Interpretability and Explainability of LSP Evaluation Criteria*

Jozo Dujmović, *LSM, IEEE*

***Abstract*— Logic Scoring of Preference (LSP) is a soft computing decision method for evaluation and selection of complex objects and alternatives using logic criteria. LSP criteria are based on graded logic aggregation structures that in most cases have a canonical form of tree. Such trees aggregate degrees of truth or degrees of fuzzy membership. The aggregators are graded logic functions. Each node in the tree of logic aggregators has a specific semantic identity – it has the interpretation, role, meaning, and importance for the decision maker. The semantic identity of all arguments can be used to develop explainable LSP criteria, and to provide explanation of all results generated in the process of evaluation, comparison, and selection of complex objects and alternatives. In this paper we propose explainability parameters and use them in decision making to provide the explainability of both the results of evaluation of a single object, and the results of comparison and selection of multiple competitive alternatives.**

I. INTRODUCTION

All decisions can and should be explained. Any method for selecting the best among several alternatives has limited credibility unless it can provide a verbalized explanation of the proposed decision. Stakeholders regularly want to know a convincing answer to the "why question." Unfortunately, explanations of decisions do not come automatically, and in many cases it is not obvious what qualifies as an acceptable interpretation of results and explanation of reasons that support the credibility of results. In this paper we propose indicators that can be used in the process of interpreting and explaining results generated by LSP criteria.

LSP criteria are functions $G:\mathbb{R}^n \to I=[0,1]$ that use n input suitability attributes $a_1,...,a_n,\ a_i \in \mathbb{R},\ i=1,...,n$ of an evaluated object/alternative and return the resulting overall suitability degree $X \in I$, as shown in Fig. 1. Input suitability attributes are not arbitrarily selected – they include only those attributes that provably affect the capability of evaluated object to satisfy stakeholder's goals and requirements. The resulting overall suitability score $X=G(a_1,...,a_n)$ can be interpreted as the degree of truth of the value statement claiming that an evaluated object completely satisfies specific stakeholder's requirements [1]. So, $X=1$ denotes the complete satisfaction of all requirements and $X=0$ denotes a fully unacceptable object/alternative that should be rejected. Alternatively and equivalently, we can interpret X as the degree of fuzzy membership of evaluated object in the set of those objects that completely satisfy all requirements [2]-[4].

The process of computing the overall suitability X using the LSP method consists of two steps. In the first step we define the set of attribute criteria $g_i:\mathbb{R}\to I,\ i=1,...,n$ and use them to compute the attribute suitability degrees $x_i=g_i(a_i),\ i=1,...,n$. In the second step the attribute suitability degrees are aggregated to compute the overall suitability degree, as follows:

$$X=L(x_1,...,x_n)=L(g_1(a_1),...,g_n(a_n))=G(a_1,...,a_n)\ .$$

The function $L:I^n\to I$ is a compound graded logic function: both inputs and the output are degrees of truth, i.e. graded logic variables. In other words, L is a formula of the graded logic propositional calculus. According to the graded logic conjecture [2], the L function can be organized as a superposition of ten necessary and sufficient basic types of graded logic functions: hyperconjunction, full conjunction, hard partial conjunction, soft partial conjunction, neutrality, soft partial disjunction, hard partial disjunction, full disjunction, hyperdisjunction, and negation. These basic logic functions integrate models of means (from full conjunction to full disjunction) and selected t-norms and t-conorms belonging to hyperconjunction and hyperdisjunction groups of aggregators. Except for negation, all other basic types of graded logic functions have the status of logic aggregator. All logic aggregators are observable in human intuitive reasoning and fully cover all regions of the unit hypercube I^n.

By definition [2],[8] a logic aggregator $A(x_1,...,x_n)$ is nondecreasing in all arguments, satisfies the boundary conditions $A(0,...,0)=0$, $A(1,...,1)=1$, and is sensitive to positive and incomplete truth: if $x_i>0,\ i=1,...,n$ then $A(x_1,...,x_n)>0$; if $x_i<1,\ i=1,...,n$ then $A(x_1,...,x_n)<1$. If a logic aggregator is idempotent, then it is a mean [5]-[7]: $\min(x_1,...,x_n)\le A(x_1,...,x_n)\le \max(x_1,...,x_n)$, and if it is not idempotent, then it can be either hyperconjunctive (satisfying $\lfloor x_1\cdots x_n\rfloor \le A(x_1,...,x_n)<\min(x_1,...,x_n)$) or hyperdisjunctive (satisfying $\max(x_1,...,x_n)<A(x_1,...,x_n)\le 1-\lfloor(1-x_1)\cdots(1-x_n)\rfloor$).

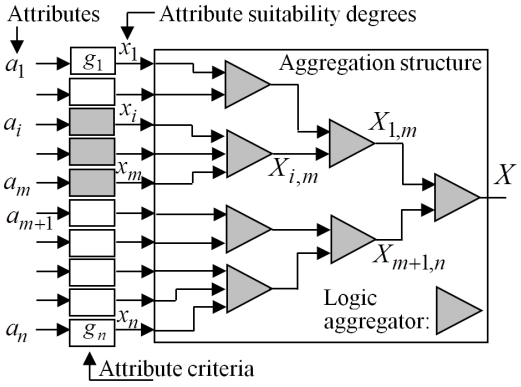

Figure 1. The structure of LSP criteria

*J. Dujmović is with the Dept. of Computer Science, San Francisco State University, San Francisco, CA 94132 USA (e-mail: jozo@sfsu.edu).

978-1-7281-6932-3/20/$31.00 ©2020 IEEE

The goal of this paper is to analyze LSP evaluation criteria, and to provide techniques that decision makers can use to explain the reasons for evaluation and comparison results. The paper is organized as follows. Interpretability and explainability parameters are discussed in Sections II and III. Sections IV and V are devoted to the explainability of evaluation and comparison. Section VI investigates the certainty of suitability evaluation results. Section VII presents a verbalizer (an explanation tool) and Section VIII includes a summary and conclusions.

II. Inherent Interpretability of Aggregation Trees

LSP criteria include a suitability aggregation structure that has the form of a tree. In such a structure each node has a *shadow*, which is defined as a subset of inputs that affect the suitability in that node. E.g., in Fig. 1, $X_{i,m}$ denotes the node suitability where the shadow is $(a_i,...,a_m)$. The role of a node can be interpreted by aggregating the roles and properties of its shadow. The shadow of the root of the aggregation tree $(X = X_{1,n})$ is the set of all suitability attributes $(a_1,...,a_n)$. Obviously, some sets of adjacent input attributes cannot be a shadow; e.g., in Fig. 1, $(a_i,...,a_{m+1})$ is not a shadow of any node in the aggregation tree.

The suitability aggregation tree can be interpreted as a feedforward neural network [11]. Generally, multilayer feedforward neural networks have interpretation (explainable role and meaning) only of output nodes; intermediate nodes cannot be interpreted. That makes the explainability of results a very difficult problem. Similarly, the lack of interpretability is a fundamental problem of many ML black box models [9].

In contrast to general neural networks, each node of the tree-structured suitability aggregation networks has a clearly defined shadow-based interpretation (semantic identity). That is a natural consequence of the fact that the aggregation tree is developed from the root towards the leaves, by decomposing compound subsystems into their components. The aggregation process uses the same structure but in the opposite direction. Consequently, tree structures are inherently interpretable: each node has a clear identity and role, and its suitability can be used to prove and explain the reasons why the overall suitability is high or low.

III. Explainability Parameters

Let us consider the logic function $X = L(x_1,...,x_n)$, $n>1$, where all variables are degrees of truth and belong to [0,1] (or [0,100%]). That function can be the complete LSP aggregation structure or any part of it, including a single aggregator. We assume that this function satisfies basic conditions for logic aggregators: $L(0,...,0)=0,\ L(1,...,1)=1,$ continuous nondecreasing monotonicity in all arguments, and the sensitivity to positive and incomplete truth.

All arguments of logic aggregators have the status of *contributors* to the aggregated value *X*. In the most frequent case of idempotent logic aggregators, $\min(x_1,...,x_n) \le X \le \max(x_1,...,x_n)$. In such cases, we can differentiate two subsets of arguments: *high supporters* and *low supporters*. The high supporters $\overline{x}_{(i)} \in \{x_1,...,x_n\},\ i \in \{1,...,n\}$ are all arguments that have suitability degree greater than or equal to *X*. The low supporters (denoted $\underline{x}_{(i)},\ i \in \{1,...,n\}$) are all arguments that have suitability degree less than or equal to *X*. We assume that the arguments are sorted so that $x_{(i)}$ denotes the ith smallest value: $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$. In this case the logic aggregator is $X = L(\underline{x}_{(1)},...,\underline{x}_{(k)},\overline{x}_{(k+1)},...,\overline{x}_{(n)})$, where $\min(x_1,...,x_n) = \underline{x}_{(1)} \le \cdots \le \underline{x}_{(k)} \le X \le \overline{x}_{(k+1)} \le \cdots \le \overline{x}_{(n)} = \max(x_1,...,x_n)$. Therefore, the first explanation of the reasons for result *X* is that *X* has *high supporters* $\overline{x}_{(k+1)},...,\overline{x}_{(n)}$ and *low supporters* $\underline{x}_{(1)},...,\underline{x}_{(k)}$. The high supporters are responsible for attaining the suitability *X* and the low supporters are responsible for preventing the aggregated suitability to grow above *X*. This explains the reasons for specific value of *X* and the roles of contributors.

The next questions are related to actions that can improve *X*. For each argument, we can define the *potential contribution* (or *suitability increment*) as follows:

$$\delta_i^+ = L(x_1,...,1,...,x_n) - L(x_1,...,x_i,...,x_n),\ i=1,...,n\ .$$

The potential contribution shows the effect attained if the ith input is improved as much as theoretically possible. We can also define the following *contributing power* of the ith input:

$$p_i = \delta_i^+ / (1-x_i),\ (x_i<1),\ i=1,...,n\ .$$

The contributing power can also be defined in selected points as the *sensitivity coefficient* $s_i(x_i) = \partial L / \partial x_i,\ i=1,...,n$.

We are sometimes interested to know what is the *potential suitability decrement* that can be obtained from the ith input:

$$\delta_i^- = L(x_1,...,x_i,...,x_n) - L(x_1,...,0,...,x_n),\ i=1,...,n\ .$$

The potential suitability decrement shows the extent of suitability reduction that can result from unsatisfying requirements of the ith input. Hard conjunctive aggregators support the annihilator 0 and for them $L(x_1,...,0,...,x_n)=0$, yielding the same value of the suitability decrement.

The *total range* of the ith input can be expressed as the sum of the potential increment and decrement, $\rho_i = \delta_i^+ + \delta_i^-$, showing the range of influence of the ith input:

$$\rho_i = L(x_1,...,1,...,x_n) - L(x_1,...,0,...,x_n),\ i=1,...,n\ .$$

The total range shows the possible variation of output suitability. In some cases this can reflect the impact of the ith input. Most indicators, including δ_i^+, δ_i^-, and ρ_i, can be multiplied by 100 and expressed as percentages.

The *relative position of output* $X = L(x_1,...,x_i,...,x_n)$ inside the range ρ_i is $\pi_i = \delta_i^- / \rho_i$ and the *potential for suitability improvement* can be expressed as $Q_i = \delta_i^+ / X$.

To exemplify these indicators, let us use a simple hard partial conjunction $L(x,y) = x^{0.2}y^{0.8}$ shown in Fig. 2. Let the low and high supporters be $x = X = 0.4$ and $y = Y = 0.8$.

Then, $L(X,Y)=0.696$, denoted in Fig. 2 as a dashed line. In this example, the suitability increments, decrements, and ranges of x and y are approximately equal: $\delta_y^+ \cong \delta_x^+ = 0.14$, $\delta_x^- = \delta_y^- = 0.696$, $\rho_y \cong \rho_x = 0.836$. However, the impact of these arguments is not the same. According to weights, the impact of y is significantly higher than the impact of x. This is visible as different contributing power indicators of these two arguments: $p_x = 0.23$, $p_y = 0.68$. The power ratio $p_y / p_x = 2.9$ indicates the significantly higher impact of y.

The study of the overall impact of arguments contributes to explainability and deserves more attention. Many complex evaluation criteria are idempotent and have conjunctive character (concave shape of sensitivity curves similar to those shown in Fig. 2). If an input argument is mandatory but not very significant, the corresponding sensitivity curve has a very high first derivative for small values of argument suitability, and then the first derivative quickly decreases (as in the case of $L(x,Y)$ in Fig. 2). This is the expected behavior: there is a strong need to satisfy such input but, after it is initially satisfied, the impact of such an argument becomes rather low. On the other hand, if an argument is very important, its first derivative shows minor variations as in the case of argument y and curve $L(X,y)$ in Fig. 2. Such sensitivity curves are close to the "line of maximum impact," $L(X_1,...,x_i,...,X_n) = x_i L(X_1,...,1,...,X_n)$ which is denoted as the dotted line in Fig. 2. Therefore, the overall impact can be quantified as the proximity to the line of maximum impact, using the following *conjunctive coefficient of impact*:

$$\gamma_i = 200\frac{L(X_1,...,1,...,X_n) - \int_0^1 L(X_1,...,x_i,...,X_n)dx_i}{L(X_1,...,1,...,X_n)},$$

$$0 \le \gamma_i \le 100\%, \quad i = 1,...,n.$$

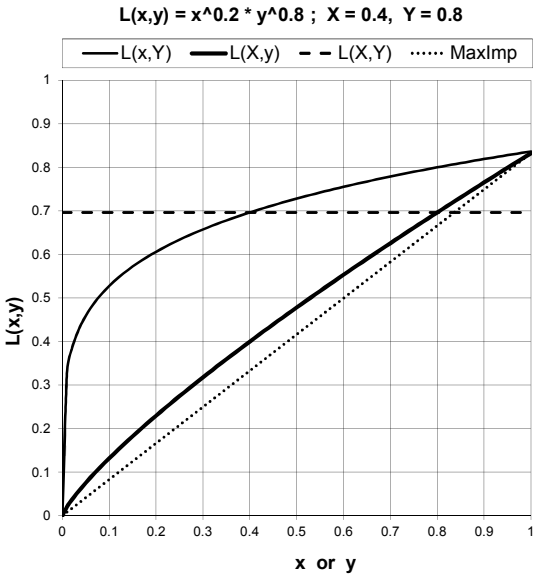

Figure 2. Sample hard partial conjunction aggregator

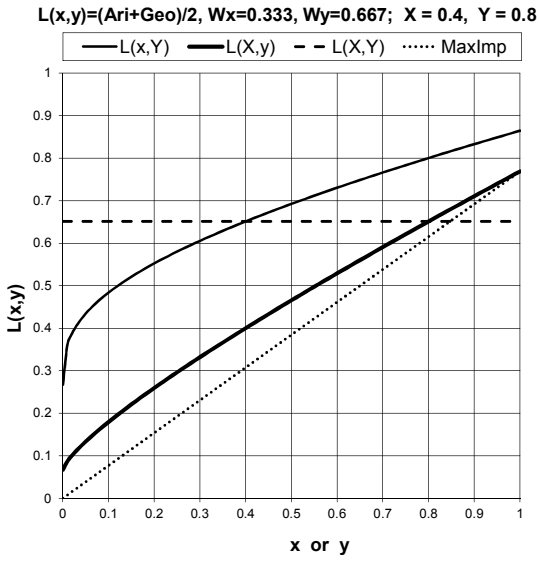

Figure 3. Sample soft partial conjunction aggregator

In an extreme case $L(X_1,...,x_i,...,X_n) = x_i L(X_1,...,1,...,X_n)$, we have $\gamma_i = 200(1-\int_0^1 x_i dx_i) = 100\%$, and for all realistic concave sensitivity curves we have $\gamma_i < 100\%$. In the case of aggregator presented in Fig. 1 the resulting coefficients of impact are $\gamma_x = 33.5\%$, $\gamma_y = 88.9\%$. That gives rather consistent ratios of coefficients of impact ($\gamma_y / \gamma_x = 2.65$), and the contributing power indicators ($p_y / p_x = 2.92$).

All presented explainability indicators can also be used with soft conjunctive aggregators. An example of soft partial conjunction is $L(x,y) = 0.5x^{1/3}y^{2/3} + 0.5(x/3 + 2y/3)$. It is presented in Fig. 3. We again use the low and high supporters $x = X = 0.4$ and $y = Y = 0.8$. The aggregated suitability $L(X,Y) = 0.651$ is denoted in Fig. 3 as a dashed line. In this example the suitability increments, decrements, and ranges of two arguments are: $\delta_x^+ = 0.21$, $\delta_y^+ = 0.12$, $\delta_x^- = 0.38$, $\delta_y^- = 0.58$, $\rho_x = 0.59$, $\rho_y = 0.70$, $\pi_x = 0.64$, $\pi_y = 0.83$. The different impact of arguments x and y is visible as follows: $p_x = 0.36$, $p_y = 0.59$, $\gamma_x = 44.2\%$, $\gamma_y = 81.8\%$, $\gamma_y / \gamma_x = 1.85$, $p_y / p_x = 1.66$. In this example we again have consistent ratios of the coefficients of impact, and the contributing power indicators.

If all arguments except x_i have fixed values, then it is interesting to solve the following equation:

$$L(X_1,...,x_i,...,X_n) = x_i$$

The solution $x_i = C_i$ is called the *concordance value*. This value is the result of aggregation obtained by aggregating all arguments except x_i. The idea of concordance value is easily visible when applied to means. For example, for geometric mean $y = x_1^{w1} x_2^{w2} x_3^{w3} x_4^{w4}$, $w_1 + w_2 + w_3 + w_4 = 1$, and the

argument x_1, the solution of equation $x_1^{w1}x_2^{w2}x_3^{w3}x_4^{w4} = x_1$ is $x_1 = x_2^{w2/(w2+w3+w4)}x_3^{w3/(w2+w3+w4)}x_4^{w4/(w2+w3+w4)} = C_1$. Therefore, the concordance value C_1 is the geometric mean of the remaining arguments x_2, x_3, x_4. If $x_1 = C_1$ then $y = x_1^{w1}x_2^{w2}x_3^{w3}x_4^{w4} = C_1$. In other words, this value of x_1 is neutral and in perfect balance (concordance) with the values of the remaining arguments. If $x_1 > C_1$ then x_1 becomes the high supporter of *y*, and if $x_1 < C_1$ then x_1 becomes the low supporter of *y*. The concordance values are suitable for computing the sensitivity coefficients $s_i(C_i) = \partial L / \partial x_i\big|_{x_i=C_i}$, $i = 1,...,n$. In addition, a useful explainability indicator is also the *coefficient of balance* $\beta_i = x_i / C_i$. This coefficient shows the extent to which x_i outperforms the remaining arguments ($\beta_i \approx 1$ denotes an "average contributor"). For example, in Fig. 2 and Fig. 3 it is easy to see the concordance values $C_x = 0.8$ and $C_y = 0.4$. So, $\beta_x = 0.5$ and $\beta_y = 2$.

The presented explainability indicators are used for concave functions that represent idempotent conjunctive aggregators. Since idempotent convex disjunctive aggregators are De Morgan duals of concave conjunctive aggregators, in most cases we can use the same explainability indicators as for conjunctive aggregators. A typical soft partial disjunction aggregator (disjunctive weighted power mean $L(x_1, x_2, x_3) = (w_1x_1^3 + w_2x_2^3 + w_3x_3^3)^{1/3}$) is shown in Fig. 4. Such aggregators can be characterized using basic explainability parameters: δ_i^+, δ_i^-, ρ_i, p_i, π_i, C_i, and β_i. Unsurprisingly, γ_i cannot be directly used for disjunctive aggregators, but it can be substituted by the power p_i.

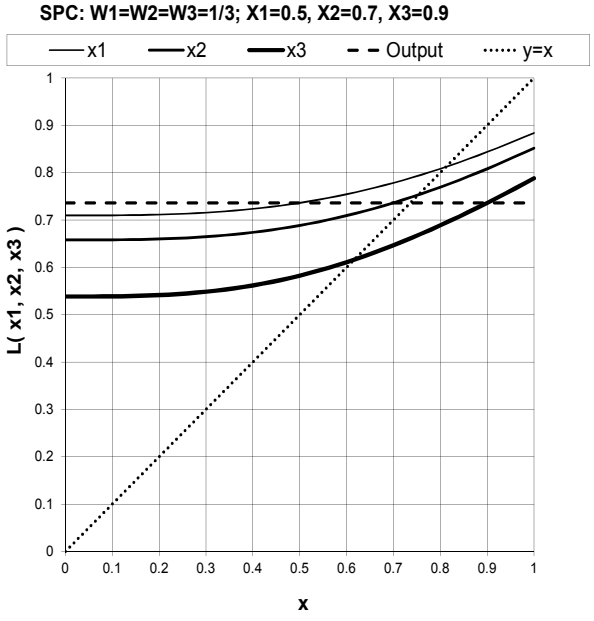

Figure 4. Sample soft partial disjunction aggregator

Hyperconjunction and hyperdisjunction are not idempotent and can combine concave and convex properties. For example $L(x_1, x_2, x_3) = x_1^{0.5}x_2x_3^{1.5}$ is a nonidempotent hyperconjunction where $L(x_1, X_2, X_3)$ is concave in x_1 but $L(X_1, X_2, x_3)$ is convex in x_3. Since all aggregators satisfy nondecreasing monotonicity, the indicators δ_i^+, δ_i^-, ρ_i, p_i, π_i, and C_i can be used for hyperconjunctive and hyperdisjunctive aggregators.

IV. Explainability of Evaluation Results

A typical structure of an LSP criterion is shown in Fig. 1. The criterion consists of *n* input attributes $a_1,...,a_n$ and *n* elementary attribute criteria $g_1,...,g_n$ that produce *n* attribute suitability degrees $x_1,...,x_n$. The attribute suitability degrees are aggregated using a tree structure of logic aggregators, yielding the overall suitability degree *X*. For each aggregator we define its shadow as the set of attributes that affect the output of the aggregator. The shadow specifies the contributors of selected aggregator and can be used in the explanation process to identify the role and significance of the aggregated output suitability.

For given values $x_i = X_i$, $i = 1,...,n$, the most frequent explainability question is to indicate which input is the most significant positive (or negative) contributor to the resulting overall suitability $X = L(X_1,...,X_i,...,X_n)$. If the i[th] input has the concordance value $x_i = C_i$, then the corresponding overall suitability $C_i = L(X_1,...,C_i,...,X_n)$ depends on inputs other than x_i; since C_i is the average of remaining $n-1$ inputs, it is in perfect balance with them. If we now increase (or decrease) this value to its actual value $x_i = X_i$ then the *individual contribution* of the i[th] input can be expressed as follows:

$$\Delta_i = L(X_1,...,X_i,...,X_n) - L(X_1,...,C_i,...,X_n) = X - C_i .$$

The input that has the highest positive Δ_i value is the primary contributor to the overall suitability. Similarly, the input that has the lowest negative Δ_i value is the primary reducer of the overall suitability. Consequently, the Δ_i values play a significant role in the explanation of results.

The explanation of evaluation results for any value of *n* and for any degree of complexity of the criterion function can be done in various ways. In the context of an LSP criterion the fundamental explainability question is to describe where the value of X comes from. The structure of LSP criterion is appropriate for several forms of explainability:

- Explainability of the whole criterion, using attribute suitability degrees $x_1,...,x_n$ as inputs and the overall suitability degree *X* as the output. The explanation can be based on selected explainability parameters Δ_i, δ_i^+, δ_i^-, ρ_i, p_i, π_i, γ_i, C_i, and β_i.

- Explainability of the whole criterion, using attributes $a_1,...,a_n$ as inputs and the overall suitability degree X as the output.
- Explainability of any intermediate suitability degree, e.g. $X_{i,m}$, with respect to the inputs in its shadow (i.e., $a_i,...,a_m$ or $x_i,...,x_m$).
- Top-down hierarchical explainability: explaining the overall suitability X as a function of its immediate inputs, and then explaining each input as a consequence of its inputs at the lower level. In the simplest case, this process of successive more detailed explanations can be based on tracing high and low supporters and continue until we reach the leaves of the aggregation tree.

The explainability process can be exemplified using a sample idempotent conjunctive aggregation structure shown in Fig 5, and its sensitivity analyses shown in Figs. 6 and 7 (we must now use the notation $x1, x2,...$ instead of $x_1, x_2,...$).

An initial explainability analysis of the LSP aggregation structure shown in Fig. 5 can be based on fixed values of input suitability degrees (typically 0.5 or 0.75). The corresponding sensitivity curves for $x1 = x2 = \cdots = x8 = 0.5$ are shown in Fig. 6. Such curves can be used to explain the general properties of the evaluation criterion. The inputs $x1, x2, x3$ are not mandatory and for $x = 0$ their sensitivity curves do not start from the origin. The remaining inputs $x4,...,x8$ are mandatory and their sensitivity curves start at (0,0). The explainability analysis results for this criterion are presented in the top table of Fig. 8. The primary interest of this analysis is the ranking of inputs according to their impact. The general impact of all inputs can be assessed using the impact indicator γ, supplemented by the results of indicators δ^+, ρ and p. The potential for suitability improvement is shown as Q. The same analysis can be performed for intermediate results $X_{13}, X_{45}, X_{68}, X_{48}$ and can be used for tuning weights and degrees of andness.

The coefficient of balance β_i shows whether the i[th] input is balanced with remaining inputs. Therefore, the coefficient of variation V of sequence $\beta_1,...,\beta_n$ is a useful indicator of the *imbalance of input attributes*, and its value should be low.

The first table in Fig. 8 explains the basic properties of the aggregation structure independently of the evaluated object. In such a case we can verify the acceptability of degrees of impact and the improvement potential of suitability attributes.

The second table in Fig. 8 corresponds to object B whose sensitivity curves are shown in Fig. 7, and the third table corresponds to object A that outperforms the object B. Now we can provide a complete explanation of the LSP evaluation results, including low and high supporters, individual contributions of attributes, balance of attributes, ranking of contributors, potential for suitability improvement, impact of all input attributes, concordance values of attributes, and ranking of attributes according to the main criteria. The presented results show that B is an object with moderate overall suitability (73%) and significant imbalance of attributes (the coefficient of variation of β indicators above 30%). The most important contributor is x8, which is also the input with the highest impact. The worst contributor of object B is x6. Therefore, if we want to improve the suitability of B, we must start by improving x6.

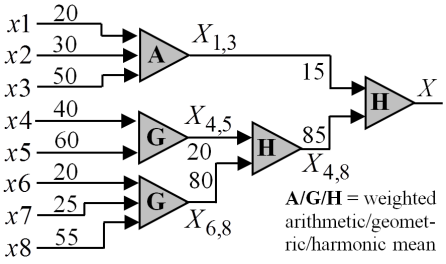

Figure 5. Sample conjunctive suitability aggregation structure

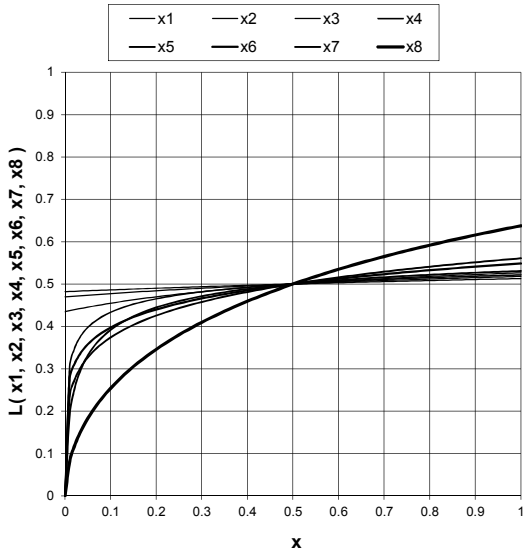

Figure 6. Sensitivity curves for fixed test values of input suitability (0.5)

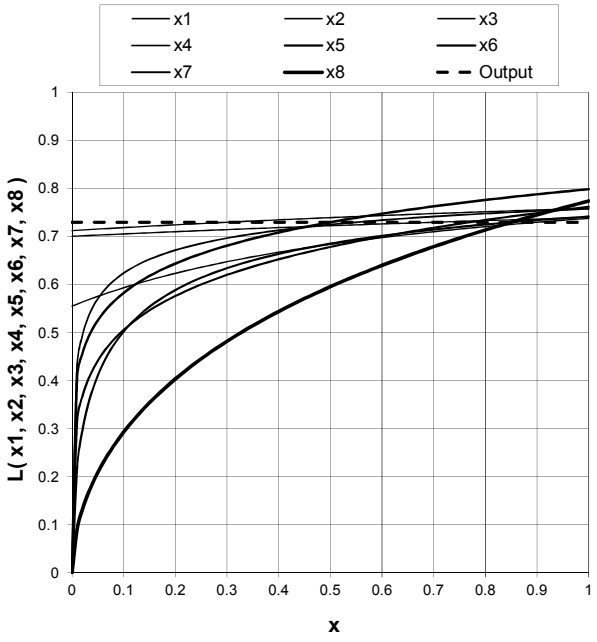

Figure 7. Sensitivity curves for given values of input suitability (object B)

```
EXPLAINABILITY ANALYSIS (all values expressed as [%])
-----------------------------------------------------
Object :  Test
Inputs :    x1    x2    x3    x4    x5    x6    x7    x8
Values :  50.0  50.0  50.0  50.0  50.0  50.0  50.0  50.0
Delta+ :   1.3   1.8   2.6   2.1   3.1   4.8   6.1  13.7
Delta- :   1.8   3.0   6.5  50.0  50.0  50.0  50.0  50.0
Ro     :   3.1   4.8   9.2  52.1  53.1  54.8  56.1  63.7
Q      :   2.6   3.6   5.3   4.3   6.1   9.7  12.1  27.5
Power  :   2.6   3.6   5.3   4.3   6.1   9.7  12.1  27.5
Impact :   5.3   7.7  12.3  14.2  21.0  24.4  29.5  54.2
Overall suitability : 50.0 %
-----------------------------------------------------

EXPLAINABILITY ANALYSIS (all values expressed as [%])
-----------------------------------------------------
Object :    B
Inputs :    x1    x2    x3    x4    x5    x6    x7    x8
Values :  70.0  30.0  90.0  55.0  85.0  50.0  77.0  85.0
C      :  73.1  75.0  71.1  74.4  71.6  77.3  72.2  66.6
DELTA  :  -0.1  -2.0   1.8  -1.5   1.3  -4.3   0.8   6.4
Beta   :  95.8  40.0 126.5  73.9 118.7  64.7 106.7 127.6
Delta+ :   1.0   2.9   0.8   2.8   1.2   6.9   3.2   4.4
Delta- :   2.9   1.7  17.5  73.0  73.0  73.0  73.0  73.0
Ro     :   3.9   4.6  18.3  75.8  74.2  79.8  76.2  77.4
Pi     :  75.2  37.8  95.6  96.3  98.4  91.4  95.8  94.3
Q      :   1.3   3.9   1.1   3.8   1.6   9.4   4.4   6.0
Power  :   3.2   4.1   8.1   6.2   8.0  13.8  13.9  29.3
Impact :   4.8   5.4  17.8  14.9  25.9  23.9  29.8  56.8
High supporters          : x3  x5  x7  x8
Low supporters           : x1  x2  x4  x6
Ranking of contributions: x8  x3  x5  x7  x1  x4  x2  x6
Ranking of attr. impact : x8  x7  x5  x6  x3  x4  x2  x1
Ranking of potential Q  : x6  x8  x7  x2  x4  x5  x1  x3
Overall suitability = 73.0%; Attribute imbalance = 31.7%
-----------------------------------------------------

EXPLAINABILITY ANALYSIS (all values expressed as [%])
-----------------------------------------------------
Object :    A
Inputs :    x1    x2    x3    x4    x5    x6    x7    x8
Values :  67.0  70.0  90.0  85.0  83.0  93.0  75.0  80.0
C      :  81.8  81.8  80.5  81.0  81.1  79.6  82.7  82.1
DELTA  :  -0.4  -0.5   0.8   0.3   0.2   1.7  -1.4  -0.8
Beta   :  82.0  85.5 111.8 104.9 102.3 116.9  90.7  97.5
Delta+ :   1.0   1.3   0.7   0.9   1.4   0.8   4.0   6.9
Delta- :   2.5   4.3  13.6  81.3  81.3  81.3  81.3  81.3
Ro     :   3.4   5.5  14.3  82.2  82.7  82.1  85.3  88.2
Pi     :  71.7  76.7  94.8  99.0  98.3  99.0  95.3  92.1
Q      :   1.2   1.6   0.9   1.0   1.8   1.0   5.0   8.5
Power  :   2.9   4.3   7.5   5.7   8.5  11.5  16.1  34.7
Impact :   3.9   5.9  13.4  16.0  24.8  25.2  30.3  57.2
High supporters          : x3  x4  x5  x6
Low supporters           : x1  x2  x7  x8
Ranking of contributions: x6  x3  x4  x5  x1  x2  x8  x7
Ranking of attr. impact : x8  x7  x6  x5  x4  x3  x2  x1
Ranking of potential Q  : x8  x7  x5  x2  x1  x4  x6  x3
Overall suitability = 81.3%; Attribute imbalance = 11.7%
-----------------------------------------------------
```

Figure 8. Results of explainability analysis for objects Test, B, and A

V. Explaining the Object Comparison Results

Comparison of objects can be based on explainability indicators summarized in Table I. A numerical example of comparison of objects A and B is presented in Fig. 8. The fundamental question in this case is to explain the reasons why the object A with overall suitability 81.3% is better than the object B whose overall suitability is 73%.

The first approach to the explanation of comparison of A and B can be based on differential effects of overall suitability obtained by explainable separate evaluation processes, as well as the fact that object A is significantly more balanced ($V = 11.7\%$) than the object B ($V = 31.7\%$). The most important contributor for object A is x6 which was the worst contributor for object B.

The second approach to explanation of reasons for ranking $A \succ B$ is to directly compare discriminant effects of all input attribute suitability scores. Let A and B have the following input and output suitability scores:

Object A: $a_1,...,a_i,...,a_n$; $X_A = L(a_1,...,a_i,...,a_n)$;

Object B: $b_1,...,b_i,...,b_n$; $X_B = L(b_1,...,b_i,...,b_n)$.

The differential effect of the i[th] input can be defined as the following discriminator:

$$D_i = L(a_1,...,a_i,...,a_n) - L(a_1,...,b_i,...,a_n),\ i = 1,...,n\ .$$

The discriminator D_i shows the contribution of the i[th] input to the ranking $A \succ B$. If $D_i > 0$, then a_i positively contributes to $A \succ B$ and if $D_i < 0$ then a_i negatively contributes to $A \succ B$. The case $D_i = 0$ shows no contribution.

The discriminant analysis is presented in Fig. 9 where the row Dis shows discriminators and the row A>B shows the ranking of inputs according to differential effects. The ranking of positive contributors (denoted +x) is x6>x4>x2. The neutral input is x3 and the negative contributors (-x) are x1>x5>x7>x8. Therefore, to improve the suitability of object A we must first allocate funds for improving x8.

TABLE I. Summary of Explainability Indiators

Symbol	Explanation of effects of attribute suitability values
δ^+	*Output suitability increment* caused by increasing input suitability from the current value to the maximum value 1
δ^-	*Output suitability decrement* caused by decreasing input suitability from the current value to the minimum value 0
ρ	*Output suitability range* if an attribute suitability changes in the range from the minimum value 0 to the maximum value 1
π	*Relative position of current output suitability* inside the output suitability range caused by the selected input attribute
Q	*Potential for max suitability incr*ease above the current value as a ratio of the output suitability increment and the output suit.
p	*Contributing power*: the output suitability increase per unit of the input suitability increase of selected input attribute
γ	*Impact*: capability of an attribute suitability to strongly affect the aggregated output suitability (proximity to the max impact)
C	*Concordance value* of an input attribute suitability, equal to the output suitability caused by other suitability attributes
β	*Balance of selected attribute*: the ratio of the actual suitability of selected attribute and the attribute concordance value
V	*Imbalance of attributes*: the coefficient of variation of all attribute balance values computed for the analyzed object
Δ	*Individual suitability contribution* caused by an attribute: the difference of output suitability and attribute concordance value

```
COMPARISON OF COMPETITIVE OBJECTS A and B
-------------------------------------------------------------------------
Object   x1    x2    x3    x4    x5    x6    x7    x8    Suitability
-------------------------------------------------------------------------
  A     67.0  70.0  90.0  85.0  83.0  93.0  75.0  80.0      81.3 %
  B     70.0  30.0  90.0  55.0  85.0  50.0  77.0  85.0      73.0 %
-------------------------------------------------------------------------
 Dis.:  -0.1   2.2   0.0   2.5  -0.2   6.7  -0.4  -1.9
-------------------------------------------------------------------------
 A>B :  +x6   +x4   +x2    x3   -x1   -x5   -x7   -x8
-------------------------------------------------------------------------
```

Figure 9. Computation of major contributors for ranking A > B

VI. A Logic Model of Certainty Estimator

Information generated by humans can be very reliable, but sometimes it can also be rather unreliable reflecting the uncertainty of information provider. A realistic assumption is that the information provider knows the origins of given information and is sufficiently qualified to assess and provide a degree of certainty that characterizes the reliability of given information. That idea was formalized in the concept of Z-numbers, introduced in 2011 by Zadeh [10]. An uncertain variable X can be characterized by a Z-number $Z=(V,R)$ where V denotes a restriction on the values which X can take (a fuzzy membership function), and R (certainty/confidence) is a measure of reliability of V.

The idea of value-certainty pairs is both natural and general and it can be used for assessment of certainty of evaluation results. When a criterion developer creates suitability attribute criteria $x_i = g_i(a_i),\ i=1,...,n$ (Fig. 1), the side effect of the process of development of each criterion is the percept of developer's certainty that the proposed criterion correctly reflects the stakeholder needs. The certainty/confidence percepts can be quantified as $r_i \in]0,1],\ i=1,...,n$ where $r_i = 1$ denotes the highest degree of certainty that the criterion $g_i(a_i)$ is appropriate and fully acceptable. The values $r_i < 1$ denote the presence of uncertainty $u_i = 1 - r_i$. We assume that in all cases $r_i > 0$ because creating criteria without any trace of certainty is obviously meaningless. The assessment of certainty (or its complement, uncertainty) that the evaluation results are correct can be based on $r_1,...,r_n$ and subsequently used as a component of the verbalized explanation process.

If an input argument has a high impact on the overall suitability, then both its suitability and the certainty that the suitability is correct should be simultaneously high. Similarly, if an input argument has low impact on the overall suitability, then the effects of low certainty can have negligible effect on the certainty of the final result. Unsurprisingly, the efforts to reduce uncertainty should be focused on inputs that have high impact.

Since LSP criteria are decision models, both the structure and all parameters of these models reflect limited certainty of criterion developer. Selection of suitability attributes can be incomplete; attribute criteria are inherently imprecise because of piecewise linear approximations, and suitability aggregators can have imprecise andness and/or weights.

The probabilistic reliability analysis described in [2] is one of techniques for assessment of the reliability of evaluation and comparison results. The probabilistic approach shows the stability of evaluation and ranking results as a function of errors in parameter assessment, without any input from the specific LSP criterion developer. However, inspired by the idea of Z-numbers, we can request that the criterion developer provides an estimate of certainty for each input in the evaluation criterion. The certainty of output depends on the certainty of inputs. If an input has a high impact on the value of output, then the certainty of that input has a high impact on the certainty of output.

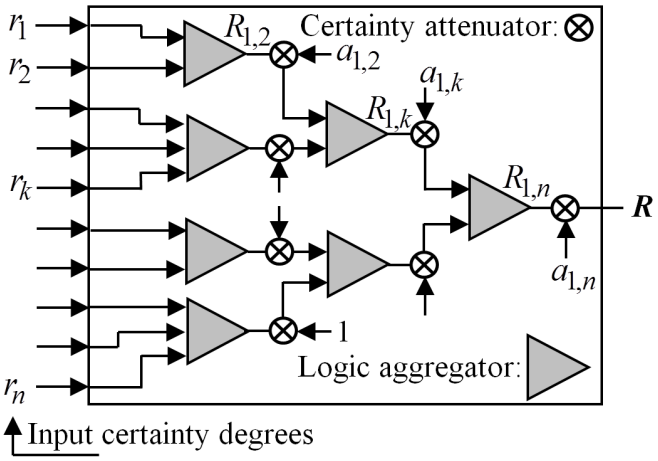

Figure 10. A sample model of certainty estimator

This reasoning yields the conclusion that input certainty degrees should be aggregated using the same logic aggregation structure as the suitability degrees. So, if the overall suitability aggregation model is $X = L(x_1,...,x_n)$ then the overall certainty estimation model should be $R = L(r_1,...,r_n)$. This model assumes that the logic aggregators do not contribute to uncertainty.

In cases where this assumption is not realistic we can use the certainty estimator model shown in Fig. 10. The new component in this model is the certainty attenuator, which is a potentiometer that multiplies each aggregated certainty degree $R_{i,j}$ by an attenuation factor $0 < a_{i,j} \le 1$ that reflects the uncertainty introduced by the selected aggregator whose certainty shadow is $(r_i,...,r_j)$. So, if a logic aggregator has components that decrease the level of certainty, then the output of the aggregator can be decremented $R^-_{i,j} = a_{i,j} R_{i,j}$ and the decremented certainty $R^-_{i,j}$ is then sent to the next aggregation block. In some cases, it can be justifiable to use $a_{i,j} = 1$, i.e. no attenuation. Finally, the overall certainty of the overall suitability is $R = R^-_{1,n} = a_{1,n} R_{1,n}$.

This indicator reflects the certainty of the specific criterion developer(s) and it can be combined with the reliability indicators generated by stochastic reliability models [2]. In the presented logic model of certainty estimator, high certainty of some inputs can compensate lower certainty of other inputs. The stochastic models based on random fluctuations of parameters have a similar property that positive errors can compensate negative errors, and that increases the overall reliability of results.

VII. Verbalizer – a Suitability Explanation Tool

The explanation parameters can be used for creating a tool for the automatic characterization and verbal explanation of evaluation and/or comparison results. The explanation should be relatively short, interactive, and based on explainability parameters introduced in Sections IV and V. An explanation of the overall suitability of a single object A can be automatically generated as the following verbalized report:

"In this report we present and explain the final results of evaluation of object A, as follows: (1) we summarize the final results of evaluation, (2) we present the properties of the evaluation criterion, and (3) we indicate ways to mitigate drawbacks and improve the obtained results.

(1) The overall suitability of object A is X. This result has certainty R and can be interpreted as the percentage of satisfied requirements. It is based on n suitability attributes, grouped in t main groups $G_{(1)},...,G_{(t)}$ *that are used in the final aggregation step for computing the overall suitability. The resulting suitability degrees of these groups are* $X_{(1)},...,X_{(t)}$. *The groups that satisfy* $X_{(i)} > X$ *are called the high supporters, and the groups where* $X_{(i)} < X$ *are called the low supporters of the overall suitability. The ranking of suitability attributes according to the decreasing* ***contribution to the achieved overall suitability*** *is the following:* $x^*_{(1)} : \Delta_{(1)},...,x^*_{(n)} : \Delta_{(n)}$ *(here* $\Delta_{(1)} \geq \cdots \geq \Delta_{(n)}$*).*

(2) The input attributes that are ***high supporters*** *of these results, in the order of increasing support are* $\overline{x}_{(k+1)},...,\overline{x}_{(n)}$ *and the highest support comes from* $\overline{x}_{(n)}$. *The* ***low supporters****, in the order of increasing support are* $\underline{x}_{(1)},...,\underline{x}_{(k)}$ *and the lowest support comes from* $\underline{x}_{(1)}$. *Attributes that must be satisfied reflect mandatory requirements and include the following:* $M = \{\hat{x}_{(1)},...,\hat{x}_{(m)}\}$. *Attributes that are capable to fully satisfy all requirements (sufficient arguments) include the following:* $S = \{\tilde{x}_{(1)},...,\tilde{x}_{(s)}\}$. *The set* $\{x_1,...,x_n\} \setminus M \setminus S$ *contains the remaining attributes that can be interpreted as optional.*

Individual suitability attributes have different role, and importance and produce different impact on the evaluation results. The ranking of suitability attributes according to their decreasing ***overall impact****, expressed as the power to contribute to the overall suitability, is the following:* $x^+_{(1)} : p_{(1)},...,x^+_{(n)} : p_{(n)}$ *(here* $p_{(1)} \geq \cdots \geq p_{(n)}$*).*

(3) The ranking of suitability attributes according to their deceasing ***potential for improvement*** *of the overall suitability is the following:* $x_{(1)} : \delta^+_{(1)},...,x_{(n)} : \delta^+_{(n)}$ *(here* $\delta^+_{(1)} \geq \cdots \geq \delta^+_{(n)}$*). Thus, the efforts to improve suitability should be focused on arguments with the high potential.*

Numerical details supporting the presented explanation summary can be found in explainability analysis tables."

In the case of comparison of multiple objects, the explainability analysis should be provided individually for each object. If the overall suitability of each object is sufficiently justified, then the comparison and selection of the best alternative follows naturally and automatically. We can also add the comparison based on discriminators: *"The object A outperforms the object B based on discriminators* $x_{(1)} : D_{(1)},...,x_{(n)} : D_{(n)}$. *The improvement of object A must follow the increasing sequence of discriminators, starting from the attribute that has the smallest discriminator value".*

The presented verbalized report has a general form and it summarizes and explains the main reasons for obtained results. It can be automatically generated for any object. In some cases this report can be abbreviated or modified, particularly in cases of very large values of *n*, where the ranking lists can be reduced to fixed number of leading components. The presented explainability analysis is focused on the overall suitability, i.e. on the root of the attribute and aggregation trees. In an interactive version of verbalizer, the user can select to perform the explainability analysis for any node of the aggregation tree, moving in the direction of more detailed explanation of the obtained results.

VIII. CONCLUSIONS

The credibility of decision methods critically depends on the interpretability, explainability, and reliability of their results. For each result, stakeholders are entitled and interested to ask for a justifiable explanation provided in a form of short, simple, and easily understandable narrative. To make such explanations provable and automatically generated by verbalizers (software tools), we introduced explainability indicators that are necessary for supporting verbal explanations in a precisely defined quantitative way.

The LSP method uses a regular and interpretable tree structure of aggregation nodes, where each node has clearly defined semantic identity. The explanation methodology introduced in this paper is primarily applied to the most important node – the root of the aggregation tree. However, the same approach can be applied to each end every node on the path from the root to the leaves of the suitability aggregation tree. That provides a strong and natural interpretability and explainability for the LSP method.

REFERENCES

[1] J. Dujmović, Weighted compensative logic with adjustable threshold andness and orness. IEEE Transactions on Fuzzy Systems 2015;23(2), (April): 270–290.

[2] J. Dujmović, *Soft Computing Evaluation Logic*. J. Wiley and IEEE Press, 2018.

[3] V. Torra, and Y. Narukawa. *Modeling decisions*. Berlin: Springer; 2007.

[4] J. Fodor and M. Roubens, *Fuzzy preference modelling and multicriteria decision support*. Dordrecht, The Netherlands: Kluwer Academic Publishers, 1994.

[5] G. Beliakov, H. Bustince Sola, and T. Calvo Sanchez. *A practical guide to averaging functions*. Studies in Fuzziness and Soft Computing 329. New York: Springer; 2016.

[6] M. Grabisch, J-L. Marichal, R. Mesiar, and E. Pap. *Aggregation functions*. Cambridge: Cambridge University Press, 2009.

[7] P.S. Bullen, Handbook of means and their inequalities. The Netherlands: Kluwer; 2003 (and 2010).

[8] J. Dujmović, Graded Logic Aggregation. Proceedings of the 15th International Conference on Modelling Decisions for Artificial Intelligence, LNAI 11144, pp. 3-12, Springer 2018.

[9] C. Rudin, Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. Manuscript based on C. Rudin, Please Stop Explaining Black Box Machine Learning Models for High Stakes Decisions. In: Proceedings of NeurIPS 2018 Workshop on Critiquing and Correcting Trends in Learning; 2018.

[10] L. A. Zadeh, A Note on Z-numbers. Information Sciences 181, Issue 14, pp. 2923-2932, 2011.

[11] J. Dujmović, Preferential Neural Networks. Chapter 7 in *Neural Networks - Concepts, Applications, and Implementations,* Vol. II. Edited by P. Antognetti and V. Milutinović, Prentice-Hall Advanced Reference Series, Prentice-Hall, pp. 155-206, 1991.