ProUCL Utilization 2020
Trend Analysis

Presenters:
Travis Linscome-Hatfield,
Polona Carson
Summary of Session 1: ProUCL A to Z

• Objective:
  • Get familiar with ProUCL and some commonly used data analysis features

• We talked about:
  • Starting ProUCL
  • Preparing data for analysis and loading in ProUCL
  • Basics of dealing with missing values and NDs
  • Exploratory Data Analysis
  • Hypothesis testing
Learning Objectives

• Objectives
  • Get familiar with ProUCL features for trend analysis

• Today we will discuss:
  • Exploring the data with Time Series Plot
  • Mann-Kendall trend test
  • Theil-Sen trend test
  • Ordinary Least Square Regression
    • Analysis of residuals
ProUCL Software

- Statistical software for environmental data analysis
- User Guide
  - Provides instructions on how to use ProUCL
- Technical Guide
  - Provides detailed background on statistical methods
Navigating ProUCL
Data Sets

- MW-1-8-9.xls
  - In ProUCL Data folder
- Trend-MW-Real-data wln.xls
  - On training website
Nondetects (NDs)

- **High non-detects** always create severe problems in analyzing data
  - They should be removed from the data set before further analysis

- **Rule of thumb to reject NDs**
  - always reject NDs greater than the largest detect
  - Use judgement to reject NDs greater than 10x the smallest detect

- **Reasonable substitutions for remaining NDs:**
  - Substitute only if there is a small number of Nds (10-15%)
  - ½ of reporting limit (RL)
  - ½ of detection limit (DL)
Explore the Data with Time Series Plot

• Diagnostic and exploratory graphical tool
  • Is there a trend?
  • Linear or exponential?

• Grouping variable allows parallel display and comparison

• Helps to evaluate:
  • The presence of temporal / spatial variability
  • Is data transformation needed?
Are you Looking at your Data the Right Way?

- **Dependent variable or response**
  - **Contaminant**
    - MW-28 column
- **Independent variable or predictor**
  - **Time**
    - Index column (time in days)

Is contaminant \(\text{(dependent variable)}\) decreasing / increasing over time \(\text{(independent variable)}\)
1. Click Statistical Tests ► Trend Analysis ► Mann-Kendall.
Advantages / Disadvantages

• Good for detecting monotone trends
• Identifies trend
• No information about the slope
Advantages / Disadvantages

• Identifies trend
• Quantifies slope
  • The rate of decrease / increase
• Robust to presence of extreme values
• Handles only 1 observation at each sampling event
**When to transform the data?**

<table>
<thead>
<tr>
<th>Visible exponential decay trend</th>
<th>Visible linear decay</th>
</tr>
</thead>
<tbody>
<tr>
<td>• 1(^{st}) order reactions</td>
<td>• 0 order reactions</td>
</tr>
<tr>
<td>• Rate of decay proportional to amount of contaminant present</td>
<td>• Decay rate constant rather than proportional to amount of contaminant</td>
</tr>
<tr>
<td>• Dilution with little or no desorption from soil</td>
<td>• No transformation needed</td>
</tr>
<tr>
<td>• Log transformation is needed</td>
<td></td>
</tr>
</tbody>
</table>
Ordinary Least Square (OLS) Regression

- Select regression variables
- Fit regression
- Check assumptions
- Evaluate model and Interpret Results
Regression Plot

Classical Regression

- **Confidence intervals**
- **Prediction intervals**

### OLS
- \( n \): 16
- Slope: -1.6372
- Intercept: 2.1637246
- R-sq: 0.9432
- R: -0.9183
- Scale Estimate: 372.9502
- Pr-value (Intercept): 0.0000
- Pr-value (Slope): 0.0000

**Mann-Whitney**
- U: -130.000
- SD of U: 26.4058
- Standardized U: -4.9999
- Approximate p-value: 0.0000

**Confidence Coefficient**: 0.9900

*Red = Prediction Interval*
*Green = Confidence Interval*
Evaluate Model – Regression Equation

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimates</th>
<th>Std. Error</th>
<th>T-values</th>
<th>p-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>intercept</td>
<td>2164</td>
<td>165.3</td>
<td>13.09</td>
<td>5.793E-10</td>
</tr>
<tr>
<td>Time (days)-6 (slope)</td>
<td>-1.637</td>
<td>0.176</td>
<td>-9.276</td>
<td>7.7292E-8</td>
</tr>
</tbody>
</table>

Contaminant = 2164 -1.637 * Time

Small p-values indicate significance p<0.05
### OLS ANOVA Table

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>SS</th>
<th>DOF</th>
<th>MS</th>
<th>F-Value</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>11952431</td>
<td>1</td>
<td>11952431</td>
<td>86.05</td>
<td>0.0000</td>
</tr>
<tr>
<td>Error</td>
<td>2222368</td>
<td>16</td>
<td>138898</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>14174799</td>
<td>17</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Small p-value indicates significance p<0.05

- % of variability explained by regression model: 84.3% (R^2)

<table>
<thead>
<tr>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Sqrt(MSE) = Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.843</td>
<td>0.833</td>
<td>372.7</td>
</tr>
</tbody>
</table>
Advantages / Disadvantages

• Standard statistical method for identifying trends
• Assumptions for residuals need to be satisfied
• Confidence bands can be used to determine compliance with fixed standards even when a trend is apparent
  • Concentration of contaminant is changing
• Sensitive to extreme values (outliers)
Assumptions for Residuals

\[ Residuals = \text{observed value} - \text{fitted value} \]

• Constant variance (homoscedasticity of residuals)
• Independent
• Normally distributed
Prepare OLS Regression Table to Check Assumptions

- Copy Regression Table to EXCEL
  - May have leading spaces!!
- Helpful trick to remove leading spaces:
  - select Y vector to Res/Scale columns
  - choose replace
  - hit spacebar in find field,
  - Click replace all

<table>
<thead>
<tr>
<th>Obs</th>
<th>Y Vector</th>
<th>Yhat</th>
<th>Residuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2880.0</td>
<td>2164.0</td>
<td>716.3</td>
</tr>
<tr>
<td>2</td>
<td>2117.0</td>
<td>2028.0</td>
<td>89.17</td>
</tr>
<tr>
<td>3</td>
<td>1633.0</td>
<td>1900.0</td>
<td>-267.6</td>
</tr>
<tr>
<td>4</td>
<td>1845.0</td>
<td>1748.0</td>
<td>97.13</td>
</tr>
<tr>
<td>5</td>
<td>1706.0</td>
<td>1587.0</td>
<td>118.2</td>
</tr>
<tr>
<td>6</td>
<td>1719.0</td>
<td>1307.0</td>
<td>411.1</td>
</tr>
<tr>
<td>7</td>
<td>1065.0</td>
<td>1154.0</td>
<td>-88.55</td>
</tr>
<tr>
<td>8</td>
<td>831.8</td>
<td>1009.0</td>
<td>-177.7</td>
</tr>
<tr>
<td>9</td>
<td>920.6</td>
<td>1009.0</td>
<td>-88.87</td>
</tr>
<tr>
<td>10</td>
<td>424.6</td>
<td>842.5</td>
<td>-417.9</td>
</tr>
<tr>
<td>11</td>
<td>181.1</td>
<td>647.6</td>
<td>-466.5</td>
</tr>
<tr>
<td>12</td>
<td>184.9</td>
<td>647.6</td>
<td>-462.7</td>
</tr>
<tr>
<td>13</td>
<td>14.0</td>
<td>511.8</td>
<td>-497.8</td>
</tr>
<tr>
<td>14</td>
<td>26.8</td>
<td>236.7</td>
<td>-209.9</td>
</tr>
<tr>
<td>15</td>
<td>5.9</td>
<td>-44.9</td>
<td>50.81</td>
</tr>
<tr>
<td>16</td>
<td>1.7</td>
<td>-349.4</td>
<td>351.1</td>
</tr>
<tr>
<td>17</td>
<td>1.8</td>
<td>-349.4</td>
<td>351.2</td>
</tr>
<tr>
<td>18</td>
<td>5.5</td>
<td>-487</td>
<td>492.5</td>
</tr>
</tbody>
</table>
Residual Plots

• Plot 1
  • Y = Residuals
  • X = Fitted values (Estimates)
  • Indicator of constant variance
    • Random scatter

• Plot 2
  • Y = Residuals
  • X = Observation order (Event #)
  • Indicator of independence
    • Random scatter of residuals
Explore Normality of Residuals

• Import Regression table from Excel to Pro UCL
  • Goodness of Fit test

• Normality required for reliable:
  • trend test, it assumes normality of residuals!!
  • confidence and prediction intervals

• Non-normal residuals
  • Likely indicate lack of fit
    • Need to improve the model
## Trend Analysis Summary

<table>
<thead>
<tr>
<th>Method</th>
<th>Assumptions</th>
<th>Information</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Theil-Sen</td>
<td>None</td>
<td>Identifies trend; Evaluates slope</td>
<td>Fair amount of data; 1 observation per sampling</td>
</tr>
<tr>
<td>Mann-Kendal</td>
<td>None</td>
<td>Is trend monotone?</td>
<td>Min 4 points</td>
</tr>
<tr>
<td>OLS</td>
<td>Normality; Constant variance; Independence</td>
<td>Quantifies slope and intercept</td>
<td>Fair amount of data; Valid assumptions</td>
</tr>
</tbody>
</table>
Final remarks

• Take time to explore the data
• Knowing something about chemical & bio processes can be helpful
• Residuals analysis provides a lot of information
• When in doubt consult statistician
• Document well steps of analysis and decisions you make
Next session: Background Level Calculations

• Objective: Explore ProUCL features for Background to Site comparison
  • Sampling design and sample size selection
  • UTLs and background threshold values
  • Comparison of BTVs to site data
Next ProUCL Webinar

ProUCL Utilization 2020: Part 3: Background Level Calculations

Mar 9, 2020
1:00PM-2:30PM EST
Contact Information for ProUCL

Felicia Barnett, EPA SCMTSC
barnett.felicia@epa.gov

Travis Linscome-Hatfield, Neptune and Company, Inc
travis@neptuneinc.org

Polona Carson, Neptune and Company, Inc
pcarson@neptuneinc.org