Analysis of forest foliar nutrition data

Mika Sulkava

T-61.6060 Special Course in Computer and Information Science VI
Data analysis and environmental informatics
1.3.2006

Contents

- Introduction
- Forest foliar nutrition data
- Foliar analysis using clustering of the self-organizing map
- Analyzing aging of needles with sparse regression
- Weighted regression and data quality
- Summary

Introduction

- Plants take up substances from environment.
- Foliar mineral composition is related to environment.
- Analysis of foliar nutrient concentrations is an important part of environmental monitoring.
- Results of cooperative foliar nutrient research done in:
  - Laboratory of Computer and Information Science
  - Finnish Forest Research Institute
  - University of Antwerp

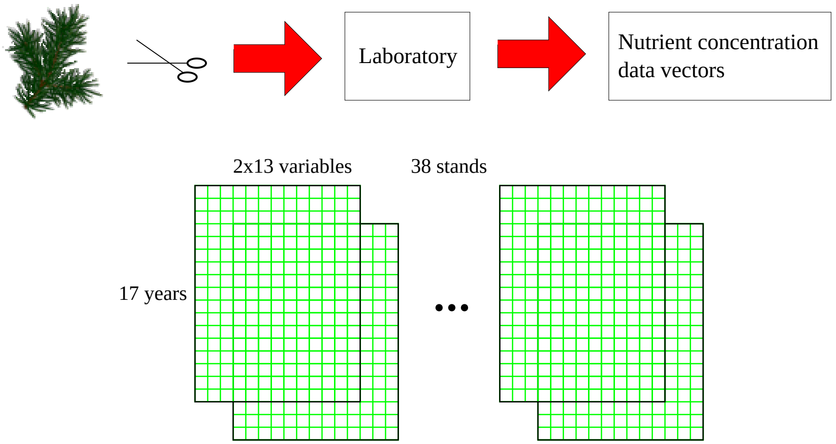
Forest foliar nutrition data

- Nutrient concentration data measured from forests of Finland.
- Data from a large-scale forest monitoring program:
  - International Co-operative Programme on Assessment and Monitoring of Air Pollution Effects on Forests (ICP) operating under United Nations Economic Commission for Europe (UNECE).
- Data collected by Finnish Forest Research Institute.

Forest foliar nutrition data

- Foliar nutrient data from 38 Finnish ICP Forests Level I stands.
  - 17 Norway spruce and 21 Scots pine stands located in different parts of Finland.
  - Annual measurements between years 1987–2003.
  - In each stand mass of needles and concentrations of 12 nutrients in pine or spruce needles were measured.
  - Measurements of both new and one year old needles.
  - 29% of measurements missing.

Forest foliar nutrition data



Laboratory

Nutrient concentration data vectors

2x13 variables

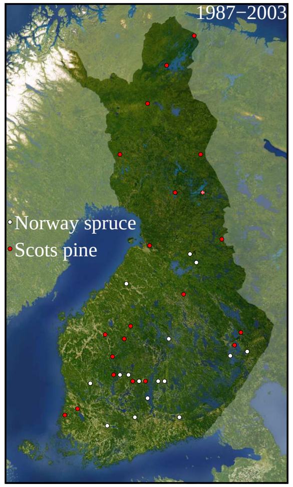
38 stands

17 years

...

Forest foliar nutrition data

Concentration measurements from the needles: Al, B, Ca, Cu, Fe, K, Mg, Mn, N, P, S, Zn.



Forest foliar nutrition data

- Environmental measurements:
  - Deposition measurements
  - Temperature
  - Precipitation
- Laboratory quality data from different sources:
  - International interlaboratory tests
    - ∗ International Union of Forest Research Organizations (IUFRO) laboratory comparisons, 1987–1994
    - ∗ ICP Forests ring tests, 1993–
  - National calibration tests with certified reference materials, 1995–

Foliar analysis using clustering of the SOM

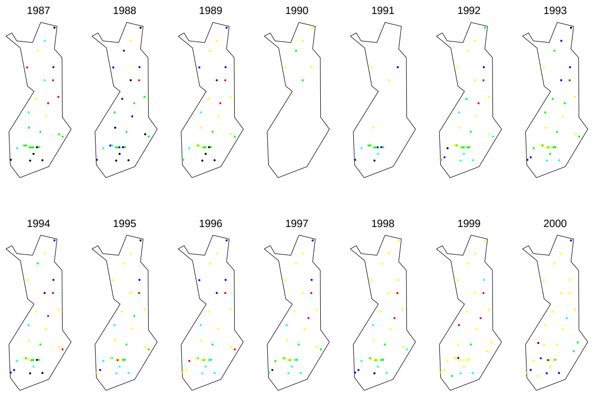
- Clustering of the SOM was used to analyze chemical composition of tree foliage.
- Exploratory analysis.
- Aims to understand the spatio-temporal mechanisms in development of foliar nutrient concentrations.
- Clusters (nutrition profiles) are a new concept in foliar analysis.
- 4D vectors: 3 concentrations (N, S, P) and needle mass (NM).

Clustering method

- An automated clustering approach.
- Results similar to the U-matrix.
- Four phases:
  - Calculate SOM and distance matrix
  - Divide map into base clusters
  - Construct cluster hierarchy
  - Select final partitioning
- Cluster hierarchy allows the data to be investigated at several levels of detail

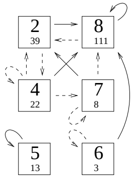
Results of clustering

- 6 clusters.
- Some correlation with location.



Results of clustering

- How about changes in time?
- Transition matrices show different cluster swithces in time
- In pine stands the clusters change with time.
  - Clusters with low N, S, P, K, Ca, Mg and Al concentrations have become more abundant.



- The effect of N and S deposition on needles has decreased between 1987-2000.

Analyzing aging of needles with sparse regression

- Understanding and predicting the development of nutrient concentrations are challenging tasks.
- Aims:
  - Predict nutrient concentrations and needle mass of one year old needles in year t using the measurements of new needles in year $t-1$ and the environmental measurements in year t.
  - Model the effect of environment and nutrients to the aging of the needles.
  - Use only a few significant regressors of total 22 for each response.
  - The models should give an understandable description of the connections between variables.

Sparse regression models

- Different multiple linear regression models were used for prediction:

$$X_{i,t,C+1} = \sum_{j=1}^{13} \beta_{i,j} X_{j,t-1,C} + \sum_{j=14}^{22} \beta_{i,j} Z_{j,t} + \epsilon_i$$

- Main advantages of linear models:
  - Easy to interpret.
  - Over short ranges, any process can be well approximated by a linear model.
- In a sparse regression model, some coefficients $\beta_{i,j} = 0$.

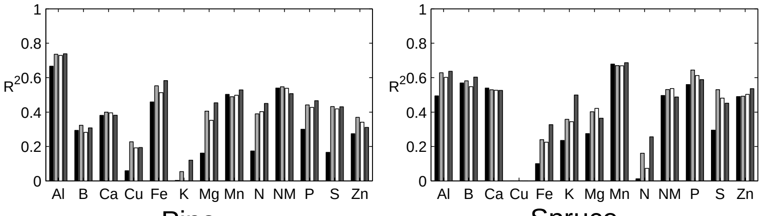
Sparse regression models

- Small number of coefficients makes the model easier to interpret and less prone to overfitting.
- Least Angle Regression (LARS) model selection algorithm and MDL information criterion were used to find the most significant regressors.
- Forward selection was used as a baseline method.
- The sparse models were compared to full regression model and a simple one-parameter model.

Results of sparse regression

- The quality of prediction was measured with the coefficient of determination R^2 and validated using 20 times 10-fold cross-validation.
- Usually, the sparse models outperform the one-parameter model, and their results are mainly comparable to the full model.
- The number of coefficients in sparse models is much lower: on average 6.1 in forward selection and 4.4 in LARS (out of 22).

Results of sparse regression

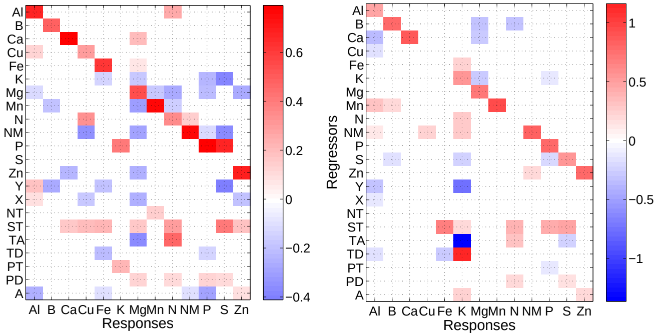


Pine

Spruce

Average R^2-values for one-parameter (black), forward selection (light gray), LARS (white) and full models (dark gray) obtained using cross-validation.

Results of sparse regression



Pine

Spruce

Values of the coefficients of LARS models.

Results of sparse regression

- A typical LARS model and full model:

$$\begin{aligned} Zn_{t,C+1} &= -0.27Mg_{t-1,C} + 0.69Zn_{t-1,C} \\ &\quad -0.20X + 0.18ST_t + 0.09A_t \end{aligned}$$

$$\begin{aligned} Zn_{t,C+1} &= 0.18Al_{t-1,C} - 0.01B_{t-1,C} + 0.02Ca_{t-1,C} + 0.09Cu_{t-1,C} \\ &\quad -0.04Fe_{t-1,C} - 0.11K_{t-1,C} - 0.19Mg_{t-1,C} + 0.05Mn_{t-1,C} \\ &\quad -0.11N_{t-1,C} - 0.07NM_{t-1,C} + 0.15P_{t-1,C} - 0.13S_{t-1,C} \\ &\quad +0.62Zn_{t-1,C} - 0.55Y - 0.35X + 0.06NT_t + 0.25ST_t \\ &\quad -0.63TA_t + 0.32TD_t - 0.13PT_t + 0.07PD_t + 0.14A_t \end{aligned}$$

- Permutation test showed that virtually always the best regressors were chosen to the LARS models.
- Given the number of coefficients, it is very hard to find a model that characterizes better the development of the foliage.

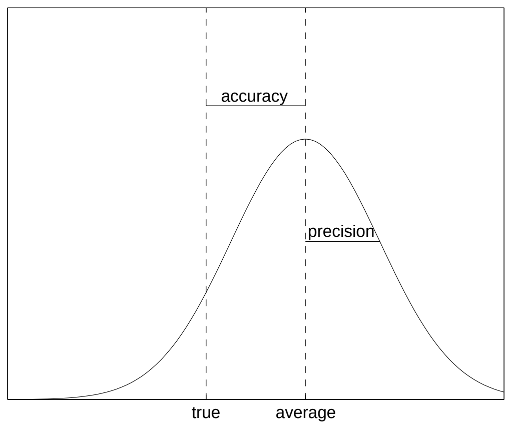
Weighted regression and data quality

- Chemical analyses of foliar samples are prone to many errors.
- Laboratory quality has improved in past two decades due to quality control and development of methods.
- Despite the improvements there are still problems in quality.
- The impact of laboratory quality on detecting changes in environment was studied.
- Theoretical computations and experiments with real-world data were used to analyze how trend detection is affected by changing data quality.

Weighted regression and data quality

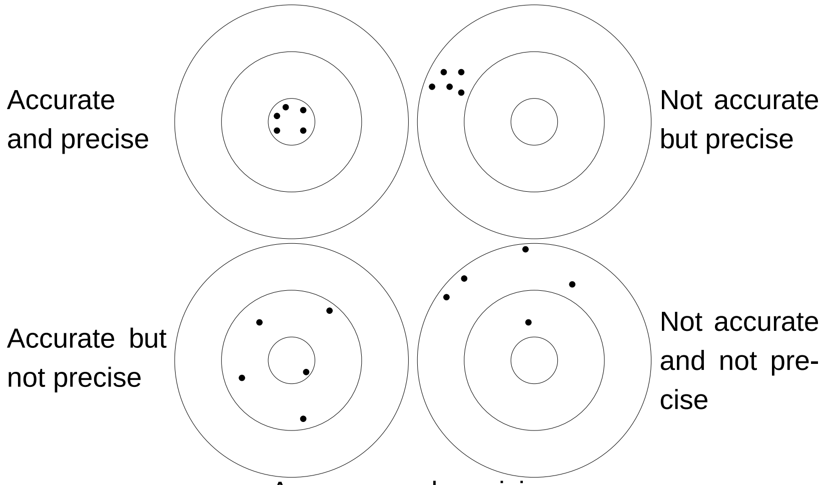
- Aims:
  - Analyze the effect of changing data quality on detecting changes in environment.
  - Study the use of weighted linear regression models in detecting trends in foliar time series data.
  - Find out how improvements in laboratory quality affect the statistical significance of trends found in foliar nutrients.
  - Calculate how much improvements in laboratory quality decrease the time needed to detect a trend.

Data quality



Distribution of measurements, accuracy and precision.

Data quality



Accuracy and precision

Weighted regression models

- Ordinary least squares regression assumes homoscedastic data.
- Weighted least squares regression can be used to analyze heteroscedastic data.

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,\ i = 1, \ldots, n$$

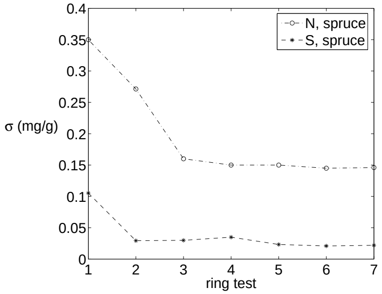
$$\epsilon_i \sim N(0, \sigma_i^2)$$

$$w_i = \frac{1}{\sigma_i^2}.$$

- Iteratively reweighted least squares regression (IRLS) can be used if variance is partially unknown.
- The hypothesis $\beta_1 \neq 0$ can be tested with the F-test.

Data quality

- The development of laboratory quality was inspected.

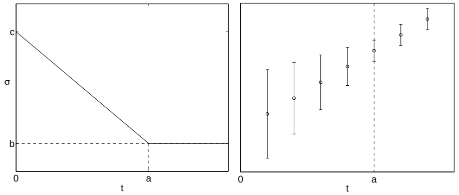


Average precision of laboratories in ICP Forests ring tests.

Model of data quality

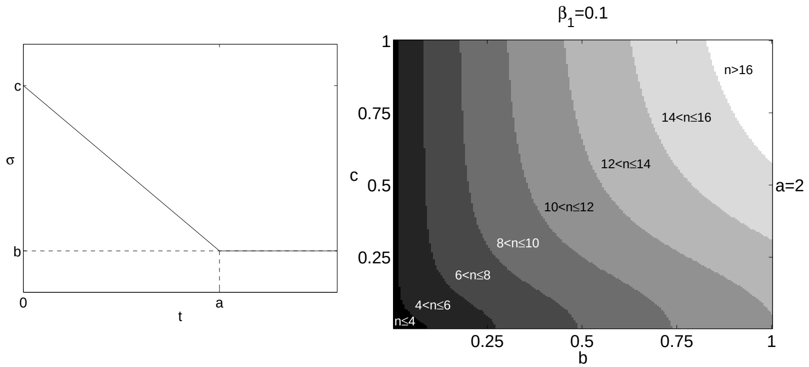
- A simple linear model for development of measurement precision was constructed.

$$\sigma_i = \begin{cases} \frac{b-c}{a}X_i + c & \text{if } X_i \le a \\ b & \text{if } X_i > a \end{cases}$$



Precision Measurements

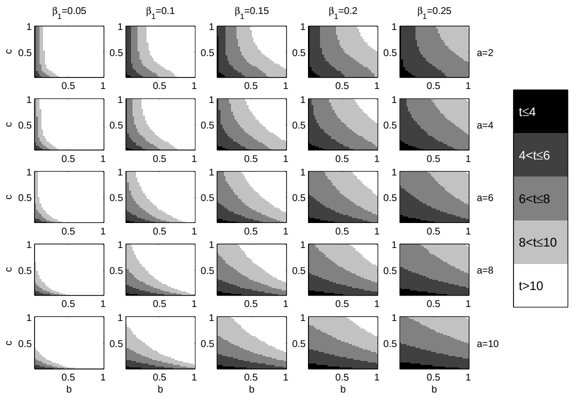
Results of weighted regression



Precision Trend detection

Trend detection with linearly changing precision.

Results of weighted regression



Trend detection with linearly changing precision and different parameter values.

Results of weighted regression

- Accuracy and precision of the laboratory in Finland was estimated using combined results of the three quality tests.
- The foliar nutrient data was analyzed using IRLS regression.
- Statistically significant ($p < 0.05$) increasing trend was found in eight nitrogen (N) and decreasing trend in 26 sulfur (S) time series (out of 38).
- The trends were detected on average in 11 years.
- If the precision of the Finnish laboratory would have been the same as in the most imprecise laboratories in Europe, none of the trends would have been detected.

Summary

- Analysis of nutrient concentrations of needles in Finland.
- Analysis using clustering of the self-organizing map:
  - Identification of profilic states from forest nutrition data.
  - Temporal cluster development: cluster switches in time
  - Decrease of many nutrient concentrations in nutrition profiles of pine needles.
  - Decreased effect of N and S deposition.

Summary

- Factors affecting aging of needles
  - Sparse models were found to be more suitable for the problem than the two other models.
  - They have comparable prediction accuracy to the full model, but with a significantly smaller number of parameters.
  - LARS models are slightly sparser than forward selection models.
  - Sparsity makes interpretation easy.
  - Helping to find the significant dependencies between different variables is an important feature of the sparse models.

Summary

- Measurement quality
  - Experiments with weighted regression show that measurement precision strongly affects trend detection.
  - The results from theoretical computations and experiments with real world data highlight the importance of quality in laboratory analyses.
  - Improving data quality can decrease the time needed for finding statistically significant trends.
  - With better quality smaller trends can be detected.