Applying Bipartite Network Approach to Scarce Data: Validation of the Habitat Suitability Model of a Marine Mammal Species

ChinYing Liew[1] and Jane Labadin[2]
[1]*Faculty of Computer and Mathematical Sciences,*
Universiti Teknologi MARA, 94300 Kota Samarahan, Sarawak, Malaysia.
[2]*Faculty of Computer Science and Information Technology,*
Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia.
liewchinying@hotmail.com

***Abstract*—This paper presents the validation of the bipartite habitat suitability network (BiHSN) model formulated for a marine mammal. The model formulation published earlier resulted in the ranking of location nodes of the concerned area of possible habitats. Thus, the validation of the model is achieved by comparing the result produced by the BiHSN Model with the result acquired i) using another sample of actual data; and ii) from an ecological survey conducted by another researcher. Spearman's Rank Correlation Coefficient (SRCC) is used to quantify the similarity of the comparison where a threshold value of at least 0.70 is set in order to signify an acceptable validation analysis. In the former validation analysis, this study reports an SRCC of 0.976 whereas the later validation analysis reports an SRCC of 0.914. Due to the high values of SRCC obtained, we conclude that the BiHSN Model is thus validated.**

***Index Terms*—Model Validation; Bipartite Network Modeling; Network Modeling; Computational Modeling; Habitat Suitability; Irrawaddy Dolphin; Marine Mammal.**

I. INTRODUCTION

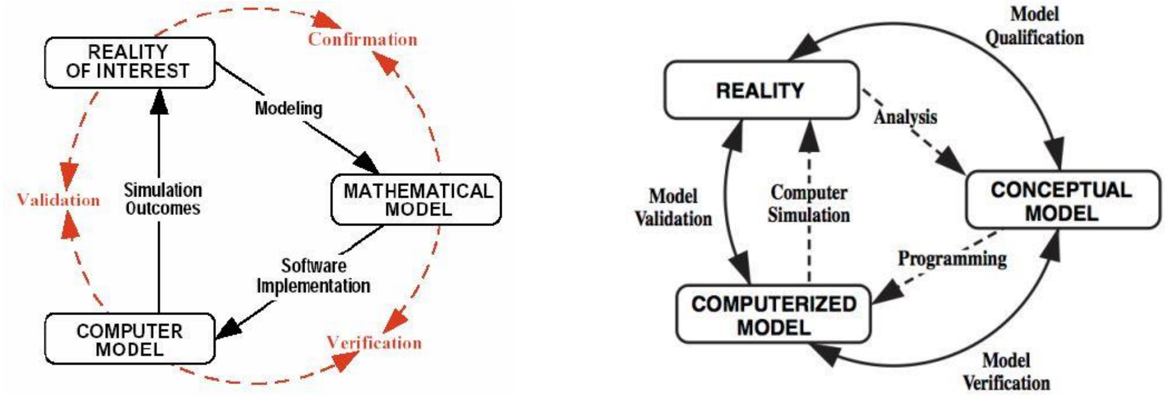
Validation is an important analysis in typical modeling effort regardless of discipline or field the model is formulated for. Though accepted by general researchers that it is almost impossible to obtain absolute validated model, i.e. model without error, model validation remains a pertinent process required for almost all modeling work [1]-[4]. Researchers are concerned if prediction produced by a model is reliable and applicable in the real world by the end users of the model. Therefore, the output of a validation analysis is closely related to the accuracy and credibility of the model developed particularly in the aspect of the model potential predictive capability. It has become an enabling methodology [3] in the modeling processes that ascertain the accuracy of predictions by the model with quantified confidence. As stipulated by Thacker et al. [3], model validation is expected to inform of the "…quantified level of agreement between experimental data and model prediction, as well as the predictive accuracy of the model…".

In relating the role of the validation process within the computational modeling activities, researchers in the field of sciences and engineering have illustrated how validation play their part within the main modeling processes in a simplified diagrammatic representation as shown in Figure 1a. *Reality of Interest* represents the real problem researchers intend to solve; the *Mathematical Model* represents the possible representation that is derived by the researchers, usually in the form of mathematical equations, of the solution to the problem under studied; and *Computer Model* represents the computer programming algorithms developed by the researchers to implement the *Mathematical Model* in order to obtain the solution to the problem under studied [3].

This figure is actually adapted from a more general graphical representation that is devised to depict the verification and validation processes in common modeling activities, shown in Figure 1b. Comparison between Figure 1a and 1b reveal that the *Mathematical Model* is one of the forms of a more general *Conceptual Model* which is defined as "A description of reality in terms of verbal descriptions, equations, governing relationships or 'natural laws' that purport to describe reality." [5]. As a result, model validation is widely accepted as an analysis carried out to show that the result obtained from the implementation of computer codes reveals the actual reality scenario through comparing the predictions of the model with the experimental results.

The model we intend to validate in this study is the Bipartite Habitat Suitability Network (BiHSN) Model [6]. It is a bipartite network model formulated to represent the habitat suitability of Irrawaddy dolphin (*Orcaella brevirostris*), a marine mammal species, in Kuching Bay of Sarawak, Malaysia. The bipartite network consists of thirteen location nodes, thirteen dolphin nodes, and 38 edges as presented in Figure 2 of Liew et al. [6], p.272. Implementation of the ranking algorithm onto this network produces estimated habitat suitability index (HSI) [6] for the location nodes which has enabled the nodes to be ranked. The resulted location nodes ranking from the highest to the lowest is reported in Table 1 of Liew et al. [6], p.273 – L2, L1, L12, L8, L7, L5, L11, L6, L9, L13, L10, L4, and L3.

Consequently, this study resolves to adopt the definition of model validation as "substantiation that a computerized model within its domain of applicability possesses a satisfactory range of accuracy consistent with the intended application of the model (Schlesinger et al. [7] cited in Lim and Barlow [8], p.337). The content of this paper is presented as follow: section I introduces the theme and background of this study, and the model this study intends to validate; section II gives the methods and materials used in this study; section III reports the findings obtained and presents the corresponding discussions; and the last section summarizes and concludes the study, besides presenting suggestions for further studies.



(a)

(b)

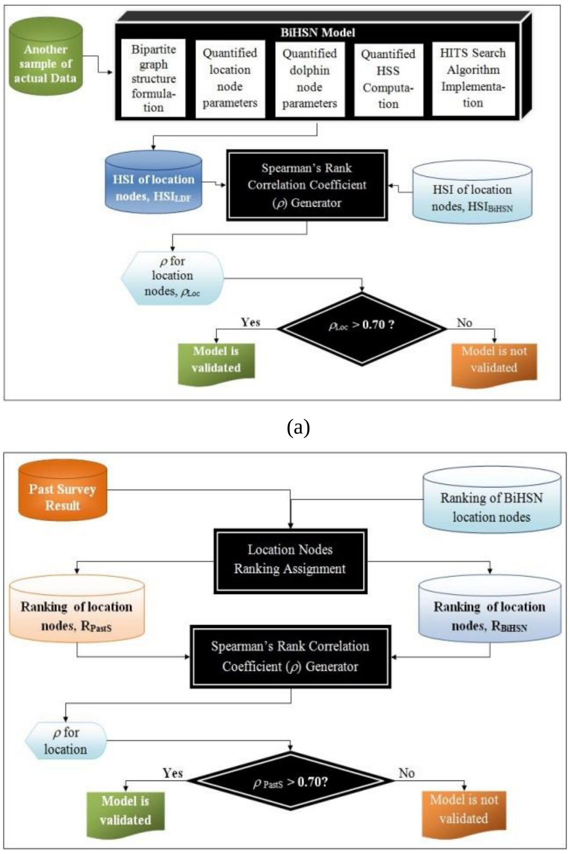
Figure 1: Role of model verification and validation process in (a) computational modeling (Source: [3], p.5); (b) general modeling (Source: [7], p.103)

II. Materials and Methods

Model validation in this study is performed through two analyses – another sample of actual data and the result of an ecological survey conducted by other research. The focus is to compare the result produced by the BiHSN Model with the results a) produced from the use of another actual data, and b) obtained from the past survey. BiHSN Model will be concluded as validated if its result is similar with the results obtained from both of these analyses. The similarity here is quantified using Spearman’s Rank Correlation Coefficient (SRCC). SRCC is valued between –1 and 1. It is able to measure the similarity between two rankings where 1 implies a perfect positive similarity (i.e. exactly the same ranking), 0 implies an absence of similarity (i.e. Similarity does not exit), and –1 implies a perfect negative similarity (i.e. exactly the opposite ranking). The rule of thumb adopted in interpreting SRCC computed states that a value of 0.90 and above signifies very high correlation; between 0.70 and 0.90 high; between 0.50 and 0.70 moderate; between 0.30 and 0.50 low; and any value less than 0.30 signifies negligible correlation relationship [9, 10]. Our study resolves to adopt a threshold value of no less than 0.70, signifying there exists a positive and high similarity between the BiHSN ranking and the ranking by another sample of actual data and the past survey result, for us to conclude that the BiHSN Model has been validated.

This sample of actual data used in the first validation is a real-world data obtained from the Sarawak Dolphin Project (SDP) research team. It is the individual ID dataset which is identified through the left dorsal fin (LDF) of the ID in Kuching Bay [11] and the re-sight maps of these individual ID as depicted in Figure 4.3a and b of Peter [11], p. 67. The individual ID dataset used in the formulation of BiHSN Model [6] is identified through the right dorsal fin (RDF) of the ID at Kuching Bay. The RDF and LDF individual ID dataset are considered and assumed as two completely different datasets [11]. The implementation design of validation using this real-world data is shown in Figure 2(a). As depicted in Figure 2(a), the sample of actual data is input to the BiHSN Model where the same quantification methods are used to generate parameters values and to calculate the HSS, and the same adapted HITS search algorithm is implemented. This produces the corresponding ranking indices – HSI_{LDF} – for the location nodes of the actual data. Nevertheless, not all of the location nodes are the same between the BiHSN Model and the actual dataset. There are eight location nodes (L2, L5, L7, L8, L9, L10, and L11) that are identical. As a result, the corresponding ranking indices for these identical nodes are used to compute the SRCC for the location nodes (ρ_{Loc}). The equation used to calculate ρ_{Loc} is given in (1) where a is a natural number, $RankHSI_{BiHSN}$ refers to the ranking of HSI_{BiHSN}, and $RankHSI_{LDF}$ the ranking of HSI_{LDF} of the respective identical location nodes, and N = 8, resembling the eight identical location nodes.

$$\rho_{Loc} = 1 - \frac{6\sum_{a=1}^{N}\left[\left\{RankHSI_{BiHSN}\right\}_a - \left\{RankHSI_{LDF}\right\}_a\right]^2}{N\left(N^2-1\right)} \tag{1}$$


Another sample of actual Data
BiHSN Model
Bipartite graph structure formula-tion
Quantified location node parameters
Quantified dolphin node parameters
Quantified HSS Computa-tion
HITS Search Algorithm Implementa-tion
HSI of location nodes, HSI_LDF
Spearman’s Rank Correlation Coefficient (ρ) Generator
HSI of location nodes, HSI_BiHSN
ρ for location nodes, ρ_Loc
ρ_Loc > 0.70 ?
Yes
No
Model is validated
Model is not validated
Past Survey Result
Ranking of BiHSN location nodes
Location Nodes Ranking Assignment
Ranking of location nodes, R_PastS
Ranking of location nodes, R_BiHSN
Spearman’s Rank Correlation Coefficient (ρ) Generator
ρ for location
ρ_PastS > 0.70?
Yes
No
Model is validated
Model is not validated

(a)

(b)

Figure 2: Implementation Design of Validation using (a) another Sample of Actual Data; (b) Past Survey Result

In the last step of Figure 2(a), the resulted ρ_{Loc} is compared with the threshold value of SRCC set in this study. The BiHSN Model will be concluded as validated only if ρ_{Loc} is not less than 0.70.

As for the validation using a past ecological survey result, the result obtained from the work of [11] is used. This past survey result recorded the relative densities (number of on-effort sightings per km searched) of ID in Kuching Bay. It is presented in a density map of Kuching Bay that is overlaid with 2 km by 2 km grid cells (Figure 3.5, [11], p. 43). Nevertheless, in order to compare with the result obtained from the BiHSN Model, we considered the past survey result that is within the scope where the BiHSN Model is formulated [6]. Figure 3 presents the visualization of overlaying the location nodes of our model onto the past survey result, following the scope of the BiHSN Model. Figure 3 shows that out of 13 location nodes defined in the BiHSN Model, six overlaps with the past survey result. These six location nodes are L2, L4, L5, L9, L10, and L12. Consequently, these six location nodes are taken as the identical location nodes and thus comparable with the above past survey result. The implementation design of this validation analysis is presented in Figure 2(b).

In the first step, assignment of ranking for location nodes for both past survey result and BiHSN result is needed as both results use different ways of ranking. A quantification process is used to assign a rank to the shades of the past survey result and the ranking of location nodes in BiHSN. The darkest shade which represents “0.200001-1.310809 ID per km searched” in the past survey result is removed from the assignment of rank here as the grid cell with this shade is outside the scope of BiHSN Model. Consequently, the shade representing “0.100001-0.200000 ID per km searched” which is ranked the first is assigned rank ‘1’, “0.050001-0.100000 ID per km searched” is assigned rank ‘2’ and “0.050001-0.100000 ID per km searched” is ranked ‘3’. The implementation of this step is shown in Figure 4(a).

As for the ranking of location nodes of BiHSN Model, it is also quantified and reassigned rank the same way as the past survey result so that both rankings could be compared using the SRCC. The common mathematical ratio rule of dividing thirteen (ranked location) by three (ranking level as in the past survey result) which is approximately equal to four is used here. Subsequently, the thirteen ranked location nodes are grouped into three groups with at least four location nodes in each of the groups, as shown in Figure 4(b).

In the second step of Figure 2(b), the two ranking results will be compared using SRCC as defined in (2) where the SRCC calculated is referred as ρ_{PastS}, $RankR_{BiHSN}$ refers to the ranking of locations nodes in the BiHSN Model, $RankR_{PastS}$ is the ranking of the BiHSN Model’s location nodes in the past survey result, a is a natural number, and N = 6 (as there are six location nodes that are identical and comparable between the BiHSN result and the past survey result). These rankings are shown in Table 1 and 2.

$$\rho_{PastS} = 1 - \frac{6\sum_{a=1}^{N}\left[\left\{RankR_{BiHSN}\right\}_a - \left\{RankR_{PastS}\right\}_a\right]^2}{N\left(N^2 - 1\right)} \tag{2}$$

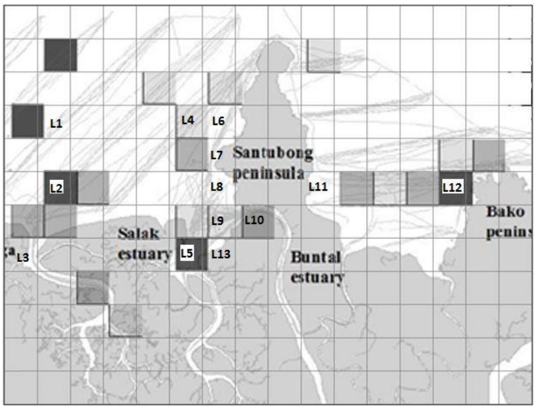


Figure 3: Modified Past Survey Result overlaid with the BiHSN Location Nodes (Source: Liew et al. [6], p.273)

Table 1
Assignment of Rank to the Shades

Shade	Rank Assigned, R_{PastS}	Location nodes
□	3	L4 L9
▨	2	L10
■	1	L2 L5 L12

Table 2
Assignment of Rank for BiHSN Result

BiHSN Rank	Rank Assigned, R_{BiHSN}
L2 L1 L12 L8	1
L7 L5 L11 L6	2
L9 L13 L10 L4 L3	3

In the last step, the resulted ρ_{PastS} is compared with the threshold value set in this study. The BiHSN Model will be concluded as validated only if ρ_{PastS} is not less than 0.70.

III. Result And Discussion

Now we are going to look at the result of these validation analyses. For the validation analysis through another sample of actual data, execution of SRCC Generator in Figure 2(a) via Equation (1) gives the SRCC (ρ_{Loc}) for the ranking of location nodes. The values of HSI computed for the location nodes in this real-world data ($\mathrm{HSI}_{\mathrm{LDF}}$), together with the HSI obtained from the BiHSN Model ($\mathrm{HSI}_{\mathrm{BiHSN}}$), are input to the SRCC Generator. Table 2 presents the values of $\mathrm{HSI}_{\mathrm{LDF}}$ and $\mathrm{HSI}_{\mathrm{BiHSN}}$, and the calculation of the terms in Equation (1) for eight location nodes. As mentioned earlier, there are eight identical location nodes that are comparable between the actual data and the BiHSN Model.

From Table 3, the ρ_{Loc} resulted is 0.976. When compared with the threshold value (no lesser than 0.70) set in this study as illustrated in Step 3 of Figure 2(a), it is concluded that the BiHSN Model is validated through another sample of actual data.

Next is the validation using the past survey result. Following the implementation design in Figure 2(b), execution of SRCC Generator is achieved via Equation (2) that produces SRCC (ρ_{PastS}) for the ranking of location nodes. The inputs to the SRCC Generator are the ranking of the location nodes obtained from the pass survey result (R_{PastS}) and the BiHSN Model (R_{BiHSN}). Table 2 shows the values of $\text{RankR}_{\text{BiHSN}}$ and $\text{RankR}_{\text{PastS}}$, and the calculation of terms in Equation (2), where $d = RankHSI_{BiHSN} - RankR_{PastS}$. From Table 4, the SRCC (ρ_{PastS}) computed is 0.914, signifying a positive and very high similarity in the ranking of location nodes between the BiHSN

Model and the past survey result. After comparing with the threshold value set (no lesser than 0.70) in this study as illustrated in Figure 2(b), we can conclude that the BiHSN Model is also validated through the past survey result.

Results obtained from both of the validation analyses show that the result produced by the BiHSN Model is positive and highly similar with the result acquired through the actual data and the past survey result. Consequently, the BiHSN Model formulated is validated.

Table 3
Calculation of SRCC for Validation Analysis through another Sample of Actual Data

Location Node	$\text{HSI}_{\text{BiHSN}}$	HSI_{LDF}	Rank $\text{HSI}_{\text{BiHSN}}$	Rank HSI_{LDF}	$\text{RankHSI}_{\text{BiHSN}}$ – $\text{RankHSI}_{\text{LDF}}$	$[\text{RankHSI}_{\text{BiHSN}} – \text{RankHSI}_{\text{LDF}}]^2$
L2	1.0000 E+00	1.0000E+00	1	1	0	0
L5	3.9070 E-03	9.8265E-05	5	6	–1	1
L7	3.5969 E-02	7.7613E-02	4	4	0	0
L8	4.8837 E-02	1.0926E-01	3	3	0	0
L9	1.0279 E-05	1.5162E-05	7	7	0	0
L10	7.8998 E-08	1.6105E-08	8	8	0	0
L11	2.7536 E-03	5.2605E-03	6	5	1	1
L12	6.4007 E-02	5.4167E-01	2	2	0	0
					Sum of $[\text{RankHSI}_{\text{BiHSN}} – \text{RankHSI}_{\text{LDF}}]^2$ =	2
					ρ_{Loc} =	0.976

Table 4
Calculation of SRCC for Validation Analysis through Past Survey Result

Location Node	R_{BiHSN}	R_{PastS}	$\text{RankR}_{\text{BiHSN}}$	$\text{RankR}_{\text{PastS}}$	d	d^2
L2	1	1	1.5	2	– 0.5	0.25
L4	3	3	5	5.5	– 0.5	0.25
L5	2	1	3	2	1	1
L9	3	3	5	5.5	– 0.5	0.25
L10	3	2	5	4	1	1
L12	1	1	1.5	2	– 0.5	0.25
					Sum of d^2 =	3
					ρ_{PastS} =	0.914

IV. CONCLUSION

In this study, we have performed the validation analysis of the BiHSN Model [6] with another actual data and the result of an ecological survey conducted by another researcher. It produces a correlation coefficient of 0.976 and 0.914, respectively. With the high values of SRCC obtained, the validation analysis performed has managed to validate the BiHSN Model. The model could thus be used to represent the preferred habitat of Irrawaddy dolphin in Kuching Bay. Further studies are suggested to investigate the use of BiHSN Model in identifying the preferred habitat of i) the same species in other location, and ii) other species.

ACKNOWLEDGEMENT

The authors wish to thank Prof. Dr. Andrew Alek Tuen, Cindy Peter, and other SDP research team members for providing the data and expert opinions for the study. C.Y. Liew wishes to thank the Ministry of Education Malaysia and Universiti Teknologi MARA (UiTM) for providing scholarship with study leave during her doctorate term of study.

REFERENCES

[1] K. M. Carley, *Validating Computational Models*. Department of Social and Decision Sciences, Carnegie Mellon University, Sep. 1996.

[2] I. Babuska, and J. T. Oden, “Verification and validation in computational engineering and science: Basic concepts,” *Computer Methods in Applied Mechanics and Engineering*, vol. 193, pp. 4057-4066, 2004.

[3] B. H. Thacker, S. W. Doebling, F. M. Hemez, M. C. Anderson, J. E. Pepin, and E. A. Rodriguez, *Concepts of Model Verification and Validation*. USA: Los Alamos National Laboratory, Oct. 2004.

[4] E. L. Urban, E. S. Minor, E. A. Trem, and R. S. Schick, “Graph models of habitat mosaics,” *Ecology Letters*, vol. 12, pp. 260-273, 2009.

[5] J. C. Refsgaard and H. J. Henriksen, “Modelling guidelines – terminology and guiding principles,” *Advances in Water Resources*, vol. 27, pp. 71-82, 2004.

[6] C. Y. Liew, J. Labadin, Y. C. Wang, A. A. Tuen, and C. Peter, “Applying bipartite network approach to scarce data: Modeling habitat suitability of a marine mammal,” *Procedia Computer Science*, vol. 60, pp. 266-275, 2015.

[7] S. Schlesinger, R. E Crosbie, R. E. Gagne, G. S. Innis, C. S. Lalwani, J. Loch, et al., “Terminology for Model Credibility,” *Simulation*, vol. 32, no. 3, pp. 103-104, 1979.

[8] H. C. Lim and M. Barlow, “Generative experimentation and social simulation: Exploring gaming for model verification and validation,” *IEEE: 2011 International symposium on Computer Science and Society*, pp. 336-340, 2011.

[9] D. E. Hinkle, W. Wiersma, and S. G. Jurs, *Applied Statistics for the Behavioral Sciences* (5th ed.). Boston: Houghton Mifflin, 2003.

[10] M. M. Mukaka, “Statistics corner: A guide to appropriate use of correlation coefficient in medical research,” *Malawi Medical Journal*, vol. 24, no. 3, pp. 69-71, 2012.

[11] C. Peter, *Distribution Patterns, Habitat Characteristics and Population Estimates of Irrawaddy Dolphins* (*Orcaella brevirostris*) *in Kuching Bay, Sarawak*. Unpublished Master Science Thesis. Sarawak: IBEC:UNIMAS, 2012.