Development of Interactive Visual Dashboard in R for Geographically Weighted Regression Model to Explore and Analyse Corn Yield in the United States

TAN Le Wen Angelina | PU Yiran | Stanley Alexander DION

Abstract—Corn has become a staple in many parts of the world, providing not only food, but also act as the raw ingredient for corn ethanol, animal feed etc. The Corn Belt in the US has about 96,000,000 acres of land just for corn production, and have characteristics of leveled land, fertile and highly organic soils. Breeders have been experimenting with various types of corn hybrids, each of them specifically created to have high yield despite the environment it is planted in. Over the years, the farmers have been using trial and error method to identify the best hybrids to plant by planting each of these hybrids in different locations with different environmental factors; this process has been proven to be slow and not very effective. Hence we designed an app in R to build Geographically Weighted Regression (GWR) models to help farmers to better analyse the relationships by exploring the meteorology and geographical factors that makes a corn. We would implement **GWmodel** package from R to generate our GWR model.

Index Terms—Geographically Weighted Regression Models, GWmodel, R, Isolines Graphs, Corn Yield, Corn Plantation, Corn Belt

1 Introduction

Corn or Maize (as called in some countries) was first grown in ancient Central America. Corn has become a staple in many parts of the world, surpassing wheat or rice. The United States accounts for about 40% of production of corn in the world [1], which makes it the largest corn producer. The major portion of production is found in the Midwestern states, such as Illinois, Iowa, Nebraska and Minnesota – these states were grouped and eventually became known as the 'Corn Belt', as seen in Figure 1.

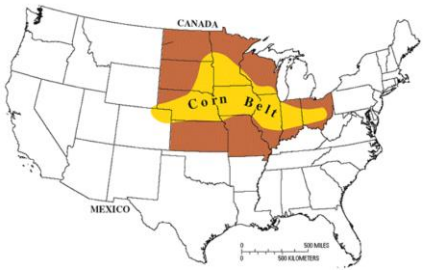


Figure 1: Corn Belt in the US

The Corn Belt has about 96,000,000 acres of land, and the states that make up the Corn Belt were selected due to levelled land, fertile and highly organic soils [2]. The growth in yield during the years 1910 to 1940 were minimal, but due to the increase interest in developing 'Hybrid Crops' from 1940 onwards, the growth in yield has been exponential [3]. Hence, breeders have been experimenting with various types of corn hybrids, each of them specifically created to have high yield despite the environment it is planted in.

1.1 Motivation

Over the years, the farmers have been using trial and error method to by planting each of these hybrids in different locations with different environmental factors; this process has been proven to be slow and not very effective [4]. This project aims to create an app using R to build **G**eographically **W**eighted **R**egression (**GWR**) Models to help farmers to better analyse the relationships by exploring the meteorology and geographical factors that makes a corn.

1.2 Scope

The scope of the project is limited to the corn produced in the US. We will analyse at the 'Environment' level (aggregated) instead of 'Hybrid' level. Corn is only grown during Summer/Autumn, hence we will only take into consideration the months that Corn is grown – this period will be termed **'Growing Period'** throughout this paper. We will also limit the weather (environmental) factors for all years provided (2008 – 2017) to the following:

- Sum and Average Precipitation,
- Sun Radiation,
- Average Temperature,
- Location (longitude and latitude) of the plantation/environment

1.3 Objective

The app aims to provide the user with the following:

- To visualise the weather patterns over the past 10 years during the growing season of corn
- To provide a user-friendly platform for people to build and visualise GWR models
- To provide a simple way to analyse the outputs of GWR models (coefficients and p values)
- To allow user to save and download GWR results for their personal use

1.4 Overview of Paper

This paper will start with a theoretical background in Section 2, where theories for GWR and Inverse Distance Weighted will be addressed. Next, Literature Review in Section 3 addresses some of current works done using GWR on corn prediction, and the visualisation tools out there publicly for corn prediction. Data preparation in Section 4 gives a summary of how the data is being processed/cleaned. The major section is Section 5, where we give a detail breakdown of how the dashboard is created. Section 6 provides with a case study to better understand how to analyse the results generated by the dashboard. Section 7 will address the conclusions, and some of the future works that can be done. Last but not least, Section 8 notes some of our learning experiences while doing this project.

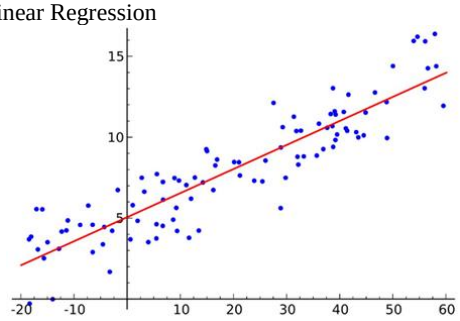
2 THEORETICAL BACKGROUND

In this section we will be delving into the theory behind the visualisation techniques and models that we use. First and most importantly, the theory behind Geographically Weighted Regression.

2.1 Geographically Weighted Regression (GWR)

The fundamental of GWR starts with Linear Regression.

2.1.1 Linear Regression



A linear regression can be summarised as fitting the best line in a cloud of points. Let us assume a dependent variable y that we want to estimate through n independent variables $= (x_j)_{1\leq j\leq n}$.
Resolving the linear regression is to find the estimator $\beta = (\beta_j)_{1\leq j\leq n}$ such that:

$$y = \beta_0 + \sum_{1\leq j\leq n} x_j\beta_j + \epsilon$$

with ϵ being the estimation error, that we wish to be minimal.
We can rewrite the above equation in Matrix form for compacity:

$$\boldsymbol{y} = \boldsymbol{x\beta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{x}$ and $\boldsymbol{\beta}$ are row and column vectors respectively.
The estimator is computed through observations, with the training data being represented by the cloud of points we want our line to pass through as accurately as possible.
If we denote $Y = (y_i)_{1\leq i\leq m}$ as the vector of our m observations of dependent variable y and $X = (x_{ij})_{1\leq i\leq m, 1\leq j\leq n}$ as the matrix of the m observations of n independent variables, we want to find the estimator β such that the error is minimal:

$$\min_{\beta} \sum_{1\leq i\leq m} (y - \hat{y})^2 = \min_{\beta} \sum_{1\leq i\leq m} (y - \boldsymbol{x\beta})^2$$

We can adopt a matrix notation where each line is an observation:

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}, \qquad X = \begin{bmatrix} (x_{1j})_{1\leq j\leq n} \\ \vdots \\ (x_{mj})_{1\leq j\leq n} \end{bmatrix}$$

and find that the estimator minimising the error is such that:

$$\boldsymbol{\beta} = (X^T X)^{-1} XY$$

More often than not the data to be analysed has an underlying spatial complexity. One could think of using the latitude and longitude of data points as variables for their regression, but most probably the dependent variable y is not linear in the coordinates. How then can we account for the location of the data points in our models?

2.1.2 Introduction of Weights

One thing we can notice from the classical linear regression is that all data points are equally considered in the regression. For instance, this means that if we want to use our estimator to evaluate the variable y near the city of Iowa, the data point of Los Angeles would have had the same influence in our estimator than the data point of Iowa.
The idea behind geographically weighted regression is that if we want to use our estimator in a specific location, the data points of areas near this location should be more influential that points far away.

The "influence" is introduced through a weight matrix W in the estimator equation:

$$\boldsymbol{\beta}(u_i, v_i) = (X^T W(u_i, v_i) X)^{-1} X W(u_i, v_i) Y$$

where (u_i, v_i) are the coordinates of the specific point i, and $W(u_i, v_i)$ is the matrix containing the geographical weights in its leading diagonal and 0 in its off-diagonal elements

$$W(u_i, v_i) = \begin{bmatrix} w_1(u_1, v_1) & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & w_n(u_i, v_i) \end{bmatrix}$$

To give an idea of what the weights do let us make the (weak) assumption that the data points are ranked by their distance to point i. Let us assume too that the weights are 1 if the distance is lower than 100 km, else is 0. We can subdivide our matrices X and Y into two submatrices $X1, Y1$ and $X2, Y2$ if they are points under 100km or not.

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}, \qquad X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix},$$

After some calculations, as shown in Annex A, our estimator can be rewritten as

$$\beta = (X_1^T X_1)^{-1} X_1 Y_1$$

We have basically filtered the points farther than 100km in our estimator, and only the points close to the specific point i are used in our estimator.

2.1.3 Different Weights

This very simple example is actually an application of a box-car weighting function of bandwidth 100km. Imagination is actually the only limit to weighting functions, though we typically acknowledge the following ones, as shown in Figure 2:

Global Model	$w_{ij} = 1$
Gaussian	$w_{ij} = \exp\left(-\frac{1}{2}\left(\frac{d_{ij}}{b}\right)^2\right)$
Exponential	$w_{ij} = \exp\left(-\frac{\lvert d_{ij}\rvert}{b}\right)$
Box-car	$w_{ij} = \begin{cases} 1 & \text{if } \lvert d_{ij}\rvert < b, \\ 0 & \text{otherwise} \end{cases}$
Bi-square	$w_{ij} = \begin{cases} (1-(d_{ij}/b)^2)^2 & \text{if } \lvert d_{ij}\rvert < b, \\ 0 & \text{otherwise} \end{cases}$
Tri-cube	$w_{ij} = \begin{cases} (1-(\lvert d_{ij}\rvert/b)^3)^3 & \text{if } \lvert d_{ij}\rvert < b, \\ 0 & \text{otherwise} \end{cases}$

Figure 2 : Six kernel functions; wij is the j-th element of the diagonal of the matrix of geographical weights W(ui, vi), and dij is the distance between observations i and j, and b is the bandwidth [5].

Without necessarily completely cutting off all data after a certain distance, weighting function can decrease the influence of data points the further they are from the point of interest.

2.1.4 Few performance remarks

We should keep in mind a few elements when using GWR. First it is much more computationally demanding as we have to calculate our estimator for each spatial point, and inverting matrix $(X_1^T X_1)$ can be a long task if we have numerous points. Secondly as some points are "filtered" in the regression if they are far from our point of interest i, we should make sure that we have enough data points around i to have a correct regression there. GWR requires thus more data than simple linear regression.
Finally the model adds a layer of complexity through the choice of the weighting function shape and the bandwidth. These two parameters should be calibrated through an eventually lengthy process to give the best fit of the model.
The R package that we use for GWR is GWmodel, with three key functions, gw.dist ( ), gwr.basic ( ), and bw.gwr ( ). gw.dist ( ) is used to create the matrix of distances between each points. gw.basic ( ) is the main GWR function where the algorithm reside, and bw.gwr ( ) is specifically used if user select auto-bandwidth.

2.2 Inverse Distance Weighted (IDW)

The data we have available is localised, the sample points where data is collected being scattered non-uniformly across the US. For visualisation purposes, it is necessary to go through a process of interpolation to evaluate values in every spatial point. Different interpolation techniques exist, but the choice made in this study is the Inverse Distance Weighted (IDW) technique. This consists on meshing the spatial area we want to display, and evaluating the missing values by interpolating between all data points within a search radius. The interpolation is not weighted linearly, but by the normalised distance to the data points.

For a point k in space, a search radius r and data points $i \in I$, we evaluate a variable v_p at point p as

$$v_k = \frac{\sum_{i \mid d(i,k) \leq r} \frac{v_i}{d(i,k)^p}}{\sum_{i \mid d(i,k) \leq r} \frac{1}{d(i,k)^p}}$$

With $d(i,k)$ being the distance between point i and k, and p a parameter that will modify the weights and lead to different interpolations, as we can see in the two examples below.

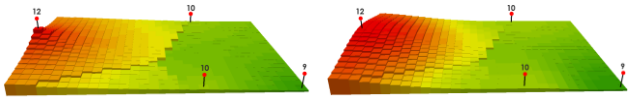


Figure 3: Examples of IDW with n=1 (left) and n=2 (right). A higher n diffuses the information further [6]

The IDW implanted in our model is from the R package called gstat, where we used the idw ( ) function. In the idw ( ) function, there are 3 methods to calculate the weight, and the method that we are using is the default Sheperd method, a slight variation of the above equation that forgo the search radius r.

3 Literature Review

Di Yang's Master thesis on spatial analysis on corn yield in the Corn Belt uses geographically weighted regression model [7]. However, Di Yang did not factor in soil properties, but mentioned in his future works that soil properties are also crucial to the yield of corn, and quoted paper from Krachenko et al..

Krachenko et al. [8] analysed the influence of soil and topography on the yield of corn in the Corn Belt region. They discovered that soil properties contribute about 30% of yield variability, with Organic Matter content influencing the yield the most. They also pointed out that elevation also contribute about 20% of yield variability. They stated that elevation is inversely correlated, with higher yields at lower landscapes. However, they did not consider the weather influence, hence we are not able to say which factor matter more. Hence, our app provides the user with a platform that not only see weather influence, but also soil influence on yield.

Veenadhari et al. [9] developed a software tool in the form of a webpage called 'Crop Advisor' with the aim to predict the influence of climatic parameters on crop yield in India. This software provides with correlations between climatic parameters with crop yield, but not considering other agro-input parameters such as soil conditions, elevation and irrigation. Moreover, the website is not intuitive, as not everybody would understand what a decision tree is, and may not understand the output generated. Hence our app aims to reach out to all, and make it simple to understand by visualising the outputs via isoline maps. In this paper, it is also stated that they found that corn is most influenced by maximum temperature. However, in our project we will discover other factors that that may have more influence that temperature.

Developing crop simulation models are not new in the research field [7,9,10]. However, to our best understanding, there is no one good app that is able to show us an easy way to understand and analyse the results of a GWR model. Most models focus on corn prediction, however our app focuses on allowing the user to be able to analyse the correlation between weather and soil properties with the yield.

4 Data Preparation

All data preparation is done in R.

4.1 Dataset

This dataset is from the 'Syngenta Crop Challenge 2019'. The main aim of this challenge is "How will we be able to grow enough food to meet the world demand?"

We have two main datasets that were used for this project: Performance Data and Weather Data. We have 10 years' worth of data, from 2008 to 2017, and we are using all the years provided.

4.1.1 Performance Data

This is the main dataset that contains the yield of the environment/plantation, and each of them is denoted by a unique ENV_ID. It can be seen that there are two yields: 'YIELD' and 'ENV_YIELD_MEAN'. 'YIELD' is the individual hybrid yield, whereas 'ENV_YIELD_MEAN' is the average yield of all hybrids planted in that particular environment. We also have the properties of soil, such as Clay, pH, Silt etc.

The growing period of corn is derived from calculating the number of days between the planted date and the harvest date. Please note growing period this is unique to each ENV_ID, and two ENV_ID may not have the same planting date and harvesting date. It is also important to note that there may be a couple of ENV_ID that have the same longitude and latitude. This is because the soil conditions in 2008 will not be exactly the same in 2016, hence though it is the same plantation, it is considered as two separate ENV_ID. In other words, one plantation could have multiple ENV_ID, but never at the same year.

We created dummy variables for 'IRRIGATION' variable, which would indicate whether the plantation is irrigated or not, so that we are able to analyse the effect of irrigation on yield.

4.1.2 Weather Data

For the Weather Data, we have data for all 365 days for each environment. However, corn does not grow throughout the year; in fact the growing period ranges from about 120 days to 180 days [3]. Hence, we will extract the growing periods for each environment based on the variable 'PLANT DATE' and 'HARVEST DATE' from the Performance Data. For example, level of precipitation at Location A on 8th January 2013 and on 8th January 2015 would be different, hence different ENV_ID despite being at the same location.

4.2 Data Cleaning

We excluded all records that have planting date after harvesting date – this does not make sense to have planting date after harvesting date. We also excluded all data points that fall outside of the US. Next, since we are given the weather data for all 365 days for each environment, we extracted out only the growing periods of each environment.

5 Dashboard Design

Link to our app: https://stanleyadion.shinyapps.io/AmazeingCrop/

User guide: https://wiki.smu.edu.sg/18191isss608g1/img_auth.php/f/fa/User_Guide.pdf

5.1 Overview Page

There are two tabs in the Overview Page:

Background: This tab is meant to give a short introduction to our project. We listed our objectives and motivation for this project, as well as the flow of the app. This page also gives an overview of both of datasets: Performance Data and Weather Data.

View Our Sandbox: This tab gives a glance at our prepared data.

5.2 Geofacet Overview Page

The geofacet package allow us to break down the data into sub regions (state or cities for example), display them in individual sub-graph (facets and position them in a way that reflects their geographical localisation.

Geofacet is a vivid way to visualise patterns in sub regions because it arranges the position of each individual sub-graph according to its geo-location in reality. As it can be seen, there are differences between states in terms of precipitation and weather.

The geofacet page aims to give the user an overall view of the weather patterns of each state that has corn plantations. The first tab (Climate Labels) shows the precipitation and temperature in barcharts, and the second tab (Climate Time Series) show the precipitation and temperature in linegraphs. Figure 4 shows an example of geofacet with Time Series.

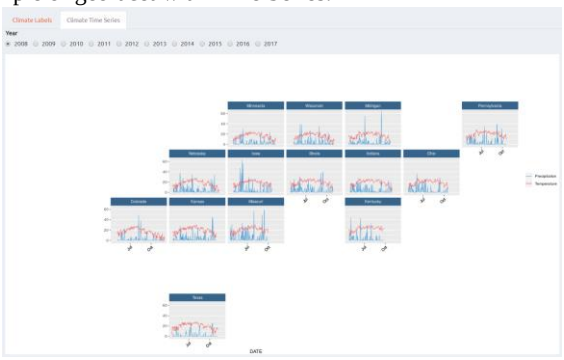


Figure 4: An example of Geofacet with Time Series

We labelled each environment using both Precipitation and Temperature data. An example is shown in Figure 5. The logic behind the labels is as follows:

For precipitation: If the 75th percentile of the individual environment is less than the 75th percentile of that year, then we label as 'LOW'. The rest would be labelled 'HIGH'

For temperature: If the 75th percentile of the individual environment is less than the 50th percentile of that year, then label as 'LOW'. If the 50th percentile of the individual environment is less than the 75th percentile of that year, then label as 'HIGH'. Those that fall in between are labelled as 'MEDIUM'.

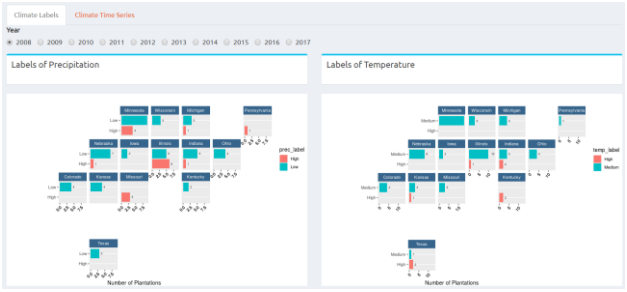


Figure 5: An example of Geofacet with Climate Labels

5.3 Climate Isoline Map Page

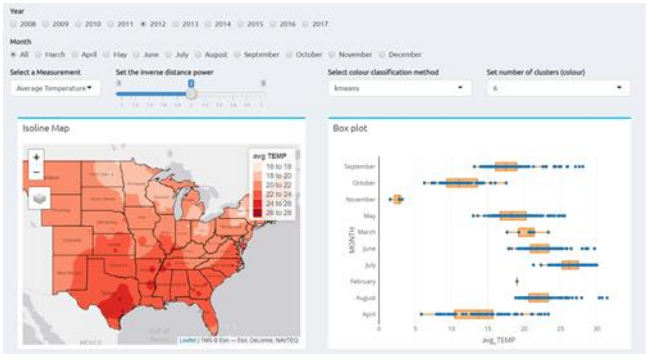


Figure 6: Example of our Climate Isoline Map Page

This page aims to give user a glimpse into the past historical weather patterns across the boundaries of states of any of the plantations that user is interested in. Figure 6 shows an example of the interface layout for this page. The user will be able to toggle among the different years, and months. Please take note that the weather data only shows data that falls within the growing season of corn in that particular environment. In the dropdown list, user will be able to select between Precipitation, Temperature and Radiation. Only Precipitation has an extra option 'SUM', as total rainfall is a common measurement of weather; it is illogical to have 'SUM' Temperature or Radiation, hence only the two options of 'AVERAGE' and 'VARIATION'.

The user is able to set the inverse distance power, p. We set it to be a slider with a step of 0.2 instead of a dropdown list so that user will be able to see the gradual changing effect of increasing the power (it will be a very long dropdown list). Since the optimal power according to Shepard [11] is 2, we set the range to be from 1.0 to 3.0 with a step of 0.2.

The user will also be able to select the method for colour classification, the methods available are kmeans, quantile and hclust. The number of classes of colour can also be set by the user, which is using the 'Set the number of clusters' selection.

Kmeans: aims to partition the points into k clusters (where k is set by user) such that the sum of squares from points to the assigned cluster centres is minimised [12].

Quantile: produces sample quantiles by segmentation, and the number of segments is indicated by the number of clusters [13].

Hclust: perform hierarchical cluster analysis using a set of dissimilarities for the n data points that are being clustered [14].

The boxplot provides more details-on-demand for the user. The user can highlight the 'outliers' for each month and the corresponding locations will be shown in the map. User can interactively explore the data by using 'Box Select', 'Lasso Select', and even download the image as a png file, and these interactive features are from plotly.

5.4 GWR Model Page

This page is about our Geographically-weighted Regression model. There are 4 tabs that will guide the user step by step to calibrate the GWR model to what the user wants. Our GWR model is solely built to have the yield as the dependent variable. We have discussed earlier that most of the plantations are planted once in the period of 10 years, it would not be sensible to take all 10 years to do a single GWR model as most plantations do not repeat. Hence we did cross-sectional analysis, where we took each year as one GWR model.

5.4.1 Model Input Data

In this tab, the user will be able to select the year, and the type of aggregation mode that user wants the yield to be. This selection will be carried forward to the other tabs. The user will also be able to see the description of the variables that are available for selection.

5.4.2 Variable Transformation

In this tab, the user will be able to see the distribution of the variables via histogram. Based on the histogram, the user can determine which variable requires transformation. This is a crucial step in calibrating any regression model. Figure 7 shows an example of the actual distribution of Elevation in 2008, which is right skewed. Hence we performed a squareroot transformation, and after which the distribution looks like a normal distribution.

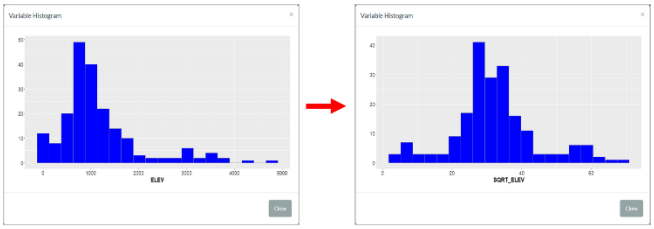


Figure 7: Before and After Transformation of variable ELEV (Elevation)

There are many reasons for performing transformation, and one reason is to reduce skewness as a distribution that is (nearly) symmetric is often easier to handle and interpret than a skewed distribution. To reduce right skewness, we mainly take roots or logarithm, and to reduce left skewness, we mainly take squares or cubes (powers). The transformation made on the selected variables will be carried forth to the next tab.

5.4.3 Variable Selection

With this tab, the user will be able to select and perform collinearity of the variables that user has chosen. This is also a vital step to take before calibrating the GWR model. This ensures that the input independent variables are not correlated. A principle danger with correlated input variables is overfitting of our GWR models. The best regression models are those with independent variables that are minimally correlated, but strongly correlated with the dependent variable, which in our case is yield.

The user will be able to 'Include' and 'Exclude' the variables, and the most interesting function of this tab is the correlation plot using corrplot package in R – this function will plot a correlation plot of the independent variables with all the variables that the user has selected. As shown in Figure 8, correlations are displayed with ellipses, actual correlation values and red-blue diverging colour. The direction and colour of the ellipse shows whether two variables are positively or negatively correlated and how strongly they are correlated: red denotes negative correlation, and blue denotes positive correlation. For the actual values, any absolute value that is greater than 0.8 would be deemed strongly correlated, and the user should drop one of the variable to ensure non-collinearity among all variables. For this given example, the 'meanYIELD' (dependent variable) is plotted against all the properties of soil that were provided in the data. We can see that 'Sand', 'KSAT' and 'AWC' are strongly correlated (both positively and negatively) with a couple of variables, hence these three should be dropped. Figure 8 also show the 'new' plot after removal of the highly correlated variables:

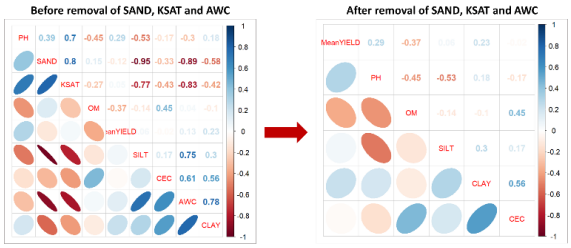


Figure 8: Correlation Plots before and after removal of highly correlated variables

Whatever that the user has selected in this tab will be carried forward to the final tab: to generate GWR model with the selected variables. From Figure 8, Krachenko's claim that OM (Organic Matter) correlates the most with the yield is supported. Among all the properties of soil, OM has the largest correlation with yield. However, this is not the only insight that we can derive from this correlation plot: OM is the most negatively correlated with yield, pH is the most positively correlated with yield.

5.4.4 GWR Model Calibration

It is important to note that the output of any GWR model are three things as listed:

i. The predicted value for dependent variable y, yield
ii. The coefficient of the selected independent variables
iii. The corresponding t statistics of the coefficient, which the algorithm will convert to p values

The y values generated are the predicted y values based on the model, and not the actual y values. The model Rsquare value is also being displayed. The higher the Rsquare value, the closer the predicted y values are to the actual y values. The user will also be able to set the bandwidth and the kernel for the GWR model, as shown in Figure 9.

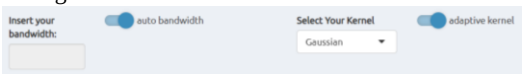


Figure 9: Bandwidth and Kernel Selection

Bandwidth: The auto bandwidth is generated by the algorithm, but the user is able to input if user has a value in mind.

Kernels: The user will also be able to choose among all the different kernels, namely Gaussian, Exponential, Bisquare, Tricube and Boxcar as shown in Figure 2. The global model is not included in the dropdown list as this means equal weightage to every point, which would translate to one single coefficient to one variable, hence the graph would only show one uniform colour. Adaptive kernel ensures all observations have the same number of points to generate a local y_i. For example, if the user input 20 into bandwidth with adaptive kernel selected, this means that each observation will have 20 nearest neighbouring points to generate a y_i. If the adaptive kernel is off, this means that each observation will have a fixed 20 KM radius to generate a y_i. The non-adaptive kernel would be slightly unfair for plantations found in Florida (very sparse plantations) as compared to Minnesota (many plantations).

Figure 10: An Example of a Coefficient map for one variable of a GWR model

Figure 10 shows an example of a GWR model for 2012, where the left-hand side shows the estimate value, and the right-hand side shows the corresponding p values. The user will be able to toggle to see the coefficient maps and the p-values of the selected variables. The user will also be able to download the data for the calibrated GWR model under the 'Data Output' tab.

We also provide the user with the results for the Global model if user wants to use it for reference. Figure 11 shows an example of how the results output is for the Global model. The estimates and corresponding p values are also given, but since this is a global (linear regression) model, there is only one estimate value per variable. The * circled in red shows the significance of the estimates to the dependent variable y. For this example, it is clear that 'IRRIGATION_IRR' is the most significant variable affecting the yield, and since the estimate is a positive value, it has the strongest positive correlation to the yield.

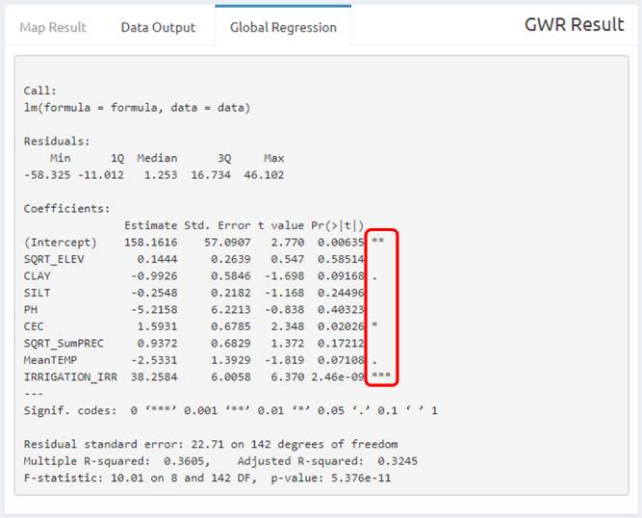


Figure 11: An Example of results for Global GWR Model

We will discuss in further details on the analysis of the results in the next section.

6 GWR Model Analysis

Global estimates may prove to be informative for climate migration, but it is definitely misleading for localised anlaysis, particularly those aimed at former adaption [15]. For this section, we will be presenting a case study. Since this app allows user to select any combination of variables which will generate different GWR models, it is difficult to present a 'generic' model. Hence, For this case study, we will focus on 2012 as there was a severe drought in the US, as well as these variables to generate our model for illustration:

i. Squareroot-transformed ELEV
ii. Clay
iii. Silt
iv. pH
v. CEC (**C**ation **E**xchange **C**apacity)
vi. Squareroot-tranformed SumPREC
vii. meanTEMP
viii. meanSRAD
ix. IRRIGATION_IRR (flag variable for normal irrigation)

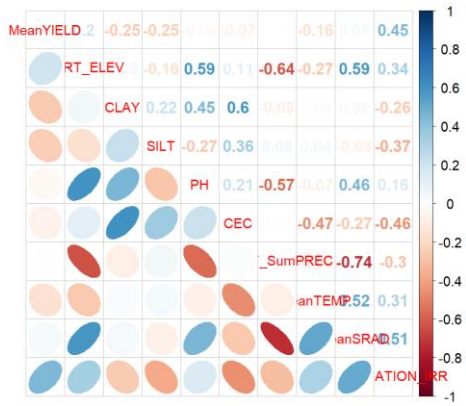


Figure 12: Correlation Plot of Chosen Variables for Case Study (2012)

From Figure 12 we can see that as per what literature mentioned, soil properties do have some influence on the yield.

6.1 Case Study: 2012 (Drought)

For our case study, we decided to pick one GWR model for illustration: 2012 as there was a severe drought in the US in 2012. The map can only show the coefficient of one variable at a time, we chose two variables to illustrate how to analyse the coefficients, we chose 'SumPREC' and 'IRRIGATION_IRR. Figure 13 shows the actual distribution of sum of precipitation in 2012. It can be seen that the circled area has relatively lesser precipitation as compared to the other regions based on the intensity of blue – the darker the shade, the more precipitation, and vice versa.

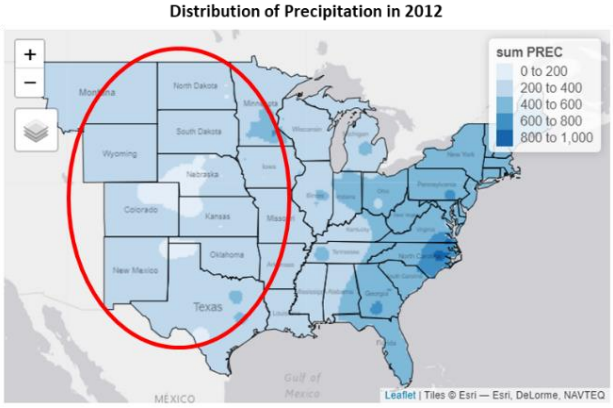


Figure 13: Distribution of Sum of Precipitation in 2012

Hence since rain is scarce in 2012, whatever rainfall that this region within the red circle receives will be vital for the survival of corn. This is supported by the coefficients generated from the GWR model, where the coefficient for variable 'sumPREC' is larger for regions within the red circle, which means stronger (positive) correlation of precipitation at regions with low rainfall with y_i as shown in Figure 14.

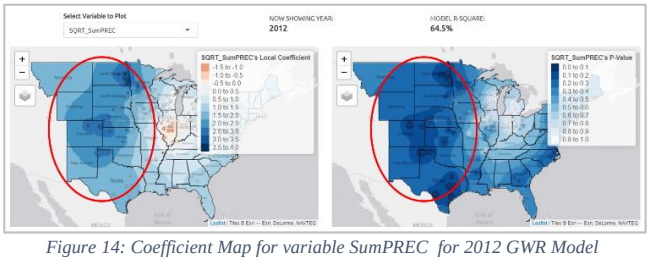


Figure 14: Coefficient Map for variable SumPREC for 2012 GWR Model

The 'IRRGATION_IRR' variable further supports this analysis. Irrigation for regions within the red circle have a stronger (and positive) correlation with y_i. Since rainfall is low, farmers will depend on irrigation to water the corn, hence areas with low rainfall will have stronger (and positive) correlation with irrigation. This is proven with Figure 15.

The p values are most significant at regions within the red circle; hence based on a 5% significance level, we are 95% confident those estimates within red circle are significant.

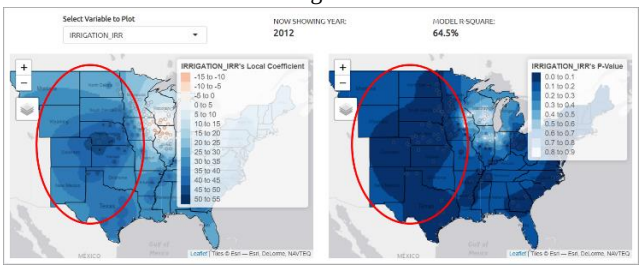


Figure 15: Coefficient Map for variable IRRIGATION_IRR for 2012 GWR Model

Figure 16 shows the results/output for the Global Model for our case study. With similar analysis with Figure 10, it can be seen that 'IRRIGATION_IRR' is the most significant variable due to ***. Similarity, since coefficient for 'IRRIGATION_IRR' is positive, it is positively correlated to yield (circled in red). In other words, an increase in IRRIGATION_IRR would generate the largest increase in yield as compared to other independent variables.

Likewise, 'MeanTEMP' has the most negative coefficient, which means that an increase in 'MeanTEMP' would generate the largest decrease in yield as compared to other independent variables.

The R-squared values for the Global model is circled in yellow, and the value hovers at around 35%. This means that the predicted y values are only 30% in similarity with the actual y values, which proves that the Global Model is not the right model for crop analysis, as Global model does not take into consideration the effect of location. Hence GWR models work much better in providing the user with analysis.

Map Result | Data Output | Global Regression | GWR Result

```
Call:
lm(formula = formula, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-58.886 -11.604   1.447  17.433  43.915

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    61.7626    83.1926   0.742   0.4591
SQRT_ELEV      -0.2659     0.3686  -0.721   0.4719
CLAY           -0.9914     0.5815  -1.705   0.0904 .
SILT           -0.2743     0.2174  -1.262   0.2092
PH             -5.2468     6.1885  -0.848   0.3980
CEC             1.6634     0.6764   2.459   0.0151 *
SQRT_SumPREC    1.7269     0.8423   2.050   0.0422 *
MeanTEMP       -5.3330     2.2445  -2.376   0.0188 *
MeanSRAD        0.3903     0.2461   1.586   0.1151
IRRIGATION_IRR 37.6686     5.9856   6.293 3.69e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22.59 on 141 degrees of freedom
Multiple R-squared:  0.3717,    Adjusted R-squared:  0.3316
F-statistic: 9.268 on 9 and 141 DF,  p-value: 5.679e-11
```

Figure 16: Results of Global Model for Case Study (2012)

We also generated different GWR models using different kernels, and the coefficients for 'SumPREC' and p values are shown below in Figure 17. The first four kernels (Gaussian, Exponential, Bisquare and Tricube) have similar looking coefficients for variable 'sumPREC', except for Boxcar kernel. Based on Figure 2, it can be seen that Boxcar takes observations within the bandwidth to have the same weightage, which is similar to the global model where all observations have equal weightage, hence this smoothens out the intensity, with larger patches of uniform colour.

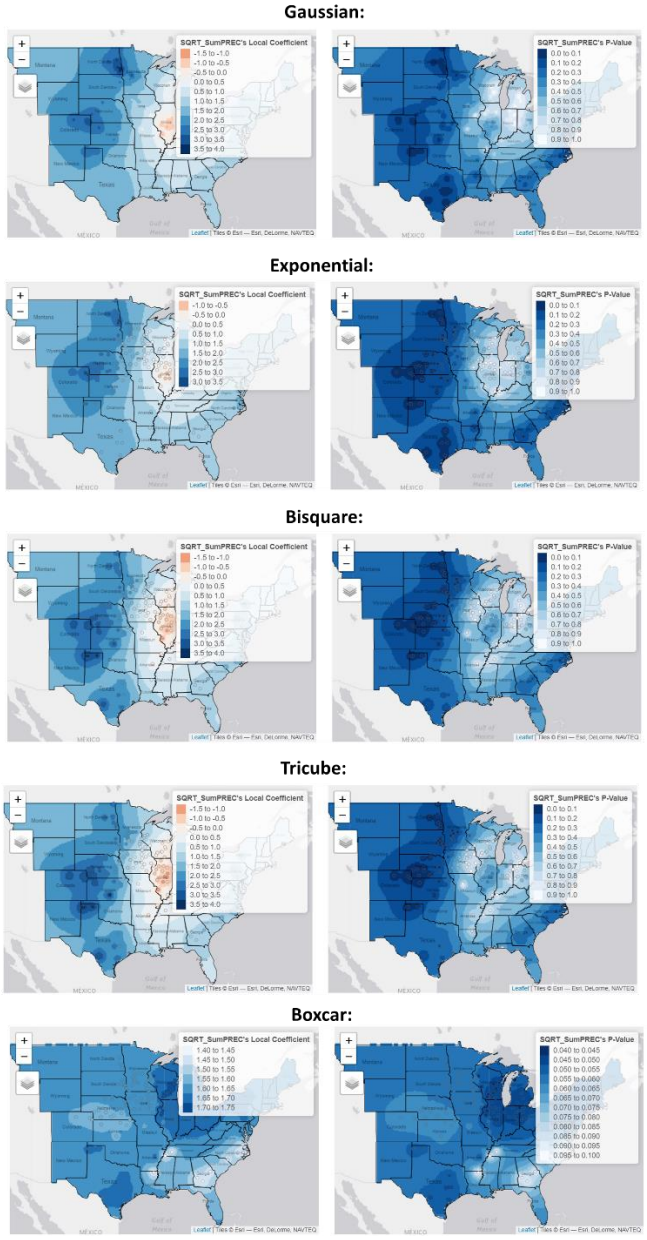


Figure 17: An example of GWR with different kernels

For this paper, we only showed two independent variables, 'SumPREC' and 'IRRIGATION_IRR', for the GWR model in 2012, but note that all 10 variables listed previously have their own respectively coefficient maps. The way to analyse the coefficients is solely based on the intensity of blue or red, depending on whether it is positively or negatively correlated respectively.

For our dataset, it is difficult to compare year on year as the plantations that were planted in one year may not be the same plantation planted in another year, hence it would not be an apple-to-apple comparison. Hence with this app, it would be more appropriate to guide user look at and analyze the independent variables affecting the yield for a particular year.

7 Future Works and Conclusion

Di Yang [7] proved that the impact of precipitation during the later stage of corn growth (reproductive state – start of growth of corn) is greater than during beginning stage of corn growth (vegetative state – growing of plant before corn). This means that precipitation is more important to have during the reproductive stage. Hence one possible future work is to segment the growing season of corn into two sections: Vegetative Stage, and Reproductive Stage. This segmentation may be able to provide deeper insights into which factors affect which growing stage of corn.

Another possible future works is to incorporate spatial multicollinearity. The variables may not be collinear with each other, but may but spatially collinear, and if not dealt with properly may lead to unreliable results [16]. Due to time constraint, we did not factor in spatial multicollinearity, which can be approached using GW Principle Component Analysis [17, 18]. GW PCA allows us to detect multivariate spatial outliers, and as well as to account for certain spatial heterogeneity.

In conclusion, corn yield is influenced by not only weather but also soil properties and presence of irrigation. The main aim of this project aims to provide user with a platform to analyse the correlations between weather and soil influence on yield, and we came up with a dashboard to fulfil this aim. This app is intuitive for user to understand the correlations, as we showed the analysis using isoline correlation maps, where the coefficients are plotted on a map, with darker shades corresponding to stronger (either positively or negatively) correlation with yield.

8 Learning Experience

The components of isoline map are actually tiny grids, the number of which is set by designer of the app. In our study, on the given US region, we created 50,000 tiny grids as the base of isoline map. It is notable that, the number of grids actually is the resolution of the displaying of the map, thus, increasing the number of grids can make the map look smoother.

Lot of learning experience is encountered when we design our application, which includes simplifying the complicated process of data preparation and calibrating the model itself. A simple example of the design thinking that we go through is to decide whether the model transform tab should come before or after the variable selection tab. In addition, small details such as the name and locations of the buttons should also be made intuitive to improve user experience.

Acknowledgments

The authors wish to thank Professor Kam Tin Seong for his guidance and help to make this project a success. The authors would also like to thank their fellow classmate Anthony Theodore for all his suggestions and help on developing the isoline graphs.

Reference

[1] Olson, R. A., & Sander, D. H. (1988). Corn production. *Corn and corn improvement*, (cornandcornimpr), 639-686.

[2] Smith, C. W. (2004). *Corn: origin, history, technology, and production* (Vol. 4). John Wiley & Sons.

[3] Shaw, R. H. (1988). Climate requirement. *Corn and corn improvement*, (cornandcornimpr), 609-638.

[4] Adee, E., Roozeboom, K., Balboa, G. R., Schlegel, A., & Ciampitti, I. A. (2016). Drought-tolerant corn hybrids yield more in drought-stressed environments with no penalty in non-stressed environments. *Frontiers in plant science*, *7*, 1534.

[5] Gollini, I., Lu, B., Charlton, M., Brunsdon, C., & Harris, P. (2013). GWmodel: an R package for exploring spatial heterogeneity using geographically weighted models. *arXiv preprint arXiv:1306.0413*.

[6] https://www.rdocumentation.org/packages/phylin/versions/1.1.1/topics/idw

[7] Yang, D. (2012). *A spatial analysis of corn and soybean yields and weather relations* (Doctoral dissertation).

[8] Kravchenko, A. N., & Bullock, D. G. (2000). Correlation of corn and soybean grain yield with topography and soil properties. *Agronomy Journal*, *92*(1), 75-83.

[9] Veenadhari, S., Misra, B., & Singh, C. D. (2014, January). Machine learning approach for forecasting crop yield based on climatic parameters. In *Computer Communication and Informatics (ICCCI), 2014 International Conference on* (pp. 1-5). IEEE.

[10] Yang, H. S., Dobermann, A., Lindquist, J. L., Walters, D. T., Arkebauer, T. J., & Cassman, K. G. (2004). Hybrid-maize—a maize simulation model that combines two crop modeling approaches. *Field Crops Research*, *87*(2-3), 131-154.

[11] Shepard, D. (1968, January). A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the 1968 23rd ACM national conference* (pp. 517-524). ACM.

[12] https://www.rdocumentation.org/packages/stats/versions/3.3/topics/kmeans

[13] https://www.rdocumentation.org/packages/stats/versions/3.1.1/topics/quantile

[14] https://www.rdocumentation.org/packages/stats/versions/3.2.1/topics/hclust

[15] Cai, R., Yu, D., & Oppenheimer, M. (2014). Estimating the spatially varying responses of corn yields to weather variations using geographically weighted panel regression. *Journal of Agricultural and Resource Economics*, 230-252.

[16] Shahid, R., & Bertazzon, S. Addressing Multicollinearity in Local Modeling of Spatially Varying Relationship using GWR.

[17] Lu, B., Harris, P., Gollini, I., Charlton, M., & Brunsdon, C. (2013). GWmodel: an R package for exploring spatial heterogeneity. *GISRUK 2013*, 3-5.

[18] Harris, P., Brunsdon, C., & Charlton, M. (2011). Geographically weighted principal components analysis. *International Journal of Geographical Information Science*, *25*(10), 1717-1736.

Annex A

Adopting the notations introduced above, let us first calculate X^TY.

$$X^TW(u_1,v_1)Y = [X_1^T \quad X_2^T] \cdot \begin{bmatrix} 1 & 0 & 0 & \\ 0 & \ddots & 0 & 0 \\ 0 & 0 & 1 & \\ & 0 & & 0 \end{bmatrix} \cdot \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}$$

$$= [X_1^T \quad X_2^T] \cdot \begin{bmatrix} Y_1 \\ 0 \end{bmatrix}$$

$$= X_1^T Y_1$$

Calculating now $X^TW(u_1,v_1)X$ we find:

$$X^TW(u_1,v_1)X$$

$$= [X_1^T \quad X_2^T] \cdot \begin{bmatrix} 1 & 0 & 0 & \\ 0 & \ddots & 0 & 0 \\ 0 & 0 & 1 & \\ & 0 & & 0 \end{bmatrix} \cdot \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$$

$$= [X_1^T \quad X_2^T] \cdot \begin{bmatrix} X_1 \\ 0 \end{bmatrix}$$

$$= X_1^T X_1$$

Using these two results we have proven that

$$(X^TW(u_1,v_1)X)^{-1}(X^TW(u_1,v_1)Y)$$

$$= (X_1^T X_1)^{-1} X_1 Y$$