**4.2. Real topographic data**

After the code development of different regression and resample technics from previous steps, it becomes useful to look at real data. For this purpose, a digital elevation raster image is provided and used for fitting different models. The dataset is located in Norway, on the Telemark region, and it was surveyed by the Shuttle Radar Topography Mission (SRTM) 1 Arc-Second Global (<https://www.usgs.gov/centers/eros/science/usgs-eros-archive-digital-elevation-shuttle-radar-topography-mission-srtm-1>). The resolution of this USGS satellite mission is 1 arc-second (30 meters).

The raster image is downsampled to 1000x1000 points by selecting the upper left corner, focusing on the Møsvatn Austfjell conservation area.

For the analysis, the elevation data is parameterized, and different regression functions were fitted: Ordinary-Least-Squares, Ridge and Lasso regression. A Min-max scaler is used to normalize the data. This normalization translates each value into a range between zero and one. Rescaling our data improves the efficiency of our calculations.

The models were tested initially up to a 10th order polynomial degree of approximation. However, the model errors escalated largely after the 5th order polynomial. Therefore, all the calculations were limited up to the 5th polynomial order (as tested in previous sections with the Franke function).

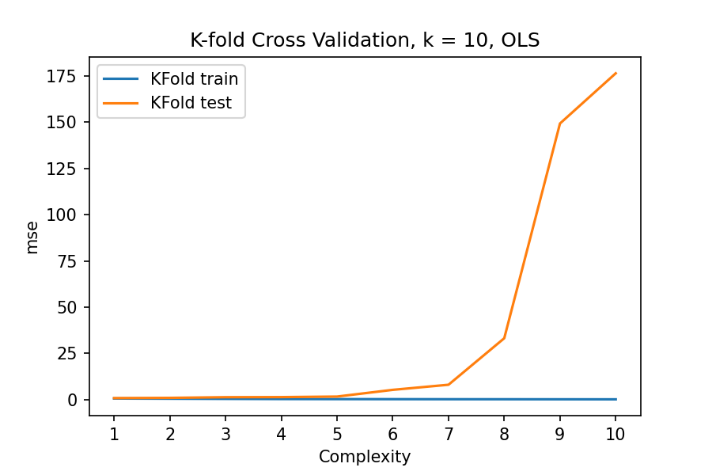
In order to evaluate which model fits the elevation data best, it was used a cross-validation resampling technique with 10 k-folds. In other words, we divided the dataset into k=10 groups, being one of them the test data where the model is tested, and the rest the training dataset where the model is fitted. In a loop, every time a different test and train groups are set, and the averaged model error is returned.

In Ridge and Lasso regression, the hyperparameter λ it’s introduced, ranging from different values and tested the model for 9 different lambdas. These parameters are not learned by the model, and they are tested in each case to find the values that minimize the model error.

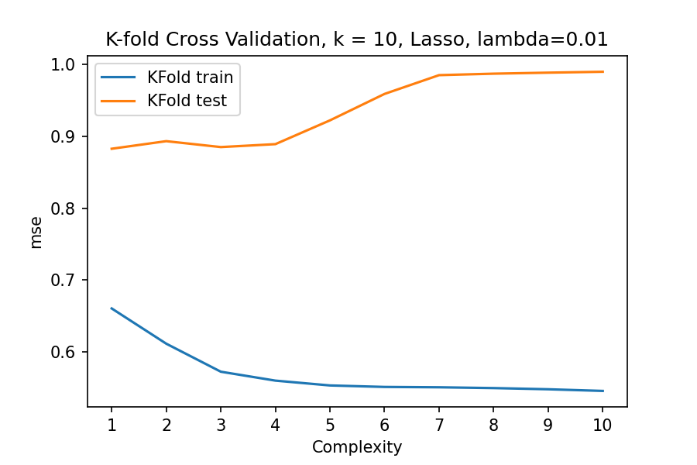
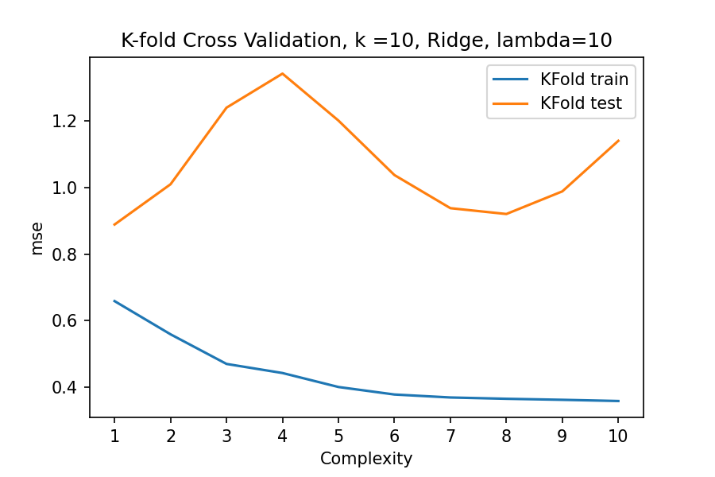
Several figures are produced to compare the mean square error for train and test dataset, depending on the complexity of the model (polynomial degree), and/or the hyperparameter lambda. Finally, a 3D representation of the fitted data is produced and compared with the original dataset, for some selected polynomial degree and lambda. All of this would help deciding which model performs better and under which parameters.

**5.2 Analysis of real topographic data**

Cross-validation method with Kfold=10 was performed for polynomial degrees 1 to 10 on a loop, and for a range of hyperparameter lambda in the case of Ridge and Lasso. A fast look to the obtained mean squared error (MSE) by polynomial degree on the OLS regression shows us that the MSE for the test datasets increases with increasing degree of complexity, specially with a degree higher than 5. Therefore, a smaller polynomial degree should be chosen for avoiding overfitting. In other words, the fitted model would work better with higher complexity on the train dataset, but the same model applied to another dataset would given worse results. Polynomial degree 1 to 5 will be assessed in OLS. For Ridge and Lasso regression, this relation is a bit different, and we can assess the degree of complexity from 1 to 10 (see Figure 1).

*Chart, line chart

Description automatically generated*

**

*Fig 1: OLS, Ridge and Lasso Cross-validation (10 fold) MSE for different polynomial degrees for train and test dataset.*

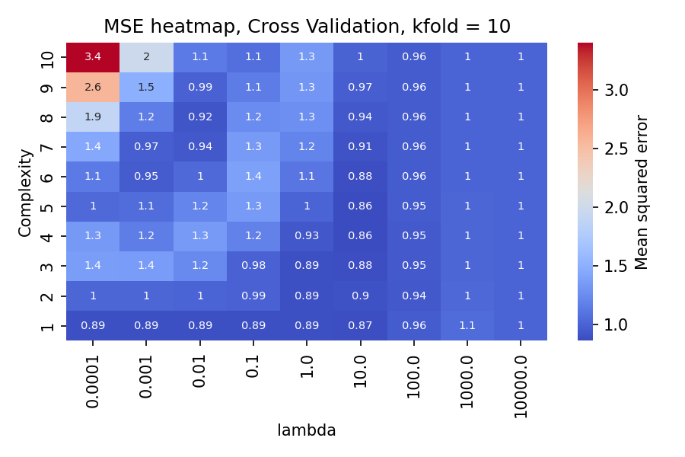
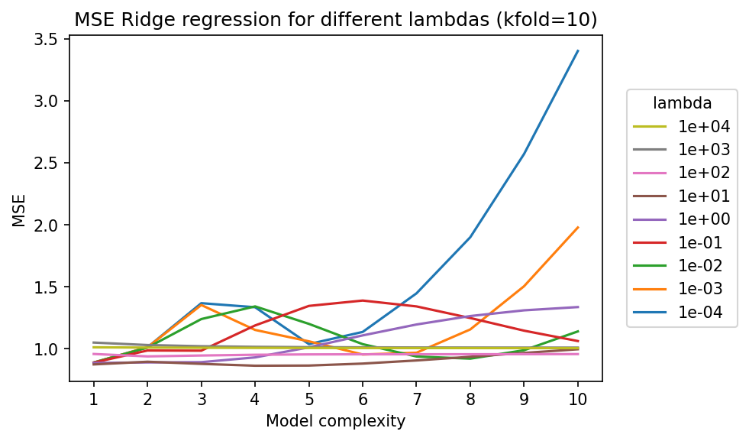
When we apply the regression model to the whole dataset and visualize it in a 3D map, we can compare it with the original terrain dataset and to other regression methods.

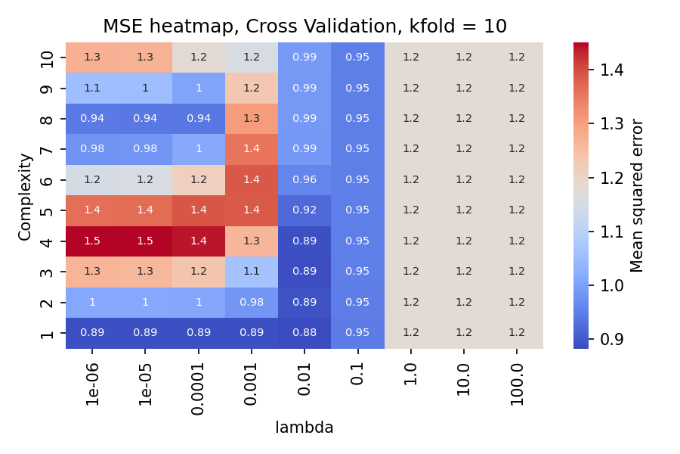
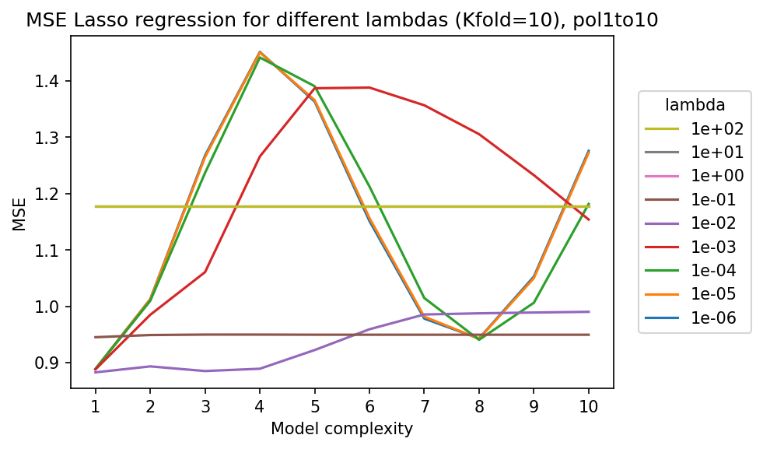
For Ordinary-Least-Squares, a degree of complexity of 3 was chosen as a good compromise between the train/test MSE and the visualization of the data.

In the case of Ridge regression, and after some testing, the range for lambda was set between 10-4 and 104, conducting 9 tests on the logscale (10-4, 10-3 … 104). In this case a polynomial degree of 4 and a lambda of 102 was chosen.

The same approach as in Ridge was conducted for Lasso regression. After some testing, the chosen range of the hyperparameter lambda for testing was 10-6 to 102, while the chosen optimal parameters were polynomial degree 4 and lambda 10-1.

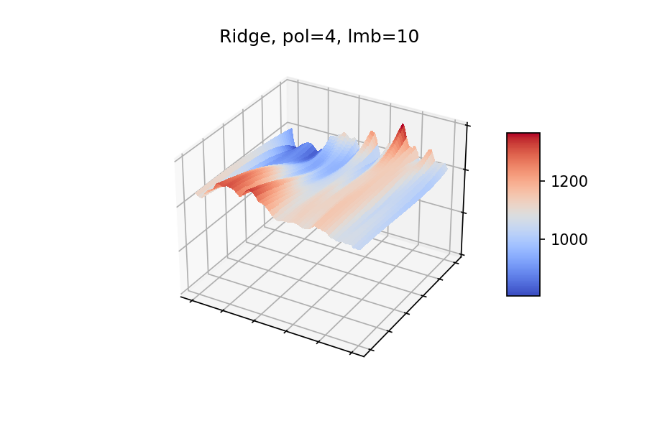
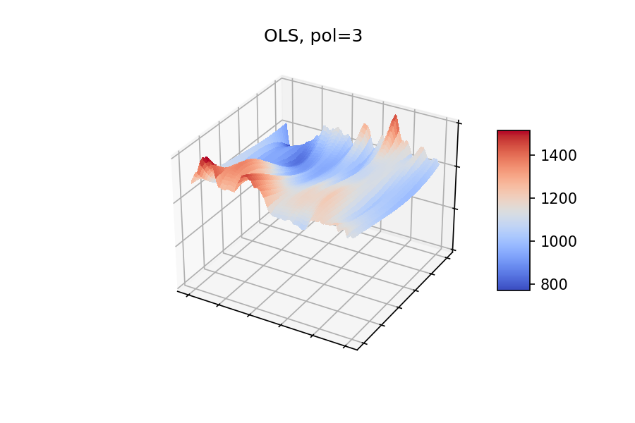
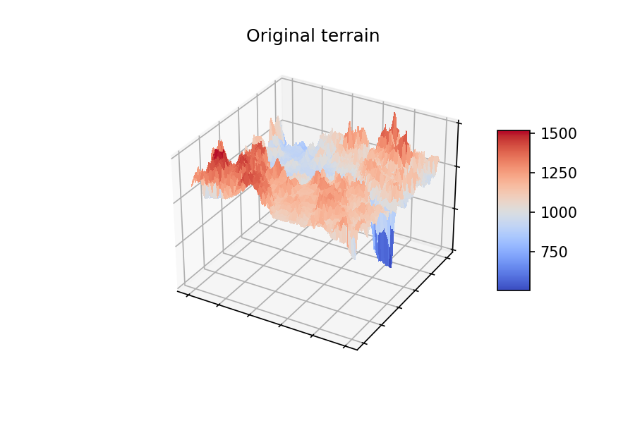
Figure 2 shows the performance of each lambda regarding MSE when increasing the complexity of the model, both as a line and in a numerical way, colored in a heatmap for easy identification of low/high MSE values.

**

**

*Figure 2: Ridge (higher) and Lasso (lower) regression MSE of each tested lambda for different polynomial degrees, showed as a line (left) and heatmap (right).*

The Ordinary-Leas-Squares method performed significantly better than Ridge and Lasso in terms of the MSE. The 3d representation (Figure 3) shows that the model predicts some general terrain features, but not all the variatin in elevation, which implies that a different model could be tested for fitting the terrain data better. The Lasso and Ridge regression performed fairly good but worse than OLS, both with a similar model error (MSE) and 3D representation.

Chart

Description automatically generated

*Figure 3: Visualization of the original terrain data (upper-left), the modeled one with OLS and polynomial degree=3 (upper-right), the Ridge modeled terrain with polynomial degree=4 and lambda=10 (lower-left) and the Lasso modeled terrain with polynomial degree 4 and lambda = 10-1 (lower-right).*