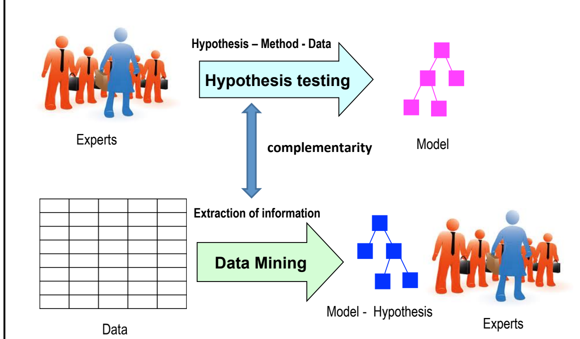
Introduction of data mining to forest ecology and forestry

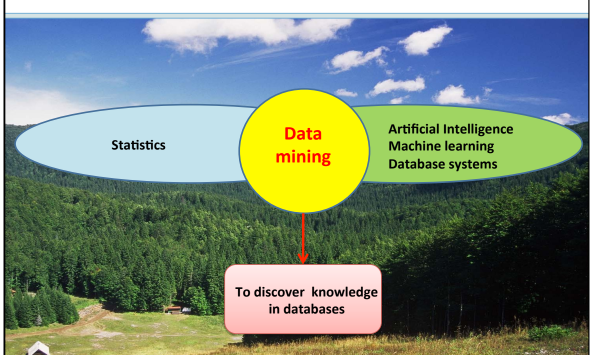
Marko Debeljak

Jožef Stefan Institute, Slovenia

Introduction



Introduction

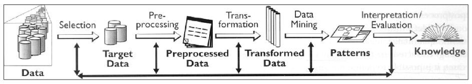


Knowledge discovery in data bases (KDD)

What is KDD?

Frawley et al., 1991: "KDD is the non-trivial **process** of identifying valid, novel, potentially useful, and ultimately understandable **patterns** in data"

The key task is the discovery of **previously unknown knowledge**



How to find patters in data?

Data mining (DM) – **central step** in the KDD process concerned with applying computational techniques to actually find **patterns** in the data (15-25% of the effort of the overall KDD process).

- step 1: **data preprocessing (50%)**
- step 3: **evaluation** of discovered patterns **(25%)**

Data mining and machine learning

Data mining focuses on the discovery of **previously unknown knowledge** and integrates **machine learning**.

Machine learning focuses on **descriptions** and **prediction**, based on known properties **learned** from the training empirical data (examples) using computer algorithms.

Learning from examples is called **inductive learning**

If the goal of **inductive learning** is to obtain model that **predicts** the value of target variable from learning examples, the it is called **predictive or supervised learning.**

Data mining

The most relevant notions of data mining:

1. **Data**
2. **Patterns**
3. **Data mining algorithms**

Data

Data stored in **one flat** table.

Each example represented by a **fixed number** of attributes.

PROPOSITIONAL data mining

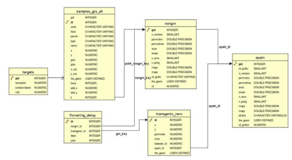
Loos of information due to aggregation

	Properties of objects			
	Distance (m)	Wind direction (^{0})	Wind speed (m/s)	Out-crossing rate (%)
Objects	10	123	3	8
	12	88	4	7
	14	121	6	3
	18	147	2	4
	20	93	1	5
	22	115	3	1
	…	…	…	..

Data stored in **original** tables or relations.

No loss of information, due to aggregation.

RELATIONAL data mining



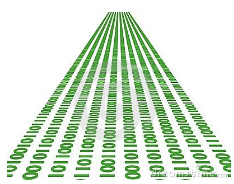
Marko Debeljak

Data

Data are **not stored at all** but they continuously flow through algorithm.

Each example can propositional or relational.

DATA STREAM mining



Marko Debeljak

Pattern

2. What is a pattern?

A pattern is defined as: "A statement (**expression**) in a given language, that describes (**relationships** among) the facts in a **subset** of the given data and is (in some sense) simpler than the enumeration of all facts in the subset" (Frawley et al. 1991, Fayyad et al. 1996).

Classes of patterns considered in data mining

A. **equations,**

B. **decision trees, relational decision trees**

C. association, classification, and regression **rules.**

Selection of the pattern type depends on the data mining task at hand.

Pattern

A. Equations

To predict the value of a **target** (dependent) variable as a **linear or non linear combination** of the **input** (independent) variables:

- **Algebraic equations**

To predict the behavior of **dynamic** systems, which change their rate over time:

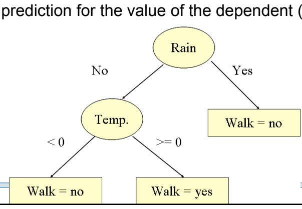
- **Difference equations**
- **Differential equations**

Pattern

B. Decision trees

To predict the value of **one or several target dependent variables** from the **values** of other **independent variables** by **decision tree**.

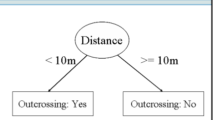
Decision tree has a hierarchical structure, where:
- each internal **node** contains a test on an independent variable,
- each **branch** corresponds to an outcome of the test (critical values of independent variable,
- each **leaf** gives a prediction for the value of the dependent (predicted) variable.



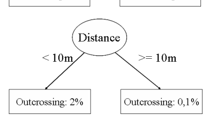
Pattern

Decision tree is called:

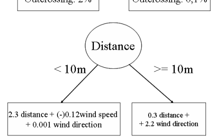
- A **classification** tree: value of dependent variable in leaf is **discrete** (**finite** set of nominal values): e.g., (yes, no), (spec. A, spec. B, …)



- A **regression** tree: value of dependent variable in leaf is a **constant** (**infinite** set of values): e.g.,120, 220, 312, …



- A **model** tree: leaf contains **linear model** predicting the value of piece-wise linear function:
  out-crossing rate= 12.3 distance - 0.123 wind speed + 0.00123 wind direction



Pattern

C. Rules
To perform association analysis between variables discovered by **association rules**.

The **rule denotes** patterns of the form:
IF „Conjunction of conditions“ **THEN** „Conclusion.”

- For **classification rules**, the conclusion assigns one of the possible **discrete** values to the dependent variable (finite set of nominal values): e.g., (yes, no), (spec. A, spec. B, spec. D)
- For **predictive rules**, the conclusion gives a prediction for the value of the dependent variable (**infinite** set of values): e.g.,120, 220, 312, …

Algorithm

3. What is data mining algorithm?
Algorithm in general:
- a **procedure** (a finite set of well-defined instructions) for accomplishing some task which will terminate in a defined end-stat.

Data mining algorithm:
- a computational process for finding patterns in data

Data mining (DM) - algorithm

Selection of algorithm depends on problem at hand:

1. **Equations = Linear and multiple regressions, equation discovery**
2. **Decision trees = Top/down induction of decision trees**
3. **Rules = Rule induction**

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES Jožef Stefan Institute

Institut "Jožef Stefan" Ljubljana, Slovenija

DATA MINING – CASE STUDIES

Applications – forest ecology and forestry

POPULATION DYNAMICS

HABITAT MODELLING

GENE FLOW MODELLING

RISK MODELLING

Applications – forest ecology and forestry

Propositional and relational supervised data mining:

- **Simple data mining**
- **Data mining of time series**
- **Spatial data mining**

1. Equations:

- **Algebraic equations**
- **Differential equations**

2. Single and multi target decision trees:

- **Classification trees**
- **Regression trees**
- **Model trees (single target only)**

Algebraic equations: POPULATION DYNAMICS

Algebraic equations: CIPER

ECOLOGICAL MODELLING 215 (2008) 180–189

available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/ecolmodel

ELSEVIER

Modeling radial growth increment of black alder (*Alnus glutionsa* (L.) Gaertn.) tree

Jana Laganis[a,*], *Aleksandar Pečkov*[b], *Marko Debeljak*[b]

[a] Laboratory for Environmental Research, University of Nova Gorica, Vipavska 13, Nova Gorica, Slovenia

[b] Department of Knowledge Technologies, "Jozef Stefan" Institute, Jamova 39, Ljubljana, Slovenia

Algebraic equations: POPULATION DYNAMICS

Dataset

Measured **radial increments**:
- 8 trees
- 69 years old

Hydrological conditions
(*HMS Lendava*; monthly data on minimal, average and maximum values)
- Ledava River levels
- groundwater levels

Management data
(thinning; m³/y removed from the stand; *Forestry Unit Lendava*)

Meteorological conditions
(monthly data, HMS Lendava):
- Time of solar radiation (h),
- precipitation (mm),
- ET (mm)
- Number of days with white frost
- Number of days with snow
- T: max, aver, min
- Cumulative T>0ºC, >5ºC, and >10ºC
- Number of days with:
  - minT>0ºC
  - minT<-10ºC
  - minT<-4ºC
  - minT>25ºC
  - maxT>10ºC
  - maxT>25ºC

- Monthly data + aggregated data (AMJ, MJJ, JJA, MJJA etc.)
- **Σ**: **333** attributes; 35 years

Algebraic equations: POPULATION DYNAMICS

- 52 different combinations of attributes were tested.

Σ: 124 models

Experiment	**RRSE**	**# eq. elements**
jnj3_2m	0,7282	6
jnj3_3s	0,7599	6
jnj3_1s	0,7614	6
jnj3_4m	0,76455	3
jnj2_2	0,7685	5
jly_4xl	0,7686	6

Algebraic equations: POPULATION DYNAMICS

Model jnj3_2m:

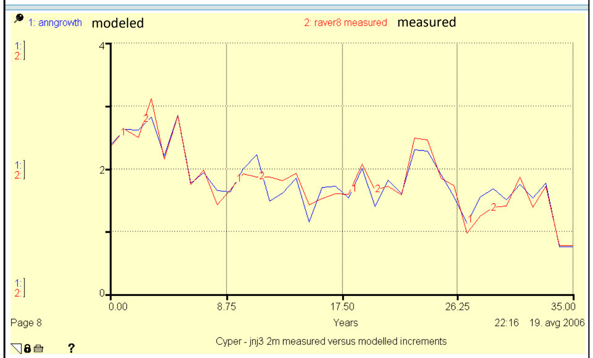
RadialGrowthIncrement =
+ -0.0511025526922 minL8-10^1
+ -0.0291795197998 maxL8-10^1
+ -0.017479975134 t-sun4-7^1
+ 0.0346935385853 t-sun8-10^1
+ -1.950606536e-05 t-sun8-10^2
+ -2.01014710248 d-wf-4-7^1
+ 9.35586778387e-05 minL4-7^1 t-sun4-7^1
+ -0.000179339939732 minL4-7^1 t-sun8-10^1
+ 6.45688563611e-05 minL8-10^1 t-sun8-10^1
+ 3.06551434164e-05 maxL8-10^1 t-sun4-7^1
+ 0.00282485442386 t-sun4-7^1 d-wf-4-7^1
+ -0.00141078675225 t-sun8-10^1 d-wf-4-7^1
+ 7.91071710872

Relative Root Squared Error
= 0.728229824611

Correlation between average measured (r-aver8) and modeled increments: linear regression:
R^2 = 0.8771

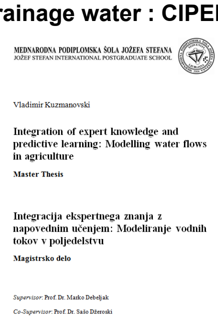
8 out of 333 attributes

Algebraic equations: POPULATION DYNAMICS



Algebraic equations: WATER CYCLE

Prediction of drainage water : CIPER


MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA
JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL

Vladimir Kuzmanovski

Integration of expert knowledge and predictive learning: Modelling water flows in agriculture

Master Thesis

Integracija ekspertnega znanja z napovednim učenjem: Modeliranje vodnih tokov v poljedelstvu

Magistrsko delo

Supervisor: Prof. Dr. Marko Debeljak

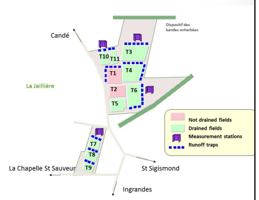
Co-Supervisor: Prof. Dr. Sašo Džeroski

Ljubljana, Slovenia, September 2012

PCQE Database

- Experimental site La Jaillière
- Western France
- Owned by ARVALIS

- Shallow silt clay soils
- 11 fields are observed
- Field size about 0.3 - 1 ha



PCQE Database *(continued)*

- Agricultural practices
  - Fertilization
  - Irrigation
  - Phytochemical protection
  - Harvesting
  - Tillage
- Slope
- Water flow
  - Drainage
  - Runoff

- 25 campaigns (1987 - 2011)
- Campaign is defined as period starting from 01.09 and finishing on 31.08, following year

DRAINAGE predictive model - CIPER

- Polynomial equations induced on data for a whole campaign - CIPER algorithm
- Evaluation
  - "Leave one out" approach

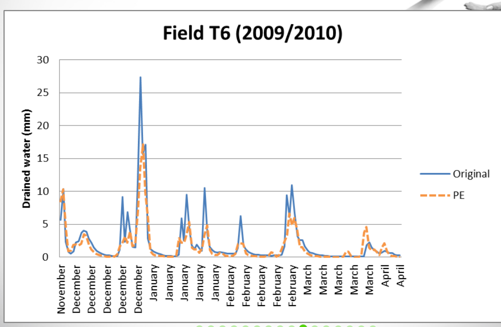
Fields	Test field	Std. Dev.	RMSE	RRSE	Corr. coeff. (r)
All	T3	3.187	2.1119	66.26 %	0.7855
All	T4	3.188	1.7220	54.01 %	0.8273
All	T5	3.163	2.2478	71.06 %	0.7467
All	T6	3.096	2.1784	70.36 %	0.8096
All	T7	3.229	1.3286	41.15 %	0.7812
All	T8	3.210	1.3813	43.03 %	0.7839
All	T9	3.210	1.5927	49.62 %	0.7434
All	T10	3.130	1.6672	53.26 %	0.7766
All	T11	3.108	1.6274	52.36 %	0.7841
T3	T6	3.056	2.1251	69.54 %	0.8125
T6	T3	3.600	2.0961	58.22 %	0.7745

Predictive models

Model (All/T4)

$$
\begin{aligned}
Drainage = {} & 0.0196445 * RainfallA1 * Temp \\
& + 0.33246 * DrainageN1 \\
& + 0.000662861 * CDCoef^{2} * RainfallA1^{2} * Slope \\
& + 0.00000107253 * Runoff * DrainageN1 * Temp^{2} * RainfallA1^{2} \\
& - 0.00115983 * Runoff^{2} * DrainageN1 * Slope^{3} \\
& - 0.00114057 * Temp^{2} * RainfallA1 \\
& + 0.00153725 * RainfallA1^{2} \\
& + 1.63563 * Runoff * Slope \\
& - 1.90622 * Runoff \\
& - 0.0231748 * Slope^{2} * RainfallA1 \\
& + 0.0654042 * RainfallA1 * Slope \\
& - 0.00755737 * Slope^{3} * CDCoef^{3} * Runoff^{2} \\
& + 0.0675951 * Slope \\
& - 0.146702
\end{aligned}
$$

Predictive models



Algebraic equations: GENE FLOW

Algebraic equations: Lagramge

Ecological Modelling 220 (2009) 1063–1072

Contents lists available at ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel

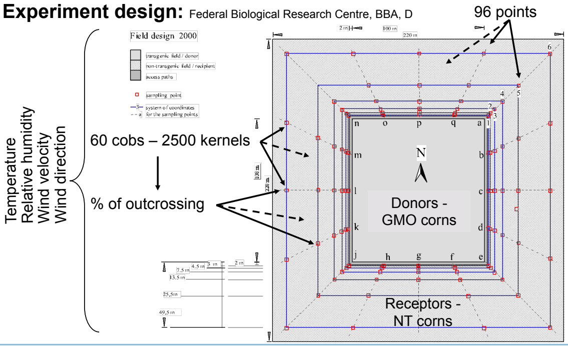
Modelling the outcrossing between genetically modified and conventional maize with equation discovery

Aneta Ivanovska[a,*], Ljupčo Todorovski[b], Marko Debeljak[a], Sašo Džeroski[a]

[a] Jožef Stefan Institute, Department of Knowledge Technologies, Jamova cesta 39, SI-1000 Ljubljana, Slovenia
[b] University of Ljubljana, Faculty of Administration, Gosarjeva 5, SI-1000 Ljubljana, Slovenia

Algebraic equations: GENE FLOW

Experiment design: Federal Biological Research Centre, BBA, D



Algebraic equations: GENE FLOW

Table 5
The context free grammar used by Lagramge to model the outcrossing between GM and non-GM maize. The grammar specifies the space of possible equations to be considered by Lagramge.

Outcrossing → const × (Distance Influence$^{\alpha}$) × (Wind Influence$^{\beta}$)
Distance Influence → 1
Distance Influence → F
Distance Influence → F
$F \to e^{-\text{Distance}}$
F → 1/Distance
F → 1/Distance^2
F → Distance$^{-\gamma}$; $(0 \leq \gamma \leq 1000)$
Distance → variable_minDistance
Distance → variable_distanceCenter
Wind Influence → 1
Wind Influence → PWind
PWind → (PWind) × Wind + const|const
Wind → variable_appropriateWindProc
Wind → variable_windTunnelLength

Algebraic equations: GENE FLOW

Table 8
Correlation coefficients (r), relative mean squared error (reMSEs) and best equations of the experiments carried out on BBA2000, BBA2001, KIS2006, all BBA, and all three datasets.

	Correlation coefficient (reMSE)	Best equation
BBA2000	0.89 (0.50)	$\text{Outcrossing} = \frac{0.02}{\text{minDistance}^{1.8}} \times [0.007 \times \text{windTunnelLength}^2 \times \text{appropriateWindProc} + 602.93]$
BBA2001	0.68 (0.90)	$\text{Outcrossing} = \frac{0.01}{\text{distanceCenter} \times \text{minDistance}^2} \times [\text{windTunnelLength}^3 + \text{windTunnelLength}^2 + \text{windTunnelLength} + 1]$
KIS2006	0.83 (0.33)	$\text{Outcrossing} = \frac{531.12}{\text{distanceCenter} \times e^{\text{minDistance}}}$
BBA2000 + 2001	0.86 (0.48)	$\text{Outcrossing} = \frac{0.01}{\text{distanceCenter} \times \text{minDistance}^2} \times [\text{appropriateWindProc} \times \text{windTunnelLength}^2 + \text{windTunnelLength}^2 + \text{windTunnelLength} + 1]$
ALL	0.64 (1.52)	$\text{Outcrossing} = \frac{0.01}{\text{distanceCenter} \times \text{minDistance}^{0.1}} \times [\text{appropriateWindProc}^2 + \text{appropriateWindProc} + 1]$

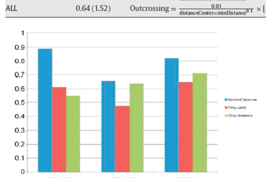


Fig. 5. Comparison of the correlation coefficients for the equations learned for each dataset, when using only distance variables, only wind variables and using all the variables (distance and wind influence).

Table 9
Predictive performance of the models learned on data from one region and tested on data from other region.

Train	Test	Correlation coefficient
BBA2000 + 2001	KIS	0.77
KIS	BBA2000 + 2001	0.63

Differential equations: COMMUNITY STRUCTURE

Algebraic equations: Lagramge

ECOLOGICAL MODELLING 194 (2006) 37–48

ELSEVIER

available at www.sciencedirect.com

SCIENCE DIRECT

journal homepage: www.elsevier.com/locate/ecolmodel

Automated modelling of a food web in lake Bled using measured data and a library of domain knowledge

Nataša Atanasova[a,*], Ljupčo Todorovski[b], Sašo Džeroski[b], Špela Rekar Remec[c], Friedrich Recknagel[d], Boris Kompare[a]

[a] Faculty of Civil and Geodetic Engineering, University of Ljubljana, Slovenia
[b] Jožef Stefan Institute, Slovenia
[c] Environmental Agency of the Republic of Slovenia, Slovenia
[d] University of Adelaide, Australia

Differential equations: COMMUNITY STRUCTURE

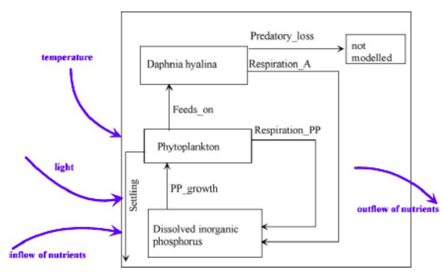


Fig. 2 – Simple conceptual model for lake Bled.

Differential equations: COMMUNITY STRUCTURE

Table 2 – Measured data (variables) in lake Bled used for model induction

Variable name	Description
q_krivica (m^3/day)	Inflow to the lake
q_misca (m^3/day)	Inflow to the lake
q_radovna (m^3/day)	Inflow to the lake
q_jezernica (m^3/day)	Outflow (at surface)
q_natega (m^3/day)	Outflow (syphon)
ps_krivica, ps_misca, ps_radovna (mg/l)	Nutrient (orthophosphate) concentration in the inflows
temp (°C)	Water temperature of the streams and lake
light (J/(cm^2 day))	Calculated underwater light
ps, no, silica (mg/l)	Inorganic nutrients' concentration in the lake (ps is soluble phosphorus and NO is nitrate)
phyto (mgDW/l)	Phytoplankton biomass concentration in the lake
daph (No. ind/ml or mgDW/l (see text))	Zooplankton (*Daphnia hyalina*) biomass concentration in the lake

Table 4 – Description of the variables influences on the phytoplankton dynamics equations induced on 1-year data sets (1995–2002)

	Process/term							
	Growth					Respiration	Settling	Grazing
	Temp[a]	Light[a]	ps[a]	Silica[a]	no[a]	temp[a]	temp[a]	
1995	exp	mon	no	mon	no	lin	lin	No
1996	exp	inh	mon	no	mon	no	no	No
1997	lin	no	mon	mon	mon	exp	no	No
1998	lin	no	exp	mon	exp	exp	no	Yes
1999	exp	inh	mon	mon2	mon	lin	exp	Yes
2000	exp	inh	mon	exp	mon	lin	no	Yes
2001	exp	inh	mon2	exp	no	lin	lin	No
2002	lin	mon	no	mon2	mon2	lin	no	Yes

The influence is described using the following labels: no denotes no influence, yes denotes presence of influence and other labels (exp, lin, mon, mon2) denote the specific influence model (exponential, linear, Monod, second-order Monod, respectively).

Differential equations: COMMUNITY STRUCTURE

Phosphorus

$$\frac{d(ps)}{dt} = ps\_krivica \cdot \frac{q\_krivica}{7 \cdot 10^6} + ps\_misca \cdot \frac{q\_misca}{7 \cdot 10^6} + ps\_radovna \cdot \frac{q\_radovna}{7 \cdot 10^6} - ps \cdot \frac{q\_jezernica}{7 \cdot 10^6} - ps \cdot \frac{q\_natega}{7 \cdot 10^6} + 0.0022 \cdot phyto^2 \cdot 0.072 \cdot \frac{temp - 2.7}{20.4 - 2.7} + 0.07 \cdot daph \cdot 0.0026 \cdot \frac{temp}{12.3} - 0.0023 \cdot phyto \cdot 0.21 \cdot \frac{ps}{ps + 0.00042} \cdot \frac{temp}{16.7} \cdot \frac{light}{170} \cdot e^{\left(1 - \frac{light}{170}\right)} \quad (10)$$

water in-flow
out-flow

respiration

growth

Differential equations: COMMUNITY STRUCTURE

Phytoplankton

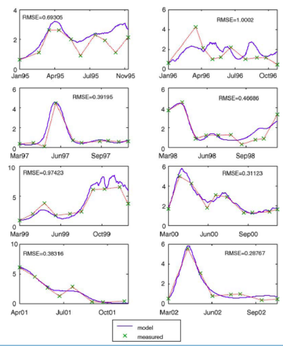
$$\frac{d(phyto)}{dt} = phyto \cdot 0.21 \cdot \frac{ps}{ps + 0.00042} \cdot \frac{temp}{16.7} \cdot \frac{light}{170} \cdot e^{\left(1 - \frac{light}{170}\right)} - phyto^2 \cdot 0.072 \frac{temp - 2.7}{19.7 - 2} - phyto \cdot \frac{0.5}{10} \cdot \frac{temp - 2}{18 - 4} - daph \cdot 0.5 \cdot \frac{temp - 2.6}{18 - 4} \cdot (1 - \exp(-0.58 \cdot phyto)) \cdot 0.56 \cdot phyto \quad (13)$$

growth

respiration

sedimentation

grazing



Differential equations: COMMUNITY STRUCTURE

Zooplankton

$$\frac{d(daph)}{dt} = 0.14 \cdot daph \cdot 0.5 \cdot \frac{temp - 2.6}{18 - 4} \cdot (1 - \exp(-0.58 \cdot phyto)) \cdot 0.56 \cdot phyto - daph \cdot 0.026 \cdot \frac{temp}{12.3} - 0.01 \cdot \frac{daph^2}{0.001 + daph}$$

Feeds on phytoplankton

respiration

mortality

Differential equations: COMMUNITY STRUCTURE

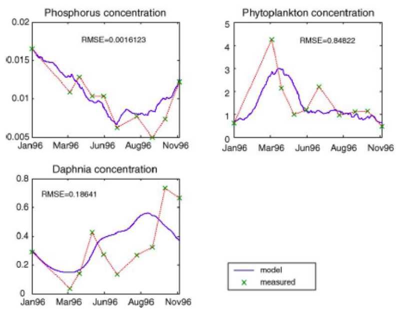


Fig. 6 – Performance of the food web model for phosphorus, phytoplankton and daphnia.

Decision trees: HABITAT MODELS

Classification trees: J48

ELSEVIER
Available online at www.sciencedirect.com
SCIENCE @ DIRECT®
Ecological Modelling 170 (2003) 453–469
ECOLOGICAL MODELLING
www.elsevier.com/locate/ecolmodel

Modeling the brown bear population in Slovenia
A tool in the conservation management of a threatened species

Klemen Jerina[a,*], Marko Debeljak[b], Sašo Džeroski[c], Andrej Kobler[d], Miha Adamič[a]
[a] Department of Forestry, Biotehnical Faculty, University of Ljubljana, Večna pot 83, 1000 Ljubljana, Slovenia
[b] Nova Gorica Polytechnic, School of Environmental Sciences, Vipavska 13, 5000 Nova Gorica, Slovenia
[c] Jozef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
[d] Slovenian Forestry Institute, Večna pot 2, 1000 Ljubljana, Slovenia

Decision trees: HABITAT MODELS

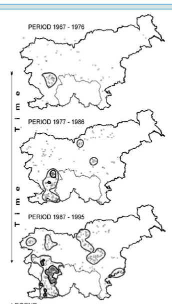
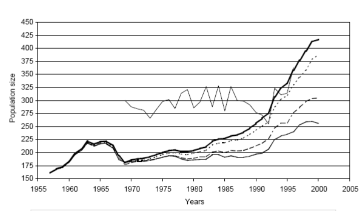


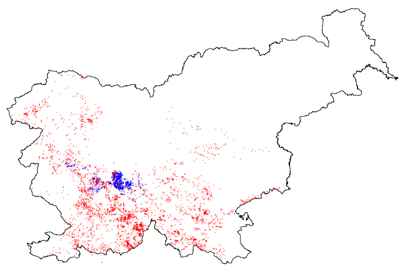
Fig. 3. The spatial expansion of the brown bear in Slovenia in the period 1967–1995.

Population growth



Decision trees: *HABITAT MODELS*

Observed locations of BBs



Decision trees: *HABITAT MODELS*

The training dataset

- **Positive examples:**
  - Locations of bear sightings (Hunting association; telemetry)
  - Females only
  - Using home-range (HR) areas instead of "raw" locations
  - Narrower HR for optimal habitat, wider for maximal

- **Negative examples:**
  - Sampled from the unsuitable part of the study area
  - Stratified random sampling
  - Different land cover types equally accounted for

Decision trees: *HABITAT MODELS*

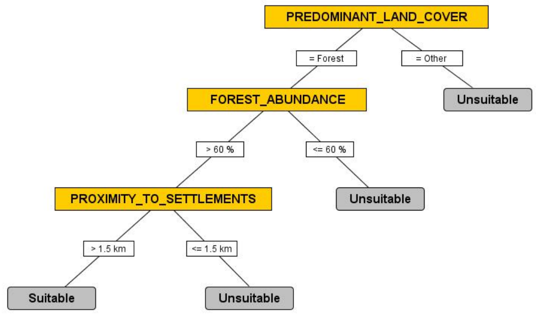
Dataset

1,73,26,0,0,1,88,0,2,70,7,20,1,0,0,1,0,60,0,0,0,0,0,2,**0**,0,0,4123,0,0,0,0,63,211,11,11,11,83,213,213,0,0,4155.
1,62,37,0,0,2,88,0,2,70,7,20,1,0,0,1,0,60,0,0,0,0,0,2,**1**,53,0,3640,0,0,0,-1347,63,211,11,11,11,83,213,213,11,89,3858.
2,0,99,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,6,82,**0**,10404,0,2074,-309,48,0,0,11,11,11,83,83,83,0,20,3862.
2,0,100,0,0,1,76,0,16,71,0,12,0,0,0,0,0,0,0,0,0,0,0,**1**,6,82,0,7500,0,1661,-319,-942,0,0,11,11,11,0,0,0,0,20,4088.
1,8,91,0,0,1,52,0,59,41,0,0,0,0,0,0,0,4,0,0,0,0,5,**1**,6,82,0,6500,0,1505,-166,879,9,57,11,11,11,281,281,281,0,20,3199.
4,3,0,86,9,0,75,0,33,67,0,0,0,0,0,0,1,2,0,0,0,0,0,0,**1**,2,54,0,0,0,465,-66,-191,4,225,11,31,31,41,72,272,60,619,4013.
1,34,65,0,0,2,51,9,76,9,5,1,4,1,0,1,0,29,0,0,0,0,0,0,**1**,2,54,0,3000,0,841,-111,-264,34,220,11,41,41,151,141,112,60,619,3897.
1,100,0,0,0,3,52,0,86,6,3,5,9,6,7,38,40,0,0,0,0,0,0,**0**,1,17,64,0,8062,0,932,-603,-71,100,337,11,41,41,171,232,202,4,24,3732.
....
....
....

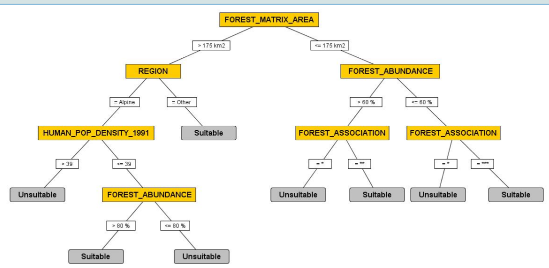
Present: 1

Absent: 0

The model for maximal habitat



The model for optimal habitat



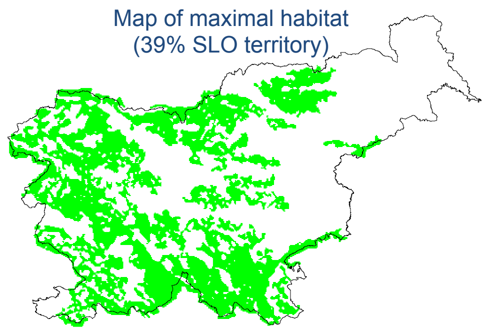
* QUERICO ROBORI - CARPINETUM, CARICI ELATE - ALNETUM GLUTINOSAE, CARICI BRIZOIDI - ALNETUM GLUTINOSAE, ALNETUM GLUTINOSO - INCANAE, ALNETUM INCANAE, SALICI - POPULETUM, SALICETUM GR., QUERCO - CARPINETUM VAR. HACQUETIA, QUERCO - CARPINETUM VAR. LUZULA, etc ...

** MELAMPYRO VULGATI - QUERCETUM, ADENOSTYLO - FAGETUM, FAGETUM SUBALPINUM, QUERCO - FAGETUM, ABIETI - FAGETUM DINARICUM

*** ADENOSTYLO - FAGETUM, FAGETUM SUBALPINUM

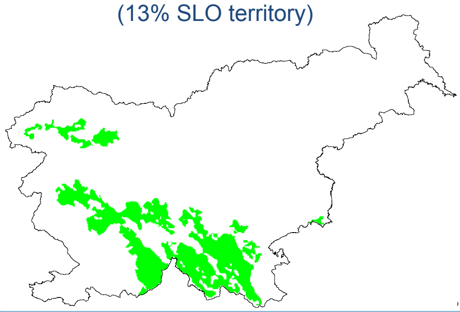
Decision trees: HABITAT MODELS

Map of maximal habitat
(39% SLO territory)



Decision trees: HABITAT MODELS

Map of optimal habitat
(13% SLO territory)



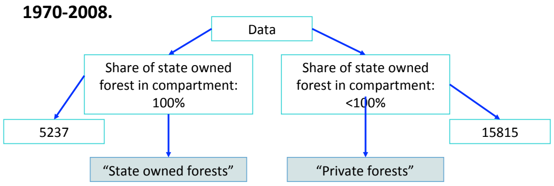
Multi-target decision trees: POPULATION DYNAMICS

Predictive models of forest development in Slovenia

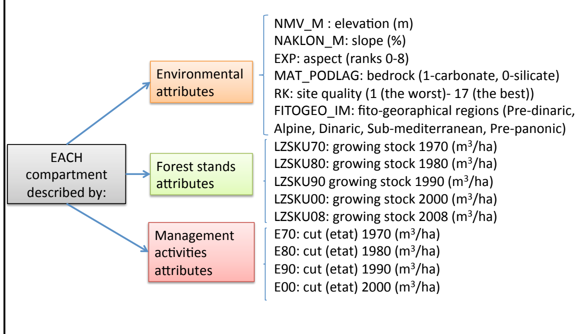
Data

Database *Silva* 1970-2008:

- data from years 1970, 1980, 1990, 2000 and 2008,
- data unit is a **permanent compartment**
- data from **21052 permanent compartments for the period 1970-2008.**

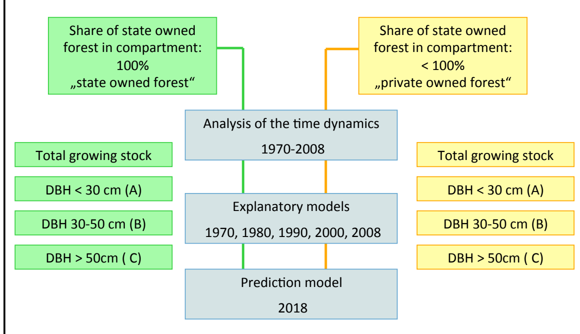


Data



Marko Debeljak

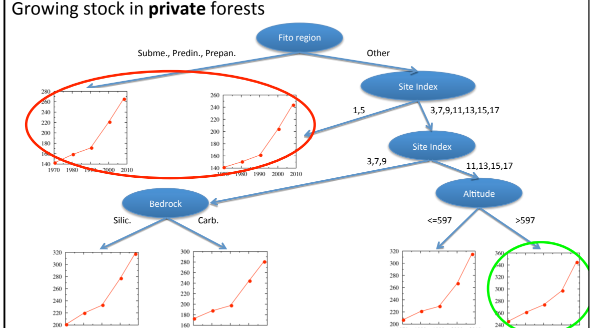
Experimental design



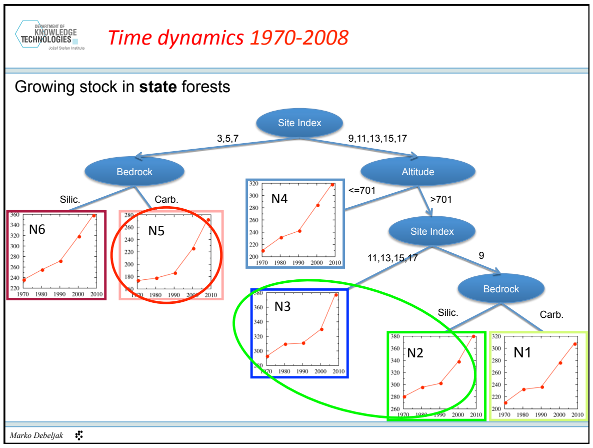
Marko Debeljak

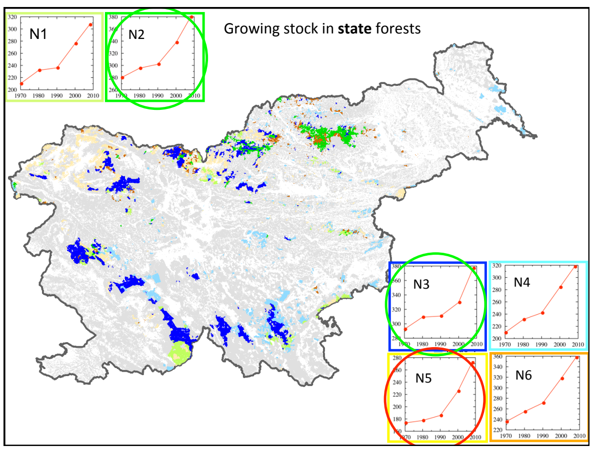
Time dynamics 1970-2008

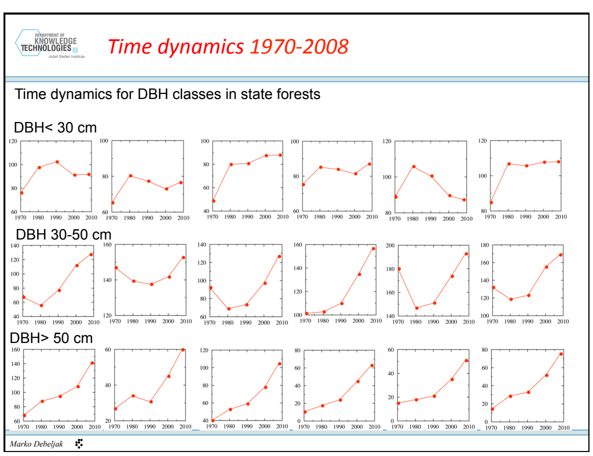
Growing stock in **private** forests



Marko Debeljak

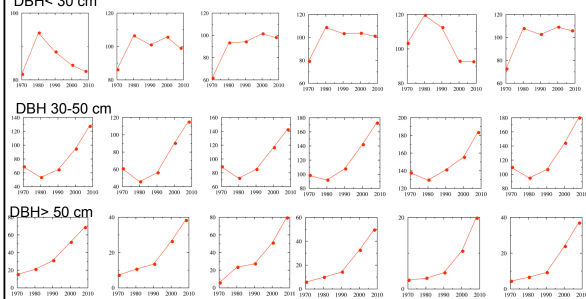

Time dynamics 1970-2008
Growing stock in **state** forests
Site Index
3,5,7
9,11,13,15,17
Bedrock
Altitude
Silic.
Carb.
<=701
>701
Site Index
11,13,15,17
9
Bedrock
Silic.
Carb.
N6
N5
N4
N3
N2
N1
Marko Debeljak


Growing stock in **state** forests
N1
N2
N3
N4
N5
N6


Time dynamics 1970-2008
Time dynamics for DBH classes in state forests
DBH< 30 cm
DBH 30-50 cm
DBH> 50 cm
Marko Debeljak

Time dynamics 1970-2008

Time dynamics for DBH classes in private forests



Marko Debeljak

*Prediction model for year 2018 – **MODEL decision trees***

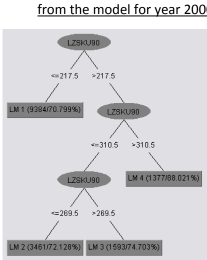
Step 1: Verification of methodology on prediction for 2008:
- Extrapolation of the linear trends for model 2000 on 2008
- Verification on real data for 2008

Step 2: Prediction of growing stock for 2018:
- Extrapolation of linear trend of the model for 2008 to the year 2018

Marko Debeljak

Prediction for private forests in 2018

Step 1: Verification of methodology on year 2008 using extrapolation of linear trend from the model for year 2000



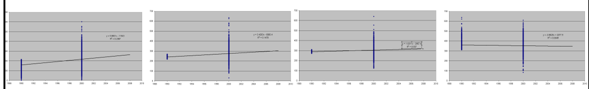

```
LM 1:
LZSKU00 =
-0.1233 * NAKLON_M
    + 9.3081 * MAT_PODLAG=0
    + 20.5809 * RK=13,9,11,3,7,15,17
    + 4.0773 * RK=11,3,7,15,17
    - 2.7453 * RK=3,7,15,17
    - 0.0116 * RK=7,15,17
    + 6.7562 * FITOGEO_IM=Predpanon,Dinarsko,Predalpsko,Alpsko
    - 6.685 * FITOGEO_IM=Dinarsko,Predalpsko,Alpsko
    + 17.1923 * FITOGEO_IM=Predalpsko,Alpsko
    + 0.0249 * LZSKU70
    + 0.0001 * LZSKU80
    + 0.5976 * LZSKU90
    + 87.0434

LM 2:
LZSKU00 =
0.0002 * NAKLON_M
    + 11.9402 * MAT_PODLAG=0
    + 0.0405 * RK=13,9,11,3,7,15,17
    + 0.0132 * RK=11,3,7,15,17
    + 0.0088 * RK=3,7,15,17
    + 0.0091 * RK=7,15,17
    + 0.0099 * FITOGEO_IM=Predpanon,Dinarsko,Predalpsko,Alpsko
    - 0.0047 * FITOGEO_IM=Dinarsko,Predalpsko,Alpsko
    + 11.7842 * FITOGEO_IM=Predalpsko,Alpsko
    + 0.0284 * LZSKU70
    + 0.0007 * LZSKU80
    + 0.5109 * LZSKU90
    + 130.2721

LM 3:
LZSKU08 =
0.0002 * NAKLON_M
    + 0.1377 * MAT_PODLAG=0
    + 0.0405 * RK=13,9,11,3,7,15,17
    + 0.0132 * RK=11,3,7,15,17
    + 0.0088 * RK=3,7,15,17
    + 0.0281 * RK=7,15,17
    + 0.0099 * FITOGEO_IM=Predpanon,Dinarsko,Predalpsko,Alpsko
    - 0.0047 * FITOGEO_IM=Dinarsko,Predalpsko,Alpsko
    + 0.0822 * FITOGEO_IM=Predalpsko,Alpsko
    + 0.0005 * LZSKU70
    + 0.1804 * LZSKU80
    + 0.0074 * LZSKU90
    + 254.9865

LM 4:
LZSKU00 =
0.0105 * NMV_M
    + 0.0002 * NAKLON_M
    + 0.1034 * MAT_PODLAG=0
    + 0.0405 * RK=13,9,11,3,7,15,17
    + 0.0132 * RK=11,3,7,15,17
    + 0.0088 * RK=3,7,15,17
    + 0.0185 * RK=7,15,17
    + 0.0099 * FITOGEO_IM=Predpanon,Dinarsko,Predalpsko,Alpsko
    - 0.0047 * FITOGEO_IM=Dinarsko,Predalpsko,Alpsko
    + 0.0212 * FITOGEO_IM=Predalpsko,Alpsko
    + 0.0006 * LZSKU70
    + 0.162 * LZSKU80
    + 0.4448 * LZSKU90
    + 132.4082
```

2000	LM1	LM2	LM3	LM4
Average real growing stock	217.1	275.1	304.9	352.3
Average model growing stock	253.4	350.9	304.8	352.3
Mean absolute error (MAE)	61.0	93.4	43.3	50.9

Marko Debeljak

Prediction for private forests in 2018

Step 1: - Extrapolation of linear trends from 1990-2000 to 2008
- Verification on real data for 2008

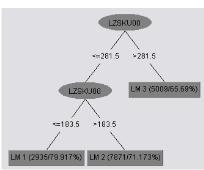


2008	LM1	LM2	LM3	LM4
Linear regression a (k, n)	5.8801, - 11543	3.4203, - 6565.4	1.6311, - 2957.4	-0.8628, 2077.9
Average real growing stock 2008	264.2	312.1	339.5	380.5
Model prediction for 2008	264.2	302.6	317.8	345.4
Difference (%)	0.0	-3.0	-6.4	-9.2

Marko Debeljak

Prediction for private forests in 2018

Step 2: - Prediction of wood stock for 2018 with the model for 2008



LM 1:
LZSKU08 =
-0.0387 * NMV_M
+ 0.0845 * RK=11,3,7,15,17
+ 0.0795 * FITOGEO_IM=Predalpsko,Alpsko
+ 0.0001 * LZSKU70
+ 0.2302 * LZSKU90
+ 0.6201 * LZSKU00
+ 120.3242

LM 2:
LZSKU08 =
0 * NMV_M
+ 14.6206 * RK=11,3,7,15,17
+ 15.1998 * FITOGEO_IM=Predalpsko,Alpsko
+ 0.0203 * LZSKU70
+ 0.2167 * LZSKU90
+ 0.3798 * LZSKU00
+ 127.4198

LM 3:
LZSKU08 =
0.0399 * NMV_M
+ 10.8066 * RK=11,3,7,15,17
+ 0.0283 * FITOGEO_IM=Predalpsko,Alpsko
+ 0.0001 * LZSKU70
+ 0.1484 * LZSKU90
+ 0.591 * LZSKU00
+ 80.9512

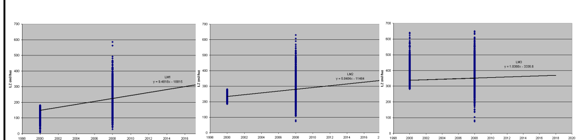
- Correlation coefficient: 0.7002,
- Mean absolute error: 41.9041 m3/ha,
- Root mean squared error: 55.0751 m3/ha,
- Relative absolute error: 69.7019 %,
- Root relative squared error: 71.3877 %.

2008	LM1	LM2	LM3
Average real growing stock (m^3/ha)	224.4	280.0	351.7
Average predicted growing stock (m3/ha)	224.4	280.0	351.8
Mean absolute error (MAE)	49.1	43.3	35.5

Marko Debeljak

Prediction for private forests in 2018

Step 2: - Extrapolation of linear trends from 2000-2008 to the year 2018

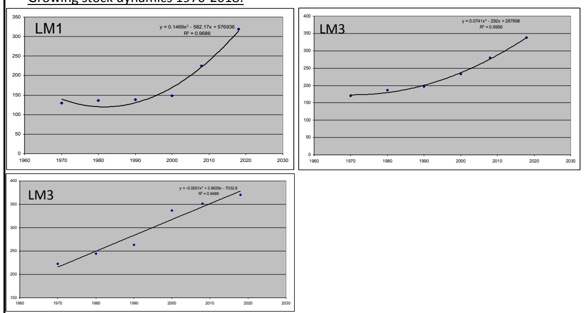


2018	LM1	LM2	LM3
Linear regression (k, n)	9.4818,- 18815	5.8484, - 11464	1.8368, - 3336.6
Average real growing stock 2008 (m3/ha)	224.4	280.0	351.7
Predicted growing stock 2018 (m3/ha)	**319.3**	**338.1**	**370.1**
Index (2008=100)	**142.3**	**120.8**	**105.2**

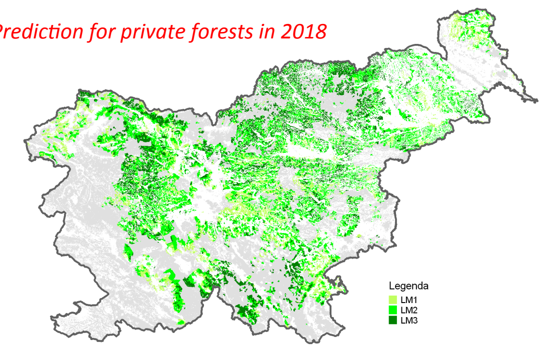
Marko Debeljak

Prediction for private forests in 2018

Growing stock dynamics 1970-2018:

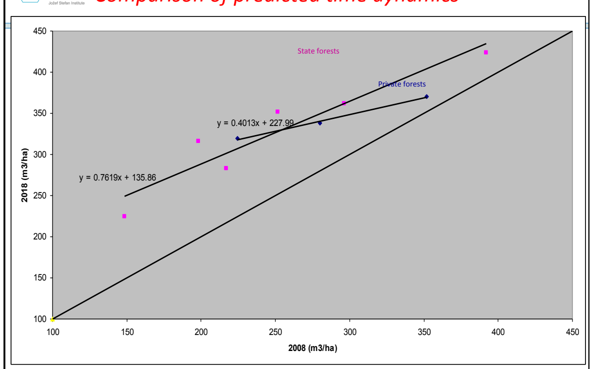


Prediction for private forests in 2018



	LM1	LM2	LM3
Average real growing stock 2008 (m3/ha)	224.4	280.0	351.7
Predicted growing stock 2018 (m3/ha)	319.3	338.1	370.1
Index (2008=100)	142.3	120.8	105.2

Comparison of predicted time dynamics



GIS data mining: COMMUNITY STRUCTURE

Ecological Informatics 5 (2010) 256–266

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Estimating vegetation height and canopy cover from remotely sensed data with machine learning☆

Daniela Stojanova[a], Panče Panov[b,*], Valentin Gjorgjioski[b], Andrej Kobler[a], Sašo Džeroski[b]

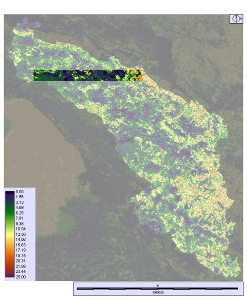
[a] Slovenian Forestry Institute, Večna pot 2, SI-1000 Ljubljana, Slovenia
[b] Jožef Stefan Institute, Department of Knowledge Technologies, Jamova cesta 39, SI-1000 Ljubljana, Slovenia

GIS data mining: COMMUNITY STRUCTURE

Data

- Locations: Kras region (Karst)
- Attributes:
  - Statistical information (max, min, avg, std) from Landsat, IRS, SPOT & aerial photographs
  - Normalized Difference Vegetation Index (NDVI)
  - Textures
  - Relief: Aspect, Slope, Elevation
- Targets (forest properties) from LiDAR data:
  - Vegetation height (H)
  - Canopy Cover (CC)

Landsat and LiDAR data



GIS data mining: COMMUNITY STRUCTURE

Machine Learning Methodology

- WEKA
  - Regression (RT) and Model (MT) trees
  - Bagging of Model Trees (BagMT)
- CLUS
  - Single Target Regression Trees (STRT)
  - Multi Target Regression Trees (MTRT)
  - Ensembles: Bagging of Model Trees (MTBG) and Random Forest (MTRF)

SELECTED: Random Forest Multi Target Regression Trees

Vegetation height

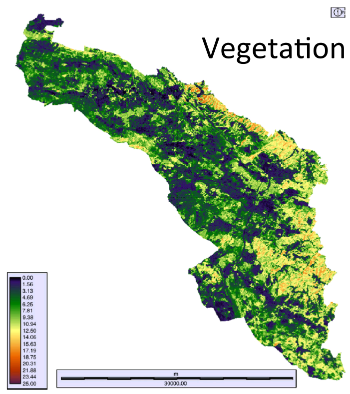


Fig. 5. Map of vegetation height for the Kras region generated by using a random forest of multi-target regression trees model. The legend shows the vegetation height in meters.

Canopy cover

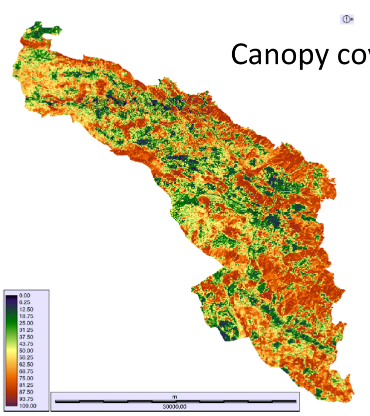


Fig. 6. Map of canopy cover for the Kras region generated by using a random forest of multi-target regression trees model. The legend shows the percentage of canopy cover.

GIS data mining: COMMUNITY STRUCTURE

- The integration of LiDAR and RS promises detailed estimation of forest parameters
- Detail forest vegetation maps can be generated and used for forest management

Data mining of time series: COMMUNITY STRUCTURE

Ecological Modelling 222 (2011) 2524–2529

Contents lists available at ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel

Analysis of time series data on agroecosystem vegetation using predictive clustering trees

Marko Debeljak[a,*], Geoffrey R. Squire[b], Dragi Kocev[a], Cathy Hawes[b], Mark W. Young[b], Sašo Džeroski[a]

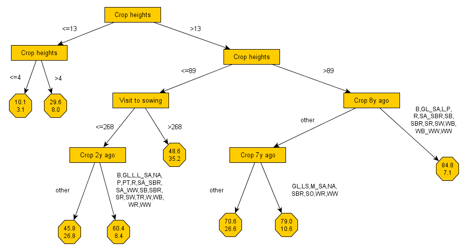
[a] Jozef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia
[b] Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA, Scotland, UK

Data

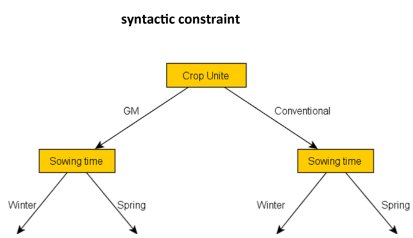
- 130 sites, monitoring every 7 to 14 days for 5 month (2665 samples: 1322 conventional, 1333, HT OSR observations)
- Each sample (observation) described with 65 attributes
- Original data collected by Centre for Ecology and Hydrology, Rothamsted Research and SCRI within Farm Scale Evaluation Program (2000, 2001, 2002)

Results scenario B: Multiple target regression tree

Target: Avg Crop Covers, Avg Weed Covers			**Excluded attributes: /**			
Constraints: MinimalInstances = 64.0; MaxSize = 15						
Predictive power:	Corr.Coef.:	0.8513, 0.3746	RMSE:	16.504,12.6038	RRMSE:	0.5248,0.9301

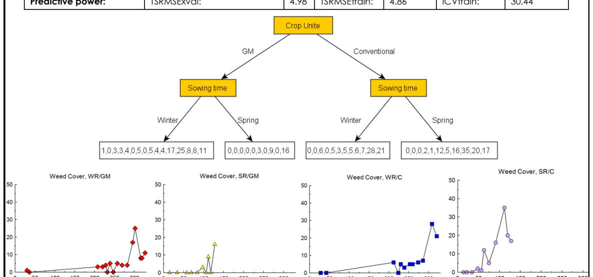


Results scenario D: Constraint predictive clustering trees for time series including TS clusters for crop (CLUS)

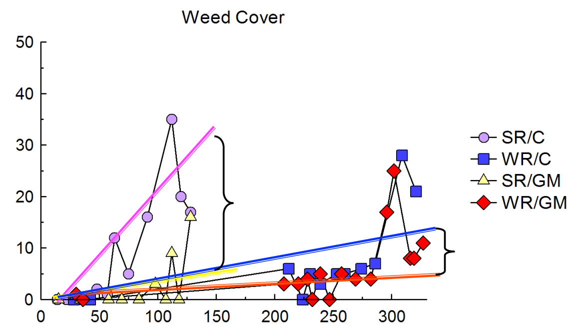


Results scenario D: Constraint predictive clustering trees for time series including TS clusters for crop (CLUS)

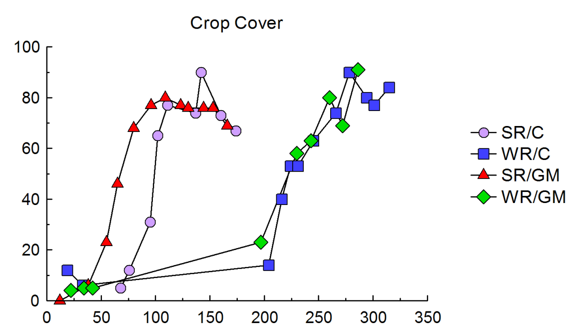
Target: Avg Weed Covers (Time Series)			Scenario 3.9			
Constraints: Syntactic, MinInstances = 32						
Predictive power:	TSRMSExval:	4.98	TSRMSEtrain:	4.86	ICVtrain:	30.44



Results scenario D: Constraint predictive clustering trees for time series including TS clusters for crop (CLUS)



Results scenario D: Constraint predictive clustering trees for time series including TS clusters for crop (CLUS)



Relational data mining: GENE FLOW MODELLING

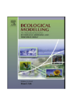
Ecological Modelling 245 (2012) 75–83

Contents lists available at SciVerse ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel

ELSEVIER



Using relational decision trees to model out-crossing rates in a multi-field setting

Marko Debeljak[a,b,*], Aneta Trajanov[a], Daniela Stojanova[a,b], Florence Leprince[c], Sašo Džeroski[a,b,d]

[a] Jozef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
[b] Jozef Stefan International Postgraduate School, Jamova 39, 1000 Ljubljana, Slovenia
[c] ARVALIS-Institut du Végétal, 21, Chemin de Pau, 64121 Montardon, France
[d] Centre of Excellence for Integrated Approaches in Chemistry and Biology of Proteins, Jamova 39, 1000 Ljubljana, Slovenia

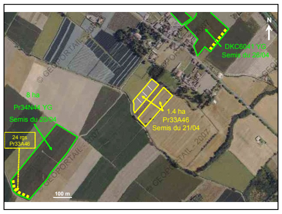
Marko Debeljak

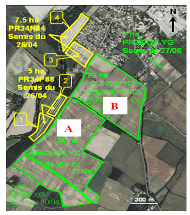
Relational data mining: GENE FLOW MODELLING

Initial questions:

To ***what extent*** will GM maze grown on **Geens** genetically ***interfere*** with the maize on **Yelows**?

Will this interference be small enough to ***allow co-existence***?



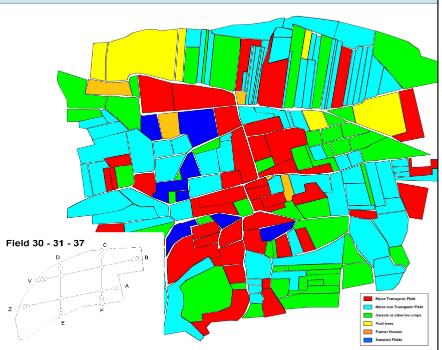


Spatial temporal relations

2004:
40 GM fields
7 non-GM fields
181 sampling points

2005:
17 GM fields
4 non-GM fields
127 sampling points

2006:
43 GM fields
4 non-GM fields

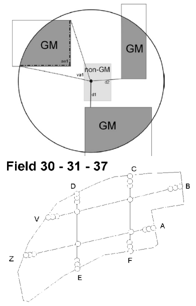


Relational data mining: GENE FLOW MODELLING

Data scattered over several **tables** or **relations**:

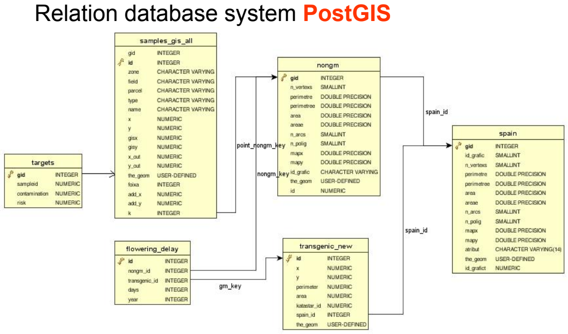
- A table storing **general information** on each field (e.g., area)
- A table storing the **cultivation techniques** for each field and each year
- A table storing the **relations** (e.g., distance) between fields

Relational data mining: GENE FLOW MODELLING



Relational data mining: GENE FLOW MODELLING

Relation database system **PostGIS**



Relational data mining – building model

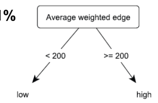
Relation data analysis:

- Algorithm **Tilde** (Blockeel and De Raedt, 1998; De Raedt et al., 2001) => upgrade of algorithm **C.4.5** (Quinlan, 1993) for classification decision trees

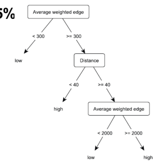
- The algorithm is included in the **ACE-ilProlog** data mining system

Relational data mining – results

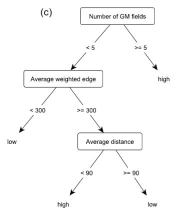
Threshold 0.01%



Threshold 0.45%



Threshold 0.9%

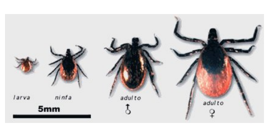


GIS data mining: RISK MODELLING

HABITAT MODELING OF TICK *Ixodes ricinus*, MAJOR VECTOR FOR LYME BORRELIOSIS and TICK-BORNE MENINGOENCEPHALITIS IN THE UPPER SOČA VALLEY

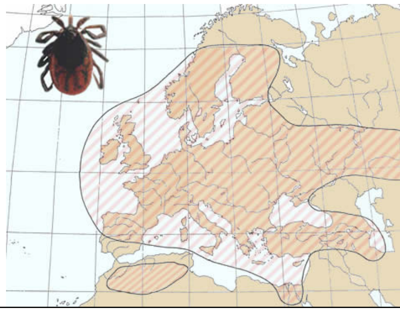
Prof. dr. Marko Debeljak

Tick *Ixodes ricinus L.*:



- Common European ticks
- Ectoparasites (external parasites)
- Living by hematophagy on the blood of mammals, birds, and occasionally reptiles and amphibians
- Vector for Lyme disease and tick-borne meningoencephalitis in humans

- Present in most parts of Europe
- Habitat: mixed, shadow forest with diverse tree composition with shrubs and dense vegetation cover with a lot of litter

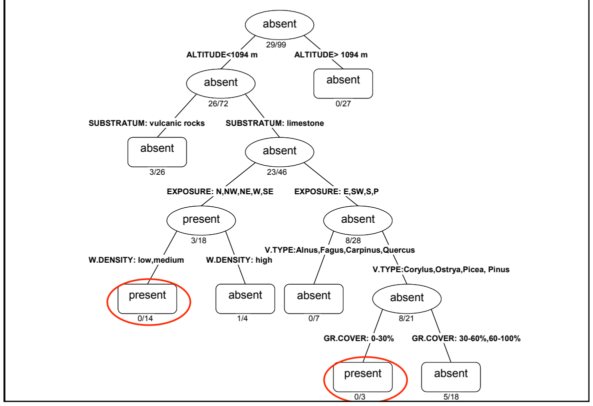


Visitors of natural heritage in upper valley of Soča are exposed to the risk of tick-borne diseases.

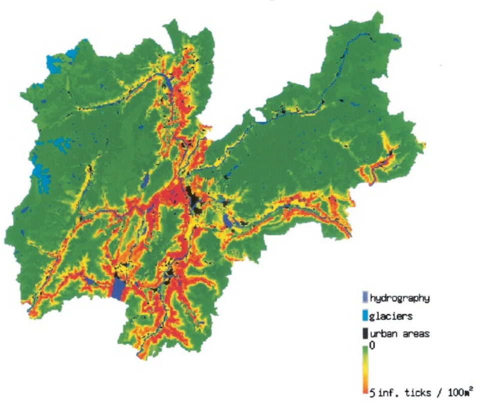
The goal: To identify areas of tick habitats where visitors can be infected with tick-born diseases.

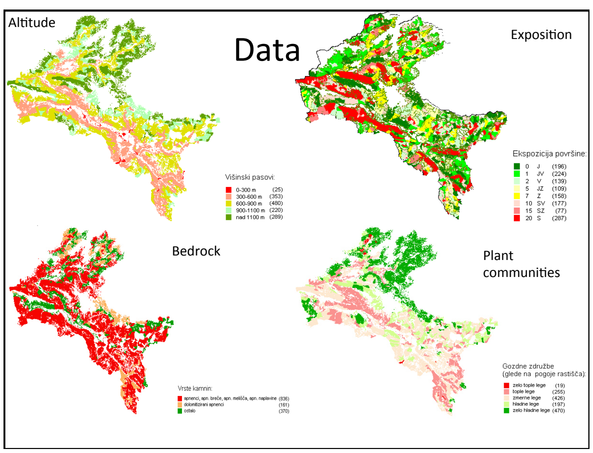
Study area: Zgornje Posočje

Classification tree - Trentino region:



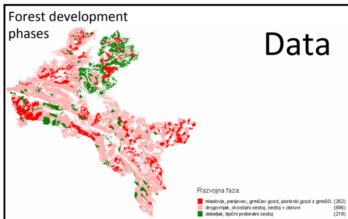
Classification tree – Trentino region :



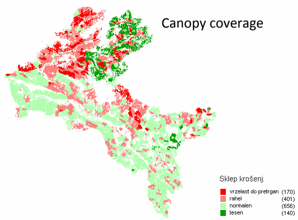


Forest development phases

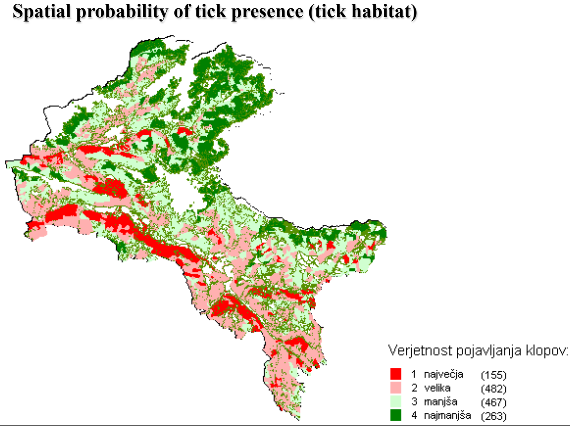
Data



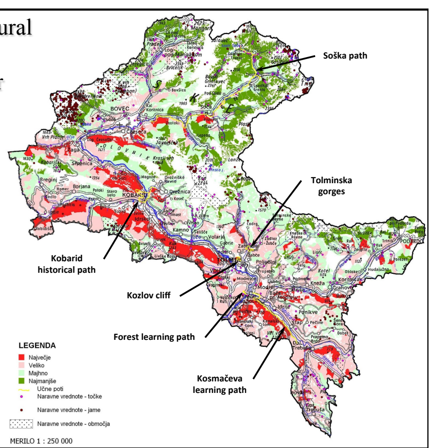
Canopy coverage



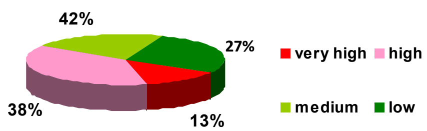
Spatial probability of tick presence (tick habitat)



Locations of natural heritages and exposure of their visitors to ticks

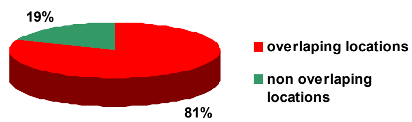


Distribution of **natural heritage** according to the risk of infections with tick-borne diseases in upper Soča valley



Verification of the model:

- Overlapping with the locations of confirmed infections



- Locations with the highest confirmed infections with tick-borne diseases: Tolmin, Kobarid, Volče, Zatolmin, Idrsko, Most na Soči, Poljubinj, Borjana, Bovec, Ljubinj, Žabče, Trebuša, Vrsno.

Multi-target regression model: RISK MODELLING

Soil Use and Management

Soil Use and Management, March 2009, **25**, 66–77 doi: 10.1111/j.1475-2743.2009.00196.x

Potential of multi-objective models for risk-based mapping of the resilience characteristics of soils: demonstration at a national level

M. Debeljak[1], D. Kocev[1], W. Towers[2], M. Jones[2], B. S. Griffiths[3,*] & P. D. Hallett[3]
[1]*Department of Knowledge Technologies, Jozef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia,* [2]*Macaulay Institute, Craigiebuckler, Aberdeen AB15 8QH, UK, and* [3]*Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA, UK*

Multi-target regression model: RISK MODELLING

The dataset: soil samples taken on 26 location throughout SCO

The dataset: The flat table of data:26 by 18 data entries

Multi-target regression model: RISK MODELLING

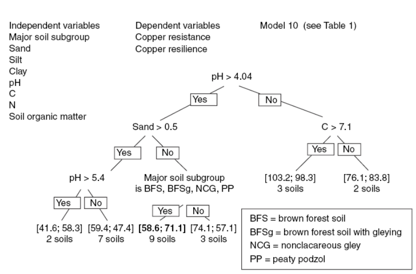
The dataset:

- **physical properties:** soil texture: sand, silt, clay
- **chemical properties:** pH, C, N, SOM (soil organic matter)
- **FAO soil classification:** Order and Suborder
- **physical resilience:** resistance to compression: 1/Cc, recovery from compression: Ce/Cc, overburden stress: eg, recovery from overburden stress after two days cycles: eg2dc
- **biological resilience**: heat, copper

Multi-target regression model: RISK MODELLING

Different scenarios and multi-target regression models have been constructed:

A model predicting the resistance and resilience of soils to copper perturbation.



Multi-target regression model: RISK MODELLING

The increasing importance of mapping soil functions to **advice on land use** and **environmental management** - to make a **map of soil resilience for Scotland.**

The **models = filters** for existing GIS datasets about physical and chemical properties of Scottish soils.

Multi-target regression model: RISK MODELLING

Macaulay Institute (Aberdeen): soils data – attributes and maps:

Approximately 13 000 soil profiles held in database

Descriptions of over 40 000 soil horizons

Application

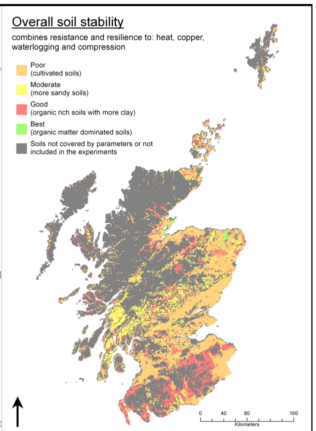
Experiment 1

Independent Attributes
- MAJOR_SOIL_SUBGROUP
- Sand
- Silt
- Clay
- pH
- C
- N
- SOM

Dependent attributes
- Heat_resist
- Heat_Resil
- Cu_resist
- Cu_Resilience
- Over_resist_eg
- Over_resil_2dc_eg
- 1/Cc
- Ce/Cc

Validated Parameters:
RMSE: **[15.4825,20.8952,11.0126,18.5897,0.0809,0.2114,0.3469,0.0098]**
Correlation Coefficient: **[-0.3097,-0.0784,0.8062,0.4445,-0.1455,0.2886,0.355,0.8447]**

```
C > 7.6
+--yes: [56.47,110.4,103.24,98.3,0.789667,1.359333,0.371243,0.055778]: 3
+--no:  C > 4.58
     +--yes: pH > 4.09
     |      +--yes: [48.836,93.08,70.264,64.36,0.8828,1.1358,0.626798,0.016037]: 5
     |      +--no:  [39.8,91.733333,73.483333,79.1,0.805333,0.875,0.847719,0.040065]: 3
     +--no:  N > 0.14
          +--yes: [53.946667,86.833333,54.366667,57.35,0.8165,0.9415,0.914021,0.011523]
          +--no:  [69.673333,81,59.116667,59.866667,0.865,0.911,1.54998,0.02033]: 3
```



Application

Experiment 3A

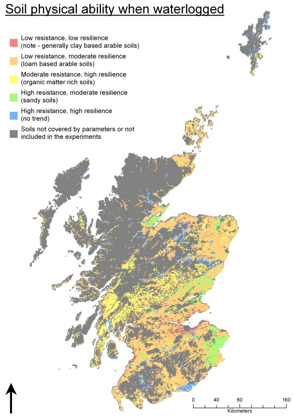
Independent Attributes
- MAJOR_SOIL_SUBGROUP
- Sand
- Silt
- Clay
- pH
- C
- N
- SOM

Dependent attributes
- Over_resist_eg
- Over_resil_2dc_eg

Validated Parameters:
RMSE: **[0.0965,0.2644]**
Correlation Coefficient: **[-0.3021,0.1116]**

```
C > 7.6
+--yes: [0.789667,1.359333]: 3
+--no:  N > 0.21
      +--yes: pH > 5.5
      |       +--yes: [0.742,0.9415]: 2
      |       +--no:  [0.863077,1.042462]: 13
      +--no:  C > 2.75
             +--yes: [0.739,0.788]: 2
             +--no:  [0.840167,0.887333]: 6
```

Soil physical ability when waterlogged



Application

Experiment 3B

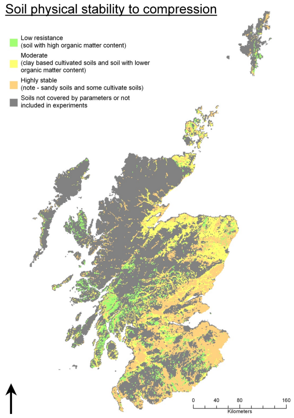
Independent Attributes
- MAJOR_SOIL_SUBGROUP
- Sand
- Silt
- Clay
- pH
- C
- N
- SOM

Dependent attributes
- 1/Cc
- Ce/Cc

Validated Parameters:
RMSE: **[0.2433,0.0164]**
Correlation Coefficient: **[0.708,0.4774]**

```
C > 5.92
+--yes: MAJOR_SOIL_SUBGROUP in {Regosol,Peaty_Podzol,Humus_Iron_P
|       +--yes: Sand > 0.54
|       |       +--yes: [0.718108,0.059179]: 2
|       |       +--no:  [0.299543,0.055084]: 2
|       +--no:  [0.438268,0.025836]: 2
+--no:  N > 0.14
      +--yes: MAJOR_SOIL_SUBGROUP in {Brown_Forest_Soil,Noncalcareo
      |       +--yes: [0.858735,0.00968]: 13
      |       +--no:  [0.920932,0.024988]: 4
      +--no:  [1.54998,0.02033]: 3
```

Soil physical stability to compression



Conclusions

What can data mining do for you?

Knowledge discovered by analyzing data with DM techniques can help:
- Understand the domain studied
- Make predictions/classifications
- Support decision processes in environmental management

Conclusions

What can data mining do for you?

Knowledge discovered by analyzing data with DM techniques can help:

- Understand the domain studied
- Make predictions/classifications
- Support decision processes in environmental management

Conclusions

What data mining cannot do for you?

- The law of information conservation (garbage-in-garbage-out)
- The knowledge we are seeking to discover has to come from the combination of data and background knowledge
- If we have very little data of very low quality and no background knowledge no form of data analysis will help

Conclusions

Side-effects?

- Discovering problems with the data during analysis
  - missing values
  - erroneous values
  - inappropriately measured variables
- Identifying new opportunities
  - new problems to be addressed
  - recommendations on what data to collect and how