Margherita Piana, Avery Bantrup

Final Project EK 125

Biases

https://www.powermetrix.com/2019/03/common-reasons-your-customers-energy-bills-may-be-inaccurate/

Electricity and energy companies are directly affected by biases in large sets of data, if they are not carefully addressed, these biases can result in the company losing money from these errors. It's important that businesses involved in the energy industry keep a close eye on the data they receive from different sources, especially so, in order to accurately charge customers for the amount of energy that they are actually consuming and should by paying for. An article written by Powermetrix reports some of the specific holes that these companies can fall into. If inaccurate data is collected from whatever source and communicated incorrectly to the company, the company can either upcharge or undercharge their customers for their services. This issue may appear easier to fix than it actually is, once the inaccurate billing information is sent out, resolving the issue becomes complicated; the customer may cost the company money from the process of calling the company to get the correct bill or even just from losing a customer due to the mishap.

In this instance, false data is typically collected from faulty or damaged equipment. Because the equipment is not fully functional, it uses more energy to do the tasks it would normally perform with a low level of electricity. One example of this equipment might be a utility meter that would typically be installed by the company upon purchase. If this installation was done incorrectly or happens to break over time, this will lead to inaccurate data collection. This is what causes the big issues and leads to money being lost.

If data is collected manually, this causes even larger gaps in accuracy with the data, simply due to human error.

It is vital for electricity and energy companies to ensure that the data they collect and use is accurate in order to protect the integrity of the company and consistently charge their customers for the energy they use.

Background on the problem

Link to data set: http://data.un.org/Data.aspx?d=EDATA&f=cmID%3aEL.
Population data link: https://datatopics.worldbank.org/world-development-indicators/

Electricity use all over the world has consistently grown over time causing an even greater strain on the environment. This energy consumption can be broken down into categories of the different sources that cause this strain. The data set used in this project includes data from a variety of these sources, such as: gross electricity production, combustible fuels, hydroelectricity, net electricity production, imports, households, agriculture. Values for these sources are reported for each country in the world by year from 2009-2019. This overall increase in electricity consumption will over time begin to affect the world more and more drastically, many systems are currently in place to fight the environmental destruction that is produced. Electricity is one of these factors that will increase the greenhouse gas emissions in our atmosphere.

It is important to understand how these trends will continue to grow in order to better predict and prepare for the future state of our environment. Through this project, we are able to make these predictions and elaborate on what the most concerning factors are for specific countries. For those working to fight climate change, these conclusions are especially useful and applicable to every part of the world. Individual countries will be able to see a physical

representation of their growth, and understand what they should do towards becoming a more sustainable part of the world.

More specifically, this project observes electricity consumption from the five smallest and largest countries in the world, giving us a wide range of circumstances that affect this consumption. However, the entire data set includes data for every country in the world, and our project can be used to analyze these important things for every country as well. In smaller underdeveloped countries, we are able to better understand the distribution of this energy and make changes to combat the aspects adding to the deterioration of the environment.
Inferences:

Scrubbing is one of the most important parts of coding when analyzing a set of data, it includes considering outliers in the set of data, understanding how to address them, and normalizing the data. We started by downloading our data file into an .xlsx file, which had to be done in two steps since the website was only able to download half of the data at a time. We then had to transfer it into matlab in the form of a structure. Vectors have been created for each field of the structure and then they have been randomized.

```
undata= table2struct(readtable('UN_data_final.xlsx'));
```

```
Warning: Column headers from the file were modified to make them valid MATLAB identifiers before creating variable names
for the table. The original column headers are saved in the VariableDescriptions property.
Set 'VariableNamingRule' to 'preserve' to use the original column headers as table variable names.
```

```
footnotes=[];
comodity=[];
year=[];
uni=[];
amount=[];
country=[];

country=[country,undata(:).CountryOrArea];
comodity=[comodity, undata(:).Commodity_Transaction];
year=[year,undata(:).Year];
uni=[uni,undata(:).Unit];
amount=[amount,undata(:).Quantity];
footnotes=[footnotes,undata(:).QuantityFootnotes];
```

The following image is an example of the lines of code used to randomize the data for each of the vectors created.

```
rand_position=randperm(length(country));
for k= 1:length(country)
    country_randomply_placed(k)= country(rand_position(k));
end
```

After having randomized the data, the NaN values were counted for each vector of data and replaced when needed. For footnotes it has been decided to delete all the NaN values since they will have no significance or effect on our prediction.

```
footnotes_randomly_placed(isnan(footnotes_randomly_placed))=[];

count=0;
for i=1:length(footnotes_randomly_placed)
    if isnan(footnotes_randomly_placed(i))==1
    count=count+1;
    end
end
fprintf("The number of NaN values in the footnotes vector are now %d",count)
```

```
The number of NaN values in the footnotes vector are now 0
```

The NaN were counted to be zero for country and commodity and uni. The NaN values for the year vector were four. Since their exact position is not known, we had to loop through and calculate the mean of the fourth and fifth values closest to the NaN value to replace them. We used this technique because each year represents a different country, so replacing them with the overall mean might create some error in the prediction. The same code was used to find the NaN values present in the amount vector, since again they were counted to be four.

```
count=0;
for a=1:length(year)
    if isnan(year(a))==1
    count=count+1;
    end
end
fprintf("The number of NaN values in the year vector are %d",count)
```

```
The number of NaN values in the year vector are 4
```

```
for a=1:length(year)
    if (isnan(year(a)))
        if (a>1 && a<length(year))
            mean_temp=(year(a+4)+year(a-4))/2;
         elseif (a==1)
             mean_temp=(year(a+4)+year(a+5))/2;
         else
            mean_temp=(year(a-4)+year(a-5))/2;
    end
        year(a)=mean_temp;
    end
    end

count=0;
for a=1:length(year)
    if isnan(year(a))==1
    count=count+1;
    end
end
fprintf("The number of NaN values in the year vector are now %d",count)
```

```
The number of NaN values in the year vector are now 0
```

The randomized and scrubbed data were converted to a structure again and then we started normalizing the data by considering the year and electricity consumed for five of the largest and five of the smallest countries in the world. The countries considered are: China, India, United States, Indonesia, Pakistan, Singapore, Denmark, Finland, Slovakia and Pakistan.

The Machine Learning Toolbox app Regression Learner has then been used to make a prediction of the electricity consumption for each country. The responsive variable is the electricity while the predictor has been set to be the years. Looking at the models, it has been decided to use the squared exponential for all the countries because it had a low value for the RMSE test. The model for each country has then been used to make a prediction on the electricity consumption from 2020 to 2030. The following lines of code have been used

to create the prediction graph and download the image taken of the squared exponential model for each country.

```
china_prediction= 2020:1:2030;
model_china=[];
for i=1:length(china_prediction)
china_future_table=table(china_prediction(i),'VariableNames',{'year'});
model_china(i)= electricity_china_prediction.predictFcn(china_future_table);
end
plot(china_prediction,model_china,'.-')
title('Prediction of the consumption of electricity of china')
xlabel('year')
ylabel('Electricity')
%image of model used to predict china through machine learning toolbox
china_image=imread('china_e_graph.jpg');
imshow(china_image,[])
```

Here follow the prediction graphs and the squared regression model for the five biggest countries.

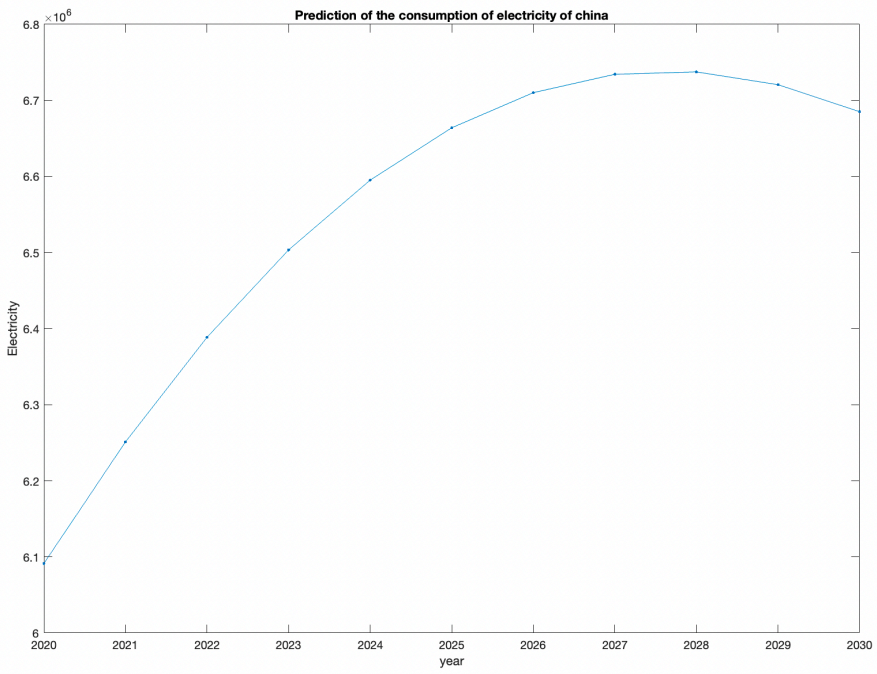

Figure 1: Prediction of electricity consumption from 2020 to 2030 for China

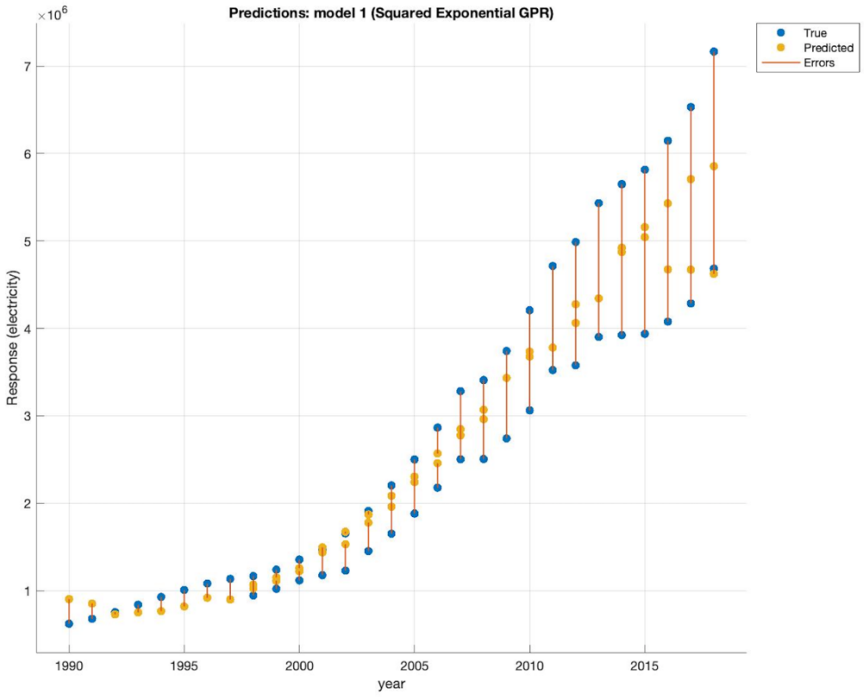

Figure 2: Squared exponential model for China

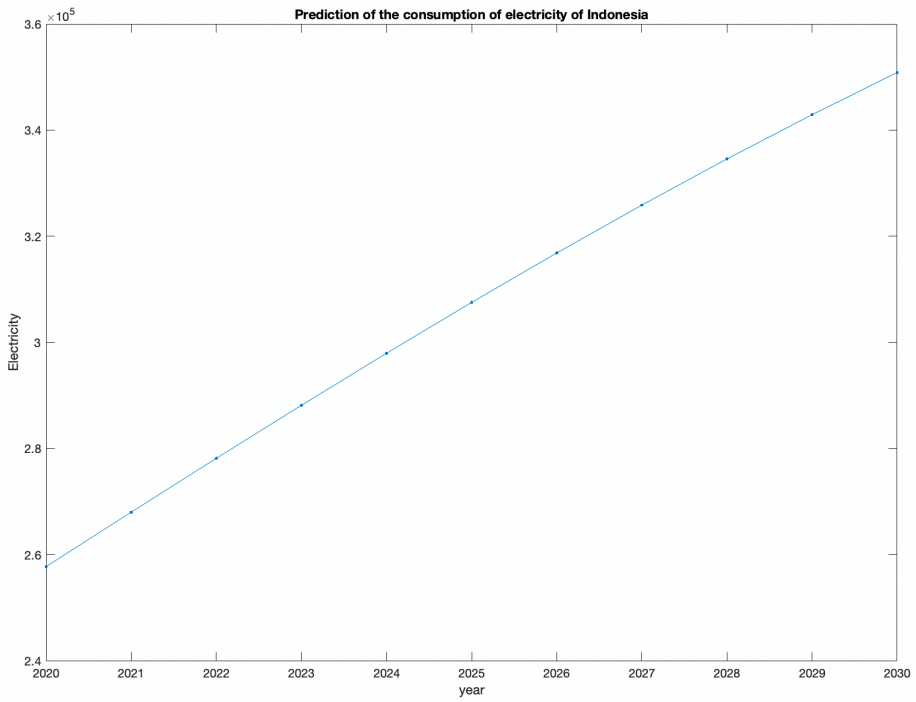

Figure 3: Prediction of electricity consumption from 2020 to 2030 for Indonesia

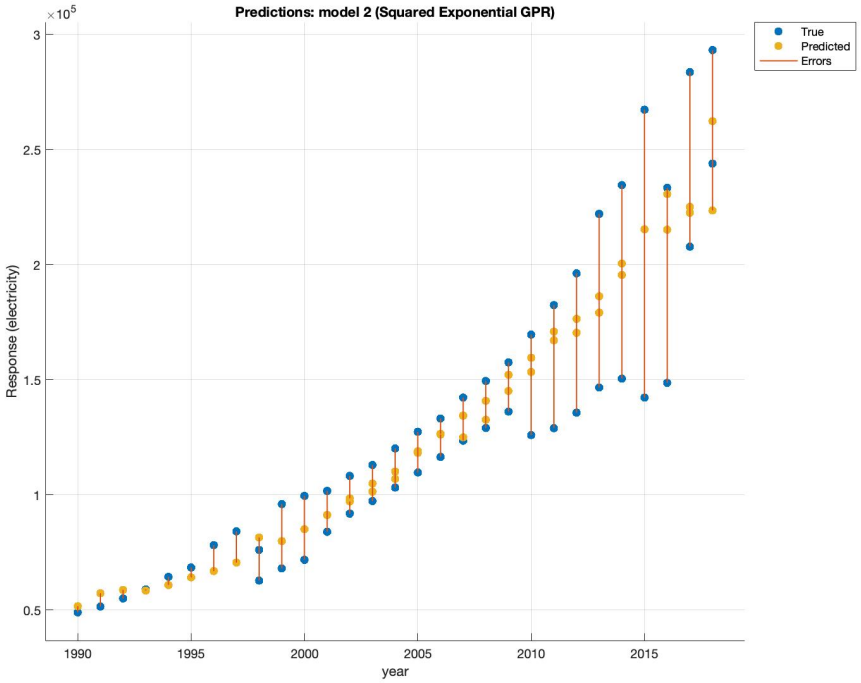

Figure 4: Squared exponential model for Indonesia

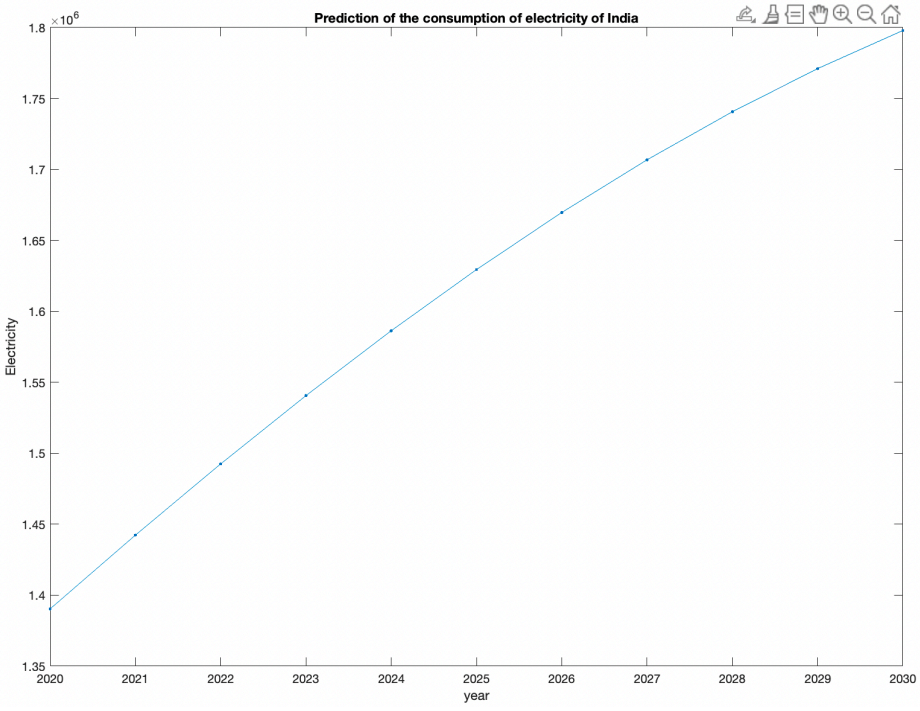

Figure 5: Prediction of electricity consumption from 2020 to 2030 for India

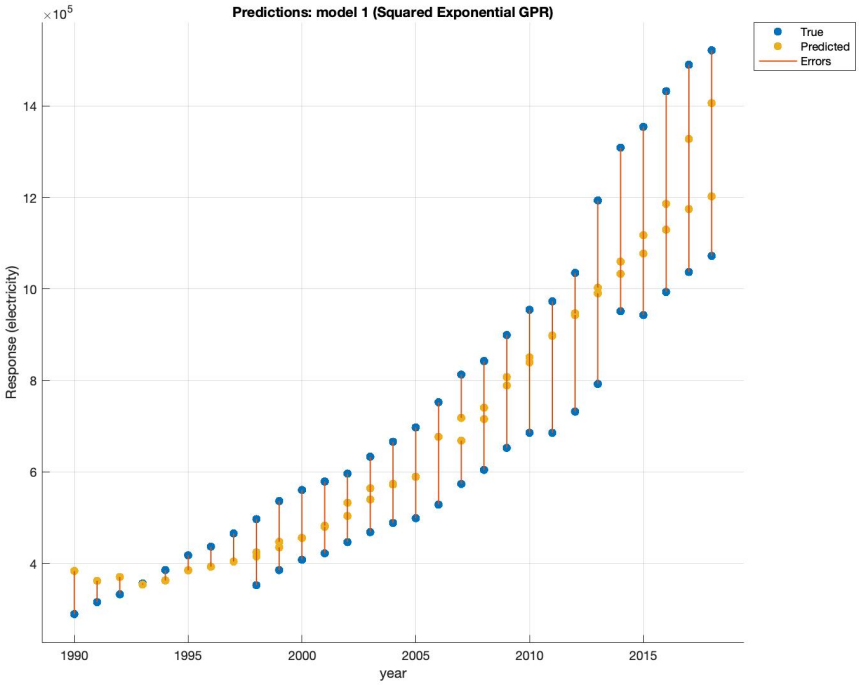

Figure 6: Squared exponential model for India

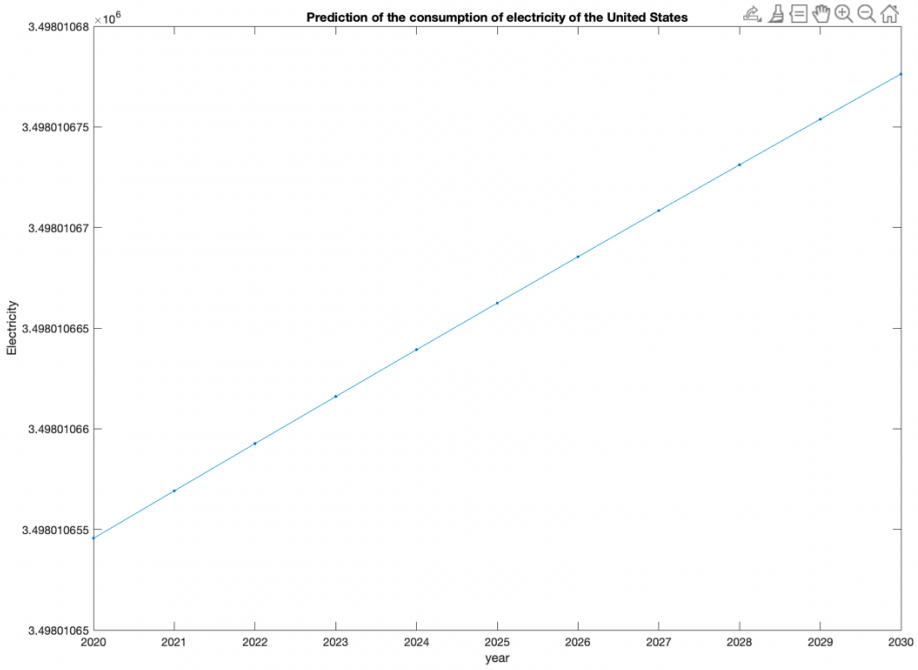

Figure 7: Prediction of electricity consumption from 2020 to 2030 for United States

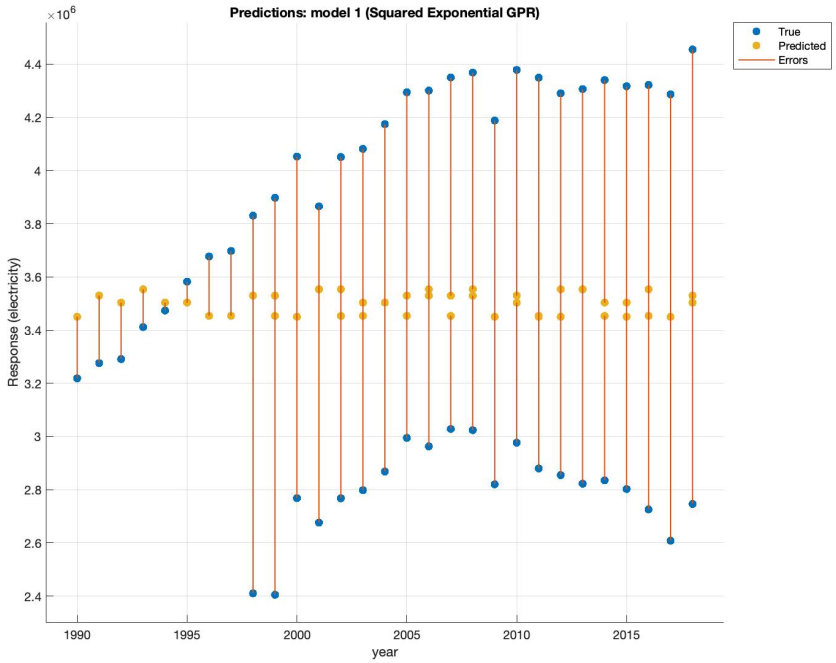

Figure 8: Squared exponential model for the United States

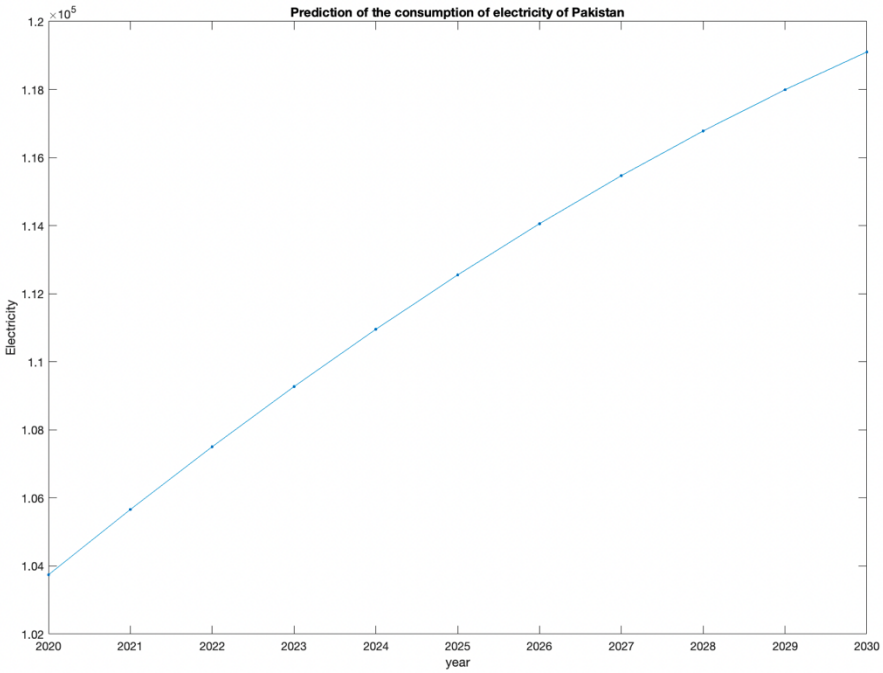

Figure 9: Prediction of electricity consumption from 2020 to 2030 for Pakistan

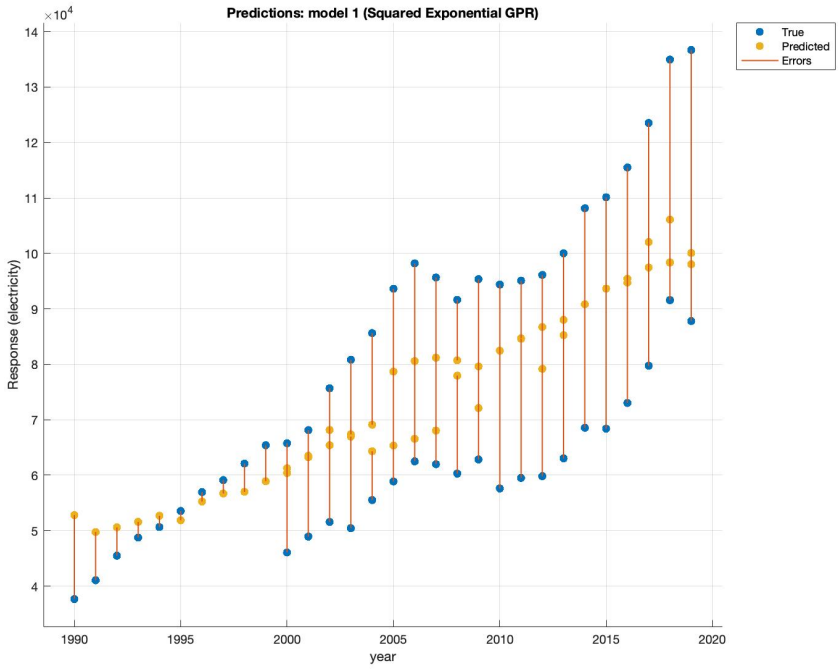

Figure 10: Squared exponential model for Pakistan

The five largest countries we observed were China, India, United States, Indonesia, and Pakistan. For each country, we used the squared exponential model to show a line graph to make a prediction about the future use of electricity. Another version of the graph allowed us to look at the current electricity consumption for each country. India and China have similar trends for their future prediction graphs and their squared exponential models are also similar. Unfortunately, they are similar in the way that both of their predicted data shows how drastically poor their production rates are for the environment. This conclusion is not surprising, seeing as China has 1.4 billion people living there and India has 1.38 billion people; these huge masses will surely be the driving force behind these predictions. More specifically, our model predicts that China will hit its maximum consumption in 2028 with a value of 6.7 x 10^6 kilowatt hours per million; the data then begins to decrease slowly until 2030, which could be a result of a data bias found in the overall set used. India is reported to hit its maximum in 2030 at 1.8 x 10^6 kilowatt hours per million. While the graphs for China and India are very similar, the difference in the maximum values is a result of China being a more technology oriented country.

We also noticed very close similarities between The United States, Indonesia, and Pakistan, which all produce increasing prediction graphs, although they vary within their squared exponential models. The variation in the exponential models is likely due to cultural differences within each country and to what extent each of them uses technology in everyday life. The United States will continue to increase its consumption, with a maximum predicted value of 3.49 x 10^6 KWh/million in 2030, Indonesia and Pakistan with the same trends of increase have max predicted values of 3.5 x 10^5 KWh/million and 1.19 x 10^5 KWh/million, respectively, both in 2030. All of these similar values and trends are supported by the populations of these countries as well, United States with 329.5 million people, Indonesia with 273.5, and Pakistan with 220.9.

The yellow points used to make predictions in the squared exponential graphs are in between the amount of consumed electricity, because for each year there are multiple versions of the

data due to the changes of different types of consumed electricity. Importance of looking at types divides it into sustainable vs non sustainable energy being used. The squared exponential model indicates the strongest results for China and India.

Here follow the prediction graphs and the squared regression model for the five smallest countries.

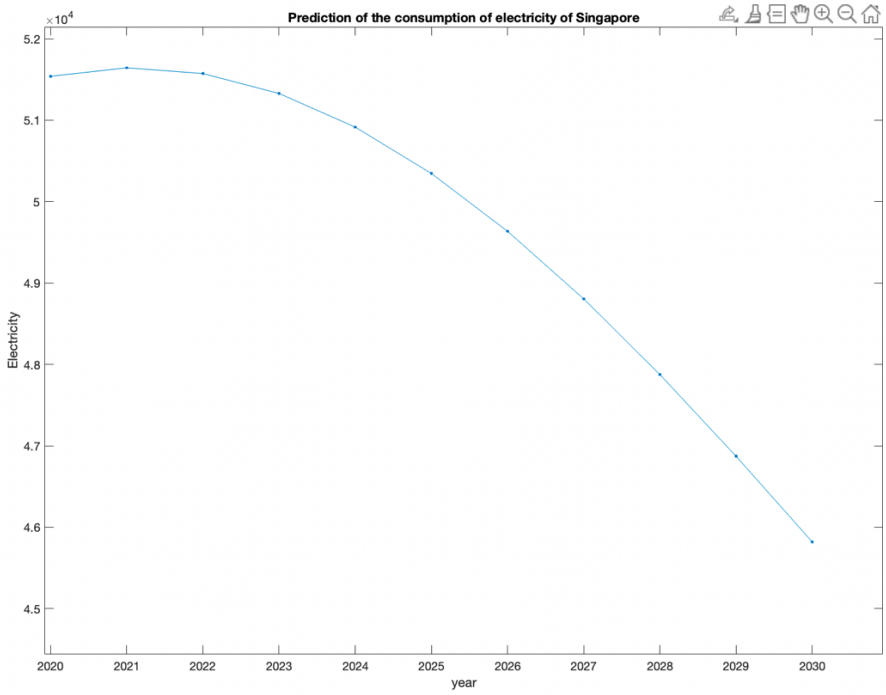

Figure 11: Prediction of electricity consumption from 2020 to 2030 for Singapore

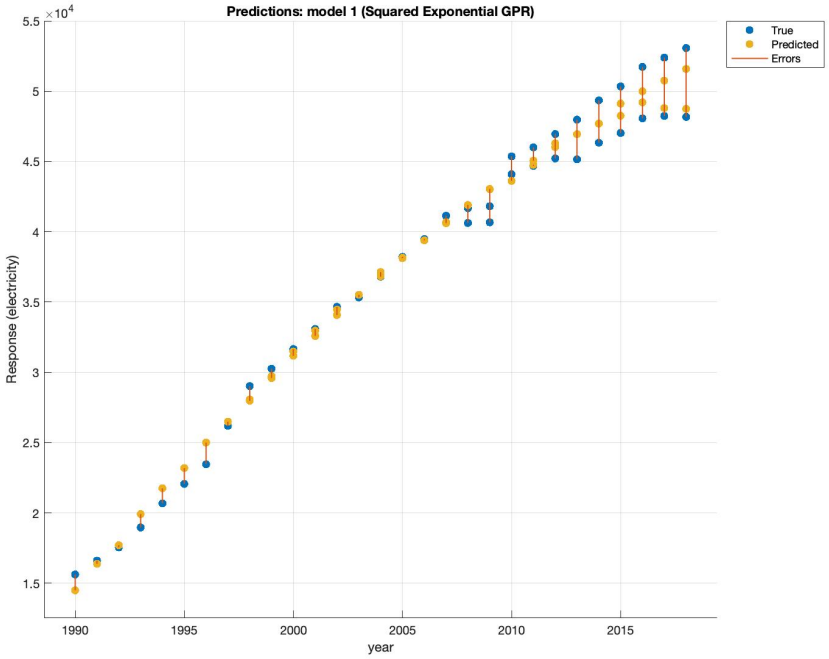

Figure 12: Squared exponential model for Singapore

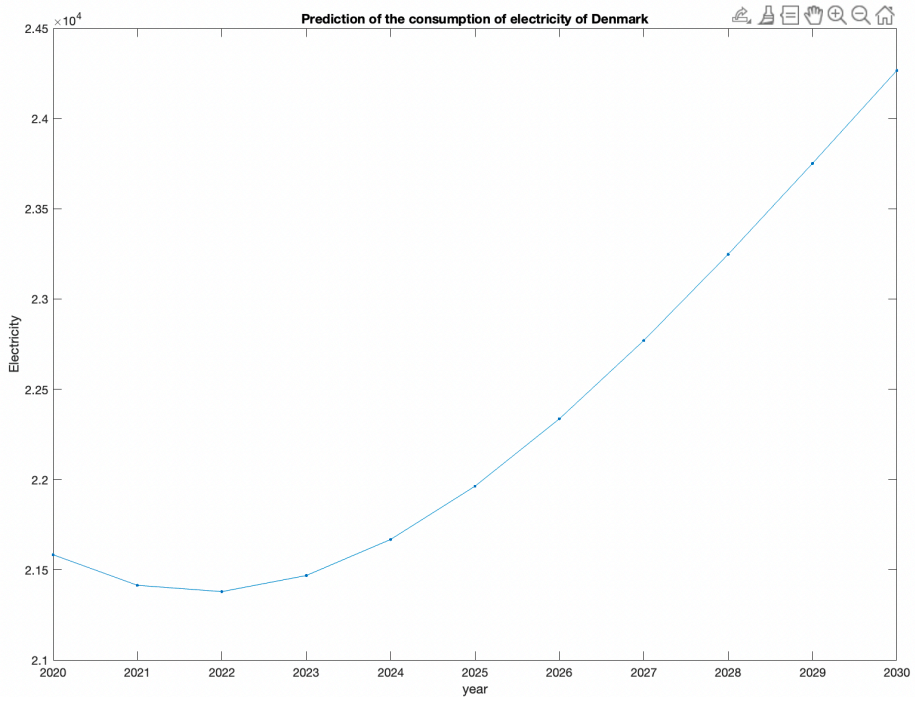

Figure 13: Prediction of electricity consumption from 2020 to 2030 for Denmark

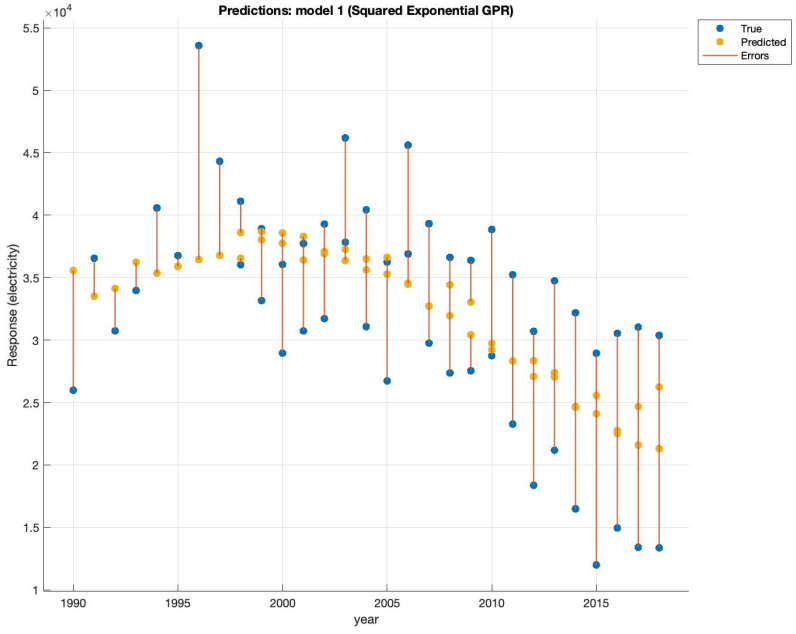

Figure 14: Squared exponential model for Denmark

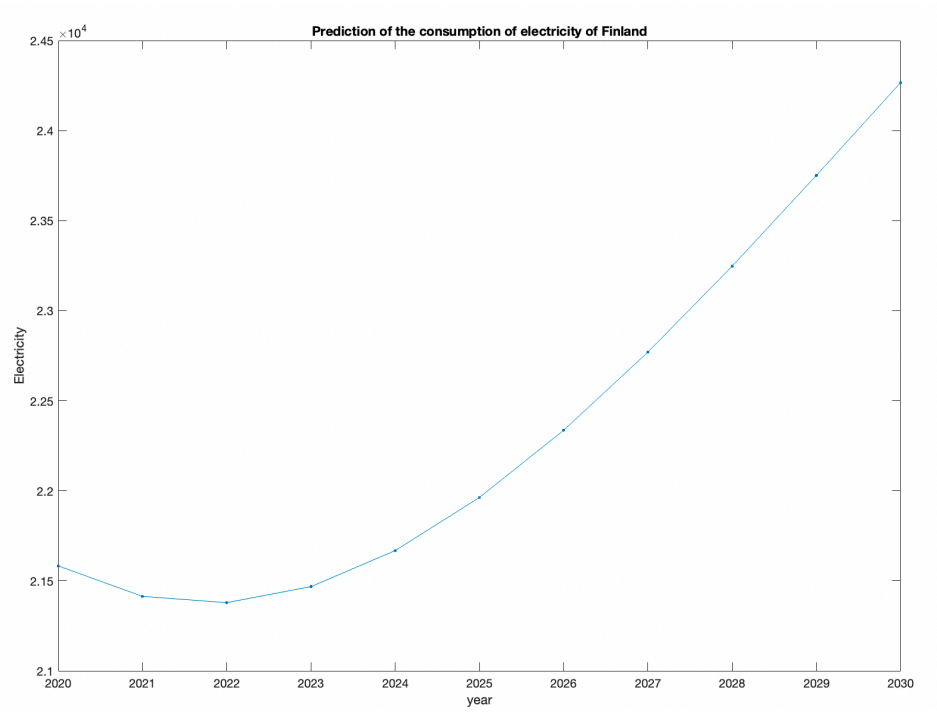

Figure 1: Prediction of electricity consumption from 2020 to 2030 for Finland

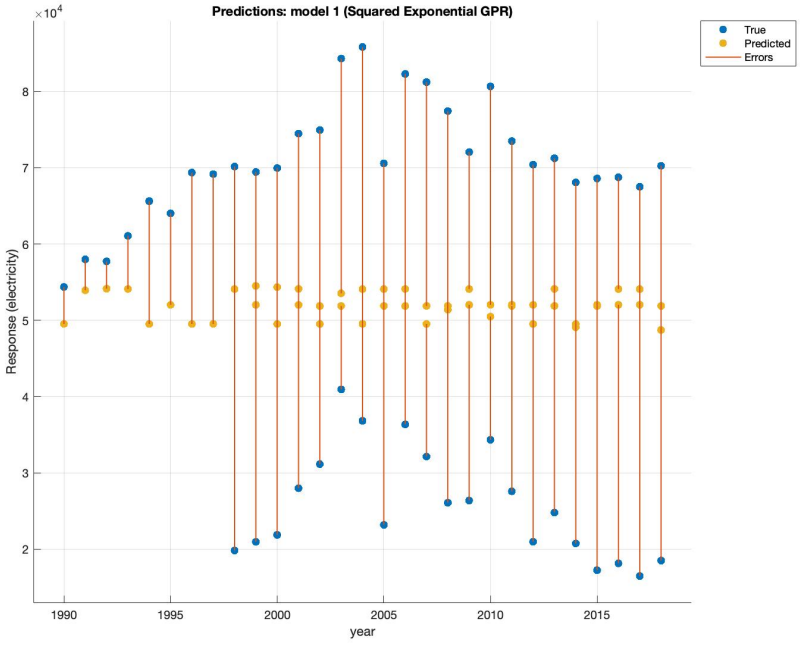

Figure 16: Squared exponential model for Finland

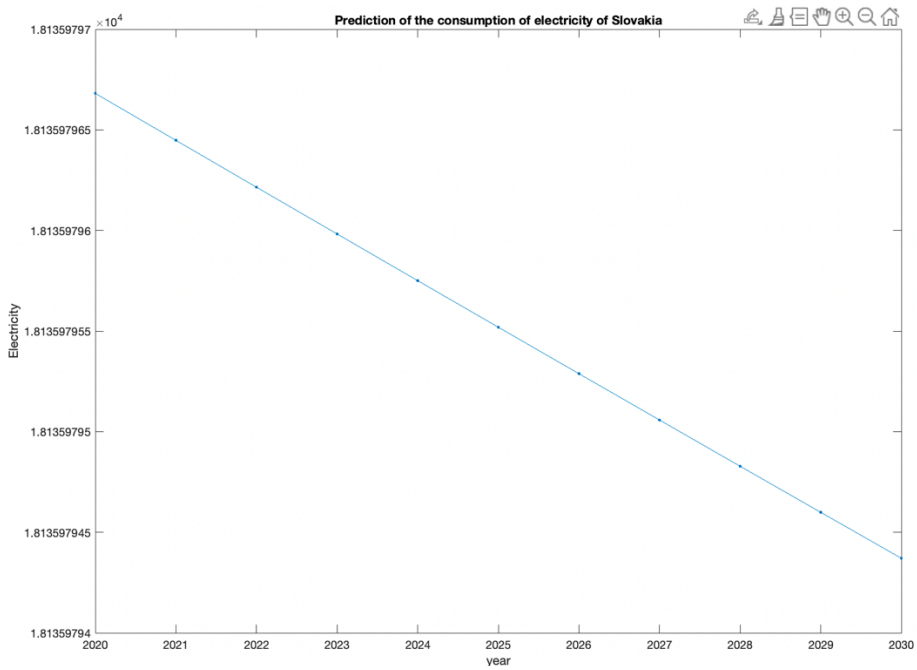

Figure 17: Prediction of electricity consumption from 2020 to 2030 for Slovakia

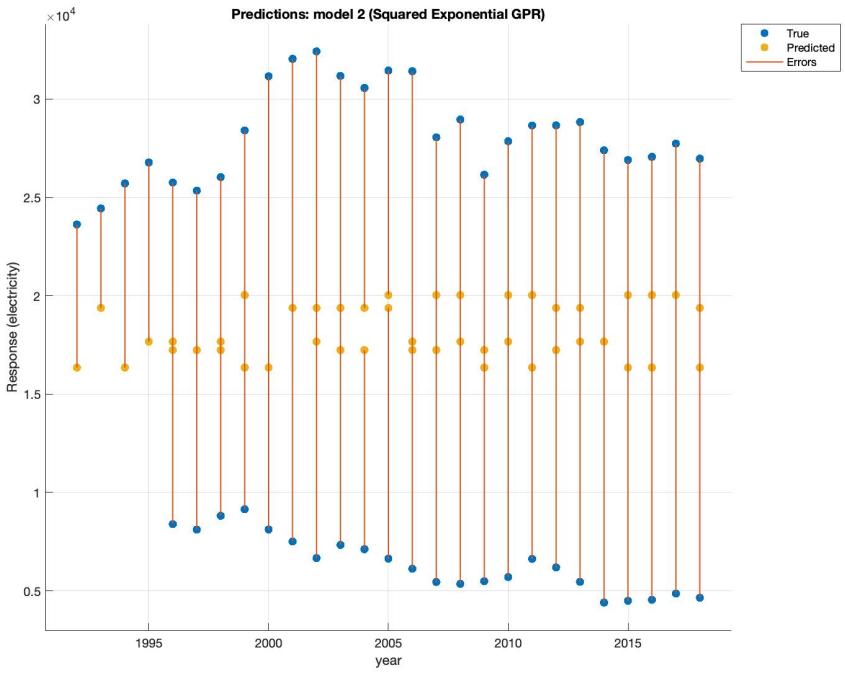

Figure 18: Squared exponential model for Slovakia

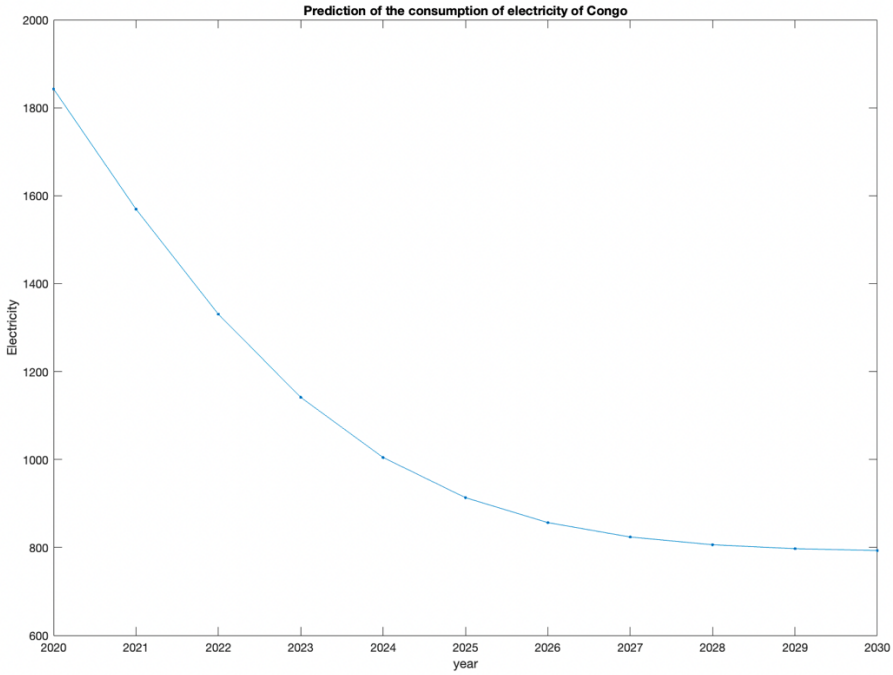

Figure 19: Prediction of electricity consumption from 2020 to 2030 for Congo

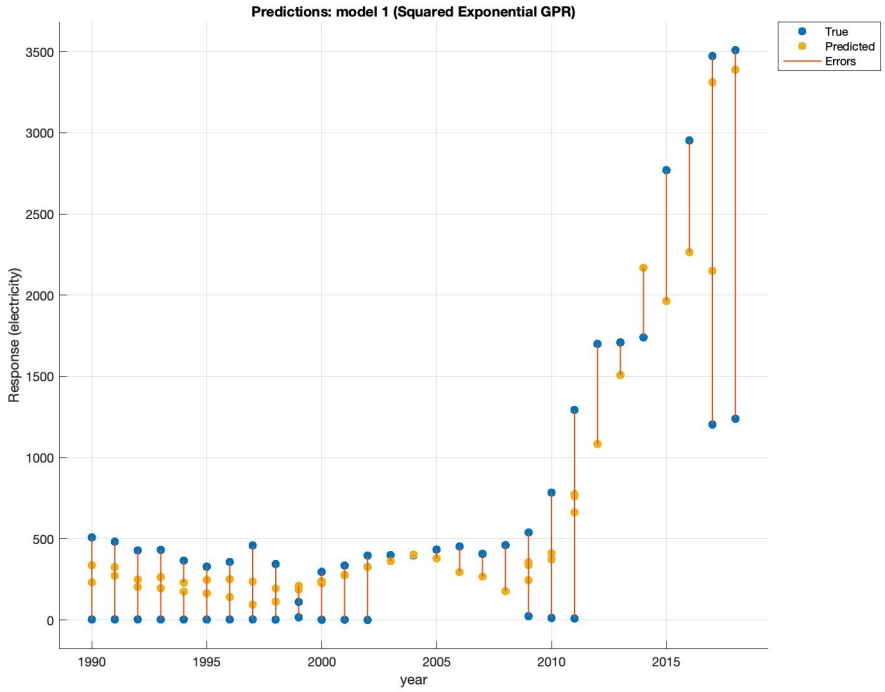

Figure 20: Squared exponential model for Congo

The five smallest countries we observed and their given populations were: Singapore, 5.686 million, Denmark, 5.831 million, Finland, 5.531 million, Congo, 89.56 million, Slovakia, 5.459 million. As we found similarities in the first analysis of the five largest countries, we were able to find trends within this new set of countries as well. Denmark and Finland had the exact same prediction model which decreased slightly until 2022, followed

by a significant increase in consumption. Both countries predicted maximum values in 2030 to be 2.15 x 10^4 KWh/million. This reasonable conclusion is supported by the extremely similar values in population as well as the close proximity of the countries. Both of these countries, specifically, experience months of darkness because of their northern location, which may be a driving force behind their electricity consumption.

Although Singapore is close to the population of these two countries, its prediction model varied and showed a maximum predicted value in 2021 of 5.2 x 10^4 KWh/million. After this, the graph began to decrease at a slow rate. It wouldn't follow the global trend of increased electricity consumption, but would still be possible. However, it is more likely that this was the result of another data bias within the overall set that was used. Slovakia ended up in a similar position to Singapore, with a population close to that of Denmark and Finland, but with varying prediction models. The model showed to be more linear than Singapore and included a constant slow decrease in its consumption values. The reported maximum value occurred in 2020 with a value of 1.8 x 10^4, the smallest found in our analysis so far. The reasoning behind this lower value is again most likely from the importance of technology for the average person, being less than that of other countries. Congo had the highest population in the group of small countries, but had the lowest reported values by a significant difference. The prediction model decreases fast and doesn't include large numbers to begin with. The maximum reported value was found in 2020 at 1900 KWh/million; it's possible this data set included an unfortunate abundance of error, but also important to remember that Congo would be the most underdeveloped country studied in this project. It is unclear to what extent this affects their data, but it is certainly an important factor to take into account. For the set of data including small countries, the squared exponential model showed the greatest accuracy for Singapore's data.

This prediction model supports the universal growth rates of electricity consumption and gives specific insight on what these values will look like for each country in the future. In order to take accurate precautionary measures to fight climate change, this data is absolutely necessary for countries to use to determine the best way to make the changes that need to be made to work towards a more sustainable environment. In the graphs that show decreasing values over time, it's important to consider where various errors could have possibly been made to analyze the data in the most accurate way possible. These errors and variations can be the result of a few different conditions such as errors in the actual data collected, or cultural differences among different countries that produce lower or higher rates of electricity based on what a common electronic user looks like. This information and predicted values from the project are vital to better understanding the direction that our environment is headed towards, what is concerning it, and, most importantly, what changes can be made to improve this. In some cases, it may be nearly impossible to make any large changes, simply due to the large population, but ideally countries in this position would use this data to recognize that issue and take steps towards using more renewable energy.