Leveraging ESG Propensity in Fixed Income Investing

Philip Thoenen

Contents

1 Problem Statement

Investors looking to maximize the profit of their portfolio often turn to prediction models to estimate future prices of securities. This analysis will attempt to do so with an eye on a less talked about market: long term sovereign bonds. Specifically, the yield of 10 year treasury bonds will forecasted. Countries' ESG propensity will also be incorporated in order to asses how this newer investor focus may assist predictions. Ultimately, this analysis attempts to leverage ESG metrics to cluster countries and use this information to assist predictions of bond yields.

2 Background

Traditionally, an investor's goal is to make the highest returns they can given the risk profile they feel comfortable with. This requires a delicate balance as the highest returns often come from the riskiest investments. If one can predict a future imbalance between these two factors, aka high returns coming from a safe investment, they should be able to take advantage and position themself to make money. This paper will explore making such predictions in a specific sector of investing called fixed income investing.

Fixed income investing refers to lending a sum of money in return for fixed payments, like dividends, until an agreed upon maturity date. Upon this date, the investor gets back the initial amount they gave the borrower. These types of investments are often bonds, which have a fixed interest rate, given out by companies or countries. Investors may seek these out as a way to balance out a portfolio as they provide a known return with relatively low risk.

One of the largest portions of the fixed income market is sovereign bonds. These are bonds issued by countries' governments to raise money for efforts like financing government programs or paying down old debt. While these bonds can last for different amounts of time, one of the most common bonds is the 10-year bond. These will be the focus of this analysis.

As mentioned above, investors usually decide if sovereign bonds should be added to their portfolio by comparing the bonds' risk and return. A bond's risk can be thought of as the likelihood of a country defaulting on the bond, meaning they cannot pay back the money they borrowed. This is typically measured using a credit rating. Just as individuals have credit scores, countries receive credit ratings which help investors assess the likelihood of default.

Three major institutions that assign these ratings are S&P, Fitch, and Moody. Each institution has slightly different criteria for assigning a rating but in general there is consistency across the rankings. As an example of the scales, S&P's scale goes from AAA to D where AAA corresponds to obligations that are of the highest quality and therefore have the lowest risk. Anything rated BB or below is not considered investment grade and is subject to substantial credit risk. [1]

A bond's return can be calculated in several different ways, but a common one is its yield. The yield is the return an investor receives on the bond. The simplest way to calculate this is by dividing the interest rate the bond is paying by its face value. For example, if an investor receives \$60 each year for a \$1000 bond they purchased, the yield would be 6%.

A bond's yield tends to be inversely related to a country's credit rating. This is because investors demand higher returns from countries which are more likely to default. They are only willing to take on extra risk for the promise of extra return. While both the yield and credit rating for a country's 10-year bond can be easily obtained for the current day, the ability to predict both of these data points in the future would give an investor an advantage in determining when to buy and sell a particular bond.

Additionally, over the past 10 years, Environmental, Social, and Governance (ESG) investing has become more popular across the globe. This philosophy entails thinking about these three features in addition to just considering the financial return an investment might provide. People are starting to prioritize investing in companies and countries which produce social good. Some also claim that ESG investing can produce higher returns than traditional strategies, although there is inconclusive evidence as to whether this is true.

ESG investing is typically mentioned in reference to stocks and large corporations; however, it is starting to take hold in fixed-income markets. One could argue that a country's ESG friendliness has a direct relationship to its bond performance, as these bonds are handed out by the government which is usually weighted the highest of the three ESG factors. Also, many of these bonds are used to fund infrastructure projects, which could have significant impacts on the environment or on the social welfare of the country's citizens. Therefore, incorporating ESG metrics in the analysis of sovereign bonds may prove beneficial.

3 Data Sources

The data for this analysis was collected from multiple sources. The ESG data comes from the World Bank which has a database of sovereign ESG data that dates back to 1960. This data ranges across 67 different factors that fall into the environmental, social, or governance categories. The data can be found on the world bank website. [2]

The historical yield data comes from the Federal Reserve Bank of St. Louis, which maintains a database called FRED that contains economic data. [3] The FRED database has an API that was used to access the yield data for all the countries. This ended up being the limiting factor in the countries used for this analysis. FRED only contained data for the long term bonds of 39 sovereigns; therefore, the analysis focuses on these nations.

4 Methodology

The data sources were quite messy and required a significant amount of pre-processing to extract the relevant information. The ESG data came from a CSV file which made extracting the data simple, but it was quickly evident that lots of data was missing. As mentioned above, the data contained 67 different indicator tags per country, per year. However, when looking at recent years such as 2017 and 2018 only around 40 of those indicators contained data for the majority of countries.

Indicators which contained more than 5 missing values in a given year were dropped from the data set. Imputation was done for the other indicators which contained just a few missing values. The median of other countries was chosen for the imputation value. The median was chosen because many of the factors had skewed distributions so it was felt that the median would be less biased than the mean. Regression was also considered for imputation but it was ruled out due to the added complexity along with the fact that it could lead to over-fitting as the data is being used twice to fit models.

As mentioned above, the bond yield data was the limiting factor in determining the countries for this analysis. The online FRED database only contained long term government bond yields for 40 entities. One of these was the Euro Area which was removed as it is not an individual country like the rest of the data. FRED had recorded data at monthly, quarterly, and yearly intervals. Granularity was desired to hopefully catch seasonal trends, however, monthly data was missing for a few countries so quarterly data was selected for the analysis.

The starting point at which data had been collected varied widely between the countries, with some going back as far as the 1950s. Given that data this far back was unlikely to bear significant impact on a country's bond yield today, all data from before 2001 was removed. This also had the benefit of leaving all but 5 countries with an equal length series of data. For these countries that did not go back to 2001, they all possessed data from 2012 and beyond which was deemed enough to include.

Similar to the ESG data, the yield data also contained some missing values. Luckily, no two consecutive values were missing for any country which meant a relatively simple interpolation method could be used. Linear interpolation between the previous yield and the next yield was chosen. While the yield data tended to fluctuate from quarter to quarter, those changes tended to be small between adjacent quarters. Therefore, it was determined that this would be a close enough approximation as the yield was likely to fall with the range of its adjacent measurements.

Ultimately, based on the cleaned data that was available, it was chosen to use information from 2018 and before as training data and data for 2019 as test data. This had two advantages: one, the 2019 ESG data was missing data for more than half the factors which would have reduced the analysis capability; two, it avoids using 2020 data which is likely greatly affected by the Covid-19 pandemic.

Once the data had been cleaned and divided into training and test sets, it was checked for compliance with the various assumptions that are necessary for the models that were used. First, the covariance matrix of the ESG was calculated. This produced the insight that there was a large amount of co-linearity present in the data. This was unsurprising, as it makes sense that countries that are environmentally friendly rate highly in many of the environmental categories or similar logic for the other categories. To counteract this, principal component analysis (PCA) was performed.

The main benefit of PCA, like mentioned above, is to transform the data into dimensions which are independent of one another. Many models rely on the assumption of independent factors. Specifically, in respect to clustering, removing co-linearity prevents two highly correlated features from dominating the model. Additionally, instead of applying future models to almost 40 factors, the dimensionality can be reduced to just those principal components which explain a significant portion of the variability in the data. This speeds up model training and produces a simpler model because less parameters are fit.

The number of principal components to reduce to was chosen by looking at the variance explained by each component. While each successive component will explain more variance, often there is a point of diminishing returns at which the next component does not explain significantly more variance than what has been explained by the previous components. An elbow graph, as shown below, was created to visually assess this point.

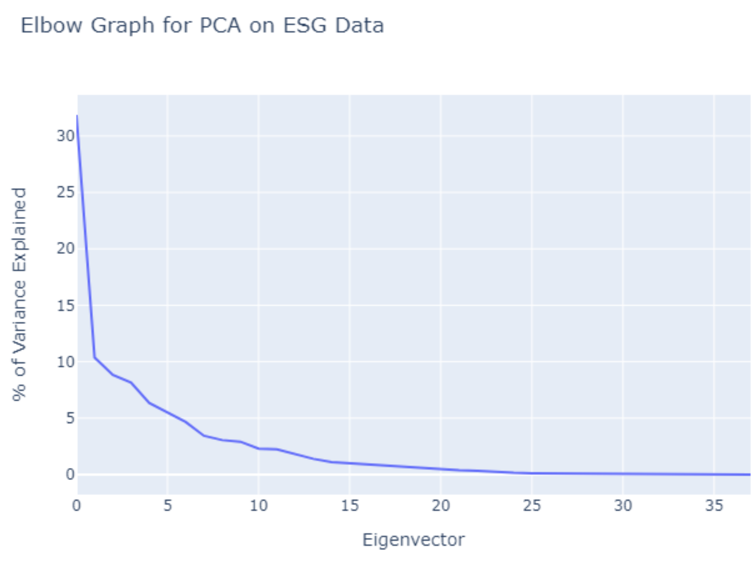

Figure 1: PCA Elbow Graph

Based on the elbow graph, the first 11 principal components were chosen. As the graph shows, these components make up the majority of the variance in the data set. One other benefit of PCA is that it takes highly dimensional data, such as this, and transforms it into something that can be plotted in two dimensions and provide a good amount of signal. While 11 components, will be used for the analysis, the first two represent more than 40% of the variance explained so looking at just those can give some insight. Figure 2 shows the countries plotted on the top 2 principal component axes.

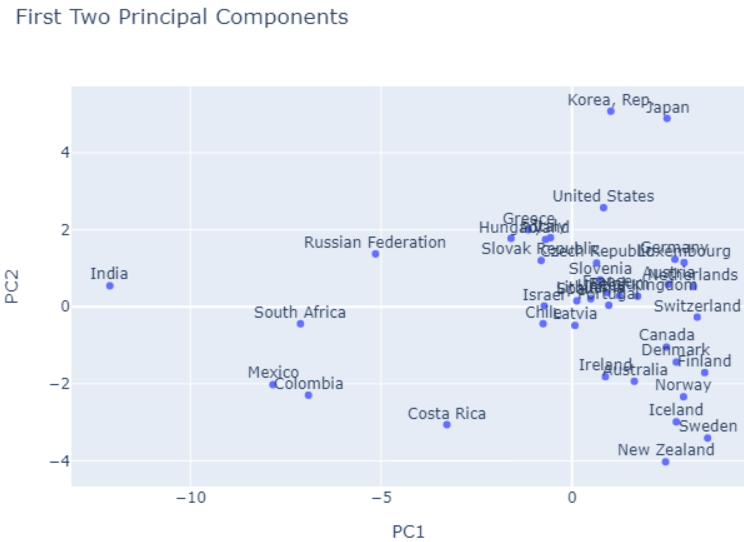

Figure 2: Countries plotted in Top 2 Principal Component Dimensions

Some of the labels are hard to read in the closely clustered areas but this helps to show the spread of the data. It also allows for exploration into what original factors are driving the principal components. For example, India is a significant outlier in the first principal component. Some of the factors driving this are a low proportion of the population using the internet, a large amount of the GDP deriving from agriculture, and a high prevalence of undernourishment. In the scaled data, India is at least 4.5 standard deviations away from 0 in all of these categories and they all contribute significantly in the first principal component.

This graph also shows off one of the benefits of PCA, as these very significant factors do not play a large role in the second principal component. That way the clustering algorithm will be able to detect differences in other factors to give a fuller picture of the separations of the countries. In principal component 2, Japan scores very highly due to a high number of hospital beds per 1,000 people. On the other side, New Zealand's high renewable energy usage puts it on the far negative side.

The ESG data was then transformed into the 11 principal components and K-means clustering was performed on this reduced-dimension data. In order to determine the best number of clusters, 4 metrics were employed. The metrics are briefly described below [4].

1. Silhouette Score

   This compares the mean distance between a sample and the points in its cluster to the mean distance between a sample and the next nearest cluster.

2. Calinski-Harabasz Index

   Also known as the variance ratio criterion, this is the ratio of between-clusters dispersion and within-cluster dispersion.

3. Davies-Bouldin Index

   This index compares the distance between clusters with the size of the clusters themselves.

4. Within-Cluster-Sum-of-Squares

   This is the sum of the squared distances from all points to their cluster centers.

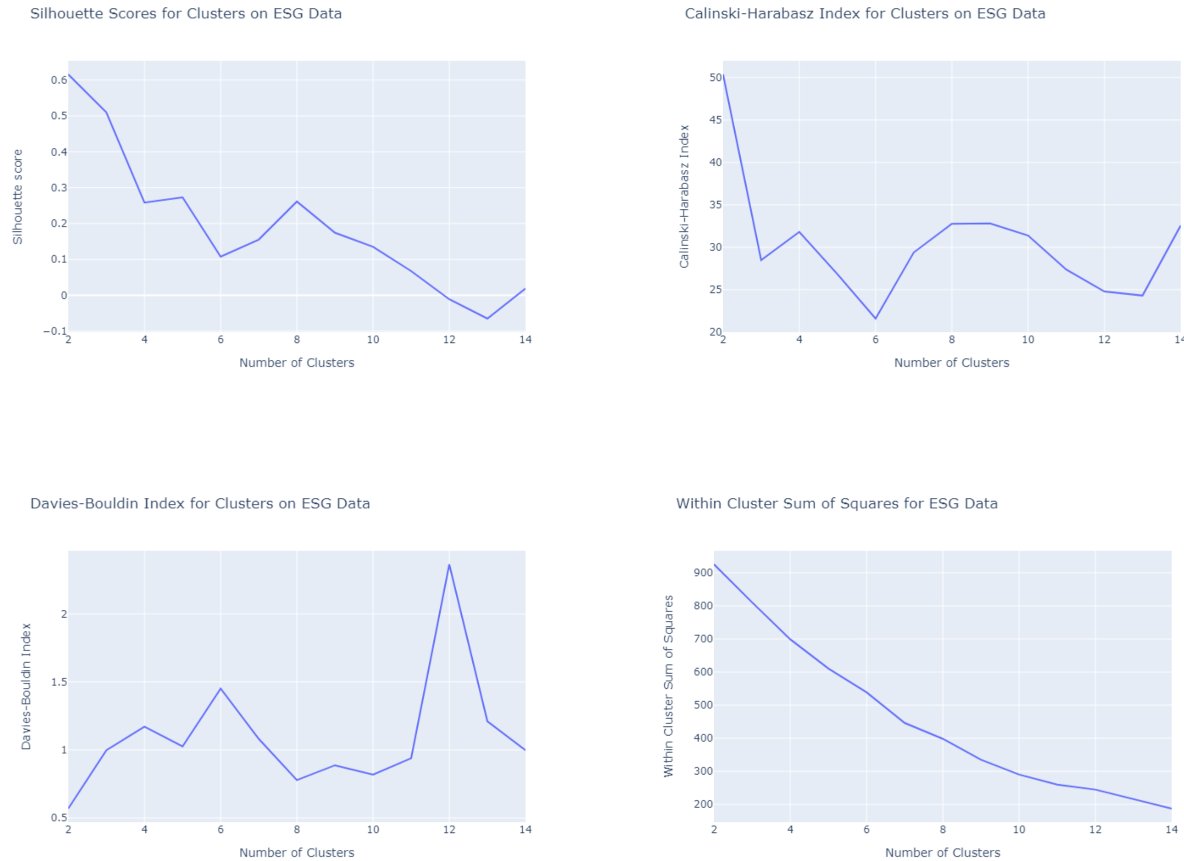

Figure 3: Clustering Performance Metrics

With silhouette scores and CH index, higher numbers represent better performance. DB index and WCSS, on the other hand, see better performing models with lower numbers. Plots for each metric against number of cluster are shown in Figure 3.

From the figures, it is clear the scores aren't always in agreement. One number of clusters does seem to perform well across all the metrics. That is 8 clusters. This seems to give relative segmentation of the countries without creating too many clusters that contain only 1 country.

Once the countries were assigned to clusters, predictions on future yields were performed. The clusters were leveraged in the validation of the time series models used for predictions. Instead of just fitting a model to each country and looking at the parameters which optimize the fit for that country, the parameters which optimized the fit across all the countries in a cluster were chosen to be in the final model. This was done in order to grow the sample size in an attempt to reduce the variability in the yield data.

Originally, it was thought that a similar process could be done to predict a country's bond rating for the next year as well. However, it was discovered that there is very little variability in country bond ratings from year to year. Therefore, the best prediction of the next bond rating is simply whatever the current rating is. These were not looked at further in this analysis.

To predict the future bond yields, the first method explored was exponential smoothing. Simple exponential smoothing applies weights to previous data that exponentially decrease over time. This allows more recent data to weigh more heavily in the current prediction. Commonly, exponential smoothing incorporates additional terms that represent trend and seasonality. This version of exponential smoothing is referred to as triple exponential smoothing or Holt-Winters.

In this application, it was necessary to use Holt-Winters. The seasonality component is a natural addition given that the yield data is collected quarterly. The trend is also important as bond yields have steadily declined over the past 20 years for most countries. This violates a property called stationarity that must be present for a simple exponential smoothing model. Incorporating a trend into the model, like in Holt-Winters, often solves this requirement because after the trend is accounted for the data is more likely to meet the requirements for stationarity. The equations for the additive method of Holt-Winters are shown below.

$$\hat{y}_{t+h|t} = \ell_t + hb_t + s_{t+h-m(k+1)}$$

$$\ell_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(\ell_{t-1} + b_{t-1})$$

$$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1}$$

$$s_t = \gamma(y_t - \ell_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m}$$

Defining the overall terms we have: $\hat{y}_{t+h|t}$ is the forecast, ℓ_t is the level, b_t is the trend, and s_t is the seasonality component all for a specified time t. The smoothing parameters α, β^*, and γ correspond to the level, trend, and seasonality components respectively. The other terms in the equations are m which represents how many seasonal components there are and h which is how many time periods ahead the forecast is.

To fit the models, the initial parameters were fixed based on country, and the smoothing parameters were allowed to vary. This is because each country starts at a different bond yield and it is important to account for these different starting points to avoid skewing the model results. Models were first fit on the countries themselves without taking into account the clustering results. This was done to establish a baseline, but in the process it turned out the result that the smoothing parameters for all the countries were basically the same. As shown in the table below, no smoothing ended up being done. Once the initial parameters were set, the historical yields did not matter beyond the previous quarter. Therefore, clusters were not incorporated into any exponential smoothing model fits as it would not have made a difference.

Country	Initial Level	Initial Trend	Initial Seasons
India	8.24	-0.05	(0.08, 0.08, 0.01, 0.05)
Colombia	15.75	-0.11	(-0.23, -0.07, -0.07, -0.16)
Mexico	10.56	-0.05	(0.18, 0.08, 0.16, 0.02)

Country	Smoothing Level	Smoothing Trend	Smoothing Seasons
India	1	9.3e-18	1.4e-8
Colombia	1	2.2e-11	6.6e-10
Mexico	1	0	7.8e-9

Table 1: Exponential Smoothing Parameters For Select Countries

Figure 4 shows an overlay of Mexico's actual yields with the smoothing model. The ES predictions appear to be a mirror image of the actual yield, just one quarter off and slightly lower. Overall, this is still an okay prediction but it likely could be improved.

A second time-series modeling approach was used to try to predict the future bond yields. This was Auto-Regressive Integrated Moving Average (ARIMA) models. ARIMA models are similar to exponential smoothing in that they combine the historical data with different weights in order to make a prediction. The typical ARIMA model is made up of three parts: Auto-regression (AR), Integrative (I), and Moving average (MA). These have parameters p, d, and q respectively which dictate how far back the model looks

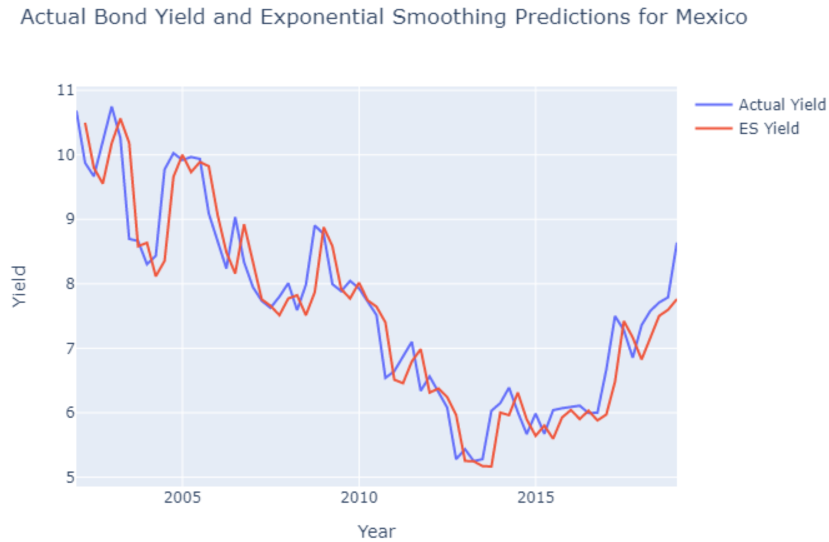

Figure 4: Exponential Smoothing Fit for Mexico

for AR and MA and how many degrees of differencing for I. ARIMA can also be made to be seasonal by adding 4 additional terms P, D, Q, and m where the first three perform similar functions to their lowercase counterparts and m represents the number of seasons in a full period.

In initial fits of these models, it was found that different countries were fit better under different parameters unlike the exponential smoothing models. Therefore, to train the models, the clusters from the ESG data were used to split up the countries. The training procedure is as follows:

1. Loop through yield data, filtering by a specific cluster
   (a) Loop through yield data for each country within the cluster
      i. Fit ARIMA model for a given set of parameters
      ii. Determine Mean Squared Error (MSE) for specific parameter set
   (b) Sum MSEs for specific parameter sets over all countries in a cluster
   (c) Determine which parameter set produced the lowest total MSE
   (d) Refit final models for each country within the cluster using the parameter set from above

Mean squared error was chosen as the optimization metric because it discourages being extremely wrong on a given prediction. Because the terms are squared, errors of more than 1% yield are heavily penalized. This is important because when making these investments, being off by a large amount once can devastate someone financially while being slightly wrong often is easier to recover from. From the training procedure above, the following ARIMA parameters were found as the best performing for each cluster.

Cluster	p	d	q	P	D	Q	s
0	0	1	3	3	1	0	4
1	0	1	0	3	1	0	4
2	0	1	3	1	1	0	4
3	0	1	2	1	1	0	4
4	0	1	2	1	1	0	4
5	0	1	3	3	1	0	4
6	1	1	2	3	1	2	4
7	3	1	0	3	1	1	4

Table 2: ARIMA Parameters for Each Cluster

Looking at Mexico once more, we can see how the ARIMA model performed. In this case, the model cannot infer a starting position from the data and must adapt over the first few years. Once this initialization period ends it seems to be performing quite well. One key difference between this and the exponential smoothing model is that the smoothing model was often under predicting yield because of the negative trend. The ARIMA model is able to better adapt as the yield starts going up around 2013.

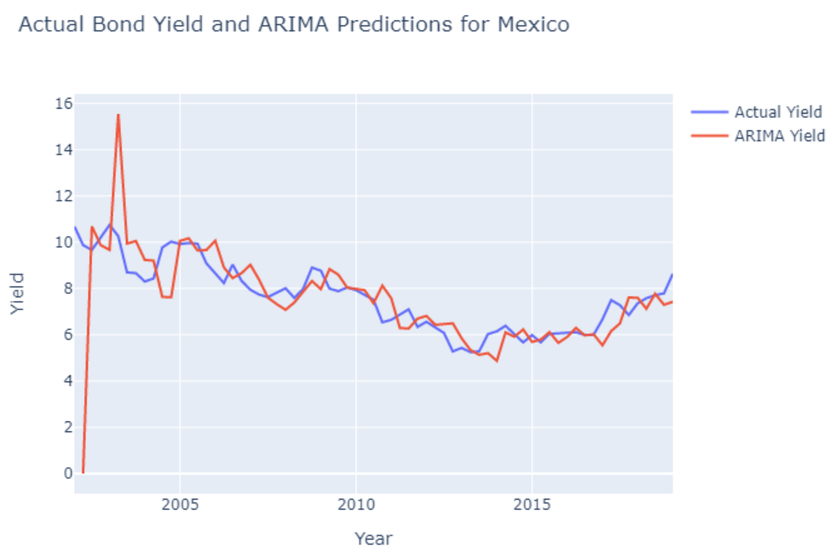

Figure 5: ARIMA Fit for Mexico

5 Evaluation and Final Results

The exponential smoothing and ARIMA models were used to forecast yields for all 4 quarters of 2019. With time series models, it is expected that the more time periods that are forecasted, the worse the error becomes. Therefore, a value of 4 time periods was chosen as it allows us to view one full seasonal period but not more which may artificially drive the error up. These models were also compared to simply using the yield from 2018 Q4 as the prediction. This provides a baseline to see if the models provide any value. Mean squared error for the different prediction methods is shown below. The MSE for just Q1 is shown, in addition to the MSE for all of 2019, due to the note above that error increases the further out the forecast goes.

Model	MSE for 2019 Q1	MSE for all of 2019
2018 Q4	0.082	0.512
Exponential Smoothing	0.072	0.498
ARIMA	0.079	0.479

Table 3: Yield Forecast Model Performance

The table shows that Exponential smoothing performs the best when just predicting the next quarter. However, the ARIMA model which incorporates the ESG clusters performs better when extending the forecast to a full year. This is an interesting conclusion. It seems that the ESG propensity may add meaningful value to long term bond yield predictions although the performance is close enough that it is hard to say for sure.

6 Future Work

The above analysis was just the tip of the iceberg when it comes to looking at ESG metrics and sovereign bonds. One of the largest drawbacks with the analysis, in the author's eyes, was the limited amount of countries that were ultimately analyzed. Only 40 countries of the more than 200 that exist were analyzed and this was a highly skewed group of mainly developed European countries. If data were to become available for additional countries, it would add a lot of value to this analysis.

Additionally, this analysis could be improved by looking at alternative methods. Some ways that could be explored are kernel PCA, clustering via a Gaussian Mixture Model, or an ARIMA model with exogenous regressors. Alternatively, the country credit ratings could be looked at in more depth given that they were left out of this analysis. Instead of looking at them as a time series, they could be looked at through the lens of regression with data such as the ESG data used as the factors.

7 References

[1] S&P Ratings Direct Global Ratings Definitions https://www.maalot.co.il/Publications/GMT20160823145849.pdf

[2] World Bank Sovereign ESG Data https://datatopics.worldbank.org/esg/

[3] Federal Reserve Economic Data https://fred.stlouisfed.org/

[4] Scikit-Learn Clustering evaluation https://scikit-learn.org/stable/modules/clustering.html#clustering-evaluation