Influence Analysis Simulation Study

Peter Radvanyi

November 16, 2022

1 Objectives

The primary objective is to compare the performance of different influence analysis metrics against well-based cross-validation for ordering wells based on their influence on model predictions.

The secondary objective is to identify factors that could affect the performance of influence analysis metrics, such as plume complexity, number of monitoring wells, arrangement of monitoring wells and the assumption of measurement error type (additive or multiplicative). We would also like to examine the estimated influence of wells on the boundaries of the monitoring site.

2 Simulated Contaminant Plume Data

There are three hypothetical plumes: simple, mid and complex. Simulated 'true' data from the PDE corresponding to the selected plume is loaded. This simulated data includes coordinates, time and concentration measurements of 22500 locations covering a large domain of 100x35 units. There are 20 measurement times for each location to cover a long enough period, making the number of observations $22500 * 20 = 450000$. The concentration measurements range from 0 to 100 with means of 5.70, 6.60 and 4.53 for the simple, mid and complex plumes respectively. This indicates a heavy right-skew in the response.

3 Network designs

A network design is made up of two factors: the number of wells and the well placement strategy. The number of wells is either 6, 12 or 24 to mimic similar GWSDAT example designs, while the placement strategy is either random, grid or expert. The loaded network design data includes the well IDs, which are the numbers of the wells as they appear in the list, and their coordinates.

4 Creating Well Data

The simulated 'true' plume data loaded in section 2 is filtered for the coordinates of the monitoring wells which are loaded in section 3. This can be expressed by the following notation:

$$W \subset P, \tag{1}$$

where W is the well data and P is the simulated 'true' data. W is the subset of P where the coordinates match the coordinates of the monitoring wells.

5 Adding Random Noise to Well Data

Random noise is added to the simulated well data to represent measurement errors. The added noise is either additive or multiplicative. In groundwater quality monitoring it is commonly assumed that the observation data has multiplicative measurement errors but in this case both scenarios were tested to assess the difference in outcome. Additive noise was added using the following equation:

$$y_i = z_i + \epsilon_i, \tag{2}$$

where y_i is the i-th observation with added measurement noise, z_i is the i-th well data point without measurement noise (simulated 'true' data), ϵ_i is a random variable drawn from $N(\mu, \sigma)$, a normal distribution with mean, $\mu = 0$ and standard deviation, $\sigma = 0.1$ to represent 10% measurement error.

Multiplicative noise was added using the following equation:

$$y_i = z_i * \epsilon_i, \tag{3}$$

where y_i is the i-th observation with added measurement noise, z_i is the i-th well data point without measurement noise (simulated 'true' data), ϵ_i is a random variable drawn from $N(\mu, \sigma)$, a normal distribution with mean, $\mu = 1$, because the mean of the noisy data should be equal to the mean of the original data, and standard deviation, $\sigma = 0.1$ to represent 10% measurement error.

6 Log Transforming Observation Data

The distribution of observations is right-skewed, because there are many low concentration measurements and relatively few high concentration ones. Therefore, the observation data are log transformed by:

$$log(1 + y_i), \tag{4}$$

where, y_i, $i = 1, 2..., n$ are the observations. 1 is added to each observation before the log transformation to account for values close to zero as they could potentially introduce negative values when additive noise is applied. Moreover,

log-transforming the response variable which has multiplicative noise prior to modelling allows for an additive interpretation of the error.

7 Well-Based Cross-Validation

Well-based cross-validation is used to determine the true well influence order.

7.1 Aim

We would like to measure how much influence a particular well has on the model fit. We can measure this by assessing how the model fit changes if the observations of the well are removed from the underlying data prior to model fitting. Thus, we use a special case of cross-validation [6] where the number of folds equals the number of monitoring wells in the data set. In each step, all observations from a single well are omitted, and the remaining data are used as the training set, while the omitted observations are used as the test set. The well influence order is then given by the prediction errors corresponding to the omitted wells.

Thus, the influence of well k ($k = 1, 2, ..., w$, where w is the number of monitoring wells) is estimated by the prediction error calculated at the coordinates of k, using a model that is fitted to a subset of the observation data that does not include any observations from k.

7.2 Code Step-by-Step

A list is created with length equal to the number of wells. Each item in the list is a duplicate of the observation data.

From each item in the list, all observations of the well whose ID corresponds to the number of the item in the list are removed. The removed observations are stored in a separate list.

In each iteration, a P-splines model [4] is fitted to the observation data that excludes observations from well k. The resulting model objects are stored in a separate list. The model takes the following form:

$$y_{(-k)i} = \sum_{j=1}^{m} b_j(x_{(-k)i})\alpha_j + \epsilon_i, \tag{5}$$

where $i = 1, 2, ...n$, $k = 1, 2, ..., w$ and $j = 1, 2, ...m$. $y_{(-k)i}$ are the contaminant concentrations excluding the response from well k, b_j are the p-spline basis functions (either quadratic, which is the gwsdat default or cubic), $x_{(-k)i}$ are the corresponding explanatory variables (spatial coordinates and time of measurement), α_j are the basis coefficients and ϵ_i are the measurement errors, assumed to be independent and normally distributed $N(\mu, \sigma^2)$.

Predictions for the coordinates of the corresponding deleted well from each model are calculated and stored in a list.

Prediction error is calculated for each model using the above predictions and observations from the removed wells using the following equation:

$$RMSE_k = \sqrt{\frac{\sum_{i=1}^{n_k} (y_{ki} - \hat{y}_{ki})^2}{n_k}}, \tag{6}$$

where $RMSE_k$ is the root mean squared prediction error for the k-th well, calculated using the model which was fitted to a data set that excluded the observations of the k-th well, y_{ki} is the i-th observation from the k-th well, $\hat{y}_{ki}$ is the i-th fitted value for the k-th well and n_k is the number of observations from the k-th well.

The well influence order is given by the resulting prediction error values. Well influence increases with increasing prediction error since the removed well would have had a significant effect on the model fit at that location.

8 Computing Influence Analysis Metrics

Different influence analysis metrics are used to estimate the well influence order using information from a model fitted to the complete observation data set.

8.1 Aim

We would like to calculate influence analysis metric values for each observation in the data and then average these values across the monitoring wells the observations came from. Thus the wells will be ordered by the average influence analysis metric values of their observations. A higher average value means higher influence and consequently a higher placement.

The analysed influence analysis metrics are leverages, standardised residuals, Cook's distance, DFFITS, Hadi's influence measure and COVRATIO.

8.2 Code Step-by-Step

A P-splines model is fitted to the complete observation data using equation 5 with y_i and x_i.

The hat matrix is computed from the fitted model. The fitted values, $\hat{y}$, are given by:

$$\hat{y} = Hy, \tag{7}$$

where the hat matrix H is:

$$H = B(B^T B + \lambda D_d^T D_d)^{-1} B^T \tag{8}$$

where B is the matrix of B-spline basis functions, λ is a non-negative smoothing parameter and D_d is a matrix that computes the successive d-th order differences across the sequence of α-s in each of the 3 covariate dimensions. Each row in the hat matrix corresponds to one of the observations in our data.

8.2.1 Leverages

The leverages are the diagonal elements of the hat matrix:

$$h_{11}, h_{22}, ...h_{nn} \tag{9}$$

They are saved in a new data frame in which the rows correspond to the observations and the columns to the influence analysis metrics, first of which is the leverage. All influence analysis metric values will be saved in this data frame.

8.2.2 Standardised Residuals

The standardised residuals for the observations are computed using the following equation:

$$r_i^s = \frac{r_i}{\sqrt{\frac{\sum_{i=1}^n r_i^2}{n-p}}\sqrt{1-h_{ii}}}, \tag{10}$$

where r_i^s represents the internally studentised (or standardised) residual [7] of the i-th observation, r_i is the residual, n is the number of observations, p is the effective degrees of freedom which is given by the trace of the hat matrix and h_{ii} is the leverage. The standardised residuals are added to the results data frame.

8.2.3 Cook's Distance

Cook's distance [2] is calculated using the following equation:

$$CD = \frac{1}{p}(r_i^s)^2 \frac{h_{ii}}{1-h_{ii}}, \tag{11}$$

where p is the effective degrees of freedom which is given by the trace of the hat matrix, r_i^s is the internally studentised (standardised) residual of the i-th observation and h_{ii} is the leverage of the fitted value as given by the i-th row and i-th column of the hat matrix H. Cook's distance values corresponding to the observations are added to the results data frame.

8.2.4 DFFITS

DFFITS [1] values are calculated using the following equation:

$$DFFITS = r_i^e \sqrt{\frac{h_{ii}}{1 - h_{ii}}}, \tag{12}$$

where h_{ii} is the leverage of the i-th observation, and r_i^e is the externally studentized residual [7], which is calculated using the following formula:

$$r_i^e = \frac{r_i}{\sqrt{\frac{\sum_{\substack{j=1 \\ j \neq i}}^{n} r_j^2}{n-p-1}} \sqrt{1 - h_{ii}}}, \tag{13}$$

where r_i represents the i-th residual and the first term in the denominator is an estimate of the standard deviation, σ of the i-th residual based on all but the i-th residual. DFFITS values corresponding to the observations are subsequently added to the results data frame.

8.2.5 Hadi's Influence Measure

Hadi's influence measure [3] is calculated using the following equation:

$$H_i^2 = \frac{p a_i^2}{(1 - (1 - h_{ii}) a_i^2} + \frac{h_{ii}}{1 - h_{ii}}, \tag{14}$$

where, p is the effective degrees of freedom which is given by the trace of the hat matrix, h_{ii} is the leverage of the i-th observation and a_i is i-th adjusted residual calculated by:

$$a_i = \frac{r_i}{\sqrt{q - h_{ii}}}. \tag{15}$$

Hadi's influence measure values are added to the results data frame.

8.2.6 COVRATIO

COVRATIO [5] is calcualted using the following equation:

$$COVRATIO = \frac{1}{[\frac{n-p-1}{n-p} + \frac{r_i^{e\,2}}{n-p}]^p (1 - h_{ii})}, \tag{16}$$

where n is the number of observations, p is the effective degrees of freedom, r_i^e is the externally studentized residual of the i-th observation and h_{ii} is the leverage of the i-th observation. The COVRATIO values are added to the results data frame.

8.2.7 Creating Well Influence Order Based on Influence Analysis Metrics

The influence analysis metric values (calculated for each observation) are grouped by well ID. Then, an average value for each well can be calculated. The chosen averaging functions are described below:

- leverages - the median value is calculated
- standardised residuals - the median absolute deviation (MAD) is calculated
- Cook's distance - the median value is calculated
- DFFITS - the median value is calculated
- Hadi's influence measure - the median value is calculated
- COVRATIO - the median value is calculated

The median and median absolute deviation were chosen, because they are robust measures when the data is non-normal, which might be the case with groundwater monitoring data even after the log-transformation. The wells are subsequently arranged by their average influence analysis metric values in a decreasing manner, resulting in one ordered list per influence analysis metric type. These orders can then be compared to the order created by the well-based cross-validation method.

9 Comparing Well Influence Orders

The influence analysis-based well influence orders are compared by calculating a difference score D, that measures the total number of well placement differences between an influence analysis-based order and the well-based cross-validation-based order. This method is shown in the following equation:

$$D = \sum_{i=1}^{w} |o_i^{wbcv} - o_i^{ia}|, \tag{17}$$

where w is the number of monitoring wells, o_i^{wbcv} is the position of the i-th well in the well-based cross-validation influence order and o_i^{ia} is the position of the same well in the influence analysis metric-based influence order.

The number of wells, w, can be different in scenarios, therefore D needs to be standardised to allow for scenario-independent comparisons. This is accomplished by dividing D by the maximum of D, which is a function of the number of wells. Then,

$$D_s = \frac{D}{D_{max}}, \tag{18}$$

is the fraction of misplaced wells with a value of 0 meaning the two well influence orders are equivalent and a value of 1 meaning the two orders are complete opposites.

References

[1] David A Belsley, Edwin Kuh, and Roy E Welsch. *Regression diagnostics: Identifying influential data and sources of collinearity*. John Wiley & Sons, 2005.

[2] R. Dennis Cook. "Detection of Influential Observation in Linear Regression". In: *Technometrics* 19.1 (1977), pp. 15–18. ISSN: 00401706. URL: http://www.jstor.org/stable/1268249 (visited on 11/11/2022).

[3] Ali S. Hadi. "A new measure of overall potential influence in linear regression". In: *Computational Statistics & Data Analysis* 14.1 (1992), pp. 1–27. ISSN: 0167-9473. DOI: https://doi.org/10.1016/0167-9473(92)90078-T. URL: https://www.sciencedirect.com/science/article/pii/016794739290078T.

[4] Wayne R. Jones et al. "A software tool for the spatiotemporal analysis and reporting of groundwater monitoring data". In: *Environmental Modelling & Software* 55 (2014), pp. 242–249. ISSN: 1364-8152. DOI: https://doi.org/10.1016/j.envsoft.2014.01.020. URL: https://www.sciencedirect.com/science/article/pii/S1364815214000309.

[5] Marcello Merli. "Outlier recognition in crystal-structure least-squares modelling by diagnostic techniques based on leverage analysis". In: *Acta Crystallographica Section A* 61.4 (July 2005), pp. 471–477. DOI: 10.1107/S010876730501809X. URL: https://doi.org/10.1107/S010876730501809X.

[6] Payam Refaeilzadeh, Lei Tang, and Huan Liu. "Cross-Validation". In: *Encyclopedia of Database Systems*. Ed. by LING LIU and M. TAMER ÖZSU. Boston, MA: Springer US, 2009, pp. 532–538. ISBN: 978-0-387-39940-9. DOI: 10.1007/978-0-387-39940-9_565. URL: https://doi.org/10.1007/978-0-387-39940-9_565.

[7] Sanford Weisberg and J Fox. *An R Companion to Applied Regression*. English. 2nd ed. Thousand Oaks: Sage, 2011.