

Recommended Analytical Workflow: Vastopolis Outbreak — Origin and Transmission Mechanism

The task decomposes into three linked analytical questions that should be addressed in sequence but revisited iteratively:

1. **When and where did the outbreak begin?** (spatio-temporal origin detection)
2. **How did it spread?** (diffusion pattern in space and time)
3. **What is the likely transmission mechanism** — person-to-person, airborne, waterborne, or other — inferred from how the spread pattern relates to human mobility infrastructure, water bodies, and wind?

Because microblog text is unstructured, population data is tabular, and weather is a bivariate time series, the workflow combines text-mining/entity-extraction logic (from **Jigsaw**), event-to-place extraction (from **"From Movement Tracks through Events to Places"**, Andrienko et al. 2011), spatio-temporal grouping and modelling (from **"A Visual Analytics Framework for Spatio-Temporal Analysis and Modelling"**, Andrienko & Andrienko 2013), calendar/temporal-cluster reasoning (from **van Wijk & van Selow's Cluster and Calendar-Based Visualization**), and lagged cross-series analysis (from **"Detecting Stable Cross-Impact Patterns in Bivariate Time Series"**, Andrienko et al. 2026) to test the wind/airborne hypothesis specifically.

Stage A — Extract candidate illness events from microblog text

Goal: turn raw geolocated text into a structured set of "illness-report events," each with a location, timestamp, and semantic content.

- Build a lexicon of flu-related symptom terms (fever, cough, vomiting, headache, chills, "sick," "dying," "hospital," "ER," etc.) and apply keyword/entity extraction to every message, analogous to Jigsaw's entity extraction step where persons, locations, organizations, and dates are pulled from text. Here the extracted "entities" are symptom mentions, severity indicators (e.g., mentions of death or hospitalization), and place/landmark references embedded in the text (which may be more precise than the GPS tag, e.g., "at the lake festival").
- As in Jigsaw, treat extraction as imperfect: the analyst samples and corrects false positives/negatives (e.g., "flu-like" sarcasm, unrelated "sick of my job" posts) before trusting the derived event set. This mirrors Jigsaw's entity-correction loop and the human-in-the-loop feature-validation step from **"Feature Engineering for Behaviour**

Pattern Recognition" (Andrienko et al. 2024), where engineered indicators are checked visually against representative examples before being trusted for further analysis.

- Tag each retained message with a severity level (mild symptom mention vs. severe/death mention), since severity will later help disambiguate early point-source cases from general community spread.
 - Output: a table of geolocated, time-stamped "illness events," each with symptom type(s), severity, and raw text (kept for later drill-down reading).
-

Stage B — Temporal overview: detect onset and growth pattern

Goal: establish *when* the sharp increase began and whether it has a single abrupt onset (point-source signature) or a gradual ramp (person-to-person diffusion signature).

- Aggregate illness-event counts into hourly (or 6-hour) bins, city-wide and per zone, producing a short multivariate time series (13 zones × time).
 - Use a timeline/heatmap overview in the spirit of **IVESA**: sort zones by onset time, event density, or burst intensity; visually inspect for gaps, sudden bursts, and regularity versus irregularity of report accumulation.
 - Where the data spans enough days to show daily rhythm (e.g., normal daytime microblogging vs. night silence), apply the **van Wijk Cluster-and-Calendar** logic: split the series into daily profiles, cluster days by shape (using a normalized or offset-corrected distance so that the outbreak's abnormal shape stands out from the "normal" pre-outbreak baseline), and display cluster membership per day. This immediately flags the specific day(s) that break from the baseline pattern — the operational start of the outbreak — without requiring the analyst to predefine a threshold.
 - Judgment point: identify the earliest zone(s) whose case curve deviates from baseline. These are candidates for the spatial origin, to be confirmed in Stage C.
-

Stage C — Spatial hotspot / significant-place extraction

Goal: convert scattered illness events into a small number of interpretable "significant places," and identify which one is earliest (candidate origin).

This stage directly follows the workflow in "**From Movement Tracks through Events to Places**" (Andrienko et al. 2011), substituting illness-report events for the paper's m-events:

- Apply density-based clustering (DBSCAN-like) to the illness events in space and time jointly, using a distance function that combines spatial proximity and temporal proximity (and optionally symptom similarity), so that events close in space but far apart in time are *not* merged, and events in different unrelated locations are not merged. Isolated events are treated as noise.
 - Build spatial buffers/hulls around each event cluster to delineate candidate "outbreak places" (e.g., a park, a water treatment plant vicinity, a stadium, a hospital catchment).
 - For each place, compute a small time series of event counts, enabling a "when did this place become active" comparison across places — mirroring the paper's aggregation of events by place into time series.
 - Cross-reference the earliest, most severe cluster(s) with the city map's landmark layer (hospitals, water bodies, event venues) to see whether the earliest cluster coincides with a specific fixed feature (e.g., a reservoir, a festival ground) — a signature consistent with a point-source (waterborne or common-exposure) event, as opposed to diffuse residential spread consistent with person-to-person transmission.
 - Normalize event counts per place/zone by the 13-zone **population statistics**, producing incidence-per-capita rather than raw counts, since a raw count of reports may simply reflect population density or Twitter usage density rather than true attack rate.
-

Stage D — Spatio-temporal grouping and growth modelling across zones

Goal: determine whether the 13 zones behave as one expanding front (diffusion) or as several independently-seeded clusters (multiple point sources / common exposure).

Following the **Spatio-temporal Analysis and Modelling** framework (Andrienko & Andrienko, 2013):

- Treat each zone's per-capita incidence time series as one spatial time series. Cluster zones by similarity of temporal profile (e.g., k-means or hierarchical clustering on normalized incidence curves), using a linked time-graph + map display so the analyst can check whether temporally similar zones are also spatially adjacent (supporting contiguous diffusion, e.g., along highways) or spatially disconnected (supporting a shared non-contact cause, e.g., a common water source or wind-borne plume).

- Fit a simple growth model (e.g., exponential/logistic onset curve) to each group's representative curve and inspect residuals over time and space, as in the source workflow, to detect zones that deviate from the dominant growth pattern — these deviating zones are candidates for a distinct secondary source or a differently-timed exposure (e.g., a downwind zone affected only after a wind-driven delay).
 - Overlay zone clusters on the map together with the highway network: if the incidence-similar/temporally-lagged zones are strung along major highways or transit corridors, this favors person-to-person/mobility-driven spread; if they are unrelated to the road network but spatially coincide with a shared water body or its service area, this favors waterborne transmission; if they align with a downwind sector regardless of road connectivity or population contact, this favors airborne spread.
-

Stage E — Wind–spread cross-impact analysis (testing the airborne hypothesis specifically)

Goal: rigorously test whether the direction and timing of spread is statistically consistent with wind advection from the candidate origin, using a lagged-correlation approach rather than eyeballing alone.

This stage directly adapts "**Detecting Stable Cross-Impact Patterns in Bivariate Time Series**" (Andrienko et al., 2026):

- Define the **base series** as the daily wind vector (speed and direction) at/near the candidate origin place identified in Stage C.
- Define the **response series** as a directional exposure measure per zone: e.g., the daily per-capita incidence in each of the other 12 zones, or better, a projected "downwind alignment score" per zone (how well the zone's bearing from the origin matches the day's wind direction).
- Slide a time window across the base series and compare it with time-shifted windows of each zone's response series over a range of plausible incubation lags (using a tolerance-based association measure such as the paper's tolerance-Kendall's-tau, which is robust to noisy daily fluctuations).
- Visualize the resulting time×lag impact matrix with an lkat-style plot for each zone (or for an aggregated "downwind vs upwind" comparison), and calibrate window length, lag range, and tolerance interactively while cross-checking against the actual wind and incidence series behind any cell.
- **Interpretation:** if downwind zones show a consistent positive impact at a lag plausible for airborne incubation/dispersal, and upwind/crosswind zones do not, this is strong evidence for airborne transmission from the identified origin. If no stable

directional lag pattern emerges (spread is equally likely upwind, downwind, and cross-wind, or tracks population/road proximity instead), airborne transmission from a single point source is less supported, and the analyst should weight person-to-person or waterborne explanations more heavily.

Stage F — Person-to-person and waterborne hypothesis checks

Goal: directly test the two remaining mechanisms using complementary evidence, since Stage E only tests the wind/airborne hypothesis.

- **Person-to-person check:** using the corrected illness-event set from Stage A, extract social/co-location mentions from the text (e.g., "my coworker," "my whole family," shared venues, shared bus/train lines) — the same entity/relation-extraction logic as **Jigsaw**, here applied to build a light co-occurrence network of people/places mentioned together. If cases chain along social/mobility links (school, workplace, transit routes) and the spatial growth in Stage D tracks the highway/transit network rather than a single fixed landmark, this supports person-to-person spread following normal human contact and commuting patterns.
 - **Waterborne check:** overlay the earliest and largest clusters from Stage C directly on the map's water-body layer and any known water-distribution/service-area boundaries for the 13 zones (if available via the population/zone metadata). A waterborne signature would show: (a) an origin cluster physically coincident with a water source (reservoir, treatment plant, river) rather than a crowded public venue; (b) case distribution matching the water utility's service zones rather than geographic proximity or road adjacency; (c) no dependence on wind direction (ruling out Stage E's airborne signature) and no dependence on social/mobility chains (ruling out Stage F's person-to-person signature).
 - Because these are complementary, mutually-exclusive-leaning hypotheses, structure this stage as a comparison table/checklist: for each candidate mechanism, list the diagnostic spatial/temporal/contextual signature expected, and whether the data satisfies it — supporting transparent, auditable reasoning rather than a single black-box conclusion.
-

Stage G — Synthesis, drill-down, and reporting

Goal: consolidate findings into a defensible narrative the public-health team can act on.

- Maintain a coordinated multi-view dashboard throughout: (1) a map view showing hotspot clusters, water bodies, highways, hospitals, and wind vectors per day; (2) a

timeline/calendar view showing zone-level incidence and cluster membership over time; (3) a text drill-down panel to re-read raw microblog content behind any cluster or spike (essential domain-knowledge check, as emphasized throughout Jigsaw's Document View and the raw-record drill-downs in DPVis and Kokiri); (4) the Ikat-style wind-impact panel from Stage E; (5) a zone table with population-normalized rates.

- Use an iterative "propose visualize validate against raw text/map refine" loop, consistent with the general pattern seen across the library (e.g., **explAIner's** understand–diagnose–refine cycle and **SOMFlow's** provenance-tracked iterative refinement): each hypothesis about origin or mechanism should be checked against raw messages and map context before being accepted, and parameter choices (clustering distance, lag window, tolerance) should be logged so the reasoning is reproducible.
- Deliverable: a short evidence-linked report identifying (a) the most probable origin place and time window, (b) the dominant spread pattern (contiguous/road-following vs. wind-directional vs. water-service-area-bound), and (c) the best-supported transmission mechanism with the specific visual/statistical evidence (Stage C cluster map, Stage D growth/residual plots, Stage E Ikat wind-lag plot, Stage F checklist) cited for each claim, so emergency-management resources can be targeted (e.g., quarantine/contact-tracing near roads, water-supply intervention, or air-quality/evacuation measures) accordingly.