Project X. Spatial Interpolation for Digital Elevation Generation

Abstract

Spatial interpolation serves as a cornerstone technique in geospatial data analytics, offering a methodology to predict unknown point values within a specific domain. In this study, I employed Inverse Distance Weighting (IDW), universal kriging, and K-Nearest Neighbors (KNN) to interpolate a spatial dataset consisting of 10,000 sample points across Indiana. The primary objective was to generate a comprehensive Digital Elevation Model (DEM) that accurately represents the state's topographical variations. Preliminary data cleaning involved the removal of anomalies and outliers, ensuring a robust dataset for analysis.

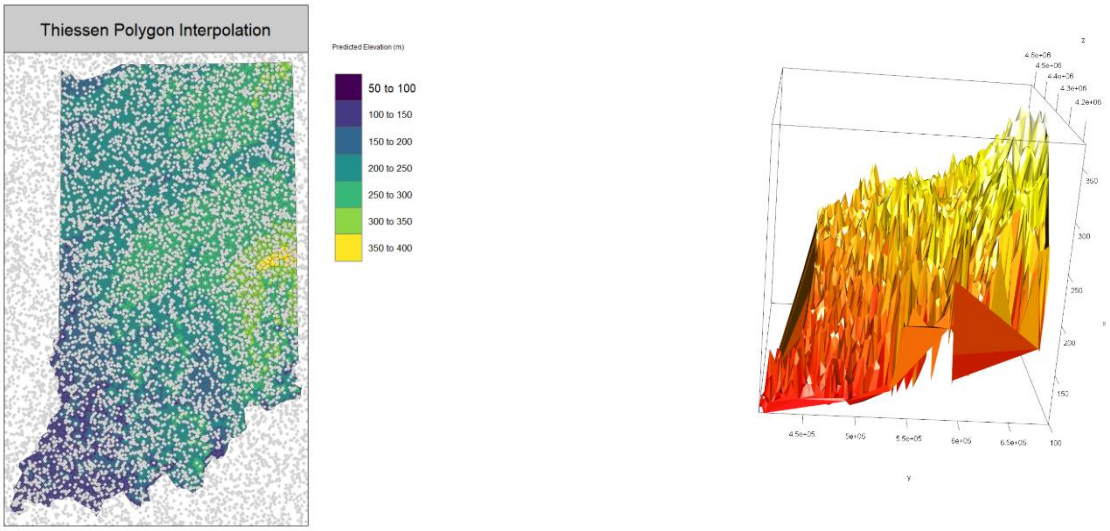
1. Data Cleaning and Exploration

In this analysis, we applied spatial interpolation techniques to a dataset of 10,000 sample points across Indiana to predict elevation differences and model the state's topography. Using the Dirichlet-Voronoi diagram, the regions were divided into tessellated areas around each point to provide a base for interpolation.

The two-dimensional Thiessen polygon map (Figure 1) depicts a granular elevation profile of Indiana with a varied color scheme, ranging from light blue (50 to 100 meters) to yellow (350 to 400 meters). This visualization highlights a distinct pattern of elevation; the central region is characterized by intermediate elevations, transitioning to higher altitudes towards the east. The 3D topography may (Figure 01) provides a view of the state's terrain, with notable elevation in the southwestern regions are lower while the northern region are relatively higher.

Figure 01

2D Thiessen Polygon and 3D Topography



2. Apply the IDW Method

In this section of the spatial interpolation study, the Inverse Distance Weighting (IDW) method was implied with varying power parameters (p = 1, 2, and 3) to a sample of 10,000 points across Indiana. Upon increasing the power parameter from 1 to 3, the results demonstrated a discernible refinement in the interpolated elevation data. For p = 1, the resultant Digital Elevation Model (DEM) displayed broad generalization, with large areas sharing similar elevation bands, as seen in the broad swaths of color. Adjusting the power to p = 2 yielded a more differentiated elevation model, with finer distinctions between adjacent elevation zones. At p = 3, the interpolation produced the most detailed and localized variations in elevation, resembling the more topographical complexity of Indiana. The finer nuances of the state's elevation are captured, and smaller geographic features become more apparent (Figure 02).

Figure 02

IDW Interpolation Results.

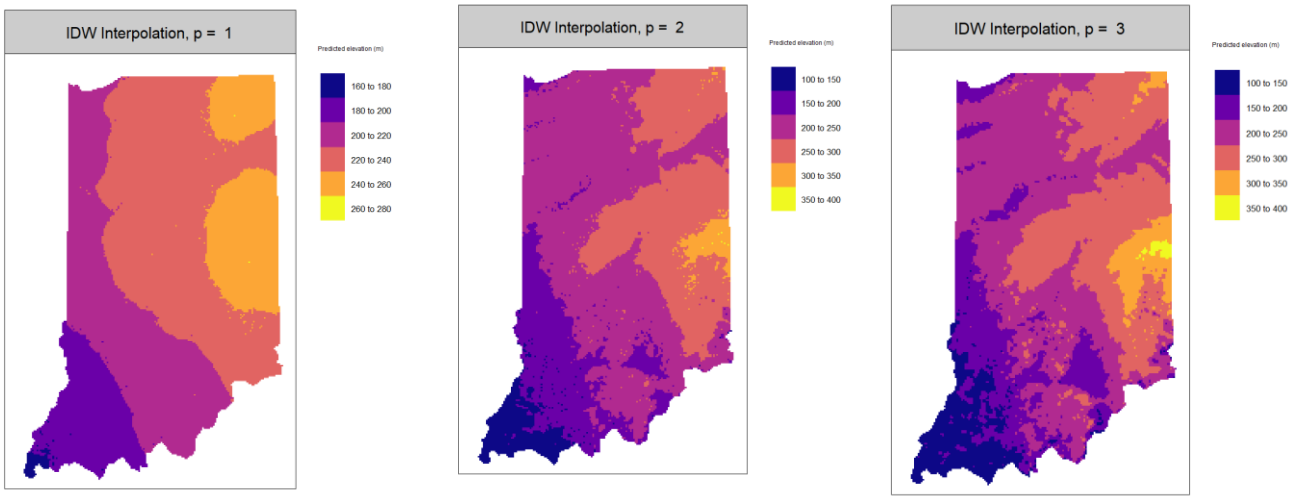


Table 01 below shows the result of Root Mean Square Error (RMSE) of the IDW interpolation method under different power.

	P = 1	P = 2	P = 3
RMSE	36.58053	17.33036	13.48662

The RMSE values obtained were 36.58053 for p = 1, 17.33036 for p = 2, and 13.48662 for p = 3. These results indicated that as the power increases, the RMSE decreases, implying a more accurate interpolation. A lower RMSE value suggests a closer fit of the interpolated values to the actual data, thus confirming that a higher power parameter enhances the precision of the IDW method. The validation of this trend is critical for ensuring the reliability of spatial predictions. The improvement

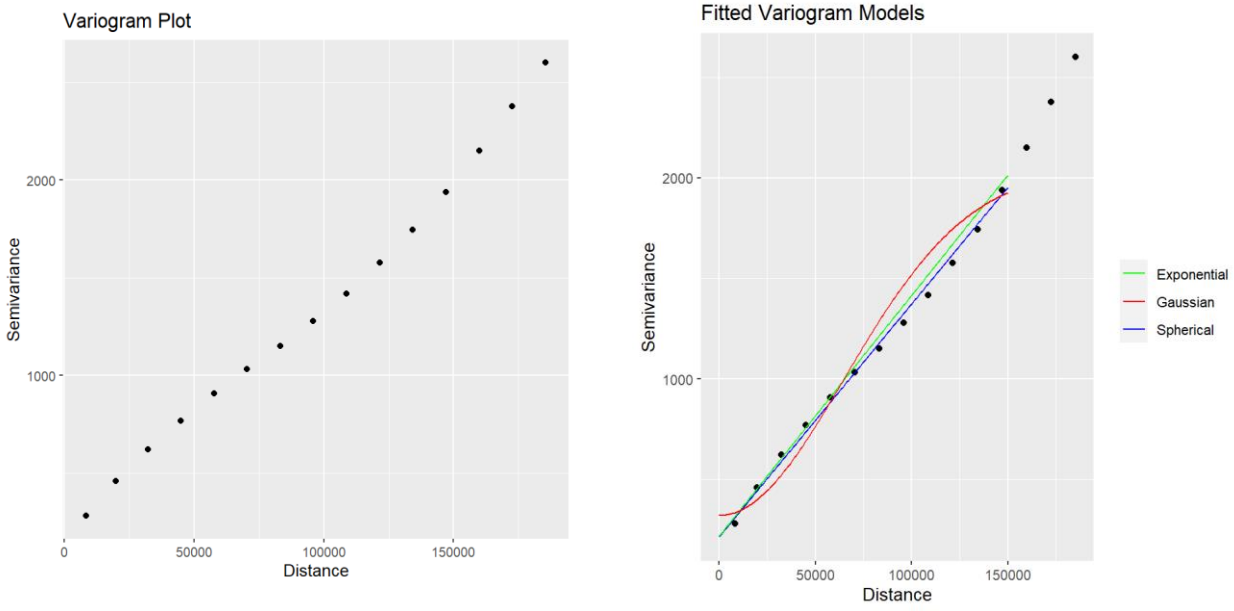
in RMSE with increasing power values provides a strong validation for the use of a higher power parameter in the IDW method, especially when high accuracy is imperative.

3. Apply the Universal Kriging Method

For the first step, the initial step in the analysis was to evaluate the experimental variograms to identify the spatial autocorrelation in the data. To fit the theoretical models, spherical, exponential, and Gaussian models were employed to the empirical variogram. Figure 03 showed that all three theoretical models captured the increasing trend of semivariance with distance, but with distinct curvatures. Upon comparison with the automatic model selection by the automap package there was a close match in the performance, with the exponential model fitting the data slightly better than the others. This was determined to be the most appropriate model for our dataset.

Figure 03

Variograms of All Three Theoretical Models

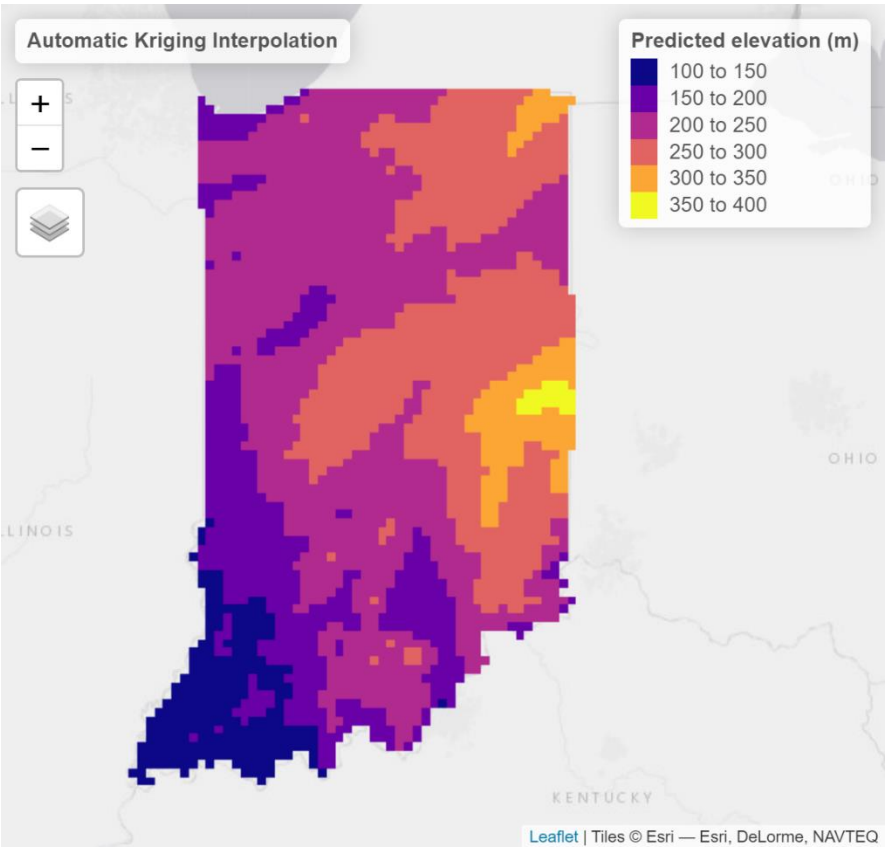


With the exponential variogram model selected, the next step is to perform universal kriging to create a 90m resolution Digital Elevation Model (DEM) of Indiana. The generated DEM (see Figure 04) revealed a detailed and smooth representation of the state's topography, with a clear delineation of elevation ranges. The map's color gradient, from deep purple to bright yellow, corresponds to elevation levels ranging from 100 to 400 meters. This visualization suggests a topographical gradient, with lower elevations in the southwest gradually rising towards higher elevations in the northeast. The average RMSE across all folds can be considered as an overall measure of the model's accuracy. In

this case, the average RMSE is approximately 13.602 (In reference), which can be interpreted as the model having an average error of around 13.602 meters in the predicted elevations. This level of accuracy is generally acceptable.

Figure 04

Auto Kriging



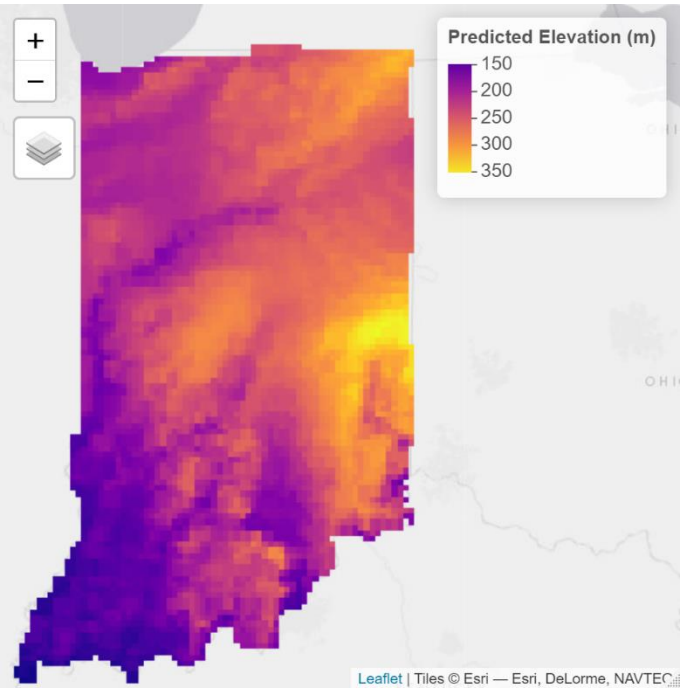
4. Choose a different interpolation method of your own choice.

The KNN method relies on the premise that locations with similar attributes are often in close proximity. Therefore, it predicts unknown values based on a weighted average of the nearest known data points. The KNN interpolation map (Figure 05) presents a vivid gradient of predicted elevations, ranging from 150 to 350 meters. The cross-validation process provided the average RMSE values of around 14.12875. Compared to IDW and kriging, KNN offers a distinct approach by considering only the most similar points rather than a smooth function across all points. It can be particularly effective in regions where elevation changes are abrupt or non-linear. However, the slightly higher RMSE

values in comparison to the kriging results suggest that while KNN provides a good general estimate, it may not capture all the subtle variations in elevation as effectively as kriging. (RMSE in appendix)

Figure 05

KNN Method

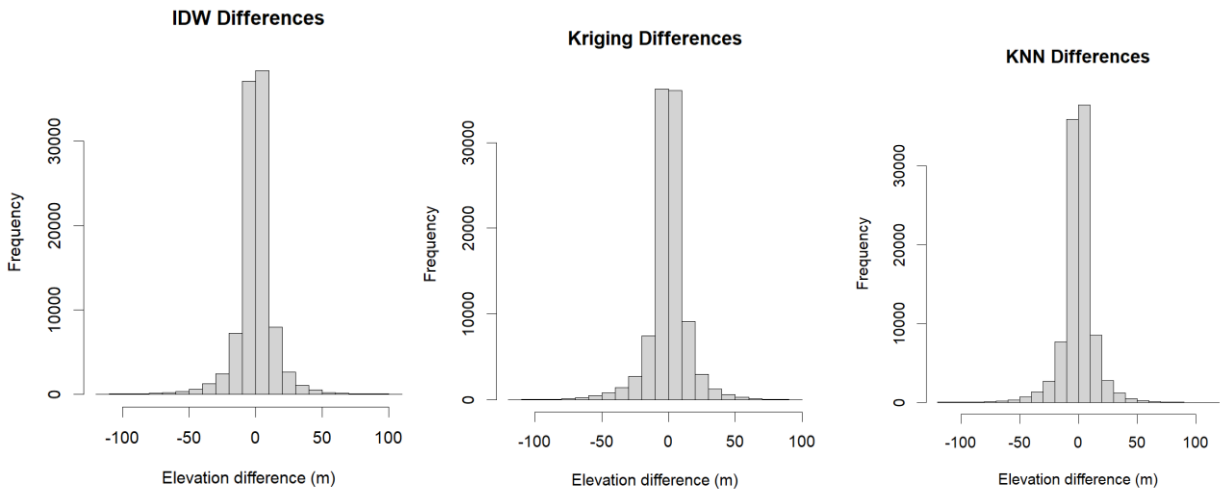


5. External Evaluation

The quality of the generated DEMs was evaluated against a ground truth DEM with a resolution of 30m. Each method's performance was assessed by examining the differences in elevation between the generated DEM and the ground truth. The Histograms below (Figure 06) depicting the frequency of elevation differences for IDW, kriging, and KNN show a central tendency around zero, indicating that on average, the predictions are close to the ground truth. However, the spread of the IDW differences is slightly tighter than that of kriging and KNN, suggesting a marginally better performance in capturing the elevation variations.

Figure 06

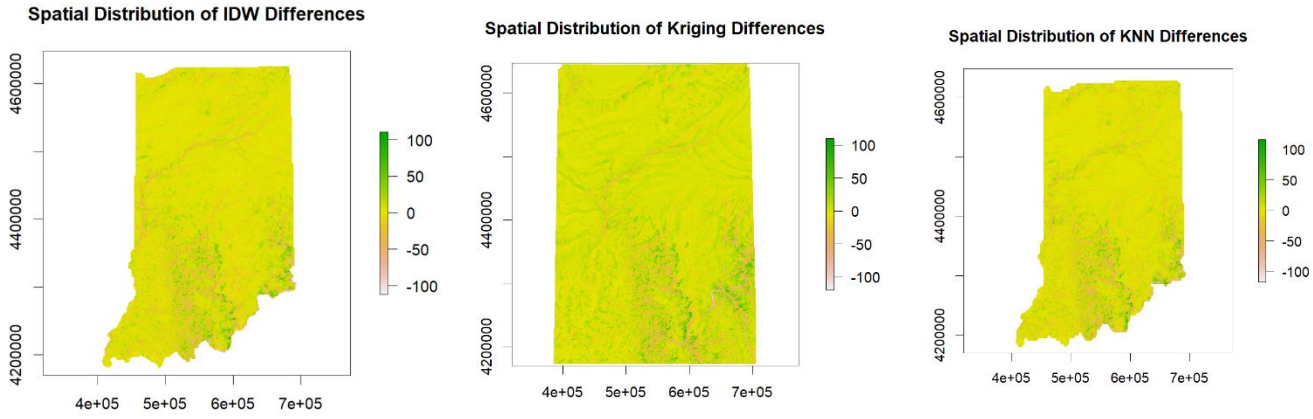
Frequency of the Elevation Differences



The spatial distribution maps (Figure 07) highlight where the discrepancies between the generated DEMs and the ground truth are most pronounced. For the IDW and KNN methods, the differences are relatively homogenous across the state, with few areas of significant differences. For the kriging spatial distribution map, the state of Indiana was not correctly represented as the corner of the state was excluded. There can be errors in the interpolation process.

Figure 07

Spatial Distribution of All Three Methods



The IDW method shows a strong performance, with a relatively symmetrical distribution of elevation differences, implying consistent accuracy across the state. Kriging, despite the potential error in spatial representation, generally follows a similar pattern but with a broader spread of differences,

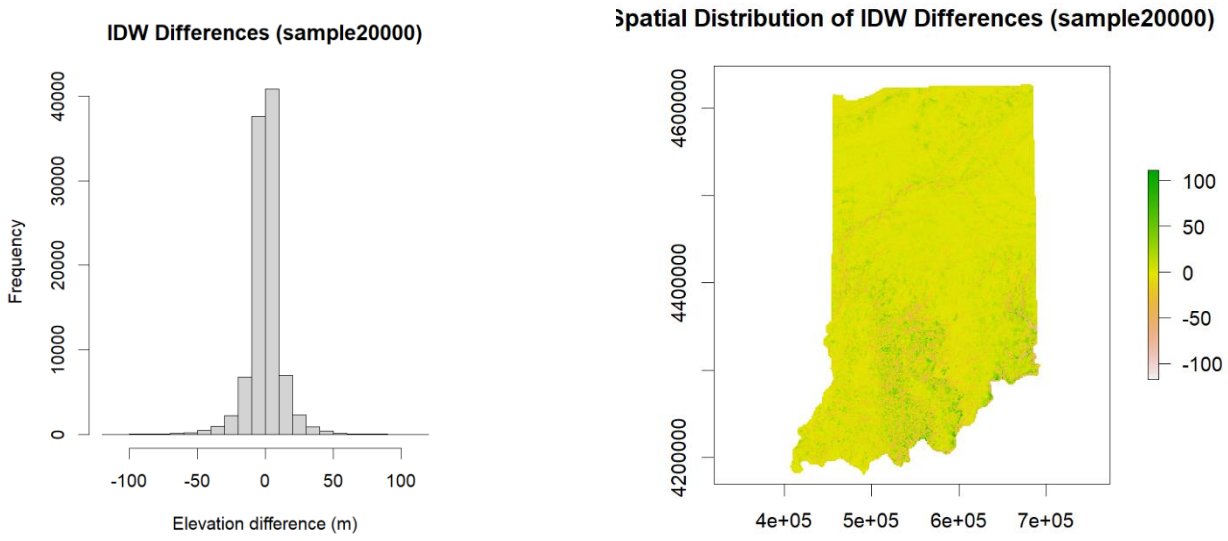
which might reflect a smoother, less detailed surface representation. The KNN method also performs comparably, with its histogram closely resembling that of IDW.

6. Bonus (20000 samples)

While changing the sample size to 20,000, the histogram (Figure 08 left) for the IDW differences with 20,000 samples shows a very sharp, narrow peak centered around zero, indicating that the majority of predicted elevations are very close to the actual values. The narrowness of the peak suggests a high level of accuracy in the interpolation process, with most differences falling within a small range of the true elevations. The symmetry around zero indicates that the IDW method does not have a systematic bias; it is equally likely to overestimate as it is to underestimate the true elevation. The spatial distribution map (Figure 08 right) shows the elevation differences across the state. The predominance of yellow, indicating minor differences close to zero, suggests that the IDW method has performed well across most of the area. There are some green fringes, representing areas where the differences are greater, but these appear to be relatively sparse and isolated.

Figure 08

Histogram and Spatial Distribution of IDW 20,000 Samples



By doubling the sample size from 10,000 to 20,000, the IDW interpolation has potentially become more accurate and consistent. A larger sample size generally provides a denser data network, which can help to reduce the uncertainty in the interpolation and provide a more representative model of the actual terrain.

7. Summary/Conclusion/Concluding Remarks

The evaluation in this project revealed that the IDW method, particularly with an increased sample size of 20,000 points, yielded the most accurate DEM, as indicated by the narrow distribution of elevation differences centered around zero. The spatial distribution of these differences was predominantly minor across the state, demonstrating the method's effectiveness in predicting true elevations.

Acknowledgement

References

IDW RMSE

```
> results <- idw_CV(p = 1, folds = folds, data_sp = data_sp)
[inverse distance weighted interpolation]
[inverse distance weighted interpolation]
[inverse distance weighted interpolation]
[inverse distance weighted interpolation]
[inverse distance weighted interpolation]
> mean(results$idw_RMSE)
[1] 36.58053
> results <- idw_CV(p = 2, folds = folds, data_sp = data_sp)
[inverse distance weighted interpolation]
[inverse distance weighted interpolation]
[inverse distance weighted interpolation]
[inverse distance weighted interpolation]
[inverse distance weighted interpolation]
> mean(results$idw_RMSE)
[1] 17.33036
> results <- idw_CV(p = 3, folds = folds, data_sp = data_sp)
[inverse distance weighted interpolation]
[inverse distance weighted interpolation]
[inverse distance weighted interpolation]
[inverse distance weighted interpolation]
[inverse distance weighted interpolation]
> mean(results$idw_RMSE)
[1] 13.48662
```

Kriging RMSE

```
> # Output the RMSE for each fold
> kriging_RMSE
[1] 13.62944 13.52761 14.14631 13.08216 13.62422
>
```

KNN RMSE

```
> knn_RMSE
[1] 14.15825 14.02500 14.47887 13.66544 14.31618
> data_df <- as.data.frame(data_sp)
```