BigData IoT in environmental modelling

Andreja Naumoski and Miroslav Mirchev

E-mail: andreja.naumoski@finki.ukim.mk

Faculty of Computer Science and Engineering – Ss. Cyril and Methodius University in Skopje, Republic of Macedonia

Data-driven environmental modelling

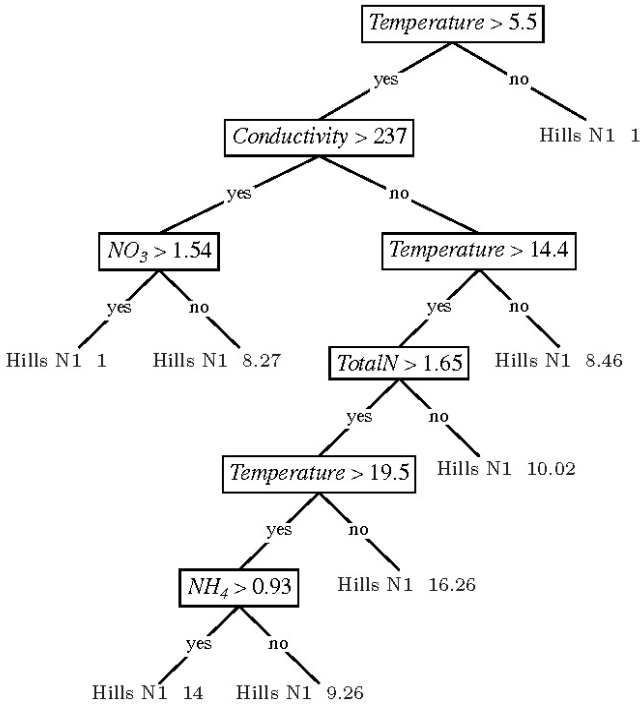
- Laboratory of Eco-Informatics at FINKI, part of it does, focus on applying computer science method, namely, machine learning algorithms to solve environmental problems with the power of data computer analytics.

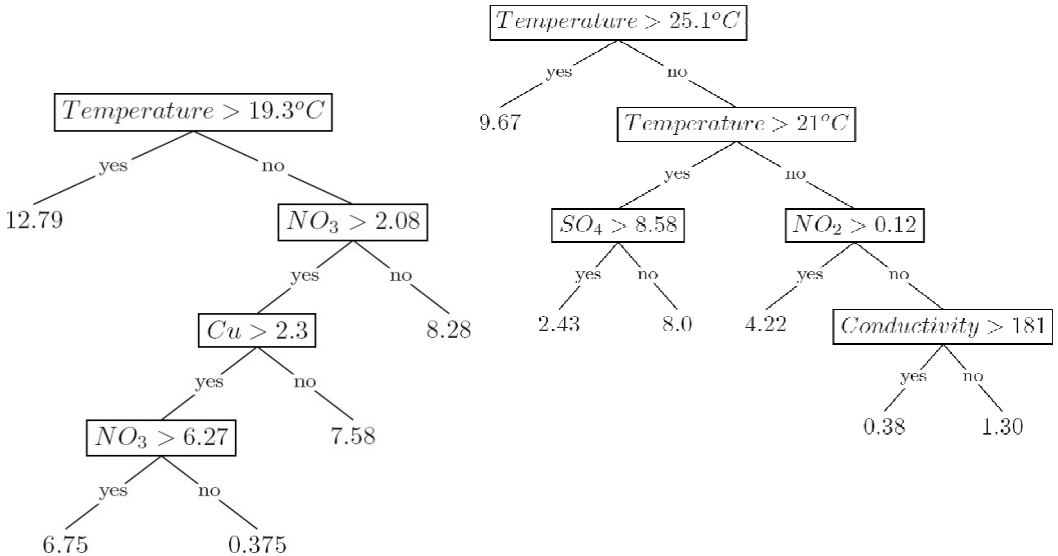
- Models are data-driven based on data from various sources.
  - Habitat suitability modelling of organisms in water ecosystems
  - Bio-diversity modeling.
  - Early warning flood prediction.
  - Ambient air pollution with dispersion modelling.
  - Car traffic pollution modelling.
  - Predicting the trend and factors in air pollution gases.
  - Noise pollution in cities and etc.

- Modelling only with data have to always taken with caution, because the world is presented to the method only through the data point of view.

Data-driven environmental modelling

- Most of the environmental modelling is done by using supervised (white box) machine leaning algorithms that are easy interpretable.
- Example of data driven models based on single target classical decision tree for habitat suitability modelling.





* A. Naumoski, "Multi–target modelling of diatoms diversity indices in Lake Prespa," Applied Ecology and Environmental Research, 10(4), pp. 521–529,2012.

Single vs Multi target modelling

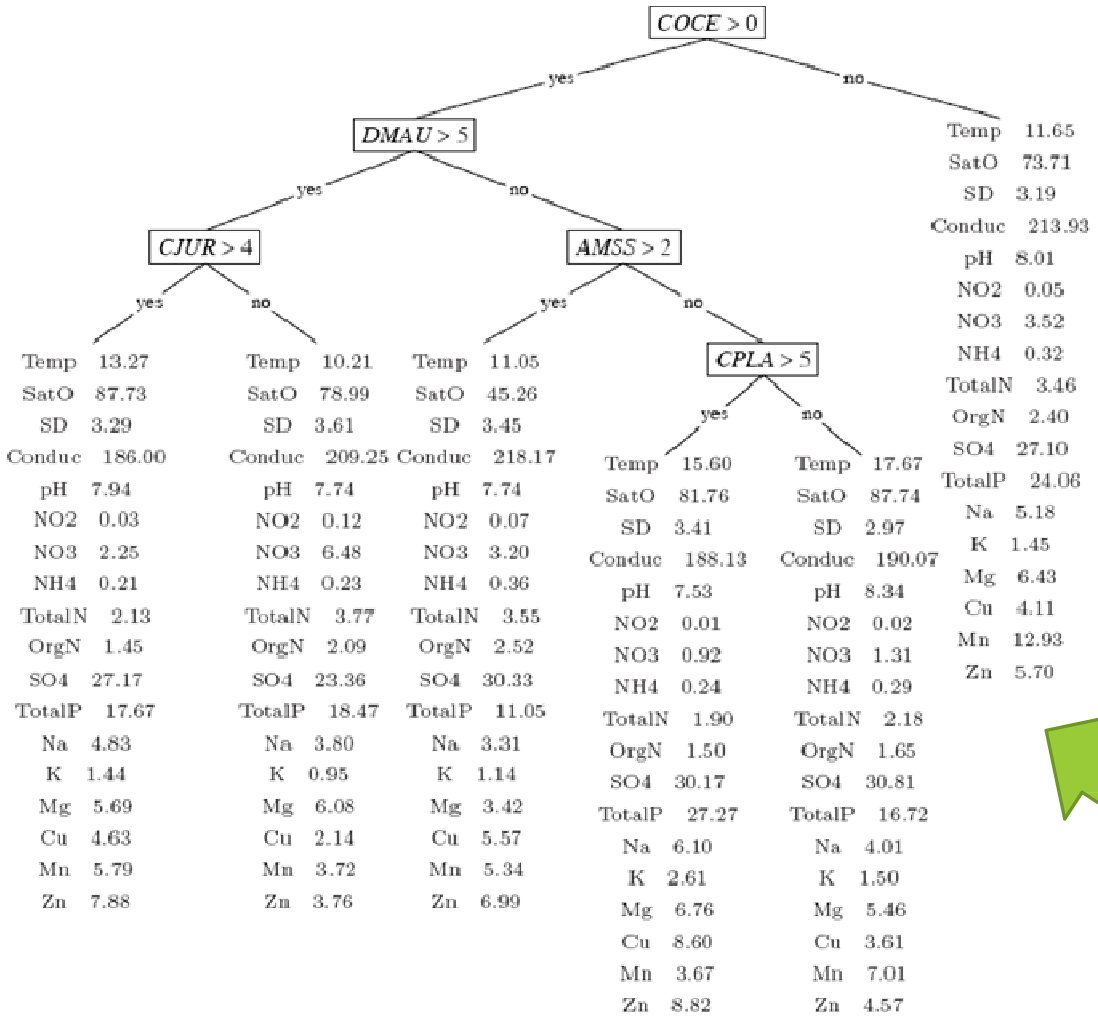
- Single target models -
  Model for each variable

								M1	M2		...		M6
A1	A2	A3	A4	A5	A6	A7	A8	P1	P2	P3	P4	P5	P6
12	3	6	5	8	9	245	5	0,5	6,5	12	2	C3	C1
42	12	2	4	65	12	6	15	1,2	51	5,9	1,1	C2	C1
3	6	5	12	5	12	2	5	22	65	5,1	25	C1	C2

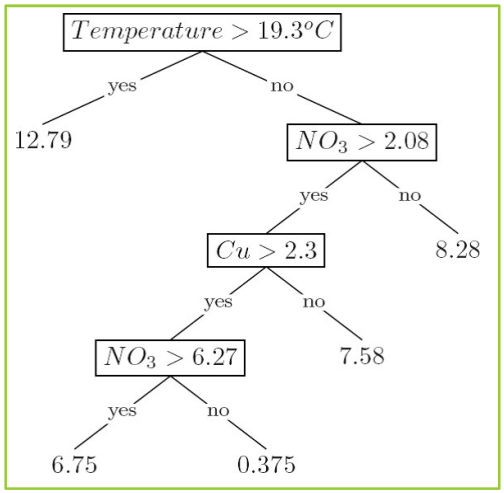
- Multi target models
  Several outputs in one model

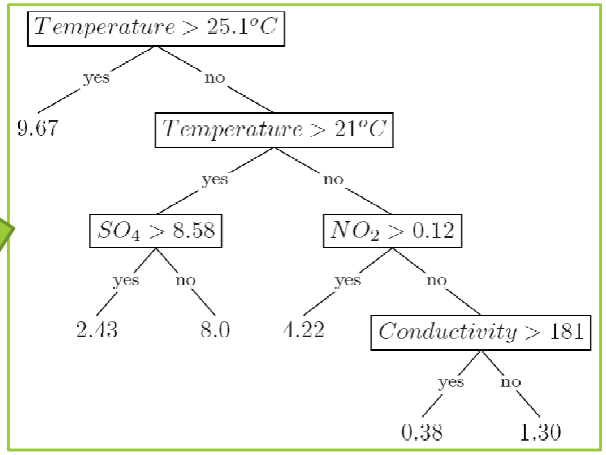
								M1					
A1	A2	A3	A4	A5	A6	A7	A8	P1	P2	P3	P4	P5	P6
12	3	6	5	8	9	245	5	0,5	6,5	12	2	C3	C1
42	12	2	4	65	12	6	15	1,2	51	5,9	1,1	C2	C1
3	6	5	12	5	12	2	5	22	65	5,1	25	C1	C2

Predicting environmental conditions



Multi--target regression tree for all physico--chemical parameters

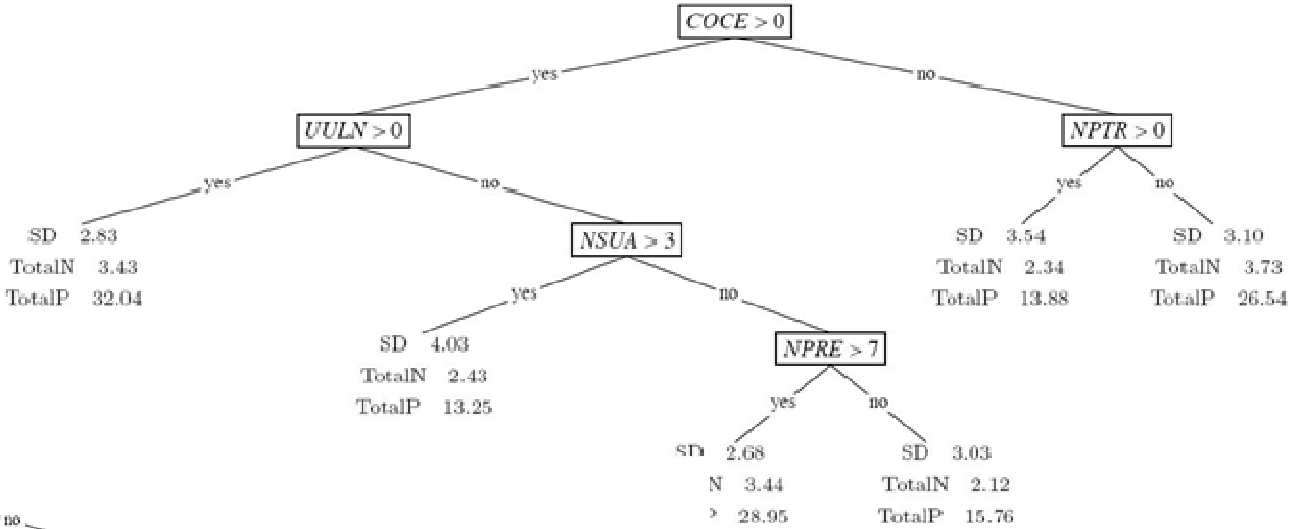


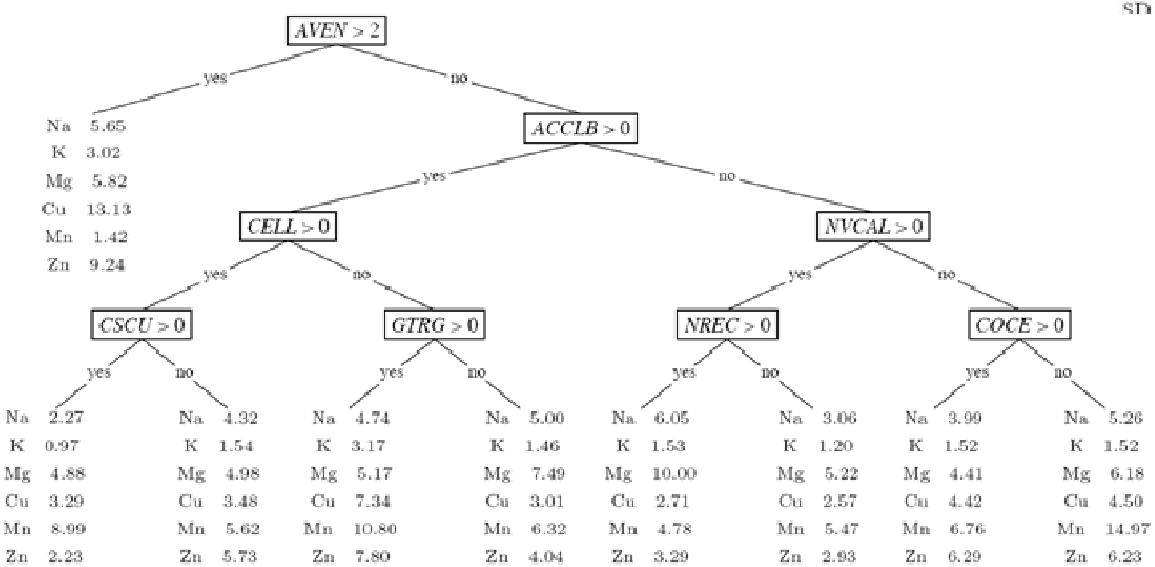


* These are unpublished results.

Predicting environmental conditions

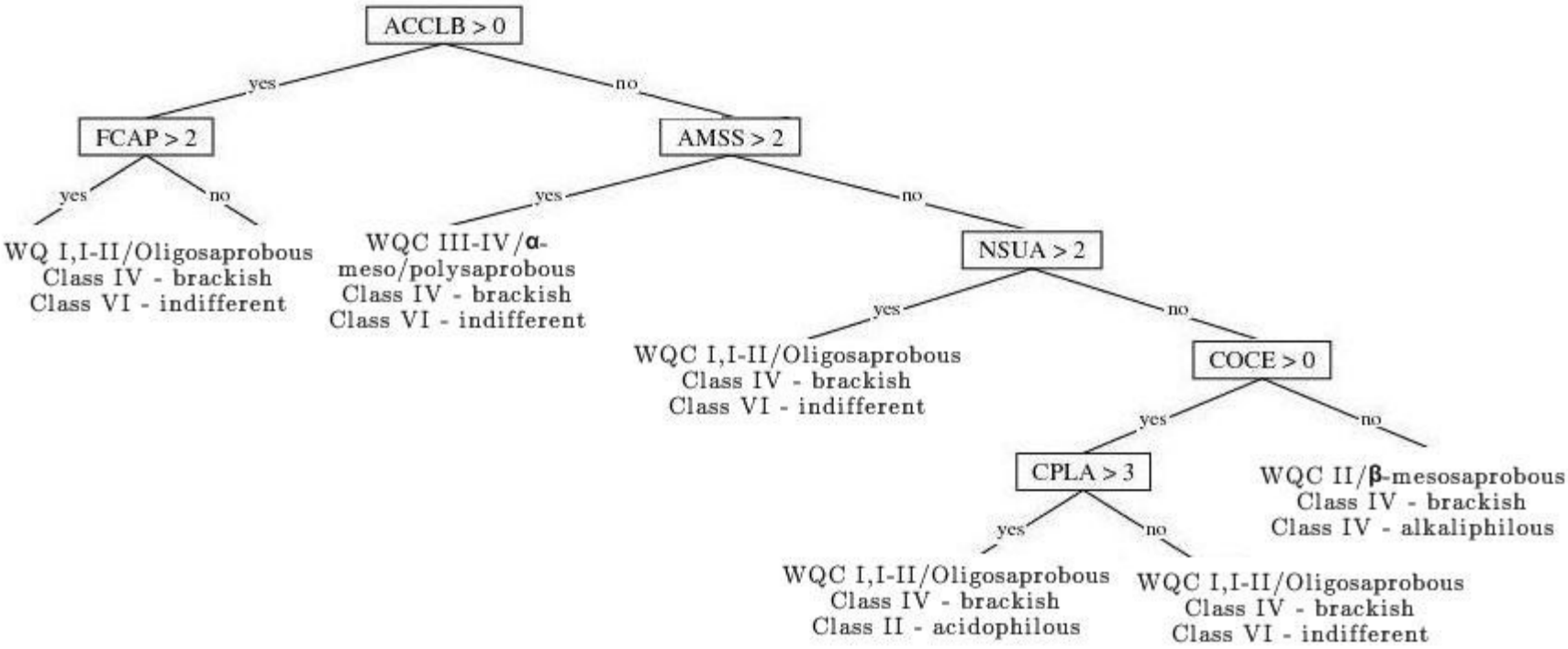
Multi–target regression tree for predicting trophic parameters





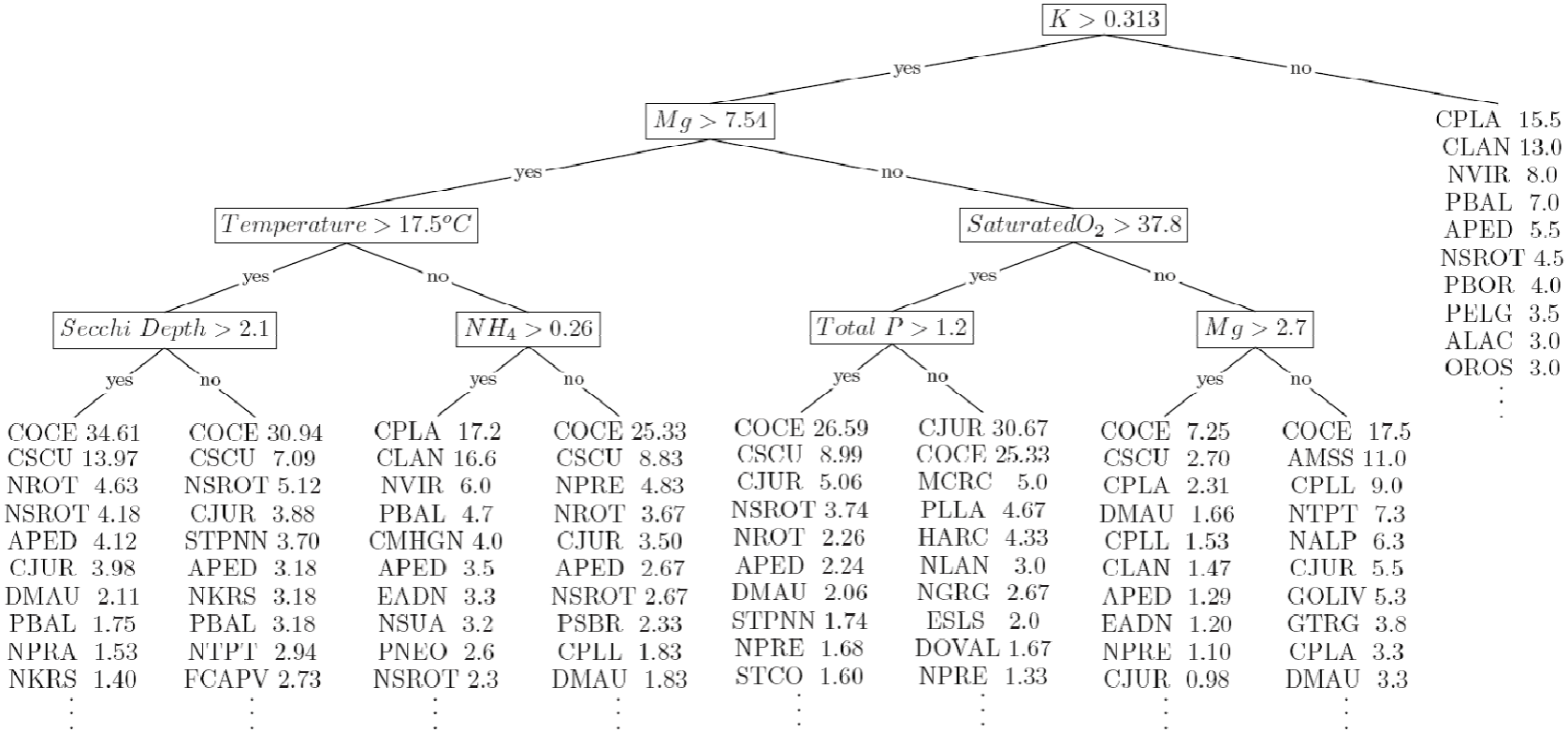
Multi–target regression tree for predicting concentration of metals

Predicting water quality using entire community



- A multi--target regression tree predicting the water quality class from the entire lake diatom community

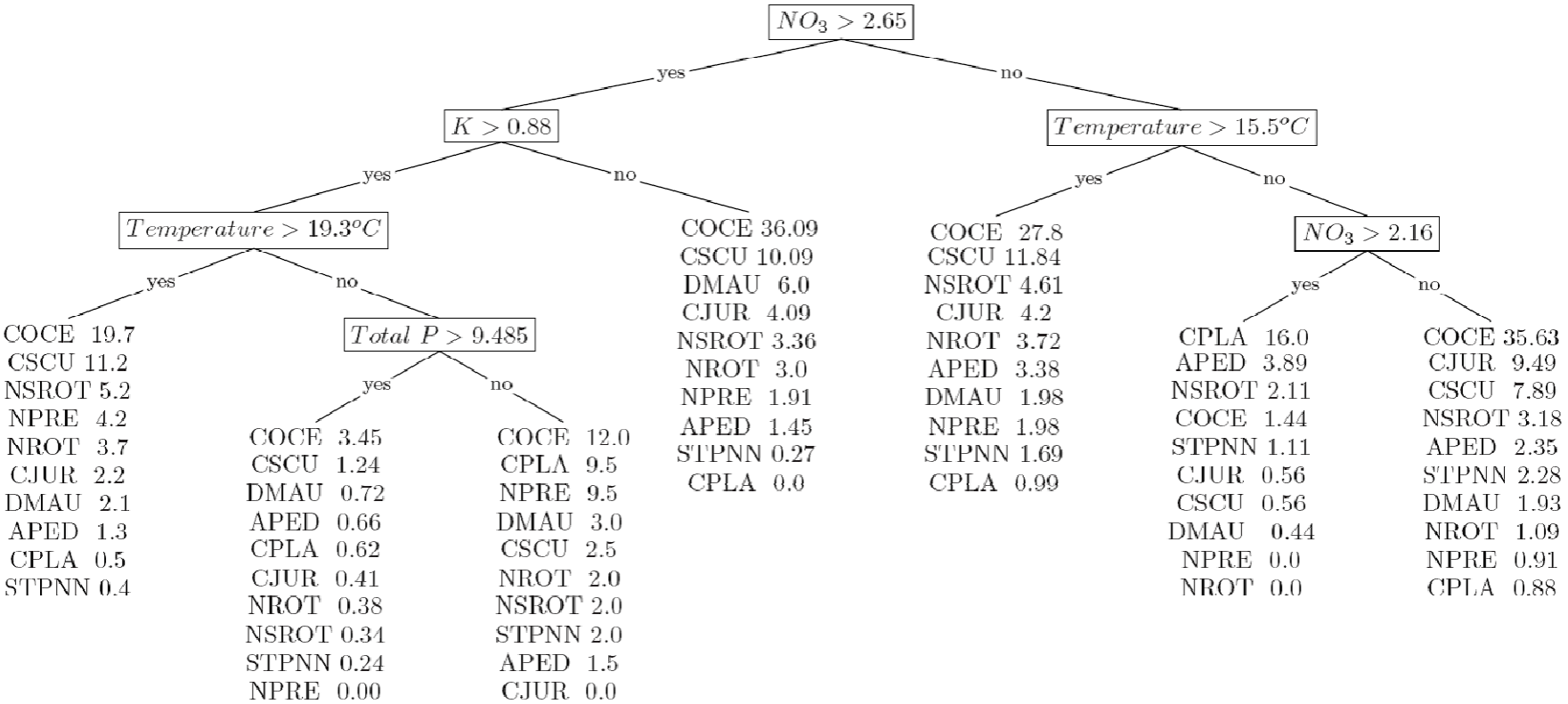
Predicting the entire community dynamic



- A multi–target regression tree predicting the structure of the entire lake diatom community

* D. Kocev, A. Naumoski, K. Mitreski, S. Krstic, and S. Dzeroski, "Learning habitat models for the diatom community in Lake Prespa, "Ecological Modelling, 221(2), pp. 330–337,2010.

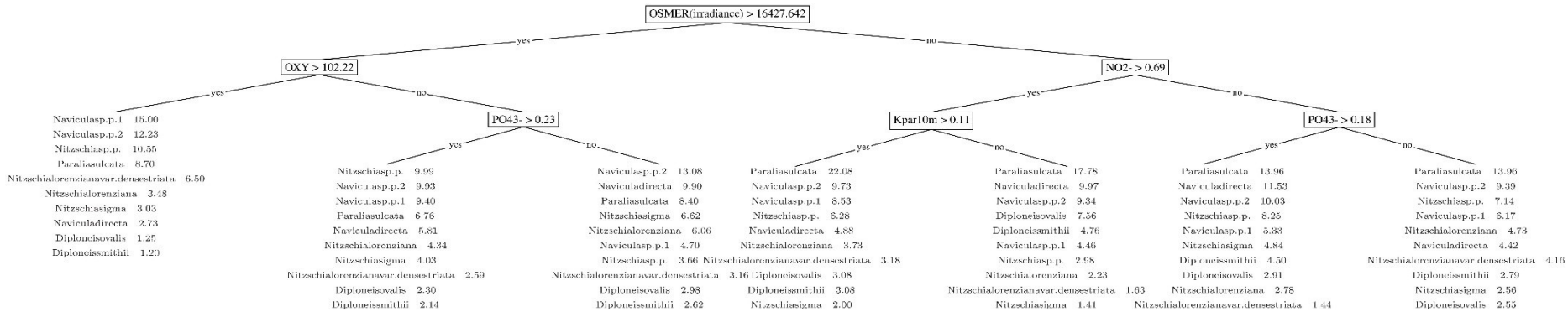
Predicting TOP10 diatom dynamic



- A multi–target regression tree predicting the relative abundance of the 10 most abundant lake diatoms

* D. Kocev, A. Naumoski, K. Mitreski, S. Krstic, and S. Dzeroski, "Learning habitat models for the diatom community in Lake Prespa, "Ecological Modelling, 221(2), pp. 330--337, 2010.

Predicting the entire community dynamic



- The most interesting from the model is the separation of two groups (clusters) of diatom species influenced by several different parameters. Actually, the OSMER irradiance is the most influencing factor on the abundance of the this diatom community according to the model, while Kpar10m and PO_4^{3-} are the less influence factors on the distribution of thediatoms.

- Furturemore, the *Paraliasulcata* species are mostly influenced by the NO_2, while the *Naviculasp.sp.1, Naviculasp.sp.2* and *Nitzschiasp* diatoms are influences by the Oxygen demand parameter.

- * These are unpublished results. (Data from Venice, Italy)

Rule induction models

Rule 1: IF NROT <= 19 AND CSCU > 19.0 AND CPLA <= 0
THEN [*oligosaprobous*] [196/16: 22/3] [*]

Rule 2: IF (Navicula krsticii (NKRS)) <= 0THEN

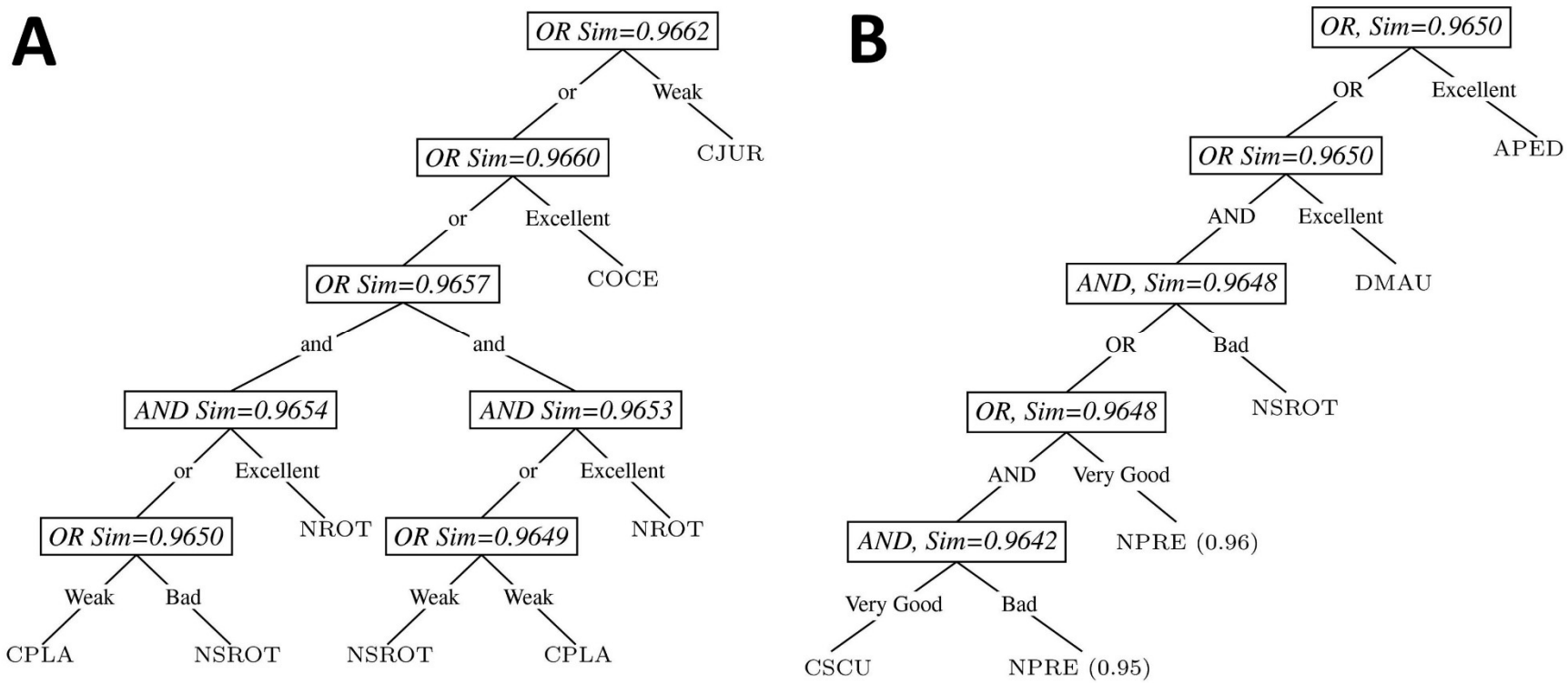
Temp	***SatO***	***SD***	***Conduc***	***pH***	***NO_2***	***NO_3***	***NH_4***	***TotalN***
11.6	75.2	2.42	212.1	6.5	0.7	4.5	2.5	8.5
OrgN	***SO_4***	***TotalP***	***Na***	***K***	***Mg***	***Cu***	***Mn***	***Zn***
4.56	21.31	28.9	7.49	2.5	8.5	7.2	8.5	4.7

Rule 3 :IF (Sellaphora pupula (SPUP) > 4 AND Cyclotella juriljii nom. nud.(CJUR) <= 2) THEN

Temp	***SatO***	***SD***	***Conduc***	***pH***	***NO_2***	***NO_3***	***NH_4***	***TotalN***
9.22	68.94	3.132	223.03	7.86	0.01	33.89	0.3	3.87
OrgN	***SO_4***	***TotalP***	***Na***	***K***	***Mg***	***Cu***	***Mn***	***Zn***
2.74	30.31	28.64	5.48	1.56	6.29	4.59	16.59	6.46

* A. Naumoski, and K Mitreski, "Diatom indicating property discovery with rule induction algorithm," In ICT Innovations 2012, Advances in Intelligent Systems and Computing, vol. 207, Markovski, Smile; Gusev, Marjan (Eds.).Springer–Verlag Berlin Heidelberg, pp.225–235, 2013.

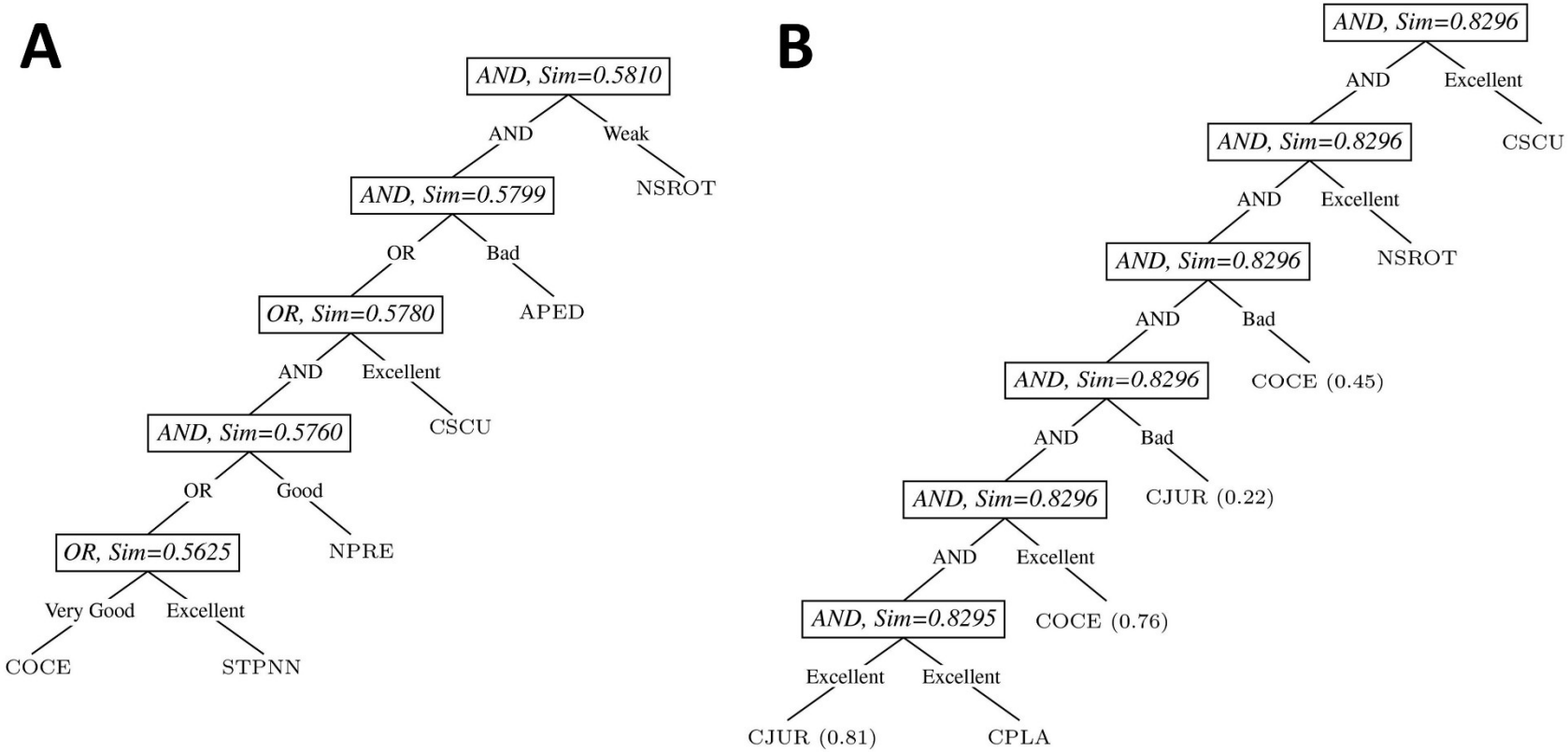
Fuzzy based decision tree models



- (A) PT model for the *mesotrophic* class based on TSI_TP
- (B) PT model for the *eutrophic* class based on TSI_TP

* A. Naumoski, G. Mirceva, and K. Mitreski, " A novel fuzzy based approach for inducing diatom habitat models and discovering diatom indicating properties," Ecological Informatics 7(1), pp. 62--70, 2012

Fuzzy based decision tree models



- (A) PT model for the *brackish* class based on Conductivity parameter.
- (B) PT model for the *alkalibiontic* class based on pH parameter.

- **Fuzzy Pattern Trees are more accurate, but they demand more processing power.**

Dataset description so far…., However!!

- The dataset used in these experiments is collected during 16 month period on 14 lake measurement stations (freq = 1 month) in total **224 collected samples.**

- Physico–chemical parameter set includes (temperature, dissolved oxygen, Secchi depth, conductivity, alkalinity (pH), nitrogen compounds (NO2, NO3, NH4, inorganic nitrogen), Sulphur oxide ions SO4, and Sodium (Na), Potassium (K), Magnesium (Mg), Copper (Cu), Manganese (Mn) and Zinc (Zn)).

- Biological information of the ecological system was represented with 116 different diatoms and their relative abundance.

- Today environmental monitoring and thus more precise modelling requires data with high frequency sampled, high in volume and variety of different type of data.

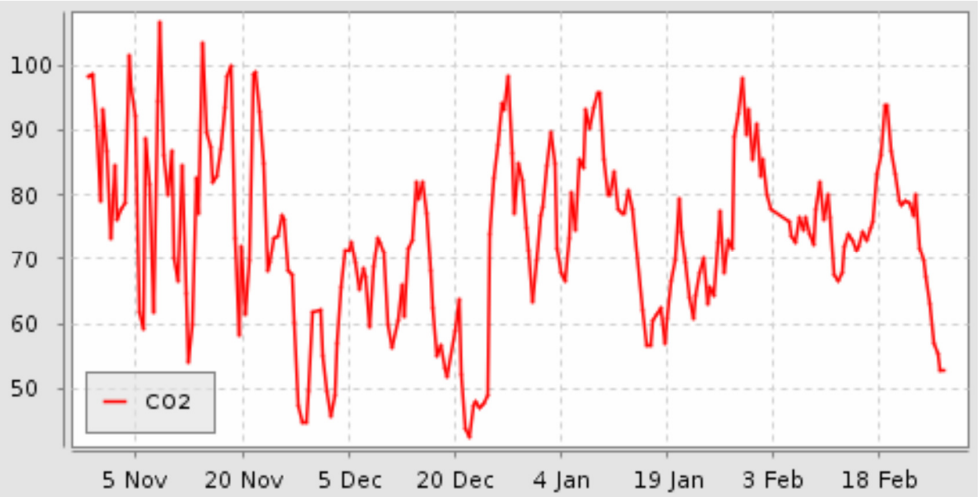
- System based on the IoT.

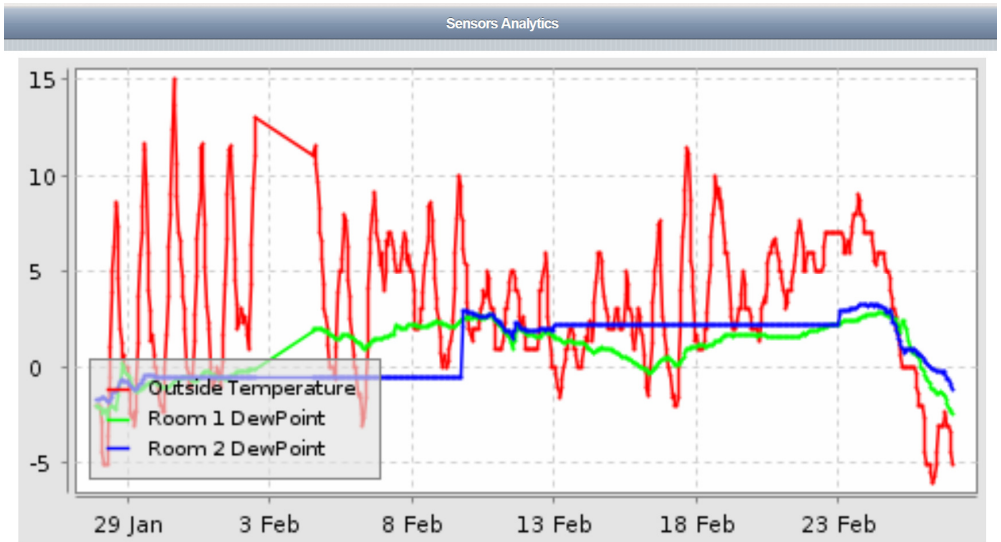
Environmental monitoring using IoT

Last Update	Mon 23:56
Temperature	13.10 °C
Humidity	38.20 %
Abs Humidity	4.36 g/m3
Bar Pressure	976.06 mb
BarPres Temp	12.38 °C
Humindex	No significant
Humindex	10.74 °C
HeatIndex	No Heat Risk
HeatIndex	11.46 °C
Kitchen DewPoint	-0.84 °C
AirQualityIndex	Moderate Pollution
Dust Density	122.73 ug/m3
CO_2 CO2	53.93 ppm
MQ-2	379.6 mV
MQ-7	281.8 mV
MQ-135	310.3 mV

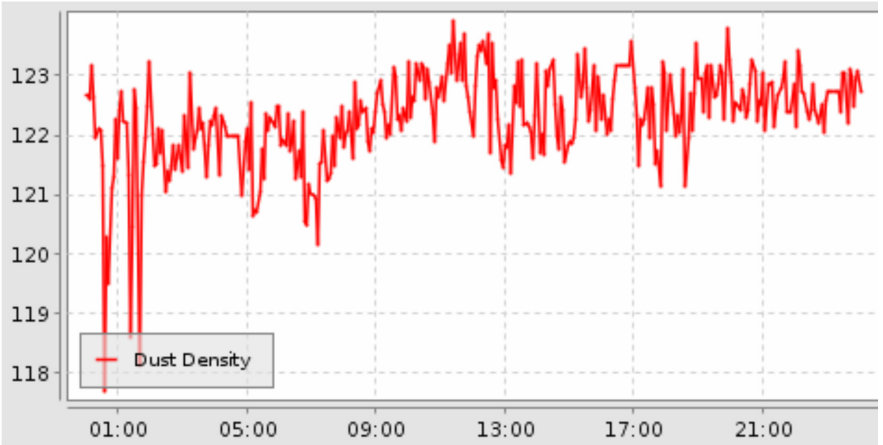
- Working live prototype of multi sensor environmental monitoring suite

Sensors generates a lot off data - BigData





- Traditional environmental monitoring usually measures every 8 hours till 2 weeks, depending on what you measure.



- Frequency of measurement ranges from every 20s, 1 min, 5 min or 20 min, depending of the nature of the measuring parameter.

Automated environmental monitoring station

Typical monitoring station will have:

- Weather station
- Chain of temperature sensors
- Dissolved oxygen
- pH
- Turbidity
- Chlorophyll fluorescence
- All every 2-5 minutes.

Credit to NETLAKE Project.

Pre-processing of the sensor generated data

- In order to process the BigData we need more processing power, and to build better models we need more data and more accurate machine learning algorithms (Fuzzy decision trees, but more power hungry).

- One way is to tamper the sampling frequency, but this is strongly depends from the nature of the measured parameter and the purpose of the measurement itself.

- Another way is to overcome the problem of volume is to remove the unnecessary attributes and reduce the size of the dataset (dimensional reduction) using feature selection.

- Methods like PCA, CCA, DCA, Attribute elevator methods (CorrelationAttributeEval technique, InfoGainAttributeEval Attribute Evaluator, Genetic algorithm, ant optimization and etc.), wrapper subset eval methods based on particular machine learning method.

Pre-processing of the sensor generated data

- **OR**, we can use more advance methods like fuzzy set theory concepts and fuzzy-rough redact concept to further improve the feature selection process.
  - A. Naumoski, and M. Mircev, "**Novel fuzzy measure of similarity for fuzzy-rough feature selection**," In ICT Innovations 2015, S. Loskoska, S. Koceski (Eds.), ISSN 1857-7288, pp. 11-21, Ohrid, 2015.
  - Naumoski A., Mirceva G., Lameski P. (2018) **Influence of Fuzzy Tolerance Metrics on Classification and Regression Tasks for Fuzzy-Rough Nearest Neighbour Algorithms**. In: Stojanov G., Kulakov A. (eds) ICT Innovations 2016. ICT Innovations 2016. Advances in Intelligent Systems and Computing, vol 665. Springer, Cham, doi: https://doi.org/10.1007/978-3-319-68855-8_11
  - A. Naumoski, G. Mirceva., and K. Mitreski**.," Influence of Algebraic T-norm on Different Indiscernibility Relationships in Fuzzy-Rough Rule Induction Algorithms**". In: Trajanov D., Bakeva V. (eds) ICT Innovations 2017. ICT Innovations 2017. Communications in Computer and Information Science, vol. 778. Springer, Cham doi: 10.1007/978-3-319-67597-8

- We have to be careful not to remove the features that describes the target attribute.

- And if this fails, there is always the way…

Massive data environmental analysis. Future work.

- Brute Force with pure CPU processing power, and hope for the best.

- Incremental learning of decision trees from streams of data.
  - The algorithm performs online and in real-time, observes each example only once at the speed of arrival, and maintains at any-time a ready-to-use model tree.

- Using ensemble's of decision trees for data streams.
  - Online Bagging*
  - Online Boosting*
  - Online RandomForest*

- Any many other machine learning methods for classification, regression, clustering, outlier detection, concept drift detection and recommender systems, which are not mention in this presentation, but deserve their place in this research area!!

*Ikonomovska E, Gama J, Džeroski S. Online tree-based ensembles and option trees for regression on evolving data streams. Neurocomputing. 2015 Feb 20;150:458-70.

Conclusion

- The presented research mainly focuses on water ecosystems, however air, soil or any other type of environmental system can be similar modelled with these machine learning algorithms.

- Many of the classical decision tree methods are easy and fast to be used in rather small datasets, but for more accurate modelling fuzzy methods provide excellent platform with a cost of performance price attach to them.

- Feature selection methods and methods for incremental learning of data streams provide interesting option to tackle some of the properties behind the concept of Bigdata in environmental modelling.

- There is no easy way to tackle the BigData when it comes to environmental modelling.

Q&A Section

Thank you for yourattention

Any Questions?

Contact E-mail: andreja.naumoski@finki.ukim.mk