Use of Rough Sets Analysis to Classify Siberian Forest Ecosystems According to Net Primary Production of Phytomass

Matti Flinkman[1], Wojtek Michalowski[2], Sten Nilsson[3], Roman Slowinski[4], Robert Susmaga[4], and Szymon Wilk[4]

Abstract

This paper attempts to identify attributes that are considered essential for a development of sustainable forest management practices in the Siberian forests. This goal is accomplished through an analysis of net primary production of phytomass (NPP), which is used to classify the Siberian ecoregions into compact and cohesive NPP performance classes. Rough Sets (RS) analysis is used as a data mining methodology for the evaluation of the Siberian forest database. In order to interpret relationships between various forest characteristics, relationships known as *interesting rules* are generated on a basis of a reduced problem description.

Keywords: Rough sets, knowledge discovery, sustainable development, forest ecosystem functioning

[1] Department of Forest Industry Market Studies, Swedish University of Agricultural Sciences, Uppsala, Sweden.

[2] School of Business, Carleton University, Ottawa, Canada.

[3] Sustainable Boreal Forest Resources Project, International Institute for Applied Systems Analysis, Laxenburg, Austria.

[4] Institute of Computing Science, Poznan University of Technology, Poznan, Poland.

1. Problem Statement

Two key issues in the development of sustainable forest management practices in the boreal forest zone are:

a) Identification and evaluation of the current and desired state of forest ecosystems that are essential for the proper functioning of the ecosystem; and

b) The study of the impact of alternative forest management regimes on the functioning of the ecosystem.

This paper deals primarily with the analysis of the current ecosystem functions of the Siberian forests by using a comprehensive database maintained at the International Institute for Applied Systems Analysis (IIASA). Knowledge established through such an analysis might, in turn, form a basis for further work on the development of sustainable management practices.

The analysis presented here draws on the framework established by the Statement of Principles on Sustainable Forest Management at the 1992 UN Conference on Environment and Development (UNCED, 1992). This framework stipulates that analysis should be conducted on a number of theme areas. Specific criteria and indicators have been suggested for each of these theme areas. The following theme areas have been proposed:

- Global carbon cycles
- Health and vitality
- Wood and non-wood productive functions
- Biological diversity
- Functions that protect soil and water
- Socio-economic functions and conditions

Despite the comprehensive nature of this framework, its principal shortcoming is that the theme areas and performance indicators associated with them are treated in isolation rather than in a comprehensive manner (Nilsson, 1997a). In order to address this shortcoming, this paper emphasizes that the theme areas and the interconnections between them should be considered simultaneously in the analysis of various ecosystem functions. Therefore, *ecosystem functioning* should be used as a core concept, implying that the appropriate and desirable functioning of all theme areas is necessary to support ecosystem services.[1] Use of ecosystem functioning as a core concept also improves the overall understanding of the consequences of natural or anthropogenic changes within a specific theme area.

[1] Delivery of ecosystem services involves: (1) Capture of solar energy and conversion into biomass that is used for food, building materials and fuels; (2) Breakdown of organic wastes and storage of heavy metals; (3) Maintenance of gas balance in the atmosphere that supports human life: absorption and storage of carbon dioxide and release of oxygen for breathable air; (4) Regeneration of nutrients in form essential to plant growth, e.g. nitrogen fixation and movement of those nutrients.

In this study, the general framework for the identification of forest ecosystem attributes is therefore based on the premise that the possible impact of descriptive attributes identified within different theme areas should be examined by using the core concept of ecosystem functioning. The explanatory attributes, chosen from abiotic, biotic, and human induced factors, should thus describe the interactions between land-uses, vegetation types, forest density, site-class, age, and different aspects of human activities. The identification of the attributes that contribute the most to the explanation of the ecosystem functioning is a first step towards developing sustainable forest management practices. It is this step that is described in this paper in greater detail.

The data component of this study is described in Section 2. The data set contains information on a number of attributes recorded at ecoregion level. The decision as to which attributes should be selected is complex because there are a significant number of possible attributes. In keeping with the idea of a comprehensive approach, it is necessary to consider cross-classifications reflecting the different roles of attributes in describing different conditions. This task will be accomplished through the combination of a Rough Set (RS) analysis with a heuristic evaluation of the possible sets of attributes providing similar descriptive accuracy. The principles for RS analysis are introduced in Section 3 and in discussed in greater detail in Appendix 2. Use of the RS methodology on a data set derived from the Siberian forest database is also described in Section 3. Section 4 presents a discussion of the results.

2. Siberian Forest Database

The Siberian forest database contains information relevant to the cornerstone areas of the Sustainable Boreal Forest Resources Project at IIASA (Nilsson, 1997b). Nearly 5000 attributes describing abiotic, biotic, and human induced conditions are included in the database. The spatial coverage of the collected information is aggregated at different levels. The highest level covers the whole of Siberia. There are sub-levels for 65 administrative regions, 65 ecological regions (ecoregions), 360 landscapes, and 2500 forestry enterprises. All database items can be related to some spatial aggregation level that allows spatial descriptions of abiotic, biotic, and anthropogenic conditions.

For purposes of this study, data aggregated at the ecoregion level was extracted from the Siberian forest database (see Appendix 1). This data set contains a sample of the original abiotic and biotic attributes and attributes for human induced conditions. In addition, a number of modified attributes known as CODE-descriptors and SHDI-descriptors, describing the structure of certain distributions have been developed for each ecoregion. In creating the CODE-descriptor, the original distribution data (for example, the age distribution of a forested area) has been categorized into few (4-7) share classes. This allows the creation of a number of distribution "profiles". The SHDI-descriptors were created based on Shannon diversity index formula (Shannon and Weaver, 1962). The SHDI-descriptor represents the degree of diversity of the attribute under consideration. For example, an attribute with only a few dominating classes results in a low diversity value for the SHDI-descriptor, while an evenly distributed share results in a high value.

3. Net Phytomass Production

The study is based on the hypothesis that the classification of Siberian ecoregions into different classes based on the net primary production of phytomass (NPP) will reflect different types of land-use and biogeophysical conditions (Shvidenko, *et al*. 1997). The net primary production of phytomass is an estimated measure of an ecoregion's total production potential of phytomass in t/ha/year, calculated according to Bazilevich (1993). The NPP measure includes all land uses, including agricultural land, within an ecoregion. Therefore, such a classification will capture a number of the factors assumed to be associated with the level of ecosystem functioning. It is not a straightforward exercise to create a cohesive description of each ecoregion in terms of its ecosystem functioning, because there are a significant number of attributes that might be considered as the candidates for such a description. Therefore, RS analysis (Pawlak, 1991; Slowinski, 1992) was used to develop such a description. Methodological considerations associated with this issue are presented in Section 3.1, whereas the application of RS analysis to the NPP classification problem is described in Section 3.2.

In order to create a compact and cohesive description of the NPP classification problem, we proceed with the identification of a smaller subset of the attributes that need to be evaluated. Following RS principles, we focus on the identification of subsets of the attributes with desired characteristics. In Appendix 2, Section 2 we describe a heuristic procedure which we use to generate a "good" subset (known as a *good* reduct) for a given classification. Identification of a good reduct results in a significant reduction of the number of attributes to be considered in describing the NPP classification problem.

An important aspect of any policy analysis is the explanation of the relationships between problem components. One of the best methods for conveying such information is provided by decision rules that are logical statements of the type *if… then…* We use them in our study to generate interesting rules for the good reduct.[1] The interesting rules provide a helpful explanation of the role of attributes and the significance of their specific values, and allow us to draw conclusions in terms of knowledge statements.

3.1. Methodological Considerations

The methodology used to analyze the relationships among the attributes describing the ecosystem functioning of the Siberian forests is based on RS theory. This theory was first proposed by Pawlak (1991) to study classification problems in a computer science. In order to obtain the most useful results from a basic RS analysis, it is considered best to use symbolic (qualitative) data rather than continuous-valued (quantitative) information. If quantitative information is used, the domains of continuous-valued attributes should be discretized (categorized) prior to the analysis (see Appendix 2 Section 4). The data set under consideration consists of *objects* (also known as *examples* or *cases*) representing Siberian ecoregions. The characteristics of these ecoregions are described by discrete values of the attributes. The set of attributes is usually divided into two disjoint subsets, called *condition* and *decision* attributes. Condition attributes express some descriptive information about the ecoregions, whereas the decision

[1] The general principles for generating rules are described in Appendix 2 Section 3.

attributes describe a classification assigned to the ecoregion. The set of ecoregions described by attributes and represented in a table format is called *a decision table*. This table is then further analyzed to reduce the number of condition attributes while maintaining a good approximation of the original set[1].

It is important to stress that the NPP classification problem considered in this paper could also be analyzed using statistical methods based on discriminant analysis. The main goal of discriminant analysis is to create functions that can later be used to assign a given ecoregion to a predefined class depending on the scores of these functions associated with the classes. Discriminant analysis can also be used to reduce the number of attributes and to select the most important ones. Most discriminant analysis methods are applicable only for continuous-valued attributes. Only a few methods (for example, the "location model approach") can deal with the mixture of continuous-valued and symbolic attributes.

RS theory has several advantages over discriminant analysis (Stefanowski, 1992) when considering the properties of the Siberian data set and the comprehensibility of the generated output. These are:

a) Discriminant analysis methods are very demanding in regard to the quality of the input data. A normal distribution of continuous attributes is assumed, and the considered classes should contain comparable number of objects. Neither of these requirements is satisfied for the Siberian forest data. In contrast, RS theory does not impose these requirements.

b) In the location model approach, all qualitative attributes have to be transformed into binary attributes. When the qualitative attributes can take on many values (as in the case of the CODE-descriptor attributes), the resulting number of attributes increases rapidly. In practical applications, it is suggested that there should be no more than six binary attributes (Krzanowski, 1983). However, there are 22 qualitative attributes in the analyzed data set. RS methodology, in contrast, does not require any transformation of qualitative attributes and does not limit their number.

c) The methods of discriminant analysis generate a final result in a form of discriminant functions, which aggregate the input information in a non-transparent way. Methods based on the RS theory produce decision rules that are much more transparent than the discriminant functions and can be easily interpreted by a prospective user.

3.2. The NPP Classification Problem

The set of condition attributes used in the classification problem consists of[2]: MOUNTAIN, PERMAFROST, AV_AIR_TEM, AV_SOIL_TE, AV_MAX_SOI, AV_MIN_SOI, TOT_PRECIP, WIND, SUM_T10, SUM_T5, SUM_PREC10, SUM_PREC5, DURATION_1, DURATION_5, SNOW_COVER, Vext-SHDI, FA/Area, FF-CODE, FF-SHDI, BON-CODE, BON-SHDI, DENS-CODE, DENS-SHDI, AgAr-CODE, AgAr-SHDI, AgVo-CODE, AgVo-SHDI,

[1] Basic notions of the RS theory are described in greater detail in Appendix 2.

[2] See Appendix 1 for a full list of the attribute and for an explanation of their abbreviations.

POP/sqkm, Autow/sqkm, Railw/sqkm and Riverw/sqkm. RS analysis, together with a heuristic procedure to generate a good reduct (see Appendix 2 Section 2) was applied to the data set containing the above condition attributes. The results are described in the following section.

3.2.1 The Reduct

Each ecoregion was assigned into one of three NPP classes (L, M and H, denoting the *low*, *medium* and *high* class of the NPP, respectively), according to ecoregion's potential phytomass production capacity calculated according to Bazilevich (1993). Through the RS analysis, the original set of 31 attributes was then reduced and the following good reduct was identified: *relief conditions* (MOUNTAIN); *snow cover conditions* (SNOW_COVER); *share of forested area of total ecoregion area* (FA/Area); *forest fund profile consisting of forest land, non-forest land and lease* (FF-CODE); *age profile of growing stock consisting of 5 age class categories* (AgVo-CODE); and *density of railway network* (Railw/sqkm)

The *age profile of growing stock, share of forested area of total ecoregion area* and *forest fund profile consisting of forest land, non-forest land and lease* are all forest-related attributes. The *relief conditions* and *snow cover conditions* describe biogeophysical conditions, and the *density of railway network* can be considered as an indicator of ecoregion development. The reduction of the original set of 31 attributes to the 6 most relevant attributes constitutes a significant improvement over other studies due to the formal reduction of the original dataset without loss of information.

3.2.2 Generation of Interesting Rules

General knowledge statements were built using the interesting rules (shown in Table 1) generated for the good reduct[1]. Each row in Table 1 represents one decision rule. The conditional part of the rule is a conjunction of elementary conditions on those attributes for which values are specified (the elementary condition has the syntax *attribute* = *value*) and the decision part reflects assignment of an ecoregion to the specified NPP class. For example, rule 7 should be read as:

if AgVo-CODE equals to ABDBC and MOUNTAIN equals to 1, **then** NPP class is M.

An interpretation of this rule is as follows: **if** the distribution of growing stock into age classes is such that 0-5% of the growing stock is in the age class "youngest forest", 5-20% is "young forest", 40-60% is "middle aged forest", 5-20% is "immature forest", 20-40% is "mature and overmature forest", and relief conditions are mountainous, **then** the NPP class is medium.

The forest fund[2] profile (FF-CODE) appears to be the most frequent attribute present in conditional part of the interesting rules. This is particularly true for the high NPP class,

[1] see Appendix 2, Section 3

[2] The "Forest Fund" consists of all forests and all land allocated for forest purposes

where it appears in the conditional part of all decision rules. For the two other NPP classes, it appears in combination with a number of other attributes in most of the rules.

Table 1. Interesting rules for the NPP classification problem[1]

Rule no.	NPP class	Elementary conditions						Relative rule strength
		AgVo-CODE	FA/Area	FF-CODE	MOUNTAIN	Railw/ sqkm	SNOW COVER	
1	L	AABAF						12%
4	L		0				LONG	56%
5	L			ECA	1		LONG	12%
6	L			ECA			LONG	12%
7	M	ABDBC			1			10%
8	M	AABBE			2			16%
9	M	ABDBC	1					10%
10	M		1	ECA		1		13%
11	M		1	GAA		0		13%
12	M	AACBD	1			1		10%
13	M		1	FBA	2		SHORT	10%
14	M		1		2	0	SHORT	16%
15	M			FBA	2	0	SHORT	10%
16	H		0	FBA	2			19%
17	H			FBA	2	1		19%
18	H		0	FBA		1		25%

The column "relative rule strength" gives the percentage of all the ecoregions "covered" by a given rule (i.e., those that are classified into appropriate class by this rule). While generating the interesting rules, we use a threshold of 10%. That is, only those rules that "cover" at least 10% of the cases (ecoregions) are considered interesting rules.

[1] Values for the AgVo-CODE attribute represent different distributions of growing stock into age classes (youngest forest, young forest, middle aged forest, immature forest, and mature and overmature forest). Values of the FF-CODE attribute represent different distributions of land within forest fund into land use classes (forest land, non-forested lands, and 'long-term lease lands. The letter gives the share percentage ranging from <5% (A) to >95% (G)). Values of the MOUNTAIN attribute reflect different relief conditions, with 1 denoting mountain relief and 2 denoting plain relief condition. Values of the SNOW COVER attribute reflect different duration of a snow cover, with value LONG denoting long winter and SHORT denoting short winter. Values 0 and 1 for the attributes FA/Area and Railw/sqkm indicate either first or second interval generated by *Recursive Minimal Entropy Partitioning* discretization method (Fayyad and Irani, 1993) applied for these two attributes. All other attributes were discretized according to the value intervals provided by an expert.

3.2.3 Extracting Knowledge from the Rules

The analysis of the NPP classification problem suggests that ecoregions classified into the high (H) NPP class are characterized by a low amount of forested areas (FA/Area = 0). The existing forest fund (FF-CODE = FBA) within these ecoregions consists of mainly forest land and, to a lesser extent, non-forest land. These ecoregions seem to be well developed (Railw/sqkm = 1, MOUNTAIN = 2). Climate conditions appear to be relatively favorable for a high net primary production of phytomass, due to their southern location (mainly in West and Southwest Siberia).

Ecoregions classified into the low (L) NPP class are characterized by mountainous and harsh climatic conditions, and are therefore relatively inaccessible. Much of the land in the forest fund in these regions is non-forest land (FF-CODE = ECA). Therefore, the production of phytomass is based, to a large extent, on growing potential outside of the forests. This is also confirmed by the low amount of forested areas in these ecoregions (FA/Area = 0). In cases where the conditions described above do not apply, the low net primary production of phytomass is due to the uneven distribution of the growing stock into different age classes (AgVo-CODE = AABAF). The amount of forested lands in the mature and overmature age class is clearly dominant over other age classes; this implies that the forested area of such an ecoregion is approaching the "climax" stage of its development cycle.

The ecoregions classified into medium (M) NPP class represent "forested" regions because the forest cover of the total ecoregion area is clearly predominant (FA/Area = 1). This is also supported by the fact that the forest fund consists, to a large extent, of forest land (FF-CODE = ECA or GAA). In addition, the age class distribution of growing stock within forested area (AgVo-CODE = ABDBC or AABBE) represents a low amount of mature and overmature forest, which indicates a certain degree of utilization of the forest resources or possible "natural management" through disturbances like fires and insect attacks which have brought down the volume of the "old growth". Such a distribution of growing stock results, therefore, in a higher net primary production of phytomass.

In conclusion, ecosystem services in the ecoregions belonging to low and high NPP class are, to a large extent, not delivered by forests. Ecosystem functions in these classes are delivered by non-forested areas. At the same time, the ecoregions classified into medium NPP classes are predominantly characterized by forest areas. Forests in these regions therefore appear to play a crucial role in supplying ecosystem services.

From the point of view of forestry and forest management practices, the interest should be focused on ecoregions belonging to medium class NPP. The findings of our study confirm the importance of forests for ecosystem functioning in medium class NPP ecoregions. This, in turn, implies considerable potential for implementation of desirable forest management policies.

4. Discussion

We evaluated the classification of the Siberian forests from the point of view of net primary production of phytomass. This required the incorporation of several descriptive

aspects considered as essential for evaluating ecosystem functioning. Analysis of complex situations, characterized by many decision attributes of different character and different levels of detail, calls for a methodology that allows simplification of the descriptive requirements of the problem. In the case of the Siberian forest database, the RS methodology, enhanced with the procedure for identification of good reduct, enabled the development of a reduced (i.e., having fewer attributes) and compact description for the classification problem. The creation of such a compact description has advantages from a data mining perspective, as it requires less information to be collected and accessed, and it also facilitates analysis of data dependencies. Generation of the interesting rules demonstrates that it is possible to identify certain common features for ecoregions belonging to the same class. We attempted to translate these commonalties into general knowledge statements. A promising aspect that emerged in the creation of these statements is that the regularities discovered in the Siberian forests are in line with forces shaping ecosystem functions in other boreal regions, outside Siberia.

There are some limitations to the data used in the analysis. As pointed out earlier, the net primary production of phytomass is an estimated measure of an ecoregion's total production potential of phytomass (Bazilevich, 1993). Therefore, this measure does not give the actual phytomass and is not measured in situ. However, this was the only information available at the time of the study, and this data has been used in many international studies (e.g. Kolchugina and Vinson; 1993; Dixon *et al*., 1994; and Krankina *et al*., 1996). On the other hand, it can be pointed out that aggregated Russian forest inventory and forest ecological data have been evaluated to be of the same quality as the inventories and data for other countries in the boreal zone (Raile, 1994).

One of the principal issues related to studying forest ecosystems is the importance of evaluating several aspects of the problem, as exemplified by the appropriate theme areas. In order to address this issue, one needs to consider a set of diversified attributes. In the study we accomplished this and, moreover, identified the relations between specific attributes related to different theme areas. Our findings should assist future forest studies in focusing on those aspects of theme areas that are deemed to be important, and thus create a basis for sustainable forest management policies.

In future research, some recent extensions of the RS methodology can be used for a more detailed study of Siberian forest database. In particular, extensions of RS methodology that concern attributes with preference ordered domains, and the approximation of classes by what are known as *dominance relations* instead of the classical indiscernibility relation (see Greco *et al.*, 1999), should prove useful for this kind of analysis.

Acknowledgements

The authors would like to thank Michael Gluck for comments on the first draft of the paper and Keith Compton for his editorial assistance.

R. Slowinski, R. Susmaga and S. Wilk wish to acknowledge financial support from the Polish Committee for Scientific Research (KBN). The Decision Analysis and Support Project at IIASA and Natural Sciences and Engineering Research Council of Canada

(NSERC) supported research conducted by W. Michalowski while he was a senior research scholar at that institute.

References

Bazilevich, N.I. (1993). *Biological Productivity of Ecosystems of Northern Eurasia* (in Russian), Nauka, Moscow.

Dixon, R.K., Brown, S., Houghton, R.A., Solomon A.M., Trexler, M.C. and Wisniewski, J. (1994). Carbon Pools and Flux of Global Forest Ecosystems. *Science* Vol. 263, pp. 185-190.

Dougherty, J., Kohavi, R. and Sahami, M. (1995). Supervised and Unsupervised Discretizations of Continuous Features. In: *Proceedings of the 12th International Conference on Machine Learning*, Morgan Kaufmann Publishers, New York, pp. 194-202.

Fayyad, U.M. and Irani, K.B. (1993). Multi-interval Discretization of Continuous-valued Attributes for Classification Learning. In: *Proceedings of the 13th International Conference on Artificial Intelligence*, Morgan Kaufmann Publishers, New York, pp. 1022-1027.

Greco, S., Matarazzo, B. and Slowinski, R. (1999). The Use of Rough Sets and Fuzzy Sets in MCDM. In: Gal, T., Hanne, T. and Stewart, T. (eds.), *Advances in Multiple-Criteria Decision Making*, Kluwer Academic Publishers, Boston, pp. 14.1-14.59.

Grzymala-Busse, J.W. (1992). LERS: A System for Learning from Examples Based on Rough Sets. In: Slowinski, R. (ed.), *Intelligent Decision Support. Handbook of Applications and Advances of the Rough Set Theory*. Kluwer Academic Publishers, Dordrecht, pp. 3-18.

Kolchugina, T.P. and Vinson, T.S. (1993). Comparison of Two Methods to Assess the Carbon Budget of Forest Biomes in the Former Soviet Union. *Water Air and Soil Pollution* Vol. 70, pp. 207-221.

Krankina, O.N., Harmon, M.E. and Winjum, J.K. (1996). Carbon Storage and Sequestration in the Russian Forest Sector. *Ambio* Vol. 25, pp. 396-404.

Krzanowski, W.J. (1983). Stepwise Location Model Choice in Mixed-variable Discrimination. *Applied Statistics* Vol. 32, pp. 260-266.

Mienko, R., Stefanowski, J., Tuomi, K. and Vanderpooten, D. (1996). Discovery-oriented Induction of Decision Rules. *Cahier du LAMSADE*, No. 141.

Nilsson, S. (1997a). Challenges for the Boreal Forest Zone and IBFRA. A Keynote address presented at the 7th International Boreal Forest Research Association Conference, Duluth, Minnesota, USA. pp. 16-20.

Nilsson, S. (ed.) (1997b). Dialogue on Sustainable Development of the Russian Forest Sector - Volume II. IIASA Interim Report (IR-97-010), International Institute for Applied Systems Analysis, Laxenburg.

Pawlak Z. (1991). *Rough Sets. Theoretical Aspects of Reasoning About Data*. Kluwer Academic Publishers, Dordrecht.

Predki B., Slowinski R., Stefanowski J., Susmaga, R. and Wilk, Sz. (1998). ROSE: Software Implementation of the Rough Set Theory. In: Polkowski, L. and Skowron, A. (eds.), *Lecture Notes in Artificial Intelligence 1414. Rough Sets and Current Trends in Computing. (Proceedings of RSCTCí98)*, Springer Verlag, Berlin-Heidelberg, pp. 605-608.

Raile, G. (1994). Evaluation of Russian Forest Inventory Data. International Institute for Applied Systems Analysis, Laxenburg. Unpublished manuscript.

Shannon, C.E and Weaver, W. (1962). *The Mathematical Theory of Communication*. University of Illinois Press, Urbana.

Shvidenko, A., Nilsson, S. and Roshov, V. (1997). Possibilities for Increased Carbon Sequestration through the Implementation of Rational Forest Management in Russia. *Water, Air and Soil Pollution* No. 94, pp. 137-162.

Slowinski, R. (ed.) (1992). *Intelligent Decision Support. Handbook of Applications and Advances of the Rough Set Theory*. Kluwer Academic Publishers, Dordrecht.

Slowinski, R. and Stefanowski, J. (1994). Rough Set Reasoning about Uncertain Data. *Fundamenta Informaticae*, Vol. 27, No 2/3, pp. 229- 244.

Stefanowski, J. (1992) Rough Sets Theory and Discriminant Methods as Tools for Analysis of Information Systems. A Comparative Study. *Foundations of Computing and Decision Sciences* No. 2, pp. 81-98.

UNCED, (1992). *The Global Partnership for Environment and Development*. United Nations, Geneva.

Appendix 1

List of attributes used in the study

Attribute Name	Description
PhyProClass	Net Primary Production classes of Phytomass
MOUNTAIN	Relief conditions: mountain, plain or far east mountain
PERMAFROST	Permafrost: year round, seasonally or no frozen ground
AV_AIR_TEM	Average air temperature
AV_SOIL_TE	Average soil surface temperature
AV_MAX_SOI	Average max soil surface temperature
AV_MIN_SOI	Average min soil surface temperature
TOT_PRECIP	Average total precipitation
WIND	Average wind speed
SUM_T10	Total number of days during the growing season with average temperature above 10°C
SUM_T5	Total number of days during the growing season with average temperature above 5°C
SUM_PREC10	Total precipitation during the growing season for days with average temperature above 10°C
SUM_PREC5	Total precipitation during the growing season for days with average temperature above 5°C
DURATION_1	Duration of vegetation period where average temperature is above 10°C
DURATION_5	Duration of vegetation period where average temperature is above 5°C
SNOW_COVER	Duration of snow cover
Vext-SHDI	Shannon diversity index for vegetation types
FA/Area	Forested area of total ecoregion area in %
FF-CODE	Forest fund profile distributed by forest land, non-forest land, and lease
FF-SHDI	Shannon diversity index for forest fund profile
BON-CODE	Site class profile for all age classes
BON-SHDI	Shannon diversity index for site class profile of all age classes
DENS-CODE	Density class profile for all age classes
DENS-SHDI	Shannon diversity index for density class profile of all age classes
AgAr-CODE	Age class profile of total forested area
AgAr-SHDI	Shannon diversity index for age class profile of total forested area
AgVo-CODE	Age class profile of growing stock within total forested area

Attribute Name	Description
AgVo-SHDI	Shannon diversity index for age class profile of growing stock within total forested area
POP/sqkm	Population density per square kilometer
Autow/sqkm	Road density per square kilometer
Railw/sqkm	Railways density per square kilometer
Riverw/sqkm	Waterway density per square kilometer

Appendix 2

2.1. Basic notions of Rough Sets theory

The fundamental notion of RS theory when it is applied to analyze a given classification is that of an *indiscernibility relation* among the objects. An indiscernibility relation exists when objects are indiscernible from one another when only a given set of attributes is taken into account. The relation is based on the assumption that the values of attributes are the sole source of knowledge about the objects. Because indiscernibility is an equivalence relation, it defines a *partition* (also called *classification*) of objects into disjoint subsets called *elementary sets*. The main function of the RS theory is to examine different partitions of objects induced by different sets of condition attributes and decision attributes, and the relationship between these partitions.

Two particular partitions of the objects are most frequently studied. One of them is the partition induced by the set of all the decision attributes. The elementary sets of this partition are called *classes* - sets of objects that are described by the same value of a decision attribute. The other partition of interest is induced by the set of all condition attributes. The elementary sets of this partition, called *atoms*, contain objects that are indiscernible from one another with regard to all condition attributes. The name "atom" is used to stress that it represents the smallest indivisible granule of knowledge that can be used to approximate (build) another knowledge, namely the partition of objects into classes. The definitions of atoms and classes are followed by the next step of the RS analysis, in which the different partitions are matched and analyzed.

Any subset of objects (called a *concept*) is *definable* by a set of attributes if this concept can be represented as a union of the elementary sets generated by these attributes. If this is not possible, the RS theory introduces the notion of a concept approximation, which consists of the *lower approximation* and the *upper approximation*. The lower approximation of a concept is the union of all elementary sets that are included in this concept, while the upper approximation is the union of all elementary sets that have non-empty intersection with the concept. Thus, the lower approximation is always a subset of the concept, while the upper approximation is a superset of the concept. Concepts for which the lower and the upper approximations are equal are called *crisp* sets; otherwise they are referred to as *rough* sets. Every rough set is characterized by a non-empty *boundary region*, which is defined as the difference between its upper and lower approximation.

If each class is definable by the set of all condition attributes then the values of the condition attributes provide sufficient information to distinguish between objects belonging to different classes. Otherwise the non-definable classes are represented in form of approximations. The situation is referred to as *data inconsistency*.

To control the level of inconsistency in the data, the RS theory introduces a special measure called the *quality of approximation*, which is defined as the ratio of all objects belonging to lower approximations of all classes to all objects in the decision table. The maximum value of this measure, equal to *1.0*, indicates that all the classes may be fully

distinguished from one another using the information supplied by the condition attributes.

It may be interesting to explore whether there are some proper subsets of condition attributes, which are sufficient to generate the same quality of approximation as the whole set. This leads directly to the idea of attribute reduction.

2.2. Reducts and Their Computation

A *reduct* is defined as a subset of attributes that includes a minimal number of attributes and that ensures the same quality of approximation as the whole set of attributes. In general, it is possible that there is more than one reduct for a given decision table. In that case, the set called the *core* of attributes is defined as the intersection of all reducts. In other words, the core consists of those common attributes that belong to all reducts. As far as data consistency is concerned, the core is the set of the most relevant and indispensable attributes in the table – removal of any of the core attributes from the decision table leads to an increase in data inconsistency, manifested by a drop of the quality of approximation.

Generating the core is easy because it does not involve finding all the reducts and producing their intersection. A convenient method is to remove, one by one, each of the attributes and to check the quality of approximation: if the quality drops then the given attribute should be included in the core.

The process of generating reducts, on the other hand, is computationally complex (NP-complete). As a result, apart from exact algorithms designed for generating all reducts from a decision table, there exist approximate algorithms, designed for generating a single reduct or a population of reducts, with the aim of decreasing computing time. The main disadvantage of the approximate algorithms is that it is not possible to state that reducts generated in such a way are indeed minimal.

In many practical situations the difficulties associated with core and the reducts are:

- The number of reducts is usually very large; often, too large to be effectively analyzed; and
- The core is often empty.

This indicates that the regularities in the data are not clear enough to be captured in a form of a core of the attributes or reducts. It does not mean, however, that such regularities do not exist.

In an attempt to express those regularities, the notion of a β-core is introduced, which is a natural generalization of the classical RS core. Assuming that β is a real number from the [0,1] interval, the β-core is the set of all attributes whose relative frequency of occurrence in all reducts is not lower than β. This definition ensures that the β-core is equivalent to the core in the classical RS sense. Using the notion of β-core and β-reducts, it is possible to generate a *good* reduct. The heuristic procedure to generate a good reduct is given below.

1. Generate all existing reducts.
2. For every conditional attribute, calculate its relative frequency of occurrence in the

reducts.

3. Establish the threshold $\beta \in [0, 1]$ (this may be done, for example, after analyzing the histogram of relative frequencies), and establish the β-core, i.e. select those attributes whose frequencies (calculated in 2) are not lower than β.
4. Find all β-reducts, i.e. reducts that include the β-core. If such a reduct does not exist, modify the β-core by dropping the attribute with smallest relative frequency and repeat this step.
5. Using the β-reducts identified in 4, find those which have the smallest cardinality (i.e., the smallest number of attributes). If there is only one such reduct, then it is the *good* one. Otherwise go to the next step.
6. For each reduct identified in 5 test its ability to construct an accurate classifier representing the data set in terms of decision rules[1]. Identify the reduct with the best result of that test. This is the *good* reduct.

It is important to stress that unlike the classical RS core, the β-core must be generated by computing all reducts and calculating the attributes' frequencies. This may be quite difficult, especially when the number of reducts is very large. The β-core may be useful, however, in exposing interesting regularities in the decision table. Additionally, the β-core may prove helpful in handling the large number of reducts: the reducts that do not include the β-core are discarded and only a small set of reducts remains to be analyzed. Following the β terminology, these remaining reducts may be referred to as β-reducts.

2.3. Decision Rules

A decision rule is a logical statement defined as "***if*** *some conditions are met,* ***then*** *some decisions are recommended*", where the conditional element is a conjunction of elementary conditions (i.e. elementary tests on attribute values), and the decision element is a disjunction of recommended decisions (i.e. assignments to classes). A rule is said to cover an object if all conditions in the condition part are matched by the attribute values of an object.

Decision rules are generated by induction. During this process, two sets of objects are considered: a set of positive objects and a set of negative objects. For the decision rule being induced, positive objects covered by the rule are *supporting* it, and negative objects covered by this rule are *contradicting* it. The ratio of the number of positive objects covered by the rule to the number of all objects covered by the rule is called *discrimination level*.

[1] This test is called cross-validation (CV) test. Presence of "noise" in data suggests giving priority to a self-test while selecting the attributes for further analysis. At the same time, due to large variance associated with the CV test results, the cognitive validity of this particular test should be downplayed and used only as a last resort. Nevertheless, discrimination among β-reducts with the smallest cardinality (step 5 of the procedure) using the results of CV test encourages the consideration of β-reducts with the best predictive powers.

The set of negative objects is always defined as the complement of the set of positive objects. The set of positive objects may be defined in one of two ways:

1. Indirectly, as either lower approximation or the boundary of a given class; in this case, *exact* and *approximate* rules are induced respectively. The discrimination level of induced rules is equal to 1 (Grzymala-Busse, 1992; Predki *et al.*, 1998); or
2. Directly, as a given class. In this case, the discrimination level of the induced rules is less than or equal to 1.

It is possible to generate the rules using the following induction strategies:

1. Induction of all possible rules. This approach provides the best insight into the analyzed data set (all existing relationships between attribute values and definition of classes are shown), but may be computationally inefficient even for small data sets.
2. Induction of a minimal set of rules (known as a *minimal covering*). This approach provides a minimal number of rules that cover all objects from the analyzed data set.
3. Induction of rules satisfying some user requirements (so-called *interesting rules* or *satisfactory description* (Mienko *et al.*, (1996)). This approach provides a set of rules that represents some information patterns and regularities in the analyzed data set, and as such can be helpful in understanding and explaining relationships between attribute values and definition of classes.

For the *interesting* rules, user requirements are defined in terms of:

- The minimal strength of a rule. This can be either absolute, as the number of positive objects covered by the rule; or relative, as the ratio of the number of positive objects covered by the rule to a number of all objects in the class. Rules that are weaker than a given threshold are not induced.
- The maximal length of a rule, defined through a number of elementary conditions in the condition part of the rule. Rules longer than a specified threshold are not induced.
- The minimal discrimination level of a rule. Rules with discrimination level smaller than a given threshold are not induced.

The induction of *interesting* rules (discovery-oriented induction) is based on a breadth-first exploration of the rule space restricted through the thresholds defined above.

The process of inducing *interesting* rules starts with the shortest rules (length equal to 1) and the rule length increases in next steps. In each step all rules are evaluated against the threshold values of length and strength. Rules that are too long or too weak are discarded. Then the level of discrimination of remaining rules is evaluated and all rules with an acceptable value of this measure are stored. Rules with unacceptable value of discrimination level are further specialized by adding new elementary conditions. The process stops when there are no more rules to consider.

It must be stressed that there is no claim that the *interesting* rules constitute a complete description of the classification in terms of the condition attributes. The interesting rules represent only a part, although a well-founded part, of domain knowledge. This is because the *interesting* rules do not cover all the objects from the decision table, and/or the decision table need not contain a representative sample of objects,

2.4. Discretization of Continuous Attributes

From the practical point of view, the indiscernibility relation may be applied only if the values of the attributes are symbolic (qualitative, discrete) as even very small differences in continuous values affect considerably the definition of atoms. To prevent this from happening, the continuous attributes should be discretized. As a result of discretizing, the precision of the original data is decreased (in the sense that the original values of the attributes cannot be reconstructed from the discrete values), but the generality of the data is increased.

It should be also stressed that discretization of continuous values is deeply embedded in human reasoning. For example, a decision maker often groups actual values together and considers discretized values such as ”low”, ”medium” or ”high”.

Discretization information typically consists of a finite set of numbered subintervals defined over the range of continuous attribute values, resulting in a *hard discretization*. This type of discretization is also referred to as *norm*-based discretization, because the subintervals are often defined following norms in the subject domain. Subintervals are used to discretize the continuous values by substituting the interval number to which the value belongs for the original value. A more advanced form of discretization involves subintervals represented as fuzzy numbers with overlapping bounds. This fuzzy form of discretization requires different (usually more advanced) techniques for processing the discretized decision tables (Slowinski and Stefanowski, 1994).

When a domain expert, following his/her judgment, specifies the subintervals for the discretization, the process is designated *expert discretization*. On the other hand, when the intervals are defined automatically, then the process is designated *automatic discretization* (for a review of automatic discretization procedures see Dougherty *et al*., 1995; Fayyad and Irani, 1993).