



Politecnico di Milano  
*Dipartimento di Elettronica ed Informazione*  
Dottorato di Ricerca in Ingegneria dell'Informazione

---

# Environmental Modelling via learning-from-data techniques

A Ph.D. Dissertation by:  
**Giorgio Corani**

Advisor:  
**Prof. Giorgio Guariso**

Tutor:  
**Prof. Sergio Rinaldi**

Supervisor of the Ph.D. Program:  
**Prof. Stefano Crespi Reghizzi**

---

**Giorgio Corani:**

*Environmental Modelling via learning-from-data techniques.*

Dipartimento di Elettronica e Informazione.

Politecnico di Milano.

Ph.D. Thesis, 2005.

**Giorgio Corani**

Dipartimento di Elettronica e Informazione

Politecnico di Milano

Via Ponzio 34/5

20133 Milano, Italy

corani@elet.polimi.it

# Abstract

The thesis deals with learning algorithms, and how they can be applied to environmental modelling. A major separation can be drawn within learning problems, depending on whether linear or nonlinear approximating functions are considered.

In the linear case, the key-problem is the selection of the supposedly best model from among the set of candidates. The contribution of the thesis on this topic is an exhaustive comparison between a number of traditional model selection criteria, typically derived under restricting asymptotic assumptions, and the Structural Risk Minimization (SRM) approach; SRM, derived under hypotheses of great generality, is almost unknown in fields others than machine learning, and it is used in this thesis for the first time with time series. The investigation has been performed both on artificially generated time series and real datasets. Under practically all the huge amount of simulations investigated, our findings show that SRM recognizes with the highest probability the true model underlying the data, and that it also leads to the lowest prediction error in out-of-samples simulations. As worked case studies, several datasets of animal populations already known in the ecological literature are re-analyzed; models selected by SRM are shown to provide lower prediction errors on out-of-samples simulations with reference to the models selected in previous works.

As for nonlinear learning algorithms, different approaches are rigorously compared in environmental case studies of great practical interest, such as flood forecasting and prediction of air pollutants. In particular, several algorithms alternative to feed forward neural networks (FFNNs), recognized as state-of-the-art approach, have been analyzed, trying to get rid of their known problems (such as overparameterization and difficulty to assess the relevance of the different

---

input variables) and to improve the quality of the generalization.

As for flood prediction, two Italian basins have been considered as case studies. We show that pruned neural networks (PNNs) allow to detect irrelevant rain gauges by removing their parameters from the architecture of the network and that, despite the use of a smaller set of information, PNNs provide the same predictive accuracy than FFNNs. This allows for instance to move unnecessary instruments to more urgent locations if one is interested in improving the design of the monitoring network, minimizing the costs; on the other hand, the pruned predictor is more robust, since it requires polling a smaller number of gauges, and therefore it is less subject to downtimes due to data acquisition failures. We show that the best predictive accuracy has been however obtained, in both basins, by using a novel neuro-fuzzy (NF) framework, which constitutes an original algorithmic contribution of the thesis. Such a framework is based on a set of local neural networks, each specialized on a certain condition of basin saturation. It is known, in the hydrological literature, that the nonlinear form of the rainfall-runoff relationship strongly changes, depending on the state of saturation of the basin.

As for air pollution, we analyzed the prediction of  $PM_{10}$  and  $O_3$  in Milan; such two pollutants constitute a major concern for the air quality of the city. In this case, we compare FFNNs and PNNs with the lazy learning (LL) approach, which is known to be suitable for time series prediction, and which is used here for the first time in environmental modelling. LL is a local linear modelling approach; it gives linearity a chance via locality, allowing to reuse a large amount of procedures taken from the linear statistics. The results obtained on both  $PM_{10}$  and  $O_3$  show that all the considered models techniques provide a satisfactory prediction reliability; for instance, correlation between true and predicted values are around 0.9. What deserves particular consideration, besides the performances issues, is however that, differently from any neural network, LL allows to easily interpret the relevances of the different inputs; moreover, it is much faster to design and easier to be kept up-to-date.

Challenges for future studies are given by the use of SRM for model selection also among nonlinear models, and by the improvement of the selection of the linear regressor within Lazy Learning using the Local

---

Risk Minimization approach, which is an extension of SRM to local modelling.

# Contents

<b>Riassunto</b>	<b>iv</b>
<b>Abstract</b>	<b>vii</b>
<b>1 Introduction</b>	<b>4</b>
1.1 Learning unknown relationships from data . . . . .	5
1.1.1 Environmental case studies . . . . .	7
1.2 Model selection criteria . . . . .	9
1.2.1 Model selection in ecological modelling . . . . .	10
1.3 Publications . . . . .	13
<b>2 Learning from data algorithms</b>	<b>16</b>
2.1 Soft Computing Techniques . . . . .	17
2.1.1 Feed-forward neural networks (FFNN) . . . . .	19
2.1.1.1 Backpropagation . . . . .	21
2.1.1.2 Local approximation of square error . . . . .	22
2.1.1.3 Levenberg Marquardt (LM) training algorithm . . . . .	23
2.1.1.4 Early stopping and regularization . . . . .	25
2.1.1.5 FFNN architecture selection . . . . .	26
2.1.2 Pruned neural networks (PNN) . . . . .	26
2.1.2.1 Optimal Brain Surgeon pruning algorithm . . . . .	28
2.1.3 Lazy learning (LL) . . . . .	32
2.1.3.1 Considerations on Lazy Learning vs. global approaches . . . . .	34
2.1.4 A neuro fuzzy framework for time series prediction	35

2.1.4.1	LM training algorithm for weighted least square (LM-wls) . . . . .	38
2.1.4.2	Local models identification and prediction computation . . . . .	39
2.1.5	Performances assessment via cross validation . . . . .	40
2.2	Model selection criteria . . . . .	41
2.2.1	VC-dimension . . . . .	43
2.2.2	Model selection problem statement . . . . .	45
2.2.3	PBLR (Parametric Bootstrap Likelihood Ratio) . . . . .	48
<b>3</b>	<b>Environmental modelling via soft computing</b> . . . . .	<b>50</b>
3.1	Air pollution case studies . . . . .	51
3.1.1	Ozone prediction . . . . .	54
3.1.1.1	Input variables significance . . . . .	58
3.1.2	PM <sub>10</sub> prediction . . . . .	60
3.1.2.1	Input relevances . . . . .	63
3.1.3	Discussion . . . . .	64
3.2	Hydrological prediction . . . . .	65
3.2.1	Neural network modelling of the rainfall runoff relationship . . . . .	69
3.2.2	Rainfall time delays configuration . . . . .	70
3.2.3	A remark on architecture selection for one-step-ahead recursive prediction . . . . .	71
3.2.4	Basin saturation issues . . . . .	72
3.2.5	Performances assessment . . . . .	74
3.2.6	Olona River . . . . .	75
3.2.6.1	Predictors analysis . . . . .	77
3.2.7	Tagliamento River . . . . .	82
3.2.7.1	Predictors analysis . . . . .	84
3.2.8	Discussion . . . . .	88
<b>4</b>	<b>Stat. Learning Theory for model selection</b> . . . . .	<b>90</b>
4.1	Experiments on density-dependence detection . . . . .	95
4.1.1	Drift model recognition . . . . .	96
4.1.2	Ricker model recognition . . . . .	98
4.1.3	Discussion . . . . .	102
4.2	Model selection within a suite of candidate models . . . . .	102

*Contents*

---

4.3	Results . . . . .	107
4.4	Discussion . . . . .	112
4.5	The <i>Alpine ibex</i> case study . . . . .	113
	4.5.1 The findings of Jacobson et al. [27] . . . . .	115
	4.5.2 Analysis by SRM . . . . .	116
4.6	Discussion . . . . .	117
<b>5</b>	<b>Conclusions</b>	<b>121</b>
	<b>Bibliography</b>	<b>127</b>

# Chapter 1

## Introduction

Modern science and engineering are based on using *first-principles* models to describe physical, biological and social systems. Such an approach start with a formalized law of the phenomenon to be described (as for instance the Newton's law of mechanichs); then, experimental measurements are used to verify the appropriateness of the models, or to estimate model parameters especially hard to measure. However, in many applications physical systems are too complicated to be described in their first principles; in these cases, a viable alternative approach is constituted by *learning-from-data* methods [1], which build the model by estimating a suitable approximating function on the available input/output samples. Once such a relationship has been estimated and validated, it can be used for the prediction of the future system behavior. Generally, learning approaches are *tabula rasa* methods, since they learn from scratch from the training data without any initial knowledge besides the representation of hypotheses [2].

Usually, learned models can be developed more quickly than first-principles (or *physically-based*) models, obtaining at the same time a high approximation quality. Clearly, their main drawback is the limited explanation of the system under study provided to the investigator, since they lack an explicit representation of the causalities of the real system. However, their effectiveness in tasks such as for instance input/output simulation, time series prediction, classification is nowadays recognized from large part of the scientific community; in fact, they have been increasingly used in different research areas over

the last years, also thanks to the availability of low-cost computers with great computational power.

In particular, techniques such as neural networks, genetic algorithms, fuzzy systems are thoroughly studied also within the community of environmental scientists, and for instance the International Environmental Modelling and Software Conference ([www.iemss.org](http://www.iemss.org)) or the Environmental Applications of Machine Learning Conference ([www-ai.ijs.si/SasoDzeroski/ECEMEAML04/ecem.html](http://www-ai.ijs.si/SasoDzeroski/ECEMEAML04/ecem.html)) pay great attention to advances in learning approaches.

This thesis collects the results of three years of PhD research, devoted to the application of different learning methods (neural networks, fuzzy logic, lazy learning, statistical learning theory) to different environmental case studies, such as flood forecasting, air quality prediction, analysis of time series of animal populations.

## 1.1 Learning unknown relationships from data

If the system under study is complex and its constitutive input/output relationship is unknown at all, one has to randomly generate many different nonlinear input/output mapping functions, until a satisfactory degree of approximation is reached. Learning unknown relationships from data actually constitutes the key problem addressed by Soft Computing techniques.

Among them, feed forward neural networks (FFNN) are recognized to constitute a state-of-the-art approach; their flexibility and their ability in capturing the non-linearities underlying the data are nowadays well-known also outside the machine learning community. As for ecological modelling, for instance, a review of neural networks applications in water resources and atmospheric sciences can be found respectively in [3] and [4]. However, FFNN have also well known drawbacks, such as the tendency to overfitting because of the heavy parameterization, and the time consuming procedures required in order to identify via trial and error the optimal architecture for a given modelling task. Moreover, their “very black-box” nature makes a hard task the interpretation of the model parameters; for instance, it is

quite difficult to assess the relevances of the input variables within the network.

The need for addressing the above cited drawbacks and the attempt of further improving the quality of the approximation lead to the adoption of alternative learning approaches; in particular, in this thesis we investigate pruned neural networks (PNN) and lazy learning (LL). Moreover, we introduce a novel neuro-fuzzy framework, based on a set of local neural networks, which constitutes an original algorithmic contribution of the thesis.

Pruned neural networks (PNN) constitute a recognized research field in the machine learning area (see [5] for a review); nevertheless, they are still not widely used for applications. The basic idea of pruning algorithms is to remove the redundant parameters from a fully connected neural network; pruned networks can contain one order of magnitude less parameters than fully connected ones and, as such, they are much less prone to overfitting, constituting a parameter-parsimonious neural networks approach. Remarkably, unuseful inputs are disconnected from the network during the pruning session, and hence the really meaningful input variables are highlighted.

The proposed novel neuro-fuzzy framework constitutes an original algorithmic contribution of the thesis, suitable for modelling systems characterized by very different forms of the input/output approximating function in different regions of the input space. The framework is based on a set of local neural networks, each providing a map between inputs and output on a certain region of the input space. The outputs of all the local models are combined in a fuzzy way, in order to have smooth transitions between the local models. Although combinations of FFNN have been already subject of investigation [6] [7], the peculiarity of our approach is that the coefficients of the combination are not fixed: on the contrary, they vary dynamically according to the values of the provided inputs.

While FFNN and PNN are global, nonlinear methods, Lazy Learning (LL) is a local linear modelling approach. Probably, the most relevant contribution to LL development and diffusion has been done by the research group working at IRIDIA (<http://iridia.ulb.ac.be/~lazy>), which continuously works over LL algorithmic enhancements and applications, and which also releases the LL implementa-

tion as open-source code. LL outperformed [8] several alternative approaches (including FFNN) in a number of benchmark datasets available from the UCI repository (<http://www.ics.uci.edu/~mlearn/MLRepository.html>) and therefore it appears as a suitable approach for time series prediction. Moreover, it is worthwhile citing that it ranked second out of 17 participants at the International Competition for Time Series Prediction, held in Leuven in 1998 [9]. The development of a LL model is much shorter than in the neural networks case, since the (local) linear nature of the model greatly reduces the need for trial and error; moreover, LL models can be updated in a very fast and easy way. A final relevant strength of such an approach is that the simplicity of the local linear regressor allows to easily evaluate the relevances of the different input variables.

### 1.1.1 Environmental case studies

The Soft Computing methods introduced in Section 1.1 are applied to real world case studies in environmental modelling, such as hydrological prediction, air quality prediction, air quality recovery planning. In each case study, we identify first a FFNN model, which then constitutes the fundamental benchmark, and then we assess the performances of the alternatives methods.

As for the hydrological prediction, two different basins have been considered: Olona and Tagliamento. River Olona is located in Lombardia, Northern Italy; it floods areas around Milan quite often under heavy rainfalls and the study of such a catchment has been encouraged also by the Civil Protection. The average flow is about  $2.5 m^3/sec$ , while the maximum expected flow over a time period of 10 years is about  $108 m^3/sec$ . The basin sizes about  $200 km^2$ ; it can be divided in an upper part, mountainous and weakly anthropized, and a lower one, flat and strongly urbanized. Many artificial inflows in the lower part of the basin and a series of small reservoirs built in order to mitigate flood events alter significantly the natural behavior of the basin. The data available to us refer to 13 flood episodes occurred within the period 1999-2001; the dataset comprises about 1100 water levels and rainfall measures taken at hourly steps. Given the small size of the basin, a forecast horizon of three hours is judged as suitable by Civil

Protection technicians.

The second case study refers to river Tagliamento, located in Friuli, North-East of Italy. The average flow is about  $90 m^3/sec$ , while the maximum flood peak over the last decades reached  $4000 m^3/sec$  in 1966. The basin measures about  $1950 km^2$ ; the dataset comprises 20 flood events occurred over the years 1978-1996, for an overall length of 2000 hourly time steps. The dataset has been obtained by contacting the authors of a previous paper [10], who proposed a feed-forward neural network for the prediction of flood events. Therefore, the performances of such a model will constitute the natural term of comparison for the obtained results. According to [10], a forecast horizon of 5 hours in advance can be considered satisfactory on this basin.

As for air pollution prediction, the case of ozone and PM10 time series in Milan has been studied. The Milan urban area, located at the center of the Po Valley, is the most industrialized and populated district in Italy. According to the State of the Environment Report [11], the yearly average of pollutants such as SO<sub>2</sub>, NO<sub>x</sub>, CO, TSP has decreased respectively of about 90%, 50%, 65%, 60% during the last decade. Also the yearly averages [11] of micro-pollutants such as benzene and lead are largely under the thresholds established for human health protection. However, a major concern for the air quality of the city is constituted by high levels of both PM10 and ozone; these pollutants constitute a major concern since they have been associated in the epidemiological literature with increase in the mortality and cardiorespiratory hospitalizations [12, 13]. The yearly average of PM10 has been substantially stable (about  $45 \mu g/m^3$ ) since the beginning of monitoring in 1998; on average, it exceeds the limit value of  $50 \mu g/m^3$  on about 100 days every year. On the other hand, ozone began to rise in the Milan area in the early 90s, partly as a consequence of the reduced SO<sub>2</sub> and CO, which caused a more oxydant atmosphere. In the period 1997-2002 about 10 yearly exceedances of the attention threshold have been recorded. A system able to predict ozone and PM10 concentrations can provide useful early warnings, allowing Public Authorities to manage the emergency, for instance by planning an increase in the public transports in the case of an incoming traffic block, or by warning people to avoid exposures to unhealthy air.

## 1.2 Learning from data the best model: model selection criteria

Soft Computing techniques assume generally that the input/output relationship is unknown at all, and therefore it has to be found by some kind of random search, trying many different, complex, nonlinear approximating functions. However, there also different situations, where the system can be modeled by quite simple (i.e., linear or linearizable by suitable transformations) input/output approximating functions; in such cases, model selection can be accomplished by analytical techniques, which provide an estimate the prediction error of the model on future unknown data. The mathematical formulation of such estimates depend on the assumed hypotheses and hence different *model selection criteria* are referenced in the literature.

A classical approach is constituted by Information Criteria, which try to achieve an optimal trade-off between the quality of the approximation on the training data and model complexity by choosing the model that minimizes the product of the training square error and a penalization factor, computed as an increasing function of the ratio of parameters of the model to the number of data. For instance, criteria such as FPE [14] and SIC [15] are widely referenced in the literature; it should be however pointed out that their very restricting constitutive hypotheses are generally not met in real case studies.

As a viable alternative to classical approaches, the model selection criterion developed within Statistical Learning Theory (SLT), and defined as Structural Risk Minimization (SRM), is investigated in the thesis. SLT, due to the joint work of Vapnik and Chervonenkis [16], provides a theoretical framework for learning with finite samples derived under very general assumptions. The very core of SLT is the idea of VC-dimension, which is a complexity index for classes of functions. VC-dimension can be easily known for linear models, but it is generally unknown for nonlinear models; this constitutes in fact a major obstacle to the application of SLT findings in nonlinear contexts. With reference to linear regression problems, it has been shown [17] to consistently outperform traditional Information Criteria for different dataset sizes and noise levels, choosing with higher probability model with better *generalization* performances, i.e. with better performances

on predictions issued outside the calibration dataset. However, despite its relevant properties, SRM has been up to now rarely used in real case studies.

### 1.2.1 Model selection in ecological modelling

The application area of model selection criteria considered in this thesis is constituted by the analysis of the observed course of ecological population abundances. Indeed, in these cases, one usually considers a limited set of demographic models, which quite often can be linearized by suitable transformations (e.g., taking logarithms). However, we do not consider in this application non-linearizable demographics, because neither they can be managed by ICs, nor their VC-dimensions is known. An attempt to estimate the VC-dimension of nonlinear demographic models can be however found in [18].

T. R. Malthus, the founder of modern demography, in his famous work of 1798 [19] proposed a simple linear model, according to which the population can increase indefinitely in an exponential way or tend to extinction. The main assumption underlying the Malthusian model is that the environment can provide each individual with the same amount of resources necessary to survival and reproduction, regardless of the population density; this is actually the *density-independence* hypothesis.

However, no population grows indefinitely; as the density rises, some competition takes place between individuals (for example for food, water, or reproduction), slowing down or halting the population increase, which therefore depends in this case on the population size (*density-dependence*). For instance, a density-dependent population can move towards an equilibrium (*carrying capacity*), fluctuating around it over time.

To recognize whether a population is growing in a density-dependent or independent way is important since it can allow to predict the future animal abundances, which is of paramount relevance in order to design correct policies for the sustainable exploitation of natural populations. In fact, to statistically distinguish density-dependent from independent time series stimulated a great research effort over the past three decades [20–22] in the ecological literature.

Earlier works were based on hypothesis-testing approaches; a milestone in this context is for instance the work of Dennis and Taper [22]. Despite their statistically soundness, hypothesis tests have the main drawback of comparing just a couple of models at a time and that on the other hand managing many models through hierarchical pairwise hypothesis testing does not necessarily lead to the selection of the best model [23]; therefore, they have been finally recognized to convey just limited information.

Further works [23–26] performed model selection among demographic models using Information Criteria, such as FPE or SIC; a weakness of such approaches is however that ICs are based on asymptotic arguments, while ecological datasets are composed generally of a just few tens of data.

The contribution of the thesis on this topic is a thorough comparison of the above traditional approaches (hypothesis tests and ICs) with SRM, the model selection criterion developed on the base of the findings of Statistical Learning Theory and up to now rarely used in real world case studies.

A great amount of experiments has been done on artificial data, simulating different demographic models stochastically (i.e., with noise), under a wide variety of settings (model parameters, noise levels, simulation length). On each generated time series a set of alternative models are identified, and model selection criteria are then asked to choose one among the candidates. This way we statistically assess the ability of the model selection criteria in recognizing the model underlying the data; then we also evaluate the generalization of the chosen models, simulating them on data generated by the same stochastic mechanism, but not included in the calibration samples.

Our findings show that SRM chooses the model which really underlies the data with significantly higher frequency than traditional approaches and, as a direct consequence, it allows also the best generalization outside the training set. Such a result is consistent in the sense that it is found under almost all the parametric settings investigated.

As real case study, the ungulates population of Alpine ibex (capra ibex) living in Gran Paradiso National Park (Italy), has been analyzed. The available dataset is unusually long for ecological applica-

## *1.2 Model selection criteria*

---

tions, containing a 45-years time series (1956-2000) of both censuses and meteorological data. Over this period, the population ranged between 2500 and 5000 individuals; there has been no hunting inside or outside the park, and large predators such as lynx and wolf have been absent over the last 100 years. Since several studies (see the literature overview in [27]) suggested that, if large predators are rare or absent, the changes in ungulates populations can be explained by considering climate forcing and density-dependence, we analyze thoroughly the dependence of the population growth rates on both population size and meteorological variables, and then we compare our results with the statistical analysis carried out in [27] on the same dataset.