The ITEA Journal of Test and Evaluation 2019; 40: 249-255

The Purpose of Mixed-effects Models in Test and Evaluation

John Haman, Ph.D.
Matthew Avery, Ph.D.
Heather Wojton, Ph.D.
Institute for Defense Analyses, Alexandria, Virginia

Mixed-effects models are the standard technique for analyzing data that exhibit some grouping structure. In defense testing, these models are useful because they allow us to account for correlations between observations, a feature common in many operational tests. In this article, we describe the advantages of modeling data from a mixed-effects perspective and discuss an R package—ciTools—that equips the user with easy methods for presenting results from this type of model.


John Haman, Ph.D.


Matthew Avery, Ph.D.


Heather Wojton, Ph.D.

Characteristics of Operational Test Data

By design, operational test data are often noisy. Scenarios with real operators, conducting operationally realistic missions against a responsive opposing force, generate data that reflect realistic combat environments and include operationally important sources of variance. The further we move from lab experiments, the more uncontrollable circumstances and conditions will influence the numbers our instruments and data collectors report. One strategy for dealing with noisy data is simply to collect more of it. As the sample size n increases, our uncertainty bars shrink and the risk to the program is reduced. But this is an inefficient use of taxpayer dollars and, because test budgets are limited, often infeasible. An alternative is to do more with the data we have.

Mixed-effects models, or simply mixed models, are a well-studied statistical technique used regularly across diverse fields of research, including pharmacology, agriculture, image analysis, and biology. Mixed-effects models are used in cases where researchers suspect their data contain systematically correlated errors. By properly accounting for these correlations, mixed-effects models produce estimates with smaller uncertainties.

A canonical example from agriculture is a comparison of crop yields from different seeds planted in multiple fields. Each field is divided into some number of plots, one type of seed is planted in each plot, and at the end of the test the yield of each plot is measured. Each field has unique characteristics, including exposure to sun, irrigation runoffs, and which plants are adjacent to the field. Since these characteristics affect crop yield, the attributes of each field introduce noise to the data. Because our goal is to determine how different types of seeds will perform in the future, the unique properties of the fields we used for our test are not relevant. By using a mixed-effects model, we can estimate the field-level variation separately from random plot-to-plot variation. This makes it easier to make comparisons, narrows the confidence bounds around our estimates of the average yield of each seed type, and

reduces the chance that we will confuse random variation due to a particularly good or poor field with changes in yield due to seed type.

While Defense Test and Evaluation (T&E) is quite different from agriculture, the attributes of the data can be surprisingly similar. Consider an evaluation of a radar-equipped unmanned aerial vehicle (UAV) operating in a maritime environment. The radar system provides maritime surveillance and intelligence by detecting targets at sea and reporting their locations to a host platform. The goal of the test is to compare the radar system's target location error (TLE, the straight-line distance between the true location of a target and the location provided by the radar system) against a requirement of 250 meters.[1]

The test plan calls for collecting six days' worth of data spread over one month. The primary factor driving radar system performance is the distance between the UAV and the target and, therefore, the test requires that performance be measured along a range of distances to the target. Distance is analogous to seed type in the above example. Environmental factors (the state of the water over which the aircraft is flying, atmospheric conditions, etc.) will also affect radar performance. These will vary from day to day, meaning that the TLE measurements will be correlated within each Day. Day is analogous to the field factor in the agricultural example; that is, both factors partition our data into groups.

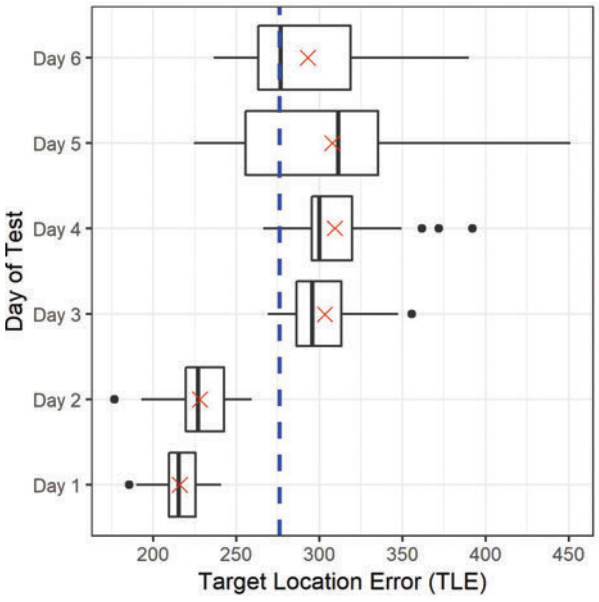

Figure 1: Variation in target location error is observed across the six days of testing. Average TLE for each day is represented by a red x. The global average is denoted by the blue dashed line

It is clear from Figure 1 that observations from single days are grouped. For example, the conditions on Days 1 and 2 seem to have provided the best set of environmental factors for the UAV's radar system. Ignoring the day-to-day variability could cause us to draw the wrong conclusions about the system's performance. When we analyze the results from this test, we must properly account for the within-day correlations or risk misidentifying important drivers of performance and reporting inflated uncertainty estimates.

Linear Models and Linear Mixed-effects Models

Linear models are commonly used in test and evaluation when testers want to determine the effect of different conditions on critical measures. Examples of linear models include linear regression and analysis of variance (ANOVA), both of which can be expressed using the following framework:

Let $\boldsymbol{y} = (y_1,\ldots,y_n)^T$ be a vector of responses; $\boldsymbol{X}_{nx(p+1)}$ a matrix containing the covariate information; $\boldsymbol{\beta} = (\beta_0,\ldots,\beta_p)^T$, a vector of model coefficients; and $\boldsymbol{\varepsilon} = (\varepsilon_1,\ldots,\varepsilon_n)^T$, a vector of random noise. The effects of the variables in $\boldsymbol{X}$ are modeled on the response $\boldsymbol{y}$ through the linear equation

$$\boldsymbol{y} = \underbrace{\boldsymbol{X\beta}}_{fixed} + \underbrace{\boldsymbol{\varepsilon}}_{random}.$$

Mixed-effects models posit that the data are "grouped" in some way such that observations in the same group are more similar than observations from different groups. This grouping causes within-group observations to be correlated, which is accounted for statistically by introducing "random effects" to the linear model. Random effects[2] account for the group-to-group variability that would otherwise be ignored by a classical linear model.

For mixed-effects models, we additionally let $\boldsymbol{Z}_{n\times q}$ be a design matrix that describes the group structure of the model and let $\boldsymbol{\gamma} = (\gamma_1,\ldots,\gamma_q)^T$ be the vector of random effects. The linear mixed-effects model relates $\boldsymbol{\gamma}$ to the covariates through the linear equation

$$[\boldsymbol{y}|\boldsymbol{\Gamma} = \boldsymbol{\gamma}] = \underbrace{\boldsymbol{X\beta}}_{fixed} + \underbrace{\boldsymbol{Z\gamma} + \boldsymbol{\varepsilon}}_{random}.$$

This model assumes that $\boldsymbol{\varepsilon} \sim MVN(\boldsymbol{0},\sigma_\varepsilon^2\boldsymbol{I})$, $\boldsymbol{\gamma} \sim MVN(\boldsymbol{0},\sigma_\gamma^2\boldsymbol{I})$, and that $\boldsymbol{\varepsilon}$ and $\boldsymbol{\gamma}$ are independent. *MVN* denotes the multivariate normal distribution. In many applications, the covariance matrices $\sigma_\varepsilon^2\boldsymbol{I}$ and $\sigma_\gamma^2\boldsymbol{I}$ need not be so simple, but they suffice here. The model may be expressed directly in terms of the distribution of $\boldsymbol{y}$ conditioned on the random effects, $\boldsymbol{\Gamma}$. Conditioning on $\boldsymbol{\Gamma}$ is statistical language that means looking at

the model from the group level ("Day" in the TLE example). When we look at the model from the group level, we may write it as

$$[\boldsymbol{y}|\boldsymbol{\Gamma}=\boldsymbol{\gamma}]\sim MVN(\boldsymbol{X\beta}+\boldsymbol{Z\gamma},\sigma_\varepsilon^2\boldsymbol{I}).$$

The model above describes a scenario in which we test on the six specific days, and the resulting data are contingent on the environmental conditions of those six days. (That is, $\boldsymbol{\Gamma}$ is a vector of unknown random quantities representing the effect of the environmental conditions on TLE, and $\boldsymbol{y}$ is the specific realized conditions experienced during testing.)

Alternatively, we may look at the model "in aggregate." Doing so, we do not condition on $\boldsymbol{\Gamma}$. This is called the population level – e.g., all possible days on which the test may have occurred. In this case, we write the model as

$$\boldsymbol{y}\sim MVN\left(\boldsymbol{X\beta},\sigma_\varepsilon^2\boldsymbol{I}+\sigma_\gamma^2\boldsymbol{ZZ}^T\right).$$

When we speak of the model at the population level, we are looking at the data or model without considering their grouping structure.

The structure of the matrix $\boldsymbol{Z}$ describes the dependence/grouping structure of the data. For instance, if

$$\boldsymbol{Z}_{n\times q}=\begin{bmatrix}1&0&&0&0\\ \vdots&\vdots&&\vdots&\vdots\\ 1&0&\cdots&0&0\\ 0&1&&0&0\\ \vdots&\vdots&&\vdots&\vdots\\ 0&1&&0&0\\ &\vdots&\ddots&\vdots&\\ 0&0&&0&1\\ \vdots&\vdots&\cdots&\vdots&\vdots\\ 0&0&&0&1\end{bmatrix},$$

the mixed-effects model is called a random intercept model, a different intercept term is supplied to each of the groups ("Days" in the UAV example). For added interpretability, we can write the random intercept model with summation notation,

$$[y_{ij}|\boldsymbol{\Gamma}=\gamma_i]=\boldsymbol{x}_{ij}^T\boldsymbol{\beta}+\gamma_i+\varepsilon_{ij}=$$
$$(\beta_0+\gamma_i)+\beta_1x_{ij}^1+\ldots+\beta_px_{ij}^p+\varepsilon_{ij}.$$

In this expression, y_{ij} is the response (e.g., TLE) measured on the j^{th} observation (e.g., a single target) in the i^{th} group (e.g., the third day of testing); $\boldsymbol{x}_{ij}^T$, the covariates of the j^{th} observation in the i^{th} group (e.g., the operating mode of the radar, the speed of the UAV, etc.); and γ_i is the random intercept of the i^{th} group (e.g., the difference between the average performance for the third day of testing and the overall average across all days of testing). This model supposes that the effects of the covariates are identical across groups and that each group differs from the population average response by a flat amount, γ_i.

In the random intercept model, the variance of y_{ij} may be expressed as

$$\text{var}(y_{ij})=\sigma_\varepsilon^2+\sigma_\gamma^2,$$

while $\text{cov}(y_{ij},y_{ik})=\sigma_\gamma^2$ and $\text{cov}(y_{ij},y_{kl})=0$. That is, observations in the same group (e.g., observations from the same day) are correlated with each other and observations from different groups (e.g., observations from different days) are independent.

The two variances σ_ε^2 and σ_γ^2 are commonly called variance components. Further, the correlation of two observations in the same group may be summarized by the intra-class correlation coefficient,

$$\text{ICC}(y_{ij},y_{ik})=\frac{\sigma_\gamma^2}{\sigma_\gamma^2+\sigma_\varepsilon^2},\quad j\neq k.$$

An ICC much greater than 0 indicates substantial group-to-group variability.

For a complete discussion of mixed-effects models, including their assumptions, algorithms, and analysis, the authors recommend Pinheiro and Bates (2006) and Gelman and Hill (2006).

UAV Example Analysis Using a Linear Mixed Mode

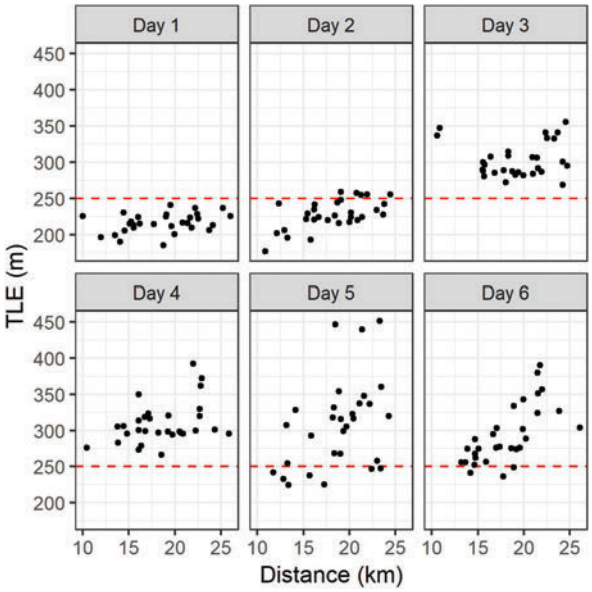

Figure 2: Scatterplot facetted by Day. Requirement denoted with a red dashed line

Returning to our UAV example, Figure 2 shows a clear picture of our data, including the day-to-day

variance of TLE. The main feature we notice is the variation between days (σ_γ^2) in TLE compared to the variance of observations within single days (σ_ε^2). We fit a linear mixed model to the data to quantify the amount of between-day variance and determine its importance.

Figure 3 shows the resultant mixed-effects model fit. The mixed model is considerably better than a simple linear model. The solid red line on the plot shows the mixed model fit to the data. On Days 3 – 6, environmental conditions were so poor that the average TLE never meets the requirement. On Days 1 and 2, the opposite occurs. Conditions on these days were optimal for the performance of the radar system. The high ICC (0.6) indicates that the majority of variation in TLE is attributable to day-to-day variation in environmental conditions. An ordinary linear regression that ignores the day-to-day variation would not identify this trend.

The slope of the solid red lines represents the effect that distance to target has on the performance of the UAV radar. The slope is positive because the UAV radar is less effective at farther distances from the target (as expected). This effect is statistically significant.

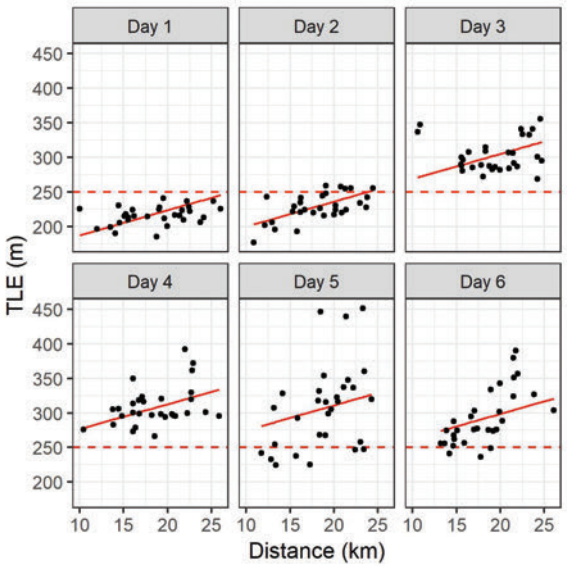

Figure 3: Scatterplot facetted by Day. Linear mixed model fit displayed as solid red lines

The Mixed-effects Model Advantage

Instead of using a random slope mixed model, we could also analyze these data using a simple linear regression, with Day as a fixed effect, or by fitting a series of separate models for each day of testing.

Day as a Fixed Effect

Treating Day as a fixed effect simplifies the analysis to a standard multiple linear regression model. This approach allows us to control the day-to-day variation, but there are drawbacks. The resulting model cannot make inferences that do not depend on a specific day, and predictions for data that may be observed on a new day or an unknown day cannot be made. Although the decision of whether to treat an effect as fixed or random can be difficult, day is more naturally treated as a random effect in this example (as discussed in endnote 1).

Mixed-effects Models Versus Separate Regressions for Each Day

Another alternative to the mixed model approach is to build a series of separate regressions for each day. However, this requires much more data to reach the same level of precision for a given effect size. Using separate models also precludes using information from one model to inform the next model. There is no way to share "statistical strength" between the separate models in this approach. Worst of all, you may get contradictory results in your separate models, with no way to reconcile them!

Comparing Methods

Figure 4 shows a comparison of three methods. The single linear regression model (black lines) doesn't fit any individual day particularly well. The separate regression models (blue lines) provide adequate fit for each group, but aren't consistent and, as a result, would be difficult to interpret and report on concisely. The mixed effects model (red lines) provides a compromise between these two, showing a consistent relationship between target range and TLE, while fitting each Day reasonably well.

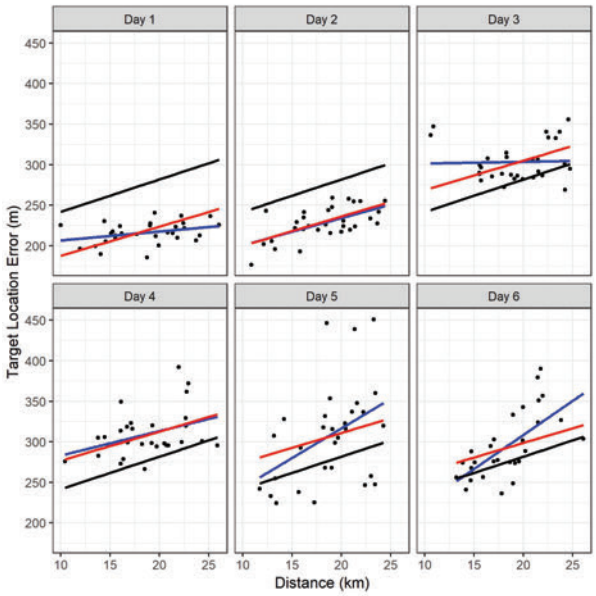

Figure 4: Comparison of random intercept model (red) with a single linear regression (black) and separate linear regressions for each group (blue)

Multiple Levels of Inference

Figure 5 compares the simple linear model with the linear mixed-effects model at the population level. The black line (simple linear regression) and the red line (mixed model) nearly overlap. This is because, at the population level, these two models produce similar estimates of mean TLE. The mixed model provides better estimates at the Day level while preserving performance at the population level, giving analysts the best of both worlds.

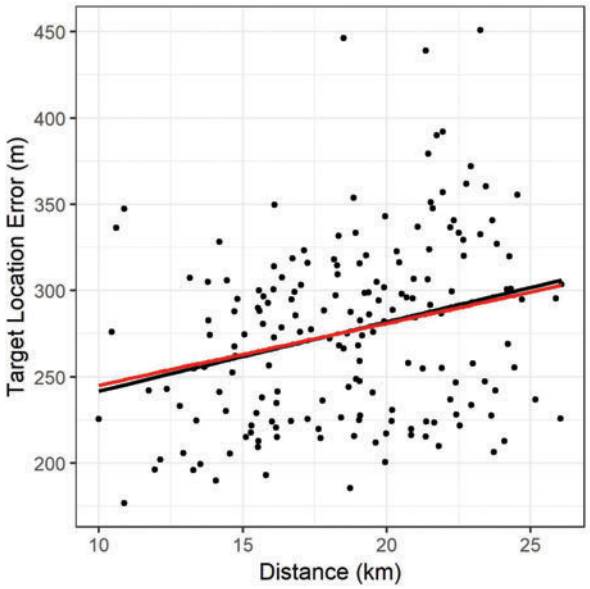

Figure 5: Linear model fit (black), population level linear mixed effects model fit (red) plotted against the full data set

Visualizing Uncertainty for Mixed-effects Models

Analysis based on mixed effects models must reflect the multiple sources of variation described by the model. Interval estimates are commonly used to convey prediction uncertainty, and they work well with mixed models. When constructing uncertainty intervals for mixed models, analysts must decide whether to build confidence intervals or prediction intervals and whether their intervals are conditional on the value of the random effect.

Confidence Intervals vs. Prediction Intervals

Confidence intervals reflect uncertainty estimates of model parameters. In our UAV example, they bound uncertainty for estimates of average TLE. Using the notation from the model described in section 2, this uncertainty is $var(\boldsymbol{x}_{ij}^T \hat{\boldsymbol{\beta}})$.

Prediction intervals show us the uncertainty in raw TLE (black dots in Figure 6). In addition to uncertainty about the model parameters used to estimate average performance, they also reflect the variation in TLE from one target to the next. In terms of the model presented in Section 2, this is σ_ε^2. Combining the uncertainty about the estimated average TLE and the uncertainty about future observations, the total uncertainty used for prediction intervals is $var(\boldsymbol{x}_{ij}^T \hat{\boldsymbol{\beta}}) + \hat{\sigma}_\varepsilon^2$.

Prediction intervals define the region into which we expect new observations to fall and confidence intervals define a region in which we believe the average value lies. There is more uncertainty in predicting new TLEs compared to average TLEs, so prediction intervals are wider.

Population-Level Confidence and Prediction Intervals

Mixed models have an additional source of variation (σ_γ^2) beyond the two discussed above. Depending on the system under test, the objectives of the analysis, and other factors, we may wish to include this source of variation in uncertainty bounds. Unconditional (or population-level) confidence and prediction intervals include this source of variation. For our UAV radar analysis, unconditional interval estimates describe radar performance regardless of Day, or on a new day when we are uncertain about what conditions will be like.

Figure 6 shows the estimated TLE as a function of range to target, along with unconditional confidence and prediction intervals. The confidence interval is the darker, narrower band. We can see both the uncertainty about average system performance as compared to the requirement and the uncertainty about individual TLEs.

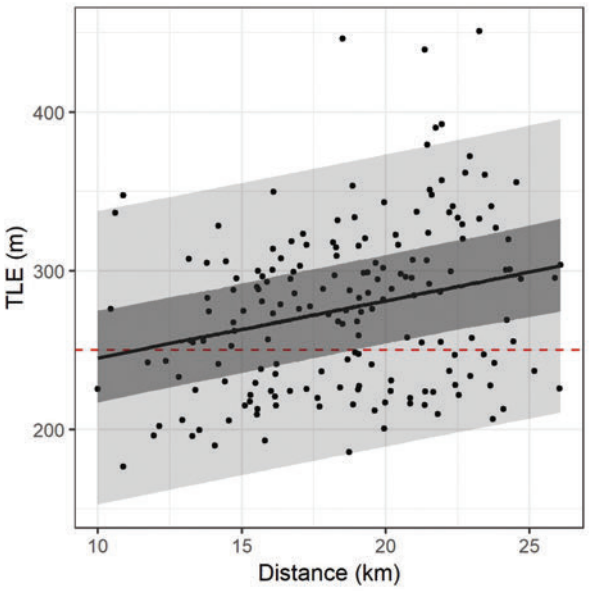

Figure 6: Population-level effect of distance on TLE (black line). 90% confidence interval (dark gray) and 90% prediction interval (light gray)

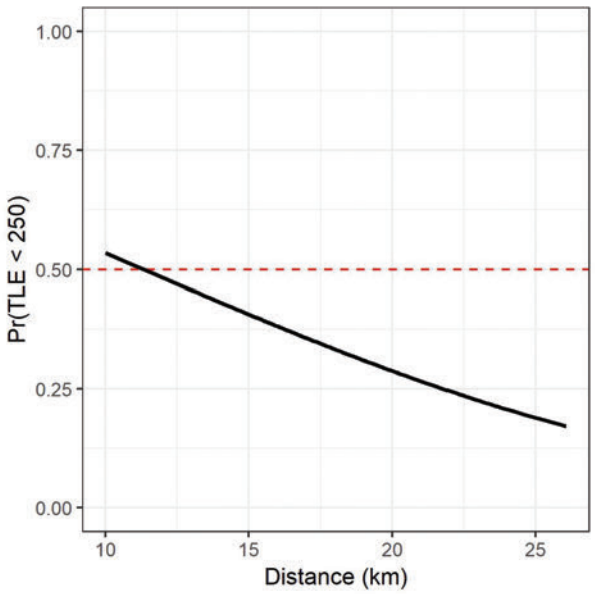

Figure 7: Estimated unconditional probability that TLE is less than the requirement of 250 meters

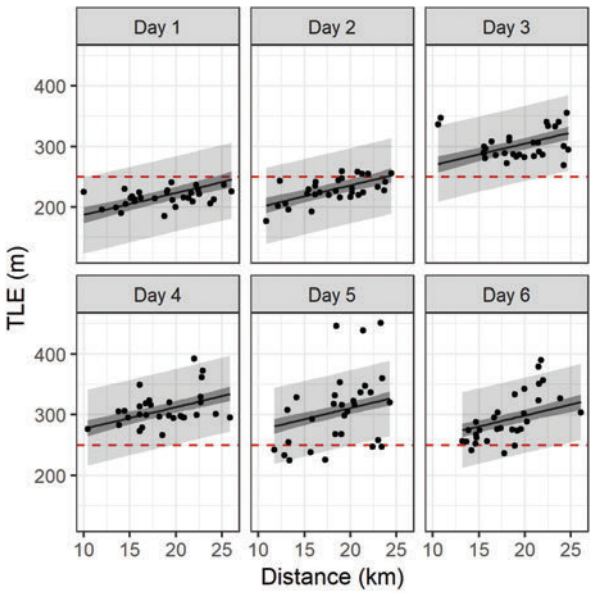

Figure 8: Group-level model fits for the first six days of testing (black lines). 90% confidence intervals (dark gray) and 90% prediction intervals (light gray)

An alternative way to visualize the probability that individual observations meet the threshold is shown in Figure 7. Rather than show an interval into which we expect new observations to fall, we can plot the probability that a new observation has a TLE below the threshold requirement. From this plot, we can see that to have at least a 50% chance that the TLE for a new target is less than 250 meters requires that the UAV closes to nearly 10 km of the target.

Group-level Confidence and Prediction Intervals

Alternatively, we may want to illustrate group-to-group (or in our case, day-to-day) variation in system performance. One way to do this is to show estimated results for each observed group. Since these results are now contingent upon the particular value of the random term observed, they are known as conditional results. Figure 8 shows conditional (group-level) confidence and prediction intervals for each of the six days of testing. Note that both sets of uncertainty bounds are narrower than the unconditional estimates shown in Figure 6. This is because the random effect is now shown as the differences from one day to the next rather than as additional interval width.

These results show that the average TLE of Days 3 – 6 is above 250 meters but below 250 meters for Days 1 and 2. Similarly, the prediction intervals show that on days with poor conditions, very few TLEs will be lower than the required value.

Figure 9 estimates these probabilities explicitly, showing the estimated probability that a new observation will meet the requirement under conditions equivalent to each of the six days of testing. This plot illustrates the large effect that environmental conditions have on system performance, emphasizing that on some days this radar system will perform very well while on others it may not meet the needs of the warfighter. Contrast this with Figure 7, which shows the same inconsistent performance but with an emphasis on the overall variation rather than the day-to-day variation.

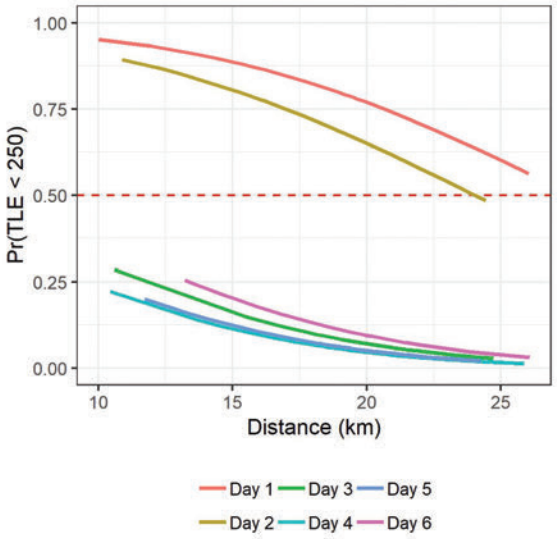

Figure 9: Probability that TLE will be less than the requirement of 250 meters

Implementation

Fitting mixed models requires software tools with more sophistication than is required for simple linear regression. To estimate the additional variance terms, fitting mixed models often requires an algorithm known as Restricted Maximum Likelihood (Bartlett (1937), Patterson and Thompson (1971)). While there are many popular software packages capable of fitting mixed models, the authors recommend R (R Core Team, (2019)). Among other advantages, R is free to use and has excellent visualization tools, such as ggplot2, which was used to build the graphics in this article.

In R, mixed-effects models are usually fit with the packages lme4 (Bates et al., 2015) or nlme (Pinheiro, (2019)). Both packages are trustworthy and yield correct, reliable model inferences. The lme4 documentation (https://cran.r-project.org/web/packages/lme4/index.html) provides a comparison of these two packages.

Neither of these packages provides easy methods for uncertainty estimation, however. For that, we recommend ciTools (Haman and Avery, 2019). The uncertainty estimates shown above were calculated with ciTools, which provides a simple and consistent interface for analysts to use for prediction inference. ciTools works well with packages such as ggplot2 to help analysts craft data displays that accurately reflect model estimates and uncertainties around those estimates. ciTools is compatible with mixed effects models, simple linear models, generalized linear models, and reliability models, all of which are commonly seen in test and evaluation. Visit https://cran.r-project.org/web/packages/ciTools/vignettes/ciTools-demo.html for more details. ❑

JOHN HAMAN, Ph.D. is a statistician at the Institute for Defense Analyses. He is a member of the Test Science team and is interested in statistical methodologies for operational evaluation. Currently, John is supporting electronic warfare systems for the Navy and Air Force.

MATTHEW AVERY, Ph.D. is a statistician at the Institute for Defense Analyses, where he develops methods and tools for analyzing test data. He has worked with a variety of Army, Marine Corps, and Navy systems, including unmanned aerial vehicles (UAVs) and ground vehicles.

HEATHER WOJTON, Ph.D. is a Research Staff Member at the Institute for Defense Analyses. She provides expertise in the evaluation of human-system interactions. She currently aids in the test and evaluation of a broad range of major defense systems, including both training and operational aircraft, and information systems.

References

Bates D., M. Maechler, B. Bolker, and S. Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67 (1): 1-48.

Bartlett, M. S. 1937. "Properties of Sufficiency and Statistical Tests." Proceedings of the Royal Society of London. *Series A-Mathematical and Physical Sciences 160* (901): 268-282.

Gelman A. and J. Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge University Press.

Haman J. and M. Avery. 2019. "ciTools: Confidence or Prediction Intervals, Quantiles, and Probabilities for Statistical Models." R package version 0.5.1. Accessed July 25, 2019, https://CRAN.R-project.org/package=ciTools.

Patterson H. D. and R. Thompson. 1971. "Recovery of Inter-Block Information When Block Sizes are Unequal." *Biometrika* 58 (3): 545-554.

Pinheiro J. and D. Bates. 2006. *Mixed-effects Models in S and S-PLUS.* Springer Science & Business Media.

Pinheiro J., D. Bates, S. DebRoy, D. Sarkar, and R Core Team. 2019. "nlme: Linear and Nonlinear Mixed Effects Models." R package version 3.1-139, Accessed July 25, 2019, https://CRAN.R-project.org/package=nlme.

R Core Team. 2019. "R: A Language and Environment for Statistical Computing." R Foundation for Statistical Computing, Vienna, Austria. Accessed July 25, 2019, https://www.R-project.org/.

Endnotes

[1] This requirement is notional and used only for illustrative purposes in this document. The requirement has no relation to an actual system under test.

[2] The independent variables in typical regression models are referred to as "fixed effects", in contrast to "random effects", which appear in mixed models. Factors or independent variables are treated as "fixed" if (1) the individual levels observed are of interest beyond the specific experiment they were observed in, (2) the experiment aims to quantify differences in the response for specific levels of that factor, and (3) the researcher would choose the same levels in a future experiment. Effects should be treated as random if (1) the individual levels observed during test were incidental, (2) these levels would not be repeated if the test were replicated, and (3) the primary outcome of analyzing these variables is to quantify the variation in the response attributable to them. In our UAV example, range to target would be a fixed effect since researchers would always want to observe "Near" and "Distant" targets and understand the difference in average TLE for near targets as compared to that for distant targets. Day is treated as Random because the specific Days of test are not of interest. A new test would be conducted on different Days, and the goal of incorporating Day into the analysis is to understand the variability in TLE attributable to Day rather than compare TLE on specific Days.