The Correlation between Poverty and various Internet infrastructure indicators

Contents

Abstract

Through this project, we examing the various econmic, social, and technological indicators in detail using American Census Data and measure it with respect to poverty rates. Using state by state data from 2017, we built several exploratory models before examing the regression coefficients for their correlary information. We repeated this process using state expenditures data in addition to the technological indicators from above. Using generalized linear models and random forest regressors, we were able to build descriptive models that all had R^2 values of .80, meaning the chosen indicators explained 80% of the variance in poverty. Additionally, all models had a root mean square error of 1-2% of the actual poverty rate. We found that technological

indicators like a STEM education, home computer access, broadband internet, and smarthphones were most highly correlated with poverty rates. Computer access

Questions

What economic indicators (race, occupation, community poverty rate) are most strongly correlated with internet access rates? Can we build a model that accurately predicts said rates?

Are internet access rates a stronger predictor of poverty rates than other forms of social investment (ie roads, schools, hospitals)?

Do these effects extend across internet technologies (cell phones and broadband internet)? If not, which type of infrastructure investment is better.

Motivation

We are interested in this problem as data scientists because our field is a mixed bag. On one hand, big data can be used to influence elections, spread hateful propaganda, and be used to track every purchase and decision we make. These political consequences are well known. However, the Internet has a history of advancing economies, and those without the internet tend to be left behind. To speak about this in particular, we need to investigate the ways in which internet access influences occupational outlook while controlling for other confounding factors like geography, race, and infrastructure investment more generally.

Literature Review

Poverty studies have a long history and ranges from the confirmation that poverty is a natural process to the fact that it was artificially created by the society itself. There is no doubt that poverty is more complex concept than just separation of people who live below set poverty line. Kuznets argues that growth and inequality go hand in hand, only stabilizing as an economy matures. However, [1] Kuznets' analysis is heavily critiqued by Binerjee et all "Inequality dynamics depend primarily on the policies and institutions adopted by governments and societies as a whole"[2]. Almost every study, despite the methodology and whether it was cross-country or single country, found a positive economic impact from fixed broadband [3][4]. However the "World Development Report 2016" raises the question of causality[5] – does internet access exclude poverty or is it merely an indicator of a growing economy?

The contribution of internet to development economic, and more specifically to the reduction of poverty has been studied by scholars and policy makers. Most of them try to get a response to one specific question: Does the increase of internet services and applications will increase GDP and improve the well-being of the poor? In 2006 Gillet examine the impact of broadband availability on economic activity in the US and the zip code. Using Ordinary Least Square model and data from 1998-2002 the study shows that the availability of broadband services adds as much as 1.4% to the employment growth regardless of actual adoption" (H. Galperin 2006).

In general there appears to be agreement with most studies that the impact is only noticeable after a certain threshold of broadband penetration (though the exact level remains imprecise) with several exceptions. One study found that of allICTs (i.e., fixed telephones, mobile, Internet use and broadband), broadband has the biggest economic impact (Qiang et al. 2009)[6]. However another study found that in a low-income economy, mobile has a bigger impact, both in terms of basic subscriptions and mobile broadband (Katz and Koutroumpis 2012)[7]. One of the studies found that mobile broadband actually has a negative impact possibly due to its complimentary effect and non-productive application (Thompson and Garbacz 2011) [8]

A different approach to appreciate the impact of the internet on development is to consider its distributed effect. Several studies take into account the effect on labor demand to examine the impact of internet

technologies. For instance, Atasoy in 2013 conduct a research about broadband diffusion and the labor market in the US between 1999 and 2007. Their finding show that broadband availability has a positive impact on the country employment rate (H. Galperin (2017). Those impacts are more accrued in country with large portion of college-educated workers. In general, most of the studies made around the distributive of internet agree that internet diffusion positively affect wages. The lack of data to study profoundly the issue at hand makes it difficult to appreciate thoroughly the impact of internet on economic growth. But most of the research point out that the advance countries are the principal beneficial of the propagation of the internet and for less advanced economies the impact of internet on the reduction of poverty is questionable

American Community Survey

Poverty: Percent below poverty line; Estimate

Median.Age: Median age; Estimate;

Percent Male: (number of males per 100 females) /2; Estimate

Population: Population; Estimate

Percent.White: Percent of people identifying as white alone; Estimate

Percent.Black: Percent of people identifying as black alone; Estimate

Percent.Native: Percent of people identifying as Native alone; Estimate

Percent.Asian: Percent of people identifying as Asian alone; Estimate

Percent.Other: Percent of people identifying as other alone; Estimate

Percent.Two.or.more.races: Percent of people identifying as two or more races; Estimate

Percent.Foreign.Born: Count of Foreign Born / Count of Native

Percent.in.Public.School: Percent in public school; Estimate; Population 3 years and over enrolled in school

Median.Annual.Income: Median income (dollars); Estimate; Households

Employment.Rate: Employment/Population Ratio; Estimate; Population 16 years and over

Median.Monthly.Housing.Costs: Occupied housing units; Estimate; MONTHLY HOUSING COSTS - Median (dollars)

Percent.Computer.In.Home: Percent; Estimate; TYPES OF COMPUTER - Has one or more types of computing devices: - Desktop or laptop

Percent.Smartphone.In.Home: Percent; Estimate; TYPES OF COMPUTER - Has one or more types of computing devices: - Smartphone

Percent.Internet.Subscriptions: Percent; Estimate; TYPE OF INTERNET SUBSCRIPTIONS - With an Internet subscription:

Percent.STEM.education: Percent; Estimate; Trained/Educated in Science and Engineering

Employee of private company workers; Estimate; Service occupations: Percent in Service Industry

Percent.Occupied.Single.Family.Homes: Percent owner-occupied housing units; Estimate; UNITS IN STRUCTURE - 1, detached

Percent.With.Broadband:: With a computer - Percent Broadband Internet Subscription; Estimate; Total population in households

Percent.Education level: Percent; Estimate; Population 25 years and over with respective education levels, broken into high school, some college, bachelor's, and graduate degrees

Percent.Limited.English.Households: Percent limited English-speaking households; Estimate; All households

Annual Survey of State Finances
For the financial analysis section, we investigated state spending on social programs and infrastructure as divided into the following categories: Hospitals, Health, Highways, Parks.and.recreation, Police protection, Public Welfare, Governmental Administration, Correction, Natural Resources and State debt. Additionally, we included the technological infrastructure indicators from above so that we could see how they compare to state spending efforts to combat poverty.

World Bank Data was used for the map data including high tech exports, internet subscription rates, cell phone subscription rates, and broadband subscription rates.

IEE MAC Address Blocks are were used as an indicator of the number of internet-related manufacturers in a country.

This list of Internet Exchange Points was used as an indicator to which countries have the most internet infrastructure available to

Methodology

First we will examine the problem on a global scale using chloropleth maps that will inform our future choices.

We will build several models for predicting poverty rate, using a support vector machine, a neural net, a random forest regressor and the generalized linear model. In this way, we'll see how things like internet access and infrastructure investment influence poverty rates. The American Community Survey includes internet access rates, poverty, race, industry, language, occupation, place of birth, and familial origin. Using this data alone, we should be able to see if race or occupation is a better indicator of aggregate poverty than internet access rates. Using the financial data, we can investigate if state spending on infrastructure affects poverty more or less than technological infrastructure.

Hypothesis

Pew Research says that 20% of teens are unable to finish their homework due to the digital divide. The end result of this is likely low-skill careers and lower incomes. In fact, the internet tends to raise the tide for all, as a breadth study (also by Pew) showed that per capita income and access rates are highly correlated. We'd like to investigate the relationship between technology and the economy and see if we can build models resilient to the particle type of device. Previous work has used infrastructure investment to build logistic models for poverty using satellite images of infrastructure. It is also well known that poverty and broadband access rates are highly correlated. However, it is unknown if there is an underlying causal factor or if internet can, *by itself*, lift people out of poverty. The McKinney Global Institute did a massive study on the economic potential of internet investment in China that will inform our approach in this matter. Finally, the Internet Society, a global organization that builds internet infrastructure (mostly in the developing world), has compiled a list of internet penetration rates and other such metrics by country across the world. However, due to data collection limitations and the quality of data sources across continents, it would be impossible to investigate these things with respect to more generic features like race and infrastructure. Since the United States has a non-uniform income distribution across states, this should allow us to draw from a breadth of circumstances. Due to the multiplicative of effects in education, business opportunities, and spending opportunities available on the Internet, we suspect that governmental investment in digital infrastructure will have at least as much affect as road or school spending. Additionally, we suspect that this multiplier is reduced for cellular infrastructure relative to fixed (broadband) infrastructure because of the productivity gains associated with PCs over smartphones. This research will reveal to governments (both local and national) what kinds of infrastructure investment yields the most economic gains in the digital age. To our knowledge, this particular question has not been answered.

Correlation between Various Technology Indicators and Poverty Rates

Data Initialization and Preprocessing

Fixing Missing Data

To fill in any missing values, we replaced the empty fields with the mean of the column in which they come from. The has the result of not changing the variance of the column, improving the fit of our models.

Evaluation of metrics

Our data occurs across many orders of magnitude. For the best fit, the data should be centered and scaled. Additionally, we have to impute some values because they are missing and we don't have enough data to discard those lines.

Correlation Plot of Predictors

We can see significant covariance in the data. Additionally, many data points have near-zero variance. Excluding the confounding variables will improve our model while reducing the number of irrelevant variables will make our correlation analysis stronger.

Data Pre-processing

For preprocessing of data, we remove near zero predictors, fill in missing values with KNN method, and transform predictors using the Yeo-Johnson transformation method. We also center and scale the data. Additionally, we remove the covariant terms and the ones with near zero variance here, since they will not improve our models.

Support Vector Machine

Comparing SVM with Neural Networks (NN), both are non-linear algorithms. A Support Vector Machine with different kernels is comparable to a Neural Network with different layers. One advantage SVMs have over NNs is that NNs need large amounts of data to train, SVMs work with smaller-sized data with less computing power. Finally SVM usually only have 2-3 parameters to tune, they are easy to code, but the parameters don't correspond to a given

Model	RMSE.train	RSquared.train	RMSE.test	RSquared.test
Support Vector Machine	2.8493	0.8711891	1.5309	0.5936714

The support vector machine performed decently well, but since it optimizes for two introduced variables, it does little to help us explain poverty rates in the context of the indicators. Support vector machines are used for reducing dimensionality, but have no corresponding descriptive value. So, it is inappropriate for our purposes, even if it is relatively accurate.

Random Forest Model

Random forests are a modification of bagging that builds a large collection of de-correlated trees [1]. They are considered to belong in the category of non-parametric models since the number of parameters grows with the size of the training set. They are considered to be an improvement to the use of CART (Classification and Regression Tree) models because they do not suffer from some of the problems associated with CART models, such as the fact that CART models are unstable: small changes to the structure of the input data can have large effects on the CART model [2]. Random forests are designed to be low-variance estimators.

Random forests are based on the basic idea of aggregating uncorrelated sets of predictors, since one way to reduce the variance of an estimate is to average several estimates together [2]. A random forest trains a randomly chosen set of input variables over a randomly chosen subset of the data, and aggregates together several such trees to produce an overall estimator. Random forests have proven to be quite successful in a variety of real-world applications and often are seen to generalize very well to unseen real-world data.

```
## Percent.Occupied.Single.Family.Homes                          White
##                                2.838490                       3.111520
##                                Black                Percent.High.School
##                                3.678276                       3.810775
##                               Native                         Population
##                                4.692691                       5.512052
##                           Median.Age        Percent.Less.Than.High.School
##                                7.103013                      13.537891
##                          Two.or.more            Percent.in.Public.School
##                               14.584543                      18.911294
##                           Percent.male                 Percent.Bachelors
##                               26.580138                      37.131630
##                  Percent.Foreign.Born          Percent.In.Service.Industry
##                               49.146696                      54.888750
##      Percent.Smartphone.In.Household                              Asian
##                               56.404661                      64.775034
##                  Percent.Some.College               Percent.With.Broadband
##                               66.898220                      70.038735
##        Percent.Computer.in.Household               Percent.STEM.education
##                               72.870669                      82.809302
## Percent.No.of.Internet.Subscriptions                  Population.Graduate
##                               82.848777                      83.030706
##         Median.Monthly.Housing.Costs                             Other
##                               83.573760                      86.814715
##   Percent.Limited.English.Households                   Employment.Rate
##                               96.580679                      96.973178
##                  Median.Annual.Income
##                              134.790094
```

The variable importance plot is composed of two subplots. The left one indicates the mean square error introduced by excluding a given variable and is a measure of an indicator's relevance to the model. The right one indicates the nodes that allow for the fewest number of decision splits in our model and are very similar to the linear model below, likely due to our minimization of model complexity (as measured by AIC). The random forest regressor works by using automation and sampling to make many regression models before investigating the resulting output relative to the sensitivity and precision of the input. In effect, it builds hundreds or thousands of generalized regression models to determine the variables that affect the output the most. Since the random forest model is iterative and recursive, it seems like an appropriate model for describing our observations although some authors dispute it's integrity in predicting time-series data [3]. Since our purpose here is descriptive rather than prescriptive, the RMSE and R^2 will determine its efficacy.

Model	RMSE.train	RSquared.train	RMSE.test	
Random Forest	1.876554	0.9704061	1.655401	
Like the general	ized linear m	odel below, the R	andom Forest	regressor is able to predict poverty to within 1-2% ar

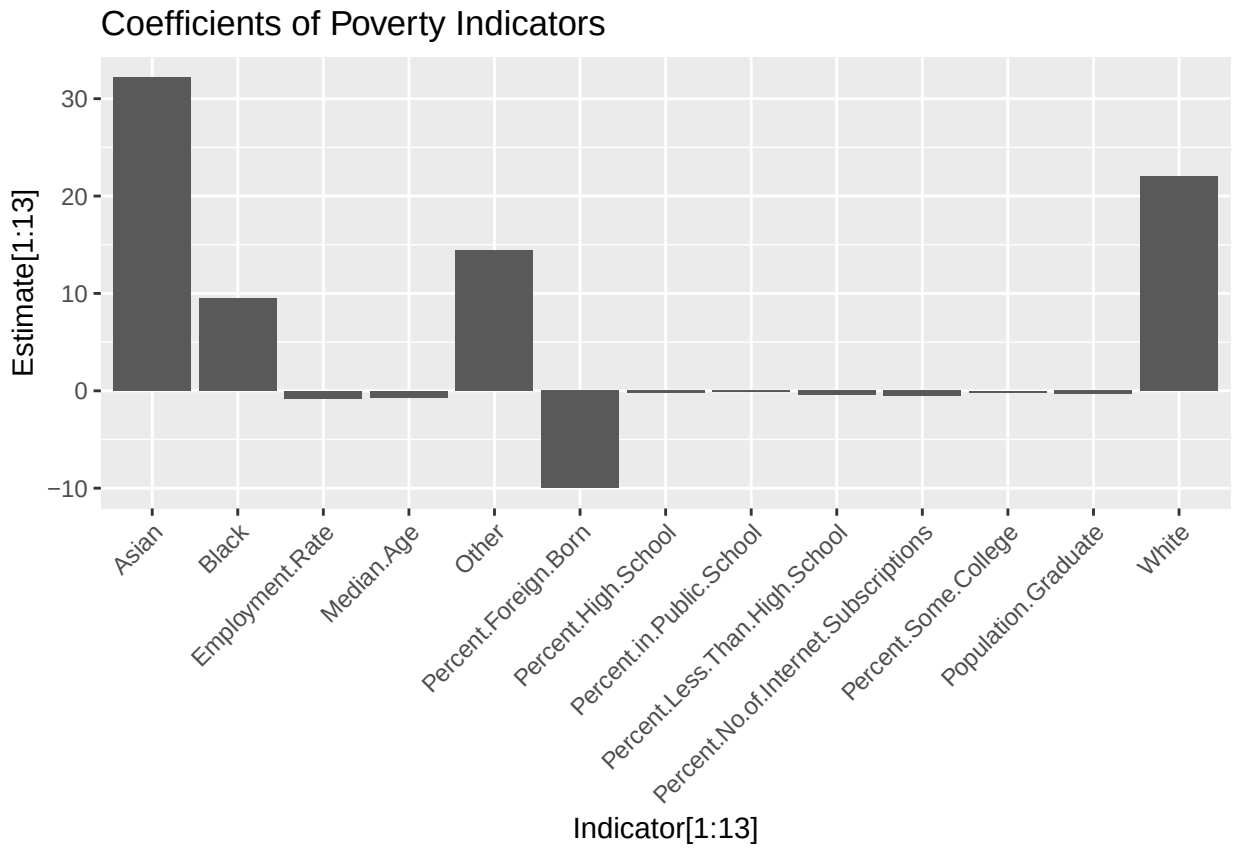
Neural Network Model

Neural Networks are a powerful nonlinear technique inspired by theories about how the human brain works [5]. Neural Networks can be classifiers (when the output variable is categorical) or regression (when the output variable is numeric). In this problem we use a regression artificial neural network (ANN) using the nnet package in R. Below we build and evaluate a Neural Network model of the regression problem.

Model	RMSE.train	RSquared.train	RMSE.test	
NeuralNet	5.997393	0.2715674	3.939332	
The neural n	et, however,	has inferior RMSE	and %R^2% v	alues and is rather poor as a descriptive model. Since t

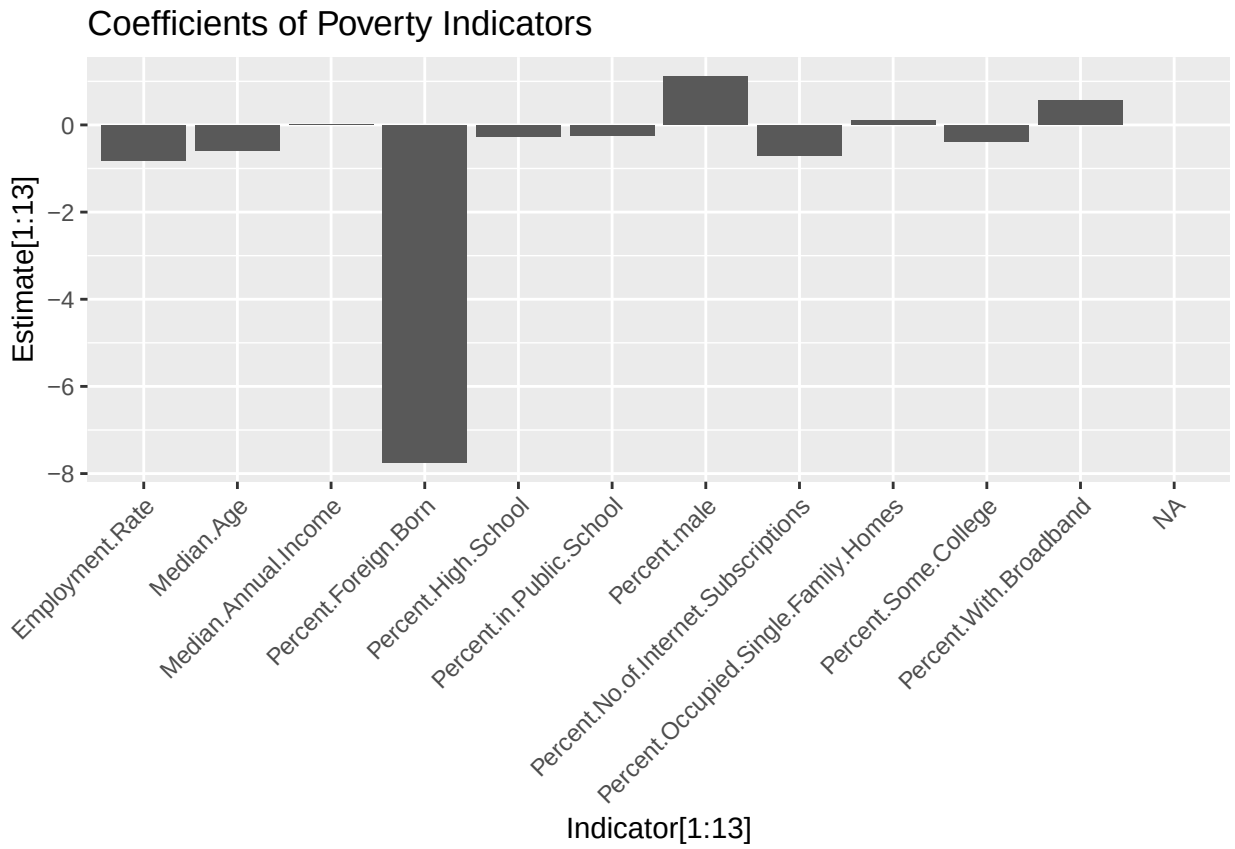
Generalized Linear Model

Because our output value is continuous and our data is numeric, we can use a generalized linear model to compute the pH. These models are generic and assume linearity in response. We will use the "Gaussian" type which assumes normally distributed variables. Mwabu et al. (2000) used regression analysis and identified the following variables as the key determinants of poverty: size of household, places of residence(urban or rural), level of schooling and livestock [9]. We tried to pick indicators specific to the culture in which they were measured– that of America in 2017. So, we looked at race, education, and occupational indicators.



Model	RMSE.train	RSquared.train	RMSE.test	RSquared.test
GLM	0.722634	0.9854818	2.196015	0.8323232

This model shows a somewhat even distribution of the value of poverty indicators other than race indicators and percent foreign born. However these appear to encode geography which may be a confounding factor. In particular, New York and California have large and diverse populations and the strongest economies. More generally, the model shows that all that all racial indicators have positive effects on poverty rates. Interestingly, immigration appears to be correlated with a decrease in poverty rates, suggesting again that this model appears to be encoding the more diverse and urban states. Despite that the model is able to predict poverty rates to within 1% for the training set and 2% for the training set, which is an acceptable deviation. Additionally, the test set has an R^2 value of 83%, meaning that 83% of the variance in the poverty rate is explained by the model. Below, we run the model again, excluding racial indicators. The number of immigrants seems to be the largest indicator of poverty rates and is inverse related to poverty. Somebody should inform the president. Additionally, the technological indicators seem to be influenctial as median income or median age, which is what we expected.



Model	RMSE.train	RSquared.train	RMSE.test	RSquared.test
GLM	0.8127761	0.9816339	2.903837	0.6964135

Conclusion: Social Indicators

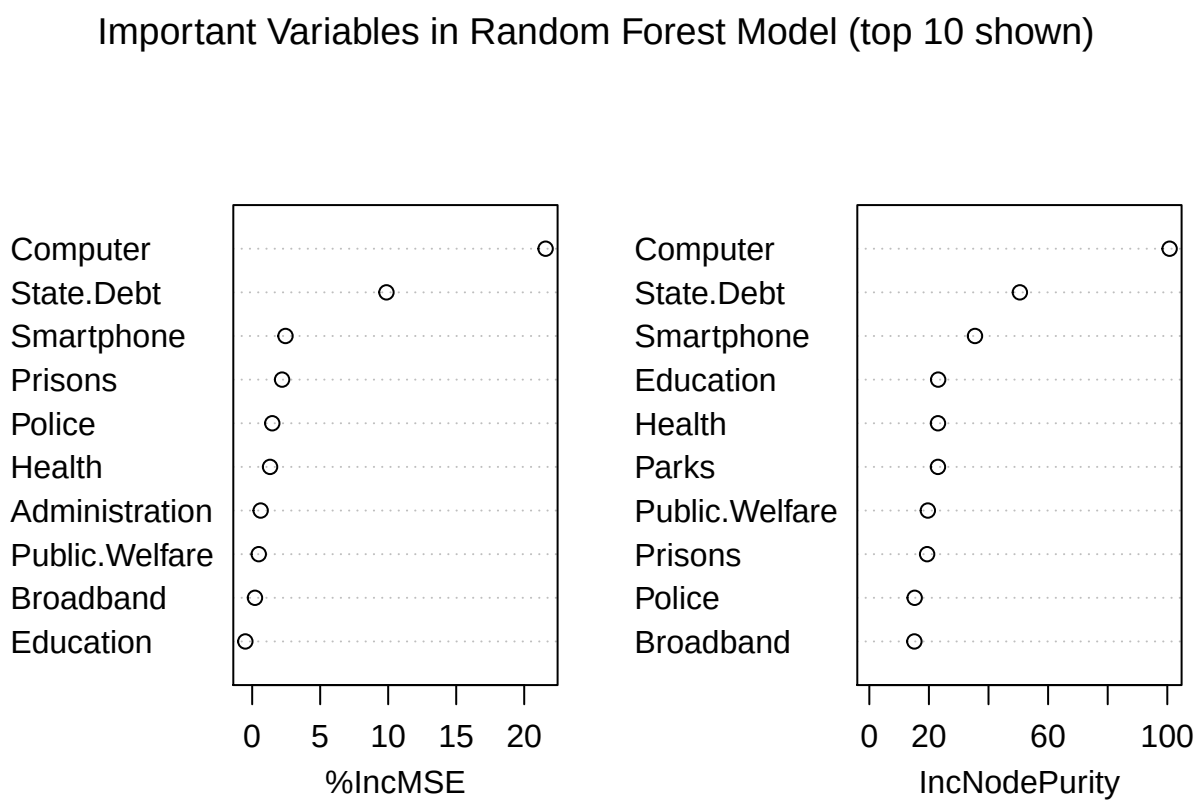
We were successfully able to evaluate the efficacy of several different models. The non-parametric models did not perform as well as the parametric one, perhaps because we had reduced our dimensional during pre-processing. Additionally, they both suffer from large test/train splits. The linear model performed very well, but the strongest indicators were raced based– seeming to an indicate geography as a confounding factor. The random forest method performed only slightly worse than the generalized linear model, but due to the iterative nature of the algorithm, it had a superior test/train split indicating more generalizability. The generalized linear model indicates that race is the most important factor, but is optimized for the smallest number of parameters. Additionally, our results could be explained by the confounding effects of geography, since race appears to be a significant factor across the board. The random forest model corroborates this idea, by highlighting the estimation purity factors as similar to the results from the linear model. However, the data vectors that explain poverty rates the most are not at all surprising – housing prices, median income, and the unemployment. However, computer access seems to be a stronger indicator than any given education level and either the broadband metric or the smartphone metric. Maybe there's so economic that we can isolate to the creative potential of having a computer in the house. Below, we see the RMSE and R^2 values for all models. Because the glm and random forest have human-readable coefficients/factors and performed the best, we will continue using them moving forward.

TODO: Remove this code chunk from the pdf. I couldn't figure out why/how it's staying when the others have the same options.

Finance Analysis

The American Community Survey also collects data from state governments about their spending habits. Below, we charted all social expenditures against the 'Broadband,' 'Smartphone,' and 'Computer' indicators. Additionally, because these numbers occur at very different scales, we applied the same preprocessing techniques above to build normally distributed vectors as required by the linear model.

First, we use the random forest model because the variable importance that it measures will inform our linear model below. As above, this model works by alliteratively building regression models and measuring the resulting change to the response variable.

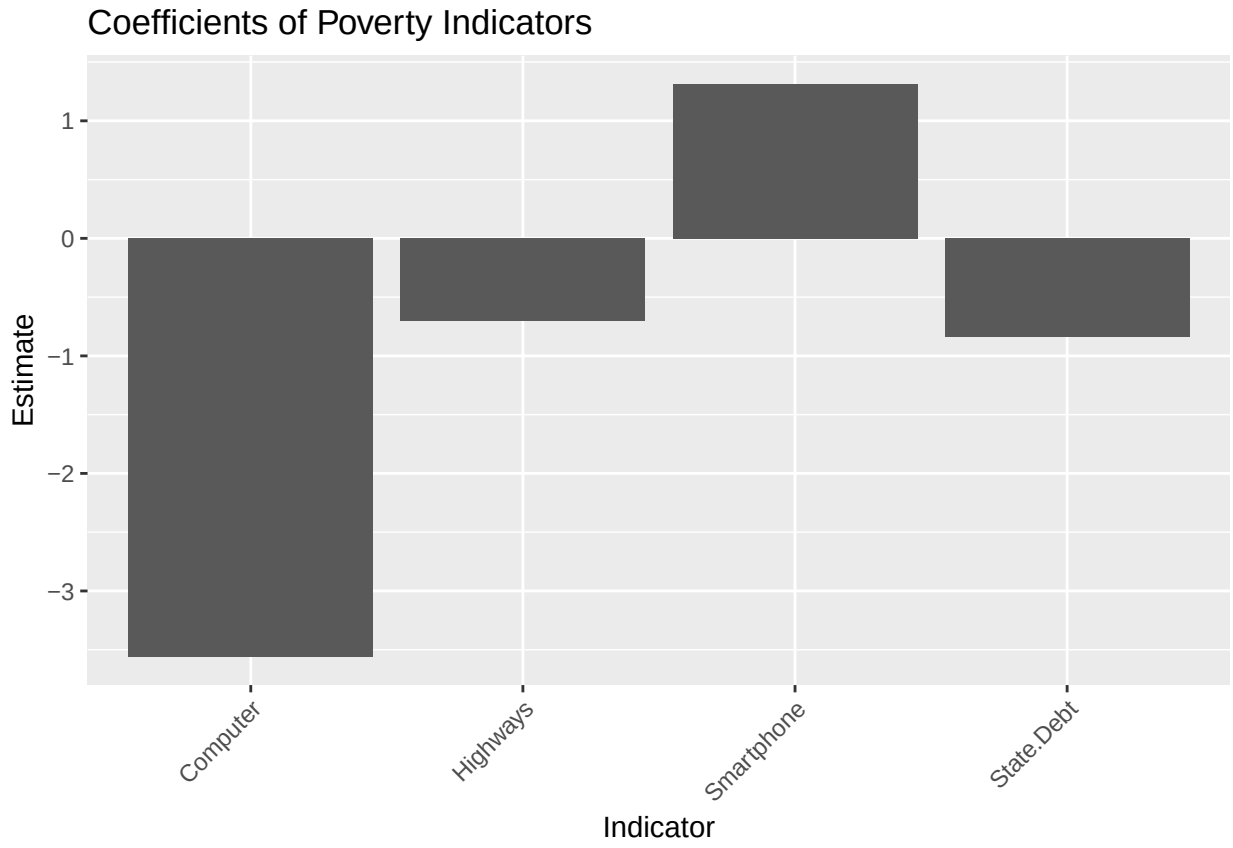


Model	RMSE.train	RSquared.train	RMSE.test	RSquared.test
Random Forest	1.083638	0.9656586	2.220804	0.8002246

The most important variables on the financial data set are computer, smartphone, and state debt. That is true for both precision and accuracy. Notably, the computer indicator is a better indicator of poverty status than broadband access, suggesting that the ability to learn and create technology is far more valuable than mere access. Additionally, state debt's strength corroborates our hypothesis above about confounding geography. Unfortunately, the scale of prison and police funding is a stronger indicator of poverty rates than the spending on health, welfare, and education. This does not mean that poor people are criminals, just that states tend to appropriate more funds for policing in poor areas. Education is a far more stable predictor as indicated by the chart on the right.

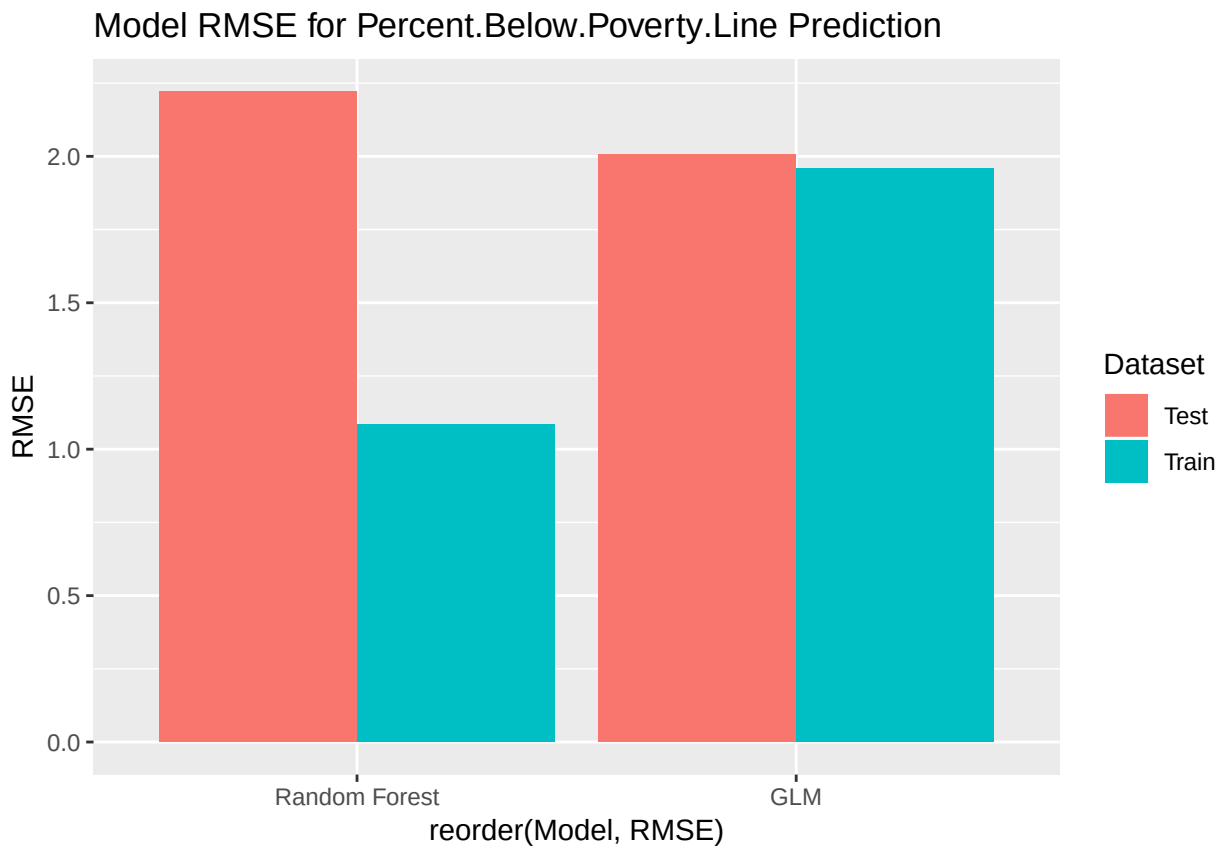
Model	RMSE.train	RSquared.train	RMSE.test	RSquared.test

Model	RMSE.train	RSquared.train	RMSE.test	RSquared.test
GLM	1.957209	0.6317318	2.005328	0.7493103

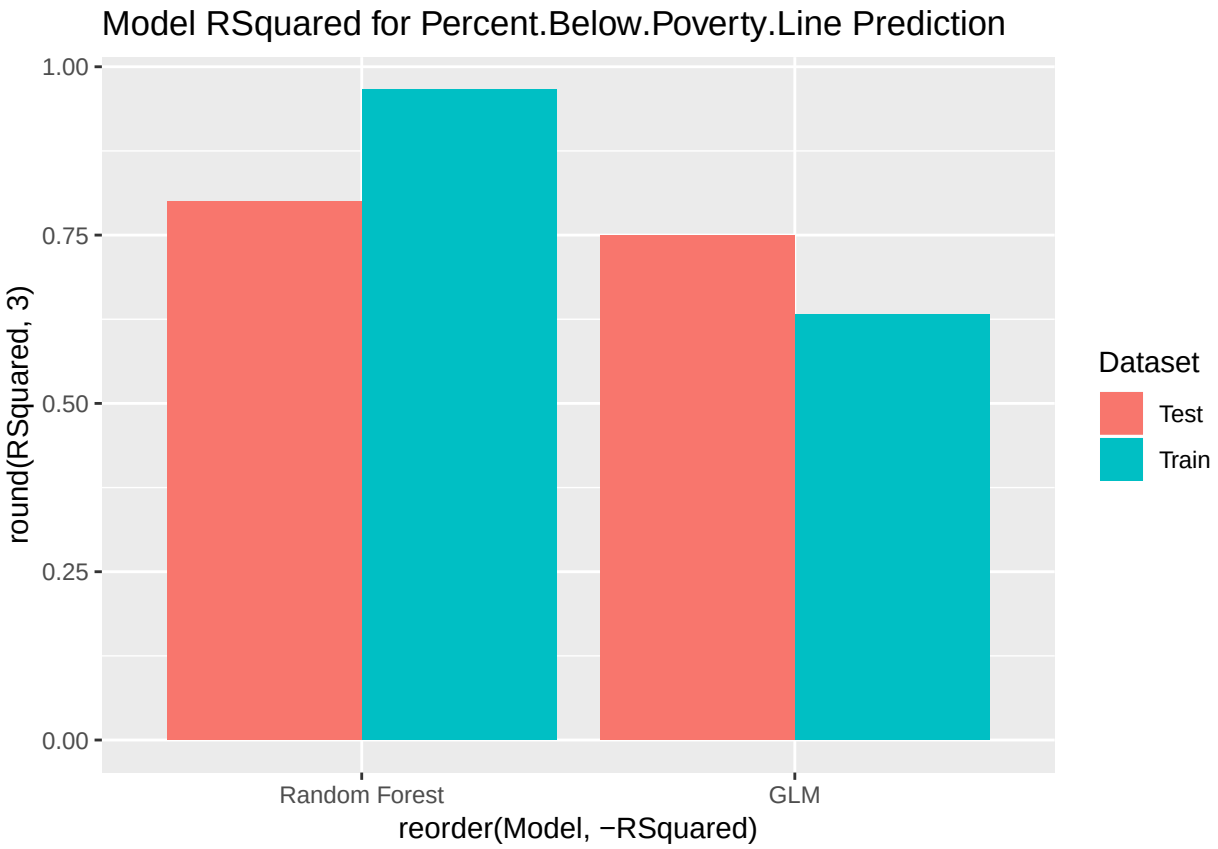


When examined from the context of state spending, computers, highways, smart phones, and state.debt are the most important measures for poverty rates. Interestingly, a higher computer ownership rate indicators a lower poverty rate while the opposite is true of smartphones. Additionally, state debt spending (particularly on highways) has a larger effect on explained variance than any social indicator (like education spending, hospital budgets, or prison spending).

```
modelperf = data.frame(matrix(ncol=4, nrow=4))
colnames(modelperf) = c("Dataset", "Model", "RMSE", "RSquared")


modelperf[1,] = list("Train", "Random Forest", rmse.rf.train, r2.rf.train)
modelperf[2,] = list("Test", "Random Forest", rmse.rf.test, r2.rf.test)
modelperf[3,] = list("Train", "GLM", rmse.glm.train, r2.glm.train)
modelperf[4,] = list("Test", "GLM", rmse.glm.test, r2.glm.test)

ggplot(data=modelperf, aes(x=reorder(Model, RMSE), y=RMSE, fill=Dataset)) +
    geom_bar(stat="identity", position=position_dodge()) +
    ggtitle("Model RMSE for Percent.Below.Poverty.Line Prediction")
```




```
ggplot(data=modelperf, aes(x=reorder(Model, -RSquared), y=round(RSquared,3), fill=Dataset)) +
    geom_bar(stat="identity", position=position_dodge()) +
    ggtitle("Model RSquared for Percent.Below.Poverty.Line Prediction")
```



As we can see above, both models are good enough to be relied on for describing the relationships between our data vectors. However, the test/train gap present in both indicates a lack of generalizability in the models.

Conclusion

Across the board, our models seem to indicate that having a laptop or desktop computer is a far more valuable indicator of poverty rate than smartphone or broadband subscription rates. While the social models are dominated by racial considerations, the positive value of all racial coefficients seems to favor diversity over a given race. Additionally, this indicates that geography is a confounding factor in these models as the wealthiest states have the most immigration and the most diverse populations. When it comes to the economic models, technological access is a much strong indicator of poverty rates than state spending on education, schools or police. Since our models are not totally consistent, it's impossible to draw a strong conclusion about correlation or causality. However, it's clear that technological access is important to a healthy economy and a reduction of poverty rates. Luckily, there are ways to measure these things more directly. Unfortunately, they are outside the scope of this investigation.

Next Steps

There are many ways we can continue improving the model performance, one method could be running more times of cross validation on more folds than 3 times 3-fold we have now for SVM and GBM models. It would take a longer time to compute, but the results would likely be better. Finally, more data would help us build a better model, in particular because the gap between the test and train sets tends to be relatively large

across all of the models, with the exception of the Random Forest regressor because it builds its model over 1000 iterations.

We could also use time-series data to examine causal relationships between the data using either Bayesian networks or Granger inference. The former requires a much larger amount of data (perhaps all county level census data since 2010) and the Granger analysis requires continuous time-series data. We simply didn't have time to collect the sheet volume of data needed to do those analysis, but that is planned for the future.

References

TO DO: APA format
1. Random Forests. https://uc-r.github.io/random_forests
2. Kevin Murphy (2012). Machine Learning a Probabilistic Perspective.
3. Support Vector Machine. https://uc-r.github.io/svm
4. Support Vector Machines. http://web.mit.edu/6.034/wwwbob/svm.pdf
5. Kuhn et al (2013). Applied Predictive Modeling

References: Atasoy, H. (2013) 'The Effects of Broadband Internet Expansion on Labor Market Outcomes', Industrial and Labor Relations Review 66(2): 315–45 Galperin, H.; Fernanda Viecens (2017) Connected for Development. Theory and evidence about the impact of internet technologies on poverty alleviation. Development Policy Review, o(o) :1-22 Gillett, S.; Lehr, W.; Osorio, C. and Marvin, S. (2006) Measuring the Economic Impact of Broadband Deployment. Washington, DC: US Department of Commerce, Economic Development Administration. Koutroumpis, P. (2009) 'The Economic Impact of Broadband on Growth: A simultaneous approach', Telecommunications Policy 33(9): 471–85. StatCounter. (2014). StatCounter Global Stats. http://gs.statcounter.com/

UNPACS. (2014). United Nations Public Administration Country Studies Data Center.

We Are Social. (2014). Global Digital Statistics 2014. http://wearesocial.net/tag/sdmw/.

WHO. (2014). World Health Statistics 2014. Geneva: World Health Organization.

World Bank. (2009). Information and Communications for Development 2009:

Extending Reach and Increasing Impact. Washington. World Bank. (2014). Enterprise Surveys: What Business Experience.
http://www.enterprisesurveys.org/

World Bank. (2014). World Bank Open Data. http://data.worldbank.org/.

World Wide Worx. (2012). Internet Matters: The Quiet Engine of the South African Economy

Review of Development Economics, 6(2), 183–203, 2002

Abhijit Banerjee , Roland Benabou, Dilip Mookherjee (2006), "Understanding Poverty".

European Scientific Journal, August 2014 edition vol.10, No.24 ISSN: 1857 – 7881 (Print) e - ISSN 1857-7431

European Journal of Social Sciences – Volume 13, Number 1 (2010) : "A Logistic Regression Model to Identify Key Determinants of Poverty Using Demographic and Health Survey Data"

World development report (2016), "Exploring the Relationship Between Broadband and Economic Growth".

Christine Zhen-Wei Qiang and Carlo M. Rossotto with Kaoru Kimura (2009) Economic Impacts of Broadband.

Dr. Raúl L. Katz (2012), Economic and Social Impact of Broadband and Development of Digital Agendas

Thompson and Garbacz (2011), "Economic impacts of mobile versus fixed broadband".

Thomas Pave Sohnesen Niels Stender (2016) "Is Random Forest a SuperiorMethodology for Predicting Poverty?" Journal of Economic Literature Classification: I30, I32, N97 " Determinants of Poverty in Kenya"

Appendices:

R Source Code:

```
library(caret)
library(caTools)
library(corrplot)
library(e1071)
library(fastDummies)
library(forecast)
library(ggplot2)
library(imputeTS)
library(lattice)
library(knitr)
library(ModelMetrics)
library(nnet)
library(randomForest)
library(readxl)
library(reshape2)
library(tidyr)
library(tidyverse)
library(xlsx)
library(rworldmap)
library(tidycensus)
# Show loaded packages.
(.packages())
training <- read_csv(file = "Cleaned_Community_Data.csv")
states <- training$State
training$State <- NULL
training$X1 <- NULL
f=function(x){
   x<-as.numeric(as.character(x)) #first convert each column into numeric if it is from factor
   x[is.na(x)] =mean(x, na.rm=TRUE) #convert the item with NA to median value from the column
   x
}
training <- data.frame(sapply(training, f))
means <- sapply(training, mean, na.rm = TRUE)
sds   <- sapply(training, sd, na.rm = TRUE)
explore <- as.data.frame(cbind( means, sds))
p <- ggplot(explore, aes(x = row.names(explore), y = means))+
  geom_bar(stat = 'identity') +
  labs(title = "Means of Various Features") +
  xlab("Data Features") +
  ylab("Mean of Data") +
  theme(panel.background = element_blank()) +
  geom_errorbar(aes(ymin = means - sds, ymax = means + sds))
p + theme(axis.text.x = element_text(angle = 90))
results <- cor(training, method = 'pearson')
corrplot::corrplot(results, method = 'circle')
x_train <- subset(training, select = -Percent.Below.Poverty.Line )
y_train <- training$Percent.Below.Poverty.Line
transformed <- preProcess(x_train, method = c("center", "scale", "YeoJohnson", "nzv", "corr"))
```

```
x_train <- predict(transformed, x_train)

set.seed(100)

y_train.orig = y_train
sample = sample.split(y_train.orig, SplitRatio = .75)
y_train_svm = subset(y_train.orig, sample == TRUE)
y_test_svm = subset(y_train.orig, sample == FALSE)

training_svm = subset(x_train, sample == TRUE)
test_svm = subset(x_train, sample == FALSE)

fitControl <- trainControl(## 10-fold CV
                           method = "repeatedcv",
                           number = 10,
                           ## repeated ten times
                           repeats = 10)
svmFit <- train(training_svm, y_train_svm,
                method = "svmRadial",
                trControl = fitControl,
                tuneLength = 8,
                metric = "RMSE")
model1 <- svmFit

rmse.svm.train = rmse(predict(svmFit, training_svm), y_train_svm)
r2.svm.train = R2(predict(svmFit, training_svm), y_train_svm)
rmse.svm.test = rmse(predict(svmFit, test_svm), y_test_svm)
r2.svm.test = R2(predict(svmFit, test_svm), y_test_svm)


kable(data.frame(Model=c("Support Vector Machine"), RMSE.train=min(rmse.svm.train), RSquared.train=min(

# Split the training data into a portion that is withheld from the model and used to evaluate
# the model.

set.seed(123)
sample = sample.split(training$Percent.Below.Poverty.Line, SplitRatio = .75)
training_forest = subset(training, sample == TRUE)
test_forest = subset(training, sample == FALSE)

set.seed(123)
rfFit <- randomForest::randomForest(Percent.Below.Poverty.Line ~ ., data = training_forest, importance =
                                    ntree = 1000, keep.forest = TRUE)
model2 <- rfFit

importance <- importance(rfFit, scale = TRUE)
sort(importance[,2])

training_forest2 = dplyr::select(training_forest, -Percent.Below.Poverty.Line)
rfPred.train = predict(rfFit, training_forest2)
rmse.rf.train = rmse(training_forest$Percent.Below.Poverty.Line, rfPred.train)
r2.rf.train = R2(training_forest$Percent.Below.Poverty.Line, rfPred.train)
test_forest2 = dplyr::select(test_forest, -Percent.Below.Poverty.Line)
```

```
rfPred.test = predict(rfFit, test_forest2)
rmse.rf.test = rmse(test_forest$Percent.Below.Poverty.Line, rfPred.test)
r2.rf.test = R2(test_forest$Percent.Below.Poverty.Line, rfPred.test)
kable(data.frame(Model=c("Random Forest"), RMSE.train=c(rmse.rf.train), RSquared.train=c(r2.rf.train), I
set.seed(123)
training_nn = training_forest2
training_nn_y = training_forest$Percent.Below.Poverty.Line
test_nn = test_forest2
test_nn_y = test_forest$Percent.Below.Poverty.Line

nnetFit <- nnet(training_nn, training_nn_y,
                size = 4,
                decay = 0.01,
                linout = TRUE,
                trace = FALSE,
                maxit = 500, # Iterations
                ## Number of parameters used by the model
                MaxNWts= 4 * (ncol(training_nn) + 1) + 5 + 1)
model3 <- nnetFit
rmse.nnet.train = rmse(training_nn_y, predict(nnetFit, training_nn))
r2.nnet.train = mean(cor(training_nn_y, training_nn)^2)
rmse.nnet.test = rmse(test_nn_y, predict(nnetFit, test_nn))
r2.nnet.test = mean(cor(test_nn_y, test_nn)^2)
kable(data.frame(Model=c("NeuralNet"), RMSE.train=c(rmse.nnet.train), RSquared.train=c(r2.nnet.train), I



set.seed(123)
training_glm = subset(training, sample == TRUE)
test_glm = subset(training, sample == FALSE)

glm_train_label <- training_glm$Percent.Below.Poverty.Line
glm_test_label <- test_glm$Percent.Below.Poverty.Line
glm_train <- select(training_glm, -Percent.Below.Poverty.Line)
glm_test <- select(test_glm, -Percent.Below.Poverty.Line)

model4 <- glm(glm_train_label ~., glm_train, family = "gaussian")
model5 <- step(model4, direction = "backward", trace = FALSE)

sum4 <- summary(model4)
sum5 <- summary(model5)

sum4 <- as.data.frame(sum4$coefficients,)
sorted <- sum4[order(abs(sum4$Estimate), decreasing = TRUE),]
sorted <- add_rownames(sorted, "Indicator")
sum5 <- as.data.frame(sum5$coefficients,)
sorted <- sum5[order(abs(sum5$Estimate), decreasing = TRUE),]
sorted <- add_rownames(sorted, "Indicator")


plot <- ggplot(data = sorted[2:14,], aes(x=Indicator[1:13], y=Estimate[1:13])) + geom_bar(stat="identity
plot + theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
rmse.glm.train <- rmse(predict(model5, glm_train), glm_train_label)
r2.glm.train <- R2(predict(model5, glm_train), glm_train_label)
rmse.glm.test <- rmse(predict(model5, glm_test), glm_test_label)
r2.glm.test <- R2(predict(model5, glm_test), glm_test_label)
kable(data.frame(Model=c("GLM"), RMSE.train=c(rmse.glm.train), RSquared.train=c(r2.glm.train), RMSE.tes
set.seed(123)
training.without.race <- select(training, -White, -Other, -Black, -Asian, -Native, -Two.or.more)
training_glm = subset(training.without.race, sample == TRUE)
test_glm = subset(training.without.race, sample == FALSE)

glm_train_label <- training_glm$Percent.Below.Poverty.Line
glm_test_label <- test_glm$Percent.Below.Poverty.Line
glm_train <- select(training_glm, -Percent.Below.Poverty.Line)
glm_test <- select(test_glm, -Percent.Below.Poverty.Line)

model4 <- glm(glm_train_label ~., glm_train, family = "gaussian")
model5 <- step(model4, direction = "backward", trace = FALSE)

sum4 <- summary(model4)
sum5 <- summary(model5)

sum4 <- as.data.frame(sum4$coefficients,)
sorted <- sum4[order(abs(sum4$Estimate), decreasing = TRUE),]
sorted <- add_rownames(sorted, "Indicator")
sum5 <- as.data.frame(sum5$coefficients,)
sorted <- sum5[order(abs(sum5$Estimate), decreasing = TRUE),]
sorted <- add_rownames(sorted, "Indicator")


plot <- ggplot(data = sorted[2:14,], aes(x=Indicator[1:13], y=Estimate[1:13])) + geom_bar(stat="identit
plot + theme(axis.text.x = element_text(angle = 45, hjust = 1))

rmse.glm.train <- rmse(predict(model5, glm_train), glm_train_label)
r2.glm.train <- R2(predict(model5, glm_train), glm_train_label)
rmse.glm.test <- rmse(predict(model5, glm_test), glm_test_label)
r2.glm.test <- R2(predict(model5, glm_test), glm_test_label)
kable(data.frame(Model=c("GLM"), RMSE.train=c(rmse.glm.train), RSquared.train=c(r2.glm.train), RMSE.tes
modelperf = data.frame(matrix(ncol=4, nrow=8))
colnames(modelperf) = c("Dataset", "Model", "RMSE", "RSquared")


modelperf[1,] = list("Train", "Random Forest", rmse.rf.train, r2.rf.train)
modelperf[2,] = list("Test", "Random Forest", rmse.rf.test, r2.rf.test)
modelperf[3,] = list("Train", "SVM", rmse.svm.train, r2.svm.train)
modelperf[4,] = list("Test", "SVM", rmse.svm.test, r2.svm.test)
modelperf[5,] = list("Train", "GLM", rmse.glm.train, r2.glm.train)
modelperf[6,] = list("Test", "GLM", rmse.glm.test, r2.glm.test)
modelperf[8,] = list("Train", "Nnet", rmse.nnet.test, r2.nnet.train)
modelperf[7,] = list("Test", "NNet", rmse.nnet.train, r2.nnet.test)

ggplot(data=modelperf, aes(x=reorder(Model, RMSE), y=RMSE, fill=Dataset)) +
    geom_bar(stat="identity", position=position_dodge()) +
    ggtitle("Model RMSE for Percent.Below.Poverty.Line Prediction")
```

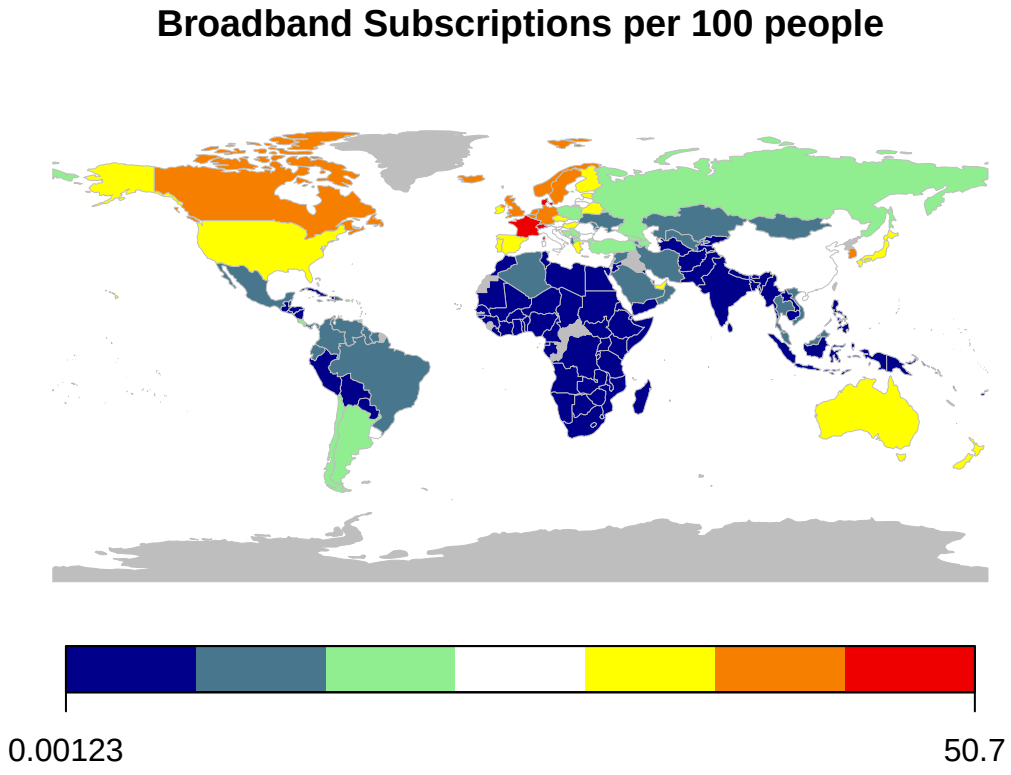
```r
ggplot(data=modelperf, aes(x=reorder(Model, -RSquared), y=round(RSquared,3), fill=Dataset)) +
    geom_bar(stat="identity", position=position_dodge()) +
    ggtitle("Model RSquared for Percent.Below.Poverty.Line Prediction")
finances <- read.csv("Cleaned_Finance_Data.csv")

target <- finances$Poverty

finances$X <- NULL
finances <- as.data.frame(cbind(finances$Education,
finances$Hospitals,
finances$Health,
finances$Highways,
finances$Parks.and.recreation,
finances$Police.protection,
finances$Governmental.administration,
finances$Correction,
finances$Natural.resources,
finances$Internet,
finances$Smartphone,
finances$Computer,
finances$Public.welfare,
finances$Debt.at.end.of.fiscal.year))


preProcValues <- preProcess(finances, method = c("center", "scale", "YeoJohnson", "nzv", "corr"))
finances <- predict(preProcValues, finances)


colnames(finances) <- c("Education", "Hospitals", "Health", "Highways", "Parks", "Police",  "Administra

finances$Poverty <- target


# Split the training data into a portion that is withheld from the model and used to evaluate
# the model.

set.seed(123)
set.seed(123)
sample = sample.split(finances$Poverty, SplitRatio = .75)
training_forest = subset(finances, sample == TRUE)
test_forest = subset(finances, sample == FALSE)
rfFit <- randomForest::randomForest(Poverty ~ ., data = training_forest, importance = TRUE,
                                    ntree = 1000, keep.forest = TRUE)
model6 <- rfFit

varImpPlot(rfFit, n.var=10,
           main="Important Variables in Random Forest Model (top 10 shown)")
training_forest2 = dplyr::select(training_forest, -Poverty)
rfPred.train = predict(rfFit, training_forest2)
rmse.rf.train = rmse(training_forest$Poverty, rfPred.train)
r2.rf.train = R2(training_forest$Poverty, rfPred.train)
test_forest2 = dplyr::select(test_forest, -Poverty)
```

```
rfPred.test = predict(rfFit, test_forest2)
rmse.rf.test = rmse(test_forest$Poverty, rfPred.test)
r2.rf.test = R2(test_forest$Poverty, rfPred.test)
kable(data.frame(Model=c("Random Forest"), RMSE.train=c(rmse.rf.train), RSquared.train=c(r2.rf.train), R

set.seed(123)
training_glm = subset(finances, sample == TRUE)
test_glm = subset(finances, sample == FALSE)

glm_train_label <- training_glm$Poverty
glm_test_label <- test_glm$Poverty
glm_train <- select(training_glm, -Poverty)
glm_test <- select(test_glm, -Poverty)

model4 <- glm(glm_train_label ~., glm_train, family = "gaussian")
model5 <- step(model4, direction = "backward", trace = FALSE)

sum4 <- summary(model4)
sum5 <- summary(model5)
sum5 <- as.data.frame(sum5$coefficients,)
sorted <- sum5[order(abs(sum5$Estimate), decreasing = TRUE),]
sorted <- add_rownames(sorted, "Indicator")


rmse.glm.train <- rmse(predict(model5, glm_train), glm_train_label)
r2.glm.train <- R2(predict(model5, glm_train), glm_train_label)
rmse.glm.test <- rmse(predict(model5, glm_test), glm_test_label)
r2.glm.test <- R2(predict(model5, glm_test), glm_test_label)
kable(data.frame(Model=c("GLM"), RMSE.train=c(rmse.glm.train), RSquared.train=c(r2.glm.train), RMSE.test

plot <- ggplot(data = sorted[-1,], aes(x=Indicator, y=Estimate)) + geom_bar(stat="identity", position=po
plot + theme(axis.text.x = element_text(angle = 45, hjust = 1))

modelperf = data.frame(matrix(ncol=4, nrow=4))
colnames(modelperf) = c("Dataset", "Model", "RMSE", "RSquared")


modelperf[1,] = list("Train", "Random Forest", rmse.rf.train, r2.rf.train)
modelperf[2,] = list("Test", "Random Forest", rmse.rf.test, r2.rf.test)
modelperf[3,] = list("Train", "GLM", rmse.glm.train, r2.glm.train)
modelperf[4,] = list("Test", "GLM", rmse.glm.test, r2.glm.test)

ggplot(data=modelperf, aes(x=reorder(Model, RMSE), y=RMSE, fill=Dataset)) +
    geom_bar(stat="identity", position=position_dodge()) +
    ggtitle("Model RMSE for Percent.Below.Poverty.Line Prediction")


ggplot(data=modelperf, aes(x=reorder(Model, -RSquared), y=round(RSquared,3), fill=Dataset)) +
    geom_bar(stat="identity", position=position_dodge()) +
    ggtitle("Model RSquared for Percent.Below.Poverty.Line Prediction")

df <- read.csv("Global/InternetIndicators.csv")
df2 <- read.csv("Cleaned_Community_Data.csv")
```

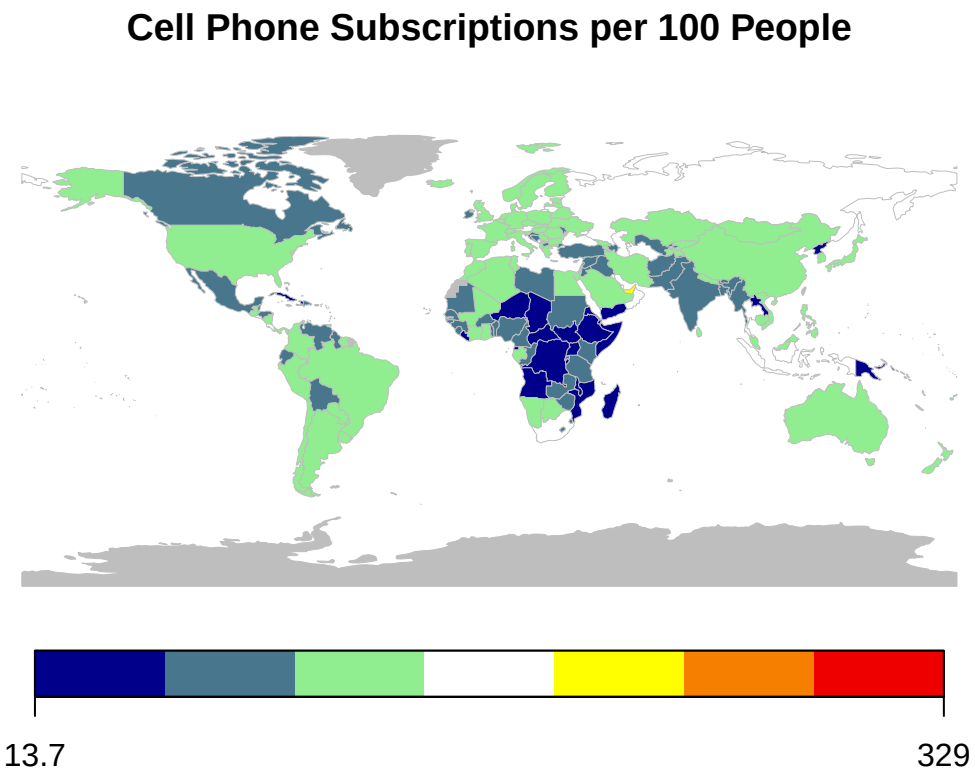
```
map <- suppressWarnings( joinCountryData2Map(df, nameJoinColumn = "Country", joinCode = "ISO3"))
df2
df$IXP[is.na(df$IXP)] <- 0

mapCountryData(map, nameColumnToPlot = "Broadband", mapTitle = "Broadband Subscriptions per 100 people"
mapCountryData(map, nameColumnToPlot = "Cells", mapTitle = "Cell Phone Subscriptions per 100 People", ca

mapCountryData(map, nameColumnToPlot = "Servers", mapTitle = "Servers per 10,000 people", catMethod = '

mapCountryData(map, nameColumnToPlot = "mac.addresses", mapTitle = "Mac Addresses Blocks Assigned per C
mapCountryData(map, nameColumnToPlot = "IXP", mapTitle = "Internet Exchange Points by Country", catMeth
mapCountryData(map, nameColumnToPlot = "Exports", mapTitle = "High Tech Exports (2017 USD)", catMethod =
```

Internet Indicators: A Global Perspective

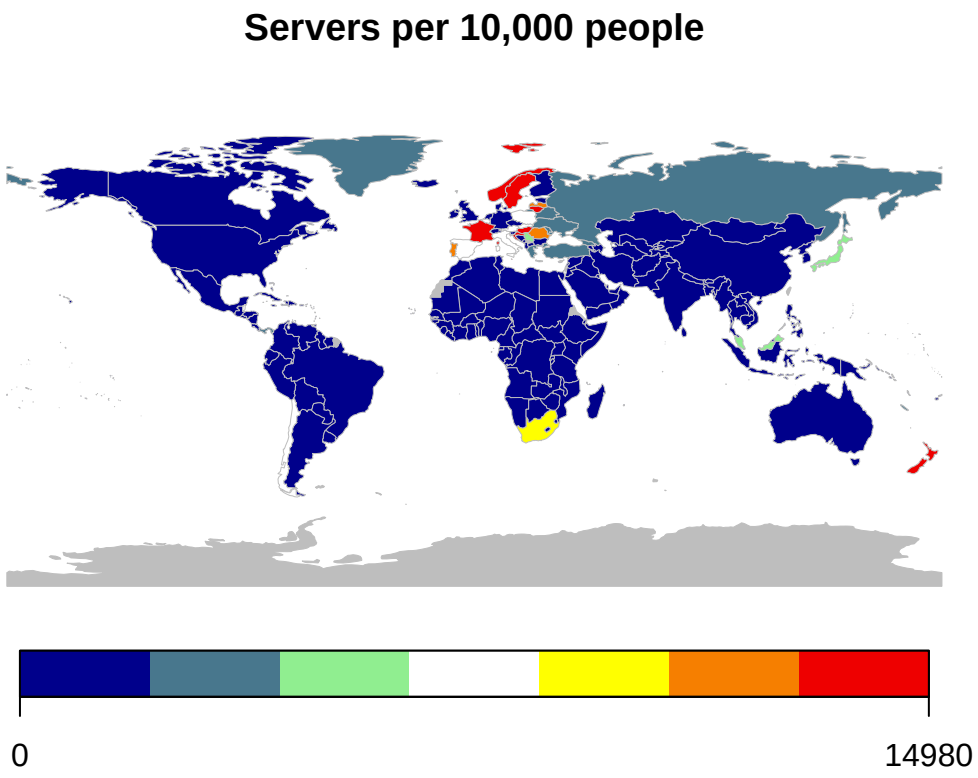
In general, global data is particularly useful for buildings models because currencies, cultures, and indicators vary so much for the topics covered above. However, there are a few things that seem to be measured reliably. Below, we have collected maps of those indicators and provided a wider context for our studies.



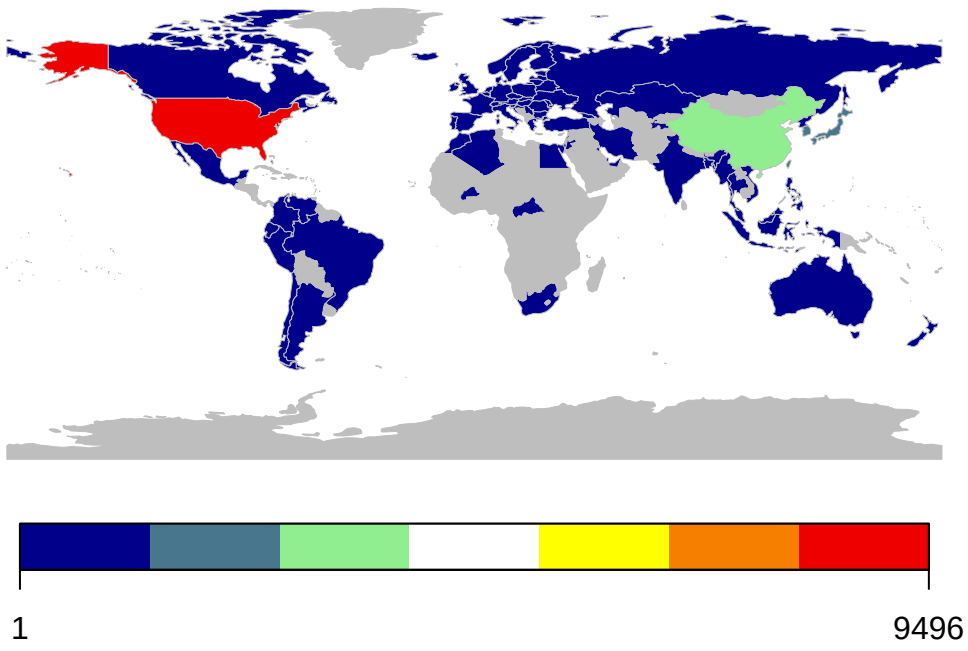
Immediately, we can see that broadband subscription rates are higher in strongly developed places like North America, Western Europe, and Australia. Conversely, poor countries across South America, Africa, and South Asia have significantly lower broadband access rates. Please note that countries in grey have unknown values.



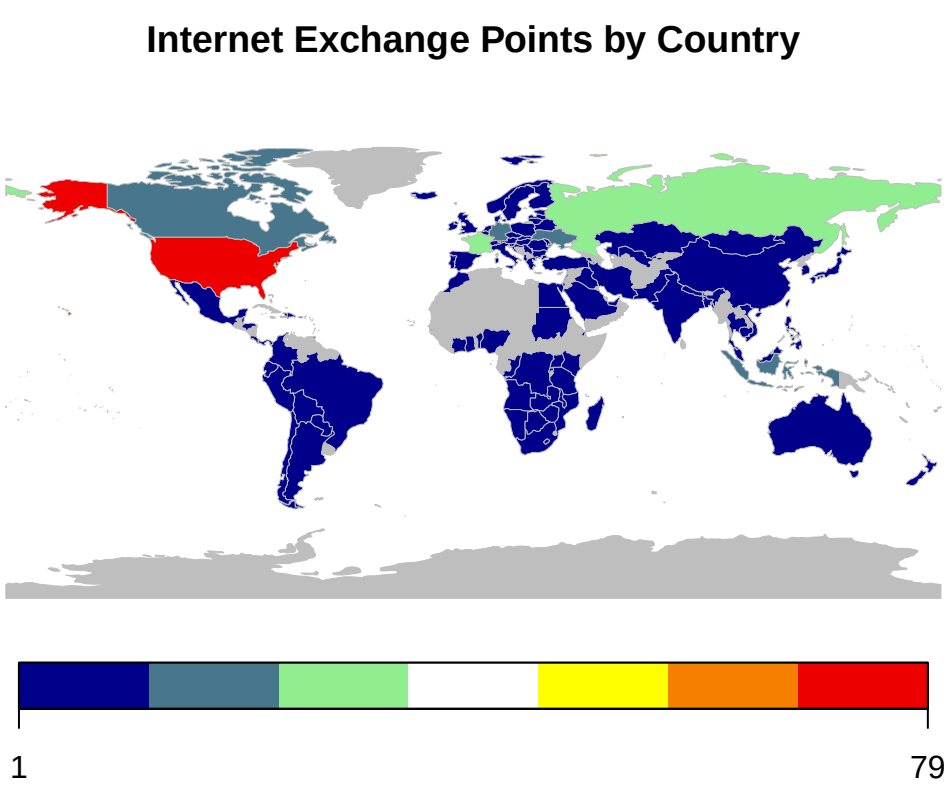
However, the number of cell phone subscriptions per 100 people is much more uniform. That is due to the lower cost of wireless network deployment compared to the capital-intensive processes of digging trenches to lay copper or fiber.



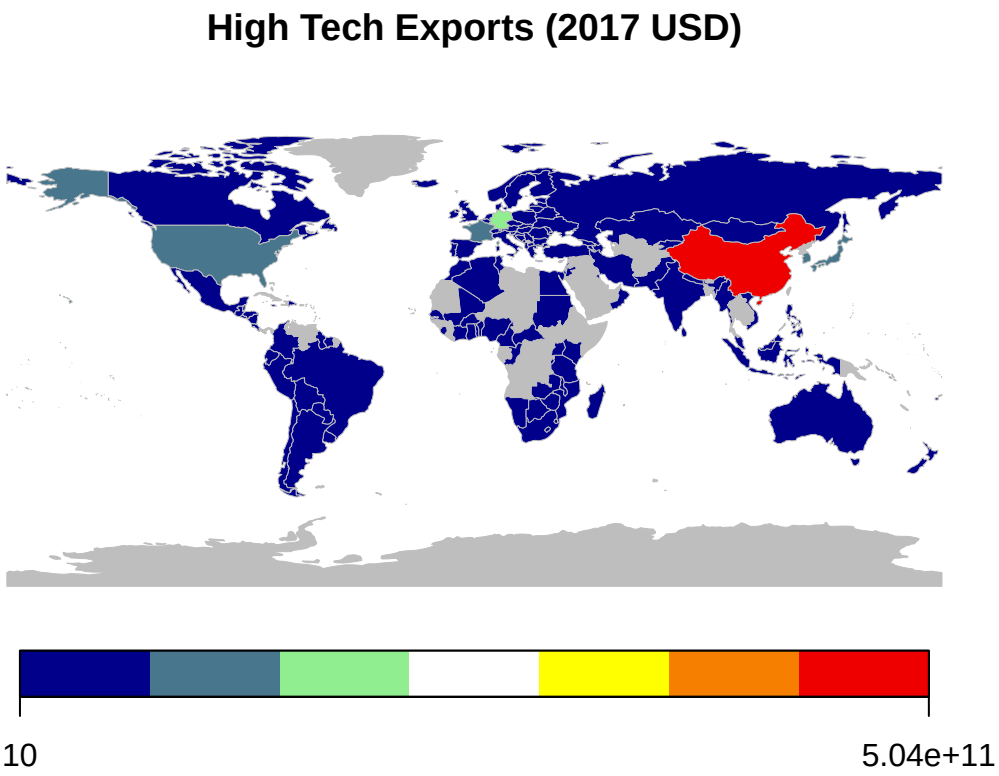
When we look at the number of servers available in each country, we find that Western Europe has the highest per capita server load. Countries like New Zeal and and South Africa are also high because they are conveniently located for undersea cables that compose the back-bone of the internet. TODO Source



The IEEE is a global organization that manages technological standards, publishes and circulates literature about electronics and and electrical engineering. In addition, they allocate MAC addresses which are the physical address of every bluetooth/WiFi radio, Ethernet port, and fiber cable on the internet. As we can see, a relatively small number of countries have original electronics manufacturers, with the US registering more than twice the number of devices as the next country (China).



The undersea cables mentioned earlier wind up at one of 600 buildings around the world where network operators connect their computers to their peers and create what we think of as the 'inter' net. These 600 buildings are not even distributed, with most countries only have a single access point to the Internet. In addition, regimes known for censorship (ie Egypt, Turkey, and China) have relatively few internet exchange points, allowing for centralized control and censorship. TODO: Source



The net result of the modern Internet infrastructure is a centralized model with a few players making all of the profits. Above we see the total amount of high tech exports as measured in 2017 USD. Three countries account for the bulk of the profit here, seeming to indicate that a centralized Internet infrastructure does not raise the standards for everybody. It is apparent that today's paradigm encourages consumption over creation.