FDEP Order No. BAA752
Florida International University

Kristen Jacobs Coral Reef Ecosystem Conservation Area Water Quality Data Analysis

Final Report
June 8, 2022

Henry Briceño, Joseph N. Boyer, and Ian L. Dryden

Supplier:
Florida International University - SERC Recharge Centers
11200 SW 8 St. OE-148
Miami, FL 33199
United States
Phone: 1-305-348-3095
PI: Henry O Briceño

This report was funded by the Florida Department of Environmental Protection, Office of Resilience and Coastal Protection award No. BAA752. The views, statements, findings, conclusions and recommendations expressed herein are those of the author(s) and do not necessarily reflect the views of the State of Florida or any of its sub agencies.

EXECUTIVE SUMMARY

The Kristin Jacobs Coral Reef Ecosystem Conservation Area (Coral ECA) Water Quality Assessment (WQA) was designed in 2014 by a collaborating body of National Oceanic and Atmospheric Administration (NOAA) scientists, Florida Department of Environmental Protection's Coral Reef Conservation Program (CRCP) staff, and partners from the Southeast Florida Coral Reef Initiative (SEFCRI). The goal of the WQA was to provide data for managers to assess the status of the Coral ECA, an area which historically did not have a consistent water quality monitoring program.

The goal of this data analysis project is to evaluate and prepare the available water quality data of the Coral ECA for future assessments aimed at identifying both the constituents and the impacts of land-based sources of pollution (LBSP) on the Coral ECA. The main objective of the project is to inform resource managers and decision-makers on the status of water quality in the Coral ECA.

The available data, collected from 2017 to 2021, contain many non-detects (ND) due to high and variable laboratory detection limits. In some instances, censoring reaches 100% of some data sets. High percentages of ND pose a challenge to the use of the database and to the interpretation of results, especially in use for the development of nutrient criteria and in determining compliance to those criteria.

A NOAA Technical Memorandum (Whitall et al. 2019), analyzed the Coral ECA water quality data from September 2016 through December 2018. In that report, the authors used the methodology described by Flynn (2010) to populate a time series by replacing the non-detects with imputed "dummy" values. This technique could be questioned because substitution of non-detects with dummy values creates an artificial dataset which may be biased. To avoid this problem, there are other well-known methods such as survival statistics which are deemed to be more appropriate (Helsel 2006, 2010, 2011).

The tasks of the present project are to: 1) Reformat the available dataset because their current WIN/WQX data format is not easily ingested by standard statistical packages; and 2) Calculate

descriptive statistics for the dataset distributions using both Flynn's methodology and Survival Statistics and provide a comparison between results of these approaches. The resulting comparison of five statistical methods was carried out for estimating the mean, median, sd, and inter quartile range when censored data are present. A Monte Carlo simulation study was performed to investigate the performance of the estimators under known distributions.

From these simulations we determined that the Censored Maximum Likelihood Estimation using the Weibull distribution (MLE-W) is the most unbiased and most efficient method when the underlying distributional family is correct. The dummy imputation methods can perform poorly in some cases.

In addition, increasing the level of censoring in the dataset resulted in large positive biases in the dummy estimates in some examples while the Weibull MLE remained stable. Therefore, for highly censored datasets, the Weibull MLE was the best overall estimator.

Future analyses on the reported datasets from this project are aimed to answer the following questions:

- Are there differences in the data between the individual ICA's, and if possible, also compared to land use coefficients?
- Are there differences between site types – inlet vs reef vs outfall samples?
- Is there a significant difference in analyte concentrations between bottom vs surface samples?
- How do the available concentration data compare to any relevant published thresholds, especially to those of southeast Florida waters?

We recommend that the Weibull MLE be used in performing these future analyses, with suitable checks that the distribution is appropriate.

Contents

1 BACKGROUND

The counties of Southeast Florida (Miami Dade, Broward, Palm Beach, and Martin) are highly urbanized, inhabited by 6.29 million people (citation). Most development occurs directly along the coast, and Florida's Coral Reef lies to the east, just 1.5 km from the urbanized shoreline. Therefore, southeast Florida's coral reefs are directly impacted by anthropogenic stressors, especially terrestrial runoff, and from failing wastewater disposal systems, which degrade coastal water quality.

The Kristin Jacobs Coral Reef Ecosystem Conservation Area (Coral ECA) Water Quality Assessment (WQA) was designed in 2014 by a collaborating body of National Oceanic and Atmospheric Administration (NOAA) scientists, Florida Department of Environmental Protection's Coral Reef Conservation Program (CRCP) staff, and partners from the Southeast Florida Coral Reef Initiative (SEFCRI).

Since 2016 a water quality monitoring program was implemented, whose main objectives are:

1) Determining the trends of coastal and offshore water quality in Southeast Florida; and
2) Assessing potential links between water quality and land-based sources of pollution, and/or changes in coral condition.

The focus of this study extends from the St. Lucie Inlet in the north to offshore Biscayne Bay in the south, containing nine major inlets, namely, St. Lucie (STL), Jupiter (JUP), Lake Worth (ILW), Boynton (BOY), Boca Raton (BOC), Hillsboro (HIL), Port Everglades (PEV), Baker's Haulover (BAK), and Government Cut (GOC) (Fig 1).

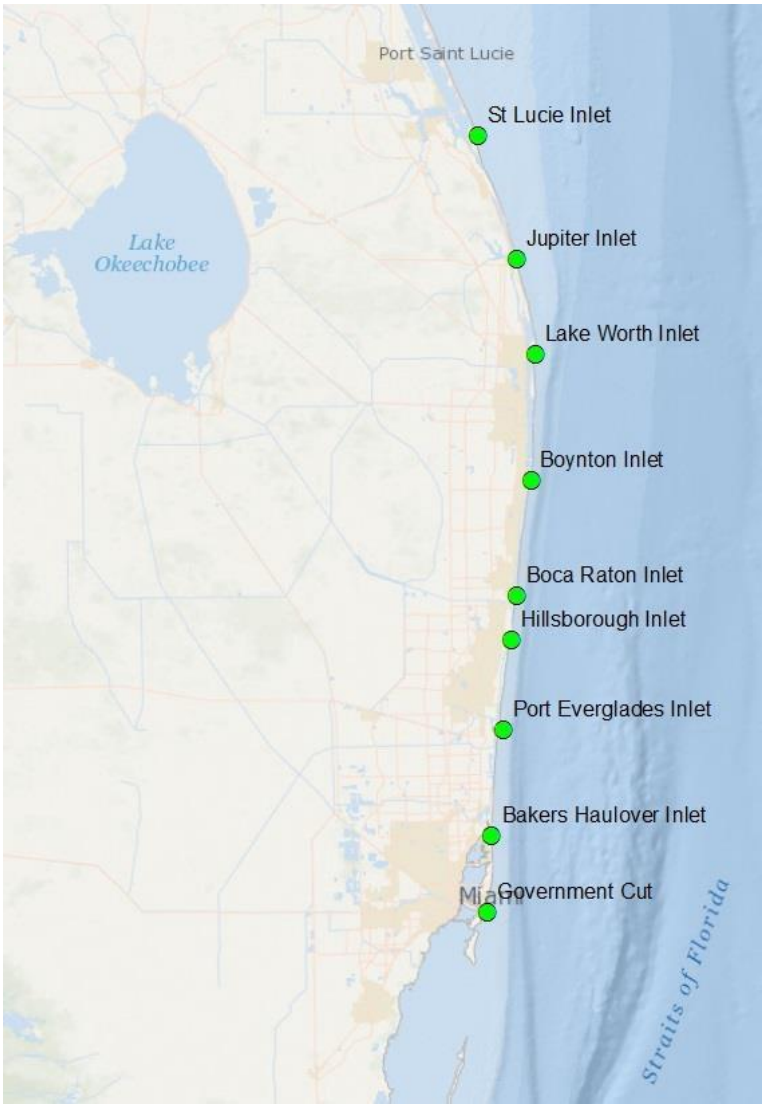


Figure 1. *Map of study area, showing the location of Inlet Contributing Areas (ICAs). From Whitall et al. 2019*

Data gathered from 2017 to 2021, contain many non-detects (ND) due to high and diverse detection limits. In some instances, censoring reaches 100% of the data sets (Fig 2). High percentages of ND pose a challenge to the use of the database and to the interpretation of results. In order to overcome that limitation, Whitall et al. (2019) used Flynn's (2010) estimation method. This method is appropriate to identify potential statistical distributions of the data, and estimate statistics for such distributions, resorting to optimization algorithms.

Well known methods of survival statistics (Helsel 2006, 2010, 2011) have been recommended to

be used instead of Flynn's replacing methodology, mostly when statistical methods for handling censored data and computing statistics are well established in medical and industrial statistics, under survival or reliability analysis (Klein and Moeschberger, 2003; Meeker and Escobar, 1998).

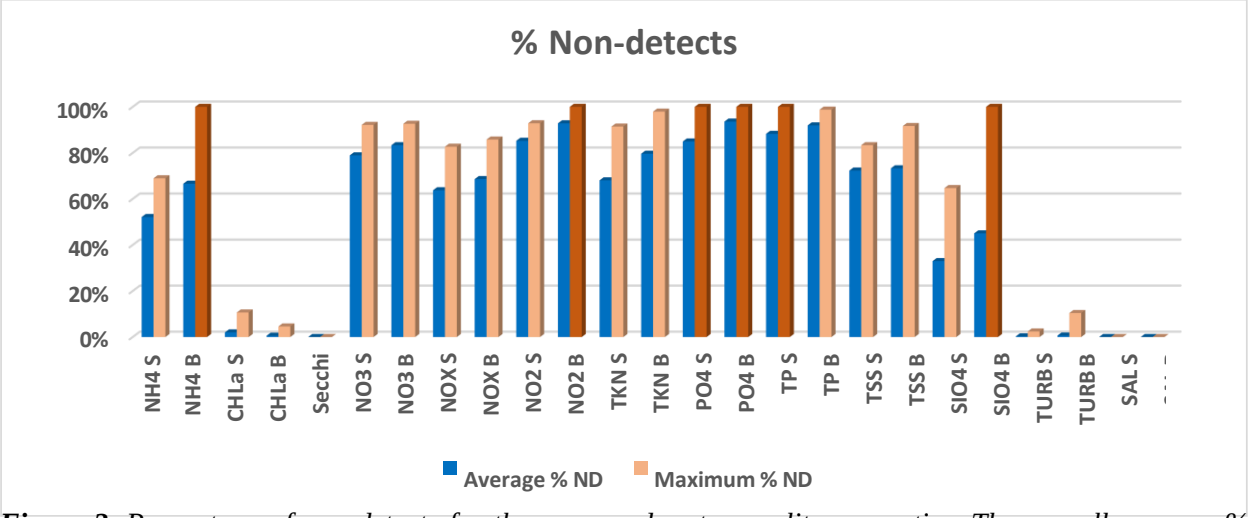


Figure 2. *Percentage of non-detects for the measured water quality properties. The overall average % (blue) is about 53%, and the maximum % (pink) is usually above 80%, and reaches up to 100% (red) in some species.*

2 PROJECT TASKS

This project entails a re-evaluation of methodologies to deal with the existing water quality data, and the calculation of statistical parameters to be used in the assessment of water quality conditions in the Southeast Florida Coastal Zone, along Florida's Coral Reef. For its execution, this project was subdivided into two phases.

Phase 1. The first phase included the following tasks:

Task 1: Reformat and Prepare Data for Analyses

Task 2: Statistical distribution identification and derivation of Survival Statistics for whole dataset

Results from this initial phase were delivered on May 10, 2022.

Phase 2. The second phase of this study included the following tasks:

Task 3: Create "dummy" values for all results below MDL using Flynn Method

Task 4: Summary of outcomes from the above analysis to include figures, graphs and maps

Results from this second phase are the objective of the present report.

3 BROWARD LAB ANALYSIS USING IMPUTATION AND CENSORED MAXIMUM LIKELIHOOD ESTIMATION

3.1 INTRODUCTION

This report describes in brief the methods and analysis of environmental data collected by the Broward County Environmental Monitoring Laboratory gathered between 2017 and 2021. In the first method censored values that are below the Minimum Detection Limit (MDL) are estimated by maximizing the Shapiro-Wilk statistic for normality of the logarithms of the observations. This method imputes dummy values for the censored observations by aiming to make the distributions as log-normal as possible according to the Shapiro-Wilk statistic. The method is described in Flynn (2010, *Ann. Occup. Hyg*., 54, 263–271) and the Shapiro-Wilk values are computed using the approximation of Royston (1992, *Statistics and Computing*, 2, 117-119).

3.2 EXAMPLE: NH4 S AT BAKER'S HAULOVER INLET

In the following example for NH_4 S (surface) at Baker's Inlet there are 80 measured values and 124 censored values. The optimization is carried out in R using simulated annealing, and plots are given in Figure 3.

Sample statistics are then computed using the combined data of the imputed and observed values. In this example the mean is 0.0115, the median is 0.007, the standard deviation is 0.0125 and the interquartile range (IQR) is 0.0108. Note that there are many optimal solutions (due to invariance of reordering of many of the dummy values). However, from Monte Carlo simulations with different starting values the statistics can be reasonably reliably estimated. Out of 100 Monte Carlo simulations for NH_4S at Baker's Inlet the 5% -95% ranges are mean (0.0109-0.0119), median (0.0063-0.007), sd (0.0114-0.0130), IQR (0.0100-0.0124), and Shapiro-Wilk (0.9910-0.9933).
The full set of statistics for all variables and ICAs can be found in the csv file **Broward-imputed-results.csv**, and the plots are in the pdf file **Broward-analysis-imputation-ICA-plots.pdf**.

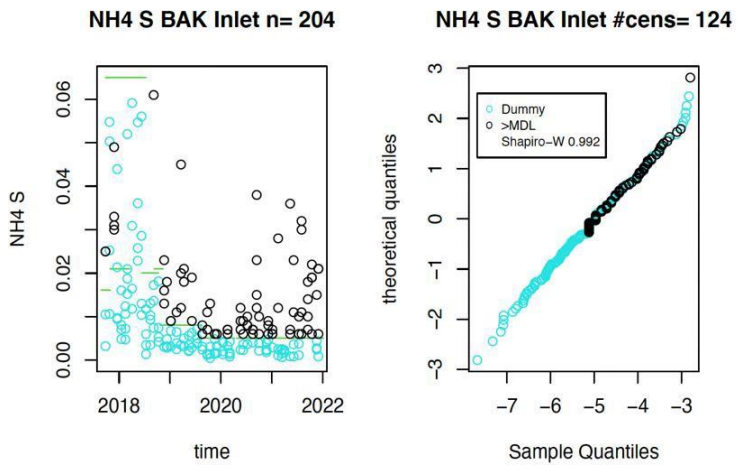


Figure 3. *(left) The observations (black) and the dummies (cyan) obtained by maximizing the Shapiro-Wilk statistic on the logarithms such that the dummies remain below the MDL (green). The aim is to choose dummy values which make the distribution as log-normal as possible. The imputed values are not unique - there are invariances with respect to ordering and there are small numerical differences in different runs due to random starting values. (right) A Q-Q normal plot for the natural logarithm of the data. A straight line indicates that a log-normal distribution would be appropriate. The black points are observations, the cyan ones are the optimized filled-in dummy values.*

3.3 COMPARISON OF METHODS

A comparison of five methods has been carried out for estimating the mean, median, sd, IQR when censored data are present.

1. Dummy L. The log-normal dummy imputation method as described above.
2. Dummy W. A new method of Weibull imputation. Maximization of the square of the correlation of log data with log(-log(1-probability)) in order to fill in dummy values below or equal to the MDL. Hence the observations are filled in according to an optimal Weibull distribution.
3. MLE L. Censored Maximum Likelihood Estimation (MLE) using the log-normal distribution.
4. MLE W. Censored MLE using the Weibull distribution.
5. KM. The nonparametric Kaplan-Meier method

Example plots for NH_4S at Baker's Inlet are given below in Figures 4, 5 and the full set is given in **Broward-analysis-ICA-all-methods.pdf**.

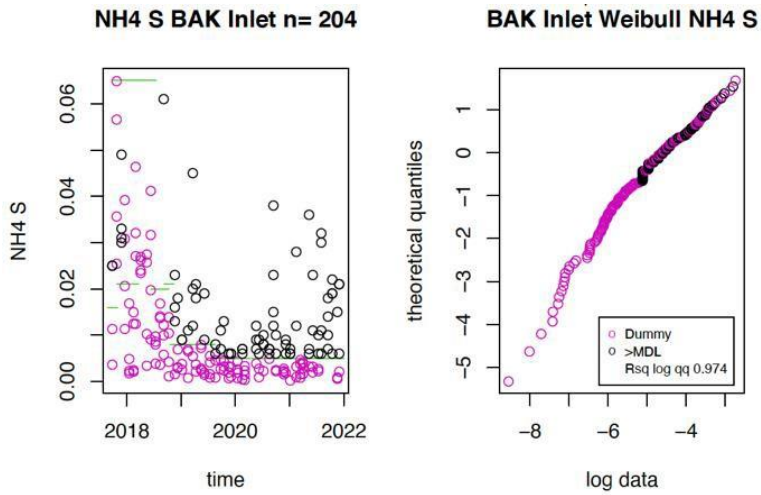
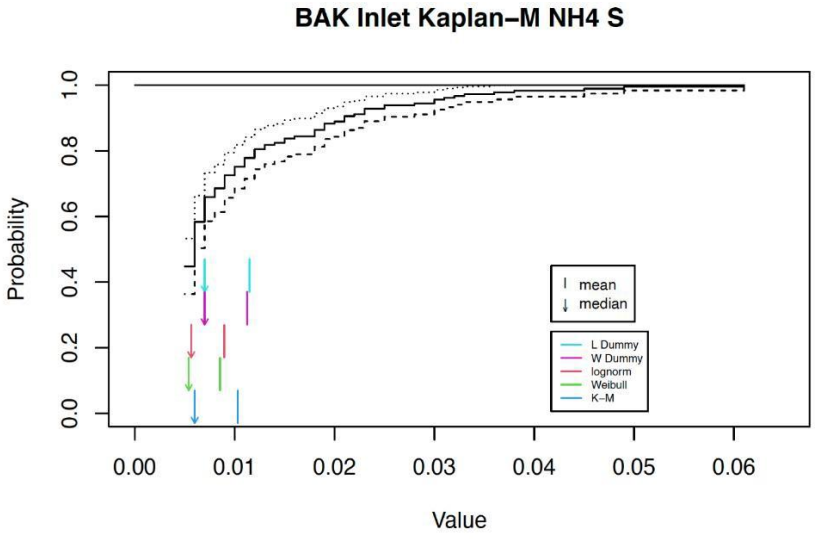


Figure 4. *These plots show: (left) The observations (black) and the Weibull dummies (purple) obtained by maximizing the correlation in the QQ plot such that the dummies remain below the MDL (green). The aim is to choose dummy values which make the distribution as Weibull as possible. As for the Flynn method the imputed values are not unique - there are invariances with respect to ordering and there are small numerical differences in different runs due to random starting values. (right) A QQ plot of the quantiles of the Weibull for the observed (black) and dummy values (purple). This should be as straight as possible.*




```
## l dummies: mean 0.0115 med 0.007 sd 0.0125 IQR= 0.0108
## w dummies: mean 0.0113 med 0.007 sd 0.012 IQR= 0.0121
## lognorm: mean 0.009 med 0.0057 sd 0.011 IQR= 0.0078
## Weibull: shape= 0.9004 scale= 0.0081
## Weibull: mean 0.0085 med 0.0054 sd 0.0095 IQR= 0.0096
## Kaplan-M: mean 0.0103 med 0.006 sd 0.0094 IQR= NA
```

Figure 5. *The Kaplan-Meier estimate with confidence intervals. Also, the medians using all the methods are displayed with colored down-pointing arrows, and the means are given by colored vertical line segments. Some estimates might not be available, especially for the Kaplan-Meier.*

Comparing all methods with Dummy L we consider Figure 6. For Dummy L vs Dummy W: these estimates are very similar indeed, and so whether we use lognormal or Weibull for imputation makes little difference. For Dummy L vs MLE L: there are differences in mean (MLE higher or lower) and the median often being lower using the lognormal MLE. The standard deviation is often larger for the lognormal MLE. For MLE W vs Dummy L the Weibull MLE will

often give smaller mean/median estimates than the imputation method, especially for the median. For Dummy L vs KM: these estimates are generally similar, but the Kaplan-Meier mean is often higher for small values and the KM standard deviation lower for small values. The KM median statistics may not be computed when there are many censored values, which is an advantage for the imputation and MLE methods.

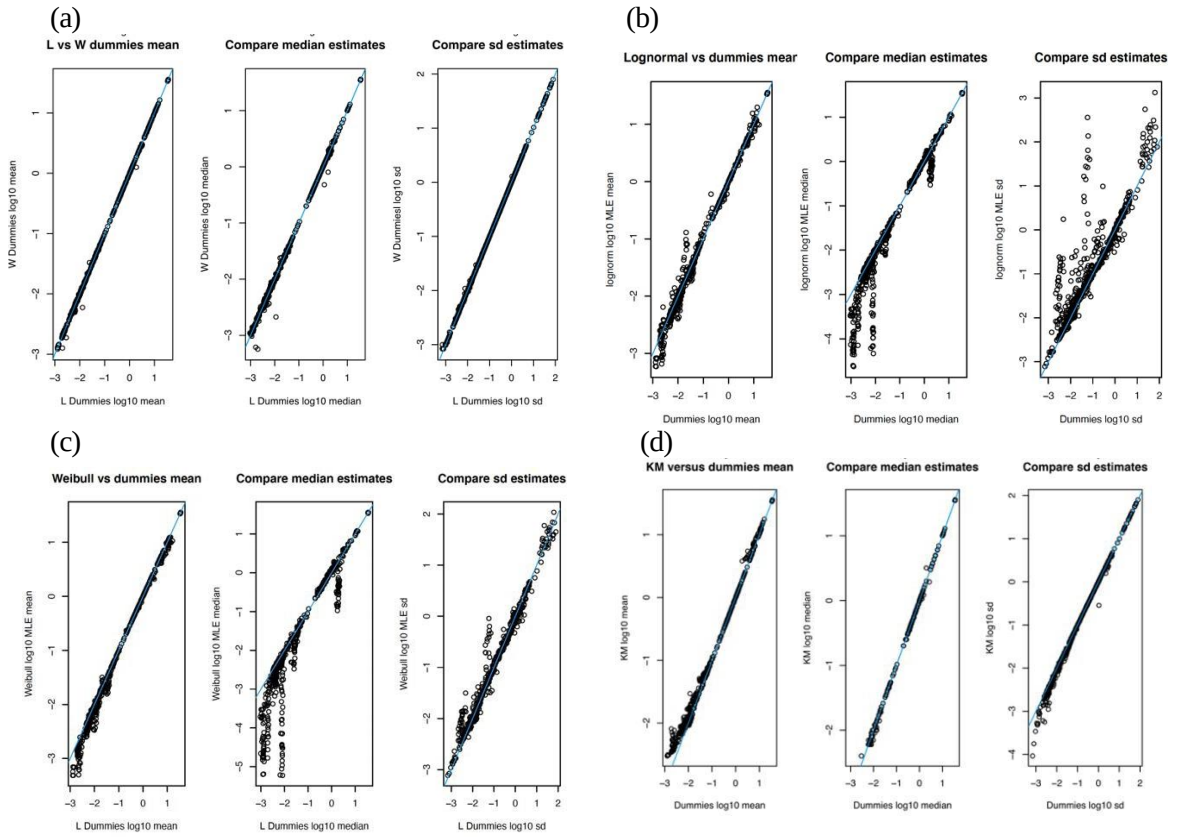


Figure 6. *The plot shows log_10 values of the alternative methods versus log-normal dummy imputation method in four sub-panels. Plots for the mean (left), median (center) and standard deviation (right) are given in each sub-panel. The methods are (a) Method 2 versus Method 1, (b) Method 3 versus Method 1, (c) Method 4 versus Method 1, (d) Method 5 versus Method 1.*

3.4 SIMULATION STUDIES

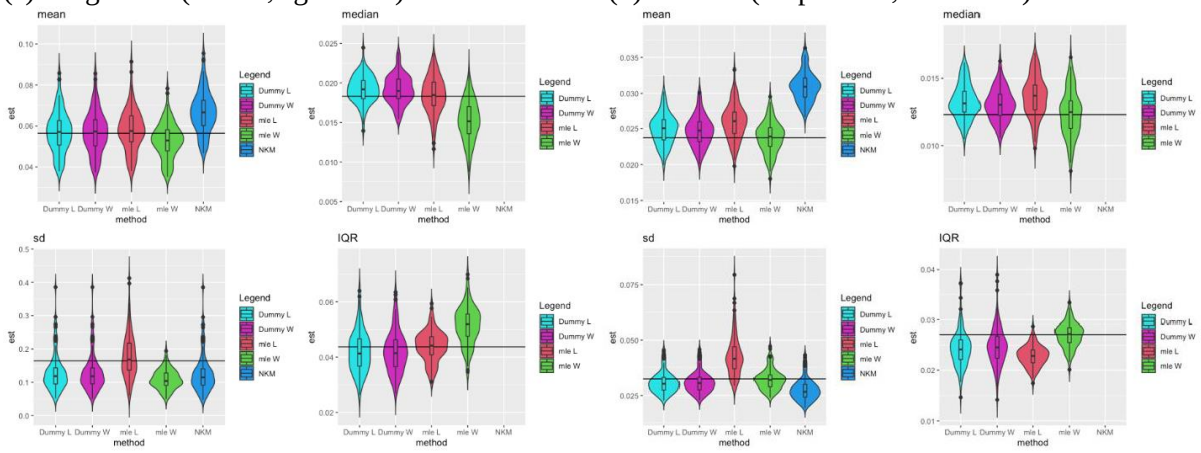
If the distribution really is log-normal or Weibull then we would expect the censored MLE methods to provide the best estimates, due to the asymptotic efficiency of MLE. A Monte Carlo simulation study was carried out to investigate the performance of the estimators under known distributions. Some data are simulated from a known true distribution, and then values below the MDL treated as censored. In Figure 7 we present results from four distributions with n=200 and the MDL at the 60% quantile of the true distribution. Estimates using the methods were computed and violin plots of the estimates from the 100 Monte Carlo simulations given. The true theoretical values are indicated by a black horizontal line.

From these simulations we observe that the censored MLE is unbiased and is the most efficient method when the underlying distributional family is correct. When the underlying distribution is incorrect the Weibull MLE has some bias, and the Lognormal MLE can be very biased for the mean and standard deviation. The dummy imputation methods can perform poorly in some cases. The Kaplan-Meier method did not perform well here generally, and the KM median and IQR estimates are not available for this level of censoring.

In addition, we consider changing the level of censoring to the 20% and 80% theoretical quantiles for two of the distributions, as seen in Figure 8. This figure further highlights the lack of robustness to distributional assumptions for the lognormal MLE estimates of mean and sd. Also there is large positive bias in the dummy estimates for the median with 80% censoring for both examples. From Figures 7 & 8. the Weibull MLE is shown to be the best overall estimator for these scenarios.

(a): Lognormal(mu=-4,sigma=1.5)

(b): Weibull(shape=0.75,scale=0.02)



(c): Weibull(shape=0.4,scale=0.04)

(d): normal(mu=0.3,sigma=0.05)

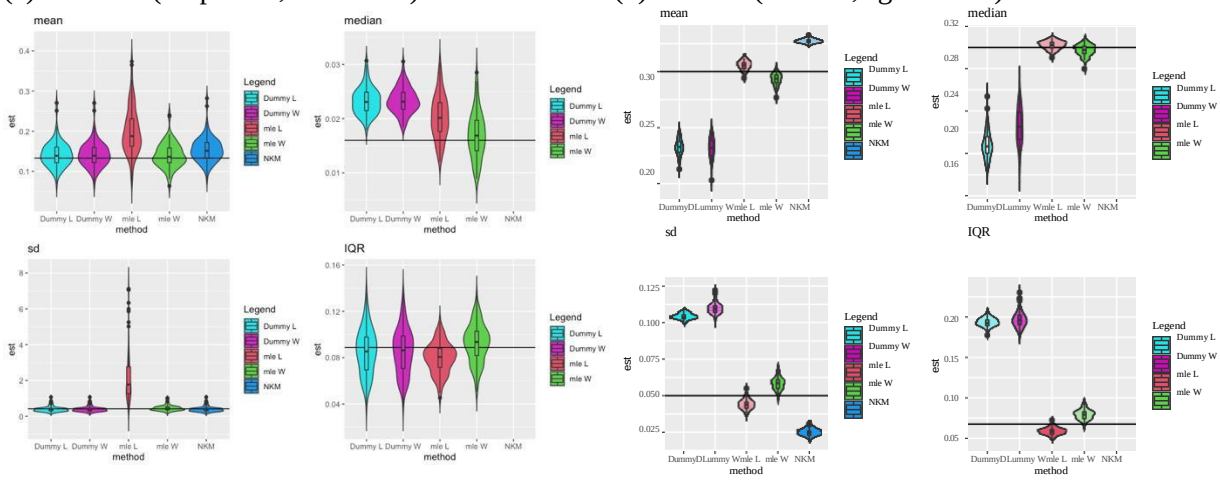
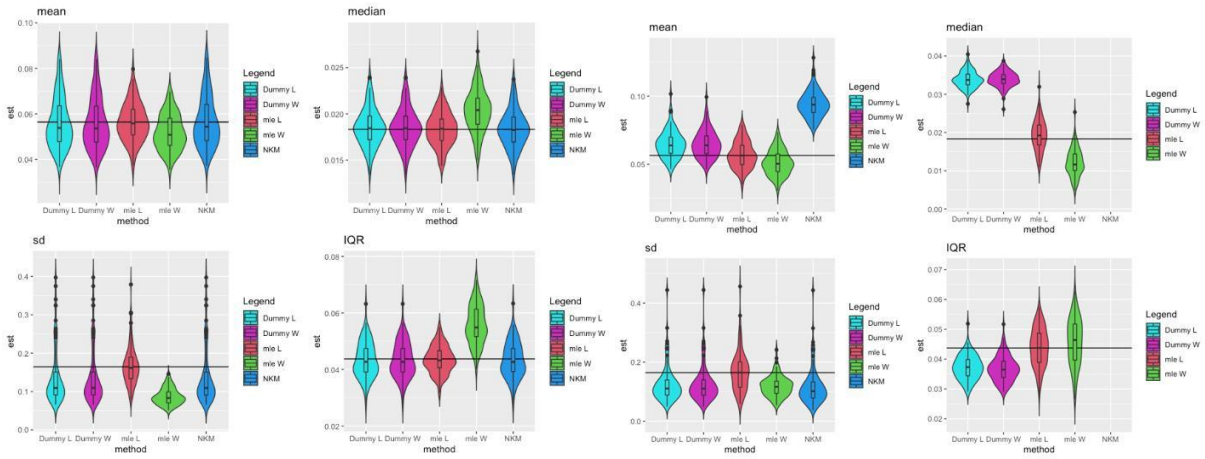


Figure 7. *Monte Carlo simulation results for different distributions. (a), (b), (c) are long-tailed positively skew distributions, and (d) is symmetric. In each case the censoring is 60% on average. Each plot shows violin plots for each method, which display the shape of the estimate distribution and a boxplot. The true value for each parameter is given by a straight line, and so good estimates have the line intersect the middle of the violin plot. The dummy lognormal, dummy Weibull, lognormal MLE, Weibull MLE and nonparametric Kaplan-Meier (when available) estimates are denoted by dummy L, dummy W, mle L, mle W, NKM respectively.*

(a): Lognormal(mu=-4,sigma=1.5) 20% cens. (b): Lognormal(mu=-4,sigma=1.5) 80% cens.



(c): Weibull(0.4,0.04) 20% censored (d): Weibull(0.4,0.04) 80% censored

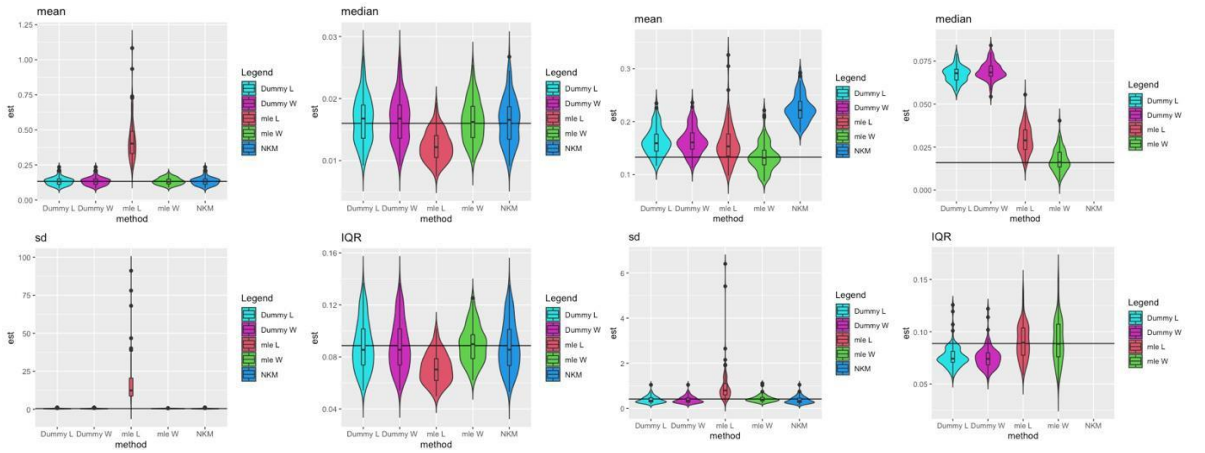


Figure 8. *Panels showing the performance for the mean, median, sd and IQR estimates under (a), (c) 20% censoring on average and (b), (d) 80% censoring on average.*

4 DELIVERABLES

Deliverables for this report include:

1- The full set of statistics for all variables and ICAs, which can be found in the csv file **Broward-imputed-results.csv**

2- The plots for the imputed ICA results, which are in the pdf file **Broward-analysis-imputation-ICA-plots.pdf**

3- The water quality data for all parameters, including calculated "dummies" are given in Excel file **ICA_Broward_Lab_WQ_censored.xlsx**

4- Example plots for NH4 S at Baker's Inlet are given in pdf file **Broward-analysis-ICA-all-methods.pdf**

5 RECOMMENDATIONS

From these investigations we recommend:

1. The censored MLE should be used for further analysis for the Broward data, with checks for goodness of fit of the appropriate distribution. From probability plots the Weibull distribution is appropriate in many cases, and the Weibull MLE is a good general choice.
2. Besides the distribution statistics, we provide the "dummies" generated during the performance of Flynn method, but warn against the imputation of those dummies to a specific sampling event and site. That would be incorrect.
3. Assess the links between land-use-cover on the watershed and the water quality characteristics at each ICA, requires good quality information on water circulation at each inlet and freshwater contributions, from surface and groundwater, as well as water chemistry of leaving/incoming through those inlets. If that type of data is missing it would be impossible to have an acceptable product.
4. There are statistically robust methods to compare censored data among ICAs and sampling site type (inlet, outfall and reef) (Helsel 2012), but such robustness certainly depends upon the percentages of Non-detects affecting the data set.

This is contribution #1459 from the Institute of Environment at Florida International University

Citation:
Briceño, Henry, Joseph N. Boyer and Ian L. Dryden. 2022. Kristen Jacobs Coral Reef Ecosystem Conservation Area Water Quality Data Analysis. Final Report. FDEP Order # BAA7752. FIU/IoE Contribution # 1459.

6 REFERENCES

Boyer, Joseph N., Henry O. Briceño, Jeff Absten, David Gilliam and Dick Dodge. 2012. 2011 Annual Report of the Water Quality Monitoring Project for the Southeast Florida Coral Reef Initiative (SEFCRI).Southeast Environmental Research Center at Florida International University FIU/SERC Technical Report #542, and National Coral Reef Institute at Nova Southeastern University Publication #141. 17 p.

Chambers, Raymond L.; Steel, David G.; Wang, Suojin; Welsh, Alan (2012). Maximum Likelihood Estimation for Sample Surveys. Boca Raton: CRC Press. ISBN 978-1-58488-632-7.

Dudley, William N., Rita Wickham, and Nicholas Coombs. 2016. An Introduction to Survival Statistics: Kaplan-Meier Analysis. J Adv Pract Oncol. 2016 Jan-Feb; 7(1): 91–100. doi: 10.6004/jadpro.2016.7.1.8 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5045282/

Flynn, Michael. 2010. Analysis of censored exposure data by constrained maximization of the Shapiro–Wilk W statistic. Ann. Occup. Hyg., Vol. 54, No. 3, pp. 263–271, 2010.

Helsel, D. R. (2006). Fabricating Data: How substituting values for nondetects can ruin results, and what can be done about it. Chemosphere 65 (11), 2434-2439

Helsel, D.R., 2010. Much Ado About Next to Nothing: Incorporating Nondetects in Science. Ann Work Exposures and Health 54, 257-262. http://dx.doi.org/10.1093/annhyg/mep092

Helsel, D.R., 2011. Statistics for censored environmental data using Minitab and R, 2nd edition. John Wiley and Sons, New York. 344 p. https://doi.org/10.1002/9781118162729

Jain, Ram B., Samuel P. Caudill, Richard Y. Wang, and Elizabeth Monsell. 2008. Evaluation of Maximum Likelihood Procedures To Estimate Left Censored Observations. Anal. Chem. 2008**,** 80, 1124-1132

Klein, J.P. and M.L. Moeschberger, 2003, Survival Analysis: Techniques for Censored and Truncated Data,2nd edition.Springer, New York, 536 pp.

Meeker, W.O. and L.A. Escobar, 1998, Statistical Methods for Reliability Data.Wiley, New York, 680 pp.

Royston, Patrick . 1992. Approximating the Shapiro-Wilk W-test for non-normality. Statistics and Computing, 2, 117-119.

Whitall, D., S.Bricker, D. Cox, J. Baez, J. Stamates, K. Gregg and F. Pagan. 2019. Southeast Florida Reef Tract Water Quality Assessment. NOAA Technical Memoradum NOS NCCOS 271. Silver Spring. 116 pages.