Spatial Data Analysis for Temperature in Third Quarter in California

Name	Code
Sohila Reda Mohamed Hassan	5190267
Mariam Khairy Mansour	5190875
Nagwa Ahmed Mohammed Nasr	5190539

Under supervision of

Dr.Sara Osama

Spring 2023

Contents

Introduction

Spatial data describes things in terms of their location on Earth. These could be people, animals, objects, natural phenomena, crime rates, or business outcomes. Spatial analysis is "*The process of examining the locations, attributes, and relationships of features in spatial data through overlay and other analytical techniques in order to address a question or gain useful knowledge*". We are going to analyze a spatial dataset about temperature values in California using descriptives measures to explore the data and assess the autocorrelation among these locations and, we're going to model this data using different methods of interpolation to see how different stations are affected by their neighbors.

Objectives:

The objective of this project is to apply a complete spatial data analysis on temperature in the third quarter in California state. This is achieved through exploring the data to discover general pattern in the dataset. Assuring the existence of autocorrelation is an important step to determine whether spatial analysis is needed or not. Fitting trend surface models which are simply regression model to determine the effect of Longitude and Latitude on our variable of interest. Interpolating unsampled locations within the area under investigation is an important objective that can be reached through applying Inverse Distance Weighting and Ordinary Kriging. Finally, we aim to get a detailed view of the temperature status in California and determine which areas that need more monitoring stations to enhance this existing network thus, improve our prediction.

Data description:

Our data consist of 456 stations in California state for one year and their corresponding coordinates. For our project, we are going to use the average of the months in quarter 3 (JUL, AUG, SEPT) as an indicator about temperature status in this quarter.

Section_1: Data Exploration and Visualization

Normality and outliers

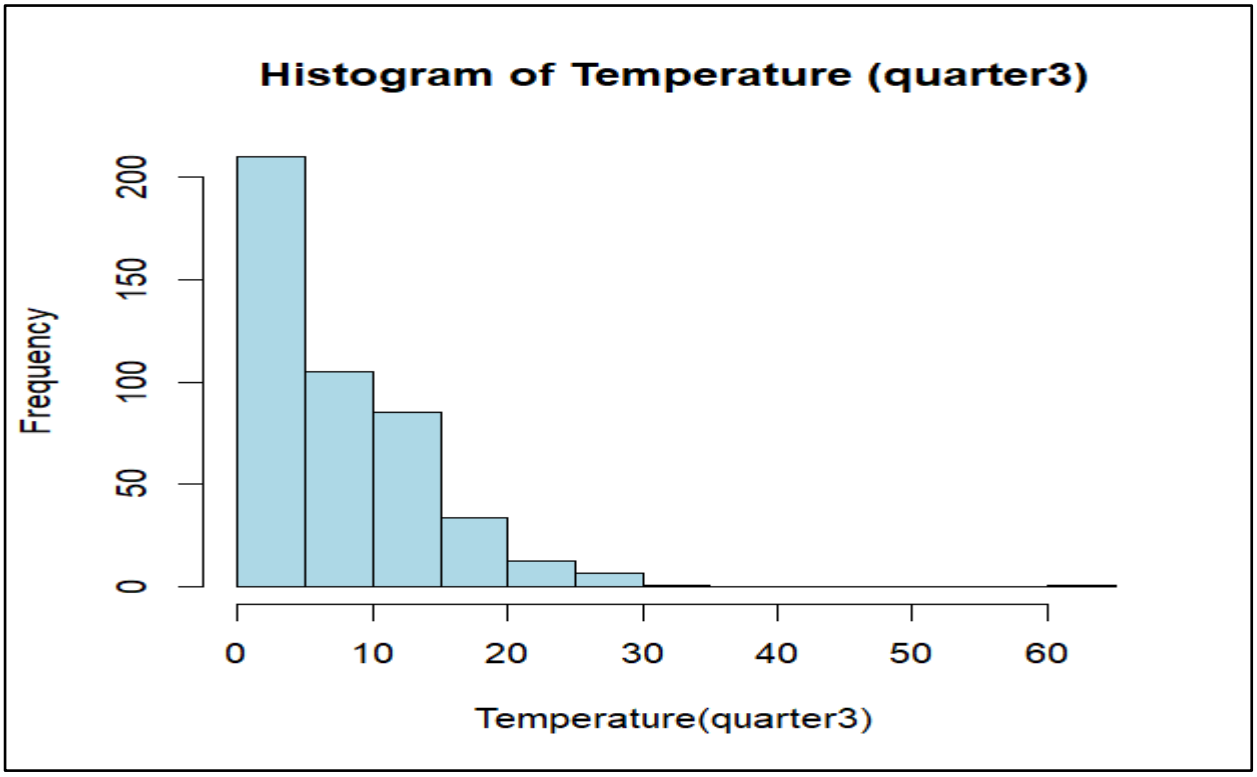

Figure 1.Histogram of Temperature in quarter 3

The data is right skewed with one distributional outlier of 61.66 in CORONA station that needs further investigation and this outlier is removed

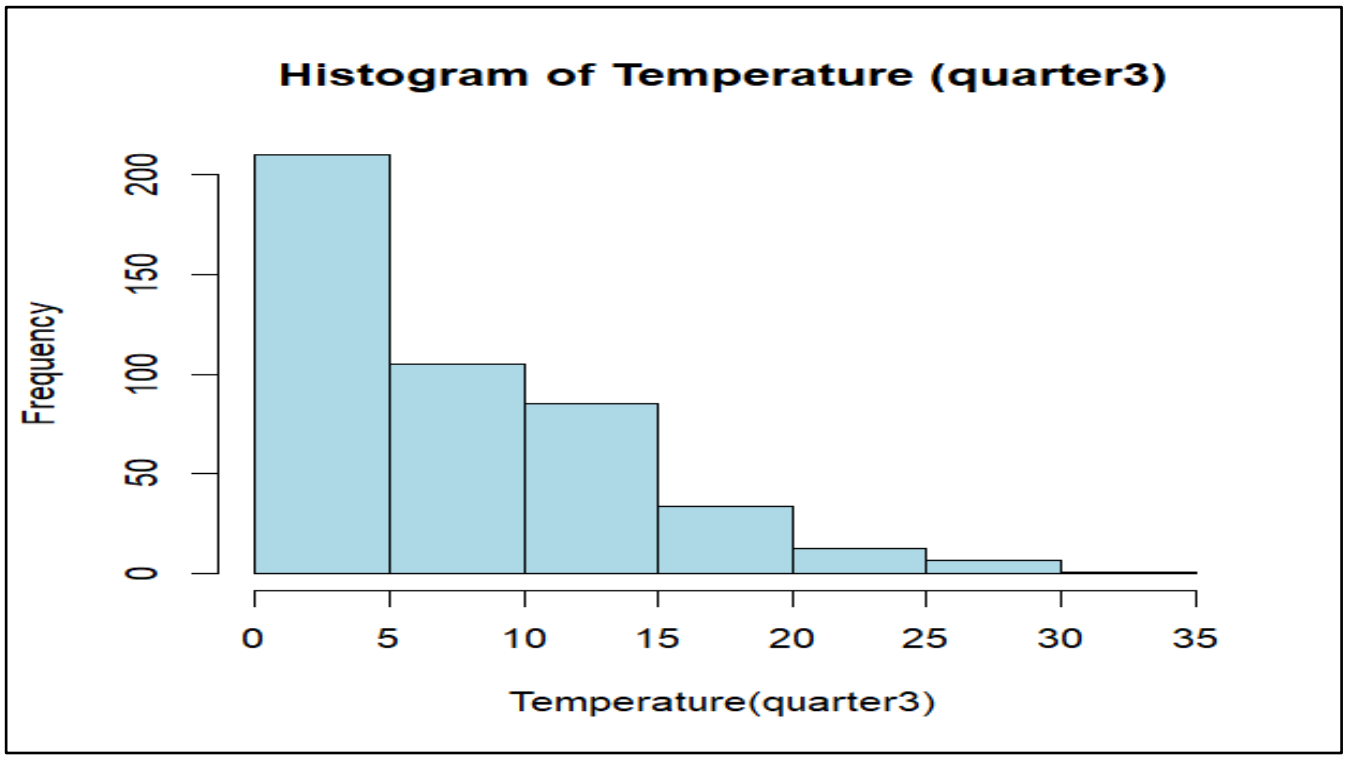

Figure 2.Histogram of temperature in quarter 3 after removing the outlier

Normality doesn't hold so transformations are needed to achieve it.

The transformation (sqrt (Temperature)) don't yield normality since it leads to p-value<0.05 in Shapiro test for normality.

Note that: log(Temperature) and 1/Temperature transformations are not available since the data contains an observation = 0.

Therefore, we are going to proceed with original data and assume that normality holds.

Bubble Plot

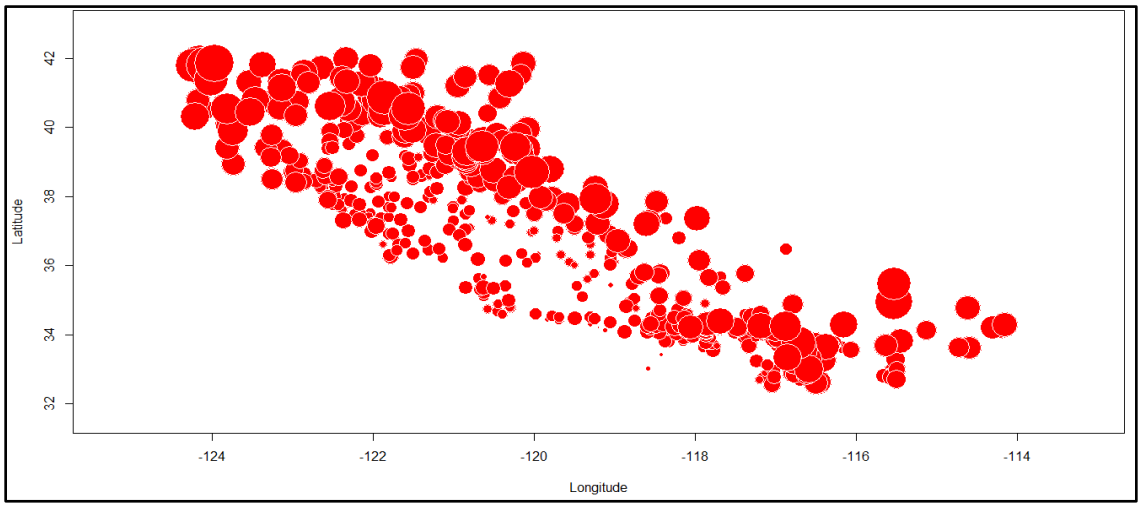

Figure 3. Bubble Plot for Temperature (Quarter 3)

We can notice from the bubble plot that temperature increases when we go east and there is a concentration of high values for temperature in the highest north.

Spplot

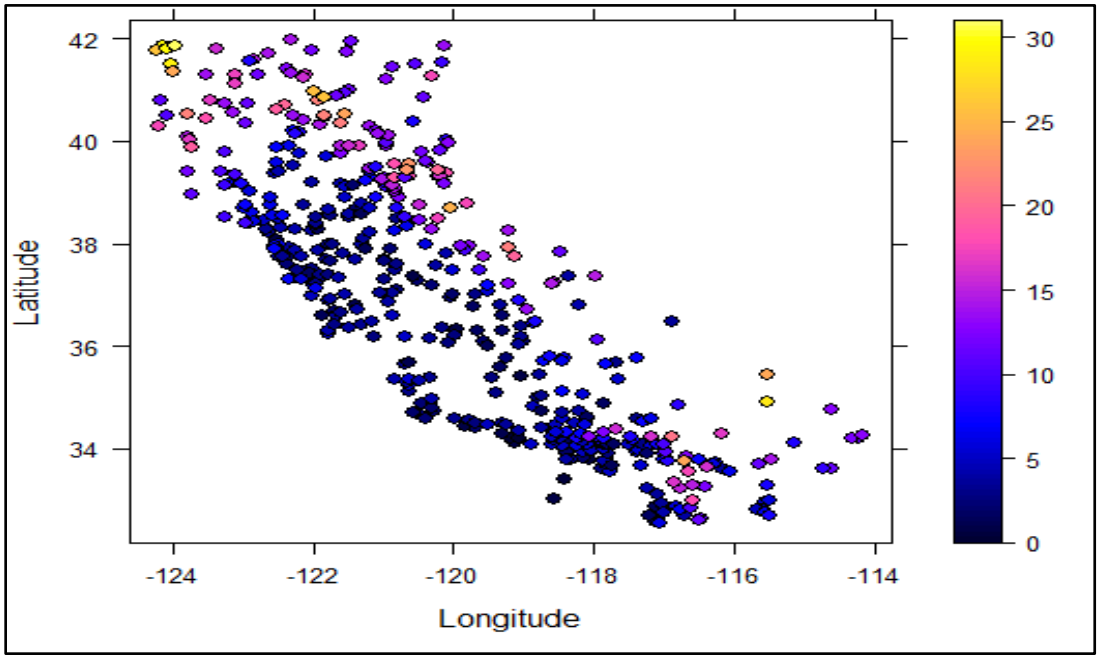

Figure 4.Spplot for Temperature (Quarter 3)

We can see from the spplot that temperatures ranges from 0 to 30 degree and go low when we go south especially at the west, there are some high temperatures at the highest northwest.

Autocorrelation

We will measure the autocorrelation on global and local level.

1) **Moran. I:-**

   We find that Moran-I= 0.2474155 > -0.002202643(Expected)
   and P-value =0 < 0.05 then there is significant weak positive(clusters) autocorrelation at $\alpha = 0.05$.

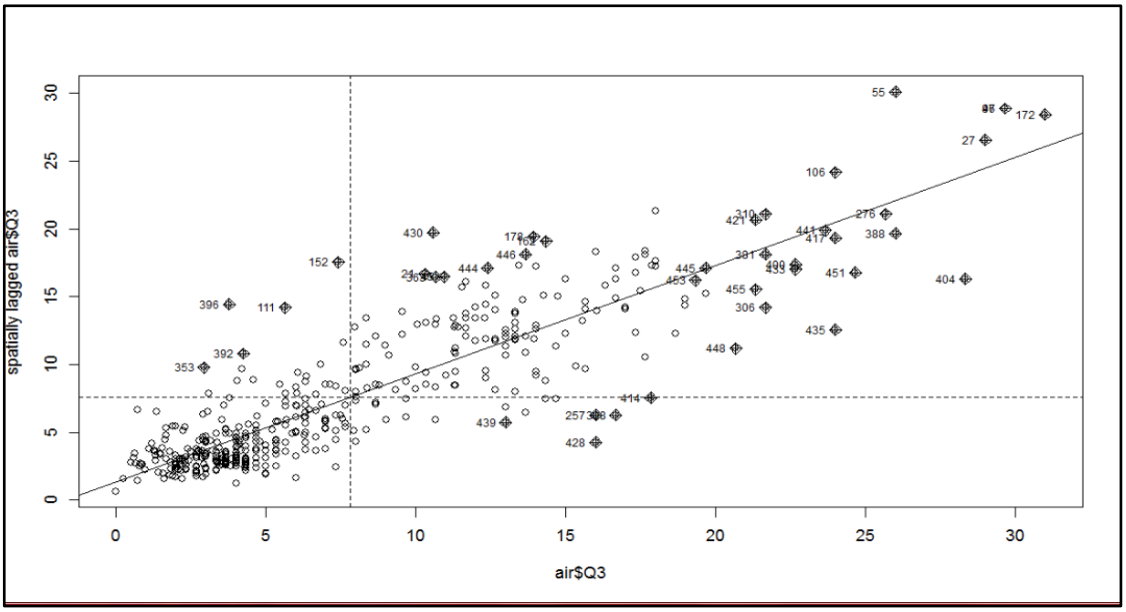

Figure 5.Moran-I plot

From figure (5):- we note from slope positive autocorrelation, there is a strong cluster in low-low quarter (low surrounded by low) and there is a moderate cluster in high-high quarter (high surrounded by high), there is a number of spatial outliers in high-high quarter (high surrounded by high) & low-high quarter (low surrounded by high) and high-low quarter (high surrounded by low).

2) Localized (Lisa plot):-

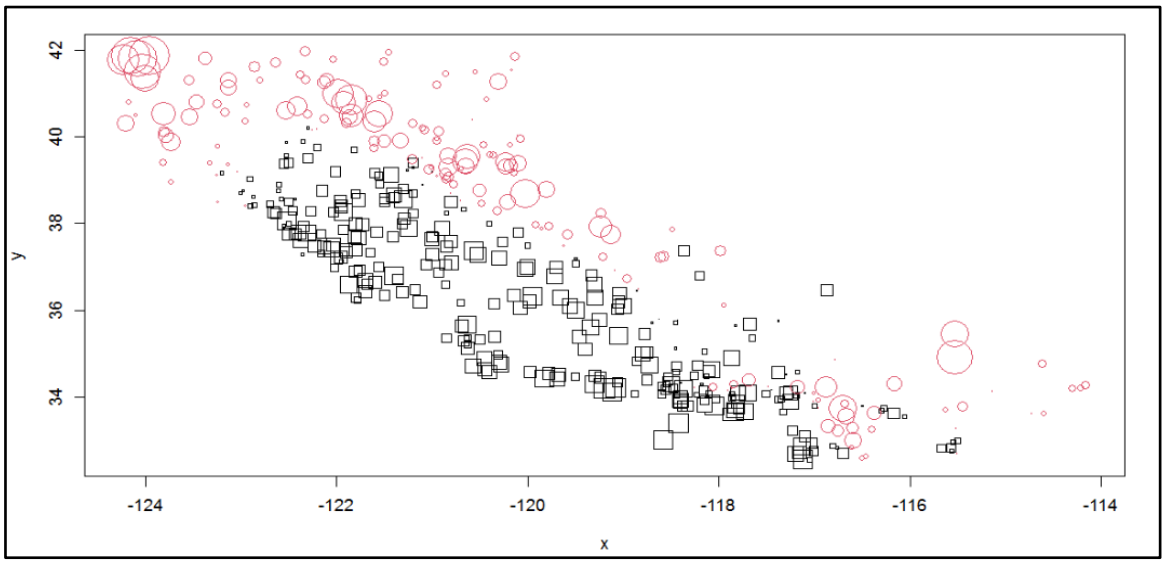

Figure 6.Lisa plot

Red circles, they are significant and above mean values (positive autocorrelation "clusters"). Black squares, they are significant and below mean values (negative autocorrelation).

Section_2: Data Analysis

- Inverse Distance Weighting:

We have applied IDW on our data using different values of P that produces lowest MSE.

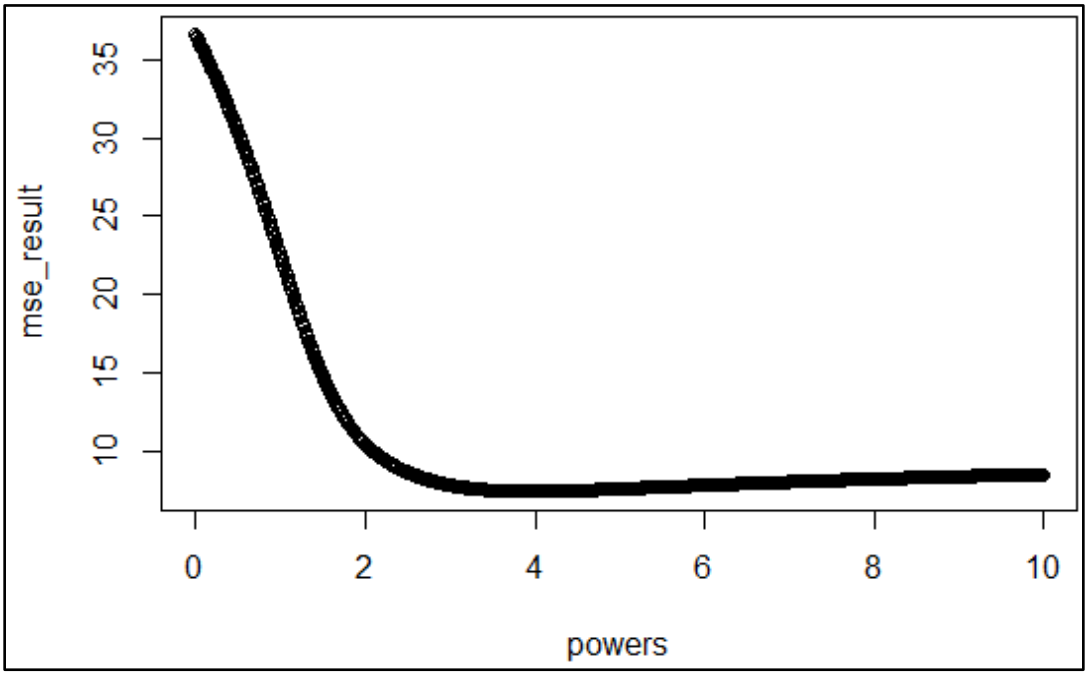

Figure 7.MSE for IDW Interpolation at different P values

We have found that the best p that gives lowest MSE is 3.951 which isn't close to the value zero which indicates that the stations are spatially autocorrelated and the closer points are highly weighted.

Applying IDW at P = 3.951

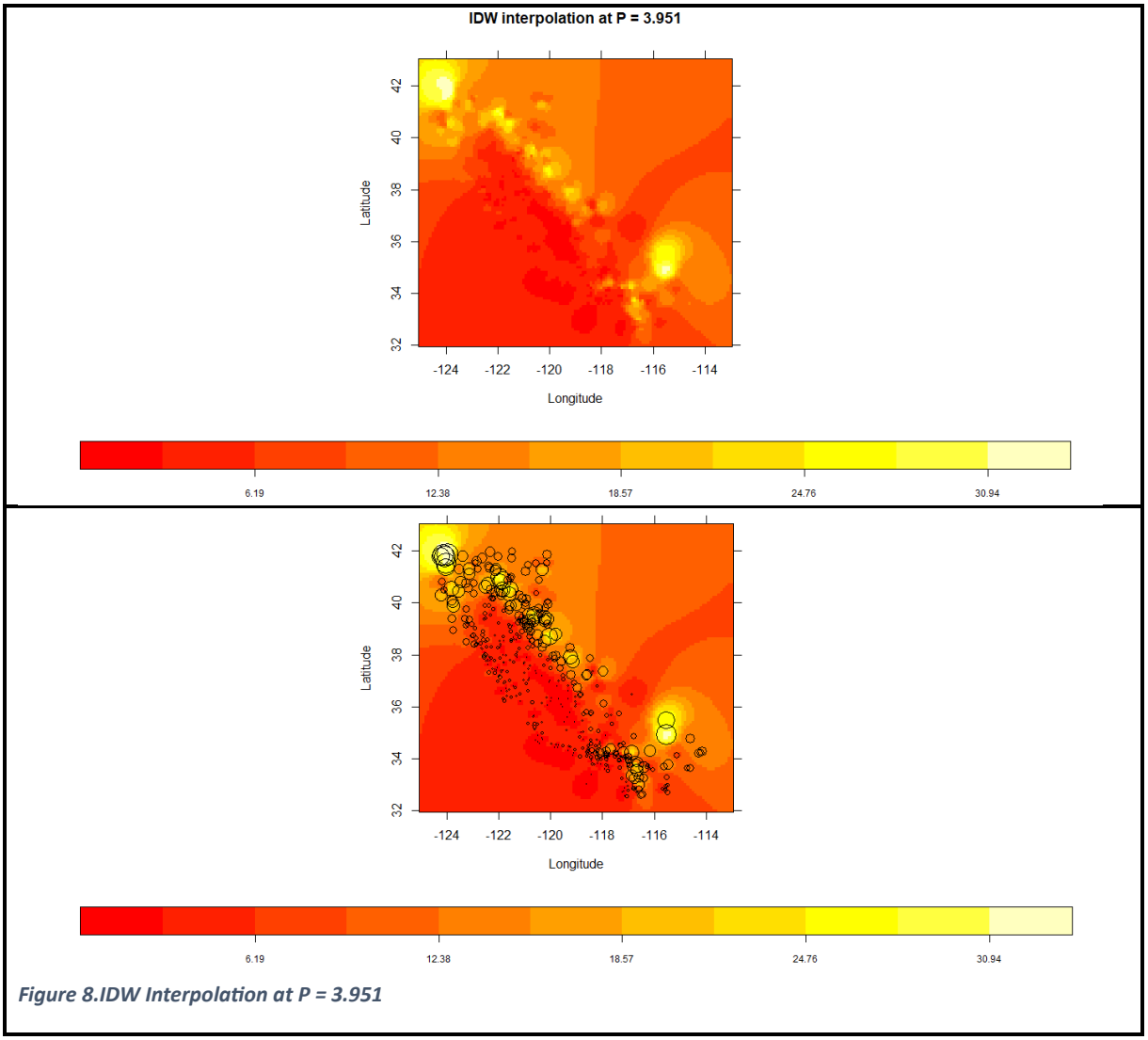

Figure 8.IDW Interpolation at P = 3.951

From figure (8), we can see that lowest values of temperature nearly less than 10 degrees concentrated at the south west of California while the highest values that are nearly higher than 25 degrees along the 45degree line from right especially at highest northwest and lowest southeast.

- Trend Surface Model

We do trend surface model at different P (1, 2, 3) and compare between them using AIC. We choose the lowest AIC at P=3 (AIC= 1210.77).

Number of parameters= (p+1)*(p+2)/2

Then number of parameters here = 10

$$\widehat{\textbf{Temperature}} = 9.996311 + 28.354335\,\text{Long} - 1.453182\,\text{Lat} - 52.029072\,\text{Long} * \text{Lat} + 21.226841\,\text{Long}^2 - 12.562872\,\text{Lat}^2 - 101.434762\,\text{Long}^3 - 4.778572\,\text{Lat}^3 - 96.461400\,\text{Long}^2 * \text{Lat} - 31.465894\,\text{Long} * \text{Lat}^2$$

Equation 1: Fitted cubic trend surface model for Temperature (quarter3)

P-value for model = 2.22e-16 < 0.05 then the model is significant.

R-Squared: 0.6234 and Adjusted R-squared: 0.6158, this mean that 62% from variation in temperature is explained by our explanatory variables.

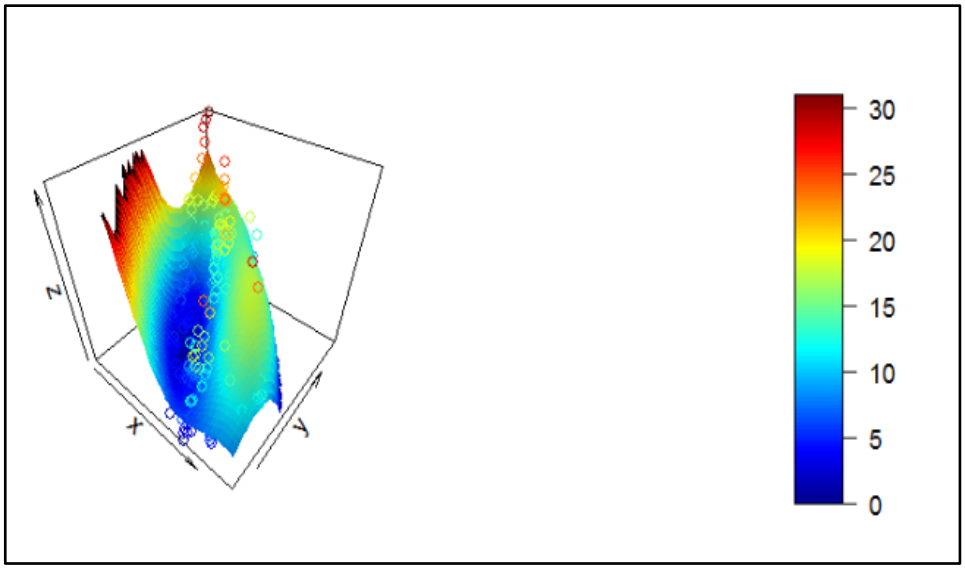

Figure 9.Trend surface model at P=3

From figure (9):- We find that in the west, the some values of temperature is high.

Note: - we shouldn't depend on trend surface model for prediction because it assumes that there is no autocorrelation.

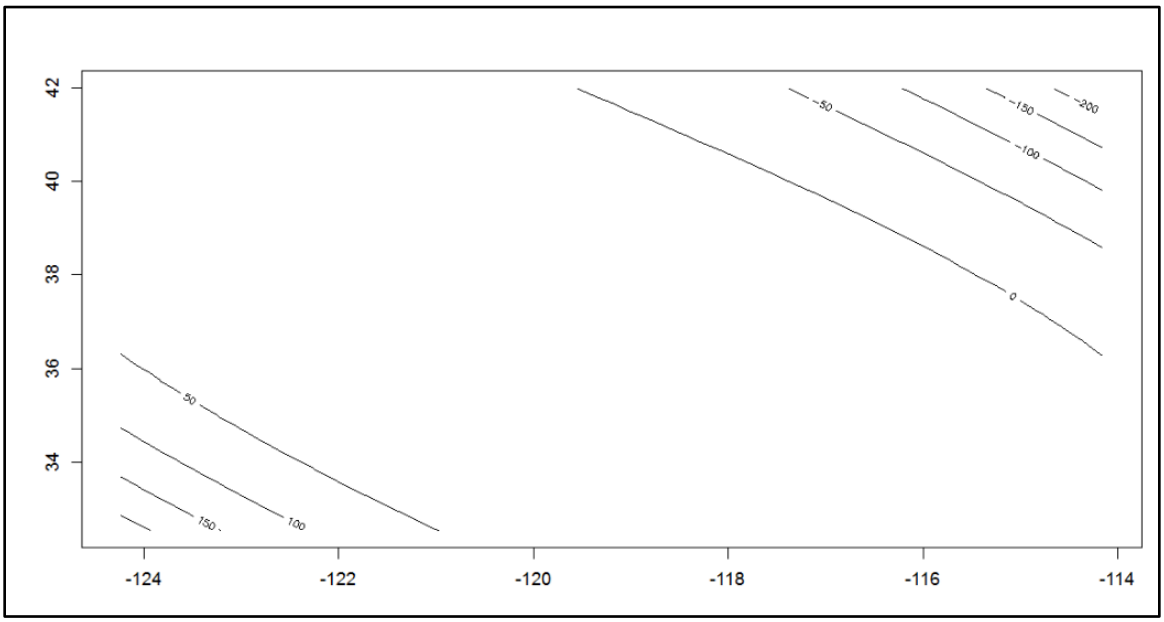

Figure 10.Contour

From figure (10):- the contour interval = 50

- Kriging

1)Variogram

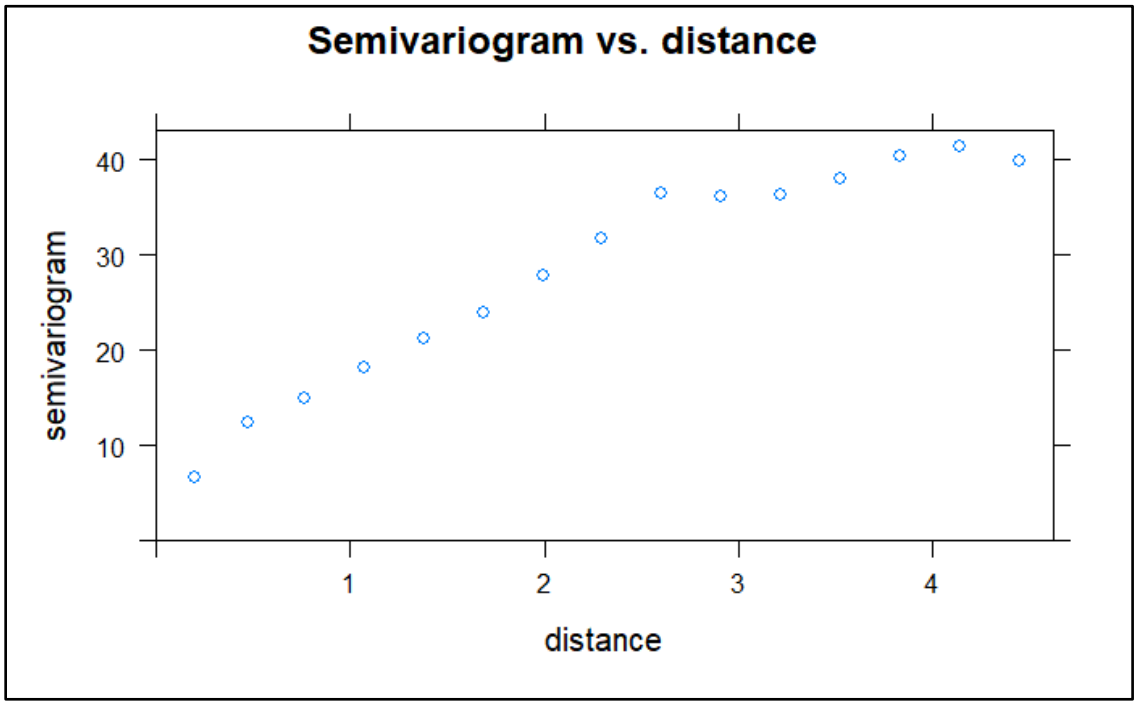

Figure 11. Semivariogram vs. distance

The empirical classical semi variogram is calculated and a suitable parametric model for it is aimed.

We have fitted different models and the best one that resulted in minimum SSE is the Exponential model.

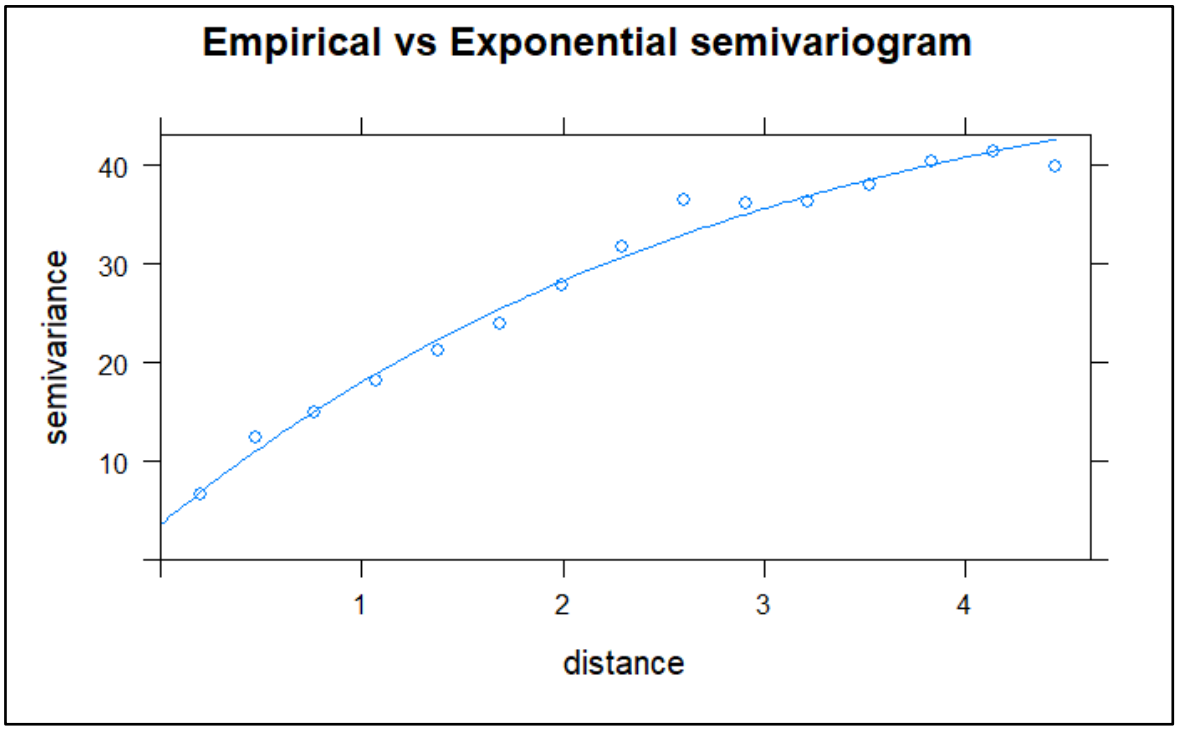

Figure 12. Exponential model for the semivariogram

The fitted semivariogram equation is:

$$\widehat{\gamma(h)} = 3.596 + 50.1516\,(1 - \exp\left(-3 * \frac{h}{2.947}\right))$$

Equation 2: Fitted semivariogram exponential model for Temperature (quarter3)

- The semivariogram of the difference between tends to 3.596 (nugget) when the distance between any two points approaches zero.
- The maximum variation in the data between any two stations that are h distance apart is (3.596+50.1516) = 53.7476 (sill).
- This maximum variation is reached at distance 2.947 (range).

This semivariogram model is used to fit ordinary kriging

2)Kriging

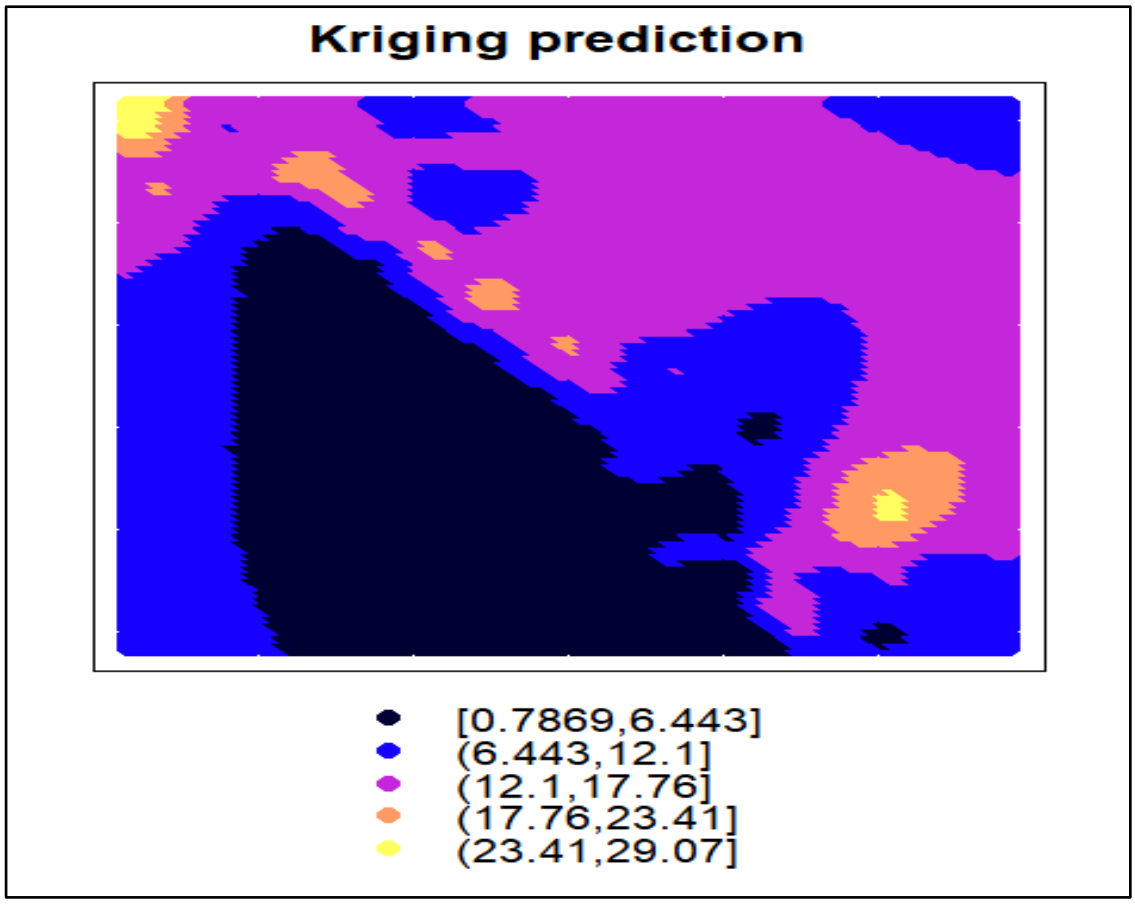

Figure 13.Kriging prediction map for Temperature (quarter 3)

- The predicted temperature ranges from 0.787 to 12.1 in south California while in northeast of California at the third quarter, it ranges from 12.1 to 17.76 in most areas. From northwest to southeast, it seems that there is a line of relatively high predicted temperature ranging from 17.76 to 29.07.

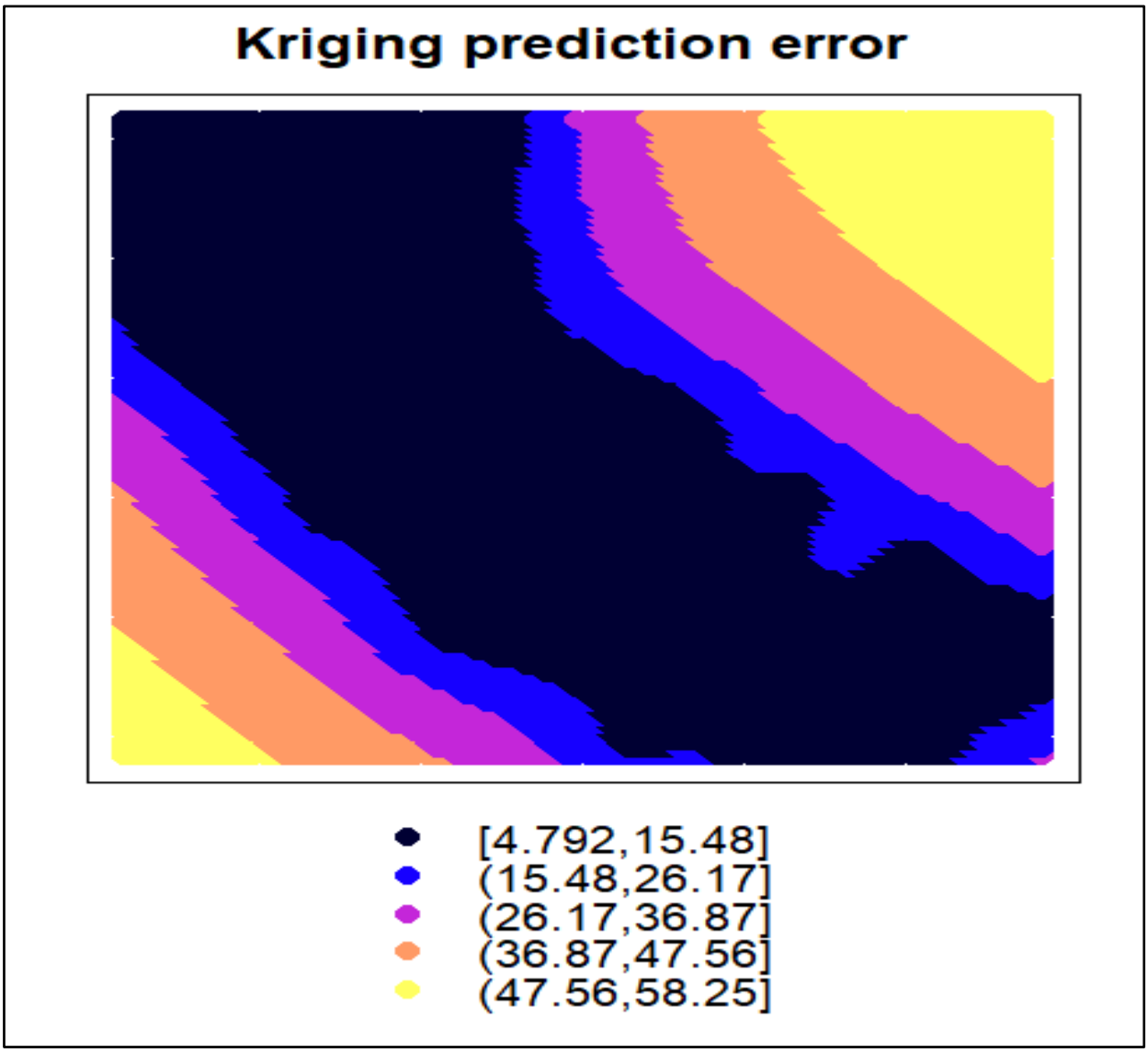

Figure 14.Kriging prediction error map for Temperature (quarter 3)

- The prediction error is at its minimum in the middle (from northwest to southeast) of California ranging from 4.792 to 15.48 and as we move northeast or southwest, the prediction error increases until it ranges from 47.56 to 58.25 at the borders of California. This prediction error map may establish that temperature monitoring stations are concentrated at the middle of California and few stations exist as we move northeast or southwest as shown in IDW interpolated map with points representing stations.

Section_3: Conclusions and Recommendations

- Conclusion

Finally, we know some of information about California, we need to study it locally since there are spatial outliers and P=3.951 in IDW. There are clusters (positive autocorrelation), strong cluster in low-low quarter (low value of temperature surrounded by low values of temperature). The highest values of temperature in northwest and east of California and the less values in southwest and south of California. We can't depend on ordinary kriging in southwest and northeast of California since they have the highest prediction error.

- Recommendation

1. Using kriging for interpolation is better since the inverse distance weighting assume no error (mathematical function) and the trend surface model assume no autocorrelation.
2. We note that we need more stations in southwest and northeast of California since they have the highest prediction error [(36.96, 47.61) & (47.61, 58.26)] based on ordinary kriging and the best prediction in the middle of California.
3. We can do things which need more temperature in northwest and east of California & less temperature in southwest and south of California.