Predicting groundwater levels using linear regression and neural networks

Sara Maatta

December 15, 2011

Abstract

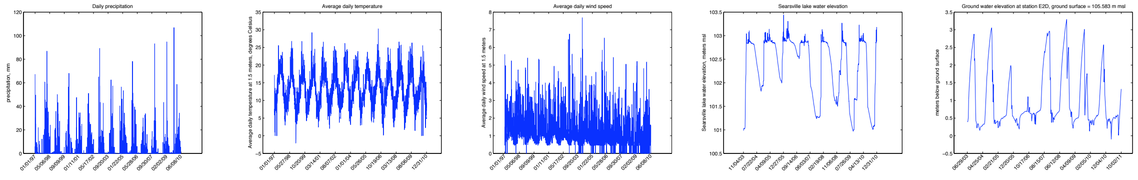
Water resources managers can benefit from accurate prediction of the availability of groundwater. In this project I present two models to predict groundwater levels in an unconfined shallow aquifer in the Searsville basin, part of the Jasper Ridge Biological Preserve. The input data (ie, features) for the models includes local weather, lake stage, and stream flow data, and moving averages of the weather, stage and stream flow data taken over time-frames of one week, one month, three months and six months. When moving averages are included as features, a linear regression model does well at predicting summer groundwater levels. In contrast, a feed-forward time-delay neural network does well at predicting winter groundwater levels. In combination, these models can provide useful predictions for groundwater levels throughout the year. Feature analysis indicates that the most important features are the longer time-frame moving averages that measure the "seasonality" of the example.

Motivation

In California, groundwater levels follow a roughly sinusoidal pattern: high in the winter and low in the summer. In a perfect world, we could use a mass balance to calculate groundwater levels using $\text{Mass}_{\text{in}} = \text{Mass}_{\text{out}} + \Delta\text{storage}$, but practically, we can only estimate global mass fluxes from measurements taken at discrete locations at discrete times. Precipitation, recharge from stream and lake levels and subsurface flows bring mass (water) into the system, and evapotranspiration, withdrawals (well pumping), discharge to streams, and subsurface flows take mass (water) out of the system. In an unconfined aquifer, the change in storage in the aquifer is represented by change in groundwater table as measured by piezometers. The change in groundwater table elevation is an approximately linear function of volume stored, depending on aquifer geometry.

In the field of hydrology, models of subsurface water flows typically require knowledge or estimation of the hydrologic parameters of the basin. The hydraulic conductivity and porosity of a soil define how water will flow through the subsurface, but these parameters are difficult to determine empirically; it requires thorough three-dimensional knowledge of the subsurface. Thus most hydrologic models estimate the effective hydraulic conductivity and porosity from relatively few soil samples.

This project will attempt to model groundwater levels without any explicit knowledge of the soil parameters of the basin. By training a machine learning model, the parameters and weights will implicitly represent the effective soil parameters of the basin. Instead of basin soil parameters, the inputs for the machine learning model for this project will include daily weather, lake stage, and stream flow data.



(a) (b) (c) (d) (e)

Figure 1: Hydrologic data from Jasper Ridge Biological Preserve (please zoom in to view)

Eventually, technology like groundwater prediction can help agriculture and water resources managers plan for availability of groundwater. The goal would be for this technology to be able to use weather predictions for a coming season to predict change in the groundwater levels.

Data

The weather station providing weather data for the project is located at the Jasper Ridge Biological Preserve. The lake stage data is from the Searsville reservoir, located on the Preserve. The stream flow data is from Corte Maderas creek, which flows into Searsville reservoir. The piezometer used to measure groundwater level is located less than 1km upstream of the reservoir.

The weather data include precipitation, average temperature, wind speed, and humidity. [1] The data for this project were provided by various student and faculty projects located at Jasper Ridge, and compiled by Dr. David Freyberg. Figure 1 shows timeseries of the data used in this project. Note that the date ranges for each series is slightly different, so a common subset (Nov 4, 2003 to June 8, 2010) will be used as the input (training and testing) for the models.

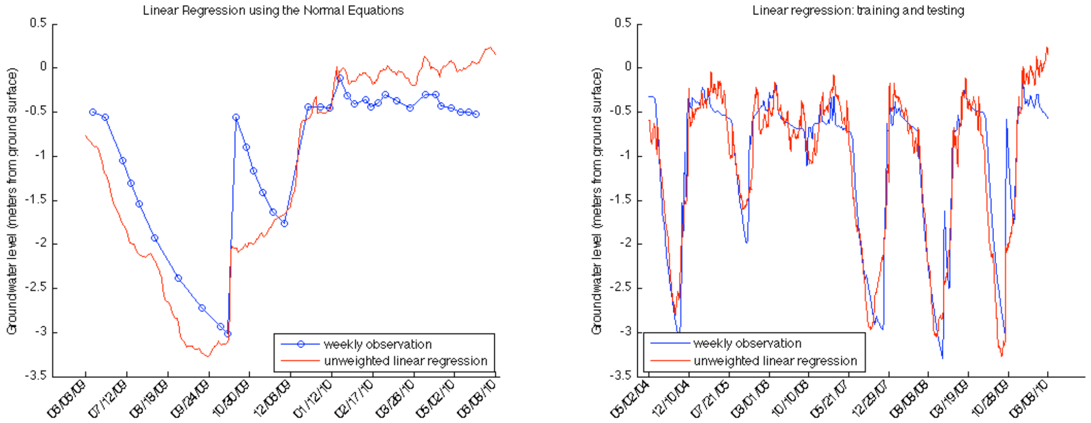
Data is preprocessed to have zero mean and unit variance, except precipitation, which is in units of centimeters.

Model results

Linear Regression

The first model implemented to predict groundwater levels was unweighted linear regression. For the training set, I used the subset of data from Nov 4, 2003 to June 8, 2009. For the test set, I used the subset of data from June 9, 2009 to June 8, 2010. The training matrix X is composed of rows containing the $x^{(i)}$'s, each with the data inputs for day i (weather, lake stage and stream flow), and $x_0^{(i)} = 1$. The results vector Y is composed of the groundwater level $y^{(i)}$ observed on day i. The unweighted linear regression finds θ according to to the normal equations: $\theta = \left(X^T X\right)^{-1} X^T Y$. Figure 2 shows the training and test results using unweighted linear regression. With 6 features, an evaluation of the model using the mean squared test errors results in $mse \approx 0.525$ meters squared.

[1] Evapotranspiration data are not collected at the Jasper Ridge weather station, though evapotranspiration is expected to play an important role in governing mass flows out of the aquifer system. Typically, when evapotranspiration data are not measured directly with evaporation pans and crop coefficients, they can estimated from average daily temperature, wind speed, humidity, and solar radiation. For this project, instead of explicitly calculating evapotranspiration using one of these methods, we will simply use temperature, wind speed and humidity data as inputs into our model, and allow the model to develop appropriates weights for these inputs.



(a) Test results

(b) Test and training results. Training occured from Nov 11, 2003 to June 8, 2009, and testing occured from June 9, 2009 to June 8, 2010.

Figure 2: Test and training results using unweighted linear regression

Model evaluation and improvement

An evaluation of testing and training error versus number of training examples indicates that the 6-feature system has high bias. Figure 3 shows the test and training error against number of training examples. To correct for the high bias, I created more features; specifically, features that give a measure of “seasonality” to each example. The features I selected were moving averages of the six original features, calculated for the previous one week, one month, three months, and six months. I also included a variable $M \in \{1 : 12\}$ to represent the month that the example is in, as another measure of seasonality.

With these 31 features, the linear regression model does much better, with mean test error of $mse \approx 0.246$ meters squared.

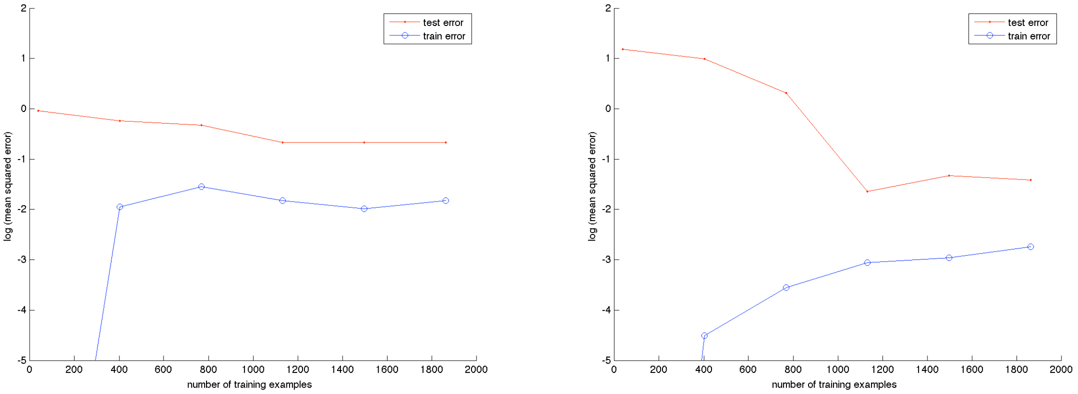
Inspection of the 31-feature model indicates that the model does worst during the winter months, when the groundwater levels are highest. In an attempt to help the model correct for this issue, I created another feature that is an indicator of “high groundwater” likelihood, based on the height of the lake stage (high lake stage is strongly correlated with high groundwater table). Thus if lake stage is above $102.7m$ mean sea level, the indicator variable is 1, and 0 otherwise. The mean squared error for the model including the high lake stage indicator is $mse = 0.243$ meters squared, so we can see that inclusion of this feature did not help the model predict winter levels any better.

Next I performed an ablative analysis to investigate which features and subsets of features are the most important. Table 1 shows the results of the ablative analysis. Interestingly, the six-month moving averages appear to be the most important features for the system. This results implies that the most important aspect of the system to consider is not the recent weather, but the past season’s weather. For an area with an arid climate like California, this finding makes intuitive sense, because groundwater levels are expected to be strongly influenced by the season.

Neural Network

I trained a feed-forward time-delay neural network with backpropagation using a sigmoid transfer function. The network architecture has two layers of hidden nodes: each layer had 10 nodes. I trained for 10 iterations (ie 10 epochs), with a delay vector of $(1 : 30, 90, 180)$.

The test error is $mse = 0.532$ meters squared, which is significantly worse than the linear



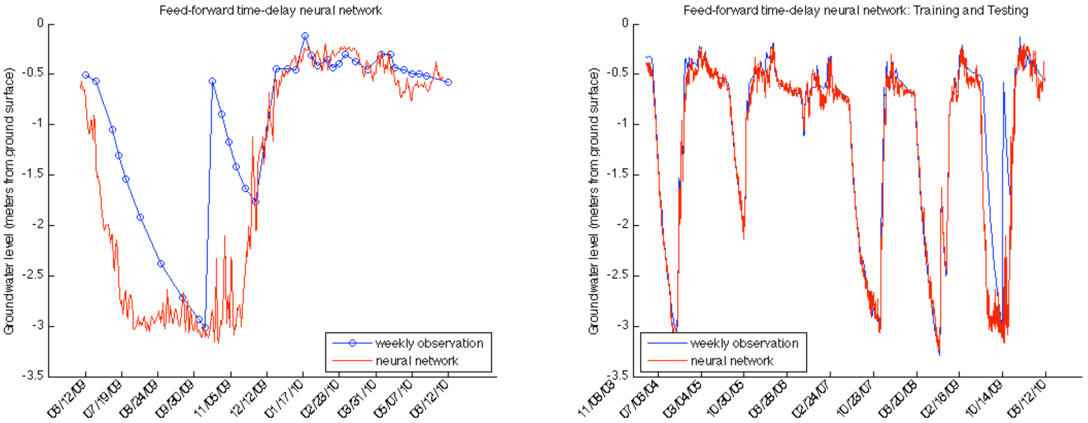
(a) 8 features

(b) 30 features

Figure 3: Test and training error for linear regression

Ablated feature	ΔTest error (%)	ΔTrain error (%)
– (all features included)	0 (by definition)	0 (by definition)
6 month moving averages	125%	79%
3 month moving averages	42%	26%
1 month moving averages	−1%	−7%
1 week moving averages	−2%	1%
Precipitation features	17%	5%
Lake stage features	−29%	71%
Temperature features	15%	26%
Wind speed features	−13%	4%
Humidity features	19%	18%
Stream flow features	31%	36%

Table 1: Ablative analysis of linear regression model: the higher the change in test error, the more important the feature(s) for the success of the model



(a) Test results

(b) Test and training results. Training occured from Nov 11, 2003 to June 8, 2009, and testing occured from June 9, 2009 to June 8, 2010.

Figure 4: Test and training results using a neural network

regression with 32 features, but comparable to the linear regression with 6 features. As you can see in Figure 4, the neural network severely overfits to the training set, and doesn't do very well on the test set.

However, the neural network does perform better than the linear regression in one respect: the neural network is much better at predicting the winter groundwater levels.

Model comparison and conclusion

Linear regression does a good job of predicting groundwater levels in the summer, when water levels are low, while the neural network does a good job of predicting groundwater levels in the winter, when water levels are high. This result supports the combination of linear regression and neural networks for predicting hydrologic response up to one year in advance.

Long time-frames (ie, 6-month and 3-month moving average) are extremely important features for the linear regression algorithm, while short term (1-month and 1-week moving average) are less important.

Acknowledgements

1. Data from David Freyberg, Department of Civil and Environmental Engineering, Stanford University

2. "Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications" by Maier and Dandy, Environmental Modelling & Software, 15: 101–124, 2000.