Safe Water
Optimization Tool

Unpaired Data Modelling with the SWOT

Version 1.0
April 4, 2023
Michael De Santi, Dr. Usman T. Khan, Dr. Syed Imran Ali

Background and Problem Description

Waterborne illnesses are a leading cause of outbreaks of infectious disease in refugee and internally displaced person (IDP) settlements. To prevent these illnesses from spreading, it is crucial that water system operators provide sufficient free residual chlorine (FRC) at the tapstand to ensure that water is protected against recontamination throughout the post-distribution period (including collection, transport, and storage/use). Maintaining at least 0.2 mg/L of FRC throughout the post-distribution period is typically sufficient to prevent faecal recontamination. The Safe Water Optimization Tool (SWOT) uses machine learning (ML) and process-based (PB) models to forecast post-distribution FRC concentrations and provide tapstand FRC guidance to ensure that water remains protected against pathogenic recontamination up to the final point-of-consumption in the household.

One of the current limitations of the SWOT is that it requires paired water quality measurements from the tapstand and household (i.e., each household sample must have a corresponding tapstand sample from the same unit of water). The SWOT's PB models require this because empirical FRC decay models assume samples reflect the change in a single unit of water over time. However, paired sampling is not the normative practice in humanitarian response settings. Instead, FRC measurements are collected from the tapstand and household in an unpaired fashion: there is a set of tapstand measurements and a set of household measurements, but no guarantee that they are from the same unit of water, or even that an equal number of household and tapstand samples are collected. This poses problems for the ML models which, while not bound to physical assumptions, require data be structured (i.e., there must be a set of observed target values, a corresponding set of predictors, and an equal number of targets and predictors). Thus, unpaired data cannot be accepted into the SWOT's current ML models either. This is problematic because unpaired data collection is much more common, so many sites cannot benefit from the SWOT's optimized FRC guidance. Additionally, collecting paired data is more time consuming and requires more training than collecting unpaired data, creating barriers to switching from unpaired to paired data collection in order to use the SWOT.

This investigation presents a first attempt to develop modelling tools for the SWOT that can use unpaired data. To achieve this, we propose to use, instead of individual tapstand FRC measurements, the *distribution* of tapstand FRC concentrations from the day that water was collected to forecast the distribution of household FRC concentrations. Thus, instead of pairing individual measurements, we pair household measurements with the date that the water was collected and the distribution of tapstand water quality observations measured on that day.

Objectives

This investigation sought to achieve the following three objectives:

1. Develop a rules-based approach to pairing household measurements to the corresponding date the water was collected.
2. Generate predictor variables based on the distribution of tapstand measurements for a given day.
3. Produce a probabilistic model to forecast household FRC as a probability distribution and compare this model's performance to a comparable paired-data model.

Approach

Site and data description

For this investigation, we used an existing SWOT dataset collected using paired data. The use of a paired dataset allows us to assess factors such as errors in the rules-based pairing approach and performance decrease due to using unpaired data. The dataset we used was collected from the Kutupalong-Balukhali Extension Site, Camp 1 refugee settlement in Cox's Bazaar, Bangladesh. The dataset was collected over a six-month period from July to December 2019 and contains 2130 samples. At the tapstand and household several water quality parameters were measured including FRC, total residual chlorine (TRC), electrical conductivity (EC), water temperature, pH, and turbidity. The storage duration for each sample was measured using the timestamps from the paired data.

Rules-based data pairing

The overall goal of this study was to predict the distribution of household FRC concentrations based on the distribution of tapstand water quality measurements. To achieve this, we used a series of rules to associate each household FRC measurement to the date that that the water was collected. We developed these rules based on the data collection patterns and patterns of water handling behaviour on site. Observations from the site indicated that drinking water was typically collected at two times during the day: once in the morning and once in the afternoon. Water collected in the morning was typically used on the same day and water collected in the afternoon was typically stored overnight and into the next morning. The data collection on the site attempted to mirror this behaviour: when a tapstand measurement was collected in the morning, the household follow-up measurement was collected that afternoon. When the tapstand FRC measurement was collected in the afternoon, the household follow-up measurement was collected the next morning. Thus, to assign a date of water collection to each household FRC measurement, we can use the timestamp data associated with each household measurement. If the timestamp is from the afternoon (PM) then we know the water was likely collected that morning so the date of water collection is the same as the date of the household FRC timestamp. If the timestamp on the household measurement is from the morning (AM) we know that the water was likely collected the afternoon before so the date of water collection is set one day before the date in the household timestamp.
It is important to note at this point that to conduct this rules-based pairing, two pieces of information must be available: the data collection follow-up behaviour and the household timestamps. The storage durations cannot be used because with unpaired data we do not know which tapstand observations correspond to which household observations so we cannot know the precise storage duration for a given household measurement. Additionally, while water handling behaviours around storage duration are important, they are not as critical for this rules-based pairing as the actual data collection behaviours because it is the data collection follow-up (not the water handling behaviours themselves) that determine patterns in the measurement timestamps that allow us to infer the date of collection.

Predictor Variables

With each household observation associated to a date of water collection, we next need an approach to handling the multiple tapstand measurements associated with that given date of collection to form a set of predictor variables. Since we have no way to know which tapstand measurement is associated with which household measurement, we instead treat all tapstand measurements from a given day as a distribution representing that day's tapstand water quality

and extract statistics to summarize this distribution. For this study, we used four summary statistics:

1. The mean tapstand FRC;
2. The maximum tapstand FRC;
3. The minimum tapstand FRC;
4. The standard deviation of tapstand FRC measurements;

Using these summary statistics we obtain a structured dataset: each household FRC measurement has, as its set of predictors, the mean, maximum, minimum and standard deviation of tapstand FRC measurements from the date of collection.

Modelling Approach

To model the structured dataset developed through the steps described in sections 2.2 and 2.3, we used quantile regression neural networks (QRNNs). Neural networks are a type of ML model which identify and replicate non-linear patterns between input and output variables. QRNNs are a special type of neural network that are designed to predict a specific quantile of the distribution of output variables (a quantile is simply a percentile divided by 100, so the 95th percentile is equivalent to the 0.95 quantile). This ability to predict quantiles make QRNNs particularly useful for probabilistic modelling. A group, or ensemble, of QRNNs can be used to predict an entire probability distribution with each QRNN predicting a different quantile. This is beneficial for the SWOT because post-distribution FRC decay is highly variable, and probabilistic modelling is necessary to capture and quantify this variability and to accurately represent the nature of chlorine decay to water system operators.

We used an ensemble of 101 QRNNs to forecast the daily distribution of household FRC concentrations. These QRNNs predicted all quantiles from 0.01 to 0.99 in increments of 0.01, as well as the 0.0001 and 0.9999 quantiles, which were treated as upper and lower forecast limits. The individual QRNN models were designed as multilayer perceptron (MLP) type neural networks with a single hidden layer; the same architecture as used in the SWOT. The hidden layer used a hyperbolic tangent activation function and the output layer used a linear activation function. The hidden layer size was 32 neurons – 8 times the size of the input layer. This has proven effective in past modelling with the Bangladesh dataset. The cost function used to train the QRNNs is the quantile, or pinball, loss function (Equation 1). This is an asymmetric loss function defined as:

$$pinball\ loss = \begin{cases} \tau * e\ if\ e > 0 \\ (\tau - 1) * e\ if\ e < 0 \end{cases} \quad (1)$$

Where e is the difference between the true and predicted value of a household FRC measurement and τ is the selected quantile. Thus, the model penalizes overprediction errors at a rate of τ and penalizes underprediction errors at a rate of $1 - \tau$.

Performance Metrics

The quality of the probabilistic forecasts generated by the QRNN ensemble was evaluated using ensemble verification metrics, which are a special type of performance metric that assess the reliability and sharpness of a forecast. In probabilistic forecasting, reliability refers to the similarity between the forecast probability distribution and the underlying distribution of the observations. Sharpness refers to how small the spread of each forecast is around each observation (sharper forecasts have less spread around predictions). The specific metrics we used were:

- Percent capture (PC), measured for the overall dataset and for high-importance observations where the household FRC is below 0.2 mg/L
    - Measures the percentage of observations captured within the upper and lower bounds of the model forecast (defined by the 0.0001 and 0.9999 quantiles). Ideally 100% of observations will be captured, and higher PC is better.
- PI reliability score
    - Evaluates forecast reliability by comparing the percentage of observations captured within different prediction intervals (PIs) to the ideal value for that interval. For example, the 90% PI would be defined by the 0.05 and 0.95 quantiles – ideally 90% of observations would be captured within the 90% PI (and 80% within the 80% PI, etc.). The PI reliability score takes the sum of the squared difference between the ideal and actual capture in all PIs from 10% to 100% in 10% increments. An ideal score for the PI reliability score is thus 0 and lower scores are better.
- Rank histogram δ-score
    - Rank histograms are formed by identifying the rank that an observation would have within the sorted values of an ensemble forecast had it been a prediction (i.e., how many QRNNs predicted household FRC lower than the observed household FRC). A reliable model will have a flat rank histogram as an observation is equally likely to fall anywhere within the ensemble forecast and therefore the forecasting system is able to reproduce the underlying data generation process (Hammil, 2001). The δ-score evaluates the flatness of the rank histogram (Candille and Talagrande, 2005). The ideal score is 1 with scores much higher or lower than 1 indicating poor performance.
- CRPS and CRPS reliability term
    - The continuous ranked probability score (CRPS) is a probabilistic equivalent of mean absolute error. CRPS evaluates forecast sharpness, reliability, and uncertainty simultaneously, but tends to be dominated by sharpness (Ferro, 2014; Hersbach, 2000). The CRPS reliability term is based on a decomposition of the CRPS for ensemble forecasts and directly evaluates the ensemble reliability (Hersbach, 2000). An ideal score for both is 0, and lower is better.
- Quantile Score:
    - The quantile score takes the average quantile loss (Equation 1) over all quantiles. The ideal quantile score is 0 and lower is better.

To evaluate the generalization performance of the model, the following scores were only considered on an independent testing dataset which was obtained from sampling 25% of the original data and omitting this during the training of the QRNNs.

Results

Figure 1 shows the forecast median and 90% PI and the observations plotted against the average tapstand FRC from the date of water collection. This figure shows that the forecasts were able to capture the overall trends between the average tapstand FRC on the date of water collection and household FRC. Additionally, we see that the majority of observations are captured within the 90% PI, though there are several outliers, notably when the average tapstand FRC is 0.7 mg/L. However, since this is the 90% PI, 10% of observations should actually fall outside of the PI., so these outliers may not be a problem. Figure 2 shows the PI reliability diagram which plots the PC within each PI. We see that all the markers are very close to the 1:1 line shown in black, indicating that at all PIs the percent capture was very close to the ideal value, thereby confirming that the outliers in Figure 1 represent very nearly 10% of the observations. Thus, the outliers in Figure 1 are acceptable as, for all PIs, the QRNN ensemble was able to capture the targeted percentage of observations. This indicates that the model has good reliability (it can reproduce the underlying distribution of observations very well).

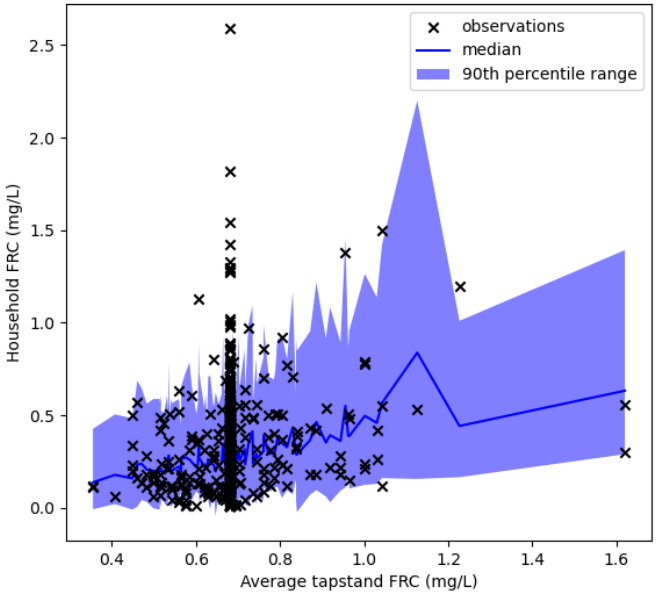

Figure 1: Forecast median and 90th percentile range and observed household FRC by average tapstand FRC.

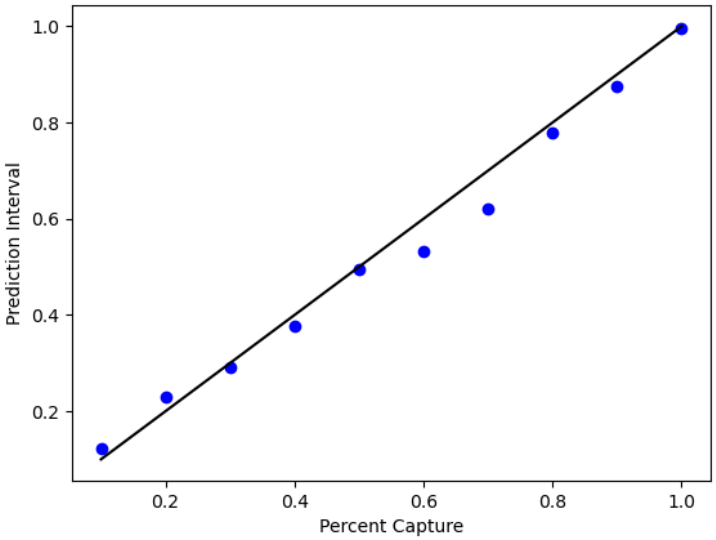

Figure 2: The PI reliability diagram using unpaired data

Table 1 shows the performance obtained by the unpaired QRNN ensemble model for the seven performance metrics described in Section 2.5. This table shows that even using unpaired data the model produced very good performance across most of the metrics. The PC, both overall and for observations with household FRC below 0.2 mg/L, was near-perfect. The PI reliability score was very close to a perfect 0, as was the CRPS reliability score. The only poor score was the rank histogram δ score which was much higher than 1 – indicating that despite a near perfect PI reliability score, there is some aspect of the reliability that requires improvement. Not only are several scores very close to their ideal value, this performance surpasses several past research attempts where the paired Bangladesh data was used with other types of neural networks (i.e., not QRNNs) (De Santi et al., 2021; 2022).

Table 1 also shows the performance using a comparable QRNN ensemble model with the paired data. As can be seen from Table 1, the paired model is still substantially better, with the improvement in the paired model relative to the unpaired model as high as 74%. This indicates that while the proposed method for modelling unpaired data is very effective, using the same QRNN approach with paired data still produces better performance. This also reinforces our choice to use the QRNN model as this too surpasses the performance of all previous modelling attempts using this dataset.

Table 1: Ensemble verification performance for the unpaired data model and a QRNN trained with paired data

Score	Ideal score	Unpaired Data Model	Paired Data Model	Percent Improvement with paired QRNN
PC (All)	100%	99.6%	99.6%	0%
PC (household FRC below 0.2 mg/L)	100%	100%	100%	0%
PI reliability score	0	1.39×10^{-2}	3.69×10^{-3}	74%
δ score	1	6.77	2.20	68%
CRPS	0	1.49x10-1	9.55×10^{-2}	36%

CRPS reliability score	0	1.34×10^{-3}	3.93×10^{-4}	73%
Quantile score	0	6.87×10^{-2}	4.52×10^{-2}	34%

One other key finding is that, in our Bangladesh field trial analysis, we found that 126 of the 2,130 samples were invalid due to FRC increasing from tapstand to household, household timestamps earlier than the tapstand timestamp, or improbably long storage durations. However, with an unpaired data approach, none of these errors are caught (since we cannot associate a given household measurement to a tapstand measurement). Future research should be conducted to catch potential data errors as many of the preprocessing checks the SWOT conducts cannot be conducted with unpaired data.

Conclusions and Next Steps

This investigation intended to develop a rules-based approach to pairing household measurements to their corresponding date of water collection, generate predictor variables based on the distribution of tapstand measurements on that water collection date, and then probabilistically forecast the distribution of household FRC concentrations. The approach we took to achieving these objectives, using the household timestamps to determine the date of collection and then using summary statistics of each day's tapstand FRC measurements to feed into a QRNN ensemble, demonstrated good performance on key ensemble verification metrics. However, we see that the QRNN modelling approach is still much more effective when using paired data. This is to be expected as the paired data provides more precise upstream information than summary statistics, however the performance of the unpaired model is still good and this reduced performance may be acceptable for some potential SWOT users as it allows for cheaper and easier data collection (potentially meaning more data for the SWOT). Further refinements to our modelling and rules-based pairing approaches may also decrease the performance deficit of the unpaired model. We intend to investigate the following improvements to the current modelling approach:

- Investigate alternative summary statistics to determine the most useful predictors for the distribution as a whole and for predicting individual quantiles
- Investigate the effect of adding additional water quality data (EC, water temperature, etc.)
- Investigate approaches to refine the rules-based pairing, using multiple clusters for morning vs afternoon collection, or individual clusters per tapstand or water distribution network.

There are several future tests that should be conducted prior to deploying this modelling approach for unpaired data. First, since the Bangladesh dataset is one of the larger SWOT datasets, this unpaired approach should be reviewed using a smaller dataset. Next, this approach should be validated on a true unpaired dataset that was collected without paired sampling. If acceptable performance can be obtained in these test cases, this would demonstrate the validity of this modelling approach for unpaired data.

References

Candille, G., Talagrand, O., 2005. Evaluation of probabilistic prediction systems for a scalar variable. Q. J. R. Meteorol. Soc. 131, 2131–2150. https://doi.org/10.1256/qj.04.71

De Santi M., Khan U. T., Arnold M., Fesselet J.-F., & Ali S. I. 2021. Forecasting point-of-consumption chlorine residual in refugee settlements using ensembles of artificial neural networks. *npj Clean Water*, 4, 35. https://doi.org/10.1038/s41545-021-00125-2

De Santi M., Ali S.I., Arnold M., Fesselet J-F., Hyvärinen A. M. J., Taylor D., & Khan, U. T. 2022. Modelling point-of-consumption residual chlorine in humanitarian response: Can cost-sensitive learning improve probabilistic forecasts? *PLOS Water*, 1(9), e0000040, https://doi.org/10.1371/journal.pwat.0000040

Ferro, C.A.T., 2014. Fair scores for ensemble forecasts. Q. J. R. Meteorol. Soc. 140, 1917–1923. https://doi.org/10.1002/qj.2270

Hamill, T.M., 2001. Interpretation of Rank Histograms for Verifying Ensemble Forecasts. Mon. Weather Rev. 129, 550–560.
https://doi.org/10.1175/15200493(2001)129<0550:IORHFV>2.0.CO;2

Hersbach, H., 2000. Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. Weather Forecast. 15, 559–570