Linear Optimization for Ecological Indices Based on Aggregation Functions

Gleb Beliakov[1], Andrew Geschke[1], Simon James[1](✉), and Dale Nimmo[2]

[1] Deakin University, Geelong, Australia
{gleb,sjames}@deakin.edu.au, geschke.andrew@gmail.com
[2] Charles Sturt University, Albury, Australia
dnimmo@csu.edu.au

Abstract. We consider an optimization problem in ecology where our objective is to maximize biodiversity with respect to different land-use allocations. As it turns out, the main problem can be framed as learning the weights of a weighted arithmetic mean where the objective is the geometric mean of its outputs. We propose methods for approximating solutions to this and similar problems, which are non-linear by nature, using linear and bilevel techniques.

Keywords: Aggregation functions · Linear programming · Weight learning · Ecology · Biodiversity

1 Introduction

We consider the problem of distributing a human population across a finite land area in such a way that negative impact to local flora and fauna is minimized. A simple version of the problem, optimization of abundance for a single species, is one that is easily solved with a linear programming approach, however ecologists are usually more interested in how the land-use allocations affect *biodiversity*. A number of quantitative indices exist for biodiversity, which incorporate both the number of species present (the richness) along with how evenly distributed the species are. As has been observed in [1], many of these can be expressed in terms of common aggregation functions. For instance, the geometric mean of species abundances is being increasingly used as a proxy for biodiversity [2], providing an average abundance that is more sensitive to smaller values (rare species). Whilst the optimization of these more complicated indices is non-linear in nature, we will show that close approximations can be achieved using mostly linear techniques and capitalizing on the ability to express quasi-arithmetic means in terms of generating functions. We illustrate the techniques using bird occupancy data from surveys conducted in Melbourne, Australia, and include details of our implementations as an appendix.

The article will be set out as follows: In Sect. 2, we outline our notation along with the necessary underlying concepts from the field of aggregation functions. In Sect. 3, we introduce the ecological context and go through the associated

J.P. Carvalho et al. (Eds.): IPMU 2016, Part II, CCIS 611, pp. 411–422, 2016.
DOI: 10.1007/978-3-319-40581-0_33

problems with our proposed optimization solutions. We then provide an example in Sect. 4, before concluding in Sect. 5.

2 Preliminaries

We will give an overview of the preliminary concepts of aggregation functions as relevant to the problem of land-use allocation. We consider an input dataset consisting of an $m \times n$ matrix where the entries x_{ij} denote the predicted abundances of the i-th species for the j-th land-use type. In practice, such abundances are measured by reporting rates calculated after conducting surveys. In our case, the n land-use types correspond with increasing densities of human population but these need not be numeric or even ordered. With respect to a total human population P and available land area A, the values w_j denote the percentage allocation to each land-type, so that $\sum_{j=1}^{n} w_j = 1$. These values will correspond with the weights of our aggregation functions.

Aggregation functions are employed in various contexts for summarizing data. Overviews of the important families, properties and definitions can be found in [3–6].

Definition 1. *An aggregation function* $f : [a,b]^n \to [a,b]$ *is a function monotone in each argument and satisfying the boundary conditions* $f(a,\ldots,a) = a$ *and* $f(b,\ldots,b) = b$ *(with* $a < b$*).*

Of particular interest to us is the weighted arithmetic mean, perhaps the most commonly employed aggregator across various contexts. It is expressed,

$$WAM(x_1,\ldots,x_n) = \sum_{j=1}^{n} w_j x_j. \tag{1}$$

In our case, for a given species i, the aggregated value $WAM(x_{i1},\ldots,x_{in})$ denotes its abundance per unit of area.

Another aggregation function important in ecology is the geometric mean. For an input vector $\mathbf{x}$, the geometric mean is given by,

$$G(x_1,\ldots,x_n) = \left(\prod_{j=1}^{n} x_j\right)^{\frac{1}{n}}. \tag{2}$$

In ecology, the geometric mean of species abundance is often used to give a measure of abundance that is more sensitive to rare species. So if we have $\mathbf{s} = (s_1, s_2, \ldots, s_m)$ denoting the set of species abundances for each of the m species, $G(\mathbf{s})$ is a proxy measure for biodiversity. We note that $G(\mathbf{s}) \le AM(\mathbf{s})$ where AM is the weighted arithmetic mean with equal weights, and that the values will be closer the more even the species abundances are.

The geometric mean can also be obtained as a special case of the quasi-arithmetic mean, which generalizes[1] the WAM. Specifically, we have

$$G(x_1, \ldots, x_n) = g^{-1}\left(\sum_{j=1}^{n} w_j g(x_j)\right), \tag{3}$$

where $g(t) = \ln t$ is the generating function, its inverse is $g^{-1}(t) = \exp(t)$ and in our context we have $w_j = 1/n$ for all j.

3 Finding Optimum Land-Use Allocations with Respect to Species Diversity

The process of urbanization is a major contributor to biodiversity loss [7], with the expansion of cities leading to habitat loss, climatic changes in temperature as well as other disruptions to local species dynamics. However while some species respond negatively to increases in human population density, other species (pigeons for example) can actually benefit. In planning for the development of cities and towns, two theories of conservation have arisen in ecology literature [8]: *land-sharing*, whereby the human population is spread as evenly as possible over a given area; and *land-sparing*, which fits the human population to small areas of high density so that the remaining area can be reserved to preserve flora and fauna.

The way individual species respond to changes in human population density can be considered in terms of response curves (see Fig. 1).

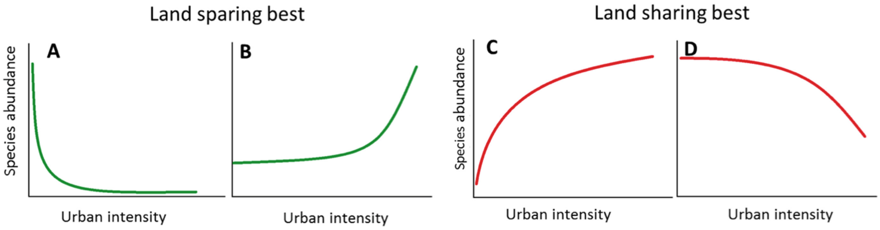

Fig. 1. Examples of response curves for species that benefit most from either a land-sparing or land-sharing approach to urban development. Species that respond to urban density according to curves like A and B are best suited to land-sparing, since it aims to segment a city into either very high or very low density living (where the abundances for these species are highest). On the other hand, response curves like C and D relate to species who would be better off with a land-sharing approach, since they have high abundance for mid-range urban intensity.

[1] More information about such generalizations can be found in any of [3–6], however we will restrict ourselves to the relevant cases to our problem.

We consider different ranges of population density to constitute a 'land-type', with each species having a predicted abundance. For example, if we considered 5 levels of density, for each species we would have data of the following form.

Land type (j)	1	2	3	4	5
Human population density (per 25ha)	0	50	100	250	500
Species abundance (per 25ha)	20	14	12	10	7

In the following subsections, we will present methods for finding the best allocation of each land-type subject to area and human population constraints where we are interested in optimizing either (1) the total abundance of all species, (2) the geometric mean of species abundances, and (3) the biodiversity as calculated using Shannon's diversity index. We focus on linear methods over more general approaches for two reasons. Firstly, although the dataset we use here is relatively small, both the number of species and the number of land-types can potentially be very large in practice and we want the method to be scalable. Secondly, we have a number of constraints that are more difficult to incorporate in more general optimisation models[2], such as the land area and population.

3.1 Optimization of Total Abundance

In previous works we have used linear optimization to learn the weights of various aggregation functions from data [9–14]. In those cases, we considered a set of input and output pairs with the aim of minimizing differences between observed and predicted outputs. Our aim here is to find the best allocation of land types. The percentage allocations which correspond with our aggregation weights w_j are our decision variables. We denote by d_j the population density for the j-th land-type. For any given species i, we have

$$\begin{aligned} \underset{\mathbf{w}}{\text{Maximize}} \quad & \sum_{j=1}^{n} w_j x_{ij}, \\ \text{s.t.} \quad & A\sum_{j=1}^{n} w_j d_j \geq P, \\ & \sum_{j=1}^{n} w_j = 1, w_j \geq 0, j = 1, \ldots, n. \end{aligned} \tag{4}$$

In order to maximize the total abundance, we note that we can simply sum the total abundances across all species for each land-type, so that the objective equation becomes,

[2] However since the constraints are linear, quadratic programming formulations would also be fine.

$$\underset{\mathbf{w}}{\text{Maximize}} \sum_{j=1}^{n} w_j \left(\sum_{i=1}^{m} x_{ij} \right). \tag{5}$$

3.2 Maximizing the Geometric Mean of Species Abundances

As discussed previously, we are often more interested in maximizing the geometric mean of abundances, which is more sensitive to rare species. This is so that the impression of abundance is not inflated by having a very common species. Our objective becomes,

$$\underset{\mathbf{w}}{\text{Maximize}} \prod_{j=1}^{m} \left(\sum_{j=1}^{n} w_j x_{ij} \right)^{\frac{1}{n}}. \tag{6}$$

We can ignore the $1/n$ power since the product and geometric mean will have the same maximum. This is still a non-linear objective, however we can use Eq. (3) and consider maximizing the sum of the logarithms of each species. In terms of the decision variables we have,

$$\underset{\mathbf{w}}{\text{Maximize}} \sum_{i=1}^{m} \ln \left(\sum_{j=1}^{n} w_j x_{ij} \right), \tag{7}$$

and whilst this representation remains non-linear, we can find an approximate solution to any desired precision by taking advantage of the fact that the log function is concave and hence can be expressed as the maximum value with respect to a set of bounding linear equations.

We transform the log function and write it as,

$$\ln t = \lim_{K \to \infty} \min \left(f_1(t),\, f_2(t),\, f_3(t), \ldots, f_K(t) \right),$$

where $f_k(t)$ denote the tangent lines of $\ln t$ across its domain, with $f_k(t) = \alpha_k t + \beta_k, \alpha_k = \frac{d}{dt}(\ln t_k) = 1/t_k, \beta_k = \ln t_k - \alpha_k t_k$ where t_k are the points at which log is evaluated. In other words, the logarithm is expressed in terms of the minimum of its K affine functions. Figure 2 helps demonstrate this visually.

Equation (7) hence becomes piecewise linear and the objective can be reduced to a linear program if the constraints are also linear. For each species i and each of our tangent functions given by $f_k(t) = \alpha_k t + \beta_k$, we introduce constraints of the form, $-\alpha_k s_i + y_i \leq \beta_k$, where s_i is the abundance of the i-th species, i.e.

$$-\alpha_k(w_1 x_{i1} + w_2 x_{i2} + \ldots + w_n x_{in}) + y_i \leq \beta_k.$$

The variables y_i now become decision variables in the optimization formulation. We optimize for the maximum sum of these values, however each y_i is bounded from above by the tangent lines described by the K constraints.

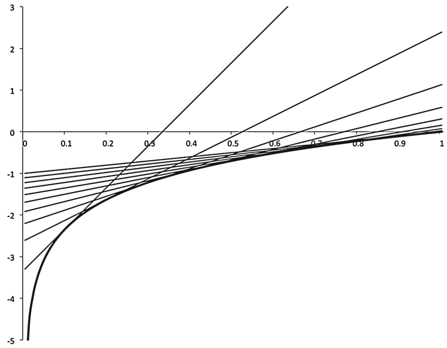

Fig. 2. Example of the natural log function being bound from above by 10 approximating affine functions, equispaced over the interval $[0, 1]$.

Summarizing, we have the following linear programming formulation for this problem.

$$\begin{gathered}\underset{\mathbf{w},\mathbf{y}}{\text{Maximize}} \ \sum_{i=1}^{m} y_i \\ \text{s.t.} \quad A\sum_{j}^{n} w_j d_j \geq P. \\ -\alpha_k A(w_1 x_{i1} + w_2 x_{i2} + \ldots + w_n x_{in}) + y_i \leq \beta_k, k = 1, \ldots, K, i = 1, \ldots, m \\ \sum_{j=1}^{n} w_j = 1, \quad w_j \geq 0, j = 1, \ldots, n. \end{gathered} \tag{8}$$

3.3 Maximization of Shannon's Diversity

Whilst the approach of the previous section could be adapted to the optimization of any convex function of abundance values that can be expressed as the sum of generating functions, there are a number of ecological indices that are not based on species abundances but rather on proportional abundance, i.e. the values p_i such that,

$$p_i = \frac{\sum\limits_{j=1}^{n} w_j x_{ij}}{\sum\limits_{i=1}^{m}\sum\limits_{j=1}^{n} w_j x_{ij}}.$$

Shannon's diversity index is one such example, expressed in terms of the p_i,

$$\sum_{i=1}^{m} -p_i \ln p_i. \tag{9}$$

For a given number of species (m) present in a given community, it reaches a maximum $\ln m$ when all species have equal abundance, and approaches zero if a single species dominates, i.e. if $p_i = 1$ for any i.

The use of p_i values makes it impossible to express this in terms of linear multiples of the constraints as we did previously, however we can take a different approach that results in a bi-level optimization problem. As we will see, it remains feasible for finding practical solutions with real datasets.

We introduce a variable M which indicates the total abundance, i.e.

$$M = \sum_{i=1}^{m} w_j \left(\sum_{j=1}^{n} x_{ij} \right).$$

With M known, we can therefore use this to scale our variables so that they are equivalent to proportions. We then have the capacity to solve the optimization, *provided* we know M. As before, we create affine functions from the curve, $-t \ln t$ and maximize such that the y_i values are bounded by these lines. In this case, accuracy bounds pose less of a problem since we know that all p_i are less than 1. The constraints will now be of the form,

$$-\alpha_k(w_1 \frac{x_{i1}}{M} + w_2 \frac{x_{i2}}{M} + \ldots + w_n \frac{x_{in}}{M}) + y_i \leq \beta_k.$$

We then can find the M that gives the best result for Shannon's diversity, which we implement as a bilevel problem. We have,

$$\underset{M}{\text{Maximize}}\ Z$$

$$Z = \max_{\mathbf{w},\mathbf{y}} \sum_{i=1}^{m} y_i$$

$$\text{s.t. } A \sum_{j}^{n} w_j d_j \geq P$$

$$-\frac{\alpha_k A}{M}(w_1 x_{i1} + w_2 x_{i2} + \ldots + w_n x_{in}) + y_i \leq \beta_k, k = 1, \ldots, K, i = 1, \ldots, m$$

$$\sum_{j=1}^{n} w_j = 1, \quad w_j \geq 0, j = 1, \ldots, n. \qquad (10)$$

4 Example: Bird Surveys Data

Survey data reporting presence or absence of bird species across 28 landscapes in the wider Melbourne area was collected over a period of four months (May to August) in 2015. All landscapes were one hectare in area and the human population densities were determined from census data. The report rates for each species were the result of four separate observation rounds. An example of the fitted response curves (showing the probability of a species occurring in a landtype with that human density) across the 28 sites for three bird species are shown in Fig. 3.

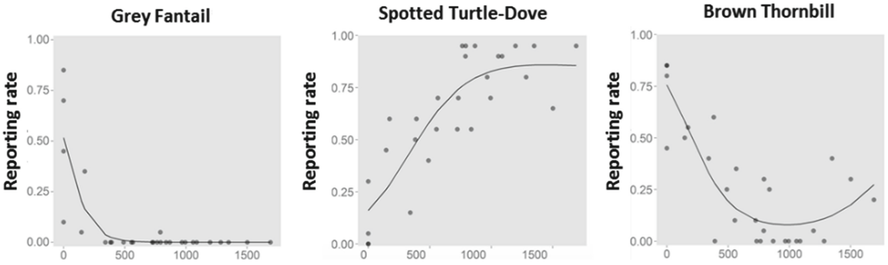

Fig. 3. Data collected for three bird species at sites with increasing human population density. While the *grey fantail* virtually only lives in reserves (i.e. all vegetation and no housing), the *spotted turtle dove* benefits from high urban population density and the *brown thornbill* was present across the range.

The dataset we will use to illustrate the methods proposed here includes that relating to 21 native species, with response rates calculated at densities from 0 to 1600 at intervals of 100 ($n = 17$ land types). We consider allocating the optimum allocation of a human population of 2.744 million, i.e. the current population of the Melbourne residential area (outside the central business district). The area under consideration is 964 km^2.

Table 1. Summary results from applying the methods for optimizing total abundance, the geometric mean, and Shannon's diversity index respectively.

Objective	Densities					
	w_1	w_2	w_3	Total	Geometric	Shannon's
	0	1000	1600	Abundance	Mean	Diversity
Abundance	0.5551	0	0.4449	17445	549.4	2.7425
Geometric mean	0.5210	0.0911	0.3879	17068	564.3	2.7443
Shannon's diversity	0.5247	0.0813	0.3941	17108	564.0	2.7443

* weights and Shannon's diversity rounded to 4 dp, geometric mean rounded to 1 dp

The results are shown in Table 1. These do not vary greatly based on the objective used, however we do note slight changes. Obviously all three measures will be somewhat correlated, with each essentially capturing some overall measure of how many individuals are present. We have only displayed three weights because in all models the remaining densities were all given zero allocation, regardless of the method. In terms of the ecological interpretation, we see that the *land-sparing* approach to biodiversity conservation is preferred overall for this particular set of species, allocating areas of high population density as well as reserves for wildlife, with only small amounts of land at medium population density (Table 2).

Table 2. Summary results from applying the methods for optimizing total abundance, the geometric mean, and Shannon's diversity index respectively with a smaller subset of the data (only 4 species).

Objective	Densities							
	w_1	w_2	w_3	w_4	w_5	Total	Geom.	Shan.
	0	600	800	1000	1600	Abund.	Mean	Div.
Abundance	0.1102	0	0.8898	0	0	5736	990.0	1.1106
Geometric mean	0.3582	0	0	0.5251	0.1167	5534	1079.6	1.2156
Shannon's diversity	0.3499	0.3284	0	0	0.3217	5244	1054.3	1.2289

* weights and Shannon's diversity rounded to 4 dp, geometric mean rounded to 1 dp

To help give some insight into the difference between these approaches, we have also optimized for a smaller set of species. We used the four species with response curves shown in Fig. 4. We note that in this case, we have four very different response curves including quite common species, e.g. the *Australian magpie*, and rare species such as the *eastern rosella*.

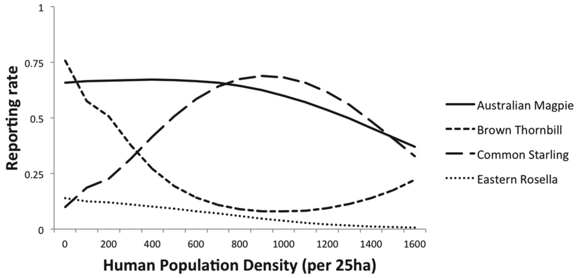

Fig. 4. Four species and their response curves. The *Australian magpie* is present at high levels across all population densities while the *brown thornbill* and the *eastern rosella* both generally decrease as the human population becomes more and more dense.

In Fig. 5 we can observe how each of the species change individually with respect to the different optimization objectives. When the overall abundance is maximized, low abundance in the *brown thornbill's* population is compensated for by high abundance with the *Australian magpie* and *common starling*[3].

[3] As a side note the *common starling* is an introduced species and was not included in the previous example of 21 native bird species.

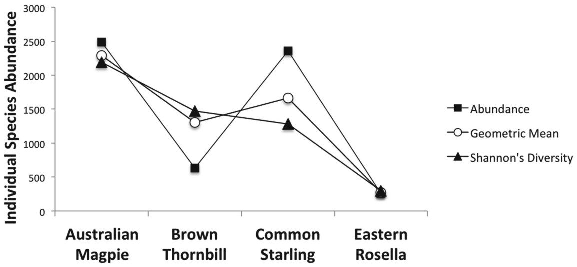

Fig. 5. Comparison of individual species' abundances obtained with each of the optimization objectives.

Both the geometric mean and Shannon's diversity approaches obtain a more evenly spaced distribution for these three species, however we note that in all cases we are not able to raise the abundance of the *eastern rosella.*

5 Conclusion

We have focused on an optimization problem that arises in ecology where land management decisions can have an impact on the local biodiversity. In this context, the optimization problems take on a similar form to what we have in learning parameters for aggregation functions, however with objective functions whose arguments are expressible in terms of weighted arithmetic means of our data. We have shown that these non-linear objectives can be approximated with linear techniques, the advantages of which are that they are quickly solvable, are guaranteed to reach a global optimum and are scalable in terms of time and computation complexity. In our main problem involving the geometric mean of species abundances, we represented its additive generating function as the maximum of bounding affine functions. We have made all R algorithms available at our website.

Appendix: Implementations

We have implemented all three approaches to optimization as functions in an R library available at our website[4].

Optimizing total abundance: `eco.opti()`.

Description of inputs

`species.data` - matrix of species abundances per unit of land area, i.e. with x_{ij} denoting the i-th species and its abundance for land type j;

[4] http://aggregationfunctions.wordpress.com.

`densities` - vector of densities per unit of land area, i.e. with d_j denoting the human population density for a given land-type j;
`tot.pop` - the total population required to be fit into the given land area;
`tot.land` - the total area over which we need to distribute the population.

Additional optional constraints

`w.min / w.max` - vectors denoting minimum or maximum bounds on the land-types, e.g. if we want to ensure that at least 20 % of the land is populated at minimum density we incorporate the constraint $w_1 \geq 0.2$ (assuming w_1 is the land-type with minimum density), or alternatively we may wish to limit high density housing to at most 40 % of the land area etc.;
`spec.min / spec.max` - vector placing minimum or maximum bounds for a particular species, for example, if we want to make sure that a rare species is above a given threshold γ_i, the linear constraint $\sum\limits_{j=1}^{n} w_j x_{ij} \geq \gamma_i$ is added for that species.

The function also gives as output a number of ecological indices such as the individual species abundances, and the Simpson and Shannon diversity indices.

Maximizing the geometric mean of species abundances: `eco.opti.gm()`

Description of inputs

In addition to all inputs and constraints used with `eco.opti()`, this function has two additional optional parameters to control the precision.

`fprec` - a positive integer giving the number of tangent functions to be defined. The default setting is 100 linear segments, and so gains in accuracy can be achieved with settings of 500, 1000 etc., however obviously at the cost of computation time;
`max.x` - a real number giving the maximum value for the domain over which the tangent functions are calculated, the default setting is 10000, and so depending on the scale given it could be necessary to increase this value (or decrease it for finer accuracy) or the optimization will be the same as it would be for maximizing abundance.

For the number of tangent lines K, optimizing over 5 species with 100 linear segments will require $5 \times 100 = 500$ additional constraints, use of 1000 linear segments will require 5000 additional constraints and so on. We need to be careful when reducing the precision, since the log function's gradient changes more drastically for values closer to zero than it does for large values. Rather than taking equal step sizes in calculating our tangent lines, they were distributed using $t_k = \exp(-k \cdot \max(s)/K)$ where $\max(s)$ is the `max.x` parameter above.

We note also that by default the setting for `spec.min` will be 1 for all species. It could be adjusted to a fractional value if desired.

Maximizing Shannon's diversity index: `eco.opti.sh()`

This function uses the same inputs as with the previous two. The program first solves a maximum and minimum problem using `eco.opti()` in order to find the feasible bounds to search for M. Note that $-t \ln t$ is concave for $t \in (0, 1]$.

Another biodiversity index used as an objective and included in the code made available online is Simpson's diversity index $1/(\sum_{i=1}^{m} p_i^2)$. This is performed in a similar manner, however now we are minimizing for a convex function $y = t^2$ rather than maximizing for a concave function and so we need to make the appropriate changes when using the linear framework above.

References

1. Tuomisto, H.: An updated consumer's guide to evenness and related indices. Oikos **121**, 1203–1218 (2012)
2. Kelly, L.T., Bennett, A.F., Clarke, M.F., McCarthy, M.A.: Optimal fire histories for biodiversity conservation. Conserv. Biol. **29**, 473–481 (2015)
3. Beliakov, G., Bustince, H., Calvo, T.: A Practical Guide to Averaging Functions. Springer, New York (2015)
4. Beliakov, G., Pradera, A., Calvo, T.: Aggregation Functions: A Guide for Practitioners. Springer, Heidelberg (2007)
5. Grabisch, M., Marichal, J.-L., Mesiar, R., Pap, E.: Aggregation Functions. Cambridge University Press, Cambridge (2009)
6. Torra, Y., Narukawa, V.: Modeling Decisions. Information Fusion and Aggregation Operators. Springer, Heidelberg (2007)
7. Grimm, N.B., Faeth, S.H., Golubiewski, N.E., Redman, C.L., Wu, J., Bai, X., Briggs, J.M.: Global change and the ecology of cities. Science **319**, 756–760 (2008)
8. Lin, B.B., Fuller, R.A.: FORUM: sharing or sparing? how should we grow the world's cities? J. Appl. Ecol. **50**, 1161–1168 (2013)
9. Beliakov, G., James, S.: Citation-based journal ranks: the use of fuzzy measures. Fuzzy Sets Syst. **167**, 101–119 (2011)
10. Beliakov, G., James, S.: Using linear programming for weights identification of generalized Bonferroni means in R. In: Torra, V., Narukawa, Y., López, B., Villaret, M. (eds.) MDAI 2012. LNCS, vol. 7647, pp. 35–44. Springer, Heidelberg (2012)
11. Beliakov, G., James, S., Gómez, D., Rodríguez, J.T., Montero, J.: Learning stable weights for data of varying dimension. In: Proceedings of the 8th International Summer School on Aggregation Operators, University of Silesia, Katowice, Poland (2015)
12. Beliakov, G., James, S., Gómez, D., Rodríguez, J.T., Montero, J.: Approaches to learning strictly-stable weights for data with missing values. Fuzzy Sets and Systems (2016, submitted)
13. Beliakov, G., James, S., Nimmo, D.: Learning aggregation weights from 3-tuple comparison sets. In: Proceedings of IFSA 2014, Edmonton, Canada. pp. 1–6 (2013)
14. Beliakov, G., James, S., Nimmo, D.: Using aggregation functions to model human judgements of species diversity. Inf. Sci. **306**, 21–33 (2015)