GECCO 2019 Industrial Challenge: Monitoring of drinking-water quality

Frederik Rehbach, Steffen Moritz, Thomas Bartz-Beielstein [1]

February, 2019

Goal of the GECCO 2019 Industrial Challenge is to develop capable procedures for online monitoring and change detection in water quality data. Precise detection of changes in water quality is a crucial task for public water companies and urgently required for a timely reaction to these changes.

To be suitable for its designated use, methods must be accurate and computationally efficient. This document provides a set of rules and regulations for the GECCO IC, a detailed problem description, as well as contact and submission information.

[1] Cologne University of Applied Sciences, 51643 Gummersbach, Germany
frederik.rehbach@th-koeln.de,
steffen.moritz@th-koeln.de,
thomas.bartz-beielstein@th-koeln.de

1 *Introduction*

Water covers 71% of the Earth's surface and is vital for all known forms of life.
The holistic consideration of water as an important means of nourishment as well as the general protection of lakes and rivers are a central basis for the growth and further development of human civilization. At the same time, civilization itself, with its steady growth, is a menace to the purity of water resources used for drinking water supply and its distribution network. They are highly sensible to any kinds of contaminations. The provision of clean and safe drinking-water is an essential task for water supply companies all over the world.
To deal with this scenario, highly sensible sensors monitor relevant water- and environmental data at several measuring points, on a regular basis. The monitored data can be analyzed to discover any kinds of anomalies. This allows for early recognition of undesirable changes in the drinking water quality and enables the water supply companies to counteract in time.
This year's industrial partner is Thüringer Fernwasserversorgung (TFW)[2], which provides the dataset used in this challenge.

Thüringer Fernwasserversorgung
Mehr als reines Wasser

[2] The Thüringer Fernwasserversorgung, located at the heart of Germany, is a public water company with its headquarters in Erfurt. Thüringer Fernwasserversorgung operates more than 60 dams and reservoirs, 2 central water treatment plants and 550 km of bulk water transport network. With about 200 employees Thüringer Fernwasserversorgung transfers more than 50 million cubic meters of raw water and drinking water to its clients, local and municipal water supply companies, thus ensuring a reliable supply of highest quality drinking water to more than 1 million people.

The Goal of the GECCO 2019 Industrial Challenge is to develop a change detection system to accurately predict any kinds of changes in time series of drinking water composition data. An adequate and accurate alarm system that allows for early recognition of all kinds of changes is a basic requirement for the provision of clean and safe drinking-water.

Although many different methods can be used for time series forecasting, Computational Intelligence (CI) methods, such as Evolutionary Computation and Artificial Neural Networks, offer an attractive option. CI methods have been successfully applied to time series prediction and analysis problems in the past, which makes CI-based systems an interesting alternative to the classical time series analysis methods more widely applied in energy consumption

forecasting, and motivated this competition.[3]

[3] J.D. Hamilton. Time Series Analysis. Princeton University Press, 1 edition, January 1994. ISBN 0691042896

HIGHLIGHTS of the GECCO IC include:

- Interesting Problem Domain: Change detection based on drinking water data offers a challenging test case for modern time series prediction methods.
- Real-world Data: Real drinking-water time series are provided for training, testing, and assessing event- and change detection methods.
- Fair Submission Assessment: Prediction accuracy is determined on test data available to the organizers only, which will be made public after the competition ends.
- Direct Link to Industry: The Thüringer Fernwasserversorgung will evaluate the winning submissions for an implementation in real-world applications. Moreover, a direct contact with the winning participants, who will keep all rights to their detection system, is highly appreciated by Thüringer Fernwasserversorgung.

THE REMAINDER of this document specifies the information needed to take part in this competition. It is organized in three parts: Section 2 introduces the problem of water monitoring and analysis, as well as the water quality data set provided. Section 3 presents the set of rules and regulations. Finally, Section 4 gives information on how to participate in the industrial challenge.

2 *Problem Description*

The objective of this competition is to develop an online monitoring tool to detect changes in water quality. The data provided for this competition consists of one time series of water quality data. Participants of the challenge should implement a system that accurately detects any kinds of changes in the water quality, based on the training data that is supplied, and that meets the requirements specified hereafter. This years data is from sensors that were for testing purposes not continuously maintained - making event detection even harder because of outliers and sensor drift.

2.1 Data collection for water quality monitoring

For the monitoring of the water quality, the Thüringer Fernwasserversorgung performs measurements at significant points throughout the whole water distribution system, in particular at the outflow of the waterworks and the in- and outflow of the water towers. For this purpose, a part of the water is bypassed through a sensor system where the most important water quality indicators are measured. The data that is supplied for this challenge has been measured at different stations near the outflow of a waterworks.

2.2 *Training- and Test-Datasets*

Column name	Description
`Time`	Time of measurement, given in following format: yyyy-mm-dd HH:MM:SS
`Tp`	The temperature of the water, given in °C.
`pH`	pH value of the water
`Cond`	Electric conductivity of the water, given in S/m
`Turb`	Turbidity of the water, given in FNU
`SAC`	Spectral absorption coefficient, given in 1/m
`PFM`	Pulse-Frequency-Modulation, given in Hz
`EVENT`	Marker if this entry should be considered as a remarkable change resp. event, given in boolean.

Table 1: Description of the given time series data

The data for the GECCO IC contains one time series denoting water quality data and operative data on a minutely basis. Given is the the pH value, the electric conductivity, the turbidity and the spectral absorption coefficient of the water. These values are the water quality indicators, any changes here are considered as events. The PFM value of the sensor panel and the temperature of the water are considered as operational data, changes in these values may indicate changes in the related quality values but are not considered as events themselves. Table 1 gives an overview of the data provided.

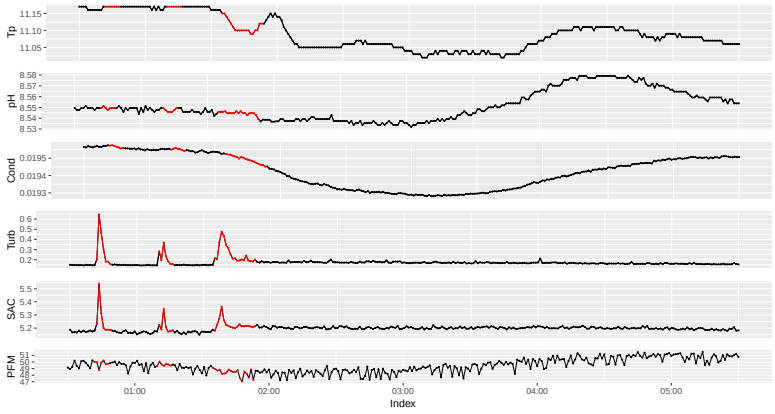

Figure 1: The plot shows an of the given time series data with several original events marked.

The original data only contained a limited amount of events. For this challenge additional events have been imputed into the data to simulate the difficulties that analysis methods have to cope with in case of an event.

The simulated events resemble the original water quality data

with different levels of variation during specific time periods. The training data includes such kind of events and Figure 2 is one such example. [4]

All data constellations considered as remarkable changes in the water quality resp. events are marked in the column EVENT. These changes have to be detected by an online method as accurately as possible. It should be taken into consideration that single outliers and baseline changes in the data are not considered as events.

A first impression of these time series is given by Figure 1.

[4] Sean A. McKenna, David B. Hart, Regan Murray, and Terra Haxton. Handbook of Water and Wastewater Systems Protection, chapter Testing and Evaluation of Water Quality Event Detection Algorithms, pages 369–396. Springer New York, New York, NY, 2011. DOI: 10.1007/978-1-4614-0189-6_19. URL http://dx.doi.org/10.1007/978-1-4614-0189-6_19

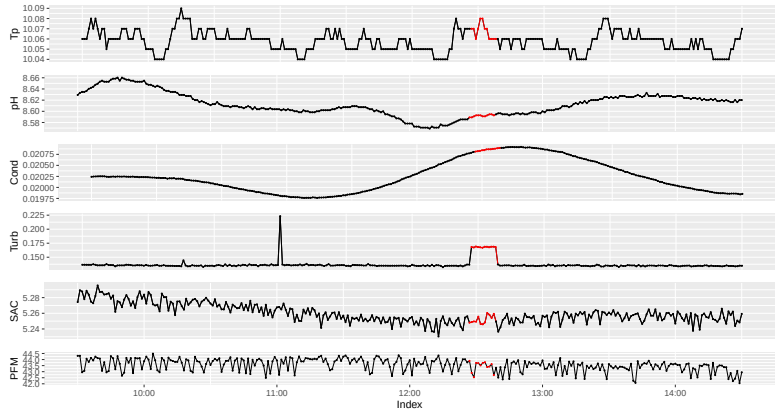

Figure 2: The plot shows an extract of the given time series data with an artificial event marked.

2.3 *Competition assignment*

To participate in the competition an online event detector has to be implemented in R. The detector method has to be uploaded through the automatic challenge evaluation tool. This process is later explained in more detail.

An example code outline for a submission is shown in listing 1. It only consists of a single function which MUST be named *detect*.

Listing 1: Example detector code

```
detect ← function(singleRowOfTheDataSet){
  ## Load your pre-trained model (if you use one):
  ## The model has to be saved in a file with this
  ## EXACT name!
  load("model.Rdata")

  ## random guess with 50% probability
```

```
  probability ← runif(1)
  event ← probability > 0.5

  ## return prediction
  return(event)
}
```

The function *detect* specifies the actual detector function. This function will be called with one single row of the data at each call and has to return a boolean indication if an event is occurring at this single point of time. The example detector shown in Listing 1 performs a random guess with a 50% probability for an event without further considering the given data.

A FRAMEWORK, also containing this dummy detector, is supplied with the software package to allow for appropriate testing of the submission. This framework consists of the training data (*Data > waterDataTraining.RDS*), a dummy detector (*Detectors > DummyEventDetector.R*) and the main evaluation method (*EvaluationMain.R*) as well as one supplementary file (*f1score.R*) which is used for the calculation of the prediction quality.

The main evaluation file will, when executed, automatically source, execute and evaluate all detectors that are deposited in the *Detectors* folder.

ALL SUBMISSIONS have to be tested against this method since the evaluation also will be done with this framework. To do so the detector files only have to be added to the *Detectors* folder and the *EvaluationMain.R* has to be run. Please take into consideration, that for your uploaded model and detector to the challange evaluation tool, the response time of the prediction method is not allowed to exceed a maximum of 30 seconds per prediction. An evaluation which runs for too long is automatically killed and will stop without a result. Also each participant can at maximum run 2 submissions at the same time on the server, not more.

2.4 *Detector Quality Rating*

For this competition the quality of the detector is then calculated using the F1 score, the harmonic mean of precision and recall, which is defined as:

$$F_1 = 2 \cdot \frac{PPV \cdot TPR}{PPV + TPR},$$

with

$$PPV = \frac{\sum TruePositive(TP)}{\sum PredictionPositive} \quad \text{and} \quad TPR = \frac{\sum TruePositive(TP)}{\sum ConditionPositive}$$

so that PPV describes the ration between the number of true positive predictions and the number of all positive predictions while TPR the ratio between the number of true positive predictions and

the number of actual events. Per definition this score ranges from values from 0 to 1, with 1 being the best achievable score.

Submissions are ranked by the F1 score the detector achieves on a part of the data only known to the organizers. This part of the time series will be published after the competition ends. See Section 4 for instructions on how to download reference material, source code, and documentation.

3 *Rules and Regulations*

In order to participate in the competition, an online event detector has to be supplied featuring the interface as specified in Section 2.3. We expect the runtime of the detectors to be reasonable. All required packages have to be accessible by the organizers and be installable from CRAN. Submissions will be ranked using the F1 score that is calculated as defined in Section 2.4. The winner of the GECCO IC will be the participant whose detector achieved the largest F1 score. The last working detector that was uploaded through the submission tool by each participant will be evaluated.

Finalists selected by the organizers will be invited to present their submission at the competition session, held during the GECCO conference. The winner of the competition will be announced at the SIGEVO meeting ceremony, on July 17, 2019.

4 Submission

Submissions will be handled through an automated online evaluation tool. You can access the tool via:

`http://owos.gm.fh-koeln.de:3838/geccoICWebpage/`

The entrance password to the tool is 'idea@gecco'. After entering the page you will have to create an account on the registration page. Please remember your password as we have NO mechanisms for resetting a forgotten password! Once you created an account, you can start uploading your submissions as often as you like. The submissions and scores will be saved on our servers. The only things you have to upload are: your event detector (a file containing the 'detect()' function). And optionally a model that you fitted to the training data that was available in the resource package. The F1-Score of your submission will automatically be evaluated on the server. You can access the results and potential errors in your submission through the 'Refresh Results' button. The F1-Score is evaluated on a validation dataset that is not available for download. The final scores which determine the challenge winners are calculated on a third separate testing dataset. For this evaluation, the last working upload of each participant will be evaluated.

4.1 Organizing Committee

- Frederik Rehbach, TH Köln
- Steffen Moritz, TH Köln
- Thomas Bartz-Beielstein, TH Köln

List of References

J.D. Hamilton. Time Series Analysis. Princeton University Press, 1 edition, January 1994. ISBN 0691042896.

Sean A. McKenna, David B. Hart, Regan Murray, and Terra Haxton. Handbook of Water and Wastewater Systems Protection, chapter Testing and Evaluation of Water Quality Event Detection Algorithms, pages 369–396. Springer New York, New York, NY, 2011. DOI: 10.1007/978-1-4614-0189-6_19. URL `http://dx.doi.org/10.1007/978-1-4614-0189-6_19`.