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

MPNFLA001

Plagiarism Declaration

Declaration: We (names) Flassie Mpanza .... are the authors of this work, using our own words (except where attributed to others) We know that plagiarism is to use another's work and pretend that it is one's own, and that this is wrong. We have used APG 4013C Notes .... convention for citation and referencing. We have provided citations and references in all cases where we 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**
Flassie Mpanza	MPNFLA001	Everything

[paste a copy of your signatures here]

Signed: FS Mpanza

Date: 28/08/2023

The three standard declarations are as follows:

(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 .APG 4013C Notes convention for citation and referencing. Each contribution to, and quotation in, this essay/report/project/ ......report.... 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/......report............. 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 FS Mpanza

(ii) For group work resulting in a single authored essay/report/project

(Where each member of the group submits his/her own essay/project/report and a separate declaration) **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 ....APG 4013C Notes...........convention for citation and referencing.
3. The text of this essay/project/report is my own work, using my own words (except where attributed to others).

This essay/project/report uses the work of the group (list names).

Flassie Mpanza

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. Each group member has undertaken to submit his/her own essay/project/report and acknowledge the work of other group members.

I acknowledge that by copying someone else's assignment or essay, or a part of it, is wrong, and declare that this essay/report/project is **my own work**, and is based on the work of the group.

Signature FS Mpanza

(iii) For group work resulting in a multiple-authored essay/project

(One declaration signed by all authors to be submitted with the essay/assignment/project)

Declaration:

We (names) Flassie Mpanza are the authors of this work, using our own words (except where attributed to others)

We know that plagiarism is to use another's work and pretend that it is one's own, and that this is wrong.

We have used APG 4013C Notes convention for citation and referencing. We have provided citations and references in all cases where we have quoted from the work of others, or used other's ideas or reasoning in this essay/project/report.

The writer of each of the sections of this essay/report/project is listed below:

Name	Section(s) authored
Flassie Mpanza	Everything

Signatures of all authors: FS Mpanza

NOTES:

(i) Individual academics and departments may adapt this declaration to suit particular needs, and may apply the rule with discretion. Declarations submitted in the case of group work assignments must be suitably adapted.

(ii) The declaration must be used for any substantial work that a student does unsupervised.

(iii) The declaration is neither suitable for reports submitted at the end of a three-hour science laboratory session, nor required in such cases.

(iv) A suitably altered version should be used for students in first-year (Foundation) courses submitting an assignment before having been taught about referencing conventions.

The declaration must be used regularly. It is not sufficient to require students to sign the declaration on annual basis. Students need to be constantly reminded. The onus is on academics to ensure that before work is marked, it is accompanied by signed declaration in the standard form, or if the lecturer has provided one, in a contextually altered form.

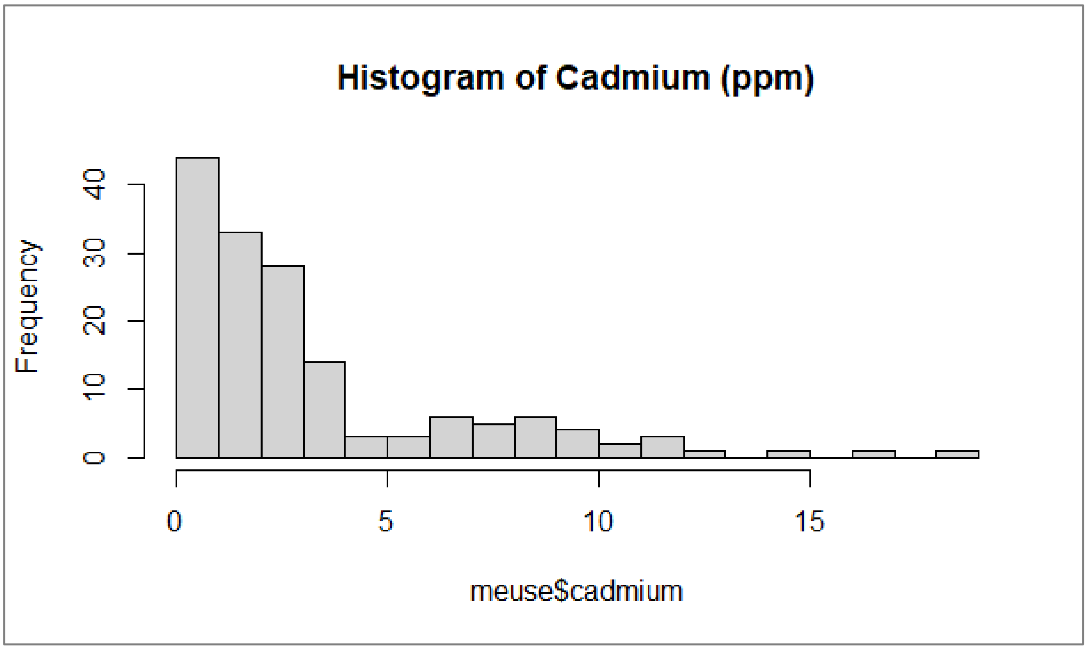
Question 1

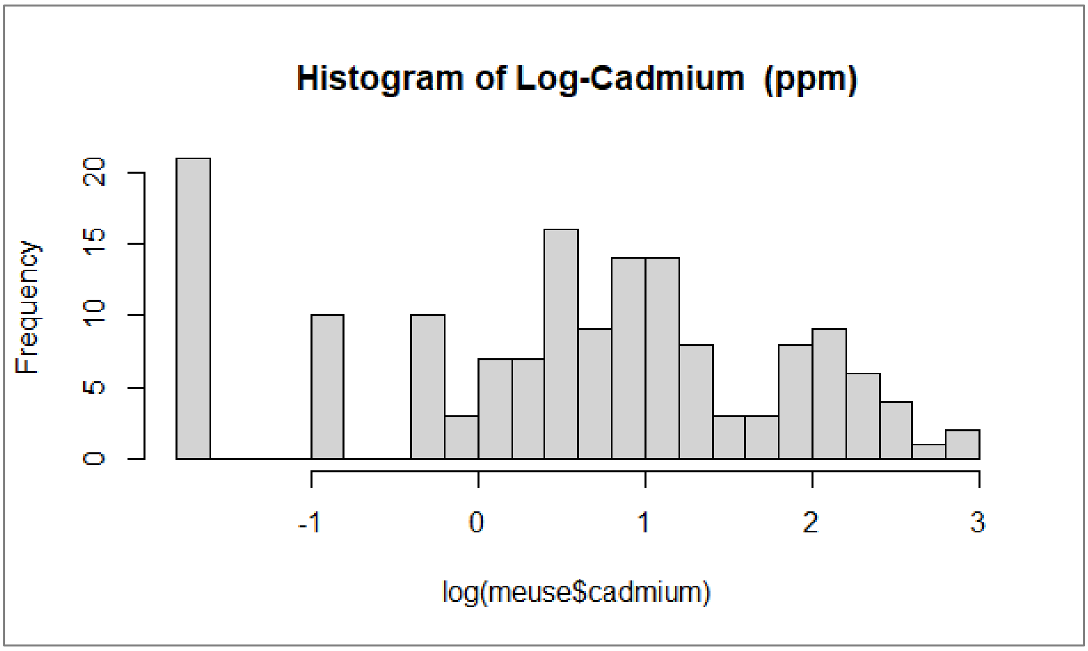
Summary Statistics

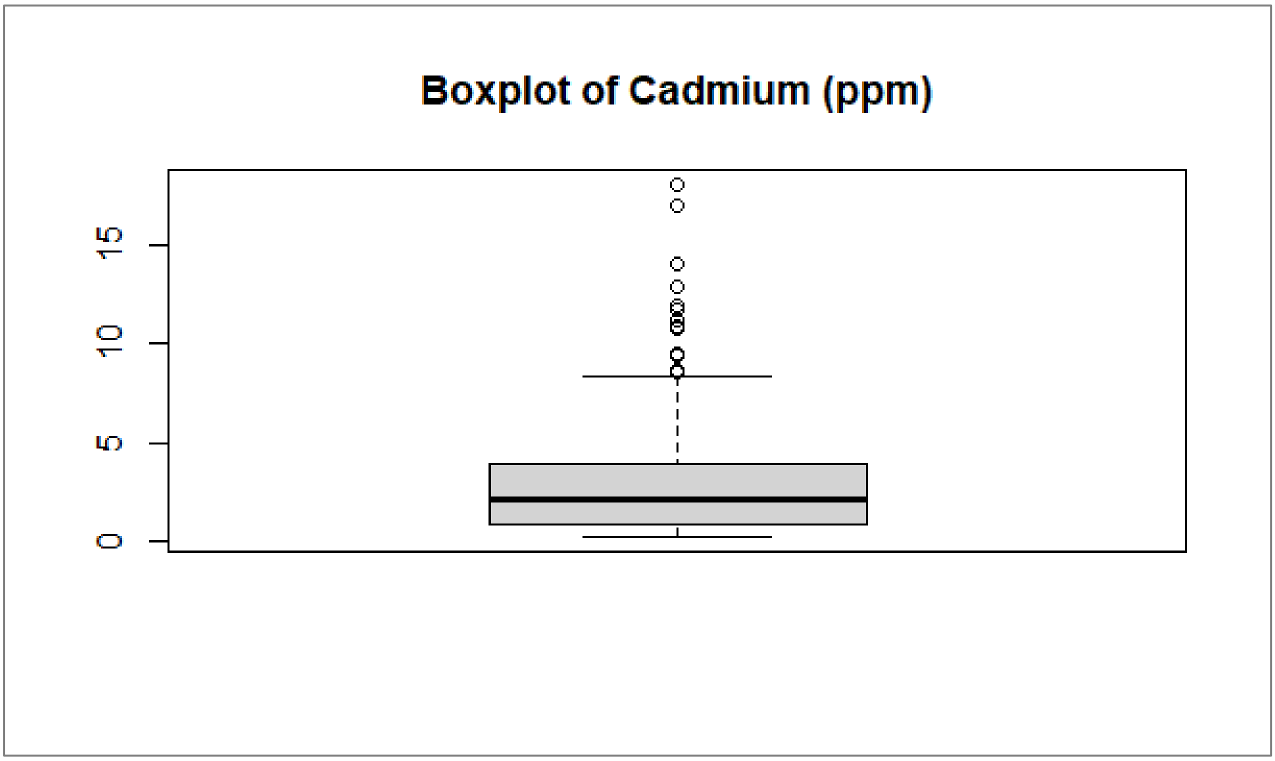
```
summary(meuse$cadmium)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.20    0.80    2.10    3.25    3.85   18.10
```

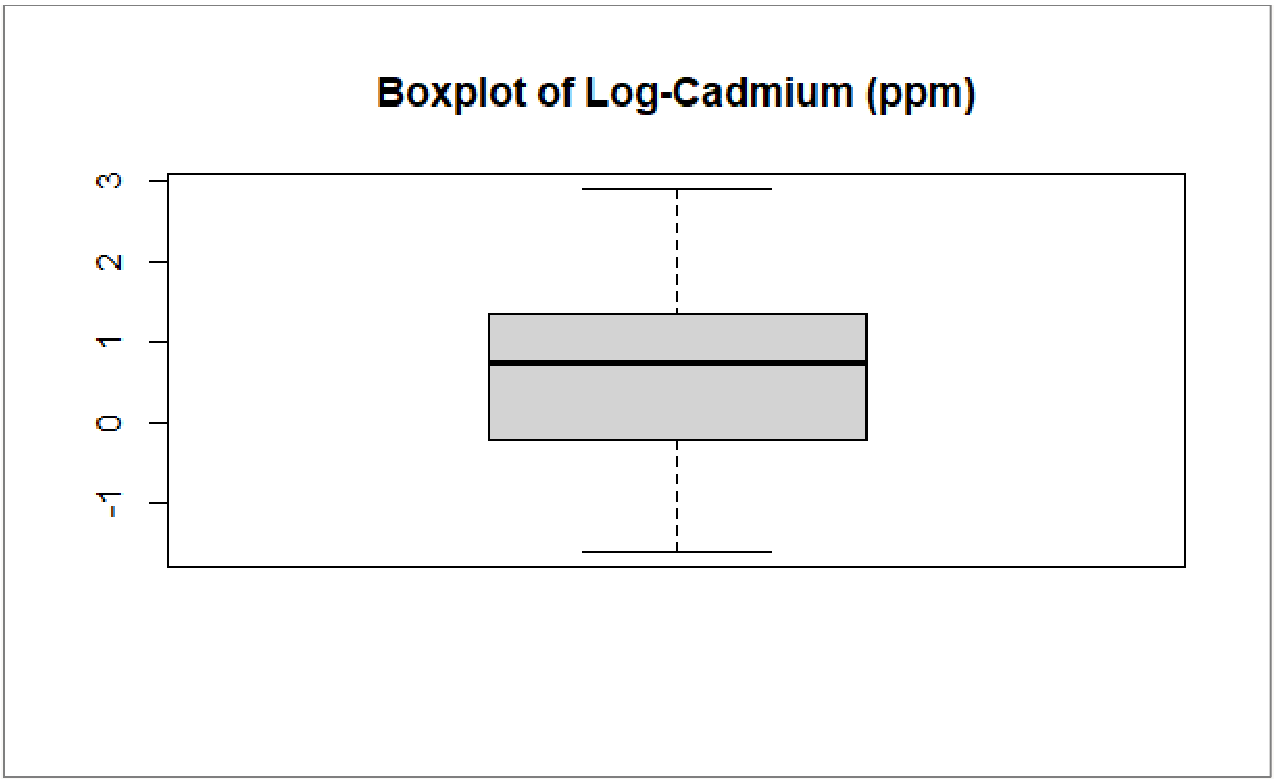
```
summary(log(meuse$cadmium))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -1.609  -0.223   0.742   0.561   1.348   2.896
```

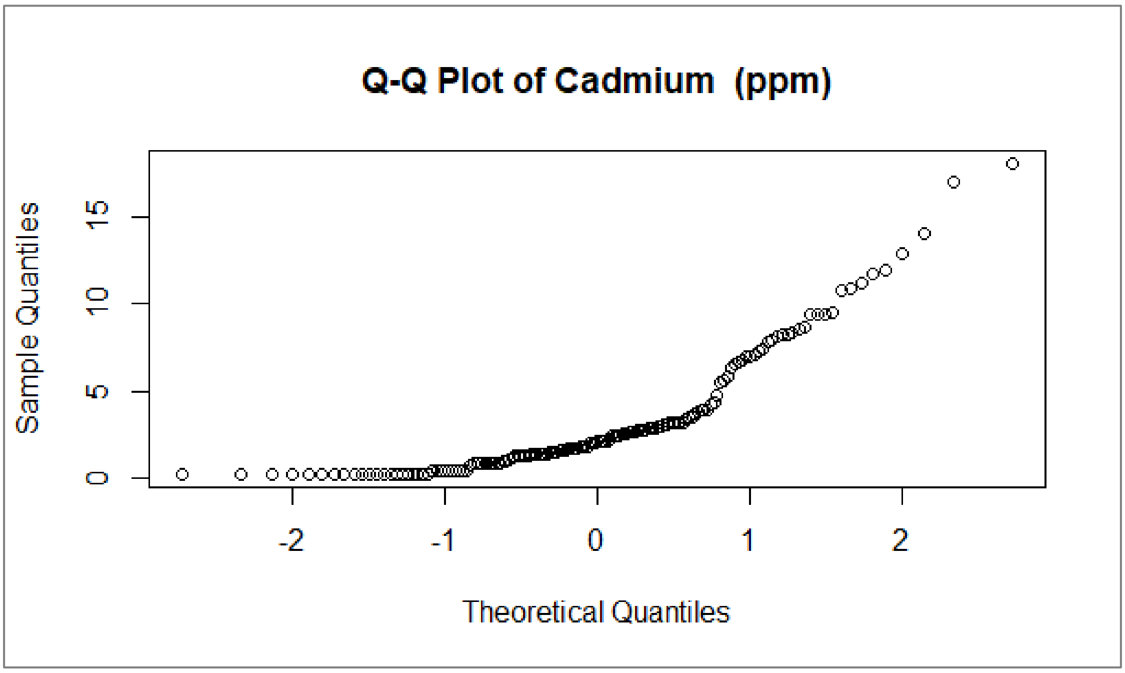
Plots

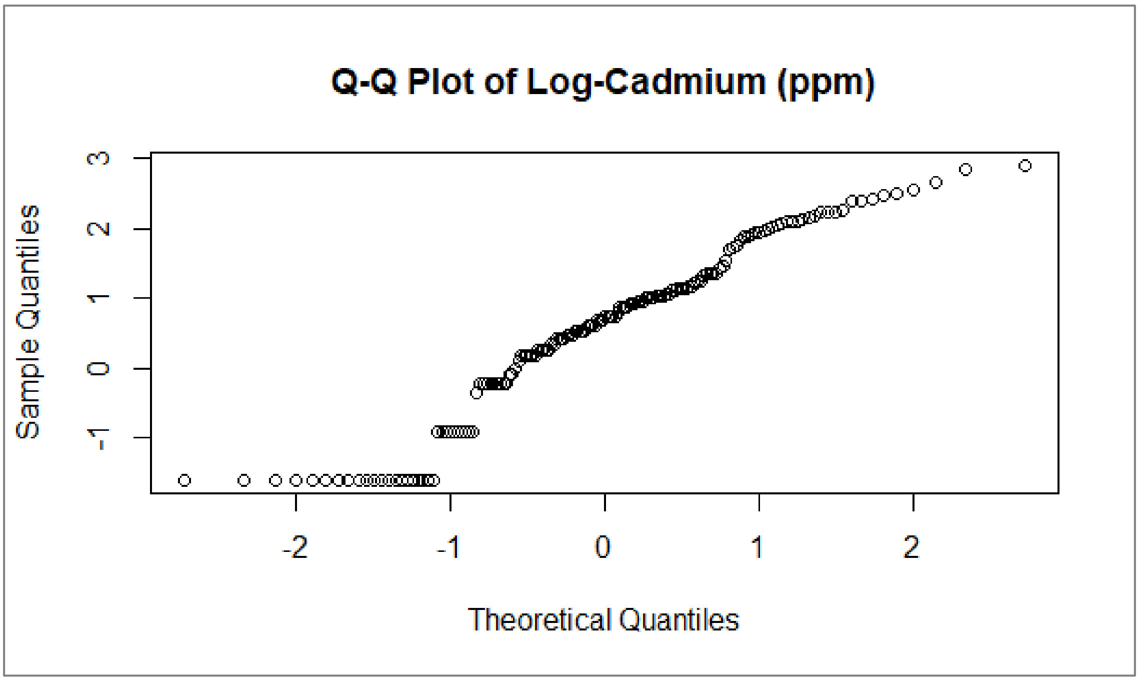





Boxplot of Cadmium (ppm)


Boxplot of Log-Cadmium (ppm)





Results (summary statistics and plots)

Stem-and-leaf plots offer a more detailed view of data distribution. The original "Cadmium" plot shows a right-skewed distribution with a few significantly higher values. The log-transformed plot is more symmetric and centred around 0.

Histograms provide a detailed data distribution view. The original "Cadmium" variable has a positively skewed distribution with a long right tail due to higher values. The log-transformed variable's histogram is more symmetric and resembles a normal distribution.

Boxplots reveal data spread, outliers, and central tendency. The original "Cadmium" variable's boxplot indicates potential outliers above the upper whisker. The log-transformed variable's boxplot has a smaller interquartile range and fewer outliers, suggesting the transformation reduced the impact of extreme values.

Q-Q plots assess data normality. The Q-Q plot for the original "Cadmium" variable shows non-normality indicated by deviations from the diagonal line at both tails. The Q-Q plot of the log-transformed variable aligns more closely with the diagonal line, suggesting improved normality.

Comment

The original Cadmium data has a skewed distribution with some extreme values, as evident from the summary statistics and plots. The log transformation helps mitigate the effect of extreme values and brings the distribution closer to a normal distribution. This is reflected in the summary statistics and the shape of the histograms and Q-Q plots. The log transformation appears to be appropriate in this case as it reduces the skewness, improves normality, and reduces the influence of outliers on the distribution.

Question 2

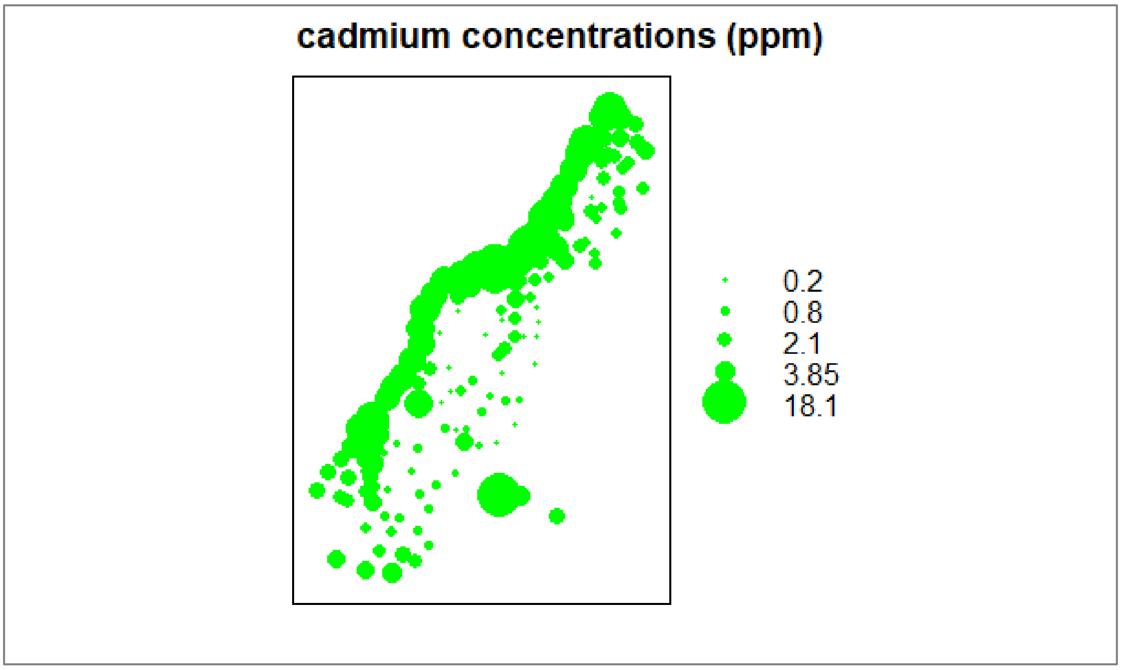
Comment

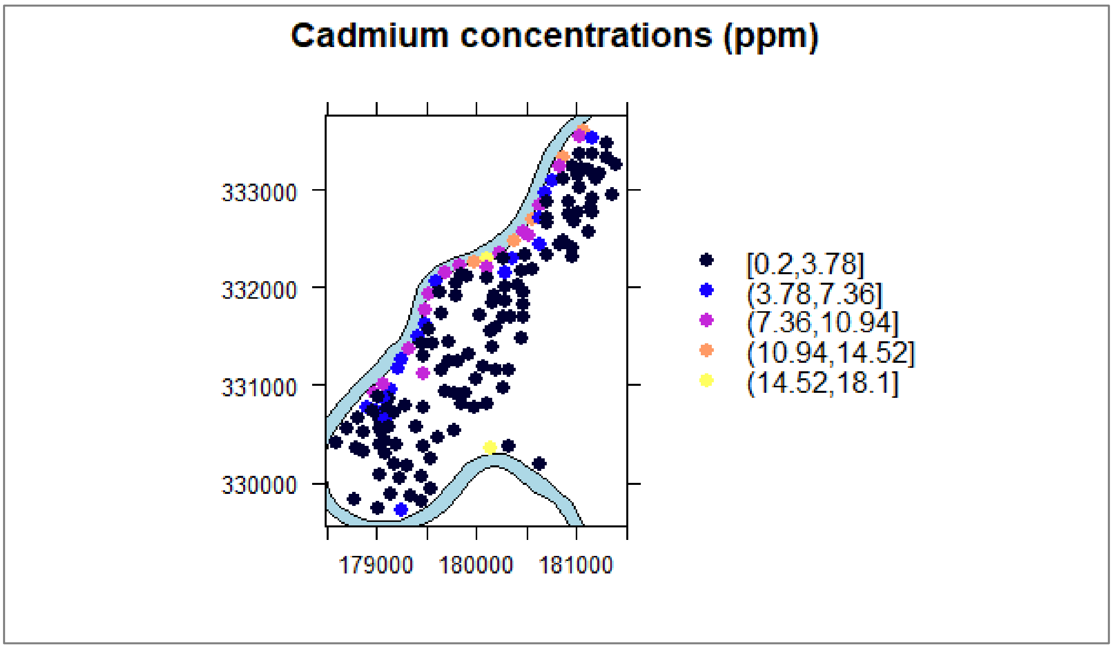
The bubble plot illustrates the spatial distribution of cadmium concentrations using scaled dots coloured in green. The plot's title is Cadmium concentrations (ppm). The dot plot also displays cadmium concentration data as dots. It incorporates geographical context by adding light blue polygons depicting the Meuse river boundaries in the background. A legend is included on the right side.

Interpretation of Spatial Distribution:

The bubble plot uses scaled green dots to represent concentrations, with larger dots indicating higher levels. In contrast, the dot plot incorporates geographical context by overlaying light blue polygons that outline the Meuse river boundaries. This plot also depicts cadmium concentrations using dots and includes a legend. In both visualizations, regions with elevated cadmium concentrations are evident, especially around the Meuse river. The distribution appears somewhat scattered, with certain areas exhibiting higher concentrations while others display lower values. Possible outliers are indicated by larger dots in the bubble plot. The presence of higher concentrations near the river suggests a potential link between cadmium levels and proximity to the water body, influenced by factors like industrial activities, land use, and natural processes.

Plots



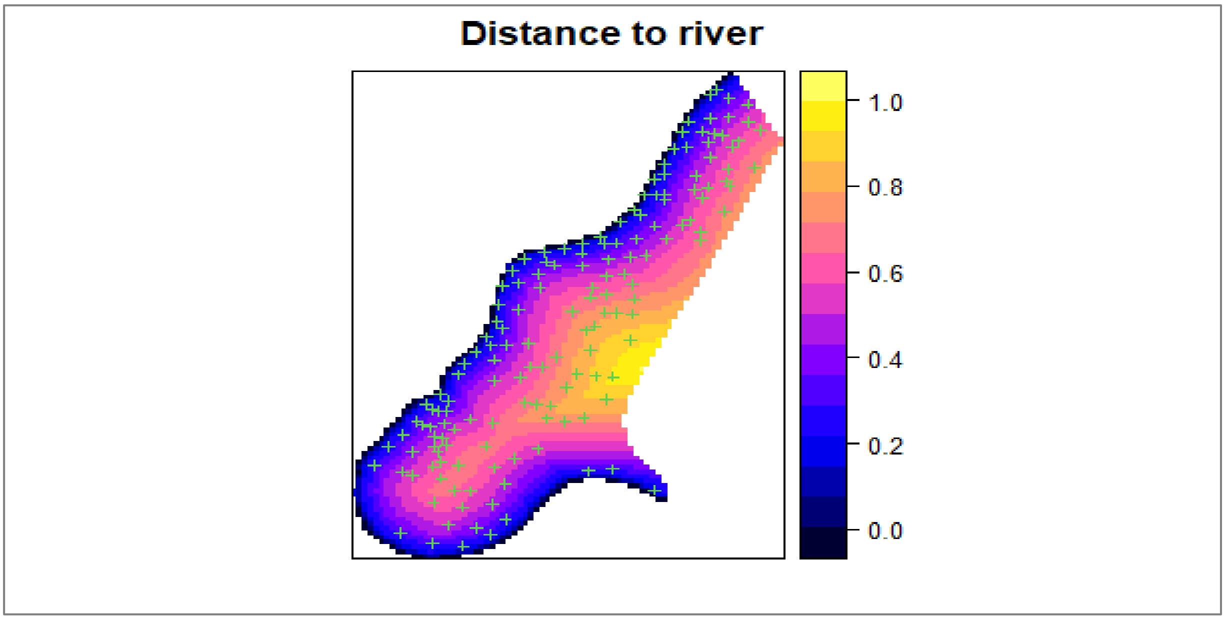


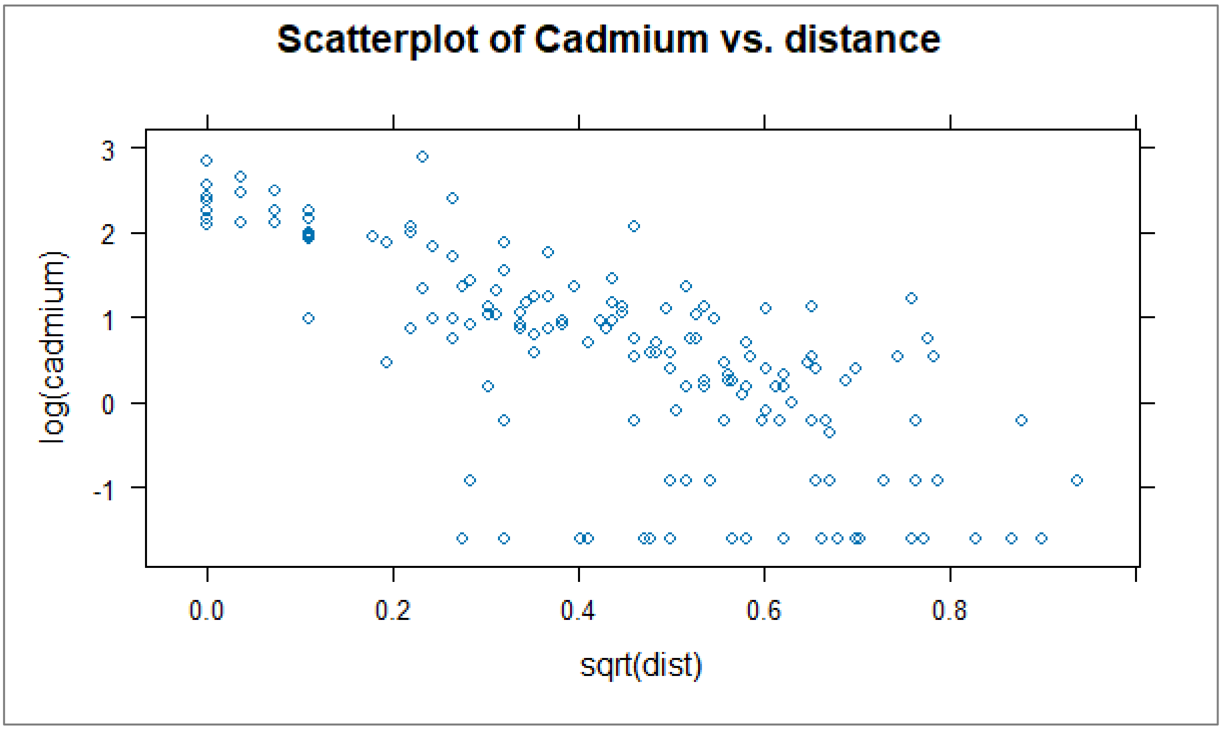
Question 3

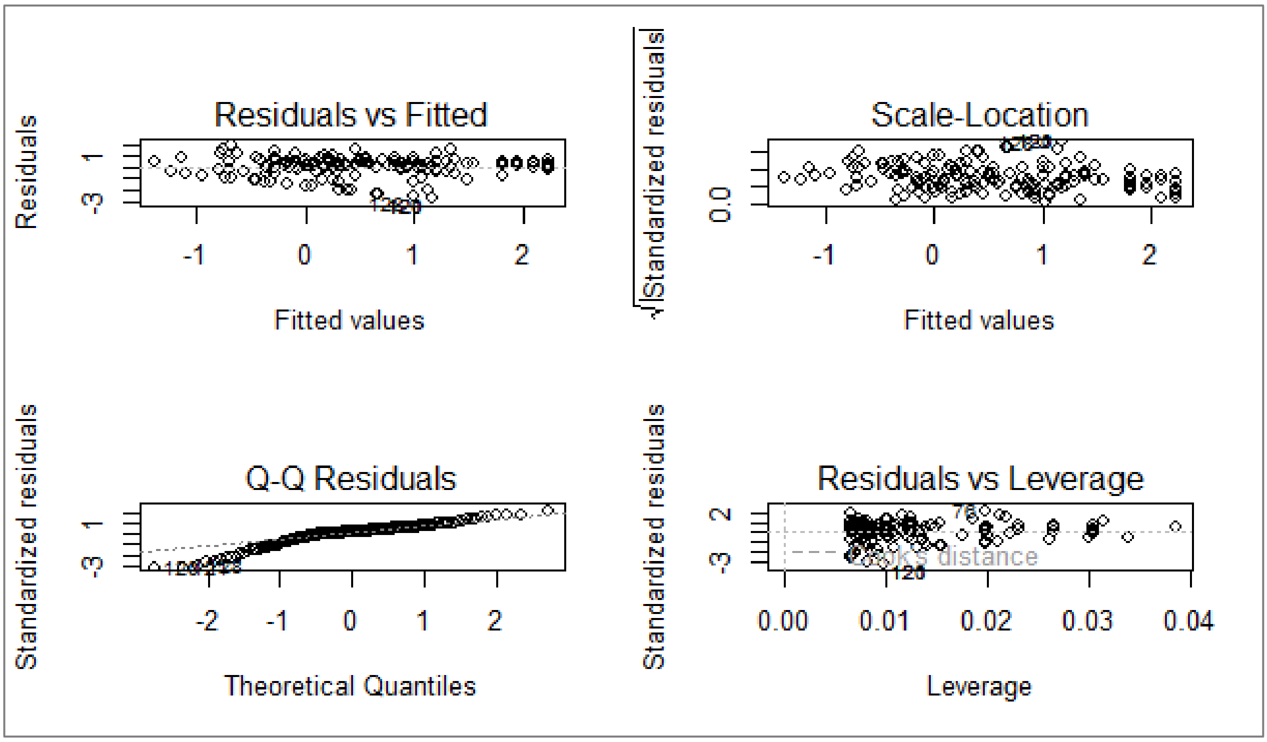
Comment

The scatterplot indicates a potential negative linear relationship between log-transformed cadmium concentrations and the square root of the distance to the river, suggesting decreasing concentrations as distance from the river increases. To confirm this relationship, a linear regression model is used. The summary of the regression provides insights into significant coefficients, particularly the p-value linked to the square root of distance, which explains variations in log-transformed cadmium concentrations. Diagnostic plots help assess assumptions like non-constant variance or non-linearity that might impact the model's reliability. The quality of interpolation depends on the model's fit, as indicated by the R-squared value in the summary. Predicted value and standard error spatial plots visualize the model's interpolation of cadmium concentrations across the study area, with lower standard errors reflecting higher prediction confidence and better alignment with real-world patterns.

Plots





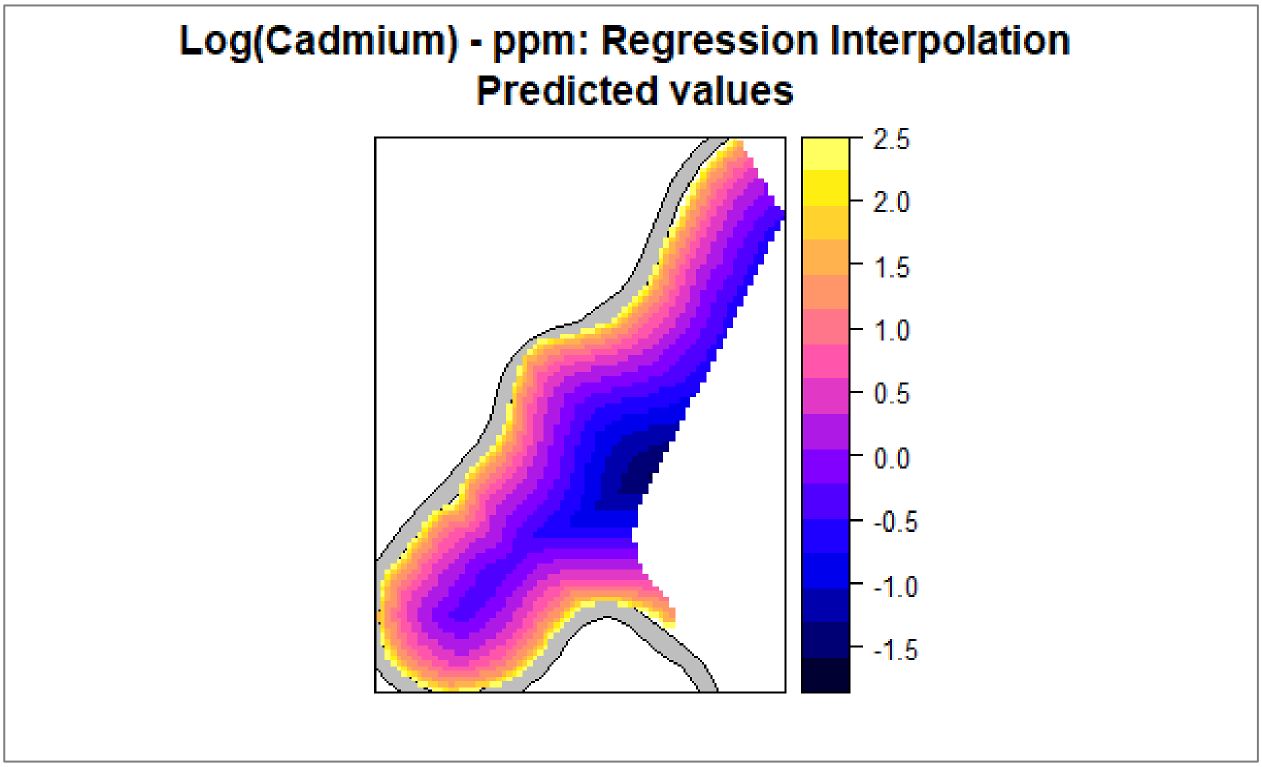


Residuals vs Fitted

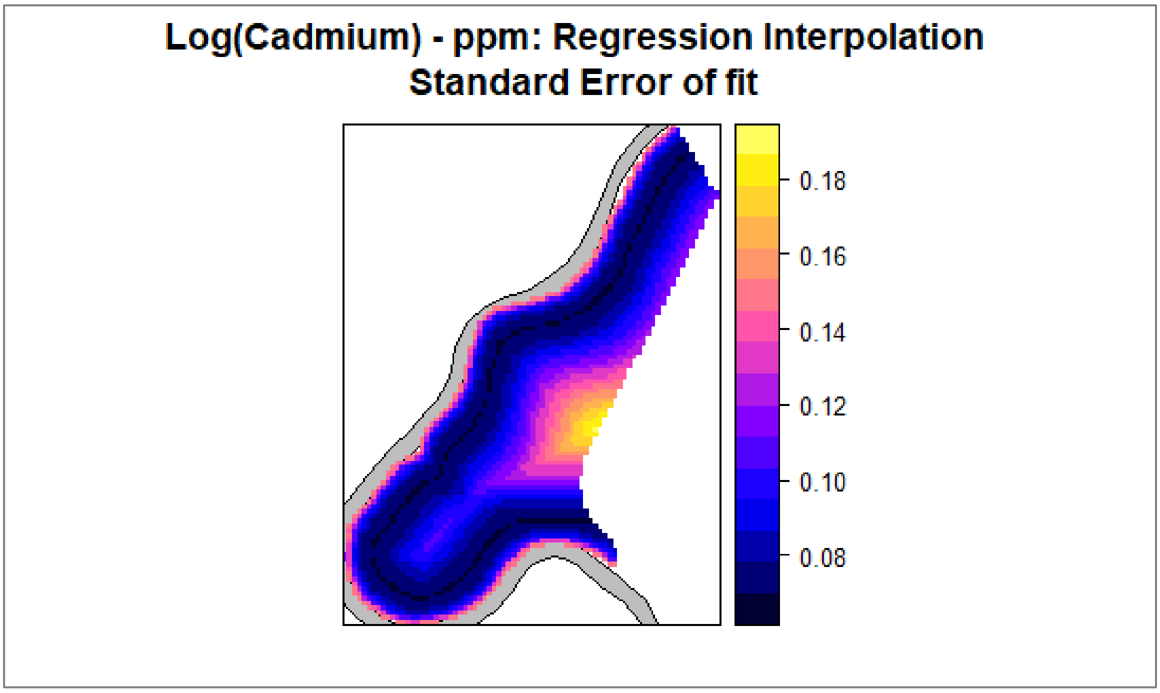
Scale-Location

Q-Q Residuals

Residuals vs Leverage



Log(Cadmium) - ppm: Regression Interpolation Predicted values



Log(Cadmium) - ppm: Regression Interpolation
Standard Error of fit

Question 4

Comment

In comparing and interpreting different trend surfaces for data distribution, the choice of trend order plays a crucial role. A linear trend surface with a degree of 1 assumes a simple linear pattern in the data, while quadratic and cubic trends (degrees 2 and 3) offer greater flexibility to capture curved patterns and fluctuations. The appropriateness of the trend order hinges on the actual behaviour of the data. When the data demonstrate a linear relationship, a degree-1 trend might suffice, while nonlinearities or intricate patterns might necessitate higher-degree trends. The quality of interpolation depends on control point distribution and density; having more control points can enhance interpolation accuracy. However, caution must be exercised with higher-degree trends (2 and 3) to prevent overfitting. If the underlying relationship doesn't warrant such complexity, overfitting can introduce noise into the interpolated surface, potentially diminishing the reliability of the results.

When contrasting the outcomes of the three maps, it's evident that the degree 1 interpolation is likely to yield a smoother surface with reduced fluctuations. On the other hand, the interpolations using degrees 2 and 3 capture finer local variations, but there's a risk of accentuating noise, particularly when the underlying data isn't inherently intricate. Opting for the degree 1 interpolation could offer a more favourable fit when the spatial variation of the data adheres to a simpler linear pattern.

Code

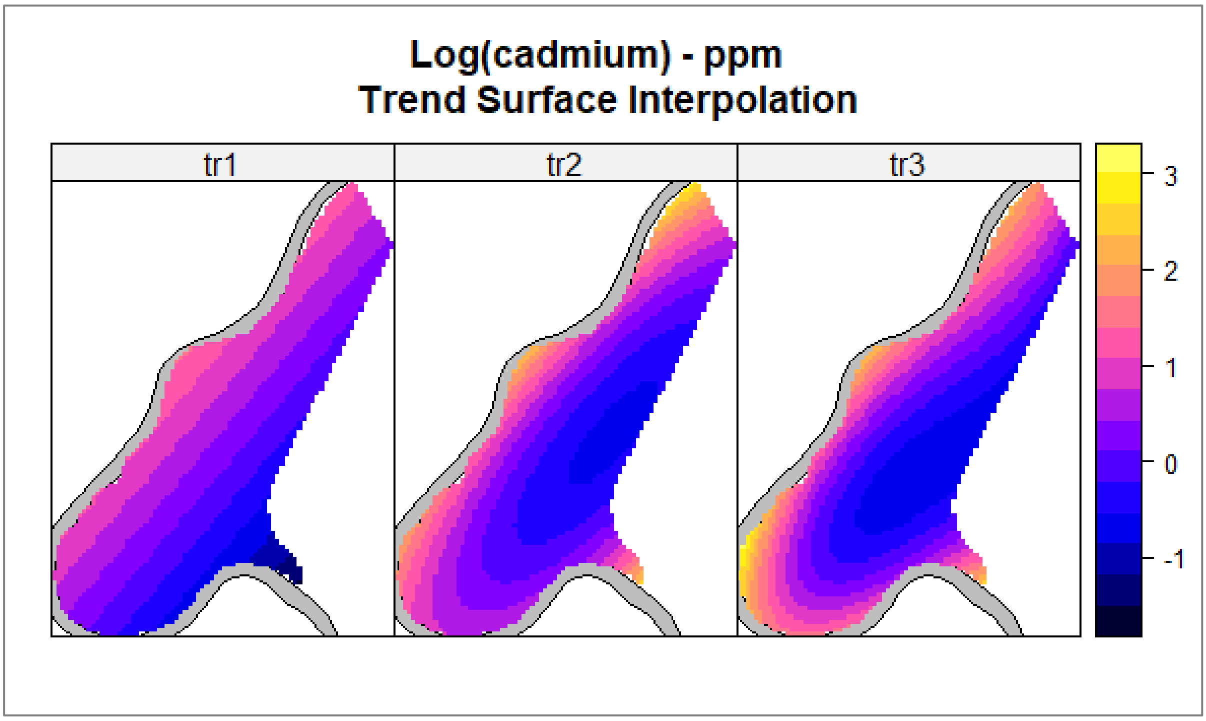
```
# Performing kriging and creating predictions for degree 2 and 3

predictions_deg2 <- krige(log(cadmium) ~ 1, meuse, meuse.grid, degree = 2)

predictions_deg3 <- krige(log(cadmium) ~ 1, meuse, meuse.grid, degree = 3)
```

```
# Storing the predictions in the meuse.grid data frame
meuse.grid$tr2 <- predictions_deg2$var1.pred
meuse.grid$tr3 <- predictions_deg3$var1.pred
spplot(meuse.grid, c("tr1", "tr2", "tr3"), sp.layout = meuse.lt,
      main = "Log(cadmium) - ppm \n Trend Surface Interpolation")
```

Plot



Question 5

Comment

The IDW interpolation technique is utilized with varying power values (1, 2.5, 5, 10) for log-transformed cadmium concentrations. The resulting predicted values are stored in distinct columns (idwp1, idwp2.5, idwp5, idwp10) within the meuse.grid data frame. To visualize and contrast the interpolated surfaces based on different power values, a spatial plot (spplot) is generated. This plot overlays control points from the meuse data frame, providing a reference for comparison.

1. **The Role of the Power Function:**

   - The power value (p) determines the weight assigned to each control point based on its distance. Higher p values decrease the impact of distant points, resulting in localized interpolation.

   - Lower power values (near 0) create uniform weighting and smoother interpolations, while higher values (far from 0) accentuate nearby points, leading to abrupt transitions.

2. **Quality and Prediction Behaviour:**
   - Lower power values (e.g., p = 1) prioritize nearby points, yielding smoother surfaces but potentially losing local detail due to over smoothing.
   - Higher power values (e.g., p = 5, p = 10) emphasize nearby points, producing sharper, localized changes that offer greater detail but might introduce noise.
   - Selecting the power value depends on data distribution, control point density, and desired interpolation characteristics. An evenly distributed control point setup might benefit from a power value near 1, while sparser control points might necessitate higher values to capture local nuances effectively.

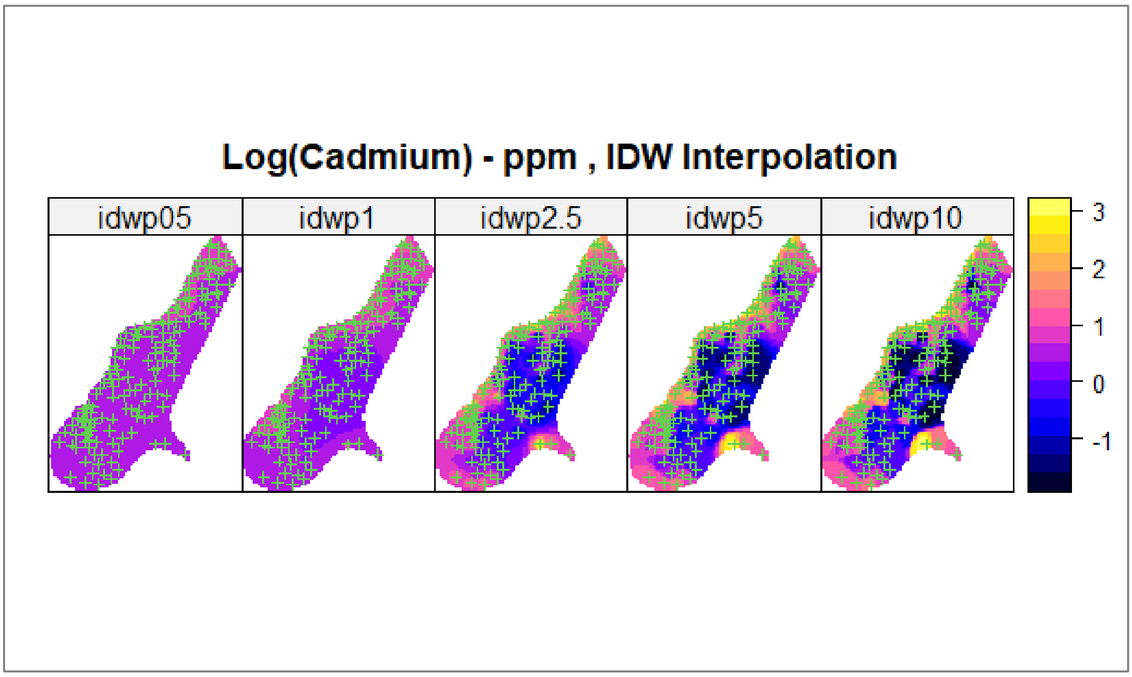
In conclusion, the choice of power value in the IDW interpolation significantly impacts the results:

A lower power value (p = 1) yields smoother transitions between control points, although it might miss finer local variations.

Higher power values (p = 5, p = 10) provide more localized predictions, capturing local features but potentially being influenced by noise.

Selecting the appropriate power value should be influenced by the data's characteristics and the intended utilization of the interpolation outcomes.

<u>Plot</u>



Question 6

Comment

h-scatterplots show how semi variance changes with distance, aiding in assessing spatial correlation across different lags.

Variogram Cloud displays pairs of distances and semi variance values, offering insights into variogram structure.

Sample Variogram Plot demonstrates semi variance behaviour with lag distance, helping understand spatial correlation.

Variograms in Different Angles assess variograms for various directions, aiding in detecting anisotropy.

Override Default Cutoff and Interval Width customizes cutoff and interval width for finer or coarser variogram views.

Specifying Interval for Distance Vector focuses on semi variogram behaviour within specific distance ranges.

Variogram Plot visually represents the calculated variogram with sample points and a fitted model.

Initial Variogram Fitting fits a model to the sample variogram using initial parameter guesses.

Partial Fitting of Variogram Coefficients selectively fits variogram coefficients, useful with prior parameter knowledge.

REML Fitting uses restricted maximum likelihood for accurate parameter estimation and robustness.

Anisotropy Variogram models variograms considering direction and anisotropy parameters.

Anisotropy Plot displays anisotropy variograms and fitted models, visualizing spatial correlation variation.

Variogram Map generates a map showing the spatial distribution of semi variance for identifying correlation patterns.

Simple Kriging and Ordinary Kriging Interpolations perform predictions at unobserved locations based on spatial correlation using different kriging methods.

Assumptions and Implications: Variogram analysis helps us understand the spatial correlation structure of the data and identify an appropriate variogram model.

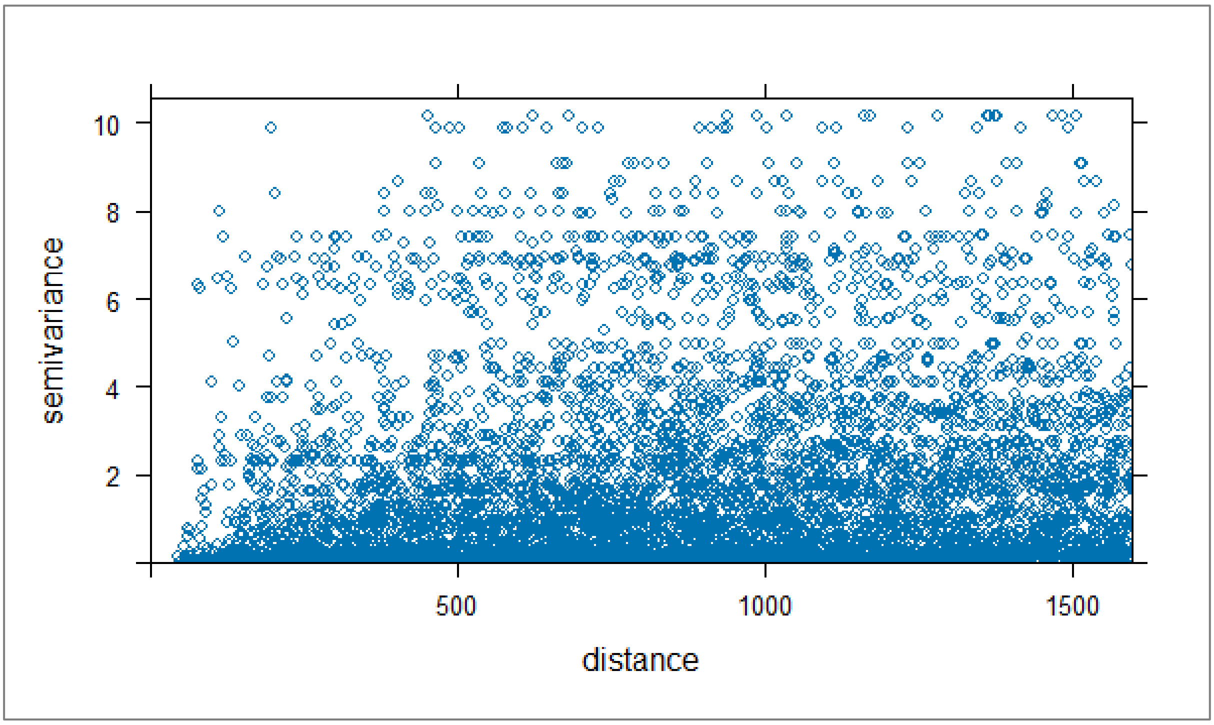
Fitting variogram models involves making assumptions about the nugget, sill, and range parameters. These assumptions affect the quality of predictions.

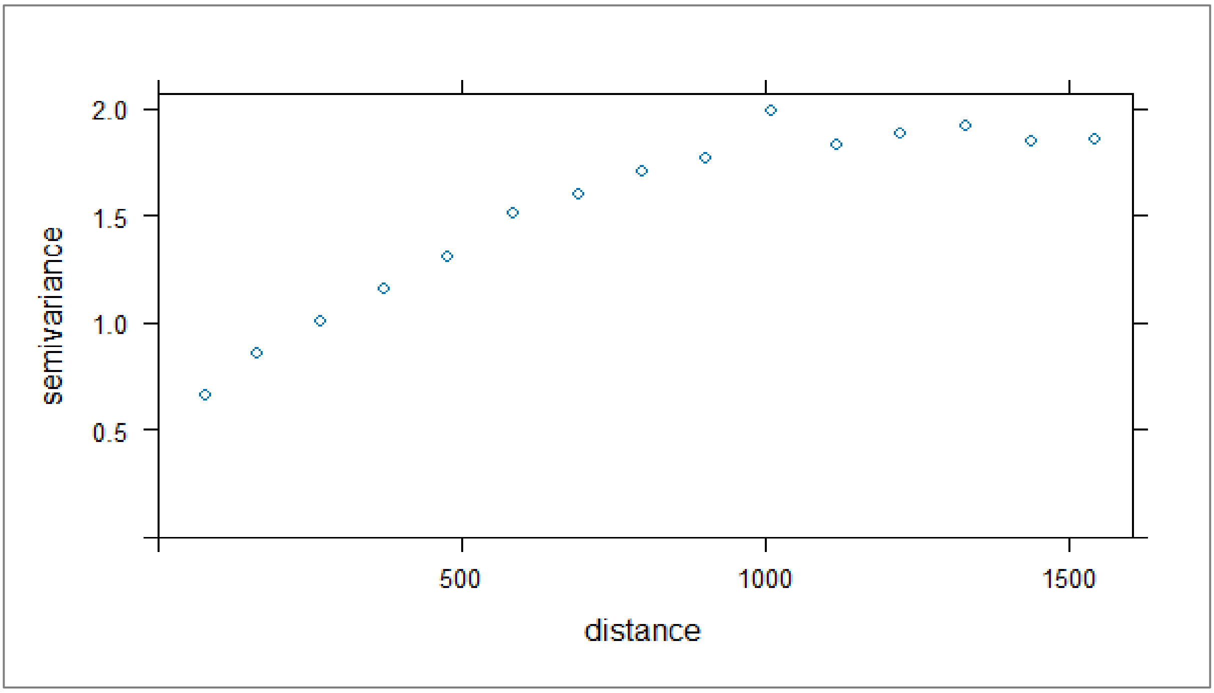
Anisotropy modeling accounts for directional variability in spatial correlation.

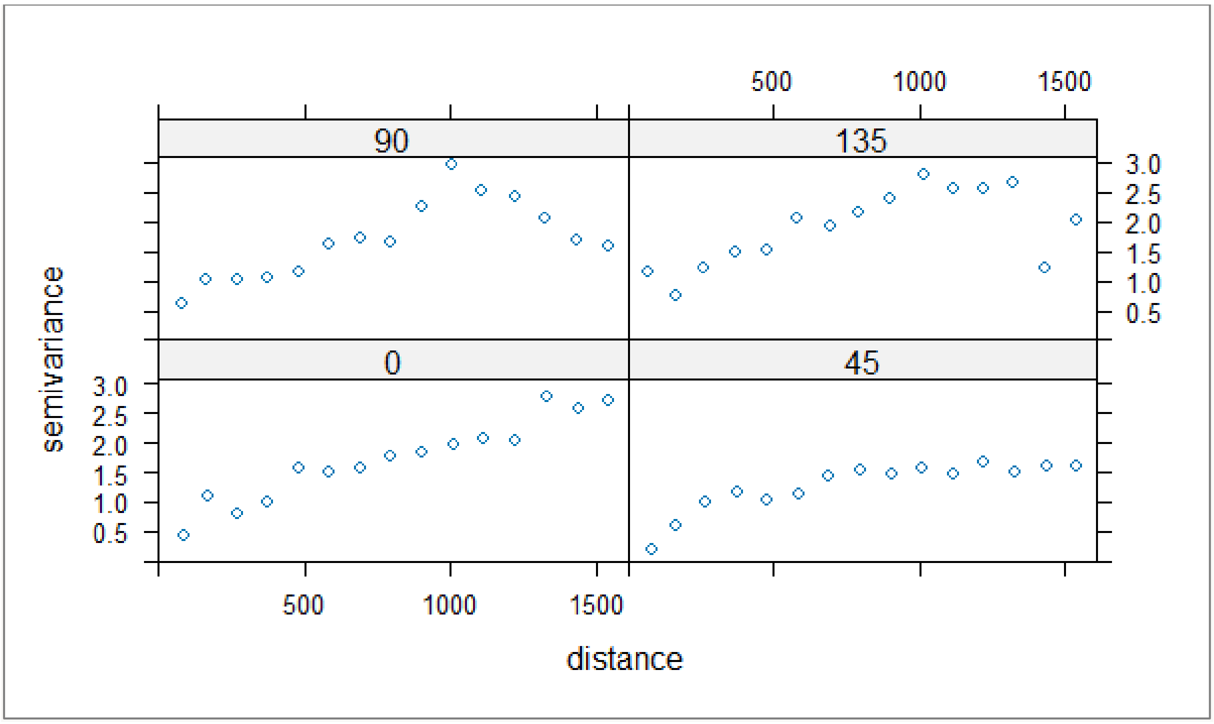
Kriging interpolations are based on the variogram model and provide predictions for unobserved locations.

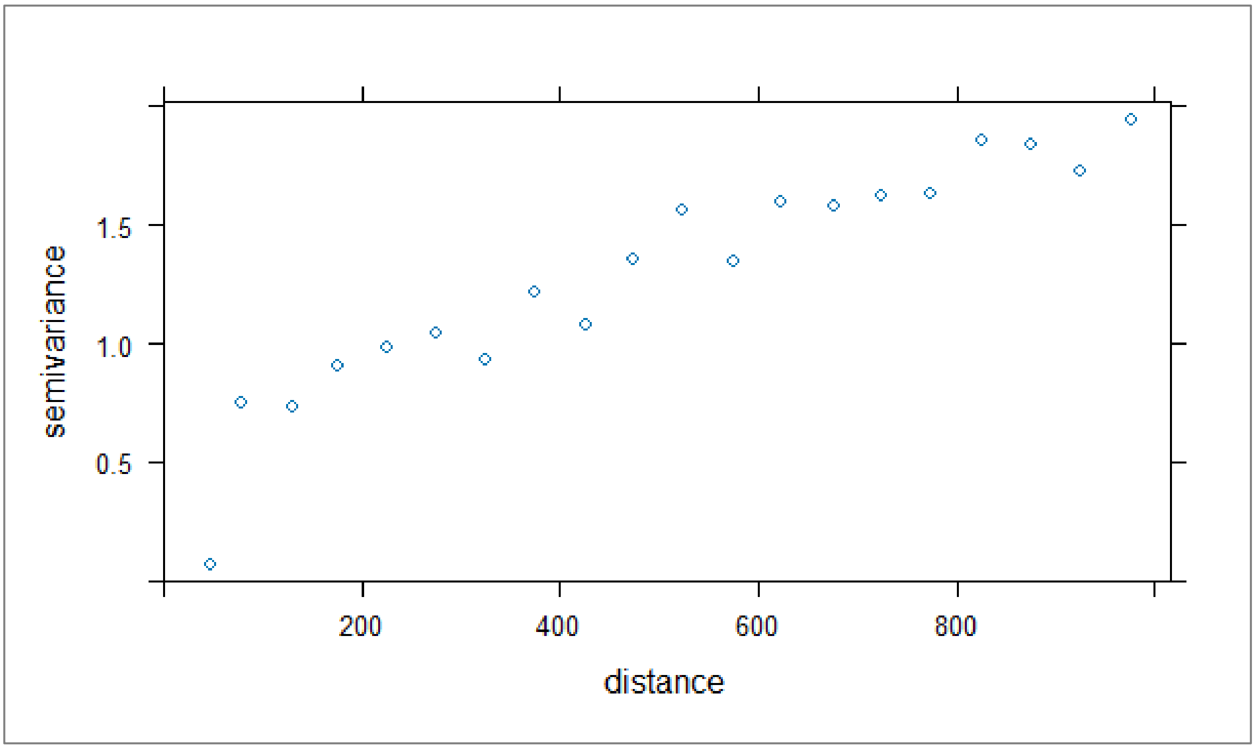
The choice of interpolation method (simple kriging vs. ordinary kriging) depends on whether you have information about the mean of the variable at unsampled locations.

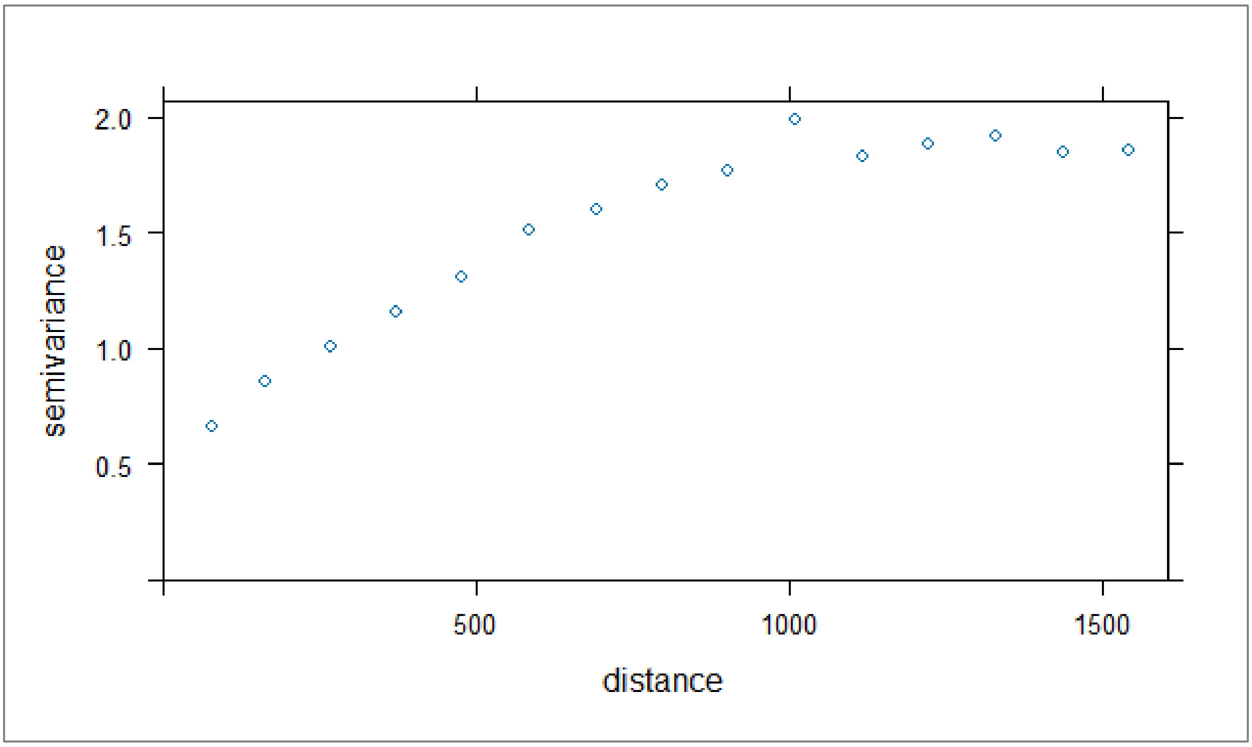
Plots

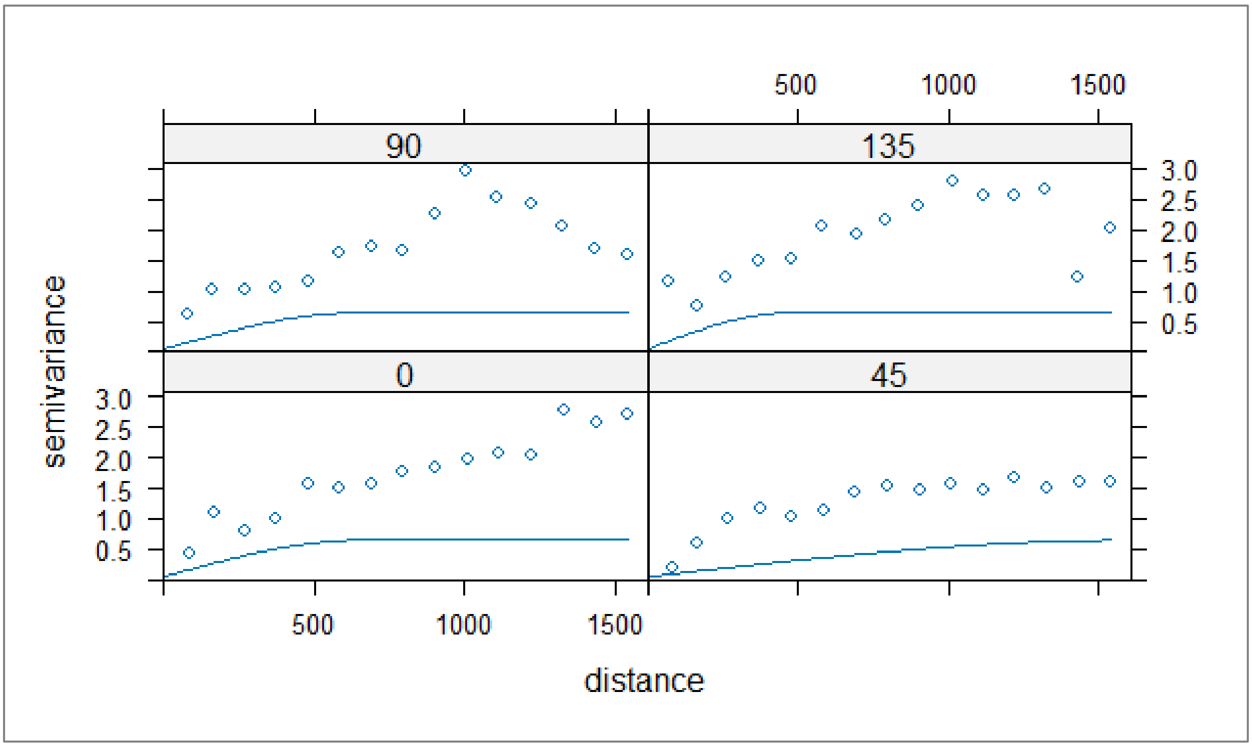


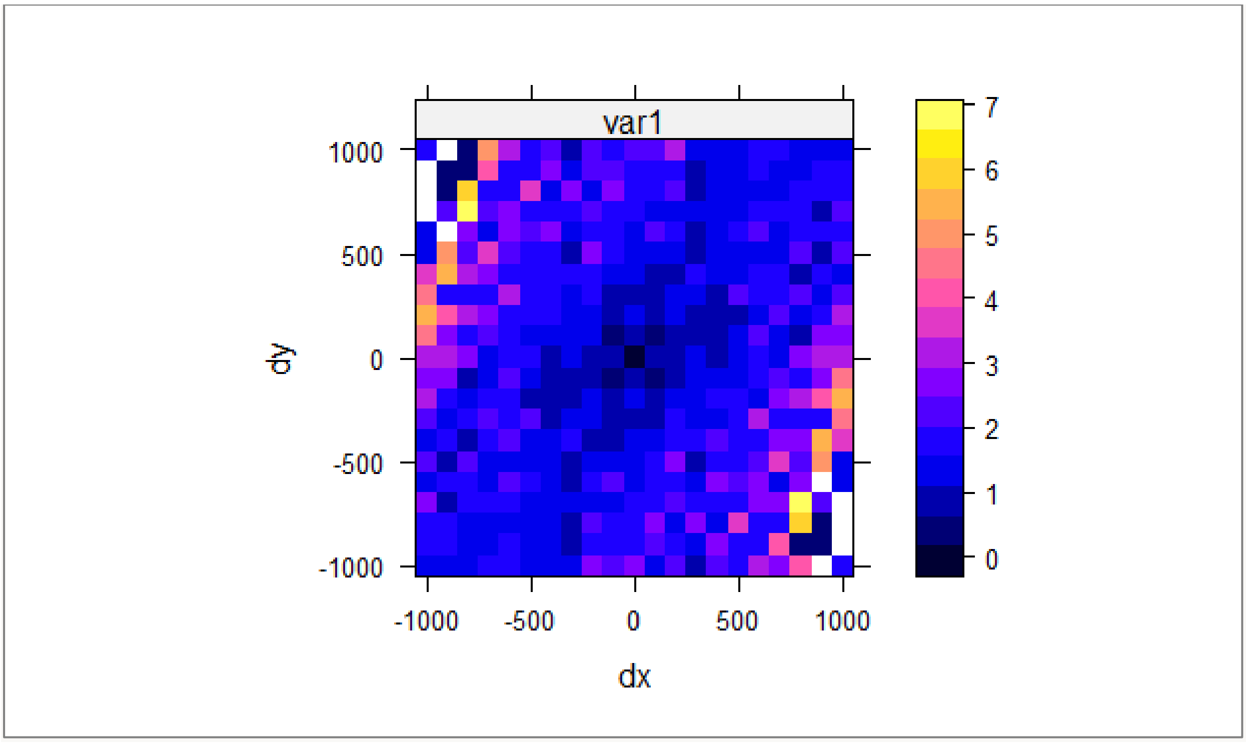



90
135
0
45
semivariance
distance
500
1000
1500
0.5
1.0
1.5
2.0
2.5
3.0


semivariance
distance
0.5
1.0
1.5
200
400
600
800


semivariance
2.0
1.5
1.0
0.5
500
1000
1500
distance


500
1000
1500
90
135
3.0
2.5
2.0
1.5
1.0
0.5
semivariance
0
45
3.0
2.5
2.0
1.5
1.0
0.5
500
1000
1500
distance



Question 7

Comment

Prediction Variance: Kriging not only offers predictions but also provides estimates of prediction variance. Higher prediction variance signifies increased uncertainty in the prediction. This aspect aids in pinpointing areas where interpolation might be less dependable due to sparse or unevenly distributed control points.

Residual Analysis: Residuals are discrepancies between observed and predicted values. Examining the distribution of residuals assists in assessing whether the interpolation captures the underlying data variability. If residuals display a random pattern without systematic trends, it suggests the interpolation successfully captures the spatial data patterns.

Variogram Fitting Quality: The quality of the fitted variogram model influences kriging prediction quality. A visual comparison between the fitted and experimental variograms helps determine how well the model captures spatial correlation structure.

Local Measures: Instead of solely focusing on global measures, local prediction accuracy metrics can be computed and displayed on a map. Local measures draw attention to regions with heightened prediction uncertainty or exceptional interpolation performance.

Outlier Analysis: Identifying outlier values within residuals aids in pinpointing problematic areas or control points. This step contributes to the accuracy assessment of the interpolation.

By examining these output measures and diagnostic tools, we can make a more comprehensive assessment of the quality of Simple and Ordinary Kriging interpolations. This assessment guides confidence in the interpolated results and helps identify potential issues that need further attention or exploration.

Question 8

Comment

Visual Comparison of Interpolation Results

Smoothness and Local Variations: In both Simple Kriging (SK) and Ordinary Kriging (OK) maps, the interpolated surfaces exhibit a relative smoothness, effectively capturing the underlying spatial trends of the variable. Nonetheless, within areas of denser control point distribution, the OK map showcases slightly more detailed local variations compared to the SK map.

Control Point Influence: Within regions featuring higher control point densities, interpolated values closely align with observed data points in both SK and OK maps. OK, relying on a weighted average of nearby control points, better adapts to local changes in control point density and distribution.

Edge Effects: Along the study area's edges, where control points are sparse, both SK and OK methods rely more on global trends captured by the variogram model. The SK method, incorporating a local mean estimate, potentially presents a smoother transition from observed to unobserved areas along these edges.

Nugget Effect: The nugget effect, responsible for small-scale variability, can lead to localized fluctuations in interpolated values. The SK method might exhibit fewer nugget effects due to its integration of local mean estimates.

Prediction Uncertainty: Within regions with limited control point influence, both SK and OK maps demonstrate higher uncertainty, evident by wider prediction intervals. Notably, the OK map might show slightly narrower prediction intervals than the SK map in areas with sparse control points, reflecting OK's adjustment of weights based on neighboring points.

The comparative analysis yields valuable insights into the behaviour of predictions concerning control point distribution and density: Both Simple Kriging (SK) and Ordinary Kriging (OK) adeptly capture the overarching spatial trends of the variable, resulting in relatively smooth interpolated surfaces. OK displays superior adaptability to local variations and changes in control point density due to its weighted averaging approach. On the other hand, SK, incorporating local mean estimates, offers smoother transitions from observed to unobserved areas along the study area's edges. Notably, regions with sparse control points exhibit heightened prediction uncertainty, emphasizing the necessity for cautious result interpretation in such contexts.

Question 9

Comment

Cross-Validation: Is executed for each interpolation method, involving the exclusion of a subset of data points to compare predicted values against observations. The outcomes of cross-validation are analysed to gauge accuracy and method robustness.

Distribution of Residuals: (disparities between predicted and observed values) for each method is examined. A normal distribution of residuals centred around zero signifies a well-fitted model.

Spatial Patterns of Residuals: Are assessed to identify consistent overestimation or underestimation regions for each method. This reveals method strengths and weaknesses across different study area segments.

Prediction Variance: Maps generated by each interpolation method allows for an evaluation of prediction uncertainty. Higher variance regions indicate areas of less reliable predictions.

Model Fit Quality: Is analysed for each method. This entails assessing how well each model captures the spatial correlation structure within the data.

Sensitivity Analysis: Involves varying crucial parameters like the power value in IDW or variogram model parameters in kriging. This examination observes how parameter changes influence prediction quality for each method.

Local Measures of Accuracy: Measures, such as local Root Mean Squared Error (RMSE) or Mean Absolute Error (MAE), are computed to appraise method performance across different study area regions.