Grant Agreement: 644715

AQUASMART

"Aquaculture Smart and Open Data Analytics as a Service"

Deliverable D2.3

Data Analytics in Aquaculture

Work package:	WP2 – Open Cloud Data Platform
Prepared By/Enquiries To:	João Pita Costa (jsi@aquasmartdata.eu) – JSI Matjaž Rihtar (jsi@aquasmartdata.eu) – JSI Ioannis Zarifis (i2s@aquasmartdata.eu) – I2S Gerasimos Antzoulatos (i2s@aquasmartdata.eu) – I2S John McLaughlin (tssg@aquasmartdata.eu) – I2S Derek O'Keefe (tssg@aquasmartdata.eu) – I2S
Reviewers:	John McLaughlin (TSSG), Ruben Costa (Uninova)
Status:	Final
Date:	29/01/2016
Version:	1.0
Classification:	Public

Authorised by:	**Reviewed by:**	**Reviewed by:**
Steven Davy		Ruben Costa
Steven Davy WIT	John McLaughlin WIT	Ruben Costa Uninova

Authorised date: 02/02/2016

Disclaimer:

This document reflects only authors’ views. Every effort is made to ensure that all statements and information contained herein are accurate. However, the Partners accept no liability for any error or omission in the same. EC is not liable for any use that may be done of the information contained therein.

Aquasmart Project Profile

Contract No.: H2020-ICT-644715

Acronym:	Aquasmart
Title:	Aquaculture Smart and Open Data Analytics as a Service
URL:	www.AquaSmartdata.eu .org .com
Twitter	@AquaSmartData
LinkedIn Group	AquaSmartData
Facebook Page	www.facebook.com/AquaSmartdata
Start Date:	02/02/2015
Duration:	24 months

Partners

WATERFORD INSTITUTE OF TECHNOLOGY (TSSG) COORDINATOR	IRELAND
INTERGRATED INFORMATION SYSTEMS (I2S)	GREECE
UNINOVA - INSTITUTO DE DESENVOLVIMENTO DE NOVAS	PORTUGAL
GRAMMOS S.A. (GRAMMOS)	GREECE
ARDAG COOPERATIVE AGRICULTURAL SOCIETY LTD (ARDAG)	ISRAEL
NIORDSEAS SL (ANDROMEDA)	SPAIN
Q-VALIDUS LIMITED (Q-VALIDUS)	IRELAND
INSTITUT JOZEF STEFAN (JSI)	SLOVENIA

ICT-15-2014: Big data and Open Data Innovation and take-up
H2020-ICT-2014-1

PROJECT PARTNER CONTACT INFORMATION

WATERFORD INSTITUTE OF TECHNOLOGY ArcLabs Research & Innovation Building, WIT, West Campus, Carriganore, Co. Waterford, Rep. of Ireland. T: +353 51 302920 E: info@tssg.org http://www.tssg.org	**INTEGRATED INFORMATION SYSTEMS SA** Mitropoleos 43, Metropolis Centre, 15122 Marousi, Athens, Greece. T: +30 210 8063287 E: i2s@i2s.gr http://www.aqua-manager.com	**INSTITUTO DE DESENVOLVIMENTO DE NOVAS TECNOLOGIAS** Quinta da Torre, 2829-516 Caparica, Portugal. T: +351 212948527 E: rg@uninova.pt / jfss@uninova.pt http://www.cts.uninova.pt/group_C2_objetives	**GRAMMOS** Ag. Apostolon & Pargas 2, 46100, Igoumenitsa, Greece. T: +30 26650 29231 E: info@grammos-sa.gr http://www.grammos-sa.gr/
ARDAG Ashdod Farm: Ashdod Port, Israel. Main Office: North Shore, Po.B. 1742, 88116, Israel. T: +972-8-6303200 E: nir@ardag.co.il ido@ardag.co.il http://www.ardag.co.il	**ANDROMEDA IBERICA ANDROMEDA GROUP** C/ Zinc s/n, Parque Empresarial Carabona. Spain. T: +34 964 587 068 E: AquaSmart @andromedagroup.es http://www.andromedagroup.es	**Q-VALIDUS LIMITED** NexusUCD, Blocks 9 & 10, Belfield Office Park , University College Dublin, Belfield. Dublin 4, Ireland. T: +353 1 716 5428 E: info@q-validus.com http://www.q-validus.com	**JOŽEF STEFAN INSTITUTE** Jamova cesta 39, SI-1000 Ljubljana, Slovenija. T: +386 1 477 33 77 E: dunja.mladenic@ijs.si http://www.ailab.ijs.si

Table 1.1: Partner Contact Information

Aquasmart is co-funded by the European Commission - Agreement Number 644715 (H2020 Programme)

Document Control

This deliverable is the responsibility of the Work Package Leader. It is subject to internal review and formal authorisation procedures in line with ISO 9001 international quality standard procedures.

Version	Date	Author(s)	Change Details
0.1	22/10/15	Matjaž Rihtar	Table of Content.
0.2	4/11/15	Giannis Zarifis	Draft of data analytics.
0.3	25/11/15	João Pita Costa	Initial draft for review.
0.4	22/12/15	Matjaž Rihtar	Development of FCR models
0.5	28/12/15	João Pita Costa	State of the art in aquaculture.
0.6	12/1/16	João Pita Costa	Neural networks approach to aquaculture.
0.7	13/1/16	Gerasimos Antzoulatos	Draft of the presentation of results.
0.8	28/1/16	Derek O'Keefe, Gerasimos Antzoulatos	Draft of data visualisation.
0.9	28/1/16	João Pita Costa	Final draft of the deliverable.
0.9.6	29/1/16	Matjaž Rihtar	Final formatting.
1.0	2/02/16		Approved version release.

Executive Summary

Objectives

Following the increasing adoption of Big Data analytics in aquaculture, it is now the time to bring the techniques of aquaculture to a new level of development and understanding. However, the specific challenges of the business questions in aquaculture highlight specific needs and problems that must be appropriately addressed. Considerations must be given to the state-of-the-art methods of statistics and data mining that permit deeper insights into the aquaculture reality through the collected datasets, either from daily data or from sampling to sampling data. This must also be tuned to the expert knowledge of the fish farmers, their procedures and technology in use today. Moreover, in this deliverable we further address the Aquasmart data visualisation tools that have been developed and will contribute to the success of the end user businesses.

Results

In this deliverable, we review the state-of-the-art of Big Data analytics methodology in aquaculture, the data available deriving from the procedures characteristic to this business, and propose mathematical models that permit a deeper insight on the data. This will enable the user to discover valuable information from the data that can be made usable, relevant and very actionable.

Table of Contents

TABLE OF FIGURES

TABLES

1 INTRODUCTION

This deliverable presents the Aquasmart approach to data analytics. It contributes directly to the Task 2.5 - Machine Learning Component - by achieving the objective to "develop and incorporate machine learning and classification for open data to enable prediction based analytics" in Work Package 2. This deliverable also takes into account the technical architectural requirements defined in the deliverable D2.2, and provides input to the KPIs defined in the deliverable D2.4.

In particular, this deliverable is aimed at the application of data mining methodology to aquaculture. The deliverable builds on from both the Grant Agreement and Consortium Agreement procedures and it defines the required data collection, data analytics and data visualisation. In that we will use the Cross Industry Standard Process for Data Mining (CRISP-DM): a comprehensive process model - independent of both the industry sector and the technology used - for carrying out data mining projects [She00].

In Section 3 we discuss the input data collected from the fish farms, assigning different importance factors to its features according to the business questions and KPIs discussed with the industrial partners. We discuss the characteristics and the quality of the data collected daily and within the sampling procedure. Moreover, we also discuss the pre-processing of that data, including the cleaning procedure and the missing data problem.

In Section 4 we address the methodology of the data analytics in this project. We consider the Big Data approach in aquaculture today and the specific challenges in aquaculture. We then review the state-of-the-art of data analytics methods in aquaculture, including the analysis of the mathematical modelling and prediction, the analysis of data streams from sensors, the computer vision for automatic feeding control, the forecasting using artificial neural networks, and the case-based reasoning for fish disease diagnosis. Finally, we discuss the Big Data analytics methodology and technology already available, and it's potential and challenges in aquaculture.

In Section 5 we describe in more detail the mathematical models and algorithms used within Aquasmart to take profit of the data mining potential in aquaculture. We start by discussing the KPIs and available datasets in the light of the business questions considered. This also includes the formulation of some of the important KPIs to the data analytics component. We then discuss the usage of generalised linear regression to determine impact factors in aquaculture production, and consider the application of the stream story software to the specific needs of Aquasmart. We further discuss model selection with an example of a model to count the fish mortality and losses for other reasons. Moreover, we discuss the modelling of the FCR tables and the SFR calculations, and finally we consider artificial neural networks to forecast time series data in aquaculture.

In Section 6 we present the analysis of Aquasmart results from the data analytics procedures. We start by describing how we will use summary statistics to obtain structured information within the analysis of results. Then we present the potential of exploratory data analysis to provide the representation of those results maximising the insight into the information they provide. Finally, we discuss the procedure for the evaluation of the results based on a Gaussian approach.

In Section 7 we show how the data visualisation is provided to the end-user, describing the filtering associated with it, presenting some visualisation tools and methods that will be used.

Constructive comments for the improvement of the Project Plan and the method for managing the project are always welcomed. Please contact the Project Coordinator initially by email:

Steven Davy, Research Unit Manager 3MT.
Waterford Institute of Technology (WIT)
Telecommunications Software & Systems Group (TSSG).
ArcLabs Research & Innovation Building,
WIT, West Campus, Carriganore, Co. Waterford,
Ireland.
E: sdavy@tssg.org
W: http://www.tssg.org

2 ABBREVIATIONS AND ACRONYMS

2.1 Abbreviations and short description

Abbreviation	Description
BWG	Body Weight Gain
LTD	Life to Date
BWG	Body Weight Gain
CV	Coefficient of Variation
FCR	Feed Conversion Rate
SFR	Suggested Feeding Rate
SGR	Specific Growth Rate
GPD	Growth Per Day
ADG	(Average) Daily weight Gain
MR	Mortality Rate
FR	Feed Rate
SR	Survival Rate
KPI	Key Performance Indicator

Table 2.1: Table of abbreviations and acronyms

2.2 Long description and formulation

A long description of some of the acronyms used in representing KPIs is presented, discussed and carefully formalized in Section 5.1.1.

3 DATA

3.1 Input data features

In the following section, we describe some of the features of the data including an assessment on its quality and measures to overcome obstacles to the analysis.

3.1.1 Types of Data

As discussed in the deliverable D 2.2, the input and output variables of the dataset include: numerical and categorical. Numerical variables can be:

- Continuous – measured quantities expressed as a float (e.g. 'av. weight');
- Discrete – count expressed as an integer (e.g. 'number of fish');

while categorical variables can be:

- Regular categorical – data including non-ordered classes (e.g. species Bream/Bass);
- Ordinal – classes that can be ordered in levels (e.g. estimations poor/fair/good).

From the variables that can be measured it is important to distinguish between:

- Variables that do not change over time, often identifying population attributes (e.g. identifications such as 'year' or 'hatchery');
- Variables that can change over time but do not change within a sampling period (e.g. 'batch');
- Variables that change daily, taken into account when samplings occur (e.g. 'average weight')

In the following table, we identify the variables changing between samplings in blue, and changing through time but not between samplings in green.

VARIABLES	TYPE
Fish No	Integer
Av. Wt.	float
Biomass	Float
Model Feed	Float
Actual Feed	Float
Temperature	integer
FR	Float

Harvest (No)	Integer
Harvest (Kg)	Float
Mortality (No)	Integer
Mortality (Kg)	Float
Adj. (No)	Integer
Adj. (Kg)	Float
Culling (No)	Integer
Culling (Kg)	Float
Transfer No (-)	Integer
Transfer No (+)	Integer
Sampling Av. Wt.	Float
Fasting	Integer

Table 3.1: Type and nature of the input data

3.1.2 Input data classification

All the data within this project is collected by hand by the aquaculture professionals and handled in a CSV or an Excel file to be pre-processed. In the exploitation of this project, when opening the platform to other fish farmers other than the Aquasmart partners, we can expect the input of sensor data, much as discussed in Section 4.2.2.

Essentially we have three types of input data according to the impact they assure:

1. **Identification data.** This is the data that permits the fish farmer to manage the production and correctly identify the fish;
2. **Daily data.** This is the data that is provided by the fish farmers resulting from their everyday data input (e.g. ‘date’, ‘av. wt.’, ‘actual feed’, etc.);
3. **Sampling data.** At predetermined points of the fish growth timeline, a sample of the fish is done to confirm the model values and make the appropriate adjustments (this will be further discussed in Section 3.2);
4. **Life To Date** [LTD]. This is cumulative data that is calculated from the time when the fish enters the net as a fry to the date of data collection, and will last until the date of the harvest.

The identification data in input 1 is rather unspecific, as we cannot at this date in time identify the fish one by one as it is done in other animal farming such as cows and pigs. The data in this input category is composed of two codes:

- **Unit.** The group of production indicating localization (also known as *cage*).

- **Batch.** The individual production series of fish.

There is no further distinction in the identification. Batch has to go with Unit. Aquafarmers may have different batches in one unit or fish from one batch in many units. There are also cases where the aquafarmers move units from one site to another (rare but it has happened), so we cannot eliminate site. Region is not necessary.

The daily data, the sampling data and the LTD data in inputs 2, 3 and 4 fall into three categories:

- Direct values – values that correspond to the direct observation of the aquafarmers on either variables values including small errors measured in the field (e.g. sampling measures such as average weight) or precise values provided by external sources (e.g. water temperature or oxygen level);
- Calculated values – values that are dependent of a number of other observed values (e.g. LTD values calculated from the daily data);
- Derived values – values deriving from previously available calculation tables (e.g. FCR calculated from the table, given average weight and water temperature).

The daily data in (input 2) is recorded by the aquafarmers on a daily basis. Each of these data instances is recorded as time-series data in a column of the input excel file. These data columns follow the development of the fish since day one when it enters as a fry. The data inputted mostly follows one batch of fish from the beginning till the end of the production. One input data can have several units but, for purposes of the algorithms used, we consider only the time spent in one unit. For some of the algorithms used, the data is split this way (some data tables don't have values in the column 'harvest') with clear input/output within one unit. This is also checked during the data cleaning as described in Section 3.3.

The sampling data in (input 3) serves the aquafarmers to improve/tune their initial FCR model with the real data. That raw data is being collected and provides the potential to adjust the measurements to the reality of the aquaculture production. It includes features that can be learned by a specific set of data. Those features will later be important for the algorithms as discussed in Section 5.2.2. They often correspond to columns with potential effect on the end result. Also, they can influence the production (e.g. 'feeder'). The software produced within Aquasmart will adapt to the data and will try to do the analysis and prediction from the available data. Note: that the input will also include data columns unknown to the system and optional to the aquafarmer (e.g. these can include pH level that is important in closed systems aquaculture [Sti07]). We cannot predict the relevance of the data on those columns (neither their nature) but will consider them in the overall global analytics.

3.1.3 Accuracy of the data

There is the potential for error from human observation that can be expressed in the data and consequently in the output on the fish production optimization. Sometimes that error can be measured and controlled by imposing limits to that same error. That brings us to three categories of data:

- **Inaccurate data**. Data that can be measured and evaluated but may contain errors (e.g. 'av. weight');
- **Accurate data**. Data that can be measured and evaluated but cannot contain errors (e.g. 'food');
- **Unclassified data**. Data that cannot be measured, quantified or evaluated (e.g. 'feed').

Note that the method of counting and averaging with margin of error is acceptable for the aquafarmers in general. Clearly, the identification data such as 'unit' or 'batch' are not considered in the analysis of the exactness of the data.

In the input data, each fish farm will be handled separately in order to avoid mixing with the data from other farms. Only some of the common features will be used for global models with appropriate data privacy measures.

3.2 Sampling methods and data

The sampling at the fish farms is a common procedure that permits an evaluation of the production process. It consists of the collection of a sample of the fish in a certain unit during a period of time. The output of this process is expressed in 'adjustments' where the number of fish is adjusted. Furthermore, the average weight is corrected providing a more precise assessment to FCR and SFR. In particular, the aquaculture sample is perceived to gain the following significant benefits:

- costs savings in frame updating and maintenance;
- aquaculture indices (i.e. indices for sustainable aquaculture: e.g. biological, environmental, etc. [PFP07]) can be selected more quickly and economically;
- improvement of the reliability of aquaculture indices from different aspects;
- possibility of integration and substantive understanding of the global production.

The classic sampling method is the following: the fish are crowded by a swipe net and then a selection of fish (the number depends on the fish size) are taken out, either for anesthetizing in a specific bowl to avoid the measurement stress, or straight to the sampling bucket. In the bucket fish are sampled in groups of 2-4 kg, then the fish are counted back to the unit and the total weight divided by the fish number to create the average fish weight of the cage. The average weight is entered into the system only after the farmer decides whether the number is good or bad, according to previous samplings and according to how the fish appear in the unit. Sometimes there is adjustment of the result before being entered into the system.

Figure 3.1: Classical sampling methods in aquaculture

The sample data includes features like weight, water temperature and oxygen, among others. Though, features like biomass are calculated, not sampled. The number of fish is a good example for the need of sampling as it is often a number we cannot measure exactly. This is because of the way the fish farms do the sampling, the fact that the 'live' material is sensitive to handling, and often the lack of appropriate tools. Even in fry delivery there is an tolerated inexactness of up to 10% (corresponding to the additional fry received when ordered) due to the potential death of some fries during the trip to the fish farm.

Some companies don't carry out sampling, mostly because the stress the process induces in the fish. Sampling is done approximately on a monthly basis. It provides inside knowledge permitting adjustments to the input of data guided by models. These measurements are often intrusive to the livestock. An example is the sample of the average weight, which is typically done by weighting a full bucket of fish and counting the fish. In this case, the fish are counted to exact number, but the weight, if the sampling is done on the water, can be approximate.

3.3 Data collection cleaning and fusion methods

The aim of this module is to create an intelligent automation for the ETL stage – Extract, Transform and Load. The original data received from the fish farmers must be cleaned and processed before analytic computations. At present, the data is received from the farmers as an excel file, which is then converted to CSV and loaded into the Aquasmart system (in time a user interface will be provided to upload the data). All special characters are converted to standard ANSI characters. After that, it goes through a first cleaning process correcting obvious errors and a second cleaning process correcting less obvious errors against the formulas and known variable dependencies. Recall that the Aquasmart processes don't work locally but rather they work via message bus through query to the big data storage machinery. We will now describe in detail the steps of the data cleaning process and related problems.

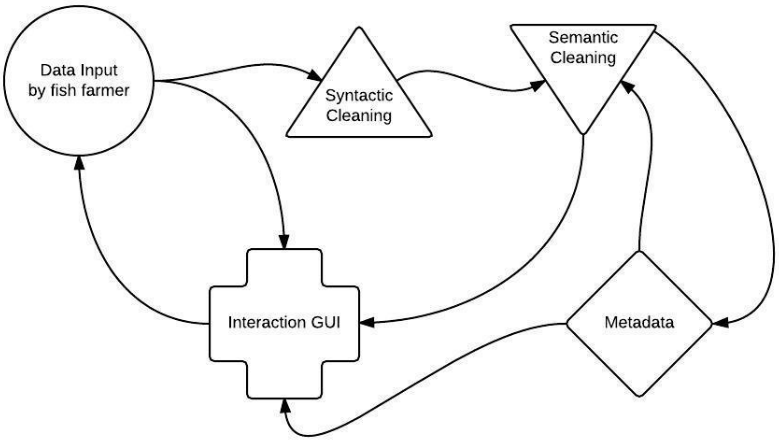


Figure 3.2: The Aquasmart data cleaning process

The original data is entered mostly following one batch of fish from the beginning when the fish are still fry, until the end when the harvest is complete. In between there may be adjustments and transfers corresponding to the splitting of units. As that procedure requires significant resources in terms of manpower and time, not all farms carry out splitting.

The two main typical problems with the data are (1) data availability and (2) data quality. Data availability, refers to the percentage of data that we were expecting vs. the actual data that we have in the datasets. On the other hand, data quality refers to the percentage of good data vs damaged data. In that, the problems of data cleaning that can be found in the original data are the following:

- The data can include characters/words in the local alphabet which need to be converted to standard ANSI;
- Problems can occur when converting Excel files to CSV files (CSV is a preferred file format for data analytics);
- Files can be corrupted during the transfer or input.

After having the data harmonised with all numbers converted to float variables and all text in standard ANSI, we can proceed with the verification and correction of the data.

The obvious errors identified in the data provided by a preliminary *syntactic* analysis of that data are of three types:

Err 1. Inconsistencies of values according to the agreed formulas;

Err 2. Obvious outliers (e.g. FCR with a value of 20000);

Err 3. Wrong type of data in a column (e.g. data for ‘av. wt.’ placed in the column ‘date’)

The harmonisation of the data is then essentially done in three steps:

Har 1. Ignore the outliers that are obviously not valid (this is done considering predefined limitations of the data;

Har 2. Identify the variables important for the Aquasmart data analytics (e.g. ‘average weight’ is primarily important but the name in the variable ‘Feed’ is not) to which we call *analytic variables*;

Har 3. Correct dependent values, taking profit of the fact that some of the columns/variables derive from other variables.

Experienced user will know the importance of the features and their context, limiting to what is important and proceeding appropriately with the outliers as done automatically in (Har 1). The information collected in this process will be stored in the data store in the form of metadata.

For point (Har 2) the identification of (in)dependent variables is the key to reduce computational effort and increase the level of correction of the calculations.

Automatic checks with the available formulas sometimes do not correspond to the numbers entered by the aquafarmers in the tables. This makes step (Har 3) essential to calibrate our Aquasmart system. Notice that this checking will not be done in the optional columns. In the following, we list the independent analytic variables and their dependencies. The choice of which variables to consider as independent variables is internal to the Aquasmart approach and could change by simply changing a combination of the dependent variables (sometimes a single one) with the independent one.

INDEPENDENT VARIABLES	DEPENDENCIES
Fish No	FCR, Mortality (No), Transfer No (-), Transfer No (+)
Av. Wt.	LTD Econ. FCR, LTD Biol. FCR, Sampling Av. Wt.
Biomass	Econ. FCR, Biol. FCR
Model Feed	Biomass, Temperature
Actual Feed	FCR, SFR, Model Feed, Fasting
Temperature	LTD Econ. FCR, LTD Biol. FCR
FR	Actual Feed, Average Weight, Water Temperature
Harvest (No)	LTD Harvest No, LTD Harvest %
Harvest (Kg)	LTD Harvest Kg,
Mortality (No)	Mortality %, LTD Mortality No, LTD Mortality %
Mortality (Kg)	LTD Mortality Kg
Adj. (No)	LTD Adj. No, LTD Adj. %
Adj. (Kg)	LTD Adj. Kg
Culling (No)	LTD Culling No
Culling (Kg)	LTD Culling Kg
Transfer No (-)	Fish No
Transfer No (+)	Fish No
Sampling Av. Wt.	Av. Wt.
Fasting	Actual Feed

Figure 3.3: The dependencies within the KPIs of Aquasmart

3.3.1 Semantic data cleaning

After the above syntactic analysis of the data, we perform a *semantic* analysis of the data corresponding to an automatism to understand each column and determine outliers that agree with the calculations and slip through the syntactic corrections. This is built on a knowledge-based decision that includes the metadata in the data store and the eventual access to expert knowledge.

In this process there are several hidden variables that can be predicted but not measured exactly. An experienced aquafarmer for example, will know that the reason the economic FCR grows in cold temperatures with the growing average weight is due to the energy spent in fish reproduction.

3.3.2 The missing data problem

The original data provided by the aquafarmers has variances/holes and is not precise because it is not measured automatically but rather entered by manually (with some exceptions such as 'temperature'). Sometimes it is not entered for 1 or 2 days due to the bad weather, which complicates the access to the measurements and to the units themselves (sometimes this adds up to 4 days without entries). Sometimes this is due to intentional fasting to readjust features and in that case the data measurements stay the same as the ones in the previous fields, just before fasting takes place. The major discrepancies should be pushed to the user as a compromise. If the data is missing up to a certain threshold, the data will be sent back to the user in order to be re-entered after appropriate corrections.

The data will be reported to the Aquasmart system in CSV format with a periodicity according to the type of data that corresponds to (much as discussed in Section 3.1.2.). The daily data will be inserted daily, with the exception of days of impossible access to the units such as days of storm, etc. The sample data will be inserted with a periodicity that corresponds to the sampling procedure adopted by the fish farm. There is also occasionally inserted data referring to:

- Fasting – the fasting of the fish prior sampling or other aquaculture activity;
- Harvesting – the collection of the fish ending the production cycle;
- Transfer – the transfer of grown fish to other units according to their average weight;
- Fish disease – the information of the disease outbreak in a unit;
- Fish death or disappeared – the information of the dead/disappeared fish in a unit.

The user interface will make a conformance checking upon data uploading and provide feedback to the user about the submission status and quality of the data set. The options for the missing data problem are to consider it as an error and report it to the user requesting the missing data, or to interpolate the missing data on a fixed mesh grid. This will be further discussed in Section 5.2.2.

4 METHODOLOGY

4.1 Unique challenges of Aquaculture

In this section, we review the current impact of Big Data on general agriculture as a motivation for what follows in future overall Big Data approach to aquaculture. We also present the state-of-the-art in data analytics within aquaculture, taking into account the technology and infrastructure in use at present. We then put all that in context with the Aquasmart approach.

4.1.1 Big Data in agriculture

Big Data in agriculture is focused towards unlocking the economic potential of improved management decisions. The benefits of precision in the agriculture production pipeline provide a relevant improvement in the optimisation of resources. This includes the usage of geospatial maps on imagery and fertility, of publicly available data sources for weather or soil, and appropriate analytics on crop models, data co-ops and markets. While it is true that agriculture has been traditionally used some of that raw Big Data before, the availability of data analytic tools to visualise and extract full value of real time data greatly enhances the possibilities.

The influence of the transformations made by precision agriculture (also known as site specific crop management) is starting to have impact on aquaculture and on the seafood sector in general. This is boosted by the recent developments on sensors, robotics, computer vision, satellite imaging and Big Data analytics. Nowadays, the available technology enables the usage of drones and driverless boats, or interconnected devices powered by advanced data mining in the context of the Internet of Things adapted to the needs of the sector.

It has been announced that aquaculture is the next target for the advances boosted by Big Data technology [ECAq], much as what happened to agriculture, as discussed previously. At a fundamental level, agriculture and aquaculture share several common objectives (e.g. optimization of production costs). The manipulation of available data to those common features can often profit from the adaptation of already available technology towards aquaculture. Examples of that technology are the forecast of specific weather features such as the prediction of storms.

4.1.2 Uncertainties and challenges in aquaculture

It is well known that the production in aquaculture has specific features and objectives associated with it. When talking about the adaptation of existing technology, the features important to the production in aquaculture come from weather prediction. These are the oxygen levels and water temperature, which are very specific to this activity.

The tasks in fish farming carry several uncertainties – often expressed by measurements or even evaluations – that permit further optimization. A classic example is the aim for a better control on the food loss and food quality. A contribution of data mining in this context would be of interest to the aquafarming industry, saving or relocating resources.

An important variable that remains undetermined during the complete production pipeline is the exact number of fish. A margin of up to 10% of number of fries is added to the initial production at time t=0 due to uncertainty of number of deaths in the transport. That means that we already have a maximum of 10% more fish than our estimations (assuming that no fries die during transport or adaptation at t=0). Other than that we can only have less fish than we estimated due to the lost fish because of unknown reasons. This is already an open problem at the level of the bounds for total amount of harvested fish and the description of best-case scenario and worst-case scenario. This represents a big lack of knowledge about production. In fact, the unknown number of the fish until the end of the production is important for the amount of food given and, consequently, for the resources spent.

The SFR table allows the fish farmer to assess the amount of food to give to the fish according to their average weight and the temperature of the water. Each farm has its own SFR table. This is an opportunity to create our own table/model by tweaking the numbers accordingly. Also specifying the influence of sexual maturity and the lack of oxygen, which are done by hand/intuition, have features to take in consideration by the math model. The FCR and SFR models currently consider only temperature and average weight. This will be further discussed in Section 5.2.2.

Modern research and commercial aquaculture operations have begun to adopt new technologies, including computer control systems. Aquafarmers realize that by controlling the environmental conditions and system inputs (e.g. water, oxygen, temperature, feed rate and stocking density), physiological rates of cultured species and final process outputs (e.g. ammonia, pH and growth) can be regulated. These are exactly the kinds of practical measurements that will allow commercial aquaculture facilities to optimize their efficiency by reducing labour and utility costs. Anticipated benefits for aquaculture process control and artificial intelligence systems are:

1. increased process efficiency;
2. reduced energy and water losses;

3. reduced labour costs;
4. reduced stress and disease;
5. improved accounting;
6. improved understanding of the process.

[Lee00] reviews the technologies and implementation of the technologies necessary for the development of computer intelligent management systems for enhanced commercial aquaculture production. Today's artificial intelligence (AI) systems (i.e. expert systems and neural networks) offer the aqua culturist a proven methodology for implementing management systems that are both intuitive and inferential. There have been many successful commercial applications of AI (e.g. expert systems in cameras and automobiles). The major factors to consider in the design and purchase of process control and artificial intelligence software are functionality/intuitiveness, compatibility, flexibility, upgrade path, hardware requirements and cost. Of these, intuitiveness and compatibility are the most important. The software must be intuitive to the user or they will not use the system. Regarding compatibility, the manufacturer should be congruent with open architecture designs so that the chosen software is interchangeable with other software products.

4.2 State-of-the-art

The use of machine learning tools in aquaculture is itself not new, but it has not been explored in its full potential. In the following paragraphs we present an overview of the advances of aquaculture analytics at this level, including what was developed in the past and what is available today.

4.2.1 Modelling and prediction in aquaculture

Mathematical modelling aims to describe the different aspects of the real world, their interaction, and their dynamics through mathematics. It constitutes the third pillar of science and engineering, achieving the fulfilment of the two more traditional disciplines, which are theoretical analysis and experimentation. Nowadays, mathematical modelling has a key role also in aquaculture. In the following section, we present an overview of that. Growth and reproductive modelling of wild and captive species is essential to understand how much of food resources an organism must consume, and how changes to the resources in an ecosystem alter the population sizes.

The study of growth means basically the determination of the body size as a function of age. Therefore, all stock assessment methods work essentially with the age composition data. This has been an important topic in the aquaculture research and development. Several numerical methods have been developed, which allow the conversion of length-frequency data into age composition. Already in 1920, Pütter developed a growth model which can be considered the base for most other models off growth including the one developed as a mathematical model for individual growth by von Bertalanffy in 1934, and which has been shown to conform to the observed growth of most fish species. Later on, the von Bertalanffy growth model of body length as a function of age has become one of the cornerstones in fish biology because it is used as a sub-model in more complex models describing the dynamics of fish populations. In the following we present the von Bertalanffy growth equation (on the left) and a family of growth curves with different curvature parameters, different K values (on the right). This model will be further discussed in Section 5.2.3.

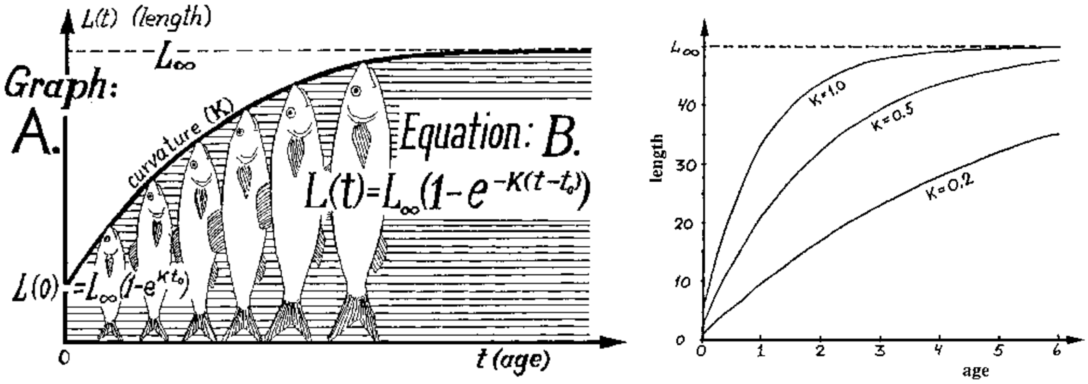


Figure 4.1: The von Bertalanffy growth equation (on the left) and a family of growth curves with different curvature parameters, different K values (on the right)

Already in [HH89] the authors develop a model for the growth of the eel in aquaculture. It was designed to take the growth rate variability into account by classifying the population, step by step in time, into size boxes (classifying the fish in size intervals) on which grading and grouping is considered. In that model it is allowed the fit to other growth conditions and other species by selecting options. Furthermore, that growth model was a predictive model for aquaculture, constituted by adding algometric functions related to physiological features. The methodology for length-age data of several species of fish can consider models fitted to each dataset other than the von Bertalanffy growth model (VBGM). These are: generalized VBGM, Gompertz growth model, Schnute-Richards growth model, and Logistic [Kat06].

A statistical model has been considered in [LASHGQ11] to forecast fish growth, with uncertainty, providing good predictions of future biomass of Norwegian farmed salmon. The model is based on the number of fish in each weight class and their average weight. The model, which is related to standard size-structured models, computes the number of fish growing into the next weight class the next month and the number of fish remaining in the same weight class. In addition, the number of new fish stocked; fish lost, slaughtered and wasted, as well as the sea temperature related to the growth, were modelled. [RS96] describes optimal management in a single-farm model for sea bass, based on weight classes and biological sub-models. Any animal production involving cycles may benefit from this tool, since it takes quite some time between the time a producer decides to expand the production, and the time those animals have reached their slaughter weight.

Accurate characterization of temperature and dissolved oxygen stratification in units used for aquaculture is of critical importance in understanding how these units may be constructed, oriented, or otherwise managed biophysically when one wishes to provide optimal environmental conditions for the organisms being cultured. While field studies can provide characterizations of water quality stratification at a single locale, to date there have been few attempts at developing reliable models which can be used at a variety of sites after initialization with appropriate local geographic and atmospheric data. Advances in model structure and reduction of data requirements relative to previous models reflect the desire to provide for culturists the ability to predict stratification events with commonly available data, obtained either by hand or from a simple weather station located at or near the unit site. A series of simulation runs was performed in [CP96] to assess the quantitative effects on temperature and dissolved oxygen concentration generated by varying pond depth and phytoplankton density input values.

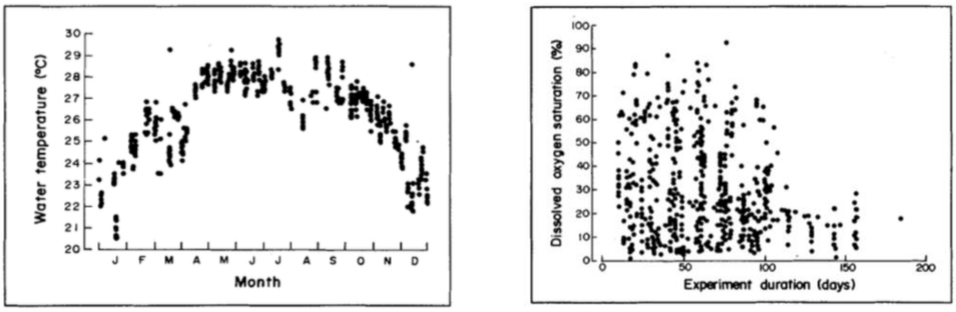


Figure 4.2: Seasonal variation of daily early morning water temperature (on the left) and distribution of early morning dissolved oxygen (on the right)

Feed composition has a large impact on the growth of animals, particularly marine fish. In [BR09], the authors developed a quantitative dynamic model to predict the growth and body composition of marine fish for a given feed composition over a timespan of several months. The model takes into consideration the effects of environmental factors, particularly temperature, on growth, and it incorporates detailed kinetics describing the main metabolic processes (protein, lipid, and central metabolism) known to play major roles in growth and body composition. That showed that multiscale models in biology can yield reasonable and useful results. The model predictions are reliable over several timescales and in the presence of strong temperature fluctuations, which are crucial factors for modelling marine organism growth. The model provides important improvements over existing models.

Advanced artificial intelligence techniques such as support vector machines are used in environmental modelling in [DRG06] considering the multidimensional nature of the study cases. The use of Bayesian hierarchical models [GCSR04] within a state-space framework [DK01] allows for separate modelling of the system process and the data process. The system process is essentially the theoretical model for the standing stock and the sub-models. The data process describes the observable quantities, measurement errors and how the data are connected to the system process. Still, how to properly balance the numbers and weight of fish, as well as the short and long-term performance of the model, continues to be an open problem.

4.2.2 Environmental time-series analysis from sensors

Today, the available sensor technology offers real-time environmental monitoring system for aquaculture in a wide range of areas. The most common systems can monitor current speed, current direction, water temperatures and oxygen inside and outside the unit. More advanced systems also offer control aerators, pumps, alarms and communication devices. These systems can be configured to serve the smallest to the largest aquaculture enterprises. Real-time data to the feeding system can optimize the feeding process thereby reducing not only cost, but also the environmental impact. The information is used by aquaculture farms to plan and adjust feeding and thereby avoid feed spillage. By monitoring the conditions our customers operate more efficiently and save money on feed. The systems have low maintenance cost since the sensors will be stable over very long periods without any recalibration.

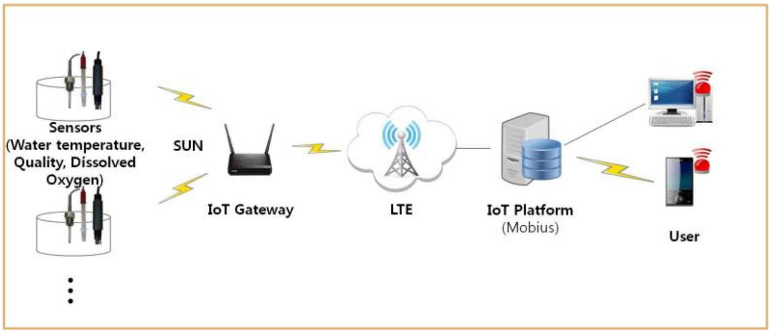


Figure 4.3: A variety of sensors in aquaculture

In general, a time series is any function varying with time, including all sequences of daily measurements and sampling measurements in aquaculture. Time series data often arise when monitoring aquaculture processes (such as average weight or water temperature) or tracking their metrics. The essential difference between modelling data via time series methods or using the process monitoring methods is that time series analysis accounts for the fact that data points taken over time may have an internal structure (such as autocorrelation, trend or seasonal variation) that should be accounted for. This section provides a brief overview of some of the more widely used techniques in the rich and rapidly growing field of time series modelling and analysis in general, and in the context of aquaculture. By combining sensor data with historical cases in [SC14], the authors propose two approaches using time series and machine learning for prediction. In one approach, they consider time series prediction and then use expert rules. In the other approach, they use time series classification for prediction. Both approaches exploit a dynamic data-driven technique where prediction models are updated with the update of new data to predict closure decisions

Sensor networks deployed in the field to monitor aquaculture environments often suffers from failure due to sensor bio-fouling, communication failure, etc. This leads to missing sensor readings that are required by the machine learning based decision-making systems. The authors in [ZRD13] and [RDT13] have designed a multiple classifier based method to deal with missing sensor data. Instead of inputting missing data, the prediction of events is based on available sensor readings. Ensemble approach is shown to perform better than imputation methods. The authors assume equal weights for all features, which may not be true in real world scenarios (see [RM04] and [RM09]). The data produced by the sensors is sometimes faulty. The decisions based on faulty sensor reading will result in wrong conclusion. In [RST14] and [RST13], the authors have presented a novel ensemble classifier approach for assessing the quality of sensor data. The base classifiers are constructed by random under–sampling of the training data where the sampling process is guided by clustering. The inclusion of cluster based under–sampling and multi–classifier learning has shown to improve the accuracy of quality assessment.

The research in [DM11] builds a wireless network sensor system to empower the means for aquaculture water quality monitoring that currently have a weak infrastructure. It includes embedded computing technology, MEMS technology, distributing information processing technology and wireless communication technology. That system is a digital, networked, intelligent real-time dynamic for monitoring the aquaculture water quality. The system not only can deal the normal detection of the aquaculture environment indicators (temperature, PH, dissolved oxygen, turbidity, ammonia, etc.) in real-time monitoring, but it can also detect indicators of data fusion and data mining to establish a history database of aquaculture environmental monitoring indicators. The system can also gather the monitoring data by a local or remote way, and realize the real-time, dynamic display and analysis. So as to improve the process of aquaculture, water resources utilization, the quality of the culture environment and reduce emissions of pollutants. The system can provide an important technical means and scientific basis.

The authors in [PDCJ11] examine the capabilities of seasonal autoregressive integrated moving average (SARIMA) models to fit, forecast, and monitor the landings of data-poor fisheries. Despite the limited sample size, a SARIMA model could be found that adequately fitted and forecasted the time series of meagre landings (12-month fore- casts; mean error: 3.5 tons (t); annual absolute percentage error: 15.4%). Model-based prediction intervals are derived to detect problematic situations in the fishery. Over the course of one year the meagre landings remained within the prediction limits of the model and therefore indicated no need for urgent management intervention.

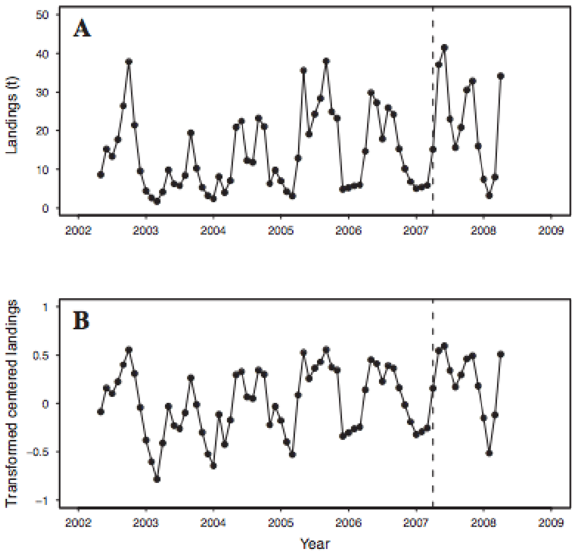


Figure 4.4: The time series analysis in [PDCJ11] for the Portuguese coast. The dashed vertical line is the forecast origin and separates the fitting period from the hold-out period. (A) Raw data. (B) Log10-transformed mean-centered data

The technological advances lead to the dominance of artificial neural networks (ANNs) for classification and forecasting of time series data [BMSSA07], which will be further discussed in Sections 4.2.4 and 5.2.3 in the context of aquaculture.

4.2.3 Computer vision for automatic feeding control

In [ASL15], the authors present an efficient visual signal processing system to continuously control the feeding process of fish in aquaculture tanks. The aim is to improve the production profit in fish farms by controlling the amount of feed at an optimal rate. The automatic feeding control includes two components: 1) a continuous decision on whether the fish are actively consuming feed, and 2) automatic detection of the amount of excess feed floating on the water surface of the tank using a two-stage approach. The amount of feed is initially detected using the correlation filer applied to an optimum local region within the video frame, and then followed by a SVM-based refinement classifier to suppress the falsely detected feed. Having both measures allows the authors to accurately control the feeding process in an automated manner. Experimental results show that their system can accurately and efficiently estimate both measures.

Some work has been done in [Z08] to determine if an automated feeding system can be developed through image analysis of fish feeding behaviour using a submerged surveillance camera. The proposed method is to obtain video images of fish feeding behaviour and subsequently applying data analysis. That data analysis entails image processing, followed by pattern recognition and machine learning methods.

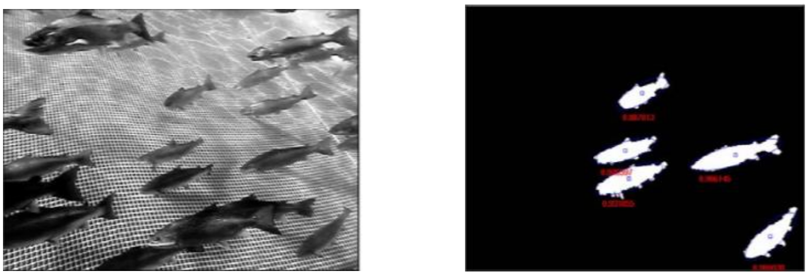


Figure 4.5: Fish detection using neural network approach in [MZZ11]: image after histogram equalization (on the left) and labelled image after segmentation and confidence estimation has been performed (on the right)

There are many common tasks any fish farmer would find difficult to do, such as counting, sorting, measuring and weighing of fish without having to individually handle and stress the fish. These are critical needs in fish farming, as this is essential information for financing, insurance, stock management and feeding activities. Past ongoing research has developed numerous fish feed monitoring, counting and measurement techniques without having to handle or stress the fish. In the past, the implemented techniques included acoustic equipment and signal processing and even the x-raying fish fed with spiked iron powder. Today, the conventional practice of hand-feeding is based on the use of feed tables and the experienced eye of the feeder adjusting the feed quantity to suit the needs of the fish. As units and holding units have become larger and deeper, accurate visual observations of the fish have become more difficult. The information feedback of feed consumption can be improved by implementing methods such as:

- the airlift pump with pellet counters to provide an automatic feeder cut-off, and a facility for recycling the uneaten pellets (see [BPR92], [BJLF93] and [B04]);
- the underwater video camera to observe the stock during feeding (see [PBR85], [KMHT91] and [TH92]).

One of the fundamental challenging problems in computer vision is detecting object inside an image or video frames. The authors in [MZZ11] present a neural network-based approach for detecting fish object inside a digital image. The ANN is used to recognize the fish object from a small window that scans for the object all over the image. The ANN will examine the windows of the image and decides whether each window contains a fish. This approach is aimed to eliminate the difficult task of manually selecting the fishes, which must be chosen and localized to be analysed.

4.2.4 Artificial neural networks to forecast water quality and temperature

Artificial neural networks have long been used for weather forecast [MKA04] and to generate probabilities of precipitation and quantitative precipitation forecasts [HBD99]. Neural network forecasts exceeded other traditional forecasting methods (such as linear or logistic regression systems) boosting controversy surrounding the value of a priori knowledge in determining predictor variables. However, in many cases, incorporating a priori weather knowledge is not feasible because it is very difficult to quantify prior knowledge of weather processes as input to a neural network

[FDL07]. Additionally, research revealed that neural networks outperformed logistic regression, discriminant analysis, and rule-based prediction systems in the classification of tornado events [MS96]. Similarly, ensemble-based neural networks combining the outputs of neural network subcomponents have been shown to outperform each individual neural network subcomponent in the prediction of wind-speed, temperature, and humidity [MKA04].

Artificial neural networks are massively parallel processors that have the ability to learn patterns through a training experience. Because of this feature, they are often well suited for modelling complex and non-linear processes such as those commonly found in the heating system. They have been used to solve complicated practical problems. The simulation results indicate that, the control unit success in keeping water temperature constant at the desired temperature by controlling the hot water flow rate in closed aquaculture systems [AFAD11]. This methodology is also useful in estimating the water temperatures in small river streams [US11].

In [AFAD] neural network control is used for dissolved oxygen of aquaculture pond aeration system. In particular, for controlling the speed of air flow rate from the blower to air piping connected to the pond through control blower speed. This is also the approach in [LYTXL11] that analyses the important factors for predicting dissolved oxygen of Hyriopsis Cumingii ponds, and finally chooses solar radiation (SR), water temperature (WT), wind speed (WS), PH and dissolved oxygen (DO) as six input parameters. Alternatively, the authors in [BGMC13] study the oxygen dissolved in water using machine learning techniques based on Bayesian inference that can be used to enhance the computer simulation of molecular materials, focusing here on water. Moreover, they train their machine-learning algorithm using accurate, correlated quantum chemistry and predict energies and forces in molecular aggregates ranging from clusters to solid and liquid phases.

As a decision system, ANNs are an important tool for forecast in aquaculture to forecast the freshwater fish caught [BB15] and for detecting fish object inside a digital image in the context of computer vision [MZZ11]. The neural network is used to recognize the fish object from a small window that scans for the object all over the image. The neural network will examine the windows of the image and decides whether each window contains a fish.

Remote sensing is another example where ANNs are used to manage the available data, complemented with other artificial intelligence methods and techniques. Local empirical neural network algorithms can be used to evaluate the potential impact of aquaculture, assessed using remote sensing data [BB14]. It is possible to determine the highest particle concentrations and lowest light penetration occurred in the spring and summer. Water circulation patterns can be identified as the major force determining the distribution and hence the source of particles and are also applied to reflect the particle loads introduced by feeding activity performed in aquaculture facilities.

Interest in using artificial neural networks (ANNs) for forecasting has led to a surge in research activities. While ANNs provide a great deal of promise, they also embody much uncertainty. Researchers to date are still not certain about the effect of key factors on forecasting performance of ANNs. ANNs have powerful pattern classification and pattern recognition capabilities, being able to learn from and generalize from experience. As opposed to the traditional model-based methods, ANNs are data-driven self- adaptive methods in that there are few a priori assumptions about the models for problems under study. They learn from examples and capture subtle functional relationships among the data even if the underlying relationships are unknown or hard to describe. Thus, ANNs are well suited for problems whose solutions require knowledge that is difficult to specify but for which there are enough data or observations. Recent studies have shown the classification and prediction power of the Neural Networks. ANNs can approximate any continuous function and have been successfully used for forecasting of data series in aquaculture. While ARIMA assumes that there is a linear relationship between inputs and outputs, ANN have the advantage that can approximate nonlinear functions. An ANN can be used to solve problems involving complex relationships between variables. The problem with the data-driven modelling approach is that the underlying rules are not always evident and observations are often masked by noise. It nevertheless provides a practical and, in some situations, the only feasible way to solve real-world problems.

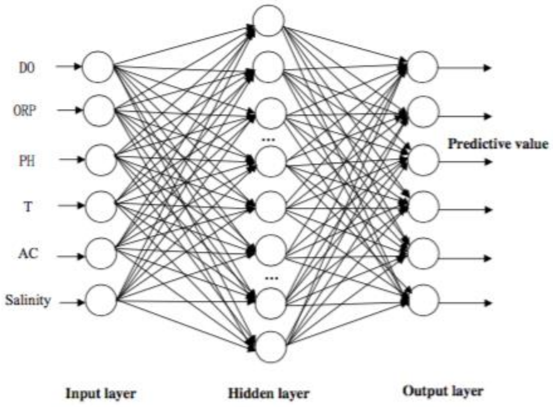


Figure 4.6: The representation of an artificial neural network for water quality monitoring

4.2.5 Fish disease diagnosis and reasoning

Case-Based Reasoning (CBR) is a reasoning method that solves a new problem by getting a reminder to a similar problem solved before, with a range of more specific methods for accomplishing that task. CBR systems perform remarkably well on complex and poorly formalized domains. CBR classifiers use a database of problem solutions to solve new problems. Unlike nearest-neighbour classifiers, which store training tuples as points in Euclidean space, CBR stores the tuples or “cases” for problem solving as complex symbolic descriptions. The decisions of a fish farmer are based on both knowledge and intuition. The knowledge is a combination of sensor data and the understanding of the historical data leading up to the relevant day. The intuition is mainly experience with similar situations and some good old gut feeling.

CBR is intuitive, relatively simple to implement, transparent and it learns. Decision support system developers have problems with the knowledge elicitation bottleneck, the dynamics of decision support, the constant maintenance that systems require, the fact that systems must be accepted and that advice must be justified. CBR addresses each of these problems. A case-based reasoner solves new problems by using or adapting solutions that were used to solve old problems. It also offers a reasoning paradigm that is similar to the way many people routinely solve problems. Cases are several features describing a problem plus an outcome or a solution. Cases can be very rich (text, numbers, symbols, plans, multimedia, etc.), are not distilled knowledge, are records of real events and are excellent for justifying decisions. Unlike CBR, neural nets and genetic algorithms cannot justify their decisions.

The five areas that CBR can contribute to the rest of the AI-community [David B. Leak, 1996]:

Knowledge Acquisition - it is not necessary to do extensive knowledge acquisition while creating a CBR system, as we rely on specific cases instead of an all-including domain model.

Knowledge Maintenance - The CBR system is always updated, and learns from experience. It starts out with a collection of “start-cases” which will be maintained as the system encounters new cases.

Increasing problem-solving efficiency - Reusing of old solutions increases the problem-solving efficiency and there is no need to create solutions to similar problems from scratch each time when you can use an old, maybe adjusted, solution.

Increasing quality of solutions - With bad or non-existing domain models

User acceptance - A user will more likely approve of a solution that once worked on a similar problem given in a case format, than derived rule-chains. Cunningham et al. show that Case-Based Explanation is considered more convincing than the rule-based alternative through a series of tests on faculty staff and students at Trinity College Dublin [Cunningham et al., 2003].

A large amount of work is done in fish disease diagnosis with the help of CBR. To mention some of it, [S97] describes work done on a DSS for hatchery production management for Atlantic salmon in Norway. Moreover, in [BNE00] were developed decision support tools for aquaculture to assess economic and ecologic impacts of alternative decisions on aquaculture production (a system based on simulation models and enterprise budgeting). In particular, [LFD02] describes a web-based expert system for diagnosing fish disease in aquaculture facilities in China using short message service with great success. It considers the usage of the mobile phone instead of the computer systems, widely used in rural areas of China. They implemented a SMS platform in Java with the high accurate diagnosis rate of 93.57% for fresh-water-fish diseases validated by experts. CBR can also be used to capture the information on the sorting, and to also encapsulate the knowledge and intuition of the fish farmer with it, combining sensor data with the knowledge of the fish farmer [GS11]. This system can diagnose 48 kinds of fresh-water-fish of in Tian Jin area; the system has been tested with 140 fish diseases cases that were diagnosed by fish expert. Nine couldn't be diagnosed, the rest diagnosis results agreed with the experts', so the accurate diagnosis rate was as high as 93.57%. It was built upon the existing web-based system and the SMS platform composed of Wave com GSM/GPRS modem, SIM card, central computer and common mobile phone that can send short message, using the nearest-neighbour search model.

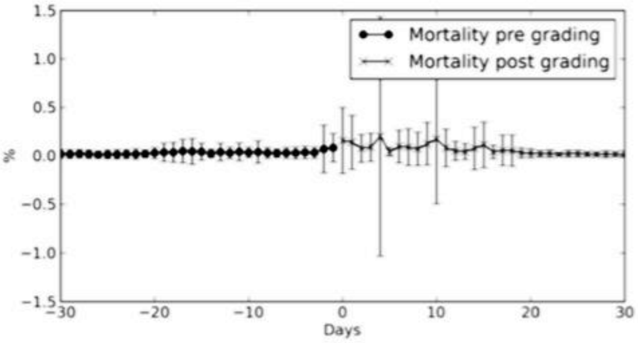


Figure 4.7: Case-based reasoning for lower fish mortality during grading operations in [TBA12]

Furthermore, the main focus of the CBR application developed in [TBA12] is to lower fish mortality during grading operations. The cases in the knowledge base consist of an object (fish), the given symptoms and the treatment given. The similarity assessment between a query and the case-base is done by a clustering algorithm to find which part of the case-base the query belongs to. A simple nearest neighbour algorithm is then used to find the closest matching cases. The evaluation of three different methods for case retrieval: kNN, linear programming for setting feature weights (using the simplex algorithm), and echo state network (neural networks). It considers the *leave one out* cross validation (LOOCV) with only 74 cases. The results show that classification rates using LOOCV of the CBR retrieval mechanisms reveal that the LP approach is the best (82.4%), with the kNN and ESN doing slightly worse (both ∼75%). The similarity prediction error is very low for the LP method, only 1.7%. Thus, 81.1% of the cases belong to one class, while the remaining classes contain 12.1% and 6.8% of the cases, respectively.

4.3 Big Data analytics

In the following paragraphs we will review the methodology and technology in use today in Big Data analytics. In particular, we will emphasize the methodology with potential to be used in aquaculture. In this project we will be using data mining techniques that deal with extracting of information from data and will be using it to predict trends and behaviour patterns. This section will also briefly describe the state-of-the-art of that area of research. Moreover, we shall address here the specific problematic of data analytics in the context of Big Data.

4.3.1 Overview

At the time of writing, 2.5 quintillion bytes of data are created daily—so much that 90% of the data in the world today has been created in the last two years [GR12]. Most of the successful decisions that were made in the world of business were based on the interpretation of available data. Correct analysis of the data is the key success factor in being able to make better decisions that are based on the data. Given the quantity and complexity of the data that is being created, traditional database management tools and data processing applications simply cannot keep up, much less make sense of it all. The challenges for handling big data include capture, storage, search, sharing, transfer, analysis, and visualization. The trend to larger data sets is due to the additional information that can be derived from analysis of a single large set of related data, compared to separate smaller sets with the same total amount of data. Some estimates for the data growth are as high as 50 times by the year 2020 [GR12].

Having a large amount of data available enables us present a wider perspective of the overall aquaculture production. It provides a clearer assessment to the aquafarmers and helps them to make better judgements and more appropriate decisions. In fact, with such a vast amount of data we often do not need sophisticated statistical methods to obtain an insight on the data. We only need simpler models, in particular if we know how to scale the data. The risk with Big Data mining is to discover patterns that occur by chance (Bonferroni's principle - in large datasets we can find statistical artefacts). Therefore, the validation of experts is important to the veracity of the data.

The methods and techniques of Big Data have been developing fast in the past years. When talking about Big Data usually we refer to data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time. It requires techniques and technologies with new forms of integration to reveal insights from datasets that are diverse, complex, and of a massive scale [SMR12]. The data growth challenges and opportunities are used to describe Big Data by the following characteristics:

Volume - the quantity of generated and stored data (measured in bits, objects, rows, etc.) can determine the value and potential insight;

Velocity - the speed at which the data is generated and processed relative to the demands and challenges that lie in the path of growth and development;

Variety - the type and nature of the data (the diversity of sources, formats, quality, structures, etc.) can help people who analyse it to effectively use the resulting insight;

Veracity - the validation of the correctness of the large amount of rapidly arriving data;

Variability - the inconsistency of the data set can hamper processes to handle and manage it;

Value - the quality of the collected data can vary greatly, affecting accurate analysis.

As a result, big data solutions are characterized by real-time complex processing and data relationships, advanced analytics, and search capabilities. These solutions emphasize the flow of data, and they move analytics from the research labs into the core processes and functions of enterprises. Moreover, Big Data is often a cost-free by-product of digital interaction. It often doesn't ask why and simply detects patterns [MC13].

Today, big data is becoming a business imperative because it enables aquaculture professionals to accomplish several objectives:

- Apply analytics beyond the traditional analytics use cases to support real-time decisions, anytime and anywhere;
- Tap into all types of information that can be used in data-driven decision making;
- Empower people in all roles to explore and analyse information and offer insights to others;
- Optimize all types of decisions, whether they are made by individuals or are embedded in automated systems by using insights that are based on analytics;
- Provide insights from all perspectives and time horizons, from historic reporting to real-time analysis, to predictive modelling;
- Improve production outcomes, forecast the weight of measured impact factors and manage risk, in the moment of assessment and in future instances.

The systems infrastructure must capitalize on real-time information optimized for analytics, to respond dynamically to evermore increasing demands. To achieve efficiency, analytics must run close to the data while in motion. Therefore, the storage infrastructure must embody a defensible disposal strategy that reduces risk, rate of storage and legal expense. Privacy and data protection are also relevant, safeguarding all the data and insights on which a business relies. When infusing analytics, strong security measures are needed to be built in order to guard against internal and external threats.

In short, big data provides the capability for aquaculture to reshape itself, dynamically adapting to the changing needs of its customers by using information from a wide range of sources. Big Data within aquaculture points to five major categories:

- Developing a 360-degree view of the production;
- Understanding operational analytics developing their business intelligence;
- Addressing threats, fraud, and security in a more efficient manner;
- Analysing information that before was not thought as usable;
- Offloading and augmenting data warehouses.

Each of these broad categories can lend itself to a different architecture and mix of technologies. Therefore, each calls for different priorities for performance and capacity.

Big data technology must support search, development, governance and analytics services for all data types—from transaction and application data to machine and sensor data to social, image and geospatial data, and more. The Big Data landscape is dominated by two classes of technology:

- systems that provide operational capabilities for real-time, interactive workloads where data is primarily captured and stored;
- systems that provide analytical capabilities for retrospective, complex analysis that may touch most or all of the data.

These classes of technology are complementary and frequently deployed together.

Operational and analytical workloads for Big Data present opposing requirements and systems have evolved to address their particular demands separately and in very different ways. Each has driven the creation of new technology architectures. Operational systems, such as the NoSQL databases, focus on servicing highly concurrent requests while exhibiting low latency for responses operating on highly selective access criteria. Analytical systems, on the other hand, tend to focus on high throughput; queries can be very complex and touch most if not all of the data in the system at any time. Both systems tend to operate over many servers operating in a cluster, managing tens or hundreds of terabytes of data across billions of records.

NoSQL encompasses a wide variety of different database technologies that were developed in response to the demands presented in building modern applications. Relational databases were not designed to cope with the scale and agility challenges that face modern applications, nor were they built to take advantage of the commodity storage and processing power available today. Aquasmart takes profit of the advantages of MongoDB, a cross-platform document-oriented database classified as a NoSQL database. It eschews the traditional table-based relational database structure in favor of JSON-like documents with dynamic schemas, making the integration of data in certain types of applications easier and faster.

For operational Big Data workloads, NoSQL Big Data systems such as document databases have emerged to address a broad set of applications, and other architectures, such as key-value stores, column family stores, and graph databases are optimized for more specific applications. NoSQL technologies, which were developed to address the shortcomings of relational databases in the modern computing environment, are faster and scale much more quickly and inexpensively than relational databases. Critically, NoSQL Big Data systems are designed to take advantage of new cloud computing architectures that have emerged over the past decade to allow massive computations to be run inexpensively and efficiently. This makes operational Big Data workloads much easier to manage, and cheaper and faster to implement. In addition to user interactions with data, most operational systems need to provide some degree of real-time intelligence about the active data in the system. For example, in a multi-user game or financial application, aggregates for user activities or instrument performance are displayed to users to inform their next actions. Some NoSQL systems can provide insights into patterns and trends based on real-time data with minimal coding and without the need for data scientists and additional infrastructure.

Analytical Big Data workloads, on the other hand, tend to be addressed by MPP database systems and MapReduce. These technologies are also a reaction to the limitations of traditional relational databases and their lack of ability to scale beyond the resources of a single server. Furthermore, MapReduce provides a new method of analysing data that is complementary to the capabilities provided by SQL. As applications gain traction and their users generate increasing volumes of data, there are a number of retrospective analytical workloads that provide real value to the business. Where these workloads involve algorithms that are more sophisticated than simple aggregation, MapReduce has emerged as the first choice for Big Data analytics. Some NoSQL systems provide native MapReduce functionality that allows for analytics to be performed on operational data in place. Alternately, data can be copied from NoSQL systems into analytical systems such as Hadoop for MapReduce.

4.3.2 Opportunities for aquaculture

In the following paragraphs, we turn our focus to the specific characteristics of aquaculture and discuss what are the concrete advantages of considering Big Data methods in that context. We describe the sources of Big Data in that context. We also discuss the Aquasmart approach to Big Data analytics in general. The latter is a topic to be developed in Section 5.

According to the latest available statistics collected globally by FAO [FAQ14], the global fish production has grown with the world aquaculture production attaining 90.4 million tonnes (live weight equivalent) in 2012, of which 66.6 million tonnes are food fish (aquatic animals produced for the intended use as food for human consumption). The food fish supply increased at an average annual rate of 3.2%, outpacing world population growth at 1.6%. This development has been driven by a combination of population growth and rising life standards and facilitated by the strong expansion of fish production and more efficient distribution channels. In Europe, the largest market for fish in the world with increasing consumption, aquaculture accounts for about 20% of fish production with 65% of the seafood consumed in the EU being imported [ECAq]. The future demand for fish is expected to increase due to increasing population and income and health benefits associated with fish consumption. With it, the increase of aquaculture is expected to provide the opportunity for the access to Big Data in the sector.

The sources of Big Data in aquaculture are often originated from:

- the ability to track much more information than we used to;
- new and pervasive sensors, measuring water temperature, oxygen level, etc.;

- the ability to collect all recorded data, whether or not we need it or can extract relevant information from it.

The problem is the complex modelling other than counting. Modelling and reasoning with data of different kinds can get extremely complex. Data streams are common source of Big Data. The (problematics of) data stream management include blocking query operators that need the entire input to produce any result (e.g. sort, sum, max) or use approximations, sampling, window of data [ZE11]. They must consider issues brought up by data operators – collect, prepare, represent, model, reason, visualize – and additional issues – usage, quality, context, streaming, scalability. The approaches to process data streams include maintaining simple statistics (mean, standard deviation, etc.), or using time windows [WR15]. Time windows in data stream processing include:

- sliding (fixed size – e.g. the last 100 values);
- landmark (fixed start – e.g. from the start of the day);
- tilted (recent data in more details – e.g. last hour in 15 mins, last day in 24 hours, last month in days, last year in months).

Data streams in aquaculture are often sensors producing big data by a continuous flow of sensor readings often at a very high speed, in dynamic and time changing environment, large number of sensors on different locations. Sensor data samples the population. It is noisy and often requires cleaning. Moreover, it can be duplicated, and it has spatial and temporal attributes playing a major role in the later interpretation of their data [LRU14]. Queries in sensor networks can be:

- one shot vs long-running;
- all data vs aggregate data;
- accurate vs approximate;
- urgent vs delay tolerant;
- pull vs push.

Big Data analytics within the Aquasmart data streams analysis considers:

- smart sampling of data, reducing the original data while not losing the statistical properties of data;
- finding similar items with efficient index;
- incremental updating of the models;
- distributed linear algebra, dealing with large sparse matrices.

AI can be used for solving a wide spectrum of problems:

- optimization (e.g. diet, temperature, light, etc.)
- pattern recognition (e.g. anomalies, size, etc.)
- prediction (e.g. growth, diseases, etc.)
- automation (e.g. feeding, water quality, etc.)

sampling on big data enables off-line data analysis, enabling performing expensive operations

As mentioned in the section 4.9.6 of the technical architecture deliverable 2.2, the primary focus of this project is in providing the end users with tools to explore the analytics of their data, meeting their needs much as data can do. As for the long term scalability, Apache Spark will possibly be used for fast clustering computing. Spark is an open source cluster computing framework well-suited to machine learning algorithms, permitting multi-stage in-memory primitives and providing performance up to 100 times faster [Z11].

In this project, we will be using Python libraries that permit the manipulation of Big Data and can deal with machine learning algorithms. In particular, we use Scikit-learn [PVGT11], a set of simple and efficient tools for data mining and data analysis, widely used by the machine learning community. It is built on the Python libraries NumPy, Scipy and matplotlib permitting classification, regression, clustering, dimensionality reduction, feature extraction and normalization. Moreover, it is open source but is also commercially usable under a BSD license.

5 MODELLING AND ALGORITHMS

5.1 Determining influential factors

In this section, we discuss how we determine the factors that influence aquaculture production (e.g. temperature, season, size, av. weight, oxygen, pH, local features) in accordance with the business questions to be answered and business KPIs to be considered. Moreover, we describe the generalised linear regression used for that aim and a novel method – the stream story – to identify the relevance of features by simultaneous analysis of time-series.

5.1.1 KPIs and datasets

In the following paragraphs we briefly discuss the business key performance indicators [KPIs] and business questions to be considered in the context of Aquasmart Big Data analytics. The project's KPIs were introduced, described and exemplified in the section 4 of the deliverable 2.4. Along with this, we also described their associated metrics for KPI modelling and business analytics. The formulas to determine and quantify the measurements of these KPIs in the context of the available data are expressed in that section and complemented by the Section 5.1.2 of this deliverable.

5.1.1.1 Business KPIs in analysis and Business Questions to be answered

The following paragraphs identify the core business KPIs to be tackled in this deliverable, and describe the relation between those KPIs and the business questions to be answered.

Core set of business KPIs

The following KPIs will be used:

- FCR
- SGR
- GPD
- Mortality %
- Production time

There could be some other KPIs like *Protein Efficiency Ratio* but we will focus on the ones above.

There is an obvious relation between business questions and user stories. In the following list we present some of the user stories to be considered for implementation within Aquasmart. In **bold**, we represent the user stories that directly relate with the business questions to be answered. Also with a (*) we represent the user stories that were considered of high priority by the end users that are partners in this project.

- Evaluate feeding policy*

- Evaluate feed performance*

- Evaluate fry quality*

- **Evaluation of production practices***

- Evaluate the influence of the environment*

- Estimate average weight*

- Estimate fish count*

- Evaluate feeding process

- Define extra parameters

- Request analysis of data

- View a graph of my data

In the following we briefly describe the business questions to be answered in the context of this deliverable with the help of the Big Data analytics made available in this project.

1. **Evaluation of feed performance.** This is done vs. model or other feeds. We want to evaluate feed suppliers, feed types and feed composition, taking into account also the time dimension (winter, summer).
2. **Evaluation of feeding process.** We want to evaluate people (feeders), feeding rates and feeding times.
3. **Evaluation of fry quality.** We want to evaluate different hatcheries, brood stock origins and hatchery qualities.
4. **Evaluation of production strategies.** Stocking month, feeding policies (e.g. time to change feed size), protocols for grading, unit type and size, vaccinations.
5. **Evaluation of the influence of the environment.** Farm, anchor, unit, other environmental data like oxygen level, water currents, weather.
6. **Estimation of fish number and average weight.** Based on the feed consumption.

5.1.1.2 Datasets

In the following paragraphs we discuss the datasets considered in Aquasmart and furthermore discuss their intersection with KPIs and business questions.

The LTD datasets can be distinguished within an operational point of view as:

1. LTD datasets (normalized) until the first "good" harvest (i.e., "enough" fish have been harvested);
2. LTD datasets (normalized) until final harvest;
3. Daily datasets including feeding, model feeding and all the daily transactions.

Normalization or pre-processing is very important in aquaculture. Sometimes there are big differences between the estimated and actual number of fish. This is usually found in the final harvest. In such cases, it is necessary to manually process the dataset and either adjust the initial fish number or split the difference into different time periods. All the dataset attributes including the KPIs are recalculated and then the dataset reflects the reality. The datasets of type (1) can help us answer business questions that remain stable thorough out the lifetime of a fish population.

- Hatchery evaluation
- Stocking month evaluation
- Region evaluation

On the other hand, the datasets of type (2) need careful handling on the Data after the first harvest. It may help us evaluate "big size fish" strategies. The whole handling strategy is the same as LTD dataset

Finally, the dataset of type (3), the daily dataset, is the best candidate for "Big Data" strategies. Most columns are calculated and "follow" a deterministic model. The challenge here is to use some pattern recognition, time series and outlier detection techniques in order to use the daily dataset. We have sampling information that we can use to identify the real 'Av. weight' or get a better estimation of it. Then we can recalculate our core KPIs.

In the following two tables we present the correlation matrices that allow us to compare datasets, business questions and KPIs globally for Aquasmart.

Dataset **Question**	Sampling-to-sampling dataset	LTD Dataset until first good harvest	LTD Dataset of finished populations	Daily dataset
Evaluation of feed performance	√			√
Evaluation of feeding	√			√
Evaluation of fry quality	√	√	√	
Evaluation of production strategies	√	√	√	
Evaluation of the influence of the environment	√	√	√	√
Estimation of fish number and average weight	√			√

Table 5.1: Correlation matrix: Datasets - Business Questions

KPI **Question**	FCR	SGR	GPD	Mortality	Production time
Evaluation of feed performance	√			√	
Evaluation of feeding	√			√	
Evaluation of fry quality		√	√		√
Evaluation of production strategies	√	√	√		
Evaluation of the influence of the environment	√	√	√	√	
Estimation of fish number and average weight	√				

Table 5.2: Correlation matrix: Business questions – KPIs

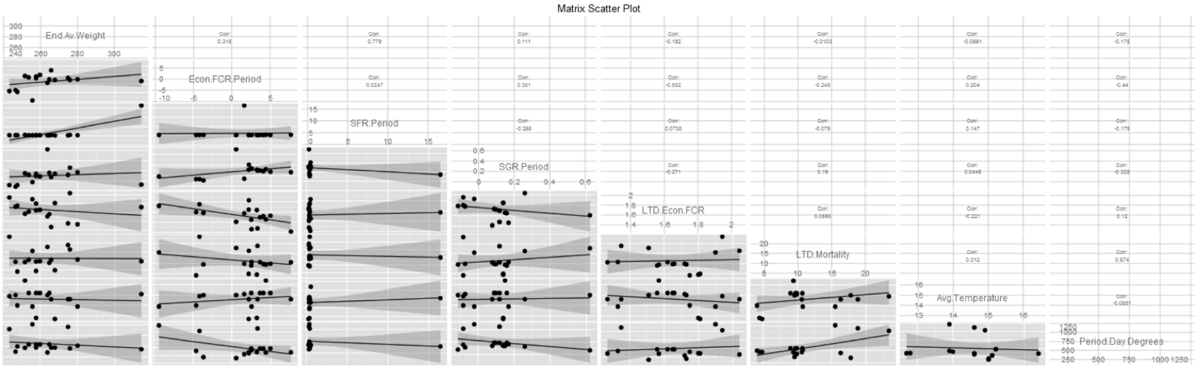


Figure 5.1: Example: scatterplot showing the relationships between all parameters affecting the main KPIs

5.1.2 Formulation of KPIs for data analytics

In the following paragraphs we present, describe and further formalize the entities and KPIs in present in this deliverable, associated to data analytics in aquaculture. This complements the information presented in section 4.2.3 of the deliverable D 2.4. This inventory of KPI formulations was made based on the expert information of the aquafarmers, and on the online resources [aqu16] and [glo16].

BWG – Body Weight Gain

$$BWG = Final\ body\ weight(g) - Initial\ body\ weight(g)$$

CV – Coefficient of Variation (or weight Coefficient of Variance)

$$CV = \frac{Standard\ deviation}{Mean} * 100$$

DGR – Daily Growth Rate

$$DGR = \frac{Final\ body\ weight(g) - Initial\ weight(g)}{Period(days) * Inital\ weight(g)} * 100$$

FCR – Feed Conversion Rate

Net growth = Biomass at the end of the period – Initial Biomass + Biomass of harvests + Biomass transferred to other units – Biomass transferred from other units
Gross growth = Net growth + Biomass of mortalities + Biomass of culling + Biomass of adjustments

$$Economical\ FCR = \frac{Total\ dry\ feed\ given}{Net\ growth}$$

$$Biological\ FCR = \frac{Total\ dry\ feed\ given}{Gross\ growth}$$

$$FE(\%) = \frac{1}{FCR * \ 100}$$

SFR – Suggested Feeding Rate

$$SFR = \frac{Daily\, feed}{Biomass} * \ 100$$

Daily feed could also be period feed → period SFR

SGR – Specific Growth Rate

$$SGR(\%/day) = \frac{lnWf - lnWi}{t} * \ 100$$

lnWi = the natural logarithm of the initial average body weight(g)

lnWf = the natural logarithm of the final average body weight(g)

t = time(days) between lnWi and lnWf

Note: If the Food Conversion Rate is known, the SGR can also be calculated by dividing the Percentage body weight fed per day by the food conversion rate. This calculation can be turned round to predict growth if the SGR is known.

$$SGR(\%/day) = \frac{\%body\ weight\ fed/day}{FCR}$$

GPD – Growth Per Day, ADG – Average Daily weight Gain

Average weight gain = Average body weight at the end of the period – Average body weight at the start of the period

$$ADG(g/fish\,/day) = \frac{Average\ weight\ gain(g)}{Period(days)}$$

MR – Mortality Rate

$$MR = \frac{LTD\ mortalities}{Initial\ fish\, number} * 100$$

FR – Feed Rate

The amount of food given to fish over a specified period of time. The most common way of expressing this is as percentage of the animal's body weight per day. For example a 1000 gram fish, being fed 20g of food per day would be on a 2% feed rate [(20 / 1000) x 100)].

$$FR = \frac{Feed}{Biomass} * 100$$

$$Mortality\ biomass(kg)\ = \frac{Mortality\ No. *\ Av.Weight(g)}{1000}$$

SR – Survival Rate

$$SR = \frac{Number\ of\ fish\ harvested}{Initial\ fish\ number} * 100$$

It is not correct to say that $Survival\ \ Rate\ = \ 100 - Mortality\ Rate$ because there are also other reasons that reduce the fish numbers (e.g. adjustments which are missing or extra fish and culling).

5.1.3 Generalised linear regression models to determine impact factors.

In the following section we discuss how to apply generalized linear models (GLMs) to determine the influential factors in the aquaculture production. Having the features previously determined by the expert knowledge and the data collected we will use automatic feature selection methods to determine those influential factors by constructing a classifier. This consists on the process of selecting a subset of relevant features (variables, predictors) for use in model construction [GWHT13]. That will permit to:-

- reduce the noise in the output of the forecast model;
- simplify it to make it easier to interpret by researchers/users;
- enhance generalization by reducing overfitting;
- shorten the training times;
- improve its overall computational efficiency.

The main argument when using a feature selection technique is that the data contains many features which are either redundant or irrelevant, and can thus be removed without incurring much loss of information [BPS15]. Notice that redundant or irrelevant features are two distinct notions, since one relevant feature may be redundant in the presence of another relevant feature with which it is strongly correlated [GE03].

We will start by distinguishing the roles of two classical methods: correlation vs. regression. Correlation refers to any of a broad class of statistical relationships involving any statistical relationship between two random variables or two sets of data. Correlation examines if there is an association between two variables, and if so to what extent.

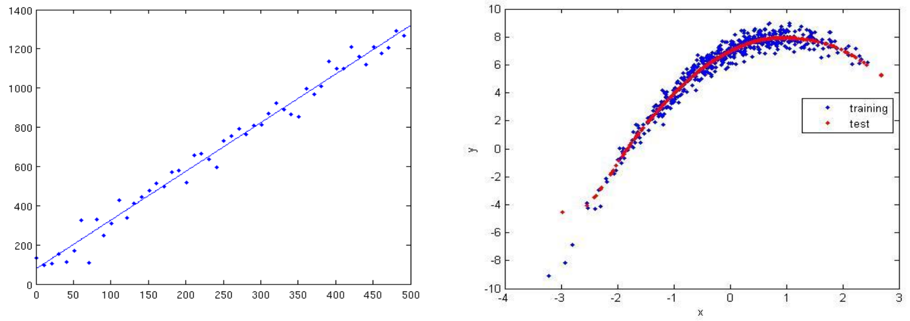


Figure 5.2: Example of correlation (on the left) and regression (on the right)

Regression establishes an appropriate relationship between the variables. Its aim is to discover how a dependent variable Y is related to one or more independent variables X. Ordinary linear regression predicts the expected value of a given unknown quantity (the response variable, a random variable) as a linear combination of a set of observed values (predictors).

The goal of statistical classification is to use an object's characteristics to identify which class (or group) it belongs to. A linear classifier achieves this by making a classification decision based on the value of a linear combination of the characteristics. An object's characteristics are also known as feature values and are typically presented to the machine in a vector called a feature vector. Such classifiers work well for practical problems such as document classification, and more generally for problems with many variables (features), reaching accuracy levels comparable to non-linear classifiers while taking less time to train and use [GCC12]. Examples of training of linear classifiers include:

- Logistic regression — maximum likelihood estimation the feature vector assuming that the observed training set was generated by a binomial model that depends on the output of the classifier;
- Support vector machine — an algorithm that maximizes the margin between the decision hyperplane and the examples in the training set.

The most common method for analysing binary response data is logistic regression which is used to model relationships between the response variable and several explanatory variables which may be categorical or continuous. Logistic regression has been generalized to include responses with more than two nominal categories [DB08]. We now turn our attention to two types of models where the response variable is discrete and the error terms do not follow a normal distribution, namely logistic regression and Poisson regression. Both belong to a family of regression models called *generalized linear models* (GLMs).

In a GLM, each outcome of the dependent variables, Y, is assumed to be generated from a particular distribution in the exponential family, a large range of probability distributions that includes the normal, binomial, Poisson and gamma distributions, among others. In linear regression, the use of the least-squares estimator is justified by the Gauss-Markov theorem, which does not assume that the distribution is normal. When considering generalized linear models it is useful to suppose that the distribution function is the normal distribution with constant variance and the link function is the identity, which is the canonical link if the variance is known. The data is divided in three parts: 80% of the data will be the training set, 15% will be test data, and 5% will be assigned to cross-validation.

5.1.4 Stream story for aquaculture

The in-house developed software *stream story* (freely available at http://streamstory.ijs.si) provides a qualitative representation of sensory data or any time-series data. It permits data visualization as a hierarchical state machine. It also permits the user to explore the distribution of states and transitions. Moreover, it offers prediction and anomaly detection services.

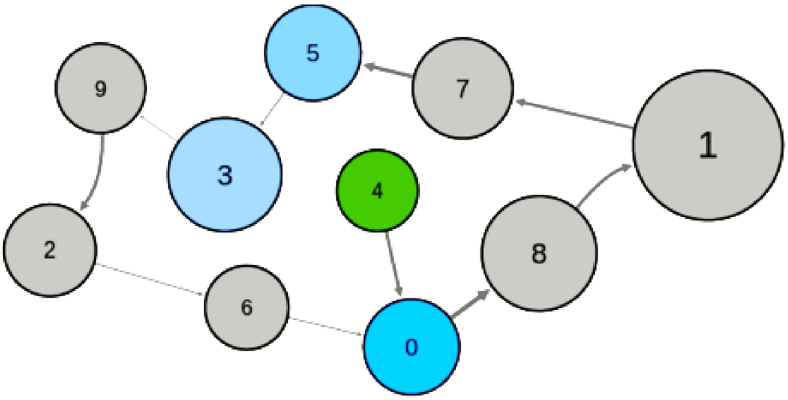
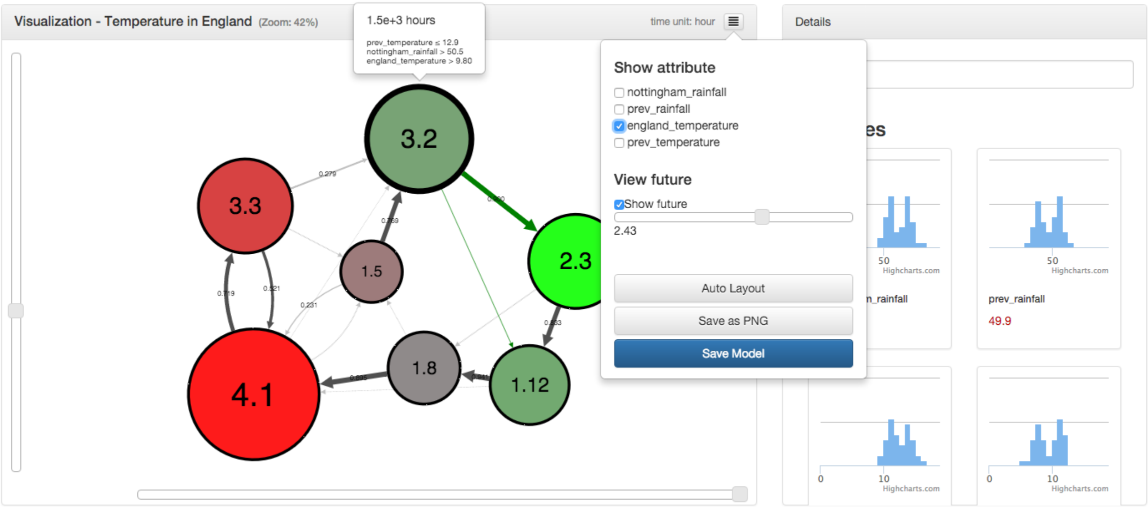


Figure 5.3: Example of a stream story diagram of states for the temperature in the UK

It permits to analyse simultaneous time-series data streams corresponding to features that are being studied and determine the impact of those features against one or more selected features. This is done by a hierarchical continuous-time Markovian process, where the states, transitions and hierarchy, are automatically learned from the data.

An example is the weather in UK, where the size of the states corresponds to the time spent in that states. When 'clicking' with the mouse in one state, the user can see the bar charts analysis of the attributes of the data corresponding to that state. The arrows between states are the direct relations between those states.

The user can also access the decision tree that provides the hierarchy between the features in analysis and their impact on the features being studied. It tries to explain to the user what characteristics are particular at considered state in comparison to all the other states. It is a classifier that permits the relation with other classifiers. If we get complicated trees we can extract rules to deal with the data stream in analysis.


Visualization - Temperature in England (Zoom: 42%)

1.5e+3 hours

prev_temperature ≤ 12.9
nottingham_rainfall > 50.5
england_temperature > 9.80

time unit: hour

Show attribute

□ nottingham_rainfall
□ prev_rainfall
☑ england_temperature
□ prev_temperature

View future

☑ Show future

2.43

Auto Layout

Save as PNG

Save Model

Details

Figure 5.4: Screenshot of the stream story in action exhibiting the diagram of states and the histograms analysing the selected features

It also provides predictive information based on the relations established whenever one state is selected. Moreover, the mouse over one state provides information on that state.

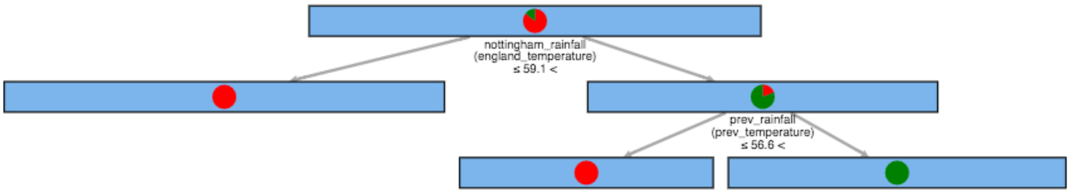


Figure 5.5: Decision tree providing the hierarchy between the features in analysis and their impact on the features being studied

The input data is a CSV file where each column is a feature in the study with its name in the first line and the first column is the time stamp. Hence, each column except the first is a sequence of time-series data. The user can then choose the features in study, the ones we want to look into the level of influence, and the ones that will be influenced. The Stream Story system splits the input data streams into two sets:

- observation parameters – that can only be observed but not directly manipulated (e.g. temperature of the water or oxygen levels);
- control parameters – that can also be observed but permit direct manipulation and may influence the behaviour of the observation parameters and, consequently, of the overall system (e.g. the amount of food given to the fish will directly influence their growth).

A typical example in aquaculture would provide the relation between the growths of the fish. The fish classified by their average weight would be represented by classes and the arrows between states would be the growth relations. The size of the state is then the time spent by the fish in a particular state. The attributes in study would be the set of features from which we would like to determine the impact of, within the fish production. In that we could manipulate the time of growth and change of state. The small size of the initial and final state corresponds to the small amount of time while being a fry and while being at the appropriate average weight for harvest, respectively.

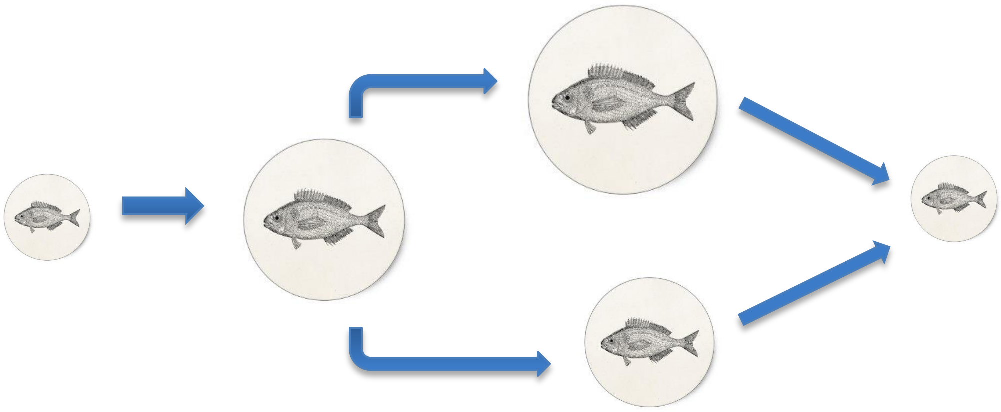


Figure 5.6: Sketch of the dynamics of the stream story in the context of Aquasmart

5.2 Modelling and Prediction

In the following paragraphs we describe various modelling techniques that will be selected and applied by Aquasmart. Their parameters are calibrated to optimal values. Some techniques have specific requirements on the form of data used.

5.2.1 Model selection for Aquasmart

5.2.1.1 General methodology

In this section, we briefly describe the methods we believe are most applicable to aquaculture an Aquasmart in particular. With each method, we also highlight its most important advantages and disadvantages relevant for our use cases. Methods from the machine learning Python library Scikit-learn will be used for implementation. This efficient open source library enables to quickly test various advanced algorithms, in order to discover the most appropriate stack of methods for our needs.

Naïve methods are considered as unsophisticated forecast approach where no forecast model is actually built [BESM00]. This method provides a baseline and will be applied using similar measurement from the past (1 hour, 1 day, 1 week...). It can produce surprisingly good results if the data is very periodic in nature. Historical sets of data have to be maintained. It is widely used in practice because of its undemanding properties, which makes it easy to develop and maintain, and is commonly used in pre-processing steps such as filling in missing data and smoothing. Due to its low computational complexity the methods are very fast, but their accuracy is also poor. So though it can offer good results, it is usually used as baseline comparison for other more sophisticated models.

Regression analysis is a statistical process for estimating the relationships between variables in order to estimate unknown model parameters from the data [DRSM98]. In particular, Ridge Regression methods require a set of historical data to learn and update the model. A recursive formula enables us to incrementally update the model instantly with every new record [VIDS05]. To estimate a single regression model with more than one outcome variable we shall use the more general multivariate regression. In particular, if there is more than one predictor variable in a multivariate regression model, the model is a multivariate multiple regression [JW92]. It is a well-known fact that despite the simplicity, this method can produce very good results, with proper feature engineering. The method is computationally undemanding and is fast. Since it uses real-time data in combination with historical data, the method is appropriate for both short and long-term predictions. The main advantage of this method is that there are no parameters that have to be optimized for the model to produce good results. By its nature, linear regression only looks at linear relationships between dependent and independent variables. Linear regression assumes that the data is independent (the scores of one subject have nothing to do with those of another). This is often, but not always, sensible. Outliers can have huge effects on the regression.

K nearest neighbours [kNN] is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The kNN algorithm is among the simplest of all machine learning algorithms. Both for classification and regression, it can be useful to assign weight to the contributions of the neighbours, so that the nearer neighbours contribute more to the average than the more distant ones. Several different variations of this method exist which can speed up the search of nearest neighbours. A relatively simple method that while usually better than baselines, can be improved upon with more complex methods. It is a non-linear method that can find some patterns in historical data that linear models cannot, and is very robust to noise (especially when inverse square of weighted distance for computing distance is used). The biggest drawback for our testbed is that kNN is not an online method, meaning that it has to iterate through entire collection of records every time we want to make a prediction, which can be costly in time and computing resources. Some indexing (e.g. K-D tree) may reduce this computational cost. An optimal parameter k (number of neighbours) has to be found for best results.

Random forests is an ensemble learning method for both classification and regression tasks, that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random forests correct for decision trees' bad habit of overfitting to their training set [BREI01]. This ease of use also makes Random Forests an ideal tool for people unexperienced in statistics, allowing them to produce fairly strong predictions free from many common mistakes, with only a small amount of research and programming. Since they have very few parameters to tune and can be used quite efficiently with default parameter settings (i.e. they are effectively non-parametric) Random Forests are good to use as a first cut when you don't know the underlying model, or when you need to produce a decent model under severe time pressure. The main limitation of the Random Forests algorithm is that a large number of trees may make the algorithm slow for real-time prediction. But number of trees can be optimized as an input parameter. This is a known trade-off between accuracy and speed. By default, Random Forests is not an online method, but variations for streaming applications exist.

Support vector machines (SVMs) are supervised learning models that analyse data and recognize patterns, used for classification and regression problems. The algorithm originally gained popularity due to its success in handwritten digit recognition. SVM computes a hyper plane (decision boundary) to separate data in two classes. In order to select the best hyper plane, SVM uses the “margin” notion. Margin is the distance between the hyper plane and the nearest training data point of any class, so the best hyper plane is one that has the largest margin. The “support vectors” are those that are situated on the margin, and they are the ones that define the model. SVR uses kernel methods and is very good at solving non-linear problems. It also has a regularization parameter, which avoids the overfitting. An SVM is defined by a convex optimization problem (no local minima) for which there are efficient methods (e.g. SMO). The main disadvantage of SVM is that it is not suitable as an online (streaming) method, and due to its complexity it takes quite a long time to train a model. Therefore, when new records are added the dataset, the model requires complete retraining, which is extremely costly in terms of time and compute resources.

Artificial Neural Networks [ANNs] are models that are inspired by the structure and/or function of biological neural networks. They are a class of pattern matching that are commonly used for regression and classification problems but are really an enormous subfield comprised of hundreds of algorithms and variations for all manner of problem types [HIPS01]. The most common in load forecasting is backpropagation method [HIBS05]. The backpropagation algorithm uses supervised learning, which means that we provide the ANN with examples of the inputs and outputs we want the network to compute, and then the error is calculated. The errors here are referred to the difference between actual and expected results. The main idea of the backpropagation algorithm is to reduce this error, until the ANN learns the training data. The training begins with random weights, and the goal is to adjust them so that the error will be minimised. Neural Networks are very suitable for streaming applications, since they are iteratively updated with each new record. They are also very appropriate for large volumes of multivariate streaming data. But the biggest benefit is that ANNs are capable of modelling complex non-linear and dynamic processes. The biggest drawback of ANNs is that there are usually several parameters that have to be optimized in order to produce meaningful results. Most standard parameters to be set are: number of input layers, number of hidden layers and number of output layers. There are also other parameters that can improve performance such as learning rate, momentum, and learning functions. The problem is that if we are developing a decentralised system, optimizing all these parameters is time consuming, while if they are not set properly; overall performance can easily be worse than other “less-sophisticated” methods that do not require so many parameter optimisations.

5.2.1.2 Example: fish count model

As previously discussed in Section 4.1.2, the number of fish in a batch during the fish production is undetermined until harvest. Though, it is an important factor in the calculation of most of the aquaculture metrics.

As discussed in Section 3.1.3, there is exact data available from external sources (e.g. temperature, oxygen level) and from sampling (e.g. average weight). But, there exist several non-exact measurements, such as fish count and its dependent variables (e.g. number of fish mortalities). To better estimate the number of fish during the production cycle, we developed a mathematical learning model of exponential nature that estimates that number built on historical data of the number of mortalities and the adjustments made.

The fish count model used for the machine learning experiments is built according to the probability of the fish dying and/or disappearing for other reasons (escape etc.), using a Poisson distribution.

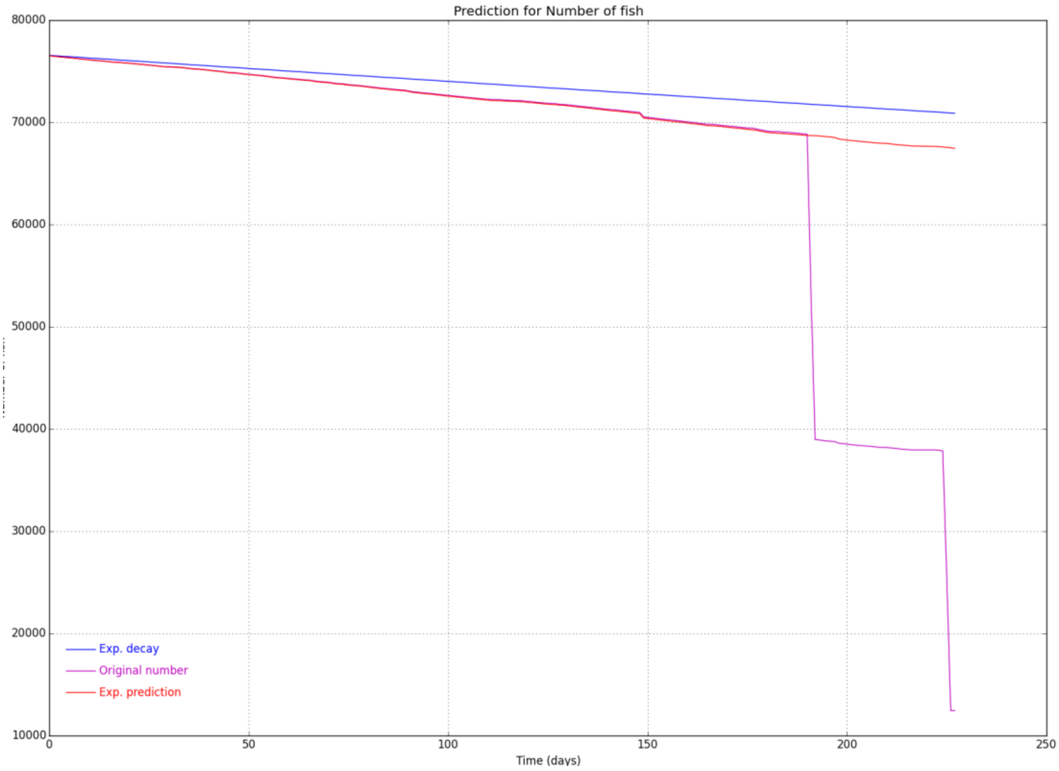


Figure 5.7: Fish count machine learning experiments: Graphical representation of one instance of our fish count prediction model where the pink line represents the original data, the red line represents our prediction and the blue line represents the ideal model.

In the above graphical representation of our fish count prediction model, the pink line represents the original data, the red line represents our prediction and the blue line represents the ideal model.

The ideal model represented refers to the number of fish dying and disappearing in an ideal way. It considers the first starting number of fries and the last number of fish harvested in the end of the production cycle.

To feed this model we have used generated data based on a growth model closely related to the von Bertalanffy growth model. The von Bertalanffy growth model is extensively used in fish farms to model the growth of the fish as a function of age from some origin t_0. If y(t) represents the growth measurement after time t, then for each fish:

$$E[y(t)] = L(1 - exp(-Kt))$$

for positive parameters L and K. With this parameterisation, L is the maximum length of the fish and K regulates the expected percentage of maximum growth measurement achieved after a particular age. Different growth curves will be created for each different set of parameters; therefore it is possible to use the same basic model to describe the growth of different species simply by using a special set of parameters for each species. This exponential model learns the parameters over the historical daily data, minimizing the error in the function. It considers the average of the two factors (mortality and adjustment). The error in the estimation is typically negative but there are cases where positive error might occur which would point out to incorrect number of fish at the beginning of the production cycle.

The growth of the fish in one unit is closely related with the earlier mentioned von Bertalanffy growth model. It simulates real data by modelling temperature and fish death count using the Poisson distribution. It includes the random death and disappearance of fish in a unit. It is as much as possible close to reality for developing algorithms and learning on real data. It is expressed as follows:

$$avg = Mg * Tc - exp(-0.008 * av\_wt * Tc + log(Mg * Tc))$$

where:

- avg = average weight gain;
- Mg = typical average weight gain;
- Tc = current temperature / average temperature (i.e., normalized temperature).

The value -0.008 was determined by learning the parameters of the model based on the historical data available. The log value inside the exponent permits us to start from zero. The average temperature was extracted from freely available online sources and parameterized accordingly.

The diverse outputs of the model and consequent incorrect estimations are a result of the problems with the input files. In fact, the transfer/harvest of fish can be done in any point of the production cycle disturbing the calculations. This is a matter that will be considered in the further developments of such model.

5.2.2 FCR and SFR Modelling

The Feed Conversion Ratio [FCR] is an important performance indicator to estimate the growth of the fish. It is widely used by the aquaculture fish farmers in pair with the Specific Feeding Ratio [SFR]. Its importance follows from the fact that 70% of the production costs in aquaculture are due to the food given to the fish during growth. Some of it will fall through the net and some will be spared. Optimization of the feeding process of the fish can thus confer great benefits to the economic development of the farms.

Specifically, the FCR permits the aquafarmer to determine how efficiently a fish is converting feed into new tissue, defined as growth [Sti07]. Recall that the FCR is a ratio that does not have any units provided by the formula:

$$FCR \ = \ dry\, weight\, of\, feed\, consumed / wet\, weight\, of\, gain$$

while the feed conversion efficiency (FCE) is expressed as a percentage as follows:

$$FCE \ = \ 1/FCR \ \times \ 100$$

There seems to be some controversy among aquatic animal nutritionists as to which is the proper parameter to measure, but in Aquasmart we use FCR (mentioning FCE here for completeness sake).

Example: if the total amount of gain over the 14-day interval of feeding was 10,500 – 10,000 kg or 500 kg, and the total amount of feed offered was 420 – 400 kg/day × 14 days = 20 kg/day × 14 = 280 kg, then the FCR and FCE are calculated as follows:

FCR = 500 kg fed/280 kg gain = 1.78

FCE = 1/1.78 × 100 = 56%

Recall that the FCR and FCE are based on dry weight of feed and fish gain, as the water in dry pelleted feed is not considered to be significant. A typical feed pellet contains about 10% moisture that will only slightly improve the FCR and FCE.

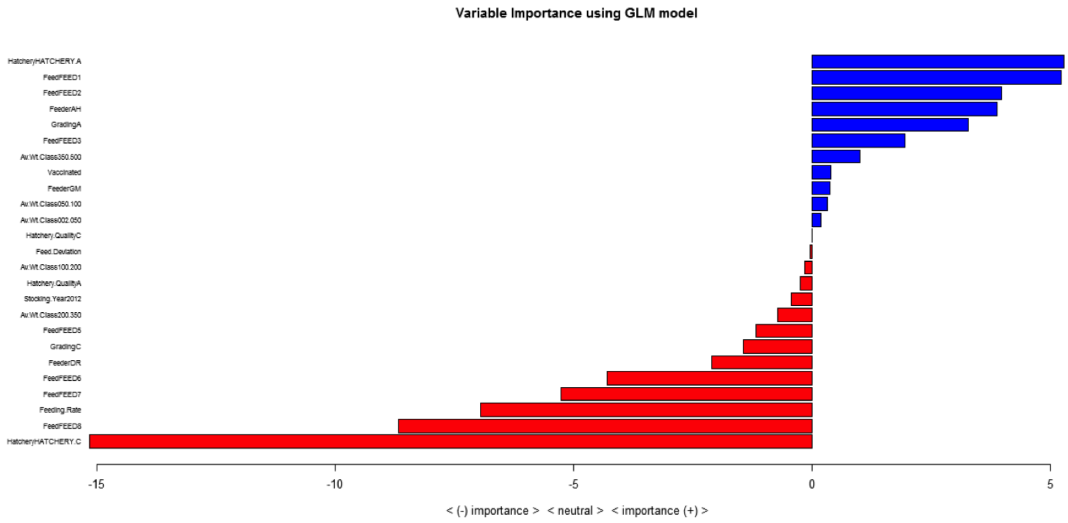


Figure 5.8: Variable importance using GLM model

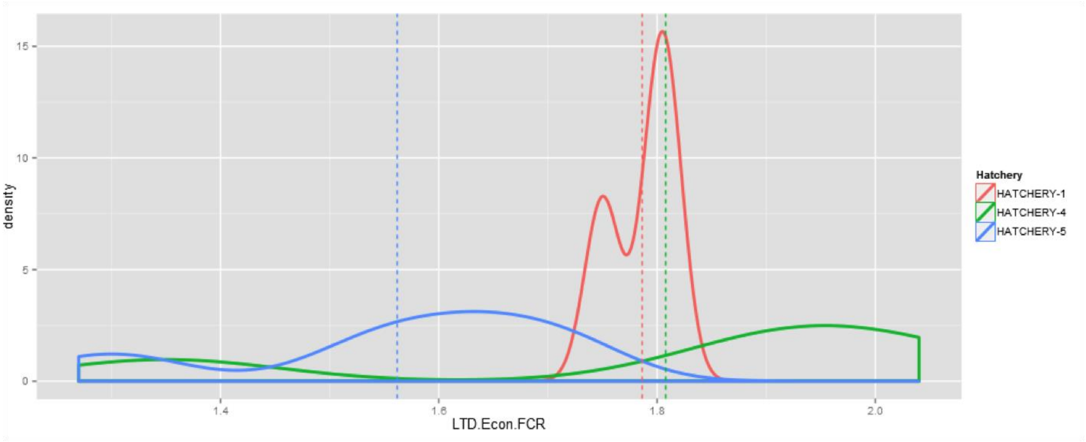


Figure 5.9: Example of evaluation of different hatcheries in terms of FCR. Hatcheries 1, 4 and 5 have the same average but 1 is unpredictable

Each aquaculture entity draws an appropriate FCR table to that batch of fish. Higher temperature leads to lower energy spent and faster growth, and consequently to a lower FCR. As the fish gets bigger, it requires more food to increase its biomass in percentage, and thus the FCR grows higher with the increase in average weight. The quality of the food and the size of the pellet are not considered at this point. At high temperatures (above 30 degrees in the case of bream and bass) low oxygen leads to low conversion to biomass. This is one of the hidden variables in the model, which should be considered separately at a later stage. One of the possibilities would be to penalize the FCR tables for the lack of oxygen. The other variable is the high reproduction of the fish in low temperatures and high average weight, which highly affects the growth of the fish.

Recall that the *Economic FCR* is the real FCR index following from the quotient between food given to the fish and the fish biomass. When the temperature is too high or too low we should ignore the data that is filled in with zeros and considered empirical data.

In the following we present the plots of the models for the three partner fish farms in Aquasmart. It includes 3 fish farms.

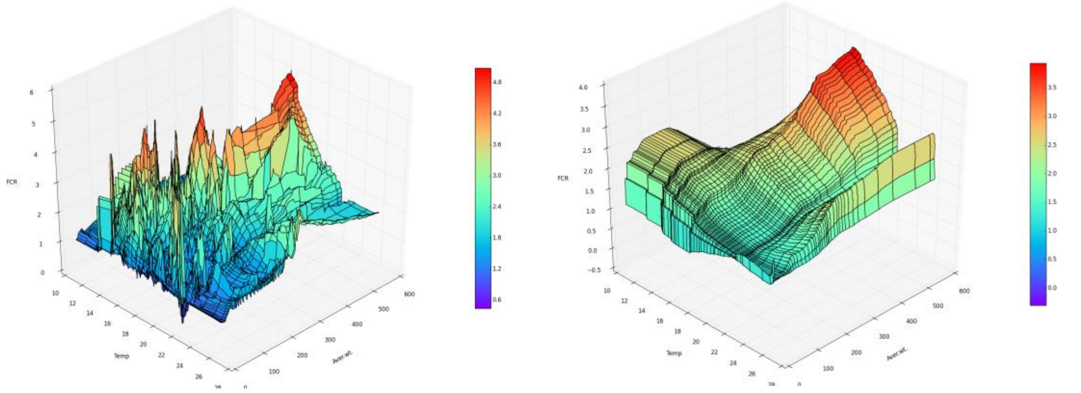


Figure 5.10: Company A: Real data (on the left) and FCR model (on the right) for the bream production

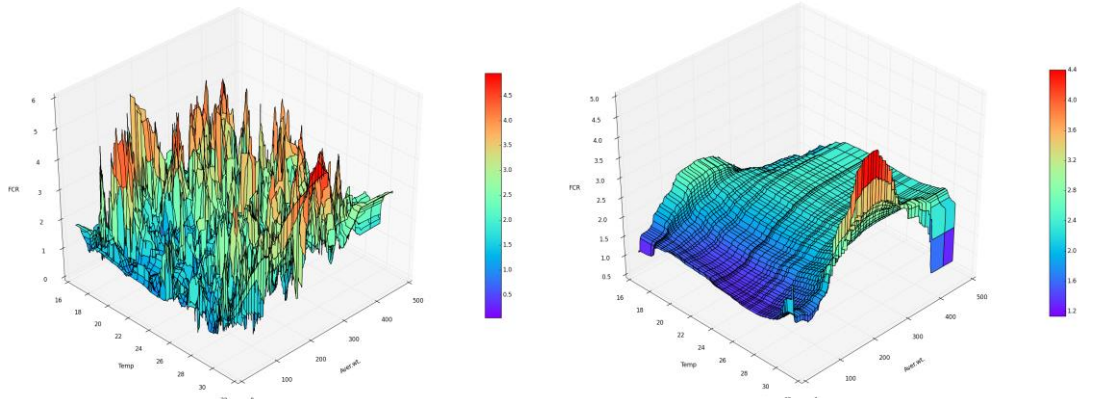


Figure 5.11: Company B: Real data (on the left) and FCR model (on the right) for the bream production

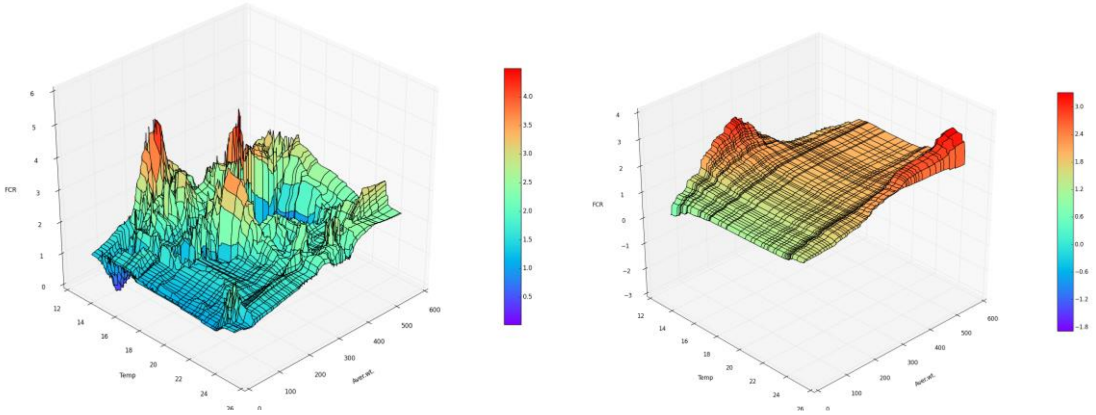


Figure 5.12: Company C: Real data (on the left) and FCR model (on the right) for the bream production

The model (on the right) produced based on the sample data (on the left) serves as a base of comparison with the historical data provided by a particular fish farm. Thus, with the new real data getting in our system, the fish farmer can compare it with the model and make an evaluation on the progress of the production. These models complement and confirm the expert knowledge: the high values on the right correspond to high fish reproduction in cold water temperatures and high average weight values. On the other hand, high temperatures represent low levels of oxygen which request higher feeding rate to maintain and increase the growth rate.

The large number of peaks in the real data, shown on the left above, corresponds to the real values. Typically, the input data can be seen within a grid. The following images show the grid view of both the real data (on the left) and the FCR model (on the right) for the company C.

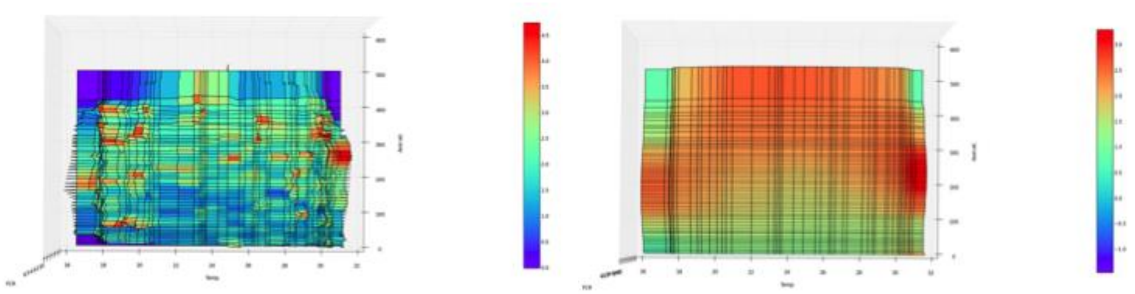


Figure 5.13: The grid view of both the real data (on the left) and the FCR model (on the right) for the company C

We then use least squares method to interpolate the missing values including all non-peak values as those interpolated values. It does so by approximate the solution of overdetermined systems. The average weight must be represented using specific values that are important in the fish production decision making process, and eventually distinct from fish farm to fish farm. Thus we consider a second interpolation to produce a final FCR table that is consistent with the systems in use by the fish farms. The nearest neighbour's algorithm is used here to find the values outside the area. That permits us to consider the complete table of measurements in line with the sample data available and the missing values calculated for the area inside the region.

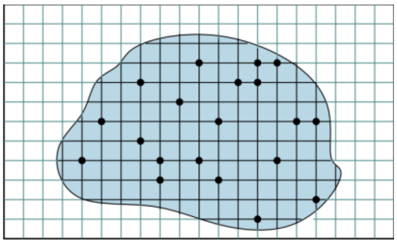


Figure 5.14: Sketch of the double interpolation in the model

The prediction can be done based on the model created here. Forecasting is a mature field with a vast amount of methods and approaches. The methods used can be grouped into 4 main classes:

- naïve approach;
- parametric approach;
- non-parametric data-driven approach;
- hybrid approach.

The Periodic SFR from the sampling data can be modelled in an analogous way to the FCR modelling described above.

We will further create a model to predict the FCR values from the sampling to sampling dataset. To do so we shall follow the classical steps of this process as the following:

Step 1 – Pre-processing

Pre-processing of the sampling-to-sampling dataset so as to create a representative dataset without faulty or extreme records/instances. This dataset will contain the interesting attributes, such as the Start average weight of the fish, the End average weight of the fish, the average temperature, the SFR, the feed type, characteristics of the farm (i.e. site, unit, batch, etc.) and other specific parameters that are important for each company. These explanatory variables are the input of the model. Also, this dataset will have to contain the response variable, the parameter that we want to predict, which is the FCR of a period.

Step 2 – Create/Evaluate the model

We will utilize the above dataset so as to create a model, which can predict the FCR values. This process is separated in two phases:

1. In the first we will train a model based on the training set of records. Specifically, a portion of records from the above dataset will be split, in order to use them as a training set. The remaining records (testing set) are used to evaluate the reliability of the created model. We shall use data mining methodologies to create a model, namely generalized linear regression models, generalized additive models, support vector machines and neural networks are among them.
2. In the second phase, we have to estimate the performance of the created model. Thus, we will use the records from the testing set, to feed the model with the inputs values and compare the model's response with the actual FCR values. The goal is to minimize the total error so as to have a reliable model. If the error is large we have to refine the model to use other methodology. This is a research issue and effort.

Step 3 – Predictions using the model

In this step, the training model predicts the FCR value based on the input values of unknown case. In other words, the users give a record with input values to the model, which returns the prediction of FCR.

The second option is to create a classification model. This approach is similar with the first one. The main difference here is that the target variable (output of the model) is discrete taking two values "good" or "bad". In the pre-processing step the end-users have to provide us a training set in which they have characterize each record as "good" or "bad". For the second step, the methodologies that we can use to create the model are classification trees, random forest, support vector machines and neural networks.

5.2.3 Time series classification and forecasting

The information provided by periodic datasets from sampling to sampling has been identified in cooperation with the user companies. This dataset gives us the basics KPIs of a unit between two samplings. With minor normalization effort, it can give us enough feedback for the evaluation on most business questions, as it involves most KPIs. Its small size is a challenge. It needs relatively small time frames between samplings, and needs relatively consistent handling (same feeder, same food type, etc.). This dataset is ideal for classification and regression algorithms. We can use it for short term forecasting of numeric KPIs with methodologies like Neural Networks (sensitivity analysis), SVM, additive generalized linear models etc.

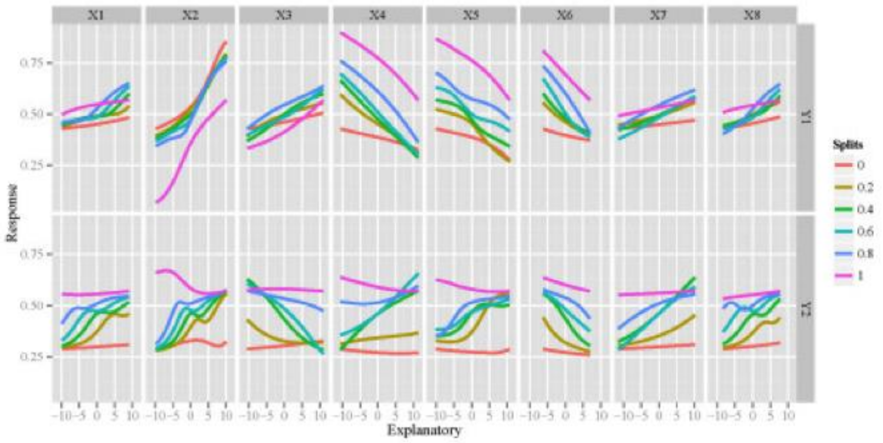


Figure 5.15: Sensitivity analysis of two response variables in the neural network model to individual explanatory variables. Slits represent the quantile values at which the remaining explanatory variables were held constant.

We can use any set of columns that users want in order to score /classify the outcome of each row. With the use of classification methods we intend not to be able to predict the class but the effect of numeric KPIs on the class (over/under performing cages etc.). Algorithms we have tested which appear appropriate are SVM classifier, GLMnet, Random Forests, and Neural Networks (Garson methodology).

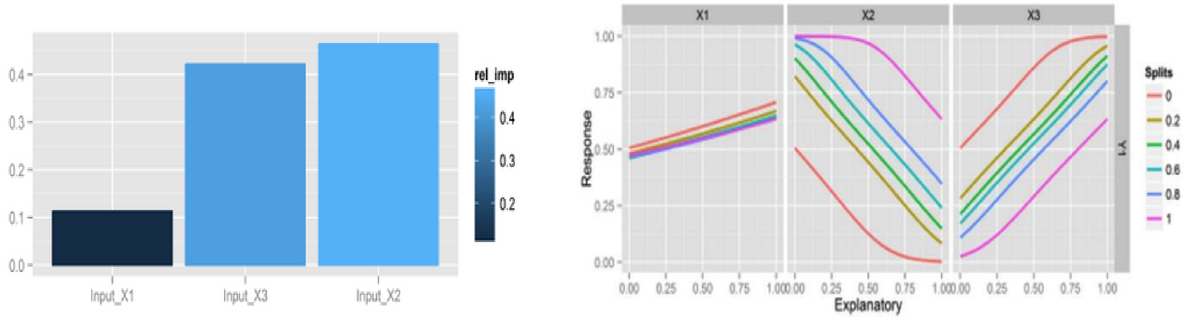


Figure 5.16: Measure of the real impact of the tree input variables X1, X2 and X3 (on the left). Detail of the sensitive analysis in the above neural network (on the right)

We can use clustering techniques to identify clusters of each KPI and maybe even common attributes (hatchery, food type) of these clusters.

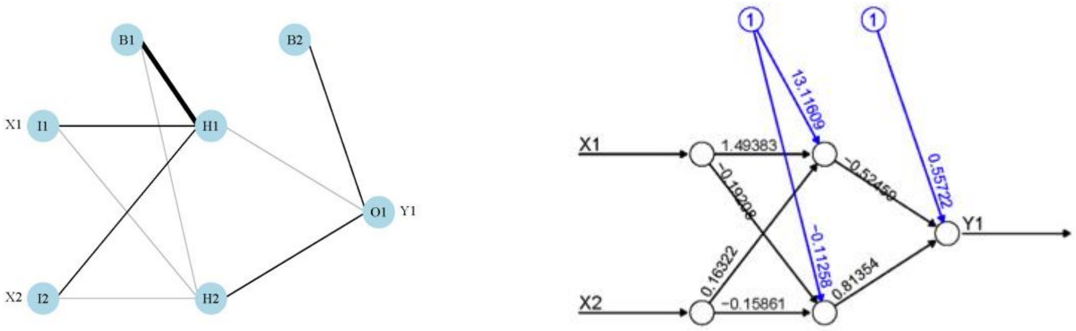


Figure 5.17: Example of a neural network with input variables X1 and X2, and output variable Y1 (on the left) also showing the weights (on the right)

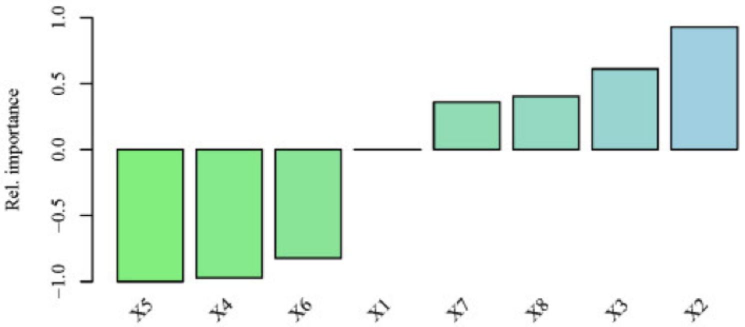


Figure 5.18: Relative importance of the eight explanatory variables for the response variable Y using the neural network created above

In harvest, there is a need for predictions and decision-making which can be of high risk and costly in short time. One of these is the problem of predicting the proportion of commercial average weight categories of fish in units that are going to be harvested. Usually, the fish farmers make these predictions based on their experience and with the contribution of static tables indicating the proportion of average harvest weight against the commercial categories of weights (such as W150.200, W200.300, W300.400, W400.600, W600.800, W800.1000, W1000, etc).

We improve the prediction of harvests' commercial weights using biological, geographical and temporal features of Life To Date data. A feed-forward neural network with one hidden layer is trained based on historical life to date harvesting data comes from an aquaculture company. As output neurons of the aNN we consider the instances of the commercial average weight category. The Feed-forward Neural Network with 30 hidden neurons has a better behaviour than other networks, with more or less hidden neurons, in terms of generalisation. It manages to predict correctly the percentage of commercial average weights categories in unknown instances. The blue bars in the presented test instances depict the training Neural Network response.

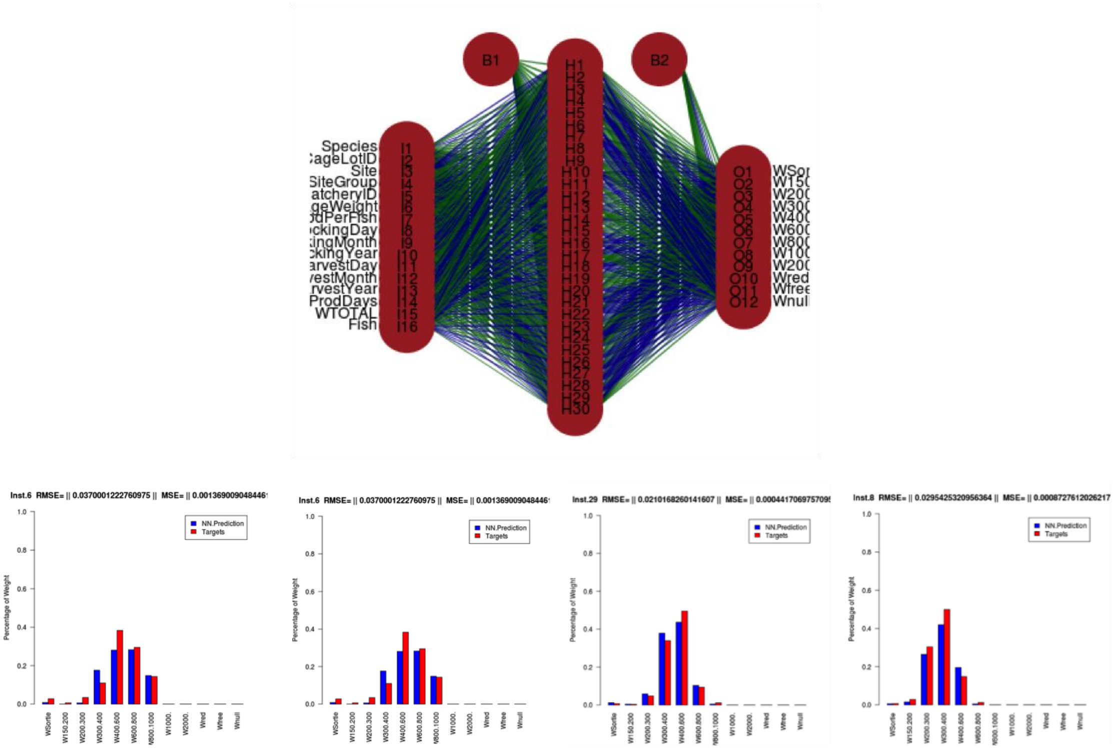


Figure 5.19: Neural network example to predict size distribution in harvesting, and the test instances.

We can then train a model having input variables of average weight, temperature, and perhaps mortality to predict LTD Econ FCR. The relative importance of the 16th input features is shown in the following figure.

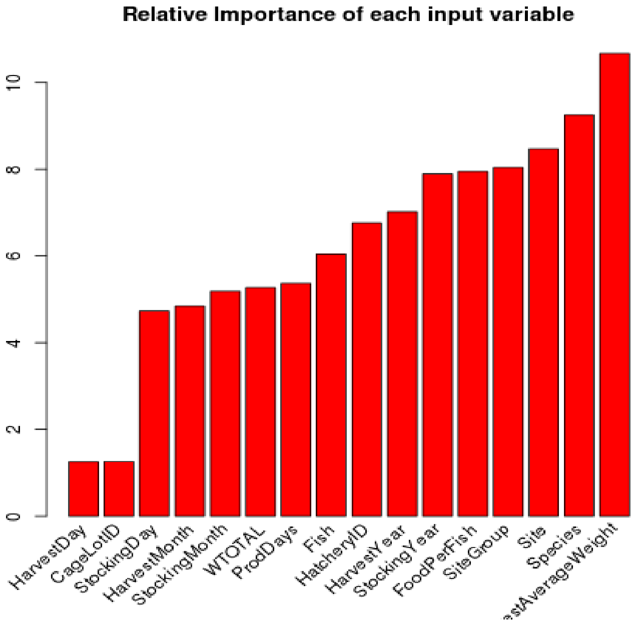
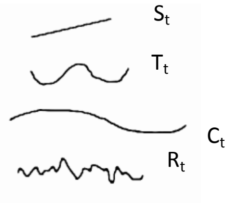


Figure 5.20: The relative importance of the 16th input features in the considered aNN.

The average weight of the harvest, the species of the fish, the region (site, site group), the food per fish quantity, the stocking and the harvest year seem to have major significant impact to the neural network training. Generally, the findings are promising when using aNNs to predict the proportions of commercial average weights.

We will now focus on the handling of time series data using neural networks. A times series can always be decomposed according to its behaviour describing the dynamics of the data. The components of a time series can be any combinations of the following:

- Trend [T_t] – a linear progression with either positive or negative slope;
- Seasonality [S_t] – a pattern repeated in time;
- Cycle [C_t] – a large pattern in time;
- Randomness [R_t] – a non-controlled behaviour.



In time, series analysis we identify the components that can be forecasted separately, and put them together in a comprehensive way to enable an overall forecast. The most common method to put this components together is the multiplicative model $X_t = T_t*S_t*C_t*R_t$. An important step in this analysis is the de-seasonalizing (i.e. the elimination of the seasonal aspect) of the time series. That is, if the time series represents a seasonal pattern of L periods, then by taking the moving average M_t of L periods, we get the mean value for the year. This would then be free of seasonality and contain little randomness. Thus, $M_t=T_t*C_t$. Now, to determine the trend, we take the de-seasonalized time series and use regression to fit a suitable trend line. The choices can be linear, quadratic, exponential, etc. After the trend T_t has been estimated, one can use $C_t = M_t / T_t$ to estimate the cycle component C_t . Finally, to isolate seasonality we simply divide the original series by the moving average

$$Xt \,/\, Mt \;=\; Tt * St * Ct * Rt \,/\, Tt * Ct = St * Rt$$

Averaging over the same month eliminates randomness and yields seasonality indices. The procedure for the forecast of a time series is then tied to its decomposition into components (identifying seasonal and trend components), and forecasting the future values of each component. The latter is done by projecting the trend component into the future and multiplying the trend component by the seasonal component.

The technological advances lead to the dominance of artificial neural networks (ANNs) for classification and forecasting of time series data, which is further discussed in Section 4.2.4 in the context of aquaculture. In general, the predictive power of ANNs for time-series data is based on the idea that we can use the past to predict the future.

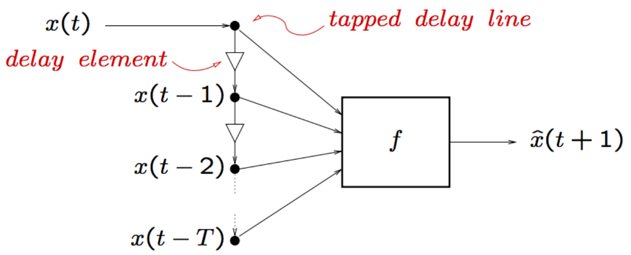


Figure 5.21: Schematic neural network forecasting of time series data

In detail, if we set up a shift register of delays, we can retain successive values of our time series. Then we can treat each past value as an additional spatial dimension in the input space to our predictor. The input space to our predictor must be finite. At each instant 't', truncate the history to only the previous d samples. Autoregressive Moving Average (ARMA) models are appropriate to the study of time series data influenced by cycles and seasons. It is possible to generalize ARMA using the ANN approach by considering the following diagram:

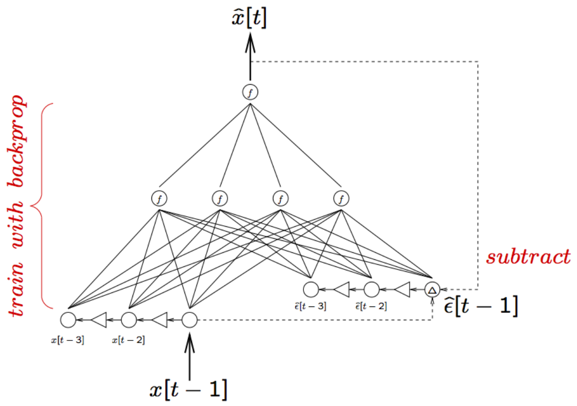


Figure 5.22: Schematic neural network extension of a ARMA model based on time series data

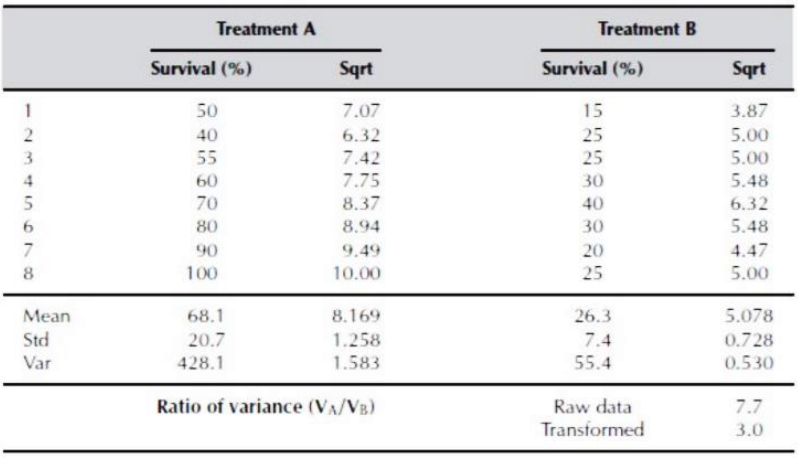
6 ANALYSIS OF RESULTS

6.1 Summary statistics

Aquasmart will use summary statistics in order to communicate the largest amount of information as simply as possible. We will rely on traditional methods to summarize a set of observations as described in detail in the following paragraphs.

In that, a large part of the procedures within aquaculture that can be modelled with the Aquasmart provide outputs that, for the sake of summary statistics, are themselves considered as observations. These observations are described in:

- a measure of location, or central tendency, such as the arithmetic mean;
- a measure of statistical dispersion like the standard deviation;
- a measure of the shape of the distribution like skewness or kurtosis;
- a measure of statistical dependence such as a correlation coefficient (when more than one variable is measured).



	Treatment A		Treatment B	
	Survival (%)	Sqrt	Survival (%)	Sqrt
1	50	7.07	15	3.87
2	40	6.32	25	5.00
3	55	7.42	25	5.00
4	60	7.75	30	5.48
5	70	8.37	40	6.32
6	80	8.94	30	5.48
7	90	9.49	20	4.47
8	100	10.00	25	5.00
Mean	68.1	8.169	26.3	5.078
Std	20.7	1.258	7.4	0.728
Var	428.1	1.583	55.4	0.530
	Ratio of variance (V_A/V_B)		Raw data	7.7
			Transformed	3.0

Table 6.1: Summary statistics in aquaculture

Summary statistics techniques can be adopted into Big Data analytics with the following objectives:

- Suggest hypotheses about the causes of observed phenomena in aquaculture data;
- Assess assumptions on which statistical inference will be based;
- Support the selection of appropriate statistical tools and techniques;
- Provide a basis for further data collection through surveys or experiments.

We will use the common collection of order statistics known as the five-number summary, extended to a seven-number summary, and the associated box plot:

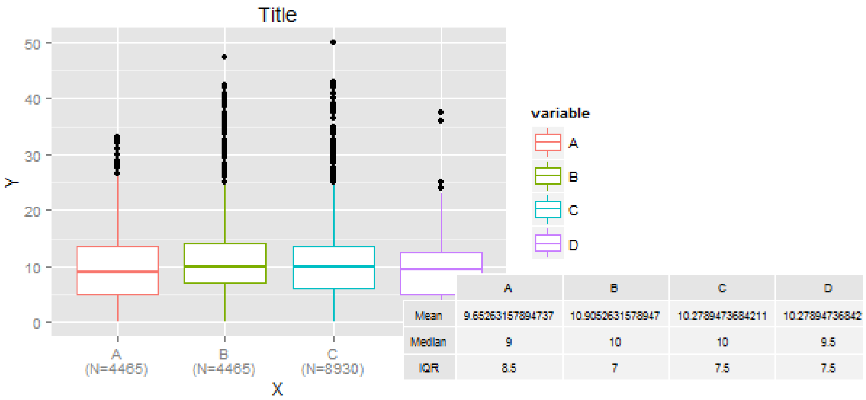
- **Location** - measures of location, or central tendency, as the arithmetic mean, median, mode, and interquartile mean.
- **Spread** - measures of statistical dispersion as the standard deviation, variance, range, interquartile range, absolute deviation and the distance standard deviation. Measures that assess spread in comparison to the typical size of data values include the coefficient of variation. A simple summary of a dataset can be given by quoting particular order statistics as approximations to selected percentiles of a distribution.
- **Shape** - Common measures of the shape of a distribution as skewness or kurtosis, while alternatives can be based on L-moments. A different measure is the distance skewness, for which a value of zero implies central symmetry.
- **Dependence** - measure of dependence between paired random variables as the Pearson product-moment correlation coefficient, or the Spearman's rank correlation coefficient. A value of zero for the distance correlation implies independence.

6.2 Representation of the results

An appropriate visualisation of the dataset in analysis is critical for the success of any business intelligence and, in particular, within aquaculture. It profits of the advantages provided by statistics to discover valuable information from the data and make it usable, relevant and actionable, clarifying communications.

Aquasmart will use exploratory data analysis (EDA) to summarize the main characteristics of the datasets in analysis, often with visual methods. Primarily EDA provides us with a preview of what the data can tell us beyond the formal modelling or hypothesis testing task. EDA is an approach for data analysis that employs a variety of data visualisation techniques to:

- maximize insight into a data set;
- uncover underlying structure;
- extract important variables;
- detect outliers and anomalies;
- test underlying assumptions;
- develop parsimonious models; and
- determine optimal factor settings.



	A	B	C	D
Mean	9.65263157894737	10.9052631578947	10.2789473684211	10.27894736842
Median	9	10	10	9.5
IQR	8.5	7	7.5	7.5

Figure 6.1: Exploratory data analysis techniques for data visualisation

The graphical techniques used to represent the main characteristics of our data sets are:

- **Box plot** - a convenient way of graphically depicting groups of numerical data through their quartiles (as in Figure 6.1).
- **Histogram** - a graphical representation of the distribution of numerical data estimating the probability distribution of a continuous (quantitative) variable.
- **Multi-vari chart** - a visual way of presenting variability through a series of charts.
- **Run chart** - a graph that displays observed data in a time sequence
- **Pareto chart** - a type of chart that contains both bars and a line graph, where individual values are represented in descending order by bars, and the cumulative total is represented by the line.
- **Scatter plot** - a type of diagram using Cartesian coordinates to display values for a set of data (eventually with points color-coded to increase the number of displayed variables).
- **Odds ratio** - one of three main ways to quantify how strongly the presence or absence of property A is associated with the presence or absence of property B in a given population
- **Multidimensional scaling** - a means of visualizing the level of similarity of individual cases of a dataset. It refers to a set of related ordination techniques used in information visualization, in particular to display the information contained in a distance matrix.
- **Principal component analysis** - a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.
- **Median polish** - finds an additively-fit model for data in a two-way layout table (usually, results from a factorial experiment) of the form row effect + column effect + overall median.
- **Trimean** - a measure of a probability distribution's location defined as a weighted average of the distribution's median and its two quartiles.
- **Gradient analysis** – orders objects that are characterized by values on multiple variables (i.e., multivariate objects) so that similar objects are near each other and dissimilar objects are further from each other.

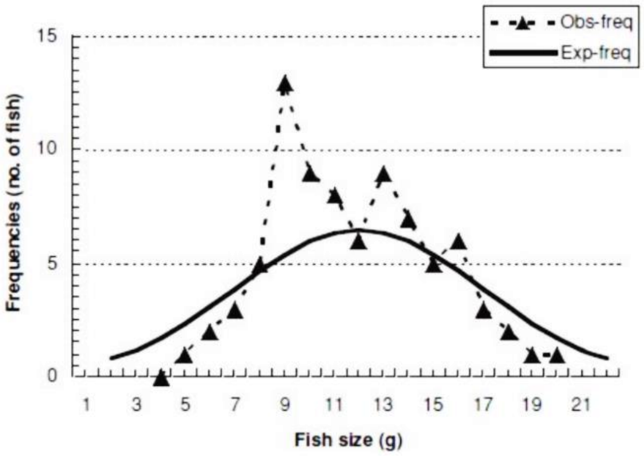


Figure 6.2: Exploratory data analysis in aquaculture

6.3 Evaluation of the results

For the evaluation of the results we will use a Monte Carlo approach measuring our models, including repeated random sampling, the fitting of Gaussian noise and cross-validation as we will further describe in the next paragraphs.

The evaluation of polynomial regression based models used in Aquasmart follows a procedure that permits a better understanding of the strengths of the model in analysis. In the following we will refer to the error of the model as its global distance towards the plot of the real data. In detail, the error (or disturbance) of an observed value is the deviation of the observed value from the (unobservable) real data value of the measured data (such as average weight or water temperature), and the residual of an observed value is the difference between the observed value and the estimated value of the quantity of interest (for example, a sample mean).

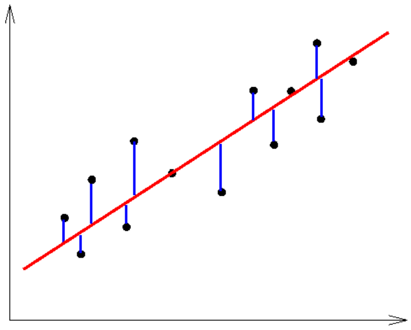


Figure 6.3: The error on the regression model as the global distance to the real data (in blue)

To analyse the model complexity against the error we shall use the elbow method. In that we compare the amount of error in the model and the model complexity. Increasing the model complexity leads to minimization of the error. The optimized model can be identified by the point of optimization known as elbow, as identified in Figure 6.5 by the point p.

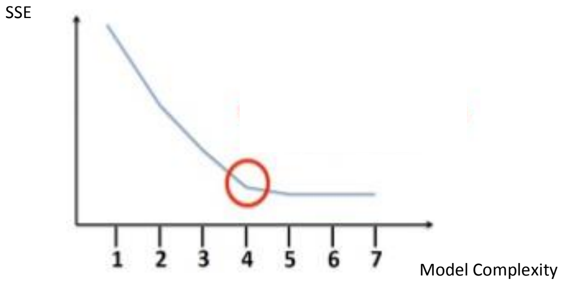


Figure 6.4: The elbow method for optimization of model complexity

The elbow method permits us to determine the optimal number of clusters in a dataset. The clusters here correspond to the levels of model complexity considered. The idea of the elbow method is to run our model on the dataset for a range of values of ’k’ (say, k from 1 to 7 in the examples above), and for each value of k calculate the sum of squared errors (SSE).

The best fit for such a model can be identified by the distribution of the error on it. A Gaussian distribution indicates that the regression approach is the most appropriate unbiased estimator.

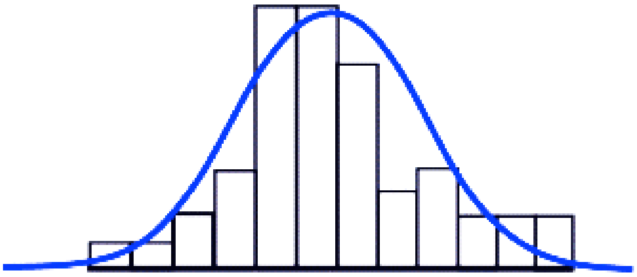


Figure 6.5: The Gaussian distribution of the error in the model

It is often difficult to obtain the variance of the error in the model in analysis. A procedure that often contributes in the evaluation of the model is to generate Gaussian noise and fit it to the model. Let us discuss that in detail. Consider the original model:

$$Y = aX2 + bX + c + \mu$$

for the estimation of the parameters a,b & c where µ is a random variable originated from a normal distribution with an associated Gaussian error. Now consider Gaussian noise, i.e., statistical noise having a probability density function equal to that of the normal distribution, (also known as the Gaussian distribution). Fitting that noise to the original model will provide a random variable ‘X’ = model + noise. We can then directly generate data from the new model and compare it. If the confidence interval contains most of the real data points we can then assure the validity of our model.

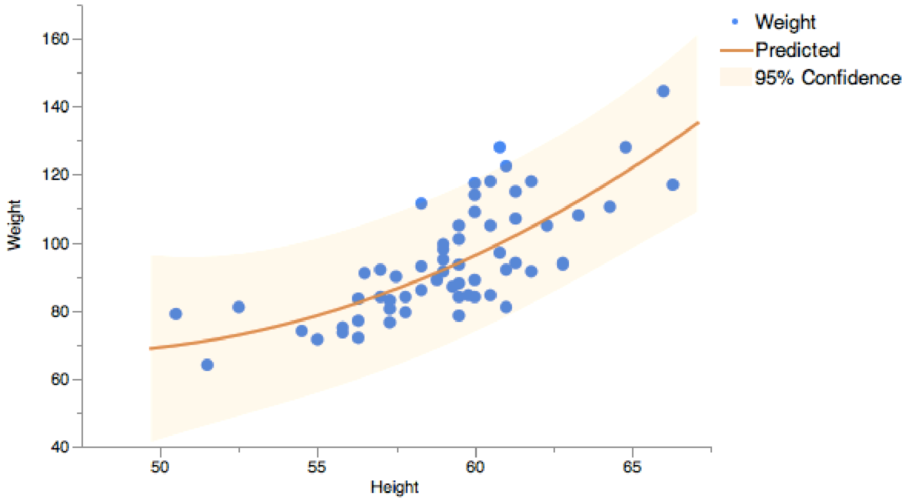


Figure 6.6: The confidence interval for experiments with height and weight measure as an example

The confidence intervals belong with the factors and not the response. Indeed, different factor values could produce the same predicted response value but with different confidence intervals. We can consider the factors categorically and plot each factor combination separately. In Figure 6.8, we present an example of that based on a prediction model for oxygen, rotated to make it easier to read the four factor values.

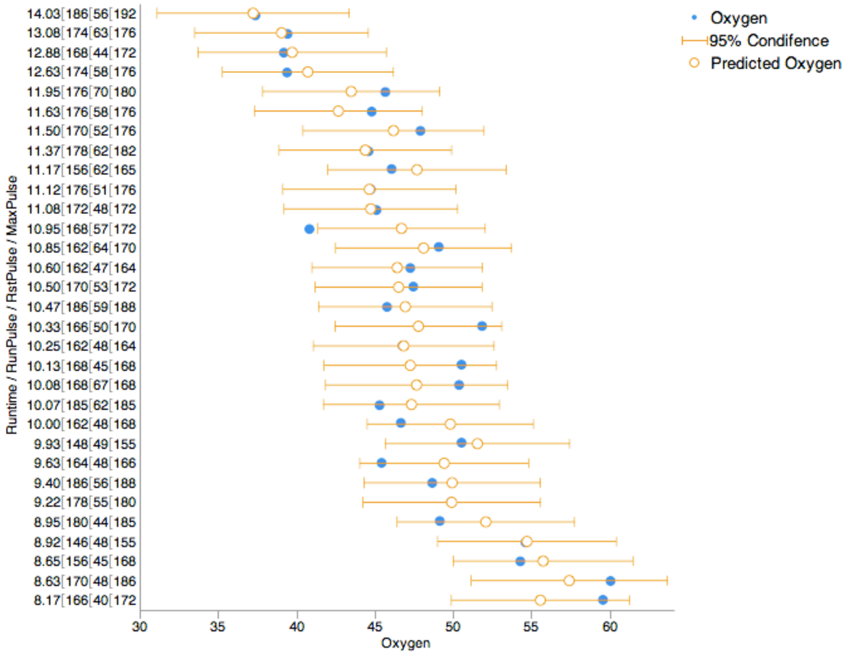


Figure 6.7: The confidence interval for oxygen level forecast as an example

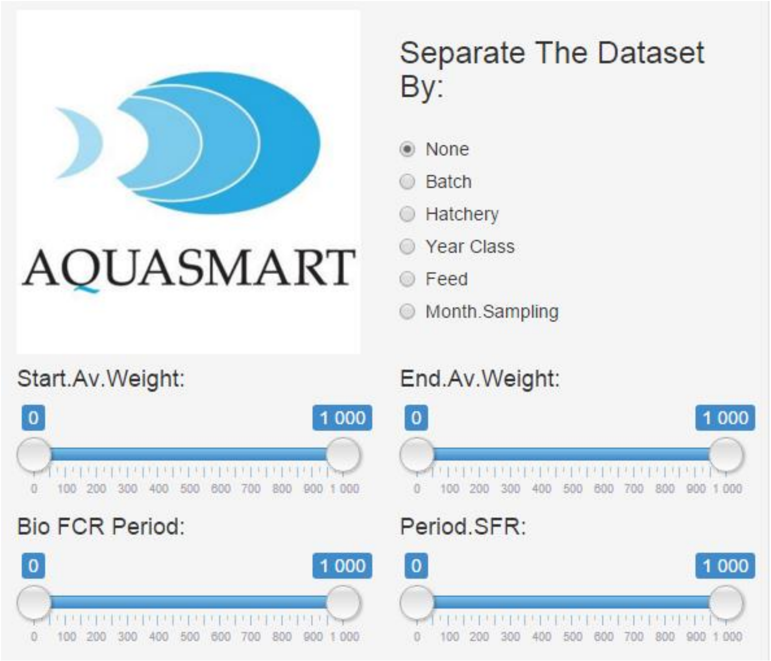
7 DATA VISUALIZATION

Aquasmart will provide the end user with appropriate visualisations for the outputs of the analytics. This section describes some of visualisation techniques that will be employed. Please note that the methods presented here are not exhaustive and further appropriate methods may be introduced as the project progresses. This will include a number of plots/graphs representing the average weight, average temperature and other deciding factors as well as representing the data produced when these values are taken into account.

Filtering options for the data will also be available to the users. These will be a precursor to the analysis and will determine the dataset that is used in the analysis.

7.1 Filtering

The user will be able to interactively filter their data in order to drive the queries that determine what data are fed into the analytics routines. An example is shown below



Separate The Dataset By:

- ● None
- ○ Batch
- ○ Hatchery
- ○ Year Class
- ○ Feed
- ○ Month.Sampling

Start.Av.Weight: 0 – 1 000

End.Av.Weight: 0 – 1 000

Bio FCR Period: 0 – 1 000

Period.SFR: 0 – 1 000

Figure 7.1: Data filters in the Aquasmart GUI

The available filters will be:

- Radio buttons
- Sliders – specify min/max values for a given attribute
- Constrained lists – modals (e.g. select a given year, batch etc.)
- Date ranges – start/end dates

The values for the filters (e.g. max/min values, list contents etc.) will be driven by the actual user data.

7.2 Visualisation

Visualisation will be rendered on the client – not the server, thus, the server will simply return a set of results in JSON format and these will be rendered locally. This has the following advantages:

- More efficient – typically the raw data will be much more compact than an image rendered remotely;
- Responsive – the visualisation can adapt to the platform on which it is being viewed in terms of resolution;
- Better looking – modern browsers supporting HTML5 have sophisticated graphic rendering capabilities via WebGL, canvases etc.;
- Possibilities for interactivity – when the raw data for a visualisation is available interactive functionality can be given to the end users (e.g. the ability to rotate 3D plots as discussed below).

The rest of this section explores some of the techniques which may be used to explore the results.

7.2.1 Tabular

The simplest visualisation is a table, though this is not to say that this is not effective:

Samplings	Econ. FCR Period	Model Econ FCR Period	% Deviation
1	0,31	1,77	**82,67**
2	2,69	2,32	**-16,17**
3	3,66	3,35	**-9,28**
4	2,89	2,56	**-13,08**
5	3,25	2,56	**-27,14**
6	1,71	1,85	**7,58**
7	2,22	2,32	**4,30**
8	7,94	2,83	**-180,74**
9	3,33	2,50	**-33,08**
10	1,09	1,88	**42,18**
11	1,30	1,20	**-8,35**
12	0,95	0,92	**-2,23**
13	1,55	2,73	**43,43**
14	1,18	1,24	**4,91**
15	4,00	3,47	**-15,26**

Table 7.1: Tabular visualisation for the economic FCR period and the model FCR period in the sample to sample data

The example shown above compares the Economic FCR Period of a selection of samplings to the model and displays the deviations. In this example, deviations of 10% or less are acceptable and colour coded blue, deviations greater than this are highlighted in red. The acceptable deviation can be user controlled so the table (above) can update dynamically without necessitating a round trip to the server.

7.2.2 Pie Chart

To follow from the previous example, the overall percentage of samplings that fall within the acceptable deviation can be summarised in a pie chart. Again, this can be made dynamically updatable from a user driven deviation threshold.

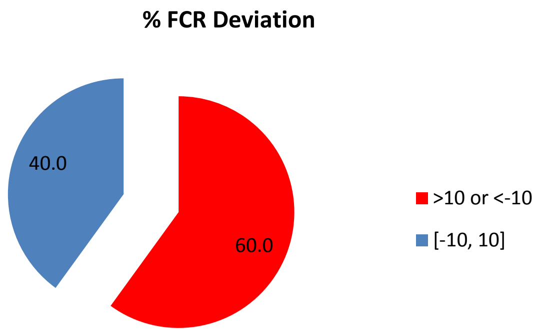


Figure 7.2: Pie chart showing the FCR deviation in percentage

7.2.3 Bar chart

This example shows the samplings and their deviations plotted as a bar chart, again with red/blue colour coding:

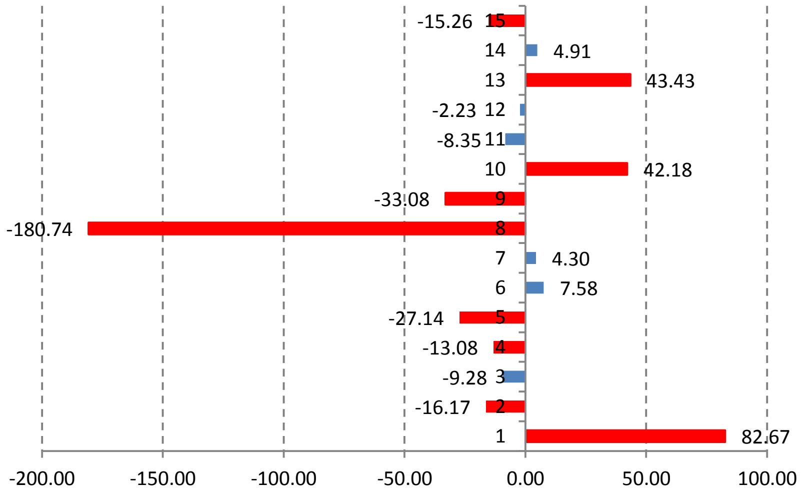


Figure 7.3: Bar chart describing the samplings and their deviations

7.2.4 Scatter Plots

This example shows a scatter plot of SFR-FCR grouped by % Deviation:

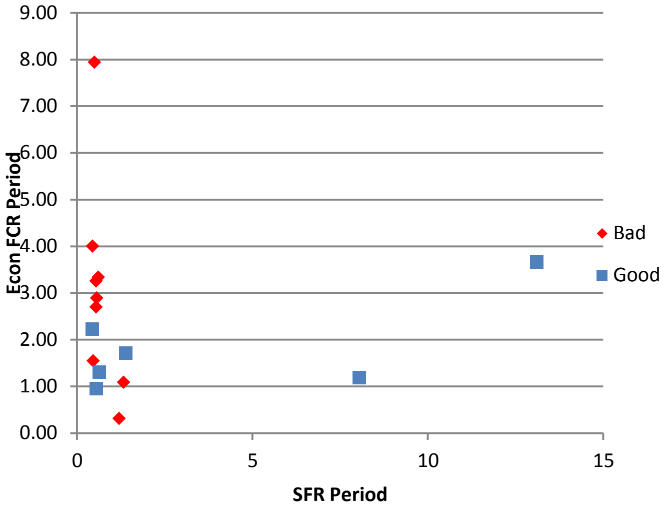


Figure 7.4: Scatter plot comparing the SFR with the FCR grouped by deviation in percentage

7.2.5 Box Plots

Basic statistics measures such as mean, median, standard deviation, etc. of FCR, SFR, Avg. Temperature grouped by % Deviation (one group consists of the samplings where absolute % Deviation is above 10% and the rest of the samplings belong to the second group) can be visualised as boxplots of two groups on FCR, SFR, Avg. Temperature measures.

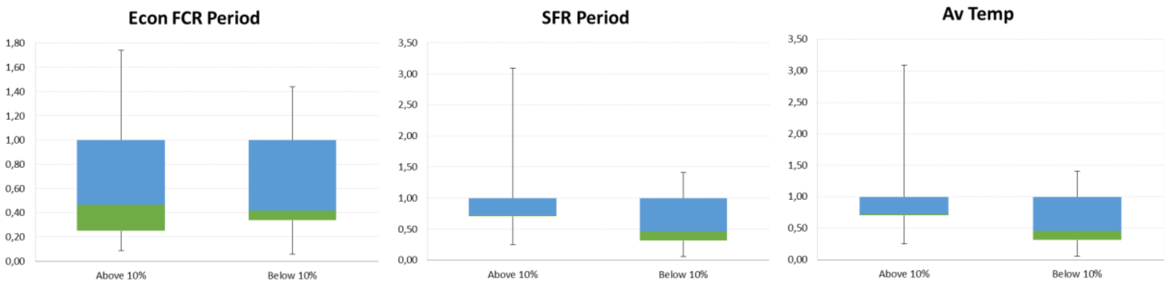


Figure 7.5: Box plots describing the economic FCR period (on the left), the SFR period (in the center) and the average temperature (on the right) for the sample to sample data

7.2.6 3D Plots

For more complex data, 3D plots can be used. The examples below illustrate visualisations for the interpolated economical FCR model depending on the weight and temperature.

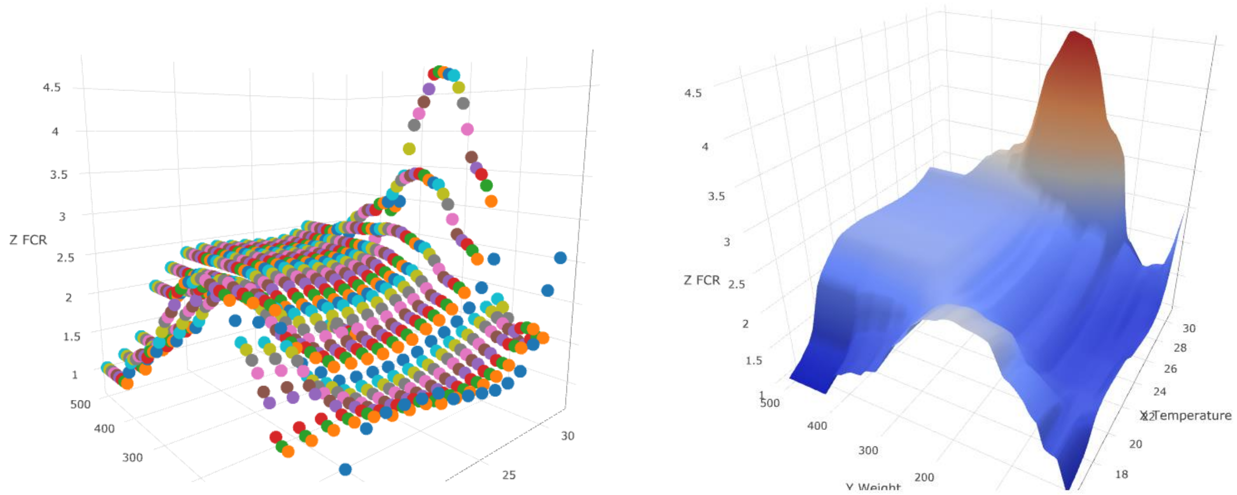


Figure 7.6: 3D scatter plot (on the left) and 3D surface plot (on the right) for the FCR table dependent on the average weight and water temperature

In both cases, the end user can rotate, pan and zoom on the plot as necessary to inspect the data more closely and by hovering the mouse over any point on the plots will give them a numerical summary of the data represented by that point.

8 CONCLUSIONS

The challenges of aquaculture for data analytics are very specific and must be addressed with the appropriate methodologies and technology, in line with the expertise of the fish farmers. The uncertainty of measurements, such as the number of fish until the time of harvest, derives in variances that do not permit a complete accuracy of some of the calculations. This is particularly important to some of the available tools to monitor the business, such as the feed conversion rate tables in use by the fish farmers to optimize the production costs. The mathematical models developed in this project and discussed in this deliverable aim to contribute to the improvement of the aquaculture procedures, providing a deeper insight on the information retained in the collected data, using state-of-the-art methods of data mining in line with the expert knowledge of the field transferred to the metadata in the data store. In that way, we are able to have a better insight over the collected data and potentially to identify new patterns in that data that can be disseminated through the community. Moreover, the statistical analysis of the results permits a clearer visualisation of the important features in the data that can boost the production and optimize the processes related to it. This will permit an important contribution to the open data cloud platform developed in WP2 with the development and incorporation of a machine learning component enabling classification and forecast based on the analytics of the available data.

9 REFERENCES

[AK15] Alagappan, M., & Kumaran, M. (2015). Expert system for shrimp aquaculture–an ICT aided tool for knowledg management. Indian Journal of Fisheries.
[AN14] Allahyari, M. S., & Noorhosseini, S. A. (2014). Agro-Economic Factors Determining on Adoption of Rice-Fish Farming: An Application for Artificial Neural Networks. Journal of Advanced Agricultural Technologies, 1(2), 1–6.
[AFAD11] Atia, D. M., Fahmy, F. H., Ahmed, N. M., & Dorrah, H. T. (2011). Solar Thermal Aquaculture System Controller Based on Artificial Neural Network. Engineering, 03(08), 815–822.
[AFAD] Atia, D. M., Fahmy, F. H., Ahmed, N. M., & Dorrah, H. T. (n.d.). Mathematical Modeling and Neural Network Control for Dissolved Oxygen of Aquaculture Pond Aeration System. The Online Journal on Electronics and Electrical Engineering 4(2).
[ASL15] Atoum, Y., Srivastava, S., & Liu, X. (2015). Automatic Feeding Control for Dense Aquaculture Fish Tanks. IEEE Signal Processing Letters, 22(8), 1089–1093.
[aqu16] AQUATEXT Dictionary http://www.aquatext.com/dicframe.htm (accessed in 28.1.2016).
[BMSSA07] Balakrishnan, M and Meena, K and Sethi, S N and Sarangi, Aditya N (2007) Neural network and its application in aquaculture. Bioinformatics and statistics in Fisheries Research, 3. pp. 145-151.
[BR09] Bar, N. S., & Radde, N. (2009). Long-term prediction of fish growth under varying ambient temperature using a multiscale dynamic model. BMC Systems Biology, 3(1), 1.
[B14] Barbedo, J. (2014). Computer-Aided Disease Diagnosis in Aquaculture: Current State and Perspectives for the Future. Revista Innover.
[BGMC13] Bartók, A. P., Gillan, M. J., Manby, F. R., & Csányi, G. (2013). Machine-learning approach for one- and two-body corrections to density functional theory: Applications to molecular and condensed water. Physical Review B, 88(5), 054104.
[BESM00] T. Bellemans, B. De Schutter, and B. De Moor (2000) Data acquisition, interfacing and pre-processing of highway traffic data. Proceedings of Telematics ..., vol. 19, pp. 0–2.
[BB14] Bengil, F., & Bizsel, K. C. (2014). Assessing the impact of aquaculture farms using remote sensing: an empirical neural network algorithm for Ildırı Bay, Turkey. Aquaculture Environment Interactions, 6(1), 67–79.
[BB15] Benzer, R., & Benzer, S. (2015). Application of artificial neural network into the freshwater fish caught in Turkey. International Journal of Fisheries and Aquatic Studies, 2(5), 341–346.
[Bhu11] Bhujel, R. C. (2011). Statistics for Aquaculture. John Wiley & Sons.
[BJLF93] Bjordal A, Juell JE, Lindem T, A. Ferno A (1993); Hydroacoustic monitoring and feeding control in cage rearing of Atlantic salmon, Fish Farming Technology, pp. 203-208, Balkema, Rotterdam.
[BPR92] Blyth PJ, Purser, GJ & Russell JF (1992) Boosting profits with adaptive feeding: letting the fish decide when they're hungry. AustAsia Aquaculture, Vol. 6 pp 33-38.
[BNE00] Bolte John, Nath Shree, and Ernst Doug (2000). Development of decision support tools for aquaculture: the pond experience. Aquaculture Engineering, 23:103–119.
[BPS15] Bermingham, Mairead L.; Pong-Wong, Ricardo; Spiliopoulou, Athina; Hayward, Caroline; Rudan, Igor; Campbell, Harry; Wright, Alan F.; Wilson, James F.; Agakov, Felix; Navarro, Pau; Haley, Chris S. (2015). Application of high-dimensional feature selection: evaluation for genomic prediction in man. Sci. Rep. 5.
[B04] Beveridge M (2004). Cage Aquaculture. Third Edition, Oxford, UK.
[BREI01] L. Breiman (2001) Random forests. Machine learning, pp. 5–32.
[CP96] Steven D. Culberson and Raul H. Piedrahita (1996). Aquaculture pond ecosystem model: temperature and dissolved oxygen prediction—mechanism and application. Ecological modelling 89.1: 231-258.

[DM11] Ding, W., & Ma, Y. (2011). The Application of Wireless Sensor in Aquaculture Water Quality Monitoring. In Computer and Computing Technologies in Agriculture V. Vol. 370, pp. 502–507. Berlin, Heidelberg: Springer Berlin Heidelberg.
[DEB] Ernst, D. H., Bolte, J. P., & Nath, S. S. (2000). AquaFarm: simulation and decision support for aquaculture facility design and management planning. Aquacultural Engineering, 23(1), 121-179.
[DRG06] Drake, J. M., Randin, C., & Guisan, A. (2006). Modelling ecological niches with support vector machines. Journal of Applied Ecology, 43(3), 424–432.
[DRSM98] N. R. Draper and H. Smith (1998). Applied regression analysis, 3th editio. New York: Wiley.
[DK01] Durbin, J., and Koopman, S. J. (2001). Time Series Analysis by State Space Methods. Oxford University Press, Oxford, United Kingdom.
[ECAq] European Comission, Aquaculture. http://ec.europa.eu/fisheries/cfp/aquaculture/index_en.htm (accessed in 28.1.2016)
[FDL07] D. Fabbian, R. De Dear, and S. Lellyett (2007). Application or neural network forecasts to predict fog at Canberra International Airport," Weather and Forecasting, vol. 22, no. 2, pp. 372-381.
[FTPBPE11] Farmaki, E., Thomaidis, N. S., Passias, I. N., Baulard, C., Papaharisis, L., & Efstathiou, C. E. (2011). Artificial neural networks for the diagnosis of the aquaculture impact on marine sediments. Proceedings of the 12th International Conference on Environmental Science and Technology, pp. 531–539.
[FAQ16] Food and Agriculture Organisation of the United Nations. The State of World Fisheries and Aquaculture 2014. http://www.fao.org/fishery/sofia/en (accessed in 28.1.2016)
[GR12] John Gantz and David Reinsel (2012). The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East." IDC, for EMC Corporation, December.
[GS11] Garaas, M., & Stevning, G. H. (2011). Case-Based Reasoning in identifying causes of fish death in industrial fish farming.
[GCSR04] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). Bayesian Data Analysis, 2nd Edition. Chapman & Hall/CRC, Boca Raton, Florida.
[glo16] Fish Farming Glossary http://www.aquaculture.co.il/getting_started/glossary.html (accessed in 28.1.2016).
[GE03] Guyon, Isabelle; Elisseeff, André (2003). An Introduction to Variable and Feature Selection". JMLR 3.
[GCC12] Guo-Xun Yuan; Chia-Hua Ho; Chih-Jen Lin (2012). Recent Advances of Large-Scale Linear Classification. Proc. IEEE 100 (9).
[HBD99] Hall, T., Brooks, H. E., & Doswell, C. A. I. (1999). Precipitation Forecasting Using a Neural Network. Weather and Forecasting, 14(3), 338–345.
[HKP11] Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques. Elsevier.
[DB08] Dobson, A. J., & Barnett, A. (2008). An Introduction to Generalized Linear Models, Third Edition. CRC Press.
[HH89] Houvenaghel, T., & Huet, T. (1989). A model for eel growth in aquaculture. European Aquaculture Society.
[GWHT13] Gareth James; Daniela Witten; Trevor Hastie; Robert Tibshirani (2013). An Introduction to Statistical Learning. Springer. p. 204.
[JW92] Johnson, Richard Arnold, and Dean W. Wichern (1992). Applied multivariate statistical analysis. Vol. 4. Englewood Cliffs, NJ: Prentice hall.
[KMHT91] Kadri, S., Metcalfe, N.B. Huntingford, F.A. & Thorpe J.E. (1991) Daily feeding rhythms in Atlantic salmon in sea cages. Aquaculture, vol. 92, pp. 219 – 24.
[Kat06] Katsanevakis, S. (2006). Modelling fish growth: Model selection, multi-model inference and model selection uncertainty. Fisheries Research, 81(2-3), 229–235.
[Lee00] Lee P.G. 2000. Process control and artificial intelligence software for aquaculture. Aquacultural Engineering, 23, 13-36.
[LRU14] Leskovec, Jure, Anand Rajaraman, and Jeffrey David Ullman (2014). Mining of massive datasets. Cambridge University Press.
[LFD02] Li, D., Fu, Z., & Duan, Y. (2002). Fish-Expert: a web-based expert system for fish disease diagnosis. Expert Systems with Applications, 23(3), 311–320.

[LYTXL11] Liu, S., Yan, M., Tai, H., Xu, L., & Li, D. (2011). Prediction of Dissolved Oxygen Content in Aquaculture of Hyriopsis Cumingii Using Elman Neural Network. In Computer and Computing Technologies in Agriculture V (Vol. 370, pp. 508–518). Berlin, Heidelberg: Springer Berlin Heidelberg.
[LASHGQ11] Løland, A., Aldrin, M., Steinbakk, G. H., Huseby, R. B., Grøttum, J. A., & Quinn, T. J. (2011). Prediction of biomass in Norwegian fish farms. Canadian Journal of Fisheries and Aquatic Sciences, 68(8)
[MZZ11] Man, Mustafa, et al. FishDTecTools: Fish Detection Solution Using Neural Network Approach. Journal of Communication and Computer 8.2 (2011): 96-102.
[MS96] C. Marzban and G.J. Stumpf (1996). A neural network for tornado prediction based on Doppler radar-derived attributes. Journal of Applied Meteorology, vol. 35, no. 5, pp. 617-626.
[MKA04] I. Maqsood, M.R. Khan, A. Abraham (2004). An ensemble of neural networks for weather forecasting, Neural Comput & Applic 13, 112-122.

[MC13] Mayer-Schönberger, V., & Cukier, K. (2013). Big data: a revolution that will transform how we live, work and think. London: John Murray.
[MKT08] Moghim, M., Kor, D., & Tavakolieshkalak, M. (2008). The effects of feeding frequency on FCR and SGR factors of the fry of rainbow trout, Oncorhynchus mykiss. Back to Nature.
[MF04] Mosig, J., & Fallu, R. (2004). Australian Fish Farmer. Landlinks Press.
[NumPy] NumPy Developers, NumPy. Available: http://www.numpy.org/. (accessed 8.1.2016).
[PVGT11] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. The Journal of Machine Learning Research, 12, 2825-2830.
[PBR85] Phillips, M.J., Beveridge, M.C.M & Ross L.G. (1985). The environmental impact of salmonid cage culture on inland fisheries: present status and future trend. Journal of fish Biology, Vol. 27 pp 123 – 37.
[PDCJ11] Prista, N., Diawara, N., Costa, M. J., & Jones, C. (2011). Use of SARIMA models to assess data-poor fisheries: a case study with a sciaenid fishery off Portugal. Fishery Bulletin.
[PFP07] Pullin, Roger SV, Rainer Froese, and Daniel Pauly (2007). Indicators for the sustainability of aquaculture. Ecological and Genetic Implications of Aquaculture Activities. Springer Netherlands. 53-72.
[RDT13] A. Rahman, Claire D' Este, and G. Timms (2013). Dealing with Missing Sensor Values in Predicting Shellfish Farm Closure. Proceedings IEEE Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), pp. 351–356, Melbourne.
[RM04] A. Rahman and M. Murshed (2004). Feature weighting methods for abstract features applicable to motion based video indexing. IEEE International Conference on Information Technology: Coding and Computing (ITCC), vol. 1, pp.676–680.
[RM09] A. Rahman and M. Murshed (2009). Feature Weighting and Retrieval Methods for Dynamic Texture Motion Features. International Journal of Computational Intelligence Systems, vol. 2, no. 1, pp. 27–38.
[RST14] A. Rahman, D. Smith, and G. Timms (2014). A Novel Machine Learning Approach towards Quality Assessment of Sensor Data. IEEE Sensors Journal, 14.4:1035 - 1047.
[RST13] A. Rahman, D. Smith, and G. Timms (2013). Multiple Classifier System for Automated Quality Assessment of Marine Sensor Data. Proceedings IEEE Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), pp. 362–367.
[REC13] Rahman, A., D'Este, C., & McCulloch, J. (2013). Ensemble Feature Ranking for Shellfish Farm Closure Cause Identification (pp. 13–18). Presented at the Workshop, New York, New York, USA: ACM Press. arXiv preprint arXiv:1405.1304
[RS96] Rizzo, G., and Spagnolo, M. 1996. A Model for the Optimal Management of Sea Bass Dicentrarchus Labrax Aquaculture. Mar. Resour. Econ. 11: 267–286.
[SciPy] SciPy Developers, SciPy. Available: http://www.scipy.org/. (accessed 8.1.2016).
[S97] G. Schulstad (1997). Design of a computerized decision support system for hatchery production management. Aquaculture Engineering, 16:7–25.

[She00] Shearer C.(2000). The CRISP-DM model: the new blueprint for data mining, J Data Warehousing, 5:13—22.
[Sti07] Stickney, R. R. (2007). Aquaculture: An Introductory. (C. Publishing, Ed.). CABI Publishing.
[SMR12] Snijders, C.; Matzat, U.; Reips, U.-D. (2012). 'Big Data': Big gaps of knowledge in the field of Internet. International Journal of Internet Science 7: 1–5.
[SC14] Sumon Shahriar and John McCulluch (2014). A Dynamic Data-driven Decision Support for Aquaculture Farm Closure. ICCS 2014, Procedia Computer Science, Volume 29, Pages 1236–1245.
[TBAS11] A. Tidemann, F. O. Bjørnson, A. Aamodt and Sintef Fisheries (2011). Case-Based Reasoning in a System Architecture for Intelligent Fish Farming. Scai.
[TBA12] Tidemann, A., Bjørnson, F. O., & Aamodt, A. (2012). Operational Support in Fish Farming through Case-Based Reasoning. Iea/Aie, 7345(Chapter 12), 104–113.
[TH92] Thorpe JE & Huntingford FA (1992); The importance of Feeding Behaviour for the Efficient Culture of Salmonoid Fishes. World Aquaculture Society, Baton Rouge, LA.
[US11] Government, U. S. (2011). Estimating Water Temperatures in Small Streams in Western Oregon Using Neural Network Models. Books LLC.
[VIDS05] S. Vijayakumar, A. D'Souza, and S. Schaal (2005). Incremental online learning in high dimensions. Neural computation, vol. 17, no. 12, pp. 2602–34.
[WL09] Guirong Wang and Daoliang Li. A Fish Disease Diagnosis Expert System Using Short Message Service (2009).
[WR15] Works, Karen, and Elke A. Rundensteiner (2015). Practical Identification of Dynamic Precedence Criteria to Produce Critical Results from Big Data Streams. Big Data Research 2.4: 127-144.
[Z11] Matei Zaharia (2011). Spark: In-Memory Cluster Computing for Iterative and Interactive Applications. Invited Talk at NIPS 2011 Big Learning Workshop: Algorithms, Systems, and Tools for Learning at Scale.
[Z08] Dunn, Zelda (2008). Improved feed utilisation in cage aquaculture by use of machine vision. Diss. Stellenbosch: Stellenbosch University.
[ZRD13] Q. Zhang, A. Rahman, and C. D'Este (2013). Impute vs. Ignore: Missing Values for Prediction. Proc. IEEE International Joint Conference on Neural Networks (IJCNN), pp. 2193–2200.

[ZE11] Zikopoulos, Paul, and Chris Eaton (2011). Understanding big data: Analytics for enterprise class hadoop and streaming data. McGraw-Hill Osborne Media.

10 APPENDICES

Deliberately left blank