stat-471-final-proj

Ethan Kallett

12/19/2021

Executive Summary

Problem

In 1933, after Josef Stalin replaced the then-deceased Vladimir Lenin as premier of the Soviet Union, he looked around at the much wealthier nations to his West and noticed one key commonality – Great Britain, France and Germany all had well-developed industrial sectors. Stalin made the reallocation of resources from agriculture to industry perhaps the defining feature of his three decades of rule, even when it came at the cost of the lives of 35+ million of his own citizens.

This study sought to measure the impact of a "Big Push" – a massive reallocation of capital and labor from less productive (lower value-added) segments of the economy to higher value-added

Data

The majority of the data came from the World Bank Development Indicators. They are one of the world's most comprehensive datasets on economic development by country, with nearly 1500 indicators across 90 topics gathered annually at the country level since 1960. However, since there was a sparsity of data on some indicators, I chose indicators which contained a minimum threshold of data and imputed values for the remaining NA values.

Additionally, since I believed that data relating to armed conflict would be very relevant, I obtained this data from the Uppsala Conflict Data Program.

Analysis

The first part of my project consists of an exploratory data analysis that attempts to understand the relationship between the allocation of resources across the major sectors of an economy and the level of economic development. I do this through the use of many summary statistics and graphs that explore the distribution of the labor across agriculture, industry and GDP per capita, as well as create a Recurrent Neural Network (RNN) to model this relationship.

The main focus of this assignment attempts to look at the relationship between the allocation of these resources and economic growth. I first pull 27 different covariates from the World Bank Development Indicators and a 28th focused on the level of armed conflict in different countries. First, I conducted a panel regression (through the plm package) of real GDP per capita growth (over various time periods) on the covariates identified and the different numbers of lags. Following this, I create a Recurrent Neural Network that predicts GDP per capita growth using the 28 features and a time length of 5 years for the samples.

Conclusions

First, it is quite evident that a linear model is much more better suited for modeling this specific problem than a neural network.

Second, in terms of the interpretation of the models, the most important feature in determining economic growth would be capital accumulation per worker. This would suggest there was something correct about Joseph Stalin's cruel logic, and that, in order to grow, an economy may have to reallocate resources to build up capital in a "big push". Beyond this feature, other features such as natural resource rents, Gini index (inequality) and imports of goods and services are also very influential.

Introduction

Background Information

The original motivation for this topic was my senior thesis in economics. Attempts to assess how the "China shock" heterogeneously impacts Brazil through trade. The enormity of Chinese demand for commodities, in particular soy, petroleum and iron, coupled with the prohibitively competitive efficiency of Chinese manufacturing, has led to an enormous shift in the focus of the Brazilian economy from industry to agriculture and mining over the past 20 years. This is not problematic in itself, but one of the main concerns of this "deindustrialization" of the country is that it is a huge barrier to growth. I wanted to use some of the models and methods I had learned in class to see if this deindustrialization will have economic consequences for a country in the long run.

Everything I learned about Stalinism in ECON 271: Foundations of a Market Economy just further stimulated my interest. In that class, we researched and debated quite a bit about whether or not Stalinism and the 30+ million deaths associated with it was necessary to create the relatively modern industrial superpower capable of defeating Nazi Germany in World War Two.

In addition, a huge motivation for the initial exploratory data analysis of this assignment was based on the very broad generalization of agricultural economies (e.g. Chad, Kenya, Ghana, Afghanistan) as poor economies, industrial economies (e.g. China, Thailand, Malaysia, Mexico) as middle income economies and service economies (e.g. the US, Japan, the UK, Israel) as wealthy economies. This logic follows from the fact that the amount of value added per labor hour increases in that order, so therefore the wage rate of economies should increase in that order.

Analysis Goals

To put it succinctly, my principal analysis goals were to understand the relationship between the allocation of resources in an economy and its economic growth in the medium term. As a secondary area of interest, I wanted to understand how precisely the allocation of resources across various sectors of an economy (agriculture, industry and services) was able to determine it's level of economic development.

For the initial data exploration, in which I wanted to understand the strength of the relationship between the allocation of resources across different economic sectors and the country's economic development, I used the allocation of labor across the three broad sectors of an economy (agriculture, industry and services) as the main features to predict the response. The data on this allocation of labor was surprisingly abundant. The response that I aimed to predict was GDP per capita, measured in current 2021 US dollars at purchasing power parity (PPP).

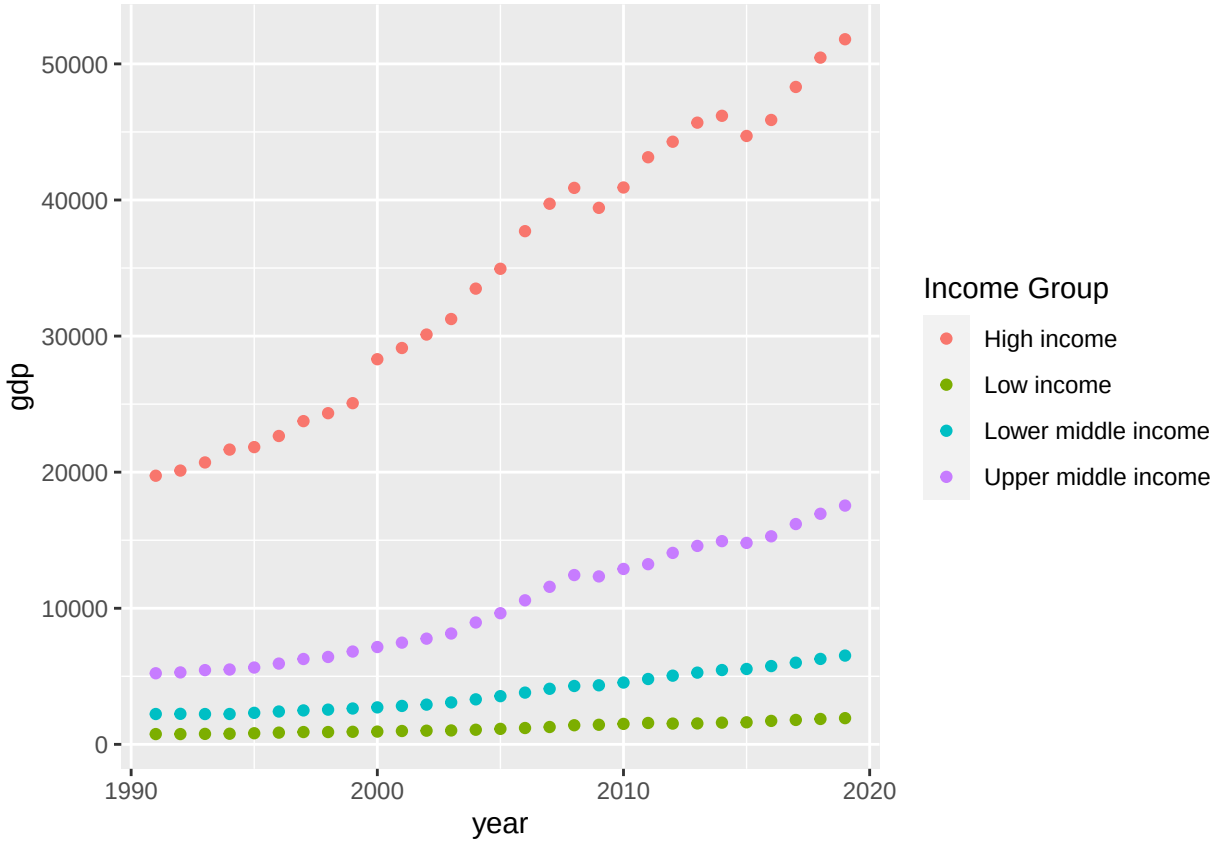
In the main focus of the assignment, I chose a selection of 28 different features from the World Bank Development Index that intended to either assess the allocation of resources within an economy or control for omitted variable bias. The outcome that I tried to predict was the percentage difference in an economy over the past several years. For the size/duration of this time frame, I tried a period of 1, 3, 5 and 10 years.

Success is evaluated in various aspects. The first metric is test MSE, i.e. how well is the model able to predict outcomes on data that it has not been trained on. This is a good approximation of how well the model was able to fit the underlying data generating process. The numerical, objective nature of this metric also makes it great to compare different models. The second criterion is how well the models can differentiate between the different importances of different features. If the end goal is to understand what allocation of resources

within an economy will help it grow the most in the medium term, it's crucial to understand what levers are available to pull, which direction to pull them, and which levers are the most powerful.

Significance

As countries across the world, particularly emerging market economies, strive to achieve sustainable growth to improve the quality of life of their citizens over the long run, studies like these are crucial. If a model can have a very low test MSE on the empirical data (meaning that it represents very well the underlying data generating process) and it is able to distinguish well between the features of which it is a part (indicating what are the important levers a "social planner" should pull), such a model would have enormous practical applications for governments across the world. Economic development leads to better economic lives for the majority of citizens of a country and aggregate gains in welfare can be huge.



RNN model

Data

Data Sources

The vast majority of the data were collected from the World Bank Development Indicators. The Development Indicators can be accessed here. The data contain over 1400 time series indicators for 217 different economies and over 40 supranational groupings, with several indicators going back over 50 years. Because the task of collecting such diffuse data is impossible, for one organization to manage, the World Bank sources a lot of the data from official sources such as national statistics organizations, United Nations agencies, academia and beyond. However, the World Bank Group does conduct several surveys and research projects of their own to add to the data set.

However, since I knew that armed conflict would have a huge influence on the macroeconomic growth of an observation country, I decided to join more data sourced from the Uppsala Conflict Data Program (UCDP). The UCDP collects data on armed conflict since 1975. They define an "event" as any instance of fatal organized violence. And to ensure truly global representation in their data set, they utilize a keyword search through the Dow Jones Factiva Indicator, which annually returns around 50 thousand unique events which are then reviewed by human evaluators, resulting in around 10-12 new events coded each year.

Data Cleaning

To clean the data, the World Bank Development Indicators were first filtered to exclude non-country observations in the data. For the indicators, since many only have observations for a small select set of (usually wealthy and developed) countries, or only a short temporal range for the time series, or the many missing values in between where data was not collected for whatever reason, I first filtered for indicators that had adequate data for the model. This reduced the number of indicators from 1443 to 497. Then, from this list, I manually chose features which lacked high levels of multicollinearity, were expressed as a percentage rather than in absolute numbers (to avoid selection bias), and were very relevant to the project. Lastly, I removed countries without a certain availability of data and restricted the temporal range to 1984-2020, since the data became less abundant earlier in the set. In order to join the World Bank Development Indicators on the UCDP conflict data, some country names had to be manually adjusted.

Following this selection of data, values were manually imputed to make sure there were no NaN values in the data set. This imputation was done by iterating through every value in the panel data set, and, if it was null, calculating a "yearly" average and an "observation" average and then averaging both. The yearly average consisted of the average value for that indicator in a given year among all countries in that income group ("High income", "Upper middle income", "Lower middle income", or "Low income"). The observation average was the mean value for a given indicator for a given country across all years in the data set.

To wrangle the data into a format suitable for the time-series fixed effects regression in Model 1: Linear Methods (Linear Models for Panel Data), the data was first transformed into a `panel data dataframe` from the `plm` package, and then subsequently transformed into a `pseries` object from the package with the `Year` and `country_id` as indices.

For the Recurrent Neural Network models, the data had to be wrangled into a three-dimensional format, where the first dimension represented observations, the second dimension represented time, and the third dimension represented the features of the model. This was done iteratively through for loops, parsing every value in the two-dimensional data and mapping it to its correct position in the three-dimensional array. Following this, each country time series was split into 5 year long strips with a stride of one (e.g. one strip for 1999-2003, one for 2000-2004, one for 2001-2005, etc.) and then aggregated into another 3-dimensional array.

Data Description

Table 1: Summary Statistics of Data with Value Imputation

Summary Statistic	Value
Number of Indicators	28
Number of Income Groups	4
Number of Years Considered	37
Number of Countries	200
Number of Rows	5600
Number of Columns	40

Summary statistics of the data can be seen from Table @ref(tab:data-description-1). There are a total of 5600 observations and 40 columns in the principal data set used, after value imputation. These observations

can be divided into time series data (from 1984 to 2020) for 200 countries, grouped into 4 different income groups, across 28 different indicators.

Table 2: List of Indicators Used in Analysis

Series Code	Indicator Name
BG.GSR.NFSV.GD.ZS	Trade in services (% of GDP)
DT.ODA.ODAT.PC.ZS	Net ODA received per capita (current US$)
FP.CPI.TOTL.ZG	Inflation, consumer prices (annual %)
FS.AST.PRVT.GD.ZS	Domestic credit to private sector (% of GDP)
NE.EXP.GNFS.ZS	Exports of goods and services (% of GDP)
NE.GDI.TOTL.ZS	Gross capital formation (% of GDP)
NE.IMP.GNFS.ZS	Imports of goods and services (% of GDP)
NV.AGR.EMPL.KD	Agriculture, forestry, and fishing, value added per worker (constant 2015 US$)
NV.AGR.TOTL.ZS	Agriculture, forestry, and fishing, value added (% of GDP)
NV.IND.TOTL.ZS	Industry (including construction), value added (% of GDP)
NV.MNF.TECH.ZS.UN	Medium and high-tech manufacturing value added (% manufacturing value added)
NV.SRV.TOTL.ZS	Services, value added (% of GDP)
NY.GDP.PCAP.CD	GDP per capita (current US$)
NY.GDP.TOTL.RT.ZS	Total natural resources rents (% of GDP)
SE.PRM.ENRL.FE.ZS	Primary education, pupils (% female)
SE.PRM.ENRR	School enrollment, primary (% gross)
SE.SEC.ENRL.GC.FE.ZS	Secondary education, general pupils (% female)
SE.TER.ENRR	School enrollment, tertiary (% gross)
SI.POV.GINI	Gini index (World Bank estimate)
SI.POV.UMIC	Poverty headcount ratio at $5.50 a day (2011 PPP) (% of population)
SL.AGR.EMPL.ZS	Employment in agriculture (% of total employment) (modeled ILO estimate)
SL.IND.EMPL.ZS	Employment in industry (% of total employment) (modeled ILO estimate)
SL.SRV.EMPL.ZS	Employment in services (% of total employment) (modeled ILO estimate)
SL.TLF.CACT.FM.ZS	Ratio of female to male labor force participation rate (%) (modeled ILO estimate)
SL.TLF.CACT.ZS	Labor force participation rate, total (% of total population ages 15+) (modeled ILO estimate)
SP.POP.1564.TO.ZS	Population ages 15-64 (% of total population)
SP.URB.TOTL.IN.ZS	Urban population (% of total population)
Conflict	Sqrt of conflict deaths given country and year

The 28 different features considered in the data set are listed in Table @ref(tab:feature-description). However, these features should be considered only 27, since the variable `NY.GDP.PCAP.CD` or `GDP per capita (current US$)` was used to construct the response variables. In Part II: Economic Structure and Economic Growth, the response variable was continuous, and was defined as a the percentage change in GDP of a given observation (country) over a certain period of time. This could be calculated as (GDP in year n - GDP in year 0)/GDP in year 0. The time periods of 1 year, 3 years, 5 years and 10 years were all tried on the data set.

For Part I: Economic Structure and Economic Development Level, the continuous feature `NY.GDP.PCAP.CD` was used directly as the response variable.

Data Allocation

To get good estimates for Model 1: Linear Methods (Linear Models for Panel Data), in this section, the values were bootstrapped and models were calculated from train/test splits multiple times. In each train/test split, 80% of the values were used for training. 20% of the values were used for testing, and test MSE, number of significant features and other metrics were calculated on this test data and aggregated.

In Model 2: Recurrent Neural Network, no bootstrapping was done, but 80% of the data was held in reserve as test data to assess model fit. During training, the neural network held 25% of the training data in reserve as validation data in order to tune the parameters.

Modeling

Part I: Economic Structure and Economic Development Level

This first portion of data looked to see at the relationship between the structure of an economy at the most aggregate, generalized level and its economic development level, measured in real (purchasing power parity) GDP per capita. It consists of Exploratory Data Analysis (Data Exploration), which looks at summary statistics to understand this relationship and Model 1: Recurrent Neural Network, which builds a RNN model to predict GDP (PPP) per capita.

Exploratory Data Analysis (Data Exploration)

From Figure @ref(fig:plot-grid-p1-p3), it is very visible that there is a clear correlation between the structure of an economy and its "income level", as classified by the World Bank. As a country gets richer, resources are very clearly taken from agriculture and reallocated to industry and services, with services being an even larger recipient than industry. Furthermore, we see that not only has this trend remains quite constant over time, but that the relative gaps between income groups in various industries also seems quite constant over the past three decades.

Figure @ref(fig:tens-boxplot) further affirms the strength of the connection between economic mix and GDP per capita. Not only are the averages so disparate across income groups, but that distributions are also very tight. Outliers in this data set are very few, and even if you look at the outliers with the highest share of agriculture in the "High income group", they all have a lower share of the economy dedicated to agriculture than the median "Upper middle income" country.

Figure @ref(fig:share-over-time-change-plot) shows us the importance of controlling for time in any model. The relative share of the economy dedicated to services has exploded across all income groups. Likewise, the share dedicated to agriculture has fallen through time in all income groups, as well, though the decline was perhaps more precipitous in poorer countries than in richer countries.

\begin{table}[H]

\caption{1991-2019 % Change in Share of Economy Dedicated to a Sector by Income Group}

Income Group	% Change Agriculture	% Change Industry	% Change Services
High income	-0.619	-0.217	0.187
Upper middle income	-0.409	-0.080	0.303
Lower middle income	-0.310	0.179	0.383
Low income	-0.186	0.191	0.608

\end{table}

Table @ref(tab:pct-chg-groups-time-tibble) again shows that the economies are not static. Subject to market forces, they have been evolving over time in an enormous way (think of how much time and resources it takes to train a smallholder farmer to be a teacher, bus driver, programmer, banker, or another professional in the service industry). Table @ref(tab:pct-chg-groups-time-tibble) shows that changes to the percentage of the economy engaged in industry have been modest, with rich countries experiencing a slight deindustrialization and poor countries experiencing slight gains in industry employment However, the data also show that the decrease in the percentage of the economy engaged in agriculture is very positively correlated with income. Conversely, again, although that the absolute gains in the service industry may be modest in poor countries, in relative terms, the poorest countries were the ones that have experienced the most dramatic shifts to a service-based economy.

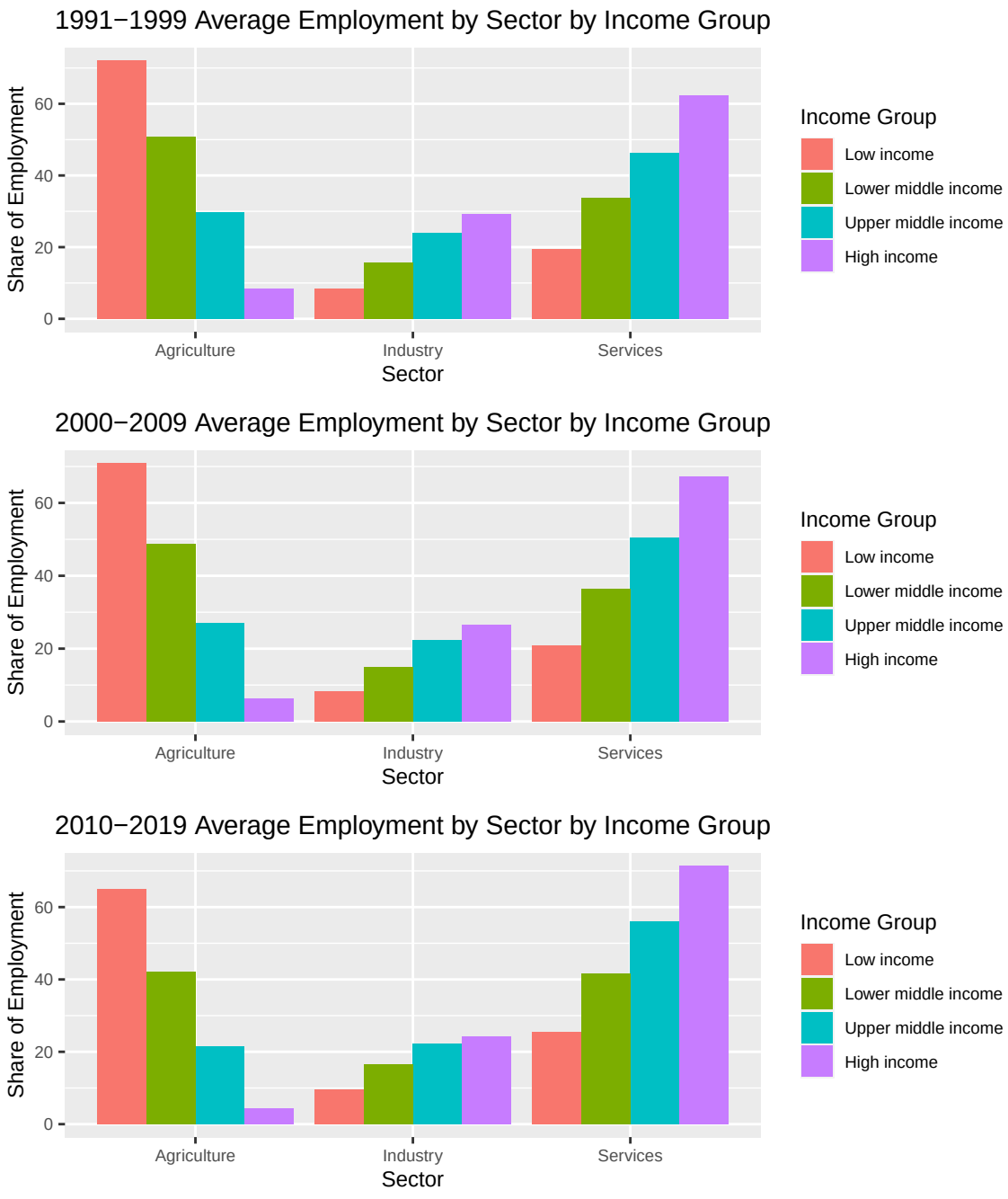


Figure 1: Average Employment by Sector by Income Group Over Time

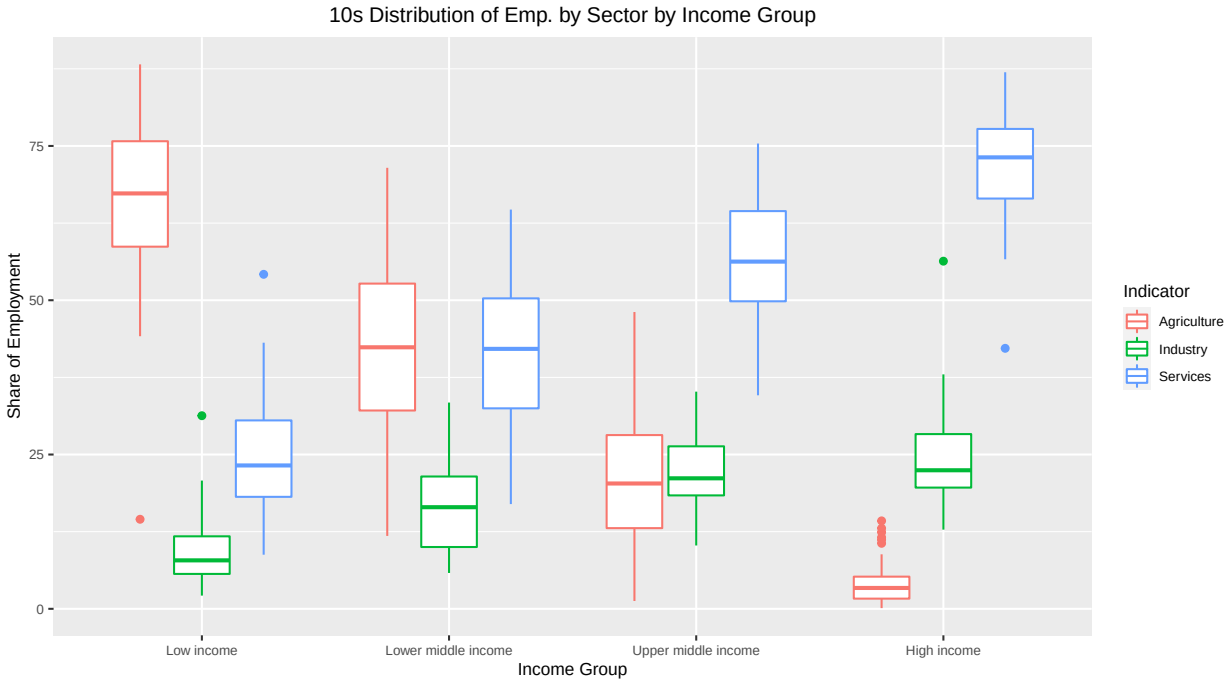


Figure 2: Distribution of Employment by Sector by Income Group 2010-2019

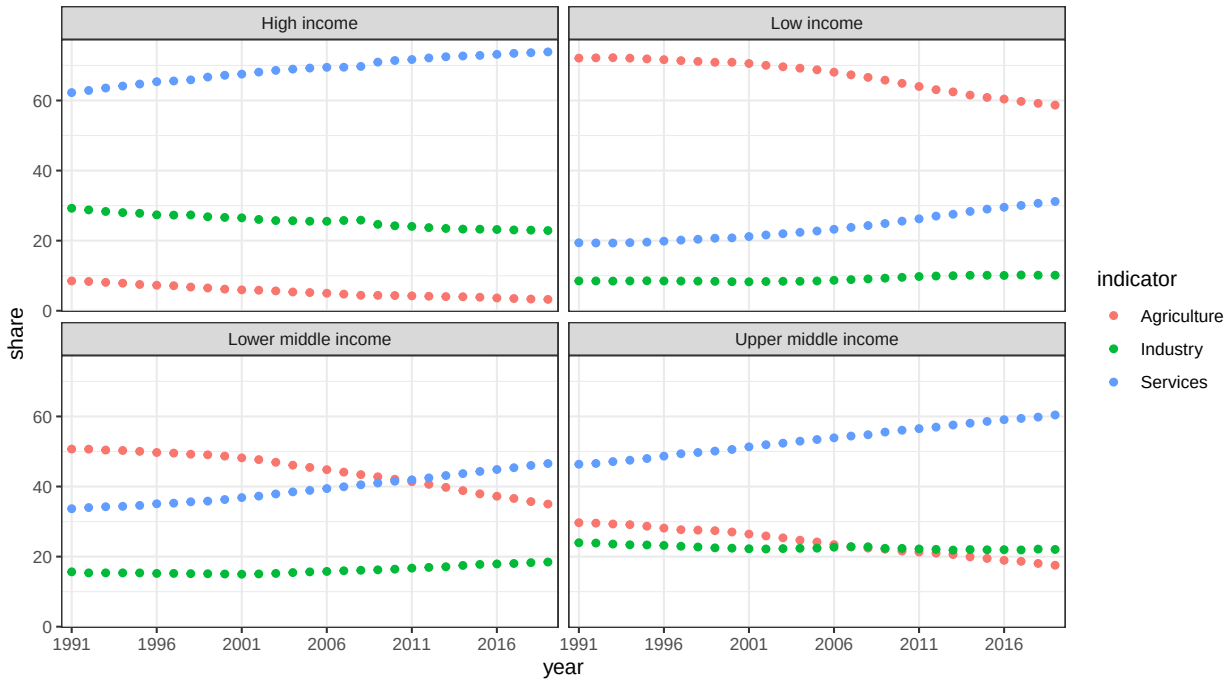


Figure 3: Sector Share Change by Income Group 1991-2019

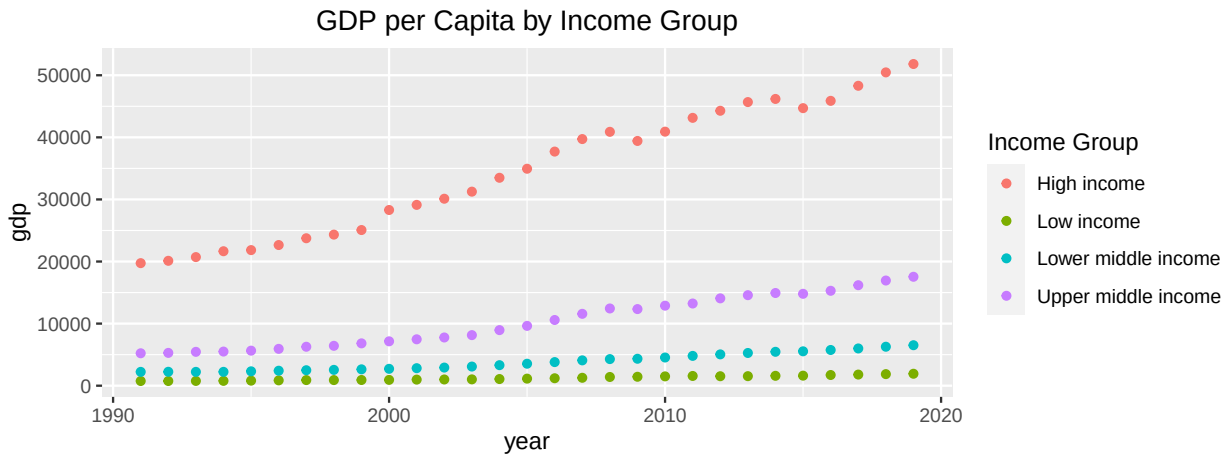


Figure 4: 1991-2019 GDP per Capita by Income Group

Figure @ref(fig:gdpcap-time-plot) shows the limitations of relying on static, 2020 classifications of country by income group. We can see from the data that rich countries have grown enormously while poor countries have remained trapped in low levels of growth. This begs the question of whether or not these poor economies have really not experienced growth, or whether there have just been a number of poor countries that have "graduated" to the status of a richer country. And, looking at the above charts, if this phenomenon exists, would the poor countries that emphasize services over agriculture be the ones that graduate to rich countries, thus giving us the massive differences across income groups that we see? These questions and many more motivate Part II: Economic Structure and Economic Growth, where I, quite ambitiously, attempt to assess how resource allocation affects economic growth over the short-to-medium term.

Model 1: Recurrent Neural Network

Given that the relationship between all of these inputs and economic growth is highly nonlinear (if we could so easily rely on coefficients to determine `X` feature's contribution to economic growth, development economists would have solved the growth trap of least developed countries already!) I decided to build a neural network to model the relationship between resource distribution and economic outcomes. And, since the data is truly panel data - it is absolutely impossible to consider any response value without understanding the identity of the country that it pertains to as well as the `n` values that came before it temporally - I chose to create and train a Recurrent Neural Network.

Considering the strength of the relationship between economic structure and GDP per capita found in Exploratory Data Analysis (Data Exploration), I decided to construct a very simple model that only used four input features to determine the output (real GDP per capita.) These four input features were: `employment in agriculture (% of total)`, `employment in industry (% of total)`, `employment in services (% of total)` and the year of the observation.

After wrangling the data into three dimensions, splicing it into sections of length 5 and then wrangling those into another three-dimensional array, I built a simple three layer recurrent neural network. It used ReLU activation functions at each step and, since the output was continuous, the output node was singular. The input shape was (5, 4) because it took 4 features (agriculture, industry, manufacturing, services and year) over 5 distinct years.

```
## Model: "sequential"
## _________________________________________________________________________
## Layer (type)                 Output Shape              Param #
## =========================================================================
## dense_1 (Dense)              (None, 5, 4)              20
```

```
## _________________________________________________________________________
## RNN_1 (SimpleRNN)                   (None, 5, 64)                   4416
## _________________________________________________________________________
## RNN_2 (SimpleRNN)                   (None, 5, 32)                   3104
## _________________________________________________________________________
## RNN_3 (SimpleRNN)                   (None, 5, 32)                   2080
## _________________________________________________________________________
## dense (Dense)                       (None, 5, 1)                    33
## =========================================================================
## Total params: 9,653
## Trainable params: 9,653
## Non-trainable params: 0
## _________________________________________________________________________
```

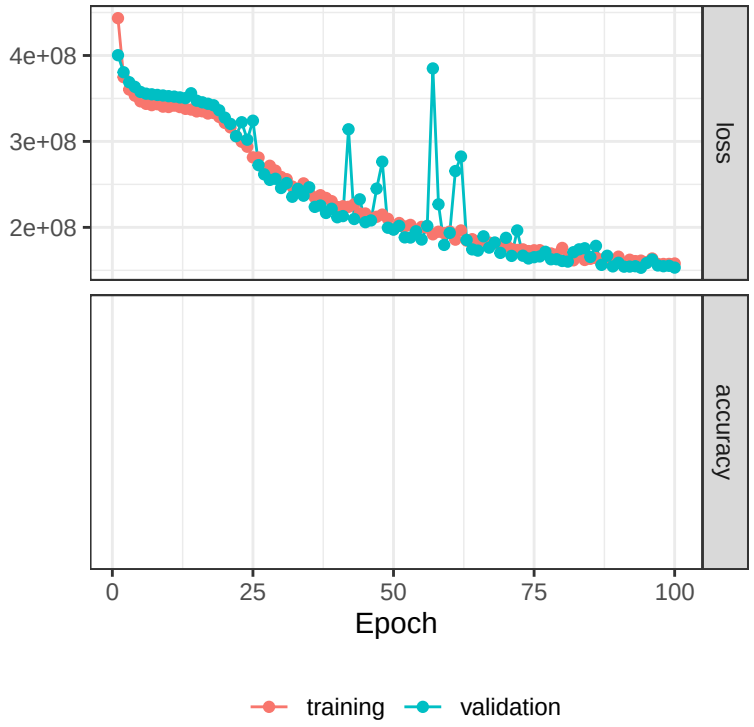


Figure 5: Training History of RNN Model from Part 1

Using the handy imported `plot_model_history` function from class, we can see that the model learns quite a bit over the 100 epochs it's run. The predictions of the RNN model actually get quite good. Unfortunately, since the problem is not a classification one, there is no data to plot for the `accuracy` portion of this graph.

Exactly how good are the predictions of the model?

Table 3: Accuracy of RNN Model on Test Data

Absolute Test Error
5900.798

The model, with very minimal tuning, Table @ref(tab:model1-abs-acc) shows that this model can predict real

GDP per capita within around $7500 USD. That's not perfect, but according to 2021 figures, that's substantially *less* than the difference between the US and Canada or the difference between Australia and the UK It's, perhaps as expected, very directionally accurate with just the above four features.

Again, this model does not have very much tuning. I decided to save time to have the vast majority of tuning be for my models for Part II: Economic Structure and Economic Growth, which is more involved and more relevant to what I wanted to study with this project.

Part II: Economic Structure and Economic Growth

This section is fundamentally different from the above Part I: Economic Structure and Economic Development Level in two ways. First, this section uses vastly many more features to try to generate a more complex and predictive model (28 in total, as opposed to Part I's 4.) Second, while the Part I looked at the relationship between the distribution of employment in an economy and its contemporaneous GDP per capita, this section aims to understand how the allocation of resources (interpreted more broadly than just labor this time) affect economic *growth* in the short-to-medium term.

Again, for this section, I manually created four different response variables. Each represented the percentage change in GDP from year 0 to year n, with the time periods being 1 year, 3 years, 5 years and 10 years.

Model 1: Linear Methods (Linear Models for Panel Data)

Since ordinary least squares is no good on panel data, I used the `plm` panel data estimators function from the `plm` package to create these linear models. Through the incorporation of time- and entity-invariant fixed effects, such a model can provide much more reliable coefficient estimates for panel data.

To tune these models, I gave each formulation an extremely high degree of personalization; I tried unique combinations of all the following features:

First, I tried the different time periods to define "economic growth" as the response variable. Second, I tried models with one to eight lags in them. Each lag is an interaction term of a value from a previous year which takes a coefficient (that may or may not be significant). Lastly, I tried three models of lags: models that contained lags of GDP per capita from previous years, lags that contained annual growth rates from previous years, and models that contained both kinds of lags.

Table 4: Features of Different Linear Model Specifications Tried

Personalizaation Feature	Types
Response Variable	1, 3, 5, 10 year growth period
Lag Number	1 to 8 lags
Lag Type	GDP per capita, annual growth or both

Because of the high variance in test MSE, the number of significant features and significant lags across all different linear model specifications, I bootstrapped the calculations. As you can see in Figure @ref(fig:plm-tuning-results), there still is a very high degree of variance in the results. However, there are a few conclusions that I drew from playing around the different tuning parameters. The first is that past annual growth rates are a much better predictor of future growth than past GDP per capita, as evidenced by the red lines in Figure @ref(fig:plm-tuning-results) hanging much lower than the green ones. The second clear conclusion is that the lags are a much better predictor of future growth than most of the other features in the model. The lags are almost always significant, while most features almost always are not. Lastly, since in absolute test MSE as the growth period extended only increased, I standardized the test MSE values with respect to the growth period that they have as a response variable. Once standardized, it interestingly enough doesn't appear that a longer growth period tends to lead to worse predictive power.

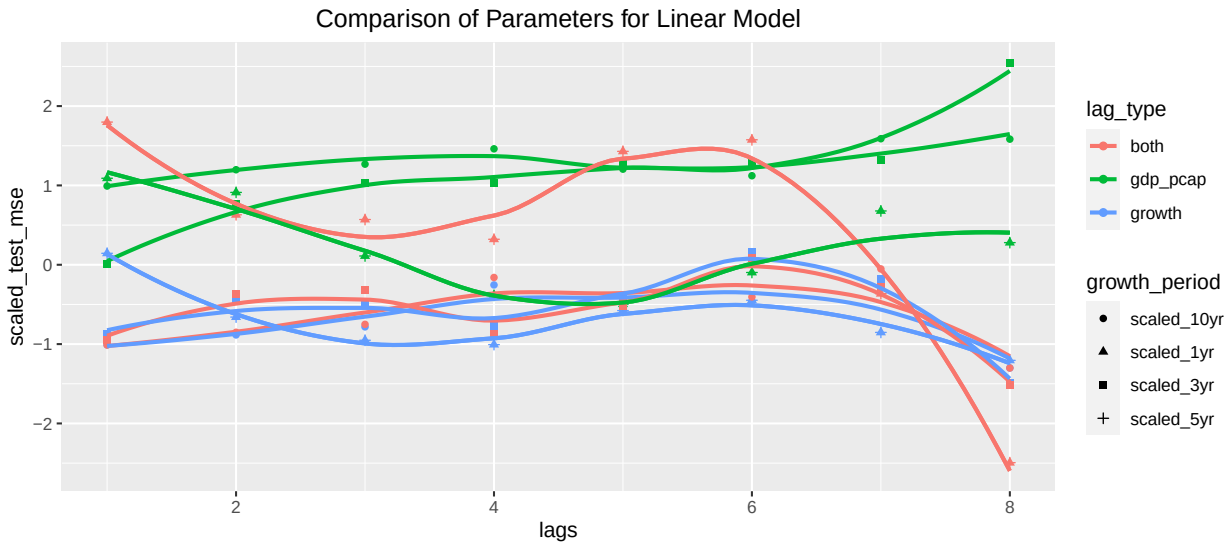


Figure 6: Comparison of the Results of Different Parameters for the Linear Model

Table 5: Features of Different Linear Model Specifications Tried

Indicator Code	Indicator Name
NE.GDI.TOTL.ZS	Gross capital formation (% of GDP)
NE.IMP.GNFS.ZS	Imports of goods and services (% of GDP)
SP.URB.TOTL.IN.ZS	Urban population (% of total population)
NY.GDP.TOTL.RT.ZS	Total natural resources rents (% of GDP)
SI.POV.GINI	Gini index (World Bank estimate)
NE.EXP.GNFS.ZS	Exports of goods and services (% of GDP)
NV.AGR.TOTL.ZS	Agriculture, forestry, and fishing, value added (% of GDP)
FP.CPI.TOTL.ZG	Inflation, consumer prices (annual %)
FS.AST.PRVT.GD.ZS	Domestic credit to private sector (% of GDP)
SL.TLF.CACT.FM.ZS	Ratio of female to male labor force participation rate (%) (modeled ILO estimate)
NV.SRV.TOTL.ZS	Services, value added (% of GDP)
NV.MNF.TECH.ZS.UN	Medium and high-tech manufacturing value added (% manufacturing value added)
SP.POP.1564.TO.ZS	Population ages 15-64 (% of total population)
SL.TLF.CACT.ZS	Labor force participation rate, total (% of total population ages 15+) (modeled ILO estimate)
BG.GSR.NFSV.GD.ZS	Trade in services (% of GDP)
NV.IND.TOTL.ZS	Industry (including construction), value added (% of GDP)
SE.SEC.ENRL.GC.FE.ZS	Secondary education, general pupils (% female)
SI.POV.UMIC	Poverty headcount ratio at $5.50 a day (2011 PPP) (% of population)
SE.PRM.ENRR	School enrollment, primary (% gross)
SE.PRM.ENRL.FE.ZS	Primary education, pupils (% female)
SE.TER.ENRR	School enrollment, tertiary (% gross)
DT.ODA.ODAT.PC.ZS	Net ODA received per capita (current US$)
NV.AGR.EMPL.KD	Agriculture, forestry, and fishing, value added per worker (constant 2015 US$)
SL.AGR.EMPL.ZS	Employment in agriculture (% of total employment) (modeled ILO estimate)
SL.SRV.EMPL.ZS	Employment in services (% of total employment) (modeled ILO estimate)
SL.IND.EMPL.ZS	Employment in industry (% of total employment) (modeled ILO estimate)

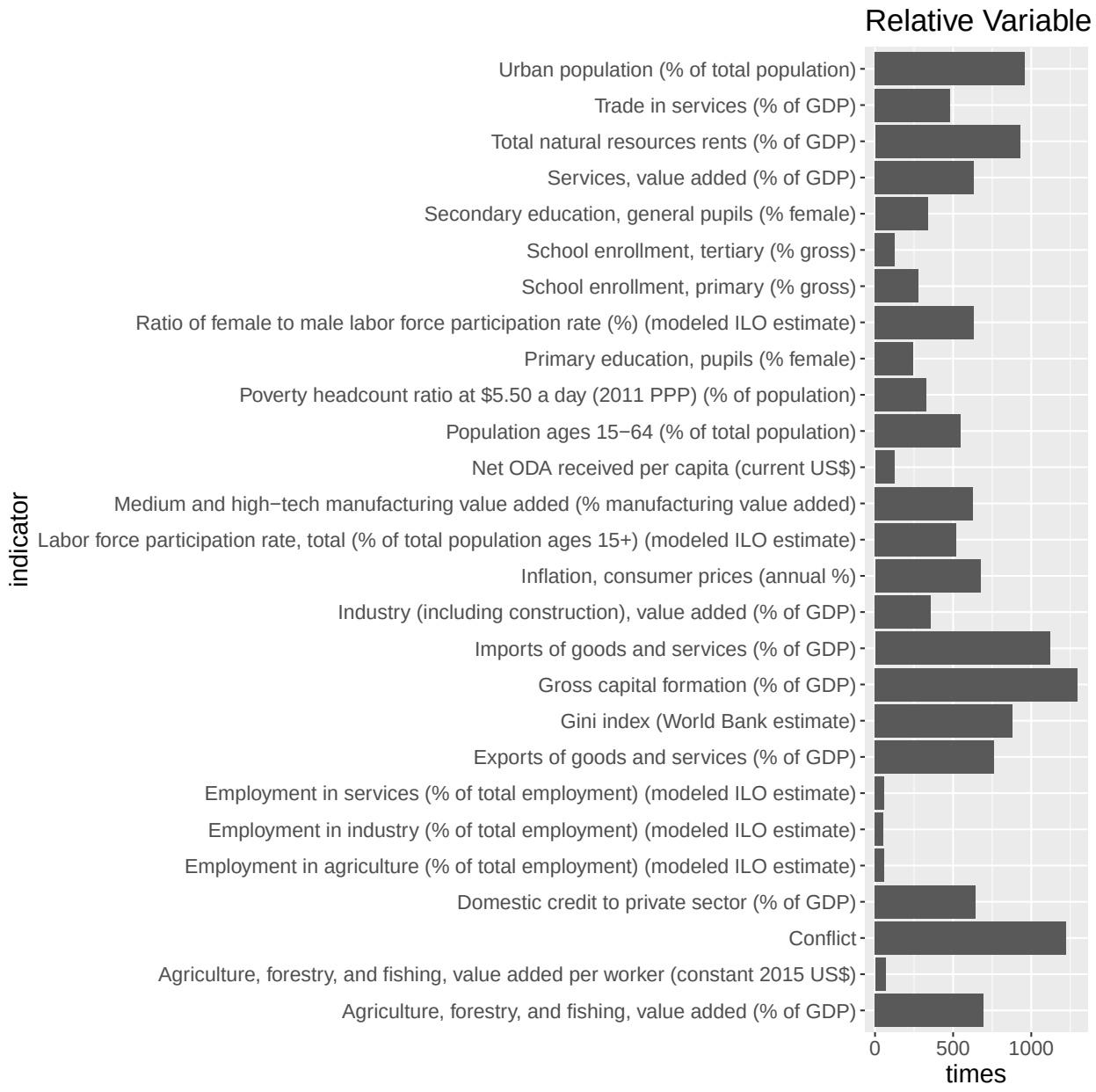


Figure 7: Relative Variable Importances Averaged over Linear Models

Table @ref(tab:plm-var-imp-tibble) and @ref(fig:plm_var_importance_graph) attempt to assess feature importance by counting the number of times that each feature was significant in a regression in the bootstrap sequence in Model 1: Linear Methods (Linear Models for Panel Data). From the findings presented in Table @ref(tab:plm-var-imp-tibble) and @ref(fig:plm_var_importance_graph), perhaps one of the most interesting conclusion is that the variables that were used to build the RNN model above, which aimed to predict economic development level, were quite literally the least important features in the data set for predicting *economic growth*. However, the models show that there is a set of other variables which are quite important when it comes to predicting GDP growth. Gross capital formation makes a lot of sense, since more capital accumulation will inherently allow for more output. This was exactly the logic behind Stalinism: extract surplus from the peasants in order to increase the capital per worker ratio. `Natural resource rents` is an interesting important variable, since traditional economics would argue that resource rents prevent a country from innovating a dynamic economy away from those resource endowments (The "resource curse"). Or, since

this chart only describes variable importance, it's very possible that the conclusion to be made from here is that the majority of resource rich countries (e.g. Saudi Arabia, United Arab Emirates, Norway, Canada) manage their mineral wealth wisely.

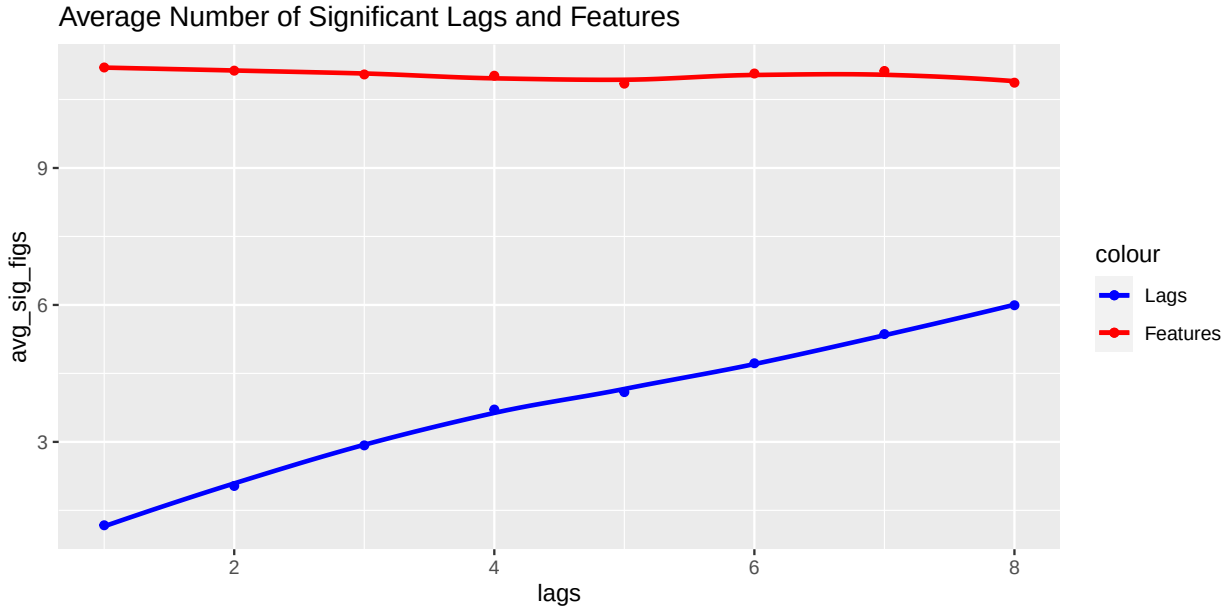


Figure 8: Number of Significant Lags and Features as Function of Lags

Figure @ref(tab:sig-lags-features) shows that, as we add more lags to the model, they are almost always significant. Given this phenomenon, though, it's quite interesting that the number of significant features only decreases incredibly slightly as the the number of lags increases. The number of significant features stays pretty consistent around 10, even when plots like @ref(fig:plm_var_importance_graph) show that those approximately 10 features must be very different in different bootstrapped models.

Model 2: Recurrent Neural Network

For the recurrent neural network portion of this assignment, I decided to construct neural networks that tune along two different parameters. The first parameter tuned for is model complexity, defined as the number of layers in the neural network and the number of nodes in each layer. For this parameter, since there are infinite many values it could take, I chose three different models, one with one layer, one with two layers and one with three. The second parameter I tuned for is the temporal length of each "strip" of data. In other words, when the RNN model is given a sequence of values to determine the predicted outcome, how many years of previous data are represented in this strip? For this component, I chose the values 3, 4 and 5 years of previous data. I tried every combination of these 2 different parameters for a total of 9 different RNN models created.

Table 6: Comparison of RNN Model test MSE against Specifications

Model Complexity	3 Previous Values	4 Previous Values	5 Previous Values
One layer	0.382	0.445	4.856
Two layers	0.219	0.240	0.180
Three layers	0.230	0.236	0.249

After training each model, each model was then run on a 20% share of the data held out as validation data.

Interestingly enough, although the simplest model which looks at 5 years of previous data had the worst test MSE, Table @ref(tab:rnn-param-comparison) shows the best parameter selection was that of the two-layer model that looks at 5 years of previous data. The reason that the one-layer 5-previous-year model is so bad compared to the other models is that, with more years of data, the model needs more parameters in order to be able to better fit the underlying trends.

Table 7: Most Importance Features according to RNN Model

Deviance	Feature Code	Feature Name
1.826	NY.GDP.PCAP.CD	GDP per capita (current US$)
0.329	NV.AGR.EMPL.KD	Agriculture, forestry, and fishing, value added per worker (constant 2015 US$)
0.010	DT.ODA.ODAT.PC.ZS	Net ODA received per capita (current US$)
0.007	FS.AST.PRVT.GD.ZS	Domestic credit to private sector (% of GDP)
0.004	FP.CPI.TOTL.ZG	Inflation, consumer prices (annual %)
0.003	BG.GSR.NFSV.GD.ZS	Trade in services (% of GDP)
0.002	NY.GDP.TOTL.RT.ZS	Total natural resources rents (% of GDP)
0.001	SL.TLF.CACT.FM.ZS	Ratio of female to male labor force participation rate (%) (modeled ILO estimate)
0.001	NE.IMP.GNFS.ZS	Imports of goods and services (% of GDP)
0.001	SE.TER.ENRR	School enrollment, tertiary (% gross)
0.001	SI.POV.UMIC	Poverty headcount ratio at $5.50 a day (2011 PPP) (% of population)
0.001	NE.EXP.GNFS.ZS	Exports of goods and services (% of GDP)
0.001	SE.PRM.ENRR	School enrollment, primary (% gross)
0.001	NV.SRV.TOTL.ZS	Services, value added (% of GDP)
0.001	SL.TLF.CACT.ZS	Labor force participation rate, total (% of total population ages 15+) (modeled II
0.000	SP.URB.TOTL.IN.ZS	Urban population (% of total population)
0.000	Conflict	NA
0.000	SP.POP.1564.TO.ZS	Population ages 15-64 (% of total population)
0.000	SL.AGR.EMPL.ZS	Employment in agriculture (% of total employment) (modeled ILO estimate)
0.000	SE.SEC.ENRL.GC.FE.ZS	Secondary education, general pupils (% female)
0.000	SL.IND.EMPL.ZS	Employment in industry (% of total employment) (modeled ILO estimate)
0.000	SI.POV.GINI	Gini index (World Bank estimate)
0.000	NE.GDI.TOTL.ZS	Gross capital formation (% of GDP)
0.000	NV.IND.TOTL.ZS	Industry (including construction), value added (% of GDP)
0.000	SE.PRM.ENRL.FE.ZS	Primary education, pupils (% female)
0.000	NV.MNF.TECH.ZS.UN	Medium and high-tech manufacturing value added (% manufacturing value added
0.000	SL.SRV.EMPL.ZS	Employment in services (% of total employment) (modeled ILO estimate)
0.000	NV.AGR.TOTL.ZS	Agriculture, forestry, and fishing, value added (% of GDP)

To determine feature importance, I scrambled the values of each f and observed how much it reduced the accuracy of the predictions. Perhaps as I should have expected, the most important feature by far according to Table @ref(tab:rnn-feature-scramble) is GDP per capita. This is interesting, because in the linear model, using GDP per capita as past lags was not nearly as powerful as using annual GDP per capita growth rates as the lags.

Even more interestingly, the second most important variable in this model is `Agriculture, forestry, and fishing, value added per worker`. This is also quite interesting and unexpected, since we've observed that richer economies transition out of agriculture and into services. However, this might make sense, since in rich economies that have transitioned out of agriculture, only the most productive workers will remain in the industry. This can easily be seen by how farmers in a country like the US have heavy machinery, fertilizers and pesticides which will increase their yields to something that smallholder farmers in low income countries with just wooden tools could only dream of.

However, the results from this model are quite alarming, since the majority of the coefficients are very close to zero. In fact, there are even a few features whose values, when scrambled, slightly *improve* the test MSE. This is ridiculous, and I conclude this is a function of there just not being enough data to create a neural network model for this problems in this domain.

Conclusions

Method Comparison

When comparing the recurrent neural network model with the linear model, it's clear that, in terms of absolute reduction in test MSE, the neural network appears to do much better. However, models are only as good as they are useful, and assessing feature importance through scrambling for the neural network revealed that the model may as well have been guessing. It didn't offer very much insight into which variables were the truly important ones driving economic growth. Most features had importances near zero in the model, and some for some features, scrambling resulted in an even *worse* prediction than on the unscrambled data.

In contrast, by counting how often a coefficient on the the linear model was significant, we were able to learn a lot more about the influence of each of the constituent variables in the model. The linear model better satisfied many of our preconceived notions that sparsity does *not* hold, and that each of these decorrelated features should have some sort of unique contribution, however large, to the economic growth and development of an economy. This, however, comes with the caveat that we assume that the relationships between these variables and the response is anything but linear. Moreover, we assume that all of these features have unique and complex interactions, which would imply that a neural network would potentially be a better model if we were given the data.

In conclusion, given that data in this problem was so limited, I believe that linear models were a much better fit for this problem than neural networks ever could be.

Takeaways

I think that perhaps one of the most powerful components of this project was Part I: Economic Structure and Economic Development Level. The simplicity of the model I used may be frowned upon by economists who say that a lack of variables would mean that my findings would be worthless because of hidden variable bias. This is without a doubt a problem with such simplicity, but on the other hand, such simplicity allows for such great model interpretability that you may learn a lot more from analyzing a lot less.

In terms of interpreting the coefficients from the linear model, the most shocking conclusion is that perhaps there was a grain of truth in the horrors committed by dictator Joseph Stalin. It shows that the most important variable by far is capital accumulation per worker. This would lend credence to the theory that a government, in order to grow it's economy, should intervene to reduce consumption and thus increase the capital stock, which can then be utilized to create more capital stock.

Another fascinating observation is that the three features I used to construct a relatively accurate recurrent neural network in Part I: Economic Structure and Economic Development Level, the share of employment in agriculture, industry and services, though great at predicting current GDP per capita, are declared by the linear model in Model 1: Linear Methods (Linear Models for Panel Data) to be the least useful features in the data set.

Limitations

Perhaps the greatest limitation of this study was just a lack of data. Before value imputation, the data was really spare and prevented me from choosing a lot of the indicators that I wanted. Moreover, I can only reasonably extrapolate a certain amount from trends in the data, so the amount of value imputation that I could do was limited. There are only so many countries in the world and years on each country for which we have data. Furthermore, through value imputation, I'm very concerned that, since I'm using each income group as the group from which to calculate the averages, I am systematically making poor countries look

more poor and rich countries look more rich. This would be an insidious kind of bias that really hinders the model.

This lack of data truly hindered the accuracy of any model on this very complex phenomenon, whether it was linear, a neural network or some other form of model. This lack of data is something that economists have been struggling with forever, and is potentially why most economists are too shy to create a "theory of everything" model that tries to predict how much an economy will grow given a set of inputs. Secondarily, in the development of the neural network, the choice of the parameters is very much more of an art than a science. It's very possible that if I added/substracted a few layers, added in a few more nodes, increased dropout et cetera, I would have developed a model that is substantially better than the one I currently have. However, since trying an infinite number of parameters is computationally infeasible, we will never know the answer to these hypothetical questions.

Follow-ups

To follow up on this, I think that I would try some tree methods on this problem, since they were not tried. Tree methods don't need nearly as much data as neural networks, so they wouldn't have as much of a problem with the scarcity of data. Moreover, tree methods can allow for much more complex interactions among the features of a data set than linear methods, so there may be a lot of merit to their value in this domain.

I think to solve the main problem, more data has to be gathered. There are only so many different countries in the world, but fortunately not every country is of the same size. Many countries that numerous regions inside of them that are as large as and economically (in)dependent as any other country. For that reason, I think it would be a phenomenal next step to repeat this methodology, but using panel data disaggregated to the level of the subnational region.

Lastly, I want to say that economists have very traditionally relied on their econometric methods, attempting to assess the significance of variables and determine causality, and have not necessarily dipped into the very valuable vault of methods that data science has to offer. With more data, suddenly the simplicity of a linear difference-in-differences methodology seems very risky, because it's hard to expect such a relationship to be simply linear.