GECCO 2018 Industrial Challenge: Monitoring of drinking-water quality

Frederik Rehbach, Steffen Moritz, Sowmya Chandrasekaran, Margarita Rebolledo, Martina Friese, Thomas Bartz-Beielstein [1]

February, 2018

Goal of the GECCO 2018 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 *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.

The Goal of the GECCO 2018 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

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

[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.

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.

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.
`Cl`	Amount of chlorine dioxide in the water, given in mg/L (MS1)
`pH`	PH value of the water
`Redox`	Redox potential, given in mV
`Leit`	Electric conductivity of the water, given in μS/cm
`Trueb`	Turbidity of the water, given in NTU
`Cl_2`	Amount of chlorine dioxide in the water, given in mg/L (MS2)
`Fm`	Flow rate at water line 1, given in m^3/h
`Fm_2`	Flow rate at water line 2, given in m^3/h
`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 amount of chlorine dioxide in the water, its pH value, the redox potential, its electric conductivity and the turbidity of the water. These values are the water quality indicators, any changes here are considered as events. The flow rate and the temperature of the water is 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.

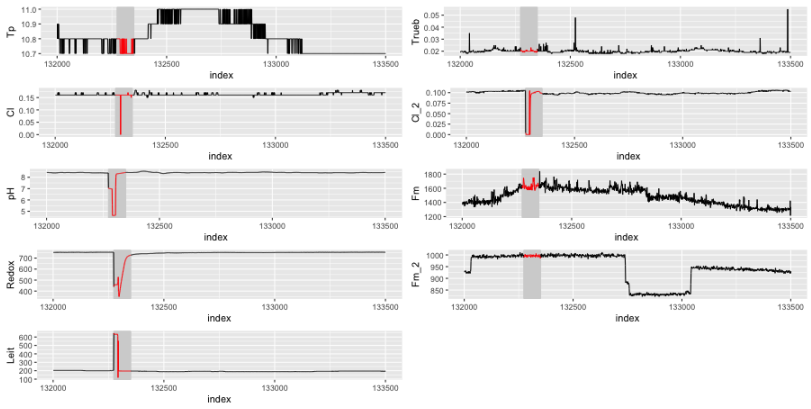

Figure 1: The plot shows an extract of about one days of the given time series data with one original event marked.

The original data only contained events related to regular checks and replacements of sensors. 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 outliers and baseline changes in the data are not considered as events.

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

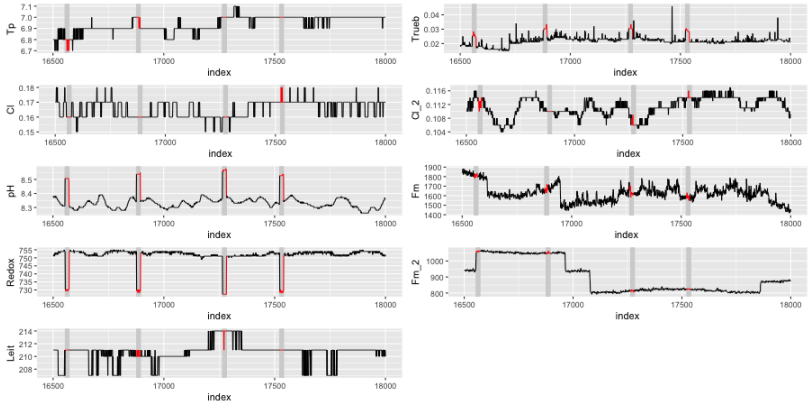

Figure 2: The plot shows an extract of about one days of the given time series data with several artificial events marked.

2.3 *Competition assignment*

To participate in the competition an online event detector has to be implemented in R. The detector method and an accompanying outline descriptor function have to be submitted in a single R-file meeting the specifications as defined hereafter. With the submission competitors have to supply a two page report describing the algorithm and naming the packages (and their versions) used.

An example code outline for a submission is shown in listing 1. It includes three functions *detect*, *destruct* and *getOutline*, the names of these functions must not be changed.

Listing 1: Example detector code

```
detect ← function(dataset){
  ## random guess with 50% probability
  probability ← runif(1)
  event ← probability > 0.5

  ## return prediction
  return(event)
}

destruct ← function(){
  ## remove global variables
  ## delete files
```

[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

```
}

getOutline ← function(){
   # Enter your data here
   competitor.name ← "Max Muster"
   competitor.institution ← "Muster University"

   # Return line must not be changed
   return (list(NAME = competitor.name
         , INSTITUTION = competitor.institution));
}
```

The function *getOutline* is needed for automatic evaluation of all submissions, hence name of the function and the return line must not be changed. Only the name and institution of the competitor has to be given. In case of multiple submissions for one competitor this entry should be made unique (i.e. by an added suffix "Max Muster (a)").

The function *detector* 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. This dummy detector is provided with the data package to allow for immediate testing of the framework provided.

The function *destruct* is called after the prediction has been completed. In this method all global variables or data files that might have been written during the prediction have to be removed from the global environment and or from the hard disk.

A FRAMEWORK, also containing a 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 the response time of the prediction method is not allowed to exceed a maximum of 30 seconds per prediction. Also the number

of submissions per participant is restricted to a maximum of two submissions.

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. Submissions are accompanied by a two page report describing the algorithm implemented and naming packages and its versions used. All packages used have to be accessible by the organizers. 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.

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 19, 2018. Therefore, each participant has to be registered at GECCO 2018.

4 Submission

Submissions to the GECCO IC should consist of:

- An online detection method, supplied in an R-file featuring the interface as specified in Section 2.3
- and a short report describing the algorithm implemented and naming packages and its versions used, also featuring your institution and contact data (two pages maximum).

PLEASE send your submission as archive file (i.e. *.zip) via email to `gecco@f10.fh-koeln.de`. You can also contact the organizers via email (`gecco@f10.fh-koeln.de`) if you have any questions.

4.1 Software and Data

Example data and source code will be available for download at `http://www.spotseven.de/gecco-challenge/gecco-challenge-2018`

4.2 Organizing Committee

- Frederik Rehbach, TH Köln
- Steffen Moritz, TH Köln
- Sowmya Chandrasekaran, TH Köln
- Martina Friese, TH Köln
- Margarita Rebolledo, 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`.