Advanced Geospatial Outlier Detection Report for Oyo State Polling Data

This report presents an in-depth analysis of polling unit data in Oyo State. By integrating advanced geospatial clustering, spatial statistical methods, machine learning validation, and demographic comparisons, this study identifies and justifies five polling units that exhibit significant deviations from expected voting patterns. The report is structured in clear sections to ensure accessibility for both technical and non-technical audiences.

1. Introduction

Background:
Elections are the cornerstone of democratic governance, yet allegations of irregularities can severely undermine public trust. In response to concerns raised by the Independent National Electoral Commission (INEC), this study focuses on detecting anomalies in Oyo State polling data. The objective is to flag polling units whose voting outcomes deviate markedly from both neighboring units and historical voting trends.

Problem Statement:
How can we systematically detect and validate polling units whose vote counts and patterns differ significantly from expected norms, both locally and historically?

2. Aims and Objectives

Primary Aim:
To build a robust, multi-faceted framework for anomaly detection in electoral data that supports enhanced election integrity efforts.

Key Objectives:

- **Dataset Preparation:**
    - Acquire and clean polling unit data for Oyo State.
    - Obtain accurate latitude and longitude coordinates via the Google Maps API, ensuring each polling unit is precisely geocoded.
- **Advanced Neighbor Identification:**

    - Apply HDBSCAN clustering to group polling units by geographic proximity.
    - Conduct sensitivity analysis to fine-tune parameters—specifically, setting the minimum cluster size (mcs) to **4** and minimum samples (ms) to **2**—to achieve an optimal balance between meaningful clusters and manageable noise.

- **Sophisticated Outlier Score Calculation:**
    - Use spatial statistical methods (Local Moran's I and Getis-Ord Gi) to capture local vote count deviations.
    - Validate and refine anomalies using the Isolation Forest machine learning algorithm.
    - Generate composite scores and vote proportion z-scores to summarize anomaly signals.

- **Comparative and Demographic Analysis:**
    - Compare current vote distributions with aggregated historical data (e.g., from 2019).
    - Integrate socio-economic and demographic indicators (e.g., unemployment, literacy, population) to provide context.

- **Interactive Visualization and Reporting:**
    - Develop interactive maps and dashboards for stakeholders, using tools such as Folium with MarkerCluster.
    - Present detailed visualizations—maps, charts, and comparative graphs—to communicate the findings clearly.

3. Methodology

3.1 Data Preparation

- **Data Import and Cleaning:**
  The dataset, including vote counts for major parties (APC, LP, PDP, NNPP) and administrative data, is imported using Pandas. Rigorous cleaning is applied to remove inconsistencies and fill missing values.

- **Geocoding:**
  Polling unit addresses are transformed into geographic coordinates using the

Google Maps API. This step ensures that each unit is accurately located, forming the basis for reliable spatial analysis.

3.2 Geospatial Clustering

- **Concept Overview:**
  Geospatial clustering groups polling units based on proximity. This helps reveal natural clusters where voting behavior is similar and highlights isolated units that might be anomalous.

- **HDBSCAN Clustering:**
  HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) is employed for its key advantages:

    - **Adaptability:** Automatically adjusts to varying densities without a fixed distance threshold.

    - **Noise Identification:** Effectively distinguishes between meaningful clusters and noise (isolated points).

    - **Robustness:** Performs well in both urban and rural settings.

- **Parameter Updates:**
  Sensitivity analysis indicated that setting the minimum cluster size (mcs) to **4** and min_samples (ms) to **2** produces optimal clustering outcomes. These settings result in a balanced distribution of clusters and noise, capturing natural groupings while reducing excessive fragmentation.

- **Implementation Details:**

    - Latitude and longitude data (obtained from the Google Maps API) are converted into radians for the haversine metric.

    - An interactive map is generated using Folium with MarkerCluster, where each cluster is color-coded and noise points are marked in gray.

3.3 Outlier Detection Techniques

A multi-method approach is used to detect anomalies:

3.3.1 Spatial Statistical Methods

- **Local Moran's I:**
  Measures spatial autocorrelation by comparing each polling unit's vote count with its neighbors. High values indicate significant local deviations.

- **Getis-Ord Gi:**
  Calculates local z-scores to determine if a unit's vote count is unusually high or low relative to nearby units.

3.3.2 Machine Learning – Isolation Forest

- **Overview:**
  Isolation Forest partitions the data to isolate anomalies. Units that are isolated using fewer splits are flagged as anomalies.

- **Application:**
  The algorithm is applied to the multidimensional vote count data. Polling units flagged as anomalous (Iso_Label = -1, then assigned Iso_Score = 1) are noted.

3.3.3 Composite and Z-Score Calculations

- **Normalization and Composite Scores:**
  Spatial statistic values are normalized using Min-Max scaling. For each party, the normalized scores (from Local Moran's I and Getis-Ord Gi) are summed to produce a composite score. A Global Composite Score is then computed by averaging these values and incorporating the Isolation Forest score.

- **Vote Proportion Z-Scores:**
  These scores compare current vote proportions with historical state-level data (e.g., from 2019). Extreme z-scores highlight significant deviations.

- **Accredited Ratio:**
  The ratio of accredited voters to total votes is computed. Ratios much greater than 1 may indicate ballot stuffing or other irregularities.

3.4 Comparative and Demographic Analysis

- **Historical Data Integration:**
  Historical election results (from 2019, 2015, and 2011) serve as benchmarks to gauge expected vote distributions.

- **Socio-Economic Indicators:**
  Indicators such as unemployment, literacy, and population are integrated to provide

context. This integration helps distinguish between genuine local shifts and suspicious anomalies.

- **Visualization:**
  Comparative charts and maps illustrate differences between current and historical data, enabling visual detection of anomalies.

4. Identification of Top 5 Outlier Polling Units

The analysis flagged five polling units as the most anomalous based on a combination of spatial, statistical, and machine learning indicators. Each unit's details are presented below.

4.1 D.C. SCHOOL, AYEDE II (Index 2631)

- **Cluster:**
  Cluster 155 (green) – Not isolated but markedly anomalous within its group.

- **Spatial Statistics:**

  - Local Moran's I: APC = 0.457665, PDP = 0.770205, NNPP = 5.297692
    These high values indicate strong local deviations.

- **Isolation Forest:**

  - Iso_Label = -1, Iso_Score = 1, confirming it as an anomaly.

- **Composite Scores:**

  - APC_Composite = 0.65643, LP_Composite = 0.662185, PDP_Composite = 0.815315, NNPP_Composite = 1.09627

  - Global_Composite_Score = 1.80755, one of the highest in the dataset.

- **Vote Proportion Z-Scores (2019 Benchmark):**

  - LP_z_score = 132.33, PDP_z_score = -9.50, indicating a drastic deviation for LP.

- **Accredited Ratio:**

  - 2.635294 (224 Total Votes vs. 85 Accredited Voters) – a major red flag.

- **Justification:**
  The extreme LP_z_score and an accredited ratio above 2.6 suggest a vote count far exceeding voter eligibility, strongly indicating potential ballot stuffing or severe data errors. Both spatial and machine learning methods confirm this unit as a prime outlier.

4.2 BLIND CENTRE (Index 2480)

- **Cluster:**
  Cluster 159 (lightblue) – Part of a cluster yet distinctly anomalous.

- **Spatial Statistics:**

    - Local Moran's I: APC = -0.096499, LP = 20.909268
      The extraordinarily high LP value highlights severe deviation.

- **Isolation Forest:**

    - Iso_Label = -1, Iso_Score = 1, reinforcing its anomaly.

- **Composite Scores:**

    - APC_Composite = 0.551315, LP_Composite = 1.694486

    - Global_Composite_Score = 1.740632.

- **Vote Proportion Z-Scores:**

    - LP_z_score = 576.27, APC_z_score = -13.01, PDP_z_score = -12.85 – indicating massive LP deviation and underrepresentation for other parties.

- **Accredited Ratio:**

    - 0.893372 (310 Total Votes vs. 347 Accredited Voters) – while normal in isolation, the skewed vote distribution is concerning.

- **Justification:**
  The extreme LP_z_score and the distorted local vote pattern indicate a manipulated vote distribution, despite a normal accredited ratio. This compelling evidence points to targeted irregularities in this polling unit.

4.3 L.A. SCHOOL, KAJOLA (Index 2259)

- **Cluster:**
  Cluster 68 (darkred) – Part of a larger grouping yet displaying pronounced deviation.

- **Spatial Statistics:**

    - Local Moran's I: PDP = 16.712605, LP_Getis_Ord_Gi = -0.486429 – a high PDP value suggests significant local deviation.

- **Isolation Forest:**

    - Iso_Label = -1, Iso_Score = 1.

- **Composite Scores:**

    - APC_Composite = 0.60927, LP_Composite = 0.168809, PDP_Composite = 1.814055, NNPP_Composite = 0.34142

    - Global_Composite_Score = 1.733388.

- **Vote Proportion Z-Scores:**

    - APC_z_score = -7.38, PDP_z_score = 6.46, LP_z_score = 21.09.

- **Accredited Ratio:**

    - 0.930921 (283 Total Votes vs. 304 Accredited Voters) – close to 1, yet the vote distribution is abnormal.

- **Justification:**
  The unusual PDP and LP vote proportions—despite a near-normal accredited ratio—demonstrate a significant deviation from historical norms. This unit's behavior is strongly flagged as anomalous.

4.4 AKINTEKUN COMPOUND (Index 2253)

- **Cluster:**
  Cluster 68 (darkred) – Shares the same cluster as L.A. SCHOOL, KAJOLA, and displays similar spatial patterns.

- **Spatial Statistics:**

    - Similar Local Moran's I values as L.A. SCHOOL, indicating comparable anomalies.

- **Isolation Forest:**
    - Iso_Label = -1, Iso_Score = 1.
- **Composite Scores:**
    - Global_Composite_Score = 1.733388, identical to L.A. SCHOOL.
- **Vote Proportion Z-Scores:**
    - Identical to L.A. SCHOOL: APC_z_score = -7.38, LP_z_score = 21.09, PDP_z_score = 6.46.
- **Accredited Ratio:**
    - 1.246696 (283 Total Votes vs. 227 Accredited Voters) – this ratio exceeds 1, indicating votes cast exceed the number of accredited voters.
- **Justification:**
  The identical spatial and statistical signals to L.A. SCHOOL, combined with a critical red flag in the accredited ratio, strongly indicate potential manipulation. The excess in recorded votes relative to eligible voters further substantiates its outlier status.

4.5 LATE OBA SABO PALACE I (Index 621)

- **Cluster:**
  Cluster 400 (orange) – Part of a distinct cluster but with pronounced irregularities.
- **Spatial Statistics:**
    - Local Moran's I: NNPP_Local_Moran_I = 17.9824 – an extremely high value that stands out.
- **Isolation Forest:**
    - Iso_Label = -1, Iso_Score = 1.
- **Composite Scores:**
    - Global_Composite_Score = 1.722513.
- **Vote Proportion Z-Scores:**

    - NNPP_z_score = 91.75 (dominant), LP_z_score = 12.33, PDP_z_score = -5.39 – showing a drastic NNPP anomaly.

- **Accredited Ratio:**

    - 0.852941 (votes are slightly below the accredited count, which is normal in isolation) – however, the anomaly is driven by the abnormal NNPP share.

- **Justification:**
  The exceptionally high NNPP_z_score indicates that the NNPP vote proportion is abnormally high relative to historical benchmarks. This strong single-party anomaly—despite a normal overall accredited ratio—implies that this unit deviates drastically from expected voting patterns. The combined evidence confirms it as a significant outlier.

5. Hypotheses on Potential Anomalies

The multi-faceted evidence leads to several hypotheses regarding the causes of these anomalies:

- **Data Entry or Transcription Errors:**
  Manual errors during data recording can lead to discrepancies where vote counts do not match accredited figures, artificially inflating anomaly scores.

- **Ballot Stuffing or Fraudulent Practices:**
  Polling units with accredited ratios well above 1 (e.g., D.C. SCHOOL, AYEDE II and AKINTEKUN COMPOUND) indicate that more votes were recorded than eligible voters, strongly suggesting ballot stuffing or other fraudulent activities.

- **Localized Socio-Economic Influences:**
  Although local socio-economic conditions (changes in unemployment, literacy, or population dynamics) can influence voting patterns, the extreme deviations observed—particularly in party-specific z-scores—are unlikely to be solely explained by these factors.

- **Systematic Manipulation:**
  Consistent anomalies in neighboring units (as seen in cluster 68 for L.A. SCHOOL, KAJOLA and AKINTEKUN COMPOUND) could indicate coordinated efforts to manipulate vote counts in a specific area.

- **Technical Failures in Vote Tallying:**
  Malfunctions in vote-counting equipment or errors in data transmission may result in incorrect totals, contributing to abnormal ratios and composite scores.

Limitations in Hypothesis Testing

- **Granularity Mismatch:**
  Historical data are aggregated at the state level, and polling unit–level historical data are not always available. This may mask local nuances.

- **Assumption of Stability:**
  The analysis assumes that the 2019 historical voting proportions remain relevant for 2023. Shifts in demographics or political dynamics may naturally alter expected outcomes, potentially leading to false-positive anomaly signals.

- **Inexact Variance Estimation:**
  The standard error for z-scores is derived from global proportions, which might not capture the variance inherent at the polling unit level.

- **Contextual Changes Over Time:**
  Changes in political sentiment or socio-economic conditions between 2019 and the current election might result in genuine shifts in vote patterns that the current model could mistakenly classify as anomalies.

6. Recommendations for Election Authorities

Based on these detailed insights, the following robust recommendations are proposed:

- **Conduct Comprehensive On-Site Audits:**

    - **Target Extreme Outliers:**
      Prioritize polling units with exceptionally high Global Composite Scores, extreme vote proportion z-scores, and concerning accredited ratios (notably D.C. SCHOOL, AYEDE II and AKINTEKUN COMPOUND).

    - **Forensic Examination:**
      Engage specialized audit teams to physically verify ballot counts, cross-check voter registers, and review polling procedures. Detailed forensic audits should be implemented in areas flagged as outliers.

- **Enhance Real-Time Data Validation:**

    - **Automated Cross-Checks:**
      Implement systems that automatically flag discrepancies between reported vote counts and accredited voter numbers in real time.

    - **Immediate Alerts:**
      Establish protocols that trigger prompt investigations when pre-set

thresholds (e.g., accredited ratios above 1 or extreme z-scores) are breached.

- **Strengthen Training and Standardization:**
    - **Staff Training:**
      Regularly train polling personnel on accurate data entry, effective use of vote-counting equipment, and error-checking procedures to minimize manual errors.
    - **Standardized Procedures:**
      Establish uniform protocols across all polling units to ensure consistency in data reporting and processing.
- **Increase Transparency Through Public Dashboards:**
    - **Interactive Visualizations:**
      Develop public dashboards using tools such as Folium, Tableau, or Power BI to display real-time electoral data and clearly mark anomalous polling units.
    - **Independent Oversight:**
      Enable independent observers and the media to access these dashboards, fostering community and stakeholder oversight.
- **Integrate Updated Socio-Economic Data:**
    - **Contextual Analysis:**
      Collaborate with local agencies to incorporate recent demographic and socio-economic data. This integration will help distinguish between genuine local shifts and suspicious anomalies.
    - **Regular Updates:**
      Periodically refresh historical benchmarks and socio-economic indicators to reflect current conditions, ensuring ongoing relevance of the analysis.
- **Establish a Dedicated Electoral Integrity Task Force:**
    - **Continuous Monitoring:**
      Form a task force dedicated to monitoring polling data anomalies, coordinating follow-up investigations, and ensuring accountability.
    - **Structured Reporting:**
      Implement a structured reporting mechanism that communicates findings directly to oversight bodies and the public.

- **Review and Upgrade Voting Infrastructure:**
    - **Technical Audits:**
      Conduct regular audits of vote-counting equipment and digital systems to ensure operational accuracy.
    - **Redundancy Protocols:**
      Introduce backup systems and cross-verification methods to minimize the risk of technical failures affecting vote counts.

7. Conclusion

This comprehensive report integrates advanced geospatial clustering (with updated HDBSCAN parameters: mcs = 4 and ms = 2), robust spatial statistics, machine learning validation, and comparative historical analysis to identify five outlier polling units in Oyo State. Each flagged unit exhibits a unique combination of spatial deviations, extreme vote proportion z-scores, and, in some cases, alarming accredited voter ratios. The multi-pronged evidence from independent analytical methods robustly indicates that these units deviate significantly from both local and historical voting patterns.

While the anomalies could stem from data entry errors, ballot stuffing, systematic manipulation, or technical failures, the inherent limitations of the available historical data and variance estimation require cautious interpretation. Nonetheless, the strong, insight-driven recommendations provided herein offer a clear roadmap for targeted audits, enhanced data validation, and overall improvements in electoral infrastructure—crucial steps for safeguarding the integrity of the electoral process and restoring public confidence.