Last Judgement Report

Rodrigo Alejandro Chavez Mulsa

Maastricht University

1 Problem Statement and Motivation

A meteorite has been detected on its way to Maastricht, using the MASA (Maastricht Space Agency) radars the Gemeente wants to predict where the meteorite will fall to prepare and take early measures. This paper analyzes the radar data and provides with an estimate of where the meteorite will fall with a 90% interval.

2 Introduction and Description of Data

For this task 3 datasets were provided, two radar datasets (where the second one is 5 times more accurate) that contains n (130 for radar1 and 120 for radar2) position datapoints of the meteorite in x, y, z coordinates in intervals of 10 seconds. The third dataset contains the positions of buildings in Maastricht with their residents, beds and baths count. Some resident datapoints where missing but after trying multiple predicting models, the approach used to fill this was to use the median of residents which was really close to the mean.

3 Modeling Approach

For this task two models were used, a simple linear regression as baseline and a polynomial linear regression as main model. Since the objective was to predict x and y when z is 0, the first approach was to use z as an independent variable and predict the coordinates based on this altitude. The second approach consisted on using z to predict at which moment in time the meteorite will hit the city and train the x and y models based on the time feature. The reasoning for this is that since the time intervals are fixed in 10 seconds each, there will be less variance since time is independent from the meteorite position, while if z is used to predict x and y, the variance of z could induce a bigger error.

3.1 Simple Linear Regression

This baseline model was implemented using the sklearn library with no parameter tuning. Two approaches where used to predict the x and y coordinates, in the first one z was used as independent variable and in the second one time was the independent variable. To measure the performance of the model, the root mean squared error and R^2 scores were used.

Include at least one baseline model (perhaps a simple one?) for comparison and then describe your implementations beyond the baseline model and the design choices that you made along the way. As with the clinics, it is highly important to properly motivate your choices, justify any selection and comment, interpret and judge the results.

3.2 Polynomial Regression

This model is used because after looking at the simple linear regression model it was clear that the coordinates cold not be predicted with a linear model and a polynomial was required to better describe the data trend. Besides using the previous approaches for independent variables, different polynomial degrees were tested where a 3rd degree resulted in the best fit for the data, the same measurement features were used as in the linear regression model.

3.3 Splitting and resampling data

In order to test the models, the data was split in train and test datasets where 20% was allocated to the test set with and without shuffling but after multiple tests shuffling was discarded because the model scores were misleading, since the model should predict future points ahead of the training dataset the last 20% of the datapoints were used.

Due to the small amount of data points and the severity of the situation, a prediction with a confidence interval was required. To accomplish this, ten thousand training datasets of same length were created using random resampling with repetition, where the datapoints that were not in this dataset were used for testing.

Thereafter, a 90% confidence interval was computed using the ten thousand models that were trained and tested in the new resampled data.

4 Results and Interpretation

The polynomial regression gave promising results with the confidence interval, in the following figure it's visible in red the area corresponding to the interval where the meteorite could fall, and the gray circles represent the buildings.

Using the radar1 data it was estimated that 7 residents would be affected while radar2 predictions shows no building would be impacted.

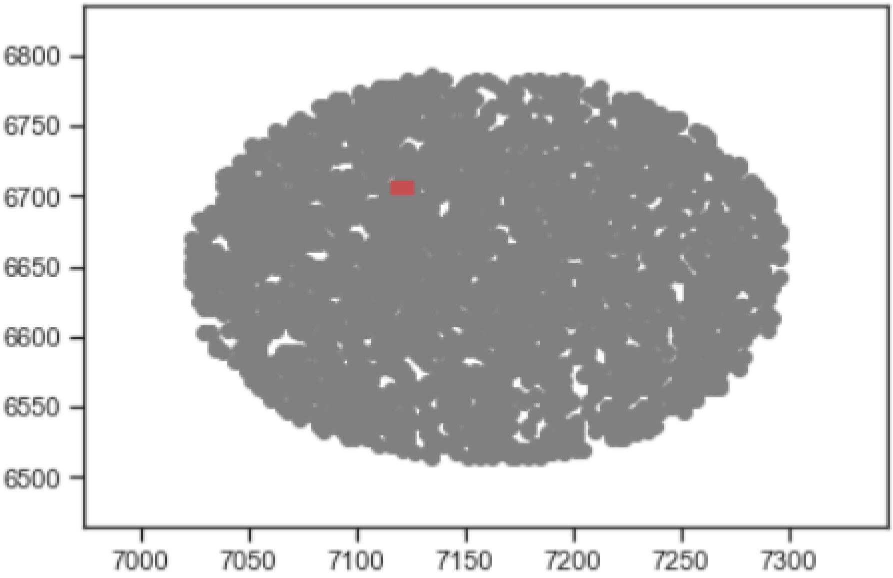


Fig. 1. Predicted hit area from radar1 with 90% confidence interval

Fig. 2. Predicted hit area from radar2 with 90% confidence interval (outside buildings zone)

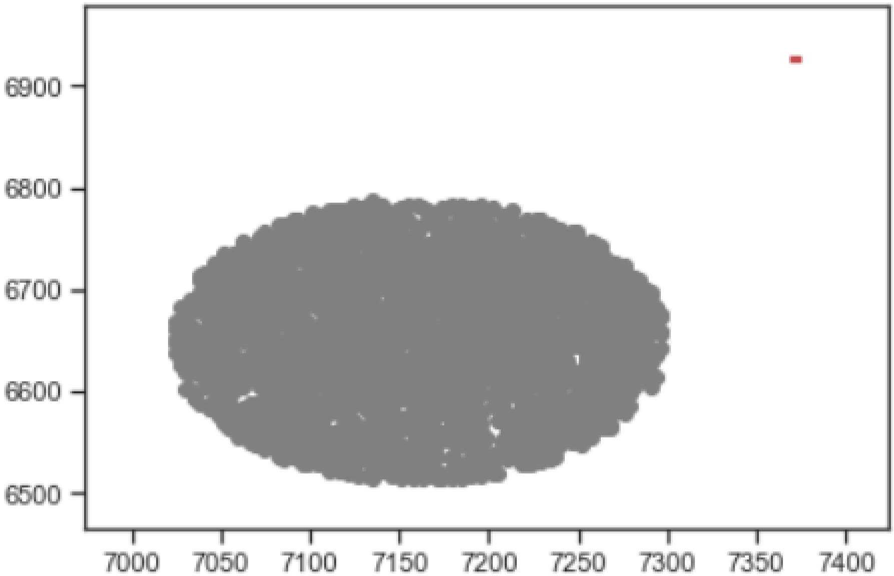


Fig. 3. Predicted hit area with 90% confidence interval from when radar1 and radar2 data is concatenated

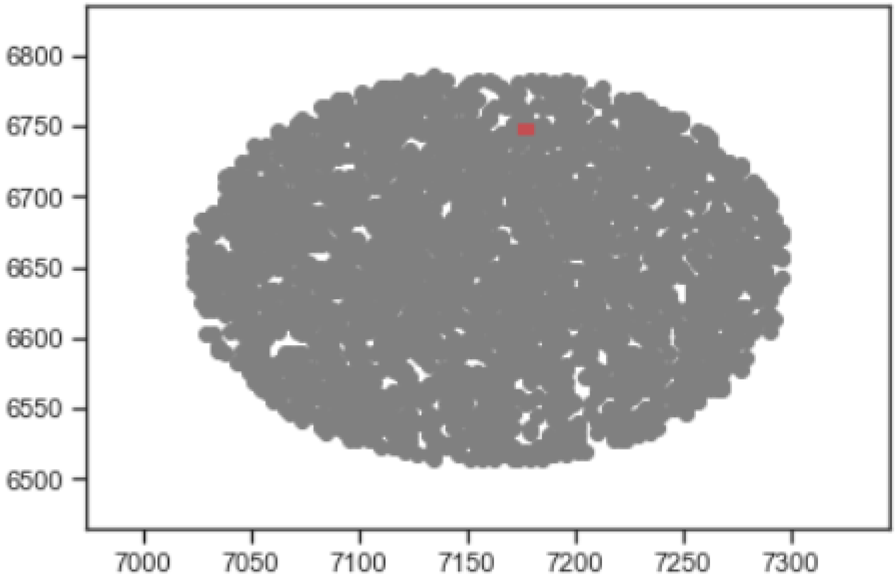
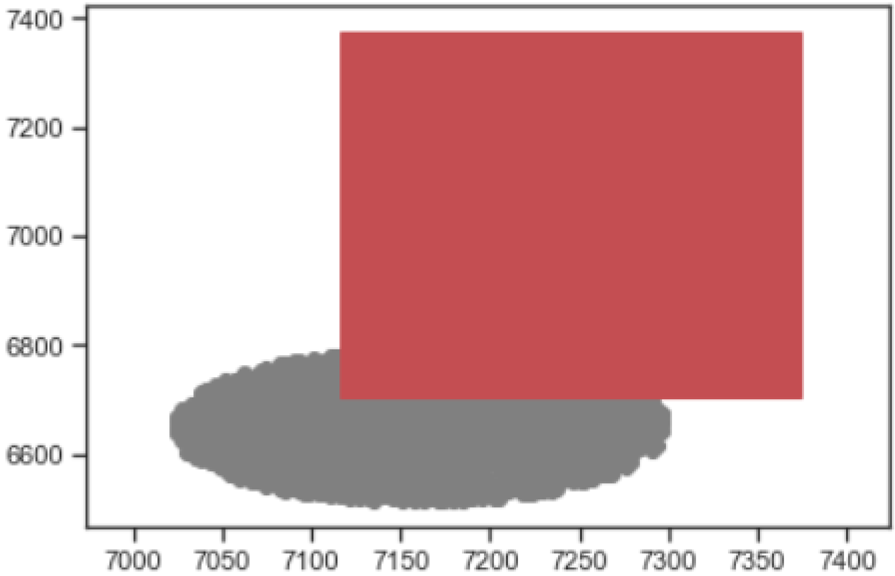


Fig. 4. Predicted hit area that its between radar1 and radar2 predictions



5 Conclusion

In this paper it was analyzed radar datasets and predicted the impact point of a meteorite with a 90% confidence interval. The results of radar2 model show that the meteorite will not impact any building, but more testing and different approaches should be done to avoid any preventable causalities in case the model is overfitted to the first dataset.

Some possible approaches could be using the more precise dataset for testing or compute intermediate points between the given data and when coordinate z is zero.

Deciding on which model to use when the amount of data available is small is hard and testing in a robust way is important. Through the experiments it was visible multiple times how small changes on the approach or model have a big influence in the results. Situations like a meteorite hit prediction

References