Automatic online monitoring scheme optimization in the presence of daily water quality patterns

M. Gabrielli[1], F. Trovò[2], M. Antonelli[1]
[1] Dipartimento di Ingegneria Civile e Ambientale (DICA), Politecnico di Milano
[2] Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB), Politecnico di Milano

POLITECNICO MILANO 1863

1. Introduction

The quality of most water matrices (including surface, drinking and waste waters) presents daily patterns caused by cyclic environmental conditions and anthropic activities. While the development of online automatic instrumentation has allowed to monitor potentially at any desired instant, the monitoring frequency adopted in monitoring campaigns is often dictated by to budget contraints. However, **the arbitrary selection of both the monitoring frequency and sampling instants does not guarantee to properly identify the daily concentration variability** (maximum and minimum concentrations), **possibly leading to an inaccurate evaluation of the stability of the monitored process** (e.g. water treatment) **or**, worse, **a misleading estimate of the environmental and human health risks**. While some methods have been proposed to optimize water sampling, they either rely on the measurement of an external variable, an option which is not always possible, or do not properly take into account the potential dynamicity of the daily patterns.
For this reasons, **the goal of this work is to develop automatic algorithms to guide monitoring schemes to correctly sample the instants in which the maximum and the minimum pollutants concentrations are present, while, at the same time, reduce the costs linked to high frequency sampling**. Such algorithms will allow to perform more efficient monitoring schemes at lower costs favouring a more throughout environmental control, aligning with the goal of ensuring save water included in the Sustainable Development Goal #6.

2. The developed algorithms

Two algorithms were developed to guide the choice of sampling instants to target the instant which result in the maximum daily concentration variability of any pollutant.

The main components of the algorithms are:
- **Gaussian process** (GP)
Used to exploit the autocorrelation between sampling instants
- **Upper confidence bound** (UCB) criterion
Used to choose the next istant to sample. Such criterion does not rely on the observed concentration means, but on the value of appropriately-selected confidence bounds
- **Seq meta-algorithm**
Used to select multiple sampling instants per day

Figure 1 illustrates the outcome of the combination of the three components of the algorithms discussed above. At time *t* the selected sampling times (indicated by a red cross) are both around 20:00 as the absence of any measurement (shown as blue dots) in that period of the day resulted in a larger uncertainty (displayed by the grey area). The next day, the sampling instants for the maximum concentration is selected around 11:00, in correspondence with the highest measured concentration due to the reduction of the confidence bounds at 20:00.

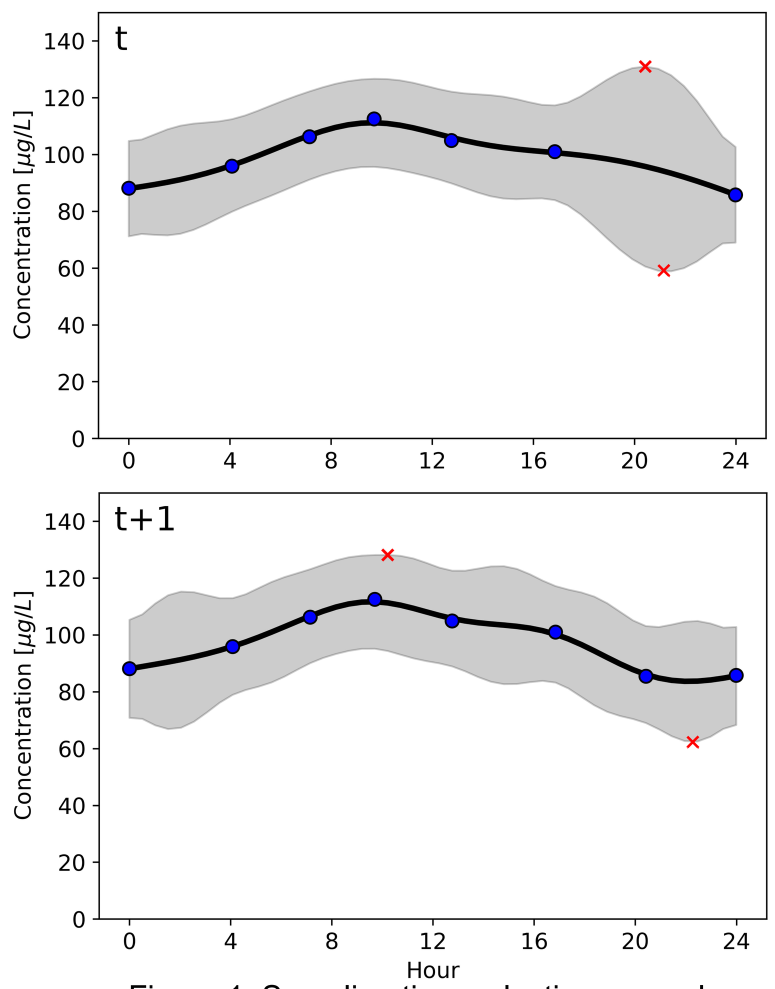
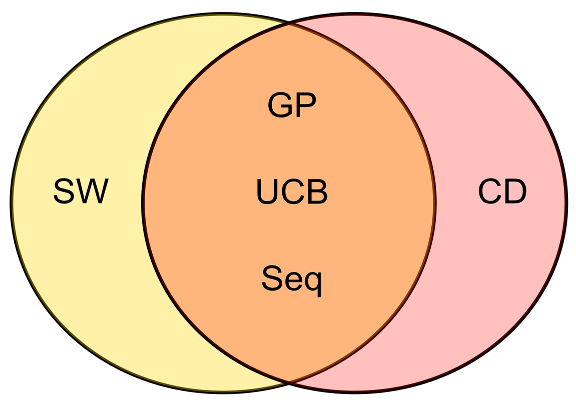


Figure 1. Sampling time selection example

The algorithms share the components mentioned above, but **differ in the way they adapt to the possible patters dynamicity**:

- **Seq(GP-UCB-SW)**
Considers a sliding window on the last *n* samples to select the next sampling instant



- **Seq(GP-UCB-CD)**
Employs a change point model to actively detect changes in the distribution of maximum and minumum values

3. Performance evaluation metrics

The performance of the algorithms was evaluated using two conflicting metrics:
- **Relative difference between the observed and the true maximum daily variation ($RDOT_{delta}$)**
Such metric allows to test the error performed in the identification of the proper sampling instants.

RDOT is calculated as: $$\frac{(max_{obs} - min_{obs}) - (max_{true} - min_{true})}{max_{true} - min_{true}}$$

- **Number of samples per day**
This metric is used as a proxy for the costs due to the samples analyses and the instrumentation maintenance

4. Case study

The developed algorithms were compared to traditional monitoring schemes (i.e. random sampling and fixed-time sampling) **on a synthetic (SynS) and real-world (RW) scenario**. The traditional schemes were tested with different number of samples per day and, in case of fixed-time sampling, evaluating, also, all the the combinations of sampling instants (SIC).
The **synthetic scenario has been derived from a monitoring campaing of intact bacterial cells of a full-scale distribution network (DWDN)**. A plausible shift in the daily pattern was simulated to test the resilience of both algorithms and traditional schemes.
Successively, the **real-world test** was then performed using directly the **measurements of total bacterial concentrations in a DWDN** collected every two hours for about 5 months.

5. Results

Figures 2 and 3 show the performance of the algorithms and traditional schemes in both tests. In SynS the performance is split between before and after the daily pattern change, as indicated by the arrows.

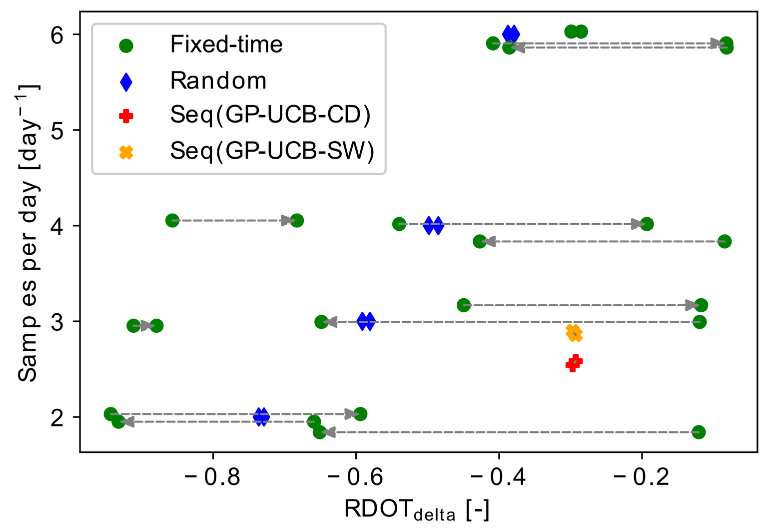


Figure 2. Synthetic scenario results

Regarding the traditional monitoring schemes, **random sampling is not affected by the pattern change, but provides only mediocre $RDOT_{delta}$** in both tests. Differently, fixed-time $RDOT_{delta}$ depends on the chosen SIC, differing greatly between the best, median and worst combinations, as shown in Figure 2. Furthermore, the performance of each SIC varies before and after the change. So even though **fixed-time sampling** might achieve the best $RDOT_{delta}$, **its performances suffer from the need to identify the best SIC and are not resisilient to pattern changes**.

On the contrary, **Seq(GP-UCB-CD) and Seq(GP-UCB-SW) provide performances which are resilient to pattern changes** which, in addition, are also correctly identified by the change detection test in Seq(GP-UCB-CD). Even in the RW dataset, where complex dynamics are present, **both algorithms result in $RDOT_{delta}$ which can be achieved by traditional schemes only at a higher number of samples or only one fixed-time SIC**, highlighting their good applicability in real world monitoring applications.

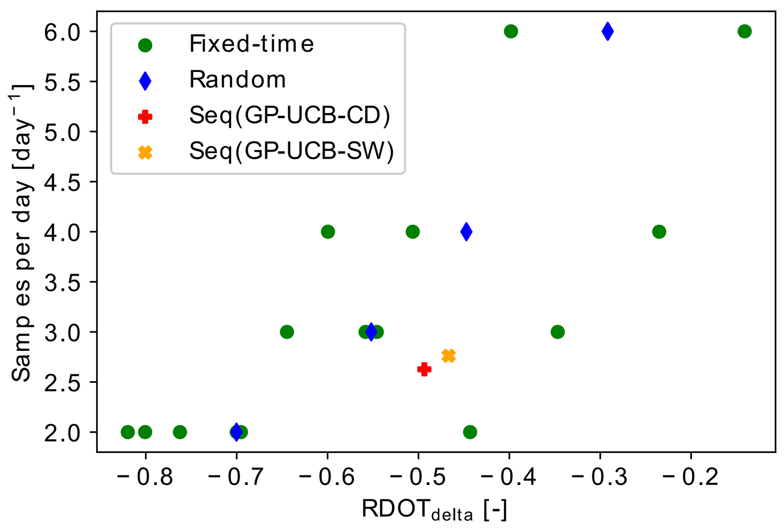


Figure 3. Real-world results

6. Conclusive remarks

This work presents two algorithms which automatically optimize the monitoring scheme performed by online instruments. **The algorithms result in sampling schemes which are resilient to changes in the concentration patterns and provide increased performances compared to standard monitoring schemes, while, coincidentally, reducing the cost of samples analyses**. The use of such algorithms will allow a more widespread water quality monitoring due to the reduced costs and, also, a more accurate estimate of the pollutants concentrations and the linked implications (e.g. human health risk, process stability, etc.).