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

Neural Network Prediction in Lake and Marine Ecosystems

Matthieu Molinier

15.03.2006

Material

- 2 chapters of the book by F. Recknagel : *Ecological Informatics - Understanding Ecology by Biologically-Inspired Computation*
- Additional material (1 journal paper, links to websites)

Part I

C.H. Reick, A. Grünewald, B. Page

Multivariate Time Series Analysis Prediction of Marine Zooplankton by Artificial Neural Networks

Environmental data

- Low quality data
  - Noisy : environmental <=> non-laboratory conditions
  - Non-representativity : one measures what one can get, not what one wants
  - Temporal/spatial extent : limited because expensive
- High number of simultaneous variables
  - Only improves information on particular system states
  - Too many inputs can harm prediction with NN

Neural Networks

- High adaptivity, ability to generalize
- Experiments : input data / pre-processing / NN structure
- Selection : model with lowest *prediction error*

Works well with high quality data (successful generalisation)

- Reliability of the prediction with low quality data ? How to detect generalisation success / failure ?

5 causes of bad predictive performance (1/2)

- 2 causes independent of Neural Networks
  - Unpredictability : nature of the phenomenon (e.g. stock returns)
  - Poor data :
    - Not enough data
    - Too noisy data
    - Incomplete dataset / phenomenon not completely represented
    - Wrong information due to e.g. instationarity
- Under-adaptation
  - If training is stopped too early, reduced ability to distinguish different system states

5 causes of bad predictive performance (2/2)
Causes relating to generalisation failures

- Over-adaptation (over-training)
  - Training error decreases monotonically, then validation error rises
  - Good adaptation and good generalisation are conflicting goals
- Unsuited network structure (how to choose it ?) : number of neurons, network connectivity, activation and output functions
  - –> change also the topology during training
- General *performance failure* and *generalisation failure* in practice

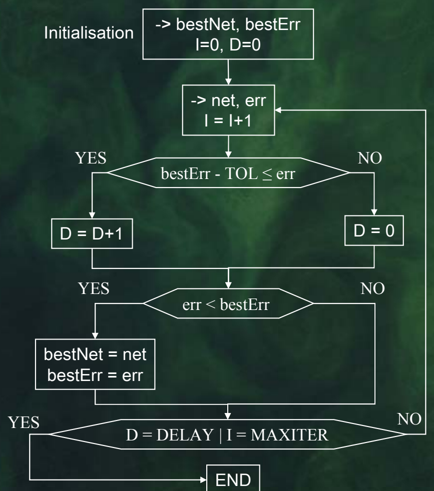
Other aspects of generalisation

- Reliability of predictions (gen. failures independent of data)
  - Train a NN on high quality data, test on poor quality (very noisy)
  - -> Prediction would be poor but reliable
- Model correctness
  - systematic deviations from correct values
  - relates to correlation between errors and data
- How to detect (un)reliability or model (in)correctness for the two types of *generalisation failures* ?

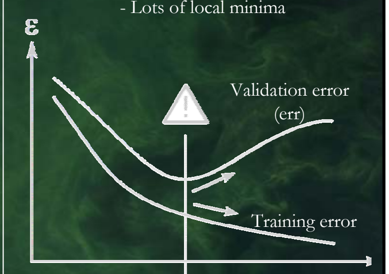
2 types of generalisation failures

- Over-adaptation
  - Reliability : is prediction error stationary ? (needs large dataset)
  - Model correctness : are predictions errors uncorrelated to data ?
- Unsuited network structure
  - Consider a family of networks, assume "ideal training"
  - Wrt network structure, reliability = "predictive performance is independent of training data" (CV / leave-1-out)
  - Small fluctuations of the *mean prediction errors* indicate a good ability to generalize wrt to network structure

Automatic termination of training – early stopping



Cross-validation <=> lots of training
Validation error curve :
- No global minimum
- Lots of local minima



Source : Wikipedia

Training is stopped if for DELAY cycles, err improves less than TOL

Case study - Zooplankton prediction

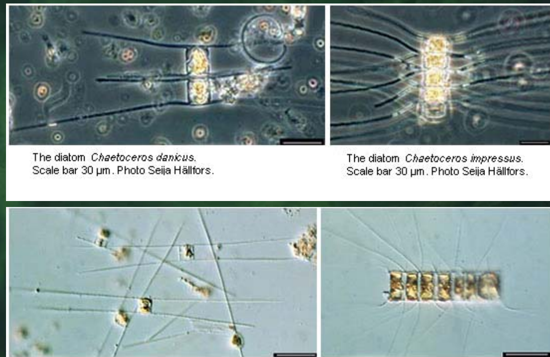
- Zooplankton development 1975-1994 in Helgoland (North Sea)
- Every 2-3 day at <u>the same location</u> :
  - Plankton fished (net), visual inspection/identification (microscope
  - Water flow (-> density of organisms indiv./m^3) => **abundance**
  - 7 physical parameters measurements : w**ater temperature, salinity, phosphate concentration […]**

=> Abundances of **45** groups of zooplankton organisms ("taxa"), 3200 points of time (averaged over ~7 days, 52 points/year)

Cruel laws of nature…

Cirripedia nauplius, Barnacle larva

eats



The diatom *Chaetoceros danicus*. Scale bar 30 µm. Photo Seija Hällfors.

The diatom *Chaetoceros impressus*. Scale bar 30 µm. Photo Seija Hällfors.

The diatom *Chaetoceros similis*. Scale bar 30 µm. Photo Seija Hällfors.

The diatom *Chaetoceros wighamii*. Scale bar 30 µm. Photo Seija Hällfors.

- 2 phytoplankton groups (diatoms and flagellates) -> carbon mass/m3

Known facts

- Data taken from a single point -> not representative
- Short-time prediction of *abundance* is not possible

=> Prediction of magnitude of abundance : $m = \log_{10}(x+1)$

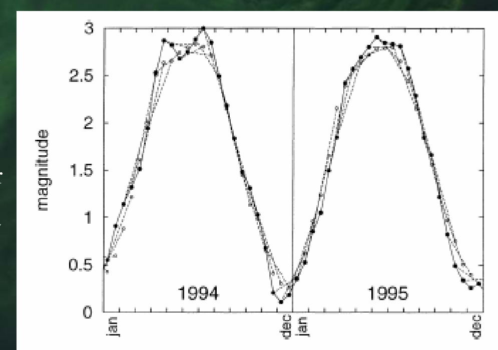
- Complex network of interactions within plankton ("food web")
  - simultaneous (multivariate) prediction
  - massive computation (several combinations of inputs)

Prediction of Barnacle larvae abundance

- Inputs : magnitude of abundances
  - Last 8 weeks of *Cirripedia nauplius*
  - Last 8 weeks of diatom data (one of the preys)
- 16 x 5 x 2 x 1 feedforward Neural Network trained by *resilient backpropagation*
- First 16 years in-sample, (12 y training, 4 y validation), 30 CV
- MAXITER = 1000, DELAY = 100, TOL = 10^{-4}

Results

- 364 iterations
- Predicted vs actual data
  - corr = 0.91
  - E[error] < data
  - In winter, same order
- Prediction error vs data
  - corr = 0.46
  - Not independent (model incorrectness)
- Generalization fails



Conclusions (I)

- Relevant inputs to the NN are not known (plankton interactions)
- Reliability and model correctness have to be evaluated independently from prediction quality
  - –> supposes automation of training, lack of tools for it
- C. Reick, B. Page, *Time Series Prediction by Multivariate Next Neighbor Methods with Application to Zooplankton Forecasts*. Mathematics and Computers in Simulation 52 (2000), pp. 298-310.

Part II

H. Wilson, F. Recknagel

A Generic Artificial Neural Network Model for Short-Term Predictions of Algal Blooms in Lakes and Reservoirs

Background

- Growth of algal biomass -> water blooms affects water quality
  - –Taste
  - –Odour
  - –Toxicity
- Complicate relationships between nutrients, physical / biological / chemical processes
- Classical methods fail to predict timing, magnitude and algal species of significant bloom events

Neural Network for algal bloom modelling

- Neural Networks applied to algal bloom modelling
  - symbolic expression of domain knowledge is not required
  - inherent non-linearity of the phenomenon
- 1994 : modelling phytoplankton growth dynamics (German lake)
  - Water quality -> cell counts of phytoplankton species
- Applications in other places (Australian rivers, Japanese lakes, Finnish lake Tuusulanjärvi, Turkish reservoirs)

General considerations (1)

- Aiming at a generic input layer design
  - comparison between different data sources possible
  - improved knowledge from data aggregation
  - reduces modelling effort
- Time-delay input structure
  - exploits serial correlation in the data
  - enables forecasts of algal abundance up to 4 weeks ahead (rivers)
- Claim : use the same structure for a generic model (rivers, lakes, reservoirs)

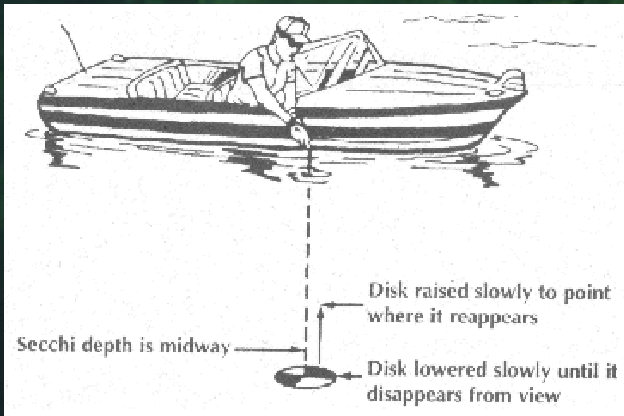
General considerations (2)

- Control of overfitting by bagging ("bootstrap aggregation")
  - a number of perturbed models are approximated then combined (by averaging) -> reduces variance component of prediction error
  - perturbed models <-> vary input data by bootstrap resampling
- Limnological domain is characterized by complex non-linearities
  - compare ANN with and without hidden layers

Input selection

- 6 water quality databases suggested 4 suitable inputs (important causal factors for algal growth and widely available)
  - T (°C) : drives rate of chemical/biological processes
  - [PO_4] (mg/L) : limiting factor for phytoplankton growth (freshwater)
  - [NO_3] (mg/L) : another nutrient, important in tropical water bodies
  - Secchi Disk Depth SDD (m)
    - * relates to phytoplantkon photosynthesis
    - * competition of phytoplankton species affected by underwater light
- Other important parameters omitted (lack in databases)
  - limit prediction to overall phytoplankton abundance (all species)

Secchi disk depth



http://www.mlswa.org/secchi.htm

Table 1
Six freshwater bodies: water quality, morphometry and database information

	Lake Biwa (Japan)	Lake Burrinjuck (Australia)	Darling River (Australia)	Lake Kasumigaura (Japan)	Myponga Reservoir (Australia)	Lake Soyang (Korea)
Water Quality						
Chl a µg/l						
Mean	9.32	7.54	21 900*	62.26	6.64	3.85
Max	38.5	30.5	196 000*	165	27.2	23.3
Water temperature °C						
Mean annual min	4.9	9.1	9.7	4.5	9.7	5.1
Mean annual max	29.5	25.6	27.2	28.8	22.3	27.0
Mean secchi depth m	1.76	1.55	101**	0.84	5.09**	3.90
Trophic state	eu-	meso-	hyper-	hyper-	meso-	meso-
Morphometry						
Depth m						
Maximum	103	63.5	n/a	7	36	118
Mean	41	56.6	n/a	4	not avail.	35.3
Area km^2	670	4.2	n/a	220	not avail.	46.5
Volume *10^6 m^3	27 800	756	n/a	900	26.8	1650
Retention time years	5.5	>2	0.002	0.55	<1	0.77
Database						
Time series length year	8	18	12.5	10.5	12	11
Number of records	103	156	375	120	156	97

* Chlorophyll a data not available. Cells/ml used instead.
** Secchi disc depth data not available. Turbidity (NTU) used instead.

2 Neural Network models

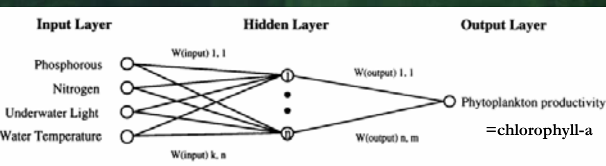
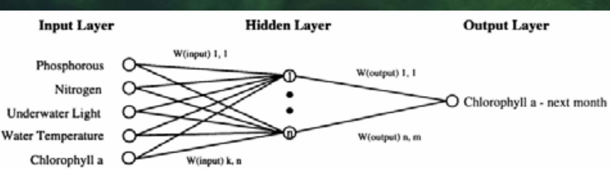


Fig. 2. Structure 1. ANN structure for same day model.

Predicts today's average algal abundance from today's inputs (irregular samplings)



Monthly averages Fig. 3. Structure 2. ANN structure for 30-days-ahead model.

Predicts next month average algal abundance from this month's average inputs & algal abund.

Training

- Stuttgart Neural Network Simulator (www-ra.informatik.uni-tuebingen.de/SNNS)

Table 2
Training variables

Tuning feature	Value
Learning algorithm	SCG (as implemented by SNNS 4.1 software package)
Weight update mode	Batch (offline)
Input scaling	Mean = 0; Standard deviation = 1.
Transfer function	Sigmoid
Weight initialisation range	Random from −0.1 to 0.1
Time to convergence on training set	1–5 s (Intel P6 class CPU)

Model selection

xperimental design	180 experiments		
atabases	Models	Number of hidden nodes	Normalised error of stopped training
ake Kasumigaura	same-day	0	0
ake Biwa	30-days-ahead	2	0.5
ake Burrinjuck		5	1.0
lyponga reservoir			1.5
arling river			2.0
ake Soyang			

- Bootstrap resampling to select training and test data : 100 times
  - Test performance of each bootstrap model
  - Bootstrap aggregate (bagged model)
    - * Min, max, mean, median, and interquartile ranges of test set predictions calculated for each value in the sample
    - * Visual assessment of mean model predictions vs observed

Results

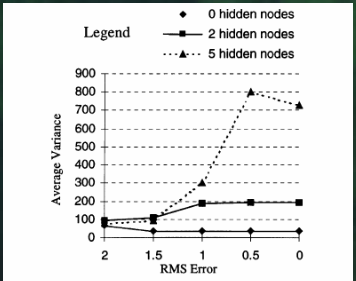


Fig. 4. Average variance per test set record vs. stopping criteria.

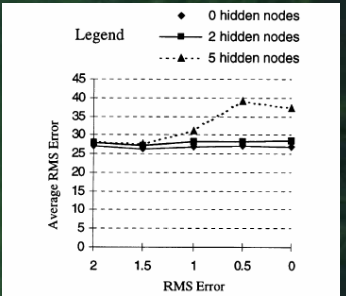


Fig. 5. Average test set RMS error vs stopping criteria.

Overfitting with 5 hidden layers (all but Lake Burrinjuck)

Stopping error

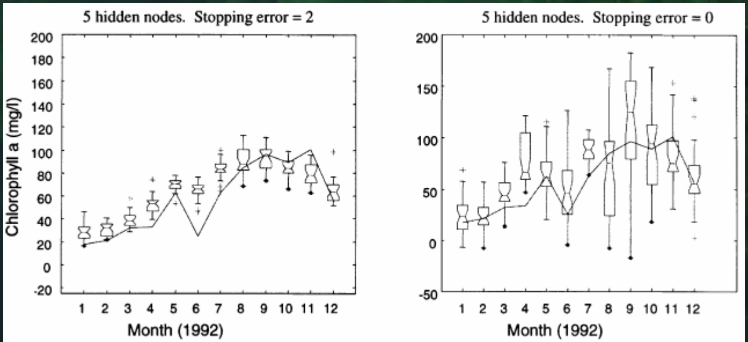


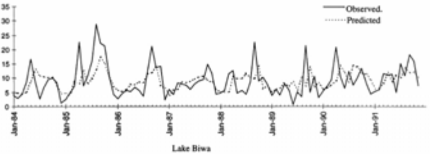
Fig. 6. Distribution of test set predictions. Lake Kasumigaura same day model 1992.

High variance but close to true value

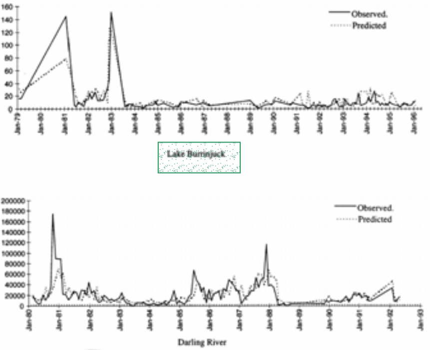
Optimum models

Table 4
Optimum selected bootstrap aggregate models

	Same-day model			30-days-ahead model		
Lake	No. hidden	Stop error	Correlation	No. hidden	Stop error	Correlation
Biwa	5	0.5	0.55	0	0.5	0.45
Burrinjuck	5	0	0.55	5	0	0.77
Darling	5	1.5	0.67	2	0.5	0.60
Kasumigaura	2	0	0.73	0	2	0.58
Myponga	2	1	0.62	2	0	0.64
Soyang	5	0.5	0.58	2	0	0.27



(Bootstrap aggregates predictions)



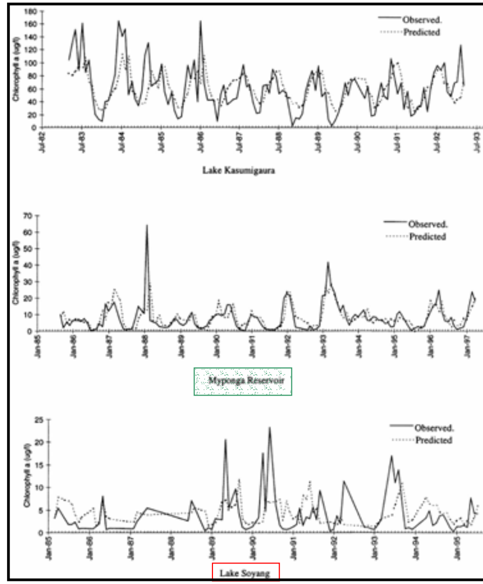
(a) Observed and predicted phytoplankton level v time. Bootstrap aggregate 30-days-ahead model. Observed and predicted phytoplankton level v time. (b) Bootstrap aggregate 30-days-ahead model.

H. Wilson, F. Recknagel

30-day model

- 2 good predictions
- other models miss some onsets, durations or magnitude, and tend to under-estimate the peak biomass

- good onset, duration, magnitude
- time-delay model is better

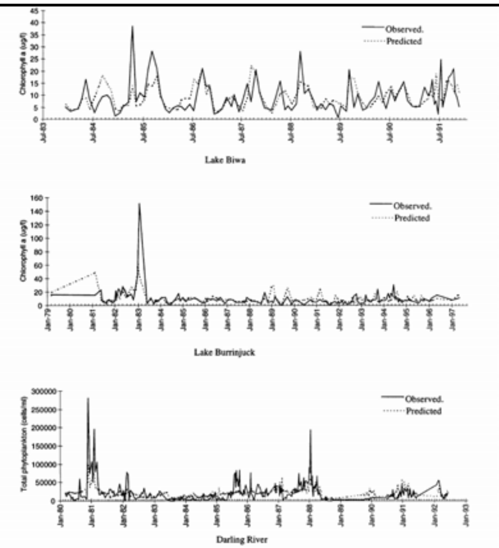


H. Wilson, F. Recknagel

30-day model

- good onset, duration, magnitude
- time-delay model is better

false positives

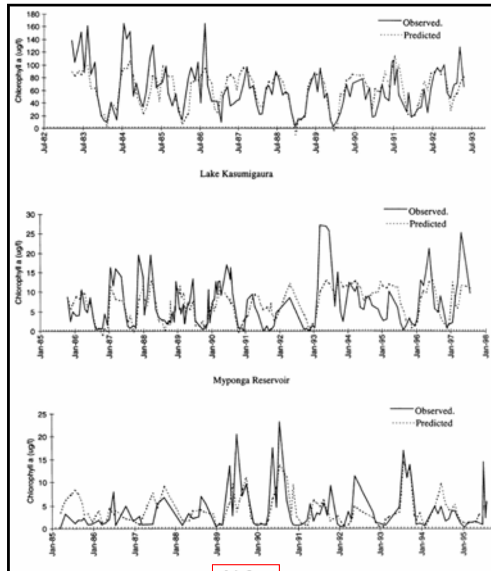


8. (a) Observed and predicted phytoplankton level vs. time. Bootstrap aggregate same-day model. (b) Observed and predicted phytoplankton level vs. time. Bootstrap aggregate same-day model.

H. Wilson, F. Recknagel

1-day model

Better model for lakes Biwa, Kasumigaura and Soyang



H. Wilson, F. Recknagel

1-day model

false positives

Conclusions (II)

Hidden nodes didn't reduce test set RMS error of boostrap aggregate models (except Myponga and Burrinjuck reservoirs)

Myponga and Burrinjuck benefited from time-delay
- temperate climates, oligotrophic - mesotrophic : limited algal growth
- ≠Warmer conditions -> eutrophic to hypertrophic (summer, monsoon)

For other sites, visual evaluation suggested hidden layers did improve prediction compared to linear prediction

Part III

Yuanzhi Zhang, Jouni Pulliainen, Sampsa Koponen and Martti Hallikainen,

Application of an empirical neural network to surface water quality estimation in the Gulf of Finland using combined optical data and microwave data.

Remote Sensing of Environment, 2002, vol. 81, no. 2-3, pp. 327-336.

Background

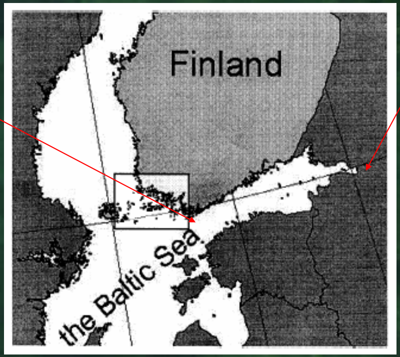
- How to relate RS to water quality parameters ? (transfer function)

- Oceanic water : light attenuation <-> phytoplankton pigments
- **Coastal** : inorganic, or dissolved/particulate organic materials
  - Optically complex situation (penetrates water)

- Radar - backscattering relates to surface roughness
  - geometry of water surface
  - material on the water surface
  - permittivity (dielectric constant) of water top layer

Study site - coastal archipelago

Rest of the Gulf :
Phosphorous-limited
Low salinity



Neva estuary :
Nitrogen-limited
Lower salinity

Shallow waters (38m average), eutrophied (lots of nutrients - anthropogenic = human activity ?)
Factors causing light attenuation (organic matter, phytoplankton, suspended sediment) vary
patially and temporally

3 data sources at a same date - 16 August 1997

- (1) Optical : Landsat TM (7 bands, 30m resolution, 8 bits)
- (2) Radar : ERS-2 (1 band, 12.5/25m resolution, 16 bits), 1h later
- (3) "Ground truth" : water samples (surface 0-0.5m) by Muikku
  - –Chl-a (Chlorophyll-a)
  - –Suspended sediments (SSC)
  - –Turbidity ("how clear water is")
  - –Secchi disk depth (transparency)

Table 1
Description of water samples used in the study (53 samples)

	Min	Max	Mean	Un
Chl-*a*	2.0	7.7	4.14	µg
SSC	1.6	11.0	4.03	mg
Turb	1.0	7.5	2.59	FN
SDD	0.67	4.2	2.60	m

- Only 53 samples (clouds, land…) analyzed within 4-10h
- Extract *observations* from TM/ERS 300m x 300m windows (μ, σ)
- External conditions : 0.39m waves, 5.5 m/s wind, 19.5°C water

Gulf of Finland, MERIS, 300m resolution, 17 July 2003 , ESA

Correlation analysis

- Turb <-> Chl-a (plankton biomass) : 0.06
- Turb <-> SSC (sediments) : 0.81
- SDD <-> Chl-a : 0.31, SDD <-> SCC : 0.49
  - –SDD significantly correlated to dissolved/particulate organic matter
- Next table : correlations between measurements and RS data
  - –Bands
  - –Transforms of bands
  - –[Linear combinations of those transforms…]

Correlations

Zhang, Pulliainen, Koponen, Hallikainen

able 2

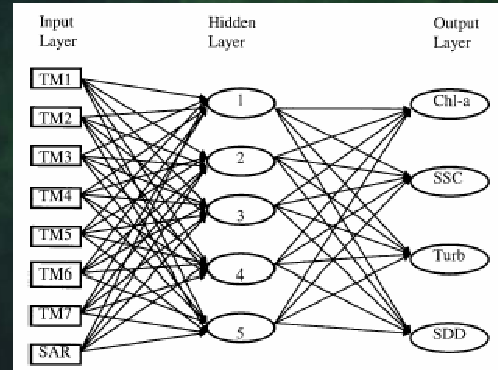
orrelation (r^2) among Chl-*a*, SSC, ln(SSC), Turb, and SDD with ratios, logarithmic transformations, and some combinations of TM bands and SAR data

	TM1 B	TM2 G	TM3 R	TM4 NIR	TM5	TM6	TM7	ERS-2	TM1/2
hl-*a*	.024	.003	.026	.019	.020	.012	.008	.357	.198
SC	.376	.467	.533	.284	.057	.007	.032	.098	.226
(SSC)	.339	.459	.503	.245	.043	.003	.022	.090	.284
urb	.476	.626	.664	.391	.044	.019	.023	.055	.323
DD	.204	.383	.530	.460	.164	.064	.073	.368	.354
	TM1/3	**TM1/4**	**TM2/1**	**TM2/3**	**TM2/4**	**TM3/1**	**TM3/2**	**TM3/4**	**TM4/1**
hl-*a*	.322	.129	.204	.239	.005	.358	.263	.017	.215
SC	.253	.067	.236	.139	.231	.327	.164	.406	.058
(SSC)	.276	.070	.299	.125	.260	.359	.148	.420	.050
urb	.292	.085	.362	.121	.313	.400	.136	.497	.080
DD	.537	.007	.411	.382	.071	.651	.416	.291	.005
	TM4/2	**TM4/3**	**TM3 − 4**	**TM2 − 4**	**TM2 + 3**	**ln(TM1)**	**ln(TM2)**	**ln(TM3)**	**ln(TM4)**
hl-*a*	.039	.006	.025	.002	.007	.071	.0001	.021	.023
SC	.175	.345	.534	.465	.493	.280	.371	.464	.256
(SSC)	.191	.367	.506	.458	.478	.255	.377	.461	.224
urb	.258	.428	.661	.622	.646	.391	.528	.603	.370
DD	.043	.245	.516	.374	.429	.116	.265	.460	.442
	TM1/(1+2+3)	**TM2/(1+2+3)**	**TM3/(1+2+3)**	**TM1/(1+2+4)**	**TM2/(1+2+4)**	**TM4/(1+2+4)**	**TM2/(2+3+4)**	**TM3/(2+3+4)**	**TM4/(2+3+4)**
hl-*a*	.272	.111	.363	.251	.165	.119	.251	.246	.018
SC	.281	.156	.292	.213	.257	.117	.039	.219	.224
(SSC)	.337	.215	.304	.273	.319	.118	.029	.202	.242
urb	.383	.263	.317	.317	.380	.171	.091	.199	.312
DD	.498	.227	.616	.404	.380	.006	.230	.471	.079

RS-2 = ERS-2 SAR, TM1/2 = TM1/TM2, TM3 − 4 = TM3 − TM4, TM1/(1 + 2 + 3) = TM1/(TM1 + TM2 + TM3), etc.

Zhang, Pulliainen, Koponen, Hallikaine

Empirical Neural Network



Scaling 0-1

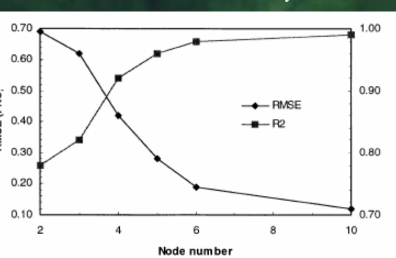
$$Y = a\tanh\left(\sum_{h=1}^{j} w_k z_k + b\right)$$

Trained 10 times per configuration, random initialisation

Zhang, Pulliainen, Koponen, Hallikainen

Results

-> 5 nodes in hidden layer



RMSE for turbidity

27 samples training, 26 testing

Table 5

Comparison of neural network training and testing

	Neural network training		Neural network testing	
	TM	TM + SAR	TM	TM + SAR
Chl-a (μg/l)				
R^2	.902	.913	.845	.970
RMSE	0.330	0.312	0.651	0.284
SSC (mg/l)				
R^2	.847	.973	.921	.986
RMSE	0.638	0.265	0.679	0.283
Turb (FNU)				
R^2	.945	.990	.942	.984
RMSE	0.248	0.104	0.401	0.208
SDD (m)				
R^2	.927	.931	.969	.977
RMSE	0.175	0.170	0.147	0.126

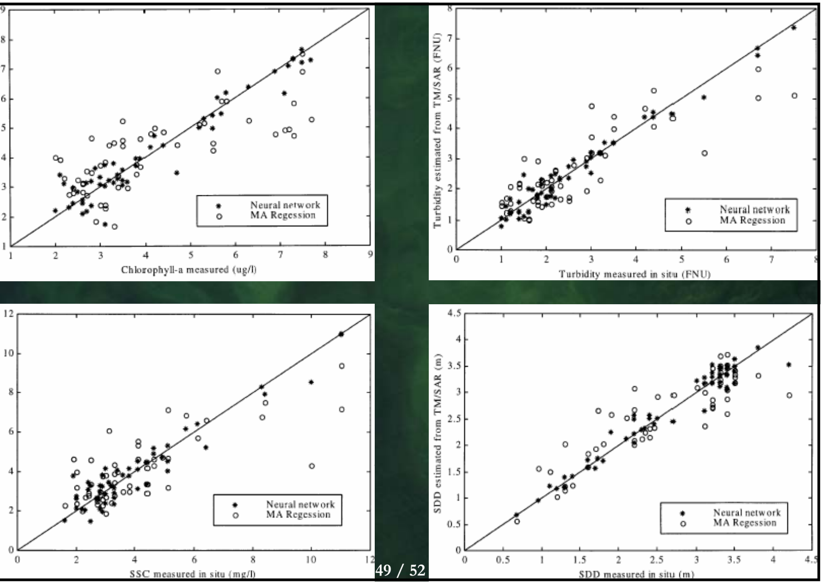
Zhang, Pulliainen, Koponen, Hallikaine

Correlation analysis vs Neural Network

Table 4

Comparison of regression analysis and neural network simulation

	K.O. Regression analysis		Neural network	
	TM	TM + SAR	TM	TM + SAR
Chl-a (μg/l)				
R^2	.65	.67	.90	.92
RMSE	0.99	0.96	0.53	0.47
SSC (mg/l)				
R^2	.54	.55	.89	.91
RMSE	1.46	1.45	0.72	0.65
Turb (FNU)				
R^2	.69	.69	.94	.96
RMSE	0.82	0.82	0.35	0.28
SDD (m)				
R^2	.72	.75	.92	.95
RMSE	0.45	0.43	0.25	0.19



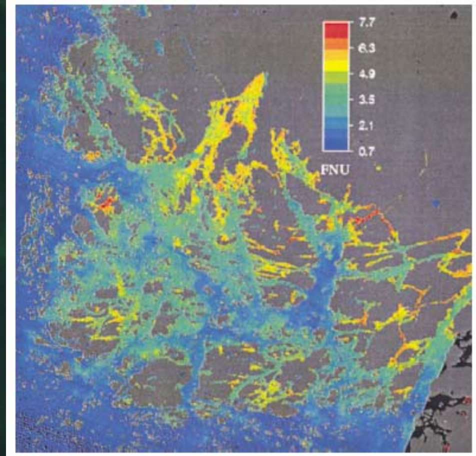


Fig. 8. An example of Turb-retrieved image (81 × 81 km) in the study area.

Conclusions (III)

- NN models quite well non-linear transfer function between RS and water quality
- SAR is a good supplement to optical data
- **Yuanzhi Zhang, *Surface Water Quality Estimation Using Remote Sensing in the Gulf of Finland and the Finnish Archipelago Sea*, D. Sc. Tech. thesis available at :**
  http://lib.tkk.fi/Diss/2005/isbn9512277190/

Links

- Baltic Sea portal
  http://www.fimr.fi/en/itamerikanta.html
- Monitoring algae bloom
  http://www.fimr.fi/en/itamerikanta/levatiedotus/levakartat.html
  http://www.eea.eu.int/Highlights/20030811104233/algalbloom
- Remote sensing of water quality in lakes and coastal waters
  http://www.ymparisto.fi/default.asp?contentid=85417&lan=en