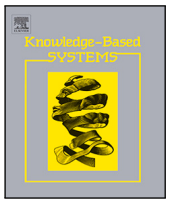
Knowledge-Based Systems 324 (2025) 113826

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



Predicting spatio-temporal dynamics in aquaculture networks: An extended Katz index approach

Michael-Sam Vidza [a,*], Marcin Budka [a], Wei Koong Chai [a], Mark Thrush [b], Mickaël Teixeira Alves [b]

[a] Bournemouth University, Computing & Informatics, Fern Barrow, Poole, Bournemouth, BH12 5BB, Dorset, United Kingdom
[b] Centre for Environment, Fisheries and Aquaculture Science (Cefas), Barrack Road, Weymouth, DT4 8UB, Dorset, United Kingdom

ARTICLE INFO

Keywords:
Link prediction
Spatial networks analysis
Similarity-based link prediction method
Aquaculture surveillance

ABSTRACT

The effective surveillance of the distribution of live fish between aquaculture farms is crucial for maintaining food security and preventing disease outbreaks. However, existing conventional models often assume the network is static and do not incorporate other factors that contribute to movement between farms, lacking the ability to accurately predict future movements, especially given the dynamic interactions within aquaculture networks. This study addresses this gap by developing the Edge-Weighted Katz Index (EWKI), an extension of the traditional Katz index that integrates spatial information to improve the accuracy of predicting fish distribution between farms. Using a comprehensive dataset on the distribution of live fish between farms in England and Wales from the year 2010 and 2023, the study evaluates the performance of the EWKI model in comparison to other similarity-based link prediction methods. The results indicate that the EWKI model significantly outperforms other methods, achieving a precision of 92.89%, a recall of 81.09%, and an F1-score of 86.59%, alongside an AUPR of 93.44% and an AUROC of 99.97%. This research has practical implications, as the developed method can accurately predict the distribution of fish between farms, supporting predictions of disease spread and facilitating targeted interventions. Furthermore, the integration of spatial information into the network analysis has broader applications across various fields where understanding and predicting spatially influenced network dynamics are crucial, including transportation networks.

1. Introduction

Network science, an interdisciplinary field that studies the structure of various interconnected systems such as biological [1], social [2], computer [3], transport [4], and climate [5] is commonly used to represent and understand how these complex systems function. The systems can be represented as a network. This network is made up of nodes, which symbolise the various entities involved. These nodes are connected by links, representing the relationships and interactions between the entities. Network modelling permits the analysis of underlying structures in the system and provides insights into their inherent dynamics. In the early 20th century, sociologists Jacob Moreno and Helen Jennings began using sociograms, visual representations of social relationships, as a tool to analyse and map human interactions [6]. Despite the visual nature of these sociograms, questions about missing links and potential future connections within these networks were raised. Kleinberg's work on decentralised search algorithms within complex networks, particularly this exploration of 'small-world' phenomena, significantly advanced our understanding of network navigation and information flow [7–9]. This has had profound implications for fields from computer science to sociology, influencing the design of link prediction algorithms and providing insights into social network dynamics.

Link prediction, which estimates the likelihood of a connection forming between two nodes based on existing links and node attributes [10], has become a cornerstone of network science. Its applications span numerous domains, including social network friend recommendation [11,12], protein–protein interaction (PPI) [13–15], transport planning [16], and e-commerce recommendation [12,16]. This study focuses on the application of network analysis in aquaculture. Network analysis has recently found increasing relevance in aquaculture, particularly in analysing distribution of live fish between farms, representing their movement and disease transmission. Aquaculture is recognised as an important contributor to world food production [17] and also supports the supply of fish for purposes such as restocking

* Corresponding author.
E-mail addresses: mvidza@bournemouth.ac.uk (M.-S. Vidza), mbudka@bournemouth.ac.uk (M. Budka), wchai@bournemouth.ac.uk (W.K. Chai), mark.thrush@cefas.gov.uk (M. Thrush), mickael.teixeiraalves@cefas.gov.uk (M.T. Alves).

https://doi.org/10.1016/j.knosys.2025.113826
Received 8 October 2024; Received in revised form 30 April 2025; Accepted 20 May 2025
Available online 5 June 2025

water bodies for recreational fishing and enhancing the aesthetic value of certain environments [18]. Effective surveillance in aquaculture is a key biosecurity measure supporting food security [19]. By representing fish farms and processing facilities as nodes and the movement of live fish as links, stakeholders can better understand the dynamics of fish distribution and identify potential bottlenecks or inefficiencies in the supply chain [20]. For instance, network analysis can reveal the central hubs in fish distribution networks, where interventions might be most effective in mitigating the spread of diseases. Studies by Green et al. [21], Tidbury et al. [22] and Murray et al. [23] have explained the structure of fish distributions and their complex interactions. These studies have contributed to the development of epidemiological and management models applied to aquaculture networks [24,25]. However, current models used in aquaculture lack predictive power because they are static models, which limits their ability to respond to the dynamic nature of the distribution of live fish. Although link prediction models are widely applied in other fields, their adoption in aquaculture remains limited.

Several studies have addressed gaps in link prediction models by developing novel algorithms. Benhidour et al. [26] proposed a method combining similarity-popularity and path patterns to improve link prediction in directed networks, while Xu et al. [27] introduced new link prediction method based on local random walks and Jensen–Shannon divergence for hyperlink prediction. Furthermore, Chen et al. [28] developed an enhanced local path index to address limitations in existing local path-based methods, and Li et al. [29] proposed new metrics incorporating community detection information for improved accuracy. While these advancements have proven effective in various domains, they often fall short in capturing the specific spatio-dynamics critical in applications in fields like aquaculture.

To address this gap, our study extends existing similarity-based link prediction algorithms – which are relatively more interpretable, computationally efficient, and suitable for sparse networks – by incorporating a spatial weight factor, aiming to provide a more agile tool that can adapt to the changing dynamics between fish farms during the distribution of live fish. For example, following an exotic disease incursion, which would result in affected farms being placed under movement restrictions by official services [30]. Among the various link prediction models, similarity-based models have the potential to enhance predictive capacities by using records on the distribution of live fish between farms. These models are founded by the concept of homophily, which suggests a tendency for connections to form between entities that share similar characteristics or relationships [31,32]. This principle is widely observed in real-world networks, where individuals with shared attributes are more likely to interact [33–35]. In the context of aquaculture, homophily suggests that farms with similar characteristics (e.g., species farmed, biosecurity measures, or proximity) are more likely to trade in the distribution of live fish. Similarity-based models are categorised into local, global, and quasi-local indices, each providing unique insights into network dynamics. Local indices, such as common neighbours [36] and Jaccard's coefficient [37], focus on the immediate neighbourhood of nodes. While these approaches offer insights based on direct connections between nodes, they may overlook the broader network dynamics and, as a result, limit the understanding of the overall interactions in the network. In contrast, global indices like the Katz index [38] consider the entire network structure, including both direct and indirect connections. However, this comprehensive approach comes with the drawback of higher computational demands. Quasi-local indices, like the local random walk [39], offer a compromise between the strictly local and fully global approaches. Despite their balanced approach, quasi-global models can sometimes inherit the limitations of both local and global indices. These limitations include reduced accuracy in extremely sparse networks or higher computational requirements than purely local models [40]. Additionally, tuning these models to optimally balance local and global information can be challenging and often requires extensive empirical testing [31,41].

The Katz index has been proven effective in many fields [31,32,42] but its application in aquaculture (and more widely in terrestrial livestock production) remains limited. However, the conventional Katz index does not consider the spatial characteristics within a network and relies solely on walks within the network, biasing predictions. Specifically, in aquaculture, farms that are geographically closer are more likely to have frequent movements between them, increasing the potential for disease spread [43,44]. This study acknowledges the complexity involved in the distribution of live fish between farms and the limitations of current models in predicting future movement considering the proximity factor between farms. It builds upon previous studies by Tidbury et al. [22] and Guilder et al. [25], and utilising data on the distribution of live fish between farms collected by the Cefas Fish Health Inspectorate for England and Wales.

The paper introduces the Edge weighted Katz index (EWKI) link prediction model, an enhancement of the traditional Katz index which integrates spatial information to account for geographical proximity, providing a more accurate prediction of the distribution of live fish between aquaculture farms. By incorporating both spatial and temporal features of the aquaculture network, the study addresses the dynamic nature of distribution between farm, which static models cannot capture. In addition, the study performs a comparative analysis of the novel model with other similarity-based models to assess the performance of the models. While focused on aquaculture, the EWKI's framework is adaptable to other domains, such as epidemiology and transportation, where understanding spatially influenced network dynamics is essential. The remainder of this paper is structured as follows: Section 2 describes the data, the prediction model, and the evaluation metrics. Section 3 discusses the results, including a comparative analysis of similarity-based methods and the implications of incorporating spatial dynamics into aquaculture networks. Finally, Section 4 concludes by summarising key findings and their broader relevance.

2. Method

2.1. Data

In this study, two datasets were utilised to evaluate the performance of the proposed approach: the live fish distribution network and the road network of the city of Peterborough [45], United Kingdom. The live fish distribution dataset comprises of 2,480 unique nodes (fish farms) and 4,696 directed edges (fish distributions) across England and Wales from the year 2010 to 2023. Each record includes the source and destination farm identifiers, geographical coordinates (longitude and latitude), the species distributed, and the year of movement. The distance between farms is calculated using the vincenty inverse method which accounts for the accurate earth model (ellipsoidal), particularly for long distances. This inherently represents the spatio-temporal dynamics of fish farm interactions.

The road network dataset is a spatially embedded, distance weighted directed network with 7,188 nodes (representing intersections or endpoints of roads) and 14,763 edges (road segments). This dataset was added to the experiment to evaluate the generalisability and robustness of the proposed link prediction method across networks with varying topological and spatial characteristics. Unlike the live fish distribution network, which incorporates temporal dynamics, the road network has a static structure that allows for the assessment of the EWKI's adaptability to networks where spatial relationships are the primary influencing factor. Table 1 summarises the basic topological properties of both datasets, highlighting their structural differences and the variety of network dynamics captured.

Table 1
Topological properties of dataset.

Property	Live fish distribution network	Peterborough road network
Degree	6.0	4.0
Average In-Degree	3.0	2.0
Average Out-Degree	3.0	2.0
Max diameter	7	102
Density	0.0042	0.0003
Average clustering coefficient	0.3565	0.04
Overall eigenvector centrality	0.0276	0.0011
Assortativity	−0.16	−0.0944

2.2. Experiment setup

The live fish movement network was modelled as a directed graph $G = (V, E, w)$ where V represents the set of nodes, with each node corresponding to a fish farm and E represents the set of directed connections between these farms, indicating the distribution of fish. Each edge, denoted as $(u, v, t) \in E$, captures the distribution of live fish from farm u to farm v at a specific time t. To capture temporal dynamics, the live fish movement network was partitioned into a series of time-based snapshots, each representing the state of the network within a specific year. This approach models the network as a sequence of temporal graphs, $G = G_{2010}, G_{2011}, G_{2012}, \ldots, G_{2023}$, where each G_t corresponds to the network at year t. Such a representation enables the analysis of the network's evolution over time and facilitates dynamic link prediction by capturing changes in fish distribution patterns across different periods. In this study, the snapshots were divided into three temporal intervals: a training set ($G_{train} = G[2010, 2021]$), validation set ($G_{val} = G[2022, 2022]$), and a test set ($G_{test} = G[2010, 2021]$) as represented in Fig. 1. This temporal snapshot partitioning ensures that the model is trained on G_{train}, validated on G_{val}d to monitor for over-fitting and assist with hyperparameter tuning, and tested on G_{test} to evaluate its generalisation ability. G_{test}, leverages historical data to evaluate the model's ability to predict future connections. This partitioning, aligned with the methodology outlined by Liben-Nowell and Kleinberg [32], facilitates a robust evaluation of the model's ability to generalise across different temporal subsets, ensuring it performs well on unseen data. Specifically, for times $t_i \leq t_j$, the subgraph $G[t_i, t_j]$ contains only edges with timestamps from t_i to t_j.

For the road network, which lacks temporal features, K-fold cross-validation was employed to divide the dataset into training, validation, and test sets. The network was split into K equally sized folds, with each fold iteratively used as a test set while the remaining $K - 1$ folds were further divided into training and validation subsets. Given the larger size of this network compared to the live fish distribution network, 10-fold cross-validation was chosen as it offers an optimal trade-off between computational cost and model performance [46]. This approach ensures that all parts of the network are systematically used for training, validation, and testing, mitigating potential biases, and reducing variance. Compared to random sampling validation, this method provides a more rigorous and consistent evaluation of model performance, especially for the larger and more complex road network.

2.3. Similarity-based methods and scoring

Within the network, each pair of nodes, u and v, is evaluated for its potential to establish a connection. This evaluation is based on a similarity score derived from various indices, reflecting the likelihood of a link forming between them. In most link prediction techniques [32,47,48], the task is treated as a ranking problem where node pairs are ordered based on their similarity scores. A threshold is established to classify pairs into positive or negative instances. Node pairs with scores above the threshold are predicted to form links (positive instances), while those below the threshold are not (negative instances). In this study, we used the precision–recall (PR) curve [49,50] to determine the optimal threshold value by evaluating the index on G_{val}. The PR curve plots the precision (the proportion of true positive predictions among all positive predictions) against the recall (the proportion of true positive predictions among all actual positives) for different threshold values. By analysing this curve, we can select a threshold that maximised the F1-score, a harmonic mean of precision and recall. This threshold was subsequently used to classify predictions in G_{test} by comparing predicted edges to actual links. Five similarity-based link prediction methods are analysed in this study:

1. The common neighbours (CN) index is a fundamental measure in link prediction that evaluates the likelihood of a link forming between two nodes, u and v, by counting their mutual neighbours [36]. The rationale behind this index is based on the principle that the presence of a greater number of shared neighbours between two nodes increases the likelihood of a direct link forming between them. This concept is based on the idea that common associations foster connectivity within the network, making it a commonly used metric for predicting future links in different network structures. It is mathematically represented as:

$$\mathrm{CN}_{(u,v)} = |\Gamma(u) \cap \Gamma(v)| \tag{1}$$

   where $\Gamma(u)$ and $\Gamma(v)$ denote the sets of neighbours of node u and v.
2. Adamic-Adar Index (AAI) [12]: This index assigns weights to common neighbours based on their connectivity. Neighbours with fewer connections have a higher contribution to the score, as it is assumed that having a less common neighbour indicates a stronger bond. Mathematically, the AAI is expressed as:

$$\mathrm{AAI}_{(u,v)} = \sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{1}{\log k_z} \tag{2}$$

   where z is the node that is a common neighbour of both u and v. k_z is the degree of node z.
3. Local path index (LPI) measures the similarity between two nodes by considering the local paths – direct connections and indirect connections via a common neighbour – that exist between them. This index balances computational efficiency and accuracy [41].

$$\mathrm{LPI}_{(u,v)} = A^2_{(u,v)} + \varepsilon A^3_{(u,v)} \tag{3}$$

   where ε represents a free parameter, $A^2_{(u,v)}$ indicates the number of paths of length 2 between nodes u and v, and $A^3_{(u,v)}$ corresponds to the number of paths of length 3.
4. The Katz index (KI) [38] is a walk-dependent index, focusing on both direct and indirect walks between nodes. The KI calculates the sum of all unweighted walks between two nodes within a network, applying an exponential damping factor to prioritise shorter walks. The KI is expressed as:

$$\mathrm{KI}_{(u,v)} = \sum_{l=1}^{\infty} \beta^l |\mathrm{walks}^{\langle l \rangle}_{(u,v)}| \tag{4}$$

$$= \sum_{l=1}^{\infty} \beta^l (A^l)_{(u,v)} \tag{5}$$

$$= \beta A_{(u,v)} + \beta^2 (A^2)_{(u,v)} + \beta^3 (A^3)_{(u,v)} + \cdots \tag{6}$$

   where $|walks^{\langle l \rangle}_{(u,v)}|$ represents the total number of walks with length l between nodes u and v, A^l is the adjacency matrix of the network raised to the power of l, the damping factor β controls the influence of walk length. A lower β value reduces the contribution of longer walks to the Katz index score, while a higher value increases their contribution. To ensure the Katz index converges, β must be assigned a value smaller than the

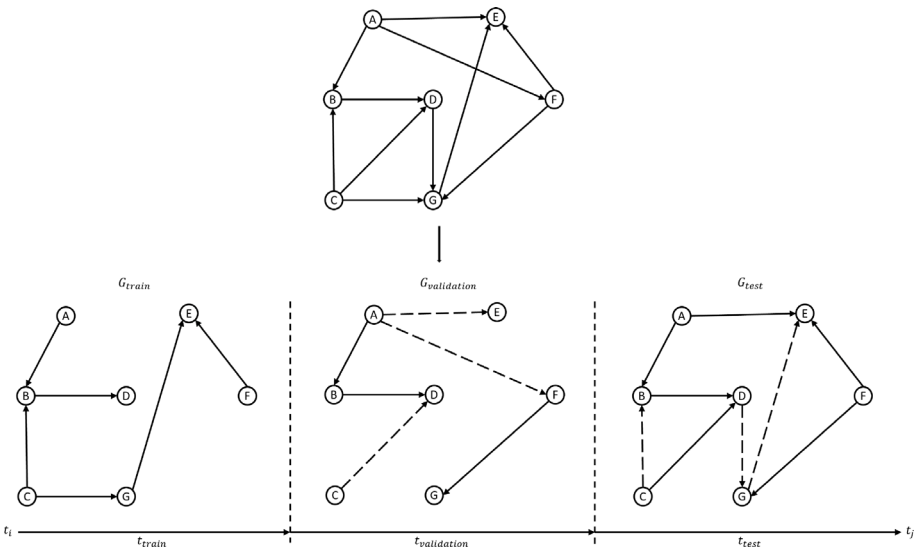


Fig. 1. Temporal snapshots of live fish movements network over time. The nodes represent individual fish farms. Solid lines indicate observed movements of fish between farms within the time interval (t_i to t_j) and dotted lines denote the predicted future movements for the interval. The horizontal axis represents the timestamps, showing the evolution of the network: training (G_{train}), validation (G_{val}), and test (G_{test}) sets.

inverse of the largest eigenvalue *lambda* of the adjacency matrix A.

$$\beta < \frac{1}{\lambda_{max}} \tag{7}$$

This condition guarantees the convergence of the infinite series that defines the Katz index. In this study, we use the upper bound value of β, calculated directly from Eq. (7), to ensure both mathematical stability and optimal utilisation of the model's expressive capacity without risking divergence.

5. The weighted Katz index (WKI) replaces the adjacency matrix in the original Katz index formula with the distance values between fish farms. This allows the model to account for the varying degrees of actual distance interaction or influence in the network. The distances are summed with a damping factor applied to longer walks, ensuring that shorter walks retain higher significance in the prediction model. The model reads:

$$\mathrm{WKI}_{(u,v)} = \sum_{l=1}^{\infty} \beta^l (A_w^l)_{(u,v)} \tag{8}$$

$$= \beta(A_w)_{(u,v)} + \beta^2(A_w^2)_{(u,v)} + \beta^3(A_w^3)_{(u,v)} + \cdots, \tag{9}$$

where A_w is the weighted adjacency matrix of the network which details the distance between farms.

6. In this study, we developed the edge-weighted Katz index (EWKI), a novel extension of the traditional Katz index designed to incorporate temporal dynamics into link prediction for spatially embedded networks. The spatial component is represented by the weight ω, defined as an exponential function of the distance $d_{(u,v)}$ between nodes, ensuring that geographically closer nodes are given higher priority. This reflects real-world patterns observed in aquaculture networks, where proximity strongly influences the movement of live fish. The decay factor γ controls the influence of distance in this weighting scheme. To identify an appropriate value for γ, we used a grid search approach [51,52], testing a predefined range of values ($\gamma \in 0.0015, 0.001, 0.015, 0.01, 0.15, 0.1$). Each candidate value was evaluated on the validation snapshot by computing the corresponding F1-score, and the threshold that yielded the highest F1-score was selected for the final model. This procedure ensured that the chosen γ optimally balanced the model's precision and recall in predicting links. The temporal dynamics described in Section 2.2 are embedded in the framework by modelling the network as a sequence of evolving graphs, with each snapshot capturing node interactions over time. This temporal segmentation allows the EWKI to account for changes in network structure and interaction patterns, moving beyond static configurations. The model reads:

$$\mathrm{EWKI}_{(u,v)} = \omega_{(u,v)} \cdot \sum_{l=1}^{\infty} \beta^l |\mathrm{walks}_{(u,v)}^{\langle l \rangle}| \tag{10}$$

$$\omega_{(u,v)} = e^{-\gamma \times d_{(u,v)}} \tag{11}$$

2.4. Combination of link prediction methods

The combination of link prediction methods in this study seeks to enhance predictive accuracy by leveraging the strengths of multiple indices that capture network properties, as each similarity index focuses on specific aspects of the network. This approach of combining predictions from multiple models is grounded in the principles of ensemble learning [53,54]. Ensemble methods, widely used in machine learning, have been shown to improve predictive performance and generalisability by reducing variance and bias [55]. In the context of link prediction, combining different methods can lead to a more accurate model by capturing diverse aspects of network structure and dynamics, particularly in sparse or complex networks where some single metrics often underperform [31,32,48]. While various sophisticated ensemble techniques exist, the adoption of an averaging approach is motivated by its simplicity, computational efficiency, and demonstrated effectiveness in prior studies. Averaging provides a straightforward method to integrate the outputs of multiple models without requiring additional hyperparameter tuning, making it well-suited for scenarios where computational resources or dataset size are limiting factors [54,55]. Moreover, it avoids over-fitting, a risk in more complex ensemble techniques, and ensures that the combined score is interpretable, a key requirement for practical applications in aquaculture networks.

The averaged score for each potential edge (u, v) will be calculated as:

$$\mathrm{C} = \frac{S^1_{(u,v)} + S^2_{(u,v)}}{2} \tag{12}$$

where C represents the composite score, $S^1_{(u,v)}$ and $S^2_{(u,v)}$ are the scores from the first and second similarity index, respectively.

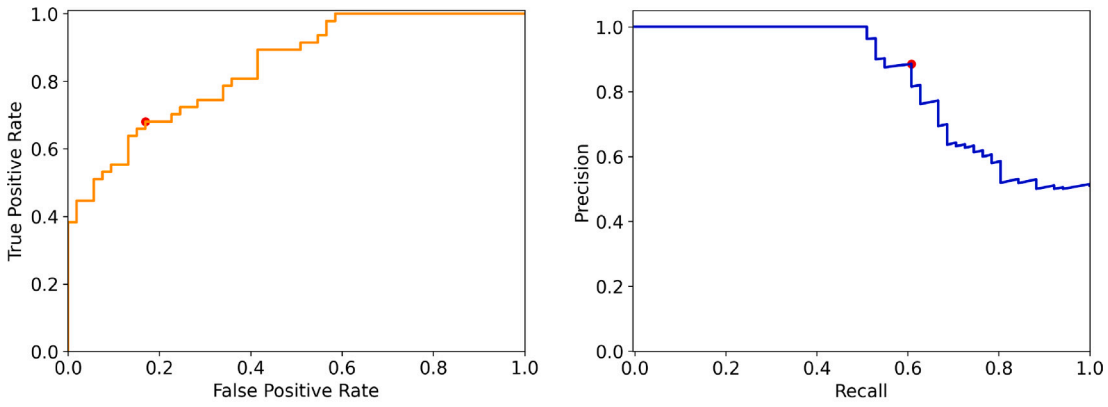


Fig. 2. Sample plots illustrating (a) Receiver Operating Characteristic curve (b) Precision–recall curve.

2.5. Evaluation metrics

Choosing the right evaluation metrics is essential for comparing the performance of the models, particularly in scenarios like ours where the dataset exhibits sparsity and class imbalance between existing and non-existing links. This study uses precision, recall, F1-score, area under the receiver operating characteristic curve (AUROC) and area under the precision–recall curve (AUPR) value to evaluate the performance of the model.

1. Precision: measures the fraction of correctly predicted positive links (i.e., actual distribution between farms) to all predicted positive links. In the context of our study, a high precision indicates that the model's predictions are predominantly accurate.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{13}$$

where TP = true positive, FP = false positive

2. Recall: measures the fraction of actual positive links that the model correctly identifies. A high recall in our study suggests that the model captures most genuine fish movements.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{14}$$

where FN = false negative

3. F1-score: is the harmonic mean of precision and recall, which means it gives a balanced measure of the two metrics. An F1-score close to 1 indicates both good recall and good precision, while an F1-score close to 0 indicates poor performance on both metrics.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{15}$$

4. AUROC: is a graphical representation of the trade-off between the true positive rate and the false positive rate at various threshold settings. The AUC value, which is the area under this curve, provides a single measure to summarise the ROC curve. It is useful for comparing the overall ranking performance of different models, irrespective of threshold choice. Fig. 2(a) is provided as a generic illustrative example to explain how the AUROC is computed and interpreted.
5. AUPR: is derived from the precision–recall curve and is especially informative for tasks like link prediction where class imbalance is significant. Unlike AUROC, AUPR focuses on the positive class, making it more relevant for evaluating performance in imbalanced datasets. Fig. 2(b) similarly serves as an illustrative sample and does not correspond to any specific model output in this study.

2.5.1. Optimal threshold

In the evaluation of models for link prediction, converting calculated scores into binary classification decisions is important for distinguishing between true and false links. This is usually done through thresholding and is challenging due to the high class imbalance often present in such networks [47,56,57]. In cases where predetermined classification thresholds are unavailable, it is common to use threshold curves like ROC (receiver operating characteristic) and PR (precision–recall) to determine the optimal threshold and assess model performance. However, when dealing with imbalanced datasets, the use of the ROC curve can result in overly optimistic results, since it may not effectively identify rare positive instances effectively. Conversely, the AUPR curve is better suited for imbalanced datasets, as it emphasises the precision–recall trade-off. For each method, F1-scores were calculated at various threshold levels. The F1-score offered a balanced metric suitable for datasets with significant class imbalances. The threshold yielding the highest F1-score was identified as the optimal threshold for each feature. By applying these optimal thresholds in the model, predicted links and non-links could be distinguished. This approach not only allowed for establishing thresholds that optimise model performance but also facilitated a detailed comparison among different Katz-based models.

3. Results & discussion

The performance of similarity-based link prediction methods was evaluated on both live fish distribution network and the road network. A baseline comparison with a random link predictor (RLP) provided a benchmark for evaluating the added value of incorporating network structure and spatial features. Tables 2 and 3 and Figs. 3 and 4 illustrate the results.

3.1. Comparative performance of the live fish distribution network

The RLP method performed the worst among all methods, with a precision of 0.3% and recall of 37.1%, resulting in an F1-score of 0.6%. Its AUPR of 0.1% and AUROC of 54.1% further highlighted its limitations. Despite a recall of 37.1% the high number of false positives (FP = 88,656) and false negatives (FN = 426) rendered the method ineffective. The CN method, with a precision of 22.6% and a recall of 2.1%, performed better than the RLP but still demonstrated limitations in predicting distributions between farms. The F1-score of 3.8% and AUPR of 1.1% reflect the limited effectiveness of the method in this context. The low recall indicates that CN fails to capture an important number of TPs, leading to a high number of FNs (FN = 663). Although the AUROC for CN is 54.1%, this value alone is misleading. The huge difference between AUPR and AUROC shows

that while CN may perform decently in distinguishing between positive and negative cases overall, it struggles with predicting TPs in this imbalanced dataset. This is evident from the low precision and recall. The performance of the method is also hindered by the sparse nature of the fish movement network, where direct neighbours are not sufficient to predict potential links accurately. Previous studies have similarly reported CN's limitations in sparse networks, where the lack of dense connectivity reduces the reliability of shared neighbours as predictors of new links [32,47,48,58].

The AAI, similar to CN, also demonstrated limited effectiveness in predicting the distribution of live fish between farms. Its precision of 31.8% was slightly higher than CN, but its recall was even lower at 1.0%. The F1-score of 2.0% and AUPR of 1.0% suggest that although AAI had slightly better precision than CN, it still struggled with recall, capturing very few TPs (TP = 7). This indicates that AAI, which is designed to enhance the predictive power of CN by giving more weight to less connected nodes, did not significantly improve link prediction in this specific network. The low AUPR and AUROC values further confirm shortcomings from AAI. The poor performance of AAI can be attributed to the same factors affecting CN, namely the sparsity of the network.

The LPI method, which considers both direct and indirect connections up to a path length of 3, demonstrated better performance compared to CN and AAI. While its recall was 16.2%, its precision was significantly lower at 6.5%. This suggests that LPI was more capable of identifying TP links but at the cost of generating a higher number of FPs (FP = 1,584). The improvement in recall compared to CN and AAI suggests that LPI, which considers both direct and indirect paths, is better suited for networks where indirect connections play a significant role. However, the low precision indicates that LPI may struggle with distinguishing between actual and potential links, leading to an increased number of FPs. This is consistent with findings in network theory, where local path-based methods are known to balance accuracy and computational efficiency [48].

The KI showed improvement compared to CN, AAI, and LPI. With a β value of 0.37, Its recall of 30.43% was the highest among the methods without spatial weights, indicating its ability to capture a greater proportion of TP links (206). Although its precision of 15.1% was low, the increase in recall resulted in a higher F1-score of 20.2%. The high AUROC value of 98.6% further supported KI's performance in distinguishing between positive and negative links. However, the AUPR of 12.8% revealed that KI still faced challenges in achieving better performance, particularly for the positive class. This was attributed to the class imbalance issue, emphasising the importance of considering both AUROC and AUPR when evaluating models in such scenarios.

With a β value of 0.37, the WKI achieved a precision of 18.2% and a recall of 8.4%, resulting in an F1-score of 11.5%. The AUPR of 12.7% was the same as the KI. This suggested that while the introduction of spatial weighting added some value in accounting for geographical proximity, it did not significantly improve the model's ability to predict TPs correctly. The AUROC for WKI was 98.8%, which, like KI, was very high but still struggled to accurately identify the minority class in an imbalanced dataset. The results of WKI indicated that simple distance-based weighting may not fully capture the complex relationship between distance and the distribution of live fish between farms.

The EWKI, a novel modification of the Katz index, outperformed all other methods, including the RLP. With a beta value of 0.37 and a gamma value of 0.01 (selected from the grid search), it achieved a precision of 92.9% and recall of 81.1%, resulting in the highest F1-score of 86.6%. The AUPR of 93.4% and AUROC of 99.9% further demonstrated its ability to accurately predict links while minimising false positives. However, it is important to interpret AUROC result within the specific characteristics of the dataset and the evaluation context. First, the live fish distribution network used in this study is highly imbalanced, with a vast majority of potential links representing non-movements (negative class) compared to actual live fish movements (positive class). In such settings, a high AUROC can arise because true negatives dominate the classification outcomes, thereby inflating the model's specificity and overall AUROC score [49]. High AUROC values in imbalanced datasets do not necessarily reflect strong performance in predicting the minority class (i.e., actual fish movements). Recognising this limitation, we complemented AUROC with more informative evaluation metrics for imbalanced datasets, including precision, recall, and F1-score. These metrics directly evaluate the model's ability to correctly identify positive links (true fish movements), providing a clearer view of performance beyond the majority class. This scenario aligns with established recommendations in the literature, where it has been shown that precision–recall metrics are often more informative than AUROC in highly imbalanced scenarios [29,31,49,50,58].

To understand the performance of EWKI, it is crucial to examine its key differences from the standard KI and the WKI. While the KI considers all paths between nodes, it lacks spatial awareness. The WKI addresses this by incorporating distance-based weights, but its linear weighting scheme does not fully capture the nuances of spatial interaction in aquaculture. In contrast, the EWKI employs an exponentially decaying weight function, prioritising local interactions while accounting for long-range connections. This approach aligns with observed network dynamics, where geographically closer farms are more likely to engage in fish distributions due to logistical and cost considerations.

3.2. Comparative performance on the road network

The performance of the similarity-based methods on the road network dataset (Table 3 and Fig. 4) reveals different trends compared to the live fish distribution network. The RLP exhibited minimal predictive capability, achieving a precision of 0.03%, recall of 11.8%, and an F1-score of 0.06%. Methods such as CN, AAI, and LPI demonstrated similarly limited effectiveness. Each of these methods identified only two true positive (TP = 2) while generating a significant number of false positives (FP $\geq$ 200) and false negatives (FN = 1,472). These results highlight the limitations of local and quasi-local methods in capturing the broader structural and spatial properties of the road network, where direct neighbour information is insufficient for accurate link prediction.

The KI method, using a decay parameter of β = 0.7, showed an improvement, achieving a precision of 66.67%, recall of 0.54%, and an F1-score of 1.08%. Its high AUPR of 88.5% and perfect AUROC of 99.99% reflect its ability to rank links effectively. However, while the ranking metrics suggest good discrimination between positive and negative links, the F1-score remained low due to a limited number of true positives (TP = 8) and a substantial number of false negatives (FN = 1,466).

The WKI method, with a decay parameter of β = 0.002, achieved perfect precision (100%), indicating that all predicted links were true positives, but its recall of 2.37% and F1-score of 4.64% also reveal the limitation experienced in other methods, that is, the method identified 35 true positive while failing to capture most actual links (FN = 1,439). The perfect AUROC and AUPR (98.2%) reflect the model's ability to rank links effectively, but these metrics are not indicative of practical utility when the F1-score is so low. This result highlights the importance of examining every metric alongside the raw counts to ensure a comprehensive understanding of model performance.

As seen in the performance of EWKI in the live fish distribution network, the EWKI method outperformed all other models. With $\beta = 0.7$ and γ = 0.01 achieving a prefect precision of 100%, recall of 72.80%, and an F1-score of 84.26%. Its high recall indicates that it captured nearly all actual links in the. Its perfect AUROC and AUPR of 99.1% further affirm its strong performance across both ranking and classification tasks. EWKI's superior performance can be attributed to its exponential weighting scheme, which integrates spatial proximity into the prediction process, enabling the model to prioritise geographically relevant links effectively.

Table 2
Performance of various link prediction methods on the live fish distribution network.

Method	Precision (%)	Recall (%)	F1-score (%)	TP	FP	TN	FN
RLP	0.3	37.1	0.6	251	88,656	156,187	426
CN	22.6	2.1	3.8	14	48	244,795	663
AAI	31.8	1.0	2.0	7	15	244,828	670
LPI	6.5	16.2	9.3	110	1,584	243,259	567
KI ($\beta = 0.37$)	15.1	30.4	20.2	206	1,157	243,686	471
WKI ($\beta = 0.003$)	18.2	8.4	11.5	57	257	244,586	620
EWKI ($\beta = 0.37$, $\gamma = 0.01$)	92.9	81.1	86.6	549	42	244,801	128

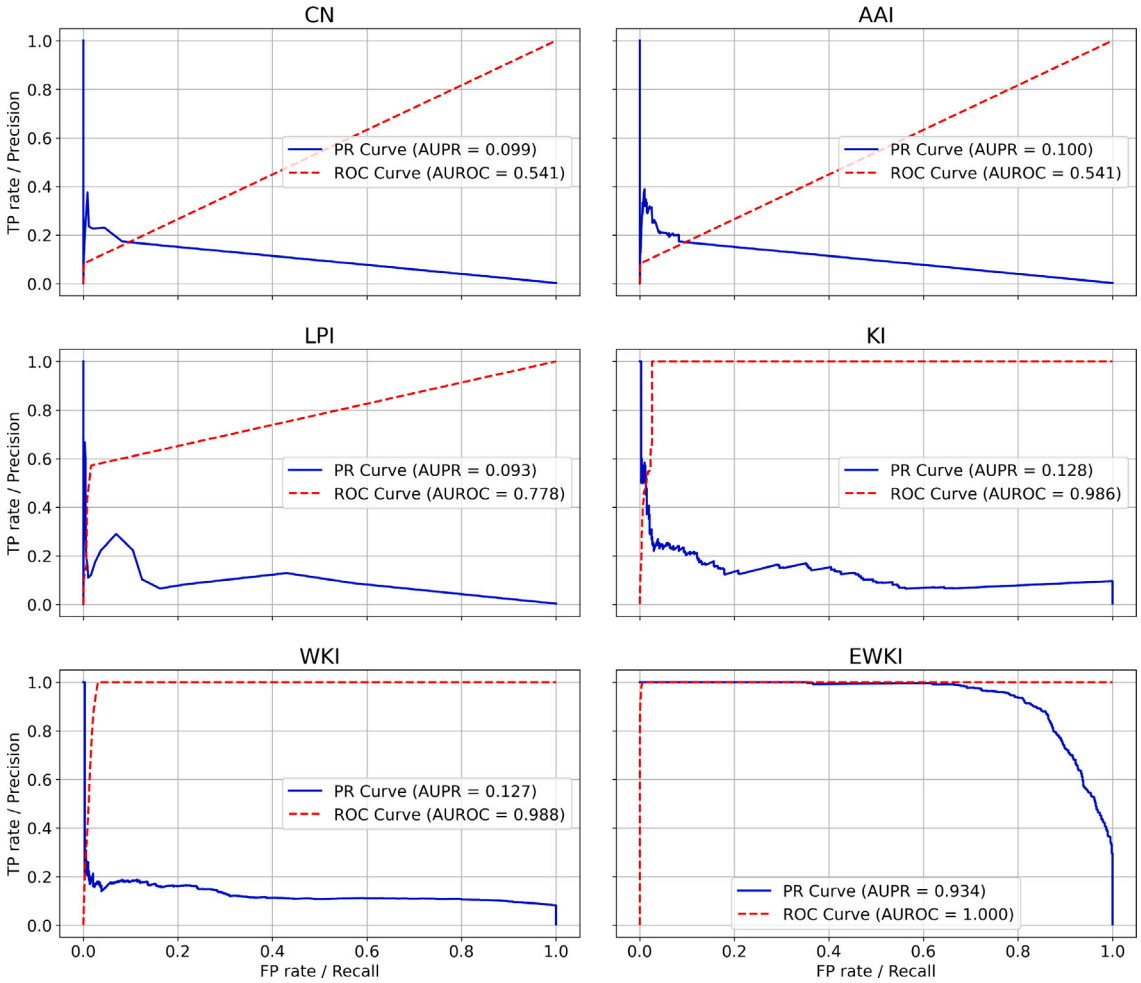


Fig. 3. Performance of ranking metrics (AUPR and AUROC) of the various link prediction methods on the live fish distribution network. Precision–recall curves (solid blue line) and receiver operating characteristic curves (dashed red line) are shown for each method, along with their respective values.

Table 3
Performance of various link prediction methods on the road network.

Method	Precision (%)	Recall (%)	F1-score (%)	TP	FP	TN	FN
RS	0.03	11.80	0.06	174	583,899	5,244,437	1,300
CN	0.99	0.14	0.24	2	200	5,828,136	1,472
AAI	0.99	0.14	0.24	2	200	5,828,136	1,472
LPI	0.89	0.14	0.24	2	223	5,828,113	1,472
KI ($\beta = 0.7$)	66.67	0.54	1.08	8	4	5,828,332	1,466
WKI ($\beta = 0.002$)	100.00	2.37	4.64	35	0	5,828,336	1,439
EWKI ($\beta = 0.7$, $\gamma = 0.01$)	100.00	72.80	84.26	1,073	0	5,828,336	401

3.3. Comparative performance of combined indices

The combination of various similarity-based link prediction methods was conducted on the live fish distribution network to evaluate whether combining indices, particularly EWKI, with other methods could enhance link prediction performance. This approach aimed to determine if model averaging could leverage the strengths of individual indices to achieve improved predictive accuracy. Table 4 and Fig. 5 summarise the performance metrics.

3.3.1. Combination with CN

The combination of CN with other indices resulted in slight improvements in predictive performance compared to CN alone, but the overall gains were minimal. For instance, when combined with AAI, CN achieved a precision of 24.4% and a recall of 3.2%, resulting in

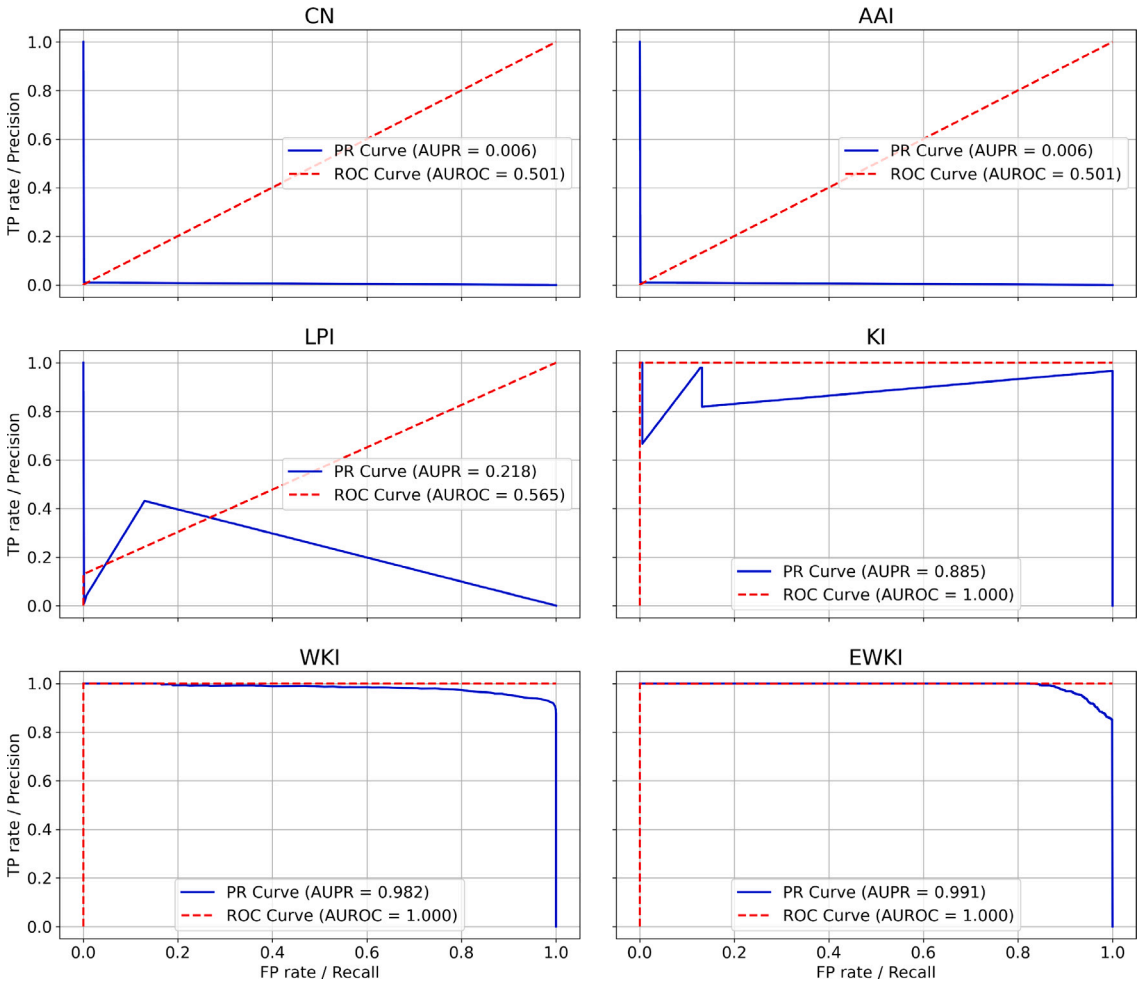


Fig. 4. Performance of ranking metrics (AUPR and AUROC) of the various link prediction methods on the road network. Precision–recall curves (solid blue line) and receiver operating characteristic curves (dashed red line) are shown for each method, along with their respective values.

Table 4
Performance of various combined link prediction methods on the live fish distribution network.

Method	Precision (%)	Recall (%)	F1-score (%)	TP	FP	TN	FN
CN*AAI	24.4	3.2	5.7	22	68	244,775	655
CN*LPI	27.7	3.8	6.7	26	68	244,775	651
CN*KI	22.2	2.1	3.8	14	49	244,794	663
CN*WKI	22.6	2.1	3.8	14	48	244,795	663
CN*EWKI	23.7	4.6	7.7	31	100	244,743	646
AAI*LPI	20.9	5.8	9.0	39	148	244,695	638
AAI*KI	34.8	1.2	2.3	8	15	244,828	669
AAI*WKI	34.8	1.2	2.3	8	15	244,828	669
AAI*EWKI	40.0	1.5	2.8	10	15	244,828	667
LPI*KI	8.4	13.1	10.3	89	966	243,877	588
LPI*WKI	7.7	14.5	10.0	98	1,177	243,666	579
LPI*EWKI	6.5	16.2	9.3	110	1,584	243,259	567
KI*WKI	15.2	21.3	17.7	144	806	244,037	533
KI*EWKI	69.3	80.4	74.4	544	241	244,602	133
WKI*EWKI	67.3	94.1	78.5	637	309	244,534	40

an F1-score of 5.7%. This marginal improvement was due to AAI's focus on less connected nodes, which complemented CN's reliance on local neighbourhoods. However, the sparse nature of the aquaculture network limited the overall effectiveness of both methods, as reflected in the low AUPR of 2%. Combining CN with LPI further improved the recall to 3.8% and the F1-score to 6.7%, highlighting the advantage of LPI's ability to capture indirect connections. The CN*LPI combination achieved an AUPR of 8% and an AUROC of 78%, indicating a better balance between precision and recall. However, despite this improvement, the high number of false negatives (FN = 651) showed that CN*LPI still struggled to identify true links accurately. The combination of CN with KI and WKI maintained precision similar to CN alone but slightly improved the AUPR to 13%, reflecting the added value of global structural information from KI and the incorporation of spatial factors in WKI. However, these combinations still exhibited low recall and F1-scores, indicating their limited effectiveness in the highly imbalanced dataset.

The most notable improvement came from combining CN with EWKI. This combination achieved a precision of 23.7%, recall of 4.6%, and an F1-score of 7.7%. The AUPR of 49% and AUROC of 99.99%

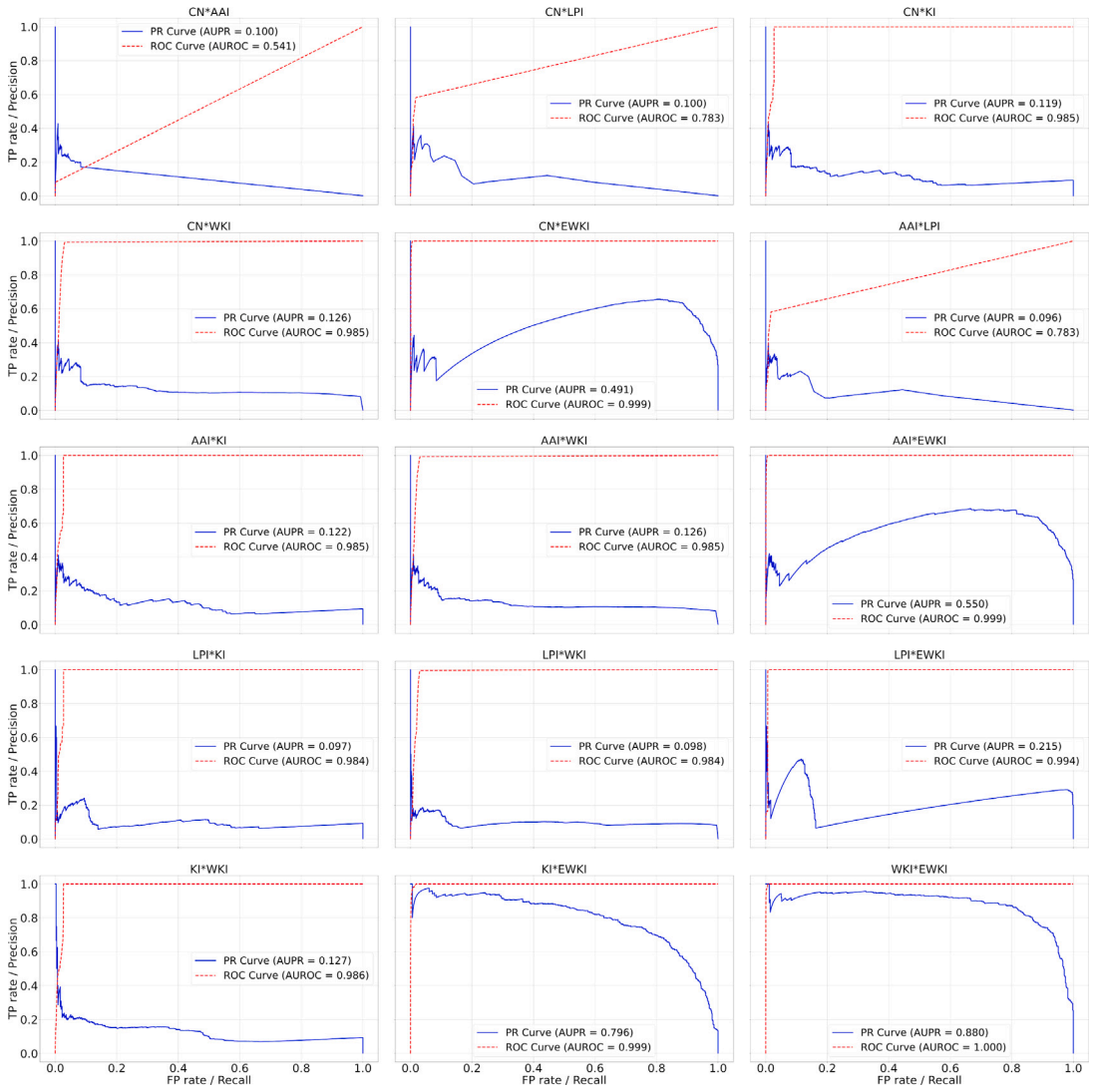


Fig. 5. Performance of ranking metrics (AUPR and AUROC) of the various combined link prediction methods on the live fish distribution network. Precision–recall curves (solid blue line) and receiver operating characteristic curves (dashed red line) are shown for each method, along with their respective values.

demonstrated that EWKI significantly enhanced CN's predictive performance. This improvement was attributed to EWKI's ability to capture spatially embedded relationships through its exponential decay weighting scheme, which prioritised geographically relevant links. By complementing CN's local information with EWKI's spatially aware structure, the model effectively addressed the limitations of CN alone in the aquaculture network, improving its capacity to predict links in a sparse and imbalanced setting. These findings demonstrate the general poor performance of CN in networks with sparse connectivity, as it relies solely on shared neighbours, which are often insufficient for predicting links in such datasets.

3.3.2. Combination with AAI

Similar to the CN combinations, the models averaging AAI with other indices also showed limited improvements. This generally poor performance can be attributed to AAI's design, which, while aiming to improve upon CN, still primarily focuses on local neighbourhood information. AAI assigns higher weights to links connecting nodes with fewer neighbours, attempting to address the bias of CN towards highly connected nodes. However, in sparse networks like the aquaculture network under study, this adjustment often proves insufficient. Many nodes have few or no common neighbours, and AAI struggles to identify potential links in these situations, leading to low recall and a high number of false negatives.

Despite the overall limited improvement, combining AAI with EWKI yielded a noticeable increase in performance, particularly in precision. The AAI*EWKI combination achieved a precision of 40%, surpassing all other AAI combinations. This improvement highlights the benefit of integrating AAI's local focus with EWKI's spatial awareness. By incorporating distance-based weights, EWKI helps AAI identify potential links that would otherwise be missed due to the sparsity of the network. For instance, two farms with no common neighbours but located in close proximity are more likely to interact. EWKI captures this spatial relationship, complementing AAI's local perspective and leading to more accurate predictions. However, even with the improved precision, the recall of the AAI*EWKI combination remained low at 1.5%. This indicates that while the combination benefits from EWKI's spatial information, it still struggles to capture the full range of true positive links. This limitation suggests that further refinements or alternative combination strategies might be necessary to fully leverage the strengths of both AAI and EWKI in sparse networks.

3.3.3. Combination with LPI

The combinations of LPI with LPI*KI, LPI*WKI, and LPI*EWKI demonstrated varying degrees of improvement, reflecting the complementary strengths of the paired models. The LP*KI model achieved

a precision of 8.4% and a recall of 13.1%, resulting in an F1-score of 10.3%. The AUPR for LPI*KI was slightly higher than LPI alone, indicating that KI's global structural information refined predictions by improving the model's ability to distinguish true links from false ones. However, the performance gains were modest, as the recall decreased slightly due to fewer true positives (TP = 89). The LPI*WKI benefited from the spatial weighting introduced by WKI, resulting in a precision of 7.7% and a recall of 14.5%, yielding an F1-score of 10.0%. However, the gains in precision and recall were limited, as the number of false positives (FP = 1,177) remained high, reflecting the challenges of balancing accuracy across metrics in sparse networks. The most notable improvement was observed in the LPI*EWKI combination, which achieved an AUPR 21.7% higher than the standalone LPI model and outperformed the other LPI combinations. This significant improvement occurred because EWKI's exponential spatial weighting complemented LPI's reliance on path-based connectivity by prioritising geographically relevant links. The combination effectively reduced the influence of long, less relevant paths while preserving the strengths of both methods in capturing true positives.

3.3.4. Combination with KI, WKI and EWKI

The combinations involving KI, WKI, and EWKI demonstrated significant improvements in predictive performance among all the combinations explored. This observation highlights the importance of incorporating both global network information and spatial awareness in predicting live fish movements. The combination of KI with EWKI yielded one of the best performances overall, with a recall of 80.3% and an F1-score of 74.4%. It also had a high AUPR of 79.7% and an AUROC of 99.9%. This suggests that EWKI's ability to incorporate spatial weighting enhances KI, making the combination highly effective in predicting links in the aquaculture network. The high recall indicates that this combination captures most TPs, while the high precision and AUPR reflect the model's ability to minimise FPs and provide accurate predictions. KI, by considering all paths in the network, captures the global connectivity patterns, while EWKI refines this information by prioritising links between geographically close farms. This synergy allows the KI*EWKI combination to effectively identify potential fish movements that are both structurally and spatially likely.

Similarly, WKI*EWKI also demonstrated strong performance with an F1-score of 78.5% and an AUPR of 88.1%. This further supports the assessment that combining spatially informed methods can be highly effective in predicting distribution between farms. Both WKI and EWKI incorporate spatial information, but their weighting schemes differ. WKI uses a simple linear weighting, while EWKI employs an exponential decay function. Combining these two approaches allows the model to capture a wider range of spatial interactions, leading to improved performance. In contrast, the KI*WKI combination showed a more moderate improvement with an F1-score of 17.7%. This indicates that while both methods capture network connectivity, their combination might not fully leverage their individual strengths. Consequently, the KI*WKI combination, while showing some improvement over the individual indices, does not achieve the same level of performance as the combinations involving EWKI.

3.4. Implications for distribution of live fish between farms

Incorporating spatial information into network analysis offers benefits for understanding live fish distribution between farms by identifying factors influencing fish movement patterns, such as transportation costs and shared resources. This knowledge informs the development of effective prediction strategies that account for the spatial dynamics of the aquaculture network. Understanding the spatial clustering of farms helps identify potential hotspots for demand or supply. Spatial analysis highlights areas with limited connectivity, revealing opportunities for infrastructure development to enhance the efficiency and resilience of the fish distribution network.

The use of EWKI extends beyond aquaculture to other fields where network analysis can benefit from spatial considerations. For example, in epidemiology, understanding disease transmission dynamics informs the development of control strategies. The EWKI model could be used with other epidemiological methods to predict the spread of infectious diseases across populations, enabling targeted interventions. In transportation and logistics, spatial network analysis can optimise routes and supply chain efficiency. In social network analysis, incorporating spatial information enhances the understanding of how proximity affects social interactions and information spread.

3.5. Limitation and future work

The study demonstrates the effectiveness of the EWKI in both the live fish distribution network and the road network, highlighting its ability to predict links with high precision, recall, and F1-score. The EWKI's incorporation of spatial weighting improved predictive performance across both datasets, showcasing its adaptability to networks with distinct structural and dynamic characteristics. However, the application of EWKI may be constrained by computational demands, particularly when scaling to larger networks. This challenge is further compounded by the need to determine the optimal value of the spatial decay parameter γ, which plays a critical role in adjusting the influence of distance in the model. Identifying the best-performing γ requires a grid search over multiple values, a process that can be computationally intensive. Another limitation concerns the use of AUROC as a performance metric in the context of highly imbalanced datasets. While AUROC can suggest strong overall discrimination, it may be overly optimistic in cases where true negatives dominate, as is typical in sparse networks like the one used in this study. Future work should consider exploring data balancing techniques such as under-sampling or over-sampling to provide a more representative evaluation of model performance. These approaches could help ensure that the AUROC reflects the model's effectiveness in identifying the minority class more accurately. Additionally, future studies could explore the integration of EWKI with dynamic graph embedding and graph convolutional approaches to further enhance its scalability, temporal adaptability, and predictive accuracy in complex evolving networks.

4. Conclusion

This study aimed to enhance the predictive accuracy of link prediction in aquaculture networks by extending the traditional Katz Index through the incorporation of spatial weighting, resulting in the development of the EWKI. The focus was to accurately predict the changing dynamics of fish farms during the movement of live fish between them. These dynamics are influenced by factors such as exotic disease incursions, which can alter movement patterns. When a disease outbreak occurs, official services may impose movement restrictions on affected farms or require the culling of entire fish stocks to prevent further spread. This creates changes in the network, as restricted farms are temporarily removed from the movement system. Accurately predicting links during these changes is essential for effective aquaculture management.

The results demonstrated that the EWKI model outperformed the traditional KI and other methods. By incorporating spatial information, the EWKI provided more accurate predictions, offering valuable insights for disease surveillance and targeted interventions in aquaculture. Furthermore, the study found that combining EWKI with other link prediction methods improved performance compared to using those methods in isolation. This observation highlights the synergistic potential of integrating EWKI's spatial awareness with the strengths of other methods, such as LPI. By leveraging both spatial weighting and global or local structural features, these combined approaches provided more accurate predictions. The outcome of this study is significant not only in aquaculture but also in transport systems.

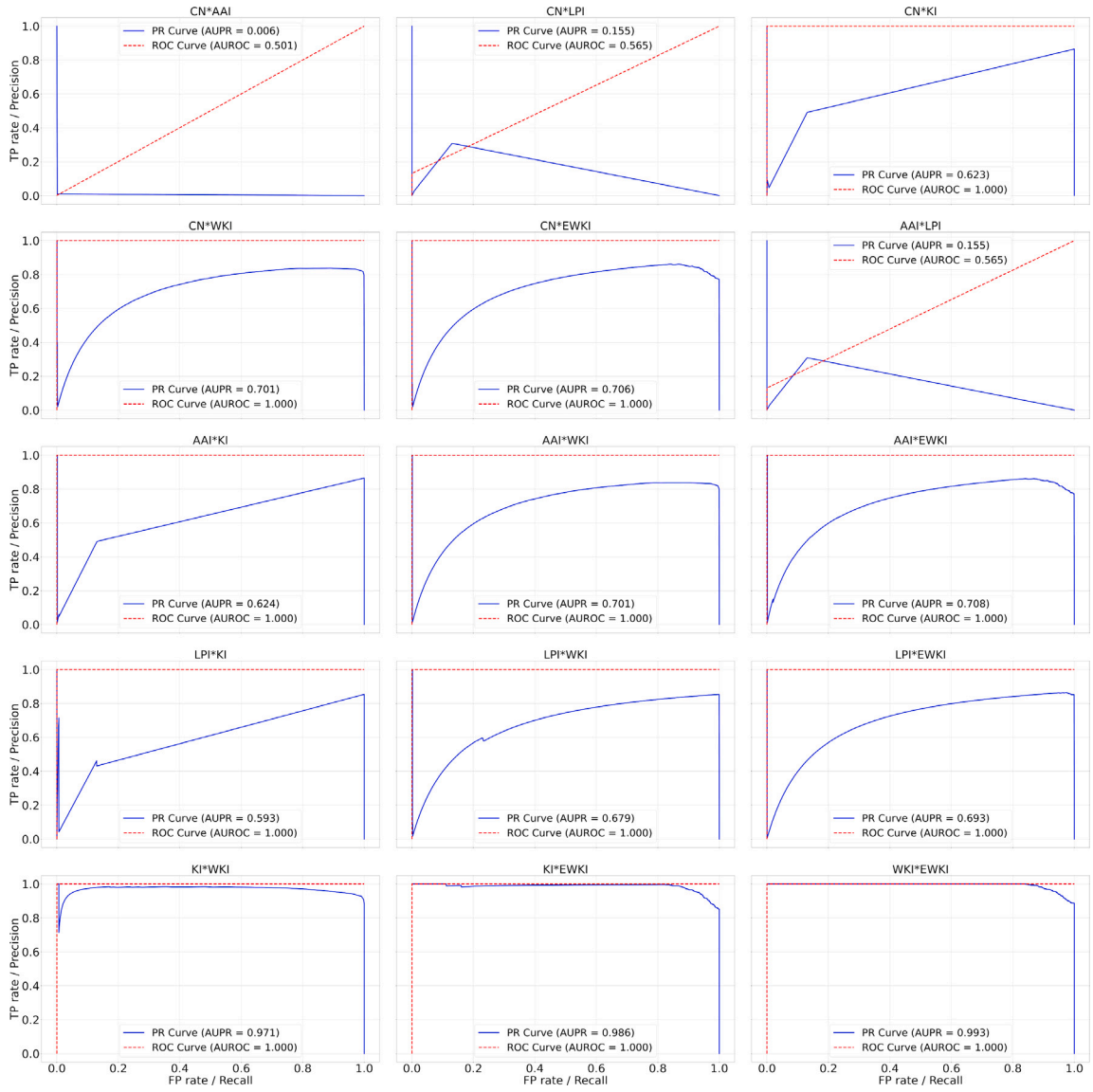


Fig. A.6. Performance of ranking metrics (AUPR and AUROC) of the various combined link prediction methods on the road network. Precision–recall curves (solid blue line) and receiver operating characteristic curves (dashed red line) are shown for each method, along with their respective values.

CRediT authorship contribution statement

Michael-Sam Vidza: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Conceptualization. **Marcin Budka:** Writing – review & editing, Validation, Supervision. **Wei Koong Chai:** Writing – review & editing, Validation, Supervision. **Mark Thrush:** Writing – review & editing, Supervision, Resources, Funding acquisition. **Mickaël Teixeira Alves:** Writing – review & editing, Supervision, Resources, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Michael-Sam Vidza reports financial support was provided by United Kingdom Department for Environment Food and Rural Affairs. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was funded by the Department for Environment, Food and Rural Affairs (Defra), United Kingdom [Project FC1215].

Appendix. Supplementary results

This Appendix presents supplementary results (Table A.5 and Fig. A.6) to provide further insights into the performance of the combined link prediction methods on the road network.

Data availability

The data that has been used is confidential.

Table A.5
Performance of various link prediction methods the road network.

Method	Precision (%)	Recall (%)	F1-score (%)	TP	FP	TN	FN
CN*AAI	0.99	0.14	0.24	2	200	5,828,136	1,472
CN*LPI	0.98	0.27	0.43	4	403	5,827,933	1,470
CN*KI	4.76	0.68	1.19	10	200	5,828,136	1,464
CN*WKI	81.87	99.59	89.87	1,468	325	5,828,011	6
CN*EWKI	81.51	96.00	88.16	1,415	321	5,828,015	59
AAI*LPI	0.98	0.27	0.43	4	403	5,827,933	1,470
AAI*KI	4.76	0.68	1.19	10	200	5,828,136	1,464
AAI*WKI	1.96	0.27	0.48	4	200	5,828,136	1,470
AAI*EWKI	12.02	1.49	2.66	22	161	5,828,175	1,452
LPI*KI	4.29	0.68	1.17	10	223	5,828,113	1,464
LPI*WKI	1.76	0.27	0.47	4	223	5,828,113	1,470
LPI*EWKI	8.61	1.42	2.44	21	223	5,828,113	1,453
KI*WKI	75.00	0.81	1.61	12	4	5,828,332	1,462
KI*EWKI	93.97	94.03	94.00	1,386	89	5,828,247	88
WKI*EWKI	100.00	73.27	84.57	1,080	0	5,828,336	394

References

[1] Hawoong Jeong, Sean P. Mason, A.-L. Barabási, Zoltan N. Oltvai, Lethality and centrality in protein networks, Nature 411 (6833) (2001) 41–42.
[2] Mark E.J. Newman, The structure and function of complex networks, SIAM Rev. 45 (2) (2003) 167–256.
[3] Stefano Boccaletti, Vito Latora, Yamir Moreno, Martin Chavez, D.-U. Hwang, Complex networks: Structure and dynamics, Phys. Rep. 424 (4–5) (2006) 175–308.
[4] Jingyi Lin, Yifang Ban, Complex network topology of transportation systems, Transp. Rev. 33 (6) (2013) 658–685.
[5] Jonathan F. Donges, Yong Zou, Norbert Marwan, Jürgen Kurths, Complex networks in climate dynamics: Comparing linear and nonlinear network construction methods, Eur. Phys. J. Spec. Top. 174 (1) (2009) 157–179.
[6] Stanley Wasserman, Katherine Faust, Social Network Analysis: Methods and Applications, Cambridge University Press, 1994.
[7] Jon M. Kleinberg, Navigation in a small world, Nature 406 (6798) (2000) 845.
[8] Jon Kleinberg, The small-world phenomenon: An algorithmic perspective, in: Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing, 2000, pp. 163–170.
[9] Jon Kleinberg, Complex networks and decentralized search algorithms, in: Proceedings of the International Congress of Mathematicians, ICM, Vol. 3, Citeseer, 2006, pp. 1019–1044.
[10] Lise Getoor, Christopher P. Diehl, Link mining: a survey, Acm Sigkdd Explor. Newsl. 7 (2) (2005) 3–12.
[11] G. Nandi, A. Das, A survey on using data mining techniques for online social network analysis, Int. J. Comput. Sci. Issues (IJCSI) 10 (6) (2013) 162.
[12] Lada A. Adamic, Eytan Adar, Friends and neighbors on the web, Soc. Netw. 25 (3) (2003) 211–230.
[13] Andrea Franceschini, Damian Szklarczyk, Sune Frankild, Michael Kuhn, Milan Simonovic, Alexander Roth, Jianyi Lin, Pablo Minguez, Peer Bork, Christian Von Mering, et al., STRING v9. 1: protein-protein interaction networks, with increased coverage and integration, Nucleic Acids Res. 41 (D1) (2012) D808–D815.
[14] Chengwei Lei, Jianhua Ruan, A novel link prediction algorithm for reconstructing protein–protein interaction networks by topological similarity, Bioinformatics 29 (3) (2013) 355–364.
[15] Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P. Tsafou, et al., STRING v10: protein–protein interaction networks, integrated over the tree of life, Nucleic Acids Res. 43 (D1) (2015) D447–D452.
[16] Linyuan Lü, Tao Zhou, Link prediction in weighted networks: The role of weak ties, Europhys. Lett. 89 (1) (2010) 18001.
[17] Agriculture Organization of the United Nations. Fisheries Department, The State of World Fisheries and Aquaculture, 2000, vol. 3, Food & Agriculture Org., 2000.
[18] Edmund J. Peeler, Nicholas G.H. Taylor, The application of epidemiology in aquatic animal health-opportunities and challenges, Vet. Res. 42 (1) (2011) 94.
[19] Simon Jennings, Grant D. Stentiford, Ana M. Leocadio, Keith R. Jeffery, Julian D. Metcalfe, Ioanna Katsiadaki, Neil A. Auchterlonie, Stephen C. Mangi, John K. Pinnegar, Tim Ellis, et al., Aquatic food security: insights into challenges and solutions from an analysis of interactions between fisheries, aquaculture, food safety, human health, fish and human welfare, economy and environment, Fish Fish. 17 (4) (2016) 893–938.
[20] Serge M. Garcia, Andrew A. Rosenberg, Food security and marine capture fisheries: characteristics, trends, drivers and future perspectives, Phil. Trans. R. Soc. B 365 (1554) (2010) 2869–2880.
[21] Darren Michael Green, Alison Gregory, Lorna Ann Munro, Small-and large-scale network structure of live fish movements in Scotland, Prev. Vet. Med. 91 (2–4) (2009) 261–269.
[22] H.J. Tidbury, D. Ryder, M.A. Thrush, F. Pearce, E.J. Peeler, N.G.H. Taylor, Comparative assessment of live cyprinid and salmonid movement networks in England and Wales, Prev. Vet. Med. 185 (2020) 105200.
[23] Anne E. Jones, Lorna A. Munro, Darren M. Green, Kenton L. Morgan, Alexander G. Murray, Rachel Norman, D. Ryder, N.K.G. Salama, Nick G.H. Taylor, Mark A. Thrush, et al., The contact structure of Great Britain2019s salmon and trout aquaculture industry, Epidemics 28 (2019) 100342.
[24] Art R.T. Jonkers, Kieran J. Sharkey, Mark A. Thrush, J.F. Turnbull, Kenton L. Morgan, Epidemics and control strategies for diseases of farmed salmonids: A parameter study, Epidemics 2 (4) (2010) 195–206.
[25] James Guilder, David Ryder, Nick G.H. Taylor, Sarah R. Alewijnse, Rebecca S. Millard, Mark A. Thrush, Edmund J. Peeler, Hannah J. Tidbury, The aquaculture disease network model (AquaNet-Mod): A simulation model to evaluate disease spread and controls for the salmonid industry in England and Wales, Epidemics 44 (2023) 100711.
[26] Hafida Benhidour, Lama Almeshkhas, Said Kerrache, Link prediction in directed complex networks: combining similarity-popularity and path patterns mining, Appl. Intell. 54 (17) (2024) 8634–8665.
[27] Xin-Jian Xu, Chong Deng, Li-Jie Zhang, Hyperlink prediction via local random walks and jensen–shannon divergence, J. Stat. Mech. Theory Exp. 2023 (3) (2023) 033402.
[28] Weilun Chen, Yinzuo Zhou, A link prediction similarity index based on enhanced local path method, in: 2021 40th Chinese Control Conference, CCC, IEEE, 2021, pp. 753–757.
[29] Min Li, Shuming Zhou, Dajin Wang, Gaolin Chen, Missing link prediction using path and community information, Computing 106 (2) (2024) 521–555.
[30] Statutory Instruments, The aquatic animal health (England and Wales) regulations, 2009, URL https://www.legislation.gov.uk/uksi/2009/463/part/1. Last accessed 16 September 2017.
[31] Linyuan Lü, Tao Zhou, Link prediction in complex networks: A survey, Phys. A 390 (6) (2011) 1150–1170.
[32] David Liben-Nowell, Jon Kleinberg, The link prediction problem for social networks, in: Proceedings of the Twelfth International Conference on Information and Knowledge Management, 2003, pp. 556–559.
[33] Miller McPherson, Lynn Smith-Lovin, James M. Cook, Birds of a feather: Homophily in social networks, Annu. Rev. Sociol. 27 (1) (2001) 415–444.
[34] Purnamrita Sarkar, Deepayan Chakrabarti, Michael Jordan, Nonparametric link prediction in dynamic networks, 2012, arXiv preprint arXiv:1206.6394.
[35] Christina Muro, Boyu Li, Kun He, Link prediction and unlink prediction on dynamic networks, IEEE Trans. Comput. Soc. Syst. 10 (2) (2022) 590–601.
[36] Mark E.J. Newman, Clustering and preferential attachment in growing networks, Phys. Rev. E 64 (2) (2001) 025102.
[37] Gobinda G. Chowdhury, Introduction to Modern Information Retrieval, Facet publishing, 2010.
[38] Leo Katz, A new status index derived from sociometric analysis, Psychometrika 18 (1) (1953) 39–43.
[39] Weiping Liu, Linyuan Lü, Link prediction based on local random walk, Europhys. Lett. 89 (5) (2010) 58007.
[40] Tao Zhou, Linyuan Lü, Yi-Cheng Zhang, Predicting missing links via local information, Eur. Phys. J. B 71 (2009) 623–630.
[41] Linyuan Lü, Ci-Hang Jin, Tao Zhou, Similarity index based on local paths for link prediction of complex networks, Phys. Rev. E—Stat. Nonlinear Soft Matter Phys. 80 (4) (2009) 046122.
[42] Fei Gao, Katarzyna Musial, Colin Cooper, Sophia Tsoka, Link prediction methods and their accuracy for different social networks and network metrics, Sci. Program. 2015 (2015) 1.
[43] Alexander G. Murray, Edmund J. Peeler, A framework for understanding the potential for emerging diseases in aquaculture, Prev. Vet. Med. 67 (2–3) (2005) 223–235.

[44] Brian Austin, Infectious Disease in Aquaculture: Prevention and Control, Elsevier, 2012.
[45] Geoff Boeing, Street network models and indicators for every urban area in the world, Geogr. Anal. 54 (3) (2022) 519–535.
[46] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, Morgan Kaufman Publ. (1995).
[47] Ryan N. Lichtenwalter, Jake T. Lussier, Nitesh V. Chawla, New perspectives and methods in link prediction, in: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2010, pp. 243–252.
[48] Víctor Martínez, Fernando Berzal, Juan-Carlos Cubero, A survey of link prediction in complex networks, ACM Comput. Surv. 49 (4) (2016) 1–33.
[49] Jesse Davis, Mark Goadrich, The relationship between Precision-Recall and ROC curves, in: Proceedings of the 23rd International Conference on Machine Learning, 2006, pp. 233–240.
[50] Takaya Saito, Marc Rehmsmeier, The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets, PloS One 10 (3) (2015) e0118432.
[51] James Bergstra, Yoshua Bengio, Random search for hyper-parameter optimization, J. Mach. Learn. Res. 13 (2) (2012).
[52] Frank Hutter, Lars Kotthoff, Joaquin Vanschoren, Automated Machine Learning: Methods, Systems, Challenges, Springer Nature, 2019.
[53] Thomas G. Dietterich, Ensemble methods in machine learning, in: International Workshop on Multiple Classifier Systems, Springer, 2000, pp. 1–15.
[54] Zhi-Hua Zhou, Ensemble Methods: Foundations and Algorithms, CRC Press, 2012.
[55] Jennifer A. Hoeting, David Madigan, Adrian E. Raftery, Chris T. Volinsky, Bayesian model averaging: a tutorial (with comments by M. Clyde, David Draper and EI George, and a rejoinder by the authors, Statist. Sci. 14 (4) (1999) 382–417.
[56] Aaron Clauset, Cristopher Moore, Mark E.J. Newman, Hierarchical structure and the prediction of missing links in networks, Nature 453 (7191) (2008) 98–101.
[57] Chao Wang, Venu Satuluri, Srinivasan Parthasarathy, Local probabilistic models for link prediction, in: Seventh IEEE International Conference on Data Mining, ICDM 2007, IEEE, 2007, pp. 322–331.
[58] Ajay Kumar, Shashank Sheshar Singh, Kuldeep Singh, Bhaskar Biswas, Link prediction techniques, applications, and performance: A survey, Phys. A 553 (2020) 124289.