Automatic measured data validation applied on hydraulic and water quality parameters in sewer systems

Nemanja Branisavljević¹
Dušan Prodanović¹
Zoran Kapelan²

¹University of Belgrade
²University of Exeter
Research challenges

Challenge 1:
Large amount of data can not be validated traditionally

Challenge 2:
Data validation procedure have to be capable to use any kind of information available

Challenge 3:
Data validation procedure have to be independent on type or number of relations between data
Measured data values

• Exact value (------) – not available
• $u$ - uncertainty
• $\varepsilon$ - error
Data validation – detecting errors if they exist

Measured data values

• Exact value (------) – not available
• $u$ - uncertainty
• $\varepsilon$ - error
Comparing measured and predicted values

Measured and calculated values are presented with intervals

Enables detection of errors in data – data validation
Comparing measured and predicted values - complex system -

Prediction results

Measured data
Suggested data validation methodology

Sorry for this mess on the screen – will make it more readable...
Suggested data validation methodology

Data preprocessing

Prediction

Validation grades

Grades interpretation

Specific for suggested methodology

Common for most methods
Data prediction
-relation errors and uncertainty-

(A) y calculated using model
(B) y calculated using measured values

(C) y calculated using model
(D) y calculated using measured values
Data prediction

-relation errors and uncertainty-
Validation grades
-data probability-

\[ p([a,b] \mid [a_1, b_1]) = \frac{p([a,b] \cap [a_1, b_1])}{p([a_1, b_1])} \]

\[
 p(x_i \mid x_i^{M_j}, X_{x_i}^{M_j}) = \begin{cases} 
 1 - \frac{\max(0, a-a_1) + \max(0, b_1-b)}{b_1-a_1}, & a_1 \leq b \\
 0, & b_1 \geq a 
\end{cases}
\]

\[
 \max\left( p(x_i \mid x_i^{M_j}, X_{x_i}^{M_j}) \right) = \max\left( P([a,b] \mid [a_1, b_1]) \right) = \frac{b-a}{b_1-a_1}
\]
Validation grades
-weighted average of data probabilities-

Prediction results
Measured data
Validation grades
-likelihood of calculated values-

**Weights**—likelihood of predicted value

$$w_i^{M_{R_j}, x_i} = p \left( x_i^{M_{R_j}, x_i} \mid X_{x_i}^{M_{R_j}, x_i} \right)$$

Maximization of a sum of likelihoods - EM algorithm

$$J = \max \sum_i \sum_j p \left( x_i^{M_{R_j}, x_i} \mid X_{x_i}^{M_{R_j}, x_i} \right)$$

**E step:**

$$p \left( x_i^{M_{R_j}, x_i} \mid X_{x_i}^{M_{R_j}, x_i} \right) = \sum_{X_{x_i}^{M_{R_j}, x_i}} \left( \frac{\partial M_{R_j, x_i} \left( X_{x_i}^{M_{R_j}, x_i}, \theta \right)}{\partial X_{x_i}^{M_{R_j}, x_i}} \right)^{-1} p_{x_i}$$

**M step:**

$$p_{x_i} = \sum_j p \left( x_i \mid x_i^{M_{R_j}, x_i}, X_{x_i}^{M_{R_j}, x_i} \right) \times \frac{p \left( x_i^{M_{R_j}, x_i} \mid X_{x_i}^{M_{R_j}, x_i} \right)}{\sum_j p \left( x_i^{M_{R_j}, x_i} \mid X_{x_i}^{M_{R_j}, x_i} \right)}$$
Validation grades

1. Not normalized
   \[ x_i^{\text{grade}} = \sum_j w_j^{M_{R_j,x_i}} \times p\left(x_i \mid x_i^{M_{R_j,x_i}}, X_{x_i}^{M_{R_j,x_i}}\right) \] (between 0 and \(a<1\))

2. Normalized
   \[ x_i^{\text{grade}} = p_{x_i}^{\text{norm}} = \left(\sum_{x_i} w_j^{M_{R_j,x_i}} \frac{p\left(x_i \mid x_i^{M_{R_j,x_i}}, X_{x_i}^{M_{R_j,x_i}}\right)}{\max\left(p\left(x_i \mid x_i^{M_{R_j,x_i}}, X_{x_i}^{M_{R_j,x_i}}\right)\right)}\right) \] (between 0 and 1)
Evaluating validation systems

\[ p = \frac{N_{\text{registered}}}{N_{\text{anomalies}} + N_{\text{missed}} + N_{\text{registered nonanomalies}}} \]

Comparing to Kappa index of coincidence:
• Higher penalty to missed anomalies, and
• Lower penalty to correct values interpreted as anomalies
Case study

Measurements of hydraulic and water quality parameters in Belgrade sewer CSO
Measurements in Belgrade sewer CSO

Measured variables: Velocity ($V$), water depth ($h$), electro-conductivity ($EC$)

### Relations between measured variables

**R₁** – Chezy-Manning equation

**R₂, R₃** – Water quality model of EC reduction

**R₄, R₅, R₆** – AR(1) models

<table>
<thead>
<tr>
<th></th>
<th>( V )</th>
<th>( h )</th>
<th>( EC )</th>
<th>( V^{t-1} )</th>
<th>( h^{t-1} )</th>
<th>( EC^{t-1} )</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>R₁</strong></td>
<td>( V^t = f_{V^t}^{R₁}(h^t) )</td>
<td>( h^t = f_{h^t}^{R₁}(V^t) )</td>
<td></td>
<td>( V^{t-1} = f_{V^{t-1}}^{R₁}(h^{t-1}) )</td>
<td>( h^{t-1} = f_{h^{t-1}}^{R₁}(V^{t-1}) )</td>
<td></td>
</tr>
<tr>
<td><strong>R₂</strong></td>
<td></td>
<td>( h^t = f_{h^t}^{R₂}(EC^t) )</td>
<td>( EC^t = f_{EC}^{R₂}(h^t) )</td>
<td></td>
<td>( h^{t-1} = f_{h^{t-1}}^{R₂}(EC^{t-1}) )</td>
<td>( EC^{t-1} = f_{EC^{t-1}}^{R₂}(h^{t-1}) )</td>
</tr>
<tr>
<td><strong>R₃</strong></td>
<td>( V^t = f_{V^t}^{R₃}(EC^t) )</td>
<td></td>
<td>( EC^t = f_{EC}^{R₃}(V^t) )</td>
<td>( V^{t-1} = f_{V^{t-1}}^{R₃}(EC^{t-1}) )</td>
<td></td>
<td>( EC^{t-1} = f_{EC^{t-1}}^{R₃}(V^{t-1}) )</td>
</tr>
<tr>
<td><strong>R₄</strong></td>
<td>( V^t = f_{V^t}^{R₄}(V^{t-1}) )</td>
<td></td>
<td></td>
<td>( V^{t-1} = f_{V^{t-1}}^{R₄}(V) )</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>R₅</strong></td>
<td></td>
<td>( h^t = f_{h^t}^{R₅}(h^{t-1}) )</td>
<td></td>
<td></td>
<td>( h^{t-1} = f_{h^{t-1}}^{R₅}(h) )</td>
<td></td>
</tr>
<tr>
<td><strong>R₆</strong></td>
<td></td>
<td></td>
<td>( EC^t = f_{EC}^{R₆}(EC^{t-1}) )</td>
<td></td>
<td></td>
<td>( EC^{t-1} = f_{EC^{t-1}}^{R₆}(EC) )</td>
</tr>
</tbody>
</table>
Chezy-Manning equation

\[ V(h_0, C, h) = C \times R(h_0, h)^{2/3} \]

\[ C = \sqrt{I_d / n} \]

\[ M_{R_1, v}: V(h) = [3.8, 1.9] \times R([239, 291], h)^{2/3} \]
AR(1) models

\[ x^t = ax^{t-1} + b \]

Autocorrelation coefficient

\[ R_{ff}(\tau) = \int_{-\infty}^{\infty} f(t + \tau) \overline{f}(t) \, dt \]

Model calibration results

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>h</td>
<td>V</td>
<td>EC</td>
</tr>
<tr>
<td>a</td>
<td>0.97427</td>
<td>0.97258</td>
<td>0.99712</td>
</tr>
<tr>
<td>b</td>
<td>[-33, 52]</td>
<td>[-0.17, 0.19]</td>
<td>[-59, 49]</td>
</tr>
</tbody>
</table>
Results

-Validation grades-
Results
-comparison with manual data validation-

Threshold value = 0.99
Nine validation methods

<table>
<thead>
<tr>
<th></th>
<th>No anomalies</th>
<th>No detected</th>
<th>No missed</th>
<th>No false</th>
<th>(p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(M_1)</td>
<td>244</td>
<td>24</td>
<td>220</td>
<td>0</td>
<td>0.052</td>
</tr>
<tr>
<td>(M_2)</td>
<td>244</td>
<td>0</td>
<td>244</td>
<td>28</td>
<td>0.000</td>
</tr>
<tr>
<td>(M_3)</td>
<td>244</td>
<td>0</td>
<td>244</td>
<td>0</td>
<td>0.000</td>
</tr>
<tr>
<td>(M_4)</td>
<td>244</td>
<td>202</td>
<td>42</td>
<td>220</td>
<td>0.399</td>
</tr>
<tr>
<td>(M_5)</td>
<td>244</td>
<td>187</td>
<td>57</td>
<td>334</td>
<td>0.294</td>
</tr>
<tr>
<td>(M_6)</td>
<td>244</td>
<td>224</td>
<td>20</td>
<td>359</td>
<td>0.360</td>
</tr>
<tr>
<td>(M_7)</td>
<td>244</td>
<td>109</td>
<td>135</td>
<td>896</td>
<td>0.085</td>
</tr>
<tr>
<td>(M_8)</td>
<td>244</td>
<td>237</td>
<td>11</td>
<td>483</td>
<td>0.321</td>
</tr>
<tr>
<td>(M_9)</td>
<td>244</td>
<td>71</td>
<td>173</td>
<td>2725</td>
<td>0.023</td>
</tr>
</tbody>
</table>

Averaging validation results:

\[
p = \frac{N_{\text{registered}}}{N_{\text{anomalies}} + N_{\text{missed}} + N_{\text{registered nonanomalies}}} = \frac{195}{244 + 49 + 58} = 0.56
\]
Conclusions

• New theoretical framework for data validation
• No restrictions about number and type of additional information that may be used in the validation process
• No restrictions of type and number of relations that may be used
• The validation system may be evaluated and improved by improving the relations between the data values
Automatic measured data validation applied on hydraulic and water quality parameters in sewer systems

Nemanja Branisavljević¹
Dušan Prodanović¹
Zoran Kapelan²

¹University of Belgrade
²University of Exeter