

Exploratory Data Analysis in Ecological modelling framework

ALEXEY MIKHAILOV*, ALEXANDER KOMAROV

Institute of Physicochemical and Biological Problems in Soil Science, Russian Academy of Sciences / Laboratory of Ecosystems Modeling
alexey.mikh@gmail.com

Abstract

The goal of this article is to discuss advantages of joining visual analytic tools and computer simulation modelling in Forestry. Advantages and possibilities of this uniting have been mentioned earlier (Chertov *et al.*, 2005). Forest ecosystem simulation model, Geographical Information System and Exploratory Data Analysis are the best tools allowing for creating a decision support system (DSS) for forestry managers and scientists. As an example CommonGIS software is applied together with EFIMOD model of forest growth and elements cycling for analysis of forest ecosystem development under different silvicultural scenarios. Such an application allows for the fast and comprehensive choice of best scenarios in respect to different criteria, for example, carbon and nitrogen dynamics in soil and vegetation, highest production of merchantable tree compartments etc.

Key words: simulation model; DSS; forest ecosystem; exploratory data analysis

1 Introduction

Applications of mathematical and computer models in ecological studies are a common practice. Any ecosystem is a very complex system. Forest ecosystems cause a lot of difficulties for investigations: a) they consist of a large set of different components (plants, mushrooms, animals etc.) with interactions, b) these systems are open (dynamics of systems depend on external impacts and environmental conditions), c) environmental conditions vary greatly, both spatially and temporally, d) life span of full development of forest ecosystem is longer than lifetime of any researcher. Nevertheless prediction of forest ecosystem dynamics is very important from economical, ecological, aesthetic and other points of view. Models in use for prognosis produce a lot of different characteristics of the simulated forest. Outputs are usually complex and multidimensional: several attributes are simulated during long time intervals accordingly to different scenarios of modelling. Respectively, analysis of outputs becomes a non-trivial task. We need to have a convenient tool to look into the data from different perspectives, calculate derived attributes, analyse the dynamics of values, reveal relationships between variables, etc. Moreover decision support system (DSS) is necessary to implement multi-purpose, sustainable forest management (Næsset, 1997; Pykalainen, 2000; Kangas, 2000; Varma *et al.*, 2000).

2 Forest ecosystem models

All forest models can be classified on three main groups in relation to the model's scale: tree models, stand models and regional models.

Tree models express the growth of a single tree with different degrees of detail of the processes. Stand models deal with single forest plot. Some stand models simulate the dynamics of forest stand as a population of trees (individual-based models), other models consider stand as a patch or a metapopulation with generalised characteristics of trees. Regional models simulate forest dynamics on landscape, regional, and national scales. Necessity in Exploratory Spatial Data Analysis is appeared when stand model is used for prediction of the dynamics of the forest consisting of numbers of separate plots placed on territory. We consider applications of a stand model in terms of implementation of Exploratory Data Analysis.

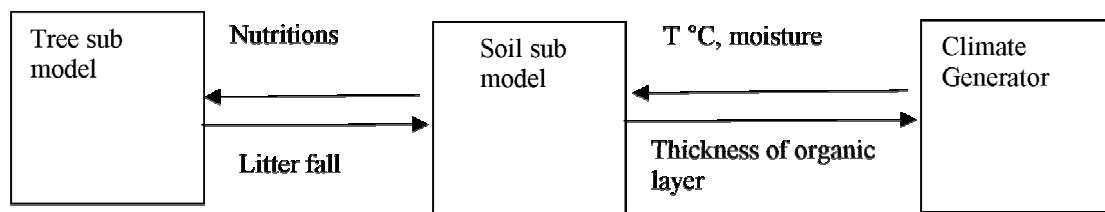


Fig. 1 Simplified chart of EFIMOD model feedbacks

EFIMOD forest model is a typical stand model, which uses individual-based approach for the description of the forest-soil system dynamics at the population level. This approach allows for the easy simulation of different types of strong external impacts: cuttings and fellings, forest fires, windfalls, insect attacks etc.). The EFIMOD model (Komarov *et al.*, 2003) consists of 3 main parts: tree sub model, soil sub model (ROMUL), statistical climate generator (SCLISS) and some additional components. The soil sub-model, ROMUL (Chertov *et al.*, 2001), simultaneously permits calculating the mineralization rate of tree litter and soil organic matter and soil organic matter humification with the corresponding carbon dioxide emission and releasing nitrogen, going into plant growth. Rates of decomposition depend on chemical composition of litter, soil temperature and moisture.

The statistical climate generator SCLISS (Bykhovets and Komarov, 2002) allows estimating of soil temperature and moisture using measured standard long-term meteorological data. Moisture and temperature of soil depends on thickness of organic layer in soil.

The tree sub model simulates the stand as a population consisting of separate trees with competition. The competition among trees takes place for available light and for available nitrogen (nutrition) coming from soil. Model allows simulating of uneven-aged and multispecies forest and utilizes standard meteorological and forest inventory data. The model outputs for every annual time step are as follows: the stand main dendrometry parameters of each forest element (even-aged pure tree cohort) (mean stand height, DBH, etc.), tree species composition, coarse woody debris, pools of soil organic matter and nitrogen. The model permits simulating of different forest management regimes. Picture 1 shows feedbacks in EFIMOD model between compartments.

3 Properties of Data

It is possible to sort out data used in the models into input data for modelling and output data.

Input data for modelling of concrete forest ecosystem dynamics may be different. Usually models apply for a simulation of the system dynamics in different conditions or/and analyzing of different impacts influence. Influence of climate change, management scenarios (type and intensity of cutting, rotation length and other), concentration of pollutions and their expansion, insect attacks, forest fires are often explored by models.

Forest ecosystem model usually produces a big amount of output data with different characteristics (numerical and non numerical, ordered and not ordered, continuous and discrete and so on). There are a lot of referrers. One of referrers is space, second important referrer is time; modelling scenarios identifier, forest type, trees species, age groups of trees and other characteristics may also be referrers.

For example forest ecosystem model EFIMOD has four subsets of input data: climate data, soil data, forest data, management and external impacts data.

Time step of output data can be different, for example, soil and climate output data has monthly time step, forest data has annual time step.

For each time step the model calculates about 30 soil output parameters, 4 climate parameters and 20 parameters for each forest element. Sometimes forest plot has the only forest element (planting) but usually natural forest has a number of forest elements. Time period of forest growth modelling depends on rotation length (time between main cuttings). Rotation length can vary (80-180 years in dependence on species and climatic zone). It is easily seen that we can calculate total amount of values for forest territory consisting of several plots:

$$V = N_{\text{steps}} * N_{\text{plots}} * (34 + 20 * N_{\text{fe}}),$$

where V - total amount of values; N_{steps} - number of steps; N_{plots} - number of plots, N_{fe} – number of forest elements.

Usual Russian forest enterprise consists of several thousands of forest plots. So we have a huge amount of data with different spatial and temporal resolution. Analysis of this dataset requests a special tool.

4 Tasks for Exploratory Data Analysis

As provided by G. & N. Andrienko (2006) we can also divide tasks, analysed in ecological research, into two groups: elementary tasks and synoptic tasks. Elementary task deals with individual elements of data. Synoptic task deals with datasets as a whole.

Prevalent elementary tasks in analysis of modelling output data are:

- To look for a plots' input/output data with minimal/maximal productivity of forest (amount of soil organic matter or other) on step X, finding outliers. Outliers can indicate model application limits and model errors.
- To compare value of forest parameter on step X with previous step value or with initial value. To compare values calculated in different scenarios with each other.
- To find high productivity forest plots with domination of species X in scenario Y on final step.

Synoptic tasks are:

- Pattern search: to group plots with similar values of parameter or similar dynamics of variables.

- Sensitivity analysis of input data: what input parameters are key parameters responsible for spatial distinction in output data.
- To find optimal scenario leading to maximization of forest productivity or increasing biodiversity or minimization of CO₂ emission (or other goal for optimization) for the whole territory.
- To find scenarios which lead to sustainable development of forest ecosystems.

Full list of typical tasks for modelling results analysis is not completed, we list here examples only.

5 Useful features for Exploratory Data Analysis system

What requirements an ideal Exploratory Data Analysis system must fulfil? The System has to allow working with large amount of complex data. It is necessary to have a program interface for links with model program (easy import data and export of analysis results). Of course the system must solve tasks described above. It will be nice if one system includes as much tools as possible for Exploratory Data Analysis (Andrienko, 2006): visualization, display manipulation, data manipulation, querying, computation (statistical analysis, data mining).

System must have friendly user interface and allow seeing data in various visual forms: graphs, plots, diagrams, maps, etc. On the one hand it had to have smart system which determines the best visual form for current dataset and task on its own; on the other hand user must have full control on visual displays which dynamically modified. It's comfortable to have possibilities of viewing maps together with graphs and diagrams. Maps must allow showing time-related data sets.

Data aggregation, receiving of new references and characteristics from existing data are very important tools of EDA system. Of course system must have flexible tool for search data with appointed attributes, create sub sets of data. It would be a good thing if the system provided easy creation and modification of query. So it's nice to have graphical interface for queries and possibilities to edit query manually using SQL language.

Powerful tool of descriptive statistics is very welcome. Sometimes it is necessary to describe big dataset in a few numbers. It is not very simple task because probability distribution of values in biological systems usually is not Gauss.

Partition data into meaningful classes, outlier analysis, trend detection, generalization are tasks require Data-mining tools affiliated with the system.

6 Examples of implementation EDA with modelling

EFIMOD model was applied together with Exploratory Data Analysis system – CommonGIS (Chertov et al, 2005) for investigation of carbon sequestration and biodiversity. CommonGIS (Andrienko and Andrienko, 1999, Andrienko et al., 2003) is a system designed to support visualization and analysis of spatially and temporally referenced data. It combines traditional GIS services with two features: tools to interactively manipulate dynamically created thematic maps; and tools for visual analysis of time-related data sets (geo-referenced time-series data tables). CommonGIS is able to handle process and visualize complex multidimensional tables describing time-series of spatially referenced data using maps and statistical charts.

Influences of forest management on forest ecosystems were investigated. Four scenarios of simulation runs were compiled for 200 years time span: 1) natural development without

any cuttings (Nat); 2) A selective forest scenario includes thinning and cutting of old large trees with 30% removing of growing stock every 30 years (SCU); 3) Clear cutting legal scenario is a scenario of four thinning and final clear cutting with burning cutting residues and successful forest regeneration (LRU); 4) Illegal clear cutting scenario represents one intensive thinning with taking all best trees and final clear cutting with burning cutting residues (ILL). The case study consisted of the part of State Forest Enterprise “Russky Les” and State Nature Reserve “Prioksko-Terrlasny” 100 kilometres south of Moscow.

Aggregation based time graph display representing (fig.2) was very useful for dynamics of Carbon stock and sustainability of forest ecosystem investigation. For producing this visualization, the value range of the attribute “Carbon stock” has been divided into 3 intervals by introducing breaks 100, 150 t C per ha. For each year, the display contains a segmented bar with the differently coloured segments showing how many values in this year over the whole territory fit in the corresponding intervals. The white segments correspond to the values below 100, the grey – to values between 100 and 150, black colour of segments presents values above 150. The aggregated display allows us to do important observations concerning the general dynamics of Carbon stock on the whole territory over the time. Natural development scenario leads to increasing number of forest compartments with amount carbon above 150 t per ha. Legal practice (LRU) results in oscillation, selective cutting scenario (SCU) demonstrates stability dynamics of system, illegal practice leads to decreasing carbon stock on the territory.

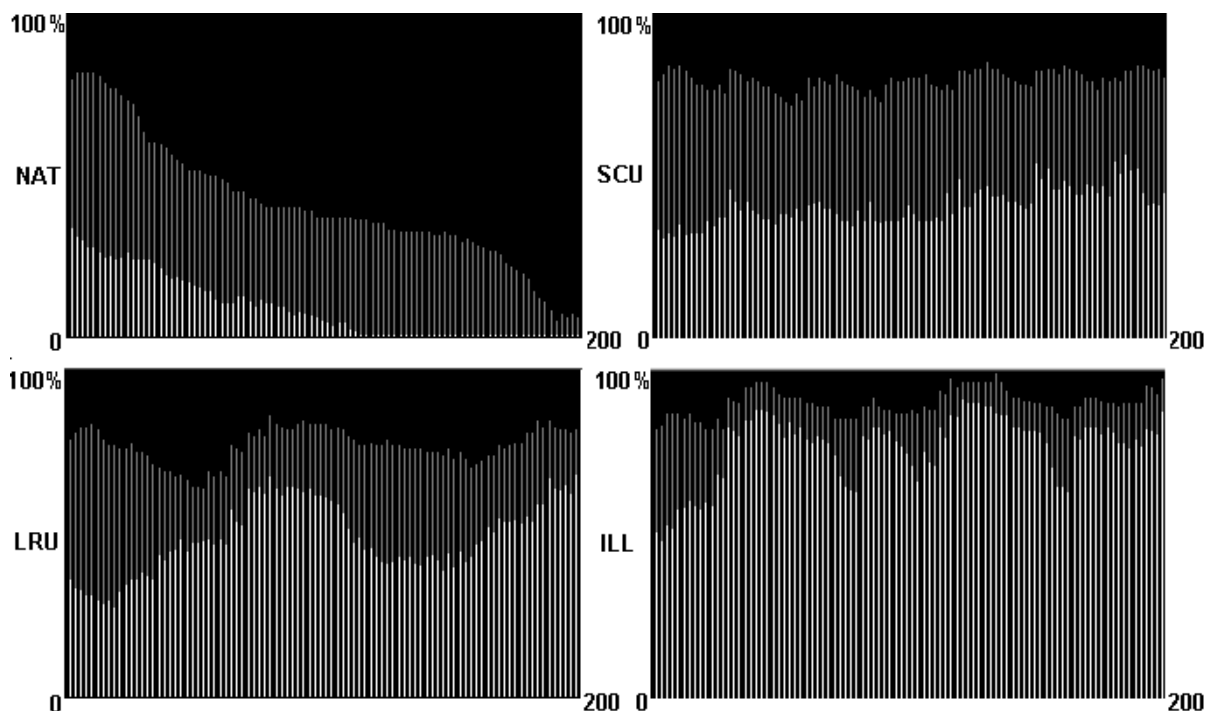


Fig.2 The value range of Carbon stock has been divided into intervals. The segmented bars show on each year the proportions of values over the whole territory fitting in these intervals. NAT, SCU, LRU, ILL – scenarios modelled; white colour - <100 t C per ha, grey – from 100 to 150 and black colour – above 150 t C per ha.

For ecosystem the natural development scenario (NAT) is the best from carbon sequestration point of view and the scenario with high intensive intermediate thinning (ILL) is the worst. On average for whole territory selective cutting scenario (SCU) is

better than clear cutting legal scenario (**LRU**). But such a conclusion is not so clear for the particular forest compartment. We define the median of carbon total amount (soil carbon, deadwood carbon, trees carbon) for 200 years on each forest compartment to answer the call. Then we classified all compartment on a criterion: what scenario leads to maximum of carbon sequestration. Thus we had to put away the natural development scenario (**NAT**) from consideration because its carbon sequestration is maximal anyhow. Selective cutting scenario (**SCU**) is better for 71% compartments, clear cutting legal scenario (**LRU**) is better for the rest. Illegal clear cutting scenario (**ILL**) is the better for none.



Fig.3 Dominant map – black colour indicates forest compartments where Selective cutting scenario (SCU) leads to maximal sequestration of carbon during modelling period; white colour – legal cutting scenario (LRU)

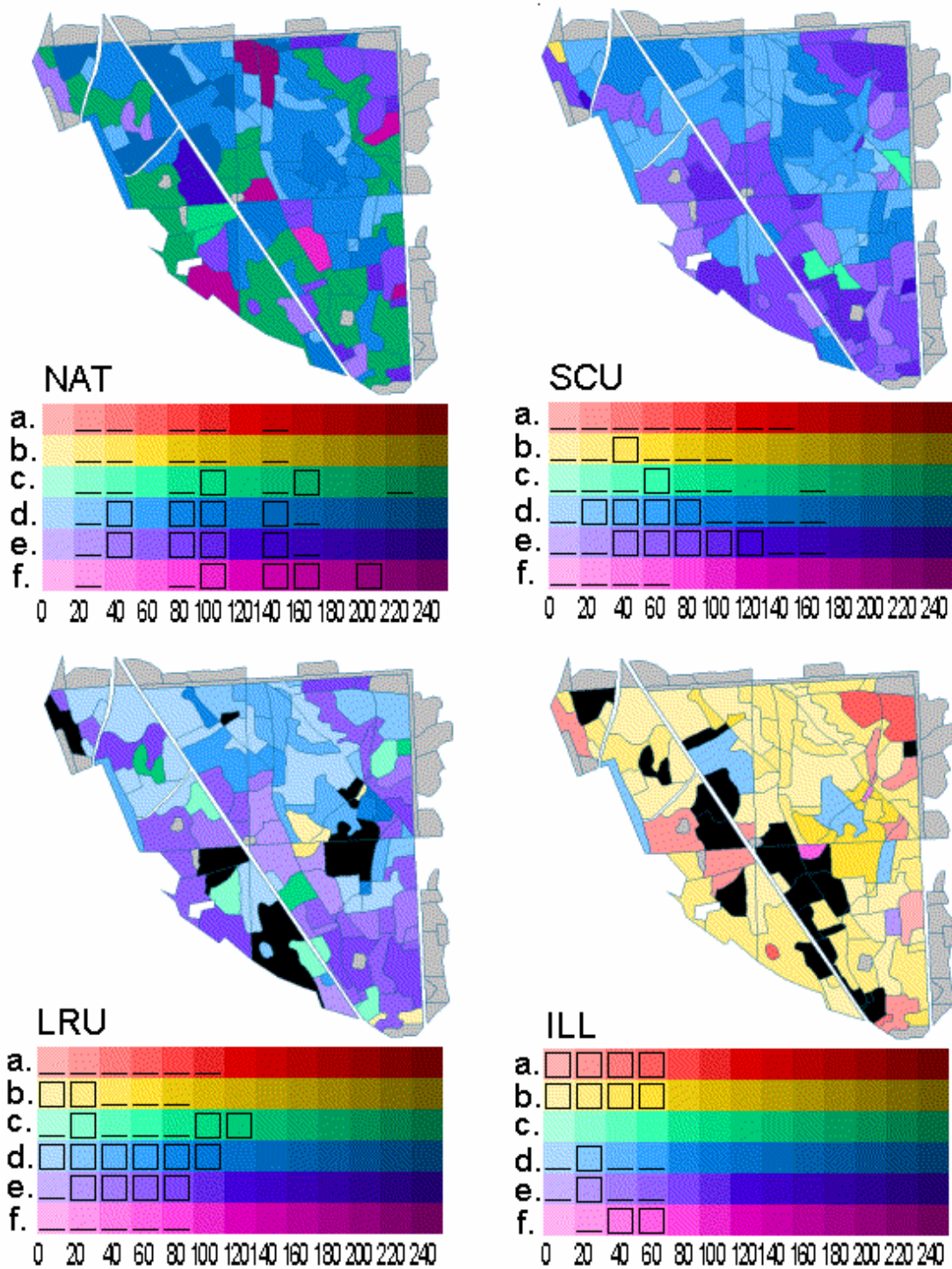


Fig.4 Species and age compositions (after 200 years): NAT – natural development scenario, SCU – selective cutting scenario, LRU – clear cutting legal scenario, ILL – Illegal clear cutting scenario. Tree species: a - aspen, b - birch, c - oak, d - pine, e - spruce, f - lime. Black colour presents area after clear cutting. 0-240 – age class of trees; square –territory has forest compartment where the tree cohort is dominated, line – presents

The vital issue for forest ecologists and managers is to analyze species composition and age groups composition of the forest. It is complete task to show this visual data at once.

There may be some trees species and some trees age groups in a certain compartment. Very convenient form for such information was created by authors of CommonGIS.

Dominant map technique (Andrienko and Andrienko, 2001) was used. The general idea of the method is to paint the map in colours corresponding to species and age groups that have maximum values of the analyzed attribute in each of the forest compartments (fig. 4).

On the map the tint indicates species; brightness indicates the belonging to age group. The black colour corresponds to the polygon with an absent of trees. The map legend has an additional data: presence/absence of trees cohorts (the even-age group of the same species) on the plot (a line) or dominance of trees cohorts (a square). To put it another way, the map legend is a diagram as well. Using such maps, one can gain some insight on forest structure. Polygons colouring enable to group the polygons of the similar structure by visual analyses.

Analysis of spatial patterns of stand composition and age groups shows that the clear-cut system results in the formation of a complex spatial mosaic of stands with different stages of post-cutting secondary succession.

We found that the strategy of natural development is the best alternative from the viewpoint of carbon sequestration. LRU is the best regime to satisfy timber production and, to some extent, forest biodiversity. Selective forestry (SCU) unifies the advantages of the Nat and LRU strategies, and may be the best strategy for the implementation of Sustainable Forest Management. The illegal practice leads to a fast decrease in productivity and biodiversity with domination of deciduous forests, and with no sequestration of soil carbon.

7 Conclusion

Exploratory analysis of simulation results allows us to learn more about the dynamics of forest growth. However, complex processes are usually described by complex multidimensional data sets. Respectively, it is rather an art than technology to find a useful sequence of analytical operations that can lead to interesting results. Only a tight integration of the modelling software with a visualisation system is not sufficient. Having huge amounts of original and derived data, one should not get lost in the complex analysis space. Our goal is to build an intelligent system that will incorporate heuristics into exploratory spatial data analysis, and automatically guide forest experts in the process of exploration and analysis of modelling results. This intelligent system will be a desired powerful decision support system.

Acknowledgments

It is a pleasure to thank Gennady Andrienko and Natalie Andrienko for effective cooperation. This work was supported by grants RFBR №05-04-49284, 04-04-48670 and grant of European Union 6th Framework Programme INCO- Russia + NIS-1 Contract № 013388 OMRISK.

References

- Andrienko, G., Andrienko, N., (1999). „Interactive maps for visual data exploration“. *Int. J. Geogr. Inform. Sci.* 13 (4), 355–374.
- Andrienko, G., Andrienko, N., 2001, Exploring Spatial Data with Dominant Attribute Map and Parallel Coordinates. *Computers, Environment and Urban Systems*, 25 (1), pp. 5-15.
- Andrienko, G., Andrienko, N., Voss, H., 2003, GIS for everyone: the CommonGIS project and beyond. In: Peterson, M. (Ed.), *Maps and the Internet*. Elsevier Science, pp. 131–146.
- Andrienko, N., Andrienko, A., 2006, *Exploratory Analysis of Spatial and Temporal Data. A systematic approach*. Berlin, Springer, 703p.
- Bykhovets, S.S., Komarov, A.S., 2002, A simple statistical model of soil climate with a monthly step. *Eurasian Soil Sci.*, 35 (4), pp. 392-400.
- Chertov, O., Komarov, A., Mikhailov, A., Andrienko, G., Andrienko, N., Gatalsky, P., 2005, Geovisualization of forest simulation modeling results: A case study of carbon sequestration and biodiversity. *Computers and Electronics in Agriculture*, 49, pp. 175–191.
- Chertov, O., Komarov, A., Nadporozhskaya, M., Bykhovets, S., Zudin, S., 2001, ROMUL – a model of soil organic matter dynamics as a substantial tool for forest ecosystem modelling. *Ecological Modelling*, 138, pp. 289-308.
- Chertov, O., Komarov, A., Andrienko, G., Andrienko, N., Gatalsky, P., 2002, Integrating forest simulation models and spatial-temporal interactive visualization for decision making at landscape level. *Ecol. Modell.*, 148 (1), pp. 47–65.
- Kangas, J., Store, R., Leskinen, P. and Mehtätalo L., 2000, Improving the quality of landscape ecological forest planning by utilising advanced decision-support tools. *Forest ecology and management*, 132, pp. 157-171.
- Komarov, A., Chertov, O., Zudin, S., Nadporozhskaja, M., Mikhailov, A., Bykhovets, S., Zudina, E., Zoubkova, E., 2003, EFIMOD 2 - the System of Simulation Models of Forest Growth and Elements Cycles in Forest Ecosystems. *Ecological Modelling*, 170(2-3), pp. 373-392.
- Komarov, A., Chertov, O., Andrienko, G., Andrienko, N., Mikhailov, A., Gatalsky, P., 2002, DESCARTES& EFIMOD: an integrated system for simulation modelling and exploration data analysis for decision support in sustainable forestry. In: Rizzoli, A., Jakeman, A. (Eds.), *Integrated Assessment and Decision Support, Proceedings IEMSS 2002*, vol. 1, Lugano, Switzerland, June 24–27, 2002, pp. 239–244.
- Næsset, E., 1997, Geographical information systems in long-term forest management and planning with special reference to preservation of biological diversity: a review. *Forest ecology and management*, 93, pp. 121-136
- Pykäläinen, 2000; Interactive use of multi-criteria decision analysis in forest planning. PhD thesis, University of Joensuu.
- Varma, V.K, Ferguson, I. and Wild, I., 2000, Decision support system for the sustainable forest management. *Forest ecology and management*, 128, pp. 49-55.