VOID FILL ACCURACY MEASUREMENT AND PREDICTION USING LINEAR REGRESSION

J. Harlan Yates, Mark Rahmes, Patrick Kelley, Jay Hackett
Harris Corporation
Government Communications Systems Division
Melbourne, Florida 32904
hyates@harris.com; mrahmes@harris.com; pkelley@harris.com; jhackett@harris.com

ABSTRACT

We present an innovative way to predict accuracy and associated error in void fill of digital surface elevation models. An answer to this question is desired: "How well does the filled data correspond with the truth values?" Typically, however, void filling is performed *because* this information is not known. It is often impractical due to time and cost to acquire truth data. This issue is typically ignored and treated as a best effort based only on visual appeal. However, by using statistical analysis, the behavior of the error given a particular output can be learned. An algorithm can learn from cases where some truth is available so that the behavior of the error can be predicted for cases where no truth data is available. For this process to succeed, each sample (void) must have a description that can be fed into a prediction model. This comes in the form of metrics computed from the void points and the surrounding non-void neighborhood. Any number of characteristic metrics can be computed for possible use in the final description collection, but hypothesis testing can then be used to determine a smaller subset of these metrics for use. We propose a methodology that predicts error produced by linear regression during the process of terrain void filling. We apply our approach to the filling of voids present in SRTM, IFSAR, LiDAR, or Seismic data. Also, while this methodology is independent of any specific algorithm, this discussion is centered on a non-image based void fill algorithm.

KEYWORDS: PDE, Exemplar Inpainting, Linear Regression, Error Prediction

VOID FILLING METHOD

Besides aesthetic appeal, the other key issue in any void fill method used for geospatial products is accuracy. How well does the filled data correspond with the truth values? Typically, however, the void filling is being processed because this information is not known. It is often impractical, from a time and cost standpoint, to acquire truth data for the missing areas (Kelley, 2008).

This issue is typically ignored, treated as a sort of best effort based on visual appeal only. There is, however, another option. By using statistical analysis, the behavior of the error is learned for a given particular output. Said another way, an algorithm can learn from cases where some form of truth is available so that in application cases the behavior of the error can be predicted where no truth data are available (Kelley, 2008).

For this process to work, each sample (void) must have a description that can be fed into the prediction model. This description comes in the form of various metrics computed on the void points and the surrounding non-void neighborhood. At this point, any number of characteristic metrics can be computed as possible additions to the final description collection (Kelley, 2008).

The purpose of this discussion is to present a methodology to predict error produced during the process of terrain void filling. Also, while this methodology is independent of any specific void fill algorithm, this discussion is centered on non-image based void fill algorithm. If statistically based error prediction can be combined with a state-of-the-art non-image based void filling process, this could provide a step towards being able to cost-effectively fill voids.

Selection of Sample Data Set Voids

Creating a statistical model for error prediction requires training on cases representative of those that will be presented during elevation model production. Here, this implies utilizing a data set that represents the full population of voids. To achieve this, the full population of voids in all of the data is sampled to create a representative subset. This sample subset is used by the system in a machine learning process to determine the parameters required based on the assumptions of the model. Independent data samples are set apart from this training process to be used in later validation of the overall process.

To attempt to predict error in a filling process for the population, the system must have the ability to determine the error present for the representative sample subset. This limits sampling for training (and later validation) purposes to geographical areas where some form of truth data are known.

Prior to metrics analysis, each cell is run through two pre-processing steps. The first step eliminates data 'islands,' which are fully surrounded by void neighbors. The second step selectively eliminates noisy void boundary posts. These two steps attempt to eliminate areas in the data that have been shown to be generally unreliable in data collection. Elimination of less reliable data near voids prior to filling (and prior to calculating metrics related to error estimation of filling) should improve the overall quality of the results of both the fill and the error estimation parameters (Kelley, 2008).

Non-Image Based Void Fill

After pre-processing, the voids comprising the sample set must be filled. The Exemplary inpainting method with Poisson Merging has the critical advantage of maintaining textures (which can often be noisy and complex) better than any other fill algorithm found in the literature (Yates, 2009). This algorithm requires significantly more processing time than other fill methods. However, this portion of the process is fully automated and is therefore acceptable in many cases, especially where manual filling would otherwise be required. The process as described up to this point is shown below in Figure 1.

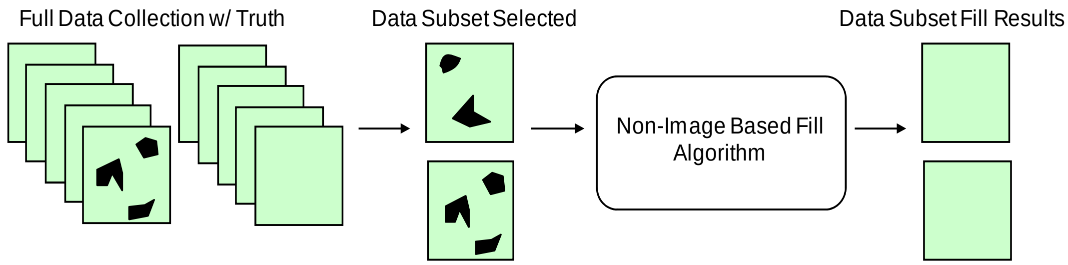


Figure 1. Void Fill of Sample Data Sets

Error Analysis

With the sample set filled, the error can be calculated with respect to the truth data. For each void (sample), an error value is calculated and recorded. The standard root mean squared error between the truth data (cells with stereo image fill for example, edited to meet specification), and the results of the candidate fill algorithm is calculated. This is computed on a void by void basis, resulting in a unique error for each void. This compilation forms an error report, as shown in Figure 2.

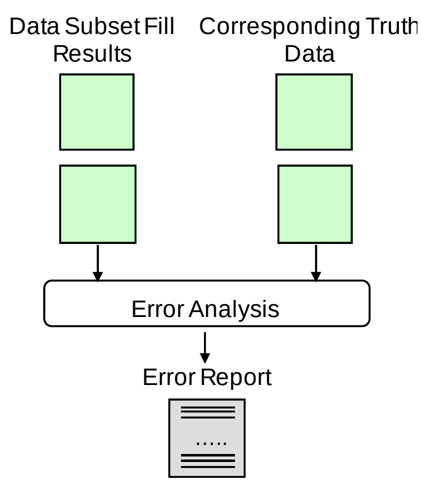


Figure 2. Error Analysis

Designing and Applying Description Metrics

To properly predict error for new samples, one must thoroughly examine the characteristics of the voids. An important task is to define a good set of description metrics that can be used for this purpose.

Initially, a full suite of metrics has been defined and implemented--approximately fifty in all. It should be noted that possible redundancy in these metrics can be tolerated at this point and will be addressed in the next stage of the process. The goal is for these metrics (or actually just a small subset of them) to predictively model the variation in the difference between the fill results and the truth data (i.e., the error).

This full collection of metrics is calculated for each void in the sample set (the selected collection of samples with truth data). The output, as shown below in Figure 3, is a report detailing these calculated metric values on a void by void basis.

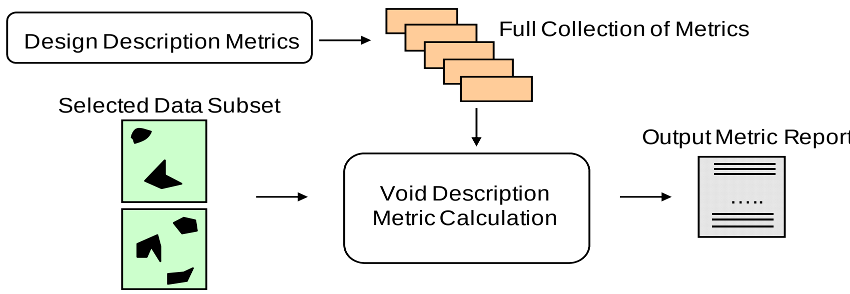


Figure 3. Design and Calculation of Metrics

Determination of Metrics Subset

With such a large collection of metrics available there is a very good chance that many of them may not be needed for prediction. Typically with a group of this size, there will be elements that contain redundant information, which is indicated by a high degree of mutual correlation. The other types of metrics that can be eliminated from the set are those that are found to not contain enough relevant information for the task of prediction (i.e., those whose values tell nothing about the expected output which, in this case, is error).

To guard against metrics in the former category--those with redundant information--correlation analysis can be used. Metrics are grouped with others where overlap may occur. To determine the amount of redundancy, metrics can be tested against each other to determine their correlation. Those that are highly correlated should remain in a group together. Those that are not may be safely separated. Only the metric that best correlates with the error is taken from each group.

A statistical test is needed to determine which metrics fall into the second category mentioned above--those that do not contribute to the prediction--to prevent them from being placed into the final metrics subset. To accomplish this, hypothesis testing may be performed, as detailed below.

A linear regression is run with the full subset of candidate metrics to define the prediction model that would result if this were the final metrics subset. From this regression, each metric receives a coefficient value that describes the relationship between it and the value being predicted (in this case, error). Next, a hypothesis is determined that will allow the elimination metrics. More specifically, for each metric it is hypothesized that its coefficient value is actually zero (i.e., no relationship exists, and it is therefore useless in prediction). This allows use of the sample set of data to test the likelihood that the calculated coefficient could have been attained if the hypothesis were true. If the metric corresponding to the coefficient in question is significant, there will be a low probability that the calculated coefficient value would have occurred by chance. Following a standard statistical threshold for any metric with a value below 5%, the hypothesis is rejected and the metric is considered significant for prediction. In this fashion, metrics are removed one at a time. This iterative removal is necessary because the interactions between metrics change as they are added and removed. This process is illustrated in Figure 4 (Friedman, 2009).

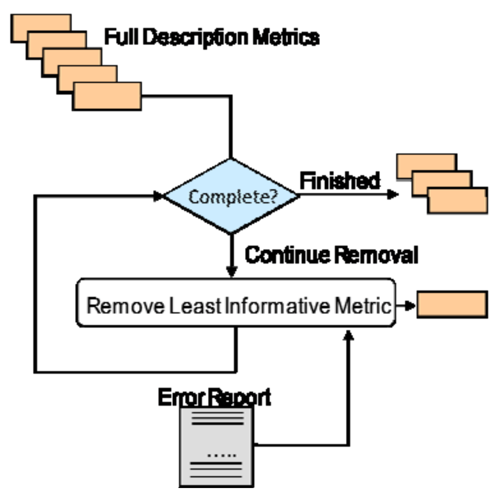


Figure 4. Reduction of Metrics Set

Creation of Training and Validation Sets

The prediction model should be independently validated on a set of data that is separate from trained system data in order to properly assess the model's ability to predict. To accomplish this, the selected sample data set with corresponding truth data are split into a training set and a validation set with no overlap between them. Each sample (void) is examined from the selected cells with all required information being recorded. These recorded entries are the actual values that are split and recorded in the separate sets. This process is shown in Figure 5 below.

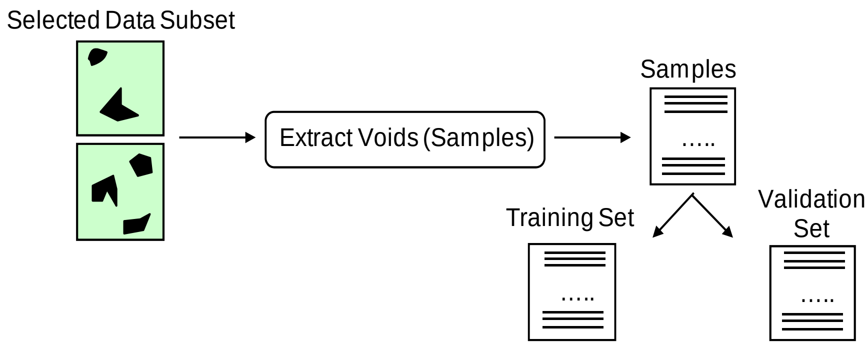


Figure 5. Training and Validation Sets

A good representation of the collection of voids must be presented to the statistical prediction model. This requires that there be a sufficient amount of sample voids with some form of corresponding truth data available. This set of truth data are split into two parts: a training set and a validation set. A common mistake, at this point, is to train the system and then test how well you did (validation) on the same set of samples. It is easy to see that this is an unfair assessment of performance and will not represent the ability to predict when truth data are not available (i.e., the interesting cases) (Kelley, 2008).

This training dataset can now be used to find the optimal (or near optimal) values for the parameters that make up the statistical prediction system, a process known as training the system. Figure 6 below displays shows this process visually (Kelley, 2008) (Friedman, 2009).

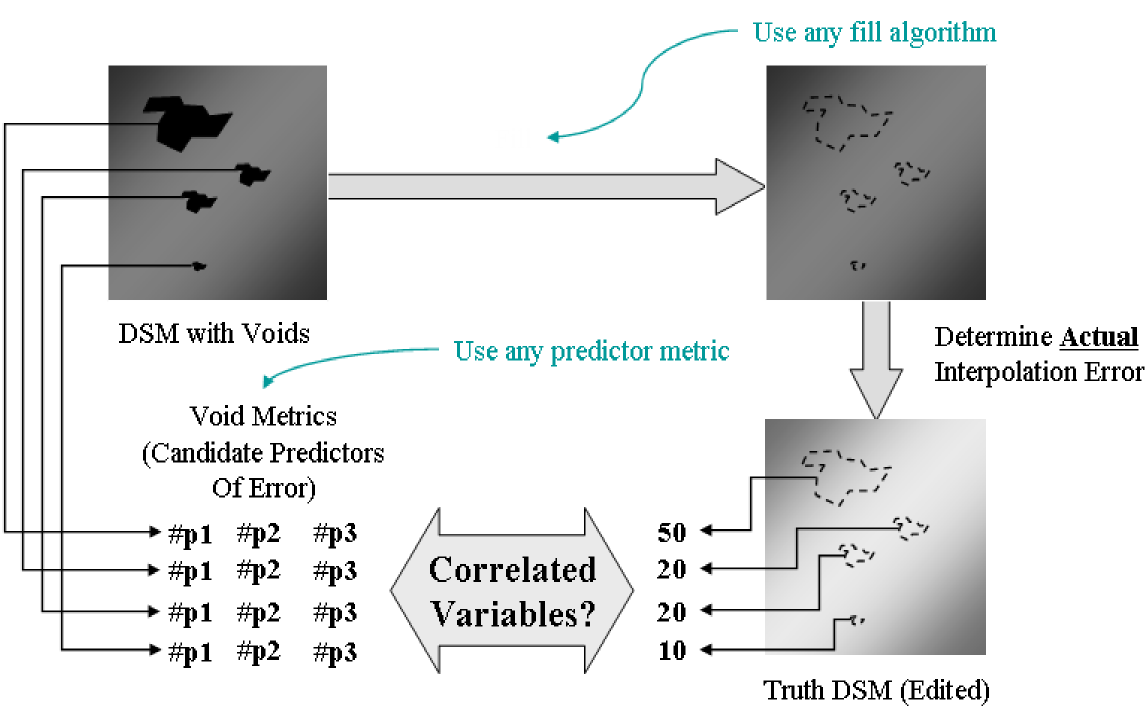


Figure 6. Statistical Model Training

Statistical Prediction Model

Using the training set described above, a form of linear regression analysis is performed to fit the prediction model to the population of voids. To do this, the set of the best description metrics available (those left after the iterative removal process previously discussed) are optimally combined to create a linear prediction equation. The performance of this method during validation (described later) determines if this statistical method can properly describe the current problem domain or if another method will be required (Friedman, 2009).

The goal of this step, regardless of the form of the prediction equation used, is to determine the coefficients of each parameter present. There is one parameter each per selected metrics in the linear model, and these coefficients model the relationship between the metrics and the error. These optimal coefficients minimize a certain residual between the predicted fill error and the actual fill error for the given training set in the given model. With the system shown in Figure 7 in place, error can now be predicted for any new void sample for which the same selected metric values can be calculated.

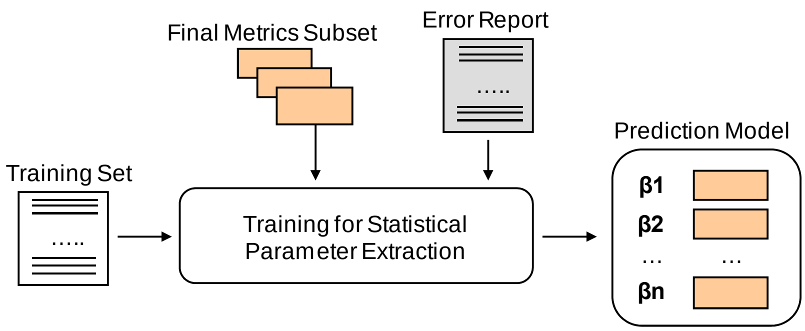


Figure 7. Prediction Model Parameters

Model Validation. Validation provides a means to assess the performance of the newly created model before applying it to the full application data set that has no corresponding truth data. As previously mentioned, this process is performed on a separate validation set that is not included during the training process.

To validate the model, each sample in the validation set is evaluated using the prediction model to produce a predicted error for each void. Along with this, confidence intervals can be calculated to further enhance the information contained in this calculation. The difference between the predicted value and the true error for the sample gives the error residual. The better the prediction model, the lower these values will be. This process is depicted in Figure 8.

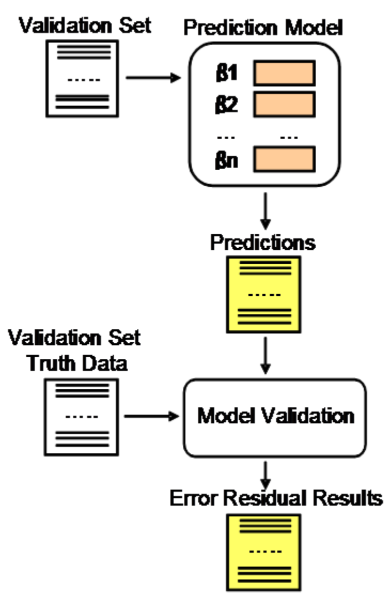


Figure 8. Prediction Model Validation

Model Prediction. If these residual values indicate that the model does not sufficiently fit the true behavior of the current problem domain, something must be changed. In the worst case, the results from validation may lead to a redesign of the whole process. Other model changes may be less severe. Once an acceptable model is generated and validated, it is ready for use in prediction of fill results for the full application data set, as shown in Figure 9.

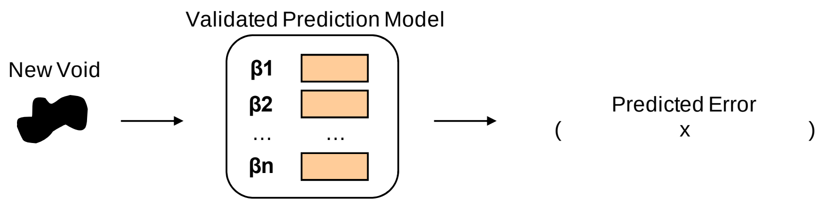


Figure 9. Application of Validated Prediction Model

At this point, the validation dataset is used to test the performance of the current model. Since truth data are available, an analysis can be made to determine how closely the model successfully predicted the true error of the fill. This can determine whether or not the current model will suffice for the current problem domain. Common problems found at this stage are that assumptions made in model type selection are invalid or that the model was made too specific (e.g., over-trained) to the training data (Kelley, 2008).

Finally, with the model trained and validated, the error that will occur in filling new samples without truth data can be predicted. Typically, these types of predictions should include some form of confidence information along with the expected value. Figure 10 shows this stage (Kelley, 2008) (Smith 2008).

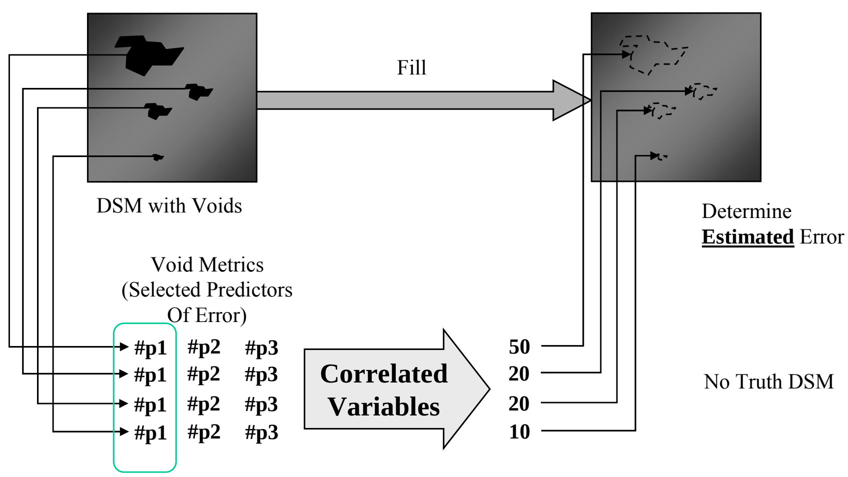


Figure 10. Error Prediction with no Truth Data

Fill or No Fill Criteria. One of the key issues is determining whether or not to fill a given void that has no truth data available. One option available is a simple cutoff at a user defined threshold above which any expected error would become unacceptable, and the corresponding void would not be filled. This method has the advantage of being straightforward and increases the total amount of voids that will be given the chance to be filled. However, the major deficiency with this method is that it does not take the uncertainty of the prediction into account. For example, a prediction model that cannot properly explain the relationship between the input description metrics and output error value will still predict an error. However, it will have a wide range of other possible output error values that are also very likely (just not as likely as the predicted value). These other likely values could range from close to, to extremely far from, the predicted value. If another model were found that somehow perfectly modeled the relationship between the input metrics and the output error, it would be able to predict the expected error without any other possible outputs. The simple cutoff approach, while it does have the advantages mentioned above, would treat these two models exactly the same, ignoring the variability in their ability to predict.

There is also another well-defined and commonly accepted statistical method that can be used for making the "fill, no fill" decision. When a new void is evaluated using the prediction model, an expected error value is provided in the same way as in the case of the simple cutoff. Using statistics on the amount of variability in the output that is not explained in the prediction model, which is unknown but can be estimated from the sample, the distribution of possible cases as described by the model can be formed around this expected value. Figure 11 illustrates this with an example where the expected value is shown as the peak of the distribution of possible error values.

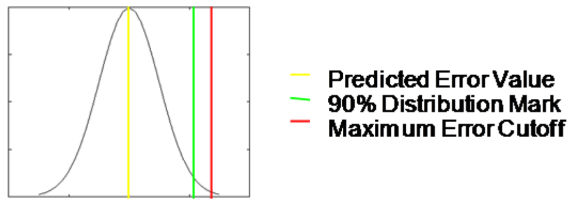


Figure 11. Estimated Error with Associated Confidence Estimate

At this point, the decision of whether or not to fill could be made by providing the same user defined error tolerance; moreover, this decision could now be accompanied by the percent confidence mark required for the application. Using this, the fill would only proceed if the total percentage of possible outcomes predicted by the model in the distribution was at or below the desired cutoff. In the example above, a 90% confidence is required (the green line). This implies that 90% of the population of possible values must reside at or below the user defined cutoff (the red line). This is true in the case shown.

As should now be clear, this is a much more conservative decision criterion. It equates to saying that in the population the training data describes, *at least* 90% of the time, a void with the given expected error will be below the desired cutoff. This method has the advantage of noticeably decreasing the number of cases where the decision is to fill when the true error is above the cutoff (termed the optimistic or worst-case scenario). This method also has the unfortunate side effect of significantly increasing the amount of cases where the fill should have been made, but was not (termed the pessimistic scenario).

VOID METRICS

Figure 12 below is a detailed description of the selected metrics set. These metrics and their corresponding beta values resulted from applying the described methodology on the selected set of training data and candidate metrics.

The selection of these metrics, from the original list of approximately fifty considered, is based on the statistics of the void data. The statistical analysis process not only selected this set of predictors, but also determined the number of predictors to use (i.e. the system is not asked for the top seven predictors, but rather for an optimal set under the current method) based on iterative analysis of the training results.

The selected metrics are based on the following:

1. Average void diameter
2. Average elevation and slope differences on opposite sides of the void
3. Average elevation and slope differences on adjacent void posts
4. Maximum elevation difference on void boundary for opposite and adjacent posts

These metrics and their corresponding beta values resulted from applying the described methodology to the selected set of training data and candidate metrics. The predictive model that incorporates these metrics is a multiple regression that has the following form (Friedman, 2009):

$$\text{Predicted Error} = \beta1*X1 + \beta2*X2 + \ldots + \beta7*X7 + \text{Error Intercept} \qquad (1)$$

The values of β (and the error intercept) are the model coefficients obtained from the training and validation data sets, and the values of X are the corresponding metrics. These are the subset of metrics selected from an original collection of approximately fifty metrics (not specifically discussed in this paper).

The selection of these metrics is based on the statistics of the void data. The statistical analysis process not only selected this set of predictors, but also determined the number of predictors to use based on the iterative analysis of the training results.

X1 = Grid Avg Distance : For each post adjacent to a void (boundary post), accumulate distance across the void in each primary direction. Divide by total number of such distances. Example: current boundary post is blue dot. The red dots are the points used to calculate distance across void as shown in Figure 12 a.

X2 = Grid Stat : Similar to previous predictor, compute average distance and average height difference between points across void. Sum the two averages.

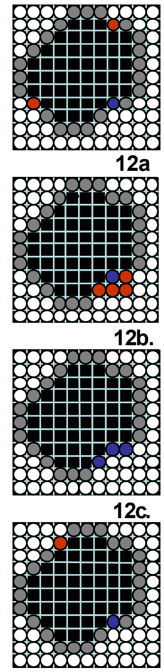
X3 = Avg Dir Range : For each void boundary post, maintain a Max and Min neighbor elevation difference for each direction (8 possible directions, direction 0 = north). Compute average Max[dir] – Min[dir] of the pairs of stored values. Example: blue post has 4 (red) neighbors (direction 2, 3, 4, and 5). Compute height difference between blue post, and each neighbor, saving the value if a new maximum or minimum value is found. After processing all boundary posts, average the difference in the pairs of Max and Min values as shown in Figure 12 b.

X4 = Slope Change %: Order the boundary posts in a contiguous clockwise fashion around the void. Compute heig ht change between consecutive posts. When consecutive height changes differ by more than 50% of grid spacing, increment a counter. After processing all height change pairs, divide by number of boundary posts as shown in Figure 12 c.

X5 = Avg Opposite Slope : Order the 'n' boundary posts in a contiguous fashion. For first n/2 points, accumulate slope between point(i) and point(i+n/2). Divide by n/2 for average. Example: the blue and red posts are 'opposite' points as shown in Figure 12d.

X6 = Max Opposite De lta: Order the 'n' boundary posts in a contiguous fashion. For first n/2 points, find maximum height difference between point(i) and point(i+n/2).

X7 = Elevation Extent: Find the highest and lowest boundary post. Subtract lowest from highest to get the elevation extent.



12a

12b.

12c.

12d.

Figure 12. Selected Metrics Set

SUMMARY

An important accomplishment of the prediction portion is achieving this quality of prediction results without the added difficulty of using different prediction models for each terrain type. The methodology described above is successful in training on a data set that attempted to describe the full population of samples (voids) without having to apply a different set of metrics to each terrain type.

The second component for success in this method, separate from the ability to predict error, is the ability to fill each void in a way that is both accurate and aesthetically pleasing. Exemplar inpainting with PDE-based seamless region merging technique for filling data has two distinct advantages over other algorithms. The first is that it has the ability to not only take the local neighborhood near the void into consideration, but also adapt to the entire input area provided to it. This drives down error. This type of methodology also has the ability to maintain natural textures. This ability overcomes the shortcoming of not only the other autonomous techniques, but also manual ones. Textures that would be very time consuming, or in some cases impossible to create or replicate using manual processes, can be achieved as a natural result of the exemplar method. Incorporation of these surface texture effects in the void fill process reduces perceptual artifacts of the fill data with respect to the surrounding data areas making the filled areas nearly indistinguishable from the output alone. With textures replicated, the PDE-based seamless region merging component of the algorithm addresses the other major cause of artifacts, mismatches in structure on the boundary and throughout the area of interest.

REFERENCES

Friedman, Jerome; "The Elements of Statistical Learning: Data Mining, Inference, and Prediction", Second Edition (Springer Series in Statistics)**,** 2009.

Kelley, P., Yates, J. H., Rahmes, M. D., Allen, J., "Seamless Exemplar Inpainting Method Using Poisson Merging and Normalization for Terrain Void Filling", *ASPRS,* May 2008.

Smith, Kelley, Yates, Rahmes, Allen, Berisford, "Exemplar /PDE-Based Technique to Fill Null Regions and Corresponding Accuracy Assessment", US Patent 7,881,913, 2008.

Yates, J. Harlan; Kelley, Patrick; Allen, Josef; Rahmes, Mark; "Advanced Terrain Processing: Analytical Results of Filling Voids in Remotely Sensed Data", *ASPRS*, Nov 2009.