Modeling Service Variability in Complex Service Delivery Operations

Yixin Diao and Larisa Shwartz
IBM Thomas J. Watson Research Center
Yorktown Heights, NY 10598, USA

***Abstract*— One of the key promises of IT strategic outsourcing is to deliver greater IT service management through better quality and lower cost. However, this raises a critical question on how to model highly variable services for diverse customers with heterogeneous infrastructure and service demands. In this paper we propose the use of statistical learning approaches for service operation variability modeling. Specifically, we use the partial least squares regression that projects service attributes to explain the service volume variability, and the decision tree approach to model the service effort based on categorical customer and service properties. We demonstrate the applicability of the proposed methodology using data from a large IT service delivery environment.**

I. Introduction

In recent years a number of expanding phenomena such as service automation and multi-sourcing have emerged in the IT industry and forced service delivery organizations to adapt their existing techniques and processes. The growth and success of service delivery also depends more on the agility in which service providers can develop and deploy new services quickly and be able to sustain profitability in the face of changing requirements. At the same time the delivery of IT services continues to heavily depend on customer specific infrastructure and business needs that define service operation volumes and efforts.

Delivering effective service operations with high level of service volume and effort variability is a challenging task. This challenge is further complicated by the accelerating need to deliver agile services, leading to an accelerating need to understand and manage the service variability in response to customer specific environments and demands. The virtualization and cloud computing technologies have progressed to largely reduce variability through enforcing standardization [1]. The manufacturing-oriented Lean Six Sigma quality management practices have also deeply impacted the management of service delivery processes to improve the speed and reduce the variability [2]. More recently, many research activities have surfaced investigating the integration of advanced analytical capabilities into service delivery solutions (e.g., [3], [4]). Our own experiences with service delivery customers have also revealed an increasing demand for using predictive analytics to derive deeper business insights from large amount of data collected during service operations.

In this paper we propose a predictive modeling methodology to study service variability in a systematic and data-driven means. The proposed modeling methodology helps to understand the underlying structure of service variability. It connects the observed operational variability (i.e., service workload volume and service effort time) to the underlying source of infrastructure and delivery variability (e.g., number of servers, types of applications, delivery skills). This helps to discover abnormal customer service behaviors in a systematic way, which further lays out the foundation for data-driven service issue diagnosis and identification in order to improve service quality and lower the management cost. Particularly, we use the partial least squares regression together with the input selection algorithm to model the service volume variability and the decision tree regression to explain the service effort variability. The proposed approaches have been implemented at a large services delivery provider with worldwide delivery locations and global customers.

The remainder of this paper is organized as follows. Section II presents the service variability modeling methodology. Section III summarizes the model evaluation results. Our conclusions and future work are contained in Section IV.

978-3-901882-77-7 © 2015 IFIP

II. Modeling Methodologies

In this section we briefly describe the statistical learning approaches that we use to model highly variable services resulting from the diversity of heterogeneous customer needs. Due to the different natures of service volume and service effort, we use the partial least squares regression together with the input selection algorithm to model the service volume variability based on the continuous service attributes (e.g., number of servers or images). Then we use the decision tree approach to explain the service effort variability based on categorical customer and service properties (e.g., industry sectors, service delivery locations).

A. Input Section

Input section determines which of the available metrics should be used as explanatory variables. There are more than 300 metrics that can be obtained from the service productivity data set. Intuitively, it seems that using more explanatory variables results in a better model. However, this is often not the case since the extra variables tend to model the “data noise” instead of the true relationship [5].

Several approaches exist for selecting explanatory variables as model input. Exhaustive search (e.g., false nearest neighbors [6]) examines all possible combinations of available metrics, but this is usually tractable only if the number of metrics is small. Ordered search explores the input possibilities according to certain importance measures, either in the incremented order or in the decremented order. One ordered search method is called stepwise regression. However, it may not necessarily produce the best model if there are redundant explanatory variables and may fail when applied to new data sets [5].

In this paper we use a two step approach to identify the explanatory variables. The first step orders the available metrics based on their correlation to the modeling residual and selects a subset of them as the explanatory variables [7]. The second step studies the orthogonal model structure and further reduces the model dimensions using a cross validation technique.

B. Partial Least Squares Regression

Although various nonlinear modeling techniques exist, we choose to use the linear modeling method in this paper to model the service volume variability. This is because linear models tend to be more robust than the more elaborate nonlinear models, especially when a large number of model inputs are involved but many of them exhibit poor data quality.

Particularly, we use the Partial Least Squares (PLS) regression approach [8], which is an extension of the commonly used Multiple Linear Regression (MLR) models. PLS regression is especially useful when there are limited number of observations as compared to the explanatory variables and when there is strong multicollinearity among the explanatory variables. Both are the case for modeling service volume variability where the number of service attributes overweights the number of service customers, and strong redundancy or correlation exists in the service attributes.

C. Decision Tree Regression

We resort to the decision tree learning approach to model the service effort variability. This is because we use customer and service properties (e.g., industry sectors, service delivery locations)) as the explanatory variables which are categorical in nature. (We have tried to explain the service effort variability using the continuous service attributes but cannot find a definite relationship; see more in the Evaluation section.)

A decision tree has a tree-like structure where each internal (decision) node represents a test on one of explanatory variables, each branch connecting to the internal node denotes an outcome of the test, and each leaf node leads to a decision on the response variable [9].

Decision tree learning can be used to build classification models when the response variable is categorical, or regression models when the response variable is continuous (which is the case in this paper). Specifically, we use the C4.5 algorithm that constructs the decision tree in a recursive manner that brakes down the dataset into smaller and smaller subsets.

III. Evaluation

In this section we provide an experimental evaluation to illustrate how the proposed modeling methodology can be used to model service volume and effort variability in a complex service delivery environment. Note that when necessary the data have been altered to preserve data privacy and simplified for the illustration purpose, though the nature of service operations has been maintained.

TABLE I
Top five activities performed in service operations ordered according to workload volume.

	Activity Name	Volume	Effort
1	Wintel Operating System Management	22,424	767,652
2	Capacity and Performance Management	11,069	346,752
3	OS Monitoring Software Management	4,852	137,927
4	Backup Management	1,744	75,212
5	Security Patch Management	1,667	75,860

A. Service Scenario

Our service scenario is characterized and represented through five sets of data: the workload data containing the incident volume, the effort data including the effort time on serving incident tickets, the service productivity data with attributes describing the delivered services, the customer directory data stating the industry sectors and geography locations of the service customers, and the service demographic data indicating the delivery countries and service skills.

Table I shows the incident volume and the total effort time for the top five activities that we are focusing on in this paper. They are ordered according to the workload volume, even if the total effort time also follows a similar order except for the last two activities. These data are collected over a three-week time period for 139 service customers. They contain 41,756 incident records out of a total of 56,910 records from all 35 activity types.

Each service activity is different to each other regarding the mean and standard deviation both for the workload volume and the effort time. They are further different for different customers, which we will explore in the paper to understand the source of their variability.

Figure 1 shows the correlation coefficient between the service attributes and the service volume and effort, respectively. The x-axis indicates the 351 service attributes from the service productivity data set. The y-axis shows the value of the correlation coefficient. The rows indicate the five top activities as shown in Table I. The columns indicate the service workload volume and the service effort time.

Note that many of the correlation coefficients for the workload volume have values close or larger than 0.5. This is an indication of potential linear relationship so that it may be possible to explain the service volume using the service attribute data. On the contrast, the correlation coeffcient for the service effort is generally much smaller so that we have to resort to other data sets (i.e., the customer directory data and the service demographic data) to study its variability.

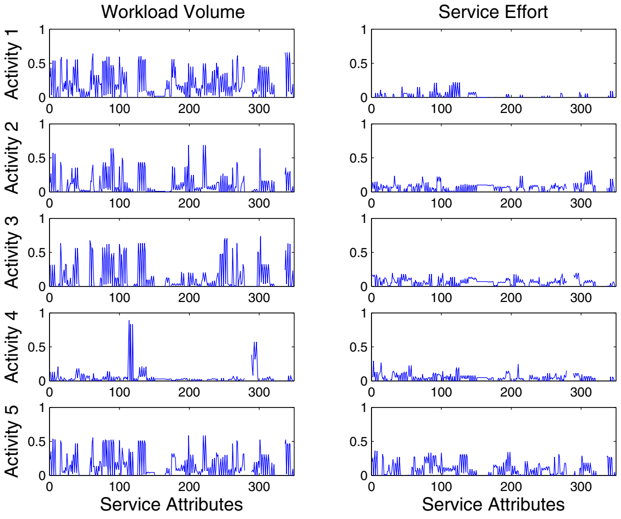

Fig. 1. Correlation analysis for service volume and effort for different activities and service attributes.

B. Service Volume Variability

To understand the underlying structure for service volume variability, we first use the input section method to select a subset of server attributes as the model input. With Wintel Operating System Management (Activity 1) as an example, Table II shows the first ten service attributes picked by the input selection algorithm as the explanatory variables. They are ordered according to their contribution to the final model (i.e., the correlation coefficient with respect to the model residual, $r_{residual}$), and thus follow a different sequence as the correlation coefficient between the service attribute and the model output, r_{output}.

Although the attribute with the highest correlation coefficient is the one that contributes the most to the model, the attribute with the second or third highest correlation coefficient may not contribute much because it can be correlated to the first so that its contribution is discounted. Indeed, this is the case for the tenth attribute (Storage Management Cost). Even if its correlation coefficient to the model output is the third highest (0.47), its contribution to the model is not ranked the third because it is strongly correlated to the first two attributes.

Based on the above variables, we build the PLS regression model to construct the orthogonal model structure in order to handle the multicollinearity among the explanatory variables, and use the cross validation technique to decide the number of latent variables to avoid model overfitting.

The modeling result is shown in Figure 2 for modeling the service volume variability of Wintel Operating System Management. The x-axis indicates the 139 different customers and the y-axis indicates the workload volume. The solid line represents the measured workload volume. The dashed line represents the predicted workload volume using the selected service attributes with examples shown in Table II, and indicates a close fit between the measured data and the predicted values.

TABLE II
TOP TEN SERVICE ATTRIBUTES FOR MODELING THE ACTIVITY VOLUME OF WINTEL OPERATING SYSTEM MANAGEMENT.

	Service Attributes	$r_{residual}$	r_{output}
1	MQ Cost	0.66	0.66
2	BackRst FTE	0.32	0.54
3	iSeries LPAR	0.24	0.08
4	SAP Systems / SAP FTE	0.22	0.13
5	BackRst Cost / Backup Server	0.18	0.03
6	SQL DBA Cost / DS SQL Database	0.18	0.09
7	Cost / Linux on zServer(Guests)	0.22	0.11
8	Cost / iSeries Server (LPAR)	0.17	0.21
9	Domino Server	0.17	0.12
10	Stg Mgmt Cost	0.17	0.47

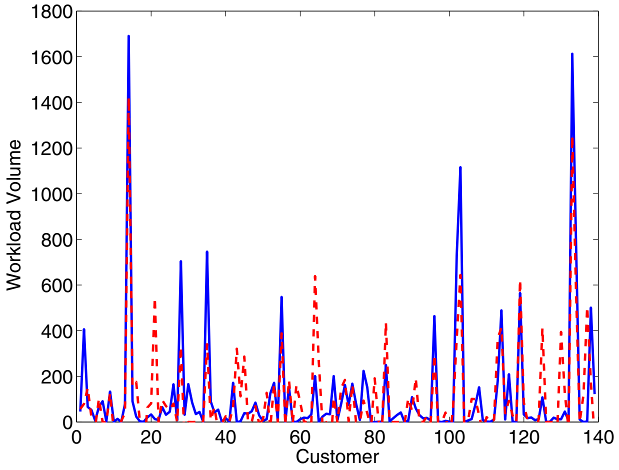

Fig. 2. Modeling result for workload volume of Wintel Operating System Management.

Table III includes the model accuracy metrics for all five top activities. For Wintel Operating System Management, the correlation coefficient is 0.82, an indication of strong linear relationship. The R^2 error is 0.67, which means the model can explain 67% of the variability in the workload volume data. The root mean square error is 147.97 as the average difference between the measured and predicted volume, which is reasonable given the workload variability across different customers.

C. Service Effort Variability

Note that we cannot get high correlation coefficients for modeling service effort variability using the service attributes from the service productivity data. This is because most of the service attributes are describing the service quantities (e.g., the number of different types servers or images to be managed). Even if the management cost also appear in the service attributes, the fact that they do not lead to strong correlation coefficients indicate that their calculation or definition may be

TABLE III

ACCURACY METRICS FOR MODELING WORKLOAD VOLUME OF TOP FIVE SERVICE ACTIVITIES.

	Activity Name	r	R^2	$RMSE$
1	Wintel Operating System Management	0.82	0.67	147.97
2	Capacity and Performance Management	0.88	0.77	105.70
3	OS Monitoring Software Management	0.90	0.81	65.38
4	Backup Management	0.79	0.62	21.98
5	Security Patch Management	0.84	0.71	1.64

over biased to the volume than the complexity of the work that is involved.

As the result, we turn to other data sets that indicate the service complexity and the effort needed to serve them. Particularly, we look at the customer directory data and the service demographic data. Figure 3 indicates the service effort variability for the top five activities . The x-axis indicates the seven industry sectors such as Computer Services, Distribution, and Financial Services. The y-axis indicates the average effort time per service incident. The different bars represent the five activities in the same order as that in Table III and are grouped together for each sector, respectively.

Since the customer directory data and the service demographic data are categorical in nature, we use the decision tree approach to identify the regression relationship for the service effort. Out of the various service properties, the decision tree uses the industry sector as the root decision node and the delivery country as the secondary decision node to construct the decision tree. This does make intuitive sense because different sectors may have different applications which affect the complexity to work on the service incidents. The service technicians from different delivery countries may possess different skill levels which also impact the effort time. On the other hand, the countries or geographies where the customers are located do not appear to be important for modeling the service effort.

Table IV shows the model accuracy metrics for all five activities. For Wintel Operating System Management, the correlation coefficient is 0.54. It is not as significant as that for the service volume, but still helps to indicate a causal relationship. The R^2 error is 0.30 which is on the smaller side, and the root mean square error is 11.06. Overall, all accuracy metrics in Table IV are worse than those in Table III. They indicate a stronger challenge for service effort modeling, but still provide useful insights in understanding the service effort variability.

We further compare the decision tree regression with the linear regression approach where dummy coding is used to turn categorical explanatory variables into continuous forms. As summarized in Table V, the decision tree regression shows better performance in this regard.

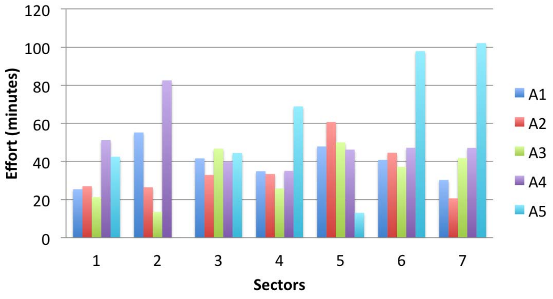

Fig. 3. Service effort variablity by industry sectors and activity types.

TABLE IV

MODELING ACCURACY FOR SERVICE EFFORT VARIABILITY.

	Activity Name	r	R^2	$RMSE$
1	Wintel Operating System Management	0.54	0.30	11.06
2	Capacity and Performance Management	0.71	0.50	10.14
3	OS Monitoring Software Management	0.44	0.19	16.22
4	Backup Management	0.47	0.22	14.82
5	Security Patch Management	0.82	0.68	7.91

IV. CONCLUSIONS AND FUTURE WORK

In this paper we proposed a predictive modeling methodology to model service operation variability. Due to the different natures of service volume and service effort, we used the partial least squares regression together with the input selection algorithm to model the service volume variability based on the continuous service attributes. Then we used the decision tree approach to explain the service effort variability based on categorical customer and service properties. We have demonstrated the applicability of the proposed methodology using data from a large IT service delivery environment.

While the initial results are encouraging, there are several areas for further improvement. First, we would like to explore different statistical and machine learning approaches. This will help to give us a better understanding of the limit of advanced learning techniques and their applicability in service operation management. Second, we would like to have more systematic study regarding the service productivity attributes. This may lead to the introduction of new service attributes that better reflect the service variability or may help to simplify the current attribute list to improve its usability.

TABLE V

MODELING ACCURACY FOR TOP FIVE ACTIVITIES.

Regression Method	r	R^2	$RMSE$
Decision Tree Regression	0.63	0.40	11.61
Linear Regression	0.53	0.28	12.64

REFERENCES

[1] Q. Zhang, L. Cheng, and R. Boutaba, "Cloud computing: State-of-the-art and research challenges," *Journal of Internet Services and Applications*, vol. 1, no. 1, pp. 7–18, 2010.

[2] M. George, *Lean Six Sigma for Service : How to Use Lean Speed and Six Sigma Quality to Improve Services and Transactions*. McGraw-Hill Education, 2003.

[3] S. Hagen and A. Kemper, "Towards solid IT change management: Auomated detection of conflicting IT change plans," in *Proceedings of the IFIP/IEEE International Symposium on Integrated Management*, 2011.

[4] J. Bogojeska, D. Lanyi, I. Giurgiu, G. Stark, and D. Wiesmann, "Classifying server behavior and predicting impact of modernization actions," in *Proceedings of International Conference on Network and Service Management*, 2013.

[5] F. E. Harrell, *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis (Springer Series in Statistics)*. Springer Verlag, 2001.

[6] C. Rhodes and M. Morari, "Determining the model order of nonlinear input/output systems," *AIChE Journal*, pp. 151–163, 1998.

[7] A. Keller, Y. Diao, F. Eskesen, S. Froehlich, J. Hellerstein, L. Spainhower, and M. Surendra, "Generic On-Line Discovery of Quantitative Models," *IEEE electronic Transactions on Network and Service Management (eTNSM)*, vol. 1, no. 1, Apr. 2004.

[8] S. Wold, A. Ruhe, H. Wold, and I. W. J. Dunn, "The collinearity problem in linear regression. the partial least squares (PLS) approach to generalized inverses," *SIAM Journal on Scientific and Statistical Computing*, vol. 5, no. 3, pp. 735–743, 1984.

[9] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2011.