GTECH 2017 Spring
Final Project Report
Chong Di

Modeling Tick (*Ixodes. Scapularis*) Density In Caumsett State Park

Abstract

The tick *I. Scapularis* (deer tick), known as the primary Lyme Disease transmitter, is widely distributed in Long Island. On one hand, the climatic conditions in Long Island is beneficial to tick habitat suitability. On the other hand, however, tick population and its life cycle are highly regulated by environmental factors. This essay first reviews several articles to have an understanding of abiotic factors that tick is intimately dependent on. Next it exemplifies Caumsett State Park as a case study in which remote sensing imagery is implemented to obtain large-scale environmental data based on microclimatic indicator discussed from literature review. Meanwhile, in a scheduled tick expedition on May 12th we collected 109 deer ticks which will be involved in a regression model to estimate tick density in Caumsett State Park. Finally, this essay is concluded with discussion on challenges in this study and further research.

Background

Tick *Ixodes. Scapularis*

I. Scapularis (***Ixodes*** refers to a hard bodied tick family) which is also known as deer tick or blacklegged tick is widely distributed in West Coast, East Coast and Northern Midwestern United States. Deer tick has a two-year life cycle during which it experiences egg, larva, nymph and adult. Attaching to its host and having a blood meal are the instinct and critical behaviors of deer tick as host blood provides energy for tick to reproduce or transform to next life stage.

A lot of local people have been brought up to fear deer tick due to its notorious reputation for transmitting various vector-borne diseases. For example, Lyme disease is one of the primary tick-borne diseases endangering public health. Human activities in tick emerging area are likely to encounter tick contacts. People may be infected by Lyme disease pathogen if they are fed on by a deer tick unnoticed for up to 36 hours. The trend of increasing number of Lyme disease cases in recent years has brought public concern nationwide, thus a lot of attempts have been made to track deer tick emergence and potential risk of Lyme disease infection.

Deer Tick Physiology & Environmental Factors

As off-host activities accounts for 98% of the two-year life cycle of deer tick, abiotic environment is playing a fundamental role in regulating its survival suitability **(Brownstein 2005)**. Tick is a cold-blooded vector which doesn't have active water absorbing behavior, thus internal temperature and water balance appeal to be essential to tick physiology **(Lindgren 2000)**. Although environmental factors such as temperature, humidity, precipitation directly influencing tick physiology are already known by tick ecologists, there are still other microclimatic factors having indirect and the complex mechanism of how these factors function as a whole in affecting tick physiology may still remain unknown.

Temperature is an important limiting factor for deer tick population. Tick becomes active under warm surface temperature in *questing*, which is a tick activity waiting for hosts on the top of vegetation. Warm temperature is a beneficial environmental factor for tick to hunt and reproduce, thus it accelerates tick’s life cycle, increases tick population and disperse tick habitat. In stark contrast, low temperature increase the duration of developmental periods and consequently increases the duration of tick’s life cycle, which is the proportion of ticks that died before completing its life cycle **(Gabriele-Rivet 2015)**.

In addition to temperature, solar radiation also influences tick ecology because solar radiation has potential influence on the dynamics of many environmental processes (i.e. air, soil temperature and moisture) **(Del Fabbro 2015)**. For example, solar radiation may has an impact on the duration of tick *questing* activity. *Questing* tick gradually loses water and turns dehydrated due to exposure of sunlight. Therefore, area with strong solar radiation shortens duration of *questing* activity and lower the chance of tick finding its hosts.

Tick cannot survive without woodland. Woodland provides open space for tick hosts’ activities. Additionally, vegetated area maintains a number of abiotic factors such as humidity, microclimatic temperature and soil profile. Thriving vegetation is more likely to hold soil and air moisture above ground and it also keeps kick from direct exposure to solar radiation which fastens internal dehydration. Therefore, vegetation also plays a fundamental role in tick habitat suitability.

Landsat-8 **Satellite Imagery**

Landsat-8 satellite imagery is relevant to this study due to the following reasons: first, *Landsat-8 satellite* image has a 30m × 30m spatial resolution. Sampling task can be completed in each 30m × 30m site with great efficiency and thus more sites can be surveyed during limited time; second, *Landsat-8* satellite has a 16 days temporal resolution. Tick life cycle and density are subject to seasonal changes thus satellite image with higher temporal resolution doesn’t provide better observation of seasonal variations; third, it is possible to derive environmental dataset from specific spectral bands from *Landsat-8* satellite images. For example, surface temperature and NDVI (Normalized Difference Vegetation Index) can be calculated using *Landsat-8 satellite* images **(Roy 2014)**. A summary of *Landsat-8* satellite imagery specification is shown below:

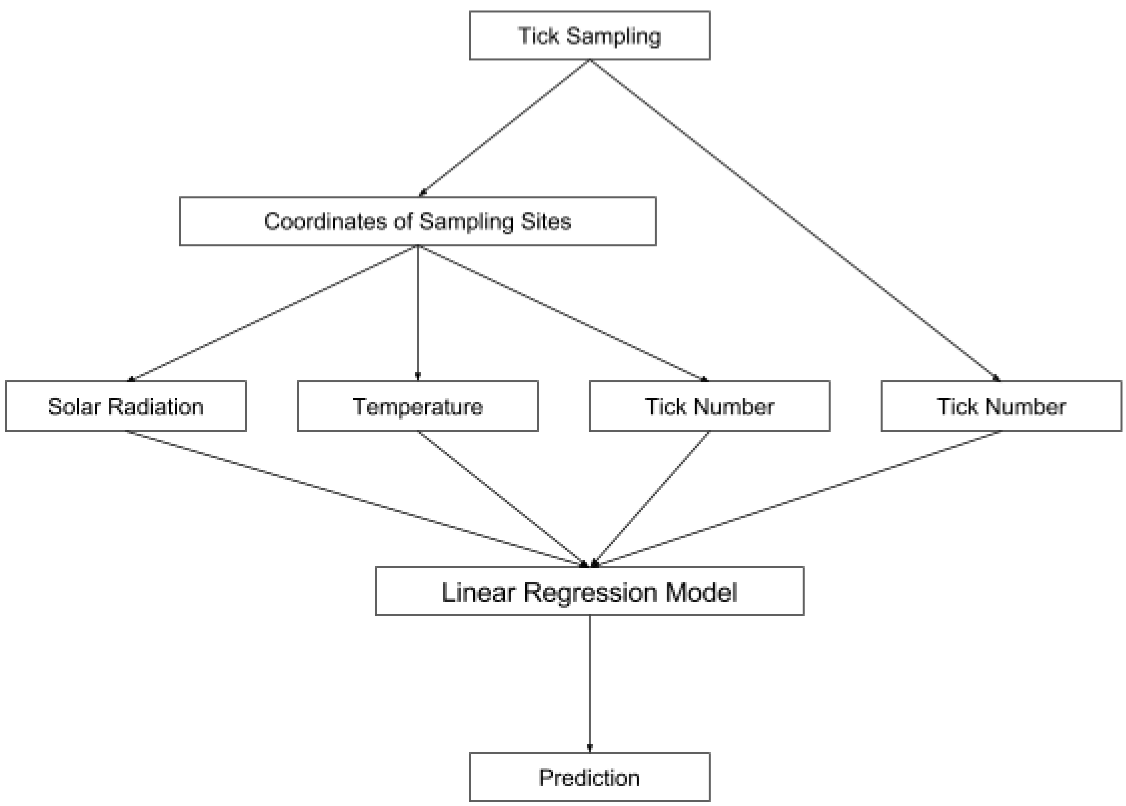
<table>
<tr><th>Temporal Resolution</th><th>Spatial Resolution</th><th>Spectral Band</th><th>Wavelength (μm)</th></tr>
<tr><td rowspan="7">16 Days</td><td rowspan="7">30 meters</td><td>Band 1 -- blue</td><td>0.43 - 0.45</td></tr>
<tr><td>Band 2 -- blue</td><td>0.45 - 0.51</td></tr>
<tr><td>Band 3 -- green</td><td>0.53 - 0.59</td></tr>
<tr><td>Band 4 -- red</td><td>0.64 - 0.67</td></tr>
<tr><td>Band 5 -- near infrared</td><td>0.85 - 0.88</td></tr>
<tr><td>Band 6 -- shortwave infrared</td><td>1.57 - 1.65</td></tr>
<tr><td>Band 7 -- shortwave infrared</td><td>2.11 - 2.29</td></tr>
</table>

	15 meters	Band 8 -- panchromatic (15m)	0.50 - 0.68
	30 meters	Band 9 -- cirrus	1.36 - 1.38
	100 meters	Band 10 -- thermal infrared	10.60 - 11.19
	100 meters	Band 11 -- thermal infrared	11.50 - 12.51

Method

As the objective of this project is using a linear regression model to predict tick density in Caumsett State Park, both environmental variables and tick number with corresponding geographic coordinates of each sampling site are required. In the previous section, three environmental factors influencing tick density have been identified, thus the next step is to discuss the strategy for data collection.

It is worth noting the order of data collection procedures in terms of the principle to apply linear regression model. Specifically, the model requires a smaller dataset including both variables and observation to conduct regression in which it first calculates the significance of each variable related to observations and then derives a linear formula consisting of variables and coefficients respectively. Then the linear formula could be applied to another dataset for prediction. The framework is shown below:



Therefore, first and foremost, tick sampling appeals to be the most essential task because this step acquires both tick number and geographic coordinates with which it is possible to collect environmental data corresponding to sampling sites.

Tick Sampling

Preparation & Objective: The tick sampling tools are listed in the following table:

Tools			
Tick prevention spray	Sampling flags × 4	Marking flags × 20	Measuring tape × 1
Anemometer & GPS	Counters × 2	Tweezers × 2	Tubes × 2

The Landsat-8 Satellite image has **30m × 30m** spatial resolution, thus the objective is to estimate tick population in each **30m × 30m** target area. However, it is impossible to completely cover a **30m × 30m** area because of two reasons:

1. Sweeping an area of 900 square meters is time consuming and labor intensive.
2. There are a lot of densely vegetated areas which are not accessible for sampling activity.

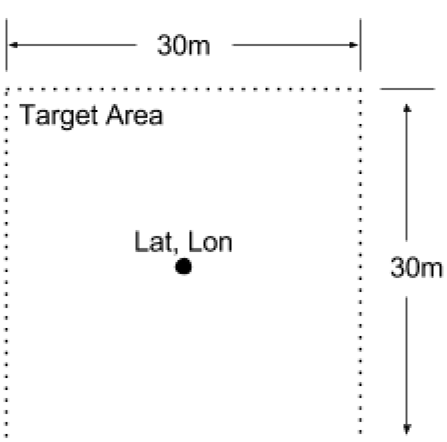
Thus sampling will be more efficient if randomly selecting smaller sampling sites within a **30m × 30m** area and estimating tick population in the target area.

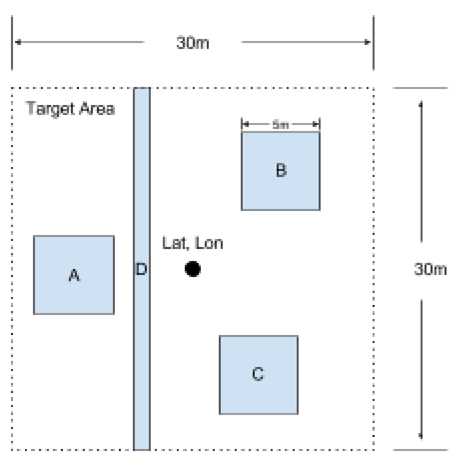
Sampling Procedure:

1. Before flagging: Counters should be reset to 0 before each trial.

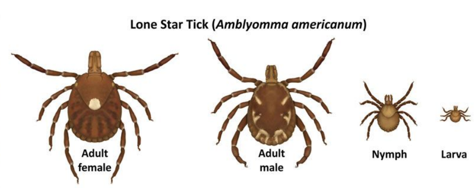
2. Select sampling area: The selection of target area should be based on landcover, e.g. grassland, bushes, trees. Once a target area is selected, its geographic coordinates should be read from GPS app on smartphone and recorded.

3. Select sampling sites: Sampling sites should be selected within the each area concerning the accessibility restricted by vegetation. The area of each sampling site should be marked by marking flags and the shape of sampling area should be recorded on the form.





4. Flagging: Each sampling sites will be swept twice. During sampling activities, what should be kept in mind is that it is possible to encounter lone star ticks which are not relevant to this project, thus an additional work is to identify deer ticks from every ticks being sampled. Every deer tick being picked with a tweezer and dropped into a tube adds one to the counter.



5. Complete sampling: After sampling work is completed in an area, all sampling flags are examined and make sure all ticks on the flag have been collected. The counters should be reset to 0 before sampling next area.

Landsat-8 **Satellite Images for NDVI & Surface Temperature**

The *Landsat-8* satellite image has 11 spectral bands and they can be used to generate both NDVI and surface temperature images. As *Landsat-8* satellite revisits the same area every 16 days and the tick sampling trip was scheduled on May 12th, the *Landsat-8* image of Caumsett State Park temporally closest to this data is May 4th. Thus the image was downloaded from USGS Earth Explorer and processed in *ENVI Desktop 5.4 with IDL (32-bit)*.

The strategies to generate NDVI and surface temperature images are different. To calculate NDVI basically requires multispectral image: selecting NDVI calculation from *Spectral Indices* window in *Band Math* collection and applying it to the multispectral image after running atmospheric correction with *Radiometric Calibration* tool.

On the other hand, however, the strategy to generate surface temperature image relies on thermal band 10 from *Landsat-8* imagery and it takes a few more steps: first, the procedure to get rid of atmospheric influence for thermal is different from multispectral image as it is based on the following formula:

$$Atmospherically\ Corrected\ Radiance\ =\ \frac{Radiance - Lu}{\varepsilon \times \tau} - \frac{1-\varepsilon}{\varepsilon} \cdot Ld$$

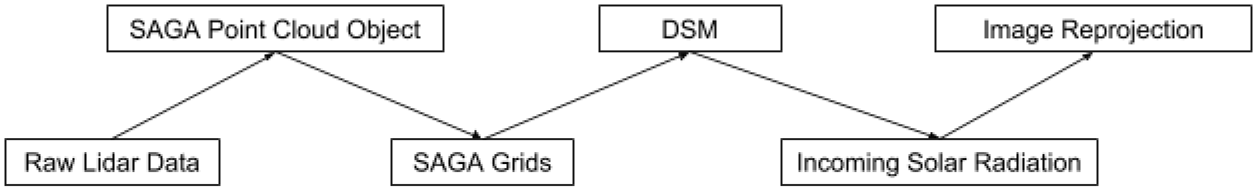
Where $Radiance$ is *TOA Radiance* which can be easily calculated by applying *Radiometric Calibration* tool to thermal band, Lu is the upwelling radiance or atmospheric path radiance, Ld is the downwelling radiance or sky radiance, ε is the emissivity of the surface (in this case ε = 1), τ is the atmospheric transmission or transmittance. Among these parameters, Lu, Ld and τ can be calculated via NASA Atmospheric Correction Parameter Calculator by specifying the date and location of sampling site; second, it still requires a little more efforts to convert radiance into surface temperature based on the following formula:

$$Surface\ Temperature\ =\ \frac{K2}{ln(\frac{K1}{L}+1)}$$

Where *K1* is the first calibration constant (in this case *K1* = 774.89 in watts/(meter squared × ster × mm)), *K2* is the second calibration constant (in this case *K2* = 1321.08 in Kelvin), *L* is the corrected spectral radiance previously calculated.

Lidar Data for Incoming Solar Radiation

The strategy to generate incoming solar radiation image is to take advantage of Lidar data and functionalities in *SAGA GIS* to generate DSM (Digital Surface Model). The Lidar data of Caumsett State Park is provided by the GTECH 734 instructor Prof. Gordon Green. The spatial resolution of final derived solar radiation image may not be the same as *Landsat-8* image, thus the new image is reprojected to 30m spatial resolution with *Arcpy* libraries. The workflow to generate incoming solar radiation image is shown below:



Data Extraction[1]

The sampling sites can be simply identified by entering the coordinates into the location box in *ENVI*. The next step is to zoom in until raster pixel becomes quite visible and then draw a small polygon around center point of the identified location to make sure the *ROI* polygon covers a single pixel on the map. Repeating this procedure until all sampling sites are identified, then the *ROI* is characterized by a group of square pixels which is used as a mask to obtain a subset of the image.

Applying this procedure to NDVI, Solar Radiation and Surface Temperature images respectively obtains three image subsets. Next step is to load three processed images and three of their subsets into *IDL* in which raster images are parsed to arrays with their pixel value stored inside. Arrays of subset images along with array converted from tick sampling result are implemented to train the model while arrays of original images are for prediction. These arrays are quite applicable to the model as they are of the same extent and spatial resolution.

Linear Regression Model

The modeling procedure is taking advantage of Python programming language which facilitates a statistical module called ***statsmodels***. We load the previously extracted training dataset with a Python module called *numpy* and construct a data frame to adopt the inputs. The conceptual linear formula is shown below:

$$Predicted\ Tick\ Population = n \times NDVI + s \times SolarRadiation + t \times Temperature + b$$

Where *n*, *s*, t are the coefficients of measured NDVI, Solar Radiation and Surface Temperature respectively, whereas *b* is the intercept of the equation. It is quite necessary to read the regression summary generated by computer after model training in order to have an understanding of the significance of each variable correlating to the observations.

[1] ***Scripts & Data:*** [***https://github.com/JohnDi0505/Tick_Modeling.git***](https://github.com/JohnDi0505/Tick_Modeling.git)

Results

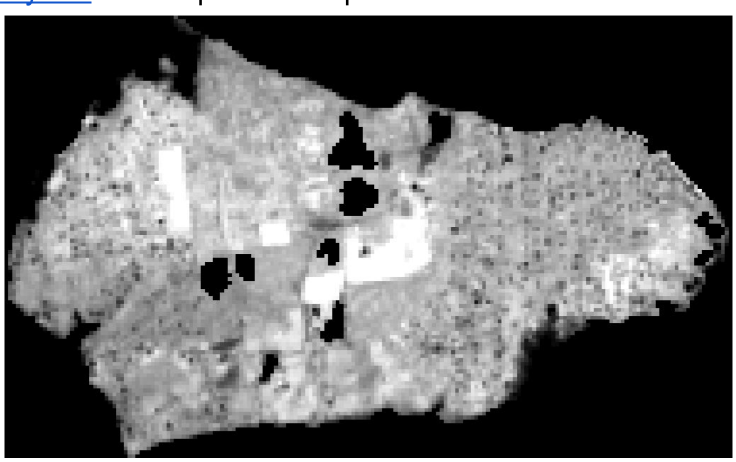
Sampling Results

The tick sampling trip was scheduled on May 12th in Caumsett State Park. A total number of 109 deer ticks were sampled and the information corresponding to each sampling site were recorded respectively (shown in the table below).

Coordinates	Number of Ticks
40.92256, -73.47001	5
40.92251, -73.47044	15
40.92251, -73.47044	9
40.92240, -73.47112	5
40.92249, -73.47151	11
40.92270, -73.47189	7
40.92272, -73.47215	5
40.92292, -73.47259	8
40.92304, -73.47270	19
40.92317, -73.47313	8
40.92339, -73.47353	17

NDVI Image Processing

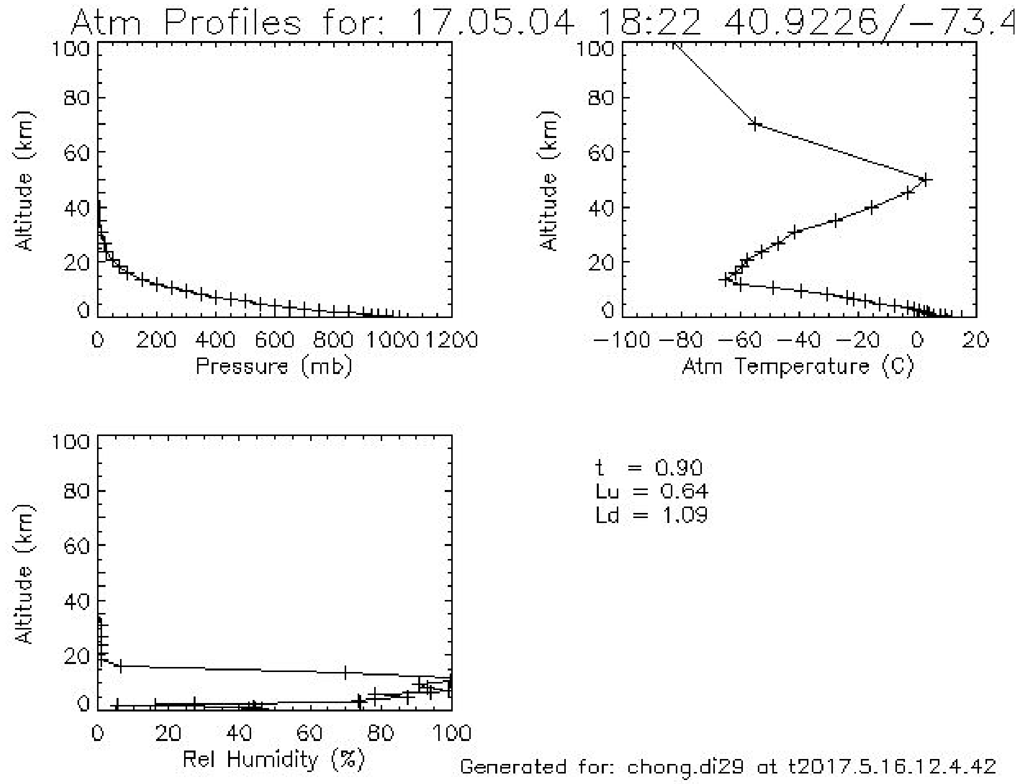
NDVI is calculated with *Spectral Indices* tool after atmospheric correction in ENVI. The raster image is also loaded to *IDL* and it is converted into a *CSV* file called Caumsett_NDVI_array.csv with its pixel value preserved.



Additionally, the sampling sites are identified in *ENVI* and with it an *ROI* is created as a mask for the extraction of NDVI training dataset. The training dataset is also converted into a *CSV* file called training_ndvi.csv.

Thermal Image Processing

After obtaining TOA radiance of the thermal band, the atmospheric correction for thermal band is based on the formula which requires specific parameters, thus the next step is to obtain atmospheric correction parameters calculated by NASA. The calculation results are shown below:



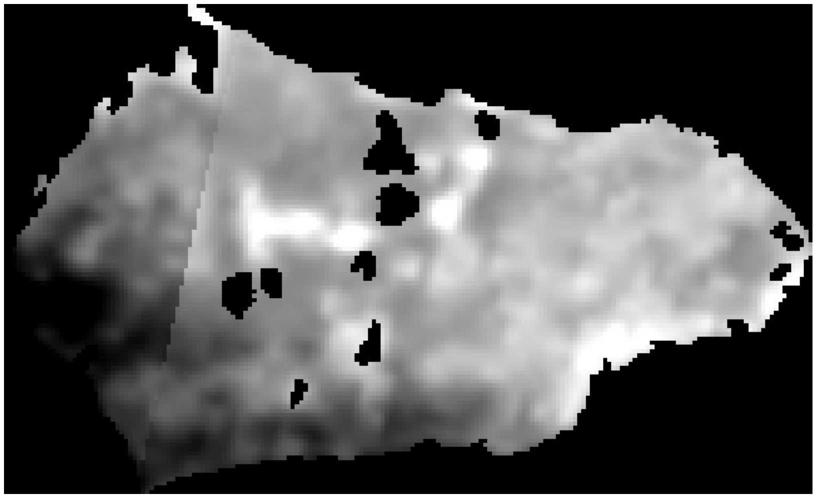
After feeding these parameters, the formula to calculate atmospherically corrected thermal band is:

$$Atmospherically\ Corrected\ Radiance\ =\ \frac{float(band\ 10)-0.64}{1\times 0.9}-\frac{1-1}{1}\cdot 1.09$$

It is also necessary to convert the surface temperature using the conversion formula introduced previously after atmospheric correction for thermal. Referring to *Landsat-8* metadata to obtain the two calibration constants ($K1$ = 774.89 watts/(meter squared × ster × mm); $K2$ = 1321.08 Kelvin) and supplying them to the formula, it is able to get:

$$Surface\ Temperature\ =\ \frac{1321.08}{ln(\frac{774.89}{float(band\ 10)}+1)}$$

Creating the formula and applying it to the thermal band, the surface temperature image is generated. Applying the mask to the image for the training dataset and loading these two in *IDL*, two *CSV* files [Caumsett_temperature_array.csv] and [training_temperature.csv] are generated.



Lidar Data Processing

The process to derive an incoming solar radiation requires *SAGA* command line functions. The pipeline is shown below with sample scripts:

1. Create saga point clouds (Module Name: "Import Point Cloud from Text File")
   - `saga_cmd -f=s io_shapes 16 -POINTS={saga point cloud} -FILE={input las file} -FIELDS='4;5' -FIELDNAMES='intensity;classification' -FIELDTYPES="2;2" -FIELDSEP=2`
2. Create grid for point cloud (Module Name: "Point Cloud to Grid")
   - `saga_cmd -f=s pointcloud_tools 4 -POINTS={saga point cloud} -OUTPUT=0 -GRID={saga grid} -AGGREGATION=3 -CELLSIZE=1.000000`
3. Interpolate grid to obtain digital surface model (Module Name: "Close Gaps")
   - `saga_cmd -f=s grid_tools 7 -INPUT={saga grid} -RESULT={digital surface model} -THRESHOLD=0.100000`
4. Calculate solar radiation (Module Name: "Potential Incoming Solar Radiation"), note that the incoming solar radiation model requires a time interval as its parameter. The time time period starts from 00:00 am on May 4th, 2017 and ends at 14:00 pm May 4th, 2017 when the sampling activity finished.
   - `saga_cmd -f=s ta_lighting 2 -GRD_DEM={digital surface model} -GRD_DIRECT={optional output} -GRD_DIFFUS={optional output} -GRD_TOTAL={incoming solar radiation} -DAY_A={start date} -MON_A={start month} -DAY_B={end date} -MON_B={end month} -DDAYS=1 -HOUR_RANGE_MIN=0.000000 -HOUR_RANGE_MAX=14.000000 -METHOD=2`

After reprojecting the solar radiation image into 30m spatial resolution using *Arcpy*, the solar radiation map is obtained:

Playing the same trick in *IDL* to this image, two *CSV* files Caumsett_Solar_array.csv and training_solar.csv were created.

In addition to environmental dataset, tick numbers are entered into cells corresponding to positions of sampling sites in an array called training_tick.csv which has the same dimension and extent as the other training dataset.

Linear Regression Model

The 4 training dataset are loaded as 4 two dimensional array and subsequently they are compiled into a data frame using a *Python* module called *Pandas* as it is straightforward to normalize the inputs and applicable to the regression module (**left.** Raw inputs from training dataset; **right.** Normalized data frame.):

	NDVI	solar	thermal	ticks
0	0.424539	0.464145	287.130829	17.0
1	0.387581	2.076295	287.174561	8.0
2	0.381647	3.399249	287.171661	19.0
3	0.360848	1.976942	287.168732	8.0
4	0.375619	3.495364	287.122070	5.0
5	0.427175	1.231600	287.031555	7.0
6	0.409164	6.177731	286.870728	11.0
7	0.414155	1.294208	287.529877	15.0
8	0.428450	0.621537	288.057770	5.0
9	0.369267	0.462883	286.905853	5.0
10	0.383062	1.926588	287.069489	9.0

	NDVI	solar	thermal	ticks
0	1.130664	-0.957836	-0.217824	1.421286
1	-0.359690	-0.015270	-0.085854	-0.382654
2	-0.598982	0.758213	-0.094606	1.822162
3	-1.437714	-0.073358	-0.103444	-0.382654
4	-0.842065	0.814408	-0.244256	-0.983967
5	1.236962	-0.509133	-0.517402	-0.583092
6	0.510658	2.382692	-1.002728	0.218659
7	0.711923	-0.472528	0.986379	1.020411
8	1.288377	-0.865815	2.579395	-0.983967
9	-1.098213	-0.958574	-0.896731	-0.983967
10	-0.541921	-0.102798	-0.402929	-0.182216

The linear regression module can be directly applied to the normalized data frame running the following line in *Python*:

```
model = ols("ticks ~ NDVI + solar + thermal + NDVI * solar + NDVI * thermal + solar * thermal", df).fit()
```

The expression inside the function is similar to the linear formula previously mentioned except for three parts: “NDVI * solar”, “NDVI * thermal” and “solar * thermal”. These 3 combinations are added as variables into the model pairwise in order to test the significance pairwise. The model generates a summary of the testing results for each variable respectively:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  ticks   R-squared:                       0.365
Model:                            OLS   Adj. R-squared:                 -0.587
Method:                 Least Squares   F-statistic:                    0.3836
Date:                Tue, 23 May 2017   Prob (F-statistic):              0.859
Time:                        19:50:35   Log-Likelihood:                -12.585
No. Observations:                  11   AIC:                             39.17
Df Residuals:                       4   BIC:                             41.95
Df Model:                           6
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
Intercept         0.6307      0.753      0.837      0.450      -1.461       2.723
NDVI              0.5255      0.949      0.554      0.609      -2.108       3.159
solar             0.7735      1.422      0.544      0.615      -3.175       4.722
thermal           1.1651      1.364      0.854      0.441      -2.623       4.953
NDVI:solar        0.3577      1.030      0.347      0.746      -2.503       3.218
NDVI:thermal     -0.5847      1.142     -0.512      0.636      -3.755       2.586
solar:thermal     0.9481      1.656      0.572      0.598      -3.650       5.546
==============================================================================
Omnibus:                        0.100   Durbin-Watson:                   2.007
Prob(Omnibus):                  0.951   Jarque-Bera (JB):                0.214
Skew:                           0.167   Prob(JB):                        0.899
Kurtosis:                       2.405   Cond. No.                         11.3
==============================================================================
```

Discussion & Conclusion

As the regression summary illustrates, all *p-values* to each variable is very high comparing with the *p-value* threshold of 0.05, which means the likelihood to reject the coefficients generated by the regression model is high. Thus the model cannot be applied to a large-scale dataset for prediction due to the low significance. However, by comparing relative *p-values*, temperature (“thermal” in summary table) has the lowest *p-value* and it implies temperature may play a relatively more significant role in regulating tick population than other variables. As a lot of literatures are pointing out that temperature is one of the fundamental variables to tick habitat, it might be an interesting topic to find out a more detailed and reliable correlation between temperature and tick population.

Limitations & Future Research

The poor results are indicating several limitations throughout the entire in this project:

- The tick sampling trip (May 12th, 2017) was not synchronized as the date (May 4th, 2017) of *Land-8* image due to some reasons: first, the weather on both May 4th and May 12th were rainy and ticks barely come out for questing on rainy days; second, the schedule of team members was tedious to coordinate and this resulted in the unsynchronized scheduling of sampling date, otherwise the efficiency of sampling would have decreased significantly.
- The amount of data is extremely limited as there are only 11 sampling sites involved in the regression model. Tick sampling proved to be a time consuming and labor intensive work as it took approximately 4 hours for 6 people to cover the a total of 11 sampling sites. However, this limitation can be solved by more sampling data added to the data repository in further research.
- **(Del Fabbro 2015)** also used a linear regression model to predict tick population. This research employed more than a dozen variables thus it had an additional procedure called *feature selection* where the significance corresponding each variable is ranked by *p-value* and the insignificant variable with *p-value* greater than the threshold of 0.05 was removed. In stark contrast, only three variables are involved in this project. Therefore, if more variables are considered in further research, it is possible to conduct the *feature selection* procedure to avoid overfitting and increase accuracy of regression model.
- Future research could still be focusing on the role that environmental variables are playing. What should be noted is that the significance of temperature has a potential for further testing. If this significance is identified, it would be possible to study tick population at large-scale. The pattern of tick developments of habitats may indicate the pattern of climate change.

Acknowledgement

Sampling Team Members in *Prof. Weigang Qiu's Lab*, *Department of Biology*, *Hunter College*

Prof. Gordon Green, Instructor of *GTECH 734 Geo Web Services*, *Department of Geography*, *Hunter College*

Prof. Wenge Ni-Meister, Instructor of *GTECH 712 Remote Sensing of Environmental*, *Department of Geography*, *Hunter College*

Landsat-8 image from [*USGS Earth Explorer*]

Reference

Brownstein, John S., Theodore R. Holford, and Durland Fish. “Effect of Climate Change on Lyme Disease Risk in North America.” *EcoHealth* 2, no. 1 (March 2005): 38–46. doi:10.1007/s10393-004-0139-x.

Lindgren, E, L Tälleklint, and T Polfeldt. “Impact of Climatic Change on the Northern Latitude Limit and Population Density of the Disease-Transmitting European Tick Ixodes Ricinus.” *Environmental Health Perspectives* 108, no. 2 (February 2000): 119–23.

Gabriele-Rivet, Vanessa, Julie Arsenault, Jacqueline Badcock, Angela Cheng, Jim Edsall, Jim Goltz, Joe Kennedy, L. Robbin Lindsay, Yann Pelcat, and Nicholas H. Ogden. “Different Ecological Niches for Ticks of Public Health Significance in Canada.” *PLOS ONE* 10, no. 7 (July 1, 2015): e0131282. doi:10.1371/journal.pone.0131282.

Del Fabbro, Simone, Sergio Gollino, Michel Zuliani, and Francesco Nazzi. “Investigating the Relationship between Environmental Factors and Tick Abundance in a Small, Highly Heterogeneous Region.” *Journal of Vector Ecology* 40, no. 1 (June 1, 2015): 107–16. doi:10.1111/jvec.12138.

Roy, D. P., M. A. Wulder, T. R. Loveland, Woodcock C.E., R. G. Allen, M. C. Anderson, D. Helder, et al. “Landsat-8: Science and Product Vision for Terrestrial Global Change Research.” *Remote Sensing of Environment* 145 (April 5, 2014): 154–72. doi:10.1016/j.rse.2014.02.001.