

Supplementary Information

Algorithm 1 Feature engineering from selected water quality variables and complaint records

Require: Set of complaint records $\mathcal{C} = \{c_1, \dots, c_N\}$, each with date t_i and spatial unit s_i (DMA or pressure zone)

Require: Water quality dataset $\mathcal{Q}$ with measurements of selected variables $\mathcal{V} = \{v_1, \dots, v_K\}$, each record with date, value, and spatial unit

Require: Time window $\Delta t = 30$ days

Ensure: Feature matrix $\mathbf{X}$ and label vector $\mathbf{y}$ for predictive modelling

— **Complaint feature extraction** —

```
1: for each complaint  $c_i \in \mathcal{C}$  do
2:   Define window  $W_i = [t_i - \Delta t, t_i)$ 
3:   for each selected variable  $v_k \in \mathcal{V}$  do
4:     Retrieve measurements  $M_{i,k}$ : query  $\mathcal{Q}$  for  $v_k$  within  $W_i$  at DMA  $s_i$ ; if empty, query by pressure zone  $s_i$ 
5:     if  $|M_{i,k}| > 0$  then
6:       Compute descriptors: mean, std, min, max, slope of  $M_{i,k}$ 
7:     else
8:       Assign missing values for all descriptors of  $v_k$ 
9:     end if
10:  end for
11:  Extract temporal features: month( $t_i$ ), season( $t_i$ ); assign label  $y_i = 1$ 
12: end for
```

— **Non-complaint (control) sample generation** —

```
13:  $n \leftarrow |\mathcal{C}|$  ▷ Balance dataset
14: Initialize empty set  $\mathcal{C}'$ 
15: while  $|\mathcal{C}'| < n$  do
16:   Sample random date  $t'_j$  within the study period
17:   Identify complaints in window  $[t'_j - \Delta t, t'_j)$ 
18:   Determine available DMAs and pressure zones not used in that window
19:   Randomly select spatial unit  $d'_j$  (DMA) or  $p'_j$  (pressure zone, if no DMA available)
20:   Define window  $W'_j = [t'_j - \Delta t, t'_j)$ 
21:   for each selected variable  $v_k \in \mathcal{V}$  do
22:     Retrieve  $M'_{j,k}$ : query  $\mathcal{Q}$  for  $v_k$  within  $W'_j$  at  $d'_j$ ; if empty, use  $p'_j$ 
23:     if  $|M'_{j,k}| > 0$  then
24:       Compute descriptors: mean, std, min, max, slope of  $M'_{j,k}$ 
25:     else
26:       Assign missing values for all descriptors of  $v_k$ 
27:     end if
28:   end for
29:   Extract temporal features: month( $t'_j$ ), season( $t'_j$ ); assign label  $y'_j = 0$ 
30:   Add instance to  $\mathcal{C}'$ 
31: end while
```

— **Dataset assembly** —

```
32: Concatenate complaint and non-complaint feature rows into  $\mathbf{X}$ ; concatenate labels into  $\mathbf{y}$ 
33: return  $\mathbf{X}, \mathbf{y}$ 
```
