PATH MODELLING ANALYSIS OF POLLUTION SOURCES AND ENVIRONMENTAL CONSEQUENCES IN RIVER BASINS

ANTÓNIO FERNANDES[1], ANA FERREIRA[1], LUÍS SANCHES FERNANDES[1],
RUI CORTES[1] & FERNANDO PACHECO[2]
[1]Centre for the Research and Technology of Agro-Environment and Biological Sciences (CITAB),
Universidade de Trás-os-Montes e Alto Douro, Portugal
[2]Vila Real Chemistry Research Centre (CQVR), Universidade de Trás-os-Montes e Alto Douro, Portugal

ABSTRACT

The Portuguese Index of Macroinvertebrates is used as a pollution index in the study of surface water in Portugal. From an environmental perspective, it is necessary to determine the pollution sources affecting surface water. Not only direct discharges of industrial and urban effluents are the cause; possibly the combination of other pollutions sources could result in ecological loss. To comprehend all the cause and effect relationships between pollution sources, water contamination and ecological integrity, it is necessary to apply complex models and possibly apply the use of thorough statistical tools. Structural Equation Modelling (SEM) has been used in the social sciences for a long time. Due to the present environmental concern and awareness of the phenomenon's complexity, SEM was used for environmental studies. In this paper are present SEM-PLS models that are applied to collected data from two different river basins, that of the Ave and the Sabor, in order to understand which are their main pollution sources and which contaminants are restraining biodiversity. The applied models from each basin reproduced different realities, as expected, since the river Ave has has the notorious impact of industry, while the Sabor basin has a higher level of water quality.

Keywords: macroinvertebrates, modelling, pollution modelling, Portuguese Index of Macroinvertebrates, river pollution, Structural Equation Modelling.

1 INTRODUCTION

To study water quality of river basins, it is crucial to understand that there is a multitude of phenomena that affect biodiversity, such as several pollution sources, different contaminants, several reactions, morphological characteristics and even climate data. To comprehend the relationships between these variables, it is necessary to use advanced statistical tools such as Structural Equation Modelling (SEM). Generically, there are two types of SEM [1]: CB-SEM (covariance-based SEM), and PLS-SEM (Partial Least Squares SEM) which is also called PLS-PM (PLS path modelling). In the first case, the estimation procedure is based on a maximum likelihood estimation, while PLS-SEM is based on ordinary least squares regression [2]. The benefits of each type of SEM differ from case study to case study, so the opinion on which type is the most appropriate method is still divergent [1], [3].

Initially, SEM was used in the social sciences; nowadays it has been increasingly used in non-social sciences, such as environmental and biological issues. In an environmental study, PLS-PM was used to understand ground-level ozone concentrations, considering meteorology, chemical reactions and the presence of primary pollutants as main causes [4]. Already in 1994 SEM was applied to study surface water quality: natural and anthropogenic effects were declared as impossible to quantify [5]. SEM was used to understand the relationship between several concentrations of substances with the total dissolved solids in ground water [6]. The comparison of 3 types of pollution (organic, sediment and eutrophication in surface waters) was applied to Feitsui Reservoir Watershed [7], with the conclusion that sediment pollution was the main cause. From a social standpoint, SEM was

WIT Transactions on Ecology and the Environment, Vol 228, © 2018 WIT Press
www.witpress.com, ISSN 1743-3541 (on-line)
doi:10.2495/WP180091

applied to comprehend community awareness about local water quality, through the usage of surveys [8]. For the Surabaya river [9], the citizens' awareness was accounted for in a model that included variables such as oxygen demands and total solids, to understand their relationship with water pollution levels using two water quality indexes, and concluding that the application of more variables would strengthen the model. The comparison of all studies about water quality using SEM led to the understanding that there are a multitude of variables to apply.

The purpose of this study is to understand the relationship between pollution sources (named/given as pressures), the concentration of contaminants in surface waters and ecological integrity, as measured by the Portuguese Index of Macroinvertebrates [10]. SEM-PLS was applied in formative models, establishing a cause and effect model for two distinct river basins, the Ave and the Sabor, to act as industrialized and rural river basins, respectively.

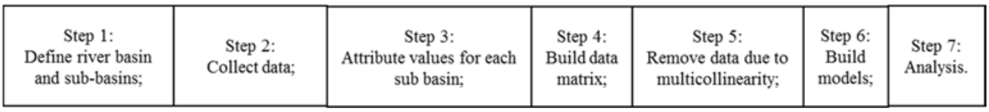
2 METHODOLOGY

To create the structural equation models, a seven-step methodology was followed (Fig. 1).

For Step 1, the river basins and sub-basins were delineated using ArcHydro [11]. The collected data for models was divided into three groups: water contamination, pressures and ecological integrity, according to the conceptual model (Fig. 2). It is known that the biodiversity in surface waters is dependent on the presence of contaminants, while the pollutions sources are the pressures in surface waters which can cause contamination. In any case, it was established that a connection between pressures and ecological integrity was established, although the effect is indirect.

For Ecological Integrity (Fig. 2), we used only one variable, IPtI_N, which is a numerical indicator that represents the biodiversity of macroinvertebrates in the north zone of Portugal: high values are indicative of water quality [10]. For water contamination, we chose 11 concentrations of contaminants, As, Cr, Cu, Fe, Pb, Zn, the oxygen demands BOD_5 and COD, the concentration of nutrients NO_3 and PO_4, and also the total suspended solids (TSS).

For the pollution sources in pressures (Fig. 2), we collected a vast number of variables such as the industrial and urban discharges in surface waters and underground waters of phosphorous, nitrogen, the oxygen demands (CBO_5 and COD), the percentage each sub-



Step 1: Define river basin and sub-basins;	Step 2: Collect data;	Step 3: Attribute values for each sub basin;	Step 4: Build data matrix;	Step 5: Remove data due to multicollinearity;	Step 6: Build models;	Step 7: Analysis.

Figure 1: Seven-step methodology.

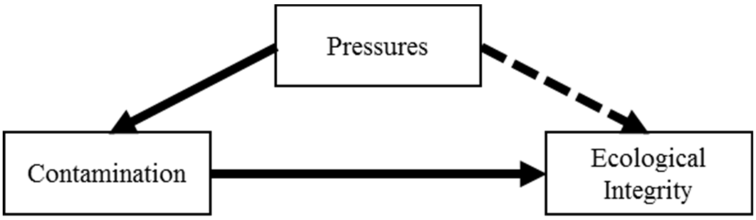


Figure 2: Conceptual model.

basin is covered with artificial and agricultural land uses, the percentage of land covered with land use conflicts [12], the population density, soil loss, wildfire risk, nitrogen and phosphorous discharges from agriculture and forest, and from livestock production. After collecting all the data and assigning a value for each sub-basin, a data matrix was built for each sub basin (step 4 of Fig. 1), where the columns are variables while each row is a sub-basin. Variables were discarded from the study in order to have a Pearson correlation coefficient between each variable of the same group below 0.8, and a variance inflation factor (VIF) below 5, because for values higher than this one, it means that a model with this data has multicollinearity [13].

In SEM-PLS there are two types of variables used: measured and latent. Latent variables are defined by the operator, while measured variables are the ones collected for the model. For this study, we chose to use 3 latent variables, according to the conceptual model: pressures, contamination and ecological integrity. We built a total of four models, two for each basin. In the first models (Ave 1 and Sabor 1) the latent variable pressures are connected to the ecological integrity, while in the second models (Ave 2 and Sabor 2) there is no direct connection between these two latent variables. SmartPLS [14] was used to build the models, which can be be either reflective or formative. In the first case, the measured variables were composed by the latent variables; while in the formative case, the latent variables are composed by the measured values. We chose to use formative models, due to the nature of the data. The aspect of a SEM is presented in Fig. 3.

Each yellow square represents a measured variable (MV), while blue circles are latent variables (LV). The algorithm attributes values for weights (w) for MVs and path coefficients for LVs, through and iterative process, in order to achieve the highest determination coefficient (R^2) for each latent variable. The model calculates a measured score for latent variables, based on the measured variables that compose it: eqn (1). For latent variables that are composed by other ones, such as B and C (Fig. 3), the model calculates a predicted score, as is demonstrated in eqn (2), for LV_C from Fig. 3.

$$\text{Measured score: } LV_i = \sum_{i=1}^{n} (MV_i \times W_i) \tag{1}$$

$$\text{Predicted score: } LV_c = LV_a \times pc_{ac} + LV_b \times pc_{bc} \tag{2}$$

The determination coefficient is calculated for latent variables that have and measured and predicted scores. Because the present study is a comparison between models with a

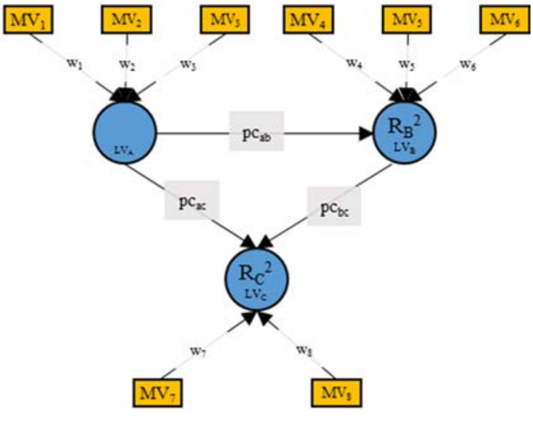


Figure 3: Structural equation model (SEM).

different number of data points in each sub-basin and even a different number of measured variables, we chose to use an adjusted R-Squared [3], instead of R-Squared, eqn (3).

$$R_{adj}{}^{2}=1-(1-R^{2})\times\frac{(n-1)}{(n-k-1)} \tag{3}$$

For this adjusted R, *n* is the total sample size, while *k* is the number of predictors. Normally, as more predictors are used in a model, the tendency of the R-squared value is to increase, so the adjustment of R-squared is useful for the kind of models that may include a high number of predictors.

3 RESULTS AND DISCUSSION

The occupied area of the Ave river basin is 1321 km^2 while for the Sabor, it is 3513 km^2. Since the Sabor is an international river, we only considered the zone within the Portuguese territory, thus reducing the study area to 2969 km^2. The number of delineated sub-basins was 92 for the Ave and 100 for the Sabor. Fig. 4 presents both river basins, their sub-basins and interpolated values of IPtI$_N$.

As shown in Fig. 4, the IPtI$_N$ in the Ave river basin is much lower than in the Sabor basin. Biodiversity lowers from upstream to downstream, not only due to an accumulation effect, but due to the fact that important pressures such as industrial discharges and higher population density are closer to the basin outlet. For the Sabor basin, the locations with low biodiversity are the sub-basins 11, 12, 13 and 14 (Fig. 4). Around these sub-basins, there is urban discharge that is possibly the cause.

Since all the variables are normalized (thus, mean of each one becomes equal to 0 and standard deviation is equal to 1), before being applied to the model, it is possible to do a comparison of path coefficients and weights. For each model, measured variables where chosen in order to eliminate multicollinearity issues, such as with inflated weights. We built a total of 4 models (Fig. 5), 2 for the river Ave (models A1 and A2) and two for the river Sabor (models S1 and S2). Initially, only models A1 and S1 were built, according to the conceptual model, and Pressures (P) were connected to Contamination (C) and Ecological Integrity (EI), with C connected to EI. We noticed that for A1 the results were reliable, had

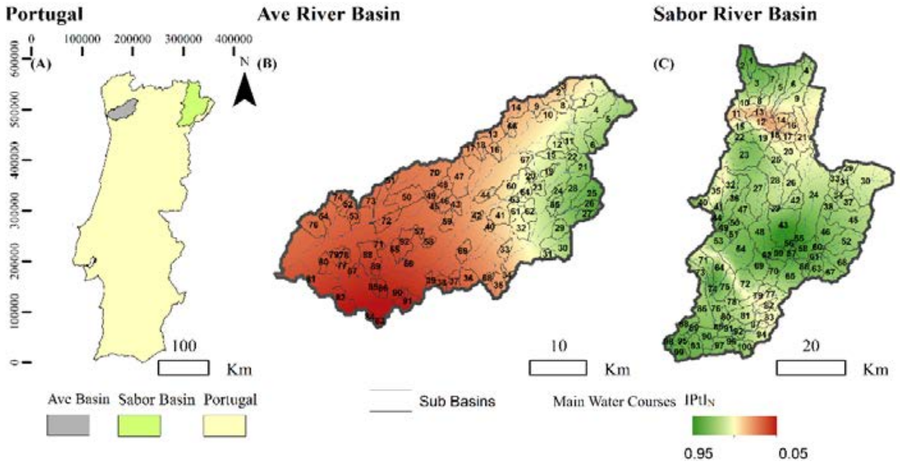


Figure 4: Map of Portugal (A) with the Ave (B) and Sabor (C) river basins, IPtI$_N$ values.

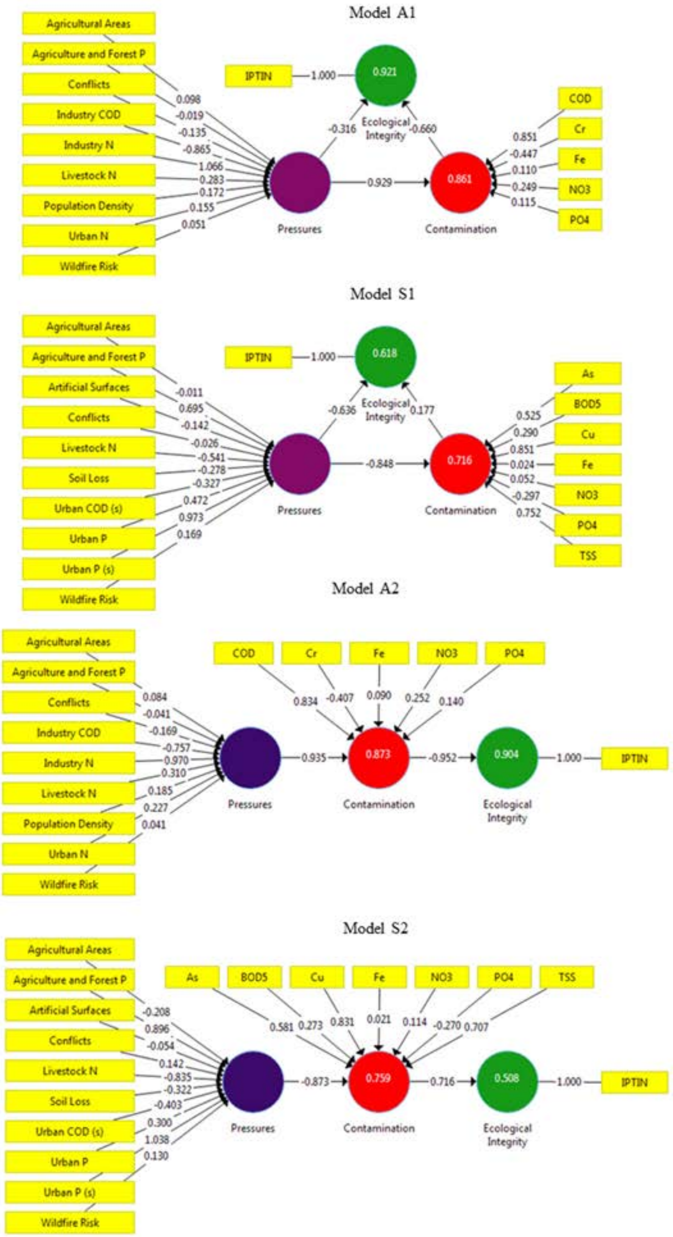


Figure 5: Structural equation models.

a highly adjusted R-squared, 0,861 for C and 0,921 for EI; and the path coefficients according to theory, from Pressures to Contamination the path coefficient is positive (0.929), which means that pressures are the cause of the high concentrations of contaminants.

Negative path coefficients from P to EI (–0.316), and from C to EI (–0.660), mean that identified pressures and contaminants are the ones that decrease EI, even more by the effect of C in EI is the higher P in EI. On the other hand, model S1 achieved lower values of adjusted R-squared: 0.716 for C and 0.618 for EI. Moreover, the path coefficients from P to C, and C to EI are unexpected, since it was foreseen that they would have the same signal as Model A1. Possibly the contaminants that lower ecological integrity in the Sabor river basin are not included in this study and that is why the path coefficients connected to contamination are unexpected; however, the chosen pressures have impact in ecological integrity, because the path coefficient between P and EI is negative. Besides, model A1 is representative of the reality, the path coefficients might be inflated, the latent variables (C and P) that compose EI appear to be collinear, as their VIF is higher than 5 (Table 1), while for model S1 the VIF values are below 5. In order to create a more reliable model, we removed the connection between P to EI, so the A1 model would have no multicollinearity, creating model A2. The same was done for model S1, in order to understand if the sign of the path coefficients would change, creating model S2.

From Model A1 to Model A2, and from Model S1 to Model S2, we noticed a small change for the adjusted R-Squared value; for the latent variable C it increases, while for EI decreases. C adjusted R-Square increases due to the fact that for latent variable P the weights that compose it are calculated in order to increase the determination of the coefficient for C (in Model 2) while in the models A1 and S1, the weights of MVs that compose P were calculated in order to increase both determination coefficients (for C and EI); that is why in the models, the adjusted R-squared lowered from Models 1 to 2. In Model A2, the path coefficients remain according to theory, in the path coefficient between P and C there are no significant changes, but the path coefficient between C and EI increases in the module, due to the fact that it absorbs the effect of P in EI. For Model S2, the path coefficients are totally discordant, because the pressures can’t decrease C, and C cannot increase EI. For the river Sabor, the best model is certainly Model S1, because for this river basin, there is at least one reliable path coefficient from P to EI (Fig. 5).

We analysed the product of total effect and the weight of each variable (Table 2). For example, in Model A1 the direct effect of P in EI is –0.316, but since P is connected to C, there is and indirect effect of P in EI that is equal to –0.613 (–0.660×0.929), so the total effect of P in EI is equal to the sum of direct and indirect effect, which results in –0.928. As an example, the product of total effect by Population Density is –0.160. Besides, if the weight of some measured variables is positive, it does not mean that in reality they increase biological diversity, but that the model attributes positive values for these variables, since in order to maximize the R-Squared value some of the weights must have a positive sign, or because they have a positive correlation and do not explain the variation. Besides the unreliable path coefficients in models S1 and S2, it is still possible to identify which are the measured variables that have a negative effect in ecological integrity. Through Table 2 it is possible to identify variables that lower EI (for Models S1 and S2) are Agriculture and Forest P, Urban P, Urban P (s), Wildfire Risk and PO_4. Since Conflicts have a positive effect in S1 and a negative effect in S2, it is uncertain if the real effect of this variable is positive or negative. For the Ave river basin, the variables that do not lower EI are Cr, Agriculture and Forest P, Conflicts and Industry COD.

For both river basins, the models show that effluent discharges of nutrients are a cause to lower EI, but not oxygen demands. In water treatment stations, one of the major concerns is to lower oxygen demands: in order to do this some biological treatments require the addition of nutrients, which results in low values of oxygen demand, but a high concentration of nutrients. For the Ave river basin, all contaminants have a negative impact in EI; only Cr does lower EI, probably due to its low concentration in surface waters, while PO_4 is the only contaminant that restrains biodiversity for the Sabor river basin.

Table 1: Inner model VIF for model Ave 1 and Sabor 1.

Model Ave 1			Model Ave 2		
	C	**EI**		**C**	**EI**
C		**7.263**	**C**		3.556
P	**1.000**	**7.263**	**P**	**1.000**	3.556
Model Sabor 1			Model Sabor 2		
	C	**EI**		**C**	**EI**
C		**1.000**	**C**		**1.000**
P	**1.000**		**P**	**1.000**	

Table 2: Product total effect and the weight for each measured variable.

	A1	A2
COD	-	-
Cr	+	+
Fe	-	-
NO3	-	-
PO4	-	-
Agricultural areas	-	-
Agriculture and forest P	+	+
Conflicts	+	+
Industry COD	+	+
Industry N	-	-
Livestock N	-	-
Population density	-	-
Urban N	-	-
Wildfire risk	-	-

	S1	S2
As	+	+
BOD5	+	+
Cu	+	+
Fe	+	+
NO3	+	+
PO4	-	-
TSS	+	+
Agricultural areas	+	+
Agriculture and forest P	-	-
Artificial surfaces	+	+
Conflicts	+	-
Livestock N	+	+
Soil loss	+	+
Urban COD (s)	+	+
Urban P	-	-
Urban P (s)	-	-
Wildfire risk	-	-

Through the analysis of the models, it is possible to establish an order of pressures which require environmental intervention. By decreasing the order of severity, the pressures that require intervention in the Ave River basin are industrial discharges of nutrients, livestock production discharges in underground water and urban discharges. Besides wildfire risk and agricultural areas having a negative impact, they should not be considered as dangerous, since the weight of these variables in the models A1 and A2 are considerably low. For the Sabor River Basin, the pressures of most concern are urban discharges, phosphorous in soil, forest and agriculture-based flows of nutrients, and urban discharges into surface waters.

4 CONCLUSIONS

The application of SEM-PLS models to the Sabor River and the Ave River datasets, which describe multiple pressures, surface water quality and biodiversity loss, proved to be efficient in discriminating factors that explain biodiversity loss. The chosen dataset for the Ave basin proved to be descriptive of the reality, while for the Sabor river basin, it appears that other phenomenona and variables should be considered in order to improve the models' reliability, such as morphological data. Possibly the fact that the Ave river basin is polluted resulted in an explanatory and reliable model; while for a clean basin such as the Sabor, it is harder to explain biodiversity loss. For the application of these models in other river basins, we advise to always collect as much data as possible and then proceed to the analysis of the correlation matrix, as some variables can explain each other. For further studies, we recommend the use of the transformation of variables and even the use of reflective models, in order to compare their results.

ACKNOWLEDGEMENTS

This research was funded by the INTERACT project: "Integrated Research Environment, Agro-Chain and Technology", no. NORTE-01-0145-FEDER-000017, in its line of research entitled BEST, which was co-financed by the European Regional Development Fund (ERDF) through NORTE 2020 (North Regional Operational Program 2014/2020). For the authors integrated in the CITAB research centre, the research was further financed by the FEDER/COMPETE/POCI Operational Competitiveness and Internationalization Programme, under Project POCI-01-0145-FEDER-006958, and by National Funds from FCT, the Portuguese Foundation for Science and Technology, under the project number UID/AGR/04033/2013. For the author who is integrated into the CQVR, the research was additionally supported by National Funds from the FCT, under the project number UID/QUI/00616/2013.

REFERENCES

[1] Binz, A.C., Patelb, V. & Wanzenried, G., A comparative study of CB-SEM and PLSSEM for theory development in family firm research. *J. Fam. Bus. Strat.,* **5**(1), pp. 116–128, 2014.

[2] Hair, J.F., Hult, G.T.M., Ringle, C. & Sarstedt, M., *A Primer on Partial Least Squares Structural Equation Modeling,* Sage Publications Inc., 2014. https://doi.org/10.1016/j.lrp.2013.01.002.

[3] Garson, G.D., *Partial Least Squares: Regression and Structural Equation Models*, 2016. www.statisticalassociates.com/pls-sem.htm.

[4] Kumar, G., Tuluri, A.F. & Tchounwou, P.B., Development of PLS: path model for understanding the role of precursors on ground level ozone concentration in Gulfport, Mississippi, USA. Atmos. *Pollut. Res.,* **6**(3), pp. 389–397, 2015.

[5] Zou, S. & Yu, Y.-S., A general structural equation model for river water quality data. *J. Hydrol.,* **162**(1), pp. 197–209, 1994.

[6] Chenini, I. & Khemiri, S., Evaluation of ground water quality using multiple linear regression and structural equation modeling. *Int. J. Environ. Sci. Technol.,* **6**(3), pp. 509–519, 2009

[7] Wu, E.M.-Y., Tsai, C.C., Cheng, J.F., Kuo, S.L. & Lu W.T., The application of water quality monitoring data in a reservoir watershed using AMOS Confirmatory Factor Analyses, 2014.

[8] Levêque, J.G. & Burns, R.C., A structural equation modeling approach to water quality perceptions. *J. Environ. Manag.,* **197**, pp. 440–447, 2017.

[9] Nugroho, A., Structural Equation Modelling as an instrument for water pollution factor analysis: Study case on the Surabaya River. *The Fourth International Conference on Advances in Applied Science and Environmental Technology (ASET)*, Institute of Research Engineers and Doctors: USA, pp. 28–32, 2016.

[10] INAG, Critérios para a classificação do estado dasmassas de água superficiais - rios e albufeiras. Technical Report. [Criteria for classifying the condition of the volumes of surface waters in rivers] Ministério do Ambiente, do Ordenamento do Território e do Desenvolvimento Regional [Ministry of the environment, care of the territory and regional development], Instituto da Água [Water Institute], IP in Portuguese, 2009. www.apambiente.pt.

[11] ESRI, ArcHydro Tools for ArcGIS 10 – Tutorial, 2012.

[12] Pacheco, F.A.L., Varandas, S.G.P., Sanches Fernandes, L.F. & Valle Junior, R.F., Soil losses in rural watersheds with environmental land use conflicts. *Sci. Total Environ*, pp. 110–120, pp. 485–486, 2014.

[13] Monecke, A. & Leisch, F., SemPLS: Structural Equation Modeling using partial least squares. *J. Stat. Softw.,* **48**(3), pp. 1–32, 2012.

[14] Ring, C.M., Wende, S. & Will, A., Smart PLS, 2005. www.smartpls.de.