Sugeno integrals in subjective mental workload evaluation
Application to flying personnel data

Éric Raufaste
CNRS, UTM, 31058 Toulouse Cedex
raufaste@univ-tlse2.fr

Henri Prade
IRIT-CNRS, UPS, 31062 Toulouse Cedex
prade@irit.fr

Abstract

Classical methods for subjective mental workload evaluation are based on multiple criteria approaches and use weighted sum aggregations. This paper proposes a new method, based on Sugeno integral. This setting has the advantage to be more qualitative and enables us to detect different aggregation policies. The paper presents an elicitation method for Sugeno fuzzy integrals and provides an illustration on flying personnel data.

Keywords: Multiple criteria aggregation, Sugeno integral, Mental workload assessment.

1 Introduction

Measuring the mental workload associated with various situations is a very important topic in cognitive ergonomics and human factors. Of two work environments to be compared (e.g., two human-computer interfaces), the one inducing a lower level of subjective mental workload is generally preferred by ergonomists. However, measuring mental workload is not a trivial task since subjective workload is generally defined as a multidimensional construct. Among the most widely used methods are the National Aeronautics and Space Administration-Task Load Index (NASA-TLX, Hart and Staveland, 1988) and the Subjective Workload Assessment Technique (SWAT, Reid and Nygren, 1988). Contrary to the Cooper-Harper scale (Cooper and Harper, 1969) where operators are asked a single workload estimate, the NASA-TLX and SWAT both assume that workload is a multidimensional concept, with six and three workload sources respectively. Supposedly, a multidimensional approach provides a richer and less biased picture of workload. In turn, a multidimensional approach makes difficult to directly compare various settings (work situations, interfaces, and so on).

Subjective workload assessment is viewed here as a practical example of a multiple criteria aggregation problem where each criterion corresponds to one of the mental workload sources. The data for each criterion / dimension being provided by participants by means of a specific rating scale, such data are clearly subjective in nature. Arguably, they have a qualitative flavor. This has motivated a few approaches extending NASA-TLX to weighted sum of fuzzy scores (Chen, 1996; Liou and Wang, 1994). In this paper we follow another road based on a qualitative aggregation setting provided by Sugeno integrals (Sugeno, 1974; Sugeno 1977).

2 The NASA-TLX

The NASA-TLX rating procedure provides an overall workload score based on a weighted average of the ratings on six subscales: Mental Demands, Physical Demands, Temporal Demands, Performance, Effort, and Frustration. Depending on situations, the various sources may differently contribute to the op-
erator’s subjective workload. Taking into account the relative weights of the sources first requires obtaining a measure of their relative importance. For example, during the standard NASA-TLX procedure participants provide the 15 possible pairwise comparisons of the six subscales. In each comparison, subjects select the source that contributed to the workload more than the other. Each source receives one point for each comparison where it was deemed to contribute more. The relative weight of a source is then given by the sum of those points, divided by 15 for normalization purposes. In order to avoid confusion, in this paper we will call ”rating” the value provided for each workload source, and ”weight” the relative importance of that source.

After information about ratings and weights is collected, the question is to choose the aggregation method. The NASA-TLX makes use of a classical weighted mean, which simply sums the products of ratings by their normalized weights (Σw_i = 1). Thus, noting a_i the rating about the i^th source and w_i the relative importance of the same source, the subjective workload SW in the NASA-TLX method is provided by

\[ SW = \sum_{i=1}^{6} w_i a_i \]

where w_i and a_i respectively denote the weight and rating associated with the i^th workload source.

The weighted average is easy to compute and familiar to most users. On the other hand, despite its apparent simplicity, it is built upon several strong mathematical assumptions that are not necessarily verified in workload assessment. For example, it requires that weights do not depend on ratings. This condition could be attained, for example, by having operators providing the ratings and external experts providing the weights independently. Unfortunately, in the standard NASA-TLX procedure, each operator provides both the ratings and weights. Second, weighted average does not allow taking into account interactions between sources. Is it a reasonable choice to neglect dependencies and interactions between workload sources? By neglecting such interaction effects, a weighted average model might induce measurement biases.

The question addressed here is: is it possible to find an aggregation scale that does not require the elicitation of both ratings and weights by the same subjects, nor demand external expertise and still differentiates between aggregation policies? Moreover, one could want to process subjective data in a qualitative fashion since subjective rating scales are not proved to possess full numerical properties.

The paper extends previous work (Raufaste et al. 2001) suggesting that the Sugeno integral provides a potential solution to those problems.

3 Sugeno integrals: background and elicitation

3.1 Definition

Sugeno (1974, 1977) introduced fuzzy measures and fuzzy integrals, and advocated the latter as a general multiple criteria aggregation device. Let us first restate the necessary background in a finite setting. Let N = \{1,...,n\} be the index set of criteria.

A fuzzy measure is a set function from 2^N to [0,1] such that

i) \( \mu(\emptyset) = 0 \);

ii) \( \mu(N) = 1 \);

iii) if \( A \subseteq B \subseteq N \) then \( \mu(A) \leq \mu(B) \).

A fuzzy integral associated with a function f from N to [0,1] over a non-fuzzy subset A \subseteq N is defined as

\[ S_\mu(f, A) = \sup_{\alpha \in [0,1]} \min(\alpha, \mu(A \cap F_\alpha)) \]

where \( F_\alpha = \{i : f(i) \geq \alpha\} \). In the following \( A = N \) and we shall write \( S_\mu(f) \) for short. A fuzzy integral can be put under the form of
the median of $2n - 1$ terms, as pointed out by Kandel and Byatt (1978). Namely

$$S_\mu(f) = \max_{i=1,n} \min(f(\sigma(i)), \mu(F_{\sigma(i)}))$$

with $F_{\sigma(i)} = \{\sigma(i), \sigma(i+1), ..., \sigma(n)\}$, provided that $N$ has been reordered through the permutation $\sigma$ such that $f(\sigma(1)) \leq ... \leq f(\sigma(n))$. Observe that $F_{\sigma(1)} = N$, $\mu(F_{\sigma(1)}) = 1$, and more generally $\mu(F_{\sigma(i)}) \leq \mu(F_{\sigma(j)})$ if $j \leq i$ since then $F_{\sigma(i)} \subseteq F_{\sigma(j)}$. Then, $S_\mu(f)$ is the median of the set of $2n - 1$ terms, namely $S_\mu(f) = \text{median}\{f(\sigma(i)) : i = 1, n\} \cup \{\mu(F_{\sigma(i)} : i = 2, n\}$. 

If $\mu$ is a possibility measure $\Pi$ (Zadeh, 1978), then it can be shown that

$$S_\Pi(f) = \max_{i=1,n} \min(f(i), \pi(i)),$$

which is the expression of a weighted maximum aggregation of the $f(i)$’s where the $\pi(i)$’s are degrees of importance.

Similarly, the expression

$$\min_{i=1,n} \max(f(i), 1 - \pi(i))$$

of the weighted minimum of the $f(i)$’s is another example of Sugeno integral (based on a necessity measure).

Applying $S$ to multiple-criteria aggregation, the $f(i)$’s correspond to the criteria values and $\mu(F_{\sigma(i)})$ is the degree of importance of the coalition of criteria $F_{\sigma(i)}$. Note that this understanding agrees with the monotonicity of fuzzy measures w. r. t. set inclusion (the larger the subset of criteria, the more important it is globally). See Dubois et al. (2001) for details.

### 3.2 Elicitation

The above multicriteria decision model can be readily applied to workload assessment, and more generally to any subjective evaluation problem. It could be used instead of the NASA-TLX, taking the set of ratings used in the NASA-TLX as the set $N$ of criteria. Then provided $\mu$ is known, one can calculate the overall score (subjective workload $SW$) of any vector $f$ of scores on criteria.

The problem that remains to be solved is the identification of $\mu$. Contrarily to the NASA-TLX, these coefficients are not elicited from the subjects (this would be too difficult, and most probably meaningless).

The set $N$ of criteria corresponds to the set of 6 ratings $(f_1, ..., f_6)$ used in the NASA-TLX. In the NASA-TLX approach, the subjective workload $SW(f)$ associated with a vector $f$ of ratings $(f_1, ..., f_6)$ is assumed to be a weighted sum of the $f_i$’s, where the weights can be elicited from the subjects. Here we assume that for a set of subjects, there exists a Sugeno integral $S$ such that

$$SW(f) = S_\mu(f),$$

where the measure $\mu$ is unknown and to be determined. Thus, a priori, $\mu$ may be the same for all the subjects doing the same task. When it is not the case, we shall see how the elicitation method can detect it.

Here $n = 6$. There are $2^6 = 64$ subsets of $N$ for which the measure $\mu$ has to be determined, in fact $2^n - 2 = 62$, since $\mu(\emptyset) = 0$ and $\mu(N) = 1$. As already said, $S_\mu(f)$ is a median, here the median of $2n - 1 = 11$ terms, namely $f_1, ..., f_6$, and 5 other unknown terms among the 62 values of $\mu$ to be determined. The procedure that we describe works on tuples of the form $(f_1, ..., f_6, SW(f))$ where $f_1$ to $f_6$ are the six ratings provided by a participant in a given measurement wave, and $SW(f)$ is the global rating, also provided by the participant. So the principle of the elicitation procedure is the following.

1. For each tuple $(f_1, ..., f_6)$, corresponding to a piece of individual recorded data, rank-order the $f_i$’s such that $f_{\sigma(1)} \leq ... \leq f_{\sigma(6)}$.

2. Note that if $SW(f) < f_{\sigma(1)}$ or $f_{\sigma(6)} < SW(f)$, there is no solution since $SW(f)$, which is the median of 11 terms, cannot have 6 terms above or below it. Assume $f_{\sigma(k)} \leq SW(f) \leq f_{\sigma(k+1)}$.
As already said, the \( \mu(F_{\sigma(i)}) \)'s are such that 
\[
\mu(F_{\sigma(6)}) \leq \mu(F_{\sigma(5)}) \leq \mu(F_{\sigma(4)}) \leq \mu(F_{\sigma(3)}) \leq \mu(F_{\sigma(2)}), \quad \text{with} \quad F_{\sigma(i)} = \{\sigma(i), \sigma(i + 1), \ldots, \sigma(n)\}. 
\]
Then the \( 5 - k \) smallest terms of the \( \mu(F_{\sigma(i)}) \)'s are below \( SW(f) \) since together with \( f_{\sigma(1)}, \ldots, f_{\sigma(k)} \) it will make \( 5 \) terms below. The other terms are above. This clearly induces constraints on the involved \( \mu(F_{\sigma(i)}) \)'s.

3. These constraints are progressively updated when considering each new tuple of data. If an inconsistency is found, it is processed as a potentially new aggregation strategy, corresponding to another fuzzy measure \( \mu \), also to be determined. However, the inconsistency may also be due to measurement error, which is expected to appear in the end as an isolated fuzzy measure, i.e., applying to a very small subset of observations. Besides, the approach discriminates between different classes of aggregation behaviors.

NB. The detection of inconsistencies is based on the comparison of numerical values obtained experimentally. Thus, only sufficiently meaningful differences should be taken into account.

4 Application to mental workload data

4.1 Method

4.1.1 The situation under study

The data used in the present paper were collected during a series of five rotations of planes pertaining to a big European airline company. Overall, the rotations covered 48 flights. Three types of planes were used (Airbus A319, A320, and A321). Each rotation covered three days.

4.1.2 Participants

twenty-two participants of flying personnel, onboard of the planes, participated to the study. They were either stewards / stewardesses or cabin chiefs.

4.1.3 Data collection

All participants responded to a subjective mental workload assessment questionnaire once in each phase of the flight: preparation, taking off, cruise, and landing. Overall, 840 observations were collected.

4.2 Implementation

We implemented the elicitation method described above in a computer program. The algorithm introduces the observations so as to first select the least informative one. Indeed, introducing an informative constraint too early may prevent the introduction of a subsequent piece of data, which would have been compatible otherwise. Thus, the algorithm computes the amount of information potentially brought by each observation in terms of reduction in interval size. For example, if before introducing the observation we know that 
\[
\mu(F_{\sigma(i)}) \in [a, b], \quad \text{and after its introduction we know that} \quad \mu(F_{\sigma(i)}) \in [a + \epsilon, b], 
\]
the amount of information brought by the observation is \( \epsilon \). Indeed, there is a monotony constraint among coalitions: if \( j < i \), \( \mu(F_{\sigma(i)}) \leq \mu(F_{\sigma(j)}) \). Thus, the information brought by an observation to the lower (resp. upper) bound of the measure of a coalition can be propagated to the lower (resp. upper) bound of the measure of superordinate (resp. subordinate) coalitions. For a given observation, the amount of information is summed over all coalitions. Finally the observation with the lowest sum is actually introduced.

As observations are introduced, the fuzzy measure under processing is more and more precise, and therefore becomes compatible with less and less potential new observations. When no new available observation is compatible with the fuzzy measure under processing, a new fuzzy measure is started. In the end, each observation is assigned to one and only one fuzzy measure. In fact, for each set of consistent observations, we only obtain a bracketing of the parameters of the corresponding fuzzy measure. Thus, it becomes possible to compute two Sugeno integrals for any given observation, respectively based on the
min and max boundaries of the fuzzy measure that was assigned to this observation.

4.3 Results

Overall Quality of the aggregation. The Spearman rank correlation between global estimates and Sugeno integrals based on these estimates were \( \rho(840) = .721, p < .001 \) and \( \rho(840) = .719, p < .001 \) for the integrals based on the lower and upper boundaries respectively. Also the nonparametric Wilcoxon test showed no significant difference between between global scores and Sugeno Integrals. As a consequence, and given the high level of noise commonly observed in subjective rating data, it can be said that the quality of the aggregation provided by the method reported here was fair.

Classification. The program sorted the observations into 13 classes, i.e., generated 13 fuzzy measures. There were referred to in subsequent analyses as Class #1. The three most frequent classes included 591 (70.4%) observations but some classes contained too few observations and were rejected as outliers. The rejection threshold was set at 4%, i.e., any fuzzy measure that represented less than 4% of observations was taken as measurement noise and then ignored from further analyses. Finally, eight fuzzy measures were included in subsequent analyses, respectively class #1 to #5, #7, #8, and #12. They represented 97.6% (\( N = 820 \)) of all observations.

Individual differences. With a contingency coefficient value of .528 (\( p < .001 \)), it can be said that the classification differentially affected participants. In other words, the eight aggregation policies that have been retained were differentially distributed among participants. For example, the class #1 represented 36.2% of the observations but the actual rate for given participant varied from 10.0% to 80.0%. Thus, some participants seemed to be quite homogeneous in their aggregation policy whereas other participants varied more.

Flight phases influence. The moment of the rating had an influence on the aggregation policy, \( \chi^2(21, 820) = 52.8; p < .001 \). Since there was 4 flight phases, one could expect 25% of the observations of each class in each phase. In contrast, 40.7% of the class #3 and 41.0% of the class #3 were related to the sole taking-off phase. Only 12.1% of same class observations were related to the cruise phase, which collected 44.1% of the class #5.

Season and mental workload. Interestingly, our results suggest that the season of the flight had an influence on the aggregation policy. Two seasons were contrasted (summer and winter). 61.5% of observations were made in summer vs. 38.5% in winter and therefore, one could have expect to find those base-rate within each class. As one can check in Table 1 below, some classes differ substantially from this pattern, \( \chi^2(7, 820) = 63.1; p < .001 \). This result may appear surprising at first sight but it is also a common subjective experience that people feel themselves differently as a function of the season. Hence, it is a good point for the classification method proposed here that it be able to capture this subjective pattern.

Day within the rotation. Rotations were accomplished over three days. Supposedly, fatigue may accumulate over the days. As a consequence, the pattern of mental workload could be affected. This phenomenon is detected by our classification of aggregation policies: The moment of the rating had an influence on the aggregation policy, \( \chi^2(14, 820) = 25.3; p = .032 \).

5 Discussion - Conclusion

The paper has provided a preliminary study of the potentials of Sugeno integral as a qualitative aggregation operation to be used for describing the assessment of subjective mental workload. By contrast with classical methods based on weighted sums where the weights

\(^1\)Note for non-experimentalist readers: the \( \chi^2 \) statistics are reported under the form \( \chi^2(df, N) \) where \( df \) (Degrees of freedom) is the product of the number of columns - 1 by the number of rows - 1 in the contingency table. \( N \) is the total number of observations.
Table 1: Percentages of observations falling in each aggregation class as a function of the season of the flight.

<table>
<thead>
<tr>
<th>Class</th>
<th>Summer</th>
<th>Winter</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>147</td>
<td>149</td>
<td>296</td>
</tr>
<tr>
<td>% in Class 1</td>
<td>49.7%</td>
<td>50.3%</td>
<td>100%</td>
</tr>
<tr>
<td>N</td>
<td>120</td>
<td>83</td>
<td>203</td>
</tr>
<tr>
<td>% in Class 2</td>
<td>59.1%</td>
<td>40.9%</td>
<td>100%</td>
</tr>
<tr>
<td>N</td>
<td>66</td>
<td>25</td>
<td>91</td>
</tr>
<tr>
<td>% in Class 3</td>
<td>72.5%</td>
<td>27.5%</td>
<td>100%</td>
</tr>
<tr>
<td>N</td>
<td>24</td>
<td>21</td>
<td>45</td>
</tr>
<tr>
<td>% in Class 4</td>
<td>53.3%</td>
<td>46.7%</td>
<td>100%</td>
</tr>
<tr>
<td>N</td>
<td>22</td>
<td>12</td>
<td>34</td>
</tr>
<tr>
<td>% in Class 5</td>
<td>64.7%</td>
<td>35.3%</td>
<td>100%</td>
</tr>
<tr>
<td>N</td>
<td>55</td>
<td>4</td>
<td>59</td>
</tr>
<tr>
<td>% in Class 7</td>
<td>93.2%</td>
<td>6.8%</td>
<td>100%</td>
</tr>
<tr>
<td>N</td>
<td>25</td>
<td>14</td>
<td>39</td>
</tr>
<tr>
<td>% in Class 8</td>
<td>64.1%</td>
<td>35.9%</td>
<td>100%</td>
</tr>
<tr>
<td>N</td>
<td>45</td>
<td>7</td>
<td>52</td>
</tr>
<tr>
<td>% in Class 12</td>
<td>86.5%</td>
<td>13.5%</td>
<td>100%</td>
</tr>
<tr>
<td>N</td>
<td>45</td>
<td>7</td>
<td>52</td>
</tr>
<tr>
<td>Total</td>
<td>61.5%</td>
<td>38.5%</td>
<td>100%</td>
</tr>
</tbody>
</table>

are elicited from the subject, the proposed approach is able to identify different aggregation policies among the measurement waves from the data. Those aggregation policies seem to be related to variations in the context of the measurement, which is a desirable property for studies in cognitive ergonomics.

Note that in the proposed procedure, the grouping of the data tuples in consistent subsets remains heuristic in the sense that tuples expressing constraints that are not very restrictive, could be consistent with other consistent subsets of tuples, and even consistent with more than one. The reported results are not affected by this remark. However, determining the complete subset of data that is consistent with a particular pair of minimal/maximal Sugeno fuzzy integrals, should enable a finer analysis of the experimental data.

Further research will focus on a practical comparison with the merits of a Choquet integral-based approach (Grabisch et al. 2006; Raufaste et al. 2001), as a representation tool (Grabisch and Labreuche, 2004).

Another open question is to provide a meaningful interpretation of the aggregation policies. In this regard, it would be interesting to take advantage of an if-then rules-based representation of Sugeno integrals (Greco et al. 2004) for explaining the way the mental workload may depend on the six basic evaluations.

Acknowledgements

This research was granted by the INRS, contract #5033704 “Évaluation subjective de la charge mentale : développement et mise au point d’une méthode de traitement des données par les ensembles flous”. We particularly thank Daniel Liévin, from INRS.

References


