

Recommended workflow: outbreak-origin and transmission-hypothesis analysis

Important limitation: microblog data are a proxy for illness and public concern, not confirmed cases. This workflow should identify and rank **operational hypotheses** for deployment, not claim a definitive epidemiological cause. Confirm findings with clinical surveillance, environmental sampling, and field investigation.

1. Prepare trusted spatiotemporal signals

1. Clean the posts

- Remove duplicates, obvious bots, repost cascades, and posts with unreliable coordinates or timestamps.
- Retain uncertainty flags for low GPS accuracy, ambiguous text, and potentially automated accounts.
- Aggregate data into several spatial resolutions: city zones for population-normalized reporting, plus a fine grid or adaptive hexagons for local hotspot detection.
- Use several temporal bins, such as 1-, 3-, and 6-hour intervals, to check that conclusions do not depend on one bin size.

2. Classify message content with analyst review

- Create categories such as:
 - self-reported flu-like symptoms;
 - reports about another ill person;
 - death or severe-illness reports;
 - healthcare-seeking or hospital mentions;
 - concern, news, or misinformation;
 - references to locations, events, water, ventilation, or travel.
- Begin with keyword rules and a topic model, then have public-health analysts inspect representative messages and correct misleading terms, sarcasm, news reposts, and non-flu conditions.
- Use the iterative pattern from **Human-Steered Topic Modelling**: inspect topics and documents, merge or split categories, adjust terms, retrain, and review the changed results.
- Extract named places, landmarks, hospitals, and organizations from text and verify them in context, following the evidence-grounding approach of **Jigsaw**.

3. Construct outbreak indicators

- For each spatial unit and time bin, calculate:
 - high-confidence symptom-post count;
 - symptom-post rate per population;
 - symptom posts as a share of all local posts, to reduce bias from variation in platform activity;
 - severe/death-associated post counts, kept separate from symptom counts;
 - uncertainty intervals reflecting classification confidence and low message volume.
- Treat hospital mentions primarily as a possible **care-seeking/reporting effect**, not immediate evidence that hospitals are infection sources.

Visual checkpoint: Show linked maps, temporal heatmaps, zone time series, and sampled original posts. The analyst should confirm that apparent peaks represent plausible illness reports rather than news amplification or a single viral post.

2. Detect the first sustained outbreak locations

1. Map onset and excess activity

- Visualize symptom intensity over time as an animated map and space–time cube.
- For each cell or zone, estimate its onset time: the first time when symptom activity becomes elevated and remains elevated for a defined duration.
- Show both raw counts and normalized rates. Dense central areas may have more posts but not necessarily higher illness prevalence.

2. Identify persistent spatiotemporal clusters

- Extract high-confidence symptom events and apply density-based clustering in space and time.
- Require a cluster to be persistent across adjacent time bins and supported by multiple independent accounts where possible.
- Adapt the logic of **From Movement Tracks through Events to Places**: identify meaningful event concentrations first, then turn recurring concentrations into candidate outbreak places rather than relying only on predefined city zones.

3. Rank candidate origin areas

- Rank clusters by:
 - earliest sustained onset;
 - intensity after normalization;
 - persistence;

- evidence that later hotspots emerge after the candidate;
 - robustness across text-classification thresholds and spatial/temporal resolutions.
- Display temporal expansion contours from each candidate area and inspect nearby landmarks, water bodies, highway access, and hospitals.

Decision rule: distinguish the **origin of the observed social-media signal** from the true infection origin. A hospital-area cluster may be the first visible signal because sick people seek care there, even if transmission began elsewhere.

3. Build explicit competing transmission hypotheses

Create a small set of testable hypotheses rather than searching for one pattern without a comparison baseline.

Hypothesis	Expected observable pattern	Main visual and analytical tests	Operational implication if supported
Person-to-person spread	Serial local growth; neighboring zones rise after earlier hotspots; spread may follow dense population and major travel corridors	Compare lagged symptom rates among adjacent zones and highway-connected zones; inspect spatial waves and residual maps	Deploy testing, clinical outreach, public messaging, and crowd-reduction measures in expanding hotspot zones
Airborne or aerosol transport	New cases occur preferentially downwind of a candidate source; direction changes with wind changes; distance and wind strength matter	Overlay wind vectors and downwind exposure plumes on onset maps; compare upwind and downwind zones at plausible lags	Inspect candidate emission sites, ventilation systems, industrial sites, and enclosed venues; issue targeted exposure guidance

Hypothesis	Expected observable pattern	Main visual and analytical tests	Operational implication if supported
Waterborne exposure	Elevated activity near the same water body, catchment, distribution area, or downstream corridor; less dependence on wind or road adjacency	Compare zones by proximity to water bodies and—if available—water-service boundaries; inspect local text references	Collect water samples, inspect treatment and distribution infrastructure, consider targeted boil-water or avoidance guidance
Common event or point source	Abrupt, near-simultaneous onset among people associated with one place or event; limited serial spatial progression initially	Inspect landmark/event mentions, early temporal synchrony, and co-occurring place references	Investigate venue, event, food source, workplace, or institution; identify and notify exposed groups
Reporting or media artifact	Sharp message volume without corresponding self-reported symptoms, spatial consistency, or persistence	Compare symptom reports with news/concern posts and total posting volume; inspect source accounts	Avoid misallocating emergency resources based on social-media amplification

4. Test the hypotheses with linked visual models

For each zone and time period, construct explanatory variables:

- prior symptom intensity in the same zone;
- prior intensity in adjacent or highway-connected zones;
- distance from candidate origin locations;
- downwind exposure, using wind direction converted to the direction of air transport;
- wind speed;

- distance to or relationship with water bodies;
- population and total microblog activity;
- proximity to hospitals and landmarks, as potential reporting confounders.

Fit several transparent spatiotemporal models, for example count or rate models with alternative transmission terms:

1. **Contact-propagation model:** local prior rate plus lagged neighboring/highway-connected rates.
2. **Airborne model:** source strength plus downwind alignment, wind speed, distance, and lag.
3. **Water-exposure model:** water proximity or service-area term, adjusted for population and posting activity.
4. **Mixed model:** combines plausible mechanisms, but penalizes unnecessary complexity.

Use the residual-driven process from **Partition-based Regression Modelling**:

- inspect how symptom rates vary against each candidate factor;
- fit an initial model;
- map and plot its residuals;
- add a factor only when residual structure suggests a meaningful unexplained pattern;
- compare models using held-out periods or locations, not only in-sample fit.

Use the grouping-and-residual logic of **Spatio-temporal Analysis and Modelling** to:

- cluster zones with similar outbreak trajectories;
- map those groups;
- determine whether trajectory groups align with wind exposure, water proximity, or road-connected spatial propagation;
- inspect whether residual hotspots remain after accounting for each proposed mechanism.

For wind-related claims, compare multiple periods with different wind direction. A genuine airborne pattern should move or strengthen in the new downwind direction. The time-lag calibration concepts from **Detecting Stable Cross-Impact Patterns in Bivariate Time Series** can help visually inspect whether wind changes consistently precede changes in downwind symptom activity.

Visual checkpoint: For every model, show:

- observed versus predicted zone-time rates;

- a map of residuals;
- uncertainty/confidence intervals;
- example messages and locations underlying key anomalies.

A model that predicts well but leaves concentrated residual hotspots is insufficient; those hotspots may reveal another source, a reporting bias, or a mixed transmission mechanism.

5. Investigate candidate sources in detail

For the top-ranked origin candidates:

1. Select the earliest local cluster and inspect all associated messages in temporal order.
2. Review referenced landmarks, facilities, events, workplaces, water locations, and travel mentions.
3. Compare early candidate areas with matched areas that had similar population and posting volume but did **not** develop high symptom activity.
4. Trace later hotspots backward:
 - Did they arise after nearby hotspots?
 - Were they downwind at the relevant time?
 - Do they share a water feature or service boundary?
 - Are they connected by highways or major mobility corridors?
5. Record each judgement, supporting evidence, contradictory evidence, and confidence level.

If user identifiers and multiple posts per user become available, derive anonymized aggregate origin–destination flows. At that point, the spatial- and temporal-aggregation strategy of **MobilityGraphs** can help assess whether mobility flows plausibly explain the observed zone-to-zone progression. Without user linkage, do **not** infer person-level movement or contact chains from individual posts.

6. Produce an emergency-management dashboard and update loop

Deliver four linked outputs:

1. **Current risk map:** population-normalized symptom activity, severe-report signal, uncertainty, and emerging hotspots.
2. **Origin candidates panel:** ranked locations with onset time, evidence, contradictory evidence, and sensitivity results.

3. **Transmission evidence panel:** comparative support for person-to-person, airborne, waterborne, point-source, and mixed explanations.
4. **Resource-priority map:** zones prioritized by current intensity, growth rate, vulnerable population count, proximity to hospitals, and uncertainty.

Re-run the workflow as new posts and weather observations arrive. Preserve an analysis log recording text-model changes, thresholds, cluster settings, selected hypotheses, and analyst interpretations.

Essential external data to request immediately

The current data can generate strong leads, but these additional data are needed to validate them:

- de-identified confirmed-case line lists, including symptom onset date and approximate residence or exposure area;
- emergency-department visits, lab test positivity, ambulance calls, and death records;
- water distribution, sewer, treatment, and sampling data;
- higher-resolution wind and air-quality observations;
- major-event, school, workplace, and transit information;
- anonymized aggregate mobility data, if legally and ethically available.

The recommended decision posture is therefore: **deploy resources to high-confidence, rapidly growing hotspots immediately; investigate the leading transmission hypotheses in parallel; and update interventions as clinical and environmental evidence confirms or rejects the social-media-derived patterns.**