Patent Volume Prediction Using Google Trends

Vishnu Murthy, Maria-Ann Conti

September 5, 2018

Abstract

This project aims to develop a method to forecast patent application volume using Google Trends information. One component of USPTO revenue is patent application filing fees therefore prediction of patent application volumes is important for budgeting purposes. This project utilized a mathematical model called a seasonal autoregressive model. as described in "Predicting the Present with Google Trends", Hal Varian et al. [1]. The model had parameters of previous patent filing volumes processed from the USPTO's Patent Examination Research Dataset (PatEx), as well as search query Google Trends data to fit a linear regression model. Two versions of the model were created, one model trained and evaluated on a random split while the other model train and evaluated on a data split by year. The first version of the model received a 9.66% testing error, while the other received a 7.47%. This approach can hopefully be combined with other forecasting models currently used by the USPTO Office of the Chief Financial Officer to more accurately respond to real-time end user data via Google Trends.

1 Introduction

1.1 Patent Volume Prediction

A large component of USPTO revenue is patent application filing fees. Basic filing fees range from $200-300$ depending on the type of patent and whether or not the filer is seeking a reissue [2]. Creating a model that can not only accurately predict the volume of patent application filings but also is sensitive to fluctuations in the market is very valuable, resulting in higher quality budget planning by the Office of the Chief Financial Officer and better preparation for large changes in volume.

Current models use an econometrics for prediction. The book, "Forecasting Innovations: Methods for Predicting Numbers of Patent Filings" by Hingley and Nicholas , summarizes the results of an external research programme that aimed to improve the forecasting methods in the European Patent Office. The models described in the book, which is similar to the models used in the USPTO, uses economic indicators such as Gross Domestic Product, Research & Development expenditures of various companies across a broad range of industries and number of employees in a particular year [3]. Although these models are helpful for forecasting patent filing volumes in the long-term, the economic models have difficulty in capturing short-term trends. Additionally, these indicators are relatively accurate at predicting patent volumes that come from amateurs or not from established companies that utilize patent attorneys.

Augmenting Google Trends information with econometric models can help capture these microtrends. Google is the largest search engine in the world, and Google Trends provides trends data in real time. To receive data from Google Trends, search a Google query in the Google Trends website and download a .csv file with a time period and its normalized query share; "query share" meaning that similar search queries are grouped together, and normalized to a maximum of 100. Google's Chief Economist Hal Varian describes a method of using a seasonal autoregressive model on Google Trends data for predicting the present. The example model is a linear regression used for predicting car sales. The model uses the previous month's sales, the month in the previous year's sales, as well as Google Trend search queries in the first week of the current month to predict the current month's sales. This project uses the seasonal autoregressive model with Google Trends in order to predict a given month's patent application volume. [1]

1.2 Data Processing

The two sources of data used were the USPTO's Patent Examination Research (PatEx) Dataset and Google Trends. The PatEx dataset contains detailed information on each patent filed per line of the file. Preprocessing required counting the number of patents filed per month, and used data from 2006-2015. A dataframe was created with columns containing shifts of this data every month from one month to twelve months before. The Google Trends data was taken from the Google Trends website, and the model used the following nine search queries: "Cost of Patent", "Patent Application Process", "Patent Application Search", "USPTO", "File for Patent", "Patent Filing Fees", "PatentsView", "Utility vs. Design Patent", and "EFS-web" numbed 1-9 as per Table 1. These terms were devised from the fact that a potential patent filler, especially an inventor who has not filed for a patent before, would search these terms on Google. In future models, more terms can be added. Google Trends provides query share data in a monthly format, so preprocessing required to use only 2006-2015 data. Both Trends and PatEx were normalized by subtracting the mean and dividing by standard deviation of each variable, allowing for comparison of coefficients to determine variable importance.

Trend Number	Trend	Explanation of Usage
1	Cost of Patent	Filers want to find the cost of a patent application
2	Utility vs. Design	Filers want to know the difference between a utility and design patent
3	USPTO	Filers search the Patent and Trademark Office's website
4	Patent Application Process	Filers want to learn about the application process
5	Patent Application Search	Filers want to search other patents
6	EFS-Web	Patent submission portal
7	PatentsView	PatentsView is the patent data visualization and analysis platform created by the USPTO
8	File For Patent	Filers want to learn how to file for a patent
9	Patent Filing Fees	Files want to patent know filing fees

Table 1: Google Trend number n query

1.3 Approach

The data was split into training and testing data. In one version of the model, the train test split was random. In the other version of the model, training data ranging from January 2006 to December 2013 and testing data ranging from January 2014 to November 2015. December 2015 was not included in the data due to the fact that it was a large outlier compared to the rest of the data. The mathematical model used for this project was a seasonal auto-regressive model, Figure 1. SciPy's *curve_fit* was used to fit the function to the training data.

$$\overline{x} = ax_{t-1} + bx_{t-1} + ... + kx_{t-11} + lx_{t-12} + mg_{1,t} + ng_{2,t} + ... + tg_{8,t} + ug_{9,t}$$

Figure 1: Model

where $\overline{x}$ is forecasted patent filing volume for month t, x_t is real patent filing volume in month t, $g_{n,t}$ is the normalized query share for Google Trend number n in month t, $a, b, c, ..., s, t, u$ are trainable parameters.

2 Results

The fitted parameters for both versions of the model are below.

Version 1 Model's Fitted Parameters: [0.201926, 0.102239, 0.119384, 0.15901, 0.162669, 0.11809, 0.083266, -0.101351, 0.163696, -0.478877, -0.214215, 0.486767, 0.153088, 0.022055, 0.311381, -0.019408, 0.200401, 0.056629, -0.095644, -0.004067, 0.02397]

Version 2 Model's Fitted Parameters: [-0.11460571, 0.04640811, 0.08898412, 0.14869581, 0.1430395, 0.08730534, 0.09995036, -0.10291923, 0.07270733, -0.25949317, -0.18018593, 0.67907019, 0.06264297, 0.01429775, 0.15065317, -0.06863694, 0.05723635, 0.02610932, -0.0636133, 0.00365002, 0.01983938]

Two metrics were used to measure the accuracy of the model - mean squared error and mean absolute percent error. These metrics were evaluated on the train and test data, recorded in Table 2.

Version	**Train/Test**	**Mean Squared Error**	**Mean Absolute Percent Error**
1	Train	20485631.96	5.07%
1	Test	28892846.94	9.66%
2	Train	20791118.27	5.616 %
2	Test	21208389.92	6.574 %

Table 2: Results - Mean Squared Error and Percent Error of train and test data

The test data's percent error for the first version of the model was a 9.66% while the second version of the model was a 6.57%. In both cases, the training data's mean squared error and percent error is slightly less than the test data's, indicating minor overfit.

Figure 2 & 3 is a plot showing the predicted volume against the actual volume. Although both graphs are plotted from Jan 2014, the first version of the model contains a mix of both train and test data due to its random split while the second version contains only test data.One important factor to notice is that both models are sensitive, meaning that it can forecast large changes in data. One can visually see that the models capture the microtends, or "spikes" in the data, very well.

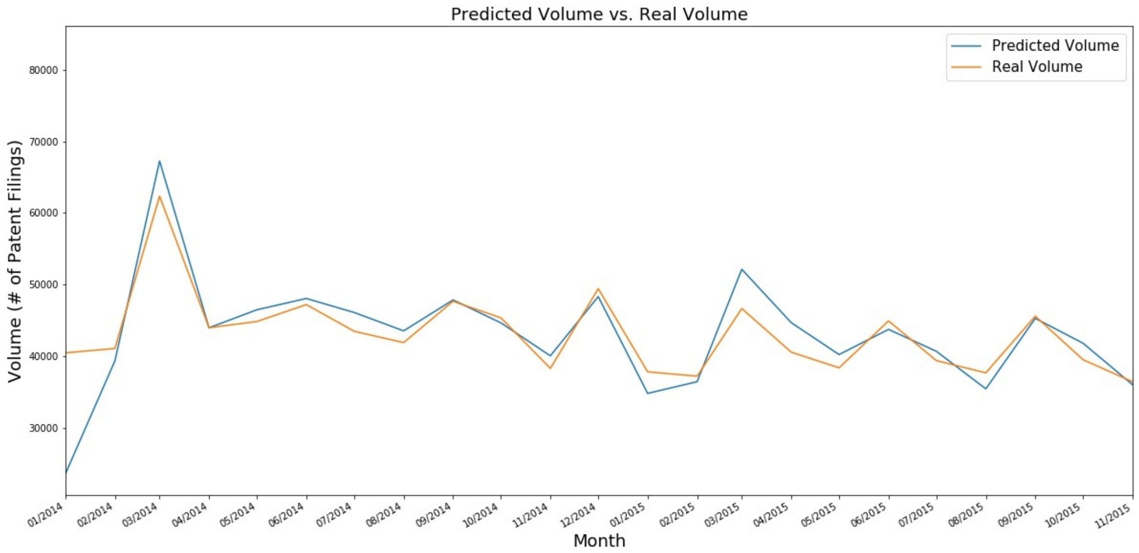

Figure 2: Predicted Volume vs. Real Volume Version 1

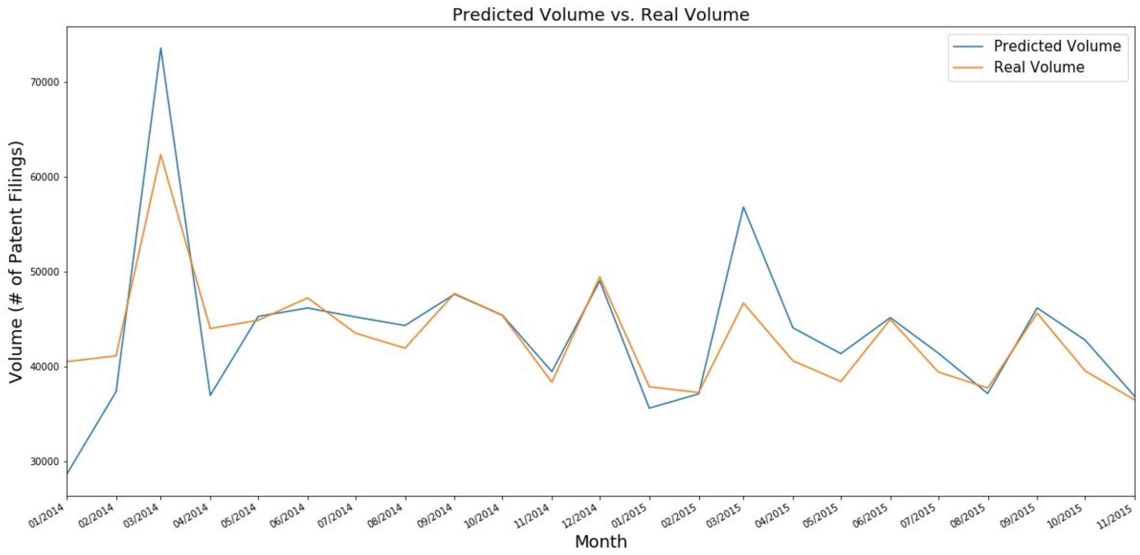

Figure 3: Predicted Volume vs. Real Volume Version 2

Comparing the absolute values of the coefficients can indicate variable importance, or correlation to patent application filing volumes, as the data is on the same scale such that each variable has a mean of 0 and standard deviation of 1. Tables 3 & 4 are the top five variables with the highest coefficient. Tables 5 & 6 are the top five trends with the highest coefficient.

Rank	**Standardized Coefficient**	**Variable**
1	0.486767	x_{t-12}
2	0.478877	x_{t-10}
3	0.311381	“USPTO”
4	0.214215	x_{t-11}
5	0.201926	x_{t-1}

Table 3: Top Five Most Correlated Variables Version 1

Rank	**Standardized Coefficient**	**Variable**
1	0.632613	x_{t-12}
2	0.318228	x_{t-10}
3	0.241018	x_{t-11}
4	0.21827	“USPTO”
5	0.177142	x_{t-4}

Table 4: Top Five Most Correlated Variables Version 2

The two models interestingly had similar results, with many of the terms and their orders being the same. Both models had a high correlation with the terms x_{t-12}, x_{t-11}, and x_{t-10}, implying that a yearly seasonal trend occurs. The Google Trend query “USPTO” also had a relatively high correlation.

3 Conclusion

This seasonal autoregressive model used for this project achieved optimal results. It received a 9.662% testing percent error on a random train test split model and a 6.57% testing percent error on a split based on year. Importantly, the model used minimal data for training and was sensitive to changes in the data, which is important for budgeting purposes. The model can be further improved by selecting higher quality Google Trends queries and using more granular data,

Rank	**Standardized Coefficient**	**Trend**
4	0.311381	USPTO
6	0.200401	Patent Application Search
7	0.153088	Cost Of Patent
12	0.095644	PatentsView
15	0.056629	EFS-Web

Table 5: Top Five Most Correlated Trends Version 1

Rank	**Standardized Coefficient**	**Trend**
4	0.21827	USPTO
6	0.154433	Patent Application Search
7	0.12249	Cost Of Patent
12	0.093497	PatentsView
15	0.077276	Patent Application Process

Table 6: Top Five Most Correlated Variables Version 2

resulting in a larger amount of data and possibly a higher accuracy model. Other machine learning or deep learning models can be explored such as Recurrent Neural Networks or Long Short-Term Memory networks, however these types of models usually require millions of data points. Because this model is optimal for finding microtrends, this model can be augmented with USPTO's current model to provide more accurate forecasting in the future.

4 Acknowledgements

This project was created as a part of the Student Innovation Challenge externship in the USPTO created by Mrs. Maria Conti. Special thanks to Mrs. Conti, Dr. Brantley, Chief Economist Mr. Nicholas Pairolero, Mr. Gerard Torres, Mr. Joseph Wolf, Mr. Gardy Rosius for help and support throughout the project. Email: vishnurmurthy@gmail.com, vishnu.murthy@uspto.gov, Maria-Ann.Conti@uspto.gov
Data: Google Trends, Patent Examination Research Dataset (Public PAIR)
Code Repository: https://github.com/vishnurmurthy/uspto/

References

[1] H. Varian and H. Choi, "Predicting the present with google trends," Apr 2009.

[2] O. of the Chief Financial Officer, Aug 2018.

[3] P. Hingley and M. Nicolas, *Forecasting innovations: methods for predicting numbers of patent filings*. Springer, 2006.