SUPPLEMENTAL DATA

ASCE *Journal of Water Resources Planning and Management*

Arid Inland Community Survey on Water Knowledge, Trust, and Potable Reuse. II: Predictive Modeling

Lauren N. Distler, Caroline E. Scruggs, and Kellin N. Rumsey

DOI: 10.1061/(ASCE)WR.1943-5452.0001219

www.ascelibrary.org

Cumulative Link Models (CLMs) refer to a class of models designed for data where the response variable falls in an ordered set of J categories. Let $Y_i = j$ if the i^{th} observation belongs to the j^{th} category and let $\pi_{ij} = P(Y_i = j)$ represent the probability that the i^{th} observation falls into category j. This implies that

$$\sum_{j=1}^{J} \pi_{ij} = 1, \quad \text{for all } i = 1, 2, \cdots n$$

Next we define the cumulative probabilities

$$\gamma_{ij} = P(Y_i \leq j) = \pi_{i1} + \cdots \pi_{ij}$$

so that

$$\pi_{ij} = \gamma_{ij} - \gamma_{i,j-1}$$

At the edge cases then, we have $\gamma_{i1} = \pi_{i1}$ and $\gamma_{iJ} = 1$ for each $i = 1, 2, \cdots n$. These cumulative probabilities are now modeled with respect to the predictor variables using a *link function*. In this paper, we choose to use a standard logit link function, which produces a proportion log-odds CLM (i.e. ordered logistic regression). The assumption is stated mathematically as follows.

$$\gamma_{ij} = \frac{1}{1 + \exp\left(-\theta_j - \sum_{k=1}^{p} x_{ik}\beta_k\right)}$$

Where x_{ik} represents the value of the k^{th} predictor variable for the i^{th} observation, the β parameters denote the usual regression coefficients and $-\infty \equiv \theta_0 \leq \theta_1 \leq \cdots \leq \theta_{J-1} \leq \theta_J \equiv \infty$ are known as the *threshold coefficients*. In total, there are $J-1$ threshold coefficients and p regression coefficients (where p is the number of predictor variables) and these coefficients are estimated with maximum likelihood. Maximum likelihood estimates (MLE's), are the maximizers of the Log-likelihood function. For this problem, the Log-likelihood function can be written using *indicator functions* as follows.

$$\ell(\theta_1, \cdots \theta_{J-1}, \beta_1, \cdots \beta_p | y_1, \cdots y_n) = \sum_{i=1}^{n} \sum_{j=1}^{J} \mathbb{1}(Y_i = j) \log \pi_{ij}$$

This optimization problem is solved efficiently and accurately using the `clm` function in the R package `ordinal`. Once these parameters have been estimated, one can easily extract estimates for the probabilities π_{ij} as a function of the predictor variables. If we desire to predict which class an observation belongs to, we can do this by choosing the j which maximizes the probability π_{ij}.

$$\hat{Y}_i = \arg\max_j \pi_{ij}$$

Figure S1. Estimation and interpretation of Cumulative Link Model (CLM) parameters, use of models for prediction

The ordered logistic regression models obtained LOOCV accuracies of 49.5% and 59.8% for DPR and IPR respectively. We claim that the demographic variables retained by the models (Table 1) have a significant impact on the predictive power of the models. To justify this claim, we compare the results to a simple but reasonable probabilistic model and conduct a formal hypothesis test to show that the ordered logistic regression models are effective. The null classifier makes predictions at random with probabilities proportional to how often each class occurs the data. For $i = 1, 2, \cdots n$, we have

$$(DPR\ Null\ Prediction)_i = \begin{cases} \text{Willing to accept} & , \text{ with probability} = 0.484 \\ \text{Neutral} & , \text{ with probability} = 0.221 \\ \text{Unwilling to accept} & , \text{ with probability} = 0.295 \end{cases},$$

$$(IPR\ Null\ Prediction)_i = \begin{cases} \text{Willing to accept} & , \text{ with probability} = 0.580 \\ \text{Neutral} & , \text{ with probability} = 0.237 \\ \text{Unwilling to accept} & , \text{ with probability} = 0.183 \end{cases}.$$

Formally, we would like to conduct the following hypothesis test.

$$H_0: \text{ There is no difference in accuracy between the two models}$$

$$H_1: \text{ The CLM model has higher accuracy that the null model}$$

Using the null classifier, we classify each of the n observations and calculate the accuracy. Repeating this process $10,000$ times, we construct empirical distributions for the accuracy of the classifier under the null Hypothesis. The figure below shows both of these empirical distributions. The null distributions are approximately Normal with a standard deviation of 0.012 in each case and a mean of 0.365 and 0.445 for DPR and IPR respectively. The vertical dotted line in each panel illustrates the actual prediction accuracies of the ordered logistic regression models. It is clear that the observed prediction accuracies, for both IPR and DPR, are far greater than they would be under the null hypothesis ($p - val \approx 0$). Thus we have strong evidence to reject the null hypothesis and we conclude that our models gain significant predictive power by using demographic information.

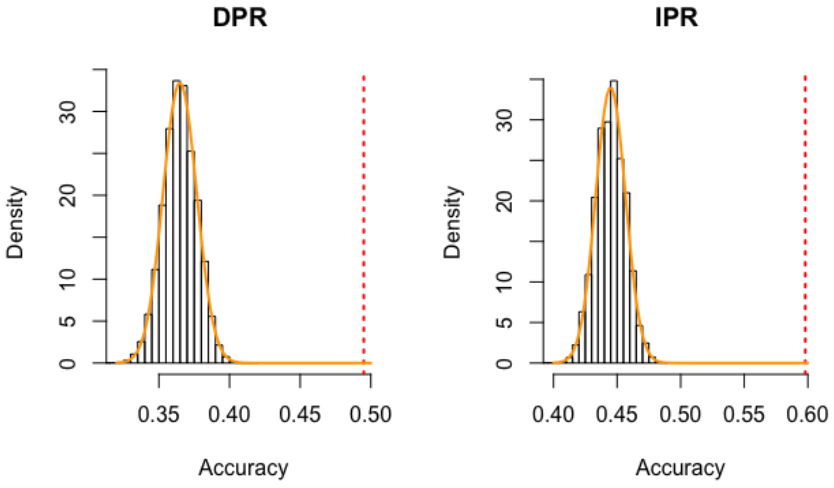

Figure S2. Justification of ordered logistic regression LOOCV accuracies

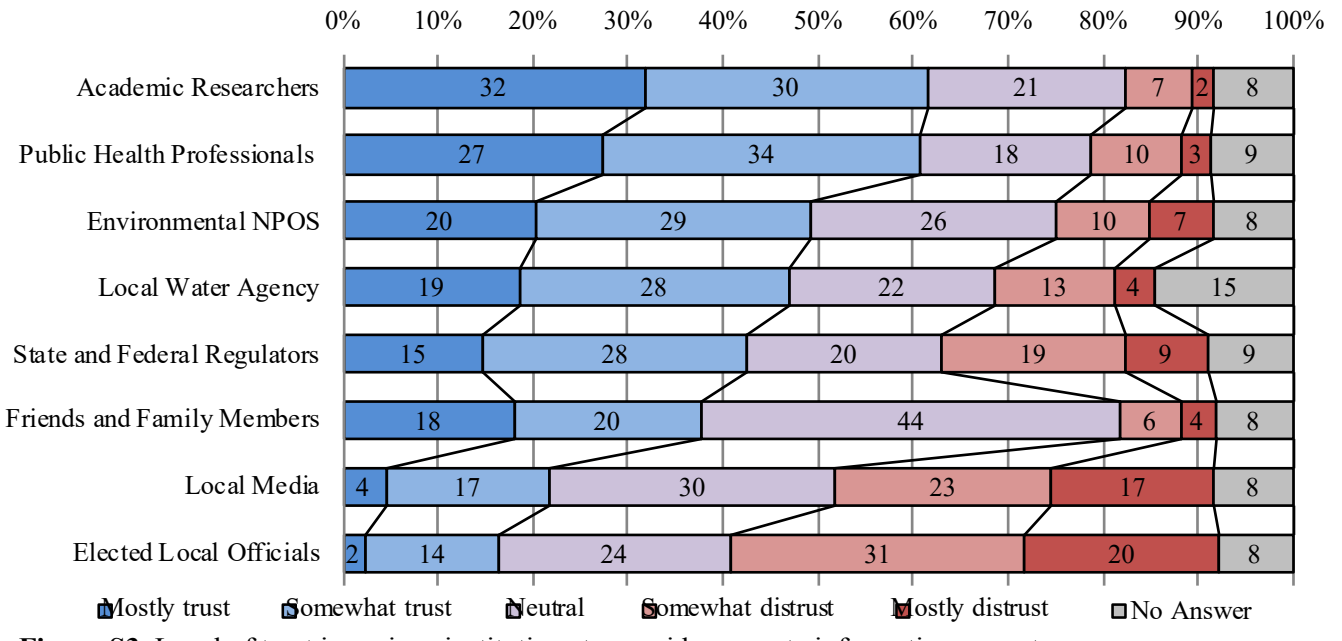

Figure S3. Level of trust in various institutions to provide accurate information on water reuse

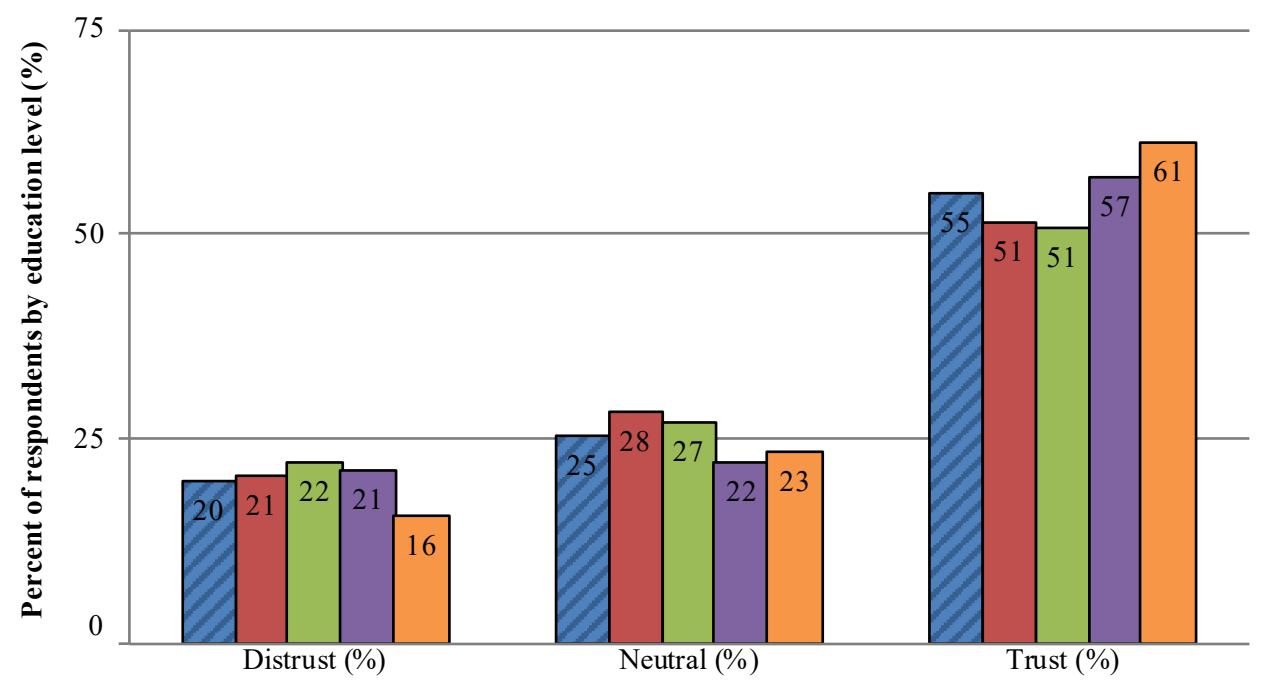

Figure S4. Level of trust in ABCWUA by education level compared to survey sample as a whole