CHALMERS
UNIVERSITY OF TECHNOLOGY

Modelling biodiversity in highway stormwater ponds

Master’s thesis in the Master’s Programme of Infrastructure and Environmental Engineering

RICARDO FRANCISCO HERMIDA CALVEIRO

Department of Civil and Environmental Engineering
Division of Civil and Environmental Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2014
Master Thesis 2014:127

MASTER’S THESIS 2014:127

Modelling biodiversity in highway stormwater ponds

Master of Science Thesis in the Master’s Programme Infrastructure and Environmental Engineering

RICARDO FRANCISCO HERMIDA CALVEIRO

Department of Civil and Environmental Engineering
Division of Civil and Environmental Engineering

CHALMERS UNIVERSITY OF TECHNOLOGY

Göteborg, Sweden 2014

Modelling biodiversity in highway stormwater ponds
Master of Science Thesis in the Master's Programme Infrastructure and Environmental Engineering
RICARDO FRANCISCO HERMIDA CALVEIRO

Examensarbete / Institutionen för bygg- och miljöteknik,
Chalmers tekniska högskola 2014:127

Department of Civil and Environmental Engineering
Chalmers University of Technology
SE-412 96 Göteborg
Sweden
Telephone: + 46 (0)31-772 1000

Chalmers Reproservice / Department of Civil and Environmental Engineering
Göteborg, Sweden 2014

Modelling biodiversity in highway stormwater ponds
Master's thesis in the Master's Programme Infrastructure and Environmental Enineering

RICARDO FRANCISCO HERMIDA CALVEIRO
Department of Civil and Environmental Engineering
Division of Water Environment Technology
Chalmers University of Technology

ABSTRACT

The development of road infrastructures causes great disruptions in the biodiversity of the natural areas. The Norwegian Public Roads Administration is investigating the possibility of employing stormwater ponds for compensating the loss of biodiversity due to the construction of the E39 highway. To define the guidelines for the design of biodiversity-promoting stormwater ponds, a model predicting biodiversity in stormwater ponds based on abiotic and biotic factors is needed. The literature review performed in this thesis showed that specific examples regarding biodiversity prediction models are scarce. However, several modelling approaches were described and one of them was identified as the most suitable: the Machine Learning methods. Using this approach, a model for predicting biodiversity in stormwater ponds was constructed. The model was based on the monitoring data collected during a sampling campaign performed within the NORWAT project at the Norwegian Public Roads Administration. During the sampling campaign several stormwater ponds along several major roads near Oslo in Norway were studied. Due to the different number of samples for water and sediment quality, two different models were built. In order to measure biodiversity three indices were defined: Species richness, Shannon diversity index and inverse Simpson's index. The models were feedforward Artificial Neural Networks trained with the backpropagation algorithm. The results showed that the prediction capabilities were rather poor in all the cases but one, which performed well. The two models that were built showed very similar performances. The performances were in accordance with other results found in literature. Out of the three biodiversity indices, the species richness presented the best performance. This model confirmed that the Machine Learning models can be useful for biodiversity prediction.

Key words: Highway, Stormwater, Stormwater Pond, NORWAT, Ecology, Biodiversity, Machine Learning, Artificial Neural Network.

Contents

Preface

This Master's Thesis supposes the conclusion of the Master's Programme in Infrastructural and Environmental Engineering at Chalmers University of Technology, Sweden. This project was performed at the Department of Civil and Environmental Engineering at Chalmers, in collaboration with the Norwegian Public Roads Administration, Statens vegvesen. The partnership between Chalmers University of Technology and Statens vegvesen started as a result of the project of construction of the new E39 highway in Norway. This thesis is part of this collaboration. Specifically, this Master's thesis is a part of an ambitious project that will study how to convert the new E39 highway into the first biodiversity neutral road.

Academically, this thesis is worth 30 ECTS. This project was conducted over a period of approximately 20 weeks, from January till June. The supervisor of this Master's Thesis was Dr. Ekaterina Sokolova, Division of Water Environment Technology at Chalmers University of Technology.

I would firstly want to thank my supervisor Ekaterina Sokolova who gave me the freedom to take this research through the path I found more interesting, and also for the help and support I received from her.

I also want to extend my gratitude to the members of Statens vegvesen and the Natural History Museum of Oslo for their help and contribution in this project. I would especially want to thank Sondre Meland and Turid Hertel-Aas for making the stay of James Clarke and me in Oslo more easy, productive and fun. My special thanks also to James Clarke, who helped me to find my way through the wild world of biodiversity.

Finally, I remember the help and support of my colleagues and friends. In special, Jorge and Ángela, who made the bad moments run faster, and the good moments last longer. I want to dedicate my last words to Anna, who stayed at my side always giving me the strength to finish this thesis.

Notations

AADT	Average Annual Daily Traffic
ABM	Agent based model
Ag	Silver
Al	Aluminium
ANN	Artificial Neural Network
As	Arsenic
Ba	Barium
BFGS	Broyden-Fletcher-Goldfarb-Shanno algorithm
BMP	Best Management Practice
BOD	Biological Oxygen Demand
Ca	Calcium
CCA	Curvilinear Component Analysis
Cd	Cadmium
CIRIA	Construction Industry Research and Information Association
Cl^-	Chloride
Cl	Chlorine
Co	Cobalt
Cr	Chromium
Cu	Copper
CV	Cross Validation
CWA	Clean Water Act
DO	Dissolved oxygen
$D_{Shannon}$	Shannon Diversity Index
$D_{Simpson}$	Simpson’s diversity index or Inverse Simpson index
EPA	Environmental Protection Agency
FDD	Feedforward Neural Network
Fe	Iron
FS	Principal Components of the Sediment Quality data
FW	Principal Components of the Water Quality data
GA	Genetic Algorithm
Hg	Mercury
IBM	Individual based model
ICA	Independent Component Analysis
IVS	Input Variable Selection

K	Potassium
LID	Low-Impact Development
LOOCV	Leave-One-Out Cross Validation
Mg	Magnesium
Mn	Manganese
Mo	Molybdenum
MSE	Mean Squared Error
N	Nitrogen
Na	Sodium
NaCl	Sodium Chloride
Ni	Nickel
NO_3^-	Nitrate
NORWAT	Nordic Road Water
NURP	Nationwide Urban Runoff Program
O	Oxygen
P	Phosphorus
PAH	Polycyclic Aromatic Hydrocarbon
Pb	Lead
PCA	Principal Component Analysis
R	Regression factor
S	Sulphur
Sb	Antimony
Si	Silicon
SO_4^{2-}	Sulphate
Sr	Strontium
SuDS	Sustainable Drainage System
SVR	Single Variable Regression
TOC	Total Organic Carbon
WFD	Water Framework Directive
WSUD	Water-Sensitive Urban Design
Zn	Zinc

1 Introduction

Roads and highways, as any other type of modern infrastructure, play a very important role in today's society. Infrastructure is a major contributor to the economy of a region by allowing the transport of goods and people.

The Norwegian Western Coast is nowadays connected by the highway E39. The route connects the cities of Kristiansand, in the South of Norway, and Trondheim. The road, with a distance of almost 1100 km, crosses several fjords with the use of ferries. This causes great increases of the travel time, spanning between 21 and 22 hours between the two cities. A new highway has been proposed in order to eliminate the ferries and reduce the total travel time to around 12-13 hours. Despite these advantages, the Norwegian Public Roads Administration (Statens vegvesen) is aware that the construction of a new road imposes also some damages to the environment, especially to the biodiversity. Thus, the aim of Statens vegvesen with the new highway E39 is to build a biodiversity-neutral road.

The effect that roads have on the environment is a recent area of research. However, it is already well known that road development contributes to the loss of biodiversity. There are several causes to the loss of biodiversity along the roads. One of the negative effects for biodiversity is water pollution. The European Water Framework Directive (WFD), implemented in 2003, and incorporated into the Norwegian Law in 2007, was introduced to enforce the protection of the natural water bodies, in terms of the chemical and ecological quality. Since its implementation, a great effort has been made to reduce the pollution generated by the road traffic in the surrounding water bodies. A remediation for this problem has been the installation of Best Management Practices (BMPs) along roads, especially stormwater ponds. These constructed devices eliminate the majority of the pollutants carried by the water runoff generated on the surfaces of the roads.

After the progressive increase in the number of Best Management Practices (BMPs) along roads, some researchers discovered that these systems unexpectedly support high species biodiversity (Bishop et al., 2000a, Bishop et al., 2000b, Wall, 2007, Le Viol et al., 2009, Kazemi et al., 2009, Kazemi et al., 2011, Moore and Hunt, 2012, Le Viol et al., 2012). Despite supporting biodiversity, the effect that BMPs apply to biodiversity on a regional scale has not been agreed. While some think that BMPs can cause damages to the regional ecosystems (Bishop et al., 2000a, Bishop et al., 2000b), others discuss the contribution that these aquatic systems can provide to the nature (Oertli et al., 2002, Le Viol et al., 2009, Kazemi et al., 2009, Kazemi et al., 2011). The importance of the new human-created environments are reinforced considering the progressive decrease that natural water ponds, a similar ecosystem, have experienced over the last century (Le Viol et al., 2009).

The negative effects of road construction on biodiversity can be minimized by developing measures that fight against them. But, even with the best preventions the impact on biodiversity cannot be completely avoided. Hence, in order to build a biodiversity-neutral road, it is not possible to reduce the footprint of roads to zero. Statens vegvesen, in collaboration with the Chalmers University of Technology, are investigating the possibilities to compensate the loss of biodiversity with the provision of new ecosystems with high biodiversity. For this purpose, Statens vegvesen has considered the use of BMPs (specifically stormwater ponds) as the source of biodiversity.

1.1 Aim and objectives

The main aim of the project started by Statens vegvesen and Chalmers University of Technology is to gain knowledge on the use of BMPs for the promotion of biodiversity-neutral roads. While other students have focused on the understanding of the variables involved in the process of biodiversity development, this Master's Thesis investigates how to simulate the biodiversity creation capacity of stormwater ponds. In order to achieve this aim, the following objectives have been proposed:

- Review previous literature to find which approaches have been attempted to model biodiversity.
- Analyse the measurements provided by the NORWAT project conducted by Statens vegvesen.
- Determine which of the possible modelling approaches found in literature is the most suitable for the application on highway stormwater ponds.
- Implement such model using the data collected during the NORWAT project.
- Analyse the results of the model and make recommendations on the possible applicability of the model to predict biodiversity.

2 Background

2.1 Biodiversity

2.1.1 Definition

The definition of biodiversity is not easy to give. There is not a short explanation for the concept of biodiversity in scientific literature. According to United Nations (1992), the biodiversity is formally defined as "the variability among living organisms from all sources including, inter alia, terrestrial, marine and other aquatic ecosystems and the ecological complexes of which they are part; this includes diversity within species, between species and of ecosystems". Thus, the term biodiversity applies to all the living organisms, including animals, plants and microorganisms, as well as, ecosystems and the processes happening inside them.

The concept of biodiversity tries to represent the number and frequency of all these elements in a certain area. The definition of biodiversity can be cut down to three levels: genetic, species and ecosystem diversity. The genetic diversity accounts for the number and variation of the genetic information within the organisms living in a region. The species diversity refers to the variety of organisms in a certain region. The ecosystem diversity is the frequency and variety of ecosystems (European Environment Agency, 2010).

The keystone for the assessment of biodiversity and its importance relies on the number of species. The total number of species on the Earth is to date unknown. Despite having a good knowledge of the total number of vertebrates and plants, the scientists have only been able to guess the total number of insects. Insects represent a very important share of the total amount of the already discovered species and almost the totality of the species to be discovered. To date, the total number of discovered species is under 2 million species. Nevertheless, the estimations of the total number of species is believed to range from 10 million up to 30 or 50 million (Abe et al., 1997).

The human presence and activities have endangered the biodiversity on the Earth. The consumption of both renewable and non-renewable resources has caused abrupt changes in the habitats of many species in the world. These changes, combined with the fragility and singularity of some species, have led to the extinction of many of them. The increasing population and the growth of the demand of those resources are deteriorating the biodiversity even further and at a rate that increases continuously (Winiger, 1998).

The necessity of protecting the biodiversity is not easy to justify by scientific facts. The protection of biodiversity can be understood as an exercise of responsibility of the human race (Winiger, 1998). Also, conservationists have also considered biodiversity as a resource itself to be managed for the future. As said in Winiger (1998), the uses that the society could find in the future for the variety of genes, species and ecosystems are unpredictable. Furthermore, the understanding of how ecosystems work is very limited, and the roles played by the different species are yet to be fully discovered. Hence, endangering one single species of whose importance society is not aware, can lead to unpredicted damages and high costs for us in the future (Winiger, 1998).

2.1.2 Effects of roads on biodiversity

The construction of roads is one of the contributors to the loss of biodiversity. The effects that road development causes to the environment is a field very well documented

(Andrews, 1992, Carr et al., 2002, Coffin, 2007, Forman, 1998, Forman, 2003, Seiler, 2001, Spellerberg, 1998, Trombulak, 2000). In Andrews (1992), the author classifies the harmful effects of roads on the biodiversity:

- Alteration and loss of habitats. The alteration of the orography of the region and the cut of vegetation cause the direct loss of ecosystems and contributes to the alteration of others by the modification of hydrology.
- Edge effect. The natural gradation of species habitats (ecotones) is modified by the inclusion of a road. The road causes the ecosystems to interrupt abruptly. Edge areas hinder the species with poor dispersal abilities and attract those who are more capable of invading and colonizing. Thus, the edge effect leads to areas where a few species dominate.
- Barrier effect. The inclusion of physical barriers obstructs the freedom of movement of the species, causing the cut off of vital resources as water and disrupts social organizations.
- Disturbance. Roads cause the species to avoid the areas surrounding the roads. Noise from traffic also causes certain species to abandon the area.
- Road kills. Collisions between traffic and animals crossing the roads increase mortality.
- Increased human access. The development of a new road causes an increase of the human activity in the area. This can lead to increased hunting, increase in the fires, and a notable increase of the pollution.

Improving the construction design process, paying special attention to minimizing the effects on ecosystems, can help to mitigate some of these undesired effects. Other elements of the list, such as the disturbance effect or the increase of human access are more difficult to solve. The addition of systems such as the Best Management Practices (BMPs) can help to reduce the effect of the contamination on the aquatic ecosystems around the roads. These systems receive the runoff water collected on the road pavement with the sediments and contaminants that its flow carries. The main purpose of the BMPs is to reduce the pollution load and turbidity when the runoff water reaches the natural recipient water bodies.

2.2 Best Management Practices (BMPs)

2.2.1 Definition

In recent years, there has been an increased concern about the damage that the runoff from human modified surfaces can have on the natural water bodies. First in the U.S. and afterwards in Europe, new regulations have been implemented. The U.S. was the pioneer in the development of laws and regulations for the protection of water bodies from this type of pollution (Hvitved-Jacobsen et al., 2011). The Clean Water Act (CWA) passed in 1972 was a keystone in the control and implementation of pollution control programs. Two additional programs, the Nationwide Urban Runoff Program (NURP) in 1983 and the Stormwater Program in 1990 and 1999, followed the CWA. With them, the commitment to protect the environment from the stormwater pollution was further reinforced. In Europe, the equivalent of those three programs was introduced with the name of Water Framework Directive (WFD). This set of laws and norms was introduced in 2000, and implemented in 2003. Its aim is to improve the quality of all the water bodies in Europe by 2015. The WFD has been introduced as a law in all the countries members of the EU and Norway. The WFD was implemented in the Norwegian law in 2007, and since then, work has been done in order to fulfil it.

In the urban environments, as well as in roads and highways, the soil, vegetation cover and the irregularities of the terrain are substituted by an impervious and smooth surface. This change causes an increase in the percentage of water that becomes runoff, which otherwise would evaporate and infiltrate. Furthermore, the inclusion of a drainage network with low roughness accelerates the movement of the surface water in the catchment. The combination of these two modifications leads to an increment of the quantity of the water and to the shortening of the time in which that water is discharged. Thus, the runoff is discharged to the water bodies in the form of high peaks of flow. Furthermore, over the urban surfaces, the traffic and other human activities deposit dust, sediments and garbage. This increased pollution on the surface is washed off by the runoff and transported directly into the receiving natural waters. The negative effect of the contaminants discharged into the environment is further increased by the first wash. The first wash is the peak of pollutant concentrations created by the erosion of the first and softest layer of dirt on the surfaces of the catchment. The first wash can cause severe harm to the organisms inhabiting the natural water bodies.

The Best Management Practices (BMPs) are systems designed for the mitigation of the harmful effects of stormwater runoff in urban environments. The main aims of the BMPs are the removal and reduction of the water pollutant content and the providing an increased buffer capacity for the stormwater peak flows. The BMPs are usually implemented inside the cities to collect and treat the stormwater separately from the sewage water. The BMPs are also used in the treatment and regulation of the runoff from the pavements of roads and highways.

The term for BMPs is not unique, and the notation for this type of systems has not been normalized. Some publications use such terms as Sustainable Drainage Systems (SuDSs), which is mainly used in the UK, Low-Impact Development (LID) and Best Management Practices (BMPs), which are used mainly in the U.S. and Canada, and finally Water-Sensitive Urban Design (WSUD), often applied in Australia. In this report, the term chosen for referring to these systems will be Best Management Practices (BMPs).

The design of BMPs is usually done following the guidelines marked by public institutions. Among these guidelines, some state design manuals in the U.S. and some publications from the CIRIA association can be highlighted (Woods-Ballard et al., 2007, Schueler and Claytor, 2000, Atlanta Regional Commission, 2001a, Atlanta Regional Commission, 2001b, Bishop et al., 2000a).

2.2.2 Types of Best Management Practices

BMPs are divided in different types, depending on the treatment that is provided to the water. Again, the division and the nomenclature of the different BMPs are not well defined. In The SuDS Manual by Woods-Ballard et al. (2007), 11 types of BMPs are defined:

- Filter strips: Provide treatment by infiltration and settling of particles. They are installed adjacent to big impervious surfaces as a linear structure between one of the borders of the area and a receiving water body, water collection structure or an additional BMP. They are covered with vegetation and have a width between 7.5 and 15m. The runoff is forced to move towards the filter strip where it is evenly distributed.

- Trenches: Provide treatment by infiltration or filtration. They consist of trenches filled with void-creating material, such as stones, pebbles or rubble. They can receive either lateral inflow, which is preferable, or point sources. If the soil is permeable and it is allowed, the trenches are designed to filter and infiltrate runoff water. When the soil is impermeable the trenches are designed to filter and convey water for drainage or additional treatment.
- Swales: Provide treatment by settling of particles and, in some cases, infiltration. They are linear drainage systems consisting of a vegetated channel that is used to convey water to a receiving water body or to an additional BMPs. The swales can be designed to be standard, wet or dry swales, which are differentiated by the presence of temporary, permanent or complete lack of surface water.
- Bioretention: Provide treatment by particle removal and filtration. They consist of shallow depressions installed in small catchments with predesigned soil profiles and vegetation to provide improved treatment of water. Usually, the bioretention areas are drained with perforated pipes. This BMP design allows the effective reduction of both volume and rate of runoff.
- Pervious pavements: Provide treatment by filtration and infiltration. They consist of pavements designed to allow the infiltration of the surface water to lower layers of the pavements, in which the water is stored. After storage the water can be infiltrated into the soil, transferred to an additional BMP or discharged to a water body. The main disadvantage of this system is the limitations that this type of pavements applies to the traffic, lowering speeds and axle loads.
- Geocellular or modular systems: These systems provide no water quality treatment. The main purpose of these systems is to store runoff water for posterior infiltration or conveyance to an additional treatment step. The geocellular systems are formed by smaller modules of plastic material with a very high void ratio that are installed underground. These systems provide a cheap and easy to install method for runoff control that can be used under pedestrian and street pavements, or under public open spaces.
- Sand filters: Provide treatment by filtration. The sand filters are structures in the form of boxes that contain sand as a primary filter medium. This configuration provides great pollutant removal and is mostly used when high pollution is expected in the runoff, due to the high cost of installation and maintenance. The capacity for runoff peak flow and volume reduction is rather limited in the sand filters, by only allowing a small amount of ponding in top of filter.
- Infiltration basins: Provide treatment by infiltration. They consist of large vegetated depressions that provide enough volume for storage and surface for infiltration. The fact that the system infiltrates the runoff water excludes its use in locations where groundwater is vulnerable. The design has a low cost both on construction and installation but it is very surface demanding.
- Detention basins: Despite providing some water treatment by particle settling, the detention basins are mainly used for peak flow reduction. The detention basins are surface depressions that provide some buffer volume for the incoming runoff water. Most of the detention basins remain dry between rain events, filling up with water during these events. There can be

a small permanent pool that can help to prevent re-suspension of sediments. Infiltration is not allowed in these systems.

- Ponds: Provide treatment by sedimentation and biological uptake. The ponds are constructed depressions with a permanent pool of water. The ponds work by storing and treating the runoff water between consecutive rain events. The existence of a permanent pool of water allows the growth and development of plant and animal life, which help in the treatment of the pollution both by stimulating the sedimentation of particles and the uptake of nutrients. The design of the ponds includes usually a smaller basin for pre-treatment. This slam-basin allows the sedimentation of coarser sediments and other type of large residues. The shoreline of the ponds is designed for the support of vegetation, which is done by decreasing the slope of the bench or preventing the consolidation of the soil.

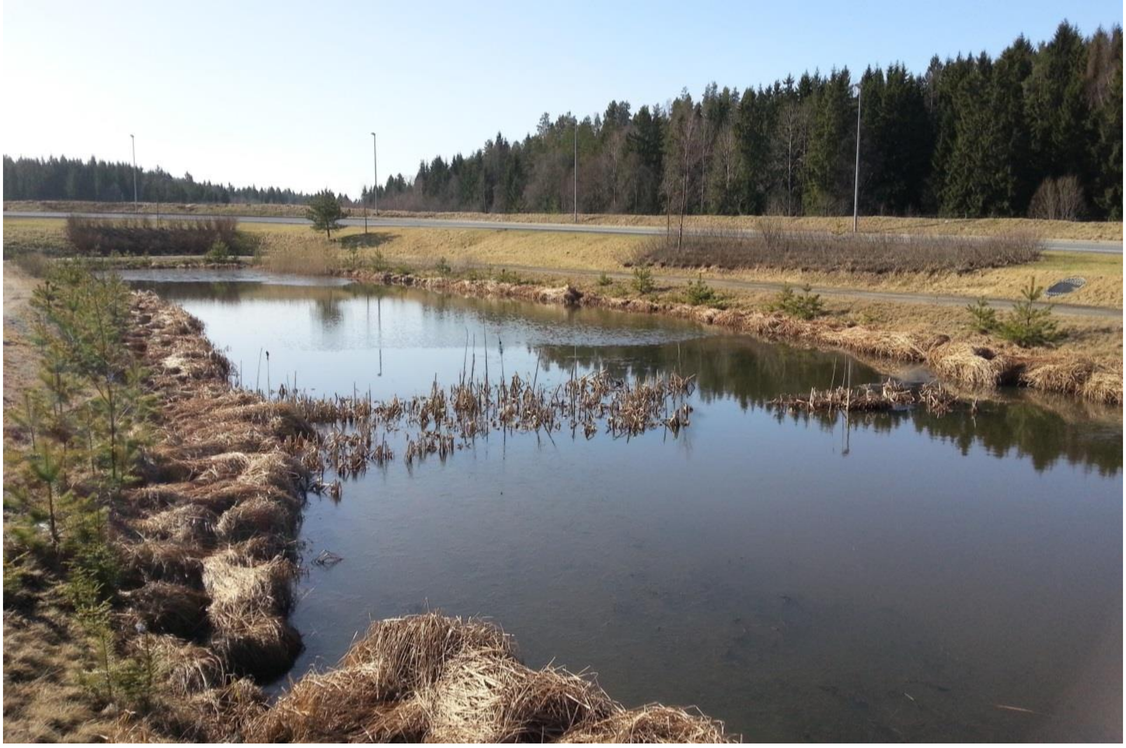


Figure 1. Pond at Taraldrud Junction (59.79703555; 10.84086138) near Oslo (Norway).

- Stormwater wetlands: Provide treatment by sedimentation and biological uptake. The stormwater wetlands combine shallow ponds with parts almost completely covered with vegetation. The wetlands are designed to allow long retention periods, time enough for sediment settling and aerobic decomposition of nutrients. This BMP method requires great extensions of land, which in many cases could not be available. On the other hand, the great areas occupied by the wetlands allow a big buffering capacity for extreme events, which in other cases would not be treated. Furthermore, the maintenance cost of the stormwater wetland can be quite low, once the system is established.

2.2.3 Effect of BMPs on biodiversity

The selection of the type of BMP most suitable for an individual case is usually based on five different criteria. According to Woods-Ballard et al. (2007), these five factors are:

1. Land use characteristics.
2. Site characteristics.
3. Catchment characteristics.
4. Quantity and quality performance requirements.
5. Amenity and environmental requirements.

The last of this criterion, *Amenity and environmental requirements*, includes the capacity of habitat creation as one of the factors to consider. The use of BMPs is known to provide an improved area for wildlife. Not all the BMPs are equal in their habitat creation capacity. The ponds and wetlands, when well designed and maintained, provide an excellent habitat for vegetation and aquatic life. Also, the presence of grass strips in swales, trenches or filter strips can be used as green corridors connecting two habitats (Woods-Ballard et al., 2007).

This capacity for habitat creation and biodiversity boosting is a very recent field of study. Bishop et al. (2000a) and Bishop et al. (2000b) are some of the first documents that cover this aspect of the wetlands and ponds. In these two documents, the authors study the biodiversity of 15 stormwater ponds and 1 wetland in order to determine the contribution of the wetlands to the creation of habitats for wildlife.

Despite that these two studies found that the ponds' created habitats were low quality with low species richness, several studies appeared in recent years, agreeing that the biodiversity capacities of constructed wetlands and ponds should not be underestimated. In a study conducted on 6 stormwater ponds in south-eastern France, the results indicated that the ponds' biodiversity was very similar to that found in natural water ponds in the same region (Scher et al., 2004). The authors stated the high benefit of the highway stormwater ponds, not for providing a better quality habitat for species, but for increasing the number of ponds when the number of natural ponds has decreased progressively during the last century. The same conclusions were also drawn by Scher et al. (2004) in a study performed in 25 ponds along 56 km of a highly used highway in France. More recent studies have even explored the possibilities of the use of wetlands and stormwater ponds for carbon sequestration by organic uptake of plants and animals (Moore and Hunt, 2012).

Despite of mainly being focused on stormwater ponds and wetlands, the study of the biodiversity benefits of BMPs has also been extended to bioretention basins, swales and filter strips. In two different studies, Kazemi et al. (2009) and Kazemi et al. (2011), the authors concluded that the construction of such structures generated a positive effect on the biodiversity of the areas where they were installed.

Given the positive effect of the highway stormwater ponds and wetlands, the study field has now moved towards the identification of the factors that contribute to and harm the development of wildlife in these ponds. Increasing our knowledge in the factors affecting biodiversity could lead to development of new enhanced designs for ponds and wetlands. With these new designs, the stormwater ponds could not only reduce the pollution of the receiving waters, but also increase the biodiversity and species richness in the region.

2.3 Factors affecting biodiversity in BMPs

The factors affecting biodiversity can be divided into two groups: abiotic and biotic factors. Abiotic factors are non-living physical and chemical properties of the environment that affect the ability of survival of the organisms present. Biotic factors are the effect that the living species in a habitat impose to the rest of fellow species.

The study of the factors that affect biodiversity in BMPs is a very recent field of study among biologists. There are few research publications specifically regarding the study of the biodiversity factors in BMPs (Bishop et al., 2000a, Bishop et al., 2000b, Le Viol et al., 2009, Le Viol et al., 2012, Thygesen, 2013, Clarke, 2014). In many cases, the study is limited to the analysis of the differences found between natural and artificial water ponds (Le Viol et al., 2009).

In this section, the classification of factors reviewed by Clarke (2014) will be followed, describing the effects that several abiotic and biotic factors have on the environment and, particularly, on the biodiversity of BMPs and, particularly, stormwater ponds.

2.3.1 Abiotic factors

According to Clarke (2014), ten different abiotic factors can be identified for having a major responsibility for the biodiversity of stormwater ponds. Half of the factors account specifically for water quality properties.

- Salinity: The use of salts, specially sodium chloride (NaCl), is common along roads as de-icing agent during winter time (Le Viol et al., 2009, Seiler, 2001). As a result, high levels of salinity have been found in stormwater ponds. The effect of salinity on biodiversity depends greatly on the species involved and the sensitivity and tolerance of these species to salt (Snodgrass et al., 2008). However, in the vast majority of the cases, such effect is of negative consequences, and tend to affect the most those species categorized as sensitive (Snodgrass et al., 2008)

- Conductivity: Conductivity is also a factor related to the presence of salts in the water. As with salinity, conductivity values are often found to be greater in stormwater ponds than in natural ponds. The effect of high conductivities on biodiversity has not been completely elucidated (Clarke, 2014). However, the effect seems to be mostly negative.

- pH: The pH values that are found in stormwater ponds are often different that those measured in natural water ponds. The most accepted theory states that the pH in stormwater ponds is lower than in natural ponds because of the much lower presence of vegetation litter, which decomposes generating humic acids (Le Viol et al., 2009). The difference of pH values has not to be found to affect biodiversity importantly (Clarke, 2014).

- Nitrogen oxides: The presence of nitrogen oxides has been found to be higher in stormwater ponds than in natural ponds. The source of this nitrogen oxide surplus is caused mainly by traffic emissions and agricultural fertilizers. Very high levels of nitrogen oxides can lead to eutrophication of the waters, which has very negative consequences for the biodiversity. Hence, despite that moderate levels of nitrogen oxides have no clear effect on biodiversity levels, the presence of nitrogen oxides should be controlled (Clarke, 2014).

- PAHs and heavy metal accumulation: The impact of Polycyclic Aromatic Hydrocarbons (PAHs) and heavy metals on biodiversity is a branch of research that it is being explored today. Out of the many PAHs and heavy metals existing in the environment, only a few have been analysed in detail. Thus, the source of heavy metals and PAHs has been found to be vehicle combustions and by-products of traffic (Sternbeck et al., 2002) The conclusion tends to determine that the presence of elevated levels of both heavy metals and PAHs has a negative impact on biodiversity (Clarke, 2014).
- Average Annual Daily Traffic (AADT): AADT measures the volume of traffic that a specific road holds. Several studies relate biodiversity with AADT (Thygesen, 2013, Clarke, 2014). However, in other cases, the effect of AADT is decomposed into the individual factors that are consequence of the volume of traffic (Clarke, 2014)
- Basin size, depth and shape: The size, depth and shape of the pond have a great effect on biodiversity. The size of stormwater ponds is apparently positively correlated with biodiversity. Regarding depth and shape of the stormwater ponds, researchers tend to indicate that ponds should present variety of slopes and depths to accommodate as many species as possible (Clarke, 2014).
- BMPs substrate type: The type of substrate base of stormwater ponds has been identified as a possible factor for biodiversity development. In general, the use of a natural base types increases the levels of biodiversity (Le Viol et al., 2009).
- Age: Several studies point out a positive correlation between the age of the stormwater pond and the levels of biodiversity present in the pond. Le Viol et al. (2009) found that old enough stormwater ponds could hold as much biodiversity as natural ponds. This is considered to be caused by a progressive naturalization of the pond (Clarke, 2014).
- Noise: Noise is an effect that is present during the various processes of road development, from construction to final use. The impact of noise on some species has been long studied, and in any case the effect appears to be negative (Coffin, 2007).

2.3.2 Biotic factors

The biotic factors are the living components of an ecosystem that affect the ecosystem and the rest of organisms existing in that environment. Despite the numerous possible biotic factors, two are identified by Clarke (2014) as most relevant: vegetation and human influence.

- Vegetation: Vegetation is a key component of any ecosystem. This key role entails, as well, the great importance of vegetation on the biodiversity of an ecosystem. The own plant biodiversity constitutes a relevant factor affecting biodiversity. Thus, the greater the diversity of vegetation, the more diverse the number of taxa it can support (Clarke, 2014). However, the presence of vegetation can affect also negatively some species while increasing levels of a different one (Clarke, 2014). Thus, special attention must be paid to balance the contributions of different vegetation ecosystems in stormwater ponds. In this situations, researchers have found that the preservation of natural woodland contributes positively to achieving high biodiversity levels (Clarke, 2014). In essence, vegetation is generally regarded as a key factor for biodiversity provision. Vegetation can not only be effectively used for increasing biodiversity in a stormwater pond, but also for promoting specific desired species.

- Human influence: It has been long demonstrated that improved accessibility of natural spaces by humans tend to create important disruptions and increased levels of pollution (Andrews, 1992). The consequences of human presence on biodiversity tend to be negative. The impact of human influences on stormwater ponds are difficult to account, but there is evidence that the effect is slightly negative (Clarke, 2014).

3 Review of ecological modelling approaches

3.1 Ecological modelling

The first ecological model ever recorded corresponds to Lotka-Volterra and their predator-prey equations, and Streeter-Phelps, with their study of water pollution relating Dissolved Oxygen (DO) and Biological Oxygen Demand (BOD), both of them in the 1920s. Nevertheless, the greatest development and explosion of ecological modelling did not occur until the 1970s. During this decade and the following, ecological models evolved answering questions never solved before. The development of such models carried also problems and difficulties on how to represent real world topics into the scheme of a model.

According to Jørgensen (1999), the main problems that the ecological modelling have been facing since its beginning and that are to be solved yet are:

a. The scarcity of sampling data for its use in model development.
b. The parameterization of the model, which is usually the weakest point.
c. The complexity of the real world problems is not fully represented by current models.

From this short list, the third point is the only one that ecological modellers can participate in its improvement. In recent years, new types of models have been developed in order to address some of these issues, such as the modelling of spatial problems, and the development of dynamic models. Regarding the two first points, the major work lays over the modeller by increasing the frequency and resolution of the measurements to be applied in their models. However, some new model types have been created or translated from different study areas into the ecology field allowing the user to represent the reality prioritising the economy of data over the details included in the model.

The range of topics that the ecological modelling covers is huge. Due to this, several different types of models have been developed during the last 40 years. These models differ in many aspects such as type of available data, type of problem, type of required resolution … It is not possible to study the changes in the distribution of a certain species of trees in a natural park and the evolution of the population of a water pond in time with the same type of model (Mladenoff and Baker, 1999). These two examples have two completely different aims and, however, both of them can be referred to as ecological models. Also, they do not share most of the variables and parameters, and, of course, the output of the model is completely alike. Thus, the aspects that distinguish one type of model from the others will be further discussed in this chapter.

3.2 Model approaches in ecological modelling

In a review of all the publications in the journal *Ecological Modelling* (Salski, 2006, Jørgensen and Fath, 2011), the authors distinguished nine different types of models. The list of ecological model types created by Jørgensen and Fath (2011) focus mainly on the quality and quantity of available data, and on the key feature the model seeks. A definition and a list of pros and cons of each model, as well as the most suitable scenario for each of the models were included. The list of the nine models considered is presented below accompanied by a brief description of each of them.

Despite distinguishing these nine model classes, the authors also recognised the existence of hybrid models combining two or more of these basic types. These hybrid models mostly are a combination of biogeochemical models with another model (Jørgensen and Fath, 2011).

3.2.1 Dynamic biogeochemical

This type of model is the most widely applied of the nine, with an application rate of 32% (Jørgensen and Fath, 2011). The aim of the dynamic biogeochemical models is to represent the biogeochemical and geochemical processes occurring in a certain environment. The processes are computed dynamically by the use of differential equations. They are based on mass or energy conservation principles and usually based on causality. The dynamic biogeochemical models are often useful as a prediction tool and are easy to understand, interpret and develop. The main disadvantage is the high number of parameters required when the model becomes slightly complex. Hence, a large and good quality set of data is required for calibration.

3.2.2 Steady-state biogeochemical

A steady-state biogeochemical model, as the dynamic biogeochemical mode, is based on the representation of the biogeochemical and geochemical processes happening in the study case but simplifying them disregarding the effect of time. This results in an easier to build and calibrate model, which can provide useful results for worst-case and average scenarios (Jørgensen and Fath, 2011).

3.2.3 Population dynamics

In this category fall all the models that represent the evolution of the population or populations of different species sharing the same space. Population dynamics study how a certain population distributes, grows or interacts with other species. The main processes that are represented in this kind of model are factors as the natality, the mortality or predation. The models can be built using deterministic or stochastic approaches, being the former category the most broadly used. The main disadvantages are the difficulty of the calibration and the need of a good and homogenous database (Jørgensen and Fath, 2011).

3.2.4 Structurally dynamic

The structurally dynamic models are a very complex model type that can be used for an accurate representation of ecosystems. When studying an ecosystem, two approaches can be taken, reductionism and holism. The first one aims the analytical study of all the processes underlying behind the apparent behaviour of the ecosystem. Holism states that it is impossible to represent all these processes analytically, both because of the amount of them and because of the high level of interaction existing. Because of this, the holism approach looks at ecosystems trying to examine the whole system and the most relevant reactions of it. Thus, structurally dynamic models focus on the most relevant processes occurring in an ecosystem as a whole, rather than as the sum of smaller processes of different species. Hence, the adaption and the changes in the species composition play a very important role in structurally dynamic models.

The evolution of the ecosystem is usually achieved by defining a goal function to which the ecosystem is forced to adapt. This goal function is often a unit of energy, like the exergy and eco-exergy variables. The exergy of an ecosystem can be defined as the difference of energy an ecosystem presents against a reference condition, normally established by the surrounding environment. The complex definition of this goal function is one of the main disadvantages of this model kind. Other disadvantages include the lack of specific software to develop the model, the high resource consumption of the model when it is run and the need of data describing the structural changes aimed by the model (Jørgensen and Fath, 2011).

3.2.5 Fuzzy

Fuzzy models are based on fuzzy algorithms. This type of algorithms developed by Zadeh (1968) are designed to process uncertain or incomplete data. The fuzzy set theory defines fuzzy sets as elements that are classified by a grade of membership intermediate between full and none membership. This theory differs with the classic set theory because the fuzzy sets have not sharp boundaries. This difference can be exemplified with the definition of colour. In classic set theory, the shades of a certain colour are defined as specific names such as, in the case of blue, baby blue, light blue, dark blue, navy blue… On the other hand, the fuzzy set theory would define all this colours as *more or less blue*. The definition of the statement *more or less blue* is not immediate and even more difficult to define. It is not a closed boundary or definition, but a fuzzy set, which contains all the members of the group *blue*. Fuzzy models allow to use fuzzy sets to build statements and algorithms for its application on regular logical and arithmetical operations (Salski, 2006).

The use of fuzzy models is quite suitable to ecological modelling. Ecological data is usually heterogenic and uncertain, and many times the only available knowledge is subjective or expert knowledge. Fuzzy models can be easily combined with quantitative information to provide reliable results with poor or incomplete data. The main inconvenient of fuzzy models are the lack of specific software, the absence of a precise numerical result or the limited complexity of the models build with this methodology (Jørgensen and Fath, 2011).

3.2.6 Artificial Neural Networks

The Artificial Neural Networks (ANNs) have their origin on the idea of the neural networks present on the human and animal brains. The millions of nerve cells present on the brain are interconnected forming groups with different fashions and number of members. The connection or synapses between neurons are the main responsible of the process of learning in our brain. The ANNs have translated this idea into the computer science.

In an ANN, a neuron is defined as a nonlinear, parameterized, bounded function (Dreyfus, 2005). If the function is of linear order, it is called linear neuron. Normally, the parameterization of the function is performed by two methodologies: parameterizing the inputs by including a weight to each of them or by parameterizing the nonlinearity of the function included in the neuron (Dreyfus, 2005).

A neural network can also be classified by the type of connections existing between two different neurons. The Feedforward Neural Networks (FNNs) are sets of neurons connected together that process inputs that at the same time are composition of the functions of its neurons. Information in this type of networks only can flow from the input to the output. The networks are formed by two or more layers of neurons. Each of the neurons in a layer is connected to every neuron in the previous layer and, if desired, to neurons in previous layers. The neurons of the last layer are called output neurons and the neurons in the intermediate layers are called hidden neurons (Dreyfus, 2005). The recurrent neural networks are ANNs where the connection between neurons can be cyclic, meaning that the information available in a certain layer can be returned to a previous neuron layer. The definition of recurrent neural networks imply the definition of an additional dimension such time, since a neuron can only have one value at a certain time. Hence, each connection in a recurrent neural network is assigned a time step or delay. In order to present causality, each cycle of connections must have at least one connection with zero delay (Dreyfus, 2005).

The ANNs are black-box models in which a set of inputs is connected with a set of layers of neurons that process the information to provide a certain output or set of outputs. As any other black-box model, the ANNs need a process of training. The training of an ANN consists of the estimation of the parameters of every neuron in the network. With a supervised training, the ANNs can be used for the approximation of any given function in a finite region of space, given that the function is bounded and sufficiently regular (Dreyfus, 2005).

By the process of training and the use of nonlinear functions, the ANNs can find relations between variables out of a heterogeneous database. This, of course, means that the result of the model will not present a relation of causality. The model will behave just as a black-box. Furthermore, the use of nonlinear functions allows this method to be quite parsimonious, this means that it will work sufficiently well despite the quantity and quality of the database. Another benefit of the ANNs is the ease of implementation and use compared with the rest of models. On the other hand, the main disadvantage of the model is that the capacity of prediction of the model will be very limited. Therefore, if a sufficiently broad and homogeneous database is available, the use of other method based on causality should be considered (Jørgensen and Fath, 2011).

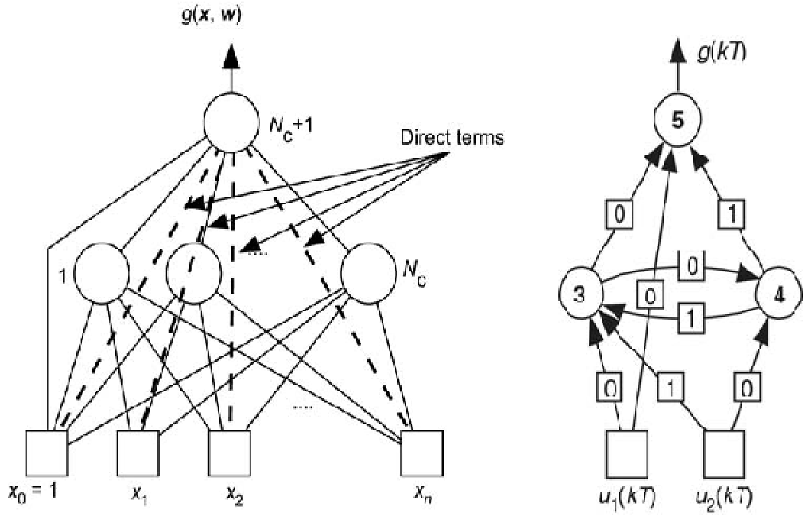


Figure 2. Schematic description of a feedforward neural network (at the left) and a recurrent neural network (at the right). The feedforward neural network has an output g(x,w) that depends on the input vector x and the neuron layer N defined by the weights w. The recurrent neural network have an output g(kT) that depends on the input vector u, the weights w and the time unit T. The boxes in the graph represent the delay in the connections (Dreyfus, 2005).

3.2.7 Spatial

There are many cases in which processes in nature present important differences in space. For example, movement, dispersion and distribution of species (Jørgensen and Fath, 2011). Spatial models are models in which the spatial dimension of those processes is accounted. The processes and variables accounted in spatial models are similar to the rest of the models presented, with the difference that the spatial dimension and time are always considered.

The addition of supplementary dimensions is, in many cases, the only method to represent and study some specific processes and conditions. However, as a general rule, the inclusion of a new dimension means adding complexity to a model. The addition of spatial dimensions increases the number of parameters and, hence, requires more knowledge of the processes (Jørgensen and Fath, 2011). Thus, the main disadvantage of spatial models is the requirement of large databases that are employed for a more difficult calibration and validation of the model. With this extra difficulties, the spatial models are reserved to studies in which the spatial distribution is the key variable and in which its consideration is crucial for the analysis (Jørgensen and Fath, 2011).

3.2.8 Individual-based or agent-based

Some areas of research in ecology cannot consider an ecosystem from a holistic approach as the structurally dynamic models do. Sometimes, the individuals of one or two species in an ecosystem are the centre of that investigation. An individual based model is a model that focuses on the behaviour and interaction of the members of one or more individual species in a system. In an agent-based model, all the individuals, or agents, differ from each other and from themselves in time as the life cycle continues. This constitutes the main difference between individual-based models and the previously described population dynamics models, in which all the individuals of the same species are defined uniformly. The properties of each agent in the system determine how it interacts, both with the other agents and with the environment (Jørgensen and Fath, 2011).

The rules set for every agent can be simple or very complex. However, independently of the complexity of the rules defining the agent behaviour, the agent-based models seek for a higher degree of behaviour. The aim of agent-based models is to obtain a complex behaviour of a system from the combination of simple rules assigned to individuals, which can interact between themselves and the environment (Bandini et al., 2009).

The definition of the agent is, therefore, key for the development of this model type. An agent is referred in modelling to a software entity with some level of autonomy and with a certain capability to learn from the interaction with other agents and the environment (Politopoulos, 2007). The construction of an agent has to consider the agent behaviour, the agent-agent interaction and the environment. First, the behaviour of the agent has to be adaptive. In other words, the agent has to come with different behaviours depending on the conditions of the system, to produce situation-specific decisions (Jørgensen and Fath, 2011). The approach to this behaviour can be in the form of deliberate or reactive decisions (Bandini et al., 2009). Reactive agents are simple agents that adjust their behaviour directly and automatically from the condition of other agents and the environment, while deliberative agents produce behaviours not only based on the present conditions but also from their knowledge and from past experiences.

Finally, these models are most suitable for the simulation of systems where the individuality of the members of the species in that system plays a key role. Nevertheless, the agent-based models have as an inconvenient that the definition of a sufficiently advance behaviour of the agent can be very complex and that the database required for setting up and validating the model must be large.

3.2.9 Ecotoxicological

Ecotoxicology models are a class of models that is used in ecotoxicology research. Their main differences with other models relay in the limited information and knowledge of the parameters, the use of safety factors and the inclusion of an effect component (Jørgensen and Fath, 2011).

Table 1. Recommended data set characteristics and most suitable problem for the main nine ecological models distinguished by Jørgensen and Fath (2011)

Model type	Data set recommended	Problem studied
Dynamic biogeochemical	High quality, homogeneous	Exchange of matter/energy
Steady-state biogeochemical	Low quality, homogeneous	Exchange of matter/energy
Population dynamics	High quality, homogeneous	Population dynamics
Structurally dynamic	High quality, homogeneous	Structural changes and adaptation are significant
Fuzzy models	Uncertain or only-rules data	Any
Artificial Neural Networks	Medium-high quality, heterogeneous	Any
Spatial	High quality	Spatial differences
IBMs[1] or ABMs[2]	-	Individuality is relevant
Ecotoxicological	-	Toxic substances, distribution and effect

[1] IBM: Individual based model

[2] ABM: Agent based model

4 Methodology and model theory

In this chapter, the methodology and proceedings followed for the acquiring of the final results and subsequent conclusions are explained. The purpose of this chapter is to explain in full detail all the steps undertaken in order to provide the resources for being able to replicate the experiments and obtain results comparable to the ones extracted in this Master Thesis.

4.1 Data collection

In this section a brief description of the sampling methodology is provided. The measurements employed in this Master's Thesis were not made by the author. The sample data used for the model of this project was obtained by Thygesen (2013). For her thesis a four-month sampling campaign was performed. A more detailed description of her sampling methodology can be found in her thesis.

The NORWAT project has been simultaneously taken samples in a project that will extend in duration for more than those 4 months. During the development of this Master's Thesis, the NORWAT group proposed the participation in one of the sampling campaigns. The NORWAT sampling campaigns are usually undertaken in a single day, during which all the studied stormwater ponds are visited. The methodology that is followed by NORWAT differs slightly from the sampling procedures made by Thygesen (2013).

4.1.1 Chemical sampling

The chemical sampling has been performed with a multi-parameter water quality sonde. The sonde that has been used for the measurements is the model 6600V2-4 from the company YSI. A multi-parameter quality sonde is an instrument equipped with sensor that directly registers measurements of several quality parameters from the water source. This specific model has the capacity to detect Dissolved Oxygen concentration, conductivity, salinity, temperature, pH, turbidity, nitrate, ammonia, and chloride, along other parameters. The sonde cannot measure metal concentrations of water.

The measurements are performed in the surroundings of the intake of the stormwater ponds. The sensor of the multi-parameter sonde is introduced in the water of the pond. The measurements taken by the sensor are continuous, and can be checked in the screen of a small handheld controller. This controller serves, as well, as a memory stick in which the results of all the measurements are stored. The procedure can be observed in Figure 3.

During the sampling of the stormwater ponds, no other water quality samples were taken. However, during Thygesen (2013) campaign, the water quality sampling was complemented with several water samples. The water samples were analysed by ALS Laboratory Group (Oslo) in order to measure the concentration of metals, oils and Polycyclic Aromatic Hydrocarbons (PAHs) in water. The metals that were analysed were Al, Sb, As, Ba, Cd, Ca, Cr, Co, Cu, Fe, Pb, Mg, Mn, Hg, Mo, Ni, P, K, Si, Ag, Na, Sr and Zn. Three anions were also analysed, chloride (Cl^-), nitrate (NO_3^-) and sulphate (SO_4^{2-}). Also the total organic carbon (TOC) was measured.

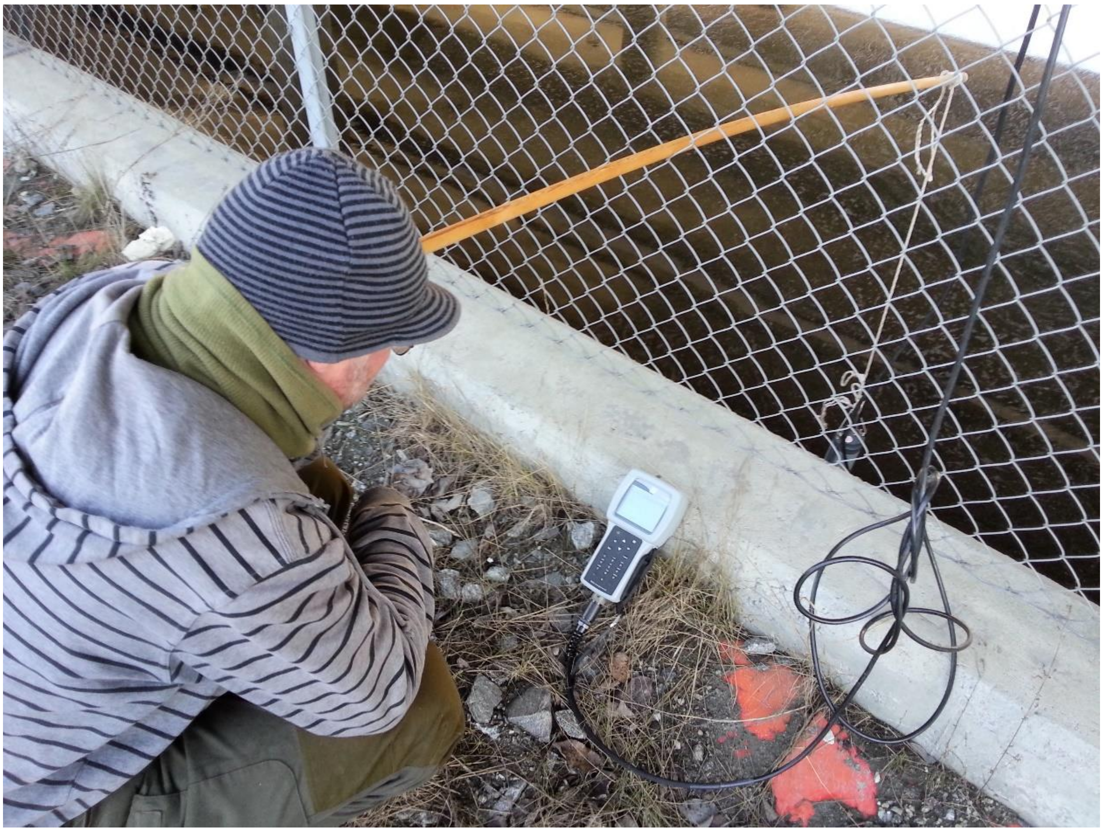


Figure 3. Picture of the water quality sampling with a multi-parameter water quality sonde.

4.1.2 Biological sampling

The purpose of the biological sample is to determine which species of organisms live in the stormwater ponds and measure the abundance of each of the species. To guarantee that the sampling is representative of the reality, between two and three samples are taken in different sites within the same pond.

The samplings are performed with a kick net. The specifications of the kick net are according to Thygesen (2013) of an opening of 30x30 cm and a mesh size of 0.45 mm. The procedure of the sampling is the same in any case. The net is introduced in the water at one side of the pond. The net is swept five times in the same spot, trying to cover all the depth. The procedure can be observed in Figure 4. In the case of the presence of ice on the pond's surface, the place with the thinnest ice cover is sampled. The ice cover is broken and the pieces of ice are removed so not to enter in the net.

The organic material caught with the net is then placed in a sampling tray. To preserve the samples for the posterior analysis, ethanol is added to the organic material and the solution is poured into a plastic bag. The bags containing the biological samples can then be stored without risking the validity of the sample. The process can be seen in Figure 4 and Figure 5.

While the biological samples taken by Thygesen (2013) were almost analysed immediately, the samples taken by NORWAT have not been analysed yet. The biological samples taken by NORWAT will be sorted to the species level when possible. In Thygesen (2013), due to time restrictions, the samples were sorted to family level, and in some cases to species level.

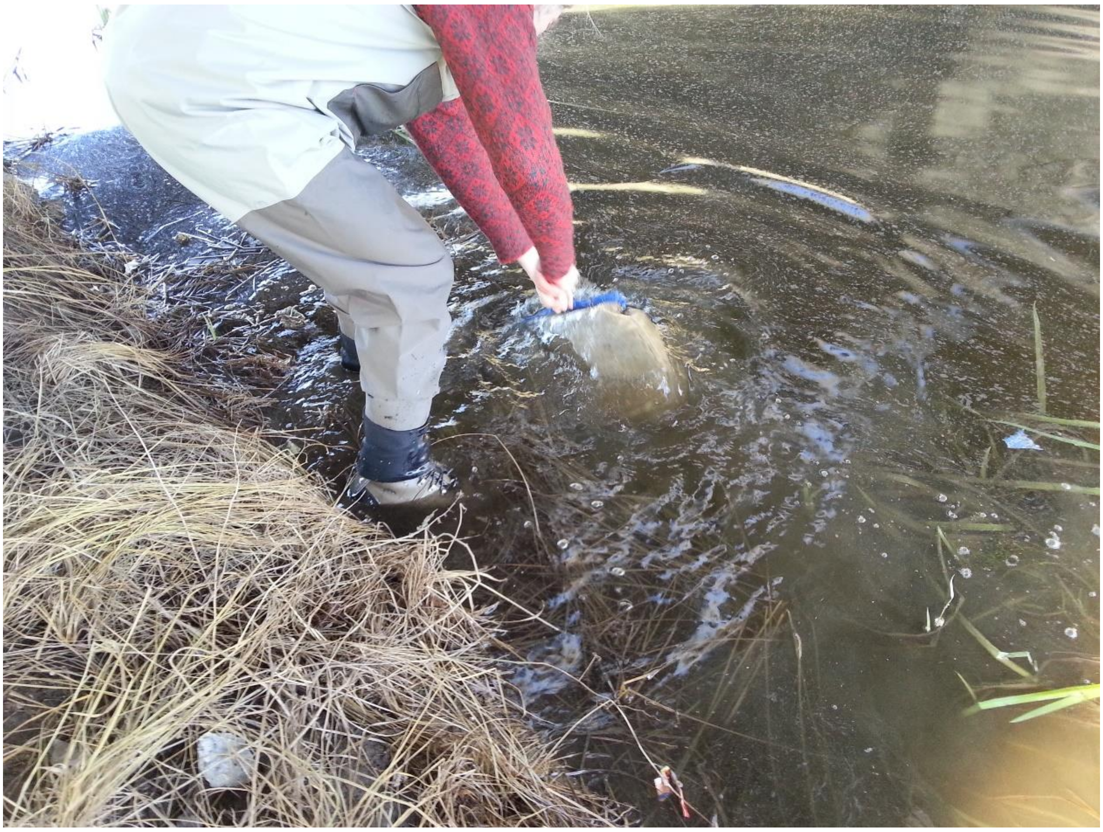


Figure 4. Picture of the biological sampling process. The kick net is swept five times from one side to the other in the same spot, covering all the depth of the pond.

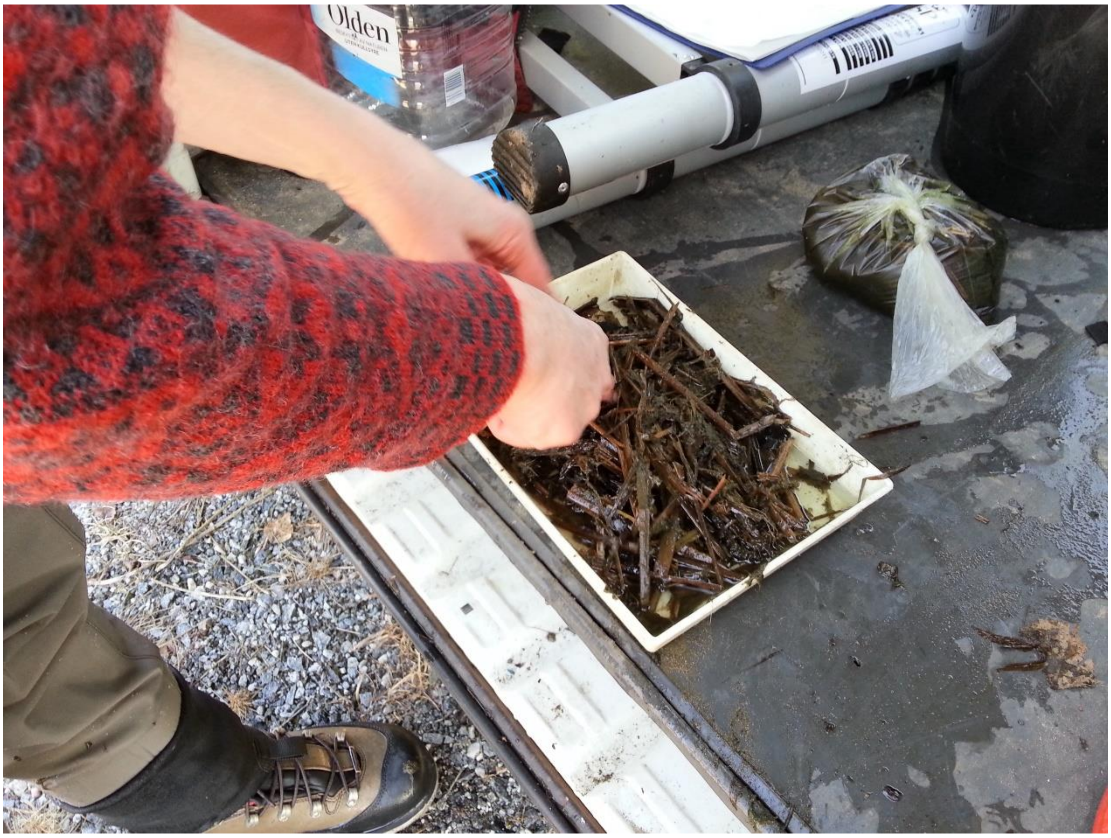


Figure 5. Picture of the biological sampling process. The biological sample is placed in a sampling tray, where ethanol is added for preserving the sample. Afterwards, the sample is stored in a plastic bag.

4.2 Literature review

In first place, a literature review was conducted. The aim of the review was to provide a guide with the information that is required for a basic understanding of the motives, development and initial conditions in the field that this document approaches.

The background is mostly focused on the review of the different types of ecological modelling that have been developed in recent years. It is not the intention of this section to be a reference document, but to provide enough resources to acquire a basic knowledge of this particular topic. Also, it has the aim of supplying references to additional documents supporting the links that allow the broadening of this piece of knowledge.

4.3 NORWAT data

The raw data were available through the work performed by Thygesen (2013). This document contains several datasheets containing the results of several water quality and biological analyses performed during four different months during 2013. Additional data for sediment quality were provided by NORWAT members, although no reference was given, since the data have not yet been published.

In this section, the different proceedings applied to the raw data provided by Thygesen (2013) and sediment quality data are described.

4.3.1 Water quality data

The water quality data were inside a excel file with extension .xlsx with the name *WaterQuality*. The file contained two different worksheets, *WaterQuality* and *Vanndata*. The first worksheet contained the raw data, while in the second worksheet the data were slightly modified for its introduction in the software program *CANOCO*. The two sheets contained the same chemical analysis results, so the *WaterQuality* worksheet was employed, since it contained the complete names and information about the month in which the samples were collected.

Water quality data were measured in twelve different stormwater ponds: Skullerud, Taraldrud North, Taraldrud Junction, Taraldrud South, Nøstvedt, Vassum, Idrettsveien, Nordby, Ennebakk, Fiulstad, Såstad and Karlshusbunn. A detailed description of the studied stormwater ponds can be found in Thygesen (2013) . The measurements were taken once per month during the months of April, June, August and October in 2012. Generally, only one sample was taken. However, in three ponds, Idrettsveien, Nordby and Karlshusbunn, two samples were taken, one per side of the pond.

The first column contained the complete name of the pond that was analysed and the month in which the sampling was performed. The second column contained an abbreviated code for the pond and month. According to Thygesen (2013), the code was introduced for simplifying the analysis of data in *CANOCO*. The rest of the columns contained the numerical values of the concentrations of 28 variables. The variables contain 23 different elements and 5 additional data: Total Organic Carbon (mg/l), Temperature (C°), Oxygen (mg/l), pH and Conductivity (µs/m).

The water quality data supplied were not modified and no work was performed over it.

4.3.2 Sediment quality data

The sediment quality data were also provided in an excel file (.xlsx file extension) named *Sediment_chemical data*. The file contained a single worksheet, *Tulostaulukko*.

The sheet contains 55 columns, from which the 4 first are filled with information about the samples, *Sample ID*, *Sample name*, *Product Code* and *Place*. The 51 remaining columns contain the results of the chemical analysis of the sediment. There are two different groups of chemical compounds, single element concentrations and hydrocarbon compounds.

The data provided in the data file contain the analysis results of 32 samples. Each sample corresponds to a different pond. Hence, the sheet contains data of 32 different ponds. From these 32 ponds, there are only 9 sets of data corresponding with the ponds contained in the water quality data file. Thus, the sediment quality data of 3 pounds, Fiulstad, Såstad and Karlshusbunn, were not recorded. According to the NORWAT project coordinators, some of the ponds were not monitored after the first analyses due to some of them not working properly at the time when the samples were taken. Only the sediment data of the ponds where the water quality data were studied were used.

The sediment data contained only one analysis per stormwater pond. The sediment samples were taken between 2013/04/30 and 2013/06/27. According to NORWAT researchers, sediment quality does not vary drastically between different months; having remained stable after a first set of samples was obtained. This is consistent with literature, where German and Svensson (2005) found very little variations in sediment quality provided no extreme contamination events occur. Considering these facts, the quality data for the sediments were assumed to be constant and equal in the 4 months of measurements of water quality data. Furthermore, the neural network models with which the data were modelled, introduce an element of randomness into data, as they assume that there is an associated uncertainty to them (Dreyfus, 2005).

For modelling purposes, the analysis data were modified and processed. Some of the element and hydrocarbon columns presented data with the less-than sign. The meaning of this was not provided in the excel file. Nevertheless, it can be safely assumed that the less-than sign corresponds to measurements where concentrations were below the detection limit of the sampler. The treatment of data below the detection limit is not obvious. Since the exact registered values for the below the detection data were not provided, the only statistical treatment possible was to ignore these data or use a simple substitution method (Helsel, 1990). The decision was made to use the detection limit values as if the sampler registered the original data. However, some chemical elements and the Polycyclic Aromatic Hydrocarbons (PAHs) presented a vast majority of data points below the detection limit. Considering that substituting the below detection limit with the exact detention limit value would drop the statistical significance of the sample, the decision was made to not use those concentrations. The elements that were eliminated from the sample data were mercury (Hg), silver (Ag) and selenium (Se). In the list of hydrocarbon compounds, all the individual PAHs were eliminated, leaving the sum of 16 US EPA PAHs as the only source of data for PAH.

4.3.3 Biological data

Two sets of biological data were provided by NORWAT, one excel file containing data of taxa sampling and another containing fuzzy data of the vegetal populations in the stormwater ponds.

4.3.3.1 Taxa sampling data

The taxa sampling data were included in an excel file with the name *Artliste taxa, species numbers*. The excel file contained two worksheets, *ORIGINAL* and *BRUK*. The two sheets contained the same data and in a similar fashion to that observed in the water quality excel file. The sheet *ORIGINAL* was the one employed for the extraction of data.

The data were distributed in 118 columns. The first column contained information about the sampling site, the month when the sample was collected and the part of the pond where the sampling was taken. The second column contained a code for the identification of each sampling for its use in *CANOCO*. The remaining 116 columns contained the number of individuals of each of the 116 found species for every sample. If a sample contained individuals of a specific species or taxa then the number of observed individuals was introduced in the corresponding cell. If no individuals were observed then the cell was left blank.

The taxa sampling was performed in the exact same ponds as the water quality sampling was done. Nevertheless, the number of sampling tests carried on differed. While in the water quality analysis one sample was taken per month and pond during the taxa sampling 5 samples were generally taken. As explained in section 4.1, the sampling methodology consisted of three samples with the net, one in the inlet and two in the main basin, and two traps. However, for some months, one or more of the subsamples are missing. Also, in some stormwater ponds 6 subsamples were taken in a single month (Idrettsveien and Karlshusbunn).

The taxa data were processed considering each of the ponds as a single habitat, and hence, unifying the data from the subsamples into a set of values representing the totality of the pond. Due to the high variability in the number of subsamples taken in each pond and month, the sample mean value of each species was considered. This value was obtained by summing the number of individuals found in every subsample collected and dividing that number by the total of subsamples. With this modification, the data from different months and ponds could be compared.

Using the modified data, three biodiversity indicators, the Species Richness, the Shannon index and the Inverse Simpson index were calculated.

4.3.3.2 Aquatic plant sampling data

The aquatic plant survey data were included in an excel file with the name *Damundersøkelse-planter-data*. The excel file contained two worksheets.

In the first worksheet, with the name *Totalliste damdeler mengde*, an account of the species encountered in each pond and their frequency are presented. The aquatic species in this sheet are divided depending on if they were seen in the water or at the shore of the stormwater pond. The first column of the sheet presents the common name of the aquatic plant accounted, while the second column contains the scientific name. The following columns contain the frequency of appearance of the species in the different ponds. The frequency of appearance in some of the ponds is independently accounted in the inlet and the main body of the pond. For the two groups of aquatic plants (water and shoreline plants) the total number of species present in each pond and the total number of ponds in which the aquatic plant was found were accounted. Hence, the worksheet contains information of aquatic plant species in every pond, the number of species present in every pond and the number of ponds in which each plant species was found.

In the second worksheet, named *Forekomst damkompleks*, the two groups of plant species are mixed together and the total number of aquatic plant species for each pond is calculated. Neither the data contained in the first worksheet nor the data in the second were modified.

4.4 Biodiversity measurement

Biodiversity can be defined as "the structural and functional variety of life forms at genetic, population, community, and ecosystem levels" (Sandlund et al., 1992). This definition includes two important concepts. First, the biodiversity is not only the measurement of the number of individual species, but of a more complex structure of interrelated forms of life. Second, there is not only one type of biodiversity, but several biological levels of biodiversity (Van Dyke, 2008).

Considering these two concepts, any measurement of biodiversity must include information about what feature of biodiversity is measured, the level of biodiversity explored, and the dimension of that level (Van Dyke, 2008). There are several levels of biodiversity such as genetic, species, community or ecosystem diversity. Despite this, the term biodiversity is most often used at the species level (Colwell, 2009).

The dimension of the biodiversity levels is generally divided in three, which are defined by the alpha, beta and gamma diversities. The alpha diversity measures the mean species diversity within an ecological community, or, in other words, at a local scale, for example, sample sites. The beta diversity is the diversity among different ecological communities. This provides an understanding of biodiversity at a regional scale. The gamma diversity is defined as the measurement of the total diversity across a landscape level. The gamma diversity is the product of the alpha and beta diversities across the landscape. Thus, the gamma diversity measures the diversity of different types of ecological communities across a landscape level (Van Dyke, 2008). In the case of this Master's thesis, the biodiversity will only be measured at a local scale. Hence, only the alpha diversity will be considered.

The first approach to the measurement of biodiversity within an ecological community is to obtain a list of species. In this list the number and name of species identified at a particular sample site is noted down. The information provided by this measurement is quite limited. An improvement can be made by standardizing the measurement of the number of species per sampling area or per a determined number of observations (Van Dyke, 2008). This biodiversity measurement is called species richness. The species richness is defined as "the number of species of a particular taxon or life form that characterize a particular biological community, habitat, or ecosystem type" (Colwell, 2009). However, the only measurement of the species richness would not give any information about the abundance of each of the species. The evenness is the measurement of the abundance of each species relative to the others in the same community, habitat, or ecosystem type (Maurer and McGill, 2011).

Thus, the species richness and the evenness are the basis for other measurements of biodiversity. There are many other formulations, which mostly measure both species richness and evenness, and in some cases, even other biodiversity qualities. This biodiversity metrics are called *biodiversity indices* (Van Dyke, 2008). In this thesis, two of the most relevant and widely employed biodiversity indices will be employed: the Shannon index and the Inverse Simpson's diversity index.

The Shannon index (Equation 4.1) is probably the most widely used diversity index. The Shannon index is based on the Shannon's information theory. Thus, the Shannon index measures the uncertainty of an individual selected at random from a population being of a certain species. The more species there are and the more nearly even their distribution, the greater the uncertainty, and, hence, the greater the diversity (Peet, 1974).

$$D_{Shannon} = -\sum_i p_i \ln\left(p_i\right) \tag{4.1}$$

where:

$D_{Shannon}$ is the Shannon diversity index.
p_i is the percentage of abundance within the community by the *i*th species.

The Simpson's index (Equation 4.2) is the first heterogeneity index used in ecology (Peet, 1974). It measures the probability that two individuals drawn from a sample, which can be an infinite or finite community, would belong to the same species (Maurer and McGill, 2011). As Simpson's index is inverse of the diversity, several new versions were created to solve this problem. The most common approach is the Simpson's diversity index, which is the inverse of the Simpson's index.

$$D = \sum_i p_i^2 \tag{4.2}$$

$$D_{simpson} = \frac{1}{D} = \frac{1}{\sum_i p_i^2} \tag{4.3}$$

where:

D is Simpson's index.
$D_{Simpson}$ is the Simpson's diversity index or Inverse Simpson index.
p_i is the percentage of abundance within the community by the *i*th species.

4.5 Model theory

4.5.1 Static black-box modelling

Out of the several model types analysed in Section 3 the black-box models have been selected for its application in this specific case. Black-box models are models that find a relation between observed explanatory variables and target variables. They do not rely on specific analytical formulations, therefore, not requiring a full understanding of the problem to be modelled. The use of black-box models combined with statistical analysis allows the modeller to define models with a given accuracy, even when not all the variables that might interfere are measured (Dreyfus, 2005).

As black-box models are only based on observed data, the importance of observed data grows compared with analytical methods. The precision and range of application of the model depend almost exclusively on the data used for the definition of the model. Hence, the use of black-box models should be preceded by a well-defined sampling campaign. First, the modeller should define the level of precision that the model will require. The error of the measurements will define the overall error of the model. Second, the sampling should define variables as frequently and widely as possible, at least in the range of values in which the model is most likely to be used. Black-box models are not good estimators outside the range of values with which they are created. Hence, for creating a model with good predictive capabilities, the observed data must cover all the probable values that the variable can have. Also, when the quantity of data samples is large, the model can learn more examples and, thus, provide better performances. If measurements are already taken, the modeller should check the validity of data for its use in a black-box model, and if necessary undertake extra measurements.

In essence, black-box models search for the regression function. The regression function uses the value of the measured variables to provide an expectation value of the target variables. The first step in the construction of a black-box model consists of making an assumption on how the function will behave (Dreyfus, 2005). If the variation of the target variable against the explanatory variables is assumed to be linear, then a linear or an affine model might be used. If the relationship is thought to be more complex or the results using a linear model are not satisfactory, a non-linear behaviour can be assumed.

According to Dreyfus (2005), the development of a black-box model (non-linear) includes three main tasks:

1. Input variable selection. Choose from all the variables measured only those that explains the variation of the target variables. It is done in two steps: reduction of the number of variables to eliminate redundancy and rejection of the variables not contributing to the model.
2. Estimation of model parameters. For the development of the model, one must choose which type of model (functions defining it) to use and then decide which conditions within the chosen model type are best.
3. Performance analysis. If the calibrated model is not satisfactory, a different type of model might be chosen. With the new type of model the process is repeated from point 2.

First, the input variables for the model are selected. Afterwards, a linear model is created. If the results are regarded as not satisfactory, a non-linear model is built. The non-linear model can be created using different sets of function, from which only neural networks will be used. After calibration, the results are analysed, and the calibration process is repeated until the results are as good as desired.

4.5.2 Input variable selection

Input variable selection is one of the major problems in the application of neural networks in real world problems (Giordano, 2014). Some applications of models are focused in cases in which the processes and variables are clearly defined or at least bounded. Nevertheless, neural networks and other black-box models are applied when the knowledge of the processes and variables interfering in them are often unknown. In such problems, sampling methodologies usually try to be as broad as the budget and technical possibilities allow. Thus, the modeller starts with an ample quantity of variables from which the model must be developed. In a set of sampling data with numerous variables, three main difficulties may arise: excessively large number of variables, the existence of correlated variables (redundant variables) and of variables with little or no predictive power (May et al., 2011).

The number of input variables of a model defines its prediction capacity and general performance. An under-specified model is a model with insufficient variables or defined by uninformative variables. On the other hand, a model with an excessive number of variables, in which many are redundant or uninformative, is called an over-specified model. According to May et al. (2011), the input variable selection has a major impact in the relevance of the model, the computational effort, the training difficulties, the dimensionality and comprehensibility. The *relevance* of the model is affected by the selection or not of explanatory variables within those among the input of the model; the absence of one, several or even all relevant variables derive in a model with bad performance. The *computational effort* in any model is dependent on the number of parameters of the model. In neural networks, the presence of an extra input variable adds at least one parameter to the model, which increases the size of the network and contributes to extend the computational burden. Additionally, the *training* becomes more difficult if redundant and uninformative variables are included in the model. In one hand, redundant variables add more combinations of possible solutions to the network, increasing the number of local minima, which might not be the ones yielding the best error. On the other, irrelevant variables add noise to the network and decrease the efficiency of the algorithms. *Dimensionality* relates the number of parameters in the model with the number of samples required to maintain a given precision. Thus, the greater the number of parameters the more samples would be required, relation that grows rapidly. Finally, the *comprehensibility* or capacity of the network to discover relationships between input and output variables is decreased when the number of input variables is increased.

Regarding all these effects of the input variables in the model, a definition of a perfect input selection can be yielded. A desired input variable would be a highly informative explanatory variable and independent to the other input variables. Consequently, the set of input variables would be minimal in the number of variables, reducing redundancy, and with maximum prediction capacity over the output, thus minimizing the number of uninformative variables. (May et al., 2011).

There exist numerous algorithms and methodologies for both reducing redundant variables and highlighting unimportant variables. However, the existent reviews and compilations of methods fail in their effort to classify the methodologies comprehensively (May et al., 2011, Hamby, 1994).

In this section, the classification defined by May et al. (2011) will be followed and some of the methods will be shortly explained. May et al. (2011) review is more recent, including more methods, and provides a more reasonable classification of the Input Variable Selection (IVS) groups.

4.5.2.1 Dimension reduction

Dimensionality reduction methods are not exactly defined as IVS methods. Their main aim is the reduction of the number of input variables for minimization of the computational burden. The field of application of this type of methods is essentially multivariate data analysis, but are often employed for input selection.

Several methods exist inside this description, being the Principal Component Analysis (PCA) the most well known. Other methods have been developed with PCA in focus but modifying some restrictions and limitations of the former, such as Independent Component Analysis (ICA) and Curvilinear Component Analysis (CCA).

4.5.2.1.1 Principal Component analysis

Despite that its precise origin is difficult to trace, PCA was one of the first statistical analysis to be developed (Jollife, 2002). The concept of principal components is to reduce the number of variables in a problem to a set of newly created uncorrelated variables derived from the original set with the maximum explanatory capacity.

The PCA method uses the input space or representation space, which contains as many dimensions as variables exist in the model, as starting point. Each observation is represented as a point in the multidimensional space according to the values that each of the variables has. The PCA reduces the dimension of the representation space by creating sub-spaces with fewer dimensions where the distribution of the observations is as close as that in the input space (Šmilauer and Lepš, 2014). The similarity between representations is measured by the total inertia of the scatter diagram, which in statistical terms is defined as variance. Thus, the PCA consists of the linear projection of the observations in sub-spaces that maximize the inertia of the scatter diagram (McGarigal et al., 2000). In this way, the first axis in PCA would be the axis to which variance of data is maximum, followed by a second orthogonal axis to which the variance of the data projected from the first axis is maximum, etc.

From a mathematical perspective, the PCA of a data set *X* is developed as follows.

1. The mean value of each variable is subtracted.
2. The covariance matrix is calculated.
3. The eigenvectors of the matrix are calculated.
4. According to the eigenvalues, a ranking of eigenvectors is performed.
5. Select the number of desired principal components from the ranking.

PCA are usually represented in 2 dimensions, the first principal component in the abscissa and the second component in the ordinate. The immediate plot in this situation would be the projection of the observations in the new plane. The information of such a representation is quite limited. Due to this, the PCA are always represented in *biplots*. A *biplot* represents the observations but also display the relative positions of the variables in the two dimensions defined by the two first principal components. Representing the two plots simultaneously provides useful additional information about the relationship between variables and observations (Jollife, 2002). An example of a PCA *biplot* can be seen in Figure 6.

For dimension reduction purposes, the PCA is employed by substituting the initial variables with the principal components of the analysed data. However, not all the principal components are required, since only the first principal components retain the vast majority of the variability. Hence, for the dimension reduction, a set of the most relevant principal components is chosen. Nevertheless, it is not clear how to decide whether a principal component is relevant enough or not. Thus, the number of principal components is chosen by the percentage of explanation of the overall data that each of the principal components contain. A common method for determine the number of components to choose is based on the *scree plot*. The *scree plot* is the representation of the percentage of explanation of the principal components against the order of the principal components. One method establishes that the last relevant principal component is the principal component in which the gradient of the cumulative variability curve changes in slope. Another method, more conservative, declares that an appropriate limit for the principal component selection is that where the cumulative variability reaches the 95%. In Figure 7, an example of a scree plot is shown.

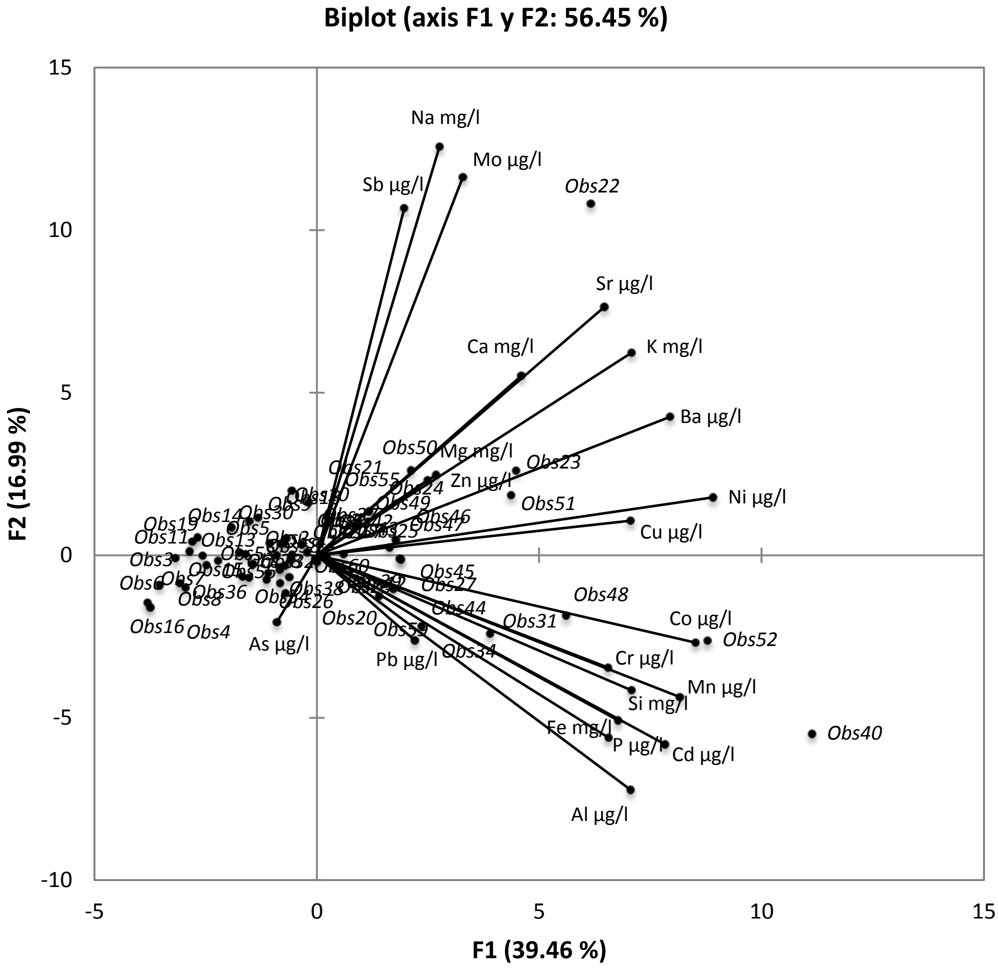


Figure 6. Example of a PCA biplot. The isolated dots represent the projections of the observations in the new axes, while the lines display the variables.

Nevertheless, the use of PCA for dimension reduction has some flaws. First, the PCA assumes linearity in the mixing of input variables, and between them and the output variables. If there is any non-linear relationship between data, the PCA will fail in finding it and, even more, will linearize the relationship after the linear projections. Second, the PCA transforms the original variables into a new set of orthogonal uncorrelated vectors with explanatory basis. Hence, after the transformation the identity of the original variables is lost. Thus, it is not possible to account the contribution of each variable to the variance of the final output (May et al., 2011). Thus, if only the first principal components are chosen, but a later principal component explains the variance of a single important variable, the information contained in that variable would be lost (McGarigal et al., 2000).

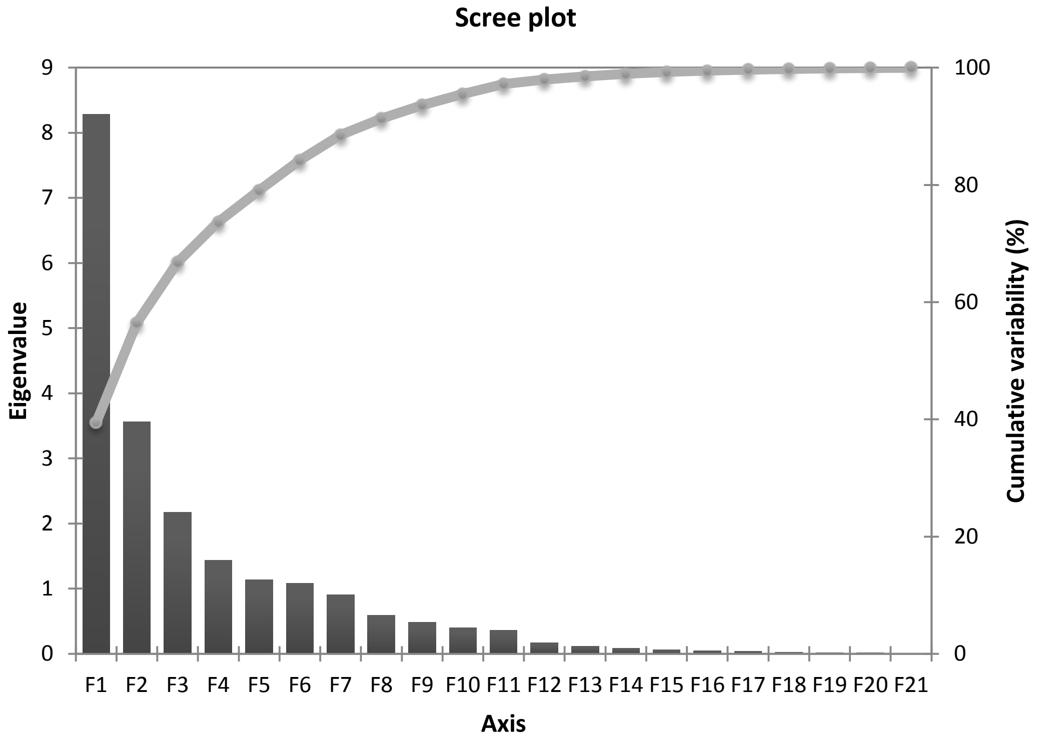


Figure 7. Example of a scree plot. The bars represent the eigenvalues of the principal component axes, while the line represents the cumulative variability explained by the PCA axes.

4.5.2.1.2 Other dimension reduction methods

Considering the flaws of PCA, data analysts have developed several different dimension reduction methods. Mainly, the new methods focus on the problem of non-linear data. The Independent Component Analysis (ICA) is an alternative methodology to PCA that is usually employed in signal processing. Since ICA is not restricted to linear correlations, it has been more often used with non-linear datasets (McGarigal et al., 2000). The Curvilinear Component Analysis (CCA) is also a nonlinear dimensionality reduction method, often employed in data analysis. The method is employed to represent data structures distributed in a nonlinear manner (Dreyfus, 2005).

4.5.2.2 Variable selection

The Input Variable Selection (IVS) methodology consists of the use of algorithms for selection of the input variables that maximize the explanatory capacity of the input minimizing the total number of variables. The IVS algorithms have been usually classified into three kinds, *wrapper*, *embedded*, and *filter* algorithms (May et al., 2011).

Wrapper algorithms are the simplest IVS algorithms. Wrapper algorithms treat IVS as part of the optimization process of the model. Thus, the efficiency of a wrapper model for IVS depends on the ability of the selected model to learn the relationships between input and output variables. One of the most simple and used algorithms among the wrapper algorithms is the Single Variable Regression (SVR) method. The SVR algorithm consists of the training of a model with just one variable at a time, and measuring the error of the model with a test data set. Depending on how well the input variable explains the output, a ranking of input variables can be built. Furthermore, the algorithm includes a statistical bootstrap method for determining whether a variable contributes to the explanation of the result or not. In this case, the bootstrap method consists of the random selection with replacement of the samples of the input variable. Thus, the model is trained several times, once using the original variable, and several (many) more using the randomized variable. When the error of the model containing the original variable is greater than a certain percentile of the randomized variable error, then that variable is rejected. The main flaw of the SVR method is that it does not consider the existence of redundant variables, and using the method can yield in a set of many redundant variables. Therefore, a dimensionality reduction pre-processing step is required (May et al., 2011).

Embedded algorithms are algorithms that are *embedded* or directly incorporated inside the training algorithm of the model. Embedded algorithms are similar to wrapper algorithms with the difference that only one model is trained. Both wrapper and embedded algorithms are based on iterative processes. In embedded algorithms, instead of iterating different models for each variable, as in the wrapper algorithms, only a single model, containing all variables is iterated. Also, while wrapper algorithms consider model performance of each variable at a time, the embedded algorithms can account for the impact of each variable in the model performance (May et al., 2011).

Finally, filter algorithms are algorithms not based in any model. This means, that the process of IVS can be performed even if the type of model has not yet decided. Filter algorithms test relevance of individual or combinations of variables independently of the model to indicate which the most important variables are. Within the filter algorithms to different classes can be distinguished: the ones based on linear correlations and the algorithms based on mutual information (MI) criteria, which is a theoretic measurement of the dependence between variables. In the group of linear correlation algorithms, two methods are highlighted: the rank correlation method and the partial correlation method. The two methods are based on the classification of variables in base of the Pearson correlation between the input and the output variables. The difference between the rank and the partial correlation methods relies on the consideration of redundancy. While rank correlation does not consider it, the partial correlation includes a term for testing the correlation between input variables. The linear correlation methods have the flaw of only considering linear relations. Due to this, the mutual information methods were developed. Mutual information methods are more capable of identifying relationships when data seems to be chaotic or non-linear (May et al., 2011).

4.5.3 Artificial Neural Networks (ANNs)

4.5.3.1 Simple neuron

A neuron is defined as a nonlinear, parameterized, bounded function (Dreyfus, 2005). The neuron or simple neuron is the basic unit that composes a neural network. A single neuron is composed of three basic elements or operations:

1. Synapsis. The synapsis is the link that connects the neuron with the previous neurons or inputs. The synapsis is characterized by the *weight* of the bond between the two elements. The weight is used for prioritizing inputs that the model finds to be more important and degrading those not affecting the goodness of the model. There are several weight functions, although the most employed is the product. Thus, the weight function would consist of the product of the neuron input *p* and the weight *w*. The

result is called *weighted input* and is passed to the following element in the neuron as an input.

In a structure with more than one input, there would be a weight scalar for every input to the neuron. The weight function would be the scalar product of the input and the weigh vector.

2. Bias. The bias is introduced in the neuron to account for the uncertainty of the input. Usually the bias is added to the weighted input as a scalar. The result is called *net input, n*. The bias is applied to the weighted input as a single element and not to each of the input elements. Hence, developing the equation for an input vector of *R* elements:

$$n = w_{1,1} \cdot p_1 + w_{1,2} \cdot p_2 + \ldots + w_{1,\mathrm{R}} \cdot p_R + b \qquad (4.4)$$

3. Transfer or activation function. The activation function is a function that limits the amplitude range of the output of the neuron. Usually this range is limited to the closed spaces [0,1] or [-1,1], depending on the function used. Although there are several different transfer functions, three of them are most important.

The *threshold function* transforms the net input into two groups. The neurons using this transfer function are referred as all-or-none neurons, following the expression:

$$f(n) = \begin{cases} 1 & \text{if } v \geq 0 \\ 0 & \text{if } v < 0 \end{cases} \qquad (4.5)$$

The *piecewise-linear function* transforms the net input using three different linear functions. The linear transfer functions are usually employed in the final layer of multilayer layers. For a piecewise-linear function with an amplification factor of 1, the expression of the function is:

$$f(n) = \begin{cases} 1 & v \geq \frac{1}{2} \\ v & \frac{1}{2} > v > -\frac{1}{2} \\ 0 & v \leq -\frac{1}{2} \end{cases} \qquad (4.6)$$

The *sigmoid functions* are a group of functions that provide an s-shaped graph as a result. It is the most common employed type of transfer function, as they provide an ideal balance between linearity and non-linearity. Among the sigmoid functions, one can find the logistic function, logarithmic sigmoid function or the hyperbolic tangent function.

The result of applying these three elements to the input of the neuron gives an output, *a*, that can be used by a new neuron as an input or it can be the final result of the model. The structure of the neuron can be finally defined by the equation:

$$a = f(w \cdot p + b) \qquad (4.7)$$

A scheme of a simple neuron with all its components can be observed in Figure 8.

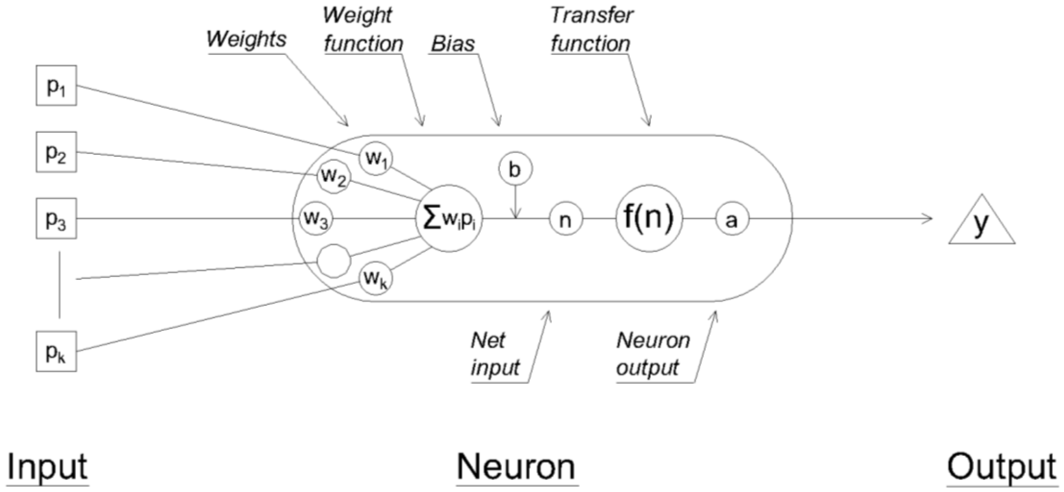


Figure 8. Schematic figure of a single simple neuron neural network.

4.5.3.2 Network

An Artificial Neural Network (ANN) can be constituted by just one single neuron, as the one explained in the former section, or a high number of them. When an ANN is formed by more than one neuron the neurons can be organized in two dimensions, number of layers of neurons and number of neurons per layer.

When an ANN contains more than one layer of neurons, the structure of the network becomes more complex. In a single neuron, the input is combined and computed by the neuron and an output of the model is obtained. In a multi-layered network, the first of the neuron layers functions exactly the same as in a single neuron model. However, the output of this neuron is not the output of the model. The result of the neuron is transmitted to other neurons, which use the output of the model as its input.

Depending on how the output of the neurons are transmitted, two different neural network classes can be defined, the *feedforward* and the *recurrent neural networks.* The difference between the two types of networks is the direction in which the information flows.

The feedforward neural networks are networks where the information only goes in a forward direction. The flow of information starts with the input variables and is propagated from one layer of neurons to the next until the final output of the model is obtained. The result of the transfer function in the first layer of neurons is used as an input for the following layer or layers of neurons, in which the process is repeated until the output layer of the neural network is reached.

In a multi-layered feedforward neural network, three different types of layers can be found. The input layer is a layer that contains the values of all the variables for every sample taken, which are fed to the model. The input layer is not a neural layer, since it does not contain neurons. Nevertheless, when accounting the number of neuron layers, some authors consider the input layer as a neuron layer. The hidden neuron layers are the layers of the model that contain neurons but do not provide an output for the model. The hidden neural layers process the output of the previous layer of neurons (or the input layer) and fed the following neuron layer with their output. Normally, the transfer functions employed in the hidden layers are sigmoid transfer functions, which account for the non-linearity to the model. Finally, the output layer is the layer of neurons that provide the output or result of the model. The output layer contains as many neurons as target variables are studied in the model. The transfer function of the output layer is usually configured to be a linear function for function approximation. A scheme of a multilayer neural network can be seen in Figure 9.

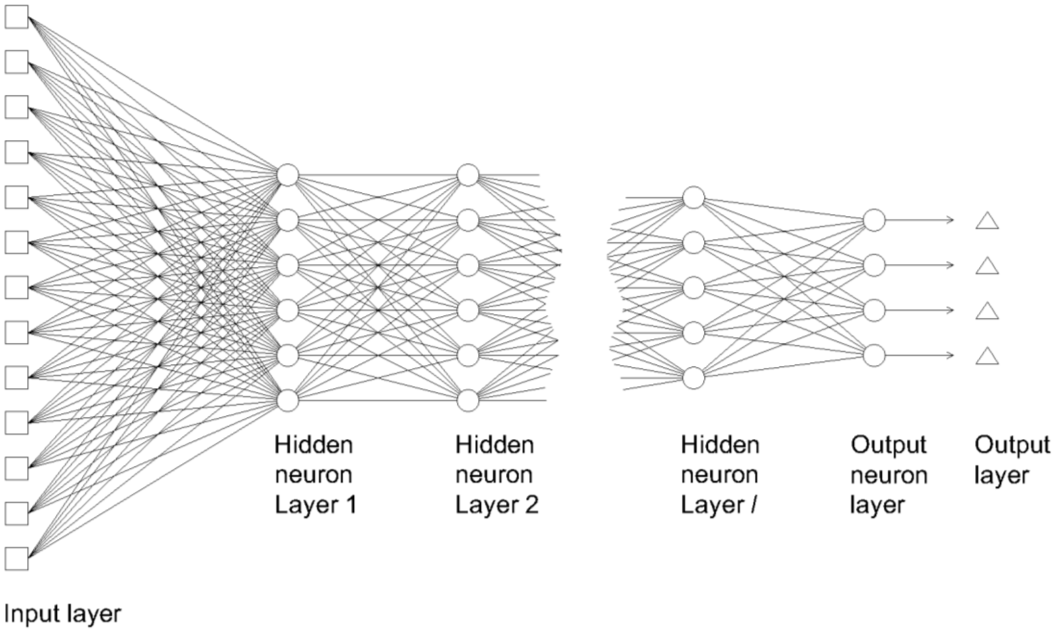


Figure 9. Scheme of the connectivity of a multi-layered artificial neural network. Each square represent the input for each variable to the model, the circles represent the neurons in the model, and the triangles represent each of the output variables of the model.

The recurrent or feedback neural networks are the most general neural network architecture (Dreyfus, 2005). These types of networks allow the information not only to flow forward but also backward, to neurons in the same layer or prior layers. The definition of such architecture requires the definition of an additional dimension. If a neuron of a recurrent network sends the output to a previous layer, in what is called a *cycle*, and no new dimension is added, there would be two inputs simultaneously for this neuron. Thus, in order to solve this problem, the dimension of time has to be defined. With the inclusion of time and time steps or *delays*, the information can be divided between current values and values from the past. The delays determine the moment in time when the connection between two neurons takes place. In every cycle there must be at least one connection with delay different than zero.

4.5.3.3 Preparation of data

Before undergoing the design of a new neural network, the data to be used in it has to be collected and then prepared. The preparation of the sample data can be divided in two steps: *pre-processing* and *division of data* into subsets. Even before the preparation, the user has to check that the data that has been collected will be useful if introduced in a neural network. Neural networks are not accurate predicting data outside the range in which they have been trained with. Hence, it is important to verify that the collected data covers sufficiently the range of values in which the networks is going to be used.

The pre-processing of data ensures that the network will behave efficiently during the process of training. The pre-processing of data usually consists of the normalization of data prior to the training. The normalization of the input data has as its aim to prevent the transfer functions to become saturated, especially when the functions are sigmoid. The normalization of data is performed both with the input and the target data. The network's output is afterwards reverse-transformed to match the units of the original target data.

The pre-processing of data also verifies that the input and target variables fed to the neural networks have no elements repeated. The repetition of samples provides no improvement in the network but increases the computational time. Another task carried out with the pre-processing concerns the unknown or "don't care" values in the target variables. When one or several of the target variables are missing, the input and target values of that sample are eliminated.

The division of data into subsets is a general proceeding before the training of the network. The subsets into which the data is divided are the *training*, the *validation* and the *test* sets. The training set is the group of examples used for the training of the network. The validation set is a complementary data set that is used for avoiding overfitting of neural networks by the early-stopping method. Finally, the test set is a set completely separated from the training process. The purpose of the test set is the comparison of the prediction capacity of different models.

The division of data can be performed in different manners. A common methodology consists of the division of observations randomly (Lek, 1996, Karul et al., 2000, Lock et al., 2014). Other data division method consists of splitting data in three contiguous sets (Singh et al., 2009). In other cases, a more complex sort of data division is employed. For instance, in Bowden et al. (2006), a Genetic Algorithm (GA) is used for data division. With the GA, the data division guarantees that all the patterns in the available will be present both in the training and test sets.

4.5.3.4 Training

The most relevant property of the neural networks is the ability of the network to learn from the information that is provided by the user. With the learning, the network improves its performance. The process of learning in neural networks is usually referred as *training*.

The training of a neural network consists of the progressive adjustment of the weights and bias in each of the neurons in order to explain or replicate the information provided to the network. In the case of recurrent neural networks, the training would also adapt the delays of the synapses to achieve the same aim. According to Haykin (1994), the learning process involves three steps. The first step consists of the feeding of an existing neural network with new information. With this new information, the neural network modifies its parameters in order to "remember" the data. Finally, the neural network, with the changes in its structure, responds in a new way.

For the adjustment of the parameters of the model, first, a measurement of the "distance" between the predicted and the sampled values must be defined. This measurement is usually called *cost function* and its purpose is to quantitatively account the accuracy of a model by comparing the outputs with the measured data. The training of neural networks consists hence of the minimization of the cost function, and thus, minimizing the distance between predicted and sampled output. One of the most common cost functions, and the one that is going to be used in this thesis, is the *least squares* cost function. The least squares cost function consists of the total summation of the squared residuals of the observations in the training set.

According to Dreyfus (2005), the training of a neural network can be performed in two different forms depending on how the samples are managed during training. The *non-adaptative* training, or *batch* training, is a training method that only updates the parameters in the model once all the examples in the training set have been processed. On the other hand, the *adptative* or *on-line* training, updates the parameters of the model every time a new example is found. Therefore, the model can be improved over time, even when the training has already taken place. Usually, the training process is performed in the first place with a batch training method, and, if later on, more examples are available, an off-line training method can be used to further improve the model. During the development of this project there was no prediction of new examples for the model, so only batch training was employed. Thus, only the insides of the batch training processes are discussed in this section.

The batch training process is quite different if the network that is modelled is linear or non-linear to their parameters. In a linear model, the least squares cost function is quadratic with respect to their parameters. This means that an absolute minimum value might be obtained, and that minimum can be obtained analytically just by solving a system of linear equations. If the model is non-linear then the least squares cost function is not well defined and can have more than one single minimum. Hence, the selection of the most accurate model is not direct, since several different local minima can be found during the training process. Also, the process of identifying those local minima is not a direct solution of a system of equations but part of a more complex technique.

The batch training of non-linear neural networks is approached through an iterative process. In each iteration, the cost function is computed and a gradient of the cost function is estimated. With the gradient, the parameters of the model are updated and the process is repeated successively. The most complex element of the training of non-linear networks relies on the computation of the gradient of the cost function. The most common and computational economic method for the estimation of the gradient is the *backpropagation algorithm*.

The backpropagation algorithm is an algorithm for the computation of the cost function’s gradient in feedforward non-linear neural networks. The algorithm consists basically of two steps. The *propagation step* consists of the calculation of the output of the network using the input and a random set of parameters in the neural network. With the results, the cost function and its gradient are calculated. The *backpropagation step* consists of the calculation of the gradient by the application of the rule of chained derivatives repeatedly backwards in the network. A step-by-step development of the algorithm can be found in Dreyfus (2005).

The last step in the training of a neural network is the update of the parameters. This is performed using the gradient of the cost function and an iterative minimization algorithm. The most common minimization algorithms are the *Simple Gradient Descend*; and several second-order algorithms, such as the *Levenberg-Marquardt* (Hagan and Menhaj, 1994) and the *Broyden-Fletcher-Goldfarb-Shanno algorithm* (BFGS algorithm). Among all of them, the Levenberg-Marquardt algorithm is the preferred algorithm in small networks (Dreyfus, 2005, Karul et al., 2000, Hudson et al., 2013).

4.5.3.5 Generalization performance

The selection of the model to be used is the third pillar of the neural network design (Dreyfus, 2005). When designing a neural network, the user is interested in achieving the greatest generalization properties out of the model. In other words, when designing a neural network, the modeller must search for the perfect balance between learning capacity and generalization. If the learning capacity of a neural network is set too high, the model can learn even the noise of the training set. Thus, a neural network is said to generalize when the model learns the main average patterns of data avoiding falling into the random error that the data contain. When the opposite happens, the neural network is *overfitted* or *overtrained*. If the neural network reduces the learning capacity too much in order not to train the random noise, the networks is *underfitted*.

Therefore, from a mathematical point of view, the modeller must find a method to select the model in which the number of parameters is enough to approximate the regression functions but limiting the sensitivity of the model to data noise. This can be performed in two different ways, by reducing the number of parameters of the model or reducing the effective size of each parameter dimension (Abu-Mostafa et al., 2012). There exist several methods for the reduction of the number of parameters such as *pruning*, *greedy construction learning*, or *weight sharing*. (Abu-Mostafa et al., 2012). For reducing the size of the parameter dimension, two methods are usually employed, *regularization* and *early-stopping*.

The first group of algorithms have as an aim the reduction of the effective number of parameters inside the network. The pruning algorithms deal with this problem eliminating some of the parameters when the algorithm detects that that connections are not used for the final result and, thus, could have been adding noise to the solution (Reed, 1993). The greedy construction learning proceeds almost opposite, increasing the size of the network successively until a good solution is found. With this method, the algorithm adds hidden neurons progressively but in the form of a cascade, every time a neuron is added the rest of the neurons are kept fixed. The new neuron is set to improve the result obtained by the last neuron to be added (Fahlman and Lebiere, 1989). Finally, the weigh sharing algorithm reduces the size of the parameter space by sharing the parameter values of different neuron connections.

The second group of algorithms is the most frequently employed generalization algorithms. All the methods based on parameter dimension constriction require of a special set of data called *validation set*. The validation set is an independent set of data from the training and the test sets. The purpose of the validation set will be explained in each case, for the regularization method and the early-stopping method.

The regularization methodology is based on the concept of the cost function and the effect of the function in training. The cost function, most often the least squares function, is the function that measures the difference between the output of the model and the measurements. During the training of the network, the cost function is minimized. When the complexity of the model is too high, the model can *overfit* the data used for learning and then provide a poor generalization of the processes modelled. The solution that the regularization algorithm provides consists of the modification of the cost function to account for the complexity of the model (Wall, 2007). With this new term, the complexity of the model is penalized, even when the error is minimum, and favours the simpler models, despite having higher errors. The restriction to the complexity is achieved by limiting the value of the squared sum of all the weights of the network under some prior decided limit. The most used regularization algorithm is the *weigh decay algorithm*. The main flaw of the system comes from the selection of the limit value, for which there are no analytical formulations. The importance of the limit selection is crucial since a value too low can result on the overtraining of data, and a value too high can result on underfitting of data (Abu-Mostafa et al., 2012).

The solution to the constraint optimization comes with the validation set. The validation set is a complementary data set to the training set. By itself, the validation data is a training set, but instead of training the weights of the network, the new data is employed for training the value of the constraint. The validation set is hence used for the computation of the *out-of-sample* error. The out-of-sample error is the error of the model outside the data employed during training. The out-of- sample error serves as indicator of the generalization capacity of the model. The lower the out-of-sample-error is, the greater the generalization of the model will be. However, since the validation set is limited in size, the out-of-sample error cannot be exactly calculated but estimated. The accuracy of this estimation depends on the size of the validation set, the greater the number of points, the more accurate the estimation will be. Nevertheless, there is a limited amount of data that has to be shared between training, testing and now validating the model. If the validation set is too big, then the training of the model will have to be performed with less data and the learning capacity of the model will be reduced. If the opposite happens, either overfitting or underfitting might happen. Again, no direct mathematical solution for the data split decision is available. Although, as a rule-of-thumb, the validation set size is recommended to be set between one fifth and one sixth of the total size of data (Abu-Mostafa et al., 2012).

Early stopping is a different approach to the generalization problem. Whilst the regularization method searched generalization by restricting the value of the parameters of the model, the early stop method tries to identify the point of the training in which the model starts to overfit. As it was explained in the previous section, the training of neural networks is done in a recursive way. Every time the weights of the neural network are updated during training is usually called *epoch*. Thus, early stopping seeks the epoch of the training where the model has trained the most but before overfitting can happen. As with regularization, the identification of this exact point is done with the help of a validation set, which is used for the *training* of the epoch in which the model should stop. Hence, the model would be trained as usual, but, additionally to the test set, the mean squared error of the validation set would be calculated. The training algorithm would normally reduce the mean square error of the training set progressively, until a certain limit, either a minimum value or a number of repetitions, would be reached. In early stopping, the validation error is also calculated to estimate the out of sample error of the model. During a normal training, both errors, the test and validation errors, would descend simultaneously firstly. Then, once the model has learned the general patterns, the model would start to overfit the training data. When that happens, the validation error stops the descending to slowly start to grow, separating its path from the training error. When this happens, the algorithm decides that the model at the epoch before the overtraining starts is the one that generalizes data the best. Since there are some cases when the training algorithm despites finding a local minimum can later converge to a better minimum, the algorithm usually continues training the model for several epochs more. This is performed to verify with certainty that the model has identified its final minimum (Abu-Mostafa et al., 2012).

4.5.3.6 Model selection

In black-box modelling is usual to have several different models. They can be models of different type, a same model with different complexities or a model with a given complexity and different starting conditions, or even models with different regularization parameters (Dreyfus, 2005). The final step in the modelling process consists of the selection of one out of the different models that have been built. Again, the selection is done following the bias-variance trade-off of the model. The model with the best predictability and with the least overfitting is the wanted model.

Despite that there are several model selection methodologies, all of them are based on the same principle (Dreyfus, 2005). This principle is the generalization error. The generalization error is a measurement of the error that the model would present over unseen data (Murphy, 2012). Thus, the model with the best predictability and with the least overfitting would have the lowest generalization error. A perfect model would hence have no generalization error. However, since prior to the building the functions governing the model are unknown, the generalization error cannot be analytically calculated. The generalization error has to be estimated.

Estimation of the generalization error requires an independent set of data not used during training (Abu-Mostafa et al., 2012). The previously mentioned validation set, despite not being used during training, is employed for the process of early-stopping or regularization. Hence, this data set cannot be used for the estimation since the estimation of the generalization error would be biased (Haykin, 1994). There is a need for an additional independent data set, the *test set*.

The introduction of a new data set can become problematic when the available amount of data samples is limited, for the total data samples have to be divided in training, validation and test set. Thus, with the introduction of the test set, the number of data samples used for training purposes has to be reduced (Abu-Mostafa et al., 2012). Hence, the generalization performance of the trained model would decrease. Furthermore, the estimation of the generalization error is also dependent on the size of the test set. The larger the test set is, the more accurate the estimation of the generalization error would be. This constitutes a problem difficult to solve, since either the performance of the model or the accuracy of the generalization error has to be prioritized over the other. However, with the use of a technique called *cross-validation* this dilemma can be solved (Abu-Mostafa et al., 2012).

Cross-validation is a technique used for the estimation of the generalization error of a model based on the use of a test set. The cross-validation technique allow to simultaneously use a proportionally large training set and to be able to estimate accurately the generalization error (Dreyfus, 2005). This is achieved by dividing the available data into *D* subsets. Then, the model is trained iteratively using *D*-1 subsets of data. For each of the trained models, the generalization error is estimated using the remaining subset of data, which in this case plays the role of test set. In the end, *D* different models are trained and *D* estimations of the generalization error are obtained. Finally, the estimation of the generalization error is calculated as the average value of the *D* generalization error (Murphy, 2012).

Any number of subsets can be used in the cross-validation technique. However, it must be noted that the accuracy in the estimation of the generalization error increases with the number of subsets (Abu-Mostafa et al., 2012). Also, the number of models to be trained increases with the number of subsets, increasing simultaneously the computational effort. Hence, the *leave-one-out* cross-validation (LOOCV), which is the cross-validation technique when the number of subsets equals the number of data samples, would yield the best estimation of the generalization error. Nevertheless, the computational effort required for that accuracy might not be justified. Thus, it is desired to have a balance between accuracy and computational effort. In literature, the most common cross-validation techniques are the 5-fold and 10-fold, which divide data into five and ten subsets respectively (Dreyfus, 2005, Abu-Mostafa et al., 2012, Murphy, 2012). These two divisions are said to preserve computational time while providing a sufficiently accurate estimation of the generalization-error (Murphy, 2012).

4.6 MATLAB

MATLAB is a computer program developed by MathWorks. MATLAB is a numerical computing environment broadly employed in academy and industry. The software includes a programming language, which is backed up by a great quantity of predefined functions. The combination of the MATLAB programming language and the functions included in it, simplifies in a great degree the development of quite complex programs.

For this thesis, the version of MATLAB 7.12.0 (R2011a) was employed. The software was provided by Chalmers.

4.6.1 Neural Network Toolbox

The functionality of MATLAB is further boosted by the inclusion of Toolboxes. The MATLAB toolboxes are complements to the MATLAB basic package that extend the amount of included functions in MATLAB in a specific area.

For this specific thesis, the Neural Network Toolbox was used. The Neural Network Toolbox implements into MATLAB a series of functions that allow the user to create most of the types of ANNs of almost any type of complexity. In fact, the toolbox defines four types of user depending on the level of complexity and changes that they apply into the toolbox. Thus, a user can just simply employ a basic graphic interface, which provide a limit range of modelling options, or go as deeper as desired in order to modify the functions designed by MATLAB behind the basic interface.

5 Results

5.1 Artificial Neural Networks (ANNs)

5.1.1 Input variable selection

5.1.1.1 Input variables

The original input data to the model consisted of water quality measurements, sediment quality measurements, vegetation presence, macro invertebrate diversity, and generic information about the properties and location of the stormwater ponds. The number of observations was not uniform, being 12 ponds for water quality and 9 ponds for the sediment data. Due to this difference, a decision had to be made between using or not, the sediment data. The use of the sediment data would imply that the unmatched water quality observations should be discarded, and therefore, the size of the training set would be reduced.

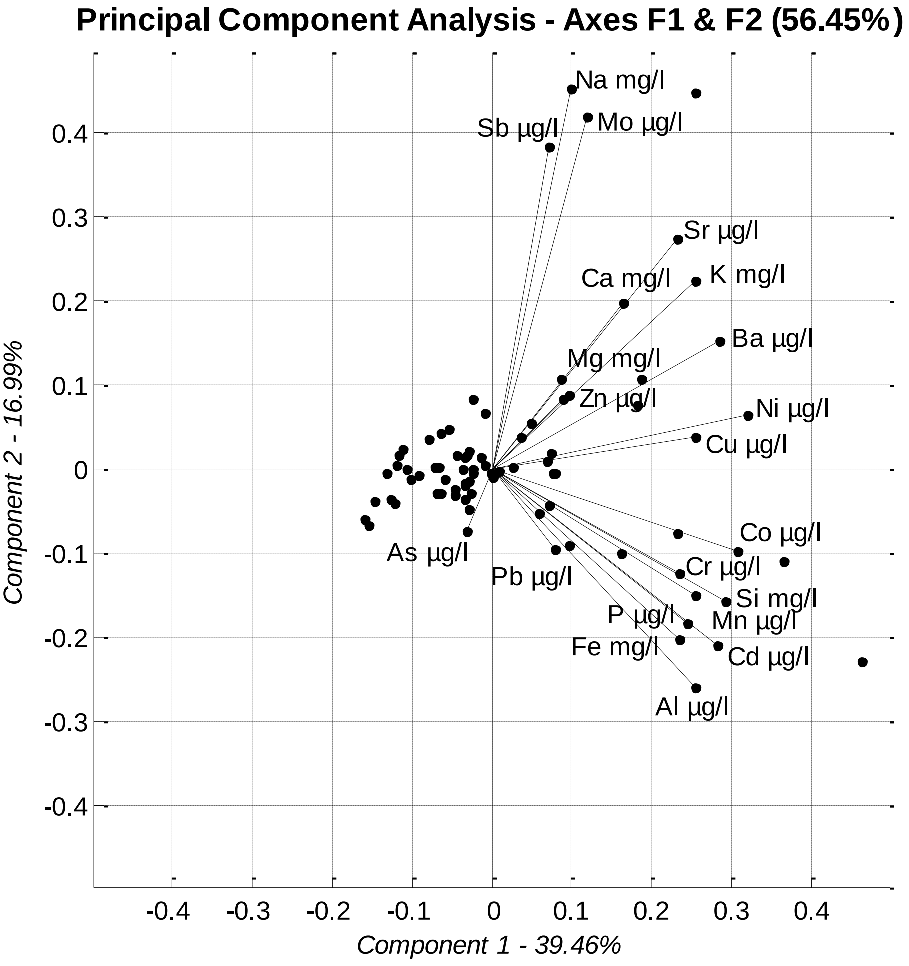
5.1.1.2 Dimension reduction

The dimension reduction technique that was used for the building of the model is Principal Component Analysis (PCA). As explained in the previous chapter, PCA is a method used for dimension reduction of the input variable space. The technique consists of the recursive projection of the observations on planes, orthogonal among them, defined by the axis of maximum inertia, defined by the variance. Thus, the new axes or principal components can hence be used for explaining the input data in a more efficient manner, since they are orthogonal, thus independent, and they maximise the variance of input data.

The purpose of the PCAs was to reduce the total number of variables, reducing redundancy of data. Due to the separation between water and sediment quality data, two different PCAs were made. In Figure 10, the results of the PCA analysis of the water quality data is presented, while in Figure 11, the sediment quality data results are displayed.

In Figure 10, the biplot of the PCA of the water quality input data is displayed. In the figure, the dots represent the projection of the observations in the plane defined by the first and second principal components. The vectors represent the projection of the variance of the input variables in the new plane. The first principal component is the axis with maximum explanation, with a 39.95%. Most of the data are positively correlated with this axis. The second principal component explains the 16.99% of the data variance. This axis is highly correlated with the group of variables formed by Na, Mo and Sb. Three different groups of variables are formed due to common correlation. The group formed by Na, Mo and Sb, correlated to the second component; a group represented by K, Ni and Cu and another represented by Al, Cd and Mn, mainly correlated to the first component.

The explanatory capacity of the two principal components displayed is quite limited, only representing a 56.45% of variance. It seems necessary to increase the number of principal components to explain the input data with sufficient accuracy. An explanatory capacity of 95% is regarded as optimal for the representation of data. Thus, using the scree plot of the PCA of water quality data in Figure 10, nine components are selected. With the selection of these 9 components, approximately the 95% of the variability is explained.



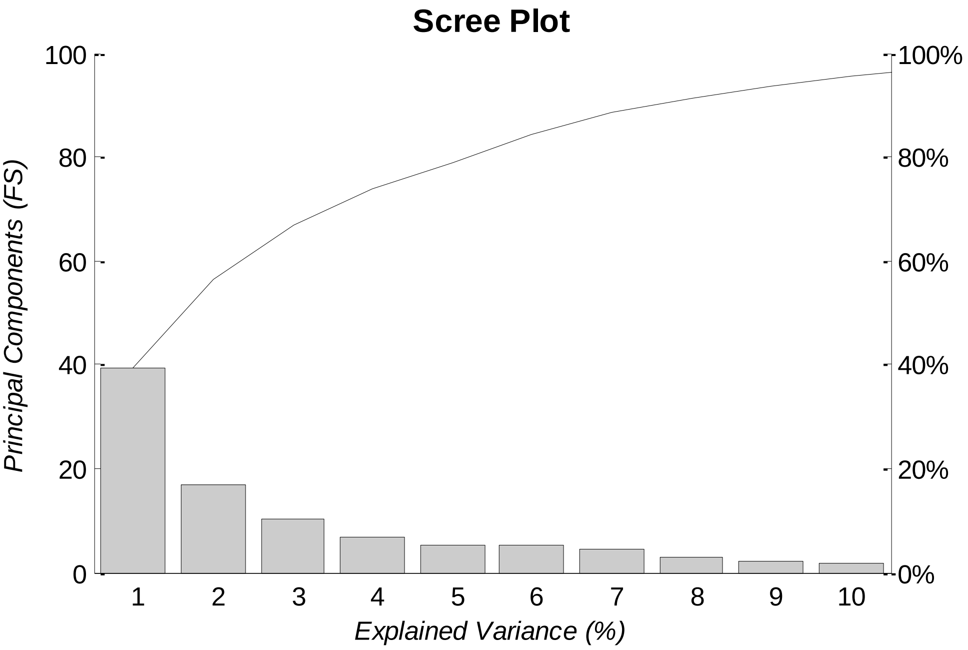
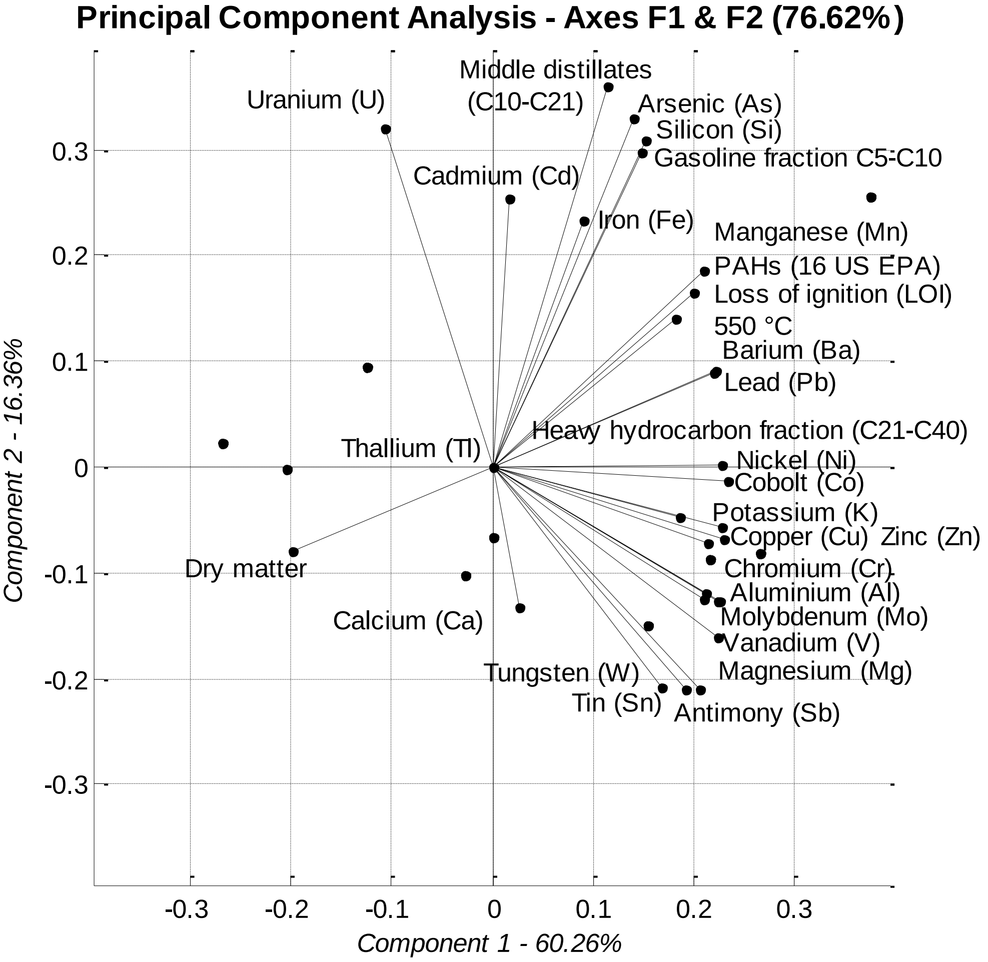


Figure 10. Principal Component Analysis (on top) and scree plot (on bottom) of the principal components of the water quality input data. In PCA the dots represent the projection of the observations whilst the vectors represent the projection of the variance of the variables. In the scree plot, the bars represent the variance explained by each principal component and the line represents the accumulated variance of the principal components.



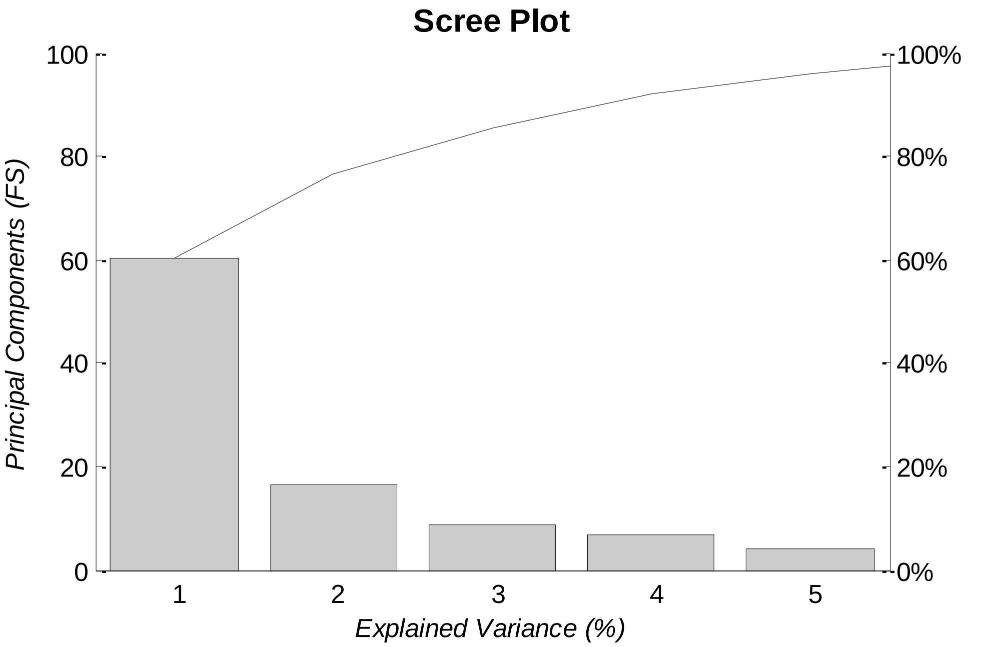


Figure 11. Principal Component Analysis (on top) and scree plot (on bottom) of the principal components of the sediment quality input data. In PCA the dots represent the projection of the observations whilst the vectors represent the projection of the variance of the variables. In the scree plot, the bars represent the variance explained by each principal component and the line represents the accumulated variance of the principal components.

Finally, the results of the PCA of the sediment quality input data are presented in Figure 11. In this case, the first principal component explains the 60.26% of data variance. Again, most of the variables display a positive correlation with the axis that represents the first principal component; only Uranium and Dry matter show negative correlation. The second principal component lowers the explanatory capacity to a more modest 16.36% of data variance. Here, the correlation of the variables with the second principal component is divided. Several groups of variables with similar behaviours can be defined. One group represented by Fe, Si and Gasoline Fraction is mainly correlated with the second principal component. Another group is represented by Pb, Ba and PAHs, and mostly correlates with the first principal component. The same correlation is found in a third large group represented by Cu, Zn, K and Ni.

Despite that the two first principal components explain a 76.62% of data variance, it is still not sufficient for an accurate description of the input data. As it has been said, a level of 95% of variance explained is regarded as an optimal value. For achieving this level of explanatory capacity, a total of 5 principal components have to be selected, as seen in Figure 11.

5.1.1.3 Variable selection

The variable selection has been made following a filter method. The method that has been selected is the simple rank correlation method. The rank correlation is based on the relevance measure determined by the Pearson correlation. The Pearson correlation is defined by the formula:

$$R_{XY} = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\cdot\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \cdot \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}} \qquad (5.1)$$

where *x* corresponds to an input variable and *y* corresponds to a target variable.

The variable selection is performed by ranking the input variables according to their Pearson correlation values R_{XY}. Again, the system was divided depending or not the sediment data were included. The classification was performed using 3 different indicators.

First the Pearson correlation was calculated for every diversity index and input. Then, a ranking of input variables was constructed for each diversity index. Afterwards, a number of input values were chosen for each index. Two methods could be followed, greedy selection or a Person correlation limit value. The greedy selection consists of picking a determined quantity of input variables from the ranking, regardless of any other consideration. The limit value of the Pearson correlation is approximated by the rule of thumb that says that the absolute correlation of a significant variable has to be greater than $2/\sqrt{n}$, with *n* being the total number of observations. Since this last method was more specific than greedy selection, the input variables selected for the model were chosen by the rule of thumb.

Applying the rule of thumb to the obtained Pearson correlations showed to be very strict. This can be due to either the non-linearity of data or due to the little significance of the explanatory variables. Thus, the use of the rule of thumb left very few variables as relevant, and those were different for each of the target variables. Due to this, the selection of the relevant variables was modified to widen the number of them. Also, for the simplification of the modelling process, the same explanatory variables were selected for the three cases.

The final results of the variable selection process for the Water Quality and the Water and Sediment Quality models are presented in Table 2 and Table 3. The results are presented for the three biodiversity indicators. As explained, some basic parameters of the stormwater ponds are added as variables to the model. These variables are the month, Total Organic Carbon (TOC), temperature, dissolved oxygen concentration, pH, conductivity, Average Annual Daily Traffic (AADT), age and size of the pond, the main basin base type, the presence of tunnel wash, the presence of a pre-basin and the presence of a wetland. These variables are compared with the selected principal components from the PCA dimension reduction. The principal components are named as FW, if the principal component comes from the water quality data, and FS, if it comes from the sediment quality data. In the end, the variables selected for the modelling of the neural networks were: temperature, oxygen, pH, conductivity, AADT, main basin type, tunnel wash, pre-basin, wetland, FW3, and FW8, for the water quality variables. For the sediment variables, the variables selected were the same as in water quality plus month, FW2 and FW4, FS1 and FS2, and without FW8.

Table 2. Variable selection for the Water Quality model input data by the Simple Rank Correlation method. Three different biodiversity indicators were used: Species Richness, Shannon and Inverse Simpson. The values of the Pearson correlation greater than the relevancy limit are filled in grey.

	Richness	Shannon	Inv. Simpson
Month	0.070024	0.176744	0.0625287
TOC	0.113789	0.123323	0.0790544
Temperature	0.229134	0.167737	0.1639703
Oxygen	0.339204	0.236533	0.1690786
pH	0.392570	0.250015	0.1472604
Conductivity	0.146081	0.095876	0.3269010
AADT	0.286772	0.370699	0.2137642
Age	0.027684	0.216685	0.1063507
Size	0.025403	0.123065	0.1792595
Main Basin Base Type	0.213663	0.259373	0.2251645
Tunnel Wash	0.006564	0.253139	0.2941399
Pre-basin	0.256170	0.275562	0.3632902
Wetland	0.287597	0.408786	0.0759527
FW1	0.020926	0.027798	0.0787338
FW2	0.069463	0.155503	0.1777572
FW3	0.145075	0.266656	0.0792712
FW4	0.158225	0.175828	0.033959
FW5	0.010363	0.072725	0.1715205
FW6	0.07859	0.185774	0.0203984
FW7	0.183992	0.085986	0.0330937
FW8	0.326304	0.072092	0.0197281
FW9	0.160794	0.023316	0.1537084

Table 3. Variable selection for the Water and Sediment Quality model input data by the Simple Rank Correlation method. Three different biodiversity indicators were used: Species Richness, Shannon and Inverse Simpson. The values of the Pearson correlation greater than the relevancy limit are filled in grey.

	Richness	Shannon	Inv. Simpson
Month	0.059480	0.268860	0.1274302
TOC	0.107384	0.179496	0.1589702
Temperature	0.222589	0.226668	0.0899987
Oxygen	0.335649	0.152752	0.3522663
pH	0.447999	0.392763	0.2756770
Conductivity	0.146014	0.174646	0.2438978
AADT	0.252857	0.294554	0.2601740
Age	0.033166	0.253126	0.0075476
Size	0.098871	0.009512	0.1241008
Main Basin Base Type	0.208168	0.200635	0.2284864
Tunnel Wash	0.021630	0.243983	0.2127753
Pre-basin	0.231152	0.239219	0.2407430
Wetland	0.237969	0.334459	0.0536630
FW1	0.037012	0.032099	0.0499674
FW2	0.039818	0.138068	0.3417106
FW3	0.151969	0.420004	0.2387651
FW4	0.373015	0.323847	0.0563036
FW5	0.037048	0.115842	0.1220343
FW6	0.211058	0.099311	0.0136232
FW7	0.191022	0.033772	0.0759485
FW8	0.235548	0.033887	0.0621896
FW9	0.157334	0.068131	0.1539246
FS1	0.233671	0.291777	0.3184034
FS2	0.021168	0.296446	0.1123885
FS3	0.137614	0.100380	0.0505906

5.1.2 Model 1: Water quality data

5.1.2.1 Description

In this section the first models are built and their performance is assessed. These models were created employing the water quality variables and general data. The specific variable values of all the 12 water ponds used for the creation of this model can be found in Appendix 3. However, the most relevant statistics of the variables can be seen in Table 4. The input variables of the model were selected in the previous section.

The target variables of the model are three biodiversity indices: the Richness factor, the Shannon Index and the Inverse Simpson Index. The specific target values used for the creation of this model can be found in Appendix 2.

Table 4. List of input variables for the Water Quality models. The table includes information about the mean, maximum and minimum values, as well as the standard deviation for each variable.

Variable	Mean Value	Max Value	Min Value	Std Deviation
Temperature (℃)	11.963	19.500	5.800	3.732
Dissolved Oxygen (mg/l)	9.729	17.500	4.550	2.361
pH	7.099	9.740	4.340	1.029
Conductivity (µs/m)	590.862	1836.000	59.000	473.779
AADT	33179.800	66500.000	22735.000	12146.418
Main Basin Type	0.600	0.750	0.250	0.222
Tunnel Wash	0.356	0.667	0.333	0.084
Pre-Basin	0.444	0.667	0.333	0.158
Wetland	0.489	0.667	0.333	0.168
FW3	0.000	4.475	-4.938	1.488
FW8	0.000	2.792	-2.010	0.776

There are two different manners of creating an artificial neural network model in order to provide three different outputs. One is to create a single artificial neural network with an output layer of three neurons and three outputs. The other would consist of building three independent neural networks, each one with a single output. This way, the calibration of the number of hidden neurons in the network would be performed individually for each network, and better results might be achieved. The drawback of the second procedure would be the increased computational time required to train three networks instead of just one. Nevertheless, due to the reduced size of the network, and the limited amount of samples to be processed by the networks, the computational time should not be a problem. Therefore, the decision was to build three different models for each of the diversity indices.

The type of model that was employed is a feedforward artificial neural network. This type of model technique, which is further explained in section 4.5.3, consists of a series of neurons located in one or several layers and in which data only go in one direction.

The feedforward ANN is a two layered ANN. Hence, the ANN contains an input layer, with as many elements as there are variables in the model; one hidden neuron layer, with an indeterminate number of neurons; and an output neuron layer, containing as many neurons as outputs are required by the model, which in this case is just one. The number of hidden neurons was determined by the calibration of the model, which was performed by the use of the cross-validation procedure.

5.1.2.2 Training

The training of the models was done following a 10-fold cross-validation method. Hence, the whole data set was divided in 10 smaller data sets, forming, thus, a 10-fold cross-validation system. This means that the model was trained 10 times, each time with nine different sets for training and one for validation. The squared error of the validation set was calculated in the training of each of the folds and the mean value of the 10 squared errors, the Mean Squared Error (MSE) was obtained. For the training, a two hidden layer neural network was used. The initial weights and bias were randomly selected in the beginning of the process and the values were kept fixed during the rest of the process of training. The cross-validation method was used to determine how many hidden neurons were required for obtaining the best generalisation performance.

A first estimation of the number of hidden neurons required was necessary. This was needed for determining the range in which the calibration was performed. Thus, the number of neurons in the hidden layer was estimated to be between 1 and 20. Hence, for the calibration of the hidden neurons quantity, a total number of 200 different trainings were required. Considering the complexity of the model, such a proceeding is computationally affordable.

The results of the cross-validation process are displayed in Figure 12, Figure 13 and Figure 14. The representation of results is the same in the three figures. The abscise axis represents the number of neurons used in the training of the model. The ordinates axis represents the performance of the neural network in terms of the measured MSE. Three different curves are plotted in the figures, first, the training performance curve (simple dashed line), then the validation performance curve (point-dashed line), and, finally, the total performance curve (filled line). In order to determine the model with the best prediction performance, the number of hidden neurons for which the validation performance is minimized is marked with a circle. The MSE for that specific case is showed in the legend box.

In Figure 12, the performance results of the Richness factor model are shown. In the graph, it can be seen that the validation performance is almost steady for any value of hidden neurons. It can also be seen that the training and the overall performance of the model decreases with the number of neurons. Thus, the best performance results for the training are recorded for the smallest number of hidden neurons. This behaviour can be the result of the noise generated by the number of hidden neurons that are not efficiently employed by the model. Thus, the higher the number of hidden neurons are in the model, the higher the number of hidden neurons that are not used by the model will be, and, therefore, the higher the noise in the output will become. Despite this, the lowest validation performance is recorded for a number of hidden neurons equal to 16.

Figure 13 displays the results of the model with Shannon Index as target variable. The curves in this case are a bit more unsteady compared with those in Figure 12. Nevertheless, the same behaviour can be observed. Neither improvement nor worsening of the validation performance is observed with the increase in number of the hidden neurons. The effect of the increase of the number of hidden neurons in the training and overall performance is the same as in the former case. The increase of neurons implies a reduction of the performance. It is likely that the reason behind this behaviour is, again, the noise of the unemployed hidden neurons. In this case, it is noticeable that the performance of the model for a number of neurons between 1 and 3 is increased. This might indicate that the model requires a larger amount of neurons to explain the target variable. In the end, the best performance of the validation is achieved for the model with 12 hidden neurons.

In Figure 14, the performance of the Inverse Simpson model is displayed. The three performance curves are quite unstable, presenting numerous maximum and minimum peaks. However, again, the behaviour is, on average, similar to the observed in the two previous cases: the validation performance is more or less steady, while the training and overall performances tend to increase with the number of hidden neurons. The best validation performance is obtained for the model containing 6 hidden neurons.

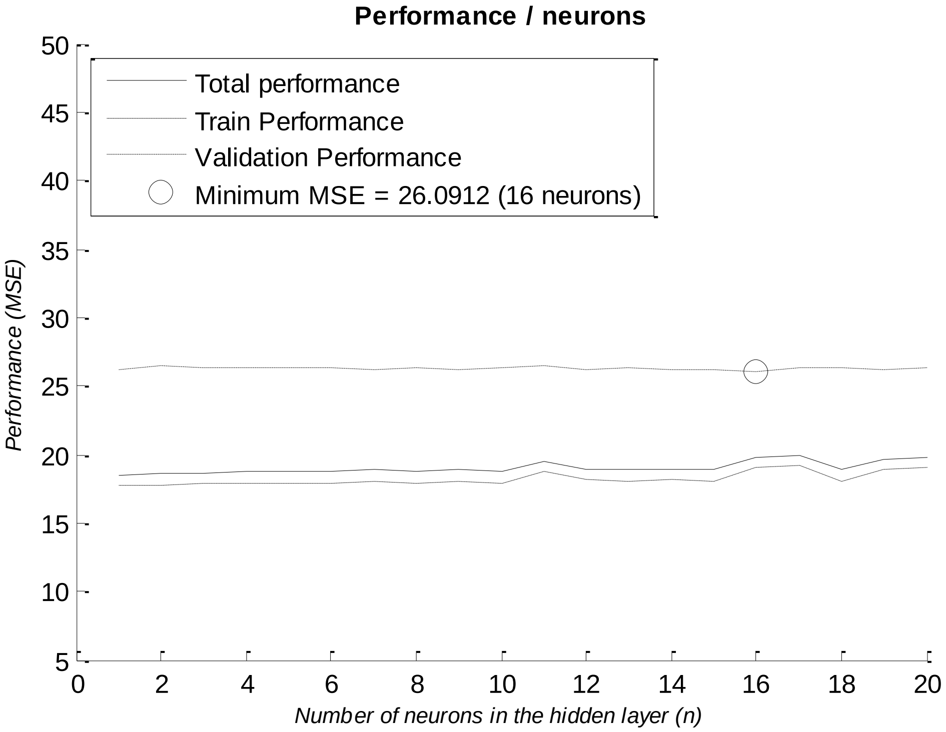


Figure 12. Calibration of the number of hidden neurons for Richness based on the cross-validation error using the water quality data. The simple dashed line represents the error of the training set, the point-dashed line represents the validation error and the filled line represents the combined total error. The minimum cross-validation error is marked with a circle marker.

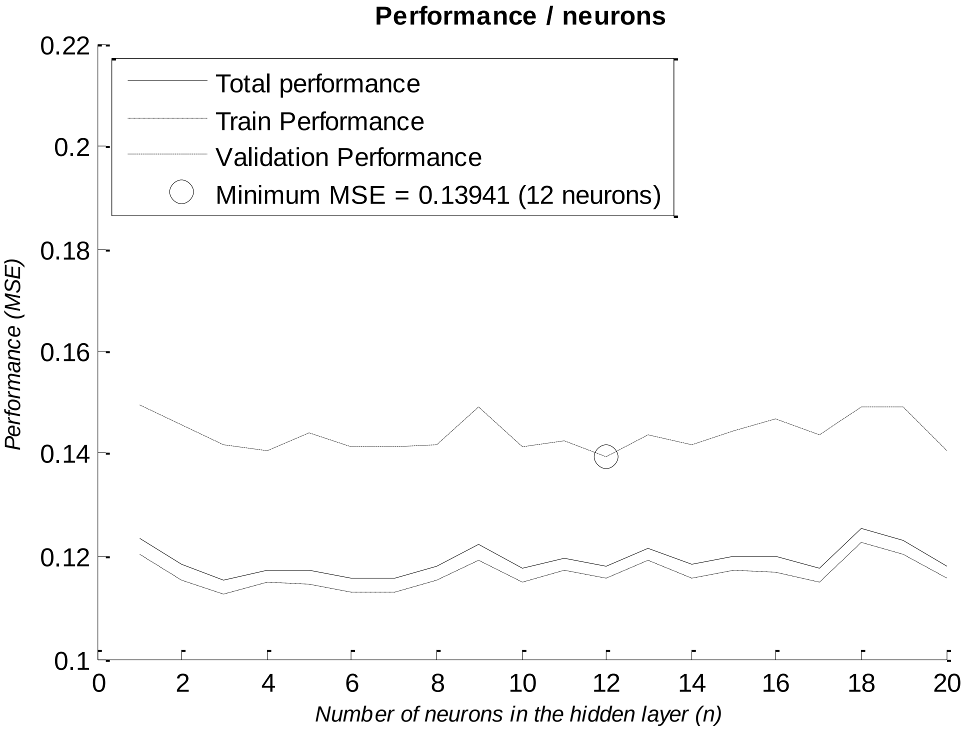


Figure 13. Calibration of the number of hidden neurons for the Shannon Index based on the cross-validation error using the water quality data. The simple dashed line represents the error of the training set, the point-dashed line represents the validation error and the filled line represents the combined total error. The minimum cross-validation error is marked with a circle marker.

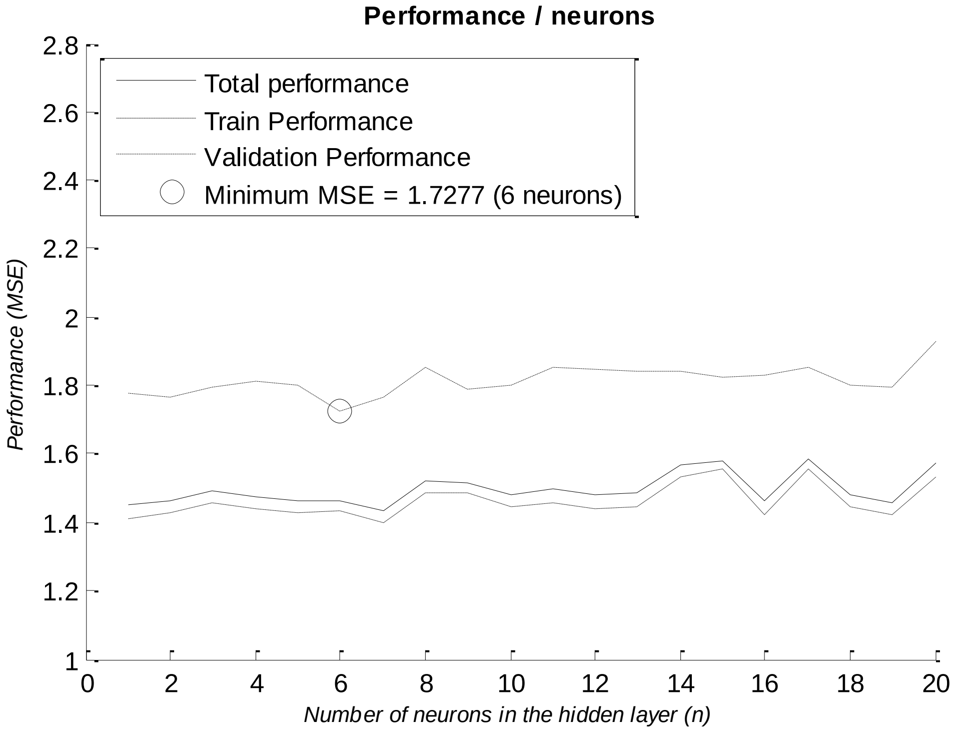


Figure 14. Calibration of the number of hidden neurons for the Inverse Simpson index based on the cross-validation error using the water quality data. The simple dashed line represents the error of the training set, the point-dashed line represents the validation error and the filled line represents the combined total error. The minimum cross-validation error is marked with a circle marker.

5.1.2.3 Model performance

The final performance of the model was measured with the regression factor, *R*, of the outputs of the model and the measured values of the target variables. With the regression factor (*R*), the performance of the three models can be compared between them and, also, with the results of other models found in literature. The advantage of the regression factor compared to the MSE is that the regression factor is dimensionless, and, hence, it can be used to compare models with different outputs.

For the final assessment of the model performance, only the models with the number of hidden neurons that performed the best in the calibration process are employed in this section. Thus, in this section, the mean regression factor values obtained by the application of the cross-validation method are calculated. With those values a comparison can be made between the performances of the different target variables. Also, the regression plots of the models that had the best validation performance among the 10-fold cross validation method are showed.

The mean regression values of the models representing the three target variables, Species Richness, Shannon Index and Inverse Simpson Index are presented in Table 5. The results show that the variable that presents the best predictability is the Species Richness. The regression factor for Species Richness is the best in the overall, the training and the validation performances. This is especially true for the training and overall regressions, which values are almost twice the obtained for the Shannon and Inverse Simpson. The results of the validation regression for Species Richness are more modest, but also better than the Shannon and the Inverse Simpson indices. In the case of Shannon and Inverse Simpson, the results suggest that the Inverse Simpson index might behave better, most notably the validation regression. Nevertheless, the fact that the training regression is lower than the validation regression indicates that the model might not be working as well as the raw numbers state. The validation regression cannot be higher than the training regression. Thus, in this case, the most probable reason for this difference is the lower number of samples used for the estimation of the validation performance. Overall, the results of the regression factors for the three models show to be rather disappointing. The regression factors that were obtained do not predict the target variables with sufficient accuracy. Hence, the use of the models for biodiversity prediction would prove to be unsuccessful.

Table 5. Mean regression values of the ANN models targeting Richness, Shannon Index and Inverse Simpson Index in a 10-fold cross-regression method over the Water quality data. The regression of all the data set is R_{all}; the regression of the training data is R_{train}; and the regression of the validation set is $R_{validation}$. The number of hidden neurons used for the models is specified for each of the three variables.

	Richness (n = 16)	Shannon (n=12)	Inv. Simpson (n=6)
R_{all}	0.6046	0.3222	0.3271
R_{train}	0.6225	0.3195	0.3230
$R_{validation}$	0.4624	0.2966	0.3411

A better glance of what the regression factors displayed in Table 5 look like can be seen in Figure 15, Figure 16 and Figure 17, for the Species Richness; Figure 18, Figure 19 and Figure 20, for the Shannon Index; and Figure 21, Figure 22, and Figure 23, for the Inverse Simpson Index. For each target variable, three figures are provided, one for the training regression, one for the validation performance, and a final figure for overall performance.

In general terms, the behaviour of the three models can be explained using the same terms. The training regression figures show that the performance achieved in this process is quite deficient. In every case, the variance of the model outputs is lower than the variance of the sample data. Graphically, this means that the regression line of the target versus output data has a slope lower than the unity. The models with this behaviour tend to perform well for the average cases but fail when estimations outside the average are required.

The figures presented correspond with those models that present the highest validation performance. The fact that one or more of the cases of the 10 models trained have a high validation performance does not necessary indicate that the model presents good prediction behaviour. One has to take into account when observing these graphs that the displayed validation performance is not the real performance, but that of one of the multiple possible cases employing the available sample data set. A very good validation performance might be due to the use of a set of samples that randomly provide good match with the target variables. A closer approximation to the real performance of a model is obtained with the mean regression values of the cross-validation method.

In spite of this, the results for the training, validation and overall performance are clearly better in the case of the Species Richness variable. This is the same behaviour observed with the mean regression values in Table 5.

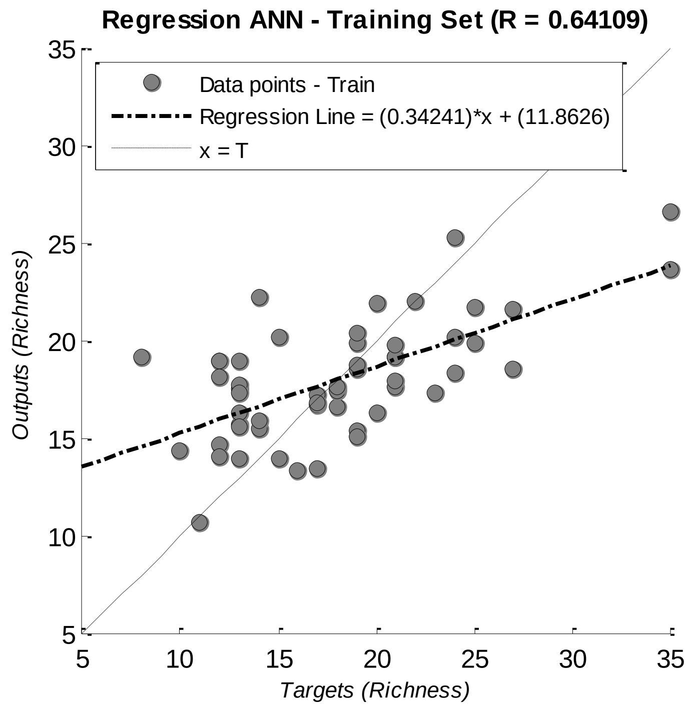


Figure 15. Regression plot of the results of a neural network with 16 neurons and Richness as target variable for the training data set. The filled circles represent the data points, being the abscise axis the observed values and the ordinates axis the outputs of the model. The point-dashed line represents the linear regression, which equation is stated in the legend. The regression factor R is 0.6411.

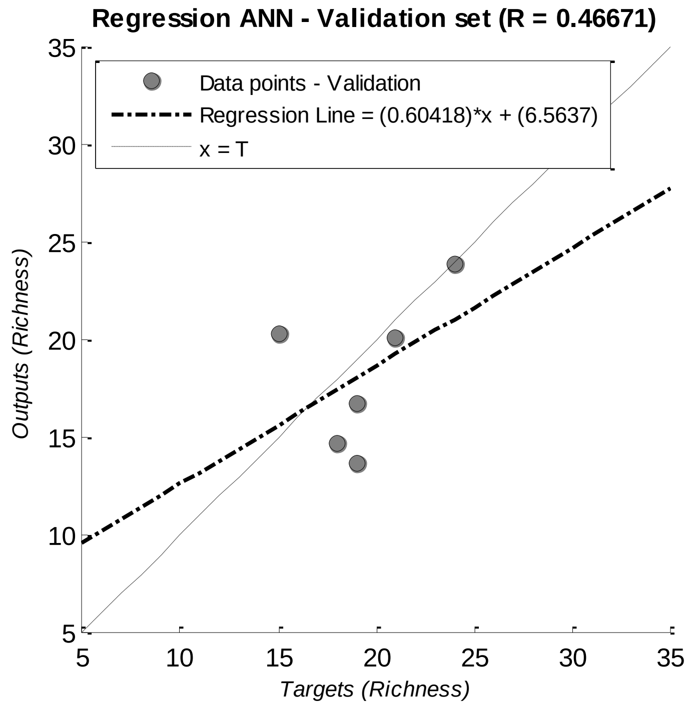


Figure 16. Regression plot of the results of a neural network with 16 neurons and Richness as target variable for the validation data set. The filled circles represent the data points, being the abscise axis the observed values and the ordinates axis the outputs of the model. The point-dashed line represents the linear regression, which equation is stated in the legend. The regression factor R is 0.4667.

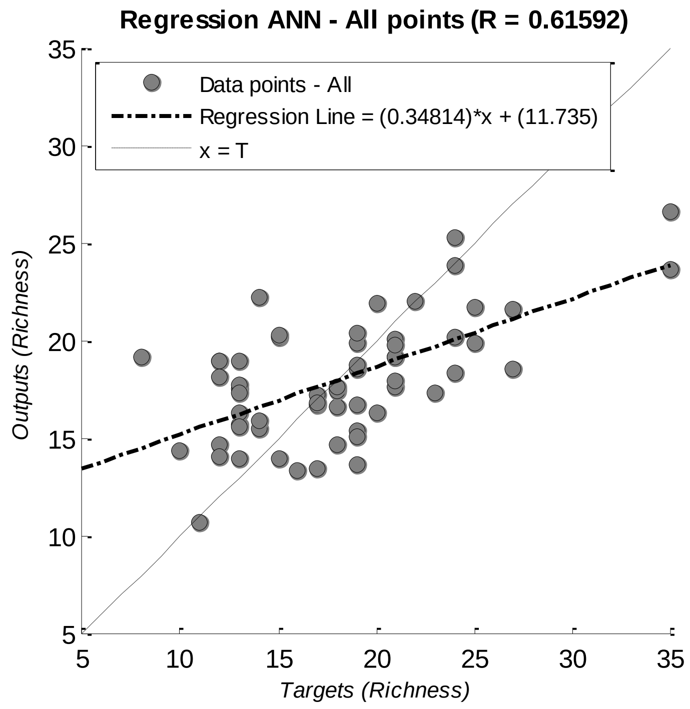


Figure 17. Regression plot of the results of a neural network with 16 neurons and Richness as target variable for the all data. The filled circles represent the data points, being the abscise axis the observed values and the ordinates axis the outputs of the model. The point-dashed line represents the linear regression, which equation is stated in the legend. The regression factor R is 0.6159.

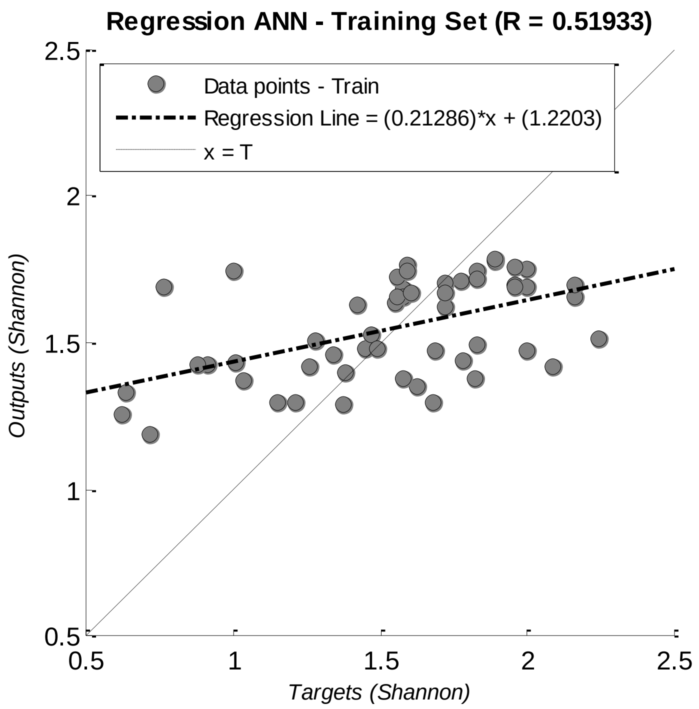


Figure 18. Regression plot of the results of a neural network with 12 neurons and Shannon index as target variable for the training data set. The filled circles represent the data points, being the abscise axis the observed values and the ordinates axis the outputs of the model. The point-dashed line represents the linear regression, which equation is stated in the legend. The regression factor R is 0.5193.

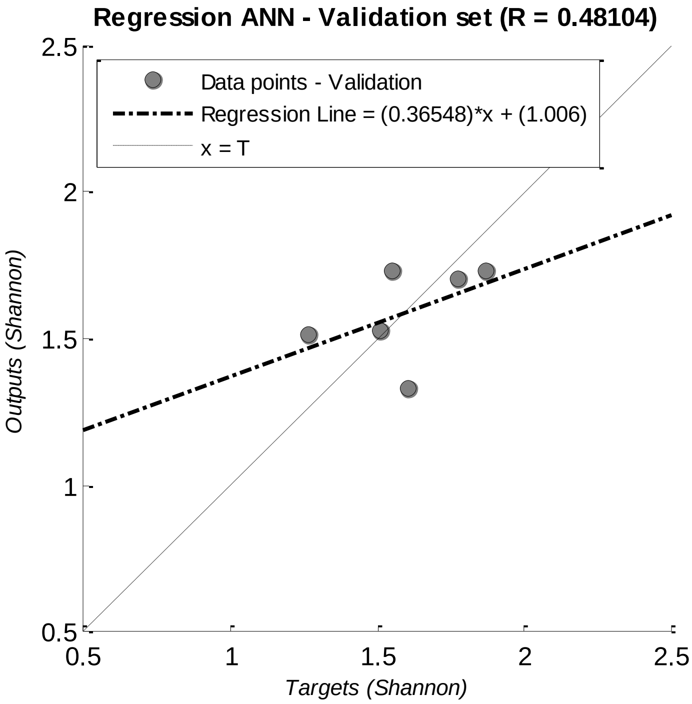


Figure 19. Regression plot of the results of a neural network with 12 neurons and Shannon index as target variable for the validation data set. The filled circles represent the data points, being the abscise axis the observed values and the ordinates axis the outputs of the model. The point-dashed line represents the linear regression, which equation is stated in the legend. The regression factor R is 0.4810.

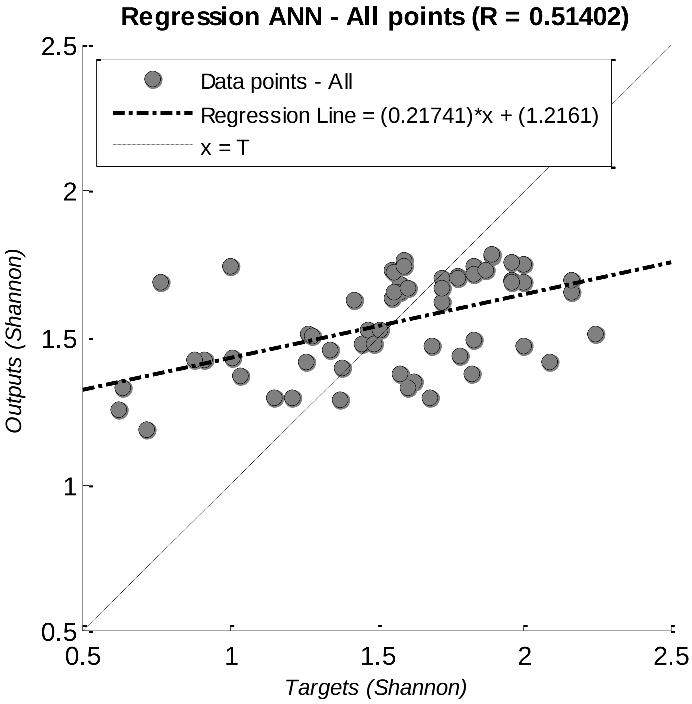


Figure 20. Regression plot of the results of a neural network with 12 neurons and Shannon index as target variable for the all data. The filled circles represent the data points, being the abscise axis the observed values and the ordinates axis the outputs of the model. The point-dashed line represents the linear regression, which equation is stated in the legend. The regression factor R is 0.5140.

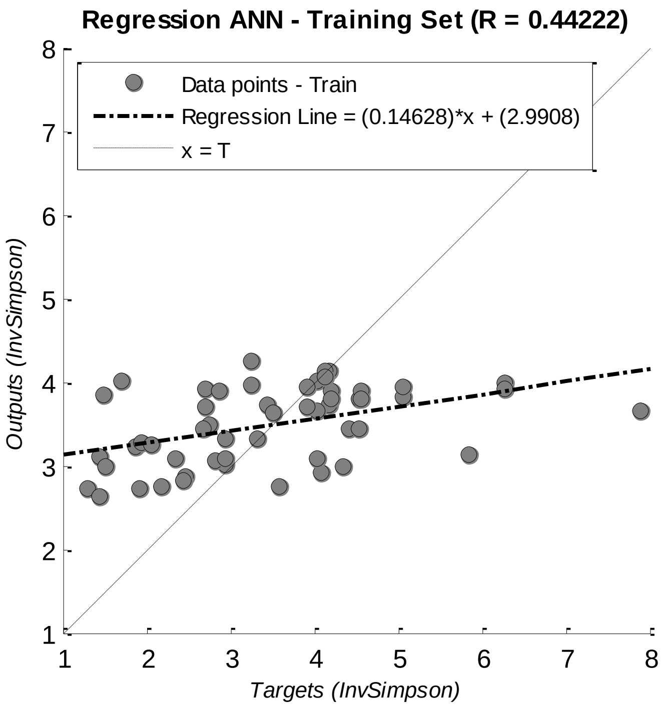


Figure 21. Regression plot of the results of a neural network with 6 neurons and Inverse Simpson index as target variable for the training data set. The filled circles represent the data points, being the abscise axis the observed values and the ordinates axis the outputs of the model. The point-dashed line represents the linear regression, which equation is stated in the legend. The regression factor R is 0.4422.

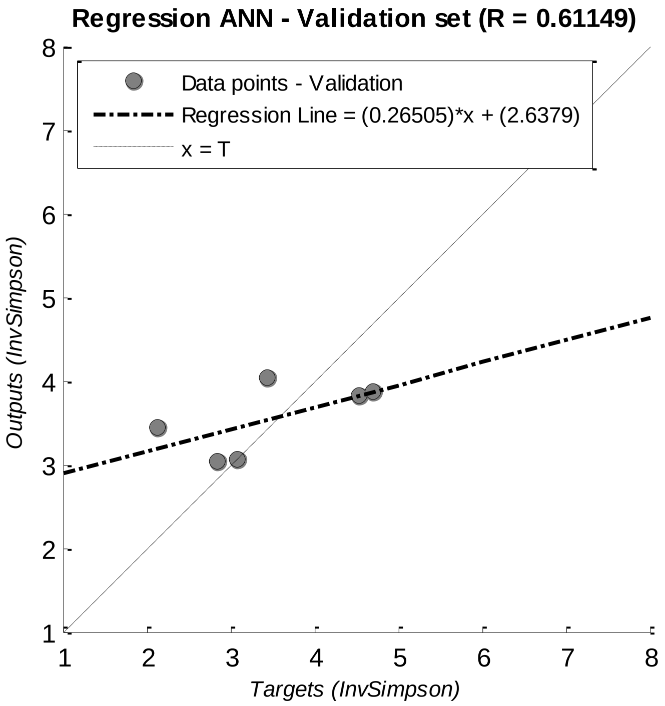


Figure 22. Regression plot of the results of a neural network with 6 neurons and Inverse Simpson index as target variable for the validation data set. The filled circles represent the data points, being the abscise axis the observed values and the ordinates axis the outputs of the model. The point-dashed line represents the linear regression, which equation is stated in the legend. The regression factor R is 0.6115.

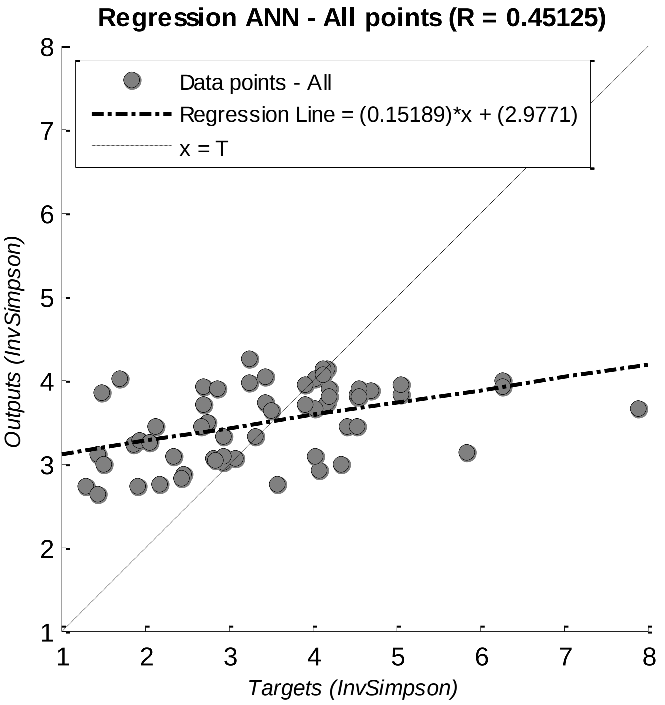


Figure 23. Regression plot of the results of a neural network with 6 neurons and Inverse Simpson index as target variable for the all data. The filled circles represent the data points, being the abscise axis the observed values and the ordinates axis the outputs of the model. The point-dashed line represents the linear regression, which equation is stated in the legend. The regression factor R is 0.4513.

5.1.3 Model 2: Water and sediment quality data

5.1.3.1 Description

In this section the models were created employing the data samples that include water and sediment quality variables, and general data of the ponds. The specific variable values of the 10 ponds used for the creation of this model can be found in Appendix 3. However, the most relevant statistics of the variables can be seen in Table 6. The number of data samples available including all variables is smaller than those only including water quality variables. As in the previous case, the final input variables of this model were selected in the previous section.

Table 6. List of input variables for the Water and Sediment Quality models. The table includes information about the mean, maximum and minimum values, as well as the standard deviation for each variable.

Variable	Mean Value	Max Value	Min Value	Std Deviation
Month	7.000	10.000	4.000	2.262
Temperature (℃)	12.275	18.880	5.800	3.630
Dissolved Oxygen (mg/l)	9.445	17.500	4.550	2.333
pH	7.067	8.590	4.340	0.991
Conductivity (µs/m)	520.191	1420.000	59.000	405.308
AADT	35007.000	66500.000	22735.000	13367.517
Main Basin Type	0.545	0.750	0.250	0.237
Tunnel Wash	0.364	0.667	0.333	0.097
Pre-Basin	0.424	0.667	0.333	0.150
Wetland	0.485	0.667	0.333	0.168

Variable	Mean Value	Max Value	Min Value	Std Deviation
FW2	-0.006	10.817	-5.502	2.088
FW3	-0.405	2.649	-4.938	1.290
FW4	-0.011	1.191	-2.082	0.676
FS1	0.000	7.241	-5.164	4.155
FS2	0.000	4.931	-2.883	2.165

As previously, the target variables of the model are three biodiversity indices, the Richness factor, the Shannon Index and the Inverse Simpson Index. The specific values used for the creation of this model can be found in Appendix 4.Three independent neural networks, each one with a single output, was created and calibrated.

As for the Water Quality model, in section 5.1.2, the type of model that was employed is a feedforward artificial neural network. Again, the feedforward ANN is a two layered ANN. The number of hidden neurons was determined by the calibration of the model, which was performed by the use of the cross-validation procedure.

5.1.3.2 Training

Exactly as for the Water Quality model, the training was done using a 10-fold cross-validation system. The squared error of the validation set was calculated in the training of each of the folds and the mean value of the 10 squared errors, the Mean Squared Error (MSE) was obtained. Again, as a first estimation, the number of hidden neurons required varies between 1 and 20.

In Figure 24, Figure 25 and Figure 26, the calibration of the number of hidden neurons for the three models presented is displayed. The representation of results is the same in the three figures. The abscise axis represents the number of neurons used in the training of the model. The ordinates axis represents the performance of the neural network in terms of the measured MSE. Three different curves are plotted in the figures, first, the training performance curve (simple dashed line), then the validation performance curve (point-dashed line), and, finally, the total performance curve (simple line). In order to determine the model with the best prediction performance, the number of hidden neurons for which the validation performance is minimized is marked with a circle. The MSE for that specific case is showed in the legend box.

The results of the hidden neuron calibration for the Species Richness target variable are presented in Figure 24. It can be observed that the three curves have the same tendency of decreasing performance (increasing MSE) when the number of hidden neurons is increased. This same behaviour was obtained in the previous section, where the water quality samples were employed. The explanation for this phenomenon is again, most likely, due to the noise added by the redundant hidden neurons. These neurons do not contribute to the explanation of the output of the model in any way. However, the value of these neurons after the training process is not exactly zero. This means that there is some contribution of these hidden neurons to the final output, but instead of explaining the output, it consists of random noise.

In Figure 24 it is also noticeable that the training of the model presents a maximum performance around the 2 or 8 neurons in the hidden layer. The maximum performance of the validation set is also obtained in the same range. In the end, the minimum MSE is achieved with a model with 7 neurons, as displayed in the legend.

In Figure 26, the performance graphs of the Shannon Index model are shown. In this case, the behaviour of the curves differs from the seen previously. Thus, the training performance seems to increase with the number of neurons in the hidden layer. The validation performance seems not to be greatly modified by the number of neurons. This curve shows a quite unstable behaviour with several maximum and minimum peaks. One of these peaks presents the absolute minimum MSE, specifically for 20 neurons. Despite that the training performance tends to increase with the number of neurons, the results displayed in all the other cases suggest that the number of parameters in the model is sufficient. Hence, there should not be any improvement in the model if the number of neurons were increased.

The performance graphs of the Simpson Index are represented in Figure 26. The training performance for the Simpson Index target variable shows again a tendency to slightly increase the MSE with the number of neurons. This behaviour suggests that the number of neurons that are fully employed by the model is quite limited. At the same time, all the unemployed neurons contribute to the output with random noise. Thus, the more unemployed hidden neurons, the greater the error gets. Nevertheless, this tendency is not as relevant as in the two other models observed before. The maximum performance is obtained for 6 neurons in the hidden layer.

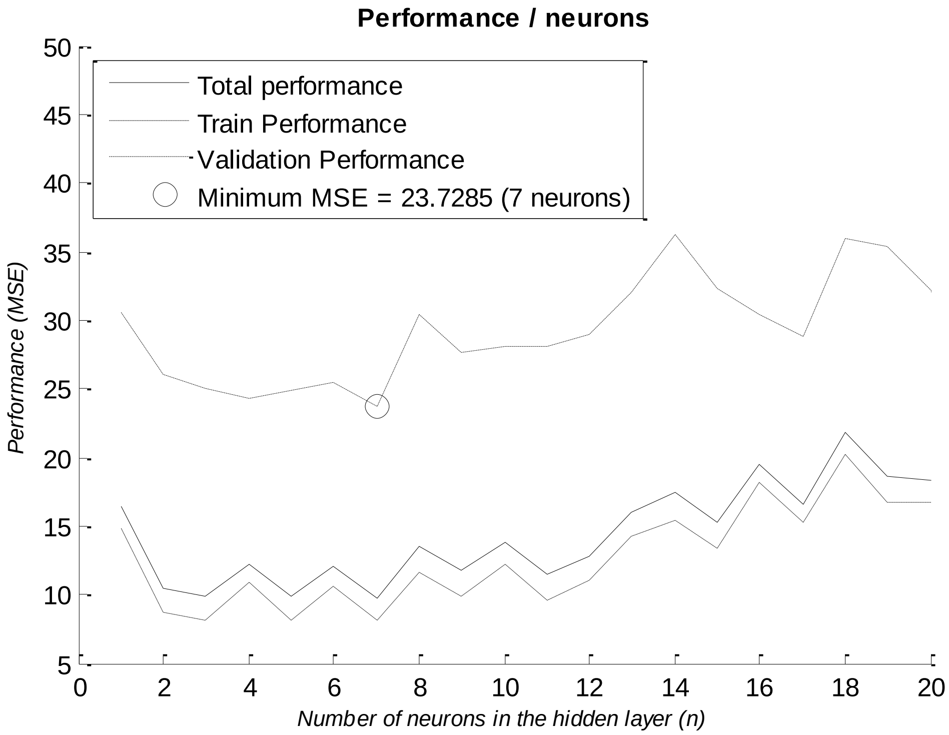


Figure 24. Calibration of the number of hidden neurons for the Richness index based on the cross-validation error using the Water and Sediment quality data. The simple dashed line represents the error of the training set, the point-dashed line represents the validation error and the filled line represents the combined total error. The minimum cross-validation error is marked with a circle marker.

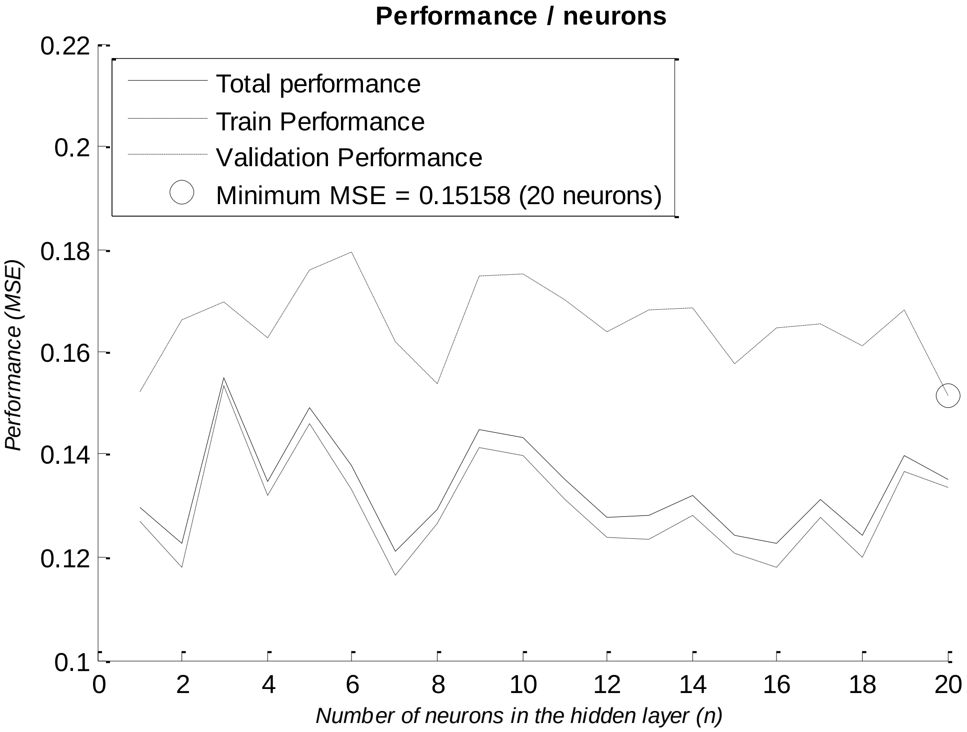


Figure 25. Calibration of the number of hidden neurons for the Shannon index based on the cross-validation error using the Water and Sediment quality data. The simple dashed line represents the error of the training set, the point-dashed line represents the validation error and the filled line represents the combined total error. The minimum cross-validation error is marked with a circle marker.

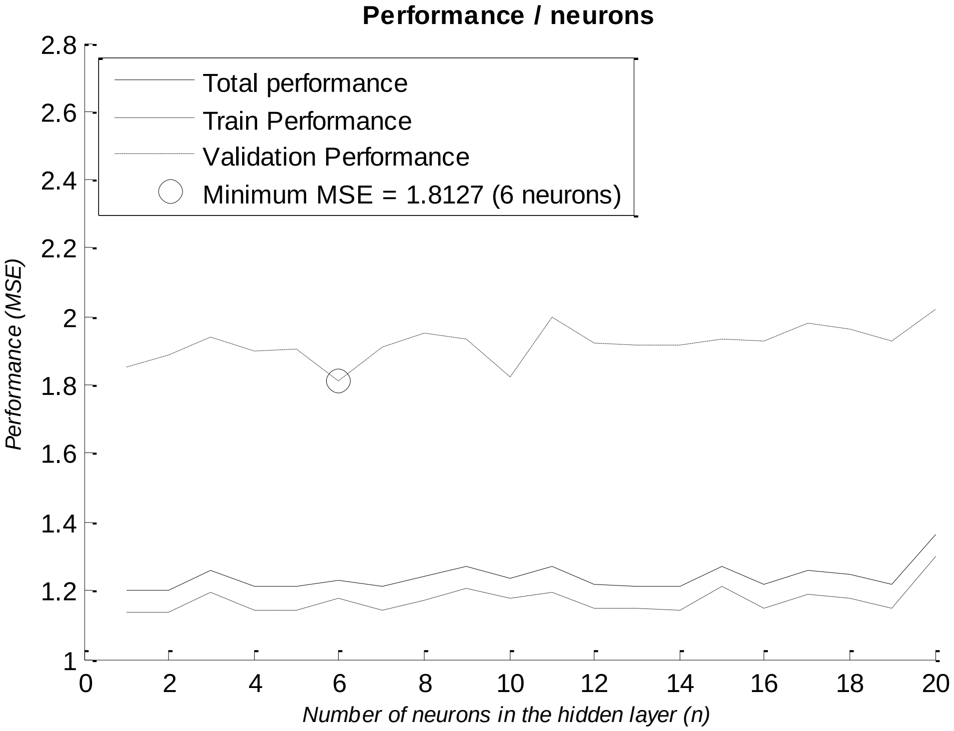


Figure 26. Calibration of the number of hidden neurons for the Inverse Simpson index based on the cross-validation error using the Water and Sediment quality data. The simple dashed line represents the error of the training set, the point-dashed line represents the validation error and the filled line represents the combined total error. The minimum cross-validation error is marked with a circle marker.

5.1.3.3 Model performance

The final performance of the model was measured with the regression factor, *R*, of the outputs of the model and the measured values of the target variables, as it was done in section 5.1.2.3. Again, only the regression plots of the models that had the best validation performance among the 10-fold cross validation method are showed.

The mean regression values of the models predicting the three target variables, Species Richness, Shannon Index and Inverse Simpson Index, are presented in Table 7. As with the water quality models, the results show that the variable that presents the best predictability is the Species Richness. The regression factor for Species Richness is best in the three cases, training, validation, and overall. The difference between the results is remarkable, with a regression up to 0.8763 in the training set. The validation regression is, however, more discrete, with only 0.4688, but still notably higher than the Shannon and Inverse Simpson indices. With all said, the overall performance of the Species Richness model is regarded as quite satisfactory.

Comparing Shannon and Inverse Simpson indices, the results suggest that, in this case, the Shannon index might behave better. This is mainly due to the higher training regression of this index. Overall, the results of these two regression factors showed to be rather disappointing, especially for the training sets. In all the cases, the regression factors that were obtained do not predict the target variables with sufficient accuracy. Hence, the use of models for biodiversity prediction would prove to be unsuccessful.

Table 7. Mean regression values of the ANN models targeting Richness, Shannon Index and Inverse Simpson Index in a 10-fold cross-regression method over the Water and Sediment quality data. The regression of all the data set is R_{all}; the regression of the training data is R_{train}; and the regression of the validation set is $R_{validation}$. The number of hidden neurons used for the models is specified for each of the three variables.

	Richness (n = 7)	Shannon (n = 20)	Inv. Simpson (n = 6)
R_{all}	0.8612	0.5524	0.5646
R_{train}	0.8763	0.5713	0.6028
$R_{validation}$	0.4688	0.3744	0.3419

All the regression plots from Figure 27 to Figure 35 represent the best results of each of the three models. Three figures are presented for each model. First, the training regression, Figure 27, Figure 30 and Figure 33; second, the validation performance, Figure 28, Figure 31 and Figure 34; and last, the overall performance, in Figure 29, Figure 32 and Figure 35.

One has to take into account when observing these graphs that the displayed validation performance is not the real performance, but that of one of the multiple possible cases employing the available sample data set. A very good validation performance might be due to the use of a set of samples that randomly provide good match with the target variables. A closer approximation to the real performance of a model is obtained with the mean regression values of the cross-validation method.

In general terms, the different figures show the same behaviour indicated by the average values in Table 7. Thus, the results of the Species Richness performance are quite good, whilst the Shannon and Inverse Simpson performances are poor. However, as seen in the water quality models, the variance of the model outputs is lower than the variance of the sample data. Graphically, this means that the regression line of the target versus output data has a slope lower than the unity. The models with this behaviour tend to perform well for the average cases but fail when estimations are outside the average.

The results for the training, validation and overall performance are clearly better in the case that the Species Richness variable is used. It is remarkable that the regression factor of the training set is quite close to the mean regression for the same set in Table 7. This might indicate that for this target variable the neural network model shows some stability. This, at the same time, is a sign that the model is working efficiently and that the neurons can find a relationship between variables. A reflection of the good behaviour of this model can be found in Figure 28, where the validation performance is analysed. The regression factor for the validation set is R = 0.86, which is a good result, especially when it is compared with any of the validation performances of the other studied models. With the combination of good training and validation performances, it can be said that the model is capable of predicting with moderate accuracy the Species Richness in a stormwater pond with the given variables.

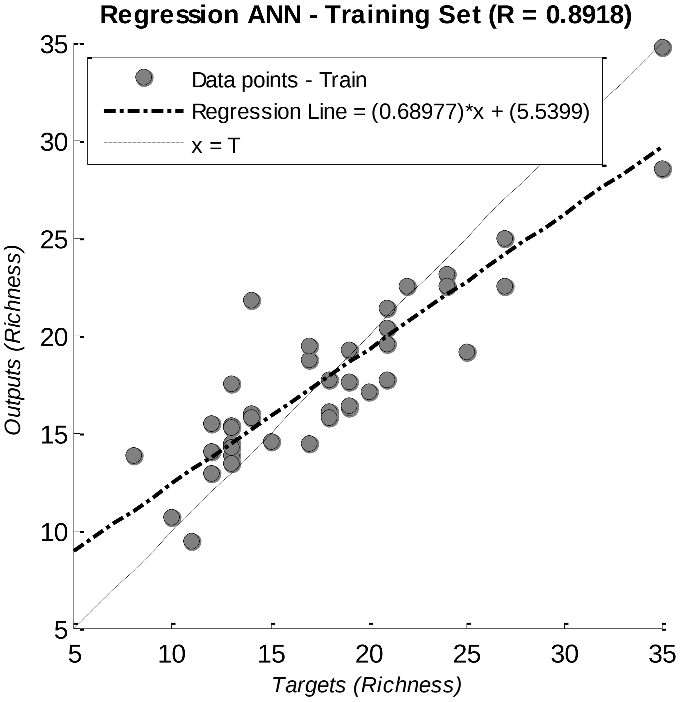


Figure 27. Regression plot of the results of a neural network with 7 neurons and Richness as target variable for the training data set. The filled circles represent the data points, being the abscise axis the observed values and the ordinates axis the outputs of the model. The point-dashed line represents the linear regression, which equation is stated in the legend. The regression factor R is 0.8918.

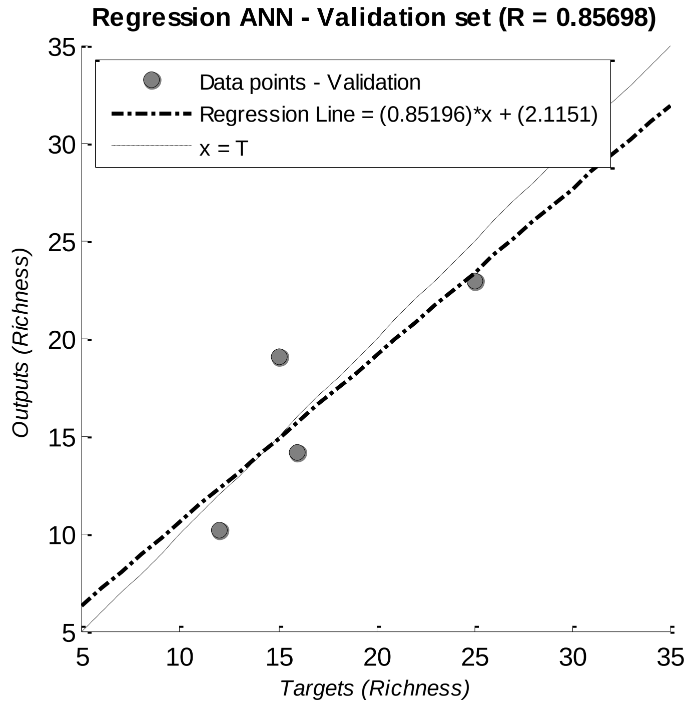


Figure 28. Regression plot of the results of a neural network with 7 neurons and Richness as target variable for the validation data set. The filled circles represent the data points, being the abscise axis the observed values and the ordinates axis the outputs of the model. The point-dashed line represents the linear regression, which equation is stated in the legend. The regression factor R is 0.8570.

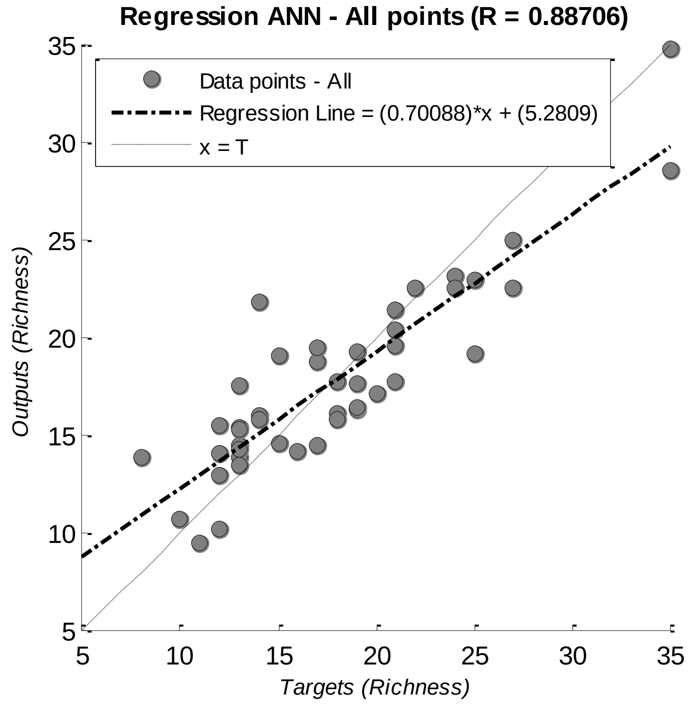


Figure 29. Regression plot of the results of a neural network with 7 neurons and Richness as target variable for the all data. The filled circles represent the data points, being the abscise axis the observed values and the ordinates axis the outputs of the model. The point-dashed line represents the linear regression, which equation is stated in the legend. The regression factor R is 0.8871.

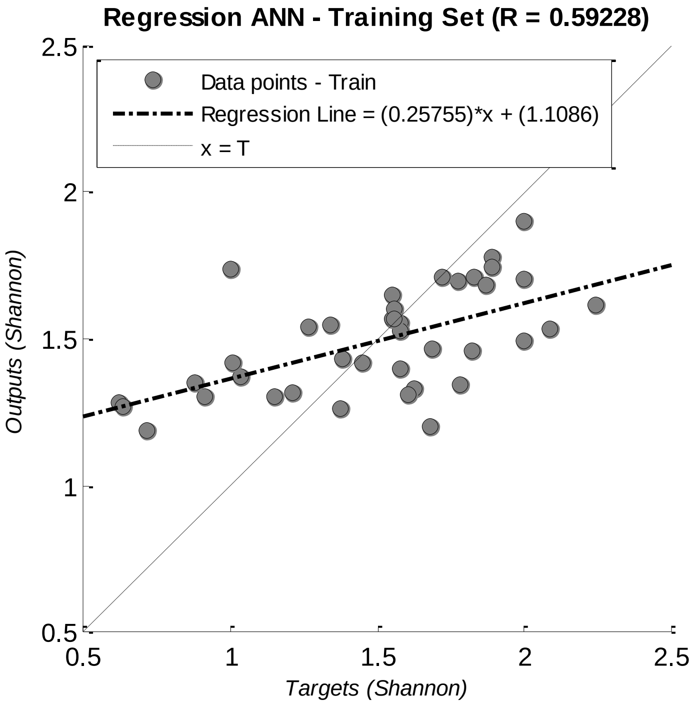


Figure 30. Regression plot of the results of a neural network with 20 neurons and Shannon index as target variable for the training data set. The filled circles represent the data points, being the abscise axis the observed values and the ordinates axis the outputs of the model. The point-dashed line represents the linear regression, which equation is stated in the legend. The regression factor R is 0.5923.

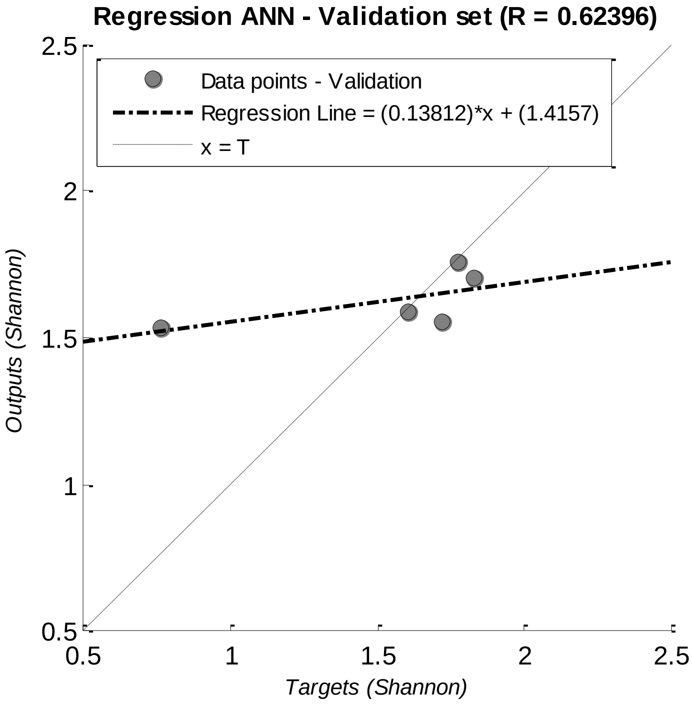


Figure 31. Regression plot of the results of a neural network with 20neurons and Shannon index as target variable for the validation data set. The filled circles represent the data points, being the abscise axis the observed values and the ordinates axis the outputs of the model. The point-dashed line represents the linear regression, which equation is stated in the legend. The regression factor R is 0.6240.

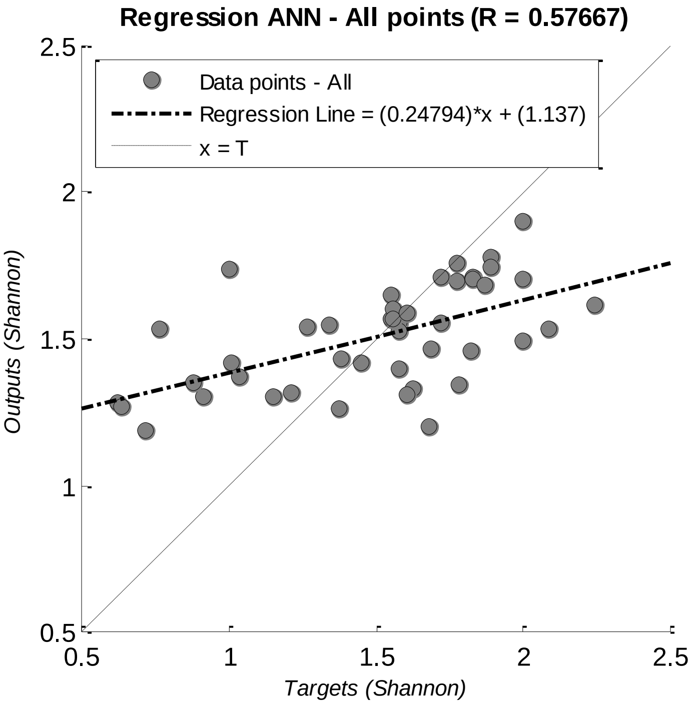


Figure 32. Regression plot of the results of a neural network with 20 neurons and Shannon index as target variable for the all data. The filled circles represent the data points, being the abscise axis the observed values and the ordinates axis the outputs of the model. The point-dashed line represents the linear regression, which equation is stated in the legend. The regression factor R is 0.5767.

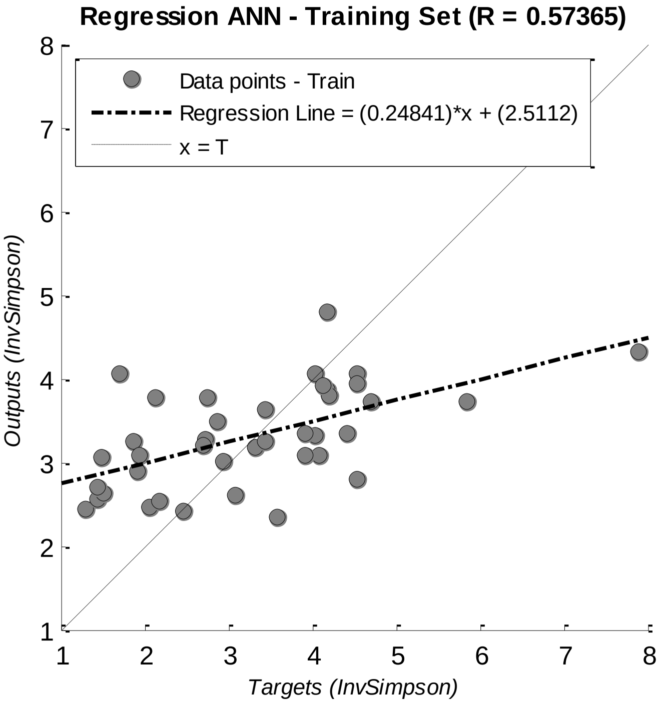


Figure 33. Regression plot of the results of a neural network with 6 neurons and Inverse Simpson index as target variable for the training data set. The filled circles represent the data points, being the abscise axis the observed values and the ordinates axis the outputs of the model. The point-dashed line represents the linear regression, which equation is stated in the legend. The regression factor R is 0.5737.

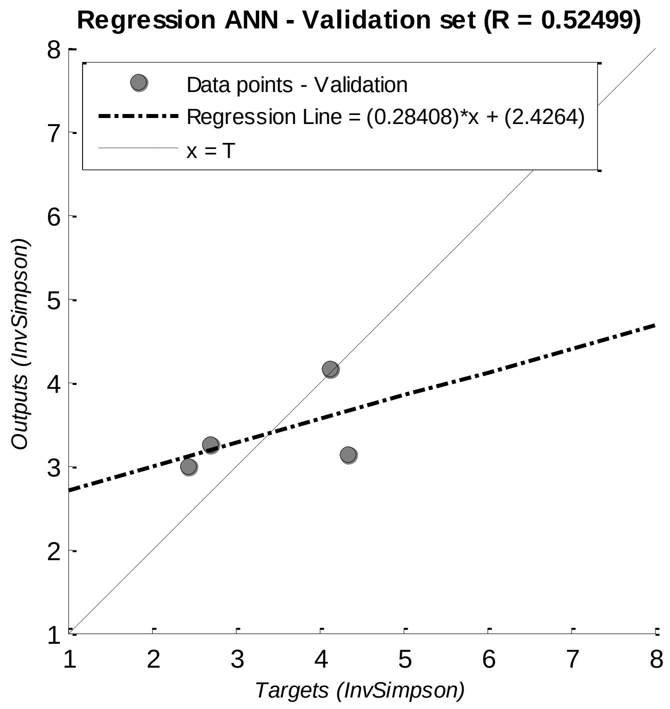


Figure 34. Regression plot of the results of a neural network with 6 neurons and Inverse Simpson index as target variable for the validation data set. The filled circles represent the data points, being the abscise axis the observed values and the ordinates axis the outputs of the model. The point-dashed line represents the linear regression, which equation is stated in the legend. The regression factor R is 0.5250.

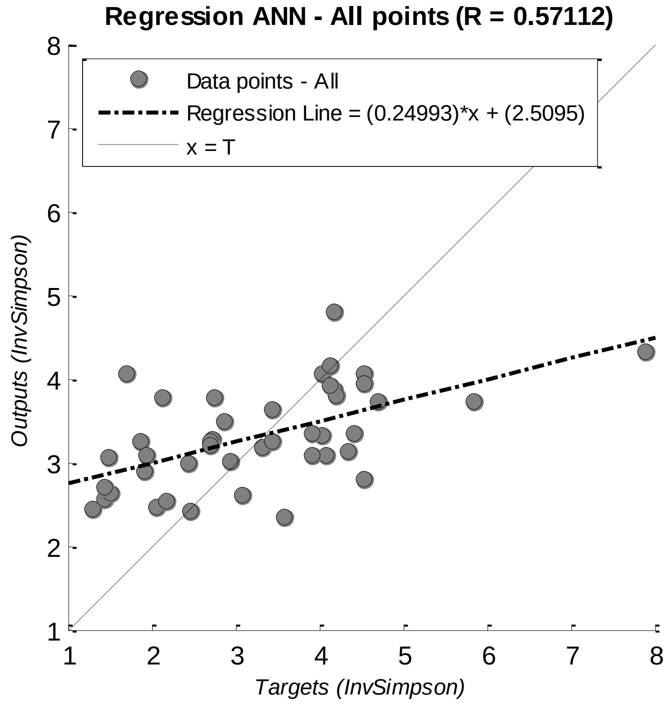


Figure 35. Regression plot of the results of a neural network with 16 neurons and Inverse Simpson index as target variable for the all data. The filled circles represent the data points, being the abscise axis the observed values and the ordinates axis the outputs of the model. The point-dashed line represents the linear regression, which equation is stated in the legend. The regression factor R is 0.5711.

6 Discussion

The discussion is divided in five different points, covering the most relevant parts of this Master Thesis work.

6.1 NORWAT data

All data employed in the models built in this thesis were provided by the NORWAT project. The provided data, which details can be seen in section 4.3, consisted essentially of three major parts, water quality, sediment quality and biological sampling data. Measurements of water quality and biology samples were taken in 11 different stormwater ponds along important roads near Oslo (Norway). The sediment quality was sampled in some of these ponds but not in all of them. Also, the sediment samples were taken only once during the campaign.

The primary aim of these samplings was the monitoring and study of biodiversity in stormwater ponds and the effect that different abiotic and biotic factors play in the development of biodiversity in these ecosystems. The collected data allowed a basic knowledge of the conditions in these ponds and the estimation of average values of different values. In Thygesen (2013) a careful analysis of the results of this campaign was recorded. The outcomes of this research allowed to reveal some relationships between some variables and biodiversity indicators. Further work was developed by Clarke (2014), with the study of a different set of variables. In both cases, the authors pointed out the need of further sampling and study of the possible relations, as the conclusions drawn by the two were not conclusive.

The use of these data in the models developed in this thesis revealed similar problems. One of the most relevant difficulties with the data was the limitation of range of the samples. Most of the variables displayed short ranges of values, and in some of these ranges only some values were taken. One example is the age of the stormwater ponds. The measured ponds had ages between 4 and 13 years. This means that there are 9 different years that can be measured between the 4 and 13 years. However, only 3 years of those 9 were measured, 8, 9 and 12 years. Any possible trend regarding this variable would be regarded as inconclusive. The same would happen with other variables such as the month when the measurements were taken or the type of bottom material. In other cases, the variables present a reasonable range of values, but the number of sample points in between the two range limits is insufficient to draw the real trends of such variables.

The fact that most of the variables might be highly interrelated makes the problem even more complex. A correct definition of these relationships would require accurate and abundant samples, at least, an enough of number of samples to determine with enough precision the relations between the different variables and the biodiversity indices.

Nevertheless, the number of possible factors affecting the biodiversity is enormous. It is possible that most of them might not be discovered, and if they are, it is probable that the resources for measuring them would be too high. Hence, it has to be assumed that a perfect representation of all conditions in these ecosystems is, in fact, impossible.

6.2 Modelling approach

In Chapter 3, a review of the different approaches to model ecology was performed. There was a great amount of material regarding ecological modelling, a science that started around the 1920s and that continues to grow. Several groups of model techniques were identified. Each one has a different aim, a different mathematical approach and a different application range (Jørgensen and Bendoricchio, 2001).

However, the main aim of this project was to study what types of models have been applied to biodiversity prediction. In this specific field very little literature was encountered, and, the few found were very recent. This means that this research is still in its beginning (Jørgensen and Fath, 2011).

As it has been said many times, the representation of any process in a model requires the modeller to identify all variables and relationships (or at least most of them) between them and the desired outcome of the model. In the particular case of biodiversity this is almost impossible to achieve. The biodiversity is affected by numerous, almost uncountable variables, most of them highly interrelated (Ingram and Steel, 2010). Therefore, with situation as described, a model is not easily built following a traditional deterministic empirical approach. It is not possible to define the exact equations and processes happening in reality. Hence, a different approach was required. The answer was provided by the Machine Learning systems (Fielding, 1999). These mathematical tools do not need the user to input the specific equations relating explanatory and target variables. Instead, they find those relationships by themselves. Out of the several Machine Learning methods, the Artificial Neural Networks (ANNs) were chosen for the development of the final model.

Despite the great benefit of not having to introduce specific equations in the model, the Machine Learning methods presented some important drawbacks. The first, and maybe most relevant, is that they are not based on any sort of physical, chemical or biological basis. This is a direct consequence of the former statement. The Machine Learning methods are based on learning, which means that they form relations between variables with examples provided by the user. Hence, the accuracy and validity of the model is related with the training data, which limits the applicability of the model in great manner. For example, when used for function approximation, as it is the case in this project, the Machine Learning methods can only be effectively used for prediction inside the range of values of the data set employed for training the model (Murphy, 2012). Outside these margins, the model has no information to guess how the relationships work. As it is obvious, this downside implies a great limitation to its use, since for a real application it would be required to learn the model every possibility and extreme case.

This leads to another drawback, the learning of data. The Learning Methods require a great amount of training data on examples of real situations to let the model copy the patterns. The number of examples required to accurately define a model depends on the complexity of the process to model, the quality of the samples and the precision required to the model (Murphy, 2012). However, in any of the cases, the amount of examples is very high. In some situations in which the process might not be complex, the amount of resources required for sampling and building the model would be more effectively used trying to implement a different type of model. The cost of the sampling would only be justified if the complexity of the process to model is high enough to be the only available possibility to make the model. Furthermore, since the knowledge of the processes is in many times very limited, there are some variables that are not acknowledged and, hence, not measured and introduced in the model.

A third and final drawback is also related to the sampled data and the involved error. Any sample is always accompanied by some uncertainty. The more accurate a measurement is, the lower the related uncertainty is. However, the uncertainty will never be zero; the measurements will always have some error. This becomes an inconvenience when the only information with which a model is built comes from samples with error. The model could interpret the error as truth and train the model to represent this error as real. A model trained in such a way would lead to inaccurate predictions and bad performance. This behaviour in which the model assumes the error as real information is called *overfitting* (Murphy, 2012). Despite several solutions that are available to reduce or even avoid the problem, it is always an issue that has to be taken into account when employing these methods.

Hence, in the specific case of this project, and with the limited resources of time and a limited knowledge in the ecological and biological field, particularly in biodiversity, the decision was made to proceed with a Machine Learning method. The literature review showed that this approach has been used frequently in recent years, in some cases the situations were comparable to the one in this project (Tamvakis et al., 2014). The complexity of biodiversity modelling justifies the use of this type of model, even with such drawbacks. The data available were regarded as sufficient for a first attempt to create this type of models, with a sufficient amount of measured variables and examples.

6.3 Modelling methodology

The methodology followed for the creation of the models was defined after a thorough review of literature and several examples of usage of Artificial Neural Networks. The construction of a Machine Learning method requires two basic steps: input variable selection and training of the model.

6.3.1 Input variable selection

The input variable selection is a very important step in the definition of the model. The aim of this procedure is to optimise the explicatory capacity of the variables minimising the total number of variables. Essentially, the aim is to produce a set of data that would optimize the training capacity of the model, which is penalised by correlated variables and by the size of the variable space. In section 4.5.2, a detailed description of the specifics of the methodology can be found. In this same section, it is also asserted that the methodology for input variable selection was not closely defined and that several methodologies were found (Reed, 1993, Haykin, 1994, Stoppiglia et al., 2003, Dreyfus, 2005, May et al., 2011, Giordano, 2014). Out of the different classifications of these input variable selection methods, only the clearest one was analysed for the final application (May et al., 2011). This means that there could be more complex or precise methods for this purpose not mentioned in the section.

The input selection method followed in the project consisted of two different parts: dimension reduction and variable selection. The dimension reduction consisted of the reduction of the number of input variables by selecting those that maximise the variance in the variable space and eliminate redundant variables. The variable selection consisted of the selection of the variables, out of the defined in the former step, that were more correlated with the target variables. The chosen methods were Principal Component Analysis (PCA), for dimension reduction, and the Rank Correlation method, for the variable selection.

In essence, these two methods, PCA and Rank Correlation, were chosen due to their relative simplicity and because of multiple references in literature relating to these methods. Despite this fact, the two procedures have quite a few of disadvantages that could penalise the ultimate performance of the models.

In first place, the PCA is a method that consists of linear projections of the variable space in order to maximise the variance of the samples. To achieve that, the PCA creates new variables, which are meant to preserve the variance of the real variables (May et al., 2011). This leads to the substitution of physical, chemical or biological variables into some dimensionless variables, and the knowledge of the importance of the original variables is lost. If no information of the variable relevance can be obtained, any new sample campaign would require measuring every variable, instead of the variables that are really being used. The other important issue with the PCA comes from the linearization of the variables. By performing linear projections based on the variance, the non-linear behaviour of any of the variables, which might be important for the explanation of the biodiversity, would be lost (May et al., 2011). Hence, any exponential or logarithmic trend in the variables would be transformed into a linear trend.

The main drawback of the Rank Correlation method is also related to the linearity of the method. The Rank Correlation method is based on the Pearson Correlation, which is a measure of the linear correlation between two variables. Again, any non-linear behaviour between explanatory and target variables would be penalised (May et al., 2011). It could happen that two variables highly non-linearly correlated were left out of the final model simply because the linear correlation of the two of them was less than the established limit value.

There are several methodologies that account for non-linearity of variables for both of the methods. Some of them also allow the preservation of the original variables during all the process, allowing the identification of those important variables. These methods are usually combinations of the two steps in just one step, such as some complex wrapper methods (*Forward selection, backward selection* or *GA-ANN*), filter methods (*Mutual Information*) and embedded methods (*Recursive Feature Elimination* or *Evolutionary ANNs*). In all cases, the gain in performance that might be achieved by their use comes together with a much higher complexity in calculation (May et al., 2011).

6.3.2 Artificial Neural Network

The definition of an artificial neural network is quite complex. There are multiple configurations for the creation of artificial neural networks. These methods are quite flexible and present a great number of alternative functions and parameters with which to define them. First, the type of network can be decided, feedforward or recurrent, how many layers, how many neurons per layer, which algorithms to use for calibration…

The type of model that was finally employed was a feedforward artificial neural network trained by the backpropagation algorithm. This is a classical and simple type of neural network, which is frequently found in literature (Lek, 1996, Lek, 1999, Karul et al., 2000, Bowden et al., 2006, Kuo et al., 2007, Singh et al., 2009, Lock et al., 2014, Tamvakis et al., 2014). The model was defined with one hidden neural layer. With this configuration an artificial neural network can, as said by the universal approximation theorem, approximate any continuous function given that a sufficient number of neurons is available.

One of the key decisions in the creation of the model was the calibration of the number of neurons in the hidden layer. The procedure to determine the optimal value was the cross-validation method. A 10-fold cross-validation was applied in order to determine the average Mean Squared Error (MSE) of the validation sets of the model. The results obtained after applying the method showed some uneven curves that could be improved by increasing the number of folds (Abu-Mostafa et al., 2012). By doing this, a more accurate approximation of the validation performance could be obtained. On the other hand, increasing the number of folds would, as well, increase the computational time. This increase in resources might also not cause any improvements to the results.

6.4 Analysis of results

Two different types of models were built and tested. The models differed in the variables and number of samples. One of the models employed the water quality data in combination with generic data on the stormwater ponds. The other model added to these variables some sediment quality measurements. Nevertheless, the number of stormwater ponds where the sediment quality samples were taken was fewer than the ponds where the water quality samples were obtained. Hence, the available number of samples containing all variables was fewer than the number of samples only containing water quality variables.

Furthermore, three different biodiversity indices were used: the Species Richness, the Shannon Index and the Inverse Simpson Index. The three indices are commonly employed for the assessment of biodiversity. The three biodiversity measures point to different aspects of biodiversity and are calculated using different procedures.

The purpose of employing different models and biodiversity indices was to determine in which of all these cases the prediction capabilities of the neural networks worked the best. Also, also it could resolve if the models trained and tested could be employed for a real application for the design of new stormwater ponds in order to maximise biodiversity.

The model performance was approximately the same both for Water Quality data and Water and Sediment Quality data. In the case of the MSE, the Water Quality model presents the best performance with the only exception of the Species Richness model, as it can be seen in Table 8. The predictive performance measured with the regression factor provides a different scenario. The results of the regression factor, presented in Table 9, indicate that the best performance is achieved by the Water and Sediment Quality data. Hence, it can be said that there is no clear winner regarding which model performs the best. Thus, the additional variables in the Water and Sediment Quality model seemed not to cause any improvement to the model. It can be inferred as well, that a larger number of samples might cause no noticeable improvement. However, further work should be placed in order to sustain that conclusion.

Table 8. Comparison of the predictive performance, in terms of Mean Squared Error (MSE), between the results of the Water Quality model and Water and Sediment Quality model for the Species Richness, Shannon Index and Inverse Simpson Index.

	Water Quality model	Water and Sediment Quality model
Species Richness	26.09	23.73
Shannon Index	0.14	0.15
Inverse Simpson	1.73	1.81

The results obtained for the model of Species Richness and using Water and Sediment Quality data were the only that showed some useful prediction capacity of biodiversity. Despite the relative success of the model, the results are far from perfect but can be acceptable in some cases.

There are very few examples in literature with the obtained results can be compared. In Tamvakis et al. (2014) a comparison between the performances of three machine learning methods used to model a set of biodiversity indices is performed. In Tamvakis et al. (2014) several biodiversity indices are tested. Luckily, the indices used in this thesis are among the reviewed in Tamvakis et al. (2014)and, hence, their results can be compared. This comparison is presented in Table 9. Nevertheless, the object of the research was the biodiversity in coastal water in the Aegean Sea. Thus, the explanatory variables do not coincide with those employed in this project. Also, despite being the same biodiversity indices, the target variables represent different ecosystems that might influence the final performance. Therefore, the results of Tamvakis et al. (2014) must be taken carefully into consideration when comparing the performances of the models.

Table 9. Comparison of the predictive performance, in terms of regression factor (R), between the results of the Water Quality model, Water and Sediment Quality model and the results presented in Tamvakis et al. (2014) for the Species Richness, Shannon Index and Inverse Simpson Index.

	Tamvakis et al. (2014)	Water Quality model	Water and Sediment Quality model
Species Richness	0.47 – 0.56	0.46	0.47
Shannon Index	0.40 – 0.44	0.30	0.37
Inverse Simpson	0.35 – 0.51	0.34	0.34

The predictive performance obtained in the models built in this project is, in general, worse than the performance in Tamvakis et al. (2014). However, the results reflected in Tamvakis et al. (2014) are not good. The maximum performance obtained is a little over 0.50, and in most of the cases, below that value. The only results that are comparable are the Species Richness and Inverse Simpson.

The comparison of the results also suggests that the Species Richness is the best of the biodiversity indices for model usage. Both Shannon and Inverse Simpson indices present indistinctive performances, neither better nor worse than the other, and, hence, it is unclear which one would be better to use in a model. The same behaviour can be observed in the Water Quality and Water and Sediment Quality model performances.

6.5 Further work

As it has been discussed in the previous sections of this chapter, there are numerous aspects of the project that can be improved. This improvement may lead to a better predictive capacity of the models, but this assertion cannot be proved. It might just occur that after the improvements and changes, the models would continue to have rather poor performance.

The first important improvement that can be made is to use a different type of Machine Learning method. In Tamvakis et al. (2014), the author concluded that the artificial neural networks might not provide the best results among other Machine Learning methods. In fact, the *Instance Based Learning* method (IBk) scored much higher performances than the artificial neural networks. Another Machine Learning method employed in Tamvakis et al. (2014) was the *Model Trees* (MTs), which performance was also better than that of the artificial neural network. Some additional Machine Learning methodologies are further described in Fielding (1999) and Dominguez-Granda et al. (2011).

The sampled data can also be improved. The number of samples plays a great influence in the final performances of the models. Insufficient data can lead to models with poor predictive performance. Other improvement could be to increase the number of variables to analyse and include in the models. There could be factors, not acknowledged yet, that might play a great influence on the biodiversity of the ponds. Therefore, a broader sampling campaign might be advisable, both in number of samples and in number of variables. The precision of the measurements should also be assessed, especially in the case of species counting, which might include a great amount of uncertainty.

Another area that could be well improved is the input variable selection. As explained previously, the methods used for the variable selection in this project present some important drawbacks, as the linearization of the variables and the loss of the original variable identification after this process. Some proposed methods, such as *Forward selection*, *Backward selection* or *GA-ANN*; *Mutual Information*; or *Recursive Feature Elimination*, or *Evolutionary ANNs*, do not present any of those disadvantages (May et al., 2011). However, the complexity of the process would be much increased.

Finally, if the decision is made to use artificial neural networks, some improvements could also be implemented. One enhancement could be the improvement of the generalisation methodology. The methodology employed in this thesis might have been too conservative when applying generalisation methodologies, which might cause the training of the model to be underfitting the variables. By using a more refined generalisation methodology, the training of the model could be driven further to achieve a more accurate model without entering on the grounds of overfitting.

7 Conclusions

A literature review of biodiversity modelling approaches showed that this area of research is still in development. The methodology suggested for this type of model is based on Machine Learning methods. This branch of modelling is quite broad and there are several different methodologies that can be employed. Also, it has its basis in complex statistics that might be not easy to follow from zero.

After the implementation of one of these models, Artificial Neural Network (ANN), some results were obtained and then compared with literature. The results showed predictive performances that are regarded as far from good. The factors of regression (R) are in all the cases below 0.5 and, in most of the cases, even below 0.3. The comparison with literature indicated that the model performance results obtained in the project were worse. Only one of the models performed acceptable, with a regression factor of 0.48.

The final results also indicated that the Water Quality model and the Water and Sediment Quality model offered very similar performances. Hence, further work has to be done on how to improve the performance of the models.

Comparing the three studied biodiversity indices, the one that presented a greater performance was the Species Richness; this is consistent with other studies. The other two indices, the Shannon and Inverse Simpson indices, presented similar results. In this situation, no preference of one of the two indices over the other can be established.

Finally, despite the apparent rather poor performance of the models, it was showed that the Machine Learning methods can be applied to the biodiversity prediction and with some acceptable results. The performance of these models could be improved by using a different type of Machine Learning method, by improving the number of variables and samples, or by enhancing the input variable selection procedures.

8 References

ABE, T., LEVIN, S. A. & HIGASHI, M. 1997. *Biodiversity : an ecological perspective,* New York, Springer.

ABU-MOSTAFA, Y. S., MAGDON-ISMAIL, M. & LIN, H.-T. 2012. *Learning from data,* AMLBook.

ANDREWS, A. 1992. Fragmentation of habitat by roads and utility corridors: A review:A. Andrews, Australian Zoologist, 26(3-4), 1990, pp 130-141. *Biological Conservation,* 59**,** 77-77.

ATLANTA REGIONAL COMMISSION, G. E. P. D. 2001a. *Georgia stormwater management manual. Volume 1,* Atlanta, Ga., Atlanta Regional Commission.

ATLANTA REGIONAL COMMISSION, G. E. P. D. 2001b. *Georgia stormwater management manual. Volume 2,* Atlanta, Ga., Atlanta Regional Commission.

BANDINI, S., MANZONI, S. & VIZZARI, G. 2009. Agent Based Modeling and Simulation. *In:* MEYERS, R. A. (ed.) *Encyclopedia of Complexity and Systems Science.* Springer New York.

BISHOP, C. A., STRUGER, J., BARTON, D. R., SHIROSE, L. J., DUNN, L., LANG, A. L. & SHEPERD, D. 2000a. Contamination and wildlife communities in stormwater detention ponds in Guelph and the Greater Toronto Area, Ontario, 1997 and 1998. Part 1. Wildlife communities. *Water Quality Research Journal of Canada,* 35**,** 399-435.

BISHOP, C. A., STRUGER, J., SHIROSE, L. J., DUNN, L. & CAMPBELL, G. D. 2000b. Contamination and wildlife communities in stormwater detention ponds in Guelph and the Greater Toronto Area, Ontario, 1997, and 1998. Part 2. Contamination and biological effects of contamination. *Water Quality Research Journal of Canada,* 35**,** 437-434.

BOWDEN, G. J., NIXON, J. B., DANDY, G. C., MAIER, H. R. & HOLMES, M. 2006. Forecasting chlorine residuals in a water distribution system using a general regression neural network. *Mathematical and computer modelling,* 44**,** 469-484.

CARR, L. W., FAHRIG, L. & POPE, S. E. 2002. Impacts of landscape transformation by roads. *Applying landscape ecology in biological conservation.* Springer.

CLARKE, J. M. 2014. *A review of the factors affecting the biodiversity of constructed stormwater management systems along roads.* Bachelor of Science Thesis, Chalmers University of Technology.

COFFIN, A. W. 2007. From roadkill to road ecology: A review of the ecological effects of roads. *Journal of Transport Geography,* 15**,** 396-406.

COLWELL, R. K. 2009. Biodiversity: concepts, patterns, and measurement. *The Princeton guide to ecology***,** 257-263.

DOMINGUEZ-GRANDA, L., LOCK, K. & GOETHALS, P. 2011. Application of classification trees to determine biological and chemical indicators for river assessment: case study in the Chaguana watershed (Ecuador). *Journal of Hydroinformatics,* 13**,** 489-499.

DREYFUS, G. 2005. *Neural networks methodology and applications* [Online]. Berlin; New York: Springer.

EUROPEAN ENVIRONMENT AGENCY 2010. *EU 2010 biodiversity baseline,* Copenhagen, European Environment Agency.

FAHLMAN, S. E. & LEBIERE, C. 1989. The cascade-correlation learning architecture.

FIELDING, A. 1999. *Machine learning methods for ecological applications,* Springer.

FORMAN, R. T. T. 2003. *Road ecology : science and solutions,* Washington, DC, Island Press.

FORMAN, R. T. T. A. L. E. 1998. Roads and their major ecological effects. *Annual review of ecology and systematics.,* 29**,** 207.

GERMAN, J. & SVENSSON, G. 2005. Stormwater pond sediments and water—characterization and assessment. *Urban Water Journal,* 2**,** 39-50.

GIORDANO, F. L. R. M. P. C. 2014. Input Variable Selection in Neural Network Models. *Communications in statistics theory and methods,* 43**,** 735-750.

HAGAN, M. T. & MENHAJ, M. B. 1994. Training feedforward networks with the Marquardt algorithm. *Neural Networks, IEEE Transactions on,* 5**,** 989-993.

HAMBY, D. 1994. A review of techniques for parameter sensitivity analysis of environmental models. *Environmental Monitoring and Assessment,* 32**,** 135-154.

HAYKIN, S. 1994. *Neural networks: a comprehensive foundation,* Prentice Hall PTR.

HELSEL, D. R. 1990. Less than obvious-statistical treatment of data below the detection limit. *Environmental Science & Technology,* 24**,** 1766-1774.

HUDSON, M., HAGAN, M. & DEMUTH, B. 2013. *Neural network toolbox for use with MATLAB® : user's guide,* Natick, MathWorks.

HVITVED-JACOBSEN, T., VOLLERTSEN, J. & NIELSEN, A. H. 2011. *Urban and highway stormwater pollution: Concepts and engineering,* CRC Press.

INGRAM, T. & STEEL, M. 2010. Modelling the unpredictability of future biodiversity in ecological networks. *Journal of theoretical biology,* 264**,** 1047-1056.

JOLLIFE, I. 2002. Principal component analysis. Second Edition ed. New York: Springer.

JØRGENSEN, S. E. 1999. State-of-the-art of ecological modelling with emphasis on development of structural dynamic models. *ECOLOGICAL MODELLING,* 120**,** 75-96.

JØRGENSEN, S. E. & BENDORICCHIO, G. 2001. *Fundamentals of ecological modelling,* Elsevier.

JØRGENSEN, S. E. & FATH, B. D. 2011. *Fundamentals of ecological: Modelling applications in environmental management and research* [Online]. Amsterdam; Boston: Elsevier.

KARUL, C., SOYUPAK, S., ÇILESIZ, A. F., AKBAY, N. & GERMEN, E. 2000. Case studies on the use of neural networks in eutrophication modeling. *Ecological modelling,* 134**,** 145-152.

KAZEMI, F., BEECHAM, S. & GIBBS, J. 2009. Streetscale bioretention basins in Melbourne and their effect on local biodiversity. *Ecological Engineering,* 35**,** 1454-1465.

KAZEMI, F., BEECHAM, S. & GIBBS, J. 2011. Streetscape biodiversity and the role of bioretention swales in an Australian urban environment. *Landscape and Urban Planning,* 101**,** 139-148.

KUO, J.-T., HSIEH, M.-H., LUNG, W.-S. & SHE, N. 2007. Using artificial neural network for reservoir eutrophication prediction. *Ecological modelling,* 200**,** 171-177.

LE VIOL, I., CHIRON, F., JULLIARD, R. & KERBIRIOU, C. 2012. More amphibians than expected in highway stormwater ponds. *Ecological Engineering,* 47**,** 146-154.

LE VIOL, I., MOCQ, J., JULLIARD, R. & KERBIRIOU, C. 2009. The contribution of motorway stormwater retention ponds to the biodiversity of aquatic macroinvertebrates. *Biological conservation,* 142**,** 3163-3171.

LEK, S. D. M. B. P. D. I. L. J. A. S. 1996. Application of neural networks to modelling nonlinear relationships in ecology. *ECOLOGICAL MODELLING,* 90**,** 39-52.

LEK, S. G. J. F. 1999. Artificial neural networks as a tool in ecological modelling, an introduction. *ECOLOGICAL MODELLING,* 120**,** 65-73.

LOCK, K., ADRIAENS, T. & GOETHALS, P. 2014. Effect of water quality on blackflies (Diptera: Simuliidae) in Flanders (Belgium). *Limnologica - Ecology and Management of Inland Waters,* 44**,** 58-65.

MAURER, B. A. & MCGILL, B. J. 2011. Measurement of species diversity. *Biological Diversity: Frontiers in measurement and assessment. Oxford University Press Oxford***,** 55-65.

MAY, R., DANDY, G. & MAIER, H. 2011. Review of input variable selection methods for artificial neural networks. *Artificial neural networks—methodological advances and biomedical applications***,** 19-44.

MCGARIGAL, K., CUSHMAN, S. & STAFFORD, S. G. 2000. *Multivariate statistics for wildlife and ecology research*, Springer.

MLADENOFF, D. J. & BAKER, W. L. 1999. *Spatial modeling of forest landscape change: approaches and applications*, Cambridge University Press.

MOORE, T. L. & HUNT, W. F. 2012. Ecosystem service provision by stormwater wetlands and ponds - a means for evaluation? *Water Research,* 46**,** 6811-23.

MURPHY, K. P. 2012. *Machine learning: a probabilistic perspective*, MIT press.

OERTLI, B., DOMINIQUE AUDERSET, J., CASTELLA, E., JUGE, R., CAMBIN, D. & LACHAVANNE, J.-B. 2002. Does size matter? The relationship between pond area and biodiversity. *Biological Conservation Biological Conservation,* 104**,** 59-70.

PEET, R. K. 1974. The measurement of species diversity. *Annual review of ecology and systematics***,** 285-307.

POLITOPOULOS, I. 2007. Review and analysis of agent-based models in biology. *University of Liverpool*.

REED, R. 1993. Pruning algorithms-a survey. *Neural Networks, IEEE Transactions on,* 4**,** 740-747.

SALSKI, A. 2006. Ecological Applications of Fuzzy Logic. *In:* RECKNAGEL, F. (ed.) *Ecological Informatics. Scope, Techniques and Applications* Berlin: Springer.

SANDLUND, O., HINDAR, K. & BROWN, A. 1992. *Conservation of biodiversity for sustainable development:[based on the international conference on the conservation of genetic resources for sustainable development, Roeros, Norway in September 1990]*.

SCHER, O., CHAVAREN, P., DESPREAUX, M. & THIÉRY, A. 2004. Highway stormwater detention ponds as biodiversity islands? *Archives des Sciences,* 57**,** 121-130.

SCHUELER, T. R. & CLAYTOR, R. A. 2000. Maryland Stormwater Design Manual. *Maryland Department of the Environment. Baltimore, MD.*

SEILER, A. 2001. *Ecological Effects of Roads. A review.* Introductory Research Essay, Swedish University of Agricultural Sciences.

SINGH, K. P., BASANT, A., MALIK, A. & JAIN, G. 2009. Artificial neural network modeling of the river water quality—A case study. *Ecological Modelling,* 220**,** 888-895.

ŠMILAUER, P. & LEPŠ, J. 2014. *Multivariate Analysis of Ecological Data Using CANOCO 5,* Cambridge university press.

SNODGRASS, J. W., CASEY, R. E., JOSEPH, D. & SIMON, J. A. 2008. Microcosm investigations of stormwater pond sediment toxicity to embryonic and larval amphibians: variation in sensitivity among species. *Environmental Pollution,* 154**,** 291-297.

SPELLERBERG, I. 1998. Ecological effects of roads and traffic: A literature review. *Global Ecol Biogeography Global Ecology and Biogeography,* 7**,** 317-333.

STERNBECK, J., SJÖDIN, Å. & ANDRÉASSON, K. 2002. Metal emissions from road traffic and the influence of resuspension—results from two tunnel studies. *Atmospheric Environment,* 36**,** 4735-4744.

STOPPIGLIA, H., DREYFUS, G., DUBOIS, R. & OUSSAR, Y. 2003. Ranking a random feature for variable and feature selection. *The Journal of Machine Learning Research,* 3**,** 1399-1414.

TAMVAKIS, A., TRYGONIS, V., MIRITZIS, J., TSIRTSIS, G. & SPATHARIS, S. 2014. Optimizing biodiversity prediction from abiotic parameters. *Environmental Modelling & Software,* 53**,** 112-120.

THYGESEN, H. 2013. *Biodiversity in wet sedimentation ponds constructed for road runoff.* Norwegian University of Life Sciences.

TROMBULAK, S. C. F. C. A. 2000. Review of Ecological Effects of Roads on Terrestrial and Aquatic Communities. *COBI Conservation Biology,* 14**,** 18-30.

UNITED NATIONS 1992. Convention on biological diversity.

VAN DYKE, F. 2008. *Conservation biology: foundations, concepts, applications,* Springer.

WALL, R. 2007. *Ecological conditions of stormwater retention ponds,* University of Delaware.

WINIGER, M. 1998. *Biodiversity: a challenge for development research and policy*, Springer.

WOODS-BALLARD, B., KELLAGHER, R., MARTIN, P., JEFFERIES, C., BRAY, R. & SHAFFER, P. 2007. The SUDS manual. *CIRIA, London*.

ZADEH, L. A. 1968. Fuzzy algorithms. *Information and control*, 12**,** 94-102.

Appendix 1: Input data for the Water Quality model

Treatment pond	Temperature˚C	Oxygen mg/L	pH	conductivity µs/m
Skullerud, field no. 1, April	5,800	10,200	8,590	82,400
Skullerud, field no. 2, June	12,400	9,980	6,170	514,000
Skullerud, field no.3, August	16,220	6,400	7,070	282,000
Skullerud, field no. 4, October	10,610	9,990	7,430	59,000
taraldrud north	8,300	10,100	7,920	95,000
taraldrud north	15,300	11,420	6,060	750,000
taraldrud north	18,530	6,970	6,720	276,000
taraldrud north	10,480	7,460	8,530	976,000
Taraldrud junction	7,800	11,600	8,340	122,200
Taraldrud junction	14,900	10,700	6,200	1350,000
Taraldrud junction	17,400	7,610	7,360	602,000
Taraldrud junction	10,380	9,720	7,530	640,000
Taralrud south	11,500	10,300	8,040	155,600
Taralrud south	16,000	7,890	5,400	1420,000
Taralrud south	18,880	7,120	7,780	498,000
Taralrud south	10,020	8,090	7,620	260,000
Nøstvedt	13,500	10,750	8,330	103,400
Nøstvedt	16,300	8,700	6,230	1191,000
Nøstvedt	18,830	7,700	8,410	257,000
Nøstvedt	10,680	10,070	7,650	206,000
Vassum	10,400	10,350	8,550	133,000
Vassum	16,800	14,400	7,270	656,000
Vassum	15,740	6,250	7,410	1062,000
Vassum	10,090	9,980	7,780	392,000
Idrettsveien , left (V) towards main pond	6,200	6,400	6,550	141,900
Idrettsveien, right (H) towards main pond	7,200	8,780	7,140	250,000
Idrettsveien , left (V) towards main pond	11,200	4,550	4,340	1420,000
Idrettsveien, right (H) towards main pond	11,700	8,600	4,760	266,000
Idrettsveien , left (V) towards main pond	12,980	6,900	6,360	1234,000
Idrettsveien, right (H) towards main pond	13,720	6,810	6,700	324,000
Idrettsveien , left (V) towards main pond	11,020	8,110	7,460	700,000
Idrettsveien, right (H) towards main pond	11,400	8,010	6,910	252,000
Nordby, left (V) inlet	6,300	11,100	7,820	112,300
Nordby, right (H) inlet	6,300	9,080	6,920	164,400
Nordby, left (V) inlet	14,200	13,580	5,850	962,000
Nordby, right (H) inlet	15,400	17,500	5,730	447,000
Nordby, left (V) inlet	12,550	11,160	7,370	856,000
Nordby, right (H) inlet	16,480	9,760	7,220	412,000
Nordby, left (V) inlet	10,920	10,690	6,860	499,000
Nordby, right (H) inlet	10,870	9,890	7,330	305,000

Treatment pond	Temperature˚C	Oxygen mg/L	pH	conductivity µs/m
Enebakk/missingen, inlet	6,300	10,970	7,480	95,200
Enebakk/missingen, inlet	13,000	11,100	5,540	982,000
Enebakk/missingen, inlet	13,960	8,650	7,190	907,000
Enebakk/missingen, inlet	11,540	10,190	7,030	476,000
Fiulstad	5,800	10,100	7,200	137,300
Fiulstad	8,300	11,780	5,400	1234,000
Fiulstad	12,990	7,010	9,740	1067,000
Fiulstad	10,090	9,980	7,780	392,000
Såstad	6,100	10,370	7,590	139,600
Såstad	8,800	12,660	5,600	1812,000
Såstad	12,290	4,700	7,040	1836,000
Såstad	11,090	10,430	7,220	750,000
karlshusbunn, left (V) inlet	8,500	10,470	8,070	122,400
Karlshusbunn, right (H) inlet	6,300	10,650	7,120	328,000
karlshusbunn, left (V) inlet	17,200	12,250	5,880	1717,000
Karlshusbunn, right (H) inlet	16,400	14,500	5,580	495,000
karlshusbunn, left (V) inlet	19,500	13,470	8,530	976,000
Karlshusbunn, right (H) inlet	11,910	9,860	7,790	574,000
karlshusbunn, left (V) inlet	11,520	10,240	7,050	678,000
Karlshusbunn, right (H) inlet	10,860	9,700	7,410	305,000

Treatment pond	AADT	Main Basin Base Type	Tunnel Wash	Pre-basin
Skullerud, field no. 1, April	66500,000	0,250	0,333	0,667
Skullerud, field no. 2, June	66500,000	0,250	0,333	0,667
Skullerud, field no.3, August	66500,000	0,250	0,333	0,667
Skullerud, field no. 4, October	66500,000	0,250	0,333	0,667
taraldrud north	42900,000	0,250	0,333	0,333
taraldrud north	42900,000	0,250	0,333	0,333
taraldrud north	42900,000	0,250	0,333	0,333
taraldrud north	42900,000	0,250	0,333	0,333
Taraldrud junction	42200,000	0,250	0,333	0,333
Taraldrud junction	42200,000	0,250	0,333	0,333
Taraldrud junction	42200,000	0,250	0,333	0,333
Taraldrud junction	42200,000	0,250	0,333	0,333
Taralrud south	42200,000	0,250	0,333	0,333
Taralrud south	42200,000	0,250	0,333	0,333
Taralrud south	42200,000	0,250	0,333	0,333
Taralrud south	42200,000	0,250	0,333	0,333
Nøstvedt	35500,000	0,500	0,333	0,667
Nøstvedt	35500,000	0,500	0,333	0,667
Nøstvedt	35500,000	0,500	0,333	0,667
Nøstvedt	35500,000	0,500	0,333	0,667
Vassum	41000,000	0,750	0,667	0,667
Vassum	41000,000	0,750	0,667	0,667
Vassum	41000,000	0,750	0,667	0,667
Vassum	41000,000	0,750	0,667	0,667
Idrettsveien , left (V) towards main pond	22735,000	0,750	0,333	0,333
Idrettsveien, right (H) towards main pond	22735,000	0,750	0,333	0,333
Idrettsveien , left (V) towards main pond	22735,000	0,750	0,333	0,333
Idrettsveien, right (H) towards main pond	22735,000	0,750	0,333	0,333
Idrettsveien , left (V) towards main pond	22735,000	0,750	0,333	0,333
Idrettsveien, right (H) towards main pond	22735,000	0,750	0,333	0,333
Idrettsveien , left (V) towards main pond	22735,000	0,750	0,333	0,333
Idrettsveien, right (H) towards main pond	22735,000	0,750	0,333	0,333
Nordby, left (V) inlet	22735,000	0,750	0,333	0,333
Nordby, right (H) inlet	22735,000	0,750	0,333	0,333
Nordby, left (V) inlet	22735,000	0,750	0,333	0,333
Nordby, right (H) inlet	22735,000	0,750	0,333	0,333
Nordby, left (V) inlet	22735,000	0,750	0,333	0,333
Nordby, right (H) inlet	22735,000	0,750	0,333	0,333
Nordby, left (V) inlet	22735,000	0,750	0,333	0,333
Nordby, right (H) inlet	22735,000	0,750	0,333	0,333

Treatment pond	AADT	Main Basin Base Type	Tunnel Wash	Pre-basin
Enebakk/missingen, inlet	23837,000	0,750	0,333	0,333
Enebakk/missingen, inlet	23837,000	0,750	0,333	0,333
Enebakk/missingen, inlet	23837,000	0,750	0,333	0,333
Enebakk/missingen, inlet	23837,000	0,750	0,333	0,333
Fiulstad	33575,000	0,750	0,333	0,667
Fiulstad	33575,000	0,750	0,333	0,667
Fiulstad	33575,000	0,750	0,333	0,667
Fiulstad	33575,000	0,750	0,333	0,667
Såstad	33575,000	0,750	0,333	0,667
Såstad	33575,000	0,750	0,333	0,667
Såstad	33575,000	0,750	0,333	0,667
Såstad	33575,000	0,750	0,333	0,667
karlshusbunn, left (V) inlet	22735,000	0,750	0,333	0,333
Karlshusbunn, right (H) inlet	22735,000	0,750	0,333	0,333
karlshusbunn, left (V) inlet	22735,000	0,750	0,333	0,333
Karlshusbunn, right (H) inlet	22735,000	0,750	0,333	0,333
karlshusbunn, left (V) inlet	22735,000	0,750	0,333	0,333
Karlshusbunn, right (H) inlet	22735,000	0,750	0,333	0,333
karlshusbunn, left (V) inlet	22735,000	0,750	0,333	0,333
Karlshusbunn, right (H) inlet	22735,000	0,750	0,333	0,333

Treatment pond	Wetland	FW3	FW8
Skullerud, field no. 1, April	0,333	-0,081	0,393
Skullerud, field no. 2, June	0,333	-0,731	0,263
Skullerud, field no.3, August	0,333	-0,784	0,393
Skullerud, field no. 4, October	0,333	-0,690	0,006
taraldrud north	0,333	-0,202	-0,103
taraldrud north	0,333	-0,792	-0,082
taraldrud north	0,333	-1,199	0,037
taraldrud north	0,333	-0,865	0,039
Taraldrud junction	0,333	1,232	-0,007
Taraldrud junction	0,333	0,774	-0,056
Taraldrud junction	0,333	-0,419	0,167
Taraldrud junction	0,333	-0,426	0,365
Taralrud south	0,333	-0,714	-0,354
Taralrud south	0,333	-0,704	-0,318
Taralrud south	0,333	-0,890	-0,095
Taralrud south	0,333	-1,474	0,065
Nøstvedt	0,333	-1,899	0,478
Nøstvedt	0,333	-1,136	0,182
Nøstvedt	0,333	-1,337	0,373
Nøstvedt	0,333	-1,459	0,814
Vassum	0,333	0,324	0,310
Vassum	0,333	-2,875	-0,283
Vassum	0,333	-4,938	0,164
Vassum	0,333	-1,912	0,959
Idrettsveien , left (V) towards main pond	0,667	1,751	0,761
Idrettsveien, right (H) towards main pond	0,667	-0,365	0,028
Idrettsveien , left (V) towards main pond	0,667	2,649	2,792
Idrettsveien, right (H) towards main pond	0,667	-0,480	-0,142
Idrettsveien , left (V) towards main pond	0,667	1,348	0,617
Idrettsveien, right (H) towards main pond	0,667	-0,165	-0,138
Idrettsveien , left (V) towards main pond	0,667	-0,529	2,472
Idrettsveien, right (H) towards main pond	0,667	-0,420	0,078
Nordby, left (V) inlet	0,667	0,844	-0,383
Nordby, right (H) inlet	0,667	-0,131	-0,784
Nordby, left (V) inlet	0,667	1,152	-0,220
Nordby, right (H) inlet	0,667	0,304	-1,242
Nordby, left (V) inlet	0,667	1,018	-0,158
Nordby, right (H) inlet	0,667	0,833	-1,036
Nordby, left (V) inlet	0,667	0,725	-0,362
Nordby, right (H) inlet	0,667	-2,294	-2,010

Treatment pond	Wetland	FW3	FW8
Enebakk/missingen, inlet	0,667	-0,283	-0,232
Enebakk/missingen, inlet	0,667	-0,138	-0,524
Enebakk/missingen, inlet	0,667	-0,061	-0,128
Enebakk/missingen, inlet	0,667	-0,385	-0,262
Fiulstad	0,333	1,109	0,648
Fiulstad	0,333	2,253	0,049
Fiulstad	0,333	1,676	0,249
Fiulstad	0,333	0,214	2,156
Såstad	0,333	2,112	-1,019
Såstad	0,333	3,373	-1,007
Såstad	0,333	4,475	-0,116
Såstad	0,333	-0,137	-0,514
karlshusbunn, left (V) inlet	0,667	-0,512	-0,051
Karlshusbunn, right (H) inlet	0,667	0,734	-0,633
karlshusbunn, left (V) inlet	0,667	1,187	-0,380
Karlshusbunn, right (H) inlet	0,667	0,974	-0,358
karlshusbunn, left (V) inlet	0,667	-0,921	-0,075
Karlshusbunn, right (H) inlet	0,667	1,138	-0,174
karlshusbunn, left (V) inlet	0,667	-0,798	-0,594
Karlshusbunn, right (H) inlet	0,667	0,947	-1,018

Appendix 2: Target data for the Water Quality model

Treatment pond	Species Richness	Shannon	Inverse Simpson
Skullerud, field no. 1, April	11,000	1,381	2,945
Skullerud, field no. 2, June	14,000	1,822	4,347
Skullerud, field no.3, August	13,000	0,618	1,305
Skullerud, field no. 4, October	19,000	1,627	2,459
taraldrud north	19,000	1,684	3,315
taraldrud north	24,000	2,003	4,424
taraldrud north	24,000	1,606	3,082
taraldrud north	16,000	1,781	4,527
Taraldrud junction	14,000	1,263	2,121
Taraldrud junction	15,000	2,243	7,901
Taraldrud junction	18,000	1,035	1,872
Taraldrud junction	13,000	1,448	2,717
Taralrud south	13,000	1,005	1,932
Taralrud south	19,000	1,343	2,739
Taralrud south	17,000	0,635	1,451
Taralrud south	18,000	0,912	2,051
Nøstvedt	10,000	1,580	4,092
Nøstvedt	13,000	2,090	5,853
Nøstvedt	19,000	1,377	2,181
Nøstvedt	15,000	0,875	1,522
Vassum	13,000	1,208	1,922
Vassum	22,000	1,149	2,437
Vassum	20,000	0,715	1,445
Vassum	18,000	1,680	3,577
Idrettsveien , left (V) towards main pond	17,000	1,773	4,527
Idrettsveien, right (H) towards main pond	17,000	1,773	4,527
Idrettsveien , left (V) towards main pond	27,000	2,001	4,172
Idrettsveien, right (H) towards main pond	27,000	2,001	4,172
Idrettsveien , left (V) towards main pond	12,000	1,723	4,035
Idrettsveien, right (H) towards main pond	12,000	1,723	4,035
Idrettsveien , left (V) towards main pond	12,000	1,578	3,917
Idrettsveien, right (H) towards main pond	12,000	1,578	3,917
Nordby, left (V) inlet	21,000	1,829	4,193
Nordby, right (H) inlet	21,000	1,829	4,193
Nordby, left (V) inlet	35,000	1,889	4,130
Nordby, right (H) inlet	35,000	1,889	4,130
Nordby, left (V) inlet	25,000	1,551	3,432
Nordby, right (H) inlet	25,000	1,551	3,432
Nordby, left (V) inlet	21,000	1,559	2,702
Nordby, right (H) inlet	21,000	1,559	2,702

Treatment pond	Species Richness	Shannon	Inverse Simpson
Enebakk/missingen, inlet	13,000	1,873	4,694
Enebakk/missingen, inlet	14,000	1,000	1,694
Enebakk/missingen, inlet	13,000	1,608	2,866
Enebakk/missingen, inlet	8,000	0,764	1,478
Fiulstad	13,000	1,471	2,810
Fiulstad	15,000	1,423	2,675
Fiulstad	17,000	1,260	2,352
Fiulstad	12,000	1,488	2,934
Såstad	23,000	1,512	2,836
Såstad	20,000	1,721	3,516
Såstad	18,000	1,281	2,930
Såstad	18,000	1,833	4,045
karlshusbunn, left (V) inlet	19,000	1,959	5,066
Karlshusbunn, right (H) inlet	19,000	1,959	5,066
karlshusbunn, left (V) inlet	24,000	1,591	3,245
Karlshusbunn, right (H) inlet	24,000	1,591	3,245
karlshusbunn, left (V) inlet	19,000	2,161	6,272
Karlshusbunn, right (H) inlet	19,000	2,161	6,272
karlshusbunn, left (V) inlet	21,000	1,958	4,558
Karlshusbunn, right (H) inlet	21,000	1,958	4,558

Appendix 3: Input data for the Water and Sediment Quality model

Treatment pond	Month	Temperature ˚C	Oxygen mg/L	pH
Skullerud, field no. 1, April	4	5,800	10,200	8,590
Skullerud, field no. 2, June	6	12,400	9,980	6,170
Skullerud, field no.3, August	8	16,220	6,400	7,070
Skullerud, field no. 4, October	10	10,610	9,990	7,430
taraldrud north	4	8,300	10,100	7,920
taraldrud north	6	15,300	11,420	6,060
taraldrud north	8	18,530	6,970	6,720
taraldrud north	10	10,480	7,460	8,530
Taraldrud junction	4	7,800	11,600	8,340
Taraldrud junction	6	14,900	10,700	6,200
Taraldrud junction	8	17,400	7,610	7,360
Taraldrud junction	10	10,380	9,720	7,530
Taralrud south	4	11,500	10,300	8,040
Taralrud south	6	16,000	7,890	5,400
Taralrud south	8	18,880	7,120	7,780
Taralrud south	10	10,020	8,090	7,620
Nøstvedt	4	13,500	10,750	8,330
Nøstvedt	6	16,300	8,700	6,230
Nøstvedt	8	18,830	7,700	8,410
Nøstvedt	10	10,680	10,070	7,650
Vassum	4	10,400	10,350	8,550
Vassum	6	16,800	14,400	7,270
Vassum	8	15,740	6,250	7,410
Vassum	10	10,090	9,980	7,780
Idrettsveien , left (V) towards main pond	4	6,200	6,400	6,550
Idrettsveien, right (H) towards main pond	4	7,200	8,780	7,140
Idrettsveien , left (V) towards main pond	6	11,200	4,550	4,340
Idrettsveien, right (H) towards main pond	6	11,700	8,600	4,760
Idrettsveien , left (V) towards main pond	8	12,980	6,900	6,360
Idrettsveien, right (H) towards main pond	8	13,720	6,810	6,700
Idrettsveien , left (V) towards main pond	10	11,020	8,110	7,460
Idrettsveien, right (H) towards main pond	10	11,400	8,010	6,910

Treatment pond	Month	Temperature˚C	Oxygen mg/L	pH
Nordby, left (V) inlet	4	6,300	11,100	7,820
Nordby, right (H) inlet	4	6,300	9,080	6,920
Nordby, left (V) inlet	6	14,200	13,580	5,850
Nordby, right (H) inlet	6	15,400	17,500	5,730
Nordby, left (V) inlet	8	12,550	11,160	7,370
Nordby, right (H) inlet	8	16,480	9,760	7,220
Nordby, left (V) inlet	10	10,920	10,690	6,860
Nordby, right (H) inlet	10	10,870	9,890	7,330
Enebakk/missingen, inlet	4	6,300	10,970	7,480
Enebakk/missingen, inlet	6	13,000	11,100	5,540
Enebakk/missingen, inlet	8	13,960	8,650	7,190
Enebakk/missingen, inlet	10	11,540	10,190	7,030

Treatment pond	conductivity µs/m	AADT	Main Basin Base Type	Tunnel Wash
Skullerud, field no. 1, April	82,400	66500,000	0,250	0,333
Skullerud, field no. 2, June	514,000	66500,000	0,250	0,333
Skullerud, field no.3, August	282,000	66500,000	0,250	0,333
Skullerud, field no. 4, October	59,000	66500,000	0,250	0,333
taraldrud north	95,000	42900,000	0,250	0,333
taraldrud north	750,000	42900,000	0,250	0,333
taraldrud north	276,000	42900,000	0,250	0,333
taraldrud north	976,000	42900,000	0,250	0,333
Taraldrud junction	122,200	42200,000	0,250	0,333
Taraldrud junction	1350,000	42200,000	0,250	0,333
Taraldrud junction	602,000	42200,000	0,250	0,333
Taraldrud junction	640,000	42200,000	0,250	0,333
Taralrud south	155,600	42200,000	0,250	0,333
Taralrud south	1420,000	42200,000	0,250	0,333
Taralrud south	498,000	42200,000	0,250	0,333
Taralrud south	260,000	42200,000	0,250	0,333
Nøstvedt	103,400	35500,000	0,500	0,333
Nøstvedt	1191,000	35500,000	0,500	0,333
Nøstvedt	257,000	35500,000	0,500	0,333
Nøstvedt	206,000	35500,000	0,500	0,333
Vassum	133,000	41000,000	0,750	0,667
Vassum	656,000	41000,000	0,750	0,667
Vassum	1062,000	41000,000	0,750	0,667
Vassum	392,000	41000,000	0,750	0,667
Idrettsveien , left (V) towards main pond	141,900	22735,000	0,750	0,333
Idrettsveien, right (H) towards main pond	250,000	22735,000	0,750	0,333
Idrettsveien , left (V) towards main pond	1420,000	22735,000	0,750	0,333
Idrettsveien, right (H) towards main pond	266,000	22735,000	0,750	0,333
Idrettsveien , left (V) towards main pond	1234,000	22735,000	0,750	0,333
Idrettsveien, right (H) towards main pond	324,000	22735,000	0,750	0,333
Idrettsveien , left (V) towards main pond	700,000	22735,000	0,750	0,333
Idrettsveien, right (H) towards main pond	252,000	22735,000	0,750	0,333

Treatment pond	conductivity µs/m	AADT	Main Basin Base Type	Tunnel Wash
Nordby, left (V) inlet	112,300	22735,000	0,750	0,333
Nordby, right (H) inlet	164,400	22735,000	0,750	0,333
Nordby, left (V) inlet	962,000	22735,000	0,750	0,333
Nordby, right (H) inlet	447,000	22735,000	0,750	0,333
Nordby, left (V) inlet	856,000	22735,000	0,750	0,333
Nordby, right (H) inlet	412,000	22735,000	0,750	0,333
Nordby, left (V) inlet	499,000	22735,000	0,750	0,333
Nordby, right (H) inlet	305,000	22735,000	0,750	0,333
Enebakk/missingen, inlet	95,200	23837,000	0,750	0,333
Enebakk/missingen, inlet	982,000	23837,000	0,750	0,333
Enebakk/missingen, inlet	907,000	23837,000	0,750	0,333
Enebakk/missingen, inlet	476,000	23837,000	0,750	0,333

Treatment pond	Pre-basin	Wetland	FW2	FW3	FW4
Skullerud, field no. 1, April	0,667	0,333	0,374	-0,081	0,849
Skullerud, field no. 2, June	0,667	0,333	0,086	-0,731	0,673
Skullerud, field no.3, August	0,667	0,333	-0,171	-0,784	0,887
Skullerud, field no. 4, October	0,667	0,333	-1,605	-0,690	-0,012
taraldrud north	0,333	0,333	0,115	-0,202	-0,368
taraldrud north	0,333	0,333	-0,090	-0,792	-0,057
taraldrud north	0,333	0,333	-1,459	-1,199	-0,580
taraldrud north	0,333	0,333	-0,855	-0,865	0,055
Taraldrud junction	0,333	0,333	0,856	1,232	-0,678
Taraldrud junction	0,333	0,333	1,151	0,774	-0,524
Taraldrud junction	0,333	0,333	-0,013	-0,419	-0,014
Taraldrud junction	0,333	0,333	-0,105	-0,426	0,326
Taralrud south	0,333	0,333	0,546	-0,714	-0,375
Taralrud south	0,333	0,333	0,424	-0,704	-0,334
Taralrud south	0,333	0,333	-0,923	-0,890	-0,063
Taralrud south	0,333	0,333	-0,994	-1,474	0,353
Nøstvedt	0,667	0,333	0,001	-1,899	0,750
Nøstvedt	0,667	0,333	1,056	-1,136	0,234
Nøstvedt	0,667	0,333	-0,311	-1,337	0,502
Nøstvedt	0,667	0,333	-1,165	-1,459	1,191
Vassum	0,667	0,333	1,982	0,324	-0,013
Vassum	0,667	0,333	10,817	-2,875	-1,243
Vassum	0,667	0,333	2,597	-4,938	0,659
Vassum	0,667	0,333	1,345	-1,912	1,161
Idrettsveien , left (V) towards main pond	0,333	0,667	0,029	1,751	-1,297
Idrettsveien, right (H) towards main pond	0,333	0,667	-0,663	-0,365	-0,099
Idrettsveien , left (V) towards main pond	0,333	0,667	-0,133	2,649	-2,082
Idrettsveien, right (H) towards main pond	0,333	0,667	-0,746	-0,480	-0,110
Idrettsveien , left (V) towards main pond	0,333	0,667	-0,209	1,348	-0,727
Idrettsveien, right (H) towards main pond	0,333	0,667	-0,667	-0,165	-0,036
Idrettsveien , left (V) towards main pond	0,333	0,667	-2,412	-0,529	0,254
Idrettsveien, right (H) towards main pond	0,333	0,667	-0,854	-0,420	0,262

Treatment pond	Pre-basin	Wetland	FW2	FW3	FW4
Nordby, left (V) inlet	0,333	0,667	0,376	0,844	-0,349
Nordby, right (H) inlet	0,333	0,667	-2,191	-0,131	0,697
Nordby, left (V) inlet	0,333	0,667	0,351	1,152	-0,132
Nordby, right (H) inlet	0,333	0,667	-0,551	0,304	-0,320
Nordby, left (V) inlet	0,333	0,667	0,520	1,018	0,068
Nordby, right (H) inlet	0,333	0,667	-0,263	0,833	-0,295
Nordby, left (V) inlet	0,333	0,667	-0,049	0,725	0,398
Nordby, right (H) inlet	0,333	0,667	-5,502	-2,294	-1,344
Enebakk/missingen, inlet	0,333	0,667	-0,319	-0,283	0,006
Enebakk/missingen, inlet	0,333	0,667	0,318	-0,138	0,117
Enebakk/missingen, inlet	0,333	0,667	0,092	-0,061	0,196
Enebakk/missingen, inlet	0,333	0,667	-1,039	-0,385	0,917

Treatment pond	FS1	FS2
Skullerud, field no. 1, April	-0,501	-1,952
Skullerud, field no. 2, June	-0,501	-1,952
Skullerud, field no.3, August	-0,501	-1,952
Skullerud, field no. 4, October	-0,501	-1,952
taraldrud north	4,170	-1,678
taraldrud north	4,170	-1,678
taraldrud north	4,170	-1,678
taraldrud north	4,170	-1,678
Taraldrud junction	7,241	4,931
Taraldrud junction	7,241	4,931
Taraldrud junction	7,241	4,931
Taraldrud junction	7,241	4,931
Taralrud south	5,135	-1,581
Taralrud south	5,135	-1,581
Taralrud south	5,135	-1,581
Taralrud south	5,135	-1,581
Nøstvedt	0,002	-1,288
Nøstvedt	0,002	-1,288
Nøstvedt	0,002	-1,288
Nøstvedt	0,002	-1,288
Vassum	2,989	-2,883
Vassum	2,989	-2,883
Vassum	2,989	-2,883
Vassum	2,989	-2,883
Idrettsveien , left (V) towards main pond	-2,387	1,818
Idrettsveien, right (H) towards main pond	-2,387	1,818
Idrettsveien , left (V) towards main pond	-2,387	1,818
Idrettsveien, right (H) towards main pond	-2,387	1,818
Idrettsveien , left (V) towards main pond	-2,387	1,818
Idrettsveien, right (H) towards main pond	-2,387	1,818
Idrettsveien , left (V) towards main pond	-2,387	1,818
Idrettsveien, right (H) towards main pond	-2,387	1,818

Treatment pond	FS1	FS2
Nordby, left (V) inlet	-5,164	0,428
Nordby, right (H) inlet	-5,164	0,428
Nordby, left (V) inlet	-5,164	0,428
Nordby, right (H) inlet	-5,164	0,428
Nordby, left (V) inlet	-5,164	0,428
Nordby, right (H) inlet	-5,164	0,428
Nordby, left (V) inlet	-5,164	0,428
Nordby, right (H) inlet	-5,164	0,428
Enebakk/missingen, inlet	-3,933	-0,039
Enebakk/missingen, inlet	-3,933	-0,039
Enebakk/missingen, inlet	-3,933	-0,039
Enebakk/missingen, inlet	-3,933	-0,039

Appendix 4: Target data for the Water and Sediment Quality model

Treatment pond	Species Richness	Shannon	Inverse Simpson
Skullerud, field no. 1, April	11,000	1,381	2,945
Skullerud, field no. 2, June	14,000	1,822	4,347
Skullerud, field no.3, August	13,000	0,618	1,305
Skullerud, field no. 4, October	19,000	1,627	2,459
taraldrud north	19,000	1,684	3,315
taraldrud north	24,000	2,003	4,424
taraldrud north	24,000	1,606	3,082
taraldrud north	16,000	1,781	4,527
Taraldrud junction	14,000	1,263	2,121
Taraldrud junction	15,000	2,243	7,901
Taraldrud junction	18,000	1,035	1,872
Taraldrud junction	13,000	1,448	2,717
Taralrud south	13,000	1,005	1,932
Taralrud south	19,000	1,343	2,739
Taralrud south	17,000	0,635	1,451
Taralrud south	18,000	0,912	2,051
Nøstvedt	10,000	1,580	4,092
Nøstvedt	13,000	2,090	5,853
Nøstvedt	19,000	1,377	2,181
Nøstvedt	15,000	0,875	1,522
Vassum	13,000	1,208	1,922
Vassum	22,000	1,149	2,437
Vassum	20,000	0,715	1,445
Vassum	18,000	1,680	3,577
Idrettsveien , left (V) towards main pond	17,000	1,773	4,527
Idrettsveien, right (H) towards main pond	17,000	1,773	4,527
Idrettsveien , left (V) towards main pond	27,000	2,001	4,172
Idrettsveien, right (H) towards main pond	27,000	2,001	4,172
Idrettsveien , left (V) towards main pond	12,000	1,723	4,035
Idrettsveien, right (H) towards main pond	12,000	1,723	4,035
Idrettsveien , left (V) towards main pond	12,000	1,578	3,917
Idrettsveien, right (H) towards main pond	12,000	1,578	3,917

Treatment pond	Species Richness	Shannon	Inverse Simpson
Nordby, left (V) inlet	21,000	1,829	4,193
Nordby, right (H) inlet	21,000	1,829	4,193
Nordby, left (V) inlet	35,000	1,889	4,130
Nordby, right (H) inlet	35,000	1,889	4,130
Nordby, left (V) inlet	25,000	1,551	3,432
Nordby, right (H) inlet	25,000	1,551	3,432
Nordby, left (V) inlet	21,000	1,559	2,702
Nordby, right (H) inlet	21,000	1,559	2,702
Enebakk/missingen, inlet	13,000	1,873	4,694
Enebakk/missingen, inlet	14,000	1,000	1,694
Enebakk/missingen, inlet	13,000	1,608	2,866
Enebakk/missingen, inlet	8,000	0,764	1,478