A Comparison of Time Series Model Forecasting Methods on Patent Groups

Mick Smith
Department of Computer Systems Technology
North Carolina A&T State University
csmith715@gmail.com

Rajeev Agrawal
Department of Computer Systems Technology
North Carolina A&T State University
ragrawal@ncat.edu

Abstract

The ability to create forecasts and discover trends is a value to almost any industry. The challenge comes in finding the right data and the appropriate tools to analyze and model such data. This paper aims to demonstrate that it may be possible to create technology forecasting models through the use of patent groups. The focus will be on applying time series modeling techniques to a collection of USPTO patents from 1996 to 2013. The techniques used are Holt-Winters Exponential Smoothing and ARIMA. Cross validation methods were used to determine the best fitting models and ultimately whether or not patent data could be modeled as a time series.

1. Introduction

As innovation and technology has grown over the last several decades there has arisen a greater need for tracking, grouping, and analyzing such progress. This is satisfied through the issuance of patents. Each patent can be thought of as an index in technological advancement since they introduce a new, innovative idea or theory. If these pieces of knowledge are to be considered benchmarks in the constantly changing landscape of technology, then it may be possible to examine the trends in quantities of patents.

The goal of this paper is to show that an opportunity exists to create a technology forecasting model based on the sequence of patents issued over a given time period. To accomplish this it is necessary to demonstrate that a time series model can accurately predict the fluctuations in patent volume from month to month. Due to the overwhelmingly large amount of patent data, this research will focus on three classes of data processing patents: Generic Control Systems or Specific Applications (GCSSA), Artificial Intelligence (AI), Database and File Management or Data Structures (DFMDS). Furthermore, this subset of patents will only include patents from 1996 to 2013. Two univariate time series forecasting models will be applied to each series of patents, Exponential Smoothing and Autoregressive Integrated Moving Averages (ARIMA).

Due to a decrease in storage costs and an increase in processing power, Big Data has created a situation in which a vast amount of information has been made available. As we progress into the next several years, there will be a great need to understand the massive amounts of structured and unstructured data that is a product of the Big Data phenomenon. As it will be demonstrated by this research, analysis of patents represents an area of great analytic potential. This paper will show that patent data is certainly a prospective source for a Technology Forecasting (TF) model. This will differ from other research in TF since other techniques do not consider the sequence of patent grants as a trend. Instead, they focus only on the cumulative content of patents for a set period of time with no respect to changes over that time period. Furthermore, the creation of TF models with patent data can go a long way in helping us understand the underlying meanings within a given technological sector. The trends and analyses that result from such models would benefit other areas of government, politics, economics, and social well-being.

2. Related Work

When attempting to forecast univariate time series data, it is generally accepted that parsimonious model techniques are followed. A simple approach that has been used in many applications is the Holt-Winters Exponential Smoothing (HWES) technique. Exponential smoothing techniques are simple tools for smoothing and forecasting a time series. Smoothing a time series aims at eliminating the irrelevant noise and extracting the general path followed by the series (Fried and George 2014). It is based on a recursive computing scheme, where the forecasts are

updated for each new incoming observation and is sometimes considered as a naive prediction method (Gelper et al. 2010).

Exponential smoothing methods were originally used in the 1950's as a collection of ad hoc techniques for extrapolating various types of univariate time series (De Gooijer and Hyndman 2006). In 1960 C.C. Holt and his student Peter Winters introduced a variation to the technique which ultimately became known as the Holt-Winters technique (De Gooijer and Hyndman 2006)(Goodwin 2010). Holt's initial model extended simple exponential smoothing to allow forecasting of data with a trend. Winters would later collaborate with his mentor to produce a seasonal component (Hyndman and Athanasopoulos 2013).

While Autoregressive (AR) and Moving Average (MA) models have been in existence since the early 1900's, it was the work of Box and Jenkins in 1970 that integrated these techniques into one approach and ultimately created ARIMA (De Gooijer and Hyndman 2006). The Box-Jenkins approach allowed for non-stationary time series trends to be modeled (Shumway and Stoffer 2006). Non-stationary data can be made stationary through a process known as differencing. In some time series models there is a need to adjust for seasonality. As previously mentioned both HWES and ARIMA offer alternative methods to adjust models accordingly. However, that is not the case with the data selected for this paper.

Time series modeling has been applied in several different settings and situations. Research has been carried out in economics (Kang 1996)(Dongdong 2010)(Timmermann and Granger 2004), climate change and weather forecasting (Kumar and De Ridder 2010)(Leixiao et al. 2013), utility forecasting (Conejo et al. 2005)(Contreras et al. 2003)(De Gooijer and Hyndman 2006), and many more.

Even though the only forecasting methods mentioned here are univariate, it is worth mentioning that multivariate techniques exist as well. Some of the more popular multivariate time series models that exist include VARIMA, VARMA, VAR, and BVAR. However, the impact that one patent trend may have on another might be substantial and should not be overlooked. When considering further research in patent analysis it is possible that these modeling techniques could be used.

It should be reiterated that the main objective of this paper is to demonstrate that groupings of patent data over time can be represented as a time series and that a forecasting model can be fitted to the trend. There is a lot of value in such technology forecasting, especially as it pertains to some level of patent mining. Technology forecast modeling on patent data has been done to show areas of technological development opportunities (Jun et al. 2011)(Tseng et al. 2007). Daim et al. (2006) suggest that the use of multiple methods, including Patent Mining, Bibliometrics, and Delphi processes, improves technology forecasting. Shin and Park (2009) have demonstrated that technology forecasting methods can be a key factor in economic growth. In their methods they use Brownian agents to detect regions of technology growth.

3. Proposed Methodology

In this analysis, each patent group is being considered independently of other patents. It was important to use this approach so that it could first be shown that a sequence of patents over a given time represented a meaningful time series and that predictive modeling could be carried out. However, in building on this research it will be important to understand the relationships between each group and the effect each one may have the others.

The patent data for this project was obtained from UC Berkley Fung Institute (https://github.com/funginstitute/downloads). Their patent data has been extracted from the USPTO website and converted from XML to a SQLite table structure. The patent databases provided include patent data ranging from 1975 to 2013. From these tables it was possible to filter out the number of patents in a given classification over a period of time (1996 to 2013). While the selection of dates is somewhat arbitrary, it does coincide with a rough starting date of commercial internet use. The USPTO classes and number of patents used in this research is shown in Table 1.

Name	USPTO Class	Number of Patents (1996 – 2013)
GCSSA	700	27,503
AI	706	8,699
DFMDS	707	53,415

Table 1 – Quantities and Classifications of Patents

Each particular class has several subclasses which offer greater specificity in the classification of the patent. It should be noted that if each class were to be broken into their smaller subclass components, additional trends may appear. However, such granularity should not be necessary for this study. Every entry in the database also included the application and grant date for each patent. In this research the grant date was used to compile the total number of patents per month from January of 1996 to March of 2013. However, in generating the forecasting models only the data from January 1996 to December 2011 was used. This allowed for a portion of the actual data to be used in comparison to the proposed forecast values.

For each patent group two models will be applied, HWES and ARIMA. Two functions within R Studio were used to generate the models for each class of patents: *HoltWinters()* and *auto.arima()*. Each series was plotted and 15 month forecasts for the two models were produced. The forecast values were then compared to the actual values previously

withheld and forecast error metrics were calculated. A third Simple Exponential Smoothing (SES) forecast will be applied and graphed for purposes of providing visual comparison. However, SES models in their most basic form tend to over fit the data and may not be the best option. Furthermore, as it has been stated, the actual selection of a forecasting method is not the objective of this paper. It is the hope of this research to identify possible candidates for future patent mining/technology forecasting research.

In this paper, we make an assumption that the classifications proposed by USPTO are correct. It may be argued that other meaningful patents related to a given technology are classified elsewhere. For instance, Wu et al. (2010) suggest that most industries rely on the International Patent Classification (IPC) process too heavily. This can sometimes make searching for specific patents within a classification difficult, decrease business decision processes, and increase the possibility of patent infringement. It may be possible to cluster patents with similar content to create less arbitrary classifications. From these groupings themes could be determined and trend analysis analogous to this research could be carried out. One proposed approach is to cluster the patents using Genetic Algorithms and Support Vector Clustering (Wu et al. 2010).

4. Experimental Results

R Studio was used in this project to compile, plot, and forecast each time series trend. The first step in the process was to graph each series. Figure 1 illustrates the time series graphs of all three groupings. From each of these graphs it can be observed that there is an observable trend. Additionally it should be noted that by themselves, none of the models are stationary, which is a requirement for the ARIMA model. However, R implements ARIMA in such a manner that the level of differencing is determined automatically.

4.1 Exponential Smoothing

For each dataset both the HoltWinters and auto.arima functions were used to fit appropriate models. The smoothing parameters and Sum of Squares values for each HWES model are shown in Table 2. The alpha values were automatically generated by R and indicate how close the model will fit the actual data. The parameter can range in values from zero to one. If the value is close to one then the resulting model is influenced more by the later values of the data. However, all of the values in Table 2 indicate that both recent and less recent data points were used in creating the forecast. The coefficient value represents the final component estimate.

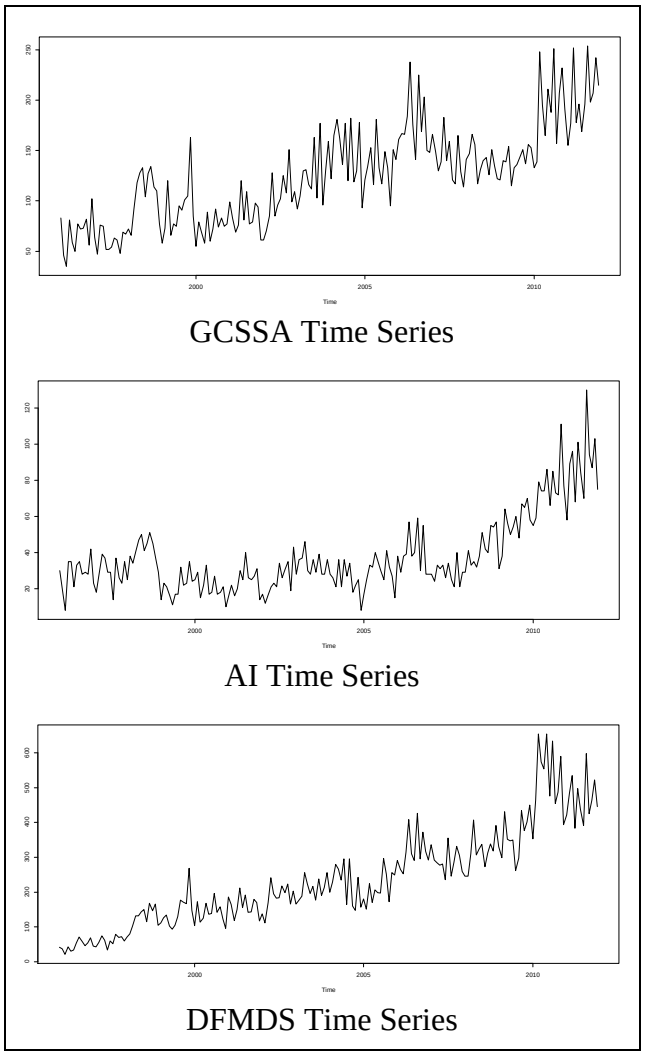

Figure 1 – Patent Time Series

Name	Smoothing Alpha	Coefficients	SSE
GCSSA	0.277	215.64	135767.9
AI	0.3	89.52	21146.18
DFMDS	0.338	472.32	515876.8

Table 2 – HW Exponential Smoothing Model Values

The trend lines generated from the HWES model appear to fit each instance very well. In fact it may be argued that they are over fitting each data series. However, for the purposes of this research such a similarity is acceptable since this study is primarily concerned with determining if modeling such data is possible to begin with. Another feature to note is that in the forecast of each HW model, the trend seems to become flat. According to Hyndman and Athanasopoulos (2013) empirical evidence suggests that Exponential Smoothing methods tend to over-forecast. To compensate for this, a technique known as damping is applied which creates a flattened forecasting line. Figures 2 through 7 show forecast for each patent group projected 15

months out for a SES and HWES model. The SES plots are being included to illustrate the predictive potential that other Exponential Smoothing models offer. Although due to the error correction options it offers, HWES will continue to be the primary model of demonstration for this paper.

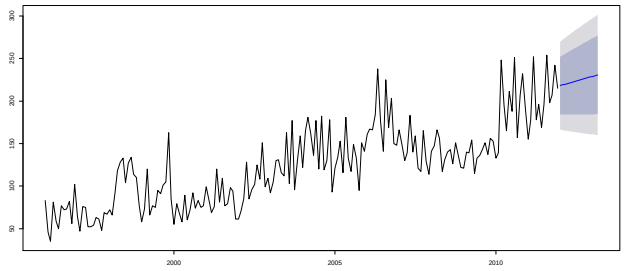

Figure 2 – SES Model and 15 month forecast for GCSSA

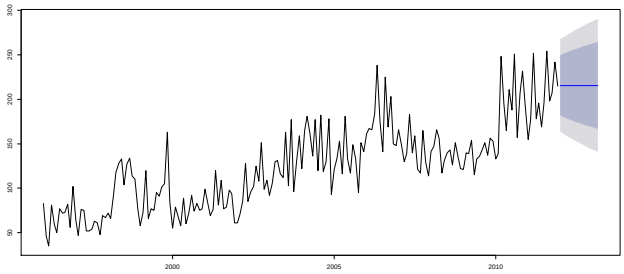

Figure 3 – HW Model and 15 month forecast for GCSSA

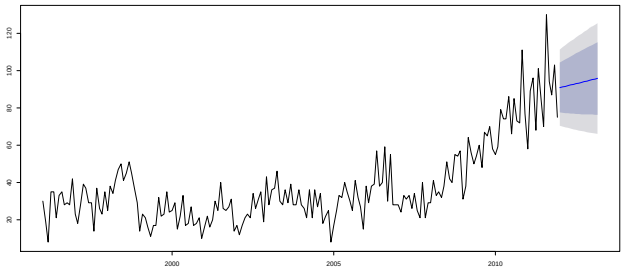

Figure 4 – SES Model and 15 month forecast for AI

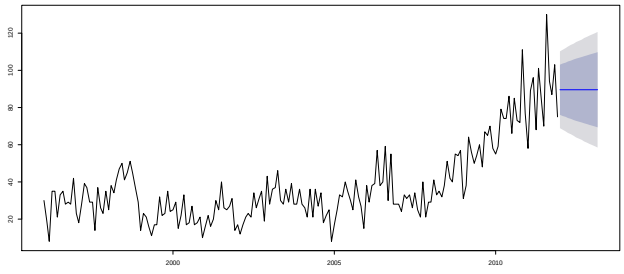

Figure 5 – HW Model and 15 month forecast for AI

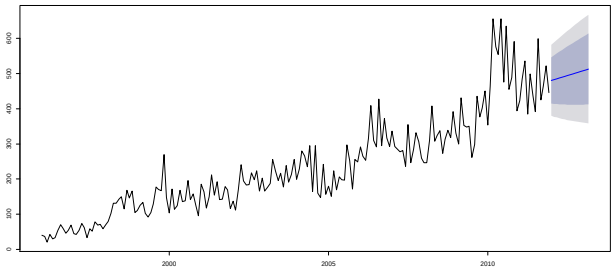

Figure 6 – SES Model and 15 month forecast for DFMDS

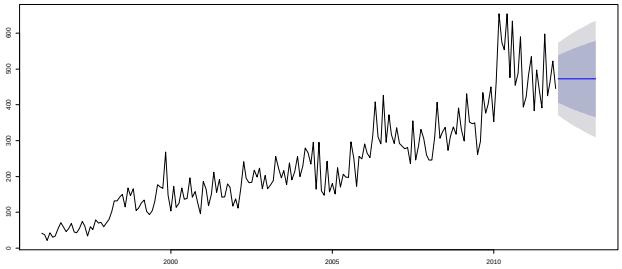

Figure 7 – HW Model and 15 month forecast for DFMDS

4.2 ARIMA

The ARIMA model has three parameters *(p, d, q)* and is often written as *arima(p, d, q)*. The Autoregressive (AR) portion of the model is based on the idea that the current value of the series, x_t, can be explained as a function of p past values, x_{t-1}, x_{t-2},...,x_{t-p}, where p determines the number of steps into the past needed to forecast the current value (Shumway and Stoffer 2006). The parameter of d represents the levels of differencing the original time series needs to undergo to become stationary. As an alternative to the autoregressive representation in which the x_t on the left-hand side of the equation are assumed to be combined linearly, the moving average model of order q, abbreviated as MA(q), assumes the white noise w_t on the right-hand side of the defining equation are combined linearly to form the observed data (Shumway and Stoffer 2006). Therefore, in the ARIMA model q represents the number of lags in the moving average.

Normally the creation of an ARIMA model requires determining the level of differencing necessary to make a time series stationary. Thankfully R has a function (*auto.arima*) that accomplishes this task in one step. It may be worthwhile to note that the middle term of each proposed ARIMA model is 1. This corresponds with the level of differencing that is needed to make each time series stationary. The model parameters for each patent group are shown in Table 3. As with the HWES and SES examples, the forecasts for each patent group were projected out 15 months and the results are shown in Figure 8.

Name	ARIMA Model	σ^2	AIC	BIC
GCSSA	(2, 1, 1)	687.5	1798.6	1811.6
AI	(2, 1, 0)	100.2	1431.1	1444.1
DFMDS	(1, 1, 3)	2407	2040.3	2056.5

Table 3 – ARIMA Model Parameters

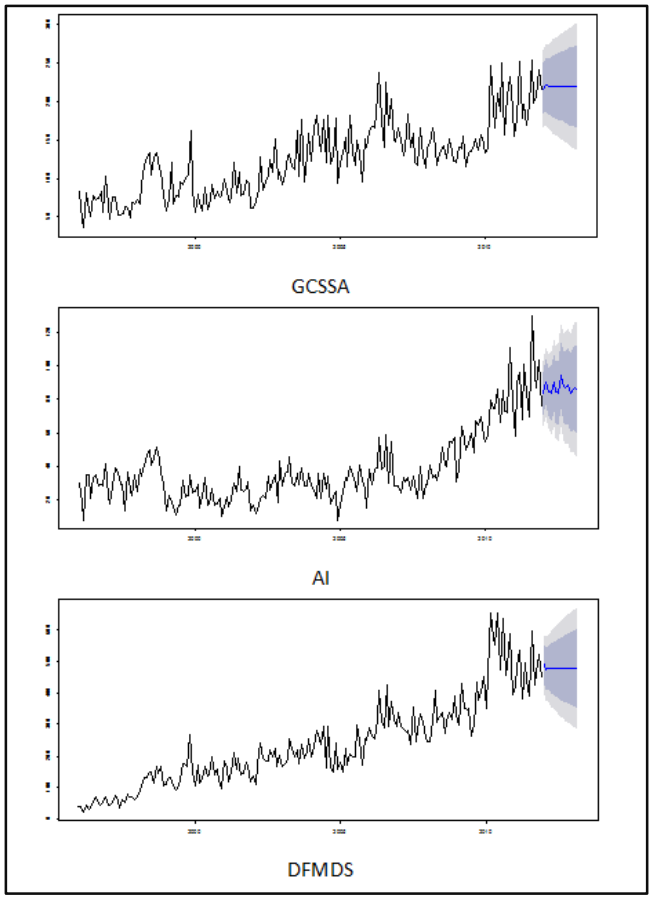

Figure 8 – ARIMA 15 Month

4.3 Model Comparison

In the early stages of time series modeling the selection of models was very subjective. Since then, many techniques and methods have been suggested to add mathematical rigor to the search process of an ARMA model, including Akaike's information criterion (AIC), Akaike's final prediction error (FPE), and the Bayes information criterion (BIC). Often these criteria come down to minimizing (in-sample) one step-ahead forecast errors, with a penalty term for over fitting (De Gooijer and Hyndman 2006). It should be noted that these model comparison techniques are only useful for selecting the best model of similar structure. For instance if there are three ARIMA models on one dataset to choose from, AIC or BIC can be used to select from those models. It is for this reason that measures of forecast accuracy like MAE, MAPE, and MASE are used to compare models of different structures.

For each model and 15 month forecast, four error statistics were calculated: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Mean Absolute Scaled Error (MASE). The results are shown in Table 4. All of these values used the 15 months not included in the original model training data as testing data. For each error calculation lower values are preferred. According to Hyndman and Koehler (2006), values of MASE greater than one indicate that the forecasts are worse, on average, than in-sample one-step forecasts from naıve (random-walk) methods. Based on this measurement, it can be seen that the MASE values indicate that all of the models have adequate forecasting capabilities.

The results from Table 4 suggest that ARIMA acts as a better predictor for the GCSSA and DFMDS data while the AI patent data seems to be better suited for an Exponential Smoothing model. Given the forecasting results, it does not seem reasonable to state that a specific time series model is best for these three patent groupings. For additional reference the full list of testing and forecasting values are listed in the appendices at the end of this paper.

4.4 Discussion

At a first glance it appears that the models generated may be over fitting the data. However, the MASE values calculated indicate that each of the models produced performs very well in predicting the testing data. It is possible that both are true. From looking at the trend lines produced, they do seem to be very similar to the actual trends. Moreover, the testing data may not have been fully representative of the full flow of each trend. In future research a different proportion of training and testing data should be considered.

Another interesting observation from the experimentation is that the Database and Control System patent groups favored an ARIMA model, while Artificial Intelligence patents fit better with a Holt Winters model. A possible explanation for this is an intuitive look at the initial time series for each classification group. In the AI trend the data seems to be fairly stationary until about 2008, when the number of patents seemed to spike rapidly. Thus it appears that not much differencing would be needed on this model and this may automatically make it a better candidate for a HWES model.

Patent Group	Model	RMSE	MAE	MAPE	MASE
GCSSA	HWES	42.52	31.23	12.11	0.6436
	ARIMA	40.66	29.66	11.62	0.6142
AI	HWES	11.61	8.03	7.84	0.731
	ARIMA	13.08	9.64	9.46	0.7906
DFMDS	HWES	85.08	65.57	11.45	0.6754
	ARIMA	80.4	60.98	10.64	0.6351

Table 4 – Model Forecast Error Statistics

5. Conclusions and Future Work

The first goal of this paper was to demonstrate that current groups of patents could be represented as a time series. From observing the initial plots it appears that this certainly is the case. An interesting observation that can be made is the consistent increase in these technology based patents over the past 20 years. The second objective of this research was to confirm that time series models could be applied to each patent group. This too was successful. Obviously it is debatable as to whether the models presented are the most optimal for the situations provided. However, it seems safe to state that with additional work patent and technology forecasting models could be produced using time series modeling techniques.

Future work would benefit from exploring the validity of the groupings of patents. A possible approach would be to use textual mining techniques to first group the patents and then conduct an analysis similar to the one carried out in this paper. It may also be worthwhile to explore multivariate autoregression techniques such as Vector Autoregression or Bayesian Vector Autoregression. As mentioned earlier in the paper, there may be associations between patent groupings that might influence the rate of change in another. Furthermore, if the patent classifications are not a good enough representation of a technological theme, then both a re-clustering of patents and a multivariate analysis may be necessary.

References

Conejo, A. J.; Plazas, M. A.; Espinola, R.; Molina, A. B. 2005. Day-Ahead Electricity Price Forecasting Using the Wavelet Transform and ARIMA models. *IEEE Transactions on Power Systems*, 20(2):1035-1042

Contreras, J.; Espinola, R.; Nogales, F. J.; Conejo, A. J. 2003. ARIMA Models to Predict Next-Day Electricity Prices. *IEEE Transactions on Power Systems*, 18(3):1014-1020

Daim, T.U.; Rueda, G.; Martin, H.; Gerdsri, P. 2006. Forecasting Emerging Technologies: Use of Bibliometrics and Patent Analysis. *Technological Forecasting & Social Change* 73:981-1012

Dongdong, W. 2010. The Consumer Price Index Forecast Based on ARIMA Model. *In Proceedings of the 2010 WASE International Conference on Information Engineering (ICIE)* 307-310

Fried, R.; George, A.C. 2014. Exponential and Holt-Winters Smoothing, *International Encyclopedia of Statistical Science*, Springer Berlin Heidelberg

Gelper, S.; Fried, R.; Croux, C. 2010. Robust Forecasting with Exponential and Holt–Winters Smoothing. *Journal of Forecasting* 29:285-300

Goodwin, P. 2010. The Holt-Winters Approach to Exponential Smoothing: 50 Years Old and Going Strong. *Foresight* 19:30-33

Jun, S.; Park, S.S.; Jang, D.S. 2011. Technology Forecasting Using Matrix Mapping and Patent Clustering, *Industrial Management & Data Systems* 112(5):786-807

De Gooijer, J.G.; Hyndman, R.J. 2006. 25 Years of Time Series Forecasting, *International Journal of Forecasting* 22:443–473

Kang, H. 1986. Univariate ARIMA Forecasts of Defined Variables. *Journal of Business & Economic Statistics* 4(1):81-86

Kumar, U.; De Ridder, K. 2010. GARCH Modelling in Association with FFT-ARIMA to Forecast Ozone Episodes. *Atmospheric Environment* 44(34):4252-4265

Leixiao, L.; Zhiqiang, M.; Limin, L.; Yuhong, F. 2013. Hadoop-based ARIMA Algorithm and its Application in Weather Forecast. *International Journal of Database Theory & Application* 6(5):119-132

Hyndman, R.J.; Athanasopoulos, G. 2013. *Forecasting: principles and practice* http://otexts.org/fpp/

Hyndman, R. J.; Koehler, A.B. 2006. Another Look at Measures of Forecast Accuracy. *International Journal of Forecasting* 22:679–688

Shin, J.; Park, Y. 2009. Brownian Agent Based Technology Forecasting. *Technological Forecasting & Social Change* 76:1078-1091

Shumway, R.H.; Stoffer, D.S. 2006. *Time Series Analysis and Its Applications*. Springer, New York

Timmermann, A.; Granger, C.W.J. 2004. Efficient Market Hypothesis and Forecasting. *International Journal of Forecasting* 20(1):15-27

Tseng, Y.-H.; Lin, C.-J.; Lin, Y.-I. 2007. Text Mining Techniques for Patent Analysis. *Information Processing and Management* 43:1216-1247

Wu, C.H.; Ken, Y.; Huang, T. 2010. Patent Classification System Using a New Hybrid Genetic Algorithm Support Vector Machine. Applied Soft Computing 10:1164-1177

Appendices

A1 – GCSSA Testing/Forecast Data

Point	Actual	HW Forecast	ARIMA Forecast
Jan 2012	243	215.6	216.6
Feb 2012	196	215.6	221.9
Mar 2012	179	215.6	219.9
Apr 2012	229	215.6	219.4
May 2012	304	215.6	220.0
Jun 2012	210	215.6	219.9
Jul 2012	288	215.6	219.8
Aug 2012	235	215.6	219.8
Sep 2012	235	215.6	219.9
Oct 2012	312	215.6	219.8
Nov 2012	230	215.6	219.8
Dec 2012	213	215.6	219.8
Jan 2013	224	215.6	219.8
Feb 2013	232	215.6	219.8
Mar 2013	244	215.6	219.8

A2 – AI Testing/Forecast Data

Point	Actual	HW Forecast	ARIMA Forecast
Jan 2012	92	89.5	83.2
Feb 2012	83	89.5	90.4
Mar 2012	83	89.5	84.7
Apr 2012	101	89.5	83.6
May 2012	126	89.5	89.7
Jun 2012	88	89.5	84.4
Jul 2012	102	89.5	83.8
Aug 2012	95	89.5	94.0
Sep 2012	99	89.5	87.4
Oct 2012	90	89.5	86.5
Nov 2012	98	89.5	88.1
Dec 2012	85	89.5	83.9
Jan 2013	97	89.5	86.2
Feb 2013	96	89.5	86.5
Mar 2013	89	89.5	85.2

A3 – DFMDS Testing/Forecast Data

Point	Actual	HW Forecast	ARIMA Forecast
Jan 2012	580	472.3	488.9
Feb 2012	486	472.3	475.9
Mar 2012	563	472.3	478.8
Apr 2012	493	472.3	476.8
May 2012	610	472.3	478.2
Jun 2012	501	472.3	477.2
Jul 2012	632	472.3	477.9
Aug 2012	516	472.3	477.4
Sep 2012	513	472.3	477.8
Oct 2012	643	472.3	477.5
Nov 2012	503	472.3	477.7
Dec 2012	472	472.3	477.6
Jan 2013	430	472.3	477.7
Feb 2013	558	472.3	477.6
Mar 2013	483	472.3	477.6