2023 IEEE 16th Pacific Visualization Symposium (PacificVis) | 979-8-3503-2124-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/PacificVis56936.2023.00015

Investigating Animal Infectious Diseases with Visual Analytics

Yun-Hsin Kuo[*]
University of California, Davis

Beatriz Martínez-López [*]
University of California, Davis

Kwan-Liu Ma[*]
University of California, Davis

Abstract

Animal infectious diseases interfere with the sustainability of livestock farming. Developing comprehensive strategies for disease prevention and control requires professionals to study livestock farms from a variety of data sources, such as veterinary medical tests, financial reports, and animal movements between farms. However, investigating animal health surveillance is challenging as the collected data is often heterogeneous, high-dimensional, and spatio-temporal. Furthermore, data missingness, one common challenge in disease surveillance, can limit the effectiveness of the analysis and induce the misinterpretation of the result due to the lack of uncertainty representation. In this paper, we present a visual analytics interface of coordinated views that supports investigating disease outbreaks by connecting the relationships of livestock farms from different aspects — geospatial, transactional, and financial. Coupled with unsupervised machine learning methods, we infer the health status of a farm, despite the absence of its diagnostic history, with uncertainty and provide interpretability to such inferences. With these functionalities, we further quantify the influence of a disease outbreak, severity and scale, guiding the user toward investigating important outbreaks. We demonstrate the analysis capability of our visual analytics interface with multiple use cases on a real-world swine production dataset.

Keywords: animal health, disease surveillance, visual analytics, machine learning

1 Introduction

Sustainable livestock farming shows promise to steady meat supply to a growing world while maintaining the environmental health. However, the sustainability hinges on the maintenance of animal health, high production efficiency, and proactive management practices. Animal infectious diseases can directly endanger animal health and indirectly affect production efficiency. Furthermore, poor management practices may lead to higher infection rates across locations over time. Understanding how animal infectious diseases interfere with the sustainability assists animal health specialists in developing comprehensive strategies for disease prevention and control. Yet, investigating the aforementioned factors requires the specialists to analyze livestock data collected from a variety of sources, such as veterinary medical tests, financial reports, or even news articles [9, 18]. While there are different analysis techniques specialized in processing certain data types [7, 35], challenges linger for integrating analysis as the data is often heterogeneous, high-dimensional, and spatio-temporal. These data characteristics impose the burden of consuming excessive information for domain experts, thus presenting challenges to disease prevention and control.

Effective management tools and analytical methods are therefore needed to support animal health surveillance. Perez et al. [34] described requirements for a system to support animal disease surveillance, including visualization design and analysis functionality. While some existing systems provide comprehensive information of infectious diseases at a global scale for epidemic control [14], there are some that focus on close monitoring by providing in-depth analysis tasks such as outbreak detection and molecular epidemiology [9, 40]. The commonality of these systems often includes geographical information support with statistical methods. Machine learning models further support such spatial analysis of heterogeneous data with growing volume [35].

Interactive visualizations assist the user in large data exploration. When coupled with analysis methods, visual analysis further facilitates the understanding of complex data. Carroll et al. [7] surveyed how visualization tools support disease surveillance with analytical methods and integrate various data sources. However, in practice, it is common to find data incompleteness, such as the lack of laboratory diagnostic history in animal farms due to high testing costs. Missing data can limit the analysis capability and may cause the user to misinterpret the result [7, 35]. It is thus critical to address how to represent missing data and uncertainty in visualizations such that valid analyses can still be carried out with limited information.

In this work, we present a visual interface that visually coordinates the heterogeneous information to support disease outbreak investigations. Our interface facilitates effective analysis via a methodology that couples interactive visualization components with machine learning methods — dimensionality reduction (DR) and contrastive learning. We utilize an existing feature learning framework for DR, FEALM [15], to extract the farms' various relationships from the high-dimensional financial reports and to further provide interpretability to the resulting visual summary via contrastive learning. This visual summary enables us to infer a farm's health status, despite the absence of one's diagnostic history, which leads to the capability of computing uncertainty. With this methodology, we investigate how disease outbreaks affect the environment over space and time by tracing animal movement. The health status inference with uncertainty allows us to measure the potential influence of disease outbreaks, supporting the user in locating their analysis target efficiently. Our visual analysis is capable of assisting disease prevention and control by revealing potentially affected farms in close proximity, risky animal movements between farms, and farms with similar management practices.

We consider our main contributions are: (1) introducing a methodology that employs an existing feature learning framework for dimensionality reduction to infer a farm's health status with uncertainty and interpretability; (2) prototyping a visual analytics interface that uncovers the relationship of farms from geospatial, transactional, and financial perspectives to investigate disease outbreaks; (3) demonstrating the analysis capability of the visual analytics interface with multiple use cases on a real-world pork production dataset.

2 Related Work

We discuss representative research on related topics. As there is little work on visual analytics for animal health surveillance, to our best knowledge, we include work on human disease surveillance and epidemiology. Visual analytics for public health often encounter similar challenges incurred by the data characteristics as we do [37]; therefore, we review relevant analytical or visualization techniques in the ensuing subsections.

2.1 Visual Analytics for Health Surveillance

EpidNews and EpidVis have analyzed news articles to support animal health surveillance [12, 19]. They utilized coordinated visualizations to facilitate spatiotemporal exploration of news items or developed visual queries to capture disease characteristics from news

[*]e-mail: {yskuo, beamartinezlopez, klma}@ucdavis.edu.

2165-8773/23/$31.00 ©2023 IEEE
DOI 10.1109/PacificVis56936.2023.00015

articles. With the assistance of sunburst and chord diagrams, they studied the hierarchical relationships between diseases, hosts, and symptoms. In addition, while LHAVA [29] integrated animal health data into their visual analysis, their goal was to support zoonotic disease analysis for human health.

In contrast, there had been thorough discussions on the requirements, tasks, and visual analytic techniques to support human disease surveillance [7, 37], including what techniques were employed to handle certain data types. For instance, geographical information support is often a vital analysis component for studying the connection between environmental factors and human health. While some visual analytic solutions focus on disease outbreak detection [29], disease spread simulation [44], or human response monitoring [28], our work focuses on the understanding of disease outbreaks, such as identifying the associations between heterogeneous spatio-temporal information. For example, to characterize groups of interest (or subpopulations), association rules [22] and subspace clustering [3] have been employed to study the commonality of a group. We utilize an existing DR framework [16] that combines contrastive learning to learn about the characteristics of a group of farms in terms of their production and financial information.

According to the relevant surveys [7, 37], many visual analytic solutions designed coordinated views in a web-based application to assist health specialists. We also adopt this architecture in our system for efficient integration of heterogeneous data. We further review two visual analysis techniques regarding disease outbreak assessment that are most relevant to our work.

2.2 Probabilistic Infection Inferences

Veterinary diagnostic tests, similar to human diagnostic tests, directly assesses the health status of an animal or group of animals and are usually a reliable indicator for the presence or absence of a disease; thus, they have become the foundation to establish the health status of an animal farm. In practice, however, it is common to find the lack of diagnostic history for several livestock animal diseases at farm level, due to high testing costs. This leads to the unavailability of determining the health status of the farm and the possibility of inferring the health status from other available information. Here, we report what other kinds of data or techniques were utilized to infer the health status.

Molecular epidemiology is one approach that studies the spatial and/or temporal distribution of genetic variants or pathogens in phylogenetic trees or dendrograms [7]. As the hierarchical relationships suggest the similarity between all the genotypic information, the health specialists can investigate different branches to determine if one node (e.g., a farm on a date) is affected by a disease source, represented by another node. Another approach is to study the networks of farms in terms of the disease exposures. Through hundreds of disease spread simulations, prior work aggregated the simulations' results and computed the probability of a patient being infected at a given time point [28, 43].

Recent work have leveraged machine learning to quantitatively predict health status over time. Supervised classifiers trained on temporal multivariate data are one of the popular techniques, surveyed by [24]. However, the excessive data dimensions have presented challenges to over-fitting problems in supervised models [13], leading to the increasing adoption of unsupervised approaches. For instance, DPVis [25] utilized Hidden Markov models (HMMs) to infer the progression of a patient's disease status, represented by different discrete states, from time-varying multivariate patient data. They characterized each state with the distribution of date attributes to provide interpretability, where uncertainty is determined by the standard deviation of an attribute. ThreadStates [41] further employed DR methods to support state identifications in HMMs and revealed disease progressions using Sankey-based visualizations. However, the user is required to determine the number of states and the set of attributes being used in HMMs.

Our work utilizes DR to infer the health status, unlike existing work which commonly uses DR as a feature engineering technique [13]. Given a farm with no diagnostic history, we support the automatic identification of its closest group (diseased or healthy farms), which is one representative pattern identification task in DR analysis [11]. We further compute uncertainty in the identification task, in which we use a probability to indicate the confidence of a farm being diseased.

2.3 Disease Propagation Visualizations

In livestock production, particularly in the swine industry, a community of highly specialized farms establishes a production system and frequently transport animals among each other. Livestock movements between farms thus represent one of the main pathways for infectious disease transmission [8]. Researchers often performed analysis on animal movements to evaluate their implications for disease transmission, where a dynamic geospatial network has been a common visual representation of a disease outbreak for exploring the data and communicating analysis results [27, 30].

We review visualization techniques for geospatial network visualizations, where a comprehensive survey can be found in [38]. While popular designs include node-link diagrams overlayed on the geospatial map and flowmaps [20], recent work have introduced novel visualizations to highlight other information in the disease spread. Employing the storyline visualization, Baumgart et al. [4] reconstructed the disease transmission and presented the patient trajectories, with a focus on individual movement to locate the potential disease source. To reduce the visual complexity caused by the varying geospatial region sizes and shapes, Dunne et al. [10] introduced three visual representations for geospatial maps, including a centroidal Voronoi tesellation technique.

We seek visualization designs that clearly show the disease spread over space and time. Therefore, we tailor node-link diagrams by adopting an abstract layout spanned by locations and time, which is free of visual clutter caused by geographical proximity. While prior work in biomedical visual analytics introduced an abstract network visualization to resolve visual clutter [33], their focus was to address the hierarchy in human brains rather than the time aspect. We integrate a Sankey-based visualization, focusing on the influence of animal movements, as an alternative representation of the disease spread. A geospatial map is incorporated to supplement the geographical information.

3 Data and Design Goals

We developed our visual interface through seven monthly meetings with three animal health specialists, including two faculty specialized in disease epidemiology and surveillance and one field veterinarian. In this section, we first describe the data used to drive our analysis and visualization design. Then, we present the design goals of our visual analytic solution, which were derived through an extensive discussion with the animal health specialists.

3.1 Data

We obtained a swine production dataset that describes a community of animal farms belonging to the same production system. The data was collected in the United States, through January 1st, 2020 to December, 31st, 2020. For confidentiality, we have anonymized the data and cannot disclose any identifiable information, such as the number of farms in the production system. In the following, the described data were already cleaned and processed by the authors.

Sites – General information of a list of more than 100 unique animal farm sites. Each entry refers to a site and records its name, premise ID, GPS coordinates, and its site type. The site type indicates the site's role in the livestock production system and includes three types: sow farms, finishing farms, and food processing companies. The production system consists of 8.5% sow farms, 83.0%

finishing farms, and 8.5% processing companies. In particular, only food processing companies do not have premise ID and GPS coordinates; thus, they are only present in the transaction record.

Diagnostic history – The diagnostic test results for diseases or antimicrobial resistance on a site. Each entry records the premise ID of the tested site, the date when the sample is received, the type of disease or antimicrobial, and the test outcome. Note that the test outcome can be "undetermined", besides "positive" or "negative". We consider the tests with undetermined outcome to be ineffective and exclude them. This results in 246 effective entries, where 36 of them show positive outcomes. Among all the unique sites, only 28.4%, composed of 6.4% sow farms and 22.0% finishing farms, have had at least one effective test results during the studied time span. The rest have no record.

Animal movement – The transaction records of animals between two sites, containing 10,337 entries in total. Each entry involves the name of the sending site (i.e., source), the name of the receiving site (i.e., target), the transaction date, the quantity of transported animals, and the description of the trade purpose.

Production and financial report – The production and financial statements of the finishing farms, related to aspects such as mortality, feeding, storage, or sales. While we use all of the 32 numerical attributes in our analysis, each statement also records the time range and the working group. As each farm may have several working groups active in different varying time periods (can be 2 weeks, 1 month, or 3 months per group; 1 month is the most common period), we further take the average information across groups and time as a farm's representative profile, in order to perform analysis with DR. We incorporate the original information in other parts of our analysis, to be illustrated in Sect. 4.2. Note that among finishing farms, 6% do not have any reports, thus only 94% of farms' data is used in our analysis.

Antibiotic history – The antibiotic usage history of finishing farms, consisting of 776 records. Each entry records the site name, the administered date, the product name, and the given dosage.

3.2 Design Goals

Our visual analytics interface supports experts, such as swine producers and field veterinarians, in analyzing disease outbreaks. The tasks of disease outbreak investigations involve going over large amounts of heterogeneous information. Through an extensive discussion with the animal health specialists , we have derived the following specific design goals to support essential analysis tasks.

DG1: Examining disease spread through animal movement. The transportation of animals is one major interaction between livestock farms. Infectious diseases may spread through the direct contact of animals [8, 27]. When a disease outbreak is identified, we should show the potential impact of an outbreak over time and space by tracing the animal movement. Furthermore, we do not assume an outbreak is necessarily the disease source; therefore, we should trace both retrospectively and prospectively.

DG2: Summarizing high-dimensional information. It is time consuming to review the high-dimensional attributes one by one. While visualizations of high-dimensional data, such as parallel coordinates and scatterplot matrices, can accurately depict the data values, their effectiveness is limited by the scalability of data dimensions. We should provide a visual summary of the high-dimensional information that is agnostic to the number of dimensions.

DG3: Inferring the health status of a site. The absence of diagnostic results can limit the capability of disease surveillance analysis. As described in Sect. 3.1, a majority of the animal farms do not have diagnostic test results; therefore, the lack of diagnostic labels induces difficulty in training an unbiased machine learning model for health status prediction. We should seek other information to facilitate health status inference. The user should also be able to validate the inferences; thus, our interface ought to provide interpretations of inferences. Uncertainty should be considered to avoid misinterpretation as much as possible.

DG4: Measuring the influence of disease outbreaks. While our system should provide comprehensive analysis support to disease outbreak investigations, the user may be overwhelmed by the number of outbreaks that need to be analyzed, especially when the user is exploring the data rather than targeting a specific site, disease, or time period. We should develop metrics to quantify the potential impact of a disease outbreak, which allows the users to locate the analysis target efficiently.

DG5: Integrating analyses. We support the computational analyses of heterogeneous data to extract important information for investigating disease outbreaks. Our system should provide and coordinate interactive visualization support for the users to perform flexible analysis and derive insights, such as risky animal movements highlighted by the health status inferences.

4 Methodology

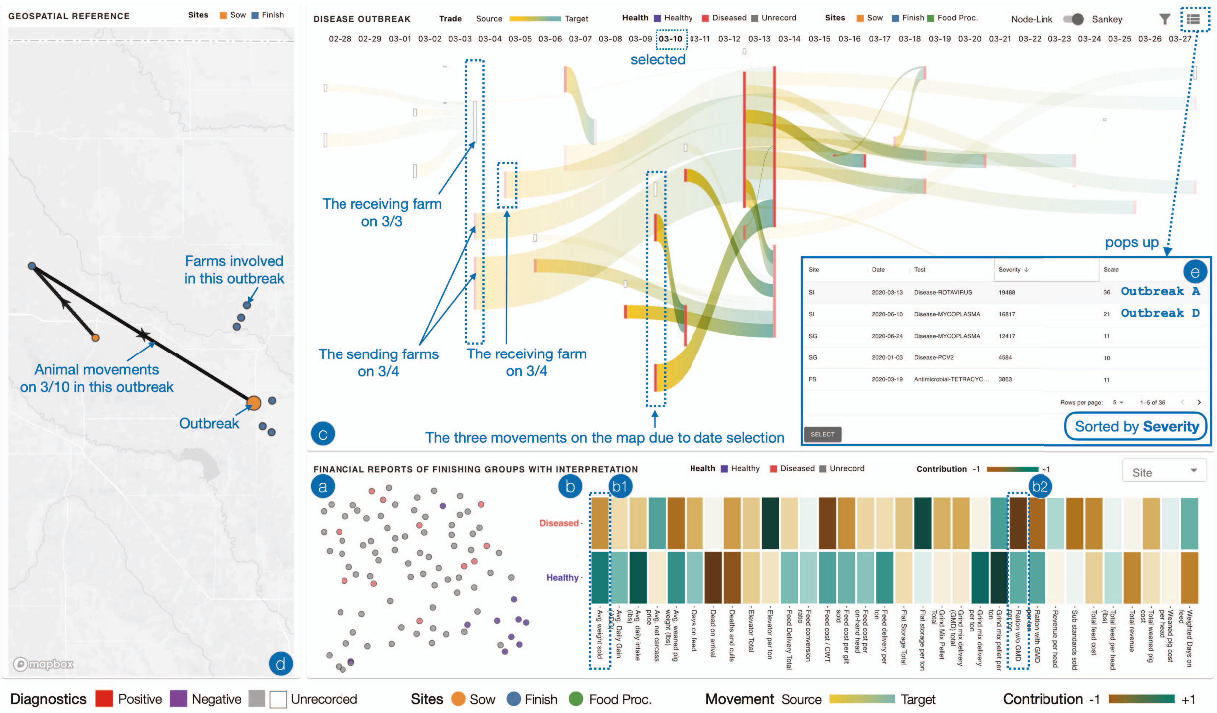
To address these **DGs**, particularly **DG1** and **DG2**, our approach is to design individual visual components and coordinate them to fully support essential analysis tasks. We extend the analytical solution of **DG2** to support **DG3**, leading to the feasibility of achieving **DG4**. To better assist the user in exploring the heterogeneous data (**DG5**), as shown in Fig. 1, we coordinate multiple visual components in a visual analytics system, including (a) visual summary of finishing farms; (b) finishing farm group characteristics; (c) disease outbreak overview; (d) geospatial map; and (e) disease outbreak selection menu. During development, our interface and visual components were reviewed by the animal health specialists in multiple monthly meetings, where visualization researchers on our team demonstrated how to perform the analysis through user interactions. We provide a video to demonstrate the functionalities of our system [2].

Implementation. Our system is a web-based application for its accessibility, available at [1]. For the back-end, we use Python to perform all the computations and integrate FEALM, with the implementation the original authors provided [15]. For the front-end interface, we use a combination of HTML5, Javascript, React, and D3 [5].

4.1 Visual Summary of Finishing Farms

We start with the visual summary of finishing farms, the most important component in the visual interface, as it drives the analysis supported by our methodology. We illustrate a complete analysis flow with a case study in Sect. 5. As mentioned in Sect. 3.1, the financial reports of the finishing farms contain high-dimensional attributes, which induces difficulty in performing analysis and understanding the subsequent results. Analyzing these financial and production reports is beneficial, as the disease outbreaks may be associated with management practices of a farm; therefore, the disease influence can reflect on certain data attributes such as mortality. With fulfilling **DG2** in mind, we seek an analytical solution to extract important information from all available data attributes.

Dimensionality reduction (DR) finds low-dimensional, often 2D, representations of entities with high-dimensional data. These low-dimensional representations are commonly visualized as points on a scatter plot, where entities with similar attribute values are in close proximity. While a variety of DR methods have been developed with a different focus on preserving certain characteristics (e.g., data variance), they can be considered as either linear or nonlinear methods based on their algorithm design. Generally, the result interpretation of linear DR methods is more accessible to humans due to linearity; in contrast, nonlinear ones are favored for their capability of handling more complex data structure. However, nonlinear DR methods may fail to capture important patterns that are apparent only in a subset of data attributes.



Site	Date	Test	Severity ↓	Scale
SI	2020-03-13	Disease-ROTAVIRUS	19488	36 Outbreak A
SI	2020-06-10	Disease-MYCOPLASMA	16817	21 Outbreak D
SG	2020-06-24	Disease-MYCOPLASMA	12417	11
SG	2020-01-03	Disease-PCV2	4584	10
FS	2020-03-19	Antimicrobial-TETRACYC...	3863	11

Rows per page: 5 ▾ 1–5 of 36 ‹ ›

SELECT

Sorted by Severity

Figure 1: The interface of our visual analytics system, consisting of the following visual components: (a) a visual summary that shows the similarity among finishing farms derived from their financial reports; (b) a heatmap that displays the unique characteristics of the diseased farms and the healthy farms in terms of the data attributes from the financial reports; (c) a disease outbreak overview presenting the outbreak in a Sankey-based visualization, including the relevant animal movements and the health status inferences, where a toggle is provided to switch to the alternative representation, the node-link diagram; (d) a geospatial map that reveals the farms in geographical proximity and shows the animal movements on a certain date with arrows; and (e) a disease outbreak selection pop-up menu, currently in descending order of the Severity metric, displaying general information of the disease outbreaks. Note that the legends at the bottom are annotated by the authors for visibility. In addition, (b1) and (b2) are two attributes related to the inference example from Fig. 3.

FEALM is a feature learning framework that addresses this challenge by discovering latent features of data to generate significantly different DR results [15]. Fig. 2 shows the 2D representations of finishing farms with different DR methods being applied to their financial reports, where we aggregate their diagnostic history into the labels and color the points accordingly. In particular, Fig. 2-(c) and (d) are representative results explored by FEALM-UMAP, i.e., UMAP [31] exemplified using FEALM. These representative results reveal the separation between the healthy farms and the diseased farms, suggesting that it is not an oversubtle pattern. Note that for the diagnostic labels, we label a farm to be **diseased** (positive) if it has ever had any positive diagnostic results; otherwise, it is considered **healthy** (negative). Among the DR results explored by FEALM or computed with other DR methods, we seek the one with clearer separation, where spatially dense groups are preferable for showing high similarity. Utilizing this visual summary, if we find a farm with an unknown status close to a group of healthy (or diseased) farms, we may assume they have similar data records; thus, the farm is inferred likely healthy (or diseased). For example, in Fig. 3-(a), since Farm FK is next to two red points (i.e., farms that had been tested positive before), we may infer the report of Farm FK is similar to the two diseased farms. This facilitates the health inference of farms with an unknown status, to be further elaborated in the ensuing subsections.

It is important to note that, linear discriminant analysis [23] could be a good candidate for the analytical solution, since it finds the weights in the linear combination of data attributes to distinguish between diseased and healthy farms. However, it is a supervised DR method taking the diagnostic labels of the finishing farms as input. As we separate between the diseased and healthy farms (each of which is less than 15) using all of the 32 numerical attributes, the number of model coefficients (i.e., weights) is larger than the number of training instances. The result then becomes untrustworthy as now we have concerns for over-fitting problems. This issue applies to other supervised classifiers as well; thus, we do not solve the health status inference as a classification problem. Last, we do not consider the combination of subspace clustering and DR techniques for achieving **DG2**, since FEALM is designed for nonlinear DR techniques and its search space includes that of subspace clustering.

4.2 Health Inference with Uncertainty and Interpretability

In this subsection, we describe how we compute the health inference of a farm with an unknown status from the aforementioned visual summary and provide explanations (**DG3**). Among all the finishing farms, while 22.0% have had positive or negative test results, the rest had none. This limits the analysis capability to investigate disease outbreaks, since the user can barely perform an analysis due to little available information. Continued from the example in Fig. 3-(a), the visual summary enables us to infer that Farm FK is more closer to the diseased farms than the healthy farms; therefore, Farm FK is likely positive. This capability connects the relationships among animal farms from financial and medical perspectives, and provides implications to the user that were not available before. However, each inference requires human judgment and may vary in its susceptibility to human perceptions [11]. For instance, one may

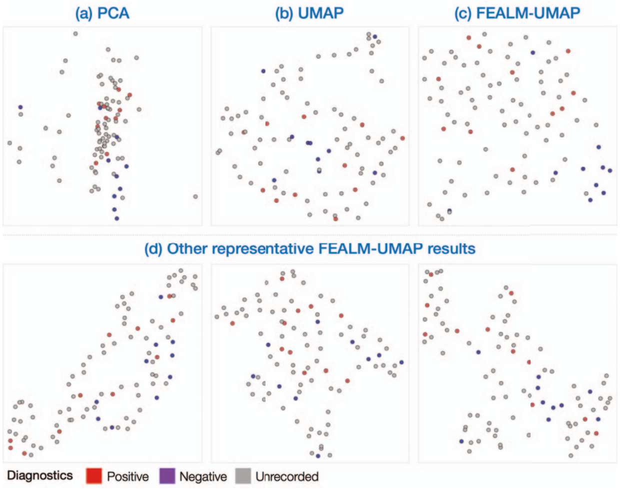


Figure 2: Comparison of the visual summaries, generated with different DR methods. (a) was computed with principal component analysis (PCA), a linear DR method; (b) was computed with UMAP, a nonlinear DR method; (c) was one representative result explored by FEALM-UMAP, adopted into our visual analytics system, that shows a clearer separation between the diseased and healthy farms when compared to (a) and (b); and (d) FEALM-UMAP reveals the separation between the diseased and health farms, despite not being clear enough as (c), in other representative results during the same search. This suggests that the separation is not an oversubtle pattern. Note that the hyperparameters used in UMAP and FEALM-UMAP are the same, i.e., min_dist = 0.1 and n_neighbors = 15.

argue that `Farm FK` is fairly close to the community of the healthy farms, and thus is possible to have a healthy status. More precisely, this ambiguity comes from the lack of a concrete decision boundary to determine the health status inference of a finishing farm.

We provide a systematic method that automatically computes the health status inferences with uncertainty from the visual summary. Given a point with unknown status (target point), we first compute its Euclidean distance to all the points labelled as either diseased or healthy. Then, we form a Gaussian probability distribution from the obtained distances, centered at 0 (i.e., the distance to the target point itself). Similar to determining neighbors in distance-based DR methods, we may interpret the resulting probability as the likelihood that the labelled point being in the neighborhood of the target point. Next, among all the labelled points, we select the top 20% of candidates with the highest probabilities for further computation. We consider this k-nearest neighbor selection to be a reasonable choice, since human judgement often considers only some nearest labelled points rather than all of them. We apply L1-normalization to the probabilities of the candidates to obtain their relative importance (summing up to 1). Finally, we sum up the relative importances of positive-labelled candidates, returned as the computed probabilistic label for the target point. For example, in Fig. 3-(a), as `Farm FK` is now given a value of 82.8%, we may interpret this as an 82.8% confidence that `Farm FK` is likely diseased. Note that when the user hovers over any farms in the visual summary, a tooltip will display the computed inference of the hovered farm.

In the absence of a farm's diagnostic history, we compute probabilistic inferences from its financial reports to complement the limited information. However, this capability derives the subsequent question — how can we trust and validate these inferences? As mentioned in the previous subsection, while the nonlinear DR methods are capable of handling more complex data structure, interpreting their DR results is hardly straightforward. Representative

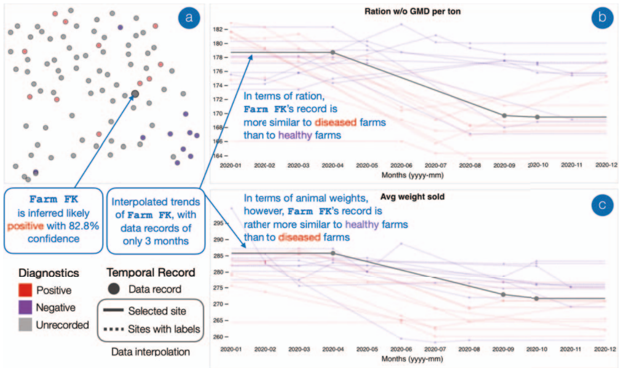


Figure 3: Understanding the health status inference of `Farm FK` with two interpretation examples. (a) in the visual summary, the inference is susceptible to human perception; however, our methodology computes the inference with uncertainty, i.e., `Farm FK` is inferred likely positive with 82.8% confidence; (b) the trends of "ration cost per ton" (Fig. 1-(b2)) suggests that `Farm FK` is more diseased-leaning; (c) the trends of "the average weight of sold pigs" (Fig. 1-(b1)) suggests that `Farm FK` is more healthy-leaning instead, which explains its positioning in the visual summary. Note that for each data attribute, the lines represent the interpolated records (i.e., the trends) of the farms, with the solidity indicating that `Farm FK` is currently selected by the user.

approaches include statistical descriptions of data attributes and axis mapping through user interactions [26]; however, they are not designed to reveal important data attributes that characterize a group of points.

FEALM integrates an existing DR method, ccPCA, that utilizes contrastive learning to study the uniqueness of a group of points in terms of data attributes [16]. By contrasting a group to others, ccPCA uses a number between -1 and +1 for each data attribute to indicate by how much is this group lower (or higher) than others in terms of their attribute values (-1 indicates lower while +1 suggests higher). We adopt this integration along with the heatmap-based visualization, as shown in Fig. 1-(b), to understand what attributes make the diseased farms more distinct than the healthy farms, and vice versa. We use a heatmap with a brown-green diverging colormap (-1 for brown and +1 for green) to display the characteristics per group, where each column refers to one data attribute and each row refers to one group of finishing farms. For example, in the leftmost column (Fig. 1-(b1)), "Average weight sold", the color for the diseased group is brown-leaning, while it is green for the healthy group. This suggests that in terms of the attribute values, the healthy group generally has higher data values than the diseased group does. The observation is reasonable since the illness can result in a lack of appetite and hence weight loss. Note that we provide the representative profiles of labelled farms to ccPCA (see Sect. 3.1).

The utilization of ccPCA provides us an overview on the unique characteristics of the healthy and diseased finishing farms; yet, one more step is required to help the users validate the computed inference of a farm. We provide the temporal data examination of the financial reports for the user to assess how a farm is similar to the healthy (or diseased) farms in terms of attribute values. The user can focus on a farm of interest by either clicking its point on the visual summary or selecting through the dropdown menu. When a farm is selected, we offer the validation as a tooltip when hovering a block on the heatmap, as shown in Fig. 3-(b) and (c). The tooltip is a multiline chart over time at month level, with respect to the hovered data attribute, where each line represents the trend of one finishing farm and is colored based on the corresponding diagnostic label. We show the trends of the labelled farms with dashed lines, where the solidity is used to highlight the selected farm. For each farm, to regularize the irregular time intervals, as described in Sect. 3.1, we first mark the report using a single month, which is in the middle of the time period. If there are multiple groups active in the same month, we aggregate the data values by taking their average. We observe that a majority of the finishing farms have monthly records for less than 6 months (e.g., `Farm FK` from Fig. 3-(b) and (c)); consequently, there are countless line segments depicting the same farm. This leads to the visual clutter and the difficulty of following a farm's temporal record. To address these issues, we employ interpolation to connect the line segments and incorporate points to indicate the presence of the actual data records. It is important to note that the purpose of the interpolation is to reduce visual clutter rather than to fill in missing data for performing analysis, since the interpolated records can be biased due to the sparsity of the time points. In addition, we are aware that the reliability of the inferences includes the inherent uncertainty in DR projections, i.e., how accurately the low-dimensional representations depict the data in the high-dimensional space, which we provide a discussion in Sect. 7.

Returning to the example of `Farm FK` being inferred to be positive with 82.8% confidence, we illustrate how we provide the validation with ccPCA and the interpolated temporal record. In the group characteristic view, as highlighted in Fig. 1-(b1) and (b2), we find that for both attributes, there are diverging colors in the blocks, where the diseased farms have a noticeably darker brown color in (b2). This finding suggests that the attribute values of the disease farms are generally lower than those of the healthy group; thus, we expect a larger gap between two groups in (b2), compared to (b1). Since (b2) is annotated as "Ration cost per ton", a possible explanation is that the ration quality (e.g., nutrition) is associated with animal health, which is also reflected in the purchase price. By selecting `Farm FK` in the visual summary or through the rightmost dropdown menu in Fig. 1-(b), we use a solid line to highlight the farm's trend in the tooltip. As shown in Fig. 3-(c), we find the difference between the diseased farms and the healthy farms and that the trend of `Farm FK` is more diseased-leaning. In contrast, in Fig. 3-(d), we can hardly decide which group `Farm FK` leans more toward. This explains `Farm FK`'s positioning in the visual summary and the health status inference of 82.81% confidence. We provide more interpretation examples in Sect. 5.

4.3 Disease Spread through Animal Movements

By analyzing animal movements together with the health status inferences, we extend our methodology to include the relationship among farms from the transactional perspective. Next, we describe how we design our disease spread visualizations to support **DG1**.

We identify each unique positive result from the 246 entries of diagnostic history as one disease outbreak, resulting in 36 outbreaks. Each disease outbreak is characterized by its location (i.e., a farm), the date, and the type of disease or antimicrobial resistance. Using this collection of information, we capture the potential disease spread by tracing through animal movements recursively and recording the farms that could be exposed to the disease either directly or indirectly. More precisely, we use a set of nodes and edges to represent the disease spread of the outbreak. While each edge is one animal movement, as described in Sect. 3.1, each node is characterized by a farm's premise ID and a date, extracted from the animal movement that we visit from. For each node, we incorporate the time-balanced health status inference to include the temporal consideration. Since we do not assume the outbreak is necessarily the disease source, we track both prospectively and retrospectively. The animal health specialists suggest two weeks to be a suitable time period of effective infection; thus, we collect relevant animal movements within two weeks for both tracking directions.

We use the retrospective tracking as an example to illustrate the complete process. Starting with the outbreak farm, we collect all of the movements involving itself as the receiving farm (target) within the past two weeks. For each movement, we create a node of the sending farm (source) along with the trade date. For every node, we compute the health status inference of the associated farm, and scale down the confidence every 3 days in the duration of the associated date and the outbreak date. If the node has never been visited before, we push it to a queue for recursive tracking. This process repeats until the queue becomes empty or a stopping criteria is met. We record all of the visited animal movements and nodes to return as the disease spread for a given outbreak. Besides the search period of two weeks, we consider another stopping criteria, the recursion depth, which is set to 2 to prioritize analyzing close farm contacts. Currently, the user cannot interactively adjust the two stopping criteria from the visual interface; however, the criteria can be updated from the back-end of the system.

We develop two visualization views, a node-link diagram and a Sankey-based visualization, to present the disease spread of an outbreak with different analysis focuses. Our first design, the node-link diagram, focuses on the disease spread over space and time. As shown in Fig. 4-(a), the diagram layout is spanned by the unique dates (*x*-axis) and the unique sites (*y*-axis). Based on the associated sites and dates, the nodes are positioned as circular points in the diagram, where their health status inferences determine the point color. We leave the circular point hollow for unavailable inferences. When hovering over a point, a tooltip displays the associated site name, date, and balanced health status inference, where the same site in the visual summary will be highlighted accordingly, if applicable. For the edges, we visualize the animal movements as the links to show how the disease spreads among the points. We overlay a yellow-green gradient on the links to indicate which node is the sender (indicated by yellow) or the receiver (suggested by green) in each movement. It is important to note that, identifying the date of an animal movement in our visualizations is heavily related to whether the movement is retrospective or prospective, due to the aforementioned tracking process. For retrospective movements, the animal movements are associated with the date in the source node; in contrast, the prospective ones are related to the date in the target node. Two interpretation examples can be found in Fig. 4-(a). We have attempted using directional pointers instead; however, this design choice resulted in visual clutter, and the user can hardly identify which node is the sender or the receiver in a movement at first glance. To reduce visual clutter and highlight critical animal movements, we add opacity to the nodes and the links, which is determined by the health status inferences, if available; thus, the nodes and the links associated with inferences of low confidence are more transparent. Furthermore, we provide a filter on the health status inference for the user to interactively highlight the nodes and the links associated with highly confident inferences, as shown in Fig. 4-(b). In addition, as annotated in Fig. 6-(d), we display the antibiotic history of farms as the cross icons to indicate what farms administered antibiotics on which date, with the details of the antibiotics provided with a tooltip.

Our alternative design, the Sankey-based visualization, focuses on influential animal movements in terms of the number of transferred animals. As shown in Fig. 1-(a), similar to the node-link diagram, we display the unique dates horizontally as the *x*-axis. While each of the bands refer to the individual animal movement, each of the blocks represents a site rather than a node (which is further characterized by a date). More precisely, instead of aligning a column of blocks to indicate the sites on a specific date, we utilize the between-columns to refer to a date. For each date , the column of blocks on the left of its label represents the sites as the source in animal movements; similarly, the column on the right describes the sites as the target in animal movements. For example, as annotated in Fig. 1-(a), on March 4th, we find three blocks on the left of the date label and one block on its right. The blocks represent either the receiving farms (target) on March 3rd or the sending farms (source) on March

4th, depending on their associated animal movements, while the receiving farm on March 4th refers to the block on the right of the date label. We adopt the aforementioned visual encodings and user interactions in this design, excluding the antibiotics history. We also encode the bandwidth with the number of transported animals per movement to highlight influential animal movements, e.g., large animal movements involving farms with highly confident health status inferences. However, compared to the node-link diagram, the user can hardly track a specific farm over time due to the optimized layout; thus, we adopt the node-link diagram as the default view and provide a toggle for the user to switch between the two visual representations.

We have described how we develop visualizations of the disease outbreaks with the abstract layouts. Since some infectious diseases can spread via airborne transmission, it is crucial to provide the spatial context for performing analysis. We incorporate the geospatial map to provide the user a geographical understanding of the disease outbreaks. When a disease outbreak is selected, as shown in Fig. 1-(d), we position all the involved farms as points, which are colored based on site types, with the outbreak farm having a larger point size. The map also displays other outbreaks for the same disease as icons, as shown in Fig. 6-(b). Upon clicking on a specific date label in the disease outbreak overview, we highlight the label in bold and show the movements associated with the date as arrows on the map, pointing from the sending farm to the receiving farm. For instance, the label of March 10th is selected in Fig. 1-(c); thus, three arrows are displayed on the map, each of which refer to each of the highlighted three blocks in the disease outbreak view. When hovering over the arrows, the points, or the icons, a tooltip provides relevant details. With these functionalities, the user may inspect the influence of the disease outbreak in terms of geospatial proximity.

4.4 Impact Quantification

Through these previous subsections, we have shown how the user can analyze one disease outbreak and relate to information incorporated from different aspects. However, it remains labor intensive if the user has to go through hundreds or thousands of disease outbreaks to gain a general understanding of the data.

Utilizing the nodes associated with the balanced health status inferences, we develop two metrics to quantify the potential impact of disease outbreaks. The first metric is **Severity**, that is the sum of transferred animal counts from the involved trades, with the animal count per trade being weighted by the corresponding health status inference. For a retrospective trade, we take the inference of the source node; otherwise, that of the target node. This metric allows us to quickly highlight outbreaks that potentially affect the most animals. Note that we exclude those movements with no inferences in computing Severity. The second metric is **Scale**, that is the number of visited nodes, regardless of the availability of their health status inferences. While the algorithm may count a site multiple times (i.e., the nodes of the same site on multiple dates), we choose not to consider the number of the unique sites, in order to highlight the outbreaks with higher disease exposure.

With Severity and Scale, we measure the potential impact of disease outbreaks (**DG4**). We provide a pop-up menu on the visual interface for the user to select a disease outbreak for analysis, as shown in Fig. 1-(e). This menu records the general information of an outbreak as follows — the site name, the date, the type of disease or antimicrobial resistance, the Severity score, and the Scale score. Sorting and filtering are provided for all of the columns, supporting the user to efficiently locate their analysis interest, such as outbreaks of the same disease at the same farm on different dates.

5 Case Studies

In the previous section, we have introduced the visual components and the user interactions in our system. We present three case studies on the real-world swine production dataset described in Sect. 3.1.

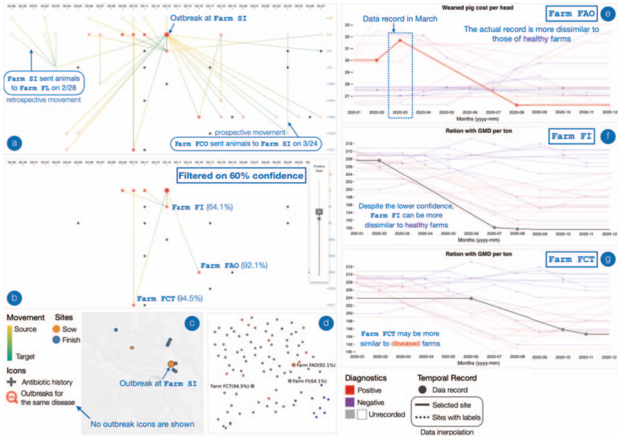


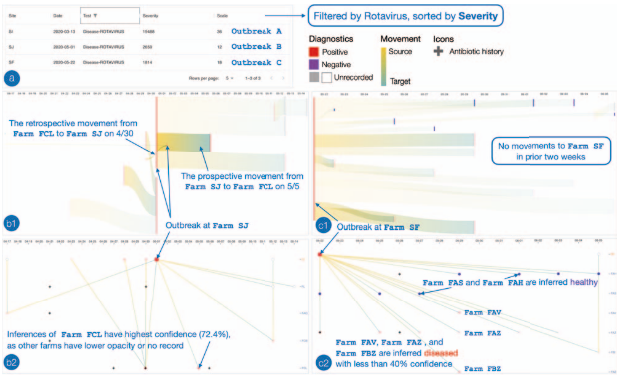
Figure 4: Revealing the risky movements through filtering and interpreting the health status inferences. (a) The node-link diagram of `Outbreak A`, happened at `Farm SI`; (b) The animal movements involve the farms with health status inferences of over 60% confidence; (c) The geospatial map shows no outbreak icons, indicating that no other Rotavirus outbreaks happened during the search; (d) The visual summary shows the three finishing farms with their health status inferences, annotated by the authors for space efficiency; (e) While `Farm FAO` is labelled positive, it has the actual record in March, providing better validation as the record is dissimilar to those of the healthy farms; (f) Despite the fact that `Farm FI` is inferred positive with 64% confidence, we find some of its records dissimilar to the healthy farms; and (g) Using the same attribute in (f), we find `Farm FCT` is diseased-leaning.

5.1 Case 1: Revealing risky movement

This case starts with the disease outbreak selection, dives into one outbreak guided by the Severity metric, and uncovers the potentially critical animal movements highlighted by the health status inferences.

As annotated in Fig. 1-(e), when sorting the outbreaks in a descending order of the Severity metric, we find a Rotavirus outbreak on `Farm SI` on March 13th, 2020 with the highest Severity score of 19488, implying the expected number of affected pigs in this outbreak (**DG4**). Upon selecting this outbreak (addressed as `Outbreak A` for simplicity), the disease outbreak view presents the relevant animal movements in a node-link diagram, as shown in Fig. 4-(a). In addition, Fig. 1-(c) shows the Sankey-based representation of the same outbreak. From the geospatial map Fig. 4-(c), the lack of the outbreak icons indicates that there was no other Rotavirus outbreaks happened in the investigated time period, suggesting the low possibility of infection from geographical proximity (**DG5**). While there are numeral animal movements shown to be potentially affected by `Outbreak A`, we can barely analyze some of them (e.g., the retrospective movement example in Fig. 1-(a)) due to the lack of the health status inferences (**DG1**). We can filter out movements involving farms with health status inferences of low or no confidence to highlight potentially critical movements (**DG5**). By setting the positive rate (i.e., confidence) to be no less than 60%, we consequently find 6 movements involving 4 animal farms, one sow farm (`Farm SI`) and three finishing farms (`Farm FI`, `Farm FCT`, and `Farm FAO`), which happened during March 9th and March 16th, as shown in Fig. 4-(b).

We can further inspect the computed health status inferences of these farms to understand more about the potentially critical movements. When hovering over the nodes of these three finishing farms in the disease outbreak view, we may find their corresponding point in the visual summary are also highlighted. We can further validate



Site	Date	Test	Severity	Scale	
SI	2020-03-13	Disease-ROTAVIRUS	19488	36	Outbreak A
SJ	2020-05-01	Disease-ROTAVIRUS	2659	12	Outbreak B
SF	2020-05-22	Disease-ROTAVIRUS	1814	18	Outbreak C

Rows per page: 5 1-3 of 3

Filtered by Rotavirus, sorted by Severity
Diagnostics
🟥 Positive
🟪 Negative
⬜ □ Unrecorded
Movement
Source
Target
Icons
✚ Antibiotic history
a
The retrospective movement from Farm FCL to Farm SJ on 4/30
The prospective movement from Farm SJ to Farm FCL on 5/5
Outbreak at Farm SJ
b1
No movements to Farm SF in prior two weeks
Outbreak at Farm SF
c1
Inferences of Farm FCL have highest confidence (72.4%), as other farms have lower opacity or no record
b2
Farm FAS and Farm FAH are inferred healthy
Farm FAV
Farm FAZ
Farm FAV, Farm FAZ , and Farm FBZ are inferred diseased with less than 40% confidence
Farm FBZ
c2

Figure 5: Comparing multiple Rotavirus outbreaks. (a) Through filtering and sorting, the outbreak selection menu shows three Rotavirus outbreaks, where Oubreak A is already explored in previous case; (b1) and (b2) presents Outbreak B in a Sankey-based visualization and a node-link diagram, respectively; and similarly, (c1) and (c2) are for Outbreak C. For Outbreak B, in (b2), we observe that Farm FCL may be crucial for being inferred with the highest confidence. For Outbreak C, in both (c1) and (c2), we see no movements to Farm SF prior to the outbreak. In (c2), we observe that while two farms are inferred healthy, the other three farms are inferred diseased but with low confidence.

the inferences of these three farms by selecting each of them in the visual summary (**DG2**, **DG3**), as shown in Fig. 4-(d). As annotated in the visual summary, Farm FAO has had positive diagnostic results, indicated by the red point color. However, since the movement between itself and Farm SI happened rather earlier prior to Outbreak A, our method balances the inference to be positive but with 92.1% confidence. In Fig. 4-(e), we can see that Farm FAO has the actual data record in March, which allows us to understand the inference better by examining the comparison with those of other labelled farms. As the data values of the healthy farms tend to be under $29 U.S. dollars, while those of the diseased farms can be more than $30, we may trust that Farm FAO is likely positive in March, 2020. On the other hand, despite their absent diagnostic history, Farm FI and Farm FCT are inferred positive with 64.1% and 94.5% confidence, respectively. When hovering over the same data attribute, i.e., the average cost of ration with delivery included, Fig. 4-(f) and (g) show that the records of Farm FI and Farm FCT are more similar to the diseased farms than the healthy ones, which provides support to high confidence behind the positive inferences. Note that for the space efficiency, we are only able to provide one interpretation for each finishing farm. The user can go through more data attributes to determine the trustworthiness of these inferences.

After validating that the inferences are reasonable, we conjecture that these critical movements together could affect the animal farms as follows: Prior to the outbreak on March 13th, the sow farm, Farm SI, sent animals to one finishing farm, Farm FCT, on March 10th and to another finishing farm, Farm FI, both on March 9th and 11st. Farm FCT sent animals to Farm SI on March 10th, while Farm FI transported animals to Farm SI on March 13th. Posterior to the outbreak, Farm SI sent animals to the other finishing farm, Farm FAO, on March 16th. We have shown that these three finishing farms are inferred to be likely positive as some of their financial records are similar to the diseased farms than the healthy farms. Thus, we understand better about Outbreak A that it may be associated with these critical animal movements as well as the three finishing farms.

5.2 Case 2: Comparing multiple outbreaks

In this case, we compare two Rotavirus outbreaks and investigate their differences. After the user analyzed a disease outbreak, they can explore other outbreaks through the disease selection menu to continue their analysis. Filtered by the keyword provided by the user, the menu displays three Rotavirus outbreaks (Fig. 5-(a)), where Outbreak A is already discussed in the previous case. One of the remaining two outbreaks (Outbreak B) happened at Farm SJ on May 1st, 2020, whereas the other (Outbreak C) happened at Farm SF on May 22nd, 2020. Compared to Outbreak A, we find both Outbreak B and C have significantly lower Severity and Scale scores. Meanwhile, in comparison to Outbreak B, the higher Scale score of Outbreak C suggests that it involves more potentially relevant animal movements while maintaining the lowest Severity score among the three (**DG4**).

To understand the differences between Outbreak B and C, we first start with the analysis of Outbreak B, which happened at Farm SJ on May 1st, 2020. As observed in Fig. 5-(b2), the finishing farm Farm FCL may have a strong association with Outbreak B as it is inferred likely positive with 72.4% confidence, while those other involved farms are with low confidence or have no record (**DG3**). The user can further perform the same analysis in the previous case to understand the inference of Farm FCL and determine if risky movements exist. When we analyze Outbreak B in the Sankey-based visualization, as shown in Fig. 5-(b1), the movement from Farm FCL to Farm SJ prior to the outbreak involves much fewer pigs; in contrast, the movements from Farm SJ tend to transport more pigs to destinations (**DG1**). This finding reflects one common management practice in the swine industry, the double stocking strategy. While the double stocking strategy can enhance production efficiency [42], the frequent deployment of such a strategy may increase the likelihood of pathogen transmissions when the animals are transported from different sources (farms).

We proceed to analyze Outbreak C, which occurred at Farm SF on May 22nd, 2020. As seen in both Fig. 5-(c1) and (c2), we find that there is no animal movement to Farm SF prior to the outbreak, implying that the disease source may be irrelevant to animal movement (**DG1**). Moreover, unlike Outbreak A and B, the farms involved in this outbreak are inferred either healthy or diseased but with less than 40% confidence (**DG3**). This finding suggests that Outbreak C may have little influence on other farms through animal movement (**DG5**). Compared to Outbreak A and B, we could speculate that Farm SF may have implemented effective prevention measures, such that the disease does not spread through animal movement.

Through this analysis flow, we have gained the understanding of Outbreak B and C as well as their differences. While additional analysis is required for validation, Outbreak B may be associated with one common management practice in the swine industry. Conversely, perhaps Farm SF in Outbreak C is equipped with effective infection control practices, since the disease source may be irrelevant to animal movements, and the prospective movements involve no farm with highly confident inferences.

5.3 Case 3: Identifying risky community of the farms

This case demonstrates a complete analysis flow for investigating an outbreak, studies the potential disease source by analyzing farms in geographical proximity, and understands the potential influence of the outbreak on other farms by associating the farms' antibiotic history with their health status inferences. In the first case, we investigated a Rotavirus outbreak (or Outbreak A) for its Severity score being the highest. We now focus on another outbreak with the second highest Severity score (**DG4**), as shown in Fig. 1-(e), which is a Mycoplasma outbreak (addressed as Outbreak D) at Farm SI on June 10th, 2020.

To gain a comprehensive understanding of Outbreak D, our first

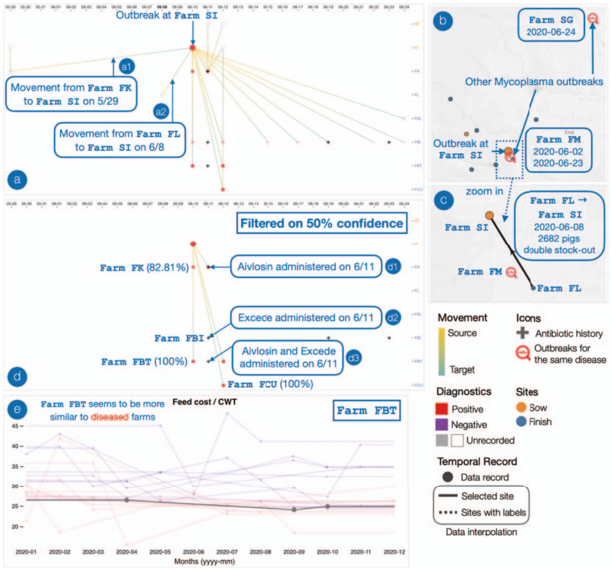


Figure 6: Investigate one Mycoplasma outbreak and identify a potentially risky community. (a) The source of Outbreak C may be relevant to two animal movements, (a1) and (a2); (b) The geospatial map discloses other three recent Mycoplasma outbreaks. Particularly, one outbreak at Farm FM is 8 days prior to the outbreak; (c) Upon selecting the date and zooming in, the movement (a2) takes place in the proximity of Farm FM; (d) Critical animal movements filtered on 50% confidence, with FBT and Farm FK inferred with over 80%. In addition, (d1), (d2), and (d3) highlights that the antibiotics were administered at three different finishing farms on June 11th, 2020; and (e) One interpretation example that reveals the interpolated trend of Farm FBT is diseased-leaning.

step is to look into what may be associated with the disease source of the outbreak. From the disease outbreak view (Fig. 6-(a)), we find that two animal movements, as annotated in Fig. 6-(a1) and (a2), may be relevant (**DG1**). The movement (a1) happened two weeks prior to Outbreak D and involved Farm FK that is inferred diseased with low confidence due to the time balance; in contrast, the movement (a2) from Farm FL happened 2 days prior to the outbreak, despite no health status inference due to the lack of its financial reports. The geospatial map reveals other three Mycoplasma outbreaks that happened in the proximity of Farm SI during the investigated time period, as indicated in Fig. 6-(b). While two of them happened posterior to Outbreak D, the other is detected at Farm FM on June 2nd, 2020, 6 days prior to the outbreak. We further examine if there is any relation between the two findings. By selecting the date of June 8th in the disease outbreak view, our system presents the animal movement (a2) in the geospatial map, as shown in Fig. 6-(c). We observe that while Farm FL in the movement (a2) has no health status inference, it is geographically close (4 miles or 6.48 kilometers) to the Mycoplasma outbreak at Farm FM (**DG5**). According to and Otake et al. [32], besides nose-to-nose contact among pigs, Mycoplasma can also be transmitted via aerosol over several miles. This suggests that the surrounding area of Farm SI may be under the risk of Mycoplasma infection.

Our next step is to investigate the farms that may be affected by Outbreak D. We narrow down our analysis targets by filtering out the movements involving inferences with less than 50% confidence (**DG5**). As shown in Fig. 6-(d), three finishing farms (Farm FK, Farm FBT, and Farm FCU) are likely under the influence due to their highly confident inferences. We focus only on the analysis of Farm FBT as one example, since we have demonstrated the interpretation process with multiple farms (Farm FK included) in Case 1 and Sect. 4.2. As shown in Fig. 6-(e), in terms of the feeding cost, the trends of Farm FBT are more similar to those of the diseased farms, supporting the high confidence behind its health status inference (**DG3**). Moreover, the cross icon on June 11th (Fig. 6-(d3)) indicates that Farm FBT administered antibiotics, Aivlosin and Excede. Hovering over other cross icons on the same date, as annotated in Fig. 6-(d1) and (d2), we find that Aivlosin was administered at Farm FK and that Farm FBI administered Excede. One of the animal health specialists explained that since Mycoplasma is a respiratory disease, it is possible that a farm, being aware of the symptoms but not the exact disease, treats for respiratory pathogens in general with the antibiotics. While Aivlosin has been approved for Mycoplasma control in swine [36], Excede is relevant to the control of other respiratory pathogens. This insight suggests that Farm FBT and Farm FK may be aware of the outbreak already and have taken prevention measures by administering antibiotics.

In this case, we have demonstrated how we can perform a comprehensive analysis of an outbreak by investigating the retrospective and the prospective movements. We reveal a risky community of farms, including Farm SI and Farm FM, by associating the animal movements with the geospatial map and validate the health status inferences with the antibiotic history.

6 Expert Review

Our system was designed to support experts in performing essential analysis tasks, as described in Sect. 3.2. Initially, the design goals were derived through multiple meetings with three animal health specialists. We subsequently presented our methodology along with the first two use cases to nine field veterinarians who are constantly working with swine production systems. In addition, this work was presented at an online national exhibition to more than 2,000 registered participants, where many share the background in disease prevention and control for either human health or animal health. With all these expert reviews, we were thus able to refine our design.

Generally, we received positive feedback from all the domain experts, who confirmed that the system effectively coordinates the heterogeneous data to support disease outbreak investigations. The three animal health specialists commented that the health status inferences drive effective analysis, since the lack of diagnostic history previously limited the capability of their analyses. In particular, they added that the utilization of the financial reports improves the applicability of our methodology in practice as it incorporates the financial perspective into analysis to generate more insights.

We also received comments from the experts on potential improvements on the system. The three animal health specialists liked how we provide the interpretations to the health status inferences; however, one of them felt the group characteristic component (i.e., the heatmap in Fig. 1-(b)) was overwhelming without any guidance provided. While the contribution values summarize the general trends of the diseased and healthy farms (e.g., on one attribute, the diseased farms tend to have higher data values than the healthy farms), these values do not effectively highlight the attributes that are most relevant to the inferences of the selected farm. Another specialist thus suggested that we could provide an attribute filter that prioritizes attributes where the diseased or healthy farms show a decreasing/increasing trend in the temporal records. For the technical aspect, one computer science researcher wondered whether the visual summary can include unseen data, while one field veterinarian was curious about the computation time of generating the visual summary. In our experience, with the default hyperparameters in FEALM-UMAP, it takes FEALM-UMAP around 3 minutes to explore the latent features of the data attributes. However, as FEALM-UMAP is a nonlinear DR method, we are required to recompute the visual summary to include unseen data. Finally, the other field veterinarian was further interested in extending our work to investigate control strategies in response to disease outbreaks, where one expert in disease prevention and control mentioned the potential incorporation of spatial analysis into our work.

Overall, the participants agreed that the analysis tasks that our system supports are essential. We plan to extend the system functionalities using another newly obtained real-world production dataset, which involves a larger community of farms and provides more data entries. Our top two priorities are: (1) extending the health status inferences to be disease-aware; (2) highlighting crucial interpretations relevant to a selected farm and improving the design of our group characteristic component. One of our long-term goals is to integrate the capabilities of our system into Disease Bioportal [40] to support swine producers and field veterinarians in near real-time fashion.

7 Discussion

We have demonstrated the effectiveness of our current system with the analysis examples and the expert review. Here we discuss the essential topics relevant to such a visual analytics system in practice.

Generalizability. Our methodology employs DR to infer the health status of the farms from their financial reports, despite the absence of their diagnostic history. This visual analytics approach is an alternative to solving the prediction problem when supervised techniques are not an option. With these health status inferences of the farms, we contextualize the disease outbreak investigations and develop two metrics that measure the potential influence of an outbreak. These two metrics provide intuitive interpretations and guide the user in locating their analysis interest efficiently, as shown in Sect. 5. This concept can inspire researchers to provide analysis guidance in the system. Finally, our workflow demonstrates how we work with limited heterogeneous data in a visual interface of coordinated views to support the user in performing valid analyses.

Lessons learned. From the collaboration with animal health specialists and the review from other domain experts, we have learned how we can further improve the applicability of our system in practice. The first is reducing the model interpretation cost. As described in Sect. 6, the experts confirm that while the interpretations are helpful, it is overwhelming for them to explore the group characteristic component. Besides our plan to provide more visual guidance to support their exploration, future work can explore the automatic summarization of the interpretations. The second is advancing exploratory analysis. We develop visualizations that are not over-complicated and encompass important information for driving effective visual analysis. In Sect. 5, we have shown how the user can utilize our system to explore and analyze different outbreaks. However, the ultimate goal of our system is to assist our target users, swine producers and field veterinarians, in their decision making. Once the user has gained trust in our system, they may prefer examining the analysis outcome directly due to time efficiency. It would be interesting to see how we can automate the analysis process in a visual analytics system and consequently generate a visual summary [6].

Spatiality and temporality. One limitation of our health status inferences is the lack of consideration of the continuity among different time points and locations. As some disease pathogens can transmit via aerosol or survive over varying time periods, it is possible that a farm gets infected by its neighborhood or at farther time points. However, for temporality, the sparse and irregular sampling intervals (mostly 3 or 4 out of 12 time points per farm) have imposed the difficulty of performing functional data analysis or data imputation. As described in Sect. 4.2, we design a tooltip, presenting the temporal distribution of farms per attribute, to provide temporal context to the user when performing analysis. In the disease outbreak overview, we also balance the health status inferences to highlight the animal movements that are most recent to the disease outbreak. With more time points in the future, besides functional analysis, we may employ other methods to extend our health status inferences, such as MulTiDR [17], a DR framework that considers the temporal context of the data, or dynamic mode decomposition. To complement the spatiality consideration, our system displays recent outbreak icons of the same disease on the geospatial map to raise the awareness of geographical proximity. We can extend our health status inferences to be spatially aware. For instance, we may balance the inference with the interpolation computed from the neighboring farms, which are faced with recent outbreaks of the same disease. While it requires expert curation to determine the neighborhood range for each disease (the transmissibility is disease-specific), future work can investigate the incorporation of spatial pattern analysis for its capability of automatically revealing spatial patterns such as a community of farms under the infection risk.

Uncertainty in health status inferences. We provide a systematic method, driven by k-nearest neighbors, that computes and aggregates the importance of the labelled points to the given entity as a probabilistic number for the health inference. Here, we discuss uncertainty derived from limited data and distortions in DR.

One assumption for the health status inference is the impact of the diseases reflects on some attributes in the financial reports, regardless of the types of infectious diseases. As mentioned in Sect. 6, while we plan to extend our methodology to be disease-aware, we should address one subsequent issue — how to maintain the reliability of the inferences when there are less effective results for a disease? Our system incorporates the farms' antibiotic history, one prevention measure for disease control, for analysis validation. Since each of the antibiotics is effective against certain infectious diseases, their administration history can inform which farms have been aware of the presence of certain diseases or symptoms.

There exists inherent uncertainty in DR projections due to algorithm accuracy. Inaccurate DR projections may degrade the reliability of the inferences, since entities of similar data now may not be placed in close proximity in DR projections. Several works [21, 39] have been dedicated to addressing the distortion issue by introducing novel interaction techniques; that is to inform the relocation of the points as they are missing or false neighbors. However, their learnability may impose cognitive burden on our target users, swine producers and field veterinarians, whose expertise tend not to include DR analysis. To maintain the intuition for performing visual analysis, we design the group characteristic component to provide interpretations instead, as described in Sect. 4.2.

8 Conclusion

We have introduced a visual analytics interface that supports animal disease outbreak investigations by analyzing heterogeneous spatio-temporal multivariate data. The analysis relies on an effective coupling of machine learning and visualization methods. Through this interface, we infer the health status of farms with uncertainty and interpretability, despite the absence of their medical information. This capability enables the impact quantification that guides the user toward analyzing important outbreaks. As demonstrated with three case studies, our system facilitates efficient and effective analysis in animal disease surveillance.

Acknowledgments

This work has been supported by the National Science Foundation Awards No. 2134901 and 1838207. Authors would like to thank swine industry collaborators for the provision of high-quality data and their suggestions/insights to improve our outcomes.

References

[1] Our visual analytics interface. http://infovis.cs.ucdavis.edu/.

[2] System demo video. https://youtu.be/KRyQDBTw6SM.

[3] S. Alemzadeh, T. Hielscher, U. Niemann, L. Cibulski, T. Ittermann, H. Völzke, M. Spiliopoulou, and B. Preim. Subpopulation discovery and validation in dpidemiological data. In *Proceedings of the EuroVis Workshop on Visual Analytics*, pp. 43–47, 2017.

[4] T. Baumgartl, M. Petzold, M. Wunderlich, M. Hohn, D. Archambault, M. Lieser, A. Dalpke, S. Scheithauer, M. Marschollek, V. M. Eichel, N. T. Mutters, H. Consortium, and T. V. Landesberger. In search of patient zero: Visual analytics of pathogen transmission pathways in hospitals. *IEEE Trans. Vis. Comput. Graph.*, 27(2):711–721, 2021.

[5] M. Bostock, V. Ogievetsky, and J. Heer. D^3 data-driven documents. *IEEE Trans. Vis. Comput. Graph.*, 17(12):2301–2309, 2011.

[6] C. Bryan, K.-L. Ma, and J. Woodring. Temporal summary images: An approach to narrative visualization via interactive annotation generation and placement. *IEEE Trans. Vis. Comput. Graph.*, 23(1):511–520, 2016.

[7] L. N. Carroll, A. P. Au, L. T. Detwiler, T.-c. Fu, I. S. Painter, and N. F. Abernethy. Visualization and analytics tools for infectious disease epidemiology: A systematic review. *J. Biomed. Inform.*, 51:287–298, 2014.

[8] M. E. Craft. Infectious disease transmission and contact networks in wildlife and livestock. *Philos. Trans. R. Soc. B: Biol. Sci.*, 370(1669):20140107, 2015.

[9] F. C. Dórea and F. Vial. Animal health syndromic surveillance: A systematic literature review of the progress in the last 5 years (2011–2016). *Veterinary Medicine: Research and Reports*, 7:157, 2016.

[10] C. Dunne, M. Muller, N. Perra, and M. Martino. VoroGraph: Visualization tools for epidemic analysis. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pp. 255–258, 2015.

[11] R. Etemadpour, R. Motta, J. G. de Souza Paiva, R. Minghim, M. C. F. De Oliveira, and L. Linsen. Perception-based evaluation of projection methods for multidimensional data visualization. *IEEE Trans. Vis. Comput. Graph.*, 21(1):81–94, 2014.

[12] S. Fadloun, A. Sallaberry, A. Mercier, E. Arsevska, M. Roche, and P. Poncelet. EpidVis: A visual web querying tool for animal epidemiology surveillance. *Inf. Vis.*, 19(1):48–64, 2020.

[13] L. Fang, H. Zhao, P. Wang, M. Yu, J. Yan, W. Cheng, and P. Chen. Feature selection method based on mutual information and class separability for dimension reduction in multidimensional time series for clinical data. *Biomed. Signal Process. Control*, 21:82–89, 2015.

[14] Food and Agriculture Organization of the United Nations. EMPRES-i+: Global animal disease information system. `https://empres-i.apps.fao.org/`. Accessed: 2022-07-10.

[15] T. Fujiwara, Y.-H. Kuo, A. Ynnerman, and K.-L. Ma. Feature learning for dimensionality reduction toward maximal extraction of hidden patterns. *arXiv preprint arXiv:2206.13891*, 2022.

[16] T. Fujiwara, O.-H. Kwon, and K.-L. Ma. Supporting analysis of dimensionality reduction results with contrastive learning. *IEEE Trans. Vis. Comput. Graph.*, 26(1):45–55, 2019.

[17] T. Fujiwara, N. Sakamoto, J. Nonaka, K. Yamamoto, K.-L. Ma, et al. A visual analytics framework for reviewing multivariate time-series data with dimensionality reduction. *IEEE Trans. Vis. Comput. Graph.*, 27(2):1601–1611, 2020.

[18] M. C. Gates, L. K. Holmstrom, K. E. Biggers, and T. R. Beckham. Integrating novel data streams to support biosurveillance in commercial livestock production systems in developed countries: Challenges and opportunities. *Front. Public Health*, 3:74, 2015.

[19] R. Goel, S. Valentin, A. Delaforge, S. Fadloun, A. Sallaberry, M. Roche, and P. Poncelet. EpidNews: Extracting, exploring and annotating news for monitoring animal diseases. *J. Comput. Lang.*, 56:100936, 2020.

[20] D. Guo. Flow mapping and multivariate visualization of large spatial interaction data. *IEEE Trans. Vis. Comput. Graph.*, 15(6):1041–1048, 2009.

[21] N. Heulot, M. Aupetit, and J.-D. Fekete. Proxilens: Interactive exploration of high-dimensional data using projections. In *VAMP: EuroVis Workshop on Visual Analytics using Multidimensional Projections*, 2013.

[22] G. Hrovat, G. Stiglic, P. Kokol, and M. Ojsteršek. Contrasting temporal trend discovery for large healthcare databases. *Comput. Methods Programs Biomed.*, 113(1):251–257, 2014.

[23] A. J. Izenman. Linear discriminant analysis. In *Modern multivariate statistical techniques*, pp. 237–280. Springer, 2013.

[24] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis. Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.*, 13:8–17, 2015.

[25] B. C. Kwon, V. Anand, K. A. Severson, S. Ghosh, Z. Sun, B. I. Frohnert, M. Lundgren, and K. Ng. DPVis: Visual analytics with hidden markov models for disease progression pathways. *IEEE Trans. Vis. Comput. Graph.*, 27(9):3685–3700, 2020.

[26] B. C. Kwon, H. Kim, E. Wall, J. Choo, H. Park, and A. Endert. Axisketcher: Interactive nonlinear axis mapping of visualizations through user drawings. *IEEE Trans. Vis. Comput. Graph.*, 23(1):221–230, 2016.

[27] K. Lee, D. Polson, E. Lowe, R. Main, D. Holtkamp, and B. Martínez-López. Unraveling the contact patterns and network structure of pig shipments in the united states and its association with porcine reproductive and respiratory syndrome virus (PRRSV) outbreaks. *Prev. Vet. Med.*, 138:113–123, 2017.

[28] Y. Livnat, T.-M. Rhyne, and M. Samore. Epinome: A visual-analytics workbench for epidemiology data. *IEEE Comput. Graph. Appl.*, 32(2):89–95, 2012.

[29] R. Maciejewski, B. Tyner, Y. Jang, C. Zheng, R. V. Nehme, D. S. Ebert, W. S. Cleveland, M. Ouzzani, S. J. Grannis, and L. T. Glickman. LAHVA: Linked animal-human health visual analytics. In *Proc. VAST*, pp. 27–34, 2007.

[30] B. Martínez-López, A. Perez, and J. Sánchez-Vizcaíno. Social network analysis. review of general concepts and use in preventive veterinary medicine. *Transbound. Emerg. Dis.*, 56(4):109–120, 2009.

[31] L. McInnes, J. Healy, and J. Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[32] S. Otake, S. Dee, C. Corzo, S. Oliveira, and J. Deen. Long-distance airborne transport of infectious PRRSV and Mycoplasma hyopneumoniae from a swine population infected with multiple viral variants. *Vet. Microbiol.*, 145(3-4):198–208, 2010.

[33] A. Pandey, H. Shukla, G. S. Young, L. Qin, A. A. Zamani, L. Hsu, R. Huang, C. Dunne, and M. A. Borkin. Cerebrovis: Designing an abstract yet spatially contextualized cerebral artery network visualization. *IEEE Trans. Vis. Comput. Graph.*, 26(1):938–948, 2019.

[34] A. Perez, M. AlKhamis, U. Carlsson, B. Brito, R. Carrasco-Medanic, Z. Whedbee, and P. Willeberg. Global animal disease surveillance. *Spat. Spatio-temporal Epidemiol.*, 2(3):135–145, 2011.

[35] D. U. Pfeiffer and K. B. Stevens. Spatial and temporal epidemiological analysis in the big data era. *Prev. Vet. Med.*, 122(1-2):213–220, 2015.

[36] Pharmgate. Aivlosin WSG approved by U.S. and Canadian regulators to control Mycoplasma hyopneumoniae in swine. `https://www.pharmgate.com/usa/aivlosin-wsg-approved-by-u-s-and-canadian-regulators-to-control-mycoplasma-hyopneumoniae-in-swine/`. Accessed: 2022-10-10.

[37] B. Preim and K. Lawonn. A survey of visual analytics for public health. *Comput. Graph. Forum*, 39(1):543–580, 2020.

[38] S. Schöttler, Y. Yang, H. Pfister, and B. Bach. Visualizing and interacting with geospatial networks: A survey and design space. In *Comput. Graph. Forum*, vol. 40, pp. 5–33, 2021.

[39] J. Stahnke, M. Dörk, B. Müller, and A. Thom. Probing projections: Interaction techniques for interpreting arrangements and errors of dimensionality reductions. *IEEE Trans. Vis. Comput. Graph.*, 22(1):629–638, 2015.

[40] University of California, Davis. Disease Bioportal. `https://bioportal.ucdavis.edu/`. Accessed: 2022-07-10.

[41] Q. Wang, T. Mazor, T. A. Harbig, E. Cerami, and N. Gehlenborg. Threadstates: State-based visual analysis of disease progression. *IEEE Trans. Vis. Comput. Graph.*, 28(1):238–247, 2021.

[42] B. Wolter, M. Ellis, J. DeDecker, S. Curtis, G. Hollis, R. Shanks, E. Parr, and D. Webel. Effects of double stocking and weighing frequency on pig performance in wean-to-finish production systems. *J. Anim. Sci.*, 80(6):1442–1450, 2002.

[43] M. Wunderlich, I. Block, T. von Landesberger, M. Petzold, M. Marschollek, and S. Scheithauer. Visual analysis of probabilistic infection contagion in hospitals. In *VMV*, pp. 143–150, 2019.

[44] C. Yang, Z. Zhang, Z. Fan, R. Jiang, Q. Chen, X. Song, and R. Shibasaki. EpiMob: Interactive visual analytics of citywide human mobility restrictions for epidemic control. *IEEE Trans. Vis. Comput. Graph.*, 2022.