**WATER QUALITY ANALYSIS**

**PHASE 2: INNOVATION**

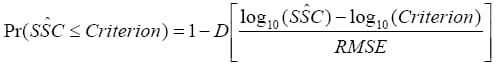
**INTRODUCTION**:

The machine learning technique we implement in this project is Regression. Regression is a statistical technique used in data analysis and machine learning to model the relationship between a dependent variable and one or more independent variables. It is a method for understanding how changes in the independent variables are associated with changes in the dependent variable. The goal of regression analysis is to create a predictive model that can make accurate predictions or estimates about the dependent variable based on the values of the independent variables.

**QUANTIFYING UNCERTAINITY:**

All regression models have uncertainty inherently associated with each computation. Uncertainty can be defined in a number of different ways, including relative percent difference, absolute error, and prediction intervals. Prediction intervals, used by this website, define a range of values for the response variable for a given level of certainty. The level of certainty presented with the modeled data is the 90-percent prediction interval. The larger the range, the more uncertainty there is associated with the regression computed value.

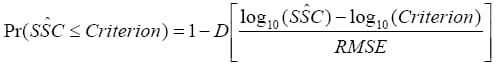
Calculating prediction intervals for regression models with two or more explanatory variables involves matrices.



* tis the value of Student's t-distribution having n-3 degrees of freedom with an exceedance probability of α/2,
* αis the level of certainty for the prediction interval (1-0.9, or 0.1 for this website),
* pis the number of explnatory variables plus one,
* sis the variance of the residuals calculated during model development,
* (X'X)-1is the X prime X inverse matrix calculated during model development — an expression of the covariance among all explanatory variables, and
* X is a vector of the explanatory variable measurements.

**PROBABILITY OF CRITERIA:**

A convenient way to interpret regression model computed concentrations in the context of water-quality criteria is the probability of exceedance. Probability of exceedance is a single value representing the percent likelihood that a criterion has been exceeded (Eq 7). This approach assumes that a normal distribution centered on the mean computed value describes the uncertainty distribution. For log-transformed response variables, this equation is applied in logarithmic space.



* Pris the probability that the criterion has been exceeded (0 <Pr< 1),
* Dis the cumulative distribution function for the standard normal curve values for it are obtained from equations that approximate the exact values.
* RMSE is the root-mean-squared-error, or standard error of the regression, or standard deviation of the residuals.

**LIMITIONS OF REGRESSION ANALYSIS:**

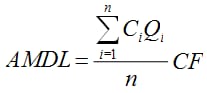
* A regression model between the response and explanatory variables.
* It may change over time if changes occur in the sources of the

constituent or an improved sensor becomes available.

* This is because of simplifying assumptions implicitly built into the regression analysis.
* For example, turbidity measurements are affected by physical properties of suspended-sediment particles such as size, color, and density.
* In turn, these physical properties are affected by complex watershed properties such as source sediment lithology, stream morphology, land use distribution, and many other things.
* Therefore, regression analysis is site-specific, and the regression model must be verified annually by continued sample collection and refined as needed.

**CALCULATION FOR CONSTITUENT LOAD:**

Annual mean daily loads (AMDL) of constituents are calculated by summation of available hourly (or more frequent time steps) loads for a given year:



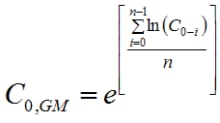
* Ciis the instantaneous, hourly concentration at the th time,
* Qiis the instantaneous, hourly streamflow at the ith time,
* CFis a conversion factor from Table 1, and
* N is the number of available hourly values for a given year (a maximum of 8,760).

**STEPS FOR CALCULATION:**

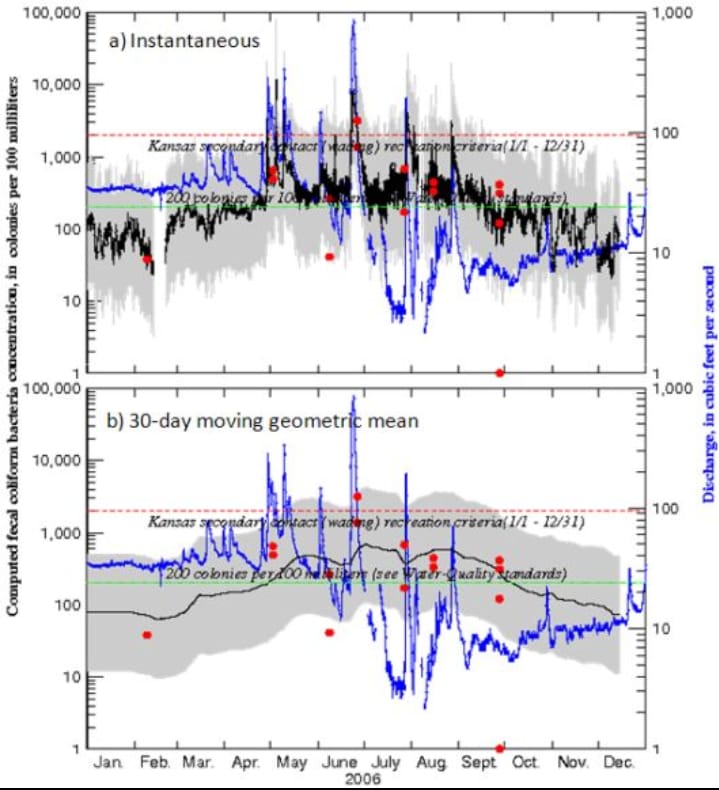
* Load computations assume that there is no error or uncertainty in the computation of streamflow.
* Although this is not true, the typical measurement uncertainty of streamflow is less than 5 percent.
* Measurement uncertainty for the sensor values typically are less than 10 percent if common protocols for operation of water quality.

**CALCULATION FOR GEOMETRIC MEAN OF INDICATOR BACTERIA:**

* This approach, a type of smoothing function, helps reveal the "central tendency" of the data and also allows for more direct comparisons with some water-quality criteria.
* A comparison between instantaneous and geometric-mean values is calculated.
* This is consistent with definitions of biological water-quality criteria used by some regulatory agencies, such as the Kansas Department of Health and the Environment.



* C0,GM is the calculated 30-day moving geometric mean for one moment in time,
* N is equal to 720, which is 24 instantaneous values per day multiplied by 30 days,
* C0 through C0-i are the current instantaneous value, and the preceding 719 instantaneous values.



**CONCLUSION:**

Regression analysis plays a vital role in water quality assessment and management. It empowers stakeholders to make informed decisions, protect water resources, and contribute to the sustainable management of this critical natural asset. However, it should be used in conjunction with other analytical methods and domain knowledge to provide a comprehensive understanding of water quality dynamics.