Strategic Node Placement using Gaussian Process

# Abstract

Gaussian Processes (GP) are frequently used to simulate environmental phenomena, such as weather, transportation of air pollution and temperature gradients. In this study, a GP was used to estimate snow depth in a square kilometer region in the Sequoia National Forest. Snow depth in this region is measured via two methods. The first is 23 sensor nodes that measure snow depth via an ultrasonic sensor 23 points in the study region. The second is via LIDAR data, which offers snow depth measurements of a resolution of up to 1 square meter. For this study, the LIDAR data was considered to be ground truth, because of its high resolution. A computational approach was used to find the optimal bandwidth parameter and covariance of a GP regression model to provide snow depth. The results of this model were then compared with the LIDAR data to discover where the locations were of greatest error. The node locations were then adjusted minimize the error between the LIDAR data, and that of the GP.

# Background and Motivation

The ulterior motive for this study is to evaluate the reduction in error by strategically adding sensors where the error in the model is the greatest. Better information about the environment leads to better knowledge of what natural resources are available. In this case, knowledge of snow depth is an important indicator of how much water is available for consumption. However, it is impossible to sense everywhere, so a balance must be found between the number of sensors, and the error reduction. This study examines placing new sensor nodes, and observing the resulting error reduction. The number of sensors when the error reduction plateaus will be considered the optimal number of sensors.

The following data was available for this study:

* A square kilometer of LIDAR data (dated 2010)
* Snow values (also from LIDAR data) at each of the node locations in the square kilometer

As in any study, it is useful to examine the assumptions made in the analysis. It was assumed that:

* there was additive noise in the data. This is generally a safe assumption since LIDAR detects other objects such as trees.
* the LIDAR data is ground truth.
* the values from the sensor nodes are the exact same values as the LIDAR data at those same locations.
* it is possible to place nodes at any location within the square kilometer. Of course in reality this is not always true because locations of rocks, bodies of water, cliffs, etc., affect where it is possible to place sensor nodes.

While not an assumption per se, it should be noted that during the addition of sensor nodes, the original 23 locations were not changed. This is because moving sensor nodes is labor-intensive, and it is more practical to add sensor nodes than it is to change their existing locations.

# Gaussian Process Model Construction (K-folds)

In a Gaussian Process regression model, a multivariate, isotropic (“multivariate” because 2-dimensional) Gaussian distribution is generated to model the spatial variation of a certain variable. From the multivariate Gaussian, a predictive distribution of the snow values is generated by two parameters - predictive mean, and predictive variance.

There are two hyper-parameters that may be varied. Since snow-depth is spatially-varying, a bandwidth (h) is used to quantify the effect of distance on the predicted values. The known values from the sensor node locations were split into training and testing data, iteratively. By iterating over a range of bandwidth values, and measuring the resulting error, the k-folds cross validation error was minimized to determine the optimal bandwidth. The same procedure was used for finding the optimal covariance (sigma squared). The optimal bandwidth was discovered to be 135 meters, and the optimal covariance was discovered to be 0.05. The Figures below validate these optimal values by plotting each parameter on the x axis, and the error on the y axis.

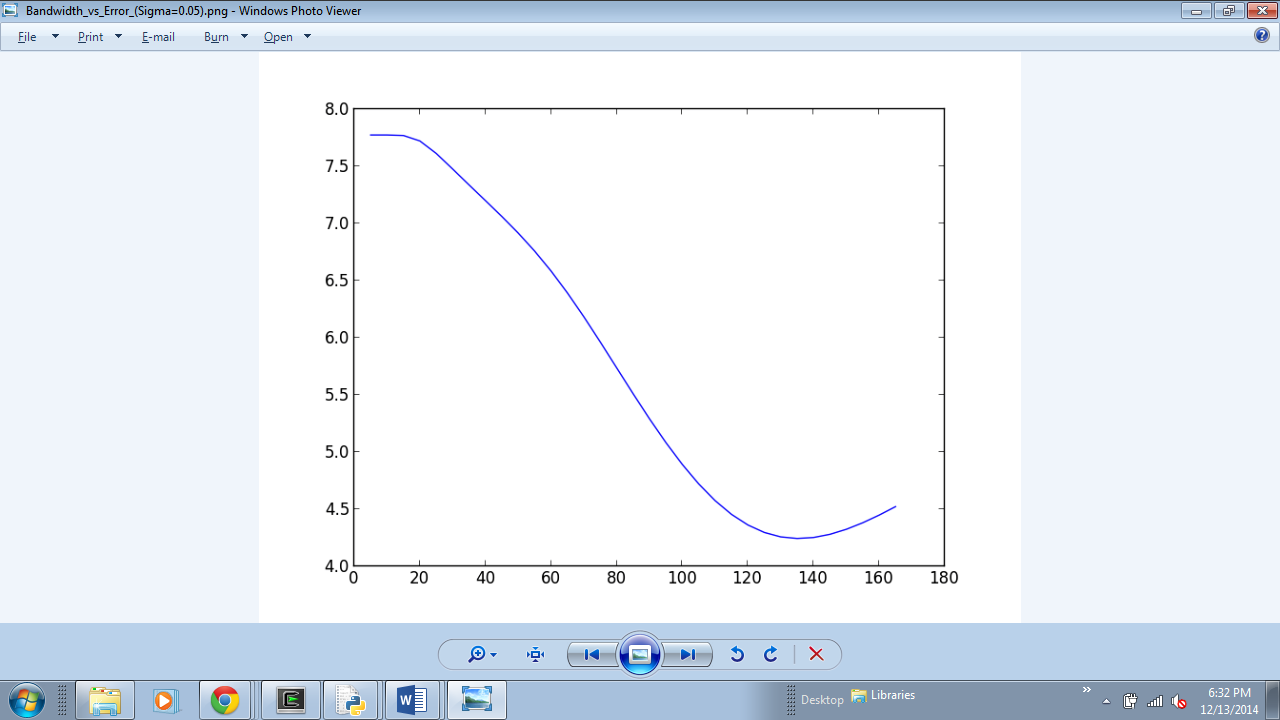
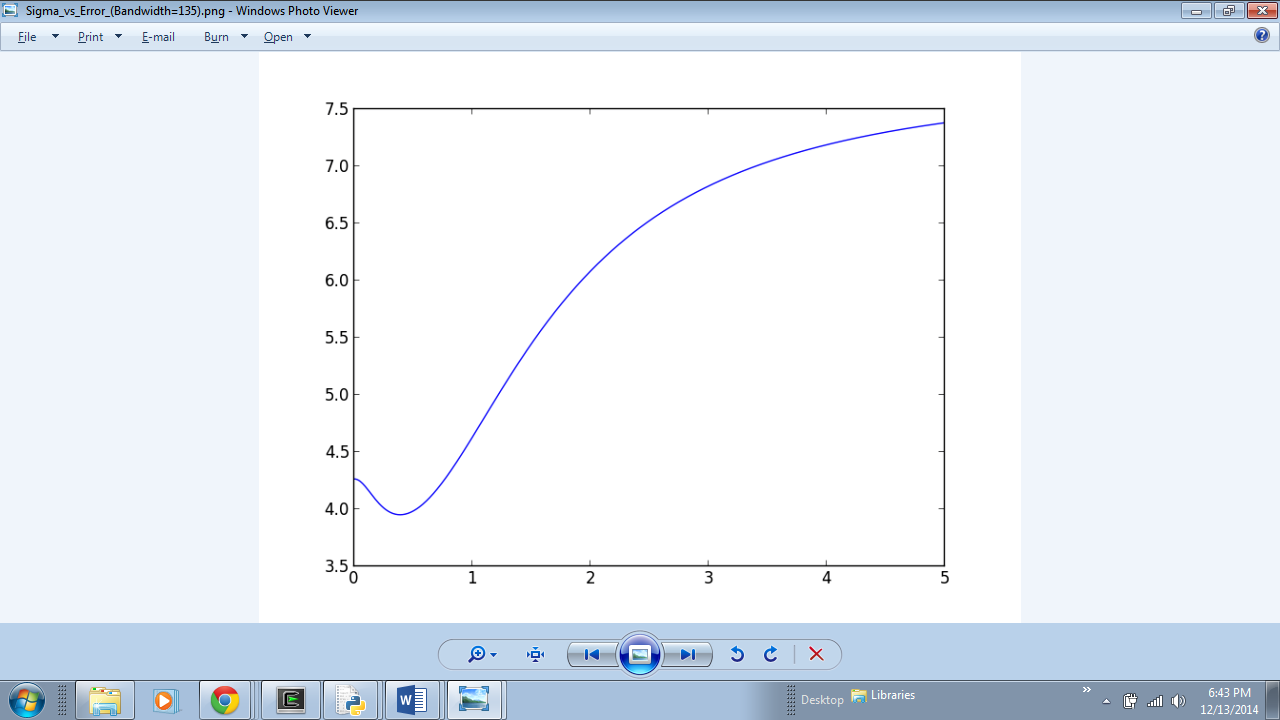
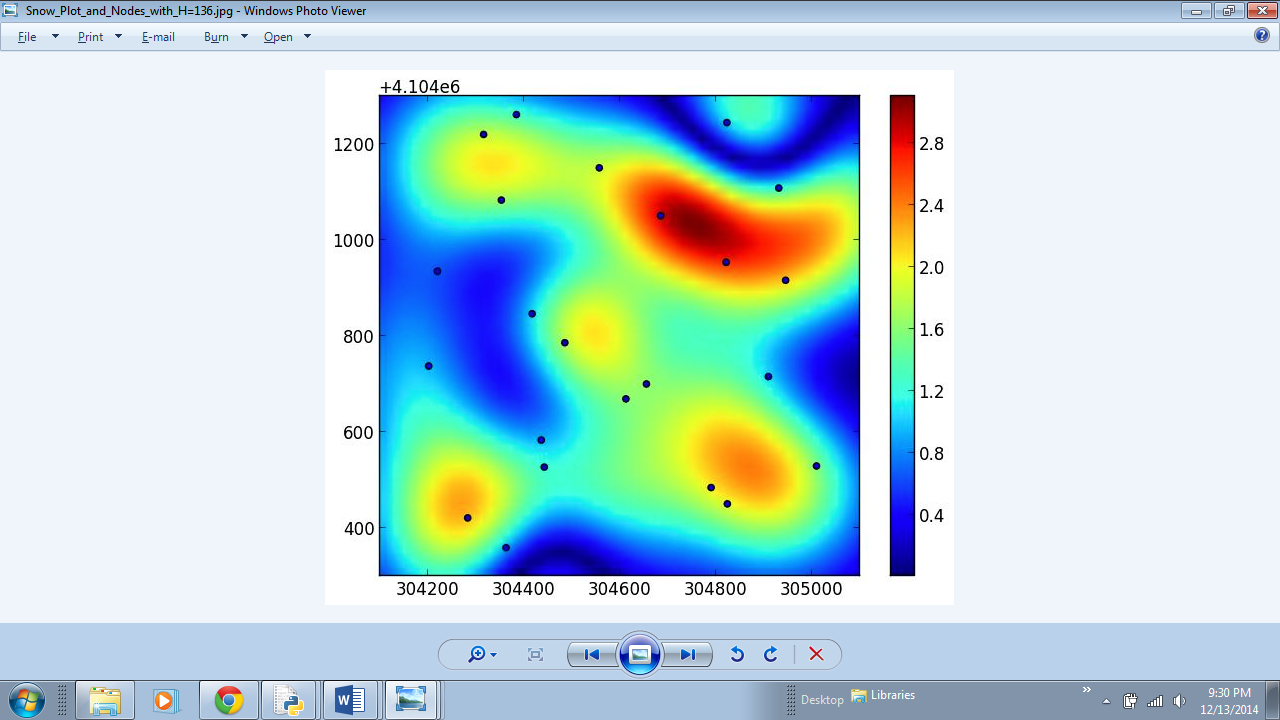
 

Figure Bandwidth versus Error Figure 2 Covariance versus Error

Using this optimal bandwidth, a stochastic simulation was generated, and thus a map of the simulated snow depth in the square kilometer region was produced. Figure 1 below shows the map generated from the bandwidth of 135 meters.



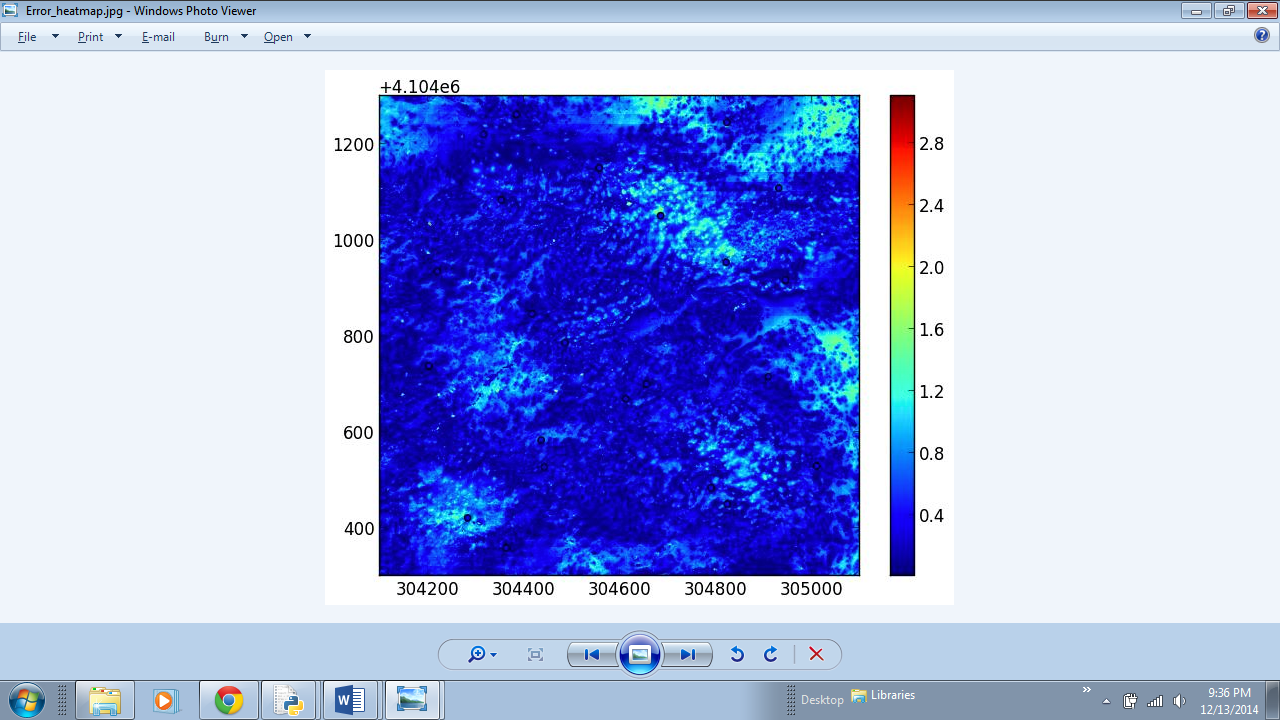


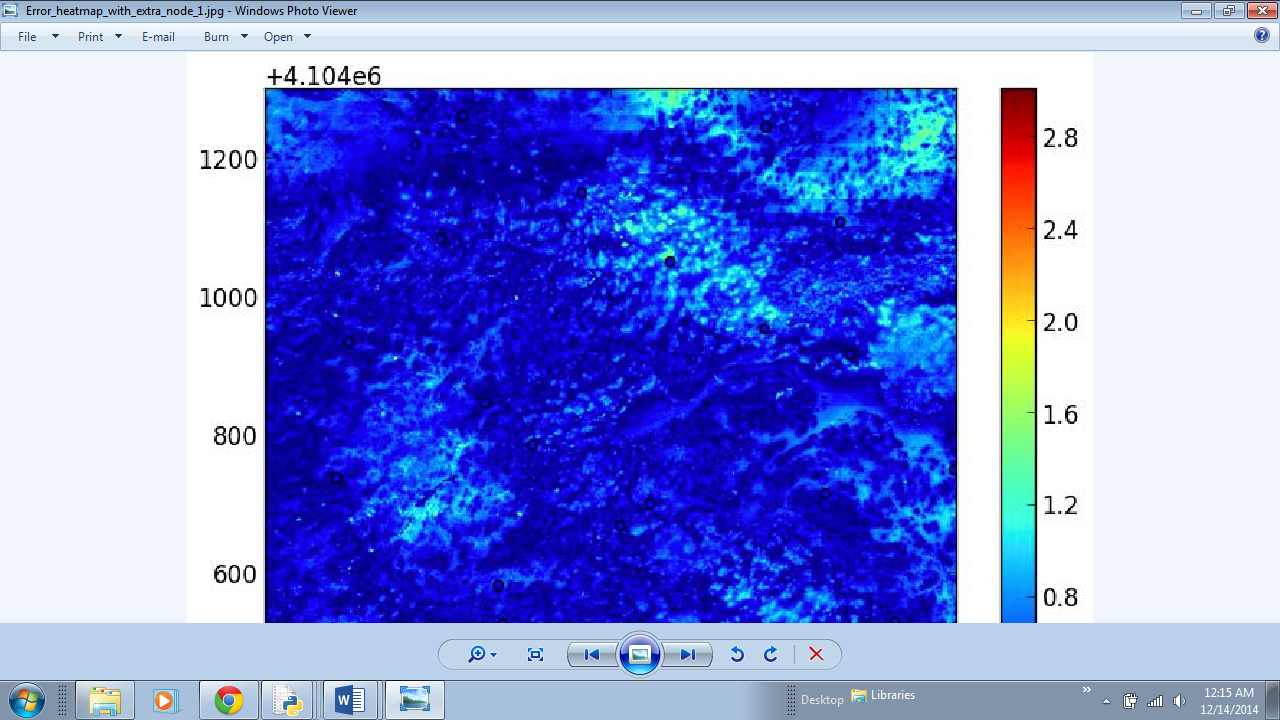
Figure 3 Snow Depth (cm) and Sensor Nodes

Figure 4 Heat map of Error between Stochastic Simulation and LIDAR Data

From Figure 2, one can observe the error between stochastic simulation, and LIDAR data. This error informs the decision of where to place an additional sensor node.

# Additional Node Placement

To construct a method for placing additional nodes to reduce total error, a few considerations were made. Firstly, it was assumed that locations in which there is a node, and a significant amount of error, will not benefit from an additional node. For example, the following figure shows a small window of the larger picture in Figure 4:



Although this region shows a notable amount of error, a node was deliberately not placed here, since there are already three nodes in the region. The places that were considered to be the best places for node placement were those where there were no nodes currently, and where the error was greatest. The following figures show the resulting heatmaps of error, with strategic placement of nodes.

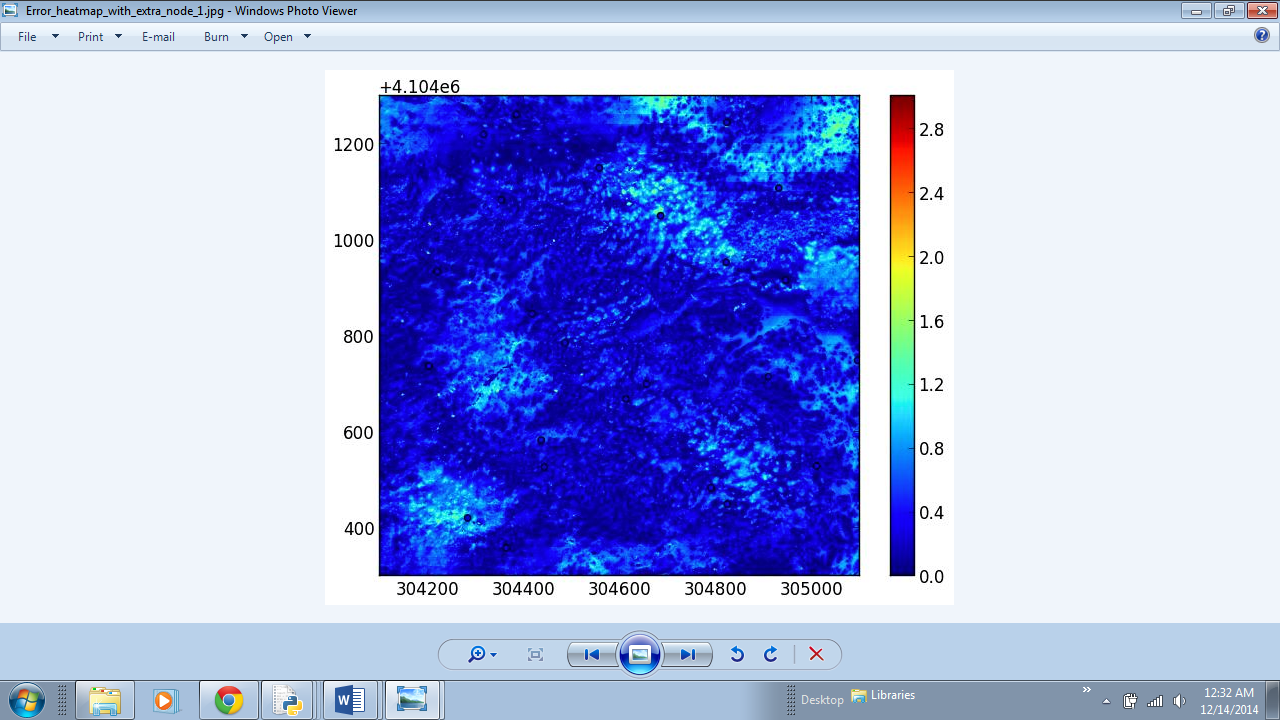


Figure Heat map of error with one node added at [305095.0, 4104750.0]

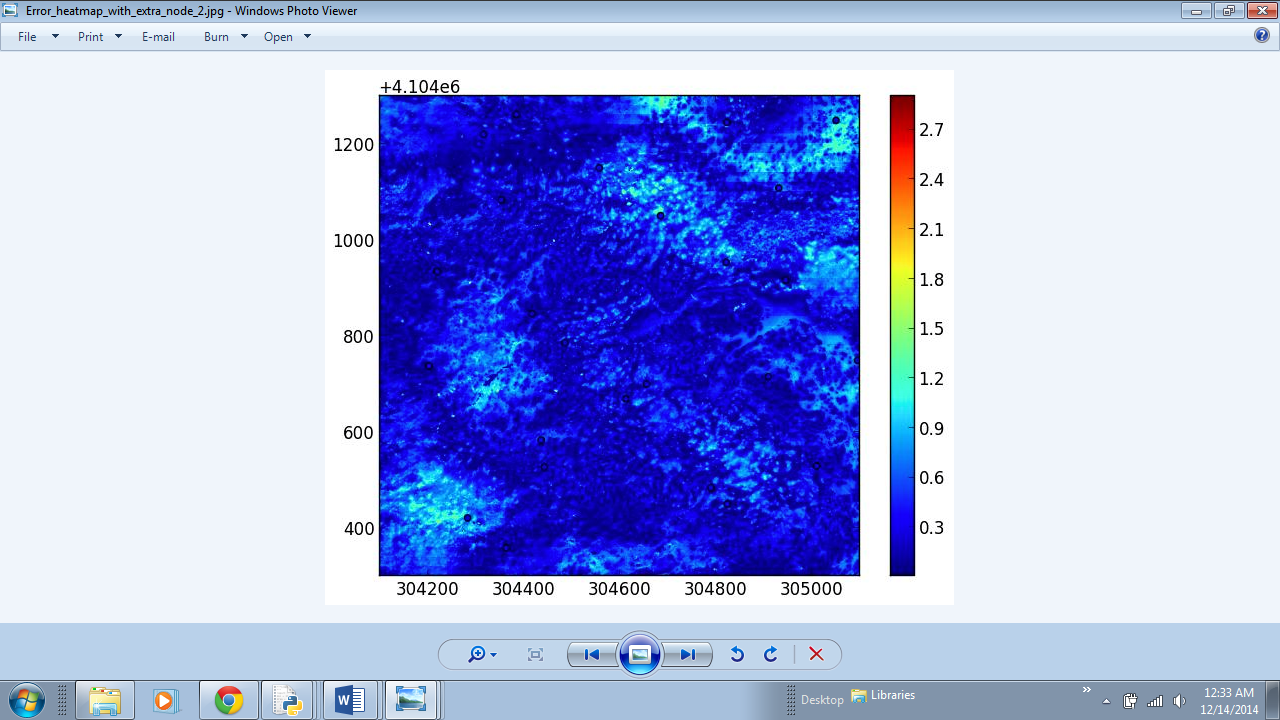


Figure Heat map of error with an additional node added at [305050.0, 4105250.0]

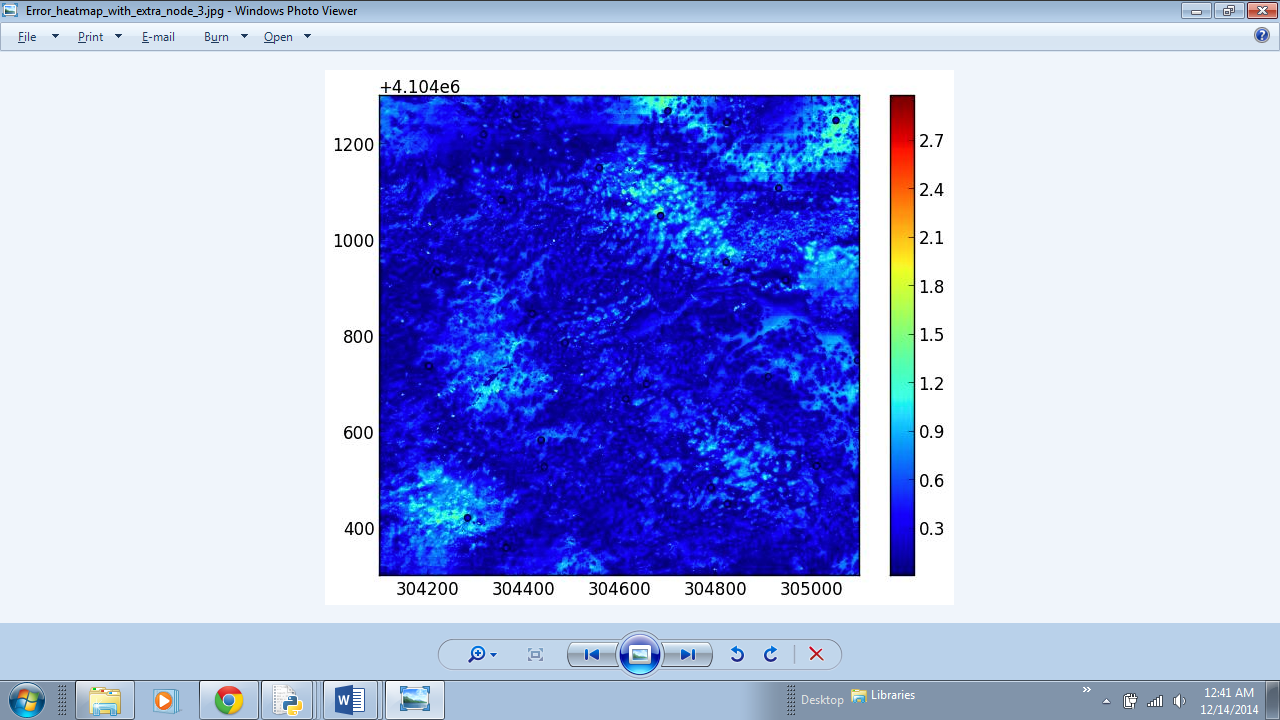


Figure Heat map of error with an additional node at [304700.0, 4105270.0]

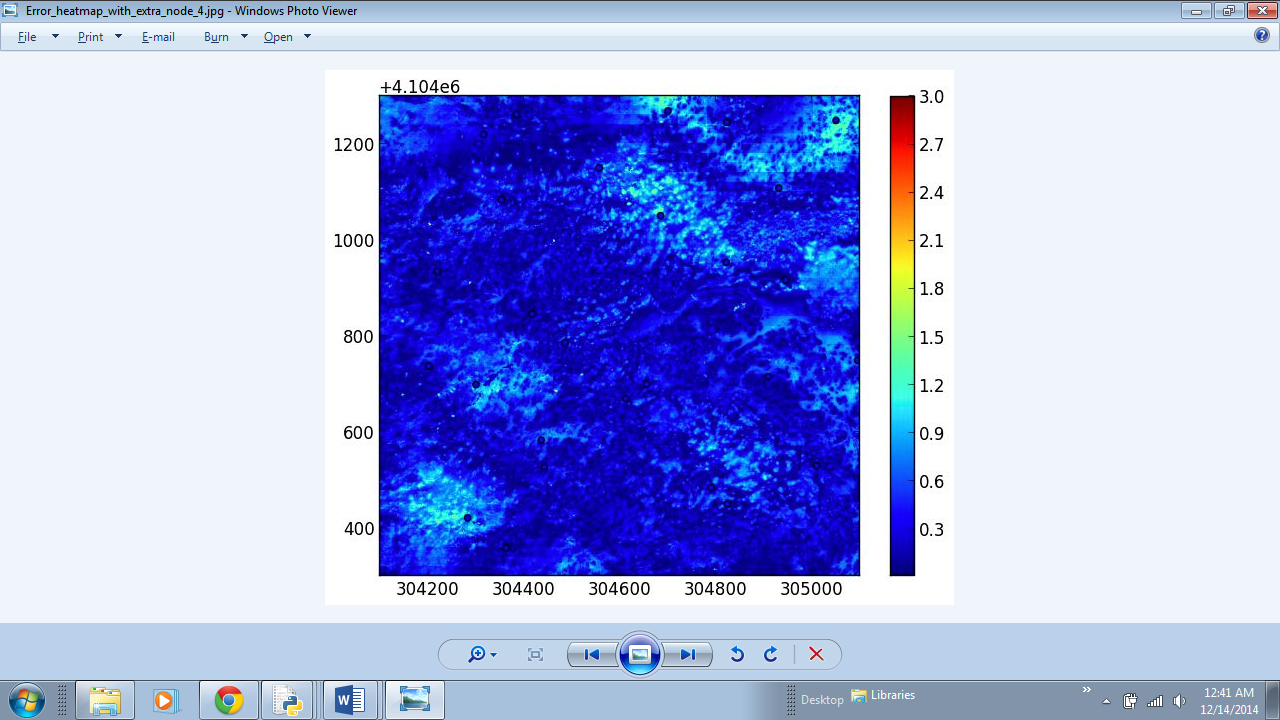


Figure Heat map with an additional node added at [304300.0, 4104700.0]

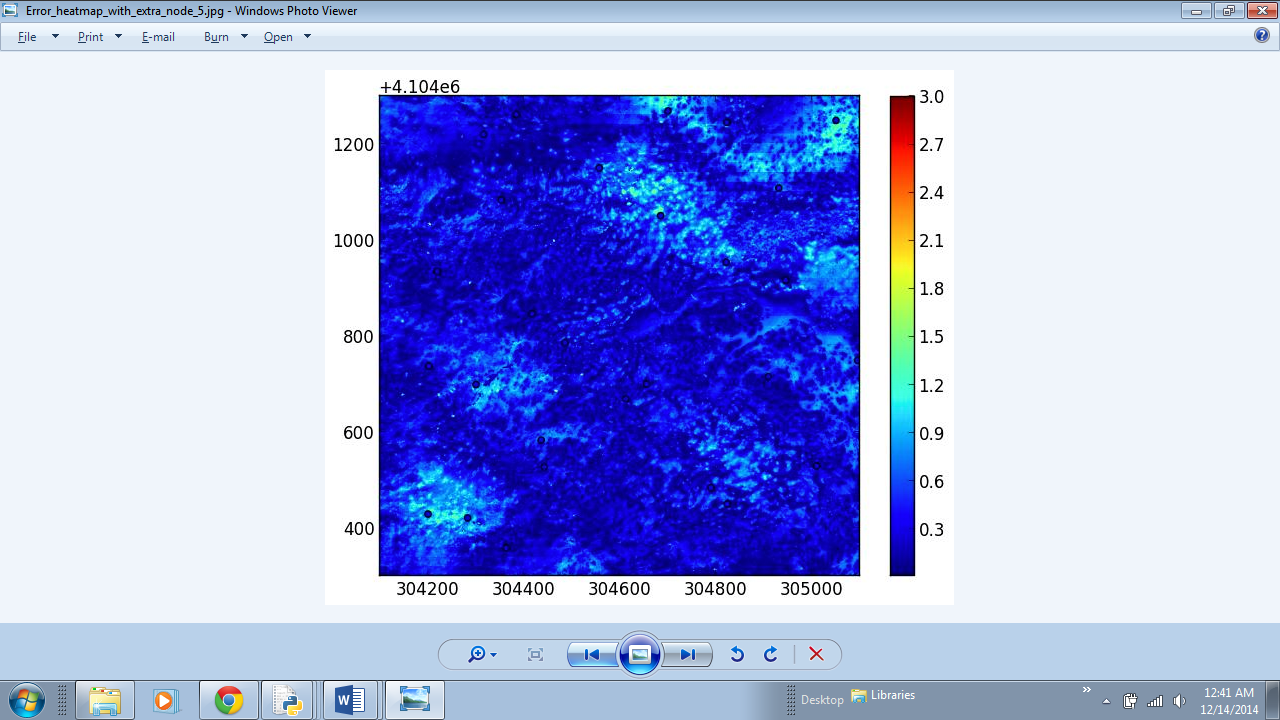


Figure Heat map with an additional node added at [304200.0, 4104430.0]

It should be noted that, for each heat map generated, a new bandwidth, and covariance generated. While the differences between plots is difficult to observe directly, a trend in the error elucidates the error reduction.

Figure Root Mean Squared Error versus number of nodes added

From Figure 7, one can observe that the error trend is not consistent, but it does show a general reduction in RMSE, with the exception of the introduction of the second node.

# Conclusion and Recommendations

All models are wrong, although some prove useful, so it is worth mentioning the limitations of this model. Due to the constraints on computation power, the grid of predictions generated by the Gaussian Process needed to be stretched to the dimensions of the grid of LIDAR data. An increase in the granularity of the prediction grid was attempted, but a “Memory Error” was encountered. The LIDAR data is a 1000 by 1000 array, whereas the grid of predictions was a 50 by 50 array. So, the 50 values along each dimension were repeated 20 times. This limitation on computation power could translate to a limitation on the accuracy of the model, and it is worth noting so.

This study might also benefit from an increase in the study area. In environmental patterns, one square kilometer is not always representative of an entire region. Furthermore, for snowpack, the ultimate goal is to estimate the amount of water stored in a region. Hence, the accuracy of an individual square kilometer is not as useful as an estimate for an entire watershed, which is often larger than one square kilometer.

The model shows that adding 3 sensor nodes will offers the greatest reduction in error. But adding more nodes than that seems to *increase* the error. Furthermore, adding the second node seems to increase the error as well. As a result, it is recommended to add 3 sensor nodes in the recommended locations, *excluding* the second node location.