Woodard & Curran

January 24
2023

Clarifying Insight: Using Machine Learning to Evaluate Secondary Clarifier Performance

PRESENTED BY

Julia Wahl (she/her)

Erik Osborn, P.E. (he/him)

Motivation

→Clarifier performance question

→Multiple parameters impact the system
- What's the driver?

→Can we leverage data analysis for new insights?

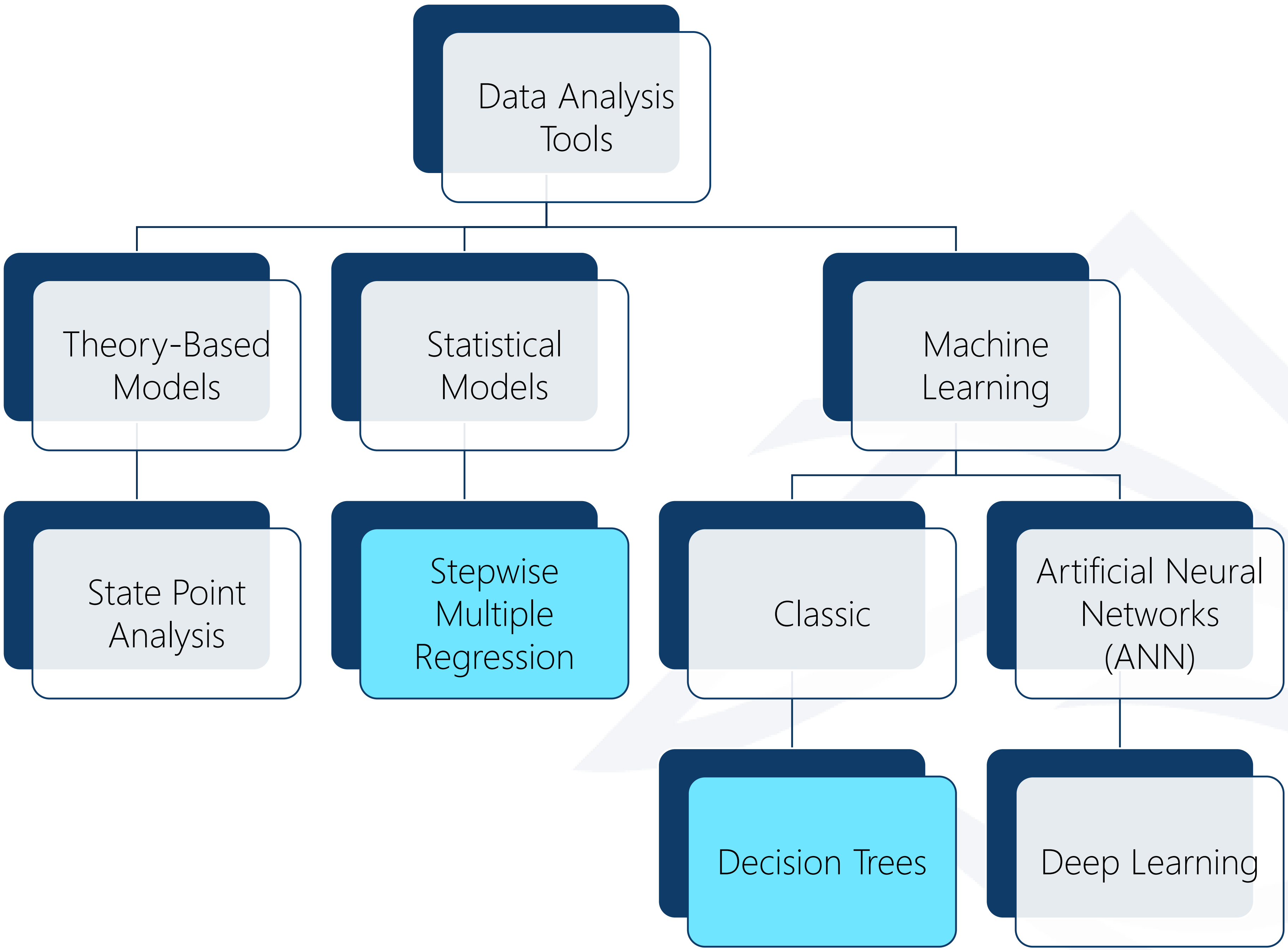
→Do data analytical techniques complement process analysis?
- Can we use this process for similar problems?

Learning Goals

Introduce two common data analysis tools: a statistical model and a machine learning model.

Learn about how these models can be used to help understand plant performance using daily data

As a case study, apply these models to actual plant data to identify factors that impact secondary clarifier performance


Data Analysis Tools
Theory-Based Models
Statistical Models
Machine Learning
State Point Analysis
Stepwise Multiple Regression
Classic
Artificial Neural Networks (ANN)
Decision Trees
Deep Learning

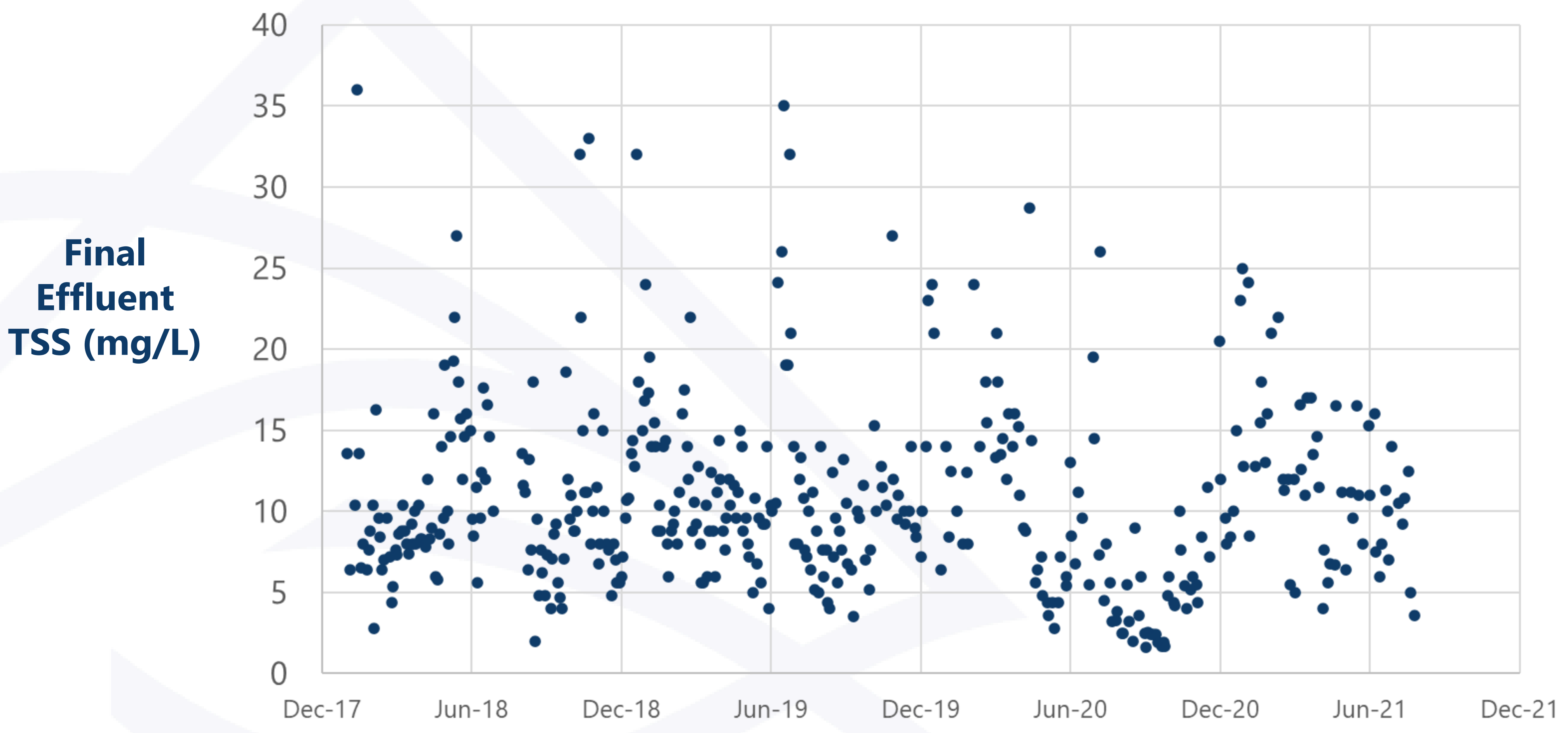
Decision Tree vs. Stepwise Multiple Regression

	Decision Tree (Machine Learning)	**Stepwise Multiple Regression (SMR)** (Statistical Model)
Can fit ("train") model with small datasets (n=100 to 1000)	**Yes**	**Yes**
Model can be interpreted by humans	**Yes**	**Yes**
Data needs to satisfy certain assumptions	**No**	**Yes**
Model can determine which parameters are significant	**No**	**Yes**
Model prone to overfitting	**Yes**	**No**

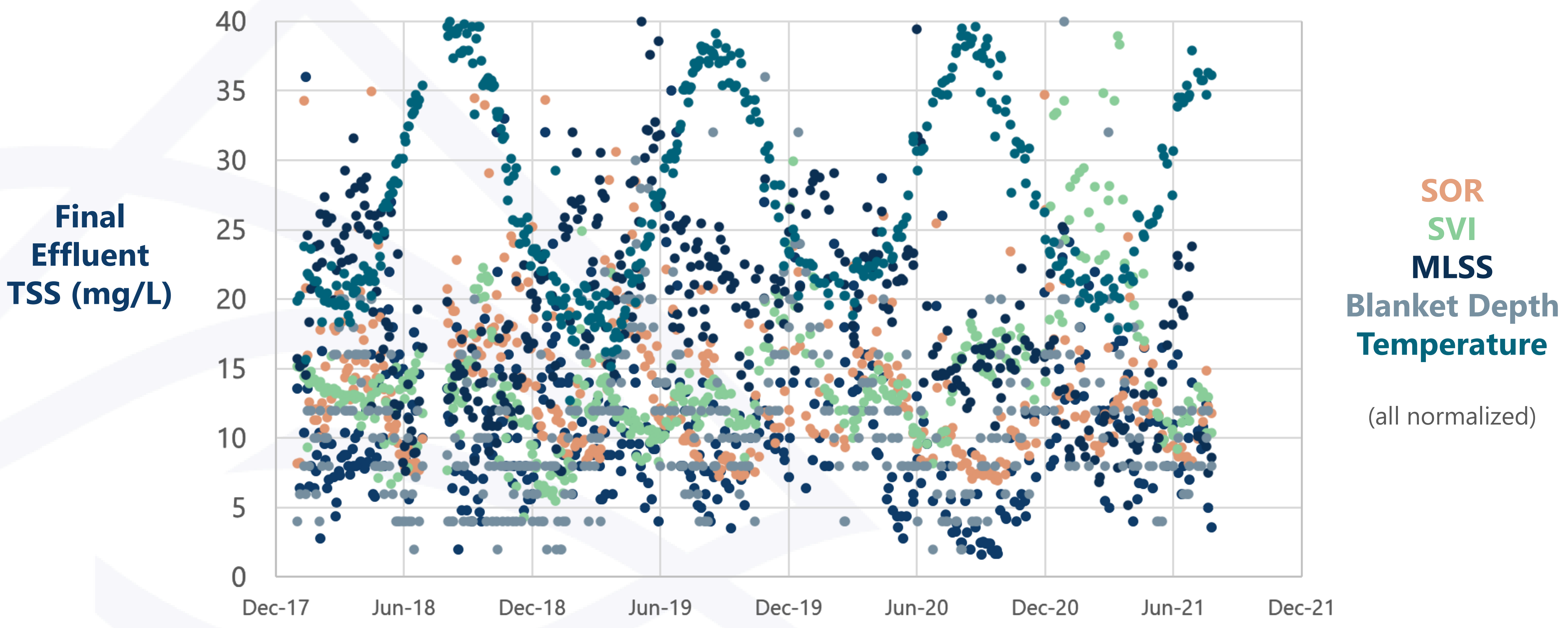
Case-Study: LAWPCA Secondary Clarifier Data

- 14 MGD secondary treatment plant
- Meets effluent TSS limits of 30/45/50 mg/L
- Clarifiers operate well below the point of solids overload predicted by SPA.
  - < 65% critical capacity 99% of the time

Case-Study: LAWPCA Secondary Clarifier Data

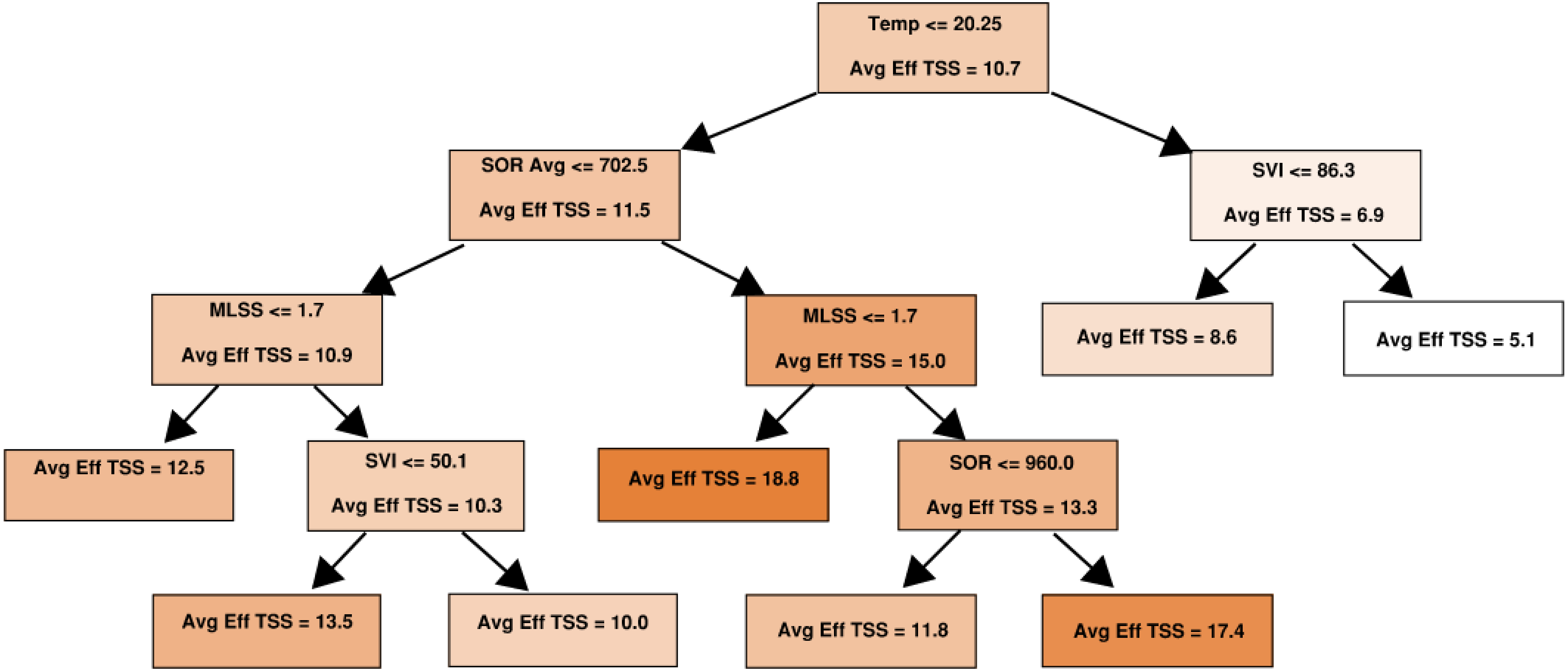


Case-Study: LAWPCA Secondary Clarifier Data

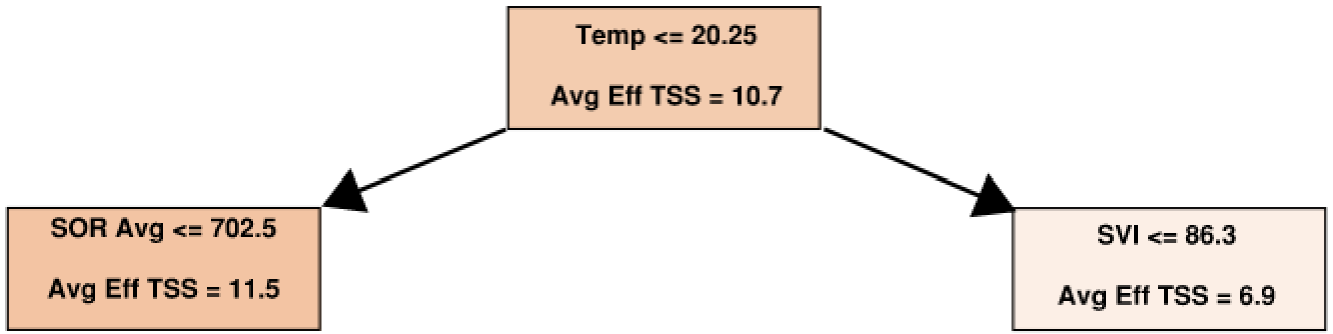


Data Analysis Tool: Decision Tree (Machine Learning)

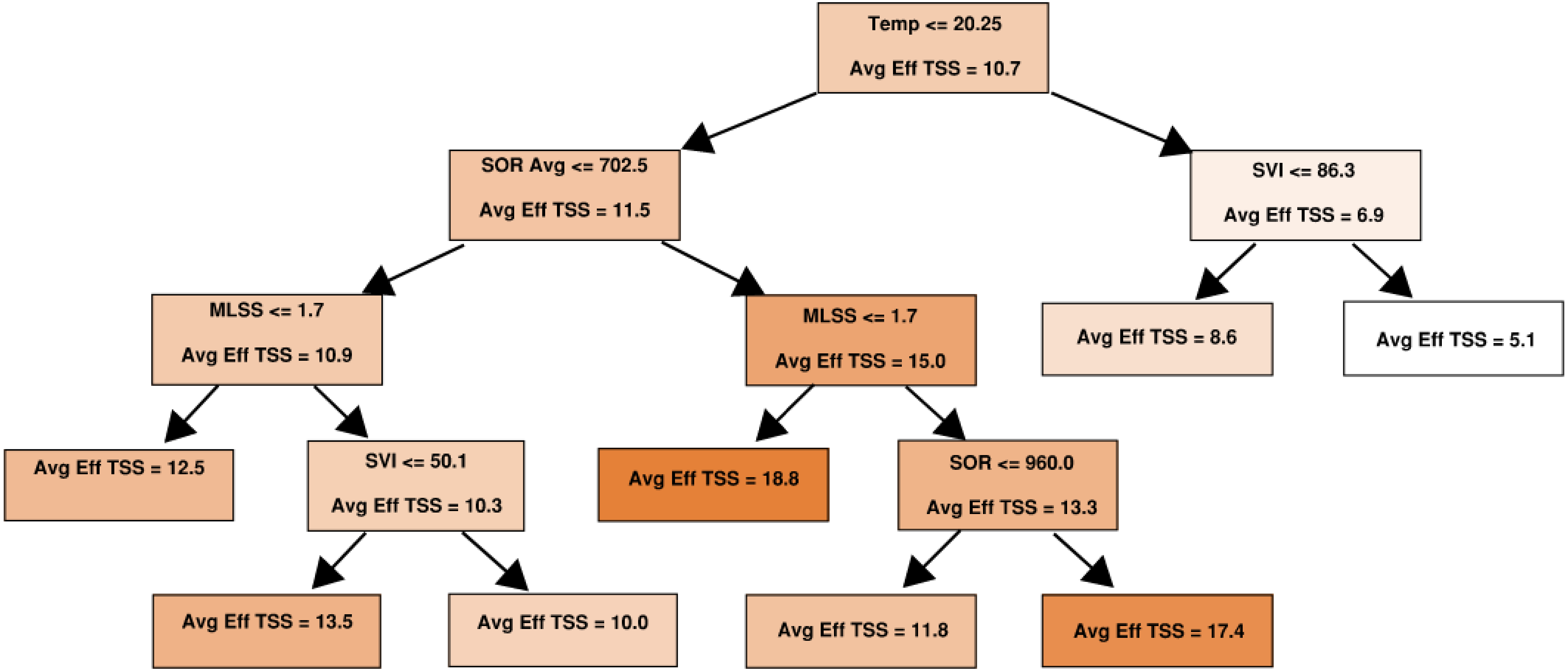
Decision Tree Results



Decision Tree Results



Decision Tree Results

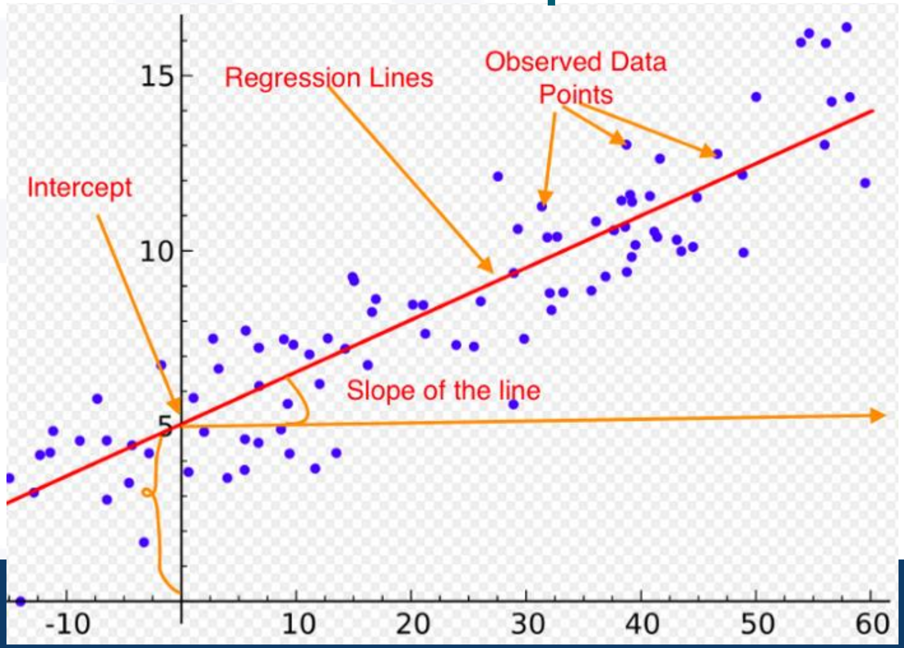


Data Analysis Tool: Stepwise Multiple Regression (Statistical Model)

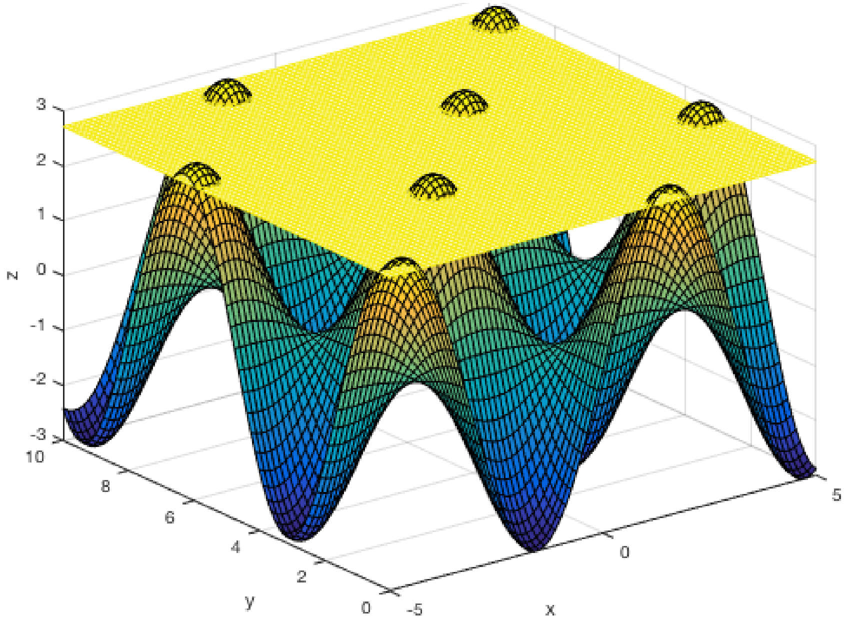
Stepwise Multiple Regression (SMR)

- → Distinct model created: do the models agree?
- → Assesses the combined effect of multiple parameters at once
- → Daylights significant parameters
- → Predicts performance

This concept



But more like this in complexity



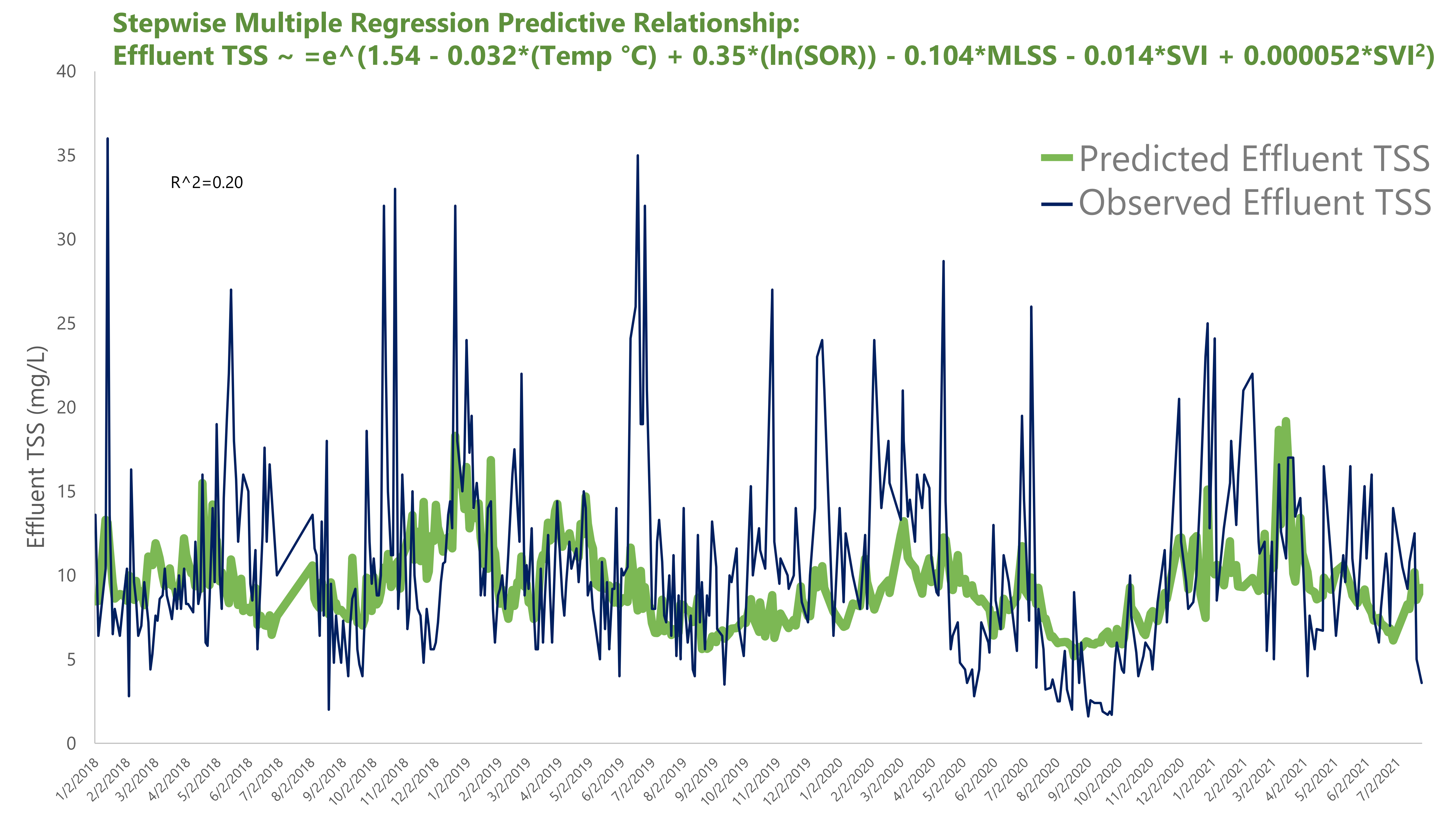
Stepwise Multiple Regression (SMR)

- Eff TSS =e^(1.54 - 0.032*(Temp °C) + 0.35*(ln(SOR))- 0.104*MLSS - 0.014*SVI + 0.000052*SVI^2)

- Parameters selected are statistically significant

- Appear in order of significance in the equation

- Most Important >>>> Less Important
  - Temp > SOR > MLSS > SVI


Stepwise Multiple Regression Predictive Relationship:
Effluent TSS ~ =e^(1.54 - 0.032*(Temp °C) + 0.35*(ln(SOR)) - 0.104*MLSS - 0.014*SVI + 0.000052*SVI²)
R^2=0.20
Predicted Effluent TSS
Observed Effluent TSS
Effluent TSS (mg/L)
0
5
10
15
20
25
30
35
40
1/2/2018
2/2/2018
3/2/2018
4/2/2018
5/2/2018
6/2/2018
7/2/2018
8/2/2018
9/2/2018
10/2/2018
11/2/2018
12/2/2018
1/2/2019
2/2/2019
3/2/2019
4/2/2019
5/2/2019
6/2/2019
7/2/2019
8/2/2019
9/2/2019
10/2/2019
11/2/2019
12/2/2019
1/2/2020
2/2/2020
3/2/2020
4/2/2020
5/2/2020
6/2/2020
7/2/2020
8/2/2020
9/2/2020
10/2/2020
11/2/2020
12/2/2020
1/2/2021
2/2/2021
3/2/2021
4/2/2021
5/2/2021
6/2/2021
7/2/2021

Model Results and Conclusions

Decision Tree vs SMR: These Models Agree

	Decision Tree (Machine Learning)	**SMR** (Statistical Model)
Parameters Selected (descending order of significance/importance)	**Temp, SOR & SVI, MLSS**	**Temp, SOR, MLSS, SVI, SVI^2**
Positively Correlated Parameters	**SOR**	**SOR**
Negatively Correlated Parameters	**Temp, MLSS, SVI**	**Temp, MLSS, SVI**
Parameters Not Selected	**DOB (max)**	**DOB (max)**
Goodness of Fit (R^2)	**0.25**	**0.20**
Significance of Model	**Unknown**	**P=2.2 x 10^-16**
Number of Coefficients (Model Complexity)	**7**	**6**

What did these data analysis tools tell us about secondary clarifier performance?

- Temperature is the most important parameter for predicting effluent TSS
  - Lower temp -> higher effluent TSS
  - Viscosity impacts, density currents, biology?
  - Correlation with high flows?
  - Other seasonal changes?
  - Easy to measure!
- Other relationships suggest discrete settling & flocculation limitations (rather than solids loading / zone settling limitations)
  - MLSS and SVI are both negatively correlated with effluent TSS
    - optimum SVI at 130 mL/g (well above the median)
  - Depth of blanket was not important
- SOR is positively correlated with effluent TSS
  - Effect of currents / short circuiting

Takeaways

- → Look for the right data analysis tool for the job
  - Can produce quick results
  - Can highlight new insight
- → Data analysis can be efficient and a broadly applicable tool
  - Leveraging readily-available *and free* data
  - Many of these types of problems
- → Hybrid data processing can drive robust system understanding
  - Combine process analytical techniques with data analysis tools and operational knowledge

Acknowledgements

LEWISTON-AUBURN
WATER POLLUTION CONTROL AUTHORITY

RStudio

python

scikit learn

Erik Osborn – eosborn@woodardcurran.com
Julia Wahl – jwahl@woodardcurran.com

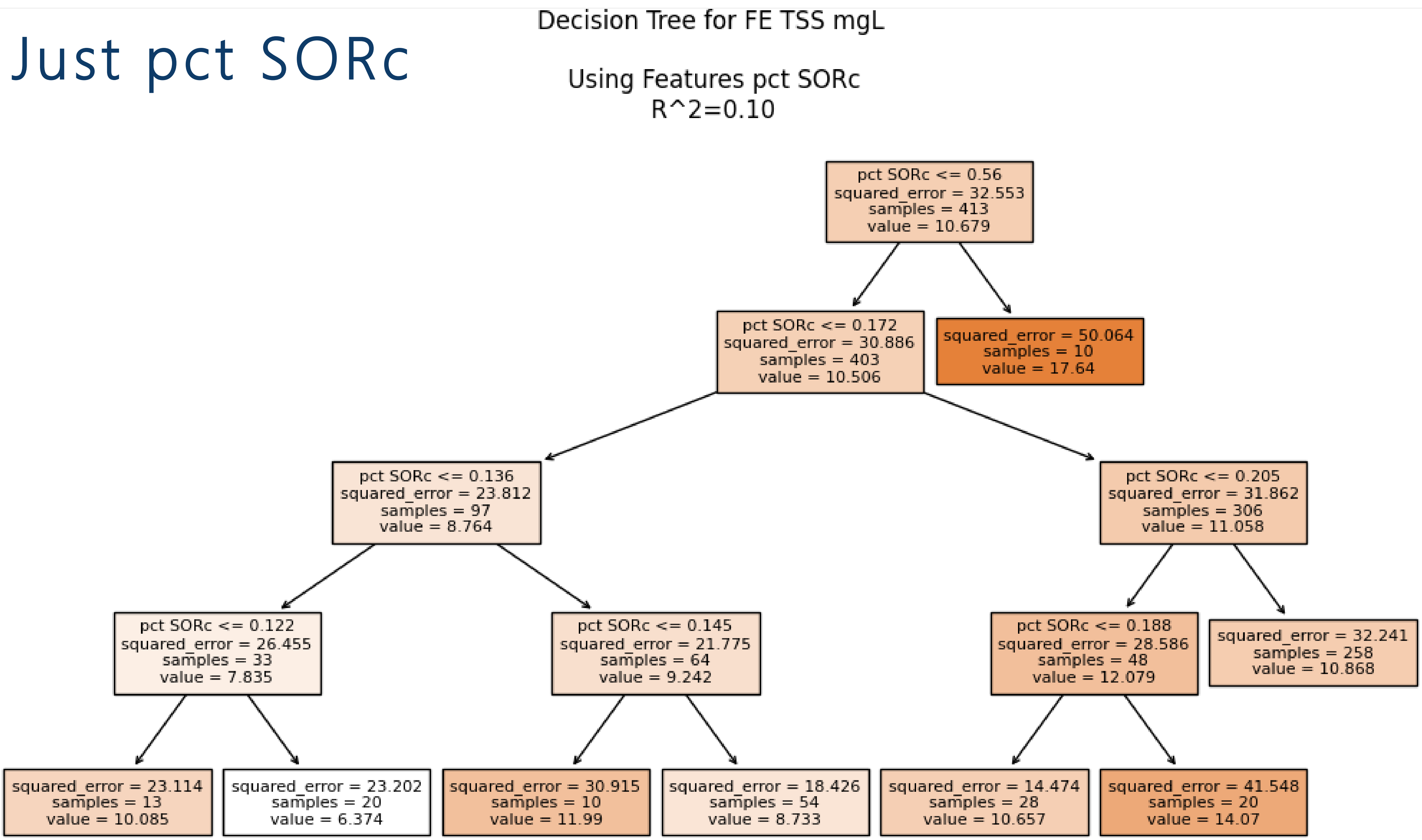
Thank you for your attention

Erik Osborn – eosborn@woodardcurran.com
Julia Wahl – jwahl@woodardcurran.com

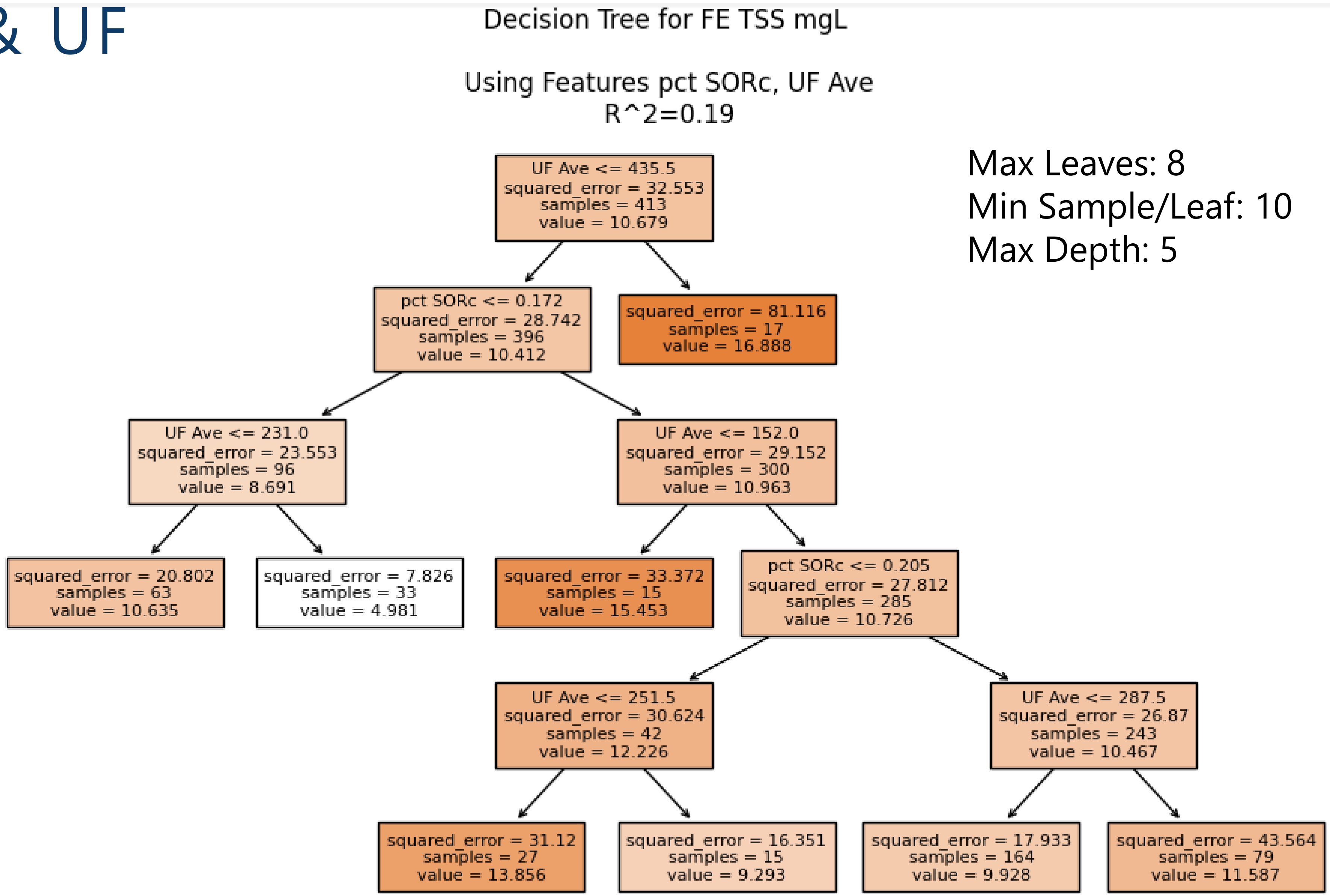
Critical Capacity

- →Create a new variable, percent of critical capacity
- →See if the decision tree can group the data based on this variable
- →Then try adding underflow rate

Just pct SORc

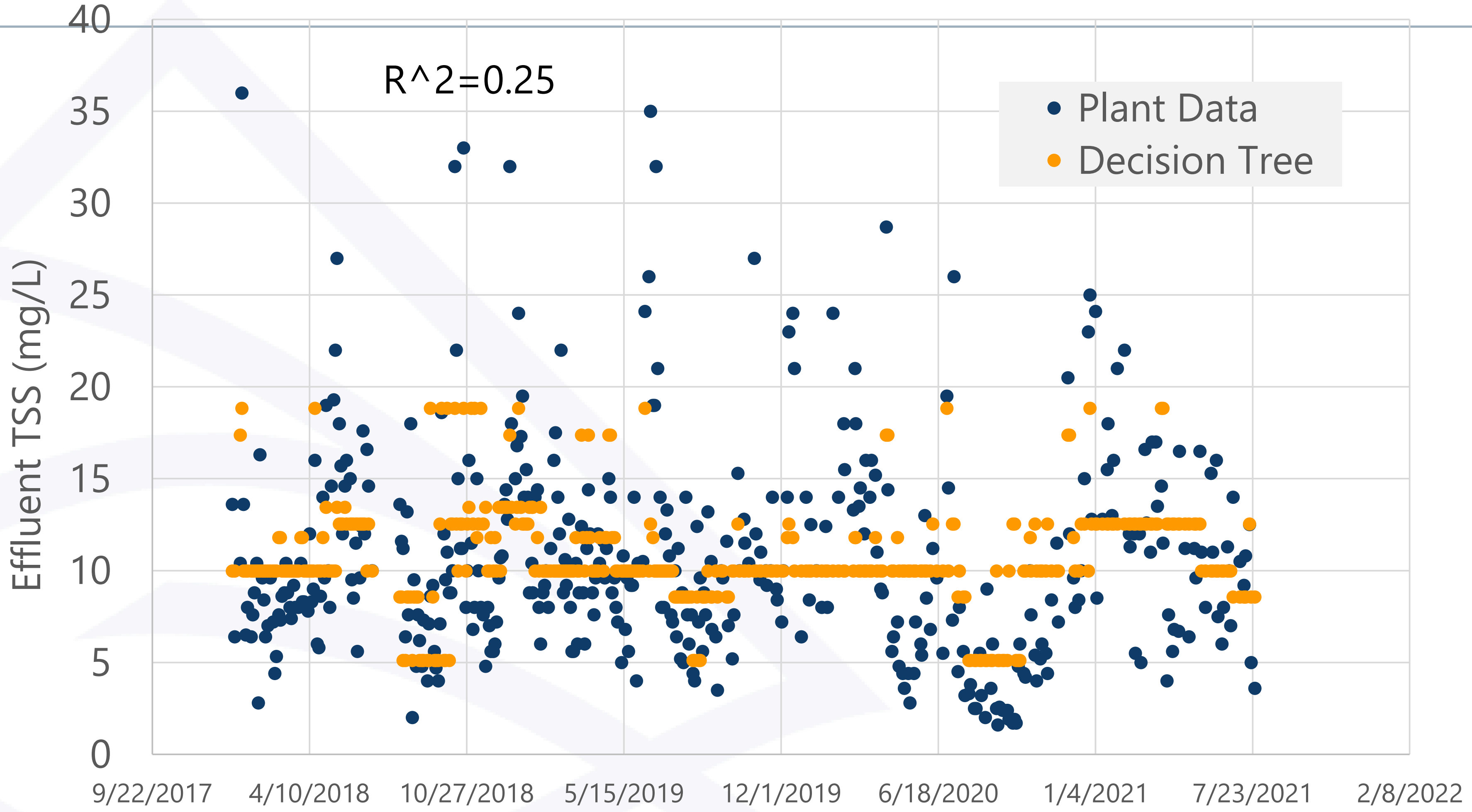


pct SORc & UF

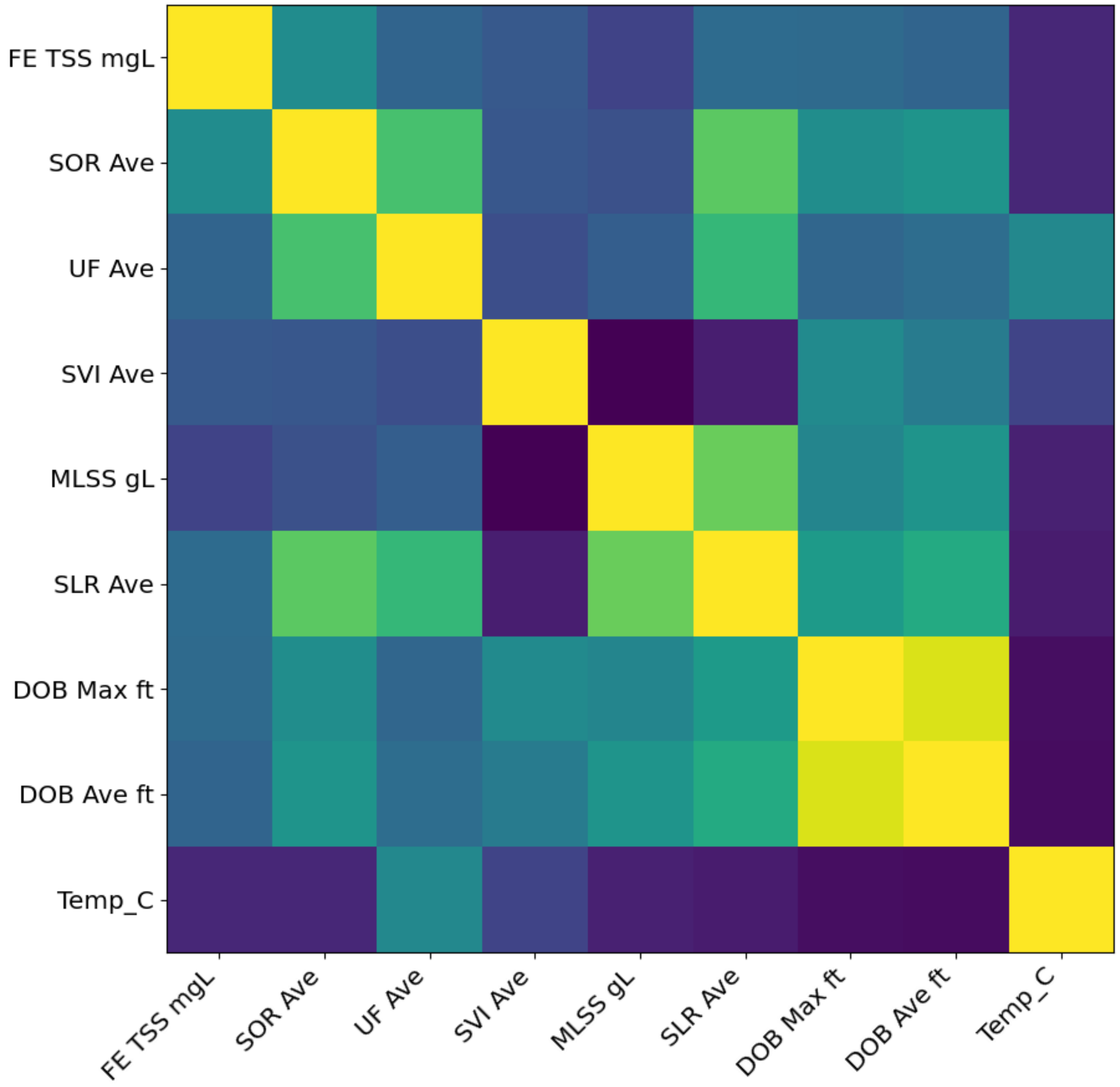


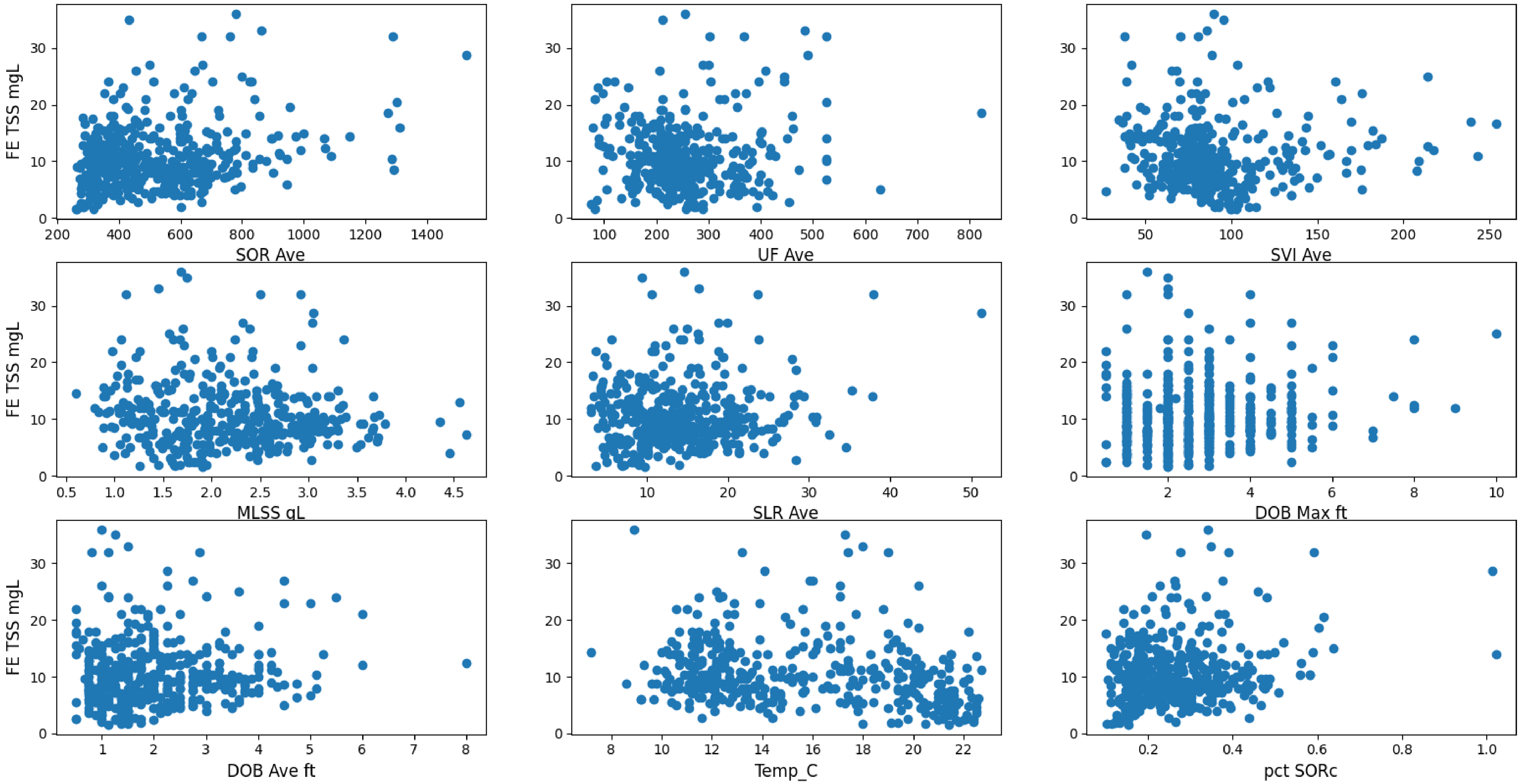
Max Leaves: 8
Min Sample/Leaf: 10
Max Depth: 5

Decision Tree Fit

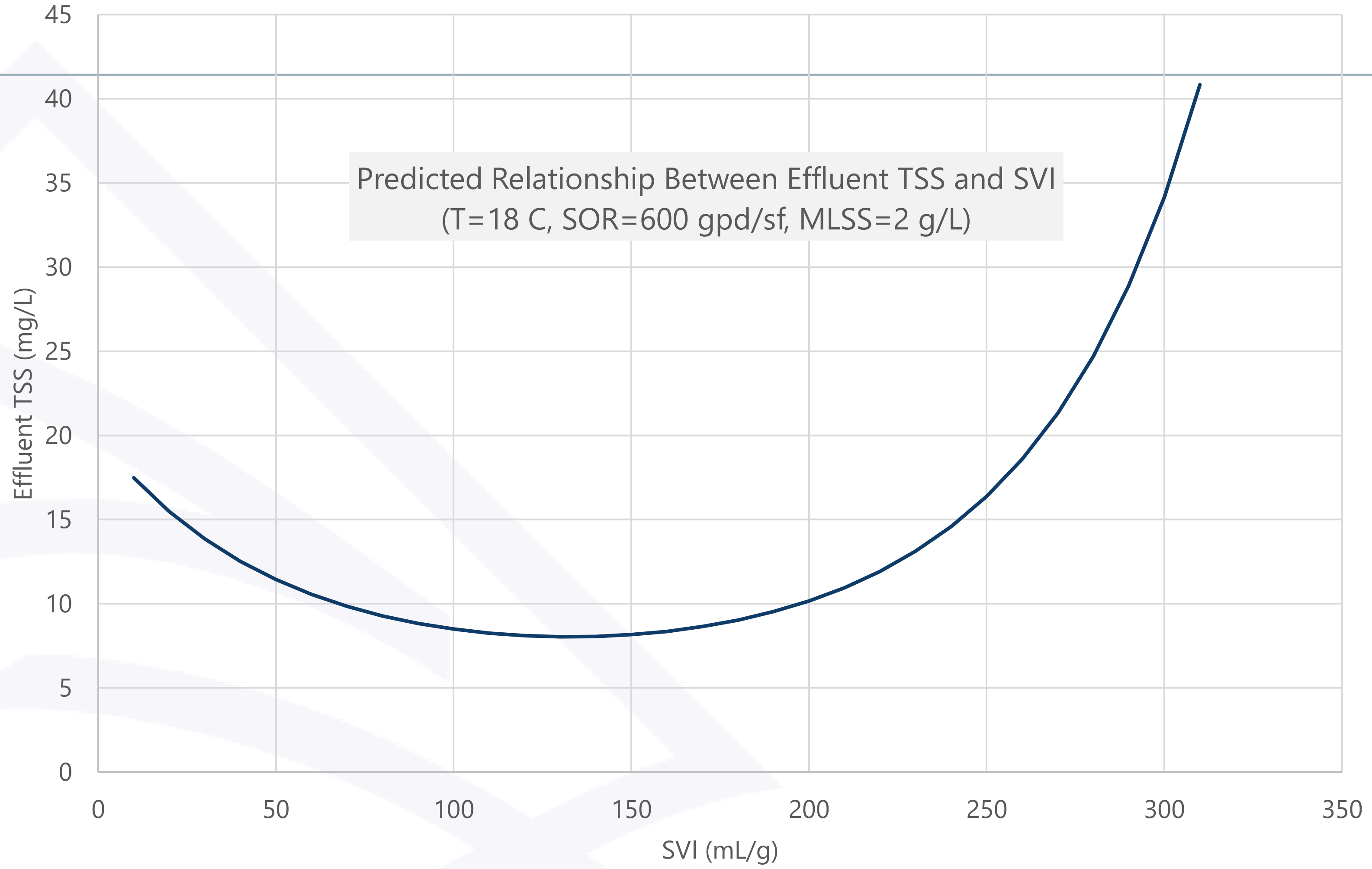


Correlation between pairs of parameters




FE TSS mgL
SOR Ave
UF Ave
SVI Ave
MLSS gL
SLR Ave
DOB Max ft
DOB Ave ft
Temp_C
pct SORc

SMR Identifies an Optimum SVI



Motivation

- Secondary clarifiers are often a limiting process for plant performance and capacity
- As a first step for evaluating clarifier performance we typically examine available plant daily data
  - Daily plant data is readily available, free, and contains lots of information!
  - We use both visual analysis of data and mechanistic models - such as state point analysis which predicts solids loading limitations
- Visual methods
  - can be subjective
  - challenging to evaluate the simultaneous impacts of multiple parameters
- While state point analysis is valuable, it isn't complete
  - it only predicts solids loading limitations
  - it doesn't account for performance impacts of:
    - hydraulic short circuiting
    - low SVI
- Can advanced data analysis tools be a valuable complement the traditional approach?
  - They are good at evaluating impacts of multiple parameters at the same time
  - Potentially less subjective, faster
  - May gain new insights
- This approach could be used to evaluate other processes in the plant too which also may depend on multiple parameters
  - Primary clarifiers
  - Biological treatment
  - Disinfection

Case-Study: What can data analysis tools tell us about secondary clarifier performance?

- Lewison-Auburn Water Pollution Control Authority
  - 8 MGD Average Flow, Secondary Treatment Plant
- Plant typically meets its effluent TSS limits of 30/45/50 mg/L
- Clarifiers operate well below the point of solids overload predicted by SPA.
  - < 65% critical capacity 99% of the time
- SVI is low (median is 85 mL/g)
- Considerable variability in effluent TSS (1 to 40 mg/L)
- 3 ½ years of daily data, 413 days with data for all parameters of interest:
  - Surface overflow rate, SVI, MLSS concentration, depth of blanket, temperature
- Can data analysis tools help us understand causes of performance (TSS) variability when operating below critical capacity?