Aldo Benini / José Cobos Romero
Data-Friendly Space (DFS)

Humanitarian severity and information gap measures using Possibility Theory

Companion note to an Excel demonstration workbook

"Knowing ignorance is strength.
Ignoring knowledge is sickness."

Lao Tzu, Chinese philosopher,
around 600 BC

Version 27 January 2022

Suggested citation:

Aldo Benini, José Cobos Romero (2022). "Humanitarian severity and information gap measures using Possibility Theory. Companion note to an Excel demonstration workbook". Data-Friendly Space (DFS). Version 27 January 2022.

Table of Contents

Tables

Figures

Formulas

Acknowledgments

Patrice Chataigner, DFS Lead Analyst, encouraged and guided this project.

The data using in the Excel demonstration workbook and for additional analyses in this note resulted from the work of the

iMMAP Columbia Team:

Xitong Zhang, former Senior Information Manager and Analyst Project Lead
Alberto Castillo, Information Management Expert
David Schoeller, Senior Information Manager and Analyst Project Lead

and of the DFS Team:

Marcela Durán, Analyst
Cindy Domínguez, Analyst
Claudia Domínguez, Analyst
Balixsy Alvarado, Analyst
Xavier Chataigner, Quality Control Analyst.

Matthew Smawfield, DFS Data Visualization Expert, checked some of our data management operations and created a fast running replication of the Excel workbook essentials on a different platform.

Attilio Benini created the map included in the Summary.

We thank those persons for their support, sharing and insights.

We are grateful for DFS' financial support.

Any errors of analysis and interpretation are solely ours.

Aldo Benini
José Cobos Romero

Summary

Possibility Theory, a variant of probability theory, is well suited to analyze large-N measurements of variable reliability, obsolescence and redundancy. Very few applications are known from the humanitarian sector, and these are about stocks and flows of relief goods measured on *continuous* scales.

We demonstrate the usefulness of Possibility Theory in testing the *binary* hypothesis that the severity of humanitarian conditions in a given region and sector is high / not high. The test relies on numerous *ordinal* severity ratings produced by coders who review humanitarian reports with location, time, sector, affected group and context information. The statistic of interest is a measure of confidence that the true severity is high / is not high. This measure is continuous-bounded in the interval [-1, +1] and thus easy to visualize in tables and maps. Also, we propose an allied measure of information gaps.

The demonstration data are from Colombia during a 14-month period in 2020-21 (18 May 2020 – 30 June 2021). Two organizations, iMMAP and Data Friendly Space (DFS), collected, excerpted and processed relevant documents of various types in a dedicated database application known as DEEP. In 2021, DEEP projects were operational in thirty countries. In Colombia, it delivered the information base for the Humanitarian Needs Overview (HNO) 2021.

The iMMAP / DFS coders parsed 357 documents ("leads") and turned them into 1,540 DEEP records ("entries"). From these we derived 24,920 observations each with a location, publication date, sector and rating on a 5-level severity scale. Our algorithm produces confidence and information gap estimates at the aggregate level (pairs of Department [Admin1] and sector). All 33 Departments and 11 sectors are represented, although not for all pairs.

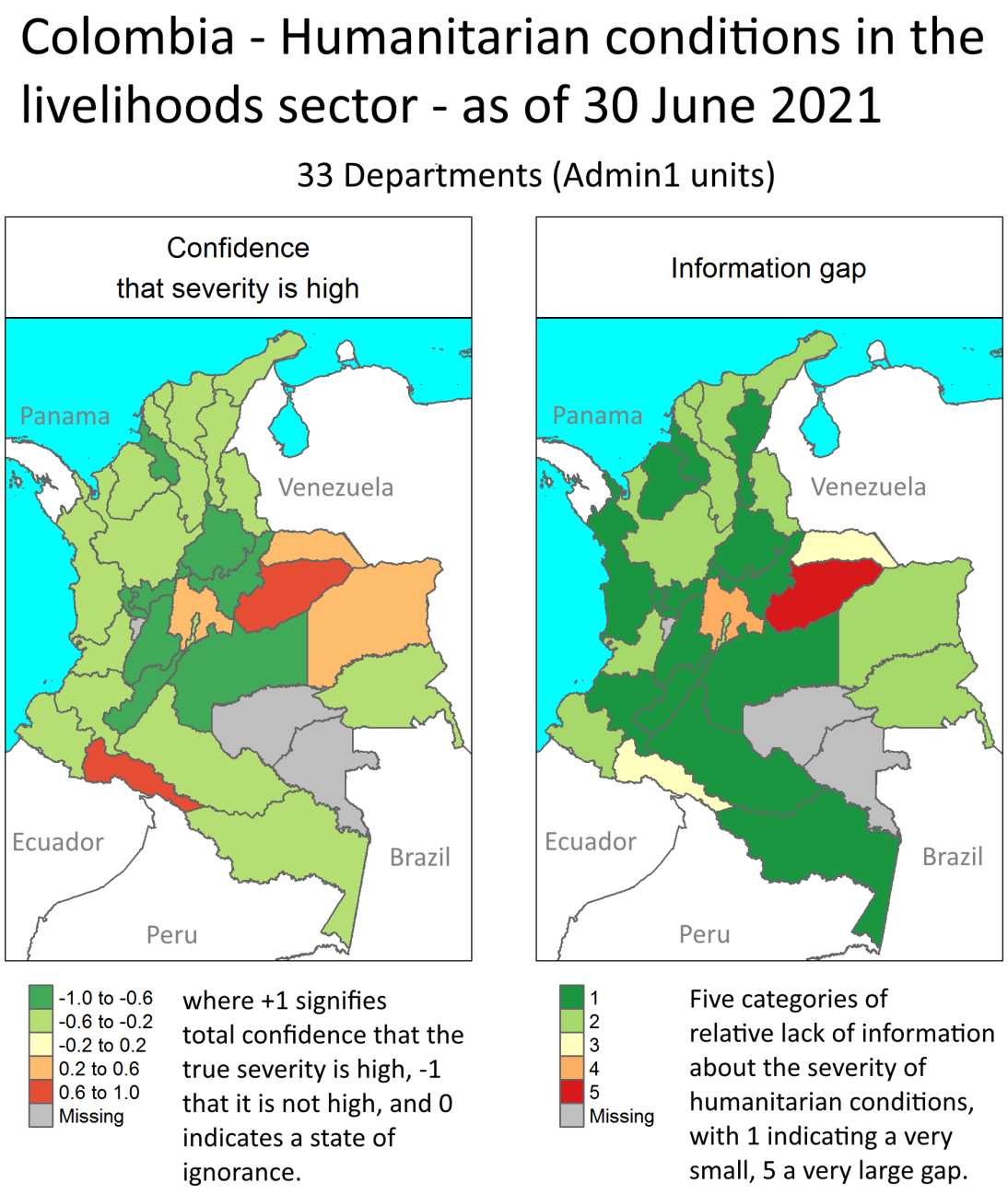
The Excel demonstration workbook uses the data from a subset of six Departments, keeping 10,624 observations. The workbook architecture is such that the user can change parameter settings, view the updated outcomes in Department X sector tables, and compare them to the ones under the initial settings, all in the same sheet. Those interested in the inner workings find explanations in column header comments and in the "back office" sheets that do the work of aggregation and gap measure calculation.

This companion note details the motivation to turn to Possibility Theory, minimal generic elements of the theory, our choices in adjusting for reliability, obsolescence and redundancy, the steps leading to the confidence measure, as well as the components of the gap measure. The note then describes the data generation and analysis workflow in the DEEP institutional environment. It presents select results from the whole-Columbia dataset. It explains the workbook structure and the functions of the various parameters, as well as some less common types of formulas. Deliberately, no VBA programming is involved.

Together, note and workbook demonstrate how an infrastructure like DEEP and unusual tools like Possibility Theory open a way to combine large sets of structured humanitarian observations in a quantitative severity measure. Our measure captures the degree of

uncertainty in a more informative way than the usual rank order-based statistics for ordinals. In addition, the gap measure points to regions and sectors needing further investigation or needs assessments.

Figure 1: Sample map of confidence and information gaps



Introduction

What this is about

This note provides conceptual and technical background to a demonstration workbook in Excel. The workbook demonstrates a method of aggregating large numbers of ordinal severity ratings obtained from Admin1 and 2-level information. The result is a numeric measure of severity for pairs of Admin1 units and sectors in a given country. The method relies on Possibility Theory, which is an alternative to classic probability. Possibility Theory is more suitable to deal with information of variable reliability, age, granularity, redundancy and incompleteness.

A further use of the aggregate severity measure is in constructing and calculating an information gap measure for all Admin1-level units and sectors that had information with ordinal severity ratings during the observation period.

The data for the Excel demonstration are from Colombia. They cover a 13-month period in 2020-21. They were collected by two organizations, iMMAP and DFS, and are held in the DFS-developed application DEEP.

The Excel workbook formulas and connections among sheets and variables provide guidance for the possible implementation of those measures in the DEEP front and backends, augmented by a mapping facility.

Motivation

The measurement of severity in needs assessments

Humanitarian needs assessments produce large numbers of severity-related statements, including in the form of ratings. Ratings are ordinal variables, usually based on a standardized scale applied to several contexts, sectors and types of affected people. What observations qualify for particular levels on the scale may be laid down in specific operational definitions, or may be left to the broad judgment of experts. The raters may be contributors, authors or reviewers of reports and statistics pertaining to unmet needs.

For the rest of this note, we assume that all severity ratings are on a common ordinal scale, applied to all localities, sectors and publication dates in a dataset of interest. The scale has a fixed number of levels, say between 3 and 7; the levels have names that verbally express increasing severity of unmet needs as well as an increasing urgency of intervention. As noted, different sectors may follow their own operational definitions for each level. Moreover, the ways of transforming relevant documents into detailed ratings may vary on unobserved informal criteria.

The common scale is fundamental to aggregating numerous ratings into a severity measure at a higher level. Still, the aggregation can be challenging for several reasons:

- The groups of affected persons overlap.
- Different assessments cover different groups and subsets of sectors. By the ideal of a full matrix (or a "balanced panel"), missing values abound.
- Reports provide information of different administrative granularity, some referring to low-level administrative-geographic units, some to higher-level units, yet others to mixtures of levels. This makes it intrinsically difficult to evaluate information gaps at lower levels[1].
- Information becomes obsolete; only part of previous measurements and expert judgments are updated.
- There is redundancy. Some of the ratings are not statistically independent; the observations are, in the language of sample surveys, "clustered".
- Errors of different types - model, sampling, measurement – compound.
- When measures are ordinal, as ratings are, many statistical procedures are not directly applicable.

Severity ratings with those limitations are still useful for the evaluation of local situations, say, of unmet needs in one or a few sectors in one locality, by decision makers well familiar with the affected groups and with the assessment process. However, their synthesis from larger numbers of communities, sectors and points in time easily overtaxes ordinary tools and understanding.

To illustrate, we look at the first of the above-enumerated challenges with a national dataset used in a demonstration and described further down. In these data, ten affected-person groups are distinguished at various levels of detail, some distinct, some overlapping, plus the undistinguished "All" and "No specifics":

[1] This is akin to the "multi-resolution problem" known from remote sensing; e.g., Azar, Engstrom et.al. (2013). See also Benini, Chataigner et.al. (2016:14-15).

Table 1: Affected-persons definitions with overlap

Group	Level0	Level1	Level2	Level3	Severity ratings
1	(no specifics)				1,747
2		Affected			6,990
3		Affected	Displaced		740
4		Affected	Displaced	Asylum Seekers	51
5		Affected	Displaced	IDP	725
6		Affected	Displaced	Others of Concern	5,315
7		Affected	Displaced	Refugees	2,312
8		Affected	Displaced	Returnees	1,740
9		Affected	Non-Displaced		2
10		Affected	Non-Displaced	Host	1,852
11		Not-affected			1
12	All				3,445
Total ratings					**24,920**

Source: iMMAP Colombia, DFS. Severity ratings from 33 Departments, 18 May 2020 - 30 June 2021.

In the following, we propose a method for aggregating such problematic sets of severity ratings. The aggregate measure takes into account:

- The **obsolescence** of the information – This has two aspects: 1. Older ratings tend to say less about the severity of the current situation than newer ones. 2. However, all we know is the combined information up to the point of the latest document captured – about the time thereafter the captured sources add no further knowledge (newer knowledge may exist elsewhere, outside those documents).
- The **reliability** of the information may vary by source and by the specificity of the ratings. Ratings specific to lower-level administrative units are generally more reliable than ones referring globally to higher-level units. Ratings of a purely intersectoral extent are to be excluded; they cannot be reliably claimed for all sectors (although their underlying information may be relevant outside this rating approach).
- **Redundancy** control: Simple repetition of a piece of information from the same source does not increase its value. Multiple statements that vary in response to subordinate aspects of the same object do add value, but the value increases less than linearly with their number. Concretely, a document with information on commune A may pack twenty severity ratings, varied by affected group and demographic profile. Another document, about commune B, provides only five ratings. Both A and B are in province X. When aggregating their ratings to a combined measure on X, a weighting function is needed. The function must define which subsets of ratings should be considered statistically dependent, e.g. all those that are from the same commune and the same sector and derived from documents published on the same date.

A measure of information gaps

Assuming that the method produces a valid severity measure, its users will still want an overview of the assessment activities in terms of information gaps for the entire humanitarian response theater. Thus, a gap measure is needed. Its scope must be appropriate: National and inter-sectoral averages are not feasible; there is no absolute gauge for information gaps at such high aggregation. A relative measure is called for, at a lower administrative level, with comparability within a given sector and between sectors for a given administrative unit.

We construct such a relative gap measure on these assumptions:

1. The measure refers to the combination of a given region (e.g., a province) and a given sector.
2. Relative to other regions and sectors, the information gap is the wider
   - The higher the severity
   - The longer the time between the latest captured document and the end of the observation period (the date when the database was closed, or the current date in ongoing data collection)
   - The larger the affected population.

It follows that the severity measure has to be ratio-level (it has to have a zero point). Moreover, the gap measure, being relative, is sample-dependent. This means that the addition or subtraction of an observed unit (e.g., a province with several assessed communes) may change the gap scores for all other units.

The gap measure must incorporate two more considerations:

- The obsolescence effect arrives in two parts. The severity measure incorporates it up to the latest document related to the unit of aggregation (e.g., the pair of a province and a sector). The severity now is the severity at the point in time when the latest document was published – by definition, there have been no new ratings since. The second part is added in the gap calculation. It is a function of the time between then and the current date (or the end of the observation period).

- The affected population may be difficult to estimate in size and composition. Generally, the accuracy decreases from registered refugees to unregistered IDPs to members of the host population significantly affected. In many situations, a proxy measure is unavoidable. Estimates of the entire host population, extrapolated from the latest national census, may be most easily available. If the proportion of affected persons tends to be lower in the more populous provinces, the population weights should be transformed by an appropriate function. Similarly, this function could translate a political will to pay smaller provinces (which tend to be neglected) more attention than the size of their populations justifies.

This approach comes with an obvious conceptual weakness. Severity and gap measures cannot be calculated for regions-sectors on which there are no severity ratings. The absence

of these may be due to a consensus that there are no humanitarian needs there. Lack of resources, of political will or access to conduct assessments may conceal real needs. The seriousness of the gaps can be evaluated with external information outside the ratings. The same raters may already have that information from other sources, but it will not directly fill the gaps in the quantitative gap matrix. In theory, area experts might be able to fill the missing cells with their estimates (duly marked as from a separate process), but no practical experiments have been made yet.

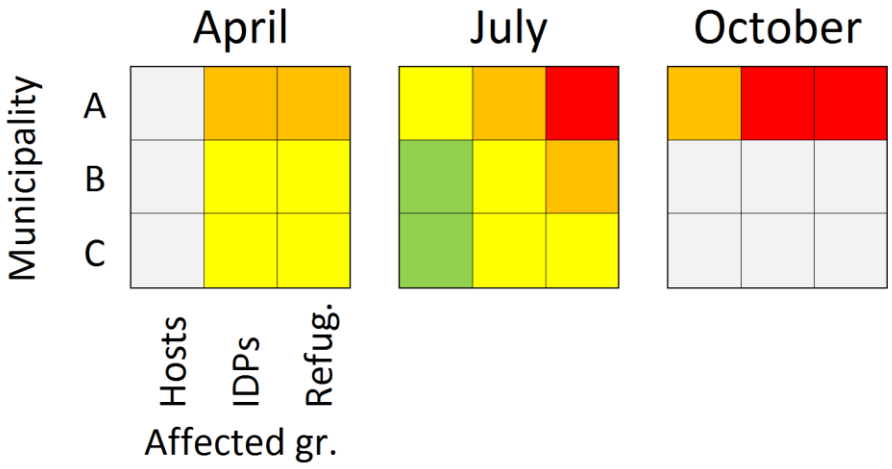
Possibility Theory for heterogeneous data

The compression of needs assessment data into ordinal severity ratings implicitly uses subjective probabilities. Raters assign severity levels considering the risks that they infer for the loss of core values in the affected group – life, dignity, prior living standards, coping ability, and more, as demanded in the rating template. These ratings can be mapped to the risks that a typical rater would assess from documents that he/she reviews. These risks could, technically, be aggregated by the tenets of classic probability theory.

However, "a large number of studies have found effects showing that subjective estimates do not conform to the requirements of probability theory. … The probabilistic framework is also weak for describing partial ignorance, that is, cases where uncertainty about an event is poorly correlated with uncertainty about the opposite event" (Raufaste, da Silva Neves et al. 2003:198). For example, we might be reluctant to publicly commit to a forecast that the humanitarian situation "will improve with a 40 percent probability, and stay the same or worsen with 60 percent". Interval estimates of "30 – 50 percent", and the complement of "50 – 70", agree better with the uncertainty that we perceive about imprecise outcomes. These are difficult to synthesize (although not impossible; see e.g., Weichselberger 2000), and it is unclear how ordinal ratings could be transformed into a uniform set of intervals.

Here a softer version of probability theory, "Possibility Theory", helps. This theory is "specifically designed to model incomplete information. … Incomplete information refers both (i) to the inability of a source to capture all information .. as well as (ii) to situations in which not all states are observable" (Holst and Lohweg 2020).

Figure 2: Incomplete coverage over time



Province X (Admin1) comprises of three municipalities A, B and C (Admin2). A sector-specific (say, shelter) assessment of refugee and IDP communities in April reveals difficult conditions in B and C, and worsening ones in A. By July, the situation of the refugees has further worsened in A and B; this assessment looks also into the situation of the host communities, for which significant difficulties have appeared in A. In October, due to limited resources, only A is reassessed, revealing further deterioration for hosts and IDPs. A synthetic judgment on the entire province, therefore, must combine current information on A with "old" information on B and C. The aggregation method must balance the ignorance of B and C's current state with the plausible assumption that they followed the same dynamic as A.

Humanitarian studies have rarely taken advantage of Possibility Theory. Rare applications concern relief logistics (Rabbani, Manavizadeh et al. 2015, Tofighi, Torabi et al. 2016), anti-personnel mine detection (Milisavljevic 2017), preparations for housing IDPs (Akimbayev, Akhmetov et al. 2020)[2]. In the Covid-19 pandemic, Huang (2020) estimated mortality in China, combining possibilistic and classic probability.

Ratings as evidence for a binary hypothesis

For simplicity, we will present this theory in a very limited way – for its use with binary hypotheses to which multiple ordinal observations can be related. Our approach to this type of situation is largely based on the outline in Lesot et al. (2011).

Our binaries are: "the true severity is high" vs. "the true severity is not high". The state deemed more likely between the two is our hypothesis. The ratings are the evidence on which the hypothesis is evaluated. The task is to so synthesize the ratings as to produce an aggregate measure of our belief that the severity indeed is high / is not high. The requirements for the synthesis to work are:

[2] Regrettably, the paper by Akimbayev et.al. was not traceable. The Tofighi et.al. paper has been cited more than 300 times. The logistics applications work with stocks and flows of relief goods; these variables are ratio-scale. As such, their possibilistic treatment, apart from being mathematically more challenging, are of no immediate interest for us.

- A common rating scale uniformly applied across localities, sectors and time
- A common set of subjective probabilities that the true severity is "high" for the scale levels

An illustration of the second point will presently follow.

In classic probability, the probabilities of "A happens" and "A does not happen" strictly sum to 1. Possibility Theory relaxes that rule. It assigns the hypothesis a score of 1 ("fully possible"); its complement gets a score from the interval [0, 1) (the half-open interval says "from impossible to less than fully possible"). The sum of their scores, therefore, is ≥ 1. For example, we may believe that the severity is high although we do not exclude that with better information it would not be high. The score for "high severity" will be 1, that for "not high" could be, for the sake of example, 0.4. Now we have two pieces of information: The choice of the favored alternative (the hypothesis), and a – so far enigmatic - fractional value (0.4) attached to its negation.

To operate on the fuzzy premise of sums of possibilities greater than one, Possibility Theory needs a second key concept – "necessity". Broadly speaking, necessity is the strength of our belief that a fully possible state will indeed happen. In our example, the necessity of "high severity" is 1 – possibility ("not high") = 1 – 0.4 = 0.6. The necessity of "high" depends on how close "not high" comes to impossibility. The necessity of state A can be measured as the strength of the evidence that speaks against "non-A".

To make this transparent, we assume – and later actually use – a severity scale with five levels. We consider the top two levels as expressions of the belief that severity is high while the three at the bottom suggest that it is not high. The ratings are observed; the high / not high severity is inferred. The table details the reasoning.

Table 2: Severity levels, subjective probabilities, possibility scores

Severity level		**Case of "high severity"?**			**Possibility scores for**		
Verbal rating	Ordinal numbered	Binary verbal	Binary numeric	Subjective probability that true condition is "high":	Severity is high	Is not high	Sum of poss. "high" and "not high"
No problem	1	No	0	0.05	0.05	1.00	1.05
Of concern	2	No	0	0.10	0.10	1.00	1.10
Major	3	No	0	0.25	0.25	1.00	1.25
Severe	4	Yes	1	0.90	1.00	0.10	1.10
Critical	5	Yes	1	0.95	1.00	0.05	1.05

The five levels are mapped to belief strengths ("subjective probabilities") ranging from a low 0.05 for "No problems" to a high 0.95 for "Critical". The green columns to the right render the possibilistic scores and their row sums. For example, "No problem – Severity is high" -> 0.05 may be understood as "In the average rater, when he/she gives a 'no problem'

rating, the probability of a false negative is 0.05". This is equivalent to saying that "in one of 20 situations given the rating 'no problem', the true severity is high". And in terms of necessity, it means that "the necessity value that the severity is not high is 1 – 0.05 = 0.95".

The reader should be able to find equivalent interpretations at the other extreme, i.e. for "Severe" and "Critical", in terms of the probability of a false positive.

A legitimate question arises from our assignment of the central level, "Major – Severity is high" -> 0.25. This implies that raters giving "major" are wrong in one of every four instances. Would it not be more appropriate to set a value of 0.5, acknowledging a high degree of ignorance about the true severity? (0.5 implies a false negative in one of every two instances.)

The problem with mapping the central level to the mid-point of the [0, 1] interval is that it washes out the distinction between high and not-high. As a result, the distribution of the aggregate severity measure over the region-sector combinations will be more centered, with flatter low and high tails. On both sides, a small number of outliers may appear, but a number of cases that in reality are low-severity will be raised to a difficult-to-decide middle range. It seems preferable to assign a possibility score to "major" that marks it clearly as believing that severity is not high, although far less firmly so than an "Of concern" rating.

Necessity and confidence

The previous table is a display of scoring rules, not an empirical dataset. Here we simulate a simple set of six key informants each rating one commune in a given sector. The communes are part of the same province.

Table 3: From subjective probability to possibility

Rating ID	**Severity level**		**Case of "high severity"?**			**Possibility scores for**	
Commune	Verbal rating	Ordinal numbered	Binary verbal	Binary numeric	Subjective probability that true condition is "high":	Severity is high	Is not high
A	Severe	4	Yes	1	0.90	1.00	0.10
B	Of concern	2	No	0	0.10	0.10	1.00
C	Important	3	No	0	0.25	0.25	1.00
D	Severe	4	Yes	1	0.90	1.00	0.10
E	Critical	5	Yes	1	0.95	1.00	0.05
F	Severe	4	Yes	1	0.90	1.00	0.10

As a simple indicator of the true state of severity in the province, we compare the arithmetic mean of the scores for "high" and "not high", 0.73 vs. 0.39. The evidence points to "high severity". However, the aggregate distribution (0.73, 0.39) violates the rule that one state must be fully possible. To satisfy it, we proportionately elevate the values such that its maximum becomes 1 – an operation called "normalization". Multiplying by 1/0.73, we get the normalized scores (1.00, 0.54). The necessity value of the "severity being truly high" is 1 – the mean possibility score of the opposite state = 1 – 0.54 = 0.46. The necessity value

of a less than fully possible state is always set to 0. So we can write the necessity vector as N("high", "not high") = (0.46, 0).

We are still dealing with *two* pieces of information: the hypothetical state at the aggregate level – with a normalized possibility of 1 – and its necessity, a value in [0, 1). In another dataset, the outcome might be different, with "not high" becoming the preferred hypothesis. The necessities might turn out as something like (0, 0.65), a strong indication that severity is not high.

Looking for a convenient *single* measure, we define a confidence score on the interval [-1, +1], where the value -1 stands for absolute certainty that the severity is not high, +1 that it is high. The midpoint, 0, stands for total ignorance or, better perhaps, undecidability.

The function to calculate the confidence is straightforward. Let P(.) denote a normalized possibility, N(.) a necessity, and C the confidence.

Formula 1: Confidence score from necessity scores

If P("is high") = 1,

then	$C = N$("is high")	= 1 - P("is not high")
Else	$C = -N$("is not high")	= -(1 - P("is high"))

In practice there is no absolute certainty; the confidence score cannot reach the exact -1 or +1 bounds. With all subjective probabilities in the rankings >0 and <1, possibilities are always > 0.

This table recapitulates the sequence of calculations.

Table 4: From severity ratings to confidence scores

1. Produce the possibility scores of the observation-level ratings:			
Rating ID	**Severity level**	**Possibility scores for**	
Commune	Verbal rating	Severity is high	Is not high
A	Severe	1.00	0.10
B	Of concern	0.10	1.00
C	Major	0.25	1.00
D	Severe	1.00	0.10
E	Critical	1.00	0.05
F	Severe	1.00	0.10
2. Aggregation by the arithmetic mean operator:			
This produces a subnormal possibility distribution (both values <0).			
		0.73	0.39
3. Normalization by division by the larger of the two values:			
One of the two states returns to "fully possible".			
		1.00	0.54
The "Severity is high" hypothesis is the preferred one.			
4. Calculation of necessity values:			
For the state with norm. poss. <1, it is 1 - 1 = 0.			
		0.46	0
5. Calculate the confidence score:			
Using the formula in the text			
		0.46	
Confidence in "severity is high" is only moderate, because of the ratings of communes B and C.			

Qualitative adjustments

So far, we have not gained much over a probabilistic treatment of a set of ordinal ratings. There are classic tests of the differences in the median ratings between two groups as well as one-sample tests of whether the median is below a certain level[3]. If the median is not significantly lower than level 4 ("Severe"), then we may be satisfied that the true severity is high. For these tests, subjective probabilities, let alone Possibility Theory, are unnecessary.

Possibility Theory comes fully into its own when the ratings require qualitative adjustments. We have already noted occasions that call for adjustments – variable reliability, obsolescence and redundancy. The adjustments are made, not on the ratings, but on the possibility scores. Those for reliability and obsolescence happen at the individual observation level (i.e., within each data table row that holds a rating). The adjusted scores

[3] E.g., Rank sum tests (Mann and Whitney 1947), Somer's D (Wikipedia 2016) for the two-group case; signed-rank tests (Wikipedia 2021c) for the matched-pair and one-sample cases.

can be aggregated to their means and normalized as shown. Adjusting for redundancy involves intermediate steps that involve multiple observations.

Reliability

To adjust for reliability, individual ratings need reliability scores from the interval [0, 1], where 1 denotes "totally reliable", and 0 "never reliable". The scores may be imported from a lookup table that assigns each information source ever used its own value. Alternatively, the scores may depend on administrative or demographic information already present in the record, such as whether the rating refers to an Admin1, Admin2, etc. unit. Commonly, the more specific the information (e.g., the lower the administrative unit), the higher the reliability. The intuitive idea is to combine the reliability score with the unadjusted possibilities such that *less reliable* ratings *diminish the necessity* of the fully possible state. The function to do that is from Lesot et al. (op.cit., 953) and is surprisingly simple:

Formula 2: Adjustment for reliability

$$\textit{Reliability-adjusted possibility} = \textit{Reliability} * (\textit{raw possibility} - 1) + 1$$

This table illustrates adjusted scores for two levels of reliability.

Table 5: Adjusted possibilities by level of reliability - Examples

Severity level	**Raw possibility scores**		**Source reliability**			**Reliability-adjusted**	
Verbal rating	Severity is high	Is not high	Source	Verbal	Score	Severity is high	Is not high
No problem	0.05	1.00	A	Very reliable	0.9	0.15	1.00
No problem	0.05	1.00	B	Mostly reliable	0.7	0.34	1.00
Of concern	0.10	1.00	A	Very reliable	0.9	0.19	1.00
Of concern	0.10	1.00	B	Mostly reliable	0.7	0.37	1.00
Major	0.25	1.00	A	Very reliable	0.9	0.33	1.00
Major	0.25	1.00	B	Mostly reliable	0.7	0.48	1.00
Severe	1.00	0.10	A	Very reliable	0.9	1.00	0.19
Severe	1.00	0.10	B	Mostly reliable	0.7	1.00	0.37
Critical	1.00	0.05	A	Very reliable	0.9	1.00	0.15
Critical	1.00	0.05	B	Mostly reliable	0.7	1.00	0.34

Obsolescence

Needs assessment information loses value with age. While some pieces in a report may grow obsolete sooner than others, specific rates of **obsolescence** for sectors or regions are hard to determine. Practically, one is reduced to working with a constant rate for all. This can be expressed as the uniform half-life of severity judgments. Technically, we multiply the possibility scores by a factor that combines the chosen half-life parameter with the days lapsed since the publication of the assessment report. The exponential function

Formula 3: Obsolescence factor

$$\text{Obsolescence factor} = 0.5^{(\text{days since publication} \,/\, \text{halflife})}$$

makes the constant-rate assumption practical. It shares with all exponential functions the generic property:

$$x^{(a+b)} = x^a * x^b$$

This is the key to an important result of the subsequent aggregation and the normalization of the aggregate scores. In the normalized aggregate possibility scores, the obsolescence adjustment affects *only those observations older than the newest observation* in the region-sector pair in point.

> Assume, for example, the adjusted scores for "severity is high" / "is not high" are based on just three observations. These were published 150, 100, and 50 days ago, respectively. If the half-life is 50 days, then the calculated obsolescence factors are (1/8, 1/4, 1/2).
>
> These can be written has (1/4 * 1/2, 1/2 * 1/2, 1 * 1/2). In each observation, both scores, for "high" and "not high", are multiplied by the same factor. In the normalization, we divide the adjusted aggregate scores by the higher of the two. Therefore, the common part of the factors – in this example it is 1/2 – cancels out. The effective obsolescence factors are (1/4, 1/2, 1).
>
> The result therefore is the same as if the observations were published 100, 50 and 0 days ago. Or, if you will, 180, 130, and 80 days ago, etc., as long as the differences up to the newest remain the same.

This property is desirable for a severity measure. Since the latest observation, no new rating information has been added to our knowledge about the severity in that region and that sector in point. The severity estimate at hand does grow obsolete with time, but it does not change until the arrival of new information. (The same property is undesirable for the information gap measure – a point to which we will come back later.)

Redundancy

The amount of detail in assessments may vary greatly by locality, sector and time. Key to the redundancy adjustment is an analytic device that we call the "cluster". We define a cluster as the subset of rankings from the same lowest-level observed administrative unit with its distinct code in the database (e.g., sub-district, municipality, etc., with its official administrative code), same sector and same age (determined by the publication date.).

For illustration, we again turn to the Colombia dataset, specifically to a cluster of six ratings generated from a health sector report[4]. For simplicity, we pretend that the ratings are perfectly reliable, and that the report has been published today. Adjustments for reliability and obsolescence thus are not needed.

The ratings are based in part on verbal summaries, in part on values of standardized variables. They show a higher than usual variation, from "Of concern" all the way to "Critical". There is no definitive way for an outside reader to grasp the bits that are redundant across the six observations. Instead, general assumptions have to be translated into formulas that produce a quantitative factor to adjust the possibility scores. We present two methods.

[4] Lead Id = 44766, 29 November 2021. Department = Valle de Cauca ("VdC" in the table). No municipality codes; the ratings refer to the whole Department.

Table 6: Two methods to adjust for redundancy - Example

[Context] / Demogr. group/ special needs	Outcome	Severity rating	Subj. probability	P("is high")	P("is not high")	Redundancy adjustment factor	
						Method 1	Method 2
[Dengue fever update, entire Dept.]		Of concern	0.10	0.10	1.00	0.408	0
[Covid-19 upodate] Chronically Ill		Of concern	0.10	0.10	1.00	0.408	0
[Dengue, several Depts., incl. VdC]		Major	0.25	0.25	1.00	0.408	0
[Deaths from Dengue, entire Dept.]	Deaths	Major	0.25	0.25	1.00	0.408	0
[Gunshot wounds, several Depts.]	Injuries	Severe	0.90	1.00	0.10	0.408	0
[Malnutrition, several Depts.] <5 years old)	Deaths	Critical	0.95	1.00	0.05	0.408	1

Method 1 creates identical weights for all ratings in the cluster. We make two assumptions. First, the real-world processes that result in the different ratings are fairly autonomous the ones from the others, despite some common contextual factors. The six ratings, therefore, have equal importance. Second, while the information in larger clusters is prone to more redundancy, its value for the aggregated severity measure still increases with the number of ratings ("the cluster size"), though less than proportionately. Therefore, it is appropriate to set the weight of rating *j* in cluster *i* to

Formula 4: Weights for redundancy adjustments, Method 1

$$W_{ij} = (\text{size of cluster } i)^{D}$$

where D is a dampening exponent from the interval [-1, 0]. At D = 0, all weights within and across clusters are 1. At D = -1, the sum of weights in any cluster, regardless of size, is 1. The values in the table above use D = -0.5; the weights turn out as $6^{(-0.5)} = 0.408$; the weight sum as $6 * 6^{(-0.5)} = 6^{(+0.5)} = 2.45$; at this D value, the weight sum is always equal to the square root of the cluster size. These redundancy weights are applied equally to the "high" and "not high" possibilities, as the obsolescence factor before was.

Method 2 assumes that the ratings in a cluster are closely interconnected. Usually, the same rater processes all the information that is the basis of the cluster. His/her personal knowledge underpins all the ratings. Moreover, we assume that similar processes have produced the various aspects of the situation that the ratings capture. This is likely so even when ratings within the cluster differ. For example, a low rating for one demographic group at present does not eliminate the risk that in future it will progress to the higher severity level already reached by another group in the cluster. To err on the side of caution, the cluster's contribution to the aggregated measure should be biased towards "severity is high". A weight of 1 is given to one instance of this combination:

- the largest "is high" possibility score and,
- within this subset, the lowest "is not high" score.

For all other observations in the cluster, the redundancy adjustment factor is set to zero.

In this illustration, unsurprisingly, it is the one observation with the "critical" rating that receives the weight of 1 under Method 2. This is not necessarily always so with adjusted scores; when reliability scores vary substantially within a cluster, a "severe" rating based on a reliable piece of information may take away the place from a less reliable "critical" rating.

Both methods have their **pros and cons**:

- **Method 1** rewards a more differentiated report with greater effect on the aggregate measure than another report that produced few ratings. Its major drawbacks are the arbitrary exponent D and the dilution of high-severity situations when raters translate less critical aspects into numerous observations with low ratings.

- **Method 2** makes poor use of the total information, homing in on just one observation in the cluster that will contribute to the aggregate. It is, however, closer to the spirit of Possibility Theory:

The theory postulates that the possibility score of a union of subspaces in a state space is equal to the maximum of the scores of the subspaces. Such subspaces are easily seen in the above table. For example,

- the complete set of persons in the Department
- those exposed, sick or dead from Dengue fever
- only the dead from Dengue
- the injured from gunshot wounds,

etc. They are overlapping (in this particular cluster, the union is the entire Department population, but this need not be so in every cluster). Using the maximum, the cluster inherits the "Critical" rating from infant deaths from malnutrition. This approach is easy to implement in our demonstration workbook[5].

While it seems important to present both methods in outline, in the following we work with Method 2. At low administrative levels, its assumptions seem to be more plausible, overall, than those made in Method 1.

[5] On a finer point, the *necessity* value of the union of subspaces, too is the *maximum* of the subspace values. This implies in our binary "high" / "not high" situation, the "non-high" *possibility* score of the union is the *minimum* of the subspace values. The proverb "The chain is no stronger than its weakest link" expresses this intuitively. For the formal theorems, see nos. 5 and 6 in Table 20: Properties linking possibility and necessity measures in the appendix on page 43.

A final note on the redundancy adjustment. The six observations in our sample cluster underline the advantage of Possibility Theory. From the report, the raters distilled an observation on the subspace "Persons injured by gunshots". There is no observation on the complement "Persons not injured by gunshots". If the situation of this group had to be rated, it could not simply be "no problem"; many of these persons may suffer other, unobserved deprivations. Similarly, infants not (yet) dead from malnutrition were not observed for any complementary information. Possibility Theory makes sense of such incomplete information more readily than classic probability does – at the price of lesser confidence in the findings.

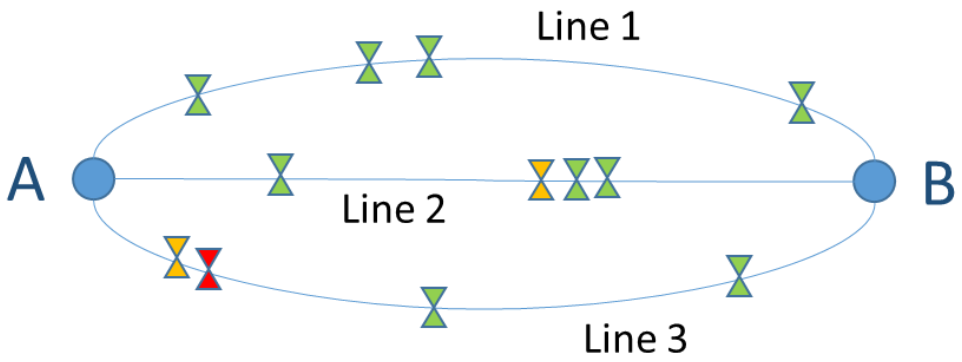
Aggregation

In order to obtain a severity measure on a higher-level administrative unit (e.g., a district), for a given sector, the adjusted possibility scores of the lower level units (localities) are aggregated. The assumption at this stage is that the humanitarian situations *between* localities are less strongly correlated than those of groups *within* localities. The observed localities, therefore, can be viewed as equally important for the aggregate. Then the arithmetic mean is a legitimate aggregation operator. The aggregate value is the weighted means of the reliability- and obsolescence-adjusted scores, with the redundancy factors as the weights.

[Sidebar:] The railroad network analogy

Two cities A and B are connected by three railway lines. Movement control on each line is aided by numerous sensors, which continuously report on a number of safety-relevant parameters. Green sensor values allow trains to move at normal speed. Yellow imposes reduced speed. Red signals a level of accident risk that suspends traffic on the line.

Figure 3: A railroad network with three lines



The possibility of an accident is ever present, not least for causes that the sensors do not cover. Within the current state of knowledge, however, it seems intuitive to measure the state of the line by the reported maximum risk, green for line 1, yellow for 2, red for 3.

How to suitably measure the state of the entire system calls for some more consideration. If the interest is in how speedily trains can move between A and B, an obvious candidate

is the minimum, over the three lines, of the maximum risks in each of them. If so, the system is at level “green” since on line 1 trains can move at normal speed throughout.

However, if we seek a measure of the average quality of the lines, then an averaging operator on the individual line maxima is appropriate. Since the sensors produce ordinal values, the median of the maxima is the appropriate statistic. The system as a whole is in state *median*(“green”, “yellow”, “red”) = “yellow”.

The aggregation method chosen for humanitarian severity ratings follows a similar logic. At the lower level (Admin2), we take the maximum, within each cluster, of the adjusted possibility scores for “the severity is high” (and the minimum for “not high”). This is the redundancy adjustment under Method 2, not yet the aggregation proper. The justification for the maximum is that the particular group rated highest may be a precursor of the severe conditions that await other groups in that locality.

At the next level – the pair of an Admin1 unit and a sector -, we take the arithmetic mean of the cluster maxima (not the median; the scores are ratio-level). At first sight, the mean operation seems to deviate from the theorems of Possibility Theory on combining the scores of multiple subspaces. For the union and intersection operations, the theory allows only minimum and maximum. However, at this level there is only one event space, not multiple subspaces – only the binary “high” and “not high”. We relate observations to it with the intent to measure *“the average quality”* of the Admin1-sector pair - not the possibility and necessity that the severity is high *anywhere* within the Admin1 unit, nor whether it is high *everywhere*. Therefore, all relevant possibility scores for the pair are equally important. The mean operator handles that.

There are mean functions for so-called fuzzy numbers that could be justified in Possibility Theory (Dubois, Prade et al. 1999). Their difficulty places them out of reach for this note.

The above, of course, is a very simplistic, purely allegoric, model of a train control system. A Google Scholar search for “Possibility theory” + “railway” + “train movement” returns some forty references, many of demanding mathematical sophistication.

Normalization and confidence

These are the last two steps to perform under Possibility Theory; the calculation of the gap scores does not need the theory.

On both sides of “high” / “not high”, the aggregate possibility scores will be in the open interval (0, 1). They have to be normalized (see page 16). The operation is simple: dividing by the larger of the two values.

The confidence that the severity in this Admin1 unit and sector is high, or not high, is based on the necessity values, which were introduced earlier. The confidence score is calculated using Formula 1, on page 17. Its range is the open interval (-1, +1); the exact values -1 and +1 are never reached because reliability < 1 for all observations implies aggregate necessities < 1. Mathematical purists would prefer a measure in (0, 1), with 0.5 as the point of undecidability. However, the (-1, +1) representation is more readily understood.

Calculation of the information gap measure

The information gap measure is a multiplicative index incorporating severity, age of the information, and size of the affected population. It is normalized to the half-open interval (0, 1], and in this final form is dimensionless.

- The severity enters as the confidence score. For the purpose of this index, we re-scale it to the interval (0, 1) (to avoid negative values when multiplied by the other two factors). The rescaling loses no information.
- The age of the information is summarily gauged from the date of the most recent cluster. The exponential decay function with the same half-life assumption is used as in the obsolescence adjustment. However, instead of multiplying by its value, we divide by it (we are not concerned with the diminished value of a piece of information, but with the size of the gap that the collection of information leaves open).
- The size of the affected population results from counts and estimates of persons in need. Alternatively, it is based on a proxy indicator, such as the host population. Its effect on the gap score can be tamed with a power function, with the exponent in [0, 1], depending on the desired bias to areas with smaller populations. In many countries these are the more rural and poorer ones.
- Normalization of the product of the three transformed variables is by division by the maximum value over all areas and sectors. By design, the maximum gap value will always be 1. The measure is relative.

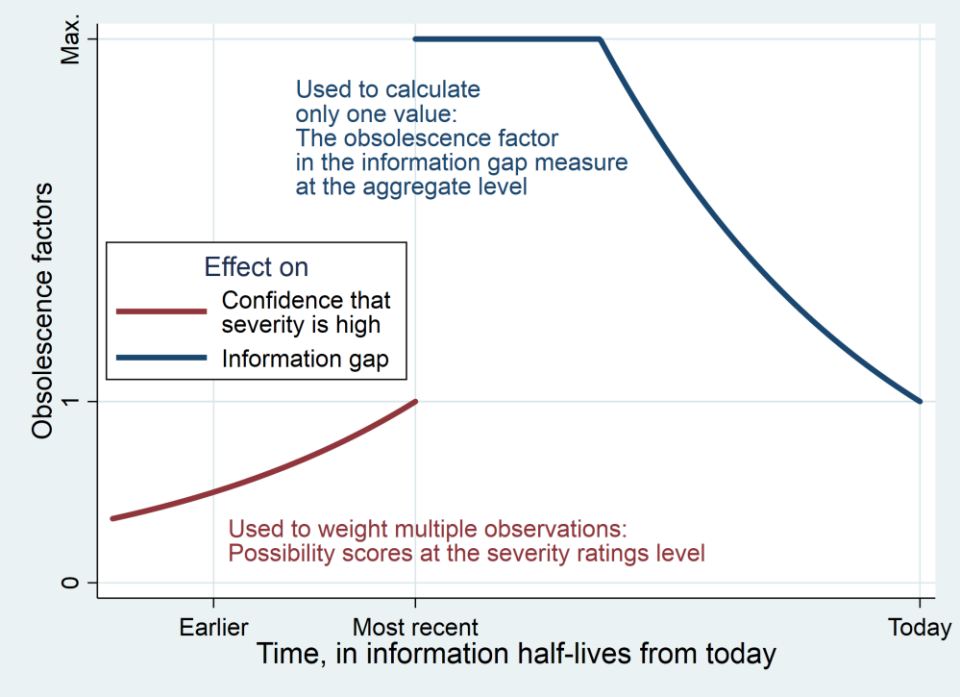
Some sectors have comparatively little assessment activity. In some areas, therefore, the periods since their latest activity tend to be much longer than in other sectors. Extreme values in one area and sector distort the distribution of gap values of most other areas and sectors. The simple exponential function is not optimal. Therefore, a permissible maximum is set for the obsolescence factor, as in

Formula 5: Ceiling on the obsolescence factor in the gap measure

$$\text{Max(ObsFactor)} = \text{Min}\left(\text{MaxObsolesc}, 1 / 0.5^{(\text{days since the most recent cluster / halflife})}\right)$$

The ceiling will prevent extreme values in the factor and therefore will not push the gap scores in most other areas-sectors down to levels where their uniformity would make them useless. Values for MaxObsolesc between 2 and 4 seem reasonable. This means that, going further and further backward in time, beyond 1-2 half-lives the obsolescence factor for the *gap measure* no longer increases. The calculation of obsolescence for the *confidence scores* is not affected by this. This diagram clarifies the two uses of the obsolescence factors.

Figure 4: Obsolescence factors



At this point, the gap measure is very experimental. The choices of population size exponent and obsolescence factor ceiling are hard to justify. The measure is purely relative; the way forward to an absolute one is not obvious.

Empirical demonstration

Data

Institutional framework

The data used in this demonstration workbook are the result of a collaboration between two organizations, specifically their joint work in Colombia. iMMAP is "an international non-for-profit organization that provides information management services to humanitarian and development organizations" (iMMAP 2021); it has actively been supporting humanitarian clusters for years. Data Friendly Space (DFS) is a US-based INGO focused on strengthening modern data systems and data science in the humanitarian and development communities (DFS 2021). Their work in Colombia was part of the *iMMAP COVID-19 Situation Analysis in Syria, Burkina Faso, Nigeria, DRC, Colombia, and Bangladesh, alongside broader global efforts* project. *COVID-19 Situation Analysis*[6]. It was funded by U.S. Agency for International Development (USAID) between July 2020 and June 2021.

The strategic objective of the project was to strengthen the assessment and analysis capacity of countries affected by humanitarian crises and the COVID-19 pandemic, by

[6] https://immap.org/global-covid-19-situational-analysis-project/ .

addressing issues related to comprehensiveness of data and information, data consistency, analytical value, historical data, timing, and focus. Unlike traditional epidemiological approaches with their narrow focus on disease dynamics, *COVID-19 Situation Analysis* worked in a multisectoral needs assessment perspective. It drew on best practices and analytical standards developed in recent years for humanitarian analysis.

Sources

IMMAP / DFS collected information from a large variety of sources. A listing of 114 main sources used in 2021 includes organizations and clusters of the United Nations system, international NGOs, Colombian government statistical services, humanitarian monitoring groups as well as national newspapers and media organizations of recognized prestige (iMMAP and DFS 2021)[7]. Those sources were complemented by monthly Web reviews using both generic and academic search engines. They were further deepened by regular interaction between iMMAP staff and personnel of the different humanitarian partners in Colombia, including the Inter-Agency Mixed Migration Flows Group (GIFMM) or the Humanitarian Country Team.

Resources

Personnel

The Colombia part of the project employed an in-country team of four iMMAP experts as well as six DFS counterparts based in Spain and Venezuela. The full teams were active between July 2020 and September 2021.

The DEEP database

The project's vital infrastructure was a database, analysis and visualization application known as DEEP[8]. DFS and Togglecorp[9] have been developing DEEP since 2016 as a "web-based platform offering a suite of collaborative tools tailored towards humanitarian crisis responses". DEEP is intended to support the analysis workflow across all its major stages, from design and planning, through data collection and preparation, exploration and analysis, interpretation and sense-making, to dissemination and feedback. In its current version, DEEP is most effective in the collection, preparation and exploration phases. Mapping and plotting components add spatial and temporal dimensions to data exploration.

Thus, DEEP essentially is a platform to facilitate and speed up secondary data review and analysis. Teams collaboratively identify relevant sources, extract snippets of information from sources and tag them by relevant categories. In the process, information of different nature is converted to a grid of records, observations and variables while maintaining access to the original sources. The information in a source document may be sparse or voluminous; it may refer to multiple entities that belong to different dimensions of humanitarian interest; the original formats may be text, statistical tables and images.

[7] The list is included in the demonstration workbook.

[8] Initially intended as an acronym for Data Entry and Exploration Platform, it has become a self-standing name. A complete guide on the use of DEEP, complemented by a series of video tutorials, can be found at https://deephelp.zendesk.com/hc/en-us.

[9] https://www.togglecorp.com/

The value added comes in the shape of categorical, date and text (summaries) variables. The former two provide a firm scaffolding for exploration and analysis; the latter aid both narrow verbal queries and the interpretation of broader findings. DEEP's assessment registry creates catalogs of available needs assessments by country, listing what was assessed, by whom, where, with type of respondents and sample sizes. Each catalogued report is assessed for quality, on criteria of purpose, trustworthiness, analytical rigor, analytical writing and analytical density. Observers of a humanitarian crisis can thus form an idea of the extent, quality and practical value of the assessments at hand.

A consortium of nine humanitarian organizations (IFRC, IMMAP, UNOCHA, UNICEF, OkularAnalytics, UNHCR, OHCH, JIPS, IDMC) manages DEEP. The Danish Refugee Council is the administrative host. By the end of 2020, DEEP hosted more than 1,200 humanitarian secondary data review projects involving 1,700 individual registered users worldwide. In 2021, DEEP was actively used in thirty countries to support strategic planning such as Refugees Response Plans or Humanitarian Needs Overviews.

In Colombia, DEEP has helped manage a large volume of information generated by the humanitarian consequences of the Covid-19 pandemic as well as of multiple longstanding conflicts inside and outside the country. From the humanitarian community, 56 individuals have direct access to the iMMAP-DFS DEEP project for Colombia; more than 1,300 documents were processed during the project (of which, as noted, 357 led to the severity ratings in this analysis). Situation reports built on DEEP data informed the 2021 Humanitarian Needs Overview (OCHA 2021). By June 30, 2021, the Colombia data had been looked up 2,300 times.

Only a minority of the 357 documents are reports with "needs assessment" in the title or abstract. Henceforward, therefore, we refer to "documents" while keeping in mind that they all contribute to needs analyses.

[Sidebar:] Historic precedents to DEEP

DEEP was not begotten in an eureka moment in one individual or organization. Rather, it grew out of cumulative developments in the humanitarian community. At their origin was the need for better coordination of needs assessments. The ancestor of DEEP's assessment registry was known as "Survey of Surveys", centered on Humanitarian Information Centers in the late 1990s and early 2000s in places like Kosovo and Angola. Nomenclature, strategies and hosting arrangements have since gone through some iterations, but motivation and inventiveness have endured. OCHA created its first template for "Who is assessing what where" entries in 2009; ACAPS did a review of "Survey of Surveys" efforts in 2011 (ACAPS 2011). In the following years the civil war in Syria strengthened the development of tools for this purpose. Turning from what information was there to what was missing, explicit attempts to visualize gaps in assessments were made in the Central African Republic in 2014 and were further developed in the response to the Nepal earthquakes in 2015 and 2016 (Benini, Chataigner et al. 2016).

The classic "an organization produces a report" style of needs assessment has never been supplanted, but has been significantly enlarged and enhanced by recent networking and information processing advances. In the humanitarian community's own terms, the

growing capacity for "secondary data review" and analysis is the key development to note here. It is not entirely clear what exactly its intellectual roots were. Before the year 2000, secondary data analysis in the social and medical sciences overlapped considerably with what statisticians nowadays call meta-analysis. As such, the analytic focus was rather narrow, dependent on variables that were already commensurate across datasets (such as persons with the same illness, comparable attributes, treatments vs. controls, a single outcome of interest). There may have been methodological stimuli radiating from social service studies and the like (e.g., Gaber 2000), but they must have been sparse and weak if at all noticeable to the humanitarian community.

By contrast, the kinds of secondary data review (SDR) and analysis that increasingly took hold of the humanitarian needs assessment community in the 2010s has tended to broaden the focus of what the analysts should consider. ACAPS (2014) defined an SDR as a "rigorous process of data collation, synthesis and analysis building on a desk study of all relevant information available from different sources such as the government, NGOs, UN agencies, media, social media, etc". A multi-sectoral outlook and the dearth of pre-existing commensurability in assessment designs favored this wide angle. Also, it was increasingly understood that "primary data during coordinated assessments in emergencies [was] not the main source of information, rather secondary data is the key information source during the initial days and weeks after a disaster" (ibd.). In 2015 the UN Inter-Agency Standing Committee made secondary data analysis a requirement in multi-sector rapid assessments, with an emphasis on pre-crisis collection to obtain useful baselines (IASC 2012).

Since then, with the continued better access to the Internet and thereby to public data, the growth of text- and data-processing tools, and a pool of analysts with greater skills in mixed (quantitative-qualitative) methods, the need for, and acceptance of, a comprehensive tool like DEEP, too have increased. Also, needs assessment information has grown more commensurate, by the adoption of common typologies (e.g., affected groups) and of common measures (e.g., severity scales), which in turn have made DEEP more productive.

Work flow

This paragraph details the work process in DEEP as far as it matters for the understanding of this note[10]:

1. Both the iMMAP and DFS teams identify and acquire relevant documents (e.g., assessment reports) and datasets (e.g., statistical tables published by governments). The principles of reliability, credibility and confidentiality guide the selection (ACAPS 2014).
2. When a document or a dataset is uploaded to DEEP, it becomes a *lead*[11] after the analyst assigns it a series of metadata and basic information tags. Tags are values of categorical variables selected from drop-down menus or from clickable cells in

[10] A complete guide on the use of DEEP, complemented by a series of video tutorials, can be found at: https://deephelp.zendesk.com/hc/en-us

[11] A lead is a document or data set that can be stored in DEEP in several formats (plain text, PDF, etc.), and it is the element hosting entries.

a DEEP analysis framework. This first categorization conforms the new additions to the common DEEP structure.

3. Once the *lead* has been initially categorized, the analyst begins the process of tagging the remaining information against the analytical framework designed for the particular project. These tags attach to selections from the raw text or to images in cases where the relevant information in the analyst's judgment is best captured graphically.
4. The lead is then broken down into *entries*. An entry is the basic autonomous unit of information in DEEP. It is a snippet of information from the lead tagged with pertinent elements of the analytical framework, notably the concerned humanitarian sectors, their dimensions and the operational environment. This completes the first level of tagging.

Table 7: Analytical Framework, segment

		Agriculture	Education	Food security	Health	Livelihoods	Nutrition	Protection	Shelter	WASH	Logistics	Analytical Outputs
Impact	Drivers & aggravating factors											Crisis impact:
	Impact on people											Humanitarian profile
	Impact on services and systems											Affected people
Humanitarian Conditions	Living standards											Severity of conditions:
	Coping mechanisms											Persons in need, by severity class
	Physical / mental wellbeing											
People At Risk	People at risk / vulnerable											Number of people at risk

The coders rated entries for impact, humanitarian conditions, and/or people at risk, depending on the information content, and then within each by dimension. The severity measure uses only the ratings of humanitarian conditions (area within dashed line). Attempts to supplement it with estimates of persons-in-need were abandoned over concerns of partner organizations. Adapted from iMMAP and DFS (2020:4).

5. Second-level tagging assigns to the entry information that is not necessarily tied to a particular analytical framework, but is still relevant. The attributes to be tagged at this level are geographic locations, population groups affected and their demographic characteristics, and the presence of people with special needs.
6. Each entry receives a reliability rating. Ideally, the rating is the result of evaluating the source on four criteria: motive for bias, technical expertise, track record for accuracy, and method. In countries with a rich supply of good data, the process can be simplified algorithmically, such as giving a higher rating to information that originates from known specific lower administrative units, and a lower rating for summary information on higher units.
7. When the entry relates to living standards, coping mechanisms or physical and mental well-being among affected groups, the analyst adds a severity rating. The

rating is on an interpreted five-level scale. Its primary function is to filter entries on the urgency of humanitarian interventions. In the context of this note, the severity ratings are the basis on which information on humanitarian conditions is aggregated to a quantitative measure applied to pairs of Admin1 units and sectors.

Table 8: Severity levels - Names, meanings, numbering

Level	1	2	3	4	5
Severity	**No problem** or minor problem	Of **concern**	**Major** problem	**Severe** problem	**Critical** problem
Implied urgency	No intervention required	Monitoring required	Middle-term intervention required	Short-term intervention required	Urgent intervention required

8. Quality control: a dedicated DFS team member is responsible for an ongoing review of the quality of the tagged content and its compliance with the proposed analytical framework. Every lead is validated by a quality controller[12].
9. A defined set of validated leads and its associated entries can be exported to an Excel workbook. It comes with two sheets. In sheet "Grouped entries", every entry, regardless of the number of its locality, sector, affected group, etc. tags, occupies one row. In "Split entries", a row is created for each combination of tags. Thus, the fictitious entry "In *municipalities X and Y, food insecurity* among children is aggravated by the lack of *school feeding* programs" would dissolve into 2 localities * 2 sectors = 4 split entries. The split entries sheet, data-wise, is the departure point for our analysis outside the DEEP.

Working dataset

Entries, leads and clusters

The dataset exported from DEEP to Excel comprises 116,114 observations (split entries in the above terminology). However, by far not all are suitable for the calculation of the severity and information gap measures. The observations finally retained resulted from a multi-step exclusion process:

[12] Given the importance of personal judgment in a distributed network of coders, the validation process in DEEP merits brief explanation:

> For every lead added to DEEP a quality controller reviews all the entries and, if satisfied, marks the lead as "validated". Every workday, the quality controller picks a small sample of entries from different analysts and checks that the severity has been scored homogeneously (across the information reviewed). When disparities are manifest, a consultation with other quality controllers and with senior analysts is held. Once clarity about the interpretation of tagging criteria is reached, they are shared with the concerned analysts. The refined guidelines are incorporated in a quality control master sheet available for all coders working under the same framework. In addition, quality controllers periodically arrange tagging sessions with several analysts. These tag the same leads independently. The variations are discussed, and final versions are decided, in plenary (Skype calls).

Table 9: Sequential reduction of dataset

Reduction of split entries	Excluded	Remaining
Exported from DEEP		**116,114**
Has no severity rating	53,270	62,844
Rating is not about humanitarian conditions	24,601	38,243
Sector subdimension is not about coping mechanisms, living standards, or physical and mental well-being	1,665	36,578
Rating refers to cross-sector	1,028	35,550
Has no Admin1 (i.e., Department) code	10,630	**24,920**
Of the remaining observations,		
with both municipality and department codes		9,592
with department codes, but without municipality code		15,328
Working dataset		**24,920**

The retained entries span the 409 days from May 18, 2020 to June 30, 2021. They are derived from 357 uploaded documents and datasets (the leads). They are about all 33 Departments (Admin1) in Colombia and explicitly touch upon 251 of the 1,122 municipalities (Admin2) (Wikipedia 2021a). Humanitarian conditions in the remaining 871 municipalities were indirectly captured in Department-level entries.

The cluster is a pivotal analytic construct for this analysis. It is the basis for the redundancy adjustment to the possibility scores. It is not part of the DEEP. As discussed and illustrated on pages 20 - 23, a cluster is the set of observations with identical locality, sector and publication date. The working dataset holds 2,477 clusters; the mean cluster size is 10.1 observations. This distribution is highly skewed (median = 3, min = 1, max = 323 entries). The largest clusters are composed of Department level-only entries, e.g., Cauca has a Department-level cluster with as many as 323 entries, compared to a maximum of 44 entries among municipality-specific clusters in that Department.

Elsewhere in the country, some municipality-specific clusters are large. The Federal District of Bogotá has a cluster with 196 entries. Large clusters are problematic in the aggregation of possibility scores. With a weighted mean operator for the cluster, the value passed on to the Department-sector pair aggregation is likely too low to send a clear signal about the presence of severely affected groups – the less severely affected ones will dampen down the aggregate value too much.

The maximum operator avoids that. Rarely it may tend in the other direction, meaning: its contribution to the aggregate is too high. This happens when the cluster contains relatively small proportions of severe or critical ratings that may speak to an unknown, but similarly

small proportion of persons in need[13]. Still, because it errs on the side of high severity, the maximum operator is preferable.

Select results for all 33 Departments

Time until new information

As noted, there are 2,477 clusters - distinct combinations of locality, sector and publication date. When we distinguish observations by sector and publication date, but replace locality by the respective Department, 1,925 distinct combinations remain. This is the number of occurrences when new information about any Department-sector pair arrived during the observation period.

The length of time between arrivals in the same pair is a statistic of interest to characterize the dynamic of needs assessments. However, we cannot consider first occurrences because they most likely were preceded by others before our observation period (the data are so-called left-censured). Thus, excluding the interval from beginning of observation period to first occurrence in every pair, 1,678 subsequent arrivals of new information occurred (1,925 total arrivals – 247 pairs).

The 1,678 intervals vary greatly in length, from 1 to 252 days (mean = 20.5; median = 12.5). Not surprisingly, the mean intervals by sector are of unequal length, with sectors with numerous occurrences showing shorter inter-arrival times within department-sector pairs, and sectors with scant activity taking longer on average. The mean and standard deviations by sector are therefore correlated, and the only statistic of non-trivial interest, besides the number of occurrences, is the coefficient of variation, the ratio of standard deviation to mean. Sectors with relatively low CoVs – education, nutrition, logistics – seem to have been updated at more regular intervals across all departments in which they had assessment activity. Others like food security and protection, although their mean intervals were much shorter, had considerably higher variation.

[13] For instance, in the Department of La Guajira, six communities, including the sizeable cities of Riohacha and Uribia, each have a cluster of 136 entries. These are part of an assessment of the situation of indigenous and other minorities in the WASH sector; the entries are derived from one lead, the assessment report. Of the combined 816 severity ratings, 74 percent are below the level "Severe", which is the highest used in this subset (i.e., zero "Critical" ratings). The rating distribution is identical for all six communities. This raises the question whether the municipality-specific ratings really add value over global Department-level ratings, or whether they should be assigned the same lower reliability scores as for the Department level.

Table 10: Days until new information arrived on Department-sector pair

Sector	Occurrences (except first)	Min	Median	Mean	Max	C.o.V.
Agriculture	2	13	61.5	61.5	110	1.12
Education	71	1	42.0	42.7	162	0.79
Food security	149	1	23.0	31.9	252	1.21
Health	388	1	12.0	17.1	97	0.98
Livelihoods	294	1	14.5	18.6	178	1.07
Nutrition	10	6	47.5	64.0	180	0.79
Protection	517	1	6.0	11.5	91	1.26
Shelter	87	1	21.0	30.0	168	0.99
WASH	114	1	21.0	27.9	152	0.94
Logistics	46	1	41.0	44.9	116	0.74
Total	**1678**	**1**	**12.5**	**20.5**	**252**	**1.21**

Severity ratings and possibility scores

The effect of administrative levels

The overall distribution of severity ratings has two peaks, at the "major" and "critical" levels. The latter is pushed up primarily by Department-level ratings. The median level for municipality-specific ratings is "major"; for the Department-level it is "severe".

Table 11: Severity ratings, by administrative level

Severity	Administrative level		Total
	Municipality	Department	
No problem	344	293	637
	3.6%	1.9%	2.6%
Of concern	2,172	3,250	5,422
	22.6%	21.2%	21.8%
Major	4,532	3,511	8,043
	47.3%	22.9%	32.3%
Severe	1,271	2,385	3,656
	13.3%	15.6%	14.7%
Critical	1,273	5,889	7,162
	13.3%	38.4%	28.7%
Total	**9,592**	**15,328**	**24,920**
	100.0%	**100.0%**	**100.0%**

The reasons for this bias are speculative. Plausibly, coders working with municipality-specific documents detect and rate greater variability whereas Department-level sources induce them to go by the more severe elements.

A more informative comparison relies on the adjusted possibility scores for "severity is high" vs. "is not high". The difference between these two variables, once they are aggregated to Department-sector pairs, determines the confidence that the true severity is high, resp. not high.

Table 12: Difference "high" - "not high" adjusted possibility scores

Admin. level	Clusters	Mean	St.dev.	Min.	Max.
Municipality	1,166	-0.15	0.48	-0.81	0.82
Department	1,311	0.06	0.43	-0.67	0.71
All	**2,477**	**-0.04**	**0.47**	**-0.81**	**0.82**

Note: Using one instance of max. "high" score – min. "not high" score per cluster

This reinforces the impression that there are opposite biases at work depending on the administrative level of the ratings. It hardly matters; in the global statistic, the two appear to almost cancel out, with an absolute mean difference of 0.04, about 1/12 of the standard deviation. This is reassuring about the reliability.

The confidence that severity is high, is not high

To aggregate the possibility scores (adjusted for reliability, obsolescence as well as redundancy), they are averaged for each Department-sector pair. Remember that with the maximum-based redundancy control, only one value per cluster is > 0 (see Table 6 on page 21, Method 2). The arithmetic mean is calculated, on both sides of "high" / "not high", as

Formula 6: Aggregate possibility score, before normalization

Unnormalized aggregate score =
sum(adjusted poss. scores in given Dept.-sector pair) /
Number of clusters in that pair

The aggregate scores are then normalized (normalization was introduced on pages 16-18). The larger of the two means becomes 1, signifying that this side is fully possible. The smaller value is increased by the same factor and by design remains <1. At this point, the key quantity for the severity score is extracted, the necessity value, equal to (1 – the smaller of the two normalized scores). In pseudo-Excel notation:

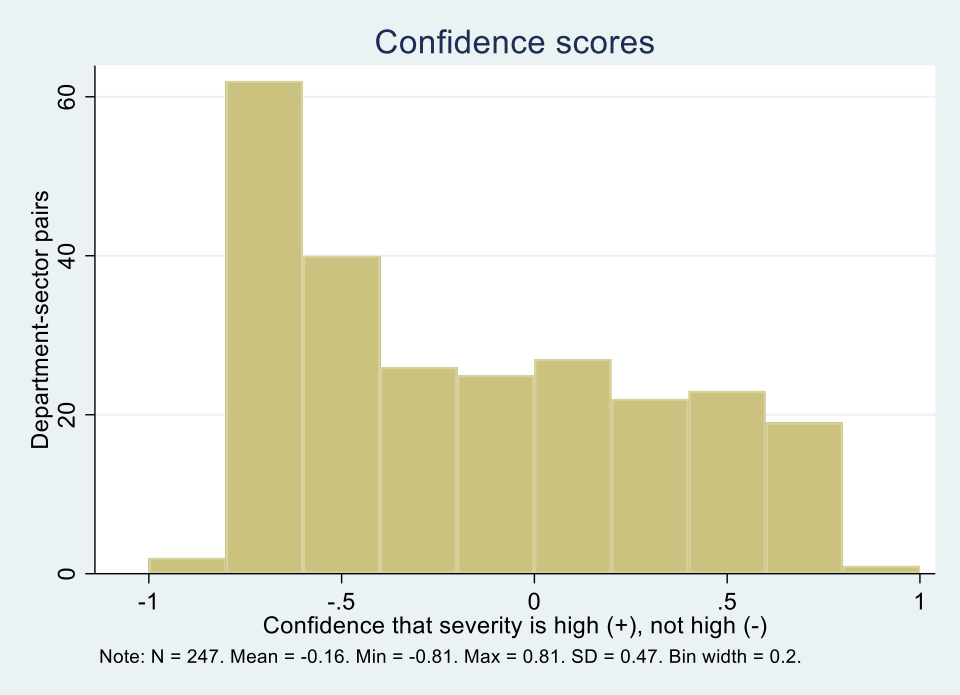
Formula 7: Confidence score, Excel notation

Confidence score = IF(possHigh = 1,1- possNotHigh, -(1-PossHigh))

The confidence scores are in the interval (-1, +1). Mathematically, this is awkward, but it makes for easier visualization, with the extremes denoting near-certainty, and 0 expressing ignorance.

Ratings occurred in 247 of the 33 Departments * 10 sectors = 330 pairs. The global distribution of confidence scores tilts towards the "severity is not high" side:

Figure 5: Distribution of confidence scores



The table of confidence scores for all 247 pairs is included in the workbook sheet "Confidence_33Dept". The spatial distribution, exemplified for one sector (livelihoods), is shown in the left panel of the map in the Summary.

Demonstration workbook

Purpose and scope

The workbook demonstrates an arrangement of connected worksheets, named ranges and formulas that produces the desired aggregate measures of severity and information gaps. Its structure is suitable to translation to a dedicated application like DEEP, eventually with additional features. These might include a variety of customizable tables, maps and graphs. At this point, our purpose is to let the users understand the internal mechanics as well as the parameters that embody key assumptions. Users can vary several of the parameters and see the effect of variation on the aggregates side by side in the same worksheet.

For file size and calculation time reasons, the workbook is limited to a subsample of the working dataset. It uses a purposive sample of six out of the 33 Colombian Admin1 units, the Departments. The number of observations (severity ratings) drops from 24,920 in the full dataset to 10,624 in this subsample. The observation period remains the same, 18 May 2020 – 30 June 2021. As before, the key input variables are: Location, sector, date published, 5-level severity rating, plus (for the information gap part) the Department's 2020 population projection imported from outside the DEEP. A further key input, the

reliability scores given the individual observations, is a calculated variable, determined by the observation level (municipality vs. Department), in the same way as in the full dataset.

The main interest is not in the distributions of the inputs, but in the calculated variables. All sectors from the full dataset and all rating levels are represented, with a numeric dominance of the protection and WASH sectors.

Table 13: Severity ratings by sector and level, demonstration sample

Sector	No problem	Of concern	Major	Severe	Critical	Total
Agriculture	0	0	8	1	0	9
Education	12	160	206	23	0	401
Food security	11	129	135	238	21	534
Health	29	121	521	220	83	974
Livelihoods	4	246	659	132	12	1,053
Nutrition	0	0	1	17	6	24
Protection	39	409	994	667	1,777	3,886
Shelter	216	474	222	92	2	1,006
WASH	59	1,324	844	373	11	2,611
Logistics	20	47	49	9	1	126
Total	**390**	**2,910**	**3,639**	**1,772**	**1,913**	**10,624**

These ratings result from 265 documents (the leads). Grouped by identical location, sector and publication date, the ratings form 1,042 clusters. Just over half of the ratings (51 percent) are at the municipality level. In order to produce uniformly formatted cluster codes, an artificial municipality code was created for observations at the Department level, adding "000" to the Department code, as in, e.g., "CO05000" for Antioquia.

The selection of the six Departments is purposive, with three pairs each selected for a particular set of humanitarian problems.

Departments	Humanitarian characteristics
Bogotá / Antioquia	Host the two biggest cities. Large numbers of both IDPs and refugees. Similar needs linked to urban context, and similar access to services.
Norte de Santander / Arauca	Close to the Venezuelan border. Many migrants in transit. Historically affected by conflict.
Chocó / La Guajira:	Mostly rural. High proportion of Black/Indigenous communities. High poverty rates. Deficient infrastructure.

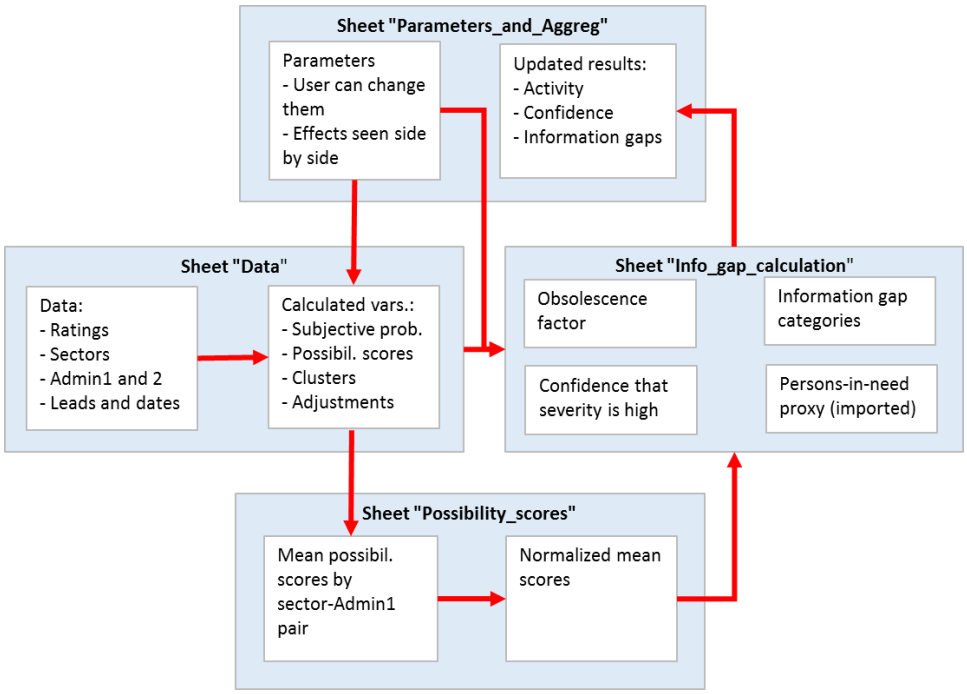
Workbook structure

The essential calculations are carried out in the four sheets shown in this diagram. The red arrows show the flows of parameter values and of the data among sheets. Flows within sheet "Info_gap_calculation" are not shown; the arrows would make the diagram unwieldy.

Users can grasp the flows by looking at the various tables in the sheet, arranged from left to right, with formulas explained above the tables.

While "Info_gap_calculation" is, so to speak, the back office of the workbook, its essential outputs are dynamically reflected in "Parameters_and_Aggreg". In this sheet, users see the aggregates of interest updated in response to changes that they make in parameters, and can compare the updated values to those under parameter defaults.

Figure 6: Connections between worksheets



Three additional sheets help the understanding, but no calculations of consequence take place there. "Sample" details, as already earlier in this note, the steps reducing the raw DEEP extract to the working sample. "Variables" adds to the short variable names used in "Data" the fully written-out labels, the counts of non-missing observations as well as, for binary and continuous variables, their means. Scrolling to columns 10 – 16 and thence downward reveals two text boxes that the detailed-oriented user may want to read. Finally, "Named_ranges" offers a convenience table of all such ranges.

Definition of parameters

"Parameters_and_Aggreg" gives each parameter with its functional description, short name, value(s) and destination column in "Data". Some parameters specify more than one value in lookup tables.

For the severity measure

Table 14: Subjective probabilities

To calculate the raw possibility scores			
Subjective probability (belief strength) that problems are severe or critical	Severity short	Ordinal severity level	Probability
[Lookup table]	No problem	1	0.05
	Of concern	2	0.10
	Major	3	0.25
	Severe	4	0.90
	Critical	5	0.95

The probabilities must be strictly monotonously increasing along the severity levels. There is little gain in changing them. The key to a well-discriminating confidence measure is setting the probabilities in levels 1 – 3 far below 0.5, and those of levels 4 and 5 far above it. As argued on page 16, the setting for the central level ("Major") should be well below the 0.5 midpoint.

Table 15: Reliability adjustment

To adjust for reliability			
depending on admin. specificity of severity rating			
[Lookup table]		DeptLevel (col 6)	Score
Is record at Department level?	Yes	1	0.75
	No, at municip. level	0	0.90

"Is the rating for an observation at the Department or the municipality level?", or more generically, at the Admin1 or Admin2 level. Experimenting with the reliability scores is meaningful; plausibly they should by above 0.50, and the more specific municipality level score above the Department level one.

Table 16: Obsolescence adjustment

To adjust for obsolescence		
Half-life of information (in days)	halflife	120
End date of observation period	enddate	6/30/2021

The half-life of information (in days) and end-of-observation-period settings work for both measures, confidence and information gap. The end date is fixed; the half-life is open to experiments (120 days is the initial setting). A longer 180 days may be appropriate when the primary concern is for agencies to update their operational environments in six-months funding cycles. In the perspective of faster changing policies (e.g., Covid-19, nutrition surveillance), information has shorter half-lives, perhaps as short as 30 days.

Redundancy adjustment

Redundancy adjustment using the maximum operator on the reliability and obsolescence-adjusted severity scores does not need additional parameters. In their lieu, the cluster tag variable (column 47) is used to ensure that from all observations with the highest adjusted

scores in a given cluster, exactly one receives a weight of one, and all other observations in the cluster have zero weights[14].

For the information gap measure

Table 17: Additional parameters of the information gap measure

To adjust for obsolescence		
Ceiling on obsolescence factor (informtion gap measure only)	MaxObsolesc	3
Population senitivity		
of information gap measure	popexpon	0.2

with MaxObsolesc as explained on page 25, with plausible values between 2 and 4. popexpon is legitimate in [0, 1], 1 produces the full population effect on the measure, and 0 means that the Departments are treated equally regardless of population size.

Formulas

The calculation of adjusted observation-level scores in "Data", of the aggregates in "Possibility_scores" and of the tables in "Info_gap_calculation" relies on the extensive use of a small number of Excel features. Besides Excel's standard functions, they include named ranges and array functions.

All data columns in "Data" are named by their short variable names in row 2. For example, "clust_size" is the short name of the variable "Cluster size" as well as of the range of cells "Data!R3C48:R10626C48", which holds all 10,624 values below "clust_size".

Array functions are recognized by the curly brackets that surround them. They are entered by pressing Shift + Ctrl + Enter together. For example, the identical formula for the unnormalized possibility scores in every cell in "Possibility_scores!R5C3:R10C12"

Formula 8: Excel array function, syntax example

```
{ =IFERROR(AVERAGEIFS(possHighRedund, dep_code,RC2, Sector,R4C, clust_tag,1),"") }
```

makes use of

- the standard functions `IFERROR` and `AVERAGEIFS`,

[14] Redundancy adjustments using the mean operator would be more involved. They would need a weighting parameter that affects all observations in the given cluster equally. One possibility would be to use an exponent on the cluster size (number of observations in the cluster), *clSizeExp* in [-1, 0], to create weights

$$w_{redundancy} = cluster\_size^{clSizeExp}$$

clSizeExp = 0 means that all observations across clusters carry a weight of one, and -1 means that every cluster, regardless of size, counts as one observation only. A middling value, such as -0.5, balances both concerns – maximum use of observations and penalty for redundancy.

In earlier experiments using the mean operator for redundancy adjustments, we found that the resulting confidence distribution was less discerning between Department-sector pairs with known high severity and others than the results of the maximum operator. Therefore, we do not pursue this alternative in this note.

- the named range `possHighRedund` as the variable to be averaged,
- the named ranges `dep_code`, `Sector` and `clust_tag` as the IF-variables,
- for which the mixed-type cell references `RC2` and `R4C` and the constant 1 point to the values that, in combination, determine the observations to be included.

Repeated use is made also of another array function type when standard functions are not available for the purpose. Excel 2016, for example, does not offer `MAXIF` or `MAXIFS` functions for conditional maxima, in analogy to `AVERAGEIFS`. Such functions are needed, among others, in columns 76 and 77 in "Data" for the maxima of the (reliability and obsolescence adjusted) possibility scores in each cluster. Thus, in column 76 we find a work-around with

Formula 9: Array function syntax, with IF-clause

```
{ =MAX(IF(cluster=RC45,possHighObsoles)) }
```

where

- the maximum of `possHighObsoles` has to be found in each cluster,
- for which the cluster ID is held in in the same row in column 45 (`RC45`),
- and passed identical to all cells in column 76 belonging to that cluster.

The key to this construction is to write the operator on the substantive variable (`MAX`) first, followed by `(IF(conditioning variable range = value to use, substantive variable range))`. It fills certain gaps in Excel standard functions, but easily gets complicated when more than one condition are to be met. Wherever standard functions fill the need, they are preferable.

With these special feature descriptions in mind, mid-level Excel users should be able to find their way through the demonstration workbook.

Conclusion

Together, demonstration workbook and companion note provide a first proof of concept for these related ambitions:

- Large collections of severity ratings can be aggregated, despite challenges from variable reliability, obsolescence and redundancy.
- Possibility Theory offers a basis for an aggregation algorithm that takes account of those challenges.
- The key output is a (continuous) measure of confidence in binary hypotheses of the kind "The severity in Admin1 unit X and sector Y is high / is not high".
- The algorithm can be implemented in Excel, at a moderate level of complexity (no VBA programming!) accessible to mid-level users.

In addition, building on the confidence measure and on the obsolescence of information, and outside the purview of Possibility Theory, a measure of information gap at the same aggregation level has been proposed. This is still very experimental and sample-dependent, with the largest gap among the Admin1-sector pairs dictating the levels of all others.

The workbook invites the user to experiment with different parameter settings. The calculations are relatively slow; on the main author's computer each change would take between 5 and 10 seconds to be reflected in the results section. Clearly, there is a case to make the back office part invisible and running in a faster application, with a friendly interface for user input and key outputs. In fall 2021, a colleague of ours (Matthew Smawfield) made a first experiment on a Java-based platform, with identical values of the confidence measure produced in a fraction of a second. DFS is discussing the possible use of the algorithm, with additional features, in DEEP.

At the same time, the limitations of our approach to aggregating large numbers of severity ratings are considerable. The reliability measure, considering only the administrative level, is coarse. The severity scale at the unit level is plausible, but the aggregate of interest – the confidence measure – has not been validated. At present, population estimates of any kind do not go into the calculation of the severity measure; and only Department-level host population figures enter the gap measure, as a proxy for persons in need.

Looking into the future, the measurement of severity would become more accurate if assessments of large-city municipalities were specific of major subdivisions. Other improvements will be possible when municipality-level estimates of persons in need become available.

Meanwhile, our hope is that readers will be stimulated to consider applications of Possibility Theory in their own work, to improve on it for this particular application, or to advance better ways to aggregate large numbers of severity ratings by other means.

[Sidebar:] Is there a cheaper alternative at hand?

Some readers will be uncomfortable with the transformation of a five-level scale to a binary hypothesis. The severity scale, as described, is an *interpreted* measure, with meanings defined at each level. "Severity is high / is not high", by contrast, is interpreted as stark extremes. These are not directly observable. What is "high"? What is "not high"?

Moreover, the scale interpretation is rooted in a realist philosophy. This is understood best by considering the central category, "major problem", with the implication that some "middle-term intervention" will be required. The uncertainty that goes with it is in the reality – whether the intervention will happen and whether, if it doesn't, conditions will further worsen. It is not in the judgment of the experts who rate the problem as "major".

Technically, it is feasible, though not straightforward in Excel, to create unit-level combined weights of reliability, obsolescence and redundancy and use them in order to calculate weighted rank-order statistics, such as the weighted median severity level for any Admin1-sector pair. In the same vein, if one wanted to err on the side of prudence, a weighted percentile (75 or 90) would be achievable. This method produces results in terms of the five-level scale and is correct for ordinal data.

This table illuminates the change in weighted percentiles of severity levels, where levels are numbered as in Table 8 on page 31, and observation-level weights are the products

of reliability score (column 70 in sheet “Data”), obsolescence weight (col 73) and cluster tag (col 47).[15].

Table 18: Weighted percentiles of severity levels in the food security sector

Department	Clusters	Ratings	Weight sum	Median	75th	90th	Max
Antioquia	21	45	9.94	4	4	4	4
Bogotá, D.C.	7	42	3.71	3	4	4	4
Chocó	21	83	9.86	4	4	5	5
La Guajira	16	164	7.03	4	5	5	5
Norte de Santander	16	134	8.57	3	4	4	5
Arauca	15	66	8.67	3	4	4	5
Total	**96**	**534**	**47.78**	**4**	**4**	**5**	**5**

However, three objections arise, with increasing seriousness in this order:

- Differently from the handling of obsolescence in our application, this rank order-based aggregation method factors obsolescence at the aggregate level strictly from the end-of-observation date, not from the latest observation. Admin1-sector pairs with a long interval since the latest observation will be underweighted.
- The choice of percentile is entirely arbitrary. Which should we take?
- While this method captures unpredictability in the objective conditions, it does not consider the uncertainty in the expert judgments that produce the severity ratings. In particular, “major problem”, the most frequently used level in the dataset, while cautious in individual ratings, does not help to clarify the true severity in the aggregate.

Possibility Theory helps us to mitigate those drawbacks. It does so by sharply distinguishing the subjective probabilities assigned to the five levels that the true severity is high. For example, the probability of 0.25 set for “major problem” implies that on average the expert choosing that level is correct in 3 out of 4 times that the severity is not high. In one out of 4, he/she misses a situation of truly high severity. The theory works with the dual measures of possibility and necessity. The possibilities point to the more plausible side of the binary; from the necessities we derive the confidence that, in the Admin1-sector pair of interest, the true severity is high or not high. The theory captures both uncertainties, in the reality and in its observers.

---

[15] Calculated in Stata. Akinshin (2021) discusses several methods and rejects the Wikipedia article and several R and Python implementations as faulty (Wikipedia 2021b). Members of the “Excel Forum” may download a spreadsheet at the bottom of the thread https://www.excelforum.com/excel-formulas-and-functions/1294249-weighted-median-for-large-dataset-with-many-varying-weights.html , but we have not examined this.

Appendices

Further notes on Possibility Theory

Status of the theory

Building on earlier sources, the French mathematicians Didier Dubois and Henri Prade elaborated Possibility Theory starting in the 1980s (Dubois and Prade 2015). The short Wikipedia article, which spells out definitions and some theorems of possibility and necessity, calls the theory an "imprecise probability theory" (Wikipedia 2020). Bronevich and Klir (2010) compare the axiomatic bases of possibility and probability theories. In a perspective on a wider variety of uncertainty theories, Klir (2006) anchors Possibility Theory in a tradition that *predates* probability. While these references are of interest chiefly to mathematicians and are of no direct practical value in our context, they underline the seriousness of the theory as an intellectual heavyweight rather than an arbitrary fix in struggling with difficult data. A segment from a table in Klir (op.cit., 159) contrasts the two theories on some of their basic properties.

Table 19: Probability vs. Possibility Theories

Probability Theory	Possibility Theory
Based on measures of *one type*: probability measures, *Pro*	Based on measures of *two types*: possibility measures, *Pos*, and necessity measures, *Nec*
Body of evidence consists of *singletons*	Body of evidence consists of a *family of nested subsets*
Unique representation of *Pro* by a *probability* distribution function $p: X \to [0, 1]$ via the formula $Pro(A) = \sum_{x \in A} p(x)$	Unique representation of *Pro* by a *basic possibility function* $r: X \to [0, 1]$ via the formula $Pos(A) = \max_{x \in A} r(x)$
Normalization: $\sum_{x \in X} p(x) = 1$	Normalization: $\max_{x \in X} r(x) = 1$
Additivity: $Pro(A \cup B) = Pro(A) + Pro(B) - Pro(A \cap B)$	Max/Min rules: $Pos(A \cup B) = \max \{Pos(A), Pos(B)\}$ $Pos(A \cap B) \leq \min \{Pos(A), Pos(B)\}$ $Nec(A \cap B) = \min \{Nec(A), Nec(B)\}$ $Nec(A \cup B) \geq \max \{Nec(A), Nec(B)\}$
Not applicable	Duality: $Nec(A) = 1 - Pos(\bar{A})$ $Pos(A) < 1 \Rightarrow Nec(A) = 0$ $Nec(A) > 0 \Rightarrow Pos(A) = 1$

where A and B are subsets in the relevant event space, and $\bar{A}$ is A's complement.

Possibility Theory has not been without some fundamental critiques. These come from two opposite sides:

- Defenders of classic probability theory argue that the duality of possibility and necessity makes it difficult to interpret results in practical decision-making (Aven 2011, Flage, Aven et al. 2014), and this would then also be true of measures derived from them like confidence.
- Some adherents of imprecise probability consider Possibility Theory suitable for incomplete, but nested evidence, but not versatile enough to fuse evidence collected from disjoint and overlapping state subspaces (Kikuchi and Chakroborty 2006).

The argument *against* the former group of critiques is that the choice of Possibility over probability theory is not capricious, but dictated by the kind of evidence at hand. The latter group has a strong pragmatic argument *for* it: The versatile Dempster-Shafer Theory (DST) (Shafer 1976, Wikipedia 2011, Hensher and Li 2014), the main rival to Possibility Theory, has been implemented in the statistical application R (in packages like DST, evclust, ipptoolbox, ibelief). Hardly any are available with functions in the direct ambit of Possibility Theory (with the exception of a few in the FuzzyNumbers package, which are irrelevant for our situation). However, the prerequisites for DST to work with ordinal variables transformed and adjusted as in our situation are unknown.

Literature

For readers seeking a deeper understanding without entering Klir's theoretical labyrinth, Solaiman and Bossé (2019) contributed the first textbook-length treatment. It is a demanding text, of uneven didactic quality and geared towards applications far from humanitarian concerns, e.g. automated image recognition. The book is valuable as a detailed introduction to the theory, and in particular to mathematical tools needed to work with possibility distributions over *continuous* variables. On page 32, it offers a handy list of relevant properties linking possibility and necessity measures (for normalized possibility distributions):

Table 20: Properties linking possibility and necessity measures

1. $\Pi(A) \geq \mathcal{N}(A)$;
2. $max\{\Pi(A), 1 - \mathcal{N}(A)\} = 1$;
3. $\Pi(A) + \Pi(A^C) \geq 1$, and $\mathcal{N}(A) + \mathcal{N}(A^C) \leq 1$;
4. $max\{\Pi(A), \Pi(A^C)\} = 1$, and $min\{\mathcal{N}(A), \mathcal{N}(A^C)\} = 0$;
5. $\Pi(A \cup B) = max\{\Pi(A), \Pi(B)\}$, and $\mathcal{N}(A \cap B) = min\{\mathcal{N}(A), \mathcal{N}(B)\}$;
6. $\Pi(A \cap B) \leq min\{\Pi(A), \Pi(B)\}$, and $\mathcal{N}(A \cup B) \geq max\{\mathcal{N}(A), \mathcal{N}(B)\}$;
7. $\Pi(A) + \mathcal{N}(A^C) = 1$, and $\Pi(A^C) + \mathcal{N}(A) = 1$;
8. $\Pi(A) < 1 \Rightarrow \mathcal{N}(A) = 0$, and $\mathcal{N}(A) > 0 \Rightarrow \Pi(A) = 1$.

The notation is different from Klir's (op.cit., above). Here Π, the Greek capital letter Pi, stands for possibility $\mathcal{N}$ for necessity, and A^C for the complement of A.

Key inspiration and guidance for our application comes from a small paper on semi-automatic possibilistic information scoring (Lesot, Delavallade et al. 2011, Lesot and d'Allonnes 2017). The authors contribute two key elements:

- The treatment of evidence in support of a strictly binary hypothesis such as "The true severity of conditions in region X and sector Y is high" vs. "is not high", including the arithmetic mean as the aggregation operator on the possibility profiles of the relevant evidence pieces (followed by normalization of the aggregate possibilities);
- The necessity to adjust, at the observation level, for reliability, obsolescence and redundancy, including the function for the reliability adjustment that we use (but without specific guidance on the other two adjustments, for which we eventually chose functions on our own that seem plausible for our purpose).

We note that the important concept of "cluster" – as the set of observations that share the same locality, sector and publication date – is not from Lesot et.al., nor from anywhere in Possibility Theory. We borrowed it from sampling theory, which applies it to observations that are not statistically independent. In our application, the clusters are primarily used in the redundancy adjustment. Similarly, the maximum operator that we chose for that step, while superficially consonant with the a.m. $\Pi(A \cup B) = max\{\Pi(A), \Pi(B)\}$ rule,
is dictated by the desire to err on the side of caution (or pessimism) about future developments in the concerned locality and sector, and only marginally by Possibility Theory.

Holst and Lohweg (2020, op.cit.), in the context of multi-sensor systems, develop an algorithm for redundancy adjustments, but the maximum operator we chose is preferable for its simplicity.

To obtain the raw data on all 33 Departments

The data used in the demonstration is part of the Excel workbook.

The full dataset (all of Colombia's 33 Departments for the same observation period) from which the demonstration set was extracted may be requested from David Schoeller, iMMAP's project lead in Colombia (dschoellerdiaz@immap.org), or from José Cobos Romero, DFS's analysis coordinator (jose@datafriendlyspace.org).

References

ACAPS. (2011). "Survey of Surveys. Technical note [July 2011]." Geneva. Retrieved 12 August 2015, from https://www.acaps.org/sites/acaps/files/resources/files/technical_note-survey_of_surveys_july_2011.pdf.

ACAPS (2014) "Technical Brief: Secondary Data Review. Sudden Onset Natural Disasters." Retrieved 11 November 2021, from

https://www.acaps.org/sites/acaps/files/resources/files/secondary_data_review-sudden_onset_natural_disasters_may_2014.pdf.

Akimbayev, Y. Z., Z. K. Akhmetov, M. S. Kuanyshbaev, A. T. Abdykalykov and R. V. Ibrayev (2020). "Determination of integral indicators characterizing the possibility of accommodating the evacuated population in the territory of the Republic of Kazakhstan." The Journal of Defense Modeling and Simulation: 1548512920970288.

Akinshin, A. (2021) "Weighted quantile estimators." Retrieved 17 December 2021, from https://aakinshin.net/posts/weighted-quantiles/.

Aven, T. (2011). "Interpretations of alternative uncertainty representations in a reliability and risk analysis context." Reliability Engineering & System Safety **96**(3): 353-360.

Azar, D., R. Engstrom, J. Graesser and J. Comenetz (2013). "Generation of fine-scale population layers using multi-resolution satellite imagery and geospatial data." Remote Sensing of Environment **130**(15): 219-232.

Benini, A., P. Chataigner, N. Noumri, L. Tax and M. Wilkins. (2016). "Information gaps in multiple needs assessments in disaster and conflict areas. With guidelines for their measurement, using the Nepal 2015 earthquake assessment registry." Geneva, Assessment Capacity Project (ACAPS). Retrieved 23 May 2016, from http://www.acaps.org/sites/acaps/files/resources/files/info_gaps.pdf.

Bronevich, A. and G. J. Klir (2010). "Measures of uncertainty for imprecise probabilities: An axiomatic approach." International Journal of Approximate Reasoning **51**(4): 365-390.

DFS. (2021). "About Us." Retrieved 5 November 2021, from https://datafriendlyspace.org/about/.

Dubois, D. and H. Prade (2015). Possibility theory and its applications: Where do we stand? Springer handbook of computational intelligence, Springer**:** 31-60.

Dubois, D., H. Prade and R. Yager (1999). Merging fuzzy information. Fuzzy sets in approximate reasoning and information systems, Springer**:** 335-401.

Flage, R., T. Aven, E. Zio and P. Baraldi (2014). "Concerns, Challenges, and Directions of Development for the Issue of Representing Uncertainty in Risk Assessment." Risk Analysis **34**(7): 1196-1207.

Gaber, J. (2000). "Meta-needs assessment." Evaluation and Program Planning **23**(2): 139-147.

Hensher, D. A. and Z. Li (2014). "A scoping inquiry into the potential contribution of Subjective Probability Theory, Dempster–Shafer Theory and Possibility Theory in

accommodating degrees of belief in traveller behaviour research." Travel Behaviour and Society **1**(2): 45-56.

Holst, C.-A. and V. Lohweg (2020). A Redundancy Metric based on the Framework of Possibility Theory for Technical Systems. 2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA), IEEE.

Huang, C. (2020). "Analysis of Death Risk of COVID-19 under Incomplete Information1." Journal of Risk Analysis and Crisis Response **10**(2): 43-53.

IASC (2012) "Multi-Cluster/Sector Initial Rapid Assessment (MIRA). Provisional Version March 2012 [final version in 2015]." Retrieved 16 December 2021, from https://www.unocha.org/sites/dms/CAP/mira_final_version2012.pdf.

iMMAP. (2021). "Who We Are." Retrieved 5 November 2021, from https://immap.org/who-we-are/.

iMMAP and DFS. (2020). "Secondary Data Analysis for OFDA Situation Analysis project. Concept, Approach & Method. August 2020. Version 1.1." Retrieved 13 January 2022, from https://www.dropbox.com/s/fvgl0tq3kyo67ya/210209%20Secondary%20Data%20Situation%20Analysis%20Framework%20iMMAP_DFS.docx?dl=0.

iMMAP and DFS. (2021). "211109 iMMAP-DFS Main Data Sources.xlsx." Bogotá. Retrieved 13 December 2021, from [permanent site to be created].

Kikuchi, S. and P. Chakroborty (2006). "Place of possibility theory in transportation analysis." Transportation Research Part B: Methodological **40**(8): 595-615.

Klir, G. J. (2006). Uncertainty and information: Foundations of generalized information theory. Hoboken, New Jersey, John Wiley & Sons, Inc.

Lesot, M.-J. and A. R. d'Allonnes (2017). Information quality and uncertainty. Uncertainty Modeling, Springer**:** 135-146.

Lesot, M. J., T. Delavallade, F. Pichon, H. Akdag, B. Bouchon-Meunier and P. Capet (2011). Proposition of a semi-automatic possibilistic information scoring process. Proceedings of the 7th conference of the European Society for Fuzzy Logic and Technology, Atlantis Press.

Mann, H. B. and D. R. Whitney (1947). "On a test of whether one of two random variables is stochastically larger than the other." The annals of mathematical statistics: 50-60.

Milisavljevic, N. (2017). "Data Fusion for Close - Range Detection." Mine Action: The Research Experience of the Royal Military Academy of Belgium: 83.

OCHA (2021) "Humanitarian Needs Overview 2021 - Colombia." Retrieved 11 November 2021, from https://www.humanitarianresponse.info/es/operations/colombia/document/colombia-pnh-hno-panorama-de-necesidades-humanitarias-mar-2021.

Rabbani, M., N. Manavizadeh, M. Samavati and M. Jalali (2015). "Proactive and reactive inventory policies in humanitarian operations." Uncertain Supply Chain Management **3**(3): 253-272.

Raufaste, E., R. da Silva Neves and C. Mariné (2003). "Testing the descriptive validity of possibility theory in human judgments of uncertainty." Artificial Intelligence **148**(1): 197-218.

Shafer, G. A. (1976). A mathematical theory of evidence. Princeton, NJ, Princeton University Press.

Solaiman, B. and É. Bossé (2019). Possibility Theory for the Design of Information Fusion Systems, Springer.

Tofighi, S., S. A. Torabi and S. A. Mansouri (2016). "Humanitarian logistics network design under mixed uncertainty." European Journal of Operational Research **250**(1): 239-250.

Weichselberger, K. (2000). "The theory of interval-probability as a unifying concept for uncertainty." International Journal of Approximate Reasoning **24**(2): 149-170.

Wikipedia. (2011). "Dempster–Shafer theory." Retrieved 29 March 2011, from http://en.wikipedia.org/wiki/Dempster%E2%80%93Shafer_theory.

Wikipedia. (2016). "Somers' D." Retrieved 27 November 2016, from https://en.wikipedia.org/wiki/Somers%27_D.

Wikipedia. (2020). "Possibility theory." Retrieved 28 November 2020, from https://en.wikipedia.org/w/index.php?title=Possibility_theory&oldid=978949635.

Wikipedia. (2021a). "Municipalities of Colombia." Retrieved 17 November 2021, from https://en.wikipedia.org/w/index.php?title=Municipalities_of_Colombia&oldid=1034871821.

Wikipedia. (2021b). "Percentile." Retrieved 17 December 2021, from https://en.wikipedia.org/wiki/Percentile#The_weighted_percentile_method.

Wikipedia. (2021c). "Wilcoxon signed-rank test." Retrieved 30 November 2021, from https://en.wikipedia.org/wiki/Wilcoxon_signed-rank_test.

Author information

Aldo Benini has a dual career in rural development, with a focus on Bangladesh and on organizations of the poor, and in humanitarian action. In the latter capacity, he has worked for the International Committee of the Red Cross, for the Global Landmine Survey as well as for organizations building needs assessment and data analysis capacity. He has a Ph.D. in sociology from the University of Bielefeld, Germany, based on field research in community development in West Africa. He can be reached at aldobenini @gmail.com.

José Cobos Romero divides his career between humanitarian analysis and the study of conflict and peacebuilding. In recent years, both the subject of the PhD he has been pursuing and the context in which he has worked have been focused mainly on Colombia. For the last three years, he has been working as a humanitarian analyst and consultant for Data Friendly Space and OkularAnalytics, in different projects for clients such as the United Nations High Commissioner for Refugees, the United Nations Development Coordination Office, the International Federation of Red Cross and Red Crescent Societies and the US Agency for International Development. His email address is jose@datafriendlyspace.org.