Bihar Electoral Analysis - Additional Analyses Implementation Guide

Executive Summary

This guide details 15 categories of additional analyses that can be conducted using existing Bihar electoral datasets. All suggested analyses utilise available data and do not require external sources, predictive modelling, or text mining.

Available Data Sources

* **booth\_analysis**\_india\_nda.xlsx – Contains 232,432 booth-level records including electoral outcomes.
* **surname\_counts.csv** – Includes 68,719 surname records with counts.
* **guardian\_surname\_c**ounts.csv – Provides 77,214 guardian surname records.
* **AlldeletionsBoothsform20.csv** – Lists detailed deletion data by booth as well as alliance outcomes.
* **Bihar\_Surnames\_Cas**tes\_preview.csv – Reference for caste-surname mapping.
* **Existing analysis documents** – Pre-calculated statistics are available in Word reports.

Priority 1: High-Impact Analyses (Implement First)

1. Surname Vulnerability Analysis

**Data Required**: surname\_counts.csv, guardian\_surname\_counts.csv, AlldeletionsBoothsform20.csv

**Implementation**:

```

* Calculate deletion rate per surname (deleted count / total count)
* Develop Surname Vulnerability Index (SVI) = (deletion\_rate - mean) / std\_dev
* Identify top 100 surnames most affected
* Compute surname diversity metrics (Simpson's Index) by constituency
* Assess concordance rates between guardian and voter surnames

```

**Visualisations**:

* Interactive searchable table with risk scores
* Heatmap of surname vulnerability by region
* Treemap of surname proportions

2. Spatial Clustering Analysis

**Data Required**: AlldeletionsBoothsform20.csv with booth coordinates

**Implementation**:

```

* Calculate Global Moran's I for spatial autocorrelation
* Apply Local Moran's I (LISA) for cluster detection
* Use Getis-Ord Gi\* for hot spot analysis
* Examine distance effects from district/AC headquarters
* Review constituency contiguity patterns

```

**Visualisations**:

* Cluster maps indicating hot and cold spots
* Moran's I scatter plot
* Distance decay curves

3. Electoral Impact Quantification

**Data Required**: AlldeletionsBoothsform20.csv

**Implementation**:

```

* Calculate margin erosion: (deletions × expected\_vote\_share) / total\_votes
* Measure changes in competitiveness: margin\_with\_deletions vs margin\_without
* Generate Alliance Vulnerability Score for each constituency
* Identify "flippable seats" based on deletion patterns

```

**Visualisations**:

* Scatter plot: Deletions versus Victory Margin
* Sankey diagram: Deletion bins to Electoral outcomes
* Bubble chart: X=deletions, Y=margin, Size=booth\_count

4. Composite Integrity Index

**Data Required**: All deletion and outcome data

**Implementation**:

```

Electoral Integrity Index (EII) = weighted average of:

* Deletion intensity (25%): deletions per 1000 voters
* Gender skew (25%): deviation from population baseline
* Caste skew (25%): deviation from census proportions
* Spatial clustering (25%): local Moran's I value

Score range: 0–100 (100 = highest integrity)

```

**Visualisations**:

* Traffic light dashboard
* Ranked bar chart of constituencies
* Spider/radar charts showing multi-dimensional profiles

Priority 2: Statistical Deep-Dives

5. Multi-level Hierarchical Analysis

**Implementation**:

```

Level 1: Booth dataset

Level 2: Assembly Constituency

Level 3: Alliance Outcome (INDIA/NDA/Other)

* Variance decomposition via ANOVA
* Random effects modelling
* Intraclass correlation coefficients
* Cross-level interactions

```

6. Advanced Cross-tabulations

**Implementation**:

```

Three-way tables:

* Religion × Deletion\_Bin × Winner
* Surname\_Type × AC\_Type × Outcome
* Gender × Caste × Deletion\_Intensity

Include:

* Chi-square tests
* Cramér's V for effect size
* Standardised residuals
* Odds ratios

```

7. Distribution Analysis Suite

**Implementation**:

```

For each key variable:

* Mean, Median, Mode, Variance
* Skewness, Kurtosis, Jarque-Bera test
* Percentiles (P5, P10, P25, P50, P75, P90, P95)
* Gini coefficient, Lorenz curve
* Outlier detection: Z-scores, Modified Z-scores, IQR method

```

Priority 3: Advanced Visualisations

8. Flow Visualisations

**Sankey Diagrams**:

* Deletion Bins to Electoral Outcomes
* Religion Composition to Alliance Winners
* Surname Types to Deletion Categories

**Alluvial Diagrams**:

* Multi-level flows illustrating relationships

9. Distribution Visualisations

**Violin Plots**: Display distribution shapes by category

**Ridge Plots**: Overlapping distributions across categories

**Box-Whisker Plots**: Highlighting outliers by constituency

**Hexbin Plots**: For dense scatter plot data

10. Multi-dimensional Visualisations

**Spider/Radar Charts**: Compare 6–8 metrics per constituency

**Parallel Coordinates**: Detect multi-variable patterns with interactive filtering

**Bubble Charts**: Represent 3–4 dimensions using size and colour encoding

Priority 4: Booth-Level Analytics

11. Anomaly Detection

**Implementation**:

```

* Compute booth-level z-scores for all metrics
* Flag booths exceeding three standard deviations
* Employ Isolation Forest for multivariate outlier detection
* Construct an anomaly index combining multiple flags

```

12. Booth Clustering

**Implementation**:

```

K-means clustering based on:

* Deletion rate
* Gender ratio
* Religion composition
* Electoral outcome

Determine optimal k using elbow method

Profile each resulting cluster

```

Priority 5: Statistical Validation

13. Correlation Analysis

**Implementation**:

```

* Full Pearson correlation matrix
* Spearman rank correlations
* Partial correlations controlling for confounders
* Distance correlations for non-linear associations
* Correlation heatmaps with indicators of statistical significance

```

14. Effect Size Calculations

**Implementation**:

```

* Cohen's d for group differences
* Eta-squared for ANOVA
* Phi coefficient for 2×2 tables
* Confidence intervals via bootstrap (1,000 iterations)

```

15. Quality Metrics

**Implementation**:

```

* Data completeness score by constituency
* Logical consistency checks
* Reliability coefficients when applicable
* Sensitivity analysis for key assumptions

```

Implementation Roadmap

Phase 1 (Week 1–2)

* Surname Vulnerability Analysis
* Spatial Clustering Analysis
* Basic Composite Index

Phase 2 (Week 3–4)

* Electoral Impact Quantification
* Advanced Cross-tabulations
* Flow Visualisations

Phase 3 (Week 5–6)

* Multi-level Modelling
* Distribution Analysis
* Booth-level Analytics

Technical Requirements

Statistical Libraries Needed

* **Basic Stats**: mean, median, mode, standard deviation, percentiles
* **Spatial Stats**: Moran's I, Getis-Ord Gi\*
* **Clustering**: K-means, hierarchical clustering
* **Visualisation**: D3.js for Sankey diagrams, Chart.js for standard charts
* **Statistical Tests**: Chi-square, t-tests, ANOVA

Data Preprocessing Steps

* Standardise surname spellings and address variations.
* Aggregate data at the constituency level.
* Calculate per-capita metrics (per 1,000 voters).
* Generate binary flags for relevant categories.
* Bin continuous variables into quintile or decile groups.

Expected Outputs

Dashboard Enhancements

* **New Tabs**:
* Surname Analysis
* Spatial Patterns
* Integrity Scoring
* Deep Statistics
* **Interactive Features**:
* Surname lookup tool
* Constituency profiling
* Custom cross-tab builder
* Export options for all analyses
* **Key Metrics to Display**:
* Surname Vulnerability Index (SVI)
* Electoral Integrity Index (EII)
* Spatial Clustering Coefficient
* Multi-level Variance Decomposition
* Top 10 Anomalous Booths

Success Metrics

* All analyses are based exclusively on existing data.
* Each analysis generates actionable insights.
* Visualisations are exportable in PNG/CSV formats.
* Statistical tests include calculations of effect size.
* Adheres to academic standards of rigour.

Notes for Implementation

* Use only provided data; do not fabricate any data.
* Include confidence intervals where appropriate.
* Clearly document all assumptions.
* Supply interpretation guides for complex statistics.
* Ensure mobile compatibility for all visualisations.
* Maintain consistent colour schemes.
* Document methodology for each analysis.

Priority Order for Maximum Impact

* Surname Vulnerability Dashboard
* Spatial Hot Spot Analysis
* Electoral Integrity Index
* Sankey Flow Diagrams
* Advanced Cross-tabulation Suite

This implementation guide presents a structured methodology for expanding the Bihar electoral analysis dashboard with robust analyses using available datasets.