Identifying the Correlation between Meteorite Landings and Natural Disasters on Earth

Lauren Liao[1] and Jiwon Yoo[2]

[1]A12831428, [2]A11277388

December 5, 2018

0.1 Abstract

According to NASA, no recording of human fatality is attributed to meteorite landings in the past one thousand years. However, meteorite landings can create impact events that have hazardous effects on the environment and subsequently cause natural disasters that lead to thousands of deaths of human and animals alike. Therefore, we are interested in examining the correlation and possible factors attributing to the relationship between meteorite impacts and number of subsequent natural disasters such as earthquakes, tsunami, and forest fires. We hypothesize there is a significant relationship where more instances of natural disasters occurring after the impact events, and specific attributes of the meteorites are positively correlated number of natural disasters that follows. We are using Meteorite Landings data from NASAs Open Data Portal and Disaster Declarations from Federal Emergency Management Agency (FEMA) to address our problem. First, we used simple and multiple linear regression to consider whether a relationship exist or not and identifying, if any, which aspect of the meteorite attribute to a significant correlation to the number of natural disasters; then we used cross validation to do model selection and bootstrap re-sampling to estimate the distribution of our model coefficients We found that the best model contains none of the meteorite landing data as an explanatory variable, in fact, we found only correlation with which year the natural disaster data was recorded. More information should be obtained to will enable further understanding of the impact meteorites have on Earth.

0.2 Introduction

Meteorite landings often contain lack of understanding on the impact it has on Earth. Popularized by media, meteorite landings often are associated with large impacts that can possibly wipe out the human existence as it did for dinosaurs. Known as the Cretaceous Paleogene extinction event addressed in Schulte et al. [3] (2010), the large meteorite landing caused global environmental change, cooling the Earths temperature by 10 degrees Celsius. However, most meteorite landings do not cause such significant impact.

Due to the randomness of meteorite landing timing and location, predictions on the location and direct implications to human survival and well-being is difficult to quantify. Therefore, researchers attempt to quantify the impact and highlight the direct effects caused by the meteorite

landings. In Rumpf et al. [2] (2017), the authors address the direct causation of asteroid landings in affect to human populations within 15-400 meters. Specifically, they stated that despite the potential tsunami as a result of asteroid landing in water, the impact on land has more harm to the human population with an increase in casualty numbers. Furthermore, Collins et al. [1] (2005) address the regional environmental consequences from the meteorite landings. They emphasized the impact-induced seismic shaking as a significant feature causing wide-range environmental effects and thermal radiation as a significant feature in short-range environmental effects. Even though National Aeronautics and Space Administrations (NASA) Jet Propulsion Laboratory, meteorite landings have not directly caused human fatality, the geophysics researchers had indicated the environmental changes as a direct cause by the meteorite landing impacts. Therefore, we are interested in understanding whether there is a correlation between the meteorite impacts and the number of subsequent natural disasters? If so, what are the possible factors?

We hypothesize there is a significant relationship where more instances of natural disasters occurring after the impact events, and specific attributes of the meteorites are positively correlated number of natural disasters that follows. This is important because we can use this information to understand the meteorite landings further in consideration of the extended possible consequences in relation to natural disasters. Meteorite landings may cause a ripple of subsequent events that has serious consequences to the human population. Although the effects may not be as serious as extinction, it is still important to consider the environmental factors that are associated with meteorite landings.

0.3 Materials and Methods

0.3.1 Data sets

We are using the Natural Disasters data set from Federal Emergency Management Agency, or FEMA. The data set has 48390 observations and 15 columns: Disaster Number, IH Program Declared, IA Program Declared, PA Program Declared, HM Program Declared, State, Declaration Date, Disaster Type, Incident Type, Title, Incident Begin Date, Incident End Date, Place Code, and finally, Designated County. Among these data, we focused on 4 columns: Incident Begin date (in years), Incident Type, Title, and Declared County. We took out rows with missing data and

restricted observed years up to 2012 to match with other data sets we have that will be used during our analysis. The final data for Natural Disaster contains 43206 rows and 4 columns. This data set contains a lot of different types of incidents classified as natural disasters, and it is difficult for us to identify which natural disasters are, if any, affected by or related to meteorite impacts. Even there are some relation, there are many unknown variables affecting natural disasters; thus, our analysis will have irreducible noise caused by these variables.

We are using the meteorite landings data set from NASA open data portal. The original data as seen in figure 1 has 45716 rows and 10 columns. After taking out the rows with missing values, we have 38116 rows of data entry left for analysis. However, since our corresponding data set has a limited range in years, we have restricted our data to match the number of years. We took out possible entry errors, such as latitude and longitude of 0N/0E, we note that this place is off of the western coast of Africa which is difficult to discover any meteorite pieces. We also restricted our data to only United States of America to match the data set above as seen in the figure 1. In the end, we conclude our data set with 1020 row entries and 7 columns: meteorite name, recclass (as classified by NASA), mass (units in grams), year, and geographic location. In the final set of data, we have considered the full set using summary statistics to use as explanatory variables each year. For example, we are keeping the class of the meteorite that landed the most in the given year and the count for that particular class. We are also keeping the average mass and the maximum mass of the meteorite landing in a particular year. These summary statistics help us to identify possible contributors to the natural disasters and possible relations.

The final data set we are using consists of both the data sets as stated above. The final data set contains 54 rows and 7 columns: years, ndo (natural disaster occurrences), mio (meteorite impact occurrences), class_name, max_count, avg_mass , and max_mass. Each row is based on specific year from 1959 to 2012. Class_name refers to the class with the maximum count of meteorite landings. Max_count refers to the maximum count of the meteorite landings, corresponding to the class_name. Avg_mass records the average mass in grams of all the meteorite landings of the given year. Max_mass refers to the maximum mass in grams of all the meteorite landings of the given year.

Introspect

To understand our data, first we plotted the meteorite impact occurrences against the natural disaster occurrence. We considered the bar plots of the meteorite landing events and natural disasters over time to visualize the data. The data shows that there are abnormally large number of meteorite landings in 1979 with over 3000 accounts and large number of natural disaster occurrences in 2005 with near 5000 incidents as possible outliers. However, we cannot draw a clear conclusion on the trend in visualizing initially at a glance 2.

0.4 Methods

0.4.1 Linear Regression

Since we are fitting a continuous predictor, we choose to use linear regression. Linear regression is trying to fit the continuous response variable using the response variable with random normal errors. This means that we are building a model that assumes possible linear relationship between the response variable and the explanatory variable(s). Without loss of generality, let us consider the simplest case for linear regression. The equation $\hat{y} = \beta_0 + \beta_1 X$ is modelled. y is the response variable, in our case, the number of natural disaster occurrences, $\hat{y}$ is the estimated y from the linear model. β_0 is the intercept, and β_1 is the change in X contributing to the change in y, where X is the explanatory variable. The explanatory variable can be any possible contributing variables we are considering.

Since the response variable is the number of natural disaster occurrences, we plotted possible explanatory variables to consider whether there are any visible relationships at first glance in the figure 3. The only visible relationship is that the data tend to gather at the bottom right corner, which creates an exponential distribution for some explanatory variables. We also considered the residual plot (not shown) to consider if there exists any patterns in the errors. However, we do not see any visible patterns. Therefore, we choose to use linear regression, considering if we were to choose other non-linear regressions, then the non-linear regression would be over-fitting the data and capturing the random pattern of the noise instead of the underlying relationship of the data.

0.4.2 Cross Validation

We choose to use cross validation to compare different models to select the best number of variables that should be included in the final model. Since models with more explanatory variables always results in lower training set mean squared error, we choose to use cross validation to use a subset of the data as the testing model to evaluate the performance of the model. We then will obtain the testing set mean squared error to consider the performance of the model to make comparisons. We note that testing and training set of the model refers to supervised learning where we know the true values in the testing set compare the estimated parameters (generated using the training set) to calculate mean squared errors. Because our data is small, we use 3 fold cross validation, where we take a third of our data first then using that set as the testing set and the rest as training set. Then we use the second set of third, then last set of third as training set and repeating the process. Lastly we use the generated mean square errors to compare the different models and evaluate which model is the most appropriate.

0.4.3 Bootstrap Re-sampling

To estimate the standard error for the estimated coefficient(s), we choose to use bootstrap re-sampling. The non parametric bootstrap re-sampling uses the original data and sample with replacement to create a simulated data set of the same size. Since we have no prior assumption on how the data should look like, this re-sampling method presents well estimated the standard deviation according to the original data. Therefore, we choose to use this to estimate our standard error for the estimated coefficient(s).

0.5 Results

We fitted a simple linear regression line using ordinary least square model. We consider our results significant at type I error of level 0.05. We tested the hypothesis that the coefficient for meteorite impact occurrences is the same as 0, without any correlation. We considered both simple and multiple linear regression to consider the contributing factors modeling NDO. First question we needed is whether any of the explanatory variables are significant. Therefore, we use the F statistics p-value to determine any, if at all, the variables could be significant. Using 3 different predictors

from our data: mio (meteorite impact occurrences), years, and avg_mass, we generated 3 simple linear models and 4 multiple linear models. Following are 3 simple linear models we used:

Table 1: Simple Linear Regression Results summary

Model	p-value	Training MSE	Testing MSE	Bootstrap Confidence Interval
ndo $\sim$ 1 + mio	0.087	491444.385837	487348	[0.283669, 0.348720]
ndo $\sim$ 1 + years	0.000	3.466295e+09	3.432455e+0	[36.481105, 39.051040]
ndo $\sim$ 1 + avg_mass	0.272	6.526792e+08	6.216565e+08	[-0.027507, -0.017506]

First model (ndo 1 + mio) shows p-value $>$ 0.05 suggesting that mio may not be significant to ndo. And since reported p-value was greater than 0.05, we cannot reject the hypothesis that the MIO does not have a significant effect on NDO (Natural Disaster Occurrence). Similarly, the third graph (ndo $\sim$ 1 + avg_mass) shows p-value $>$ 0.05. As seen in the figure**??**, we tested the hypothesis that the coefficient for the avg_mass of the meteorite of the given year is the same as 0, without any correlation. The resulting p-value of the variable avg_mass is 0.272. Since reported p-value was greater than 0.05, we cannot reject the hypothesis that the avg_mass does not have a significant effect on NDO. We now look at years as the explanatory variable. Within simple linear regression models, years is the only predictor that shows significant effect on ndo with a p-value of $<$ 0.05. We found that years is the best contributing factor but there might be more predictors contributing to ndo. So we considered multiple linear regression also. Following are 4 multiple regression models we used:

Table 2: Multiple Linear Regression Results summary

Model	p-value	F-value	Testing MSE	CI
ndo $\sim$ 1 + mio + years	mio: 0.040 years: 0.000	1.73091 e -08	267964	mio: [-0.410011, -0.294609] years: [45.111177, 50.581814]
ndo $\sim$ + mio + avg_mass	mio: 0.137 avg_mass: 0.4666	0.17994	268292820	mio:[0.197771, 0.272834] avg_mass:[-0.024233, -0.013549]
ndo $\sim$ 1 + years + avg_mass	years: 0.000 avg_mass: 0.975	1.44285	268025020	years:[35.284239, 37.966268] avg_mass:[-0.007456, -0.003737]
ndo $\sim$ 1 + years + avg_mass + mio	years: 0.000 avg_mass: 0.813 mio: 0.636	9.93394e-08	178861935	years:[44.115732, 49.119838] avg_mass: [-0.410011, -0.294609] mio: [-0.398091, -0.283549]

The Table 2 shows the model, p-value of each predictors, F-statistic-p-value of the model, and 95% confidence interval of a selected model after running bootstrapping. From the table, we can conclude that years predictor is significant because it shows small value in every models with years.

In model 1, mio reports small p-value that is less than 0.05.

From the above experiment, we were able to figure out two possible models that may fit best to our analysis. A model with a single predictor: years, and another model with 2 predictors: years and mio. Since two models have different number of predictors, we used cross-validation to select best models. The resulting mean squared error, or MSE, for the simple linear model (ndo $\sim$ 1 + years) was 165.5015 while multiple linear model (ndo $\sim$ 1 + mio + years) was 267964.9143. Reported MSE values are very high, and we suspect that this is due not normalizing our data before-hand. However, normalizing our data would produce the same results as to which model has the lowest MSE. Therefore, our method is still valid. Based on these results, we concluded that years is the only predictor that should be in our finalized model that is, simple linear model.

0.6 Discussion and Conclusions

We learned from our analysis that the best model with the least mean squared errors is the simple linear model with year as the explanatory variable. This result is surprising because it contradicts our hypothesis that the meteorite impacts are a significant contributor to the natural disaster counts. We learned that meteorites average mass and occurrences does not contribute as significant factors to the natural disaster counts. In fact, none of predictors other than years were significant to include in our final models. We had 54 rows and 7 columns: years, ndo (natural disaster occurrences), mio (meteorite impact occurrences), class_name, max_count, avg_mass , and max_mass. Among these, we used 3 predictors: years, mio, and avg_mass and concluded that years was the best factor that contributes to our model. Although we considered all possible combinations of models that made sense, there were many confounding variables that we we were not able to capture such as environmental features like temperature that might act as an important contributor to our final model. As Rumpf et al. [2] (2017) addressed the impact of range 15-400 meters, perhaps moving forward we can consider the particular locations of the specific impacts and note the type of natural disasters with the location that happened. For example, a meteorite fallen in an ocean could have potentially caused a tsunami in near city or even small villages. Using this information, we can better assess the impacts. However, the current data we have is very limited. Perhaps using summary data provides too general of a calculation to the impact a meteorite landing may

cause. Since we have restricted the meteorite data to only locations of the United States, if we were to have more data on the natural disaster around the world, we may obtain a better result. Our data is also restricted to only the particular year; with further information about the month that it landed and estimated effects given the mass of the meteorites, we will be able to reach a better conclusion to the environmental impact and natural disasters that are linked to the meteorite landings. Since the year factor is very important as an explanatory factor, we can consider that as the year increases, there are more natural disasters. This suggests that other factors that are not correlated to meteorite landings may be significant that contributes to the natural disasters. With more time, we can consider other data sets such as greenhouse gas emission or the global temperature of each year to address the increase of natural disaster occurrences.

0.7 Acknowledgements

We thank Professor Mukamel, teaching assistants Justin and Connie, in COGS 109: Modeling and Data Analysis class, for teaching us and enriching our learning experiences at UC San Diego.

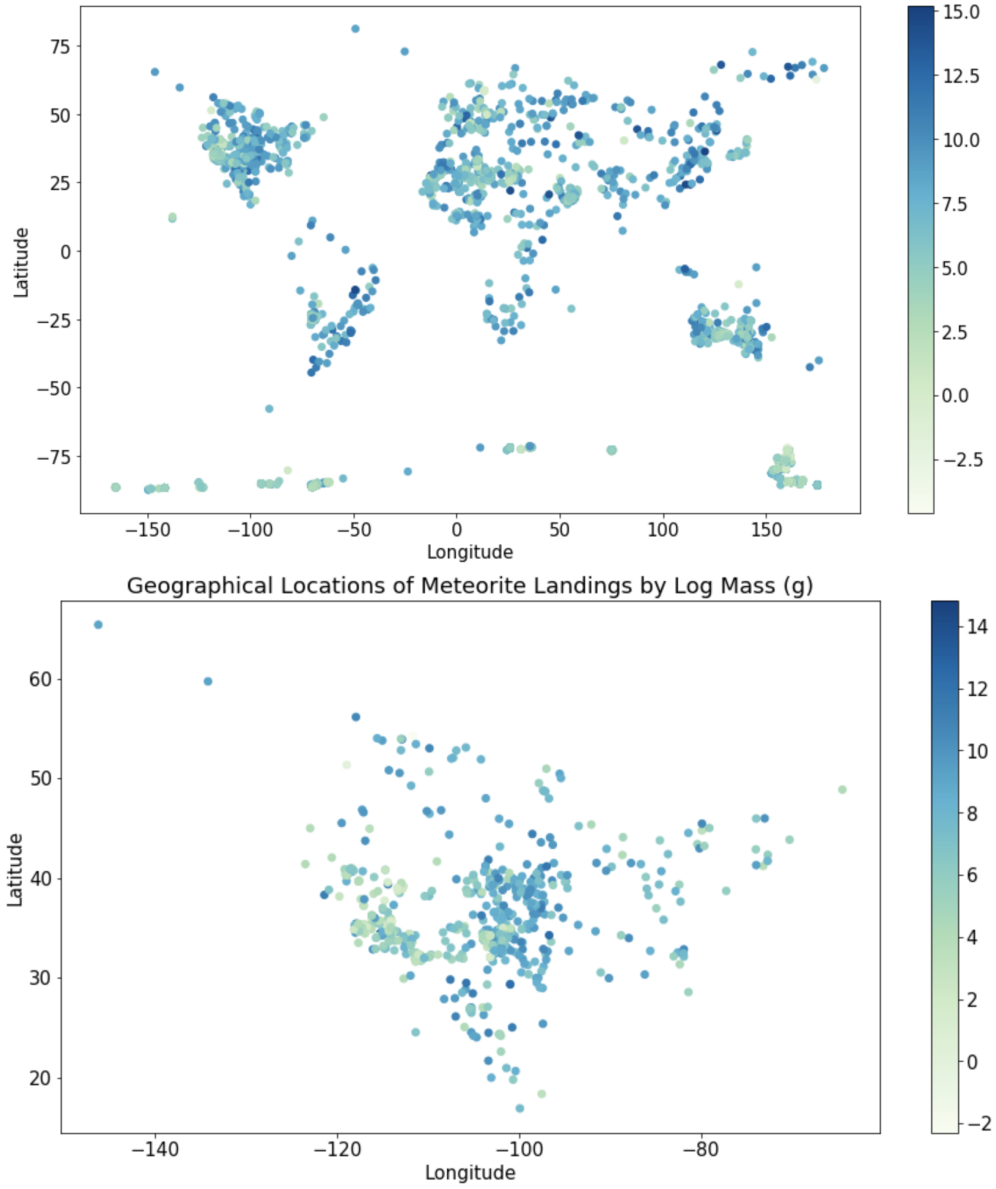

Figure 1: The figure shows a scatter plot of the geographical locations of the occurrences with log10 scaled mass to visualize, where darker color blue indicates heavier meteorites. Light green indicates lighter meteorites. Top figure shows the meteorite landing locations. Bottom figure shows the restricted the data for meteorite landings to only addressing the US since the natural disaster data only addresses the US.

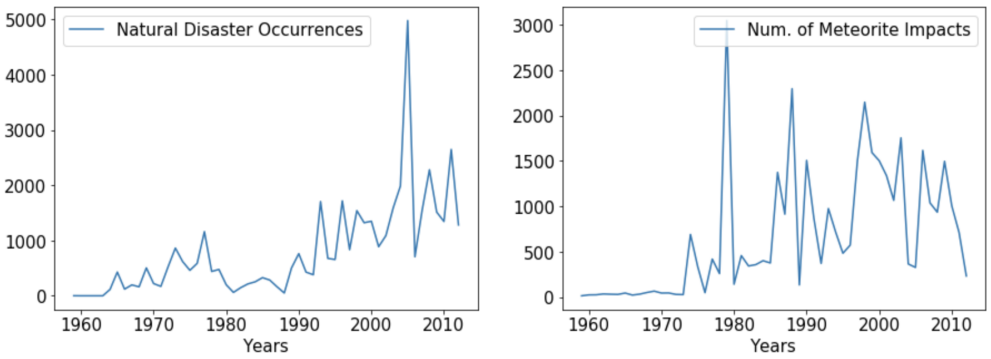

Figure 2: Line trends over time (in years). The figure show the number of natural disaster occurrence (left) and the number of meteorite impacts (right). Line trends plotted to provide initial comparisons of the occurrences side by side.

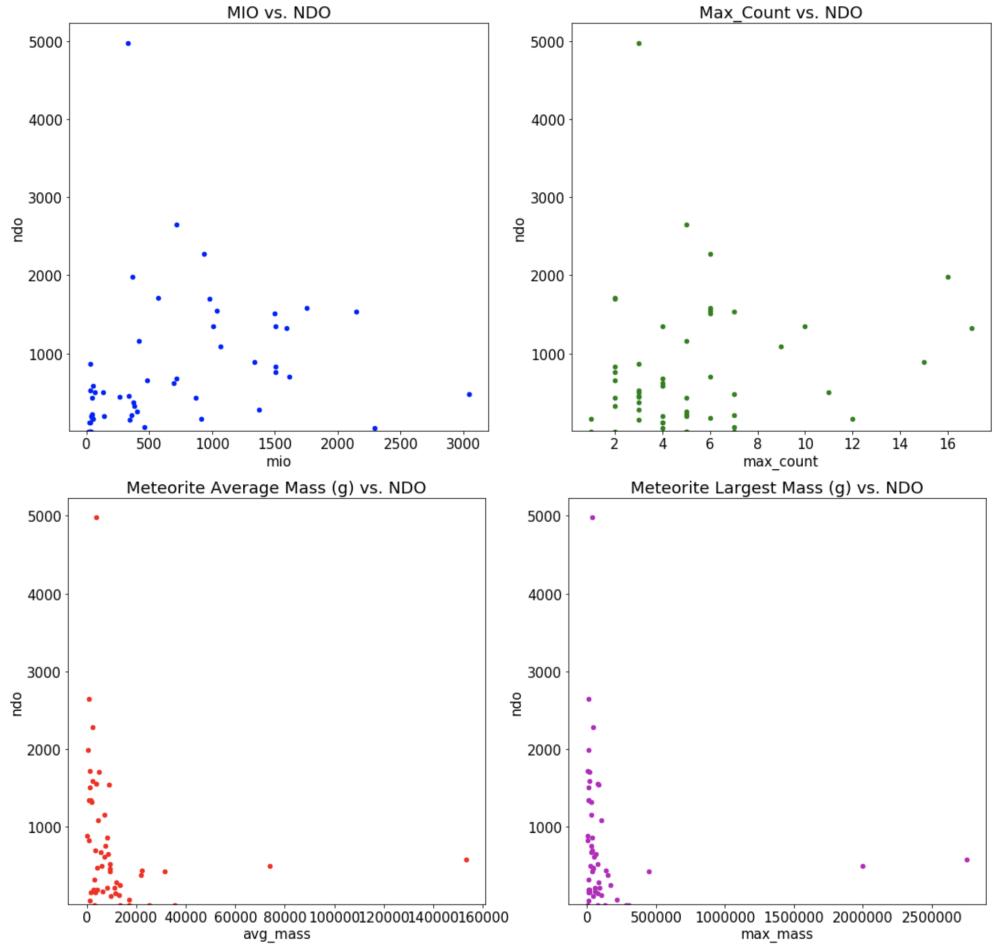

Figure 3: Scatter plots of explanatory variables against natural disaster occurrences on regular axes. The figure shows four scatter plots of meteorite vs. natural disaster occurrences (blue dots, top left), maximum count vs. natural disaster occurrences (green dots, top right), meteorite average mass vs. natural disaster occurrences (red dots, bottom left), and meteorite largest mass vs. natural disaster occurrences (purple dots, bottom right). Each dot is representative of the data recorded in a given year, with the corresponding data.

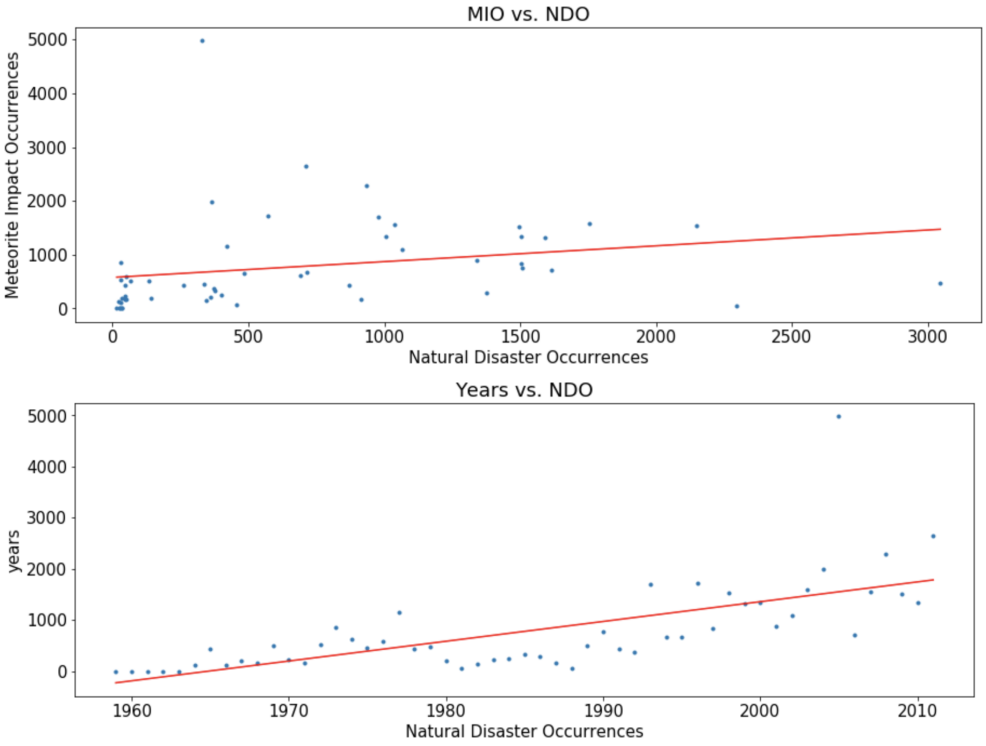

Figure 4: Top: Correlation between meteorite impact occurrences and natural disaster occurrences of each given year. The figure shows a scatter plot of meteorite vs. natural disaster occurrences (blue dots), as well as a linear regression line fit by least squares (red line). Bottom: Correlation between the year and natural disaster occurrences. The figure shows a scatter plot of years vs. natural disaster occurrences (blue dots), as well as a linear regression line fit by least squares (red line).

Bibliography

[1] Collins, G. S., Melosh, H. J., & Marcus, R. A. (2005). Earth impact effects program: A webbased computer program for calculating the regional environmental consequences of a meteoroid impact on Earth. *Meteoritics & planetary science, 40(6)*, 817-840.

[2] Rumpf, C. M., Lewis, H. G., & Atkinson, P. M. (2017). Asteroid impact effects and their immediate hazards for human populations. *Geophysical Research Letters, 44(8)*, 3433-3440.

[3] Schulte, P., Alegret, L., Arenillas, I., Arz, J. A., Barton, P. J., Bown, P. R., ... & Collins, G. S. (2010). The Chicxulub asteroid impact and mass extinction at the CretaceousPaleogene boundary. *Science, 327(5970)*, 1214-1218.