ON THE INCONSISTENCY OF PAIRWISE COMPARISONS: AN EXPERIMENTAL STUDY

Jiří Mazurek, Radomír Perzina

Abstract*: A problem of the inconsistency of pairwise comparisons is of focal interest in the analytic hierarchy process (AHP), but, up to date, we know only little about how much are real decision makers inconsistent, and whether the number of objects to be compared influences the inconsistency of their judgments. Therefore, the aim of this paper is to experimentally assess how the inconsistency of pairwise comparisons in the AHP framework changes when the number of objects to be compared (alternatives and/or criteria) increases. In our study, the method of a blind experiment was selected: subjects of the study, who were familiar with the AHP, were instructed to pairwise compare from 3 to 7 objects not knowing the true objective of the study. The main result obtained via ANOVA method is that the consistency ratio was not affected by the increasing number of compared objects, the result that might be likely attributed to the apparent redundancy of pairwise comparisons in the AHP which "corrects" inconsistent judgments. Also, it was found that only 3% of pairwise comparison matrices provided by decision makers were fully consistent, while for 36% of pairwise comparison matrices the consistency ratio CR exceeded the threshold of 0.10.*

Introduction

Pairwise comparisons as a tool for a decision making or a measurement were already considered in the works of Franciscan tertiary Ramon Llul, see (Llul, 1275) or Marquis de Condorcet, see (Condorcet, 1785). The theory of pairwise comparisons was provided for the first time by L. L. Thurstone in 1927, see (Thurstone, 1927).

Since the early 1980s, the pairwise comparisons became the central point of the analytic hierarchy process (AHP) and the analytic network process (ANP) introduced by T. L. Saaty along with his fundamental scale for pairwise comparisons ranging from 1 to 9 (Saaty, 1980 and 1989). AHP/ANP proved to be a useful tool in many areas of human action where a multiple criteria decision making is involved, such as economics, management and marketing, construction, medicine, politics, environmental protection, etc. An overview of the AHP/ANP applications can be found e.g. in (Zahedi, 1986), (Vargas, 1991 and 2001) or (Vaidya and Kumar, 2006). The latter paper alone provides a list of more than 150 papers on application of the the analytic hierarchy process.

Pairwise comparisons of more than two objects give a rise to the problem of inconsistency of these comparisons. If, for example, comparing objects A, B and C, an

expert may say that A is two times better than B, and B is three times better than C. Then, A should be exactly six times better than C by a transitive property. In such a case pairwise comparisons are considered *consistent*. Any value different from six would mean *inconsistency*.

To measure inconsistency of pairwise comparisons, (Saaty, 1980, 1989, 2004 and 2008) proposed to use his consistency index (*CI*) and consistency ratio (*CR*) (see hereinafter). Later, many other indices were proposed, see e.g. (Koczkodaj, 1993), (Alonso and Lamata, 2006), (Brunelli and Fedrizzi, 2015), (Koczkodaj and Szybowski, 2016) or (Mazurek, 2016). Recently, various comparative studies on inconsistency indices emerged, see e.g. (Brunelli et al., 2013), (Kazibudzki, 2016) or (Mazurek, 2016). In the last years, the study on inconsistency of pairwise comparisons focused mainly on the problem of axiomatic properties of inconsistency indices in general, see e.g. (Koczkodaj and Szwarc, 2014), (Brunelli and Fedrizzi, 2015) or (Mazurek, 2016)

However, there is a problem dealing with the pairwise comparisons which was not studied so far: is inconsistency growing when the number of compared objects, for example alternatives or criteria, is getting larger? Is it simpler (or more common) for a decision maker to be consistent when comparing only a small number (3 or 4) of objects rather than a large number (5 and more)?

Therefore, the aim of this paper is to experimentally assess how the inconsistency of pairwise comparisons changes when the number of objects to be compared increases. In our study, the method of the blind experiment was selected: subjects of the study were instructed to pairwise compare from 3 to 7 objects not knowing the true objective was to infer the consistency index and the consistency ratio from their evaluations (both values were inaccessible during the experiment).

The paper is organized as follows: in section 1 the analytic hierarchy process and pairwise comparisons are briefly described, in section 2 the experiment setting and its results are provided, and conclusions close the article.

1 The analytic hierarchy process, pairwise comparisons and inconsistency

In the AHP, objects are organized into a hierarchy with a goal on the top, and criteria and alternatives in the following lower levels. All objects from the same level are considered independent, and are pairwise compared with regard to a superior element of a hierarchy on the so called Saaty's fundamental scale from $\{1/9, 1/8, ..., 1, ..., 8, 9\}$, where $s_{ij} \in \{1/9, 1/8, ..., 1, ..., 8, 9\}$ expresses relative importance of an object *i* with respect to an object *j*, see Table 1.

Tab. 1: Saaty's fundamental scale

Intensity of importance	Definition
1	Equal importance
2	Weak or slight
3	Moderate importance
4	Moderate plus
5	Strong importance
6	Strong plus
7	Very strong importance
8	Very, very strong importance
9	Extreme importance

Source: (Saaty, 2004).

In the AHP it is assumed the pairwise comparisons s_{ij} are *reciprocal*:

$$s_{ij} = \frac{1}{s_{ji}}, s_{ii} = 1, \forall i, j .\qquad (1)$$

A reciprocal matrix S (s_{ij}), the so called *pairwise comparison matrix*, is formed:

$$\begin{pmatrix} 1 & s_{12} & ... & s_{1n} \\ s_{21} & 1 & ... & s_{2n} \\ ... & ... & 1 & ... \\ s_{n1} & s_{n2} & ... & 1 \end{pmatrix}$$

Then, weights w of all objects (criteria and alternatives) are usually determined by Saaty's eigenvalue method as the principal right eigenvector w of the matrix S (Saaty, 1980, 1989, 2004 and 2008):

$$Sw = \lambda_{\max} w,\qquad (2)$$

Also, several other methods for deriving weights from a pairwise comparison matrix were proposed, such as the geometric mean method (the least logarithmic squares method), which gives the same vector of weights in the case of $n = 3$, but slightly different vector for the larger n.

The aggregation of preferences proceeds as follows: let the weight of the i-th criterion be v_i and the weight of the j-th alternative with respect to a criterion f_i be $w_j(f_i)$, then the overall weight E_j of j-th alternative (in a three level hierarchy) is:

$$E_j = \sum_{i=1}^{m} v_i \cdot w_j (f_i),\qquad (3)$$

where $j=1, 2, ..., m$.

At the end, all alternatives are ranked according to their value of E_j.

Pairwise comparisons are *consistent*, if and only if the following condition is satisfied:

$$s_{ij} \cdot s_{jk} = s_{ik}; \forall i, j, k\qquad (4)$$

However, decision makers are seldom fully consistent in their judgments, so the following measures of consistency, the consistency index *CI* and the consistency ratio *CR* were proposed (Saaty, 1980,1989 and 2004):

$$CI = \frac{\lambda_{max} - n}{n - 1}, \tag{5}$$

$$CR = \frac{CI}{RI}, \tag{6}$$

where *n* in (5) is the order of a pairwise comparison matrix, and *RI* in (6) is the so called *random consistency index*, its values are provided in Table 2.

It should be mentioned that values in Table 2 were obtained via Monte Carlo simulations (a generation of a large number of random matrices of a given order) by Saaty, but other authors claim to obtain slightly different results, see (Alonso and Lamata, 2006).

According to (Saaty 2004 and 2008), the acceptable degree of inconsistency is if *CR* is smaller or equal to 0.10. If this condition is not met, a decision maker is asked to revise his or her judgments. However, this "rule of thumb" was criticized by some authors, see e.g. (Dyer and Forman, 1992) or (Koczkodaj, 1993).

During recent decades, the AHP was extended to the fuzzy AHP or interval AHP to encompass uncertainty often present in a real-world decision making, see e.g. Buckley (1985), (Demirel et al. 2008), (Ramík and Korvíny, 2010), (Zadnik and Groselj, 2013) or (Ramík, 2016).

For the number *n* of compared alternatives with respect to a given criterion there is $\binom{n}{2} = \frac{n(n-1)}{2}$ pairwise comparisons. Hence, for larger *n* the task of comparisons becomes more tedious and time consuming.

Moreover, according to the pioneer study of (Miller, 1956), a human brain is capable of processing only up to 7 pieces ("chunks") of information at the same time. This indicates that the more alternatives are compared, the more inconsistent these comparisons will be. Nevertheless, the proof for this claim is missing, as there are no studies known to authors investigating the issue.

Tab. 2: Random consistency index (RI)

number of alternatives	1	2	3	4	5	6	7	8
RI	0	0	0.58	0.90	1.12	1.24	1.32	1.41

Source: (Alonso and Lamata, 2006).

2 The experiment

2.1. Experiment setting

The experiment was conducted on 42 students in a role of decision makers, including 32 women and 10 men of the undergraduate course "The decision analysis

for managers" who were familiar with the AHP. The experiment took place in a computer classroom.

Decision makers were instructed by a teacher to pairwise compare from 3 to 7 alternatives of a fictional problem not knowing the true objective of their task: the evaluation of consistency of their pairwise comparisons (the values of *CI* and *CR* were not available to the students during the experiment).

Each decision maker acted independently and utilized DAME (*Decision Analysis Module for Excel*) software, which is the free Excel built-in module substitution of commercial products for the AHP. DAME offers two-language environment (Czech and English), and it is constructed for the 3-level hierarchy: goal-criteria-alternatives. The use of DAME is described in a more detail in (Ramík and Perzina, 2014).

To conclude, each and every decision maker (a student) provided one 3×3 pairwise comparison matrix, one 4×4 pairwise comparison matrix, and so on, with the 7×7 pairwise comparison matrix being the last.

2.2. Results: descriptive statistics

Every decision maker (DM) pairwise compared from 3 to 7 alternatives, so 210 pairwise comparison matrices were constructed in total.

Only in 5 cases (2.4% of all cases) DMs were fully consistent (their *CI* = 0), in 205 cases (97.6%) they were inconsistent. Furthermore, 134 (64%) pairwise comparisons matrices had the consistency ratio *CR* smaller or equal than 0.10, for the remaining 76 (36%) pairwise comparison matrices the consistency ratio *CR* exceeded the value of 0.10, so, in practice, decision makers who provided these matrices would be asked to revise their judgments.

Average values of *CI* and *CR* of all 42 DMs and for 3-7 alternatives are provided in Table 3. Moreover, Table 3 provides the number of DMs with consistency ratio smaller or equal to 0.10, the inconsistency still acceptable according to Saaty. As can be seen from Table 3, the consistency ratio was rather decreasing with the increasing number of alternatives, and the number of DMs with *CR* up to 0.10 was increasing.

2.3. A correlation between *CR* and the number of alternatives

The correlation between *CR* and the number of alternatives (k) was examined via Pearson's correlation coefficient: r = –0.836, indicating indirect proportionality between *CR* and k.

Statistical significance of the value r was examined via t-test as well:

The null hypothesis H_0: $r = 0$.

The t-statistics $t = \dfrac{|r| \cdot \sqrt{n-3}}{\sqrt{1-r^2}} = 2.64$ (where $n = 5$) was compared with the critical value for $\alpha = 0.05$: $t_{crit.} = 3.18$. The result is that the null hypothesis cannot be rejected.

Hence, the correlation coefficient is not statistically significant at 0.05 level. This outcome was rather expected due to the very small sample of 5 pairs.

2.4. ANOVA analysis of *CR*

To find whether the number of alternatives was a factor behind the changes of the consistency ratio *CR*, one-way ANOVA (analysis of variance) method was employed.

The method divides the variance into two parts, into treatments $S_{y,m}$ and error $S_{y,v}$ parts, and their comparison via F-test with $k-1$ and $n-k$ degrees of freedom, where k denotes the number of values of a given factor (in our case $k=5$), and n denotes the total number of cases ($n=210$).

The null hypothesis H$_0$: *The mean value of CR for all k is equal.*

The F-test with the values $S_{y,v} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \overline{y}_i\right)^2 = 15.42$, $S_{y,m} = \sum_{i=1}^{k} n_i \left(\overline{y}_i - \overline{y}\right)^2 = 0.678$, $n=210$ and $k=5$ yields:

$$F(4,205) = \frac{\dfrac{S_{y,m}}{k-1}}{\dfrac{S_{y,v}}{n-k}} = 2.25. \tag{7}$$

The value of the test criterion (7) is (slightly) lower than the critical value $F_{0.05}(4,205) = 2.37$. Therefore, the null hypothesis that the sample means for 3-7 alternatives are equal cannot be rejected at 0.05 level. In other words, the effect of the number of alternatives on consistency ratio *CR* was not confirmed.

Tab. 3. The results of the experiment: the last three columns give the number of pairwise comparisons matrices satisfying given values of CR.

Number of alternatives	Average *CI*	Average *CR*	*CR* > 0.1	*CR* ≤ 0.10	*CR* = 0
3	0.1376	0.2373	19	23	3
4	0.1085	0.1206	17	25	1
5	0.1075	0.0960	14	28	0
6	0.1270	0.1024	14	28	1
7	0.10033	0.0782	12	30	0

Source: authors.

Conclusions

The aim of this paper was to investigate the inconsistency of pairwise comparisons provided by decision makers for different numbers of compared objects (alternatives) in a blind experiment.

Somewhat surprisingly, the inconsistency measured in terms of the consistency ratio *CR* was not statistically significantly increasing with the growing number of alternatives. By the ANOVA method, the influence of the number of alternatives on *CR* was not found significant at 0.05 level. The correlation between *CR* and the number of alternatives was not statistically significant either.

Perhaps the most surprising result of the study is that out of 210 pairwise comparison matrices provided by decision makers and examined in this study, only 5 were completely consistent. This inconsistency is understandable for 6 or 7 alternatives, when decision makers provided 15 and 21 pairwise comparisons

respectively, the numbers well above the "magical number 7" threshold found in the study of (Miller 1956), who states that humans are only capable of handling up to 7 pieces of information at one time. Otherwise, the so called "cognitive overload" makes it impossible to process information correctly. In this study, however, even in the case of only 3 alternatives, when decision makers carried out just 3 pairwise comparisons in total, astonishing 93% of decision makers were inconsistent. Therefore, this outcome indicates some other explanation of inconsistent judgments is needed. Perhaps, the consistency defined by relation (4) is too strong for a practical use, and might be substituted by a simple transitivity of preferences (if A is better than B, and B is better than C, then A is better than C).

Further research may focus on more than 7 alternatives, and the research sample could be expanded. Also, the examination of other measures of inconsistency (than *CI* and *CR*), such as the inconsistency proposed in (Koczkodaj, 1993), would be useful. Another interesting possibility is to examine whether the inconsistency depends on the gender, which couldn't be examined in this study due to the limited sample of decision makers.

Acknowledgments

References

Alonso, J. A., Lamata, M. T. (2006). Consistency in the analytic hierarchy process: a new approach. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 14(4), pp. 445-459. doi: dx.doi.org/10.1142/S0218488506004114.

Buckley, J. J. (1985). Fuzzy hierarchical analysis, *Fuzzy Sets and Systems*, 17, pp. 233–247.

Condorcet, M. (1785). *Essai sur l'Application de l'Analyse `a la Probabilit´e des D´ecisions, Rendues `a la Pluralit´e des Voix*, Paris.

Brunelli M., Fedrizzi M. (2015). Axiomatic properties of inconsistency indices for pairwise comparisons. *Journal of the Operational Research Society*, 66(1), pp. 1-15. doi:10.1057/jors.2013.135.

Brunelli, M., Canal, L., Fedrizzi, M. (2013). Inconsistency indices for pairwise comparison matrices: a numerical study. *Annals of Operations Research*, 211(1), pp. 493-509

Demirel, T., Demirel, N. C., Kahrman, C. (2008). Fuzzy Analytic Hierarchy Process and Its Application. In: *Fuzzy Multi-Criteria Decision Making*, Vol. 16 of the series Springer Optimization and Its Applications, pp. 53-83

Dyer, R. F., Forman, E. H. (1992). Group decision support with the analytic hierarchy process. *Decision support systems*, 8(2), pp. 99-124.

Kazibudzki, P. T. (2016). An examination of performance relations among selected consistency measures for simulated pairwise judgments. *Annals of Operations Research*, 244, pp. 525–544. doi: 10.1007/s10479-016-2131-6.

Koczkodaj, W. (1993). A new definition of consistency of pairwise comparisons, M*athematical and Computer Modeling*, 18(7), pp. 79–84.

Koczkodaj, W., Szwarc, R. (2014). On Axiomatization of Inconsistency Indicators for Pairwise Comparisons. *Fundamenta Informaticae*, 132(4), pp. 485-500. doi: 10.3233/FI-2014-1055.

Koczkodaj, W., Szybowski, J. (2016). The Limit of Inconsistency Reduction in Pairwise Comparisons. *Int. J. Appl. Math. Computer Science,* 26(3), pp. 721–729. doi: 10.1515/amcs-2016-0050.

Llull, R. (1275). *Artifitium electionis personarum.*

Mazurek, J. (2016). On Inconsistency Indices and Inconsistency Axioms in Pairwise Comparisons. arXiv:1703.05204v2.

Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information, *Psychological Review*, 63(2), pp. 81–97.

Ramík, J. (2016) Ranking Alternatives by Pairwise Comparisons Matrix with Fuzzy Elements on Alo-Group. In: *Czarnowski I., Caballero A., Howlett R., Jain L. (eds) Intelligent Decision Technologies 2016. Smart Innovation, Systems and Technologies*, vol 57. Springer, pp. 371-380.

Ramík, J. Korvíny, P. (2010). Inconsistency of pair-wise comparison matrix with fuzzy elements based on geometric mean. *Internat. J. Fuzzy Sets and Systems,* 161, pp. 1604–1613.

Ramík, J. Perzina, R. (2014). Solving Multicriteria Decision Making Problems using Microsoft Excel. In: *Proceedings of the 32nd International Conference Mathematical Methods in Economics.* Palacky University Olomouc, pp. 697-702.

Saaty, T. L. (1980). *Analytic Hierarchy Process.* New York: McGraw-Hill.

Saaty, T. L. (1989). Group decision making and the AHP. In *Golden et al. (Eds.), The Analytic Hierarchy Process: Applications and Studies*, pp. 59–67.

Saaty, T. L. (2004). Decision making – The analytic hierarchy and network processes (AHP/ANP). *Journal of Systems Science and Systems Engineering*, 13(1), pp. 1–34.

Saaty, T. L. (2008). Decision making with the analytic hierarchy process. *Int. J. Services Sciences,* 1, pp. 83–98. doi: doi.org/10.1504/IJSSci.2008.01759.

Thurstone, L .L. (1927). A law of comparative judgement. *Psychological Review*, 34, pp. 273-286.

Vaidya, O. S., Kumar, S. (2006). Analytic Hierarchy Process: An overview of applications. *European Journal of Operational Research*, 169, pp. 1-29.

Vargas, L. G. (1991*). Prediction, Projection and Forecasting: Applications of the Analytic Hierarchy Process in Economics, Finance, Politics, Games and Sports*, Boston: Kluwer Academic.

Vargas, L.G. (2001). *Models, Methods, Concepts & Applications of the Analytic Hierarchy Process*, Boston: Kluwer Academic.

Zadnik, S. L., Groselj, P. (2013). Estimating Priorities in Group AHP Using Interval Comparison Matrices. *Multiple Criteria Decision Making*, 8, pp. 143-159.

Zahedi, F. (1986). The Analytic Hierarchy Process—A Survey of the Method and its Applications, *Interfaces*, 16(4), pp. 96-108.

Contact address:

Mgr. Jiří Mazurek, Ph.D.
Obchodně-podnikatelská fakulta v Karviné
Univ. náměstí 1934/3
Karviná
73301