Building a benchmarking model

Methodology

AHP

AHP is the shorthand of Analytic Hierarchy Process. When facing a problem of decision making, for example, selection of the best alternative or benchmarking of alternatives, one has to consider many factors, objectives, and criteria with respect to the goal (e.g. cost, available resources, …). Afterwards, based on selected objectives, one has to consider different alternatives and choose the one that seems to be the best with respect to meeting all the requirements. Additionally, there might be multiple decision makers who are addressing the same issue with different points of view on the problem.

For example, consider buying an apartment. One follows objectives such as price, location, number of rooms, etc. The alternatives are then different offers of various apartments. When buying as a family, one incorporates different decision makers, possibly even with different significance for each.

For complex problems, all these factors should be considered before making the final decision. AHP provides a method how to address crucial decision making problems (although it may be used for simple ones as well). It does not solve the problem for us. It helps us determine valuable and important roles and provides us with a tool to mathematically evaluate different aspects of the final goal.

The process consists of three crucial parts: dividing the goal into a hierarchy of smaller sub-objectives, evaluating each alternative with respect to each of the lowest level sub-objectives, and aggregating the results to get the final score. Forming a hierarchy in the first step to decompose the problem is crucial because smaller parts can then be addressed individually and more precisely.

The inputs into the method are, among others, the hierarchy of objectives and the alternatives. As a measure, judgments about preferences made by each decision maker (with a corresponding weight of the decision maker) are used. The output of the process is the final weighting of the criteria and scores of the alternatives. Besides, it provides us with a measure of consistency of individual judgments.

(Forman and Selly 2001) describe the AHP as a model allowing "decision makers to model a complex problem in a hierarchical structure showing the relationships of the goal, objectives, sub-objectives, and alternatives. Uncertainties and other influencing factors can also be included". Furthermore, it "allows for the application of data, experience, insight, and intuition in a logical and thorough way". Lastly, "AHP is a compensatory decision methodology because alternatives that are deficient with respect to one or more objectives can compensate by their performance with respect to other objectives. AHP is composed of several previously existing but unassociated concepts and techniques such as hierarchical structuring of complexity, pairwise comparisons, redundant judgments, an eigenvector method for deriving weights, and consistency considerations”.

In the following text, each of the three crucial parts above is described in detail together with how it is applied in the effort that creates an interactive European Contraception ATLAS which benchmarks country performance in terms of provision of access to contraceptives through national health systems.

Hierarchy

The first step in the AHP is to decompose the problem into different objectives and sub-objectives.

AHP applied in Contraception ATLAS

To apply AHP in the effort that creates an interactive European Contraception ATLAS, we set the following terminology:

- The goal is to assess the “access to contraception”.
- The alternatives are different countries.
- The decision making problem is then to evaluate each country with respect to “access to contraception”.
- The “access to contraception” consists of different objectives and sub-objectives. We call these objectives and sub-objectives the criteria or sub-criteria. For example, the “access to contraception” splits into criteria called “access to online information” and “access to contraceptive supplies”.

Hierarchy tree

With the goal being “access to contraception”, the aim is now to break the goal down into smaller components (criteria). We may interpret this break-down as forming a hierarchy tree in which the root node of the tree is the goal. The components of each node are so called branches that lead to other, subsequent nodes. Each subsequent layer of nodes is called a level of the tree. The last level of the tree is called a leaf level (or leaf nodes or simply leaves). The leaves represent the criteria with respect to which the final evaluation of the alternatives is done (the leaves stand for the lowest-level sub-objectives). A visualization of the hierarchy tree is displayed in Figure 1.

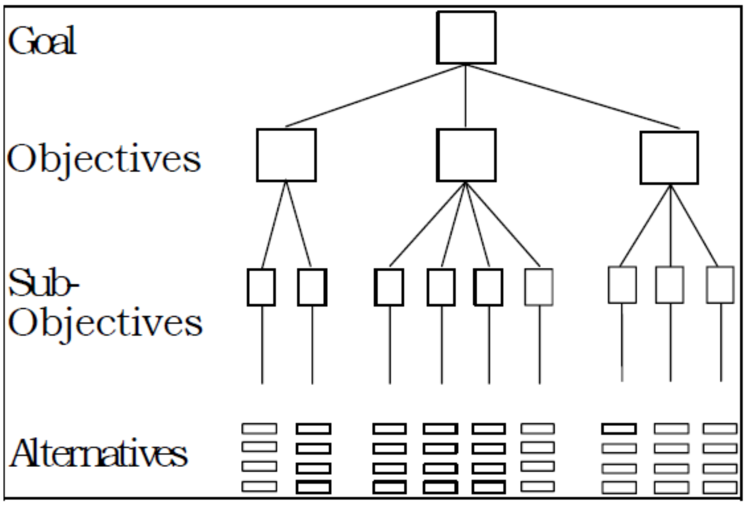

Figure 1: Layout of the hierarchy tree, the goal is the root. See (Forman and Selly 2001)**.**

Following the example above, the root node "access to contraception" splits into two branches leading to two nodes "access to contraceptive supplies" and "access to online information". These two nodes form the second layer (or the second level) of the hierarchy tree. They are also called child nodes of the node "access to contraception". From the other point of view, the node "access to contraception" is the parent of "access to contraceptive supplies" and "access to online information".

To create the hierarchy, one needs to take into account all possible criteria, how much they differ in importance, what are the possible values, etc. The final hierarchy of "access to contraception", according to which these examples are made, is displayed in the results section.

Once the criteria are settled, proper weights need to be derived. These weights express how important each criterion is. There are 2 types of weights: local weights and global weights.

Local weights are weights of nodes that express the importance of the nodes with respect to their parent node. For example, node "access to online information" has 3 child nodes: "type of online information", "information coverage", and "user friendliness". These child nodes are differently important so their weights could be, for example, 50 %, 30 %, and 20 %, respectively (saying that "type of online information" is as important as "information coverage" and "user friendliness" together). These weights express importance with respect to the parent node and are therefore called local weights. Local weights of all children with respect to their parent node must always sum up to 100 %.

Global weights are weights of nodes that express the importance of the nodes with respect to the root node (overall goal). For example, our root node (or the goal) splits into "access to online information" and "access to contraceptive supplies". Since their parent is the root node, the local weights would be the same as the global weights. If we go one step deeper (one more level) in the hierarchy, both "access to online information" and "access to contraceptive supplies" split into 3 child nodes so the root node has 6 child nodes in total on this level. By this relation, they are all related to the root node. In other words, they all contribute to the overall goal and their global weights should sum up to 100 % since we are splitting the root node into these 6 criteria. This process goes further until we reach the leaves, so in general, the global weights at the same level of the tree always need to sum up to 100 %.[1]

As the conclusion:

- Local weights make sense at each node and express the relative importance of the sub-criteria to the parent criterion
- Global weights make sense on the lowest level and express the relative importance of the evaluated criteria to the overall goal

The evaluation (or the benchmarking) of the countries is done according to the leaves so for the calculation one needs their global weights. Since global weights are derived from local weights, one needs to derive local weights firstly. There are many ways how to do this. According to (Forman and Selly 2001) and (Saaty 2008 Vol. 1 No. 1), the most accurate one seems to be using pairwise comparisons.

Pairwise comparisons

The pairwise comparison is a technique to assess relative preference of one criterion over another (with respect to given objective). "Pairwise" means that only two objects are compared. Human mind can accommodate and assess up to seven things only, with the rule the less the better, see (Forman and Selly 2001). Therefore, the pairwise comparisons are believed to suit the best to our minds. One may observe that they are generally used in various questionnaires. The disadvantage is that for *n* objects, the number of comparisons grows with n^2 because the total number of comparisons is

$$\binom{n}{2} = n(n-1)/2.$$

The choice of preferences in the pairwise comparison is described in Figure 2. Note that if one of the criteria in the pairwise comparison is preferred by, say, a factor of 6, then the other is preferred by a factor of 1/6. An example of the pairwise comparison is: How preferred is "access to contraceptive supplies" to "access to online information" with respect to "access to contraception"? The answer

[1] To be mathematically correct, this is true only if each leaf is at the same level, which is our case. If leaves are on different levels, then global weights sum up to 1 from nodes at the same level plus the leaves from the previous levels.

may be that the first is moderately more important than the second, so the first is preferred by a factor of 3 over the second, or the second is preferred by a factor of 1/3 over the first (with respect to "access to contraception").

Since the pairwise comparisons are defined with respect to an objective, they are made only within child nodes. This greatly lowers the total number of combinations. For a node with 3 children, only 3 comparisons are made. From these comparisons, one derives the (local) weights. Afterwards, the global weights are calculated and eventually used for the evaluation of the final goal.

This structured form of pairwise comparisons (structured in a way that the comparisons are assessed only within child nodes in the hierarchy) enables assessing "incomparable" criteria.[2] If one criterion is much more important than another (by more than an order of magnitude), then they should be placed on different levels in the tree because the pairwise comparison does not allow such different preferences. Indeed, when calculating the global weights and, hence, multiplying the weights of subsequent levels, one can get to large distinctions among weights, which may represent even greater dissimilarities than an order of magnitude. Thanks to the hierarchy, these differences are rationalized.

Intensity of Importance	*Definition*	*Explanation*
1	Equal Importance	Two activities contribute equally to the objective
2	Weak or slight	
3	Moderate importance	Experience and judgement slightly favour one activity over another
4	Moderate plus	
5	Strong importance	Experience and judgement strongly favour one activity over another
6	Strong plus	
7	Very strong or demonstrated importance	An activity is favoured very strongly over another; its dominance demonstrated in practice
8	Very, very strong	
9	Extreme importance	The evidence favouring one activity over another is of the highest possible order of affirmation
Reciprocals of above	If activity *i* has one of the above non-zero numbers assigned to it when compared with activity *j*, then *j* has the reciprocal value when compared with *i*	A reasonable assumption
1.1–1.9	If the activities are very close	May be difficult to assign the best value but when compared with other contrasting activities the size of the small numbers would not be too noticeable, yet they can still indicate the relative importance of the activities.

Figure 2: Enumeration of preferences in the pairwise comparison, (Forman and Selly 2001). Scale 1-9 corresponds to the scale 0-4 that was used in the provided application called SCOUT collector.

[2] One of the axioms of AHP is the homogeneity axiom: the elements being compared should not differ by more than an order of magnitude.

Derivation of weights and a consistency index

In this paragraph, we describe how the local weights are derived from the pairwise comparisons. For each pairwise comparison, the decision maker chooses a preference (one number out of 1, 2, … 9 or 1/2, 1/3, … 1/9). If there are n objects to compare, there are $n(n-1)/2$ comparisons. Resulting preferences form so called Saaty's matrix

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix},$$

where $a_{ii} = 1$ (by definition), $i \in \{1, 2, \dots n\}$, and elements a_{ij} ($i \neq j$) express that *i*-th criterion is preferred by a factor of a_{ij} compared to the *j*-th criterion. By the definition of pairwise comparison, it holds that $a_{ji} = 1/a_{ij}$.

Having defined the Saaty's matrix, the eigenvector corresponding to the highest eigenvalue of the matrix expresses the vector of weights, provided it is rescaled to sum up to 1. Additionally, denoting the highest eigenvalue as λ_{max}, the consistency index *CI* is defined as $CI = (\lambda_{max} - n)/(n-1)$, where *n* is the number of rows of the matrix. The consistency ratio $CR = CI/RI_n$ then serves as a measure of consistency of the set of comparisons, where RI_n is so called random index of order *n*. These random indices are tabularized. It is supposed, but not required, that the consistency ratio is below 0.1. If the *CR* is too high, it suggests inconsistent preferences of the evaluator.

Here in our example (node "information coverage"), we have a matrix of 3 rows and 3 columns, each row and each column stands for the criteria "number of contraceptives", "financial information", and "logistical information", respectively. The appropriate matrix could look like this, for instance:

$$\begin{pmatrix} 1 & 2 & 1/3 \\ 1/2 & 1 & 1/5 \\ 3 & 5 & 1 \end{pmatrix}.$$

Here, "number of contraceptives" is preferred by a factor of 2 to "financial information", and so on. The resulting local weights would be approximately 0.23, 0.12, and 0.65 for "number of contraceptives", "financial information", and "logistical information", respectively.

Once all the local weights are derived, the hierarchy tree is used to calculate the appropriate global weights for the criteria. The most important global weights are those of the leaf nodes (or lowest-level nodes), which are used for the evaluation of the alternatives.

Lowest-level criteria

The lowest-level criteria (or the leaf nodes) are used for the evaluation of the alternatives. There are two approaches to that. First approach requires pairwise comparisons among the alternatives with respect to each leaf node. This, however, generates enormous amount of comparisons, especially when there is a large number of alternatives.

The second approach is so called ratings approach (or a scoring approach or an open system approach). In this case, scores are defined for each leaf node. These scores express how an alternative performs with respect to the criterion. The particular score can also be called a measure level. During the evaluation, each alternative receives (by individual decision makers) appropriate scores. The scores are then multiplied with the global weight of the criterion to obtain the total score of the alternative.

However, there are many possible ways how to set up the scores. Moreover, they need to satisfy many conditions. They must represent the whole range of all possible values of the criterion. One must take into account that if one alternative receives a bad score for one criterion, then the difficulty to compensate this loss by doing better in other criteria should be set justly as well (so the scores do not favor only one alternative). In other words, the scores for different criteria should be comparable and compensatory. Furthermore, one distinguishes between qualitative and quantitative criteria. Both are driven by different methods how to calculate the scores. Last but not least, there is a question whether a score allowing zero point evaluation shall be included.

Qualitative versus quantitative criteria

From the mathematical point of view, a qualitative measure is a discrete measure, while quantitative measure is continuous. In the first case, there are levels or exact values of the variable that are usually not described by numbers but rather by words (e.g. high / medium / low). The problem is how to properly transform these scales into numbers, for example, how many times is “high” better than “medium”. What is more, these numbers must reflect the “compensation problem” described above. A general solution to that is, again, to use pairwise comparisons of the individual levels. These comparisons generate weights that can be consequently interpreted as scores of the levels.

In the second case (quantitative criteria), there is a range of (infinite number of) possible values (e.g. interval (0, 10)). Again, there is a problem how to assess the importance of a given value, for example, how much is 8 better than 2 (is it really 4 times better or is it different?). General solution is to split the interval into several sub-intervals and generate the weights using pairwise comparisons of the center points of the sub-intervals. The more intervals we set, the more comparisons, but also the better (more granular) approximation we get. If a continuous approximation is desired, one can interpolate the generated weights by a smooth curve. The type of the interpolating function depends on the nature of weights. Since they basically form some kind of an arithmetic progression, one should choose the interpolating function in accordance with the kind of the progression.

Yes / No criteria

Yes/No criterion is a special version of a qualitative criterion with only 2 possible values. Typically, “Yes” gets 100 points and “No” gets 0 points or vice versa. In the other case, one can use again pairwise comparison or intuitively assign the weights since there are only 2 levels.

Zero level criteria

One of the axioms of the AHP is the reciprocal axiom: if $\square_{\square}(\square_{\square}, \square_{\square})$ is a pairwise comparison of elements *A* and *B* with respect to their parent, element *C*, representing how many times more the element *A* possesses a property than does element *B*, then $?_?(?_?, ?_?) = 1/?_?(?_?, ?_?)$. For example, if objective *A* is 5 times larger than objective *B* with respect to objective *C*, then *B* is one fifth as large as *A* with respect to *C*. Therefore, pairwise comparisons cannot generate zero weight (since division by zero is not defined) and even the worst possible option still receives a non-zero weight (or score). From the application point of view, this can be an issue because often decision makers simply want to rate an object with zero points. This can be overcome as follows: using pairwise comparisons for all the non-zero levels and, subsequently, complementing these levels and their weights with an "artificial" zero-weight level. In this approach, the power of the pairwise comparisons is kept.

Aggressive versus conservative approach

This modification relates to the "compensation problem" described above (it is mainly an issue of measures and not an issue of criteria). When comparing only a few options (namely 2 or 3) using pairwise comparisons, the resulting weights could have large differences, which result in favoring slightly better alternatives too much. According to (Forman and Selly 2001) and (Saaty 2008 Vol. 1 No. 1), the numerical preferences used in pairwise comparisons (see Figure 3) can be altered so the resulting weights depict the measure levels more evenly. Specifically, if we bound the maximum preference to be, for example, 5 instead of 9 (which basically means that the chosen preferences cannot be that extreme), then the derived weights also cannot differ extremely from each other. Hence, it becomes easier for an alternative to compensate a loss in one criterion by doing better in other criteria. This alteration of numerical preferences in pairwise comparisons was used in our "conservative" approach (-es). If there is no alteration, we call it "aggressive".

Putting it altogether

To summarize the thoughts above, for each lowest-level criterion (or leaf node), a qualitative or a quantitative measure with appropriate and fair weights must be defined. The weights should reflect the nature of the criterion as well as the possible values that the alternatives could have.

Now, the weights can be transformed and interpreted as scores by dividing them by the maximum weight (maximum weight within the criterion). Consequently, the new maximum score becomes 1 (or 100 points if multiplied by 100) and can be interpreted as an ideal or serve as a level of reference. This ideal level then receives the whole global weight of the criterion (because the global weight is multiplied by the score, which is 1 in this case). The lower scores now reflect how well they compare to the ideal level. This interpretation is more useful since it is more natural to work with.

Once the scores are set (in the rating approach), each alternative is assigned appropriate score in each of the leaf node. To calculate the total score of an alternative, the global weight of the criterion is multiplied with the score and, consequently, all partial scores are summed up.

Incorporating different decision makers

If the decision making problem is evaluated by multiple experts, each one may have different opinion and, therefore, would match the preferences differently. If all the experts are to be taken into account, a method how to aggregate the preferences shall be incorporated (so called group decision making). According to (Saaty 2008 Vol. 1 No. 1), each expert assigns the preferences on his/her own and each individual's weights are calculated. Then, weighted geometric mean is used to aggregate the results. The weighted version of the geometric mean enables users to enter different importance (or weights) for the decision makers. These weights can, once again, be derived using pairwise comparisons among the decision makers. In context of the effort that creates an interactive European Contraception ATLAS, the weights for the experts are equal, that is, no evaluator or a group of evaluators is preferred to the others. The aggregated results are used as the final weights which fairly reflect the whole group decision making. Note that the aggregated weights should be rescaled to sum up to 1, if necessary.

Bibliography

Forman, Ernest H., and Mary Ann Selly. *Decision by Objectives: How to Convince Others that You are Right.* World Scientific, 2001.

Saaty, Thomas L. "Decision making with the analytic hierarchy process." *International Journal of Services Sciences*, 2008 Vol. 1 No. 1: 83-98.

Results

Hierarchy

Discussions with the experts led to the final hierarchy tree displayed in Figure 3.

```
Access to Contraception
 ¦--Access to online information
 ¦   ¦--Type of online information
 ¦   ¦   °--Type of online information
 ¦   ¦--Information Coverage
 ¦   ¦   ¦--Number of contraceptives
 ¦   ¦   ¦--Financial information
 ¦   ¦   °--Logistical information
 ¦   °--User friendliness
 ¦       ¦--Minority languages in the same website
 ¦       ¦--Visualization or Ease of navigation
 ¦       °--Ease of finding the website
 °--Access to contraceptive supplies
     ¦--Reimbursement
     ¦   ¦--Level of reimbursement
     ¦   ¦--Special reimbursement for minors
     ¦   °--Special reimbursement for vulnerable groups
     ¦--Procedure
     ¦   ¦--Reimbursment of cost of consultation
     ¦   ¦--Need for parental consent
     ¦   °--Availability without disclosure of legal status
     °--Alternatives
         ¦--Availability of non-prescription emergency contraception
         °--Availability of non-prescription non-barrier contraceptives
```

Figure 4: Final hierarchy tree. In this figure, the root is in the top left corner, the leaves are the rightmost nodes. Each leaf is at the 4th level which is convenient for further visualizations of the tree.

In a questionnaire, the experts expressed their preferences, resulting in the following global weights. The aggressive approach corresponds to the maximum preference of a factor of 9, the conservative to the maximum preference of a factor of 5.

Leaves	Aggresive	Conservative
Type of online information	3.34%	6.09%
Number of contraceptives	7.28%	7.31%
Financial information	12.13%	10.29%
Logistical information	4.52%	5.30%
Minority languages in the same website	0.53%	1.24%
Visualization or Ease of navigation	2.21%	3.44%
Ease of finding the website	5.34%	6.17%
Level of reimbursement	11.66%	9.43%
Special reimbursement for minors	6.61%	6.40%
Special reimbursement for vulnerable groups	5.87%	5.89%
Reimbursment of cost of consultation	5.83%	5.77%
Need for parental consent	6.90%	6.54%
Availability without disclosure of legal status	6.77%	6.45%
Availability of non-prescription emergency contraception	14.22%	12.34%
Availability of non-prescription non-barrier contraceptives	6.81%	7.34%

Lowest-level criteria

Next, for all the leaves, we needed to set measure levels (or scores). Measure IDs are shown in Figure 4.

Leaves	Measure ID
Type of online information	DZ1
Number of contraceptives	DZ5
Financial information	DZ4
Logistical information	DZ4
Minority languages in the same website	YN
Visualization / Ease of navigation	DZ4
Ease of finding the website	DZ4
Level of reimbursement	DZ2
Special reimbursement for minors	YN
Special reimbursement for vulnerable groups	YN
Reimbursment of cost of consultation	DZ2
Need for parental consent	DZ3
Availability w/o disclosure of legal status	YN
Availability of non-prescription emergency contraception	DZ6
Availability of non-prescription non-barrier contraceptives	DZ6

Figure 5: Measure IDs of leaves.

Particular measure levels and their scores using the conservative approach (maximum preference was 3, lower preferences are distributed equidistantly) are shown in the following figures:

DZ1	**Score**		**DZ2**	**Score**
Gov't supported standalone website	100		Superior to other	100
Gov't supported integrated website	69		Similar to other	74
Non-gov't supported websites	41		Less than other	44
Other online resources	26		No reimbursement	0
No online resources	0			

DZ3	**Score**		**DZ4**	**Score**
No	100		Exceptional	100
Yes (indirect consent)	42		Good	61
Yes (direct consent)	0		Insufficient	28
			Not avalaible	0

DZ5	**Score**		**DZ6**	**Score**		**YN**	**Score**
Superior	100		Yes (legal)	100		Yes	100
Standard	71		Yes (illegal)	46		No	0
Weak	36		No	0			
Insufficient	27						
Not avalaible	0						