19th International Congress on Modelling and Simulation, Perth, Australia, 12–16 December 2011
http://mssanz.org.au/modsim2011

Coupling Bayesian networks and geospatial software for environmental risk assessment

A. Jolma [a], A. Lehikoinen [b] and I. Helle [b]

[a] *Department of Civil and Environmental Engineering, School of Engineering, Aalto University, Espoo, Finland*
[b] *Department of Environmental Sciences, Helsinki University, Helsinki, Finland*
Email: ari.jolma@aalto.fi

Abstract: The Gulf of Finland, the easternmost part of the Baltic Sea, is a fairly narrow and shallow gulf that has witnessed a significant increase in maritime traffic, especially in oil transportation in the last years. While several measures have been taken to improve the safety of sea traffic in the area, a risk of a major oil spill accident exists. Although oil spills can be combated at sea at least to some extent, substantial amounts of spilled oil can reach the shoreline, especially in small and narrow sea areas like the Gulf of Finland. This may cause significant environmental damage. The risks that oil transportation poses to environmental values are spatially distributed because some locations within the transport network, are more prone for accidents, and because ecological values are not evenly distributed along the shores.

In this paper we describe a method and tool that was developed to assess the risks that oil transportation poses to environmental values of the Finnish shore of the Gulf of Finland. In a recent research project (SAFGOF) a Bayesian network (BN) describing oil tanker accidents in the Gulf of Finland was developed. The network encodes a set of initial conditions that was considered to cover all realistic situations. A large number of oil spill simulations were carried out using the initial conditions. The environmental values for the assessment were available from a previous study, where information about shoreline habitats was collected and sensitivity indicators for oil spill combating were developed.

The HUGIN software for BNs was used in SAFGOF and it was also chosen for this study. Geoinformatica was chosen as the geospatial software for the study. A Perl foreign function interface was developed for the HUGIN C library so that it could be easily coupled with Geoinformatica. The internal plug-in technology of Perl (Perl modules) along with some graphical widgets was used to couple the risk assessment tool to the Geoinformatica graphical user interface application.

The risk assessment tool was implemented as a Perl module, which presents a dialog-box to the user and contains callback-functions for the interaction with the BN and for the computations. The dialog-box is also used for specifying which layer to use as the environmental values for the computations. Use of the plug-in requires some preliminary coupling work for linking a BN with the set of geospatial environmental scenarios as raster files. When the user opens a BN into the tool, the tool looks for the coupling file. Additionally, the tool looks for an image file of the network.

The coupling is defined in a text file, which contains a file name template for the scenario files and tuples, which link a specific state of a specific variable in the BN to a specific match in the file name template. The file name of each scenario must therefore contain codes that reveal the state that each scenario variable had when the scenario was computed. The plug-in contains several checks for the correctness of the coupling of the scenario rasters with the BN. The risk assessment tool supports an interactive workflow, where a value layer is created from, e.g., a vector layer created by valuation tool, a scenario is defined by a BN, and an overlay computation is initiated to produce a risk map. Generic visualization tools can then be used to prepare the map.

Keywords: *Bayesian network, Geospatial software, Decision Support System (DSS), risk assessment, oil spill*

INTRODUCTION

Ecological risk assessment (ERA) is a field of study which evaluates the risks associated with an eco-environmental hazard under uncertainty (Xu et al 2004). The risks are often spatially distributed because some locations, e.g., within a transport network, are more prone to accidents, and because ecological values are not evenly distributed in space. Geospatial information technology supports obtaining and managing data for ERA and geospatial analysis provides methods for examining events, patterns, and processes that operate on or near the surface of our planet (de Smith et al 2009).

The definition of risk varies, but in a decision theoretic framework it usually combines the probability of an undesirable event and the severity of consequences. A thorough assessment of the probabilities and consequences likely reveals that they themselves are combinations of various factors. In geospatial context risk assessment leads to combining various geospatial datasets to obtain maps of hazards and ecological values. Combination, or overlaying, of geospatial datasets (or layers) is a common and widely used method in geospatial analysis. Two overlay methods: a computational one, which produces a new layer from two or more existing ones, and a visual method, which produces a map by rendering layers on top of each other, are used. Straightforward overlay methods fail to take into consideration the reliability of datasets and ignore the uncertainties associated with the combination itself (Stassopoulou et al 1998).

A Bayesian network (BN) is a tool for fusing uncertain information from disparate sources according to the rules of probability theory (e.g., Jensen 2001). Formally BNs are directed acyclic graphs, whose nodes represent random variables and edges represent probabilistic dependencies between the variables. Random variables in BNs are often discretized to a small set of states. Arcs often depict causal relationships making the model an attractive way to describe a domain. Any variable in the network can be an input or output variable, and efficient algorithms exist to update information, i.e. the probability density functions (pdfs) of the variables, when new information is entered into the network. BNs are increasingly used in decision support and risk assessment contexts and also in environmental management (Uusitalo 2007). BNs have been used in geospatial context only to a limited extent. Stassopoulou et al (1998) used BNs for the purpose of classification of sites according to their risk desertification. Taylor (2003) developed a conceptual framework for assessing risk to habitat using BNs. Pullar and Phan (2007) used a BN to create a spatial model of the threats to koala populations. Using BNs in geospatial context poses both methodological and software problems, and to our knowledge there are no readily usable systems coupling BNs with geospatial software.

The Gulf of Finland, the easternmost part of the Baltic Sea, is a fairly narrow and shallow gulf that has witnessed a significant increase in maritime traffic, especially in oil transportation in the last years. The amount of oil transported via the area has grown from approximately 20 million tons in 1995 to over 140 million tons in 2007 and it is expected to be 250–300 million tons by the year 2015 (Hietala and Lampela 2007). While several measures have been taken to improve the safety of sea traffic in the area, a risk of a major oil spill accident exists. The ecological consequences of an oil spill depend strongly on the behavior of oil after the accident. The spreading and drifting of the oil slick is affected by several factors such as the type and the amount of the spilled oil, prevailing weather conditions as well as waves and currents. Although oil spills can be combated at sea at least to some extent, substantial amounts of spilled oil can reach the shoreline, especially in small and narrow sea areas like the Gulf of Finland. This may cause significant environmental damage. There are various ways to classify shoreline habitats according to their sensitivity to oil. Whatever the classification approach is, the sensitivity is usually spatially distributed.

In this paper we describe a method and a tool that was developed to assess the risks that oil transportation poses to environmental values of the Finnish shore of the Gulf of Finland. In a recent research project (SAFGOF, http://www.merikotka.fi/safgof/ 24.1.2011) a BN describing oil tanker accidents in the Gulf of Finland was developed (the structure of the BN is shown in Figure 1.). The BN encodes a set of initial conditions that was considered to cover all realistic situations. A large number of oil spill simulations were carried out to cover the initial conditions. During a previous study (Ihaksi et al. 2011) information about shoreline habitats was collected and sensitivity indicators for oil spill combating were developed.

The aim of this study was to develop a methodology and a tool for preparing risk maps by combining spatial oil spill simulation results and spatial ecological data guided by a BN.

1. MATERIALS AND METHODS

1.1. Bayesian network

We chose to use the proprietary HUGIN library for BNs in the tool development from HUGIN EXPERT A/S (http://www.hugin.com, 6.7.2011). The reasons behind the selection were (i) the SAFGOF project was already using it and thus the tool could use the same BN file format, (ii) the library is available for both MS Windows and Linux platforms in both 32 and 64 bit versions, (iii) the library has a clean C API, and (iv) there is only a limited selection of free or free and open source software for BN and our experience with them was not very positive. However, the HUGIN library is rather expensive and thus limits the usability of the tool.

The HUGIN library C API contains several opaque pointers to structures and many functions that operate on those pointers making it a kind of object-oriented system.

A version of the final SAFGOF BN for oil spill accidents is shown in figure 1. The main issue in the network from the point of view of the computations and the tool is that it contains all the oil spill simulation scenarios that have been prepared as combinations of discrete states of the variables. Besides that, the network can be arbitrarily large and contain other variables.

1.2. Geospatial software

Geospatial software is software for capturing, storing, analyzing, managing, and presenting data that are linked to location. There are two main technologies for geospatial data: raster and vector. A raster layer represents a rectangular area divided into equally sized small rectangles (cells). A raster, which has one band, has value for one attribute for each cell. Vector data consists of features that contain a geometry (a point, a line, an area, or a collection of geometries) and zero, one, or more attributes with values.

Features can be grouped in many ways, but typically vector layers are made of features that have the same data model. A raster layer can also be thought of as a vector layer that consists of rectangular features with one or more (the case of multi-band rasters) attributes. Raster attributes are always numeric, but integer values can be look-up table indexes leading to attributes with arbitrary type. A raster can have a value designated as ”no value”, thus removing the difference between rasters and vectors, where rasters have a value everywhere and vector layers represent object collections.

Algebras exist for raster layers (map algebra, Tomlin 1990) and for vector geometries (OpenGIS Simple feature geometry object model, http://portal.opengeospatial.org/files/?artifact_id=829 6.7.2011).

1.3. Geoinformatica

The Geoinformatica software (Jolma 2007) was used in this study. The reason for selecting it was its familiarity to the developers and it was also used in a previous, related study (Kokkonen et al 2010). Geoinformatica is a software stack that is based on GEOS, GDAL, GTK+, and Perl, which are all free and open source software (FOSS). GEOS (http://geos.osgeo.org 6.7.2011) provides predicate functions and operators for vector geometries. GDAL (http://gdal.osgeo.org 6.7.2011) provides access to several geospatial raster and vector data formats through a generic programming interface. GDAL does not provide map algebra functionality for rasters but Geoinformatica contains a small C library for in-memory raster math (libral). GTK+ (http://www.gtk.org 6.7.2011) is a graphical user interface (GUI) toolkit. Perl is a general high level programming language. Perl modules exist for using GDAL and GTK+ through foreign function interfaces (FFIs). Geoinformatica is, in a narrow sense, a set of geospatial Perl modules that build mainly on Perl modules for GTK+ and GDAL. In particular, Geoinformatica contains a singleton class Gtk2::Ex::Geo::Glue (referred to as Glue below), which links display objects with geospatial data objects.

1.4. Oil spill simulations

Altogether 1080 simulations, which covered the potential spreading of oil after various possible accidents under various weather conditions, were run by the Finnish Environment Institute for the SAFGOF project. The simulations were run using the program SpillMod (Ovsienko 2002). More specifically, the simulation results provide the probability of oil to be found within a certain area in the Gulf of Finland within 10 days after the accident. The simulations were run separately for three seasons (excluding winter), two oil types, six

oil spill volumes and five accident locations by using weather data from years 1996-2001. The exact parameters used in the computation are available from the authors by request.

The oil spill simulations were provided as MapInfo vector layers, consisting of adjacent geographic rectangles with attribute values. The set of rasters for the tool were computed from these layers by rasterizing and averaging over the simulation years. The resulting set had thus one raster for each accident scenario: season, oil type, spill volume, and location.

1.5. Habitat data

Habitat data consists of a vector feature layer describing the locations of species classified nationally as threatened (using a scheme of three classes: "Critically endangered", "Endangered" and "Vulnerable") or near threatened according to the criteria developed by the International Union for Conservation of Nature (IUCN 2994, Rassi et al. 2001).

1.6. Coupling geospatial with Bayesian networks

Coupling geospatial with Bayesian networks involves a methodological and a software part. The methodological part is involved with associating geospatial features or layers with variables or their discrete values in BNs. To our knowledge this is a poorly researched area and good overviews do not exist. In our case each accident scenario was associated with a specific combination of discrete values of variables in the BN. Thus, assuming a certain state of the BN, the probability for that scenario can be computed as a product of the probabilities of those values. Computationally, this means multiplying the scenario probability raster with the probability value obtained from the BN. Then, assuming an accident happens, the probability of any location being polluted by oil is the sum of the probabilities associated with each scenario. The last sum is a sum of raster layers in map algebra sense.

The technical coupling of geospatial with Bayesian networks depends of course partly on the methodology of the coupling, but in general the requirement is to use the BN computational engine with data obtained from geospatial layers and/or perform geospatial computations (map algebra and/or vector geometry computations) with beliefs obtained from the BN. In our case the requirement was to do map algebra computations with beliefs obtained from the BN.

2. RESULTS

The main result of this study is the spatial risk assessment tool, which was implemented as a plug-in for Geoinformatica. Achieving the result required two preparatory tasks: (i) development of plug-in technology for Geoinformatica and (ii) development of Perl FFI for HUGIN. With these available, the tool itself – a dialog-box and a set of callback-functions – could be developed.

2.1. Plug-in technology

The Geoinformatica distribution contains a simple graphical program (gui.pl referred to as GUI from now on) for viewing and manipulating spatial data. The GUI is written in Perl and it uses the Perl FFI for GTK+ and the PERL FFI to GDAL. The main object in GUI is Glue, which among other things is used to create and maintain a GTK+ toolbar. The Glue provides a method for registering commands, which usually means adding a button to the toolbar and hooking a callback function to it. It is a requirement in Geoinformatica that all geospatial layer classes provide certain methods so that the Glue object can publish their capabilities to the users through menus and other widgets.

Perl modules are kind of plug-ins in that a Perl program can start using new modules at will during runtime. This is simply achieved by calling the Perl function *require* using a Perl module name as an argument. The opposite, removing a module from a running Perl program is possible by simply removing the module from the internal array of modules. Thus, the technology existed and all that was required was to add the functionality to the GUI. This was achieved by (i) adding one call to the *register_command* method of Glue, and (ii) defining a new subroutine *extend* to the GUI. The call to *register_command* adds a *Extend* button to the GUI and sets the subroutine *extend* as a callback for the button. The *extend* subroutine displays a *file open* dialog-box for the user to select a plug-in – a Perl module file. If the user presses *ok*, the selected file is *required*, i.e., its code is added to the running program. In our case the developed SAFGOFExt.pm Perl

module is added to Geoinformatica and all other modules that it requires. The most important additional module is the Perl module for HUGIN. During the addition of the plug-in module its initialization code is executed. In our case the initialization adds a new button to the Geoinformatica toolbar for showing the dialog-box of the tool and builds the dialog-box, with which the user controls the BN used in the risk assessment and initiates the computation.

2.2. Foreign function interface to HUGIN

The Perl FFI for the HUGIN library (Hugin.pm) was developed within this project (Hugin Expert A/S provides FFIs for some other languages) (https://geoinformatics.tkk.fi/trac/browser/bntools/trunk/Hugin 7.7.2011).

Hugin.pm is faithful to the C API, except that it is truly object-oriented and supports perlish syntax in some cases making the use of the module easier. The module does not yet completely cover all aspects of the HUGIN library, but implements enough of the API to be usable for simple/commonly used networks. The module defines four classes Class, Collection, Domain, and Node. A Class is an object-oriented Hugin model and a Collection is a collection of models. A Domain is a runtime model and a Node is a node in a network. Node class has a method get_beliefs, which returns the beliefs as a list, basically all that is needed in the risk computation.

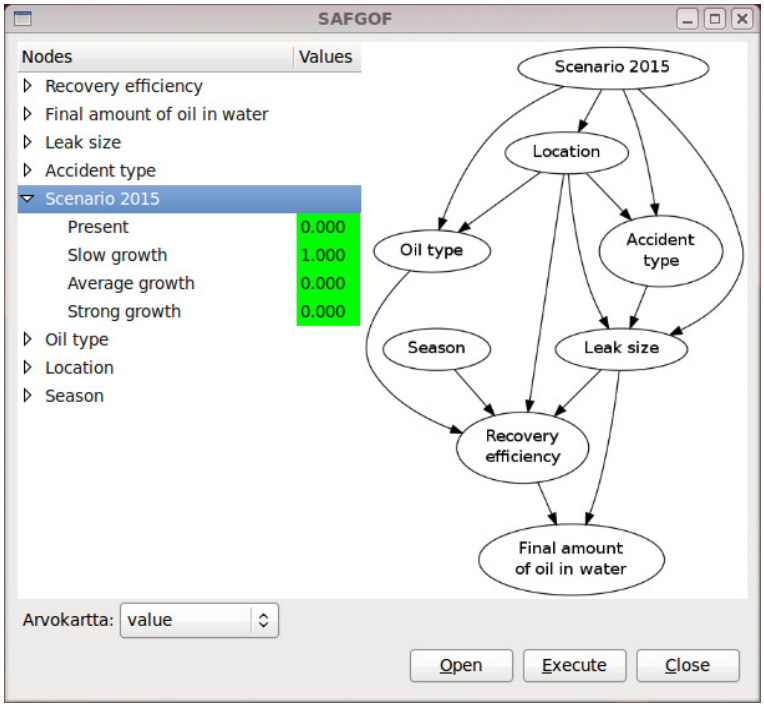

Figure 1. The plug-in dialog box. The "slow growth" option is set to occur at probability 1. "Arvokartta" is for selecting the value layer.

2.3. The Bayesian risk assessment tool

The risk assessment tool was implemented as a Perl module (code 1.), which builds a dialog-box (figure 1.) and defines callback-functions for it. The dialog-box provides an interactive table for manipulating the BN, an image of the network, a drop down list for value raster, and buttons to open a BN, compute the risk, and close the tool.

Use of the plug-in requires some preliminary work for coupling a BN with the set of geospatial environmental scenarios (raster files). When the user opens a BN into the tool, the tool looks for the coupling file. Additionally, the tool looks for an image file of the network. The coupling is defined in a text file, which contains a file name template for the scenario files and tuples, which link a specific state of a specific variable in the BN to a specific match in the file name template. The file name of each scenario must therefore contain codes that reveal the state that each scenario variable had when the scenario was computed. The plug-in contains several checks for the correctness of the coupling of the scenario rasters with the BN.

The image file was created with program "dot" from the free ImageMagick (http://www.imagemagick.org 6.7.2011) package with the help of a simple Perl program that parses a HUGIN network and writes out a dot-file. The image file is not interactive, it provided only for information. The tool mimics the graphical Hugin software. States can be instantiated by double clicking and more complex evidence can be entered by clicking and entering value to each state. Implementing this functionality required three subroutines, 60 lines of code in total.

```
declaration of the file as a Perl module;
list of referenced modules;
plug-in requirements;
dialog-box initialization;
dialog-box callback-functions;
subroutine for syncronizing between dialog-box and Bayesian network;
subroutine for reading in the mapping between raster data sets and BN states;
```

Code 1. Pseudo code for the plug-in. See text for explanation. The Perl module is available at https://geoinformatics.tkk.fi/trac/browser/safgof/trunk (7.7.2011)

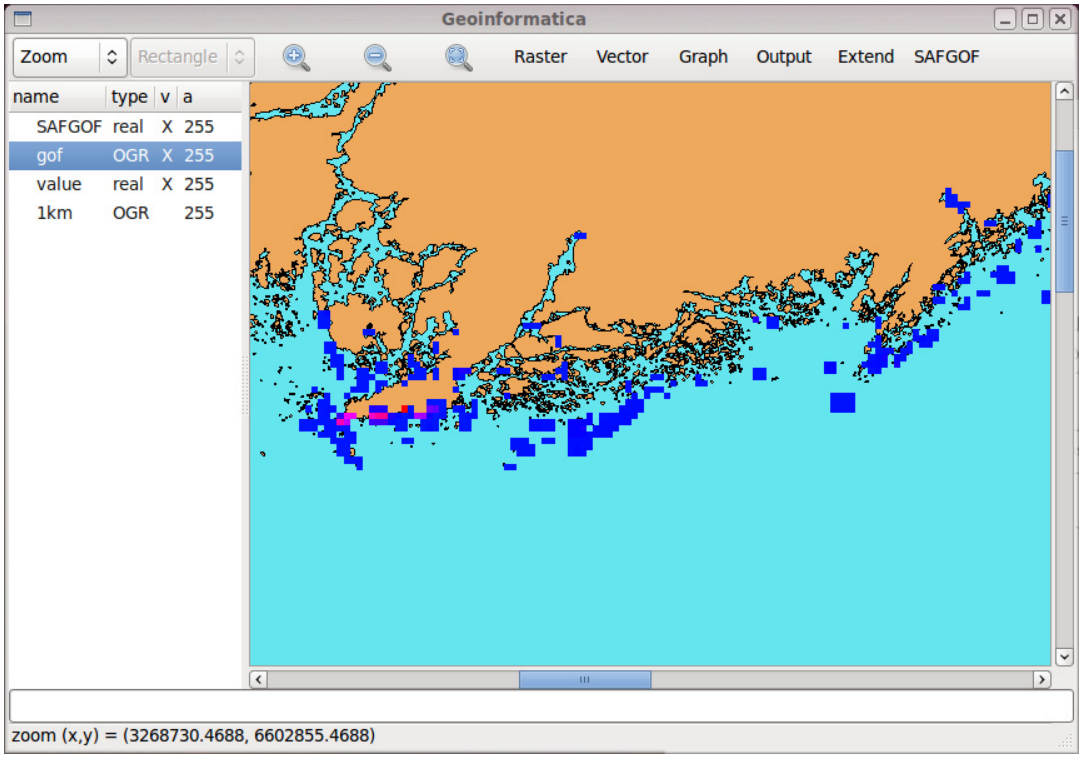

Figure 2. A risk map, where the risk layer (blue to red rectangles) are created using the SAFGOF tool. The legend for the colors is in the color-setting dialog-box, not in the main window. See text for more explanation of the workflow for creating the risk map.

The risk analysis computation is implemented in one subroutine, the most complex in the plug-in and containing 60 lines of code. The computation proceeds as follows: (i) retrieve all probability data from the network and create and associative array of tuples (node_name, state_name, probability), (ii) get the value raster from Glue using the name from the drop down list, (iii) compute a combined probability map using the coupling (raster files are read in and discarded as the computation proceeds), (iv) multiply the combined probability raster with the values raster, and (v) add the result to the Geoinformatica interface or replace the data of an existing result (preserving the no data value if one is set). Despite seemingly rather heavy computation, using the tool is very fast.

2.4. Simple risk assessment workflow

Figure 2. depicts the result of an example risk assessment workflow. The user has first extended the basic Geoinformatica GUI with the SAFGOF tool and opened a vector layer with pre-computed environmental values (the *1km* vector layer). She has also added a map of the coastal area (the *gof* vector layer). She has then opened a BN into the SAFGOF tool and executed it once (the *SAFGOF* raster). She has then rasterized

the *1km* vector layer using *SAFGOF* raster as a model for the size and discretization (the *value* raster). The *value* raster is then available as an environmental value raster in the tool (this requires a closing and opening the dialog-box). She has then used the *value* raster in the tool and after setting a state of interest in the BN she has executed the tool. She has then adjusted the colors of the resulting risk map (*SAFGOF* raster with large areas having no-data values) and the background map for better visual impression.

3. DISCUSSION AND CONCLUSIONS

Many environmental risks have a significant spatial aspect due to uneven spatial distribution of hazards and values. Risk assessment inherently requires probabilistic computations, which are commonly done using Bayesian networks. It is somewhat surprising that coupling geospatial software with Bayesian network engines seems to be a methodologically and technically poorly researched area.

In this paper we have described a tool that was developed for a specific case study – assessment of environmental risks created by increasing oil transportation in the Gulf of Finland. The results of the case study, and thus the use of the tool, are still forthcoming, but the initial impression of the methodology and the tool is that they are effective and useful. The computations and assessments can be carried out in a graphical interactive environment using MS Windows or Linux based workstations or distributed systems, and they can be carried out also in batch mode with little additional work.

A high-level language (Perl in this case) proved to be a fast development platform for creating complex information systems that couple different types of software tools.

REFERENCES

de Smith, M.J., Goodchild, M.F. and Longley, P.A. (2009). *Geospatial Analysis,* 3rd ed. Matador. (http://www.spatialanalysisonline.com/ 6.7.2011)

Hietala, M. and Lampela, K. (2007). Oil pollution preparedness on the open sea – Final report of the working group. The Finnish Environment 41/2007. 42 pp. In Finnish, abstract in English.

Ihaksi, T., Kokkonen, T., Helle, I., Jolma, A., Lecklin, T., and Kuikka, S. (2011). Combining conservation value, vulnerability, and effectiveness of mitigation actions in spatial conservation decisions: an application to coastal oil spill combating. *Environmental Management.* 47(5): 802–813.

IUCN (1994). The IUCN Red List of Threatened Species. Categories and Criteria. (http://www.iucnredlist.org/technical-documents/categories-and-criteria 8.3.2011)

Jensen, F.V. (2001). *Bayesian Networks and Decision Graphs*. Statistics for Engineering and Information Science, Springer–Verlag, New York.

Jolma, A. (2007). Geoinformatica: a modeling platform built on FOSS. In: D. A. Swayne and J. Hrebicek, editors, Proceedings of the 6th International Symposium on Environmental Software Systems (ISESS 2007), 22-25 May, Prague, Czech Republic.

Kokkonen, T., Ihaksi, T., Jolma, A., Kuikka, S. (2010). Dynamic mapping of nature values to support prioritization of coastal oil combating. *Environmental Modelling & Software.* (25)2, 248–257.

Ovsienko, S. (2002). An updated assessment of the risk for oil spills in the Baltic Sea area. (http://www.helcom.fi/stc/files/shipping/RiskforOilSpillsReport2002.pdf 8.3.2011)

Pullar, D.V. and Phan, T.H. (2007). Using a Bayesian network in a GIS to model relationships and threats to koala populations close to urban environments. In: Les Oxley and Don Kulasiri, MODSIM 2007: International Congress on Modelling and Simulation, University of Canterbury, Christchurch, New Zealand, 1370–1375.

Rassi P., Alanen A., Kanerva T. and Mannerkoski, I. (eds) (2001). The 2000 Red List of Finnish species. Ministry of the Environment. 432 pp.

Stassopoulou, A., Petrou, M. and Kittler, J. (1998). Application of a Bayesian network in a GIS based decision making system. *Int. J. Geographical Information Science.* 12 23–45.

Taylor, K. (2003). Bayesian belief networks: a conceptual approach to assessing risk to habitat. MSc thesis. Utah State University. 126 p.

Tomlin, C.D. (1990). *Geographic Information Systems and Cartographic Modeling*. Prentice Hall. 572 p.

Uusitalo, L. (2007). Advantages and challenges of Bayesian networks in environmental modeling. *Ecological Modelling.* 203: 312–318.

Xu, X., Lin, H. And Fu, Z. (2004). Probe into the method of regional ecological risk assessment—a case study of wetland in the Yellow River Delta in China. *J Environ Manage.* 70(3) 253–262.