Outlier Detection in Election Data Using Geospatial Analysis: Case Study: Ensuring Election Integrity (Kwara State) presented by Winner Obayomi

Introduction

This project was focused on detecting potential voting irregularities using geospatial and statistical methods. The objective was to identify polling units (PUs) in Kwara State whose voting results significantly deviated from their neighboring units, potentially signaling anomalies or irregularities in the election process. By combining location-based proximity and statistical outlier detection techniques, this study provides an evidence-based framework to improve election transparency and data integrity.

Methodology

1. Data Preparation

The dataset used for this analysis was obtained from the {Kwara_crosschecked2.csv} file, which contained detailed election results at the polling unit level. The data included:

- Polling Unit (PU) names and codes
- Votes obtained by various political parties
- Administrative information (Wards, LGAs, etc.)

Since the raw file did not include geographical coordinates, I added latitude and longitude values using the Geocode by Awesome Table add-on in Google Sheets. Each polling unit's address or name was passed to the geocoder, and the resulting coordinates were exported back into the dataset.
This provided the necessary spatial component required for the geospatial analysis.

2. Neighbor Identification

To identify each polling unit's neighbors, I used geospatial proximity based on their latitude and longitude coordinates. A radius of 1 kilometer was defined, meaning all polling units located within 1 km of a target unit were considered its neighbors. This was implemented in Python (Jupyter notebook) using the Haversine formula, which accurately measures the distance between two geographic coordinates on the earth's surface. This allowed each polling unit to have a list of nearby units with which its vote patterns could be compared.

3. Outlier Score Calculation

Once neighbors were identified, I computed outlier scores for each party within every polling unit.
For each PU:

- The mean and standard deviation of votes received by a party in neighboring PUs were calculated.
- A Z-score was then computed using the formula:

$$Z = \frac{(PU\_Votes - Neighbor\_Mean)}{Neighbor\_StdDev}$$

- The absolute value of the Z-score represented how much that polling unit's votes deviated from the normal pattern of nearby units.
- A higher Z-score indicated a stronger deviation meaning that polling unit's result was potentially irregular or influenced by external factors.

Each PU was then assigned a maximum absolute Z-score across all parties (max_abs_z_1km), representing the most extreme deviation found within that unit.

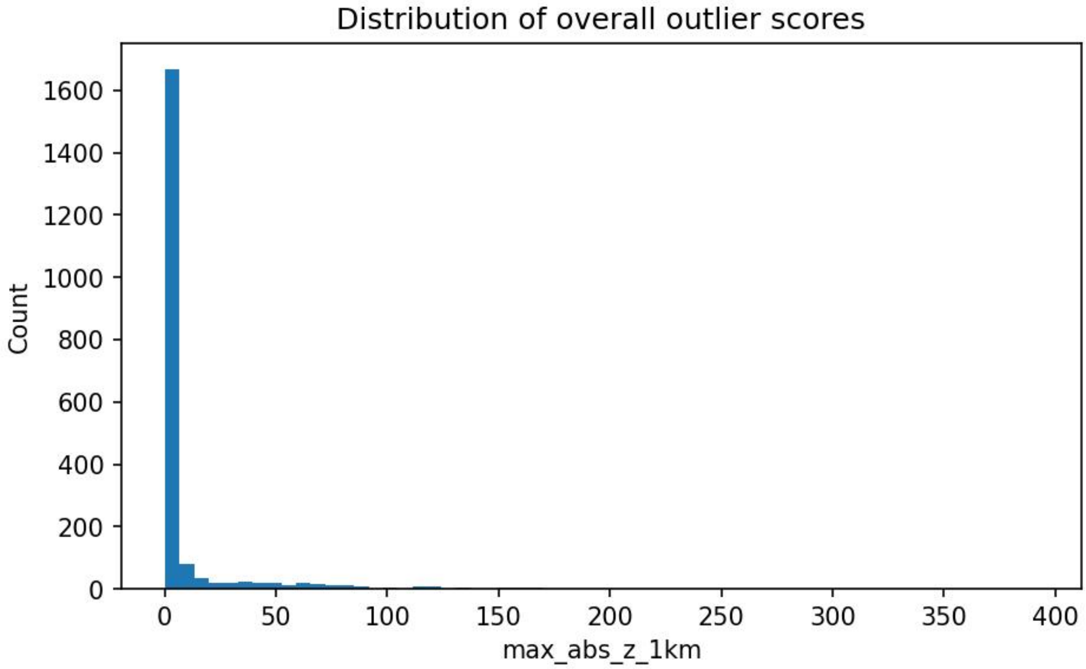
4. Tools and Environment

- Python (Jupyter Notebook): for data analysis, computation, and visualization
- Pandas: for data cleaning and manipulation
- NumPy: for numerical computation
- Matplotlib & Seaborn: for generating visualizations
- Geocode by Awesome Table (Google Sheets): for retrieving longitude and latitude
- Excel: for organizing and sorting outlier scores by party

Visualization and Interpretation

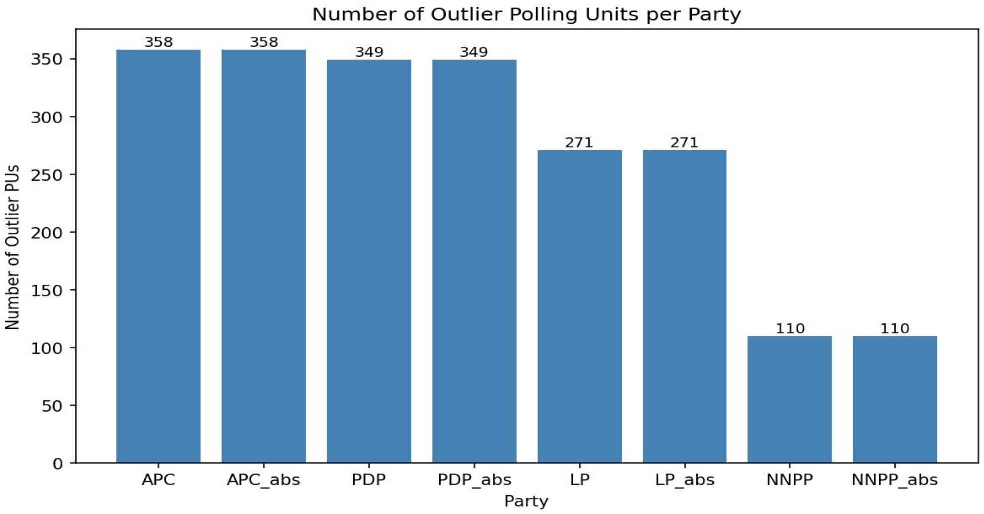
1. *Distribution of Outlier Scores*

This histogram shows how the overall outlier scores are distributed across all polling units. Most of the polling units have very low outlier scores, which suggests that their results fall within normal or expected ranges. However, a small number of units have much higher scores, visible on the far right side of the chart. These extreme values indicate polling units where the results deviate sharply from surrounding patterns, making them potential outliers. Overall, the graph highlights that irregularities are rare but significant in magnitude where they occur.



2. Number of Outlier Polling Units per Party

A bar chart was created showing how many polling units were outliers for each political party based on a Z-score threshold (e.g., |Z| > 2.5).



The chart reveals that APC had the highest number of outlier polling units than others. This could indicate regional strongholds where the party dominated, or possible inconsistencies that differ significantly from nearby results.

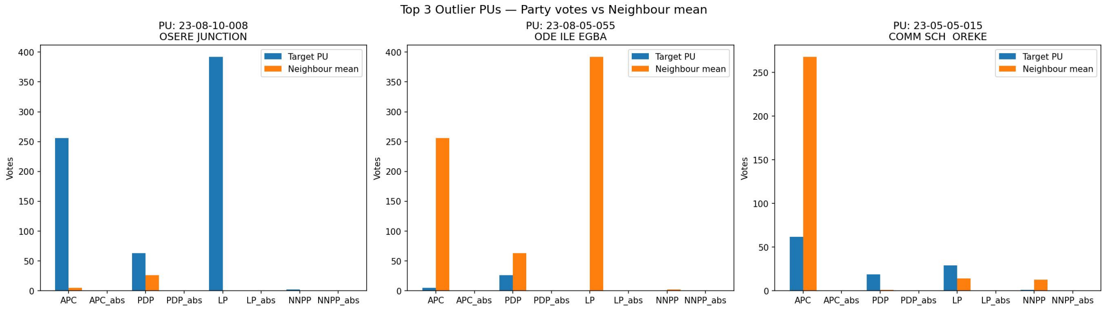
3. The Top 3 Outliers

The chart below illustrates the three polling units (PUs) that displayed the most significant deviations from their neighbouring PUs within a one-kilometre radius. Each subplot compares the vote distribution among major political parties at the target PU (in blue) with the average votes of its neighbouring PUs (in orange). The main reason for the comparison is to highlight unusual voting patterns that differ from surrounding polling units.

At PU: **23-08-10-008 (OSERE JUNCTION)**, the target PU recorded an high number of votes for the Labour Party (LP), reaching almost 400 votes, while the neighbouring average for LP was very low. The APC and PDP votes at this PU were also higher than the neighbouring averages, though the gap was most pronounced for LP. This indicates that OSERE JUNCTION is a strong positive outlier, showing high support for LP compared to nearby polling units.

At **PU: 23-08-05-055 (ODE ILE EGBA)**, had the opposite pattern. Here, the target PU recorded very low votes across all major parties, while its neighbouring polling units showed higher averages, around 250 votes for APC and nearly 400 for LP. This wide disparity suggests that ODE ILE EGBA is a negative outlier, where the votes are drastically below the expected range based on surrounding results.

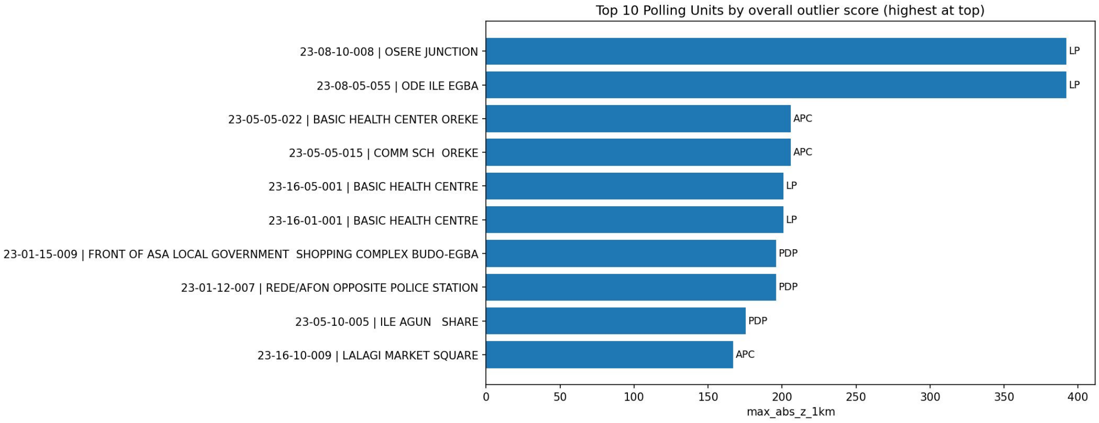
At **PU: 23-05-05-015 (COMM SCH OREKE),** shows a similar but less extreme trend. The target PU has moderately lower votes for most parties compared to its neighbours. For example, the neighbouring average for APC is about 270 votes, while the target PU has only around 60. The differences for LP and PDP are smaller but still evident, suggesting that this PU underperformed relative to its surrounding area, particularly for APC.



Overall, the three polling units show voting results that are far from what is typical in their areas. While OSERE JUNCTION recorded unusually high votes, the other two showed significantly lower figures. These patterns could be due to reporting errors, unique local factors, or possible irregularities. The analysis helps highlight polling units that deserve further review because of their unusual vote distributions.

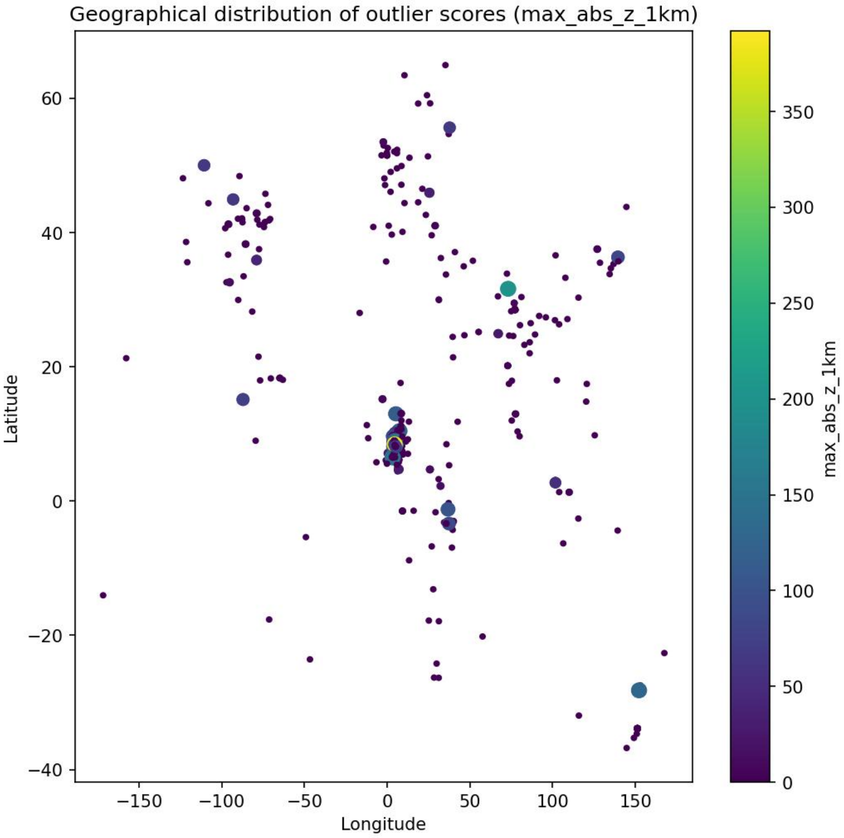
4. The Top 10 polling units by overall outliers score

The chart showed the strongest outlier scores in the election, meaning these locations had unusually high or irregular voting patterns compared to others nearby. Each bar represents a polling unit, with the length showing how extreme its outlier score is. The labels on the right indicate which political party was responsible for the largest deviation at that PU. From the chart, the most of the highest outlier scores were associated with the Labour Party (LP), followed by the APC and PDP in some cases. This visualization helps identify polling units that may need further review or verification due to their unusually high results.



5. Geographical distribution of outliers scores

This scatter plot shows how outlier scores are distributed geographically across different polling units. Each point represents a polling unit, with its position determined by latitude and longitude. The size and color of each point reflect how extreme its outlier score is, larger and brighter points indicate stronger deviations from the norm. From the map, while most polling units have relatively low outlier scores, a few clusters stand out with very high values. These concentrated areas of unusually high scores may suggest irregularities or anomalies that deserve closer investigation or verification.



Conclusion

This analysis has demonstrated how geospatial and statistical methods can be used to detect potential irregularities in election data. By comparing each polling unit's results with those of its neighbouring units within a one-kilometre radius, it was possible to identify areas where voting patterns deviated sharply from the norm. Most polling units in Kwara State showed consistent and reasonable results, suggesting that the election data is generally reliable.

However, a few units displayed strong deviations, particularly OSERE JUNCTION, ODE ILE EGBA, and COMM SCH OREKE, where vote distributions differed significantly from nearby polling units. These outliers may not necessarily indicate manipulation, but they highlight locations that deserve closer review or verification to ensure data accuracy.

Across the political parties, the Labour Party (LP) and the All Progressives Congress (APC) recorded the highest number of outlier polling units, showing a mix of both high and low vote counts. This pattern may reflect localized political dynamics, voter concentration, or possible inconsistencies in result reporting.

Overall, this project reinforces the importance of data analytics in promoting transparency and accountability in electoral processes. The integration of geospatial tools and outlier detection techniques provides election observers and policymakers with a practical framework for identifying and investigating anomalies. By applying similar analyses across other states, election stakeholders can strengthen public trust and improve the credibility of future elections.