Assignment IV - Kriging

By

Parsa Besharat

An assignment's handout submitted as part of the requirements

for the lecture, Geomodeling, of MSc Mathematics of Data and Resources Sciences

at the Technische Universität Bergakademie Freiberg

December, 2024

Supervisor: Prof. Jörg Benndorf

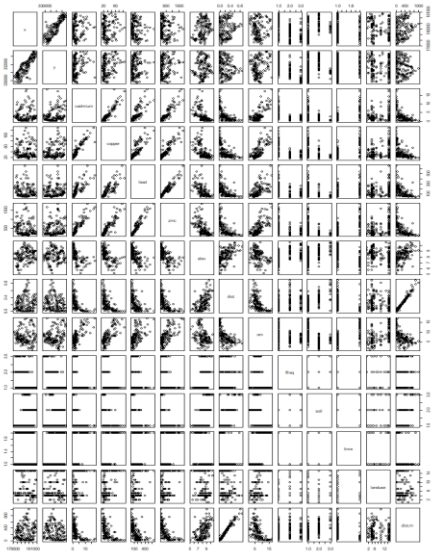
Contents

Page

Explanation of the MeUse.r code

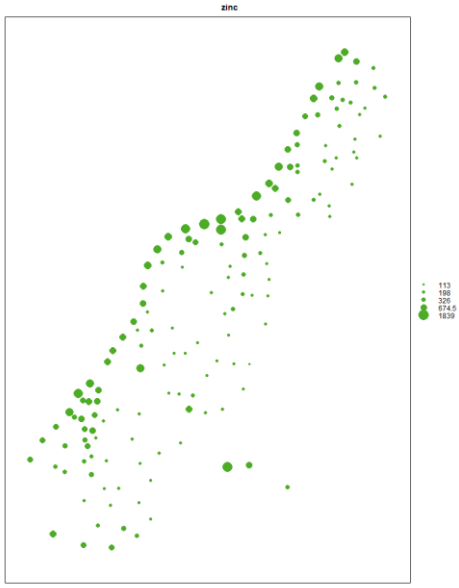
The R script performs a geostatistical analysis using Kriging to model the spatial distribution of zinc concentrations in the Meuse dataset. It begins by installing and loading necessary libraries, including ***sp, gstat, MASS, rgdal, and sf,*** which provide functions for handling spatial data, performing geostatistical modeling, and working with coordinate reference systems. The dataset meuse is then examined to check its coordinate reference system (CRS) using the ***st_crs()*** and ***proj4string()*** functions, ensuring spatial consistency between datasets. The script assigns the EPSG:28992 projection (Amersfoort / RD New) to both the meuse dataset and ***meuse.grid***, which is a necessary step before performing spatial interpolation. If there is a mismatch in projections, the script applies a transformation using ***spTransform()*** to align the datasets. Once the spatial data is properly set up, exploratory data analysis (EDA) is conducted by visualizing the distribution of zinc concentrations with bubble plots and histograms. The script also assesses the relationship between zinc levels and the distance to the river Meuse by plotting log-transformed zinc concentrations against the square root of distance. To better understand the spatial variation in zinc concentrations, the script calculates summary statistics such as the mean, variance, and standard deviation. Outliers are identified using histograms and boxplots, and the log transformation of zinc values helps normalize the distribution.

Plot Analysis

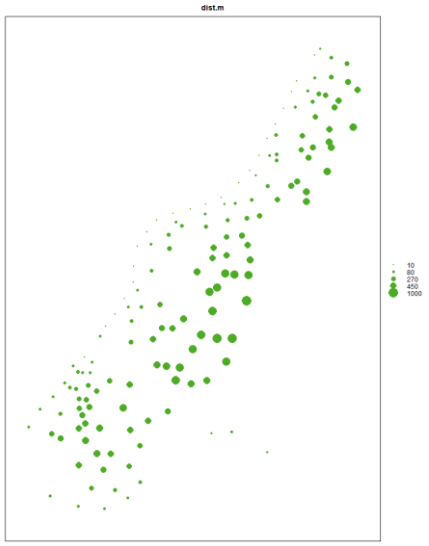


This image is a pairwise scatter plot matrix, also known as a correlation plot, displaying relationships between multiple variables in the dataset. Each cell in the matrix represents a scatter plot between two different variables, allowing for visual inspection of potential correlations. For example, variables like zinc, cadmium, copper, and lead may show positive correlations if their scatter plots form an upward trend. The diagonal typically contains variable names or histograms of each variable's distribution. This plot is useful for identifying trends, dependencies, and potential multicollinearity between different

environmental factors, such as heavy metal concentrations and geographic features. Let me know if you need a more detailed interpretation of any specific relationships in this plot!

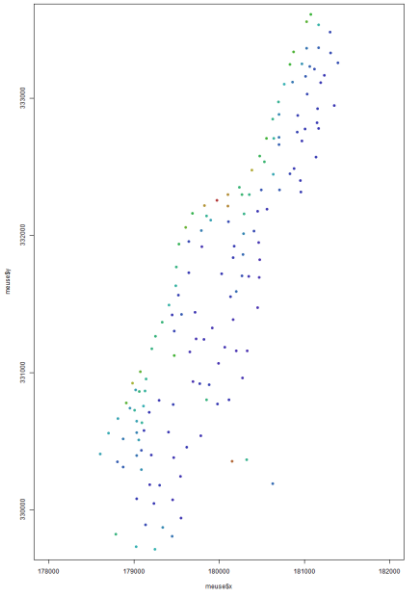


This is a bubble plot representing the spatial distribution of zinc concentrations across the study area. Each point corresponds to a sample location, and the size of the bubbles is proportional to the zinc concentration at that point. Larger bubbles indicate higher zinc levels, while smaller ones represent lower concentrations. The legend on the right provides reference values for interpretation. This visualization is useful for identifying hotspots of high zinc contamination, which could be linked to environmental factors such as proximity to industrial sites, water sources, or geological formations. If needed, additional layers like river locations or land use types could enhance the interpretation.

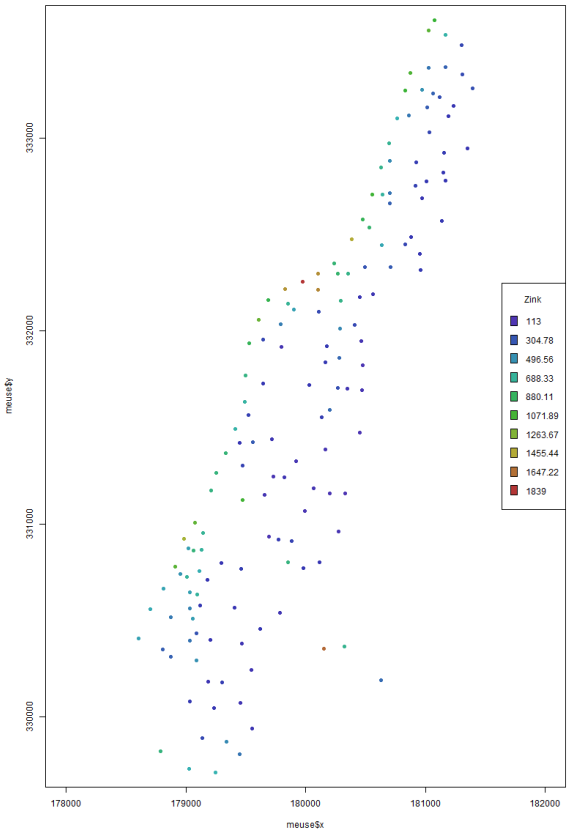


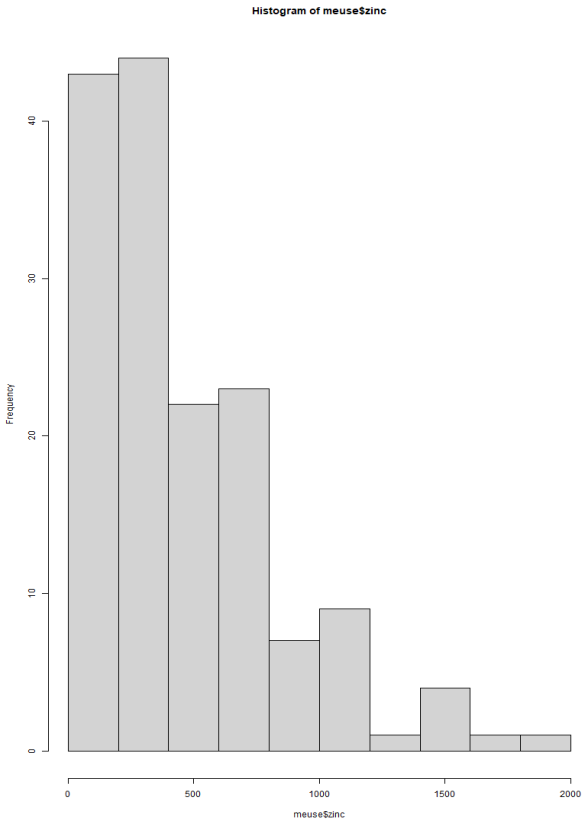
This is a bubble plot visualizing the distance to the river (dist.m) at various sampling locations. Each point represents a measurement site, and the size of the bubbles corresponds to the distance from the river, with larger bubbles indicating greater distances. This plot is crucial in understanding how environmental variables, such as zinc concentration, may vary with proximity to the river. If zinc pollution is linked to water sources, we would expect higher concentrations closer to the river and lower concentrations further away. Comparing this plot with the zinc concentration bubble plot can help assess spatial dependencies and contamination sources.

This image represents a heatmap or raster visualization of a spatial variable, likely showing the distance to the river or another environmental factor across the study area. The color gradient from red (higher values) to yellow (lower values) suggests that areas closer to a key feature (such as the river) are depicted in yellow, while those farther away are shown in red. This type of visualization helps in spatial interpolation and is useful for identifying patterns such as pollution dispersion or the effect of proximity to natural features on environmental variables like zinc concentration. If this represents distance, we can compare it with the zinc distribution to evaluate how pollution levels vary with proximity to the river.



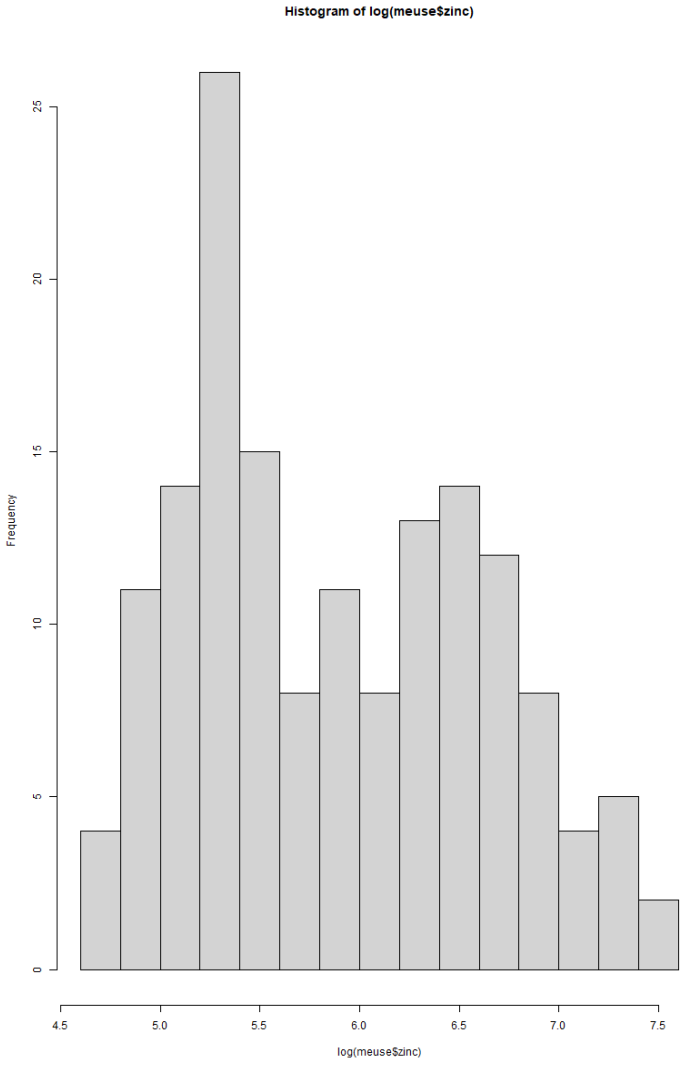
This is a spatial scatter plot that represents the distribution of zinc concentrations across different geographic coordinates (meuse$x and meuse$y). The points are color-coded based on zinc levels, with a gradient indicating increasing concentrations. Dark blue represents lower zinc values, while green, yellow, and red indicate higher concentrations. The spatial clustering of colors suggests that zinc contamination is not uniform across the area but rather concentrated in specific locations. This visualization helps in identifying contaminated hotspots, potential sources of pollution, and spatial trends in zinc dispersion, which can later be analyzed using variogram modeling and Kriging interpolation.

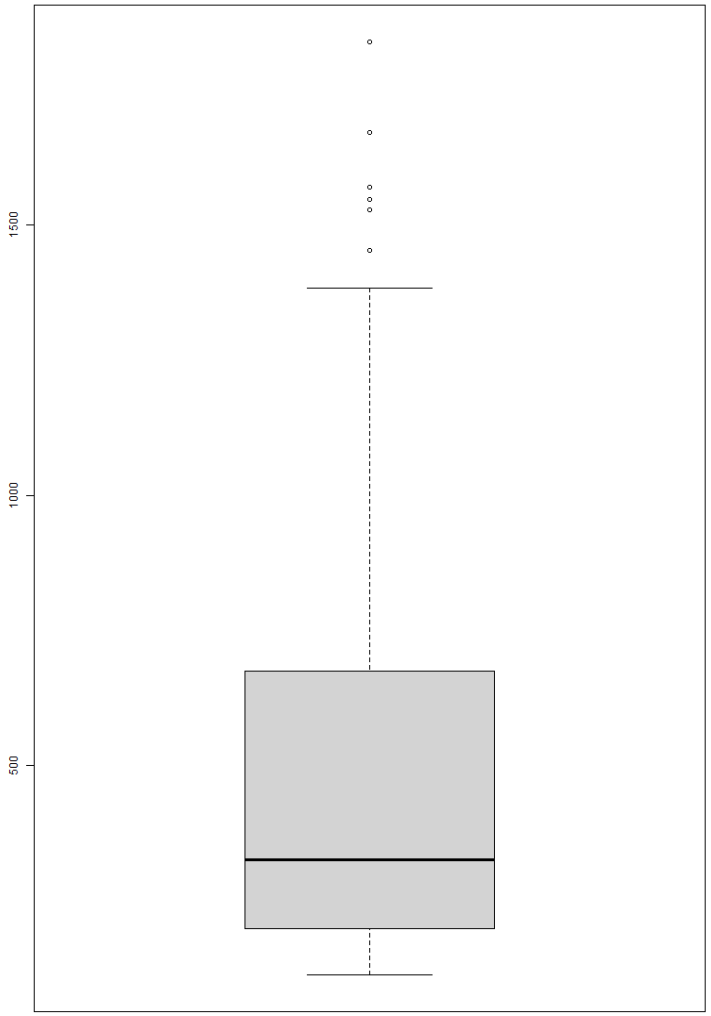




The Right figure, is a spatial scatter plot with a color gradient representing zinc concentrations in the study area. Each point represents a measurement location, with colors ranging from blue (low zinc levels) to red (high zinc levels). The accompanying legend provides numerical values for reference. The distribution suggests that zinc contamination is not uniform, with certain regions exhibiting higher concentrations, potentially due to local pollution sources or natural geological factors. This visualization is valuable for identifying hotspots of heavy metal accumulation, which can be further analyzed through variogram modeling and Kriging interpolation to estimate contamination in unmeasured locations**.**The histogram of zinc concentrations (meuse$zinc) displays the frequency distribution of zinc levels in the dataset. The distribution is right-skewed, indicating that most

locations have relatively low zinc concentrations, while a few sites exhibit very high zinc levels, possibly due to localized contamination or natural mineral deposits. The presence of extreme values suggests potential outliers, which can influence geostatistical modeling and require careful consideration during interpolation. Log-transforming the data might help normalize the distribution, making it more suitable for further analysis, such as Kriging.

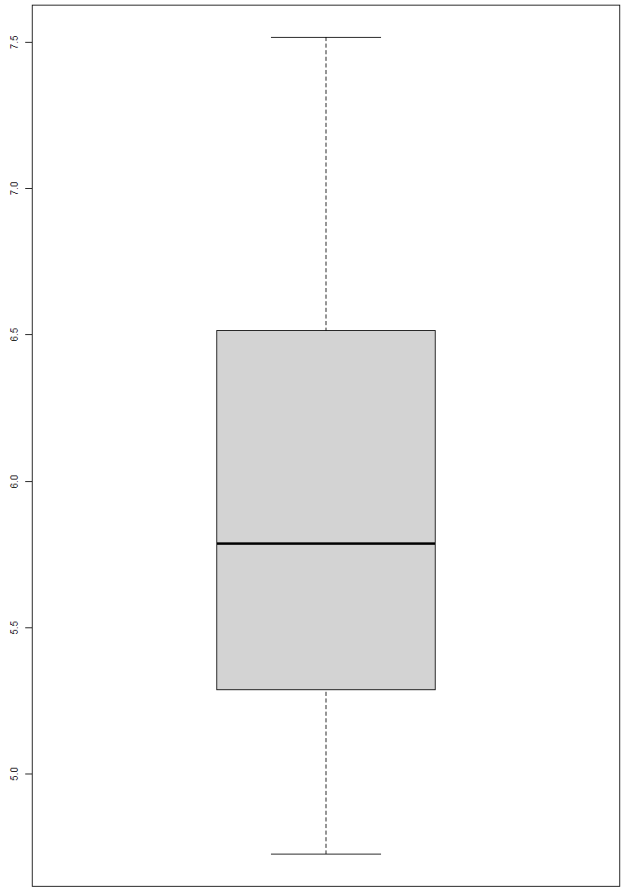


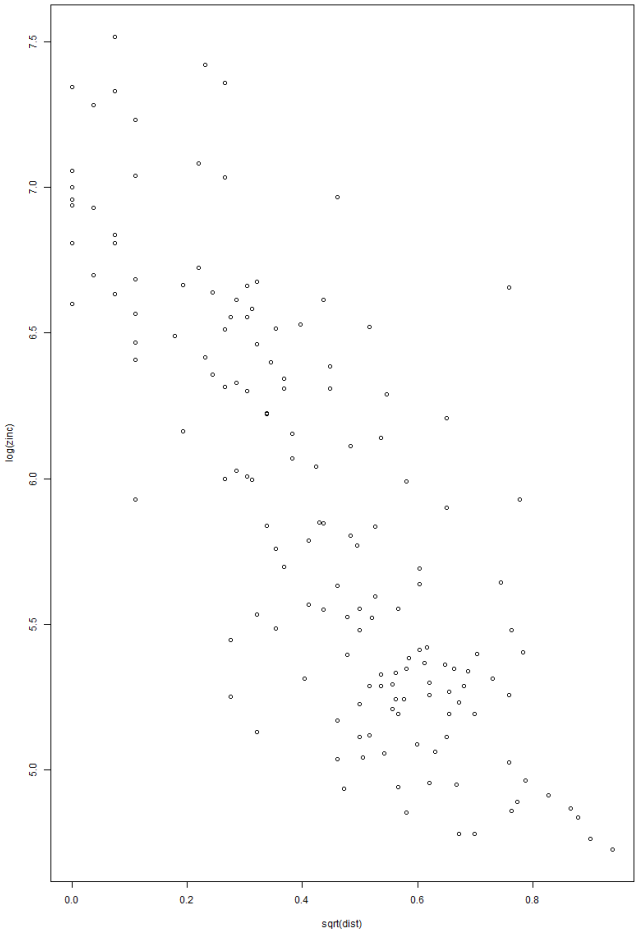


The histogram of log-transformed zinc concentrations (log(meuse$zinc)) shows a more symmetrical and approximately normal distribution compared to the raw zinc values. The log transformation helps to reduce the skewness observed in the original dataset (seen in the previous histogram) by compressing the higher values and

spreading out the lower ones. This is beneficial for geostatistical modeling, as many spatial interpolation techniques, such as Kriging, perform better with normally distributed data. The transformation ensures that extreme zinc values do not disproportionately influence the model while maintaining meaningful spatial variations.

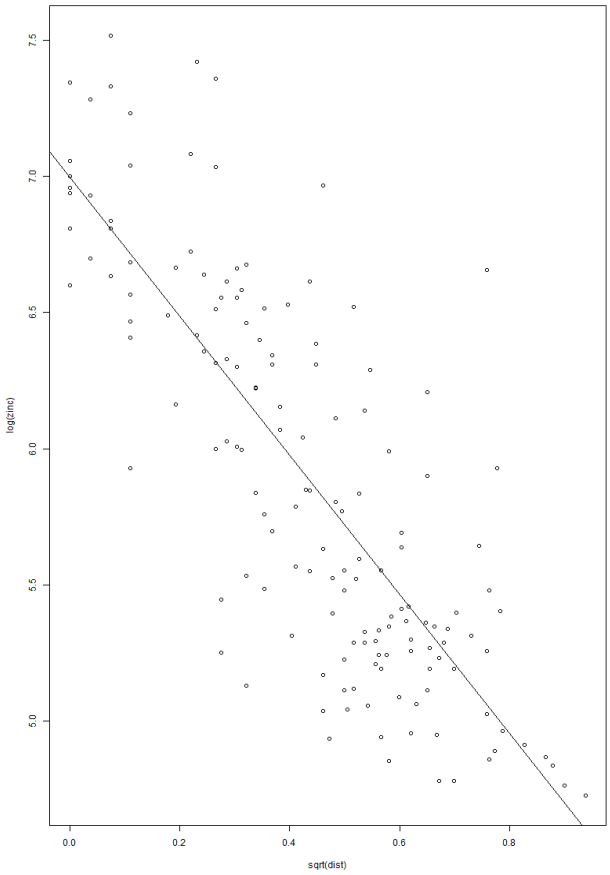
The boxplot of zinc concentrations (meuse$zinc) provides a clear visualization of the spread and distribution of zinc values, highlighting outliers in the dataset. The box represents the interquartile range (IQR), where the middle 50% of values are concentrated, while the horizontal line inside the box indicates the median zinc concentration. The whiskers extend to the minimum and maximum values within 1.5 times the IQR, beyond which outliers are plotted as separate points. The presence of multiple outliers above the upper whisker suggests significantly high zinc concentrations in certain locations, indicating localized contamination. These high values could disproportionately affect statistical models, so a log transformation (as seen in the previous histogram) may be necessary to normalize the data.

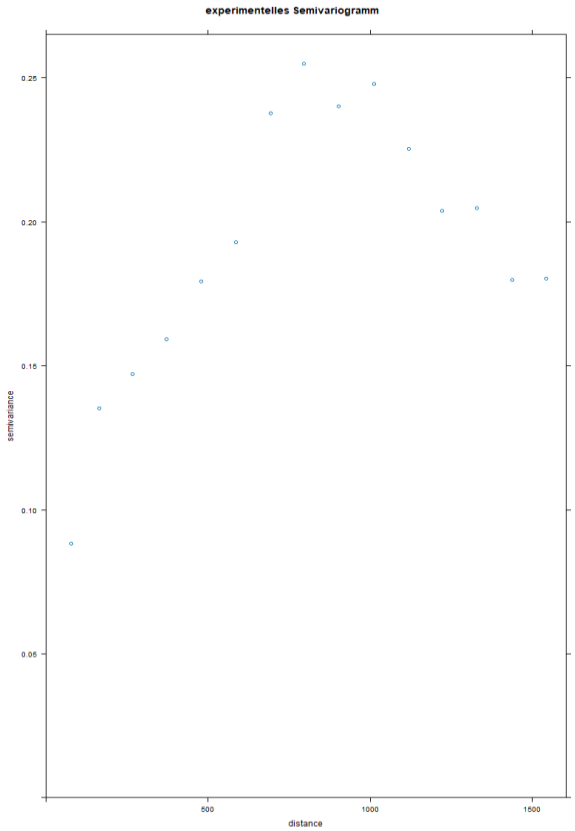




The boxplot of log-transformed zinc concentrations (log(meuse$zinc)) shows a much more balanced and symmetric distribution compared to the raw zinc data. The interquartile range (IQR) and whiskers indicate that the log transformation has effectively reduced the influence of extreme outliers, making the data more suitable for statistical modeling. Unlike the previous boxplot of raw zinc values, there are no extreme outliers present, confirming that the transformation successfully normalized the data. This is particularly useful for geostatistical techniques like Kriging, which assume a more normal distribution of the variable being interpolated.

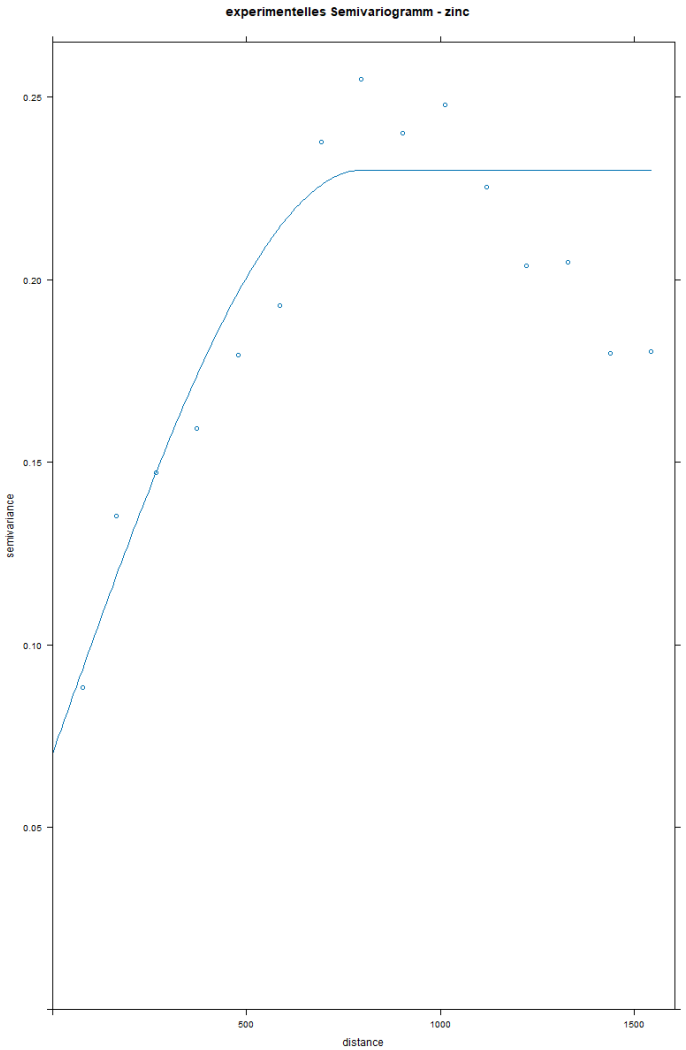
This scatter plot of log-transformed zinc concentrations (log(zinc)) versus the square root of distance (sqrt(dist)) shows a clear negative correlation, meaning that zinc levels tend to decrease as the distance from the river increases. The strong downward trend suggests that zinc contamination is concentrated near the river, likely due to industrial runoff, mining activity, or natural sediment transport. The transformation of both variables (log for zinc and square root for distance) ensures that the relationship is more linear, making it easier to model. This finding supports the hypothesis that proximity to the river is a key factor influencing zinc pollution in the area.

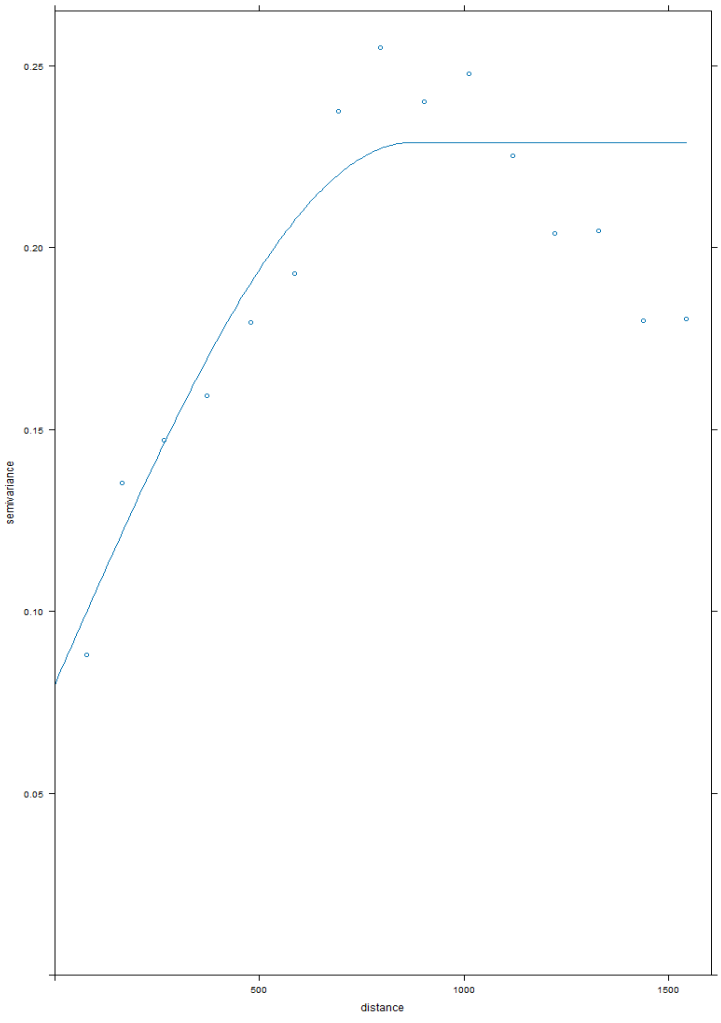




The right scatter plot of log-transformed zinc concentrations (log(zinc)) versus the square root of distance (sqrt(dist)), now with a fitted linear regression line, reinforces the strong negative correlation between zinc levels and distance from the river. The regression line highlights the downward trend, confirming that zinc concentrations are highest near the river and decrease significantly with distance. The spread of points around the line suggests some variability, meaning other factors might also influence zinc levels. However, the overall trend strongly supports the hypothesis that proximity to the river plays a key role in zinc contamination, making this relationship important for predictive modeling and environmental assessment.

The left experimental semivariogram shows how the semivariance of zinc concentrations changes with increasing spatial distance. The semivariogram is a key tool in geostatistics and Kriging, helping to quantify the spatial correlation of data points. Initially, the semivariance increases with distance, indicating that closer points have more similar zinc concentrations, while more distant points show greater variation. The curve reaches a plateau, suggesting a range beyond which spatial correlation weakens and zinc values become uncorrelated. This range is critical for interpolation models, as it defines the effective spatial influence of measurements, guiding optimal Kriging predictions.

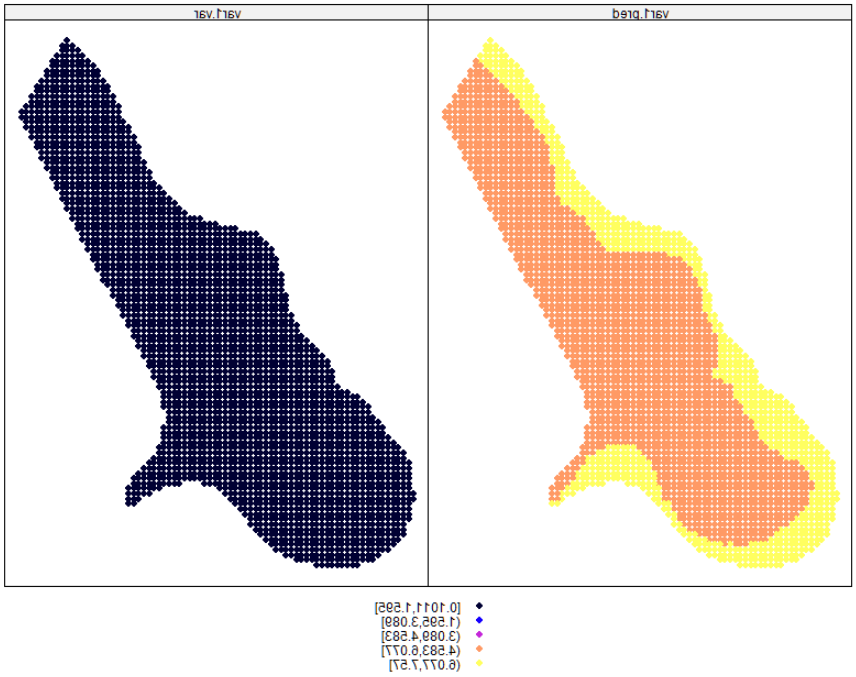


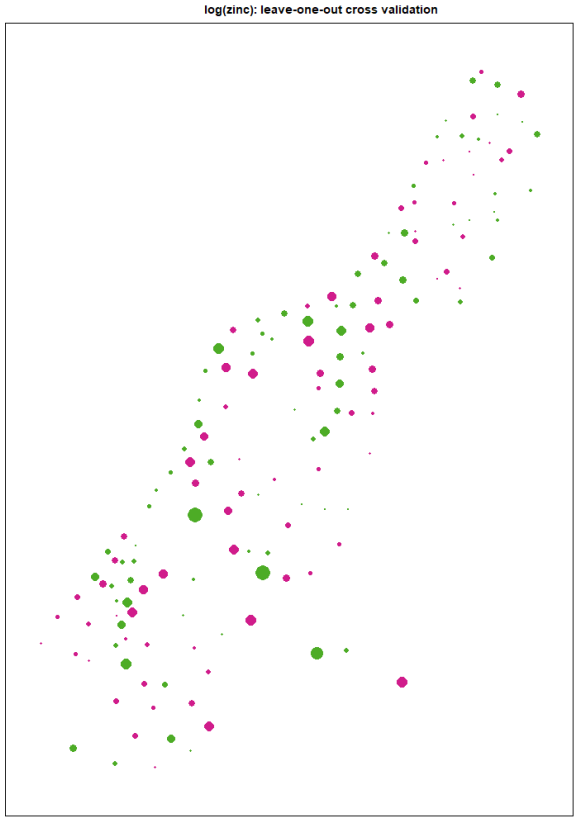


The right semivariogram with a fitted model illustrates the spatial dependence of zinc concentrations. The empirical semivariogram points (dots) represent observed semivariance at different distances, while the fitted curve represents a theoretical variogram model (likely a spherical or exponential model). The curve increases initially, indicating that nearby points are more correlated, and then flattens out, marking the range—the distance beyond which spatial correlation becomes negligible. The sill (plateau level) represents the total variance, while the nugget effect (y-intercept) accounts for measurement errors or micro-scale variability. This model is essential for Kriging interpolation, as it defines how spatial predictions are weighted across distances.

The left fitted semivariogram is a refined version of the previous plot, showing the relationship between semivariance and spatial distance. The curve follows a spherical model, where semivariance increases with distance before stabilizing at the sill, indicating the range beyond which spatial correlation is negligible. The smoothness of the fitted model suggests a well-structured spatial dependency in the zinc concentration data, meaning predictions based on Kriging interpolation will be reliable. The nugget effect (near-zero variance at

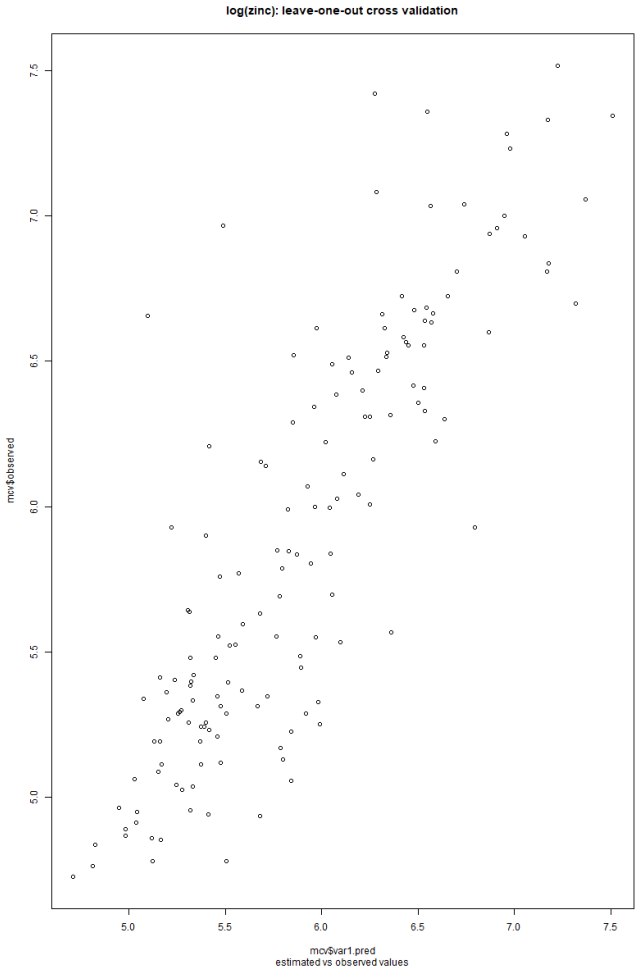
cale variation. This semivariogram is crucial





The right dual-panel spatial prediction map represents the results of ordinary Kriging interpolation for zinc concentrations. The left panel (var1.pred) shows the predicted zinc concentration across the study area, with colors ranging from low (yellow) to high (red/orange), indicating spatial variation. The right panel (var1.var) displays the kriging variance, which quantifies the uncertainty in predictions—darker areas indicate lower variance, meaning more confidence in predictions, while lighter areas suggest higher uncertainty. The variance is likely higher towards the edges where fewer observed data points are available. This visualization helps assess the spatial distribution and reliability of interpolation results.

The left leave-one-out cross-validation (LOOCV) plot evaluates the accuracy of spatial interpolation for log-transformed zinc concentrations. Each point represents a location where the prediction error was computed by excluding that point from the model. Green circles indicate positive errors (overestimations), while pink circles indicate negative errors (underestimations). The size of the circles corresponds to the magnitude of the error, with larger circles representing greater discrepancies between predicted and observed values. The distribution suggests localized biases in predictions, which can help refine the spatial model to improve accuracy.

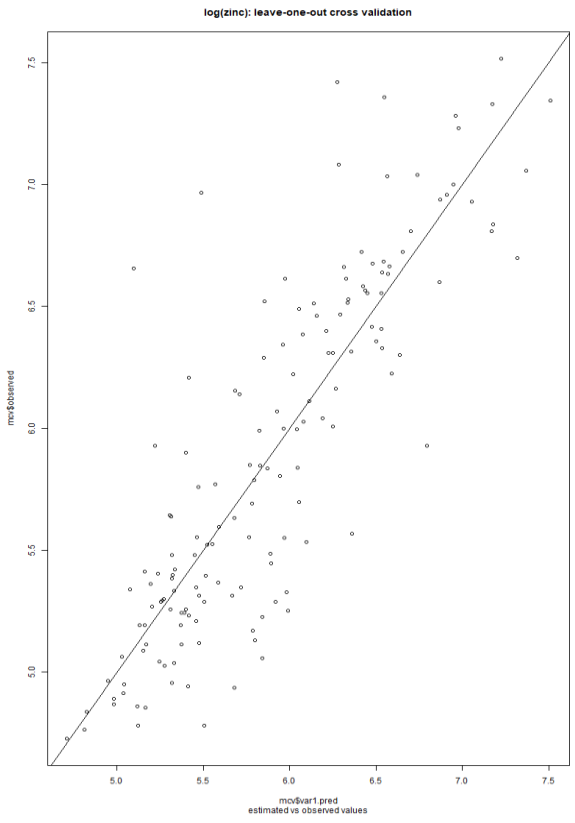


This scatter plot visualizes the leave-one-out cross-validation (LOOCV) results for the log-transformed zinc concentrations.

- The x-axis represents the predicted values from the model.
- The y-axis represents the actual observed values.
- A perfect prediction would align all points along the diagonal line y = x.

Interpretation:

- If points are close to the diagonal, the model provides accurate predictions.
- If points deviate significantly, the model has prediction errors.
- Any systematic bias (e.g., overestimation or underestimation trends) can be observed from clustering patterns.



This scatter plot shows the leave-one-out cross-validation (LOOCV) results for the log-transformed zinc concentrations, with a diagonal reference line (y = x) indicating perfect predictions.

Interpretation:

- The x-axis represents the predicted log(zinc) values.
- The y-axis represents the actual observed log(zinc) values.
- The diagonal line represents the ideal scenario where predicted values match observed values.

Key Observations:

1. Close to the diagonal → Predictions are accurate.
2. Above the diagonal → The model underestimates zinc concentration.
3. Below the diagonal → The model overestimates zinc concentration.
4. Spread from the diagonal → Higher deviation means lower prediction accuracy.

Insights:

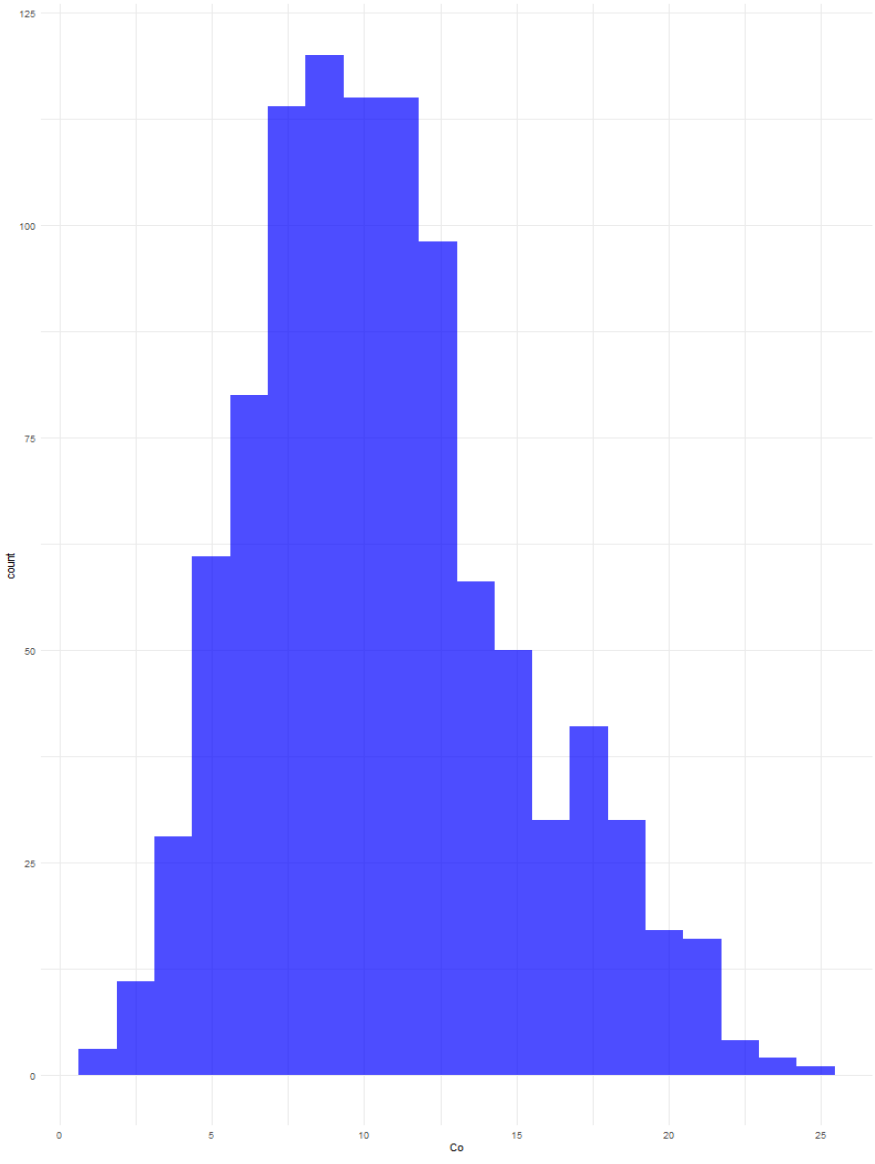
- The model seems to perform reasonably well, with most points aligning near the diagonal.
- Some outliers suggest instances where predictions significantly deviate from actual values.
- The spread of points indicates some level of uncertainty, which may require model tuning or additional data preprocessing.

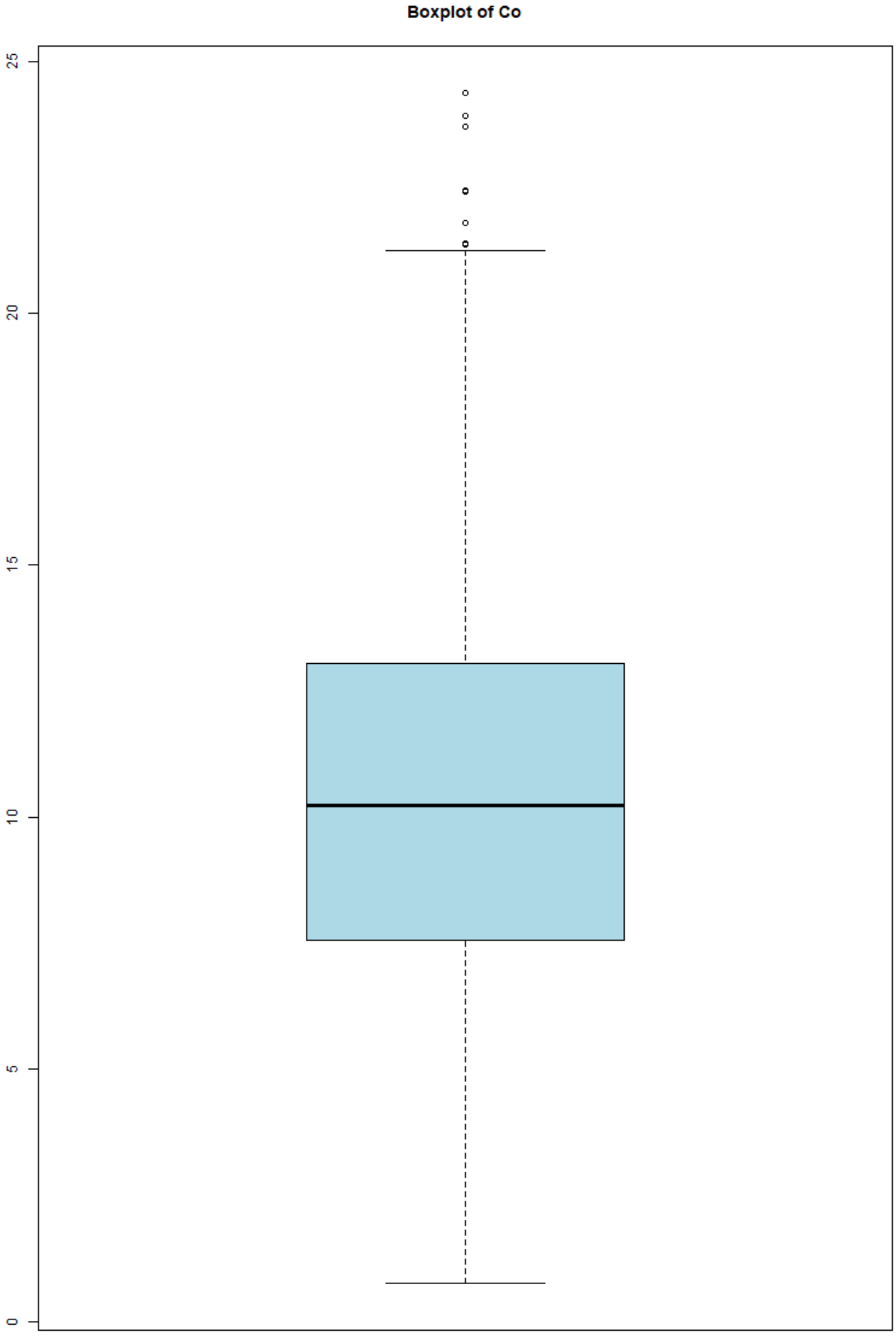
Geostatistical analysis for the testData123

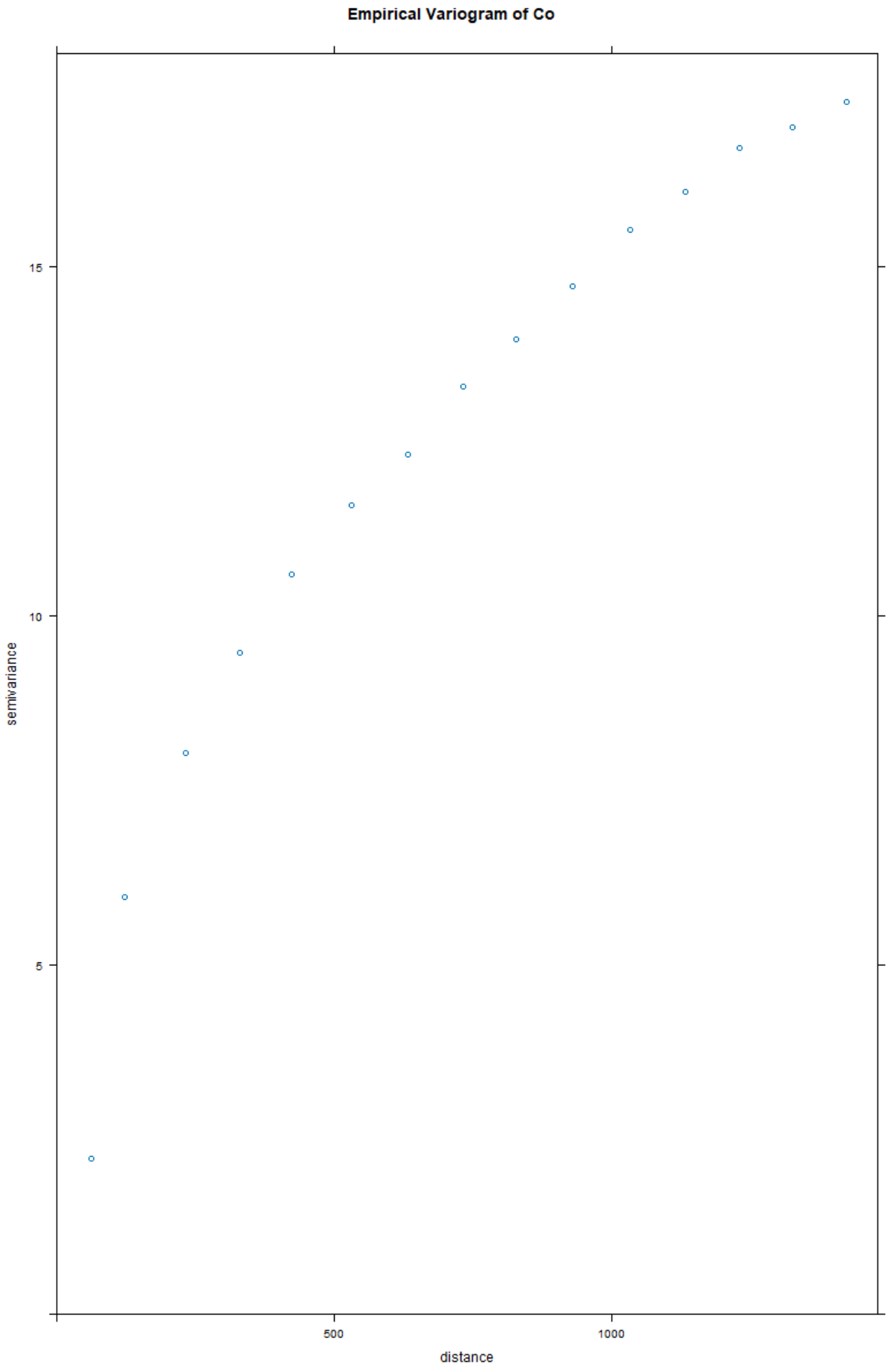
The R script written in order to:

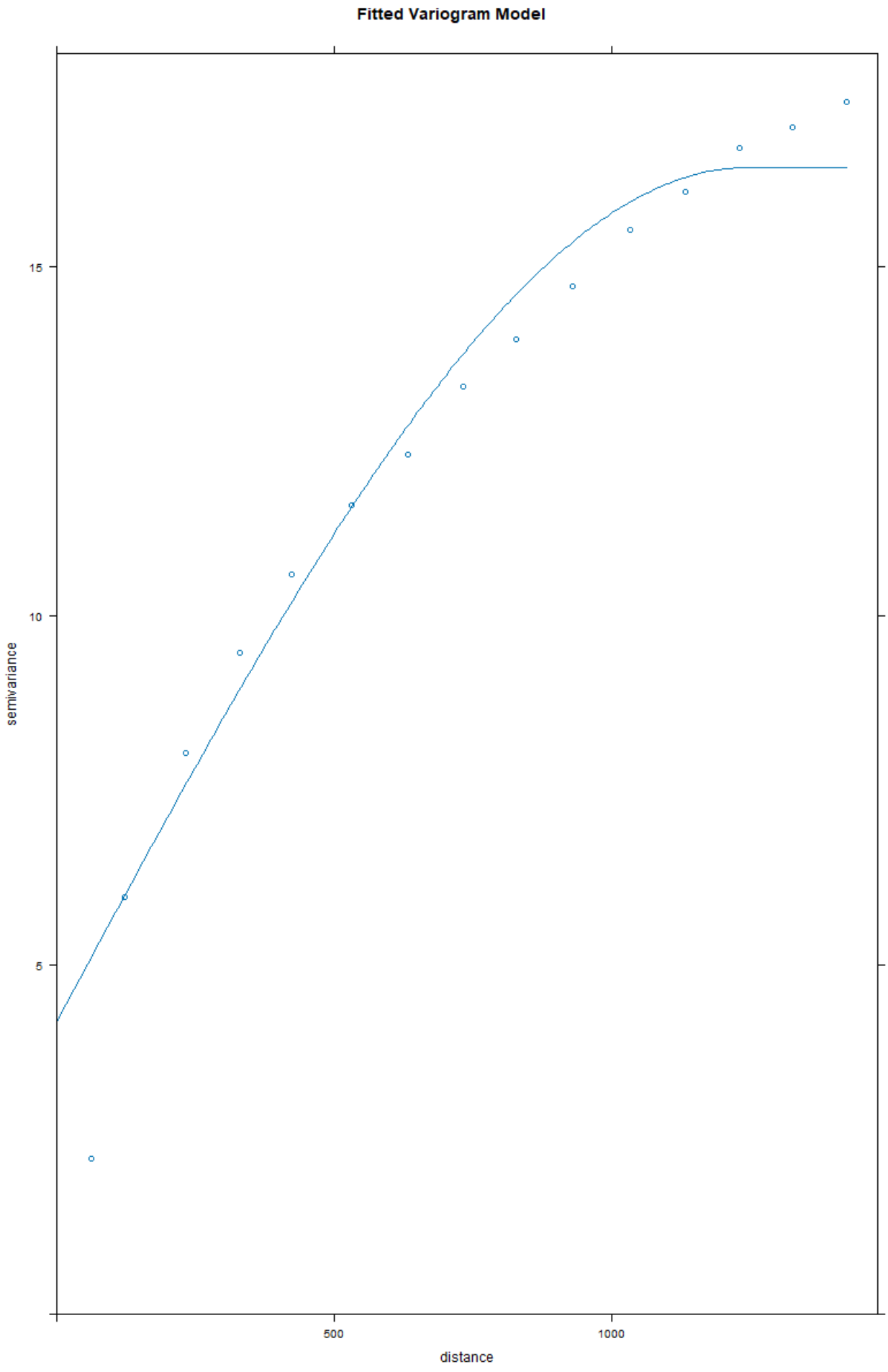
- Loading the dataset and removes unnecessary columns.
- Performing exploratory data analysis with summary statistics and plots.
- Computing the empirical variogram and fits a spherical model.
- Using Ordinary Kriging for interpolation.
- Conducting leave-one-out cross-validation (LOOCV) and computes standard deviation.
- Saving the Kriging predictions as a CSV file.

Plots and results

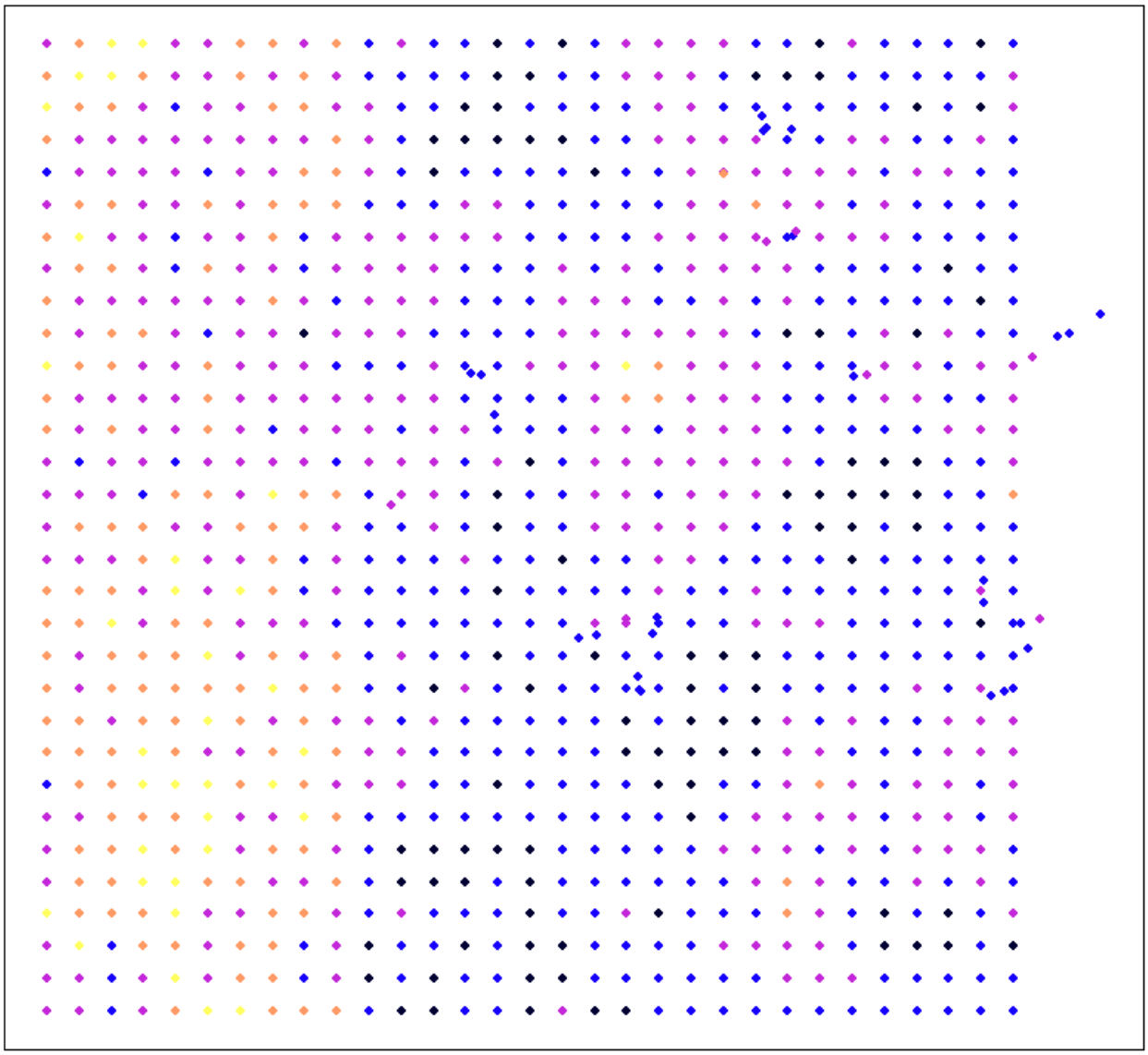



Boxplot of Co
25
20
15
10
5
0


Empirical Variogram of Co
semivariance
distance
5
10
15
500
1000


Fitted Variogram Model
semivariance
5
10
15
500
1000
distance

Kriging Prediction of Co



[0.766,5.486]
(5.486,10.21]
(10.21,14.93]
(14.93,19.65]
(19.65,24.37]

The visualizations present a thorough analysis of the dataset, revealing crucial insights into its distribution, variability, and spatial dependence. The histogram provides an overview of the distribution of cobalt concentrations, showing a skewed pattern where most values are concentrated around a central range but with a noticeable tail extending toward higher concentrations. This skewness suggests that a transformation might be needed before applying Kriging to achieve a more normal distribution. The presence of multiple peaks or clusters in the histogram could indicate spatial heterogeneity, which should be explored further through spatial statistics.

The boxplot highlights the presence of outliers in the dataset, particularly in the higher range of cobalt concentrations. The whiskers extend to a considerable range, but several extreme values are marked as distinct points beyond the upper whisker. This suggests that certain locations have significantly higher cobalt concentrations than the rest, which could be due to localized geological phenomena, measurement errors, or anomalies. Identifying these outliers is crucial before performing geostatistical modeling, as they could influence the variogram structure and the subsequent Kriging results.

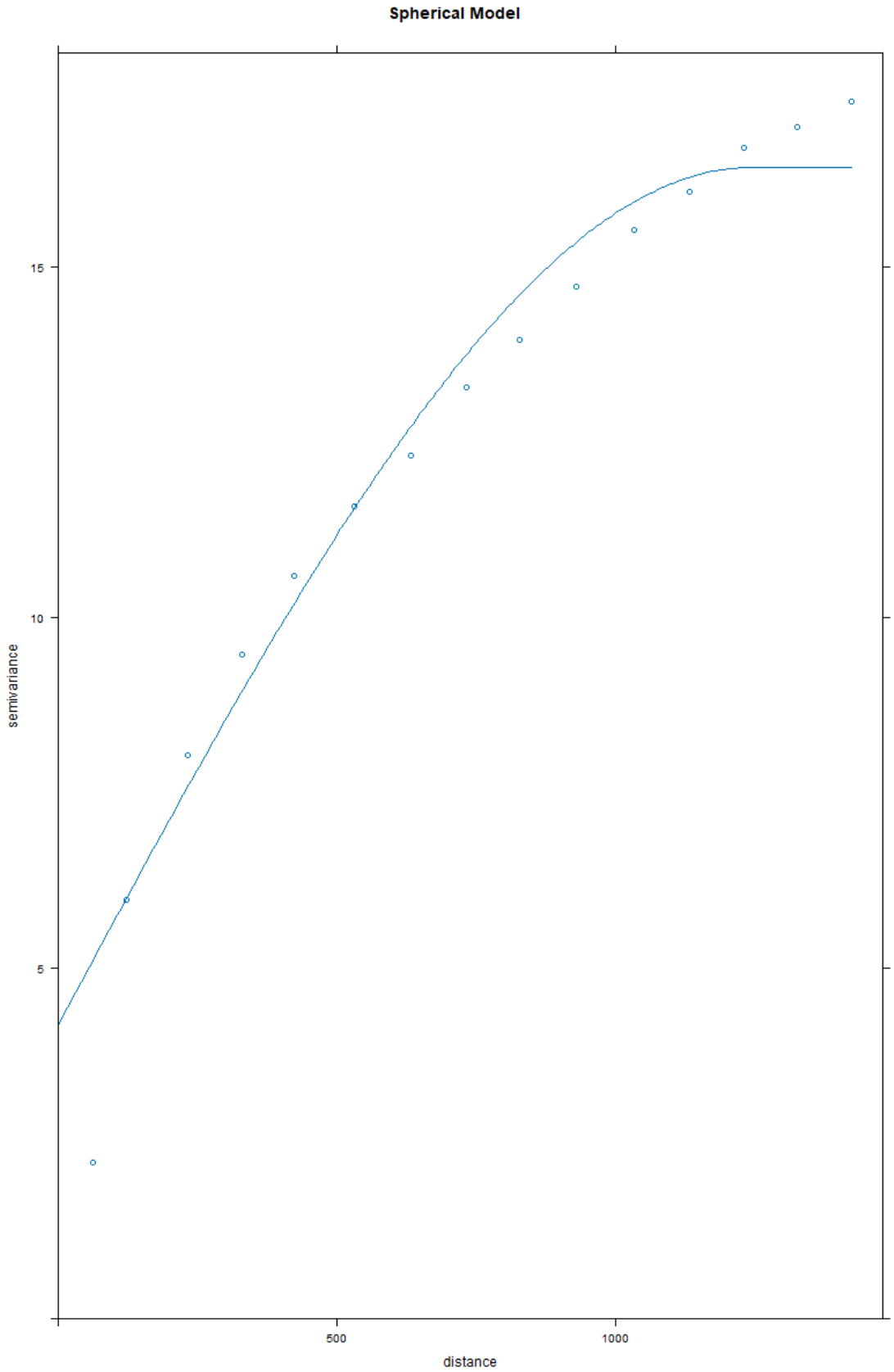
The empirical variogram is a key step in spatial analysis, depicting how the spatial correlation of cobalt concentrations changes with distance. The increasing trend of semivariance with distance suggests spatial dependence, meaning that closer points tend to have similar values while those further apart exhibit greater differences. The general pattern of the variogram appears to reach a plateau, indicating the range beyond which spatial correlation diminishes. This helps in determining the appropriate spatial model to fit for Kriging.

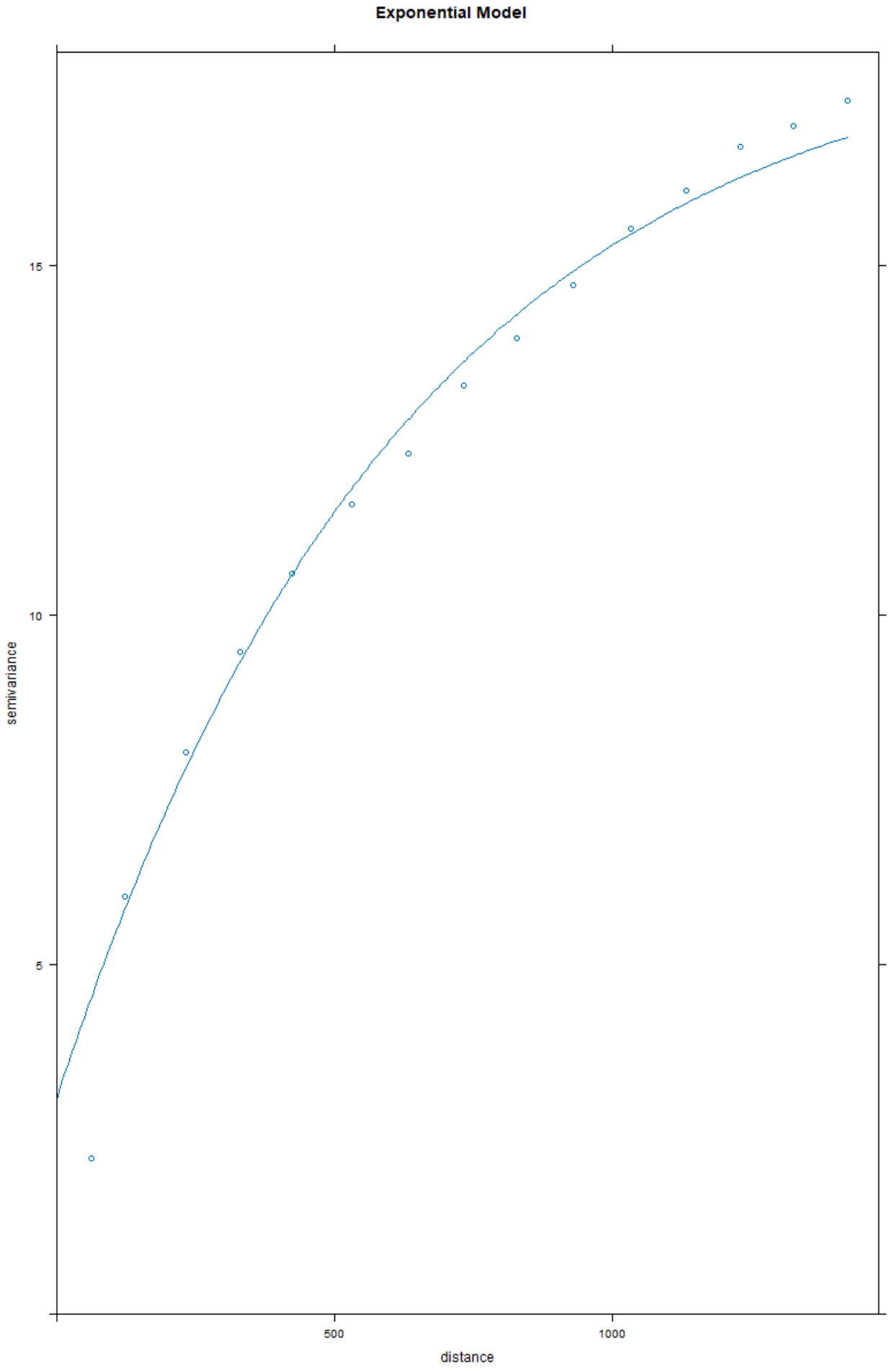
Fitting a theoretical variogram model is essential for interpolation, and the fitted variogram model effectively captures the spatial dependence structure observed in the empirical variogram. The curve shows a gradual increase in semivariance up to a sill, suggesting that the spherical model is a reasonable choice. The parameters derived from this model, such as the nugget (indicating measurement error or microscale variability), range (maximum distance of spatial correlation), and sill (total variance), are crucial for understanding the spatial behavior of the dataset and ensuring an accurate Kriging prediction.

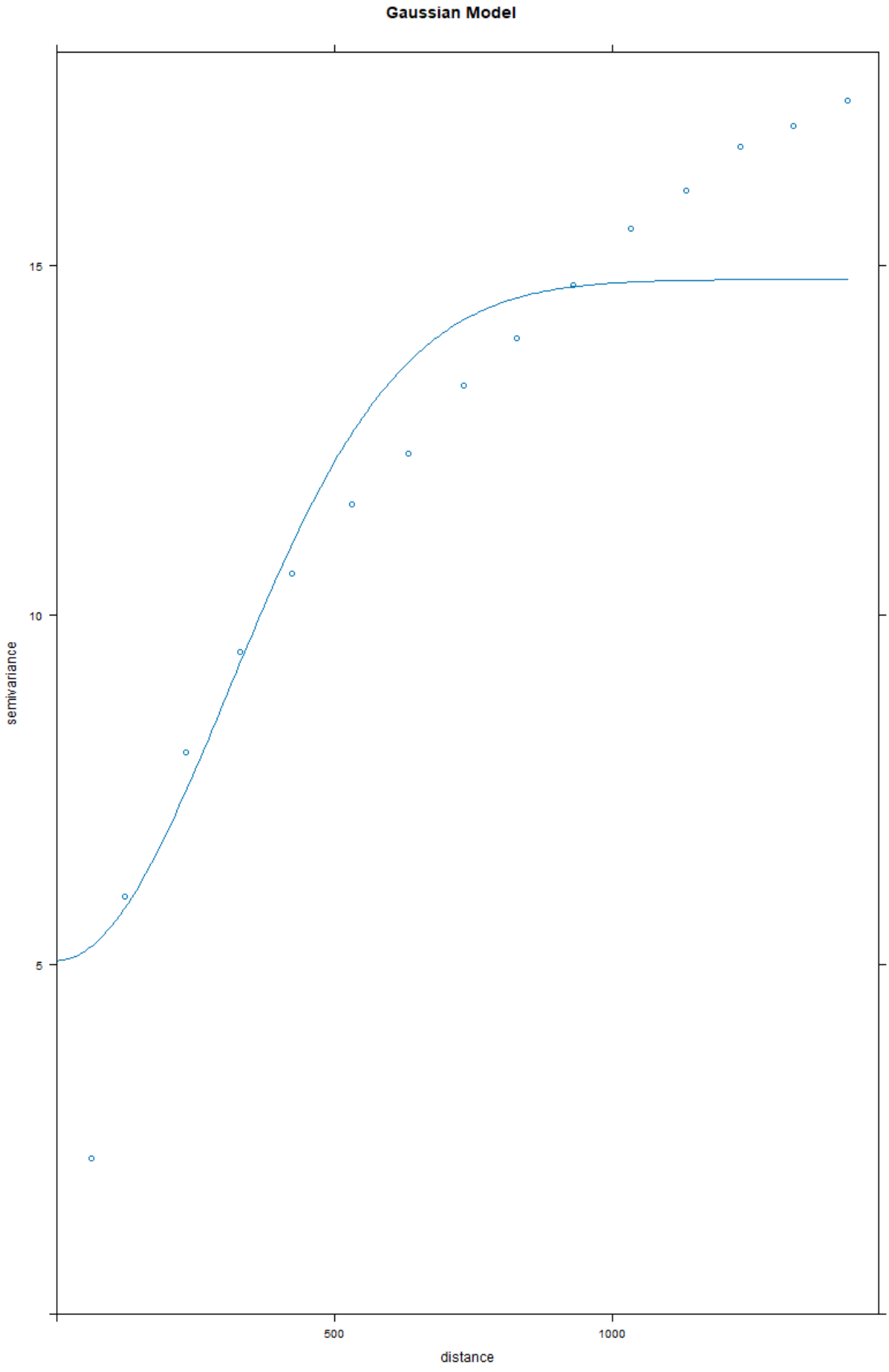
The Kriging prediction map provides a visual representation of the estimated spatial distribution of cobalt concentrations across the study area. The variation in color intensity suggests regions with higher and lower predicted values, reflecting spatial trends and potential hotspots. The influence of observed data points is visible, with smooth transitions in estimated values across space. This map is critical for decision-making in resource evaluation, environmental monitoring, and further geostatistical analysis. Additionally, assessing the uncertainty in these predictions would be beneficial, as Kriging also provides an estimation variance that helps quantify confidence in different regions of the map.

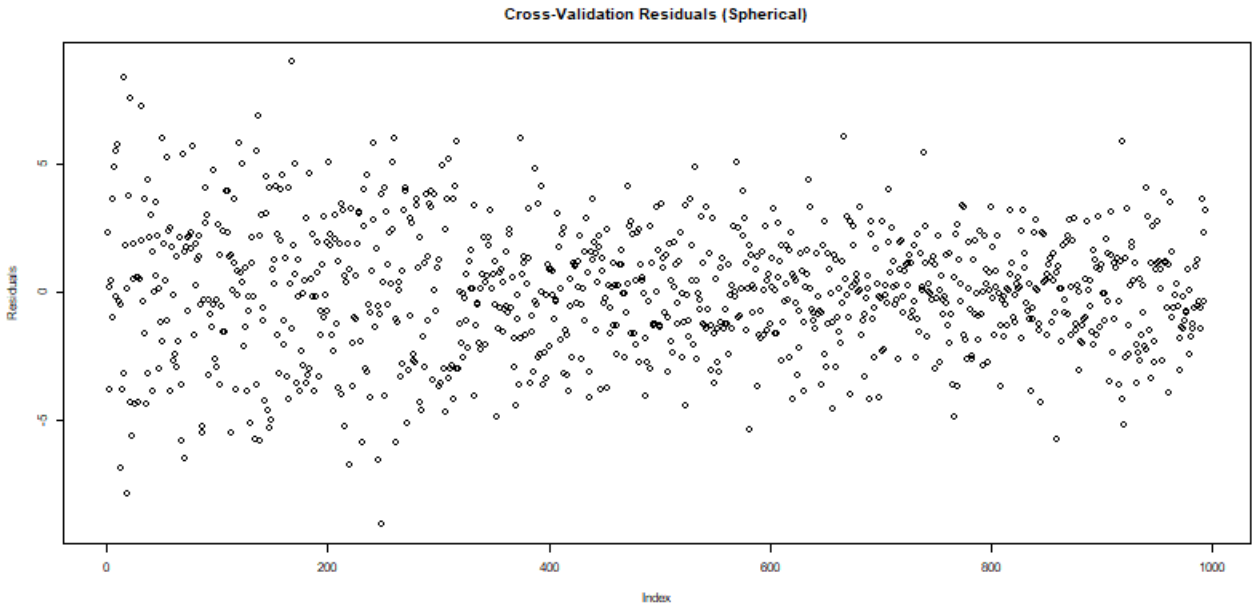
Overall, the analysis effectively explores the dataset, ensuring that spatial structures are well understood before applying Kriging. The combination of statistical summaries, exploratory data analysis, and spatial modeling allows for a robust geostatistical approach. The next step should involve cross-validation to assess the accuracy of the Kriging model and confirm that the chosen variogram parameters provide reliable predictions.

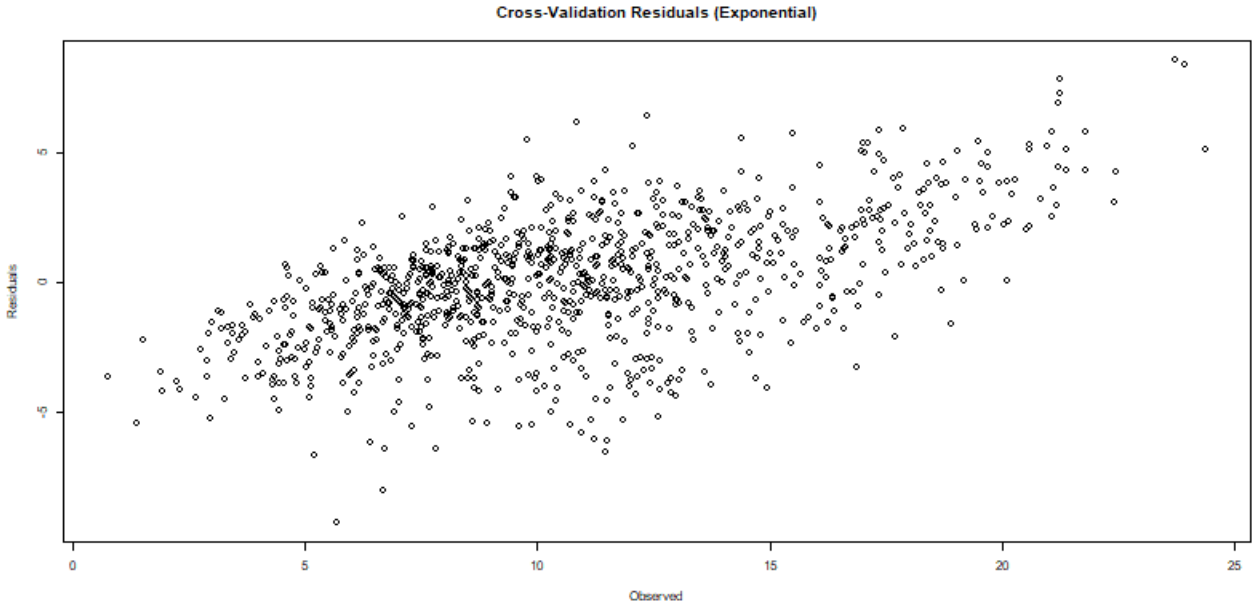
Kriging Evaluation

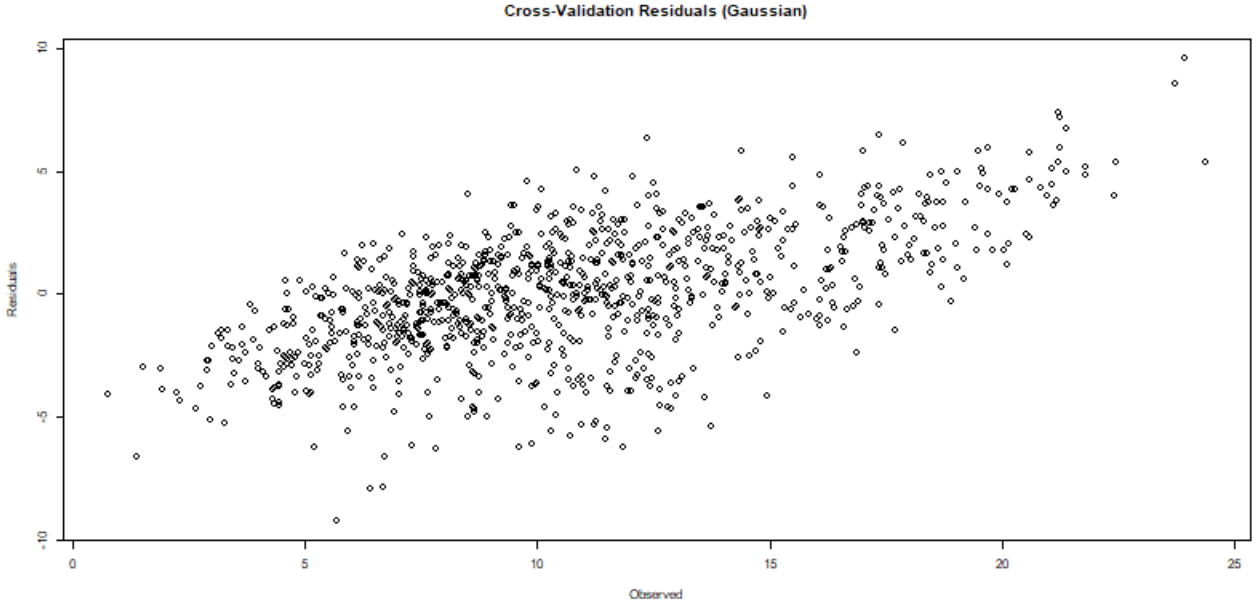



Exponential Model
semivariance
5
10
15
500
1000
distance


Gaussian Model
15
10
5
semivariance
500
1000
distance


Cross-Validation Residuals (Spherical)
Residuals
Index
5
0
-5
0
200
400
600
800
1000


Cross-Validation Residuals (Exponential)
Residuals
Observed
5
0
-5
0
5
10
15
20
25


Cross-Validation Residuals (Gaussian)
Residuals
Observed
10
5
0
-5
-10
0
5
10
15
20
25

Analysis of Plots

The variogram models provide insight into the spatial dependence of the dataset and are critical in selecting the most appropriate model for Kriging interpolation. Each model—Spherical, Exponential, and Gaussian—exhibits unique characteristics that influence the interpolation results.

The Spherical model demonstrates a characteristic behavior where the semivariance increases with distance before leveling off at a sill. This pattern suggests that beyond a certain range, additional separation between points does not significantly increase the variance. The model appears to fit the empirical variogram well, particularly in the mid-range distances, though there may be slight deviations at shorter lag distances. The leveling off of the curve indicates a clear range within which spatial correlation is significant. The Spherical model is often used in geostatistical applications when the assumption is that spatial correlation increases up to a limit and then stabilizes.

The Exponential model exhibits a continuously increasing semivariance that does not sharply reach a sill but gradually asymptotes towards it. This implies that spatial correlation exists over longer distances but decreases progressively rather than abruptly. The Exponential model fits the empirical variogram well and appears smoother compared to the Spherical model. It suggests that the dataset may contain long-range spatial dependencies, which are not captured as distinctly in the Spherical model. The fit is particularly strong in the short to mid-range distances, making it a strong candidate for interpolation, especially if data points have continuous correlation beyond the range.

The Gaussian model, unlike the previous two, initially rises more gradually before increasing more steeply and eventually reaching a sill. This suggests that nearby points exhibit a high degree of similarity, but the correlation structure changes more abruptly at greater distances. The fit to the empirical variogram is relatively good, though it may not be as strong in capturing mid-range dependencies as the Exponential model. The Gaussian model is typically used when smooth transitions in spatial variation are expected. However, it can sometimes overestimate correlation at very short distances while underestimating it at longer distances.

The cross-validation residual plots further provide an evaluation of each model's predictive accuracy. The residual distribution in the Spherical model appears to be fairly uniform, with no significant clustering or patterns, suggesting that the model does not introduce systematic bias. However, the presence of a few extreme residuals suggests that some locations may not be well-represented by the model. The Exponential model shows a more compact spread of residuals, with fewer extreme outliers compared to the Spherical model. This supports the previous conclusion that the Exponential model offers the best fit for the dataset, as it minimizes prediction errors. The Gaussian model residuals display a slightly wider spread, with more points exhibiting higher absolute residual values. This indicates that while the Gaussian model provides a reasonable interpolation, it tends to overfit or underfit certain spatial variations in the data.

Taken together, these results suggest that the Exponential model provides the best balance between capturing spatial correlation and minimizing prediction errors. While the Spherical model is also a strong contender, particularly when a well-defined correlation range is observed, its performance is slightly weaker in terms of residual distribution. The Gaussian model, while smooth and suitable for gradual transitions, appears to have larger prediction deviations, making it less ideal for this dataset. Ultimately, the Exponential model emerges as

the most effective for Kriging interpolation in this case, as it successfully balances local and global spatial dependencies.

Variogram Model	Standard Deviation of Residuals	Min Residual	Max Residual	Mean Residual	Median Residual
Spherical	2.443781	-9.085612	8.964888	0.000784	0.060138
Exponential	2.409647	-9.24025	8.53923	0	0.04579
Gaussian	2.523507	-9.265885	9.593744	0.001137	0.035376

The Exponential variogram model demonstrates the best performance for Kriging interpolation, with the lowest standard deviation of residuals (2.409647), indicating greater accuracy and stability. The Spherical model follows closely (2.443781), while the Gaussian model has the highest residual spread (2.523507), making it the least reliable.

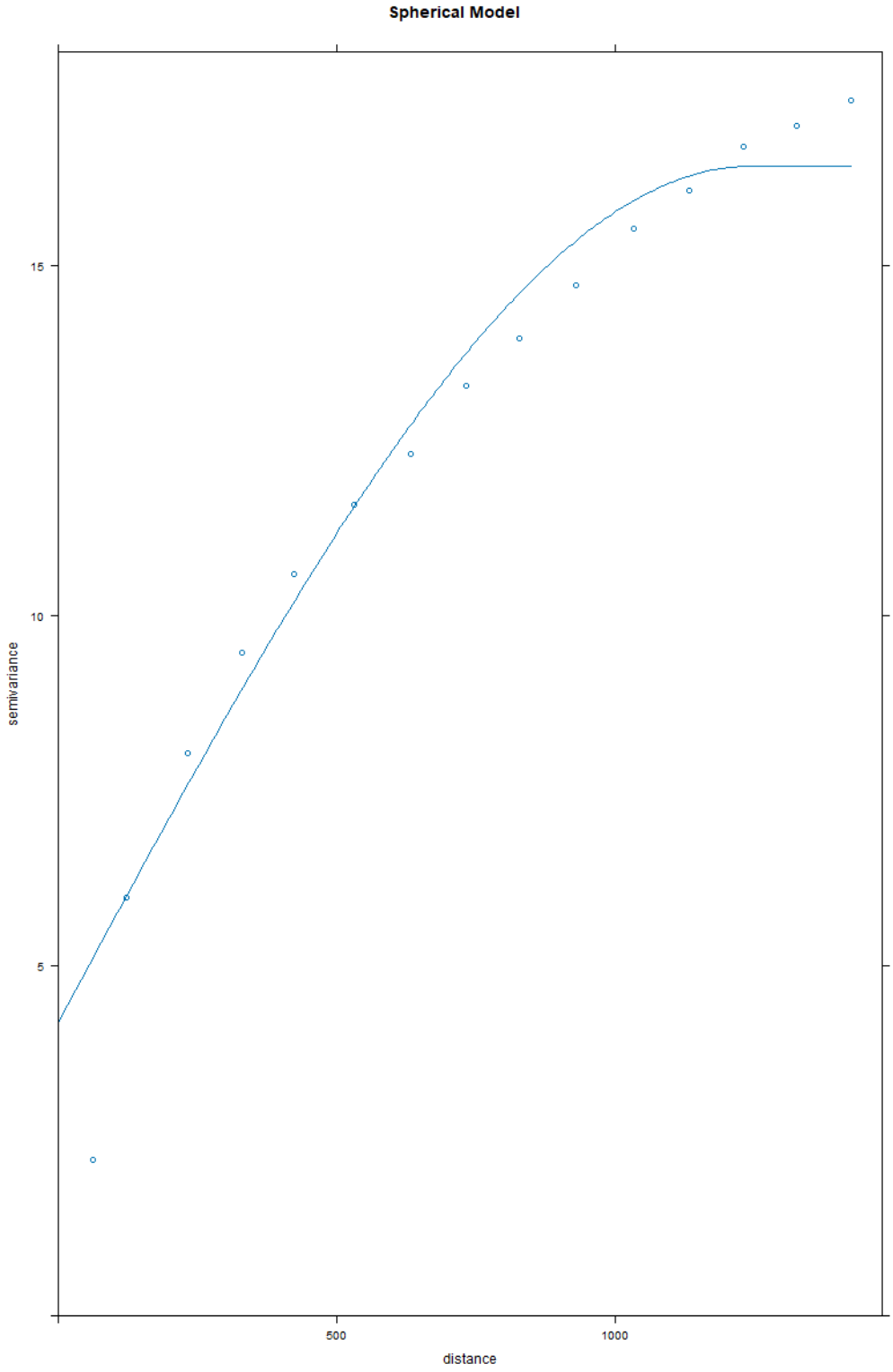
Looking at minimum and maximum residuals, the Exponential model has the smallest range, meaning it produces fewer extreme errors. The Gaussian model has the largest range, suggesting greater variability in prediction accuracy. The Spherical model falls between these two, making it a viable alternative.

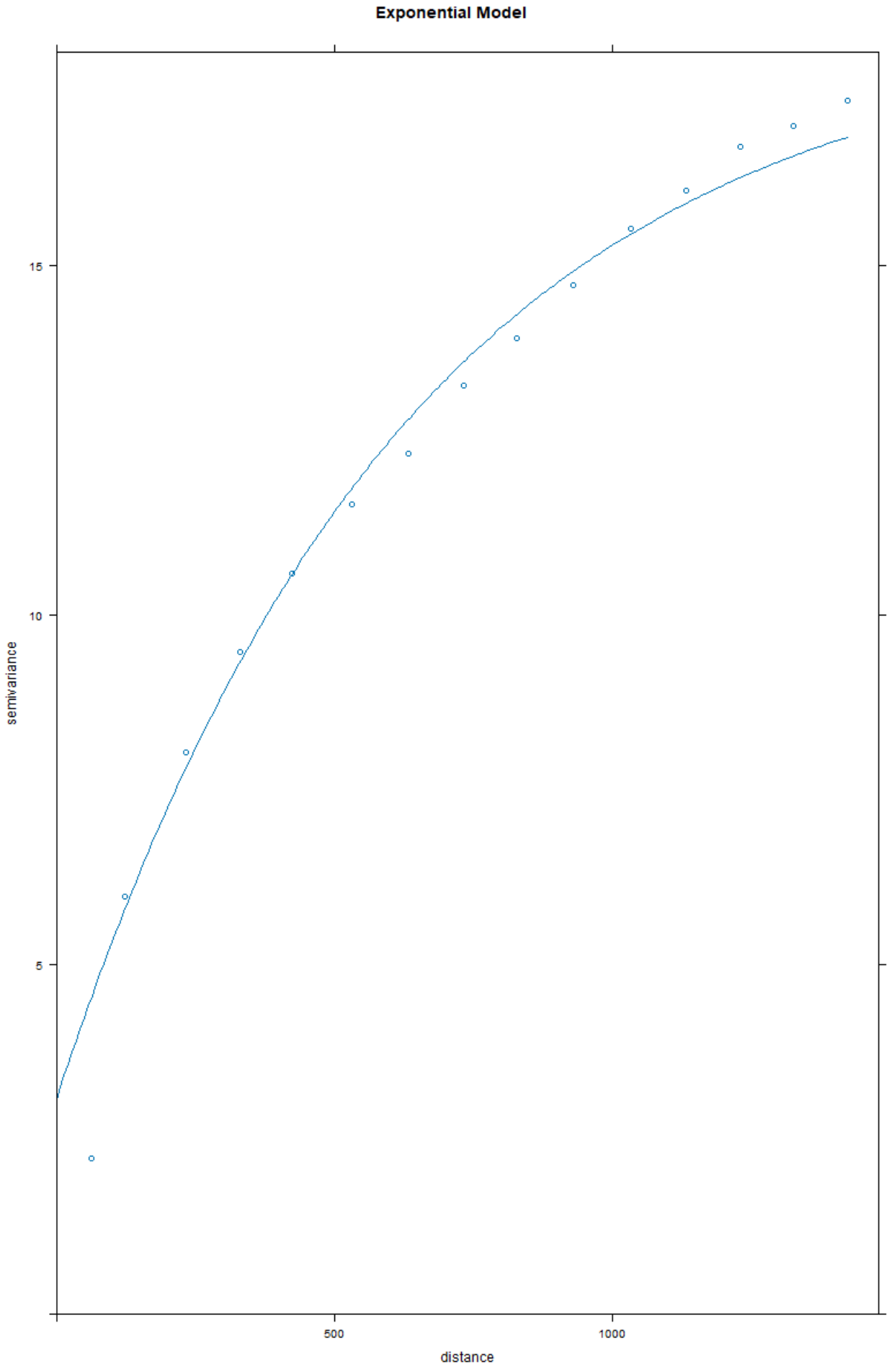
Cross-validation results confirm that the Exponential model provides the most balanced error distribution, with fewer large deviations. The Spherical model also performs well but shows slightly higher variability. The Gaussian model struggles with localized variations, leading to a wider spread of residuals.

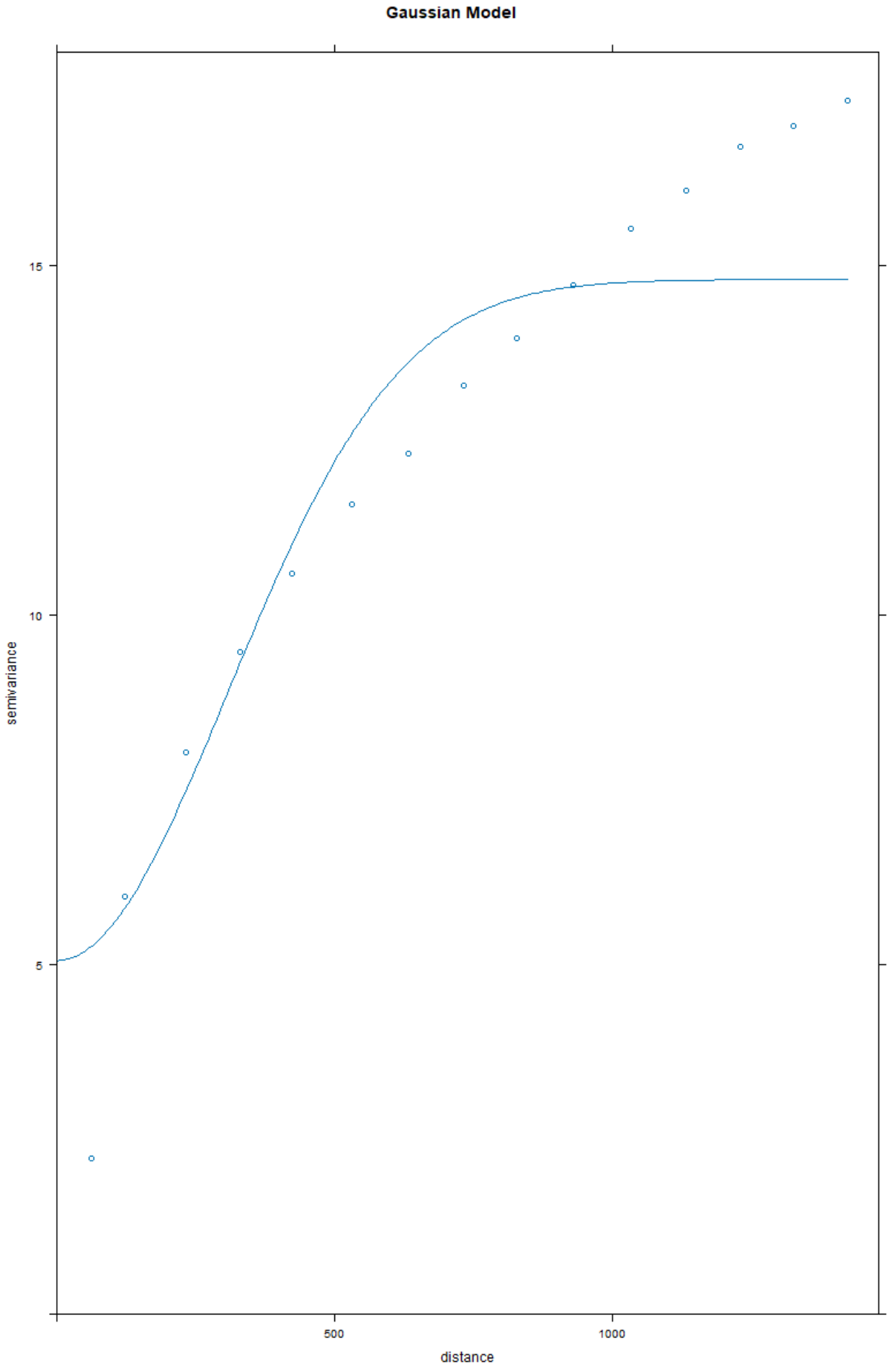
Overall, the Exponential model is the most suitable for interpolation, offering better accuracy, lower prediction errors, and a more stable variogram fit. The Spherical model is a reasonable alternative, while the Gaussian model introduces higher uncertainty and is less reliable for this dataset.

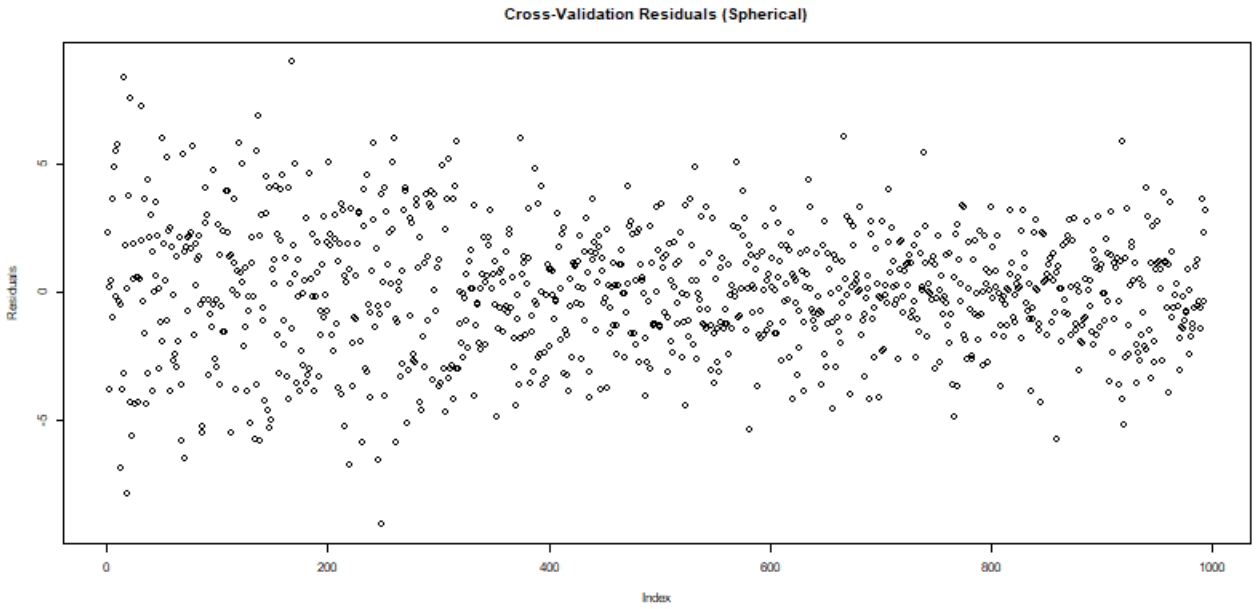
Kriging comparison results

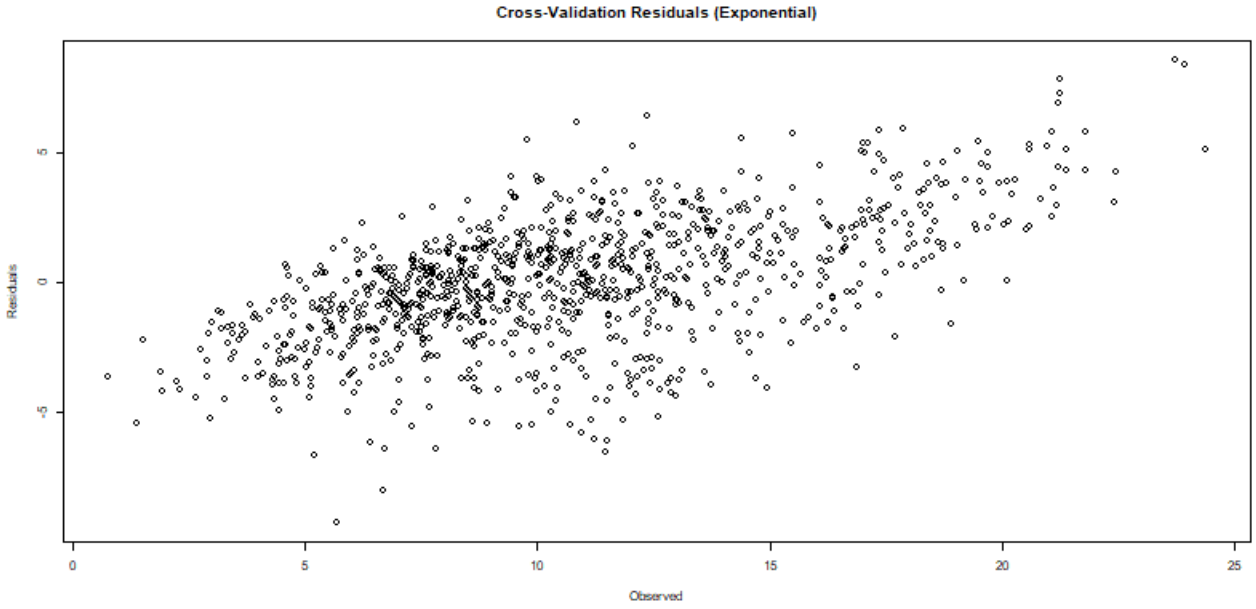
Whole dataset Variograms and Kriging

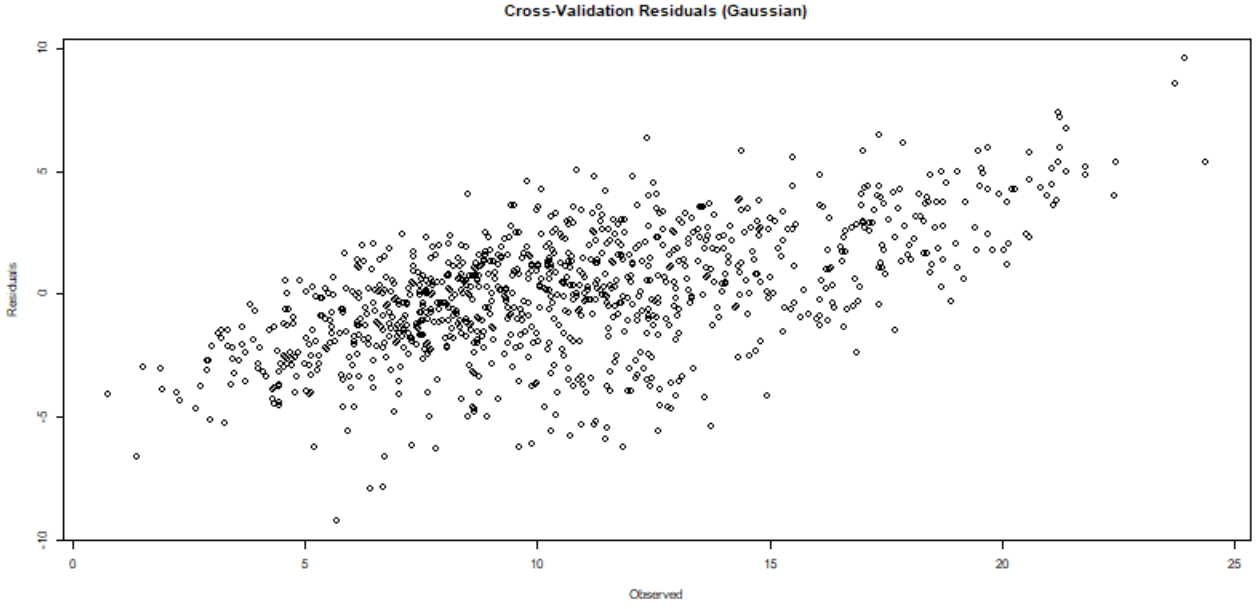



Exponential Model
semivariance
5
10
15
500
1000
distance

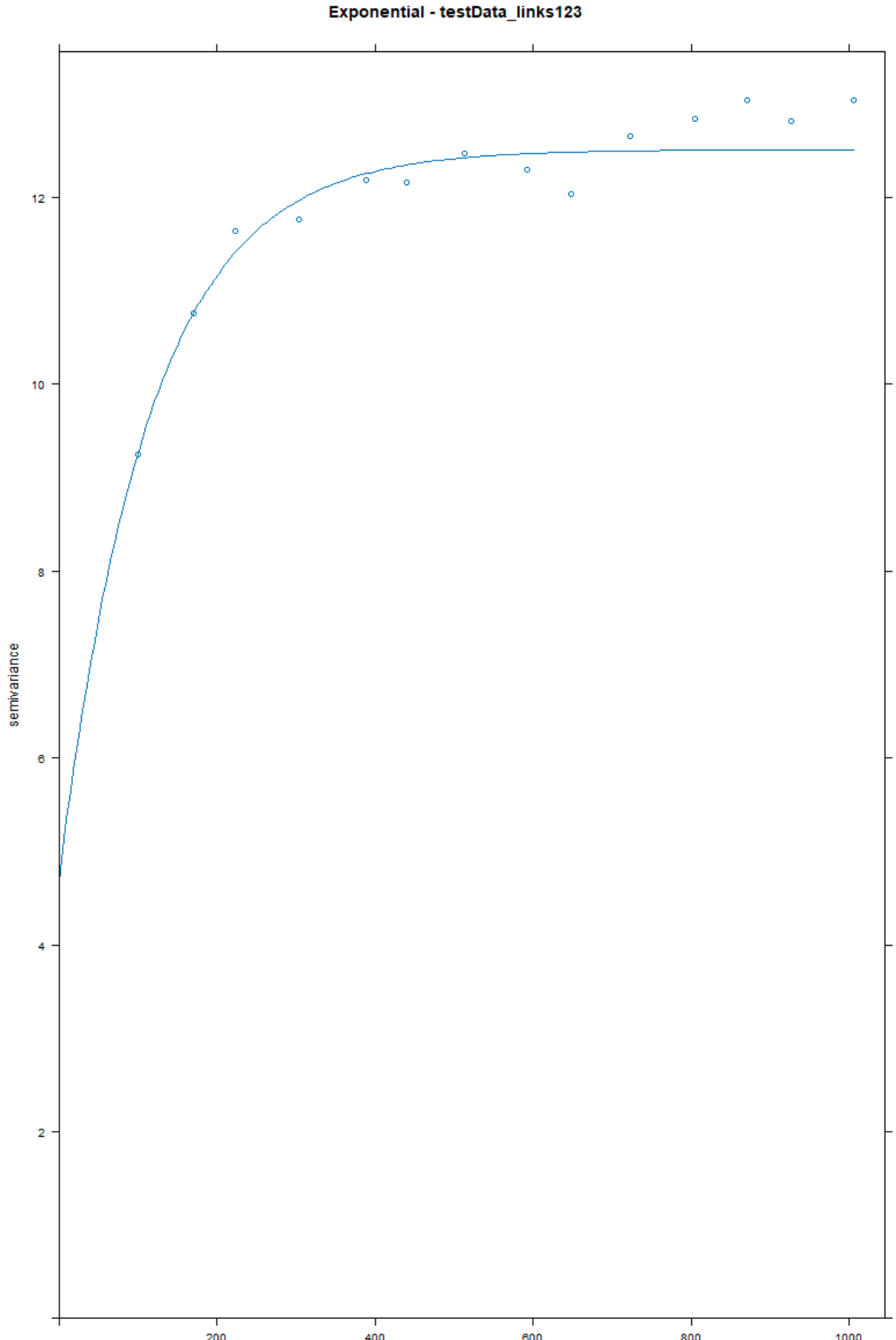

Gaussian Model
15
10
5
semivariance
500
1000
distance

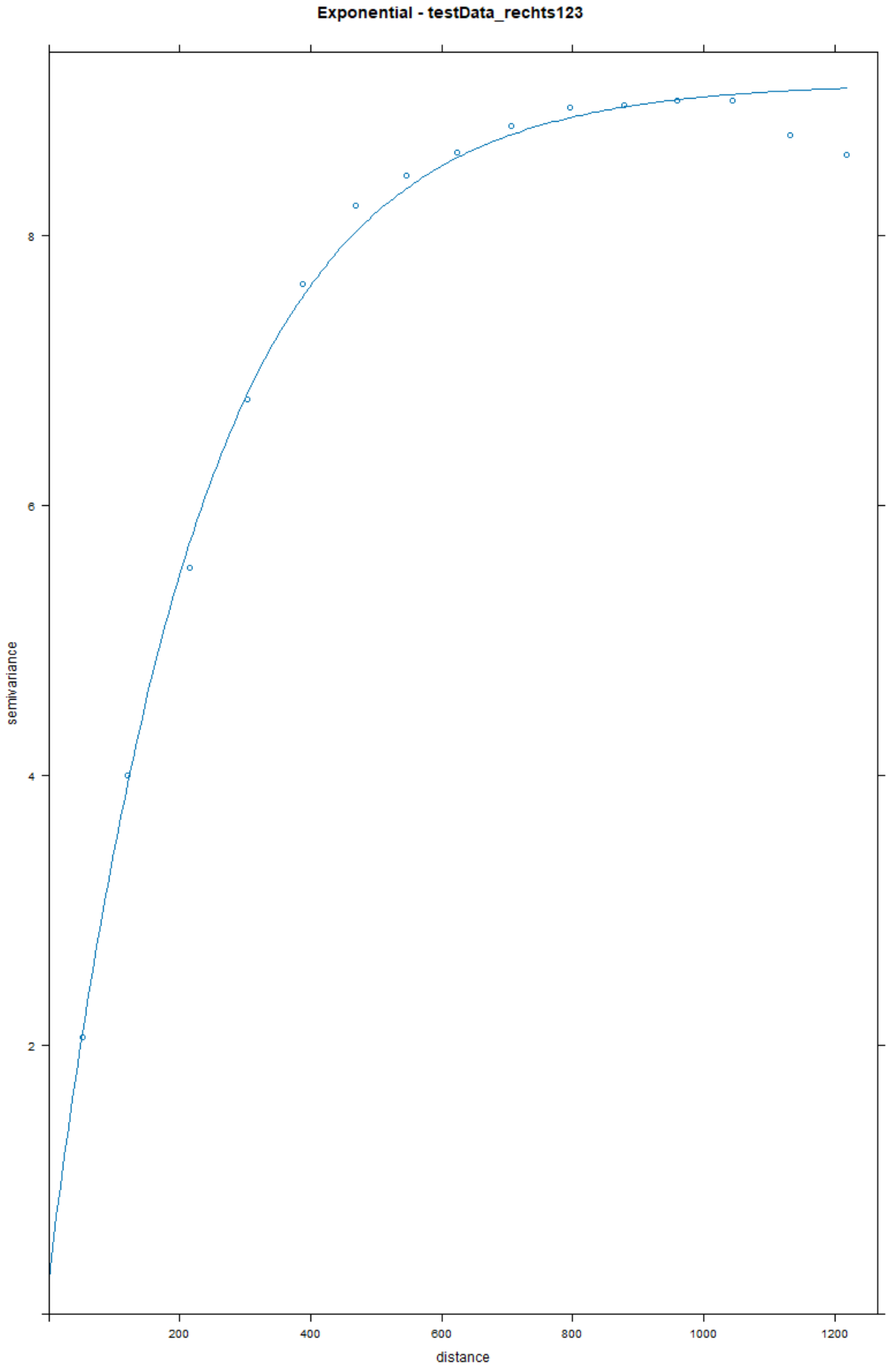

Cross-Validation Residuals (Spherical)
Residuals
Index
5
0
-5
0
200
400
600
800
1000

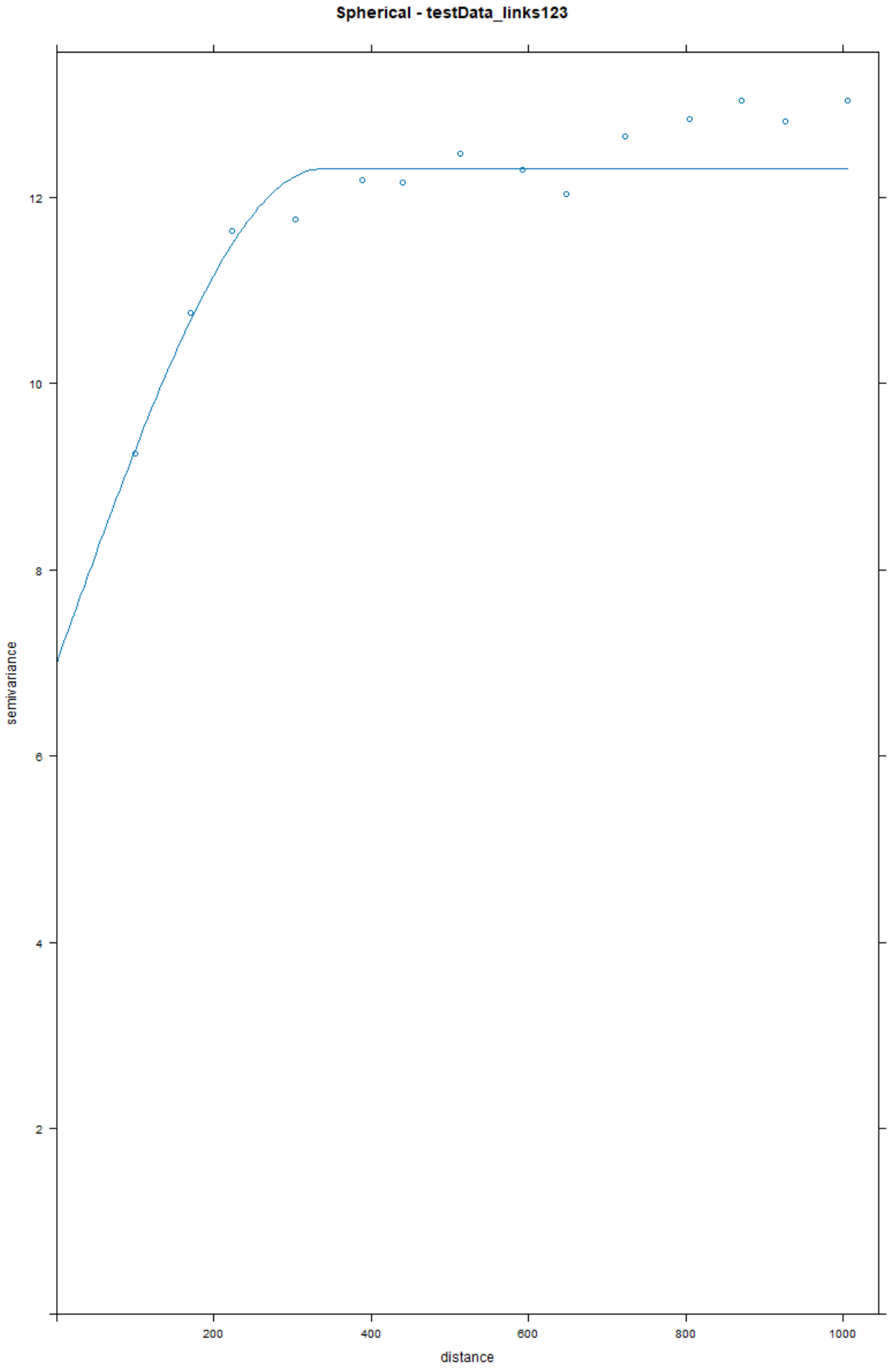

Cross-Validation Residuals (Exponential)
Residuals
Observed
5
0
-5
0
5
10
15
20
25

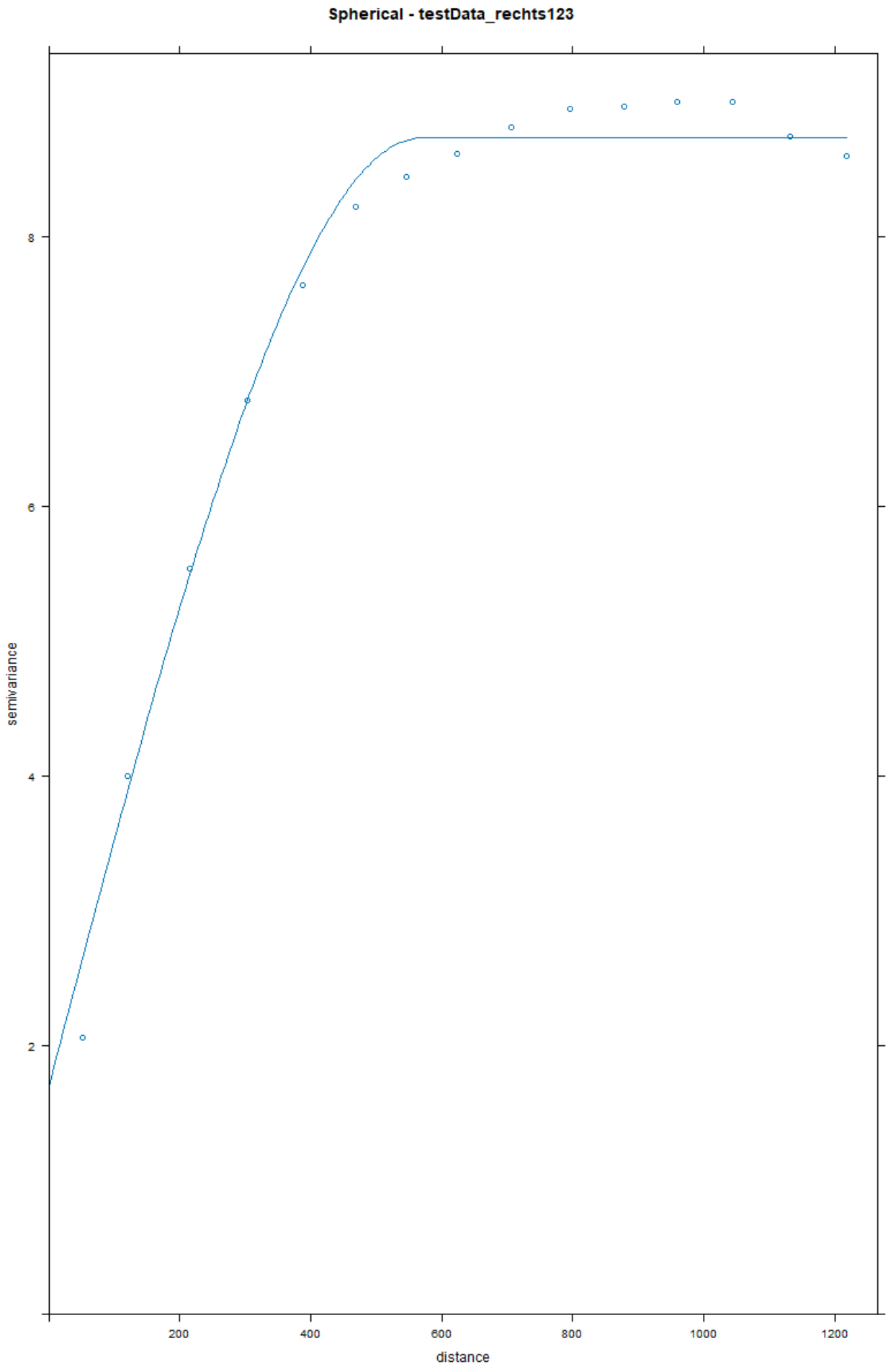

Cross-Validation Residuals (Gaussian)
Residuals
Observed
10
5
0
-5
-10
0
5
10
15
20
25

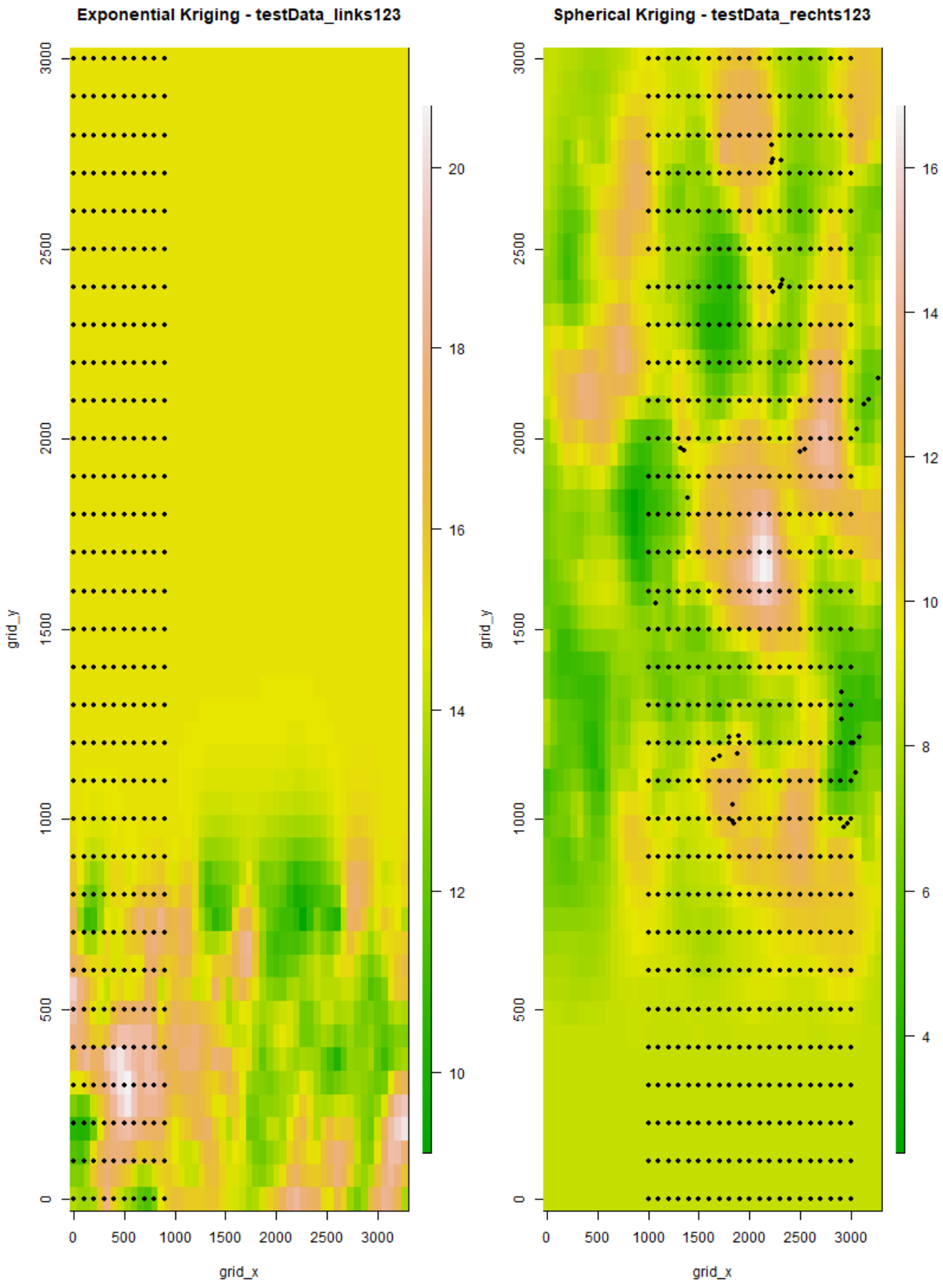
Rechts and Links datasets




Exponential - testData_rechts123
semivariance
distance
2
4
6
8
200
400
600
800
1000
1200


Spherical - testData_links123
semivariance
distance
2
4
6
8
10
12
200
400
600
800
1000


Spherical - testData_rechts123
semivariance
distance
2
4
6
8
200
400
600
800
1000
1200


Exponential Kriging - testData_links123
grid_y
grid_x
0
500
1000
1500
2000
2500
3000
10
12
14
16
18
20
Spherical Kriging - testData_rechts123
grid_y
grid_x
0
500
1000
1500
2000
2500
3000
4
6
8
10
12
14
16

Analysis

The variogram analysis of the whole dataset compared to the two subsets (links and rechts) reveals distinct spatial correlation patterns, particularly in the Spherical and Exponential models. For the whole dataset, the Spherical variogram shows a gradual increase in semivariance before reaching the sill at a larger range, indicating that spatial correlation extends over a longer distance. This suggests that the whole dataset exhibits a more continuous spatial structure where observations remain correlated over a wider area. In contrast, when looking at the Spherical variograms of the links and rechts subsets separately, the range is significantly shorter, meaning spatial correlation diminishes more quickly. The sill is reached at a lower distance, confirming that within these smaller regions, variations occur more abruptly.

The Exponential variogram of the whole dataset further reinforces this contrast by showing a faster rise toward the sill, though it still maintains a longer range than the subset variograms. This model captures more localized variations compared to the Spherical model but still indicates that some degree of spatial dependence exists over a broader area. When examining the Exponential variograms of the links and rechts subsets, the spatial correlation is even more localized, with a much sharper increase in semivariance at short distances. The range for the subsets is much smaller than in the whole dataset, showing that in these regions, the variations are more abrupt, and spatial continuity does not extend far. The Exponential variograms for links and rechts depict a scenario where observations lose their correlation quickly, highlighting the need for a model that prioritizes local spatial relationships over broader trends.

When considering the Kriging results, the whole dataset produces smoother interpolations for both the Spherical and Exponential models. The Spherical Kriging results for the whole dataset show a gradual transition in predicted values, reflecting its longer spatial correlation range. Predictions do not change drastically over short distances, indicating that the spatial process captured by this model assumes a relatively even distribution of variations. However, when the same Spherical Kriging approach is applied to the links and rechts subsets, the predicted values exhibit sharper transitions, aligning with the shorter correlation range found in their variograms.

The Exponential Kriging results, on the other hand, behave quite differently. In the whole dataset, Exponential Kriging captures more localized variations compared to the Spherical Kriging, but it still maintains a level of spatial smoothness due to the dataset's broader trends. However, when applied separately to the links and rechts subsets, the Exponential Kriging results display more abrupt changes in predicted values, closely following the short-range spatial correlation observed in their variograms. The rapid shifts in predictions emphasize the Exponential model's strength in handling localized variations.

The differences between the whole dataset and the subsets highlight the importance of selecting the right variogram model. The Spherical model suits long-range correlations, making it ideal for the whole dataset, while the Exponential model captures short-range variations better, as seen in the links and rechts subsets. Kriging results confirm this, with Spherical Kriging providing smoother predictions and Exponential Kriging offering more localized detail. For overall consistency, the Spherical model is preferable, but for higher accuracy in smaller regions, the Exponential model is more effective.