AGH
University of Science and Technology in Krakow

DEPARTMENT OF MINING GEODESY AND ENVIRONMENTAL ENGINEERING

Theme: "DTM uncertainty analysis in Matlab/Octave"
Author: Tymoteusz Maj
Field of study: Remote Sensing and GIS

Kraków, 2024

Technical Report: Geospatial Data Comparison and Statistical Analysis

Introduction: This MATLAB script conducts a comprehensive analysis of two geospatial datasets, specifically comparing data derived from the Geoportal1 and Up42 sources. The aim of the analysis is to evaluate statistical properties of both datasets, identify potential differences, and assess their correlation. The script is structured to handle missing data, compute key statistical metrics, and visualize the relationship between the datasets through scatter plots and histograms.

1. Data Loading and Preprocessing: The code begins by importing the geospatial data from an external CSV file (data.csv). Each dataset, representing measurements from geoportal1 and up421, is assigned to corresponding variables for further analysis. The data in both geoportal1 and up421 columns are rounded to the nearest integer to normalize the measurements and reduce potential noise introduced by floating-point values. Additionally, the script addresses missing data (NaN values) in the up421 dataset by imputing these values with the column's mean. This ensures a complete dataset for statistical analysis without introducing significant bias from missing entries.

2. Statistical Metrics Calculation: For both the geoportal1_rounded and up421_filled datasets, the code computes four fundamental statistical metrics that are critical in geospatial data analysis:

- **Mean Absolute Error (MAE):** This measures the average magnitude of the errors in the datasets without considering their direction, providing an overall sense of the deviation from an ideal value.
- **Standard Deviation (SD):** The SD quantifies the amount of variation or dispersion in the dataset. A higher SD indicates that the values are more spread out, whereas a lower SD signifies that the values are more clustered around the mean.
- **Mean Error (ME):** This metric represents the average error, taking into account both positive and negative deviations from the mean, which allows for understanding whether the dataset tends to over- or underestimate values.
- **Root Mean Square Error (RMSE):** The RMSE calculates the square root of the average of squared differences between values, providing a more sensitive measure of larger errors in the dataset. It is often used to assess the overall accuracy of predictive models or measurements.

3. Dataset Comparison: The script compares the statistical measures calculated for both datasets, identifying the differences in MAE, SD, ME, and RMSE between the geoportal1_rounded and up421_filled datasets. This comparison serves to highlight the magnitude of variability between the two geospatial sources. By understanding these differences, the user can infer which dataset may be more consistent or reliable for geospatial analysis.

4. Validation: An important validation step is included in the code to assess whether the Standard Deviation (SD) is equal to the Root Mean Square Error (RMSE) for both datasets. This condition holds true if the Mean Error (ME) is zero, indicating that the error distribution is symmetric around the mean. The script checks this condition and reports the outcome, providing additional insight into the error structure of the datasets.

5. Explanation of DTM Uncertainty Analysis:

- **DTM Accuracy:**
    - The accuracy of a Digital Terrain Model (DTM) is typically assessed using statistical metrics like **Root Mean Square Error (RMSE)** and **Standard Deviation (SD)**. These metrics provide insights into the quality of elevation data in the DTM by comparing the predicted values to actual elevation measurements.
    - **RMSE** measures how far the values deviate from the ground truth on average, with smaller values indicating higher accuracy.
    - **SD** reflects the variability or spread in the dataset. A low SD suggests that the data is more consistent and reliable.
- **Assessment of DTM Accuracy in the Code:**
    - In the code, we calculate the RMSE and SD for the up421 dataset (assumed to be DTM data). These values are then compared to thresholds that determine the accuracy level:
        - If both RMSE and SD are less than 1, the DTM is considered highly accurate.
        - If the values fall between 1 and 2, the DTM accuracy is moderate.
        - Values greater than 2 indicate low DTM accuracy, signaling the need for improved data or model corrections.

6. Which Statistic is Reliable?

- **Root Mean Square Error (RMSE)** is often considered a reliable measure of accuracy because it gives more weight to larger errors, providing a robust assessment of how the model performs across different scales.
- **Standard Deviation (SD)** is also useful for assessing data consistency. When combined with RMSE, SD can provide a comprehensive understanding of the uncertainty in the model.

7. What is DTM Accuracy?

- **DTM Accuracy** refers to how well the Digital Terrain Model represents the true terrain. It is commonly measured using the **RMSE** between the model's predicted elevations and the true ground elevations. The lower the RMSE, the more accurate the DTM.
- In practice, a highly accurate DTM will have low RMSE and low SD, indicating that the model closely matches actual terrain elevations and does not exhibit significant variability.

8. Visualization:

8.1 Scatter Plot: The first visualization is a scatter plot that illustrates the relationship between the geoportal1_rounded and up421_filled datasets. Each point on the plot corresponds to a pair of values from the two datasets, providing a visual representation of their correlation. A tighter clustering of points along a diagonal line would suggest a strong linear relationship between the two datasets, whereas a more scattered distribution could indicate significant differences in their measurements.

8.2 Histograms: The script also generates two histograms, each displaying the frequency distribution of values in geoportal1_rounded and up421_filled. These histograms offer valuable insight into the distribution characteristics of the datasets, such as:

- The concentration of data points within specific value ranges.
- The spread and skewness of the data.
- Potential outliers or unusual data patterns.

By comparing the histograms, it is possible to observe how the datasets differ in terms of their internal distribution. For instance, one dataset may exhibit a more concentrated distribution, while the other may show greater variability across its value range.

9. General Analysis and Interpretation: The combined visualizations (scatter plot and histograms) provide a comprehensive view of the similarities and differences between the geoportal1_rounded and up421_filled datasets. The scatter plot offers a preliminary indication of correlation, suggesting whether the datasets align closely or diverge. Meanwhile, the histograms delve deeper into the data's frequency distribution, revealing how spread out or concentrated the values are.

Through this detailed analysis, the script enables the user to assess the quality and reliability of geospatial datasets and make informed decisions regarding their suitability for further use in geospatial modeling or other applications.

Conclusion: This MATLAB script serves as a robust tool for comparing geospatial datasets, providing key statistical metrics and visualizations to aid in understanding the underlying data structure. The statistical measures offer a quantitative comparison, while the visual representations allow for an intuitive interpretation of dataset relationships and variability.