Malik, A., Mohan, D., & Sinha, S., 2004: Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomti River (India) - A case study. *Water Research*, *38*, 3980–3992. doi:10.1016/j.watres.2004.06.011

[16] Giri, S., & Qiu, Z., 2016: Understanding the relationship of land uses and water quality in Twenty-First Century: A review. *Journal of Environmental Management*, *173*, 41–48. doi:10.1016/j.jenvman.2016.02.029

[17] Kardos, M.K. & Clement, A. 2019: Similarities among small watercourses based on multiparameter physico-chemical measurements. *Central European Geology* (accepted for publication)

[18] Chapra, S.C., 1997: *Surface Water-quality modeling*. McGraw-Hill, New York, 1–844 p.

[19] Arnold, J.G., Srinivasan, R., Muttiah, R.S., & Williams, J.R., 1998: Large area Hydrologic Modeling and Assessment Part I: Model development "Basin scale model called SWAT (Soil and Water speed and storage, advanced software debugging policy to meet the needs, and the management to the tank model)." *American Water Resources Association*, *34*, 73–89. doi:10.1111/j.1752-1688.1998.tb05961.x

[20] Tsakiris, G., & Alexakis, D., 2012: Water quality models: An overview. *European Water 37*, 33–46.

[21] Jaafari, A., Najafi, A., Rezaeian, J. & Sattarian, A, 2015: Modeling erosion and sediment delivery from unpaved roads in the north mountainous forest of Iran. *Int J Geomathematics 6*. 343–356. doi:10.1007/s13137-014-0062-4.

[22] Xie, X., Norra, S., Berner, Z., & Stüben, D., 2005: A GIS-supported multivariate statistical analysis of relationships among stream water chemistry, geology and land use in Baden-Württemberg, Germany. *Water, Air, and Soil Pollution*, *167*, 39–57. doi:10.1007/s11270-005-0613-2

[23] Rothwell, J.J., Dise, N.B., Taylor, K.G., Allott, T.E.H., Scholefield, P., Davies, H., & Neal, C., 2010: Predicting river water quality across North West England using catchment characteristics. *Journal of Hydrology*, *395*, 153–162. doi:10.1016/j.jhydrol.2010.10.015

[24] Angyal, Z., Sárközi, E., Gombás, Á., & Kardos, L., 2016: Effects of land use on chemical water quality of three small streams in Budapest. *Open Geosciences*, *8*, 133–142. doi:10.1515/geo-2016-0012

[25] Allan, D.J., & Arbor, A., 2004: The Influence of Land Use on Stream Ecosystems. *Annual Review of Ecology and Systematics*, *35*, 257–284.

[26] Mehaffey, M.H., Nash, M.S., Wade, T.G., Ebert, D.W., Jones, K.B., & Rager, A., 2005: Linking land cover and water quality in New York City's water supply watersheds. *Environmental Monitoring and Assessment*, *107*, 29–44. doi:10.1007/s10661-005-2018-5

[27] Barclay, J.R., Tripp, H., Bellucci, C.J., Warner, G., & Helton, A.M., 2016: Do waterbody classifications predict water quality? *Journal of Environmental Management*, *183*, 1–12. doi:10.1016/j.jenvman.2016.08.071

[28] Varol, M., Gökot, B., Bekleyen, A., & Şen, B., 2012: Spatial and temporal variations in surface water quality of the dam reservoirs in the Tigris River basin, Turkey. *Catena*, *92*, 11–21. doi:10.1016/j.catena.2011.11.013

[29] Zhou, P., Huang, J., Pontius, R.G., & Hong, H., 2016: New insight into the correlations between land use and water quality in a coastal watershed of China: Does point source pollution weaken it? *Science of the Total Environment*, *543*, 591–600. doi:10.1016/j.scitotenv.2015.11.063

[30] Bostanmaneshrad, F., Partani, S., Noori, R., Nachtnebel, H.P., Berndtsson, R., & Adamowski, J.F., 2018: Relationship between water quality and macro-scale parameters (land use, erosion, geology, and population density) in the Siminehrood River Basin. *Science of the Total Environment*, *639*, 1588–1600. doi:10.1016/j.scitotenv.2018.05.244

[31] Kändler, M., Blechinger, K., Seidler, C., Pavlů, V., Šanda, M., Dostál, T., Krása, J., Vitvar, T., & Štich, M., 2017: Impact of land use on water quality in the upper Nisa catchment in the Czech Republic and in Germany. *Science of the Total Environment*, *586*, 1316–1325. doi:10.1016/j.scitotenv.2016.10.221

[32] Vigiak, O., Grizzetti, B., Udias-Moinelo, A., Zanni, M., Dorati, C., Bouraoui, F., & Pistocchi, A., 2019: Predicting biochemical oxygen demand in European freshwater bodies. *Science of the Total Environment*, *666*, 1089–1105. doi:10.1016/j.scitotenv.2019.02.252

[33] Hosmer, D.W., & Lemeshow, S., 1989: *Applied logistic regression*. John Wiley & Sons, New York, 1–307 p.

[34] Hastie, T., Tibshirani, R., & Friedman, J., 2009: *The Elements of Statistical Learning*, Springer, 1–745 p. doi:10.1007/b94608

[35] O'Dwyer, J., 2014: Microbiological contamination of Private Water Wells in the Midwest region of Ireland: investigation of water quality, public awareness and the application of Logistic Regression in contaminant modelling. *THESIS PhD*,. University of Limerick, 1–240 p.

[36] Avila, R., Horn, B., Moriarty, E., Hodson, R., & Moltchanova, E., 2018: Evaluating statistical model performance in water quality prediction. *Journal of Environmental Management*, *206*, 910–919. doi:10.1016/j.jenvman.2017.11.049

[37] Wunderlin, A.D., Díaz, M., Amé, M. V., Pesce, F.S., Hued, A.C., & Bistoni, M., 2001: Pattern recognition techniques for the evaluation of spatial and temporal variations in water quality. A case study: Suquía River basin (Córdoba-Argentina). *Water Research*, *35*, 2881–2894. doi:10.1016/S0043-1354(00)00592-3

[38] Hatvani, I.G., Clement, A., Kovács, J., Kovács, I.S., & Korponai, J., 2014: Assessing water-quality data: The relationship between the water quality amelioration of Lake Balaton and the construction of its mitigation wetland. *Journal of Great Lakes Research*, *40*, 115–125. doi:10.1016/j.jglr.2013.12.010

[39] Wang, Y. Bin, Liu, C.W., Liao, P.Y., & Lee, J.J., 2014: Spatial pattern assessment of river water quality: Implications of reducing the number of monitoring stations and chemical parameters. *Environmental Monitoring and Assessment*, *186*, 1781–1792. doi:10.1007/s10661-013-3492-9

[40] General Directorate of Water Management, Hungary, 2016: Hungarian Part of the Danube River Basin - River Basin Management Plan. (In Hungarian: A Duna- vízgyűjtő magyarországi része - Vízgyűjtőgazdálkodási terv) 2015. Downloadable from http://www.vizugy.hu/vizstrategia/documents/E3E737A3-3EBC-4B6F-973C-5DD9B8A6DBAB/OVGT_foanyag_vegleges.pdf. Accessed 01/Nov/2019.

[41] Clement, A., Jolánkai, Zs., Kardos M.K., 2015: River Basin Management Planning results concerning urban water management: The role of municipal wastewater treatment in surface water quality and the planned measures. (In Hungarian: A vízgyűjtőgazdálkodási tervezés települési vízgazdálkodással kapcsolatos eredményei: A kommunális szennyvíztisztítás szerepe a felszíni vízminőség alakulásában és a tervezett intézkedések). Hírcsatorna 5. pp 1-11.

[42] The working group on water bodies 2003: Guidance Document No 2. - Identification of Water Bodies (Common Implementation Srategy for the Water Framework Directive) Report. Downloadable from https://circabc.europa.eu/sd/a/655e3e31-3b5d-4053-be19-15bd22b15ba9/Guidance%20No%202%20-%20Identification%20of%20water%20bodies.pdf. Accessed 01/Sep/2015

[43] Copernicus, L.M.S., 2016a: European Digital Elevation Model (EU-DEM), version 1.1. URL https://land.copernicus.eu/imagery-in-situ/eu-dem/eu-dem-v1.1 (accessed 6.1.19).

[44] Tarboton, D.G., 1997: A new method for the determination of flow directions and upslope areas in grid digital elevation models. *Water Resources Research*, *33*, 309–319.

[45] Copernicus, L.M.S., 2016b: Corine Land Cover (CLC) 2012, Version 18. URL https://land.copernicus.eu/pan-european/corine-land-cover/clc-2012 (accessed 1.1.18).

[46] European Environment Agency, 2015: Waterbase - UWWTD: Urban Waste Water Treatment Directive – reported data. Downloadable from eea.europa.eu/themes/water/european-waters/water-use-and-environmental-pressures/uwwtd. Accessed 01/Sep/2019

[47] Somlyódy, L., & Patziger, M., 2012: Urban wastewater development in Central and Eastern Europe. *Water Science and Technology*, *66*, 1081–1087. doi:10.2166/wst.2012.289

[48] General Directorate of Water Management, Hungary, 2016: Wastewater Load data. Supplement no. 3-1 to the Hungarian River Basin Management Plan. (In Hungarian: 3-1. melléklet az Országos Vízgyűjtőgazdálkodási Tervek 2015. évi felülvizsgálatához: Szennyvízterhelés jellemzői: kommunális és ipari szennyvízkibocsátás). Downloadable from http://www.vizugy.hu/vizstrategia/documents/10B9EE2E-D889-4C94-815D-5CB2D53C846A/3_1_melleklet_szennyvizterheles.xls. Accessed 01/Aug/2019.

[49] R Core Team, 2019: R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria, https://www.R-project.org/.

[50] Venables, W.N., & Ripley, B.D., 2002: *Modern Applied Statistics with S*. Springer, 1–495 p.

[51] Wickham, H., 2009: *ggplot2: Elegant Graphics for Data Analysis*. Springer Verlag, New York.

[52] Venohr, M., Donohue, I., Fogelberg, S., Arheimer, B., Irvine, K., & Behrendt, H., 2003: Nitrogen retention in a river system under consideration of the river morphology and occurrence of lakes Diffuse Pollution Conference , Dublin 2003 1C Water Resources Management. *Diffuse Pollution Conference*, 61–67.

[53] Clement, A., & Buzás, K., 1999: Use of ambient water quality data to refine emission estimates in the Danube basin. *Water Science and Technology*, *40*, 35–42.

[54] General Directorate of Water Management, Hungary, 2018: Study required to comply with the nitrate directive - Physico-chemical status assessment - Summary (In Hungarian: Nitrát Irányelvnek történő megfeleléshez szükséges vizsgálatok - Általános kémiai állapotértékelés - összefoglaló). Project Report.