American Journal of Mathematics and Statistics 2018, 8(2): 36-49
DOI: 10.5923/j.ajms.20180802.02

# Using Fuzzy Logic or Probability Approach in Revising Unknown, Invalid, or Missing Data Points: Application to Shrimp Data Files in the Gulf of Mexico, Years 2005 and 2006

**Morteza Marzjarani**

National Marine Fisheries Service, Southeast Fisheries Science Center, Galveston Laboratory, USA

**Abstract** Probability and fuzzy logic have played important roles in data analysis, prediction, and estimation. Fuzzy logic can help to solve complex problems using approximations and allows users to analyze incomplete and imprecise data sets. Fishery data in general contain missing data points (see for example, [1, 2] or shrimp data files, Gulf of Mexico, 1984 to present). It is of great importance to estimate such points accurately using reliable scientific oriented techniques. The purpose of this study is to introduce the theories of probability and fuzzy logic especially the latter to the shrimp data files for estimating missing/invalid/unknown data points. In the article, these theories along with the statistical mode and multiple imputation were deployed to revise such data points. As an application, shrimp data in the Gulf of Mexico collected by the National Marine Fisheries Service for the years 2005 and 2006 were selected due to the existence of unknown, invalid, or missing values in the *species*, *fathomzone*, and *subarea* fields and the similarity in missing patterns. The methods mentioned above were deployed to revise these fields. The probability approach deployed a discrete multivariate probability distribution developed based on the shrimp data files 2000-2001, statistical mode, and imputation. The fuzzy logic approach also deployed a special form of a Gaussian membership function based on 2000-2001 data files, statistical mode, and imputation. In general, analyses showed that both theories estimated the *species*, *fathomzone*, and *subarea* in a satisfactory manner. However, it was concluded that the fuzzy logic showed more robustness when a large number of data points were to be estimated.

**Keywords** Estimation, Modeling, Probablity, Fuzzy logic

## 1. Introduction

Fuzzy logic is an extension of Aristotelian or Boolean logic (true or false) where partial membership is allowed through membership functions. The idea was formulated by [3, 4], a professor of computer science at Berkley and it became known as the fuzzy logic. Since then, this theory has been addressed by many authors [5-11], for example.

Classical or Boolean logic is good for some applications but does not provide flexibilities needed in some other cases. Fuzzy logic on the other hand, allows solving complex problems using approximations. In some situations, minor approximations could be helpful in reducing the degree of difficulty of the problem. Imagine that you want to park your car in a parking spot. How often do you park your car exactly in the middle of the spot? Theoretically, you are expected to park your car exactly in the middle. However, in practice, a slightly parked car to the left or the right would be acceptable. Therefore, you use this approximation to reduce the degree of difficulty in parking your car. Terms like “heavy”, “light”, and “dark” are called fuzzy terms. You turn on the light in the room if you feel it is too dark. The next person come in and turns off the light since he/she feels that there is enough light in the room.

The fuzzy logic has been used in many areas of science and engineering. Japan heavily makes the use of fuzzy logic in its technology [5]. For example, the Sendai subway speedo train was the first to make the use of fuzzy logic [5, 12].

The fuzzy logic is somewhat similar to the probability theory, but the two theories are not the same. Probability theory takes a conservative approach and does not allow approximations. However, as mentioned above, some problems can be solved in a reasonable time if some approximations are used. In situations like these, fuzzy logic can be very helpful.

* Corresponding author:
morteza.marzjarani@noaa.gov (Morteza Marzjarani)
Published online at http://journal.sapub.org/ajms

For handling missing data, [13] used a statistical method known as "Imputation." For invalid or unknown data points, one could assume that these points are missing and estimate them using the above method. However, when dealing with categorical variables, alternative methods should be deployed [14, 15]. In this research, two of these alternatives were presented and applied to estimate some missing, invalid, or unknown data points in shrimp data files 2005 and 2006. The first alternative was a probabilistic approach, which was built on some variables of interest. The second alternative used here was known as fuzzy logic. In addition, since neither theory was capable of estimating all the missing, invalid, or unknown data points, for the purpose of completing the estimation process, the statistical mode, and another probabilistic approach, namely multiple imputation were used to handle the remaining such data points.

**Method**

The National Marine Fisheries Service (NMFS) is responsible for shrimp effort estimation in the Gulf of Mexico (GOM). NMFS port agents and state trip tickets record the daily operations and shrimp production of the commercial fisheries fleet operating within the boundaries of the U.S. GOM [16]. For assigning fishing activity to a specific geographical location, scientists have subdivided the continental shelf of U.S. Gulf of Mexico into 21 statistical *subarea*s [17]. Subareas 1-9 represent areas off the west coast of Florida, 10-12 represent Alabama/Mississippi, 13-17 represent Louisiana, and 18-21 are designated to Texas (Figure 1). These subareas are further subdivided into 5-fathom depth increments from the shoreline out to 50 fathoms [18]. These divisions are used by port agents and the state trip ticket system to assign the location of catches and fishing effort expended by the shrimp fleet on a trip-by-trip basis [16]. The shrimp data files included several fields of interest to this study. Table 1 gives the fields used in this research and the corresponding descriptions.

For this study, the shrimp data files 2005 and 2006 were selected along with the shrimp data files 2000-2001, the latter two as the basis for developing the probability distributions and the fuzzy logic memberships for estimating missing, invalid, or unknown data points in 2005 and 2006 data files. The 2000-2001 shrimp data files were selected since there were no missing data points in the fields under study for estimation. Each data file consisted of several fields of interest to this study. Table 1 gives the fields used in this research (not in a specific order) and the corresponding description for each.

The shrimp data files contained two fields, size1 and size2 representing the lower and upper boundaries used for the shrimp classification. The algorithm in the Appendix converts these into one categorical variable, *size*, used in this study.

This study focused only on the offshore data in the northern portion of the GOM. Since the shrimp data files contained both inshore and offshore data, the first step was to identify and then remove the inshore data from these files. The field *shore* contained a 1 (offshore) or 2 (inshore) data points. All records with a *shore* code 2 in this field were deleted from shrimp data files 2000-2001 and 2005- 2006.

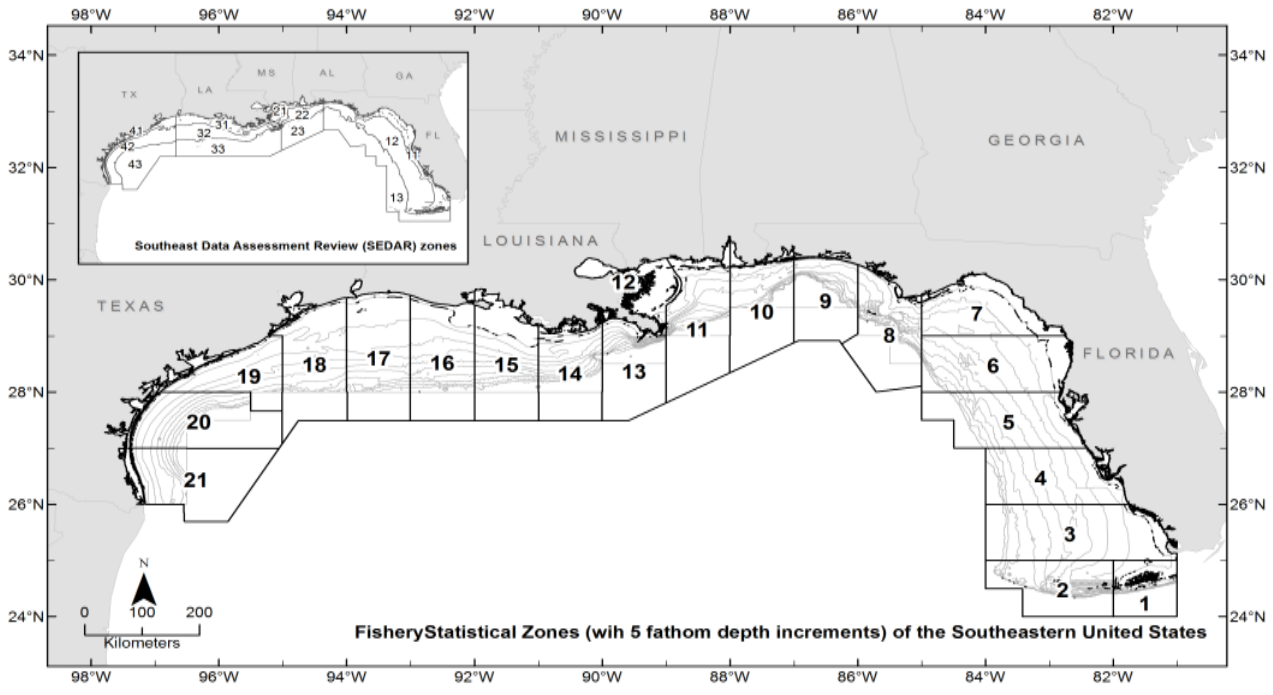


**Figure 1.** The Gulf of Mexico is divided into twenty-one statistical *subarea*s (1-21) as shown

**Table 1.** Description of fields in the shrimp data file used in this research

| Field name | Description |
|---|---|
| *Port* | The shrimp port of landing |
| *vessel id* | US Coast Guard vessel identification number |
| *yearU, monthU, dayU* | Date of unloading shrimp at a designated port. The concatenation of these three was generated and called *edate* |
| *subarea* | Division of the GOM into 21 statistical *subareas* (1 to 9, 10 to 12, 13 to 17, and 18 to 21) |
| *fathomzone* | *Depth* of water where the shrimp was caught (1 to 2, 3 to 6, and 7 to 12 fathoms) |
| *daysfished* | Based on 24 hours/day fished. Included all interviewed by port agents and non-interviewed records |
| *species* | 1=Brown, 2=Pink, 3=White, 4=Seabobs, 5=Royal red, 8=Rock, 9=Trachypenaus |
| *pounds* | Pounds of shrimp harvested |
| *priceppnd* | Average real price per pound of shrimp in 2005 or 2006 (Not adjusted for inflation) |
| *shore* | Code for inshore or offshore (2=inshore, 1=offshore) |
| *size* | Code for size1 and size2 fields (see Appendix for more details)* |
| *geartype* | Type of gear used during fishing trip |

In order to prepare the data files 2005 and 2006 for subsequent applications such as shrimp effort estimation, some issues related to the 2005 and 2006 shrimp files had to be addressed. A number of records in these files had 0's (invalid) in the *subarea* field and unknown (99) in the *fathomzone* field. Table 2 displays the number of *subarea* data points with 0 values and unknown *fathomzone* values (99). This table also includes the total number of offshore records in the shrimp data files 2005 and 2006 and total number of *species* 0, 4, 5, 8, and 9. Table 3 displays the percentages of *species*, *fathomzone*, and *subarea* with respect to the corresponding total data points in the shrimp 2005 and 2006 data files. It is of interest to note that the percentage of the unknown *fathomzone* (99) was relatively high.

**Table 2.** Number of invalid/unknown/missing and total offshore records in shrimp data files 2005 and 2006

| Year | *Subarea* 0 | *Fathomzone* 99 | *Species* 0,4,5,8,9 | Total offshore records |
|---|---|---|---|---|
| 2005 | 6 | 18,179 | 2,564 | 77,755 |
| 2006 | 40 | 9,791 | 1,989 | 71,871 |

**Table 3.** Percentages of *species* 0, 4, 5, 8, 9, *fathomzone* 99, and *subarea* 0 with respect to the total records in shrimp data files 2005 and 2006

| Year | *Subarea 0* | *Fathomzone 99* | *Species* 0,4,5,8,9 |
|---|---|---|---|
| 2005 | 0.01 | 23.38 | 3.30 |
| 2006 | 0.06 | 13.62 | 2.77 |

Since the study was limited to the shrimp fishing in the United States portion of the GOM (that is, statistical *subarea*s 1-21, see Figure 1), all *subarea* data points over 21 were deleted from the shrimp files 2005 and 2006. There were many records with *fathomzone* values 12 or higher (not including those coded as 99). Table 4 was used to replace all the *fathomzone* data points greater than 12 (except those with the unknown code 99) with the corresponding *fathomzone* values listed in columns 1 and 2 of this table. For example, a value between 16 and 20 in the *fathomzone* field was converted to 4 replacing the corresponding value in this field.

**Table 4.** Fathom zones (1-12), fathom, and corresponding *depth* zones (1-3) in the Gulf of Mexico

| Fathom zone | Fathom | *Depth* zone (*depth*) |
|---|---|---|
| 1 | 00-05 | 1 |
| 2 | 06-10 | 1 |
| 3 | 11-15 | 2 |
| 4 | 16-20 | 2 |
| 5 | 21-25 | 2 |
| 6 | 26-30 | 2 |
| 7 | 31-35 | 3 |
| 8 | 36-40 | 3 |
| 9 | 41-45 | 3 |
| 10 | 46-50 | 3 |
| 11 | 51-55 | 3 |
| 12 | >55 | 3 |

Since shrimp *species* 1, 2, and 3 account for about 97% of the Gulf landings of *species* 1, 2, and 3 (See Tables 3 and 5, also http://www.st.nmfs.noaa.gov/commercial-fisheries/commercial-landings/annual-landings/index), all data points with the value 0 or greater than 3 in the *species* field in the shrimp data files 2000-2001 were deleted.

**Table 5.** Frequency (percentage) of *species* 0, 1, 2, 3, 4, 5, 8, and 9 in shrimp data files 2005-2006

| | Year | |
|---|---|---|
| *Species* | 2005 | 2006 |
| 0 | 1,172 (1.51) | 1,350 (1.88) |
| 1 | 43,545 (56.00) | 39,752 (55.31) |
| 2 | 9,385 (12.07) | 6,527 (9.08) |
| 3 | 22,261 (28.63) | 23,603 (32.84) |
| 4 | 260 (0.33) | 206 (0.29) |
| 5 | 46 (0.06) | 44 (0.06) |
| 8 | 1,082 (1.39) | 389 (0.54) |
| 9 | 4 (0.01) | 0 (0.00) |

The combined 2000-2001 offshore data set of *species* 1, 2, 3 consisted of 276,956 records, which provided a large sample for generating the required distributions. Similarly, for the reason of low percentage of *species* other than 1, 2,

or 3, all the data points 0 or greater than 3 in this field in the shrimp data files 2005 and 2006 were assumed missing, but not removed (See Tables 3 and 5). Table 6 shows the results of total pounds before and after removing the invalid/unknown/missing data points in the *species, fathomzone* and *subarea* fields justifying the reason for keeping these records in the corresponding files.

**Table 6.** Reduction in pounds in shrimp data files 2005 and 2006 before and after removing the invalid/unknown data points in the *fathomzone* and *subarea* fields

| Year | Total pounds due to unknown *fathomzone* (99) | Total pounds due to invalid *subarea* (0) | Total pounds due to *species* 0 or greater than 3 | Total reduction in pounds |
|---|---|---|---|---|
| 2005 | 31,451,863 | 4,207 | 1,810,745 | 33,266,815 |
| 2006 | 21,109,349 | 74,863 | 1,393,763 | 22,577,975 |

The shrimp data files 2000 and 2001 were used to determine the multivariate probability distribution functions and the fuzzy membership functions to estimate the unknown *fathomzone* values (99), invalid *subarea* data points (0), and missing *species* data points in the data files 2005 and 2006. The 2000-2001 shrimp files were chosen out of several shrimp data files, because there were no unknown (99) values in their *fathomzone* and no invalid (0) data points in their *subarea* fields.

The following two approaches were deployed to handle some unknown, invalid, or missing data points in the *subarea* and *fathomzone,* and *species* fields. The probability approach consisted of a multivariate probability distribution, the statistical mode (most frequent observation), and imputation. The fuzzy logic approach also consisted of the fuzzy logic, statistical mode, and imputation.

### 1.1. Fuzzy Logic Approach

Unlike classical logic, the most interesting and important feature of fuzzy logic is its ability to handle incomplete, uncertain, imprecise, or vague data sets. In fuzzy logic, the membership function is a continuous function and defines the magnitude of the classical membership (member or not a member). A fuzzy set is a set of elements where the boundaries cannot be defined clearly. To each member of such a set, a membership value is assigned which shows the degree to which the member belongs to the set.

More formally, if A is a fuzzy set and $f_A$ is the associated membership function, then $f_A$ is a mapping from A to the set [0, 1]. That is,

$$f_A: A \rightarrow [0,1] \quad (1)$$

In other words, a fuzzy set is a set of ordered pairs $(x, f_A(x))$ such that:

$$\{(x, f_A(x))| \, x \, \varepsilon \, A\} \quad (2)$$

In the case of classical logic, the symbols [] are replaced with {}. Although the range of this function looks like the range for the probability measure, the membership function is not a probability measure. That is, the sum of mutually exclusive events in the sample space does not have to add up to 1. In other words, from a probability perspective, the degrees of freedom is n-1, but in fuzzy logic, the degrees of freedom is n.

There were several possibilities for the fuzzy membership functions including user-defined functions. In this article, a special case of the Gaussian function with μ and 1 as the location and scale parameters respectively were selected and used.

$$f_A(x:\mu) = e^{-(x-\mu)^2}, \; x \in \mathbb{R}, \quad (3)$$

As an example of a Gaussian fuzzy membership function, assume that the fuzzy set consists of all possible *fathomzone* values. As mentioned earlier, the *fathomzone* is a variable with 3 depth levels, 1 to 2, 3 to 6, and 7 to 12 fathoms. Suppose a vessel is fishing somewhere at a depth between 3 and 6 fathoms (the far left and far right points in Figure 2). Since it is impossible or at the least very hard to measure the precise *depth* of the vessel, fuzzy logic can help with this situation. Assume that the fuzzy membership function represents the depth between an arbitrary point and the mid-point of 3 to 6 fathoms, that is, 4.5 (the mid-point in Figure 2). If the current fishing depth of the vessel is less than 3 fathom, then such depth has a membership function equal to 0 (that is, it is not a member of the set). On the other hand, if the current fishing depth of the vessel is greater than 6 fathom, then such depth has a membership function equal to 1 (that is, it is a member of the set). For values between 3 and 6, the membership values are determined by the points on the curve. Figure 2 is a representation of this example.

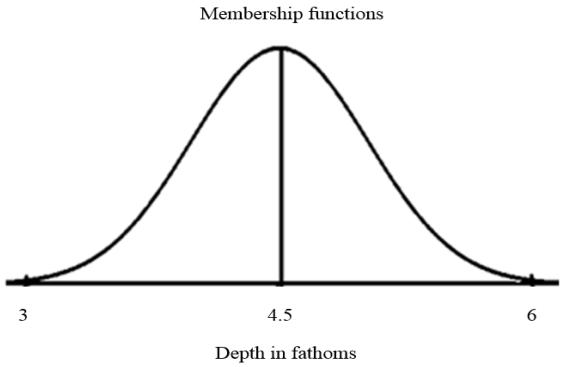


**Figure 2.** An example of fuzzy membership function

In this paper, the fuzzy logic was deployed to revise the missing, invalid, or unknown data points in the *species*, *fathomzone*, and *subarea* fields. The well-known Gaussian curve (3) was selected as the fuzzy membership function in revising the invalid or unknown values in the shrimp 2005 and 2006 data files.

$$f_A(\text{species}:\mu) = e^{-(species-\mu)^2}$$

$$f_A(\text{fathomzone}:\mu) = e^{-(fathomzone-\mu)^2}$$

$$f_A(\text{subarea}:\mu) = e^{-(subarea-\mu)^2}, \quad (4)$$

where μ is the average of the lower and the upper bounds of the *species*, *fathomzone*, or *subarea* subintervals. The next step was to find the maximum (in other words, the most

likely value for the *species*, *fathomzone,* or *subarea*) of the fuzzy membership functions over all possible values of *species*, *fathomzone,* or *subarea* from shrimp data files 2000 and 2001 and estimate/revise these variables.

max $\{f_A(\text{species})|$ species ε all species values in data sets 2000-2001$\}$

max $\{f_A(\text{fathomzone})|$ fathomzone ε all fathomzone values in data sets 2000-2001$\}$

max $\{f_A(\text{subarea})|$ subarea ε all subarea values in data sets 2000-2001$\}$. (5)

To develop the fuzzy membership functions for estimating the data points in the *species*, *fathomzone,* or *subarea* fields, it was assumed that the variables *fathomzone* and *subarea* were each a function of *monthU*, *port, species, size,* and *geartype*. This selection is up to the researcher to modify (if needed). In the case of *species*, the list was the same except the role of this variable was changed to the response variable. That is, the method was applied to the *species* first and then the results were applied to the *fathomzone* and *subarea* fields next. A backward elimination method for *species*, *fathomzone*, and *subarea* was used to select the significant variables contributing to the *species, fathomzone* and *subarea* fields (Adj. R-Squared 0.46, 0.70, and 0.89 respectively). For information on R-Squared, the reader is referred to [19].

The final variables for estimating *species* included *monthU*, *port, size,* and *geartype.* For the variables *fathomzone* and *subarea,* the covariate *species* was added to the list. Of course, as mentioned earlier, the choice of including or excluding a variable in or from this list is at the discretion of the reader and it can easily be modified using alternative lists or different selection method (s).

### 1.2. Probability Approach

Similar to the case of fuzzy logic explained above, the variables *monthU*, *port, species, size,* and *geartype* were selected through the backward elimination method mentioned earlier for generating the multivariate probability distributions for the *fathomzone, subarea,* and *species* (with the exception of *species* where this variable was the dependent variable). Again, it is important to note that the choice of independent variables is completely up to the reader and can be modified using alternative selection method (s). These probability distributions then were used to determine the most likely values for the invalid, unknown or missing data points in these fields.

To develop the probability distributions for either *fathomzone* or *subarea*, suppose $A_i$, *i=1, 2, 3,...* represent all possible values for variables *monthU*, *port, species, size,* and *geartype*. Define the corresponding random variable $X_i$, *i=1, 2, 3, ...* as follows:

$$X_i : A_i \rightarrow \mathbb{R},\ i=1,\ 2,\ 3,\ \ldots. \quad (6)$$

Also, the joint multivariate probability distribution function of $X_i$, *i=1, 2, 3, ...* is defined as:

$$f_A(x_i \mid i= 1, 2, 3,\ldots) = P(X_i = x_i),\ i= 1,2,3,\ldots. \quad (7)$$

The maximum over all possible values of the variables *monthU*, *port, species, size,* and *geartype* was used to revise the invalid *subarea* or unknown *fathomzone* data points.

$$\{\max\{ f_A( x_i )\} \ |X_i^{-1}(x_i) \ \epsilon\ (A_i),\ i = 1,2,3,\ldots\}. \quad (8)$$

For the variable *species*, formulas (6) through (8) can easily be modified by removing the *species* from the list of independent variables and making it a dependent variable.

The shrimp data files 2000-2001 were used to generate the multivariate probability distributions and fuzzy logic functions for *species, fathomzone,* and *subarea*.

### 1.3. Handling *Species*

Throughout this article, software packages such as MS Excel [(1)], SAS [(1)], and R [(1)] were utilized. The multivariate probability and the fuzzy logic approaches described earlier were developed based on the shrimp files 2000 and 2001 and then used to estimate as many as missing *species* (0 or greater than 3 as stated previously) as possible in shrimp data files 2005 and 2006. For the remaining missing data points in this field, the statistical modes of the existing data points in the data files 2005 and 2006 for each triple (*vessel*, *edate*, *port*) in the *species* field were calculated and used to revise as many data points as possible. The argument for determining mode based on the triple (*vessel*, *edate*, *port*) was the fact that this triple was used to reorganize records in the shrimp data files as trips [13]. All the remaining missing *species* (if any) were estimated using a monotone imputation method [20].

### 1.4. Handling *Fathomzone*

The multivariate probability distributions mentioned above were developed based on the shrimp data files 2000-2001. As mentioned earlier, Table 4 represents the interval boundaries for the *fathomzone field* (middle column). This field involves measurements and prone to errors. Even, with the use of electronic devices such possibility of errors exists. For example, a slight move or shake of the vessel by nature or otherwise, would cause the electronic device to misread a measurement. As appeared in [17], there are possibilities of errors in delineating statistical *subarea*s and *depth* contour and possible errors in the planimeter measurements. Therefore, the fuzzy logic seemed an appropriate alternative method for revising an unknown value (99) in the *fathomzone* field. Upon the application of either probability or fuzzy logic, the modes of the existing *fathomzone* values based on the triple (*vessel*, *edate*, *port*) in 2005 and 2006 shrimp files were used to replace the unknown *fathomzone* data points where possible. In either case of probability or fuzzy logic, all the remaining unknown *fathomzone* data points (if any) were estimated using a monotone imputation method [13]. For a couple of reasons the imputation method was deployed last. First, the variables in question were categorical and the use of imputation in such case (s) should be minimized [14, 15]. Second, the

imputation method if applied, would have estimated all the missing data points with no need for an additional estimation method (s).

### 1.5. Handling *Subarea*

The only issue with this field was a limited number of invalid records with 0 in the *subarea* field (see Table 2). Similar to the case of unknown *fathomzone* values, the probabilities generated via the multivariate distribution for *subarea*, fuzzy membership functions in the fuzzy logic approach, the mode of the existing *subarea* data points, and imputation (if either of the latter two needed) were used to replace the 0's in this field.

**Table 7.** Conversion of *subarea* field in the Shrimp file to *area*

| Subarea | Area |
|---|---|
| 1 through 9 | 1 |
| 10 through 12 | 2 |
| 13 through 17 | 3 |
| 18 through 21 | 4 |

Upon the completion of the steps listed above, the *species, subarea*, and *fathomzone* fields in each shrimp 2005 and 2006 data file satisfied the following relations.

$$\text{species} \in \{sp \mid sp \in \{1, 2, 3\}\}$$
$$\text{fathomzone} \in \{fz \mid fz \in \{1, 2, 3, \ldots, 12\}\}$$
$$\text{subarea} \in \{sub \mid sub \in \{1, 2, 3, \ldots, 21\}\} \quad (9)$$

In the next step using Table 4 (Columns 1 and 3) and Table 7, *fathomzone* and *subarea* fields were converted to two categorical variables *depth* (a categorical variable with 3 levels as $depth_1$, $depth_2$ and $depth_3$) and *area* (with 4 levels as $area_1$, $area_2$, $area_3$, and $area_4$), respectively. As appeared in [13], these categorical variables were used in the model estimating the shrimp effort in the GOM.

## 2. Analysis and Results

To measure the effectiveness of the probability and fuzzy logic approaches in revising unknown, invalid, or missing data points in the shrimp data files, analysis was performed on the *species*, *fathomzone*, and *subarea* in the shrimp 2005 and 2006 data files.

### 2.1. *Species*

As mentioned earlier, the probability distribution, the fuzzy logic, and statistical mode or imputation (where needed) were used to estimate the missing data points in the *species* field. Tables 8 through 11 display some statistics of interest to this study regarding this variable. The hypothesis:

$$H_0: p_{prob}\text{-}p_{logic}=0 \text{ vs. } H_a: p_{prob}\text{-}p_{logic} \neq 0 \quad (10)$$

was set up to compare the proportions of *species* assigned by the probability ($p_{prob}$) or fuzzy logic ($p_{logic}$) approach to this variables. Tables 9 and 11 display the Test-statistics, the 95% confidence interval, the *p-value* and the power of each test. In the case of failing to reject the null hypothesis, there was simply not enough evidence to support the alternative. The equivalency of the probability and fuzzy logic approaches was justified when applied to the variable *species* with a moderate number of missing data points.

### 2.2. *Fathomzone*

The number of unknown *fathomzone* codes (99) was relatively high (over 23%) compared to the number of invalid *subarea* or missing *species* data points. The coefficient of variation (CV) for columns 3 and 4 in Tables 12 and 13 were computed and shown below.

$$CV_{\text{Table12, Column 3}}=85\% \qquad CV_{\text{Table12, Column 4}}=84\%$$
$$CV_{\text{Table12, Column 5}}=228\% \qquad CV_{\text{Table12, Column 6}}=184\%$$
$$CV_{\text{Table13, Column 3}}=91\% \qquad CV_{\text{Table13, Column 4}}=88\%$$
$$CV_{\text{Table13, Column 5}}=173\% \qquad CV_{\text{Table13, Column 6}}=136\% \quad (11)$$

As the above CVs show, it can be concluded that the fuzzy logic allocated the unknown data points to different levels of the *fathomzone* more uniformly (that is, the variation was less). Table 14 summarizes the results of the hypotheses of the equality of the corresponding proportions. When combined with the existing data points, the percentage of the times the null hypothesis could not be rejected in 2005 and 2006 were 67% and 58% respectively. That is, more than half the times, we had to assume that the proportions were the same. In other words, we could say that in the case of the *fathomzone* with a relatively large number of unknown data points, the two approaches performed similarly more than they did otherwise, but not completely. Overall, looking at the estimated proportions, the results of the tests were significant meaning that the two approaches performed differently.

Tables 15 and 17 display the distributions of *fathomzone* to the variable *depth* (1 through 3) and *area* (1 through 4) respectively.

In Table 16, the analysis showed that there was no significant difference between the probability and fuzzy logic proportions at $depth_3$. In Table 18, there was no significant difference between the probability and fuzzy logic proportions allocated to the variable *area*.

### 2.3. *Subarea*

Analysis was performed on the *subarea* and the results are listed in Tables 19 through 21. Here, the number of invalid data points was the lowest compared to the variables *species* and *fathomzone*. As expected (due low a low percentage) and displayed in Table 21, in either 2005 or 2006 data file, there were no significant differences between the corresponding proportions.

### 2.4. Summary

Tables 22 through 25 display the overall distributions of missing, unknown, or invalid *species*, *fathomzone*, and

*subarea* allocated to *depth* and *area* along with the corresponding analyses. In Table 23, the only significant difference was observed in 2006 in the *subarea* field at $area_1$ and $area_2$. In Table 25, the impact of *fathomzone* was significant meaning that the allocations of estimated *fathomzone* to the variable *depth* were different. It can be concluded that from this perspective, the two theories performed differently.

**Table 8.** Actual and estimated frequencies of *species* in the 2005 shrimp data file

| Species | Actual Frequency* | Total of species 1, 2, 3 under probability approach** | Total of species 1, 2, 3 under fuzzy logic approach** | Frequency of estimated species data points-probability** | Frequency of estimated species data points-fuzzy logic** |
|---|---|---|---|---|---|
| 0 | 1172 | | | | |
| 1 | 43,545 | 44,445 | 44,427 | 524 | 505 |
| 2 | 9,385 | 10,592 | 10,581 | 534 | 524 |
| 3 | 22,261 | 22,718 | 22,747 | 196 | 225 |
| Subtotal (1,2,3) | 75,191 | 77,755 | 77,755 | | |
| 4 | 260 | | | | |
| 5 | 46 | | | | |
| 8 | 1082 | | | | |
| 9 | 4 | | | | |
| Subtotal (0,4,5,8,9) | 2,564 | | | | |

*Actual numbers **Estimated numbers

**Table 9.** Test-stat, power, *p-value*, and 95% confidence interval on the equality of the proportions of total and estimated *species* by the probability and fuzzy logic approaches (2005)

| | Total *species* (Existing+ estimated) | | | | | Estimated | |
|---|---|---|---|---|---|---|---|
| | Test-stat | *p-value* | 95% Confidence | Interval | Power | Test-stat+ | *p-value*+ |
| | | | Lower bound | Upper bound | | | |
| *Species* 1 | 0.09 | 0.92651 | -0.005 | 0.005 | 0.051 | 0.77 | 0.440 |
| *Species* 2 | 0.08 | 0.93517 | -0.003 | 0.004 | 0.051 | 0.40 | 0.686 |
| *Species* 3 | -0.16 | 0.87156 | -0.005 | 0.004 | 0.053 | -1.55 | 0.121 |

+ : For the estimated *species* only, test-stat and *p-value* are given.

**Table 10.** Actual and estimated frequencies of *species* in the 2006 shrimp data file

| Species | Actual Frequency* | Total of species 1, 2, 3 under probability approach** | Frequency estimated under fuzzy logic** | Frequency of estimated species data points-probability** | Frequency of estimated species data points-fuzzy logic** |
|---|---|---|---|---|---|
| 0 | 1,350 | | | | |
| 1 | 39,752 | 40,303 | 40,295 | 172 | 165 |
| 2 | 6,527 | 7,581 | 7,551 | 311 | 281 |
| 3 | 23,603 | 23,987 | 24,025 | 114 | 151 |
| Subtotal (1,2,3) | 69,882 | 71,871 | 71,871 | | |
| 4 | 206 | | | | |
| 5 | 44 | | | | |
| 8 | 0 | | | | |
| 9 | 389 | | | | |
| Subtotal (0,4,5,8,9) | 1,989 | | | | |

*Actual numbers **Estimated numbers

**Table 11.** Test-stat, power, *p-value*, and 95% confidence interval on the equality of the proportions of total and estimated *species* by the probability and fuzzy logic approaches (2005)

| | Total *species* (Existing+ estimated) | | | | | Estimated | |
|---|---|---|---|---|---|---|---|
| | Test-stat | *p-value* | 95% Confidence interval | | Power | Test-stat+ | *p-value*+ |
| | | | Lower bound | Upper bound | | | |
| *Species* 1 | -0.18 | 0.86078 | -0.006 | 0.005 | 0.054 | 0.45 | 0.665 |
| *Species* 2 | -0.48 | 0.63058 | -0.004 | 0.002 | 0.007 | 1.74 | 0.082 |
| *Species* 3 | 0.50 | 0.61864 | -0.004 | 0.006 | 0.079 | 2.58 | 0.009 |

+ : For the estimated *species* only Test-stat and *p-value* are given.

**Table 12.** Actual and estimated frequencies of *fathomzone* in the 2005 shrimp data file

| Fathomzone | Frequency* | Frequency after estimation under probability approach** | Frequency after estimation under fuzzy logic approach** | Frequency of estimated fathomzone data points-probability** | Frequency of estimated fathomzone data points-fuzzy logic** |
|---|---|---|---|---|---|
| 1 | 3,785 | 14,084 | 12,046 | 10,087 | 8,059 |
| 2 | 7,631 | 9,912 | 10,338 | 1,895 | 2,329 |
| 3 | 10,936 | 11,890 | 13,437 | 450 | 1,993 |
| 4 | 13,000 | 15,829 | 15,723 | 2,248 | 2,114 |
| 5 | 6,877 | 7,602 | 7,796 | 190 | 400 |
| 6 | 5,942 | 6,597 | 6,489 | 266 | 158 |
| 7 | 5,263 | 5,520 | 5,547 | 22 | 53 |
| 8 | 3,658 | 3,777 | 3,812 | 0 | 33 |
| 9 | 1,287 | 1,329 | 1,339 | 0 | 5 |
| 10 | 674 | 689 | 702 | 0 | 14 |
| 11 | 461 | 463 | 463 | 0 | 0 |
| 12 | 62 | 63 | 63 | 0 | 0 |
| Subtotal (1-12): | 59,576 | | | | |
| Fathomzone 99: | 18,179 | | | | |

*Actual numbers **Estimated numbers

**Table 13.** Actual and estimated total frequencies of *fathomzone* in the 2006 shrimp data file

| Fathomzone | Frequency* | Frequency after estimation under probability approach** | Frequency after estimation under fuzzy logic approach** | Frequency of estimated fathomzone data points-probability** | Frequency of estimated fathomzone data points-fuzzy logic** |
|---|---|---|---|---|---|
| 1 | 12,363 | 15,914 | 15,125 | 3,318 | 2,530 |
| 2 | 9,413 | 11,717 | 11,437 | 2,190 | 1,914 |
| 3 | 7,356 | 7,729 | 9,115 | 226 | 1,610 |
| 4 | 9,436 | 12,473 | 11,726 | 2,867 | 2,119 |
| 5 | 8,494 | 8,734 | 9,015 | 100 | 382 |
| 6 | 7,203 | 7,378 | 7,404 | 68 | 96 |
| 7 | 3,776 | 3,843 | 3,857 | 2 | 12 |
| 8 | 1,932 | 1,962 | 2,032 | 0 | 69 |
| 9 | 1,222 | 1,233 | 1,259 | 0 | 26 |
| 10 | 684 | 687 | 700 | 0 | 13 |
| 11 | 189 | 189 | 189 | 0 | 0 |
| 12 | 12 | 12 | 12 | 0 | 0 |
| Subtotal (1-12): | 62,080 | 71,871 | 71,871 | | |
| Fathomzone 99: | 9,791 | | | | |

*Actual numbers **Estimated number

**Table 14.** Results of comparing proportions of total *fathomzone* data points for the years 2005-2006

| | | Excising + estimated | | Estimated | |
|---|---|---|---|---|---|
| Year | *Fathomzone* | Significant difference ($p$-*value* $\leq 0.05$) at *fathomzone* | Non-significant difference ($p$-*value* $> 0.05$) at *fathomzone* | Significant$^{+}$ effects | $p$-*value*$^{+}$ |
| 2005 | 1 through 12 | 1, 2, 3, 5 | 4, 6, 7, 8, 9, 10, 11, 12 | All (11, 12 -no report) | All $\leq 0.05$ (11, 12 -no report) |
| 2006 | 1 through 12 | 1, 2, 3, 4, 5 | 6, 7, 8, 9, 10, 11, 12 | All (11, 12 –no report) | All $\leq 0.05$ (11, 12-no report) |

+: For the estimated *fathomzone*, only Test-stat and *p-value* are given.

**Table 15.** Distribution of total *fathomzone* data points under probability or fuzzy logic approaches allocated to the variable *depth* for the years 2005 and 2006

| Year | *Depth* | Probability approach | Fuzzy logic approach |
|---|---|---|---|
| 2005 | 1 | 23,996 | 22,384 |
| | 2 | 41,918 | 43,445 |
| | 3 | 11,841 | 11,926 |
| 2006 | 1 | 27,631 | 26,562 |
| | 2 | 36,314 | 37,260 |
| | 3 | 7,926 | 8,049 |

**Table 16.** Results of comparing proportions of total *fathomzone* data points allocated to the variable *depth* for the years 2005-2006

| | | Test-stat | *p-value* | 95% Confidence interval | | Power |
|---|---|---|---|---|---|---|
| Year | *depth* | | | Lower bound | Upper bound | |
| 2005 | 1 | 8.94 | 0.000 | 0.016 | 0.025 | 1.000 |
| | 2 | -7.78 | 0.000 | -0.025 | -0.015 | 1.000 |
| | 3 | -0.60 | 0.549 | -0.005 | 0.002 | 0.092 |
| 2006 | 1 | 5.82 | 0.000 | 0.010 | 0.020 | 1.000 |
| | 2 | -4.99 | 0.000 | -0.018 | -0.008 | 0.999 |
| | 3 | -1.03 | 0.302 | -0.005 | 0.002 | 0.178 |

**Table 17.** Distribution of total *fathomzone* data points under probability or fuzzy logic approaches allocated to the variable *area* for years 2005 and 2006

| Year | *Area* | Probability approach | Fuzzy logic approach |
|---|---|---|---|
| 2005 | 1 | 13,864 | 13,866 |
| | 2 | 5,233 | 5,231 |
| | 3 | 28,030 | 28,030 |
| | 4 | 30,628 | 30,628 |
| 2006 | 1 | 9,391 | 9,401 |
| | 2 | 6,225 | 6,215 |
| | 3 | 27,337 | 27,337 |
| | 4 | 28,918 | 28,918 |

**Table 18.** Results of comparing proportions of total *fathomzone* data points allocated to the variable *area* for the years 2005-2006

| | | Test-stat | *p-value* | 95% Confidence interval | | Power |
|---|---|---|---|---|---|---|
| Year | *area* | | | Lower bound | Upper bound | |
| 2005 | 1 | -0.01 | 0.989 | -0.004 | 0.004 | 0.050 |
| | 2 | 0.02 | 0.984 | -0.002 | 0.003 | 0.050 |
| | 3 | 0.00 | 1.000 | -0.005 | 0.005 | 0.050 |
| | 4 | 0.00 | 1.000 | -0.005 | 0.005 | 0.050 |
| 2006 | 1 | -0.08 | 0.938 | -0.004 | 0.003 | 0.051 |
| | 2 | 0.09 | 0.925 | -0.003 | 0.003 | 0.051 |
| | 3 | 0.00 | 1.000 | -0.005 | 0.005 | 0.050 |
| | 4 | 0.00 | 1.000 | -0.005 | 0.005 | 0.050 |

**Table 19.** Actual and estimated frequencies of *subarea* in the 2005 shrimp data file

| subarea | Frequency* | Frequency after estimation under probability approach** | Frequency after estimation under fuzzy logic approach** | Frequency of estimated subarea data points-probability | Frequency of estimated subarea points-fuzzy logic |
|---|---|---|---|---|---|
| 0 | 6 | | | 0 | 3 |
| 1 | 289 | 289 | 292 | 0 | 1 |
| 2 | 5,064 | 5,064 | 5,065 | 0 | 0 |
| 3 | 377 | 377 | 377 | 0 | 0 |
| 4 | 859 | 859 | 859 | 0 | 0 |
| 5 | 694 | 695 | 694 | 0 | 0 |
| 6 | 2,230 | 2,230 | 2,230 | 0 | 0 |
| 7 | 3,412 | 3,412 | 3,412 | 0 | 0 |
| 8 | 775 | 775 | 775 | 0 | 0 |

| subarea | Frequency* | Frequency after estimation under probability approach** | Frequency after estimation under fuzzy logic approach** | Frequency of estimated subarea data points-probability | Frequency of estimated subarea points-fuzzy logic |
|---|---|---|---|---|---|
| 9 | 161 | 163 | 162 | 0 | 0 |
| 10 | 1,108 | 1,108 | 1,108 | 0 | 0 |
| 11 | 3,475 | 3,477 | 3,475 | 0 | 0 |
| 12 | 647 | 648 | 648 | 0 | 0 |
| 13 | 4,819 | 4,819 | 4,819 | 0 | 0 |
| 14 | 5,771 | 5,771 | 5,771 | 0 | 0 |
| 15 | 5,576 | 5,576 | 5,576 | 0 | 0 |
| 16 | 4,050 | 4,050 | 4,050 | 0 | 0 |
| 17 | 7,814 | 7,814 | 7,814 | 0 | 0 |
| 18 | 4,193 | 4,193 | 4,193 | 0 | 0 |
| 19 | 8,231 | 8,231 | 8,231 | 0 | 0 |
| 20 | 5,580 | 5,580 | 5,580 | 0 | 0 |
| 21 | 12,624 | 12,624 | 12,624 | 0 | 0 |
| Subtotal (1-21) | 77,749 | 77,755 | 77,755 | | |

*Actual numbers **Estimated numbers

**Table 20.** Actual and estimated frequencies of *subarea* in the 2006 shrimp data file

| Subarea | Frequency* | Frequency after estimation under probability approach** | Frequency after estimation under fuzzy logic approach** | Frequency of estimated subarea data points-probability | Frequency of estimated subarea data points-fuzzy logic |
|---|---|---|---|---|---|
| 0 | 40 | | | | |
| 1 | 180 | 180 | 180 | 0 | 4 |
| 2 | 3,477 | 3,478 | 3,478 | 0 | 6 |
| 3 | 372 | 373 | 373 | 0 | 2 |
| 4 | 963 | 963 | 963 | 0 | 2 |
| 5 | 229 | 229 | 229 | 0 | 0 |
| 6 | 1,699 | 1,701 | 1,701 | 0 | 0 |
| 7 | 1,698 | 1,701 | 1,701 | 0 | 0 |
| 8 | 446 | 448 | 448 | 0 | 0 |
| 9 | 316 | 318 | 318 | 0 | 0 |
| 10 | 950 | 957 | 957 | 0 | 0 |
| 11 | 4,483 | 4489 | 4489 | 0 | 0 |
| 12 | 774 | 779 | 779 | 0 | 0 |
| 13 | 5,041 | 5,042 | 5,042 | 0 | 0 |
| 14 | 4,977 | 4,983 | 4,983 | 0 | 0 |
| 15 | 5,446 | 5,450 | 5,450 | 0 | 0 |
| 16 | 3,212 | 3,212 | 3,212 | 0 | 0 |
| 17 | 8,650 | 8,650 | 8,650 | 0 | 0 |
| 18 | 4,101 | 4,101 | 4,101 | 0 | 0 |
| 19 | 6,908 | 6,908 | 6,908 | 0 | 0 |
| 20 | 6,856 | 6,856 | 6,856 | 0 | 0 |
| 21 | 11,053 | 11,053 | 11,053 | 0 | 0 |
| Subtotal (1-21) | 71,831 | 71,871 | 71,871 | | |

*Actual numbers **Estimated numbers

**Table 21.** Results of comparing proportions of total *subarea* data points allocated to the variable *subarea* for the years 2005-2006

| | | Existing + | Estimated | Estimated | |
|---|---|---|---|---|---|
| Year | *Subarea* | Significant difference ($p\text{-value} \leq 0.05$) at *fathomzone* | Non-significant difference ($p\text{-value} > 0.05$) at *fathomzone* | Test-stat$^{+}$ | *p-value*$^{+}$ |
| 2005 | 1 through 21 | | All non-significant | No report | No report |
| 2006 | 1 through 21 | | All non-significant | No report | No report |

+ : For the estimated *subarea* only Test-stat and *p-value* are given.

**Table 22.** Frequencies of invalid/unknown/missing *species*, *fathomzone*, and *subarea* by approach for the years 2005 and 2006 allocated to the variable *area*

| Year | Approach | | | | | | |
|---|---|---|---|---|---|---|---|
| | | Probability | | | Fuzzy logic | | |
| | *area* | *species* | *fathomzone* | *subarea* | *species* | *fathomzone* | *subarea* |
| 2005 | 1 | 1,855 | 618 | 3 | 1,855 | 620 | 5 |
| | 2 | 70 | 2,732 | 3 | 70 | 2,730 | 1 |
| | 3 | 372 | 14,484 | 0 | 372 | 14,484 | 0 |
| | 4 | 267 | 345 | 0 | 267 | 345 | 0 |
| 2006 | 1 | 1662 | 440 | 11 | 1662 | 446 | 21 |
| | 2 | 49 | 4,376 | 18 | 49 | 4,370 | 8 |
| | 3 | 205 | 4,888 | 11 | 205 | 4,888 | 11 |
| | 4 | 73 | 87 | 0 | 73 | 87 | 0 |

**Table 23.** Results of comparing proportions of missing/unknown/invalid data points allocated to the variable *area* for the years 2005-2006

| Year | Significant difference ($p$-*value* $\leq 0.05$) | Non-significant difference ($p$-*value* $> 0.05$) |
|---|---|---|
| 2005 | | |
| *species* | | All non-significant |
| *fathomzone* | | All non-significant |
| *subarea* | | No value to report |
| 2006 | | |
| *species* | | All non-significant |
| *fathomzone* | | All non-significant |
| *subarea* | $area_1$ ($p$-*value* $\approx 0.02$), $area_2$ ($p$-*value* $\approx 0.01$) | $Area_3$, $area_4$ (No value to report) |

**Table 24.** Frequencies of invalid/unknown/missing *species*, *fathomzone*, and *subarea* by approach for the years 2005 and 2006 allocated to the variable *depth*

| Year | Approach | | | | | | |
|---|---|---|---|---|---|---|---|
| | | Probability | | | Fuzzy logic | | |
| | *depth* | *species* | *Fathomzone* | *Subarea* | *species* | *fathomzone* | *subarea* |
| 2005 | 1 | 1,399 | 12,580 | 4 | 1,425 | 10,835 | 4 |
| | 2 | 1,025 | 5,163 | 2 | 1,002 | 6,370 | 1 |
| | 3 | 140 | 436 | 0 | 137 | 974 | 1 |
| 2006 | 1 | 1,566 | 5,855 | 29 | 1,588 | 4,786 | 29 |
| | 2 | 393 | 3,825 | 10 | 362 | 4,771 | 10 |
| | 3 | 30 | 111 | 1 | 39 | 234 | 1 |

**Table 25.** Results of comparing proportions of missing/unknown/invalid data points allocated to the variable *depth* for the years 2005-2006

| Year | Significant difference ($p$-*value* $\leq 0.05$) | Non-significant difference ($p$-*value* $> 0.05$) |
|---|---|---|
| 2005 | | |
| *species* | | All non-significant |
| *fathomzone* | All significant (Power ≈1) | |
| *subarea* | | All non-significant |
| 2006 | | |
| *species* | | All non-significant |
| *fathomzone* | All significant (Power ≈ 1) | |
| *subarea* | | All non-significant |

## 3. Discussion and Concluding Remarks

The purpose of this paper was to propose two approaches to the fishery data files when there was a need to revise or estimate missing, invalid, or unknown data points in these data files. Clearly, these points play a significant role in the estimation/prediction and care must be taken to assure that a correct method (s) is used. The question of finding the most suitable method is an open-ended one and research for finding a proper and advanced method (s) is necessary. The issue becomes more complex and more challenging (hopefully more accurate) as the more advanced techniques are deployed. All possibilities must be explored, and care must be taken when selecting a method for estimating such points. In this article, the probability and fuzzy logic theories were considered for handling this important issue.

The study focused on the shrimp 2005-2006 data files and the *species* 1, 2, 3, due to their high percentage compared to the other *species*. To maintain all the records in the shrimp data files 2005 and 2006, it was assumed that all *species* data points in these files except 1, 2, and 3 were missing but keeping the corresponding records in the files. Removing such points would have reduced the number of records in these data files, which in turn, would have had an impact on any application of such files such as shrimp effort estimation. Table 6 displays the reduction in pounds if such records were eliminated from the 2005-2006 shrimp data files. For estimating the probability or fuzzy logic membership functions, the *species* 0 or greater than 3 in the data files 2000-2001 were removed still providing a very large sample (276,956 records) and the required distributions were determined based on these records in the data files 2000-2001.

The issues with *fathomzone* and *subarea* in the 2005 and 2006 shrimp data files were the presence of invalid *subarea* (0) and unknown *fathomzone* data points (99). The data files shrimp 2000 and 2001 seemed comparable with these files with the advantage of having no *subarea* (0) data points in the *subarea* filed and no code (99) in the *fathomzone* field. These two files were selected as the basis for generating the multivariate probability distributions for *species*, *fathomzone*, and *subarea* and then used to estimate/revise the missing/invalid/unknown data points in these fields. The readers might choose alternative files for this purpose.

Tables 8 through 11 address the frequencies of *species* (original versus after allocation) and the allocation of this variable to the variables *area* and *depth*. As displayed in Tables 9 and 11, the proportions assigned under the fuzzy logic or the probability approach in both cases of total *species* (existing and estimated), and estimated *species* were relatively close (that is, the differences were not statistically significant).

Tables 12 through 18 address the distributions and allocations of the estimated *fathomzone* data points. As displayed in Tables 14, 16, and 18, the two approaches performed somewhat differently. In Table 14, the results of the tests showed a significant difference in proportions at the lower depths (5 and below) when estimated and existing data points were added. Considering estimated data points only, the proportions were different with no comparisons at depths 11 and 12 in both years 2005 and 2006 due to the 0 frequencies. This implied that the two approaches performed somewhat differently. This might be due to the large number of unknown data points in the field *fathomzone* (Tables 2 or 3). Looking at the CVs given in (11), one could conclude that the fuzzy logic distributed the estimated data points more consistently. This is consistent with the fact that the probability approach is precise, allows no approximation, and requires precise and compete data sets. Fuzzy logic on the other hand, is more flexible in that respect and is built upon the assumption of being able to handle imprecise and incomplete data sets.

As mentioned earlier, analyses showed some sensitivities with respect to the number of data points to be estimated. In the case of *species* with a moderate number of points to be estimated, the two theories performed equivalently (Tables 9 and 11). However, in handling a large number of data points to be estimated, we observed some discrepancies between the two approaches (Tables 14, 16, and 18). Tables 19 through 21 address the estimated data points in the field *subarea*. The analysis showed that the two approaches performed equivalently when applied to the total proportions (existing and estimated). However, due to a very low number of invalid data points in this field, it was difficult to compare the performances of the two theories in this case and it was not something one can heavily emphasize or draw any definite conclusions from these tables. Tossing a fair coin does not necessarily produce the expected 50% heads or invalidates the expected frequency. Overall, for a relatively large number of data points to be estimated, fuzzy logic seems to respond reasonably well.

Tables 22 through 25 provide the allocations of missing/invalid/unknown data points to the variables *area* and *depth*. These two variables play a role in models dealing with shrimp effort estimation [13] or similar applications. In Table 23, one can observe the equivalency of the approaches with a minor discrepancy regarding the *subarea* field even though such conclusions for this field are not conclusive due to a very low number of invalid data points in this field. In Table 25, the proportions of *fathomzone* allocated to the variable *depth* were all statistically different.

Along with the probability and fuzzy logic, two important other statistical methods were deployed in this research, that is, statistical mode and imputation. Statistical mode is a measure of location and represents the most frequent observation in a given data set. Here, following the applications of the probability theory and/or fuzzy logic, this measure was deployed to estimate additional data points. As appeared in [13], a trip was formed based on the triple (vessel id, edate, port). A real example was a case where a vessel was at the port say, A, on May 2, 2005, four times recorded at *fathomzone* 1 and once at *fathomzone* 99 (unknown). From a statistical perspective, it was reasonable to assume that the code 99 was recorded in error and very likely the vessel was

at the *fathomzone* 1. Unfortunately, none of the three methods, the probability, the fuzzy logic and the statistical mode were able to estimate all the missing/invalid/unknown data points. Since the main objective of the research was to compare the probability and fuzzy theories, these two implemented first followed by the statistical mode and imputation to complete the estimation of missing/invalid/ unknown data points. For the obvious reason, imputation was used last. To do otherwise, imputation would have estimated any and all the missing/invalid/unknown data points.

The probability theory was expected to be more sensitive towards the number of data points to be estimated losing robustness as the number of such data points increased. Fuzzy logic also displayed the same, but with less severity. Under a reasonable number of data points to be estimated and roughly speaking, both theories of fuzzy logic and probability performed equivalently in this research. The probability approach would be recommended for its accuracy and precision for cases where a low or moderate number of data points are to be estimated (See Table 4 and discussion on the CVs). For a very large number of such data points, perhaps fuzzy logic is an option due to its flexibility towards handling imprecise or incomplete data sets. For brevity purposes and to avoid extending the scope of this paper further, the issue of "large" was not explored beyond this point in this article. One possibility is to conduct a simulation study and determine the cutoff point for a number of missing data points to be called "large."

Previously, [13] used multiple imputation in handling missing data points where covariates used in the estimation process were continuous. In this article, it was possible to use this method and estimate all the missing, invalid, or unknown data points. The concern about this approach was the fact that the covariates representing these missing data points were all categorical in nature. The issue and concern about applying the imputation method to the categorical variables was addressed in [14, 15]. In this article, the application of this approach was limited to the cases where the other methods would not help with estimating the missing data points.

Both the probability and fuzzy logic have been used in research in one form or another. For example, [21] used the probability theory in their book to address the estimation of parameters in a fishery stock assessment model (Chapter III, pages 157-434). Authors in [22] used fuzzy logic for estimating the parameters of weight-length relationship. In addition, authors in [23] used fuzzy logic to build an expert system to estimate intrinsic extinction vulnerabilities of marine fishes to fishing. Here, these theories were deployed from a different perspective that is, estimating missing/invalid/unknown data points in shrimp data sets. The ideas are wide open and it is entirely up to the readers to modify these approaches. Up to now, no "perfect" method has been proposed for handling missing data points. The imputation method has shown to be a good candidate, but the issue with categorical variables and rounding presents a problem to the researchers. Unfortunately, in a research work of this nature, it is difficult to say that one method is superior to another. Comparisons are all relative, but not absolute. There are always advantages and disadvantages in any approach we select. For example, the probability approach is accurate, precise and allows no approximation. The probability distribution once defined is unique and offers no flexibility in that respect. Fuzzy logic on the other hand, uses approximation and offers flexibility in defining/selecting a proper membership function.

The methods deployed here have been chosen out of many possibilities and by no means have they constituted the best set of options. It is up to the reader to modify the selection of independent variables used to estimate *species*, *fathomzone*, or *subarea* or to replace them with some alternatives. Another possibility is to use a different method (s) for selecting the predictors, which might suggest different sets for different response variables.

The methods for estimating the missing data points as written in (5) and (8), were intentionally defined to be similar so that if there were any differences , it would not have been due to the selection of membership functions. In other words, any discrepancy would mainly have been due to the difference between the two theories and again, not completely due to the selection of fuzzy membership.

In conclusion, the focus of this article was the introduction of the probability and fuzzy logic theories in estimating unknown/invalid/missing data points when dealing with categorical variables. Although the application was limited to the shrimp data, it could easily be extended and applied to other data sets where categorical data need to be estimated.

## ACKNOWLEDGEMENTS

The author would like to thank Mr. James Primrose of NMFS for providing the data sets for this research and Dr. Ryan Kitts-Jensen of NMFS, and an anonymous referee for their excellent editorial comments.

## Disclaimer

The scientific results and conclusions, as well as any views or opinions expressed herein, are those of the author and do not necessarily reflect those of NOAA or the Department of Commerce.

## Appendix

Each original shrimp data file contains two fields called size1 and size2. These variables represent the lower and upper bound for the shrimp size range. The following algorithm was used to calculate the size variable included in this analysis.

```
Begin
calc := ((size1 + size2) / 2) + 0.5;
        if (size1 = 999) and (size2 = 999) or (calc is null)
        then size=0
        else if (size1 = 999) and (size2 = 999) then size=9
        else if (calc < 15) then size=1
        else if (calc >= 15) and (calc <= 20) then size=2
        else if (calc > 20) and (calc <= 25) then size=3
        else if (calc > 25) and (calc <= 30) then size=4
        else if (calc > 30) and (calc <= 40) then size=5
        else if (calc > 40) and (calc <= 50) then size=6
        else if (calc > 50) and (calc <= 67) then size=7
        else
                size=8
end.
```

(Source: Shrimp Oracle Database)

# REFERENCES


[1] Browder, J. A., Restrepo, V. R., Rice, J., Robblee, M. B., Zein-Eldin Z. (1999). Environmental Influences on Potential Recruitment of Pink Shrimp, Farfantepenaeus duorarum, from Florida Bay Nursery Grounds, *Estuaries*, Volume 22, Issue 2, 484–499.

[2] Whitmore, k., Richards, A., Carloni, J., Hunter, M., Hawk, M., Drew K., Jacobson, L., Chen, Y., Jie Cao, J. (2014). ASMFC Northern Shrimp Technical Committee, C. Northern Shrimp Stock Assessment for 2014, pp. 529-784, 58th SAW Assessment Report, Northeast Fisheries Science Center Reference Document 14-04, 529-784.

[3] Zadeh, L. A. (1965). Fuzzy sets. Information and Control, 8 (3), 338–353.

[4] Zadeh, L. A. (1968). Fuzzy algorithms. Information and Control, 12 (2), 94–102.

[5] Pollack, A. (1989). Fuzzy Computer Theory: How to Mimic the Mind? Special to the New York Times. http://www.nytimes.com/1989/04/02/us/fuzzy-computer-theory-how-to-mimic-the-mind.html?pagewanted=all (No page numbers).

[6] Gerla, G. (2016). Comments on some theories of fuzzy computation. International Journal of general Systems, 45 (4), 372–392.

[7] Novák, V. (2005). Are fuzzy sets a reasonable tool for modeling vague phenomena? Fuzzy Sets and Systems, 156, 341–348.

[8] Novák, V., Perfilieva, I. and Močkoř, J. (1999). Mathematical principles of fuzzy logic. Dodrecht: Kluwer Academic, ISBN 0-7923-8595-0.

[9] Pelletier, F. J. (2000). Review of metamathematics of fuzzy logic. The Bulletin of Symbolic Logic. 6 (3), 342–346, JSTOR 421060.

[10] Valliant, L. (2013). Probably Approximately Correct Nature's Algorithms for Learning and Prospering in a Complex World New York: Basic Books, ISBN 978-0465032716.

[11] Zaitsev, D. A., Sarbei, V. G., Sleptsov, A. I. (1998). Synthesis of continuous-valued logic functions defined in tabular form. Cybernetics and Systems Analysis, 34 (2), 190–195.

[12] Smith, E. T. (1993). Why the Japanese are Going in for this 'Fuzzy Logic, Business Week, Feb. 20, 39.

[13] Marzjarani, M. (2016). Higher Dimensional Linear Models: An Application to Shrimp Effort in the Gulf of Mexico, Years 2007-2014. International Journal of Statistics and Application, 6(3), 96-104.

[14] Allison, P. D. (2001). Missing Data. Thousand Oaks, CA, Sage Publications.

[15] Horton N. J., Lipsitz, S. R., and Parzen, M. (2003). A potential for bias when rounding in multiple imputation. American Statistician 57, 229-232.

[16] Hart R. A., Nance, J. M. (2013). Three Decades of U.S. Gulf of Mexico White Shrimp, Litopenaeus setiferus, Commercial Catch Statistics, Marine Fisheries Review, 75 (4), 43-47.

[17] Patella, F. (1975). Water surface area within statistical subareas used in reporting Gulf coast shrimp data. Mar. Fish. Rev. 37(12), 22–24.

[18] Nance, J., Keithly, W., Caillouet, C., Cole, J., Gaidry, W., Gallaway, B., Griffin, W., Hart, R., Travis, M. (2008). Estimation of Effort, Maximum Sustainable Yield, and Maximum Economic Yield in the Shrimp Fishery of the Gulf of Mexico, NOAA Technical Memorandum NMFS-SEFSC-570.

[19] Cameron, A. C., Windmeijer, F. A. G. (1997). An *R*-squared measure of goodness of fit for some Common nonlinear regression models, Journal of Econometrics, Volume 77, Issue 2, April 1997, 329-342.

[20] Yuan, Y. (2011). Multiple Imputation Using SAS Software, Journal of Statistical Software, Vol. 45, Issue 46, 1-25.

[21] Hilborn, R., Wlaters, C. J. (1992). Quantities Fisheries Stock Assessment: Choice, Dynamic, and Uncertainty, Springer, ISBN 978-1-4020-1845-9.

[22] Bitar, S. D., Campos, C. P., Freitas, C. E. C. (2016). Bras. J. Biol. Applying fuzzy logic to estimate the parameters of the weight-length relationship, 611-618.

[23] Cheung, W. L., Pitcher, T. J., Pauly, D. (2005). A fuzzy logic expert system to estimate intrinsic extinction vulnerabilities of marine fishes to fishing, Biological Conservation 124 (2005), 97–111, Springer.