Growth model for white shrimp in semi-intensive farming using inductive reasoning methodology

Raúl Carvajal V. *
Escuela de Informática.
Universidad Autónoma de Sinaloa
Av. Universidad y Av. Ejército Mexicano, C.P. 82000
Mazatlán, Sinaloa, México
E-mail: carvajal@ic.upc.es
FAX: (343) 401-60-00, Telephone: (343) 401-60-76

Angela Nebot
Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya
Jordi Girona Salgado, 1-3, Barcelona 08034, Spain
E-mail: angela@lsi.upc.es
FAX: (343) 401-60-50, Telephone: (343) 401-56-42

*Present address : Institut de Cibernètica. Universitat Politècnica de Catalunya. Av. Diagonal 647, Planta 2. Barcelona 08028. Spain.

ABSTRACT

A growth model for occidental white shrimp (*Penaeus vannamei*) in semi-intensive farming using Fuzzy Inductive Reasoning (FIR) methodology is presented. The model was developed using data from 17 cycles of culture (1990-1995) at the "El Remolino" shrimp farm located on the Northwest Pacific Coast in Sinaloa, Mexico. Due to the nature of the available data, special routines for handling missing data had to be used. Several qualitative relationships were found using the FIR methodology. The significant variables detected were: temperature, salinity, oxygen, and weight. The model was validated with two cycles that had not been used in deriving the model. The forecast results exhibited an error near 10%, which is a considerable improvement over the error of 20% obtained by classical statistical methods. The use of FIR methodology in aquaculture seems very promising. It can help farmers find good farming strategies for obtaining better profits.

Key words: Growth models, shrimp farming, inductive reasoning, fuzzy systems.

1. INTRODUCTION

Shrimp farming is a branch of agriculture representing annual sales of millions of dollars and giving employment (direct and indirect) to thousands of people in the world. The economic success depends on many factors, including characteristics of the site, climate, water quality, type of farming, technology used, shrimp species farmed, feeds, shrimp diseases, farm management, market prices, production costs, government support, capital, and human resources.

Farmers need to plan the dates for seeding and harvesting the ponds, taking into account all the aforementioned factors, in order to get the best profits. It is necessary to analyze the available data to propose models that can help the producers accomplish this. Few attempts have been made so far to construct growth models for shrimp in farms. The few reports found in the open literature use classical statistical methodologies, such as multiple regression, and are based on structural model assumptions that are not supported by deep knowledge as the system is quite poorly understood (Carvajal, 1993). The prediction errors obtained when using these models were at a level of approximately 20%, which is quite high because they can lead to incorrect farming decisions that may jeopardize the economic basis of the farming enterprise in serious ways.

In this paper, a new attempt at modeling shrimp growth in a farm is presented that uses an inductive qualitative modeling methodology, called *Fuzzy Inductive Reasoning (FIR)*. FIR is based on the *General System Problem Solving (GSPS)* paradigm proposed by Klir (1985). The FIR methodology has been proven to be valuable when modeling and

predicting those systems which structure is either totally or at least partially unknown, as it is usually the case of biomedical applications, such as, anesthesiology and cardiology (Nebot *et al.*, 1996, 1997).

One of the prime advantages of this methodology is that it is not based on structural knowledge of the system under study, but derives its models strictly by means of behavioral patterns observed during its knowledge acquisition phase. This feature is particularly desirable in applications, such as, shrimp farming, where little if any structural knowledge is available *a priori*. Another advantage of this methodology, related to its inherent imprecision, is that it is able to deal with uncertainty. This is very important for the task at hand, because the data records obtained through measurements in the farm are particularly imprecise. It will be shown that the prediction errors obtained using FIR models are significantly lower than those reported earlier obtained by means of classical methodologies, reducing this error from 20% to about 10%.

2. SHRIMP FARMING

Shrimp farming has become increasingly important in recent years. In 1980, only 2% of the world shrimp production of 1.6 millions of metric tons was produced on farms, whereas in 1995, already 27% of the 2.6 million metric tons of shrimp produced worldwide were farmed. Seventy eight per cent of the farmed shrimp production originated in Asia, predominantly produced by countries such as Thailand, Indonesia, China, India, and Vietnam, whereas the remaining 22% were produced in the Americas, primarily in Ecuador, Mexico, Colombia, and Honduras (Rosenberry, 1996).

In shrimp farming, the term *extensive farming* refers to low density farming (less than 25,000 juveniles per hectare), *semi-intensive farming* relates to farming at medium density (between 25,000 and 200,000 juveniles per hectare) and *intensive farming* denotes high density farming (with more than 200,000 juveniles per hectare). As the density increases, the size of the estate needed for the farm decreases, while the technology necessary to maintain such high density of animals in the ponds becomes increasingly sophisticated, the capital investment grows, and the production cost per unit space increases dramatically.

The farm studied in this research employs semi-intensive farming with carefully laid out ponds of 5 to 25 hectares, with shrimp feeding equipment and diesel pumps for water exchange. The pumps exchange between 5 and 10% of the water every day. Using the stocking rates that are characteristic of semi-intensive farms, there is already too much competition for the natural food available in the pond, and therefore, the farmers have to augment the natural food by actively supplying shrimp feeds. The construction cost for such a farming venture ranges from US$10,000 to US$25,000 per hectare. Wild or hatchery-produced post-larvae or juveniles (0.1 to 2.0 g) are stocked in growout ponds. It takes between three and six months to produce a crop of market-sized shrimp. Depending on the temperature of the site, it is possible to obtain one, two or three crops per year. The farmer harvests by draining the pond through a net, or by using a harvest pump. Yields range from 500 to 5,000 kilograms (head-on) per hectare per year, with 2,000 kilograms per hectare per year being a much sought after goal. The production cost ranges from US$2.00 to US$6.00 per kilogram of live shrimp.

After the harvest, the shrimp is frozen and packed. For the US market, the shrimp is packed in units of five pounds (2.27 kg) with head-off shrimp of uniform size. The number of homogeneous head-off shrimp (tails) that together weigh one pound, determines the size classification commonly used. Shrimp prices depend on the size classification, and are regulated by market laws. Table 1 shows recent shrimp prices in the US (Zimmerman, 1996).

For the European market, the shrimp is packed in units of two or three kilograms of frozen head-on shrimp of uniform size. As in the case of the American market, the price depends on the size and is in fact quite similar, taking into account that a tail weighs approximately 65% of the whole animal, and one kilogram corresponds to 2.2 lb.

3. DATA FROM THE "EL REMOLINO" FARM

Mexico produced 72,000 metric tons of shrimp in 1995, 17% of which came from farms. Most of the 250 Mexican shrimp farms are located along the northwestern Pacific shore. Most farmers use semi-intensive farming and grow the occidental white shrimp *Penaeus vannamei*. This shrimp was identified by Boone in 1931, and keys for identification, diagnosis, and taxonomy can be found in Dore and Frimodt (1988). Additional information about the biology of the species can be found in Barreiro (1970) and Lizárraga (1976). In semi-intensive farming, *P. vannamei* reaches sizes between 61..70 and 31..35 (head-on shrimp with an average weight of 10 to 20 g), exhibiting an average growth of 0.5 to 1.5 g/week, a feed conversion from 1.0 to 3.0 and a global mortality between 10 and 40%, depending mostly on factors such as those mentioned

in the next section.

The data presented in this paper has been collected from the "El Remolino" farm, property of the enterprise *Camaricultores de Sinaloa*. The farm is located on the northwestern Pacific coast of Mexico, close to a coastal lagoon, from which the water is pumped. In this geographic zone, the climate is sub-tropical with two annual rain periods. The first and most important one occurs during the summer (June to October), whereas the other rainy season, with occasional rains only, occurs during the winter (December to February). The salinity of the water in the ponds depends heavily on the rainfall, varying from 15 parts per thousand towards the end of the summer rainy season to 90 ppt at the beginning of the summer rainy season, as it is shown in the upper graph of Fig. 1. The minimal temperatures in the ponds are registered in February (18°C), whereas the maximal temperatures occur in August (39°C), as can be seen in the lower graph of Fig. 1.

The farm usually processes two cycles per year, and the annual production is approximately 400 tons. Table 2 shows the minimal, maximal, and mean values of the principal variables that were recorded for the 16 ponds of the farm, from 1987 to 1995.

As can be seen in Table 2, the most important output variables, i.e., the final weight and the yield, have varied considerably over these years. It is important to notice that the input variables are subject to wide variations due partly to externally controllable factors, such as stock density, and partly to uncontrollable factors such as the temperature and amount of rainfall. It is important to be able to identify which of

these variables are most significant, and to determine how they affect the production, in order to conceive models that make it possible to predict what would happen if these variables were to assume any given set of values. A growth model is essential to knowing how the shrimp will grow, and therefore to be able to plan the best seeding and harvesting strategy that will optimize the profit obtained.

4. SYSTEM VARIABLES AFFECTING GROWTH

There are many factors affecting the production. In this paper, only the most significant among the physical (non-controllable) as well as technological (controllable) factors are mentioned. More detailed information can be found in Cook and Rabanal (1978), Gulland and Rothschild (1984), Korringa (1976), Lawrence (1983), and Weng-Young (1981).

Feed. The quality of feed is very important, because the shrimp obtain most of their nutrition from it. High-quality feeds offer several advantages over lower-quality feeds: better feed conversion, faster growth, lower mortality, and improved water quality. Shrimp in ponds must be fed between one and three times a day with pellets. The protein content of the feed varies from 20 to 40%.

Density. The number of shrimp present in a pond is an important factor influencing the growth process. As the shrimp density increases, there is more competition among the animals for space, feed, and oxygen, with the result of more organic residuals, and thereby a decrease in the growth of the animals. These difficulties can be overcome by increasing the water exchange and feeding rates.

Temperature. Most shrimp are farmed between 22°C and 34°C. In general, as temperature increases, the growth rate also increases. However, temperatures above 35°C produce mortality.

Salinity. Adult shrimp can exist in a wide range of salinity, from 10 to 60 parts per thousand. However, they do not tolerate quick changes in salinity, especially at low values. Optimal values are between 25 and 35 ppt. Values of salinity over 40 ppt reduce the growth rate.

Oxygen. The normal amount of oxygen in the water ranges from 2 to 10 parts per million. Values below 2 parts per million cause stress, a low growth rate, and mortality. Water with high temperature and salinity contains less oxygen.

Visibility. This is a measure of the transparency of the water. It is directly related with the presence of plankton. It also provides information about water quality. It is measured with a Secchi disc, a bar with a mobile disc that is placed in the water. The disc is moved down until the user lose sight of it. The space left between the beginning of the bar and the disc indicates the centimeters of visibility of the pond. Recommended values are between 25 and 35 cm.

5. FIR METHODOLOGY

As has been mentioned earlier, the approach used in this paper to obtain a shrimp growth model is the *Fuzzy Inductive Reasoning (FIR)* methodology. Inductive reasoning was first developed by George Klir (1985) as a tool for general system analysis. Fuzzy measures were introduced into the General System Problem Solver in the late eighties

(Li and Cellier, 1990), resulting in the Fuzzy Inductive Reasoning (FIR) methodology.

FIR is a data driven methodology based on system behavior rather than structural knowledge. It is able to derive causal qualitative relations between the variables of the system, and to infer future behavior of that system from observations of its past behavior. It is therefore a useful tool for modeling and simulating those systems for which no *a priori* structural knowledge is available, including agricultural systems. The FIR methodology is composed of four main processes, namely: *fuzzification*, *qualitative modeling*, *qualitative simulation*, and *defuzzification*. These four processes are shown in Fig. 2.

SAPS-II (System Approach Problem Solver, Version II), which is the implementation of the FIR methodology, has been used in this study. It is available as a Matlab toolbox (Cellier, 1991).

5.1 Fuzzification

FIR is fed with data measured from the system under study, that are then converted into fuzzy information by means of the *fuzzification* function. The fuzzification function converts quantitative values into qualitative triples, as is shown in Fig. 3. The first element of the triple is the class value, the second element is the fuzzy membership value, and the third element is the side value. The class value represents a discretization of the original real–valued variable. The fuzzy membership value denotes the level of confidence expressed in the class value chosen to represent a particular quantitative value. Finally, the side value tells us whether the quantitative value is to the left, to the right or to the center of the peak

value of the membership function. The side value, which is a peculiarity of FIR methodology since it is not commonly introduced in fuzzy logic, is responsible for preserving the complete knowledge in the qualitative triple that had been contained in the original quantitative value.

Figure 3 shows an example of fuzzification of the variable temperature. For instance, a quantitative temperature value of $23°C$ is discretized into a qualitative class value of 'normal' with a fuzzy membership function value of 0.895, and a side function value of 'right' (since 23 is to the *right* of the maximum of the bell–shaped membership function that characterizes the class 'normal'). Thus, a single quantitative value is recoded into a *qualitative triple*. As it is also shown in Fig. 3, the landmarks 13 and 27 define the class value 'normal' (*landmarks* being the points where the class value changes). Therefore, any temperature with a quantitative value between 13 and 27 will be recoded into the qualitative class value 'normal'.

In order to convert quantitative values to qualitative values, it is necessary to provide to the fuzzification function the number of classes into which the definition domain of each variable is going to be divided as well as the landmarks that separate neighboring classes from each other. Once this information has been provided, the fuzzification engine of FIR is capable of automatically fuzzifying the quantitative data values using either Gaussian or triangular fuzzy membership functions.

5.2 Qualitative Modeling

The *qualitative modeling* function of the FIR methodology is responsable for finding causal spatial and temporal relations between variables

that offer the best likelihood for being able to predict the future system behavior from its own past, thereby obtaining the best model (called a *mask* in the FIR terminology) that represents the system.

At this point, the continuous trajectory behavior recorded from the system has been converted into an episodical behavior (qualitative data) by means of the fuzzification function. In the process of modeling, it is desired to discover causal relations among the recoded variables that make the resulting state transition matrices as deterministic as possible. A mask represents a possible relation among the qualitative variables. Let us introduce the concept of a mask by means of a simple example composed of two inputs and three outputs.

$$
\begin{array}{c}
{}_t\backslash^x \\
t-2\delta t \\
t-\delta t \\
t
\end{array}
\begin{array}{c}
\begin{array}{ccccc} u_1 & u_2 & y_1 & y_2 & y_3 \end{array} \\
\left(\begin{array}{ccccc}
0 & 0 & 0 & 0 & -1 \\
0 & -2 & -3 & 0 & 0 \\
-4 & 0 & +1 & 0 & 0
\end{array}\right)
\end{array}
\tag{1}
$$

The negative elements in this matrix are referred to as m–inputs (mask inputs), which denote input arguments of the qualitative functional relationship. They can be either inputs or outputs of the subsystem to be modeled, and they can have different time stamps. The above example contains four m–inputs. The sequence in which they are enumerated is immaterial. They are usually enumerated from left to right and from top to bottom. Therefore, the -1, -2, -3 and -4 elements of the mask do not represent numerical information but a direct causal relation between these m–inputs and the output to be predicted. The single positive value

denotes the m–output. The terms m–input and m–output are used in order to avoid a potential confusion with the inputs and outputs of the plant. In the above example, the first m–input corresponds to the output variable y_3 two sampling intervals back, $y_3(t - 2\delta t)$, whereas the second m–input refers to the input variable u_2 one sampling interval into the past, $u_2(t - \delta t)$, etc.

A mask denotes a dynamic relationship among qualitative variables. It has the same number of columns as the episodical behavior, and it has a certain number of rows, the *depth* of the mask. It represents the temporal domain that can influence the output. Each row is delayed relative to its successor by a time interval of δt representing the time lapse between two consecutive samplings. δt may vary from one application to another. In the previous example the output, $y_1(t)$, will be derived from the values of the four m–inputs that compose the mask and, therefore, the output is influenced by the values of different variables at different points in time (t, $t - \delta t$ and $t - 2\delta t$).

How is a mask found that, within the framework of all allowable masks, represents the most deterministic state transition matrix? This mask will optimize the predictiveness of the model. In SAPS–II, the concept of a *mask candidate* matrix has been introduced. A mask candidate matrix is an ensemble of all possible masks from which the best is chosen by a mechanism of exhaustive search.

The mask candidate matrix contains '-1' elements where the mask has a potential m–input, a '$+1$' element where the mask has its m–output, and '0' elements to denote forbidden connections. Thus, a mask candi-

date matrix to determine a predictive model for variable y_1 in the previously introduced five–variable example might be:

$$
\begin{array}{c} {}_t\backslash^x \\ t-2\delta t \\ t-\delta t \\ t \end{array}
\begin{array}{c}
\begin{array}{ccccc} u_1 & u_2 & y_1 & y_2 & y_3 \end{array} \\
\begin{pmatrix} -1 & -1 & -1 & -1 & -1 \\ -1 & -1 & -1 & -1 & -1 \\ -1 & -1 & +1 & 0 & 0 \end{pmatrix}
\end{array}
\tag{2}
$$

Here again, the -1 elements of the mask candidate matrix do not represent numerical information but possible causal relations with the output. The function of the optimal mask of SAPS–II is to find those causal relations among all the possible relations allowed (-1 elements). Corresponding mask candidate matrices are used to find predictive models for y_2 and y_3. In all three mask candidate matrices, the instantaneous values of the other two output variables are blocked out in order to prevent *algebraic loops* to occur between the output variables that are to be estimated.

The optimal mask function of SAPS–II searches through all legal masks of complexity two, i.e., all masks with a single m–input and finds the best one; it then proceeds by searching through all legal masks of complexity three, i.e., all masks with two m–inputs and finds the best of those; and it continues in the same manner until the maximum complexity has been reached. In all practical examples, the quality of the masks will first grow with increasing complexity, then reach a maximum, and then decay rapidly. A good value for the maximum complexity is usually five or six (Nebot, 1994; Mugica and Cellier, 1994).

Each of the possible masks is compared to the others with respect to its potential merit. The optimality of the mask is evaluated with respect to the maximization of its forecasting power, that is quantified by means of the *quality* measure. Let us focus on the computation of the quality of a specific mask.

The overall quality of a mask, Q_m, is defined as the product of its uncertainty reduction measure, H_r, and its observation ratio, O_r:

$$Q_m = H_r \cdot O_r \tag{3}$$

The uncertainty reduction measure is defined as:

$$H_r = 1.0 - H_m / H_{\max} \tag{4}$$

where H_m is the overall entropy of the mask and $H_{\max}$ the highest possible entropy. H_r is a real number in the range between 0.0 and 1.0, where higher values usually indicate an improved forecasting power. The masks with highest entropy reduction values generate forecasts with the smallest amounts of uncertainty. The highest possible entropy $H_{\max}$ is obtained when all probabilities are equal, and a zero entropy is encountered for relationships that are totally deterministic.

The overall entropy of the mask is then computed as the sum:

$$H_m = -\sum_{\forall i} \mathrm{p}(i) \cdot H_i \tag{5}$$

where $\mathrm{p}(i)$ is the probability of that input state to occur and H_i is the Shannon entropy relative to the i_{th} input state. The Shannon entropy relative to one input state is calculated from the equation:

$$H_i = \sum_{\forall o} \mathrm{p}(o|i) \cdot \log_2 \mathrm{p}(o|i) \tag{6}$$

where $\mathrm{p}(o|i)$ is the "conditional probability" of a certain m–output state o to occur, given that the m–input state i has already occurred. The term probability is meant in a statistical rather than in a true probabilistic sense. It denotes the quotient of the observed frequency of a particular state in the episodical behavior divided by the highest possible frequency of that state.

The observation ratio, O_r, measures the number of observations for each input state. From a statistical point of view, every state should be observed at least five times (Law and Kelton, 1991). If every legal m–input state has been observed at least five times, O_r is equal to 1.0. If no m–input state has been observed at all (no data are available), O_r is equal to 0.0. The optimal mask is the mask with the largest Q_m value.

5.3 Qualitative Simulation

Once the best model (mask) has been identified, it can be applied to the qualitative data matrices that were previously obtained in the fuzzification process, resulting in a 'rule base' that, in the FIR terminology, is called the *behavior matrix*. Once the behavior matrix and the mask are available, a prediction of future output states of the system can take place using the FIR inference engine. This process is called *qualitative*

simulation.

The FIR inference engine is based on the k–nearest neighbor rule, commonly used in the pattern recognition field. In particular, the 5–NN pattern matching algorithm is the core of the FIR inference process. The forecast of the output variable is obtained by means of the composition of the potential conclusion that results from firing the five rules whose antecedents have best matching with the actual state.

The forecasting procedure is presented in diagram of Fig. 4 (with an example containing three inputs and one output). The mask is placed on top of the qualitative data matrix, in such a way that the m–output matches with the first element to be predicted. The values of the m–inputs are read out from the mask, and the behavior matrix ('rule base') is used, as it is explained latter, to determine the future value of the m–output, which can then be copied back into the qualitative data matrix. The mask is then shifted further down one position to predict the next output value. This process is repeated until all the desired values have been forecast. The qualitative simulation process predicts an entire qualitative triple from which a quantitative variable can be obtained whenever needed.

The fuzzy forecasting process works as follows, the membership and side functions of the new input state (input pattern in Fig. 4) are compared with those of all previous recordings of the same input state contained in the behavior matrix. For this purpose, a normalization function is computed for every element of the new input state and an Euclidean distance formula is used to select the 5 nearest neighbors, the ones with

smallest distance, that are used to forecast the new output state.

The contribution of each neighbor to the estimation of the prediction of the new output state is a function of its proximity. This is expressed by giving a distance–weight to each neighbor, as shown in Fig. 4. The new output state values can be computed as a weighted sum of the output states of the previously observed five nearest neighbors.

5.4 Defuzzification

Defuzzification is the inverse function of fuzzification. It converts qualitative triples into quantitative values. As has been mentioned earlier, no information is lost in the process of fuzzication. The qualitative triple contains exactly the same information as the original quantitative value, and it is thus possible to defuzzify the quantitative value from the qualitative triple precisely, i.e., without any error or uncertainty, at any point in time.

For a deeper insight of the FIR methodology the reader is referred to Cellier *et al.* (1996) and Nebot (1994).

6. GROWTH MODEL

FIR, just like all other inductive modeling methodologies, is based on observations of patterns of input/output behavior, rather than on structural information. Therefore it needs rich data in order to find a model. In the application described in this paper, 19 sets of data corresponding to 19 production cycles in the farm were available. Nine data sets were recorded from pond number 1, whereas the other ten were recorded from pond number 5. Both ponds have a size of approximately

10 hectares. The recording took place in the years 1990-1995. In both cases, there were about the same number of cycles registered during summer and winter.

During each cycle, the following variables were recorded on a weekly basis: shrimp weight (system output), feed, density, temperature, salinity, oxygen, and visibility (system inputs). All input variables were sampled daily, whereas the value recorded was the mean value over the entire week. Shrimp weight was sampled weekly, because from a biological point of view it is an adequate period to detect shrimp growth. In order to illustrate the growth patterns, a subset of the available data are presented in Fig. 5. The first plot corresponds to a winter recording of pond 1, whereas the second plot refers to a summer recording of pond 5. From the available 19 data sets, 17, corresponding to approximately 340 samples, were used to obtain the growth model, whereas the remaining two sets, containing about 40 samples, were used to validate the FIR model.

Because each cycle corresponds to a different data set, and because FIR needs consecutive data to construct a model, the *missing data feature* available in the methodology has been used. This process, a knowledge combination technique, allows the concatenation of different data sets by adding a group of pre-defined missing data values between separate sets of available data (Nebot, 1994). All input variables recorded were considered as potentially useful for the model, allowing the inductive reasoner to discard those that were less causally relevant. These variables are described in section 4 of the paper. Previously recorded values of shrimp weight were considered as additional potential input variables.

The first step in constructing the FIR model is to convert the quantitative variables to qualitative data using the fuzzification function, as has been described in section 5. To accomplish this, it is necessary to decide on the number of discrete levels (classes) into which each of these variables will be recoded. For the example at hand, it was decided that all seven variables could be sufficiently well characterized by three levels. The landmarks between neighboring classes were established taking all data of each variable and dividing them into three groups of approximately equal size. The results of this process are presented in Table 3. The resulting landmarks were shown to an expert farmer who found the division into classes satisfactory and reasonable, from the farming point of view, for all the variables recorded (Carvajal, personal communication, 1997).

The next step is to find a model that describes the system. The model is represented by a mask through which the causal relations (both spatial and temporal) between input and output variables are described. The process starts with the definition of the mask candidate matrix encoding an ensemble of all possible masks from which the best is to be chosen.

In the study presented in this paper, a value of δt of one week was selected. The mask candidate matrix used in this application is of depth 3. It is shown below :

$$
\begin{array}{c} {}_t\backslash^x \\ t-2\delta t \\ t-\delta t \\ t \end{array}
\begin{array}{c}
\begin{array}{ccccccc} D & S & V & O & Te & F & W \end{array} \\
\begin{pmatrix} -1 & -1 & -1 & -1 & -1 & -1 & -1 \\ -1 & -1 & -1 & -1 & -1 & -1 & -1 \\ -1 & -1 & -1 & -1 & -1 & -1 & +1 \end{pmatrix}
\end{array}
$$

where D stands for density, S for salinity, V for visibility, O for oxygen, Te for temperature, F for feed, and W for weight. It was decided to propose a depth of 3, because it was considered unlikely that any of the variables in question would have a delay of more than two weeks in their relation with the output variable.

As a means to accelerate the search, the optimal mask function of SAPS-II offers the possibility to specify an upper limit to the acceptable *mask complexity*, i.e. the largest number of non-zero elements that the mask may contain. In the present case, a maximum complexity of five was chosen. The optimal mask function of SAPS-II returned the following suboptimal masks of complexities three, four, and five. The qualities of these masks are also indicated:

$$
\begin{array}{c} {}_t\backslash^x \\ t-2\delta t \\ t-\delta t \\ t \end{array}
\begin{array}{c}
\begin{array}{ccccccc} D & S & V & O & Te & F & W \end{array} \\
\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & -2 & 0 & +1 \end{pmatrix}
\end{array}
$$

$$Quality \;=\; 0.786$$

$$
\begin{array}{c|c}
{}_t\backslash^x & \begin{array}{ccccccc} D & S & V & O & Te & F & W \end{array} \\
\begin{array}{c} t-2\delta t \\ t-\delta t \\ t \end{array} & \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & -2 & 0 & 0 & -3 \\ 0 & 0 & 0 & 0 & 0 & 0 & +1 \end{pmatrix}
\end{array}
$$

$$Quality \;=\; 0.781$$

$$
\begin{array}{c|c}
{}_t\backslash^x & \begin{array}{ccccccc} D & S & V & O & Te & F & W \end{array} \\
\begin{array}{c} t-2\delta t \\ t-\delta t \\ t \end{array} & \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & -2 & 0 & 0 & -3 \\ 0 & -4 & 0 & 0 & 0 & 0 & +1 \end{pmatrix}
\end{array}
$$

$$Quality \;=\; 0.550$$

The interpretation of a mask has been described already in the previous section. For example, the mask of complexity 5 can be interpreted as follows, the weight W at the current time t somehow depends on last week's values of the degree of visibility V, the level of oxygen O in the pond, as well as the weight that the shrimp had a week ago. It also depends on the current value of salinity S in the lake.

It can be seen that a mask depth of two provides a model that sufficiently represents the system, because the output variable at time t is influenced primarily by the values of some of the measured variables at that time and at one week earlier.

SAPS-II determined that the density D and the feed F are less important variables for explaining the observed growth patterns available

for producing the qualitative model. For this reason, these two variables do not appear in the best masks selected and, therefore, are of lesser importance for explaining the observed output patterns. The elimination of these two variables from the set of important inputs may at first sight be surprising, but can in fact be reasonably explained. When the shrimp density is increased, the farmers always add more feed to the ponds, and also enhance the water exchange, in order to maintain the oxygen content of the water at an acceptable level. Also, the feed is always supplied proportional to the current biomass in the pond. Thus, the seven variables are by no means independent of each other, and it is not necessary to know them all to explain the output. SAPS-II simply eliminated redundant information.

It is also worth noticing that the masks of complexity 3 and 4 are characterized by quite similar quality values, yet make use of different input variables. Whereas in the mask of complexity 3 the current temperature was chosen as an important input variable, the mask of complexity 4 proposed the previous week values of salinity and oxygen contents to be used instead. Although these three variables are uncontrollable, they are not independent of each other. Moreover, the selected optimal mask always represents a compromise taking into account all of the available data. However, some of the data records represented summer cycles, whereas others were obtained from winter cycles. As will be discussed later, the mask of complexity 3 is generally better suited for summer cycles, whereas the mask of complexity 4 leads to better predictions in the context of winter cycles.

7. MODEL VALIDATION

With the masks obtained in the previous section, two different cycles were forecast to validate the models. The first data set corresponded to pond number one, cycle fifteen, whereas the second corresponded to pond number five, cycle seven. The results of the prediction with each of the three models for the first data set (1-15) are presented in Figs. 6a, 6b, and 6c. Figures 7a, 7b, and 7c show forecasting results for the second data set (5-7). The results obtained for both cycles are quite good when using the mask of complexity 3 in cycle 1-15 and the mask of complexity 4 in cycle 5-7. In order to evaluate quantitatively the degree of agreement between the forecast and the real data, a normalized mean square error (MSE) was used, defined by:

$$MSE = \frac{E[(y(t) - \hat{y}(t))^2]}{y_{\text{var}}} \cdot 100\%$$

where E denotes the mean value and y_{var} is the variance. The MSE values for the six predictions of Figs. 6a, 6b, 6c as well as 7a, 7b, 7c are given in Table 4.

For cycle 1-15, the mask of complexity 5 could not forecast after week 12, because an input pattern resulted in week 13 that had never been seen before, as can be seen in Fig. 6c. Thus, SAPS-II prematurely ended the forecasting process. According to the results presented in table 4, the lowest error for cycle 1-15 is 6.8%, whereas it is 10.3% for cycle 5-7. These errors are acceptably low and significantly smaller than those obtained earlier for the same data using classical statistical techniques, which were

around 20% (Carvajal, 1993). The reduction of the prediction error of roughly 50% is quite important considering its economic impact on the production.

8. DISCUSSION

As can be seen in Figs. 6a, 6b, and 6c as well as 7a, 7b, and 7c, the best predictions for cycle 5-7 are obtained using the mask of complexity four, whereas for cycle 1-15, a smaller error is obtained using the mask of complexity three. Why are the best predictions in each cycle obtained using different masks? Examining the data, it is possible to come up with a reasonable explanation. In the case of the prediction of cycle 1-15, the shrimp were stocked in the pond during the month of August, coinciding with the main rainy season, therefore the salinity, which is one of the primary factors limiting shrimp growth, maintains adequate levels throughout the entire cycle, and therefore, temperature is the most important variable in this case. By contrast, in cycle 5-7, the shrimp were seeded during the month of March, i.e., when the salinity becomes most problematic, and therefore is the dominant factor to be considered. Although higher temperature values are also advantageous in this period, the hypersalinity of the water does not permit the shrimp to grow as much as would be desired.

The above discussion suggests that it might be useful to have available a specialized model for each of the two main seasons of the region, i.e., for the *wet season* (extending from June to December), and for the *dry season* (January to May). The most adequate model for the wet season could be the mask of complexity 3 that includes temperature as main

variable, whereas for the dry season, the mask of complexity 4 may be the most satisfactory, because it includes salinity and oxygen as the principal variables.

In some cases, it might be preferable for the farmers to have available a global model that is valid for any season of the year. In this situation, it would be possible to apply a voting procedure (implemented in SAPS-II) as explained in (Nebot and Cellier, 1994). This technique, instead of working with a single optimal mask, as described in the previous section, determines three separate masks of high-quality. During the prediction process, three separate forecasts are computed at each step. Let M_a, M_b, and M_c be the three best masks. Each of these masks leads to a different forecast. Let them be called F_a, F_b, and F_c. Three distance measures can be computed in the following way:

$$\begin{aligned} D_a &= \text{abs}(F_a - F_b) + \text{abs}(F_a - F_c) \\ D_b &= \text{abs}(F_b - F_a) + \text{abs}(F_b - F_c) \\ D_c &= \text{abs}(F_c - F_a) + \text{abs}(F_c - F_b) \end{aligned}$$

denoting the distance of each forecast to its two competitors. Once the distance measures have been computed, the predicted value with the largest distance measure is rejected, whereas the mean value of the predictions obtained with the two remaining masks is accepted as the true forecast.

In the study at hand, the three selected masks were those introduced in section 6, because of their high quality values. The predictions ob-

tained for cycles 1-15 and 5-7, using the proposed voting procedure, are shown in Figs. 8 and 9, respectively. It should be noticed that it would be appropriate to include additional structural knowledge in the model, e.g. preventing the model from predicting decreasing values of shrimp weight, since such predictions do not make any sense for the application at hand.

As can be seen in these plots, the results are not as good as those obtained using the specialized models, shown on Figs. 6a, 6b, 6c and 7a, 7b, 7c. It is evident that the predictive power of the combined model is poorer than that of the individual models (as would be expected); however the *generic model* could be useful when farmers do not know exactly the dates for farming, or when these dates do not match exactly either of the two defined seasons. It should be mentioned though that even the results obtained using the generic model are still better than those obtained previously using classical statistical methods (Carvajal, 1993). The MSE errors are 10.8% for cycle 1-15 and 17.1% for cycle 5-7, respectively.

As a last remark, a more complete evaluation of the predictive power of the previously obtained models would be desirable. Undoubtedly, the use of more than two data sets during model validation process would considerably increase the evidence of the predictive capability of the models. However, within the registered data available for this study, it was not possible to use more than two data sets for validation purposes, due to the fact that this would imply a reduction of the amount of the training data available. That is to say, the more data sets we use for validation purposes, the less data sets will be left for training purposes. As has

been already explained, FIR is a data driven methodology that infers the model from system m easured trajectories. FIR is not able to obtain 'good' models with poor and/or scarce data. This is the reason why different kinds of growth patterns have been chosen for model validation.

9. CONCLUSIONS

Shrimp farming is an agricultural activity that recently has gained in significance. The shrimp farmers need to plan the dates for seeding and harvesting the ponds taking into account many different factors, in order to maximize their profits. It is beneficial for them to have available models that allow them to make informed decisions.

In this paper, qualitative models of shrimp growth have been obtained by means of the *Fuzzy Inductive Reasoning (FIR)* Methodology. This methodology seems well suited for predicting shrimp growth in a farm. The main variables selected by FIR as being significant for predicting shrimp growth are: temperature, salinity, oxygen level, and previous shrimp weight. The obtained models were validated using data from two different ponds of the "El Remolino" shrimp farm located on the northwestern Pacific coast of Mexico (Sinaloa). The two predicted cycles represent different seasons of shrimp stocking into the ponds. One of them is characteristic of wet season farming (June to December), whereas the other represents a cycle of dry season farming (January to May). FIR found two different qualitative models, each one representing one of the two cycles. The wet-season model uses temperature as the most relevant input variable, which is reasonable, because during the rainy season, the salinity levels are always maintained at relatively low values, allowing

the shrimp to grow as fast as the temperature allows. The dry-season model uses salinity and oxygen levels as the most important input variables, as the high salinity levels characteristic of this period limit the shrimp growth and prevent higher temperature values from having the same positive effect on shrimp growth as during the wet season. The prediction errors obtained with these models for both cycles are less than 10.5%, which is a good result taking into account that predictions obtained previously for the same data using classical statistical techniques exhibited errors of somewhere around 20%. This significant improvement in forecasting can have an important economic impact when planning production in shrimp farms.

It would be desirable to use more than two data sets for validation purposes in order to maximize the evidence of the predictive capability of the models. With the reduced amount of registered data available, it was not possible to use more than two validation data sets. However, the two data sets chosen represent two different kind of growth patterns.

In this paper, a generic model useful for all seasons has also been obtained using a voting procedure, available in the FIR methodology. This model can be helpful for farmers when they do not know exactly the dates of farming or when these dates do not match either of the two defined seasons. The results obtained using the generic model are slightly better than those obtained using classical methodologies, offering errors below 18%.

ACKNOWLEDGEMENTS

The authors are thankful to Dr. Rafael Huber of the Institut de Robòtica i Informàtica Industrial, UPC-CSIC, and Dr. François E. Cellier of The University of Arizona, for their continuous interest in and support of this work. Special thanks are due also to Biologist Fernando Berdegué, General Manager of Camaricultores de Sinaloa, who provided the data analyzed in this paper. This project was partly sponsored by the Consejo Nacional de Ciencia y Tecnología, México.

REFERENCES

Barreiro, M.T. (1970) Biología del Camarón blanco Penaeus vannamei. FAO, México, 15 pp.

Carvajal, R. (1993) Modelo de Producción para una Granja de Cultivo de Camarón. Master Thesis. Centro de Estadística y Cálculo. Colegio de Postgraduados, Montecillo, México.

Cellier, F.E. (1991) Continuous System Modeling. Springer-Verlag, New York, 755 pp.

Cellier, F.E., Nebot, A., Mugica, F. and Albornoz, A. (1996) Combined Qualitative/Quantitative Simulation Models of Continuous-Time Processes Using Fuzzy Inductive Reasoning Techniques. International Journal of General Systems, 24 (1-2): 95-116.

Cook, H.L. and Rabanal, H.R., Editors (1978) Manual of Pond Culture of Penaeid Shrimp. FAO, Philipinnes, 130 pp.

Dore, I. and Frimodt., C. (1988) An Illustrated Guide of Shrimp of the Word. Osprey books, New York, 230 pp.

Gulland, J. and Rothschild, B. (1984) Penaeid Shrimps: Their Biology and Management. Fishing New Books, San Diego, 308 pp.

Klir, G.J. (1985) Architecture of Systems Problem Solving. Plenum Press, New York, 540 pp.

Korringa, P. (1976) Farming Marine Fishes and Shrimps: A multidisciplinary treatise. Elsevier Science Pub. Co., New York, 208 pp.

Lawrence, A.L. (1983) Shrimp Mariculture: State of Art, Texas University Press, 12 pp.

Law A.M. and Kelton W.D. (1991) Simulation Modeling and Analysis 2nd Edition, McGraw–Hill, New York, 759 pp.

Li, D. and Cellier, F.E. (1990) Fuzzy Measures in Inductive Reasoning. In: Proceedings 1990 Winter Simulation Conference, New Orleans, LA, pp. 527–538.

Lizárraga, S.M. (1976) Recursos pesqueros: Camarón. Facultad de Ciencias, UNAM, México.

Mugica, F. and Cellier, F.E. (1994) Automated Synthesis of a Fuzzy Controller for Cargo Ship Steering by Means of Qualitative Simulation. In: A. Guasch and R. Huber (Editors), Proc. ESM'94, European Simulation MultiConference, 1-3 June 1994, Barcelona, Spain, pp. 523–528.

Nebot, A. (1994) Qualitative Modeling and Simulation of Biomedical

Systems Using Fuzzy Inductive Reasoning. Doctoral Thesis. Departament de Llenguatges i Sistemes Informàtics. Universitat Politècnica de Catalunya, Barcelona.

Nebot, A. and Cellier, F.E. (1994) Preconditioning of Measurement Data for the Elimination of Patient-Specific Behavior in Qualitative Modeling on Medical Systems. In: Proceedings CISS'94, First Joint Conference of International Simulation Societies, August 1994,Zurich: 584-588.

Nebot, A., Cellier, F.E. and Linkens, D.A. (1996) Synthesis of an Anaesthetic Agent Administration System Using Fuzzy Inductive Reasoning. Artificial Intelligence in Medicine, 8(3): 147–166.

Nebot, A., Cellier, F.E. and Valverdú, M. (1997) Mixed Quantitative/ Qualitative Modeling and Simulation of the Cardiovascular System. Computer Methods and Programs in Biomedicine (under revision).

Rosenberry, B. (1996) World Shrimp Farming 1995. Shrimp News International, San Diego.

Weng-Young, T. (1981) Shrimp Mariculture: A Practical Manual, Chieng Cheng Pub, Hong Kong, 300 pp.

Zimmerman, J. (1996) 1995 Shrimp Prices, Horsager Trading Co., Minneapolis.

Captions of the illustrations

Fig. 1 Temperature and Salinity at the "El Remolino" farm during 1995.

Fig. 2 FIR main processes.

Fig. 3 Fuzzification of a temperature value of $23°C$.

Fig. 4 Qualitative simulation process diagram.

Fig. 5 Examples of shrimp growth patterns recorded at the "El Remolino" farm.

Fig. 6a Real data and qualitative forecast for cycle 1-15 (Mask of complexity 3).

Fig. 6b Real data and qualitative forecast for cycle 1-15 (Mask of complexity 4).

Fig. 6c Real data and qualitative forecast for cycle 1-15 (Mask of complexity 5).

Fig. 7a Real data and qualitative forecast for cycle 5-7 (Mask of complexity 3).

Fig. 7b Real data and qualitative forecast for cycle 5-7 (Mask of com-

plexity 4).

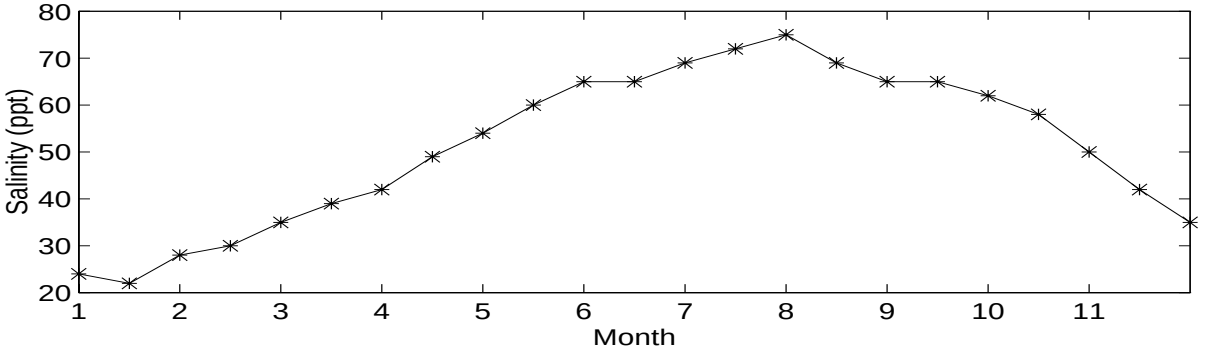
Fig. 7c Real data and qualitative forecast for cycle 5-7 (Mask of complexity 5).

Fig. 8 Real and forecast data with voting procedure for cycle 1-15.

Fig. 9 Real and forecast data with voting procedure for cycle 5-7.

Table 1: Head-off shrimp prices in US market (Zimmerman, October-1995)

Tails / lb.	Dollars / lb.	Tails / lb.	Dollars / lb.
12	10.25	36..40	4.75
15	9.00	41..50	4.00
16..20	7.75	51..60	3.30
21..25	6.75	61..70	3.00
26..30	5.75	71..80	2.70
31..35	5.25	80..over	2.30


Salinity (ppt)
80
70
60
50
40
30
20
1
2
3
4
5
6
7
8
9
10
11
Month

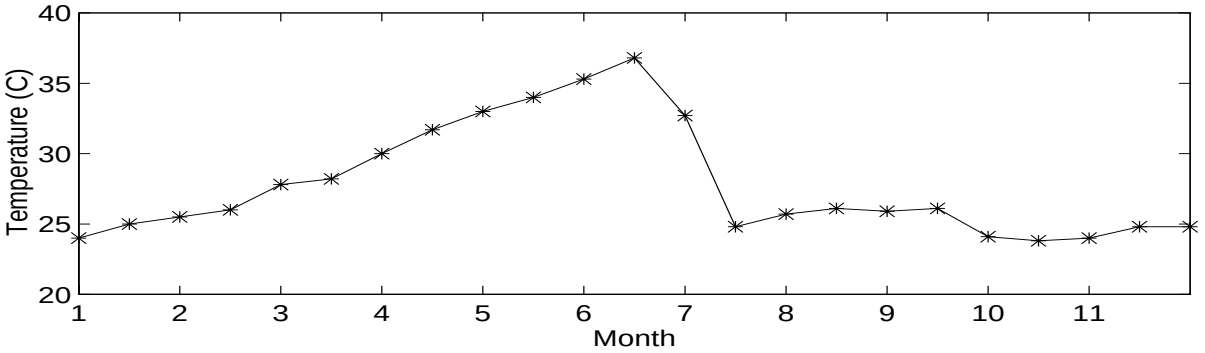
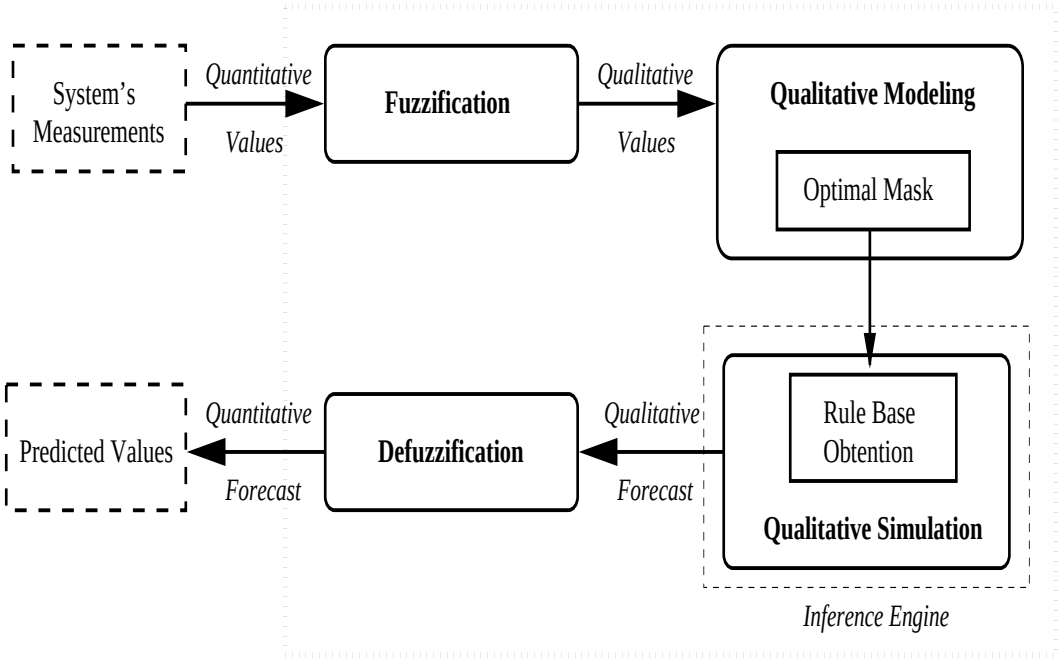
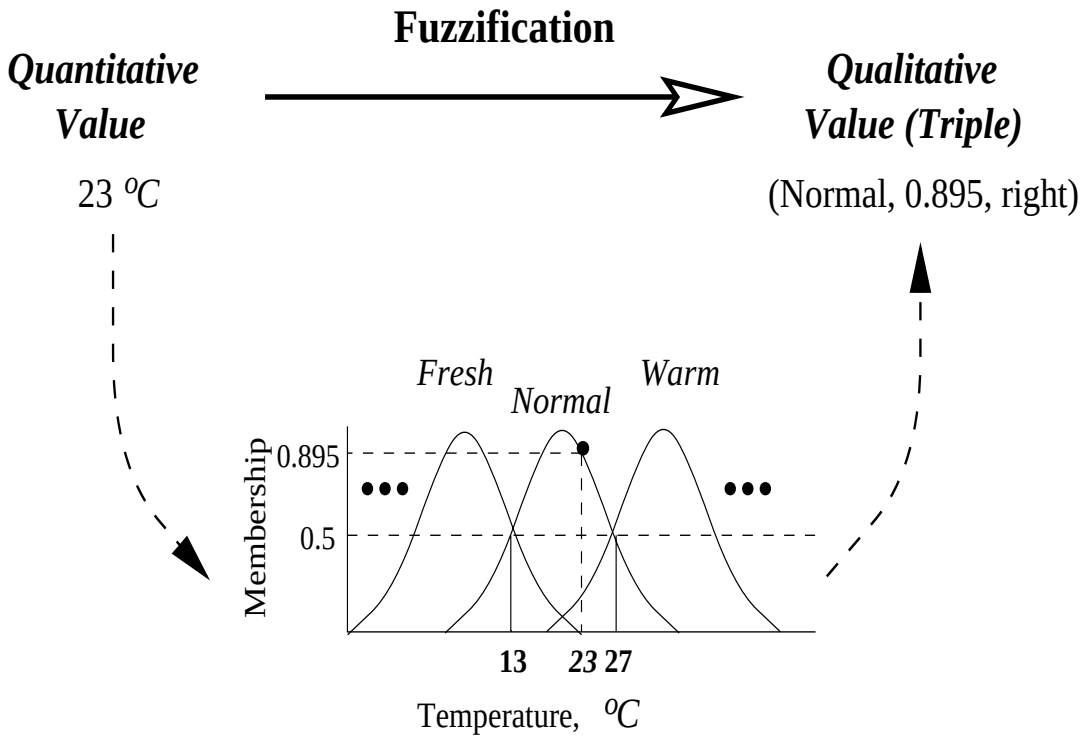

Temperature (C)
40
35
30
25
20
1
2
3
4
5
6
7
8
9
10
11
Month

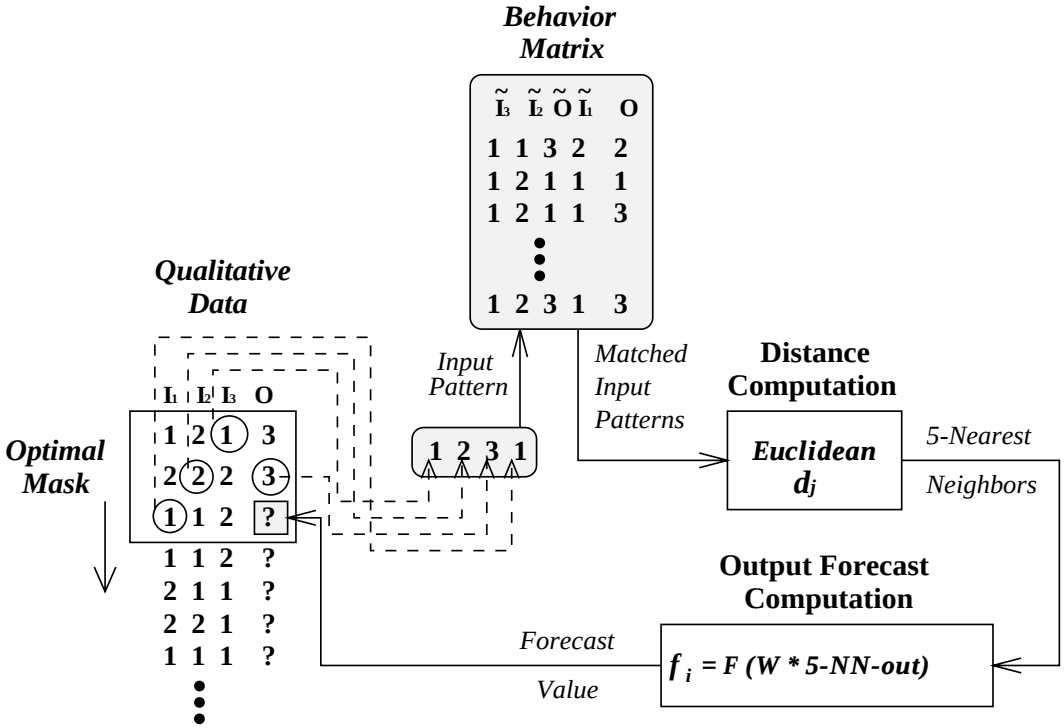
Table 2: Values of the principal variables recorded at the "El Remolino" farm (1987-1995)

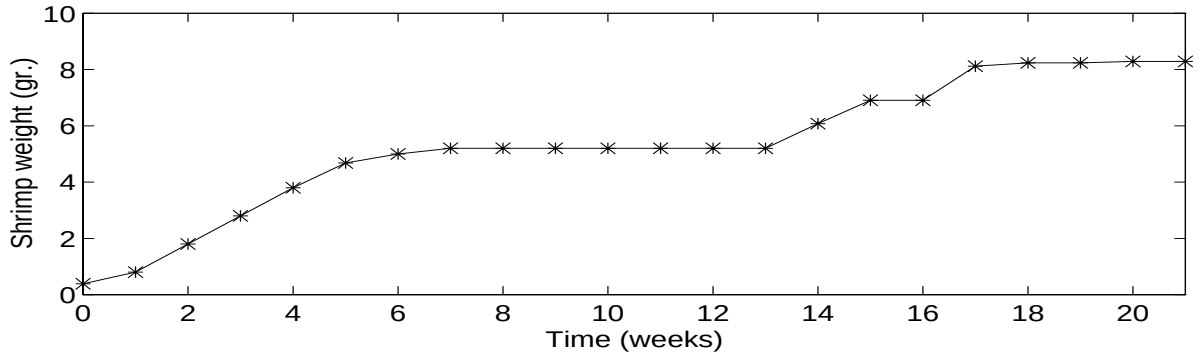
Variable	Minimum	Maximum	Mean
Density (shrimp/m^2)	36	340	137
Days in pond	60	201	132
Survival (%)	48	100	75
Feed conversion	0.6	4.4	2.3
Salinity (ppt)	15	92	48
Temperature ($^\circ C$)	18	39	28
Oxygen (ppm)	0.4	12	4.0
Visibility (cm)	12	70	30
Final weight (gr)	6.0	21.4	12.8
Yield (Tonne/ha/year)	0.6	3.5	2.2


System's Measurements
Quantitative
Values
Fuzzification
Qualitative
Values
Qualitative Modeling
Optimal Mask
Rule Base Obtention
Qualitative Simulation
Inference Engine
Qualitative
Forecast
Defuzzification
Quantitative
Forecast
Predicted Values

FIR


Fuzzification
Quantitative Value
23 °C
Qualitative Value (Triple)
(Normal, 0.895, right)
Fresh
Normal
Warm
Membership
0.895
0.5
13
23
27
Temperature, °C


Behavior Matrix
Ĩ3 Ĩ2 Õ Ĩ1 O
1 1 3 2 2
1 2 1 1 1
1 2 1 1 3
1 2 3 1 3
Qualitative Data
I1 I2 I3 O
1 2 1 3
2 2 2 3
1 1 2 ?
1 1 2 ?
2 1 1 ?
2 2 1 ?
1 1 1 ?
Optimal Mask
Input Pattern
1 2 3 1
Matched Input Patterns
Distance Computation
Euclidean dj
5-Nearest Neighbors
Output Forecast Computation
f i = F (W * 5-NN-out)
Forecast Value


Shrimp weight (gr.)
Time (weeks)
0
2
4
6
8
10
12
14
16
18
20

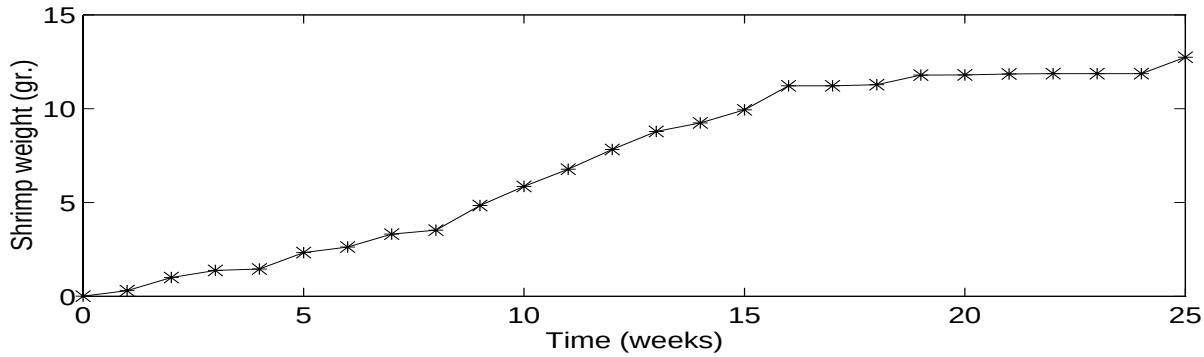
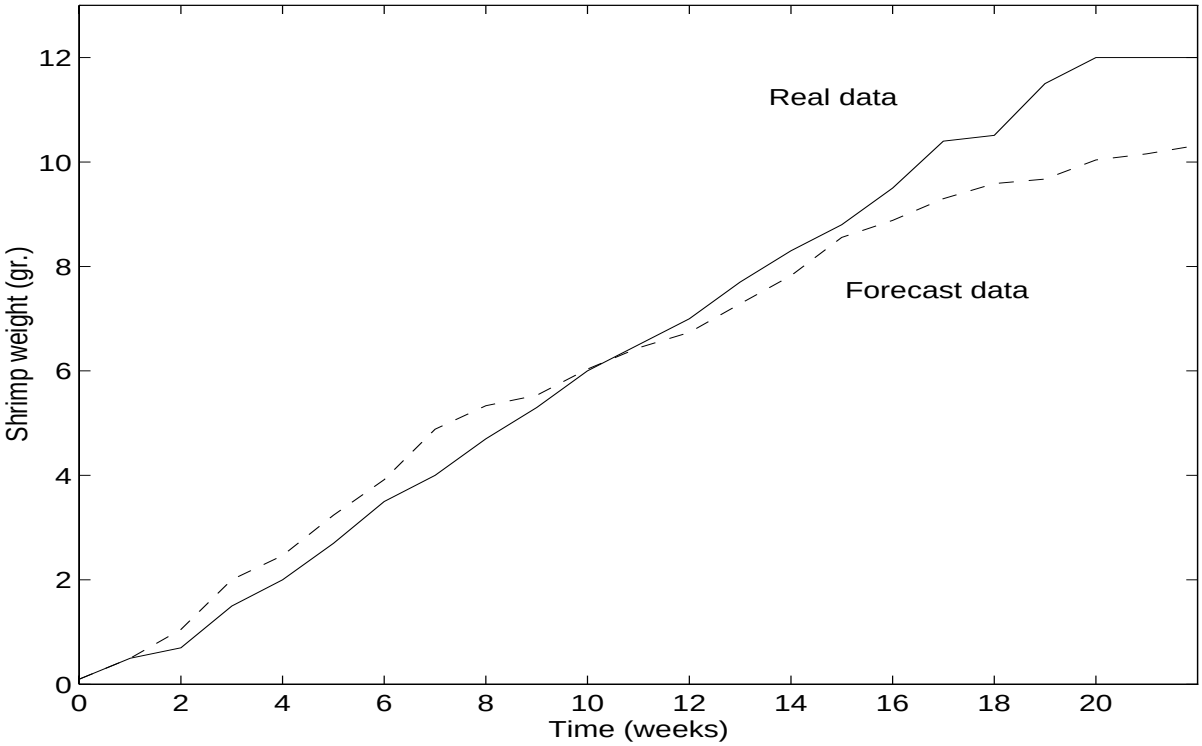
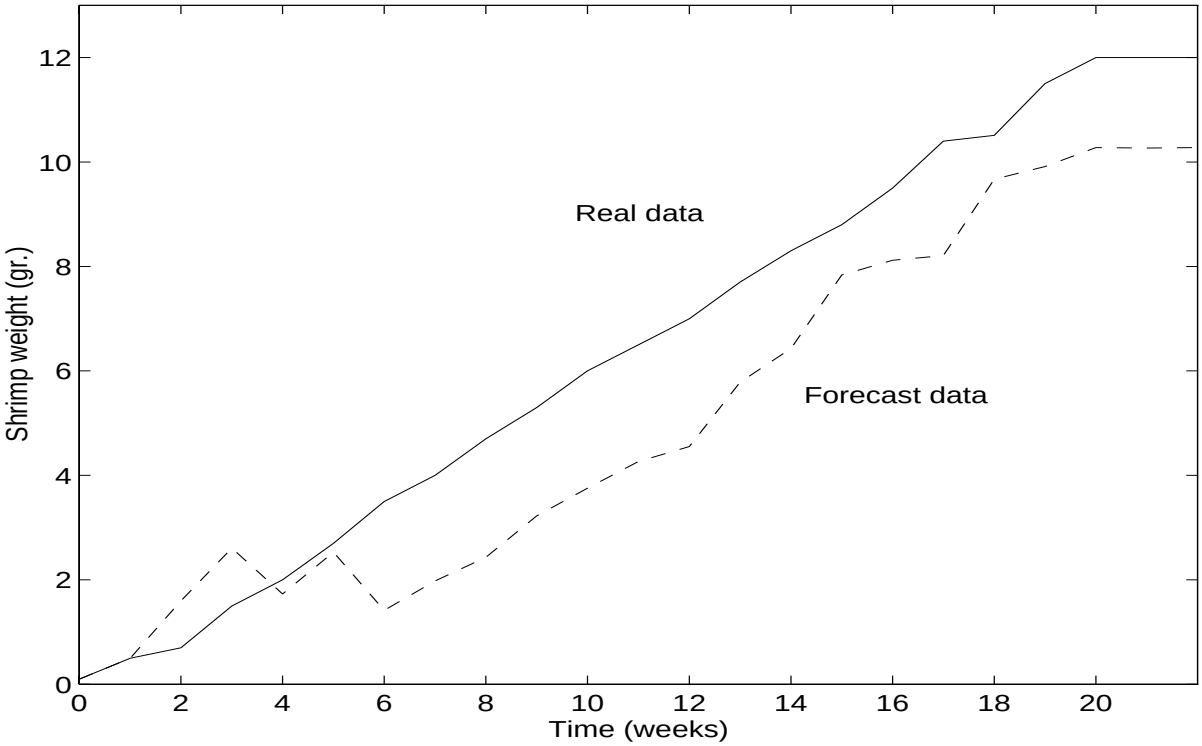
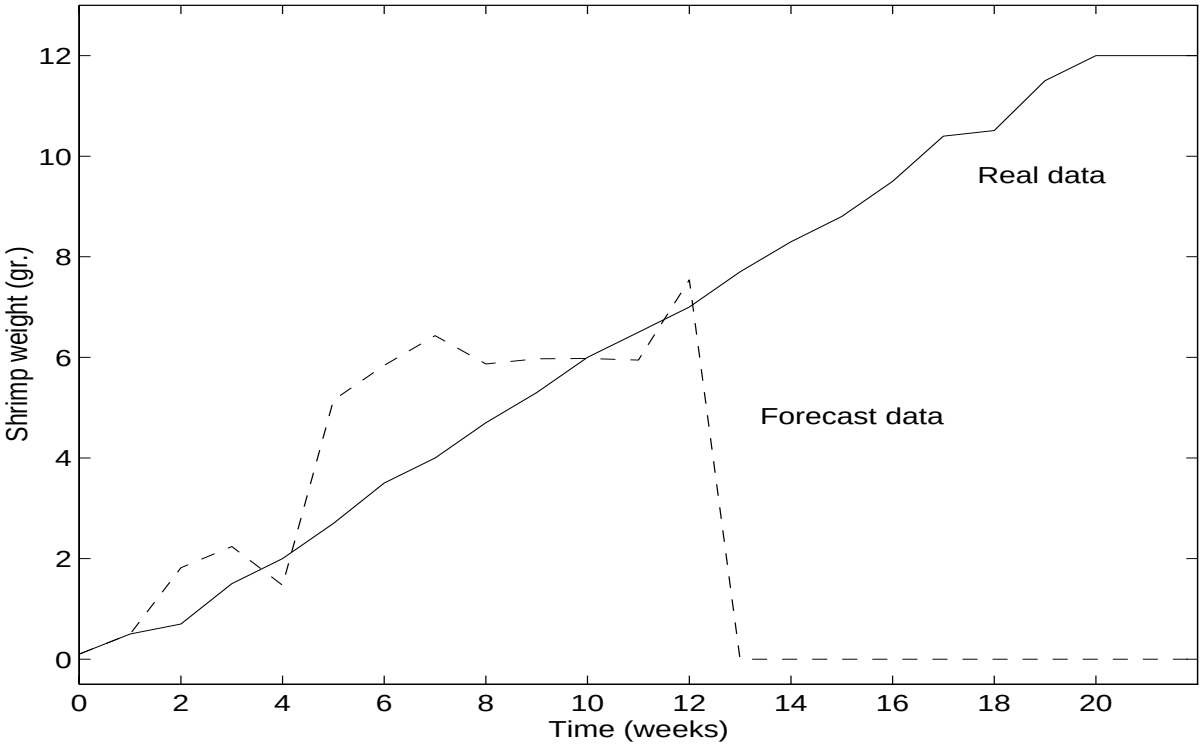

Shrimp weight (gr.)
Time (weeks)
0
5
10
15
20
25

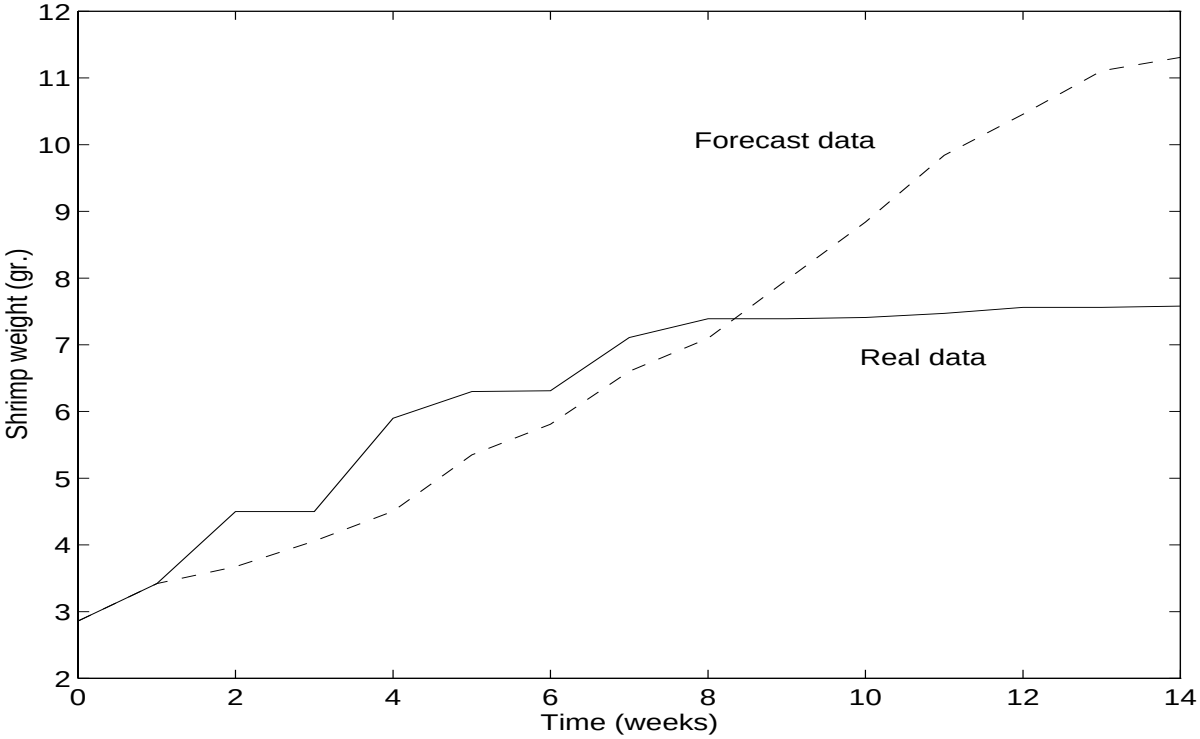
Table 3: Landmarks of the three discrete classes defined for each input variable

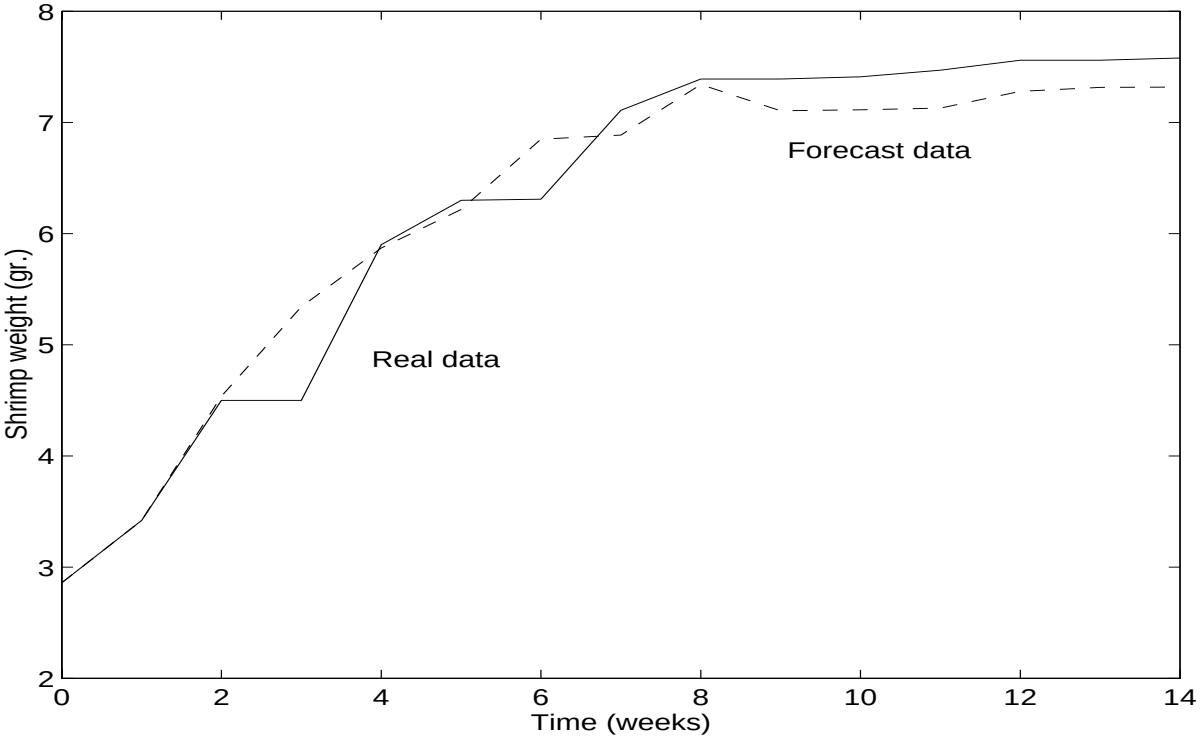
Variable	Land. 1	Land. 2	Land. 3	Land. 4
Feed	0.0	0.12	0.19	0.60
Density	2.97	8.5	18.8	24.8
Temperature	21.2	25.6	29.6	35.0
Salinity	6.7	30.9	49.0	90.7
Oxygen	0.9	3.7	5.0	11.1
Visibility	12.4	22.9	28.4	88.5
Weight	0.01	4.5	8.2	13.7


Real data
Forecast data
Shrimp weight (gr.)
Time (weeks)
0
2
4
6
8
10
12
14
16
18
20


Real data
Forecast data
Shrimp weight (gr.)
Time (weeks)
0
2
4
6
8
10
12
14
16
18
20


Shrimp weight (gr.)
Time (weeks)
Real data
Forecast data
0
2
4
6
8
10
12
14
16
18
20


Forecast data
Real data
Shrimp weight (gr.)
Time (weeks)
12
11
10
9
8
7
6
5
4
3
2
0
2
4
6
8
10
12
14


Shrimp weight (gr.)
Time (weeks)
Forecast data
Real data
8
7
6
5
4
3
2
0
2
4
6
8
10
12
14

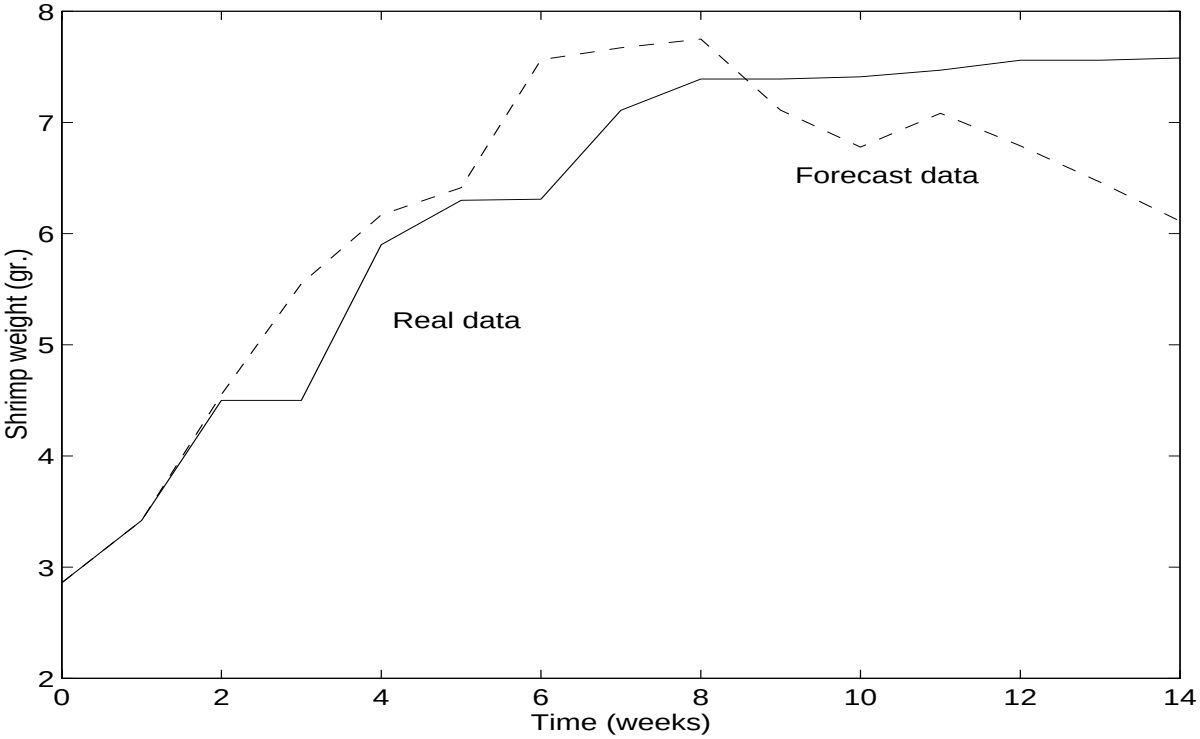
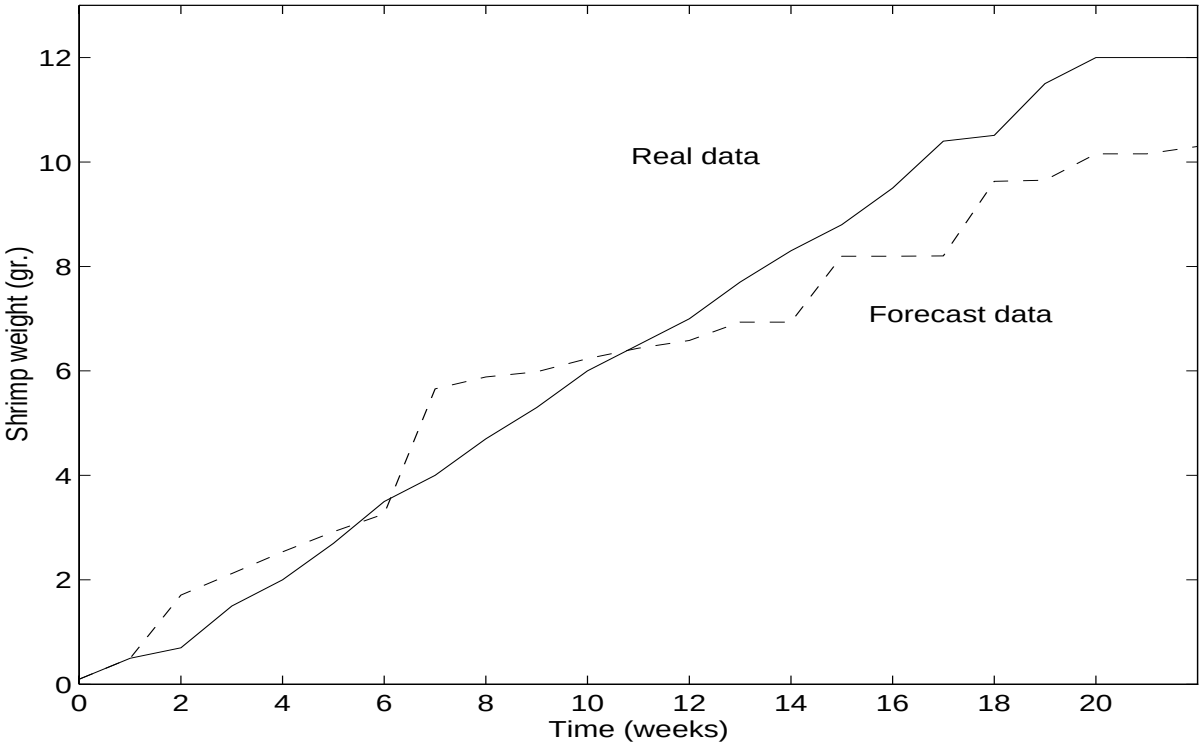
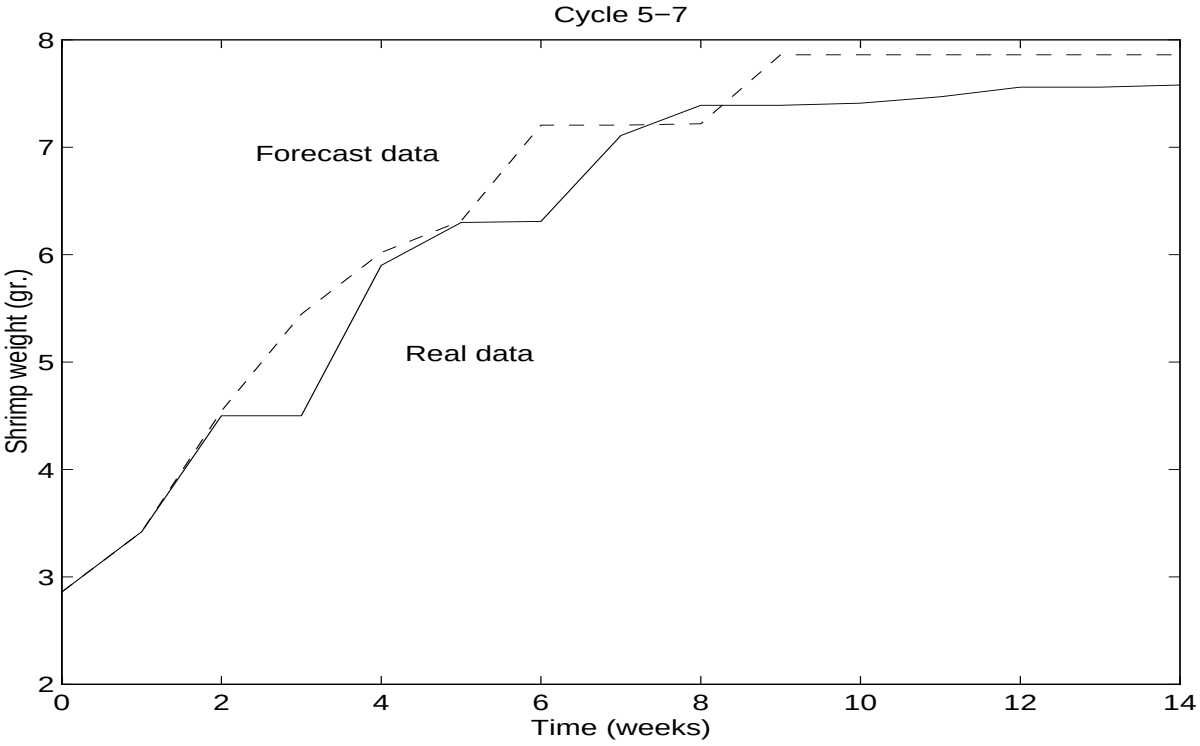

Shrimp weight (gr.)
Time (weeks)
Real data
Forecast data
2
3
4
5
6
7
8
0
2
4
6
8
10
12
14

Table 4: Mean Square Errors for masks of complexities 3, 4 and 5, for two different forecast cycles

Mask	MSE Cycle 1-15	MSE Cycle 5-7
Complex. 3	6.8 %	> 100 %
Complex. 4	22.9 %	10.3 %
Complex. 5	xxx	51.5 %


Real data
Forecast data
Shrimp weight (gr.)
Time (weeks)
0
2
4
6
8
10
12
14
16
18
20


Cycle 5−7
Forecast data
Real data
Shrimp weight (gr.)
Time (weeks)
8
7
6
5
4
3
2
0
2
4
6
8
10
12
14