Promoters:

Prof. Dr. ir. Peter L. M. Goethals
Prof. Dr. Niels De Pauw
Research Unit Aquatic Ecology
Department of Applied Ecology and Environmental Biology
Ghent University

Dean: Prof. Dr. ir. G. Van Huylenbroeck

Rector: Prof. Dr. P. Van Cauwenberge

Habitat suitability modelling of pike (*Esox lucius*) in rivers

MSc. Rahmat Zarkami

Thesis submitted in fulfilment of the requirements for the degree of
Doctor (PhD) in Applied Biology Sciences

Dutch translation of the title:

Habitatgeschiktheidsmodellering van de snoek (*Esox lucius*) in rivieren.

This work should be referred as:

Zarkami, Rahmat (2008). Habitat suitability modelling of pike (*Esox lucius*) in rivers. PhD thesis, Ghent University, Ghent, Belgium, 235 pp.

Cover illustrations:
Front: Isolated Scheldt meander with a very rich pike population.
Back: Ditch with a high potential for pike, but affected by agricultural activities.

ISBN 978-90-5989-272-9

The author and the promoters give the authorization to consult and to copy parts of this work for personal use only. The laws of copyright limit any other use. Permission to reproduce any material contained in this work should be obtained from the author.

Acknowledgements

First of all I would like to express my gratitude to my promoters, prof. Dr. ir. Peter Goethals and prof. Dr. Niels De Pauw, for their friendship, continuous support, expert advises and encouragement. During my study, the promoters gave me the opportunity to be part of the Aquatic Ecology Research Unit.

I am also indebted to the examination committee of this PhD dissertation for their critical comments and valuable advices; Chairman: Prof. Dr. Jean-Pierre Ottoy (UGent-FBW, BW10), Secretary: Prof. Dr. ir. Peter Bossier (UGent-FBW, BW13), Prof. Dr. ir. Olivier Thas (UGent-FBW, BW10), Dr. Alain Vandelannoote (Aquafin NV, Afdeling Milieu en Kwaliteit), Dr. ir. Wim Gabriels (VMM, Afdeling Rapportering Water).

Many thanks to Dr. Koen Lock (AECO group) for his help during the last phase of my PhD study.

I would like to thank Prof. Colin Janssen (Department Head) and to all friends of the Laboratory of Environmental Toxicology and Aquatic Ecology: Andy, Ans, Argaw, Brita, Diederik, Eleni, Emmy, Frederik, Gisèle, Guido, Jill, Karel, Leen, Lien, Marc, Marianne, Marjolein, Martine, Michiel Vandegehuchte, Michiel Claessens, Minh, Nele, Peggy, Peter, Spyros, Roel, Thu Huong, Tom, Veronique, Wim, Wouter and Yblin.

And last, but as always certainly not least, a special word of thanks to my wife Jila! I will certainly never forget her worries when my work was facing difficulties and her pleasure when my work was in progress. I now know that being a supporter is much harder and far more stressful than being a player. And of course, I would like to thank our son Ali who has always been a source of joy and inspiration for me.

Rahmat Zarkami
Ghent University, 12th December 2008

Table of contents ... 1
List of abbreviations ... 5

General introduction

Problem definition ... 9
Need for quantitative approaches to describe pike habitats ... 12
Scope and objectives ... 13

Chapter 1: Biology and ecology of pike

1.1 Introduction ... 19
1.2 Biological and ecological knowledge of pike ... 19
1.2.1 Taxonomy ... 19
1.2.2 Morphology ... 20
1.2.3 Origin, distribution and abundance ... 21
1.2.4 Growth rates ... 21
1.2.5 Life cycle of pike ... 22
1.2.6 Role of vegetation in the life cycle of pike ... 23
1.2.7 Feeding ... 24
1.2.8 Habitat requirements ... 25
1.3 Quantitative approaches to model pike populations ... 26
1.3.1 Habitat suitability modelling ... 26
1.3.2 Predator-prey modelling ... 33
1.3.3 Competition modelling ... 36
1.3.4 Food web modelling ... 36
1.4 Status of pike in Flanders, Belgium ... 38

Chapter 2: Materials and methods

2.1 Introduction ... 41
2.2 Monitoring and data collection methods ... 42
2.2.1 General characteristics of the river basins in Flanders and location of the monitoring sites for pike ... 42
2.2.2 River typology classification ... 45
2.2.3 Data collection methods ... 47
2.2.3.1 Available data concerning the river basins in Flanders for pike habitat suitability modelling ... 47
2.2.3.2 Abiotic and biotic charachtreistucs of Flemish river basins ... 48
2.2.3.3 Monitoring of pike in Flemish river basins ... 51
2.3 Database set-up and data analysis methods ... 52
2.3.1 Bandwidth and distribution of input and output variables ... 52
2.3.2 Correlation analysis between input and output variables ... 53
2.3.3 Visual relation analysis between the input and output variables ... 54
2.4 Modelling methods ... 55
2.4.1 General methodology ... 55
2.4.2 Dataset splitting for training and validation ... 59
2.4.2.1 General aspects of dataset splitting for training and validation ... 59

2.4.2.2 Dataset spliting and validation in thids study…….....60
2.4.3 Model evaluation 62
2.4.3.1 General overview of model evaluation 62
2.4.3.2 Selection of model evaluation procedures in this study 64
2.4.4 Ecological model development 64
2.4.4.1 Classification trees 65
2.4.4.1.1 General aspects of classification trees 65
2.4.4.1.2 Specific parameter settings in this study 68
2.4.4.2 Artificial neural networks 68
2.4.4.2.1 General aspects of artificial neural networks 68
2.4.4.2.2 Data processing and bandwidth 71
2.4.4.2.3 Learning method 71
2.4.4.2.4 Model architecture 71
2.4.4.2.5 Input variable selection 72
2.4.4.2.6 Number of hidden layers and hidden neurons 73
2.4.4.2.7 Model interpretation 75
2.4.4.2.8 Specific parameter settings in this study 75
2.4.4.3 Support-vector machines (SVMs) 76
2.4.4.3.1 General aspects of support-vector machines 76
2.4.4.3.2 Specific parameter settings in this study 79
2.4.4.4 Logistic regressions (LRs) 80
2.4.4.4.1 General aspects of logistic regressions 80
2.4.4.4.2 Specific parameter settings in this study 82
2.4.4.5 Genetic algorithms (GAs) 83
2.4.4.5.1 General aspects of genetic algorithms 83
2.4.4.5.2 Specific parameter settings for this study 84
2.5 Application of habitat suitability models for decision support in river management 85

Chapter 3: Database set-up and exploration

3.1 Introduction 91
3.2 Results 91
3.2.1 Bandwidth and distribution of input and output variables 91
3.2.2 Correlation analysis between input and output variables 95
3.2.3 Visualisation of the input and output variables 98
3.3 Discussion 103
3.4 Conclusion 104

Chapter 4: Development of habitat suitability models for pike using classification trees in combination with genetic algorithms

4.1 Introduction 107
4.2 Results 108
4.2.1 Assessment based on model performance criteria 108
4.2.2 Ranking of major input variables for pike habitat requirements by classification trees 111
4.2.3 Variables selected by genetic algorithms 113
4.3 Discussion 117
4.4 Conclusion 120

Chapter 5: Development of habitat suitability models for pike using artificial neural networks in combination with genetic algorithms

5.1 Introduction 125
5.2 Results 125
5.2.1 Assessment based on model performance criteria 125
5.2.2 Variables selected by genetic algorithms 128
5.3 Discussion 130
5.4 Conclusion 132

Chapter 6: Development of habitat suitability models for pike using support-vector machines in combination with genetic algorithms

6.1 Introduction 137
6.2 Results 138
6.2.1 Assessment based on model performance criteria 138
6.2.2 Variables selection by genetic algorithms 139
6.3 Discussion 141
6.4 Conclusion 143

Chapter 7: Development of habitat suitability models for pike using logistic regressions in combination with genetic algorithms

7.1 Introduction 147
7.2 Results 148
7.2.1 Assessment based on model performance criteria 148
7.2.2 Variables selection by genetic algorithms 149
7.3 Discussion 153
7.4 Conclusion 154

Chapter 8: Comparison of habitat suitability model for pike using classification trees, artificial neural networks, support-vector machines and logistic regressions in combination with genetic algorithms

8.1 Introduction 159
8.2 Results 160
8.2.1 A comparative discussion of the obtained results 160
8.2.2 Comparison of the selected variables for the 4 modelling techniques 162
8.3 Discussion 166
8.4 Conclusion 169

Chapter 9: Application of predictive models for decision support in river management

9.1 Introduction 173
9.2 Site description and restoration options in the Dender basin 174

9.3 Application of the models for simulations related to the impact of land us changes on pike in the Dender basin 175
9.3.1 Simulations based on actual river conditions 175
9.3.2 Simulation exercise based on land use changes 178
9.4 Discussion 180
9.5 Conclusion 181

General discussion, conclusions and further research

Introduction 185
Collection of data related to river conditions 185
Towards (integrated) data management for model development and application 187
Model development 189
Model applications for decision support in water management 191
Further research, conclusions and recommendations 193

References 197
Summary 229
Samenvatting 235
Curriculum vitae 241
Annexes 247

List of abbreviations

ANNs	Artificial neural networks
BBI	Belgian Biotic Index
BBNs	Bayesian Belief Networks
BOD	Biological Oxygen Demand
CCI	Correctly Classified Instances
COD	Chemical Oxygen Demand
CTs	Classification Trees
DO	Dissolved Oxygen
DSS	Decision Support System
DST	Decision Support Tool
DTs	Decision Trees
EC	Electrical Conductivity
ESRI	Environmental System Research Institute
FN	False Negative
FP	False Positive
GAs	Genetic Algorithms
GIS	Geographical Information System
GUI	Graphical User Interface
HSI	Habitat Suitability Index
IBI	Index of Biotic Integrity
IBW	Institute for Forestry and Game Management
ICI	Incorrect Classified Instances
INBO	Institute for Nature and Forest Research
K	Cohen's Kappa
LRs	Logistic Regressions
MAE	Mean Absolute Error
MLPs	Multilayer Perceptrons
MLRs	Multinomial Logistic Regressions
MSE	Mean Squared Error
NMI	Normalized Mutual Information statistic
PCF	Pruning Confidence Factor
PhD	Philosophy doctor
r	Correlation coefficient
r	Determination coefficient
RAE	Root Absolute Error
RBF	Radial Basis Function
RIVPACS	River Invertebrate Prediction And Classification System
RMSE	Root Mean Squared Error
RRSE	Root Relative Squared Error
RSE	Relative Squared Error
RWQM	River Water Quality Management
SMO	Sequential Minimal Optimisation
Sn	Sensitivity
Sp	Specificity
SSE	Sum of Squared Error
SWAT	Soil and Water Assessment Tool
SVMs	Support Vector Machines

TN	True Negative values
TP	True Positive
TSS	True Skill Statistics
VMM	Flemish Environment Agency
WAECO-DSS	Water Ecology Decision Support System
WFD	Water Framework Directive

General introduction

Problem definition

The rates of extinction of freshwater species and the number of endangered freshwater species are at present alarming, especially those animals which occupy the top of food chains or food webs. A particular example in this case is the northern pike (*Esox lucius* L.), hereafter called pike, of which the population is endangered in Flanders, Belgium. The pike population, within its native range is declining due to loss of spawning and nursery habitat (Casselman and Lewis, 1996).

With the intention of addressing the complicated interrelationships between human activities and the situation of freshwater health, legislation initiatives have been taken and freshwater directives have been adopted at different levels e.g. Water Framework Directive (WFD) at the European level and the Decree of Integrated Water Policy at the local Flemish level (Adriaenssens, 2004). The Water Framework Directive (also known as the WFD or Directive 2000/60/EC) is a legislative framework to protect and improve the quality of all water resources such as rivers, lakes, groundwater, transitional and coastal waters within the European Union.

The WFD is a vital piece of European legislation designed to integrate the way water bodies are to be managed across Europe. The WFD intends to guard and enhance the water environment, promote sustainable water consumption, reduce water pollution and lessen the effects of floods and droughts. The WFD updates all existing European legislation and promotes a new approach to water management through river basin planning. The aim is also to stop the deterioration of wetlands and improve aquatic habitats for wildlife.

The first step in this regard is to identify water bodies, which are defined by the WFD as discrete and significant elements of surface water such as a lake, a reservoir, a stream, river or canal, part of a stream, a transitional water or a stretch of coastal water. These are the primary units of water management. The river basin means the area of land from which all surface run-off flows through a sequence of streams, rivers and, possibly, lakes into the sea at a single river mouth, estuary or delta.

In general, the main objectives of the WFD can be summarized as follows:
- to enhance the status and prevent further deterioration of aquatic ecosystems and associated wetlands. There is an obligation for almost all inland and coastal waters to achieve 'a good status' by 2015;
- to endorse the sustainable use of water;
- to reduce pollution of water, in particular focusing on 'priority' and 'priority hazardous' substances;
- to lessen the effects of floods and droughts;
- to update existing water legislation and introduce a co-ordinated approach to water management in terms of the concept of river basin planning.

Recently, modelling techniques are becoming more and more important in water quality and river restoration management, due to rapid development of computational problem-solving tools and the enhancement of scientific approaches (Huang and Xia, 2001). The availability of proper datasets and modelling techniques allows the development of ecosystem models with high reliability (Recknagel, 2003). Ecological modelling and data mining, also known as 'knowledge discovery in databases' are few of the many terms for finding structure in large-scale datasets. Several techniques emanating from computer sciences are at the moment increasingly used to analyse the behaviour of aquatic ecosystems. For instance, classification and regression trees (Dzeroski et al., 1997; Blockeel et al., 1999a,b; Dzeroski and Drumm, 2003), artificial neural networks (Lek and Guégan, 1999), support-vector machines (Decoste and Scholkopf, 2002; Guoa et al., 2005), logistic regressions (Hosmer and Lemeshow, 1989; McCullagh and Nelder, 1989), fuzzy logics (Barros et al., 2000; Adriaenssens et al., 2004b), Bayesian belief networks (Adriaenssens et al., 2004a), etc. proved to be useful tools in ecological water management. They combine reliable predictions with improved insight in ecosystem interactions (Recknagel, 2001, 2003). Mathematical methods are in this context considered to be valuable tools for water quality simulation and prediction. They can scheme consequences of alternative management, planning, or policy level activities, and as such identify effective management schemes.

In Flanders, for instance, numerous water system models are already being applied by different governmental administrations and institutions (Dedecker, 2005). These models describe different parts of the water system like sewers (e.g. Hydroworks, infoWorks CS), river hydraulics (e.g. Mike 11, ISIS) and river water quality (e.g. SIMCAT, SENTWA, PEGASE, SEPTWA, RWQM, SWAT).

The strong relapsing of the fish communities in Flemish rivers is mainly a result of the deterioration of their habitats. Weirs in Flanders' river basins, which are used for the sake of water quantity control, hamper fish migration and are considered to be one of the most important problems. An in depth study was conducted on the development of fish migration channels and natural overflow systems to achieve an ecologically friendly water quantity management in the near future (Soresma, 2000).

During the last decades, pike populations in many surface waters decreased tremendously. In Flanders for instance, pike was monitored in several sites during the last decades. This highly deteriorated state of pike in most surface waters in Flanders is associated with the severe water pollution, habitat deterioration and the intensive sport fishing (Goethals et al., 2006a). During recent years, many efforts have been made to bring back fish populations in Flanders but owing to the very limited remaining pike populations, this species is being threatened with extinction in Flanders. For this reason, it is very important to gain more insight in the specific habitat needs of pike to increase its remaining populations in the near future; in particular, meeting the requirements set by the Water Framework Directive (WFD) can be taken into account as an urgent task. In the literature, one can find, however, only limited (quantified) information with respect to the habitat requirements of pike.

In countries such as Belgium, Poland, Finland, Germany and Denmark, where more attention is being paid to restoring pike populations as well as to biomanipulation, imperfect information is a major concern regarding lake management (Skov et al., 2002). Pike, owing to its dependence on the aquatic vegetation, is a particularly good indicator for the general water quality.

Need for quantitative approaches to describe pike habitats

Habitat suitability index models for pike were developed by Inskip (1982), and spawning and nursery habitat models were refined and improved by Anderson (1993). Deltour (2004) developed a fuzzy logic model for the prediction of habitat preferences of pike in some river basins in Flanders. In this study, prediction of pike populations with some relevant variables was carried out to define habitat suitability rules.

In another study (Kerle et al., 2001), univariate and multivariate fuzzy rules were defined in terms of 4 habitat requirements: vegetation cover, flow velocity, water depth and substrate, for 5 different life stages of pike: spawning, adult, subadult, juvenile and fry. Glasbergen (2001a) used 6 variables of the pike's habitat requirement: water depth, oxygen concentration, vegetation cover, temperature, flow velocity and substrate. The author described two sets of habitat suitability rules: one for the reproduction stage (spawning, eggs and larvae) and another one for the growth stage (juveniles and adults). This study showed a higher correlation with habitat prediction for fuzzy based simulations than those based on preference functions (Jorde et al., 2000). Glasbergen (2001b) found similar findings to define habitat suitability rules with 6 variables (mentioned above) for pike and other species in two life stages (reproduction and growth). Casselman and Lewis (1996) examined the nursery and spawning requirements in the first year of the pike's life using a system for classifying and ranking the major physical characteristics and requirements. According to this study, the depth of the nursery habitat was linearly correlated with fish size and fish age. Spawning habitat was generally less vital or limiting but more easily manipulated and restored than nursery or juvenile-adult habitats. Adult abundance was related to the quantity of the vegetation cover, which was most favourable between 35 and 80% but reciprocally associated to the body size. Aadland (1993) proved that local velocities, depths or dominant roughness are strongly determining factors for pike and their size classes.

On the basis of the problems described, the scope and objectives of this thesis were formulated.

Scope and objectives

The main scope of this PhD thesis is to predict the habitat suitability of pike in Flanders by means of different modelling techniques, in particular Classification Trees (CTs), Artificial Neural Networks (ANNs) and Support Vector Machines (SVMs) and Logistic Regressions (LRs) and their combination with Genetic Algorithms (GAs).

Using these modelling techniques enables managers to derive rules for getting more insight in river ecosystems and hereby support their management. The development of these ecological modelling techniques would allow for supporting decision making in water management in Flanders. This could help to persuade river managers, stakeholders or other responsible organisations to make the necessary investments in river ecosystems. In other words, this information could guide river managers in the planning and the design of restoration and protection actions. In addition, this information could also serve as a basis to construct numerical habitat models of this species.

Taken as a whole the objective of the presented research can be summarized as follows (Figure 1):

- the establishment of monitoring networks and ecological databases to develop models predicting pike in 6 river basins in Flanders;
- the development, optimization and validation of habitat suitability models of pike based on integrated ecological modelling (classification trees J48, artificial neural networks, support-vector machines, logistic regressions and their combination with genetic algorithms);
- the simulation exercise based on predicted restorations.

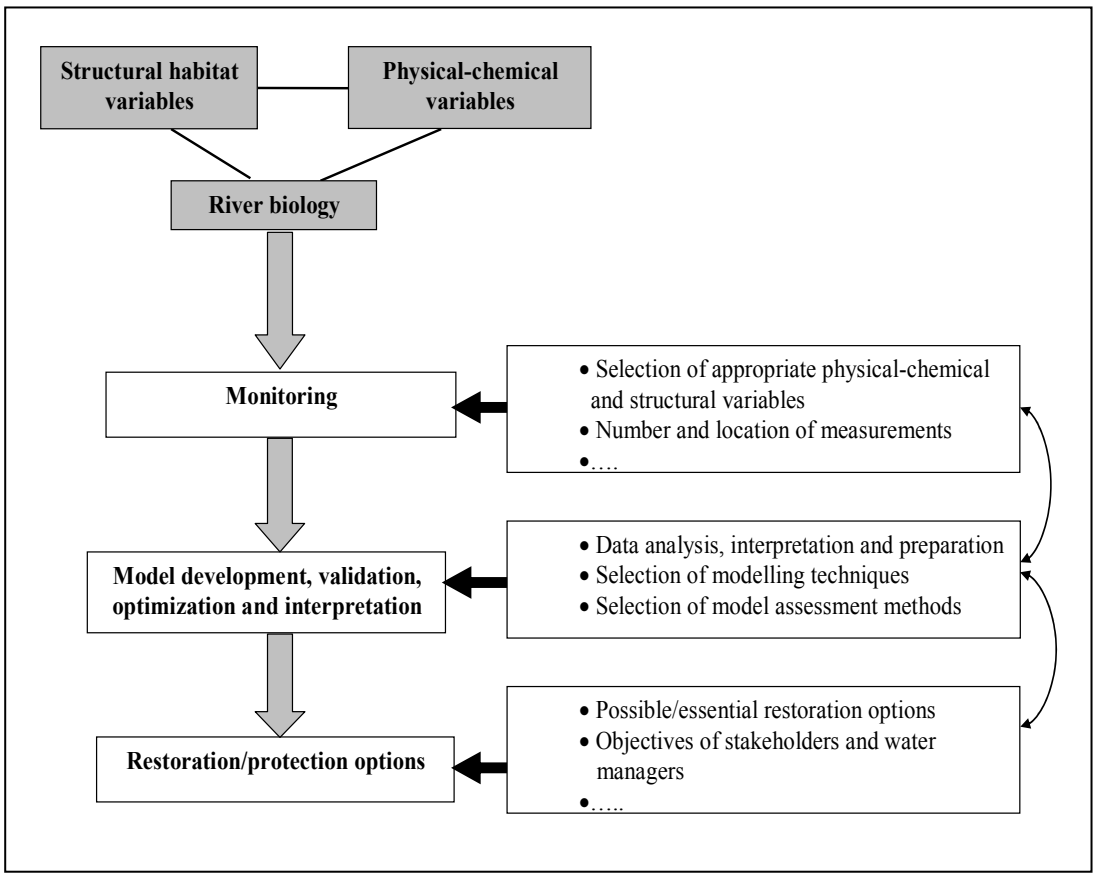


Figure 1. The main steps in relation to the development and application of ecological models for decision support in water management that have been considered in this PhD study.

In summary, this thesis is composed of a general introduction and 9 main chapters which are in short explained in the next paragraphs and followed by a general discussion, conclusions and further research.

In the **general introduction**, very briefly an overview is given regarding the problem definition, the need for quantitative approaches to describe pike habitats and scope and objectives.

In **Chapter 1** the biological and ecological knowledge about pike is described. In this review, some issues discussed among others are the life cycle, the feeding habitat requirements and the quantitative approaches needed to model pike populations. The latter section considers some matters regarding the habitat suitability modelling of pike and also the synecological modelling (relation with preys, competitors and the whole food web). In this chapter, the condition of the pike populations in Flanders is also reviewed.

Chapter 2 gives a description of the study area in Flanders, the data collection methods and the database set-up. Also the data analysis and ecological modelling methods are described here in a general manner. The detailed settings are given per chapter, since these were dissimilar.

Chapter 3 details the database set-up and exploration regarding bandwidth and distribution of input and output variables, correlation analysis between input and output variables and visualisation of the input and output variables.

Chapter 4 is about the development of habitat suitability models for pike by means of classification trees (J48) in combination with genetic algorithms.

Chapter 5 deals with the development of habitat suitability models of pike by means of artificial neural networks in combination with genetic algorithms.

Chapter 6 considers the development of habitat suitability models of pike by means of support-vector machines in combination with genetic algorithms.

Chapter 7 describes the development of habitat suitability models of pike by means of logistic regressions in combination with genetic algorithms.

Chapter 8 concerns the comparison of the habitat suitability of pike based on classification trees, artificial neural networks, support-vector machines and logistic regressions and their combination with genetic algorithms.

Chapter 9 focuses on practical applications of the predictive habitat suitability models for decision support in river management in Flanders.

The thesis ends with a **general discussion, the conclusions and the need for further research** regarding the improvement of data collection, model development and management options for pike in Flanders.

Chapter 1

Biology and ecology of pike

1.1 Introduction

The northern pike (*Esox lucius* L. 1758*)*, hereafter called pike, is a piscivorous fish that has been the objective of many studies during the last century as illustrated by Raat (1988) and Craig (1996). Many studies of pike ecology, in particular population dynamics and the role of pike in a community, have been based on long-term field studies.

Pike populations, like many other fish populations, and communities have been subject to considerable habitat changes (often human induced) and other perturbations including climate change (Casselman, 2002; Craig, 2008). It is well known that pike is a keystone piscivore in cool-water habitats and can influence species composition, abundance and distribution of many species including other pike, through cannibalism and competition in a fish community (Casselman and Lewis, 1996; Berg et al., 1997). This can have significant implications for the management and conservation of temperate freshwater fisheries (Craig, 2008).

In this chapter, general biological and ecological insights concerning the life cycle, feeding habits and habitat requirements of pike are described. In addition, quantitative approaches to model pike populations are reviewed. As a last part of this chapter, the status of pike in Flanders is briefly discussed.

1.2 Biological and ecological knowledge of pike

1.2.1 Taxonomy

Other common names for pike are 'common pike', 'great northern pike', 'jack', 'jackfish' and 'northern pike' (www.itis.usda.gov). Table 1.1 shows the taxonomical classification of pikes.

Table 1.1. Taxonomical classification of pike (www.itis.usda.gov).

- Kingdom *Animalia*
  - Phylum *Chordata*, animals with a spinal chord
  - Subphylum *Vertebrata*, animals with a backbone
    - Superclass *Osteichthyes*, bony fishes
    - Class *Actinopterygii*, ray-finned and spiny rayed fishes
    - Subclass *Neopterygii*
    - Infraclass *Teleostei*
      - Superorder *Protacanthopterygii*
      - Order *Esociformes*, the pikes and mudminnows
        - Family *Esocidae*, the pikes and pickerels
          - Genus *Esox*, the pikes
            - Species *Esox lucius*

1.2.2 Morphology

Pike is one of the most easily recognisable fish species (Figure 1.1) and is characterized by its elongated body and 'duck-bill' snout, large mouth with many sharp teeth and the rearward position of dorsal and anal fins. Gill rakers are only present as patches of sharp teeth on gill arches. The lateral line is notched posteriorly. The dorsal fin is located far to the rear, while the anal fin is located under and arises a little behind the dorsal one. The pectorals are located low on the body, with their base under the opercula. The paddle-shaped pelvic fins are also located low on the body. The caudal fin consists of 19 rays. Generally greenish in colour, they mature at 25-100 cm and 5-8 kg; exceptionally old fish, up to thirty years, may even reach 150 cm and 35 kg (Fitter and Manuel, 1986).

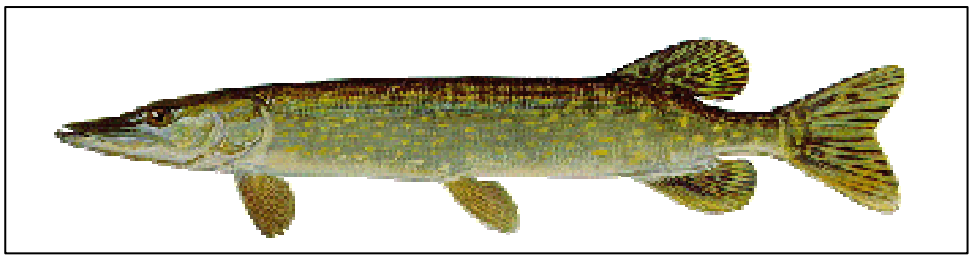


Figure 1.1. The pike (Esox lucius L. 1785) (www.fishbase.org).

1.2.3 Origin, distribution and abundance

Pike has a Holarctic range. Pike is distributed in temperate climates between 74°N-36°N, 167°W-180°E in North America, Europe and parts of Asia (www.fishbase.org). The broad distribution of pike and its occurrence in many different habitats give some indication of pike's influence on freshwater communities. Pike and its relatives radiated in freshwater before Eurasia and North America became isolated (Crossman, 1996; Craig, 2008). European populations of pike appear to be more variable than populations from North America and Siberia (Launey et al., 2006).

Pike is a cool water species (Casselman, 1978) and a popular freshwater game fish. Pike can inhabit rivers, lakes and weakly saline waters (Craig, 2008). In addition to severe impacts from water pollution, habitat degradation and fishing, also climate change is expected to have effects on the population biology of pike and its distribution (Winfield et al., 2008).

1.2.4 Growth rates

The growth of pike fluctuates with latitude. Pike is vulnerable to stunting due to lack of prey and this is typically associated with high population densities and limited water transparency, while water temperatures higher than 21°C restrict growth and related size distribution (Casselman and Lewis, 1996). A study comparing growth and mortality of pike between river and lake populations at the same latitude, however, did not show any differences (Griffiths et al., 2004; Craig, 2008). Flow is probably unimportant compared to other abiotic factors. Growth of pike in lakes is strongly affected by both abiotic and biotic factors, such as temperature, water transparency, productivity, prey type and abundance, pike density and other competitors (Casselman and Lewis, 1996; Margenau et al., 1998; Pierce et al., 2003; Craig, 2008).

1.2.5 Life cycle of pike

Pike in the southern parts of its range spawns sometimes at the age of one year, while the majority of them spawn initially at two years (McClane, 1998). In general, during their second or third year of their life, pike individuals become sexually mature. However, according to some specialists, body size is a more crucial factor than age. The spawning period of pike is in March and April and thus relatively early in the year (De Nie, 1993). The spawning usually takes place in a temperature range between 6°C and 14°C, with an optimum of 8°C to 12°C (Glasbergen, 2001b). However, small numbers of pike are able to spawn at any time from late February until late May. Spawning is preferred in the shallow vegetation rich water. According to Casselman and Lewis (1996), also dense vegetation is most favourable for the nursery habitat of the pike larvae. Figure 1.2 represents different larval stages of pike (www.fishbase.org). Also in the later years, pike is to a high extent depending on vegetation as an ambush for its attack, and also to hide from larger pike. Only after four years, pike is moving to the open water, since then the risk for cannibalism has become relatively low. Chapman and Mackay (1984) and Vehanen et al. (2006) pointed out that the size of an individual pike influences its habitat choice. This is more elaborated in the next section.

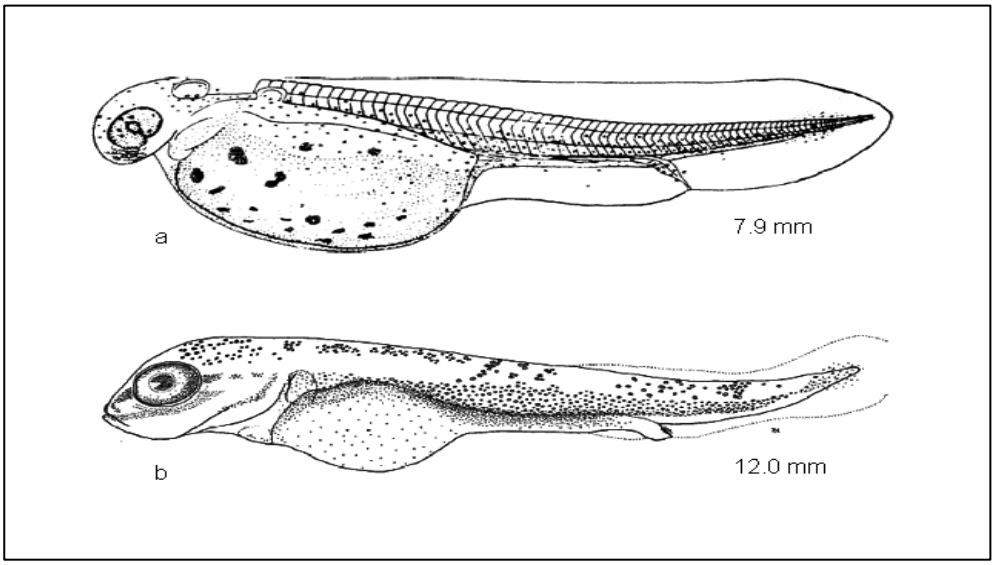


Figure 1.2. Different larval stages of pike (www.fishbase.org).

1.2.6 Role of vegetation in the life cycle of pike

Pike is strongly associated with vegetation rich habitats. Especially in the first year of life, the young pike depends on vegetation (Casselman and Lewis, 1996) and particularly complex habitats are selected (Skov and Berg, 1999; Skov et al., 2002). Various reasons can be considered for the pike's dependence on the vegetated cover, e.g. protection against cannibalism (Eklov, 1997), cover against predators (Skov and Berg, 1999) and trophic relationships (Bry, 1996). Numerous plant taxa can have a significant role on the pike spawning grounds e.g. decaying vegetation, hygrophytes, amphiphytes, helophytes and temporary vegetation, as well as partially submerged terrestrial vegetation. Meeting spawning habitat requirements (including the presence of adequate vegetation) is the most critical condition for establishing a sustainable pike population (Inskip, 1982; Raat, 1988). Figure 1.3 shows a typical and optimal pike habitat consisting of mixed plants (floating, submerged and emerged).

Figure 1.3. A typical pike habitat: shallow brook with rich vegetation.

Figure 1.4 reveals that pike is strongly associated with the relative vegetation coverage. Based on Figure 1.4, there is a direct relation between vegetation cover and pike's spawning. When there are low plant densities only large pike individuals are encountered. On the other hand, vegetation coverage of higher than 50 % leads to seriously reduced pike populations.

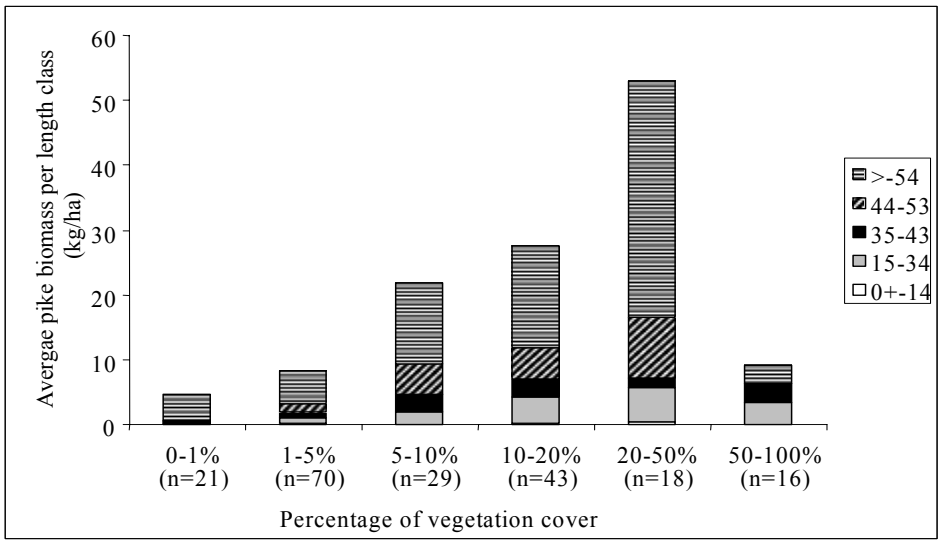


Figure 1.4. Average pike biomass per length class (cm) in relation to percentage of vegetation cover; n=number of sampling sites (Goethals et al., 2006b).

1.2.7 Feeding

Pike are opportunist feeders and catch the most common species in the easiest way. In this way, pike exploit seasonally abundant food resources. Feeding behaviour of pike is associated with the water depth and time (season). The most favourable feeding habitat temperatures for pike (in the summer) are 18°C to 20°C (Raat, 1988). In summer, pike individuals are often found in weedy, shallow waters of 1.5-2 meter deep. Raat (1988) argues that in a well vegetated area pike can spend more time feeding than in a less vegetated habitat, since it takes more energy in watching for protection against other larger pike. This cannibalistic interaction between different life stages of pike (Craig and Kipling, 1983) can have an intensive influence on the habitat preferences (Raat, 1988; Bry, 1996). Several biotic and abiotic factors have the potential to influence the extent of cannibalism of pike, including the presence of alternative prey, size variations, population densities, refuges, water clarity and light intensity (Folkvord, 1997).

1.2.8 Habitat requirements

Main habitat requirements of pike can be divided in 3 categories: spawning, nursery and juvenile-adult habitat. These are mainly related differences in vegetation cover, substrate, water temperature, oxygen, transparency and turbidity (Casselman and Lewis, 1996).

The relations between habitat preferences of pike with age in different seasons showed that in summer a range of habitat types including open water (Diana, 1980) and vegetated areas (Grimm, 1981) are used by large pike but in early summer young-of-the-year pike settle in dense emergent and submerged vegetation near the edge of water (Grimm, 1981), until reaching a size of around 15 cm total length (Grimm, 1981).

Kobler et al. (2008) studied contrasting movement rates and habitat choice of pike in mid-summer and mid-winter in a lake. The authors stated that pike moved significantly more in summer, and when it had activity peaks in twilight periods. In winter, prominent activity peaks at specific daytime periods were missing and pike chose habitats significantly closer to shore.

Pike spawns in shallow waters with vegetation in spring. They are inclined to migrate up tributaries of flooded marshes and wetland or shallow shoreline inundation. Optimal spawning substrate is the flooded vegetation in a shallow and sheltered area (Casselman and Lewis, 1996). Spawning is associated with the water-level changes. High water levels during spawning with stable levels after the incubation period are associated with large year classes. Spawning habitat of pike may be affected by other fish species (Casselman and Lewis, 1996).

Nursery (young-of-the-year) requirements of pike are much less critical than the spawning habitat requirements. Young-of-the year pike individuals are often solitary, difficult to catch, require specialized sampling techniques (Casselman, 1978). Young fish in the nursery environment grow quickly and increase in size and activity. Once pike individuals are distributed into deeper water, they are usually found in moderately dense vegetation.

Anderson (1993) developed suitability indices for water depth, vegetation type and density of young pike individuals in their nursery habitat. The author concluded that intermediate

densities of vegetative cover were optimal and that young-of-the-year pike, measuring 65 mm preferred a mixture of submerged and emergent vegetation with densities ranging from 20 to 50 %.

Hawkins et al. (2003) investigated the spatial distribution and social behaviour of young-of-the-year pike in early winter. Reed beds were selected in the early phase but thereafter, pike congregated in pool habitats. During their maturation, pike altered their vegetative associations from emergent for fry, to emergent, floating and submerged for young and finally submerged for adult individuals. Consequently, macrophyte density influences the size structure of pike populations. Since large pike require more open water, lakes with dense vegetation are usually dominated by smaller pike (Grimm and Klinge, 1996). For pike, however, it is likely that there is some adaptability in habitat selection, depending on availability of prey and other factors (Inskip, 1982).

Pike populations can be negatively affected by eutrophication. Casselman and Lewis (1996) concluded that turbid water leads to a decline in macrophytes and the loss of macrophytes may be crucial for the young pike. Similarly Bry (1996) points out that high turbidity can reduce the number of spawning habitats. Bry (1996) also states that eutrophication results in unfavourable hunting conditions for pike due to the lack of cover and low visibility. Once the oxygen concentration falls below 30-35 % saturation, hatch of pike eggs and survival of embryos and larvae are significantly reduced (Siefert et al., 1973). Hummocks and vegetative mats are optimal spawning substrates because they help keep the eggs suspended in better oxygenated water. Pike embryos are sensitive to heavy siltation.

1.3 Quantitative approaches to model pike populations

1.3.1 Habitat suitability modelling

Physical habitat models have become an important tool in river management. Habitat based fish population models are needed by both fisheries and habitat managers to assess the effects of exploitation and habitat change on fish populations. Therefore, it is necessary to determine quantitative linkages between habitat and fish populations.

Already 26 years ago, Inskip (1982) developed Habitat Suitability Index (HSI) models for pike in riverine and lacustrine systems. The author used seven habitat variables for the description of pike habitat suitability: 1) ratio of spawning habitat to summer habitat (area that is less than 1 m deep and vegetated (spring) divided by the total midsummer area); 2) drop in water level during embryo and fry stages; 3) percent of midsummer area with emergent and/ or submerged aquatic vegetation or remains of terrestrial plants (bottom debris excluded); 4) Total Dissolved Solids (TDS) in the epilimnion during midsummer; 5) least suitable pH in spawning habitat during embryo and fry stages; 6) average length of frost-free season; and 7) maximal weekly average temperature of the epilimnion (1 to 2 m deep).

Luz and Loucks (2003) described a method for assessing the relative quality of a coastal wetland of Lake Ontario as habitat suitability for a bioindicator fish like pike (focusing on its early life stages) based on hydrological factors: daily water levels and water temperature. Results showed that at least for the early life stages of pike a less variable water level regime was more beneficial than a regime of that natural variation.

Later on, Minns et al. (1996 and 1999) constructed an age structured fish population model that includes density-dependent effects related to the supply of suitable habitats for pike. The model considers various life stages: spawning, fry, juveniles and adults (Figure 1.5). The model was used to exhibit the effect of habitat supply and lake morphometry on fish populations and was used to simulate the potential effects of habitat restoration efforts.

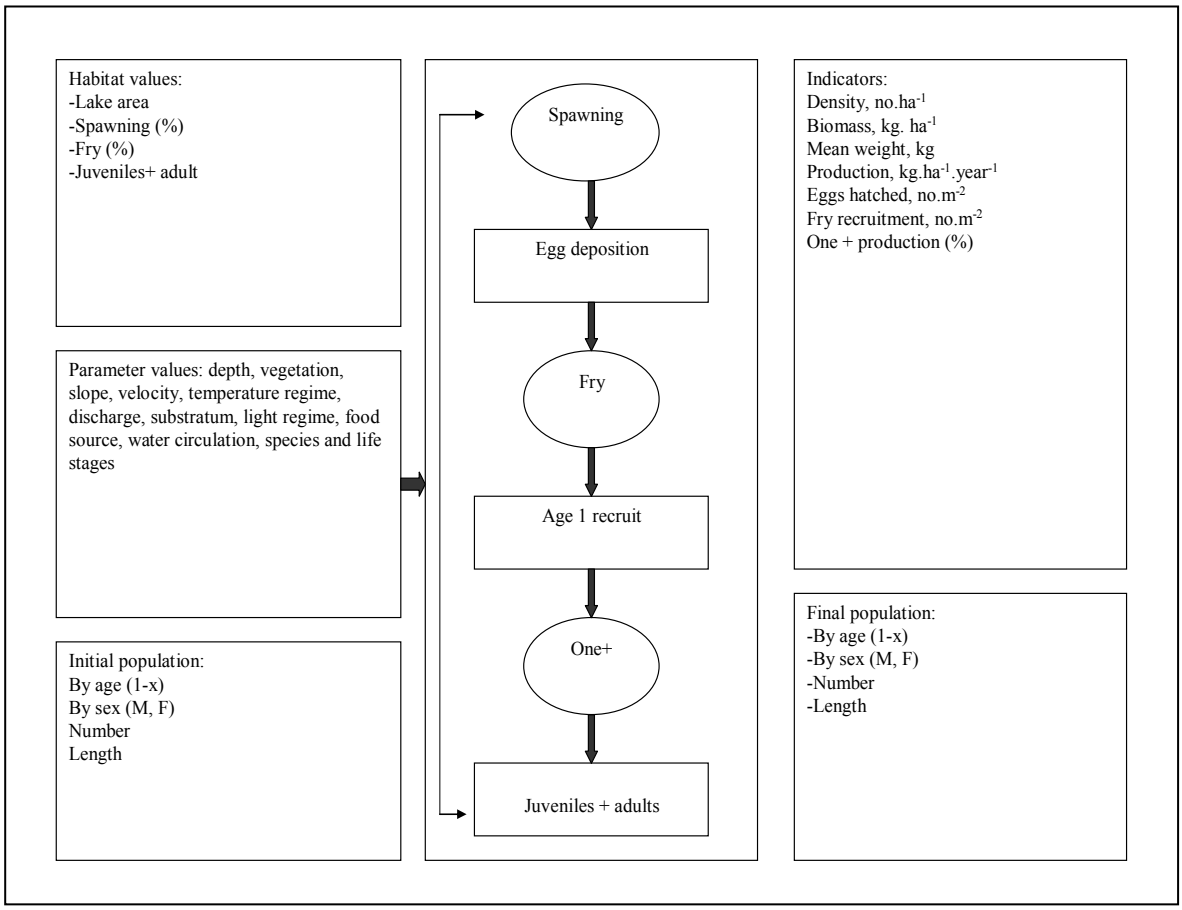


Figure 1.5. Schematic of a pike population and a habitat modelling framework (after Minns et al., 1996).

In another study, Kerle et al. (2001) defined univariate and multivariate fuzzy rules in terms of some habitat requirements: vegetation cover, flow velocity, water depth, substrate, water temperature and dissolved oxygen for 5 different life stages of pike: spawning, fry, juvenile, subadult and adult. The habitat suitability index reflecting the overall habitat quality of pike ranged between 0 and 1. Table 1.2 and 1.3 represent particular examples of fuzzy sets constructed for the HSI respectively the reproduction and growth stage of pike (Kerle et al., 2001). Besides, Figure 1.6 represents a particular example of the division of water depth (an important parameter of the pike's habitat suitability) in fuzzy sets for pike. The division of water depth in fuzzy sets is similar to flow velocity. Very shallow water, with an average depth under 10 cm, is in general less suitable for juvenile and adult fish, while it can be very suitable for larvae. Shallow water

up to about 100 cm deep is usually preferred by fish that prey on sight. Water depth with an average depth of 400 cm is considered to be too high (Kerle et al., 2001). Figure 1.7 shows the fuzzy classes of the HSI.

As can be visualized in Tables 1.2 and 1.3, there are several limitations in the study of Kerle et al. (2001) in which habitat suitability models of pike were constructed based on fuzzy logic. This technique is based on knowledge of experts and is represented by IF and THEN rules. In this method, field data were not used for the model development. The application of this technique is moreover very time-consuming and does not consider the information that can be extracted from field observations. For this reason, only a limited number of river characteristics can be considered. Most of the variables used in that study were related to the structural-habitat characteristics, while only one water quality variable (dissolved oxygen) was considered for the development of the pike habitat suitability models. These underline some weaknesses of the work of Kerle et al. (2001) in relation to its relevance in rivers in Flanders, where water quality still has a substantial impact on the presence of pike.

Table 1.2. Habitat suitability rules for the reproduction stage of pike (Kerle et al., 2001). DO = dissolved oxygen concentration.

IF	Vegetation cover	is	Low	**THEN**	Habitat suitability	is	Medium
IF	Vegetation cover	is	Medium	**THEN**	Habitat suitability	is	High
IF	Vegetation cover	is	High	**THEN**	Habitat suitability	is	Very high
IF	DO	is	Low	**THEN**	Habitat suitability	is	Very low
IF	DO	is	Medium	**THEN**	Habitat suitability	is	Medium
IF	DO	is	High	**THEN**	Habitat suitability	is	Very high
IF	Substrate	is	Silt	**THEN**	Habitat suitability	is	Medium
IF	Substrate	is	Sand	**THEN**	Habitat suitability	is	High
IF	Substrate	is	Gravel	**THEN**	Habitat suitability	is	High
IF	Temperature	is	Very Low	**THEN**	Habitat suitability	is	Medium
IF	Temperature	is	Low	**THEN**	Habitat suitability	is	Very high
IF	Temperature	is	Medium	**THEN**	Habitat suitability	is	Medium
IF	Temperature	is	High	**THEN**	Habitat suitability	is	low
IF	Temperature	is	Very high	**THEN**	Habitat suitability	is	Very low
IF	Flow velocity	is	Very Low	**THEN**	Habitat suitability	is	Very high
IF	Flow velocity	is	Low	**THEN**	Habitat suitability	is	Medium
IF	Flow velocity	is	Medium	**THEN**	Habitat suitability	is	Very low
IF	Flow velocity	is	High	**THEN**	Habitat suitability	is	Very low
IF	Water depth	is	Very Low	**THEN**	Habitat suitability	is	Very high
IF	Water depth	is	Low	**THEN**	Habitat suitability	is	Medium
IF	Water depth	is	Medium	**THEN**	Habitat suitability	is	Very low
IF	Water depth	is	High	**THEN**	Habitat suitability	is	Very low

Table 1.3. Habitat suitability rules for the growth stage of pike (Kerle et al., 2001). DO = dissolved oxygen concentration.

IF	Vegetation cover	is	Low	**THEN**	Habitat suitability	is	Medium
IF	Vegetation cover	is	Medium	**THEN**	Habitat suitability	is	High
IF	Vegetation cover	is	High	**THEN**	Habitat suitability	is	High
IF	DO	is	Low	**THEN**	Habitat suitability	is	Low
IF	DO	is	Medium	**THEN**	Habitat suitability	is	Medium
IF	DO	is	High	**THEN**	Habitat suitability	is	Very high
IF	Substrate	is	Silt	**THEN**	Habitat suitability	is	High
IF	Substrate	is	Sand	**THEN**	Habitat suitability	is	Very High
IF	Substrate	is	Gravel	**THEN**	Habitat suitability	is	Very High
IF	Temperature	is	Very Low	**THEN**	Habitat suitability	is	Low
IF	Temperature	is	Low	**THEN**	Habitat suitability	is	Medium
IF	Temperature	is	Medium	**THEN**	Habitat suitability	is	High
IF	Temperature	is	High	**THEN**	Habitat suitability	is	Very high
IF	Temperature	is	Very high	**THEN**	Habitat suitability	is	Low
IF	Flow velocity	is	Very Low	**THEN**	Habitat suitability	is	Very high
IF	Flow velocity	is	Low	**THEN**	Habitat suitability	is	High
IF	Flow velocity	is	Medium	**THEN**	Habitat suitability	is	Low
IF	Flow velocity	is	High	**THEN**	Habitat suitability	is	Very low
IF	Water depth	is	Very Low	**THEN**	Habitat suitability	is	Low
IF	Water depth	is	Low	**THEN**	Habitat suitability	is	Very high
IF	Water depth	is	Medium	**THEN**	Habitat suitability	is	High
IF	Water depth	is	High	**THEN**	Habitat suitability	is	Medium

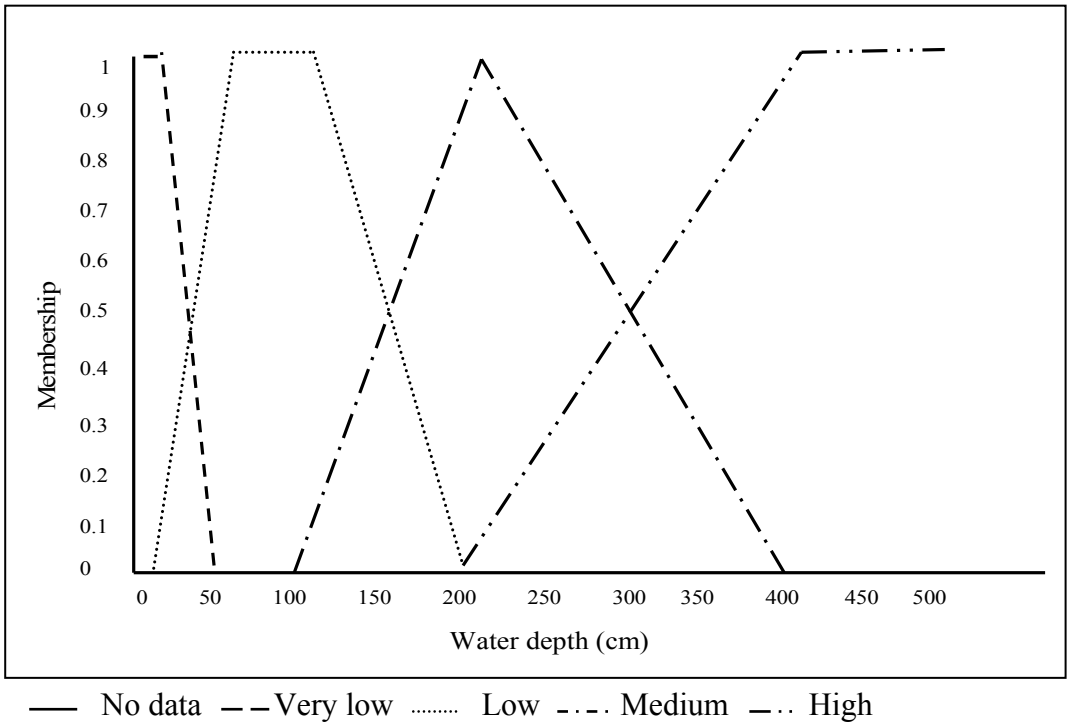


Figure 1.6. Example of fuzzy sets for water depth, as part of the Habitat Suitability Index (HSI) model for pike (Kerle et al., 2001).

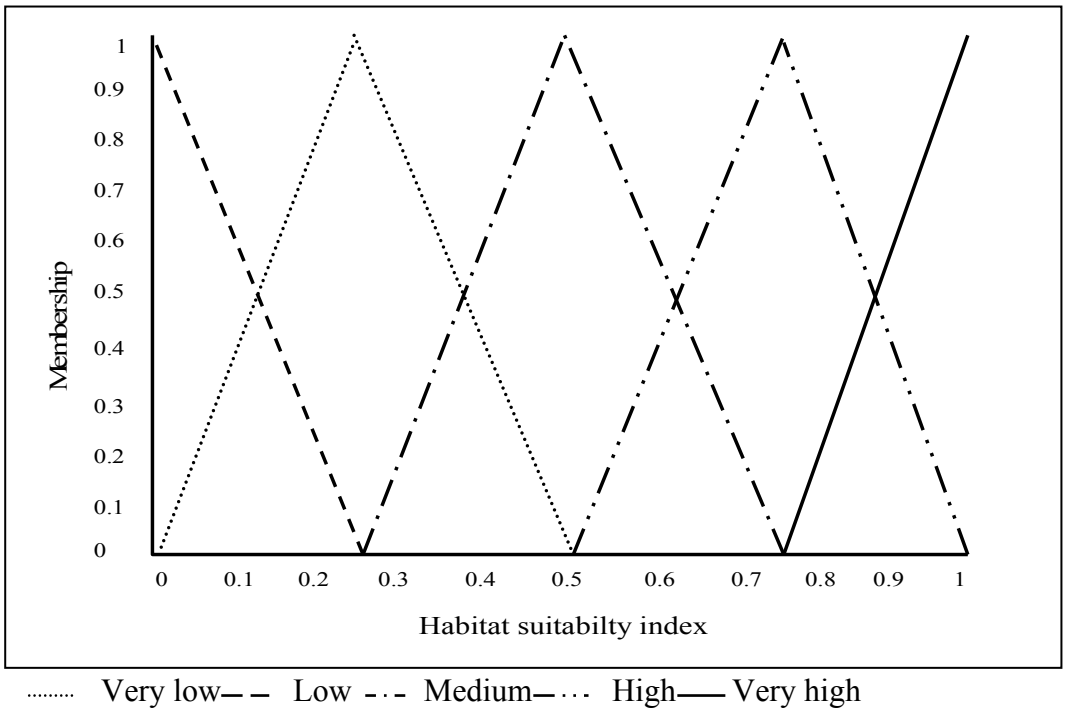


Figure 1.7. Fuzzy sets of the Habitat Suitability Index for pike (Kerle et al., 2001) which ranges between 0 and 1 (0 is not suitable and 1 is maximum suitability). The HSI is divided in five fuzzy sets, ranging from 'very low' to 'very high'.

Glasbergen (2001a) used six factors to determine the pike's habitat requirements: water depth, oxygen concentration, vegetation cover, temperature, flow velocity and substrate. The author developed rules for two life stages: one for the reproduction stage (spawning, eggs and larvae) and another for the growth stage (juveniles and adults). The author's study showed a higher correlation with habitat prediction for fuzzy based simulations than those based on preference functions (Jorde et al., 2000). Glasbergen (2001b) defined similar habitat suitability rules for the six variables mentioned above for pike and other species in two life stages (reproduction and growth).

Lafayette (2003) used Habitat Suitability Index (HSI) functions for the most crucial life stages of pike: spawning, egg incubation, and juvenile growth. The HSIs were constructed in 5 stages: 1) water depth and flooding frequency during the spawning period; 2) water depth during egg incubation and larval period; 3) water level stability during growth rate; 4) water temperature during spawning, and 5) water temperature during the growth period. These 5 stages were used for the construction of composite habitat suitability indices.

Most studies were, however, performed on a local basis (individual lakes) and mostly for stagnant waters. Also data-driven approaches which have been successfully applied on other communities (e.g. Goethals, 2005), like classification trees and neural networks, have thus far not been used to predict pike on the basis of habitat conditions. Such approaches being tested on a larger scale are therefore interesting research goals, both from an academic and from a river management perspective.

1.3.2 Predator-prey modelling

Pike can continue eating invertebrates as adults under certain conditions. Although pike can consume a wide variety of food items from invertebrates to fishes, they are extremely well-adapted piscivores for most of their lives. Their capability, as predators, to capture and consume prey is determined by both their size and that of the prey, the prey-to-predator size ratio (Craig, 2008).

According to Elliott and Hurley (2000), a shift to piscivory is energetically advantageous because a fish diet brings about a faster growth rate than an invertebrate diet. Hence, fishes

should preferably shift to piscivory at the smallest possible size. Predation by pisicivorous fish such as pike can have direct effects on prey fish populations e.g. a reduction of population density or the change of size structure (Turesson et al., 2002). Predator preference of habitat, diet or prey size may verify the predator-prey dynamics and the community structure (Fryxell and Lundberg, 1998; Nilsson, 2001).

Predator behaviour may seriously affect the impact on prey. Furthermore, the size- and density-dependent interactions between predators might be a major key to the understanding of predator-prey dynamics and community composition. For many fish, diet shifts ultimately result in piscivory (e.g. largemouth bass, *Micropterus salmoides*, Olson, 1996 and yellow perch, *Perca flavescens*, Fullhart et al., 2002). This shift to piscivory is considered as favourable for individuals because of the mutual increase in growth and survival (Persson and Bronmark, 2002).

Day-active predators such as pike and perch are important piscivorous predators in European lakes and can cause major habitat shifts in roach (Eklov and Persson, 1996; Schulze et al., 2006; Eklov and Jonsson, 2007). Eklov and Persson (1996) showed that roach individuals move into or stay close to the vegetation in the presence of piscivorous perch. In contrast, in the presence of pike, which stays mainly within or near macrophyte beds, roach uses the open water habitat. In the presence of both predators, roach uses vegetated areas and therefore becomes more vulnerable to pike predation.

The success of prey capture by a predator is affected by the behaviour of the prey. There are numerous examples of anti-predatory behaviour by prey fishes including pike themselves in the presence of pike (Kelley and Magurran, 2003; Skov et al., 2003; Craig, 2008). The response can be caused by visual e.g. turbidity that can play an important role (Lehtiniemi, 2005; Skov et al., 2007), olfactory e.g. chemical (Brown, 2003) tactile and auditory detection (Kelley and Magurran, 2003).

Predator-prey interactions may have a significant role in regulating the freshwater fish community structure. Most studies have focused on predator-prey or competitive interactions between pairs of species. Rising prey size with rising predator size is a common phenomenon

among predatory species, because since fishes grow, their energy demand increases and larger prey items become energetically more beneficial (Elliott and Hurley, 2000).

Nilsson (2001) has modelled size-structured predator-prey systems. The effects were evaluated for gape-size limited predation on prey population growth and density when predators were non-interacting, cannibalistic, interfering and cannibalistic and interfering. Based on this model, predation from non-interacting predators obviously reduces prey density, compared with prey densities in the absence of predation. The absolute gape-size limit in pike is high and that maximum ingestible prey size is limited by prey body depth (Nilsson and Bronmark, 2000; Dörner et al., 2007). For instance, as pike is a selective predator, it prefers the shallow-bodied roach over the deeper-bodied common bream and in general, small prey sizes within each prey species. Another point is that the handling time in pike increases with prey body depth. For instance, since bream is deeper-bodied than roach, the handling time is also longer. Therefore, prey vulnerability is identified by risk of predation and intraspecific interactions, and behavioural preferences in pike and not by the pike's gape-size limits (Nilsson, 2001).

Pikeperch and pike are important top predators in temperate freshwater ecosystems (Maitland and Campbell, 1992). Pikeperch is a pelagic piscivorous fish that unlike perch and pike is most active during twilight and at night (Craig, 1987). They both prefer small fish prey to large in spite of pike's remarkable ability to swallow large prey. Pikeperch and pike are solitary predators. While pikeperch combines pursuit and ambush techniques in its foraging tactics, pike relies solely on ambush predation (Maitland and Campbell, 1992). Pikeperch and pike are also cannibalistic, both increasing cannibalism in response to decreasing abundance of alternative prey (Mehner et al., 1996).

Pikeperch and pike are suitable model species for a size-structured predator–prey model that evaluates effects of different degrees of intraspecific interactions and interference, i.e. effects of cannibalism and kleptoparasitism, on predator functional responses and predator and prey population growth and density (Nilsson, 2001). The author used roach as suitable model for pike's prey because roach individuals are common preys for pike and pikeperch (Maitland and Campbell, 1992). The reason is that roach is shallow-bodied and rarely exceeds 25 cm body length, and is therefore the preferred prey (Maitland and Campbell, 1992).

Predator–prey interaction studies have usually concentrated on only one predator species (Kahilainen and Lehtonen, 2003). The authors applied logistic regressions (LRs) on 4 predator fishes: pike, burbot, Arctic charr and brown trout. Their study showed that the specialist piscivorous pike and burbot consumed more prey species and took a wider range of prey sizes than the other two fish species. The prey length for all predators increased in relationship to predator length, which meant that the specialist piscivorous pike and burbot switch to piscivory at a smaller size than the secondary piscivorous ones (the Arctic charr and brown trout). Pike and burbot are opportunistic feeders, using several prey fish species compared to the brown trout and Arctic charr.

Regarding the relation between pike populations and prey species, the applied techniques are very limited, and similar to the habitat suitability modelling, not yet applied on a larger scale. This could be very important regarding pike stocking, since the efficiency of the stocking is probably related to the availability of prey. When prey is limited, cannibalism might occur to a high extent, leading to very unfavourable effectiveness of the stocking efforts.

1.3.3 Competition modelling

Pike might compete with some piscivorous fish such as other members of the esocid family e.g. walleye, pikeperch, perch and burbot. Pike can coexist perfectly with rainbow trout in deep areas (mean depth 12-15 meters) and moderately in littoral areas (Rutz, 1999). Nilsson (2001) presented the synecological relationship between pikeperch and pike in ecosystems with use of the predator model. Based on this model, pike and pikeperch represented different effects on ecosystems. Since pikeperch are not likely to interfere similarly to pike, their per capita functional responses would not be affected by predator densities. Therefore, they would maintain a high functional response throughout predator densities. Pike, on the other hand, do interfere, and therefore their per capita functional responses decrease with increased predator density. Recently, more and more ecological studies have revealed that roach, perch and pike dominate in degraded waters (Wolter, 2001; Penczak et al., 2002; Kruk et al., 2007). This dominance is often judged an indicator of the destruction of riverine habitats, particularly when convoyed by the marginal dominance of rheophils (Wolter, 2001; Aarts and Nienhuis, 2003; Kruk et al., 2007).

1.3.4 Food web modelling

Pike can have important effects on the structuring of fish communities e.g. size and condition, abundance and longevity of both prey and other predators (Hinch et al., 1991; Bertolo and Magnan, 2005; Craig, 2008). These effects have been mostly observed when pike have been accidentally or intentionally introduced into a community (Craig, 1996; Craig, 2008). However, as a predatory fish at the top of the food chain, pike populations are small in comparison to other fishes further down the food chain (Craig, 2008).

Effects of changes in populations of piscivorous keystone species might cascade down the food web, affecting other organisms at lower trophic levels. For example, if the piscivorous populations are numerous, they could forage all the planktivorous fish. As a result of this, the zooplankton community could increase and this can lead to lower phytoplankton concentrations and thus clearer water. The reverse effect occurs as a result of eutrophication, leading to a limited vegetation where too few piscivorous fish are present.

Also habitat structure can affect food availability (Crowder and Cooper, 1982), predator-prey relationships (Sih and Chirstensen, 2001) and competitive interactions. Prejs et al. (1997) studied food web modelling in shallow and eutrophic lakes in the context of biomanipulation strategy development. This study aimed at introducing young-of-the-year pike to control the density of juvenile stages of the dominant planktivores (e.g. roach and white bream). The successive introductions of the juvenile pike were accompanied by selective removal of the large pike, roach, bream and white bream.

According to Hansson et al. (1997), the effects on community structure caused by pikeperch might remain for long periods after stocking of pike, e.g. in order to manipulate the cyprinid density. The effect of increased density of pike will possibly last for only short periods of time (Meijer and Hosper, 1997), and it seems that continuous stocking of pike is necessary to achieve a long-term clear-water phase in lakes (Hansson et al., 1997). Pike might alter fish communities, primarily through predation (Craig et al., 2003).

1.4 Status of pike in Flanders, Belgium

According to Vandelannoote et al. (1998), pike is in particular considered as rare in streams in Flanders but not in stagnant waters. Since last decades, pike populations have substantially decreased, in particular in running waters. This poor status of pike in most streams in Flanders is associated with severe water pollution, habitat deterioration and intensive sport fishing. Moreover, weirs and other morphological impacts, which are used for the control of water quantity, can hamper fish migration and are among the most severe problems for fish communities in the Flemish river basins. For this reason, it is very important to gain more insight in the specific habitat needs of pike to increase the remaining populations in the near future.

Van Hees (2003) reported that the pike population status in the Nete basin (Flanders) is moderate. The ratios of predator and prey fluctuate significantly from one place to another. In the Kleine Nete basin, however, there are no migration barriers but the bad structure of the watercourses and the loss of suitable habitats are now alarming. Deltour (2004), on the other hand, developed a fuzzy logic model for prediction of the pike's habitat preferences in some river basins in Flanders. The priority must urgently be given to the reconstruction of natural riverbanks and the introduction of spawning habitat zones with suitable aquatic vegetation. The Grote Nete in Flanders, on the other hand, is facing some other problems related to water mills along the course of the river. These barriers prevent the natural migration of pike and therefore recolonisation. The increase of metal pollution entering the Nete and the upper course of the Grote Nete seems to be another drawback for pike populations and other fish species as well. In the case of different restoration projects, an intelligent and well-organized reintroduction in potentially appropriate environments for pike is required.

Chapter 2

Material and methods

2.1 Introduction

Given the intensive use of freshwater resources, the ecological quality of catchments and watercourses in Flanders has been significantly lowered. Watercourses with natural stream characteristics and vulnerable species are rare and mainly consist of isolated headwaters. To maintain the natural variety of river types and the networks within the catchments and to protect the vulnerable species, in particular fish communities in Flanders, there is an urgent need for river restoration programmes. In this context, numerical relations between river conditions and fish communities are necessary to guide decision support in the selection of relevant options.

In the light of the problem mentioned, this chapter is focusing on the monitoring and data collection techniques for both biotic and abiotic variables of a set of relevant river basins in Flanders, Belgium. The biotic data were collected by means of electrofishing methods by the Institute for Forestry and Game Management (IBW), recently included in the newly created Research Institute for Nature and Forest (INBO). Additionally, some chemical data were provided by the Flemish Environment Agency (VMM). This chapter also deals with the applied data statistical techniques, the ecological modelling methods and the application of these habitat models for decision support in river management.

In summary, Figure 2.1 outlines the conceptual framework of this PhD dissertation for the development and application of ecological models for decision support in water management of the river basins in Flanders. As illustrated in this figure, this dissertation consists of 4 main parts: 1) monitoring and data collection; 2) data analysis; 3) modelling methods, and 4) restoration management. The first part (monitoring and data collection) is entirely described in this chapter. The general aspects of the second part (data analysis) are also explained here, but the results are given in Chapter 3. The general aspects and specific parameter settings of the third part (modelling methods) are described here, while their results will be allocated in Chapter 4, 5, 6 and 7 and the comparison of these models will be described in Chapter 8. The restoration management (fourth part) is very briefly explained in this chapter but will be discussed in Chapter 9 with more details.

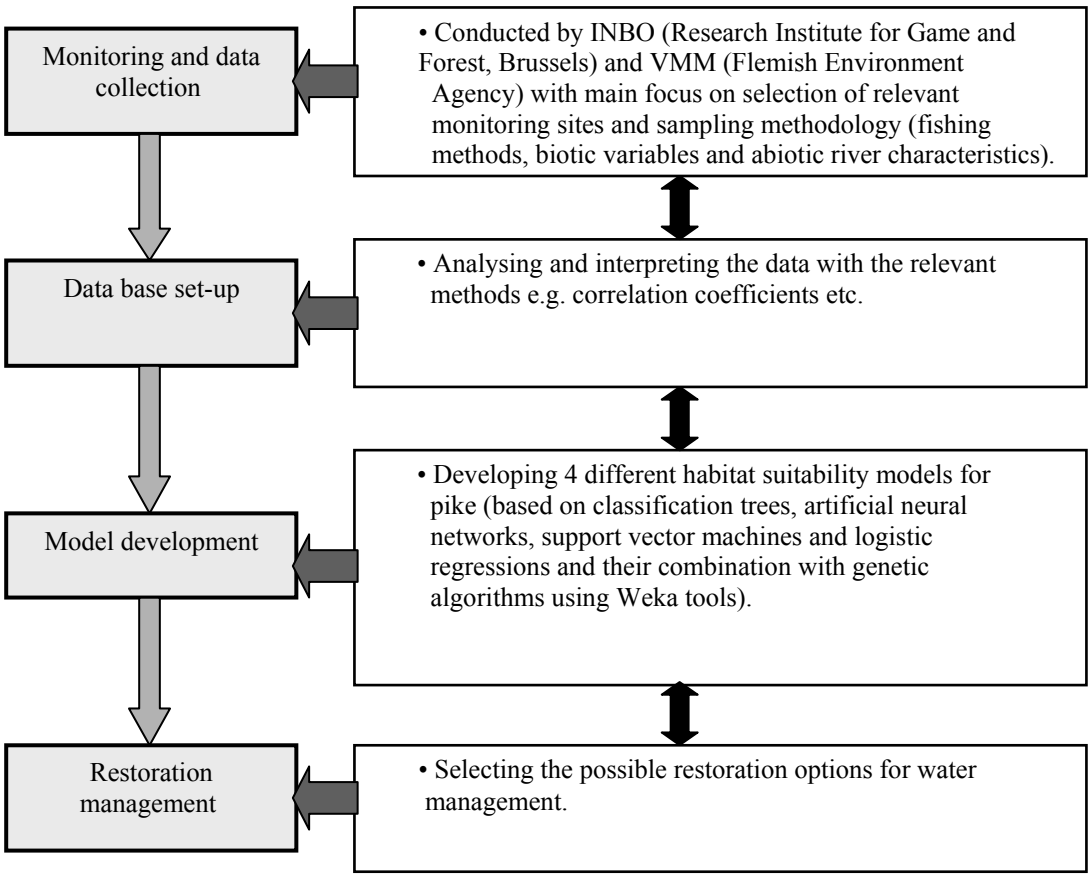


Figure 2.1. The conceptual framework for the development and application of ecological models for decision support in water management in Flemish river basins.

2.2 Monitoring and data collection methods

2.2.1 General characteristics of the river basins in Flanders and location of the monitoring sites for pike

Flanders is located in the northern part of Belgium. It has several major river basins (Figure 2.2). The main river basin covering almost the whole of Flanders is the Scheldt with a total surface of about 19141 km^2 (which is about 70 % of Flanders). For management aims, the Scheldt River basin itself is divided in several subbasins e.g. upper Scheldt, lower Scheldt, Leie, Nete, Demer, Dijle and canals around Ghent. The Ijzer, the second river basin, which is located in the coastal region, has a surface area of about 734 km^2. The third river basin is the one of the Meuse River, of which a smaller part (1730 km^2) is flowing in Flanders before continuing its course in the Netherlands. Most running waters in Flanders can be

considered as lowland streams. Only the lower part of the Scheldt and the Ijzer are estuarine with tidal influences. The lowland area (e.g. Brugse Polders) of Flanders is also characterized by a large number of ditches and slow flowing canals.

Many fresh water fish populations in these river basins decline because of pollution and habitat degradation and their migration paths being hampered by weirs, water mills and dams, which lead to isolated and fragmented populations and difficulties of recolonisation after restoration.

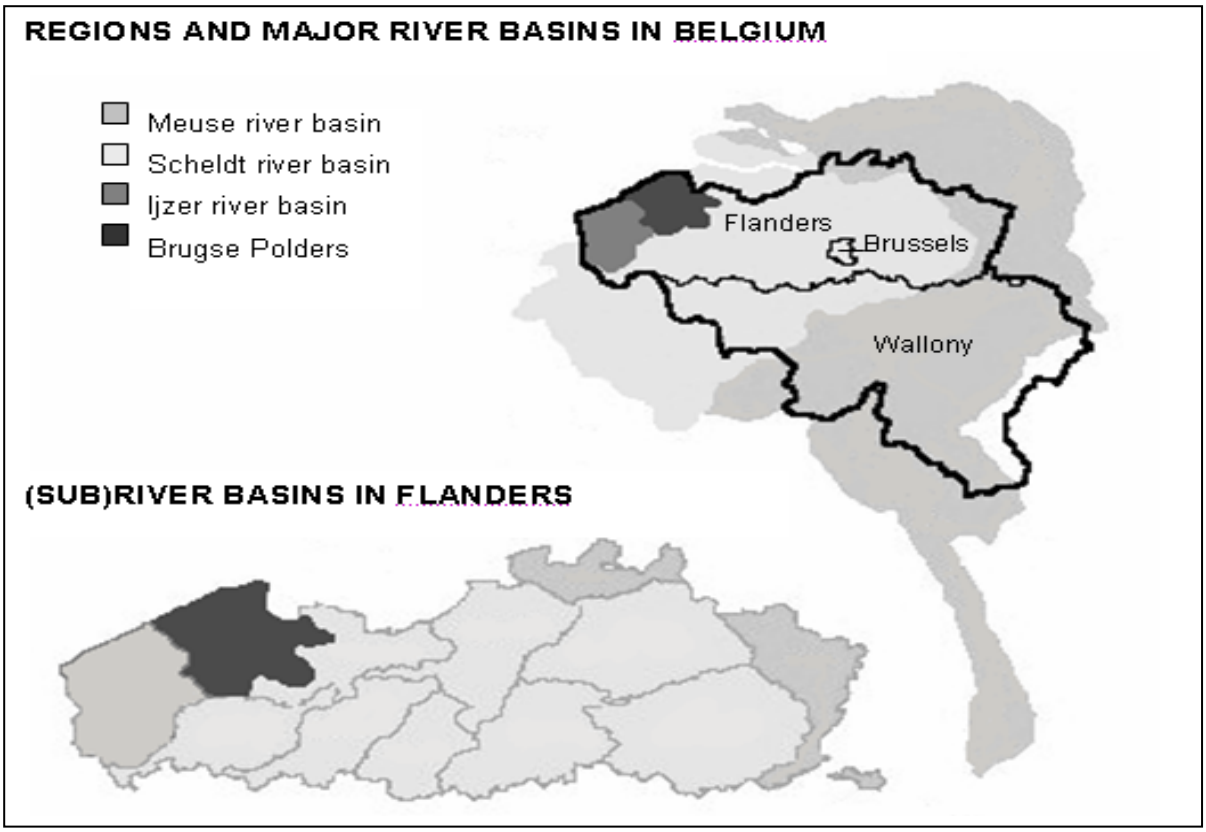


Figure 2.2. Location of main river basins in Flanders, Belgium.

Figure 2.3 shows the sites in 6 main river basins namely Demer, Dender, Dijle, Ijzer, Nete and upper Scheldt in Flanders in which pike populations were monitored. Within these 6 river basins, about 110 sites were sampled to be surveyed between 1991 and 2002. At the end, a total of 150 sets of observations were available for data analysis and model development. Some important site selection criteria could be considered by INBO (Institute for Nature and Forest Research, Brussels) and the Flemish Environment Agency (VMM) in order to:

- fulfil scientific and ecological requirements,
- be of relatively feasible access for electrofishing,
- be distributed over the important river basins,
- cover different river types based on Huet's zonation scheme (Huet, 1949, 1954) in Flanders.

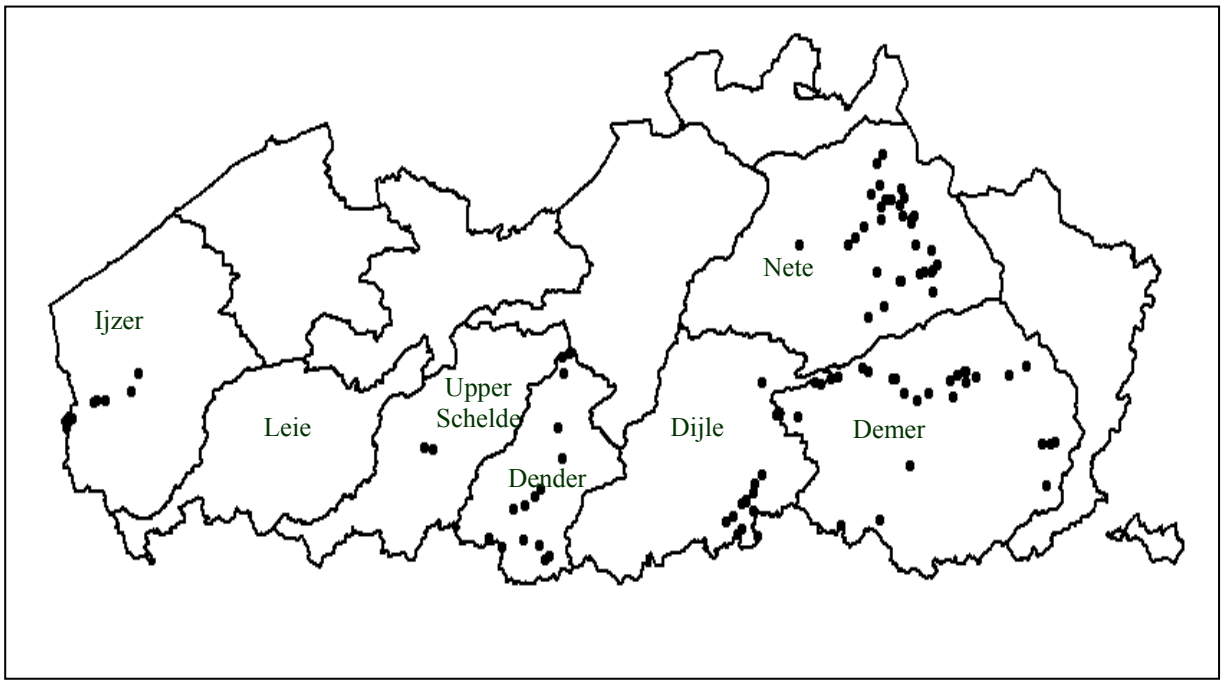


Figure 2.3. Monitoring sites (110) for pike in 6 river basins (Demer, Dender, Dijle, Ijzer, Nete and upper Scheldt) in Flanders during the period 1991-2002 (Leie was not monitored).

Table 2.1 shows the relative contribution of the sampling of pike to each river basin in Flanders during the period 1991-2002. As seen in this table, the number of samples was not equally distributed over different river basins. The Demer basin contributed to most of the pike monitored in the sampling sites (36.7 %), followed by the Nete basin (28.0 %). The Dender and Dijle collectively had the third rank (12 % for each of them). Ijzer and upper Scheldt had the lowest rank (9.3 % and 2.0 % respectively). Figure 2.4 also illustrates the relative contribution of each river basin for the monitoring of pike when taking presence/absence observations into account.

Table 2.1. Relative contribution of the sampling of pike for each river basin in Flanders (based on 150 datasets) during the period 1991-2002.

Flemish river basins	Total number and percentage of observations recorded in each river basin	Total number and percentage of observations in relation to pike presence/absence	
		Present	Absent
Demer	55 (36.7 %)	28 (37 %)	27 (37 %)
Nete	42 (28.0 %)	21 (28 %)	21 (29 %)
Dender	18 (12.0 %)	9 (12 %)	9 (12 %)
Dijle	18 (12.0 %)	9 (12 %)	9 (12 %)
Ijzer	14 (9.3 %)	7 (9 %)	7 (9 %)
Upper Scheldt	3 (2.0 %)	2 (2 %)	1 (1 %)
Total	**150 (100 %)**	**75 (100 %)**	**75 (100 %)**

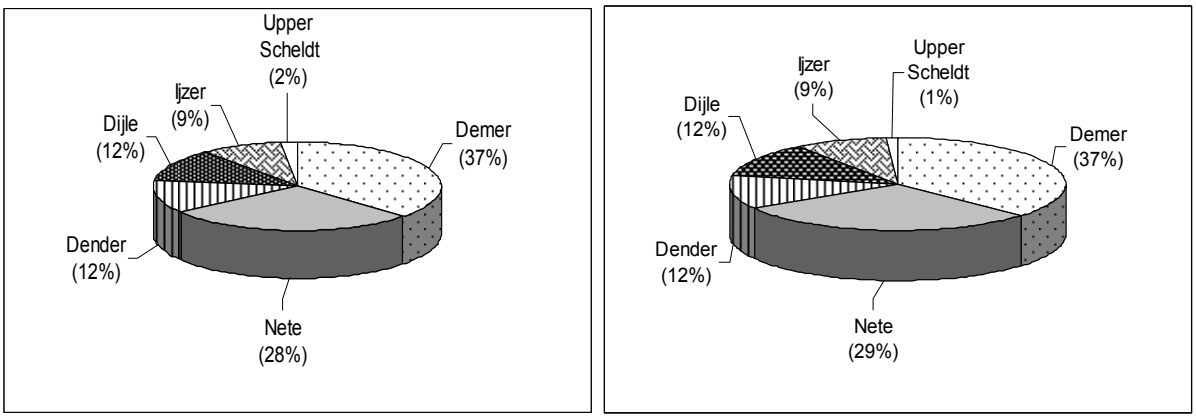


Figure 2.4. Relative contribution of the sampling of pike for each river basin (based on 150 datasets) during the period 1991-2002 (on the left side pike is present and on the right side pike is absent).

2.2.2 River typology classification

Huet's scheme of river typologies (Huet, 1949, 1954) is still currently used in Belgium. In this approach river slope and width are used to define the different river zones based on the fish community structure. Based on the Huet's zonation scheme, there are some specific zones e.g. the trout zone, grayling zone, barbel zone and bream zone.

In this study, the sampling sites were classified in 3 zones according to Breine et al. (2004), which is a slight modification of Huet zonation scheme (Huet, 1949, 1954): bream, barbel and upstream (trout and grayling zone). Of the sets of observations gathered in the Huet zonation, the maximum sampling of pike occurred in the bream zone (83), followed by the barbel zone (53) and the remaining samplings (14 observations) were recorded in the upstream zone (Table 2.2).

Of the 83 datasets located in the bream zone, pike occurred in 43 and was absent in 40 instances. From the total number of instances in the barbel zone (53), pike was present in 27 and was absent in 26 instances. The upstream zone had the lowest contribution for the monitoring of pike so that only 14 observations were found here. In this zone pike was absent in 9 while present in 5 instances. Therefore in the 3 zones, in 75 cases pike was present and in 75 cases pike was absent (Figure 2.5).

Table 2.2. Relative contribution of the sampling of pike (based on 150 datasets) according to Huet zonation scheme in 6 river basins in Flanders during the period 1991-2002.

Huet zonations as adapted by Breine et al. (2004)	Total number and percentage of data recorded in each zone	Total number and percentage of the data in relation to present and absent	
		Present	Absent
Bream zone	83 (55.4 %)	43 (57 %)	40 (53 %)
Barbel zone	53 (35.3 %)	27 (36 %)	26 (35 %)
Upstream*	14 (9.3 %)	5 (7 %)	9 (12 %)
Total	**150 (100 %)**	**75 (100 %)**	**75 (100 %)**

(*) Consisting of the trout and grayling zone as described by Huet (1949).

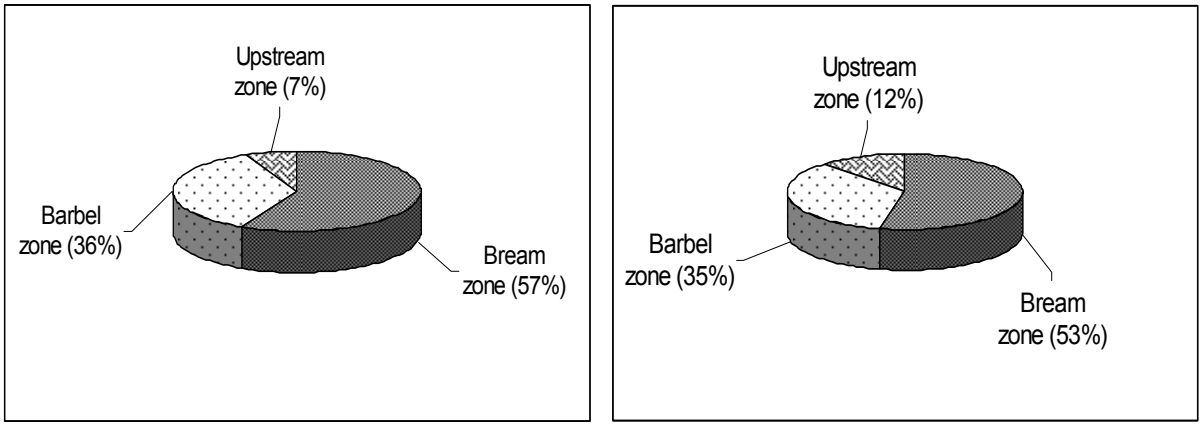


Figure 2.5. Relative contribution of the sampling of pike (based on 150 datasets) according to Huet-zonation scheme in the river basins in Flanders during the period 1991-2002 (on the left side pike is present and on the right side pike is absent).

2.2.3 Data collection methods

2.2.3.1 Available data concerning the river basins in Flanders for pike habitat suitability modelling

Since the last decades, a large number of data have been collected in Flemish watercourses. In spite of the huge amounts of data monitored annually in the river basins in Flanders, due to the different monitoring sites, sampling strategies and timings chosen for biological and environmental surveys, all the monitored data cannot be directly used for model

development purposes. Besides, the data are distributed over different institutes in Flanders using different format types, other co-ordinate systems etc. This was particularly problematic for the data of this study, of which most of the data could not be used to develop habitat suitability models, due to missing variables (e.g. water temperature, habitat quality etc).

2.2.3.2 Abiotic and biotic characteristics of Flemish river basins

The following environmental and biological variables were recorded such as slope, distance from source, wetted width, average depth, flow velocity, locality-length (the distance of length for electrofishing), locality-width (the distance of width for electrofishing), fished-area, pH, dissolved oxygen concentration, electric conductivity, nitrate, nitrite, ammonium, ortho-phosphate, total phosphorus, suspended solids, chloride, chemical oxygen demand (COD), biological oxygen demand (BOD) and the Belgian Biotic Index (BBI). Among the biological variables related to fish, only the presence and absence of pike was considered (Table 2.3).

Water temperature and habitat quality were not used in the model development because many data were missing. The variables locality-length, locality-width and fished-area were not directly considered for the model development, but were included in the data analysis to check their impact on the obtained data, related to the interpretation of the models and the general discussion regarding the data acquisition methods. To achieve this, a logistic regression model was made to see the impact of the locality-width (as an example) on pike presence/absence and how this variable could affect the success of the recorded variable in the datasets of Flanders.

The ranges of all the input variables were continuous, except for only one biotic variable, pike presence (1) and absence (0). In addition to these variables, some other information also was considered such as the Huet zonation, site-code, co-ordinates (x)-(y) and land-use. Site-code and co-ordinates (x)-(y) were employed for visualising the geographical distribution of pike in Flanders (Figure 2.3). The relevant map was created by means of ArcView 3.2a, which is a product of the Environmental System Research Institute (ESRI). In total 20 abiotic and 1 biotic variables could be used for the model development. Some structural habitat variables such as distance from the source and slope were measured using overlays in a geographical

information system (GIS) and a topographic map (Breine et al, 2004). Width and depth were checked and measured in the field. The physical-chemical variables were collectively determined in the field and laboratory on the basis of standardised and quality controlled methods (Table 2.3).

Table 2.3. Recorded river characteristics in the different river basins in Flanders.

River characteristics	**Units**
Slope	%
Distance from source	km
Width	m
Depth	m
Flow velocity	cm/s
Locality-length*	m
Locality-width*	m
Fished-area*	m^2
Land-use	5 classes (forest, industrial, arable-land, pasture, urban)
Huet-zonation	3 classes (bream, barbel, upstream)
pH	-log[H+]
Dissolved oxygen	mg/l
Conductivity	µS/cm
Nitrate	mg/l
Nitrite	mg/l
Ammonium	mg/l
Ortho-phosphate	mg/l
Total phosphorus	mg/l
Suspended-solids	mg/l
Chloride	mg/l
COD	mg/l
BOD	mg/l
Biological variable (BBI)	0 (extremely bad quality) to10 (very good quality)
Biological variable (pike)	0 (absence) and1 (presence)

(*) The variables fished-area, locality-length and locality-width were not used in the model development.

The surface of the fished-area for the monitoring of pike ranged from 48 m^2 to 2000 m^2. The distance of width for electrofishing had a minimum range of 0.5 m up to a maximum range of 10.5 m. The distance of length for electrofishing ranged from 40 m to 500 m. The variable fished-area is a choice of the field data collectors in order to collect relevant and representative data in an efficient manner. Actually within one sampling station, this variable is positively correlated to the probability of pike occurrence. In other words, probability of occurrence increases when a larger area is monitored. As illustrated in Figure 2.6, a trend can also be observed between the locality-width and the presence/absence of pike. Here, one can at first conclude that when the locality-width increases, the probability of pike occurrence decreases. Apparently this is in contrast with the obvious. In other words, an opposite trend would be expected based on expert knowledge. The reason why the actual data are in contrast with expected ones is probably based on an increased river pollution in these broader sites. This implies that probability of pike occurrence in the Flemish river basins not only depends on physical habitat variables as already described by Kerle et al. (2001), but also water quality can affect the presence of pike in Flanders.

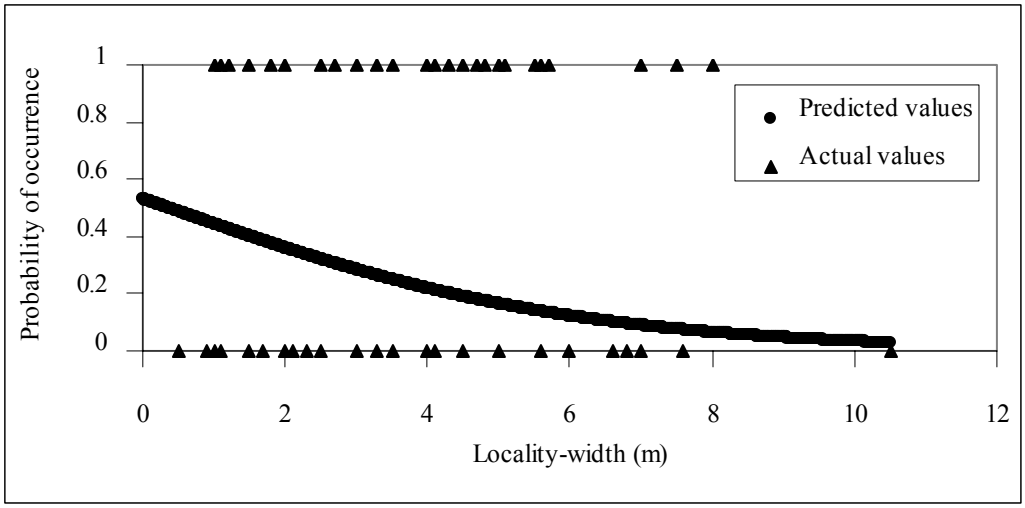


Figure 2.6. Probability of presence of pike in relation to locality-width in the river basins in Flanders (μ_y = 0.50).

2.2.3.3 Monitoring of pike in Flemish river basins

Pike data were obtained by electrofishing using a 5 kW generator with an adjustable output of 300 to 500 V and a pulse frequency of 480 Hz (Figure 2.7). In this method, two hand-held anodes (handnets) were employed. The type of anode used was ring-shaped and had a mesh size of 8 mm. Electrofishing was applied in barbel, bream and upstream zones within a distance of 100 m.

Figure 2.7. Electrofishing device (left) and handnet (right).

After monitoring in the field, pike was measured based on length and biomass and then was put back in the same site (Figure 2.8). The electrofishing method was conducted in the sampling sites in 6 major river basins in Flanders during the period 1991-2002 (Breine et al., 2004). The relevant biological database used in this study consisted of information about the presence and absence of pike.

	Fish species are caught by means of the electrofishing method.
	The length and biomass of fish are measured (in this figure only the pike's length is measured).
	After measuring, fish species are put back into the water.

Figure 2.8. Monitoring of pike by electrofishing with two handnets in Flemish river basin, measuring the pike's length and releasing the fish back into the water.

2.3 Database set-up and data analysis methods

The data analysis methods consisted of 3 parts: 1) calculation of bandwidth and distribution of input and output variables; 2) correlation analysis between input and output variables, and 3) visual relation analysis between the input and output variables. Each part is briefly described with emphasis on their application and importance. The results are presented in Chapter 3.

2.3.1 Bandwidth and distribution of input and output variables

According to Lek and Guégan (1999), data-driven models are unable to extrapolate beyond the range of the data used for training. Maier and Dandy (2000), on the other hand, stated that

poor predictions would result from the validation data containing values outside of the range of those used for training. Therefore, getting insight in the range of inputs and outputs is of initial and fundamental importance before starting the model development and application. To do so, calculation of some general statistical methods is important e.g. the calculation of the minima, maxima, mean, and standard deviations of the variables. By doing this, the results and analyses can be incorporated in visualisation graphs (see section 2.3.3). If these calculations are perfectly carried out, it is easy to understand whether or not high standard deviations are a result of a wide span of most data or are more related to some outliers. Also by comparing the median with the average, minimum and maximum can be a good indication. Therefore analysis of bandwidth and distribution of input and output variables is considered as a first and fundamental step before model development and application. The given data can also be displayed with boxplots (also known as box-and-whisker diagrams) because they are a convenient way of graphically depicting groups of numerical data through their five-number summaries (the smallest observation, lower quartile (Q1), median (Q2), upper quartile (Q3), and largest observation). They can also indicate which observations, if any, might be considered as outliers.

2.3.2 Correlation analysis between input and output variables

When considering the output of a model consisting of presence/absence, abundance, richness, diversity, density and etc. several performance measures such as the correlation coefficient (r) or determination coefficient (r) and the root mean squared error (RMSE) or a derivative between observed (O) and predicted (P) values are very often applied. Among these measures, verifying the correlation coefficients (the following formula) between a pair of input variables is very important.

$$Correlation\ coefficient = r = \frac{n\sum PO - (\sum P)(\sum O)}{\sqrt{n(\sum P^2) - (\sum P)^2}\sqrt{n(\sum O^2) - (\sum O)^2}}$$

Correlation analysis is a powerful statistical tool for assessing the relationship between two variables (Gotelli and Ellison, 2004). This statistical technique is frequently applied in bioassessment studies to determine significant relationships between or within biological and environmental measures.

In many ecological studies, the Pearson product moment correlation is broadly applied for the analysis of relationships between biological and environmental data (Legendre and Legendre, 1998). The Pearson correlation is a way to quantify the linear relationship between two random variables. It ranges from - 1 to + 1, indicating respectively a perfect negative or positive correlation. Normally, a value close to zero indicates that there is not any correlation and an increasing coefficient in absolute value means an increasing degree of correlation. Correlation coefficient values higher than 0.4 are considered as a good predictive performance indication for aquatic communities.

In addition to the correlation coefficient, it would be of great importance to simultaneously compute a measure of the significance of the relationship (the *p*-value). In ecological studies, *p*-values lower than 0.05 are considered to reflect enough significance of correlation coefficients to reject the null hypothesis and, in that case, one can accept that there is a significant linear association between the variables. In this present research, however, this measurement was not considered because it was beyond of the scope of this study. A potential drawback when using Pearson correlations is to use inconsistent data over a wide range of values to calculate correlation coefficient that can be skewed by a pair of large values of data, e.g. outliers (Rong, 2000). Therefore a precise viewing of the data to detect these extreme values should be performed previous to the correlation analysis, for instance by graphical (e.g. scatterplots and boxplots) and numerical (e.g. standard deviation).

According to Walczak and Cerpa (1999), a correlation coefficient with an absolute value of 0.20 or higher indicates a possible noise source in data driven models. Therefore, it is strongly recommended to remove one of these variables. In that case, if one of these highly correlated variables is left out, the model performance should not be affected. Nevertheless, from a practical application point of view, these correlated variables can be left in because their ecological relevancy is not pertinent.

2.3.3 Visual relation analysis between the input and output variables

A visual relation analysis between the input and output variables could have several advantages regarding getting insight into outliers, data clusters and missing or scarce variable combinations in certain ranges (Goethals, 2005). Therefore these methods can be very

important to realise the difficulty of the development of well performing models or finding out why some models do not fit to perform strongly and why some data can be classified as outliers. Recently analyses of these methods are becoming more and more popular and used as standard tools in most data mining and analysis software packages e.g. Weka software (Witten and Frank, 2000). These types of data analysis have two major advantages: 1) the distribution of the input variables, and 2) the observed output classes (e.g. presence and absence of pike) are clearly visualised for the particular species. This can be achieved for instance with two types of popular graphs e.g. histograms and scatter plots.

In summary, data visualisation is considered as an interesting method to get acquainted with the model performance as well as to realise what type of measurements must be carried out in the future to improve the dataset.

2.4 Modelling methods

2.4.1 General methodology

Models developed with the intention of describing the habitat of a specific organism in a predictive way are called habitat suitability models. The aim of the habitat suitability modelling approach is to determine either the suitability of a habitat for a particular species or use the information concerning the physical habitat of a species in order to predict its absence/presence or abundance. Often these models are also referred to as habitat distribution models or simply habitat models. Hirzel and Guisan (2002) suggested some important issues to improve habitat suitability modelling:

- enhance the sample size;
- choose systematic rather than random sampling;
- include environmental information in the design of the sampling strategy.

River mangers are curious to predict the impact of particular changes in land use, hydrology and geomorphology on the aquatic ecosystem and more specifically on the biological communities. Like many interdisciplinary efforts, problems arise when all these components need to be linked together. Hydrology, geomorphology and ecology differ fundamentally in the types of knowledge they produce, the space and time scales over which that knowledge can be applied and the format in which knowledge is communicated (Benda et al., 2002). In

addition to these interdisciplinary gaps, there still exist important intradisciplinary gaps, and these gaps in knowledge and data in each separate discipline prevent their easy linkage in an integrated forecasting model (Nilsson et al., 2003). Especially, this is true for the species-environment relationships part of the model e.g. ecological change as illustrated in Figure 2.9. Though many studies have explored the environmental responses of biological river communities to specific conditions, there is a clear need for models that quantify the species-environment relationships. The prediction of species distributions based on data describing the biotic environment called habitat suitability modelling has been recognised as a significant component of conservation planning (Franklin, 1995; Guisan and Zimmerman, 2000; Austin, 2002; Adriaenssens, 2004; Hein et al., 2007; Degraer et al., 2008) and is now considered as the core of predictive ecology. In effect, habitat suitability is the quantification of species-environment relationships. Undoubtedly, river managers can benefit from these predictive models as decision support tools to improve their efficiency in monitoring and assessment and active management for example, by choosing the most optimal solution for a given river restoration scenario. To do so, the cost-efficiency ratio for a given action or intervention could be kept at a low level.

The increased development of models is coherent with the view of 'more rigorously scientific, more informative and more useful ecology' (Peters, 1991). Therefore, prediction is considered as an aid in building scientific knowledge to get more insight leading to improved predictions. Moreover, on top of their role in decision support, the developed habitat suitability models can have an additional but not less important function in basic ecological research.

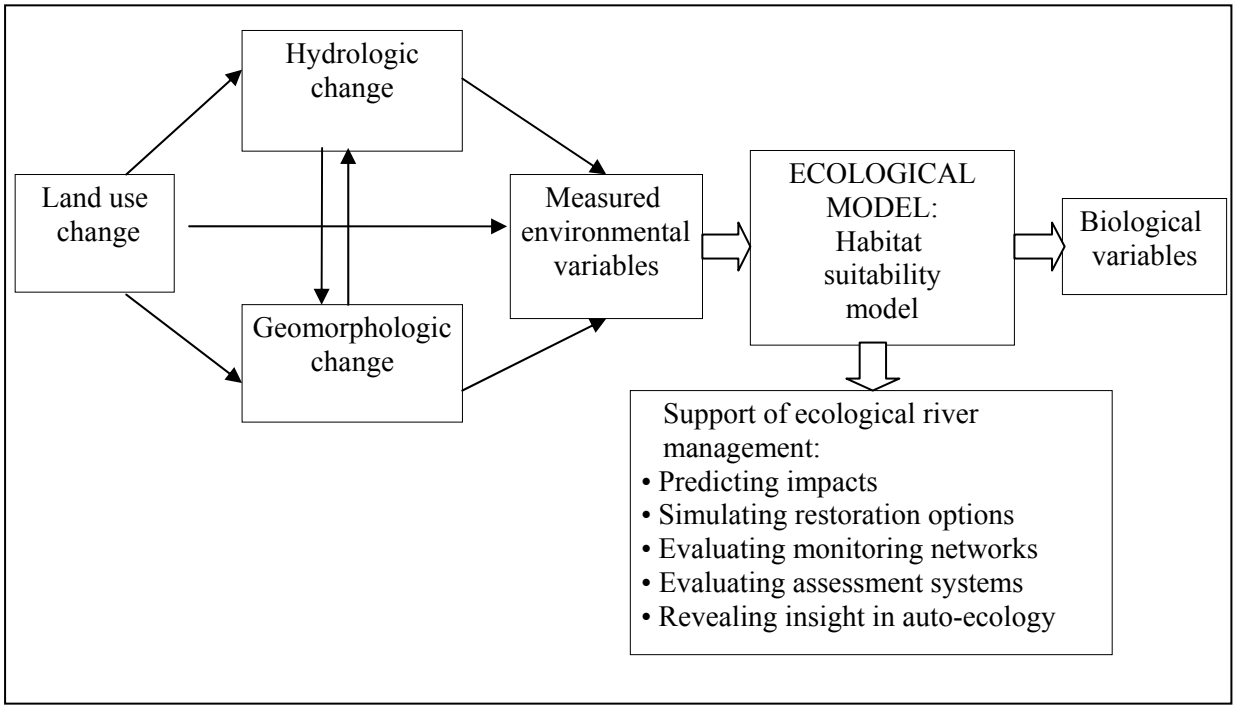


Figure 2.9. Components and objectives of the predictive habitat suitability models for use in ecological river management within the broader environmental modelling framework (Adriaenssens, 2004).

A wide array of habitat suitability models has been developed to cover aspects as diverse as biogeography, conservation biology, climate change research and habitat or species management (Guisan and Zimmermann, 2000). The selection of a proper method should not depend only on statistical performance considerations. Some models are more suitable to reflect theoretical findings on the shape and nature of the species' response, while others are more convenient to support decision making in river management because of a good visual interface and ease to understand the model principles. Verdonschot and Nijboer (2002) reviewed model design for habitat models. This demonstrates that many techniques were not yet applied to relate river characteristics with biological communities. There is still a serious lack of studies comparing different techniques. Therefore, it is still very hard to realise what are the strengths and weaknesses of the different techniques and under what conditions the specific techniques can be applied and are best performing. In addition to this, the data preparation and model validation processes are often performed in different ways, making the comparison even more difficult between the available studies (Goethals, 2005).

Figure 2.10 shows another scheme for habitat suitability model development. The first step comprises ecosystem component selection including input and output variables (left) to link habitat characteristics to community variables, followed by model training (middle) and model validation (right).

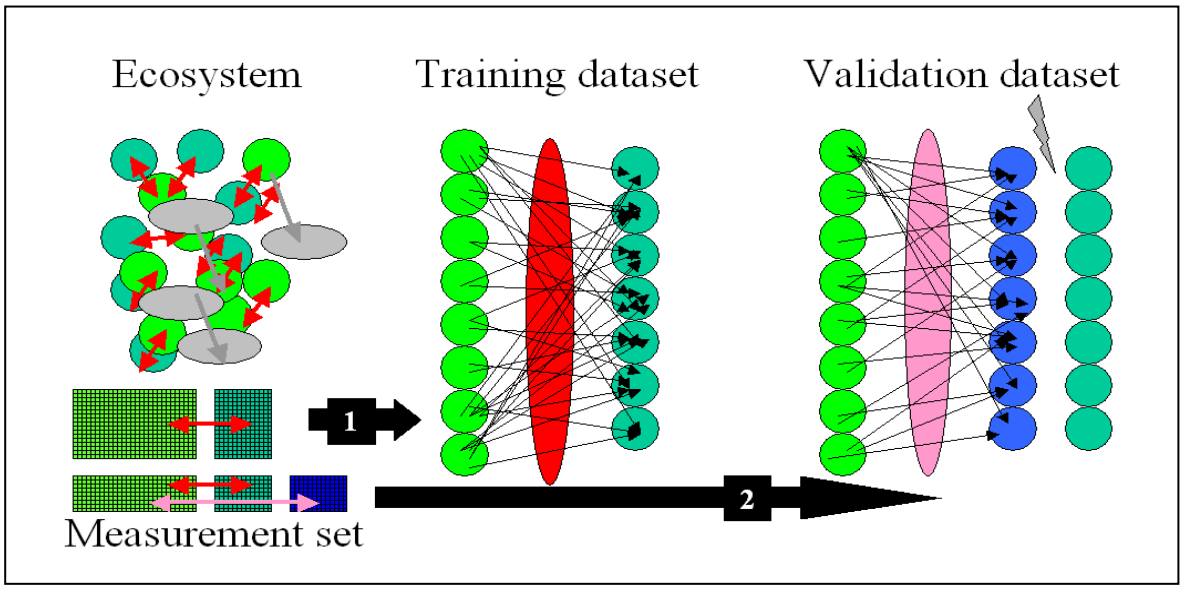


Figure 2.10. Habitat suitability model development scheme: first step consisting of ecosystem component selection to link habitat characteristics to community variables (left), model training (middle) and model validation (right) (Goethals, 2005).

As demonstrated already in chapter 1, some drawbacks could be observed in the work of Kerle et al. (2001) regarding the construction of the pike habitat suitability models by fuzzy logic methods. In particular, lack of field data and use of a limited set of input variables (mainly focused on structural-habitat variables) were the major constraints in the development of habitat suitability models for pike. Due to the lack of expert knowledge and long model development procedure related to these fuzzy logic models, this PhD research, analysed the relevance of a set of data driven approaches to predict pike in streams in Flanders. These techniques allow considering a much higher number of the input variables. In total, 20 variables could serve as inputs to these pike habitat suitability models. A selection was made based on a pool of structural-habitat and physical-chemical variables, as well as some integrating river characteristics (e.g. Huet-zonation, land-use and BBI). In the work of Kerle et al. (2001), pike models were specifically developed for different life stages. However, due

to the relatively small datasets available for this study (150 instances), there was opted for only presence/absence models for pike in general.

2.4.2 Dataset splitting for training and validation

2.4.2.1 General aspects of dataset splitting for training and validation

Very often, the high number of data needed is a major constraint in the training and validation phase. Especially for aquatic ecosystems, the data collection can be very expensive often leading to relatively small datasets. In other words, the extent of the folds (subsets) in each training and validation phase seems to be very crucial and should be balanced to ensure that the training and validation is carried out in a 'globally' optimal manner (Goethals, 2005).

If a lot of data is available, simply two independent samples will be taken: one for the training and one for the testing. More training will lead to an improvement of the model. More testing will result in more accurate error estimates but as mentioned already, the problem is that obtaining data is often expensive and time consuming. To cope with this problem, a possible solution could be to obtain a limited dataset and use a holdout procedure. In this procedure, a certain quantity of data is maintained for the testing set and the remainder for the training set. However, to use all the data for training and validation, a subset swapping method is commonly applied. This technique is known as 'cross-validation'. Cross-validation is the statistical practice of partitioning a sample of data into subsets such that the analysis is initially performed on a single subset, while the other subset(s) are retained for subsequent use in confirming and validating the initial analysis. This method is often applied for predicting the error rate of a learning algorithm. In the Weka software (Witten and Frank, 2000), a standard way is based on stratified 10 folds cross-validation. Nevertheless, data or time constraints can make 5 folds or 3 folds cross-validation more convenient.

According to Witten and Frank (2000), in most cases one-third of the data must be held out for the testing and the remainder (two-thirds) for the training. It is not important how the model evaluation is measured e.g. by precision, recall, mean squared error (MSE), CCI, Kappa or correlation etc, but it is important to measure on the testing set, not on the training

set. Performance on the training only implies that the model learns what it is supposed to learn. It is not a good indicator of performance on unseen data (Witten and Frank, 2000).

2.4.2.2 Settings of dataset splitting for training and validation in this study

In this thesis, a 3 folds cross-validation was applied on the all the dataset (150 instances). To achieve this, two-thirds of the dataset (100 instances) was allocated for the training and one-third (50 instances) for the validation set (Figure 2.11). As can be inferred from this figure, for making a 3 folds cross-validation, the first step is to split up the original dataset (150 instances) into the classes of presence and absence (in 75 instances pike was absent and in 75 instances pike was present) and then each class (presence/absence) is split in three equal parts. The second step is to create 3 folds (folds 1, 2 and 3) as indicated by different colours respectively. Figure 2.12 represents the optimal distribution of the 3 folds cross-validation in this study, which was obtained by trial and error. As shown here, when the training dataset increases, the validation dataset decreases.

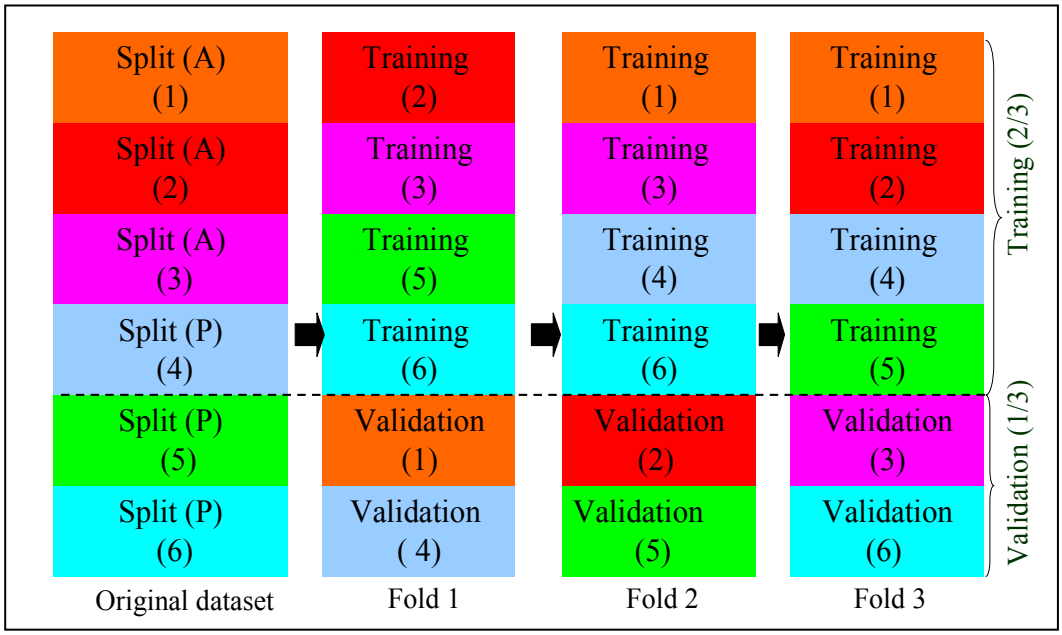


Figure 2.11. Data preparation of this study: splitting up the datasets for creating 3 folds cross-validation (A = absence, P = presence).

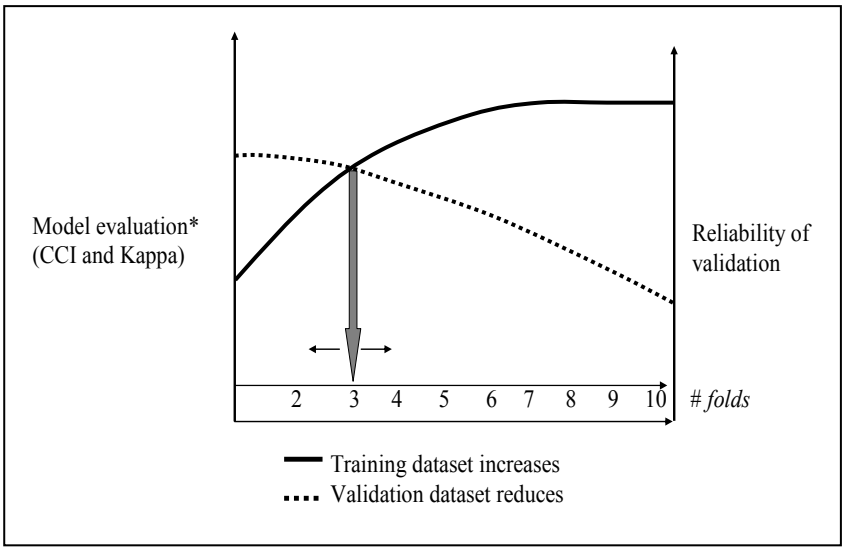


Figure 2.12. Optimal distribution of 3 folds cross-validation (see model evaluation in the next paragraph, CCI = correctly classified instances).*

2.4.3 Model evaluation

2.4.3.1 General overview of model evaluation

The precision of the classification of all ecological models and the quality of the model predictions can be confirmed by analysing the different model evaluations. Model evaluation is a key component of the model training and validation procedures. This can be an important starting point to assess the quality of the models. Model evaluation can deal with either continuous or discrete model outputs, or with both. If a model generates discrete predictions, these outputs can be represented in a confusion matrix (Fielding and Bell, 1997; Manel et al., 2001). A confusion matrix compares the model predictions to the observations (Table 2.4). A number of model evaluations have been derived from this confusion matrix (Table 2.5) e.g. the percentage of Correctly Classified Instances (*CCI*) (Buckland and Elston, 1993; Fielding and Bell, 1997), Cohen's Kappa (Cohen, 1960), Sensitivity (*Sn*), Specificity (*Sp*), the Normalized Mutual Information statistic (*NMI*) (Forbes, 1995), the odds ratio (Fielding and Bell, 1997) and True Skill Statistics (*TSS*) (Allouche et al., 2006). The odds ratio and true skill statistics range from -1 to 1 while all other criteria mentioned above range from 0 to 1 (it is 0, where models are completely inaccurate, and it is 1, where presence-absence is perfectly predicted). The percentage of Correctly Classified Instances (*CCI*) is the rate of correctly classified cells. Sensitivity (*Sn*) is the probability that model will correctly classify a presence. Specificity (*Sp*) is the probability that the model will correctly classify an absence. Normalized Mutual Information (*NMI*) quantifies the information included in the model predications compared to that included in the observations. The Cohen's Kappa statistics (*K*) and True Skill Statistics (*TSS*) normalise the overall accuracy by the accuracy that might have occurred by chance alone. In all formulae used in Table 2.5 the value N = a + b + c + d.

Table 2.4. The derivation of a confusion matrix as a basis for performance measures with true positive (TP), false positive (FP), false negative (FN) and true negative values (TN).

		Observed	
		Present	Absent
Predicted	Present	*a (TP)*	*b (FP)*
	Absent	*c (FN)*	*d (TN)*

Table 2.5. Measures based on the confusion matrix to assess the performance of presence/absence models (after Fielding and Bell, 1997). CCI is the percentage of Correctly Classified Instances, NMI is the normalized mutual information, TSS is the true skill statistics and N is the total number of instances.

Measure	Formula
Cohen's Kappa (*K*)	$\frac{[(a+d)-(((a+c)(a+b)+(b+d)(c+d))/N)]}{[N-(((a+c)(a+b)+(b+d)(c+d))/N)]}$
CCI	*(a + d)/N*
Sensitivity *(Sn)*	*a/(a + c)*
Specificity *(Sp)*	*d/(b + d)*
TSS	*Sn + Sp – 1*
Odds-ratio	*(ab)/(cd)*
NMI	$\frac{[-a.\ln(a)-b.\ln(b)-c.\ln(c)-d.\ln(d)+(a+b).\ln(a+b)+(c+d).\ln(c+d)]}{[N.\ln(N)-((a+c).\ln(a+c)+(b+d).\ln(b+d))]}$

2.4.3.2 Selection of model evaluation procedures in this study

During the last decades, several statistical methods have been developed to assess model evaluation (Zweig and Campbell, 1993). Among them, Cohen's Kappa has been frequently used in many studies. Cohen's Kappa is strongly recommended by many authors (e.g. Fielding and Bell, 1997; Manel et al., 1999; Degraer et al., 2008) as a reliable model evaluation procedure for assessing the presence/absence of organisms, because it is a more robust judge than simple per cent agreement calculation. The effect of prevalence on Cohen's kappa is negligible (Dedecker et al., 2004; Goethals, 2005; Degraer et al., 2008). In other words, Cohen's kappa is compensated for the prevalence of the entity to predict. It takes into account the chance that a sample would be attributed to a community. A high level of Kappa statistic occurs when the Kappa values are above 0.5 and a poor level is less than 0.3 (Cohen, 1960; Congalton, 1991).

The CCI is another model evaluation procedure, which is defined as the number of sites for which the modelled habitat suitability class was the same as the monitored one, divided by the total number of sites. When predicting the presence of freshwater organisms, the percentage of CCI would be a good index for evaluating the model. However, according to Fielding and Bell (1997), Manel et al. (1999) and Goethals et al. (2002), the frequency of occurrence of the test organism(s) being modelled, affects the correctly classified instances. For instance, Goethals et al. (2002) and Dakou et al. (2006) demonstrated that in some cases, when the taxa were very common or extremely rare, the CCI was very high while Kappa was very low.

In summary, however, different model evaluation procedures can evaluate the quality of model predictions, but a proper selection will be necessary to get more insight in which soft computing methods perform best for each type of problem. For that reason, a combination of both model evaluations (CCI and Kappa) was applied in this study to overcome this sort of problem. More information on the model evaluation can be found in the study of Fielding and Bell (1997).

2.4.4 Ecological model development

Several techniques such as classification trees (CTs), artificial neural networks (ANNs), support-vector machines (SVMs) and logistic regressions (LRs) were used for ecological modelling purposes. Moreover, genetic algorithms (GAs) were applied to select the relevant input variables. All these modelling techniques were constructed on the basis of the Weka software developed at the University of Waikato in New Zealand. Weka stands for *Waikato Environment for Knowledge Analysis* (Witten and Frank, 2000). It is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to datasets or called from one's own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules and visualisation. It is also well-suited for developing new machine learning schemes.

One way of using Weka is to apply a learning method to a dataset and analyse its output to learn more about the data. Another one is to use learned models to generate predictions on new instances. A third one is to apply several different learners and compare their performance in order to choose one for prediction. The learning methods are called *classifiers* and in the interactive Weka interface it is possible to select one from the menu. Many classifiers have tuneable parameters, which one can have access to through a property sheet or *object editor*. A common evaluation module is used to measure the performance of all classifiers. Some general aspects as well as the specific settings of these modelling methods are described in the following sections of this chapter.

2.4.4.1 Classification trees

2.4.4.1.1 General aspects of classification trees

Classification trees (Breiman et al., 1984), often referred to as decision trees (Quinlan, 1986), are efficient tools for the solution of classification and regression problems. Decision tree analysis is one of the main techniques used in so-called data mining. The common ways to perform the algorithms are top-down induction of decision trees (Quinlan, 1986). The C4.5-decision tree algorithm (Quinlan, 1993) is a supervised learning approach of machine

learning. The J48 algorithm (Weka.classifiers.trees.J48) is a Java reimplementation of C4.5 (Witten and Frank, 2000).

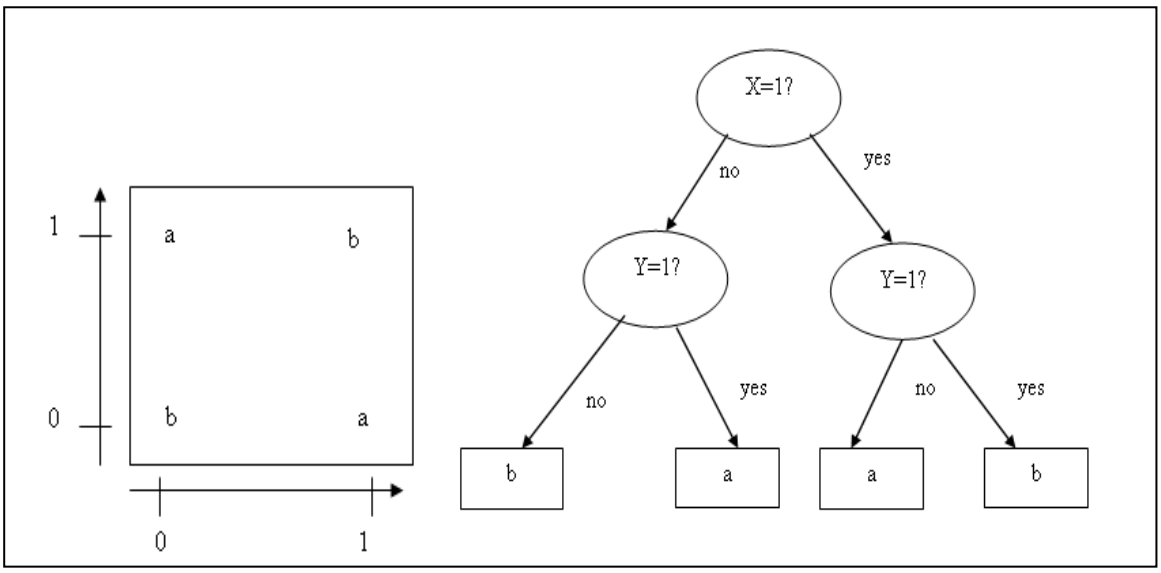
Classification trees categorise variables of a hierarchical decision scheme or multidimensional feature space into classes. In classification trees, a feature is checked by each internal node of a tree and a class or category is assigned by each leaf node and the arcs out of a node are labelled with the possible values of the features of this node. An important aspect in classification tree learning is the amount of branches. When there are many branches, the classification trees are difficult to interpret and often these last branches do not contribute significantly to the reliability of the tree.

When modelling species presence/absence, the procedure begins with the entire data set, also called the root node and formulates split-defining conditions for each possible value of the explanatory variables to create candidate splits. Next, the algorithm selects the candidate split that minimizes the misclassification rate and uses it to partition the data set into two subgroups. The algorithm continues recursively with each of the new subgroups until no split yields a significant decrease in the misclassification rate or until the subgroup contains a small number of observations.

In order to reduce the noise in the data and to improve the predictive results with regard to complexity and accuracy of the predictions, several optimisation methods can be applied like pruning, bagging and boosting. Pruning is a labour negations problem in classification trees. The simple classification trees perform better than the more complex ones and it makes more sense too (Witten and Frank, 2000). The confidence factor, which is often used for this purpose, is a parameter that has an effect on the error rate estimate in each node. When the confidence factor is increased, the difference between the error estimate of a parent node and its splits decreases. In this way, it is less likely that the split will be pruned. The smaller the value of the confidence factor is, the larger is the difference between the error rate estimates of a parent node and its potential splits. Optimal pruning is an important mechanism as it improves the transparency of the induced trees by reducing their size as well as enhances their classification accuracy by eliminating errors that are present due to noise in the data.

An important advantage of classification trees is that relationships and hidden structures in the data are uncovered because it is possible to derive general rules for the absence or presence of organisms based on trees constructed in each subset. General rules by means of 'IF and THEN' can be subtracted from classification trees by following the pathway from a leave to the root of a tree.

Figure 2.13 presents simple classification rules which are a popular alternative for decision trees (Witten and Frank, 2000). The diagram on the left of this figure shows an *exclusive-or* function for which the output is *a* if x = 1 or y = 1 but not both. To make this into a tree, one has to split on one attribute first, leading to a structure like the one shown on the right. In contrast, rules can reflect the true symmetry of the problem with respect to the attributes, as shown at the bottom of the figure.



If x = 1 and y = 0 then class = a

If x = 0 and y = 1 then class = a

If x = 0 and y = 0 then class = b

If x = 1 and y = 1 then class = b

Figure 2.13. The exclusive-or function (Witten and Frank, 2000).

2.4.4.1.2 Specific parameter settings in this study

The default settings were applied based on trial and error except for the pruning confidence factors (PCFs), which had a very important effect on the selected variables and the number that were used for the classification. This factor was tested at four levels: 0.5, 0.25, 0.1 and 0.01 (default = 0.25). Therefore, the following settings were developed for the classification trees (Witten and Frank, 2000):

- 'binary splits' (whether to use binary splits on nominal attributes when building the trees): false;
- 'pruning confidence factor' (the confidence factor is used for pruning; smaller values indicate more pruning): 0.5, 0.25, 0.1 and 0.01;
- 'minimum number of instances per leaf': 2;
- 'number of folds' (determines the amount of data used for reduced-error pruning: one fold is used for pruning, the rest for growing the tree): 3;
- 'reduced error pruning' (whether reduced-error pruning is used instead of C.4.5 pruning): false;
- 'seed' (the seed used for randomising the data when reduced-error pruning is used): 1;
- 'subtree raising' (whether to consider the subtree raising operation when pruning): true;
- 'unpruned' (whether pruning is performed): false.

2.4.4.2 Artificial neural networks

2.4.4.2.1 General aspects of artificial neural networks

Artificial neural networks (ANNs) are non-linear mapping structures that can be applied for predictive modelling and classification (Patuwo et al., 1993; Piramuthu et al., 1994; Kanellopoulos and Wilkinson, 1997). The choice of the type of network depends on the nature of the problem to be solved. The most popular ANNs are multilayer feedforward neural networks with a backpropagation algorithm e.g. backpropagation networks (Rumelhart et al., 1986; Hagan et al., 1996) and Kohonen self-organizing maps (SOMs) (Kohonen, 1982 and 1995).

The backpropagation network is based on the 'supervised' procedure. The network constructs a model based on examples of data with known outputs. To make reliable predictions, it is better to rescale the input variables, because they can have very different orders of magnitude. The architecture of the backpropagation network is a layered feed-forward neural network in which the non-linear elements, the neurons, are arranged in successive layers and the information flows from input layer to output layer, through the hidden layer(s) (Figure 2.14). As can be seen from this figure, nodes from one layer are connected to all nodes in the following layer, but no lateral connections within any layer, nor feedback connections are possible. With the exception of the input neurons, which merely connect one input value with its associated weight values, all neurons can be visualised with their connections (Figure 2.15). The inputs are indicated as x_1, x_2, … x_n, each associated with a quantity called weight or connection strength w_{j1}, w_{j2}, … w_{jn} for the input to the j-th neuron. The net input for each neuron is the sum of all input values, each multiplied by its weight and z_j a bias term which may be considered as the weight from a supplementary input equalling one:

$$a_j = \sum w_{ji}x_i + z_j$$

The output value, y_j, can be calculated by feeding the net input into the transfer function of the neuron:

$$y_j = f(a_j)$$

The number of input and output nodes depends on the number of the input and output objects. For determining the values of weights and biases in a backpropagation network, all the weights and biases are initially set to small random numbers. Subsequently, a set of input/output ensembles is presented to the network. For each of the input sets, the output is calculated by the ANNs and an error term is calculated by comparing the calculated output with the desired output (the 'target'). Using this error term, the weights and biases are updated in order to decrease the error, so future outputs are more likely to be correct. This procedure is repeated until the errors become small enough or a predefined maximum number of iterations is reached. This iterative process is termed 'training'. After the training, the ANNs can be

tested using independent data. A more detailed description of ANNs can be found in Lek and Guegan (1999).

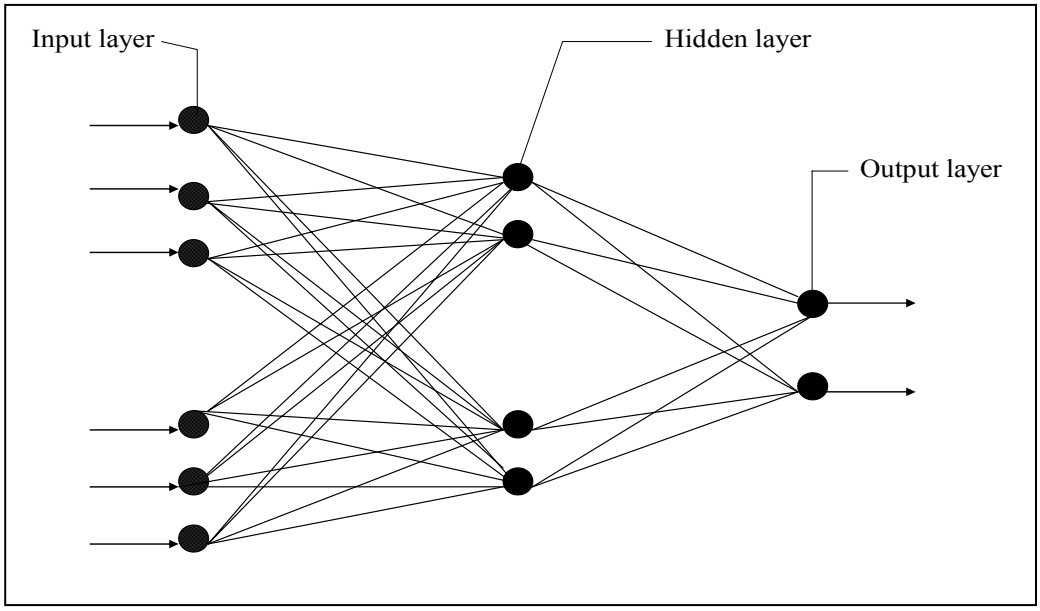


Figure 2.14. Illustration of a three-layered artificial neural network, with one input layer, one hidden layer and one output layer (Lek et al., 2004).

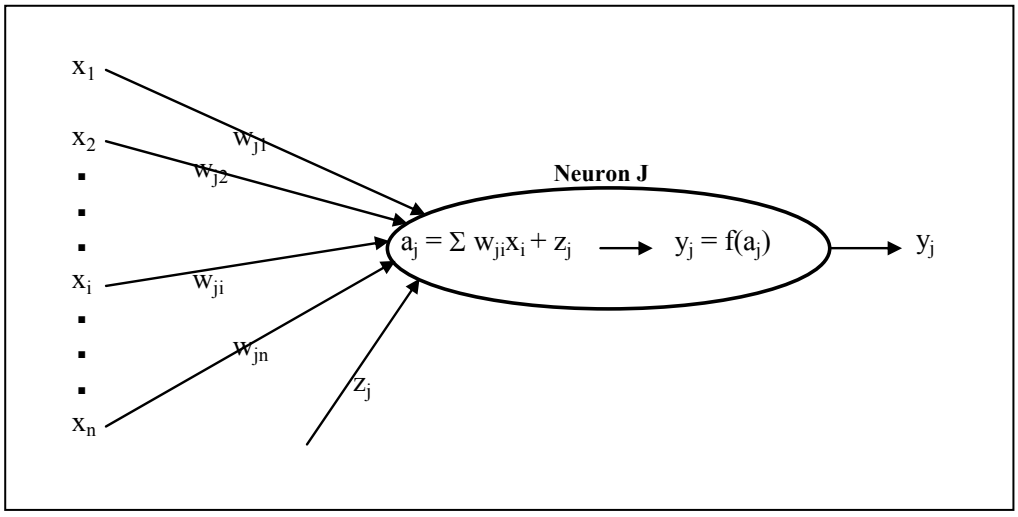


Figure 2.15. Scheme of a neuron in a backpropagation network receiving input values from n neurons, each associated with a weight, as well as a bias z_j. The resulting output value y_j is computed according to the presented equations (Goethals, 2005).

2.4.4.2.2 Data processing and bandwidth

Different variables span different ranges. In order to ensure that all variables receive equal consideration during the training process, it would be of great importance to standardise them. Besides, the variables can be scaled in such a way as to be proportionate with the limits of the activation functions used in the output layer (Maier and Dandy, 2000). For instance, the data can proportionally be normalised between zero and one [0 1] in the range of the maximum and minimum values (Schleiter et al., 1999, 2001; Wagner et al., 2000; Chon et al., 2001, 2002; Gabriels et al., 2002).

According to Lek and Guégan (1999), ANNs are built solely from the examples presented during the training phase, which are together assumed to implicitly contain the information necessary to establish the relation between input and output. As a result, ANNs are unable to extrapolate beyond the range of the data used for training. Consequently, poor predictions can be expected when the validation data contains values outside of the range of those used for training (Maier and Dandy, 2000).

2.4.4.2.3 Learning method

Artificial neural networks can be divided into supervised and unsupervised training methods. Unsupervised learning methods do not need output values for the training process and are mainly used for classification problems. The suitability of a particular method is often a trade-off between performance and calculation time. The majority of the ANNs used for prediction are trained with the backpropagation method (e.g. Cherkassky and Lari-Najafi, 1992; Maier and Dandy, 2000). Because of its generality (robustness) and ease of implementation, backpropagation is the best choice for the majority of ANNs. Backpropagation is the superior learning method when a sufficient number of relatively noise-free training examples are available, regardless of the complexity of the specific domain problem (Walczak and Cerpa, 1999). In order to optimise the performance of backpropagation networks, one has to know that the performance is a function of several internal parameters including the transfer function, error function, learning rate and momentum term. The learning rate is directly proportional to the size of the steps taken in weight space. Traditionally, learning rates remain

fixed during training (Maier and Dandy, 2000) and optimal learning rates are determined by trial and error. A momentum term is usually included in the training algorithm in order to improve learning speed (Qian, 1999) and convergence (Hagan et al., 1996). The momentum term should be less than 1.0, otherwise the training procedure does not converge (Dai and Macbeth, 1997). Dai and Macbeth (1997) suggest a learning rate of 0.7 with a momentum term of at least 0.8 and smaller than 0.9 or a learning rate of 0.6 with a momentum term of 0.9.

2.4.4.2.4 Model architecture

The generalisation capability of a neural network is influenced by three factors: the size of the training set and how representative it is of the environment of interest, the architecture of the neural network and the complexity of the problem (Haykin, 1999). The architecture is the only of these three factors that can be influenced in the modelling process, making it a crucial step, which should be considered carefully. Four design criteria have been distinguished for artificial neural networks which should be decided upon in subsequent steps: knowledge-based selection of input values, selection of a learning method, design of the number of hidden layers, and selection of the number of hidden neurons for each layer (Walczak and Cerpa, 1999).

2.4.4.2.5 Input variable selection

ANNs have the ability to determine which model inputs are crucial. Nevertheless, presenting a great number of inputs to ANNs and relying on the network to determine the critical model inputs, generally increases the network size. This has a number of disadvantages, for example decreasing processing speed and increasing the amount of data required to estimate the network parameters efficiently (Maier and Dandy, 2000). In this way, selection of input variables can be considered as an important task. It can considerably reduce the necessary labour of data collection.

Several methods can be followed to determine the optimal set of input variables. The first one is to perform standard knowledge acquisition. ANNs input variables should not be correlated.

Correlated variables degrade the ANNs' performances by interacting with each other as well as with other elements to produce a biased effect.

A first pass filter to help identify 'noise' variables is to calculate the correlation of pairs of variables. If two variables are strongly correlated, then one of these two variables can be removed without adversely affecting the ANNs performance. The cut-off value for variable elimination is a heuristic value and must be determined separately for every ANN application, but any absolute correlation value of 0.20 or higher indicates a probable noise source to the ANNs (Walczak and Cerpa, 1999).

Besides, there are distinct advantages in using analytical techniques to help to determine the inputs of ANN models (Maier and Dandy, 2000). Obach et al. (2001) and Beauchard et al. (2003) used a stepwise procedure to identify the most influential variables. A third technique which is frequently used is genetic algorithms (e.g. Obach et al., 2001; Schleiter et al., 2001; D'heygere et al., 2003). In this study, only genetic algorithms were used to select the input variables for ANNs as well as other modelling techniques. For this reason, other techniques of input variable selection will not be further considered.

2.4.4.2.6 Number of hidden layers and hidden neurons

A greater number of hidden layers enables ANNs to improve their closeness-of-fit, while a smaller quantity improves the smoothness or extrapolation capabilities of the ANNs (Walczak and Cerpa, 1999). ANNs with one hidden layer can approximate any function as long as sufficient neurons are used in the hidden layer (Hornik et al., 1989). Flood and Kartam (1994) suggest using two hidden layers as a starting point.

The number of neurons in the input layer is fixed by the number of model inputs, whereas the number of neurons in the output layer equals the number of model outputs. The crucial factor, however, is the choice of the number of neurons in the hidden layer. More hidden neurons bring about a longer training period, while fewer hidden neurons provide faster training at the cost of having fewer feature detectors (Bebis and Georgiopoulos, 1994). For two networks with similar errors on training sets, the simpler one (the one with fewer hidden units) is likely to produce more reliable predictions in new cases, while the more complex model implies an

increased chance of overfitting of the training data and reducing the model's ability to generalise on new data (Hung et al., 1996).

Many authors suggest rules of thumb for determining the number of hidden neurons. Some of these rules are based on the number of input and/or output neurons, while others are based on the number of training samples available. Table 2.6 shows the rules that suggest the number of hidden neurons based on the number of input (N_i) and/or output (N_o) nodes.

Table 2.6. Rules suggesting the number of hidden neurons based on the number of input (N_i) and/or output (N_o) nodes.

Rule	Reference
$(2/3) * N_i$	Wang, 1994
$0.75 * N_i$	Lenard et al., 1995
$0.5 * (N_i + N_o)$	Piramuthu et al., 1994
$2 * N_i + 1$	Fletcher and Goss, 1993; Patuwo et al., 1993
$2 * N_i$ or $3 * N_i$	Kanellopoulos and Wilkinson, 1997

Some rules are suggested to determine the necessary number of training samples (S) based on the number of connection weights. Given the number of training samples is fixed, inverting these rules can give an indication of the maximum number of connection weights to avoid overfitting (Table 2.7).

Table 2.7. Indication of the maximum number of connection weights to avoid overfitting based on the number of training samples (S).

Maximum number of connection weights	Reference
S	after Rogers and Dowla, 1994
S/2	after Masters, 1993
S/4	after Walczak and Cerpa, 1999
S/10	after Weigend et al., 1990
S/30	after Amari et al., 1997

The number of hidden neurons necessary can be calculated given the number of connection weights and the number of input and output neurons. According to Walczak and Cerpa

(1999), the number of hidden neurons in the last layer should be set equal to the number of decision factors used by domain experts to solve the problem.

2.4.4.2.7 Model interpretation

In many studies, ANNs have shown greater predictive power compared to traditional approaches, but they have also been labelled as a 'black box' because they give little explanatory insight into the relative influence of the independent variables in the prediction process (Olden and Jackson, 2002). This lack of explanatory power is a major concern to ecologists since the interpretation of statistical models is desirable for gaining knowledge of the causal relationships driving ecological phenomena (Karul et al., 2000). As a result, many authors proposed several algorithms to illustrate the role of variables in ANNs models like sensitivity analysis (Olden and Jackson, 2002; Brosse et al., 2003) and is based on a successive variation of one input variable while the others are kept constant at a fixed value (Lek et al., 1995).

2.4.4.2.2 Specific parameter settings in this study

The default settings were applied for ANNs (weka.classifiers.functions.MultilayerPerceptron), except for the number of hidden layers which was set at 1 and the number of neurons in this layer. This factor was tested between 2 to 9 and thus optimised via trial and error. Other settings in the ANNs toolbox were default and were determined through experience.

The following settings were developed for the ANNs (Witten and Frank, 2000):

-'GUI' (brings up a Graphical User Interface (GUI). This will allow the pausing and altering of the neural network during training. If the GUI is not set, the network will not require any interaction): false;
-'auto build' (adds and connects up hidden layers in the network): true;
-'debug' (if set to true, classifier may output additional information to the console): false;
-'decay' (this will cause the learning rate to decrease. This will divide the starting learning rate by the epoch number, to determine what the current learning rate should be): false;
-'learning rate' (the amount the weights are updated): 0.3;

-‘momentum’ (momentum applied to the weights during updating): 0.2;
-‘nominal to binary filter’ (this will preprocess the instances with the filter. This could help improve performance if there are nominal attributes in the data): true;
-‘normalise attributes’ (this will normalise the attributes. This could help improve performance of the network): true;
-‘normalise numeric class’ (this will normalise the class if it is numeric. This could help improve the performance of the network. It normalises the class to be between -1 and 1): true;
-‘random seed’ (seed used to initialise the random number generator. Random numbers are used for setting the initial weights of the connections between nodes, and also for shuffling the training data): 0;
-‘reset’ (this will allow the network to reset with a lower learning rate): true;
-‘training time’ (the number of epochs to train through; if the validation set is non-zero then it can terminate the network early): 500;
-‘validation set size’ (the percentage size of the validation set): 0;
-‘validation threshold’ (used to terminate validation testing. The value here dictates how many times in a row the validation set error can get worse before training is terminated): 20.

2.4.4.3 Support-vector machines (SVMs)

2.4.4.3.1 General aspects of support-vector machines

Support-vector machines (SVMs) were developed by Vapnik (1995). They are a new generation of learning algorithms that implement Platt's sequential minimal optimisation algorithm for training a support-vector classifier (Keerthi et al., 2001). This implementation replaces all missing values and transforms nominal attributes into binary ones. Multi-class problems are solved using pairwise classification (Witten and Frank, 2000). They consist of input and output layers connected with weight vectors. The arrangements of neurons in the input layer gets the input vectors while the output layer comprises a two-dimensional network of neurons arranged on a hexagonal net.

In the past few years, there has been a lot of interest in support-vector machines (Vapnik, 1995; Burges, 1998; Keerthi, 2001) because they have yielded excellent generalization

performance on a wide range of problems (Keerthi, 2001). A remarkable feature of support-vector machines is that they are extremely easy to implement. They are often much faster and have better scaling properties (Keerthi, 2001). They make very competitive results with the best accessible classification methods and need only the smallest amount of model tuning (Decoste and Scholkopf, 2002; Guoa et al., 2005). An advantage of SVMs over artificial neural networks is that with SVMs, the function to be minimised is very well shaped. During the learning process of the SVMs if an input vector x is presented through the network, every output neuron k of the network will be able to calculate the distance between the input vector x and the weight vector w.

Support-vector machines have been designed for two-class problems including positive and negative objects (Guoa et al., 2005). They can be used for classification and regression. A complete analysis of SVMs entails three steps: model selection, fitting and validation. The basic idea of SVMs is to transform the data in such a way they become more or even completely linearly separable, and to perform a linear separation (Figure 2.16), that is, there is a hyperplane in instance space that classifies all training instances correctly. The maximum margin hyperplane is the one that gives the greatest separation between the classes (marked by filled and open circles). The instances that are closest to the maximum margin hyperplane are called support-vector. There is always at least one support-vector for each class and often there are more. The set of support-vectors uniquely defines the maximum margin hyperplane for the learning problem. Given the support-vectors for the two classes, it is possible to easily construct the maximum margin hyperplane. All other training instances are irrelevant (they can be deleted without changing the position and orientation of the hyperplane). Among all hyperplanes that separate the classes, the maximum margin hyperplane is the one that is as far away as possible from both convex hulls (it is the perpendicular bisector of the shortest line connecting the hulls, which is shown dashed in Figure 2.16).

A hyperplane separating the two classes is written in the two-attribute case, where a_1 and a_2 are the attribute values and there are three weights w_i to be learned:

$$x = w_0 + w_1a_1 + w_2a_2$$

Nevertheless, the equation defining the maximum margin hyperplane can be written in another form, in terms of the support-vectors. The class value y of a training instance is written as either 1 (for *yes*, it is in this class) or -1 (for *no*, it is not in this class). Then the maximum margin hyperplane is:

$x = b + \Sigma_i \alpha_i y_i \mathbf{a}(i).\mathbf{a}$

where y_i is the class value of training instance $\mathbf{a}(i)$; while b and α_i are numeric parameters that have to be determined by the learning algorithm and **a**(i) and **a** are vectors. The vector **a** represents a test instance-just as the vector $[a_1, a_2]$ represented a test instance in the earlier formulation. The vector $\mathbf{a}(i)$ is the support-vector (those circled in Figure 2.16), which is a selected member of the training set. The term $\mathbf{a}(i).\mathbf{a}$ represents the dot product of the test instance with one of the support-vectors.

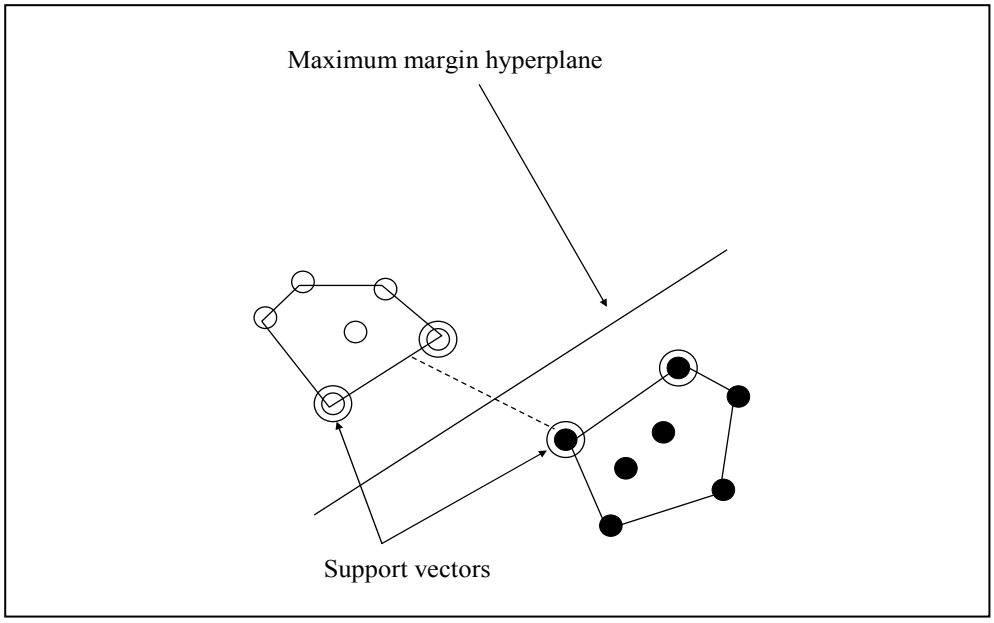


Figure 2.16. A maximum margin hyperplane for support-vector machines.

Figure 2.17 represents a particular example of support-vector machines for two fish species: the black dots represent sardines, the open circles are herrings, and the two x-variables are size and weight, respectively. The data in Figure 2.17 (a) are linearly separable. It is possible to draw a straight line, such that all open circles are on one side of the line, and all filled circles are on the other side. In Figure 2.17 (b), however, the data cannot be separated by a straight line. In this situation, it would help if a (nonlinear) transformation could be found,

such that the transformed data were again linearly separable or at least more linearly separable (Figure 2.17 c).

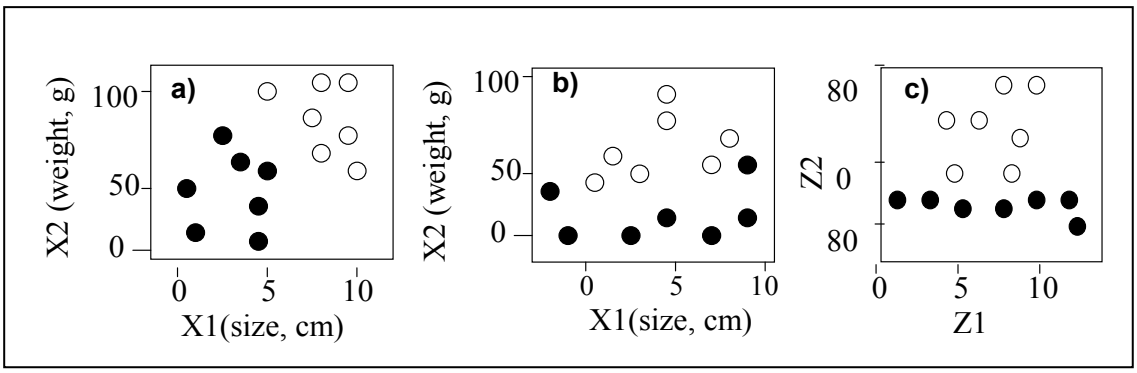


Figure 2.17. Linear and nonlinear separation of herrings and sardines. (a) Linear separable herrings and sardine, (b) linear separation is not possible and (c) after suitable transformation of x, linear separation is again possible (Lek et al., 2004).

Four standard kernels are usually used in classification problems and also used in regression cases: 1) linear kernel; 2) polynomial kernel; 3) radial basis function (RBF) network and 4) sigmoid. A support-vector machine with the RBF network and one with a sigmoid kernel implements another type of neural network, a multilayer perceptron with one hidden layer. The last three kernel functions are used to implement different non-linear mappings. Which one produces the best results depends on the application, although the differences are rarely large in practice (Witten and Frank, 2000). In comparison with other methods, such as decision tree learners, even the fastest training algorithms for support-vector machines are slow when applied in non-linear settings.

2.4.4.3.2 Specific parameter settings in this study

The default settings were used for support-vector machines (weka.classifiers.functions.SMO) except for the exponent values. To achieve this, the exponent was repeated 10 times.
The following settings were developed for the SVMs (Witten and Frank, 2000):
-‘build logistic models’ (whether to fit logistic models to the outputs): false;
-‘c’ (the complexity parameter C): 1.0;

-‘cache size’ (the size of the kernel cache; this should be a prime number. 0 is used for full cache): 250007;
-‘debug’ (if set to true, classifier may output additional information to the console): false;
-‘epsilon’ (the epsilon for round-off error. This should not be changed): 1.0E-12;
-‘exponent’ (the exponent for the polynomial kernel): 1.0;
-‘feature space normalisation’ (whether feature-space normalization is performed (only available for non-linear polynomial kernels): false;
-‘filter type’ (determines how/if the data will be transformed): normalize training data;
-‘gamma’ (the value of the gamma parameter for RBF kernels): 0.01;
-‘lower order terms’ (whether lower order polynomials are also used; this is only available for non-linear polynomial kernels: false;
-‘number folds’ (the number of folds for cross-validation used to generate training data for ‘logistic models’): -1;
-‘random seed’ (random number seed for the cross-validation): 1;
-‘tolerance parameter’ (the tolerance parameter; this should not be changed): 0.0010;
-‘use RBF’ (whether to use a radial basis function (RBF) kernel instead of a polynomial one): false.

2.4.4.4 Logistic regressions (LRs)

2.4.4.4.1 General aspects of logistic regressions

Logistic regressions can be used with two types of target variables: 1) a categorical target variable that has exactly two categories (e.g. a binary or dichotomous variable) and 2) a continuous target variable that has values in the range 0.0 to 1.0 representing probability values or proportions. Figure 2.18 shows a simple example of a logistic regression function in one dimension. The logistic regression curve has an S (sigmoidal) shape. It builds a linear model based on a transformed target variable.

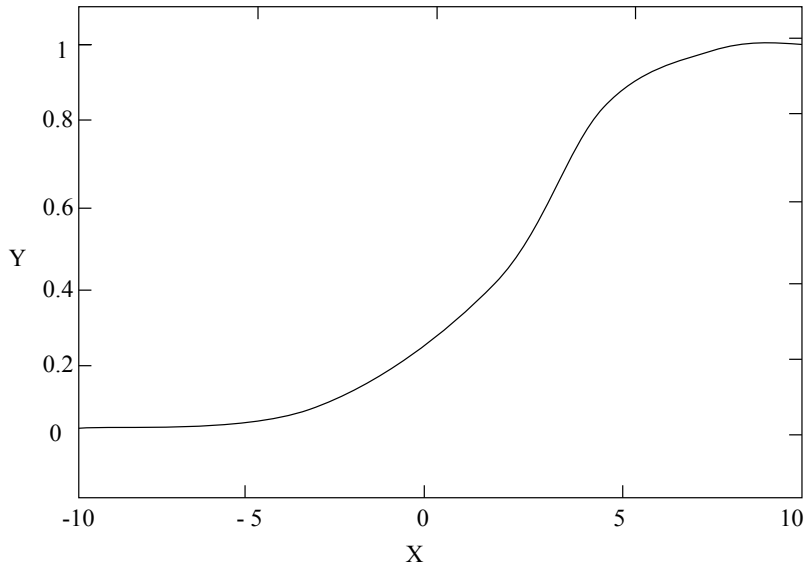


Figure 2.18. An example of logistic regression function.

The form of the logistic regression model formula is:

P = 1(1 + exp(-(B0 + B1X1 + B2X2 + …. + BkXk)))

where B0 is a constant and Bi are coefficients of the predictor variables (or dummy variables in the case of multi-category predictor variables). The computed value, P, is a probability in the range 0 to 1. The exp( ) function is *e* raised to power.

Logistic regressions, indicated as multinomial logistic regressions (MLRs) in the Weka software, are the class for building and using a multinomial logistic regressions model with a ridge estimator. For building a model for the prediction of community composition from environmental data a two-step approach is often used: in the first step, sites are clustered into groups on the basis of the biological data alone. This is generally a so-called hard-clustering in which every site is assigned to one organism group only. In the second step, these groups are related to the environmental data. Two important methods to establish this organism group-environmental relationship are discriminant analysis and multinomial logistic regressions. For the logistic regressions model, the whole range of methods and techniques for generalized linear models is available, such as selection of environmental variables. Another advantage of logistic regressions is that they directly model the probability of occurrence of each organism group as a function of the environmental data, while in discriminant analysis,

this is a by-product of the analysis. Logistic regressions are therefore the preferred method for relating organism groups to environmental data. The results can be used to predict the occurrence of organism groups from environmental variables.

However, the fitting of all possible regression models is very computing-intensive, especially for multinomial logistic models. A practical approach is to perform an iterative model selection. Firstly, the variables are subdivided into a few groups and model selection is performed. The best predictors from each group are then combined in a new model selection step, which yields a few best variables. With these few variables fixed in the model, the remaining variables are again subdivided into a few groups and the next iteration starts. This eventually results in a number of best candidate models. The predictive power of these candidate models can be assessed by means of leave-one-out and also cross-validation. In the first leave-one-out step, the first observation is temporarily deleted from the data and the model is fitted to the remaining observations. This model is then used to calculate a leave-one-out predication for the first observation. In the same way, leave-one-out predictions are obtained for all observations, by subsequently removing them from the data. The mean leave-one-out probability of predicting the correct organism group can then be used as a criterion for choosing among the candidate models. Further information about logistic regression can be found in McCullagh and Nelder (1989) and in Hosmer and Lemeshow (1989).

2.4.4.4.2 Specific parameter settings in this study

In recent years, logistic regressions have been widely used in different sciences e.g. ecological and medical (Sadat-Hashemi et al., 2005). In contrast to the other modelling techniques applied in this study, the whole default settings were used for logistic regressions (weka.classifiers.functions.Logistic) because this model requires less parameters for tuning (Witten and Frank, 2000).

The following settings were developed for the LRs (Witten and Frank, 2000):
-‘debug’ (output debug information to the console): false;
-‘maximum iterations’ (maximum number of iterations to perform); -1
-‘ridge’ (set the ridge value in the log-likelihood):1.0E-8.

2.4.4.5 Genetic algorithms (GAs)

2.4.4.5.1 General aspects of genetic algorithms

Genetic algorithms are a part of evolutionary computing, which is a rapidly growing area of artificial intelligence. Holland (1975) first explored GAs on chromosomes. The specific features of genetic algorithms make them novels for ecological modelling. They are general purpose search algorithms inspired by Charles Darwin's principle of "survival of the fittest" to solve complex optimisation problems (Holland, 1975; Goldberg, 1989; Vose, 1999). The basic idea is to maintain a population of chromosomes, which represent candidate solutions to the concrete problem being solved. A chromosome is represented by a string of genes. A possible value of a gene is called an allele. The chromosomes evolve through a process of competition and controlled variation. Genetic algorithms begin with a population of randomly generated chromosomes and advances towards better chromosomes by applying genetic operators that are based on genetic processes occurring in nature. The population undergoes evolution in the form of natural selection. During successive iterations, called generations, chromosomes in the population are rated for their adaptation or associated fitness. On the basis of these evaluations, a new population of chromosomes is formed using a selection method and genetic operators such as crossover and mutation. Figure 2.19 shows the simple structure of genetic algorithms. Chromosomes for the new population are selected with a probability proportional to their fitness. Mutation occurs only occasionally with a very low probability and means that the value in a string position may be changed. Crossover means that pairs of chromosomes are recombined by swapping parts of them from a randomly selected point in order to create two new chromosomes.

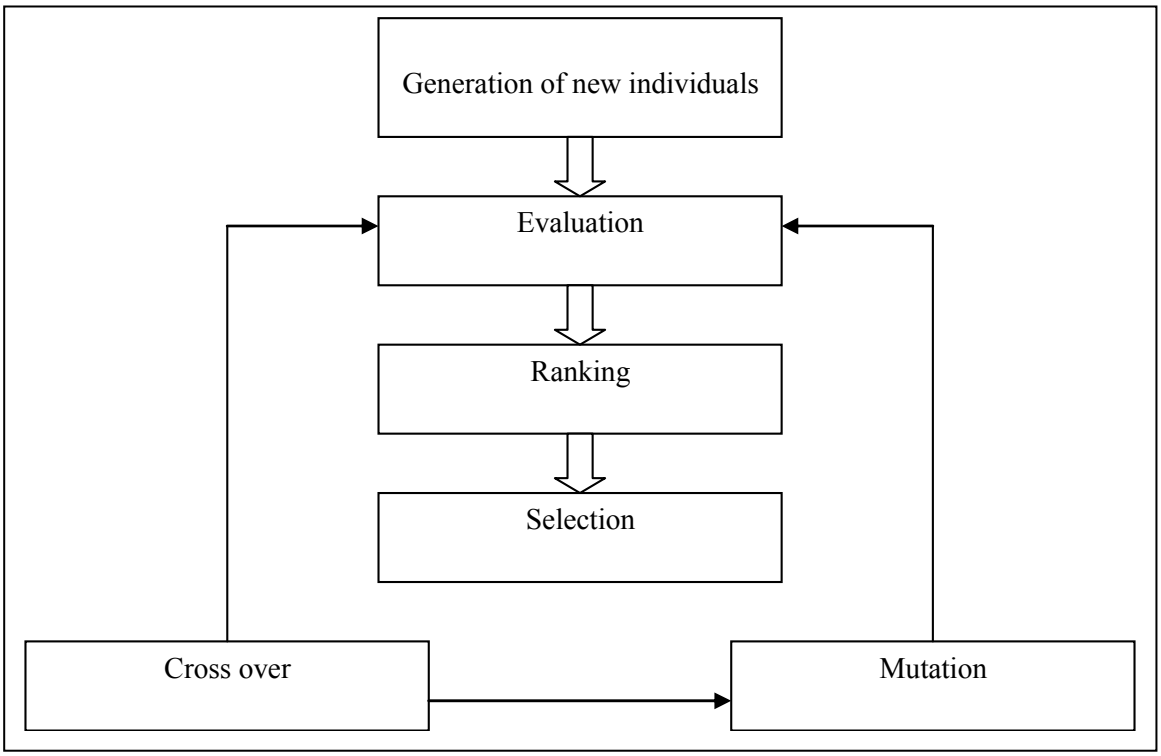


Figure 2.19. The structure of simple genetic algorithms (D'heygere et al., 2003).

2.4.4.5.2 Specific parameter settings for this study

In this study, a Goldberg genetic algorithm (Goldberg, 1989) was implemented in the Weka toolbox (weka.attributeSelection.GeneticSearch). A chromosome with 20 input variables was introduced to machine learning (each variable was equal with one gene). The initial population consisted of 20 chromosomes that were evolved through a maximum of 20 generations. Crossover was set at a probability of 60 %. Probability of mutation was set at 3.3 %.

The alphabet of encoding of a gene was binary digits and marked as pike absence (0) and presence (1). The attribute evaluator was based on *wrapper subset evaluator function* (weka.attributeSelection.WrapperSubsetEval) with full training set (150 instances). Here, a 3-folds cross-validation was considered (the default value is 5). Cross validation is used to estimate the accuracy of the learning scheme for a set of attributes. The Wrapper Subset Evaluator function evaluates attribute sets by using a learning scheme. In this wrapper

method, the variable selection algorithm functions as a wrapper for 4 modelling techniques (classification trees, artificial neural networks, support-vector machines and logistic regressions). After the new variables were selected by genetic algorithms, the model performance criteria were compared with and without genetic algorithms. The following settings were developed for the genetic algorithms (Witten and Frank, 2000):

- 'crossover probability' (set the probability of crossover. This is the probability that two population members will exchange genetic material): 0.6;
- 'max generations' (set the number of generations to evaluate): 20;
- 'mutation probability' (set the probability of mutation occurring): 0.033;
- 'population size' (set the population size. This is the number of individuals (attribute sets) in the population)): 20;
- 'report frequency' (set how frequently reports are generated): 20;
- 'seed' (set the random seed): 1;
- 'start set' (set a start point for the search. The start set becomes one of the population members of the initial population).

For the wrapper subset evaluator function, the following settings were used (Witten and Frank, 2000):

-'folds' (number of *x* validation folds to use when estimating subset accuracy): 3 (the default value is 5);
-'seed' (seed to use for randomly generating *x* validation splits): 1;
-'threshold' (repeat *x* validation if standard deviation of mean exceeds this value): 0.01.

2.5 Application of habitat suitability models for decision support of river management

Until now, in Flanders predictive ecological habitat suitability models for pike that can be used in decision support of river management have not been developed. Biological monitoring for river quality assessment based on the Belgian Biotic index (BBI) (De Pauw and Vanhooren, 1983) is accomplished by the Flemish Environment Agency (VMM). However the information which is obtained by monitoring is rarely used for further impact analysis, evaluation of monitoring networks or river restorations and simulation of river restoration

options. Undoubtedly, this can be a weakness concerning river management and hence paying no attention to a lot of information. An integrated ecological river management can be very valuable in the future and as such will be an environmental policy goal, as set by the WFD, and implemented in the Decree of Integrated Water Policy in Flanders. With the aim of an integrated water management in Flanders, the challenge will be to extend the ability of the existing monitoring framework though the implementation of predictive habitat modelling. Applications of ecological modelling can as such be considered as an added value for decision support of river management aside the existing bioassessment tools.

Particular challenges exist when predictive ecological modelling is directly linked to environmental management. The main challenges that river managers are facing, are to define ecosystem needs that are clear enough to guide policy formulation and management actions that attempt to balance the competing demands and visions. One of the challenges is the transformation of ecologically based models to make them suitable as decision support tools for river managers. Due to the decision support objectives of these predictive models, different dimensions to the objectives and challenges should be accounted for (Elith et al., 2002; Holling and Allen, 2002; Regan et al., 2002; Poff et al., 2003; Adriaenssens, 2004).

In summary, the developed habitat suitability models which serve as a decision support should:

- be clear and easy to use;
- be close to human reasoning;
- reach a compromise between policy relevance and predictive precision;
- be practical in providing management information with minimum cost and effort;
- be derived from a sound theoretical concept in ecology;
- integrate and assess existing knowledge;
- have the possibility to integrate qualitative data;
- have the possibility to integrate judgement from scientific experts;
- predict policy-relevant ecosystem attributes;
- be as precise as possible by separating different types of human impact on a given ecosystem;
- be as general as possible regarding its geographic location.

The combination of these objectives and challenges is important but habitat suitability models like classification trees, artificial neural networks etc. have the capacity to facilitate predictive modelling based on knowledge and judgement. Most of the recent predictive modelling techniques rely mostly on monitoring data and a wide variety of data-based statistical and machine-learning methods that have been introduced during the last decades in combination with geographical information systems (Guisan and Zimmerman, 2000; Austin, 2002).

As a conclusion, application of habitat suitability models using ecological modelling methods can be very helpful for decision support in river management. To meet this, a simulation exercise was done in the Dender river based on the applied models (see Chapter 9). Here, the main focus was paid to classification trees because of easy interpretation of the results.

Chapter 3

Database set-up and exploration

3.1 Introduction

This chapter is devoted to the exploratory analysis and preparation of the environmental input and biological output datasets for pike in the Flemish river basins. A first and important step before model development and application is to get acquainted with the range of inputs and outputs, what determines also the practical application range of the models. This can be of major importance in relation to river restoration management. Additionally, the mutual correlation between input variables and between input and output variables is useful to recognise 'noise' variables. A visual relation analysis between input and output variables is determined to recognize outliers, data clusters and missing or scarce variable combinations in certain ranges. To do so, these data analyses can give advice regarding new or additional data that require to be collected and how this should be done to be able to develop more related models and to gather important information for river management afterwards.

This chapter consists of three main parts. The first part describes the bandwidth and distribution of input and output variables. The second one deals with the correlation analysis between input and output variables. The third part presents the visualisation of the input and output variables. In the end, a discussion and conclusion is given.

3.2 Results

3.2.1 Bandwidth and distribution of input and output variables

In this part of the data analysis, calculation of some general statistical parameters was carried out e.g. the calculation of the minima, maxima, averages and standard deviations of the given variables (Table 3.1). To achieve this, the results and analyses were incorporated in the visualisation graphs provided in Figures 3.2 and 3.3. Analysis of the output variables, however, seemed to be rather irrelevant because the outputs are only expressed by 0 and 1 (as pike absence and presence respectively). As stated already in Chapter 2 (see Section 2.3.1), if these calculations are performed, one can directly see whether or not high standard deviations are a result of a wide span of the data or are more related to some outliers. It is also very interesting to use the median and compare it with the average. Since the visualisation graphs will be presented in the coming paragraphs, the median was not included in Table 3.1. As

indicated here, analysis of bandwidth and distribution of input variables was taken into account as an initial and fundamental step before model development and application.

Based on the analysis of 6 river basins in the Flanders dataset (Table 3.1 and Figure 3.1), no extreme outliers (indicated with very high values and standard deviations) can be observed in the datasets during the year 1991-2002. Some mild outliers can be seen in the variables slope, distance from source, electric conductivity, chloride, ammonium and BOD. In all the given variables these outliers are skewed to the upper part of the boxplots. For the values of physical-chemical variables illustrated in the boxplots, the measurements were most likely accurate so that the sites can be contaminated by nutrients.

It is obvious that the broad range of the variables can lead to a relative compression of the majority of the measurements. To make the models applicable in the broadest span of cases as well as to make a tryout on data that are as natural as possible, these outliers were left in the datasets. In other words, no outliers and instances were removed from the datasets.

Table 3.1. Minima, maxima, averages and standard deviations of the numeric input variables that were used in the river basins database (1991-2002).

Variable	Minimum	Maximum	Average	Standard deviation
Slope (%)	0.0	6.3	1.1	1.2
Distance from the source (km)	1.2	83.2	21.1	19.5
Width (m)	0.5	32.0	8.5	7.7
Depth (m)	0.1	2.5	0.8	0.6
Flow velocity (m/s)	0.0	1.6	0.6	0.4
Locality-length (m)	40	500	157	114
Locality-width (m)	0.5	10.5	3.7	1.6
Fished-area (m)	48	2000	559	378
pH	6.4	8.2	7.5	0.4
Dissolved oxygen (mg/l)	1.3	10.8	8.1	1.4
Conductivity (µS/cm)	158	2173	666	344
Nitrate (mg NO_3^-/l)	0.2	12.1	3.8	2.9
Nitrite (mg NO_2^-/l)	0.0	0.3	0.1	0.1
Ammonium (mg NH_4^-/l)	0.0	15.6	1.7	1.9
Ortho-phosphate (mg P/l)	0.0	13.9	0.6	1.8
Total phosphorus (mg P/l)	0.0	14.4	0.8	1.7
Suspended solid (mg/l)	6.9	113.5	36.9	22.1
Chloride (mg/l)	8.4	401.4	67.0	61.8
COD (mg O_2/l)	6.8	79.3	32.2	12.3
BOD (mg O_2/l)	0.0	24.4	3.6	3.5
BBI	0.0	9.0	5.6	1.7

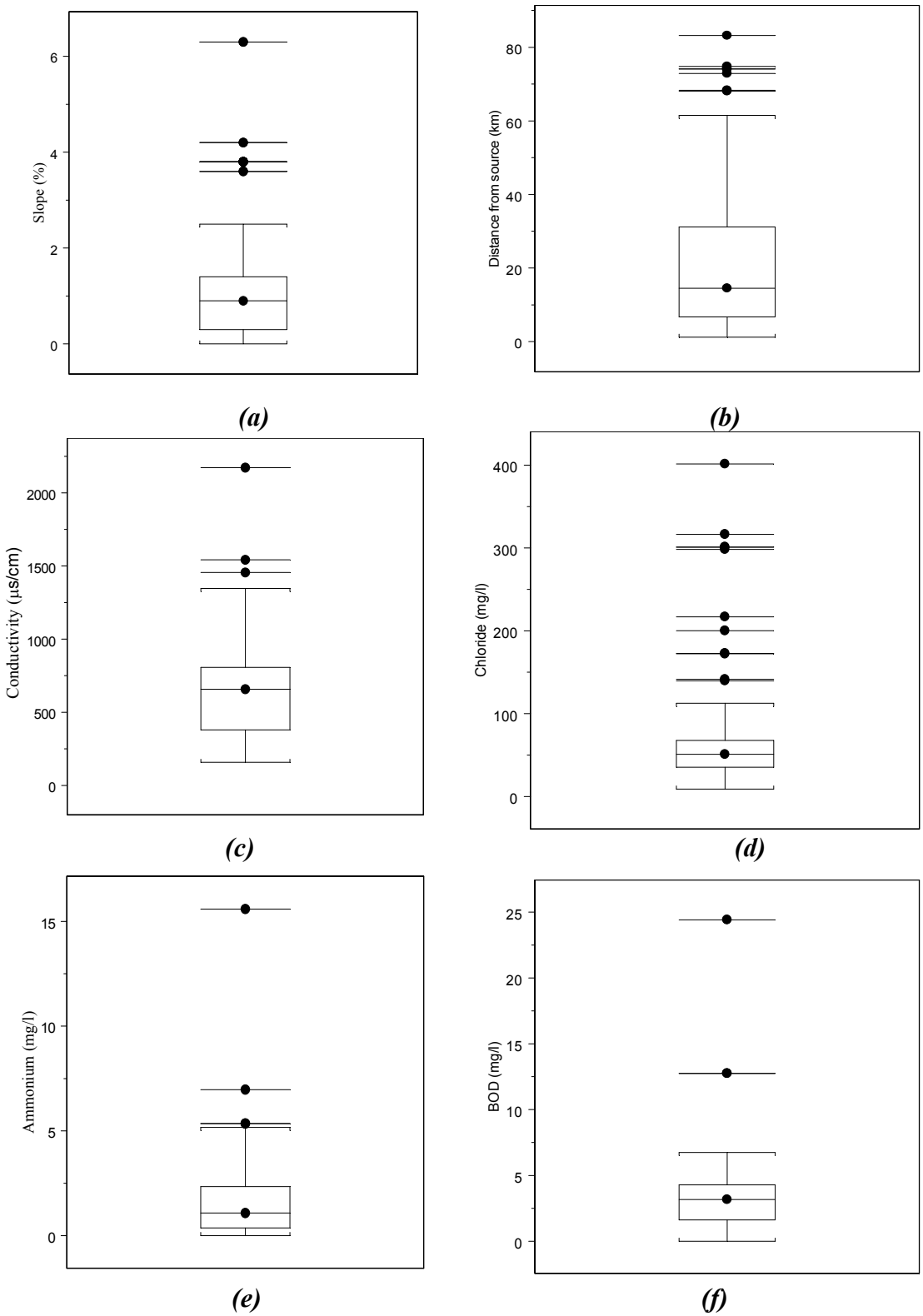


Figure 3.1. Boxplots of some structural-habitat and physical chemical variables applied for pike in the river basins in Flanders.

3.2.2 Correlation analysis between input and output variables

The second stage of the data analysis consisted of determining how related some variables might be, based on their correlation coefficients (*r)*. Here, also only the *r* between input variables was considered because only one output variable was registered in the dataset (in terms of pike presence/absence). As stated already, any *r* with an absolute value of 0.20 or higher indicates a possible noise source to data driven models and it is recommended to remove one of these correlated variables (Walczak and Cerpa, 1999). Table 3.2 shows the correlation coefficients (*r)* of 21 variables calculated for 110 sampling sites in Flanders. Correlation coefficients with an absolute value of at least 0.20 and lower than 0.50 are marked in yellow, higher values are marked in orange. Among the entire variables, high correlation values ($r > 0.50$) are observed between some river characteristics: e.g. fished area and locality-length ($r = 0.82$), width and depth ($r = 0.72$), fished area and width ($r = 0.62$), width and locality-length ($r = 0.51$), fished area and depth ($r = 0.62$), depth and locality-length ($r = 0.53$). These variables are positively correlated, while slope and locality-width ($r = -0.51$) are negatively correlated.

Among the physical-chemical variables, a reasonable correlation was identified between the electric conductivity and some pollutant variables such as chloride ($r = 0.84$), nitrite ($r = 0.68$), nitrate ($r = 0.54$) and ammonium ($r = 0.52$). Quite a high *r* can also be seen between other pollution variables: e.g. total phosphorus and ortho-phosphate ($r = 0.55$), total phosphorus and BOD ($r = 0.81$), COD and BOD ($r = 0.76$), the COD and ammonium ($r = 0.59$), pH and nitrate ($r = 0.75$), pH and nitrite ($r = 0.53$), and ammonium and dissolved oxygen ($r = -0.59$).

A logic correlation was also observed between the Belgian Biotic Index (BBI) and some chemical variables e.g. positive correlation between the BBI and dissolved oxygen ($r = 0.52$). In contrast, negative correlations are seen between BBI and BOD ($r = -0.59$), COD ($r = -0.51$) and ammonium ($r = -0.65$). The BBI method uses macroinvertebrates as indicators for the level of pollution (De Pauw and Vannevel, 1993). This methodology points out that increasing pollution can lead to a loss of diversity and a progressive elimination of certain pollution-sensitive groups. As indicated in Table 3.1, the BBI ranges from 0 to 10 (here, only minima, averages and maxima of this variable are presented). The BBI can be interpreted as follows: 0

- 2 = very heavily polluted, 3 - 4 = heavily polluted, 5 - 6 = moderately polluted, 7 - 8 = slightly polluted and 9 - 10 = unpolluted.

Although some variables in relation to others were characterised by relatively high correlation coefficients, all variables were kept in the dataset for both ecological and practical reasons (Table 3.2). The reason was to find out whether the data driven model development methods would succeed in removing these highly redundant variables or not and how these variables would be ranked by these methods.

Table 3.2. Pearson correlations (r) for the 21 variables in the 6 river basins in the Flanders. Correlation coefficients with an absolute value of at least 0.20 and lower than 0.50 are marked in yellow, higher values are marked in orange.

	Slope	**Distance from source**	**Width**	**Depth**	**Flow velocity**	**pH**	**Dissolved oxygen**	**Conductivity**	**Locality-length**	**Locality-width**	**Fished-area**	**Nitrate**	**Nitrite**	**Ammonium**	**Ortho-phosphate**	**Total phosphorus**	**Suspended solids**	**Chloride**	**COD**	**BOD**	**BBI**
Slope	1.00	-0.25	-0.18	-0.35	0.15	-0.13	0.09	-0.27	-0.30	-0.51	-0.38	0.06	-0.19	-0.16	-0.18	-0.16	-0.11	-0.31	-0.25	0.09	0.10
Distance from source		1.00	0.49	0.39	0.45	0.02	-0.48	0.37	0.44	0.11	0.40	0.04	0.51	0.39	-0.12	0.31	0.36	0.59	0.22	0.18	-0.19
Width			1.00	0.72	0.27	0.43	-0.32	0.65	0.51	0.18	0.62	0.24	0.49	0.46	0.19	0.51	0.06	0.47	0.29	0.08	-0.29
Depth				1.00	0.26	0.38	-0.19	0.53	0.53	0.11	0.62	0.37	0.57	0.33	0.28	0.20	0.22	0.36	0.25	0.05	-0.28
Flow velocity					1.00	0.16	0.04	0.42	0.35	-0.04	0.31	0.35	0.38	0.34	0.35	0.31	0.21	0.36	0.00	0.17	0.01
pH						1.00	0.06	0.65	0.29	0.16	0.41	0.75	0.53	-0.06	0.20	0.18	0.04	0.26	-0.09	-0.17	0.05
Dissolved oxygen							1.00	-0.41	-0.18	0.05	-0.15	0.17	-0.28	-0.59	-0.07	-0.47	-0.14	-0.49	-0.30	-0.29	0.52
Conductivity								1.00	0.47	-0.08	0.42	0.54	0.68	0.52	0.44	0.62	0.08	0.84	0.23	0.12	-0.4
Locality-length									1.00	-0.15	0.82	0.15	0.37	0.20	0.05	0.25	0.20	0.47	0.15	0.04	-0.14
Locality-width										1.00	0.35	0.10	0.05	-0.12	-0.06	-0.11	-0.15	-0.14	-0.23	-0.39	0.16
Fished-area											1.00	0.24	0.38	0.12	0.08	0.20	0.05	0.30	0.00	-0.14	-0.11
Nitrate												1.00	0.66	-0.07	0.30	0.14	0.04	0.20	-0.09	-0.18	0.01
Nitrite													1.00	0.37	0.25	0.40	0.35	0.45	0.24	0.21	-0.38
Ammonium														1.00	0.39	0.82	0.02	0.54	0.59	0.46	-0.65
Ortho-phosphate															1.00	0.55	-0.07	0.28	0.31	0.54	-0.44
Total phosphorus																1.00	0.07	0.55	0.67	0.81	-0.67
Suspended solids																	1.00	0.13	0.31	0.28	-0.21
Chloride																		1.00	0.36	0.15	-0.35
COD																			1.00	0.76	-0.51
BOD																				1.00	-0.59
BBI																					1.00

3.2.3 Visualisation of the input and output variables

The third and last method applied in data analysis of pike habitat suitability concerns the visualisation graphs. As described already, two types of graphs are presented in this part: 1) the histograms in which the relevant datasets and variables of the pike's habitat requirements are visualised by means of the Weka tool (Witten and Frank, 2000), and 2) the scatter plots in which the relationship between two variables are plotted for visualising the pike presence/absence. Figure 3.2 and 3.3 show two types of these graphs. The application of such graphs was very interesting to compare the model outcomes.

According to Figure 3.2, for many structural habitat and physical-chemical variables (e.g. distance from source, wetted-width and most pollution variables such as ammonium, ortho-phosphate, total-phosphorus, chloride and suspended solid) a logic relation can be observed between the given variables and the classes of pike presence/absence. Pike was mainly found when the values of these variables were low. In this regard, a good example can be given for the variable ammonium. When the ammonium concentration ranged between 0 and 1.3 mg/l (the first column on the left side of Figure 3.2) about 78 instances were grouped in the distribution graph. Among 78 instances, pike was present in 39 cases (red colour) and was also absent in 39 instances (blue colour), while in other columns a clear reduction of the presence class was detected when the concentration of the ammonium increased.

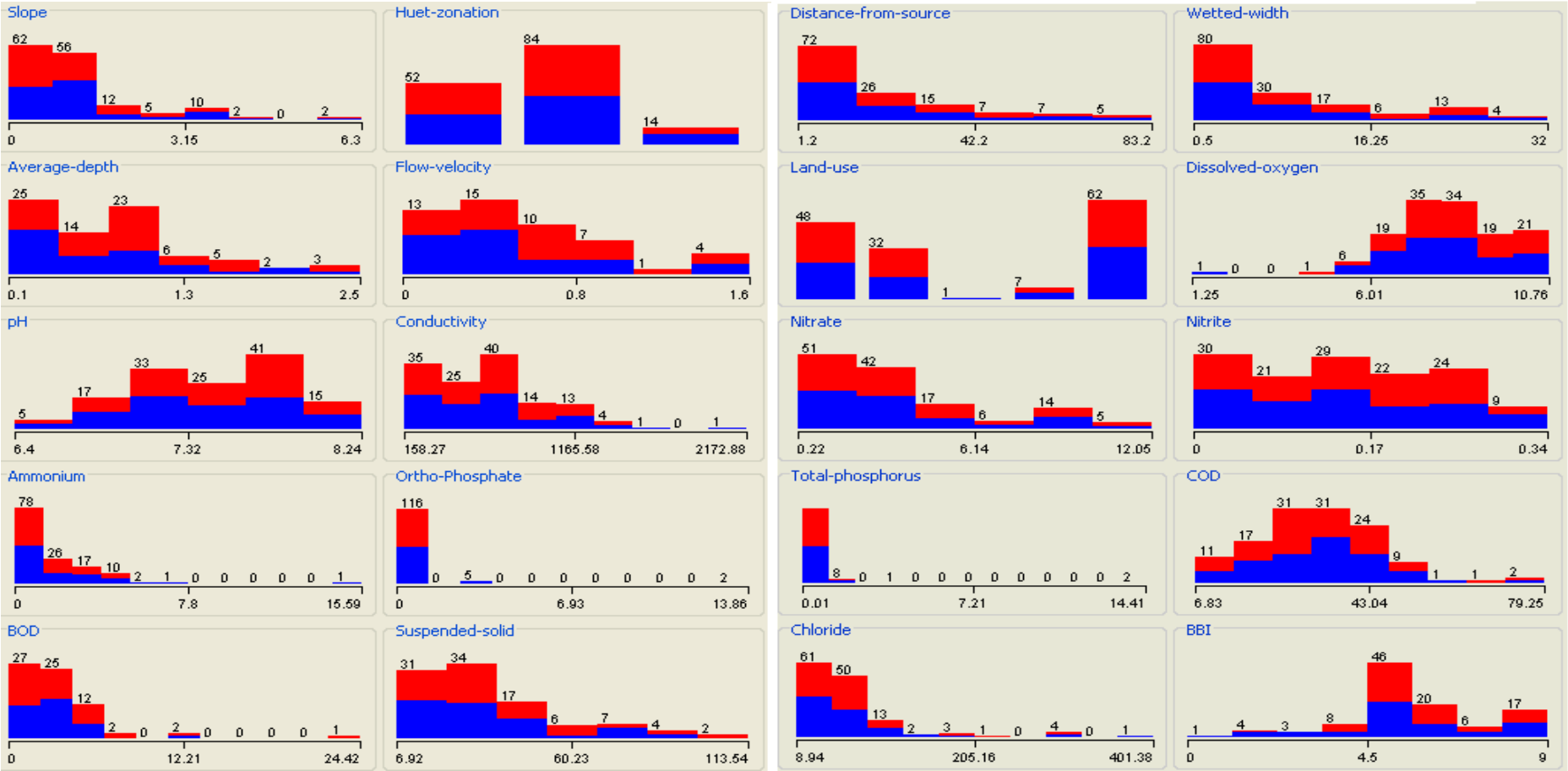


Figure 3.2. Histograms for pike presence/absence in 6 river basins in Flanders (in total 150 instances) in relation to the environmental variables (pike absent in 75 instances (blue), and present in 75 instances (red)).

Figure 3.3 shows the results of the second type of visualisation (scatter plots). The scatter plots used in this approach were allocated for visualising the distribution of input and output variables (the classes of pike presence/absence). Based on these graphs, the effect of input variables on the output variable could be observed (in these graphs the presence is indicated as filled circles and absence is indicated as open circles). Similar conclusions could be drawn as in section 3.2.1. For instance, some outliers can be detected for the variables conductivity etc.

In this part, only a few variables were plotted. As can be inferred from Figure 3.3, for most variables, a rational relation was perceived so that high concentration of most pollutants such as conductivity and chloride resulted in a decrease in pike populations (Figure 3.3a and b). Almost the same trend was observed regarding most structural habitat variables e.g. depth and wetted width (Figure 3.3c and d) so that a reduction of pike (in terms of presence/absence class) occurred when the depth and wetted width increased. Here most pike populations showed the tendency of being present in depths less than 2 m and also most of them were present in widths less than 15 m. Thus, by applying this method, those variables with high correlation coefficients or instances with extreme outliers could be left out. By removing such instances, the distributions over all classes and values of the input variables would become more even. Nevertheless, the variables and datasets were kept as natural as possible. This was mainly carried out to find out whether the models could also cope with this bottleneck or not.

Besides, for getting a better visualisation, the geographical distribution of the data in the river basins of Flanders (represented as pike presence/absence) over the study areas was plotted by means of ArcView 3.2a, which is a product of the Environmental System Research Institute (ESRI). This method was already mentioned in Chapter 2 (see Figure 2.3) and therefore is not further discussed in this part.

To get a better insight into the temporal variation over different years, the variables were additionally plotted as illustrated in Figure 3.4 for the variables dissolved oxygen, conductivity and pH (plots of some appropriate variables are presented in the Appendix 13 for the datasets of 6 river basins). Here the annual average of the given variables was considered. By doing so, the possible temporal trends could be observed for each variable. For instance, the temporal variation of the dissolved oxygen fluctuates between 6.5 and 10.5 mg/l during the period 1991-2002.

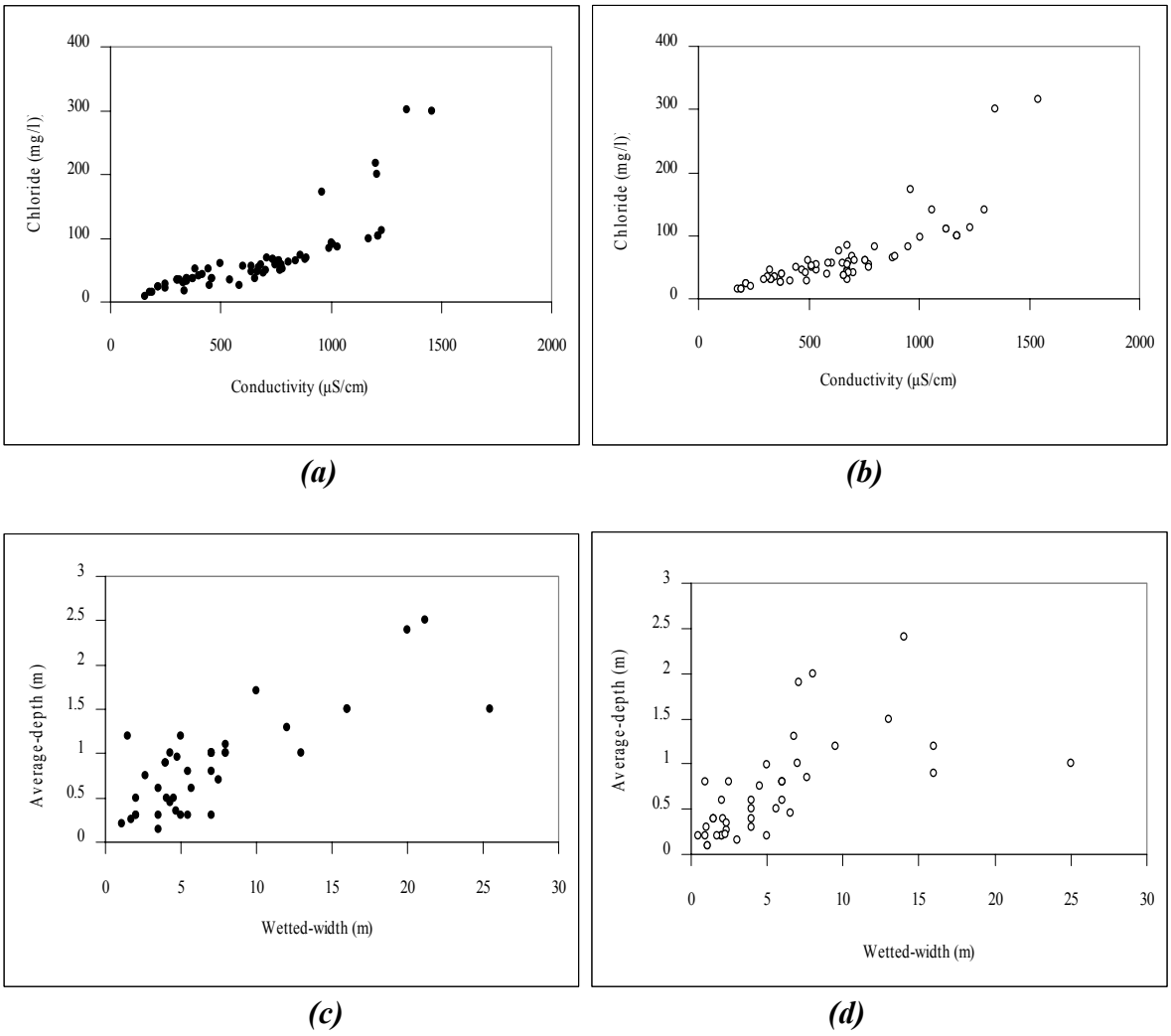


Figure 3.3. Scatter plots of some variables for pike (pike presence is indicated as filled circles and pike absence is indicated as open circles) in the Flemish river basins (in total 132 instances were available for the first two scatter plots and 84 instances for the last two scatter plots).

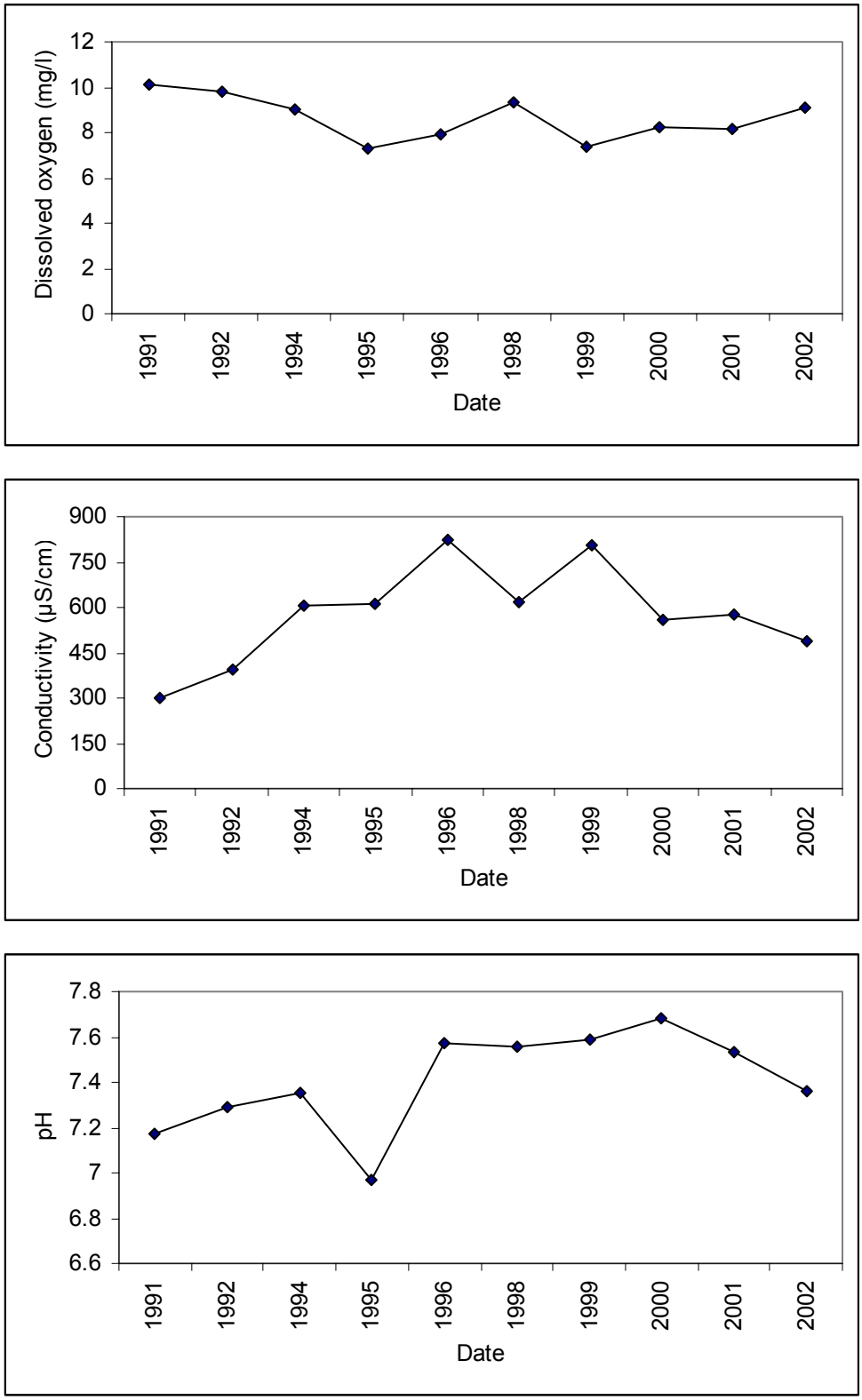


Figure 3.4. Temporal variation of the annual average of 3 variables (dissolved oxygen, conductivity and pH) during the period 1991-2002 in the 6 river basins in Flanders (in total 136, 132 and 137 instances were recorded for dissolved oxygen, conductivity, and pH respectively).

3.3 Discussion

The aim of this part was to analyse and prepare datasets for pike's habitat requirements in the 6 river basins in Flanders. In essence, this part was taken into account as the first starting point for developing and optimising the ecological modelling techniques. Gathering data is an important task regarding data analysis and variables. As stated by Witten and Frank (2000), missing values are endemic in real-world datasets. Therefore, for each measurement, there might be some missing values. A lot of missing values of important variables leads to the removal of some instances as encountered in the visualisation graphs presented in Figure 3.3 and Figure 3.4. Here, some instances had to be eliminated because of missing values. One way of handling missing values is to treat them as just another possible value of the attribute; this is appropriate if the fact that the attribute is missing, is significant in some way. Sometimes the attributes, whose values are missing, play no part in the decision, in which case these instances are as good as any other (Witten and Frank, 2000).

In the datasets gathered in the river basins, numerous electrofishing data as well as other important environmental variables were recorded for pike but due to many missing values in the given variables, only a set of 150 observations and more than 20 variables could be provided for the development of the habitat suitability model for pike. Several critical variables were missing, in particular those variables closely associated with physical habitat e.g. vegetation cover and habitat type. Water temperature was also eliminated for the same reason. In that case, undoubtedly several constraints such as financial and time dilemmas could play an essential role regarding the deficit of the datasets. Therefore, all of these problems could certainly influence pike prediction.

In all ecological monitoring, it is very important to be acquainted with the field and sampling approach. By doing so, one is able to measure the most relevant variables or ignore measuring the superfluous variables. Besides, it can also help to recognize what variables play a significant role in the ecosystems and for river managers.

In summary, the number of variables and electrofishing records introduced to a model seemed to be enough. In spite of this, however, some crucial variables were not available for the model development and also missing a lot of values in the variables could certainly prevent

the models to make a reliable prediction. Leaving some variables (or instances) out and or leaving other variables in can be important for practical and theoretical purposes. In this research, however, some variables applied in the developed model didn't seem to be relevant for the pike habitat requirements.

As illustrated in Table 3.2, some river characteristics such as pollution-related variables showed relatively high correlation coefficients. However, all these variables were used to find out how the different model development techniques could cope with these problems, and what the added value can be of variable selection by genetic algorithms in this context. In that way, it was possible to analyse whether genetic algorithms remove the redundant variables or not and how these variables would be ranked by these methods for the four applied modelling techniques.

3.4 Conclusion

This part of the study aimed at analysing and preparing the datasets for developing and optimising the ecological modelling techniques for the prediction of the pike's habitat suitability (based on 150 presence/absence data) in 110 sampling sites in Flanders.

It seemed that to make an ultimate decision concerning the number of attribute/instances for developing modelling techniques, many factors are involved such as the data collection, the knowledge of how to measure different aspects of the ecosystem, financial and time constraints and also measurement problems. Several theoretical reasons came into sight to eliminate some variables and instances, but also some practical reasons to leave them in to investigate how data driven modelling techniques handle them. In this thesis, the latter option was chosen.

The results obtained by these data analyses can be used to interpret the results of the modelling techniques in the coming chapters. Besides, they could be very helpful to detect hot spots for river restoration and therefore to collect new or additional data and how this has to be carried out to develop better models and to gather helpful information for river management afterwards.

Chapter 4

Development of habitat suitability models for pike using classification trees in combination with genetic algorithms

4.1 Introduction

Habitat suitability models as applied in river management are predictive ecological models that are capable of establishing a scientifically sound link between biotic and abiotic river components. These models should be meaningful to a broad range of people engaged in the decision making process and to that end, a clear presentation of the model structure and inference process is needed. Currently, the assessment of running water bodies relies on a set of physical, chemical and biological features. The latter are based on a set of characteristics of the biological communities inhabiting the watercourse such as richness, abundance, presence/absence etc. In this context, pike plays a major role in structuring freshwater communities (Spens et al., 2007) and has been used in stocking programmes to improve water quality (biomanipulation) (Craig, 2008). Despite many recent studies, quantifying and modelling the role of pike as a top predator in many ecosystems have been problematic, perhaps due to the difficulties of sampling natural populations (Craig, 2008).

As briefly reviewed before, classification trees provide qualitative discrete outputs of a system under certain conditions represented by parameters (Witten and Frank, 2000; Solomatine and Dulal, 2003; Chen and Mynett, 2004). It is only for a few years that classification trees became recognised as simple but powerful tools in the analysis of ecological data (Stankovski et al., 1998; Debeljak et al., 1999; Blockeel et al., 1999a; Blockeel et al., 1999b, De'ath and Fabricius, 2000).

The aim of this chapter was to develop predictive habitat suitability models based on classification trees in combination with genetic algorithms for pike in 6 river basins in Flanders. First, an assessment based on model evaluation is presented, followed by an overview of the selected variables and the ranking of their importance. The trees are presented in the Appendices 1-12. For each of the three subsets, different pruning confidence factors were examined. Here, a genetic algorithm (Goldberg, 1989) was applied in combination with classification trees (J48 algorithm) for selecting the relevant variables. Finally, an ecological interpretation and discussion for practical use of the induced classification trees and genetic algorithms is given.

4.2 Results

4.2.1 Assessment based on model performance criteria

In this part of the results, the independent variables consisted of 20 input variables, while the dependent variable was considered as the presence or absence of pike. In Figure 4.1a, the best performing classification tree (according to the CCI and Kappa statistics as marked in green in Table 4.2) is presented for pike. The classification tree induced for the prediction of pike gave the best predictive performance in the subset 1 (PCF = 0.5), with the highest percentage of correctly classified instances (78 %) and Kappa statistics (0.56). As can be inferred from Figure 4.1a and Table 4.3, the tree obtained for pike with the highest PCF used only 9 out of 20 input variables. Here, the complex trees were generated because increasing PCF would result in a greater complexity of trees and then leading to a more difficult interpretation. This phenomenon has already been described by Witten and Frank (2000). In contrast, the simple classification trees were presented in Figure 4.1b when applying the intermediary pruning confidence factors (PCFs = 0.25 and 0.1). At PCF levels 0.25 and 0.1, the size of trees were equal (10) and three variables (land-use, chloride and wetted-width) were used by the classification trees. At these PCF levels, a CCI 60 % and Kappa 0.20 were resulted (Table 4.2), while at PCF 0.01, no trees were detected. Here, also the lowest model performance was detected (CCI = 50 % and Kappa = 0.00).

```
Land-use = forest
|   Chloride <= 57.18
|   |   BOD <= 2.17
|   |   |   BOD <= 1.62
|   |   |   |   Wetted-width <= 3.5: PIKE PRESENT (5.01/2.03)
|   |   |   |   Wetted-width > 3.5: PIKE ABSENT (4.84/1.03)
|   |   |   BOD > 1.62: PIKE PRESENT (4.93/1.42)
|   |   BOD > 2.17: PIKE ABSENT (12.32/3.77)
|   Chloride > 57.18: PIKE PRESENT (5.89/0.18)
Land-use = pasture: PIKE PRESENT (18.0/6.0)
Land-use = industrial: PIKE ABSENT (0.0)
Land-use = urban
|   Wetted-width <= 12: PIKE PRESENT (4.0/1.0)
|   Wetted-width > 12: PIKE ABSENT (2.0)
Land-use = arable-land
|   Huet-zonation = barbel
|   |   BOD <= 2.83
|   |   |   COD <= 25.01: PIKE PRESENT (3.27/0.27)
|   |   |   COD > 25.01: PIKE ABSENT (2.73/1.0)
|   |   BOD > 2.83: PIKE ABSENT (6.0/1.0)
|   Huet-zonation = bream
|   |   Flow-velocity <= 0.71
|   |   |   Average-depth <= 0.9
|   |   |   |   Belgian biotic Index (BBI) <= 6: PIKE ABSENT (4.64/1.07)
|   |   |   |   Belgian biotic Index (BBI) > 6: PIKE PRESENT (2.95/0.99)
|   |   |   Average-depth > 0.9: PIKE PRESENT (9.56/3.44)
|   |   Flow-velocity > 0.71: PIKE ABSENT (12.86/3.86)
|   Huet-zonation = upstream: PIKE ABSENT (1.0)
```

(a) (PCF = 0.5)

```
Land-use = forest
|   Chloride <= 57.18: PIKE ABSENT (27.11/11.29)
|   Chloride > 57.18: PIKE PRESENT (5.89/0.18)
Land-use = pasture: PIKE PRESENT (18.0/6.0)
Land-use = industrial: PIKE ABSENT (0.0)
Land-use = urban
|   Wetted-width <= 12: PIKE PRESENT (4.0/1.0)
|   Wetted-width > 12: PIKE ABSENT (2.0)
Land-use = arable-land: PIKE ABSENT (43.0/18.0)
```

(b) (PCF = 0.25 and 0.1)

Figure 4.1. Classification trees for pike in the river basins in Flanders (Subset 1, (a) PCF = 0.5, (b) PCF = 0.25 and 0.1). (Values between brackets indicate instances in which rules are true/false).

For pike prediction, 17 leaves and 12 nodes (tree size of 29 as marked in green in Table 4.1) was the best outcome among the three subsets (CCI = 78 % and Kappa = 0.56). These high model performances indicated that reliable models were learned in the subset 1 (CCI > 70 % and Kappa > 0.40). For the other subsets, the results obtained based on model performances did not meet the threshold values. In other words, no reliable model was obtained for the other two subsets.

Table 4.1 and 4.2 show how pruning can influence tree size and model evaluation, respectively. Pruning has a tremendous effect on the size and related complexity of the trees. The tree size dropped from 28 to 2 (SS Average row marked in yellow in Table 4.1). Nevertheless, in subsets 2 and 3, the CCI and Kappa seemed to stay rather constant under different pruning levels. An exception can be seen at PCF 0.01 in the subset 1 (indicated in orange in Tables 4.1 and 4.2), where the Kappa dropped to zero and no tree was produced. In that case, the tree size was equal to one, meaning that a 'strange rule' was induced by the classification trees. This means that a CCI 50 % could be obtained without using any information from the environmental variables.

Table 4.1. Tree size of the induced classification trees (for different pruning confidence factors (PCFs)) for pike in 6 river basins in Flanders.

Tree size	**PCF=0.5**	**PCF=0.25**	**PCF=0.1**	**PCF=0.01**	**PCF Average**	**PCF Stdev**
Subset 1	29	10	10	1	13	12
Subset 2	30	3	3	3	10	14
Subset 3	26	17	17	3	16	10
SS Average	28	10	10	2	13	11
SS Stdev	2	7	7	1		

By looking at the three subsets, the pruning differs enormously from one subset to another, but when assessing the averages (PCF Average columns), one can conclude that the average tree size and their average model evaluation were more or less stable. Based on the average CCI and Kappa (average values marked in grey), the classification trees didn't yield any reliable prediction for pike. For the four PCFs, the highest pruning was obtained at PCF 0.5, leading to a CCI 60 % and Kappa 0.20.

Table 4.2. Performance of the induced classification trees (for different pruning confidence factors (PCFs)) for pike in 6 river basins in Flanders.

CCI	**PCF=0.5**	**PCF=0.25**	**PCF=0.1**	**PCF=0.01**	**PCF Average**	**PCF Stdev**
Subset 1	78	60	60	50	62	12
Subset 2	52	54	54	54	53	1
Subset 3	50	52	52	50	51	1
SS Average	60	55	52	51	55	4
SS Stdev	13	3	4	2		
Kappa	**PCF=0.5**	**PCF=0.25**	**PCF=0.1**	**PCF=0.01**	**PCF Average**	**PCF Stdev**
Subset 1	0.56	0.20	0.20	0.00	0.48	0.23
Subset 2	0.04	0.08	0.08	0.08	0.07	0.02
Subset 3	0.00	0.04	0.04	0.00	0.02	0.02
SS Average	0.20	0.11	0.11	0.03	0.11	0.07
SS Stdev	0.26	0.07	0.07	0.04		

4.2.2 Ranking of major input variables for pike habitat requirements by classification trees

Ranking input variables has many advantages. By doing this, one can easily interpret the outcomes of different subsets in which variables are ranked based on their importance. The ranking of these variables is based on the generated trees in each subset. In the Appendices 1-12, these trees are presented with 3 subsets and at 4 pruning levels (0.5, 0.25, 0.1 and 0.01). Only the highest pruning confidence factor (PCF = 0.5) was considered for ranking of the input variables because when applying other PCFs, the major variables remained constant over different PCFs. This is obvious from Figure 4.1b, indicating that lowering the PCF only leads to the getting less input variables.

Table 4.3 reveals the ranking of the major input variables induced by the classification trees. The variables obtained at each splitting level are listed in each column for each subset separately. When a variable occurs at several levels, only the first level has to be considered. In the end, the results of the three subsets are incorporated to find out to what extent the stability of the result would be. As an example, the first variable visualised in Figure 4.1a (subset 1), was the type of land-use (forest, pasture, industrial, urban, and arable-land) and the second ones were chloride, wetted-width and the type of Huet-zonation (barbel, bream and

upstream). The variables from the third level were BOD and flow velocity. COD and average-depth were present in the fourth level and finally the BBI presenting in the fifth one.

Table 4.3. Major variables of the induced classification trees for pike in the river basins in Flanders.

Variables	**First**	**Second**	**Third**	**Fourth**	**Fifth**
Subset 1	Land-use	Chloride Wetted-width Huet-zonation	BOD Flow velocity	COD Average-depth	BBI
Subset 2	Land-use	BOD Flow velocity Average-depth	Huet-zonation BBI	Slope	Total-phosphorus
Subset 3	BBI	Huet-zonation	BOD Average-depth Wetted-width	Land-use	Distance from the source
Three times	-	-		-	-
Two times	Land-use	Huet-zonation	BOD	-	-
One time	BBI	Chloride Wetted-width BOD Flow velocity Average-depth	Flow velocity Huet-zonation BBI Average-depth Wetted-width	COD Average-depth Slope Land-use	BBI Total-phosphorus Distance from the source

Based on Table 4.3, most of the significant variables ranked for pike were associated with the type of land-use (forest, pasture, industrial, urban, and arable-land), followed by structural-habitat and chemical variables. In the first level, the land-use was repeated 2 times. In the second and third level, the Huet-zonation and BOD were also repeated two times. The results ranked over three subsets in the fourth and fifth level were instable. This means that no variables were repeated two or three times in these levels.

As a whole, for the pike's habitat, the type of land-use, BBI, Huet-zonation, chloride, wetted-width, BOD, flow velocity, average-depth were the main predictors. Distance from the source and total-phosphorus were ranked in the fifth level. Since pike populations are strongly dependent on physical-habitat variables, these variables were expected to be ranked by the induced classification trees. Nevertheless, from an ecological point of view some of the rules induced by the classification trees may not be so relevant for prediction of the pike's habitat. As can be concluded from Figure 4.1a (where PCF = 0.5 was applied), a complex rule was

induced for some variables in particular for BOD. Here, pike seemed to be absent when an extremely low BOD concentration is observed (BOD > 2.83 mg/l). This rule can be doubted from an ecological point of view (however, the implicit meaning could be that pike is not present in systems with very low concentrations organic material, e.g. very close to sources), but the selection of this extremely low value is more questionable, since the relevance of extremely low concentrations is low due to the limitations of the BOD-analysis method. In spite of this, only a few instances were classified as pike absence by the induced trees. The same counts for the variable COD where pike was found absent when COD concentrations were also relatively low (> 25.01 mg/l). Nevertheless, in this case also the number of instances assigned to the class 'pike absent' was low (meaning that the rule is not very important). In Figure 4.1b, where PCF = 0.25 and 0.1 were used as settings, these two variables were not selected. This is an indication of the relatively low importance of these variables. In these PCF levels, the rule for the chloride concentrations can be interpreted as relatively crucial because the number of instances assigned to the class of pike absence is high. This implies that pike requires chloride concentrations of higher than 57.18 mg/l, what can be considered as ecological relevant, since extreme low chloride concentrations are probably only typical for upstream water systems in which pike is usually not present. The effect of land-use on pike seemed to be quite logical (except for the type of industrial since no data were available for this type of activity). When the activities are more directed towards arable-land, pike was absent and in case of pasture (often resulting in a much lower impact on the river systems) pike was considered to be present according to the model.

4.2.3 Variables selected by genetic algorithms

In this part of the results, a chromosome with 20 input variables was introduced to machine learning (each input variable was equal to one gene). The standard settings were used for the genetic algorithms. The only exception was the number of folds when using the wrapper subset evaluator. Here, a 3-fold was considered for estimating subset accuracy. For simplicity in the interpretation of the results, the pruning was set at its default value (the default pruning is 0.25). The first step was to apply the genetic algorithms for the 20 input variables. Here, the full training dataset (150 instances) was used for selecting the major important input variables to predict the pike's habitat requirements. After the new variables were selected by the genetic algorithms, they were tested in terms of the model evaluation (CCI and Kappa) as well as the

trees induced by the model. This was achieved to compare the results obtained with and without a variable selection procedure based on a genetic algorithm.

An important reduction in the number of variables was observed when applying the genetic algorithms with the full training dataset. Here, the number of variables dropped from a total of 20 to 9. Based on the classification trees, 3 structural-habitat variables (slope, distance from source and wetted-width) were the main predictors for the pike's habitat. In total, the dominant variables selected by genetic algorithm were as follows:

- Slope
- Distance from source
- Wetted-width
- Nitrite
- Ortho-phosphate
- Total phosphorous
- BOD
- Suspended solids
- Huet-zonation

When analysing the ecological relevance of classification trees in combination with genetic algorithms, it seems that the importance of the three (mainly) structural variables distance from source, wetted-width and slope are very well confirmed. The effect of the average-depth on the presence/absence of pike was not confirmed after variable selection stage, while this variable was used in the study of Kerle et al. (2001). In addition to the selected structural-habitat characteristics, water quality variables also seemed to be of major importance for pike. The variable suspended-solids (related to water transparency) was for instance also selected by the genetic algorithm. Figure 4.2 shows the induced classification trees after the variable selection stage. Here, the best performing tree (according to the CCI and Kappa) is presented for pike. In this stage, the tree considered only 4 out of 9 input variables that were selected by genetic algorithms: distance from the source, Huet-zonation, wetted-width and total phosphorus.

```
Distance from source <= 4.5: PIKE ABSENT (9.64/0.87)
Distance from source > 4.5
|   Huet-zonation = barbel
|   |   Wetted-width <= 8: PIKE PRESENT (21.9/6.0)
|   |   Wetted-width > 8: PIKE ABSENT (4.0/1.0)
|   Huet-zonation = bream
|   |   Total-phosphorus <= 1.03: PIKE PRESENT (52.17/22.52)
|   |   Total-phosphorus > 1.03: PIKE ABSENT (5.58/0.68)
|   Huet-zonation = upstream: PIKE ABSENT (6.71/1.9)
```

Figure 4.2. Classification trees for pike after variable selection by genetic algorithms in 6 river basins in Flanders (Subset 1, PCF = 0.25). (Values between brackets indicate instances in which rules are true/false).

Table 4.4 gives a comparison of the model performances with and without variable selection stage. After variable selection, the subset 1 gave the highest performance (CCI = 72 % and Kappa = 0.44). After applying the genetic algorithms, however, the overall averages CCI and Kappa (SS Average marked in grey) were enhanced but the prediction by the induced trees still was not so good.

Table 4.4. Comparison of the model evaluation of the induced classification trees after and before variable selection in the river basins in Flanders (PCF = 0.25).

CCI	**Before variable selection**	**After variable selection**
Subset 1	60	72
Subset 2	54	60
Subset 3	52	54
Subset Average	55	62
Subset Standard deviation	3	9
Kappa	**Before variable selection**	**After variable selection**
Subset 1	0.20	0.44
Subset 2	0.08	0.20
Subset 3	0.04	0.08
Subset Average	0.11	0.24
Subset Standard deviation	0.07	0.18

Table 4.5 compares the effect of pruning optimisation (PCF = 0.25) on the tree size with and without variable selection stage. Here, the subset 1 gave the equal tree size after and before applying the genetic algorithms. After variable selection, a tree size 10 was noticed as the best outcomes (CCI = 72 % and Kappa = 0.44).

Table 4.5. Comparison of the tree size of the induced classification trees (PCF = 0.25) after and before variable selection in 6 river basins in Flanders (PCF = 0.25).

Tree size	**After variable selection**	**Before variable selection**
Subset 1	10	10
Subset 2	8	3
Subset 3	3	17
Subset Average	7	10
Subset Standard deviation	4	7

Table 4.6 ranks the most important predictors for the classification trees after the variable selection. Ranking of these variables was based on the trees generated at PCF 0.25 for each subset. According to information obtained in Table 4.6, the main variables obtained for pike were Huet-zonation, distance from the source and BOD. In the first level, the Huet-zonation was repeated 2 times, followed by distance from source (one time). In the second rank, BOD was repeated two times. Here, again the type of Huet-zonation was categorised as an important predictor. In the fourth and fifth level, no variables were ranked by the classification trees. By comparing the ranking of the variables in Table 4.6 (with variable selection) and Table 4.3 (without variable selection), one can see that some predictors such as the type of Huet-zonation, BOD, total phosphorus and wetted-width were used in both trees.

Table 4.6. Major variables of the classification trees for pike after variable selection by genetic algorithms (PCF = 0.25) in the river basins in Flanders.

Variables	**First**	**Second**	**Third**	**Fourth**	**Fifth**
Subset 1	Distance from the source	Huet-zonation	Wetted-width Total phosphorus	-	-
Subset 2	Huet-zonation	BOD	-	-	-
Subset 3	Huet-zonation	BOD	Suspended solids	-	-
Three times	-	-	-	-	
Two times	Huet-zonation	BOD	-	-	-
One time	Distance from the source	Huet-zonation	Wetted-width Total phosphorus Suspended solids	-	-

4.3 Discussion

Predicting presence and absence of organisms in ecology by ecological modelling techniques has parallels in other fields e.g. in remote sensing (Helmer et al., 2000) and medical diagnostics (Walker et al., 1999). According to Humpl and Pivnicka (2006), for ecological communities, species presence/absence data emphasise a coarser grain of environmental tolerance, whereas relative abundance data present fine-grain information because they emphasise local peaks in species performance. In particular, prediction of the habitat suitability of a species based on the habitat characteristics is an interesting objective in basic and applied ecology (Whitehead et al., 1997) and can be of high interest to managers and engineers dealing with freshwater ecosystems (Mastrorillo et al., 1997).

The aim of current work was to evaluate the abilities of classification trees in combination with genetic algorithms to predict the pike's habitat requirements in relation to their main influencing environmental variables in 6 river basins in Flanders. Ecological models are indeed considered as essential tools in modern water management. Many mistakes, however, can make models unreliable e.g. careless use of input data, insufficient calibration and validation, departing from wrong model assumptions. These difficulties should lead to more

time spent on the model design, before models can be used for practical applications in order to avoid these drawbacks.

For model development, the important thing is that models have to be tested based on reliable model evaluation. In the literature review, serious gaps can be found regarding several model performances. Several statistical methods were developed to evaluate model performance. For instance, methods used to evaluate the performance of presence/absence models in a sample of ecological publications (1989-99) revealed that Kappa was used in 11 % of the cases (Manel et al., 2001). Most modelling techniques also included logistic regressions (79 %), discriminate analysis (22 %), artificial neural networks (4 %) and other methods (3 %). Contribution of fish in these presence/absence modelling was only 9 % (Manel et al., 2001). A value of Cohen's Kappa statistic higher that 0.4 indicates a reliable model, while lower values show poor model performance. Landis and Koch (1977) presented the degree of agreement that exists when using the Kappa statistics, hence defining the performance of Kappa values as: ≤ 0 poor; 0-0.2 slight; 0.2-0.4 fair; 0.4-0.6 moderate; 0.6-0.8 substantial and 0.8-1 almost perfect.

The correctly classified instances (CCI), on the other hand, are the most logical performance measure due to its easy interpretation towards water managers and policy makers. It is a simple measure for the accuracy of the prediction. Predictions can be deceptive if only the overall prediction success is used as performance measure. For instance, CCI itself is not able to give a good prediction when species are found commonly or only rarely in the monitored site. In these cases, models are inclined to predict that very common species are always present, while very rare ones are always absent. Accordingly, in the present work, a combination of both CCI (from an ecological point of view) and Kappa statistic (from a mathematical point of view) were used to evaluate model performances.

On the basis of the guideline presented by Landis and Koch (1977), the model developed for pike in the Flanders can be considered of satisfactory performance only in the subset 1 at PCF 0.5 (CCI 78 % and Kappa 0.56), while the models developed for other subsets did not yield a promising prediction. Noise in the data and missing values in the variables could probably be the main reasons for the incapability of the developed model to cope with these drawbacks.

When developing a tree model, pruning algorithms can be applied to reduce the complexity of trees by eliminating less informative variables. This optimisation procedure is of special interest when the models are used in decision making of river restoration and conservation management (Goethals and De Pauw, 2001; Dakou et al., 2006). In this study, pruning of the models through the reduction of the confidence factor affected the tree sizes. In Figure 4.1 (subset 1) the tree sizes dropped from 29 (PCF = 0.5) to 1 (PCF = 0.01). Here, very simple trees were induced when a low pruning confidence factor was applied. Furthermore, in most cases, lowering the pruning confidence factor resulted in a reduction of model performance.

In the Flemish river basins, some variables were able to predict the pike's habitat requirements: the type of land-use and Huet-zonation, wetted-width, average-depth, BOD etc. From ecological point of view, these factors are considered to be structuring pike populations. In particular, in the case of the land-use, several activities play the key roles e.g. forestry, arable-land and so on. Pike populations were absent when the type of land-use was arable. In contrast, they were present when the type of land-use was pasture. One of physical-habitat variables used by the classification trees was the average-depth. This variable forms an essential feature in pike habitat preference considering the need for shelter against predation. According to Casselman and Lewis (1996) and Vuorinen et al. (1998), there is a positive correlation between depth at the habitat and size of the young-of-the-year pike, at least until they reach 150 mm in length. Eklov (1997) stressed that the deep water and steeply sloping parts areas are usually occupied by top predators such as pike and perch (*Perca fluviatilis* L.).

The genetic algorithms have been successfully applied in ecological modelling e.g. for modelling the abundance of algae (Recknagel and Wilson, 2000) and selecting input variables for the prediction of macroinvertebrates in Flanders (D'heygere et al., 2003). Considering the influence of each environmental variable on the prediction of the pike's habitat with use of genetic algorithms, the classification trees gave valuable information concerning some predictors and that the number of selected variables dropped from 20 to 9. The overall success of prediction relatively increased depending on the chosen variable selection method.

Distance from the source and wetted-width were determined as important predictors before and after applying genetic algorithms. From ecological perspectives, these variables seem to be quite logic and the relevance of these (integrating) structural-habitat variables is confirmed

in the literature (e.g. Kerle et al., 2001). The variable 'distance from source' reveals that pike populations prefer to inhabit downstream part of rivers. This variable also was in accordance with the study of Brosse et al. (1999). These authors demonstrated that the variables which mostly influenced the abundance of pike were distance from the source and flooded vegetation cover. Some variables were expected to be induced by the classification trees but they were not. After and before the variable selection procedure, the classification trees did not present some variables such as dissolved oxygen in the rivers in Flanders. The effect of critically low oxygen concentration on the survival of pike has been widely studied. The relation between critical oxygen concentration and temperature is highly significant, curvi-linearly related (Casselman, 1978). After the variable selection procedure by means of the genetic algorithm, some predictors were again selected by the classification trees e.g. the type of Huet-zonation (in particular in upstream part of rivers where pike was absent), distance from the source and wetted-width. This confirms the importance of these variables for pike in the rivers and also the stability of the method. Moreover, the genetic algorithm had a positive effect on the reliability of the developed classification tree models.

4.4 Conclusion

This part of the study aimed to develop and optimise classification trees in combination with genetic algorithms for the prediction of the pike's habitat suitability based on presence and absence data in 6 river basins in Flanders. Given the input variables, the prediction of pike (based on presence and absence data) was assessed by two model performances. The pruning had an effect on the tree size and complexity of trees. Looking at the average predictive results (CCI 55 % and Kappa 0.11), the classification trees gave a poor prediction. The variables used by the classification trees (before the variable selection was applied) were land-use, BBI, the type of Huet-zonation, wetted-width, flow velocity, average-depth, chloride and BOD. However, in some cases the results were instable when looking for the important variables to develop the classification trees.

With the use of genetic algorithms, ten variables for the pike's habitat requirements were selected. The elimination of variables by this method could slightly improve prediction performance. The results obtained in this chapter also showed that classification trees can be

helpful methods to present the relationship between environmental river characteristics and presence/absence of pike.

Chapter 5

Development of habitat suitability models for pike using artificial neural networks in combination with genetic algorithms

5.1 Introduction

It was demonstrated that artificial neural networks imitate aspects of biological information processing for data modelling and can be useful in many domains e.g. ecology (Whitehead et al., 1997; Recknagel, 2001) and geographical information system (Wang, 1994). In this sense, the prediction of aquatic communities such as fishes, algae and macrophytes by ANNs has in recent times been argued by many authors (Maier and Dandy, 2000; Scardi, 2001; Olden and Jackson, 2002; Lee et al., 2003). According to Haykin (1999) and Walczak and Cerpa (1999), some criteria are important for the development an ANN: 1) selection of relevant environmental input variables; 2) choosing the backpropagation algorithm as appropriate learning method and 3) searching for the optimal ANN architecture (the number of hidden layers in the ANN architecture and the number of neurons in the hidden layer(s)).

The aim of this chapter is to demonstrate the relationship between river basins characteristics and the presence and absence of pike. More specifically, this chapter mainly aims to discuss the development and optimisation of artificial neural networks to obtain the best model configuration for prediction of pike in 110 sampling sites in the Flemish river basins. First, predictive performances in terms of correctly classified instances and Cohen's Kappa statistics are presented. For each of the three subsets considered for ANN models, different hidden neurons were tested. After that, a genetic algorithm was applied in combination with ANNs for determining the appropriate environmental input variables. Finally, an interpretation and discussion of networks and genetic algorithms is given.

5.2 Results

5.2.1 Assessment based on model performance criteria

The average model performances of ANN models obtained by applying different hidden neurons (in one hidden layer) are presented in Table 5.1. Here, a feed-forward neural network called multilayer perceptron was trained with a back-propagation algorithm (Rumelhart et al., 1986) using the software package Weka 3-4 (Witten and Frank, 2000). Other parameter settings in the artificial neural networks toolbox were default and were determined through experience: Learning rate = 0.3, momentum = 0.2, random seed = 0, training time = 500 (see methodology). The training was stopped when the error per epoch in the validation set started

to increase in order to avoid overfitting. The optimal amount of neurons was obtained by trial and error, meaning that the number of neurons was varied (Figure 5.1). On the basis of this, it was possible to evaluate the effect of different hidden neurons on the model performances. As visualised in Figure 5.1 and Table 5.1, among 3 subsets, the best predictive performances were obtained in subset 1. Here, the CCI ranged from 82 % to 86 % and the respective Kappa from 0.60 to 0.72. The average CCI and Kappa of different hidden neurons were obtained 84 % and 0.66 respectively. This enabled ANN models to make a satisfying prediction for pike in the rivers.

In contrast, in the other two subsets (subset 2 and 3), the obtained results did not meet the minimum threshold value of CCI > 70 % and Kappa > 0.40. As can be seen in Figure 5.1, the majority of instances were not correctly classified for the major part of the model architectures in these subsets and also very low Kappa values were observed. Here, the CCI and Kappa seemed to be rather constant in most hidden neurons but when comparing the results over three subsets, the large standard deviations indicated a high instability of predictive power. The overall average CCI and Kappa in all different hidden neurons (as marked in grey in Table 5.1) showed that the models gave fair prediction (CCI = 66 % and Kappa = 0.30).

Table 5.1. Average model performances of ANNs obtained with applying different model architectures in the river basins in Flanders. Model stability is based on the standard deviation.

	CCI
Subset 1	84
Subset 2	57
Subset 3	56
Subset Average	66
Subset standard deviation	16
	Kappa
Subset 1	0.66
Subset 2	0.12
Subset 3	0.12
Subset Average	0.30
Subset standard deviation	0.31

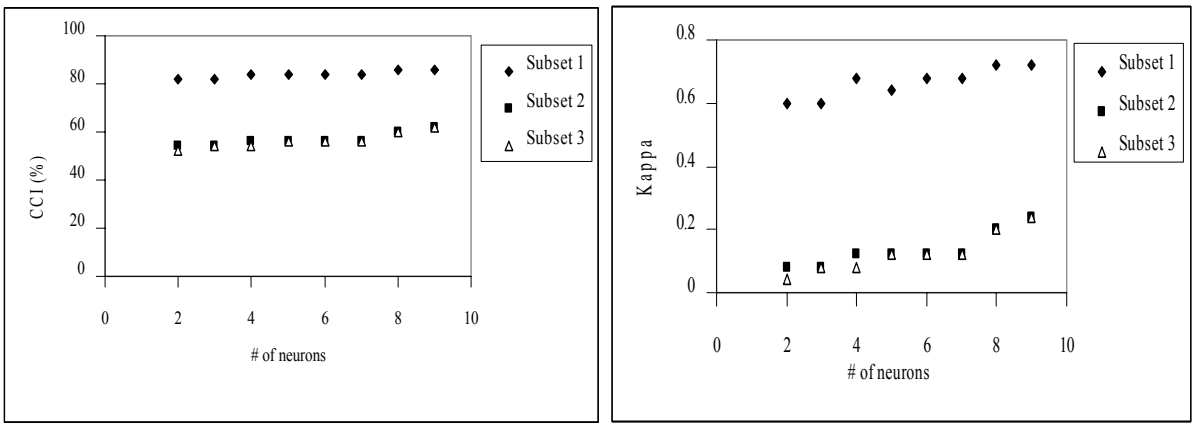


Figure 5.1. Comparison of the model performances of ANNs with different model architectures in the river basins in Flanders.

The optimal ANN model architecture was constructed with 9 hidden neurons, with an average CCI 70 % and Kappa statistics 0.40 (Figure 5.1). In that way, the network consisted of one input layer (with 20 neurons each representing one physical-chemical and structural-habitat variables) and one hidden layer with 9 neurons. The output neuron computed the value of the dependent variables (with 2 neurons indicating the presence and absence of pike). Therefore, for predicting the presence and absence of pike, the best ANNs architecture was a three-layered (20→ 9→2) feed-forward network with bias for the hidden and output layers (Figure 5.2).

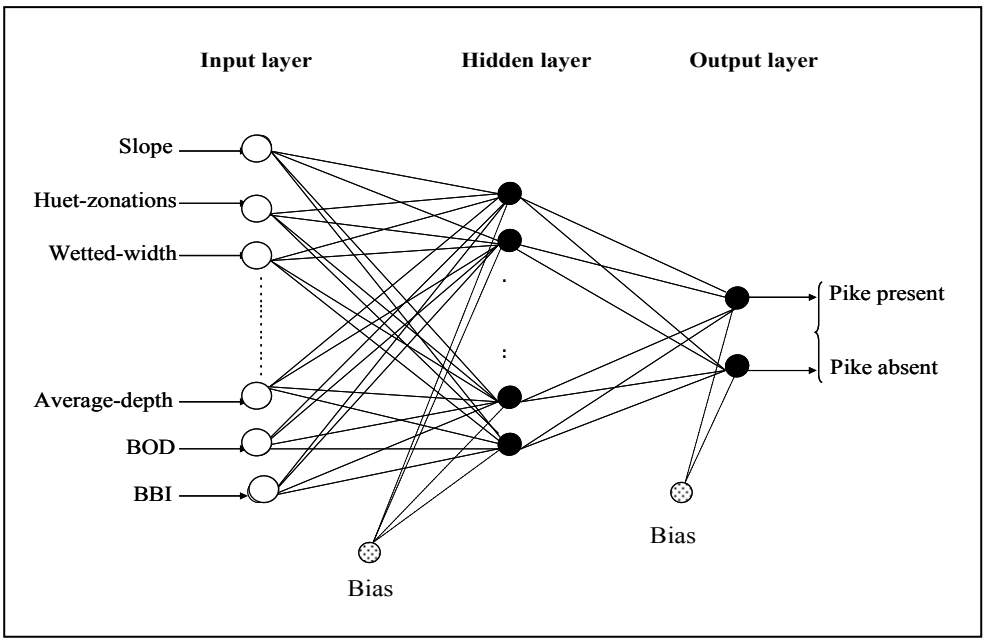


Figure 5.2. The optimal model architecture of the artificial neural networks used to model the presence and absence of pike with three-layered feed-forward. One input layer with 20 input neurons, one hidden layer with 9 neurons and one output layer with 2 neurons (pike present and absent).

5.2.2 Variables selected by genetic algorithms

As for classification trees, the full training dataset (150 observations) was used for selecting the most important input variables. For genetic algorithms, whole parameters were kept as default values except for the folds when using the wrapper subset evaluator. In that way, subset accuracy could be estimated using 3-fold cross-validation. For this part of the results, the standard model architecture (see methodology) also was used for the artificial neural networks. This was done either to compare the outcomes of ANNs before and after the variable selection stage or to make a clear interpretation of the obtained results.

After genetic algorithms were combined with ANNs, approximately 50 % of the input variables could be removed. Here, the selected variables were a mixture of some structural-habitat and pollution variables. The structural-habitat variables that were very crucial for predicting the pike's habitat requirements in the river basins were slope, average-depth,

wetted-width and distance from the source. Some selected predictors were related to pollution like nitrite, total-phosphorus, ortho-phosphate, conductivity and BOD. Additionally, other variables were Huet-zonation, Belgian Biotic Index (BBI) and type of land-use. After ANNs were combined by genetic algorithms, the developed models were characterised by an improved Kappa, and thus more reliable.

Table 5.2 and Figure 5.3 compare the predictive performances after and before the variable selection procedure. The model performances increased for all three subsets when applying a genetic algorithm. Among these subsets, again subset 1 gave the highest outcomes (CCI = 88 % and Kappa = 0.76). Looking at the average predictive results obtained over three subsets (as marked in grey in Subset Average), one can see that the ANNs were able to present good prediction based on a CCI 70 % and Kappa 0.40 for the pike's habitat suitability in the rivers.

Table 5.2. Comparison of model performances of the ANN models before and after the variable selection by genetic algorithms in the river basins in Flanders.

CCI	**Before variable selection**	**After variable selection**
Subset 1	82	88
Subset 2	54	62
Subset 3	52	60
Subset Average	63	70
Subset Standard deviation	17	16
Kappa	**Before variable selection**	**After variable selection**
Subset 1	0.64	0.76
Subset 2	0.08	0.24
Subset 3	0.04	0.20
Subset Average	0.25	0.40
Subset Standard deviation	0.34	0.31

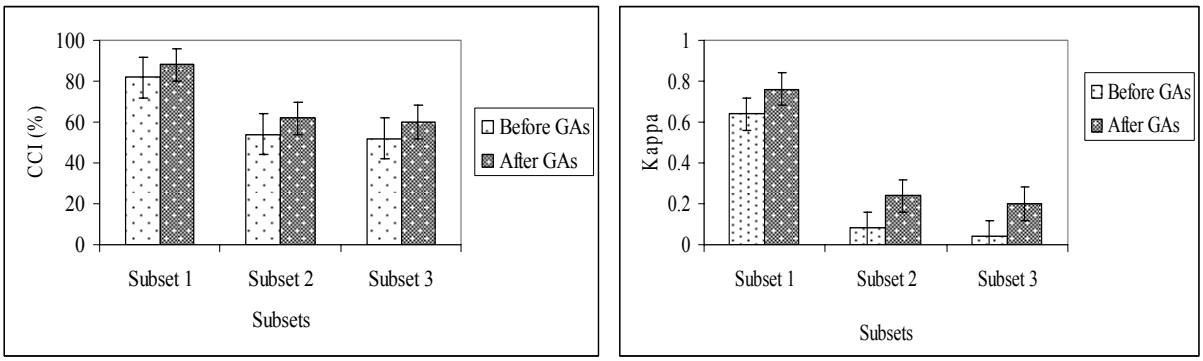


Figure 5.3. Comparison of model performances of the ANNs after and before the variable selection by genetic algorithms in the river basins in Flanders.

5.3 Discussion

It is possible to design an ANN with no hidden layers but these types of ANNs can only classify input data that is linearly separable (Haykin, 1994), which rigorously limits their application. Artificial neural networks that contain hidden layers have the ability to deal with nonlinear and complex problems and therefore can handle on more interesting problems (Haykin, 1994). Therefore, it is very important to optimise and develop different neural network architectures to acquire the best model configuration for a specified problem (Walczak and Cerpa, 1999; Olden, 2003; Dedecker et al., 2004; Dakou et al., 2006).

In the present research, several hidden neurons were examined by trial and error. Olden (2003) followed the same procedures so that 1 to 25 hidden neurons were tested to determine the best model configuration for neural networks. In this study, this was carried out to find the best model architecture for artificial neural networks. As visualised in Figure 5.2, the optimal model configuration of the networks were based on a three-layered (20→ 9→2) feed-forward network, determined as the optimal one giving the lowest error in the training and testing sets of data with minimal computing time (Lek et al., 1996; Brosse et al., 1999). The optimal number of hidden neurons found was in between the studies of Olden (2003) applying 7 hidden neurons and Brosse et al. (1999) and D'heygere et al. (2006) applying 10 hidden neurons.

The performance of the ANNs model was acceptable only in one of three subsets (subset 1), while models did not present any desirable prediction in the other two subsets for pike. It is thus likely that in these situations, the ANNs failed to recognise patterns in the data. After the variable selection stage by genetic algorithms, the selected variables reached to 12 for ANNs. The variables that were most selected were associated with the structural-habitat variables (slope, average-depth, wetted-width and distance from the source) and physical-chemical variables (conductivity, nitrite, ortho-phosphate, total phosphorus and BOD). In addition to these, the type of land-use, Huet-zonation and Belgian Biotic Index (BBI) had a contribution for the prediction of pike. The dependence of pike on the structural-variables is well studied e.g. distance from the source (Brosse et al., 1999; Brosse and Lek, 2000; Santoul et al., 2004; Humpl and Pivnicka, 2006; Spens et al., 2007), depth (Casselman and Lewis, 1996; Eklov, 1997; Vuorinen et al., 1998) and slope (Spens et al., 2007). Many physical-chemical variables (e.g. dissolved oxygen) were not selected as important predictors for the prediction of pike. If wastewater will be sufficiently treated in the future, structural variables will become increasingly important.

The lack of illustrative power of the ANNs models is a major concern to ecologists since the interpretation of statistical models is desirable for gaining knowledge of the relationships driving ecological phenomena (Olden and Jackson, 2002). In this way, ANN models have been labelled as 'black box'. To make them more transparent and enlarge their explanatory capacity, different methods can be used such as contribution or sensitivity analysis (Goh, 1995; Scardi and Harding, 1999; Olden and Jackson, 2002; Dedecker, 2005; Goethals, 2005) and variable selection methods by genetic algorithms (e.g. D'heygere et al., 2003). The selected variables by genetic algorithms as such could help to identify the major environmental factors influencing the habitat suitability of pike in the river basins in Flanders. On one hand, the relevance of the structural-habitat variables was confirmed by genetic algorithms, as these appeared in the list of selected variables. On the other hand, also some water quality variables seemed to play a major role for the presence of pike in the studied river basins. These water quality variables have moreover a close relation with the variables land-use and Belgian Biotic Index (BBI) that were also selected by this method.

An assessment of model performance (based on CCI and Kappa) demonstrated that when the ANNs were integrated with a variable selection stage, the predictive performance increased

for the prediction of pike in the rivers. D'heygere et al. (2006) derived similar conclusions when predicting macroinvertebrate taxa using artificial neural networks in Flanders. The highest performance after the variable selection stage can be explained by the way artificial neural networks work. In artificial neural networks, the unrelated information is sent through the nodes and can as such slightly alter the connection weights and influence the overall performance of artificial neural networks. Another advantage of the variable selection stage is the decreased network size. This can result in higher processing speeds, hence less data are needed to estimate the connection weights efficiently (D'heygere et al., 2003). The number of input nodes is fixed by the number of input variables, while the number of output nodes is fixed by the number of output variable(s) represented as the presence and absence of pike. The number of input nodes was determined by the results of the genetic algorithms. Adding a variable can lead to 10 extra connection weights to be calculated.

5.4 Conclusion

The main aim of this chapter was to develop and optimise the artificial neural networks in conjunction with genetic algorithms for the prediction of the pike's habitat suitability based on presence and absence data in 6 river basins in Flanders.

ANN models are known as 'black box'. To solve this drawback and make ANNs more transparent, different methods have been suggested such as contribution or sensitivity analysis and genetic algorithms. These models can play an important role to find general trends on habitat suitability of aquatic communities. The application of artificial neural networks is an advantage if relations between environmental input variables are unknown, very complex or nonlinear. In combination with a specific procedure for the selection of the most important variables by a genetic algorithm, the complexity of the models could be diminished. This could result in an improvement of the generalisation of the induced models, which finally bring about a simplification and a better understanding of the underlying relationships in the data.

Network architecture is generally known to be highly problem dependent. A good model performance was obtained only in one of three subsets (subset 1), while the others did not make use of the environmental variables to predict pike. Here, most different model

architectures gave similar results. The best model architecture over three subsets was obtained by trial and error. This led to an optimal architecture with 9 hidden neurons. The average CCI and Kappa over different model architectures showed that the predictive performances of ANNs were in the neighbourhood of threshold values.

After the variable selection stage, 12 important predictors could be used for the prediction of the habitat suitability of pike in the rivers. This set of variables consisted of a combination of structural-habitat, water quality and integrative variables. After removing the irrelevant variables, the predictive performances of the developed ANN-models improved according to the Kappa.

Chapter 6

Development of habitat suitability models for pike using support-vector machines in combination with genetic algorithms

6.1 Introduction

Application of computational tools to predict fish habitat requirements is of great importance in ecology. In this context, support-vector machines are among the best methods, giving excellent generalisation performance on a wide range of problems (Keerthi et al., 2001; Burbidge et al., 2001). SVMs determine a separating hyperplane in order to perform pattern recognition and prediction by classification. More recently, they have also been applied successfully to regressions and time series. Mathematically, SVMs are a range of classification and regression algorithms that have been formulated from principles of statistical learning theory developed by Vapnik (Vapnik, 1995). Empirical performance of SVMs is generally as good as the best ANN solutions. This is because there are fewer model parameters to optimise in the SVMs approach, reducing the possibility of overfitting the training data and thus increasing actual performance (Brown et al., 1999). In recent years, a number of nonlinear classification and regression SVMs have been developed and these have been benchmarked against artificial neural networks. Despite many advantages, two problems in the SVMs models are still far from resolved: feature selection and parameter settings. These have been proved to have important impacts on the efficiency and accuracy of SVM classification (Huang and Wang, 2006; Yang et al., 2008). Proper SVM parameter settings can help to improve the SVM classification accuracy (Huang and Wang, 2006; Yang et al., 2008). Due to the importance of these problems, there has been much recent work in this domain. Many algorithms have been suggested to solve these problems separately such as genetic algorithms (Lucasius and Kateman, 1993) and recursive feature eliminations (RBF) (Guyon et al., 2002).

The aim of the current chapter is to develop habitat models for pike in Flanders based on SVMs in combination with genetic algorithms. The first part of the results assesses the predictive performance of SVMs based on correctly classified instances and Cohen's Kappa. In the second part, the application of the integrated genetic algorithms and support-vector machines is presented for selecting the appropriate variables. The third and last part of the results compares the predictive performance with and without a variable selection procedure. Eventually, a short discussion and conclusion is given regarding the applied techniques.

6.2 Results

6.2.1 Assessment based on model performance criteria

The average predictive results of support-vector-machines are presented in Table 6.1. To obtain the optimal results regarding predictive performances, the support-vector machines were optimised based on applying different exponents. Other parameter settings in Weka toolbox (weka.classifiers.functions.SMO) were default and were determined through experience (see methodology). As illustrated in Table 6.1 and Figure 6.1, compared with 3 subsets, the best and highest predictive performance was obtained in subset 1, resulting in an average CCI 81 % and Kappa 0.62. The CCI ranged from 60 % to 86 % and their related Kappa from 0.20 to 0.68 (Figure 6.1). In subset 2, the range of CCI was between 56 % and 62 % and Kappa between 0.12 and 0.24 and roughly the same outcomes were observed in subset 3 (CCI varied from 50 % to 62 % and Kappa from 0.00 to 0.24). The last two subsets did not present a good prediction compared to subset 1. The outcomes obtained with different exponents revealed that when only the default value of this parameter was used, the model evaluation was lower (in WEKA tool, the default exponent is indicated with 1). On the contrary, the prediction power was higher when applying different exponents. The overall average CCI and Kappa (Subset Average as marked in grey) were obtained 66 % and 0.33 respectively, which is fair prediction for pike occurrence.

Table 6.1. Average model performances of SVMs obtained by applying different exponents in the river basins in Flanders. Model stability is based on the standard deviation.

	CCI
Subset 1	81
Subset 2	60
Subset 3	58
Subset Average	66
Subset standard deviation	13
	Kappa
Subset 1	0.62
Subset 2	0.20
Subset 3	0.16
Subset Average	0.33
Subset standard deviation	0.25

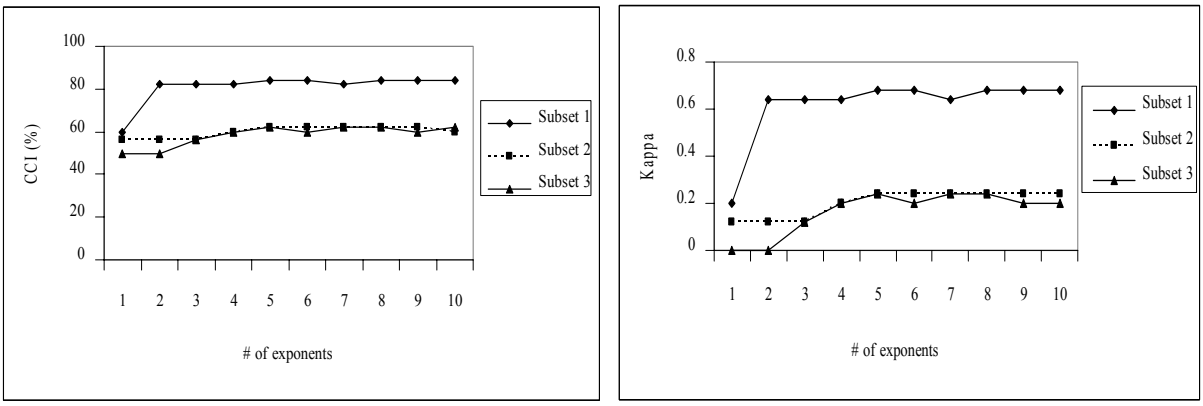


Figure 6.1. Comparison of the predictive performances of SVMs with applying different exponents in the river basins in Flanders (1 is the default exponent in the Weka toolbox).

6.2.2 Variables selection by genetic algorithms

A set of 150 observations (full training datasets) were employed for selecting the key input variables for pike. For genetic algorithms, the parameter settings were left as default values except for the number of folds when using the wrapper subset evaluator. In that case, a 3-fold was applied to estimate subset accuracy. Besides, in this part of the results, the whole standard settings also were used for support-vector machines. This was carried out to compare the outcomes of support-vector machines before and after the variable selection procedure. Furthermore, this could ease the interpretation of the obtained results.

A combination of the genetic algorithms and SVMs resulted in the elimination of 11 input variables. A large deal of the selected attributes is associated with the physical river habitat such as slope, average-depth, wetted-width and distance from the source, but also physical-chemical variables (conductivity, nitrite, ortho-phosphate, suspended-solids and BOD) were selected. Other selected variables were the integrated river characteristics such as type of land-use and Huet-zonation. Strikingly, the BBI did not appear in the list, this could mean that this index is too general, and that pike presence is to a high extend related to the river type.

In Table 6.2 and Figure 6.2, a comparison is made regarding the predictive performances of SVMs with and without variable selection procedure. After the unrelated variables were eliminated, the model performance increased for all three subset; particularly the improvement was very noticeable in subset 1. The overall average prediction improved (as marked in grey in Subset Average) so that the CCI and Kappa were more or less acceptable (CCI = 67 % and Kappa = 0.32 were the best outcomes), but the high standard deviation obtained after the variable selection (SD CCI = 67 % ± 15 and SD Kappa = 0.32 ± 0.28) indicated that the results were less stable than those obtained before the variable selection (SD CCI = 55 % ± 5 and for SD Kappa = 0.11 ± 0.10).

Table 6.2. Comparison of CCI and Kappa of the support-vector machines before and after variable selection stage by genetic algorithms in the river basins in Flanders.

CCI	**Before genetic algorithms**	**After genetic algorithms**
Subset 1	60	84
Subset 2	56	60
Subset 3	50	56
Subset Average	55	67
Subset standard deviation	5	15
Kappa	**Before genetic algorithms**	**After genetic algorithms**
Subset 1	0.20	0.64
Subset 2	0.12	0.20
Subset 3	0.00	0.12
Subset Average	0.11	0.32
Subset standard deviation	0.10	0.28

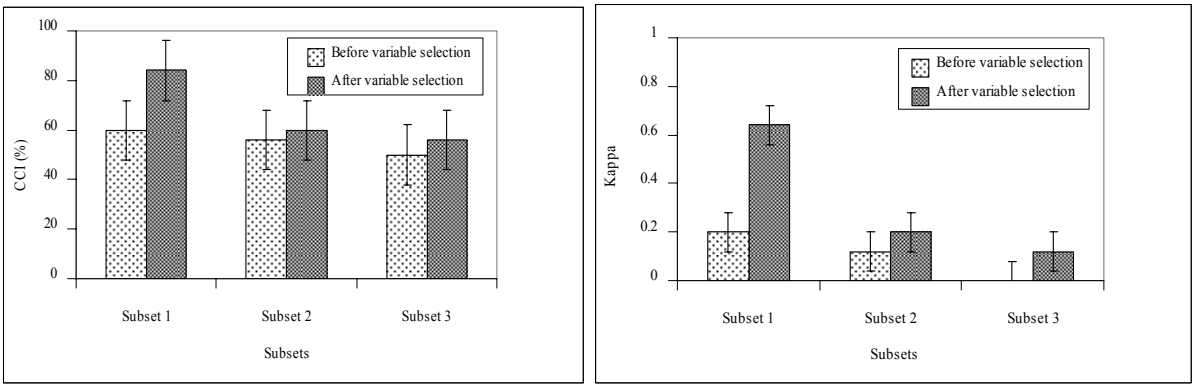


Figure 6.2. Comparison of the predictive performance of the support-vector machines before and after variable selection in the river basins in Flanders.

6.3 Discussion

SVM seems a promising tool for classification, particularly for ecological data, where the difficulties of dimensionality are often present; a technique that can deal with many predictors is very useful. Another advantage of SVMs for ecological data is that they do not assume relationships to be linear (Akkermans et al., 2005).

With datasets examined in this chapter, SVMs gave fair results when optimising them with different exponents so that the overall predictive performances improved more than 10 % compared to their standard settings. Dibike et al. (2000) studied on the overall performance of SVMs and stated that these models provide an attractive approach to data modelling in many domains such as civil engineering and hydroinformatics. Determining the proper parameter settings in SVMs is still a heuristic process and almost surely suboptimal (Dibike et al., 2000). Akkermans et al. (2005) compared the predictive performances of two classification methods namely support-vector machines and logistic regressions to predict macro-fauna community types from environmental variables. It was demonstrated that support-vector machines gave higher prediction relative to logistic regressions. Predictive performances of support-vector machines can work out reasonably well in small classes. Besides, multicollinearity does not seem to be a large problem for support-vector machines. When a variable selection stage was integrated with support-vector machines, 9 out of 20 variables could be omitted. Therefore, the remaining variables were used to compare the outcomes of predictive performances after and before the variable selection. According to SVMs, the most important variables for pike were average-depth, wetted-width, distance from the source and slope. Pike's abundance is negatively correlated with the gradient slope. In addition to this, some of the variables were related to eutrophication such as suspended-solids, conductivity, nitrite, ortho-phosphate and BOD. As for other modelling techniques discussed in the previous chapters, land-use and river typologies based on Huet's scheme (Huet, 1949, 1954) also were recognised as crucial predictors for the prediction of pike occurrence. Unlike ANNs and CTs, the relevance of BBI and total-phosphorus for pike in the rivers was not confirmed by SVMs after variable selection stage. It seems that the high correlation coefficient observed between total phosphorus and ortho-phosphate prevented GAs to make a selection for both variables.

An assessment of model evaluation after variable selection demonstrated that the elimination of variables in SVMs could enhance CCI and Kappa. Yang et al. (2008) evaluated the use of SVMs combined with genetic algorithms to predict pharmacokinetic properties of drugs. They demonstrated that the integration of genetic algorithms and SVMs significantly improve the overall prediction accuracy of the models. A similar study also was conducted by Ma et al. (2008) when assessing the effect of applying genetic algorithms in combination with SVMs. The prediction accuracy increased when a variable selection scheme was integrated with support-vector machines.

6.4 Conclusion

SVMs still are new tools of the machine learning community that have recently gained popularity in many domains. They are considered as very successful methods, in particular in classification problems.

This chapter integrated genetic algorithms and support-vector machines to establish a prediction model for habitat requirements of pike based on presence and absence data in Flemish river basins. After optimising the SVMs, good model performances (according to CCI and Kappa) were obtained only in one of three subsets (subset 1). Applying the different exponents of SVMs yielded the highest overall predictive results (CCI = 66 % and Kappa = 0.33) compared with the standard setting values, but choosing the proper parameters in SVMs still is considered as a big problem.

When genetic algorithms were combined with support-vector machines, the most essential variables were determined for the prediction of pike. As expected, most of these variables were related to the habitat and pollution ones. In that stage, the average predictive performances of SVMs improved (CCI reached from 55 % to 67 % and Kappa from 0.11 to 0.32).

This part of results showed that SVMs, as data-driven techniques, can be helpful in many domains, in particular in ecology. As such, they can provide managers an effective way and clear understanding for predicting fish communities.

Chapter 7

Development of habitat suitability models for pike using logistic regressions in combination with genetic algorithms

7.1 Introduction

Conventionally, models that are applied in ecology to predict species presence and absence as well as abundance have been based on linear relationships with environmental variables. Datasets are in turn presumed to have normal errors, suitable for in linear regressions, multiple regressions and multiple discriminant analysis. Complications in fulfilling these assumptions have often raised statistical and theoretical concerns (Austin and Meyers, 1996; Lek et al., 1996), so that new modelling paradigms have been evolving. They consist of linear methods e.g. logistic regressions, which contain binomial error, and are already broadly applied (Osborne and Tigar, 1992; Green et al., 1994; Austin and Meyers, 1996).

If one wants to understand the relationship between organisms and environmental predictors, two important methods are commonly used: 1) binary (dichotomous) and multinomial logistic regressions (it is binary when there are only two classes and is multinomial when there are multiple classes) and 2) discriminant analysis. In logistic regression methods, the whole range of methods and techniques for generalised linear models is available, such as the selection of environmental variables. Logistic regressions can model the probability of occurrence of each organism as a function of the environmental data. They are the preferred methods to link biotic to abiotic variables. In such a way, the outcomes can be applied to predict the occurrence of organisms from environmental variables. They are similar to linear regression models, but are suited to model where the dependent variables are dichotomous. McCullagh and Nelder (1989), Cessie and van Houwelingen (1992) and Hosmer and Lemeshow (1989) were among those who comprehensively described logistic regressions.

This chapter aims to develop habitat models for pike in Flanders based on logistic regressions together with the genetic algorithms. As applied for the previous modelling techniques, first of all, the attention is paid to the model performance of the logistic regressions. Then, the most informative variables induced by genetic algorithms will be presented. As already described, logistic regressions can directly model the occurrence of an organism, therefore some part of this chapter will additionally discuss the predicted probability of pike occurrence in relation to some important predictors. Finally, this chapter ends with a brief discussion and conclusion regarding the applied techniques.

7.2 Results

7.2.1 Assessment based on predictive performance criteria

The results obtained based on predictive performances (in terms of CCI and Cohen's Kappa) for logistic regressions are presented in Table 7.1. In contrast to the other models developed for pike, all the default options were used for these models because compared to other techniques, logistic regressions require less tuning of parameters (Witten and Frank, 2000). As indicated in this table, neither the CCI and nor Kappa were good for the prediction of the presence or absence of pike. The highest predictive results, however, were obtained in subset 1, leading to a CCI 60 % and Kappa 0.20, but looking at the average of the three subsets (marked in grey in Subset Average), one can directly see that the models gave poor predictions (CCI = 55 % and Kappa = 0.09). As illustrated in Figure 7.1, the low values of Kappa, particularly in subsets 2 and 3, prevented the models to make a reliable prediction for pike occurrence.

Table 7.1. Model performance of logistic regressions for pike in 6 Flemish river basins. Model stability is based on the standard deviation.

	CCI
Subset 1	60
Subset 2	52
Subset 3	52
Subset Average	55
Subset standard deviation	5
	Kappa
SS1	0.20
SS2	0.04
SS3	0.04
Subset Average	0.09
Subset standard deviation	0.09

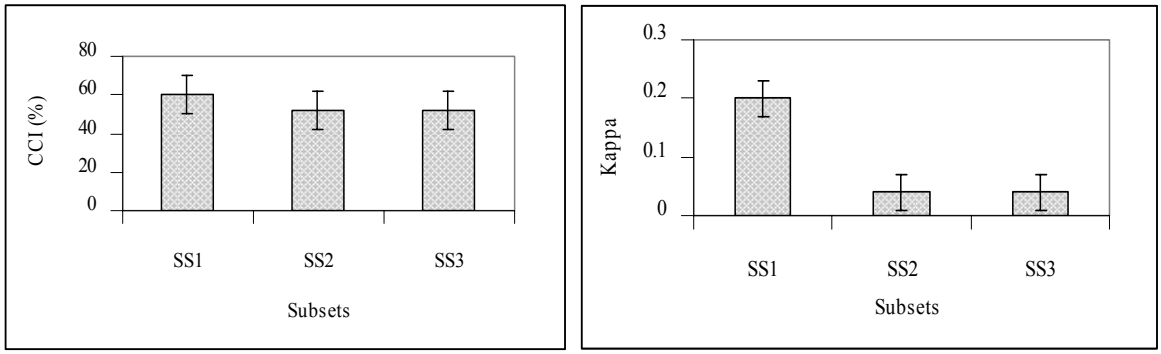


Figure 7.1 CCI and Kappa of logistic regressions for pike in 6 river basins in Flanders (SS = Subset).

7.2.2 Variables selection by genetic algorithms

The whole training datasets were introduced to genetic algorithms to determine the most informative environmental variables influencing the occurrence of pike in the sampling sites. The default options were kept as constant as possible. The only exception was the number of folds for the wrapper subset evaluator. In that way, a 3-fold was employed to estimate subset accuracy. When genetic algorithms were combined with logistic regressions, the number of variables was reduced by 8. According to this model, the slope, wetted-width, average-depth and distance from the source were the most explanatory variables that were used to predict habitat requirements of pike. Other important input variables were those related to eutrophication and pollution in general like conductivity, nitrite, ammonium, ortho-phosphate, total-phosphorus and BOD. Type of land-use and Huet-zonation also played a key role for the prediction of pike in the rivers. In total, 12 variables were available for the logistic models. Similar to other models discussed in the previous chapters, the dependence of pike on all habitat variables was confirmed by logistic regressions combined with genetic algorithms.

To unravel the information on the impact of each of the input variables on the prediction of pike occurrence, logistic regressions were only applied on some variables. Figure 7.2 (*a* to *f*) shows the impact of some structural-habitat and physical-chemical variables on the probability of occurrence of pike in 6 Flemish river basins. Logistic models for P independent variables (such as those selected by the genetic algorithms) can be written as follows:

$P = 1(1 + \exp(-(B0 + B1X1 + B2X2 + \ldots + BkXk)))$

where B0 is a constant and Bi are coefficients of the predictor variables. The P is a probability in the range 0 to 1. The exp( ) function is *e* raised to power.

All graphs presented in Figure 7.2, were calculated based on this formula. The step function was used in the statistical package SPSS 16. Here, the probability of the presence and absence of pike was modelled as a linear function of the given variables. Models were fitted using the maximum likelihood method from Hosmer and Lemeshow goodness-of-fit test (Hosmer and Lemeshow, 2000).

According to Figure 7.2, a high gradient slope reflects the importance of a particular variable in that range. In the variable selection stage, the importance of some selected variables by genetic algorithms was highlighted. As illustrated in Figure 7.2, the effect of some of habitat and pollution variables can be visualised on the pike's prediction. A decrease in the probability of the pike's occurrence could be observed when the pollution variables such as total-phosphorus and biological oxygen demand (BOD) increased, but these two latter variables do not seem to be influencing the occurrence of pike as strong as the habitat variables such as the distance from the source and the wetted-width. In the case of distance from the source, it seems that the actual data are opposite with the expected ones, but the trend that was discovered by this multivariate technique is ecological relevant. This illustrates the added value of these methods, since they can filter out the effect of 'confusing' variables, which are in this case mainly related to water quality.

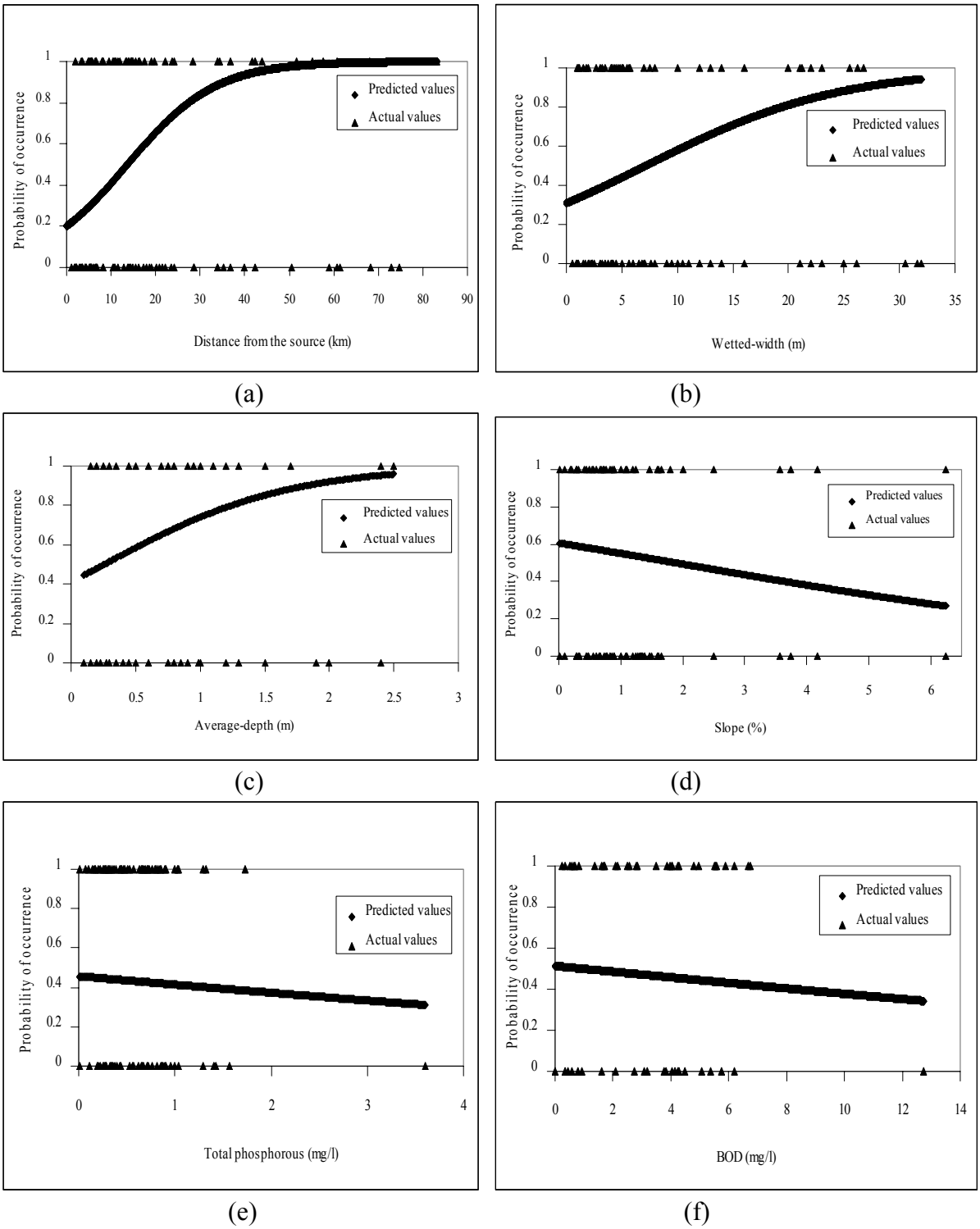


Figure 7.2 (a to e). The impact of some structural-habitat and water quality variables on the probability of occurrence of pike in 6 Flemish river basins (150 instances were available for the variables wetted-width and slope and 133 for the distance from the source and total-phosphorus, 78 for the average-depth and 70 for BOD).

In Table 7.2 and Figure 7.3, the predictive results of the logistic regressions are summarized and compared before and after the variable selection procedure. 12 out of 20 variables were selected by genetic algorithms, but the average predictive performances (as marked in grey in Subset Average) improved slightly (CCI reached from 55 % to 58 %) and Kappa from 0.09 to 0.16), nevertheless, they were not good indicators for the prediction of pike occurrence.

Table 7.2. Comparison of the model evaluation of the logistic regressions before and after variable selection stage by the genetic algorithms in the river basins in Flanders.

CCI	**Before variable selection**	**After variable selection**
Subset 1	60	62
Subset 2	52	56
Subset 3	52	56
Subset Average	55	58
Subset standard deviation	5	4
Kappa	**Before variable selection**	**After variable selection**
Subset 1	0.20	0.24
Subset 2	0.04	0.12
Subset 3	0.04	0.12
Subset Average	0.09	0.16
Subset standard deviation	0.07	0.07

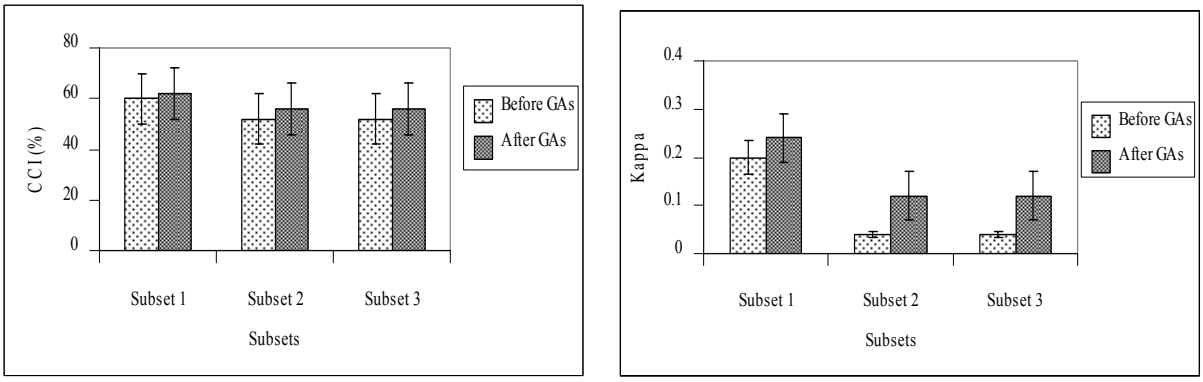


Figure 7.3. Comparison of the predictive performance of logistic regression models before and after the variable selection stage by genetic algorithms in 6 river basins in Flanders.

7.3 Discussion

The main aim of river ecology is to conserve and manage aquatic communities, therefore precise and accurate predictions are of great importance, so one should use suitable models in order to predict aquatic communities with minimum errors and maximum confidence.

In the present study, the average predictive performances of logistic regressions over three subsets showed that these models were not able to make a trustworthy prediction for pike occurrence. Logistic regressions are a parametric method for prediction and classification, but their predictive performances depend on the distribution of variables, size and quality of data (Sadat-Hashemi et al., 2005). When more predictors are nominal and responses are nonlinear, predictions will be confusing and will lack power (Sadat-Hashemi et al., 2005). Logistic regression analysis may be used to develop a prognostic model for a dichotomous outcome. In particular, when limited data are available, it is difficult to determine an appropriate selection of covariables for inclusion in such models (Steyerberg et al., 2000). Since these problems were encountered in this study, these models failed to make a reliable prediction for pike in the surveyed sites.

An important decrease in the number of selected variables was noticed when genetic algorithms were combined with logistic models, so that 8 variables were removed. The most important variables were related with structural-habitat variables such as slope, average-depth, wetted-width and distance from the source. The pollution variables (conductivity, nitrite, ortho-phosphate, total phosphorus, ammonium and BOD), on the other hand, were also important predictors that were recognized for the prediction of pike occurrence. The type of land-use and Huet-zonation also were placed in this category. The importance of river typology based on Huet-zonation (Huet, 1949, 1954) was already distinguished as an important characteristic to classify the rivers in terms of width and slope. The curves made by the logistic models for some structural-habitat variables (i.e. for the distance from the source) showed that pollution can affect the probability of occurrence of pike in the river basins. The logistic models developed in the study also revealed that nutrient loads in the river basins can restrict the pike presence.

The selection of input variables can be essential either for improving model performance or for policy and management objectives. It is therefore essential that key variables can be recognised which determine the presence or absence of pike. It has been shown that the developed automated variable selection scheme can trace these key variables. This method could also make data collection more efficient because some variables can be irrelevant. The problem is that when the river water quality improves, other variables that were eliminated before could become important in the future.

In the Flemish river basins, an assessment on the model evaluation showed that after variables selection, the predictive results increased somewhat, but the models still were not considered reliable. The low predictive power of logistic models can most likely be attributed to the noise in the datasets and missing values for the important variables on the one hand and the small datasets on the other hand.

7.4 Conclusion

This chapter aimed to develop logistic regressions in accompany with genetic algorithms for the prediction of pike occurrence based on presence and absence data in 6 river basins in Flanders. Choosing an analytical model depends on the theoretical limitations and the quality and quantity of the data and availability of adequate software. Distribution of variables, size and quality of data can seriously influence the predictive performance of logistic regressions. When more variables are nominal or where variables are nonlinear, prediction will lack power for such logistic models.

The highest predictive performances were obtained in subset 1, but the developed model was not able to predict the occurrence of pike in the surveyed sites. After variable selection, the most important variables (such as structural-habitat and pollution variables) were determined for the prediction of the pike's habitat requirements. After the less informative input variables were eliminated, the predictive results increased a little. However, considering the average CCI and Kappa of the logistic regressions, reliable predictions were not achieved for pike occurrence.

The outcomes obtained by logistic models revealed that in addition to structural-habitat variables, water quality variables also are considered as determinant predictors influencing the

presence/absence of pike. This technique is in particular interesting to understand the individual effect of each variable, what is not possible on the visualisation of the raw data.

Chapter 8

Comparison of habitat suitability model for pike using classification trees, artificial neural networks, support-vector machines and logistic regressions in combination with genetic algorithms

8.1 Introduction

When dealing with the prediction of species presence/absence, models are considered as important tools regarding resource assessment, environmental conservation and biological monitoring (Jongman et al., 1995; Fielding and Bell, 1997). Models developed with the aim of describing the habitat of a particular organism in a predictive way are referred to as habitat suitability models which are closely related to population dynamics modelling (Dzeroski, 2001). In the habitat suitability modelling approach, one is willing to either describe the suitability of a habitat for a species or use the information concerning the physical habitat of a certain species in order to predict the species absence/presence or abundance. Often these models are also referred to as habitat distribution models or simply habitat models. These models, either deterministic or stochastic, based on the relationships between environment variables and characteristics of fish populations are excellent tools for managers (Lek et al., 1996).

In recent years, with the advance of computational efficiency combined with sophisticated statistical methods, Machine Learning (ML) methods have been increasingly used as powerful tools in a wide variety of science disciplines including planetary science, computer science, bioinformatics and environmental science (Mjolsness and DeCoste, 2001). Among these techniques, artificial neural networks (Lek and Guégan, 1999), classification and regression trees (Dzeroski and Drumm, 2003), support-vector machines (Bahlmann et al., 2002; Guoa et al., 2005; Hsu et al., 2008) and logistic regressions (King and Zeng, 2000) have shown a high capability in ecological modelling because they combine trustworthy predictions with a better understanding of ecosystem interactions (Recknagel, 2001). Until now, however, a lot of ecological modelling techniques have been developed, but the practical application of them does not seem to meet the particular needs to support river management. Especially, for river restoration management, there is a need for tools to guide the investments essential to reach a good ecological status as set by the Water Framework Directive (WFD).

The main aim of this chapter was to compare the results of the 4 modelling techniques including classification trees (CTs), artificial neural networks (ANNs), support-vector machines (SVMs) and logistic regressions (LRs) in combination with genetic algorithms. In this way, the average predictive results of the 4 modelling techniques discussed in the

previous chapters (Chapters 4, 5, 6 and 7) are brought together to make a clear interpretation of the obtained results. In summary, this chapter consists of two main parts. The first part compares the outcomes of predictive performances of all modelling techniques for pike. In this part, the aim was to optimise all modelling techniques based on different parameter settings and then to get insight whether the changes made on parameter settings can affect the model performances of the 4 models or not. In the second part, the relevance of the most important variables selected by genetic algorithms will be assessed and compared for the 4 models. Here, the model performance criteria of 4 models with 20 variables were assessed with default parameter settings. After the variables were selected, the 4 modelling techniques were again assessed based on the model performances. In the end, a short discussion and conclusion is given about the applied techniques.

8.2 Results

8.2.1 A comparative discussion of the obtained results

In Table 8.1 and Figure 8.1, the outcomes of predictive performances of the 4 modelling techniques for predicting the presence/absence of pike are presented. This table was made on the basis of the averages obtained for each model in Chapters 4 (Table 4.2), 5 (Table 5.1), 6 (Table 6.1) and 7 (Table 7.1). As described in these chapters, all models were optimised with different parameter settings. The only exception was for logistic regressions of which no optimisation efforts were made. Among the 4 models, the predictability of pike's presence/absence based on artificial neural networks and support-vector machines was comparatively high. Both models gave similar prediction based on the overall average CCI and Kappa (CCI was 66 % for both models and Kappa was 0.33 and 0.30 for support-vector machines and artificial neural networks, respectively). The best results were obtained for subset 1, while the other two subsets didn't meet the threshold values for the prediction of pike occurrence. For artificial neural networks, the CCI ranged from 56 % to 84 % and Kappa from 0.12 to 0.66. For support-vector machines, the CCI ranging from 58 % to 81 % and Kappa from 0.16 to 0.62 were the best outcomes. These predictive results seem to be rather stable in subsets 2 and 3. Nevertheless, the models seemed to be quite instable when considering the large standard deviation over the three subsets. Looking at the overall average CCI and Kappa, one can conclude that the models were in the vicinity of the threshold values

(CCI 70 % and Kappa 0.4). This demonstrated that they presented a fair prediction for pike occurrence based on the presence and absence datasets.

The poorest predictive results were obtained with logistic regressions and classification trees. For both models, however, subset 1 gave the best results compared to the other two subsets. Looking at their average predictive performances (CCI 55 % for both models Kappa 0.19 for classification trees and Kappa 0.09 for the logistic regressions), one can see that both classification trees and logistic regressions failed to make satisfying prediction of pike occurrence.

Table 8.1. Comparison of model performances of the 4 models for pike in 6 river basins in Flanders (CCI and Kappa are based on the averages obtained from the previous chapters. Model stability is based on standard deviation).

	CCI				***Kappa***			
	CTs	**ANNs**	**SVMs**	**LRs**	**CTs**	**ANNs**	**SVMs**	**LRs**
Subset 1	62	84	81	60	0.48	0.66	0.62	0.20
Subset 2	53	57	60	52	0.07	0.12	0.20	0.04
Subset 3	51	56	58	52	0.02	0.12	0.16	0.04
Subset Average	55	66	66	55	0.19	0.30	0.33	0.09
Subset Standard deviation	6	16	13	5	0.25	0.31	0.25	0.09

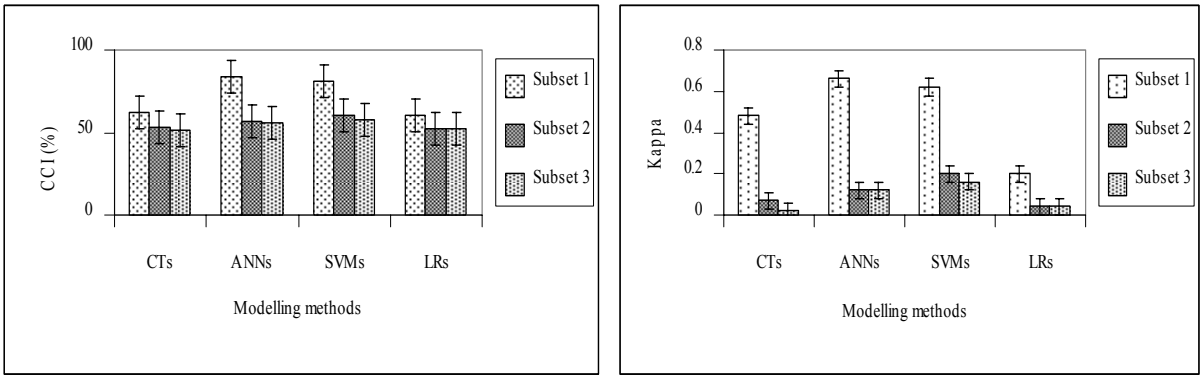


Figure 8.1. Comparison of model performances of the 4 models (CTs = classification trees, ANNs = artificial neural networks, SVMs = support-vector machines, LRs = logistic regressions).

8.2.2 Comparison of the selected variables for the 4 modelling techniques

Figure 8.2 illustrates the input variable selection scheme used for the 4 modelling techniques. As described already, a genetic algorithm was combined with each modelling technique for selecting the most important input variables. The dataset for each model consisted of a set of 150 observations collected in 6 river basins during the year 1991-2002 in Flanders. A chromosome representing 20 input variables was introduced to machine learning (each input variable was equal to one gene). The standard settings were used for the genetic algorithms as well as for each modelling technique.

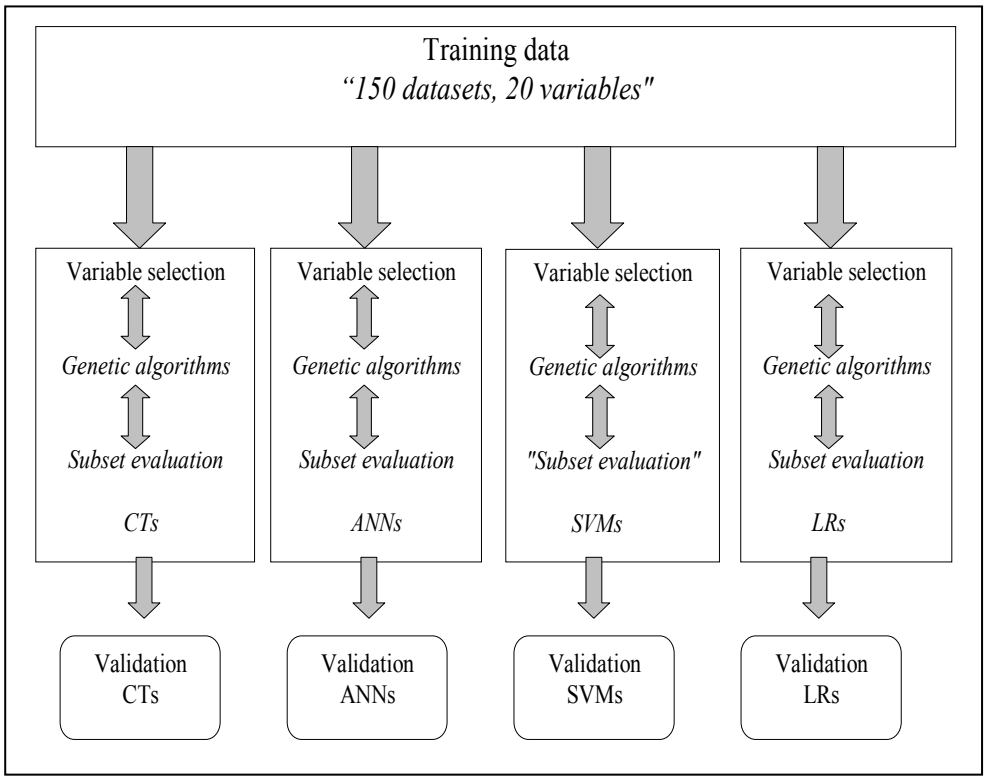


Figure 8.2. Modelling setup of the developed input variable selection scheme for the 4 models.

The variables of the 4 modelling techniques selected by genetic algorithms are presented in Table 8.2. An important decrease in the number of selected variables was noticed when applying genetic algorithms. For instance, the number of variables dropped from a total of 20 to 9 for classification trees and to 8 for artificial neural networks. Approximately 50 % of the input variables also were removed for support-vector machines and logistic regressions. Table 8.3 presents the variables which were selected for each model. For all models, the 3 structural-habitat variables distance from the source, slope and wetted-width seemed to have ecological

relevance for the prediction of the presence and absence of pike. Another habitat variable (average-depth) was selected 3 times, but this variable was not important for classification trees. From ecological point of view, the selected habitat variables for the prediction of pike can be considered relevant since the impact of these variables on pike has long been studied (Nilson, 2008). Also 3 pollution variables (ortho-phosphate, nitrite and BOD) were always selected for the prediction of pike in the rivers. Conductivity and total phosphorus were among the pollution variables that were selected three times. These variables also were not important for classification trees after variable selection. All models also confirmed that the Huet-zonation can have an important effect on the prediction of pike occurrence in the Flemish rivers since this river characteristic is related to gradient slope and width. Suspended-solids were the only variable that was selected two times for classification trees and support-vector machines. Finally, the variables being only once selected for the prediction of pike occurrence were Belgian Biotic Index (BBI) and ammonium. These two variables were important for ANNs and LRs, respectively. According to the 4 models, the pH, chloride, COD, dissolved oxygen, nitrate and flow velocity were never distinguished as important variables for the prediction of habitat suitability of pike. These are striking results, since in the literature on pike modelling one can find that they have been widely used for making habitat suitability rules in different pike's life stages (i.e. Kerle et al., 2001). This is most probably related to correlation of these variables with other variables distance from the source, slope, ... In total, 14 out of 20 predictors were chosen for the 4 modelling techniques by genetic algorithms.

Table 8.2. Comparison of the selected variables by GAs with CTs, ANNs, SVMs and LRs as an evaluation of pike occurrence in 6 river basins in Flanders during the years 1991-2002.

Input variables	***CTs***	***ANNs***	***SVMs***	***LRs***
Slope	*	*	*	*
Average-depth		*	*	*
Wetted-width	*	*	*	*
Distance from the source	*	*	*	*
Flow velocity				
Conductivity		*	*	*
Dissolved oxygen				
pH				
Nitrate				
Nitrite	*	*	*	*
Ammonium				*
Ortho-phosphate	*	*	*	*
Total phosphorus	*	*		*
Suspended-solids	*		*	
Chloride				
COD				
BOD	*	*	*	*
BBI		*		
Land-use		*	*	*
Huet-zonation	*	*	*	*

Table 8.3. Most important variables selected by the genetic algorithms with the 4 models for pike based on presence and absence data in 6 river basins in Flanders during the year 1991-2001.

Number of repeated variables in all models			
4 times	**3 times**	**2 times**	**1 time**
Slope Distance from the source Wetted-width Ortho-phosphate Nitrite BOD Huet-zonation	Average-depth Conductivity Total phosphorus Land-use	Suspended-solids	BBI Ammonium

Table 8.4 and Figure 8.3 compare the predictive results of the 4 models before and after the variable selection by genetic algorithms. As explained in chapters 4-7, all applied settings of 4 models were kept as default as possible before and after the variables selection. After the

variables were selected by genetic algorithms, the 4 models were again evaluated in terms of the two model performances criteria. Compared to the other modelling techniques (as illustrated in Table 8.4), the predictive results of logistic regressions were identical with previous chapters because no optimisation efforts were done for the logistic regressions.

After the less important variables were removed, the average model performance increased for each model. Both CCI and Kappa were more or less acceptable for artificial neural networks (CCI 70 % and Kappa 0.40), but the large standard deviation showed that there was instability between the subsets. For support-vector machines, the average CCI and Kappa were in the neighbourhood of the threshold values (based on CCI 67 % and Kappa 0.32), but the high standard deviation indicated that the models were not stable over the three subsets.

For classification trees and logistic regressions, the model performance was low. In particular, removal of variables in logistic regressions was negligible so that there was an increase of CCI and Kappa only by 2-3 %. In conclusion, after the variable selection stage, the average predictive performance of all models increased, but they still were not able to give a satisfying prediction for pike occurrence.

Table 8.4. Comparison of model performance for the 4 models after and before the variable selection stage in the river basins in Flanders (No GAs = before the variable selection by genetic algorithms, GAs = after the variable selection by genetic algorithms).

	CTs		***ANNs***		***SVMs***		***LRs***	
CCI	**No GAs**	**GAs**	**No GAs**	**GAs**	**No GAs**	**GAs**	**No GAs**	**GAs**
SS1	60	72	82	88	60	84	60	62
SS2	54	60	54	62	56	60	52	56
SS3	52	54	52	60	50	56	52	56
SS Average	55	62	63	70	55	67	55	58
SS Stdev	3	9	17	16	5	15	5	4
Kappa	**No GAs**	**GAs**	**No GAs**	**GAs**	**No GAs**	**GAs**	**No GAs**	**GAs**
SS1	0.20	0.44	0.64	0.76	0.20	0.64	0.20	0.24
SS2	0.08	0.20	0.08	0.24	0.12	0.20	0.04	0.12
SS3	0.04	0.08	0.04	0.20	0.00	0.12	0.04	0.12
SS Average	0.11	0.24	0.25	0.40	0.11	0.32	0.09	0.16
SS Stdev	0.07	0.18	0.34	0.31	0.10	0.28	0.07	0.07

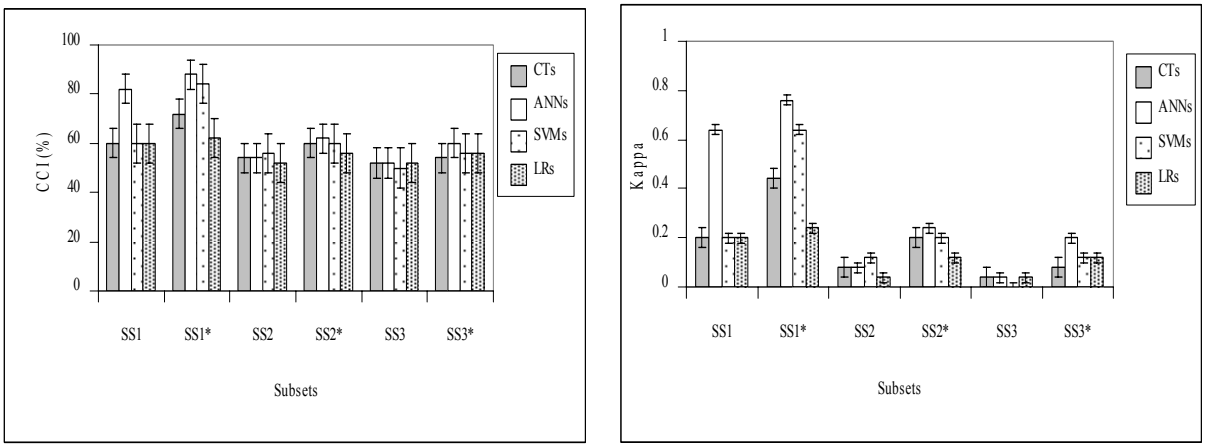


*Figure 8.3. Comparison of model performances of the 4 models before and after the variable selection stage (*SS = *subsets and* SS* = *the subsets after the variable selection).*

8.3 Discussion

A broad range of habitat suitability models has recently been developed to examine the relationships between habitats and species. Essentially, the physical habitat of a species relies on more than one variable and a combination of several suitability indices is needed to define a composite index (Ahmadi-Nedushan et al., 2006). Ecological modelling dealing with habitat requirements and prediction of species is considered as a helpful tool to support decision-making in river restoration management (Goethals and De Pauw, 2001). These models are

considered as supportive machine learning techniques for predicting the absence/presence of organisms in freshwater ecosystems.

Until now, there are only a few comprehensive empirical studies to compare learning algorithms. Learning algorithms are used in many domains and different performance metrics are appropriate for each domain. According to Caruana and Niculescu-Mizil (2006), although some methods clearly perform better than others, there is significant variability across the problems and metrics. Even, the best models sometimes perform poorly, and models with poor average performance occasionally perform exceptionally well.

In the present study, the advantages/disadvantages of several modelling techniques were confirmed. For example, one can find in the literature that models like support-vector machines and boosting and bagging achieve excellent performance. Feed-forward neural networks show the best performance (Caruana and Niculescu-Mizil, 2006; Kurt et al., 2008). Drake et al. (2006) explored the accuracy and reliability of ecological niche models with support-vector machines. The authors stated that the SVMs perform comparably well and are superior to other methods where only moderate amounts of data are available, while avoiding common problems and limitations. According to Tax and Duin (2004), methods based on support-vector machines have been particularly successful in applications with datasets with a large set of variables. These methods are more stable, require less model tuning and have fewer parameters than other computational optimisation methods such as neural networks. Among many machine-learning methods, they have several appealing characteristics for modellers (Guoa et al., 2005).

Artificial neural networks (ANNs) are better suited for analysis of nonlinear relationships between species distribution and environmental variables. Another advantage of artificial neural networks is that they give good performance. Logistic regressions are a popular method often used for modelling species and their relationships with environment (Ahmadi-Nedushan et al., 2006). As mentioned already, these methods are parametric for prediction and classification, but their performance depends on the distribution of variables, size of the datasets and quality of the data. When more predictors are nominal or responses are nonlinear, predictions will be confusing and will lack confidence (Sadat-Hashemi et al., 2005). Neural networks can be considered as an alternative technique to overcome these problems, because

training will not affect the results (Sadat-Hashemi et al., 2005). Logistic regressions have been effectively used to predict distributions of species and communities. These methods (Keating, 2004) and the other models based on probability densities (Robertson et al., 2001) illustrate the relative frequency of habitat utilization. They are therefore more strongly related to resource utilization or resource selection. According to Keating (2004), logistic regressions are important tools for wildlife habitat-selection studies, but the methods frequently have not been applied because they are difficult to interpret and due to the influence of sampling design.

Classification tree models, on the other hand, are powerful methods for modelling complex datasets (Breiman et al., 1984). Dzeroski et al. (2000) showed the predictive power of classification and regression trees to predict chemical parameters of river water quality from bioindicator data. Dzeroski (2001) used regression trees to model the growth of a dominant species of algae (*Ulva rigida*) in relation to some physical-chemical variables.

Until now, not many studies seem to have been carried out using different modelling techniques predicting the habitat suitability of pike. In the present research, a combination of 4 ecological modelling techniques (evaluated with two model performances) was compared in order to predict the habitat suitability of pike in 6 Flemish river basins. The outcomes obtained for the prediction of pike habitat proved the power of the SVMs and in particular ANNs to analyse nonlinear relationships in ecosystems as the advantages of ANNs have been explored in many ecological studies (e.g. Lek et al., 1996; Brosse et al., 1999; Dedecker et al., 2002; Dedecker et al., 2005).

As can be derived from the obtained results, by using a 3 fold cross-validations and evaluation of predictive performance based on two methods, the overall prediction of pike differed based on the modelling techniques. The SVMs and ANNs achieved better results than the two others. For getting more insight and a better interpretation, the integration of the two predictive performances seemed to yield better results than separately. However, many different methods applied to evaluate the performance make the comparison between studies very difficult. The practical applicability of the models for decision support is also a crucial qualitative aspect.

Variable selection stage (as applied by genetic algorithms in the present research) can be an essential factor either for improving model performance or for policy and management goals (D'heygere et al., 2003). These automated methods can trace these key variables. They could also serve for data collection optimisation, because some variables can be irrelevant for the particular purposes. For instance, in all models developed in this study, an important decrease in the number of selected variables was detected when applying genetic algorithms. According to most models, the structural-habitat variables (distance from the source, slope, wetted-width and average-depth) were important predictors for the pike's habitat suitability. The importance of some of these physical-habitat variables for pike was already confirmed in the work of Kerle et al. (2001). Besides, water quality variables (e.g. ortho-phosphate, nitrite and BOD) seemed to be crucial for the prediction of pike in the rivers. All models also confirmed that the Huet-zonation can have an important effect on the prediction of pike occurrence in the Flemish rivers. This was in line with the study of Huet (1949, 1954) and Breine et al. (2004). An evaluation on the model performance criteria showed that after genetic algorithms were integrated with 4 models, the average predictive results of all models increased. This explicitly shows the importance of variable selection stages by genetic algorithms. In this context, artificial neural networks and support-vector machines presented relatively high predictive performances when these are combined with GA, most probably because these techniques allow the highest degree of freedom (but also risk of overtraining) to fit data.

8.4 Conclusion

The present chapter aimed to compare different predictive models for pike in 6 river basins in Flanders. Particularly, the attention also was paid to the ecological relevance of these models as well as their performance in practical river management applications. The selection of input variables introduced in the modelling procedures is an important element for this type of approach (Fausch et al., 1988; Brosse et al., 1999).

The back-propagation of ANNs and SVMs constituted more efficient tools than the other modelling techniques to predict the pike's habitat suitability from the environmental characteristics of the river basins in Flanders because these methods had better predictive power. Classification trees and logistic regressions had less predictive power as the ANNs and

SVMs. The overall average CCI for all modelling techniques ranged from 55 % to 66 % and the range of Kappa from 0.09 to 0.33.

After the variable selection procedure, the number of variables could be reduced by 8 to 11, depending on the modelling techniques. In total, 14 out of 20 variables were selected by genetic algorithms for the 4 models. Among these variables, 3 structural-habitat variables, 3 pollution variables and Huet-zonation were more often selected for predicting the presence and absence of pike since they were selected for all modelling techniques. From ecological point of view, most models showed the dependence of pike habitat on the structural-habitat variables in the rivers. Moreover, water quality variables that can restrict the survival of pike in the sampling sites also contributed to the prediction of pike. Huet-zonation was an important river characteristic that could play a key role for the prediction of pike occurrence in the Flemish rivers. After the less important variables were removed, model performances showed an increase for each model (e.g. from CCI 58 % to 70 % and Kappa from 0.16 to 0.40). In spite of this, the models were still not possible to predict pike in a highly reliable manner. However, one has to consider the complexity of conditions that affect the presence of this top predator and also the lack of some key variables such as vegetation.

Chapter 9

Application of predictive models for decision support in river management

9.1 Introduction

In river basin management, sustainability aims at considering and integrating three main aspects: ecological, economic and social development. The sustainability of ecological or natural resources such as physical habitat and water quality and quantity status entails that the environmental conditions should be maintained or improved. The sustainability from economical point of view demonstrates that an equilibrium should exist between the value of a resource and the costs of its development. The sustainability from social aspects implies maintaining a human society that provides a quality of life for current and future generation. When making a proper decision for river restoration, the complex interaction between physical, chemical and biological components should be determined. The achievement of river restoration is dependent on steering the suitable key factor(s), which vary from river to river and from site to site. However, theories on river ecology are complex and not easy to use in the practise of the stream management (Verdonschot and Nijboer, 2002). For a successful restoration action at a site, river managers require a decision support system as simple as possible to cope with the ecological complexity. Application of ecological modelling techniques enables decision makers or stakeholders to predict aquatic communities on the basis of the results of the water quality simulations together with habitat data. These applications can thus be very useful to demonstrate their practical use and also to make simulation exercises relevant for the implementation of the European Water Framework Directive in Flanders.

The main objective of this chapter was to illustrate and validate the practical applications of the data driven habitat suitability models to support decision making in river management. This was done for the Dender basin in Flanders. At first, an overview is considered regarding the site description and restoration options of the catchment of river Dender. Second, the applied models are presented for simulation exercises. As demonstrated in Chapter 2, for river restoration management, the priority was given to the classification trees since the outcomes of the method can be directly read from these tree models. To fulfil this, the pike's habitat suitability models were considered based on presence/absence data of an example site in the Dender basin over 20 years (1989-2008).

9.2 Site description and restoration options in the Dender basin

As described in Chapter 2, Flanders has several main river basins of which 6 basins were monitored in this study. One of these is the Dender basin with a total area of 1384 km^2. It has an average discharge of 10 m^3/s at its mouth (Vandenberghe et al., 2005b). Since roughly 90 % of the flow results from storm runoff and the sources have a relatively small contribution, the Dender has a very irregular flow velocity with high peak discharges during intensive rain events and very low discharges during dry periods. However, to enable boat transport, this river is canalized and regulated by 14 sluices. Figure 9.1 illustrates one of these sluices. Sluices have a significant impact on the habitat characteristics and indirectly also influence the water quality. During dry periods, the Dender functions as a series of reservoirs with a (mid)depth of 3-5 m, a width of 12-50 m and a length of 2-8 km. During the periods of high flows, all locks are opened and the river regains a more natural stream profile (Vandenberghe et al., 2005b).

Figure 9.1. Illustration of the Dender river basin. This picture shows one of the several sluices installed for water quantity control and boat transport.

For the last few years, intensive research on pollution has been conducted in the Dender, which demonstrates a high variability of the pollution loads in this river (Vandenberghe et al., 2005a). D'heygere et al. (2002), on the other hand, made additional biological measurements during this period of on-line monitoring.

There are various types of direct discharges originating from industry, treated and untreated sewage from households. In addition to this, nutrient releases coming from agricultural activities can intensify this problem. The high nutrient loads can cause severe algae blooms during summer, leading to complex diurnal processes. Furthermore, many structural habitat modifications were established to alleviate flood control and guarantee boat traffic (Figure 9.1). These modifications can have a rigorous impact on the habitat characteristics and can result in a completely different fish community in comparison with the natural conditions. These problems led to the development of various models for the Dender river: e.g. the ESWAT model (Vandenberghe et al., 2005a). Here, 17 sampling places were monitored on the Dender between 1999 and 2000. Three of those measurement places were respectively in Aalst, Pollare and Denderbelle. The ESWAT model, which stands for "Extended Soil and Water Assessment Tool", can simulate several water quality characteristics in any combination desired by the user. This model was developed to allow an integral modelling of the water quantity and quality process in the basin. Based on the model, several important variables were considered for the catchment of the Dender such as dissolved oxygen, biological oxygen demand, nitrate, ammonium and so on, which are related to eutrophication problems and sewage discharges.

9.3 Application of the models for simulations related to the impact of land use changes on pike in the Dender basin

9.3.1 Simulations based on actual river conditions

Figure 9.2 shows the sampling site along the Dender, which is one of the sites, situated downstream the city of Geraardsbergen. In this site, several physical-chemical and structural-habitat variables were measured during the period 1989-2008, however in Table 9.1, only the variables necessary for the tree models are presented for simulation exercises in the Dender basin. In total, 11 input variables were used for the simulation exercises, which were a

combination of 4 structural-habitat variables (flow velocity, slope, average-depth and wetted-width), 4 pollution variables (chloride, COD, BOD and total phosphorus) and the other 3 predictors were land-use, Huet-zonation and BBI. The values of the actual conditions of the 4 structural-habitat variables were supposed to be similar over the 20 years of monitoring: flow velocity = 0.1 cm/s, slope = 0.4 %, average depth = 2.5 m and wetted-width = 22 m. The target site was located in the bream zone and the actual type of land-use was pasture. The data listed in Table 9.1 are merely used for the model simulations not model development. Data on pike presence/absence were not available in the site of Dender for simulations exercises during 20 years (1989-2008).

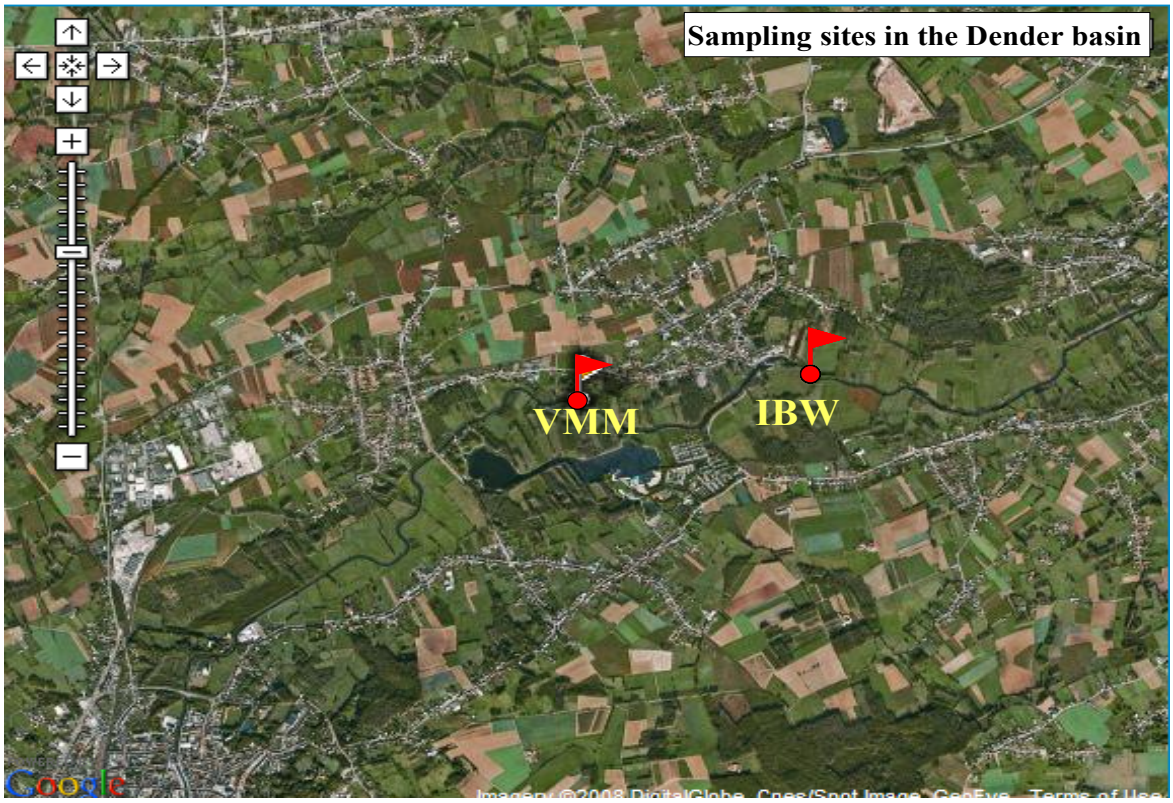


Figure 9.2. The Dender basin with indication of the sampling sites. These sites are situated downstream the city of Geraardsbergen (sites related to chemical water analysis and fish/physical habitat sampling stations are indicated as VMM and IBW respectively).

Table 9.1. Actual conditions of structural-habitat and physical-chemical variables, which were used to simulate the habitat suitability of pike based on presence/absence data in the Dender basin during the period (1989-2008) (? indicates missing data).

Date	Land-use	Huet-zonation	BBI	Flow* Velocity (cm/s)	Slope (%)*	Average-depth (m)*	Wetted-width (m)*	Chloride (mg/l)	COD (mg/l)	BOD (mg/l)	Total phosphorus (mg/l)
1989	Pasture	Bream	1	0.1	0.4	2.5	22	132.88	48.38	?	2.54
1990	Pasture	Bream	2	0.1	0.4	2.5	22	138.67	51.78	?	3.76
1991	Pasture	Bream	?	0.1	0.4	2.5	22	99.50	34.75	?	2.01
1992	Pasture	Bream	5	0.1	0.4	2.5	22	75.88	24.00	?	0.81
1993	Pasture	Bream	5	0.1	0.4	2.5	22	81.31	29.00	?	1.08
1994	Pasture	Bream	5	0.1	0.4	2.5	22	60.29	24.29	?	0.48
1995	Pasture	Bream	4	0.1	0.4	2.5	22	69.12	25.12	?	0.65
1996	Pasture	Bream	6	0.1	0.4	2.5	22	92.17	27.33	?	0.84
1997	Pasture	Bream	6	0.1	0.4	2.5	22	99.62	27.88	?	0.54
1998	Pasture	Bream	3	0.1	0.4	2.5	22	70.45	22.90	5.00	0.92
1999	Pasture	Bream	6	0.1	0.4	2.5	22	68.08	28.83	4.25	0.57
2000	Pasture	Bream	6	0.1	0.4	2.5	22	56.81	36.45	4.67	0.66
2001	Pasture	Bream	6	0.1	0.4	2.5	22	50.31	18.79	3.93	0.51
2002	Pasture	Bream	5	0.1	0.4	2.5	22	57.18	23.92	3.98	0.43
2003	Pasture	Bream	6	0.1	0.4	2.5	22	65.83	25.51	4.31	0.57
2004	Pasture	Bream	5	0.1	0.4	2.5	22	80.92	29.58	5.13	0.63
2005	Pasture	Bream	6	0.1	0.4	2.5	22	89.75	26.92	1.83	0.38
2006	Pasture	Bream	5	0.1	0.4	2.5	22	86.33	30.08	3.05	0.20
2007	Pasture	Bream	?	0.1	0.4	2.5	22	88.08	23.58	3.72	0.51
2008	Pasture	Bream	?	0.1	0.4	2.5	22	70.38	22.00	3.17	0.31

(*) Data based on the supposition that the conditions remained constant over the years.

9.3.2 Simulation exercise based on land use changes

The aim of river management is to focus on those actions providing effective results obtained by the applied models. All 4 models used for the prediction of pike occurrence (in the previous chapters) could have been applied for the restoration actions of the 6 river basins, but as already stated, the main focus was paid to the Dender river. The applied methods were based on classification trees. For simulation exercise in this river, 3 subsets with 4 pruning confidence factors were applied (the details are presented in the Annexes 1-12). In that way, the models could deliver insight for decision makers. The developed trees revealed that the land-use was the key factor influencing the ecological status of pike occurrence in the target site, and therefore, the effect of land use changes was studied. In particular the effect of modifying the pasture to arable land and forested areas was simulated.

The results obtained with genetic algorithms also showed that for majority of modelling techniques, the type of land-use can be a key predictor for the prediction of pike occurrence in rivers (3 out of 4 models confirmed this variable). In Tables 9.2, the outcomes of the predicted pike conditions made on the Dender river (at the sampling site downstream of the wastewater treatment plant of Geraardsbergen) are presented during the years 1989-2008. This table illustrates what the potential effects of a change of type of land-use could have meant for the pike population over this time period. Looking at the average predicted results, one can see that arable-land can lead to an extra stress for pike, resulting in its absence in the period between 1989 and 1990. Nevertheless, it seems that this period is not so relevant for land-use simulations, since in these years extremely poor water quality in terms of Belgian Biotic Index (BBI = 1 and 2) and BOD probably played a major role for pike absence. The models tend to search for the major variables for whole period of the available data, and probably these first years of simulations would need other more specific models for that period.

Table 9.2. Predicted pike presence/absence values according to the classification trees with 4 different pruning confidence factors in the Dender river basin during the period 1989-2008 (in brackets the amount of folds out of a total of 3 that supports the outcomes, P *= Present,* A *= Absent).*

Date	**Predicted pike conditions (Pasture)**					**Predicted pike conditions (Forest)**					**Predicted pike conditions (Arable-land)**				
	PCFs					**PCFs**					**PCFs**				
	0.5	**0.25**	**0.1**	**0.01**	**Average**	**0.5**	**0.25**	**0.1**	**0.01**	**Average**	**0.5**	**0.25**	**0.1**	**0.01**	**Average**
1989	P (2/3)	P (2/3)	P (2/3)	A (2/3)	P (7/12)	P (2/3)	P (2/3)	P (2/3)	A (2/3)	P (7/12)	P (2/3)	A (2/3)	A (2/3)	A (2/3)	A (7/12)
1990	P (2/3)	P (2/3)	P (2/3)	A (2/3)	P (7/12)	P (2/3)	P (2/3)	P (2/3)	A (2/3)	P (7/12)	P (2/3)	A (2/3)	A (2/3)	A (2/3)	A (7/12)
1991	P (3/3)	P (3/3)	P (3/3)	P (2/3)	P (11/12)	P (3/3)	P (3/3)	P (3/3)	P (2/3)	P (11/12)	P (3/3)	P (2/3)	P (2/3)	?	P (7/12)
1992	P (3/3)	P (3/3)	P (3/3)	P (2/3)	P (11/12)	P (3/3)	P (3/3)	P (3/3)	P (2/3)	P (11/12)	P (3/3)	P (2/3)	P (2/3)	P (2/3)	P (9/12)
1993	P (3/3)	P (3/3)	P (3/3)	P (2/3)	P (11/12)	P (3/3)	P (3/3)	P (3/3)	P (2/3)	P (11/12)	P (3/3)	P (2/3)	P (2/3)	P (2/3)	P (9/12)
1994	P (3/3)	P (3/3)	P (3/3)	P (2/3)	P (11/12)	P (3/3)	P (3/3)	P (3/3)	P (2/3)	P (11/12)	P (3/3)	P (2/3)	P (2/3)	P (2/3)	P (9/12)
1995	P (3/3)	P (3/3)	P (3/3)	P (2/3)	P (11/12)	P (3/3)	P (3/3)	P (3/3)	P (2/3)	P (11/12)	P (3/3)	P (2/3)	P (2/3)	P (2/3)	P (9/12)
1996	P (3/3)	P (3/3)	P (3/3)	P (2/3)	P (11/12)	P (3/3)	P (3/3)	P (3/3)	P (2/3)	P (11/12)	P (3/3)	P (2/3)	P (2/3)	P (2/3)	P (9/12)
1997	P (3/3)	P (3/3)	P (3/3)	P (2/3)	P (11/12)	P (3/3)	P (3/3)	P (3/3)	P (2/3)	P (11/12)	P (3/3)	P (2/3)	P (2/3)	P (2/3)	P (9/12)
1998	P (3/3)	P (3/3)	P (3/3)	P (2/3)	P (11/12)	P (3/3)	P (3/3)	P (3/3)	P (2/3)	P (11/12)	P (3/3)	P (2/3)	P (2/3)	P (2/3)	P (9/12)
1999	P (3/3)	P (3/3)	P (3/3)	P (2/3)	P (11/12)	P (3/3)	P (3/3)	P (3/3)	P (2/3)	P (11/12)	P (3/3)	P (2/3)	P (2/3)	P (2/3)	P (9/12)
2000	P (3/3)	P (3/3)	P (3/3)	P (2/3)	P (11/12)	P (2/3)	P (2/3)	P (2/3)	P (2/3)	P (8/12)	P (3/3)	P (2/3)	P (2/3)	P (2/3)	P (9/12)
2001	P (3/3)	P (3/3)	P (3/3)	P (2/3)	P (11/12)	P (2/3)	P (2/3)	P (2/3)	P (2/3)	P (8/12)	P (3/3)	P (2/3)	P (2/3)	P (2/3)	P (9/12)
2002	P (3/3)	P (3/3)	P (3/3)	P (2/3)	P (11/12)	P (2/3)	P (2/3)	P (2/3)	P (2/3)	P (8/12)	P (3/3)	P (2/3)	P (2/3)	P (2/3)	P (9/12)
2003	P (3/3)	P (3/3)	P (3/3)	P (2/3)	P (11/12)	P (3/3)	P (3/3)	P (3/3)	P (2/3)	P (11/12)	P (3/3)	P (2/3)	P (2/3)	P (2/3)	P (9/12)
2004	P (3/3)	P (3/3)	P (3/3)	P (2/3)	P (11/12)	P (3/3)	P (3/3)	P (3/3)	P (2/3)	P (11/12)	P (3/3)	P (2/3)	P (2/3)	P (2/3)	P (9/12)
2005	P (3/3)	P (3/3)	P (3/3)	P (2/3)	P (11/12)	P (3/3)	P (3/3)	P (3/3)	P (2/3)	P (11/12)	P (3/3)	P (2/3)	P (2/3)	P (2/3)	P (9/12)
2006	P (3/3)	P (3/3)	P (3/3)	P (2/3)	P (11/12)	P (3/3)	P (3/3)	P (3/3)	P (2/3)	P (11/12)	P (3/3)	P (2/3)	P (2/3)	P (2/3)	P (9/12)
2007	P (3/3)	P (3/3)	P (3/3)	P (2/3)	P (11/12)	P (3/3)	P (3/3)	P (3/3)	P (2/3)	P (11/12)	P (3/3)	P (2/3)	P (2/3)	P (2/3)	P (9/12)
2008	P (3/3)	P (3/3)	P (3/3)	P (2/3)	P (11/12)	P (3/3)	P (3/3)	P (3/3)	P (2/3)	P (11/12)	P (3/3)	P (2/3)	P (2/3)	P (2/3)	P (9/12)

9.4 Discussion

The modification and restoration of rivers has become a common practice in the management of natural resources (Junker and Buchecker, 2008). Based on the data driven models applied in the present work, calculating the effect of land use change on aquatic ecosystems was illustrated. The simulation exercises for the Dender revealed that in order to establish a river restoration plan for the Dender basin, the identification of current and future land uses can be of great importance (but amongst others also depending on the water quality of the system).

Based on this information, environmental objectives can be effectively selected to guarantee the sustainability of the environment to support the target functions. These targets have to be compared with the current status of the aquatic ecosystem and recognizing possible environmental problems which are resulting from these activities. Therefore, management options that intend to improve or solve these problems have to be selected and implemented. However, it has to be noted that restoration project proposals may fail as a result of lack of public support. Therefore, it will be of scientific and practical value to pay more attention to the factors that affect people's evaluations of the aesthetics of river restoration management (Junker and Buchecker, 2008).

Validation of the models for making simulation exercises is not easy. The studies related to these exercises for river restoration management need more models concerning other aspects of biological communities to make a complete evaluation of the overall ecological effects. In this context, other fish species and their interactions in an ecosystem (e.g. from food web perspectives also) need to be considered. Besides, the combination of habitat with water quantity and quality models is compulsory. The coupling of models is also important to get insight in the interactions occurring due to changing habitat characteristics or pollution levels. Determining the environmental input variables of the rivers, is difficult merely based on available (scarce) data, and model coupling is therefore necessary. In particular the interaction between variables can be included via integrated models (e.g. the effect of land use change on BOD, nutrients, ... was in the present study not considered).

The data driven models (such as the used classification trees for the Dender) can deliver ecological suitable information, but the stability of models is very important for river

restoration management. In other words, when the developed models have instability in terms of input variables or parameter settings (as was encountered for most modelling techniques in the previous chapters), it is difficult to make a proper decision for river restoration management. When the applied models were tested with different pruning confidence factors, the illustrated effects of the land use changes only slightly altered the overall prediction of presence/absence of pike.

Further research is required to find out and solve various sorts of problems with regard to the number of instances, outliers, parameter settings and modelling techniques themselves etc. In that way, they can help to find the major gaps in our knowledge of river systems and help to set up cost effective monitoring programmes to improve the models.

9.5 Conclusion

The simulation exercises at the site of Geraardsbergen in the Dender basin demonstrated that the type of land-use is an important predictor of pike occurrence. The simulation exercises illustrated that a shift from pasture to arable-land lead to an extra stress for pike, leading to its absence in the period close to 1990, whereas the modification to forest probably would not lead to much better conditions. Nevertheless, further studies are required to identify whether other types of land use (e.g. urban and industry) can influence the ecological quality of the river basins or not. In particular, in the case of industry, no data were available.

Before the models can be practically used for decision support in river management, they need to be improved and become more sensitive for other input variables (e.g. explicit relations with BOD, nutrients, ... could be interesting to simulate more scenarios such as the installation of wastewater treatment facilities and buffer zones). The developed models have the ability to give ecological insight and hence, they can predict habitat suitability for pike. However, to achieve the optimal restoration actions, the models should be assessed in terms of various issues. For instance, stability of the models needs to be increased and also the number of input variables has to be expanded. The selection of suitable parameter settings can also serve in this respect. The coupling of water quantity and quality with habitat models is of great importance as well.

General discussion, conclusions and further research

Introduction

This thesis aimed to develop and optimise data driven models to predict the habitat suitability of pike in Flemish river basins. For this, four model development techniques were applied in combination with genetic algorithms. All modelling techniques were based on 21 variables (20 input and 1 output variables) and a set of 150 observations collected in 110 sampling sites in six river basins in Flanders during the years 1991-2002.

The aim of this chapter is to link the results and discussions in the previous chapters and raise some general and practical issues regarding the development and application of predictive models for decision support in water management. At the end of this chapter, some recommendations and considerations are given for further research in this field.

This chapter consists of the following components:
- collection of data related to river conditions;
- towards (integrated) data management for model development and application;
- model development;
- model application for decision support in water management;
- further research, conclusions and recommendations.

Collection of data related to river conditions

Selection of proper model input variables is crucial for models dealing with prediction and forecasting (Faraway and Chatfiled, 1998). Identification of the suitable environmental input variables is an important but complex issue (Soulie, 1994). Artificial intelligence techniques are highly dependent on the specification of input variables (Pakath and Zaveri, 1995). Therefore, the first step is to get acquainted with the type of the variables which will be considered in the field (or existing databases, cf. this research). When selecting the input variables, two major problems may occur: too many or too little environmental variables are available to develop a relevant relation between inputs and outputs (respectively environmental river conditions and habitat suitability of pike in this research). However, in many cases the latter can be solved based on expert knowledge or via feature (variable) selection methods.

In this study, a major problem was the limited set of variables that were available in the fish database. For several decades, various variables and fish community data have been monitored in the river basins of Flanders. However, abiotic river conditions were often not recorded, or not in a standardized manner. By coupling this dataset to a dataset consisting of chemical data of the Flemish Environment Agency, 20 river characteristics (one variable was actually the land use in the vicinity of the river) could be finally used in the present work to predict the habitat suitability of pike as discussed in Chapter 3. These variables consisted of a combination of physical-chemical, structural-habitat and three 'integrating' variables. Among the latter, the land use played an important role in the input of nutrients and habitat deterioration. The results in the previous chapters confirmed the relationship between the structural characteristics (physical habitat) and presence/absence of pike in the rivers. For the last few years, the water quality in Flanders has improved due to investment in sewerage and wastewater treatment plants (VMM, 2003), nevertheless, the water quality is still a limiting factor in several streams in Flanders.

The results from previous chapters showed that some relevant variables were missing (i.e. water temperature, vegetation cover…) and that more standardisation and quality control is required to warrant a suitable dataset for data driven model development. Careful attention should be paid to recognize which variables are relevant and which ones can be neglected (Auger et al., 2000). Loading a great number of attributes to data driven models (i.e. artificial neural networks) increases the network size without the guarantee for an added value of the predictive relevance of the model. Therefore, it is recommended to use only a limited number of variables to predict the output(s), for example based on optimisation results with feature selection methods such as genetic algorithms.

In addition to the need of extra variables, it is moreover important that these data cover a relevant range and that enough data (instances) are gathered. In other words, it is not always relevant to include new variables when datasets seem to be too small for model development. Pertaining to the range of the data, particularly in Flanders, there is a major lack of river ecosystems of good quality, which makes it difficult to develop well performing models for restoration options and prediction of reference conditions. Thus, to cope with these drawbacks, one possible solution is that more data from different river basins (e.g.

international data) should be gathered in the future to apply the modelling techniques described in the previous chapters.

In chapter 3, some theoretical explanations were given to exclude some variables because of high correlation coefficients between them or to remove instances due to outliers, but also some practical reasons were given to leave them in. In this research, all variables and instances were kept in. To test the model development techniques for theoretical (e.g. testing the added value of optimisation procedures like the genetic algorithms) and practical purposes (e.g. prediction the effects of river restoration scenarios based on models trained with raw data), all variables were kept in the dataset.

Pike monitoring was conducted in 6 main river basins. The related information was given in Chapter 2. Of 150 observations recorded in these river basins, the frequency of occurrence of pike was 50 %, in which pike was absent in 75 instances and present in 75 cases. Samples were not evenly distributed over the given river basins, ranging from 2 % (upper Scheldt) to 37 % (Demer). The same problem was encountered with the temporal patterns of pike monitoring in the sampling sites during the study period (1991-2002). More than two-third of pike monitoring took place in 1995, 1996, 1999 and 2001. In contrast, during the years 1991, 1992, 1994, 1998, 2000 and 2002, less monitoring data were available. No monitoring data were available of the years 1993 and 1997. The differences in monitoring ultimately could hamper to make a proper decision about the temporal and spatial patterns of pike in the sampling sites.

Towards (integrated) data management for model development and application

According to the European Union (EU) water policy, fish is an element that has to be evaluated to assess the ecological condition of water bodies (EU, 2000). When monitoring fish communities, certain factors, for instance fish species composition, abundance and age structure have to be included. Fish assemblage attributes precisely reflect the overall biotic integrity of aquatic ecosystems (Breine et al., 2004; Schmutz et al., 2007). Thus, it is essential to monitor fish assemblages with more caution (in particular pike populations due to their dynamic behaviour and dependence on the rest of the food web). Belpaire et al. (2000) developed a multimetric fish Index of Biotic Integrity (IBI) to assess the biotic integrity of

Flemish water bodies. The development of fish-based methods to assess human-induced impacts on aquatic ecosystems has been strongly influenced by the index of biotic integrity (e.g. Belpaire et al., 2000; Roset et al., 2007; Schmutz et al., 2007). The majority of the indices combine a reference condition approach and pertinent biological variables or metrics to illustrate the fish assemblage characteristics and to quantify the impact of human activities on the biota (Noble et al., 2007; Roset et al., 2007). However, the applied sampling techniques used for describing fish assemblage in river basins in Flanders vary according to research institutions and water types. In the current study, fish assemblage surveys in 110 sampling sites in Flanders were mainly carried out from 1991–2002 by the Institute of Forestry and Game and Management. The applied technique to collect fish assemblage data was electrofishing, using a 5 kW generator with an adjustable output voltage of 300 to 500 V and a pulse frequency of 480 Hz. The number of hand-held anodes used was 2, except when the river had a width of less than 1 m. Electrofishing was carried out in 3 Huet zones. This method has been broadly used in ecological fields for fish monitoring and a wide range of river types (e.g. Belpaire et al., 2000; Breine et al., 2004). Sometimes a supplementary method is used e.g. gill nets, fykes (90 cm diameter and 22 m long) and seine nets (variable sizes) (Belpaire et al., 2000), but these methods were not used in the present study. At each station, pike was counted and measured in terms of abundance, biomass and length but in the present work, only presence and absence data were considered for pike. Insufficient information was provided with regard to length and age classes of pike therefore these variables could not be used. Based on the models, Huet's typology (Huet, 1949, 1954) was often used to classify pike habitat suitability of river stretches. The gradient slope of all sampling sites ranged from 0.0 % to 6.0 % and the range of width was between 0.50 and 32 m. The 110 sampling sites belonged to the four different zones of Huet, but a simplified classification system (Breine et al., 2004) was here applied making use of only three zones: barbel, bream and upstream (trout and grayling) zones. The bream zone resulted in more monitoring data for pike, followed by the barbel zone, while for the upstream zone almost no data were available. This latter zone is in general also not a common place for pike to occur, due to the high flow velocities and absence of vegetation in most cases.

Though a large amount of data has been gathered in Flemish river basins, there are still some very practical gaps to be filled before the resulting databases meet the requirements of model development to support river managers. The datasets are distributed over different institutes in

Flanders using various format types and different location codes. This problem was also encountered in the present research. The data were gathered by two institutes namely, INBO (Institute for Nature and Forest Research, Brussels), and VMM (Flemish Environment Agency). Structural-habitat data for pike were gathered by INBO, while the chemical data were provided by VMM. During the year 1991-2002, more than 1000 electrofishing events took place in Flanders, however a set of 150 observations collected in 110 sampling sites of 6 Flemish river basins remained after removing the events with too many missing variables and aiming for a similar number of cases where pike was present or absent. Another drawback was that samples were not equally distributed over the 6 river basins and over the Huet zones.

Model development

Although modelling techniques can contribute significantly to the solution of a numerous of real-world problems, many studies confirm their limited adoption and use (Carter, 1987; Franz, 1989). Failure of modelling efforts is often related to user perception of model complexity and inadequate user involvement. Users find maintaining models hard and slight changes in the software often require time-consuming and expensive modifications by technical experts (Franz, 1989).

All models discussed in Chapters 4 to 8 were evaluated on the basis of both overall predictive success expressed by correctly classified instances (CCI) and Cohen's Kappa. CCI is the most commonly applied predictive performance criterion because of its easy interpretation for water managers and policy makers. Frequency of occurrence can however affect the prediction of an organism. For example, predictions of very common or very rare organisms can be misleading if only the overall prediction success is used as predictive performance. For this, the number of cases where pike was present was made the same as the absent cases. Moreover, another model evaluation criterion, namely Cohen's Kappa has widely been recommended in the literature (Fielding and Bell, 1997; Manel et al., 2001; Dedecker et al., 2004; Goethals, 2005) to assess presence/absence models. Therefore this is also applied to evaluate the 4 models in the present work. Since prevalence of pike was 50 %, both model performance criteria gave reasonable and easy-interpretable results. The developed models (before applying a variable selection) did not yield a very good predictive power. After variable selection, the predictive power improved for all models but they still did not meet the

threshold values (CCI > 70 % and Kappa > 0.4) for good and reliable models. Artificial neural networks, however, were in the vicinity of these threshold values. When comparing the overall results (after less informative variables were removed), it seemed that the artificial neural networks and support-vector machines were performing slightly better than the classification trees and logistic regressions. In spite of this, the ecological meaning of the classification trees can be directly seen, while it needs more efforts to analyse the ecological meaning of the ANNs models. For instance, in the case of classification trees developed in chapter 4, it was possible to find the ecological relevance of some structural-habitat and water quality variables for prediction of the pike's habitat suitability in the river basins of Flanders.

So far, there are still several uncertainties when developing data driven models. When data are prepared for data driven models, many problems can be encountered e.g. leaving outliers out or keeping them in. As a result, it is not easy to decide how to develop well performing data driven models based on a general set of rules. Small changes in the methodology can lead to different outcomes, in particular for artificial neural networks and support-vector machines. For artificial neural networks, it was not so easy to determine the best model architecture. Furthermore, the rules suggested for determining the optimal set of hidden neurons varies according to the used literature. Dedecker et al. (2004) compared different architectures and training methods, but the results were not compatible with a set of simple rules of thumb. In the literature, one can hardly find a specific rule for SVMs (Ma et al., 2008; Yang et al., 2008). In this study for optimisation of SVM models, however, the exponent approach was used, but one has to be careful because a higher exponent enables more complex data to be fit and hence also increases the complexity of the model.

Recently, the ensembles of decision trees generated using boosting (Geoffrey, 2000), and bagging (e.g. Dakou et al., 2006) and random forest (e.g. Breiman, 2001) have been shown to be very good all-round predictors because they require less parameter tuning than support-vector machines and artificial neural networks (Witten and Frank, 2000). Also other techniques (e.g. genetic algorithms), can lead to model improvements. This was already applied by D'heygere et al. (2003) for the development of classification trees and artificial neural network models for macroinvertebrates taxa in Flanders. Their relevance, even for small datasets, was an important result of this study. On the other hand, Bayesian Belief Networks (BBN) (Adriaenssens et al., 2004a) and fuzzy logic (Adriaenssens et al., 2004b) are

becoming popular techniques applied for ecological modelling since these models can be to a high extend based on expert knowledge. This could also be a relevant method to deal with missing variables to develop models. Nevertheless some data are necessary to be able to validate these knowledge based models. Also the practical relevance of models needs to be considered during the model development process. A possible option would be to involve stakeholders from different organizations e.g. river managers, mathematicians, applied informatics specialists, ecosystem scientists and data collectors (Goethals, 2005).

Model applications for decision support in water management

In recent years, various studies have been developed regarding ecological models predicting biological communities in aquatic ecosystems and these models are considered as valuable tools (e.g. Brosse et al., 1999; Brosse et al., 2007; Kruk et al., 2007). Among these models, one can find the model of River Invertebrate Prediction And Classification System (RIVPACS) as the first appealing model (Wright et al., 1993). Such assessment systems are able to predict the reference communities based on a set of local river features as a basis for the assessment.

Making a reliable model application (as applied in this study) is very important for decision support in water management purposes. These issues include an understandable description of the purposes, suitable data collection, analysis of the existing data, testing of the model structure and parameters, model validation etc. Various decision support tools have been recently developed for river basin management such as scenario simulation and modelling systems (known as DSS), expert systems, GIS applications and databases, visualisation, role plays and gaming (Welp, 2001). Due to the improvement of the reliability of management activities in Flanders, decision support systems could have a significant role in this context. When such systems are designed, careful considerations should be paid to the possibilities and limitations of the available water system models. The major problem is that there are not many modelling studies which are validated in practice and hence are not working appropriately and they are also not capable to give the information of interest to the managers. According to Van der Molen and Boers (2001), the credibility of models by water managers is principally defined by their perception of the practical value of the model. On the basis of this, to convince the river managers and stakeholders of the validity and practical usefulness of the

developed models in this thesis, some practical simulations were made in the Dender basin (Chapter 9). This was merely done by means of the induced classification trees during the periods 1989 till 2008. These scenario analyses can help to decide which restoration options to choose, and moreover mean a crucial validation of these models from a practical point of view. The simulation exercises showed that the type of land-use can play a key role for prediction of pike occurrence in particular when the activities are more directed towards arable-land in the river basins of Flanders. Based on this simulation exercise on the Dender, when water quality issues are combined with type of activities, pike populations are facing a combined stress. However, for the study of these impacts, integrated models are needed to be able to consider interactions between abiotic variables and the human activities, as a basis for a relevant interpretation of the outcome of the ecological models. Also further studies are needed to identify whether other types of land use (i.e. urban and industry) can influence the ecological quality of the river basins or not, since in this study in particular insufficient data were available for industry so it was not possible to draw a relevant conclusion in these simulation exercises.

There are several requirements for improvement of water management. This can be achieved either by habitat suitability models (as developed for pike in this research) or other more complex models. Besides, there is a need to make simulations of whole communities (e.g. fish communities), but also other models are essential to determine the input variables of these habitat suitability models, such water quantity and quality models. This type of practical studies also can give insights in how to improve the data collection for developing models that are of practical use for decision support in river management. One of the main problems for the model development process is how to engage the river managers (Goethals, 2005).

Recently, there has been a growing demand for effective public participation in river basin management. These are posed, among others, by Water Framework Directive and international conventions and policy documents. The demands will increase a need for guidance on how the public is to be engaged (Welp, 2001). Goethals (2005) demonstrated the potential link between ecological models and socio-economical models and stakeholder information requirements, since economic assessments can play a key role to analyse the costs and benefits of river restoration options within the policy area of water management. Particularly in Flanders, it would be crucial to combine the existing hydraulic and river quality

models and the ecological habitat suitability models. These models are able to provide valuable information regarding river restoration scenarios based on the environmental river characteristics used as input variables for the habitat suitability models. In particular, the effect of agricultural activities can be modelled (use of mineral fertilisers and applications of manure from livestock), which are considered as one of the main sources of nutrient input to European water bodies (European Environment Agency, 2002).

Further research, conclusions and recommendations

This research demonstrated that habitat suitability modelling by means of the developed models combined with genetic algorithms can help to better understand the relationship between environmental river characteristics and pike occurrence. Applying the pike's habitat models in this study can be practical and used for the development of other models such as food web and synecological (e.g. competition) models. These habitat suitability models can have ecological valuable outcomes concerning river restoration or deterioration decisions, hence they can support river management. However, many challenges and difficulties still have to be solved to develop well-performing models. In this research, difficulties regarding deficiency of data were obvious so that the need for good data and appropriate variables seemed to be one of the main constraints. Up to now, a large amount of data from the Flemish river basins has been gathered but these data do not meet the specific requirements for developing models. As a consequence, an applicable and update monitoring network has to be set up for the development of trustworthy and useful predictive habitat suitability models for pike or other fish species in Flanders. Increasing the number of instances and measuring more relevant variables for fish communities may enhance the reliability of the habitat suitability models. As demonstrated already, for pike only 150 samples and 21 variables (20 as inputs and one as output) could be used, while several variables and instances were inevitably eliminated due to too much missing values. As a result, for improving the reliability of habitat suitability, it would be of a great importance to include and measure more relevant variables such as vegetation, which seems to be crucial in the whole life stage of pike. In other words, it is essential that the dataset covers a broad enough range for all variables and enough instances are gathered in the future. However, measuring new variables without gathering enough data for developing a model may not work out per se. In Flanders, in particular, there is a major lack of river ecosystems with a good ecological status. This can lead to a substantial constraint

concerning developing relevant models to predict the effect of restoration options. Providing supplementary data for other river basins from international datasets could be a good alternative to improve the reliability of the models in Flanders. Nevertheless, several issues such as financial and time constraints can have negative effects on the decision making process.

In the recent years, a great number of automated methods have proved to be practical for the selection of the most appropriate variables. For that reason, such methods are widely recommended for predicting aquatic communities and that is why they were used in this PhD research. The development of habitat suitability models for pike in this present work or whole fish communities can be very useful for the implementation of the Water Framework Directive. For development, validation and training such models described in this study, the availability of sufficient data can be the main constraint in the future. Finally, integration of river quality modelling and hydraulic modelling with habitat suitability modelling can improve the reliability and the practical applicability of the habitat suitability models. On the basis of these, an essential future step should be a partnership between modellers, river managers and stakeholders. In summary, based on the results of the previous chapters and more specifically the problems encountered in 3 main parts of this research (monitoring, modelling and management), the following recommendations would be respectively applicable for these parts:

•In any ecological studies, monitoring is the first and fundamental step before model development. More variables (i.e. vegetation cover, age classes, ...) and sampling sites need to be considered for the improvement of pike prediction. Furthermore, to make a better prediction of pike over different years, the monitoring of biotic and abiotic variables should be more continuous.

•The second step of PhD research was allocated to the model development. Some drawbacks were encountered when developing and optimising these models. For that reason development of alternative modelling techniques would be of great importance for the prediction of pike habitat suitability in the river basins in the near future. For instance some of the alternative methods can be suggested such as the ensembles of decision trees (i.e. using boosting and bagging and random forest). Linking these ecological models to river water quality and

hydraulic models, etc in Flanders would lead to outcomes of which the causes and effects can be easier linked. Moreover, development of a good habitat suitability model is not so easy for this species, due to its changing preference of habitat during its lifecycle, and dependence on other fish species. This entails that more complex models need to be developed integrating synecological and food web aspects.

•The third and very crucial part of this research was related to the relevance of these models for river management. To make a better decision for river management in Flanders, other river basins (provided more data and variables are available) should also be simulated. To do so, the priorities can be set for pike stocking and restoration in the broader perspective of the Water Framework Directive (WFD). In this study, an important restoration zone for pike populations and their preys was bream zone so this zone calls for more protection. In the end, to achieve a better management, shallow areas and potentially vegetation rich zones need more attention since these areas are very essential for pike spawning.

References

Aadland, L. P. (1993). Stream habitat types: their fish assemblages and relationship to flow. *North American Journal of Fisheries Management* **13**: 790-806.

Aarts, B. G. W. and Nienhuis, P. H. (2003). Fish zonations and guilds as the basis for assessment of ecological integrity of large rivers. *Hydrobiologia* **500**: 157-178.

Adriaenssens, V. (2004). Knowledge-based macroinvertebrate habitat suitability models for use in ecological river management. PhD thesis, Ghent University, 296 pp.

Adriaenssens, V., Goethals, P. L. M., Charles, J. and De Pauw, N. (2004a). Application of Bayesian Belief Networks for the prediction of macroinvertebrate taxa in rivers. *Annales de Limnologie-International Journal of Limnology* **40**: 181-191.

Adriaenssens, V., De Baets, B., Goethals, P. L. M. and De Pauw, N. (2004b). Fuzzy rule-based models for decision support in ecosystem management. *The Science of the Total Environment* **319**: 1-12.

Ahmadi-Nedushan, B., St-Hilaire, A., Berube, M., Robichaud, E., Thiemonge, N. and Bobee, B. (2006). A review of statistical methods for the evaluation of aquatic habitat suitability for instream flow assessment. *River Research and Applications* **22**: 503-523.

Akkermans, W., Verdonschot, P. F. M., Nijboer, R. C., Goedhart, P. W. and Braak, C. J. F. (2005). Predicting macro-fauna community types from environmental variables by means of support vector machines. In: Lek, S., Scardi, M., Verdonschot, P. F. M, Descy, J. P. and Park, Y. S. (2004). *Modelling community structure in freshwater ecosystems*, Springer, 518 pp.

Allouche, O., Tsoar, A. and Kadmon, R. (2006). Assessing the accuracy of species distribution models: prevalence, Kappa and the true skill statistic (TSS). *Journal of Applied Ecology* **43**: 1223-1232.

Amari, S., Murata, N., Muller, K. R., Finke, M. and Yang, H. H. (1997). Asymptotic statistical theory of overtraining and cross-validation. *IEEE Transactions on Neural Networks* **8**: 985-996.

Anderson, P. G. (1993). Adaptation of a habitat suitability model for prioritizing habitat rehabilitation needs of northern pike (*Esox lucius*), MS thesis, Peterborough (Ontario, Canada), Trent University, 96 pp. In: Casselman, J. M. and Lewis, C. A. (1996). Habitat requirements of northern pike (*Esox lucius*). *Canadian Journal of Fisheries and Aquatic Sciences* **53:** 161-174.

Auger, P., Charles, S., Viala, M. and Poggiale, J. (2000). Aggregation and emergence in ecological modelling: integration of ecological levels. *Ecological Modelling* **127**: 11-20.

Austin, M. P. (2002). Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling* **157**:101-118.

Austin, M. P. and Meyers, J. A. (1996). Current approaches to modelling the environmental niche of eucalyptus: implications for management of forest biodiversity. *Forest Ecology Management* **85**: 95-106.

Bahlmann, C., Haasdonk, B. and Burkhardt, H. (2002). On-line handwriting recognition with support vector machines- a kernel approach. In: *Proceeding of the 8th International Workshop on Frontiers in Handwriting Recognition (IWFHR)*, pp. 49-54.

Barros, L. C., Bassanezi, R. C. and Tonelli, P. A. (2000). Fuzzy modelling in population dynamics. *Ecological Modelling* **128**: 27-33.

Beauchard, O., Gagneur, J. and Brosse, S. (2003). Macroinvertebrate richness patterns in North African streams. *Journal of Biogeography* **30**: 1821-1833.

Bebis, G. and Georgiopoulos, M. (1994). Feed-forward neural networks. *IEEE Potentials* **13**: 27-31.

Belpaire, C., Smolders, R., Auweele, I. V., Ercken, D., Breine, J., Van Thuyne, G. and Ollevier, F. (2000). An index of biotic integrity characterizing fish populations and the ecological quality of Flandrian water bodies. *Hydrobiologia* **434**: 17-33.

Benda, L., Poff, N. L., Tague, C., Palmer, M. A., Pizzuto, J., Cooper, S. L., Stanley, E. H. and Moglen, G. (2002). How to avoid train wrecks when using science in environmental problem solving. *BioScience* **52**: 1127-1136.

Berg, S., Jeppesen, E. and Sondergaard, M. (1997). Pike (*Esox lucius L.*) stocking as a biomanipulation tool. 1. Effects on the fish population in Lake Lyng (Denmark). *Hydrobiologia* **342**: 311-318.

Bertolo, A. and Magnan, P. (2005). The relationship between piscivory and growth of white sucker (*Catostomus commersoni*) and yellow perch (*Perca flavescens*) in headwater lakes of the Canadian Shield. *Canadian Journal of Fisheries and Aquatic Sciences* **62**: 2706-2715.

Blockeel, H., Dzeroski, S. and Grbovic, J. (1999a). Simultaneous prediction of multiple chemical parameters of river water quality with TILDE. In: *Proceedings of the third European conference on principles of data mining and knowledge discovery. Lecture notes in artificial intelligence* **1704**: 32-40.

Blockeel, H., Dzeroski, S. and Grbovic, J. (1999b). Experiments with TILDE in the river water quality domain.

Breiman, L. (2001). Random forests. *Machine Learning* **45**: 5-32.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and regression trees*. Wadsworth, Belmont, 368 pp.

Breine, J., Simoens, I., Goethals, P., Quataert, P., Ercken, D., Chris, V. L. and Belpaire, C. (2004). A fish-based index of biotic integrity for upstream brooks in Flanders (Belgium). *Hydrobiologia* **522**: 133-148.

Brosse, S., Grossman, G. D. and Lek. S. (2007). Fish assemblage patterns in the littoral zone of a Europeans reservoir. *Freshwater Biology* **52**: 448-458.

Brosse, S., Arbuckle, C. J. and Townsend, C. R. (2003). Habitat scale and biodiversity: influence of catchment, stream reach and bedform scales on local invertebrate diversity. *Biodiversity and Conservation* **12**: 2057-2075.

Brosse, S. and Lek, S. (2000). Modelling roach (*Rutilus rutilus L.*) microhabitat using linear and non-linear techniques. *Freshwater Biology* **44**: 441-452.

Brosse, S., Guegan, J. F., Tourenq, J. N. and Lek, S. (1999). The use of artificial neural network to assess fish abundance and spatial occupancy in the littoral zone of a mesotrophic lake. *Ecological Modelling* **120**: 299-311.

Brown, G. E. (2003). Learning about danger: chemical alarm cues and local risk assessment in prey fishes. *Fish and Fisheries* **4**: 227-234.

Brown, M., Lewis, H. G. and Gunn, S. R. (1999). Support Vector Machines for optimal classification and spectral unmixing. *Ecological Modelling* **120**: 167-179.

Bry, C. (1996). Role of vegetation in the life cycle of pike. Pp. 45–67. In: Craig, J. F. (ed.). *Pike: Biology and exploitation*. London, Chapman and Hall. 298 pp.

Buckland, S. T. and Elston, D. A. (1993). Empirical models for the spatial distribution of wildlife. *Journal of Applied Ecology* **30**: 478-495.

Burbidge, R., Trotter, M., Buxton, B. and Holden, S. (2001). Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Computational Chemistry* **26**: 5-14.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and Knowledge Discovery* **2**: 121-167.

Carter, M. P. (1987). Preliminary findings of a survey of OR society membership. *Operational Research Society* **38**: 3-16.

Caruana, R. and Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In: *proceedings of the 23rd international conference on machine learning, Pittsburgh, PA.*

Casselman, J. M. (2002). Effects of temperature, global extremes, and climate change on year-class production of warmwater, coolwater, and coldwater fishes of the Great Lakes Basin. *American Fisheries Society Symposium* **32**: 39-60.

Casselman, J. M. and Lewis, C. A. (1996). Habitat requirements of northern pike (*Esox lucius*). *Canadian Journal of Fisheries and Aquatic Sciences* **53**: 161-174.

Casselman, J. M. (1978). Effects of environmental factors on growth, survival, activity, and exploitation of northern pike. Pp.114-128. In: Kendall, R. L. (ed.). *Selected coolwater fishes of North America*. American Fisheries Society, Special Publication 11, Bethesda, Maryland

Cessie, l. S. and van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Applied Statistics* **41**: 191-201.

Chapman, C. A. and Mackay, W. C. (1984). Direct observation of habitat utilization by northern pike. *Copeia* **1984**: 255-258.

Chen, Q. and Mynett, A. E. (2004). Predicting Phaeocystis globosa bloom in Dutch coastal waters by decision trees and nonlinear piecewise regression. *Ecological Modelling* **176**: 277-290.

Cherkassky, V. and Lari-Najafi, H. (1992). Data representation for diagnostic neural networks. *IEEE Intelligent Systems* **7**: 43-53.

Chon, T. S., Park, Y. S., Kwak, I. S. and Cha, E. Y. (2002). Non-linear approach to grouping, dynamics and organizational informatics of benthic macroinvertebrate communities in streams by artificial neural networks. In: Recknagel, F. (ed.). *Ecological Informatics: Understanding ecology by biologically-inspired computation.* Springer, Berlin, 398 pp.

Chon, T. S., Kwak, I. S., Park, Y. S., Kim, T. H. and Kim, Y. (2001). Patterning and short-term predictions of benthic macroinvertebrate community dynamics by using a recurrent artificial neural network. *Ecological Modelling* **146**: 181-193.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **20**: 37-46.

Congalton, R. G. (1991). A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment* **37**: 35-46.

Craig, J. F. (2008). A short review of pike ecology. *Hydrobiologia* **601**: 5-16.

Craig, P., Paukert, D., and Willis. W. (2003). Population characteristics and ecological role of northern pike in shallow natural lakes in Nebraska. *North American Journal of Fisheries Management* **23**: 313-322.

Craig, J. F. (1996). *Pike: Biology and exploitation*. Chapman and Hall, London, 298 pp.

Craig, J. F. (1987). *The biology of the perch and related fish*. Croom Helm, London, Timber Press, Portland (Oregon), 333 pp.

Craig, J. F. and Kipling, C. (1983). Reproduction effort versus the environment, case histories of Windermere perch (*Perca fluviatilis* L.) and pike (*Esox lucius* L.). *Fish Biology* **22**: 713-727.

Crossman, E. J. (1996). Taxonomy and distribution. Pp.1-12. In: Craig, J. F. (ed.). *Pike: Biology and exploitation*. Chapman and Hall, London, 298 pp.

Crowder, L. B. and Cooper, W. E. (1982). Habitat structural complexities and the interaction between bluegills and their prey. *Ecology* **63**: 1802-1813.

Dai, H. C. and Macbeth, C. (1997). Effects of learning parameters on learning procedure and performance of a BPNN. *Neural Networks* **10**: 1505-1521.

Dakou, E., Goethals, P. L. M., D'heygere, T., Dedecker, A. P., Gabriels, W., De Pauw, N. and Lazaridou-Dimitriadou, M. (2006). Development of artificial neural network models predicting macroinvertebrate taxa in the river Axios (Northern Greece). *Annales de Limnologie- Animal Limnology* **42**: 241-250.

De'ath, G. and Fabricius, K. E. (2000). Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology* **81**: 3178-3198.

Debeljak, M., Dzeroski, S. and Adamic, M. (1999). Interactions among the red deer *(Cervus elaphus, L.)* population, meteorological parameters and new growth of the natural regenerated forest in Sneznik, Slovenia. *Ecological Modelling* **121**: 51-61.

Decoste, D. and Scholkopf, B. (2002). Training invariant support vector machines. *Machine Learning* **46**: 161-190.

Dedecker, A. P., Goethals P. L. M., D'heygere, T., Gevrey, M., Lek, S. and De Pauw, N. (2005). Application of artificial neural network models to analyze the relationships between *Gammarus pulex* L. (Crustacea, Amphipoda) and river characteristics. *Environmental Monitoring and Assessment* **111**: 223-241.

Dedecker, A. (2005). Development of habitat and migration models for the prediction of macroinverteberates in rivers. PhD thesis, Ghent University, Belgium, 327 pp.

Dedecker, A., Goethals, P. L. M., Gabriels, W. and De Pauw, N. (2004). Optimization of Artificial Neural Network (ANN) model design for prediction of macroinvertebrate communities in the Zwalm river basin (Flanders, Belgium). *Ecological Modelling* **174:** 161-173.

Dedecker, A. P., Goethals, P. L. M., Gabriels, W. and De Pauw, N. (2002). Comparison of Artificial Neural Network (ANN) model development methods for prediction of macroinvertebrates communities in the Zwalm river basin in Flanders, Belgium. *The Scientific World Journal* **2**: 96-104.

Degraer, S., Verfaillie, E., Willems, W., Adriaens, E., Vincx, M. and Van Lancker, V. (2008). Habitat suitability modelling as a mapping tool for macrobenthic communities: An example from the Belgian part of the North Sea. *Continental Shelf Research* **28**: 369-379.

Deltour, J. (2004). The elaboration of recovery plans for the pike (*Esox lucius*) in the Zwalm basin (in Flanders). MS thesis, Ghent university (in Dutch), 167 pp.

De Nie, H. W. (1993). Atlas van de Nederlandse Zoetwatervissen. Stichting Atlas Verspreiding Nederlandse zoetwatervissen. Doetinchem. Media publishing. In: Glasbergen, M. J**.** (ed.). *Fish habitat requirements in the Dutch part of the river Rhine*. Delft Cluster Hydraulics, 68 pp.

De Pauw, N. and Vannevel, R. (1993). Macroinvertebrates and water quality. Stichting Leefmilieu, Dossier N° 11, Antwerp, pp. 316 (in Dutch).

De Pauw, N. and Vanhooren, G. (1983). Method for biological quality assessment of watercourses in Belgium. *Hydrobiologia* **100**:153-168.

D'heygere, T., Goethals, P. L. M and De Pauw, N. (2006). Genetic algorithms for optimisation of predictive ecosystems models based on decision trees and neural networks. *Ecological Modelling* **195**: 20-29.

D'heygere, T., Goethals, P. L. M and De Pauw, N. (2003). Use of genetic algorithms to select input variables in decision tree models for the prediction of benthic macroinverteberates. *Ecological Modelling* **160**: 291-300.

D'heygere, T., Goethals, P. L. M. and De Pauw, N. (2002). Optimisation of the monitoring strategy of macroinvertebrate communities in the river Dender, in relation to the EU Water Framework Directive. *Scientific World Journal* **2**: 607-617.

Diana, J. S. (1980). Diel activity pattern and swimming speeds of northern pike (*Esox lucius*) in Lac Ste. Anne, Alberta. *Canadian Journal of Fisheries and Aquatic Sciences* **37**: 1454-1458.

Dibike, Y. B., Velickov, S. and Solomantine, D. (2000). Support vector machines: review and application in civil engineering. *In: Proceeding of the 2nd joint workshop on application of AI in civil engineering, Germany.*

Dörner, H., Hülsmann, S., Hölker, F., Skov, C. and Wagner. (2007). A Size-dependent predator-prey relationships between pikeperch and their prey fish. *Ecology of Freshwater Fish* **16**: 307-314.

Drake, J. M., Randin, C. and Guisan, A. (2006). Modelling ecological niches with support vector machines. *Applied Ecology* **43**: 424-432.

Dzeroski, S. and Drumm, D. (2003). Using regression trees to identify the habitat preference of the sea cucumber (*Holothuria leucospilota*) on Rarotonga, Cook Islands. *Ecological Modelling* **170**: 219-226.

Dzeroski, S. (2001). Applications of symbolic machine learning to ecological modelling. *Ecological Modelling* **146**: 263-273.

Dzeroski, S., Grbovic, J. and Demsar, D. (2000). Predicting chemical parameters of river water quality from bioindicator data. *Applied Intelligence* **13**: 7-17.

Dzeroski, S., Grbovic, J. and Walley, W. J. (1997). Machine learning applications in biological classification of river water quality. Pp. 429-448. In: Michalski, R. S., Bratko, I. and Kubat, M. (eds.). *Machine learning and data mining: methods and applications*. John Wiley and Sons Ltd., New York, USA.

Eklov, P. and Jonsson, P. (2007). Pike predators induce morphological changes in young perch and roach. *Journal of Fish Biology* **70:** 155-64.

Eklov, P. (1997). Effects of habitat complexity and prey abundance on the spatial and temporal distribution of perch (*Perca Fluviatilis*) and pike (*Esox lucius*). *Canadian Journal of Fisheries and Aquatic Sciences*. **54**: 1520-1531.

Eklov, P. and Persson, L. (1996). The response of prey to the risk of predation: proximate cues for refuging juvenile fish. *Animal Behaviour* **51**: 105-15.

Elith, J., Burgman, M. A. and Regan, H. M. (2002). Mapping epistemic uncertainties and vague concepts in predictions of species distribution. *Ecological Modelling* **157**: 313-329.

Elliott, J. M. and Hurley, M. A. (2000). Optimum energy intake and gross efficiency of energy conversion for brown trout (*Salmo trutta)* feeding on invertebrates or fish. *Freshwater Biology* **44**: 605-615.

EU. (2000). Directive of the European Parliament and of the Council 2000/60/EC establishing a framework for community action in the field of water policy, Rep. No. PE-CONS 3639/1/00 REV 1. European Union, Luxembourg.

European Environment Agency (EEA) (2002). Environmental Signals**,** Report No.9, European Environment Agency, Copenhagen.

Faraway, J. and Chatfield, C. (1998). Time series forecasting with neural network: a comparative study using airline data. *Applied Statistics* **47**: 231-250.

Fausch, K. D., Hawkes, C. L. and Parsons, M. G. (1988). Models that predict the standing crop of stream fish from habitat variables: 1950-1985. General Technical Report. PNW-GTR-213. U.S. Portland, OR: U.S. Department of Agriculture, Forest Service, Pacific Northwest Research Station, 52 pp.

Fielding, A. H. and Bell, J. F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservations* **24**: 38-49.

Fitter, R. and Manuel, R. (1986). *The Collins guide to freshwater life*. Collins, London, 382 pp.

Fletcher, D. and Goss, E. (1993). Forecasting with neural networks: an application using bankruptcy data. *Information and Management* **24**: 159-167.

Flood, I. and Kartam, N. (1994). Neural networks in civil engineering. I: Principles and understanding. *Journal of Computing in Civil Engineering* **8**: 131-148.

Folkvord, A. (1997). Ontogeny of cannibalism in larval and juvenile fish with special emphasis on Atlantic cod (*Gadus morhua L.*). In: Chambers, R. C. and Trippel, E. A. (eds.). *Early life history and recruitment in fish populations*. Pp. 335-366. London, Chapman and Hall Ltd.

Forbes, A. D. (1995). Classification-algorithm evaluation: Five performance measures based on confusion matrices. *Journal of Clinical Monitoring and Computing* **11**: 189-206.

Franklin, J. (1995). Predictive vegetation mapping: geographic modelling of biospatial patterns in relation to environmental gradients. *Progress in Physical Geography* **19**: 474-499.

Franz, L. S. (1989). Data driven modelling: An application in scheduling department of management. *Decision Sciences* **20:** 359-377.

Fryxell, J. M. and Lundberg, P. (1998). *Individual behaviour and community dynamics*. Chapman and Hall, London, 224 pp.

Fullhart, H. G., Parsons, B. G., Willis, D. W., and Reed, J. R. (2002). Yellow perch piscivory and its possible role in structuring littoral zone fish communities in small Minnesota Lakes. *Freshwater Ecology* **17**: 37-43.

Gabriels, W., Goethals, P. L. M. and De Pauw, N. (2002). Prediction of macroinvertebrate communities in sediments of Flemish watercourses based on artificial neural networks. *Verhandlungen Internationale Vereinigung für theoretische und angewandte Limnologie.* **28**: 777-780.

Geoffrey, I. W. (2000). MultiBoosting: A technique for combining boosting and wagging. *Machine Learning* **40:** 159-196.

Glasbergen, M. J. (2001a). Fish habitat modelling in the Gamerensche Waard. WL/Delft Hydraulics, Project report, 68 pp.

Glasbergen, M. J. (2001b). Fish habitat requirements in the Dutch part of the river Rhine. WL/Delft Hydraulics, Project Report, 49 pp.

Goethals, P. L. M., D'heygere, T., Zarkami, R., Klinge, M., Jaarsma, N., Grimm, M., Belpaire, C., Van Thuyne, G. and De Pauw, N. (2006a). Restoration management of pike (*Esox lucius L.*) in Flanders, Belgium. In: *Proceeding of Biological responses to chemical and hydromorphological pressures.* Rebecca conference, WL/Delft hydraulics, Delft, Netherlands.

Goethals, P. L. M., Zarkami, R., D'heygere, T., De Pauw, N., Jaarsma, N., Klinge, M., Grimm, M., Van Thuyne, G. and Belpaire, C. (2006b). Status, ecology and conservation of pike (*Esox lucius L.*) in Flanders. Project nr. B and G/14/2001, Research report (in Dutch), 112 pp.

Goethals, P. L. M. (2005). Data driven development of predictive ecological models for benthic macroinvertebrates in rivers. PhD thesis, Ghent University, Belgium, 377 pp.

Goethals, P. L. M., Dedecker, A. P., Gabriels, W. and De Pauw, N. (2002). Development and application of predictive river ecosystem models based on classification trees and artificial neural networks. Pp. 91-107. In: Recknagel, F. (ed.). *Ecological Informatics: Understanding Ecology by Biologically-Inspired Computation*. Springer-Verlag, Berlin, 432 pp.

Goethals, P. L. M. and De Pauw, N. (2001). Development of a concept for integrated ecological river assessment in Flanders, Belgium. *Journal of Limnology* **60**: 7-16.

Goh, A.T.C. (1995). Back-propagation neural networks for modelling complex systems. *Artificial Intelligence in Engineering* **9**: 143-151.

Goldberg, D. E. (1989). *Genetic algorithms in search, optimization and machine learning*. Addison-Wesley Publishing Company, Reading, MA, 412 pp.

Gotelli, N. J. and Ellison, A. M. (2004). *A primer of ecological statistics*. Sinauer Associates Inc. Sunderland, Massachusetts, 510 pp.

Green, R. E., Osborne, P. E. and Sears, E. J. (1994). The distribution of passerines birds in hedgerows during the breeding season in relation to characteristics of hedgerows and adjacent farmlands. *Applied Ecology* **31**: 677-692.

Griffiths, R. W., Newlands, N. K., Noakes, D. L. G. and Beamish, F. W. H. (2004). Northern pike (*Esox lucius L.*) growth and mortality in a northern Ontario river compared with that in lakes: influence of flow. *Ecology of Freshwater Fish* **13**: 136-144.

Grimm, M. P. and Klinge, M. (1996). Pike and some aspects of its dependence on vegetation. Pp.125-156. In: Craig, J. F. (ed.). *Pike: biology and exploitation*. Chapman and Hall, London, UK, 298 pp.

Grimm, M. P. (1981). Intraspecific predation as a principal factor controlling the biomass of northern pike. *Fisheries Management* **12**: 77-79.

Guisan, A. and Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling* **135**: 147-186.

Guoa, Q., Kellya, M. and Graham, C. H. (2005). Support vector machines for predicting distribution of Sudden Oak Death in California. *Ecological Modelling* **182:** 75-90.

Guyon, I., Weston, J., Barnhill, S. and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning* **46**: 389-422.

Hagan, M. T., Demuth, H. B. and Beale, M. (1996). *Neural network design*. PWS Publishing Company, Boston, USA, 712 pp.

Hansson, S., Arrhenius, F. and Nellbring, S. (1997). Diet and growth of pikeperch (*Stizostedion lucioperca* L.) in a Baltic Sea area. *Fisheries Research* **31**: 163-167.

Hawkins, L. A., Armstrong, J. D. and Magurran, A. E. (2003). Settlement and habitat use by juvenile pike in early winter. *Fish Biology* **63**: 174-186.

Haykin, S. (1999). *Neural Networks*: *A comprehensive foundation*. Second edition. Prentice Hall, New Jersey, USA, 842 pp.

Haykin, S. (1994). *Neural networks: A comprehensive foundation*. Macmillan, New York.

Hein, S., Binzenhofer, B., Poethke, H. J., Biedermann, R., Settele, J. and Schroder, B. (2007). The generality of habitat suitability models: A practical test with two insect groups. *Basic and Applied Ecology* **8**: 310-320.

Helmer, E. H., Brown, S. and Cohen, W. B. (2000). Mapping montane tropical forest successional stage and land-use with multi-date Landsat imagery. *Remote Sensing* **21**: 2163-2183.

Hinch, S. G., Collins, N. C. and Harvey, H. H. (1991). Relative abundance of littoral zone fishes: biotic interactions, abiotic factors, and postglacial colonization. *Ecology* **72**: 1314-1324.

Hirzel, A. and Guisan, A. (2002). Which is the optimal sampling strategy for habitat suitability modelling. *Ecological Modelling* **157**: 331-341.

Holland, J. H. (1975). *Adaptation in natural and artificial systems*. University of Michigan Press, Ann Arbor, MI.

Holling, C. S. and Allen, C. R. (2002). Adaptive inference for distinguishing credible from incredible patterns in nature. *Ecosystems* **5**: 319-328.

Hornik, K., Stinchcombe, M. and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks* **2**: 359-366.

Hosmer, D. W. and Lemeshow, S. (1989). *Applied logistic regression*. Wiley, New York.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied logistic regression*. Second edition, Wiley, New York.

Hsu, C. W., Chang, C. C. and Lin, C. J. (2008). A Practical guide to support vector classification, Department of computer science and information engineering, National Taiwan University.

Huang, G. H. and Xia, J. (2001). Barriers to sustainable water quality management. *Environmental Management* **61**: 1-23.

Huang, C. L. and Wang, C. J. (2006). A GA-based feature selection and parameters optimization for support vector machines. *Expert systems with applications* **31**: 231-40.

Huet, M. (1954). Biologie, profils en long et en travers des eaux courantes. *Bulletin Francais de Pisciculture* **175**: 41-53.

Huet, M. (1949). Apercu dés relations entre la pente et les populations des eaux courantes. Schweizerische Zeitschrift fur *Hydrologie* **11:**333-351

Humpl, M. and Pivnicka, K. (2006). Fish assemblages as influenced by environmental factors in streams in protected areas of the Czech Republic. *Ecology of Freshwater Fish* **15**: 96-103.

Hung, M. S., Hu, M. Y., Shanker, M. S. and Patuwo, B. E. (1996). Estimating posterior probabilities in classification problems with neural networks. *International Journal of Computational Intelligence and Organizations* **1**: 49-60.

Inskip, P. D. (1982). *Habitat suitability index models: northern pike*. US department of interior, fish and wildlife service (FWS/OBS-82/10.17), Washington D.C.VS, 40 pp.

Jongman, R. H. G., Ter Braak, C. J. F. and Van Tongeren, O. F. R. (1995). *Data analysis in community and landscape ecology.* Second edition, Cambridge University Press, Cambridge, 299 pp.

Jorde, K., Schneider, M. and Zoellner, F. (2000). Analysis of instream habitat quality preference functions and fuzzy models. In: Glasbergen, M. J. (2001a). *Fish habitat modelling in the Gamerensche Waard*. WL/Delft Hydraulics. Project report, 68 pp.

Junker, B. and Buchecker, M. (2008). Aesthetic preferences versus ecological objectives in river restorations. *Landscape and Urban Planning* **85**: 141-154.

Kahilainen, K. and Lehtonen, H. (2003). Piscivory and prey selection of four predator species in a whitefish dominated subarctic lake. *Fish Biology* **63**: 659-672.

Kanellopoulos, I. and Wilkinson, G. G. (1997). Strategies and best practice for neural network image classification. *International Journal of Remote Sensing* 1**8**: 711-725.

Karul, C., Soyupak, S., Cilesiz, A. F., Akbay, N. and Germen, E. (2000). Case studies on the use of neural networks in eutrophication modelling. *Ecological Modelling* **134**: 145-152.

Keating, K. A. (2004). Use and interpretation of logistic regression in habitat selection studies. *Journal of Wildlife Management* **68**: 774-789.

Keerthi, S. S., Shevade, S. K., Bhattacharyya, C. and Murthy, K. R. K. (2001). "Improvements to Platt's SMO algorithm for SVM classifier design". *Neural Computation* **13**: 637-649.

Kelley, J. L. and Magurran, A. E. (2003). Learned predator recognition and antipredator responses in fishes. *Fish and Fisheries* **4**: 216-226.

Kerle, F., Zollner, F., Kappus, B., Marx, W. and Giesecke, J. (2001). Fish habitat and vegetation modelling in floodplains with CASIMIR. CFR project report 13, IWS, University of Stuttgart, 75 pp.

King, G. and Zeng, L. (2000). Logistic regression in rare events data. *Political Analysis* **9**: 137-163.

Kobler, A., Klefoth, T., Wolter, C., Fredrich, F. and Arlinghaus, R. (2008). Contrasting pike (*Esox lucius L.*) movement and habitat choice between summer and winter in a small lake. International pike symposium, *Hydrobiologia* **601**:17-27.

Kohonen, T. (1995). *Self-organizing maps*. Springer-Verlag, Berlin, 362 pp.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics* **43:** 59-69.

Kruk, A., Lek, S., Park, Y. S. and Penczak, T. (2007). Fish assemblages in the large lowland Narew river system (Poland): Application of the self-organizing map algorithm. *Ecological Modelling* **203***:* 45-61.

Kurt, I., Ture, M. and Turhan, K. A. (2008). Comparing performances of logistic regression, classification and regression tree and neural networks for predicting coronary artery disease. *Expert Systems with Applications* **34**: 366-374.

Lafayette, D. L. (2003). Developing habitat suitability criteria for water management: A case study. *International Journal River Basin Management* **1**: 283-295.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* **33**: 159-174**.**

Launey, S., Morin, J., Minery, S. and Laroche, J. (2006). Microsatellite genetic variation reveals extensive introgression between wild and introduced stocks, and a new evolutionary unit in French pike (*Esox lucius* L.). *Journal of Fish Biology* **68**: 193-216.

Lee, J. H. W., Huang, Y., Dickman, M. and Jayawardena, A. W. (2003). Neural network modelling of coastal algal blooms. *Ecological Modelling* **159**: 179-201.

Legendre, P. and Legendre, L. (1998). *Numerical ecology*. Second English edition. Elsevier Science. Amsterdam, 870 pp.

Lehtiniemi, M. (2005). Swim or hide: predator cues cause species specific reactions in young fish larvae. *Journal of Fish Biology* **66**: 1285-1299.

Lek, S., Scardi, M., Verdonschot, P. F. M. Descy, J. P. and Park, Y. S. (2004). *Modelling Communities Structure in Freshwater Ecosystems.* Springer, 518 pp.

Lek, S. and Guégan, J. F. (1999). Artificial neural network as a tool in ecological modelling, an introduction. *Ecological Modelling* **120**: 65-73.

Lek, S., Delacoste, M., Baren, P., Dimopoulos, I., Lauga, J. and Aulagner, S. (1996). Application of neural networks to modelling non-linear relationships in ecology. *Ecological Modelling* **90**: 39-52.

Lek, S., Beland, A., Dimopoulos, I., Lauga, J. and Moreau, J. (1995). Improved estimation, using neural networks, of the food consumption of fish populations. *Marine and Freshwater Research* **46**: 1229-1236.

Lenard, M. J., Alam, P. and Madey, G. R. (1995). The application of neural networks and a qualitative response model to the auditor's going concern uncertainty decision. *Decision Sciences* **26**: 209-227.

Lucasius, C. B. and Kateman, G. (1993). Understanding and using genetic algorithms. Part 1. Concepts, properties and context. *Chemometrics and Intelligent Laboratory Systems* **19**: 1-33.

Luz, L. D. and Loucks, D. P. (2003). Developing habitat suitability criteria for water management: A case study. *International Journal River Basin Management* **1**: 283-295.

Ma, C.Y., Yang, S. Y., Zhang, H., Xiang, M. L., Huang, Q. and Wei, Y.Q. (2008). Prediction models of human plasma protein binding rate and oral bioavailability derived by using GA–CG–SVM method. *Journal of Pharmaceutical and Biomedical Analysis* **120**: 677-682.

Maier, H. R. and Dandy, G. C. (2000). Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. *Environmental Modelling and Software* **15**: 101-124.

Maitland, P. S. and Campbell, R. N. (1992). *Freshwater fishes*. New Naturalist Series. Walters, M., West, R., Streeter, D. and Corbet, S. A. (eds.). Pp.167-176. Harper Collins, London.

Manel, S., Williams, H. C. and Ormerod, S. J. (2001). Evaluating presence/absence models in ecology: the need to account for prevalence. *Applied Ecology* **38**: 921-931

Manel, S., Dias, J. M., Buckton, S. T. and Ormerod, S. J. (1999). Alternative methods for predicting species distribution: an illustration with Himalayan river birds. *Applied Ecology* **36**: 734-747.

Margenau, T. L., Rasmussen, P. W. and Kampa, J. M. (1998). Factors affecting growth of northern pike in small northern Wisconsin lakes. *North American Journal of Fisheries Management* **18**: 625-639.

Masters, T. (1993). *Practical Neural Network Recipes in C++*. Academic Press, San Diego, USA.

Mastrorillo, S., Lek, S. and Dauba, F. (1997). Predicting the abundance of minnow *Phoxinus phoxinus* (Cyprinidae) in the River Ariege (France) using artificial neural networks. *Aquatic Living Resources* **10**: 169-176.

McClane, A. J. (ed.). (1998). *McClane's new standard fishing encyclopedia and international angling guide.* Random House Value Publishing. Younger, R, E. (Illustrator).

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*. Chapman and Hall, London.

Mehner, T., Schultz, H., Bauer, D., Herbst, R., Voigt, H. and Benndorf, J. (1996). Intraguild predation and cannibalism in age-0 perch (*Perca fluviatilis L.*) and age-0 zander (*Stizostedion lucioperca L.*): interactions with zooplankton succession, prey fish availability and temperature. *Annales zoologici Fennici* **33**: 353-361.

Meijer, M. L. and Hosper, H. (1997). Effects of biomanipulation in the large and shallow Lake Wolderwijd, The Netherlands. *Hydrobiologia* **342**: 335-349.

Minns, C. K., Bakelaar, C. N., Brunette, P. C. E. and Schertzer, W. M. (1999). Identifying habitats essential for pike (*Esox lucius* L.) in the Long Point Region of Lake Erie: a suitable supply approach. *American Fisheries Society* **22**: 363-382.

Minns, C. K., Randall, R. G., Moore, J. E. and Cairns. V. W. (1996). A model simulating the impact of habitat supply limits on northern pike (*Esox lucius L.). Canadian Journal of Fisheries and Aquatic Sciences* **53**: 20-34.

Mjolsness, E. and DeCoste, D. (2001). Machine learning for science: state of the art and future prospects. *Science* **293**: 2051-2055.

Nilsson, P. A., Skov, C. and Farrell, J. M. (2008). Current and future directions for pike ecology and management: a summary and synthesis. International pike symposium. *Hydrobiologia* **601**: 137-141.

Nilsson, C., Pizzuto, J. E., Moglen, G. E., Palmer, M. E., Stanley, E. H., Bockstael, N. E. and Thompson, L. C. (2003). Ecological forecasting and the urbanization of stream Ecosystems: Challenges for economists, hydrologists, geomorphologists and ecologists. *Ecosystems* **6**: 659-674.

Nilsson, P. A. (2001). Predator behaviour and prey density: evaluating density-dependent intraspecific interactions on predator functional responses. *Animal Ecology* **70**: 14-19.

Nilsson, P. A. and Bronmark, C. (2000). Prey vulnerability to a gape-sized limited predator: behavioural and morphological impacts on northern pike piscivory. *Oikos* **88**: 539-546.

Noble, R. A. A., Cowx, I. G., Goffaux, D. and Kestemont, P. (2007). Assessing the health of European rivers using functional ecological guilds of fish communities: standardizing species classification and approaches to metric selection. *Fisheries Management and Ecology* **14**: 381-392.

Obach, M., Wagner, R., Werner, H. and Schmidt, H. H. (2001). Modelling population dynamics of aquatic insects with artificial neural networks. *Ecological Modelling* **146**: 207-217.

Olden, J. D. and Jackson, D. A. (2002). Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. *Ecological Modelling* **154**: 135-150.

Olden, J. D. (2003). A species-specific approach to modelling biological communities and its potential for conservation. *Conservation Biology* **17**: 854-863.

Olson, M. H. (1996). Ontogenetic niche shifts in largemouth bass: variability and consequences for first-year growth. *Ecology* **77**: 179-190.

Osborne, P. E. and Tigar, B. J. (1992). Interpreting bird atlas data using logistic models: an example from Lesotho, Southern Africa. *Applied Ecology* **29**: 55-62.

Pakath, R. and Zaveri, J. S. (1995). Specifying critical inputs in a genetic algorithm-driven decision support system: An automated facility. *Decision Sciences* **26**: 749-771.

Patuwo, E., Hu, M. Y. and Hung, M. S. (1993). Two-group classification using neural networks. *Decision Sciences* **24**: 825-845.

Penczak, T., O'Hara, K. and Kostrzewa, J. (2002). Fish bioenergetics model used for estimation of food consumption in a navigation canal with heavy traffic. *Hydrobiologia* **479**: 109-123.

Persson, A. and Bronmark, C. (2002). Foraging capacity and resource synchronization in an ontogenetic diet switcher, pikeperch *(Stizostedion lucioperca L.)*. *Ecology* **83**: 3014-3022.

Peters, R. H. (1991). *A critique for ecology.* Cambridge University Press, Cambridge, 384 pp.

Pierce, R. B., Tomcko, C. M. and Margenau, T. L. (2003). Density dependence in growth and size structure of northern pike populations. *North American Journal of Fisheries Management* **23**: 331-339.

Piramuthu, S., Shaw, M. and Gentry, J. (1994). A classification approach using multi-layered neural networks. *Decision Support Systems* **11**: 509-525.

Poff, N. L., Allan, J. D., Palmer, M. A., Hart, D. D., Richter, B. D., Arthington, A. H., Rogers, K. H., Meyer, J. L. and Stanford, J. A. (2003). River flows and water wars: emerging science for environmental decision making. *Ecological Environment* **1**: 298-306.

Prejs, A., Pijanowska, J., Koperski, P., Martyniak, A., Boron, S. and Hliwa, P. (1997). Food-web manipulation in a small, eutrophic Lake Wirbel, Poland: long-term changes in fish biomass and basic measures of water quality. A case study. *Hydrobiologia* **342**: 383-386.

Qian, N. (1999). On the momentum term in gradient descent learning algorithms. *Neural Networks* **12**: 145-151.

Quinlan, J. R. (1993). *C4.5: program for machine learning*. Morgan Kaufmann publishers, San Francisco, 302 pp.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning* **1**: 81-106.

Raat, A. J. P. (1988). *Synopsis of biological data on the northeren pike (Esox lucius L.).* FAO Fisheries Synopsis 30, 178 pp.

Recknagel, F. (2003). *Ecological informatics: Understanding ecology by biologically-inspired computation*. Springer-Verlag, Berlin, Germany, 432 pp.

Recknagel, F. (2001). Applications of machine learning to ecological modelling. *Ecological Modelling* **146**: 303-310.

Recknagel, F. and Wilson, H. (2000). Elucidation and prediction of aquatic ecosystems by artificial neuronal networks. In: Lek, S. and Guegan, J. (eds.). (2000). *Artificial neuronal networks application to ecology and evolution*. Springer, Berlin, 262 pp.

Regan, H. M., Colyvan, M. and Burgman, M. A. (2002). A taxonomy and treatment of uncertainty for ecology and conservation biology. *Ecological Applications* **12**: 618-628.

Robertson, M. P., Caithness, N. and Villet, M. H. (2001). A PCA based modelling technique for predicting environmental suitability for organisms from presence records. *Diversity and Distributions* **7**: 15-27.

Rogers, L. L. and Dowla, F. U. (1994). Optimization of groundwater remediation using artificial neural networks with parallel solute transport modelling. *Water Resources Research* **30(2)**: 457-481.

Rong, Y. (2000). Statistical methods and pitfalls in environmental data analysis. *Environmental Forensics* **1**: 213-220.

Roset, N., Grenouillet, G., Goffaux, D., Pont, D. and Kestemont, P. (2007). A review of existing fish assemblage indicators and methodologies. *Fisheries Management and Ecology* **14**: 393-405.

Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1986). Learning representations by back-propagation errors. *Nature* **323**: 533-536.

Rutz, D. S. (1999). Movements, food availability and stomach contents of northern pike in selected Susitna river drainages, 1996-1997. Fisheries Data Series. No. 99-5. Alaska Department of fish and game, Division of sport fish, Anchorage, Alaska.

Sadat-Hashemi, S. M., Kazemnejad, A., Lucas, C. and Badie, K. (2005). Predicting the type of pregnancy using artificial neural networks and multinomial logistic regression: a comparison study. *Neural Computing and Applications* **14**: 198-202.

Santoul, F., Soulard, A., Figuerola, J., Cereghino, R. and Mastrorillo, S. (2004). Environmental factors influencing local fish species richness and differences between hydroregions in south-western France. *International Review of Hydrobiology* **89**: 79-87.

Scardi, M. (2001). Advances in neural network modelling of phytoplankton primary production. *Ecological Modelling* **146**: 33-45.

Scardi, M. and Harding, L.W. (1999). Developing an empirical model of phytoplankton primary production: A neural network case study. *Ecological Modelling* **120**: 213-223.

Schleiter, I. M., Obach, M., Borchardt, D. and Werner, H. (2001). Bioindication of chemical and hydromorphological habitat characteristics with benthic macro-invertebrates based on artificial neural networks. *Aquatic Ecology* **35**: 147-158.

Schleiter, I. M., Borchardt, D., Wagner, R., Dapper, T., Schmidt, K. D., Schmidt H. H. and Werner, H. (1999). Modelling water quality, bioindication and population dynamics in lotic ecosystems using neural networks. *Ecological Modelling* **120**: 271-286.

Schmutz, S., Melcher, A., Frangez, C. and Haidvogl, G. (2007). Spatially based methods to assess the ecological status of riverine fish assemblages in European ecoregions. *Fisheries Management and Ecology* **14**: 441-452.

Schulze, T., Dorner, H., Hölker, F. and Mehner, T. (2006). Determinants of habitat use in large roach. *Fish Biology* **69**: 1136-1150.

Siefert, R. E., Spoor, W. A. and Syreet, R. F. (1973). Effects of reduced oxygen concentration on northern pike, embryos and larva. *Fisheries Research Board of Canada* **30**: 849-852.

Sih, A. and Chirstensen, B. (2001). Optimal diet theory: when does it work and when and why does it fail? *Animal Behaviour* **61**: 379-390.

Skov, C., Nilsson, P. A., Jacobsen, L. and Bronmark, C. (2007). Habitat-choice interactions between pike predators and perch prey depend on water transparency. *Journal of Fish Biology* **70**: 298-302.

Skov, C., Lousdal, O., Johansen, P. H. and Berg, S. (2003). Piscivory of 0 + pike (*Esox lucius* L.) in a small eutrophic lake and its implication for biomanipulation. *Hydrobiologia* **506–509**: 481-487.

Skov, C., Berg, S., Jacobsen, L. and Jepsen, N. (2002). Habitat use and foraging success of 0 + pike (*Esox lucius L.*) in experimental ponds related to prey fish, water transparency and light intensity. *Ecology of Freshwater Fish* **11**: 65-73.

Skov, C. and Berg, S. (1999). Utilization of natural and artificial habitats by YOY pike in a biomanipulated lake. *Hydrobiologia* **408**: 115-122.

Solomatine, D. P. and Dulal, K. N. (2003). Model tree as an alternative to neural network in rainfall-runoff modelling. *Hydrological Science* **48**: 399-411.

Soresma, (2000). Environmental impact assessment report on the development of fish migration channels and natural overflow systems in the Zwalm River basin. Soresma advise en ingenieursbureau, Antwerp (In Dutch).

Soulie, F. (1994). Integrating neural networks for real world application. In: Zurda, A., Marks, A. and Robinson, A. (eds.). *Computational intelligence: Imitating life.* IEEE Press, New York, 396 pp.

Spens, J., Englund, G. and Lundqvist, H. (2007). Network connectivity and dispersal barriers: using geographical information system (GIS) tools to predict landscape scale distribution of a key predator (*Esox lucius L.*) among lakes. *Applied Ecology* **44**: 1127-1137.

Stankovski, V., Debeljak, M., Bratko, I. and Adamic, M. (1998). Modelling the population dynamics of red deer (*Cervus elaphus L.*) with regard to forest development. *Ecological Modelling* **108**: 145-153.

Steyerberg, E. W., Eijkemans, M. J. C., Harrell, F. E. and Habbema, J. D. F. (2000). Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. *Statistical Medicine* **19**: 1059-1079.

Tax, D. M. J. and Duin, R. P. W. (2004). Support vector data description. *Machine Learning* **54**: 45-66.

Turesson. H., Persson, A. and Bronmark, C. (2002). Prey size selection in pisicivorous pikeperch (*Stizostedion lucioperca L.*) includes active prey choice. *Ecology of Freshwater Fish* **11**: 223-233.

Vandelannoote, A. and Coeck, J. (1998). Rode lijst van de inheemse en ingeburgerde zoet- en brakwatervissen en van de rondbekken in Vlaanderen**.** Pp. 259-264. In: Vandelannoote, A., Yseboodt, R., Bruylants, B., Verheyen, R., Coeck, J., Maes, J., Belpaire, C., Van Thuyne, G., Denayer, B., Beyens, J., De Charleroy, D., Vandenabeele, P. (eds). Atlas van de Vlaamse beek-en riviervissen**.** Water Energik Vlario, 303 pp.

Vandenberghe, V., Van Griensven, A., Vanrolleghem, P. A., Goethals, P., Zarkami, R. and De Pauw, N. (2005a). Coupling water quality and fish habitat models for river management: simulation exercises in the Dender basin. Pp. 375-381. In: *Proceeding of COST 626 meeting, European aquatic modelling network, Proceedings from the final meeting in Silkeborg, Denmark.*

Vandenberghe, V., Goethals, P., Van Griensven, A., Meirlaen, J., De Pauw, N., Vanrolleghem, P. and Bauwens, W. (2005b). Application of automated measurement stations for continuous water quality monitoring of the Dender river in Flanders, Belgium. *Environmental Monitoring and Assessment* **108:** 85-98.

Van der Molen, D. T. and Boers, P. C. M. (2001). Credibility and acceptability of mathematical models for environmental impacts of agriculture in the Nederland. In: Steenvoorden, J. Claessen, F. and Willems, J. (eds). (2000). *Proceedings Conference agricultural effects on ground and surface waters.* Research at the edge of science and society. Wageningen, Red Book IAHS Publication 273 pp.

Van Hees, S. (2003). Analysis and restoration management of pike in the Demer river basin. MS thesis, Ghent University (in Dutch).

Vapnik, V. (1995). *The nature of statistical learning theory*. Springer-Verlag, New York, 187 pp.

Vehanen, T., Hyvarinen, P., Johansson, K., Laaksonen, T. (2006). Patterns of movement of adult northern pike (*Esox lucius L.*) in a regulated river. *Ecology of Freshwater Fish* **15**: 154-160.

Verdonschot, P. F. M. and Nijboer, R. C. (2002). Towards a decision support system for stream restoration in the Netherlands: an overview of restoration projects and future needs. *Hydrobiologia* **478**: 131-148.

VMM (2003). Water quality-water discharges 2002. Flemish environmental Agency, VMM, Aalst, 162 pp. (in Dutch).

Vose, M. D. (1999). *The simple genetic algorithm: Foundations and theory*. MIT Press Cambridge. MA, USA, 251 pp.

Vuorinen, P. J., Nyberg, K. and Lehtonen, H. (1998). Radioactive strontium (85Sr) in marking newly hatched pike and success of stocking. *Fish Biology* **52**: 268-280.

Wagner, R., Dapper, T. and Schmidt, H. H. (2000). The influence of environmental variables on the abundance of aquatic insects: a comparison of ordination and artificial neural networks. *Hydrobiologia* **422**: 143-152.

Walczak, S. and Cerpa, N. (1999). Heuristic principles for the design of artificial neural networks. *Information and Software Technology* **41**: 107-117.

Walker, A. J., Cross, S. S. and Harrison, R. F. (1999). Visualisation of biomedical datasets by use of growing cell structure networks: a novel diagnostic classification technique. *Lancet* **354**: 1518-1521.

Wang, F. (1994). The use of artificial neural networks in a geographical information system for agricultural land-suitability assessment. *Environment and Planning* **26**: 265-284.

Weigend, A. S., Huberman, B. A. and Rumelhart, D. E. (1990). Predicting the future: A connectionist approach. *International Journal of Neural Systems* **1**:193-209.

Welp, M. (2001). The use of decision support tools in participatory river basin management. *Physics and Chemistry of the Earth* **26**: 535-539.

Whitehead, P. G., Howard, A. and Arulmani, C. (1997). Modelling algal growth and transport in rivers. A comparison of time series analysis, dynamic mass balance and neural network techniques. *Hydrobiologia* **349**: 39-46.

Winfield, I. J., James, J. B. and Fletcher, J. M. (2008). Northern pike (*Esox lucius L.*) in a warming lake: changes in population size and individual condition in relation to prey abundance. *Hydrobiologia* **601**: 29-40.

Witten, I. H. and Frank, E. (2000). *Data Mining: practical machine learning tools and techniques with Java implementations*. Morgan Kaufmann Publishers, San Francisco, 369 pp.

Wolter, C. (2001). Conservation of fish species diversity in navigable waterways. *Landscape Urban Plan* **53**: 135-144.

Wright, J. F., Furse, M. T. and Armitage, P. D. (1993). RIVPACS: A technique for evaluating the biological quality of rivers in the UK. *European Water quality Control* **3**: 15-25.

Yang, S. Y., Huang, Q., Li, L. L., Ma, C. Y., Zhang, H., Bai, R., Teng, Q. Z., Xiang, M. L. and Wei, Y. Q. (2008). An integrated scheme for feature selection and parameter setting in the support vector machine modelling and its application to the prediction of pharmacokinetic properties of drugs. *Artificial Intelligence in Medicine*. In press.

Zweig, M. H. and Campbell, G. (1993). Receiver-operating characteristics (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry* **39**: 561-577.

Summary

Summary

The main objective of the European Water Framework Directive (WFD) is to reach a good ecological status for all water bodies in the member states of the European Union by 2015. To fulfil this, each European member state has to develop an optimal management strategy.

The main part of these water bodies can be classified as running waters or rivers. To achieve a sustainable river management, a multidisciplinary approach is needed and a consistent integrated concept has to be optimised. In order to accomplish a sustainable use of water resources, water quality and quantity control per se are not sufficient, but also structural river quality and direct management of the biology and migration facilities need to be considered carefully. However, inadequate technical and financial resources often restrict the development of the integrated approaches for river management. When dealing with prediction of species presence/absence, models are considered as important tools regarding resource assessment, environmental conservation and biological monitoring. In this context, pike is considered as a keystone predator responsible for determining the structure of certain fish communities. Therefore it was selected as species for this model development and comparison exercise.

Recently, ecological models have been gaining popularity to support river management and water policy. They have several appealing applications in this respect. They are able to provide a better interpretation of the river status. These models enable managers to calculate the effect of future river restoration actions on aquatic ecosystems and support the selection of the most sustainable options. They can help to find the major gaps in our knowledge of river systems and help to set up cost effective monitoring programmes.

This thesis consisted of three main parts: monitoring, modelling and management. On the basis of the data monitored in the river basins, the habitat suitability of pike was predicted based on a set of data driven modelling techniques such as classification trees, artificial neural networks, support-vector machines and logistic regressions in combination with genetic algorithms in the river basins in Flanders. All 4 modelling techniques were developed and applied in Weka software. This PhD research mainly focused on 6 river basins namely Demer, Dender, Dijle, Ijzer, Nete and upper Scheldt in Flanders (Belgium) during the period 1991-

2002. The biotic data (based on presence/absence of pike) were provided by the Institute for Nature and Forest Research (INBO). Additionally, some chemical data were provided by the Flemish Environment Agency (VMM). In total 21 variables (20 abiotic and 1 biotic variables) were used for the model development.

The best results (according to CCI and Kappa) were obtained with artificial neural networks and support-vector machines. The lack of illustrative power of ANNs (known as black box methods) is a major concern to ecologists since the interpretation of a model is desirable for gaining knowledge about the system that is studied. Moreover, genetic algorithms were combined with the 4 modelling techniques to analyse the contribution of environmental variables to predict the presence/absence of pike in a reliable manner and to detect the major river characteristics determining the habitat suitability of pike. In total, 14 out of 21 variables were selected to predict the pike's habitat suitability in 6 river basins in Flanders. These variables were a combination of structural-habitat and pollution variables. This implies that in addition to structural-habitat variables, pollution variables also can limit the survival of pike populations in the river basins of Flanders.

These models can in such a way support the proper selection of sustainable management options and help to persuade stakeholders to make the necessary investment and activity changes by society. The developed models were applied to support decision-making in water management in Flanders. Here, a case study was calculated in the Dender basin in Flanders. Based on the developed models, the activities related to land-use seemed to play a crucial role for pike populations. However, probably during the first years of the simulations, water quality as such was of major importance. This illustrates the limitations of the models to make simulations over periods in which very drastic changes occurred.

In summary, this thesis consisted of several chapters: an overview of the ecology of pike was made in the Chapter introduction, where quantitative approaches were reviewed. In Chapter 1, a literature review described the biological and ecological knowledge of pike. Chapter 2 dealt with the study area and ecological modelling methods. In Chapter 3, the database setup and exploration was described. Chapters 4 to 7 aimed at developing, optimising and analysing the 4 data driven models including classification trees, artificial neural networks, support-vector machines and logistic regressions in combination with genetic algorithms for the development

of the pike's habitat suitability models in 6 river basins in Flanders. In Chapter 8, these 4 techniques were compared and reviewed to get a better insight in the pike's habitat requirements. In Chapter 9, the practical application of the models in the river basins (a case study for the Dender river) was described. Finally, this thesis ends with a general discussion, conclusions and recommendations for further research.

It can be concluded that the data driven models are able to contribute to a better understanding of the interactions between the biotic and abiotic components of the river basins, offering important information to support the decision making in the river basins. However, ensure such contribution, relevant databases are necessary, and this is probably the major outcome of this study: better data should be collected in the future in Flanders, not only for model development, but also for assessment purposes in general.

Samenvatting

Samenvatting

Het voornaamste streefdoel van de Europese Kaderrichtlijn Water is het bereiken van een goede ecologische toestand voor alle wateren in de lidstaten van de Europese Gemeenschap tegen 2015. Om hieraan te voldoen moet iedere Europese lidstaat een optimale strategie voor het beheer van zijn wateren ontwikkelen.

Een groot deel van deze wateren kan geklasseerd worden als rivieren. Om een duurzaam beheer van rivieren te bereiken, is er nood aan een multidisciplinaire en consistente aanpak. Alleen waterkwaliteits- en waterkwantiteitsbeheer volstaan hiervoor niet. Ook de structurele kwaliteit van een rivier, evenals het directe beheer van biologische elementen en migratiefaciliteiten moeten in acht genomen worden. Evenwel zetten de ontoereikende technische en financiële middelen vaak een rem op de ontwikkeling van een geïntegreerde aanpak van het rivierbeheer. Geïntegreerde watersysteemmodellen kunnen in die context gezien woren als belangrijke hulpmiddelen om kostenefficiënte en optimale beheersopties te bepalen. Bij deze geïntegreerde modellen ontbreekt vaak nog de kwantitatieve beschrijving van de relaties tussen de riviercondities en de biologie. Het doel van dit werk is om dergelijke modellen te ontwikkelen voor een belangrijke vissoort in rivierecosystemen. Gezien de snoek als een cruciale predator in riviersystemen mag worden aanzien, werd deze soort voor dit onderzoek uitgekozen.

Sinds kort hebben ecologische modellen als ondersteuning van rivierbeheer en waterbeleid aan populariteit gewonnen. Zij hebben immers verscheidene relevante toepassingen in deze context. Ze zijn in staat een betere interpretatie van de toestand van een rivier te verschaffen. Deze modellen maken het tevens de beheerders mogelijk om het te verwachten effect te berekenen van bepaalde herstelmaatregelen en op die manier voor de meest duurzame opties te kiezen. Ze kunnen daarnaast helpen om de grootste hiaten in de kennis van riviersystemen te detecteren en hulp bieden bij het opstellen van kosteneffectieve monitoringprogramma's.

Deze thesis is opgesplitst in drie delen, nl. monitoring, modellering en beheer. Het voornaamste objectief van deze thesis was om de habitatgeschiktheid van snoek in rivierbekkens in Vlaanderen in te schatten gebruik makend van gegevengsgebaseerde modelleertechnieken, met name *classification trees*, artificiële neurale netwerken, *support*

vector machines en logistische regressie. De combinatie van deze technieken met genetische algoritmes werd ook bestudeerd. Voor de vier modelleertechnieken werd gebruik gemaakt van Weka software. Dit doctoraatsonderzoek heeft zich hoofdzakelijk op 6 rivierbekkens gericht, namelijk deze van Demer, Dender, Dijle, Ijzer, Nete en Bovenschelde. De beschikbare gegevens werden verzameld in de periode 1991-2002. De gegevens van de afvissingen werden ter beschikking gesteld door het Instituut voor Natuur en Bosonderzoek (INBO), terwijl de fysisch-chemische gegevens voornamelijk uit het monitoringnetwerk van de Vlaamse Milieumaatschappij werden gehaald. In totaal waren 20 abiotische variabelen beschikbaar om de aan- of afwezigheid van snoek mee te associëren.

Op basis van de gebruikte betrouwbaarheidscriteria, CCI en Kappa, werden de beste resultaten bekomen met de artificiële neurale netwerken en *support vector machines*. Deze twee black-box technieken hebben evenwel het nadeel dat ze moeilijk te interpreteren zijn. Bovendien werden genetische algoritmes gecombineerd met de vier modelleertechnieken om de invloed van de milieuvariabelen te analyseren, om op een betrouwbare manier de aan- of afwezigheid van snoek in te schatten en om de voornaamste rivierkarakteristieken op te sporen die de habitatgeschiktheid voor snoek bepalen. In totaal werden 14 van de 20 variabelen als relevant beschouwd om de habitatgeschiktheid voor snoek te voorspellen in de zes bestudeerde rivierbekkens in Vlaanderen. Deze variabelen zijn een combinatie van structurele variabelen en pollutievariabelen, hetgeen aantoont dat niet alleen verontreiniging een limiterende factor is voor de snoek in Vlaanderen.

De ontwikkelde modellen kunnen steun bieden aan een duurzame besluitvorming bij waterbeheer. Zij kunnen namelijk mee instaan om stakeholders te overtuigen van noodzakelijke investeringen en bijdragen aan het veranderen van een aantal maatschappelijke activiteiten. In deze context werden de modellen ook concreet toegepast om na te gaan in welke mate ze in staat zijn om praktische beheersopties te simuleren. Daarom werd een gevalstudie uitgewerkt en doorgerekend binnen het Denderbekken. De modellen wezen uit dat het landgebruik een kritische factor is voor de aanwezigheid van snoek. Hierbij dient evenwel opgemerkt te worden dat de modellen ontwikkeld werden op basis van gegevens over meer dan een decennium, waarbinnen zeer grote wijzigingen optraden in het rivierbeheer. Wellicht zijn hierdoor de modellen minder relevant voor de eerste simulatiejaren (begin jaren negentig), waarbij vooral de zware verontreiniging als dusdanig wellicht de doorslaggevende

factor was. Dit illustreert dat dergelijk gegevensgebaseerde modellen ook zeker hun beperkingen hebben in gebruik wanneer het systeem zwaar wijzigt over de beschouwde periode.

Deze thesis is opgebouwd uit verscheidene hoofdstukken: een overzicht van de ecologie van snoek wordt gegeven in een inleidend hoofdstuk, waar de noodzaak voor kwantitatieve benaderingen in het bijzonder behandeld wordt. In het eerste hoofdstuk volgt een literatuurstudie van de biologische en ecologische kennis van snoek. Hoofdstuk 2 geeft een beschrijving van het studiegebied en de methodologie. In hoofdstuk 3 wordt de methode van gegevensanalyse beschreven. In de hoofdstukken 4 tot 7 werden vier gegevensgebaseerde modelontwikkelingstechnieken toegepast, gebruik makend van *classification trees*, artificiële neurale netwerken, *support vector machines* en logistische regressie in combinatie met genetische algoritmen voor het voorspellen van de habitatgeschiktheid voor snoek in rivierbekkens in Vlaanderen. In hoofdstuk 8 werden deze 4 technieken met elkaar vergeleken om een beter inzicht te verwerven in habitatpreferentie van snoek. In hoofdstuk 9 wordt de praktische toepassing van *classification tree* modellen in de rivierbekkens beschreven. Deze thesis besluit met een algemene discussie en een voorstel tot verder onderzoek.

Er kan geconcludeerd worden dat de gegevensgebaseerde modellen kunnen bijdragen tot een beter inzicht in de interacties tussen de biotische en abiotische componenten van rivierbekkens en dat ze belangrijke informatie kunnen bieden om de besluitvorming rond rivierbekkens te ondersteunen. Evenwel dient benadrukt te worden dat dergelijke technieken pas een goede bijdrage kunnen leveren als de vereiste gegevens hiervoor beschikbaar zijn. Dit is dan wellicht ook de belangrijkste conclusie van deze studie: in Vlaanderen dient meer aandacht besteed te worden aan de verbetering van de gegevensverzameling van de aquatische ecosystemen, niet alleen voor modelontwikkeling, maar tevens ook voor ecologische beoordeling in het algemeen.

Curriculum vitae

Personal information:

Name and surname: Rahmat Zarkami
Nationality: Iranian
Date and place of birth: 29/08/1968, Rasht (Iran)
Sex: male
Married status: married, one child (son)

Contact address:

Work address:
Rahmat Zarkami, Department of environmental science, Faculty of natural resources, Guilan University, Somehsara, Guilan province, P.O.Box:1144-43619, Iran.
Tel: 0098 18232 5399
Fax: 0098 18232 2102
E-mails:
1- rzarkami2002@yahoo.co.uk
2- rahmat.zarkami@ugent.be

Diplomas and certificates:

1988-1990: H. Sc. in natural resources, Guilan University, Somehsara, Iran.
1991-1993: B. Sc. in environmental and fishery science, University of Tehran, Karaj, Iran.
Title of B. Sc. thesis: "Temporal and spatial distribution of Gammarus in the coastal regions of Caspian Sea", Tehran University, Karaj, Iran.
1993-1996: M. Sc. in environmental science, University of Tarbiat Modarres, Tehran, Iran.
Title of M. Sc. thesis: "The survey of nutrients in Anzali lagoon", University of Tarbiat Modarres (Tehran), Iran.

Publications (articles and books):

Zarkami, R and Sabet-Rafter, K. (2003). Development of eutrophiccation model for Anzali lagoon. *Iranian Journal of Pajouhesh and Sazandegi in natural resource*s **56/57**: 38-43, (in Farsi).

Zarkami, R. (2002), Nutrients conditions in Anzali lagoon, *Iranian Journal of Pajouhesh and Sazandagi in natural resource*s. **53**: 56-66, (in Farsi).

Zarkami, R. and Hassan Nia, S. (2002). Water chemistry, Guilan University press, (translated in Farsi).

Zarkami, R. (2002). Marine ecology, Guilan university press, (in Farsi).

Zarkami, R**.** (2001). Biology of game animals, Guilan university press (in Farsi).

Zarkami, R. (1997). Basic ecology, Guilan university press (in Farsi).

Conferences, presentations and contributions:

Zarkami, R., Goethals, P. L. M. and De Pauw, N. (2007). Predictive pike (*Esox lucius)* and tench (*Tinca tinca*) population models based on classification trees. Geophysical Research Abstracts, Vol. 9, 10585, SRef-ID: 1607-7962/gra/EGU2007-A-10585 © European Geosciences Union 2007, Vienna, Austria (poster presentation).

Zarkami, R., Goethals, P. L. M, and De Pauw, N. (2006). Development of predictive habitat modelling for pike and tench. B-IWA Happy hour, Brussels (oral and poster presentation).

Zarkami, R., Goethals, P. L. M. and De Pauw, N. (2005). Development of population models for pike and tench in Flanders. Rebecca conference*, Biological responses to chemical and hydromorphological pressures*, WL/ Delft hydraulics Delft, Netherlands (poster presentation).

Zarkami, R., Goethals, P.L.M. and De Pauw, N. (2005). 70(2): 1-312, Development of population models for *Esox lucius* and *Tinca tinca* in Flanders based on classification tree methods, 11th symposium on applied biological sciences, Leuven University (poster presentation).

Goethals, P. L. M., Zarkami, R., D'heygere, T., De Pauw, N., Jaarsma, N., Klinge, M., Grimm, M., Van Thuyne, G. and Belpaire, C. (2006). Status, ecology and conservation of pike (*Esox lucius*) in Flanders. Project nr.B & G/14/2001, Research report (in Dutch), 112 pp.

Vandenberghe, V., Goethals, P.L.M., Zarkami, R and De Pauw, N. (2005). Coupling water quality and fish habitat modelling for river management: simulation exercise in the Dender basin, laboratory of environmental toxicology and aquatic ecology, Ghent University, final meeting in Silkeborg, Denmark, European aquatic modelling network, COST 626.

Goethals, P. L. M., D'heygere, T., Zarkami, R., Klinge, M., Jaarsma, N., Grimm, M., Belpaire, C., Van Thuyne. G. and De Pauw, N. (2006). Restoration management of Pike (*Esox lucius*) in Flanders, Belgium. Academic day of Flemish Fisheries, Brussels, Belgium (Poster presentation).

Work and research experience:

1996-2003: Lecturer in faculty of agricultural science (Rasht), Guilan University, Iran.
1996-2003: Lecturer in faculty of fishery science (Anzali), Guilan University, Iran.
1996-2003: Lecturer in faculty of natural resources (Somehsara), Guilan University, Iran.
1996-2002: Director of library, faculty of fishery science (Anzali), Guilan University, Iran.
1997-2002: Senior researcher in environmental agencies, water pollution and eutrophication in Anzali lagoon, Iran.

Teaching experience:

1996-2003: Basic ecology, faculty of fishery science (Anzali) and faculty of natural resources (Somehsara), University of Guilan, Iran.
1996-2003: Marine ecology, faculty of fishery science (Anzali), University of Guilan, Iran.
1996-2003: Biology of game animals, faculty of agricultural science (Rasht) and faculty of natural resources (Somehsara), University of Guilan, Iran.
1996-2003: Aquatic plants, faculty of fishery science (Anzali) and faculty of natural resources (Somehsara), University of Guilan, Iran.

1996-2000: Range management, faculty of natural resources (Somehsara), University of Guilan, Iran.

Certificates based on attendance and exam:

1996: Certificate in international workshop and congress of water and air pollution, Tehran, Iran.

1997: Certificate in the first national symposia of environmental issues, Guilan province, Iran (oral presentation).

1998: Certificate in the second national symposia of environmental issues, Guilan province, Iran (oral presentation).

Attended courses:

2004: Biological monitoring and assessment of surface water quality, Faculty of agricultural and applied biological science, laboratory of environmental toxicology and aquatic ecology, Ghent University, Belgium, lecturer: prof. dr. Niels De Pauw.

2007: Sustainable development, Faculty of agricultural and applied biological science, laboratory of environmental toxicology and aquatic ecology, Ghent University, Belgium, lecturer: prof. dr. Niels De Pauw.

2007: Conservation biology, Faculty of agricultural and applied biological science, laboratory of environmental toxicology and aquatic ecology, Ghent University, Belgium, lecturer: prof. dr. Niels De Pauw.

Annexes

Appendix 1: Classification trees for pike in 6 river basins in Flanders (subset 1, PCF=0.5)

```
land-use = forest
|   Chloride <= 57.18
|   |   BOD <= 2.17
|   |   |   BOD <= 1.62
|   |   |   |   Wetted-width <= 3.5: 1 (5.01/2.03)
|   |   |   |   Wetted-width > 3.5: 0 (4.84/1.03)
|   |   |   BOD > 1.62: 1 (4.93/1.42)
|   |   BOD > 2.17: 0 (12.32/3.77)
|   Chloride > 57.18: 1 (5.89/0.18)
land-use = pasture: 1 (18.0/6.0)
land-use = industrial: 0 (0.0)
land-use = urban
|   Wetted-width <= 12: 1 (4.0/1.0)
|   Wetted-width > 12: 0 (2.0)
land-use = arable-land
|   Huet-zonation = barbel
|   |   BOD <= 2.83
|   |   |   COD <= 25.01: 1 (3.27/0.27)
|   |   |   COD > 25.01: 0 (2.73/1.0)
|   |   BOD > 2.83: 0 (6.0/1.0)
|   Huet-zonation = bream
|   |   Flow-velocity <= 0.71
|   |   |   Average-depth <= 0.9
|   |   |   |   BBI <= 6: 0 (4.64/1.07)
|   |   |   |   BBI > 6: 1 (2.95/0.99)
|   |   |   Average-depth > 0.9: 1 (9.56/3.44)
|   |   Flow-velocity > 0.71: 0 (12.86/3.86)
|   Huet-zonation = upstream: 0 (1.0)

Number of Leaves:          17
Size of the tree:   29
=== Evaluation on test set ===
Correctly Classified Instances        39              78      %
Incorrectly Classified Instances      11              22      %
Kappa statistic                       0.56
Mean absolute error                   0.3492
Root mean squared error               0.3913
Relative absolute error               69.8398 %
Root relative squared error           78.2507 %
Total Number of Instances             50

=== Confusion Matrix ===

  a  b   <-- classified as
 20  5 |  a = 0
  6 19 |  b = 1
```

Appendix 2: Classification trees for pike in 6 river basins in Flanders (subset 1, PCF=0.25)

```
land-use = forest
|   Chloride <= 57.18: 0 (27.11/11.29)
|   Chloride > 57.18: 1 (5.89/0.18)
land-use = pasture: 1 (18.0/6.0)
land-use = industrial: 0 (0.0)
land-use = urban
|   Wetted-width <= 12: 1 (4.0/1.0)
|   Wetted-width > 12: 0 (2.0)
land-use = arable-land: 0 (43.0/18.0)

Number of Leaves:          7
Size of the tree:    10


=== Evaluation on test set ===

Correctly Classified Instances        30              60      %
Incorrectly Classified Instances      20              40      %
Kappa statistic                       0.2
Mean absolute error                   0.4439
Root mean squared error               0.4745
Relative absolute error               88.7704 %
Root relative squared error           94.9073 %
Total Number of Instances             50


=== Confusion Matrix ===

  a  b   <-- classified as
 20  5 |  a = 0
 15 10 |  b = 1
```

Appendix 3: Classification trees for pike in 6 river basins in Flanders (subset 1, PCF=0.1)

```
land-use = forest
|   Chloride <= 57.18: 0 (27.11/11.29)
|   Chloride > 57.18: 1 (5.89/0.18)
land-use = pasture: 1 (18.0/6.0)
land-use = industrial: 0 (0.0)
land-use = urban
|   Wetted-width <= 12: 1 (4.0/1.0)
|   Wetted-width > 12: 0 (2.0)
land-use = arable-land: 0 (43.0/18.0)

Number of Leaves:          7
Size of the tree:    10

=== Evaluation on test set ===

Correctly Classified Instances        30              60      %
Incorrectly Classified Instances      20              40      %
Kappa statistic                       0.2
Mean absolute error                   0.4439
Root mean squared error               0.4745
Relative absolute error               88.7704 %
Root relative squared error           94.9073 %
Total Number of Instances             50


=== Confusion Matrix ===

  a  b   <-- classified as
 20  5 |  a = 0
 15 10 |  b = 1
```

Appendix 4: Classification trees for pike in 6 river basins in Flanders (subset 1, PCF=0.01)

```
: 0 (100.0/50.0)

Number of Leaves:          1
Size of the tree:   1


=== Evaluation on test set ===

Correctly Classified Instances        25              50     %
Incorrectly Classified Instances      25              50     %
Kappa statistic                    0
Mean absolute error                  0.5
Root mean squared error                0.5
Relative absolute error             100     %
Root relative squared error          100     %
Total Number of Instances              50

=== Confusion Matrix ===

  a  b   <-- classified as
 25  0 |  a = 0
 25  0 |  b = 1
```

Appendix 5: Classification trees for pike in 6 river basins in Flanders (subset 2, PCF=0.5)

```
land-use = forest
|   Flow-velocity <= 0.5
|   |   Flow-velocity <= 0.26
|   |   |   Slope <= 1.1: 1 (4.36/1.09)
|   |   |   Slope > 1.1: 0 (2.18/0.73)
|   |   Flow-velocity > 0.26
|   |   |   Huet-zonation = barbel: 0 (7.55/1.36)
|   |   |   Huet-zonation = bream: 1 (6.91/2.73)
|   |   |   Huet-zonation = upstream: 0 (1.91/0.45)
|   Flow-velocity > 0.5
|   |   BOD <= 2.83: 1 (7.58/1.89)
|   |   BOD > 2.83: 0 (5.51/2.32)
land-use = pasture
|   BOD <= 0.92
|   |   Huet-zonation = barbel: 1 (3.5/1.0)
|   |   Huet-zonation = bream: 0 (4.0/1.0)
|   |   Huet-zonation = upstream: 0 (0.0)
|   BOD > 0.92: 1 (7.5/2.0)
land-use = industrial: 0 (1.0)
land-use = urban: 0 (6.0/3.0)
land-use = arable-land
|   Average-depth <= 0.9
|   |   BBI <= 5: 0 (12.58/2.71)
|   |   BBI > 5
|   |   |   Average-depth <= 0.6
|   |   |   |   Total-Phosphorus <= 0.37: 1 (2.37/0.86)
|   |   |   |   Total-Phosphorus > 0.37: 0 (3.19/0.82)
|   |   |   Average-depth > 0.6: 1 (6.69/2.59)
|   Average-depth > 0.9: 1 (17.18/6.32)

Number of Leaves:          18
Size of the tree:   30
=== Evaluation on test set ===
Correctly Classified Instances        26              52      %
Incorrectly Classified Instances      24              48      %
Kappa statistic                        0.04
Mean absolute error                    0.4848
Root mean squared error                0.5095
Relative absolute error               96.9541 %
Root relative squared error          101.8934 %
Total Number of Instances              50

=== Confusion Matrix ===
  a  b   <-- classified as
 15 10 |  a = 0
 14 11 |  b = 1
```

Appendix 6: Classification trees for pike in 6 river basins in Flanders (subset 2, PCF=0.25)

```
Average-depth <= 0.9: 0 (66.67/30.0)
Average-depth > 0.9: 1 (33.33/13.33)

Number of Leaves:          2
Size of the tree:   3

=== Evaluation on test set ===

Correctly Classified Instances        27              54      %
Incorrectly Classified Instances      23              46      %
Kappa statistic                       0.08
Mean absolute error                   0.496
Root mean squared error               0.4987
Relative absolute error               99.2    %
Root relative squared error           99.7397 %
Total Number of Instances             50

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  0.24      0.16       0.6       0.24      0.343    0
  0.84      0.76       0.525     0.84      0.646    1

=== Confusion Matrix ===

  a  b   <-- classified as
  6 19 |  a = 0
  4 21 |  b = 1
```

Appendix 7: Classification trees for pike in 6 river basins in Flanders (subset 2, PCF=0.1)

Average-depth <= 0.9: 0 (66.67/30.0)
Average-depth > 0.9: 1 (33.33/13.33)

Number of Leaves: 2
Size of the tree: 3

=== Evaluation on test set ===

Correctly Classified Instances	27	54	%
Incorrectly Classified Instances	23	46	%
Kappa statistic	0.08		
Mean absolute error	0.496		
Root mean squared error	0.4987		
Relative absolute error	99.2 %		
Root relative squared error	99.7397 %		
Total Number of Instances	50		

=== Confusion Matrix ===

```
 a  b   <-- classified as
 6 19 |  a = 0
 4 21 |  b = 1
```

Appendix 8: Classification trees for pike in 6 river basins in Flanders (subset 2, PCF=0.01)

```
Average-depth <= 0.9: 0 (66.67/30.0)
Average-depth > 0.9: 1 (33.33/13.33)

Number of Leaves:          2
Size of the tree:   3


=== Evaluation on test set ===

Correctly Classified Instances        27              54      %
Incorrectly Classified Instances      23              46      %
Kappa statistic                    0.08
Mean absolute error                 0.496
Root mean squared error              0.4987
Relative absolute error             99.2    %
Root relative squared error          99.7397 %
Total Number of Instances            50

=== Confusion Matrix ===

  a  b   <-- classified as
  6 19 |  a = 0
  4 21 |  b = 1
```

Appendix 9: Classification trees for pike in 6 river basins in Flanders (subset 3, PCF=0.5)

```
BBI <= 3: 0 (9.76/1.78)
BBI > 3
|   Huet-zonation = barbel
|   |   BOD <= 2.58: 1 (9.47/1.58)
|   |   BOD > 2.58
|   |   |   land-use = forest
|   |   |   |   Distance-from-source <= 39.8: 0 (9.59/2.47)
|   |   |   |   Distance-from-source > 39.8: 1 (3.2)
|   |   |   land-use = pasture
|   |   |   |   Wetted-width <= 8: 1 (2.86/0.73)
|   |   |   |   Wetted-width > 8: 0 (2.0)
|   |   |   land-use = industrial: 0 (0.0)
|   |   |   land-use = urban: 0 (1.0)
|   |   |   land-use = arable-land: 0 (7.3/3.03)
|   Huet-zonation = bream
|   |   Average-depth <= 0.95
|   |   |   land-use = forest: 1 (0.29)
|   |   |   land-use = pasture: 1 (2.3/0.58)
|   |   |   land-use = industrial: 0 (0.0)
|   |   |   land-use = urban: 1 (0.29)
|   |   |   land-use = arable-land: 0 (9.29/2.22)
|   |   Average-depth > 0.95: 1 (30.15/9.96)
|   Huet-zonation = upstream
|   |   Wetted-width <= 2: 1 (4.71/0.9)
|   |   Wetted-width > 2: 0 (7.8/1.0)

Number of Leaves:          17
Size of the tree:    26

=== Evaluation on test set ===

Correctly Classified Instances          25              50      %
Incorrectly Classified Instances        25              50      %
Kappa statistic                          0
Mean absolute error                      0.4839
Root mean squared error                  0.5327
Relative absolute error                 96.7752 %
Root relative squared error            106.5351 %
Total Number of Instances               50


=== Confusion Matrix ===

  a  b   <-- classified as
  5 20 |  a = 0
  5 20 |  b = 1
```

Appendix 10: Classification trees for pike in 6 river basins in Flanders (subset 3, PCF=0.25)

```
BBI <= 3: 0 (9.76/1.78)
BBI > 3
|   Huet-zonation = barbel
|   |   BOD <= 2.58: 1 (9.47/1.58)
|   |   BOD > 2.58: 0 (25.95/10.82)
|   Huet-zonation = bream
|   |   Average-depth <= 0.95
|   |   |   land-use = forest: 1 (0.29)
|   |   |   land-use = pasture: 1 (2.3/0.58)
|   |   |   land-use = industrial: 0 (0.0)
|   |   |   land-use = urban: 1 (0.29)
|   |   |   land-use = arable-land: 0 (9.29/2.22)
|   |   Average-depth > 0.95: 1 (30.15/9.96)
|   Huet-zonation = upstream
|   |   Wetted-width <= 2: 1 (4.71/0.9)
|   |   Wetted-width > 2: 0 (7.8/1.0)

Number of Leaves:          11
Size of the tree:    17


=== Evaluation on test set ===

Correctly Classified Instances        26              52      %
Incorrectly Classified Instances      24              48      %
Kappa statistic                          0.04
Mean absolute error                      0.4775
Root mean squared error                  0.5251
Relative absolute error                 95.4909 %
Root relative squared error            105.0129 %
Total Number of Instances                50

=== Confusion Matrix ===

  a  b   <-- classified as
  6 19 |  a = 0
  5 20 |  b = 1
```

Appendix 11: Classification trees for pike in 6 river basins in Flanders (subset 3, PCF=0.1)

```
BBI <= 3: 0 (9.76/1.78)
BBI > 3
|   Huet-zonation = barbel
|   |   BOD <= 2.58: 1 (9.47/1.58)
|   |   BOD > 2.58: 0 (25.95/10.82)
|   Huet-zonation = bream
|   |   Average-depth <= 0.95
|   |   |   land-use = forest: 1 (0.29)
|   |   |   land-use = pasture: 1 (2.3/0.58)
|   |   |   land-use = industrial: 0 (0.0)
|   |   |   land-use = urban: 1 (0.29)
|   |   |   land-use = arable-land: 0 (9.29/2.22)
|   |   Average-depth > 0.95: 1 (30.15/9.96)
|   Huet-zonation = upstream
|   |   Wetted-width <= 2: 1 (4.71/0.9)
|   |   Wetted-width > 2: 0 (7.8/1.0)

Number of Leaves:          11
Size of the tree:    17

=== Evaluation on test set ===

Correctly Classified Instances          26              52      %
Incorrectly Classified Instances        24              48      %
Kappa statistic                          0.04
Mean absolute error                      0.4775
Root mean squared error                  0.5251
Relative absolute error                 95.4909 %
Root relative squared error            105.0129 %
Total Number of Instances               50


=== Confusion Matrix ===

  a  b   <-- classified as
  6 19 |  a = 0
  5 20 |  b = 1
```

Appendix 12: Classification trees for pike in 6 river basins in Flanders (subset 3, PCF=0.01)

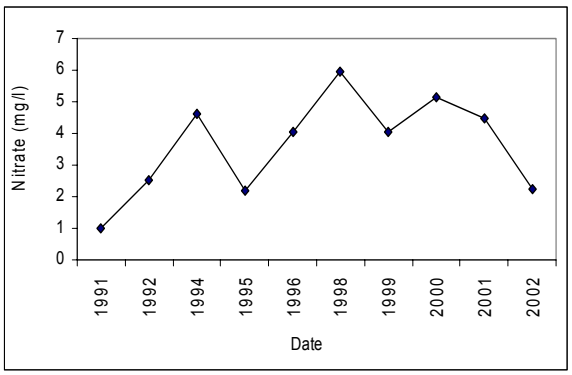
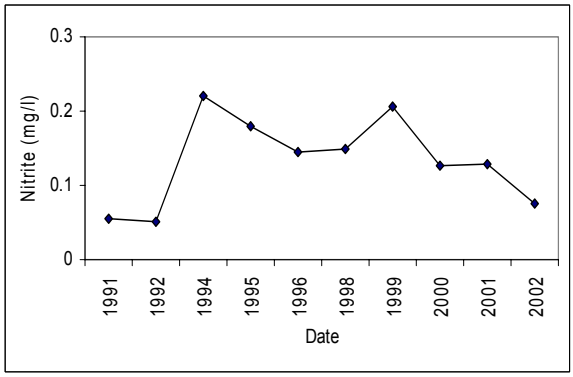
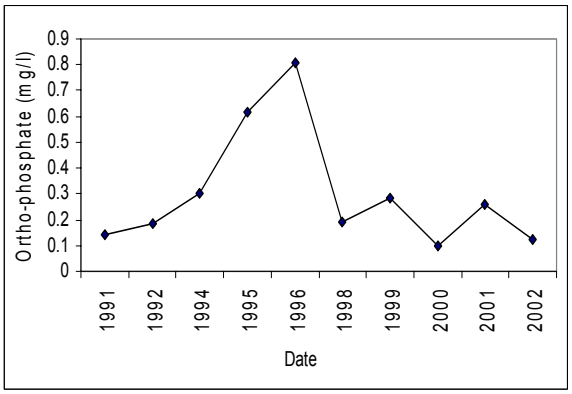
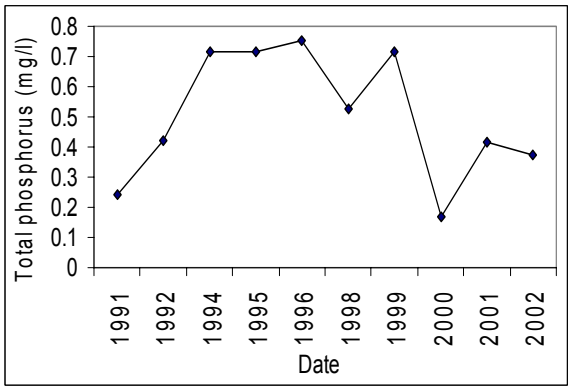
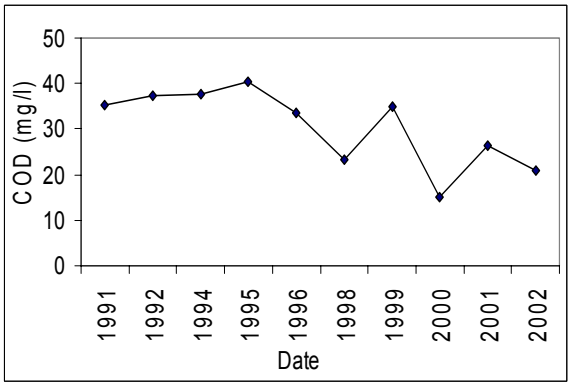
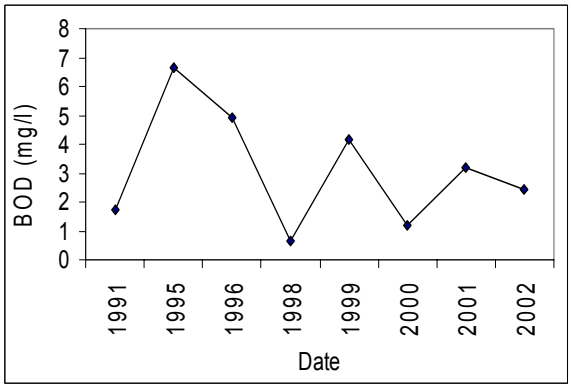
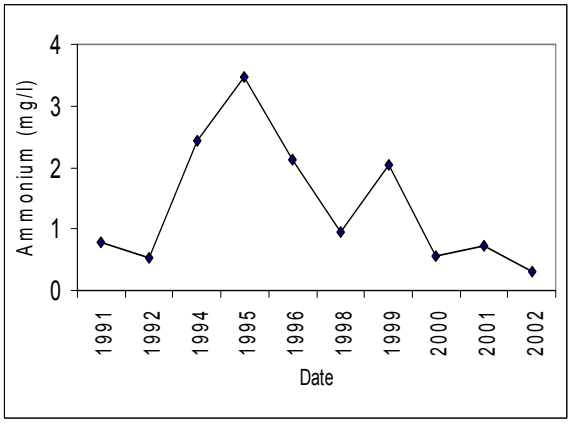
```
BBI <= 3: 0 (9.76/1.78)
BBI > 3: 1 (90.24/42.02)

Number of Leaves:          2
Size of the tree:   3

=== Evaluation on test set ===

Correctly Classified Instances        25              50      %
Incorrectly Classified Instances      25              50      %
Kappa statistic                   0
Mean absolute error                 0.5007
Root mean squared error               0.5012
Relative absolute error             100.1373 %
Root relative squared error          100.2454 %
Total Number of Instances             50

=== Confusion Matrix ===

  a  b   <-- classified as
  0 25 |  a = 0
  0 25 |  b = 1
```

Appendix 13: Temporal variation of annual average of some physical-chemical variables during the period 1991-2002 in the 6 river basins in Flanders.