Graph symbolic regression to interpret the propagation of Vesicular Stomatitis Virus across the U.S. and Mexico

Tamanna Rashme
Mississippi State University
Starkville, MS, USA
tyr17@msstate.edu

Zonghan Zhang
Mississippi State University
Starkville, USA
zz239@msstate.edu

Jason Weeks
Mississippi State University
Starkville, US
jcw1044@msstate.edu

Marouane Benbrahim
Mississippi State University
Starkville, USA
mb4194@msstate.edu

Zijian Zhang
Mississippi State University
Starkville, USA
zz242@msstate.edu

Zhiqian Chen
Mississippi State University
Starkville, USA
zchen@cse.msstate.edu

Nisha Pillai
Mississippi State University
Starkville, USA
pillai@cse.msstate.edu

Ram Ramkumar
Mississippi State University
Starkville, USA
ramkumar@cse.msstate.edu

Bindu Nanduri
Mississippi State University
Starkville, USA
BNanduri@cvm.msstate.edu

Abstract

Vesicular stomatitis virus (VSV) causes livestock disease cases that occur every year in regions in Mexico. Every few years, VSV spreads northwards into the U.S. in large outbreak events affecting hundreds of livestock premises across multiple states, leading to significant economic losses due to quarantines, trade restrictions, and veterinary expenses. VSV cases are mainly driven by biting arthropod vectors from multiple genera with different ecologies, making outbreak control challenging. The sporadic nature of outbreaks and limited understanding of transmission dynamics further hinder containment efforts, reducing the effectiveness of preemptive measures. In this paper, we propose an interpretable model to elucidate the key rules governing the spread of VSV. This model employs a sparse symbolic regression model, SINDy (Sparse Identification of Nonlinear Dynamical Systems), to identify the most significant ecological variables in spread dynamics, considering both spatial and temporal factors. Since many counties did not have VSV cases during the study period, counties were clustered into 40 regions incorporating static environmental variables land cover, soil properties, livestock density, and climate data and using spatially constrained Agglomerative Clustering based on geographic adjacency, resulting in an average region size of approximately 90 counties. Ecological variables included dynamic and static variables such as temperature, humidity, wind, soil characteristics, and altitude associated with vectors and hosts (cattle, horses, and mules). The change in cases from month to month by region was modeled with three SINDy models: a SINDy model with normal features (Normal); a SINDy model with graphical features (Graph); and a SINDy model with a subset of select graphical features (Graph (Forced)). Each alpha was chosen to minimize CV-MSE while retaining less than 11 terms. Graphical features greatly reduced model error, and the SINDy model with select graphical features had a slightly better CV-MSE score than when all graphical features were included. All models identified the infected species as important in capturing the dynamics of case differences between regions. The density of host animals, soil properties (the amount of sand in the topsoil layer; the amount of clay in the subsoil layer), land cover (urban and needleleaf deciduous tree cover) and land cover interactions with temperature were also identified as important in at least one of the three modeling approaches. The regional connectivity of the species affected by the cases was identified as an important graphical feature. We can now use these features to distinguish between regions where there are larger differences in case numbers from month-to-month (epidemic regions) and regions where there are smaller differences (endemic regions). When cases are reported outside of the endemic region, livestock managers can be alerted to the higher risk of an upcoming outbreak.

Keywords

VSV, Livestock disease,SINDy, Graphical features, Ecological variables Spatial clustering,Outbreak prediction.

ACM Reference Format:
Tamanna Rashme, Zonghan Zhang, Jason Weeks, Marouane Benbrahim, Zijian Zhang, Zhiqian Chen, Nisha Pillai, Ram Ramkumar, and Bindu Nanduri. 2018. Graph symbolic regression to interpret the propagation of Vesicular Stomatitis Virus across the U.S. and Mexico . In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (33rd ACM SIGSPATIAL).* ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Emerging infectious diseases pose a significant threat to both human and animal health, with their dynamics often influenced by complex interactions between environmental, climatic, and social factors. Among such diseases, Vesicular stomatitis virus (VSV) outbreaks pose a significant threat to food security, causing substantial economic challenges that ripple through the agricultural sector. The implementation of quarantines and trade restrictions, coupled with rising veterinary expenses, severely impacts livestock production and the food supply chain. These outbreaks affect hundreds of livestock premises across multiple states, jeopardizing the stability

33rd ACM SIGSPATIAL, Minneapolis, MN
2018. ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

of regional food systems. The 2019 outbreak exemplifies this threat, with VSV spreading to 1,144 premises across eight U.S. states and overwintering to expand further in 2020, affecting 326 premises in three additional states [**?**]. This persistence and spread underscore the virus's potential to disrupt long-term food production capabilities.

The complexity of VSV transmission is heightened by its dual propagation pathways: direct animal contact and transmission via a diverse array of arthropod vectors. These vectors, drawn from multiple genera, possess unique ecological characteristics that influence their role in disease spread [**? ?**]. This multifaceted transmission dynamic not only threatens immediate livestock health but also poses risks to sustainable food production practices. The inability to control VSV spread efficiently could lead to prolonged production losses, potentially affecting food availability and affordability for consumers. Accurate forecasting of VSV cases can aid in the timely deployment of preventive measures and resource allocation, contributing to disease management and control strategies.

The intricacy of VSV necessitates a comprehensive understanding of its underlying processes, beyond mere identification of correlations between VSV occurrences and environmental factors [**?**]. This presents several challenges: **1. Multifaceted environmental interactions**: While factors such as temperature, humidity, and soil composition play critical roles in VSV transmission, their integration into predictive models remains complex. **2. Demand for model-wide interpretability**: Current models often lack global explanatory power, providing limited insights that fail to elucidate the overarching principles governing VSV spread across various regions and timeframes. **3. Suboptimal spatiotemporal integration**: Existing methodologies frequently struggle to effectively incorporate spatial and temporal dimensions crucial for understanding VSV dynamics, resulting in predictions that are either oversimplified or excessively intricate.

In this study, we address these challenges by leveraging the Sparse Identification of Nonlinear Dynamics (SINDy) [**?**] framework, which is well-suited for discovering parsimonious models that describe the evolution of dynamic systems. SINDy models have gained traction for their ability to uncover governing equations directly from data, making them particularly appealing for systems where theoretical models are either unavailable or difficult to construct. By integrating environmental, meteorological, and graph-based convolutional features alongside historical VSV case data, our approach captures the underlying dynamics driving disease spread. The model not only predicts future case numbers but also provides interpretable insights into the roles of various features.

Our approach makes several significant contributions to the field of infectious disease modeling:

- **Interpretability through symbolic regression**: The use of symbolic regression enables the generation of mathematical expressions that clarify the connections between ecological variables and VSV propagation, fostering model-level interpretability.
- **Graph operators for spatiotemporal integration**: We introduce graph operators to incorporate spatial and temporal factors effectively. These operators encode spatial relationships among regions and the temporal evolution of ecological variables, leading to more precise and contextually informed predictions.
- **Feature selection for computational efficicency**: Given the high dimensionality of the input data, we employ the K-best algorithm to reduce the number of features while retaining predictive power. This step enhances both the efficiency and interpretability of the model.
- **Interpretable equations to describe the relationship**: Beyond its predictive capabilities, the SINDy model provides interpretable equations that describe the relationships between different features and the target variable. This interpretability is crucial for gaining a deeper understanding of the factors driving VSV outbreaks and for developing data-driven policies to mitigate their impact. The insights gained from this model can also be extended to other infectious diseases with similar dynamic properties.
- **Empirical validation using real-world data**: Extensive evaluations on a comprehensive dataset documenting VSV spread across the United States validate the efficacy of our approach in real-world scenarios.

Overall, this work makes significant contributions to the field of infectious disease modeling by introducing a novel application of the SINDy framework to the prediction of VSV cases. The combination of dynamic modeling, feature selection, and interpretability positions our approach as a valuable tool for both researchers and practitioners working in the domain of disease management and control. The findings from this study have the potential to inform future research and contribute to the development of more robust and generalizable models for infectious disease forecasting.

2 Related Work

Vesicular Stomatitis Virus (VSV). transmission is governed by intricate ecological interactions and environmental dynamics. Peters et al. [**?**] underscored the critical role of integrating big data and machine learning (ML) methodologies in ecological research, an approach that has demonstrated significant promise in VSV studies. Predictive models for VSV have effectively utilized diverse data sources, including climatic variables [**?**], vector distribution patterns [**?**], and complex host-pathogen interactions [**?**]. The application of big data approaches has markedly enhanced the precision of these models, particularly in capturing and analyzing intricate spatial and temporal patterns of virus spread [**?**]. Furthermore, in-depth investigations into vector behavior [**?**] and habitat suitability [**?**] emphasize the necessity for high-resolution, multidimensional data sets. Recent advancements have also witnessed the integration of remote sensing technologies, offering new perspectives for monitoring and predicting VSV dynamics across diverse landscapes [**? ?**]. However, significant challenges persist, particularly in the realms of model interpretability [**?**] and the seamless integration of heterogeneous data sources [**?**]. While the majority of related research typically concentrates on variable correlation and localized effects, our work distinguishes itself by adopting a more holistic approach. We aim to uncover the global governing rules for VSV spreading, considering the interconnected nature of

ecological systems. This comprehensive perspective allows for a deeper understanding of the virus's transmission dynamics, potentially revealing previously overlooked factors that influence its spread across different regions and ecosystems. By synthesizing diverse data streams and employing advanced analytical techniques, our research strives to provide a more nuanced and actionable understanding of VSV epidemiology, ultimately contributing to more effective prevention and control strategies.

Symbolic Regression and SINDy. The discovery of governing equations from data using sparse identification has become a transformative approach to understanding nonlinear dynamical systems. Brunton et al. [**?**] introduced a method that has been widely adopted in various fields. Building on this work, researchers have explored different techniques for model discovery, such as using compressed sensing [**?**] and machine learning [**?**]. The Sparse Identification of Nonlinear Dynamical Systems (SINDy) framework has been extended to include control systems [**?**] and to enhance robustness against noise [**?**]. Applications have spanned fluid dynamics [**?**], structural health monitoring [**?**], and biological systems [**?**]. Advances also include integrating deep learning methods [**?**], addressing chaotic systems [**?**], and improving interpretability [**?**]. The flexibility and power of these methods underscore their potential to reveal the underlying mechanisms of complex systems.

SINDy seeks to discover parsimonious representations of governing equations by identifying only the most relevant terms from a large set of candidate functions. This capability makes it particularly attractive for systems characterized by nonlinear and sparse dynamics. One of the key advantages of SINDy lies in its ability to provide interpretable models, which is crucial for understanding the underlying mechanisms of complex systems. Unlike traditional black-box machine learning models, SINDy explicitly identifies the mathematical relationships between state variables and their temporal derivatives.

Most existing work does not address how to incorporate graph operators in the SINDy framework. Our contribution lies in exploring a method to integrate graph structures under VSV scenarios. This novel approach enables us to capture the complex network dynamics often present in epidemiological systems without requiring pre-defined theoretical models, potentially revealing new insights into the spread and control of VSV. By combining graph theory with symbolic regression, we aim to develop a more comprehensive modeling framework that can account for the interconnected nature of host populations and environmental factors influencing virus transmission.

3 Method: Graph Symbolic Regression

In this section, we outline the target problem and demonstrate how we integrate graph and temporal factors alongside all ecological variables.

3.1 Problem Formulation

The primary objective of our study is to derive an interpretable closed-form equation that uses a limited number of terms while maintaining satisfactory predictive performance. We employ an extended version of Sparse Identification of Nonlinear Dynamics (SINDy) adapted for graph domains. This extension incorporates graph-filtered features and additional candidate functions and operators, allowing us to interpret behaviors on spatial graphs in our application, alongside conventional variables. Our goal is to obtain an equation that elucidates the mathematical relationships among these elements. In the context of , we extend the SINDy by incorporating graph-based features. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{W})$ represent our graph, where $\mathcal{V}$ is the set of vertices, $\mathcal{E}$ is the set of edges, and $\mathbf{W}$ is the weighted adjacency matrix.

Specifically, we aim to formulate an ordinary differential equation (ODE) with respect to time, describing the rate of change in infections:

$$\frac{\partial \mathbf{X}}{\partial t} = \Theta(\mathbf{X}, \mathcal{G})\Xi, \tag{1}$$

where $\mathbf{X}$ is the matrix of observed states over time (vector of infection numbers on different locations), $\mathcal{G}$ denotes the spatial networks, and $\Theta(\mathbf{X}, \mathcal{G})\Xi$ is the inner product between Θ and Ξ, and it is predicted state based on the sparse coefficients Ξ. $\Theta(\mathbf{X}, \mathcal{G})$ is the linear weighted sum of a library of candidate functions (e.g., polynomial, trigonometric and so on), including graph factor $\mathcal{L}$. This change is influenced not only by ecological factors but also by spatial graph operators. Given the initial condition $\mathbf{X}(0)$ at time 0, we can determine the constant of integration. Integrating both sides with respect to t, we have:

$$\mathbf{X}(t) = \int_0^t \Theta(\mathbf{X}, \mathcal{G})\Xi \, dt + \mathbf{X}(0). \tag{2}$$

So, the selected functions in Θ can be easily integrated, particularly when it involves typical functions like polynomials and trigonometric functions. For example, Θ can be:

$$[\mathbf{X}^n, \ldots, \sin(\mathbf{X}), \cos(\mathbf{X}), \ldots, \nabla_{\mathcal{G}}\mathbf{X}, \Delta_{\mathcal{G}}\mathbf{X}, e^{-t\mathcal{L}}\mathbf{X}, \ldots],$$

where $\mathbf{X}$ represents the state vector, $\mathcal{G}$ denotes the graph structure, $\nabla_{\mathcal{G}}$ is the graph gradient operator, $\Delta_{\mathcal{G}}$ is graph Laplacian operator, and $\mathcal{L}$ is the graph Laplacian matrix. To achieve the sparsity, the objective function is:

$$\min_{\Xi} \sum_{t=0}^{T} \|\mathbf{X}(t) - \overbrace{\Theta(\mathbf{X}(t), \mathcal{G})\Xi}^{\text{prediction at time } t}\|_2^2 + \lambda \overbrace{\|\Xi\|_1}^{\text{sparsity}}, \tag{3}$$

where $\sum_{t=0}^{T}$ captures the total error over all time steps from $t = 0$ to $t = T$, ensuring that the model minimizes the cumulative prediction error across the entire time horizon while enforcing sparsity in the coefficient matrix Ξ through the L_1 regularization term. The parameter λ continues to balance the trade-off between model accuracy and sparsity.

3.2

The Graph SINDy algorithm is a data-driven method designed to uncover the governing equations of dynamical systems on graphs. By combining the sparse identification framework with graph-structured data, this algorithm constructs a library of candidate functions comprising both regular variables and graph operators. It then iteratively identifies a sparse set of these functions that most accurately depict the system's dynamics. Through a process of sequential thresholded least squares, the algorithm enforces sparsity in the model while capturing both local node-level behavior and

graph-level interactions. This results in interpretable equations that describe the evolution of the complex system.

Candidate Function Library. The library of candidate functions $\Theta(\mathbf{X}, \mathcal{G})$ is crucial for the performance and interpretability of the model. We extend the traditional SINDy library to include graph-based terms, dividing our library into two main categories: normal variables (features on nodes) and graph operators. The normal variables operate directly on the state variables $\mathbf{X}$ and include constant terms, linear terms $\mathbf{X}$, polynomial terms such as $\mathbf{X}^2$ and $\mathbf{X}^3$, trigonometric functions like $\sin(\mathbf{X})$ and $\cos(\mathbf{X})$, exponential terms $e^{\mathbf{X}}$ and $e^{-\mathbf{X}}$, and logarithmic terms $\log(|\mathbf{X}|)$. The graph operators, on the other hand, incorporate the graph structure $\mathcal{G}$ into the dynamics. These may include the graph gradient $\nabla_{\mathcal{G}}\mathbf{X}$, graph Laplacian $\Delta_{\mathcal{G}}\mathbf{X}$, graph diffusion $e^{-t\mathcal{L}}\mathbf{X}$, normalized graph Laplacian $\mathcal{L}_{\text{norm}}\mathbf{X}$, graph curvature $\text{curv}_{\mathcal{G}}(\mathbf{X})$, and graph wavelets $\mathcal{W}_{\mathcal{G}}(\mathbf{X})$. Additionally, we may include cross-terms that combine normal variables and graph operators, such as $\mathbf{X} \cdot \nabla_{\mathcal{G}}\mathbf{X}$, $\mathbf{X} \cdot \Delta_{\mathcal{G}}\mathbf{X}$, $\mathbf{X}^2 \cdot \Delta_{\mathcal{G}}\mathbf{X}$, $\sin(\mathbf{X}) \cdot \nabla_{\mathcal{G}}\mathbf{X}$, and $e^{\mathbf{X}} \cdot \mathcal{L}_{\text{norm}}\mathbf{X}$. So, the complete library can be formally expressed as:

$$\begin{aligned}\Theta(\mathbf{X}, \mathcal{G}) = [&1, \mathbf{X}, \mathbf{X}^2, \mathbf{X}^3, \sin(\mathbf{X}), \cos(\mathbf{X}), e^{\mathbf{X}}, \log(|\mathbf{X}|), \ldots,\\ &\nabla_{\mathcal{G}}\mathbf{X}, \Delta_{\mathcal{G}}\mathbf{X}, e^{-t\mathcal{L}}\mathbf{X}, \mathcal{L}_{\text{norm}}\mathbf{X}, \text{curv}_{\mathcal{G}}(\mathbf{X}), \mathcal{W}_{\mathcal{G}}(\mathbf{X}), \ldots,\\ &\mathbf{X} \cdot \nabla_{\mathcal{G}}\mathbf{X}, \mathbf{X}^2 \cdot \Delta_{\mathcal{G}}\mathbf{X}, \sin(\mathbf{X}) \cdot \nabla_{\mathcal{G}}\mathbf{X}, e^{\mathbf{X}} \cdot \mathcal{L}_{\text{norm}}\mathbf{X}, \ldots]\end{aligned}$$

[tb] **Input**: Time series data $\mathbf{X}$, graph structure $\mathcal{G}$, library of candidate functions $\Theta(\mathbf{X}, \mathcal{G})$
Parameter: Sparsification parameter λ, max iterations $k_{\max}$
Output: Coefficient matrix Ξ [1] Collect infection numbers $\mathbf{Y}$ for each node Initialize with Least Square Solution:
$\Xi = (\Theta(\mathbf{X}, \mathcal{G})^T\Theta(\mathbf{X}, \mathcal{G}))^{-1}\Theta(\mathbf{X}, \mathcal{G})^T\mathbf{Y}$ $k = 1$ to $k_{\max}$ smallinds $\leftarrow$ $|\Xi| < \lambda$ Ξ[smallinds] $\leftarrow 0$ $i = 1$ to n biginds $\leftarrow \neg$smallinds$[:, i]$ Ξ[biginds, i] $\leftarrow \Theta(\mathbf{X}, \mathcal{G})[:,$ biginds$]\backslash\mathbf{Y}[:, i]$ support of Ξ unchanged **break** Construct sparse model: $\mathbf{Y} = \Theta(\mathbf{X}, \mathcal{G})\Xi$ **return** Ξ

Sparse Regression. The algorithm aims to discover governing equations for dynamical systems on graphs by identifying sparse representations of the infection dynamics. It combines the Sparse Identification of Nonlinear Dynamics (SINDy) framework with graph-structured data.

The algorithm takes as *input*: (1) Time series data $\mathbf{X}$ for each node in the graph. (2) Graph structure $\mathcal{G}$. (3) Library of candidate functions $\Theta(\mathbf{X}, \mathcal{G})$ including both normal variables and graph operators (e.g., graph Laplacian, graph gradient). Key *hyperparameters* are: (1) Sparsification threshold λ. (2) Maximum number of iterations $k_{\max}$.

Algorithm Steps. The algorithm begins by collecting infection numbers $\mathbf{Y}$ for each node from the time series data $\mathbf{X}$ (line 1). It then computes an initial guess for the coefficient matrix Ξ using ordinary least squares regression: $\Xi = (\Theta(\mathbf{X}, \mathcal{G})^T\Theta(\mathbf{X}, \mathcal{G}))^{-1}\Theta(\mathbf{X}, \mathcal{G})^T\mathbf{Y}$ (line 2). The main iterative process starts with identifying small coefficients in Ξ (those with absolute value less than λ) and setting them to zero (lines 4-5). For each node, a least squares regression is performed using only the non-zero terms: Ξ[biginds, i] $\leftarrow \Theta(\mathbf{X}, \mathcal{G})[:,$ biginds$]\backslash\mathbf{Y}[:, i]$ (lines 6-9). This process repeats until either the support of Ξ (i.e., the pattern of non-zero entries) stops changing or the maximum number of iterations is reached (lines 10-12). Finally, the algorithm constructs the sparse model $\mathbf{Y} = \Theta(\mathbf{X}, \mathcal{G})\Xi$ (line 13) and returns the coefficient matrix Ξ (line 14).

3.3 Integration on Time

Direct integration, while theoretically possible, is often impractical or inaccurate, especially when the identified system contains nonlinear terms (e.g., graph operators). Once we have identified the sparse coefficients Ξ, we solve the following initial value problem:

$$\frac{d\mathbf{X}}{dt} = \Theta(\mathbf{X}, \mathcal{G})\Xi, \quad \mathbf{X}(0) = \mathbf{X}_0. \tag{4}$$

Using the kth-order Runge-Kutta method [**?**], we set k=4:

$$\begin{aligned}\mathbf{k}_1 &= h \cdot \Theta(\mathbf{X}_n, \mathcal{G})\Xi,\\ \mathbf{k}_2 &= h \cdot \Theta(\mathbf{X}_n + \frac{1}{2}\mathbf{k}_1, \mathcal{G})\Xi,\\ \mathbf{k}_3 &= h \cdot \Theta(\mathbf{X}_n + \frac{1}{2}\mathbf{k}_2, \mathcal{G})\Xi,\\ \mathbf{k}_4 &= h \cdot \Theta(\mathbf{X}_n + \mathbf{k}_3, \mathcal{G})\Xi,\\ \mathbf{X}_{n+1} &= \mathbf{X}_n + \frac{1}{6}(\mathbf{k}_1 + 2\mathbf{k}_2 + 2\mathbf{k}_3 + \mathbf{k}_4),\end{aligned} \tag{5}$$

where h is the time step, $\mathbf{X}_n$ is the current state, and $\mathbf{X}_{n+1}$ is the next state. Each $\mathbf{k}_i$ represents an estimate of the slope at different points within the time step. This process is repeated for each time step to obtain the full trajectory $\mathbf{X}(t)$ from $t = 0$ to the desired end time.

3.4 Agglomerative Clustering

Agglomerative clustering is a bottom-up hierarchical method that combines data points that are similar to each other until a full tree model is formed [**?**]. Ward's method is one of the best ways to link things together because it keeps the variance within clusters to a minimum [**?**]. At each iteration, the pair of clusters that minimizes a specified dissimilarity measure is merged. This process continues until all data points belong to a single cluster or until a predefined number of clusters is reached.

In our study, we employ *Ward's linkage method*, which is particularly effective in minimizing the total within-cluster variance. The merging criterion is defined to select the pair of clusters A and B that minimizes the increase in within-cluster sum of squares (WCSS). The cost function is expressed as:

$$\Delta E = \frac{n_A n_B}{n_A + n_B} \left\|\boldsymbol{\mu}_A - \boldsymbol{\mu}_B\right\|^2$$

where n_A and n_B denote the sizes of the clusters, and $\boldsymbol{\mu}_A$, $\boldsymbol{\mu}_B$ are their respective centroids. This formulation ensures that clusters are merged in a way that retains internal compactness, producing more homogeneous groups in the feature space.To enhance geographic modeling, we adapt agglomerative clustering by integrating spatial adjacency constraints, ensuring that only neighboring regions are merged—an essential consideration for ecological validity and spatial interpretability.

3.5 Time Complexity Analysis

The time complexity of the Graph SINDy algorithm can be analyzed by examining its key components. The initialization phase, comprising the collection of $\mathbf{Y}$ and the initial least squares computation, has a complexity of $O(n)$ and $O(p^3 + np^2)$ respectively, where n is the number of nodes and p is the number of candidate functions in

$\Theta(\mathbf{X}, \mathcal{G})$. The main loop, which runs for at most $k_{\max}$ iterations, consists of finding small indices with $O(np)$ complexity and performing least squares on non-zero terms for each node, with a complexity of $O(p'^3 + mp'^2)$ per node. Here, p' is the number of non-zero terms ($p' \ll p$), and m is the number of time steps in the data. The final model construction has a complexity of $O(np)$. The dominant terms in this analysis are the initial least squares, $O(p^3 + np^2)$, and the least squares in the main loop, $O(k_{\max} \cdot n \cdot (p'^3 + mp'^2))$. Consequently, the overall time complexity can be expressed as $O(p^3 + np^2 + k_{\max} \cdot n \cdot (p'^3 + mp'^2))$. In the worst case, where $p' \approx p$, this becomes $O(p^3 + np^2 + k_{\max} \cdot n \cdot (p^3 + mp^2))$. The time complexity is heavily influenced by the number of candidate functions (p), the number of nodes (n), the number of time steps in the data (m), and the maximum number of iterations ($k_{\max}$). It's worth noting that the sparsity-promoting nature of the algorithm tends to reduce p' over iterations, which can significantly reduce the actual runtime compared to this worst-case analysis. The time integration step using RK4 has a time complexity of $O(Nmnp)$ per time step, where N is the number of time steps in the integration, m is the number of candidate functions, n is the number of nodes, and p is the number of non-zero terms in Ξ.

4 Experiments

We have assessed our model's performance using real-world data on the spread of VSV in the United States and Mexico after appropriate processing. Please refer to the detailed information below.

4.1 Data Preparation

Data Prepossessing. (1) Geographic Shapefiles: The static data contains geographic identifiers (`COUNTY_MUNI_CODE`). Geographic shapefiles for the U.S. and Mexico were acquired, providing polygonal representations of the regions. The data for U.S. counties was obtained from the 2023 TIGER/Line dataset, and Mexican administrative boundaries were downloaded from an official geographic repository.

(2) Combining Geographic Data: To integrate the geographic data, the static dataset was merged with the corresponding shapefiles based on unique geographic identifiers (`GEOID` for the U.S. and `ADM2_PCODE` for Mexico). The resulting dataset, containing both geographic and non-geographic attributes, was reprojected to the Web Mercator coordinate system (EPSG:3857) to facilitate accurate distance calculations.

(3) Centroid Calculation: For each geographic entity, the centroid was computed in the Web Mercator projection. These centroids were then reprojected back to the WGS 84 coordinate system (EPSG:4326) to ensure that geodesic distances, which represent the shortest path between two points on the Earth's surface, could be accurately calculated.

(4) Pairwise Distance Calculation: The pairwise geodesic distances between all centroids were calculated using a k-d tree, a data structure that efficiently handles nearest-neighbor searches. A threshold distance of 3,000 kilometers was applied to limit the calculations to relevant pairs, ensuring computational efficiency while maintaining accuracy. The geodesic distances were then transformed to reciprocal values, enhancing the influence of closer pairs. These reciprocal distances were stored in a sparse matrix format for efficiency.

(5) Finalizing the Distance Matrix: The distance matrix, now in DataFrame format and containing reciprocal distances, was serialized and saved for future use in modeling tasks. Let $\mathbf{L}_1$ denote the original distance matrix, representing the pairwise reciprocal distances between nodes (e.g., counties), and $\mathbf{L}_2 = \mathbf{L}_1^2$ denote the squared reciprocal distance matrix, which amplifies the influence of stronger connections in the graph. Additionally, let $\mathbf{L}^{-1}$ denote the inverse of the reciprocal distance matrix, representing a transformation that diffuses features across the graph inversely proportional to the connections. The matrix $\mathbf{X}$ represents the aggregated feature matrix from the dataset, with shape 3576×86, where 3576 indicates the number of nodes (e.g., counties), and 86 indicates the number of features for each node. The objective is to apply the transformations encoded in $\mathbf{L}_1$, $\mathbf{L}_2$, and $\mathbf{L}^{-1}$ to the feature matrix $\mathbf{X}$ to produce three new matrices $\mathbf{Y}_1$, $\mathbf{Y}_2$, and $\mathbf{Y}_{\text{inv}}$, all of which maintain the same shape 3576×86. The matrix multiplication process is described by the equations $\mathbf{Y}_1 = \mathbf{L}_1 \times \mathbf{X}$, $\mathbf{Y}_2 = \mathbf{L}_2 \times \mathbf{X}$, and $\mathbf{Y}_{\text{inv}} = \mathbf{L}^{-1} \times \mathbf{X}$. This process of multiplying the distance matrices with the feature matrix is analogous to a graph convolution operation, where features are propagated and transformed based on the graph structure. Consequently, the resulting matrices encapsulate the influence of the graph's connectivity on the features, providing a rich representation for subsequent tasks. A preliminary result is shown in Fig. **??**. Data and code for the Distance Matrix are available here [1].

(6) Ecological Region-Based Clustering: To define meaningful spatial regions, we computed statistical summaries—mean and standard deviation—of ecological and environmental variables across time for each spatial unit. These features captured both temporal trends and spatial variability. Geographic boundaries were merged and processed to identify neighboring regions using geometric adjacency. A sparse connectivity graph was then constructed to enforce spatial contiguity during clustering. Using Ward's linkage and a fixed number of clusters ($n = 40$), the algorithm produced compact, contiguous regions Figure 1. suitable for ecological interpretation and modeling.

(7)Visualization Framework for Vesicular Stomatitis Virus (VSV) Spread.

We used regionally limited clustering (K=40) to find areas that are ecologically similar for making predictions about VSV. The approach used an adjacency matrix to make sure that geographic regions were connected, and Ward linkage grouped places that were next to one other based on comparable weather conditions. The clusters that came out (**Figure. 1**) show clear geographical patterns: huge, uniform areas cover the central plains, while smaller clusters mark the coastal and border microclimates. Colour-coded visualisation shows that spatial segmentation was effective, with clusters that are clearly defined and respect both environmental similarities and physical geography. This regionalisation creates useful spatial units for future VSV modelling, with each cluster representing locations with similar environmental features.

(8)Experiment Environment. The experiments were conducted on a high-performance machine equipped with an AMD Ryzen Threadripper PRO 5955WX processor, which features 16 cores and 32 threads with support for frequencies up to 7.03 GHz. The machine also includes dual NVIDIA GeForce RTX 4090 GPUs, each with 24

[1] https://figshare.com/s/4cb54edf31e5b05fe3fa

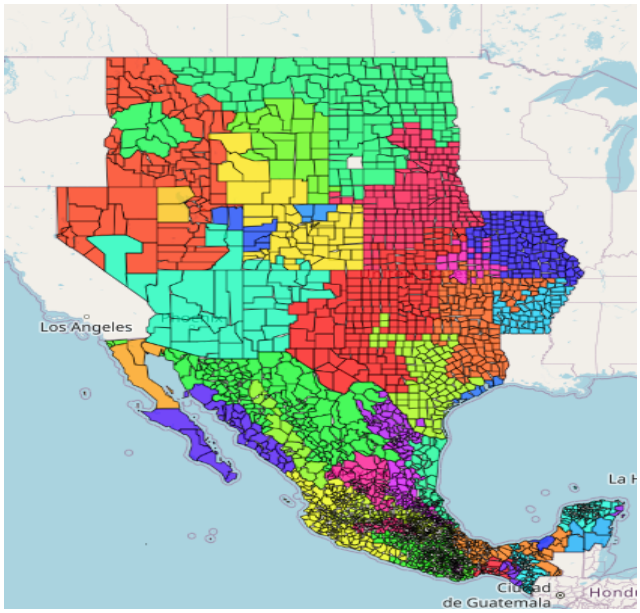

Figure 1: Spatially Constrained Clustering of U.S. and Mexican Regions into 40 Contiguous Ecological Units. Each region is colored uniquely and reflects spatial similarity based on ecological and livestock attributes.

GB of GDDR6X memory, 128 GB of DDR4 RAM, and a 3.5 TB NVMe SSD. This setup efficiently handles the complex data processing and analysis tasks required for our experiments, especially those involving intensive computations and large datasets.

4.2 Symbolic Regression for real-world VSV

Task Setup. To evaluate our approach, we selected counties that reported more than ten months of positive VSV cases during the 2004-2005 , 2014-2015 and 2019-2020 outbreaks. For each selected county, we trained an individual SINDy model using data from the 36 months of the outbreak and obtained an equation representing the relationship between the features and the derivatives of dy.

After computing the graph convolution vectors, we concatenate them with the original feature vector to form an augmented feature vector for each county, resulting in a combined vector of size 344 for each county, which includes the original features as well as the spatial information captured by the graph convolutions. The final input vector for our model is constructed by appending the timestamp associated with each data point to this augmented feature vector, resulting in a vector of size 345. We then construct the model input matrix by reorganizing the feature vectors into a temporal sequence. Specifically, for each county, we create a matrix where the columns represent the feature vectors at different timestamps, and the rows represent the 505 months in our dataset. The resulting matrix has a size of 345 × 505 and serves as the input to our predictive model, which aims to forecast the changes in the target variable over time.

Our primary objective is to predict the derivative of the output, which represents the temporal change in the target variable. Formally, for each timestamp t, the input vector X_t is used to predict the difference between the target variable at time $t + 1$ and time t, denoted as $\Delta y_t = y_{t+1} - y_t$. In other words, our model seeks to learn the relationship between the current state of a county, represented by the input vector, and the rate of change of the target variable in the subsequent time step. This formulation allows our model to capture both the temporal dynamics and the spatial dependencies inherent in the data. By incorporating graph convolution operations, we can effectively model the influence of neighboring counties on the target county, as well as the broader spatial context in which each county exists. The inclusion of the timestamp in the input vector further enhances the model's ability to account for temporal trends and seasonality, which are crucial for accurate forecasting in spatiotemporal settings.

In summary, our approach leverages a combination of feature vectors, graph convolution vectors, and temporal information to predict the derivative of the target variable. The input matrix, with a size of 345 × 505, provides a comprehensive representation of the spatiotemporal dynamics at play, facilitating a deeper understanding of the underlying mechanisms driving changes in the system.

Feature Selection. Employing the K-best method in conjunction with mutual information enhances model interpretability and accelerates execution. This technique for selecting univariate features assigns scores according to the information they provide about the target variable and then selects the top K features with the highest scores. The first input vector of the model comprises both original features and graph convolution vectors. The algorithm then assesses the significance of each attribute in forecasting temporal changes. Retaining just the 20 most significant qualities simplifies the calculations and enhances the comprehensibility of the findings((Table 1).). This strategy facilitates the identification of the most critical components for generating forecasts.

Sparse Regression via Lasso. The sparse regression technique adopted in SINDy framework is the Least Absolute Shrinkage and Selection Operator (Lasso) [**?**]. Its goal is to identify a parsimonious set of governing equations for the temporal evolution of the target variable. Lasso regression imposes an $\mathcal{L}1$-norm penalty on the regression coefficients, effectively driving insignificant coefficients to zero. This characteristic is crucial for SINDy, as it allows the model to discard unnecessary terms and retain only the most influential features governing the system's dynamics. The resulting equations are not only compact but also interpretable, making it easier to understand the role of key variables in shaping the system's behavior.

Lasso is particularly well-suited for our problem due to its ability to handle high-dimensional data while enforcing sparsity in the model. Lasso tends to shrink some coefficients to zero, eliminating less essential features and reducing the model's complexity. Although we only retained 10 features in the feature selection, the libraries adopted by the SINDy framework will increase the number of terms in the final equation significantly. For example, we adopted the second-order polynomial library in our experiment. This will increase the number of terms from 20 to 420, with 20

Table 1: The top 20 features identified by KBest mutual information.

x_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
sp_bovine	sp_equine	LC110	LC190	LC200	LC210	LC40	AWC3	S_GRAVEL	CATTLE	G1_sp_equine

x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}
G1_area	G1_WC1	G1_LC11	G1_LC120	G1_LC180	G1_LC80	G1_S_SILT	G1_T_SAND	G1_CATTLE

first-order terms and 400 second-order terms. By leveraging Lasso within the SINDy framework, we extract governing equations that provide meaningful insights into the temporal evolution of the target variable while mitigating overfitting and enhancing model interpretability.

In our implementation, the regularization parameter for Lasso is set to $\lambda = 0.1$. This parameter controls the strength of the sparsity constraint: a higher value would allow more terms to be retained, potentially reducing interpretability. Our choice of λ balances sparsity and accuracy, ensuring that the identified equations remain both informative and concise.

5 Results

Using ecological and geographically structured data, we created three variants of the Sparse Identification of Nonlinear Dynamics (SINDy) model to investigate the transmission of Vesicular Stomatitis Virus (VSV). Every model underwent assessment through region-based cross-validation, utilising a variety of regularisation parameters (α). Performance was quantified by cross-validated mean squared error (CV-MSE) and the count of active terms in the final equation. The objective was to analyse the impact of local ecological factors in relation to spatially-informed graph characteristics on the formation of the symbolic structure of outbreak dynamics.

To integrate spatial structure, we subsequently broadened the candidate feature space to encompass graph-derived variables that reflected neighborhood-averaged livestock and environmental conditions. Following the execution of the feature selection that involved ecological and graph-based variables, we generated a new SINDy equation at $\alpha = 0.01743$, which includes five active terms and demonstrates a significantly reduced CV-MSE of 0.426.

Understanding the significance of spatial relationships in the transmission of infectious diseases, we generated a third equation that maintained the integrity of spatial graph features throughout the modelling process. This enabled us to evaluate if the context at the neighbourhood level improves symbolic interpretability and predictive accuracy. With α set at 0.0329, the derived equation reached the minimum CV-MSE of 0.425 and comprised nine terms. A significant spatial interaction term was identified, which involved the product of graph-aggregated bovine density (**G1_sp_bovine**) and local bovine density. This term highlights that the risk of outbreaks can increase not just due to factors within a specific area, but also because of comparable factors in neighbouring areas, like high livestock density or common environmental conditions. The equation also included the spatial term, along with interactions between different livestock species and soil composition (e.g., sand and clay), as well as their co-occurrence with other animals like horses and mules. This comprehensive framework encompasses various factors influencing VSV dissemination, integrating ecological specificity with spatial continuity.

Figures 2 and 3 illustrate the variation in model behaviour across different regularisation values. Figure 2 illustrates that with an increase in α, the number of terms in the region-based model decreases sharply from more than 1000 to under 10, whereas the models that included graph features consistently exhibited a lower number of terms across all α values.So the model can get at a meaningful structure faster when it has spatial information. Figure 3 shows how CV-MSE has changed over time. The eq created using spatial characteristics shows a more stable and lower error curve, with a large plateau around$\alpha = 0.03$, where it reaches its lowest CV-MSE. This behaviour reinforces the chosen model, validating that the integration of spatial dependencies improves robustness and generalisation.

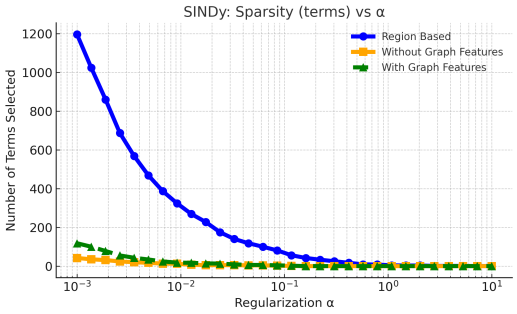

Figure 2: SINDy Sparsity (terms selected) versus α for the region-only model, graph-available model, and graph-inclusive model.

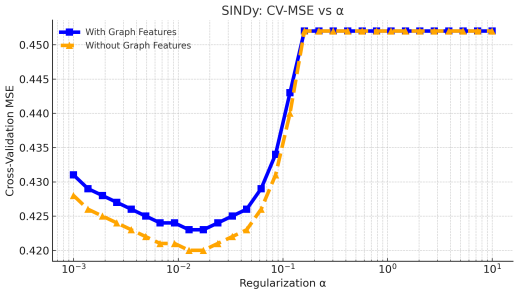

Figure 3: SINDy CV-MSE versus α for models with graph features included and retained. The graph-inclusive model shows the lowest error at $\alpha = 0.0329$.

Table 2: SINDy Equation Structures Across Features Based

Final SINDy Equations for VSV Outbreak Modeling		
Region-Based SINDy Equation	**Without Graph Features**	**With Graph Features**
$\mathbf{dy} = -1.386e+00 \cdot \text{sp\_equine}^2$ $- 6.186e-01 \cdot \text{sp\_bovine} \cdot \text{AWC1}$	$\mathbf{dy} = -3.070e-02 \cdot \text{sp\_bovine}$ $- 4.965e-02 \cdot \text{sp\_equine}$ $- 5.635e-03 \cdot \text{sp\_bovine}^2$ $- 9.120e-02 \cdot \text{sp\_bovine} \cdot \text{sp\_equine}$ $- 1.080e-02 \cdot \text{sp\_bovine} \cdot \text{LC190}$	$\mathbf{dy} = -3.781e-04 \cdot \text{sp\_equine}$ $- 4.781e-03 \cdot \mathbf{G1\_sp\_bovine} \cdot \text{sp\_bovine}$ $- 7.812e-03 \cdot \text{sp\_bovine}^2$ $- 6.401e-02 \cdot \text{sp\_bovine} \cdot \text{sp\_equine}$ $- 4.093e-03 \cdot \text{sp\_bovine} \cdot \text{t\_sand}$ $- 5.322e-02 \cdot \text{sp\_bovine} \cdot \text{horses}$ $- 2.292e-02 \cdot \text{sp\_equine} \cdot \text{s\_clay}$ $- 1.737e-02 \cdot \text{sp\_equine} \cdot \text{horses}$ $- 5.315e-03 \cdot \text{sp\_equine} \cdot \text{mules}$

6 Conclusion

provides a powerful tool for discovering interpretable, sparse dynamical models on graph-structured data. By incorporating graph-based features and operators, we can capture complex spatial-temporal dynamics while maintaining the benefits of sparsity and interpretability inherent in the SINDy approach. The theoretical analyses presented provide guarantees on the convergence of the sparse regression algorithm and the stability of the discovered systems. These analyses, combined with the computational considerations, provide a solid foundation for applying to a wide range of complex systems on graphs.

References

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009