Alex Chansky
Econometrics Project

Introduction and Literature Review

Soil liquefaction is a secondary hazard during earthquakes where an applied load causes a block of soil to act as a liquid instead of a solid for a short period of time. This can cause extreme damage to overlying infrastructure, such as buildings or roads. Several models have been created to constrain the impact of different potential explanatory variables on the occurrence of liquefaction (Zhu et al. 2015, 2017, Rashidian and Baise, 2020). However, little work has been completed to directly model the impact of different potential explanatory variables on liquefaction damage costs. In this paper, a multiple linear regression loss model is presented based on liquefaction probability, the components which go into the liquefaction probability model, population density, and several other potential explanatory variables.

A global model to evaluate liquefaction spatial extent developed by Zhu et al. (2017) and implemented by the United States Geological Survey (USGS) identified and evaluated the impact of several globally available parameters likely to cause liquefaction: water table depth (Fan et al., 2013), annual mean precipitation (Hijmans and others, 2005), distance to nearest waterbody (HydroSHEDS and OceanColor), slope-based V_{s30} (Wald and Allen 2007). The model also utilizes two parameters unique to each earthquake event: peak ground velocity (ShakeMap, Worden and Wald, 2016), and peak ground acceleration (ShakeMap). These two parameters are calculated using data from seismic stations during an event and represent quantitative measures of ground motion in response to the earthquake itself. The calculated values for these two variables vary spatially. This model's output for each cell, liquefaction spatial extent (*LSE*), is

interpreted as a probability of that cell liquefying. However, very little work has established direct relationships between explanatory variables and liquefaction damage costs.

In a current thesis project, a database has been developed of liquefaction damage costs for damaged infrastructure and their locations in all US earthquake events from 1964 – 2019 (Chansky and Baise, in progress). 173 of the 295 damage descriptors have exact coordinates associated with them, accurate to within 100 meters. Using these coordinates, values were extracted from rasters of 16 potential explanatory variables, including the six used to produce the Zhu et al. (2017) output and the output itself (*LSE*). Some variables can be extracted from national or global datasets while others such as *Magnitude*, peak ground acceleration (*PGA*), peak ground velocity (*PGV*), and *LSE* are unique to each event. Some variables unique to each event such as PGA, PGV, and LSE, also vary spatially, meaning different values can be found for different locations in the same event.

Econometrically, the null hypothesis for this project was that no variable has any significant impact on cost. The alternative hypothesis was that one or more variables are statistically significant impact. For research purposes, the null hypothesis was that some variables are expected to correlate with damage costs and evidence found would offer support for or against keeping each variable. Every variable was tested for individual and group significance. As *LSE* is the best predictor of general liquefaction, it was expected to be significant and explain much of the variance in cost. Population density is the only indicator obtained as a proxy for infrastructure, so it was also expected to explain much of the cost variance. Most of the remaining explanatory variables were used to calculate *LSE*, so they are not expected to correlate better with liquefaction damages than *LSE* itself. *LSE*, *PGV*, and *PGA* are unique shaking parameters to each earthquake, and most liquefaction damage occurred in events with high levels

of shaking, so it can be expected that these three parameters impact cost. While *Magnitude* is also unique to every event, its values are calculated on an event-wide basis and is thus less indicative of approximated applied load at each location than *PGA* or *PGV*.

High multicollinearity was also expected for some of the variables used to construct *LSE*, so some variables were expected to be eliminated during the regressions.

Data description

Cost for each damage description is provided in 2018 US Dollars from the database constructed in Chansky and Baise (in progress). Costs for every instance of damage were either found in literature review for each individual earthquake or estimated using details in reconnaissance reports and strategies detailed in Chansky and Baise (in progress) developed primarily on HAZUS-MH 2.1 estimation techniques. For instances of cost estimation found in literature review for historical earthquakes, costs were adjusted to 2018 US Dollars using the US consumer price index (CPI). Costs have also been adjusted using the area modification factor (AMF) found in Moselle (2019) of the city or state in which the damage occurred. This accounts for discrepancies in labor and material costs between different regions of the country.

Population density (*Pop_sqmi*) was calculated from a polygon shapefile in the Tufts Data Lab's M Drive representing census count of population per square mile in every block group across the US. This polygon was converted to a raster from which cell values could be extracted at every location (Census Bureau, 2018).

Soil thickness (*Soil_thickness*) is calculated in each cell using an average of combined soil, regolith, and sedimentary rock deposit thickness, in meters from the surface to solid bedrock (Pelletier et al. 2016). Bedrock is unable to liquefy, so thicker soil layers allow the potential for more liquefaction to occur than thinner layers.

The remaining variables can be categorized into three components of soil liquefaction: applied load, soil density, and saturation.

Peak ground acceleration (*PGA*) and peak ground velocity (*PGV*) are shaking parameters unique to each earthquake, estimated spatially by algorithms at the United States Geological Survey (USGS) and represent **applied load**. These parameters are published online for each event and were downloaded individually for use in this project. Liquefaction spatial extent (*LSE*) indicates the probability that each cell will liquefy. This is also unique to each earthquake, and is calculated in an algorithm using *PGA*, *PGV*, and several global parameters. *LSE* is thus expected to have some degree of collinearity with the other layers, which was checked during the empirical analysis for anything problematic. LSE was also downloaded from the USGS website individually for each event through the Shakemap archives. As discussed earlier, *Magnitude* is another variable which can be considered for applied load. However, moment magnitude used in this project is calculated as the total energy released by an event. It is not indicative of shaking experienced at any individual location. Thus, *PGA*, *PGV*, and *LSE* are expected to correlate better with cost than *Magnitude*.

Shear wave velocity (*Vs30*), elevation (*Elev*), topographic slope (*slope*), topographic position index (*TPI*), terrain roughness index (*TRI*) can all be indicative of soil density. *Vs30* is the average velocity of a type of seismic wave known as a shear wave from the Earth's surface to a depth of 30 meters. Wald and Allen established a global model for Vs30 from which all values where extracted. Elevation (*Elev*) represents the cell's average elevation above sea level, measured in meters (Danielson and Gesch, 2011. The rasters representing slope (in degrees), terrain ruggedness index (TRI) (dimensionless), and topographic position index (TPI) (dimensionless) were all calculated from the elevation raster.

Aridity Index (*AI*), precipitation (*Precip*), compound topographic index (*CTI*), water table depth (wtd), and distance to the nearest water body (*DistWater*) are different indicators of saturation levels. *AI* represents the ratio between precipitation and water needs of vegetation. High *AI* values correspond to wet, humid regions while low *AI* values correspond to dry regions that are getting below their vegetation's ideal amounts of water. Wet regions are more susceptible to liquefaction, so high *AI* values (dimensionless) are expected to correspond positively with higher liquefaction damage costs. Global AI raster was downloaded from the Global Aridity Index and Potential Evapo-Transpiration Climate Database v2 (Trabucco and Zomer, 2018). *Precip* represents annual mean precipitation for an area and higher values indicate higher saturation, so *Precip* was expected to correlate positively with cost. CTI is a function of both slope and upstream catchment area compared to upstream width. This is meant to represent a wetness index describing how much potential a cell has to receive runoff from upstream regions. Cells with high CTI values (dimensionless) represent areas receiving a lot of runoff, which are more likely to be wet. These cells are more likely to have damage due to liquefaction, so we expect a positive association with damage costs (Verdin, 2017). Low values of *DistWater* indicate a location is close to a water body while high values of *DistWater* indicate the location is further away, so this variable was expected to negatively correlate with saturation and thus negatively correlate with cost.

Lastly, a dummy variable was added into the regression, for which a 1 was assigned to all locations with damage costs and a 0 was assigned to all locations without damage costs. This variable provides a way to separate observations of damage against those without damage. Summary statistics for values extracted for each explanatory variable can be found in **table 1**.

Another point of consideration when collecting data was our effort to be wary of bias towards locations with liquefaction damage. To account for potential bias towards locations with liquefaction damage, variable values from locations of non-liquefaction-damage (NLD) were included in some empirical regressions.

NLD points used in the regression were selected randomly from a mesh of points spaced evenly apart using geographic information system (GIS) software. In 12 events with liquefaction damages, meshes were constructed of approximately XX miles around each point of liquefaction damage with NLD points spaced 1 km apart. In five events of no liquefaction, meshes were constructed of approximately 40 km by 40 km with NLD points spaced 1km apart. which had extracted values randomly from areas of potential liquefaction which did not receive any liquefaction damage.

After eliminating points in the meshes which existed over water bodies, NLD points were randomly selected from the list of over 20,000 NLD points. A ratio was used of three NLD points to one liquefaction damage point as areas of no liquefaction damage are much more common in reconnaissance reports than areas of liquefaction.

Empirical part

After completing scatter plots showing relationships between cost and all independent variabels (figure 1), it was clear that there would not be linear relationships between any of the potential explanatory variables and the explained variable, cost. Scatter plots showing relationships between the natural log of cost and all independent variables (figure 2) appeared, at least visually, to allow some relationships to be established and significance of some coefficients to be proved.

Three regressions were completed using different combinations of variables with the intent of finding a combination which maximized the r-squared and adjusted r-squared. Most rasters in this study did not provide values over water bodies because their variables related to soil properties. The pixilation of low-resolution rasters near coastlines results in some land areas near water bodies where the rasters do not have real values associated with them. Extractions of these values for liquefaction damages near coastlines resulted in missing values, which cannot be used in a regression.

For the **first regression**, only location points with damage were considered. Furthermore, only variables with values at almost all locations were considered. This led to 152 observation points under consideration of seven variables; *LSE*, *Pop_sqmi*, *AI*, *Precip*, *Soil_thickness*, *Vs30*, and *DistWater*.

A regression was conducted normally for *lnCost*. Typically, *AI* and *Precip* are expected to have high multicollinearity, but the *vif* command did not reveal high collinearity for any variables for these observations.

AI, Precip, and *Vs30* were found to be individually significant, while the other four variables were not. Several combinations of F test were conducted to observe if any combination of the other four variables were found to be jointly significant. None of them were found to be jointly significant and inclusion of any combination decreased the adjusted R^2 value. The results of the regression of the three significant variables can be found in column 1 of **table 2**.

For the second and third regressions, all 16 variables were considered. However, several variables did not have values at many locations, so only 340 total point locations could be considered for these, of which 68 are damage observations. The second regression analyzed the 68 points of liquefaction damage on as many variables as possible while the third regression

analyzed the 68 damage observations and 272 NLD points for a total of 340. The fourth regression analyzed a reduced number of variables to try maximizing the number of observations which could be included of both damage and no damage.

In the **second regression**, *Precip* and *AI* were found to have very high multicollinearity. *Precip* was not individually significant while *AI* was, so *Precipitation* was removed. *TRI* and *slope* were also found to have problematic collinearity. *TRI* was slightly less significant than slope but is more complex than slope, so *TRI* was retained in the model. An F-test was conducted to test joint significance of *slope* and *Precip,* which concluded that the two variables were not jointly significant in this regression.

In the regression without *slope* and *Precip*, the following variables were not found to be individually significant: *TPI, TRI, LSE,* and *Pop_sqmi*. F-tests for all combinations of these, including *slope* and *Precip* yieled results which were not jointly significant. Removing most of these did not impact R-squared or adjusted R-squared much. However, removing *Pop_sqmi* reduced both R-squared and adjusted R-squared by more than 0.05, so *Pop_sqmi* was retained in the regression. Coefficients for remaining variables can be found in column 2 of **table 2**.

The **third regression** was conducted normally for *lnCost*. Through the *vif* command, *AI* and *Precip* were found to have high multicollinearity. Both variables were found to be individually insignificant. *AI* had a slightly lower p-value and is a more complex variable meaning it was expected to be slightly more indicative of real-world conditions, so *Precip* was eliminated. *TRI* and *slope* were also found to have high multicollinearity and individually insignificant. *TRI* was more significant and is a more complex variable so is expected to be slightly more indicative of real-world conditions, so *slope* was eliminated.

Nine variables were found to be individually significant, while *TPI, TRI, Magnitude*, *Elev,* and *AI* were not. Through several F-test combinations, all four variables were found to not be jointly significant. Inclusion of any combination was also found to decrease R^2 and adjusted R^2. For these reasons, all four variables were eliminated from the regression. Results of the regression of the nine individually significant variables can be found in column 3 of **table 2.**

For the **fourth** and final **regression**, a smaller number of variables was considered with the goal of maximizing the number of points which could be used in the analysis. *LSE, PGV, DistWater, Vs30, AI, and Soil_thickness* were considered for 618 points, of which 144 observations represent locations of damage.

No multicollinearity was found using the *vif* command. *Soil_thickness* was the only variable found to not be individually significant. However, its removal lowered the R^2 by approximately 0.015, corresponding to 1.5% more variation explained in *lnCost* by keeping it in the regression. *Soil_thickness* was determined important to the regression as it reduces bias in other variables, shown in the coefficient increase and subsequent p-value decrease of *LSE* and *AI* with its inclusion, and its greater explanation of variance in *lnCost*. Results of this regression can be found in column 4 of **table 2**.

Limitations

It is expected that some omitted variables could improve the regression. Authors behind Chansky and Baise (in progress) are also working on rasters representing density of different infrastructure, such as roads or buildings. It is likely that these variables alone or as an interaction term in combination with *LSE* will explain much of the variance associated with infrastructure damage costs.

It is also possible that different interaction terms, quadratic terms, or log terms of variables could improve the regression. Due to time constraints, only linear terms were checked for each explanatory variable, but this could be easily improved in future analysis.

The dataset used for the regression is also impacted by sample selection bias. When identifying points with damage locations from reconnaissance reports, exact locations could not be determined for expensive, multi-site damages. For example, from the 1989 Loma Prieta reconnaissance report, it is said, “300 pipe breaks” were concluded as having occurred due liquefaction across the Marina District. Though this results in one of the more expensive damage points in the database, it does not have a precise location associated with it, so values for explanatory variables could not be extracted and this damage could not be included in the regression. These regressions are thus likely biased towards low-cost damages, which is problematic, as discussed in class.

Another large issue with the current dataset is that the resolution for some rasters is coarse enough that areas near coasts and rivers are often all considered water. So if liquefaction damage occurs in a river or coastal area, their location is assigned a value of “NaN”. The simplest way these regressions could be improved is through ensuring that all rasters of explanatory variables exist in a high enough resolution where data can be extracted at every damage location. Alternatively, for locations of missing values near coastlines, data could be extracted on the raster at the location closest to where damage occurred.

Another way the regression could be improved is through obtaining additional potential explanatory variables to explain more of the variance in the natural log of damage costs. Some possibilities for other explanatory variables are road density, building density, or other

infrastructure variables which could be direct or proxy variables for infrastructure exposed to liquefaction.

Summary and conclusion

The preferred regression is the one explaining the highest variance in the natural log of cost while including NLD points. This is the **third** regression containing all variables and 340 total points, 68 of which are observations with damage costs associated with them. While the second regression had a higher R-squared value, it is essential to sample some points without liquefaction damage. Additionally, the second regression does not consider *LSE* or *Pop_sqmi* as significant variables, when they were expected to be two drivers of damage.

Based upon the results in the third regression, nine variables are found to be at least significant at the 0.10 level, including one which was significant at the 0.05 level and five which are significant at the 0.01 level. From this, it can be concluded that *PGV*, *LSE*, *Soil_thickness*, *CTI*, and water table depth all have **very significant impacts** on liquefaction damage costs and *PGA* has a **significant impact**. While *PGA*, *PGV*, and *LSE* are unique to each earthquake, the remaining three variables are not. Thus, areas with thick soil layers between the surface and bedrock, high CTI, and low water table depth can be considered more vulnerable to liquefaction infrastructure damage than other areas. The impact of *PGA*, *PGV*, and *LSE* on cost was expected based on discussion in the introduction. While less easily predictable, obtaining these shaking parameters soon after an earthquake occurs can indicate areas of potential liquefaction damage.

Population density, expected to be one of the larger drivers of cost, was declared somewhat significant in the preferred regression and not significant in the other models. More accurate population density rasters are available and will likely improve those results in future

models. However, these rasters cannot be downloaded nationally and a fair amount of processing must be done to prepare the more detailed data, so it was not considered due to time constraints.

The R^2 for the preferred model, 0.256, is still quite low. This implies that all variables discussed only account for 25.6% of the variance in the natural logarithm of damage costs. Including some currently omitted variables will likely improve this result.

The model could be improved by several methods discussed in the limitations section. It is the author's opinion that at least some of these improvements should be made before any policy changes are made in response to these results.

Bibliography

Baise, L.G., Rashidian, V. (2020). Regional Efficacy of a Global Geospatial Liquefaction Model. *Engineering Geology.* Vol. 272

Danielson, J.J., Gesch, D.B., (2011). *Global multi-resolution terrain elevation data 2010 (GMTED2010).* USGS Publications Warehouse 10.3133/ofr20111073 http://pubs.er.usgs.gov/publication/ofr20111073

Fan, Y., Li, H., and Miguez-Macho, G., 2013, Global Patterns of Groundwater Table Depth: Science, 339, 940-943.

HAZUS Multi-Hazard Loss Estimation Methodology Technical Manual, Version 2.1, *Department of Homeland Security: Federal Emergency Management Agency,* Washington, DC, 2017

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., and Jarvis, A., 2005, Very high resolution interpolated climate surfaces for global land areas: International Journal of Climatology, 25(15), 1965–1978.

Moselle, B. (2019). 2019 National Building Cost Manual. *Craftsman Book Company.* 43rd Edition. https://www.craftsman-book.com/media/static/previews/2019_NBC_book_preview.pdf

Pelletier, J.D., P.D. Broxton, P. Hazenberg, X. Zeng, P.A. Troch, G. Niu, Z.C. Williams, M.A. Brunke, and D. Gochis. 2016. Global 1-km Gridded Thickness of Soil, Regolith, and Sedimentary Deposit Layers. ORNL DAAC, Oak Ridge, Tennessee, USA. http://dx.doi.org/10.3334/ORNLDAAC/1304

Trabucco, A., and Zomer, R.J. 2018. Global Aridity Index and Potential Evapo-Transpiration (ET0) Climate Database v2. CGIAR Consortium for Spatial Information(CGIAR-CSI). Published online, available from the CGIAR-CSI GeoPortal at https://cgiarcsi.community

U.S. Census Bureau (2018). *Selected housing characteristics, 2017 American Community Survey 1-year estimates*. Retrieved from Tufts M Drive.

United States Geological Survey: Ground Failure Scientific Background (2020). Retrieved from https://earthquake.usgs.gov/data/ground-failure/background.php.

Verdin, K.L., 2017, Hydrologic Derivatives for Modeling and Applications (HDMA) database: U.S. Geological Survey data release, https://doi.org/10.5066/F7S180ZP.

Wald, D.J., and Allen, T.I., 2007, Topographic Slope as a Proxy for Seismic Site Conditions and Amplification: Bulletin of the Seismological Society of America, 97 (5), 1379–1395.

Worden, C.B. and D.J. Wald, 2016, ShakeMap Manual Online: technical manual, user's guide, and software guide: U. S. Geological Survey.

Zhu, J., Daley, D., Baise, L.G., Thompson, E.M., Wald, D.J., Knudsen, K.L. A (2015). A Geospatial Liquefaction Model for Rapid Response and Loss Estimation. *Earthquake Spectra*, **31** (3), 1813-1837.
doi: http://dx.doi.org/10.1193/121912EQS353M.

Zhu, J., Baise, L.G., and Thompson, E.M. (2017). An Updated Geospatial Liquefaction Model for Global Application*, Bull. Seism. Soc. Am*. 107 (3), doi: 10.1785/0120160198

Appendix

Figure 1: Scatter plots of relationships between Cost and all other independent variables.

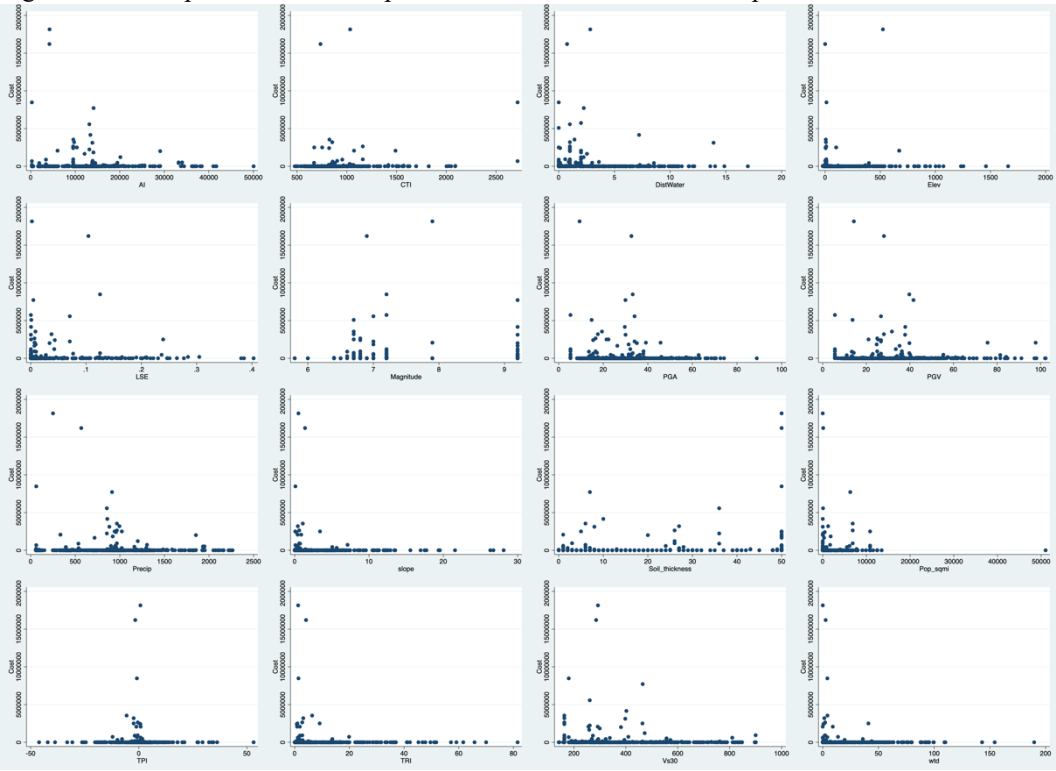

Figure 2: Scatter plots of natural logarithm of cost versus all other variables in linear format.

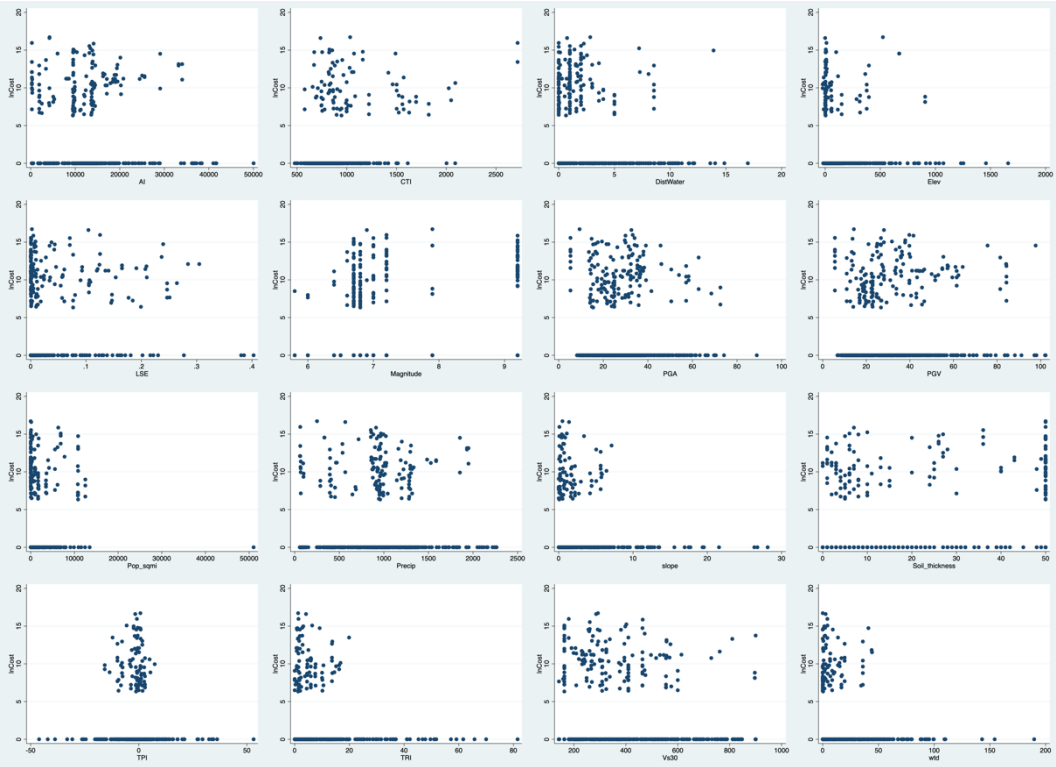

Table 1: summary statistics of Explained variables and

Summary Statistics

VarName	Obs	Mean	SD	Min	Median	Max
Cost	688	177393.29	1.14e+06	0	0	1.81e+07
lnCost	688	2.67	4.803	0	0	16.71405
Magnitude	688	7.47	1.071	5.8	6.8	9.2
PGA	688	28.33	12.720	5.27166	26.2604	88.8684
PGV	688	29.66	17.372	5.58303	25.16845	102.269
LSE	688	0.02	0.055	0	.0008953	.402934
Pop_sqmi	597	1441.85	3320.733	.1	114.2	51016.7
AI	656	10850.19	7717.624	227	11261	50052
Precip	656	850.71	460.760	56	914.5	2264
Soil_thickness	622	18.22	19.506	0	8	50
Vs30	672	452.10	215.781	141.38	416.655	900
DistWater	688	2.91	2.816	0	2	16.9706
Elev	421	161.85	247.001	-22	67	1660
CTI	414	919.48	296.776	477	854.5	2717
TPI	421	-0.47	9.324	-46.125	-.125	53.125
TRI	421	9.70	12.317	0	5.375	81.5
slope	421	2.98	4.074	0	1.48898	28.1231
wtd	420	21.97	26.699	0	11.8082	189.75
Dummy_Damage	688	0.25	0.433	0	0	1

Table 2: Output for regressions including coefficients, standard errors, p-values, number of observations, and variance explained (R^2).

VARIABLES	(1) Damage Points Only Reduced Variables	(2) Damage Points Only All Variables	(3) All Points All Variables	(4) Reduced Variables Maximized Points
lnCost				
Magnitude		4.155***		
		(1.168)		
		0.000768		
PGA		-0.0206	-0.0614**	
		(0.0411)	(0.0259)	
		0.618	0.0183	
PGV		-0.0423	0.0813***	0.0425***
		(0.0339)	(0.0218)	(0.0105)
		0.217	0.000230	6.03e-05
DistWater		-0.233	-0.159*	-0.285***
		(0.247)	(0.0836)	(0.0670)
		0.351	0.0573	2.50e-05
Pop_sqmi		-2.88e-05	9.32e-05*	
		(6.41e-05)	(4.97e-05)	
		0.655	0.0615	
AI	0.000271***	-0.000226		0.000112***
	(5.45e-05)	(0.000139)		(2.57e-05)
	1.73e-06	0.110		1.65e-05
Soil_thickness		-0.0650**	-0.0416***	-0.0155
		(0.0264)	(0.0154)	(0.0129)
		0.0171	0.00725	0.228
Vs30	-0.00386***	-0.0153***	-0.00294*	-0.00532***
	(0.00127)	(0.00349)	(0.00177)	(0.00114)
	0.00290	5.16e-05	0.0969	3.84e-06
CTI		-0.00264***	0.00345***	
		(0.000899)	(0.000790)	
		0.00475	1.70e-05	
wtd		0.0542	-0.0253***	
		(0.0354)	(0.00941)	
		0.131	0.00755	
Elev		0.00358		
		(0.00287)		
		0.217		
Precip	-0.00387***			
	(0.000933)			
	5.53e-05			
LSE			21.67***	12.29***
			(6.568)	(4.388)
			0.00108	0.00525
Constant	12.20***	-5.311	0.747	3.243***
	(0.574)	(7.881)	(1.247)	(0.799)
	0	0.503	0.550	5.59e-05
Observations	152	68	340	618
R-squared	0.170	0.484	0.256	0.179

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Do File

```
* Importing csv
*insheet using out.csv, clear
import delimited "/Users/Alex/Box/My_Work/Fall_2020/Econometrics/FinalProject/Data/ExportedValues_CSVs
> /out4.csv", case(preserve)


* Reading Stata (.dta) file
*use BWGHT, clear

describe
summarize

reg Cost Magnitude PGA PGV LSE Pop_sqmi AI Precip Soil_thickness Vs30 DC DR CTI TPI TRI wtd
reg Cost Magnitude PGA PGV LSE Pop_sqmi AI Soil_thickness Vs30 DistWater CTI TPI TRI wtd PopExp
reg Precip Vs30 wtd PGA PGV LSE DistWater Pop_sqmi AI Soil_thickness Dummy_Damage


graph matrix Cost Magnitude PGA PGV LSE Pop_sqmi AI Precip Soil_thickness Vs30 DC DR CTI TPI TRI wtd

* Damage points only, reduced variables
reg lnCost LSE Pop_sqmi AI Precip Soil_thickness Vs30 DistWater if Dummy_Damage == 1
vif
test DistWater Soil_thickness LSE Pop_sqmi LSE
test DistWater Soil_thickness
test LSE Pop_sqmi
reg lnCost LSE Pop_sqmi AI Precip Vs30 DistWater if Dummy_Damage == 1
reg lnCost LSE Pop_sqmi AI Precip Soil_thickness Vs30 if Dummy_Damage == 1
reg lnCost AI Precip Vs30 if Dummy_Damage == 1

outreg2 using myreg2.doc, replace ctitle(Damage Points Only, Reduced Variables) stats(coef se pval)

* ADDED THE FOLLOWING REGRESSION
* Damage Points only, all variables
reg lnCost Magnitude PGA PGV DistWater LSE Pop_sqmi AI Precip Soil_thickness Vs30
CTI TPI TRI wtd slope Elev if Dummy_Damage==1
vif
test TRI Precip
test Precip TRI TPI
test Precip TRI TPI LSE Pop_sqmi
test Precip TRI TPI LSE Pop_sqmi slope
reg lnCost Magnitude PGA PGV DistWater AI Soil_thickness Vs30 wtd Elev
CTI TPI wtd slope Elev if Dummy_Damage==1
reg lnCost Magnitude PGA PGV DistWater Pop_sqmi AI Soil_thickness Vs30 CTI wtd Elev if
Dummy_Damage==1
* kept Pop_sqmi because removing it lowered adjusted R-squared by 0.06, goal was
* to keep adjusted R-squared high
outreg2 using myreg2.doc, append ctitle(Damage Points Only, All Variables) stats(coef se pval)

* All points and all variables
reg lnCost LSE Vs30 wtd PGA PGV DistWater Pop_sqmi Soil_thickness CTI
reg lnCost Magnitude PGA PGV DistWater LSE Pop_sqmi AI Precip Soil_thickness Vs30 CTI TPI TRI wtd slope
Elev
```

```
vif
reg lnCost Magnitude PGA PGV DistWater LSE Pop_sqmi AI Soil_thickness Vs30 CTI TPI TRI wtd Elev
vif
test TPI TRI
reg lnCost PGA PGV DistWater LSE Pop_sqmi Soil_thickness Vs30 CTI wtd Magnitude AI
reg lnCost Magnitude PGA PGV DistWater LSE Pop_sqmi AI Soil_thickness Vs30 CTI TPI TRI wtd
test TPI TRI Magnitude
test TPI TRI Elev
test TPI TRI Magnitude AI
test Magnitude AI
test Elev AI
reg lnCost PGA PGV DistWater LSE Pop_sqmi Soil_thickness Vs30 CTI wtd

outreg2 using myreg2.doc, append ctitle(All Points, All Variables) stats(coef se pval)

* All points, reduced variables
reg lnCost LSE PGV DistWater Vs30 AI Soil_thickness
vif
test Soil_thickness DistWater
reg lnCost LSE PGV DistWater Vs30 AI
reg lnCost LSE PGV DistWater Vs30 AI
reg lnCost LSE PGV DistWater Vs30 AI Soil_thickness
outreg2 using myreg2.doc, append ctitle(Reduced Variables, Increased Points) stats(coef se pval)

sum2docx Cost lnCost Magnitude PGA PGV LSE Pop_sqmi AI Precip Soil_thickness Vs30 DistWater Elev CTI
TPI TRI slope wtd Dummy_Damage using Alex, replace stats(N mean(%9.2f) sd min(%9.0g) median(%9.0g)
max(%9.0g))


* Creating scatter plot example
scatter lnCost TRI
```