A Guide to the agop 0.1-0 Package for R

Aggregation Operators in R

Marek Gagolewski[1,2], Anna Cena[1,2]
[1] Systems Research Institute, Polish Academy of Sciences
ul. Newelska 6, 01-447 Warsaw, Poland
[2] Rexamine, Email: {gagolews,cena}@rexamine.com
www.rexamine.com/resources/agop/

June 14, 2013

The package, as well as this tutorial, is still in its early days – any suggestions are welcome!

Contents

1 Getting started

> *"The process of combining several numerical values into a single representative one is called* ***aggregation****, and the numerical function performing this process is called* ***aggregation function****. This simple definition demonstrates the size of the field of application of aggregation: applied mathematics (e.g. probability, statistics, decision theory), computer science (e.g. artificial intelligence, operation research),*

> *as well as many applied fields (economics and finance, pattern recognition and image processing, data fusion, multicriteria decision making, automated reasoning etc.)." [12, p. ??]*

R [18] is a free, open sourced software environment for statistical computing and graphics, which includes an implementation of a very powerful and quite popular high-level language called S. It runs on all major operating systems, i.e. Windows, Linux, and MacOS X. To install R and/or find some information on the S language please visit R Project's Homepage at www.R-project.org. Perhaps you may also wish to install RStudio, a convenient development environment for R. It is available at www.rsudio.org.

agop is an Open Source (licensed under GNU LGPL 3) package for R $\geq$ 2.12 to which anyone can contribute. It started as a fork of the CITAN (Citation Analysis Toolpack) package for R.

To install latest "official" release of the package available on *CRAN* we type[1]:

```
# install.packages('agop') # NOT YET AVAILABLE ON CRAN
```

Alternatively, we may fetch its current development snapshot from *GitHub*:

```
install.packages('devtools')
library('devtools')
install_github('agop', 'Rexamine')
```

Each session with agop should be preceded by a call to:

```
library('agop') # Load the package
```

To view the main page of the manual we type:

```
library(help='agop')
```

For more information please visit the package's homepage [9]. In case of any problems, comments, or suggestions feel free to contact the authors. Good luck!

2 Theoretical Background

Let us establish some basic notation convention used throughout this tutorial. From now on let $\mathbb{I} = [a, b]$, possibly with $a = -\infty$ or $b = \infty$. Note that in many practical situations we choose $\mathbb{I} = [0, 1]$ or $\mathbb{I} = [0, \infty]$. A set of all vectors of arbitrary length with elements in $\mathbb{I}$ is denoted by $\mathbb{I}^{1,2,\dots} = \bigcup_{n=1}^{\infty} \mathbb{I}^n$.

For two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{I}^n$ we write $\mathbf{x} \leq \mathbf{y}$ if and only if for all i it holds $x_i \leq y_i$. Moreover, all binary arithmetic operations on vectors $\mathbf{x}, \mathbf{y} \in \mathbb{I}^n$ are performed element-wise, e.g. $\mathbf{x} + \mathbf{y} = (x_1 + y_1, \dots, x_n + y_n) \in \mathbb{I}^n$. Similarly: $-$, $\cdot$, $/$, $\wedge$ (min), $\vee$ (max), etc. Additionally, each function of one variable $\mathsf{f} : \mathbb{I} \to \mathbb{I}$ can be extended to the vector space: we write $\mathsf{f}(\mathbf{x}) = (\mathsf{f}(x_1), \dots, \mathsf{f}(x_n))$.

Let $x_{(i)}$ denote the ith order statistic, i.e. the ith smallest value in $\mathbf{x}$. Moreover, for convenience, let $x_{\{i\}} = x_{|\mathbf{x}|-i+1}$ denote the ith greatest value in $\mathbf{x}$.

For any $n \in \mathbb{N}$ and $c \in \mathbb{I}$, we set $(n * c) = (c, \dots, c) \in \mathbb{I}^n$.

[1] You are viewing the **development** version of the tutorial. Some of the features presented in this document may be missing in the CRAN release. Please, upgrade to the **latest** development version from *GitHub* if you need the new functionality.

2.1 A Note on Representing Numeric Data and Applying Operations in R

Recall how we create numeric vectors in R:

```
(x1 <- c(5, 2, 3, 1, 0, 0))
## [1] 5 2 3 1 0 0
class(x1)
## [1] "numeric"
(x2 <- 10:1) # the same as seq(10, 1)
##  [1] 10  9  8  7  6  5  4  3  2  1
(x3 <- seq(1, 5, length.out=6))
## [1] 1.0 1.8 2.6 3.4 4.2 5.0
(x4 <- seq(1, 5, by=1.25))
## [1] 1.00 2.25 3.50 4.75
```

To obtain $(n * c)$, e.g. for $n = 4$ and $c = 1$, we call:

```
rep(10, 3)
## [1] 10 10 10
```

Please, note that in R all the arithmetic operations on vectors are performed element-wise, i.e. in a manner indicated above. This is called *vectorization*. The same holds for mathematical functions: they are extended to the vector space.

```
x <- c(1, 3, 3, 2)
y <- c(2 ,3,-1, 0)
x+y
## [1] 3 6 2 2
x*y
## [1]  2  9 -3  0
pmin(x,y)
## [1]  1  3 -1  0
pmax(x,y)
## [1] 2 3 3 2
abs(y)
## [1] 2 3 1 0
```

Thus, we calculated $\mathbf{x} + \mathbf{y}$, $\mathbf{x} \cdot \mathbf{y}$, $\mathbf{x} \wedge \mathbf{y}$, and $\mathbf{x} \vee \mathbf{y}$ (try to determine yourself what happens if we deal with vector of unequal length is R).

Moreover, for the $\leq$ relation we write:

```
all(x <= y)
## [1] FALSE
```

To get $x_{\{i\}}$ we have to sort the vector nonincreasingly.

```
(xs <- sort(x, decreasing=TRUE)) # `decresing' may be misleading
## [1] 3 3 2 1
xs[3] # the third greatest value in x
## [1] 2
```

2.2 Aggregation Operators and Their Basic Properties

Dealing with huge amounts of data faces us with the problem of constructing their synthetic descriptions. The aggregation theory, a relatively new research domain at the border of mathematics and computer science, is interested in the analysis of functions that may be used in the mentioned task.

Thus, we should start with the formal definition of objects of our interest. Here is the most general setting:

Definition 1. $\mathsf{F} : \mathbb{I}^{1,2,\dots} \to \mathbb{I}$ is called an *(extended) aggregation operator* if it is at least *nondecreasing* in each variable, i.e. for all n and $\mathbf{x}, \mathbf{y} \in \mathbb{I}^n$ if $\mathbf{x} \le \mathbf{y}$, then $\mathsf{F}(\mathbf{x}) \le \mathsf{F}(\mathbf{y})$.

Note that each aggregation operator is a mapping into $\mathbb{I}$, thus for all n we have $\inf_{\mathbf{x}\in\mathbb{I}^n} \mathsf{F}(\mathbf{x}) \ge a$ and $\sup_{\mathbf{x}\in\mathbb{I}^n} \mathsf{F}(\mathbf{x}) \le b$. By nondecreasingness, however, these conditions reduce to $\mathsf{F}(n * a) \ge a$ and $\mathsf{F}(n * b) \le b$.

Depending on problems to be solved, some authors assume (cf. [12]) that for all n we have $\inf_{\mathbf{x}\in\mathbb{I}^n} \mathsf{F}(\mathbf{x}) = a$ and $\sup_{\mathbf{x}\in\mathbb{I}^n} \mathsf{F}(\mathbf{x}) = b$.

Definition 2. We call $\mathsf{F} : \mathbb{I}^{1,2,\dots} \to \mathbb{I}$ *symmetric* if:

$$(\forall n \in \mathbb{N})\ (\forall \mathbf{x}, \mathbf{y} \in \mathbb{I}^n)\ \mathbf{x} \cong \mathbf{y} \Longrightarrow \mathsf{F}(\mathbf{x}) = \mathsf{F}(\mathbf{y}),$$

where $\mathbf{x} \cong \mathbf{y}$ if and only if there exists a permutation σ of $[n] := \{1, 2, \dots, n\}$ such that $\mathbf{x} = (y_{\sigma(1)}, \dots, y_{\sigma(n)})$.

Note that it may be shown, see [12, Thm. 2.34], that $\mathsf{F} : \mathbb{I}^n \to \mathbb{I}$ is symmetric if and only if there exists a function $\mathsf{G} : \mathbb{I}^n \to \mathbb{I}$ such that $\mathsf{F}(x_1, \dots, x_n) = \mathsf{G}(x_{\{1\}}, \dots, x_{\{n\}})$.

Definition 3. We call $\mathsf{F} : \mathbb{I}^{1,2,\dots} \to \mathbb{I}$ *idempotent* if:

$$(\forall x \in \mathbb{I})\ \mathsf{F}(n * x) = x.$$

Please note, that property introduced above – *idempotence* – is well known in algebra, where we say that elenemt x is idempotent with respect to binary operator $*$ if we have $x * x = x$. Similarly, this definition were extend for n-ary aggregation functions, cf. [12], for which it simply means that for $\mathbf{x} \in \mathbb{I}^n$ such that $x_i = x$ for all $i = 1, \dots, n$, we have $\mathsf{F}(\mathbf{x}) = x$.

Definition 4. We call $\mathsf{F} : \mathbb{I}^{1,2,\dots} \to \mathbb{I}$ *additive* if:

$$\mathsf{F}(\mathbf{x} + \mathbf{y}) = \mathsf{F}(\mathbf{x}) + \mathsf{F}(\mathbf{y}),$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{I}^n$ such that $\mathbf{x} + \mathbf{y} \in \mathbb{I}^n$.

Definition 5. We call F *minitive* if:

$$(\forall n)\ (\forall \mathbf{x}, \mathbf{y} \in \mathbb{I}^n)\ \mathsf{F}(\mathbf{x} \wedge \mathbf{y}) = \mathsf{F}(\mathbf{x}) \wedge \mathsf{F}(\mathbf{y}).$$

Definition 6. We call F *maxitive* if:

$$(\forall n)\ (\forall \mathbf{x}, \mathbf{y} \in \mathbb{I}^n)\ \mathsf{F}(\mathbf{x} \vee \mathbf{y}) = \mathsf{F}(\mathbf{x}) \vee \mathsf{F}(\mathbf{y}).$$

Definition 7. We call F *modular* if:

$$(\forall n)\ (\forall \mathbf{x}, \mathbf{y} \in \mathbb{I}^n)\ \mathsf{F}(\mathbf{x} \vee \mathbf{y}) + \mathsf{F}(\mathbf{x} \wedge \mathbf{y}) = \mathsf{F}(\mathbf{x}) + \mathsf{F}(\mathbf{y}).$$

Apart from "ordinary" minitivity, maxitivity, and modularity we may introduce their symmetrized versions, using $\mathbf{x} \overset{S}{\vee} \mathbf{y} = (x_{\{1\}} \vee y_{\{1\}}, \dots, x_{\{n\}} \vee y_{\{n\}})$ and $\mathbf{x} \overset{S}{\wedge} \mathbf{y} = (x_{\{1\}} \wedge y_{\{1\}}, \dots, x_{\{n\}} \wedge y_{\{n\}})$.

2.3 Impact Functions and The Producers Assessment Problem

The **Producers Assessment Problem** (PAP, [11]) concerns evaluation of a set of **producers** (e.g. scientists, artists, writers, craftsman) according to the **products** (e.g. articles, works, books, artifacts) each entity has produced. It is assumed that for each of the products a **rating** of quality or popularity (e.g. number of citations for articles, number of sold copies for books etc.) is given.

Tab. 1. The Producer Assessment Problem – typical instances

	Producer	Products	Rating method	Discipline
A	Scientist	Scientific articles	Number of citations	Scientometrics
B	Scientific institute	Scientists	The h-index	Scientometrics
C	Web server	Web pages	Number of in-links	Webometrics
D	Artist	Paintings	Auction price	Auctions
E	Billboard company	Advertisements	Sale results	Marketing
F	R package author	Packages	PageRank values on the citation graph	Software Engineering

PAP finds its applications in many fields, like for example scientometrics, webometrics, marketing, manufacturing or quality engineering [5], and its main interest if focused on constructing and analyzing aggregation operators which may be used in the rating task. Such functions should take into a count the two following aspects of a producer's quality:

- his/her ability to output highly-rated products,
- his/her overall productivity.

For the sake of illustration, we will consider PAP in the scientometric context, where scientists "produce" papers, which receive citations.

Let $\mathbb{I} = [0, \infty]$ represent the set of values that some a priori chosen paper quality measure may take. These may of course be non-integers, for example when we consider citations normalized with respect to the number of papers' authors.

It is widely accepted, see e.g. [26, 25, 24, 19, 16, 17, 11, 7, 6], that each aggregation operator $\mathsf{J} : \mathbb{I}^{1,2,\dots} \to \mathbb{I}$ to be applied in the PAP should at least be:

(a) nondecreasing in each variable (additional citations received by a paper or an improvement of its quality measure does not result in a decrease of the authors' overall evaluation),

(b) arity-monotonic (by publishing a new paper we never decrease the overall valuation of the entity),

(c) symmetric (independent of the order of elements' presentation, i.e. we may always assume that we aggregate vectors that are already sorted).

Conditions (a) and (b) imply that each impact function is able – at least potentially – to describe two "dimensions" of the author's output quality: (a) his/her ability to write eagerly-cited or highly-valuated papers and (b) his/her overall productivity.

More formally, axiom (b) is fulfilled iff for any $\mathbf{x} \in \mathbb{I}^{1,2,\dots}$ and $y \in \mathbb{I}$ it holds $\mathsf{J}(\mathbf{x}) \leq \mathsf{J}(x_1, \dots, x_n, y)$. (a) and (c) were defined in the previous section.

3 Predefined Classes of Aggregation Operators in agop

3.1 A Note on Storing Multiple Numeric Vectors in R

Sometimes we will store the vectors of the same length in a matrix (column-/row-wise.... col/rownames....) `apply()`....

```
expertopinions <- matrix(c(
      6,7,2,3,1, # this will be the first COLUMN
      8,3,2,1,9, # 2nd
      4,2,4,1,6  # 3rd
   ),
   ncol=3,
   dimnames=list(NULL, c("A", "B", "C")) # only column names set
)
class(expertopinions)

## [1] "matrix"

print(expertopinions)   # or print(authors)

##      A B C
## [1,] 6 8 4
## [2,] 7 3 2
## [3,] 2 2 4
## [4,] 3 1 1
## [5,] 1 9 6

apply(expertopinions, 2, mean) # on each COLUMN apply the mean() function

##   A   B   C
## 3.8 4.6 3.4
```

...or in a list, especially when they are not of the same length.... `lapply()`.... `sapply()`..... possibly named elements...

```
authors <- list(
   "John S." = c(7,6,2,1,0),
   "Kate F." = c(9,8,7,6,4,1,1,0)
)
class(authors)

## [1] "list"

str(authors)   # or print(authors)

## List of 2
##  $ John S.: num [1:5] 7 6 2 1 0
##  $ Kate F.: num [1:8] 9 8 7 6 4 1 1 0

index_h(authors[[1]]) # the h-index (see below) for 1st author

## [1] 2

sapply(authors, index_h) # calculate the h-index for all vectors in a list

## John S. Kate F.
##       2       4

index_h(authors) # index_h() expects an numeric vector on input

## Error:  argument 'x' should be a numeric vector (or an object coercible to)
```

3.2 ...

Definition 8. Let $\mathbf{w} = c(w_1, \ldots, w_n) \in [0,1]^n$ be such that $\sum_{i=1}^n w_i = 1$. Then, for any $\mathbf{x} \in \mathbb{I}^n$:

1. The *weighted arithmetic mean* $\mathsf{WAM} : \mathbb{I}^n \to \mathbb{I}$ associated with the weight vector $\mathbf{w}$ is defined as
$$\mathsf{WAM}(\mathbf{x}) = \sum_{i=1}^n w_i x_i.$$

2. The *ordered weighted averaging function* $\mathsf{OWA} : \mathbb{I}^n \to \mathbb{I}$ associated with the weight vector $\mathbf{w}$ is defined as
$$\mathsf{OWA}(\mathbf{x}) = \sum_{i=1}^n w_i x_{(i)}.$$

Definition 9. Let $\mathbf{w} = c(w_1, \ldots, w_n) \in [0,1]^n$ be such that $\bigvee_{i=1}^n w_i = 1$, and assume $\mathbb{I} = [0,1]^n$. Then, for any $\mathbf{x} \in \mathbb{I}^n$:

1. The *weighted maximum* $\mathsf{WMax} : \mathbb{I}^n \to \mathbb{I}$ associated with the weight vector $\mathbf{w}$ is defined as
$$\mathsf{WMax}(\mathbf{x}) = \bigvee_{i=1}^n (w_i \wedge x_i).$$

2. The *ordered weighted maximum* $\mathsf{OWMax} : \mathbb{I}^n \to \mathbb{I}$ associated with the weight vector $\mathbf{w}$ is defined as
$$\mathsf{OWMax}(\mathbf{x}) = \bigvee_{i=1}^n (w_i \wedge x_{(i)}),$$
with $x_{(1)} \le \cdots \le x_{(n)}$ and $w_1 \ge w_2 \ge \cdots \ge w_n$.

Definition 10. Let $\mathbf{w} = c(w_1, \ldots, w_n) \in [0,1]^n$ be such that $\bigvee_{i=1}^n w_i = 1$, and assume $\mathbb{I} = [0,1]^n$. Then, for any $\mathbf{x} \in \mathbb{I}^n$:

1. The *weighted minimum* $\mathsf{WMin} : \mathbb{I}^n \to \mathbb{I}$ associated with the weight vector $\mathbf{w}$ is defined as
$$\mathsf{WMin}(\mathbf{x}) = \bigwedge_{i=1}^n ((1 - w_i) \vee x_i).$$

2. The *ordered weighted minimum* $\mathsf{OWMin} : \mathbb{I}^n \to \mathbb{I}$ associated with the weight vector $\mathbf{w}$ is defined as
$$\mathsf{OWMin}(\mathbf{x}) = \bigwedge_{i=1}^n ((1 - w_i) \vee x_{(i)}),$$
with $x_{(1)} \le \cdots \le x_{(n)}$ and $w_1 \le w_2 \le \cdots \le w_n$.

Definition 11. The *triangle of coefficients* is a sequence $\triangle = (c_{i,n} \in \mathbb{I} : i \in [n], n \in \mathbb{N})$.

Definition 12. The *L-statistic* for a given triangle of coefficients $\triangle = (c_{i,n})_{i \in [n], n \in \mathbb{N}}$ is a function $\mathsf{L}_\triangle : \mathbb{I}^n \to \mathbb{I}$ such that
$$\mathsf{L}_\triangle(\mathbf{x}) = \sum_{i=1}^n c_{i,n} x_{(n-i+1)},$$
for any $\mathbf{x} \in \mathbb{I}^n$.

Definition 13. The *S-statistic* for a given triangle of coefficients $\triangle = (c_{i,n})_{i\in[n],n\in\mathbb{N}}$ is a function $\mathsf{S}_\triangle : \mathbb{I}^n \to \mathbb{I}$ such that

$$\mathsf{S}_\triangle(\mathbf{x}) = \bigvee_{i=1}^{n} c_{i,n} \wedge x_{(n-i+1)},$$

for any $\mathbf{x} \in \mathbb{I}^n$.

Definition 14. The *I-statistic* for a given triangle of coefficients $\triangle = (c_{i,n})_{i\in[n],n\in\mathbb{N}}$ is a function $\mathsf{I}_\triangle : \mathbb{I}^n \to \mathbb{I}$ such that

$$\mathsf{I}_\triangle(\mathbf{x}) = \bigwedge_{i=1}^{n} c_{i,n} \vee x_{(n-i+1)},$$

for any $\mathbf{x} \in \mathbb{I}^n$.

Please note that the *OWA* functions are a special case of *L-statistics*. Similarly, *OWMax* and *OWMin* operators are a special cases of *S-statistics* and *I-statistics*, respectively.

Definition 15. The *triangle of functions* is a sequence $\triangle = (f_{i,n} \in \mathbb{I}^{\mathbb{I}} : i \in [n], n \in \mathbb{N})$.

Definition 16. The *quasi-L-statistic* for a given triangle of functions $\triangle = (f_{i,n})_{i\in[n],n\in\mathbb{N}}$ is a function $\mathsf{qL}_\triangle : \mathbb{I}^n \to \mathbb{I}$ such that

$$\mathsf{qL}_\triangle(\mathbf{x}) = \sum_{i=1}^{n} f_{i,n}(x_{(n-i+1)}),$$

for any $\mathbf{x} \in \mathbb{I}^n$.

Definition 17. The *quasi-S-statistic* for a given triangle of functions $\triangle = (f_{i,n})_{i\in[n],n\in\mathbb{N}}$ is a function $\mathsf{qS}_\triangle : \mathbb{I}^n \to \mathbb{I}$ such that

$$\mathsf{qS}_\triangle(\mathbf{x}) = \bigvee_{i=1}^{n} f_{i,n}(x_{(n-i+1)}),$$

for any $\mathbf{x} \in \mathbb{I}^n$.

Definition 18. The *quasi-I-statistic* for a given triangle of functions $\triangle = (f_{i,n})_{i\in[n],n\in\mathbb{N}}$ is a function $\mathsf{qI}_\triangle : \mathbb{I}^n \to \mathbb{I}$ such that

$$\mathsf{qI}_\triangle(\mathbf{x}) = \bigwedge_{i=1}^{n} f_{i,n}(x_{(n-i+1)}),$$

for any $\mathbf{x} \in \mathbb{I}^n$.

It is easily seen that *quasi-L-statistics* generalize *L-statistics*. We obtain an *L-statistic* by taking $f_{i,n}(x) = c_{i,n}x$. Similarly, by setting $f_{i,n}(x) = x \wedge c_{i,n}$ we obtain *S-statistic* from *quasi-S-statistics*, and by setting $f_{i,n}(x) = x \vee c_{i,n}$ we obtain *I-statistic* from *quasi-I-statistics*.

Let now us introduce the class of symmetric maxitive, minitive and modular aggregation operators, cf. [2].

Definition 19. A sequence of nondecreasing functions $\mathbf{w} = (\mathsf{w}_1, \mathsf{w}_2, \dots)$, $\mathsf{w}_i : \mathbb{I} \to \mathbb{I}$, and a triangle of coefficients $\triangle = (c_{i,n})_{i\in[n],n\in\mathbb{N}}$, $c_{i,n} \in \mathbb{I}$ such that $(\forall n)\ c_{1,n} \le c_{2,n} \le \cdots \le c_{n,n}$, $0 \le \mathsf{w}_n(0) \le c_{1,n}$, and $\mathsf{w}_n(b) = c_{n,n}$, generates a nondecreasing *OM3 operator* $\mathsf{M}_{\triangle,\mathbf{w}} : \mathbb{I}^n \to \mathbb{I}$ such that for $\mathbf{x} \in \mathbb{I}^n$ we have:

$$\begin{aligned}\mathsf{M}_{\triangle,\mathbf{w}}(\mathbf{x}) &= \bigvee_{i=1}^{n} \mathsf{w}_n(x_{(n-i+1)}) \wedge c_{i,n} = \bigwedge_{i=1}^{n} (\mathsf{w}_n(x_{(n-i+1)}) \vee c_{i-1,n}) \wedge c_{n,n} \\ &= \sum_{i=1}^{n} \left(\left(\mathsf{w}_n(x_{(n-i+1)}) \vee c_{i-1,n} \right) \wedge c_{i,n} - c_{i-1,n} \right).\end{aligned}$$

We see that the above contains i.a. all order statistics (whenever $\mathsf{w}_n(x) = x$, and $c_{i,n} = 0$, $c_{j,n} = b$ for $i < k$, $j \geq k$, and some k), OWMax operators (for $\mathsf{w}_n(x) = x$), and the famous Hirsch h-index ($\mathsf{w}_n(x) = \lfloor x \rfloor$, $c_{i,n} = i$).

3.3 Bibliometric Impact Indices

Below we assume that $\mathbb{I} = [0, \infty]$.

The h-index. Given a sequence $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{I}^{1,2,\cdots}$, the *Hirsch index* [13] of $\mathbf{x}$ is defined as $\mathsf{H}(x) = \max\{i = 1, \ldots, n : x_{\{i\}} \geq i\}$ if $n \geq 1$ and $x_{\{1\}} \geq 1$, or $\mathsf{H}(x) = 0$ otherwise. It may be shown that the h-index is an zero-insensitive OM3 aggeration operator, see [8], with:

$$\mathsf{H}(\mathbf{x}) = \bigvee_{i=1,\ldots,n}^{n} i \wedge \lfloor x_{\{i\}} \rfloor.$$

Interpretation: "an author has h-index of H if H of his/her n most cited papers have at least H citations each, and the other $n - H$ papers are cited no more that H times each". The h-index may also be expressed as a Sugeno integral [21] w.r.t. to a counting measure, cf. [22].

agop implementation: `index_h()`.

```
index_h(c(6,5,4,2,1,0,0,0,0,0,0))
## [1] 3
index_h(c(-1,3,4,2)) # only for x>=0
## Error:  all elements in `x` should be in [0,Inf]
```

This is a zero-insensitive impact function.

We have $\mathsf{H}(\mathbf{x}) \leq \min\{n, x_1\}$.

The g-index. Egghe's g-index [4]: $\mathsf{G}(\mathbf{x}) = \max\{g = 1, \ldots, n : \sum_{i=1}^{g} x_{\{g\}} \geq g^2\}$, available in agop as `index_g()`. We have $\mathsf{G}(\mathbf{x}) \geq \mathsf{H}(\mathbf{x})$ with $\mathsf{G}(n * n) = \mathsf{H}(n * n) = n$

Note that this aggregation operator is not zero-insensitive, for example $\mathsf{G}(9, 0) = 2$ and $\mathsf{G}(9, 0, 0) = 3$. Thus, we also provide the `index_g_zi()` function, which treats $\mathbf{x}$ as it would be padded with 0s.

```
index_g(9)
## [1] 1
index_g(c(9,0,0))
## [1] 3
index_g_zi(9)
## [1] 3
```

The index is interesting from the computational point of view – it may be calculated on the nondecreasing vector of cumulative sums, `cumsum(sort(x, decreasing=TRUE))`, however, it cannot be expressed as a symmetric maxitive aggregation operator.

Interestingly, it might be shown that if $\mathbf{x}$ is sorted nondecreasingly, then:

$$\mathsf{G}(\mathbf{x}) = \mathsf{H}(\mathbf{x})(0 \vee \mathtt{cummin}(\mathtt{cumsum}(x) - (1:n)^2 + (1:n))),$$

where $1 : n = (1, 2, 3, \ldots, n)$.

The w-index. The w-index [26]:

$$\mathsf{W}(\mathbf{x}) = \max\left\{w = 0, 1, 2, \dots : x_{\{i\}} \geq w - i + 1, i = 1, \dots, w\right\}. \tag{1}$$

agop implementation: `index_w()`.
This is a zero-insensitive impact function.
We have $\mathsf{H}(\mathbf{x}) \leq \mathsf{W}(\mathbf{x}) \leq 2\mathsf{H}(\mathbf{x})$, $\mathsf{W}(\mathbf{x}) \leq \min\{n, x_1\}$.
Interestingly, it might be shown that if $\mathbf{x}$ is sorted nondecreasingly, then:

$$\mathsf{W}(\mathbf{x}) = \mathsf{H}(\mathbf{x})(\texttt{cummin}(\mathbf{x} + (1:n) - 1)).$$

The r_p-indices. [10] for integer vectors we have $r_1 = \mathsf{W}$ and $r_\infty = \mathsf{H}$
This is a zero-insensitive impact function.

The MAXPROD**-index.** The MaxProd-index [15]:

$$\mathsf{MP}(\mathbf{x}) = \max\left\{i \cdot x_{\{i\}} : i = 1, 2, \dots\right\}. \tag{2}$$

This index is a particular case of a projected l_∞-index, see [10].
Shilkret integral [20]
This is a zero-insensitive impact function.
agop implementation: `index_maxprod()`.

The l_p-indices. [10]

Simple transformations of the h-index. For example, The $h(2)$-index [14]:

$$\mathsf{H2}(\mathbf{x}) = \max\left\{h = 0, 1, 2, \dots : x_h \geq h^2\right\}. \tag{3}$$

Note that the $h(2)$-index is one of the many examples of very simple, direct modifications of the h-index. Many authors considered other settings than "h^2" on the right side of (3), e.g. "$2h$", "αh" for some $\alpha > 0$, or "h^β", $\beta \geq 1$, cf. [1].

It may easily be shown that these reduce to the h-index for properly transformed input vectors.....

4 Visualization

4.1 Depicting producers

The `plot_producer()` function may be used to draw a graphical representation of a given numeric vector, i.e. what is sometimes called a citation function in scientometrics.

A given vector $\mathbf{x} = (x_1, \dots, x_n)$ can be represented by a step function defined for $0 \leq y < n$ and given by:

$$\pi(y) = x_{(n - \lfloor y+1 \rfloor + 1)}.$$

This function may be obtained by setting `type == 'right.continuous'` argument in `plot_producer()`. Recall that $x_{(i)}$ denotes i-th smallest value in $\mathbf{x}$.

On the other hand, for `type == 'left.continuous'` (the default), we get

$$\pi(y) = x_{(n - \lfloor y \rfloor + 1)}$$

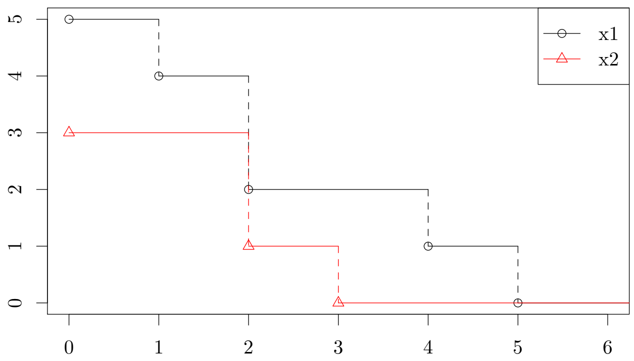
for $0 < y \le n$.

Moreover, this function may depict the curve joining the sequence of points $(0, x_{(n)}), (1, x_{(n)}), (1, x_{(n-1)}), (2, x_{(n-1)}), \ldots, (n, x_{(1)})$.

The `plot_producer()` function behaves much like the well-known R's `plot.default()` and allows for passing all its graphical parameters.

For example, let us depict the state of two given producers, $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$.

```
x1 <- c(5, 4, 2, 2, 1)
x2 <- c(3, 3, 1, 0, 0, 0, 0)
plot_producer(x1, extend=TRUE)
plot_producer(x2, add=TRUE, col=2, pch=2, extend=TRUE)
legend('topright', c('x1', 'x2'), col=c(1, 2), lty=1, pch=c(1, 2))
```



5 Pre-orders

....

Let us consider the following relation on $\mathbb{I}^{1,2,\ldots}$. For any $\mathbf{x} \in \mathbb{I}^n$ and $\mathbf{y} \in \mathbb{I}^m$ we write $\mathbf{x} \trianglelefteq \mathbf{y}$ if and only if $n \le m$ and $x_{\{i\}} \le y_{\{i\}}$ for all $i \in \min\{n, m\}$. Of course, $\trianglelefteq$ is a pre-order – it would have been a partial order, if we had defined it on the set of *sorted* vectors.

In other words, we say that an author X is (weakly) dominated by an author Y, if X has no more papers than Y and each the ith most cited paper of X has no more citations than the ith most cited paper of Y. Not that the $m - n$ least cited Y's papers are not taken into account here. Most importantly, however, there exist pairs of vectors that are *incomparable* with respect to $\trianglelefteq$ (see the illustration below).

This pre-order in agop as `pord_weakdom()`.

```
c(pord_weakdom(5:1, 10:1), pord_weakdom(10:1, 5:1)) # 5:1 <= 10:1
## [1]  TRUE FALSE
c(pord_weakdom(3:1, 5:4),  pord_weakdom(5:4, 3:1))  # 3:1 ?? 5:4
## [1] FALSE FALSE
```

We have the following result (Gagolewski, Grzegorzewski, [11]). Let $\mathsf{F} \in \mathcal{E}(\mathbb{I})$. Then F is symmetric, nondecreasing in each variable and arity-monotonic if and only if for any $\mathbf{x}, \mathbf{y}$ if $\mathbf{x} \trianglelefteq \mathbf{y}$, then $\mathsf{F}(\mathbf{x}) \le \mathsf{F}(\mathbf{y})$. Therefore, the class of impact functions may be equivalently defined as all the aggregation operators that are nondecreasing with respect to this preorder.

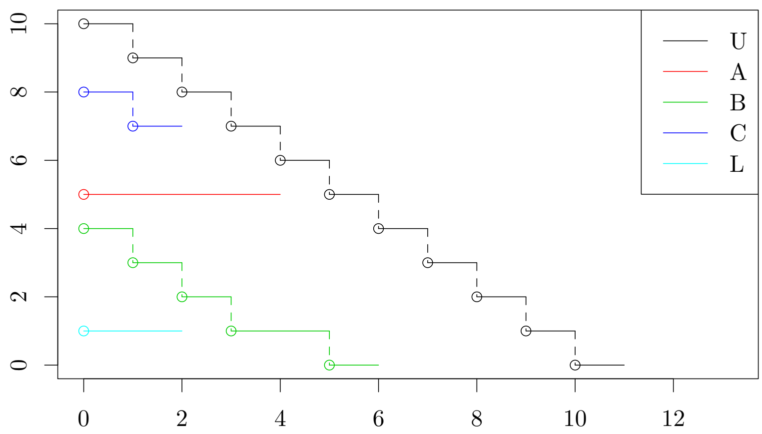
Additionally, we will write $\mathbf{x} \lhd \mathbf{y}$ if $\mathbf{x} \unlhd \mathbf{y}$ and $\mathbf{x} \neq \mathbf{y}$ (strict dominance).

Example. Let us consider the 5 following vectors.

```
ex1 <- list(
   U = 10:0,            # some upper bound
   A = c(5,5,5,5),      # moderate productivity & quality
   B = c(4,3,2,1,1,0), # high productivity
   C = c(8,7),          # high quality
   L = c(1,1)           # some lower bound
)
```

Plot of "citation" curves:

```
for (i in seq_along(ex1))
plot_producer(ex1[[i]], add=(i>1), col=i)
legend("topright", legend=names(ex1), col=1:length(ex1), lty=1)
```



get adjacency matrix for $(\{A, B, C, L, U\}, \unlhd)$....

```
ord <- rel_graph(ex1, pord_weakdom)
print(ord)
```

```
## 5 x 5 sparse Matrix of class "dtCMatrix"
##   U A B C L
## U 1 . . . .
## A 1 1 . . .
## B 1 . 1 . .
## C 1 . . 1 .
## L 1 1 1 1 1
```

```
is_reflexive(ord)  # is reflexive
```

```
## [1] TRUE
```

```
is_transitive(ord) # is transitive
```

```
## [1] TRUE
```

```
is_total(ord)      # not a total preorder...
```

```
## [1] FALSE
```

We see that we have $A??B$, $A??C$, $B??C$ (no pair from $\{A, B, C\}$ is comparable w.r.t. $\trianglelefteq$):

```
incomp <- get_incomparable_pairs(ord)
incomp <- incomp[incomp[,1]<incomp[,2],] # remove permutations: ((1,2), (2,1))->(1,2)
incomp[,] <- rownames(ord)[incomp]
print(incomp) # all incomparable pairs
```

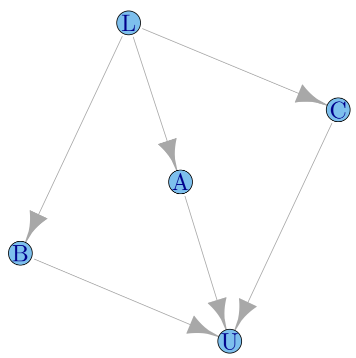
```
##      [,1] [,2]
## [1,] "A"  "B"
## [2,] "A"  "C"
## [3,] "B"  "C"
```

```
# the other way: generate maximal independent sets
lapply(get_independent_sets(ord), function(set) rownames(ord)[set])
```

```
## [[1]]
## [1] "A" "B" "C"
```

To draw the Hasse diagram, it will be good to de-transitivize the graph (for æsthetic reasons)....

```
require(igraph)
hasse <- graph.adjacency(de_transitive(ord))
set.seed(1234567) # igraph's draving facilities are far from perfect
plot(hasse, layout=layout.fruchterman.reingold(hasse, dim=2))
```



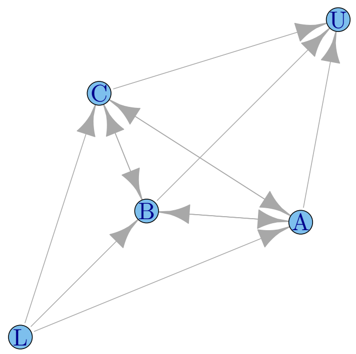
$(\{A, B, C, L, U\}, \trianglelefteq)$ is not totally ordered, let's apply fair totalization (set $x \trianglelefteq'' y$ and $y \trianglelefteq'' x$ whenever $\neg(x \trianglelefteq y$ or $y \trianglelefteq x)$ + calculate transitive closure

```
ord_total <- closure_transitive(closure_total_fair(ord)) # a total preorder
print(ord_total)
```

```
## 5 x 5 sparse Matrix of class "dgCMatrix"
##   U A B C L
## U 1 . . . .
## A 1 1 1 1 .
## B 1 1 1 1 .
## C 1 1 1 1 .
## L 1 1 1 1 1
```

```
hasse <- graph.adjacency(de_transitive(ord_total))
```

```r
set.seed(1234)
plot(hasse, layout=layout.fruchterman.reingold(hasse, dim=2))
```



...Note that each total preorder $\trianglelefteq''$ induces an equivalence relation ($x \simeq y$ iff $x \trianglelefteq'' y$ and $y \trianglelefteq'' x$; the equivalence classes may be ordered with $\trianglelefteq''$). These may be explored with the `get_equivalence_classes()` function....

```r
sapply(get_equivalence_classes(ord_total), function(set) rownames(ord)[set])
## [[1]]
## [1] "L"
##
## [[2]]
## [1] "A" "B" "C"
##
## [[3]]
## [1] "U"
```

Thus, we've obtained $L \prec (A \simeq B \simeq C) \prec U$.

6 Aggregation Operators from the Probabilistic Perspective

Theory of aggregation looks on the aggregation operators from the algebraic/calculus perspective. Of course, we should always be interested in their probabilistic properties, e.g. in i.i.d. RVs models (the simplest and the most "natural" ones in statistics), cf. [7] for discussion.

In such case we assume that input data are in fact realizations of some random samples.

In probability, an aggregation operator is simply called a *statistic* (formalism......)

Let $(X_1, \ldots, X_n)$ i.i.d. F, where $\operatorname{supp} F = \mathbb{I}$.

....

In social phenomena modeling, if F is continuous, we often assume that the underlying density f is decreasing and convex on $\mathbb{I}$, possibly with heavy-tails. E.g. in the bibliometric impact assessment problem, this assumption reflect the fact that a high paper valuation is more difficult to obtain than the lower one, most of the papers have very small valuation (near 0), and the probability of attaining a high note decreases no slower than linearly.

6.1 Some Notable Probability Distributions

6.1.1 Pareto-Type II Distribution

Many generalizations of the Pareto distribution have been proposed (GPD, *Generalized Pareto Distributions*, cf. e.g. [23, 27]). Here we will introduce the so-called Pareto-Type II (Lomax) distribution, which has support $\mathbb{I} = [0, \infty]$.

Formally, X follows the Pareto-II distribution with shape parameter $k > 0$ and scale parameter $s > 0$, denoted $X \sim \mathrm{P2}(k, s)$, if its density is of the form

$$f(x) = \frac{ks^k}{(s+x)^{k+1}} \quad (x \geq 0). \tag{4}$$

The cumulative distribution function of X is then:

$$F(x) = 1 - \frac{s^k}{(s+x)^k} \quad (x \geq 0). \tag{5}$$

TO DO: agop: `dpareto2()` – (4), `ppareto2()` – (5), and `qpareto2()`... `rpareto2()`......

Properties. The expected value of $X \sim \mathrm{P2}(k, s)$ exists for $k > 1$ and is equal to

$$\mathbb{E}X = \frac{s}{k-1}.$$

Variance exists for $k > 2$ and is equal to

$$\mathrm{Var}\, X = \frac{ks^2}{(k-2)(k-1)^2}.$$

More generally, the i-th raw moment for $k > i$ is given by:

$$\mathbb{E}X^i = \frac{\Gamma(i+1)\Gamma(k-i)}{\Gamma(k+1)} ks^i.$$

For a fixed s, if $X \sim \mathrm{P2}(k_x, s)$ and $Y \sim \mathrm{P2}(k_y, s)$, $k_x < k_y$, then X stochastically dominates Y, denoted $X \succ Y$. On the other hand, for a fixed k, $X \sim \mathrm{P2}(k, s_x)$, $Y \sim \mathrm{P2}(k, s_y)$, $s_x > s_y$, implies $X \succ Y$.

Interestingly, if $X \sim \mathrm{P2}(k, s)$, then the conditional distribution of $X - t$ given $X > t$, is $\mathrm{P2}(k, s+t)$ $t \geq 0$.

Additionally, it might be shown that if $X \sim \mathrm{P2}(k, s)$, then $\ln(s + X)$ has c.d.f. $F(x) = 1 - s^k e^{-kx}$ and density $f(x) = ks^k e^{-kx}$ for $x \geq \ln s$, i.e. has the same distribution as $Z + \ln s$, where $Z \sim \mathrm{Exp}(k) \equiv \Gamma(1, 1/k)$.

Parameter estimation. Let $\mathbf{x} = (x_1, \ldots, x_n)$ be a realization of the Pareto-Type II i.i.d. sample with known $s > 0$. The unbiased (corrected) maximum likelihood estimator for k:

$$\widehat{k}(\mathbf{x}) = \frac{n-1}{\sum_{i=1}^n \ln\left(1 + \frac{1}{s}x_i\right)}.$$

It may be shown that for $n > 2$ it holds $\mathrm{Var}\,\widehat{k}(\mathbf{x}) = k^2 \frac{1}{n-2}$.

TO DO: agop: `pareto2.mlekestimate()`

For both unknown k and s we have:

$$\begin{cases} \widehat{k} = \frac{n}{\sum_{i=1}^n \ln\left(1+x_i/\widehat{s}\right)}, \\ 1 + \frac{1}{n}\sum_{i=1}^n \ln\left(1+x_i/\widehat{s}\right) - \frac{n}{\sum_{i=1}^n \left(1+x_i/\widehat{s}\right)^{-1}} \quad = \quad 0. \end{cases} \tag{6}$$

The second equation must be solved, unfortunately, numerically. The estimation procedure has been implemented in `agop` as TO DO: `pareto2.mleksestimate()`.... It is worth noting that the above system of equations may sometimes have no solution (as the local minimum of the likelihood function may not exist, see [3] for discussion).

In this case one of the estimators worth noting (and often better than MLE) was proposed in [28]. The Zhang-Stevens MMS (*minimum mean square error*) (Bayesian) estimator has relatively small bias (often positive) and mean squared error. In `agop` it is available as TO DO: `pareto2.zsestimate()`.

Goodness-of-fit tests. TO BE DONE....

Applications. TO DO

Two-sample F-test. The following simple test was introduced in [7]. Let $(X_1, X_2, \ldots, X_{n_1})$ *i.i.d.* $\mathrm{P2}(k_1, s)$ and $(Y_1, Y_2, \ldots, Y_{n_2})$ *i.i.d.* $\mathrm{P2}(k_2, s)$, where s is an a-priori known scale parameter. We are going to verify the null hypothesis $H_0 : k_1 = k_2$ against the two-sided alternative hypothesis $K : k_1 \neq k_2$.

It might be shown that $\sum_{i=1}^n \ln(s + X_i) - n \ln s \sim \Gamma(n, 1/k)$. This implies that under H_0, the following test statistic follows the Snedecor F distribution:

$$R(\mathbf{X}, \mathbf{Y}) = \frac{n_1}{n_2} \frac{\sum_{i=1}^{n_2} \ln\left(1 + \frac{Y_i}{s}\right)}{\sum_{i=1}^{n_1} \ln\left(1 + \frac{X_i}{s}\right)} \overset{H_0}{\sim} \mathrm{F}^{[2n_2, 2n_1]}. \tag{7}$$

The null hypothesis is accepted iff

$$R(\mathbf{x}, \mathbf{y}) \in \left[\mathtt{qf}(\tfrac{\alpha}{2}, 2n_2, 2n_1),\ \mathtt{qf}(1 - \tfrac{\alpha}{2}, 2n_2, 2n_1)\right],$$

where $\mathtt{qf}(q, d_1, d_2)$ denotes the q-quantile of $\mathrm{F}^{[d_1, d_2]}$

The p-value may be determined as follows:

$$p = 2\left(\tfrac{1}{2} - \left|\mathtt{pf}(R(\mathbf{x}, \mathbf{y}), 2n_2, 2n_1) - \tfrac{1}{2}\right|\right), \tag{8}$$

where $\mathtt{pf}(x, d_1, d_2)$ is the c.d.f. of $\mathrm{F}^{[d_1, d_2]}$.

TO DO: `pareto2.ftest()`.

6.2 Stochastic Properties of Aggregation Operators

OWA, L-statistics

OWMax, S-statistics

h-index and its distribution

7 NEWS/CHANGELOG

```
                        ** agop package NEWS **

****************************************************************************

0.1-0 /under development/

* initial release
  [the package started as a lightweight fork of the CITAN package]

****************************************************************************
```

Acknowledgments. This document has been generated with LaTeX, knitr and the tikzDevice package for R. Their authors' wonderful work is fully appreciated.

The contribution of Marek Gagolewski was partially supported by the European Union from resources of the European Social Fund, Project PO KL "Information technologies: Research and their interdisciplinary applications", agreement UDA-POKL.04.01.01-00-051/10-00 (March-June 2013), and by FNP START Scholarship from the Foundation for Polish Science (2013).

Bibliography

[1] Alonso, S., Cabrerizo, F. J., Herrera-Viedma, E., and Herrera, F. *h*-index: A review focused on its variants, computation and standardization for different scientific fields. *Journal of Informetrics 3* (2009), 273–289.

[2] Cena, A., and Gagolewski, M. OM3: ordered maxitive, minitive, and modular aggregation operators – Part II: A simulation study. In *Proc. AGOP 2013*. 2013. in press.

[3] del Castillo, J., and Daoudi, J. Estimation of the Generalized Pareto Distribution. *Statistics and Probability Letters 79* (2009), 684–688.

[4] Egghe, L. An improvement of the *H*-index: the *G*-index. *ISSI Newsletter 2*, 1 (2006), 8–9.

[5] Franceschini, F., and Maisano, D. A. The Hirsch index in manufacturing and quality engineering. *Quality and Reliability Engineering International 25* (2009), 987–995.

[6] Franceschini, F., and Maisano, D. A. Structured evaluation of the scientific output of academic research groups by recent *h*-based indicators. *Journal of Informetrics 5* (2011), 64–74.

[7] Gagolewski, M. *Aggregation Operators and Their Application in a Formal Model for Quality Evaluation System of Scientific Research (Wybrane operatory agregacji i ich zastosowanie w modelu formalnym systemu jakości w nauce)*. PhD thesis, 2011. (In Polish).

[8] Gagolewski, M. On the relationship between symmetric maxitive, minitive, and modular aggregation operators. *Information Sciences 221* (2013), 170–180.

[9] Gagolewski, M., and Cena, A. *agop: Aggregation Operators in R*, 2013. www.rexamine.com/resources/agop/.

[10] Gagolewski, M., and Grzegorzewski, P. A geometric approach to the construction of scientific impact indices. *Scientometrics 81*, 3 (2009), 617–634.

[11] Gagolewski, M., and Grzegorzewski, P. Possibilistic analysis of arity-monotonic aggregation operators and its relation to bibliometric impact assessment of individuals. *International Journal of Approximate Reasoning 52*, 9 (2011), 1312–1324.

[12] Grabisch, M., Marichal, J.-L., Mesiar, R., and Pap, E. *Aggregation functions*. Cambridge University Press, 2009.

[13] Hirsch, J. E. An index to quantify individual's scientific research output. *Proceedings of the National Academy of Sciences 102*, 46 (2005), 16569–16572.

[14] Kosmulski, M. A new Hirsch-type index saves time and works equally well as the original h-index. *ISSI Newsletter 2*, 3 (2006), 4–6.

[15] Kosmulski, M. MAXPROD — A new index for assessment of the scientific output of an individual, and a comparison with the h-index. *Cybermetrics 11*, 1 (2007).

[16] Quesada, A. Monotonicity and the Hirsch index. *Journal of Informetrics 3*, 2 (2009), 158–160.

[17] Quesada, A. More axiomatics for the Hirsch index. *Scientometrics 82* (2010), 413–418.

[18] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL: http://www.R-project.org.

[19] Rousseau, R. Woeginger's axiomatisation of the h-index and its relation to the g-index, the $h(2)$-index and the R^2-index. *Journal of Informetrics 2*, 4 (2008), 335–340.

[20] Shilkret, N. Maxitive measure and integration. *Indag. Math. 33* (1971), 109–116.

[21] Sugeno, M. *Theory of fuzzy integrals and its applications*. PhD thesis, Tokyo Institute of Technology, 1974.

[22] Torra, V., and Narukawa, Y. The h-index and the number of citations: Two fuzzy integrals. *IEEE Transactions on Fuzzy Systems 16*, 3 (2008), 795–797.

[23] Villasenor-Alva, J., and Gonzalez-Estrada, E. A bootstrap goodness of fit test for the Generalized Pareto Distribution. *Computational Statistics and Data Analysis 53*, 11 (2009), 3835–3841.

[24] Woeginger, G. J. A symmetry axiom for scientific impact indices. *Journal of Informetrics 2* (2008), 298–303.

[25] Woeginger, G. J. An axiomatic analysis of Egghe's g-index. *Journal of Informetrics 2*, 4 (2008), 364–368.

[26] Woeginger, G. J. An axiomatic characterization of the Hirsch-index. *Mathematical Social Sciences 56*, 2 (2008), 224–232.

[27] Zhang, J. Improving on estimation for the Generalized Pareto Distribution. *Technometrics 52*, 3 (2010), 335–339.

[28] Zhang, J., and Stephens, M. A. A new and efficient estimation method for the Generalized Pareto Distribution. *Technometrics 51*, 3 (2009), 316–325.