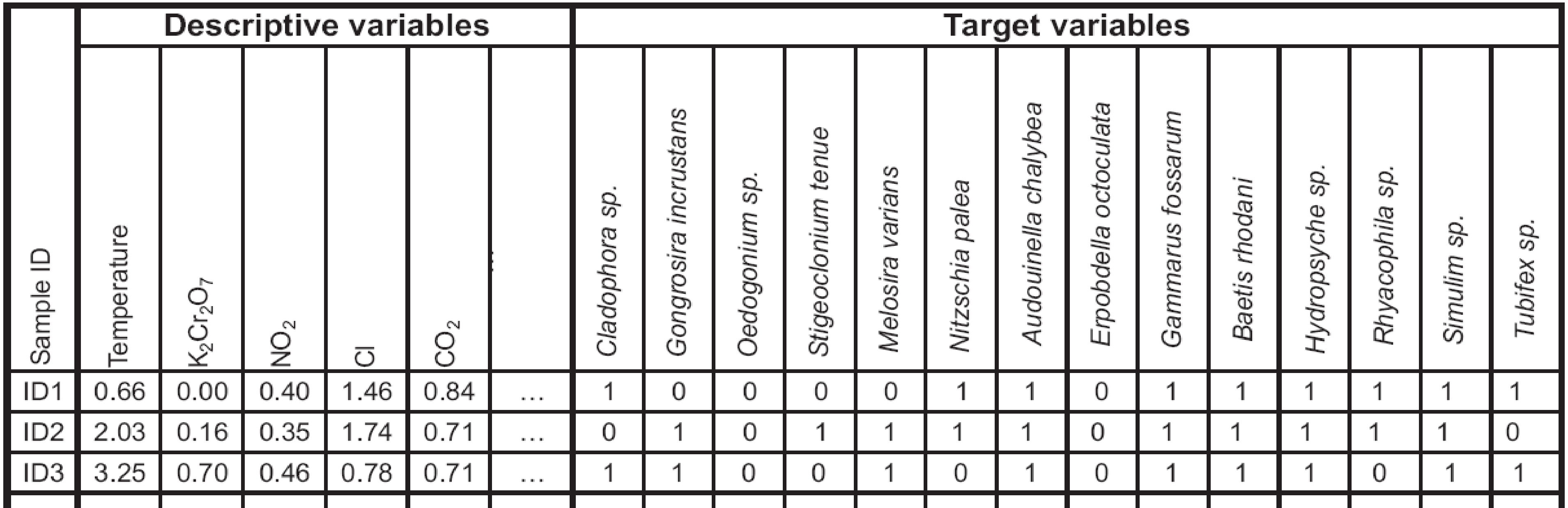
Selected Env. Apps. Of Structured Output Prediction

Sašo Džeroski
Jozef Stefan Institute, Ljubljana, Slovenia

MAESTRA
LEARNING FROM MASSIVE, INCOMPLETELY ANNOTATED, AND STRUCTURED DATA

Environment <-> Biota

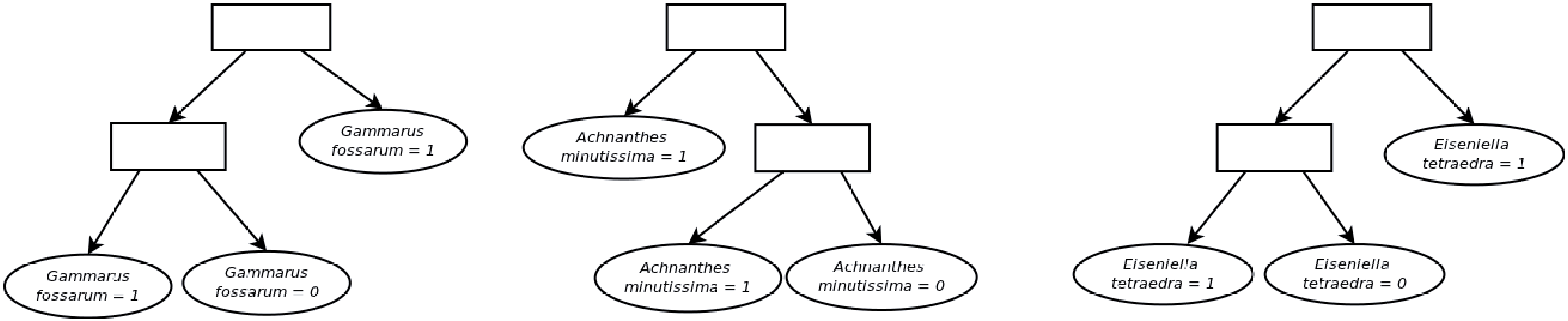
- Predict the biota (or specific components of it)
- At a given site
- From characteristics of the environment at the site
- E.g. predict river water biota from water properties



	Descriptive variables						Target variables													
Sample ID	Temperature	$K_2Cr_2O_7$	NO_2	Cl	CO_2		*Cladophora sp.*	*Gongrosira incrustans*	*Oedogonium sp.*	*Stigeoclonium tenue*	*Melosira varians*	*Nitzschia palea*	*Audouinella chalybea*	*Erpobdella octoculata*	*Gammarus fossarum*	*Baetis rhodani*	*Hydropsyche sp.*	*Rhyacophila sp.*	*Simulim sp.*	*Tubifex sp.*
ID1	0.66	0.00	0.40	1.46	0.84	...	1	0	0	0	0	1	1	0	1	1	1	1	1	1
ID2	2.03	0.16	0.35	1.74	0.71	...	0	1	0	1	1	1	1	0	1	1	1	1	1	0
ID3	3.25	0.70	0.46	0.78	0.71	...	1	1	0	0	1	0	1	0	1	1	1	0	1	1

Habitat modeling

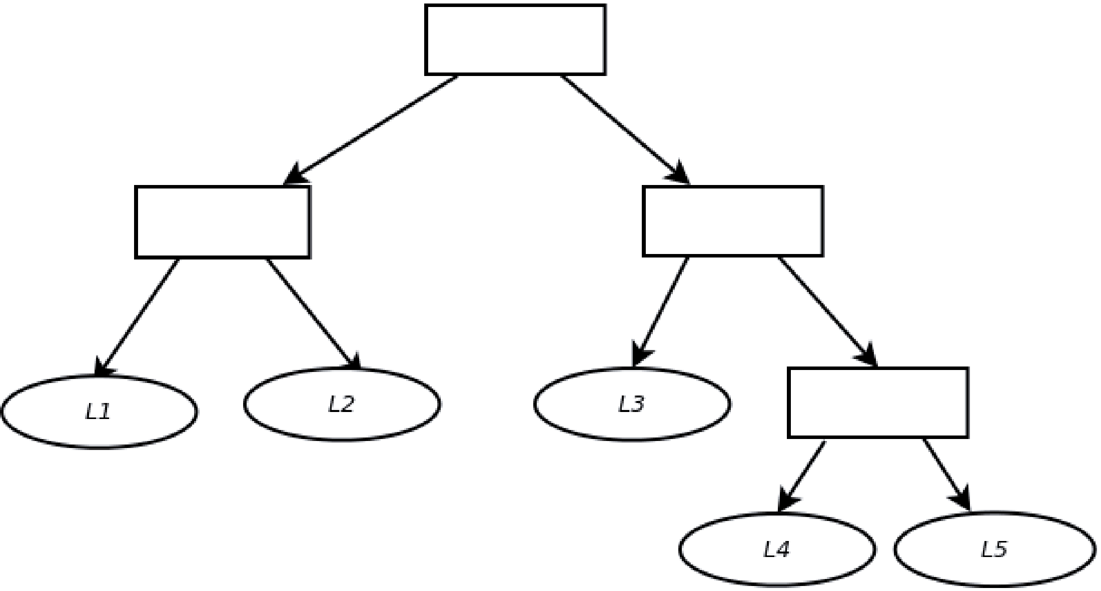
- Model the presence & absence (abundance) of each species separately



- Binary Classification (Regression)

Predicting species composition

- One model for **all the species at once**



L1:
Gammarus fossarum: 0
Achnanthes minutissima: 1
Eiseniella tetraedra: 1

L2:
Gammarus fossarum: 0
Achnanthes minutissima: 1
Eiseniella tetraedra: 0

L3:
Gammarus fossarum: 1
Achnanthes minutissima: 1
Eiseniella tetraedra: 1

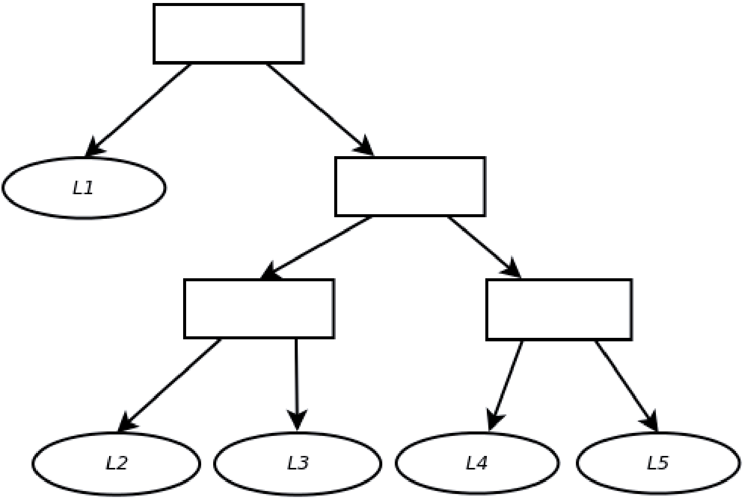
L4:
Gammarus fossarum: 1
Achnanthes minutissima: 1
Eiseniella tetraedra: 0

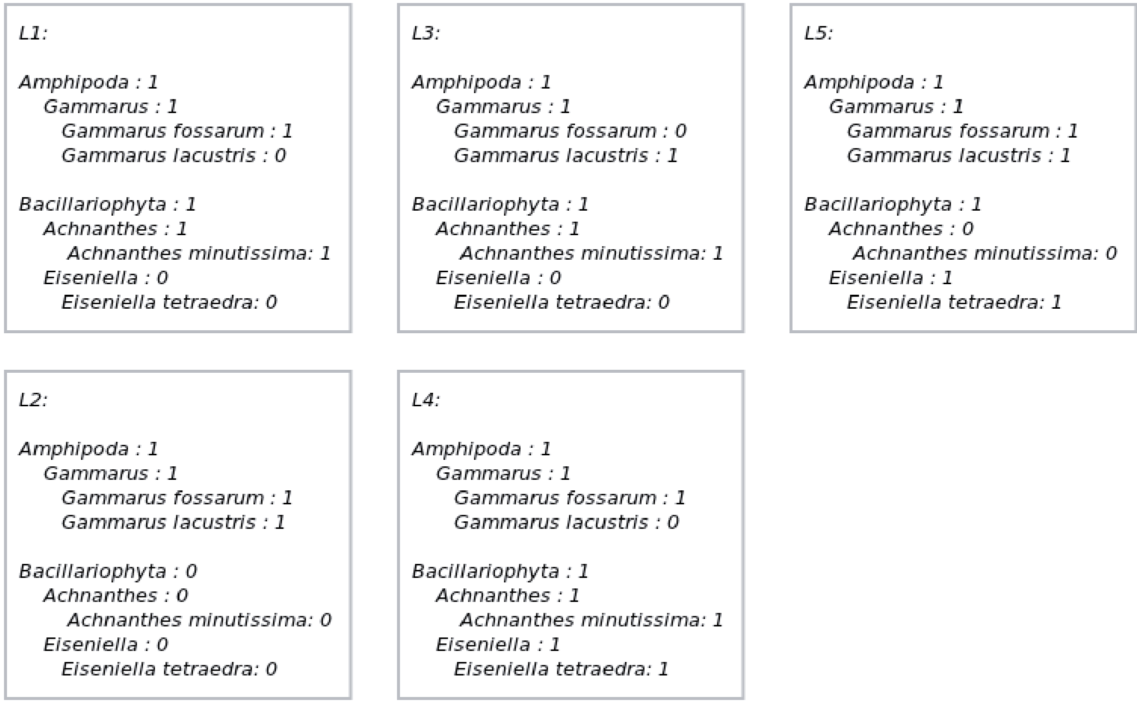
L5:
Gammarus fossarum: 0
Achnanthes minutissima: 1
Eiseniella tetraedra: 1

- **Multi-target classification/regression**

Predicting community structure

- One model for all of the species at once, additionally using the taxonomical hierarchy





L1:

- *Amphipoda : 1*
  - *Gammarus : 1*
    - *Gammarus fossarum : 1*
    - *Gammarus lacustris : 0*
- *Bacillariophyta : 1*
  - *Achnanthes : 1*
    - *Achnanthes minutissima: 1*
  - *Eiseniella : 0*
    - *Eiseniella tetraedra: 0*

L2:

- *Amphipoda : 1*
  - *Gammarus : 1*
    - *Gammarus fossarum : 1*
    - *Gammarus lacustris : 1*
- *Bacillariophyta : 0*
  - *Achnanthes : 0*
    - *Achnanthes minutissima: 0*
  - *Eiseniella : 0*
    - *Eiseniella tetraedra: 0*

L3:

- *Amphipoda : 1*
  - *Gammarus : 1*
    - *Gammarus fossarum : 0*
    - *Gammarus lacustris : 1*
- *Bacillariophyta : 1*
  - *Achnanthes : 1*
    - *Achnanthes minutissima: 1*
  - *Eiseniella : 0*
    - *Eiseniella tetraedra: 0*

L4:

- *Amphipoda : 1*
  - *Gammarus : 1*
    - *Gammarus fossarum : 1*
    - *Gammarus lacustris : 0*
- *Bacillariophyta : 1*
  - *Achnanthes : 1*
    - *Achnanthes minutissima: 1*
  - *Eiseniella : 1*
    - *Eiseniella tetraedra: 1*

L5:

- *Amphipoda : 1*
  - *Gammarus : 1*
    - *Gammarus fossarum : 1*
    - *Gammarus lacustris : 1*
- *Bacillariophyta : 1*
  - *Achnanthes : 0*
    - *Achnanthes minutissima: 0*
  - *Eiseniella : 1*
    - *Eiseniella tetraedra: 1*

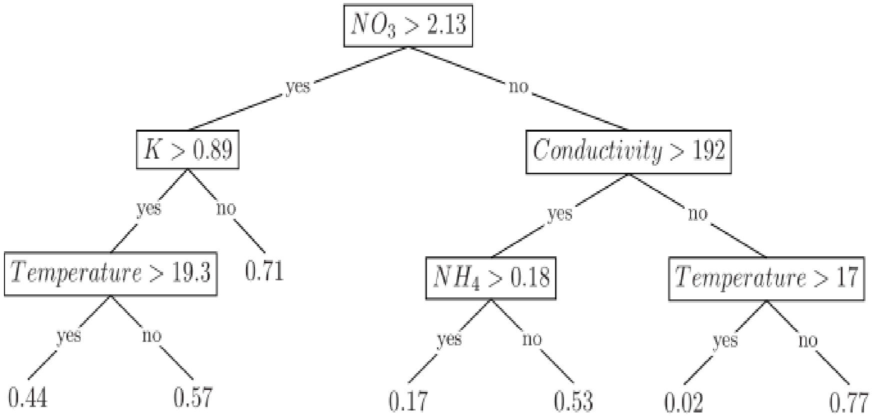
- Hierarchical multi-label classification

Slovenian rivers

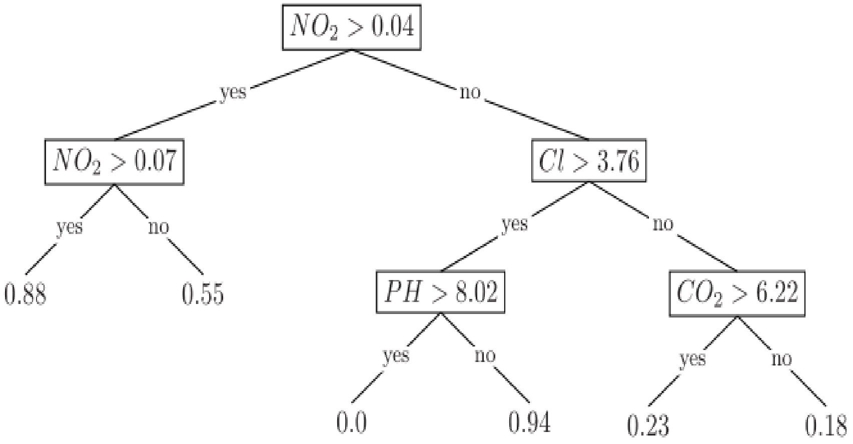
- 1.060 samples
- 16 physical and chemical props. of water, 491 species
- data collected in 1990-1995

- ephemeroptera
  - ephemeroptera_acantrella
    - ephemeroptera_acantrella_sinaica
  - ephemeroptera_baetidae
  - ephemeroptera_baetis
    - ephemeroptera_baetis_alpinus
    - ephemeroptera_baetis_buceratus
    - ephemeroptera_baetis_fuscatus
    - ephemeroptera_baetis_muticus
    - **ephemeroptera_baetis_rhodani**
    - ephemeroptera_baetis_scambus
    - ephemeroptera_baetis_vernus
  - ephemeroptera_ecdyonurus
    - ephemeroptera_ecdyonurus_forcipula
    - ephemeroptera_ecdyonurus_helveticus
    - ephemeroptera_ecdyonurus_insignis
    - ephemeroptera_ecdyonurus_torrentis
    - ephemeroptera_ecdyonurus_venosus
  - ephemeroptera_electrogena
    - ephemeroptera_electrogena_lateralis
    - ephemeroptera_electrogena_quadrilineata
- plecoptera
  - plecoptera_amphinemura
    - plecoptera_amphinemura_triangularis
  - plecoptera_brachyptera
    - **plecoptera_brachyptera_risi**
    - plecoptera_brachyptera_seticornis

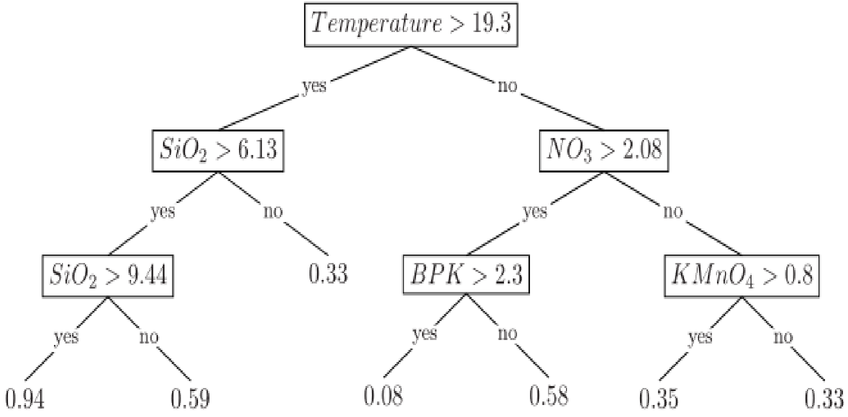
Slovenian rivers: Habitat models



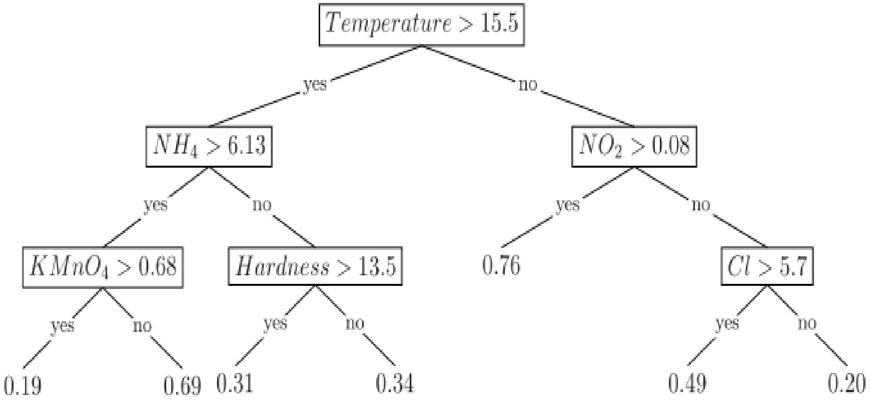
Bacillariophyta Cyclotella Comta



Bacillariophyta Nitzschia Palea



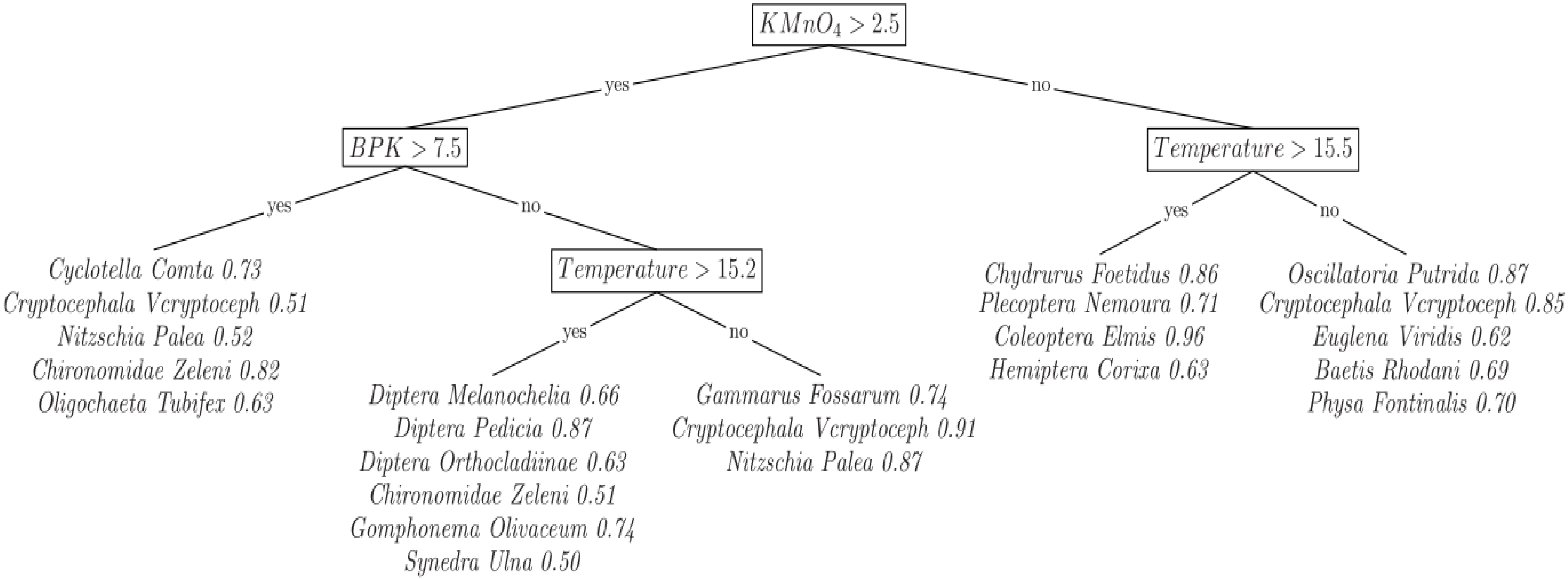
Diptera Chironomidae Zeleni



Bacillariophyta Navicula Cryptocephala Vcryptoceph

Slovenian rivers: Species comp.

- MLC: Multi-label classification tree



$KMnO_4 > 2.5$

yes

$BPK > 7.5$

yes

Cyclotella Comta 0.73
Cryptocephala Vcryptoceph 0.51
Nitzschia Palea 0.52
Chironomidae Zeleni 0.82
Oligochaeta Tubifex 0.63

no

$Temperature > 15.2$

yes

Diptera Melanochelia 0.66
Diptera Pedicia 0.87
Diptera Orthocladiinae 0.63
Chironomidae Zeleni 0.51
Gomphonema Olivaceum 0.74
Synedra Ulna 0.50

no

Gammarus Fossarum 0.74
Cryptocephala Vcryptoceph 0.91
Nitzschia Palea 0.87

no

$Temperature > 15.5$

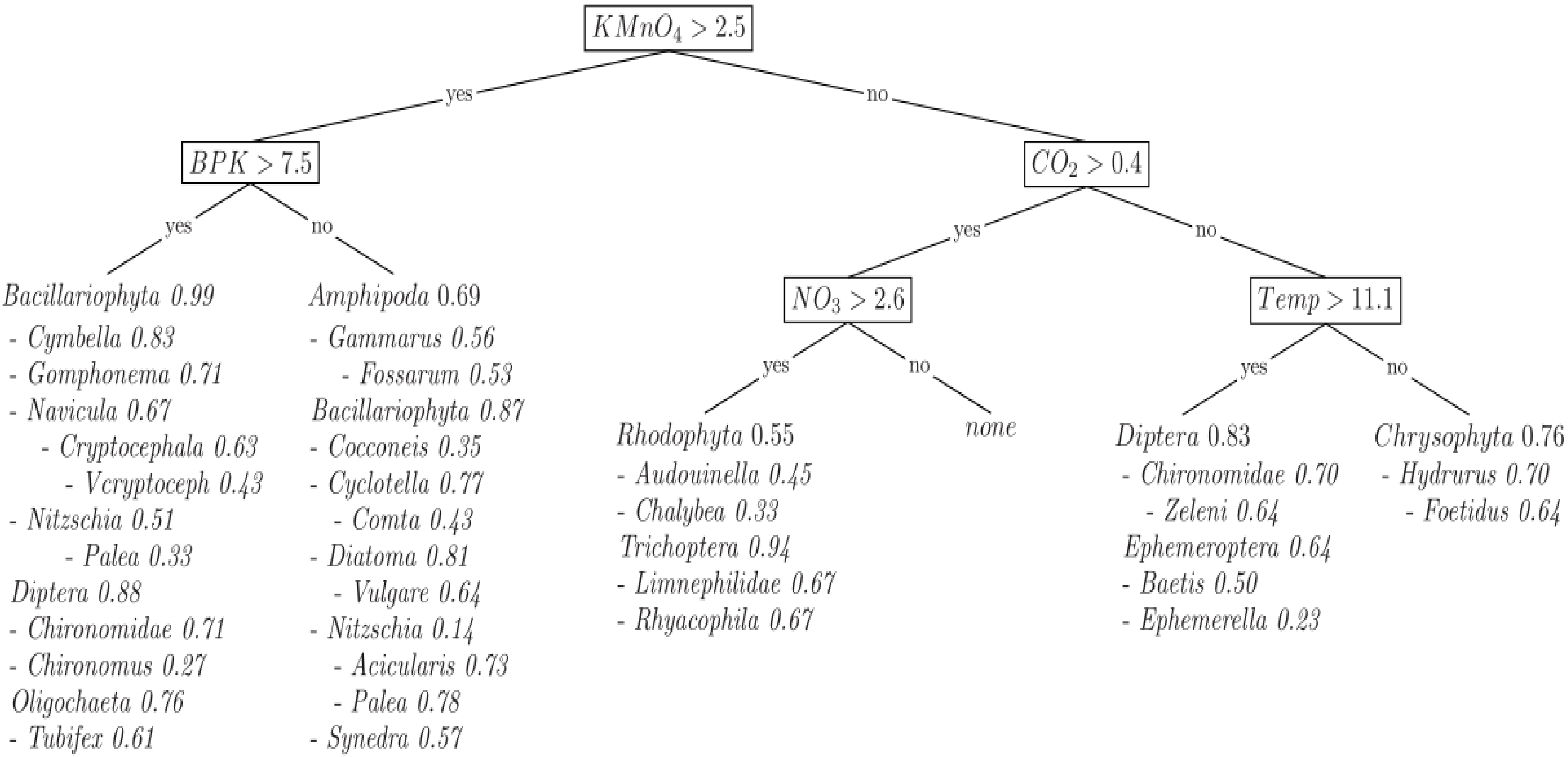
yes

Chydrurus Foetidus 0.86
Plecoptera Nemoura 0.71
Coleoptera Elmis 0.96
Hemiptera Corixa 0.63

no

Oscillatoria Putrida 0.87
Cryptocephala Vcryptoceph 0.85
Euglena Viridis 0.62
Baetis Rhodani 0.69
Physa Fontinalis 0.70

Slovenian rivers: Community struc.



$KMnO_4 > 2.5$

yes

$BPK > 7.5$

yes

Bacillariophyta 0.99
- *Cymbella 0.83*
- *Gomphonema 0.71*
- *Navicula 0.67*
  - *Cryptocephala 0.63*
    - *Vcryptoceph 0.43*
- *Nitzschia 0.51*
  - *Palea 0.33*

Diptera 0.88
- *Chironomidae 0.71*
- *Chironomus 0.27*

Oligochaeta 0.76
- *Tubifex 0.61*

no

Amphipoda 0.69
- *Gammarus 0.56*
  - *Fossarum 0.53*

Bacillariophyta 0.87
- *Cocconeis 0.35*
- *Cyclotella 0.77*
  - *Comta 0.43*
- *Diatoma 0.81*
  - *Vulgare 0.64*
- *Nitzschia 0.14*
  - *Acicularis 0.73*
  - *Palea 0.78*
- *Synedra 0.57*

no

$CO_2 > 0.4$

yes

$NO_3 > 2.6$

yes

Rhodophyta 0.55
- *Audouinella 0.45*
- *Chalybea 0.33*

Trichoptera 0.94
- *Limnephilidae 0.67*
- *Rhyacophila 0.67*

no

none

no

$Temp > 11.1$

yes

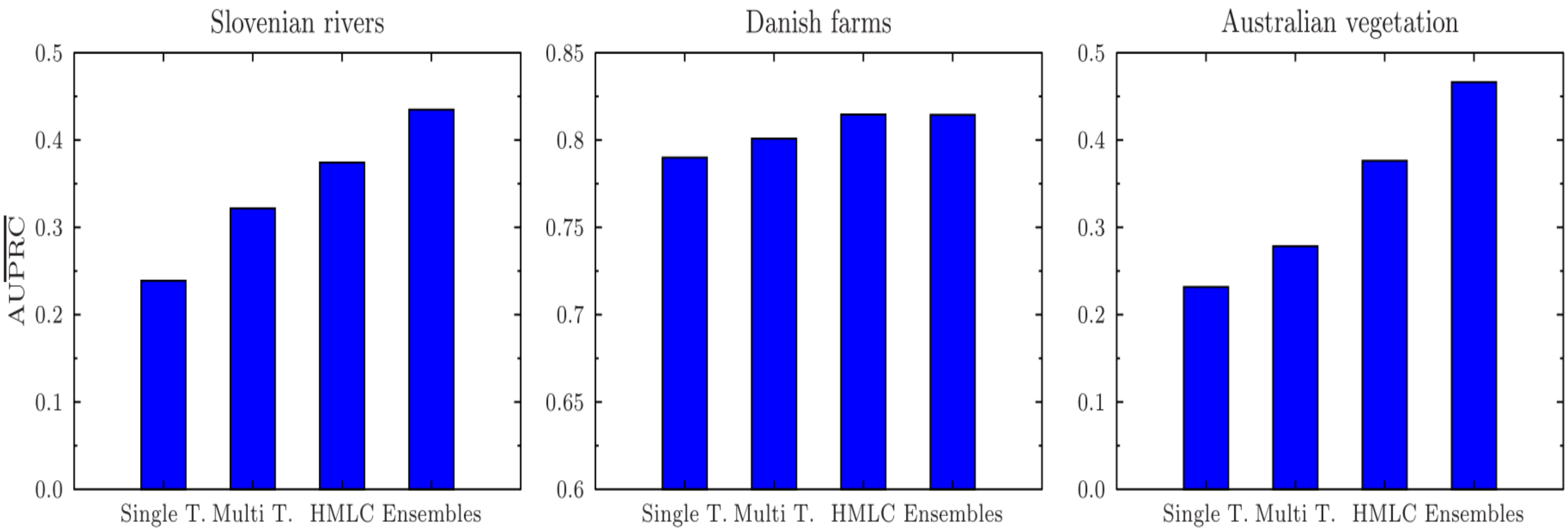
Diptera 0.83
- *Chironomidae 0.70*
  - *Zeleni 0.64*

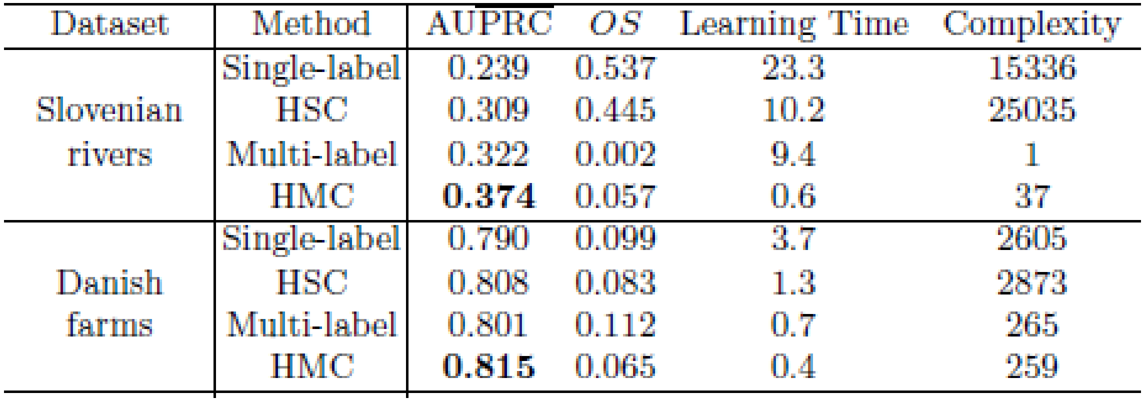
Ephemeroptera 0.64
- *Baetis 0.50*
- *Ephemerella 0.23*

no

Chrysophyta 0.76
- *Hydrurus 0.70*
  - *Foetidus 0.64*

Slovenian rivers: Overall results

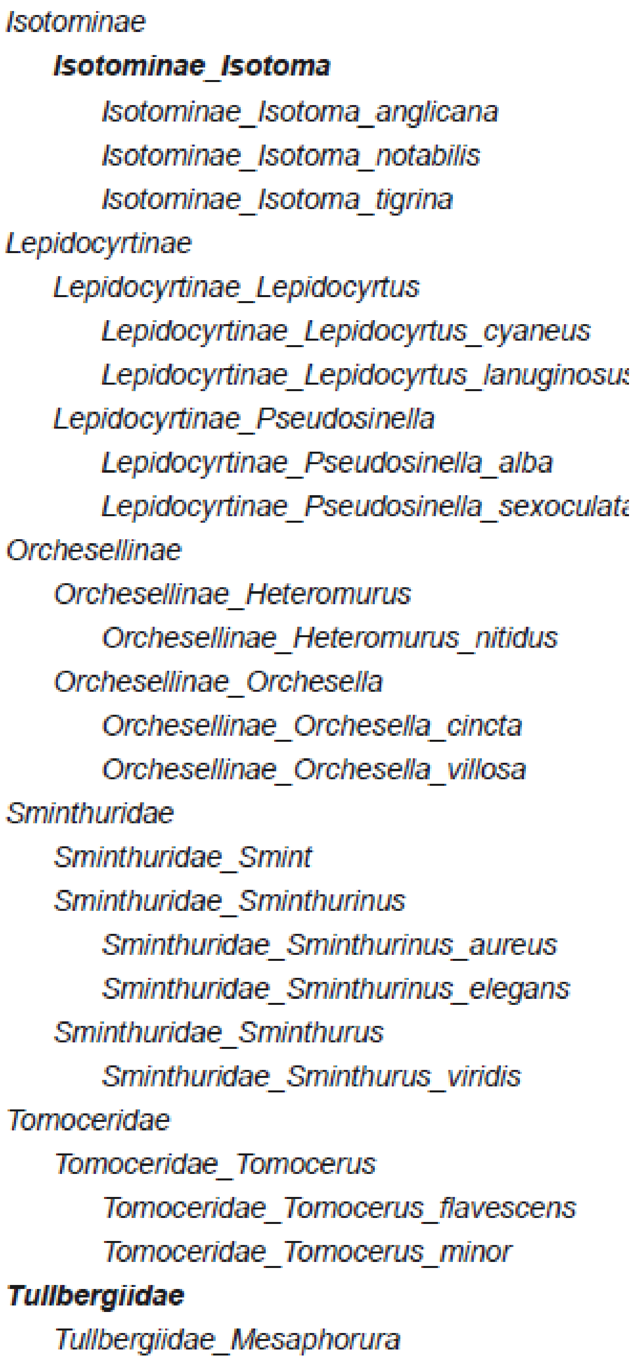




Dataset	Method	$\overline{\mathrm{AUPRC}}$	OS	Learning Time	Complexity
Slovenian rivers	Single-label	0.239	0.537	23.3	15336
	HSC	0.309	0.445	10.2	25035
	Multi-label	0.322	0.002	9.4	1
	HMC	**0.374**	0.057	0.6	37
Danish farms	Single-label	0.790	0.099	3.7	2605
	HSC	0.808	0.083	1.3	2873
	Multi-label	0.801	0.112	0.7	265
	HMC	**0.815**	0.065	0.4	259

Danish farms:
Soil Microarthropods

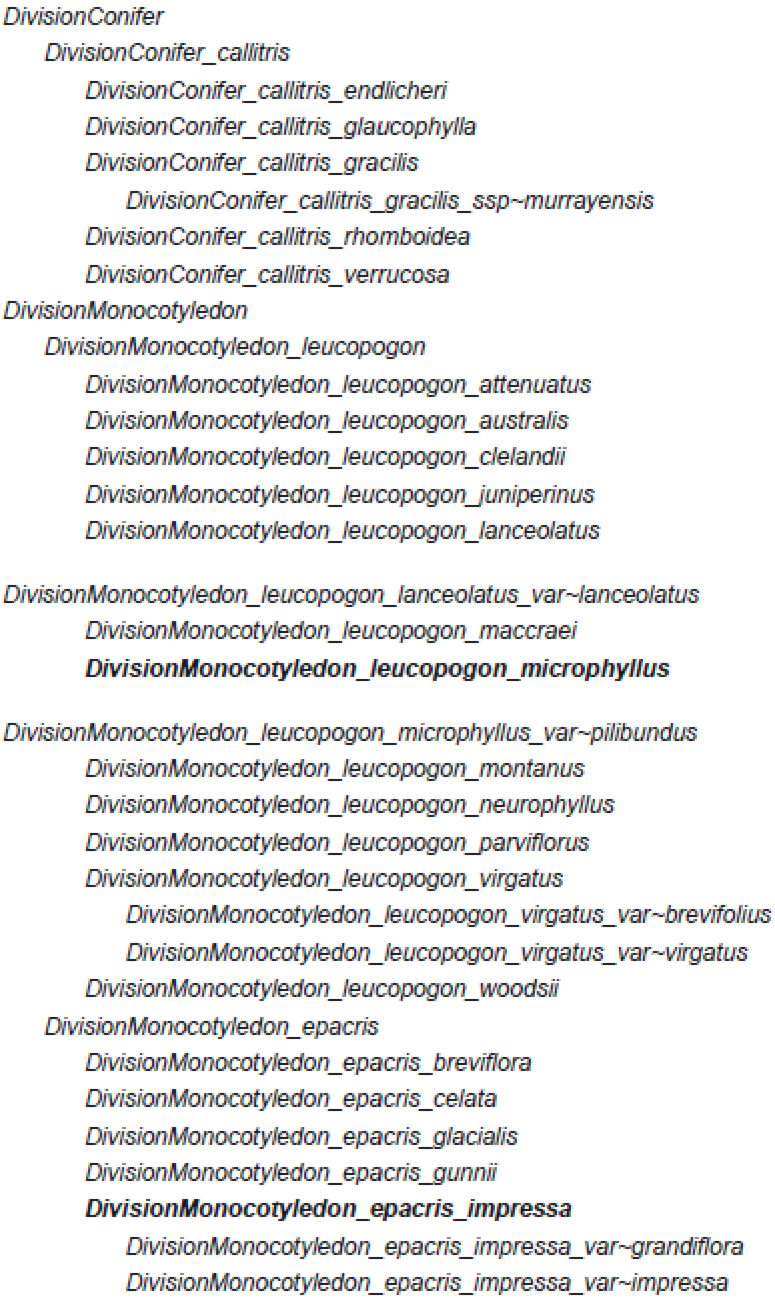
- 1.944 soil samples
- 137 attributes/agricultural events and soil biological parameters
- 35 collembolan species
- data collected 1989-1993



- Isotominae
  - **Isotominae_Isotoma**
    - Isotominae_Isotoma_anglicana
    - Isotominae_Isotoma_notabilis
    - Isotominae_Isotoma_tigrina
- Lepidocyrtinae
  - Lepidocyrtinae_Lepidocyrtus
    - Lepidocyrtinae_Lepidocyrtus_cyaneus
    - Lepidocyrtinae_Lepidocyrtus_lanuginosus
  - Lepidocyrtinae_Pseudosinella
    - Lepidocyrtinae_Pseudosinella_alba
    - Lepidocyrtinae_Pseudosinella_sexoculata
- Orchesellinae
  - Orchesellinae_Heteromurus
    - Orchesellinae_Heteromurus_nitidus
  - Orchesellinae_Orchesella
    - Orchesellinae_Orchesella_cincta
    - Orchesellinae_Orchesella_villosa
- Sminthuridae
  - Sminthuridae_Smint
  - Sminthuridae_Sminthurinus
    - Sminthuridae_Sminthurinus_aureus
    - Sminthuridae_Sminthurinus_elegans
  - Sminthuridae_Sminthurus
    - Sminthuridae_Sminthurus_viridis
- Tomoceridae
  - Tomoceridae_Tomocerus
    - Tomoceridae_Tomocerus_flavescens
    - Tomoceridae_Tomocerus_minor
- **Tullbergiidae**
  - Tullbergiidae_Mesaphorura

Victoria, Australia Vegetation

- 27.482 sites
- 81 env. attributes
- 3.173 species

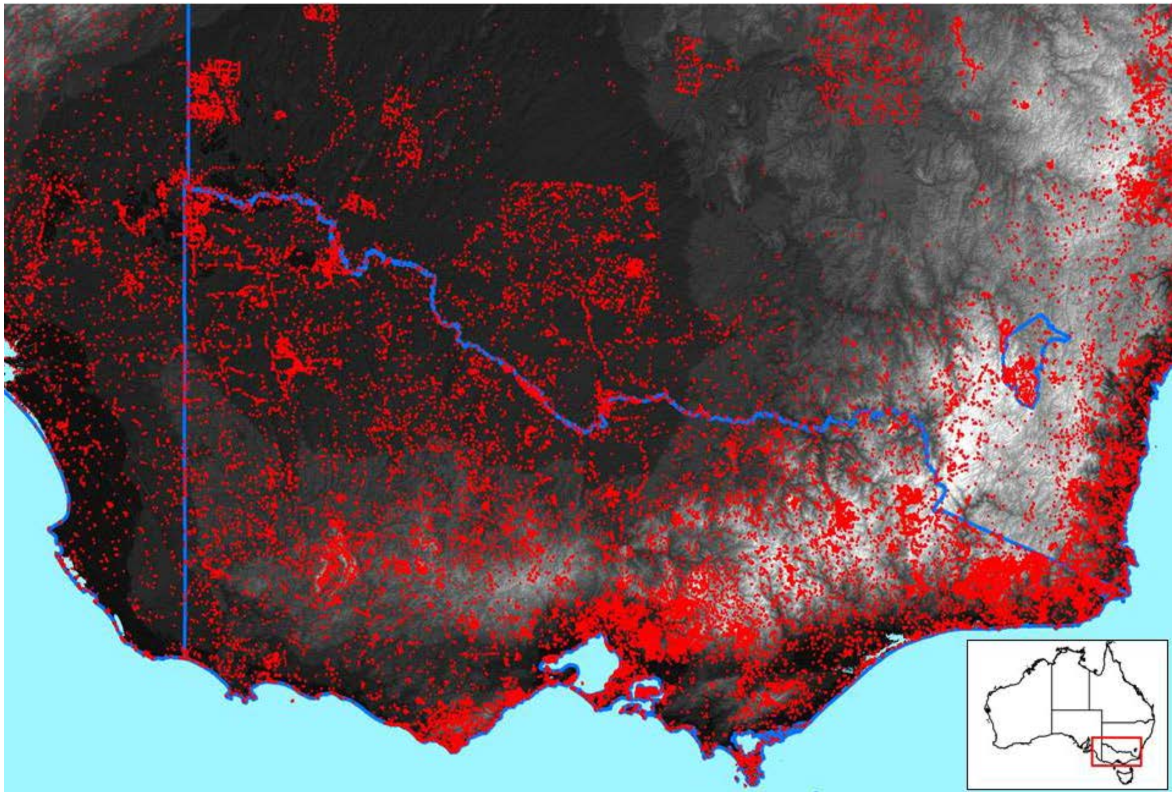


- *DivisionConifer*
  - *DivisionConifer_callitris*
    - *DivisionConifer_callitris_endlicheri*
    - *DivisionConifer_callitris_glaucophylla*
    - *DivisionConifer_callitris_gracilis*
      - *DivisionConifer_callitris_gracilis_ssp~murrayensis*
    - *DivisionConifer_callitris_rhomboidea*
    - *DivisionConifer_callitris_verrucosa*
- *DivisionMonocotyledon*
  - *DivisionMonocotyledon_leucopogon*
    - *DivisionMonocotyledon_leucopogon_attenuatus*
    - *DivisionMonocotyledon_leucopogon_australis*
    - *DivisionMonocotyledon_leucopogon_clelandii*
    - *DivisionMonocotyledon_leucopogon_juniperinus*
    - *DivisionMonocotyledon_leucopogon_lanceolatus*
- *DivisionMonocotyledon_leucopogon_lanceolatus_var~lanceolatus*
    - *DivisionMonocotyledon_leucopogon_maccraei*
    - ***DivisionMonocotyledon_leucopogon_microphyllus***
- *DivisionMonocotyledon_leucopogon_microphyllus_var~pilibundus*
    - *DivisionMonocotyledon_leucopogon_montanus*
    - *DivisionMonocotyledon_leucopogon_neurophyllus*
    - *DivisionMonocotyledon_leucopogon_parviflorus*
    - *DivisionMonocotyledon_leucopogon_virgatus*
      - *DivisionMonocotyledon_leucopogon_virgatus_var~brevifolius*
      - *DivisionMonocotyledon_leucopogon_virgatus_var~virgatus*
    - *DivisionMonocotyledon_leucopogon_woodsii*
  - *DivisionMonocotyledon_epacris*
    - *DivisionMonocotyledon_epacris_breviflora*
    - *DivisionMonocotyledon_epacris_celata*
    - *DivisionMonocotyledon_epacris_glacialis*
    - *DivisionMonocotyledon_epacris_gunnii*
    - ***DivisionMonocotyledon_epacris_impressa***
      - *DivisionMonocotyledon_epacris_impressa_var~grandiflora*
      - *DivisionMonocotyledon_epacris_impressa_var~impressa*

Victoria, Australia: Relating env. char. to plant trait profiles

New, much more extensive data: Collected 1960-2010, 53362 sites, more than 1.35 Mio indiv. spec. obs.

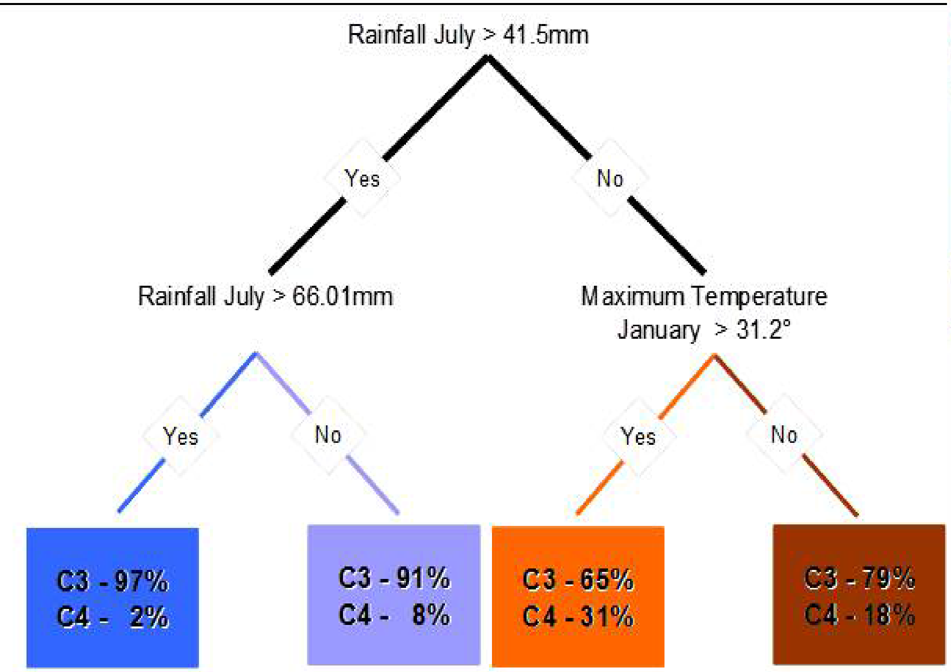
Each vascular species, recorded together with % cover

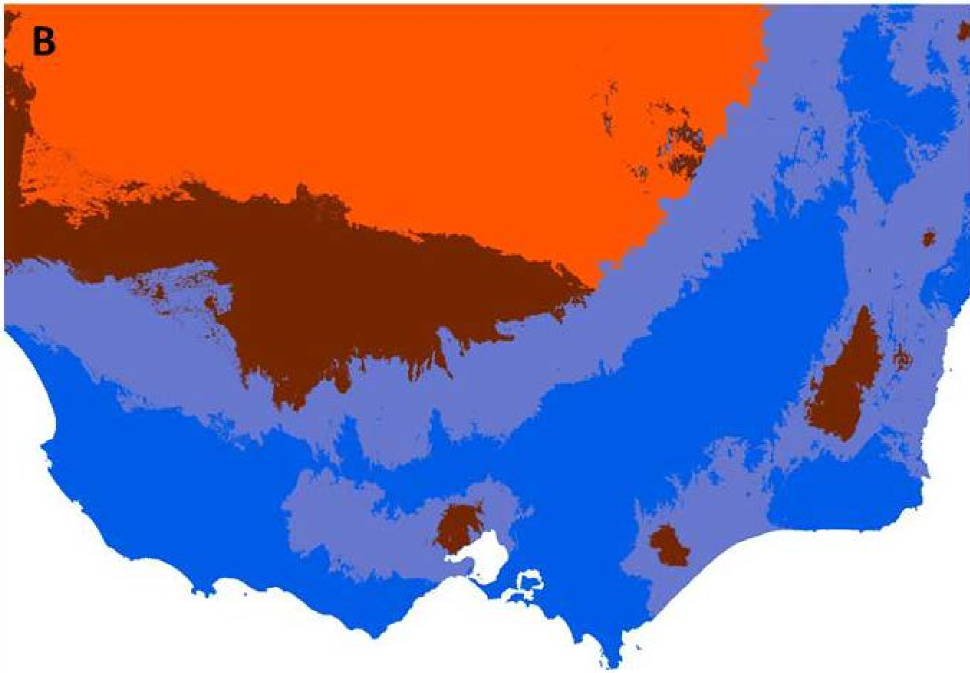


Victoria, Australia: Relating env. char. to plant trait profiles

Plant photosynthetic type (carbon fixation pathways)

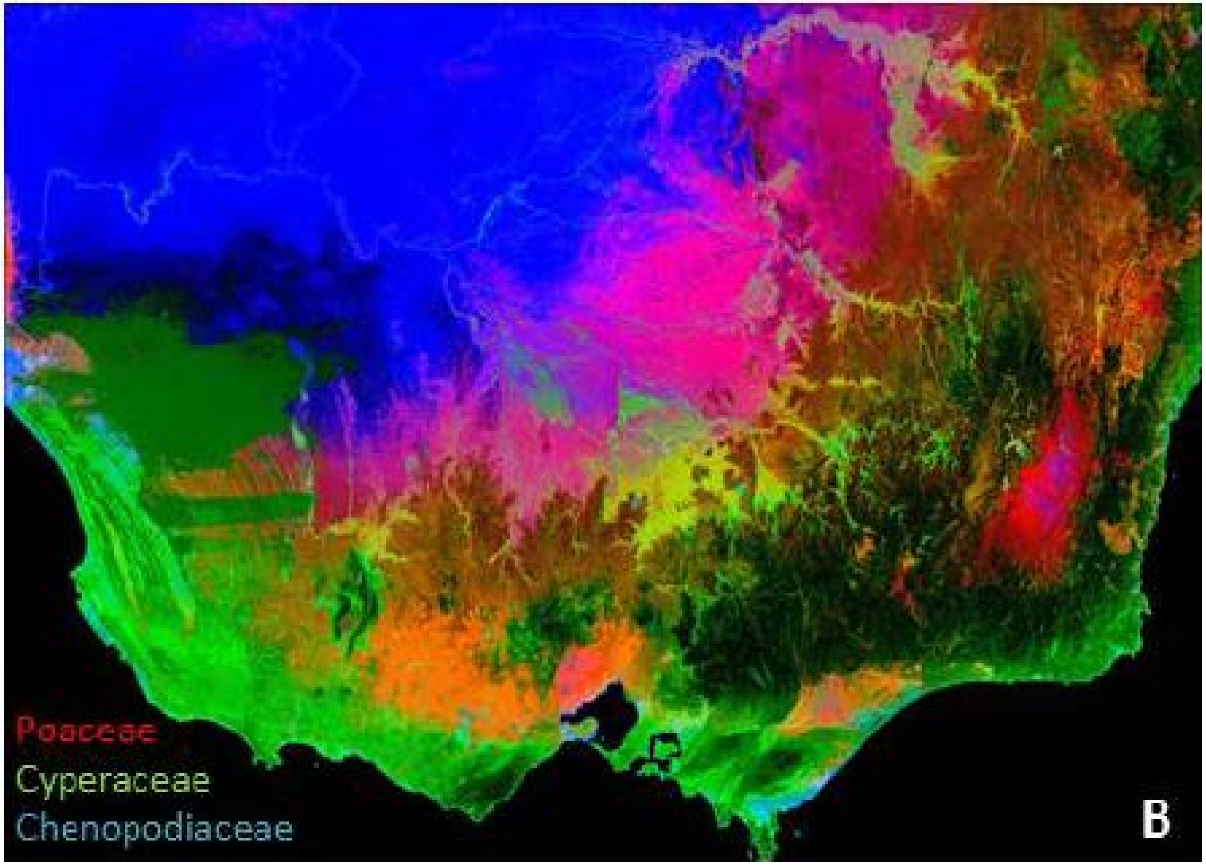
- C3: cool-season-active
- C4: warm-season-active





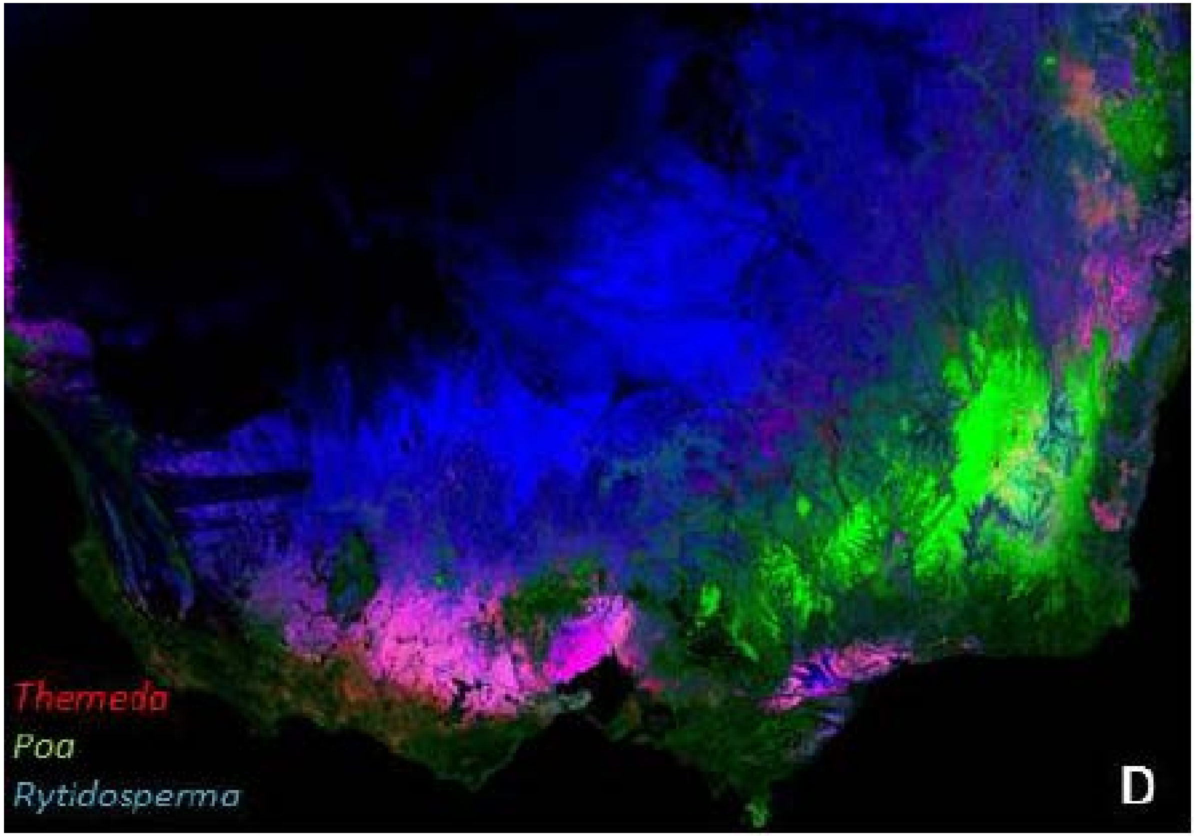
Victoria, Australia: Relating env. char. to plant trait profiles

Phylogeny via main monocot families (Poaceae=Grasses, Cyperaceae=Sedges; Chenopodiaceae=Goosefoots)



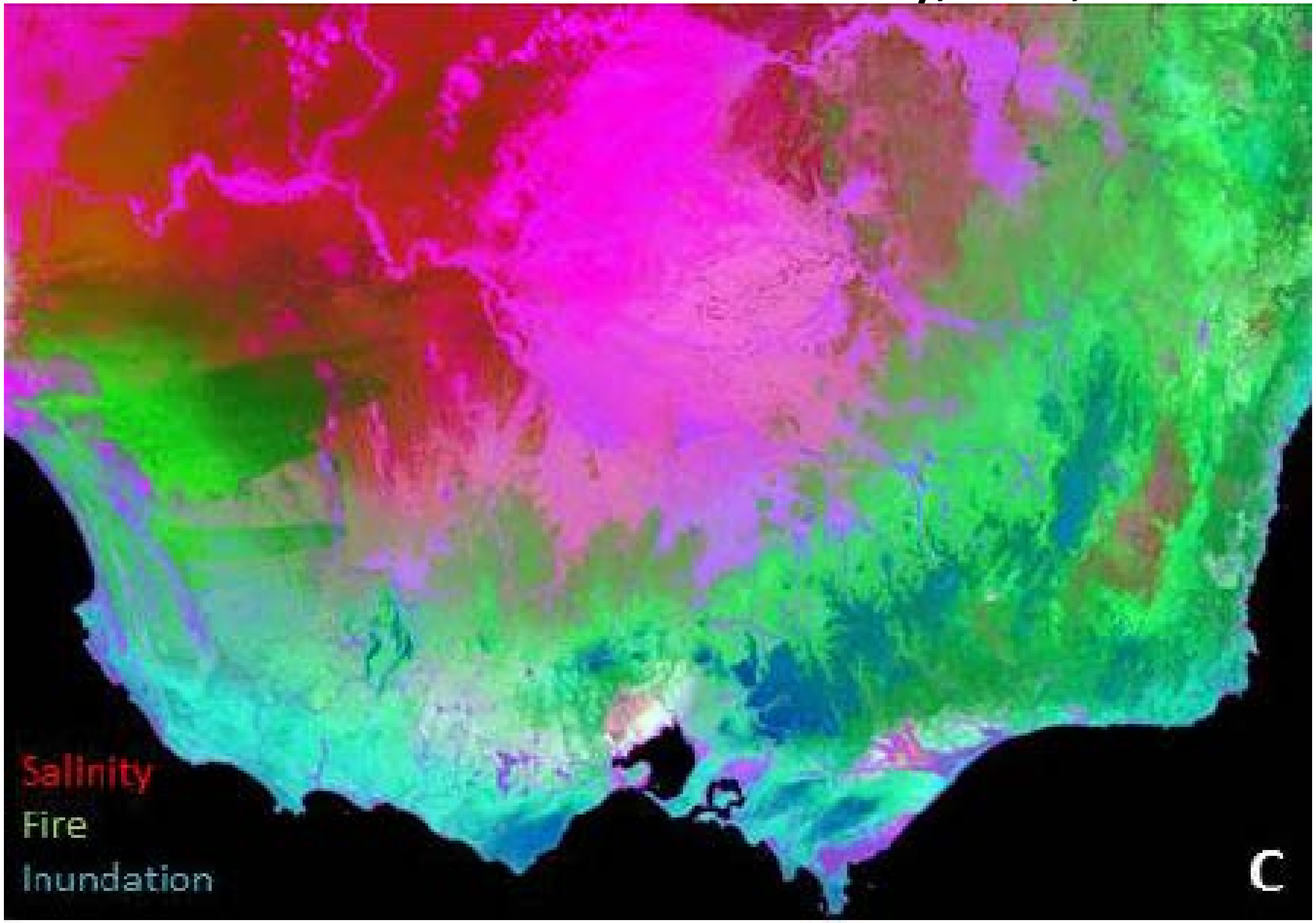
Victoria, Australia: Relating env. char. to plant trait profiles

Phylogeny via three main grass genera



Victoria, Australia: Relating env. char. to plant trait profiles

Stress tolerance: Tolerance to salinity, fire, inundation

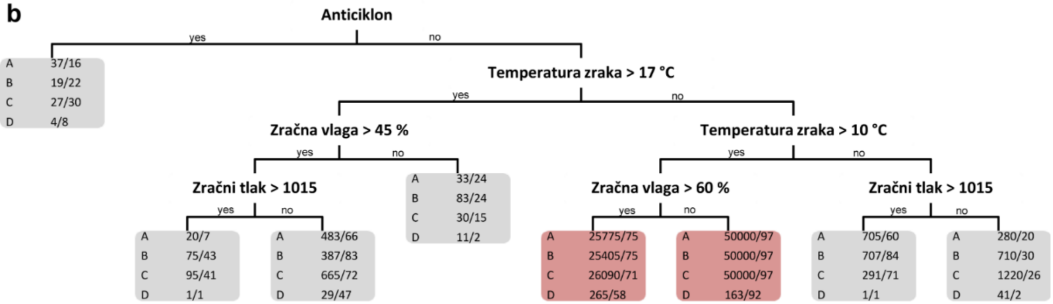


Extremo-philic and –tolerant fungi

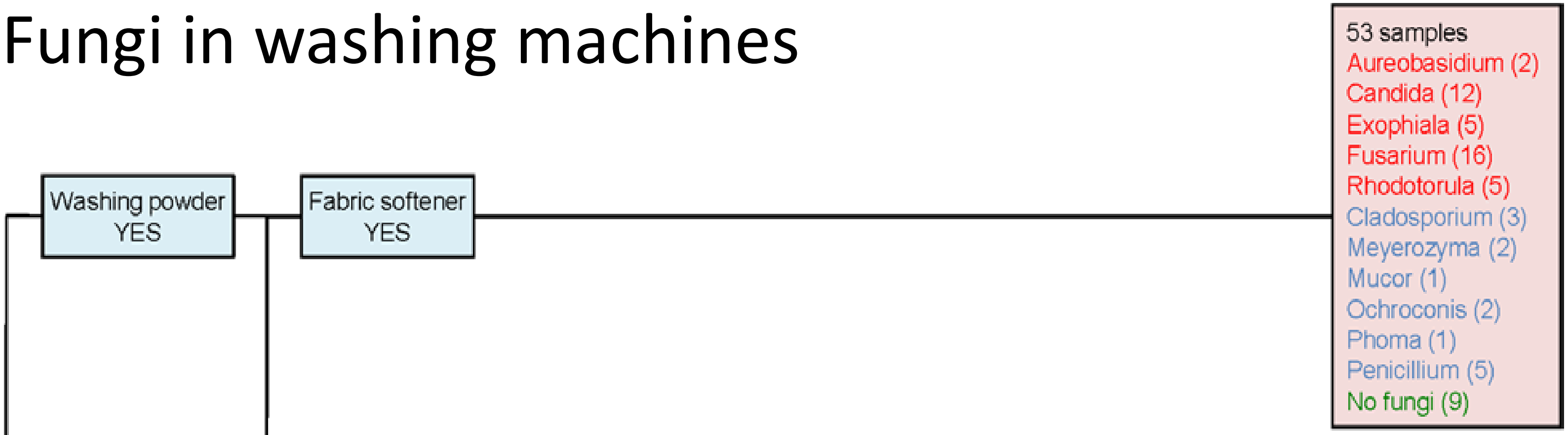
- Can be found in naturally extreme environments, like salterns and Arctic glaciers, but also in
  - The air and water (tap-water)
  - Food preserved with high concentrations of salt/sugar
  - Household appliances (dishwashers, washing machines)
- Can represent a threat to health (e.g. farmers lung disease)
- We analyzed data collected by collaborators at Uni Lj
  - Building habitat models
  - Relating species, env. factors and metabolite composition
  - Using PCTs for MTR and HMLC

Extremo-philic and –tolerant fungi

- Wallemia propagules in the air
  - Highest concentrations in agric. buildings (barns) in early spring
  - Highest concentrations expected: during a cyclone, at temperatures between 10 °C and 17 °C, at relative air humidity below 60%



- Fungi in washing machines



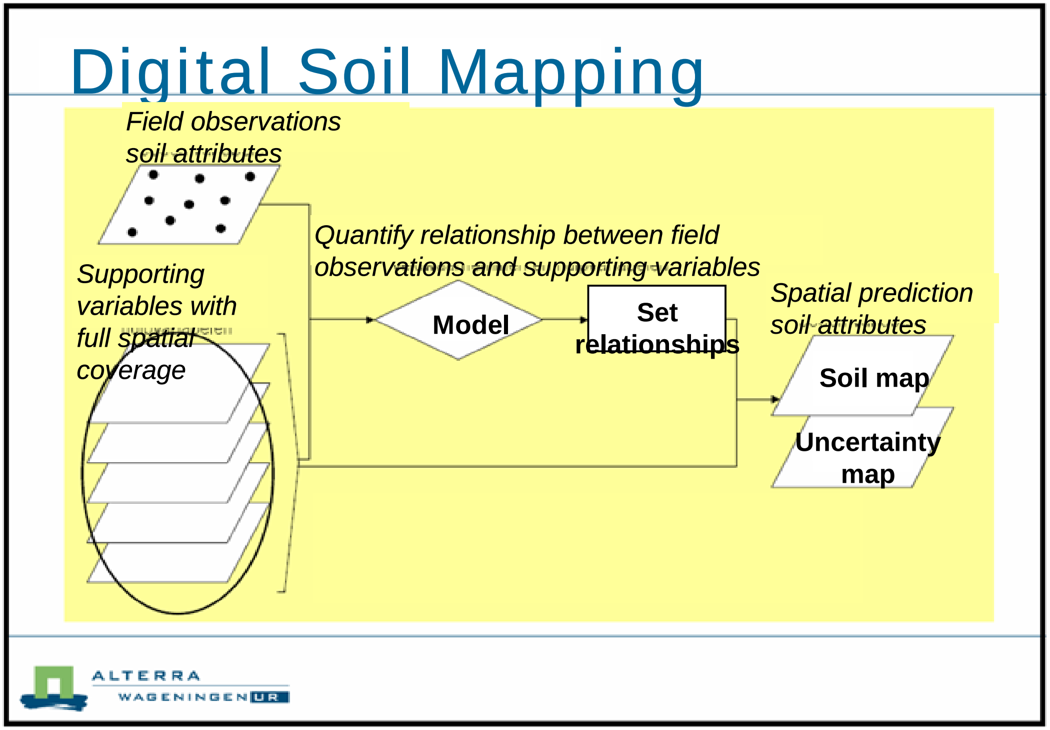
Washing powder YES

Fabric softener YES

53 samples
Aureobasidium (2)
Candida (12)
Exophiala (5)
Fusarium (16)
Rhodotorula (5)
Cladosporium (3)
Meyerozyma (2)
Mucor (1)
Ochroconis (2)
Phoma (1)
Penicillium (5)
No fungi (9)

Mining the ECOFINDERS Dataset

- Goal: Produce maps for features (related to soil functions) that are not measured globally, from those that are measured globally
- Learn mapping from data where both measured



Digital Soil Mapping

Field observations soil attributes

Supporting variables with full spatial coverage

Quantify relationship between field observations and supporting variables

Model

Set relationships

Spatial prediction soil attributes

Soil map

Uncertainty map

Mining the ECOFINDERS Dataset

EcoFinders data (transect with 81 sites)

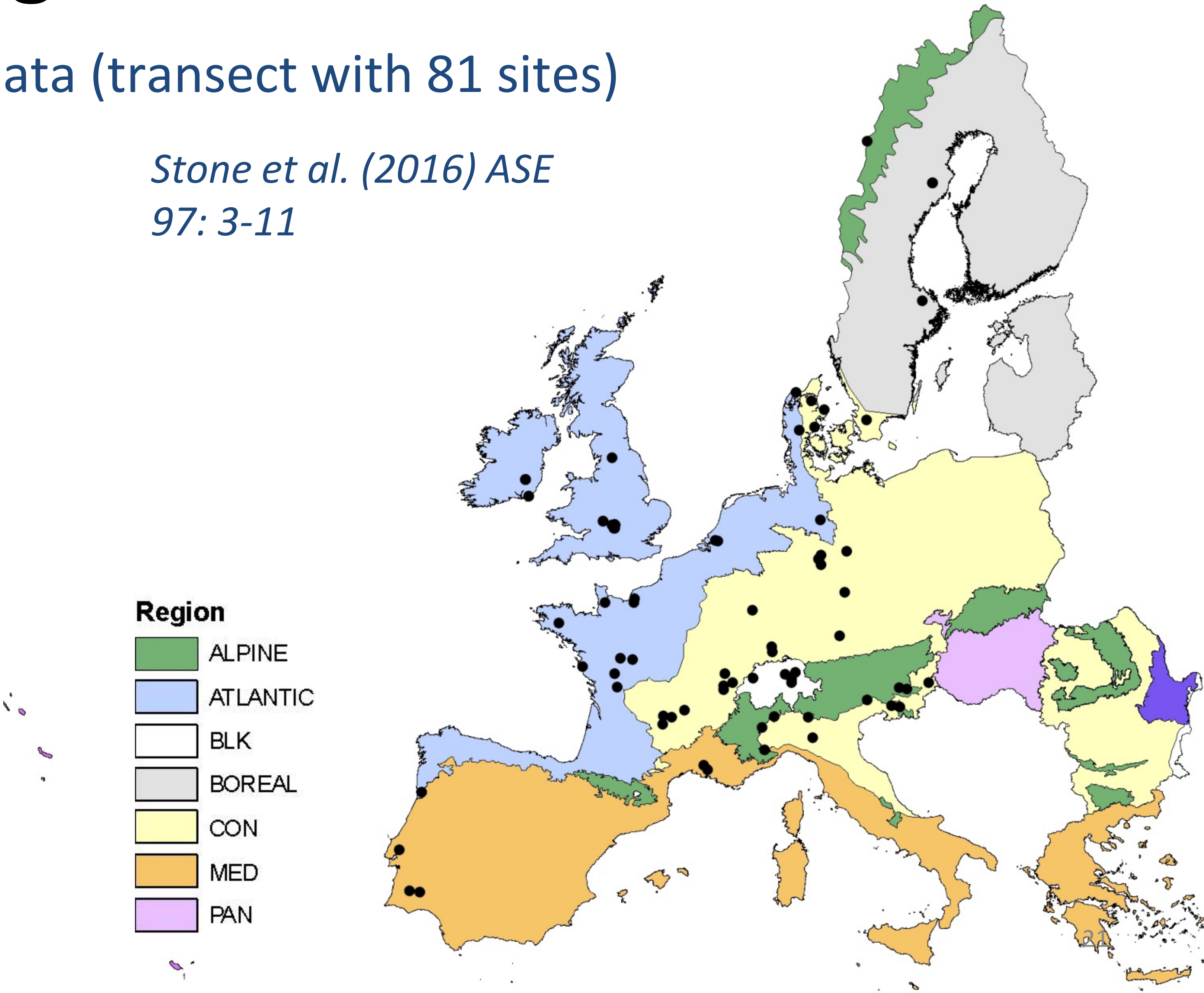
3 land uses:

- Arable
- Grassland
- Forest

4 climate zones:

- Boreal
- Alpine
- Atlantic
- Continental

Stone et al. (2016) ASE 97: 3-11



ECOFINDERS data: Attributes

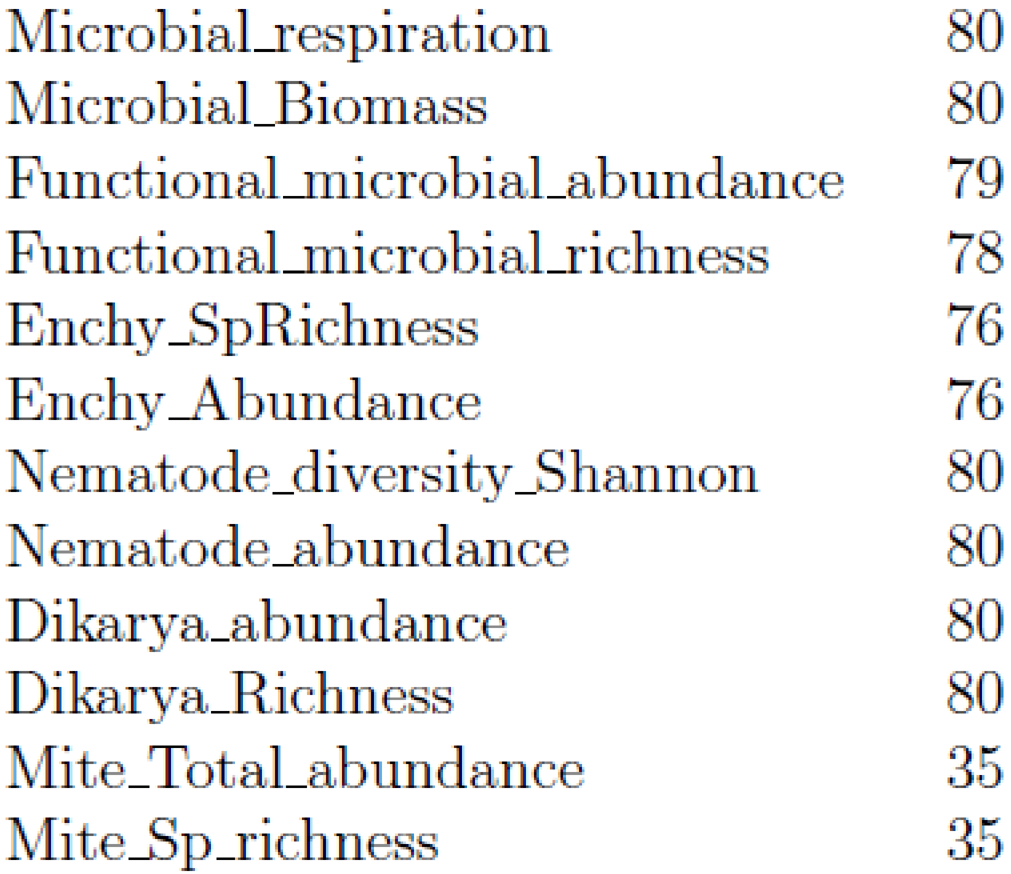
- Lat(itude), Long(itude)
- Bio-Climate Group
- Landuse Group

- Temperature: temp mean, temp min, temp max
- Precipitation: prec mean, prec min, prec max

- WHC (ml 100 g fresh soil-1), pH, Clay %
- CEC (cmol+charge kg-1), Base saturation (%), Mehlich P (mg/L), Org C, Total C, Total N, C:N ratio

ECOFINDERS DATA: Targets

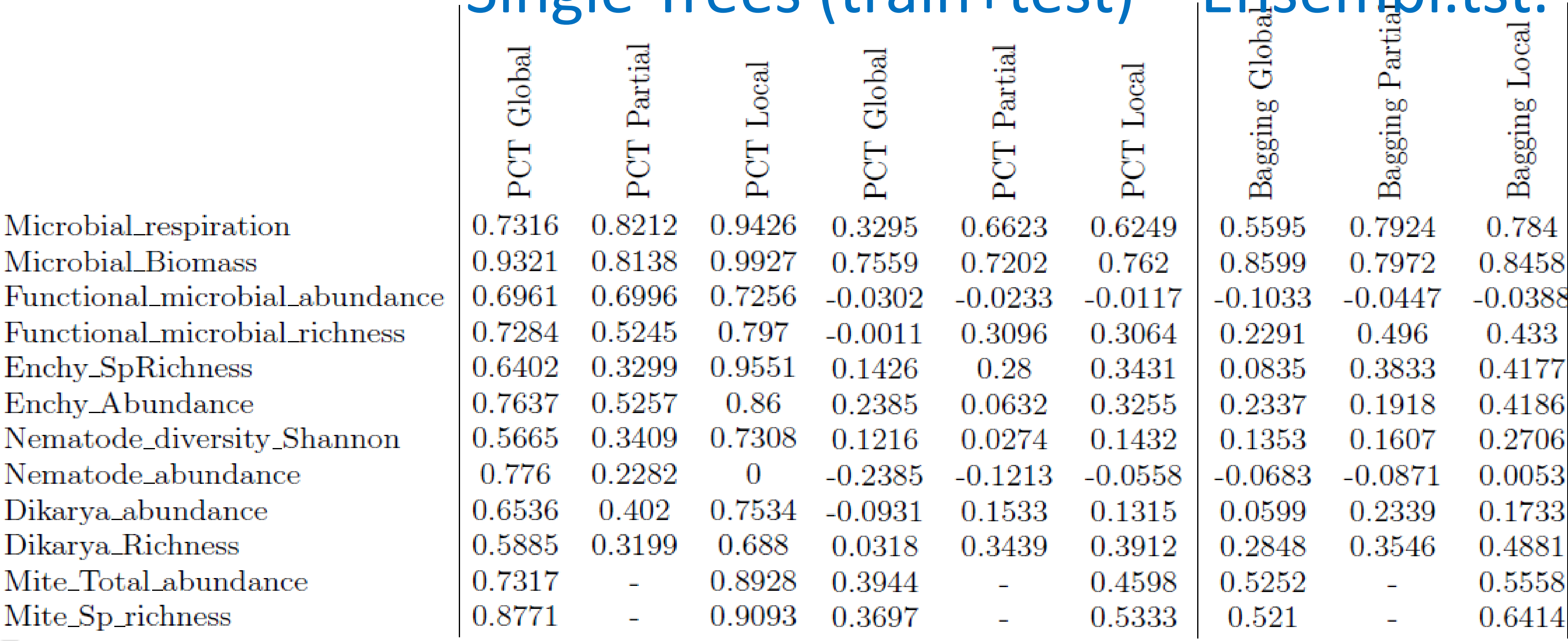
- 12 targets
- Small number of sites/samples
- Only 32 where all measured
- 75 where all but 2 measured (Mites)



Microbial_respiration	80
Microbial_Biomass	80
Functional_microbial_abundance	79
Functional_microbial_richness	78
Enchy_SpRichness	76
Enchy_Abundance	76
Nematode_diversity_Shannon	80
Nematode_abundance	80
Dikarya_abundance	80
Dikarya_Richness	80
Mite_Total_abundance	35
Mite_Sp_richness	35

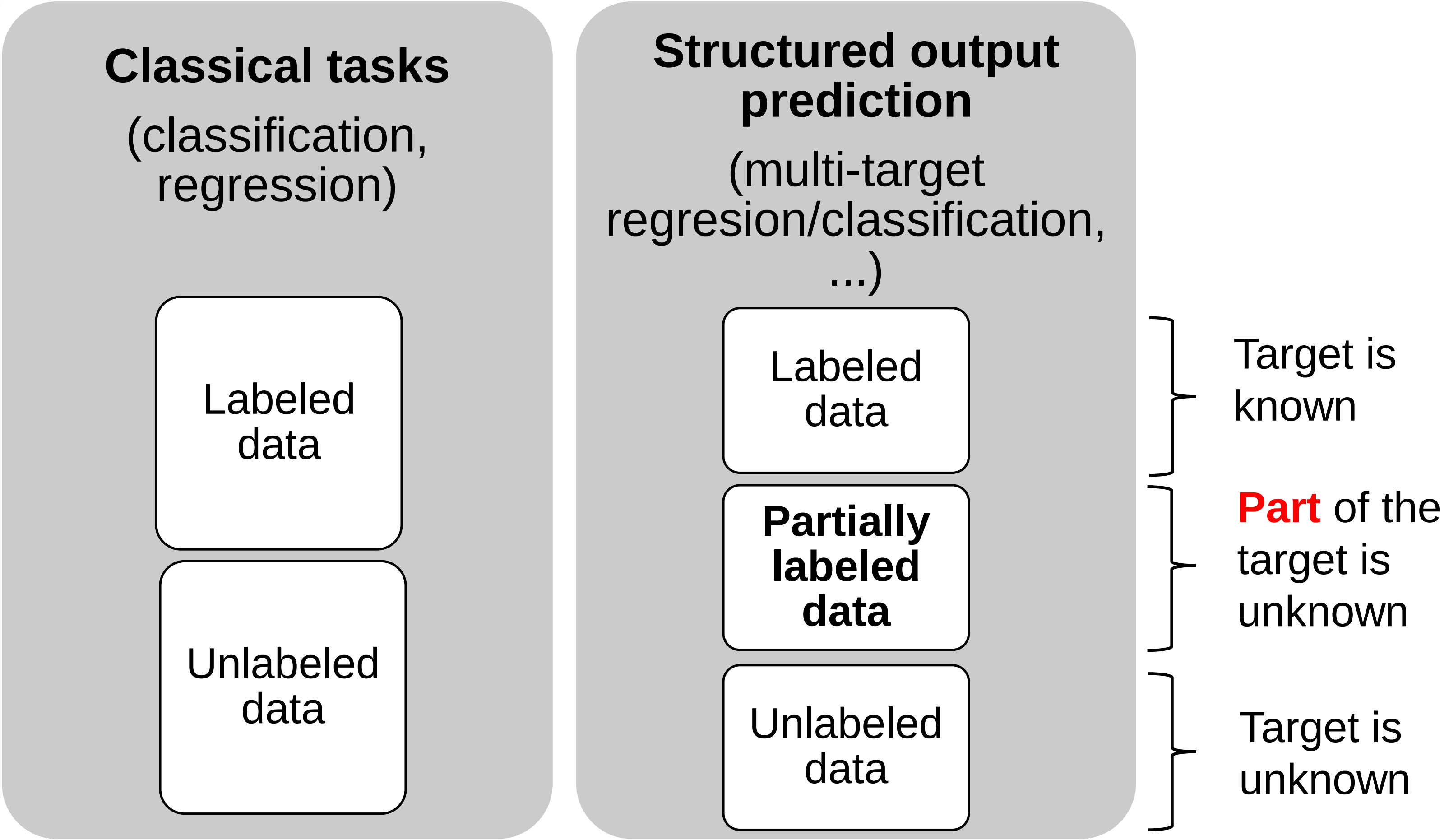
Overall Results: Correlation

- Single trees on training and testing data
- Ensembles on testing data



	Single Trees (train+test)						Ensembl.tst.		
	PCT Global	PCT Partial	PCT Local	PCT Global	PCT Partial	PCT Local	Bagging Global	Bagging Partial	Bagging Local
Microbial_respiration	0.7316	0.8212	0.9426	0.3295	0.6623	0.6249	0.5595	0.7924	0.784
Microbial_Biomass	0.9321	0.8138	0.9927	0.7559	0.7202	0.762	0.8599	0.7972	0.8458
Functional_microbial_abundance	0.6961	0.6996	0.7256	-0.0302	-0.0233	-0.0117	-0.1033	-0.0447	-0.0388
Functional_microbial_richness	0.7284	0.5245	0.797	-0.0011	0.3096	0.3064	0.2291	0.496	0.433
Enchy_SpRichness	0.6402	0.3299	0.9551	0.1426	0.28	0.3431	0.0835	0.3833	0.4177
Enchy_Abundance	0.7637	0.5257	0.86	0.2385	0.0632	0.3255	0.2337	0.1918	0.4186
Nematode_diversity_Shannon	0.5665	0.3409	0.7308	0.1216	0.0274	0.1432	0.1353	0.1607	0.2706
Nematode_abundance	0.776	0.2282	0	-0.2385	-0.1213	-0.0558	-0.0683	-0.0871	0.0053
Dikarya_abundance	0.6536	0.402	0.7534	-0.0931	0.1533	0.1315	0.0599	0.2339	0.1733
Dikarya_Richness	0.5885	0.3199	0.688	0.0318	0.3439	0.3912	0.2848	0.3546	0.4881
Mite_Total_abundance	0.7317	-	0.8928	0.3944	-	0.4598	0.5252	-	0.5558
Mite_Sp_richness	0.8771	-	0.9093	0.3697	-	0.5333	0.521	-	0.6414

Labeled and unlabeled data



Classical tasks

(classification, regression)

Labeled data

Unlabeled data

Structured output prediction

(multi-target regresion/classification, ...)

Labeled data

Partially labeled data

Unlabeled data

Target is known

Part of the target is unknown

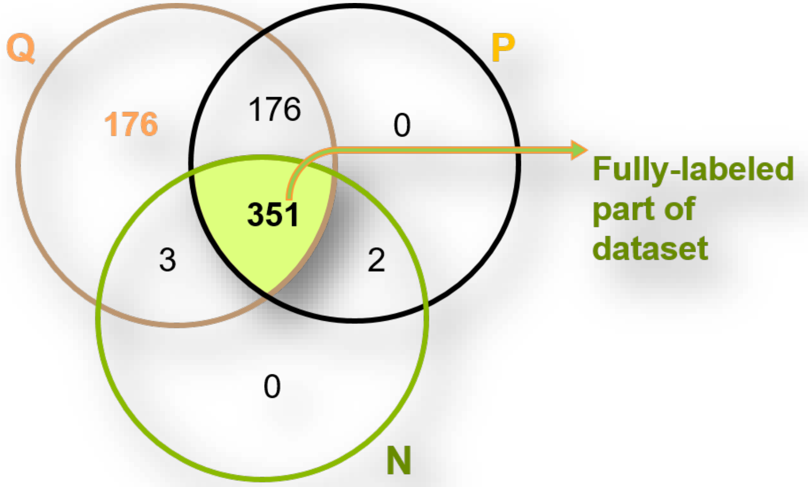
Target is unknown

Incomplete Annotations: Multi-target regression

	Descriptive space				Target space		
Example 1	1	TRUE	0.49	0.69	**?**	0.60	3.91
Example 2	2	FALSE	0.08	0.07	0.56	0.99	7.59
Example 3	1	FALSE	0.08	0.07	**?**	**?**	**?**
Example 4	2	TRUE	0.49	0.69	0.08	0.77	8.86
Example 5	3	TRUE	0.49	0.69	0.11	**?**	**?**
Example 6	4	FALSE	0.08	0.07	0.43	2.10	8.09
…	…				…	…	…

Agricultural Waters in Ireland

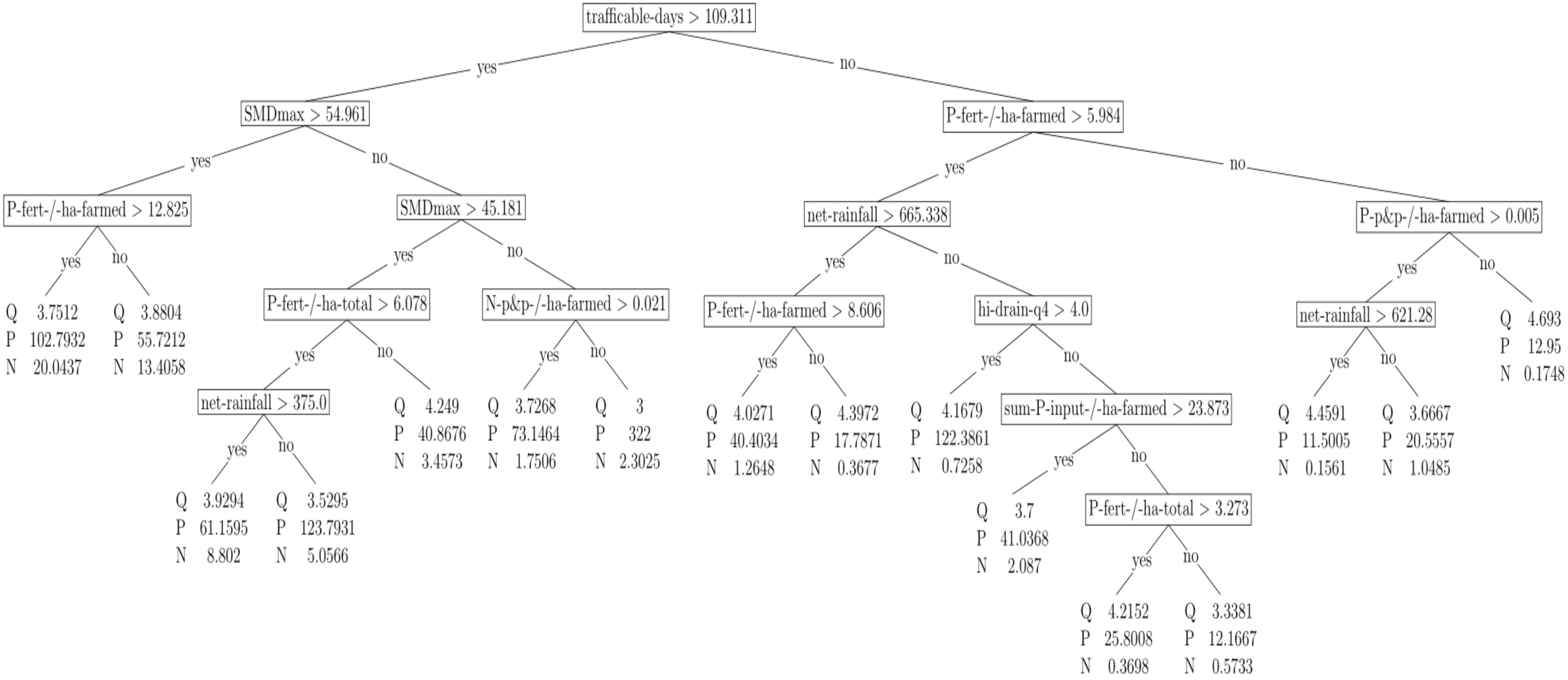
- **Task:** prediction of water quality in agricultural fields in Ireland
- 3 numeric targets
  - **Q -** ***biological water quality***
  - **P -** ***phosphorus-concentration*****;**
  - **N -** ***nitrate concentration*****.**
- 708 examples
  - observation points (10x10km grid cells)
- Not all of the 3 target variables are measured in every observation point -> **missing values!!!**
- 27 numeric attributes:
  - **Environmental pressures** (soil mineralization, drought and grass growing season)
  - **Pathways** (soil drainage, net rainfall, rainfall intensity)



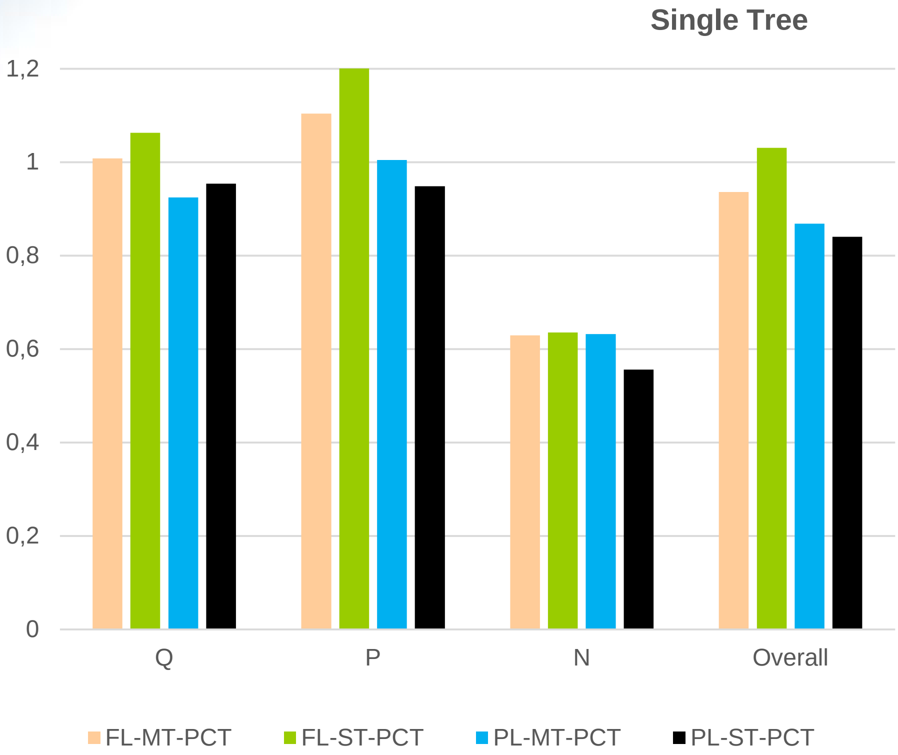
Fully-labeled part of dataset

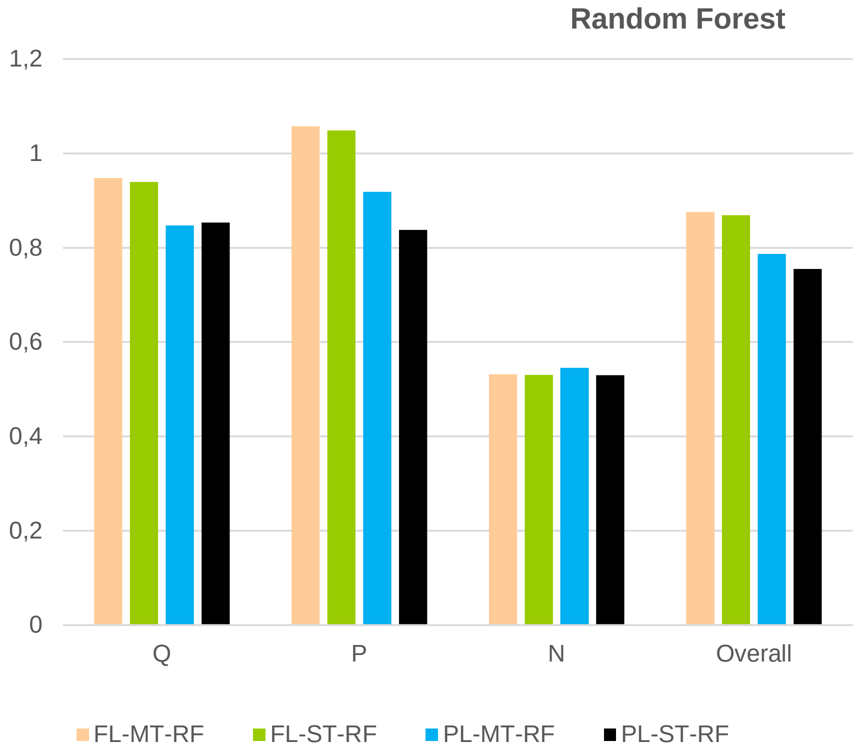
Data set suitable for methods that can handle partially labeled data

Multi-target tree from PL data: QPN



Predictive performance results





Acknowledgements and announcement

We acknowledge European Commission support through the grants

- **MAESTRA: Learning from Massive, Incompletely annotated, and Structured Data, grant 612944**
- HBP SGA1: The Human Brain Project, grant 720270
- LANDMARK: LAND Management: Assessment, Research, Knowledge base, grant 635201

As well as the Slovenian Research Agency through

- P2-0103 Knowledge technologies
- L2-7509 Structured output prediction …

And announce …

Thanks for coming to our Summer School Mining Big& Complex Data 4-8 SEP 2016, Ohrid

ECML PKDD 2017

SKOPJE, MACEDONIA

18-22 September 2017