Computationally Efficient Ranking of Groundwater Monitoring Locations

Peter Radvanyi[1], Claire Miller[1], Craig Alexander[1], Marnie Low[1], Wayne R. Jones[2], Luc Rock[3]

[1] University of Glasgow, School of Mathematics and Statistics, United Kingdom
[2] Shell Research Ltd, United Kingdom
[3] Shell Global Solutions Canada Inc, Canada

E-mail for correspondence: Peter.Radvanyi@glasgow.ac.uk

Abstract: Sampling groundwater quality monitoring wells is a costly and time intensive process that incurs health and safety risks. Reducing the number of wells whilst minimising information loss can greatly increase the sustainability of long-term monitoring. Wells that provide redundant information can be identified by assessing their observations' influence on statistical model estimates. Well-based cross-validation (WBCV) could be used to obtain such a measure of influence for each well, however, the associated computational cost renders this option unfavourable. In this paper, we propose a method based on influence statistics of regression-based, groundwater solute concentration models, as a computationally efficient, approximate alternative. The method, named well influence analysis (WIA), approximated WBCV results in a simulation study and real groundwater contaminant observations with an average 77% and 73% accuracy respectively. WIA will be implemented in the "well redundancy analysis" feature of GWSDAT, an open-source software for the spatiotemporal modelling of groundwater monitoring observations.

Keywords: Groundwater Monitoring; Groundwater Contamination; Statistical Modelling; Spatiotemporal; Influence Statistics.

1 Introduction

The aim of groundwater quality monitoring during the remediation of contaminated sites is to understand the behaviour of the solutes of concern by observing changes in their concentration levels at fixed sampling locations called wells. Spatiotemporal statistical models can be used to estimate contaminant concentrations over spatial domains of interest using these

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

observations. However, collecting and analysing samples from groundwater monitoring wells is costly, time intensive and incurs health and safety risks. Reducing sampling intensity whilst minimising the loss of information can greatly increase the efficiency and sustainability of long-term groundwater quality monitoring. Sampling intensity can be decreased by reducing the number of sampling locations. In many cases, fewer wells can be sufficient for supporting robust statistical models, provided they adequately capture the spatiotemporal heterogeneity in solute concentrations. Therefore, the choice of monitoring wells to omit from sampling is crucial, and should be based on qualitative and quantitative analyses. A possible quantitative approach using statistical models is assessing sampling wells based on their observations' impact on solute concentration estimates. Wells whose data provide redundant information to the model, could be considered for omission from future sampling campaigns. Feedback from users of the open-source, spatiotemporal groundwater quality modelling software, GWSDAT (Jones et al. 2014), highlighted the need for a tool to facilitate this well redundancy analysis. Ranking wells by influence prior to testing the impact of omitting one, aims to reduce the need for a trial-and-error approach. Assessing well influence can be done iteratively, using well-based cross-validation (WBCV). However, the computational cost associated with re-fitting the model in each iteration makes this approach unfavorable. In this work, we aim to show that for regression-based groundwater contamination models, well influence analysis (WIA) could be a computationally efficient, approximate alternative to the cross-validation-based method. WIA provides a suggested sequence for omitting wells, by ranking them using influence statistics commonly used in regression analysis. The proposed method was tested in a simulation study and on real groundwater monitoring data.

2 Simulation Study

A simulation study (Radvanyi, 2023) was designed to analyse how closely WIA approximated the cross-validation-based well influence rankings in different scenarios, and to compare different influence statistics that could be used for WIA. The simulation study was conducted using synthetic data sets.

2.1 Synthetic Data

The synthetic data (McLean et al. 2019) contained coordinates, sampling times and solute concentrations for three hypothetical contaminant plumes of increasing complexity simulated using process based models (Figure 1). 15 % multiplicative random noise was applied to the data to mimic the measurement errors of real groundwater observations. Samples were then drawn at select times and coordinates to mimic sampling from monitoring

wells. Nine monitoring network designs were created for each plume using 6, 12 and 24 wells with 3 well placement strategies. These strategies were random, grid and expert, the latter implying knowledge of plume characteristics, such as origin and groundwater flow direction. Each scenario ran for 100 iterations.

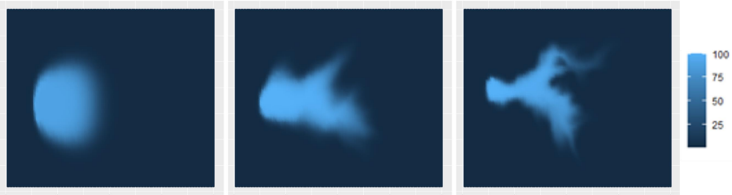

FIGURE 1. Hypothetical plumes: simple (l), moderate (c) and complex (r).

2.2 Modelling Approach

Concentration estimates over the full spatial domain were obtained using P-splines models, also used in GWSDAT (Evers et al. 2015). P-splines (Eilers & Marx, 1996) are regression splines fitted by least-squares with a roughness penalty. The P-splines model can be written as

$$y_i = \sum_{j=1}^{m} b_j(x_i)\alpha_j + \epsilon_i,$$

where y_i, $i = 1, 2, ...n$, are the natural logarithm of the solute concentrations, x_i are the corresponding coordinates and sampling times, b_j, $j = 1, 2, ...m$, are B-spline basis functions, α_j are the basis coefficients and ϵ_i are errors, assumed to be independent with $N(0, \sigma^2)$.

2.3 Well-Based Cross-Validation

The baseline ranking of well influence on estimated solute concentrations was computed via well-based cross-validation (WBCV; Evers et al. 2015). WBCV is a form of leave-one-out cross-validation, where each well (and hence associated observations) was removed sequentially and used as the test set for a model trained on the remaining data. The well ranking was given by the numerical order of corresponding root-mean-square errors (RMSE) calculated by:

$$RMSE_k = \sqrt{\frac{\sum_{l=1}^{n_k} (y_{kl} - \hat{y}_{kl})^2}{n_k}},$$

where $k = 1, .., w$ and w is the total number of wells, y_{kl} is the l-th observation from the k-th well, $\hat{y}_{kl}$ is the l-th fitted value and n_k is the number of observations.

2.4 Well Influence Analysis

Different influence metrics were compared for approximating the WBCV rankings. Cook's distance (CD; Cook, 1977), which is a measure of the sum of changes in regression estimates if an observation is deleted, produced the most informative results. It can be expressed using leverages, which are the diagonal elements of the projection matrix from the P-splines model:

$$CD_i = \frac{1}{p}(r_i^s)^2 \frac{h_{ii}}{1 - h_{ii}},$$

where p is the effective degrees of freedom, r_i^s is the standardised residual and h_{ii} is the leverage of the i-th observation. The rankings were given by the numerical orders of the median CD values for each well. The studied influence metrics were originally derived for linear regression. Their application in this case is supported by the fact that P-splines are analogous to linear regression.

2.5 Assessing Performance

The performance of WIA was quantified by calculating the normalised difference score D_n, which indicated the total difference in well ranks between WIA and WBCV. D_n is bounded by $0 \leq D_n \leq 1$ with 0 meaning the rankings were equivalent. D_n was calculated by

$$D_n = \frac{\sum_{i=1}^{w} |o_i^{wbcv} - o_i^{ia}|}{D_{max}},$$

where o_k^{wbcv} is the rank of the k-th well based on WBCV and o_k^{ia} is its rank based on WIA. The maximum difference between the two rankings, D_{max} is a function of the number of monitoring wells such that $D_{max} = \frac{w^2}{2}$.

2.6 Results

The mean D_n for CD-based WIA was 0.23, which means that on average, it approximated the baseline (WBCV) rankings with 77% accuracy. Figure 2 shows the results in the form of a boxplot categorised by scenario design features. Mean D_n values increased with plume complexity from 0.20 to 0.27. The complex plume is also associated with the highest variance. The monitoring well network design also seemed to play a role in the outcome of the analysis. The results show that WIA has better performance if well placement is done based on site characteristics as opposed to randomly or in a grid pattern. In terms of the number of monitoring wells, the smallest mean D_n results were obtained with 6 wells. However, this is most likely an artifact related to fewer possible differences in well ranks between WIA and the baseline. This effect also seems to disappear given a sufficient number of wells, since there is little difference in results between 12 and 24 wells.

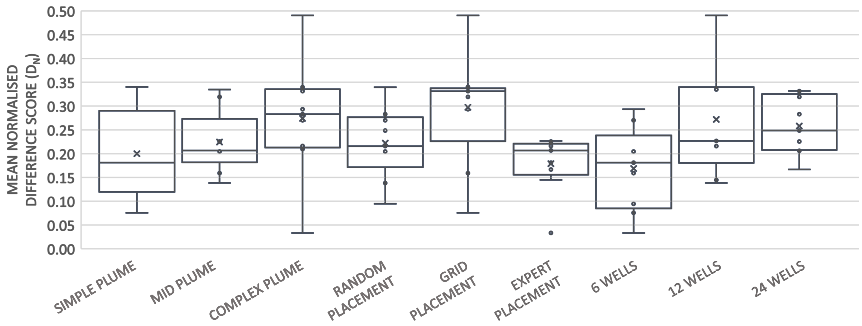

FIGURE 2. Breakdown of mean normalised difference scores (D_n; $0 \leq D_n \leq 1$) by design attributes plume complexity, well placement and the number of wells. A smaller D_n indicates a more accurate estimation of WBCV rankings by WIA.

3 Real Data Application

The comparison of WIA and WBCV was also performed on real groundwater contamination data from an undisclosed monitoring site. The data set contained the concentrations of five contaminants in groundwater samples from 32 monitoring wells collected over a 4 year period. The contaminants were modelled separately. Table 1 shows the results of the analysis by contaminant.

TABLE 1. Breakdown of normalised difference scores (D_n; $0 \leq D_n \leq 1$) by contaminant from the groundwater monitoring data. A smaller value indicates a more accurate estimation of WBCV ranking by WIA.

Contaminant	$\mathbf{D_n}$
Ethylbenzene	0.36
Toluene	0.25
Nitrate	0.18
Sulphate	0.23
TPH	0.32
Mean	**0.27**

The mean D_n was 0.27, which translated to an average of 73% accuracy in comparison to WBCV. Just as in the simulation study, most of the deviation in the WIA ranking compared to WBCV was due to an aggregation of minor rank differences. This means that wells generally occupied similar positions in both rankings.

4 Conclusions

In conclusion, empirical evidence was presented to support the application of influence statistics in the proposed context. WIA estimated WBCV results with an average 77% and 73% accuracy in the simulation study and real data examples respectively. These results were positive given the aim and the approximate nature of the analysis. The simulation study also showed that the monitoring network design and contaminant plume characteristics also affect the accuracy of WIA. WBCV would be the preferred ranking method, but it is computationally unfavorable because it requires fitting w models for each well that is considered for omission from future sampling campaigns. In contrast, WIA only requires a single model before each omission, which makes it a more efficient alternative to WBCV for ranking wells by influence on solute concentration estimates. In other words, there is trade-off between accuracy and computational efficiency, but the results indicate that in this case, the gain in efficiency is greater than the loss in accuracy. WIA is easy to implement in software built around regression-based groundwater quality models, such as GWSDAT, and it can help determine the sequence in which wells should be omitted during well redundancy analysis.

References

Cook, R.D. (1977). Detection of Influential Observations in Linear Regression. *Technometrics*, **19**, 15 – 18.

Eilers, P.H.C., Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, **11.2**, 89 – 121

Evers, L., Molinari, D.A., Bowman, A.W., Jones, W.R., Spence, M.J. (2015). Efficient and automatic methods for flexible regression on spatiotemporal data, with applications to groundwater monitoring, *Environmetrics*, **26.6**, 431 – 441.

Jones, W.R., Spence, M.J., Bowman, A.W., Evers, L., Molinari, D.A. (2014). A software tool for the spatiotemporal analysis and reporting of groundwater monitoring data. *Environmental Modelling and Software*, **55**, 242 – 249.

McLean, M.I., Evers, L., Bonte, M., Bowman, A.W., Jones, W.R. (2019). Statistical modelling of groundwater contamination monitoring data: A comparison of spatial and spatiotemporal methods. *Science of the Total Environment*, **652**, 1339 – 1346.

Radvanyi, P. (2023). Well Influence Analysis. https://github.com/peterradv/Well-Influence-Analysis