Guidance Methodology Note: Predictive Modeling of People Exposed to Protection Risks

1. Purpose and Objectives

This methodology aims to guide the development and operationalization of predictive models that estimate the number of people exposed to protection risks at a sub-national level. The goal is to support evidence-based planning and prioritization of humanitarian interventions.

2. Model Objective

To estimate the population exposed to protection risks by leveraging protection severity scores and other relevant predictors such as conflict exposure, socio-economic vulnerability, and displacement patterns.

3. Data Requirements

Minimum Dataset (Admin Level 2 or 3 preferred):

- Protection Risk Severity Scores (15 protection risks)
- Population Baseline Data (e.g., host communities, IDPs, returnees)
- Conflict Indicators (e.g., ACLED: battles, violence against civilians, fatalities)
- Vulnerability Indicators (e.g., Multidimensional Poverty Index, food insecurity IPC Phase 3+)
- Legal and Social Access Indicators (e.g., access to justice, freedom of movement)

4. Pre-Modeling Analysis

4.1 Correlation Analysis

Generate a correlation matrix to examine relationships among protection risks. Identify multicollinearity among predictors (correlation > 0.7 may indicate redundancy). Example insights:
- GBV ↔ Psychological Abuse (0.78): high co-occurrence
- Mines ↔ Attacks on Civilians (0.73): indicates shared conflict dynamics

4.2 Protection Risk Mapping

Use heatmaps or choropleth maps to visualize regional severity of each protection risk. Rank regions based on median risk scores to identify hotspots.

5. Modeling Framework

5.1 Algorithm

Use machine learning algorithms like XGBoost for regression-based prediction. XGBoost handles non-linear relationships, is robust to multicollinearity, allows for regularization, and provides variable importance scores.

5.2 Predictor Selection

Prioritize predictors based on correlation strength with the outcome, data availability and quality, and contextual relevance to the country of operation.

5.3 Target Variable

Number or proportion of population exposed to at least one protection risk.

5.4 Training and Validation

Use cross-validation to prevent overfitting. Performance metrics:
- R^2 (Explained Variance) – Target > 0.90
- RMSE and MAE – For model precision

Use scatterplots (actual vs predicted) for visual validation.

6. Post-Modeling Outputs

- - Predicted exposed population per admin unit
- - Relative influence of predictors
- - Confidence intervals (optional)

7. Interpretation and Use

Use results to prioritize regions for protection interventions. Combine predictions with qualitative assessments (e.g., field assessments, SDR etc). Cap predicted values at total population per unit to prevent overestimation. Communicate uncertainty and assumptions clearly to stakeholders.

8. Limitations

Predictions depend on input data quality and availability. The model does not account for sudden shocks or emerging crises. Risk of over-reliance on statistical outputs—should be supplemented by contextual analysis.

9. Ethical Considerations

Ensure data privacy and security. Engage local actors and affected communities in interpreting results.