information

Explainable Decision-Making for Water Quality Protection

Jozo Dujmović [1] and William L. Allen III [2]

[1] Department of Computer Science, San Francisco State University, 1600 Holloway Ave, San Francisco, CA 94132, USA; jozo@sfsu.edu
[2] The Conservation Fund, 77 Vilcom Center Drive, Suite 340, Chapel Hill, NC 27516, USA; WAllen@conservationfund.org

Abstract. All professional decisions prepared for a specific stakeholder can and must be explained. The primary role of explanation is to defend and reinforce the proposed decision, supporting stakeholder confidence in the validity of the decision. In this paper we present the methodology for explaining results of the evaluation of alternatives for water quality protection for a real-life project, the Upper Neuse Clean Water Initiative in North Carolina. The evaluation and comparison of alternatives is based on the Logic Scoring of Preference (LSP) method. We identify three explainability problems: (1) the explanation of LSP criterion properties, (2) the explanation of evaluation results for each alternative, and (3) the explanation of the comparison and ranking of alternatives. To solve these problems, we introduce a set of explainability indicators that characterize properties that are necessary for verbal explanations that humans can understand. In addition, we use this project to show the methodology for automatic generation of explainability reports. We recommend the use of explainability reports as standard supplements for evaluation reports containing the results of evaluation projects based on the LSP method.

Keywords: Explainability; LSP Method; Water Quality; Decision-Making

1. Introduction

All decisions-support systems are used to prepare justifiable decisions for a specific stakeholder/decision-maker. The stakeholder can be an organization or an individual. The evaluation decision problem consists of identification of multiple alternatives, evaluation of each alternative using a justifiable multiattribute criterion, and selection of the best alternative. In this paper, evaluation is based on the LSP method [1]. In all cases, decisions are either rejected or accepted by human decision-makers. We assume that the stakeholder must achieve a sufficient degree of confidence before accepting and implementing a specific decision. A natural way to build the stakeholder's confidence is to provide acceptable explanation of reasons for each proposed decision. The credibility of any decision depends on the justifiability and completeness of explanations. The goal of this paper is to provide methodology for automatic generation of explainability reports that can be used to justify results of evaluation decisions. All numeric results in this paper are obtained using a new LSP.XRG software tool (*LSP Explainability Report Generator*).

As the area of computational intelligence becomes increasingly humancentric, explainability and trustworthiness have become a ubiquitous research topic, simultaneously present in many AI areas [2,3,4]. The problems that are explicitly considered are loan scoring, medical imaging and related automated decision-making, reinforced learning, recommender systems, user profiling [2], legal decision-making, and selection of job candidates [4]. In addition, humans still cannot trust results and decisions generated by machines in areas such as machine learning and data science where data veracity must be taken explicitly into account [5]. AI techniques are increasingly used to extract knowledge from data and provide decisions that humans can understand and accept from automatically provided explanations. The trustworthiness of such explanations is not always sufficient. On the other hand, explanations are necessary also in multiattribute decision-making, regardless of human effort to build justifiable multiattribute criteria [6].

All decision methods are based on criteria that include a variety of input arguments and adjustable parameters. Both the selected arguments and the parameters of evaluation criterion function (piecewise approximations of argument criteria, importance weights, and logic aggregation operators) are selected by stakeholders in cooperation with decision engineers [1]. All adjustable components must reflect the goals and interests od stakeholder/decision-maker, and that cannot be done with ultimate precision. Thus, justification and explanations processes are necessary support of decision making, and the primary topic of this paper.

In the area of decision-making, the trustworthiness of resulting decisions depends on the trustworthiness of evaluation criteria. In other words, explainability methods can contribute to both the criterion development and the acceptability of results. Therefore, before accepting the results of evaluation decisions, it is necessary to provide explanations that make the proposed decisions trustworthy. The goal of this paper is to contribute to explainability of LSP method, starting from initial results presented in [6], and to exemplify proposed explainability techniques on a realistic water quality protection problem [7,8], based on strategic conservation concepts presented in [9,10].

The paper is organized as follows. The water quality protection criterion is presented and analyzed in Section 2. In Section 3 we introduce concordance values of attributes and use them to explain the evaluation results. Explanation of comparison of alternatives is offered in Section 4. The automatic generation of an explainability report is discussed in Section 5, and Section 6 provides conclusions of this paper.

2. An LSP criterion for water quality protection

The decision-making explainability problems are related to specific LSP criterion. To illustrate such problems, we will use the criterion for the Upper Neuse Clean Water Initiative in North Carolina [7,8]. The goal is to evaluate specific locations and areas based on their potential for water quality protection. The evaluation team identified 12 attributes that contribute to the potential for water quality protection resulting in the LSP criterion shown in Fig.1. The stakeholders want to protect undeveloped lands near stream corridors that have soils that can absorb/hold water so that it is possible to avoid erosion and sedimentation and promote groundwater recharge and flood protection.

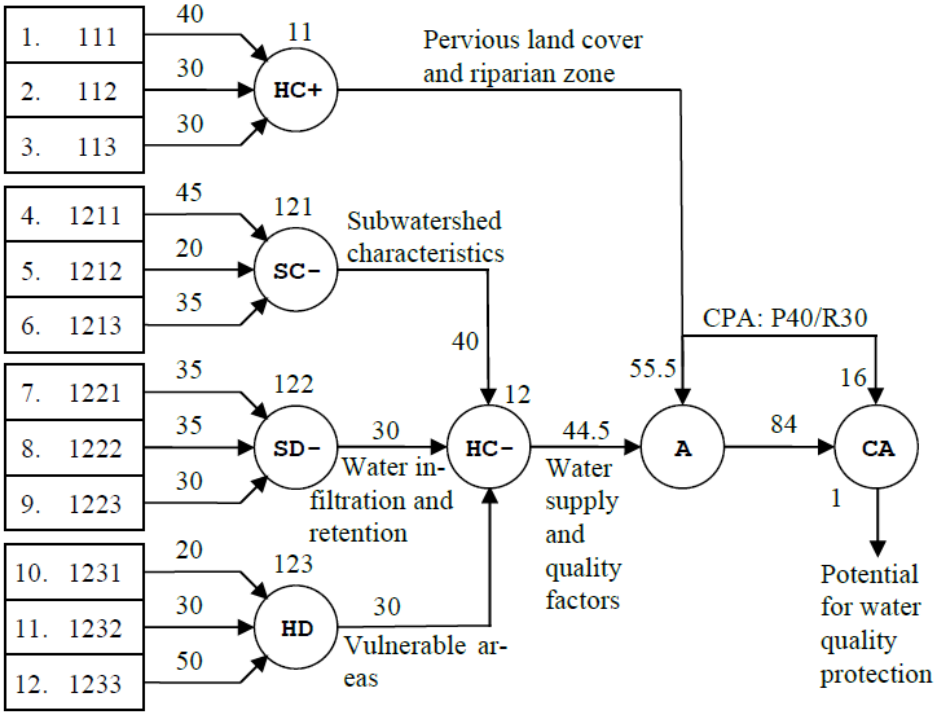
The aggregation structure in Fig. 1 is based on medium precision aggregators [1] with three levels (low, medium, high) of hard partial conjunction (HC-, HC, HC+) supporting the annihilator 0, hard partial disjunction (HD-, HD, HD+), supporting the annihilator 1, and soft conjunctive (SC-, SC, SC+) and disjunctive (SD-, SD, SD+) aggregators that do not support annihilators. These are uniform aggregators where the threshold andness is 75% (aggregators with andness or orness above 75% are hard, and aggregators with andness or orness below 75% are soft).

The nodes in the aggregation structure in Fig. 1 are numbered according to the LSP aggregation tree structure where the root node (overall suitability) is the node number 1, and generally, the child nodes of node N are denoted N1, N2, N3, and so on (e.g., the node N=11 has child nodes 111, 112, 113). In Fig. 1, for simplicity, we also numbered inputs 1, 2, …, 12, so that the input attributes are $a_1, a_2, \dots a_n,\ a_i \in \mathbb{R},\ i = 1, \dots, n;\ n = 12,$ and their attribute suitability scores that belong to $I = [0,1]$ are $x_1, x_2, \dots x_n,\ x_i \in I,\ i = 1, \dots, n$. The overall suitability is a graded logic function $L: I^n \to I$ of attribute suitability scores: $X = L(x_1, x_2, \dots x_n)$. The details of attribute criteria can be found in [8], and the results of evaluation and comparison of four competitive areas (denoted A, B, C, D), based on the criterion shown in Fig. 1, are presented in Fig. 2.

The point of departure in explaining the properties of the logic aggregation structure is the survey of sensitivity curves $X_i(x_i) = L(x_1, \dots, x_i, \dots x_n),\ x_k = c,\ k \neq i$, where c denotes a selected constant; typically, $c = 0.5$. The sensitivity curves show the impact of a single input, assuming that all other inputs are constant. Fig. 3 shows the sensitivity curves for the aggregation structure used in Fig. 1, in the case of $c = 0.5$.

The relative impact of individual inputs can be estimated using the values of the output suitability range $r_i[\%] = 100\,[X_i(1) - X_i(0)]$, $i = 1, \ldots, n$, and their maximum-normalized values $R_i[\%] = 100\, r_i/\max(r_1, \ldots, r_n)$, $i = 1, \ldots, n$. These indicators show the change of overall suitability caused by the individual change of selected input attribute suitability in the whole range from 0 to 1. Therefore, $R_i[\%]$ is one of indicators of the *overall impact* (or the overall importance) of the given suitability attribute. The corresponding ranking of attributes from the most significant to the least significant should be intuitively acceptable, explainable, and approved by the stakeholder. That is achieved in the ranking shown in Fig. 3 where the first three attributes (111, 112, 113) are mandatory, and all others are optional with different levels of impact. That is consistent with stakeholder requirements specified before the development of the criterion shown in Fig. 1. The normalized values $R_1, \ldots, R_n$ depend on the value of constant c, but their values and ranking are rather stable. In Fig. 3 we use $c = 0.5$. If $c = 0.75$, the ranking of the first six most significant inputs remains unchanged. Minor permutations occur in the bottom six less significant inputs.

111. Distance from riparian zone (M)	1221. Wetlands distance (O)
112. Pervious land cover type (M)	1222. Floodplain distance (O)
113. Percent of impervious surface (M)	1223. Groundwater recharge soil type (O)
1211. Headwaters designation (O)	1231. Wet/hydric soil type (S/O)
1212. Percent of protected land (O)	1232. Percent of steep slopes (S/O)
1213. Percent of forested land (O)	1233. Potential soil erodibility (S/O)
Coding: (M) = mandatory; (O) = optional; (S/O) = locally sufficient, globally optional	



Andness	0	1/14	2/14	3/14	4/14	5/14	6/14	1/2
Aggregator	D	HD+	HD	HD-	SD+	SD	SD-	A
Andness	8/14	9/14	10/14	11/14	12/14	13/14	1	3/4
Aggregator	SC-	SC	SC+	HC-	HC	HC+	C	CA

Figure 1. Twelve suitability attributes, the suitability aggregation structure, and the andness of medium precision hard (H) and soft (S), conjunctive (C) and disjunctive (D) aggregators used in the LSP criterion for evaluation of the potential for water quality protection.

Attribute ID	Attribute name	Area_A	Area_B	Area_C	Area_D
111	Distance from riparian zone	100.00	100.00	66.67	95.00
112	Pervious land cover type	100.00	50.00	80.00	10.00
113	Percent of impervious surface	88.00	82.00	56.67	8.57
1211	Headwaters designation	100.00	100.00	0.00	0.00
1212	Percent of protected land	40.00	80.00	20.00	10.00
1213	Percent of forested land	100.00	100.00	33.78	27.03
1221	Wetlands distance	50.00	20.00	0.00	50.00
1222	Floodplain distance	80.00	60.00	0.00	100.00
1223	Groundwater recharge soil type	80.00	100.00	20.00	20.00
1231	Wet/hydric soil type	0.00	0.00	0.00	100.00
1232	Percent of steep slopes	46.15	53.85	7.69	23.08
1233	Potential soil erodibility	50.00	100.00	10.00	80.00
121	Subwatershed characteristics	85.35	95.81	11.30	8.19
122	Water infiltration and retention	70.83	70.00	6.24	70.36
123	Vulnerable areas	44.94	100.00	7.56	100.00
11	Pervious land cover and riparian zone	95.07	59.64	63.61	9.83
12	Water supply and quality factors	63.74	87.09	8.08	16.64
1	Potential for water quality protection	83.11	69.62	41.64	12.27
OVERALL SUITABILITY SCORE [%]:		83.11	69.62	41.64	12.27

Figure 2. Results of evaluation of four areas (A, B, C, and D): suitability [%] for all inputs and for all subsystems of the aggregation structure shown in Fig. 1.

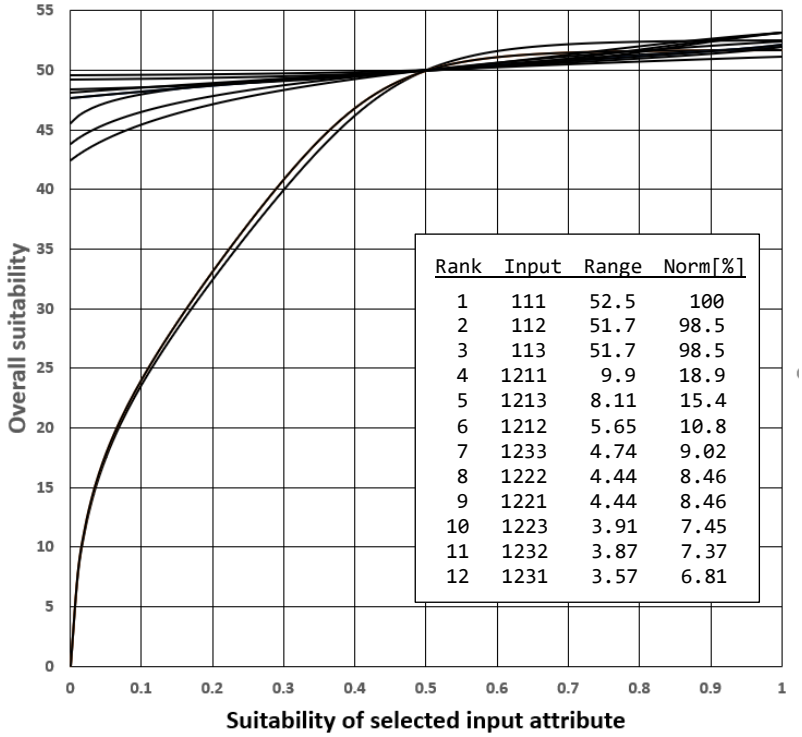


Figure 3. Sensitivity curves and ranking of attributes using the normalized range of impact.

The explainability of LSP evaluation project results is a process consisting of the following three main components:

1. Explainability of the LSP criterion
   1.1. Explainability of attributes
   1.2. Explainability of elementary attribute criteria
   1.3. Explainability of suitability aggregation structure
2. Explainability of evaluation of individual alternatives

2.1. Analysis of concordance values

2.2. Classification of contributors

3. Explainability of comparison of competitive alternatives

3.1. Analysis of explainability indicators of individual alternatives

3.2. Analysis of differential effects

The explainability of LSP criterion is defined as a general justification of the validity of criterion (i.e., the consistency between requirements/expectations and the resulting properties of criterion) without considering the available alternatives. In other words, this analysis reflects independent properties of a proposed criterion function. Most actions in the development of an LSP criterion are self-explanatory. The development of a suitability attribute tree is directly based of stakeholder goals, interests, and requirements. The selected suitability attributes should be necessary, sufficient, and nonredundant. Explainability of this step should list reasons why all attributes are necessary and sufficient. In our example, the tree is indirectly visible in Fig. 1. The attribute criteria (shown in [8]) come with descriptions that for each attribute criterion provide the explanation of reasons for a selected evaluation method. Regarding the suitability aggregation structure (Fig. 1), the only contribution to explainability consists of the sensitivity analysis for constant inputs and for ranking of the overall impact/importance of suitability attributes. All other contributions to explainability are based on specific values of inputs that characterize competitive alternatives.

3. Concordance values and explainability of evaluation results

In the case of evaluation of a specific object/alternative, each suitability attribute can provide different contributions to the overall suitability X. In the most frequent case of idempotent aggregation structures, we differentiate two groups of input attributes: *high contributors* and *low contributors*. High contributors are inputs where $x_i > X$; such attribute values are "above the average" and contribute to the increase of the overall suitability. Similarly, low contributors are inputs where $x_i < X$; such attribute values are "below the average" and contribute to the decrease of the overall suitability. Fig. 4 shows the comparison of five areas and all high contributor values are underlined. The overall suitability X shows the resulting ranking of analyzed areas: A≻ B≻ C≻ D≻ E.

SUITABILITY OF ATTRIBUTES: x[%]

	111	112	113	1211	1212	1213	1221	1222	1223	1231	1232	1233	X[%]	V[%]
Area_A	100.00	100.00	88.00	100.00	40.00	100.00	50.00	80.00	80.00	0.00	46.15	50.00	83.11	44.1
Area_B	100.00	50.00	82.00	100.00	80.00	100.00	20.00	60.00	100.00	0.00	53.85	100.00	69.62	46.5
Area_C	66.67	80.00	56.67	0.00	20.00	33.78	0.00	0.00	20.00	0.00	7.69	10.00	41.64	110.9
Area_D	95.00	10.00	8.57	0.00	10.00	27.03	50.00	100.00	20.00	100.00	23.08	80.00	12.27	86.3
Area_E	0.00	100.00	70.00	0.00	100.00	63.64	40.00	0.00	50.00	100.00	69.23	30.00	0.00	71.2

EXPLANATION: The overall suitability X of each evaluated object depends on the suitability scores of its attributes x=(x1,...,xn). The presented table shows the suitability scores of all attributes for all evaluated objects. The comparison of presented values shows strong and weak components of each object and explains why the object with the highest overall suitability outperforms other competitors. Those attributes that have suitability above the overall suitability are the primary positive contributors to the overall suitability. Attribute values below the overall suitability are negative contributors and the primary candidates for improvement.
We also show the coefficient of variation V of attribute scores. Low values of V correspond to objects that have balanced quality of attributes. Large values of V show objects that combine high and low satisfaction of attribute requirements. Generally, low values of V show balanced quality of attributes and are more desirable than high values. However, if important components are highly satisfied, and the low suitability scores correspond to those attributes that have low importance, that can be an acceptable and understandable strategy.

Figure 4. Suitability of attributes and the global suitability of five competitive areas (the underlined values indicate the high contributors).

For each attribute, there is obviously a balance point x_i^* where the i[th] input is in perfect balance with remaining inputs. This value is called the *concordance value* and it is crucial for explainability analysis. For all input attributes, the concordance values can be obtained by solving the following equations:

$$x_i^* = L(x_1, \ldots, x_{i-1}, x_i^*, x_{i+1}, \ldots, x_n), \quad i = 1, \ldots, n \ .$$

According to the fixed-point iteration concept [11], these equations can be solved, for each of n attributes, using the following simple convergent numerical procedure:

$\varepsilon = 0.0000001$; // or any other small value that defines the precision of x_i^*

$x_i^* = 0.5$; // or any other initial value inside the interval [0,1]

do

$x_i^* = L(x_1, \ldots, x_{i-1}, x_i^*, x_{i+1}, \ldots, x_n)$;

while ($|x_i^* - L(x_1, \ldots, x_{i-1}, x_i^*, x_{i+1}, \ldots, x_n)| > \varepsilon$);

The concordance values of all attributes for five competitive conservation areas, generated by LSP.XRG, are shown in Fig. 5. Note that the values of all attributes x_k, $k \neq i$, are not constants; they are the real values that correspond to the selected competitive area. The concordance value x_i^* shows the collective quality of all inputs different from i. If other inputs are high, then the concordance value of the i[th] input will also be high, reflecting the general demand for balanced, high satisfaction of inputs. Thus, the concordance values $x_i^* > X$ indicate low contributors, while $x_i^* \leq X$ characterizes high contributors as shown in Fig. 5 (in all LSP.XRG results the concordance values are denoted *c*). According to Figs. 4 and 5, the Area_E does not satisfy the mandatory requirement 111 (it is too far from the riparian zone) and therefore it is considered unsuitable and rejected by our evaluation criterion. So, the area_E will not be included in subsequent explanations.

CONCORDANCE VALUES OF ATTRIBUTES: c[%]

	111	112	113	1211	1212	1213	1221	1222	1223	1231	1232	1233	X[%]	V[%]
Area_A	75.34	77.58	80.02	82.46	83.97	82.62	83.99	83.24	83.22	88.93	89.28	89.62	83.11	5.1
Area_B	68.84	88.12	69.19	67.68	69.34	68.16	71.57	70.04	65.19	69.62	69.62	64.71	69.62	8.2
Area_C	10.69	11.96	11.96	42.34	41.84	41.78	42.92	42.92	42.58	42.48	42.56	42.66	41.64	38.6
Area_D	12.18	12.50	13.56	17.71	12.50	9.11	12.22	11.80	12.26	12.19	12.27	12.27	12.27	14.7
Area_E	48.65	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	331.7

EXPLANATION: The concordance value c of an attribute is defined as a value that is equal to the collective value of all other attributes of the evaluated object. Thus, if the actual value of an attribute (x) is equal to its concordance value (x=c), then that value is also the overall suitability: X = c = x. If the actual value of an attribute is greater than its concordance value (x > c), that yields X > c. Therefore, the concordance value of an attribute must be compared with the overall suitability X. If the concordance value is less than the overall suitability X, that indicates a desirable property where the attribute is called the "high contributor." That attribute is greater than the overall suitability. Consequently, it positively contributes to the overall suitability X. Similarly, the concordance value above the overall suitability X indicates a "low contributor," i.e., an attribute that is expected to be better, and so it negatively contributes to the overall suitability because its actual value is below X. Therefore, the above table explains the role of high and low contributors and helps in selecting those components that primarily need improvement.

We also present the value of the coefficient of variation of concordance values V[%]. If V is low, that indicates an object that has well balanced suitability of attributes. In contrast, high values of V show objects with heterogeneous suitability of input attributes. In a general case, that is less desirable.

Figure 5. Concordance values, their coefficient of variation (V[%]) and the global suitability (X[%]) of five competitive areas (the concordance values below the overall suitability show the high contributors). This is one of outputs generated by LSP.XRG.

The concordance values are suitable for explaining convenient and inconvenient properties of the specific evaluated area. Indicators that are proposed for explanation are defined in Table 1, and then applied and described in detail in Fig 6. The first question that most stakeholders ask is how individual attributes contribute to the overall suitability X. Since all values $x_1,\ldots,x_n$ contribute to the value of X, the most significant individual contributions come from inputs that have the lowest concordance values. Positive contributions shown in the individual contribution table in Fig. 6 correspond to high contributors and negative to low contributors. For example, the primary reason for the highest suitability of the Area_A (with individual contribution of 7.77%) comes from the proximity to riparian zone followed by the convenient pervious land cover type (5.53%) and low percent of impervious surface (3.1%). The individual contributions depend on the structure of the LSP criterion. For example, according to Fig. 4, the Area_A attributes 111, 112, 1211, 1213 have the highest suitability, but their individual contributions are in the range from 0.49% to 7.7%. The negative contributions of Area_A are in vulnerable areas attributes 1231, 1232, 1233 (each of them close to 6%).

Table 1. Basic explainability indicators

Explainability indicator	**Definition**
INDIVIDUAL SUITABILITY CONTRIBUTIONS OF ATTRIBUTES	$C_k = X - x_k^*$
TOTAL IMPACT OF ATTRIBUTES	$D_k = X_k^{max} - X_k^{min} = X[x_k = 1] - X[x_k = 0]$
TOTAL POTENTIAL FOR IMPROVEMENT	$D_k^+ = X_k^{max} - X = X[x_k = 1] - X$
ACCOMPLISHMENT OF ATTRIBUTES	$A_k = x_k - x_k^*$
BALANCE OF ATTRIBUTES	$B_k = x_k / x_k^*$

INDIVIDUAL SUITABILITY CONTRIBUTIONS OF ATTRIBUTES: (X-c)[%]

	111	112	113	1211	1212	1213	1221	1222	1223	1231	1232	1233	X[%]
Area_A	7.77	5.53	3.10	0.65	-0.85	0.49	-0.88	-0.12	-0.11	-5.81	-6.16	-6.51	83.11
Area_B	0.79	-18.50	0.43	1.94	0.29	1.46	-1.94	-0.42	4.44	0.00	0.00	4.91	69.62
Area_C	30.95	29.67	29.67	-0.71	-0.21	-0.14	-1.29	-1.29	-0.94	-0.84	-0.92	-1.02	41.64
Area_D	0.09	-0.23	-1.29	-5.44	-0.22	3.16	0.05	0.47	0.01	0.08	0.00	0.00	12.27

EXPLANATION: This table explicitly shows positive and negative contributions of individual attributes to the overall suitability. The highest positive contributor is the aggregator with the lowest concordance value. The lowest contributor is the aggregator with the highest concordance value.

TOTAL IMPACT OF ATTRIBUTES: (X[x=1]-X[x=0])[%]

	111	112	113	1211	1212	1213	1221	1222	1223	1231	1232	1233	X[%]	Q
Area_A	83.11	83.11	86.13	8.82	4.22	6.86	2.99	4.14	3.47	7.70	9.51	14.01	83.11	0.72
Area_B	69.62	89.80	69.79	12.09	6.71	9.77	3.73	4.53	10.86	0.00	0.00	11.10	69.62	0.62
Area_C	43.55	41.95	47.79	0.86	0.81	2.20	1.69	1.69	5.08	1.27	1.65	2.65	41.64	0.94
Area_D	12.27	12.59	13.71	13.22	7.23	13.21	0.12	0.58	0.09	0.08	0.00	0.00	12.27	0.32

EXPLANATION: Each attribute suitability is in the range from the minimum value 0 to the maximum value 1. Assuming that the values of all other attributes remain unchanged, the total impact of selected attribute on the overall suitability is Xmax-Wmin, where Xmax corresponds to the attribute suitability x=1, and Xmin corresponds to the attribute suitability x=0. This indicator reflects the overall significance of an attribute.
We are interested to have the highest satisfaction of the most significant attributes. That is evaluated using the Q factor, defined as a cosine of angle between the vector of suitability scores and the vector of total impact of attributes. The value of Q is between 0 and 1. High values of Q show a successful "first things first strategy," where evaluated objects best satisfy the most important attribute requirements. Of course, the high values of Q only show a good distribution of attribute satisfaction efforts. However, that can be achieved for any value of the overall suitability, including both high and low.

TOTAL POTENTIAL FOR IMPROVEMENT: (X[x=1]-X[x])[%]

	111	112	113	1211	1212	1213	1221	1222	1223	1231	1232	1233	max[%]
Area_A	0.00	0.00	3.02	0.00	1.07	0.00	1.51	1.01	0.96	7.70	7.70	7.70	7.70
Area_B	0.00	20.17	0.16	0.00	0.47	0.00	2.83	1.55	0.00	0.00	0.00	0.00	20.17
Area_C	1.92	0.31	6.16	0.86	0.52	0.62	1.69	1.69	1.65	1.27	1.27	1.27	6.16
Area_D	0.00	0.32	1.44	13.22	6.30	8.15	0.05	0.00	0.06	0.00	0.00	0.00	13.22

EXPLANATION: If the suitability of an attribute is x < 1 (or x < 100%), then it is a possible to improve the overall suitability X by increasing the current value of x to its maximum value 1 (100%), yielding X[x=1]. Of course, if the value of an attribute is x=1, then the potential for improvement is zero. This indicator shows what are the attributes where investing in improvement is the most needed and the most beneficial. The maximum values are shown in the last column.

INDIVIDUAL ACCOMPLISHMENT OF ATTRIBUTES: (x-c)[%]

	111	112	113	1211	1212	1213	1221	1222	1223	1231	1232	1233	min[%]
Area_A	24.66	22.42	7.98	17.54	-43.97	17.38	-33.99	-3.24	-3.22	-88.93	-43.12	-39.62	-88.93
Area_B	31.16	-38.12	12.81	32.32	10.66	31.84	-51.57	-10.04	34.81	-69.62	-15.78	35.29	-69.62
Area_C	55.98	68.04	44.70	-42.34	-21.84	-7.99	-42.92	-42.92	-22.58	-42.48	-34.87	-32.66	-42.92
Area_D	82.82	-2.50	-4.99	-17.71	-2.50	17.92	37.78	88.20	7.74	87.81	10.81	67.73	-17.71

EXPLANATION: For each attribute, the goal is to be better than other attributes of the same object. If an attribute has the concordance value c, then its value is exactly an average of the remaining attributes. Therefore, if the value x of an attribute is above the concordance value c, then the attribute is a positive contributor, and the difference x-c shows the individual accomplishment of the attribute. Of course, the individual accomplishment can also be negative. The minimum accomplishment shows the attribute that most notably needs improvement.

BALANCE OF ATTRIBUTES: x/c

	111	112	113	1211	1212	1213	1221	1222	1223	1231	1232	1233	V[%]
Area_A	1.33	1.29	1.10	1.21	0.48	1.21	0.60	0.96	0.96	0.00	0.52	0.56	46.68
Area_B	1.45	0.57	1.19	1.48	1.15	1.47	0.28	0.86	1.53	0.00	0.77	1.55	49.32
Area_C	6.24	6.69	4.74	0.00	0.48	0.81	0.00	0.00	0.47	0.00	0.18	0.23	150.73
Area_D	7.80	0.80	0.63	0.00	0.80	2.97	4.09	8.48	1.63	8.20	1.88	6.52	85.18

EXPLANATION: The ratio of the actual value of an attribute suitability x and its concordance value c shows the extent to which an attribute suitability outperforms the collective mean suitability of all other attributes. If the coefficient of variation V is low, then the attribute values of the analyzed object are well balanced.

Figure 6. Indicators used for explaining the evaluation results (output generated by LSP.XRG).

The overall impact of individual attributes is an indicator similar to the overall importance of attributes derived from sensitivity curves (defined as the range in Fig. 3). There is a difference: now we analyze the sensitivity of individual attributes based on real values of attributes of each individual alternative (areas A, B, C, D). That offers the possibility for ranking of attributes of individual alternatives according to their impact and (in cases where that is possible) to focus attention on the high impact attributes. However, the high impact is not the same concept as the high potential for improvement.

The potential for improvement is defined in Fig. 6 as a real possibility to improve the overall suitability of an alternative. For example, the highest impact attributes of Area_A are already satisfied, and the highest potential for improvement comes from attributes that are insufficiently satisfied. So, the potential for improvement is an indicator that shows (in situations where that is possible) the most impactful attributes that should have the priority in the process of improvement. Their maximum values show the highest potential for improvement of each alternative. Of course, that assumes the possibility of adjustment; unfortunately, physical characteristics of locations and areas cannot be changed.

If an attribute has the value that is significantly above the concordance value, that indicates a high accomplishment, because the quality of that attribute is significantly above the collective quality of other attributes. Exceptionally high accomplishments in a few attributes (e.g. 111, 1222, and 1231 in the case of Area_D) are insufficient to provide high overall suitability and are also an indicator of low suitability of remaining attributes, yielding low ranking of areas D and E (Fig. 4). In the case of Area_E, a single negative accomplishment in a mandatory attribute 111 is sufficient to reject that alternative.

The concordance values offer an opportunity to analyze the balance of attributes. If all attributes are close to their concordance values, that denotes a highly balanced alternative where all attributes have a similar quality. The coefficient of variation (V[%]) of the ratios of actual and concordance values of attributes shows the degree of imbalance and in Fig. 6 the lower quality areas C and D are also significantly imbalanced. Of course, the low imbalance does not mean high suitability; an alternative can have a highly balanced low quality. However, high imbalance generally shows alternatives that need to be improved. Note that the imbalance of attributes in Fig. 6 has the same ranking as the coefficient of variation of the concordance values in Fig. 5; these concepts are similar.

4. Explainability of the comparison of alternatives

Explainability of evaluation results contributes to understanding the results of ranking of individual alternatives. However, stakeholders are regularly interested in explaining the specific reasons why an alternative is superior/inferior compared to another alternative. Consequently, the comparison of alternatives needs explanations focused on discriminative properties of LSP criteria.

The superiority of the leading alternative in an evaluation project is a collective effect of all inputs and it cannot be attributed to a single attribute. However, an estimate of individual effects can be based on the direct comparison of the suitability degrees of individual attributes. Suppose that the Area_A has the attribute suitability degrees $a_1, \ldots, a_n$, and the Area_B has the attribute suitability degrees $b_1, \ldots, b_n$. Then, according to Fig. 4, we have $X_A = L(a_1, \ldots, a_n) = 83.1\%$ and $X_B = L(b_1, \ldots, b_n) = 69.6\%$. An estimate of the individual effect of attribute a_i, $i \in \{1, \ldots, n\}$, compared to the same attribute in the Area_B, can be obtained using the discriminators of attributes

$$R_i(A, B) = X_A - L(a_1, \ldots, a_{i-1}, b_i, a_{i+1}, \ldots, a_n), \; i = 1, \ldots, n$$

Similarly,

$$R_i(B, A) = X_B - L(b_1, \ldots, b_{i-1}, a_i, b_{i+1}, \ldots, b_n), \; i = 1, \ldots, n.$$

The discriminator $R_i(A, B)$ shows the individual contribution of selected attribute to the ranking A≻B. If $R_i(B, A) > 0$ then the selected attribute positively contributes to the ranking A≻B; similarly, if $R_i(B, A) < 0$, then the selected attribute negatively contributes to the ranking A≻B. If $b_i = a_i$, then there is no contribution of the selected attribute. We use n discriminators for all n attributes to explain the individual attribute contributions to the ranking of two objects/alternatives. This insight can significantly contribute to explainability reports.

If $R_i(A, B) > 0$, then a_i positively contributes to the ranking A≻B, and to condition $R_i(A, B) \times R_i(B, A) \leq 0$ (i.e., their signs are different). Since the discriminators $R_i(A, B)$, $i = 1, \ldots, n$ show the superiority of attributes of the Area_A with respect to the attributes of the Area_B, and $R_i(B, A)$ shows the superiority of attributes of the Area_B with respect to the attributes of the Area_A, it follows that these are two different views of the same relationship between two alternatives. To consider both views, we can average them and compute the *mean superiority* of the Area_A with respect to the Area_B for specific attributes as follows:

$$M_i(A, B) = [R_i(A, B) - R_i(B, A)]/2, \; i = 1, \ldots, n.$$

An overall indicator of superiority can be now defined as a "mean overall superiority"

$$M(A,B) = \frac{1}{n}\sum_{i=1}^{n} M_i(A,B).$$

The pairwise comparison of areas A, B, C, D is shown in Fig. 7. The first three rows contain the comparison of areas A and B. The first row contains discriminators $R_i(A,B)$, and the second row contains discriminators $R_i(B,A)$. The mean superiority $M_i(A,B)$ is computed in the third row. The rightmost column shows the overall suitability scores of competitive objects (X_A and X_B), followed by the mean overall superiority of the first object, $M(A,B)$. It should be noted that the individual attribute superiority indicators $M_i(A,B)$, $i = 1,\dots,n$ are useful for comparison of objects, and discovering critical issues, but they do not take into account the difference in importance between attributes. So, $M(A,B)$ shows unweighted superiority which is different from the difference in the overall suitability. Thus, we can investigate the values of the indicator $r = (X_A - X_B)/M(A,B)$. In our examples that value is rather stable (from 10.42 to 17.25), but not constant. This result shows that the overall indicator of superiority $M(A,B)$ is a useful auxiliary indicator for estimation of relationships between two competitive objects. The main contribution of discriminators to explainability is that they clarify the aggregator-based origins of dominance of one object with respect to another object.

OBJECT	111	112	113	1211	1212	1213	1221	1222	1223	1231	1232	1233	X;X;M[%]
Area_A	0.00	21.87	2.24	0.00	-0.79	0.00	0.84	0.72	-0.96	0.00	-0.71	-7.70	83.11
Area_B	0.00	-20.17	-0.08	0.00	1.31	0.00	-1.19	-0.81	2.95	0.00	0.00	7.07	69.62
Mean	0.00	21.02	1.16	0.00	-1.05	0.00	1.01	0.77	-1.96	0.00	-0.36	-7.38	1.10
Area_A	12.46	4.61	16.98	8.82	0.73	2.58	1.48	3.13	1.82	0.00	1.67	5.46	83.11
Area_C	-1.92	-0.31	-5.80	-0.86	-0.20	-0.62	-1.36	-1.58	-1.48	0.00	-0.97	-1.09	41.63
Mean	7.19	2.46	11.39	4.84	0.46	1.60	1.42	2.36	1.65	0.00	1.32	3.27	3.16
Area_A	0.96	55.45	57.21	8.82	1.42	3.04	0.00	-1.01	1.82	-7.70	1.25	-5.17	83.11
Area_D	-0.00	-0.32	-1.44	-13.22	-2.48	-8.15	0.00	0.09	-0.05	0.08	0.00	0.00	12.27
Mean	0.48	27.89	29.32	11.02	1.95	5.59	0.00	-0.55	0.94	-3.89	0.63	-2.58	5.90
Area_B	0.97	-16.46	1.88	12.09	2.50	3.85	0.89	2.99	8.75	0.00	0.00	10.63	69.62
Area_C	-1.92	4.75	-5.45	-0.86	-0.44	-0.62	-0.88	-1.45	-1.65	0.00	-1.06	-1.27	41.63
Mean	1.45	-10.61	3.67	6.47	1.47	2.24	0.89	2.22	5.20	0.00	0.53	5.95	1.62
Area_B	0.03	36.55	38.68	12.09	3.58	4.53	-1.19	-1.55	8.75	0.00	0.00	2.54	69.62
Area_D	-0.00	-0.32	-1.44	-13.22	-5.16	-8.15	0.04	0.18	-0.06	0.08	0.00	0.00	12.27
Mean	0.02	18.43	20.06	12.66	4.37	6.34	-0.61	-0.86	4.41	-0.04	0.00	1.27	5.50
Area_C	-1.85	31.11	32.17	0.00	0.13	0.16	-1.36	-1.69	0.00	-1.27	-0.54	-1.23	41.63
Area_D	0.00	-0.32	-1.44	0.00	-0.88	-1.03	0.07	0.58	0.00	0.08	0.00	0.00	12.27
Mean	-0.93	15.72	16.80	0.00	0.50	0.59	-0.72	-1.14	0.00	-0.68	-0.27	-0.62	2.44

Figure 7. Pairwise comparison of competitive areas A, B, C, and D.

From the standpoint of explainability of the comparison of objects, the individual indicators $M_i(A,B)$ explicitly show the predominant strengths (as high positive values) and predominant weaknesses (as low negative values) of the specific object. For example, in Fig. 7, the main advantage of the Area_A compared to the Area_B is the attribute 112 (pervious land cover) and the main disadvantage is attribute 1233 (potential soil erodibility). Such relationships are useful for summarized verbal explanations of a proposed decision that the protection of Area_A should have priority with respect to the protection of Area_B).

In cases where that is possible, the explicit visibility of disadvantages and weaknesses is useful for explaining what properties should be improved, and in what order. Of course, some evaluated objects (e.g. computer systems, cars, airplanes, etc.) have the possibility to modify suitability attributes in order to increase their overall suitability. In such cases, the explainability indicators such as the potential for improvement, the individual suitability contributions, and the individual superiority scores, provide the guidelines for selecting the most effective corrective actions. In the case of locations and areas that are

suitable for the water quality protection, the suitability attributes are physical properties that cannot be modified by decision-makers. In such cases the resulting potential for water quality protection cannot be changed, but the ranking of areas and explainability indicators are indispensable to make correct and trustworthy decisions about various protection and development activities.

5. Explainability report as a part of the decision documentation

Documentation of evaluation projects includes several main components. Each project starts with the specification of goals and interests of the stakeholder and the reasons for evaluating and selecting specific objects/alternatives. The next step is to develop the suitability attribute tree and elementary suitability attribute criteria that justifiably reflect the needs of the stakeholder. The suitability attributes are classified in basic groups of mandatory, optional, and sufficient inputs. These requirements are then implemented using appropriate logic aggregators in the suitability aggregation structure. This part of documentation is completed before the evaluation process. To justify the LSP criterion, it is useful to show sensitivity curves and to compute the ranking of importance of suitability attributes.

The evaluation process starts by documenting the available objects/alternatives. Then, the results of evaluation are presented as the suitability in each node of the aggregation structure, from input suitability degrees $x_1, \ldots, x_n$ to the overall suitability X. The ranking of alternatives is based on the decreasing values of the overall suitability scores. The highest suitability score indicates the alternative that is proposed for selection and implementation. In cases where alternatives have costs, the suitability and affordability are conjunctively aggregated to compute the overall value score [1] which is then used for selecting the best alternative.

In addition to the above traditional documentation, generated using LSP.NT [12], in this paper we introduced an additional *explainability report* that provides the explanation and justification of obtained results. That report is generated by the LSP Explainability Report Generator (LSP.XRG) tool. The results generated by LSP.XRG are exemplified in Figs. 2-7. The explainability report is based primarily on the following set of explainability indicators:

- Overall importance of suitability attributes (based on evaluation criterion)
- Concordance values of suitability attributes for each alternative
- Individual suitability contributions of attributes
- Total impact of individual suitability attributes for each alternative, and sensitivity analysis curves
- Total potential for improvement for each suitability attribute and for each competitive object/alternative
- Accomplishments of individual attributes for each alternative
- Balance of attribute values for each alternative
- Pairwise comparison of competitive objects/alternatives

In the case of evaluation of various locations/areas from the standpoint of their potential for water quality protection we provided the explainability indicators in Figs. 4-7. These indicators can be used in several ways. First, all tables with results can be automatically generated by LSP software support tools. Then, it is possible to compose a verbalized summary report based on explainability indicators. Finally, the information stored in explainability tables created by the LSP.XRG can be selectively inserted in executive summaries and used during stakeholder meetings and approval processes. The explainability results and explainability documentation significantly contribute to the confidence that stakeholders must have in evaluation results and proposed decisions.

6. Conclusions

Decisions are results of human mental activities, and consequently all decision methods should have a strong humancentric component. That includes the explainability of proposed decisions. Trustworthiness and explainability are currently important topics (and active research areas) particularly in cases where AI tools are used to automatically discover knowledge in large databases and propose decisions that affect human conditions and actions. In such cases, the trustworthiness of decisions becomes the critical issue.

In this paper we have shown that explainability and trustworthiness are equally important and useful also in the decision-making process that involves a permanent presence of humans as stakeholders, decision engineers, domain experts, and executives. This process includes the specification of alternatives, the development of evaluation criteria, the specification of requests to vendors or system developers, and the final evaluation of competitive alternatives, selection of the best alternative, and justifying the decision to approve its implementation.

The proposed explainability indicators and their use are developed in the context of the LSP decision method, where all explanatory presentations can be integrated in a specific explainability report. Our example of the Upper Neuse Clean Water Initiative in North Carolina was selected as a realistic decision project where explainability is important because of the large number of stakeholders, which include all interested in the protection of clean water supply in perpetuity. That includes municipalities, companies, various social organizations, and individual citizens. For all decisions in this situation, it is necessary to provide convincing evaluation results, as well as verbal and quantified explanations. In this paper we proposed a solution of that problem. The same methodology is equally applicable in practically all other decision projects based on the LSP method.

References

1. Dujmović, J.: *Soft Computing Evaluation Logic*. Wiley and IEEE Press (2018).
2. Alonso-Moral, J.M., Mencar, C., Ishibuchi, H.: Explainable and Trustworthy Artificial Intelligence. Computational Intelligence, Vol 17, No. 1, pp. 14-15 (2022).
3. Deng, Y., Weber: GW: Fuzzy Systems Toward Human-Explainable Artificial Intelligence and Their Applications. IEEE Transactions of Fuzzy Systems, Vol. 29, No 12, pp. 3577-3578 (2021)
4. Wing, J.M: Trustworthy AI. Communications of the ACM, Vo. 64, No. 10 pp. 64—71 (2021)
5. De Tré, G., Dujmović, J.: Dealing with Data Veracity in Multiple Criteria Handling: An LSP-Based Sibling Approach. Proceedings of the International Conference on Flexible Query Answering Systems, FQAS 2021, Bratislava, Slovakia. In T. Andreasen et al. (Eds.). Flexible Query Answering Systems, Springer LNAI 12871, pp. 82-96 (2021).
6. Dujmović, J.: Interpretability and Explainability of LSP Evaluation Criteria. Proceedings of the 2020 IEEE World Congress on Computational Intelligence, 978-1-7281-6932-3/20, paper F-22042 (2020).
7. Conservation Trust for North Carolina. 2020. https://ctnc.org/impressive-accomplishments-already/ Accessed March 19, 2022.
8. Dujmović, J., Allen III, W.L.: Soft Computing Logic Decision Making in Strategic Conservation Planning for Water Quality Protection. Ecological Informatics, Vol. 61, 101167. (2021).
9. Messer, K. and W. Allen: *The Science of Strategic Conservation Planning: Protecting More with Less*. Cambridge, UK: Cambridge University Press (2018).
10. Benedict, M. And E. McMahon: *Green Infrastructure: Linking Landscapes and Communities*. Washington, DC: Island Press (2006).
11. Polyanin, A.D. and A.V. Manzhirov, *Handbook of Mathematics for Engineers and Scientists*. Chapman & Hall/CRC, 2007.
12. SEAS Co, LSP.NT - LSP method for evaluation over the Internet. LSP.NT V1.2 User Manual. Accessible from http://www.seas.com/LSPNT/login.php, (2020).