Modeling the spread of loanwords in South-East Asia using sailing navigation software and Bayesian networks

František Kratochvíl[1], Václav Kratochvíl[2], George Saad[3], and Jiří Vomlel[4]

[1,3]Department of Asian Studies, Palacký University Olomouc
frantisek.kratochvil,george.saad@upol.cz
[2,4]Institute of Information Theory and Automation, Czech Academy of Sciences
velorex,vomlel@utia.cas.cz

Abstract

A loanword is a word permanently adopted from one language and incorporated into another language without translation. In this paper we study loanwords in the South-East Asia Archipelago, a home to a large number of languages.

Our paper is inspired by the works of Hoffmann et al. (2021) Bayesian methods are applied to probabilistic modeling of family trees representing the history of language families and by Haynie et al. (2014) modelling the diffusion of a special class of loanwords, so called Wanderwörter in languages of Australia, North America and South America. We assume that in the South-East Asia Archipelago Wanderwörter spread along specific maritime trade routes whose geographical characteristics can help unravel the history of Wanderwörter diffusion in the area. For millennia trade was conducted using sailing ships which were constrained by the monsoon system and in certain areas also by strong sea currents. Therefore rather than the geographical distances, the travel times of sailing ships should be considered as a major factor determining the intensity of contacts among cultures.

We use a sailing navigation software to estimate travel times between different ports and show that the estimated travel times correspond well to travel times of a Chinese map of the sea trade routes from the early seventeenth century. We model the spread of loanwords using a probabilistic graphical model - a Bayesian network. We design a novel heuristic Bayesian network structure learning algorithm that learns the structure as a union of spanning trees for graphs of all loanwords in the training dataset. We compare this algorithm with BIC optimal Bayesian networks by measuring how well these models predict the true presence/absence of a loanword. Interestingly, Bayesian networks learned by our heuristic spanning tree based algorithm provide better results than the BIC optimal Bayesian networks.

1 Introduction

This paper examines the loanword distribution in Maritime Southeast Asia, focusing on loanwords from pre-colonial contact languages. The data comes from Blust and Trussel (2013). We know, for example, that the Arabic word *arak* '(type of) alcohol' (and variants thereof) is found in 35 languages of our 124 language sample but are clueless about the borrowing pathway.

In the absence of written records and extensive archaeological and genetic evidence, the distribution of loanwords across a wide area may offer insight into past human migrations, contacts, and trade. The mapping of the loanword distribution offers an opportunity to capture large patterns of human contact that are not as readily detectable by other means. It thus supports the formulation of hypotheses that can later be verified by other disciplines such as history, anthropology, archaeology, and genetics.

Austronesian languages of Maritime Southeast Asia have been the subject of linguistic study for several centuries; their linguistic relationships (proto-languages, branches, etc.) are well described and the Austronesian family belongs to the best reconstructed and documented language families (Greenhill et al., 2008; Blust, 2009; Blust and Trussel, 2013). The languages of Maritime Southeast Asia have accumulated layers of borrowed lexicon from culturally dominant 'donor languages' through trade, conquest, religion and technological development: (i) Old Malay and proto-Malayic (mainly plant and animal terms), (ii) Sanskrit (religion, state), (iii) Tamil (trade), (iv) Arabic (religion, law, trade), (v) Tagalog, and (vi) Chinese (trade). Modern Malay has had the most profound lexical influence in the area, both as a primary source language, and as an intermediary for Sanskrit, Tamil, Indic, Persian, and Arabic words.

Examining the loanwords in Australian and Californian lanugages Haynie et al. (2014) have detected loanword networks using clustering algorithms. Such network is defined by the distribution of so called 'Wanderwörter', which usually make up a small proportion of the total vocabulary of individual languages, and only a minority of loanwords (Haynie et al., 2014). Importantly, Wanderwörter are shared by languages in areas formerly linked by trade (Haynie et al., 2014). The main difference between our work and the work of Haynie and colleagues lies in the nature of the terrain. While the Californian and Australian networks run over land where the distance can be used as the main measure, the Maritime Southeast Asian networks consist of sea routes which are subject to weather patterns (monsoons), sea currents, sea relief, and coastal features. In addition, navigation properties of the various water crafts have to be taken into account when reconstructing the routes.

Our work builds on several previous work and uses a number of publicly available resources. Our primary resource is a large database of loanwords collected from several earlier sources, such as the Austronesian Comparative Dictionary (Blust and Trussel, 2013). The database is available at `http://gogo.utia.cas.cz/loanwords/`.

The paper is structured as follows: Section 2 discusses how travel time was estimated. Section 3 describes the Bayesian network models used. Section 4 elaborates on the experiments used. Finally, section 5 rounds off the paper with some conclusions and suggestions for further research.

2 Estimation of travel time

An important resource for analyzing the historical Maritime Southeast Asian networks is the Selden map, which is the oldest surviving Chinese merchant map of the sea trade routes of East Asia from the early seventeenth century.[1] This map was rediscovered by Robert Batchelor in 2008 in Oxford University's Bodleian Library Batchelor (2013). An important property of the map is that the distances in the map correspond to travel times and therefore the map can serve as a benchmark for estimating the travel times of sailing ships.

In 2021, Perttola (2021) published a method for estimating travel times of sailing ships and compared his estimation with travel times of the Selden map. To some extent, we replicate the work of Perttola (2021) though we implement a different method for computing travel times of sailing ships from the studied historical era. Each sailing ship can be characterized by its polar graph that provides the speed of the ship for different wind angles and wind speed. Based on data presented in Perttola (2021), we estimated the polar graph of the Chinese junk rig, which was a typical sailing ship of that area; see Figure 1 where we present the polar graph of a Chinese junk rig used in our computations.

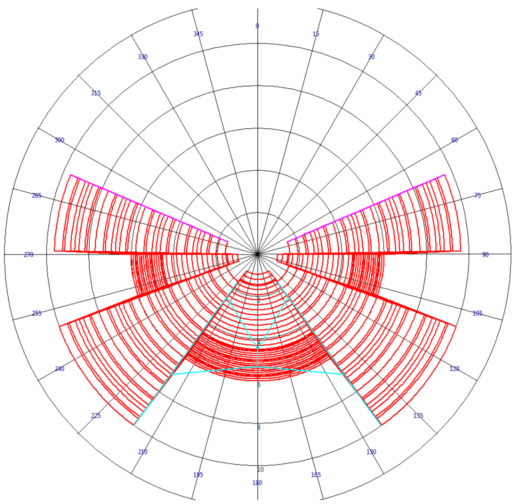


Figure 1: Polar graph of a Chinese junk rig.

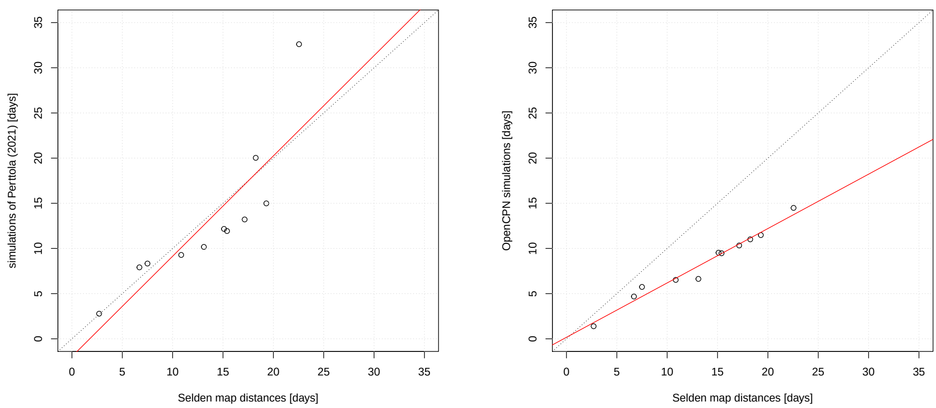
In the next step we used a sailing navigation software OpenCPN (2022) and its Weather Routing Plugin (2022) to estimate travel times of the Chinese junk rig specified by its polar graph. The weather routing plugin uses extensive data of winds and see currents from several decades and therefore it can provide reliable estimates of travel times for different seasons.

[1] The map is available at `http://seldenmap.bodleian.ox.ac.uk/` where it is possible to zoom it into a great detail.

In Figure 2 we present the comparisons of travel times from Selden map and travel estimated by Perttola Perttola (2021) and by us using the sailing navigation software OpenCPN, respectively. We can see that using the sailing navigation software we can estimate the travel times better than Perttola (2021). The Selden map travel times are computed from our estimates as $t_S = 1.66 \cdot t_O - 0.27$ meaning our estimates are systematically lower than true travel times. There can be several reasons for the optimistic estimates - the ships were not always performing as well as expected, the ships may have called at harbours along the route, etc. However, the important result is that the residual standard error of our method, which is 0.6947, is significantly lower than the residual standard error of Perttola's method, which is 4.288.

Perttola (2021) $t_P = -1.9635 + 1.1103 \cdot t_S$

Our model $t_O = 0.16469 + 0.60201 \cdot t_S$



Residual standard error: 4.288

Residual standard error: 0.6947

Figure 2: Comparison of travel time estimation methods

In our study we use this method to compute the travel times for all pairs of ports of the area. It serves as a basis for estimation of a Bayesian network (BN) model for the spread of loanwords. In Table 1 we present travel times between main ports of the corresponding language areas. The time is measured in days. Since the Malay speaking region covers a very large area, it is represented by eight different ports, namely by Palembang, Brunei, Banjarmasin, Jakarta, Singkawang, Singapore, Bukit Tengorak, and Samarinda. Similarly, Champa is represented by four different ports, namely by Indrapura, Vijaya, Kauthara, and Panduranga. When computing the distance of Malay (and Champa, respectively) to another language we used always the nearest port. In principle, this could be done also for other languages with several important ports of that language area but we decided to simplify the computations by considering only one port for other languages. We believe this simplification does not have a significant impact on the results presented, especially since many languages correspond to small geographical regions.

Table 1: Travel times (in days) between main ports of the corresponding language areas.

	Acehnese	Aklanon	Balinese	Bikol	Cas. Dumagat	Cebuano	Javanese	Kapampangan	Karo Batak	Makasarese	Manggarai	Maranao	Mongondow	Rejang	Rembong	Sangir	Simalur	Tagalog	Tiruray	Wolio	Iban	Melanau	Champa	Malay
Acehnese		16.22	14.11	18.61	18.96	16.23	11.4	16.17	2.81	15.97	16.3	16.39	20.41	6.58	16.93	17.8	1.46	16.17	16.59	17.43	10.29	10.53	12.15	7.39
Aklanon	16.22		12.28	2.8	3.29	2.59	12.39	2.03	18.7	10.04	10.94	4.02	6	17.11	11.15	4.14	17.34	2.03	3.43	11.38	7.47	6.85	4.93	3.49
Balinese	14.11	12.28		14.63	15.13	11.99	3.47	13.54	12.2	4.09	2.62	11.99	9.63	7.83	3.27	10.78	12.78	13.54	11.86	3.79	7.84	8.14	10.76	2.97
Bikol	18.61	2.8	14.63		0.76	2.67	14.83	3.5	20.95	12.66	13.53	5.88	7.71	19.24	13.17	5.15	19.72	3.5	5.88	12.34	9.84	9.23	6.68	6.1
Cas. Dumagat	18.96	3.29	15.13	0.76		3.2	15.34	3.88	21.18	13.14	14.02	6.25	8.2	19.95	13.65	5.64	20.08	3.88	6.25	12.86	10.38	9.77	6.93	6.59
Cebuano	16.23	2.59	11.99	2.67	3.2		12.35	4.61	18.42	8.36	10.75	3.73	5.59	17.19	12.96	3.66	17.34	4.61	3.1	10.89	7.38	6.77	5.59	3.3
Javanese	11.4	12.39	3.47	14.83	15.34	12.35		12.51	9.58	4.82	5.35	12.61	13.22	4.86	5.86	13.97	10.14	12.51	12.74	6.34	5.33	5.63	7.96	1.95
Kapampangan	16.17	2.03	13.54	3.5	3.88	4.61	12.51		18.42	11.39	12.27	5.97	7.96	17.02	12.48	6.08	17.28	0.5	5.42	12.88	7.53	6.92	4.37	4.78
Karo Batak	2.81	18.7	12.2	20.95	21.18	18.42	9.58	18.42		14.17	14.4	18.73	22.65	5.46	15.05	20.03	2.02	18.42	17.8	15.64	11.57	11.83	13.47	8.7
Makasarese	15.97	10.04	4.09	12.66	13.14	8.36	4.82	11.39	14.17		1.28	9.82	7.66	9.49	1.32	8.58	14.72	11.39	9.7	1.61	9.14	9.46	11.77	3.53
Manggarai	16.3	10.94	2.62	13.53	14.02	10.75	5.35	12.27	14.4	1.28		10.75	8.57	10	0.77	9.51	14.88	12.27	10.61	1.89	9.64	9.97	12.29	4.03
Maranao	16.39	4.02	11.99	5.88	6.25	3.73	12.61	5.97	18.73	9.82	10.75		4.18	17.36	9.85	1.61	17.61	5.97	1	9.16	7.66	7.06	6.44	3.33
Mongondow	20.41	6	9.63	7.71	8.2	5.59	13.22	7.96	22.65	7.66	8.57	4.18		19.78	7.92	2.91	21.53	7.96	4.98	7.1	11.6	11	10.64	4.47
Rejang	6.58	17.11	7.83	19.24	19.95	17.19	4.86	17.02	5.46	9.49	10	17.36	19.78		10.51	18.46	6.06	17.02	17.49	10.98	10.09	10.23	12.56	4.04
Rembong	16.93	11.15	3.27	13.17	13.65	12.96	5.86	12.48	15.05	1.32	0.77	9.85	7.92	10.51		8.5	15.55	12.48	9.66	1.54	10.95	10.47	12.77	4.6
Sangir	17.8	4.14	10.78	5.15	5.64	3.66	13.97	6.08	20.03	8.58	9.51	1.61	2.91	18.46	8.5		18.91	6.08	1.43	7.76	8.98	8.39	7.78	4.6
Simalur	1.46	17.34	12.78	19.72	20.08	17.34	10.14	17.28	2.02	14.72	14.88	17.61	21.53	6.06	15.55	18.91		17.28	17.71	16.2	10.47	10.72	12.33	7.61
Tagalog	16.17	2.03	13.54	3.5	3.88	4.61	12.51	0.5	18.42	11.39	12.27	5.97	7.96	17.02	12.48	6.08	17.28		5.42	12.88	7.53	6.92	4.37	4.78
Tiruray	16.59	3.43	11.86	5.88	6.25	3.1	12.74	5.42	17.8	9.7	10.61	1	4.98	17.49	9.66	1.43	17.71	5.42		8.85	7.77	7.17	6.58	3.44
Wolio	17.43	11.38	3.79	12.34	12.86	10.89	6.34	12.88	15.64	1.61	1.89	9.16	7.1	10.98	1.54	7.76	16.2	12.88	8.85		10.62	10.96	13.25	5.08
Iban	10.29	7.47	7.84	9.84	10.38	7.38	5.33	7.53	11.57	9.14	9.64	7.66	11.6	10.09	10.95	8.98	10.47	7.53	7.77	10.62		0.63	3.29	1.29
Melanau	10.53	6.85	8.14	9.23	9.77	6.77	5.63	6.92	11.83	9.46	9.97	7.06	11	10.23	10.47	8.39	10.72	6.92	7.17	10.96	0.63		2.99	3.89
Champa	12.15	4.93	10.76	6.68	6.93	5.59	7.96	4.37	13.47	11.77	12.29	6.44	10.64	12.56	12.77	7.78	12.33	4.37	6.58	13.25	3.29	2.99		3.19
Malay	7.39	3.49	2.97	6.1	6.59	3.3	1.95	4.78	8.7	3.53	4.03	3.33	4.47	4.04	4.6	4.6	7.61	4.78	3.44	5.08	1.29	3.89	3.19	

The travel times were computed using the Weather Routing Plugin (2022) of the sailing navigation software OpenCPN (2022), The Weather Routing Plugin optimizes ship routes using an isochrone method and predictive grib data or averaged gridded climate data. In this experiment, we performed computations for the date of January 1, which belongs to the period of the winter monsoon. The data from the database of the National Oceanic and Atmospheric Administration (NOAA) represents a 30 year average of winds and currents.[2] The resulting data presented in Table 1 is symmetric since we selected the direction with the shorter travel time for each pair of ports. The background assumption is that the boats can return during summer (when the winds reverse) with comparable travel time as in winter for the reverse direction. Such assumption is justified by the historical evidence showing that ships used to call at the most suitable harbours along their route where they awaited the optimal wind and weather conditions.[3]

3 Bayesian network models

Bayesian networks (Pearl, 1988; Jensen, 2001) are a popular class of probabilistic graphical models (Lauritzen, 1996; Koller and Friedman, 2009), i.e. models that use graphs to describe relations between random variables represented by nodes in the graphs. Bayesian networks model the variables' relations using directed acyclic graphs (DAGs) and their quantitative part is specified by conditional probability tables (CPTs) provided for each variable given its parents in the graph. We believe that Bayesian networks are very suitable for the task of modeling the spread of loanwords since using the graphical part we can use edges to encode frequent contacts between two languages and the quantitative relations between two languages by its nature can be modeled well by probabilistic relations. The state 0 of a model variable represents a particular loanword being absent in the corresponding language while state 1 corresponds to its presence.

To learn the graphical structure of a BN model modeling the spread of loanwords we suggest the following algorithm. Let L be the set of all considered loanwords and $a = 1, 2, 3, \ldots, 17$. The algorithm uses a training dataset which provides information on languages in which each loanword is present.

- Create an empty graph G_0 whose nodes correspond to studied recipient languages.
- For each loanword $\ell \in L$ from training data do:
  - create graph G_ℓ as a copy of graph G_0,
  - add an edge for each language pair containing loanword ℓ to graph G_ℓ. This means that the subgraph generated by languages where the loanword is present is a complete graph.

[2]It is assumed here that the weather systems of the region has been stable during the last three millennia.

[3]A more comprehensive model could include computations for several dates to select the optimal travel time for the journey and return, during the optimal season, should these fall between the monsoon peaks in January and July.

  – Evaluate all edges $X - Y$ of graph G_ℓ by the travel time $t_S(X, Y)$ from language X to language Y.
  – Find the cheapest spanning tree T_ℓ of graph G_ℓ.

- Make the union of all trees $T_\ell, \ell \in L$ by performing the union on the sets of edges for all studied loanwords. This creates an undirected graph H.
- Assign a weight to each edge computed as a number of trees in $\{T_\ell, \ell \in L\}$ containing this edge.
- Exclude from graph H edges that appear in less than a graphs.
- Direct the edges from the language with the higher number of loanwords.[4] This creates a DAG G that defines the structure of a Bayesian network.

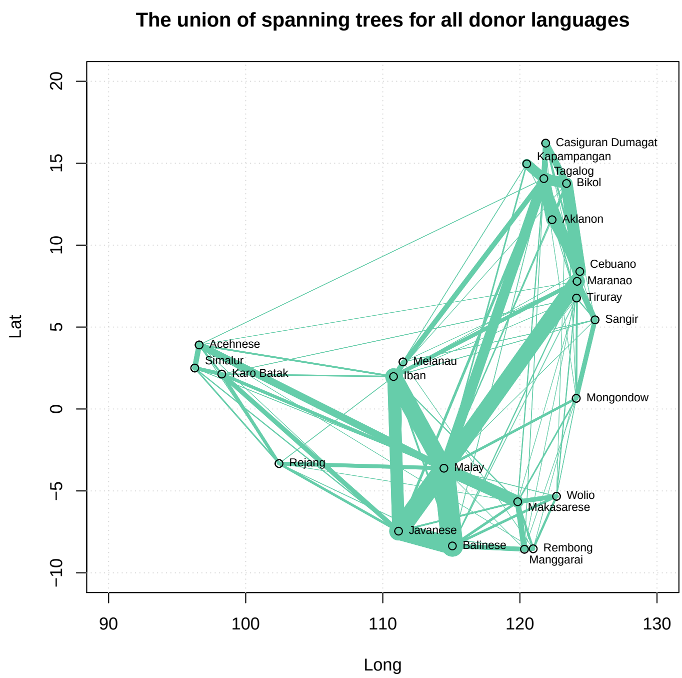


Figure 3: The underlying undirected graph H for $a = 1$.

The database of loanwords is also used to estimate probability values of CPTs which constitute the quantitative part of the learned BN. In Figure 3 we present an example of

[4]Ties are broken randomly.

the underlying undirected graph H with the threshold value $a = 1$ where the width of each edge corresponds to the number of spanning trees the edge is present. The locations of graph nodes correspond to geographical locations of corresponding languages.

In Figure 4 we present a printscreen from our linguistic tool based on a Bayesian network model. Each monitor window represents the marginal probability of a loanword being present in a studied language. The red color bars represent observed evidence for the presence or absence of a loanword while the bars printed green give a prediction probability that this loanword is present in corresponding language.

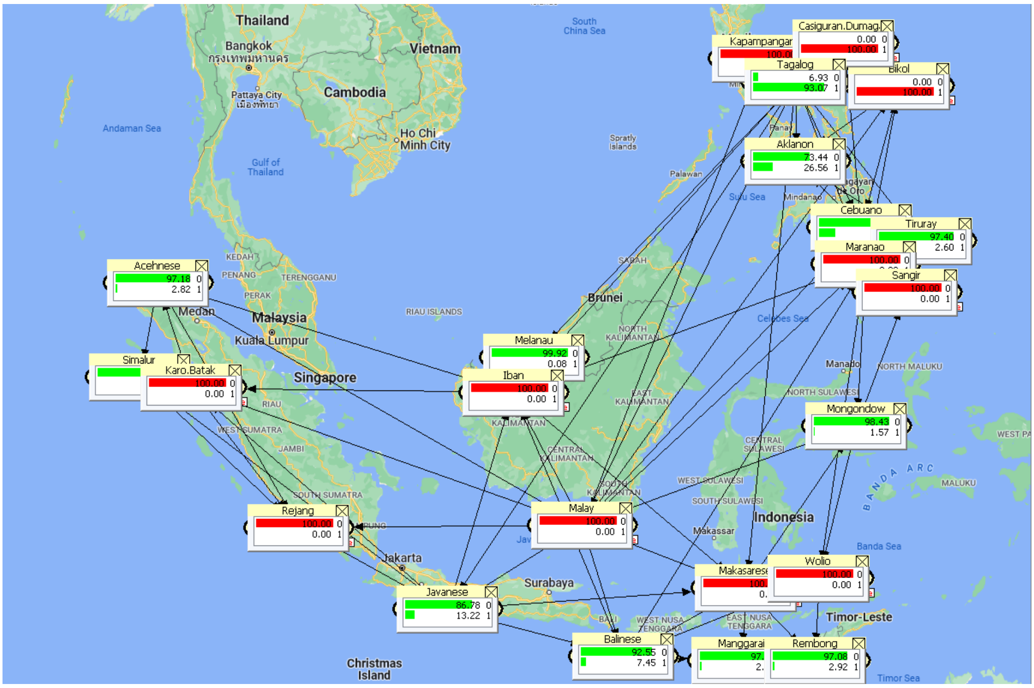


Figure 4: A printscreen from our linguistic tool based on a Bayesian network model.

4 Experiments

Standard methods for structural learning of BNs can also be applied to the studied problem. Therefore we decided to compare our proposed heuristic algorithm with the state-of-the-art learning algorithm which learns a globally optimal BN with respect to data and uses the Bayesian Information Criterion (BIC) as a BN quality criterion. For this purpose we used the Gobnilp method of Cussens and Bartlett (2018). The experiments were performed in the following way:

- The dataset D of 461 loanwords was split to ten subsets $D_1, \ldots, D_n$, all but one containing 46 loanwords and the tenth subset containing 47 loanwords.
- The following steps were repeated for each subset $D_i, i = 1, \ldots, 10$:
  - The set $D \setminus D_i$ was used to learn the BN model using the spanning trees based method (for $a = 1, 2, \ldots, 17$) and the BIC-optimal method.
  - The learned models were tested on loanwords from the set D_i. The presence/absence of this loanword for 11 randomly selected languages was entered into tested models and the conditional probability of the loanword presence in each of the remaining languages was estimated. The predicted state was compared with the true presence/absence of the loanword in D_i.
  - The number of true positives (tp), true negatives (tn), false positives (fp), and false negatives (fn) was computed and added to the results of previous subsets.
- The values of tp, tn, fp, and fn were used to compute the precision, recall, accuracy and balanced accuracy for both methods.

In this way we were able to compare models' predicting ability. In order to estimate the prediction quality limits of the models we repeated the above experiment also for the predictions based on observations of 21 randomly selected languages (instead of 11). The results are summarized in Figure 5. In Figure 6 we give an example of a spanning trees based model with the threshold value $a = 1$ and a BIC optimal model, both learned on the same training dataset.

5 Conclusions and future work

In this paper we studied loanword distribution patterns in Maritime Southeast Asia using the tools based on sailing navigation software and Bayesian networks. The sailing navigation software provided us a deeper insight into travel time between important ports of the studied region. We have shown that this method can provide good travel times estimate corresponding to travel times from a historical map of the sea trade routes in this area.

In the paper we have designed a novel heuristic Bayesian network structure learning algorithm and compared this algorithm with the Gobnilp method that learns BIC optimal Bayesian network structures. Bayesian networks learned by our heuristic spanning tree based algorithm have better prediction quality than the BIC optimal Bayesian networks. This might be attributed to the ability of our heuristic algorithm to exploit additional information provided by travel time distances. However, this needs to be further explored.

Since each language is represented by a Boolean variable, its conditional probability table $P(X|pa(X)$ can be defined as a *noisy-or*. The inhibition probabilities p_i for each $X_i \rightarrow Y$, $X_i \in pa(Y)$ can be learned from collected data. Then the conditional probability of the noisy-or is defined as:

$$P(Y = 0|X_1 = x_1, ..., X_n = x_n) \quad = \quad p_0 \cdot \prod_{i=1}^{n} (p_i)^{x_i} \quad .$$

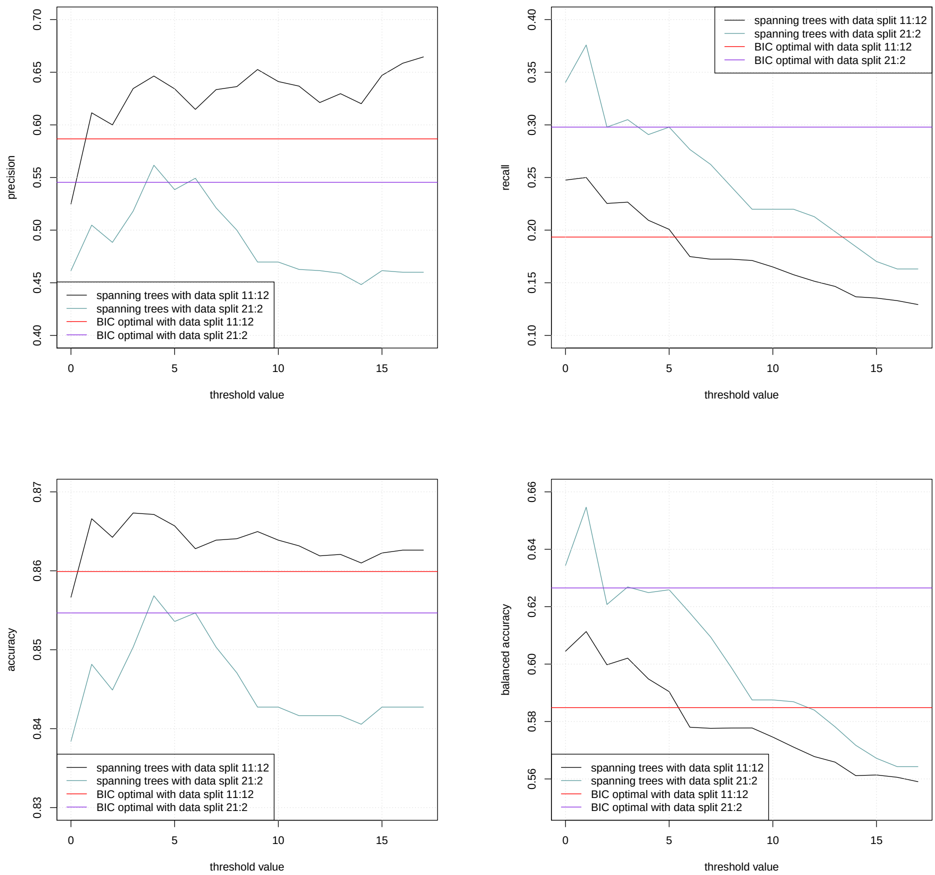


Figure 5: Comparisons of precision, recall, accuracy and balanced accuracy for the spanning trees based models with different threshold value a and the BIC optimal models.

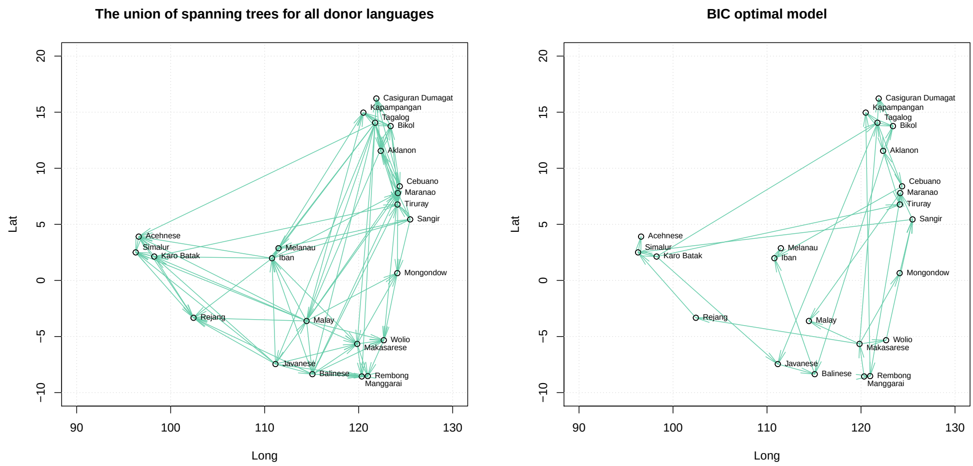


Figure 6: A spanning trees based model with the threshold value $a = 1$ and a BIC optimal model, both learned on the same training dataset.

The noisy-or model has a natural interpretation within this application. If a loanword is present in a language $X_i \in pa(Y)$ it is probable that it is present in language Y as well unless this influence is canceled for some reason. The probability of this influence being canceled is p_i. Naturally, one parent with the loanword being present could be sufficient (therefore the OR relation) but, of course, the more often the loanword is present in $pa(Y)$, the higher the probability of its presence in Y as well. A specialized BN structure learning method might be tailored well for BN models with its CPTs having this particular interpretation.

Also, from the linguistic point of view, it would be interesting to build separate models for each group of loanwords according to their donor language while refining the network of known harbours in each historical period and establishing polar graphs for older ship types than the Chinese junk rig - currently the only type for which a polar graph is available. This would allow an even more refined analysis of the spread of loanwords in different time epochs. We hope to address the above issues in our future work.

Acknowledments

We are grateful to Christian Dumard for recommending the OpenCPN tool. We are also grateful to the authors of this tool who made the weather routing plugin of the OpenCPN tool freely available. This work was supported by the Czech Science Foundation Project Nr. 20-18407S *Verb Class Analysis Accelerator for Low-Resource Languages - Robocorp*. The loanword data was compiled with the support of the ERDF grant *Sinophone Borderlands: Interactions at the Edges* (CZ.02.1.01/0.0/0.0/16_019/00 00791).

References

R. Batchelor. The Selden map rediscovered: A Chinese map of East Asian shipping routes, c.1619. *Imago Mundi*, 65(1):37–63, 2013. doi: 10.1080/03085694.2013.731203.

R. Blust. *The Austronesian languages.* Canberra: Pacific Linguistics, Research School of Pacific and Asian Studies, Australian National University, 2009.

R. Blust and S. Trussel. Australian comparative dictionary: A work in progress. *Oceanic Linguistics*, 52:493—-523, 2013. URL http://dx.doi.org/10.1353/ol.2013.0016.

J. Cussens and M. Bartlett. Gobnilp, 2018. Version version 1.6.3, https://www.cs.york.ac.uk/aig/sw/gobnilp/.

S. D'Epagnier. Weather Routing Pi, 2022. Version 1.13.38, https://opencpn.org/OpenCPN/plugins/weatherroute.html.

S. J. Greenhill, R. Blust, and R. D. Gray. The Austronesian basic vocabulary database: From bioinformatics to lexomics. *Evolutionary Bioinformatics*, 4:271—283, 2008.

H. Haynie, C. Bowerna, P. Epps, J. Hill, and P. McConvell. Wanderwörter in languages of the Americas and Australia. *Ampersand*, 1:1–18, 2014. URL https://doi.org/10.1016/j.amper.2014.10.001.

K. Hoffmann, R. Bouckaert, S. J. Greenhill, and D. Kühnert. Bayesian phylogenetic analysis of linguistic data using beast. *Journal of Language Evolution*, 8:1–75, 2021.

F. V. Jensen. *Bayesian networks and decision graphs.* Springer-Verlag, New York, 2001.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques.* The MIT Press, 2009.

S. L. Lauritzen. *Graphical Models.* Clarendon Press, 1996.

J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

W. Perttola. Digital navigator on the seas of the Selden map of China: Sequential least-cost path analysis using dynamic wind data. *Journal of Archaeological Method and Theory*, 2021. URL https://doi.org/10.1007/s10816-021-09534-6.

The OpenCPN Development Team. OpenCPN, 2022. Version 5.6.0, https://opencpn.org/.