Predicting Water Usage for Automated Irrigation System

George Gui*
Graduate School of Business
Stanford University
ggui@stanford.edu

Abstract

This paper studies whether neural networks can effectively predict the required water usage level when weather data is noisy and soil data is unobserved. By running prediction on simulated data, we showed that the variation in unobserved soil characteristics and measurement noises in ETo rates greatly affect the prediction. In contrast, the measurement error of precipitation doesn't affect the prediction significantly. For firms that are working on improving automated irrigation system, this has important strategic implication for data collection, which is generally expensive.

1 Introduction

Net Irrigation Requirement (NIR) is the amount of irrigation required for a lawn given the soil and weather conditions. If an automated irrigation system is capable of predicting the optimal irrigation amount, it will effectively help crop cultivation and lawn maintenance in terms of saving human resource and water. Based on a scientific model known as evapotranspiration (Nouri et al., 2013), the optimal irrigation amount depends on the evapotranspiration rate (ETo), the effective rainfall (ERain), and the soil water storage. Since the soil water storage depends on past rainfalls and evaporations, a sequential model is ideal for this task.

However, one challenge is that many of the above variables are unobserved or measured inaccurately in the observational data, as discussed in Torres et al. (2011) and Park et al. (2016). Despite the effort made in measuring the relevant variables and the process, it can be expensive and difficult for an automated irrigation system to collect more detailed data, as it involves installing more tracking devices per system. Therefore, we are interested in exploring how different kinds of measurement errors and unobserved features affect the prediction. Such understanding enables us to decide what data to collect to improve our prediction: we should only focus on those variables whose measurements greatly affect the prediction. Since the underlying scientific model that generates the data is known, we plan to investigate this problem by running simulation.

2 Background

2.1 Observational Data

In the process of evapotranspiration, a number of variables are involved in a given day d and a household i. We classify these variables into the following categories based on how accurately they are observed in the observational data

Measured accurately:

*This project is advised by Professor Wesley Hartmann and Kristina Brecko.

- $NIR(d, i)$: Net irrigation requirement on day d and household i.

Measured inaccurately:

- $Precip(d, i)$: Precipitation. We have the average rainfall in a region based on the weather report but don't observe the specific level of rainfall of a house. For example, resident lives in the south of a mountain should get exposed to different levels of rainfalls compared to those in the north side.
- $ETo(d, i)$: evapotranspiration rate. Similar to precipitation, we observe the average evapotranspiration rate in a region, but not at the household level. For example, if areas are exposed differently to the sun because of their surroundings, they will have different levels of evaporation.

Unobserved:

- $MaxSoil(i)$: the maximum soil water storage of a given household. This variable depends on the soil characteristics, which differ across households and is costly to measure. It is unavailable in the current observational data.
- $SoilStorage(d, i)$: the water storage in the soil that evolves over time. It is completely unobserved. If the LSTM algorithm works well on the prediction algorithm, it should implicitly capture this unobserved soil storage variable.
- $ERain(d, i)$: effective rain

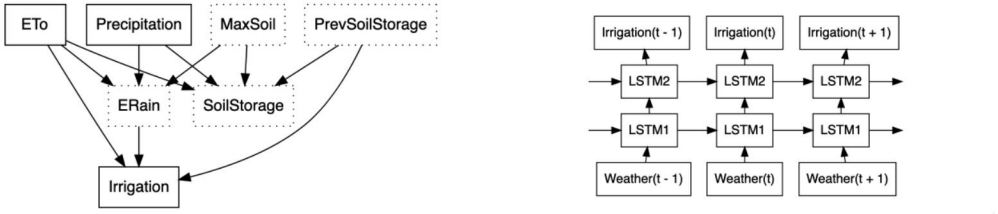
2.2 Scientific Model for Irrigation

The variable of interest, the Net Irrigation Requirement (NIR) variable evolves based on the following rules:

$$
\begin{aligned}
NIR(d,i) &= Max[ETo(d,i) - ERain(d,i) - SoilStorage(d-1,i), 0] \\
ERain(d,i) &= Min[Precip(d,i), (MaxSoil(i) + ETo(d,i) - SoilStorage(d-1)] \\
SoilStorage(d,i) &= Min[MaxSoil(i), (Precip(d) - ETo(d) + SoilStorage(d-1))]
\end{aligned}
$$

Figure 1a shows a visualization of the data generating process. The variables with dotted lines are unobserved.

Figure 1: Data Generating Process vs Neural Network



(a) Visualization of DGP

(b) Best structure used for prediction

3 Simulation

3.1 Motivation

Since many of the variables in the evapotranspiration process are not observed, the prediction algorithm doesn't work effectively given the current observational data, which has limited size and high noise. For companies who are developing an automated irrigation system, they need to decide what sensors are cost-effective in terms of measuring these variables to improve the irrigation

prediction. Indeed, one way to answer this problem is to design a system that can measure all the variables accurately, and then select different variables to test the problems. However, it is extremely expensive if we want to produce many of these devices, distribute it farmers or homeowners, and they collect data.

An alternative solution is through simulation. Unlike autonomous driving where a lot of sensors and road test are needed, the data generating process is clearly defined and known in this irrigation problem based on science. Therefore we can simulate data using and then explore how different level of measurements and amount of data will affect our prediction by simulation.

3.2 Simulation Setting

We first simulate unobserved variables as follows:

$$MaxSoil(i) \sim Unif(0.45, 0.45 + c)$$
$$SoilStorage(-1, i) \sim Unif(0, MaxSoil(i))$$

where c is a positive constant that represents how different soils are across households can be. Since $MaxSoil(i)$ is an important component that plays a nonlinear role in the data generating process, it can significantly influence the complexity of the problem.

For variables that are measured inaccurately-the evapotranspiration rate $\widehat{ETo(d,i)}$ and the precipitation $\widehat{Precip(d,i)}$, we simulate them by adding noise to the actual weather data.

$$\widehat{Precip(d,i)} = Precip(d,i) \times Max(1 + \delta_p(i) + \epsilon_p(d,i), 0)$$
$$\widehat{ETo(d,i)} = ETo(d,i) \times Max(1 + \delta_e(i) + \epsilon_e(d,i), 0)$$

where we assume

$$\delta_p(i) \sim N(0, \gamma_p^2); \quad \delta_e(i) \sim N(0, \gamma_e^2)$$

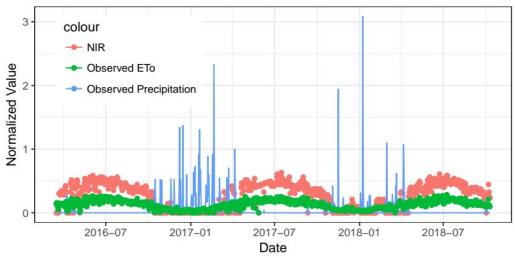
$$\epsilon_p(i) \sim N(0, \sigma_p^2); \quad \epsilon_e(i) \sim N(0, \sigma_e^2)$$

We use multiplication instead of addition for noise simulation because when there is no rain at all in a region, we are confident that any of the household didn't receive any precipitation, therefore when $Precip(d,i) = 0$, we want $\widehat{Precip(d,i)} = 0$. We still want the multiplier to be non-negative, because both $Precip$ and ETo should be positive.

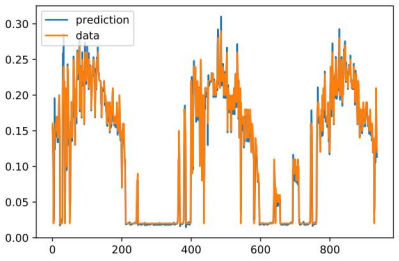
In summary, combining the three years of daily actual weather data $Precip(d,i)$ and $ETo(d,i)$ and the simulated soil data, we simulate the variable of interest, NIR, based on the evolving equations for 2000 households. In total, there are $2000 \times 365 \times 3 = 2,190,000$ observations. The task is to use the current and past noisy weather data $\widehat{Precip(d,i)}$ and $\widehat{ETo(d,i)}$, and past NIR to predict the current NIR.

4 Method

Since this problem has a sequential structure, we mainly use LSTM and hope that the network can capture hidden features that are related to the underlying scientific process. We feed ETo, Precipitation, and the NIR of the last period into an LSTM model to predict the NIR of the current period. Since the DGP has several layers, we also tried multiple stacked layers. Interestingly, a twice-stacked LSTM network performs the best, the structure of which is shown in 1b. This may be because the true data generating process also only has two layers. Figure (2b) is an example of the predicted NIR compared to the actual data over time, where the prediction is made by a 15-unit LSTM stacked by another 5-unit LSTM.



(a) Simluation Example



(b) A prediction example of NIR over time

5 Results

We test how prediction accuracy changes as we vary by the underlying parameters $\Theta = \{c, \gamma_p^2, \gamma_e^2, \sigma_p^2, \sigma_e^2\}$. We set our benchmark parameters to be $\Theta_0 = \{0.45, 0.01, 0.01, 0.01, 0.01\}$, for each parameter we vary them from a set of alternatives, while holding other prameters constant.

$$c \in \{0.1, 0.2, 0.45, 0.9, 1.5\}$$

$$\gamma_p, \gamma_e, \sigma_p, \sigma_e \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$$

5.1 Noise vs Accuracy

In order to test how different levels of noises affect the performance of the neural network, we simulate the data from the above range and compare the effect statistically by running a regression that regresses the prediction error (RMSE) on the noise parameters. If the coefficient is significantly positive, it suggests that larger noises significantly lead to higher errors. As shown in Table 1, what matters the most is the variation of soil characteristics across households and the measurement noise of evapotranspiration. In contrast, if we have constantly underestimated or overestimated the rainfall or the evapotranspiration, such bias would not affect our prediction, since the algorithm itself can almost figure it out.

Table 1: Regression Result on Accuracy vs Noise

	RMSE
Soil Noise	0.010^{*} (0.005)
Precipitaion Bias	-0.001 (0.012)
Evaporation Bias	-0.005 (0.012)
Precipitation Noise	0.003 (0.012)
Evaporation Noise	0.025^{**} (0.012)
R^2	0.333
Residual Std. Error	0.006 (df = 19)
Note:	*p<0.1; **p<0.05; ***p<0.01

Therefore, an automated irrigation system should be comfortable using county level weather data as input to the algorithm. However, getting additional information on soil characteristics and getting a better measurement on evapotranspiration can significantly improve the predcition.

5.2 Intepretation of hidden variables

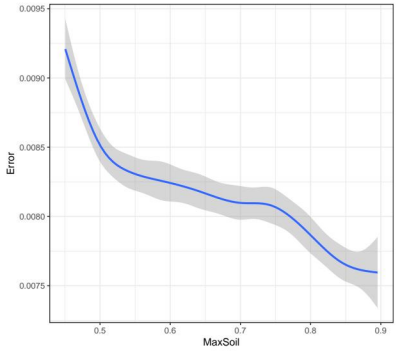
In our final output layer, the input are five latent variables that are the output of the LSTM unit. To understand whether the neural network picks up information about the underlying scientific process, we check whether the combination of these variables explain significant variation of initial soil storage, max soil storage, precipitation bias and evapotranspiration bias. These are key parameters in the data genrating process that are not observed. To do this, we first run pairwise correlational

analysis between hidden variables and the parameters of the data generating process. In the end we run a regression to test the R-squared eplained.

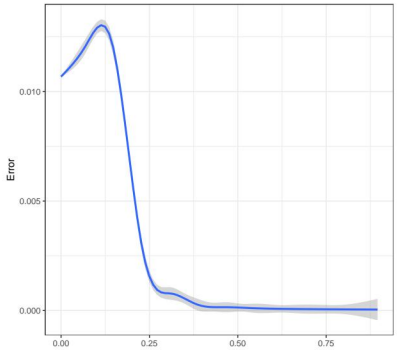
	Correlation with Hidden Features					R-Squared Explained	
	1	2	3	4	5	R2	R2 Adjusted
MaxSoil	-0.065	-0.005	-0.042	-0.091	-0.053	0.052	0.028
Initial Soil Storage	-0.048	-0.033	0.006	-0.063	0.011	0.049	0.024
ETo Bias	0.635	0.502	0.735	-0.443	0.727	0.723	0.716
Precip Bias	0.018	-0.291	-0.017	-0.304	0.288	0.297	0.279

5.3 Error Analysis

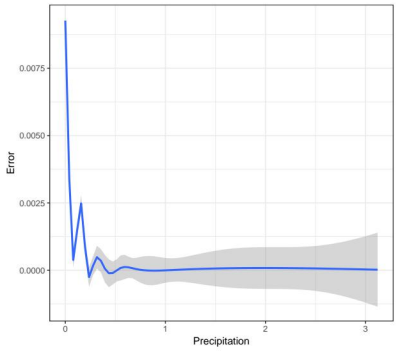
Section 5.1 compares the prediction across different datasets, generated by different parameter distributions. We are also interested in the performance of the algorithm within a dataset. This is valuable because if we can expect the cases when we know the prediction doesn't work well, we can always ask for human input in these cases and avoid unnecessary waste.



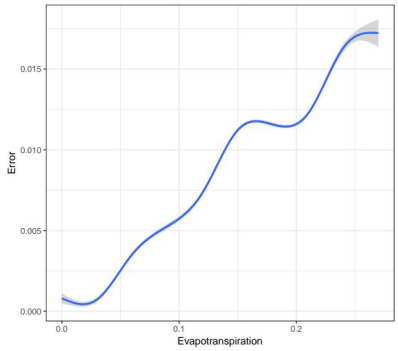
(a) Max Soil Storage vs Prediction Error



(b) Initial soil level vs Prediction Error



(a) Precipitation vs Prediction Error



(b) Evapotranspiration Level vs Prediction Error

6 Conclusion

In conclusion, we have used a stacked LSTM to successfully predict water usage, which can be applied to improve an automated irrigation system. We showed that the measurement accuracy of precipitation doesn't play a significant role, since the neural network can partially internalize these noises. Firms who are interested in developing these devices should focus on ways to measure soil characteristics, directly or indirectly. In general, it is more difficult to predict the irrigation level when the maximum soil storage for water is low, making the evapotranspiration process more non-linear. Possibly for the same reason, it is difficult to predict when the evapotranspiration is high. We should be less confident for soils with low saturation level and days when the ETo is high.

Code

`https://github.com/georgegui/CS230_Project`

References

Nouri, H., Beecham, S., Kazemi, F., and Hassanli, A. M. (2013). A review of ET measurement techniques for estimating the water requirements of urban landscape vegetation. *Urban Water Journal*, 10(4):247–259.

Park, S., Im, J., Jang, E., and Rhee, J. (2016). Drought assessment and monitoring through blending of multi-sensor indices using machine learning approaches for different climate regions. *Agricultural and Forest Meteorology*, 216:157–169.

Torres, A. F., Walker, W. R., and McKee, M. (2011). Forecasting daily potential evapotranspiration using machine learning and limited climatic data. *Agricultural Water Management*, 98(4):553–562.