STAT COE-Report-25-2017

Iterative Developmental Testing Considerations

Authored by: Michael Harman

29 January 2018

Revised 30 August 2018

The goal of the STAT COE is to assist in developing rigorous, defensible test strategies to more effectively quantify and characterize system performance and provide information that reduces risk. This and other COE products are available at www.afit.edu/STAT.

STAT Center of Excellence
2950 Hobson Way – Wright-Patterson AFB, OH 45433

Table of Contents

Revision 1, 30 Aug 2018: Formatting and minor typographical/grammatical edits

Executive Summary

Developmental testing (DT) in the Department of Defense (DOD) is an iterative process by which a system is evaluated and improved before operational testing (OT) and fielding. Employing design of experiments (DOE) in DT results in a unique iterative process. Additional considerations for test designs and regression modeling are needed for evolving configurations that purposefully change performance. Setting analytical goals helps determine how the team will deal with re-testing. DOE provides tremendous information that must be reported clearly and precisely to support the systems engineering process effectively. This paper provides practical considerations for these iterative processes.

Keywords: sequential experimentation, defense testing, regression modeling, iterative testing

Introduction

Design of Experiments (DOE) is a tool to facilitate rigorous testing and provide high quality analysis. Developmental testing (DT) in the Department of Defense DOD is an iterative process by which a system is evaluated and improved before operational testing (OT) and fielding. Employing DOE in DT requires a unique iterative process when the tested system configuration is changed. The considerations presented here primarily relate to configuration changes implemented to correct performance shortfalls. This paper provides practical details and provides several questions to ask to execute this iterative process effectively. Smaller (certainly a subjective term) changes may not be subject to the rigor detailed below and some options are touched on but are not the focus. This paper assumes the reader already has an understanding of DOE and sequential testing. For detailed DOE texts see Box, Hunter, and Hunter (2005) or Montgomery (2017).

Critical Analytical Goals

Setting analytical goals helps determine how the team will test successive configurations. Analytical goals may vary and should be specific to each test program. Several broadly applicable goals are discussed below.

Keep the Regression Model Current

A good regression model should be able to describe performance at every point inside the factor space; however, it is only relevant to the latest configuration that underwent testing. Keeping the model current with the latest configuration allows the team to use it later for engineering, mission planning, or tactical development needs. Therefore, executing only screening designs after a configuration change may not be sufficient to generate a full regression model (unless main effects are the only significant terms). Note that for small changes, late in testing, it may suffice to use only validation points with a new configuration to determine if the previous configuration’s model still applies. This should not be used as the primary method in early testing (especially when many changes are being incorporated) because it can introduce some subjectivity that the sequential screening/augmentation process strives to avoid.

Routinely Check Model Efficacy

Using validation points enables the team to assess the accuracy of the model at other points inside the factor space not already tested. Risk areas, or other regions of interest (including a particular factor combination that is meaningful to subject matter experts) may not be adequately covered by the design points alone. The addition of validation points facilitates the evaluation of requirements, risk, and corrections at locations within the factor space of the team's choosing and also bolsters confidence in the accuracy of the model.

Considerations for Iterative Testing

DT is iterative by nature. A system configuration is evaluated and then modified to address deficiencies and the process repeats until the requirements are met. Sequential DOE (screening, augmentation, and validation) still applies to testing each configuration, but evolving configurations that purposefully change performance require additional considerations. The first screening test is meant to identify significant factors. Augmenting that design to address higher order terms informs the evaluation against requirements. Finally, validation provides information about how well the regression model performs throughout the factor space. This process will need to be repeated again with a new configuration, but what other considerations should be addressed?

Questions for Iterative Re-Screening

The initial screening will identify significant factors, but how are these to be viewed after the system has been changed?

- Should current configuration screening be limited to factors significant in the last test?
- What if the change alters significant factors and others become significant?
- Are we looking to confirm or refute expectations from the latest configuration updates?
- What is the resource cost to re-screen all factors versus a reduced list?
- What is the information (decision) risk of missing significant factors?

Screening Design Considerations

Factor significance will be evaluated after the initial screening design and then decisions are made that impact the content of the subsequent augmented design. Because the configuration has changed, there is some probability the significant factors will change. However, complete re-testing consumes resources. While the team must address specifics for their system under test, here are some suggestions.

- Re-screen all factors every time. Since screening designs are very efficient and the data points become part of the augmented design, these designs are an efficient use of resources. If a new significant factor emerges after a configuration change, then this screening will be well worth its cost.
- Different significant factors between configurations indicate system performance has changed. This new information can provide a rough capability to determine if desired changes may have been successfully realized or not.

The choice of the type of screening design is important. DOE software can easily generate custom computer-generated screening designs built by optimizing only one criterion, but classical fractional factorial designs have many beneficial design properties as well. Fractional factorials can be used to screen all 2FI without the manual labor that custom designs require in some software suites. Be sure to choose at minimum a resolution IV design so that ME are not aliased with 2FI. There will be some aliasing between 2FI and other 2FI, but, if it is minimal, the results for ME significance should help discern which 2FI are truly present (for more, see Natoli [2018]).

- Include as many 2FI as possible with every screening iteration.
- As testing proceeds through many configurations, you may have more information that indicates certain 2FI will be present or are of particular interest. If so, be sure to include those with minimal aliasing.

Evaluate the root mean square error (RMSE) and recalculate the SNR after each test. Designs created using an accurate SNR will be correctly sized for model term power and will correctly allocate resources. This idea also applies to the augmented designs described below.

Questions for Iterative Design Augmentation

Similar to screening designs, questions are raised regarding the augmentation of the screening designs:

- Should previously significant factors be considered if they are not significant in the current screening?
- Should all potential two-factor interactions (2FI) be added in every iteration?
- Can some terms be objectively eliminated based on previous results?
    - Including all 2FI may require a large number of test points.

Design Augmentation Considerations

Given what was learned through screening, you will augment the design to include only the significant factors and the desired higher order terms. Factor significance will potentially/probably be different for every response. If this is the case, there are two options:

- Augment the design using all the significant factors from all responses.
    - Allows the use of a single design
    - This option increases the probability the design may be significantly larger than necessary for any single response
- Create different augmented designs for each response.
    - Results in multiple designs
    - Smaller focused designs may be less resource costly than a single larger design

Augmentation also builds subsequent designs based on the information derived from screening. If a previously significant factor does not surface in the current round of screening, it means the impact of that factor was not observed in the data.

- Augment the design with and without this additional factor and determine the resource cost to include it. With a large number of factors, adding an additional one may not impact the design size enough to justify leaving it out if it is of particular interest.
- Compare both design sizes to make an objective resource determination.

Including all 2FI may drive the design size up significantly. However, few 2FI may be truly present in the system and/or few may be of interest.

- If 2FI were NOT screened originally, then include any associated with the significant main effects.
    - If resources do not allow this, then include all that are potentially relevant based on the physics of the system.
    - There is a risk that a significant 2FI might be missed if all are not included.
- If 2FI were screened, then include the significant ones in the model.
    - If some are aliased with other 2FI, use the significant main effects to objectively weed-out the insignificant 2FI or include all that might be significant. One can also augment/fold the design strategically to break the alias chains.

A Comment on Violation of DOE Fundamentals

Typically, the transition into actual testing uncovers execution details that were previously not well understood. Issues with randomization, factor level selection or setting, equipment or measurement limitations, noise or other nuisance factors, and a myriad of others are typical. These issues may impact the ability of the test design to effectively isolate and estimate significant factors and need to be recorded for consultation during analysis. This information may also demonstrate a need to restructure the designs to address an inadequate SNR or correct for lack of randomization so the analytical goals are met. The critical point is to analyze the data as the design was actually executed, not simply how it was planned.

Questions About Model Validation

How much validation is required at each step in the iterative evaluation process?

- How will points be selected (quantity and location)?
- Should the points focus on the entire space or address identified risk areas?
- What are the resource costs and impacts?

Use of Validation Points

The number and location of validation points are not defined by any hard rules (see Kutner [2004] or NIST [2017] for discussion). The first time through you may want to choose them randomly throughout the space to evaluate model accuracy in a broad sense. As testing proceeds, the validity of the model may be more important in areas where higher risk is identified or in areas with higher prediction variance and less so in regions with better performance margins. Furthermore, risk regions addressed by the new configuration may be between design points so the only way to judge actual performance is to add points into that region. Lastly, validation points may uncover regions of interest where the model performs poorly. This may help the team focus efforts to add additional factors or model terms to improve the model or fit higher order terms if lack of fit remains significant.

Reporting Test Results

For evaluation purposes, the main benefit of DOE is the statistically rigorous determination of an empirical regression equation that includes mathematically-estimated interactions. This model facilitates estimation of the response throughout the factor space while accounting for the percent of variability in the response captured by the model (R^2) and error (RMSE). This estimate of the response can be compared to the requirement. Sufficient sequential testing should result in an accurate model that supports insightful information gathering.

However, determination of significant factors and the performance assessment can only facilitate a correction of deficiencies when it is reported clearly. The report must include regions of good margin, high risk (low margin), and failing performance along with the significant factors and interactions determined from testing. The contributing factors may help the developer determine root cause and effectively implement changes. The report can also provide the actual regression equation and the electronic analysis files (e.g., JMP, Design Expert, etc.).

A straightforward format for reporting this information is given below.

- Response name
- Factors screened
  - List all original factors
- Significant factors at XX% confidence
  - List factors/terms to include any interactions
- Significant factors comments
  - Include any noteworthy details about changes in significance between iterations
- ZZ% of the factor space passes the requirement
  - This can be accomplished with Monte Carlo simulation on the regression equation throughout the factor space (JMP software includes this in the profiler platform)
  - Provide any noteworthy comments regarding program risk in these areas
- Failing performance details
  - Driving factors (highest impact on negative performance)
  - Factors/levels where performance does not meet requirement
  - Recommendations
- Regression equation
  - The model coefficients convey the impact of each term on the response
  - Including a coded model allows relative impacts to be more easily observed. All factor levels are relative so actual units (e.g., meters, Hz, Watts) are not an issue.
  - This may be more effectively included electronically in the native DOE software or in Excel format for direct use by engineers

Repeat this listing for each response.

Conclusions

Employing DOE in DT imparts a unique iterative design generation/analytical process and requires the team to consider the application of sequential testing in a new context. The team must set analytical goals so the designs are correctly formulated and executed through the course of repetitive configuration changes. The use of DOE also facilitates clear and precise reporting of analytical results which effectively support the systems engineering process.

References

Box, George Edward Pelham., et al. *Statistics for Experimenters: Design, Innovation, and Discovery*. 2nd ed., Wiley, 2005.

Kutner, Michael H., et al. *Applied Linear Regression Models. 4th ed.*, McGraw-Hill, 2005.

Montgomery, Douglas C. *Design and Analysis of Experiments.* 9th ed., John Wiley & Sons, Inc., 2017.

Natoli, Cory. “Classical Designs: Fractional Factorial Designs”. Scientific Test and Analysis Techniques Center of Excellence (STAT COE), June 2018.

National Institute of Standards and Technology (NIST), *NIST/SEMATECH e-Handbook of Statistical Methods*, http://www.itl.nist.gov/div898/handbook/pri/section4/pri46.htm, accessed 13 December 2017.