Groundwater Modeling with Machine Learning Techniques: Ljubljana polje Aquifer

Klemen Kenda, Matej Čerin, Mark Bogataj, Matej Senožetnik, Kristina Klemen, Petra Pergar, Chrysi Laspidou and Dunja Mladenić

EWaS 2018, Lefkada, Greece
Outline

• Motivation and use-case presentation
• Introduction to data mining and machine learning
• The Quest for the best data-driven model for groundwater level
• Results
Motivation

• groundwater levels are the principal source of information about hydrologic stress
• integrate groundwater into urban design
• can we contribute or improve previous results from process models
Data Mining Process

Data Mining / Machine Learning / Stream Mining

• Cross Industry Standard Process for Data Mining
• Holistic approach to data-driven modeling – useful for real-world applications
• From understanding of needs to deployment of models
• Data Preparation is the most time-consuming step
Definitions

Data Mining / Machine Learning / Stream Mining

• Data Mining: Extraction of useful information from data

• Data Mining is application of Machine Learning techniques to solve real-life data analysis problems
Supervised Learning
Data Mining / Machine Learning / Stream Mining

• **Model** that can predict continuous or nominal attributes
• Different than process-based models (!); underlying mechanisms are **not** important
• Based only on data
• Domain knowledge introduced through feature engineering
• Stream Mining
Input Data (features)
- Daily aggregates of weather data
  - Temperature avg / min / max
  - Precipitation
  - Snow
  - Sun duration
  - Cloud cover

Label
- groundwater levels for 5 sensors in Ljubljana polje aquifer
  (1 measurement / day)
The Quest – 1/4 (direct approach)
Data – 2/4

Input Data (features)

• Daily aggregates of weather data
  • Temperature avg / min / max
  • Precipitation
  • Snow
  • Sun duration
  • Cloud cover

Label

• groundwater level changes for 5 sensors in Ljubljana polje aquifer
  (1 measurement / day)
The Quest – 2/4 (differential approach)
Feature engineering

- process of deriving new relevant features for modeling
- different moving aggregates (mean, min, max, variance)
Correlated features
Data – 3/4

Input Data (features)

- Daily aggregates of weather data
  - Temperature avg / min / max
  - Precipitation
  - Snow
  - Sun duration
  - Cloud cover

- Shifted over 1 – 100 days

- Averaged over 1 – 100 days

Label

- groundwater level changes for 5 sensors in Ljubljana polje aquifer (1 measurement / day)
The Quest – 3/4 (with feature engineering)
The Quest – 4/4 (final results)
Final results

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>$R^2$</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear regression</td>
<td>0.624</td>
<td>$2.23 \cdot 10^{-4}$</td>
</tr>
<tr>
<td>Decision trees</td>
<td>0.415</td>
<td>$3.46 \cdot 10^{-4}$</td>
</tr>
<tr>
<td>Random forest</td>
<td>0.609</td>
<td>$2.31 \cdot 10^{-4}$</td>
</tr>
<tr>
<td>Gradient boosting</td>
<td><strong>0.644</strong></td>
<td><strong>$2.11 \cdot 10^{-4}$</strong></td>
</tr>
</tbody>
</table>

* Only weather in Ljubljana has been used as input
Conclusions & Future Work

Model improvement

- Additional feature engineering weather, nearby weather, different derivatives, land use, anthropogenic features
- Better definition of use cases include also groundwater level as a feature
- Try other methods
  Deep learning, SVM
- Generalization of the models

Other directions

- Explore stream mining approach
  Big Data ready
- Opposite way
  what do the models tell us? drought?
- Implementation of the real-time platform
- Compare with process-based models
  find synergy, improvement
Support slides

Data-driven modeling of groundwater
Gradient Boosting

1. learn the model (usually regression trees)
2. calculate residuals
3. learn the model on the residuals
4. repeat step 2 until residuals are small enough

Introducing non-linear relations!