Modeling Language Endangerment with Current Geopolitical Data

Introduction

According to the *Cambridge Handbook of Endangered Languages*, there are over 7,000 languages spoken in the world today. However, roughly one third of the world's languages are in danger of extinction. Primary causes of language endangerment include natural disasters and disease, war and genocide, overt repression, and cultural, political, or economic dominance [1]. While well known, these causes can be difficult to quantify, though efforts to do so are not unprecedented. The 2014 paper "Global distribution and drivers of language extinction risk", for instance, included an in-depth analysis of factors such as GDP and language range. [2]

To better understand the causes behind language endangerment, a classification model was constructed using existing high-level quantitative measures of the causes laid out in the *Cambridge Handbook,* along with other major attributes such as linguistic family and location.

Data

An initial dataset provided by *the Guardian* contained 2,722 examples of endangered languages, including the language's central coordinates, geographic spread, number of speakers, and degree of endangerment as determined by the UNESCO language endangerment scale [3]. The UNESCO scale classifies endangered languages in one of five categories: Vulnerable (or "unsafe"), Definitely Endangered, Severely Endangered, Critically Endangered, and Extinct. The classification itself is based on nine different factors, including intergenerational language transmission, literacy, and total number of speakers.

Total number of speakers is a major factor of a language's endangerment status, but as indicated by UNESCO's nine contributing factors and the decision tree in Figure 1, it is not the only factor and does not provide a clear split between the categories. Classification of languages is nuanced, as are the causes of endangerment. The number of speakers was not selected as a final feature in the model as it may be considered a result of endangerment rather than a cause.

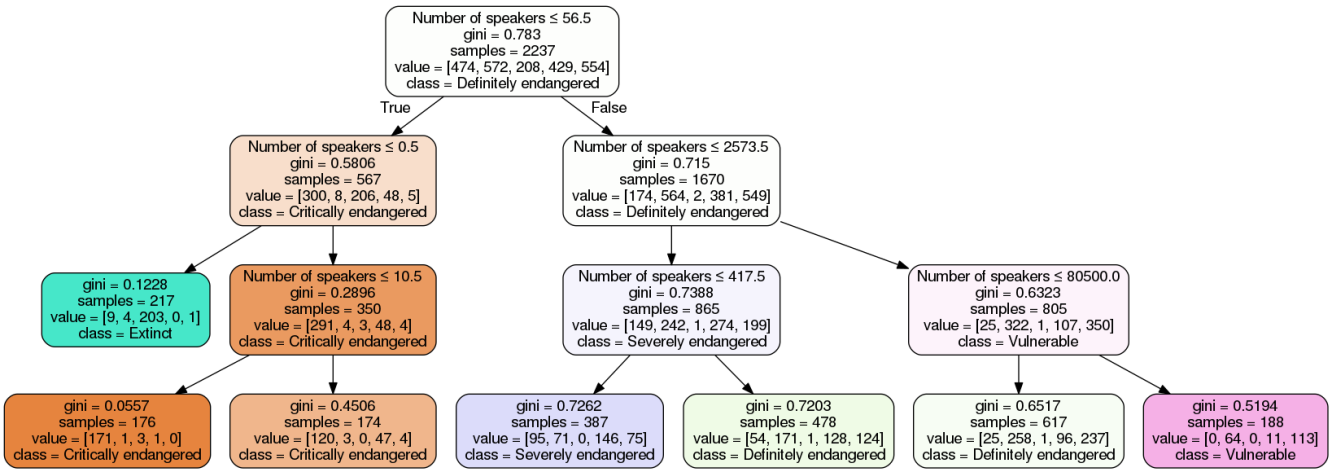

Figure 1: Decision Tree Classification of Endangered Languages based on number of speakers

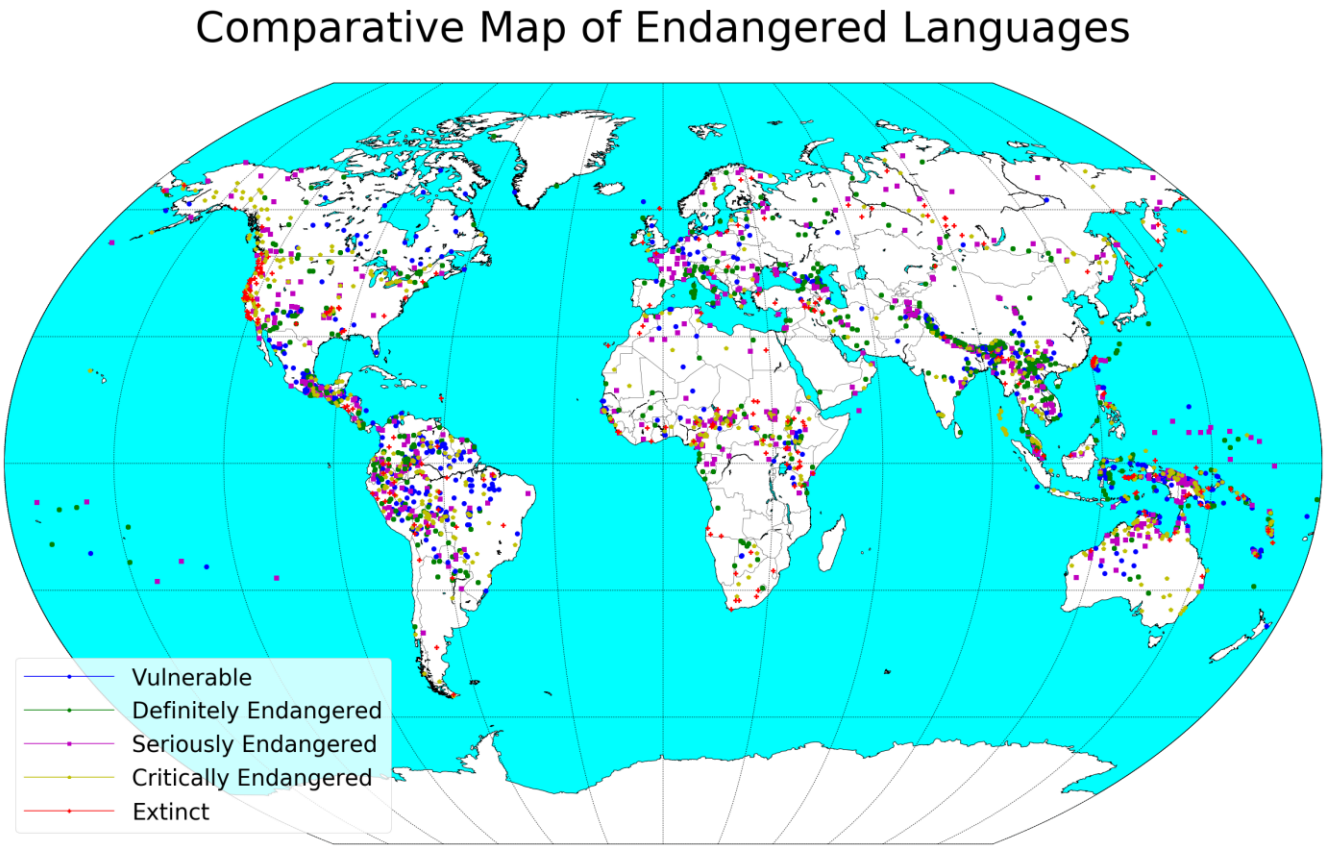

Figure 2: Map of endangered languages by endangerment status

One additional feature provided with the initial data set was latitude/longitude pairs indicating the central location of a language's speakers. As seen in Figure 2, there exist clusters of endangered languages across the globe, such as the cluster of extinct languages along the western coast of the United States and a dense group of endangered languages along the Himalayas. Geographic coordinates were also considered holistically as level 25 S2 cells via the S2 Geometry library, which allowed mapping the latitude and longitude to single values along a Hilbert curve [4].

For the purposes of this model, language range was determined by the number of countries in which a particular language is spoken. Here, language range does not refer to physical distance, but rather to the spread of a language across political boundaries.

Additional data was collected from publicly available sources on the Internet. GDP, often associated with cultural hegemony and globalization, has been considered a prime factor behind language endangerment, as confirmed in the 2014 paper “Global distribution and drivers of language extinction risk” [2]. GDP data from 2015 was collected from the World Bank's data catalog [5]. An average GDP value was calculated for languages which span multiple countries, and data for GDP was only available at the national level.

Overt repression was estimated using the Human Freedom Index, an index introduced by the Cato Institute in 2015 that considers “79 distinct indicators of personal and economic freedom” [6]. As with GDP, an average value was taken for languages which spread across international borders. Similarly, a country's propensity toward war was collected from the Global Peace Index, introduced in 2007 by the Institute for Economics and Peace [7]. Natural disasters were accounted for using the World Risk Index, which has been produced by the University of Stuttgart since 2011 and which scores

countries based on their exposure, susceptibility, coping capacity, and adaptive capacity in regard to natural disasters [8].

Two measures of linguistic diversity were also considered, both collected from the Ethnologue [9]. The family size of endangered languages was collected to account for language isolation, and linguistic heterogeneity at national levels was considered through the Greenberg Diversity Index. Greenberg's language diversity index measures the probability of two randomly selected people in the same country speaking the same mother tongue.

Results

After dropping samples with incomplete data, the data set was left with 2,227 samples. Features were scaled and scored using the scikit-learn library (see Figure 3). In agreement with previous research [2], GDP was overwhelmingly the best predictor of endangerment status with a score of 16.20. Latitude and longitude values also scored sufficiently well, with latitude scoring higher, while more specific locations as recorded in S2 cells scored lower than either.

Purely linguistic features such as linguistic diversity and language isolation scored poorly. Despite being a notable factor in the endangerment of several languages in specific instances, overall disaster risk proved to be a poor indicator as well. The final features selected for use in modeling were the GDP, Human Freedom Index, latitude, longitude, geopolitical range, and Global Peace Index values.

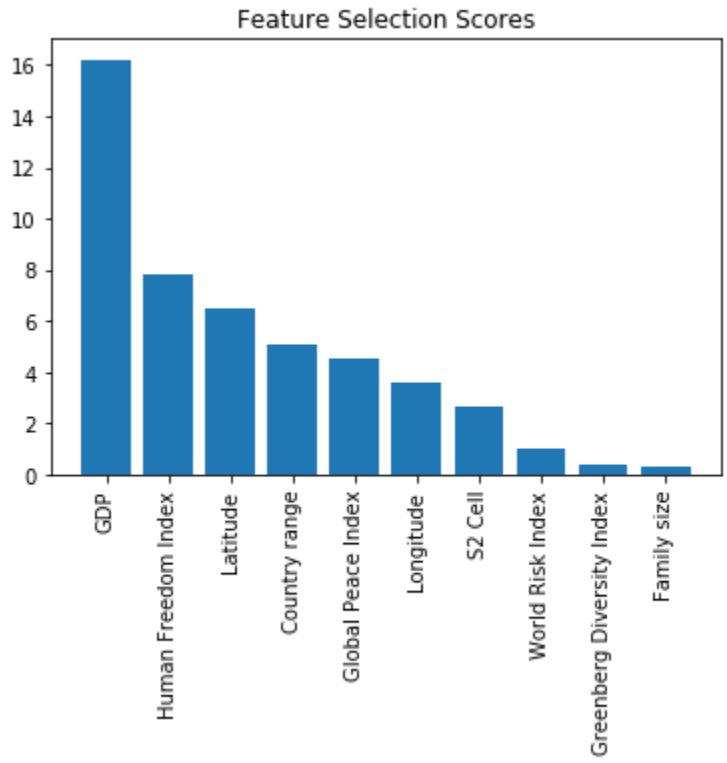

Figure 3: Features ranked by scikit-learn

However, several of the features presented are moderately correlated with other features. As shown in the correlation matrix in figure 4, features dependent on physical location or linguistic features have significant correlation between each other. For instance, values from the Human Freedom Index and the Global Peace Index have a correlation coefficient of approximately 0.33.

Selected features were tested with several estimators as recommended by the scikit-learn documentation, namely the SVC, Linear SVC, K-Neighbors, Decision Tree, and Random Forest classifiers [10]. The data was randomly split into a training set consisting of 80% of the data and a test set consisting of the remaining 20%. As the dataset was relatively small with just over 2,000 samples, the models went through 100 rounds of testing to find average accuracy. While none of the models produced were exceptionally accurate, the Random Forest had the best results with about 42% accuracy. The Random Forest was configured with a minimum of five samples per leaf and 25 trees. See Figure 5 for complete results.

	GDP	Number of speakers	Human Freedom Index	Latitude	Country range	Global Peace Index	Longitude	S2 Cell	World Risk Index	Greenberg Diversity Index	Family size
GDP	1.000000	-0.053634	0.332375	0.398862	-0.049731	0.164333	-0.330812	0.141001	-0.299564	-0.177145	-0.231516
Number of speakers	-0.053634	1.000000	0.059056	0.124218	0.347468	-0.067412	0.016656	-0.026118	-0.056664	0.005718	-0.009552
Human Freedom Index	0.332375	0.059056	1.000000	0.085917	-0.000811	0.330574	-0.310699	-0.428203	-0.268764	-0.270176	-0.241279
Latitude	0.398862	0.124218	0.085917	1.000000	0.107919	-0.183290	-0.161581	0.291437	-0.313830	-0.052393	-0.185389
Country range	-0.049731	0.347468	-0.000811	0.107919	1.000000	-0.038357	0.017412	-0.014891	-0.047767	-0.031471	-0.006614
Global Peace Index	0.164333	-0.067412	0.330574	-0.183290	-0.038357	1.000000	-0.561722	0.128950	-0.089665	-0.622188	-0.478527
Longitude	-0.330812	0.016656	-0.310699	-0.161581	0.017412	-0.561722	1.000000	-0.136543	0.349793	0.571810	0.398525
S2 Cell	0.141001	-0.026118	-0.428203	0.291437	-0.014891	0.128950	-0.136543	1.000000	-0.092855	-0.085937	-0.219722
World Risk Index	-0.299564	-0.056664	-0.268764	-0.313830	-0.047767	-0.089665	0.349793	-0.092855	1.000000	0.454572	0.406259
Greenberg Diversity Index	-0.177145	0.005718	-0.270176	-0.052393	-0.031471	-0.622188	0.571810	-0.085937	0.454572	1.000000	0.491744
Family size	-0.231516	-0.009552	-0.241279	-0.185389	-0.006614	-0.478527	0.398525	-0.219722	0.406259	0.491744	1.000000

Figure 4. Feature Correlation Matrix

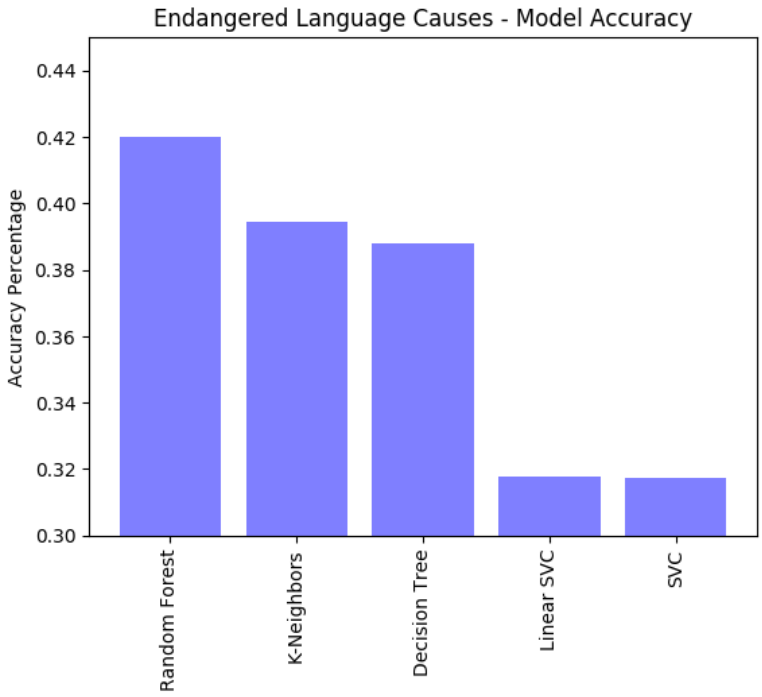

Figure 5. Modeling Accuracy

Notes and Discussion

As noted, model accuracy was relatively low: 42.00% for for classification of a language in one of five endangerment categories. While better than random chance, the high-level approach of this model is clearly insufficient for predictive classification.

The model's construction was additionally limited by the data set's composition. Only data from endangered languages was considered, so the model is not valid for examining languages which have

not been previously classified as vulnerable or worse. Further, it is difficult to maintain up-to-date data on the various languages. Sources of aggregated language data such as the Ethnologue necessarily rely on individual language studies, which studies can take years to complete and may rely on estimates depending on the spread of the language. The status of a language can also change rapidly due to the sudden occurrence of natural disasters or disease. Though the World Risk Index was not useful for this model, it has been established that natural disasters do put endangered languages at great risk. For example, the 1998 earthquake in Papua New Guinea destroyed four villages of unique ethnic populations, leaving the fate of their languages uncertain even now [11].

As such, future studies could benefit from more specific data points, such as considering economic prosperity on a local or regional scale in addition to the GDP of entire countries. Historical data would also be useful in analyzing endangerment trends, as features like GDP, relative freedom, and relative peace can vary dramatically after significant events such as the fall of the Soviet Union.

The difficulty remains that such historical and specific data often does not exist in a usable, quantitative form. The World Risk Index and Global Peace Index used in this model were first launched within the past decade, and further analysis would be required to adapt the index to historical data. Therefore, further attempts to model language risk would benefit from additional studies in related fields, such as historical trends in economics and risk of natural disasters.

Despite the low accuracy rate of this model, then, correlations between endangerment status and proposed drivers of language endangerment suggest that modeling language risk based on quantitative data will be feasible in the future.

References

[1] P. Austin and J. Sallabank, *The Cambridge handbook of endangered languages*, 1st ed. New York: Cambridge University Press, 2011, pp. 1-6.

[2] T. Amano, B. Sandel, H. Eager, E. Bulteau, J. Svenning, B. Dalsgaard, C. Rahbek, R. Davies and W. Sutherland, "Global distribution and drivers of language extinction risk", *Proceedings of the Royal Society B: Biological Sciences*, vol. 281, no. 1793, 2014.

[3] "Extinct Languages", *Kaggle.com*, 2016. [Online]. Available: https://www.kaggle.com/the-guardian/extinct-languages.

[4] O. Procopiuc, "Geometry on the Sphere: Google's S2 Library", 2011.

[5] "GDP ranking", *The World Bank*, 2017. [Online]. Available: http://data.worldbank.org/data-catalog/GDP-ranking-table.

[6] "Human Freedom Index", *The Cato Institute*, 2017. [Online]. Available: https://www.cato.org/human-freedom-index.

[7] "Global Peace Index", *Vision of Humanity*, 2017. [Online]. Available: http://static.visionofhumanity.org/#/page/indexes/global-peace-index.

[8] "World Risk Index", *University of Stuttgart*, 2017. [Online]. Available: http://www.uni-stuttgart.de/ireus/Internationales/WorldRiskIndex/.

[9] "Ethnologue: Languages of the World", *Ethnologue*, 2017. [Online]. Available: https://www.ethnologue.com.

[10] "Choosing the right estimator", *Scikit-learn 0.18.1 documentation*, 2017. [Online]. Available: http://scikit-learn.org/stable/tutorial/machine_learning_map/.

[11] D. Crystal, *Language death*, 1st ed. Cambridge [u.a.]: Cambridge Univ. Press, 2010.