

Analytical Workflow: Vastopolis Outbreak Investigation

Stage 1 – Data Preparation

- **Clean & geocode** microblog data; remove duplicates/bots; parse timestamps to consistent timezone.
- **Join population statistics** to the 13 zones (spatial join using zone boundaries).
- **Align weather data** (wind speed/direction) by day for later correlation.
- *Checkpoint:* Visual scatter of all geolocated posts on the map to confirm coverage and spot GPS errors (e.g., points outside the metro area).

Stage 2 – Text Filtering & Symptom Classification

- Apply **keyword/NLP filtering** (flu, fever, cough, sick, vomiting, etc.) to isolate illness-related posts.
- Optionally cluster symptom terms to distinguish **respiratory vs. gastrointestinal** language (clue for airborne vs. waterborne).
- *Checkpoint:* Analyst manually samples classified vs. unclassified posts to verify precision/recall of the filter.

Stage 3 – Spatiotemporal Exploration

- Build a **time-animated density map** (heatmap or hexbin) of illness-related posts by hour/day.
- Identify the **earliest cluster(s)** in time — candidate outbreak origin.
- Overlay **hospitals, landmarks, water bodies, highways** to see proximity of early cases.
- *Checkpoint:* Domain expert inspects animation to flag plausible index location(s) (e.g., near a reservoir, transit hub, or stadium).

Stage 4 – Population Normalization

- Compute **illness-report rate per capita** per zone per day (using population stats), not just raw counts.
- Rank zones by normalized incidence to correct for population density bias.
- *Checkpoint:* Compare raw vs. normalized hotspot maps — resolve discrepancies with analyst judgment.

Stage 5 – Transmission-Pathway Hypothesis Testing

Hypothesis	Expected Pattern	Data to Check
Waterborne (point source)	Cases cluster around a specific water body/supply; sudden onset near source, then diffuse	Overlay cases on water infrastructure; check GI symptoms
Airborne (wind-driven)	Case density shifts downwind over time; direction matches wind vector	Cross-correlate daily wind direction with spatial shift of new cases
Person-to-person	Gradual spatial spread from index case along transit/commute routes; growth follows population density/mobility, not wind	Check spread along highways/transit corridors, contact-network-like diffusion

- Visualize **case centroid movement per day** vs. **wind vector per day** (overlay arrows).
- Visualize **spread rate vs. distance from highways/transit** to test person-to-person diffusion.
- *Checkpoint:* Analyst determines which hypothesis best fits observed spatial-temporal signature; may combine (e.g., point-source origin + secondary person-to-person spread).

Stage 6 – Source Localization

- Trace back the **earliest normalized cluster** and cross-reference with map landmarks (e.g., convention center, water treatment plant, market).
- Validate using **text content** of earliest posts (mentions of specific location/event).

Stage 7 – Synthesis & Reporting

- Combine spatial, temporal, and textual evidence into a narrative: **likely origin point, transmission mode, and current spread trajectory**.
 - Provide **actionable map** for emergency response (zones needing resources, quarantine perimeter suggestions based on wind/mobility spread model).
-

Tools suited for this: GIS/mapping (QGIS, Kepler.gl, ArcGIS), text mining (Python/NLTK or spaCy), time-series/animation (Tableau, D3, or Python+Plotly), all with iterative human review at each checkpoint above.

Möchtest du, dass ich einen der Schritte (z. B. Textklassifikation oder die Windkorrelation) konkreter mit Beispielcode ausarbeite?