WaQuaPred

Water Quality Predictions for Safe Drinking Water Production and Recreation

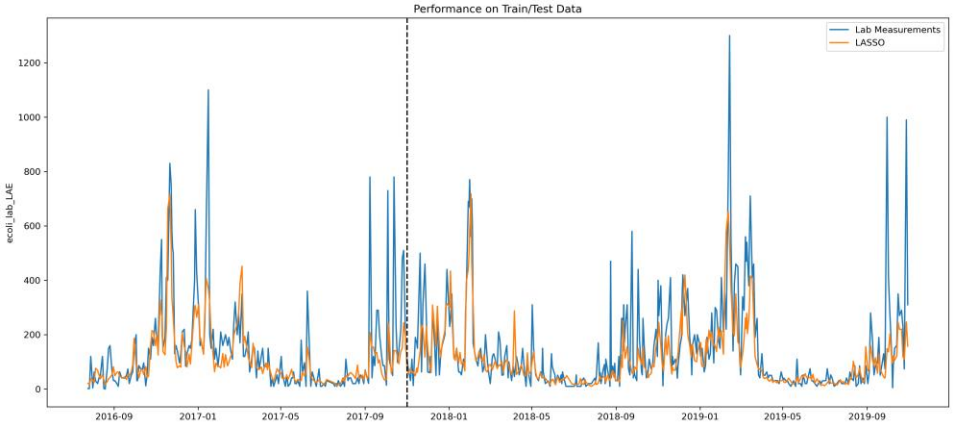
Ekaterina Sokolova, ACE – Water Environment Technology
Mia Bondelind, ACE – Water Environment Technology
Nora Speicher, DSRE
Ann Lilliestrom, DSRE
Oscar Ivarsson, DSRE

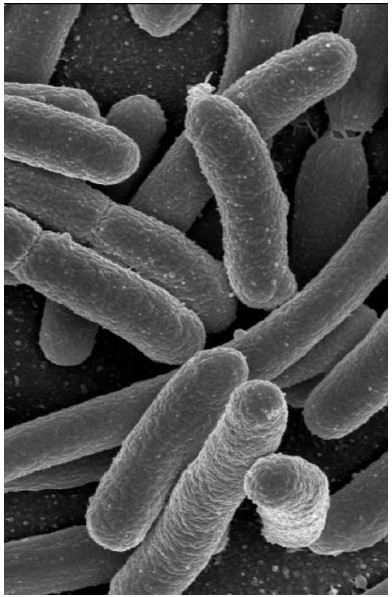
WaQuaPred

Water Quality Predictions for Safe Drinking Water Production and Recreation

Ekaterina Sokolova, ACE – Water Environment Technology
Mia Bondelind, ACE – Water Environment Technology
Nora Speicher, DSRE
Ann Lillieström, DSRE
Oscar Ivarsson, DSRE

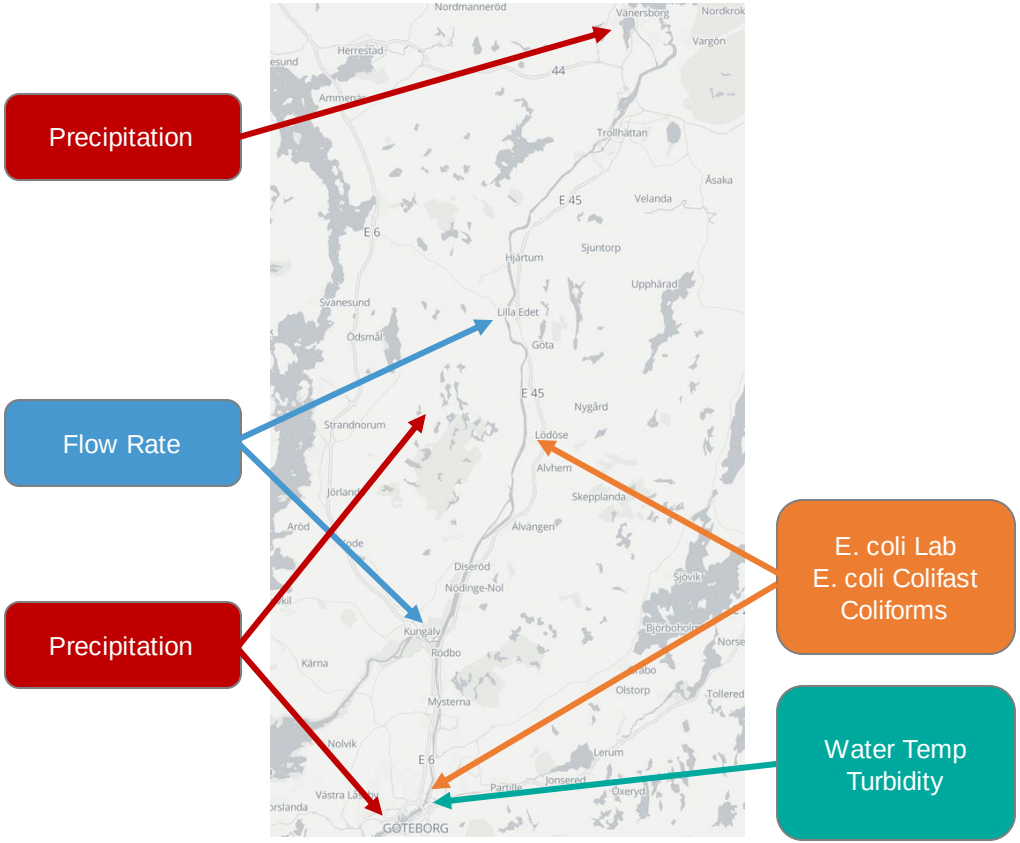
- **Background:** Pollution of water is an increasing problem globally. The overall aim of this project is to address the challenges of microbial pollution in surface water sources used for drinking water production and recreation.
- **Research Question:** Can existing monitoring data be used to develop data-driven models in order to describe and predict microbial water quality in Göta Älv?
- **Results:**
  - Multiple models implemented for forecasting E. coli levels one day ahead.
  - Inclusion of external predictors increased accuracy over several metrics.
  - Models for capturing non-linear relationships in the data performs slightly better on test data but are also more prone to overfit.
  - Upstream Colifast measurements (1-day lag) together with precipitation (1-2-day lag) have highest positive effect on E.coli levels using the LASSO model.





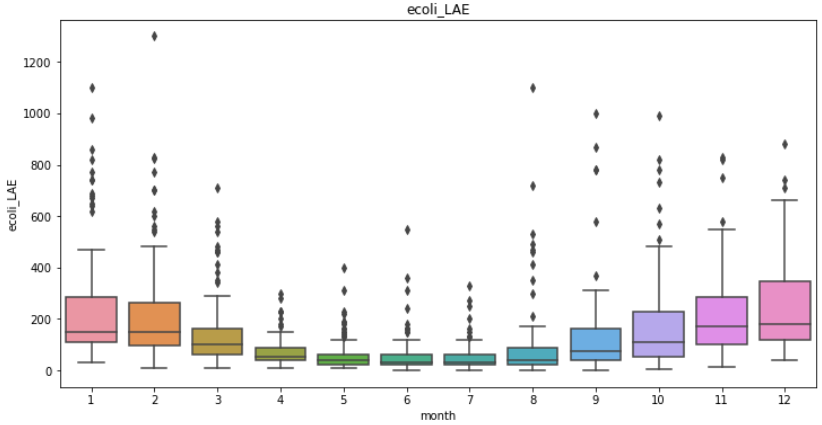
Background

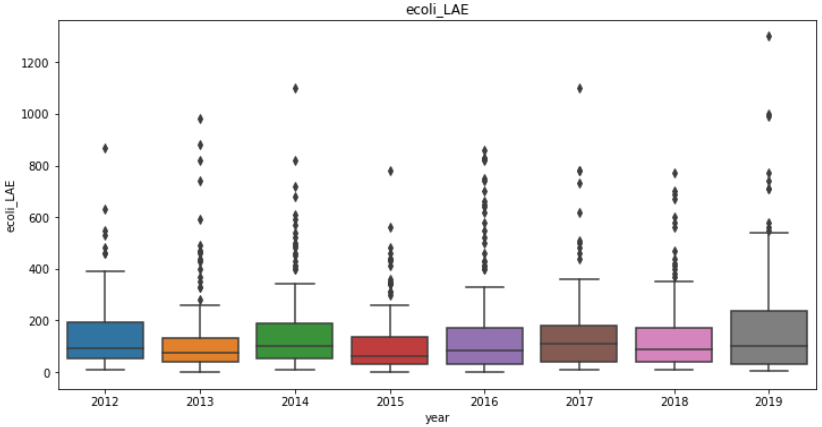
- Pollution of water is an increasing problem globally. Pathogenic microorganisms in water originating from human and animal fecal matter can infect humans and cause gastrointestinal diseases.
- The overall aim of this project is to address the challenges of microbial pollution in surface water sources used for drinking water production and recreation.
- Manual methods for analyzing water quality can be either time consuming or have low accuracy.
- Traditionally, process-based models have been used for describing water systems. It is now believed that with the current availability of data, data-driven models can be used to replace or complement these models.
- The objective is to develop a data-driven model, using AI and ML methods based on existing monitoring data, in order to describe and predict microbial water quality in surface waters.



Data Sources

- Hydrometeorological data from SMHI and water quality indicators from Göteborg Kretslopp och Vatten (GKV).
- More data is available but is considered to have little relevance for predicting water quality indicators or has too low sample frequency.
- It has been showed before that high precipitation is an indicator for higher E. coli levels.
- With combined features we have data for 2012-04-03 - 2019-10-30.
- Water quality indicator as target feature: E. coli Lab (Lärjeholm).



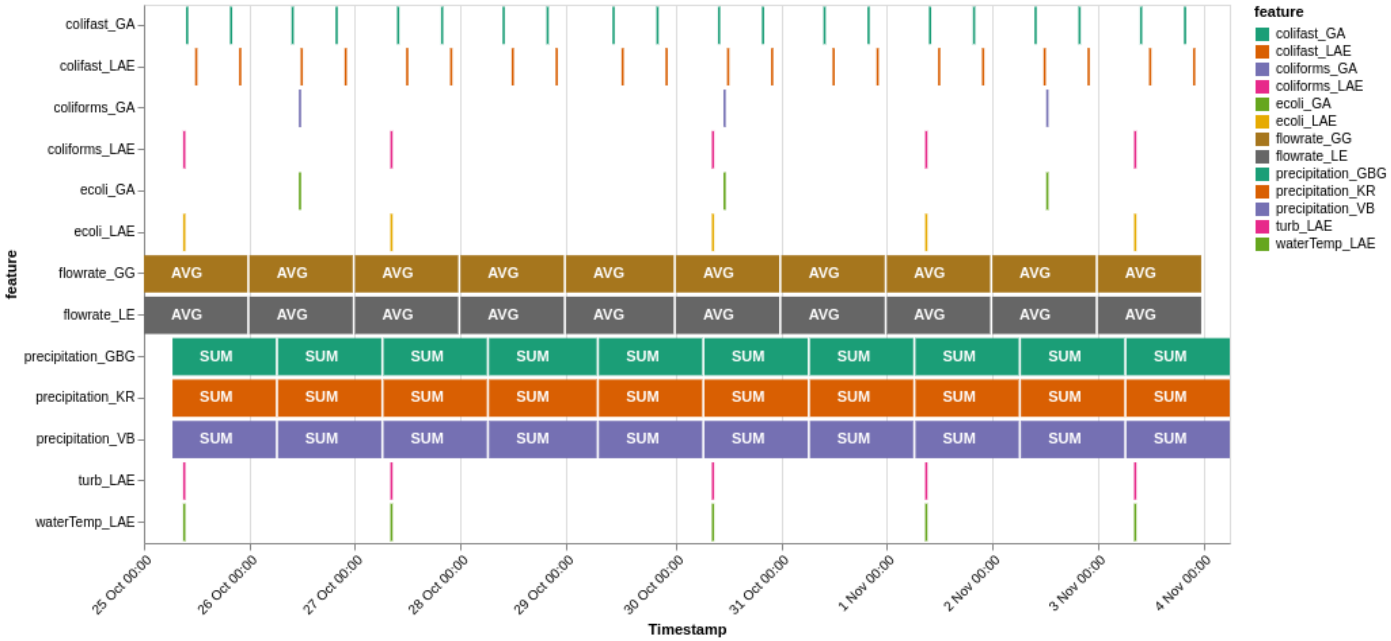


Data

Target Variable

- E. coli Lab (Lärjeholm) shows no obvious trend with a slight seasonal behavior over a year with higher extreme values and more variation during the winter months.
- The idea has been that this seasonal component can be explained by external predictors, e.g. water temperature.
- Data has been split into train/test with the test set being the last 2 years of data.
  - Train: ~5.5 years – 872 observations
  - Test: 2 years – 341 observations
- Focus on forecasting one day ahead.

Data
Temporal Format



Methods and Results

Baselines

- Naive
  - y(t+1) = y(t)
- Exponential Smoothing
  - Simple variant, could probably be improved with Holt-Winter's Seasonal Exponential Smoothing.
  - Parameter alpha (how much importance the models should allocate to its most recent observation) is optimized with statsmodels.
  - Not suitable for irregular time series.
- ARIMA
  - Parameters selected based on AIC using pyramid: p=4, d=0, q=2
  - Seasonal ARIMA tested but Out-of-memory errors. Workaround could be to use Fourier series for modeling seasonality.
  - Not suitable for irregular time series.

Train Metrics

Model	MAE	SMAPE	RMSE	R^2
Naive	89.84	66.19	159.70	0.11
ExpSmooth	84.63	62.00	143.28	0.28
ARIMA	83.81	65.50	136.50	0.35

Test Metrics

Model	MAE	SMAPE	RMSE	R^2
Naive	93.07	64.16	159.06	0.18
ExpSmooth	85.41	58.61	144.17	0.33
ARIMA	89.08	66.37	141.68	0.35

Methods and Results

Multivariate

- LASSO
  - Grid search with TSCV-5.
  - Parameters Optimized:
    - Regularization parameter alpha
    - Fit intercept
    - Target Transformer: log1p/exp-1
    - Predictors Transformer: PowerScaler, StandardScaler
- Random Forest
  - Grid search with TSCV-5.
  - Parameters Optimized:
    - Max Depth
    - Min Samples Split
    - Min Samples Leaf
    - Max Features
    - Bootstrap
    - Target Transformer: log1p/exp-1
- TPOT
  - Model pipeline optimized through genetic algorithms.
  - Produces very complex pipelines.
- VAR
  - Forward padding is performed for missing data and irregular time series.
  - AIC is used to decide lag parameter: 4

Train Metrics

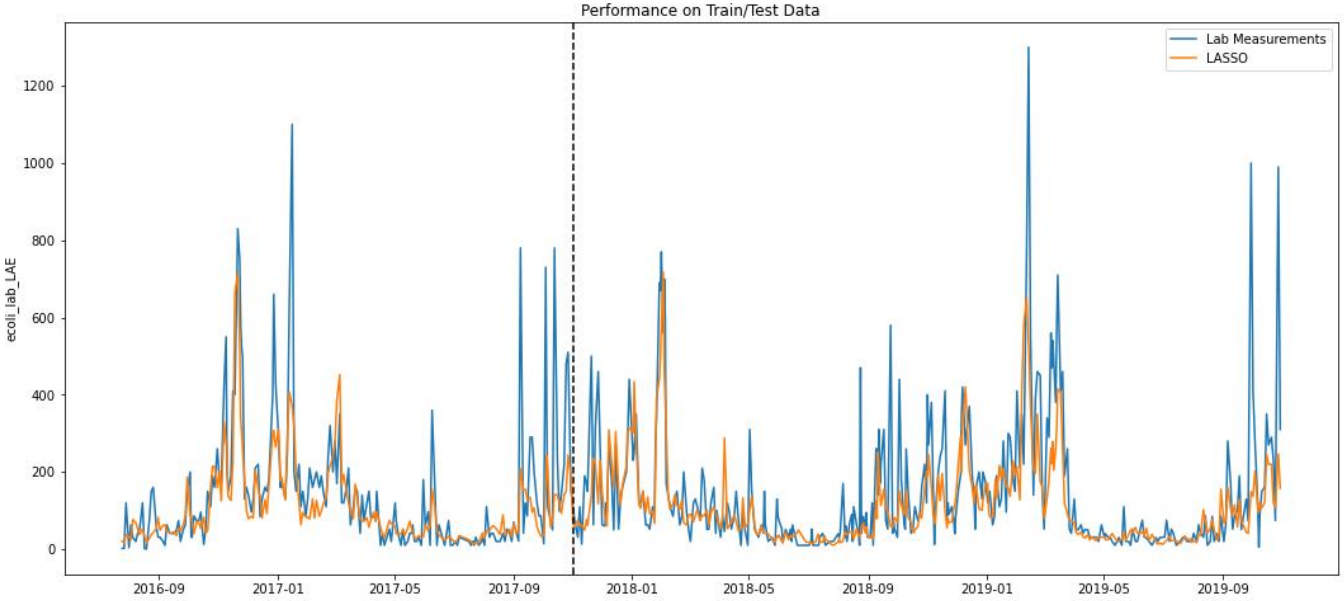
Model	MAE	SMAPE	RMSE	R^2
LASSO	60.54	48.93	110.02	0.57
RF	43.36	37.59	81.10	0.76
TPOT	38.83	36.50	74.00	0.80
VAR	65.73	59.19	111.86	0.55

Test Metrics

Model	MAE	SMAPE	RMSE	R^2
LASSO	68.55	51.02	123.34	0.51
RF	71.26	51.16	130.63	0.45
TPOT	63.66	53.27	106.18	0.63
VAR	71.59	63.56	116.90	0.56

Methods and Results

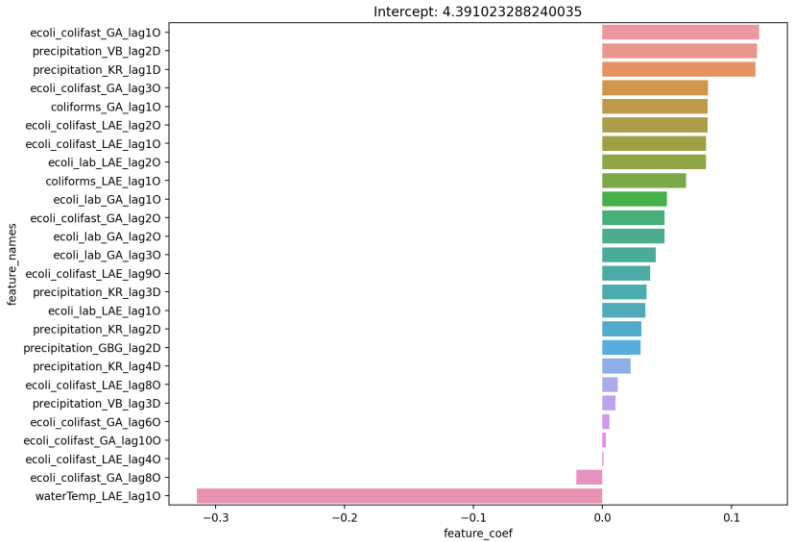
LASSO

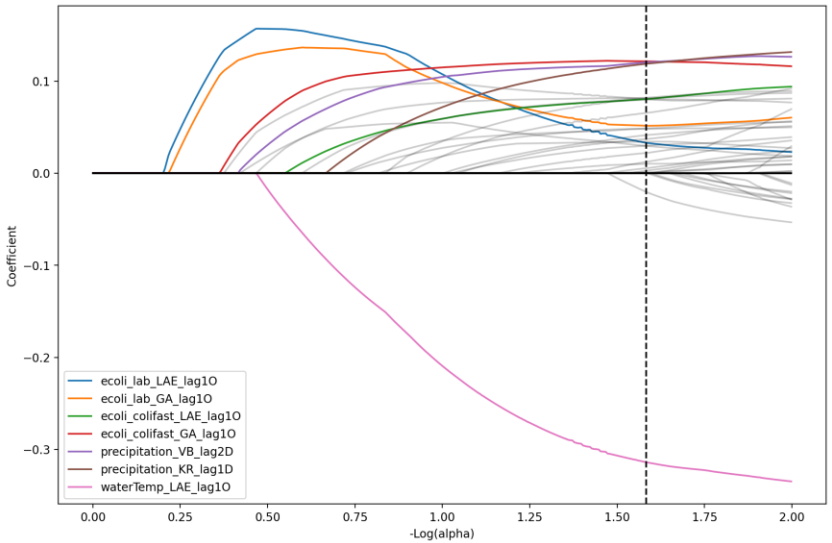


- Residuals are clearly larger during winter months (Q4-Q1).
- When model fails could also be important cases when increased E. coli levels is caused by something else that cannot be explained by data.
- However, it doesn't seem like the model can catch all peaks that it "should" catch.

Methods and Results

LASSO





Conclusion

- Inclusion of external predictors increased accuracy over several metrics.
- Models for capturing non-linear relationships in the data performs slightly better on test data but are also more prone to overfit.
- Upstream Colifast measurements (1-day lag) together with precipitation (1-2-day lag) have highest positive effect on E.coli levels using the LASSO model.
- Further work:
  - Further collaboration with GKV. Look into when the models perform bad, but they shouldn't and see if this is something that we can improve.
  - Include seasonality in autoregressive models.
  - Another approach in dealing with irregular time series. Look into models that can deal with this kind of format. More knowledge about time series regression is required.
  - Look into previously developed models for this purpose (Tornevi, 2014).

CHALMERS