Journal of Hydrology 531 (2015) 902–911

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

A statistical learning framework for groundwater nitrate models of the Central Valley, California, USA

Bernard T. Nolan [a,*], Michael N. Fienen [b], David L. Lorenz [c]

[a] U.S. Geological Survey, National Center, 12201 Sunrise Valley Drive, Reston, VA 20192, USA
[b] U.S. Geological Survey, Wisconsin Water Science Center, 8505 Research Way, Middleton, WI 53562, USA
[c] U.S. Geological Survey, Minnesota Water Science Center, 2280 Woodale Drive, Mounds View, MN 55112, USA

ARTICLE INFO

Article history:
Received 29 May 2015
Received in revised form 7 October 2015
Accepted 9 October 2015
Available online 26 October 2015
This manuscript was handled by Geoff Syme, Editor-in-Chief

Keywords:
Groundwater
Nitrate
Boosted regression trees
Artificial neural networks
Bayesian networks
Cross validation

SUMMARY

We used a statistical learning framework to evaluate the ability of three machine-learning methods to predict nitrate concentration in shallow groundwater of the Central Valley, California: boosted regression trees (BRT), artificial neural networks (ANN), and Bayesian networks (BN). Machine learning methods can learn complex patterns in the data but because of overfitting may not generalize well to new data. The statistical learning framework involves cross-validation (CV) training and testing data and a separate hold-out data set for model evaluation, with the goal of optimizing predictive performance by controlling for model overfit. The order of prediction performance according to both CV testing R^2 and that for the hold-out data set was BRT > BN > ANN. For each method we identified two models based on CV testing results: that with maximum testing R^2 and a version with R^2 within one standard error of the maximum (the 1SE model). The former yielded CV training R^2 values of 0.94–1.0. Cross-validation testing R^2 values indicate predictive performance, and these were 0.22–0.39 for the maximum R^2 models and 0.19–0.36 for the 1SE models. Evaluation with hold-out data suggested that the 1SE BRT and ANN models predicted better for an independent data set compared with the maximum R^2 versions, which is relevant to extrapolation by mapping. Scatterplots of predicted vs. observed hold-out data obtained for final models helped identify prediction bias, which was fairly pronounced for ANN and BN. Lastly, the models were compared with multiple linear regression (MLR) and a previous random forest regression (RFR) model. Whereas BRT results were comparable to RFR, MLR had low hold-out R^2 (0.07) and explained less than half the variation in the training data. Spatial patterns of predictions by the final, 1SE BRT model agreed reasonably well with previously observed patterns of nitrate occurrence in groundwater of the Central Valley.

1. Introduction

We evaluated three off-the-shelf machine learning methods for their ability to predict nitrate concentration in shallow groundwater of the Central Valley, California: boosted regression trees (BRT), artificial neural networks (ANN), and Bayesian networks (BN). We developed the models within a statistical learning framework (Hastie et al., 2009) to optimize predictive performance. The Central Valley is an intensive agricultural region and produces 8% of U.S. agricultural value on 1% of the U.S. farmland (Reilly et al., 2008) (Fig. 1). Decadal increases in groundwater nitrate concentrations have been observed in portions of the Central Valley, particularly in the eastern fans (shown as light green on the map), which typify younger, oxic conditions (Burow et al., 2013). Competition for groundwater resources in the region calls into question whether the aquifer can remain a viable source of supply to drinking water wells (Faunt, 2009).

Suitability of groundwater for drinking depends both on quantity and quality. Statistical models are commonly used at large spatial scales to identify areas with high contamination potential and to understand factors that increase contamination risk. However, modeling groundwater contaminants derived mainly from the land surface is challenging because of numerous processes that influence solute transport and fate in soils and groundwater. Transport processes frequently are nonlinear and are complicated by the spatial variability of hydraulic and geochemical conditions in aquifers. Linear regression and classification methods have been popular choices for estimating nitrate impacts on groundwater (Ayotte et al., 2006; Boy-Roura et al., 2013; Frans, 2008; Gardner and Vogel, 2005; Gurdak and Qi, 2012; Huebsch et al., 2014; Jang and Chen, 2015; Ki et al., 2015; LaMotte and Greene, 2007; Liu et al.,

* Corresponding author. Tel.: +1 703 648 4000.
E-mail address: btnolan@usgs.gov (B.T. Nolan).

http://dx.doi.org/10.1016/j.jhydrol.2015.10.025
0022-1694/Published by Elsevier B.V.

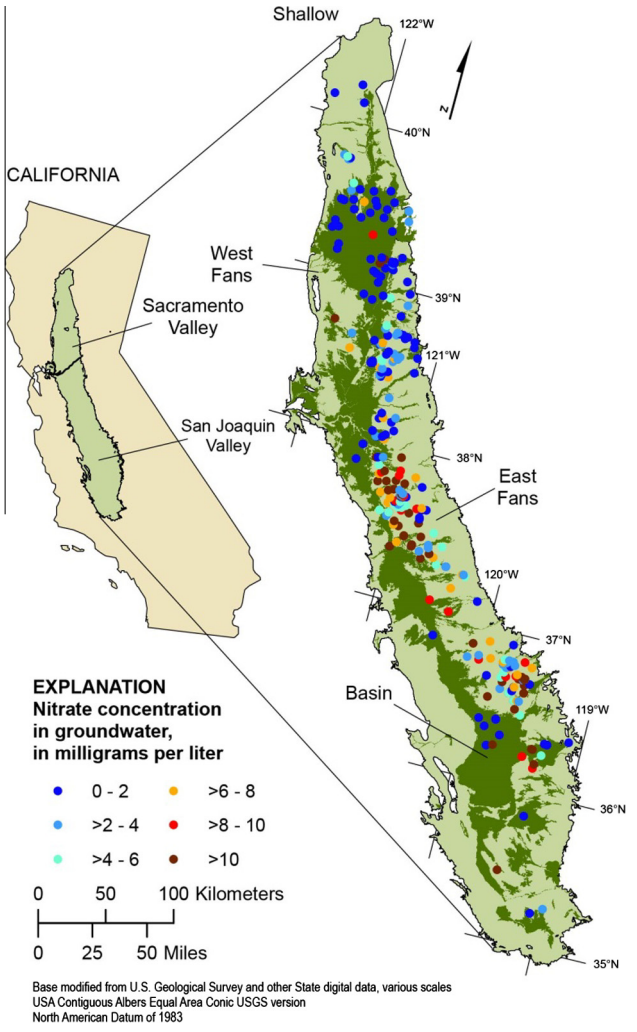

Fig. 1. Locations of shallow wells used to develop the models (modified from Nolan et al., 2014). The east and west fans are shown in light green and the basin subregion in dark green. Units of groundwater nitrate concentration are mg/L as N.

2005, 2013; Nolan et al., 2002; Rupert, 2003; Warner and Arnold, 2010). Although such methods are straightforward to apply at large spatial scales, hypothesis testing assumptions (linear and monotonic responses, assumed distributions of model residuals) are difficult to satisfy. For example, logistic regression assumes that the log odds ratio (logit) of observing some condition, such as exceeding a threshold nitrate concentration, is linearly related to a set of predictor variables.

Machine learning methods are promising alternatives that dispense with traditional hypothesis testing. For example, tree-based methods do not require data transformation, can fit nonlinear relations, and automatically incorporate interactions among predictor variables (Elith et al., 2008). Random forest regression (RFR), an ensemble tree method, was previously applied to shallow and deep wells of the Central Valley and yielded a pseudo R^2 of 0.90 for training data (Nolan et al., 2014). Random forest produces many classifiers (decision trees) and aggregates the predictions (Liaw and Wiener, 2002). The method employs bootstrap aggregating (bagging) to average the predictions over many trees, which reduces the variance of the prediction (Hastie et al., 2009). Random forest has only recently been applied to water resources data; other examples include nitrate and arsenic in aquifers of the southwestern U.S. (Anning et al., 2012), nitrate in an unconsolidated aquifer in southern Spain (Rodriguez-Galiano et al., 2014), and nitrate in private wells in Iowa (Wheeler et al., 2015).

A perceived disadvantage of machine learning methods is their "black box" nature; without estimated coefficients it is difficult to show significant relations between the response and predictor variables. However, individual classification trees can be extracted from BRT models and are easy to interpret. BRT also yields variable importance rankings and partial dependence plots. The latter can be used to infer the direction and degree of influence of predictor variables, and can provide additional insight by revealing nonlinear and non-monotonic responses. Nolan et al. (2014) used partial dependence plots to show that increasingly negative, MODFLOW-simulated vertical water fluxes (i.e., increasing downward) were related to increasing RFR-predicted groundwater nitrate concentration, particularly for deep wells during the irrigation season (see Fig. S2 in the Supporting Information of Nolan et al., 2014). Use of MODFLOW outputs as predictor variables in the RFR models constituted a multi-model, hybrid modeling approach. Variables with a high importance ranking by RFR included the depths to the top and midpoint of a well's screened interval. The first depth was a useful proxy for travel time from the land surface to the well, and the latter was a proxy for the groundwater age distribution. Bayesian networks are directed acyclic graphs comprising nodes (output and predictor variables) and edges (correlated connections between nodes) (Fienen et al., 2013). The graphic depiction of a BN is quite interpretable because the user draws the connections between predictor and response variables.

In the present study we evaluated BRT, ANN, and BN using the same data set as Nolan et al. (2014). The objective was to compare the predictive performance of the methods in the context of statistical learning, described in more detail below. The three models were then compared with the RFR model of Nolan et al. (2014) and multiple linear regression (MLR).

2. Material and methods

2.1. Data set

The Central Valley data set comprised 318 shallow domestic wells, and another 119 wells lacking screened interval data were held out for model evaluation (Nolan et al., 2014). Groundwater nitrate concentration data are summarized in Table 1. In the present study, the modeled response variable was the natural log of groundwater nitrate concentration (mg/L NO_3^- as N) in sampled shallow wells (i.e., domestic wells with depth below water table $\leqslant$46 m). The log transform reduced the influence of very high nitrate values (up to 74.7 mg/L) on model predictions. The 41 predictor variables represented soils, land use, groundwater age surrogates, and aquifer texture and MODFLOW-simulated vertical water fluxes from previous textural and numerical models of the Central Valley (Faunt, 2009) (Appendix A). All predictor variables were compiled within 500-m radius circular well buffers.

Table 1
Summary statistics of nitrate concentration in groundwater from shallow wells (from Nolan et al., 2014).

Variable	Nitrate concentration, mg/L as N
Minimum	<0.5
Maximum	74.7
Mean	6.38
Standard deviation	8.20
Median	3.61
Interquartile range	7.47
Number of observations	318