Needle in a Data Haystack
Introduction to Data Science (67978)

The Kendall Irrigation Project

Using Data to Analyze Heterogeneity in Irrigation Patterns and Plant Water Uptake

Isaac Kramer, Noam Gottlieb, Karin Osadon

The Hebrew University of Jerusalem

March 2019

This is the paper's abstract . . .

I. Kramer (341352813), N. Gottlieb (201606951), K. Osadon (203057310)

1 Introduction

In recent years, many researchers have identified the potential that data science has to improve resource efficiency and crop yields in agriculture. This is especially true with respect to irrigation management. At present, most irrigation decisions are often made based on the assumption that field conditions are homogenous. Most fields, however, have internal variations in slope, soil depth and type, and aspect, among other variables. This variation can affect the distribution of irrigation water, resulting in certain plants performing better than others.

More accurate knowledge of how heterogeneity affects the distribution of water and plant growth could help both increase plant growth and allow more efficient use of water resources. To this end, a team of researchers at the Volcani Institute is engaged in a multiyear project aimed at increasing our understanding of how spacial heterogeneity within an agriculture field affects the distribution of irrigation water, and how this in turn affects plant production.

During 2017-2018, this team conducted a field experiment in a wine vineyard located within Israel's Judean Hills region (Mevo-Beitar). The area of the vineyard is 2.4 ha and its terrain is highly variable. The vines in the vineyard were planted in 2011 in a northwest-to-southeast direction with a vine and row spacing of 1.5 m and 3.0 m, respectively.

As a part of their ongoing experiment, the researchers divided a section of the vineyard into 10 "zones." Over the course of the growing season, the different zones were subjected to different irrigation regimes and the plant health within each of the zones was measured.

2 Data

The research team provided us with the following data associated with this experiment:

I. Kramer (341352813), N. Gottlieb (201606951), K. Osadon (203057310)

1. Readings from 30 Time-Domain Reflectometry (TDR) sensors. The TDR sensors were used to measure soil volumetric water content, soil temperature, and soil electrical conductivity. Volumetric water content (θ) is defined as

$$\theta = \frac{V_w}{V_t}$$

where V_w is the volume of water in a given soil unit and V_t is the total volume of the soil unit (i.e., it includes the volume of air, solid, and water in the soil). Volumetric water content is by definition dimensionless and its values can be reported as a decimal or percent. Electrical conductivity is useful as an indication of the quantity of salts (Na^+, Ca^{2+}, Mg^{2+}, etc...) in a soil. While many salts are essential to plant development, concentrations above specific thresholds can cause significant declines in plant growth.

The 30 TDR sensors were divided such that three were placed at different locations within each of the 10 zones. Each of the TDR sensors was buried in the soil at a depth of 30 cm. Measurements were recorded every five minutes. We received data corresponding to the three month period beginning in May 2018.

2. Measurements of the Stem Water Potential (SWP) for 30 plants. SWP measurements are useful as an indicator of the degree water stress a plant is experiencing. Readings that are highly negative correspond to high levels of water stress, while readings that are closer to zero indicate that a plant is not experiencing water stress.

Each of the 30 selected plants was adjacent to one of the previously mentioned TDR sensors. The condition of the selected plants should, in theory, reflect the condition of the adjacent soil. The SWP data that we received corresponds to weekly measurements taken during the period from May to July 2018.

3. The quantity of irrigation water supplied to each of zones. Decisions about whether to irrigate each of the given zones were made weekly, based on the SWP measurements. If the SWP measurements indicated the plants in a particular zone were not experiencing water stress, they were not irrigated that week. If the SWP measurements indicated water stress, the plants were irrigated based on the degree of stress exhibited (i.e., more stressed plants larger irrigation inputs than plants with lower levels of stress).

   The irrigation quantities were reported in terms of volume of water per dunam (i.e., mm). All irrigation in the vineyard was done using freshwater.

4. Selected soil properties at the location of the TDR sensors. This data included the aspect, slope, elevation, and total soil depth in the area where each of the TDR sensors was placed.

3 Objectives

Our objectives for this project were to mine the dataset to explore possible connections between soil conditions, plant water stress, and soil properties. Specifically, we defined the following sub-objectives:

- Can we establish a connection between soil water content and the appearance of stress in that plant?

- Can we determine which soils behave most similarly? For each sensor we have data for the following variables: Water Content, Temperature, and Electrical Conductivity. If we make these three values a vector (i.e., one vector per sensor per time step) can we identify soils that are most similar to each other?

- Initial analysis of the TDR measurements indicated that following irrigation events,

some areas lost water at much faster rates than others. We investigated whether this phenomenon can be connected soil properties (slope and soil depth).

- more goals, if there are any...

4 Data Cleaning

Our first step in the project was to do a preliminary clean of the TDR data. We updated the date-time format, simplified the column headers, and ensured cells with missing values were represented uniformly. This was accomplished using the script data_clean.py.

We continued the data cleaning process by removing values at times when the sensors appeared to be malfunctioning. Figure 1 shows plots of the volumetric water content (%) for each of the 30 sensors during the duration of the measurement period. The plots also show dates of irrigation inputs (mm of water). The script facet_with_irrigation.py was used to build these plots. From these plots, we are able to see clearly that certain sensors were not functioning properly. For example, Sensor 1 in Area C shows no response to changes in water content, while Sensor 2 in Area I makes extreme jumps. Before proceeding with our analysis we removed all results from the sensors that did not appear to be working properly.

5 Results

5.1 Relationship between Plant Health and Soil Water Content

At the heart of this experiment is the expectation that higher values for Volumetric Water Content (VWC) will lead to smaller (i.e., less negative) values of Stem Water Potential (SWP). That is, that increased water availability will lead to healthier plants. We therefore sought to test this assumption.

I. Kramer (341352813), N. Gottlieb (201606951), K. Osadon (203057310)

Our code (SWP_scatter.py) compares the SWC to the SWP, restricting analysis to the days in which the SWP was measured. Plotting all the values for SWP against SWC, as shown in Figure 2 we see little visible correlation.

Dividing the data according to sensor, however, we begin to see actual trends. In Figure 3 we can start to see that as SWC declines, the SWP becomes more negative, in most cases. Table 1 shows the R-Squared values for each of the sensors, based on linear models. With the exception of area D and area H, the R-Squared values and plots indicate that there is a correlation between SWC and SWP. The fact that we do not see a clear trend when all the data is plotted together is not surprising, since heterogeneity in the soil and other factors (e.g., position with respect to the sun) can easily explain why plants in two different areas would respond differently to the same SWC. That we do see clear trends on an individual sensor basis reinforces the need to account for heterogeneity in agricultural systems.

5.2 Applications of Dynamic Time Warping and Clustering to Plant Health and Soil Water Content

We continued to explore the relationship between soil water content and stem water potential through the application of dynamic time warping (DTW) and clustering techniques. DTW is a tool used to measure the similarity between time series data. It can be used to identify similarity in cases when the sequences move at different speeds or experience different amplitudes. We believe that this was an appropriate method to use with the current dataset because it could identify identify soils (i.e., sensors) that behave relatively similarly, even when their specific wetting and drying speeds and overall maximum are different.

We began our analysis using PyPI's DTW package and working only with the SWC data (distances.py). Due to the long running time of the algorithm, we resampled the

I. Kramer (341352813), N. Gottlieb (201606951), K. Osadon (203057310)

Table 1: R-Squared values from linear models of SWP as a function of SWC

Sensor	R-Squared
A1	0.44
A2	0.70
A3	0.58
B1	0.53
B2	0.57
B3	0.41
C2	0.60
C3	0.36
D1	0.00
D2	0.00
D3	0.10
E1	0.38
E3	0.30
F1	0.09
F2	0.27
F3	0.38
G1	0.45
H1	0.341
H2	0.17
H3	0.04
I1	0.45
I3	0.63
J2	0.39
J3	0.64

data to intervals of 1 hour (rather than the original 5 minutes intervals). The results are shown in Figure 4. The matrix in this figure shows the similarity between each pair of SWC sensors, with dark squares indicating sensors that are more similar.

Next, we sought to apply a clustering algorithm to help us identify which soils were most similar. Due to high-dimensionality of our data, we used hierarchical clustering instead of K-means. We relied on the scipy package to perform the clustering (VWCclustering.py). The results of our clustering our shown in Figure 5.

Finally, we applied the same DTW and clustering techniques to the SWP dataset. If the SWP dataset showed similar clusters, then this could help future irrigation planning.

That is, if our time series for change in soil water content produced the same clusters as our time series for stem water potential, then this could indicate that the soils within each cluster behave similarly and can be irrigated with similar regimes in the future. Unfortunately, our results (Figure 6) did not show similar clusters. This is not surprising, given our findings in **??**, in which we saw that a correlation between SWP and SWC could only be established when controling for area.

5.3 Effect of Slope and Soil Depth on Water Loss Rate

The plots show that some of the soils lose water at a much faster rate following an irrigation event than others. For example, if we compare Areas C and D, we see that the volumetric water content declines much faster for the sensors located in Area C than for the sensors located in Area D. The researchers working on this project hypothesized that one reason for this might be variations in slope and soil depth.

To investigate this question, we wrote a code that calculated the "rate of decline" following an irrigation event (slopes.py). This code identifies the local maxima in water content (i.e., the water content following an irrigation event), as shown in Figure 7. It then calculates the rate of decline following these events. The code can be adjusted to calculate the rate of decline for a varying amount of time following an irrigation event (e.g., to consider just the first hour after an irrigation event, one day, several days, etc...). Because each sensor experiences multiple irrigation events, the code calculates an average based on each of the events.

Unfortunately, our results showed little connection between the "average loss rate" and the soil slope and/or soil depth. Figure 8 includes plots showing the average loss rate following an irrigation event vs. slope and soil depth (cm). As is clear in these plots, no visible relationship exists between the variables. An attempt to build a multiple linear regression model using both slope and soil depth as explanatory variables also produced

an R-squared value of nearly 0.0. The results did not change when we varied the length of time following an irrigation maximum used to calculate the average loss rate.
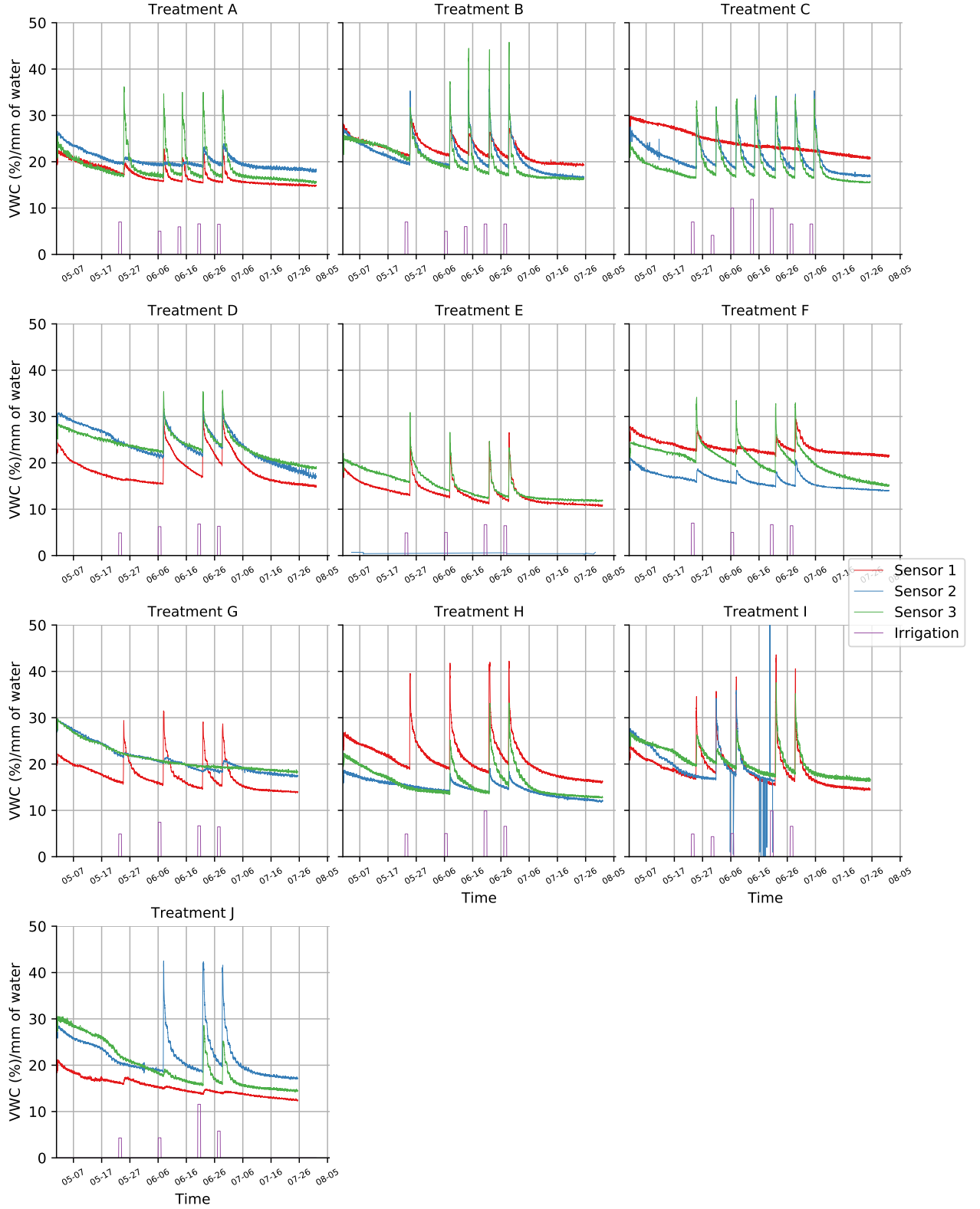
I. Kramer (341352813), N. Gottlieb (201606951), K. Osadon (203057310)

Figure 1: Plots showing the volumetric water content and irrigation inputs against time before removing data from malfunctioning sensors. The script facet_with_irrigation.py was used to build these plots.
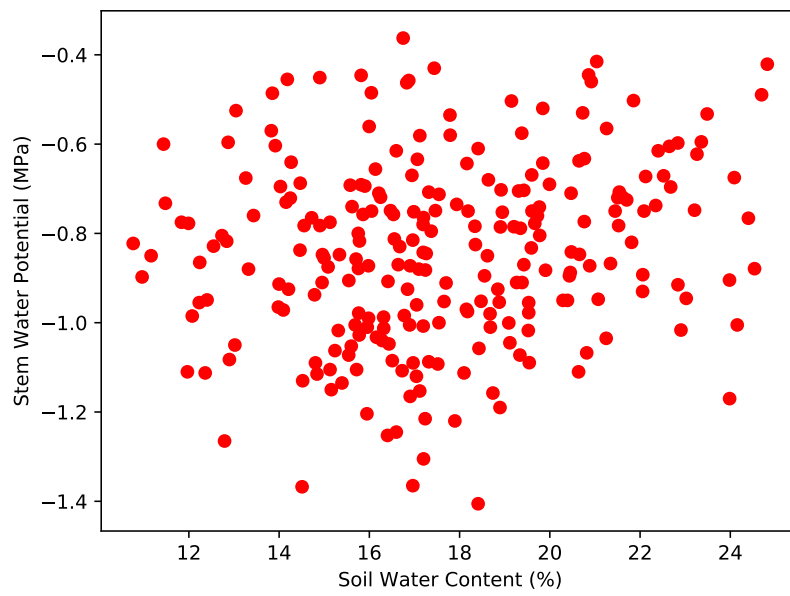
I. Kramer (341352813), N. Gottlieb (201606951), K. Osadon (203057310)

Figure 2: Scatter plot showing the Stem Water Potential against Volumetric Water Content (SWP_scatter.py).

I. Kramer (341352813), N. Gottlieb (201606951), K. Osadon (203057310)
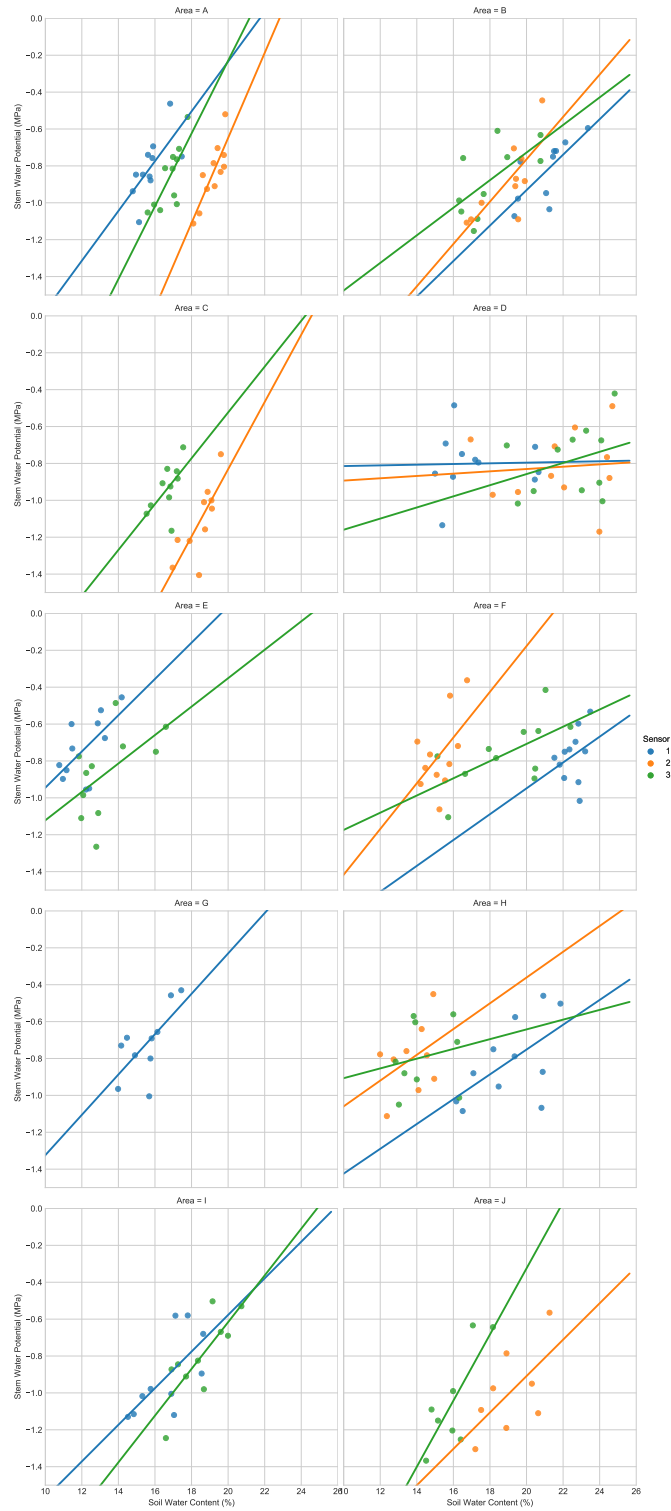
Figure 3: Scatter plot showing the Stem Water Potential against Volumetric Water Content (SWP_scatter.py).

I. Kramer (341352813), N. Gottlieb (201606951), K. Osadon (203057310)
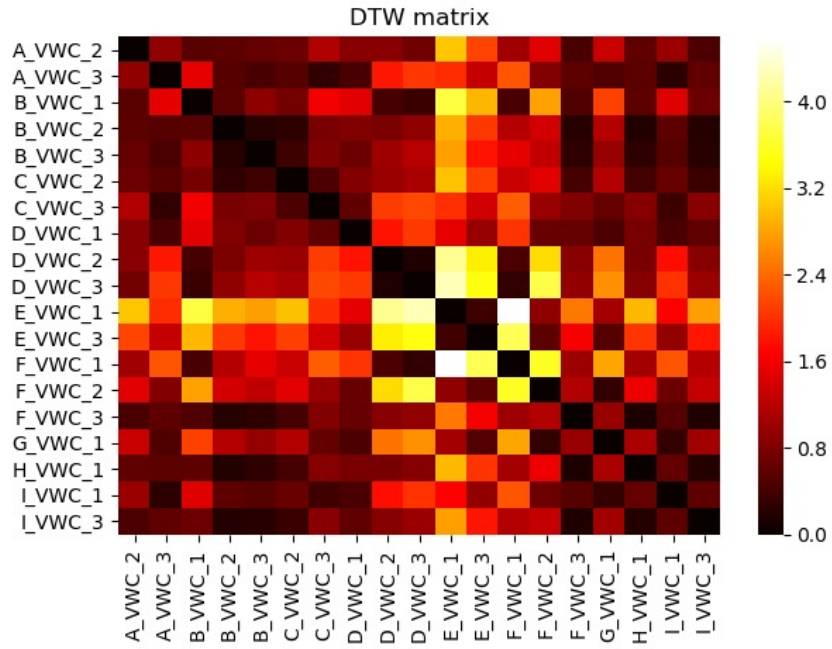
Figure 4: Matrix showing the application of DTW to the VWC data.
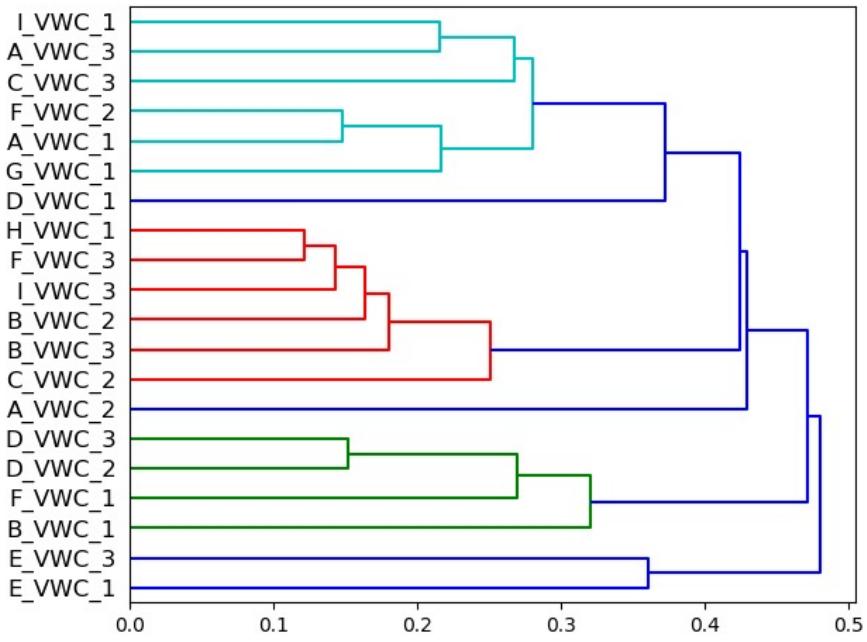
Figure 5: Plot showing the application of hierarchical clustering to the DTW output when applied to the SWC dataset.
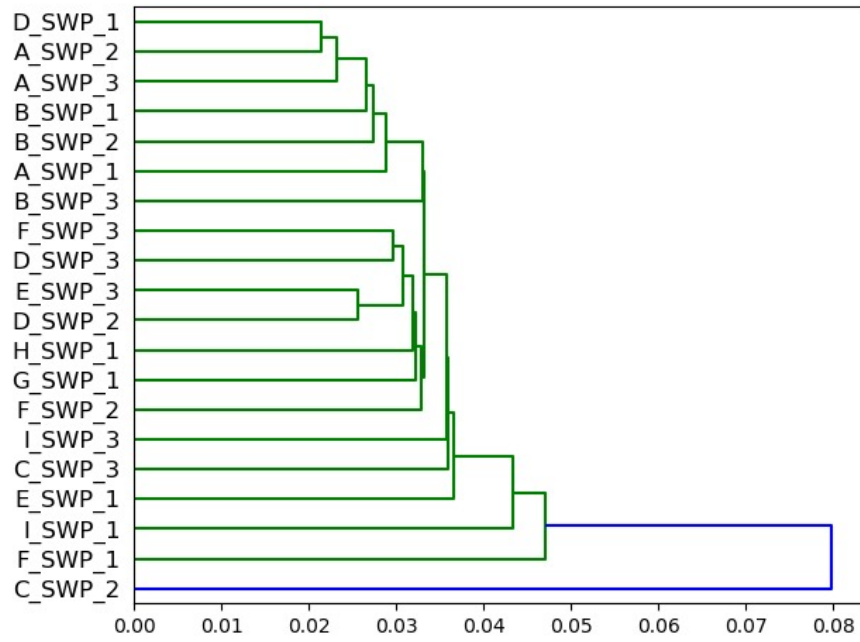
Figure 6: Plot showing the application of hierarchical clustering to the DTW output when applied to the SWP dataset.
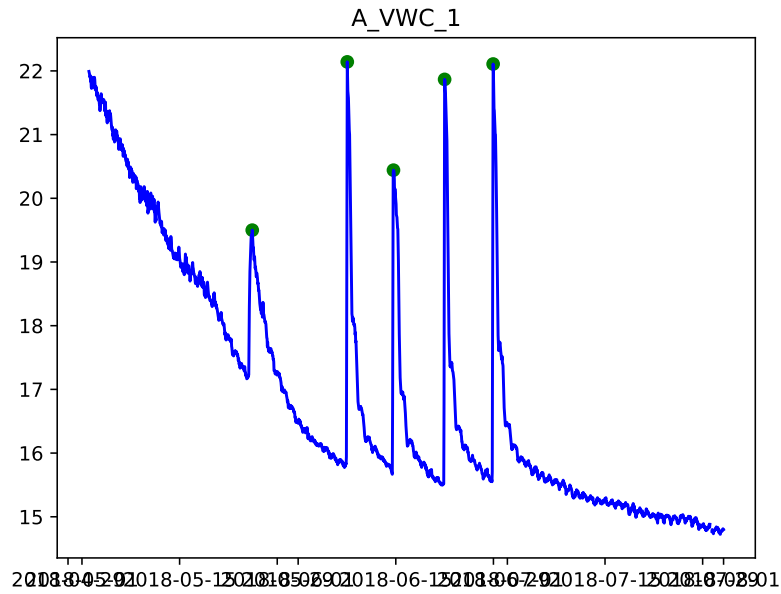

Figure 7: Plot showing the maximum volumetric water contents reached for Area A, Sensor 1 as identified by the code in slopes.py

I. Kramer (341352813), N. Gottlieb (201606951), K. Osadon (203057310)
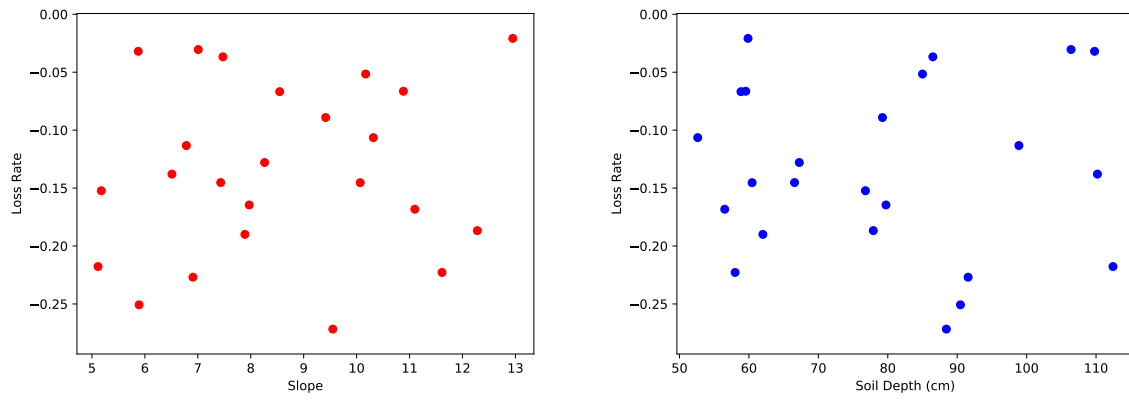

Figure 8: Plots showing average loss rate following an irrigation event vs. slope for each of the sensors and soil depth (cm).