Environmental Statistics

Spatial Data Analysis for Temperature in California

Under the Supervision of:

Dr Sara Osama

Dr Naira Mostafa

Name	**CHS Code**
Sara Emad Tawfik Nasim	5200298
Shahd Hesham ElSayed Hassan	5200384
Roshan Osama Mohamed Hosny	5200264

Table of Contents

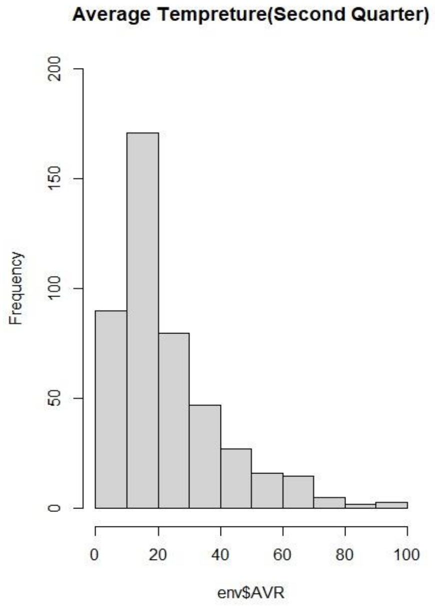
- **Introduction:**

One of the most substantial environmental factors in the natural world is temperature, affecting the whole ecosystem with all living things. Understanding the patterns and variations of environmental phenomena such as temperature across geographical regions relies heavily on spatial analysis. In this report, our main goal is to analyze the average temperature in California during the second quarter of the year using spatial analysis; We aim to explore the spatial relationships, distributions, and trends within our area of interest. The findings of the report can assist in making informed decisions related to planning, resource allocation, and policy formulation. We have taken several steps in order to reach a proper analysis; We also used the most common spatial methods and processes including (spatial interpolation and, spatial autocorrelation). The main objectives of the report are identifying the spatial autocorrelation patterns to understand the dependency structure and clustering of the average temperature values, we aim to reach different interpolated maps to draw a conclusion about the precision levels at all areas in California.

- **Exploring the Data:**

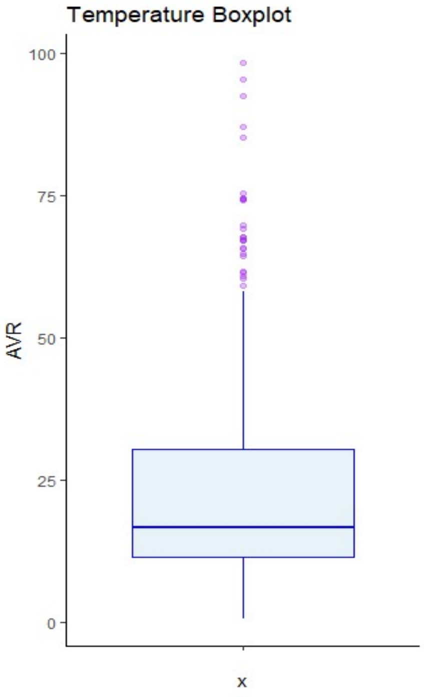
I. Graphical Methods:

- Histogram:



Histograms are a visual way to analyze and describe data distribution. Furthermore, it aids in determining the shape and central tendency of the data. This histogram depicts the distribution of average temperature in California during the second quarter; it has a skewed distribution to the right (positively skewed), so the normality assumption doesn't hold and therefore there exists outliers in the data. In addition, the graph also show that the data is unimodal.

- Box Plot:

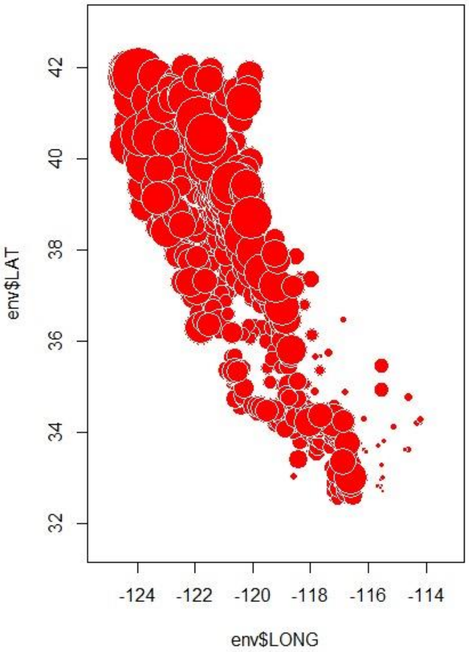


Similar to histograms, box plots are useful for data analysis, visualization, and it provides a graphical representation for the distribution of the data. Furthermore, it provides a data summary that includes the median, quartiles, and detecting outliers. This box plot displays a summary of average temperature measurements. The median is represented by the line drawn within the box and equals 16.67. It also indicates the presence of outliers, as some points are located outside the whiskers.

- Bubble Plot:

Bubble plot is a type of scatter plots that allow for the visualization of three variables, X and Y as the coordinates and Z as the variable of interest, in our case, Z is the average temperature in the second quarter of the year in California, is represented by the size of the bubble, when the average temperature increases, the size of bubble increases.

It is clear from the plot that as we move from the South-East upwards towards the North-West, the average temperature increases.



II. Geographical Weighted Regression Model (GWRM):

We’re going to study the GWR as a method for exploring our data, GWR is a non- stationary technique that models spatially varying relationships. The difference between the Global Geographical Regression and the Locally Weighted one lies mainly in the way of obtaining the regression coefficients; In Global Geographical Regression the coefficients are obtained using the OLS (Ordinary least squares) method that doesn’t take into consideration the different locations; However, in Geographical Weighted Regression the coefficients are obtained using WLS (weighted least squares) method considering the weight of each location.
So, the regression model is defined as:

$$y_i = \beta_{i0} + \sum_{k=1}^{m} \beta_{ik} x_{ik} + \varepsilon_i$$

And the coefficients are estimated using:

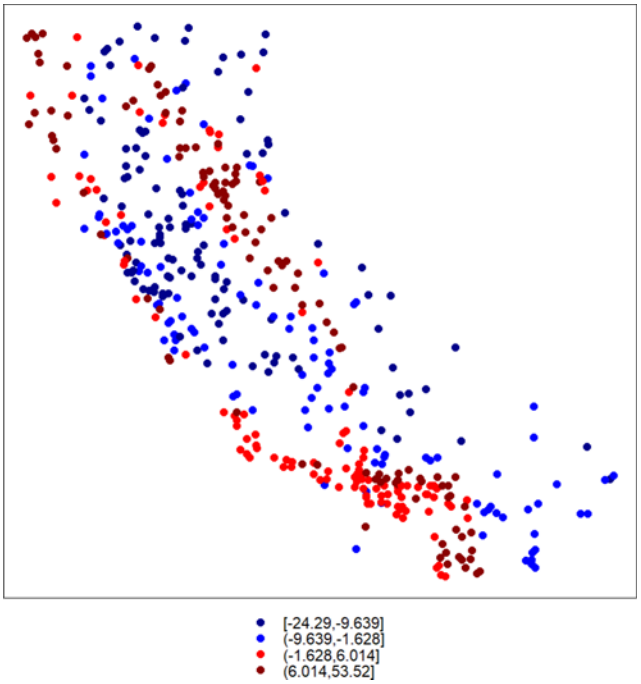
$$\hat{\beta}_i = \left(X^T W_i X\right)^{-1} X^T W_i y$$

1) For the estimation of the Global Geographical Regression:

- The fitted model was: Y_i (hat) = -106.8181 + 0.3977 LONG + 4.8000LAT

- After fitting the Global model, we obtained the residuals and plotted them:

There exists a clustering of high values of the residuals in the west centre of the map also in the south of the map exists a clustering of lesser high values; all the clusters indicating that the global geographical regression isn't the best prediction method for our data; that's why we are going to estimate the model using GWR.



2) For the estimation of the Geographical Weighted Regression:

-It's estimated using the formula given above, in which the weighting scheme is calculated by the kernel function.

-The most used Kernal function is the Gaussian Kernal and with our observed data by kernel bandwidth, we got using the adaptive bandwidth was **0.004464.**

-By increasing longitude by one unit, the average temperature is expected to range between (-9.437,8.153) holding latitude constant.

-By increasing latitude by one unit, the average temperature is expected to range between (-18.646,7.545), holding the longitude constant.

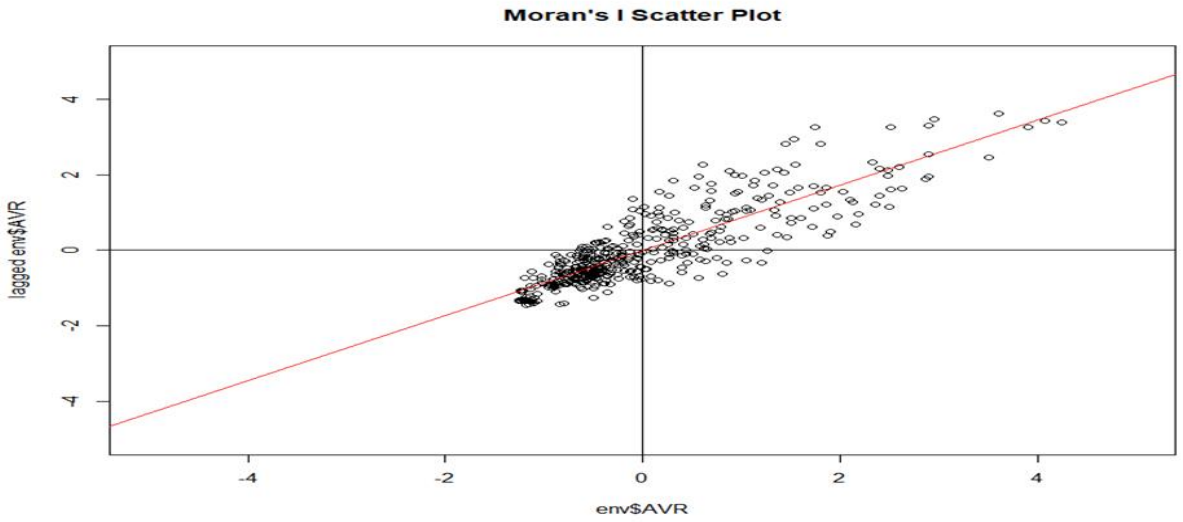
- **Spatial Autocorrelation Measures:**

Spatial autocorrelation measures assess how similar or dissimilar spatial data are in terms of average temperature across different regions of California based on their geographic proximity. It also indicates temporal spatial patterns and spatial clustering, after applying these measures, we'll be able to better understand the underlying processes that influences temperature distribution in the area under study. Moran's I and Geary's C are the most common indicators of spatial autocorrelation. In this report we'll be using Moran's I only to measure the underlying spatial autocorrelation in our data.

o Global Moran's I:

As is well known, the range of Moran's I is between (-1,1), and if the calculated value is between (0 and 1), this indicates a positive spatial autocorrelation, indicating that this location tends to have a similar value to the nearby locations so we can indicate the presence of clustering in the temperature in California. If the value falls between (-1,0), this indicates a negative spatial autocorrelation.

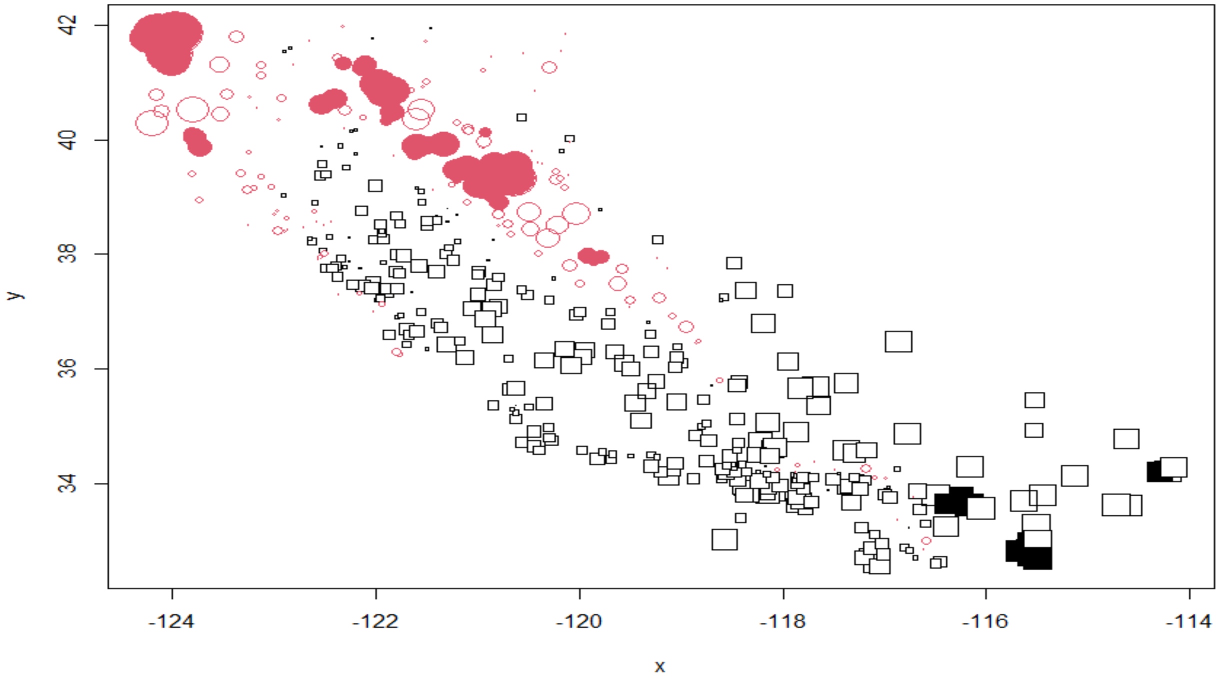
Moran's I can be measured globally and locally, the global measure only gives an indication of the autocorrelation pattern all over California. The calculated global Moran's I was equal to **0.304** and its expected value was **-0.00219.** Since the associated p-value was equal to 0 which is less than 0.05 and I>E(I), therefore, we can draw a conclusion that we have a positive spatial autocorrelation pattern(clusters) in the average temperature during the second quarter, globally.



In addition, Moran's I scatter plot can show the autocorrelation pattern. In figure, it appears that there is a positive slope indicate that similar values tend to cluster together. Moreover, it appears to have both a hot spot and a cold spot. A hotspot appears when high values of temperature tend to cluster together and there is a significantly high positive autocorrelation. While the cold spot appears in the low-low quarter of the plot where we have a very high negative autocorrelation, these areas of low values of temperature cluster together resulting in the appearance of a cold spot. There also exist some outliers in both low high and high low parts of the scatter plot. Outliers exist if a low value of temperature is surrounded by other high values, or vice versa. These outliers will probably affect the calculated value of Moran's I, therefore they must be taken into consideration.

- <u>Localized Moran's I:</u>

Because California is a very large area to study globally, we cannot provide a precise indication of the autocorrelation pattern using Global I, so we must measure spatial autocorrelation locally across different regions. To start studying the area locally, we must determine the value of the k nearest neighbours that will be taken into account while calculating I, noting that the determination of k is according to the researcher's point of view. In our case, we preferred to take the square root of the number of observations, resulting in a **<u>k=20</u>**.

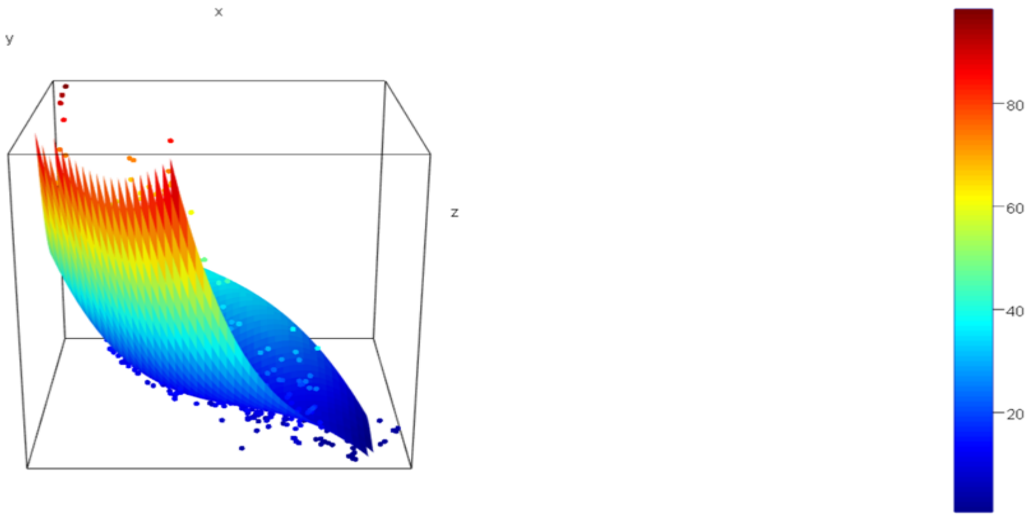


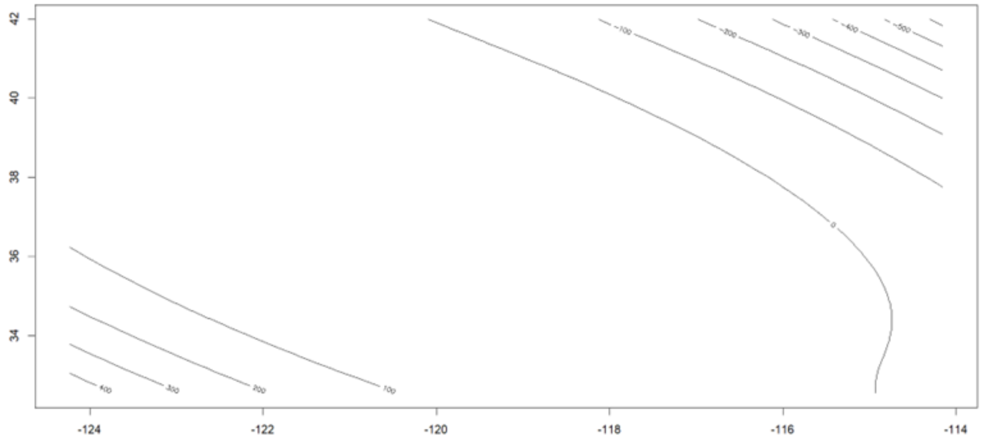
By plotting LISA, we’ll provide a visual representation of spatial clusters based on the spatial autocorrelation measures. In the above figure, spatial hot spots are represented by the pink filled circles, for illustration, we can observe that at the maximum value of latitude and minimum values of longitudes, in the North-West, we’ll be having a hotspot, where the temperature in this area is above the calculated mean of temperature. On the other hand, there exists a cold spot represented by black filled squares, in the South-East, and the values of temperature in this area is less than the mean. For the empty circles and squares, these indicate that the temperature in these areas is equal to the mean **(22.937),** attached in the appendix.

- **Trend Surface Model:**

Trend surface models are regression models having the same assumptions as regression models, they are used to estimate and visualize the underlying trends in the data. The explanatory variables are the coordinates (longitude and latitude) This will lead us to the conclusion that trend surface models are not beneficial in prediction since they aren’t suitable for spatial analysis, as spatial data violates the assumption of no autocorrelation, one of the important assumptions in regression models as well as trend surface models.

One can judge the best model by trying different orders (p’s), and calculating the R^2 and AIC for each order, then choosing the model with the lowest AIC and highest value of R^2. In most cases, we’ll be focusing mainly on the AIC value as it takes into consideration the number of parameters. In our case, we tried p=1,2,3 and decided, based on AIC value, to choose the cubic model(p=3), where its AIC = **2262.28** and R^2=**0.5577.**





- **Spatial Interpolation:**

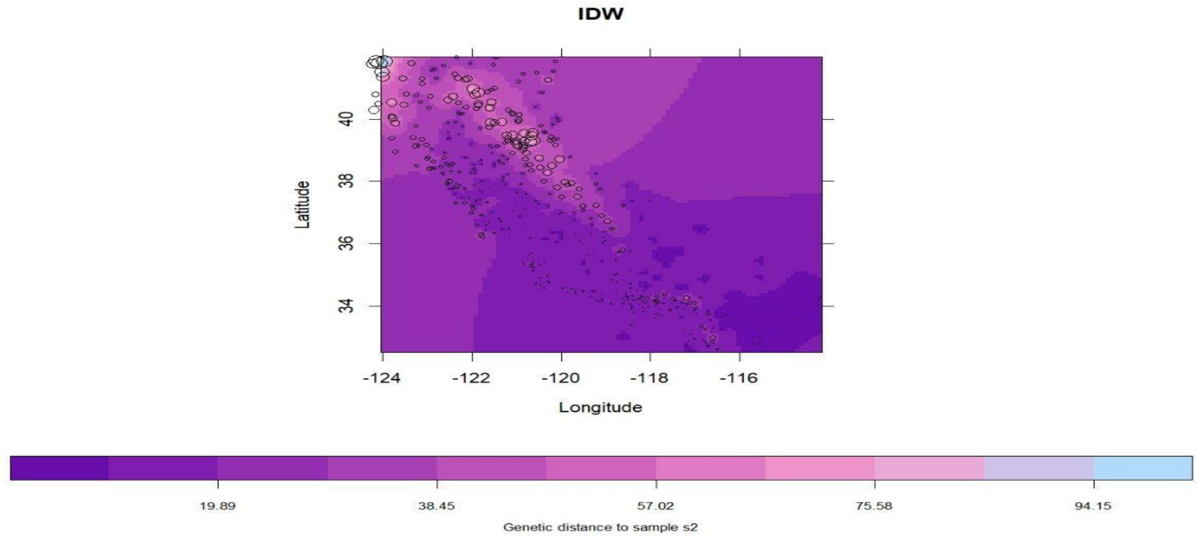
In this report, we'll be using two spatial interpolation techniques to estimate the values of unsampled locations in California based on the known temperature values at the nearby sampled locations. We'll be using the inverse distance weighting method and kriging to create the interpolated maps that will help in taking the decision of where to locate the unsampled locations in California.

- Inverse Distance Weighting:

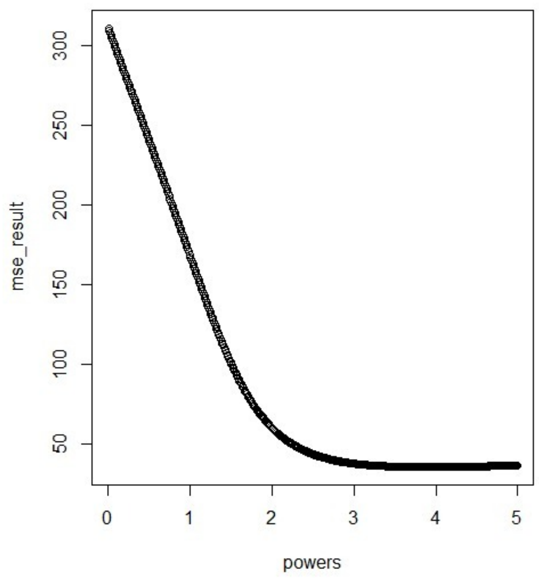
The idea behind inverse distance weighting is that the interpolating surface will be influenced more by nearby points than distant ones. It also recognises the interpolating surface as the weighted average of the points around it.

Although inverse distance weighting is not a suitable tool for prediction because it does not take into account other factors that can affect the dependent variable, represented in the error term, it does provide an overview of the dependency structure in the data. The main goal of applying this technique is to reach the optimal value of p to precisely predict the spatial dependency structure in the second quarter.

We were able to apply this by experimenting with different values of p ranging from 0.001 to 5 and calculating the MSE for each p included in the sequence. The optimal value for p is equal to **3.881** and has the least MSE of **35.72146**. Given that we rarely choose p greater than 5, the optimal value of p is large in comparison to 5, indicating that the predicted value depends on the values of the near locations and the far locations have a weight close to zero; in other words, the spatial dependency structure in the second quarter is local.



The plot shows that as we move from the South-East to the North-West, the average temperature in the second quarter increases.



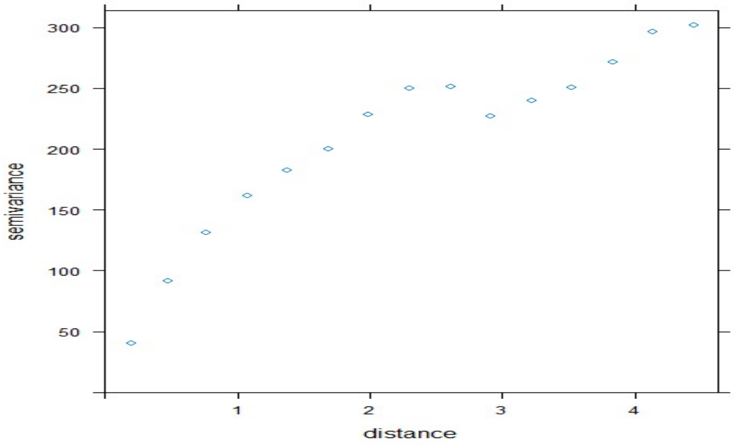
By plotting the different values of p against their calculated MSE, one can notice that as the p-value gets larger, the MSE decreases, so they're inversely related to each other.

- Kriging:

Since we can't use inverse distance weighting as a reliable technique for prediction; therefore, we'll use kriging. Kriging is the most suitable spatial interpolation technique to predict values at unsampled locations using the sampled ones, by pointing out the areas with the lowest variance and highest precision levels. There are some steps that need to be followed in order to reach the optimal level of prediction in the second quarter.

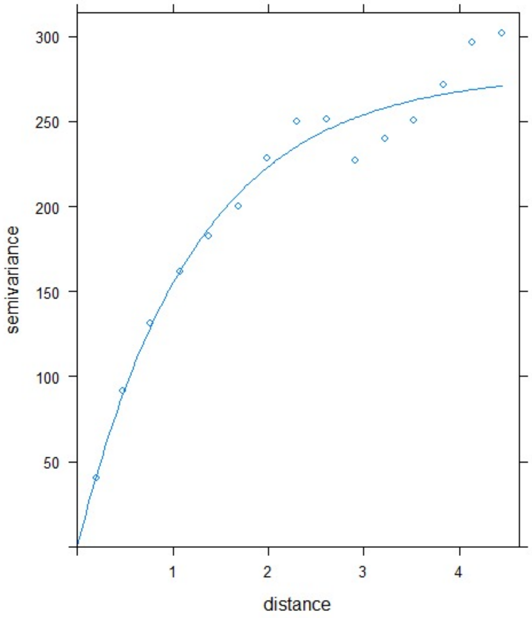
I. Fitting the empirical variogram:

We'll be using the variogram to do kriging so that we can assume intrinsic stationarity, we fitted the variogram and obtained this plot:



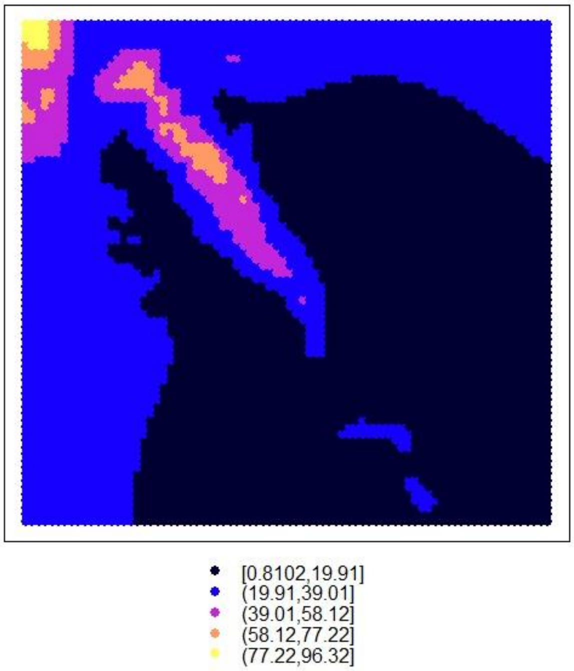
II. Choose the Suitable Parametric Variogram:

We tried three different parametric variograms (Spherical, Gaussian, and Exponential) because the empirical variogram does not satisfy the negative definiteness condition. We plotted the three variograms to find the best fit for our data, but because the plots were so subjective, we calculated the mean squared error (MSE) for each of them and chose the variogram with the lowest MSE, which was the exponential variogram whose MSE is equal to **1577298**.

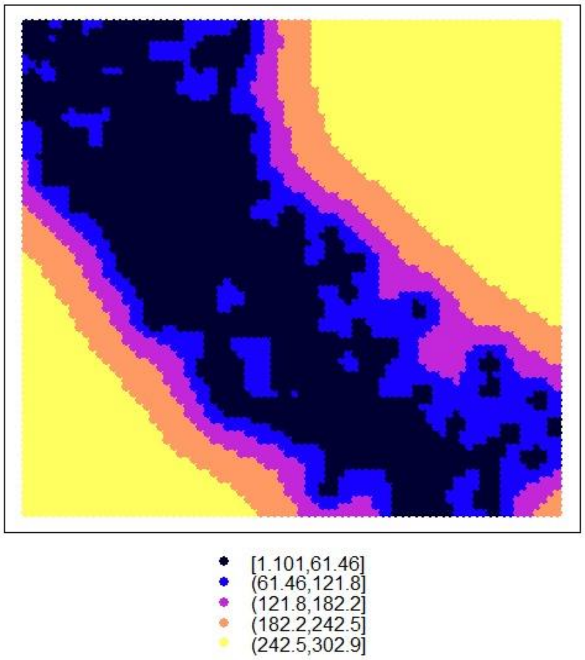


III. Create Interpolated Maps:

After choosing the suitable variogram for our data, we can reach the interpolated maps using kriging. We applied ordinary kriging as it's the most suitable kriging method since we have only one variable of interest which is temperature. Noting that ordinary kriging assumes that the mean is constant but unknown which is somehow realistic.



Prediction Map.



Variance Map.

In the variance map, we can notice that in North-East and South-West has the highest variance which ranges between (242.5 and 302.9), so in both areas we will have a low level of precision of the predicted values. From the North-East, the variance decreases gradually until it reaches the minimum variance, as we move towards the centre. Moreover, from the South-West as we move upwards, the variance decreases until it reaches the minimum variance range (1.101, 61.46). In the centre, where we have the lowest variance and different patterns of temperature, we'll have the highest levels of precision.

The lowest temperature values lie in the East, South-East, and South, where we have different patterns of variances. While the highest temperature values lie in the North-West which ranges between (77.22 and 96.32).

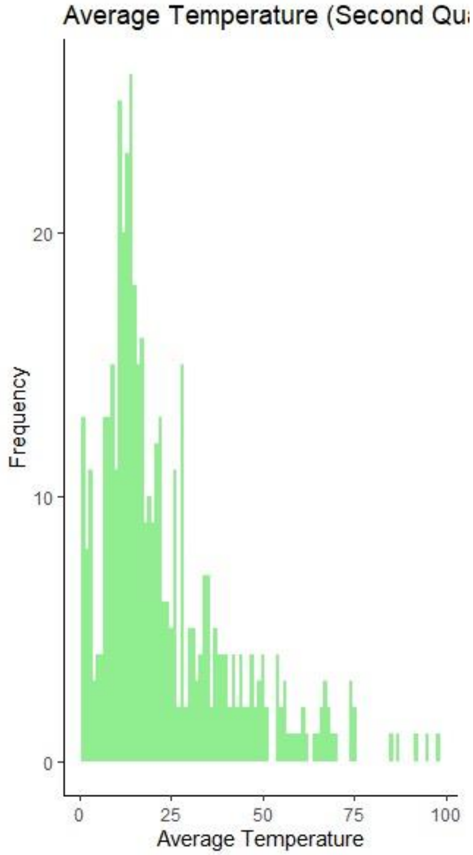
So, if we must choose where to locate our next station, we'll locate it in the area with the lowest variance range (1.101, 61.46) to obtain the highest level of prediction and accuracy. So, we could locate it in the centre.

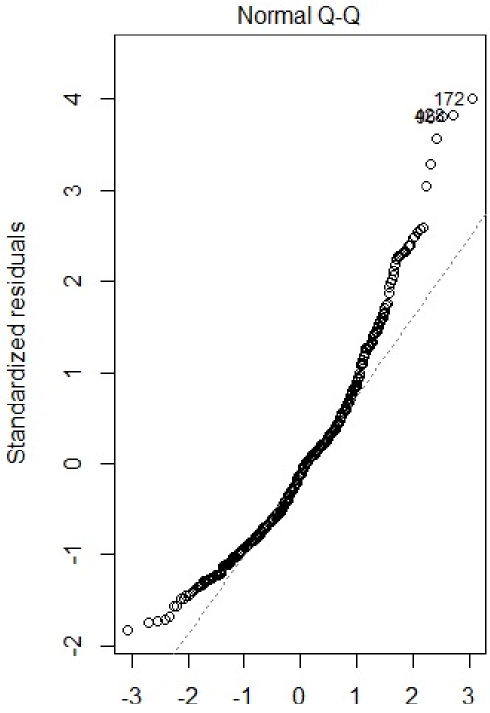
- **Conclusion:**

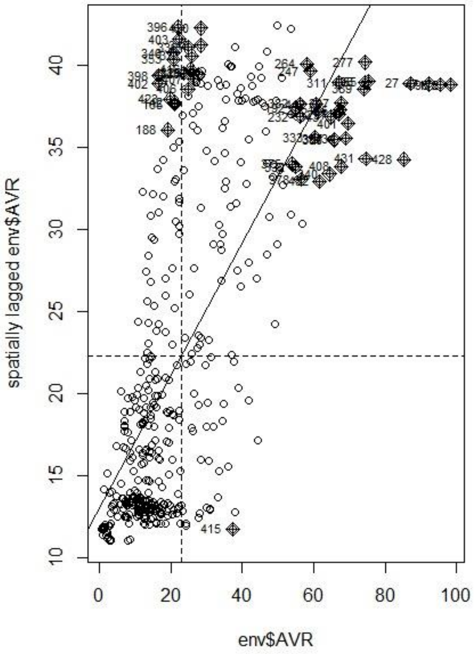
  - In the beginning of our report, we started by exploring the data and determining the shape of our data by checking the normality assumptions and conducting a bunch of plots that will represent our data graphically, also considering the GWRM as a method of exploring the data.
  - We then moved on to the objectives of our report, studying the spatial autocorrelation by getting both local and global measures of Moran's I. We conducted the required graphs to reach a conclusion about the clustering structure in the data.
  - We emphasized the importance of 3D plots by studying the trend surface model, choosing the third-degree trend surface model with the lowest AIC as it was the best fit for our data.
  - We then studied the spatial interpolation through both (Inverse Distance Weighting and Kriging), while applying the kriging technique we used the variogram function and estimated it using the exponential variogram. The kriging technique gave us an insight into where to locate our next stations.

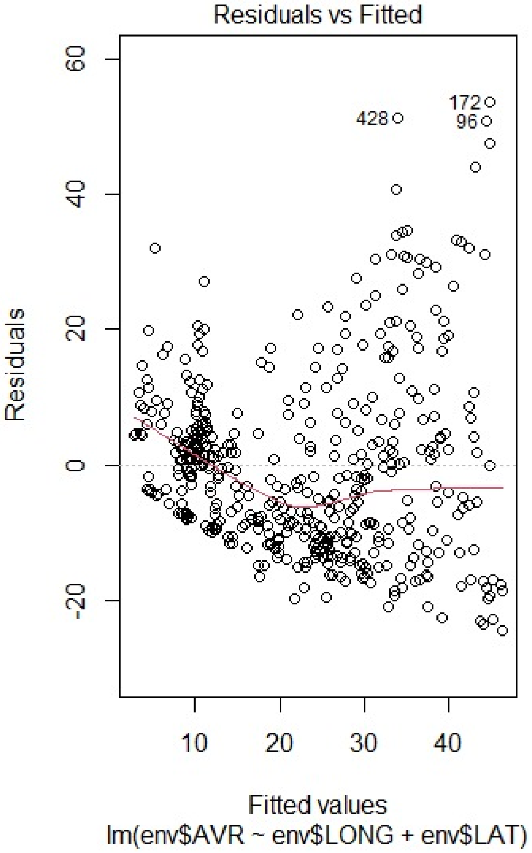
In conclusion, the report has offered a thorough review of many elements relating to studying temperature in California. Important patterns, trends, and linkages have been highlighted through the investigation of our data and the use of analytical techniques.

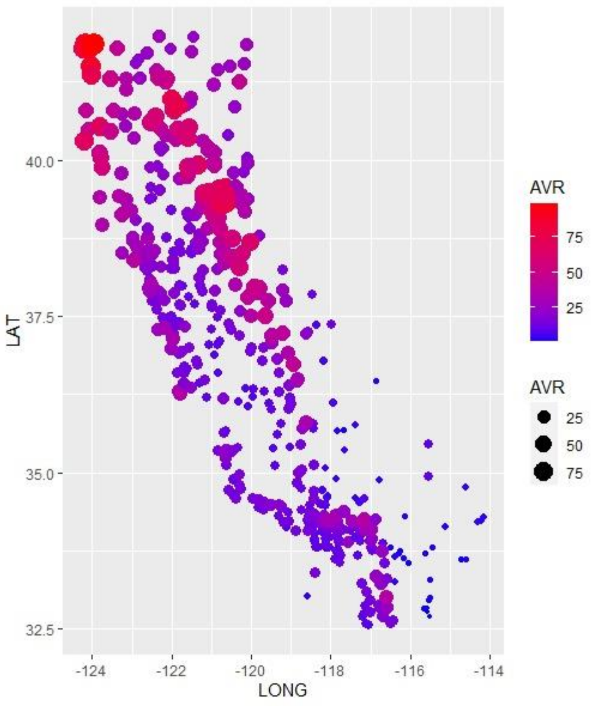
- **Appendix:**

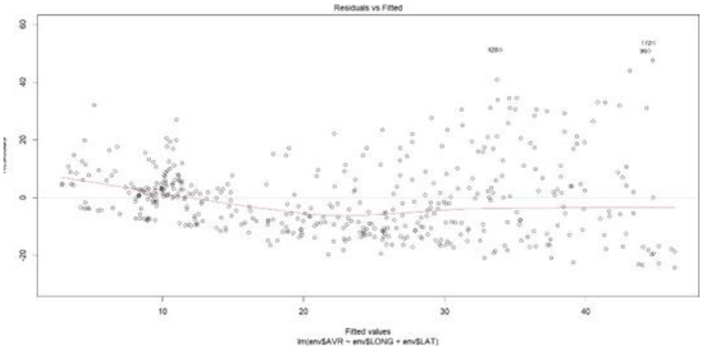

Average Temperature (Second Qu
Frequency
0
10
20
0
25
50
75
100
Average Temperature

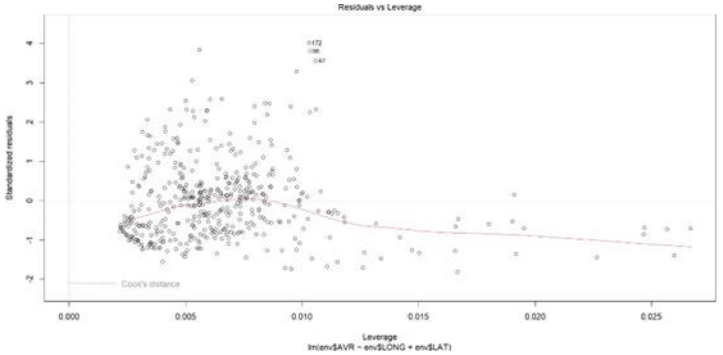

Normal Q-Q
Standardized residuals
-2
-1
0
1
2
3
4
-3
-2
-1
0
1
2
3
172
428
96

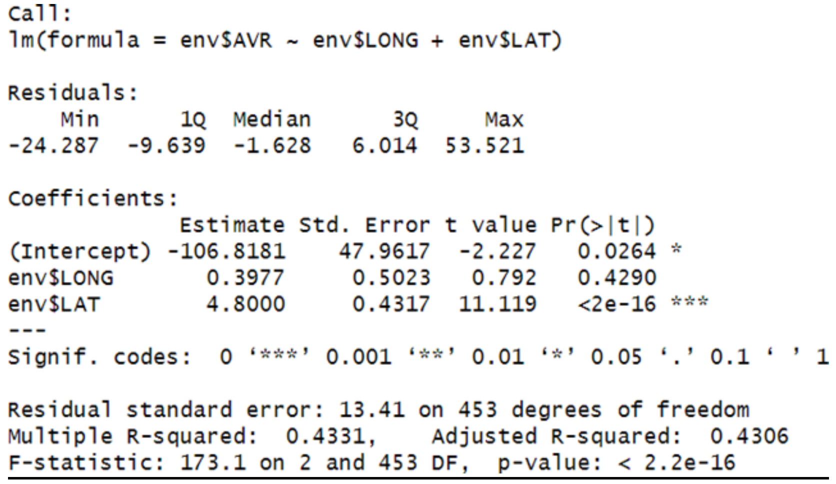

spatially lagged env$AVR
10
15
20
25
30
35
40
0
20
40
60
80
100
env$AVR
396
403
398
402
188
264
247
277
311
27
232
333
431
428
408
415


Residuals vs Fitted
Residuals
-20
0
20
40
60
10
20
30
40
Fitted values
lm(env$AVR ~ env$LONG + env$LAT)
428
172
96

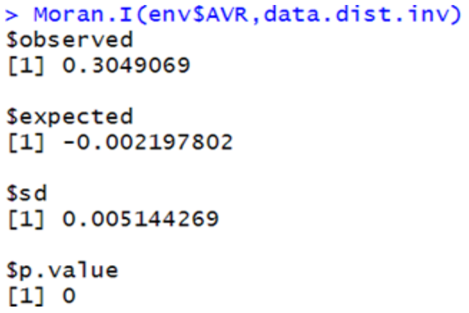




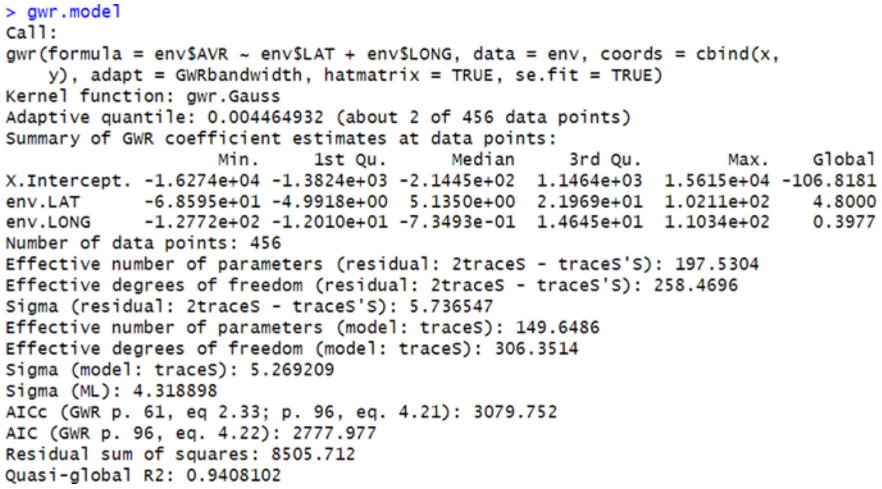




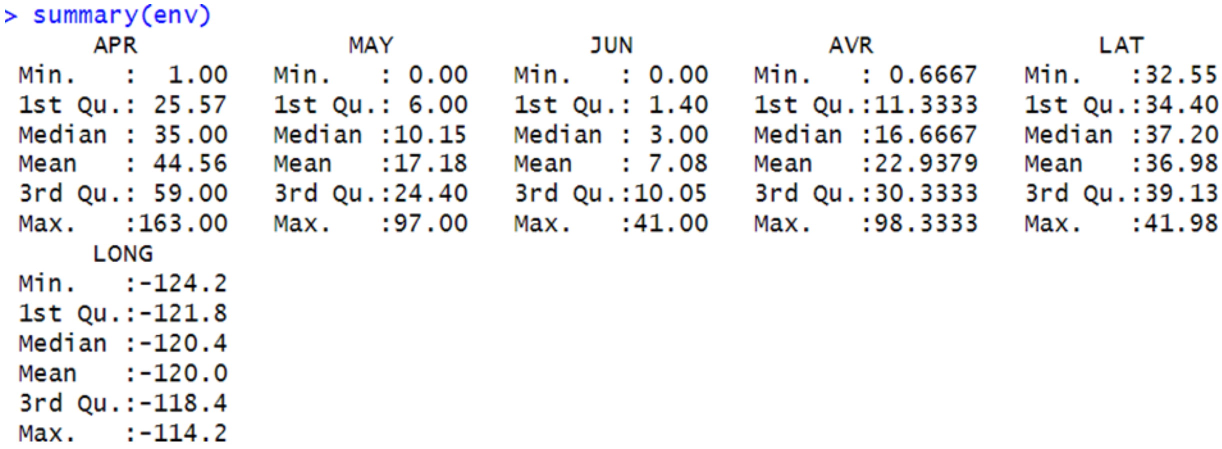
```
Call:
lm(formula = env$AVR ~ env$LONG + env$LAT)

Residuals:
    Min      1Q  Median      3Q     Max 
-24.287  -9.639  -1.628   6.014  53.521 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -106.8181    47.9617  -2.227   0.0264 *  
env$LONG       0.3977     0.5023   0.792   0.4290    
env$LAT        4.8000     0.4317  11.119   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.41 on 453 degrees of freedom
Multiple R-squared:  0.4331,	Adjusted R-squared:  0.4306 
F-statistic: 173.1 on 2 and 453 DF,  p-value: < 2.2e-16
```



```
> Moran.I(env$AVR,data.dist.inv)
$observed
[1] 0.3049069

$expected
[1] -0.002197802

$sd
[1] 0.005144269

$p.value
[1] 0
```



```
> gwr.model
Call:
gwr(formula = env$AVR ~ env$LAT + env$LONG, data = env, coords = cbind(x,
    y), adapt = GWRbandwidth, hatmatrix = TRUE, se.fit = TRUE)
Kernel function: gwr.Gauss
Adaptive quantile: 0.004464932 (about 2 of 456 data points)
Summary of GWR coefficient estimates at data points:
                    Min.     1st Qu.      Median     3rd Qu.        Max.    Global
X.Intercept. -1.6274e+04 -1.3824e+03 -2.1445e+02  1.1464e+03  1.5615e+04 -106.8181
env.LAT      -6.8595e+01 -4.9918e+00  5.1350e+00  2.1969e+01  1.0211e+02    4.8000
env.LONG     -1.2772e+02 -1.2010e+01 -7.3493e-01  1.4645e+01  1.1034e+02    0.3977
Number of data points: 456
Effective number of parameters (residual: 2traceS - traceS'S): 197.5304
Effective degrees of freedom (residual: 2traceS - traceS'S): 258.4696
Sigma (residual: 2traceS - traceS'S): 5.736547
Effective number of parameters (model: traceS): 149.6486
Effective degrees of freedom (model: traceS): 306.3514
Sigma (model: traceS): 5.269209
Sigma (ML): 4.318898
AICc (GWR p. 61, eq 2.33; p. 96, eq. 4.21): 3079.752
AIC (GWR p. 96, eq. 4.22): 2777.977
Residual sum of squares: 8505.712
Quasi-global R2: 0.9408102
```



```
> summary(env)
      APR               MAY               JUN               AVR                 LAT
 Min.   :  1.00   Min.   : 0.00   Min.   : 0.00   Min.   : 0.6667   Min.   :32.55
 1st Qu.: 25.57   1st Qu.: 6.00   1st Qu.: 1.40   1st Qu.:11.3333   1st Qu.:34.40
 Median : 35.00   Median :10.15   Median : 3.00   Median :16.6667   Median :37.20
 Mean   : 44.56   Mean   :17.18   Mean   : 7.08   Mean   :22.9379   Mean   :36.98
 3rd Qu.: 59.00   3rd Qu.:24.40   3rd Qu.:10.05   3rd Qu.:30.3333   3rd Qu.:39.13
 Max.   :163.00   Max.   :97.00   Max.   :41.00   Max.   :98.3333   Max.   :41.98
      LONG
 Min.   :-124.2
 1st Qu.:-121.8
 Median :-120.4
 Mean   :-120.0
 3rd Qu.:-118.4
 Max.   :-114.2
```