University of Cape Town
Engineering and the Built Environment
School of Architecture, Planning and Geomatics
Geomatics Division

University of Cape Town

Subject: APG 4013C

Assignment 3: Geostatistical analysis

Sanele Mteshane(MTSSAN022)

Plagiarism Declaration

Declaration: I (name) ……SANELE …. is the author of this work, using my own words (except where attributed to others) I know that plagiarism is to use another's work and pretend that it is one's own and that this is wrong. I have used ……UCT Harvard ……. convention for citation and referencing. I have provided citations and references in all cases where I have quoted from the work of others or used other's ideas or reasoning in this essay/project/report.

Name	Student no. / code	Section(s) authored
SANELE	**MTSSAN022**	**All the sections**

Signed: _____

Date: 27/08/2023

(i) **For individual work**

Declaration:

1. I know that plagiarism is wrong. Plagiarism is to use another's work and pretend that it is one's own.
2. I have used the UCT Harvard convention for citation and referencing. Each contribution to, and quotation in, this essay/report/project/ assignment from the work(s) of other people has been attributed and has been cited and referenced. Any section taken from an internet source has been referenced to that source.
3. This essay/report/project/ assignment is my own work and is in my own words (except where I have attributed it to others).
4. I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as his or her own work.
5. I acknowledge that copying someone else's assignment or essay, or part of it, is wrong, and declare that this is my own work.

Signature _____

Contents

Question 1

summary Statistics:

summary(meuse$cadmium)

```
Min.   1st Qu.  Median  Mean   3rd Qu.  Max.
0.20    0.80    2.10    3.25   3.85   18.10
```

```
summary(log(meuse$cadmium))
```

```
  Min.   1st Qu.  Median  Mean   3rd Qu. Max.
-1.609  -0.223   0.742   0.561   1.348   2.896
```

Plots :

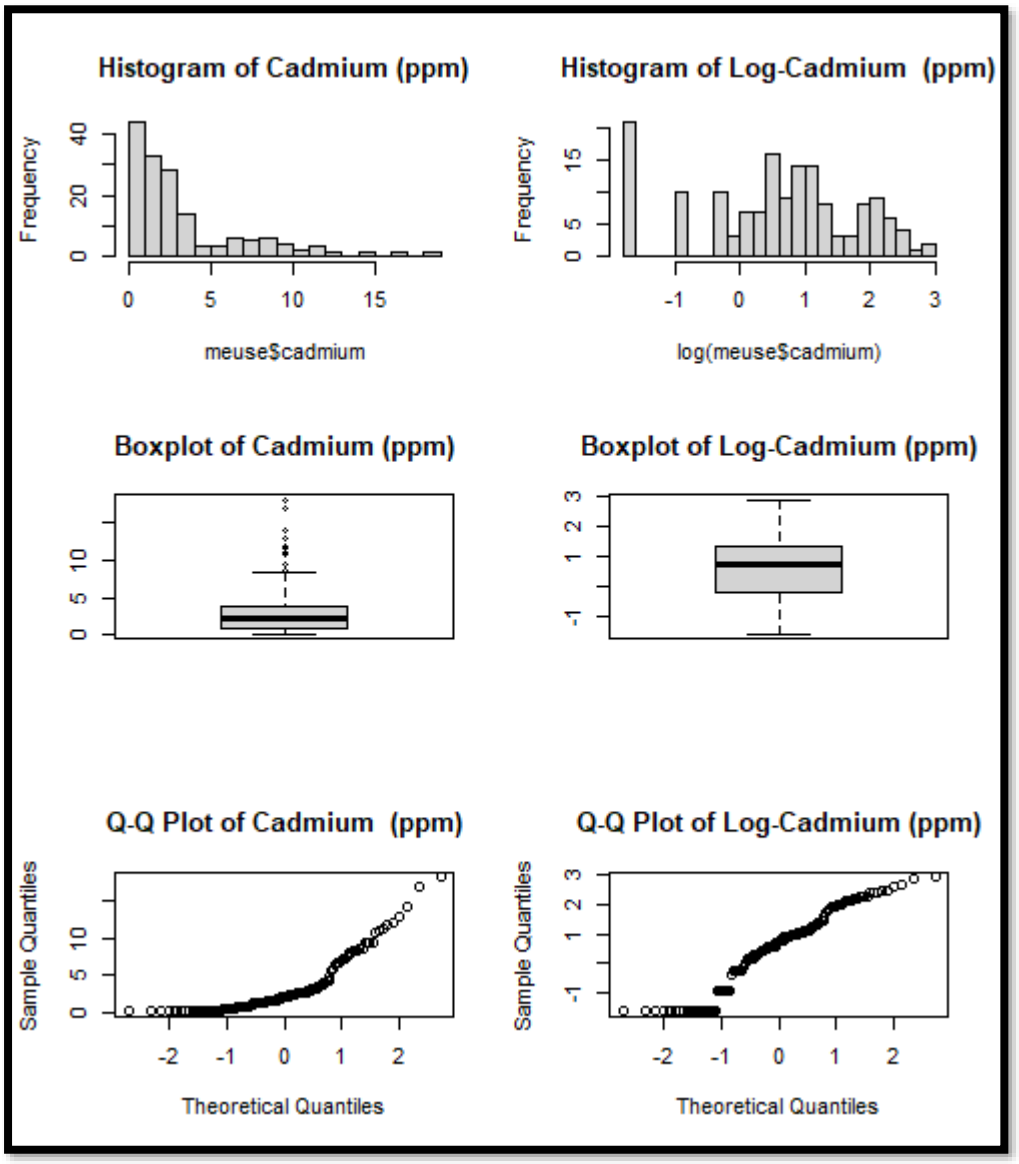


Figure 1 Shows the Cadmium Distribution Analysis

An explanation of each set of summary statistics

Summary of Original Values (meuse$cadmium):

- Minimum (Min.): The smallest value of cadmium in the dataset is 0.20.
- 1st Quartile (1st Qu.): 25% of the data points have values less than or equal to 0.80.
- Median: The middle value of the dataset, where 50% of the data points have values less than or equal to 2.10.
- Mean: The average value of the cadmium variable is 3.25.
- 3rd Quartile (3rd Qu.): 75% of the data points have values less than or equal to 3.85.
- Maximum (Max.): The largest value of cadmium in the dataset is 18.10.

Summary of Log-Transformed Values (log(meuse$cadmium)):

- Minimum (Min.): The smallest log-transformed value of cadmium is approximately -1.609.
- 1st Quartile (1st Qu.): 25% of the log-transformed data points have values less than or equal to -0.223.
- Median: The middle value of the log-transformed dataset, where 50% of the data points have values less than or equal to 0.742.
- Mean: The average value of the log-transformed cadmium variable is approximately 0.561.
- 3rd Quartile (3rd Qu.): 75% of the log-transformed data points have values less than or equal to 1.348.
- Maximum (Max.): The largest log-transformed value of cadmium is approximately 2.896.

additional relevant comments.

Log-Transformed Variable:

• After applying the log transformation (`log(meuse$cadmium)`), the distribution becomes closer to symmetric, as indicated by the mean (0.561) being approximately equal to the median (0.742).

• The range of values is compressed, with a minimum of approximately -1.609 and a maximum of approximately 2.896.

• The interquartile range (IQR), defined by the 1st quartile (-0.223) and 3rd quartile (1.348), suggests that a significant portion of the log-transformed data falls within this range.

Appropriateness of Log Transformation:

The log transformation appears to have improved the symmetry of the data distribution, which is often beneficial for statistical analyses. It has also reduced the range of values and the impact of extreme values. Whether this transformation is appropriate depends on the specific goals of your analysis and the statistical methods you plan to use. In cases where normality or symmetric distribution is an assumption, such as in linear regression, log transformation can be a useful preprocessing step. However, it should be done with careful consideration of the data and the objectives of your analysis.

Additional comments

Distribution Improvement: The log transformation has successfully improved the symmetry of the data. This is important because many statistical methods, like linear regression, assume that the data follow a normal or symmetric distribution. By applying the log transformation, you have made your data more amenable to such analyses.

Range Compression: The log transformation has compressed the range of values. This can be advantageous in cases where you want to reduce the influence of extreme values or outliers in your analysis. However, be aware that transforming the data can also impact the interpretation of results.

Question 2

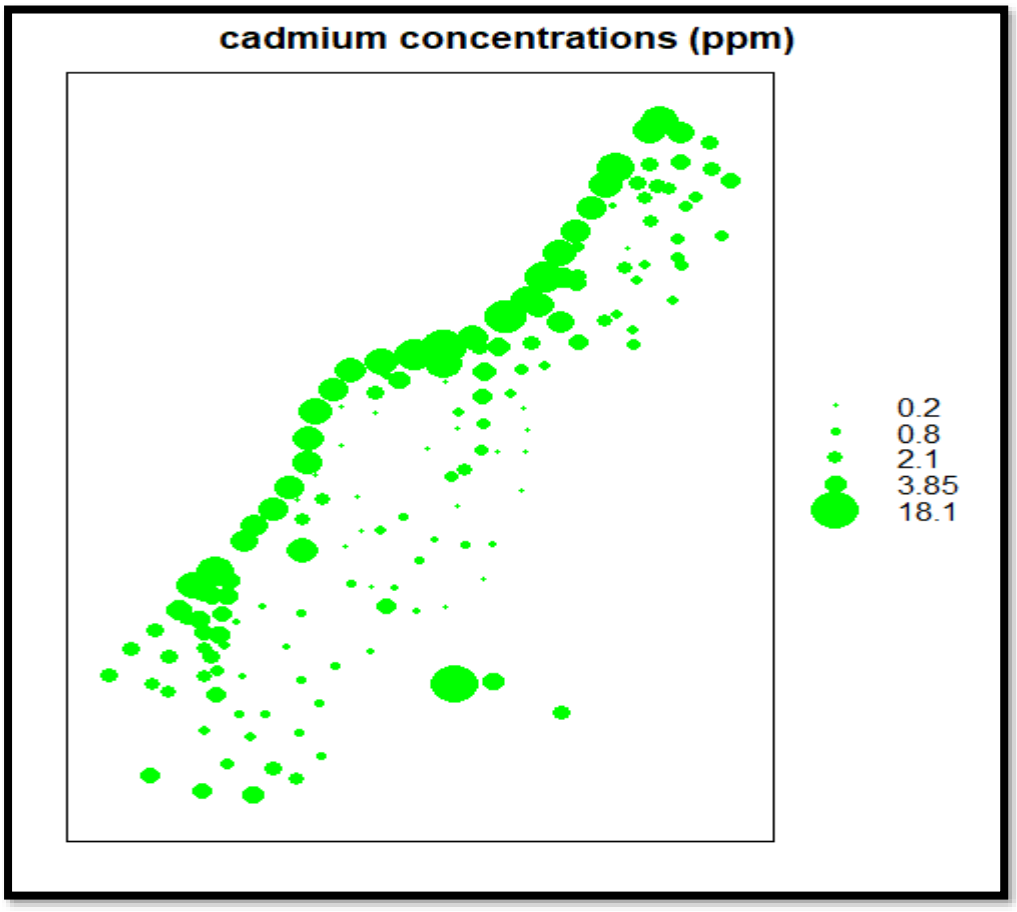


Figure 2 shows the bubble plot of the cadmium data with the dots scaled to concentration.

the bubble plot, with dots scaled based on cadmium concentration levels, indicates the following observations:

1. **Variation in Cadmium Concentrations:** The different sizes of dots indicate a wide range of cadmium concentrations across the observed locations. The larger dots correspond to higher cadmium concentrations, while the smaller dots represent lower concentrations.
2. **Localized High Cadmium Levels:** Areas with the largest dots (18.1 ppm) indicate localized spots with the highest cadmium concentrations. These areas are of particular concern due to their potential environmental impact.

3. **Spatial Clusters:** There are clusters of larger dots, suggesting spatial groupings of higher cadmium concentrations. These clusters provides insights into potential pollution sources or environmental factors influencing cadmium distribution.
4. **Variability:** The varying sizes of dots show that cadmium concentrations are not uniformly distributed. Instead, there is spatial variability, possibly influenced by factors like land use, soil type, and distance to the river.
5. **Gradient:** The gradual increase in dot size from 0.2 ppm to 18.1 ppm implies a gradual increase in cadmium concentration as you move from areas with smaller dots to areas with larger dots.
6. **Hotspots and Low-Concentration Areas:** The largest dots (hotspots) are likely areas of concern for potential contamination, while the smaller dots suggest regions with relatively lower cadmium concentrations.
7. **Presences of outliers** :Notably, a few locations appear to be outliers with exceptionally high cadmium levels.

I can conclude by saying that, the bubble plot suggests a spatially heterogeneous distribution of cadmium concentrations across the floodplain area near the Meuse River. The plot provides valuable insights into the potential areas of higher cadmium pollution and helps researchers and environmental experts identify patterns and potential sources of contamination.

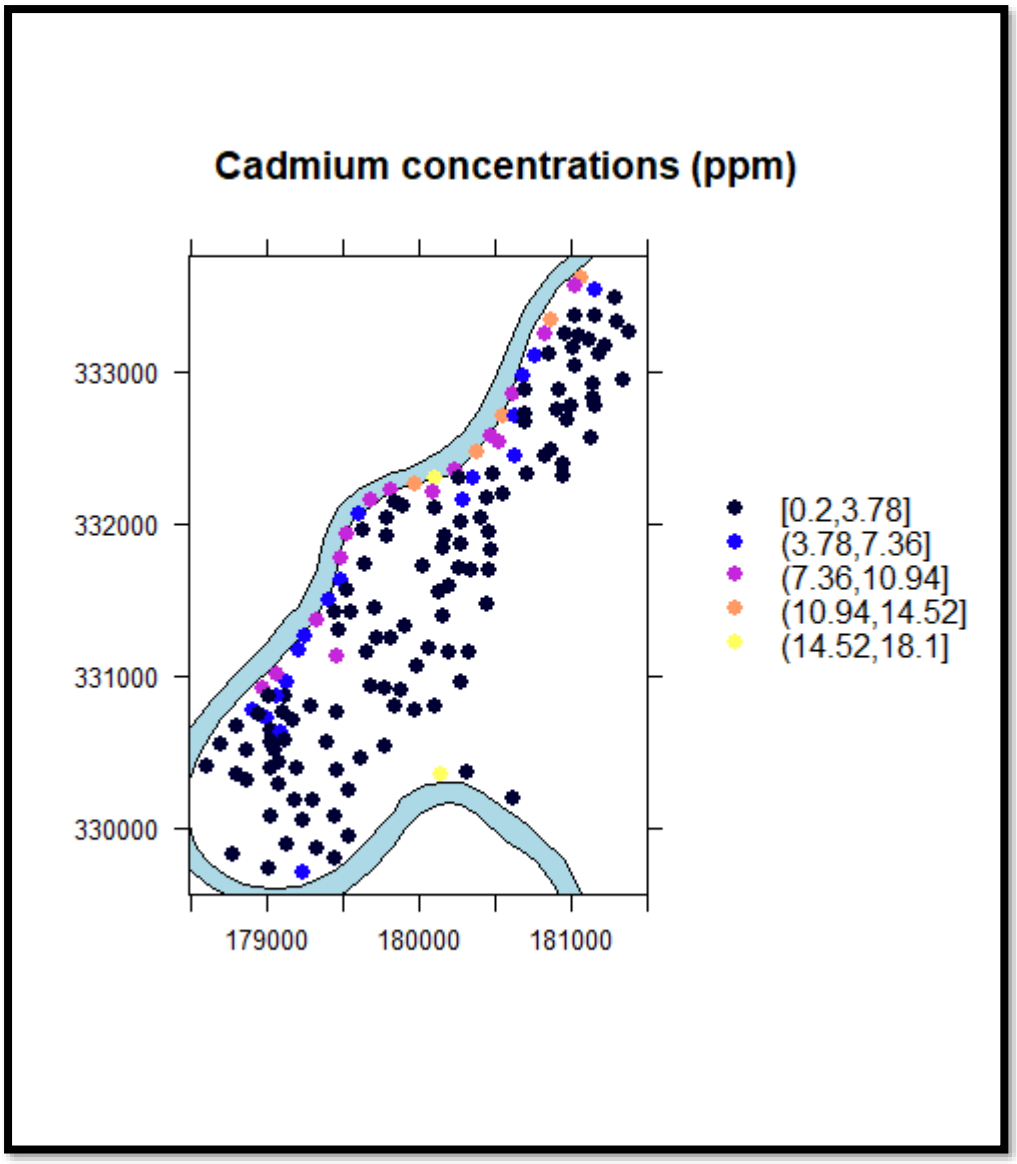
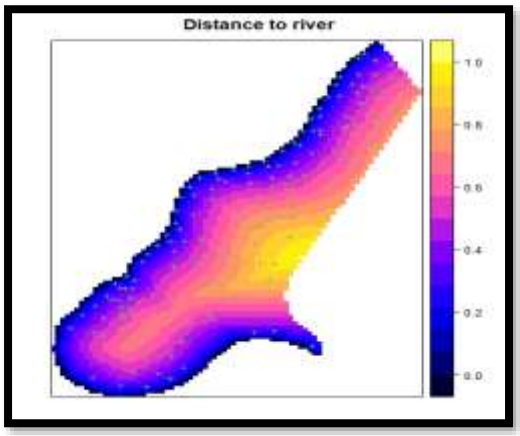


Figure 3 Shows the bubble plot of the cadmium dots scaled to the concentration

it appears that there is a spatially varying distribution of cadmium across the floodplain area near the Meuse River.

the second plot, a dot plot with geographic context, superimposes the cadmium concentration data onto a map of the Meuse River area, with dots' colour intensity reflecting cadmium levels. This plot provides valuable geographical context by displaying the Meuse River boundaries in light blue, aiding in understanding how cadmium concentrations relate to the river's geography. It becomes evident that cadmium concentrations vary both within and outside the river boundaries, with some regions near the river showing elevated cadmium levels. These visualizations collectively enable a better understanding of the spatial distribution of cadmium in the region and highlight areas of interest for further investigation or environmental assessment.

Question 3



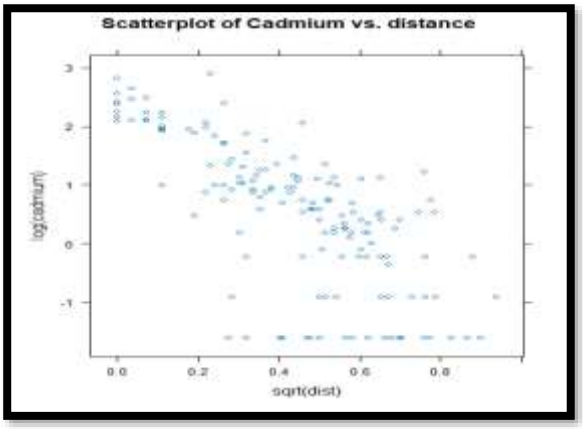
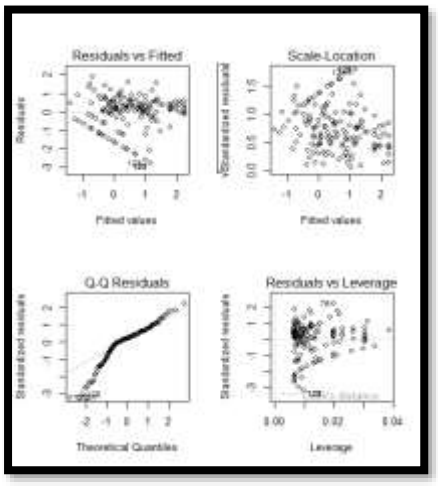


Figure 4 shows the Map of distance to river

Figure 5 shows the Scatterplot of Cadmium vs. distance.



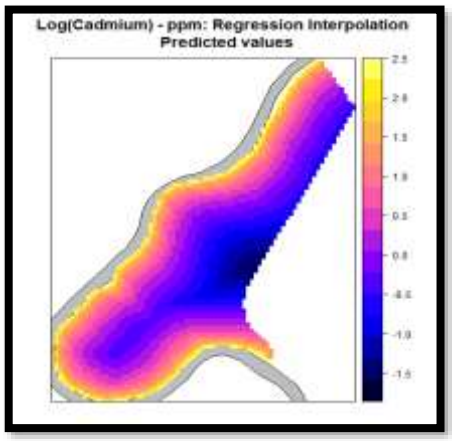


Figure 6 shows the diagnostic plots

Figure 7 shows the Log(Cadmium) - ppm: Regression Interpolation Predicted values

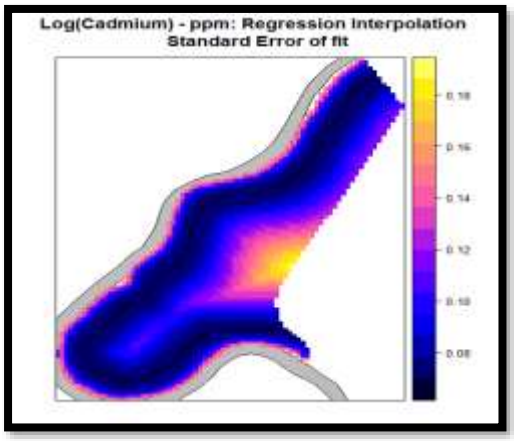


Figure 8 shows the Log(Cadmium) - ppm: Regression Interpolation Standard Error of fit

Nature of the Relationship:

The scatterplot of Cadmium concentration vs. the square root of the distance to the river provides an initial understanding of the relationship between these variables. The scatterplot indicates whether there is a linear or non-linear trend and the presence of any outliers.

Model Fit and Summary:

The regression model, as summarized by summary(cadmium.lm), provides key information about the model fit.

Coefficients: The coefficients section shows the estimated coefficients for the intercept and the square root of the distance to the river. In this case, the intercept is 2.232, and the coefficient for the square root of the distance is -3.842. These coefficients represent the estimated relationship between Cadmium concentration and distance to the river.

Significance: The p-values associated with the coefficients indicate whether each coefficient is statistically significant. In this case, both coefficients are highly significant ($p < 0.001$), suggesting that distance to the river is a significant predictor of Cadmium concentration.

R-squared: The multiple R-squared value (0.504) represents the proportion of the variance in Cadmium concentration explained by the model. This indicates that about 50.4% of the variation in Cadmium concentration can be explained by the distance to the river.

Residuals: The residual standard error (0.866) provides a measure of the typical error in predictions made by the model. It represents the spread of data points around the regression line.

3. Regression Diagnostics:

The diagnostic plots generated by plot(cadmium.lm, add.smooth = FALSE) are essential for assessing the assumptions underlying linear regression:

Residuals vs. Fitted: This plot helps check for linearity and homoscedasticity. Here the spread of residuals varies systematically across the range of fitted values, which suggests heteroscedasticity

Normal Q-Q: This plot assesses the normality of residuals. Ideally, points should follow a straight line. My plot shows a slightly curved pattern instead of a straight line, it suggests a potential departure from the normality assumption.

Scale-Location: This plot helps check for homoscedasticity. Ideally, the points should be evenly spread along the horizontal line, indicating constant variance of residuals. Here the Scale-Location plot shows scattered points, it suggests potential heteroscedasticity, which means that the spread of residuals varies systematically across the range of fitted values.

Residuals vs. Leverage: This plot identifies influential data points. High-leverage points can disproportionately affect the regression model. Here the Residuals vs. Leverage plot shows scattered points, it suggests that there may be data points with high leverage but without particularly large residuals

```
a summary of the regression model
summary(cadmium.lm)

Call:
lm(formula = log(cadmium) ~ sqrt(dist), data = meuse)

Residuals:
     Min       1Q   Median       3Q      Max
-2.784 -0.254  0.180  0.501  1.908

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.232      0.151    14.8   <2e-16 ***
sqrt(dist)    -3.842      0.308   -12.5   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.866 on 153 degrees of freedom
Multiple R-squared:  0.504,    Adjusted R-squared:  0.501
F-statistic:  155 on 1 and 153 DF,  p-value: <2e-16
```

Question 4

Trend Surface Analysis

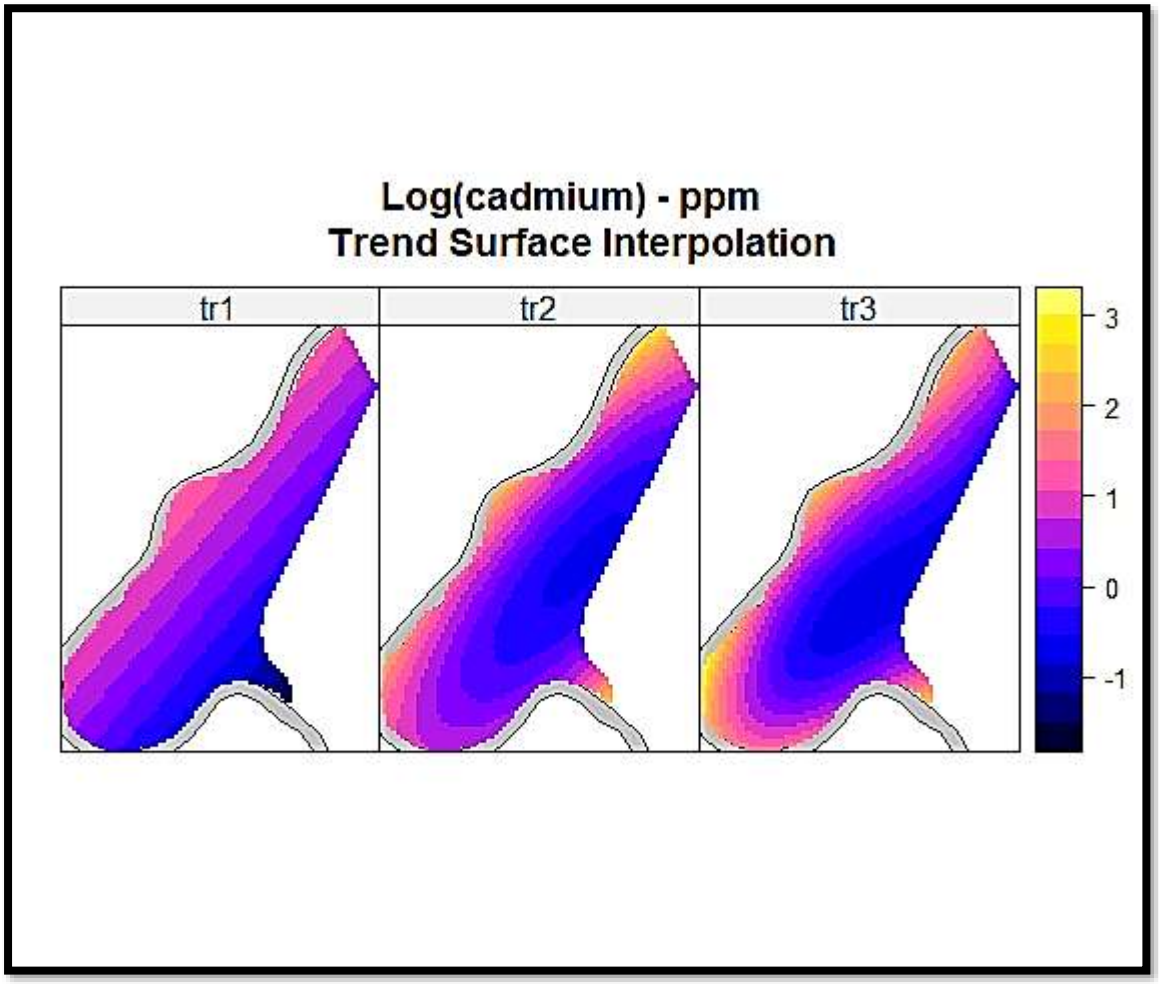


Figure 9 shows the Log(cadmium) - ppm Trend Surface Interpolation

When discussing the results and comparing the three interpolation maps, it's important to consider the appropriateness of trend orders with respect to the nature of the interpolated surface, the quality of interpolation concerning the distribution and density of control points, and the overall fit of the model.

1. **Trend Surface Orders:**
    - **Degree 1:** A linear trend surface assumes a flat surface with constant slope. It might capture simple trends but could struggle with capturing curvature or more complex patterns.
    - **Degree 2:** A quadratic trend surface introduces curvature and could provide a better fit if the true surface has a parabolic shape.
    - **Degree 3:** A cubic trend surface further introduces higher-order curvature and flexibility in fitting complex surfaces**.**

2. **Interpolation Quality:**

    - **Degree 1:** This is appropriate if the data suggests a relatively simple and linear trend, and control points are sparsely distributed.
    - **Degree 2:** This could be more appropriate when there's curvature in the underlying surface, and control points are moderately distributed.
    - **Degree 3:** A cubic surface might overfit noisy data or lead to erratic results if control points are sparse. It could be useful when the data truly exhibits complex curvatures.

3. **Comparative Analysis:**

    - In comparison tr1 being a product of linear trend surface its exhibit smoother pattern compared to tr2 & tr3.also tr2 being a product of quadratic trend surface exhibit a comparative smoother surface than tr3 & lastly tr3 being a product of cubic trend surface exhibiting clustering creating a less smooth surface.
    - Degree 2 provides a better fit since trend surface has a parabolic surface compared to degree1 & degree 3.

My own code that plots degree 2 & 3

```
# Trend surface up to degree 2
meuse.grid$tr2 = krige(log(cadmium) ~ 1, meuse, meuse.grid, degree = 2)$var1.pred

# Trend surface up to degree 3
meuse.grid$tr3 = krige(log(cadmium) ~ 1, meuse, meuse.grid, degree = 3)$var1.pred

# my own code Plotting degree 2 and 3 trend surfaces
spplot(meuse.grid, c("tr2", "tr3"), sp.layout = meuse.lt,
    main = c("Degree 2 Trend Surface", "Degree 3 Trend Surface"))
```

Question 5

Plots of the outputs and the control points

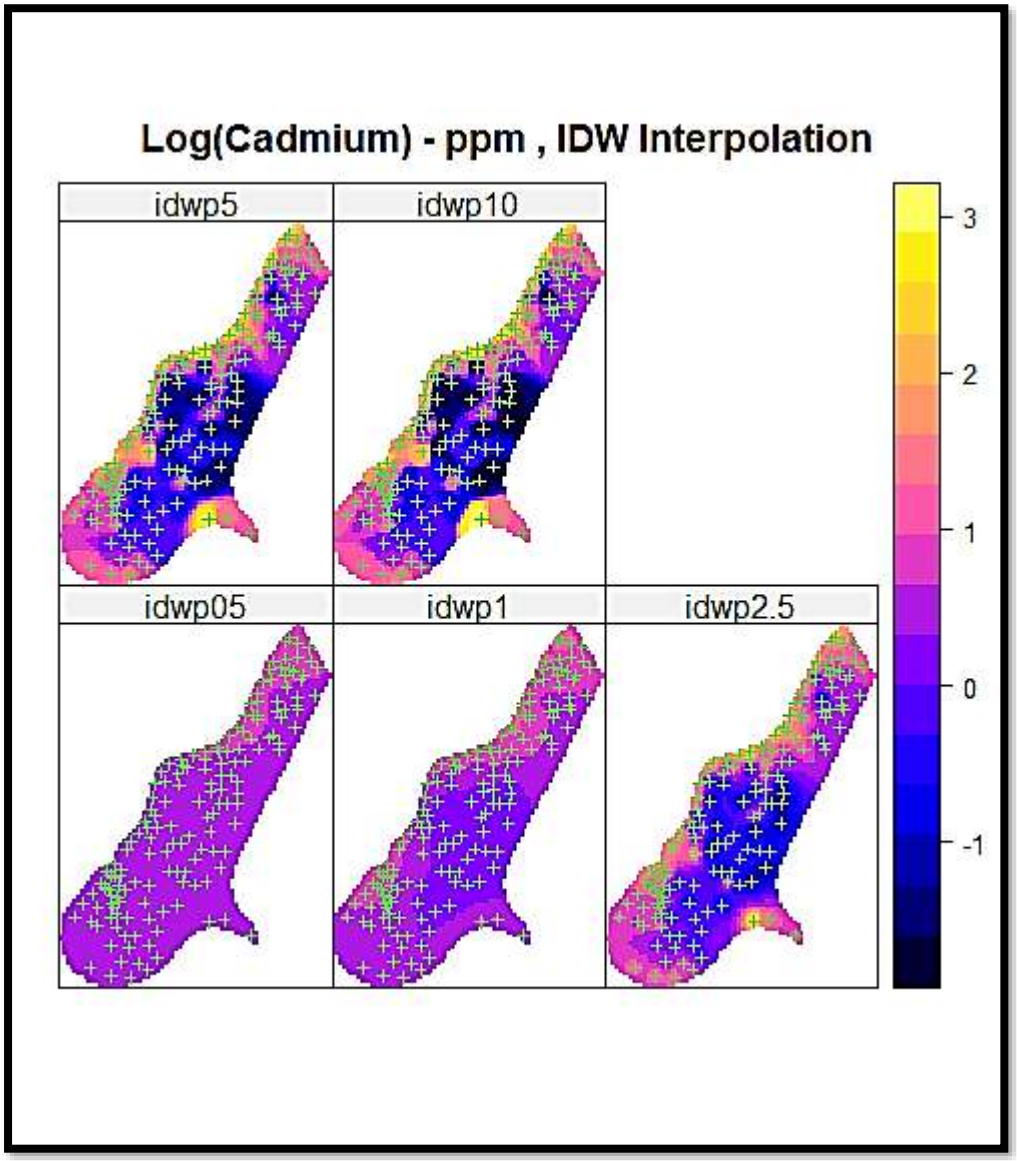


Figure 10 shows the Log (Cadmium) - ppm , IDW Interpolation

discussion considering the role of the power function (p) and the quality/behaviour of the predictions in relation to the distribution and density of control points:

Role of Power Function (p): The power function (p) in Inverse-Distance Weighting (IDW) interpolation plays a critical role in determining the influence of nearby data points on the interpolated values. Smaller values of p give more weight to closer points, resulting in localized predictions that closely match the available data points. These smaller p values are suitable for capturing fine-scale variations and rapid changes in the phenomenon being interpolated. However, this can also lead to overfitting, where the model captures noise or random fluctuations in the data.

On the other hand, larger values of p assign similar weights to more distant points, yielding smoother and more gradual interpolated surfaces. These surfaces represent a more generalized view of the phenomenon and can help highlight broader trends and patterns. However, using high p values may lead to the smoothing out of important local variations, potentially resulting in underfitting and loss of detail**.**

Quality/Behaviour of Predictions: The quality and behaviour of IDW predictions depend on the distribution and density of control points.

1. **Dense Control Points**: In regions with a dense distribution of control points, interpolation tends to perform well regardless of the chosen p-value. The abundance of nearby data points provides a reliable basis for prediction, and variations in p might not significantly impact the outcome. Predictions tend to closely resemble observed values, and small-scale variations are captured more accurately.
2. **Sparse Control Points:** In areas with sparse control points, the choice of p becomes more critical. Using a smaller p might lead to overly complex surfaces that interpolate to individual data points, introducing noise and reducing the quality of predictions. Conversely, larger p values offer smoother surfaces but may overlook localized variations**.**

In summary, the role of the power function (p) in IDW interpolation is to balance between capturing local details and providing a generalized view of the phenomenon. The quality and behaviour of predictions are highly influenced by the density and distribution of control points. For areas with sufficient control points, different p values might yield consistent results. However, in regions with sparse control points, careful consideration of p is necessary to avoid overfitting or underfitting, ensuring that the interpolation accurately reflects the underlying patterns while maintaining a balance between local and global trends**.**

Question 6

Plots:

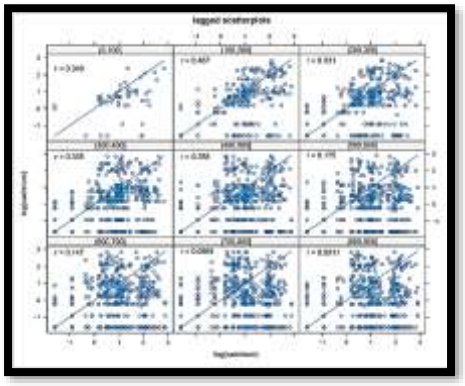


Figure 11 shows the lagged scatterplots.

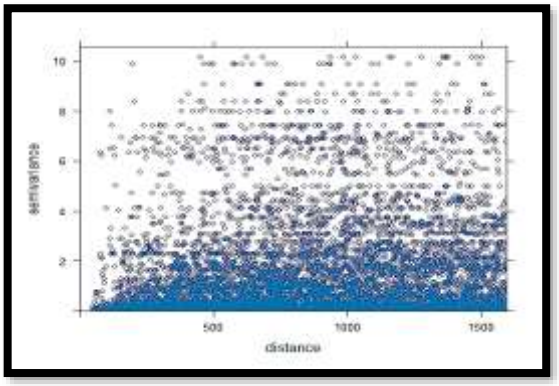


Figure 12 show the a variogram cloud

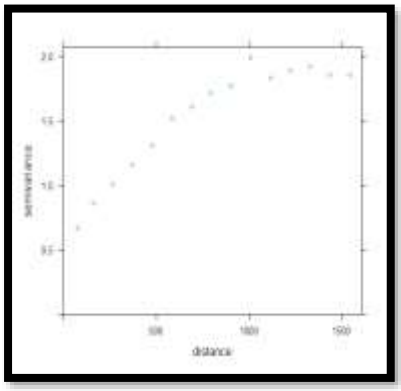


Figure 13 shows the (binned variogram) plot

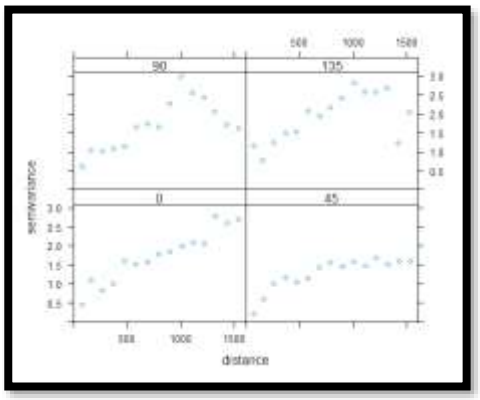
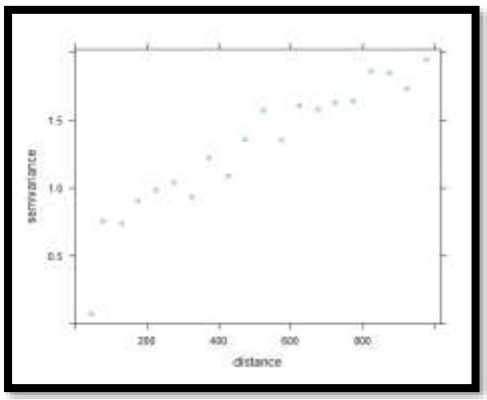


Figure 14 shows the variograms at four different angles (0, 45, 90, 135)



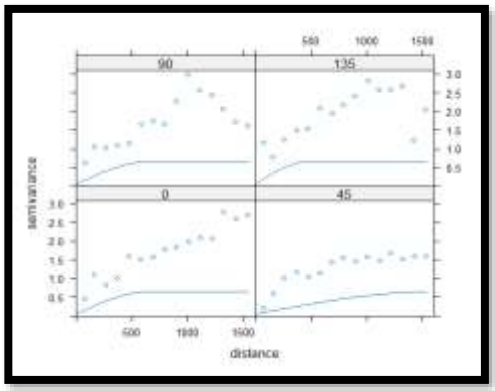


Figure 15 shows the a variogram plot

This is a lagged scatterplots, A simple way to acknowledge that spatial correlation is present or not is to make scatter plots of pairs Z(si) and Z (sj), grouped according to their separation distance.

assumptions

- here r is close to 0, so it suggests a weak or no linear relationship between the variables. In this case, changes in one variable do not systematically correspond to changes in the other variable.
- The slope of the line of best fit indicates the strength and direction of the linear relationship. If the slope is positive, it suggests a positive linear relationship, meaning that as the lagged variable increases, the current variable tends to increase as well.
- Random Residuals: In a good linear fit, the residuals should look like random scatter points around the line of best fit. This means that some residuals will be positive (above the line), and some will be negative (below the line), but they should be scattered fairly evenly. There should be no apparent pattern or trend in the residuals.

A semivariogram, also known as a variogram, is a fundamental tool in geostatistics and spatial statistics used to analyse the spatial variability and dependence of a variable across a geographic or spatial domain. Characterizing Spatial Dependence: Semivariograms help you understand how the similarity or dissimilarity of data values varies with spatial separation. In other words, they quantify the degree of spatial dependence or autocorrelation in the data.

Assumptions

No Spatial Dependence: A consistent semi variance as distance increases suggests that there is little to no spatial dependence in the data. In other words, the values of the variable at different locations are not influenced by their spatial proximity. This pattern is sometimes referred to as spatial randomness.

Spatially Homogeneous: The data exhibit spatial homogeneity, meaning that the variable's values are distributed uniformly across the study area without any discernible spatial structure.

```
Output: variogram(log(cadmium) ~ 1, meuse, boundaries = c(0, 50, 100, seq(
250, 1500, 250)))
    np   dist  gamma dir.hor dir.ver   id
1    2   46.6 0.0683       0       0 var1
2   50   78.2 0.7490       0       0 var1
3  442  184.4 0.8917       0       0 var1
4 1107  379.7 1.1332       0       0 var1
5 1317  626.8 1.5495       0       0 var1
6 1341  874.6 1.7975       0       0 var1
7 1190 1122.6 1.8848       0       0 var1
8 1057 1375.1 1.8732       0       0 var1
```

```
the fitted model parameters
```

```
  model psill range
1   Nug 0.548     0
2   Sph 1.340  1149
```

```
fit.variogram(v, vgm(1, "Sph", 800, 0.06), fit.sills = c(FALSE, TRUE))
```

output:

17

```
  model psill range
1   Nug  0.06     0
2   Sph  1.49   496
```

```
Performing REML fitting of a variogram model

Output:
  model psill range
1   Nug 0.424     0
2   Sph 1.317   800
```

```
Plotting experimental variograms and the fitted anisotropic model
```

Directional sample variogram (plus) and fitted model (dashed line) for

four directions(0 is north & 90 is east)

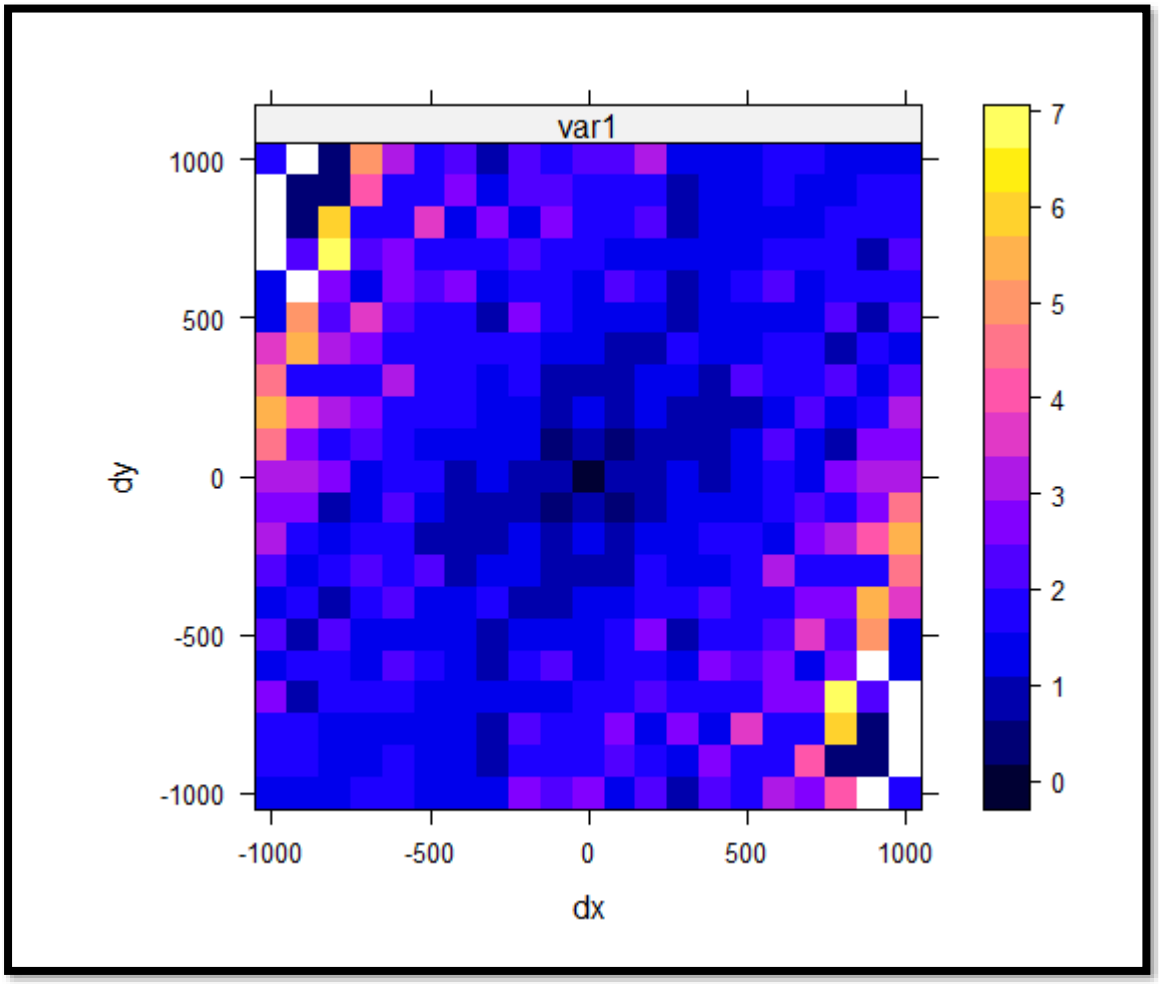


Figure 16 shows the simple kriging with a specified 'beta'

Question 7

To assess the quality of Simple and Ordinary Kriging Interpolation, you can generate additional output and perform various checks.

1. Cross-Validation: One of the most common ways to assess the quality of kriging interpolation is through cross-validation. This involves leaving out a portion of your data, interpolating those points using the remaining data, and then comparing the interpolated values to the actual values that were left out. Common metrics for assessing accuracy include Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). A low error indicates a better fit.

2. Residual Analysis: Examine the residuals, which are the differences between the observed values and the predicted values from the kriging model. Residual analysis can help identify patterns or trends in the errors. Ideally, the residuals should be random and have no systematic patterns. If patterns exist, it suggests that the model may be missing important spatial structures.

3. Cross-validation Plots: Plot the observed versus predicted values for the validation data points. This can help visualize how well the kriging predictions match the actual data. Ideally, the points should be closely clustered around the 45-degree line, indicating a good fit.

4. Variogram Model Fit: Check how well the variogram model you've chosen fits the empirical variogram. A good fit indicates that your chosen model adequately captures the spatial dependence in the data.

5. Quantify Uncertainty: Kriging provides not only predictions but also estimates of uncertainty. Investigate kriging standard errors or confidence intervals associated with the predictions. Wider intervals suggest higher uncertainty.

6. Sensitivity Analysis: Perform sensitivity analyses by varying the parameters of the kriging model (e.g., the variogram model parameters, the sill, or the nugget effect). Assess how changes in these parameters affect the interpolation results. This can help you understand the robustness of your predictions.

7. Visual Inspection: Visually inspect the kriging interpolation maps alongside the original data and any available ground truth data. Look for spatial patterns, trends, and clusters. Check if the interpolated surfaces make sense given the known spatial characteristics of the phenomenon you're studying.

8. Outliers and Anomalies: Identify and investigate any outliers or anomalies in your data. These points can significantly impact kriging results. It's important to assess whether these outliers should be included, excluded, or treated differently in your analysis.

9. Comparative Analysis: If you have multiple kriging models (e.g., Simple Kriging vs. Ordinary Kriging), compare their results in terms of accuracy and uncertainty. This can help you determine which method performs better for your specific dataset.

10. Cross-Validation Techniques: Consider different cross-validation techniques, such as k-fold cross-validation, to assess the stability of your kriging model and ensure that it is not overfitting or underfitting the data.

By exploring these extra results and analyses, I can comprehensively assess the quality and reliability of the Simple and Ordinary Kriging interpolations. These methods offer me a complete perspective on how effectively the interpolations capture the fundamental spatial patterns. This information helps me make informed judgments about whether these techniques are appropriate for my specific dataset and analysis goals.

Question8

Visual Comparison of Interpolation Results: Simple Kriging (SK) vs. Ordinary Kriging (OK)

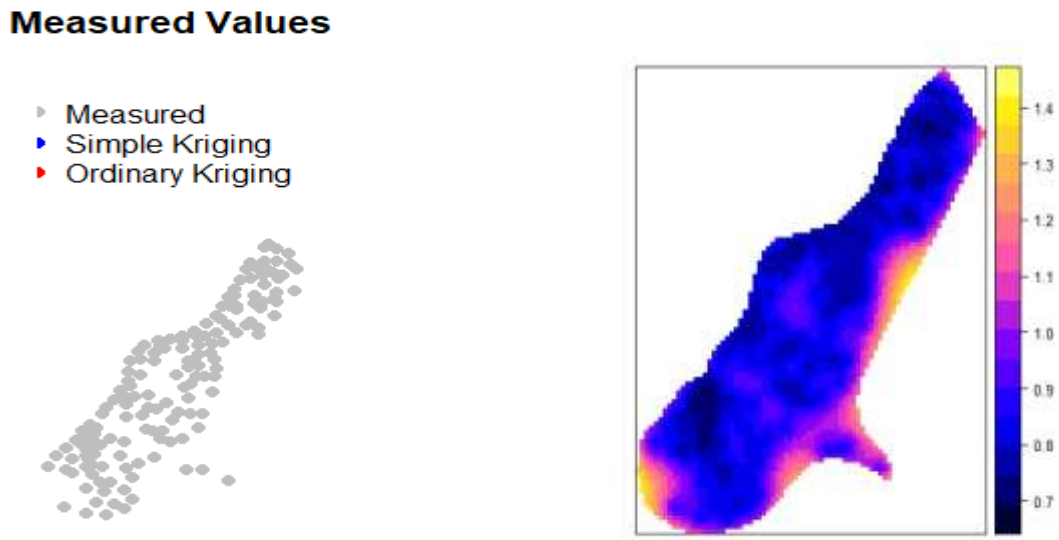


Figure 17 shows predicted output

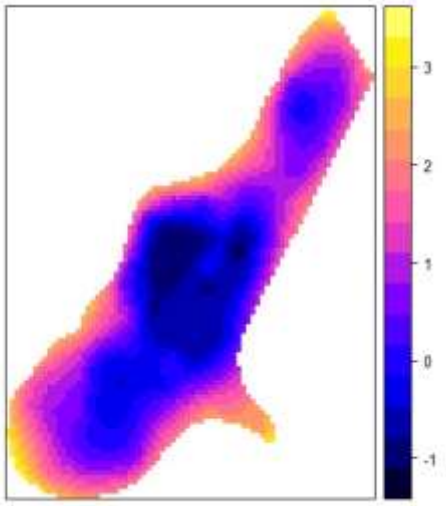


Figure 18 shows original input

Observations and Discussion:

Distribution of Control Points: Examine how the control points (measured values) are distributed across the study area. Note if there are any clusters or gaps in the control point distribution.

Behavior of Interpolated Values: Compare the behavior of interpolated values from Simple Kriging and Ordinary Kriging. Pay attention to areas where there are differences in predicted values.

Spatial Smoothing: Consider whether one method appears to exhibit more spatial smoothing than the other. Spatial smoothing refers to the extent to which the predicted values vary across space. A smoother surface indicates higher spatial correlation in the interpolation.

Outliers: Check for any outliers in the interpolated values. Outliers are values that significantly differ from their neighbors and may indicate issues with the interpolation method.

Local vs. Global Trends: Identify if either method captures local trends or global trends better. Local trends refer to fine-scale variations in the data, while global trends represent broader patterns.

Performance in Data-Sparse Areas: Assess how each method performs in areas with fewer control points. Do they exhibit different behaviors in regions with sparse data?

Prediction Uncertainty: Consider the prediction variances you calculated earlier. Are there areas where one method provides more certain predictions than the other?

Overall Quality: Based on your observations, discuss which interpolation method (Simple Kriging or Ordinary Kriging) appears to be more suitable for the specific dataset and spatial characteristics.

Question 9

Comparing various interpolation methods for the Meuse River dataset involves a multifaceted evaluation approach. This includes techniques such as cross-validation to gauge generalization performance, diagnostic analysis to assess model fit, and performance metrics like RMSE and MAE

for quantitative comparison. Sensitivity analysis aids in understanding method robustness, while case studies and expert input offer real-world context. Visualization facilitates visual comparison, and uncertainty assessment helps evaluate result reliability. Exploring ensemble methods can harness the strengths of different techniques. By integrating these methods, a comprehensive understanding of each interpolation method's suitability and performance can guide informed decisions for specific analytical objectives and dataset characteristics**.**

Resources

Chapter 15: Introduction to Geographic Information Systems [9th ed.] Kang-tsung Chan. A copy of this is available on Amathuba inside the Resources folder, Study Material & Books.

Chapter 8: Applied Spatial Data Analysis with R by Bivand et. al. (2008). A copy of this is available on Amathuba inside the Resources folder, Study Material & Books.