Predicting Societal and Economic Impact of Future Natural Disasters

Laura Griffiths, Chang Liu, and Divneet Mandair

***Abstract*—While anticipating the outcomes of natural disasters is a growing area of interest to a diverse body of organizations, little formal work has been done to make these predictions for specific countries. We propose a novel methodology that combines supervised learning on historical disaster and demographic data with intelligent clustering to expand a relevant training set for particular countries and disaster types. By clustering countries at snapshots in time, based on developmental stages and 'disaster event profiles,' we develop more specific sets of historical data that improve predictions on the number of people killed and financial cost of various disaster types.**

I. Introduction

Recent natural disasters have driven much research effort towards developing predictive models of when and where disasters will strike. An equally important aspect of disaster anticipation, however, is understanding the predicted impact of such disasters for particular countries. Organizations from the UN to Google have devoted considerable resources towards understanding the countries most at risk of high disaster impact[1]. Developing this 'risk-profile' for different countries allows international organizations to effectively tailor their aid efforts.

Few attempts, however, have been made to systematically predict the societal and economic impacts of specific natural disasters within particular countries. Part of the challenge with developing this type of model is the scarcity of historical data, for a given country, of particular disasters. Moreover, 'impact' of a natural disaster varies substantially with the development stage of a country, such that historical disaster data may no longer be particularly useful as a training instance for a country that has developed significantly over time. To resolve this, we propose a novel methodology that predicts the impact of natural disasters, for a given country, based on that country's own history of disasters of the same type and 'similar' countries that have experienced the same disaster. Our end goal is to answer the question: given a particular country and disaster that is expected to occur, can we reliably predict that disaster's impact on the country, measured by both economic and societal impact.

II. Data Description & Pre-Processing

For historical disaster data, we use the EM-DAT dataset (EM), which collects information on disasters from 1900-2008 across all countries. An event is classified as a disaster if one or more of the following criteria is met: 10 or more people reported killed, 100 or more reported affected, declaration of a stage of emergency, or call for international assistance. Features included in the dataset consist of the number of people killed (killed), financial cost (cost), country, date of the event, and a few others. We expand this set to also include a region and subcontinent code for the country of each disaster's occurrence.

Demographic information for countries, which can be used to categorize similarly developed nations, was obtained from the World Bank (WB) [2]. This data is organized into 18 categories, ranging from Climate Change to Infrastructure, with over 1300 distinct features across all topics.

Both datasets were restructured such that a row, instead of representing a single country with values for its attributes over time, is represented as a country/year pair. This facilitates comparisons of countries that, although at different periods of time, may be at similar developmental stages. Integrating both datasets with this structure yields a 12,948 x 1300 sparse design matrix.

Our analysis focuses on predicting the EM features 'killed' and 'cost' as measures of societal and economic impact.

III. Baseline Model

To develop a baseline prediction for each disaster type, initial supervised learning models consisted of either the entire EM or the combined EM and WB datasets. Linear combinations of features in these datasets were tested, such that

$$h_\theta(x) = \sum_{i=0}^{n} [\theta_i x_i)] . \quad (1)$$

Higher order polynomials, while also examined, had much poorer cross-validated mean squared test errors, likely due to overfitting. Baseline results are depicted in Table 1, for predictions on killed only.

Table 1

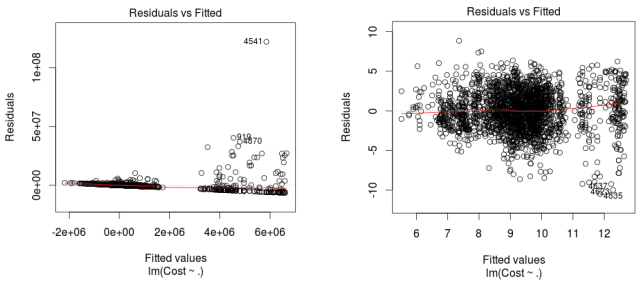
KILLED	**EM**	**EM+WB**
Epidemic		
Train R^2	0.2501934	0.2235237
Test MSE	13140730	12957340
Storm		
Train R^2	0.0926529	0.2248343
Test MSE	103318000	144051300
Flood		
Train R^2	0.1272947	0.1857909
Test MSE	588021.8	697602.9
Quake		
Train R^2	0.06555992	0.1626838
Test MSE	193371800	204932400

First, we notice the poor predictive power of using either the EM disaster features or combined EM and WB features in a linear model to estimate killed (and cost). The following are insights from our baseline analysis

- The EM dataset is sparse for a particular country/year entry and a given disaster type.
- Training, based on EM features, across the entire set of country-year pairs, as expected, leads to estimates of θ that are less likely to reflect a particular country's risk level.
- Naively combining WB features with EM features worsens the model's performance. While we might expect demographic data to be correlated with disaster outcomes, training over the entire range of data appears to hide the added value of country-specific information.

One issue immediately addressed was the underlying assumption of normality in our linear models. The plot below shows the residuals error in predicting the number of people killed for the flood disaster type. Using a log transformation of the response variable (both killed and financial cost) addressed non-normality of the errors and greatly enhanced both R^2 and testing errors in the linear models, across all disaster types.

Pre-log transformation **Post-log transformation**



IV. Further Processing: Missing Values & Imputation

To address the sparsity of our design matrix, we evaluated numerous imputation techniques. Prior to implementing this, however, it was necessary to develop a heuristic to remove rows and/or columns with excessive missing values. This helps reduce the number of necessary imputations, which can drastically skew predictions. To determine the acceptable thresholds for missing values in rows and/or columns, we iterated over a broad range of thresholds for each disaster type. The thresholds i and j, where i corresponds to a row threshold and j corresponds to a column threshold of our design matrix X, were chosen such that

$$[i, j] = \text{argmin}[(\vec{y} - \theta^T X_{i,j})^2] \quad (2)$$

where y is a vector of responses, either killed or cost. The overall result was a series of matrices with an optimized amount of missing data across rows and columns, specific to each disaster type. For most matrices, a row threshold of 0.6 (i.e. up to 60% of data entries missing), and column threshold of 0.7 was found optimal. We then evaluated both SVD and Kth nearest neighbor (KNN) imputation methods[3] on these outputs. Ultimately, KNN-imputed matrices contained fewer 'outlier' values in its imputed values.

V. Feature Selection

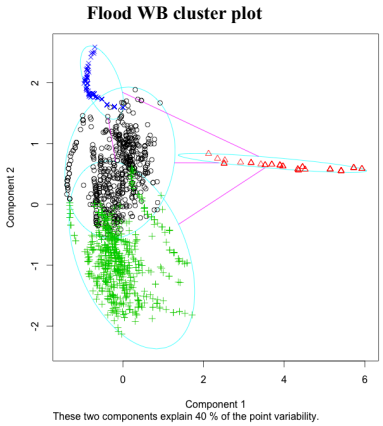
One key aspect of our prediction system is developing clusters of countries at similar development stages with *comparable risk profiles*. Thus, we do not want to simply cluster country-year entities based on all their disaster and demographic features, but rather those features that most relate to a given disaster event. We used backward selection, with a linear model, to isolate those features (across the EM and WB datasets) most predictive of disaster outcomes, for each particular disaster type. From the initial list of over 300 EM and WB features, the final number of features after all processing ranges from 30-60 across disaster types.

VI. Clustering

By clustering countries at similar risk and developmental stages, we aim to expand the set of historical disaster data available to make predictions on a specific disaster event for a particular country. Prior work has looked to classify countries by a development index [4]. However, our analysis differs from this by 1) using country-specific demographic information to inform the cluster groups and 2) allowing countries at different time points to be grouped into the same cluster. The added flexibility of 2) may allow for less intuitive combinations of particular countries.

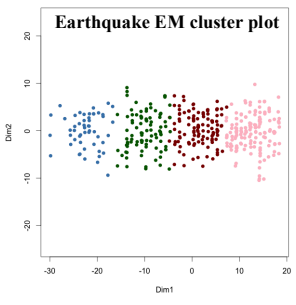
We first looked to cluster country/year pairs, for a given disaster type, based on WB features. We chose k-means as our clustering algorithm, as we view countries as belonging to distinct groups rather than probabilistically weighted across multiple groupings. Within sum of squares across cluster groups suggested the WB features partition country/year pairs into roughly 4 groups, a result that was consistent across all disaster types. The plot below, for flood WB features, depicts clusters along the first two principal components of the data. These components capture approximately 40% of the variability in the WB features for floods. The 4 distinct

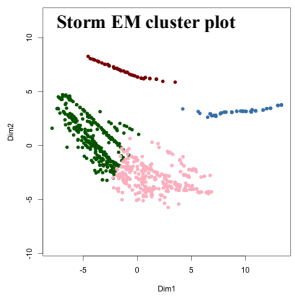
groupings confirms the within sum of squares analysis, so we chose to use 4 WB clusters for each disaster type. These groups, more qualitatively, correspond to pools of similarly 'developed' countries at particular snapshots in time.



Clustering using the unaltered WB features, however, results in skewed groupings, with nearly all the country/year pairs bucketed into 1 group. Possible explanations include the high number features (approximately 40-60 for different disaster types) and the high linear dependence among attributes. In particular, many of the features that persist in the WB set, even after backward selection, are highly correlated. Examples of such features include population density and number of live births. Clustering quickly becomes unreliable with a high number of noisy features [5]. To address this and remove co-linearity in the attributes, we employed PCA. Across all disaster types, 5 principal components consistently captured over 50% of the variability in the WB feature vectors corresponding to each disaster type. To facilitate comparison between clusters within different disaster types, all WB features were reduced to 5 principal components. Using these dimensionality-reduced components gave more reasonable cluster groups, in agreement with the plots above.

Finally, we also explored partitioning country/year pairs based solely on disaster features. Again, we used k-means as our algorithm. The plots below, for earthquake and storm disasters, rescales the kmeans distances of EM features to 2-D and then plots the clusters. Interestingly, as the plot suggests, country/year pairs can be partitioned into 4 distinct groups, similar to WB features. This partitioning along solely disaster characteristics indicates, that, independent of how developed a country is, there exist different 'disaster event profiles' for a given disaster type. To test this notion, we then employed 2 distinct clustering mechanisms for country/year pairs: 1) clustering solely off of the (dimensionality-reduced) WB features and 2) independently clustering by WB features and EM disaster features (using 4 clusters, for each). Our final analysis evaluates both mechanisms in our prediction system.





The pseudo code below summarizes the joint EM and WB cluster approach:

```
// Process both WB and EM dataset

JointX = join(WBImputed,EDImputed)
for i in 1: numDisasters {
  // get best feature subset using a linear model
  featSubset = featureSelection(JointX, disaster[i])
  subset = project(featSubset, JointX)

  // PCA
  pcaFeat = princomp(subset[,WB_feat])[,1:5]

  //Clustering using WB PCA features
  WBCInx = kmeans(pcsFeat, numWBClusters)

  //Clustering using EM features
  EMCInx = kmeans(subset[,EM_feat],numEMClusters)
  // Include clustering information into design matrix
  matrix = colBind(subset, WBCInx, EMCInx)

  for each WBClusterIndex {
    for each EMClusterIndex {
      // create model
      killedModel = linearModel(matrix, matrix$Killed)
      costModel = linearModel(matrix, matrix$Cost)
      // cross validation
      crossVal(matrix, killedModel, k=10)
      crossVal(matrix, costModel, k=10)
    }
  }
}
```

VII. Results & Analysis

A. *Model Choice*

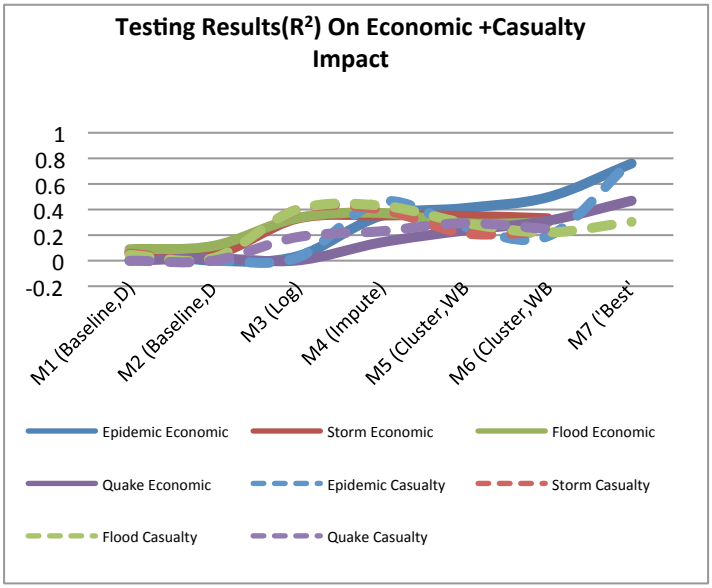
Once all processing and clustering steps were completed, we began making predictions on the two primary outcomes for a given disaster type: number of people killed and cost. Note, these predictions were made within clusters of countries with similar risk and developmental profiles. Features, as previously described, consisted of dimensionality-reduced WB features and disaster features (both narrowed by backward selection to obtain a feature set more highly correlated with a given disaster type).

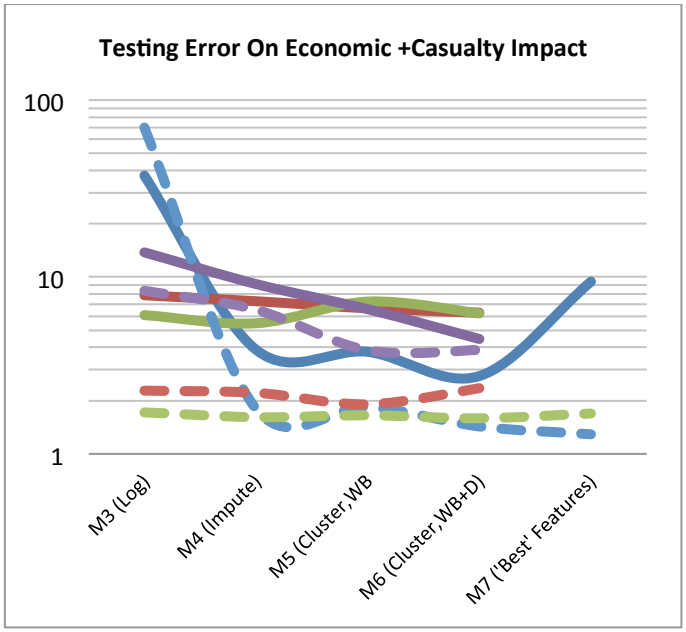
Second and third degree polynomials, for both predictions on cost and killed, drastically overfit the training set. Cross-validated test errors for these models were significantly larger than those from simpler linear models. In predicting the number of people killed, however, two alternative models were tested. Since number killed is a form of count data, Poisson regression was implemented for each disaster type. With our updated design matrix X and parameters θ, the predicted mean of the associated Poisson distribution and likelihood are as follows:

$$E[Y \mid x] = \exp(\theta^T x)$$
$$\ell(\theta \mid X, Y) = \sum_{i=1}^{m} (y^{(i)} \theta^T x^{(i)} - \exp(\theta^T x^{(i)})) \quad (3)$$

Surprisingly, the Poisson fit, across disaster types faired poorer than a base linear model. Again, comparisons were made using mean squared errors on cross-validated test sets. One reason for the poorer fit of the Poisson may be the distribution's imposed equivalence of the mean and variance. Particularly for disaster outcomes, where variation is high even within a selected disaster type, this assumption does not appear to be valid. We then attempted a negative binomial fit, which allows for a gamma-mixture in the Poisson's λ rate. While the negative binomial did outperform the Poisson regression models (allowing for a dispersion parameter $\Phi > 1$ in most cases), base linear models continued to outperform these more sophisticated models. Due to its consistent performance, the following analysis was performed with our original regression hypothesis. The overall result of our prediction system is a series of linear models, each specific to a disaster type, that are trained off clusters of similar countries and can be used to make a prediction for a given country/year pair.

The two plots below show the progression of the testing R^2 and mean-squared error as components are added to our prediction system.





In general, while R^2 measures range from moderate to poor, most of the components added to the system consistently reduced mean squared errors in test sets. This is most true in predicting the casualty from epidemics (the dotted blue line in the plot above). Key insights from our results are discussed more below.

B. Type of Clustering & Usage in Prediction Process

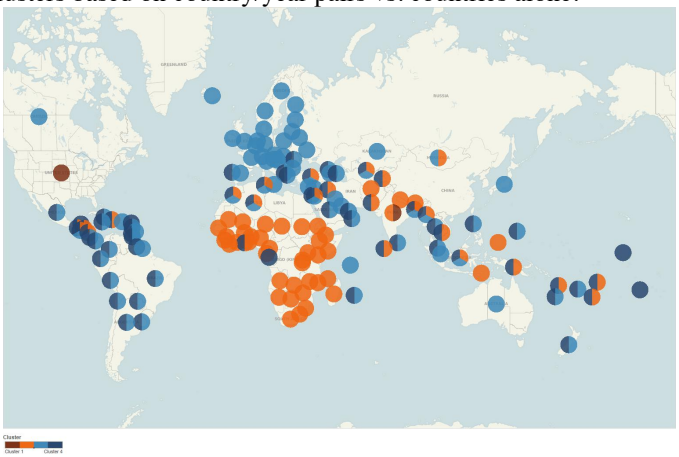
The two types of clustering performed in our analysis are 1) based solely on WB features (M3 in above plots) and 2) based on both WB and EM disaster features (M4 in above plots).

In case 2), country/year pairs are first assigned to cluster i_{WB}, based on WB features and then an independent clustering run assigns the country/year pairs to cluster j_{EM}, based on EM features. Predictions are made within each combined EM and WB cluster (i_{WB}, j_{EM}).

For both estimated generalized error (2.27 vs. 2.31) and R^2 (0.39 vs. 0.26), approach 2) outperforms 1). This confirms our clustering diagnostics, indicating that country/year pairs can be independently sorted based on 'disaster event profiles' and 'development stage' within a given disaster type.

In order to more deeply understand the physical structure underlying our clustering design, we visualized the WB clustering index assigned to countries through the years in the map below. Colors in the map represent cluster indices, and countries that experienced a development stage shift (i.e. were assigned to different WB clusters in different years) are represented by multiple colors. One insight is that many of the clusters agree with intuition: for instance, Europe, along with Japan, Australia and a few other countries, is clustered separately from central/southern Africa. Additionally, we see a variety of changes in cluster assignments. Central and South America seem to shift from being clustered with parts of Europe to its own developmental group over time. This

shifting pattern lends support for our decision to develop clusters based on country/year pairs vs. countries alone.



To further analyze the makeup of our joint EM and WB clusters (i_{WB}, j_{EM}), we simulated our prediction process, where a country, disaster type, time of occurrence were given as inputs. Imagine, for instance, we were interested in predicting the number of people killed if an epidemic were to happen in Ethiopia, in 2008. Supplying these as inputs (along with EM features such as month of occurrence, duration, etc. of the epidemic), our prediction system could then grab the most recent year, 2008 or prior, of WB features for Ethiopia. With both EM and WB features (dimensionality reduced by PCA), the system then assigns Ethiopia-2008 to an appropriate joint EM/WB cluster ($i_{WB\text{-}Eth}$, $j_{EM\text{-}Eth}$), by minimizing the Euclidean distance of this joint feature vector to each cluster's centroid. The map below highlights the countries that would be included in the extended training set for the causality 'prediction' for an epidemic in Ethiopia in 2008. As expected, other central and South African nations are included in the cluster. Interestingly, Afghanistan and Pakistan are also assigned to ($i_{WB\text{-}Eth}$, $j_{EM\text{-}Eth}$). Our system thus appears to intelligently use 'similarly' at-risk and developed countries to enhance a prediction for any particular country.

C. *Differences in accuracy across disaster type*

The accuracy of predictions vary greatly along disaster type. The linear models for estimating the number of people killed from epidemics, for instance, have an average R^2, across all clusters, of 0.78 and average mean squared test error of 1.29 This contrasts with the linear models for storm killed predictions, which have an average R^2 of 0.23 and average mean squared test error of 2.4. One reason for this large variation in performance may relate to the nature of the disaster event itself. We'd expect epidemics to occur only in certain countries, classified at a particular development stage. Moreover, when they do occur, they likely will have a consistent, quite harmful, impact on the countries they occur in. Storms, on the other hand, occur universally in almost every country. Their impact will vary highly, depending on a variety of intricate factors. We see through our clustering these intuitions are supported. When sequentially assigning country/year pairs to clusters i_{WB} and j_{EM} for epidemics, most country/year pairs (>100) are bucketed into a particular pair (i_{WB}, j_{EM}) = (1,2). Conversely, when the same clustering mechanism is applied for storm disasters, country/year pairs are more evenly dispersed across a variety of (i_{WB}, j_{EM}) pairs. Thus, the economic and societal impacts of epidemics are more consistently associated with a particular country 'risk profile' and 'development stage,' enhancing the model's predictive power.

VIII. Conclusion

Our goal was to develop a methodology to aid in the prediction of outcomes of natural disasters for particular countries. We showed that pools of countries with similar demographic and disaster event features, can be used in a supervised learning problem to enhance predictions when compared to more naïve models. While accuracy in our predictions varied, many of our estimates are typical for natural disaster prediction [6]. Ultimately, this analysis is but one step in a broader effort to quantifiably identify those countries most in need of aid when catastrophe strikes.

Acknowledgment

We would like to thank the CS 229 teaching staff for helping us frame our problem and providing insight along the way.

References

[1] "New risk index helps identify vulnerability," IRIN news, http://www.irinnews.org/report/93658/disasters-new-risk-index-helps-identify-vulnerability

[2] http://data.worldbank.org/topic

[3] Weber, Michael, and Michaela Denk. "Imputation of Cross-Country Time Series: Techniques and Evaluation." *European Conference on Quality in Official Statistics*. 2010.

[4] Asher, Jana, and Beth Osborne Daponte. *A hypothetical cohort model of human development*. United Nations Development Programme, 2010.

[5] Steinbach, Michael, Levent Ertöz, and Vipin Kumar. "The challenges of clustering high dimensional data." *New Directions in Statistical Physics*. Springer Berlin Heidelberg, 2004. 273-309

[6] Trevon L. Fuller et. al, "Predicting hotspots for influenza virus assortment," *Emerging Infectious Diseases*, vol. 19, no. 4, April 2013.