Learning 'Good Quality' Resource Allocations from Historical Data

Renuka Sindhgatta[1,2(✉)], Aditya Ghose[2], and Gaargi Banerjee Dasgupta[1]

[1] IBM Research-India, Bangalore, India
{renuka.sr,gdasgupt}@in.ibm.com
[2] University of Wollongong, Wollongong, NSW, Australia
aditya.ghose@uow.edu.au

Abstract. Effective and efficient delivery of services requires tasks to be allocated to appropriate and available set of resources. Much of the research in task allocation, model a system of tasks and resources and determine which tasks should be executed by which resources. These techniques when applied to service systems with human resources, model parameters that can be explicitly identified, such as worker efficiency, worker capability based on skills and expertise, authority derived from organizational positions and so on. However, in real-life workers have complex behaviors with varying efficiencies that are either unknown or are increasingly complex to model. Hence, resource allocation models that equate human performance to device or machine performance could provide inaccurate results. In this paper we use data from process execution logs to identify resource allocations that have resulted in an expected service quality, to guide future resource allocations. We evaluate data for a service system with 40 human workers for a period of 8 months. We build a learning model using Support Vector Machine (SVM), that predicts the quality of service for specific allocation of tasks to workers. The SVM based classifier is able to predict service quality with 80 % accuracy. Further, a latent discriminant classifier, uses the number of tasks pending in a worker's queue as a key predictor, to predict the likelihood of allocating a new incoming request to the worker. A simulation model that incorporates the dispatching policy based on worker's pending tasks shows an improved service quality and utilization of service workers.

Keywords: Resource allocation · Classification · Simulation model

1 Introduction

Service System as defined by Sphorer [11] is an important unit of analysis in support of understanding operations of an organization. A Service System (SS) comprises of resources (that include people, organizations, shared information, technology) and their interactions that are driven by a process to create a suitable outcome to the customer. In SS, the participants have to collaborate together to provide right outcome(s) effectively to the customer. In [15], the authors

F. Toumani et al. (Eds.): ICSOC 2014, LNCS 8954, pp. 84–95, 2015.
DOI: 10.1007/978-3-319-22885-3_8

argue the need for optimal allocation of resources in SS. Resources in SS are predominantly human resources and referred to as Service Workers (SW). Unlike machines or equipment, behavior and efficiency of human resources varies. In [1], the authors identify common pitfalls associated with building simulation models that includes incorrect modeling of human resources. Incorrect representation or modeling of human resources and simulation of business processes causes models to provide misleading outcome measures. Outcome measures refer to the average utilization of resources, average throughput or number of requests completed periodically, service quality that includes completing work within a specified target time.

There are several complexities in modeling human resources. Resources in a team have different efficiencies although they may have similar skills and competencies. Efficiency of a single SW is not constant and varies with the work allocated to the SW [13,20]. In this paper, we use historical task allocation data, stored in process aware information systems (PAIS) as event logs or process execution logs. Allocations that yield 'good' and 'bad' quality are identified. We use this data to build a learning based model that is able to predict allocations of tasks to resources resulting in 'good' quality. Our objective is to use historical allocation that inherently considers SW behavior to guide future allocations. The categorization of an outcome as 'good' or 'bad' quality, is generic and can be applied in the context of any SS and outcome measure.

The outline of this paper is as follows: We motivate relevance of our work with observations from a real-life SS and state our contributions in Sect. 2. Next, we present key concepts and discuss our data collection i.e. the data we used for our analysis, in Sect. 3. Section 4, presents our learning models to predict outcome of allocating tasks to resources and predicting the resources to allocate new tasks. In Sect. 5, we build a simulation model to evaluate the improvements possible when allocation of tasks to resource considers their individual performance or behavior. We discuss validity of our results in Sect. 6 and related work in Sect. 7. Section 8 concludes the paper.

2 Motivation

We analyzed the data for a real-life service system (SS) that tracks critical IT system failures (incidents) of different customers. A team of 40 service workers (SW) belonging to the service provider organization, ensure that incidents or service requests from customers are immediately responded. Customers represent different organizations, the service provider supports. As shown in Fig. 1, the service requests from customers are placed in a queue. A human dispatcher monitors the queue and assigns requests to a suitable SW. A SW could be assigned multiple requests and the number of requests currently being handled by a worker represents worker queue. The key measure of service quality is the *time taken to respond to the customer*. The total time to respond - the time elapsed between the creation of the request and response from a service worker to the customer is a measure of service quality. As these are critical system

failures, the time to respond should be lower than a set time or the *service target time* agreed by the customer and the provider. If the time to respond exceeds the service target time, the request is said to have missed or breached the service quality. After responding to the customer, the SW identifies the problem, identifies the team(s) that should work towards solving the issue and disengages from the request. While there are several other performance indicators indicating service quality, in the context of the SS under consideration, time to respond within the service target time is used as a measure of service quality. In this SS, all requests require the same skill (e.g. operating system maintenance) and expertise level (e.g. high expertise).

We make the following observations on the data analyzed for the SS.

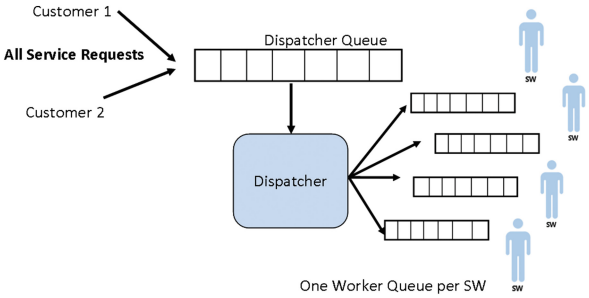

Fig. 1. Application maintenance process

Service Workers with Similar Measurable Skills or Capabilities Have Different Response Times. In the SS under study, all the SWs are of the same organizational role (e.g. Subject Matter Experts) and have same expertise and skill level (e.g. High). However, we observe different means of response time for the workers. Figure 2 shows the box plot with median, upper and lower quartile response times for different service workers (depicting the variance in their means). A one-way ANOVA test [8] for analysis of variance of response time means across different service workers yields a statistically significant difference ($p < 0.01$). Hence, we conclude that service workers with similar capabilities or skills have different efficiencies.

Queue of Pending Requests, Impacting Service Quality, Is Different for Each Service Worker. The worker queue length or the number of pending requests, of SW, impacts the time to respond to a new incoming service request and hence in meeting or breaching the service quality. Figure 2 shows the box plot of queue length of SW, measured for an incoming request that has met or breached service quality. The Work queue lengths have been shown for only 10 of the 40 service workers due to space constraints. A factorial ANOVA indicates a variance in the mean queue length across two factors - requests that meet or

miss target response time and service workers ($p < 0.05$). i.e. the mean work queue length varies for each service worker and is lower when arriving requests meet target response time.

These observations suggest that allocation of tasks to resources considering them to have similar efficiency and behavior will result in inaccurate or misleading results. In this paper, we will investigate if a learning based model can be used to guide allocation of tasks to resources. Through this work, we aim to provide the following technical contributions:

- build a learning based classification model to predict the service quality when tasks are allocated to specific resources.
- build prediction model to guide allocation of tasks to a resource.
- build a simulation model to evaluate the performance of a service system that allocates tasks to resources taking into consideration efficiencies of individual resources.

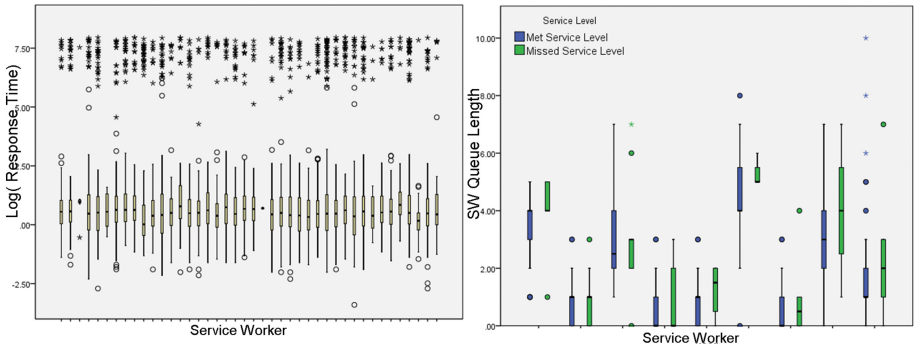

Fig. 2. Box plot indicating response time of workers and queue length at which requests meet of breach service levels

3 Background

In this section, we present concepts relevant to the service system and data collected for the system under consideration.

3.1 Service System Concepts

We define key concepts underpinning the service system below:

Service Request or Incident. Service requests (SR) or incident constitute inputs to the service system and are handled by service workers. Typically, an incident is characterized by priority. In the system we evaluate, all the request are of high priority.

Work Arrivals. The arrival pattern of service requests is captured for finite set of time intervals T (e.g. hours of a week). That is, the arrival rate distribution is estimated for each of the time intervals in T, where the arrival rate is assumed to follow a stationary Poisson arrival process within these time intervals (one hour time periods) [4,7].

Response Time. Response time refers to the time taken by a service worker to respond to a customer. This is the time interval between a customer creating request to the time a service worker responds to the customer of its receipt.

Service Time. Service time refers to the time a service worker spends on addressing the request. In the service system being studied, it is the time spent by SW in identifying suitable team(s) to handover the request. Hence, service time is the time interval, the request remains with the service worker.

Worker Queue Length. A service worker handles multiple service requests at a time, and the request remains in the worker queue till the SW hands it over to another person or team. The number of requests that remain with the worker, at any point in time, is the worker queue length.

Service Target Time and Service Level Agreement. Service levels are a measure of quality or outcome of service. Service Level Agreement (SLA), for each customer $\gamma_i = (\alpha_i, r_i), \alpha_i, r_i \subset \mathbb{R}$, is a map from each customer i to a pair of real numbers representing the service time target and the percentage of all the SRs that must be responded within this *service target time* in a month. For example, $\gamma_{Customer_1} =< 1, 95 >$, denotes that 95 % of all SRs from $Customer_1$ in a month be responded within 1 h.

3.2 Data Collection

Data from the service system is collected for a period of 8 months. The data for each request is obtained from a process aware ticketing system that contains the time a customer opened the SR, the worker allocated to the SR, the time taken to respond and the service time of the SR. Given the data, we collate following features for our prediction model measured at hourly time intervals (an example shown in Table 1):

- Hour of the day
- Number of requests arriving into the system
- Number of incoming requests assigned to each SW
- Work Queue Length of each SW (number of pending requests with SW)

Hence, for each hour, we have 82 features extracted for the model - 40 features represent the number of incoming requests assigned to SW, 40 features depict the work queue length of each SW. In addition, we have total number of requests arriving in the system (total number of incoming requests).

4 Experimental Analysis

In this section, we present the models that are built to validate our hypothesis on the suitability of using historical data to predict valid or suitable resource allocations. We use IBM SPSS Modeler 14.1 [10] to build our prediction models.

Table 1. Example set of features for learning model

Hour of day	Number of incoming requests	AgentA allocated	AgentA queue length	AgentB allo-cated	AgentB queue length	..	AgentC allo-cated	AgentC queue length
0	2	0	4	1	2	..	0	1
1	1	0	3	0	1	..	1	0

4.1 Predicting Service Quality Using Support Vector Machines (SVM)

The objective is to build a prediction model that is capable of identifying if allocation of requests to the chosen set service workers will result in meeting or missing the service quality. To train the prediction model, we use input features and define a boolean flag *Outcome*, valued 'GOOD' if all requests have met the target response time and valued 'BAD' if one or more request missed the target response time. Hence *Outcome* constitutes target feature of the prediction model, for the conditions of input volume of requests, worker queue lengths and worker assignment.

We use support vector machines (SVM) to classify and label the *Outcome* parameter. Logistic regression, Naive Bayes and SVM classifiers are very popular and widely used classification techniques. In [14], it is shown that prediction accuracy of classification techniques vary with the number of features defined in the model and the size of the training set. We carried out preliminary analysis using naive Bayes, logistic regression and SVM classifier. SVM was found to be more robust to the random samples of training and testing data sets and resulted in higher prediction accuracy.

The prediction accuracy of the model is measured by the percentage of unseen instances it correctly classifies. A good classifier must fit the training data well, in addition to accurately classifying the data it has never seen before (test data). We partition the data into training and test samples. The prediction accuracy of the training sample is 80.43 % and that of the testing sample is 78.9 %. Figure 3 shows the prediction accuracy and the confidence probability of the model for the test sample. The accuracy of prediction improves with increase in confidence probability, for the testing samples. It can be seen that, at higher confidence intervals (> 0.87), the accuracy of correct predictions is 90 %. The histogram shows the frequency distribution of confidence probability assigned by the model. A large number of predictions have higher confidence probability. Hence, historical samples can be used to learn and predict the quality outcome of requests assigned to service workers.

4.2 Predicting Allocation of Request to Service Worker

In the previous section, the objective was to categorize allocation of tasks to SWs as good or bad, considering their queue lengths and number of requests

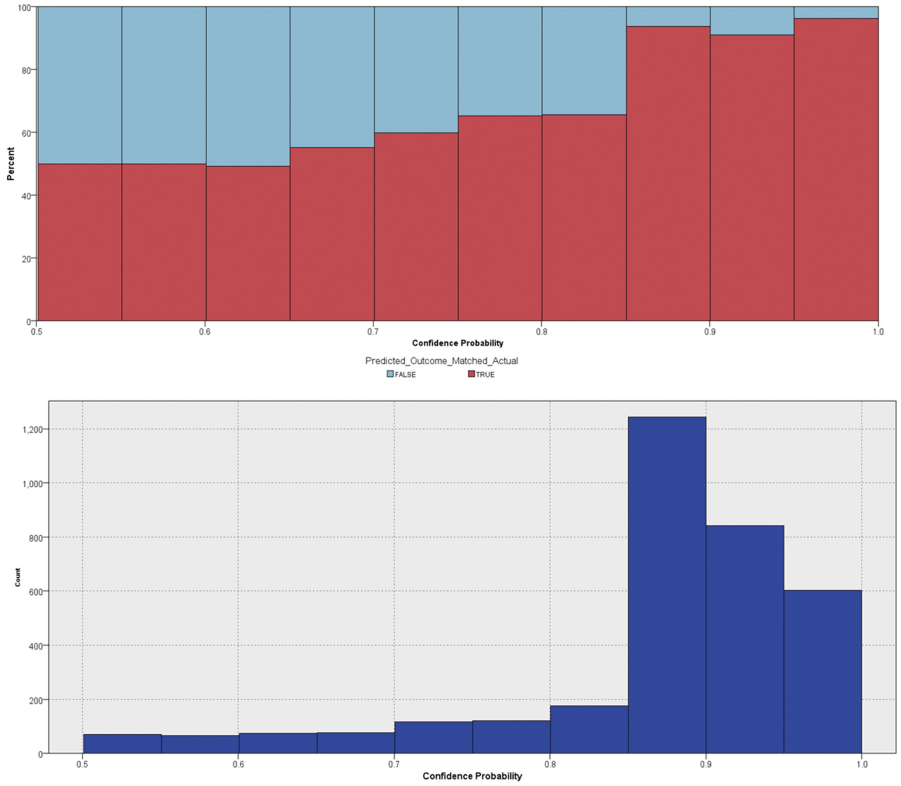

Fig. 3. Percentage of predictions that match actual outcome Vs. confidence probability and frequency distribution of confidence probability

arriving in the system. In a scenario, where a dispatcher allocates one task at a time to a SW, a model to assist the dispatcher in predicting if a request should be allocated to a SW or not, would be useful. Here, our training sample contains all allocations that have resulted in a good service quality. We build a classification model for each SW. The input to model is the number of requests and the queue length of all SW including the SW for which the model is built. A *SWAllocate* flag is the target feature which is set to 'TRUE' if a request can be allocated to the SW and 'FALSE' otherwise. In the data under consideration, for a large number of observations, a SW does not get a request allocated as the number of requests arriving in the system may be low and there are many SW. For most SW, *SWAllocate* is set to FALSE for 90 % of the observations. Hence, if we assign FALSE to all SWAllocate, it would still lead to 90 % prediction accuracy. Therefore, we evaluate a learning model that can predict TRUE allocations accurately. Logistic regression and SVM fail to make accurate predictions of SWAllocate. Linear Discriminant Analysis (LDA) based classifier [19], predicts resource allocations to a SW with 65 %–80 % accuracy. Figure 4 shows

the accuracy of predictions for the SWAllocate with 'TRUE' values for one SW. As shown, logistic regression fails to predict them with 0 % accuracy at 99 % confidence probability. We also realize that the training data needs to contain more observations where allocation of request is made to a SW. However, with the lack of large training data, LDA can be used to guide allocations to a SW when the confidence probability is high, as indicated in Figure Percentageregression. This prediction model is build for each SW.

The dominant predictors or coefficients of LDA for predicting the allocation to a SW are the number of requests arriving into the system and the work queue length of SW for whom the allocation model is built. We have seen that SW work queue length, that impacts the service target time of an incoming request, varies for each SW (Sect. 3). In the next section we build a simulation model to compare results of a model that incorporates the SW work queue length during allocation of request to SW.

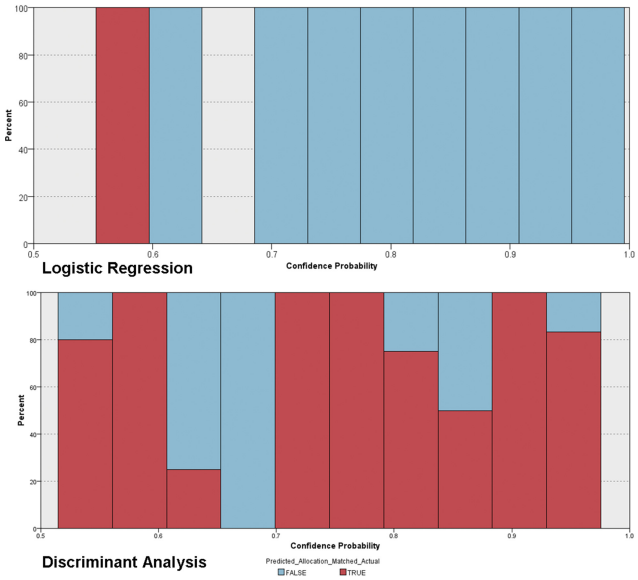

Fig. 4. Percentage of predictions that match actual allocation Vs. confidence probability for a single SW having resource allocation set to TRUE

5 Simulation Based Evaluation of SLA

In this section, we describe the simulation set up that mimics the service system being evaluated. The inputs to the model are the following:

– A finite set of time intervals for arriving work, denoted by T, containing one element for each hour of week. Hence, $|T| = 168$. Each time interval is one hour long. Work arrivals rates are defined for each time interval.

- Maximum Work Queue Length $Qmax_i$ for SW $i \in \{1, 2 \ldots n\}$: The work queue length is derived from the data for worker i (SW_i). It is the queue length of SW_i below which the worker meets target response time.
- Service Time: The mean service time for which a request remains with the SW till it is handed over.
- Response Time : The response time for a request depends on the $Qmax_i$ for SW_i. The response time is less than service target time when the worker queue length is lower than $Qmax_i$ and greater than service target time otherwise.

We build the service system model using AnyLogic simulation software [5,18] which supports discrete event simulation technique. We simulate up to 40 weeks of simulation runs. Measurements are taken at end of each week. No measurements are recorded during the warm up period of first four weeks. For our experiments, we consider request arrivals follow a Poisson model where the inter-arrival times follow an exponential distribution. In steady state the parameters that are measured include:

- SLA or the percentage of requests that meet target response time.
- Resource utilization (captures the busy-time of a resource)

We evaluate the simulation model with 10 SW (representing a single working shift). We simulate three scenarios to compare and contrast our results. The dispatching policy varies for these three scenarios. First, we have a model (Naive Dispatch Model) where a dispatcher dispatches request to a SW with minimum work queue length. This model is naive as it does not consider the $Qmax_i$ of SW_i when allocating the request to worker.

Second, we have a model considers all service workers to behave in a similar manner i.e. the $Qmax_i$ is set to an average value, for all service workers and is derived from the data ($\forall i : Qmax_i = Qmax_{mean} = 3$). The dispatcher dispatches the request to first SW with a work queue length less than $Qmax_{mean}$ (Common Behavior Model). The Common Behavior Model represents scenario where workers having same experience of skill are consider similar.

Last, we run the simulation model considering each SW_i to behave different i.e. $Qmax_i$ for each SW_i is set. The dispatcher dispatches the request to the first SW_i with a work queue length less than $Qmax_i$ (Advanced Model). The advanced model reflects our learning model where the dispatcher uses the worker queue length derived from past allocations, to decide current allocation. Based on the data, we observe and set the values as: $2 \leq Qmax_i \leq 4$ and $Qmax_{mean} = 3$. In latter two models, if there is no SW_i with a work queue length lower than the $Qmax_i$, then the request is routed to the SW with the minimum work queue length.

Table 2 shows the results obtained for the three dispatch models. The results indicate that the Advanced dispatch model outperforms the other two models in meeting the service quality. It is interesting to note that the naive model performs better than the common behavior model as the naive model tries to dispatch the request to the SW with the minimum work queue while the common behavior model assumes that workers with a work queue length lower than the

threshold will be efficient. This assumption leads to sub-optimal allocation of requests and hence the percentage SLA attained is much lower.

As discussed, our simulation model uses the parameter $Qmax_i$ to distinguish SW behavior and allocate request to compare and contrast the service quality with models that accommodate service worker behavior.

Table 2. Percent SLA and Percent Utilization for different dispatching models.

Dispatch Model	Percent SLA		Percent SW Utilization	
	Mean SLA	95 % Conf. Interval	Mean Utilization	95 % Conf. Interval
Naive Dispatch	86.84	(85.2,87.86)	64.4	(64.36,64.44)
Common Behavior	78.9	(78.08,79.72)	66.2	(66.15,66.24)
Advanced Dispatch	92.9	(91.78,94.02)	59.3	(57.71,60.89)

6 Threats to Validity

In this section, we identify the limitation of our study with respect of *construct validity, internal validity* and *external validity*.

Construct Validity. denotes that the variables are measured correctly. All the features or parameters used in the learning model been evaluated and used in earlier studies on dispatching, allocation and planning. Our study does not include additional parameters such as expertise, priority as they were not relevant to the system under study. We plan to extend our study to a service system where such parameters play a significant role.

Internal Validity. is established for a study if it is free from systematic errors and biases. During the measurement interval of 8 months, issues that can affect internal validity such as mortality (that is, subjects withdrawing from a study during data collection) and maturation (that is, subjects changing their characteristics during the study outside the parameters of the study) did not arise. Thus, we believe the extent of this threat to validity is limited.

External Validity. concerns the generalization of the results from our study. While insights can be drawn from our study, we do not claim that these results can be generalized in all instances. However, these results serve as the basis of using data driven approach for evaluating allocation of requests to workers effectively, leading to higher service quality.

7 Related Work

The problem of allocation of tasks to resources has been studied for some time now and there is a good body of literature dedicated to various aspects like

routing work to teams and dispatching tasks to resources. In [9], the authors use mixed integer programming (MIP) and a heuristic algorithm to allocate tasks to the resources based on their workload and skill, with an objective of meeting service quality. There are similar such scheduling and skill based routing of calls been addressed in the call center domain [12]. However, in all these scenarios, the inherent variations in human behavior and efficiency is not considered. Simulation models to evaluate the skill requirements of the team for a SS in the context work types discuss the improvement in service time of a SW over time through on the job learning [2,6]. Service workers of a skill and expertise level are assumed to have similar characteristics and learning factors. Given the complexity characteristics of human resources, in our work, we learn the allocation of tasks that would lead to favorable outcome from historical data and use it to guide future allocations.

Learning based predictive models, like ours, has been used for routing or dispatching work in SS. In [3], tickets or service requests are classified and routed to the right group using historical data. An approach to route the requests to multiple teams for resolving an IT problem ticket or incident, is addressed by [17]. Historical data is used to mine the sequence of groups or teams involved to further build a markov model that generates ticket transfer recommendations for an new arriving ticket. These studies focus on identifying suitable teams or groups and do not evaluate operational efficiencies of teams or workers.

In [16], the authors present an approach that uses historical data and illustrate the variance in operational productivity of workers for requests with different priorities and complexities. The variances in efficiency of workers is used to define policies for dispatching and optimally staffing teams. Our approach further demonstrates that data-driven techniques can be used to implicitly learn the efficiency of service workers and help in driving better allocation of tasks.

8 Conclusion

In this paper, we have evaluated the use of learning based model to predict and assist in allocation of tasks to resources. We observe that within a team, service workers of similar competencies vary in their efficiencies and have deterioration in the quality of service at different workloads or queue lengths. The model based on historical data has a prediction accuracy between 65 % to 80 %. The simulation model further indicates that modeling all workers as similar, results in lower quality of service. Through this work, we demonstrate that using of data-driven techniques to evaluate efficiencies of service workers, similar to ours, can serve as the basis for effective dispatching or task allocation policies and better meet the contractual service levels (quality) of the service system.

References

1. Van der Aalst, W.M., Nakatumba, J., Rozinat, A., Russell, N.: Business process simulation: how to get it right
2. Agarwal, S., Sindhgatta, R., Dasgupta, G.B.: Does one-size-fit-all suffice for service delivery clients? In: ICSOC, pp. 177–191 (2013)

3. Agarwal, S., Sindhgatta, R., Sengupta, B.: Smartdispatch: enabling efficient ticket dispatch in an it service environment. In: KDD, pp. 1393–1401 (2012)
4. Banerjee, D., Dasgupta, G.B., Desai, N.: Simulation-based evaluation of dispatching policies in service systems. In: Winter Simulation Conference, pp. 779–791 (2011)
5. Borshchev, A.: The Big Book of Simulation Modeling. Multimethod Modeling with AnyLogic 6. Kluwer, AnyLogic North America, Hampton (2013)
6. Dasgupta, G.B., Sindhgatta, R., Agarwal, S.: Behavioral analysis of service delivery models. In: Basu, S., Pautasso, C., Zhang, L., Fu, X. (eds.) ICSOC 2013. LNCS, vol. 8274, pp. 652–666. Springer, Heidelberg (2013)
7. Diao, Y., Heching, A., Northcutt, D.M., Stark, G.: Modeling a complex global service delivery system. In: Winter Simulation Conference, pp. 690–702 (2011)
8. Field, A.: Discovering Statistics Using SPSS. SAGE Publications, London (2005)
9. Gupta, H.S., Sengupta, B.: Scheduling service tickets in shared delivery. In: Liu, C., Ludwig, H., Toumani, F., Yu, Q. (eds.) Service Oriented Computing. LNCS, vol. 7636, pp. 79–95. Springer, Heidelberg (2012)
10. IBM (2008). http://www-01.ibm.com/software/analytics/spss/products/modeler/
11. Maglio, P.P., Vargo, S.L., Caswell, N., Spohrer, J.: The service system is the basic abstraction of service science. Inf. Syst. E-Bus. Manag. **7**(4), 395–406 (2009)
12. Mazzuchi, T.A., Wallace, R.B.: Analyzing skill-based routing call centers using discrete-event simulation and design experiment. In: Winter Simulation Conference, pp. 1812–1820 (2004)
13. Nakatumba, J., van der Aalst, W.M.P.: Analyzing resource behavior using process mining. In: Business Process Management Workshops, pp. 69–80 (2009)
14. Ng, A.Y., Jordan, M.I.: On discriminative vs. generative classifiers: a comparison of logistic regression and naive bayes (2001)
15. Ramaswamy, L., Banavar, G.: A formal model of service delivery. In: IEEE International Conference on Services Computing, SCC 2008, vol. 2, pp. 517–520, July 2008
16. Sindhgatta, R., Dasgupta, G.B., Ghose, A.: Analysis of operational data for expertise aware staffing. In: Sadiq, S., Soffer, P., Völzer, H. (eds.) BPM 2014. LNCS, vol. 8659, pp. 317–332. Springer, Heidelberg (2014)
17. Sun, P., Tao, S., Yan, X., Anerousis, N., Chen, Y.: Content-aware resolution sequence mining for ticket routing. In: Hull, R., Mendling, J., Tai, S. (eds.) BPM 2010. LNCS, vol. 6336, pp. 243–259. Springer, Heidelberg (2010)
18. Technologies, X.: (2011). http://www.xjtek.com/
19. Wetcher-Hendricks, D.: Analyzing Quantitative Data: An Introduction for Social Researchers. Wiley, Hoboken (2011)
20. Wickens, C., Hollands, J., Banbury, S., Parasuraman, R.: Engineering Psychology and Human Performance. Always learning, Pearson, Cape Town (2013). http://books.google.co.in/books?id=N3N0MAEACAAJ