Measures of Balance in Combinatorial Optimization

Philippe Olivier[1,2], Andrea Lodi[1,2], and Gilles Pesant[1]

Polytechnique Montréal, Montreal, Canada[1]
CERC[2]
{philippe.olivier, andrea.lodi, gilles.pesant}@polymtl.ca

Abstract. The concept of balance plays an important role in many combinatorial optimization problems. Yet there exist various ways of expressing balance, and it is not always obvious how best to achieve it. In this methodology-focused paper we study two cases where its integration is deficient and analyze the causes of these inadequacies. We examine the characteristics and performance of the balancing methods used in these cases, and provide general guidelines regarding the choice of a method.

1 Introduction

It is natural to think of *balance* as "things being as equal as possible." Yet this notion of equality is hard to define. Suppose we have candy bags of assorted sizes (5, 5, 6, 7, 9, 12, and 12 candies) which we want to distribute fairly to four children. We easily observe that a perfect distribution of 14 candies per children is not possible. What, then, constitutes a fair distribution?

We could consider as our criterion of fairness that the largest share of candies be as small as possible, which would give us handouts of 12, 12, 16, and 16 candies. Or we could instead consider an alternative criterion and ensure that the sum of candy discrepancies from the mean is minimal, giving us handouts of 12, 13, 14, and 17 candies. We can notice that the optimal solution for one criterion of fairness is not optimal for the other, and vice-versa. These options are both "fair," yet neither is intrinsically better or worse than the other.

This simple example illustrates how the notion of balance becomes more ambiguous after scratching the surface. Besides, this notion deserves special attention, as fairness is of paramount importance in several practical situations. In many jurisdictions of the United States, for instance, algorithms have taken the role of decision-makers for delicate matters such as deciding whether a defendant awaiting trial should be released or not. Racial disparities being a sensitive issue, these algorithms need to ensure fairness in this process [1]. One study found that African-Americans were one and a half times more likely than Caucasians to be wrongly classified as high risk by one of these algorithms [2].

While this paper is concerned with balance in the context of combinatorial optimization, issues of fairness also arise in the related field of game theory, where some criteria of fairness are of a different nature. *Envy-freeness* ensures that no player would want to trade his share for that of another, *Pareto efficiency* guarantees that no share can be improved without worsening some other share, and so on [3].

A tangential application combining balance types and fairness criteria is found in *social welfare functions*. These functions describe the collective welfare of a society based on the utilities, or satisfaction, of its individuals [4]. The utilitarian function measures collective welfare as the sum of all individual utilities, maximizing pure utility while disregarding any type equality between individuals. In contrast, the very fair egalitarian function considers the minimum of all individual utilities at the expense of a lower level of global welfare [5]. The Nash social welfare function maximizes the product of all utilities, which provides a sort of middle ground between the two previous functions [6].

In this methodology-focused paper, we present two cases supporting the hypothesis that unfamiliarity with the characteristics of balancing methods often leads to poor choices regarding balance in combinatorial problems. We examine the characteristics and performance of several methods, and provide general guidelines regarding the choice of a method. Section 2 introduces a few balancing methods and their characteristics, and outlines their use in the context of constraint programming and integer programming. Two problematic cases are studied in Sections 3 and 4. Finally, Section 5 discusses general guidelines for the application of the various balancing methods.

2 Balancing Methods at a Glance

Given a finite collection of real-number variables $X = \{x_1, x_2, \ldots, x_n\}$, its *balance* has been defined in several ways in the literature. Some methods only take into account the extremal values—constraining the most extreme points forces the others into a shorter interval. Other methods consider all values, and while usually more computationally expensive this often results in an improved distribution (relative to the aims of the problem). This section covers four common balancing methods.

The MINMAX method is rather crude and simply minimizes the maximum value. For a collection of n points, the global distance between these points and their arithmetic mean μ, according to a norm p, is defined by the concept of L_p-deviation

$$\sum_{i=1}^{n} |x_i - \mu|^p.$$

For example, L_1-DEVIATION minimizes the sum of absolute deviations from the mean, L_2-DEVIATION minimizes the sum of squared deviations from the mean, and L_∞-DEVIATION minimizes the maximum deviation from the mean.

These balancing methods display various characteristics. *Dispersion* represents the size of the interval within which the points are located, and by extension is one measure of the sensitivity to outliers. A distribution is *smooth* when its points appear evenly within this interval. The number of *outliers* near the edges of the dispersion interval is another measure of the sensitivity to outliers. Table 1 summarizes some characteristics of the four methods studied in this paper.[1]

Table 1. Some characteristics of balance.

	Dispersion	Smoothness	Outliers
MINMAX	Large	Uneven	Robust
L_1-DEVIATION	Medium	Varied	Robust
L_2-DEVIATION	Small	Varied	Sensitive
L_∞-DEVIATION	Small	Even	Sensitive

When optimized with MINMAX, values are only bounded on one side, and as such they show the largest dispersion. The other methods force bounds on both sides, with L_2- and L_∞-DEVIATION forcing especially tight bounds by nature. MINMAX offers little smoothness as many values will tend to be grouped around the bound. L_∞-DEVIATION is in contrast smoother—nothing is constraining the deviations apart from forcing them to be within the interval, so their associated values will appear somewhat randomly within this interval. Results are more varied for L_1- and L_2-DEVIATION, since there is a natural bias for values to be closer to the mean. The lack of minimum bound for MINMAX makes it robust against outliers, as does the linear expression of deviation for L_1-DEVIATION. Outliers have more influence on L_∞-DEVIATION since both small and large values disproportionately affect the objective, and are also more impactful on the quadratic expression of deviation of L_2-DEVIATION.

Marsh and Schilling [7] have surveyed measures of equity, in particular related to facility location. The authors record and briefly analyze some 20 balancing methods which are in use in various fields. They propose some guidelines to help in the choosing of a method.

Balancing in constraint programming (CP) is usually achieved through special constraints. L_1-DEVIATION is handled by the `deviation` constraint, introduced by Schaus et al. [8]. Pesant and Régin [9] balanced with L_2-DEVIATION using the `spread` constraint. The `dispersion` constraint proposed by Pesant [10] encapsulates multiple balancing methods, including L_1-, L_2-, and L_∞-DEVIATION. Other methods, such as MINMAX, can be expressed with classical constraints such as `minimum` and `maximum`.

[1] This table is constructed from the observations of the solutions of the Nurse-Patient Assignment Problem of Section 4. Observations of the “smoothness” characteristic were inconclusive for this problem, and so this observation instead stems from the solutions of simple bin packing problems.

Early work on balancing in mathematical programming includes a short paper by Gaudioso and Legato [11] presenting a few balancing methods, among which MINMAX. A recent paper by Olivier et al. [12] covers the L_1-, L_2-, and L_∞-DEVIATION methods in the context of integer programming, and compares these with equivalent CP approaches.

The next two sections present practical problems which include some form of balancing: the assignment of courses to periods such that the loads of the periods are balanced, and the assignment of patients to nurses such that the workloads of the nurses are balanced.[2] When these problems were initially introduced, their balancing methods were deficient; we will show how they have been improved.

3 Balanced Academic Curriculum Problem

The *Balanced Academic Curriculum Problem* (BACP) attempts to find an assignment of courses over a number of periods such that the academic load of a student is balanced throughout the curriculum and that course prerequisite constraints are respected. Let

- $\mathcal{C} = \{1, \ldots, n\}$ be the index set of courses,
- $\mathcal{P} = \{1, \ldots, m\}$ be the index set of periods,
- w_i denote the load of course i with $w = \sum_{i \in \mathcal{C}} w_i$ representing the combined loads of all the courses,
- $\mathcal{Q} \subset \mathcal{C} \times \mathcal{C}$ denote the set of prerequisites, where an element (i, j) indicates that course i is a prerequisite to course j.

The objective is to maximize the balance of an assignment of the n courses to the m periods. The BACP was originally introduced by Castro and Manzano [13], whose CP and IP models both achieved balance by minimizing the maximum academic load of the periods (MINMAX). Further papers by Hnich et al. [14, 15] introduced new CP and IP models using the same balancing criterion. Monette et al. [16] not only used MINMAX but also explored other options, namely balancing using the L_1-, L_2-, and L_∞-DEVIATION methods, all with a CP model. Let $L = \{L_1, \ldots, L_m\}$ denote the loads of the periods for an assignment. The four objectives studied by Monette et al. can be formalized as

$$\max_{k \in \mathcal{P}} L_k \tag{\textsc{minmax}}$$

$$\sum_{k \in \mathcal{P}} |L_k - w/m| \tag{L_1-\textsc{deviation}}$$

$$\sum_{k \in \mathcal{P}} (L_k - w/m)^2 \tag{L_2-\textsc{deviation}}$$

$$\max_{k \in \mathcal{P}} |L_k - w/m| \,. \tag{L_∞-\textsc{deviation}}$$

[2] A third problem, the distribution of bikes to stations in a bike sharing system such that the stations are balanced, has been removed for CP 2019 due to lack of space.

Starting with the premise that “neither criterion subsumes the others and there is no a priori reason to prefer one of them,” [16] they aim to determine how well each balance criterion approximates the others. Their findings are reproduced in Table 2, where rows represent optimized criteria and columns represent evaluated criteria. For example, at the intersection of row “L_1-DEVIATION” and column “MINMAX” is the value 2.63. This means that if the problem is optimized with regard to L_1-DEVIATION, and that we then evaluate MINMAX on that solution, it will be on average 2.63% higher than if the problem was optimized with regard to MINMAX. In other words, optimizing a problem with regard to L_1-DEVIATION is a decent approximation of MINMAX, as that solution will be on average only 2.63% worse than optimizing directly with MINMAX. The average of a row represents how well the balance criterion approximates the others in general, while the average of a column represents how well the balance criterion is approximated by the others in general.

Table 2. Comparison of the balance criteria for the BACP (reproduced from [16]).

	MINMAX	L_1-DEV.	L_2-DEV.	L_∞-DEV.	Average
MINMAX	0.00	10.62	16.53	0.06	9.07
L_1-DEVIATION	2.63	0.00	6.27	0.12	3.00
L_2-DEVIATION	0.28	0.00	0.00	0.00	0.09
L_∞-DEVIATION	10.37	18.07	23.66	0.00	17.36
Average	4.43	9.56	15.48	0.06	

Monette et al. observed that the optimal solutions of L_2-DEVIATION were often also optimal for the other methods, and thus that this method was generally a good approximation of the others. For this reason, the authors conclude that L_2-DEVIATION is the superior balancing method for the BACP. All methods exhibited similar performance except for L_2-DEVIATION whose running time was sevenfold that of the others.[3] As such, they suggest using L_1-DEVIATION as a compromise between efficiency and a good approximation of alternative methods. Further publications by various authors on this problem and its variants also use L_2-DEVIATION (see for example [10, 17, 18]).

4 Nurse-Patient Assignment Problem

The *Nurse-Patient Assignment Problem* (NPAP) seeks to assign patients to nurses within different zones in a hospital. The patients have various acuities, and should be assigned so as to best balance the workload among the nurses. The workload of a nurse is defined by the sum of his patients’ acuities. Let

[3] The filtering algorithms used for the balancing constraints have a linear temporal complexity, except for L_2-DEVIATION whose complexity is quadratic.

- $\mathcal{N} = \{1, \ldots, n\}$ be the index set of nurses,
- $\mathcal{P} = \{1, \ldots, m\}$ be the index set of patients,
- a_i denote the acuity of patient i with $a = \sum_{i \in \mathcal{P}} a_i$ representing the combined acuities of all the patients,
- $p_{\min}$ and $p_{\max}$ denote the minimum and maximum number of patients which can be assigned to a nurse.

The patients are located in different zones, and as such the NPAP is twofold: Nurses must first be assigned to zones, and then patients to nurses. The objective is to find a staffing of nurses to zones combined with a nurse-patient assignment maximizing the balance of the nurses' workloads. Let w_j be the workload of nurse j. The objectives can be formalized similarly as for the BACP

$$\max_{j \in \mathcal{N}} w_j \tag{\textsc{minmax}}$$

$$\sum_{j \in \mathcal{N}} |w_j - a/m| \tag{L_1-\textsc{deviation}}$$

$$\sum_{j \in \mathcal{N}} (w_j - a/m)^2 \tag{L_2-\textsc{deviation}}$$

$$\max_{j \in \mathcal{N}} |w_j - a/m| \, . \tag{L_∞-\textsc{deviation}}$$

The NPAP was introduced by Mullinax and Lawley [19], whose IP model expressed the measure of imbalance for a zone as the difference between its nurses' lightest and heaviest workloads. The objective was then to minimize the sum of imbalances for all the zones. Schaus et al. [20] have shown that while the previous model may do a good job in balancing the workloads within each zone, its objective function is deficient as this does not necessarily translate into a good balance of workloads between the different zones. The authors constructed CP models to solve this problem, and considered the L_1- and L_2-DEVIATION methods to minimize either the absolute or squared deviations of the workloads. They conclude that L_2-DEVIATION is more appropriate for the NPAP due to its increased sensitivity to outliers.

Table 3. Comparison of the balance criteria for the NPAP.

	MINMAX	L_1-DEV.	L_2-DEV.	L_∞-DEV.	Average
MINMAX	0.00	0.61	7.45	22.81	7.72
L_1-DEVIATION	0.19	0.00	3.79	18.07	5.51
L_2-DEVIATION	0.42	0.87	0.00	1.30	0.65
L_∞-DEVIATION	0.35	0.94	0.29	0.00	0.40
Average	0.24	0.60	2.88	10.55	

We have conducted similar experiments on the NPAP as Monette et al. did for the BACP [16] by adapting the CP model of [21] with the four objectives; Our results are reported in Table 3. The two problems share some similarities but are nevertheless unique in their own right. The BACP imposes assignment restrictions in the form of course prerequisites, while in the NPAP these restrictions are embedded in the staffing problem. Both problems have similar assignment ratios (five courses per period and six patients per nurse, on average), but the range of patient acuities in the NPAP is much wider than the range of course credits in the BACP.

L_∞-DEVIATION is the best approximator for the NPAP and the worst for the BACP, indicating that it is sensitive to the problem type. In contrast, L_2-DEVIATION is a very good approximator for both problems, suggesting robustness against various types of problems. As a general rule, MINMAX itself is not a very good approximator, but it can be well-approximated by the other methods. The opposite is true for L_2-DEVIATION, which is usually a good approximator for other balancing methods but which does not tend to be approximated very well most of the time. Performance-wise, for this CP model L_1- and L_2-DEVIATION are a few times slower than MINMAX and L_∞-DEVIATION.[4]

5 Practical Considerations

Some general guidelines concerning the choice of a balancing method can be derived from the lessons learned in the case studies. The main points to consider in the choice of a balancing method are its characteristics and performance.

Some of the many characteristics which can describe balance for the four methods studied in this paper have been outlined in Section 2, namely, dispersion, smoothness, and number of outliers. It should be kept in mind that most characteristics are neutral—they are not inherently desirable nor detrimental. For instance, the distribution of well-balanced workloads would form a narrow bell curve around the mean, since we are interested in having each worker share a similar workload. However, a distribution with a low mode would be more appropriate to balance occurrences of values representing the use of resources [22], if we are interested in mitigating resource hog.

The two previous sections conclude that most methods studied are of similar performance, except for L_2-DEVIATION which is usually more computationally expensive due to its quadraticity. Only knowledge of the time and resources available to solve a specific problem can provide guidance in choosing a method based on its efficiency. Note that a perfect balancing method constrained by time and resources such that it does not reach optimality may achieve worse solutions than an inferior method reaching optimality using these same limited resources. In a practical context, the value of a result does not solely depend on the theoretical quality of its solution.

In the two cases studied, all concluded that L_2-DEVIATION was the balancing method of choice for those problems due to its sensitivity to outliers. However, this specific characteristic is not necessarily desirable at all times. For instance,

[4] The `dispersion` constraint [10] was used for L_1- and L_2-DEVIATION, whose quadratic temporal complexity ensures domain consistency.

consider a factory with an equal number of workers and machines. Workers have various proficiencies on the machines, and a commodity is produced when all machines have been operated. Intuition may dictate that balancing the proficiencies of worker-machine pairs will ensure an efficient production. However, Fig. 1 shows that avoiding outliers may be counterproductive in some situations. The inverse of sensitivity to outliers, robustness against outliers, is itself a desirable characteristic at times as shown in the previous example.

$$\begin{array}{cccc} & m_1 & m_2 & m_3 \\ w_1 & 10 & 8 & 5 \\ w_2 & 1 & 10 & 9 \\ w_3 & 7 & 3 & 10 \end{array}$$

Fig. 1. Time required for each worker w to operate machine m. Perfect balance is achieved when all workers operate the machines on which they are the least proficient (red). In contrast, optimal throughtput is attained when the workloads are most unbalanced (blue).

As shown by Monette et al. [16] and ourselves, the fact remains that L_2-DEVIATION is a good approximation for many other methods, and could be chosen in case of doubt (performance requirements permitting). The results they presented, and which were confirmed by us in Section 4, show that no method strictly dominates the others. Finally, there is also the possibility of hybridization between multiple methods, such as bounding the minimum and maximum values, and using L_1-DEVIATION in-between these bounds, for example. With enough domain knowledge, such hybrids could present characteristics specifically tailored to a particular problem.

6 Conclusion

Oftentimes balance is attached to a problem as a side constraint or as a secondary objective without much thought. This paper shows that balancing a problem is not as straightforward as it seems, and highlights a few properties for some types of balancing methods. Two problematic modeling choices are shown and studied, after which general guidelines are proposed to prevent modelers from succumbing to the pitfalls of balancing method selection.

Acknowledgements

Financial support for this research was provided by NSERC Discovery Grant 218028/2017 and CERC, École Polytechnique de Montréal.

References

1. Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., Huq, A.: Algorithmic decision making and the cost of fairness. CoRR **abs/1701.08230** (2017)
2. Whiteacre, K.W.: Testing the level of service inventory–revised (LSI-R) for racial/ethnic bias. Criminal Justice Policy Review **17**(3) (2006) 330–342
3. Brams, S.J., Jones, M.A., Klamler, C.: Better ways to cut a cake. Notices of the American Mathematical Society **53**(11) (December 2006)
4. Pattanaik, P.K. In: Social Welfare Function. Palgrave Macmillan UK, London (2017) 1–7
5. Sen, A.: Rawls versus Bentham: An axiomatic examination of the pure distribution problem. Theory and Decision **4**(3) (February 1974) 301–309
6. Nash, J.: Two-person cooperative games. Econometrica **21**(1) (1953) 128–140
7. Marsh, M.T., Schilling, D.A.: Equity measurement in facility location analysis: A review and framework. European Journal of Operational Research **74**(1) (1994) 1–17
8. Schaus, P., Deville, Y., Dupont, P., Régin, J.C.: The deviation constraint. Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems (2007) 260–274
9. Pesant, G., Régin, J.C. In: SPREAD: A Balancing Constraint Based on Statistics. Springer, Berlin, Heidelberg (2005) 460–474
10. Pesant, G.: Achieving domain consistency and counting solutions for dispersion constraints. INFORMS Journal on Computing **27**(4) (2015) 690–703
11. Gaudioso, M., Legato, P.: Linear programming models for load balancing. Computers & Operations Research **18**(1) (1991) 59–64
12. Olivier, P., Lodi, A., Pesant, G.: The quadratic multiknapsack problem with conflicts and balance constraints. INFORMS Journal on Computing (to appear)
13. Castro, C., Manzano, S.: Variable and value ordering when solving balanced academic curriculum problems. Proceedings of 6th Workshop of the ERCIM WG on Constraints (Prague, June 2001) (November 2001)
14. Hnich, B., Kiziltan, Z., Walsh, T.: Modelling a balanced academic curriculum problem. In Jussien, N., Laburthe, F., eds.: Proceedings of the Fourth International Workshop on Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimisation Problems (CP-AI-OR'02), Le Croisic, France (2002) 121–131
15. Hnich, B., Kiziltan, Z., Miguel, I., Walsh, T.: Hybrid modelling for robust solving. Annals of Operations Research **130**(1) (August 2004) 19–39
16. Monette, J.N., Schaus, P., Zampelli, S., Deville, Y., Dupont, P.: A CP approach to the balanced academic curriculum problem. Symcon'07, The Seventh International Workshop on Symmetry and Constraint Satisfaction Problems (July 2007)
17. Chiarandini, M., Di Gaspero, L., Gualandi, S., Schaerf, A.: The balanced academic curriculum problem revisited. Journal of Heuristics **18**(1) (February 2012) 119–148
18. Ceschia, S., Di Gaspero, L., Schaerf, A.: The generalized balanced academic curriculum problem with heterogeneous classes. Annals of Operations Research **218**(1) (July 2014) 147–163
19. Mullinax, C., Lawley, M.: Assigning patients to nurses in neonatal intensive care. Journal of the Operational Research Society **53**(1) (January 2002) 25–35
20. Schaus, P., Van Hentenryck, P., Régin, J.C.: Scalable load balancing in nurse to patient assignment problems. In van Hoeve, W.J., Hooker, J.N., eds.: Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems, Berlin, Heidelberg, Springer (2009) 248–262

21. Pesant, G. In: Balancing Nursing Workload by Constraint Programming. Springer International Publishing, Cham (2016) 294–302
22. Bessiere, C., Hebrard, E., Katsirelos, G., Kiziltan, Z., Picard-Cantin, É., Quimper, C.G., Walsh, T.: The balance constraint family. In O'Sullivan, B., ed.: Principles and Practice of Constraint Programming, Cham, Springer International Publishing (2014) 174–189