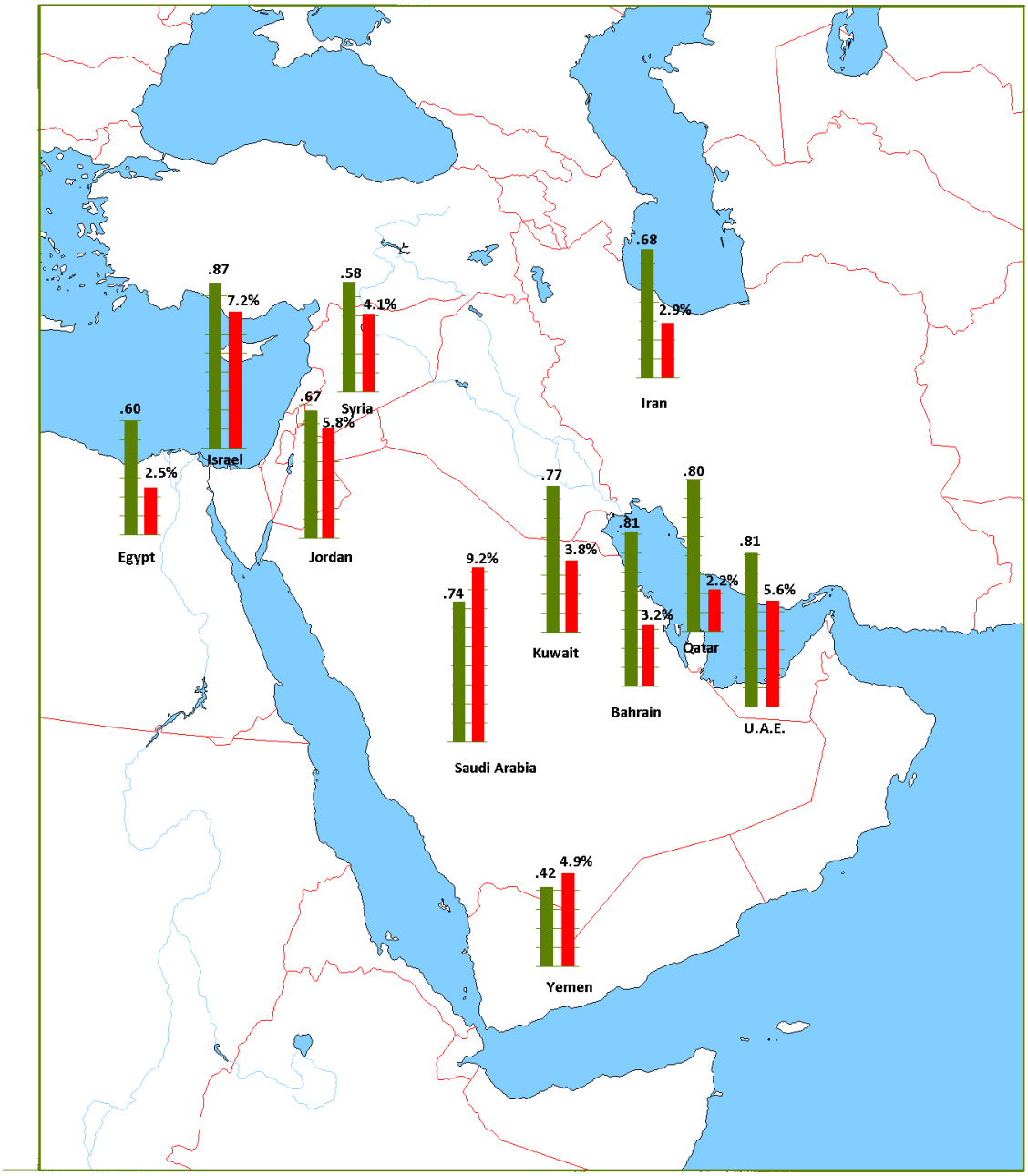
Creating a model to predict the future military expenditure (ME) and human development index (HDI) of nations



HARRY O'BRIEN – 18323075

Contents

Abstract

The objective of this research project was to analyze the relationship between HDI and ME, and create a model that can accurately predict the future HDI and ME of a country based on historical HDI and ME data. The ARIMA function produced the most accurate model.

Introduction

HDI is a unit for measuring the capabilities of people within a region, and was created to emphasize measure the development of peoples' capabilities within a region as opposed to other figures such as economic growth. ME of a nation is influenced by many factors such as the strength of allies, geography, proximity to hostile nation states, and culture. It is a topic I am particularly interested in. Seeing as HDI is considered the optimum measure of national development, and ME is an area I am particularly interested in, it seemed fit to use these 2 measures for this assignment. Once I had cleaned and analyzed my time series data I built 3 models and measured their performance.

Exploratory Analysis

To clean the dataset numerous countries with incomplete HDI and ME data were omitted, which amounted to 64 countries. These were predominantly developing regions, so the

models in this project will be stronger at predicting developed nations HDI & ME than developing nations.

Methodology

I used the ets(), arima(), and HoltWinters() functions in R because these 3 functions provide a broad overview of the functions available for time series forecasting. They are the 3 most prominent functions mentioned in the R documentation for time series analysis (Hyndman, 2021). This same documentation provided the accuracy() function, which was used to evaluate and compare the models.

Results

All of the models were pretty accurate, as can be seen by the low % of errors in the figure on the right. MAPE (mean absolute percentage error) was used as the figure for comparison due to its comprehensive view of the accuracy of the models. The ARIMA function was the most accurate, followed by the ETS and Holt Winters. This was to be expected because an ARIMA model produces forecasts based on previous values (AR terms) and the errors made by previous predictions (MA terms). This usually allows the model to quickly compensate for sudden changes in trend, producing a more accurate forecast. My hypothesis for why this applies here is because in a state of war HDI will regress as rapidly as ME will increase, so the model that is the most sensitive to these fluctuations should be expected to perform the best. This is supported by the fact that the models with the higher ACF1 performed better, because ACF1 is a measure of how much is the current value influenced by the previous

```
1. accuracy_ETS
##                       ME     RMSE      MAE       MPE     MAPE      MASE
## Training set -5.700792e-16 2.350760 1.832038 -1.450730 9.163121 0.7275384
## Test set      2.862396e-03 2.499219 1.907167 -1.679778 9.693489 0.7573733
##                    ACF1 Theil's U
## Training set 0.07622679        NA
## Test set     0.02075756 0.6372819

2. accuracy_holt
##                       ME     RMSE      MAE       MPE     MAPE      MASE
## Training set -0.1687680 2.466238 1.931718 -2.289210 9.710092 0.7671232
## Test set     -0.1416884 2.514707 1.898007 -2.382097 9.721400 0.7537358
##                    ACF1
## Training set 0.08209031
## Test set             NA

3. accuracy_arima
##                       ME     RMSE      MAE       MPE     MAPE      MASE
## Training set 0.169583006 2.343930 1.830812 -0.619336 9.082284 0.7270513
## Test set     0.002419139 2.499097 1.906639 -1.681789 9.691013 0.7571639
##                    ACF1
## Training set -0.0137157
## Test set             NA
```

Figure 1: Results from models using accuracy() function

values in a time series. However, this type of warfare is rare in our modern day world and thus the 2 other models which don't react as quickly display similar accuracy. The closeness of performance of ARIMA and HW was expected because the 2 are very similar.

Conclusion

The ETS function is best used in cases where fluctuations are rare and not long lasting. This was not the case with our dataset, which found ARIMA modelling to be the most accurate, followed closely by Holt-Winters. ARIMA allows the user to add independent regressors, but HW is more common and straightforward. For example, the UK's Office for National Statistics advises the use of HW for modelling datasets of 3-5 years because "it doesn't require the lengthy manual modelling of the 'ARIMA' approach" (Statistics, 2007).

Reflection

I found this project very interesting. I learned more about modelling theory doing these projects than I did from the lectures, probably because I learn from doing (hence why I did MSISS). I learned a lot more about how to code from the tutorials, and consequently I did nothing in this project that I hadn't done outside of the module.

I tried out a lot of different methods for analyzing this dataset, such as examining the correlation between the 2 variables overtime and k-means clustering. These methods were ultimately abandoned because they provided little new information to me and would've just been a demonstration of the fact that I can do those things, but I have demonstrated that in the tutorials and I didn't want to waste this opportunity to actually learn something new and of value. I find macroeconomics fascinating and modelling is by far the most relevant part of this module to that area so it made sense for me to focus my efforts there.

Were I to do this project again, I would have included more variables in my dataset and covered a shorter period of time. By using a dataset so large yet with so few variables my model is accurate on paper, but in reality it cannot be used to predict anything. I wish I had created a model that could at least be of some use to me in the real world. Modelling is so

cool and I am doing the derivatives module next year (against the advice of the 4th years who say it's really hard) because I want to build a high frequency trading bot, which is essentially just a very accurate model that makes bets. I would've flipped my dataset in the sense that I would've picked variables that are so new that they only go back a few months, but I would've picked dozens of them. This assignment gave me a much greater appreciation of the importance of selecting a good dataset before building the model. I feel I pigeonholed myself by picking such a poor dataset, but such is life.

Overall, I feel I learned more from this module than any other module. The 4 hour long stats labs were really tough but were some of the most productive deep work sessions I have done in my life, and the stamina alone I built from doing those labs are worth their weight in gold. I was initially confused and overwhelmed by the lack of guidance for this project which is reflected in the fact that I am submitting this about 90 minutes before the deadline, but this freedom allowed me to pick a topic I was interested in and learn as much as possible. We live in a world where there is an abundance of information but a scarcity of desire to learn, and this module and this project specifically created a strong desire in me to learn statistical methods, and for that I am extremely grateful. It is something no teacher has ever accomplished with me before.

Bibliography

Hyndman, R. J., 2021. *CRAN Task View: Time Series Analysis.* [Online]
Available at: https://cran.r-project.org/web/views/TimeSeries.html

Statistics, U. O. f. N., 2007. *Guide to Seasonal Adjustment with X-12-ARIMA,* London: ONS Methodology and Statistical Development.