CSC 558

Assignment 5: Full paper for predicting DO and pH in stream water from presence of limestone

Dr. Parson

Nathan Wilson Rew

May 10, 2023

4.2.a. Data sources:

- USGS time series builder, sites 01454700 (limestone) and 04213152 (shale)
- DCNR Pennsylvania county bedrock maps (Lehigh, Northampton, Erie) and overall stream map
- Western Pennsylvania Conservancy for overall PA bedrock geology

4.2.b. Intended goal: Analyze dissolved oxygen and pH separately as a function primarily of the presence of bedrock limestone vs. bedrock shale, and any other target variables of interest if time permits. This is an extension of prior work done on analyzing chlorophyll, dissolved oxygen, and pH in wastewater ponds by Wallace, Champagne, and Hall to determine if bedrock type has a significant impact on the presence of dissolved oxygen or pH.

4.2.c. Data cleaning:

- Main steps:
  i. Downloaded data for all USGS water sampling sites on a single day, for all sites that contained any critical attribute
  ii. Extracted site definitions, attribute definitions, and sites containing all critical attributes via Python script
  iii. Downloaded coordinates for the remaining sites, mapped them out and overlaid the result on top of a map of PA bedrock
  iv. Chose two sites, one with and one without limestone bedrock, and downloaded data from both sites for all of 2019
  v. Used another Python script to replace attribute IDs with attribute names; replace site number with limestone 0/1 boolean and add minute-of-day, minute-of-year, and day-of-year to the data sets
  vi. Used yet another python script to discard samples with timestamps not present in the other data set, samples that are missing attributes, attributes that are now redundant (original timestamps) or irrelevant metadata; and to merge the two data sets
  vii. Formatted the JSON data in a way that Weka can read it, then imported from JSON and exported to ARFF
- Problems encountered:
  i. The overlaid state maps used different projection styles, and it was hard to tell if the limestone site was not (too) contaminated by flowing over acidic features. Solution: Found county maps from the PA DCNR which show a proper overlay of streams and

bedrock (i.e. no issues with different types of projection). Stitched them together in an image-editing program where necessary and confirmed little to no potential for acidic contamination in the limestone site.

ii. During winter months, most attributes are not measured, likely to avoid ice damage. Solution: Discard entries with missing data.

iii. The two sites have different amounts of entries in their data sets. Solution: Extract all timestamps from one data set, iterate over the other and discard entries that contain timestamps not in the extracted set. Switch the two data sets and repeat.

4.2.d. Application of results to a commercial or research setting:

- I believe the overall application of my results would be in potentially providing a better understanding of the factors that contribute to the conditions of freshwater ecosystems. Any discovered effect of bedrock on water conditions could also be useful in locating sites where some of the target variables are lower or higher, or alternatively ignoring it as a factor altogether.

4.2.e. Anticipated techniques: Numeric estimation using linear regression, M5P trees, and decision trees. If time permits, then also scripting parallel Weka instances to run a wide variety of remaining numeric techniques. The models produced by these approaches will shed some light on how much, or whether or not, the variables of interest contribute to DO, pH, etc.

5.2.a. No additional data was collected for part 5.

5.2.b. The intended goal of determining the presence of a relationship between bedrock type (acidic vs. basic) and dissolved oxygen and/or pH in stream water was partially met. The results indicate that there is little to no relationship between bedrock and dissolved oxygen, but the results for pH seem inconclusive. First, the models indicate that one should expect to see more acidic stream water when a more basic type of bedrock lies beneath the stream channel. Second, only two sites were studied for this analysis. Third, the two sites have significantly different aboveground conditions, where the site with more acidic bedrock is fed by a stream that meanders through a wooded area while the site with more basic bedrock is fed by a stream that flows through several miles of significantly developed land; it is possible that any effect of bedrock on stream conditions is dwarfed by the aboveground conditions, especially where there is potential for stream pollution. Ultimately, this analysis would need to be redone with better control over independent variables to say anything conclusive about whether bedrock acidity affects stream pH.

5.2.c. Steps taken:

Dissolved oxygen analysis, all attributes (normalized)

i. Linear regression

```
Linear Regression Model

Dissolved oxygen =

     0.5605 * Discharge N +
    -4.3172 * Temperature celsius N +
     0.3502 * Specific conductance N +
     0.498  * Minute of day N +
    -0.0029 * Day of year N +
    -0.082  * Limestone +
    11.8588

Time taken to build model: 0.22 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.8027
Mean absolute error                      0.4526
Root mean squared error                  0.6685
Relative absolute error                 48.0307 %
Root relative squared error             59.6382 %
```

ii. M5P tree

```
=== Summary ===

Correlation coefficient                  0.9822
Mean absolute error                      0.0903
Root mean squared error                  0.2107
Relative absolute error                  9.5799 %
Root relative squared error             18.7929 %
```

This provides a baseline understanding of how the relationship between temperature and DO manifests in this data set. Note the extremely high correlation coefficient for the M5P tree analysis, as well as the relatively large coefficient for temp. C in the linear regression analysis. As a first step from here, we will try a sanity check to see how limestone/shale alone correlates to DO.

----

i. Linear regression

```
Linear Regression Model

Dissolved oxygen =

    -0.1597 * Limestone +
     9.449

Time taken to build model: 0.01 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.0692
Mean absolute error                      0.9354
Root mean squared error                  1.1182
```

```
Relative absolute error              99.2795 %
Root relative squared error            99.7536 %
Total Number of Instances           25440
-------------------------------------------------------------------------------------------------------------------- ------
```

Unfortunately, it does not seem like there is much to say here. Limestone, by itself or alongside other attributes, has little to no effect on dissolved oxygen levels based on the data from the two sites in this data set. Pretty much every other attribute is also known to be a predictor of dissolved oxygen levels, or is redundant with a predictor, so we cannot really bring one of them in to try to synergize with the shale/limestone boolean attribute with hope to improve its accuracy as a predictor.

```
-------------------------------------------------------------------------------------------------------------------- ------
```

pH analysis, all attributes (normalized)

i. Linear regression

```
Linear Regression Model

pH =

    -0.4241 * Discharge N +
     0.1082 * Temperature celsius N +
     0.4524 * Specific conductance N +
    -0.0932 * Minute of year N +
     0.1618 * Minute of day N +
    -0.2638 * Limestone +
     7.9417

Time taken to build model: 0.02 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient               0.897
Mean absolute error                    0.1117
Root mean squared error                  0.1469
Relative absolute error              38.5029 %
Root relative squared error            44.192  %
Total Number of Instances           25440
```

ii. M5P tree

```
=== Summary ===

Correlation coefficient               0.9927
Mean absolute error                    0.0259
Root mean squared error                  0.0401
Relative absolute error                8.9197 %
Root relative squared error             12.0614 %
Total Number of Instances           25440
```

The models for pH have very high correlation coefficients too, and limestone has some presence as a factor in the linear regression formula. However, I worry that this is a case of redundancy; note that specific conductance has one of the larger coefficients in the analysis. Specific conductance is a measure of the concentration of ions in solution, and acids/bases have greater quantities of positive/negative ions respectively.

----

pH analysis, without specific conductance (normalized)

i. Linear regression

```
Linear Regression Model

pH =

    -0.5686 * Discharge N +
     0.1351 * Temperature celsius N +
    -0.0233 * Minute of year N +
     0.1557 * Minute of day N +
    -0.4349 * Limestone +
     8.1941

Time taken to build model: 0.02 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.8853
Mean absolute error                      0.118 
Root mean squared error                  0.1546
Relative absolute error                 40.679  %
Root relative squared error             46.5051 %
Total Number of Instances            25440
```

ii. M5P tree

```
=== Summary ===

Correlation coefficient                  0.9924
Mean absolute error                      0.0269
Root mean squared error                  0.0408
Relative absolute error                  9.2655 %
Root relative squared error             12.2837 %
Total Number of Instances            25440
```

It turns out the correlation coefficient stayed almost the same while limestone’s influence increased – a good sign? Next question: Why is there such a strong correlation with discharge? Well, if we observe the limestone site...

(source)

...And the shale site...

(source)

...visually, we can see that they have vastly different amounts of water flowing through them. This seems promising however, as it suggests that there is a relationship between pH and the choice of sampling site. Let us see what happens if we remove discharge and gage height as factors (since gage height is redundant with discharge), then narrow it down from there.

----

pH analysis using minute of year and limestone (normalized)

i. Linear regression

```
Linear Regression Model

pH =

     0.1637 * Minute of year N +
    -0.5603 * Limestone +
     8.2604

Time taken to build model: 0.01 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.8522
Mean absolute error                      0.1348
Root mean squared error                  0.174
Relative absolute error                 46.4754 %
Root relative squared error             52.3256 %
Total Number of Instances            25440
```

ii. M5P tree

```
=== Summary ===

Correlation coefficient                  0.9818
Mean absolute error                      0.0462
Root mean squared error                  0.0644
Relative absolute error                 15.9297 %
Root relative squared error             19.3611 %
Total Number of Instances            25440
```

Even with just these two variables, the M5P tree still has >98% accuracy. Temperature is less of a factor here, as when we swap it in place of minute-of-year, the correlation coefficient drops to 0.8522. What is interesting is that there seems to be a negative correlation between limestone and pH even though shale = 0, limestone = 1, and limestone has a higher pH than shale. There could be issues with pollution – the river feeding into the limestone site winds through a developed area east of Allentown for several miles before the sampling station, whereas the shale site's stream is a bit more secluded in a wooded area. More sites, further from developed areas, would need to be sampled to be certain.

5.2.d. MultilayerPerceptron tests

i. All attributes

```
Correlation coefficient                  0.9519
Mean absolute error                      0.0795
Root mean squared error                  0.103
```

```
Relative absolute error              27.4026 %
Root relative squared error            30.9723 %
Total Number of Instances            25440
```

ii. All except discharge and gage height

```
Correlation coefficient              0.9315
Mean absolute error                   0.0904
Root mean squared error                0.1218
Relative absolute error              31.1613 %
Root relative squared error            36.6427 %
Total Number of Instances            25440
```

iii. Temperature, minute of year, minute of day, day of year, limestone

```
Correlation coefficient              0.8926
Mean absolute error                   0.108
Root mean squared error                0.1511
Relative absolute error              37.245  %
Root relative squared error            45.4549 %
Total Number of Instances            25440
```

iv. Minute of year and limestone

```
Correlation coefficient              0.8327
Mean absolute error                   0.1413
Root mean squared error                0.1867
Relative absolute error              48.734  %
Root relative squared error            56.1636 %
Total Number of Instances            25440
```

After trying several different setups with the MultilayerPerceptron, we can see that it is not as accurate as the M5P tree based only on minute of year and limestone. It is more accurate when we leave it more attributes to work with, but as discussed earlier these attributes are redundant and may be giving the model too much wiggle room.

5.2.e The results of this analysis could have some applications, although probably not the ones predicted in section 4.2. The unexpected negative correlation between limestone and pH could be a cause for investigating stream conditions for pollution or other issues near the limestone sampling site, if this has not already been done before. I do not think there is enough information here to make any solid conclusions about pH, as any potential expected effects of bedrock type seem to be overpowered by some other factor, but I think it is reasonable to argue that the extremely small coefficient for bedrock type when predicting dissolved oxygen indicates that either no such relationship exists or the relationship is negligibly weak.