A Study on the Relationship between Construction Projects and the Use of Virtual Technologies

Ivan Leung

Abstract

Virtual Design and Construction (VDC) consists of many useful tools to assist a construction project, including modeling technologies, process management tools, and collaboration methodologies. VDC usage data collected from 143 real construction project cases was used to predict the final performance of the project, and the best model achieved a 78.2% accuracy. However, more in-depth study of individual VDC uses is required to better understand the relationship between specific VDC use and the project outcome. Such understanding will ultimately help construction project managers apply VDC optimally to improve performance. This study tests the difference in 4 VDC uses between projects that performed well and projects that only had mediocre performance. Surprisingly, only the level of VDC formalization was found to be significantly different between the two groups, and the mediocre performance group actually adopted higher level of formalization. Further exploratory visualizations suggest substantial multiplicative effect among these 4 VDC uses, and future studies should take into account of more complex relationship between VDC use and project performance.

Introduction

Virtual Design and Construction (VDC) is a crucial component to any project in the Architecture, Engineering and Construction (AEC) industry today. Multi-dimensional modeling tools, process management protocols, and inter-stakeholder collaboration models are among the many important technologies and ideas in VDC that enable construction projects to improve their performances (Kunz, 2012). However, the uses of VDC in construction projects had not been systematically measured, thus the association between specific VDC applications and project outcomes could not be quantified on the industry level.

With the wide coverage and influence that VDC have on a construction project, it is important to understand how different uses of VDC would affect the ultimate goals of a construction project, including cost, schedule, and other project delivery objectives. There are many studies in construction management that studies the relationship between project performance and other aspects of construction; for example, Emsley et. al. (2002) predicted project outcome based on 6 strategic variables, 4 site-specific features, and 31 structural and design features (Emsley, 2002). While these studies are crucial to understanding the role of different components in a construction project, they provide few improvement opportunities for future projects: most of the variables studied are specific to the design or structure of the construction, and the owner or architects could not improve the project's outcome without imposing major design changes. In contrast, VDC comprises of many changeable parameters and processes that can be adopted on a construction project, such as the collaboration tool used for meetings, the software used for structural modeling, and the involvement of different stakeholders of the project. If the AEC industry understands how each of these changeable VDC attributes can ultimately affect a project's performance, project owners will be better equipped in deploying different VDC applications to improve project outcome.

The VDC Scorecard was developed here at Stanford University to evaluate VDC practices and the subsequent outcomes of the evaluated projects (VDC Scorecard). The VDC Scorecard is a hierarchical evaluation framework (Fig. 1): at the foundation lies a set of Metrics, which are quantitative measurements of VDC uses, such as number of quantitative VDC objectives established, number of project stakeholders involved in VDC. The response for each Metric is converted to a 0-100 percentile score based on industry experts' opinion. There are approximately 50 Metrics in the current version of the VDC Scorecard. On top of Metrics sits 10 Divisions. Each Division also has a 0-100 score, and the score is computed as a linear combination of scores of a subset of Metrics (again, the weights of the Metric scores are determined by industry experts). Similarly, linear combinations of Division scores produce 4 Area scores, and the 4 Areas are Planning, Adoption, Technology, and Performance. This hierarchical structure of the VDC Scorecard allows projects to be evaluated holistically on a broad spectrum of VDC uses while having objective quantities to support the evaluation of each aspect of VDC use.

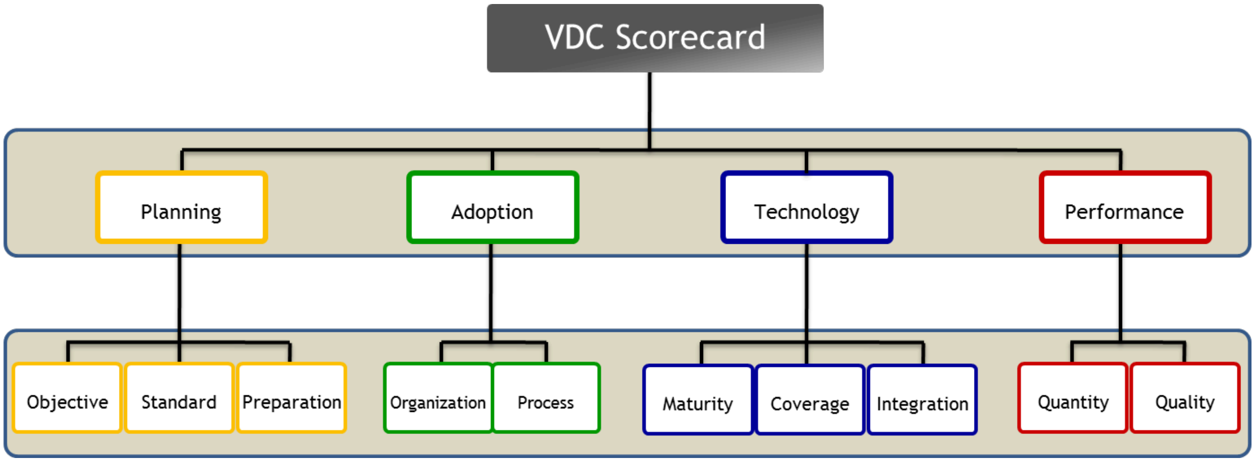


Fig. 1. Schematic of the hierarchical framework of the VDC Scorecard. The top level consists of 4 Areas, and each level consists of Divisions. Each Division is comprised of Metrics (not shown). The VDC Scorecard, in total, has 4 Areas, 10 Divisions, and > 50 Metrics. The score at each Area is a linear combination of scores of its children Divisions, and the score at each Division is a linear combination of scores of its children Metrics.

Motivation

Previous research of the VDC Scorecard has found statistically significant (but relatively weak) correlation between project performance and selected Areas and Division (publication in progress). However, these significant but weak correlations are far from offering conclusive evidence for the relationship between project performance and the surveyed VDC applications. While it is possible that (1) the surveyed VDC applications and features play important roles in affecting project outcomes but require more complex models to quantify their combined effects, it is also possible that (2) the surveyed VDC applications and features indeed have little relevance to project performance.

Thus, ongoing research is focusing on establishing more complex model to predict project performance using the VDC feature data collected. From 2012-2014, 143 real construction project cases were surveyed; after eliminating cases with obviously erroneous data and cases that responded to different versions of the Scorecard, 97 project cases are available for analysis. With each of the 96 project cases, project performance was measured as the overall satisfaction at the return of investment of using VDC. The performance is aggregated into three categories: below expectation, meeting expectation, and exceeding expectation. To test our hypothesis that a more complex model is required to quantify the relationship between VDC use and project performance, we built a customized multilayer neural network using a combination of L1 and L2 regularizations to predict the category of project performance with 51 VDC features collected from the VDC Scorecard. The model achieved a 10-fold cross-validation accuracy of 78.2% using 2 hidden layers. This relatively high prediction accuracy validates the relevance of the VDC features collected in this research and shows the association between these VDC features and subsequent project performance.

However, a more complex model tends to be less interpretable. In particular, the weights of the input features in the multilayer neural network could not be directly used to infer the features' importance to the model (even after centering the mean and standardizing the variance) because the inputs also pass through multiple hidden layers with nonlinear transformations. Although it is not possible to discern the most important features, one could look at any features that have relatively small or close-to-zero weights and discard them from our model, since the values of these features will be drowned out by other features that have significantly heavier weights. L1 regularization was used in the input layer because L1 regularizations tend to "zero-out" feature weights and is preferred for feature selection. Interestingly, the weights of the input features (Fig. 2) shows no features with weights close to 0, which seems to indicate that all the VDC features contribute significantly to the neural network model.

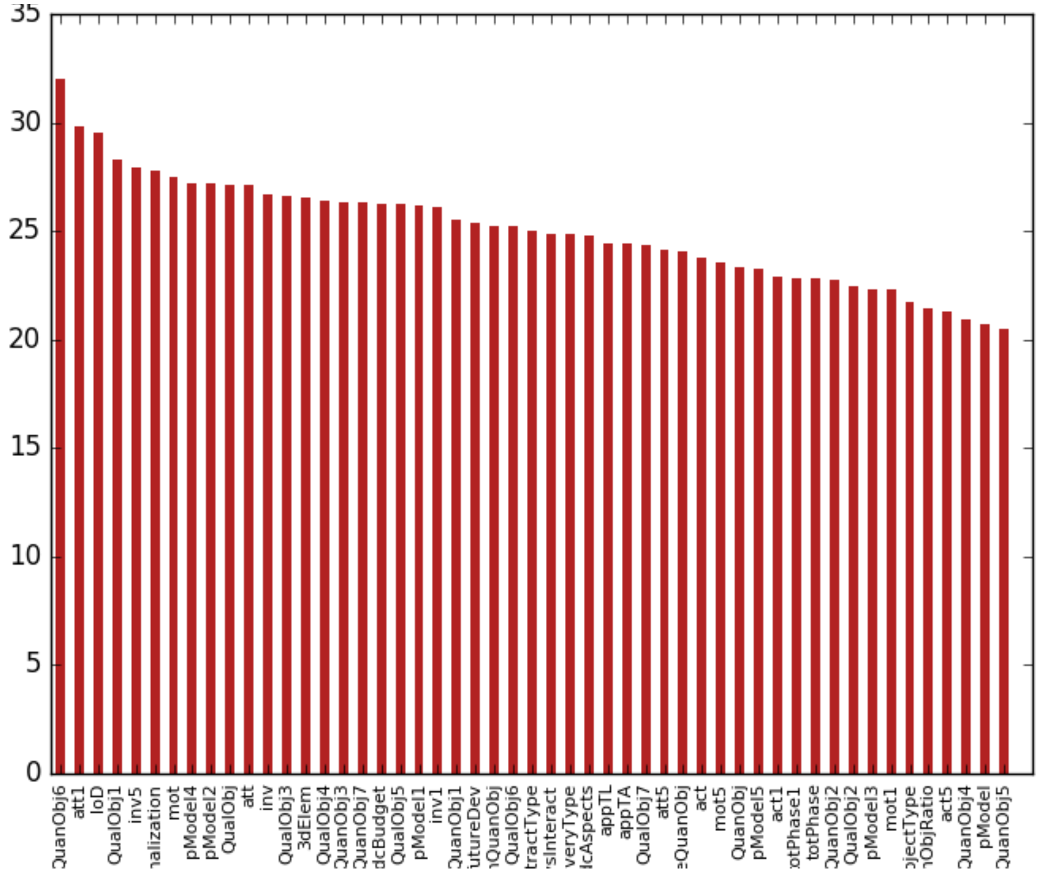


Fig. 2. Sum of input weight magnitudes in neural network that achieved highest test accuracy. Note that none of the feature weights are close to 0, thus they cannot be directly regarded as redundant to the model.

Since the neural network produces accurate prediction (78.2% accuracy) but does not rank the importance of VDC features or quantify how the distribution of feature values differ for different performance categories, this paper seeks to better understand the differences in VDC use for projects that have different outcomes.

The Data

96 construction project cases were evaluated from 2012 to 2014 using the VDC Scorecard. VDC practices as well as project outcomes are measured for each of the 96 project cases.

Project outcome is measured in the form of overall satisfaction at the return of investment of using VDC. The original research classifies the performance into 3 categories: below expectation, meeting expectation, and exceeding expectation. However, due to the limited number of samples for projects that are "below expectation", this study collapses the first two performance categories together—for the purpose of this study, a project's performance is either "below/meeting expectation" or "exceeding expectation”. This classification has advantageous practical implications as well: the AEC industry is most concerned with VDC uses that differentiates an excellent project from a mediocre one, and this classification scheme aligns with this objective.

4 features are chosen for this study. Each of the 4 features are chosen because they have relatively large variance among the project cases, contains little to no missing data among the project cases, and have been postulated by the VDC Scorecard research group to be influential to project outcome. After selecting project cases that do not have missing values in the 4 features, 95 project cases remain. The 4 features are described below:

1. Degree of VDC Formalization
   This feature takes integer values from 1 to 5, and a larger value indicates higher degree of formalization of VDC. For example, level 5 indicates that the project has "documented, shared, and contractually agreed to by multiple stakeholders".

2. # of Quantitative Objectives
   This features take any nonnegative integer value, and is defined as the number of quantitative VDC-related objectives that a project has.

3. # of Levels of Modeling Applications
   The VDC Scorecard classifies modeling applications into 5 levels, and this feature counts the number of levels that has at least 1 modeling application employed by the project. Thus, this feature takes integer values from 0 to 5.

4. Average Stakeholder Involvement in VDC
   A project typically has multiple stakeholders, who are either involved or not involved in VDC. This feature is defined as the sum of stakeholders who are involved in VDC divided by the total number of stakeholders involved in the project. Thus, this feature takes a value from [0,1].

Methods

The objective of this study is to investigate the difference in VDC usage between projects that perform well and those that have mediocre/unsatisfactory performance. The 4 VDC features and 1 performance measurement described in the Data section focuses our scope to the differences in VDC usage as quantified by the 4 features between projects that are “below/meet expectation” and those that “exceeds expectation”. *Let **Group 0** denote that group of projects that perform below or at expectation, and **let Group 1** denote the group of projects that perform beyond expectation*. In particular, this study will look at “differences” in the form of (1) differences in median, (2) differences in median, and (3) differences in distribution.

(1) Differences in median

Two nonparametric tests are relevant in testing the difference in median between two samples: Wilcoxon test for shift in location, and bootstrap test for difference in median.

Wilcoxon test for shift in location

The Mann-Whitney test statistic can be used to test whether the PDFs of two samples differ by a shift in location, assuming that the centered distribution for both PDFs are the same.

However, because this test uses ranked values, too many ties in the data would obfuscate the meaning of the test statistic ($T^+ = \#_{i,j}\{X_i < Y_j\}$ could take a wide range of values depending on the way ties are broken). Features [1] and [3] have many ties, so their test results should be considered carefully. In addition, the assumption that the both PDFs have the same centered distribution is quite debatable, especially when features [1], [2], [3] only admit a narrow range of values, and any substantial location shift would already move many data points outside of the allowed ranges.

Bootstrap test

The bootstrap test assumes the sample is representative of the population. The sample sizes (48 from Group 0, 47 from Group 1) are large enough to take a representative sample of the population. The project cases are all executed by different teams, even though some projects belong to the same broader company. But since the projects operate independently from each other, we could make the assumption that the project cases are independent from each other. The projects are also sampled from different parts of the US and from around the world. Thus, the representativeness of the samples seems to be a reasonable assumption for the bootstrap test.

For bootstrapping difference of medians between Group 0 and Group 1, the null hypothesis is:

$$H_0\text{: median of Group 0 = median of Group 1}$$

For a given VDC feature, let the values from Group 0 be $\boldsymbol{x} = x_1,\ldots,x_{n0}$, let the values from Group 1 be $\boldsymbol{y} = y_1,\ldots,y_{n1}$, and let the pooled data be $\boldsymbol{d} = d_1,\ldots,d_{n0+n1}$.To generate bootstrap samples from the null hypothesis, the data from Group 0 and Group 1 are centered to create null vectors **x'** and **y'**:

$$x'_i = x_i - median(x) + median(d)$$
$$y'_i = y_i - median(y) + median(d)$$

B bootstraps of (x'*, y'*) pairs are sampled, and the difference in medians between x'* and y'* are recorded. The bootstrap p-value is given as follows:

$$p = \frac{\#\{abs(median(x'^*) - median(y'^*) \geq abs\big(median(x) - median(y)\big)\}}{B}$$

(2) Differences in means

Bootstrap test

As opposed to differences in medians, the differences in means cannot be tested using the Wilcoxon test, since Mann-Whitney test statistic transforms the data into ranks, and thus tests for shifts in the median but not the mean. However, the bootstrap test can be adapted to test for differences in means. The formulation is the same as the test for differences in medians, except that the estimator is mean instead of median. Using the Group 0 vector ***x***, Group 1 vector ***y***, and the combined vector ***z*** as above, the null vectors **x'** and **y'** are obtained by:

$$x'_i = x_i - mean(x) + mean(d)$$
$$y'_i = y_i - mean(y) + mean(d)$$

Again, using B bootstraps, the p value is

$$p = \frac{\#\{abs(mean(x'^*) - mean(y'^*) \geq abs\big(mean(x) - mean(y)\big)\}}{B}$$

(3) Differences in distribution

Each of 4 features studied is neither strictly categorical nor strictly continuous. Feature 1, “Degree of VDC Formalization”, consists of 5 ordered categories and is ordinal data. Feature 2, “# of Quantitative Objectives”, can be viewed as data drawn from a discrete distribution. Feature 3, “# of Levels of Modeling Applications”, is ordinal data just like feature 1. Feature 4, “Average Stakeholder Involvement in VDC”, is closest to continuous data; however, the VDC Scorecard records at most 9 types of stakeholders in a project, so the data has limited continuity.

On one hand, the test of homogeneity applies to discrete RVs and assumes a multinomial distribution for a given sample. On the other hand, the two-sample Kolmogorov-Smirnov test can be applied to continuous data, but does not yield exact p-value in the presence of ties (and all our features have ties). In fact the ks.test() R function reports an error at the presence of ties.

Thus, I have decided to use the bootstrap KS test instead, which uses the pooled empirical distribution as the null hypothesis; this bootstrap version of KS test can be applied even in the presence of non-continuous data and ties (Sekhon, 2015) as long as the data has ordinal values.

(4) Alpha value

To control for false positives with multiple hypotheses, I decide to control, for each of the tests (1) – (3), to have a <= 0.05 probability of having false positive(s). Thus, for each feature, the alpha level is computed as follows:

$$\text{P(false discovery)} = 0.05$$
$$\text{P(at least 1 false discovery among 4 variables)} = 0.05$$
$$1 - \text{P(no false discovery among 4 variables)} = 0.05$$
$$[\text{P(no false discovery for 1 variable)}]^4 = 0.95$$
$$\text{alpha} = 1\text{-}0.95^{(1/4)} = 0.0127$$

Results/Discussions

(1) Difference in medians

The Wilcoxon Test shows statistically significant differences only for feature [1], "Degree of VDC Formalization". The p-values and confident interval (for median(Group 0) – median(Group 1) computed by the Wilcoxon test are tabulated below:

```
##                                  p-value   lower bound  upper bound
## Degree.of.VDC.Formalization  0.003793237  3.745067e-05 1.999998e+00
## Quantitative.Objective.Count 0.940413672 -8.479352e-05 3.821309e-06
## Levels.of.Applications       0.140652725 -5.476408e-05 9.999372e-01
## Avg..Stakeholder.Involvement 0.801514161 -1.249291e-01 1.427875e-01
```

It was actually surprising to see no significant difference in medians for all features except feature 1, "Degree of VDC Formalization", in which case Group 0 (the mediocre group) actually has a higher median than Group 1 (the exceptional performance group). However, the conclusion of the Wilcoxon Test should be taken with caution, since the two groups may not have the same centered distribution.

Even so, the bootstrap test confirms the above results:

```
##                              mean diff under null      p
## Degree.of.VDC.Formalization           -0.19640000 0.0065
## Quantitative.Objective.Count          -0.15925000 1.0000
## Levels.of.Applications                -0.51975000 0.4149
## Avg..Stakeholder.Involvement          -0.01245609 0.6605
```

Again, only "Degree of VDC Formalization" has significant difference in median at the alpha value of 0.0127.

(2) Difference in means

For this comparison, the Wilcoxon Test is not applicable since it only tests for median differences. The bootstrap test is used, and again, only "Degree of VDC Formalization" shows significant differences at the alpha value of 0.0127.

```
##                              mean diff under null      p
## Degree.of.VDC.Formalization         -0.0002355053 0.0115
## Quantitative.Objective.Count          0.0038957004 0.3768
## Levels.of.Applications               -0.0008519504 0.0796
## Avg..Stakeholder.Involvement          0.0005160528 0.2674
```

These empirical results are quite unexpected, and our research group did expect to see some significant differences between the two performance groups. Bootstrapping the confidence intervals for both groups separately gives us some insights on how the collected data is distributed.

First we start with the CI for medians:

```
##                                median (G0) lower (G0) upper (G0)
## Degree.of.VDC.Formalization  5.0000000       4.00   5.0000000
## Quantitative.Objective.Count 0.0000000       0.00   1.5000000
## Levels.of.Applications       5.0000000       4.00   5.0000000
## Avg..Stakeholder.Involvement 0.8333333       0.75   0.9285714

##                                median (G1) lower (G1) upper (G1)
## Degree.of.VDC.Formalization  3.0000000      3.000          4
## Quantitative.Objective.Count 0.0000000      0.000          1
## Levels.of.Applications       4.0000000      4.000          5
## Avg..Stakeholder.Involvement 0.8571429      0.625          1
```

Although the CIs cannot be used directly to test for median differences, they are good for exploratory analysis to see if there are any abnormal trends with the overall data distribution.

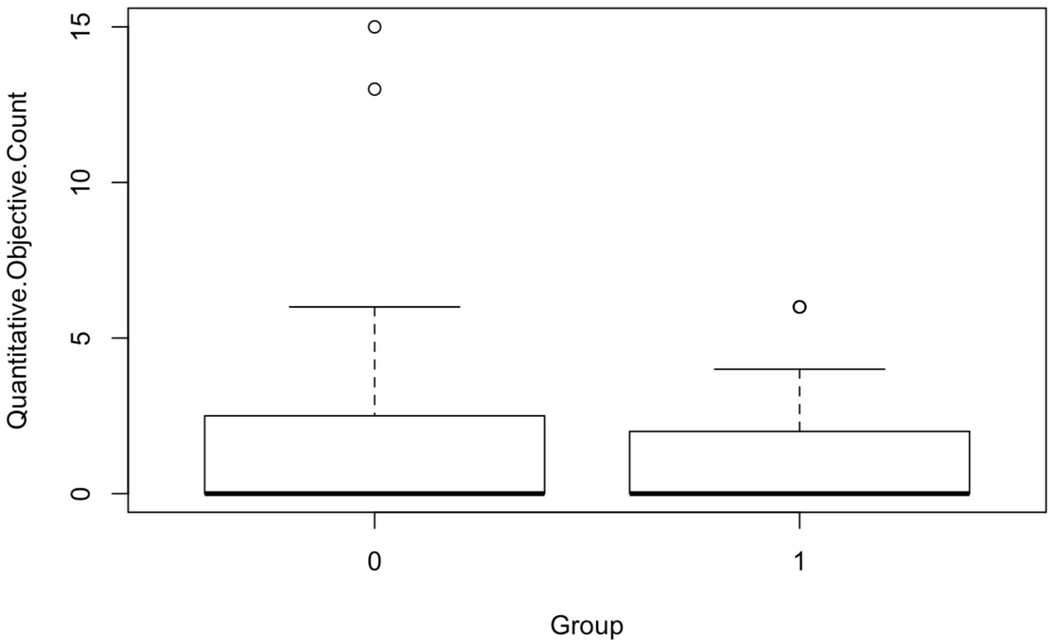
We see that the CIs for features 2-4 are mostly overlapping for the two groups. However, in feature 3, “# of Levels of Modeling Applications”, both groups attain an upper bound of 5, the maximum number of levels possible. It is possible that the categorization of modeling applications is too coarse to reflect the difference in modeling use between the two groups, if any. In this case, perhaps future versions of the survey should categorize modeling applications into more fine-grained levels, so that each project case is less likely to have applications in all levels.

Similarly, the CI for means are as follows:

```
##                                mean (G0) lower (G0) upper (G0)
## Degree.of.VDC.Formalization  4.1041667  3.6041667 4.5416667
## Quantitative.Objective.Count 1.6666667  0.7083333 2.9375000
## Levels.of.Applications       4.3750000  4.0625000 4.6458333
## Avg..Stakeholder.Involvement 0.7970899  0.7314649 0.8597966

##                                mean (G1) lower (G1) upper (G1)
## Degree.of.VDC.Formalization  3.4680851  3.0425532  3.8723404
## Quantitative.Objective.Count 1.2127660  0.6170213  1.8723404
## Levels.of.Applications       4.0212766  3.5957447  4.4042553
## Avg..Stakeholder.Involvement 0.7446049  0.6394461  0.8406535
```

Again, there is substantial overlapping for features 2-4. It is interesting to note that for feature 2, “# of Quantitative Objectives”, the interval range for Group 0 is 2.229, much higher than the interval range for Group 1, which is 1.25. A boxplot for this feature shows some outliers that contributed to the higher variance in Group 0. From this empirical observation, it seems that the median is a better measurement of average for this feature.



(3) Difference in distributions

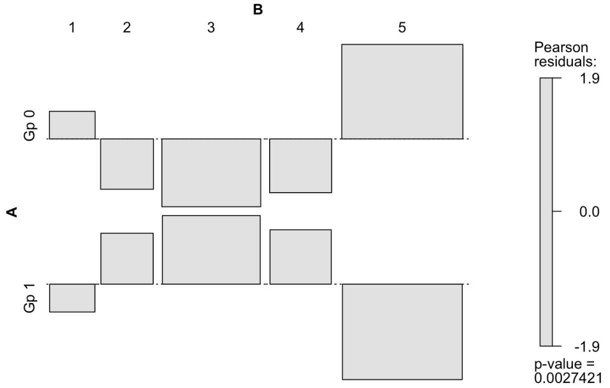
The results of the bootstrapped KS test are as follows:

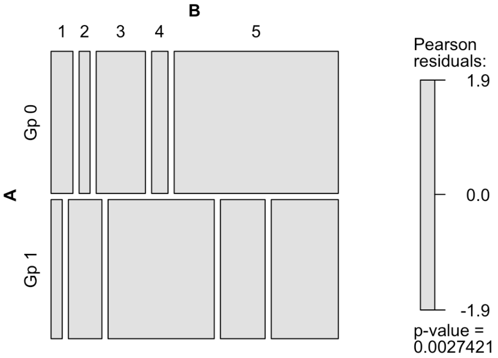
```
##                                        D p-value
## Degree.of.VDC.Formalization  0.36968085  0.0005
## Quantitative.Objective.Count 0.05851064  0.8936
## Levels.of.Applications       0.17287234  0.1267
## Avg..Stakeholder.Involvement 0.17287234  0.2733
```

Again, only “Degree of VDC Formalization” shows statistically significant difference in distribution at the alpha level of 0.0127.

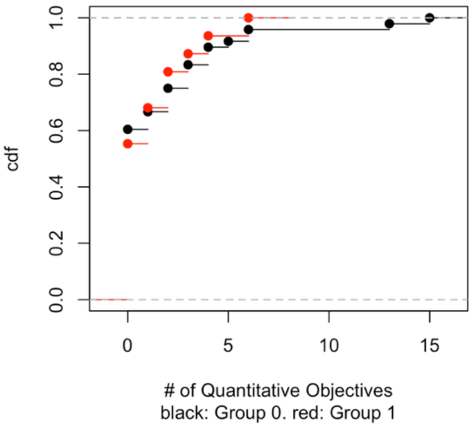
Future work

Since the hypotheses features 2-4 are all rejected, it is necessary to perform some exploratory visualizations to get some pointers for next steps. The association plots and mosaic plots for features 1 and 3 are shown below, since the data only has 5 categories. For feature 2 and 4, the data is ordinal and consists of more possible values, so association and mosaic plots would not be appropriate. Empirical CDFs are plotted instead.

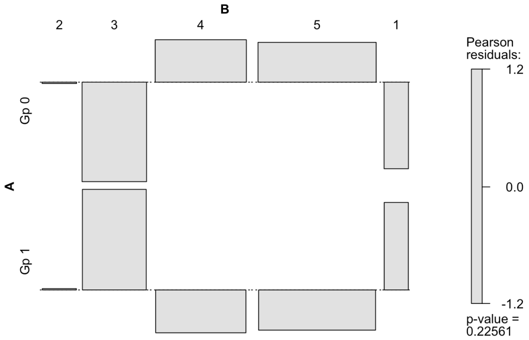


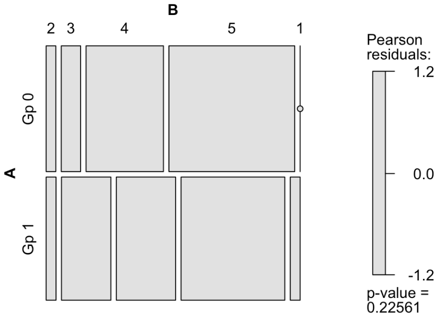


Feature 1, “Degree of VDC Formalization”. The two distributions differ most in level 5.

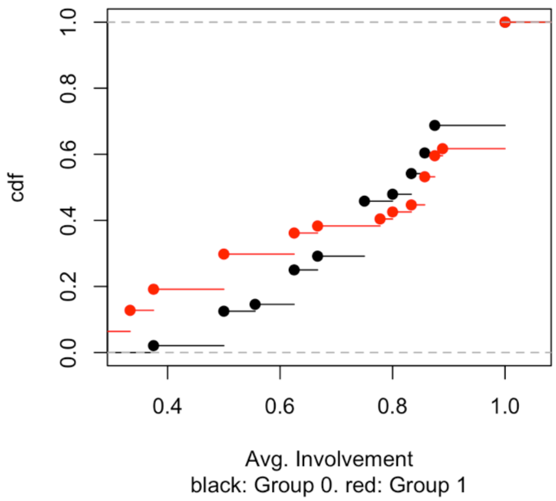


Feature 2, “# of Quantitative Objectives”. The empirical CDFs are very similar. However, note that most project cases have 0 quantitative objectives. Perhaps this feature is not a good reflection of VDC use when almost half the data concentrate on one value.



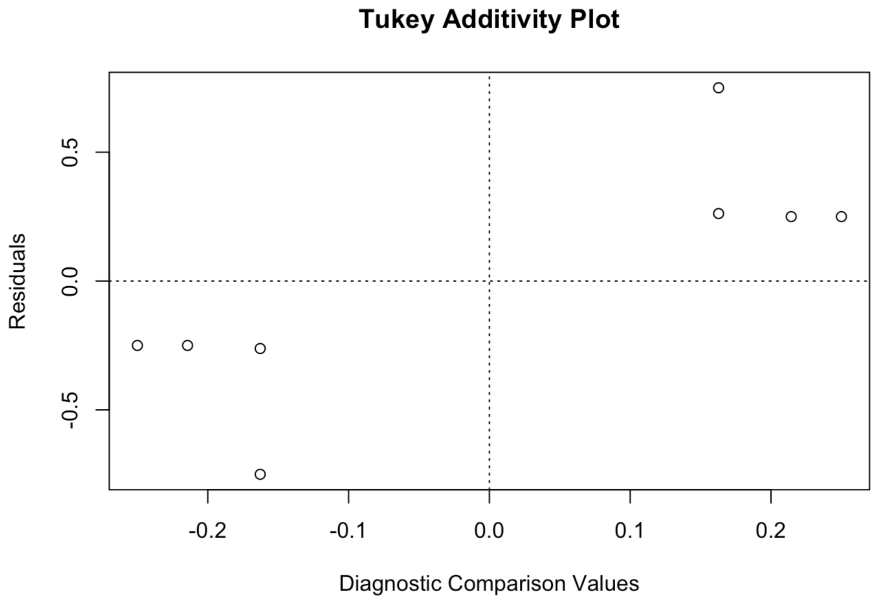


Feature 3. “# of Levels of Modeling Applications”. There is slight difference for all the levels, but none of them really deviate significantly.



Feature 4. “Average Stakeholder Involvement in VDC”. The CDFs seem to differ most at both ends. The current value is calculated by the average proportion of stakeholders involved in VDC, but does not take into account of their involvement. For future surveys, the definition of “involvement” could be further refined to include different VDC activities in which each stakeholder is involved, so that we may differentiate highly involved stakeholders from others.

Finally, the Tukey additivity plot shows rather significant trends between the comparison values and residuals. This suggests that project performance could be related to multiplicative effects among the 4 features. Future studies should definitely include a multiplicative model.



Bibliography

Emsley, M. W., Lowe, D. J., A, R. D., Harding, A., & Hickson, A. (2002). Data modelling and the application of a neural network approach to the prediction of total construction costs. *Construction Management and Economics, 20*(6), 465-472.

Kunz, J., Fischer, M. (2012). Virtual Design and Construction: Themes, Case Studies and Implementation Suggestions. CIFE WORKING PAPER #097, Stanford University.

Rostami, Afshin. L1 vs. L2 Regularization and feature selection. http://cs.nyu.edu/~rostami/presentations/L1_vs_L2.pdf

Sekhon, J., S. (2015). Package 'Matching'. R package. https://cran.r-project.org/web/packages/Matching/Matching.pdf

VDC Scorecard. https://vdcscorecard.stanford.edu/content/vdc-scorecard