Internet Presence Monitoring: Application of Heuristic Genetic Algorithm to Increase Efficiency in Work Allocation

Tricy B. Aquino, Juan Florencio C. Ferrer and Rex Aurelius C. Robielos
School of Industrial Engineering and Engineering Management
Mapua University
Intramuros, Manila, Philippines
tricyaquino2009@yahoo.com, johnflorenceferrer@gmail.com, racrobielos@mapua.edu.ph

Abstract

An effective supervisory function must allocate workload efficiently and optimize limited work hours with respect to a variety of operational requirements. These requirements often vary and entails a fast and dynamic model to be considered effective. Work allocation has typically been accomplished either through manual planning with considerations or constraints of very limited complexity, or by deterministic methods; requiring inordinate amounts of work hours which are typically incapable of adapting to changes in the environment. The methods and models presented here utilize forecasting techniques, analytic hierarchy process, and genetic algorithm (as search heuristic) to find a good sufficient allocation solution that can be performed quickly in rapidly changing process environment. Improvements introduced in this paper is projected to save 84 hours' worth of work per month.

Keywords
analytic hierarchy process, heuristics, genetic algorithm, posterior probability

1. Introduction

The finance sector is one of the most heavily regulated industries for the past decades. The instability in the financial system can have significant impact to the economy (McLaughlin, et al., 2017). Several sales malpractice and scandals prompted a number of legislative actions to mitigate such activities and protect consumers from unfair practices (Federal Trade Commission). One of the organizations leading this fight is the Financial Industry Regulatory Authority (FINRA), an independent, nongovernmental organization that writes and enforces the rules governing registered brokers and broker-dealer firms in the United States. Once such rule is the FINRA Rule 3110 that requires firms to establish, maintain and enforce a system to supervise their activities and the activities of their associated persons that is reasonably designed to achieve compliance with federal securities laws and regulations, as well as FINRA rules (FINRA Rules and Guidance).

Internet Presence Monitoring, among other communications and trading activity surveillance, is a supervisory measure adopted by some financial institutions to specifically determine if published content and other information posted on the internet pertaining to the practice or registered associate conform to company policies and guidelines. This was designed to prevent violations that may result to regulatory fines and penalties. Enforcement of these rules and regulations usually require a dedicated surveillance and control team because it takes time and effort to adapt business practices that follow these new rules and regulations correctly.

A key challenge for these surveillance and control teams is a combination of work prioritization and allocation given a multitude of constraints surrounding the process. The decision-making gets even more complex and dynamic as different parameters change over time. These conditions require planners to account for:

1) process-level considerations – review deadlines, changing conditions from the field, dynamic nature of internet content, strict allocation rules; and
2) analyst-level considerations – work capacity, productivity level, review quality, familiarity with a region or practice, etc.

Most considerations enumerated are often overlooked due to limited work hours available to sort through significant amount of data and manual discovery of the desired combination of allocation. Many organizations have similar difficulty in dealing with this challenge, and due to budget constraints and limited resources, have failed to achieve efficient means to perform the task. The sheer complexity of this problem also poses challenges in the application of conventional linear optimization techniques. Serious research on task allocation and scheduling problems began in the 1960's and has become a popular research topic in the past few decades (Chu and Beasley, 1995). While these types of problems and their solution methods are useful, many of them lack the complexity necessary for modeling any sort of real-world allocation problem that might occur in business operations. These can be classified into two categories: exact algorithms which attempt to solve the SPP to optimality, and heuristic algorithms which try to find "good" solutions quickly (Korte & Vygen, 2006). A similar approach using Heuristic allocation algorithms were conducted in distributed real-time and embedded applications such as critical areas in air defense, robotic control, and satellite constellations (Juedes et al., 2004).

In this study, the application of Heuristic genetic algorithm was explored to improve the current Internet Search Monitoring process as seen in Figure 1, with specific focus on Step 2. This is considered a major bottleneck and has most opportunity for improvement.

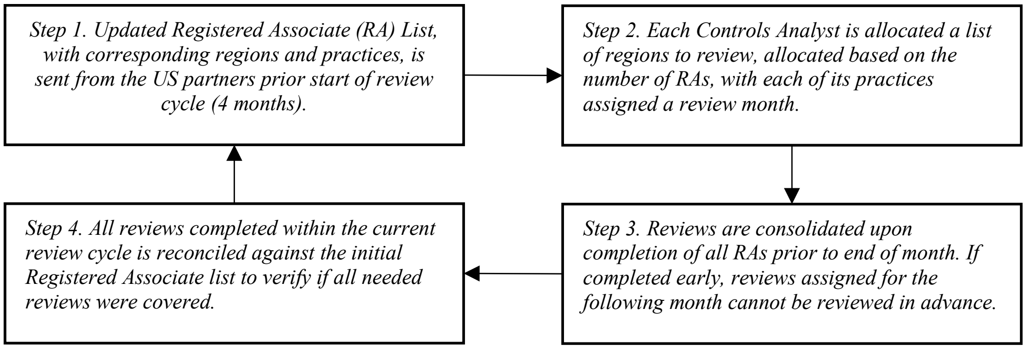
Step 1. Updated Registered Associate (RA) List, with corresponding regions and practices, is sent from the US partners prior start of review cycle (4 months).
Step 2. Each Controls Analyst is allocated a list of regions to review, allocated based on the number of RAs, with each of its practices assigned a review month.
Step 3. Reviews are consolidated upon completion of all RAs prior to end of month. If completed early, reviews assigned for the following month cannot be reviewed in advance.
Step 4. All reviews completed within the current review cycle is reconciled against the initial Registered Associate list to verify if all needed reviews were covered.

Figure 1. High-Level Process Map

2. Methodology

The dataset used in the study came from a US financial institution and covers 25 supervisory regions, 504 local practices, and 1,768 registered associates, with a total of 8,605 URLs – defining the granularity of the data. The allocation model will be for 5 review analysts. Figure 2 summarizes the methods used in the following sections.

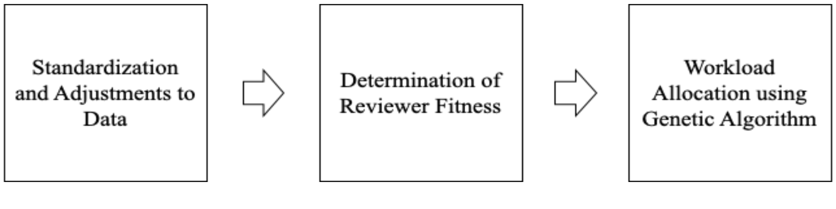
Standardization and Adjustments to Data
Determination of Reviewer Fitness
Workload Allocation using Genetic Algorithm

Figure 2. Summary of Methodology

2.1. Standardization and Adjustments to Data

To determine standard unit of work (Fin, et al., 2017), the URLs were divided into 3 groups with corresponding types affixed at the end of each group to arrive at 10 distinct classifications. Time-stamped review data for the last 8 months were used to derive the Average Review Per Classification (ARPC). The resulting ARPC was used as baseline unit of work.

URL status changes in between review cycles were modeled as a posterior probability and were computed using Bayes Theorem (Joyce, 2003). The Bayesian Formula (1) has hypothesis variable *H*, i.e. the probability of URL status changing in the next cycle, given event variable *E*, i.e. status changed from the previous cycle. The resulting probability H given E was multiplied to ARPC to arrive at a Standard Unit of Work (SUW) as seen in formula (2).

Bayes Formula: $$P(H|E) = \frac{P(E \mid H) * P(H)}{P(E)} \qquad (1)$$

Standard Unit of Work: $$P(H|E) * ARPC \qquad (2)$$

2.2. Determination of Reviewer Fitness

Determination of URL fitness to an analyst was modeled using the Analytic Hierarchy Process (AHP) as seen in Figure 3. Two main criteria used were used. First, degree of familiarity as measured in count of times the URL was previously reviewed and length of time the URL was last reviewed. Second, historical performance as measured in productivity and accuracy rate at classification level.

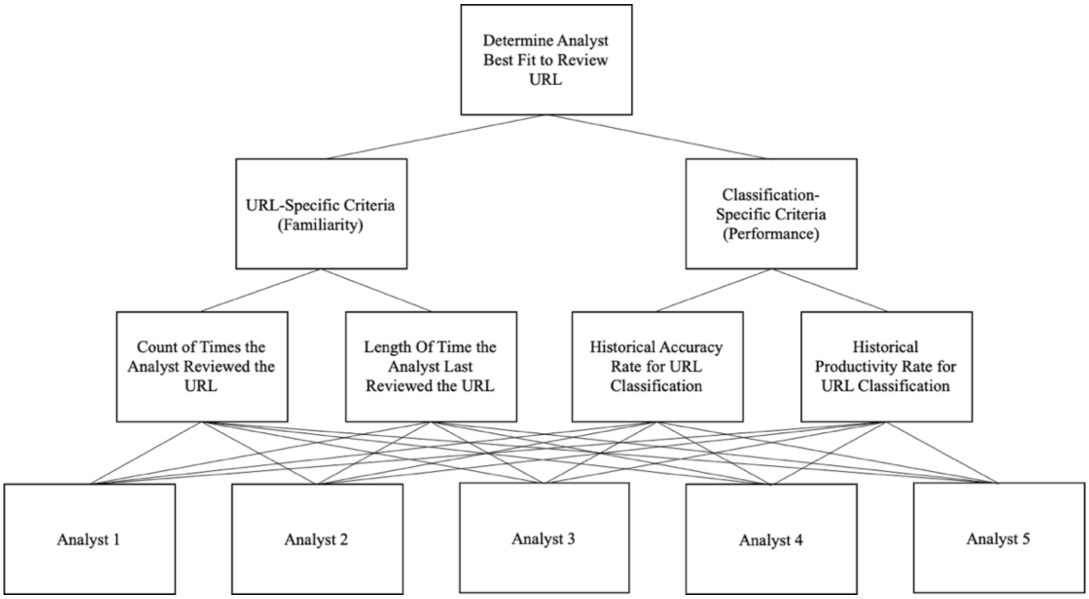
Determine Analyst Best Fit to Review URL
URL-Specific Criteria (Familiarity)
Classification-Specific Criteria (Performance)
Count of Times the Analyst Reviewed the URL
Length Of Time the Analyst Last Reviewed the URL
Historical Accuracy Rate for URL Classification
Historical Productivity Rate for URL Classification
Analyst 1
Analyst 2
Analyst 3
Analyst 4
Analyst 5

Figure 3. AHP Model

A pairwise comparison matrix was generated through the calculation of Eigenvalue (2) and Eigenvector (3). Consistency ratio was also utilized to assess the consistency of the comparison matrix (4) (Saaty, 1990).

Eigenvalue: $$\lambda_{\max} = \sum_{j=1}^{n} a_{ij} \frac{W_j}{W_i} \qquad (2)$$

Eigenvector: $$(A\text{-}\lambda \max I)\, X = 0 \qquad (3)$$

Consistency Ratio: CR = CI / RI (4)

To collect data needed for the pairwise comparison, a questionnaire (see Appendix A) was distributed across 10 respondents: 4 managers, 3 supervisors, and 3 team leads. Survey results were aggregated as a geometric average to represent a collective decision. The scale ranges from one to nine where one signifies equal importance.

The resulting criteria weights from the AHP model were multiplied to normalized values arriving at a common scale without distorting differences in the ranges of their varying unit of measures. A corresponding Normalized Performance Value (NPV) was computed for both non-beneficial (5) and beneficial (6) criteria. The sum of all NPV for each criterion determined the analyst's Fitness Ranking (FR), 1 being the highest and 5 being the lowest.

Normalized Performance Value: $Non - Beneficial \ = \left[\frac{Min(x_{ij})}{x_{ij}}\right](W_j)$ (5)

$$Beneficial \ = \left[\frac{x_{ij}}{Max(x_{ij})}\right](W_j) \qquad (6)$$

2.3. Workload Allocation using Genetic Algorithm

Baseline capacity was first calculated to determine the maximum amount of work that can be completed for each review period (1 month) and the entire review cycle (4 months). Assumptions include 7.5 working hours per day for 21 working days, converted to minutes at 9,450 total capacities for each analyst per month, a total of 37,800 minutes per cycle. Baseline deductions were forecasted using a 6-month moving average to arrive at net capacity per review cycle t_n, and monthly net capacity t_{nm}, for each analyst *n*, for each month *m*.

After determining net capacity, a two-step approach was implemented in the allocation process. First step was to designate regions based from the region-level aggregation of standard unit of work SUW_r obtained from section 2.1 to determine desired overall target allocation r_n (7). In addition, aggregated fitness ranking f_n (8) was used in conjunction with r_n (compared against t_n) to help determine allocation at analyst level. This will form first component of the objective function (9).

Target allocation per reviewer: $r_n = \sum_{i=1}^{r} SUW_r$ (7)

Total fitness ranking per reviewer: $f_n = \sum_{i=1}^{r} FR_r$ (8)

Objective function component 1: $\sum_{i=1}^{n} |t_n - r_n| + \sum_{i=1}^{m} f_n$ (9)

The second step of the allocation is to determine monthly allocation per analyst. The resulting r_n for each analyst will be divided at practice level to determine r_{nm} for reviewer *n* for each month *m* (10) compared against t_{nm}. This will form second component of the objective function (11).

Target allocation per reviewer: $r_{nm} = \sum_{i=1}^{u} SUW_u$ (10)

Objective function component 2: $\sum_{i=1}^{m} |t_{nm} - r_{nm}|$ (11)

Step 1 and 2 utilized Genetic Algorithm (GA) find an optimal combination of regions across analysts and practices across months in accordance with the following constraints:

1. Each region can only be assigned to 1 reviewer.
2. Each reviewer must be assigned to at least 1 or more regions.
3. All regions must be reviewed within a specified period (e.g. 4 months).
4. Each practice must be reviewed in a month and can't cross other months.

GA simulated the iteration process by taking an initial sample of allocation and applying genetic operators in each iteration. In optimization terms, each part of the allocation is encoded into a string or chromosome which represents a possible solution to the given problem [10]. The effectiveness of the overall allocation is evaluated with respect to the combined objective function (12).

Objective Function: $$\min = \sum_{i=1}^{n}|t_n - r_n| + \sum_{i=1}^{m} f_n + \sum_{i=1}^{m}|t_{nm} - r_{nm}| \quad (12)$$

Subject to:
$1 \leq r \leq 5$
$1 \leq m \leq 12$
$r_{mn} \geq 1$
$r_n \geq 1$

The Evolutionary engine solving method inside the Microsoft Excel Solver was utilized to run the model due to the non-smoothness of the problem. Binning techniques were applied since the said application is limited to handling only 200 variables.

3. Results

3.1. Standardization and Adjustments to Data

The resulting ARPC from time study for each of the 10 classifications are shown in Table 1. Classifications returning the highest share of workload were 3rd-Party – Listing, Social Media – LinkedIn, and Practice Website – Bio. Social Media – LinkedIn also returned the highest ARPC from the group. Classifications returning the lowest share were Social Media – Other, Practice Website - Other, and Social Media - Twitter. Social Media – Other also returned the lowest ARPC from the group. The average standard deviation across classifications is +/- 0.12. Observations outside 1.5x the Interquartile Range (IQR) were considered outliers and were removed from the dataset.

Table 1. URL Classification Summary

URL Classification	**Count or URLs**	**ARPC**	**Standard**	**% Share**
		(in min)	**Deviation**	
3rd-Party - Listing	4,371	2.24	0.19	28.69%
Social Media - LinkedIn	1,201	7.67	0.20	26.99%
Practice Website - Bio	972	6.12	0.18	17.43%
Practice Website - Home	543	6.74	0.15	10.72%
3rd-Party - Other	410	5.55	0.20	6.67%
3rd-Party - Publication	703	2.61	0.07	5.38%
Social Media - Facebook	277	3.43	0.08	2.78%
Social Media - Twitter	101	3.86	0.01	1.14%
Practice Website - Other	15	3.33	0.06	0.15%
Social Media - Other	12	1.32	0.11	0.05%

URL Status change probabilities are reflected in Table 2. There were 140 computed posterior probabilities across possible status changes for all URL classifications, with the majority 60% returning 0% probability. Out of the other 40%, only 17 or 12.14% had probabilities greater than 1%. Practice Websites for Home, Bio, and Other changing from *Under Maintenance to Active* returned the highest posterior probability at 99.93%, and *Active to Under Maintenance* at 22.86% from the same group. Status from *Active to Owner Terminated* combined probability across all classifications was at 18.29% followed by *Active to Dated* at 8.19%. Combined probability under 3rd-Party – Publication was computed at 9.83% followed by Social Media – Other at 7.43%.

Table 2. URL Status Change Posterior Probability Summary

URL Status Change	3rd-Party - Listing	3rd-Party - Publication	3rd-Party - Other	Practice Website - Home	Practice Website - Bio	Practice Website - Other	Social Media - LinkedIn	Social Media - Twitter	Social Media - Facebook	Social Media - Other
From Active to Blocked	**0.5316%**	**0.2854%**	**0.8252%**	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%
From Active to Under Maintenance	0.0000%	0.0000%	**0.9994%**	**22.8600%**	**22.8600%**	**22.8600%**	0.0000%	0.0000%	0.0000%	0.0000%
From Active to Dated	0.0000%	**7.1200%**	**0.1300%**	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	**0.9400%**
From Active to Irrelevant	**0.5500%**	**0.1600%**	**0.5000%**	0.0000%	0.0000%	0.0000%	**0.6100%**	**0.3800%**	**0.9900%**	**3.3200%**
From Active to Login Required	**0.1500%**	**0.6200%**	**0.3600%**	0.0000%	0.0000%	0.0000%	**0.3100%**	**0.5100%**	**2.4100%**	**2.2800%**
From Active to Not Found	**0.8100%**	**0.8900%**	**0.5500%**	0.0000%	**0.8600%**	**0.9000%**	**0.1100%**	**1.8400%**	**0.1300%**	**0.2500%**
From Active to Owner Terminated	**0.1600%**	**0.4600%**	**0.1900%**	**3.0900%**	**4.5700%**	**3.0900%**	**3.4500%**	**1.1200%**	**1.5200%**	**0.6400%**
From Blocked to Active	**0.4875%**	**0.2800%**	**0.8249%**	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%
From Under Maintenance to Active	0.0000%	0.0000%	**0.8425%**	**99.9333%**	**99.9333%**	**99.9333%**	0.0000%	0.0000%	0.0000%	0.0000%
From Dated to Active	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%
From Irrelevant to Active	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%
From Login Required to Active	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%
From Not Found to Active	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%
From Owner Terminated to Active	**0.0248%**	**0.0101%**	**0.1900%**	**0.0101%**	**0.0101%**	**0.0000%**	**0.0101%**	0.0000%	0.0000%	0.0000%

Regional aggregation of SUW from ARPC and corresponding posterior probabilities resulted to a total of 33,615.61 SUW computed from 8,605 URLs. There was an average of 1,171.62 SUW across 25 regions with a standard deviation of +/- 282.70.

3.2. Determination of Reviewer Fitness

Data gathered from the questionnaires were calculated using the BPMSG AHP Priority Calculator. A total of 6 pairwise comparisons were made with a corresponding principal eigenvalue of 4.046. Consistency ratio was at 1.7% which is less than the standard score of 10%. Table 3 displays the resulting weights per criteria. The accuracy rate ranked the highest with a significant weight allocation of 65.9%, more than 3x the weight of allocated for productivity rate which was at 18%. Familiarity weights *Count of Times* and *Length of Time* only had a combined weight of 16.01%.

Table 3. AHP Weight per Criterion

Category	Alternatives	Priority	Rank	(+)	(-)
1	Count of Times the Analyst Reviewed the URL	9.8%	3	1.9%	1.9%
2	Length of Time the Analyst Last Reviewed the	6.3%	4	1.4%	1.4%
3	Historical Accuracy Rate for URL Classification	65.9%	1	11.9%	11.9%

4	Historical Productivity Rate for URL Classification	18.0%	2	1.3%	1.3%

Normalized values of each criterion were multiplied to corresponding weights from the AHP model to obtain fitness ranking for each URL.

3.3. Workload Allocation using Genetic Algorithm

Monthly projected net capacity was at of 33,616 minutes for all five analysts. This is only 17.79% when compared to total baseline capacity. Majority of the deduction was for time spent on other processes which accounts for 76.17% of baseline capacity. This closely matches the total SUW at 33,615 and resulting total variance of 25 units. An average of 3,302 aggregated fitness ranking was obtained in relation to the allocation.

Table 4. Summary of Analyst Allocation

Reviewer	t_n	r_n	$\lvert t_n - r_n \rvert$	f_n
Analyst 1	6,645	6,655	5	3,384
Analyst 2	7,576	7,570	6	2,062
Analyst 3	7,159	7,155	5	2,896
Analyst 4	7,050	7,054	4	4,182
Analyst 5	5,186	5,181	5	3,988

Monthly allocation at practice level followed the analyst allocation and resulted to an average monthly variance of 75 units with a total of 302-unit variance for the entire cycle as seen in Table 5. The distribution also resulted to an average of 1,676 monthly allocations for each analyst.

Table 5. Summary of Monthly Allocation and Resulting Variance

Reviewer	t_{Sep}	r_{Sep}	t_{Oct}	r_{Oct}	t_{Nov}	r_{Nov}	t_{Dec}	r_{Dec}
Analyst 1	1,539	1,514	1,636	1,629	1,874	1,858	1,596	1,606
Analyst 2	1,324	1,317	2,122	2,148	2,398	2,406	1,732	1,755
Analyst 3	1,830	1,812	1,986	1,990	1,935	1,912	1,408	1,400
Analyst 4	1,606	1,595	1,585	1,599	1,956	1,932	1,903	1,881
Analyst 5	1,439	1,442	1,345	1,356	1,339	1,318	1,063	1,042
$\sum_{i=1}^{m} \lvert t_{nm} - r_{nm} \rvert$	*Sep = 64*		*Oct = 62*		*Nov = 92*		*Dec = 84*	

Results from both analyst and monthly allocation were competed around 125 minutes. A summary of how the genetic algorithm performed can been in Figure 3. The figure reflects variance reduction in *y* axis and minutes elapsed in the *x* axis. Genetic algorithm had minimal gain in reducing overall variance upon reaching the 95th minute.

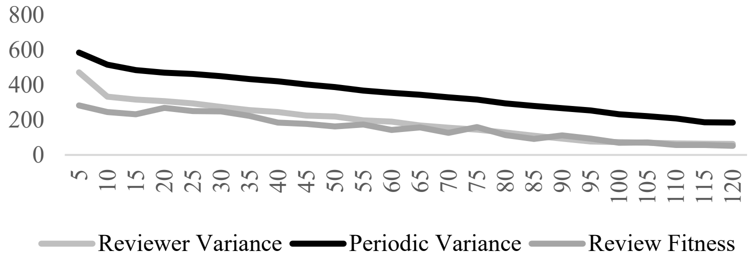

Figure 3. Genetic Algorithm Runtime Performance

4. Discussion

The current allocation process used prior to this study used a naive associate count as basis to determine unit of work. This method does not consider varying degrees of online presence, i.e. number of search hits and URLs related to each associate. In addition, each URL had varying amounts of complexity and review time. These were addressed by the standardization process that led to identifying 10 URL classifications with corresponding Average Review Per Classification (ARPC).

Apart from standardization, URL status changes in between review cycles were also considered and adjustments were made to ARPC in the form of probabilities. The resulting adjustments set the standard unit of work. Said status changes were initially thought to significantly impact the allocation since a status change from active status to non-active status would render average review time from initial ARPC to zero and vice versa. However, after posterior probabilities were computed across all possible status changes and URL classification, only 12% had probabilities greater than 1% that are worth considering when making adjustments. The amount of data and effort required to put together these numbers is significant, so is suggested to further analysis is needed to determine if the probability adjustment provides enough value to the overall allocation process.

To compliment standard unit of work, AHP was used to model "fitness" of an analyst to review a particular URL to capture possible efficiency and accuracy gains. The final weights reflected survey respondents' estimates and were used to determine fitness ranking. The decision to aggregate ranking was used due to the convenience in computing and lessen complexity in the model. There were considerations to scale the ranking based from the raw scores, however it was decided not explore this further to simplify the process. Furthermore, since this was added as a component in the objective function, it proportionately added to the GA model runtime. The effectiveness of fitness ranking will be measured in the following months and will be basis if it can remain in the model or if adjustments are needed.

The resulting allocation generated by the Genetic algorithm (GA) gained significant improvement as measured in the variance between target and actual allocations that took into account all preceding variables, i.e. SUW and FR. First, allocation from region level had a 220% improvement from 55 down to 25-unit variance compared to previous process. Second, monthly allocation from practice level even had a more significant improvement at 443% from 1,337 down to 302-unit variance. Finally, a 9-hour direct efficiency gain per month was achieved since it would now only require 2 hours to complete the process compared to 11 hours previously. In addition, an estimated indirect efficiency gains of 75 hours per month due to the efficiencies resulting from low variance between allocations. Improvements coming from probability adjustment and fitness ranking has yet to be measured but will be tracked to further enhance model.

5. Conclusion

In conclusion, the methodology introduced in the study led to significant improvements in the allocation process and is being considered as a framework to approach other processes. Several analytical tools and techniques were used to solve a number of pain points in the allocation process. The following summarizes these and corresponding pain points it was able to solve:

- Standardization dealt with inaccurate measurement of work;
- Posterior probability dealt with predicting possible changes that could impact work;
- Analytic Hierarchy Process dealt with leveraging existing capacity to increase efficiency and accuracy;
- Net Capacity dealt with estimating how much can be done; and
- Genetic Algorithm (a search heuristic) dealt with automating the process of finding the right combination of work and/or allocation variables.

Other analytical tools may be applied in lieu of any of the tools or techniques listed above to accomplish the same goal.

References

Bernhard, K. and Vygen, J. (2006), Combinatorial Optimization: Theory and Algorithms
Chu, P.C. and Beasley, J.E. (1995), A Genetic Algorithm for the Set Partitioning Problem
Federal Trade Commission, Consumer Finance https://www.ftc.gov/news-events/media-resources/consumer-finance
Fin, J.C. et al. (2017) Improvement Based on Standardized Work: An Implementation Case Study
FINRA Rules and Guidance, 3000. Supervision and Responsibilities Relating to Associated Persons, 3100. Supervisory Responsibilities
James, J. (2003), Bayes' Theorem
Juedes, D. et al. (2004), Heuristic resource allocation algorithms for maximizing allowable workload in dynamic, distributed real-time systems
McLaughlin, P. et al. (2017) Regulatory Accumulation in the Financial Sector, Mercatus Center George Mason University
Mumford, C. L. (2008), An Order Based Memetic Evolutionary Algorithm for Set Partitioning Problems
Saaty, T.L. (1990) How to Make a Decision the Analytic Hierarchy Process

Biographies

Rex Aurelius C. Robielos is the Dean of the School of Industrial Engineering and Engineering Management at Mapua University. Before joining Mapua, he was Section Manager of Operations Research Group, Analog Devices General Trias. He has a BS in Applied Mathematics from the University of the Philippines Los Baños, and a Diploma and MS in Industrial Engineering from the University of the Philippines Diliman. He is pursuing Ph.D in Industrial Management (candidate) at National Taiwan University of Science and Technology in Taiwan. He is the current Secretary of Human Factors and Ergonomics Society of the Philippines and Director of the Philippine Institute of Industrial Engineers and Operations Research Society of the Philippines.