Opportunities and limitations of avian influenza surveillance data harmonization for outbreak forecasting in Vietnam

T. Hoang Thi[1] (thi-thuy.hoang@sydney.edu.au), R. de Lara Muylaert[1], C. Thure Kirkeby[2], M. Patrick Ward[1]

[1] *School of Veterinary Science, University of Sydney, Sydney, Australia*

[2] *Department of Veterinary and Animal Sciences, University of Copenhagen, Copenhagen, Denmark*

Summary

Avian influenza is spreading globally, affecting the poultry industry and wildlife biodiversity. Vietnam has been recognised as a hotspot for highly pathogenic avian influenza (HPAI), experiencing recurrent outbreaks since its first detection in late 2003 (Gutiérrez et al., 2009). Understanding the spatiotemporal structure of outbreaks is essential for effective surveillance and control.

This study evaluated structural differences in HPAI outbreak records reported by two international surveillance systems, WAHIS and EMPRES-i, in Vietnam from 2006 to 2022. We assessed how reporting structure and data processing influence spatiotemporal cluster detection and identified hotspots that persisted across analytical parameter settings.

Following harmonisation, deduplication, and spatial validation, clustering was analysed using the SaTScan space–time permutation model under multiple spatial (10–100 km) and temporal (30–90 days) windows. Sensitivity to window choice was examined within each dataset, and cross-dataset comparison used predefined spatial and temporal matching criteria.

EMPRES consistently detected more clusters with smaller radii and shorter durations, whereas WAHIS identified fewer but broader and longer-lasting clusters. Persistent hotspots were concentrated in northern Vietnam, particularly the Red River Delta, with secondary foci in the Mekong Delta. These differences primarily reflect surveillance structure rather than underlying incidence. Rather than privileging one dataset over the other, their complementary strengths suggest that integration may provide broader insight when supported by rigorous harmonisation and sensitivity analysis. In particular, datasets with higher reporting density may be more sensitive to localised clustering, whereas more aggregated reporting structures may provide greater stability for regional or temporal trend analysis.

Methods

2.1 Data Sources

The WAHIS dataset was obtained from the World Organisation for Animal Health (WOAH) SharePoint platform (infur 20250623) and is hereafter referred to as the *WAHIS* It comprised 749 records of highly pathogenic avian influenza (HPAI) outbreaks reported between 2006 and 2022, each containing 63 variables.

The second dataset was obtained from the public interface EMPRES-i of the Food and Agriculture Organisation of the United Nations (FAO) and is hereafter referred to as the *EMPRES*. This dataset initially included 3,336 outbreak records (16 variables) spanning 2006-2022, covering both HPAI and low-pathogenic avian influenza (LPAI). After restricting the dataset to HPAI outbreaks only, 1,513 records were retained for analysis.

Each row in both datasets was assumed to represent a single outbreak event. Variables related to outbreak timing, geographic location, and key epidemiological characteristics were selected for subsequent analysis.

2.2 Data Quality Assessment

Data wrangling and harmonisation were conducted in R (version 4.5.1). Variable names were standardised across years, and Ha Tay province, merged into Ha Noi in 2008, was recoded as Ha Noi to ensure temporal consistency.

Exploratory analysis assessed data completeness and duplicate records. In EMPRES, 2.7% of outbreak start dates were missing and were imputed using report dates where appropriate. De-duplication procedures differed due to reporting structure. WAHIS was deduplicated using outbreak ID, start date, coordinates, location name, and province. EMPRES, which frequently recorded multiple entries for the same event, was deduplicated based on observation date, locality, province, and coordinates.

Spatial validation was conducted by overlaying outbreak coordinates with provincial boundaries from the Global Administrative Areas database (GADM v4.1). Records falling outside their reported province were flagged, and distances to administrative boundaries were calculated. Coordinates within 5 km of the assigned boundary were retained. Larger discrepancies were investigated through commune-level matching and external verification via local news outlets, and corrected coordinates were reassigned to commune centroids where necessary.

Outbreak duration was calculated for WAHIS records using available start and end dates to characterise temporal patterns.

2.3 Clustering Assessment

Spatiotemporal clustering was assessed using SaTScan (version 10.1) (Kulldorff et al., 2005) with the space–time permutation model. This model detects space–time interaction without requiring population-at-risk

denominators, evaluating whether cases are more concentrated in space and time than expected given the dataset's marginal spatial and temporal distributions.

Sequential Monte Carlo inference was applied to improve computational efficiency. Clustering was evaluated across multiple spatial (10, 50, and 100 km) and temporal (30, 60, and 90 days) window settings to assess scale sensitivity, yielding 9 parameter combinations per dataset. Clusters were compared across settings and between datasets using the number of significant clusters ($p \leq 0.05$), cluster radius, duration, observed case counts and test statistics, and spatial and temporal displacement between matched clusters.

Cluster agreement across parameter settings and between datasets was assessed using predefined spatial and temporal overlap criteria. Two clusters were considered matched if the centroid distance was less than or equal to 0.5 times the sum of their radii and if the temporal overlap proportion exceeded 70% of the union of their time intervals. Cluster persistence was defined as the proportion of parameter settings in which a matched cluster was detected. Clusters appearing in five or more settings (>50%) were classified as persistent signals.

Results

3.1 Data Quality

Data validation revealed structural inconsistencies in the WAHIS dataset, including approximately 5% duplicate events, primarily concentrated between December 2020 and February 2022, and 20% spatially mismatched records. Additionally, approximately 0.7% of outbreak coordinates were positioned outside the national boundary or in coastal waters. Spatial validation of WAHIS outbreak coordinates revealed substantial geolocation inconsistencies with reported administrative limits. A substantial proportion of records (17.6%; N = 125) were located 30–200 km outside their reported provincial boundary, with extreme outliers exceeding 300 km and reaching up to 1,000 km. The locality descriptions were heterogeneous and inconsistently defined, often referring to different political-geographic units, including communes, villages, streets, or informal place names. This variation substantially complicated automated geocoding and likely contributed to the misplacement of coordinates. However, cross-verification with contemporaneous outbreak reports published in online news sources enabled manual validation and correction of many affected records (8.5%, N = 61), highlighting the potential added value of triangulating surveillance data with publicly available outbreak reporting and opportunities to enhance internal troubleshooting steps in surveillance systems.

In contrast, EMPRES records showed consistent geolocation but included approximately 8% duplicate entries, mostly due to multi-species reporting. These findings indicate systematic geocoding errors in the database and highlight the need for spatial validation and correction prior to spatiotemporal modelling.

After de-duplication and harmonisation, 712 outbreak records were retained from the WAHIS dataset and 1,271 from the EMPRES dataset, covering the period 2006–2022. In the WAHIS dataset, the median outbreak duration was 25 days (mean 26.1 days; SD 8.0), and duration remained relatively stable over time. Most outbreaks were resolved within 3–4 weeks, consistent with established containment, culling, and quarantine procedures for highly pathogenic avian influenza. Prolonged outbreaks were uncommon, with a maximum recorded duration of 131 days, and likely reflect reporting delays or administrative closure processes rather than sustained farm-level transmission. Overall, the temporal stability of outbreak duration suggests a relatively consistent outbreak management cycle throughout the study period, despite marked fluctuations in outbreak incidence.

Notably, across 17 years of surveillance data, only one HPAI record involving a wild or non-domestic animal was identified in each dataset. In WAHIS, this record involved a non-avian animal, a captive civet, while in EMPRES the single non-poultry record corresponded to a farmed swift. No confirmed outbreaks in free-ranging wild birds were recorded in either dataset during the study period. This absence of wild bird records likely reflects limited wildlife surveillance and reporting rather than the true absence of infection.

3.2 Cluster Detection

Space–time permutation analysis identified marked differences in clustering structure between the EMPRES and WAHIS datasets across parameter settings. EMPRES consistently detected a higher number of significant clusters (40–55 across window combinations) compared with WAHIS (14–18), with clusters in EMPRES characterised by smaller median radii and shorter durations. Increasing spatial and temporal window sizes resulted in fewer but larger and longer clusters in both datasets, consistent with the merging of local outbreak signals into broader regional waves.

Persistent space–time clusters (detected in > 50% of 9 parameter settings) exhibited clear geographic structuring across Vietnam. Robust cluster centres were predominantly located in northern Vietnam, particularly within the Red River Delta and adjacent northeastern provinces, where multiple persistent signals were detected across both the EMPRES and WAHIS datasets. Additional stable clusters were identified in southern Vietnam, notably within the Mekong Delta region, although these were fewer and more spatially dispersed. Several persistent centres were also observed along the North Central Coast, forming a corridor of recurrent outbreak activity extending south from the Red River Delta. In contrast, the Central Highlands and much of the South Central Coast showed comparatively limited persistence, with only isolated recurrent cluster centres detected.

Notably, several northern hotspot areas were detected consistently in both datasets, suggesting recurrent amplification zones rather than parameter-dependent artefacts. However, slight spatial displacement between the EMPRES and WAHIS cluster centres within the same regions was observed, consistent with differences in geocoding precision or outbreak-aggregation. The spatial concentration of persistent clusters in northern

Vietnam and parts of the Mekong Delta broadly corresponds to provinces with higher poultry density, suggesting that structural production intensity may contribute to recurrent outbreak patterns. Although poultry density was not formally incorporated into the clustering model, this alignment supports the plausibility of identified hotspots and indicates that integrating independent production data could help refine surveillance prioritisation. Areas with substantial poultry populations but limited persistent clustering may warrant closer examination to determine whether reduced detection reflects genuinely lower risk or potential surveillance gaps.

Discussion

4.1 Implications for Risk Modelling and Policy

The results have important implications for downstream modelling and decision-making. At fine spatial scales, such as 10 to 30 km, cluster detection proved highly sensitive to reporting granularity and coordinate precision. Fine-scale risk models that do not account for these structural differences may overestimate localised hotspots or interpret reporting artefacts as meaningful transmission dynamics.

In contrast, clustering using broader spatial windows of 50 to 100 km appeared less sensitive to coordinate variability and reporting granularity (Loth et al., 2021). However, given Vietnam's elongated and geographically narrow structure, larger spatial windows may also aggregate epidemiologically distinct areas, and results at these scales should therefore be interpreted cautiously. Persistent hotspots were consistently identified in northern Vietnam, particularly within the Red River Delta, and to a lesser extent in the Mekong Delta (Artois et al., 2016). These regions remained detectable across multiple analytical configurations and across datasets, suggesting structurally stable outbreak-prone areas rather than parameter-dependent artefacts.

The persistence framework applied in this study provides a practical approach to distinguishing robust signals from window-sensitive detections. Evaluating cluster stability across multiple spatial and temporal scales improves confidence in hotspot identification and strengthens the evidence base for targeted surveillance reinforcement and control strategies.

These findings highlight the importance of harmonised reporting standards and transparent metadata documentation in international surveillance systems. Clear outbreak definitions, consistent date attribution, and explicit reporting of geocoding precision would enhance cross-dataset comparability and improve the reliability of integrated analyses. Sensitivity analysis should therefore be considered an essential component of multi-source surveillance integration to ensure that modelling outputs reflect stable epidemiological patterns rather than artefacts of data architecture.

4.2 Surveillance System Limitations and Structural Differences

This comparative analysis indicates that differences in the spatiotemporal clustering patterns detected from WAHIS versus EMPRES data are largely attributable to structural characteristics of the surveillance systems rather than to underlying epidemiological differences. Several important limitations were identified.

First, duplicate inflation was evident, particularly in EMPRES, where outbreaks were sometimes recorded separately by affected species. Without rigorous harmonisation and deduplication, such reporting practices can artificially increase cluster counts and fragment outbreak signals. Second, inconsistencies in date attribution, including differences between outbreak onset and reporting dates, introduced temporal uncertainty that may shift cluster boundaries and influence measures of time overlap across datasets. Lastly, inconsistent and heterogeneous locality descriptors increase verification difficulty and constrain effective cross-dataset integration.

Spatial precision variability emerged as a major source of divergence. Outbreaks were geocoded at heterogeneous administrative levels, ranging from commune-level coordinates to district or provincial centroids. This variation influenced cluster radius, spatial displacement, and stability under smaller spatial windows. At fine scales, minor coordinate discrepancies substantially altered overlap assessments, whereas broader windows reduced sensitivity to geocoding heterogeneity. Differences in outbreak definition, particularly whether multi-species events were recorded as single or multiple outbreaks, further contributed to variation in cluster frequency and duration.

An additional limitation concerns the under-representation of wild bird surveillance data. Both WAHIS and EMPRES primarily capture reported outbreaks in domestic poultry, with relatively sparse and inconsistent reporting from wildlife surveillance systems. Given the recognised role of wild birds in maintaining and disseminating avian influenza viruses across regions, the detected clusters in this study predominantly reflect patterns within domestic production systems rather than full ecological transmission dynamics (Jindal et al., 2025). The absence of harmonised wildlife data limits interpretation at the wild–domestic interface and may underestimate pathways of introduction linked to migratory flyways. Incorporating structured wildlife surveillance data would improve ecological interpretation and strengthen One Health integration.

Together, these findings demonstrate that surveillance data structure directly shapes spatiotemporal inference.

Conclusion

Integrating multi-source avian influenza surveillance data provides valuable opportunities to strengthen outbreak detection and risk assessment. However, this study demonstrates that differences in reporting structure, geocoding precision, and outbreak definition substantially influence spatiotemporal cluster outputs. Rigorous data quality assessment, including harmonisation, deduplication, and spatial validation, is therefore essential before conducting integrated analyses or translating results into policy recommendations.

Our findings highlight the need for greater standardisation of outbreak definitions, date attribution practices, and geospatial reporting across surveillance systems. Without such standardisation, analytical outputs may reflect structural artefacts rather than epidemiological reality. By incorporating sensitivity analysis and persistence assessment, more robust hotspot identification can be achieved, supporting evidence-based surveillance prioritisation.

Overall, these results reinforce the importance of transparent, integration surveillance frameworks to enable reliable One Health data integration. Strengthening interoperability and metadata documentation will enhance the utility of integrated surveillance systems in Southeast Asia.

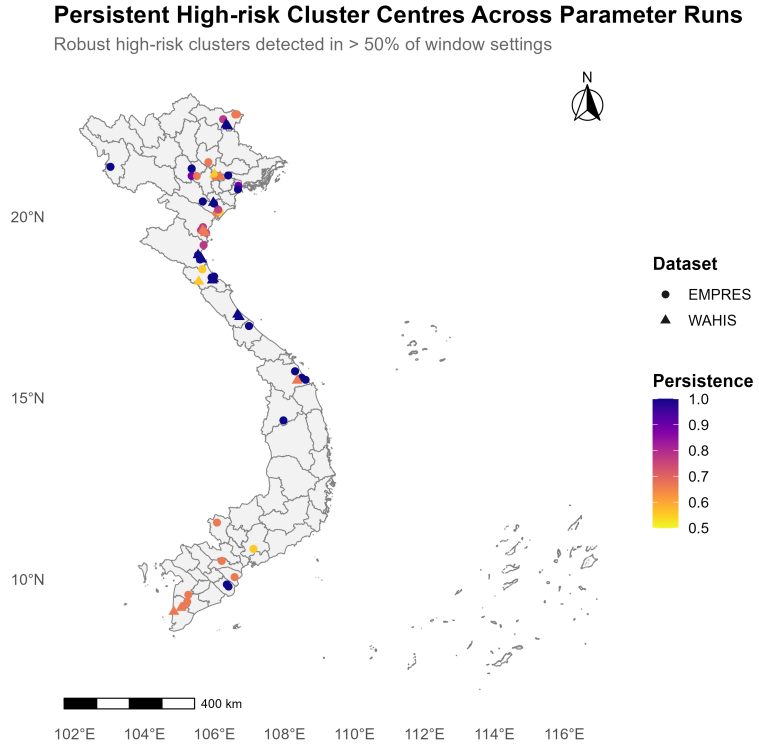


Figure 1: Spatial distribution of persistent space–time cluster centres identified across seven parameter settings in a comparative analysis of HPAI surveillance datasets in Vietnam, 2006–2022

Keywords: Panzootics, Surveillance integration, Data curation, Space-time clustering, Vietnam.

References:

1. Artois, J., Newman, S. H., Dhingra, M. S., Chaiban, C., Linard, C., Cattoli, G., Monne, I., Fusaro, A., Xenarios, I. & Engler, R. 2016. Clade-Level Spatial Modelling Of Hpai H5n1 Dynamics In The Mekong Region Reveals New Patterns And Associations With Agro-Ecological Factors. Scientific Reports, 6, 30316.
2. Gutiérrez, R. A., Naughtin, M. J., Horm, S. V., San, S. & Buchy, P. 2009. A(H5n1) Virus Evolution In South East Asia. Viruses, 1, 335-61.
3. Jindal, M., Stone, H., Lim, S. & Macintyre, C. R. 2025. A Geospatial Perspective Toward The Role Of Wild Bird Migrations And Global Poultry Trade In The Spread Of Highly Pathogenic Avian Influenza H5n1. Geohealth, 9, E2024gh001296.
4. Kulldorff, M., Heffernan, R., Hartman, J., Assunçao, R. & Mostashari, F. 2005. A Space–Time Permutation Scan Statistic For Disease Outbreak Detection. Plos Medicine, 2, E59.
5. Loth, L., Pham, L. T. & Stevenson, M. A. 2021. Spatio-Temporal Distribution Of Outbreaks Of Highly Pathogenic Avian Influenza Virus Subtype H5n1 In Vietnam, 2015–2018. Transboundary And Emerging Diseases, 68, 13-20.