The VCS C-Seam Analytics Platform: A Monte Carlo Approach to Machine Learning

Andrew L. Turscak III

Newport News Shipbuilding

Newport News, VA USA

Andrew.L.Turscak@hii-nns.com

ABSTRACT

A submarine hull join is the process of fitting and welding two large hull sections of a submarine into a single unit. Until now, the ability to predict dimensional shifts from initial cylinder fit-up through the final weld relied on the domain expertise of experienced shipbuilders. With an infinitesimal tolerance for each control point, the care put into these determinations was tedious and time-consuming. To improve efficiency, a predictive algorithm was built to learn optimal at-set coordinates from the behavior of prior joins. The algorithm took historical data from a series of Virginia Class Submarine joins, and used regression and statistical analysis to model control point shifts at each stage. Predictions were generated for the shifts in a Monte Carlo Simulation, and an Excel Solver was used to choose the at-set coordinates that minimized the probability of out-of-tolerance joins. The VCS C-Seam Analytics Platform (VCAP), written in R, employs graph theory and advanced statistical distribution fitting rules to fully automate this algorithm. The mathematics are hidden behind an intuitive user interface, passing the burden of modeling the data to the tool instead of the engineer. The platform greatly reduces the man hours needed to prepare for each join, enables speedy enhancements to the model's predictive power, and births an algorithm that progressively gets smarter from both new data and the user's ability to quickly tune its parameters.

ABOUT THE AUTHOR

Mr. Andrew Turscak is a Data Scientist in the department of Modeling & Simulation at Newport News Shipbuilding. Mr. Turscak obtained a BA in Economics at the College of William & Mary before going on to earn his Master's degree in Computational Operations Research from the same institution. He has worked with data in disciplines including manufacturing operations, human resources, telecommunications, quantitative finance, and experimental economics. The bulk of his current work consists of a series of machine learning and analytics projects initiated as part of an advanced analytics effort at Newport News Shipbuilding.

The VCS C-Seam Analytics Platform: A Monte Carlo Approach to Machine Learning

Andrew L. Turscak III

Newport News Shipbuilding

Newport News, VA USA

Andrew.L.Turscak@hii-nns.com

INTRODUCTION

A hull join is the process of fitting and welding two hull sections of a submarine into a single unit. Until now, Newport News Shipbuilding relied on the domain expertise of experienced shipbuilders to choose at-set control point coordinates such that weld distortion caused no out-of-tolerance (OOT) joins. Recently, however, these shipbuilders expressed a desire to be able to decisively predict dimensional shifts from initial cylinder fit-up through the final weld so that a deterministic method for picking these coordinates, given a desired end-state, could be established. With a very small tolerance threshold for each control point, the procedure of calculating the at-set coordinates had to be extremely meticulous. The costs of rework and schedule delays due to OOT joins vary greatly with the stage of the weld, making an approximation of these expenses difficult, and planning for them even more-so. This necessitates the prediction of at-set coordinates that minimize the probability of OOT conditions at all future stages in the process. By improving the choice of at-set conditions, the shipbuilders could redirect their focus to other aspects of the join process that demand similar precision and effort.

In a pilot project to address this need, a predictive algorithm was built to learn optimal at-set coordinates from the behavior of prior joins. The algorithm took historical shift data from 12 control points across a series of Virginia Class Submarine (VCS) hull joins, and performed a pairwise correlation test on the shift values. The historical values for the pairs of coordinates found to be significantly correlated were each fit to a simple linear regression model, and the remainder coordinates to statistical distributions of their historical behavior. Predictions from these models were generated for each of the shifts in a Monte Carlo Simulation modeling the cumulative shifts at each discrete stage of the weld and repeated for a large number of iterations. An Excel Solver was used to choose the at-set coordinates that minimized the simulated occurrences of OOT joins. The predictive accuracy of the model was noteworthy, but the manual execution of such complex analyses necessitated exceptional familiarity with the problem and a strong statistical background. Additionally, the model required the use of Minitab for the model fitting and Excel for the Monte Carlo Simulation and solver. These factors, along with the tedious reconstruction of the tool for each join, proved costly in terms of man hours (MH), averaging 90 MH per join. Over multiple joins, the compound effect of these costs was excessive. It was therefore desirable that the tool's predictive accuracy be captured in an easily repeatable, less computationally demanding way.

The Virginia Class Submarine C-Seam Analytics Platform (VCAP) is a fully automated program used to minimize man hours (MH) spent calculating the at-set control points for c-seam welds. Manually developed in R, the tool employs graph theory and a coded set of statistical distribution fitting rules to automate this algorithm and minimize man hours spent determining set point coordinates. Because the tool's purpose is to be independently deployed and managed by shipbuilders, the mathematics are hidden behind an intuitive user interface, passing the burden of modeling the data to the tool instead of the engineer. Additionally, R's vectorization capabilities were harnessed to speed up computational bottlenecks. With a runtime of less than one minute, the tool eliminates the 90 man hours that were needed to run the original algorithm. This reduction in runtime also enables quicker integration and evaluation of enhancements to the model's predictive capability, allowing the algorithm to get smarter not only from the data, but the user's ability to quickly adjust its parameters. This project's success not only reduces man hours spent performing costly operations, but serves as an inaugural paradigm for how pilot projects for predictive analytics should be deployed in manufacturing.

C-Seam Join Background

A c-seam join is a circumferential weld procedure for joining two hull sections of a submarine into a single unit. The procedure begins with the forward hull section fixed firmly in place and the aft component resting on a plank structure designed to hold it steady. Factoring in precut conditions, tolerances, the control point measurements of the forward section, and expected weld draw, the aft hull section is shimmied to an offset designed to best position the hull to hit the target coordinates and fit to the forward section.

After the desired at-set position is achieved, a tack weld --a series of short, dotted welds performed around the circumference of the seam-- is executed to temporarily hold the units in place for subsequent procedures. Following tack, a full weld is completed on half of the circumference. Finally, the second half of the circumference is welded. The discrete increments of the procedure are, in part, designed to allow for new measurements between each stage. This enables out-of-tolerance conditions to be detected or anticipated before a full weld is completed, as cuts and re-welds at later stages are far more difficult and disruptive to schedule.

DATA & ASSUMPTIONS

Data

As is common in manufacturing, the data available to initially build this tool was limited in several ways. First, key explanatory features known to drive control point shifts were either not collected, difficult to access, or incapable of being mapped to events of interest. Second, only so many Virginia Class Submarines are constructed each year, whereas predictive models typically need a large sample size to generate accurate results. Finally, the data that was available had only been consistently stored for the most recent hull joins, resulting in data sets for approximately ten hulls, with slightly more or less depending on the join type. Ultimately, these limitations dictated the design of the final predictive model.

Data used for the model was collected from engineers tasked to continuously monitor control point shifts with laser sensors. The data included precut, design, and as-built measurements for 16 control points for each historical VCS hull join of each join type, and are shown in Figure 1.

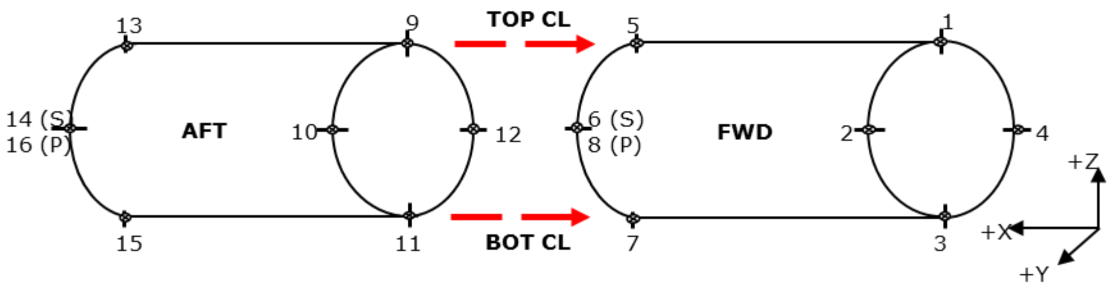

Figure 1. Dimensional control diagram for VCS Submarine.[1]

The measurements recorded in this data were taken at each stage of the join process to determine dimensional shifts that occurred between weld stages. Measurements were recorded at set, fit-up, 50% weld, and 100% weld. While not a large data set, the data for the control point shifts that was collected was verifiably accurate to one-thousandth of an inch. For this reason and the lack of other data of similar quality, the scope of the model was restricted to that data. Because the forward component of each hull is fixed in place, control point data used in the model included only measurements from

[1] Diagram obtained from NNS department O68 dimensional control report

the aft end of the hull (control points 9-16). The reported nautical directions of the shifts were transformed into Cartesian coordinates, with each shift measured in two dimensions.

Assumptions

Because data was restricted to control point shifts with no feature variables, the first assumption of this model is that shifts at certain control points are the best predictor of shifts at other control points. While it is known that factors such as cylinder shape, plate thickness, and sequence of the weld factor considerably into the direction of movement, of the data available, the initial control point movements themselves best captured this information. The inability to map events to individual shifts required the second assumption that environmental factors and weld procedures are constant across joins. While we know this not to be true, the weather, times between weld stages, and weld directions were not recorded consistently enough to incorporate these factors into our model.

Next, we assume that control point shifts at each stage of the weld process occur independent of prior and subsequent stages. For this reason, the regression modeling discussed in the next section is restricted to fitting historical control point shifts at discrete stages corresponding to each weld procedure. Finally, because the tool must be fully automated, we assume that the p-values associated with the statistical tests in the following section constitute sufficient evidence of correlation and distribution fit.

MODEL

The automated algorithm begins with a pairwise correlation test between control point shifts. R was coded to select the pairs of control points for which the test results were most significant (lowest p-value) for regression analysis. A linear regression model was fitted to each of the coordinate pairs for which the p-value fell below a user-defined threshold (defaulted to 0.025). A statistical distribution from a predefined set was fit to the historical values of those coordinates with no significantly correlated pairing. To support automation, this was done by setting the algorithm to choose the distribution with the smallest p-value in a Kolmogorov-Smirnoff test. Due to the small data set, there was often no good fit distribution. In these instances, no distribution was fit, and the discrete set of historical values was treated as the distribution for sampling purposes. Each of the following steps occur in individual iterations of a Monte Carlo Simulation.

To automatically order the regressions in the simulation, graph theory techniques were implemented to replicate the manual process of generating regression predictions. An arborescence was used to model the control points that were paired for regression, with nodes as control points, and directed edges modeling a relationship from independent to dependent variable. Figure 2 depicts this type of graph for one iteration of one stage of the weld process.

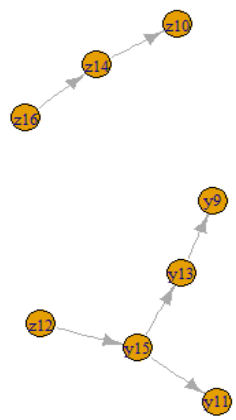

Figure 2. Arborescence ordering shift predictions from at-set to fit-up.

A control point shift was identified as a root node if, in the set of fitted regression relationships, that control point served only as an independent variable. A shift was identified as a leaf node if, in the same set of relationships, that control point existed as only a dependent variable. A random variable was generated for each root node of each subgraph, then a series of regression predictions were spawned from there, with the root node as the predictor, then each predicted variable becoming the predictor for the next regression prediction until the leaves of the tree were reached. Predictions for both those control point shifts with no significant relationship to other shifts and the residuals for the regressions were randomly sampled from a best-fit distribution or set of historical values. These predictions were generated for each control point at each stage of the weld. This process would be repeated for a user-defined N iterations in a Monte Carlo Simulation.

To optimize the at-set coordinates in a manner similar to Excel's gradient solver, an approximation algorithm was constructed to repeatedly run the simulation and adjust multiple combinations of the at-set control points until a terminating condition was reached. The chosen measure of optimality was the minimization of the Total Squared Error of the predictions. The process began with the solver executing 100 runs (default) of the Monte Carlo Simulation. Then, each at-set control point was adjusted incrementally by some step-size (defaulted to positive or negative 0.001), and the Total Squared Error of the predictions was calculated after each adjustment until no further improvements could be achieved. The simulation would then be run again with the improved at-set control points as the initial set, and this process repeated indefinitely until termination. Termination occurs when either the estimated accuracy of the predictions reaches a user-defined acceptable value (error converges to an acceptable level), or a pre-specified run time limit is reached, depending on which occurs first. The algorithm then returns the recommended at-set coordinates to achieve optimal welds, as well as the probability that an out-of-tolerance join occurs for any given control point.

PRODUCT

Because a goal of this project was to salvage man hours spent performing analysis without sacrificing mathematical complexity, a graphical user interface (GUI) was developed to run the automated algorithm in a user-friendly application designed for shipbuilders with limited mathematical backgrounds.

VCAP's user interface includes a main *Input* tab, as well as password-locked *Advanced* and *Validation* tabs. The home *Input* tab —designed for Foremen and Supervisors in the Module Outfitting Facility (MOF)— assumes no statistical background, and allows domain experts to focus solely on entering the target control points. It is depicted in Figure 3.

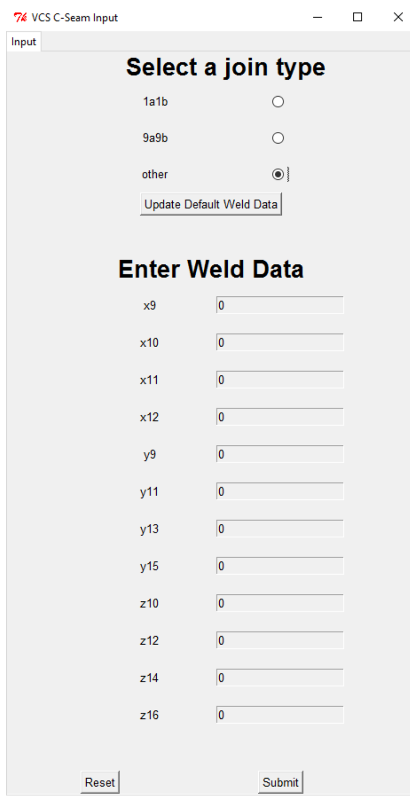

Figure 3. Screenshot of Input Tab.

This tab also allows users to default to the design conditions for specific join types to save time on data entry. When the *Submit* button is hit, a file browser enables users to select the historical data required by the model. Finally, the model is run, and another file browser allows users to name and store the optimized set-point output in an Excel file. The entire process takes roughly one to five minutes (or about one to two hours when cleaning new data).

An *Advanced* tab —designed for Modeling and Simulation Engineers with only a high-level understanding of the predictive algorithm— hands the user full command of the model's most important parameters while assuming no more than an elementary statistics background.

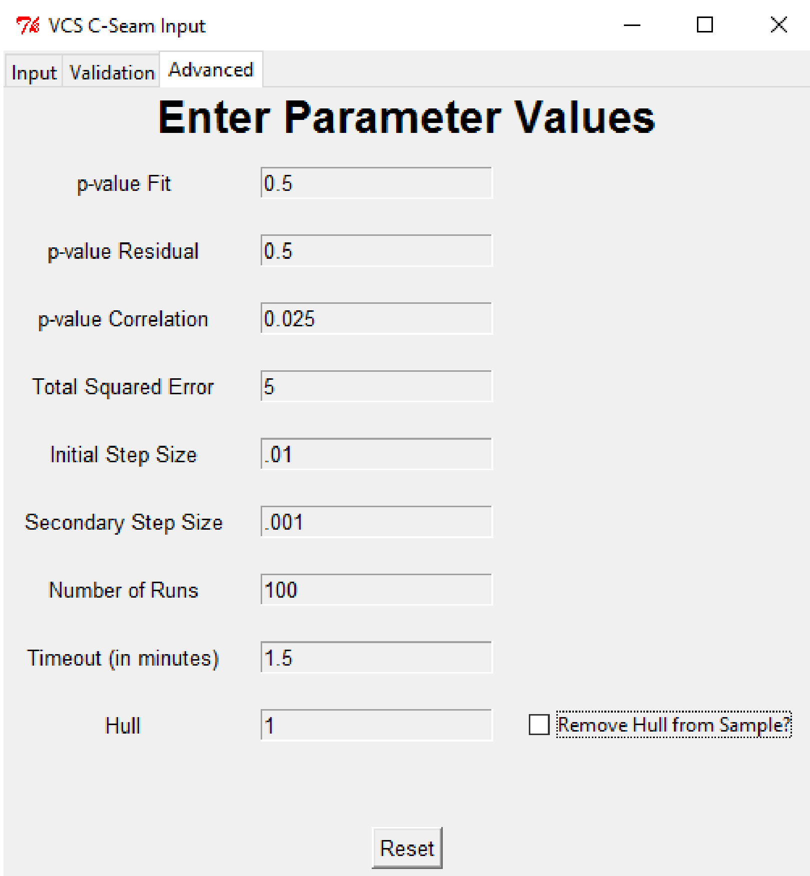

Figure 4. Advanced tab (with defaults displayed), designed for engineers to adjust model parameters.

The purpose was to isolate the most influential model parameters for easy manipulation without having to parse through 1500 lines of R code to change them.

Table 1 summarizes the functions of each of these parameters as they pertain to the simulation.

Table 1. Advanced tab parameter definitions.

Parameter	Purpose
p-value Fit	Significance threshold for accepting or rejecting a control point shift's fit to a distribution (higher is more strict)
p-value Residual	Significance threshold for accepting or rejecting an error term's fit to a distribution (higher is more strict)
p-value Correlation	Threshold for defining two control points as significantly correlated (lower is more strict)
Total Squared Error	Max tolerated total error of predictions
Initial Step Size	Size of incremental adjustments to at-set control points during the first Monte Carlo run
Secondary Step Size	Size of incremental adjustments to at-set control points during later Monte Carlo runs
Number of Runs	Number of iterations for each Monte Carlo run

Timeout (in minutes)	Cutoff time if simulation hasn't reached the acceptable Total Squared Error level
Hull	Indices of hulls to remove from the historical data

The three *p-value* options define the basis on which values are either predicted from a random variable, a linear regression, randomly sampled from historical values, or (in the case of residuals) sampled from a theoretical distribution. *Total Squared Error*, both *Step Size* options, and *Number of Runs* influence the precision with which the model accepts a solution (and therefore the model's runtime). The *Timeout* option sets a maximum runtime, should the model run slowly. The *Hull* option (and corresponding checkbox) allows the user to index hulls that they would like removed from the sample, either to perform validation testing on a single hull, or because those hulls exhibit outlier behavior detrimental to the model's accuracy.

The *Validation Tab* —designed to evaluate the model's accuracy— is visually identical to the *Input Tab*, but prompts the user for set point coordinates instead of target coordinates. The purpose is to simulate what would happen should the user set a hull at those coordinates. This feature enables a user to remove a hull from sample in the advanced tab, and then enter that hull's historical set point data into the boxes. The resulting 100% Weld simulated coordinates can be compared to the real 100% Weld coordinates to evaluate the accuracy of the simulation for that hull.

All of these features put the power of predictive analytics into the hands of the average user and maximize model flexibility to handle a range of join types. This eliminates the need for Modeling and Simulation Engineers to fit a custom model for each new join and data point. The GUI ensures that man hour savings are compounded indefinitely for every hull join that will be performed at NNS.

RESULTS & VALIDATION

Predictions for shifts in the *x* and *z* directions proved easier to predict than shifts in the *y* direction, due to the difficulty of mitigating gravity's effect on different joins. Table 2 shows the percentage of simulation iterations that resulted in out-of-tolerance conditions when using the optimal at-set recommendations for each control point in a simulation for one section join.

Table 2. Out-of-tolerance simulated probabilities.

Control Point (CP) Shift	% of Simulations Resulting in OOT Conditions
CP 9 (x-direction)	5%
CP 10 (x-direction)	1%
CP 11 (x-direction)	2%
CP 12 (x-direction)	2%
CP 9 (y-direction)	28%
CP 11 (y-direction)	25%
CP 13 (y-direction)	35%
CP 15 (y-direction)	32%
CP 10 (z-direction)	8%
CP 12 (z-direction)	12%
CP 14 (z-direction)	11%
CP 16 (z-direction)	6%

Each of the probabilities in Table 2 represent the chance that, should a shipbuilder take the model's recommendation for the optimized set, an out-of-tolerance condition will occur for that control point.

To validate the predictive accuracy of the simulation, individual hulls were pulled out of sample, and predictions generated for their end-state using the hull of interest's actual at-set values. The predictive value of the model would be based on the closeness of the predicted final weld values to the values recorded at 100% weld, and whether the predictions were within tolerance. Figure 5 shows the mean absolute deviation from real values for each control point prediction across all hulls to get a picture of the model's overall effectiveness.

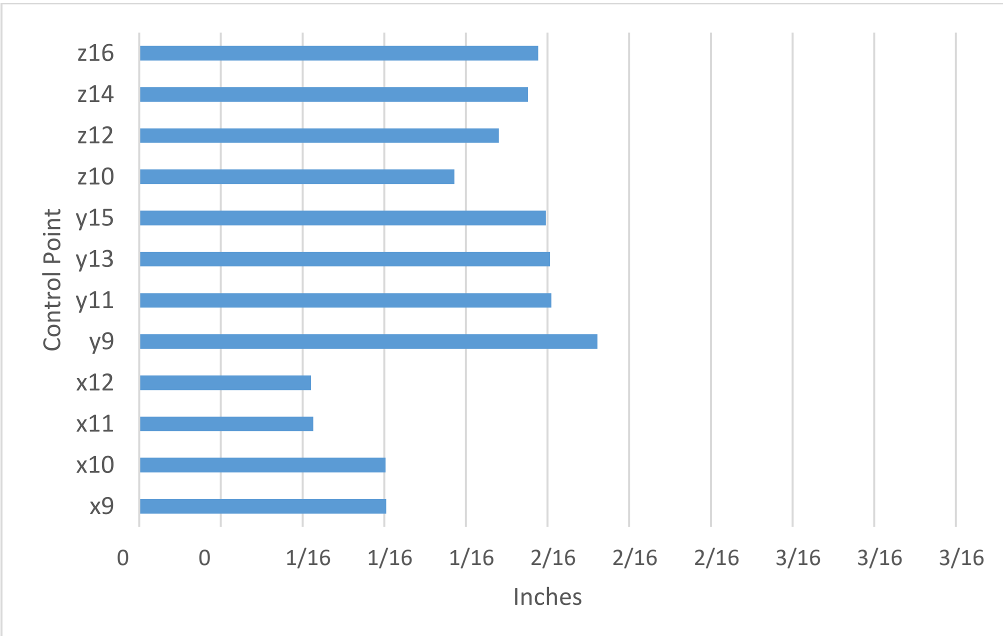

Figure 5. Mean absolute prediction error across all historic hull joins.

It is seen that, across a diverse array of join types, the model is almost always accurate to 2/16". Figure 6 shows the model's error for each control point prediction for the most recent hull join conducted at NNS.

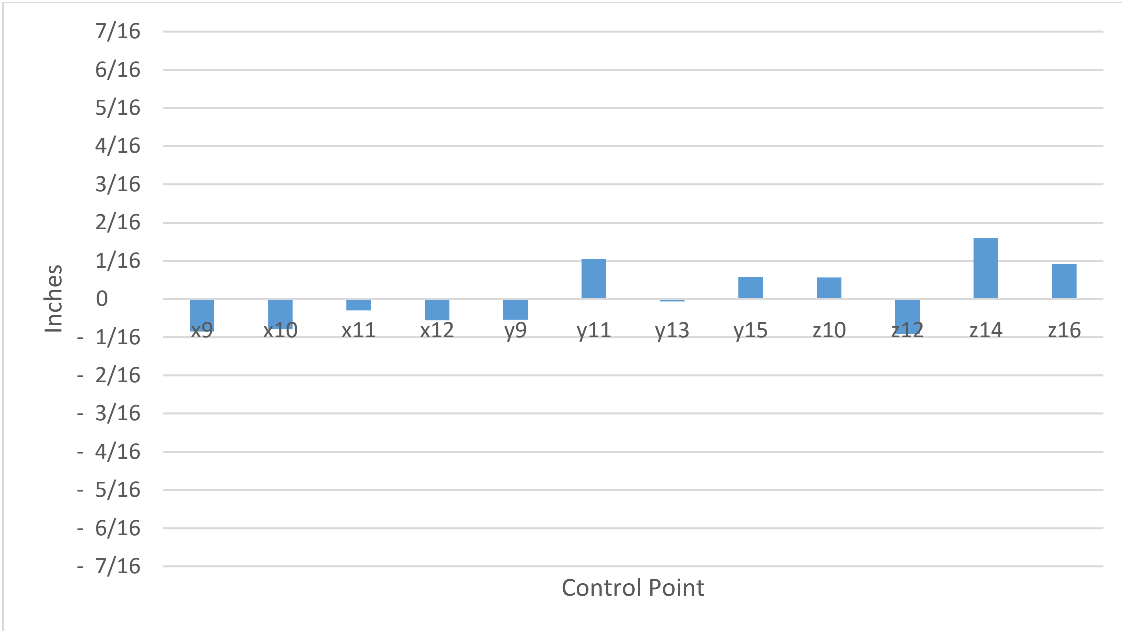

Figure 6. Prediction error (individual hull).

Figure 6 shows that the model is accurate to 1/16" for all but two control points, and that even their error is within 3/16", the tolerance threshold for this particular hull.

CONCLUSION & FUTURE WORK

The VCAP algorithm and platform provides extensive value to NNS for the increased efficiency that it adds to manufacturing operations. It also serves as a unique and creative mathematical approach to modeling a problem with limited data. Most of all, VCAP captures the predictive accuracy of a complicated and intensive algorithm without soliciting the expertise and lengthy hours required to execute it. This reduction in man hours also enables quicker integration and evaluation of enhancements to the model's predictive capability, allowing it to get smarter not only from the data, but the user's ability to quickly adjust its parameters.

While VCAP was limited to only control point shift analysis, the flexibility of linear regression to include dummy variables and other enhancements enables this tool to factor in other important features should the data become available. VCAP ensures that these savings are compounded indefinitely for every hull join that will be performed at NNS. This project's success not only saves man hours from costly operations, but serves as an inaugural paradigm for how predictive analytics should be deployed throughout the yard to achieve operational excellence and establish NNS as a leader in the deployment of smart technologies. Future work should include the collection and incorporation of known drivers of weld distortion, and the application of VCAP to other types of weld joins.

ACKNOWLEDGEMENTS

The pilot predictive algorithm on which VCAP is based was conceived by Andrew Hinton, and developed by him and Justin Wolfe. This tool would not have been possible without their intensive research, domain expertise, and Justin's dedicated mentorship in getting me acquainted with the model. Additionally, the GUI and user functionality were skillfully implemented with the swift and generous efforts of Stefani Werner. Finally, the data collection and deployment of VCAP were facilitated by engineers in O68 and the supportive team of shipbuilders in the MOF.