Social Upheaval Index: Data Collection & Validation Strategy

Phase 1: Data Collection Framework

Test Decades Selection

Recommended focus on 4-5 decades for initial validation:

- **1960s** (Expected HIGH) - Assassinations, Vietnam, civil rights upheaval
- **1970s** (Expected HIGH) - Watergate, oil crisis, economic stagflation
- **1950s** (Expected LOW) - Post-war prosperity, social conformity
- **1990s** (Expected LOW-MEDIUM) - Economic boom, end of Cold War, but some scandals
- **2010s** (Expected MEDIUM-HIGH) - Financial crisis aftermath, polarization, institutional strain

This gives you clear high/low expectations to validate against.

Component 1: Political Violence & Instability (25% weight)

High-Priority Data Sources

Major Political Assassinations

What to collect:

- President, VP, major presidential candidates
- Supreme Court justices
- Major civil rights leaders (MLK, Malcolm X level)
- Congressional leaders (Speaker, majority leaders)

Data Sources:

- Wikipedia: "List of assassinated American politicians"
- Secret Service historical reports
- FBI historical archives

1960s Expected Data:

- JFK (1963)
- Malcolm X (1965)
- MLK (1968)
- RFK (1968)
  Score: ~4 × 20 = 80 points

Domestic Terrorist Attacks

What to collect:

- Politically motivated bombings/attacks
- Mass casualty events with political goals
- Government building attacks

Data Sources:

- FBI Terrorism Database
- START Global Terrorism Database (GTD)
- "Days of Rage" book by Bryan Burrough (1970s bombing campaigns)

1970s Expected Data:

- Weather Underground bombings
- FALN Puerto Rican bombings
- SLA kidnappings/attacks
  Score: ~6-8 × 15 = 90-120 points

Major Riots/Civil Unrest

What to collect:

- Multi-day riots with federal intervention

- Death toll >5, property damage >$10M (inflation adjusted)
- National Guard deployments

Data Sources:

- Kerner Commission Report (1960s riots)
- FBI civil disorder reports
- National Guard deployment records

1960s Expected Data:

- Watts riots (1965)
- Newark riots (1967)
- Detroit riots (1967)
- MLK assassination riots (1968)
  Score: ~4 × 10 = 40 points

Large Political Protests

What to collect:

- 100,000 participants
- National significance (covered by all major networks)
- Government policy response required

1960s Expected Data:

- March on Washington (1963)
- Anti-Vietnam protests (multiple)
- Democratic Convention protests (1968)
  Score: ~5 × 5 = 25 points

Government Crises

What to collect:

- Presidential resignations/impeachments
- VP resignations

- Cabinet mass resignations
- Constitutional crises

1970s Expected Data:

- Watergate/Nixon resignation (1974)
- Spiro Agnew resignation (1973)
  Score: ~2 × 8 = 16 points

Component 2: Institutional Trust Erosion (20% weight)

High-Priority Data Sources

Major Political Scandals

What to collect:

- Presidential-level scandals requiring special prosecutor
- Congressional scandals affecting leadership
- Scandals resulting in resignations/convictions

Data Sources:

- Congressional ethics reports
- Department of Justice special counsel records
- Watergate archives

1970s Expected Data:

- Watergate (1972-1974)
- COINTELPRO revelations (1971)
  Score: ~2 × 25 = 50 points

Supreme Court Controversial Decisions

What to collect:

- 5-4 decisions on major social issues
- Decisions reversing long-standing precedent

- Decisions generating major protests/backlash

Data Sources:

- Supreme Court Database
- Constitutional law textbooks
- Gallup polling on court approval

Intelligence/Military Scandals

What to collect:

- CIA/FBI domestic surveillance programs
- Military cover-ups or war crimes
- Intelligence agency coups/interventions

1970s Expected Data:

- Pentagon Papers (1971)
- Church Committee revelations (1975)
- COINTELPRO exposure (1971)
  Score: ~3 × 15 = 45 points

Component 3: Economic Stress (15% weight)

High-Priority Data Sources

Recession Data

What to collect:

- NBER official recession dates
- Peak unemployment rates during recessions
- Duration in months

Data Sources:

- Bureau of Labor Statistics

- NBER recession database
- Federal Reserve Economic Data (FRED)

1970s Expected Data:

- 1970 recession: 8 months, 6.1% unemployment
- 1973-75 recession: 16 months, 9.0% unemployment
  Score: (8 × 6.1) + (16 × 9.0) = 193 → capped at 40

Income Inequality

What to collect:

- Gini coefficient by decade
- Changes in coefficient over decade

Data Sources:

- Census Bureau Historical Income Tables
- Piketty & Saez inequality database
- Congressional Budget Office reports

Major Financial Scandals

What to collect:

- Savings & Loan crisis scale events
- Major corporate bankruptcies with fraud
- Banking system failures

Validation Framework

Step 1: Historical Consensus Test

```
def validate_against_history():
    expected_rankings = {
        1960: 1,  # Highest upheaval
        1970: 2,  # Second highest
```

```
        2010: 3,  # Third (financial crisis)
        1990: 4,  # Lower upheaval
        1950: 5   # Lowest upheaval
    }

    calculated_rankings = calculate_all_decades()

    # Spearman correlation between expected and calculated
    correlation = spearman_correlation(expected_rankings, calculated_ranking
s)

    return correlation > 0.8  # Strong validation threshold
```

Step 2: Component Sensitivity Analysis

```
def test_component_importance():
    base_scores = calculate_base_scores()

    # Test removing each component
    for component in ['political_violence', 'institutional_trust', ...]:
        modified_scores = calculate_without_component(component)
        impact = correlation_change(base_scores, modified_scores)
        print(f"{component} removal impact: {impact}")
```

Step 3: Weight Optimization

```
def optimize_weights():
    # Try different weight combinations
    weight_combinations = generate_weight_grid()

    best_correlation = 0
    best_weights = None

    for weights in weight_combinations:
        scores = calculate_with_weights(weights)
```

```
        correlation = validate_against_history(scores)

        if correlation > best_correlation:
            best_correlation = correlation
            best_weights = weights

    return best_weights, best_correlation
```

Data Collection Tools

Recommended Approach

1. **Start with Wikipedia** for initial event lists and dates
2. **Cross-reference with official sources** for accuracy
3. **Use academic sources** for interpretation and context
4. **Create standardized coding sheet** for each event

Sample Data Structure

```
decade_1970s = {
    'political_violence': {
        'assassinations_major': 0,
        'terrorist_attacks_domestic': 8,  # Weather Underground, etc.
        'riots_major': 2,  # Kent State, etc.
        'protests_large': 6,  # Anti-war protests
        'government_crises': 2  # Watergate, Agnew
    },
    'institutional_trust': {
        'major_scandals': 2,  # Watergate, COINTELPRO
        'supreme_court_controversial': 1,  # Roe v Wade
        'intelligence_scandals': 3,  # Church Committee revelations
        'electoral_controversies': 1
    },
```

```
    # ... other components
}
```

Timeline

- **Week 1**: Collect political violence data for 5 test decades
- **Week 2**: Collect institutional trust data
- **Week 3**: Collect economic stress data
- **Week 4**: Build calculator, run initial validation
- **Week 5**: Optimize weights, refine methodology

This focused approach will give you a robust, validated upheaval metric before you apply it to any cultural analysis.