Enviromental Statistics Project

Interpolating California Temperature

Mariam Amir
Rana Osama
Malak Bayoumy

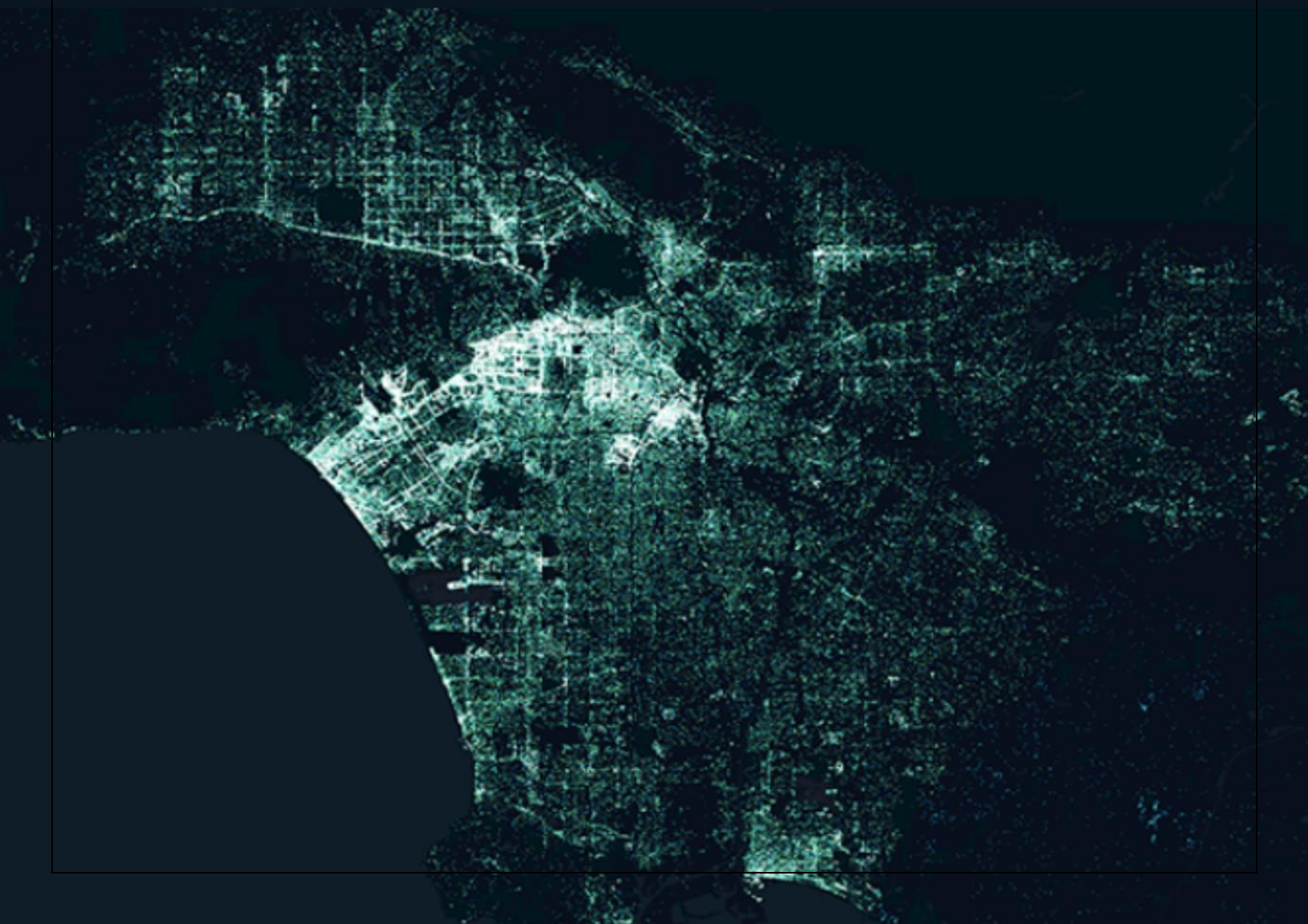

Table of Contents

Introduction:

The Environmental Statistics Project focuses on spatial interpolations to estimate temperature values in California. Spatial interpolation is a valuable technique that allows us to estimate values at unsampled locations based on observed data at sampled locations. By applying spatial interpolation to temperature data collected from 456 locations across the state, we can gain insights into the spatial distribution of temperature and uncover patterns and trends that can inform decision-making processes.

Spatial data analysis is a critical field in data science, enabling us to understand geographic data and make informed decisions. In this study, we aim to utilize spatial data analysis techniques to analyze California's temperature data comprehensively. By examining the relationships between atmospheric conditions and temperature changes, we can predict weather patterns for specific regions and identify the factors influencing temperature variations.

The findings from our analysis have practical implications for various industries and sectors. Decision-makers can leverage the unified spatial data to inform the siting of critical infrastructure such as power plants or cooling centers. Additionally, industries such as agriculture and tourism can benefit from insights into climate-related risks and opportunities. This report will provide a thorough analysis of California's temperature data, shedding light on its spatial patterns and enabling informed decision-making in various domains.

Main Objectives

The objective of this report is to employ spatial interpolation techniques to estimate temperature values in California. Specifically, we aim to:

1. Apply Inverse Distance Weighting (IDW) interpolation to estimate temperature values at unsampled locations and determine the optimal power parameter for IDW.
2. Utilize Trend Surface Models (TSM) to model the spatial variation of average temperature and identify the most suitable polynomial degree for the TSM.
3. Employ kriging, a spatial interpolation technique, to estimate temperature values across the study area based on the chosen theoretical variogram model.

Data Description & Exploration

Our data is the temperature values in California and we are working on the fourth quarter of the year, so we took the average of the three months (October, November and December) and applied the spatial data analysis

During the data exploration phase of our analysis, we examined various visualizations and statistical measures to gain insights into the characteristics of the temperature data in California. These exploratory techniques provide valuable information about the distribution, skewness, spatial autocorrelation, and relationships within the dataset. By understanding these aspects, we can uncover patterns and anomalies that will inform our subsequent analyses and interpretations.

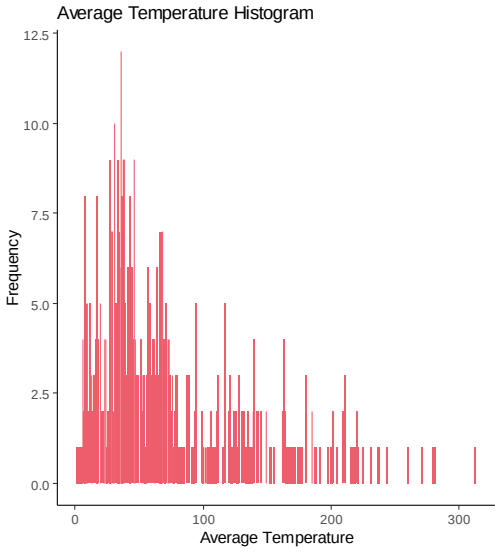

Figure 1

The histogram of the temperature data clearly indicates a positively skewed distribution, skewed to the right. This means that the majority of temperature values are concentrated towards the lower end, while a few higher values contribute to the positive skewness. Consequently, the mean of the data is expected to be greater than the median.

The box plot further supports the observation of positive skewness. The median (second quartile) is closer to the first quartile, indicating that the lower half of the data is more densely populated compared to the upper half. This skewness in the distribution has implications for our understanding of the dataset and subsequent analyses.

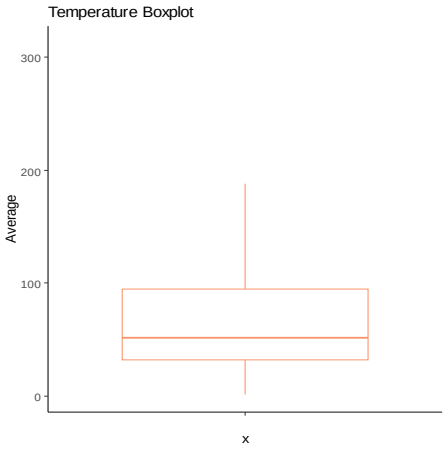

Figure 2

In the bubble plot of average temperature, the gradual decrease in the size of bubbles indicates a decrease in the values of average temperature in the fourth quarter of the year. Furthermore, the strong negative linear relationship between longitude and latitude is evident, highlighting the spatial trend in temperature values across California.

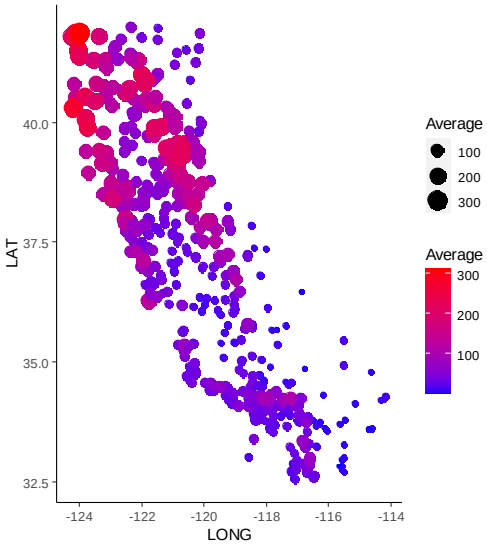

Figure 3

In order to study the spatial autocorrelation, Global Moran I was computed which was equal 0.318. Since it was less than E[I] and p – value was less than 0.05 this indicates +ve spatial autocorrelation. Moreover, in Moran I Plot, the clustering of data points suggests positive spatial autocorrelation. This means that similar temperature values tend to be located close to each other in space. This is evident from the presence of high values surrounded by other high values, indicating hot spots, and low values surrounded by similar low values, indicating cold spots. Additionally, the presence of outliers, where high values are surrounded by low values, adds further complexity to the spatial patterns observed in the dataset.

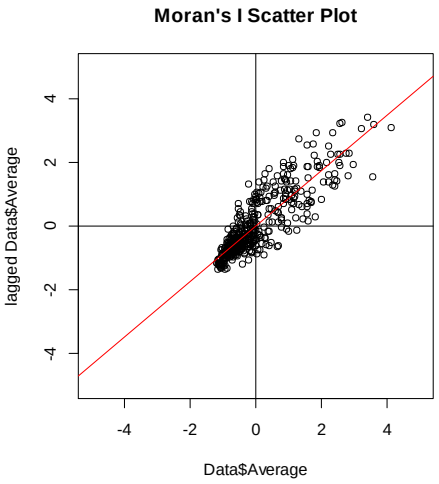

Figure 4

Lastly, the Lisa plot is a bubble-plot containing number of observations against spatial coordinates. Above mean values are signified by pink circles. Below mean values are signified by black squares. If a permutation test was performed, observations for which the associated LISA statistic is significant at a nominal (two-sided) 5%-level will be represented by filled symbols and non-significant values by open symbols. Thus spatial hot-spots are represented by pink filled circles and cold-spots by black filled squares.

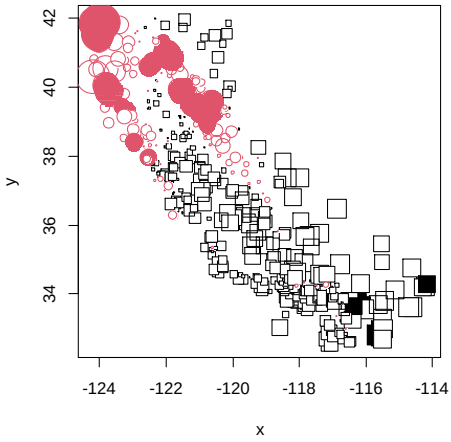

Figure 5

These exploratory visualizations and statistical measures lay the foundation for our subsequent analysis and interpretation of California's temperature data. They provide valuable insights into the distribution, skewness, spatial autocorrelation, and relationships within the dataset, enabling a comprehensive understanding of the temperature patterns across the state.

Spatial Interpolations

Spatial interpolation is a technique used to estimate values of a variable of interest at unsampled locations within an area based on known observations at sampled locations. In the context of California, for instance, we have temperature data available at 456 sampled locations across the state, spatial interpolation can be used to estimate the temperature values at other unsampled locations within California.

Inverse Distance Weighting part

The process of spatial interpolation involves the use of statistical and mathematical techniques to create a continuous surface of the variable of interest, such as temperature, over California.

One of the widely used mathematical techniques in spatial interpolation is Inverse Distance Weighting (IDW), which assumes that the values of a variable at unsampled locations are a weighted average of the values at sampled locations. The weighting factor is inversely proportional to the distance between the unsampled location and the sampled locations. The closer a sampled location is to the unsampled location, the greater its influence on the estimated value.

IDW interpolation also involves a power parameter, commonly referred to as the decay exponent, which is used to adjust the relative influence of each sampled location on the estimated value at the unsampled location. The decay exponent determines how quickly the influence of the sampled values decreases with distance from the unsampled location.

Though inverse distance weighting may be used for interpolations, but its main use is understanding the spatial dependency structure of our data. To illustrate, the power parameter in IDW plays a crucial role in determining the spatial dependency structure of the data, and by varying this parameter, we can gain insight into the spatial patterns of the data.

The power parameter controls the degree to which the influence of the sampled locations on the estimated value at the unsampled location decreases with distance. When the power parameter is set to a low value, such as 1, the influence of the sampled locations decreases slowly with distance, resulting in a more gradual decline in influence over space. This implies a strong spatial dependence structure, with nearby locations having more similar values than distant locations.

On the other hand, when the power parameter is set to a high value, such as 2 or 3, the influence of the sampled locations decreases rapidly with distance, resulting in a steep decline in influence over space. This implies a weak spatial dependence structure, with nearby locations having less influence on each other than distant locations.

By experimenting with different values of the power parameter in IDW, we can gain insight into the spatial dependency structure of our data.

To find the optimal power value, the code employs a systematic approach known as grid search. It involves defining a range of power values to test, spanning from a small value (0.001) to a large value (10) with small increments (0.01). This range ensures that a wide spectrum of power values is considered.

For each power value in the range, the code performs the following steps:

1. Performs IDW interpolation using the current power value.
2. Calculates the Mean Squared Error (MSE) between the actual values and the interpolated values.
3. Stores the MSE value for the current power in a vector.

By evaluating the MSE for each power value, the code determines the power value that yields the lowest MSE. This is accomplished by finding the index of the minimum MSE value in the vector of MSE results and extracting the corresponding power value.

The optimal power value represents the power parameter that results in the IDW interpolation with the least amount of error when compared to the actual values. It provides a data-driven indication of the power value that best captures the underlying spatial patterns in the dataset.

After executing the code, it determined that the optimal power value for the IDW interpolation method was 3.991. By selecting a power value of 3.991 as the optimal value, it suggests that there is a moderate spatial dependence in the dataset. This implies that points located closer together have a higher likelihood of exhibiting similar values or characteristics, while points that are farther apart are less likely to influence each other.

Understanding the spatial dependency of the data is crucial for selecting an appropriate power value in IDW interpolation. By considering the optimal power of 3.991, we can leverage the moderate spatial dependence in the dataset to generate interpolated values that capture the underlying patterns and spatial relationships more effectively.

Figure 6 shows the relationship between the power and the MSE.

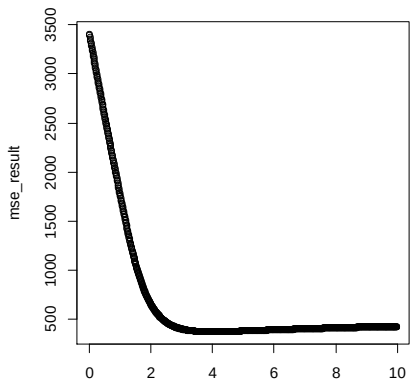

Figure 6

Observing the interpolated maps, we can discern a gradual rise in temperature as we traverse from the eastern parts of California towards the western regions. This spatial pattern suggests a westward temperature gradient, with higher temperatures found in the western areas.

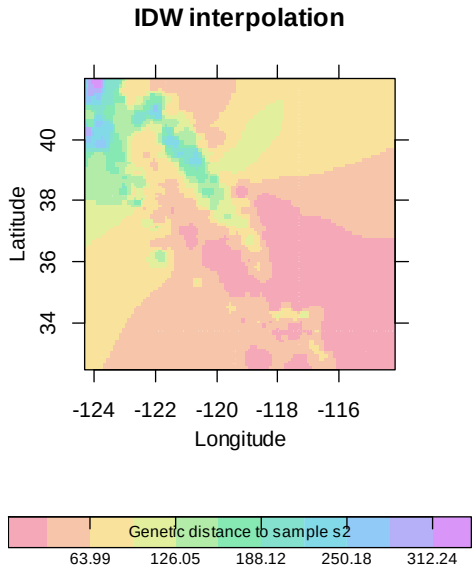

Figure 7

Trend Surface Model

The idea of a Trend Surface Model (TSM) is to estimate and visualize the spatial variation or trend in a dataset. It is a method used in spatial analysis and geostatistics to capture and represent the underlying spatial pattern or trend in a set of data points across a continuous spatial domain.

The main concept behind a Trend Surface Model is to fit a mathematical function to the observed data points in order to model the trend. This function describes the spatial variation of the variable of interest (e.g., temperature) as a smooth surface or curve over the study are and enabling the estimation of values at unobserved locations within the study area.

However, one of the drawbacks of trend surface models is Spatial Autocorrelation Ignorance. To elaborate, Trend Surface Models do not explicitly account for spatial autocorrelation, which is the tendency of nearby locations to have similar values. Ignoring spatial autocorrelation can result in biased estimates and inefficient use of data. It is essential to consider spatial autocorrelation and explore spatial statistical techniques, such as spatial regression or geostatistics, to account for the spatial dependence in the data.

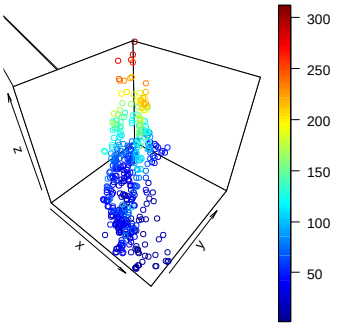

Figure 8: 3D Scatter Plot

We conducted a trend surface analysis on our dataset to capture and model the spatial variation of the average temperature. To begin, we created a 3D scatter plot (Figure 8) showcasing the distribution of data points in relation to longitude, latitude, and average temperature.

This visualization allowed us to gain insights into the overall spatial pattern of the variable.

Next, we fitted several trend surface models using different polynomial degrees to find the most suitable model for our data. We tested models ranging from the 2nd degree to the 5th degree polynomial. To assess the performance of each model, we computed various measurement errors, including R-squared and the Akaike Information Criterion (AIC).

After evaluating the R-squared values and AIC scores for the trend surface models of varying polynomial degrees, we determined that the 5th degree polynomial had the lowest AIC value of 3137.546. This indicates that the 5th degree polynomial model provides the best balance between complexity and fit for our data.

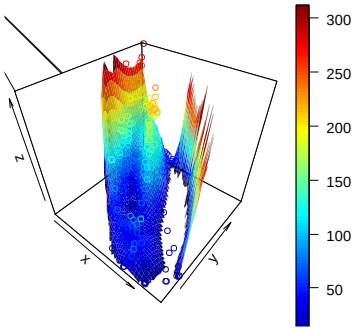

Figure 9: Fitted Model 5th order

Kriging

Kriging is a spatial interpolation technique used to estimate values at unmeasured locations based on observations at nearby locations. It takes into account both the spatial autocorrelation and the uncertainty associated with the data.

In our analysis, we utilized kriging to estimate the temperature values across the study area. To begin, we calculated the variogram, which measures the spatial dependence or correlation between temperature values at different locations. By analyzing the variogram, we aimed to identify the best fitting model that describes the spatial correlation structure of the temperature data.

We can't depend on the empirical variogram as the theoretical variogram is essential for fulfilling the positive definiteness condition of variance in spatial interpolation methods like kriging. It helps ensure the validity and reliability of the predictions by providing a suitable mathematical model that captures the spatial correlation properties of the data.

We explored various models, including the Exponential, Gaussian, Mattern, and Spherical, to determine the one that best captures the observed spatial correlation. After thorough examination, we found that both the Mattern and Exponential models exhibited the same Explained Sum of Squares (ESS), which was equal to 227496898. Additionally, their graphs showed similar characteristics.

Given these options, we selected the Mattern model as the best fitting model for our data. The Mattern model provides a reliable representation of the spatial correlation patterns observed in the temperature data, enabling accurate interpolation and estimation at unmeasured locations.

By applying kriging with the chosen Mattern model, we obtained temperature estimates throughout the study area, facilitating a comprehensive understanding of temperature distribution and aiding in decision-making processes related to various applications.

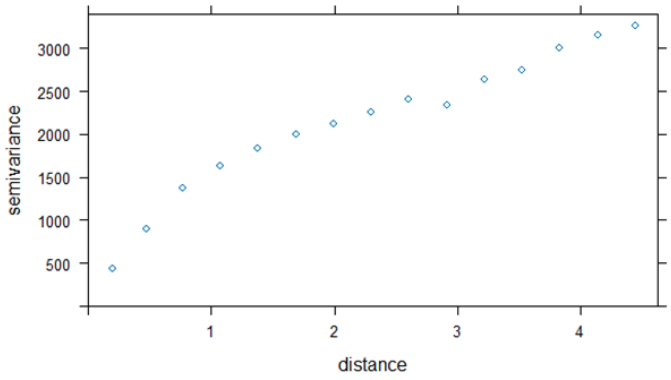

Figure 10:
Variogram

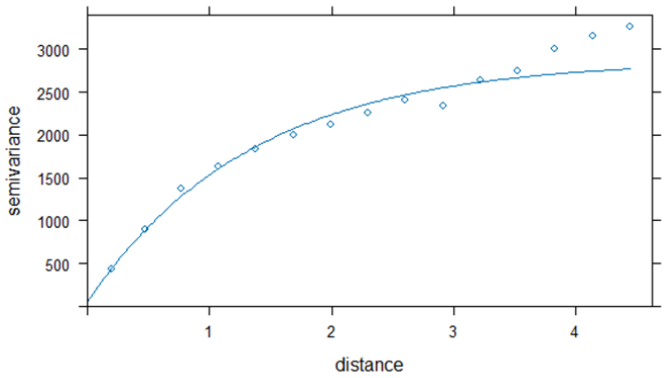

Figure :11 Mattern model on Variogram

We have chosen to conduct ordinary kriging and we computed the interpolated maps for the variance and predicted values and they showed the results as follows:

We concluded that we have at the north-east and south-west high variance which indicates that this area has low level of precision for the predicted values, also we observed that from north-west to south-east we have a lower variance which indicates higher level of precision.

When seeking to collect additional temperature data in California, prioritizing the placement of monitoring stations in areas characterized by high variances would yield the most beneficial results in terms of capturing the spatial variability and enhancing the comprehensiveness of the dataset.

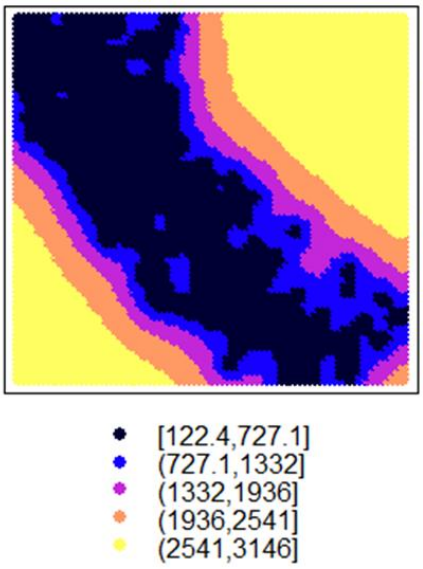

Figure 12: interpolated map for the variance

We can see at north-east that we have low predicted values of temperature and in the interpolated map of variance we were having a low level of precision, also at south-east, we have low predicted values of temperature while it had a higher level of precision at the interpolated map of variance. We have at south-west and north-west higher predicted values of temperature while the south-west region in the map of variance had low level of precision and the north-west region had higher level of precision. The regions with higher level of precision indicates that the predicted values of temperature are closer to the real values, while those with low level of precision indicates that the variance between the true values of temperature and the predicted ones is very large which means they are far beyond each other.

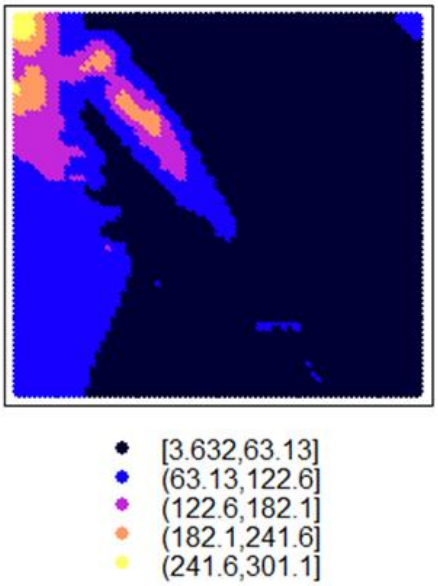

Figure 11: interpolated map for predicted values

Conclusion

In conclusion, this project focused on spatial interpolations to estimate temperature values in California. By employing Inverse Distance Weighting (IDW), Trend Surface Models (TSM), and kriging techniques, we gained insights into the spatial patterns and variations of temperature across the state.

The use of IDW allowed us to estimate temperature values at unsampled locations based on the values at sampled locations. By determining the optimal power parameter of 3.991 through a grid search approach, we found that there is a moderate spatial dependence in the dataset, with nearby locations having a higher likelihood of exhibiting similar temperature values.

The application of Trend Surface Models helped us capture the spatial variation of average temperature in California. By evaluating different polynomial degrees, we identified the 5th degree polynomial as the best fitting model based on the lowest Akaike Information Criterion (AIC) value. This model effectively represents the underlying spatial pattern of temperature in the state.

Furthermore, kriging, a spatial interpolation technique, was employed to estimate temperature values across the study area. By using the theoretical variogram and selecting the Mattern model, we were able to account for spatial autocorrelation and obtain reliable temperature estimates at unmeasured locations. Moreover, the variance maps gave us insights on where to place future stations.

Overall, these spatial interpolation techniques provide valuable tools for estimating temperature values in California and understanding the spatial patterns and variations of this important environmental variable. The results obtained from this project can aid in decision-making processes and inform further studies related to climate, environmental management, and spatial analysis in California.