Coastal Engineering 135 (2018) 16–30

Contents lists available at ScienceDirect

Coastal Engineering

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

Bayesian Networks in coastal engineering: Distinguishing descriptive and predictive applications

T. Beuzen[a,*], K.D. Splinter[a], L.A. Marshall[b], I.L. Turner[a], M.D. Harley[a], M.L. Palmsten[c]

[a] Water Research Laboratory, School of Civil and Environmental Engineering, UNSW Sydney, NSW, Australia
[b] School of Civil and Environmental Engineering, UNSW Sydney, NSW, Australia
[c] Naval Research Laboratory, Stennis Space Center, MS, USA

ARTICLE INFO

Keywords:
Bayesian network
Coastal processes
Predictive modelling
Descriptive modelling
Good modelling practice
Storm erosion

ABSTRACT

Bayesian networks (BNs) are increasingly being used to model complex coastal processes due to their ability to integrate non-linear systems, their transparent probabilistic framework, and low computational cost. A BN may be suited to descriptive or predictive application. Descriptive BNs are highly calibrated models that are useful for better understanding the physics and causal relationships driving a system. Predictive BNs are generalisations of a system that have skill at predicting outside of the training domain. The predictive and descriptive usefulness of a BN depends on its complexity and the amount of data available to train it, but there is often a trade-off; higher descriptive skill comes at the cost of reduced predictive skill. To demonstrate the differences between predictive and descriptive BNs in a coastal engineering context, a BN to predict shoreline recession caused by coastal storm events is developed and tested using an extensive 10-year dataset incorporating 137 individual storms events monitored at Narrabeen-Collaroy Beach, Australia. A parsimonious approach to BN development is used to separately determine the optimum predictive and descriptive BNs for this dataset. Results show that for this dataset two quite different BNs can be developed: one that is optimized to achieve the highest predictive skill, and a second network that is optimized to maximize descriptive skill. The optimum predictive BN is found to comprise 3 nodes (variables) and can predict the shoreline recession caused by unseen storm events with a skill of 65%. The optimum descriptive BN is composed of 5 nodes and can reproduce 88% of the training dataset, but with more limited predictive capabilities. The uses and limitations of these two different approaches to BN formulation are illustrated with example applications to coastal process modelling. It is anticipated that the insights provided in this paper will help to clarify the further development of Bayesian Networks applied to coastal modelling.

1. Introduction

Bayesian networks (BNs) are probabilistic graphical models that can be used to represent causal systems. They model interactions between variables describing a system using representative datasets and statistics founded on Bayes' rule of conditional probability. BNs originate from artificial intelligence research and are increasingly being used to model environmental systems (Aguilera et al., 2011). BNs can easily handle non-linear systems, have low computational cost, can deal with missing data and data from different sources, explicitly include uncertainties, and have a simple and intuitive graphical structure that is easily understood by non-technical users (Chen and Pollino, 2012; Uusitalo, 2007). On the other hand, BNs depend on the quality of data used to develop them and require continuous variables to be discretised. For a thorough introduction into BNs, the reader is referred to Pearl (1988) and Charniak (1991).

Recently, BNs have been used in a number of coastal engineering applications, including: predicting episodic coastal cliff erosion (Hapke and Plant, 2010), reproducing wave-height evolution in the surf zone (Plant and Holland, 2011), assessing coastal vulnerability to sea level rise (Gutierrez et al., 2011), predicting barrier island response to storms (Plant and Stockdon, 2012; Wilson et al., 2015), predicting dune retreat resulting from coastal storms (Palmsten et al., 2014) and modelling hurricane damage to urbanised coasts (van Verseveld et al., 2015).

These studies and others have shown that BNs can have considerable skill modelling a range of complex coastal processes. However, one topic that is not well clarified in the literature is that BNs may be suited to descriptive or predictive applications; that is, a BN may be skilful at

* Corresponding author.
E-mail address: t.beuzen@unsw.edu.au (T. Beuzen).

https://doi.org/10.1016/j.coastaleng.2018.01.005
Received 11 July 2017; Received in revised form 13 September 2017; Accepted 14 January 2018

representing and reproducing a unique dataset descriptively, or at generalising the causal relationships in the dataset such that they are applicable to predicting unseen data. It is therefore important to clarify whether the BN purpose is descriptive or predictive as this dictates its generic applicability. Fienen and Plant (2015) developed a *k*-fold cross-validation application using the BN software package Netica (Norsys Software Corporation, 1995–2017) for assessing the predictive and descriptive skill of a BN. In *k*-fold cross-validation a dataset is divided into *k* number of folds (or partitions), where *k* is commonly taken as 10 (Marcot, 2012). A BN is trained and tested on all but 1 fold of the data (descriptive skill) and then tested on the 1 withheld fold (predictive skill) for all *k* permutations of training and testing sets. *k*-fold cross-validation is an unbiased way of evaluating model descriptive and predictive skill (Elsner and Schmertmann, 1994), and is widely applied to test machine learning models (Refaeilzadeh et al.). Fienen and Plant (2015) and other recent coastal studies (e.g (Gutierrez et al., 2015; Poelhekke et al., 2016).,) have used cross-validation to show that predictive skill and descriptive skill vary with BN model complexity and that there is a trade-off between the two – better descriptive power usually comes at the cost of reduced predictive power (Fienen and Plant, 2015; Gutierrez et al., 2015), resulting in different optimum BN structures for both descriptive and predictive BN applications.

While cross-validation provides a useful method of distinguishing between a predictive or descriptive BN model, there remains no standard procedure to developing the optimum predictive or descriptive model for a particular dataset (Chen and Pollino, 2012; Marcot, 2012). In coastal applications to date, the typical approach taken to BN model development has been to start with a complete conceptual model of the system and then iteratively modify and evaluate this structure to investigate model skill and sensitivity (e.g (Plant and Stockdon, 2012; Wilson et al., 2015; Palmsten et al., 2014).,). An alternative and more objective approach that is often used for empirical model development, but has received less attention in the coastal BN literature to date, is the parsimonious approach to model development (Sivapalan and Young, 2005). The parsimonious model approach builds a model up, from simple to complex, using only model inputs and causal relations that are justified and optimised by the available training dataset (Sivapalan and Young, 2005). Such an approach integrates sensitivity analysis into model development and protects against the model fitting spurious relationships in the data. Practically, parsimonious BNs can be developed by constructing and evaluating a conceptualised BN model one variable at a time, based on maximising the descriptive and predictive skill at each step of construction. This not only allows identification of the optimal predictive and descriptive variable subsets to use in a BN for a given dataset but further serves the practical and physically meaningful purpose of identifying how individual variables in the dataset impact the skill of the model (Sivapalan and Young, 2005).

The aim of this paper is to explore the distinction between descriptive and predictive BNs in coastal modelling. To this end, a BN to model shoreline recession caused by coastal storm events is developed and tested using an extensive 10-year dataset from Narrabeen-Collaroy Beach, on the southeast coast of Australia. Understanding and predicting the response of the shoreline to coastal storm events remains a focus of the coastal research community (Holman et al., 2015), having important implications for both emergency and long-term coastal management. This is particularly the case for highly-developed, dynamic sandy coastlines such as Narrabeen-Collaroy Beach, where coastal storm events can place beachfront infrastructure at risk in the short term (Harley et al.) and often dominate longer-term patterns of shoreline change (Harley et al., 2011). BNs offer an appealing method of modelling the impacts of storm events that differs from the empirical or process-based model approaches that have typically been used in these coastal settings (e.g (Davidson et al., 2013; Harley et al., 2009; Karunarathna et al., 2014; Wright et al., 1985; Splinter et al., 2014a).,). Here, a parsimonious approach to BN development is used to develop the optimal descriptive and predictive BNs for modelling storm-induced shoreline change at Narrabeen-Collaroy Beach, followed by a discussion and example applications of how these models can be used in coastal settings. This paper further serves as an introduction to BN modelling and a reference point for understanding other BN studies in the coastal science and engineering community.

2. Study site

Narrabeen-Collaroy Beach (hereafter referred to simply as Narrabeen) is a sandy, 3.6 km long embayed beach bounded at its extremities by rocky headlands. It is situated on the southeast coast of Australia approximately 20 km north of the centre of Sydney. The beach is composed of fine to medium quartz sand ($D50 \approx 0.3$ mm), with ~30% carbonate fraction, and has a typical intertidal slope of 0.12. Its modal beach state ranges from dissipative-intermediate (wider, flatter beaches often with bars and adapted to high-energy wave conditions) at the exposed northern end, to intermediate-reflective (narrow, steeper beaches adapted to lower energy wave conditions) toward the sheltered southern end (Wright and Short, 1984; Harley et al., 2015a). Tides are microtidal and semidiurnal with a mean spring tidal range of 1.3 m. The Sydney region has a moderate to high energy deepwater wave climate characterised by persistent SSE swell waves ($Hs = 1.6$ m and $Tp = 10$ s) that is interrupted by ~5% significant wave height exceedance storm events ($Hs > 3$ m) typically 10–20 times per year (Short and Trenaman, 1992; Harley et al., 2010). These storm events vary in their origin, ranging from tropical cyclones from the northeast, east-coast lows from the east, and mid-latitude cyclones from the south.

Routine monitoring commenced at Narrabeen in 1976, and it is now one of the longest records of coastline variability worldwide (Turner et al., 2016). The present study focuses on one aspect of this monitoring program: 10 years of daily shoreline data spanning 2004–2014 obtained by an Argus coastal imaging station (Holman and Stanley, 2007). This coastal imaging station provides coverage of the southern half of the Narrabeen embayment, of which a 400 m alongshore region is used here (Fig. 1).

3. Methodology

3.1. Bayesian Networks

A BN is a graphical representation of the joint probability distribution of a system comprised of discrete variables. A very simple illustration of a hypothetical BN to predict the occurrence of erosion versus accretion as a response to different combinations of wave height and period is shown in Fig. 2. The BN consists of nodes representing variables in the system (e.g., *Wave Height*, *Wave Period* and *Beach Response*) that are connected with arcs representing causality between nodes. The arcs and nodes together are formally known as a directed acyclic graph (DAG), and a node is termed a *parent* of a *child*, if there is an arc directed from the former to the latter (Pearl, 1988). Each node is discretised into states (e.g., in Fig. 2, '*Beach Response*' contains the states '*Erosion*' and '*Accretion*'). Associated with each node are either prior probability tables (PPTs) for the nodes that have no parents, or conditional probabilities tables (CPTs) for child nodes. These tables quantify the probability of a node being in its particular states. As shown in Fig. 2a, while PPTs are only the same size as the number of states in a node, CPTs represent the probability of all the possible combinations of parent and child states for a given node and so increase in size exponentially:

$$size(CPT) = S\prod_{i=1}^{n} P_i \tag{1}$$

where S is the number of states of the child node, and P_i is the number of states in the *ith* parent node. For the BN in Fig. 2, the number of parameters to estimate is calculated as 2 (*Beach Response* child states) x 2 (*Wave Height* parent states) x 3 (*Wave Period* parent states) = 12