For office use only
T1 _____
T2 _____
T3 _____
T4 _____

Team Control Number
24005

Problem Chosen
C

For office use only
F1 _____
F2 _____
F3 _____
F4 _____

2013
Mathematical Contest in Modeling (MCM/ICM) Summary Sheet

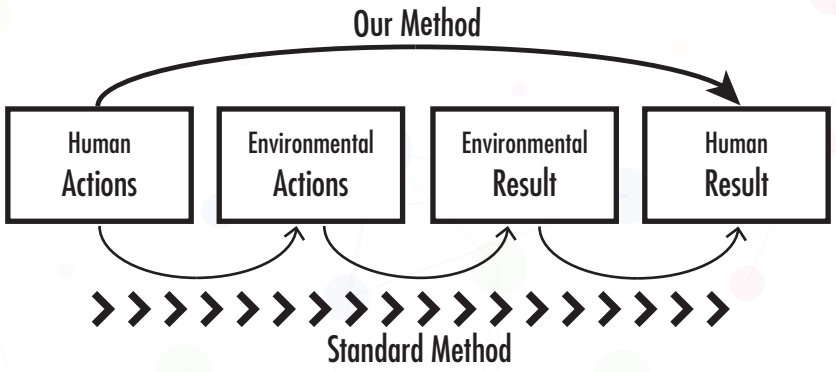
Saving the Green with the Greens

Problem Clarification

With environmental doom impending on us, governments across the globe are trying to find the best way to combat this fate. Unfortunately, they do not have access to the kinds of models necessary for the kind of decision-making they wish to achieve, due to a lack of understanding the direct human element in the situation. We based our model entirely on human relationships and influences: the thing policy makers have the most control over.

Model Design

To be especially relevant to each country's decision makers, our model directly predicts human results (measured in 2013 US dollars) with economic variables that are easily influenced by legal policy. We further improve our initial design by incorporating geographic proximity, diplomatic relations, and clustering data into a network model. All parameters were derived entirely from a data-driven approach.



Results

This design gives us excellent prediction accuracy, stable solutions based on multiple forms of sensitivity analysis, and easily interpretable results. Our network model allowed us to use the famous PageRank measure to determine the most influential nations. Additionally, running simulations on individual countries implementing optimal policy and measuring each's global effects on the total economic loss due to the environment shows exactly which countries are most influential to the fate of Earth's health and the necessary conditions on which to stabilize the world's rising environmental damage toll.

Saving the Green with the Greens

Abstract

Without the focus of human factors in current models, it is difficult to find the ideal human solution, and that was the force driving our model design. To be especially relevant to each country's decision makers, our model directly predicts human results with economic variables that are easily influenced by legal policy. We further improve our initial design by incorporating proximity, diplomacy, and clustering data into a network model. Our network model allowed us to use the PageRank measure to determine the most influential nations and running simulations on individual countries implementing optimal policy and measuring each's global effects on the total economic loss due to the environment shows exactly which countries are most influential to the fate of Earth's health and the necessary conditions on which to stabilize the world's rising environmental damage toll.

Table of Contents

Introduction.....1
- Clarification of the Problem
- Model Design
- Earth Damage Score (EDS)
- Human Actions
- Data Preparation
- Assumptions

Our Model.....3
- Tikhonov Regularization
- Geographic Network Model
- Hybrid Network Model

Sensitivity Analysis.....7
- Cross Validation
- Gaussian Noise

Conclusion.....7

References.....9

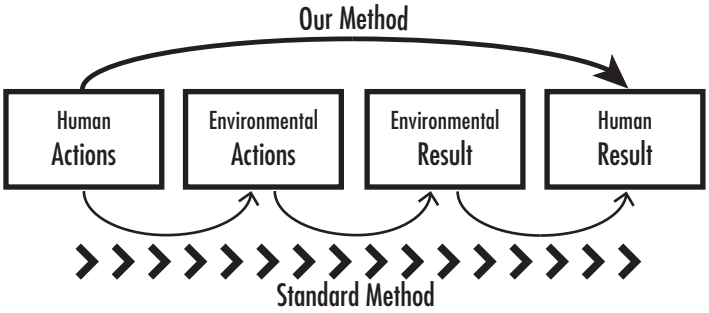
Introduction

Clarification of the Problem

Despite the best efforts of the scientific community, many aspects of environmental science are difficult to model, and it is especially difficult to see the cause and effect dynamics of human action due to the inherent randomness of the relationship between anthropogenic factors and environmental response (Pindyck, 2007). Given the difficulties of this position, we seek to create a new model that looks to the human aspect of ecological damage for better understanding of how human reactions can alter the fate of the planet.

Model Design

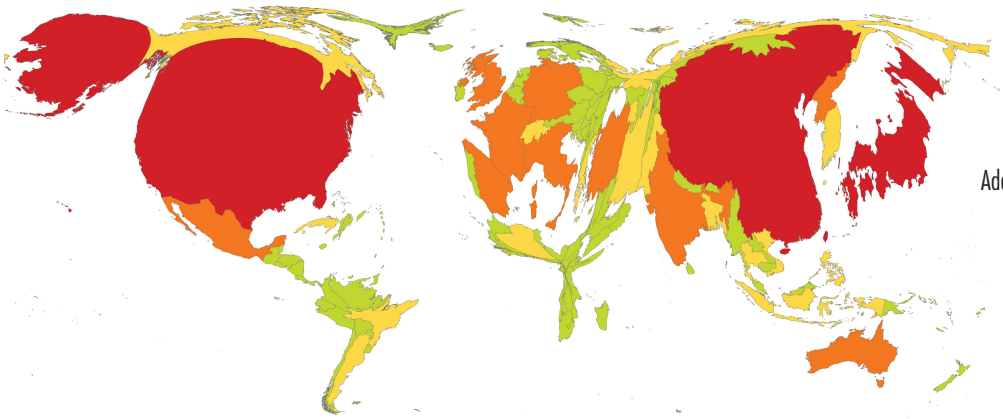
Standard approaches to environmental modeling involve measuring intermediate environmental variables to form the link between human actions towards Earth's environment and Mother Nature's response.



While there are certainly merits to that approach, there are also a number of drawbacks, such as ignoring the potential for human reaction to the environment (Chakravorty and Roumasset, 1997). Because most models do not focus on this relationship, it has not give policy makers a very clear picture of what needs to happen on their part.

We propose a model that takes a data-driven approach to directly predict human result from human actions using machine learning techniques. Using this methodology allows us to avoid the pitfalls of the standard method by using human action as our independent variable. The advantage of this is that we fit less of the noise of the environmental variables which behave indistinguishably from randomness, and we also have the possibility of capturing latent variables that may not normally be measured by environmental studies.

Visualization of Mean Annual Economic Losses Due to Natural Disasters



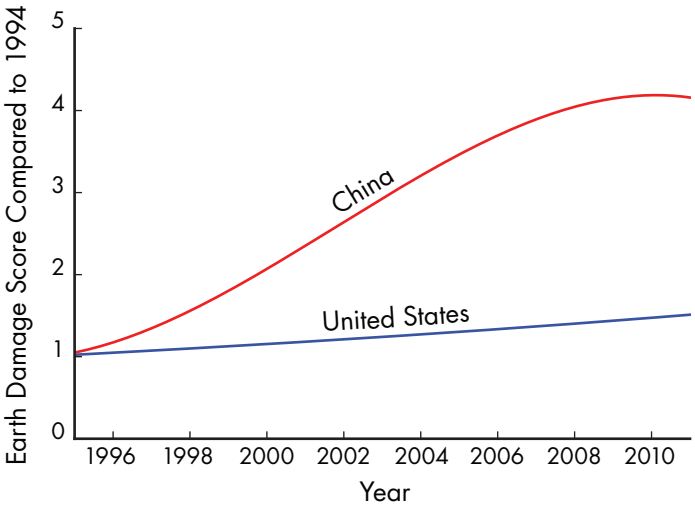
A representation of mean annual economic loss by country.

Adapted from "USA and China top global risk ranking for economic loss due to natural disasters linked to climate change."

Earth Damage Score (EDS)

In order to develop a meaningful measure for the environmental harm experienced by any particular country, we defined the yearly Earth Damage Score according to the economic loss by that country, measured in 2013 US dollars, induced by natural disasters and carbon dioxide damage using data obtained from The World Bank's Databank and Maplecroft. Such a measure was chosen because it represents environmental effects in a format that is easily digestible and directly relevant to policy makers.

We designed our models to predict the percent per year increase in each country's Earth Damage Score, allowing the models to take compounded effects into account as is necessary for environmental predictions because the effects of past actions persevere even as new measures are taken.



Comparing EDS for China and the United States during years for which data was obtained.

Throughout this paper, we use root mean square error as a metric for the inaccuracies of our predicted data. This commonly used measure for error is obtained by squaring the data, then calculating the arithmetic mean of the difference between the actual and predicted data and, finally, taking the square root. This is ideal because it ensures that small errors in differing directions do not cancel each other and the unit of measurement is kept meaningful.

Human Actions

We chose to use certain economic variables as inputs into our model as they are capable of being influenced directly by legal policy and are indicated by evidence to influence environmental factors.

Population and Change In Population

Not only does a rise in population inherently increase agricultural and industrial growth, it adds more bodies to shelter, transport, and contributes more waste (Goodland, 1992).

Agricultural Growth

Agricultural practices today result in deforestation, the release of pesticides, soil degradation, and water pollution from runoff. They are also a tremendous consumer of energy and a large contributing factor to the emission of carbon dioxide and other toxic chemicals into the atmosphere (Trautmann, 2012).

Industrial Growth

While industrial development is necessary to move away from our consumption of fossil fuels and toward a more sustainable source of energy, it nonetheless continues to contribute to the addition of toxic chemicals to the environment and carbon dioxide emissions (Grubb,

Muller, and Butter, 2004).

GDP Growth

Correlations have been found between economic growth and the production rate of pollutants by a country (Grubb, Muller, and Butter, 2004) (Friedlingstein, 2010).

Literacy Rate

Literacy rate is a good indicator of poverty within a country and it is correlated to other factors such as agriculture and industrial growth (Ahluwalia, 1976). It combines a number of factors that are subject to human influence into one easily obtained number.

Data Preparation

All of our economic variables were also obtained from the list of world development indicators in The World Bank's Databank. We collected this information by nation for the main reason that the data was already recorded that way. We use the first time derivative of EDS to run our simulation and obtain future predictions, while controlling the second derivative directly to represent policy changes in each country.

Assumptions

Environmental damage is influenced by human actions.

Our model measures the effect of human action on economic loss due to environmental factors. This is widely perceived to be true, but the assumption is necessary for the creation of our model.

Human action can be controlled, at least in part by legal policy.

This is also perceived to be true, and necessary for predicting the possible outcomes of policy change.

Countries that are near each other share similar environmental effects.

Assuming that environmental effects spread beyond borders allows us to model the interdependencies of the environment.

Policy changes are likely to propagate across diplomatic links.

This assumption allows our model to capture the ripple effect on the spreading of ideology between nations.

Each country's environmental invariants behave approximately as constant throughout the timeframe of the analysis.

Our model doesn't take environmental variables into account, and thus the model captures national invariants such as size, latitude and longitude by solving for a constant.

Our calculated Earth Damage Score is a good proxy for the true economic loss due to environmental damage.

Solving for any country's true economic loss due to the environment is an incredibly complex problem, and is difficult to measure (it is hard to truly know how something such as biodiversity loss will cost); therefore we assume that our measure approximates the true loss well.

Our Model

Tikhonov Regularization

For a simplistic model, we used Tikhonov Regularization, a machine learning algorithm and a form of linear regression which includes a regularization matrix to prevent overfitting of the data (Hoerl, 1970). This algorithm is an appropriate choice for the model because it's especially well suited to problems with limited data and a relatively small Vapnik–Chervonenkis dimension (Vapnik, 2000). Using other models in this situation would lead to much worse prediction rates for future applications.

$$\min_{\mathbf{w}} \|A\mathbf{w} - \mathbf{b}\|^2 + \|\Gamma\mathbf{w}\|^2$$

$$\hat{w} = (A^T A + \Gamma^T \Gamma)^{-1} A^T \mathbf{b}$$

The economic variables were the features of the algorithm represented by the A matrix, the changes in EDS were represented by the $\mathbf{b}$ vector, and the identity matrix was used as the regularization matrix, Γ). In addition, one binary variable for each country was added as an additional feature to allow the algorithm to find optimal constant values for each country, allowing the algorithm to take environmental and geographical invariants into account. $\mathbf{w}$ is the vector of weights for each feature, with the weights of each binary country variable equal to that country's constant value. $\hat{w}$ is the set of weights that minimizes the objective function, and is what is used for future prediction tasks.

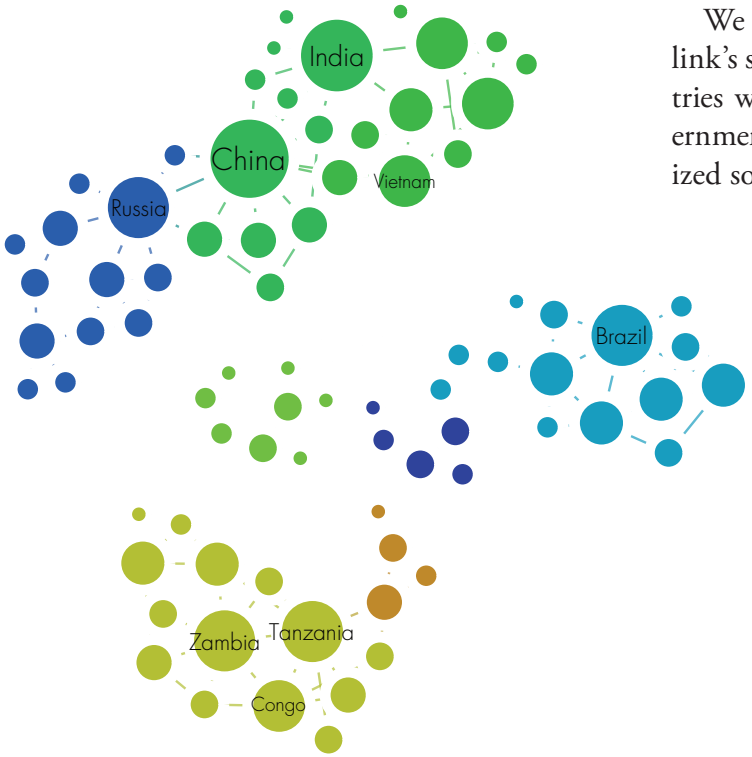
As a comparison, we implemented a similar model with environmental instead of economic variables (carbon dioxide emissions,electric power consumption, water pollution, livestock production and forest area) in order to have a baseline for model accuracy. We had two

more baselines: a simulated model where the EDS was predicted with a normally distributed random variable and another naive model which predicted the arithmetic mean regardless of input variables. As we can see in the table, the economic model does significantly better than both the random and constant model, but it is 35% less accurate than the environmental model.

Model	Root Mean Square Error (Percent Per Year)
Random	2.86
Constant (Mean)	2.39
Economic	1.55
Environmental	1.50

Geographic Network Model

Our simplistic model treated each country as an independent entity and wasn't designed to incorporate more sophisticated parameters such as geographical proximity. This led us to improve our model by creating a global network, and using a weighted modification of a k-nearest neighbor algorithm (Coomans, 1982). Each node in our network represents a country and its size is proportional to the number of countries nearby (this measure is called degree centrality, which corresponds to the influence of a country over its neighbors.) Its color and relative position within the graph correspond to the modularity class to which it belongs, a commonly used measure which intuitively gives an indication of the influential community to which a country belongs.



Geographic Network

Countries are considered adjacent if they geometrically border one another.

Node size corresponds to the number of bordering countries.

Color represents modularity class or, intuitively, the community to which a country belongs.

The graph's adjacency matrix was generated by assigning links if countries bordered each other geographically or were otherwise determined to be in extremely close geographical proximity.

Our methodology was to first perform the estimates of the simple model, and then have the final predicted value to be a weighted average of all nodes of distance two or less apart, using a standard weighting scheme of one divided by the distance of the node plus one.

$$\mathrm{EDS}_{i,\mathrm{final}} = \frac{\sum_{\|i,j\|\leq 2} \frac{\mathrm{EDS}_{j,\mathrm{initial}}}{\|i,j\|+1}}{\sum_{\|i,j\|\leq 2} \frac{1}{\|i,j\|+1}}$$

Applying this modification to our simplistic model resulted in a 25% decrease in root mean square error for predicting EDS.

Hybrid Network Model

Policies of one country can easily influence those of their allies (Hartigan, 1979). This led us to model another dynamic which we have yet to take into account: the diplomatic relationships between countries. By designing an adjacency matrix with weights corresponding to the perceived strength of the diplomatic relations and applying the same algorithm we used for the geographic network model, we obtain a model which is capable of simulating the "ripple effect", demonstrating social influence (Cialdini, 2001).

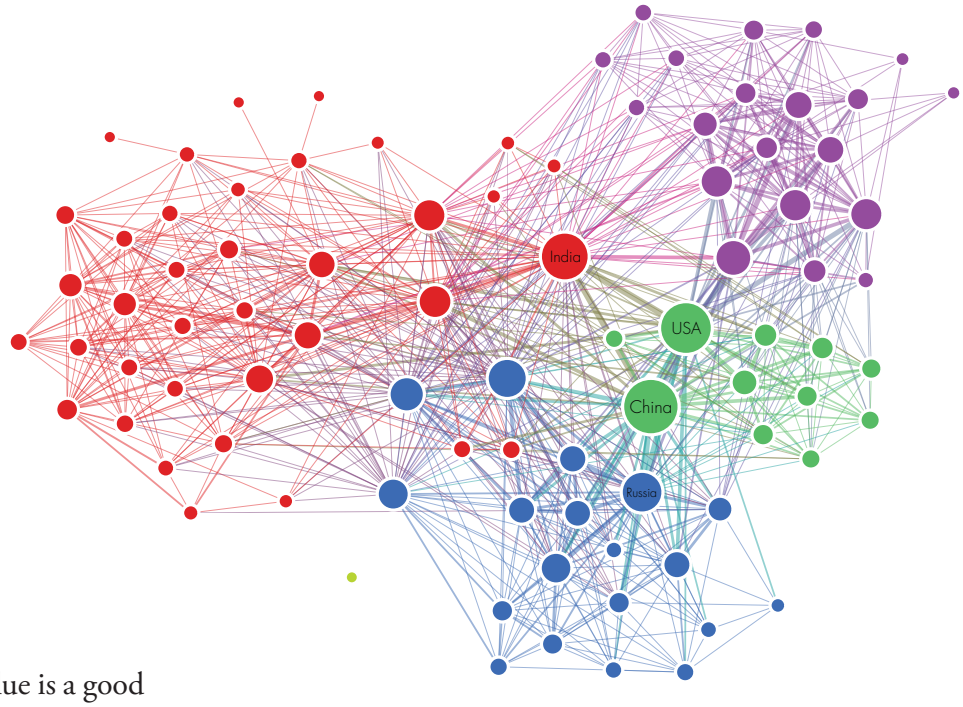
We created the adjacency matrix by weighting each link's strength by the number of times each pair of countries were in common political and economic intergovernmental organizations. This matrix was then normalized so that the maximum value of a link's weight could

Hybrid Network

Countries are considered adjacent if they are geometrically close or share diplomatic links.

Node size corresponds to the PageRank of each country, our measure of diplomatic influence.

Color represents modularity class or, intuitively, the community to which a country belongs.

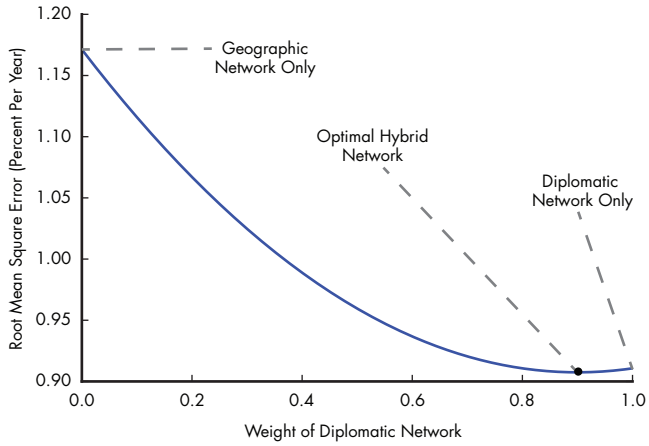


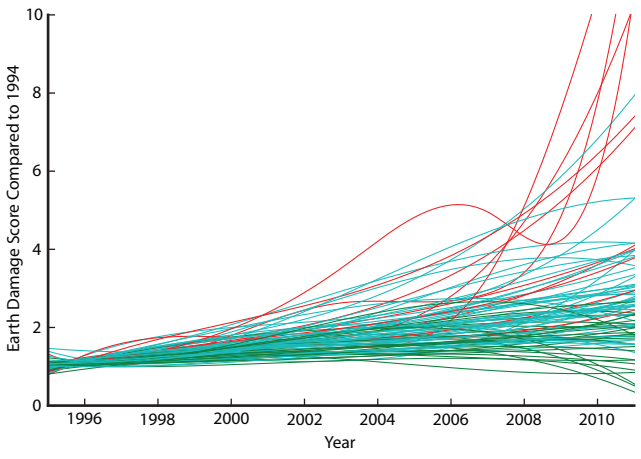
not exceed one. We are assuming that this value is a good measure of the diplomatic relationship between nations. Furthermore, we applied an unsupervised learning algorithm, k-means clustering with three means, to find relationships between any pair of countries with similar behavior (Hartigan, 1979). These similarities were also represented as edges in the modified adjacency matrix.

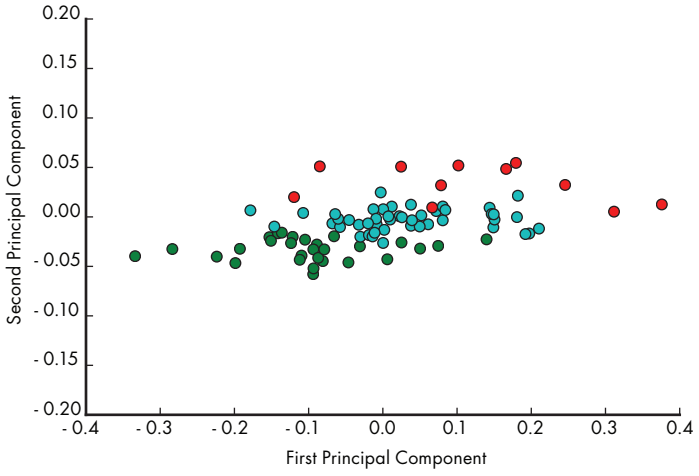
Below, we graph the 88 EDS curves, one for each country in our data, with the color of the curve corresponding to the cluster that the country belongs to. We further demonstrate the clustering by plotting the data points of each country in a two dimensional representation using a technique called randomized principal component analysis (Rokhlin, 2009).

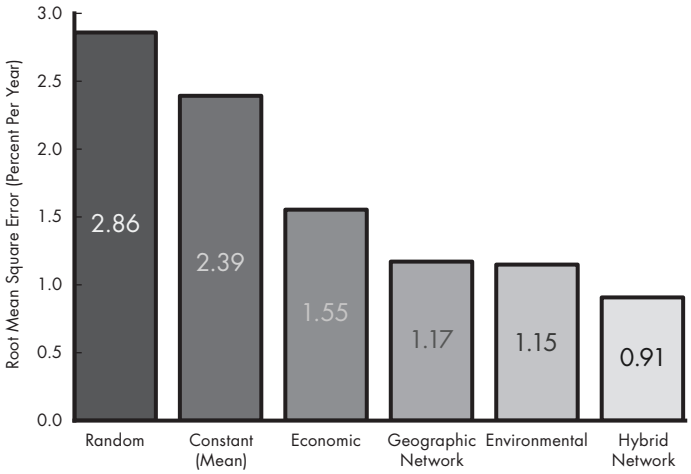
In order to recapture the influence of geographic proximity in our model, we decided to superimpose our two previous networks by assigning them parameters in order to create a realistically scaled linear combination of link weights. We found that the minimum error was obtained when the scaling parameter for the diplomatic network was roughly nine times that of the geographic network. Using these parameters, we were able to obtain the optimal combination of geographic and diplomatic relationships between countries - our hybrid network.

This allows us to obtain our optimal RMS error, 23% lower than with the geographical network model alone.









Measuring Centrality with PageRank

As our choice for a measure of a country's importance in this hybrid model, we can no longer simply use degree centrality because nodes with links of lower weight are less influential than those with higher weight. It thus becomes necessary for us to adopt a new metric for the influence of a node. A common choice is betweenness centrality, which measures how many times a node acts as a bridge. This is not a good criteria because there are a number of countries, such as Kyrgyzstan, which interact with several distinct groups yet have very little influential power. Another choice is closeness centrality, which is essentially a measure of average path length between a given node and any other node in the network. Again, this is a bad choice because although it gives some sense of how quickly policy changes could propagate through the network, it still neglects the weights of each connection and thus does not provide a very good idea of a country's influence.

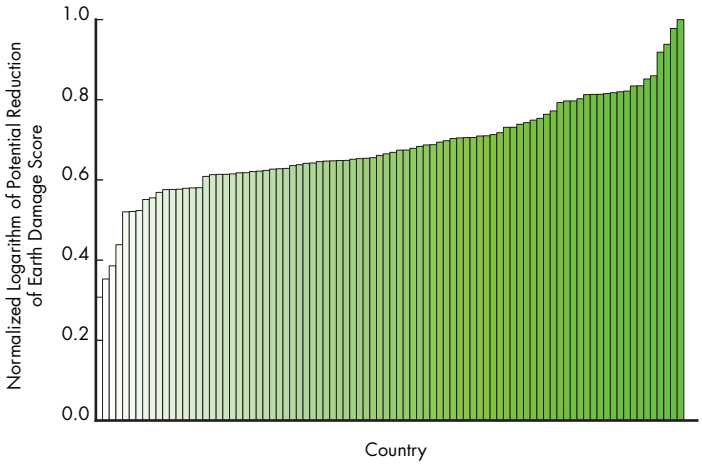
Instead, we use the PageRank link analysis algorithm to determine our metric of influential power (Page, Brin, Motwani, and Winograd, 1999). This turns out to be an excellent measure of relative importance and it is represented in our graph by the size of each node. Color and location again represent modularity class.

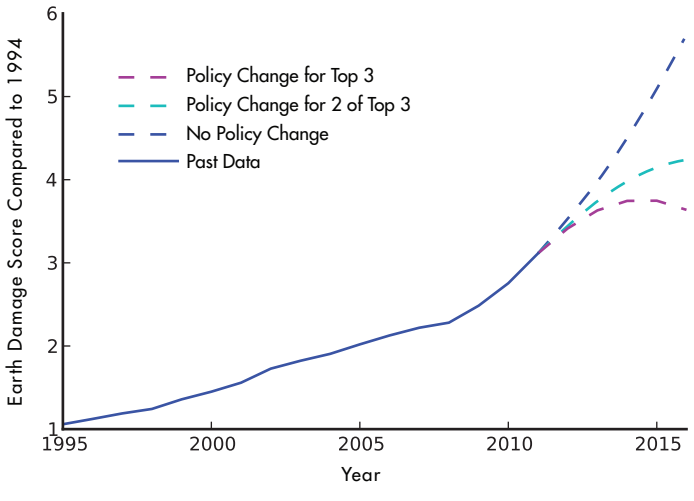
The Biggest Influences

As another method of identifying key nodes in our graph, we solved for which countries could make the largest overall effect on average EDS. In order to calculate this for each country, we solved for the direction of the second derivative of economic variables, which we assumed policy could influence, that would minimize average EDS, and found the difference between implementing a predicted optimal policy versus not implementing any policy. Not implementing any policy is equivalent to setting the second derivative of the economic variables to be zero. The results of this are shown below.

Our model predicts that the countries that could make the largest difference in the world's EDS are China, the United States, and to a lesser extent India. The difference that they make is several orders of magnitude larger than that possible by the rest of the countries. Our model predicts that implementing optimal policy in any two of the top three countries would be enough stabilize the EDS score for many generations, and implementing optimal policy in all three of China, the USA, and India, could bring the entire world's EDS down in slightly less than five years.

Furthermore, this analysis gives us a human interpretable direction of the optimal second derivative of our economic variables. Based on the data, optimal policy greatly emphasizes a decrease in the growth rate of the country's population. This knowledge can be used to further inform policy makers of the most effective policies to reduce EDS.



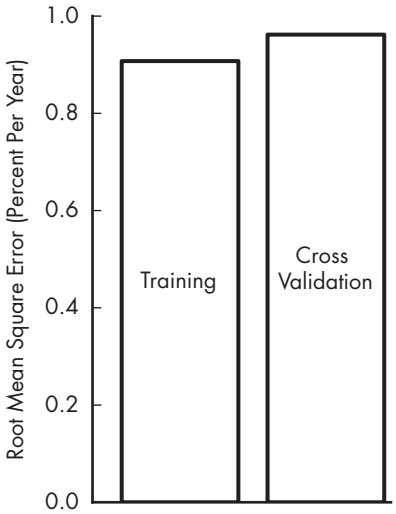


Sensitivity Analysis

A certain amount of confidence in environmental models is necessary for policy making decisions because, without it, it is difficult to determine the best policy to use, and the benefits of implementing it, if any, and currently policy makers do not have this level of confidence to make the best decisions (Pindyck, 2007). Because of this necessity, it is especially important to achieve as much confidence in our predictions as possible, thus we performed multiple forms of sensitivity analysis.

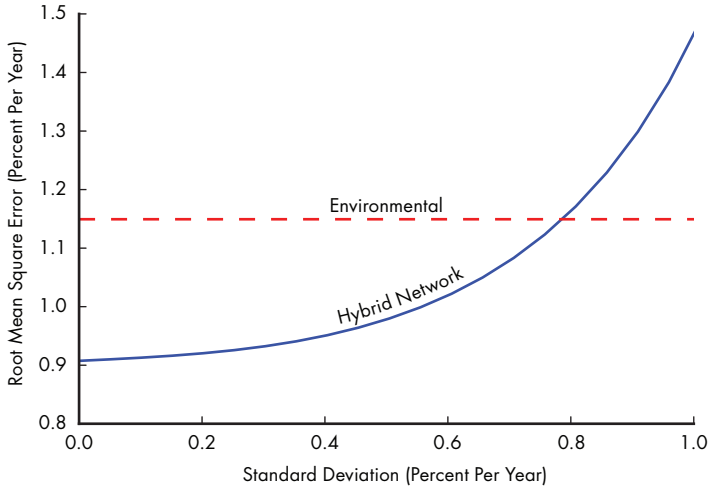
Cross Validation

We initially used 10-fold cross validation, a common standard for measuring the robustness of algorithms (Kohavi, 1995). The method involves randomly partitioning the data into tenths, and for each tenth, the remaining nine are used to predict its EDS. This method shows the cross-validation error less than 6% higher than training set error, demonstrating robustness of the model (Krogh, 1995).



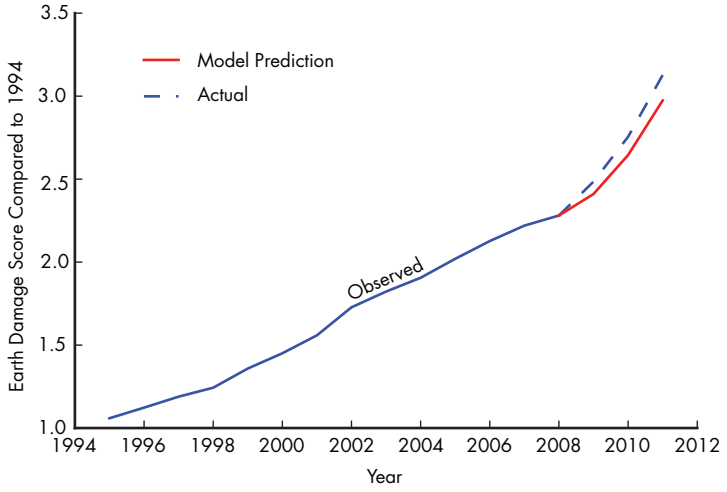
Gaussian Noise

As another form of testing the stability of our predictions, we added gaussian noise with increasing standard deviation to our input variables in order to verify that our predictions are stable with perturbations in our data. We see that for low amounts of noise, we achieve remarkably comparable accuracy to that of the model with correct data, and for standard deviations of the noise less than 0.8, we obtain comparable accuracy to the model taking environmental variables into account.



Test Set Future Prediction

As a third form of verifying model accuracy, we partitioned the last three years of our data into a test set, and only trained the model, including clustering and hyperparameter optimization, on the remaining data. We then attempted to predict the EDS of our test set data given its economic variables. This prediction was within 5% of the true EDS, showing our model to provide excellent accuracy for medium length intervals of time (less than 15 years) (Schapire, 1999).



Conclusion

Strengths

Human Relevant

Our independent and dependent variables are directly relevant to people without needing to pass through complex calculations to see potential impact.

Flexible Model

Our model was designed to be extremely flexible with the types of data it can use for prediction. This design choice was made in hope that if it's used in the future, we can provide even better predictions.

Stable Predictions

We verified with a variety of methods that our model has very stable solutions, allowing us to make predictions with confidence, which is especially important in this field.

Simplicity and Human Interpretability

By using simple but powerful models, we retain majority of the accuracy of a complex model, with the added benefit that the every parameter of the algorithm is interpretable by humans, thus empowering our decision making ability.

Weaknesses

Data Limitations

We chose to use an entirely data-driven model to attain maximum accuracy. Unfortunately, we weren't able to take countries into account whose data was too incomplete, even for state of the art data imputation techniques. This caused a selection bias within our model, and most notably, a majority of the European Union was missing. This easily could have made a difference in our predictions, but without the data, we are in no position to speculate. Thankfully, we achieved excellent prediction accuracy despite the missing data.

Domain Expertise

Like the majority of machine learning algorithms, our model can't easily take the knowledge of an expert of environmental science to enhance its predictive capabilities. We justify the usage of machine learning because most models today have an entirely different focus than ours, and thus wouldn't significantly aid our model.

Long Term Predictions

Our model suffers from the same flaw as all other environmental models: a lack of understanding of the Earth's reactions to the 'state shifts' that we are so very concerned with, i.e. global climate change. When the Earth crosses these 'critical transitions,' it tends to very abruptly override the very trends all models are based off of (Barnosky, Hadly, Bascompte, Berlow, Brown, Fortelius, Getz & Smith, 2012). We certainly do not know what will happen to the human factors when the Earth crosses this threshold.

Recommendations

Based on the results of our model, we have a few recommendations for the world, especially the policy makers.

Record More Data

Recording data for your respective countries will only help inform you and all of your fellow citizens. As computers get more and more capable of handling all the data that is collected, you put yourselves in a position to be helped as much as possible.

Tailor Your Models

By including economic and other social measures in creating ecological predictions, you not only add an important dimension to the problem itself, but it is easier to understand how we as humans can influence our environmental future.

Make a Change

Our research indicates that the best use of policy is population management. Our model has predicted this to be good for both the environment and the economy, and others have made similar claims (Bongaarts, 1992). The highest compliment that we as a team could receive is for our model to be used with even more data to help inform and make a positive change in policy.

References

Ahluwalia, M. S. (1976). Inequality, poverty and development. Journal of Development Economics, 3(4), 307-342.

Barnosky, A. D., Hadly, E. A., Bascompte, J., Berlow, E. L., Brown, J. H., Fortelius, M., Getz, W. M., & Smith, A. B. (2012). Approaching a state shift in earth's biosphere. Nature, 486(7401), 52-58. doi: 10.1038/nature11018

Bongaarts, J. (1992). Population growth and global warming. population and Development Review, 299-319.

Chakravorty, U., Roumasset, J., & Tse, K. (1997). Endogenous substitution among energy resources and global warming. Journal of Political Economy,105(6), 1201-1234.

Cialdini, R. B. (2001). Influence: Science and practice (Vol. 4). Boston, MA: Allyn and Bacon.Cialdini, R. B. (2001). Influence: Science and practice (Vol. 4). Boston, MA: Allyn and Bacon.

Coomans, D., & Massart, D. L. (1982). Alternative< i> k</i>-nearest neighbour rules in supervised pattern recognition: Part 1. k-Nearest neighbour classification by using alternative voting rules. Analytica Chimica Acta, 136, 15-27.

Friedlingstein, P., et al. "Update on CO2 emissions." Nature Geoscience 3.12 (2010): 811-812.

Goodland, R. (1992). The case that the world has reached limits: more precisely that current throughput growth in the global economy cannot be sustained. Population & Environment, 13(3), 167-182.

Grubb, Michael, B. Muller, and L. Butter. "The relationship between carbon dioxide emissions and economic growth." Oxbridge study on CO2-GDP relationships (2004).

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. Applied statistics, 100-108.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55-67.

Holdren, J. P., & Ehrlich, P. R. (1974). Human population and the global environment: Population growth, rising per capita material consumption, and disruptive technologies have made civilization a global ecological force.American Scientist, 282-292.

Kohavi, Ron (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection". Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence 2 (12): 1137–1143.(Morgan Kaufmann, San Mateo, CA)

Krogh, A., & Vedelsby, J. (1995). Neural network ensembles, cross validation, and active learning. Advances in neural information processing systems, 231-238.

List of international organizations. (n.d.). Retrieved from http://www.umdnj.edu/uroweb/international_office/general info/List of International Organizations.pdf

Maplecroft. (2009). Usa and china top global risk ranking for economic loss due to natural disasters linked to climate change. Retrieved from http://businessassurance.com/downloads/2009/03/natural-disaster-and-economic-loss1.pdf

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: bringing order to the web.

Pindyck, R. S. (2007). Uncertainty in environmental economics. Review of Environmental Economics and Policy, 1(1), 45-65.

Robert Watson and A.Hamid Zakri. UN Millennium Ecosystem Assessment Synthesis Report, United Nations Report, 2005.

Rokhlin, V., Szlam, A., & Tygert, M. (2009). A randomized algorithm for principal component analysis. SIAM Journal on Matrix Analysis and Applications, 31(3), 1100-1124.

Schapire, R. E., & Singer, Y. (1999). Improved boosting algorithms using confidence-rated predictions. Machine learning, 37(3), 297-336.

Shannon, M., Morey, D., & Boehmke, F. J. (2010). The Influence of International Organizations on Militarized Dispute

Initiation and Duration1.International Studies Quarterly, 54(4), 1123-1141.

The increasing costs of U.S. natural disasters. (2005, November). Geotimes, Retrieved from http://www.geotimes.org

Trautmann, Nancy M. (2012). Modern Agriculture: Its Effects on the Environment. Cornell Cooperative Extension. Retrived from http://psep.cce.cornell.edu/facts-slides-self/facts/mod-ag-grw85.aspx

Vapnik, V. (2000). The nature of statistical learning theory. Springer-Verlag.

World Development Indicators. (2013, 1 8). World databank. Retrieved from http://databank.worldbank.org