How Soil Moisture in North Carolina is Effected by the Amount of Solar Radiation and by the Soil Type

Meagan Bird and Rebecca Register

Statistics 380
December 4, 2009
Executive Summary

This study had two specific goals which were does soil type have an effect on soil moisture and does solar radiation have an effect on soil moisture? In order to study these goals we completed a statistical analysis based on the raw data of the parameters of soil moisture and solar radiation observed in North Carolina which were stored in the North Carolina State Climate Office CRONOS Database. We wanted to see if we could find a clear correlation of how soil moisture behaves based on these two specific factors. Based on our data we were interested in ruling out the one, if any, that played the smaller role in order to focus on one specific parameter in the event of an opportunity to further study these connections. After we gathered all of our data we input it into R and ran some analyses on it to find any correlations between the parameters and to ultimately answer the questions we posed in our goals.

When we ran the R code and fit a linear model we found that one of our goals could be answered satisfactorily due to a high R-squared value correlation, which was approaching 1. Our remaining goal however had a low R-squared value which shows us that there is no significant correlation. The goal that did well was the first one pertaining to attempting to find a link between soil moisture and soil type. The R-squared value for this goal was 0.8863. The corresponding p-value for this linear model was <2.2e-16 which is very close to 0 and a good sign. This linear model was a combination of the soil type, date, and soil type multiplied by date. When we looked at both the soil type and date individually to see which one was the larger contributor we found that the soil type accounted for 88% of the correlation and date only explained 0.2 %. On the other side, the R-squared value for the linear model of moisture with respect to the combined data was extremely low at 0.004315, meaning only about 0.4% of the data can be explained using this combination. Because of such a low combined R-squared value, we didn’t even bother looking at each component individually. Another confirmation that this combination is no good to us is the fact that the p-value is 0.4462 which as far as p-values go is significantly higher than zero, which is the desired value.
Table of Contents

I. Description of Data................................................................. 4
II. Statistical Analysis................................................................. 5
III. Major Findings....................................................................... 6
IV. Discussion............................................................................... 7
V. Appendix.................................................................................. 8
VI. Visuals.................................................................................. 15
VII. Resources............................................................................. 16
I. Description of Data

The majority of our data we collected was courtesy of the North Carolina State Climate Office CRONOS Database. From this database we were able to select which sites we wanted to look at across the state from the thirty-six available ECONet (Environment and Climate Network) stations. Of these thirty-six we chose four stations that seemed to be in different geographical regions which meant that the soil composition for each location would be different. The stations we chose, looking from west to east, were Mountain Research Station (WAYN) in Waynesville, NC, Sand Hills Research Station (JACK) in Jackson Springs, NC, Central Crops Research Station (CLAY) in Clayton, NC, and Tidewater Research Station (PLYM) in Plymouth, NC. Visuals of each location can be seen on page 15-16. Included in the visuals are topographic maps of the sites locations in the state as well as 360° panoramic views around the stations.

In order to verify if the soil types were in fact different, we used a web soil survey internet site. In order to learn the exact soil composition we input the address of each station and the site pin pointed the exact location. We then selected a large enough area surrounding that point as our area of interest and the site provided us with the soil composition of that location based on the soil records that have been stored from previous geological surveys. Upon looking up each station we found that WAYN contained two soil types which were 90% Braddock Clay Loam and 10% Edvard-Cowee Complex, JACK contained Candor Sand, CLAY was comprised of Norfolk Loamy Sand, and PLYM had Roanoke Loam. This confirmed that each station did indeed reside on areas of differing soil types and we were therefore able to move on with the data collection.

When collecting data stored in the CRONOS Database, normally you are only allowed to call up a limited amount of parameters for a time period of up to seven days prior to the current date. This was too small of a data set for us to be able to obtain a clear understanding of what we were interested in looking at so we contacted the State Climate Office and were able to be granted internal user access which allowed us to view all available data parameters and all available records going back as long as they have been kept which could be output in either hourly averages, daily averages, or monthly averages. Using this access we decided to go with the daily averages for a time frame of 1 May 2009 to 15 October 2009. We chose daily averages simply because using hourly records would have given us a data set of 15,000+ points and by
using monthly averaged data the amount of error would have increased significantly and we
would have only had twenty-four data points. By choosing daily averages we gave ourselves a
healthy sized data set of around 670.

Once we had our range we then proceeded to collect our two areas of interest which were
soil moisture measured in m^3/m^3 and solar radiation measured in W/m^2 for each location.
We had this data exported directly into an Excel spreadsheet in order for us to easily format it to
best suite our needs. Once we had all of the data in a format that we could easily work with we
were able to move it into a text file in order to input it into R and begin running our statistical
analyses on it.

II. Statistical Analysis

For our statistical analysis, we ran code comparing soil moisture versus soil type and soil
moisture versus solar radiation. However, first of all, we constructed a composite graph of all our
data (see figure 1, pg. 8). After constructing the composite graph, we calculated the summary of
our data. From the summary, we gathered the averages for each of our variables. The summary is
located in the appendix on page 8.

After looking at our data as a whole, we began analyzing the moisture. We constructed a
histogram of the moisture content, a histogram of the log of the moisture content, a scatter plot of
the log of the moisture versus the date, and a box plot of the log of the moisture for each soil type
(station). We found the box plot to be significant because it shows a significant difference in the
moisture content of each of the different types of soil (see figure 2, pg. 9). Next we created a
composite of graphs that show the moisture levels for the entire time period for each location
separately (see figure 3, pg. 10).

To summarize our analysis of moisture, we derived some linear models for moisture.
Our first linear model is for the log of the moisture versus station + date + station*date. This
linear model had each of the soil types marked as significant. The multiple R-squared for this
linear model was 0.8863 and the p-value was <2.2e-16. See appendix pages 10-11 for the
complete summary and a plot. Our second linear model is for the log of the moisture versus soil
type (station). This linear model also marked each soil type as significant. The multiple R-
squared for the second linear model was 0.88 and the p-value was <2.2e-16. See appendix page
11-12 for the complete summary and a plot.
After analyzing the moisture, we analyzed the radiation. First we constructed histograms of the radiation and the log of the radiation and a scatter plot of the moisture versus the log of the radiation. Next, we formed a composite of scatter plots that compare the log of the radiation to the moisture content (see figure 4, pg. 13). Finally, we calculated a linear model to describe the relationship between the log of the moisture and radiation + date + radiation*date. The moisture was the only significant value. The multiple R-squared was 0.004315 and the p-value was 0.4109. See appendix page 13-14 for a complete summary and plot.

III. Major Findings

While conducting our statistical analysis, we ran several different linear models in order to see if there was any correlation in data regarding soil moisture and our parameters of interest. All of the specific formulas can be seen in the appendix which starts on page 8. The first linear model we were interested in looking at was the log of moisture versus the combination of soil type, which we have listed in the code as station because they are interchangeable, the day, and the station multiplied by the day. When we output the summary of this linear model we found that there was a very high correlation between the log of moisture and this sum of parameters. The R-squared value was 0.8863 which means there is an 88% relation between them. In order to break this down more and see which is the most important variable out of those that were combined we decided to run a linear model of both the log of moisture versus just the station and then a log of moisture versus just the date. When we did this we found that the R-squared value for the station linear model was 0.88 whereas the R-squared for the date linear model was only 0.00241. This clearly shows that the most important variable thus far is strictly the station, which actually refers to the soil type.

Next we looked at the linear model of the log of moisture versus the combined data of the radiation, the day, and the radiation multiplied by the day. By running this linear model we found the R-squared value to be 0.004315, or 0.4% correlation. In other words there is absolutely no relationship as far as this study shows between the amount of soil moisture and the incoming solar radiation. Since we now know this, there is no need for us to do break down of linear models of each variable, it would just be a waste of time.

Another important output value from the fit summaries are the p-values. The p-value for the first linear model we ran was very low at <2.2e-16, which is also the same output for the
linear model of just the stations. This is to be expected since we found that the station was the most significant indicator. Just for comparison, the p-value of the linear model of just the day is 0.2057 and the p-value of the linear model for the combined data for radiation was 0.4109. Both of these values are significantly higher than zero which confirm that these are not respectable correlations.

IV. Discussion

No data is absolutely perfect…especially when it is derived from studying nature. Therefore, it’s natural to have data missing after it’s collected. Before we began analyzing our data using R, we had to clean it up by eliminating missing data lines. Fortunately, we only deleted 6 lines total from our data set. This creates error in itself. There is also human error like reading a measurement wrong. Additionally, not even modern technology can read a measurement accurate to the smallest degree. We also had error from the fact that we used daily averages versus hourly averages. However, this error did not affect our data analysis.

If this experiment were to be repeated, it should be more extensive in order to produce a more accurate result. There are several areas that should be changed in order to make this happen. First, the timeframe should be longer to include all four seasons. This will allow the experimenter to see soil moisture and solar radiation amounts for an entire year rather than a single season. Second, more samples should be taken from each soil type by choosing more than one location with each soil type. This will allow the experimenter to see if soil type really is a factor for soil moisture content. Third, samples should be taken from a larger variety of soil types. This will allow the experimenter to compare soil types to see how each soil type holds moisture. This data can be used by farmers to determine which soil types can be used for different types of crops.

While conducting our research, some questions were raised as to whether or not other factors contribute to soil moisture content. For example, amount of rain for each location can be added to the experiment as a possible variable that can affect soil moisture content. These variables, when combined with soil type, may cause an increased or decreased relationship with soil moisture.
data=read.table("clipboard", header=T)
head(data)
radiation=data$Radiation
moisture=data$Moisture
date=data$Date
station=as.factor(data$Soil_Type)
Data=data.frame(moisture, radiation, 
station, date)
Plot(Data)

Figure 1: plot(Data)

Summary(Data)

<table>
<thead>
<tr>
<th></th>
<th>moisture</th>
<th>radiation</th>
<th>station</th>
<th>date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min.</td>
<td>0.1130</td>
<td>0.0</td>
<td>CLAY:168</td>
<td>1.00</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>0.1560</td>
<td>169.2</td>
<td>JACK:168</td>
<td>43.00</td>
</tr>
<tr>
<td>Median</td>
<td>0.2350</td>
<td>226.1</td>
<td>PLYM:166</td>
<td>85.00</td>
</tr>
<tr>
<td>Mean</td>
<td>0.2517</td>
<td>216.3</td>
<td>WAYN:167</td>
<td>84.71</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>0.3310</td>
<td>275.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Max.</td>
<td>0.4550</td>
<td>352.7</td>
<td></td>
<td>168.00</td>
</tr>
</tbody>
</table>

par(mfrow=c(2,2))
hist(moisture)
hist(log(moisture))
log.mois=log(moisture)
plot(log.mois ~ date*station)
par(mfrow=c(1,1))
plot(log.mois ~ station)

plot(log.mois[station=="WAYN"] ~ date[station=="WAYN"])  
plot(log.mois[station=="JACK"] ~ date[station=="JACK"])  
plot(log.mois[station=="CLAY"] ~ date[station=="CLAY"])  
plot(log.mois[station=="PLYM"] ~ date[station=="PLYM"])  

Figure 2: plot(log.mois ~ station)
fit.mois = lm(log.mois ~ station + date + station * date)
summary(fit.mois)

Call:
  lm(formula = log.mois ~ station + date + station * date)

Residuals:  
            Min         1Q     Median         3Q        Max
  -0.4725335 -0.0796216  0.0006369  0.0737905  0.4393542

Coefficients:  
                Estimate Std. Error t value Pr(>|t|)  
  (Intercept)   -1.4260949  0.0211878  -67.307  < 2e-16 ***
  stationJACK    -0.5791645  0.0299641  -19.329  < 2e-16 ***
  stationPLYM   -0.1688091  0.0302118  -5.588  3.37e-08 ***
  stationWAYN    0.4903657  0.0299911  16.350  < 2e-16 ***
  date           0.0004764  0.0002175   2.190   0.0288 *
  stationJACK:date -0.0005114  0.0003076  -1.663    0.0968 .
  stationPLYM:date  0.0007402  0.0003092   2.394   0.0169 *
  stationWAYN:date -0.0006028  0.0003076  -1.960   0.0505 .

---

Figure 3: plot(log.mois[station == "WAYN"] ~ date[station == "WAYN"])
plot(log.mois[station == "JACK"] ~ date[station == "JACK"])
plot(log.mois[station == "CLAY"] ~ date[station == "CLAY"])
plot(log.mois[station == "PLYM"] ~ date[station == "PLYM"])

[10]
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1367 on 661 degrees of freedom
Multiple R-squared: 0.8863,   Adjusted R-squared: 0.8851
F-statistic: 736.1 on 7 and 661 DF,  p-value: < 2.2e-16

par(mfrow=c(2,2))
plot(fit.mois)

fit.mois2=lm(log.mois ~ station)
summary(fit.mois2)
Call:
  lm(formula = log.mois ~ station)

Residuals:
     Min      1Q  Median      3Q     Max
-0.491475 -0.075176 -0.004943  0.078922  0.407676

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.38584    0.01080 -128.368  < 2e-16 ***
stationJACK -0.62238    0.01527  -40.764  < 2e-16 ***
stationPLYM -0.10532    0.01531  -6.878    1.40e-11 ***
stationWAYN  0.43942    0.01529   28.738  < 2e-16 ***

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1399 on 665 degrees of freedom
Multiple R-squared: 0.8801,   Adjusted R-squared: 0.8796

[11]
F-statistic: 1628 on 3 and 665 DF, p-value: < 2.2e-16
plot(fit.mois2)

hist(radiation)
hist(log(radiation))
log.rad=log(radiation)
plot(moisture ~ log.rad)
par(mfrow=c(2,2))
plot(log.rad[station=="WAYN"] ~ moisture[station=="WAYN"])
plot(log.rad[station=="JACK"] ~ moisture[station=="JACK"])
plot(log.rad[station=="CLAY"] ~ moisture[station=="CLAY"])
plot(log.rad[station=="PLYM"] ~ moisture[station=="PLYM"])

fit.rad=lm(log.mois ~ radiation+date+radiation*date)
summary(fit.rad)
Call:
  lm(formula = log.mois ~ radiation + date + radiation * date)

Residuals:
   Min       1Q   Median       3Q      Max
  -0.760799 -0.394505  0.007816  0.364393  0.685934

Coefficients:
            Estimate Std. Error  t value  Pr(>|t|)    
(Intercept) -1.448e+00  9.955e-02 -14.5450 4.27e-16 ***
radiation   -1.326e-04  4.141e-04  -0.3200   0.749    
date         5.067e-04  8.804e-04   0.5761   0.565    
radiation:date -1.469e-06  4.184e-06  -0.3510   0.726    

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 . ’ 0.1 ’ 1

Residual standard error: 0.4033 on 665 degrees of freedom
Multiple R-squared: 0.004315,  Adjusted R-squared: -0.0001771
F-statistic: 0.9606 on 3 and 665 DF,  p-value: 0.4109

Plot(fit.rad)

fit=lm(log.mois ~ day)
summary(fit)

Call:
  lm(formula = log.mois ~ day)

Residuals:
   Min     1Q Median     3Q    Max
-0.72817 -0.38635  0.01003  0.36062  0.68961

Coefficients:    Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.4946485  0.0315106 -47.433  <2e-16 ***
day          0.0004082  0.0003223  1.267    0.206
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.403 on 664 degrees of freedom
Multiple R-squared: 0.00241,  Adjusted R-squared: 0.000908
F-statistic: 1.604 on 1 and 664 DF,  p-value: 0.2057

[14]
Resources

Parameters:  http://www.nc-climate.ncsu.edu/dynamic_scripts/cronos/map/?type=weather