DATA ANALYTICS IN AQUACULTURE

Joao Pita Costa and Matjaž Rihtar
Artificial Intelligence Laboratory
Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
joao.pitacosta@ijs.si, matjaz.rihtar@ijs.si

ABSTRACT

The specific challenges in aquaculture today reveal needs and problems that must be addressed appropriately and in sync with the most recent optimization methods. It is now the time to bring the techniques of aquaculture to a new level of development and understanding. In that, one must consider the state of the art methods of statistics and data mining that permit a deeper insight 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. In this paper we review the state of the art of 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. We also address the data unknowns and strategies developed that will contribute to the success of the business, leading to discover valuable information from the data that can be made usable, relevant and actionable.

Categories and Subject Descriptors

E.3 Data Structures; I.2 Artificial Intelligence; I.6 Simulation and Modelling

General Terms

Algorithms, Data Science, Aquaculture

Keywords

Aquaculture, data analytics and visualization,

1. INTRODUCTION

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 [2]. These are exactly the kinds of practical measurements that will allow commercial aquaculture facilities to optimize their efficiency by reducing labor and utility costs. Anticipated benefits for aquaculture process control and artificial intelligence systems are: increased process efficiency; reduced energy and water losses; reduced labor costs; reduced stress and disease; improved accounting; improved understanding of the process.

The technologies and implementation of the technologies necessary for the development of computer intelligent management systems come in a wide variety [8] and enhanced commercial aquaculture production [3]. Today's artificial intelligence (AI) systems offer the aquaculturist a proven methodology for implementing management systems that are both intuitive and inferential. 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.

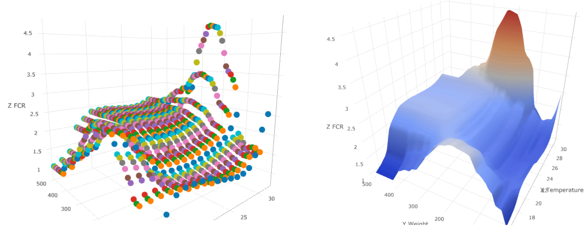

Figure 1. Dynamical plots developed for the project aquaSmart, available through a public interface where the fish farmers can upload their data and do a preliminary analysis and visualization.

The models presented in this paper were developed in the context of the EU project aquaSmart [1]. This project aims enhancing the innovation capacity within the aquaculture sector, by helping companies to transform captured data into knowledge and use this knowledge to dramatically improve performance. In particular, the tools constructed in that context (illustrated in Figure 1) serve the aquafarmers to evaluate feed performance, considering important factors such as the water temperature and average fish weight, but also underlying factors such as the oxygen level.

2. UNIQUE CHALLENGES

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 [9]. 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 fish until the end of the production is important for the amount of food given and, consequently, for the resources spent.

Feed composition has also a large impact on the growth of animals, particularly marine fish. Quantitative dynamic models exist to predict the growth and body composition of marine fish for a given feed composition over a timespan of several months [7]. 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 modeling marine organism growth.

3. UNKNOWNS IN THE DATA

It is curious that the underlying problems with the data unknowns in aquaculture represent a problem of large dimensions for the industry of aquaculture, in which the production is straightforward. In fact, it is not known at any time in production, the exact number of fish in production, and therefore it is not possible to calculate with exactness the amount of food needed to support an appropriate growth. Furthermore, there are many conditionings in the progress of the production that must be taken into account and are hard to measure with the existing and available technology. In that, it is important to describe some of the features of the data including an assessment on its quality and measures to overcome obstacles to the analysis.

The input and output variables of the dataset are classified as: numerical and categorical. Numerical variables can be: continuous measured quantities expressed as a float (e.g. 'av. weight'); discrete quantities expressed as an integer (e.g. 'number of fish'). Categorical variables can be: regular categorical data including non-ordered classes (e.g. species Bream/Bass); or 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: (i) variables that do not change over time, often identifying population attributes (e.g. identifications such as 'year' or 'hatchery'); (ii) variables that can change over time but do not change within a sampling period (e.g. 'batch'); (iii) variables that change daily, taken into account when samplings occur (e.g. 'average weight').

Table 1. Classification of values according to time dependence.

change in time?	**direct**	**calculated**	**derived**
yes	water temp.	FCR, SFR	av. weight
no	identification	av. weight at t=0	hatchery

Essentially we have four types of input data according to the impact they assure: (1) identification data that permits the fish farmer to manage the production and correctly identify the fish; (2) Daily 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, collected at predetermined points of the fish growth timeline, to confirm the model values and make the appropriate adjustments; (4) Life To Date (LTD) 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 distinguished between the group of production indicating localization - *Unit* - and the individual production series of fish - *Batch*. 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.

The daily data in input (2) is recorded by the aquafarmers on a daily basis. 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.

The sampling data in input (3) serves the aquafarmer to improve/fix his/her initial Feed Conversion Ratio (FCR) model with real data. This includes features that can be learned by a specific set of data. Those features will later be important for the algorithms. They often correspond to columns with potential effect on the end result. Also, they can influence the production (e.g. 'feeder'). The software will adapt to 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. We cannot predict the relevance of the data on those columns (neither their nature) but will consider them in the overall global analytics.

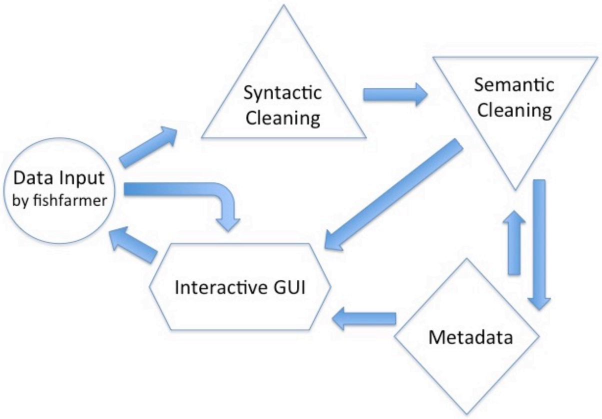

Figure 2. The proposed data cleaning process for aquaculture data, including the update of the metadata in the system and user interaction.

The daily data, the sampling data and the LTD data in inputs 2, 3 and 4 fall into three categories: (i) Direct 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); (ii) Calculated values, that are dependent of a number of other observed values (e.g. LTD values calculated from the daily data); (iii) Derived values – values deriving from previously available data (e.g. FCR calculated from the table, given average weight and water temperature).

The original data provided by the aquafarmers has variances/holes and is not precise because it is not measured automatically but instead entered by human hand (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 inputted once again after appropriate corrections. 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 consider the average from the missing data in the sense of interpolation on a fixed mesh grid.

4. DATA ANALYTICS IN AQUACULTURE

Mathematical modeling 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 fulfillment of the two more traditional disciplines, which are theoretical analysis and experimentation [4]. Nowadays, mathematical modeling has a key role also in aquaculture. In the following section we will present an overview of that. Growth and reproductive modeling 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 [6].

The 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 assigned to the food given to the fish during growth. Some of it will fall through the net and some will be spared. The optimization of the feeding of the fish can carry great benefits to the economic development of the fish farms.

Specifically, the FCR permits the aquafarmer to determine how efficiently a fish is converting feed into new tissue, defined as growth [10]. Recall that the FCR is a ratio that does not have any units provided by the formula:

$$\text{FCR} = \text{dry weight of feed consumed/wet weight of gain}$$

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

$$\text{FCE} = 1/\text{FCR} \times 100$$

There seems to be some controversy among aquatic animal nutritionists as to which is the proper parameter to measure, but in aquaSmart we used FCR (exposing here FCE for completion). Moreover, 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.

The FCR 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 FCR 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 models in this paper consider only temperature and average weight.

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, he needs more food to increase his biomass in percentage, and thus the FCR grows higher with the increase of the average weight. The quality of the food and the size of the pellet size 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 fish farms in AquaSmart. It includes 3 fish farms.

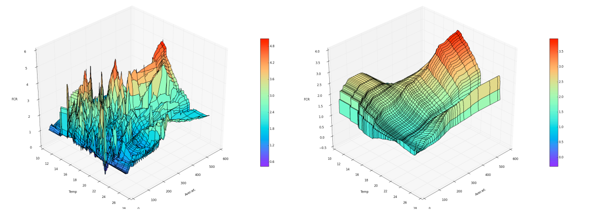

Figure 3. Company A: Real data (on the left) and FCR model (on the right) for the bream production.

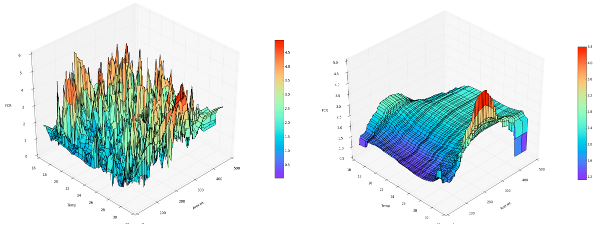

Figure 4. Company B: Real data (on the left) and FCR model (on the right) for the bream production.

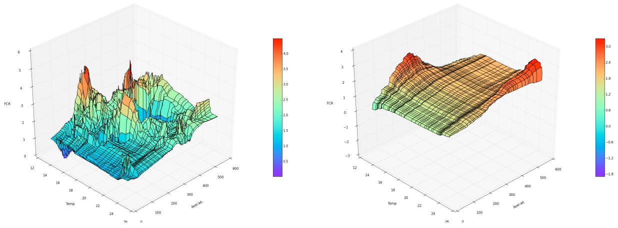

Figure 5. Company C: Real data (on the left) and FCR model (on the right) for the bream production.

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 big number of peaks in the real data, plotted on the left, correspond 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.

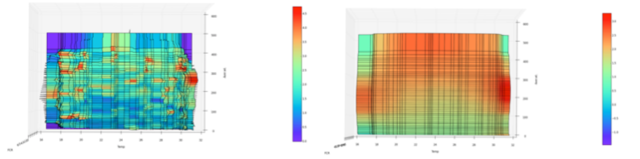

Figure 6. The grid view of both the real data (on the left) and the FCR model (on the right) for the company C.

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, 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 neighbours algorithm is used here to find the values outside the area [5]. 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.

5. CONCLUSIONS

The challenges of aquaculture for data analytics are very specific in the field and must be addressed with the appropriate methodology and technology, in tune with the expertise of the fish farmers. The uncertainty of measures, 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 the aquaSmart project and discussed in this paper 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. Moreover, the statistical analysis of the results permit a clearer visualization of the important features in the data that can boost the production and optimize the processes related to it. That will enable classification and forecast based on the analytics of the available data.

In that, future work includes the production of guidelines validated by end-users in order to facilitate the application of further advanced learning methods in aquaculture.

ACKNOWLEDGMENTS

The authors would like to acknowledge that his work was funded by the project AquaSmart, co-funded by the European Commission with the agreement number 644715 (H2020 Programme). We would also like to thank to Primož Škraba, Luka Stopar, Dunja Mladenić, Pedro Gomes de Oliveira, Ruben Costa, Gerasimos Antzoulatos, Nir Tzohari, John McLaughlin and Gary McManus.

REFERENCES

[1] aquaSmart Consortium, The aquaSmart Project [Online]. URL: www.aquasmartdata.eu. [accessed in 22.8.2016].

[2] Bhujel, R. C. (2011). Statistics for Aquaculture. John Wiley & Sons.

[3] Beveridge M (2004); Cage Aquaculture; Third Edition, Oxford, UK.

[4] N. R. Draper and H. Smith, Applied regression analysis, 3th edition. New York: Wiley, 1998.

[5] Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques. Elsevier.

[6] Dobson, A. J., & Barnett, A. (2008). An Introduction to Generalized Linear Models, Third Edition. CRC Press.

[7] 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).

[8] Lee P.G. 2000. Process control and artificial intelligence software for aquaculture. Aquacultural Engineering, 23, 13-36.

[9] Rizzo, G., and Spagnolo, M. 1996. A Model for the Optimal Management of Sea Bass Dicentrarchus Labrax Aquaculture. Mar. Resour. Econ. 11: 267–286.

[10] Stickney, R. R. (2007). Aquaculture: An introduction. (C. Publishing, Ed.). CABI Publishing.