UNIVERSITEIT GENT

FACULTEIT BIO-INGENIEURSWETENSCHAPPEN
FACULTY OF BIOSCIENCE ENGINEERING

# Macroinvertebrate based mathematical models for the prediction of microbial pathogen in rivers


Rubén Jerves-Cobo, I. Nopens, P. Goethals

Ruben.JervesCobo@UGent.Be

RESEARCH GROUP AQUATIC ECOLOGY

Laboratory of Environmental Toxicology and Aquatic Ecology



BIG DATA AND DATA MINING 2016, London, England, Date (26/09/2016)


Laboratory of Environmental Toxicology and Aquatic Ecology, Research Group Aquatic Ecology (AECO)
Ruben.JervesCobo@UGent.be

# OUTLINE

- Background and objective
- Materials and methods
- Results
- Conclusion


Laboratory of Environmental Toxicology and Aquatic Ecology, Research Group Aquatic Ecology (AECO)
Ruben.JervesCobo@UGent.be

# 1-BACKGROUND AND OBJECTIVE (I)

- The quality of the water must accomplish standards.
  - drinking water, recreational purpose, Irrigation
- The indicators used to verify microbial contamination of water are:
  - total coliforms and fecal coliforms and/or Escherichia coli


Laboratory of Environmental Toxicology and Aquatic Ecology, Research Group Aquatic Ecology (AECO)
Ruben.JervesCobo@UGent.be

# 1-BACKGROUND AND OBJECTIVE (II)

- Checking the fulfillment requires expensive and highly trained personnel in laboratories
- Biota works as a permanent monitor of water
- biological samples can :
  - reflect an increase in pollution.
  - predict average values of chemical parameters


Laboratory of Environmental Toxicology and Aquatic Ecology, Research Group Aquatic Ecology (AECO)
Ruben.JervesCobo@UGent.be

# Objective

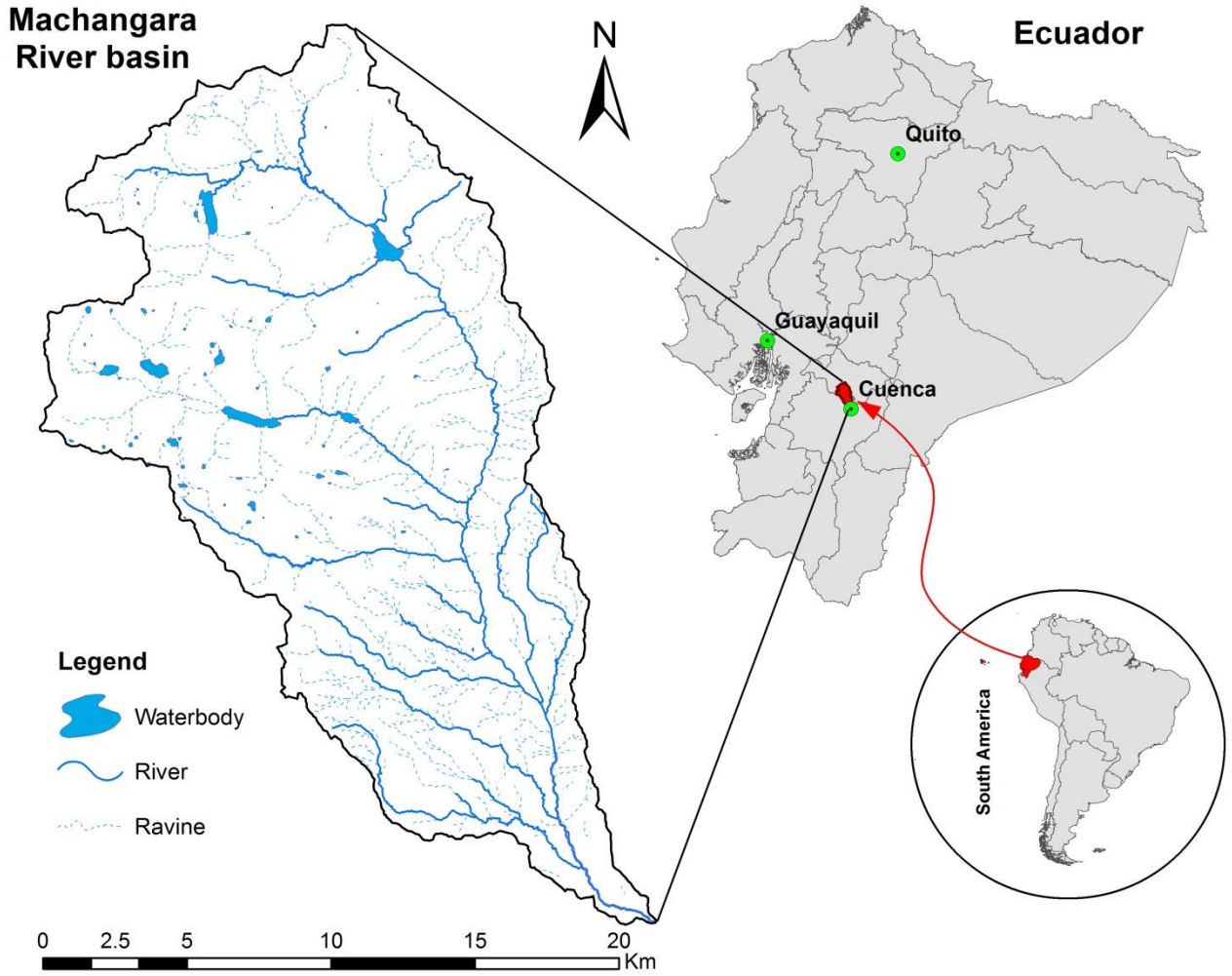
- Introduce a quick way of checking the fulfillment of fecal coliforms standards using macrobenthos.

- Analyzes the requirement to include biology and hydro-morphology aspects in Ecuadorian regulations to assess river ecosystem health.

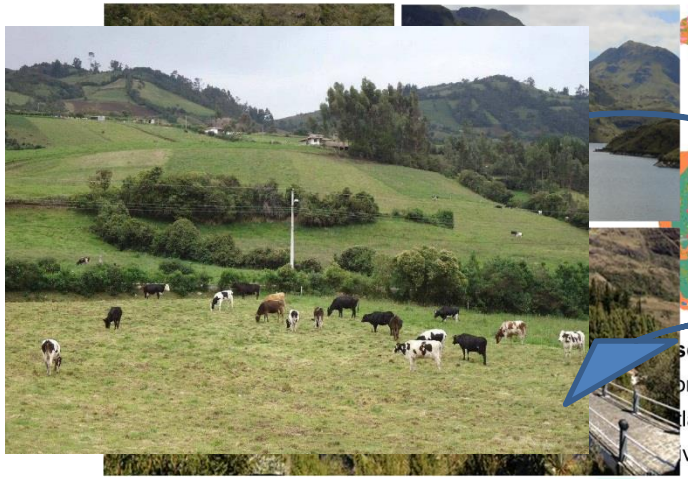

Laboratory of Environmental Toxicology and Aquatic Ecology, Research Group Aquatic Ecology (AECO)
Ruben.JervesCobo@UGent.be

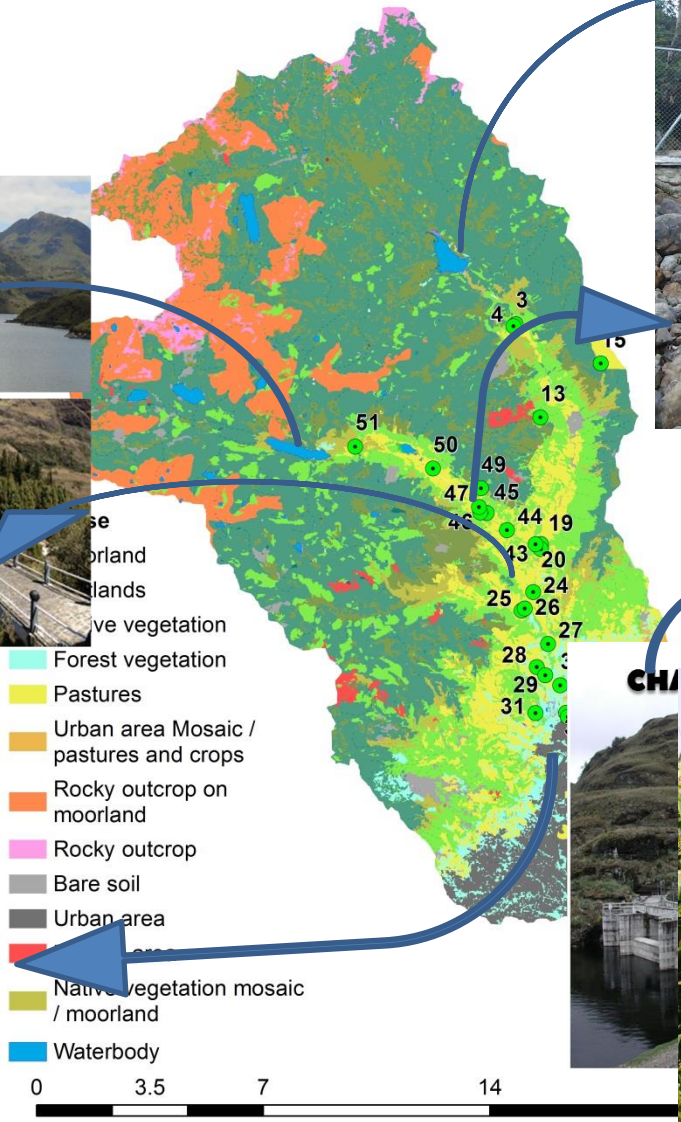
# 2- MATERIALS AND METHODS

## LOCATION



# LAND USE






UNIVERSITEIT GENT

RESEARCH GROUP AQUATIC ECOLOGY

## DATA COLLECTION:

- Completed information on 33 locations

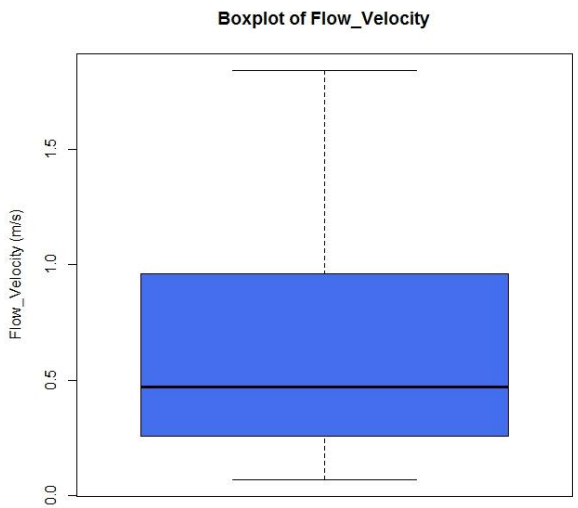
## Physicochemical, hydraulic, microbiological

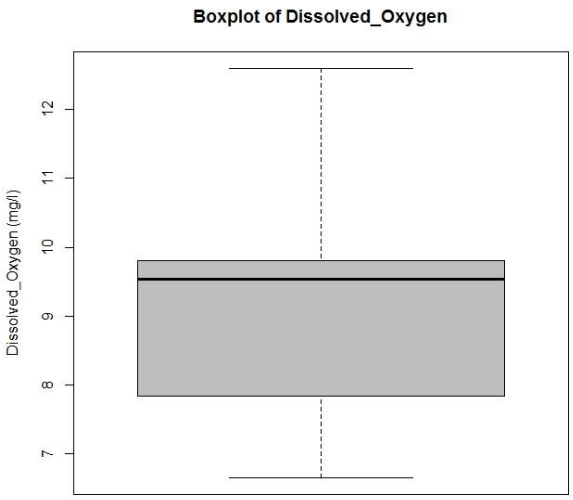
- **Laboratory**
  - $BOD_5$, COD, Nitrate + Nitrite, Ammonia, Organic Nitrogen, Phosphates, Total Phosphorus, Fecal and Total Coliforms, Real Color, Turbidity, Total Solids
- **Field:** Flow Velocity, Ph, Conductivity, Temperature, Dissolved Oxygen

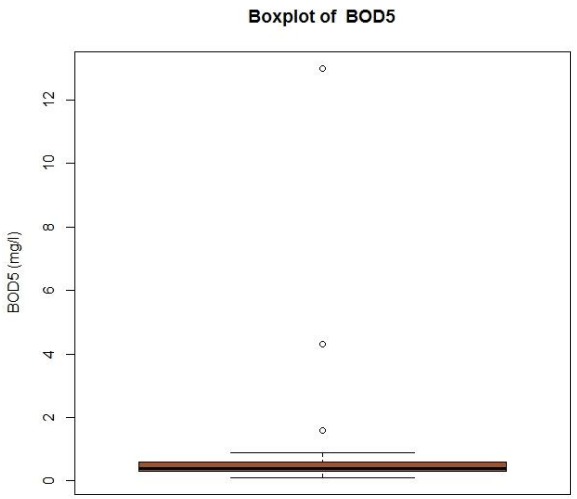
## Macrobenthos

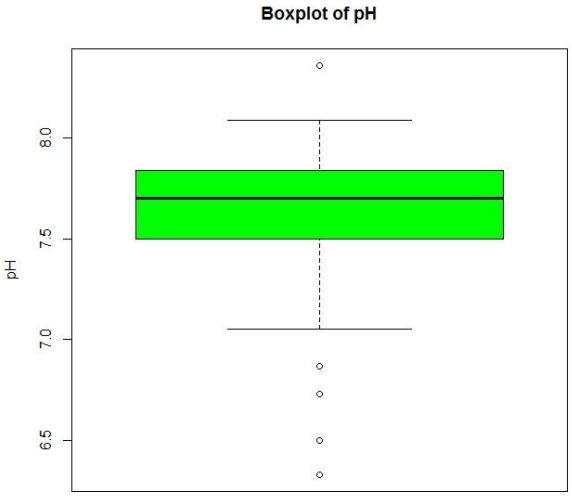
39 families (taxa) found

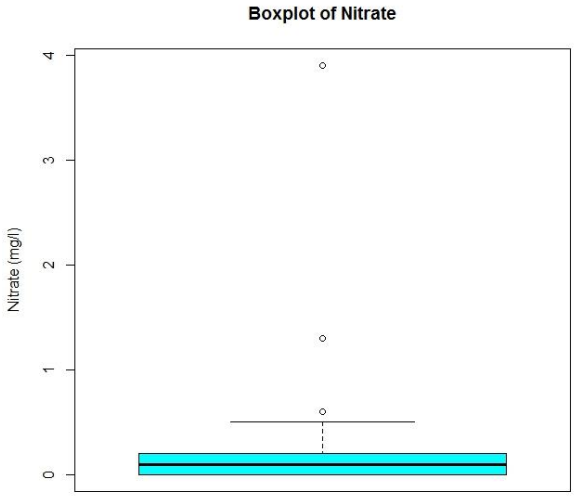
# Variables Variation

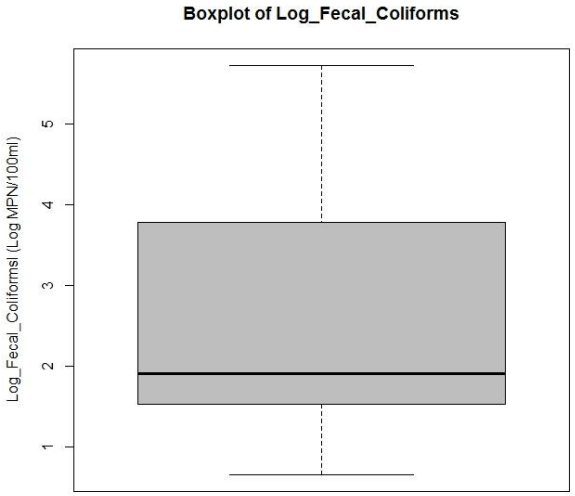


# - Ecuadorian Water Quality Regulation for Fecal Coliforms

| Regulations | Water used to | Fecal Coliforms Limited Value MPN/100 ml |
|---|---|---|
| First | Recreational with primary contact | ≤200 |
| Second | Agriculture and Livestock | ≤1,000 |
| Third | raw water previous to non-conventional treatment* | ≤2,000 |
| * Conventional treatment refers to chemical addition, rapid mixing, flocculation and sedimentation | | |


Laboratory of Environmental Toxicology and Aquatic Ecology, Research Group Aquatic Ecology (AECO)
Ruben.JervesCobo@UGent.be

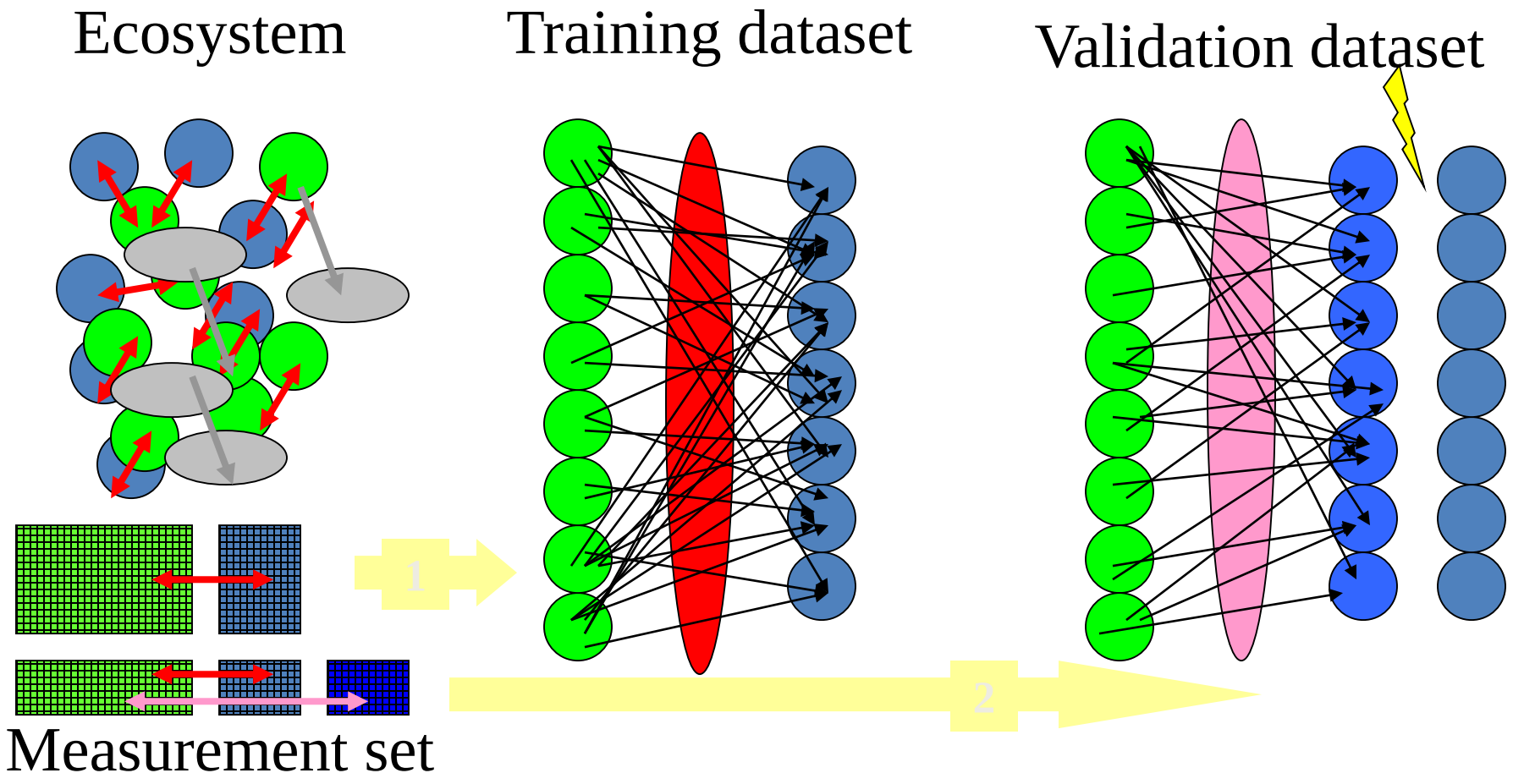
# ECOLOGICAL WATER QUALITY

- Biological Monitoring Working Party Index - Col
- BMWP-Col = f(Sensitivity of Macrobenthos)
- Sensitivity -> 1-10 (Low – High Sensitivity)

| Class | Quality | BMWP | Color |
|---|---|---|---|
| I | Very Good | > 100 | |
| II | Good | 61 - 100 | |
| III | Moderate | 36 - 60 | |
| IV | Deficient | 16 - 35 | |
| V | Bad | ≤ 15 | |

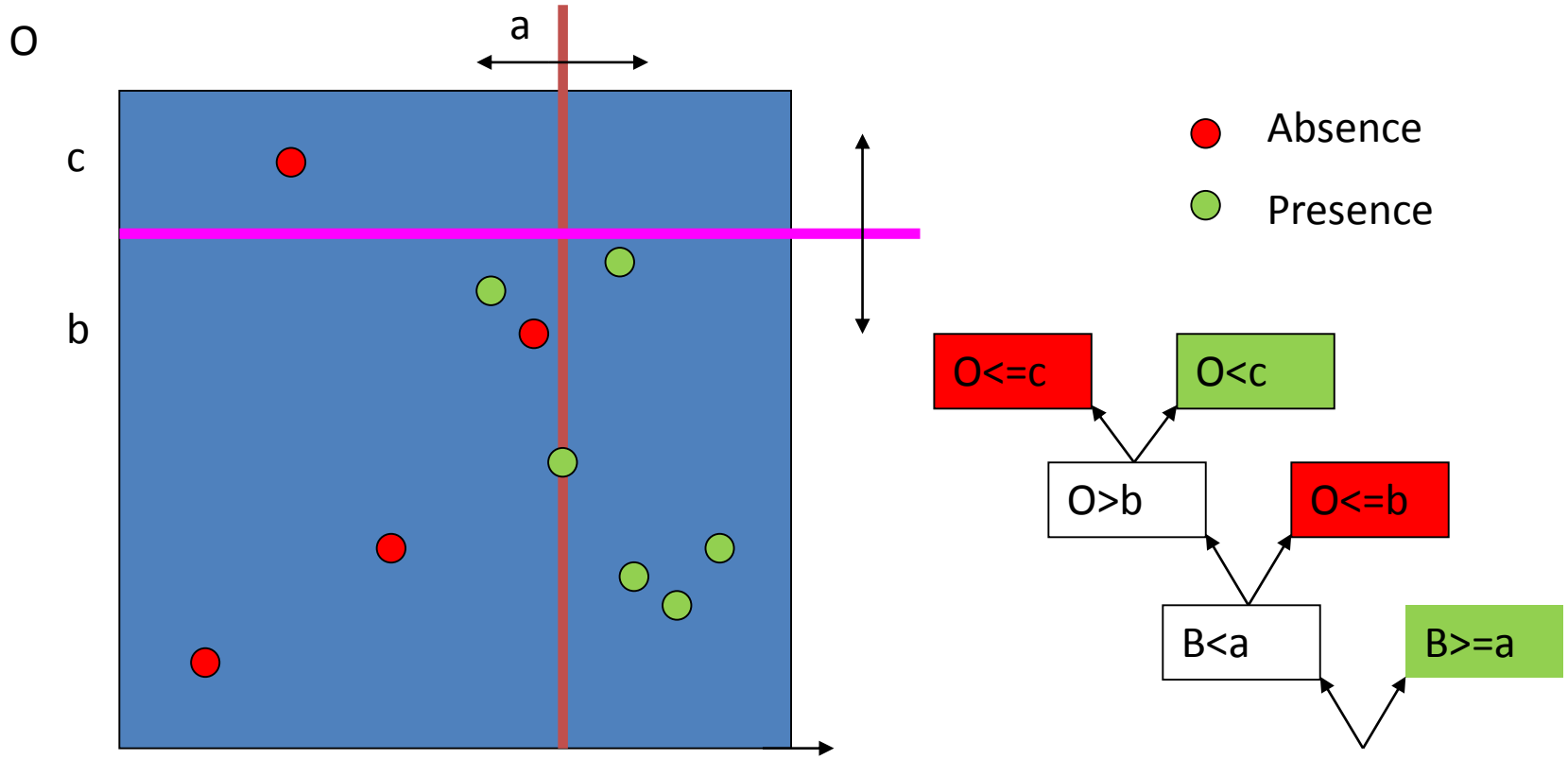

Laboratory of Environmental Toxicology and Aquatic Ecology, Research Group Aquatic Ecology (AECO)
Ruben.JervesCobo@UGent.be

# Model development



# Methodology
# Pruned Multi-target Clustering Trees (PMCT)

- Classification trees: searching for if-then rules (threshold values): 100% reliable and ‘safe’ models



## Model Performance

- Models must be evaluated based on statistical and ecological criteria.
- Models must be as clear and simple as possible.

## Settings

- Machine learning software: Waikato Environment for Knowledge Analysis (WEKA)
- Three, five, ten-fold cross validation (k fcv)
- Pruning process
  - Pruning confidence factors (PCF): 0.25, 0.10

# Model Performance

- Confusion matrix from Decision Tree Models:

| | | Predicted Class | |
|---|---|---|---|
| | | Yes | No |
| Observed Class | Yes | TP | FN |
| | No | FP | TN |

- Correctly Classified Instances (CCI):> 0.70
- Cohen´s Kappa Statistic: > 0.40
- Lowest value for the false negative (FN) will increase sensitivity


Laboratory of Environmental Toxicology and Aquatic Ecology, Research Group Aquatic Ecology (AECO)
Ruben.JervesCobo@UGent.be

# Model Optimization:

- *Cost sensitive classifier (CSC)*: gives new weights in training instances FN, FP
- *Overall confusion entropy of a confusion matrix (CEN):* evaluates the confusion level of the class distribution of misclassified samples:

$$CEN = (P_1 + P_2)CEN_j$$

$$P_1 = \frac{TP + FN}{2(TP + FN + FP + TN)} \; and \; P_2 = \frac{FP + TN}{2(TP + FN + FP + TN)}$$

$$CEN_j = -P_{FN} log_2 P_{FN} - P_{FP} log_2 P_{FP}$$


Laboratory of Environmental Toxicology and Aquatic Ecology, Research Group Aquatic Ecology (AECO)
Ruben.JervesCobo@UGent.be

# Model Optimization:

- $P_j$: confusion probability of class $j$
- $CEN_j$: confusion entropy of class $j$.

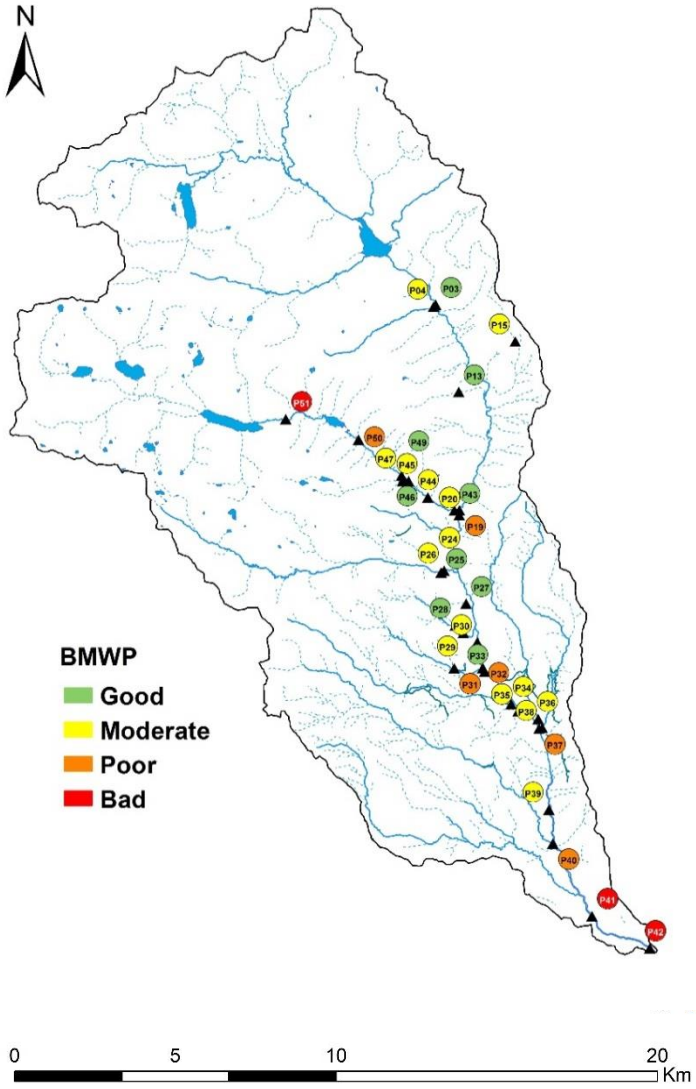
$$P_{FP} = \frac{FP}{FN + FP + 2TP} \; and \; P_{FN} = \frac{FN}{FN + FP + 2TN}$$

- $P_{FP}$ and $P_{FN}$ are the misclassification probability of classifying the samples of class $i$ to class $j$ subject to class $j$

- Higher accuracy corresponds to lower confusion entropy


Laboratory of Environmental Toxicology and Aquatic Ecology, Research Group Aquatic Ecology (AECO)
Ruben.JervesCobo@UGent.be

# 3- RESULTS




## ECOLOGICAL WATER QUALITY

BMWP-Col:

- 9 good
- 15 moderate
- 6 poor
- 3 bad

UNIVERSITEIT GENT

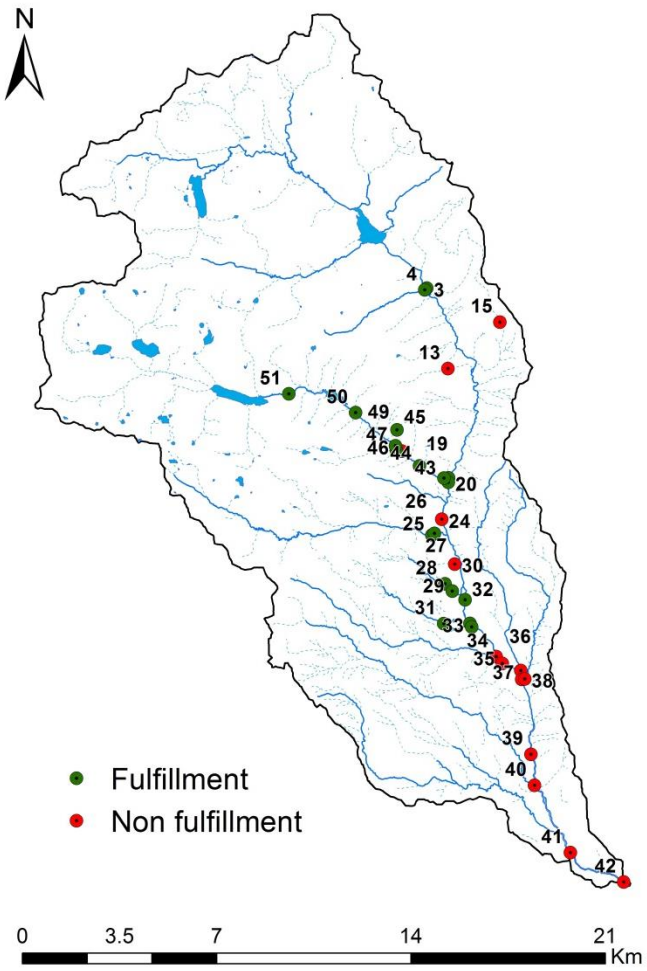
RESEARCH GROUP AQUATIC ECOLOGY

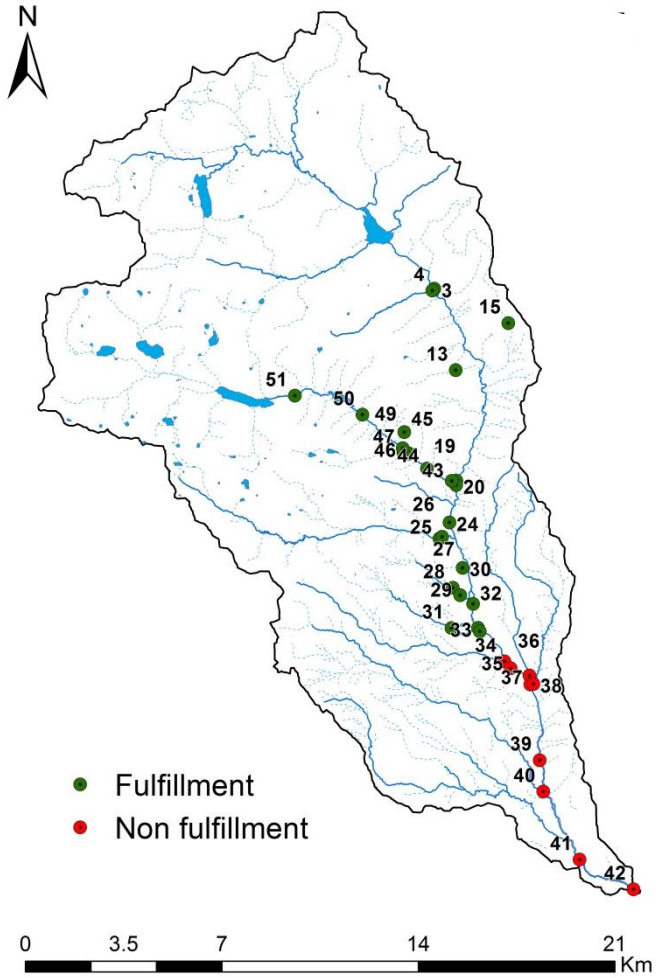
# Analysis of Ecuadorian Water Quality Regulation for Fecal Coliforms in relation to BMWP-Col

| BMWP-Col | First Preservation Regulation | Second Preservation Regulation | Third Preservation Regulation |
|---|---|---|---|
| Bad | 2 | 2 | 2 |
| Deficient | 2 | 2 | 2 |
| Moderate | 8 | 5 | 5 |
| Good or very good | 2 | - | - |
| TOTAL non fulfillment | 14 | 9 | 9 |

BIG DATA AND DATA MINING 2016, London, England, Date (26/09/2016)
Laboratory of Environmental Toxicology and Aquatic Ecology, Research Group Aquatic Ecology (AECO)
Ruben.JervesCobo@UGent.be

# Fulfillment of Fecal Coliforms limits in relation to water use



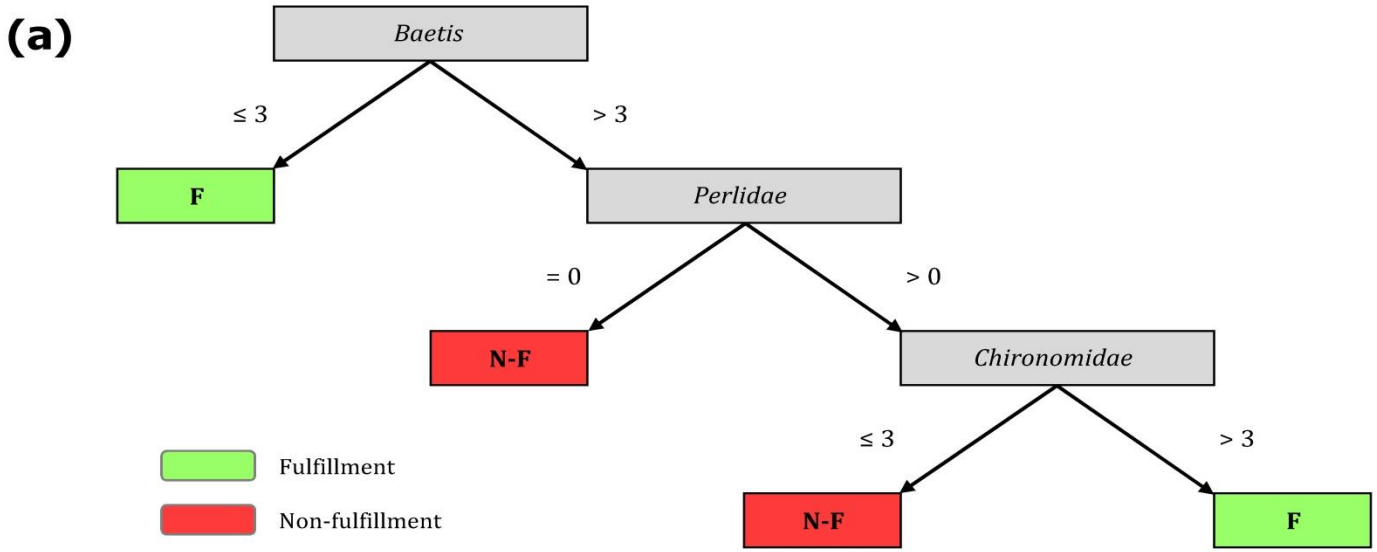




(a) Recreational with primary contact

(b) Agricultural - Livestock use and (c) raw water

# ***Model Performance (1)***

***First Model:*** *Primary Contact – Fecal Coliforms Regulation.*




CCI = 73%

Kappa = 0.46

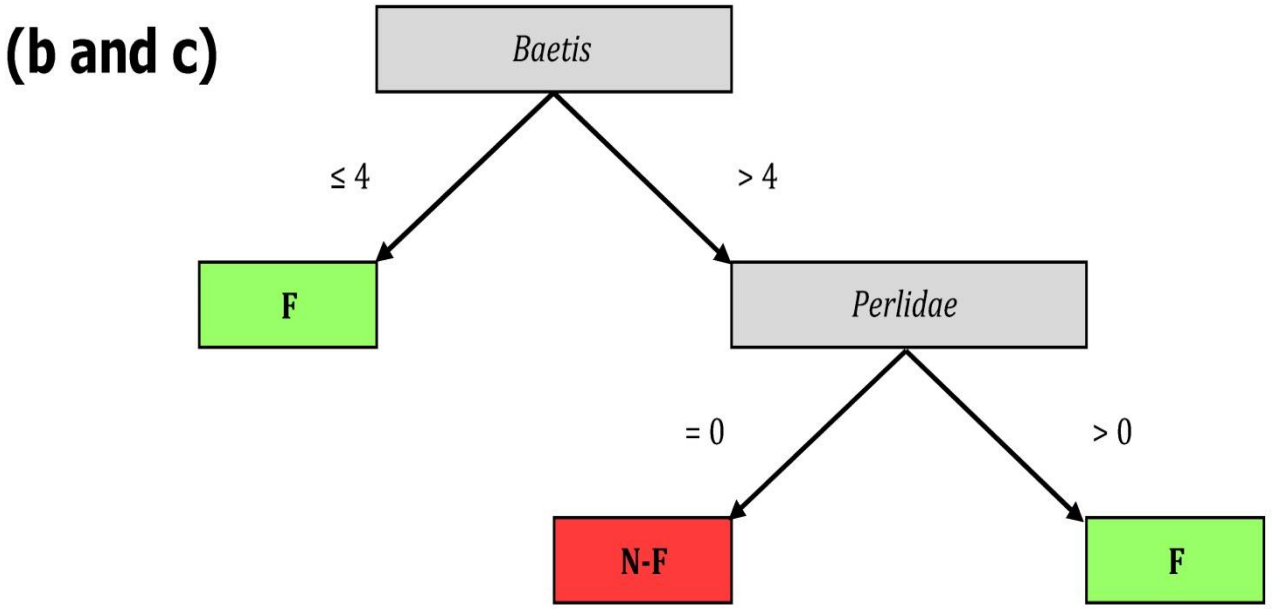
FN = 2

CEN = 0.762

*Chironomidae* families include species with large differences in tolerance to pollutants

## *Model Performance (2)*

***Second model:*** *Agriculture - Livestock Irrigation and raw water previous to non-conventional treatment*




CCI = 94%

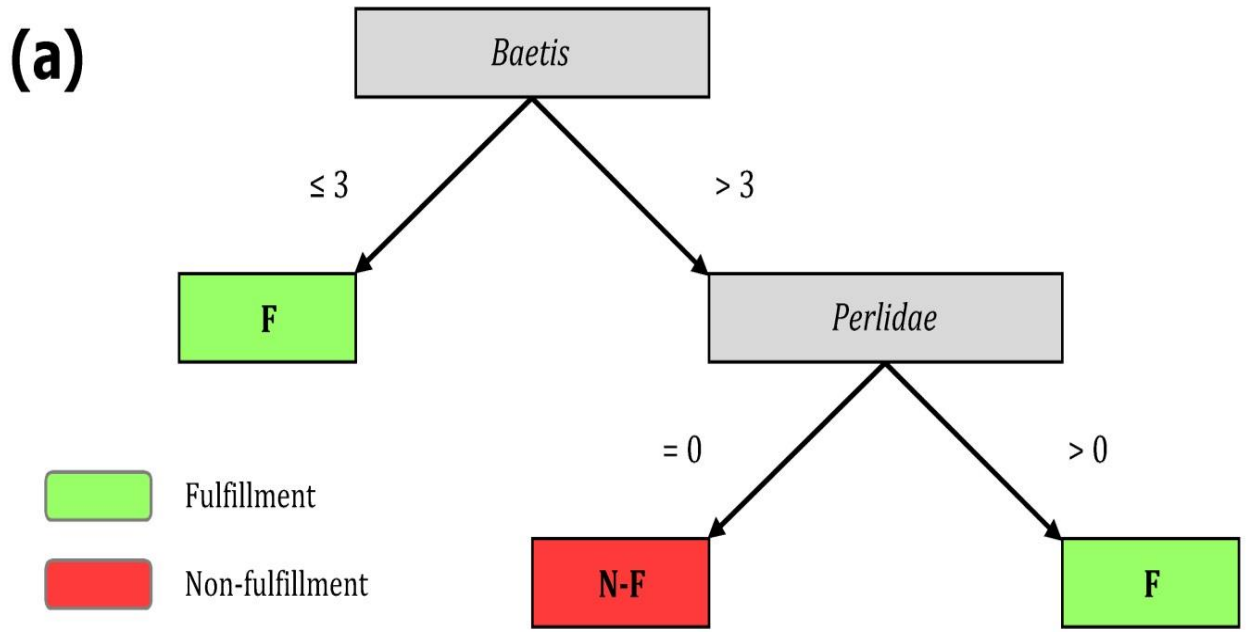
Kappa = 0.85

FN = 1

CEN = 0.348

## ***Model Optimization (2)***

***First Model:*** *Primary Contact – Fecal Coliforms Regulation.*




CCI = 91%

Kappa = 0.81

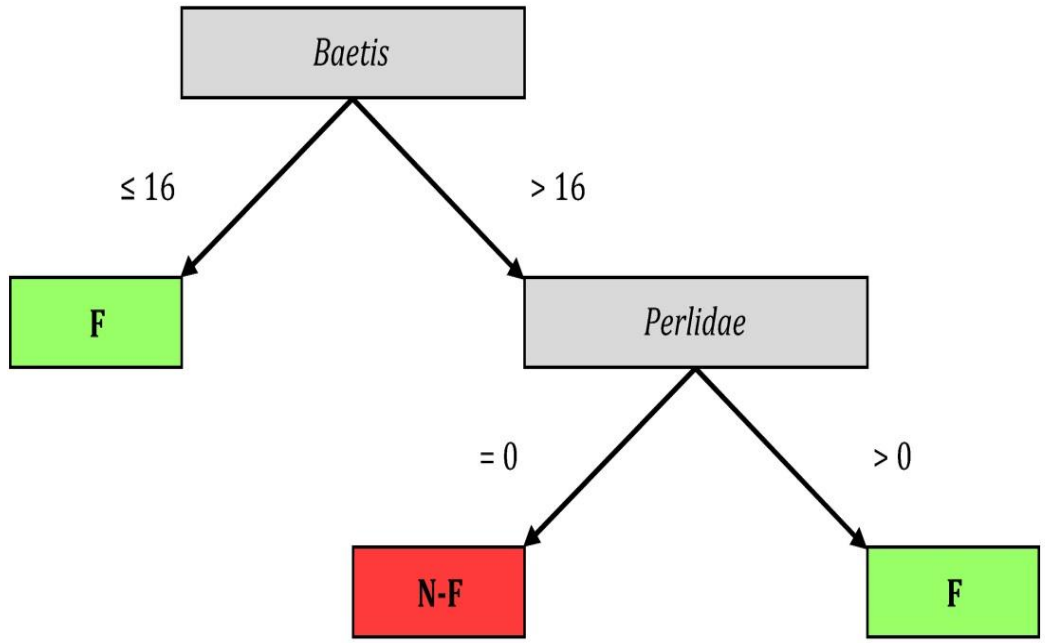
FN = 0

CEN = 0.285

# *Model Optimization (3)*

***Second model:*** *Agriculture - Livestock Irrigation and raw water previous to non-conventional treatment*

**(b and c)**




CCI = 100%

Kappa = 1.0

FN = 0

CEN = 0

# 4- CONCLUSION

- Four models were selected.
  - Two Models: fecal coliforms threshold in recreational with primary contact water use.
  - Two Models: fecal coliforms limits in agricultural - livestock water use, or raw water for drinking water treated with non-conventional processes.

- The cost-sensitive classifier (CSC) in the Weka can reduce false positives (FP) in the confusion matrix, improved the reliability of the resulting models.
- Confusion entropy of a confusion matrix (CEN) was lower when the confusion matrix had lower FN values.

# *Acknowledgement*

- VLIR-UOS IUC Programme - University of Cuenca
- VLIR Ecuador Biodiversity Network Project
- Council of the Machangara River Basin


Laboratory of Environmental Toxicology and Aquatic Ecology, Research Group Aquatic Ecology (AECO)
Ruben.JervesCobo@UGent.be

UNIVERSITEIT GENT

FACULTEIT BIO-INGENIEURSWETENSCHAPPEN
FACULTY OF BIOSCIENCE ENGINEERING

# Thank you

?


Laboratory of Environmental Toxicology and Aquatic Ecology, Research Group Aquatic Ecology (AECO)
Ruben.JervesCobo@UGent.be

# Bibliography:

- D'Heygere, T., P. L. M. Goethals, and N. De Pauw. 2003. Use of Genetic Algorithms to Select Input Variables in Decision Tree Models for the Prediction of Benthic Macroinvertebrates. Ecological Modelling 160:291-300
- Maimon, O., and L. Rokach. 2005. Data Mining and Knowledge Discovery Handbook. Springer.
- Roldán Pérez, G. 1999. Los Macroinvertebrados y Su Valor Como Indicadores de la Calidad del Agua. Academia Colombiana de Ciencia 23:375-387.
- Ting, K. M. 2002. An instance-weighting method to induce cost-sensitive trees. Ieee Transactions on Knowledge and Data Engineering **14**:659-665.
- Wei, J.-M., X.-J. Yuan, Q.-H. Hu, and S.-Q. Wang. 2010. A novel measure for evaluating classifiers. Expert Systems with Applications **37**:3799-3809


Laboratory of Environmental Toxicology and Aquatic Ecology, Research Group Aquatic Ecology (AECO)
Ruben.JervesCobo@UGent.be