# Analysis on the Correlation between Annual Vegetable Health Index and Agricultural Production

## Summary

We decided to take the analysis on the correlation between vegetable health index (VHI) and agricultural production with the country South Korea. We use a simple linear regression model to train the data and finally find that the health of the crops near Seoul and Busan could have greater impact on the yearly production value.

## Introduction

The vegetable health index could definitely contribute the impact to the annual agricultural production. However, due to the complex environment, global market, ethnic culture and different regional government policies, it is quite difficult to simply discover the correlation between them. Thus, we decided to study this type of correlation with South Korea (SK) for the following reasons:

1. SK government carries out relatively consistent polices on different regions in this country. This could help maximally decrease the impact of regional differences.
2. Most of the rice produced by SK feeds itself every year. This could help maximally decrease the impact of export / import market.
3. SK has only one nation in the country. This could help maximally decrease the impact of culture.

With all the consideration above, we decided to study the correlation between the yearly VHI and the rice production for SK. Although we have weekly VHI data from the start to end of every year, we decided to use the VHI data from the 19th week to the 42th week because in this duration it is the time for growing the crops.

## Approaches

We study the correlation based on the 2 hypothesis:

1. All the locations’ VHI could explicitly contribute to the annual rice production in the linear composition way. We need to find the ranking list of these locations by their contributions to the production.
2. The VHI data of the whole country is latently correlated with the production data. Using the past years’ both VHI data and production data could predict the future production data by future VHI data.

Based on the two hypothesis, we designed the linear model for the analysis:

For which, represents the yearly rice production data vector, represents the VHI values of SK’s different locations. Our task is to find the correct weights vector which could make the fit .

To implement this approach, we need to

1. Download the VHI images from NOAA and production data from FAO.
2. Extract the VHI data of SK from the VHI images.
3. Build and train the linear regression model.
4. Evaluate the result.

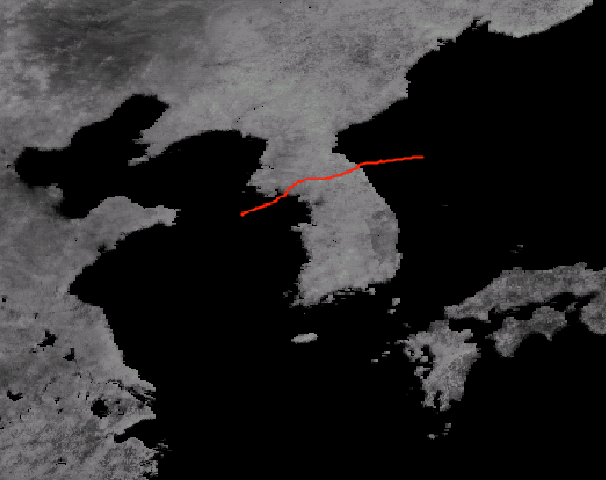
### Download Data

Downloading the data from the objective websites is not complicated but time consuming. It cost nearly 4 weeks for us to download all the VHI images which has size of 270GB.

### Extract VHI Data from VHI Images

This is the actually the most complex step during our experiments. The download images do not tell us which part of the pixels refer to SK. We have to use some programmatic tricks to find out all the pixels that belong to SK in the VHI images. When examined several VHI images, we found that all of them are in the same size. Thus, if we find the indexes of the pixels in one image, we could find the corresponding ones in all.

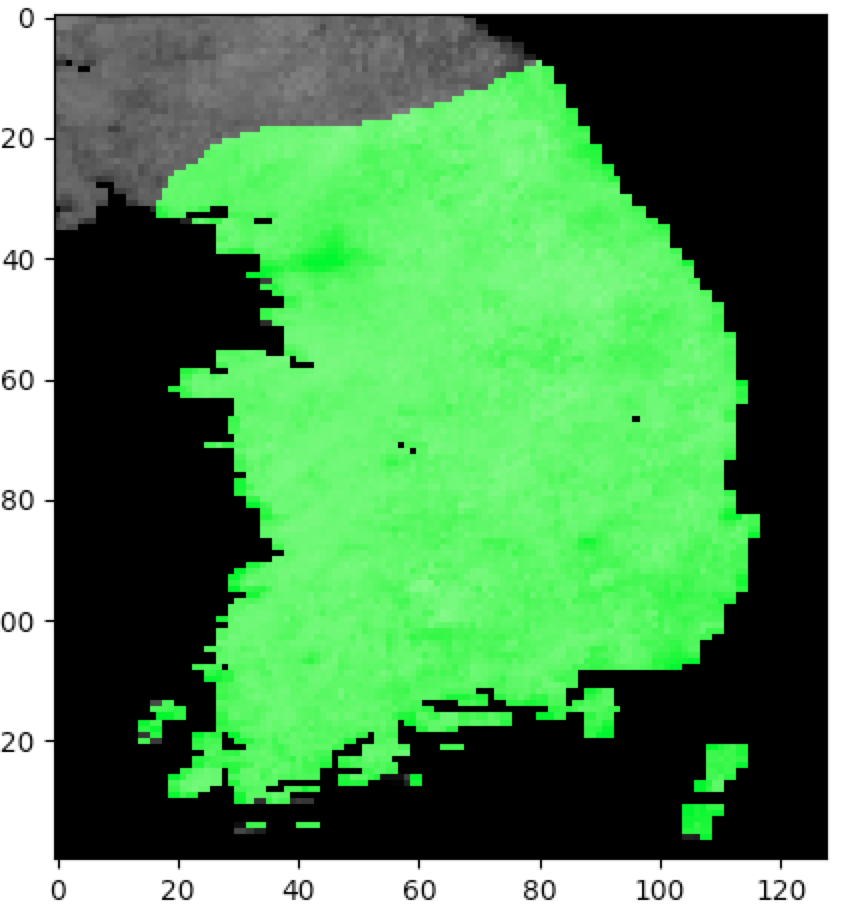
Firstly, we picked one VHI image and used some painting tools such as windows paint to clarify the border of SK. Actually, we only needed to mark the border between South Korea and North Korea. Notice that, the VHI image is only filled in black, white and grey. We must draw the border in another independent color channel such as red. The border drawn is shown as following figure:



Secondly, we found that the VHI values which fall in the sea is always 0. Luckily, SK is surrounded by North Korea and sea. So we could easily set the conditions in which we could indicate the SK pixel that belongs to SK should satisfy:

1. is under the red border that we drew.
2. and the value of the pixel is greater than 0.
3. and the pixels should be located in a general rectangle which contains SK.

By the conditions, we built a python program to extract all the pixels. The code is shown in the our github project with name of index\_extraction.py. Finally, we extract 8273 pixels that belong to SK. The result is shown in the following figure.



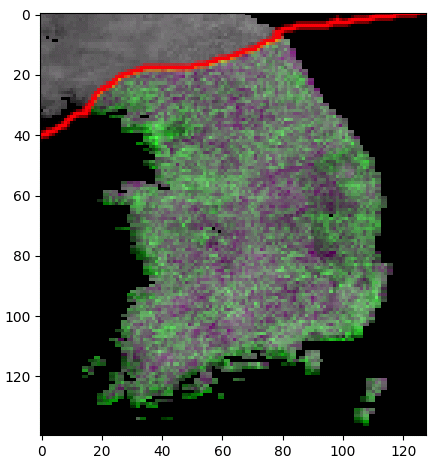
### Build and train the linear regression model

For simplicity, we use the python module sklearn to build the linear regression model. The code is shown in file analysis\_field.py which is presented in our github project.

### Evaluation

After the training process, we got the coefficient weights with length of 8327 of the linear regression model. The greater coefficient weight implies greater impact of this location for the yearly production value.

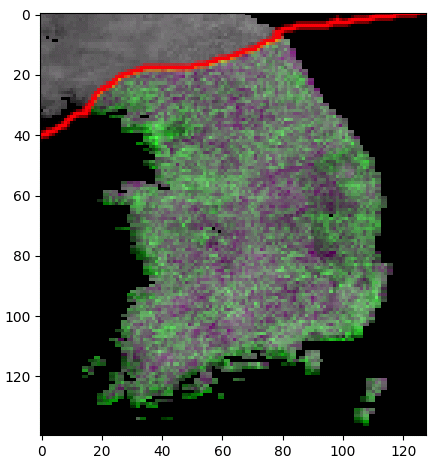
Next, we normalized the coefficient weights to (0,1), and marked them on the map of SK in green color. If the pixel has more greenness, it could have greater impact on the production value. The final output is shown in the following figure:



Let’s take a look at the figure, we could find the general phenomenon that the coastal areas of SK could have greater impact on the production value. This reveals actually the population distribution of SK.

Something more interesting, we also circled the two greenest areas concluded by our program. Compared with the real map of SK, we found that the left top circle is just right Seoul, which is the capital and largest city of SK. The right bottom circle is Busan, which is just right the second largest city of SK.

A possible explanation for this result is that, the crops near the huge city of SK could have greater impact on the yearly production sales.



Busan

Seoul

### Experiment environment

Macbook pro 14

Pycharm 2016

Scipy 0.19.1

Scikit-learn 0.19.0

Scikit-image 0.13.0

Numpy 1.13.0