Machine Learning Evapotranspiration

Anthony Nadelko

www.github.com/nad018/nad018_poster/

AGRICULTURE & FOOD - LANDSCAPES & GLOBAL CHANGE PROGRAM
www.csiro.au

Research Technician, Soil Process & Function - prior experience coding CRBASIC dataloggers, ladder logic PLC's and C+ microcontrollers. Previous work pattern involved data visualisation and analysis in Excel, further processing in open-source software, then presenting results in Excel graphs and Word tables.

My Synthesis Project

The project objective was to develop a machine learning model of evapotranspiration (ET) in irrigated cotton.

The raw data consisted of high frequency measurements from a wide variety of soil, crop and atmospheric sensor platforms. A set of 10 features and the target ET were selected from the raw data set. The target and feature data were reduced to 1 hour interval time series prior to analysis by two machine learning regression algorithms.

Hourly ET flux rates were calculated as the covariance of 10 Hz gas concentrations and wind vertical component speeds. Replicated features were averaged prior to frequency reduction. Soil water content and soil temperature measurements were natively hourly. One minute leaf temperature, dry leaf temperature, ambient temperature, humidity, vapour pressure deficit, wind speed, rainfall and solar radiation were reduced to hourly means.

Missing hourly timestamps were infilled, outliers removed and data gaps interpolated to generate time series of the target and each feature. The features were min-max scaled and hyperparameters tuned by nested cross validation prior to Lasso and Ridge regression analysis. The inital iterations were trained chronologically on 80% of data and tested on the following 20% of data.

My Digital Toolbox

Both R and Python were used in this project. R libraries used were tidyverse, padr, imputeTS, lubridate, and scales. Python modules used included numpy, pandas, scatter_matrix, pyplot, scale, cross_val_score, TimeSeriesSplit, Ridge, RidgeCV, Lasso, Lasso CV, mean_squared_error, r2_score. All of these digital tools were learned since starting Data School.

Favourite tool (optional)

My favorite Python tool was Mayplotlib for visualising data as it has capabilities not available from ggplot in R. In R I found the 'padr' library to be a very effective tool for analysing dataframes for time series gaps and inserting rows with missing timestamps infilled.

My time went ...

Processing the raw data with differing file structures and from a variety of source directories took 80% of the project time. After raw data processing most of the remaining time was used aquiring the knowledge to develop the machine learning workflow and to make appropriate subjective decisions for machine learning validation, training and testing. One very surprising finding was the regression coefficients indicated the sensor expected to be most important was 3rd in rank in Ridge and dropped entirely in Lasso.

Next steps

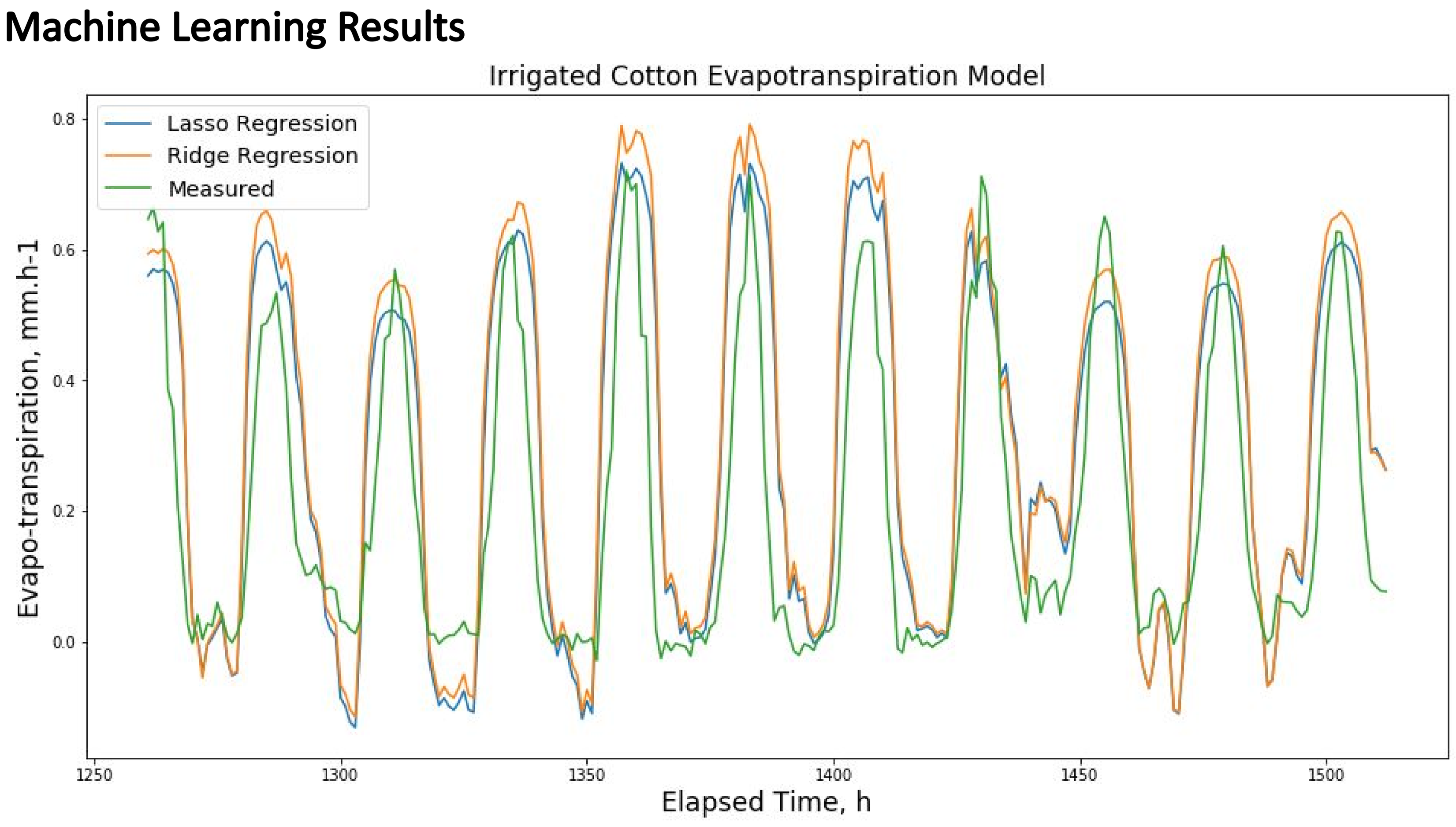
Ongoing work would include improving predicted ET accuracy by sliding window validation to better incorporate the influence of the increasing crop canopy cover over time, as well as examining the effect of reduced feature sets. Also of value would be applying CO_2 gas emmission rate measurements as the model target. Related time series models may also be developed for forecasting of ET and CO_2 emission rates.

Field Measurement Site

Target and Features Dataframe

hours	ET	leaf_temp	dry_ref_temp	ambient_temp	soil_temp	soil_wc	RH	VPD	wind_speed	rainfall	radiation
1	0.035792432	0.3800477	0.3612786	0.3464818	0.7899499	0.9444770	0.538904461	0.13506357	0.60016928	0	0.27915774
2	0.036816038	0.3305553	0.3102721	0.2999326	0.7912494	0.9448023	0.596542136	0.10213147	0.51357935	0	0.21596639
3	0.025843325	0.2964742	0.2754643	0.2613760	0.7900300	0.9456141	0.639603246	0.08228812	0.48840718	0	0.18102114
4	0.024775237	0.2743643	0.2557345	0.2415894	0.7859533	0.9460186	0.677198654	0.06857696	0.50484711	0	0.16545696
5	0.019882081	0.2508416	0.2334778	0.2215227	0.7789794	0.9468319	0.697808361	0.05991002	0.46469935	0	0.14865002
6	0.017956575	0.2237829	0.2033218	0.1971156	0.7694197	0.9462794	0.733657397	0.04852706	0.28478970	0	0.12845057
7	0.045312346	0.2427417	0.2376820	0.2137909	0.7559748	0.9465681	0.717989527	0.05243629	0.37581928	0	0.12555031
8	0.118136417	0.3096469	0.3194116	0.3097089	0.7417555	0.9561900	0.668079481	0.06821902	0.55162669	0	0.14513695
9	0.182168038	0.3780802	0.3599409	0.3521227	0.7199614	0.9605027	0.650658878	0.07694104	0.56527221	0	0.16405090
10	0.254384603	0.4497380	0.4010269	0.3877149	0.6975620	0.9614933	0.613701403	0.09356850	0.57719951	0	0.19392554
11	0.372957794	0.4974301	0.4345646	0.4139752	0.6745129	0.9621645	0.569943677	0.11572694	0.66762912	0	0.22646027
12	0.412094054	0.5323719	0.4665028	0.4409370	0.6540718	0.9627609	0.520752214	0.14266926	0.68914022	0	0.26458675

Machine Learning Results



MY DATA SCHOOL EXPERIENCE

I greatly enjoyed the ongoing support provided the network of Data School trainers, mentors and previous participants and was very impressed by the wide range of skills taught. I have been applying my new skills to recent project reporting, processing data and generating visualisations not possible before Data School. I will be presenting my Data School experiences and sythesis project results to our local laboratory group. As well, my machine learning project will be an additional outcome for the research project that provided the dataset and further model developments are anticipated.