Introduction of data mining to forest ecology and forestry

Marko Debeljak
Jožef Stefan Institute, Slovenia

Introduction

Data mining
To discover knowledge in databases

Artificial Intelligence
Machine learning
Database systems

Statistics

Data mining

Hypothesis – Method – Data

Hypothesis testing

complementarity

Model - Hypothesis

Experts

Extraction of information

Data Mining

Model

Experts

Hypothesis – Method – Data

Hypothesis testing

complementarity

Model - Hypothesis

Experts

Extraction of information

Data Mining

Model

Experts

Hypothesis – Method – Data

Hypothesis testing

complementarity

Model - Hypothesis

Experts

Extraction of information

Data Mining

Model

Experts

Hypothesis – Method – Data

Hypothesis testing

complementarity

Model - Hypothesis

Experts

Extraction of information

Data Mining

Model

Experts

Hypothesis – Method – Data

Hypothesis testing

complementarity

Model - Hypothesis

Experts

Extraction of information

Data Mining

Model

Experts

Hypothesis – Method – Data

Hypothesis testing

complementarity

Model - Hypothesis

Experts

Extraction of information

Data Mining

Model

Experts

Hypothesis – Method – Data

Hypothesis testing

complementarity

Model - Hypothesis

Experts

Extraction of information

Data Mining

Model

Experts
Knowledge discovery in databases (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 patterns 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 2: 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 a model that predicts the value of target variable from learning examples, then 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.

**RELATIONAL data mining**

No loss of information, due to aggregation.

**Data**

Data are not stored at all but they continuously flow through algorithm. Each example can propositional or relational.

**DATA STREAM mining**

**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.

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:
\[
\text{out-crossing rate} = 12.3 \times \text{distance} - 0.123 \times \text{wind speed} + 0.00123 \times \text{wind direction}
\]
### 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
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

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

Jana Leganić*, Aleškafer Pelšan*, Marko Debeljak*
* Laboratory for Environmental Research, University of Nova Gorica
† Department of Knowledge Technologies, Faculty of Civil Engineering, University of Nova Gorica, Slovenia

Hydrological conditions
- Ledava River levels
- Groundwater levels

Meteorological conditions
- 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

Measured radial increments
- 8 trees
- 69 years old

Management data
- Thinning; m³/y removed from the stand: Lendava

Dataset
- Monthly data + aggregated data (AMJ, MJJ, JJA, MJJA etc.)
- 124 models

- 52 different combinations of attributes were tested.
- 333 attributes; 35 years

<table>
<thead>
<tr>
<th>Experiment</th>
<th>RRSE</th>
<th># eq. elements</th>
</tr>
</thead>
<tbody>
<tr>
<td>jnj3_2m</td>
<td>0.7282</td>
<td>6</td>
</tr>
<tr>
<td>jnj3_3s</td>
<td>0.7599</td>
<td>6</td>
</tr>
<tr>
<td>jnj3_1s</td>
<td>0.7614</td>
<td>6</td>
</tr>
<tr>
<td>jnj3_4m</td>
<td>0.76455</td>
<td>3</td>
</tr>
<tr>
<td>jnj2_2</td>
<td>0.7685</td>
<td>5</td>
</tr>
<tr>
<td>jly_4xl</td>
<td>0.7686</td>
<td>6</td>
</tr>
</tbody>
</table>
Algebraic equations: POPULATION DYNAMICS

Model jnj3_2m:
RadialGrowthIncrement =
+ 0.05102526982 minL8-10^1
+ 0.029199909999 maxL8-10^1
+ 0.03496385863 lsun8-10^1
+ 0.005866884-91 lsun9-10^2
+ 2.014702948 d-wf-4-7^1
+ 9.356871384x-65 minL4-7^1 lsun4-7^1
+ 6.456856896x-65 minL5-10^1 lsun5-10^1
+ 3.650143419x-65 maxL6-10^1 lsun6-7^1
+ 0.005246442385 lsun6-7^1 d-wf-4-7^1
+ 0.001407867225 lsun6-10^1 d-wf-4-7^1
+ 7.915719872

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: WATER CYCLE

Prediction of drainage water: CIPER

Integrating expert knowledge and predictive training: Modelling water flows in agriculture
Model: Ciper

Integracja doświadczenia, nauki i modelowania przepływów wody w rolnictwie
Model: Ciper
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

<table>
<thead>
<tr>
<th>Fields</th>
<th>Test Field</th>
<th>Std. Dev.</th>
<th>RMSE</th>
<th>RRSE</th>
<th>Corr. coeff (r)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All T3</td>
<td>3.195</td>
<td>2.1119</td>
<td>66.26%</td>
<td>0.7855</td>
<td></td>
</tr>
<tr>
<td>All T4</td>
<td>3.188</td>
<td>1.7220</td>
<td>54.51%</td>
<td>0.8273</td>
<td></td>
</tr>
<tr>
<td>All T5</td>
<td>3.163</td>
<td>2.2478</td>
<td>71.58%</td>
<td>0.7407</td>
<td></td>
</tr>
<tr>
<td>All T6</td>
<td>3.095</td>
<td>2.1784</td>
<td>70.36%</td>
<td>0.8095</td>
<td></td>
</tr>
<tr>
<td>All T7</td>
<td>3.229</td>
<td>1.3298</td>
<td>41.15%</td>
<td>0.7812</td>
<td></td>
</tr>
<tr>
<td>All T8</td>
<td>3.210</td>
<td>1.5813</td>
<td>43.02%</td>
<td>0.7783</td>
<td></td>
</tr>
<tr>
<td>All T9</td>
<td>3.210</td>
<td>1.5627</td>
<td>49.82%</td>
<td>0.7434</td>
<td></td>
</tr>
<tr>
<td>All T10</td>
<td>3.130</td>
<td>1.6672</td>
<td>52.35%</td>
<td>0.7795</td>
<td></td>
</tr>
<tr>
<td>All T11</td>
<td>3.106</td>
<td>1.6274</td>
<td>52.35%</td>
<td>0.7941</td>
<td></td>
</tr>
<tr>
<td>T3 T6</td>
<td>3.055</td>
<td>2.1251</td>
<td>69.54%</td>
<td>0.8129</td>
<td></td>
</tr>
<tr>
<td>T6 T3</td>
<td>3.030</td>
<td>2.0861</td>
<td>58.32%</td>
<td>0.7745</td>
<td></td>
</tr>
</tbody>
</table>
Predictive models

Model (All/T4)

\[
\text{Drainage} = 0.0196445 \times \text{RainfallA1} \times \text{Temp} \\
+ 0.000032681 \times \text{CDEcoef} \times \text{RainfallA1} \times \text{Slope} \\
+ 0.00010753 \times \text{Runoff} \times \text{DrainageN1} \times \text{Temp} \\
+ 0.0015983 \times \text{Runoff} \times \text{RainfallA1} \times \text{Slope} \\
+ 0.0053725 \times \text{RainfallA1} \\
+ 1.5565 \times \text{Runoff} \times \text{Slope} \\
+ 0.02342 \times \text{Runoff} \times \text{RainfallA1} \times \text{Slope} \\
+ 0.057693 \times \text{Slope} \times \text{CDEcoef} \times \text{Runoff} \times \text{Slope} \\
+ 0.149702
\]

Algebraic equations: GENE FLOW

Algebraic equations: Lagramge

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

Ayota, F.; J. C. Tanasevicius, M. Delhez, S. Iliev; N. A. Guerit; P. C. B. Steenwerck.  
Experiment design: Federal Biological Research Centre, BBA, D

60 cobs = 2500 kernels

96 points

Temperature
Relative humidity
Wind velocity

% of outcrossing

Donors - GMO cobs

Receptors - NT cobs

Algebraic equations: GENE FLOW

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

Outcrossing = error \cdot (Distance Influence)^p \cdot (Wind Influence)^r

Distance Influence \rightarrow 1
Distance Influence \rightarrow F
Distance Influence \rightarrow F - \epsilon
F \rightarrow 1 / Distance
F \rightarrow Distance^2 \cdot (0 \leq y \leq 1000)
Distance \rightarrow variable.minDistance
Distance \rightarrow variable.distanceCenter
Wind Influence \rightarrow Wind
Wind Influence \rightarrow P \cdot Wind
P \cdot Wind \rightarrow P \cdot Wind \cdot \text{wind + condition}
Wind \rightarrow variable.appropriateWind
Wind \rightarrow variable.minWindLength

Algebraic equations: GENE FLOW
Algebraic equations: Lagrange

Differential equations: COMMUNITY STRUCTURE

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

Nataša Alenčič**, Žiga Flešer*, Sonja Dovjak, Špela Žbeker Senec*, Friedrich Beckmann†, Boris Kompark‡

† Faculty of Civil and Geodetic Engineering, University of Ljubljana, Slovenia
‡ Environmental Agency of the Republic of Slovenia, Slovenia
§ University of Ljubljana, Slovenia

Fig. 2 – Simple conceptual model for lake Bled.

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

<table>
<thead>
<tr>
<th>Variable name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chla (mg/m³)</td>
<td>chlorophyll a</td>
</tr>
<tr>
<td>Chl-b (mg/m³)</td>
<td>chlorophyll b</td>
</tr>
<tr>
<td>SEC (mg/m³)</td>
<td>suspended organic carbon</td>
</tr>
<tr>
<td>N (mg/m³)</td>
<td>total nitrogen</td>
</tr>
<tr>
<td>P (mg/m³)</td>
<td>total phosphorus</td>
</tr>
<tr>
<td>pH</td>
<td>pH value</td>
</tr>
<tr>
<td>NO₃ (µM)</td>
<td>nitrate concentration</td>
</tr>
<tr>
<td>NO₂ (µM)</td>
<td>nitrite concentration</td>
</tr>
<tr>
<td>PO₄ (µM)</td>
<td>phosphate concentration</td>
</tr>
<tr>
<td>Dosed clover</td>
<td></td>
</tr>
</tbody>
</table>
Differential equations: COMMUNITY STRUCTURE

**Phosphorus**

\[
\frac{d[\text{phosphate}]}{dt} = \text{water in-flow} - \text{out-flow} - \text{respiration} - \text{growth} - \text{sedimentation} - \text{grazing}
\]

**Phytoplankton**

**Zooplankton**

Differential equations: COMMUNITY STRUCTURE

Differential equations: COMMUNITY STRUCTURE

Differential equations: COMMUNITY STRUCTURE

Feeds on phytoplankton
Differential equations: COMMUNITY STRUCTURE

Population growth

Decision trees: HABITAT MODELS

Classification trees: J48

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

Decision trees: HABITAT MODELS
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

Observed locations of BBs

Dataset

- Present: 1
- Absent: 0
The model for maximal habitat

The model for optimal habitat

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 unit is a permanent compartment
- data from 21052 permanent compartments for the period

Share of state owned forest in compartment:
100%
5237
“State owned forests”

Share of state owned forest in compartment:
<100%
15815
“Private forests”
**Data**

- NMV_M: elevation (m)
- NAKLON_M: slope [%]
- EXP: aspect (ranks 0-8)
- MAT_PODLAG: bedrock (1-carbonate, 0-silicate)
- NKLON_M: slope (%)
- EXP: aspect (ranks 0-8)
- FITOGEO_M: fito-geographical regions (Pre-dinaric, Alpine, Dinaric, Sub-mediterranean, Pre-panonic)
- L25U70: growing stock 1970 (m³/ha)
- L25U80: growing stock 1980 (m³/ha)
- L25U90: growing stock 1990 (m³/ha)
- L25U00: growing stock 2000 (m³/ha)
- L25U08: growing stock 2008 (m³/ha)
- E70: cut (etat) 1970 (m³/ha)
- E80: cut (etat) 1980 (m³/ha)
- E90: cut (etat) 1990 (m³/ha)
- E00: cut (etat) 2000 (m³/ha)

**Experimental design**

- Share of state owned forest in compartment: 100% „state owned forest“
- Share of state owned forest in compartment: < 100% „private owned forest“
- Analysis of the time dynamics 1970-2008
- Total growing stock
- DBH < 30 cm (A)
- DBH 30-50 cm (B)
- DBH > 50 cm (C)
- Time index
- Site index
- Time index
- Site index
- Time index
- Site index

**Time dynamics 1970-2008**

Growing stock in private forests
Growing stock in state forests

Time dynamics 1970-2008

Growing stock in state forests

Time dynamics 1970-2008

Time dynamics for DBH classes in state forests

Marko Debeljak
**Time dynamics 1970-2008**

Time dynamics for DBH classes in private forests

- **DBH < 30 cm**
- **DBH 30-50 cm**
- **DBH > 50 cm**

**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

**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

<table>
<thead>
<tr>
<th>2000</th>
<th>LM1</th>
<th>LM2</th>
<th>LM3</th>
<th>LM4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average real growing stock</td>
<td>253.4</td>
<td>275.1</td>
<td>304.9</td>
<td>352.3</td>
</tr>
<tr>
<td>Average model growing stock</td>
<td>253.4</td>
<td>275.1</td>
<td>304.9</td>
<td>352.3</td>
</tr>
<tr>
<td>Mean absolute error (MAE)</td>
<td>11.0</td>
<td>10.4</td>
<td>8.3</td>
<td>5.9</td>
</tr>
</tbody>
</table>
**Prediction for private forests in 2018**

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

<table>
<thead>
<tr>
<th>Year</th>
<th>LM1</th>
<th>LM2</th>
<th>LM3</th>
<th>LM4</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear regression a, b</td>
<td>5.8801, -11543</td>
<td>3.4203, -6565.4</td>
<td>1.6311, -2957.4</td>
<td>-0.8628, 2077.9</td>
</tr>
<tr>
<td>Average real growing stock</td>
<td>264.2</td>
<td>302.6</td>
<td>317.8</td>
<td>345.4</td>
</tr>
<tr>
<td>Difference (%)</td>
<td>0.0</td>
<td>0.0</td>
<td>0.4</td>
<td>0.2</td>
</tr>
</tbody>
</table>

**Marko Debeljak**

**Prediction for private forests in 2018**

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

<table>
<thead>
<tr>
<th>Year</th>
<th>LM1</th>
<th>LM2</th>
<th>LM3</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average real growing stock (m³/ha)</td>
<td>224.4</td>
<td>280.0</td>
<td>351.7</td>
</tr>
<tr>
<td>Average predicted growing stock (m³/ha)</td>
<td>224.4</td>
<td>280.0</td>
<td>351.8</td>
</tr>
<tr>
<td>Mean absolute error</td>
<td>49.1</td>
<td>43.3</td>
<td>35.5</td>
</tr>
<tr>
<td>Relative absolute error</td>
<td>69.7%</td>
<td>58.3%</td>
<td>40.0%</td>
</tr>
<tr>
<td>Root relative squared error</td>
<td>71.3%</td>
<td>58.3%</td>
<td>40.0%</td>
</tr>
</tbody>
</table>

**Correlation coefficient:** 0.7002, **Mean absolute error:** 49.1 m³/ha, **Relative absolute error:** 69.7%, **Root relative squared error:** 71.3%.
**Prediction for private forests in 2018**

Growing stock dynamics 1970-2018:

- **LM1**
- **LM2**
- **LM3**

**Average growing stock (2008 m³/ha)**

<table>
<thead>
<tr>
<th></th>
<th>LM1</th>
<th>LM2</th>
<th>LM3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted growing stock (2018 m³/ha)</td>
<td>319.3</td>
<td>338.1</td>
<td>370.1</td>
</tr>
<tr>
<td>Index (2008=100)</td>
<td>142.3</td>
<td>120.8</td>
<td>105.2</td>
</tr>
</tbody>
</table>

**Comparison of predicted time dynamics**

- **State forests**
- **Private forests**

**Equations**

- $y = 0.1469x^2 - 582.17x + 576936$  \( R^2 = 0.9686 \)
- $y = 0.0741x^2 - 292x + 287898$  \( R^2 = 0.9956 \)
- $y = -0.0001x^2 + 3.9629x - 7032.8$  \( R^2 = 0.9489 \)
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)
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
GIS data mining: COMMUNITY STRUCTURE

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

Data mining of time series: COMMUNITY STRUCTURE

Analysis of time series data on agroecosystem vegetation using predictive clustering trees
Maksim Debovsky³, Geoffrey R. Squire³, Onut4 Kacore³, Cathy Howes², Mark W. Young³, Sola Dekrmekj³

Data

- 130 sites, monitoring every 7 to 14 days for 5 months (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

<table>
<thead>
<tr>
<th>Target</th>
<th>Avg Crop Covers</th>
<th>Avg Weed Covers</th>
<th>Excluded attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictive power:</td>
<td>Corr. Coef.: 0.8513, 0.3746</td>
<td>RMSE: 16.504, 12.6038</td>
<td>RRMSE: 0.5248, 0.9301</td>
</tr>
</tbody>
</table>

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

Target: Avg Weed Covers (Time Series)

Constraints: Syntactic, MinInstances = 32

Predictive power:
- TSRMSEexval: 4.98
- TSRMSEtrain: 4.86
- ICVtrain: 30.44
Results scenario D: Constraint predictive clustering trees for time series including TS clusters for crop (CLUS)

Weed Cover

Crop Cover

Relational data mining: GENE FLOW MODELLING

Using relational decision trees to model out-crossing rates in a multi-field setting
Marko Debeljak*, Aneta Trajanova, Daniela Stojanova*, Florence Leprince*, Salo Dzeroski*, A.

*Institute for Informatics Applications, Faculty of Chemical Engineering, University of Chemical Technology and Metallurgy, Sofia 1126, Bulgaria

†Institute of Computing Science, Faculty of Computer Science, University of Chemical Technology and Metallurgy, Sofia 1126, Bulgaria
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

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 0.9%

Threshold 0.45%

Threshold 0.01%

**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:

- Altitude
- Exposure
- Substratum
- Data
- Plant communities

Classification tree – Trentino region:

Altitude
Data
Exposure
Bedrock
Plant communities
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

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

Macaulay Institute (Aberdeen): soils data - attributes and maps:
  Approximately 13,000 soil profiles held in database
  Descriptions of over 40,000 soil horizons

Application

Overall soil stability

Soil stability is a key factor in determining the suitability of land for various uses. This map shows the stability of soils across Scotland, with different colors indicating varying levels of stability.

Legend:
- Red: Highly unstable
- Orange: Moderately unstable
- Yellow: Stable
- Green: Very stable
- Blue: Extremely stable

The map is useful for planning agricultural, forestry, and infrastructure projects, ensuring that the chosen locations are appropriate for the intended use and avoid areas with high soil instability.
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

### 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

### 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