MODELING AND SOFTWARE ENGINEERING FOR DECISION MAKING USING REGIONAL DEVELOPMENT INDICATORS

Data de aceite: 02/10/2023

Izan Fabrício Neves Calderaro
Master in Regional Development and Environment (PGDRA/UFRO). TI Analyst of Federal University of Rondônia Foundation, Porto Velho, Rondônia (Brazil). Member of GEITEC/UFRO/CNPq. Brazil.

Fabrício Moraes de Almeida.
PhD in Physics (UFC), with post-doctorate in Scientific Regional Development (DCR/CNPq). Software Engineering Specialist (FUNIP). Researcher of the Doctoral and Master Program in Regional Development and Environment (PGDRA/UFRO). Industrial Technological Development Scholarship – CNPq (DTI - A). Leader GEITEC — Federal University of Rondônia, Brazil.

David Lopes Maciel
Master of Science in Emergent Technologies in Education. MUST UNIVERSITY, MUST, EUA. Academic of the Doctoral Program in Regional Development and Environment (PGDRA/UFRO).

Carlos Alberto Paraguassú-Chaves
PhD in Health Sciences, University of Brasília - UnB, Brazil and PhD in Science - University of Havana, Cuba and Post-Doctor in Health Sciences, UnB and Degli Studi D'Aquila University, IT, Professor at the University Institute of Rio de Janeiro, IURJ, Brazil.

Fábio Machado de Oliveira
Doctor in Cognition and Language (UENF), with postdoctoral in Cognition and Language (New Information Technologies) – UENF. Rio de Janeiro, Brazil. Professor at Faculdade Metropolitana São Carlos – FAMESC.

ABSTRACT: The book chapter presents the modeling and software engineering necessary to produce analysis considering a consistent data base, a knowledge base can be developed through the use of indicators, which in turn is the link making it possible to build a data model used in the decision-making and the beginning of conscious and discriminated exploitation of resources involved. Within the context of systems that act rationally, two main approaches can be used: logical reasoning and probabilistic reasoning. The logical reasoning ponders prior knowledge about the problem and on this basis of knowledge draws its conclusions. This approach, although powerful, can not be useful in situations where not previously know the whole scope of the problem, for these cases, probabilistic reasoning comes as a good option. This raises a major challenge to development

strategies of nations and economically underdeveloped regions, since, candidates as are the economic inclusion, are faced with an environmental condition created historically by other regions which are already at the forefront of the development process. On the other hand, the new environmental awareness that is growing every day, it also brings opportunities that can and should be considered in the development strategies of these regions. Was used for this UML methodology, which is a unified modeling language that allows represent a standardized way system, using the C language for the development of algorithms and PostgreSQL as the database manager for storing variables. Therefore, it is proposed in the study, the development of a modeling of an application that measures the degree of importance of the indicators and transforms them into values that can be used as elements in decision making, making use of a system that can act in situations of uncertainty. The results are set for processing the indicators by Bayesian theory and analysis.
KEYWORDS: Modeling. Software Engineering. Indicators, Regional Development.

MODELAGEM E ENGENHARIA DE SOFTWARE PARA TOMADA DE DECISÃO USANDO INDICADORES DE DESENVOLVIMENTO REGIONAL

RESUMO: O capítulo de livro apresenta a modelagem e a engenharia de software necessária para produzir análise considerando uma base de dados consistente. Uma base de conhecimento pode ser desenvolvida através do uso de indicadores, que por sua vez são o elo tornando possível a construção de um modelo de dados empregado na tomada de decisão e início da exploração consciente e discriminada dos recursos envolvidos. Dentro do contexto dos sistemas que agem racionalmente, duas abordagens principais podem ser utilizadas: raciocínio lógico e raciocínio probabilístico. O raciocínio lógico pondera sobre o conhecimento prévio a respeito do problema e sobre esta base de conhecimento retira suas conclusões. Esta abordagem, apesar de poderosa, pode não ser útil em situações onde não se conhece previamente todo o escopo do problema, para estes casos, o raciocínio probabilístico surge como uma boa opção. Assim o objetivo do trabalho foi realizar a análise de indicadores para modelagem de sistema que forneça subsídios decisivos para a tomada de decisão, um tipo de exploração que busque minimizar o impacto ambiental e desenvolver formas mais sustentáveis de produção na utilização dos recursos naturais. Foi usada para isto a metodologia UML, que é uma linguagem de modelagem unificada que permite representar um sistema de forma padronizada, utilizando a Linguagem C para desenvolvimento dos algoritmos e o PostgreSQL como o gerenciador de banco de dados para o armazenamento das variáveis. Com esse objetivo, o estudo teve como proposta o desenvolvimento da modelagem de uma aplicação, que mensure o grau de importância dos indicadores e os transforme em valores que possam ser utilizados como elementos na tomada de decisão, utilizando para isso um sistema que possa atuar em situações de incerteza. Os resultados estão definidos pelo tratamento dos indicadores pela teoria Bayesiana e sua análise.
PALAVRAS-CHAVE: Modelagem. Engenharia de Software. Indicadores, Desenvolvimento Regional.

1 I INTRODUCTION

The indicators are essential management tools in the monitoring and evaluation activities, as well as their projects, programs and policies, because they allow to monitor the achievement of the goals, identify advances, quality improvements, correction of problems, needs for change (MPOG, 2009).

It can be said that the indicators have, minimally, two basic functions: (i) The first is to describe through the generation of information the real state of events and their behavior; (ii) The second is of an evaluative nature that consists of analyzing the information present based on the previous ones in order to make evaluative propositions.

For Sommerville (2015), "a software process is a sequence of activities that leads to the production of a software product". The software process is complex and depends on people to make decisions and carry out considerations.

2 I MODELING OF INDICATORS FOR REGIONAL DEVELOPMENT

For Prescott-Allen (1999), the realization of sustainable development is characterized by a robust economy, rich and flexible natural systems and prosperous communities. But to achieve this development requires planning and especially monitoring. And in this logic, sustainability indicators minimize the chances of obtaining unintended results.

For Moldan and Bilharz (1997), decisions are made within all spheres of society and are influenced by values, traditions and a series of in-puts from various directions. The effectiveness and rationality of the process can be increased by the appropriate use of information, and indicators can help in the decision-making process.

As Bakkes (1994) states, several steps can be identified for the decision-making process in the context of sustainability and its indicators; problem identification, policy development and control.

For Bellen (2006) models of sustainability indicators are used to draw a model of reality, evaluate conditions and trends, compare situations and places, evaluate goals and objectives, anticipate future conditions and trends.

Therefore, an analytical and well-developed structure of indicators allows to integrate, in a concise way, information of a social, ecological, economic and geographical nature, with different degrees of importance.

2.1 Indicators, regional and sustainable development

Through well-designed indicators, it is possible to set goals and objectives and measure the achievement of results in a clear and transparent way, avoiding pitfalls and distortion of planning. According to Hardi and Barg (1997), such indicators serve to identify variations, behaviors, processes and trends; establish comparisons between countries and between regions; indicate needs and priorities for policy formulation, monitoring and evaluation; and, by their capacity for synthesis, they are able to facilitate the understanding of the growing public involved with the theme.

The conception of "Development" goes through a series of theoretical and practical fields that are not confused with each other, such as History, Sociology and Economics, but it is important to highlight at the outset the specialization of the term.

Development, in principle, is generically what is in motion, in process. Narrowly, it consists of a notion of advancement and progress, in the sense of continuously reaching points of overcoming results, in a chain of interrelations and driving forces.

Thus, according to Franzin; Almeida; Souza (2014), not everything that is in Development as a process is Development as an advance of subsistence conditions and improvement of life. Regional Development is, therefore, an expression that is beyond what is produced or done in daily routines, such as actions in motion, in the circles of collective experiences.

It is also important to anticipate that "Development" is not limited to profitability. To Franzin; Almeida; Souza (2014), expectation of reducing a whole set of actions and results to the objectives of economic activity builds or feeds the ideology of alienated capitalism, in which sustaining or risk factors are taken as insignificant elements, making the human condition itself insignificant. We have clearly expressed that capitalism is not alienated but alienating.

Environmental problems are challenges for the economic sciences and their analytical instruments need to be able to provide consistent answers to harmonize the relations between environment and economy.

For Oliveira (2012), society and economic activities are inevitably dependent on the goods and services provided by the environment so that it is fundamental that economic theory considers the interconnections between economic system and external environment.

> The environmental issue and the problem of sustainable development are not restricted only to the environmental sphere, but include in the debate issues related to social and economic structures, which involve a high degree of complexity, uncertainty and ignorance, in addition to an ethical imperative related to the treatment of future generations by present generations (OLIVEIRA, 2012).

It is observed that the intention is firstly, according to Amazonas (2001), to contest the idea that environmental biophysical limits constitute limits to economic growth, arguing that technological innovations induced by the scarcity of environmental resources tend to overcome the restrictions placed by them. The other purpose is to propose that environmental damage be understood in terms of social costs, which must be internalized. Thus, while the economics of natural resources empties the concerns with social justice and equity when it presupposes that technical progress and substitution between resources are able to overcome such constraints, the economics of pollution proposes that environmental costs be internalized through the valuation of environmental resources. Thus, the two neoclassical approaches, although distinct, are complementary in their purpose.

> The first works were published in the early 1960s. In 1963, Barnett and Morse published Scarcity and Growth: The Economics of Natural Resource Availability, a seminal work. In the previous year, a statistical analysis of the historical evolution of resource prices had been published by Christy and Potter (OLIVEIRA, 2012).

Neoclassical economics, being hegemonic, has been standing out in the treatment given to environmental problems. Its theoretical formulations have as their principle methodological individualism, utilitarianism and balance, developing an understanding of the economic system as being constituted, basically of individuals who behave on the basis of a rationality maximizing well-being, which leads to an optimal result in a situation of equilibrium. Neoclassical environmental economics, according to Pearce (2002), emerged with the creation of the American institution Resources for the Future (RFF) that sought to use the theories and instruments of economics to study environmental issues. Its initial focus was the scarcity of natural resources.

> In this decade, the first environmental revolution occurred with the publication of Silent Spring, by Rachel Carson, in 1969, in which she sought to warn about the external effects of pesticides on the environment. Economists began to use the concept of externalities to interpret the growing environmental concerns (PEARCE, 2002).

2.2 The Bayesian Networks

According to Charniak (1991), the best way to understand Bayesian Networks is to imagine trying to model a situation in which chance plays an important role, but in which the understanding of what is really happening is incomplete. Thus, it is necessary to describe the situation in a probabilistic way.

According to Jensen (2001) the Bayesian Network can be defined as:

- A set of variables and a set of edges directed between the variables;
- Each variable has finite and mutually exclusive states;
- The variables and directed edges represent a directed acyclic graph (DAG);
- Each variable A, with pais B_1, B_2,... B_n, has a table of conditional probabilities $P(A \vee B_i...B_n)$, associated.

For Neapolitan et al. (2004), Bayesian networks are graphical structures to represent the probabilistic relationships between a large number of variables and to make statistical inference with these variables.

> Bayesian networks allow efficient and effective representation of the joint probability distribution over a group of random variables. The goal of probabilistic graphical models is to create a mathematical structure that unites

> graphs and probabilities and that allows modeling complex situations involving randomness or uncertainty (NEAPOLITAN et al., 2004).

"The chain rule holds for both discrete conditional distributions and continuous distributions if the Markov condition is satisfied. RB with discrete variables satisfy the Markov condition" (NEAPOLITAN et al., 2004), which is given by: each variable of the Bayesian network is conditionally independent of the set of all non-descendants of it, given the set of all its parents.

An important aspect of a Bayesian network refers to its structure (graph topology), which allows the representation of complex relationships between variables graphically and intuitively. The graphical structure of a Bayesian Network facilitates the understanding of the relationships between variables of its domain, as well as allowing the combined use of information obtained from the knowledge of experts and historical data to obtain the joint distribution of probabilities of the network.

2.3 The Bayesian inference

There are some difficulties in using the Bayesian model. According to Berg and Insua (1996) they are: the choice of the a priori distribution and the computation of the chosen model. However, the choice of a priori distribution is considered the biggest problem. This may well be the case in which subjective knowledge about unknown parameters is evaluated and can be incorporated into the subjectivity proper to the a priori densities for these parameters. This is clearly desirable if it can be accomplished. Some recent computational tools have allowed the application of Bayesian methods to models of high complexity and non-standardized. In fact, for more complicated models, Bayesian analysis has perhaps become the simplest and often the only method of analysis.

According to Berg and Insua (1996) the basis of the application of Bayesian theory is conceptual and practical: it provides a coherent framework and facilitates the analysis of decision problems over uncertainties. Criticisms of Bayesian methods are centered on three aspects: computational, imprecise, and descriptive.

Statistical inference refers to drawing conclusions about unobserved quantities θ from observed data y. Bayesian inference addresses the problem by defining probability in a subjective way, as a measure of the plausibility of a proposition, conditional on the observer's knowledge. The uncertainty with respect to θ can assume different degrees, which are represented through probabilistic models for θ. Therefore, both the observable quantities and the parameters of the statistical model are considered random quantities.

> This last characteristic constitutes a fundamental difference between the Bayesian and the classical approach, which considers the parameter as a fixed and unknown quantity, to which we approach in the inference process (BERNARDO and SMITH, 1994).

“From the practical point of view, Bayesian modeling begins with the specification of a complete probabilistic model, through the distribution of the observable and unobservable quantities of the problem” (GELMAN et al, 1995). The available information about θ, summarized in the probability density *p* (θ), is increased by observing a random quantity y that relates to θ. Bayes’ theorem provides the updating rule for this information.

According to Carlin and Louis (2000), the most basic Bayesian model has two stages, with a likelihood specification *f* (*y* ∨ θ) and an earlier specification π(θ) where both can be vectors. In this simpler Bayesian analysis π is assumed to be known, so that the probability calculation and the subsequent distribution are given as follows:

$$p(\theta \vee y) = \frac{f(y \vee \theta)\pi(\theta)}{m(y)} (1).$$

Where:

$$m(y) = \int f(y \vee \theta)\pi(\theta)d(\theta)\,(2).$$

The marginal density of the data *f* (*y* ∨ θ) is a special case of Bayes’ theorem, this evaluation of integrals used to be difficult or impossible forcing the Bayesian method into unattractive approximation. “However with the recent development of Monte Carlo computational methods allowed an accurate estimation of the value of such integrals, enabling advanced Bayesian data analysis” (CARLIN; LOUIS, 2000).

Consider an unknown amount of interest θ (typically unobservable). The information we have about θ, summarized probabilistically through *p* (θ), can be increased by observing a random quantity X related to θ. The sample distribution *p* (*x* ∨ θ) defines this relationship.

“The idea that after observing *X* = *x* the amount of information about θ increase is quite intuitive and Bayes’ theorem is the update rule used to quantify this increase in information” (CARLIN; LOUIS, 2000).

$$p(\theta \vee x) = \frac{p(\theta, x)}{p(x)} = \frac{p(x \vee \theta)p(\theta)}{p(x)} = \frac{p(x \vee \theta)p(\theta)}{\int p(\theta, x)d\theta} (3).$$

Note that 1/*p* (*x*), which does not depend on θ, functions as a normalizing constant of *p* (θ ∨ *x*).

For a value of *x*, the function *l* (θ; *x*) = p (*x* ∨ θ) provides the plausibility or likelihood of each of the possible values of θ while *p* (θ) is called an a priori distribution. These two sources of information, priori and likelihood, are combined leading to the a posteriori distribution of θ, p (θ ∨ *x*). Thus according to Carlin and Louis (2000), the usual form of Bayes is:

$p(\theta \mid x) \propto l(\theta ; x)\, p(\theta) (4).$

In words we have to;

$$distribution a posteriori \propto verisimilitude x distribution a priori$$

Note that by omitting the term p(x), the equality in equation (3) has been replaced by a proportionality. This simplified form of Bayes' theorem will be useful in problems involving parameter estimation since the denominator is only a normalizing constant. In other situations, such as model selection, this term plays a crucial (CARLIN; LOUIS, 2000).

It is also clear that the concepts of priori and posteriori are relative to that observation being considered at the moment. Thus, $p(\theta \vee x)$ is a posteriori of θ with respect to Y (which has already been observed) but is a priori of with respect to (which has not been observed yet). After observing $Y = y$ a nova posteriori (relative to $X = x e Y = y$) is obtained by applying Bayes' theorem again. But does this final posteriori depend on the order in which the *xey* observations were processed? (CARLIN; LOUIS, 2000).

Observing the quantities x_1, x_2..., x_n, independent given θ and related to θ through p_i $(x_i \vee \theta)$ follows that;

$$p(\theta|x_1) \propto l_1(\theta; x_1)p(\theta)$$

$$p(\theta|x_2 x_1) \propto l_2(\theta; x_2)p(\theta \vee x_1)$$

$$\propto l_2(\theta; x_2)l_1(\theta; x_1)p(\theta)$$

....

....

....

$$p(\theta|x_n, x_{n-1}, \dots, x_1) \propto \left[\prod_{i=1}^{n} l_i(\theta; x_i)\right] p(\theta)$$

$$\propto l_n(\theta; x_n)p(\theta|x_{n-1}, \dots, x_1)(5).$$

That is, the order in which observations are processed by Bayes' theorem is irrelevant. In fact, they can even be processed into subgroups.

3 | MATERIALS AND METHODS

These steps are respectively the processes of definition, grouping, quantification, selection, joining and qualification of the defined indicators.

3.1 Quantitative method

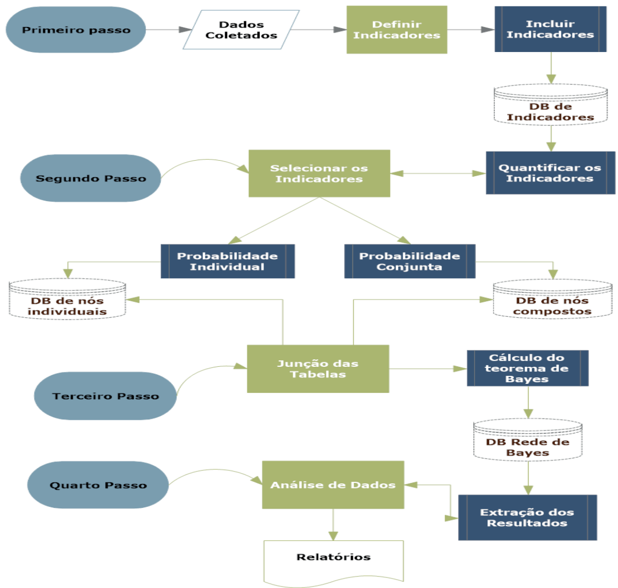

Figure 1 – Diagram of analysis of indicators. Source: Calderaro and Almeida (2014).

The analysis diagram faithfully represents the central idea of this work, the leanest scheme was defined, always avoiding unnecessary overload, that is, the graphical representation of objects that are not essential to the model to be produced with graphical representation in UML diagrams.

3.2 Qualitative method

The method of quantitative analysis obeys a pre-established plan, in order to enumerate or measure events, examines the relationships between variables by experimental or semi-experimental methods, rigorously controlled, generally employs, for data analysis, statistical instruments, confirms the hypotheses of the research or findings by deduction, that is, performs observations or experiments, uses data representing a specific population (sample), from which the results are generalized, and uses as an instrument for data choice, structured questionnaires, elaborated with questions, tests and chec-klist, applied from individual interviews, supported by a conventional (printed) or electronic questionnaire (NEVES, 2006).

Figure 2 – Storage, quantitative analysis and selection of data for the knowledge base. Source: Calderaro and Almeida (2014).

This part of the diagram is represented by schemas (Schema) in the database, each schema has its own elements.

3.3 Methodology

This article presents as research methodology: bibliographic and experimental, because it is applied in the set of pre-established indicators and adapted in the database the Bayesian network based on the collected indicators.

The bibliographic research is developed based on material already elaborated, consisting mainly of books and scientific articles. Although in almost all studies some kind of work of this nature is required, there is research developed exclusively from bibliographic sources. Most exploratory studies can be defined as bibliographic research. Research on ideologies, as well as those that propose the analysis of the various positions on a problem, are also usually developed almost exclusively through bibliographic sources (GIL, 2008).

Overall, the experiment represents the best example of scientific research. Essentially, experimental research consists of determining an object of study, selecting the variables that would be able to influence them, defining the forms of control and observation of the effects that the variable produces on the object (GIL, 2008).

Experimental research, on the other hand, consists essentially of determining an object of study, contrary to what the popular conception suggests, it does not necessarily need to be carried out in the laboratory. It can be developed anywhere, as long as it presents the following properties: manipulation, control and random distribution (GIL, 2008).

The case study that will be presented has the following research characteristics:

- Survey of bibliographic material.
- Modeling of indicators.

The indicators used for modeling in this case study were extracted from the database of the Monitoring of the Brazilian Amazon Rainforest by Satellite (PRODES) and were adapted for storage in the DBMS.

- Applied Study of Statistics, Probability and Bayesian Computing:

It seeks to define a software modeling for decision-making, based on beliefs, defining the degree of uncertainty, heuristic naivety and a simple probabilistic classifier based on the application of Bayes' theorem.

- Graphical application of Bayesian Computing with UML:

The entire computational process that will be developed in this modeling will be represented graphically through UML and its diagrams.

3.4 Bayesian indicators and networks

When it comes to modeling, the first impression is its similarity with the concept of art, that is, human activity linked to manifestations of an aesthetic order made by artists from perception, emotions and ideas. The development of systems is an art, it is the interpretation of a concept previously outlined to be the solution of the scope of a problem, but for this there are steps that must be followed for the conclusion of this goal.

3.4.1 Variables and indicators

When it comes to software modeling the starting point is the prerequisites, reasons, that led to the occurrence of the observed fact. In the scope of this study these prerequisites are the variables that influence deforestation. The bibliographic analysis pointed out the need to add some variables to the existing databases, seeking to relate deforestation with socioeconomic performance. According to Pereira and Góes (2013), Neves (2006), Santos (2007), Cavalcante (2013), Kampel and Câmare (2000) and Diniz; Junior; Grandson; Diniz (2009) shows in table 1 some variables considered for deforestation in the Amazon and the Atlantic Forest.

Authors	Variables	Methods
(PEREIRA; GÓES, 2013).	Macroeconomic and deforestation variables: 1. Consumption plus government spending 2. Government spending 3. Consumption 4. GDP 5. FBKF 6. Exports 7. Deforestation	The model proposed in this work follows the work of Hofler and Mikhail (2002) and John and Pecchenino (1994), in order to present the concept of environmental quality to dynamic and stochastic macroeconomic constructions. The most important point of the model is the introduction of the simplest possible form of trade-off between the loss of well-being generated by deforestation and the gain of aggregate product resulting from the use of wood in the production function. More production means more consumption of goods. So the classic trade-off between consumption and leisure now has one more component, which is the forest.
(NEVES, 2006).	Dependent variable: Deforestation. Variables from the following sources: 1. . Agricultural Censuses a. Personnel Employed in Agricultural Establishments b. Use of land for tillage c. Use of land for grazing d. Cattle Herd e. Tractor f. Firewood g. Wood in logs 2. Evolution of the Forest Remnants of the Atlantic Forest a. Atlantic Forest Remnants 3. IPEA Date a. Municipal GDP b. Industrial municipal GDP c. Municipal service GDP d. Municipal agricultural GDP e. Human Development Index (HDI) f. Cost of transportation from the seat of the municipality to the state capital.	The work uses two econometric tools, the correlation and the multiple linear regression model, to identify the types of relationships that the variables have. The hypothesis adopted is that the conversion of forest areas for agricultural use, that is, pastures and crops, is a determining factor of deforestation in the Atlantic Forest region. Thus, we seek to verify if this hypothesis is confirmed and if there are other factors of deforestation pressure in the Atlantic biota.
(SANTOS, 2007).	Spatial variables 1. Country 2. State 3. Municipality 4. Census tract Synthesized variables 1. Population data 2. Agricultural activities 3. Livestock activities 4. Activities of the timber sector The search for data from the implementation of policies to combat and prevent deforestation was unsuccessful because there were no systematized data on government actions.	From the list of variables obtained from other studies, the factors that could be tabulated were defined to be spatially and econometrically analyzed as those related to livestock and agriculture activities, the timber sector and population data. They are present in almost all econometric analyses, always presenting great relevance to explain/predict deforestation rates.

(CAVALCANTE; GÓES, 2013)	Quantitative data from official agencies of the Brazilian Government (IBGE and INPE/ PRODES).	The present research adopted the hypothetical-deductive method. The spatial section on which the basis of analysis of this work was constituted was based on the micro-regional perspective of Rondônia.
(KAMPEL; CÂMARA, 2000)	We used deforestation data from the Legal Amazon, from the Deforestation Project - PRODES of INPE (1999), in the form of deforestation rate (km2/year) on the municipal network of the IBGE of 1994.	Exploratory analysis of the spatial dependence of variables. Verification of spatial autocorrelation of variables. Verification of spatial associations. Spatial regression between variables.
(DINIZ; JUNIOR; NETO; DINIZ, 2009)	Deforestation Variables 1. Total hectares deforested per municipality Variables of the agricultural sector 1. Cattle herd 2. Bovine density 3. Permanent crops 4. Temporary culture 5. Occupied area Socioeconomic Variables 1. GDP per capita 2. Adult education 3. Registration 4. Agricultural credit 5. Population 6. Population density	The methodology to be employed is based on dynamic models for panel data, developed by Holtz-Eakin et al. (1988) and Arellano and Bond (1991), who developed a causality test based on Granger's seminal (1969) paper.

Table 1 – List of methods and variables that influence deforestation.

There are numerous studies related to deforestation that always relate the same variables, so in the modeling of a Bayesian Network we will adopt these variables extracted from sources such as the Institute of Applied Economic Research (IPEA), the Brazilian Institute of Geography and Statistics (IBGE) and the National Institute of Educational Studies and Research (INEP), based on the pertinent literature.

3.4.2 Modeling and the Bayesian network

When starting with the structure of a Bayesian network, the main concern one should have is with the representation of dependencies and conditional independences. Because a graph is used, only variables linked by directed arcs manifest dependency relationships.

According to Pearl (1986), one can define a procedure for the construction of a Bayesian network. Thus, given the joint distribution P (X_1, X_2,...,X_n) and a certain ordering of the variables, the construction of the graph begins by choosing the root node X_1 and specifying its a priori probability $\overset{X}{P(1)}$.

Next, the node X_2 is added to the graph. If X_2 for dependent on X_1 a directed arc should be inserted with a X_1 start point and a X_2 endpoint; Once this is done, the arc must be quantified with $\frac{X}{P}$ If X_2 is independent of X_1, one must assign to node X_2 the a priori probability $\frac{X}{P(2)}$ and leave the two nodes disconnected. Repeating the operation for the other variables obtains the network.

Once the order of insertion in the network is correctly defined, when inserting a node X_i it is known that its parents have already been inserted and that it is the nodes in $X_1,\ldots, X_{i-1}$ that directly influence X_i.

The general procedure for the construction of a Bayesian network is defined by Russell and Norvig (2004) in table 2.

1. Choose the set of relevant variables that describe the domain;
2. Define an ordering for the variables;
3. As long as there are variables left in the set:
    a. Select a variable and add a node for it to the network;
    b. Define the parents of with some minimal set of nodes that are already in the network, such that the conditional independence property is satisfied;
    c. Define the table of conditional probabilities of .

Table 2 – Procedure for building a Bayesian network. Source: Russell and Norvig (2004).

Based on the variables presented in table 1, in the referred studies, with the econometric analysis applied and taking deforestation as the dependent variable, we sought to verify which of these variables have a strong and significant relationship with deforestation.

Thus, in the investigation of the causes of the recent deforestation process of the Amazon and the correlation between deforestation and cattle ranching in the state of Rondônia, the analysis of the correlation coefficient is an important and necessary approach to verify if there is any relationship between the variables analyzed in the studies described in table 1.

The variables that make up the network according to the abbreviations adopted are: transportation cost (ct), timber (tor), firewood (len), tractor (trat), pasture area, cattle herd (bov), crop (lav), personnel employed in the rural area (po) Municipal Human Development Index (HDI), average agricultural productivity (prodmed), agricultural productivity variation (varprod), municipal agricultural GDP (pibagr), municipal services GDP (pibserv), industrial municipal GDP (pibind), municipal GDP (gdp).

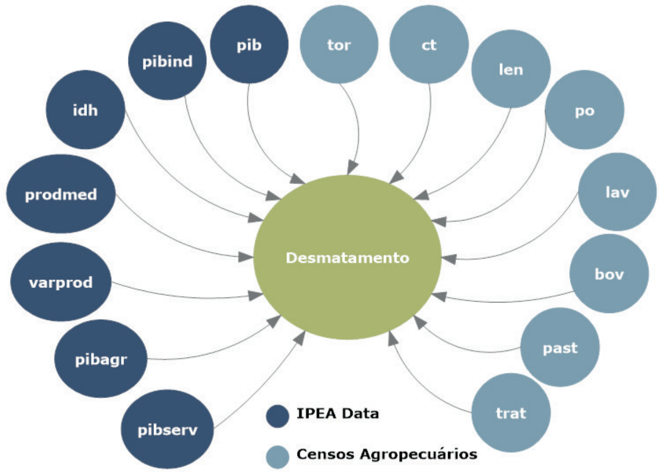

Figure 3 – Proposed model and its representation in a directed acyclic graph with convergent connection.

Where deforestation is not known then their parents are independent; otherwise, there is a correlation between their parents. The joint probability distributions for the two networks are:

Agricultural censuses:

$$P\,(ct, tor, len, trat, past, bov, lav, po, Desmatamento)$$
$$= P(ct) * P(tor) * P(len) * P(trat) * P(past) * P(bov) * P(lav) * P(po)$$
$$* P(Desmatamento \vee t, tor, len, trat, past, bov, lav, po)$$

IPEA Date:

$$P\,(idh, prodmed, varprod, pibagr, pibserv, pibind, pib, Deforestation)$$
$$= P(idh) * P(prodmed) * P(varprod) * P(pibagr) * P(pibserv)$$
$$* P(pibind) * P(pib)$$
$$* P(Deforestation \vee idh, prodmed, varprod, pibagr, pibserv, pibind, pib)$$

In several investigations it is desired to evaluate the relationship between two quantitative measures. Three main purposes of such investigations may be: (i) to verify whether the values are associated; (ii) to predict the value of one variable from a known value of the other; and, (iii) to describe the relationship between variables. The degree of linear association between two variables is assessed using correlation. While in the regression analysis the dependent and explanatory variables are treated asymmetrically, in the correlation any two variables are treated symmetrically, with no distinction between dependent and explanatory.

Despite all the positive aspects associated with the formalism of Bayesian networks, one must know the domain well before using it, because if in the domain the relationships between the variables are not of the causal type, then it cannot be guaranteed that a Bayesian network is the most appropriate structure for the representation of the dependency relations between the variables.

The model in Figure 3 does not yet define the relationship of dependence between the variables if it exists, this would be at the discretion of the causal relations that can be attributed to them, so what was proposed will undergo changes precisely to graphically define these dependencies.

4 I RESULTS AND DISCUSSIONS

According to Neapolitan (2004), the concept of Bayesian networks, originally, was developed assuming a dependence of human experts for the definition of the graph, that is, of the structure or topology of the network and for the estimation of conditional probabilities, but they can be constructed both from the knowledge of human experts and from the database, with the use of Bayesian learning algorithms.

The basic task of a Bayesian network system is to compute the conditional probability distribution for a set of query variables, given the values of a set of evidence variables, i.e. to compute the $P(variable_{consultation} \vee variable_{evidence})$;

This task is called Bayesian inference and allows you to respond to a series of "queries" about a data domain. For example, in deforestation, the main task is to obtain a diagnosis for a given deforested area by presenting certain factors (evidence). This task consists of updating the probabilities of the variables according to the evidence.

In the case of the diagnosis for deforestation, we try to know the probabilities for deforestation in a given region, of each of the possible causes of the observed variables. These are a posteriori probabilities.

4.1 Bayesian heuristics and computation with uml analysis

Describing the proposed use case diagram, which represents the data analysis process, in figure 4. In the first step, we have the storage of the indicators that are taken from the surveys presented in table 1, these indicators undergo discretization, In the second step, to be categorized in a binary way and are stored in another schema of the database, they are given the dependence on each other for a new categorization before proceeding to the third step, where there is a new selection of variables for the construction of the Bayesian model and subsequent inference performed in the fourth and final step.

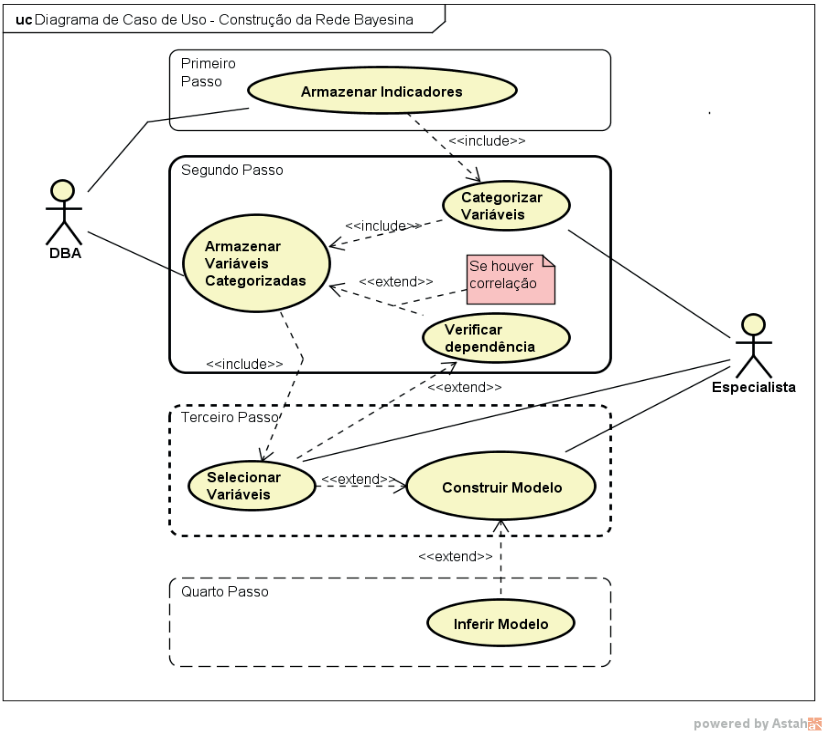

Figure 4 – Use Case Diagram. Source: Calderaro (2016).

The diagram shows the include association that is used when there is a scenario, situation, or routine common to more than one use case. Inclusion relationships indicate an obligation, that is, when a given use case has an inclusion relationship with another, the execution of the first also requires the execution of the second. An inclusion relationship can be compared to calling a subroutine or function, a device widely used in most programming languages.

The behavioral sequence diagram seeks to demonstrate the sequence of events that occur in a given process, identifying which messages should be triggered between the elements involved and in what order. Figure 4 depicts the cadence of events in the data analysis process. The two agents involved in the process: specialist and DBA exchange messages with each other demonstrating the temporality that is peculiar to this diagram and the process itself. The lifeline represents the time that an object (lifeline) exists during a process.

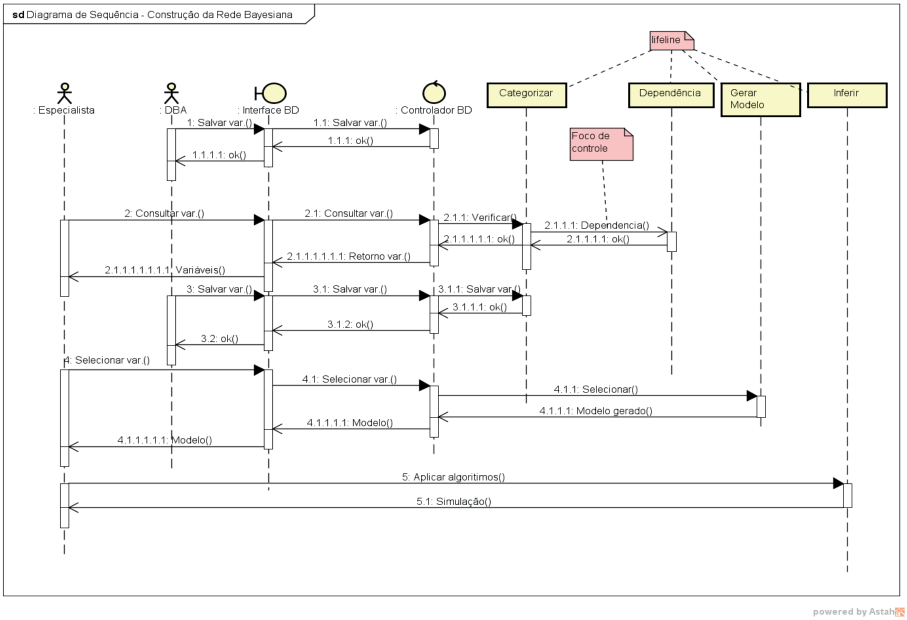

Figure 5 – Sequence Diagram. Source: Calderaro (2016).

The sequence diagram presented in figure 5 is based on the use case diagram, this is a behavioral diagram that seeks to determine the sequence of events that occur in a given process, identifying which messages should be triggered between elements involved and in order. In this diagram it is evident the exchange of messages or stimuli to demonstrate the occurrence of events, which usually force the call of a method on some of the objects involved.

As in a sequence diagram, a communication diagram focuses on a process normally based on a use case. In Figure 6 we have a communication diagram.

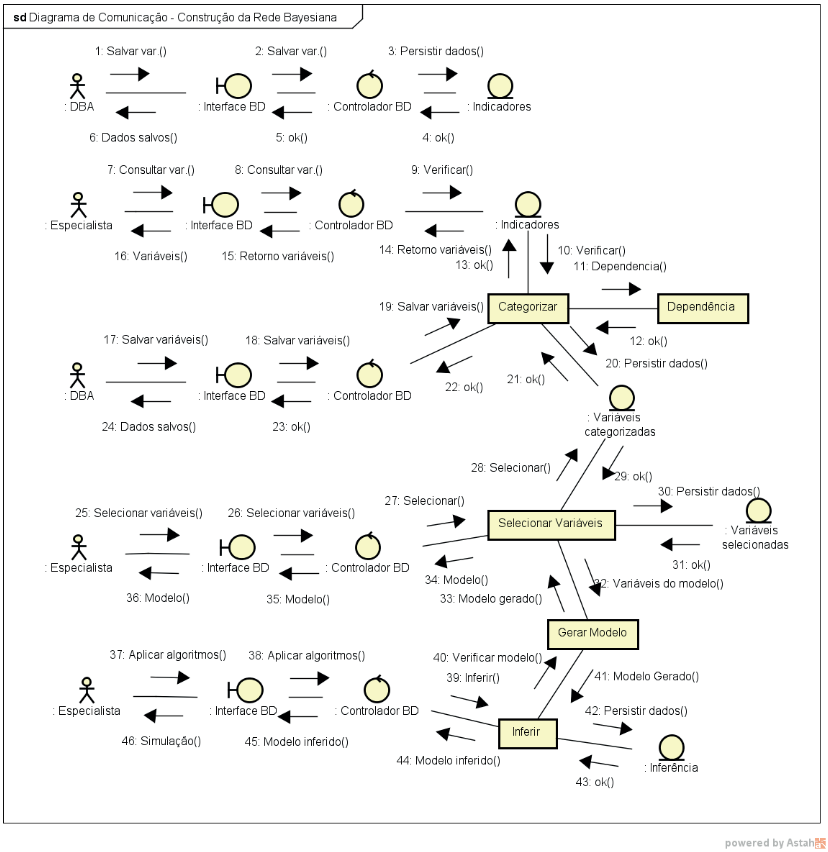

Figure 6 – Communication Diagram. Source: Calderaro (2016).

The information shown in the communication diagram is often practically the same as that presented in the sequence diagrams, but with a different focus, since these diagrams do not concern themselves with the temporality of the process, focusing on how the elements of the diagram are linked and what messages exchange with each other during a process.

The entities involved in the process of data storage, selection, categorization and inference were added to this diagram. Note that the agents involved are always exchanging messages with the database interface, performing the treatment of the indicators. The DBA agent is in charge of storing the indicators and their variables during the four steps shown in figure 4.

4.2 Database structure for Bayesian computing

The structure required to store the variables is dynamic since the indicators and their respective variables have random size, so the initial storage is at the discretion of the database administrator (DBA), just below in figure 7 there is an eventual structure of the database.

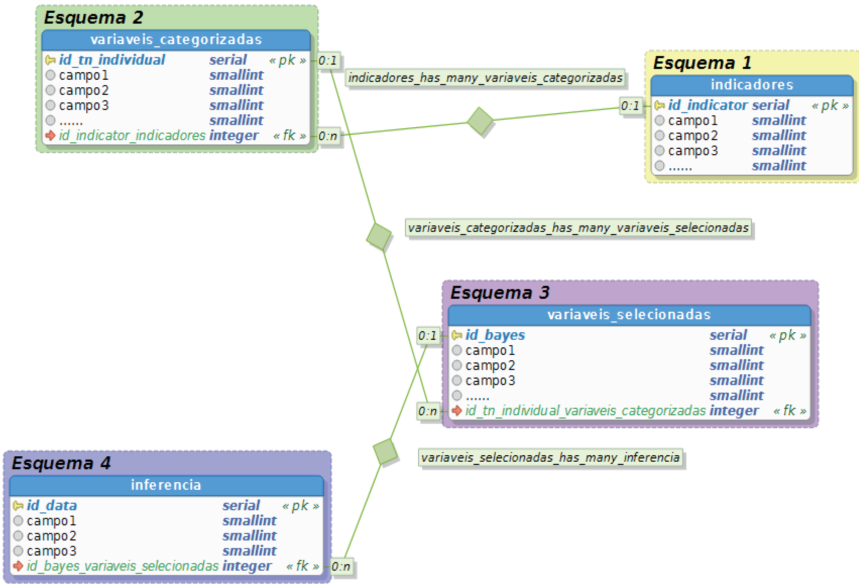

Figure 7 – Database structure. Source: Calderaro (2016).

In the relationships demonstrated between the tables, these are necessary to maintain data integrity. Other relationships and tables may be required, but these can be created dynamically during the process, only temporarily, upon return of a query.

Therefore, a possible setback is the survey of the indicators and the treatment that each variable can be conditioned in the course of the modeling experiment and the respective dependency relations between them, this defines which will be inserted in the model or not through their correlations, that is, these variables must be unconditional, just as deforestation must be dependent on these variables that influence the proposed belief network model.

A preponderant factor to be identified as risk is the inability of the model not to identify all the points necessary for the development of the application, resulting in a lack of dependence on measuring the degree of importance of the indicators and their variables, which would make their use as an element in decision making unfeasible, which may be lower than expected.

FINAL CONSIDERATIONS

Therefore, the study presented contributes to optimize modeling and software engineering approaches, that is, the development of the algorithms proposed by the modeling for the first scenario hypothesis, because all computational logic involves the applicability of these algorithms.

In the second scenario, for modeling and applications, the database structure can be developed allied to the SQL query that should be responsible for the interaction between set theory and probability and in the third and probable last scenario, it is under the responsibility of Bayesian inference applied to the database, all the heuristics for the extraction of knowledge to optimize decision making.

REFERENCES

AMAZONAS, M. C. Valor e meio ambiente: elementos para uma abordagem evolucionista. 2001, 267f. (Tese de doutorado em Economia, área de concentração: Teoria Econômica) – Instituto de Economia, Universidade Estadual de Campinas. Campinas. Disponível em: http://www.bibliotecadigital.unicamp.br/document/?code=vtls000220503. Acesso em: 21 de fevereiro de 2016.

Bakkes, J.A., Van den Born, G.J., Helder, J.C., Swart, R. J., Hope, C.W., and Parker, J.D.E. An Overview of Environmental Indicators: State of the Art and Perspectives. UNEP/EATR.94-01; RIVM/402001001. Environmental Assessment Sub-Programme. UNEP. Nairobi. 1994.

BELLEN, Hans Michael Van. Indicadores de sustentabilidade: uma análise comparativa 2.ed. Rio de Janeiro: Fundação Getúlio Vargas. 2006.

BERGER, J. O.; INSUA, D. R. Recent developments in Bayesian inference with applications in hydrology. [S.l.]: Censiglio and Nazionale Delle Ricerche. 1996.

BRASIL. Política Nacional de Desenvolvimento Regional – PNDR. Disponível em: http://www.integracao.gov.br/politica-nacional-de-desenvolvimento-regional-pndr.Acessado em: 30 julho de 2014.

CALDERARO, I. F. N., DE ALMEIDA, F.M. C-Language Functions for PostgreSQL in Modeling Software for Decision Making Through Indicators. International Science Congress Association. v. 1. p. 1-15. IVC 2014. ISBN: 978-93-83520-30-5. Disponível em: http://isca.net.co/vconference/subaccept.php. Acesso em: 10 junho de 2014.

CALDERARO, I. F. N., DE ALMEIDA, F.M. Software Development Based on Indicators. Review Of Research. v. 3. p. 1-7. 2014. DOI:10.9780/2249-894X/352014/613. Disponível em: http://ror.isrj.org/ArticleDetails.aspx?id=613. Acesso em: 05 fevereiro de 2014.

CALDERARO, I. F. N. Modelagem de software para tomada de decisão no desenvolvimento regional através de indicadores.Orientador: Prof. Dr. Fabrício Moraes de Almeida. Dissertação (Mestrado em Desenvolvimento Regional) - Fundação Universidade Federal de Rondônia. Porto Velho,Rondônia, 2016.

CARLIN, Bradley P., LOUIS Thomas A. Bayes and Empirical Bayes Methods for Data Analysis, 2nd ed. Lodon. Chapman & Hall/CRC, A CRC Press Company. 2000. 434p.

CHARNIAK, Eugene. “Bayesians Networks without Tears”. IA Magazine, 1991. 14f. Disponível em: http://www.aaai.org/ojs/index.php/aimagazine/article/view/918/836.Acesso em: 01 novembro de 2012.

DINIZ, M. B.. JUNIOR, J. N. O.. NETO, N. T.. DINIZ, M. J. T.. Causas do desmatamento da Amazônia: uma aplicação do teste de causalidade de Granger acerca das principais fontes de desmatamento nos municípios da Amazônia Legal brasileira. 2009. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-63512009000100006. Acesso em: 01 de maio de 2016.

FRANZIN, Sergio F. L., DE ALMEIDA, Fabrício M., DE SOUZA Carlos H. M. A Inovação e o Desenvolvimento Regional como Referência para Políticas públicas no Brasil. 2014. 20f. Inter Science Place. Edição 29, volume 1, artigo nº 5, Junho. 2014. D.O.I:10.6020/1679-9844/2905. Disponível em: http://www.interscienceplace.org/interscienceplace/article/view/300. Acesso em: 15 agosto de 2014.

GELMAN, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. Bayesian Data Analysis. London. Chapman & Hall/CRC, A CRC Press Company. 1995. 526p.

MPOG. Guia Referencial Para Medição De Desempenho E Manual Para Construção De Indicadores. Governo Federal. Ministério do Planejamento Orçamento e Gestão – MPOG. Brasil. Disponível em: www.gespublica.gov.br/sites/default/files/documentos/guia_indicadores_jun2010.pdf. Acessado em: 12 novembro de 2013.

HARDI, P., BARG, S. Measuring Sustainable Development: Review of Current Practice. Winnipeg. IISD. 1997.

JENSEN, F. V. Bayesian networks and decision graphs. statistics for engineering and information science. Springer. v32. p34. 2001.

KAMPEL, Silvana Amaral, CÂMARA, Gilberto. Análise Exploratória das Relações Espaciais do Desflorestamento Na Amazônia Legal Brasileira. Disponível em: http://www.dpi.inpe.br/gilberto/papers/silvana_gisbrasil2000.pdf. Acesso em: 01 maio de 2016.

NEAPOLITAN, R. E. et al. Learning bayesian networks. [S.l.]. Prentice Hall Upper Saddle River. 2004.

NEVES, Ana Carolina Marzullo. Determinantes do Desmatamento na Mata Atlântica: Uma Análise Econômica. 2006. 94p. (Dissertação de Mestrado) – Universidade Federal do Rio de Janeiro. Rio de Janeiro. Disponível em: http://www.ie.ufrj.br/images/conjuntura/Gema_Dissertaes/AnaCarolinaMarzulloNeves_2006_IE_Determinantes_do_desmatamento_na_mata_atlantica.pdf. Acesso em: 22 abril de 2016.

OLIVEIRA, Aparecida Antonia. Políticas Ambientais e Desenvolvimento Regional: A Perspectiva do Pensamento Institucionalista Evolucionário. 2012. 282f. (Tese de Doutorado em Economia) - Universidade Federal do Rio Grande do Sul. Porto Alegre. Disponível em: http://www.lume.ufrgs.br/handle/10183/69998. Acesso em: 02 fevereiro de 2013.

PEARL, J. Fusion, propagation, and structuring in belief networks. Journal of Artificial Intelligence, v.29 , p.241-288, 1986

PEREIRA, Rodrigo Mendes. GÓES, Geraldo Sandoval. O Desmatamento Amazônico e o Ciclo Econômico no Brasil. Disponível em: http://repositorio.ipea.gov.br/bitstream/11058/5563/1/BRU_n07_desmatamento.pdf. Acesso em: 22 abril de 2016.

PRESCOTT-ALLEN, R. Assessing Progress toward Sustainability: The System Assessment Method illustrated by the Wellbeing of Nations. Cambridge: IUCN, 1999.

PRODES. Monitoramento da Floresta Amazônica Brasileira por Satélite. Disponível em: http://www.obt.inpe.br/prodes/index.php. Acessado em: 30 agosto de 2015.

SANTOS, Renato Prado dos. Os Principais fatores do desmatamento na Amazônia (2002-2007) – uma análise econométrica e espacial. 2010. 129p. (Dissertação de Mestrado) – Universidade de Brasília. Brasília. Disponível em: http://repositorio.unb.br/bitstream/10482/6592/1/2010_RenatoPradodosSantos.pdf. Acesso em: 22 abril de 2016.

SOMMERVILLE, Ian. Software Engineering. 10.ed. São Paulo: Pearson, 2015.