

A Natural Language Processing approach to measuring expertise in the Delphi-based scenarios

Yuri Calleo, Simone Di Zio, Francesco Pilla

1. Introduction

In the Futures Studies context, the Delphi method (Gordon, 1994) is a very popular and empirical approach (Dalkey and Helmer, 1963) often used in combination with the scenario method (Kosow and Gaßner, 2008). Futures scenarios, support decision-makers in a long-term planning context, helping to focus on the key projections of possible/plausible futures and on the major factors that will drive those projections (Bishop et al., 2007). Both scenario and Delphi are often combined with other methodologies, but one of the most interesting and accredited combinations involves precisely these two methods, in an approach known as Delphi-based scenario (DBS), in which the results of a Delphi study are used to develop the futures scenarios (Di Zio et al., 2021).

A crucial phase in a DBS regards the building of a panel of experts, generally formed by a group of people having comprehensive or authoritative knowledge in a particular field, therefore particularly suitable for answering very specific questions regarding the topic dealt with. An old open issue – as in any experts’ consultation – regards the measurement of the *expertise* of the panel members, because each expert has a different degree of competence, and it is very difficult to quantify that degree.

In recent years, some contributions carried out to overcome this issue, most of them proceeding with a self-assessment (or “self-rating”) of the experts, asking panellists to rate their own expertise (Mullen, 2003) on the whole subject matter, or even on each item of the questionnaire. However, this approach could solve the evaluation problem only from a general perspective, specifically, we must take into account some not trivial drawbacks: 1. Self-assessment makes the decision-making process even longer, and experts may be discouraged from participating; 2. Self-evaluation can lead to several cognitive biases which greatly distort judgments on self-competence, such as, among others, overoptimism and overconfidence biases (see, for example, Bonaccorsi et al., 2020). These aspects should not be underestimated, since if we engage experts with low knowledge in a field, this may compromise the total perspective of the survey. It is important to underline here that the measurement of the expertise degree is useful to set a suitable weighting system for the proper use of the different levels of competencies in the panel. Given these premises, with the exponential increase in the use of web-based research platforms and websites on the internet, it is possible to have valuable data and information available about experts. This paper proposes to:

1. **O1:** develop a new method to evaluate the expertise degree;
2. **O2:** implement the method with text-mining techniques and statistical analysis for experts’ evaluation;
3. **O3:** overcome the problem related to the evaluation of different types of expertise coexisting in the same panel.

To showcase our method, we selected a cohort of known experts, part of the “Smart control of the climate resilience” (SCORE) H2020 European project as this would allow us to assess the production of experts with Natural Language Processing and estimate their expertise in a specific area. This paper is organised in the following sections: in Section 2 a brief literature review with a specific statement of the problem will be conducted, in Section 3, we explain the methodology used to develop our method, and in Section 4 the results will be illustrated. In Section 5 we conclude

Yuri Calleo, University College Dublin, Italy, yuri.calleo@ucdconnect.ie, 0000-0002-0190-6061

Simone Di Zio, University of Chieti-Pescara G. D’Annunzio, Italy, s.dizio@unich.it, 0000-0002-9139-1451

Francesco Pilla, University College Dublin, Ireland, francesco.pilla@ucd.ie, 0000-0002-1535-1239

Referee List (DOI 10.36253/fup_referee_list)

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Yuri Calleo, Simone Di Zio, Francesco Pilla, *A Natural Language Processing approach to measuring expertise in the Delphi-based scenarios*, © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-0106-3.29, in Enrico di Bella, Luigi Fabbris, Corrado Lagazio (edited by), *ASA 2022 Data-Driven Decision Making. Book of short papers*, pp. 163-168, 2023, published by Firenze University Press and Genova University Press, ISBN 979-12-215-0106-3, DOI 10.36253/979-12-215-0106-3

with possible future implementations.

2. Theoretical framework and related works

Given the variety of expertise involved in a Delphi panel, from the scientific literature, most of the attempts to evaluate the expertise degree are based on self-assessment or from an assessment made by researchers. In the environmental context, for example, Gorn et al. (2018), studying the climate change effects in the Region of Halle, divided the expertise competencies into two categories: i) expert type A and ii) expert type B. Type A is an expert who has specific competence and practical experience in regional planning and ecosystem services, type B is an expert with theoretical knowledge of spatial and environmental planning, regional geography, and ecosystem services.

Some scholars select experts based on their experience, considering their position within the organization and different variables identified by researchers. Gary and Von der Gracht (2015), for example, consider speaking roles at “futures” conferences and membership greater than six years in the area of interest. In these terms, it is interesting to understand how the range of time of experience in a field is important to evaluate since a member who manage the research context for many years should be more expert in comparison to who has low years of study. However, the previous approaches do not solve the issue of evaluating different types of expertise in the same panel.

As previously described, most of the time, researchers evaluate the experts based on a self-rating, for example, Varho et al. (2016), build a matrix where the experts can select from a series of variables, the areas where they have greater or familiar expertise. In this line of research, an interesting coefficient was developed by Barroso and Cabero (2013). The coefficient, named K-expert competence, is based on the self-assessment of experts, considering two components, one related to self-evaluated competence and another to the ability to argue on the subject.

That said, there is a need to develop an objective method that avoids self-assessment or evaluation by researchers or scholars in order to reduce cognitive errors and time-consuming, with enough flexibility to be applied on panels of different natures (for example environmental studies can include several participants with different expertise at both theoretical and practical level). To pursue the research aim, we apply web-mining and text-mining techniques to extract information, in order to obtain objective information in a short time, starting from objective criteria and taking into account a plurality of criteria which, in the mixed panels, are important to consider.

3. Materials and methods

We propose a new method to evaluate the expertise degree generally applicable to all participatory decision-making processes and, in particular, to Delphi panellists. We apply the method considering a list of experts in the coastal erosion context, understanding the degree of expertise of the members in the main keywords of the H2020 SCORE project: “coastal erosion”, “sensors”, “Ecosystem-Based Approaches” (EbA), “flood risk assessment”.

The first phase starts where a list of possible experts to engage is already defined and, to showcase our method, we use a list of the H2020 SCORE project members. In the DBS, the literature does not uniformly agree on the number of experts to involve, however, there is a consensus on the range of 10-30 (see Nowack and Endrikat, 2011), for that we identify a list of $N = 20$ possible experts to be involved as panellists.

The data on the selected experts are organized in a matrix including all the information useful to identify them and their personal pages on the web (e.g., name, surname, personal contacts, personal websites, personal portfolio etc.). In our case, we have different experts with different job roles and expertise, for that, we divide the panel using the following roles:

- r_1 : Academic experts;
- r_2 : Experts from industry and members of companies;
- r_3 : Local and governmental authorities.

A Delphi panel should be as varied as possible, as creativity and the differences in knowledge should be as diverse as possible. However, this opportunity turns into a challenge to be faced, as each of the categories must be evaluated with different criteria. For example, a local authority cannot be evaluated based on scientific publications, or a company manager cannot be evaluated on a social network private profile. In these terms, once we have a data repository with personal information related to each expert, we proceed to evaluate the participants on different variables of our interest, in a multi-criteria approach.

For our study, we decide to extract the number of contributions for each keyword and each expert, from publications, citations, h-index, reports, patents and policies related to the keywords. To acquire the previous information, we refer to the Google Scholar database for the publications, citations, h-index and patents, for the reports we refer to ResearchGate and personal webpages, and for the policies, we take into account the governmental webpages and portfolios of the panellists.

The procedure of data extraction cannot be carried out manually and for that we implement a Python script using the Beautiful Soup library, using text-mining in order to extract the main keywords in a webpage related to a determined topic. Beautiful Soup (Nair, 2014) is a Python library used for web scraping, it allows us to extract data from HTML and XML files obtaining a “parse tree” from the source code of the selected page. First of all, we import all the URLs acquired in the previous phase in Python, after that, we select the keywords of interest (“coastal erosion”, “sensors”, “Ecosystem-Based Approaches”, and “flood risk”) and we run the script.

The outputs show the number of times a given keyword is present on the page without repetitions, allowing us to build separate distributions of h-index, citations, publications, reports, patents, and policies for each expert.

After extracting all the data, we build a matrix, say X , with N experts on the row and p variables on the columns. The first two variables are the h-index and citation, independent of the keywords. The other four variables (publications, reports, patents, policies) are repeated within each keyword. This is because we want to take into account, for example, how many publications an expert has with “coastal erosion” as a keyword, how many reports with the same keyword, etc. Therefore, we have four variables for each of the four keywords, for a total of $p = 18$ variables.

The main shortcoming is that the column vectors of X ($X_{i1}, \dots, X_{ip}, i = 1, \dots, N$) have various locations and variabilities, so they cannot be directly combined. Therefore, the data should be made comparable by normalization and, among the various methods of normalization, here we consider the min-max:

$$Y_{ij} = \frac{X_{ij} - \min_i(X_{ij})}{\max_i(X_{ij}) - \min_i(X_{ij})}$$

To avoid computational problems, in case $\max_i X_{ij} = \min_i X_{ij} = 0$ we set $Y_{ij} = 0$, and if $\max_i X_{ij} = \min_i X_{ij} > 0$, we set $Y_{ij} = 1$.

The last phase permits to have a coefficient of production for each expert (say K_i), based on a weighted sum of the normalized variables, which represent a comprehensive measure of expertise:

$$K_i = \sum_{j=1}^p Y_{ij} w_j$$

with $i = 1, \dots, N, j = 1, \dots, p$ and $\sum_{j=1}^p w_j = 1$.

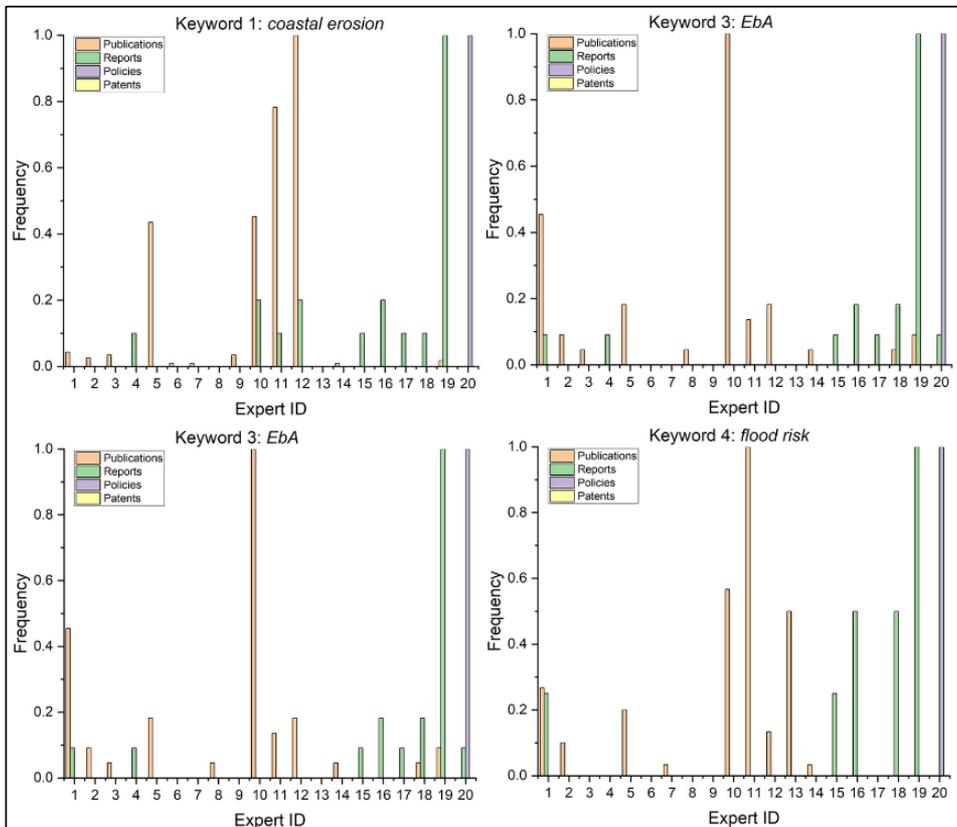
In this application, we set the weights constant to $w_j = 1/p$, but the method is very flexible, and the assessment of each weight is left to the team of researchers. After the normalisation of the variables, we proceeded with a weighted sum of the results with (as a first application and by way of example) constant weights $w_j = 0.05$. In the end, for each expert, we obtained a score for each variable and for each keyword, and a final score K_i calculated as a weighted sum, having in this way both the possibility of evaluating the experts for each keyword, understanding who has greater expertise and evaluating the degree of expertise in the macro-topic of interest.

The weighted sum at the base of the coefficient K_i is only one possible aggregation rule, but other rules can be used, such as a multiplicative one. Also, for normalization, it is possible to use other methods, such as standardization with mean and standard deviation or rank transformation. In these terms, this coefficient becomes a quantitative, flexible, and multi-criteria measure of expertise.

4. Results and discussion

The results illustrated below answered the research objectives and made it possible to have an objective evaluation of a sample of experts. The method is useful for both the evaluation of a predefined panel of experts (for example to weigh their answers in a questionnaire) and to build a new panel, in order to include the people with the highest expertise. The overall results (depicted in Figure 1), demonstrate a high level of expertise in the keywords of our interests, some of which contributed to the topic with publications, reports, and policies.

Figure 1. Normalized frequency for each keyword



Academic experts ($\#r_1 = 14$) contributed efficiently to the research in the field of coastal areas, sensors, EbA and flood risk (Table 1). Specifically, expert 10 is the academic who has contributed most to the areas of our interest with an expertise degree of 0.216 and an h-index of 51 with 16123 citations, for an overall of 95 publications and reports in the keywords analysed.

Experts from the industry sector ($\#r_2 = 5$), contribute to the areas of interest within the publications of reports and scientific articles, however, no patents have been found. In particular, expert 18, published 6 scientific papers with an average of 194 citations. For expert 19, we have found 4 scientific publications and 25 reports submitted in the context of research projects. The only local authority expert ($\#r_3 = 1$) has an overall of 17 policies, 10 in keyword 1, 5 in keyword 3 and 2 in keyword 4 with 1 report in keyword 2.

Table 1. Expertise degree estimates

<i>Expert Id</i>	<i>Role</i>	<i>Expertise degree (K)</i>	<i>Expert Id</i>	<i>Role</i>	<i>Expertise degree (K)</i>
1	Academic	0.113	11	Academic	0.152
2	Academic	0.028	12	Academic	0.166
3	Academic	0.040	13	Academic	0.029
4	Academic	0.016	14	Academic	0.013
5	Academic	0.132	15	Industry	0.022
6	Academic	0.007	16	Industry	0.044
7	Academic	0.010	17	Industry	0.010
8	Academic	0.002	18	Industry	0.046
9	Academic	0.003	19	Industry	0.155
10	Academic	0.216	20	Local authority	0.155

In our application, the high scores were identified in the experts' 10, 12, 19, 20 and 11, and with all other scores, we obtained a full ranking of the experts based on their degree of expertise (Table 1), demonstrating the efficiency of the approach in a different context of applications and for different work roles. With these results, it will be possible to select a subsample of more competent experts ("super experts") and/or weigh Delphi's responses/evaluations of the panel. In this way, there are no restrictions in terms of the choice of participants, as any expertise or work situation can be assessed by setting variables suited to the research work.

5. Concluding remarks and future works

This study proposed a new approach to evaluate the expertise degree in the participatory process, in particular in the Delphi-based future scenarios development. We applied this method to a cohort of experts' part of the "Smart control of the climate resilience" (SCORE) H2020 European project in order to estimate their expertise in the context of our interest. The results showed how the method solves one of the main problems in the decision-making process: the evaluation of participants' expertise is useful, for example, in weighing their assessments.

The method is a contribution to the objective measurement of expertise, useful in the context of panels with heterogeneous types of competencies and based on automated data retrieval.

In the application, we had no citizens that normally could be useful in the last phases of Delphi, however for the citizens we could evaluate blogs, social networks, and personal pages, referring to the main social networks (e.g., Twitter, Instagram, LinkedIn etc.).

For future work, it would be interesting to have a comparison between the objective measure described in the paper and the self-evaluation of experts. Furthermore, to consider other aggregation formulas as well as other normalization methods. Finally, to set appropriate weights for the selected variables, among the various possible approaches, we suggest the application of the Analytic Hierarchy Process (AHP), which is very efficient in generating objective weights in a multi-criteria context (Saaty, 1980).

Acknowledgements

The work carried out in this paper was supported by the project SCORE which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101003534.

References

- Barroso O. J. M., & Cabero A. J. (2013), La utilización del juicio de experto para la evaluación de TIC: el coeficiente de competencia experta. Bordón. *Revista de Pedagogía*, **65** (2), pp. 25-38. <https://doi.org/10.13042/brp.2013.65202>.
- Bishop P., Hines A., Collins T. (2007), The current state of scenario development: An overview of techniques, *Foresight*, **9**(1), pp. 5–25.
- Bonaccorsi A., Apreda R., Fantoni G. (2020), Expert biases in technology foresight. Why they are a problem and how to mitigate them, *Technological Forecasting & Social Change*, **151**, pp. 1–17.
- Dalkey N., & Helmer O. (1963), An experimental application of the Delphi method to the use of experts. *Management science*, **9**(3), pp. 458-467. <https://doi.org/10.1287/mnsc.9.3.458>.
- Di Zio S., Bolzan M., Marozzi M. (2021), Classification of Delphi outputs through robust ranking and fuzzy clustering for Delphi-based scenarios, *Technological Forecasting & Social Change*, **173**, pp. 1211400040. doi.org/10.1016/j.techfore.2021.121140
- Di Zio S., Rosas, J. D. C., & Lamelza, L. (2017), Real Time Spatial Delphi: Fast convergence of experts' opinions on the territory. *Technological Forecasting and Social Change*, **115**, pp. 143-154. <https://doi.org/10.1016/j.techfore.2016.09.029>.
- Gary J. E., & Heiko A. (2015), The future of foresight professionals: Results from a global Delphi study. *Futures*, **71**, pp. 132-145.
- Gordon T. J. (1994). The delphi method. *Futures research methodology*, **2**(3), pp. 1-30.
- Gorn L., Kleemann J., & Fürst, C. (2018), Improving the Matrix-Assessment of Ecosystem Services Provision—The Case of Regional Land Use Planning under Climate Change in the Region of Halle, Germany. *Land*, **7**(2), pp. 76. <https://doi.org/10.3390/land7020076>.
- Kosow H., Gaßner R. (2008), Methods of future and scenario analysis: overview, assessment, and selection criteria, Bonn, Deutsches Institut für Entwicklungspolitik.
- Mullen P. (2003), Delphi: myths and reality, *Journal of Health Organisation and Management*, **17** (1), pp. 37–52.
- Nair V. G. (2014), Getting started with beautiful soup. Packt Publishing Ltd.
- Nowack M., Endrikat J., & Guenther E. (2011), Review of Delphi-based scenario studies: Quality and design considerations. *Technological Forecasting and Social Change*, **78**(9), pp. 1603-1615.
- Rodríguez P. O. S. (1995), Land use conflicts and planning strategies in urban fringes: a case study of western Caracas, Venezuela.
- Rowe G., & Wright G. (1999), The Delphi technique as a forecasting tool: issues and analysis. *International journal of forecasting*, **15**(4), pp. 353-375. [https://doi.org/10.1016/S0169-2070\(99\)00018-7](https://doi.org/10.1016/S0169-2070(99)00018-7).
- Saaty T.L. (1980), The Analytic Hierarchy Process. McGraw-Hill, New York.
- Sossa J. W. Z., Halal, W., & Zarta, R. H. (2019), Delphi method: analysis of rounds, stakeholder and statistical indicators. *Foresight*. <https://doi.org/10.1108/FS-11-2018-0095>.
- Varho V., Rikkonen P., & Rasi, S. (2016), Futures of distributed small-scale renewable energy in Finland—A Delphi study of the opportunities and obstacles up to 2025. *Technological Forecasting and Social Change*, **104**, pp. 30-37.