City University London

MSc in Business Systems Analysis and Design

INM363 Individual Project Report

September 2014

"Evaluating predictive analytics methods for comparing emerging and developed markets"

Germanos Pappas

Supervised by: Dr Cagatay Turkay

25th of September 2014

By submitting this work, I declare that this work is entirely my own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the assessment instructions and any other relevant programme and module documentation. In submitting this work I acknowledge that I have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. I also acknowledge that this work will be subject to a variety of checks for academic misconduct.

Signed:

Abstract

This project applies predictive analytics algorithms to financial time series, with the aim to compare a) emerging and developed markets and b) the different predictive algorithms and methods used. It applies the three following algorithms, a) Linear Regression, b) Holt's Exponential smoothing and c) ARIMA, to datasets from suitable financial indicators from the World Bank, and compares their forecasts. The characteristics of the financial time series are investigated and a general analysis framework is proposed as a methodology for analysing financial time series, depending on these characteristics. In addition, from the comparison of the forecasts, general conclusions are drawn regarding the future trends of emerging markets and the suitability, reliability, advantages and limitations of each predictive algorithm.

Keywords: Predictive Analytics, Linear Regression, Holt's Exponential Smoothing, ARIMA, Emerging Markets

Table of Contents

Chapter 1 – Introduction and Objectives

1.1. Introduction

The last years the global economy is being revolutionised, as significant changes take place. Countries that traditionally were the foundation of global economy have suffered from recession, while new markets emerge and start to know significant growth. Many big investments are transferred from United States, United Kingdom and Europe to China, India or Brazil. As Karen Ward from HSBC has noted in *The World in 2050* "By 2050, the emerging world will have increased five-fold and will be larger than the developed world…19 of the top 30 economies by GDP will be countries that we currently describe as emerging" (Ward, 2011).

This is a hypothesis that could change the way businesses work in the future and is worth to be tested. Predictive analytics is a valuable tool in the effort to investigate the validity of that hypothesis. Predictive analytics encompasses a variety of statistical models that are able to create predictions about future values of a variable, by taking into account current and historical data for this variable. It can be used in many different fields, such as marketing, banking, insurance or manufacturing industry, but especially in the financial sector it could provide analysts with an excellent tool for enhancing economic decision making. A major question for every analyst, though, is which predictive algorithms are the most appropriate to use for financial predictions.

This project selects and explores three specific algorithms that could be used in predictive analytics with the aim a) to investigate the validity of the hypothesis that emerging markets will dominate in the future economy and b) to analyse, compare and identify advantages, reliability and limitations of predictive algorithms that could be used in financial predictions.

The beneficiaries of this research project are mainly Investment Banks and companies with activities in the banking and finance industry, Insurance companies, Financial and Business Analysts, as well as Traders and Investors who will be able to improve their decision making processes. Furthermore, other fields that use forecasts and analytics, such as marketing or retail industry, could be benefited from the general conclusions of this research.

The main reason I chose the specific project, except that it would be for me a new and exciting field, was also that the knowledge and skills I would gain through my research, would help me meet my career aspirations in the Financial Analytics and Business Intelligence sector.

1.2. Objectives of the project

For the successful completion and evaluation of the research project, by applying the SMART technique (Dawson, 2005, p. 59), the following specific, measurable, appropriate and realistic objectives were set:

- Identify suitable financial indicators from emerging and developed markets to use in the predictive analysis

- Identify predictive algorithms suitable for financial analysis
- Apply at least 3 predictive algorithms with the statistical computing software R Project to the selected financial data from emerging and developed markets
- Compare the predictions of the different algorithms
- Identify the reliability of algorithms for financial predictions
- Present their predictions with suitable visualization techniques and make recommendations for appropriate techniques of presentation
- Test the validity of the hypothesis that emerging markets will dominate in the future global economy

The suitability and reliability of the chosen algorithms will be tested by comparing their results with real observed values. Specifically, the algorithms will be applied in a smaller dataset than the available and the forecasts made will be compared with the values of the remaining dataset that were not used in the analysis (hold out sample). In addition, the comparison of the forecasts between the algorithms would give valuable indication about their reliability and accuracy. The hypothesis that the emerging markets will play a dominant role in the future global economy will be tested by comparing the slopes of the forecast lines for the selected indicators, as these slopes are representative of the rate of growth for a country.

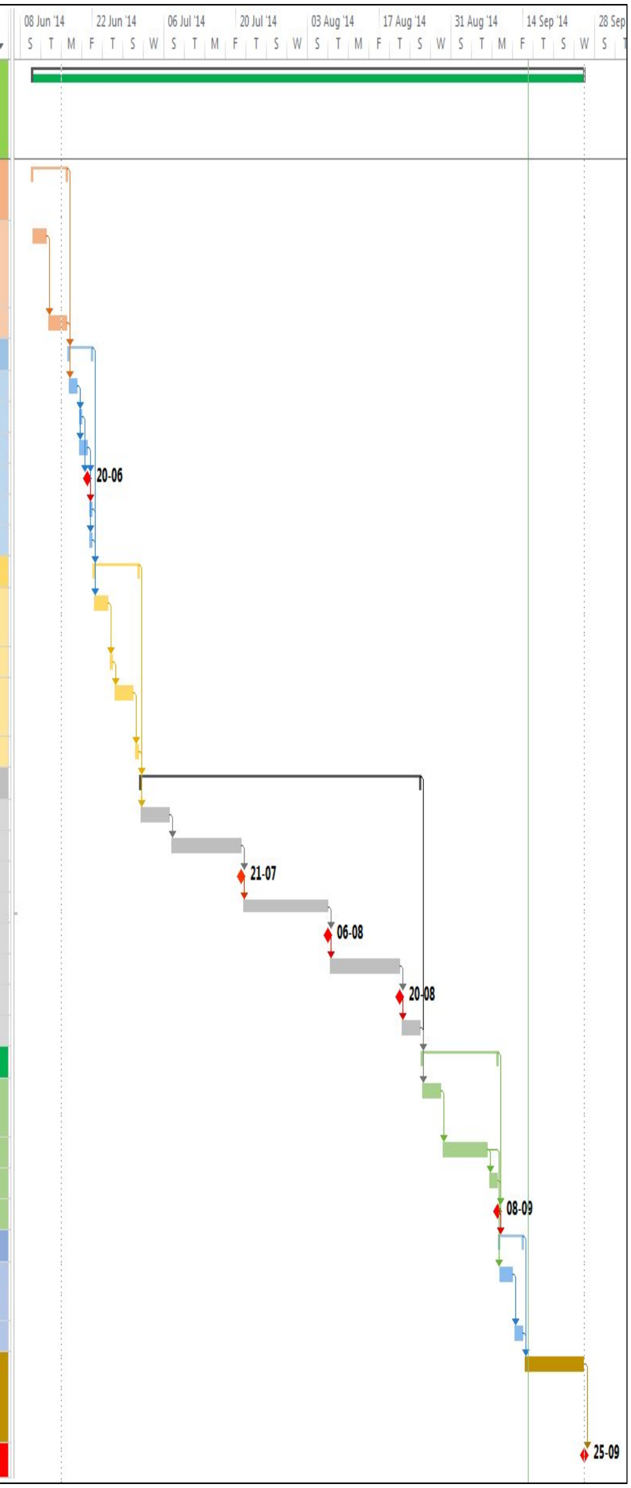
1.3. Work Plan

The real work plan of the project can be seen in the Figure 1.3.1. In general, the project consisted of the tasks that were estimated in the project plan of the proposal, which can be found in Appendix A (p.104), with only exception one major change.

In the early stages of the project, it was decided that the project should focus more on the predictive algorithms and devote more time on the analysis and comparison of results and less on the visualization techniques. More specifically, it was decided the analysis to include more indicators and more countries, in order to obtain more results that would enable the generalization and justification of the conclusions regarding the reliability and accuracy of algorithms. Due to the limitations of time, this meant that the time that could be spent on visualization techniques should be decreased. In fact, the visualization techniques were restricted only to ones applied with R during the comparison of the results process. Apart from that, a literature review for visualization methods took place and recommendations for future work were also done.

The methods used are described in detail in Chapter 3. In brief, for the data gathering was used data from open source website, the analysis made with the statistical software environment R and the comparisons were made with graphical representation and quantitative analysis of the results.

	Task Name	Duration	Start	Finish	Predecessors
1	Project: Evaluating predictive analytics methods for comparing emerging and developed markets	108 days	Tue 10-06-14	Thu 25-09-14	
2	1. General Introductory Literature Review	7 days	Tue 10-06-14	Mon 16-06-14	
3	Search for academic papers, articles and books in the area of Predictive Analytics and Finance	3 days	Tue 10-06-14	Thu 12-06-14	
4	Reading	4 days	Fri 13-06-14	Mon 16-06-14	3
5	2. Data Gathering	5 days	Tue 17-06-14	Sat 21-06-14	4,2
6	Evaluate and select sources of data	2 days	Tue 17-06-14	Wed 18-06-14	4
7	Decide which countries to compare	1 day	Thu 19-06-14	Thu 19-06-14	6
8	Decide which financial indicators to use	2 days	Thu 19-06-14	Fri 20-06-14	6
9	Milestone1: Meeting with Supervisor	0 days	Fri 20-06-14	Fri 20-06-14	8,7
10	Download and store the data	1 day	Sat 21-06-14	Sat 21-06-14	9
11	Summarize and write report	1 day	Sat 21-06-14	Sat 21-06-14	8
12	3. Selection of algorithms	9 days	Sun 22-06-14	Mon 30-06-14	5,10,11
13	Literature review (specific for algorithms) and preselection of algorithms	3 days	Sun 22-06-14	Tue 24-06-14	10,11
14	Selection of algorithms to be tested	1 day	Wed 25-06-14	Wed 25-06-14	13
15	Reading, Practice and deep understanding of selected algorithms	4 days	Thu 26-06-14	Sun 29-06-14	14
16	Write report	1 day	Mon 30-06-14	Mon 30-06-14	15
17	4. Analysis of data with statistical tool R	55 days	Tue 01-07-14	Sun 24-08-14	12,16
18	Read documentation and practice R tool	6 days	Tue 01-07-14	Sun 06-07-14	16
19	Predictive analysis with Algorithm 1	14 days	Mon 07-07-14	Sun 20-07-14	18
20	Milestone2: Meeting with Supervisor	0 days	Mon 21-07-14	Mon 21-07-14	19
21	Predictive analysis with Algorithm 2	17 days	Mon 21-07-14	Wed 06-08-14	20
22	Milestone3: Meeting with Supervisor	0 days	Wed 06-08-14	Wed 06-08-14	21
23	Predictive analysis with Algorithm 3	14 days	Thu 07-08-14	Wed 20-08-14	22
24	Milestone4: Meeting with Supervisor	0 days	Wed 20-08-14	Wed 20-08-14	23
25	Review, Conclusions and report writing	4 days	Thu 21-08-14	Sun 24-08-14	24
26	6. Project Conclusions	15 days	Mon 25-08-14	Mon 08-09-14	25,17
27	Comparison of emerging and developed markets	4 days	Mon 25-08-14	Thu 28-08-14	25
28	Comparison of algorithms	9 days	Fri 29-08-14	Sat 06-09-14	27
29	Summarize Conclusions	2 days	Sun 07-09-14	Mon 08-09-14	28
30	Milestone5: Meeting with Supervisor	0 days	Mon 08-09-14	Mon 08-09-14	29
31	5. Visualization techniques	5 days	Tue 09-09-14	Sat 13-09-14	30,26
32	Literature review (specific for visualization)	3 days	Tue 09-09-14	Thu 11-09-14	28
33	Review, Conclusions and report writing	2 days	Fri 12-09-14	Sat 13-09-14	32
34	7. Review of the whole Project, last amendments, final report writing and appendices	12 days	Sun 14-09-14	Thu 25-09-14	33,31
35	Submission of the Project	0 days	Thu 25-09-14	Thu 25-09-14	34



[Figure 1.3.1] – Real Work Plan

1.4. General analysis framework

The experimentation with the algorithms and the time series during the analysis stage had as an outcome the construction of a general analysis framework which, as it was proved to be effective, is given below as a suggested methodology for analysing financial time series:

1) Firstly, the time series with the real values is plotted.
2) From the form of the plot, specific characteristics of the time series, such as trend, seasonality or stationarity are identified.
3) Depending on these identified characteristics, the algorithms and the methods that will be used for the analysis can be selected. For instance, if the time series exhibit constant trend through time then Linear Regression can be applied. Otherwise, the time series shall be divided in smaller time periods that have relatively constant trend before Linear Regression analysis is applied. Similarly, if the time series is not stationary, it has to be differenced until it becomes stationary before ARIMA is applied.
4) After the appropriate algorithms are selected, they shall be firstly cross-validated with the hold out method. In this method, the dataset is split into two subsets, the algorithm is applied to the one of them and then, the forecasts that are made are compared with the values of the other subset that was not used in the analysis (hold out sample) in order to identify their accuracy.
5) Then the analysis for the whole dataset can be conducted for the most reliable algorithms.
6) Each analysis can be evaluated by examining the residual values (difference between fitted and real values) of the analysis.
7) Next, the forecasts from the analysis of the different algorithms and methods are compared in the same graph.
8) Finally, the forecasts should be evaluated and updated continuously with the future real values when these they will become available.

1.5. Structure of the Project Report

The rest of the project report continues in Chapter 2 with a description of the context in which the project is taking place and how it fits in this environment. Chapter 3 describes in detail the methods that were used for the successful completion of the project. In Chapter 4 the analysis process is explained and examples of the results of the analysis are quoted. The comparison of the results and a discussion regarding the conclusions that come from this comparison are included in Chapter 5. Finally, Chapter 6 contains an evaluation of the project as a whole, a summary of the conclusions, suggestions for future work and a reflections section.

The appendices include the project proposal (Appendix A), the data used for the analysis (Appendix B), all the graphs from the analysis (Appendix C), the graphs from the comparison of the results (Appendix D) and lastly one full example of the code that was implemented in R for every analysis (Appendix E).

Chapter 2 – Context

Predictive analytics algorithms have been used for many years in a lot of different fields in order to identify patterns and trends in time series and perform predictions for their future values. They have been applied widely in marketing and sales, healthcare and medical industry and of course in economics and the financial sector. The simplest model is Regression Analysis and the Method of Least Squares that originates from the nineteenth century and the works of the great mathematicians Legendre, Gauss and Galton. Since then, numerous different and more sophisticated algorithms have been deployed in an effort to forecast accurately the future values of a variable using existing data with major progress taking place the last 50 years.

In 1957 Holt presented a method of using exponentially weighted moving averages in order to forecast trends in seasonal data (Holt, 2004). A few years later, Winters used the same technique to predict seasonal sales (Winters, 1960). From these two works, the Holt-Winters Exponential Smoothing methodology was formed and created new possibilities for the predictive algorithms. At the same time, Kalman (1960) developed the Kalman Filter, an algorithm that developed significantly the accuracy of predictions. Added to that, the Box and Jenkins methodology that was published in 1970 (Box and Jenkins, 1970) had a great influence in time series analysis and forecasting and led to the widely used ARIMA methodology. From that point, many algorithms were implemented with the aim always to improve the accuracy of forecasts. A detailed historical overview of the algorithms and methods developed is given by the International Journal of Forecasting (De Gooijer and Hyndman, 2006).

Despite, though, the fact that a big number of predictive models have been developed in the last years, predictive analytics have not yet fully explored and are used much less than expected in academic research. A report from MIS Quarterly (Shmueli & Koppius, 2011) quotes that "We address a particularly large gap, namely, the near-absence of predictive analytics in mainstream empirical IS research". In this report, the two researchers made an investigation in the articles that were published in two major journals, the MIS Quarterly and the Information Systems Research. Their research revealed that although 252 articles seemed to be relevant to predictive analytics and forecasting techniques, in reality only 7 of them had really employed predictive analytics methods (Shmueli & Koppius, 2011, pp. 560–561).

This is even more emphasized in the financial sector where the majority of the work in predictive models, according to Mills (2008), has been "undertaken by financial professional and journalists rather than by academics. Indeed, this seems to have become a long-standing tradition, as, even today, much empirical research and development still originates from the financial industry itself". This is also supported by a research in Scopus and Google Scholar for academic papers on predictive analytics, which does not return significant results. The few papers that are available are mainly in the fields of sales, healthcare or retail industry. In the financial sector, the most recent significant works is a paper from Dablemont, Van Bellegem and Verleysen in forecasting High and Low of financial time series with the use of Kalman Filters and an overview of Kalman filters in mathematical finance by Date (2009).

It is, therefore, obvious that despite the big number of algorithms and methodologies that have been developed, there is limited academic research on how these algorithms could be used effectively for forecasting in financial time series. In addition, there is no clear scientific comparison between the algorithms that would reveal the advantages and limitations of each method. Even on the overview of algorithms by De Gooijer and Hyndman (2006), the focus is the historical evolution and not on the comparison of the methods. Similarly, Tsay (2005) although he provides extended information for the majority of the algorithms used in financial time series, does not attempt a comparison. As a result, the question how a financial analyst can choose the most appropriate algorithm from the total number of predictive analytics methods remains unanswered. A structured methodology that would enable an analyst to make the right choice seems to be necessary.

Moreover, little work has been done on the direction of combining different predictions algorithms in order to achieve better forecasts. A recent and excellent work has been done in this field by Firmino, de Mattos Neto, and Ferreira (2013), in which a methodology is presented for correcting and improving predictions by combining different predictive methods, but still a lot of research has to be done.

Finally, although it plays a very significant role in enhancing the effectiveness of the model and enabling the more efficient decision making process, the way that the forecasts should be communicated to the financial analyst, is often overlooked. Keim, Mansmann, Schneidewind, Thomas and Ziegler (2008) point out the novel analysis capabilities that visual analytics can offer in forecast analysis, by integrating interactive visualization methods, decision making strategies and scientific statistical models. They also question whether the current techniques "are able to meet the demands of the ever-increasing mass of information" and they emphasize the need for research in this direction.

Some of the most notable works in this field the last years are mentioned below. Javed, McDonnel and Elmqvist (2010) give an overview of different line graph techniques, such as simple line graph, braided graph, small multiples and horizon graphs, that can be used for the concurrent representation of multiple time series and present the advantages and limitations of each technique. Savikhin, Maciejewski and Ebert (2008) developed a visual analytics application for improving economic decision making, while Ichikawa, Tsunawaki, Fujishiro, and Yoon (2002) created a visualization technique for multiple daytime stock price predictions.

In addition, Buono et al (2007) created the TimeSearcher, an interactive time series visualization tool that enables analysts to change parameters and see the corresponding forecast. The latter is a very significant area that needs to be explored, as until now, as is quoted on the above paper, "little attention has been given to the visualization and the user interaction", despite the fact that it can be a powerful tool for analysts.

Lastly, a very significant work has been done recently by Zhuang, Small and Feng (2014) in the analysis of developed financial markets with visibility graphs, extending the work previously done by Lacasa, Luque, Ballesteros, Luque and Nuño (2008) in that field. Visibility graphs is a visualization technique that transforms a time series into a network that inherits the properties of the initial time series. By examining the generated network it is possible to draw quantifiable conclusions about the time series that would have been much more difficult to be drawn by examining only the plot of the time series. This method, therefore, enables analysts to make decision easier and quicker, as the networks are much simpler to analyse.

Taking into consideration all of the above, this project tries to fill some of the existing gaps in predictive analytics and wishes to draw the attention of the research community into this field. In order to do so, it explores three different predictive algorithms, compares their results and identifies advantages and limitations of the specific methods. Having identified the importance of emerging markets in the future global economy (King, 2011; Ward, 2011), it applies the algorithms in financial time series from emerging and developed markets and makes forecasts that examine the role of emerging markets in future economy. Moreover, it attempts to present a structured methodology for selecting and applying the suitable method for analysing financial time series and making forecasts. Finally, this project would wish to motivate further research in visual analytics and visualization techniques, as they can provide valuable tools for effective forecasting in predictive analytics.

Chapter 3 – Methods

3.1. Data gathering

As was suggested in the project proposal, the data gathering method that was chosen was the use of documents and more specifically the use of open data from a free access website. This choice was based on the fact that there are several highly reputable websites that provide free access to big amounts of data about development in countries from all over the world. This method had the advantage that offered quick and free access to plenty of data relevant to the project objective with no need to use other time consuming methods of data gathering. In addition all data are anonymous and can be used immediately in the analysis.

The website that was chosen as the source of the data was the http://data.worldbank.org. The World Bank is a highly reputable United Nations international financial institution that definitely satisfies the criteria set by Oates (2006, p. 241–242) for the evaluation of documents during data gathering. The purpose of the data being gathered and stored on the site is similar to the project objective, as the World Bank's aim is to store and share information and knowledge about development in countries globally. In addition, according to the organisation's Terms of Use (The World Bank, 2013) it promotes the public access and free use of these data for creative research.

The organisation selects the data and the World Development Indicators that are stored on its site, with a very meticulous process. The development professionals of the World Bank Group follow general principles for evaluating the quality of statistical data, using except from their experience and expertise, also the checklist of the Netherlands' Central Bureau of Statistics (Statistics Netherlands, 2009). In detail, the data that is stored in the site must be relevant to the development rates of countries, free accessible, offer an adequate coverage globally and over time, produced by a reliable source with regular updates and finally to allow comparison to be made between different years and countries. All these are important factors that justify the selection of the World Bank Open Data Website as the source of this project and minimize as possible the concerns regarding the reliability and suitability of the data.

From the total number of the World Development Indicators there were selected the four, that are considered the more relevant and representative for the rhythm of development for a country. These are a) GDP per capita, b) Foreign direct investment, net inflows, c) Export value index and d) Import value index. These indicators are defined from The World Bank as follow: GDP per capita is the gross domestic product divided by midyear population in current U.S. dollars. Foreign direct investment are the net inflows of investment to acquire a lasting management interest (10 percent or more of voting stock) in an enterprise operating in an economy other than that of the investor in current U.S. dollars. Export and import value indexes are the current value of exports (f.o.b.) and imports (c.i.f.) converted to U.S. dollars and expressed as a percentage of the average for the base period (2000). More information about these indicators, including their source and periodicity, can be found on the links that are provided in the table of Appendix B.1. (p. 118). In the same table the date that the data were last updated on the site, as well as the downloading date (26/07/2014) are noted.

For each Indicator there are data that span the last 30-50 years and for more than 250 different countries, economies or groupings of countries or economies. This range of data is considered adequate for the purpose of this specific project. For every one of the four selected Indicators, there was made a selection of countries and economies that, according to King (2011), are considered representative for the emerging and developed countries. From the emerging markets, there were selected China, Brazil, India, Russian Federation, Philippines and wherever that was possible the groupings of Middle East and North Africa, Latin America and Caribbean and East Asia and Pacific. From the developed countries the selection comprises the United Kingdom, the United States and the EURO zone or Germany and France wherever the grouping for EURO zone was not provided. All the data that were extracted from the World Bank web page can be found in Appendix B.2. (p. 119).

3.2. Data analysis

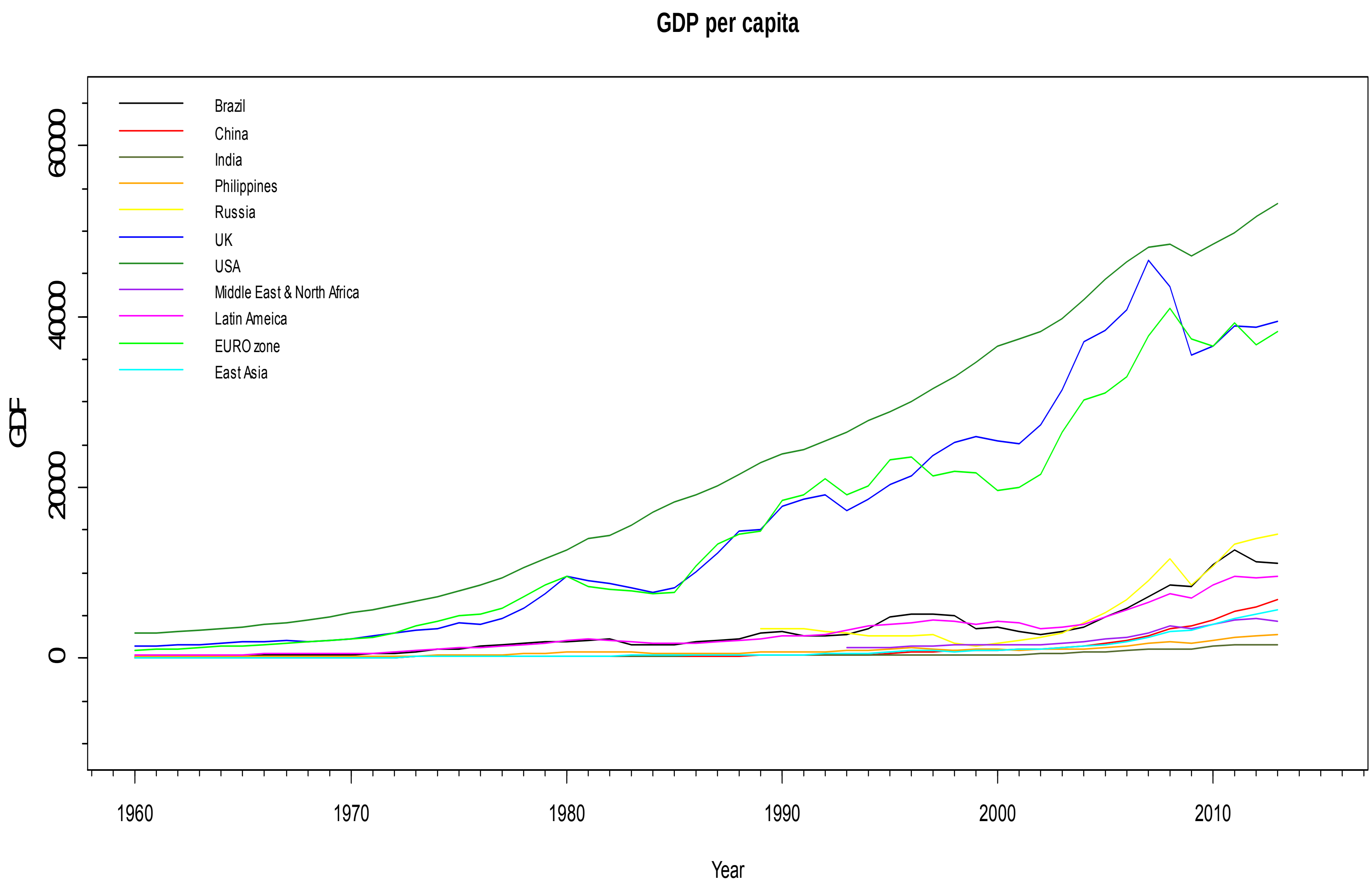
3.2.1. Software Environment for the analysis

The above data that were extracted from the World Bank Open Data web page would be analysed quantitatively with statistical models and suitable predictive algorithms in order to make predictions about the future values of the selected indicators. The analysis was chosen to be done with the open source statistical tool R Project. R is a free software tool for statistical computing and graphical presentations that can be downloaded from http://www.r-project.org/. It uses an interpreted computer language with a graphical run-time environment. The language is simple and effective and is derived from the S (Becker, Chambers & Wilks) and Scheme (Sussman) languages. Initially was created by R. Ihaka and R. Gentleman but it has since been developed with the contribution of a large group of people as an open source software.

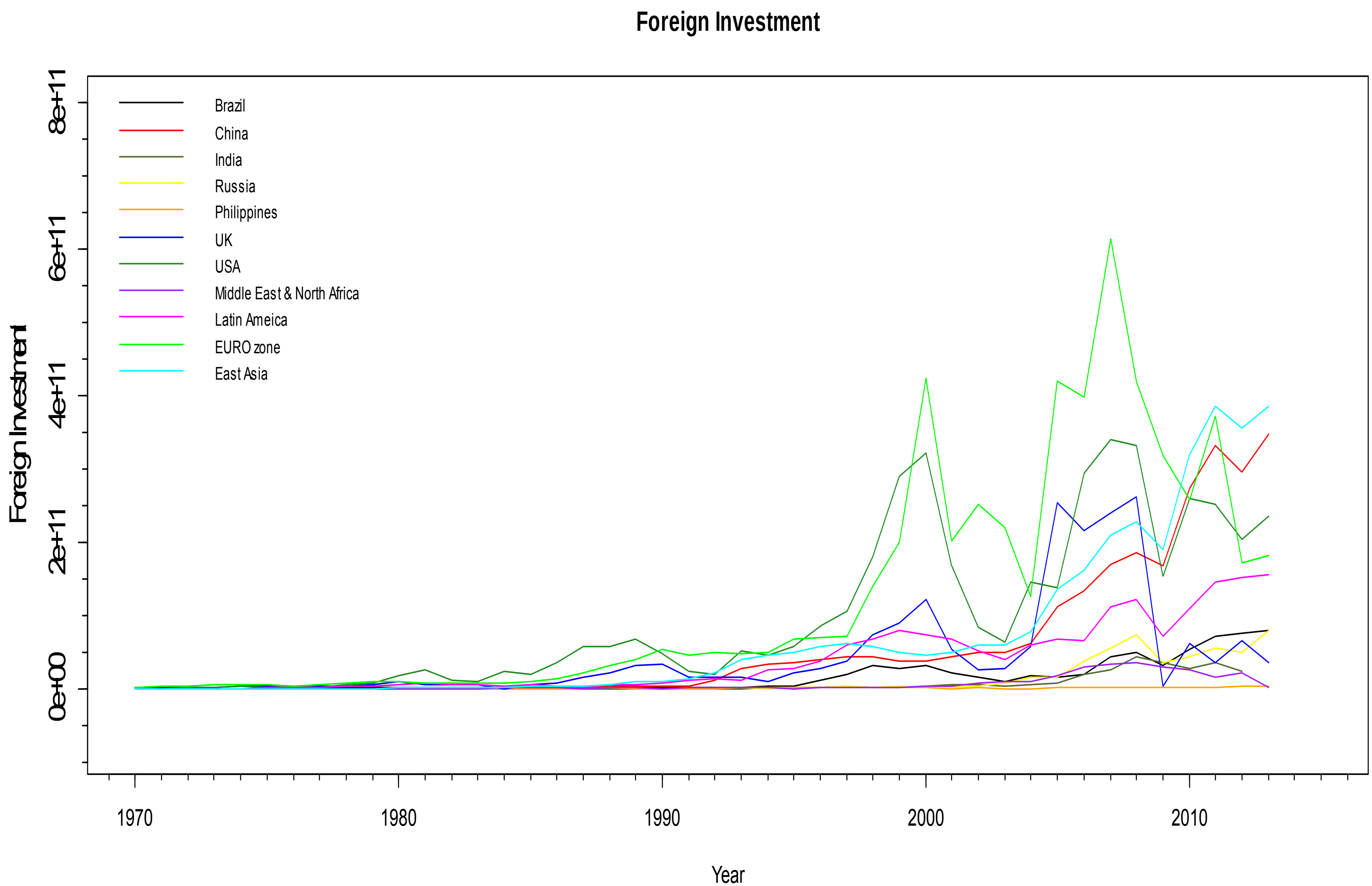
The documentation that is provided on the above web page offers the basis for the learning of the R language. Especially, the manuals *An introduction to R* (Venables, Smith, & Team, 2002) and *R Data Import/Export* (R Core Team, 2000) were a very good introduction for the first stages of the project. In addition, the book *Introductory statistics with R* (Dalgaard, 2008) provided valuable guidance. Finally, significant help on practical issues and many examples can be found on the web page *Time Series Analysis and Its Applications: With R Examples* and the *R-bloggers* blog which is a community for R users.

3.2.2. Time Series theory and selection of algorithms process

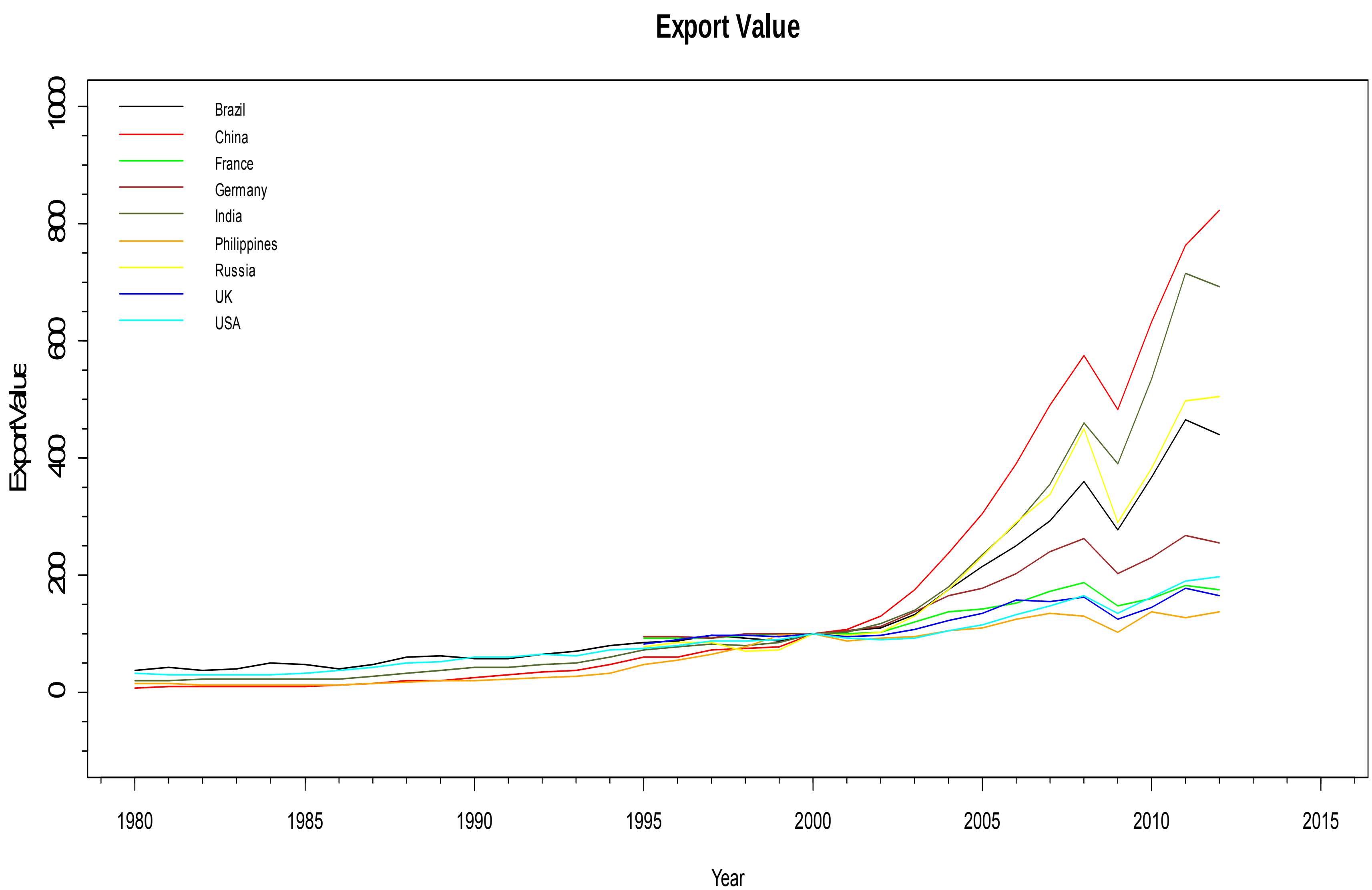
The data, as extracted from the World Bank, are univariate Time Series with one dependent variable Y (the value of the indicator) and only one independent variable, the time t. The first step in the analysis process is to plot with R all the time series for each indicator as can be seen in Figures 3.2.1 to 3.2.4.



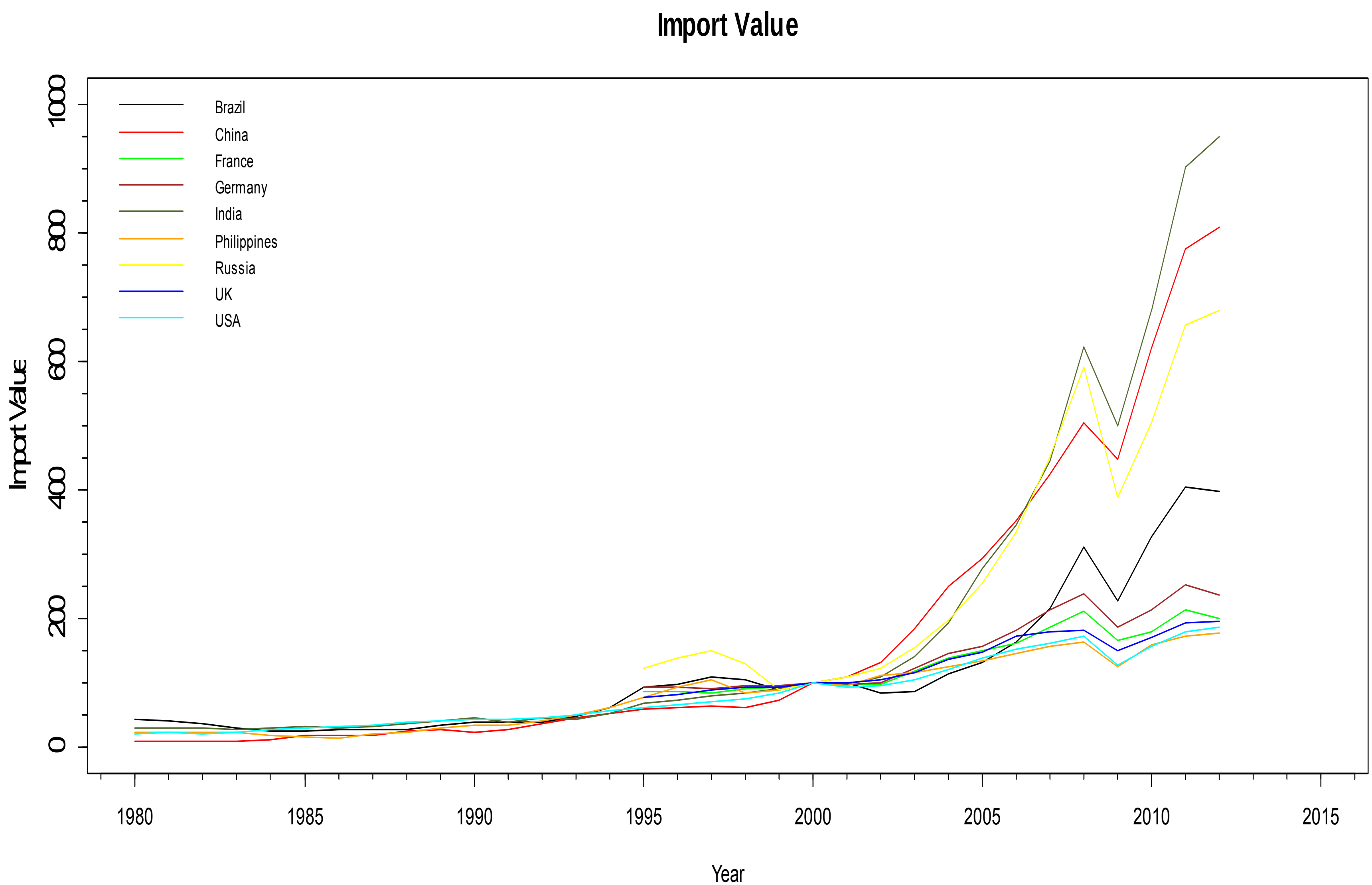
[Figure 3.2.1] – GDP per capita indicator for each country



[Figure 3.2.2] – Foreign Investment indicator for each country



[Figure 3.2.3] – Export Value indicator for each country



[Figure 3.2.4] – Import Value indicator for each country

Most time series consist of two basic components, trend and seasonality, and a random noise (error) (StatSoft, 2013). The predictive algorithms aim to filter out the noise and identify the pattern that is formed by the trend and seasonality. Once the pattern is identified, it is possible to extrapolate it in future and make forecasts for future values of the time series. The World Development Indicators have annual periodicity, which means that for each indicator and for every country only one value is recorded for each year. As a consequence the data have no seasonality and by removing the error the trend can be identified.

After plotting the time series and identifying their basic characteristics, the next step of the project included the selection of the predictive algorithms that would be used in the analysis. From the numerous existing predictive algorithms, there should be chosen those that are most suitable to apply in financial data and time series, considering the time limitations that an MSc project imposes. It was estimated that the limited time would allow the exploration of maximum three algorithms. For this selection to be successful an extended literature review was conducted. Tsay (2005) gives an excellent overview of algorithms that are used in the analysis of financial time series and it was the base of the whole algorithms' selection process. StatSoft (2013) and Lind, Marchal, Wathen, and Waite (2009) give also plenty of useful material for prediction algorithms. Other important academic works regarding forecasting in time series are given by Date (2009), Holt (2004) and Kalekar (2004).

Based on the literature review the first algorithm that was selected for the analysis was the Linear regression model, as it is the most simple algorithm that is used for the identification of trend in time series. The other two, more sophisticated, techniques that were chosen was Holt's Exponential Smoothing and ARIMA model. The following sections describe the reasons for this choice, summarize the basic theory of these algorithms, their advantages and limitations and provide the details of the analysis that was conducted with each one.

3.2.3. Linear regression

Theoretical Background

The Linear Regression represents the relationship between the dependent variable Y (the value of the indicator) and the independent variable t (the time) with a straight line and is expressed by the equation:

$$Y = a + b*t$$

where a is a constant, also referred to as the *intercept*

and b is the slope of the line, also referred to as the *regression coefficient*

The slope of the line b gives the average change in the value of variable Y for a change of one unit in t, in other words the average change of value of the indicator in every year.

The slope and intercept of Linear Regression equation can be calculated with the Least Squares Method and are given as follow:

$$b = \frac{n\Sigma tY - (\Sigma Y)(\Sigma t)}{n\Sigma t^2 - (\Sigma t)^2}$$

and

$$a = \frac{\Sigma Y}{n} - b\left(\frac{\Sigma t}{n}\right)$$

Since the parameters a and b have been estimated, the formula Y = a + b*t can be used to forecast the values Y for any future time t.

The Linear Regression is a simple technique that can represent well time series that have a constant trend over a period of time and do not present significant noise, as these time series can be represented adequately by a line. In practice, though, it has limitations as the time series do not always exhibit constant trend over long periods of time and moreover usually contain significant noise.

Analysis process

In order to overcome the first limitation, it was attempted to split the datasets and perform analysis in smaller subsets that have constant trend. The question with this way is about the reliability of the forecasts of very small subsets and how we could identify the ideal subset of data. For this, the forecasts of the different datasets should be compared with each other and the fitted values of each analysis should be evaluated against the real values by calculating the residuals.

In addition, in order to evaluate the effectiveness of the method as a whole, a cross-validation analysis was conducted in a subset of data consisting of values until the year 2008. The forecasts that were made with this subset were compared with the real values for the years 2009-2013. The selection of the year 2008 was made for a number of reasons. First of all, it is proved from practice that the predictive algorithms can give accurate predictions only for a few years and for longer periods of time can only estimate trends. In that sense, the range 2009-2013 seems reasonable to be used in the cross-validation. A bigger set would not probably add value to the validation, as it would be expected for more distant points in time the forecasts to be inaccurate.

Apart from that, though, considering that the algorithms would be applied in financial indicators, the selection of year 2008 as the border for the two subsets is even more justified from the fact that the financial crisis in the developed markets started around that year. It is challenging to investigate how accurate were the predictions that the algorithms gave for that period.

To summarize, for each indicator and every country the following Linear Regression analyses were performed:

1) analysis for the whole dataset
2) analysis for the dataset up to year 2008
3) analysis for the dataset up to year 2000
4) analysis for the dataset from year 2001 to year 2013
5) analysis for the dataset from year 2006 to year 2013

Due to the big number of different cases that had to be studied, for each indicator there was made a further selection of four countries to be used in the analysis. So, for every indicator there were chosen two countries representing the emerging markets and two countries representing the developed markets. Thus, the total number of Linear Regression analyses was:

4 Indicators * 4 Countries * 5 different datasets = 80 analyses

The whole analysis lasted two weeks from the 7^{th} July 2014 to the 20^{th} July 2014. Detailed guidance on how to apply the Linear Regression analysis in R is given in the tutorial *Using R for Linear Regression.* The basic function in R for Linear Regression is lm(), which fits a line to the time series curve and provides the slope and intercept for that line. Then the fitted and residuals values can easily be calculated by the programme, where fitted values are the values for the estimated fitted line to the time series and residual values the difference between real and fitted values.

The effectiveness of the proposed linear model can be evaluated by examining the residuals. The residuals should be normally distributed with an average mean value around zero. R offers four graphical representation techniques to examine the residuals and the effectiveness of a linear model by using the plot() command, as these are shown in the Results section (Chapter 4).

Finally, forecasts for the next years can be done by applying the command predict() on the lm() model that has been calculated before. A sample of the code that was developed for the Linear Regression analysis with explanatory comments is shown in Appendix E.2. (p. 327). The graphs for two specific examples and more details about how the method was applied are given in the Results section (Chapter 4) and all the rest graphs for the Linear Regression analyses that were conducted are included in the Appendix C.1. (p.133).

In order to overcome the second limitation of the Linear Regression Method there should be applied more sophisticated techniques that filter out the random noise by smoothing the curve of the time series. For this, first the Holt's Exponential Smoothing method was explored as described in the next section.

3.2.4. Holt's Exponential smoothing

Theoretical Background

The Exponential smoothing, according to Kalekar (2004), is a technique that continuously revises a forecast by taking into account the new observed values. It applies a weighting to the real observed values of a time series in a way that the older values are assigned exponentially smaller weights and provides a smoothed curve for the time series. For a time series that has no trend (the values fluctuate around a stable mean) and no seasonality the Simple Exponential smoothing can be applied with the following form:

$$S_t = \alpha * X_t + (1 - \alpha) * S_{t-1}$$

In this way each smoothed value depends on the previous observations with weights that decrease exponentially according to the value of α. If α is equal to zero then the previous observations are completely ignored and the smoothed value is equal to the last observed value. If α is equal to 1 then the last observation is ignored completely and the older observations have more weight on the smoothed value. In practice the value of α is between 0 and 1 producing intermediate results.

When the time series has no seasonality but exhibit a trend, then the Holt's Exponential smoothing model has to be applied. In this case except α, there is also a second parameter b to take into account the trend as well. The relevant form is:

$$S_t = \alpha * y_t + (1 - \alpha) * (S_{t-1} + b_{t-1}) \quad \text{where } 0<\alpha<1,\ 0<b<1$$

and $$b_t = \gamma * (S_t - S_{t-1}) + (1 - \gamma) * b_{t-1} \quad \text{where } 0<\gamma<1$$

It is suggested that the initial value of S_t to be equal to y_1 and the initial value of b_t to be set to $b_1 = y_2 - y_1$, where y_t the real observed values. When the values for the parameters α and b have been set, the smoothed values of the time series can be estimated. Then forecasts for future values can be made based on these previously estimated smoothed values.

In the case where the time series has both trend and seasonality there must used the Holt-Winters Exponential Smoothing where a third parameter gamma is also applied.

As has already been mentioned the time series of this project have trend but not seasonality, so the Holt's Exponential smoothing was the appropriate model to be applied.

Analysis process

To maintain consistency with the analysis that was done with Linear Regression and in order to enable the comparison of the algorithms, the same analysis process was followed with the Holt's Exponential Smoothing model (except from the analysis for the dataset up to year 2000, as it was considered that did not add value to Holt's Exponential smoothing analysis). Similarly, for each indicator and every country the following Holt's Exponential Smoothing analyses were performed:

1) analysis for the whole dataset
2) analysis for the dataset up to year 2008
3) analysis for the dataset from year 2001 to year 2013
4) analysis for the dataset from year 2006 to year 2013

Again, the same countries were selected for each indicator and were used in the analysis. The total number of the Holt's Exponential Smoothing analyses was:

4 Indicators * 4 Countries * 4 different datasets = 64 analyses

The analysis for this algorithm lasted a little more than two weeks, from the 21th July 2014 to the 6th August 2014. Valuable guidance and examples on how to apply the Holt's Exponential Smoothing model in R was found in the web page *Using R for Time Series Analysis* (2010). The basic function for Holt's Exponential Smoothing in R is HoltWinters(), where the parameter gamma of the function should be set to FALSE (the gamma parameter should be used only for time series with seasonality). With this function, R estimates the parameters α and b, the fitted values of the model and the Sum of Squared Error (SSE) which shows how well the model fits the real values and indicates the effectiveness of the model.

Forecasts for the next years can be done by applying the function forecast.HoltWinters() on the model that has been calculated before with the HoltWinters() function. Apart from the forecasts, the 80% and the 95% prediction intervals are calculated as well. The effectiveness of the model can be evaluated by calculating the in-sample forecast errors and checking whether they present non-zero autocorrelations at lags 1-20. This is done by using the acf() and the Box.test() functions. Finally, we can also, evaluate the suggested model by checking if the forecast errors are normally distributed and have a mean zero. This can be done by plotting the errors and making a histogram of their distribution.

A sample of the code that was developed for the Holt's Exponential Smoothing modeling with explanatory comments is shown in Appendix E.3. (p. 332). The graphs for a specific example and more details about how the method was applied are given in the Results section (Chapter 4) and the rest graphs for the Holt's Exponential Smoothing analyses that were conducted are included in the Appendix C.2. (p. 161).

After the completion of the analysis for Holt's Exponential smoothing, the project continued with the exploration and application of the third and final algorithm, the ARIMA model, as described in the next section.

3.2.5. ARIMA

Theoretical Background

The Autoregressive Integrated Moving Average (ARIMA) methodology provides a more complex and sophisticated model for predictions, as it takes into consideration correlations between successive values of the dataset. ARIMA models include a statistical model that takes into account the random error (irregular component of a time series) and allows for non-zero autocorrelations.

ARIMA can only be applied to stationary time series, that is, time series with a mean, variance and autocorrelation that do not change significant over time. If the time series is not stationary it has to be differenced (d) times until it becomes stationary. The model has three parameters and is referred as ARIMA(p,d,q) where (p) is the autoregressive parameter, (d) is the number of times the time series must be differenced until it becomes stationary and (q) the moving average parameter.

StatSoft (2013) and Tsay (2005) describe in detail the process of identification of the parameters (p) and (q). In practice, this is done by examining the autocorrelogram (ACF) and partial autocorrelogram (PACF) of the stationary time series and checking after how many lags these are equal to zero. This is not a straightforward process and requires experience and good understanding and experimentation with the models. After identifying the parameters (p) and (q), and having first identified the number of differencing (d) the model can now be applied. It is worth to note that the best model could be any of the following ARIMA(p, d, q) or ARIMA(0 d, q) or ARIMA(p, d, 0) models, as in the literature it is widely suggested to apply the principle of parsimony in the selection of the model (StatSoft, 2013; Using R for Time Series Analysis, 2010). This means that the model should be as simple as possible and contain the fewest parameters.

Analysis process

To maintain consistency with the analysis that was done with Linear Regression and Holt's Exponential smoothing, and in order to enable the comparison of the algorithms, the same analysis process was followed with the ARIMA model (except from the analysis for the dataset up to year 2000, as it was considered that did not add value to ARIMA analysis). Similarly, for each indicator and every country the following ARIMA analyses were performed:

1) analysis for the whole dataset
2) analysis for the dataset up to year 2008
3) analysis for the dataset from year 2001 to year 2013
4) analysis for the dataset from year 2006 to year 2013

Again, the same countries were selected for each indicator and were used in the analysis. The total number of the Holt's Exponential Smoothing analyses was:

4 Indicators * 4 Countries * 4 different datasets = 64 analyses

The analysis for this algorithm lasted two weeks, from the 7th August 2014 to the 21th August 2014. Again, valuable guidance and examples on how to apply ARIMA models in R was found in the web page *Using R for Time Series Analysis* (2010). The first step in R is to make the time series stationary with the function diff() by selecting the necessary number of differencing passes d. In order to identify the other two parameters p and q, the autocorrelogram (ACF) and partial autocorrelogram (PACF) of the stationary time series are plotted with the functions acf() and pacf() respectively. After identifying all parameters the model is applied to the time series with the function arima() in which the values of parameters are specified with the argument 'order=c(p,d,q)'.

R provides also the possibility to use the function auto.arima(), which automatically estimates the values for the three parameters (p,d,q). During the analysis for this project the identification of the p, d, q values was made primarily with the ACF and PACF plots methodology. The auto.arima() function was only used complementary to this identification process in order to evaluate its results.

Forecasts for the next years was done by applying the function forecast.Arima() on the model that has been calculated before with the arima() function. Apart from the forecasts, the 80% and the 95% prediction intervals are calculated as well. As in the Holt's Exponential smoothing the effectiveness of the model can be evaluated by calculating the in-sample forecast errors and checking whether they present non-zero autocorrelations at lags 1-20. This can be done again with the acf() and the Box.test() functions. Finally, we can also, evaluate the ARIMA model by checking whether the forecast errors are normally distributed with a mean zero. This is done by plotting the errors and making a histogram of their distribution.

A sample of the code that was developed for the ARIMA modeling with explanatory comments is shown in Appendix E.4. (p.337). The graphs for a specific example and more details about how the method was applied are given in the Results section (Chapter 4) and all the graphs for the 64 ARIMA analyses that were conducted are included in the Appendix C.3. (p. 217).

3.3. Comparison of algorithm and results

After the completion of the analysis for all three algorithms, the next stage of the project was the comparison of the algorithms and their results. The comparison was conducted from the 25th of August until 8th of September and had two aims. The first aim was to compare the results for emerging and developed markets and identify their future trends. The second was to compare the results that the different algorithms gave for the same country, in an effort to identify differences, advantages, suitability, reliability and limitations of each method that could be generalized and lead to a structure methodology for making forecasts with financial time series.

For better visualization of comparisons and in order to enable the extraction of conclusions, the next groups of comparison diagrams were plotted with R:

1) For each indicator, the forecasts were plotted from the three algorithms (using the whole dataset) for all countries in the same graph (4 diagrams)
2) For each indicator, the forecasts were plotted from the three algorithms (using the dataset from year 2001 to year 2013) for all countries in the same graph (4 diagrams)
3) For each indicator, the coefficients of slope from Linear Regression were plotted for all countries on the same graph
4) For each country and each indicator, the forecasts were plotted for all algorithms (using the dataset until 2008) and the real values from 2009 to 2013 in the same graph (16 diagrams)
5) For each country and each indicator, the forecasts were plotted for all algorithms and all datasets in the same graph (16 diagrams)

From these graphs, quantitative analysis was conducted and statistical tables were constructed that calculate the effectiveness of each algorithm and each dataset. In addition, tables and graphs were constructed with the slope coefficients of Linear Regression that enable the comparison of the rate of development between countries.

More details and the outcome of these comparisons are discussed in Chapter 5 (p.66) and all graphs can be found in Appendix D (p.290).

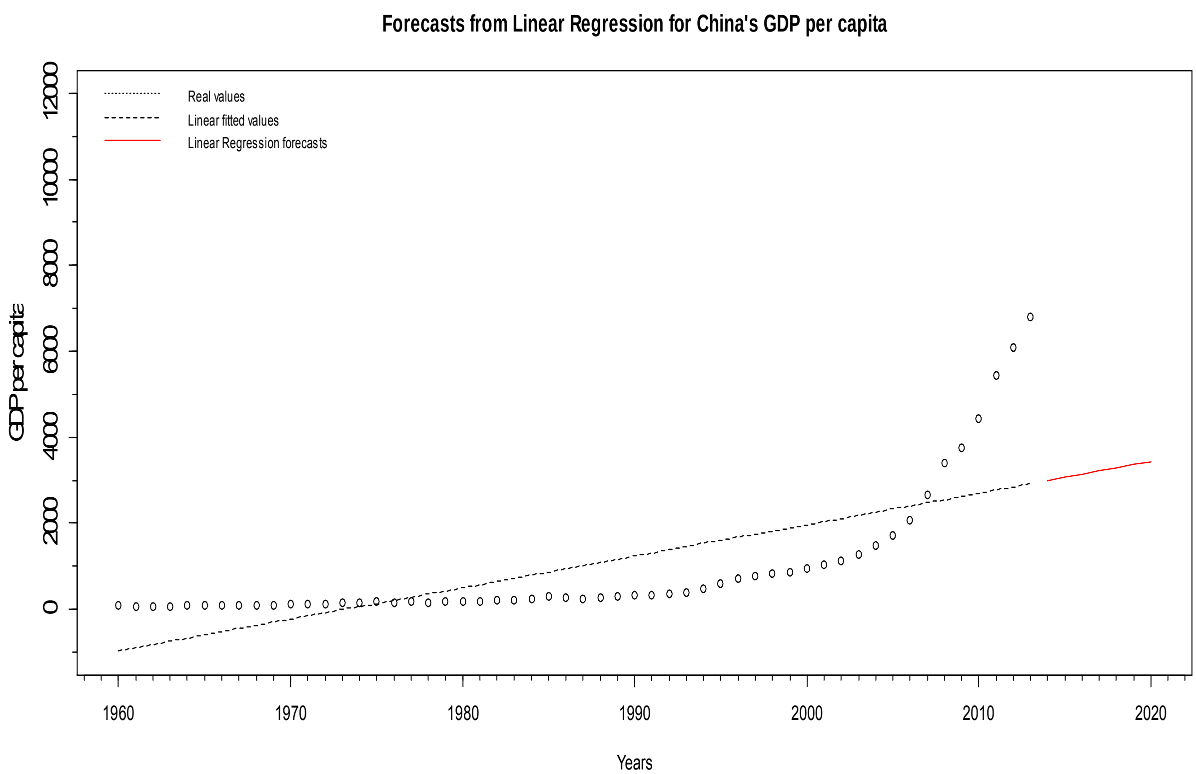
Chapter 4 – Results

This Chapter comprises the outputs of the analysis for each algorithm separately. For each algorithm is presented the complete analysis (for all dataset cases) for the GDP per capita indicator for two countries, China from emerging markets and UK from the developed. Details and comments for the analysis process and the evaluation of each method are also provided along with the graphs for each case. The graphs for the rest indicators and countries are provided in the Appendix C (p.133).

4.1. Linear Regression

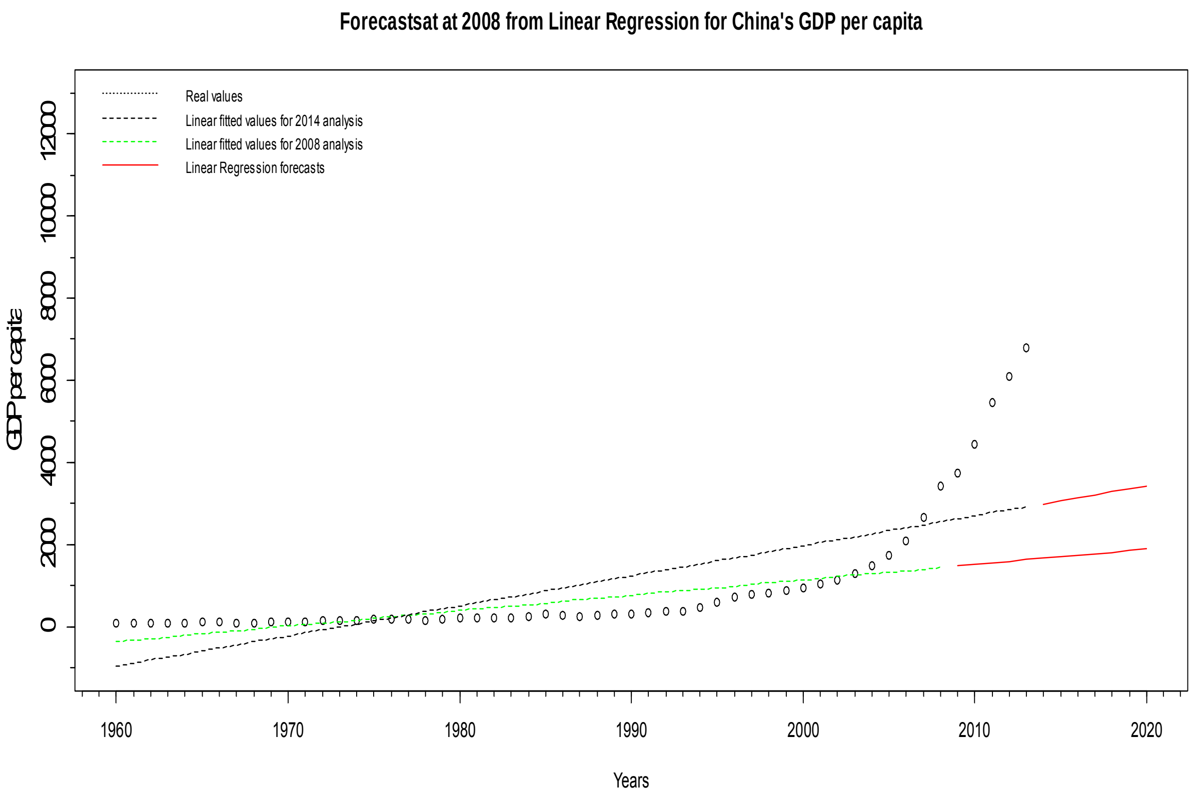
4.1.1. Analysis for China's GDP per capita

The first stage was to apply a Linear Regression analysis to the whole dataset of China values for the GDP per capita Indicator, in this case to the China values from year 1960 to year 2013. The code for this analysis can be found in Appendix E.2.1. (p. 327). The results can be seen in the Figure 4.1.1, where the real values (points), the fitted line (black dashed line) and the forecasts (red line) are plotted.



[Figure 4.1.1] – Analysis for China, GDP per capita and whole dataset

The next phase was to apply the algorithm to the dataset up to year 2008 and compare the predictions for the years 2009 to 2013 with the real values. The relevant code is in Appendix E.2.2. (p. 328) and the results in Figure 4.1.2., where the green dashed line gives the fitted line for the dataset up to 2008. In essence, this is the fitted line and the forecasts that somebody would get if he conducted the analysis in 2008. The plot allows the comparison of these forecasts with the real values (points in the graph) and the fitted line with the whole dataset which is also plotted (black line).



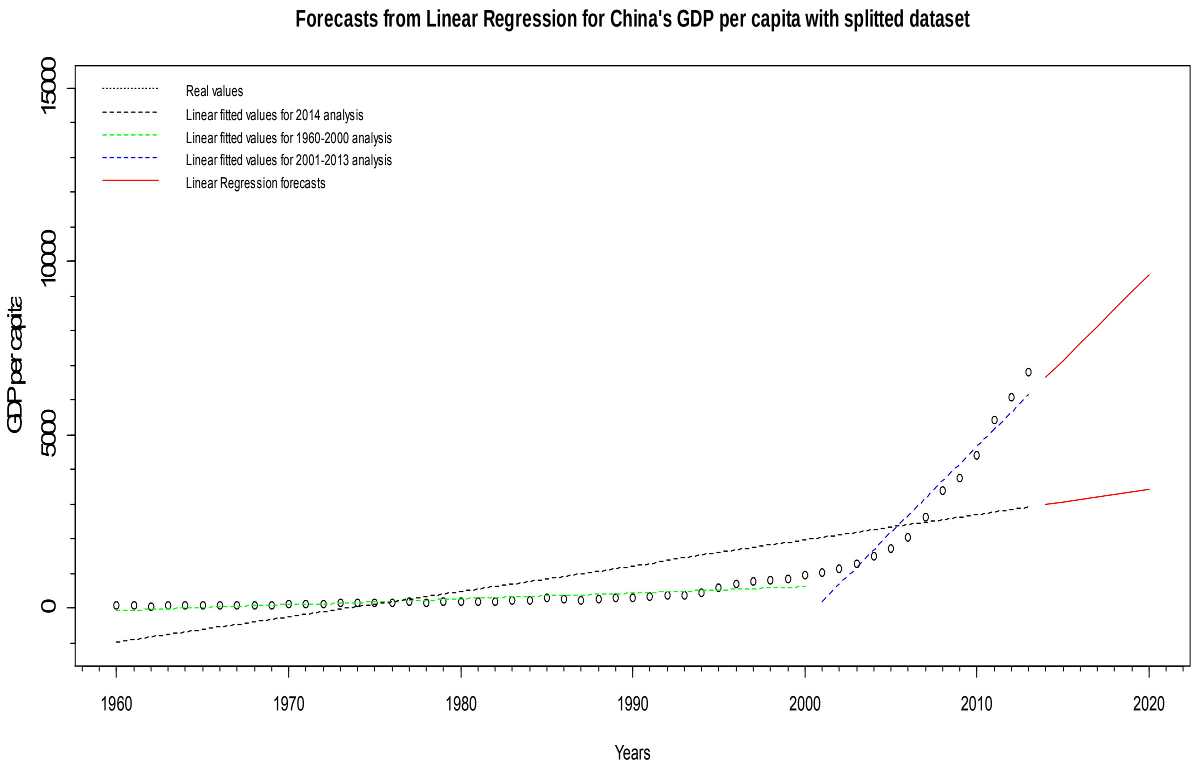
[Figure 4.1.2] – Analysis for China, GDP per capita and the dataset up to 2008

Although the green line fits well to the real values until 2004, after that point the real values start to increase and diverge from the predictions made in 2008. The graph shows that the big increase after 2005 in the values of China's GDP per capita would not have been predicted by an analysis with the whole dataset made in 2008.

Since this inefficiency was detected, the next step was to divide the dataset into two smaller datasets that seem to have constant trend, one until year 2000 and the other from year 2001 to year 2013. So, two more analyses were done, one for each subset of data. The code for these analyses is in the Appendices E.2.3. (p.329) and E.2.4. (p.329). The results can be seen in Figure 4.1.3, where the green line is the fitted line for the analysis until year 2000 and the blue line the fitted line for the analysis with subset 2001-2013.

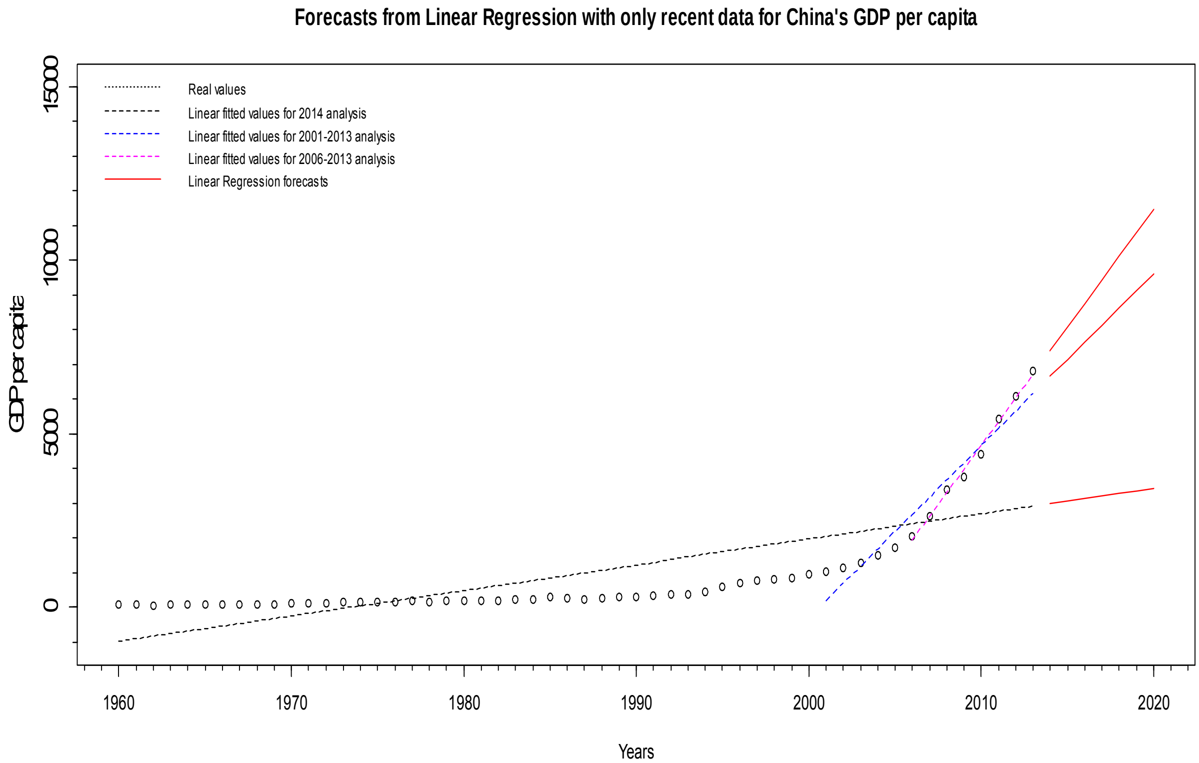
In this diagram, it can be seen that the two new fitted lines fit very well with the real values and definitely much better than the fitted line that was produced previously by the analysis with the whole dataset (and can be also seen in the graph as a black dashed line). This gives more confidence that the forecasts made with the subset 2001-2013 (blue line) are more accurate than the forecasts produced by the whole dataset.

It is also obvious that China's GDP per capita has risen dramatically after 2000. The slopes of the two fitted lines (green and blue) indicate the different rates of growth for the GDP and it can be easily seen that the last decade the rate of growth is much higher. The predictions for the future also show even bigger increase and this is a first indication for the potential growth of emerging markets.



[Figure 4.1.3] – Analyses for China, GDP per capita and the subsets up to 2000 and 2001-2013

The next question that this project attempted to answer is whether an even smaller dataset containing only a few recent observations could generate even better forecasts or not. For this, one more analysis was conducted with subset the values from year 2006 to year 2013. The code that used for this analysis is in the Appendices E.2.5. (p. 330). The next graph (Figure 4.1.4) shows the fitted line for this analysis (purple line) as well as the results for the analyses for the whole dataset (black line) and for the 2001-2013 dataset (blue line). The result is an even bigger coefficient of slope for the fitted line and the forecasts indicating even bigger increase in China's GDP per capita.



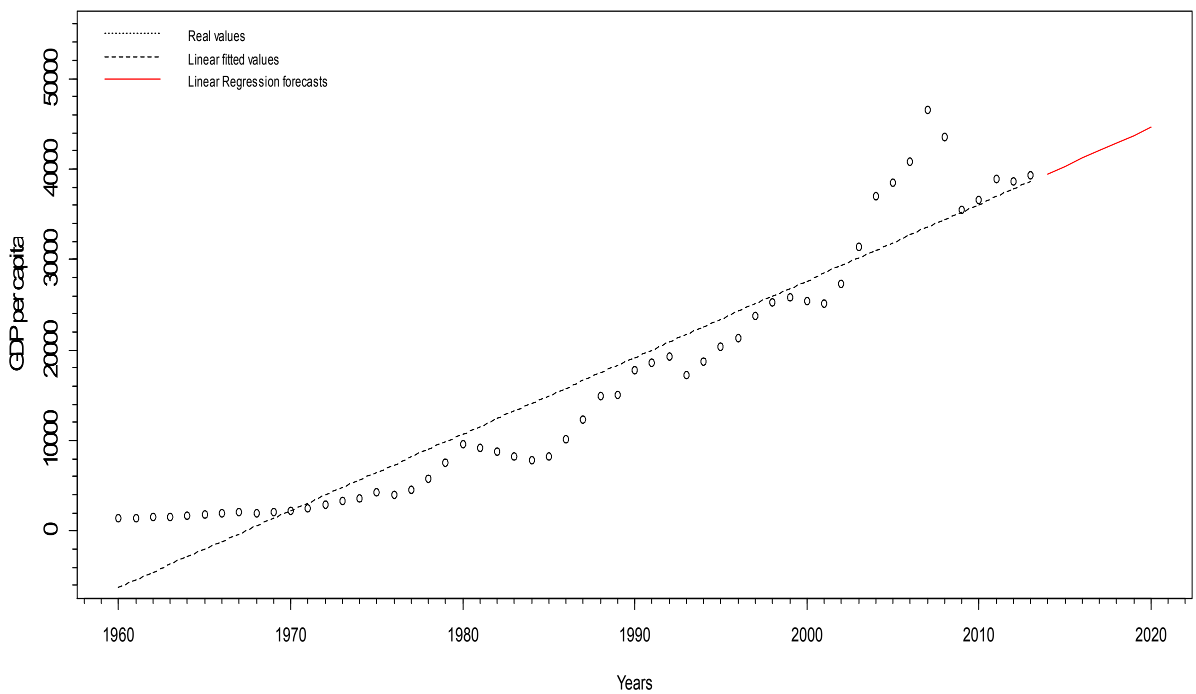
[Figure 4.1.4] – Analyses for China, GDP per capita and the subset from 2006-2013

4.1.2. Analysis for UK's GDP per capita

The same analysis process with the five different cases of datasets was implemented for the UK time series and GDP per capita Indicator. First was conducted again the analysis for the whole dataset from year 1960 to year 2013. The results can be seen in the Figure 4.1.5, where the real values (points), the fitted line (black dashed line) and the forecasts (red line) are plotted.

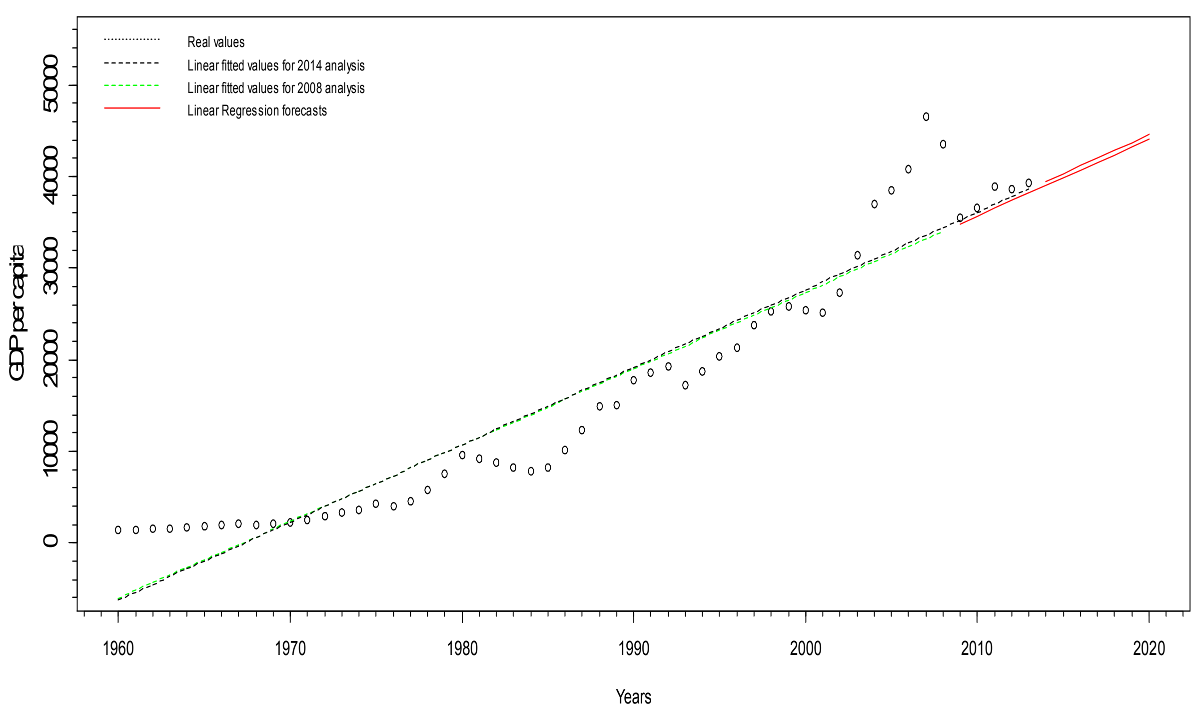
The second analysis was to apply the algorithm to the dataset up to year 2008 and the results are shown in Figure 4.1.2., where the green dashed line gives the fitted line for the dataset up to 2008 and the black line gives the fitted line from the whole dataset. In this case, on the contrary to China, the two lines are almost identical as the UK time series has an almost constant trend through time. Although the forecast misses the high peak of the real values between 2003 and 2008, it fits very well for the values after 2009.

Forecasts from Linear Regression for UK's GDP per capita



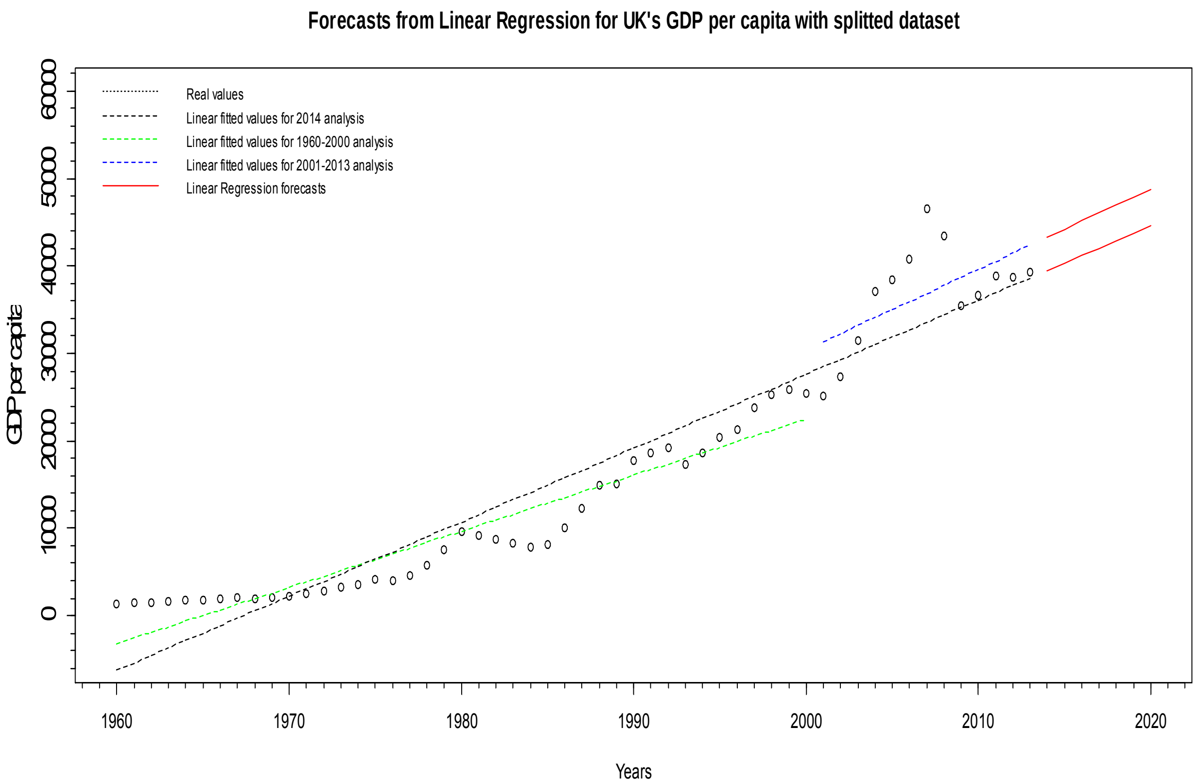
[Figure 4.1.5] – Analysis for UK, GDP per capita and whole dataset

Forecastsat at 2008 from Linear Regression for UK's GDP per capita



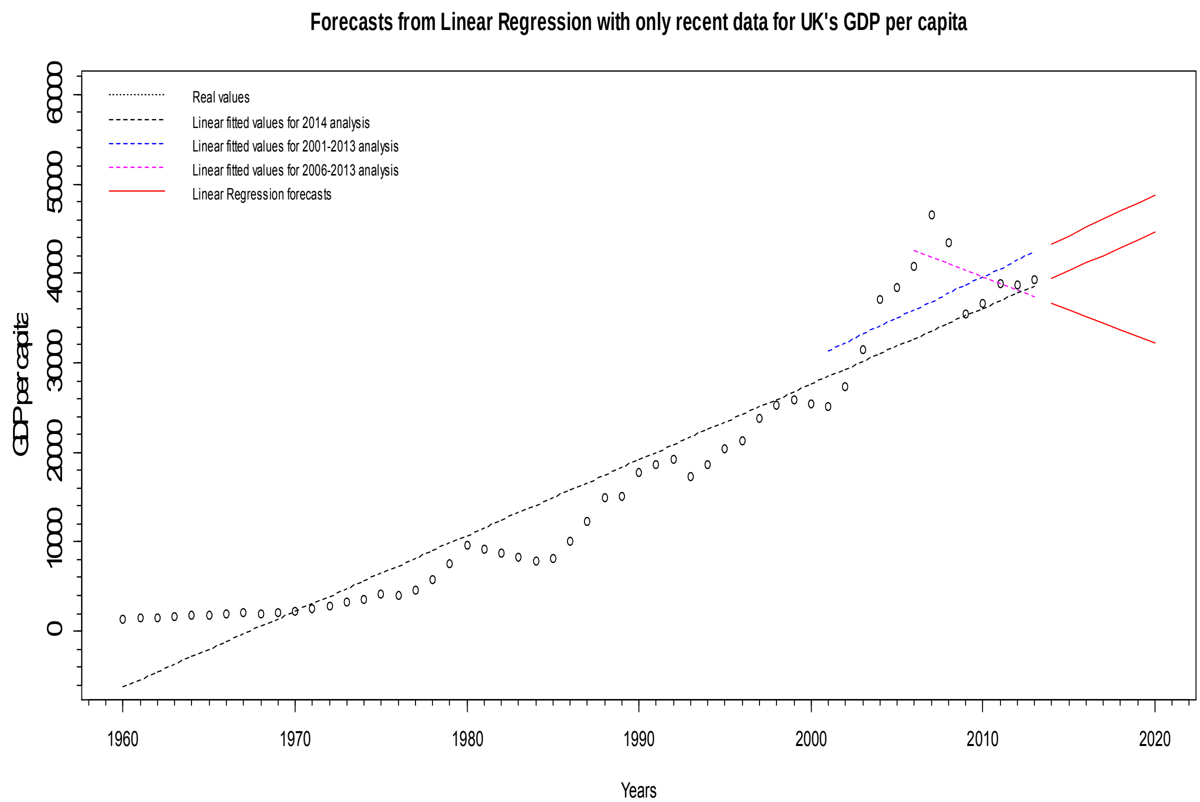
[Figure 4.1.6] – Analysis for UK, GDP per capita and the dataset up to 2008

The next figure (Figure 4.1.7) shows the results of the next two analyses for the dataset up to 2000 and for the dataset 2001-2013, where the green line is the fitted line for the analysis until year 2000 and the blue line the fitted line for the analysis with subset 2001-2013. What can be noted from the two relevant graphs for UK and China regarding the rate of growth of their GDP per capita, is that UK exhibits almost a constant rate of growth through time, while China shows a great increase in this rate after 2000, justifying its title as an emerging market.



[Figure 4.1.7] – Analyses for UK, GDP per capita and the subsets up to 2000 and 2001-2013

Finally, the analysis for the dataset from 2006 to 2013 is shown in Figure 4.1.8. Here, the fitted line for this analysis (purple line), as well as the results for the analyses for the whole dataset (black line) and for the 2001-2013 dataset (blue line) are plotted. The result is totally different from the previous analysis for dataset 2001 to 2013 and it gives a forecast with negative slope, indicating that UK's GDP will decrease the next years. So, although for China the analyses for 2001 to 2013 and 2006 to 2013 gave similar results, for UK they give totally contradicting. From this, it is obvious that the right choice of dataset range is of high importance for the forecasts, as the results may vary significantly.



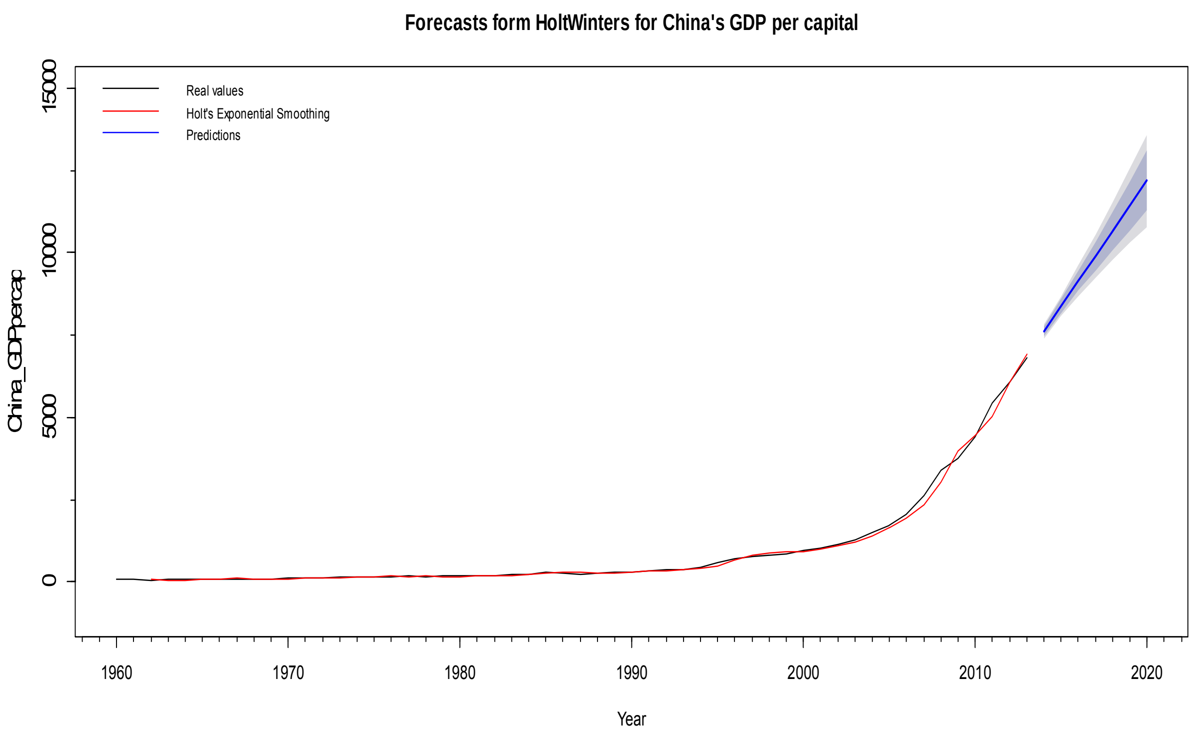
[Figure 4.1.8] – Analyses for UK, GDP per capita and the subset from 2006-2013

The rest analyses for the other indicators and the other countries in general give similar results and enable the generalisation of these first conclusions. The rest of the graphs from Linear Regression analysis are shown in Appendix C.1. (p.133), while more complete and general conclusions for the Linear Regression analysis, as well as a comparison of the coefficients of slope for all analyses, are given in the Comparison of Algorithms and Discussion Chapter (Chapter 5).

4.2. Holt's Exponential smoothing

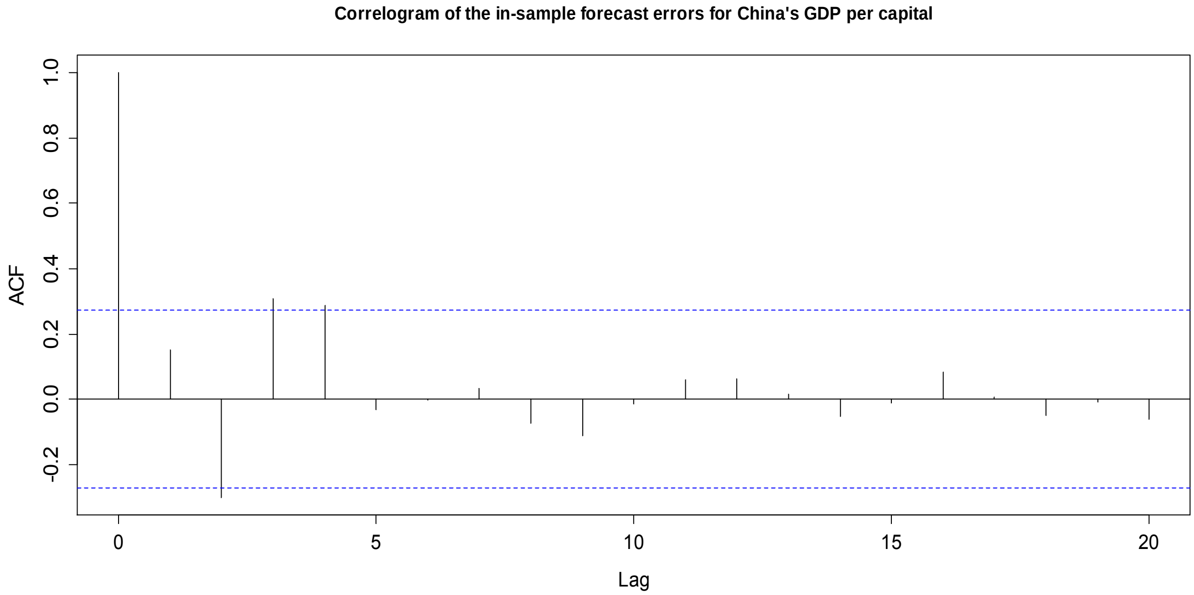
4.2.1.Analysis for China's GDP per capita

To keep consistency with the Linear Regression analysis, the same analysis process was followed as much as possible for the Holt's Exponential smoothing. Again, the first stage was to apply the model to the whole dataset of China values for the GDP per capita Indicator, from year 1960 to year 2013. The code for this analysis can be found in Appendix E.3.1. (p. 332). The results can be seen in the Figure 4.2.1, where the real values (black line), the smoothed line (red line), the forecasts (blue line) and the 80% and the 95% prediction intervals (blue shaded areas) are plotted.



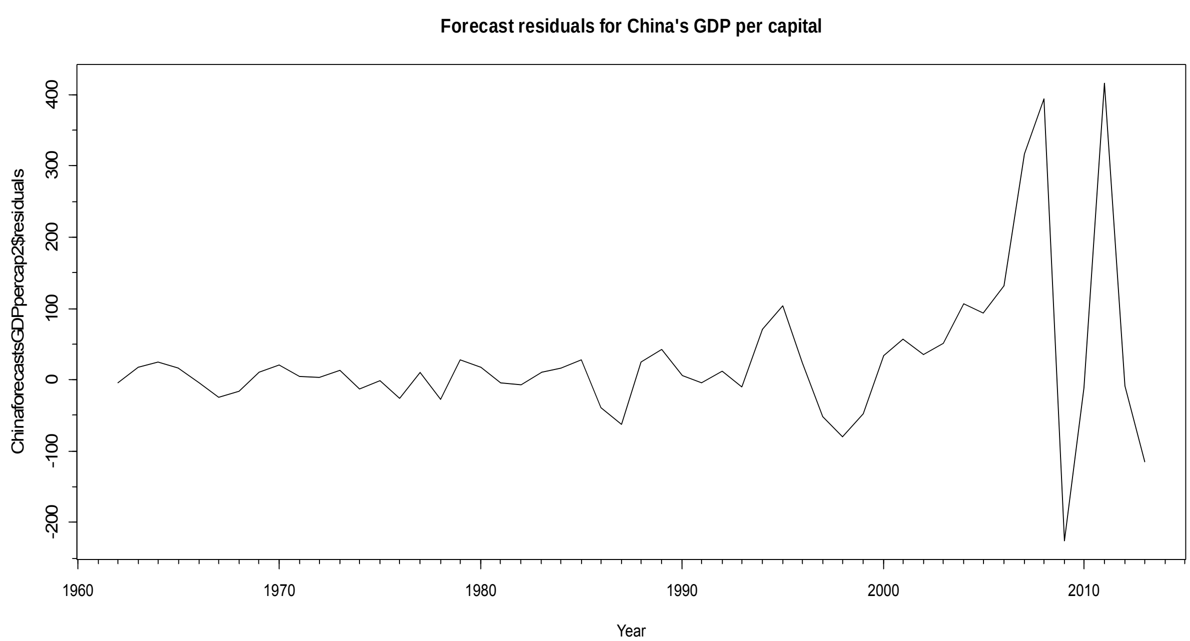
[Figure 4.2.1] – Analysis for China, GDP per capita and whole dataset

The parameters a and b are estimated by R as a=0.573 and b=1. The effectiveness of the model can be evaluated by examining the residuals (in-sample errors), where the residual values are calculated as the difference between real and fitted values. If the predictive model is adequate, the residuals must have no correlations for successive predictions. In order to check this, the correlogram of the in-sample forecast errors is plotted (Figure 4.2.2).

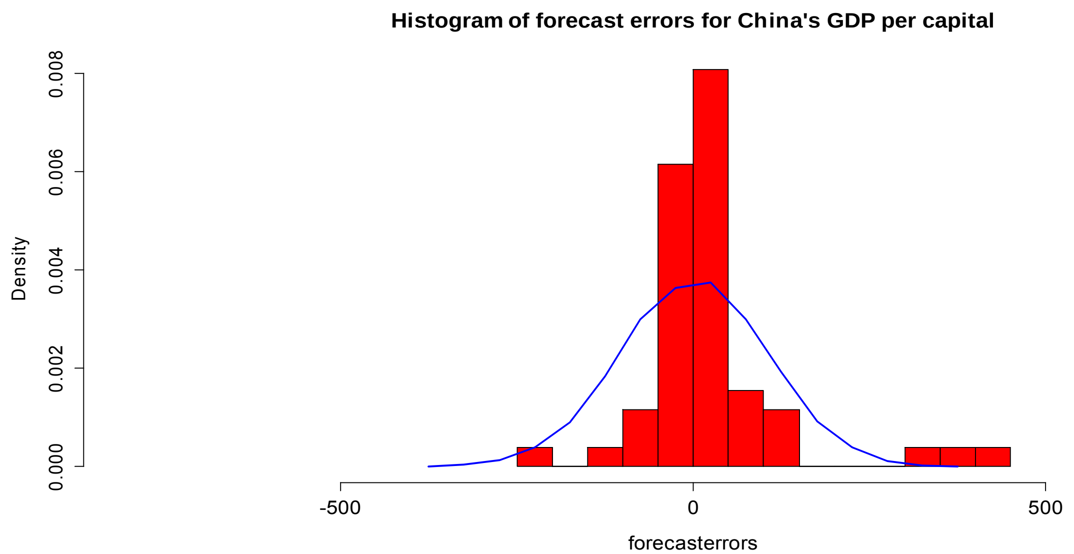


[Figure 4.2.2] – Correlogram of in-sample errors

From this we can see that the autocorrelations after lag 4 are practically zero. Also, with the Box-Ljung test the p-value was calculated to be 0.475, so there is little evidence that the residuals have non-zero autocorrelations. Finally, the suggested model is also evaluated by checking if the forecast errors have a mean zero and are normally distributed. This can be done by plotting the errors (Figure 4.2.3) and making a histogram of their distribution (Figure 4.2.4).



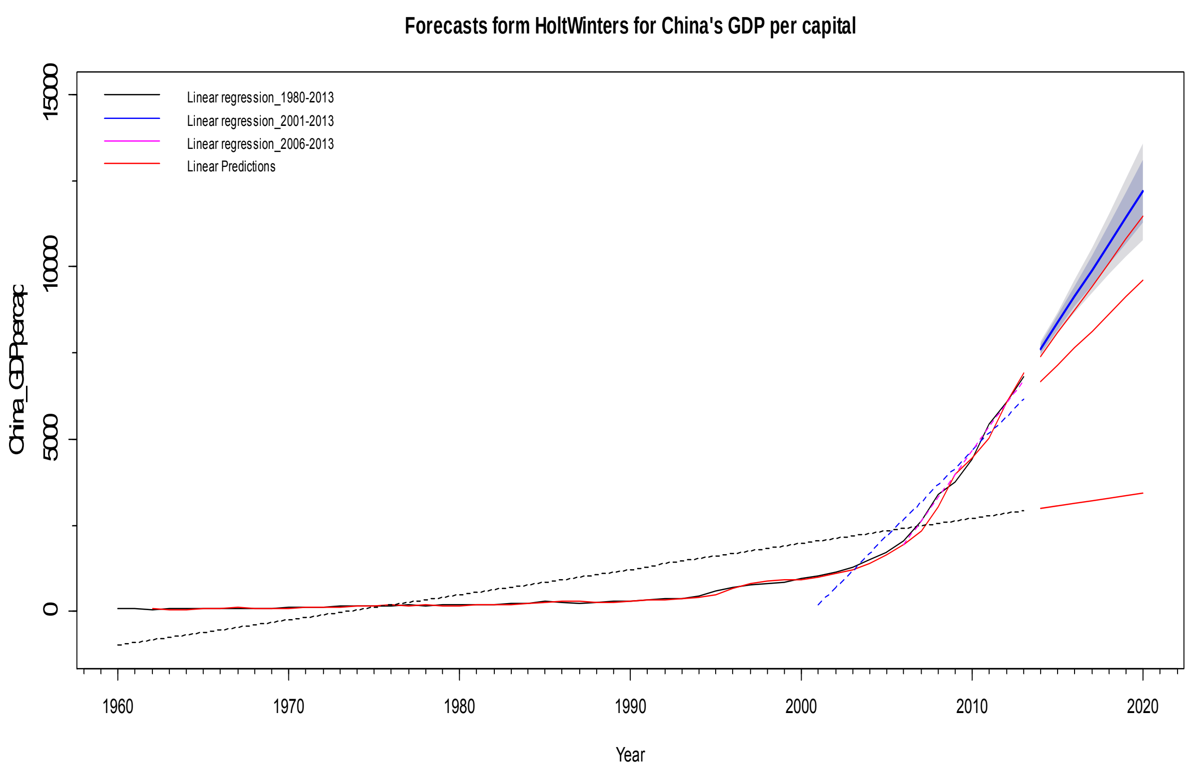
[Figure 4.2.3] – Forecast residuals



[Figure 4.2.4] – Histogram and distribution of forecast residuals

The absence of autocorrelations for the forecast errors, the high p-value and the fact the errors are normally distributed with a mean around zero indicate that the model is adequate and the confidence that the forecasts will be accurate is significant.

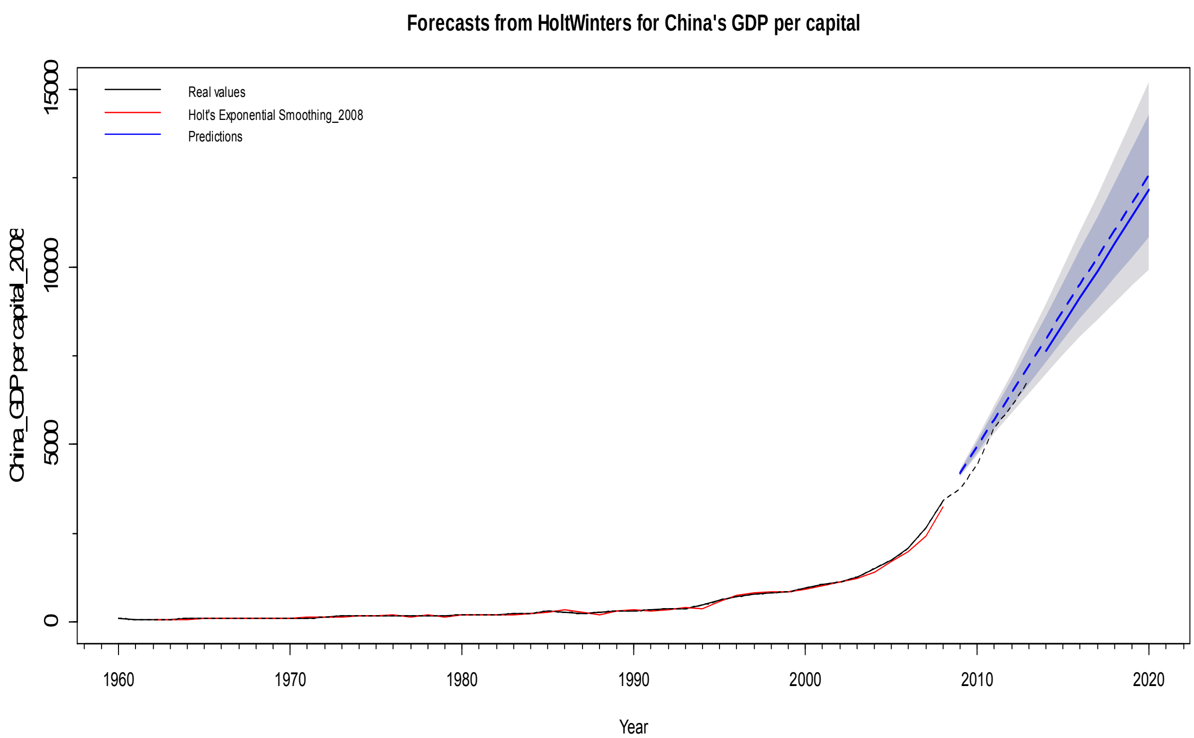
At this point, the three forecasts from Linear Regression analysis (for the whole dataset, the dataset 2001-2013 and the dataset 2006-2013) were added to the plot for comparison with Holt's Exponential smoothing model, as can be seen in the Figure 4.2.5.



[Figure 4.2.5] – Comparison of Linear Regression and Holt's Exponential smoothing

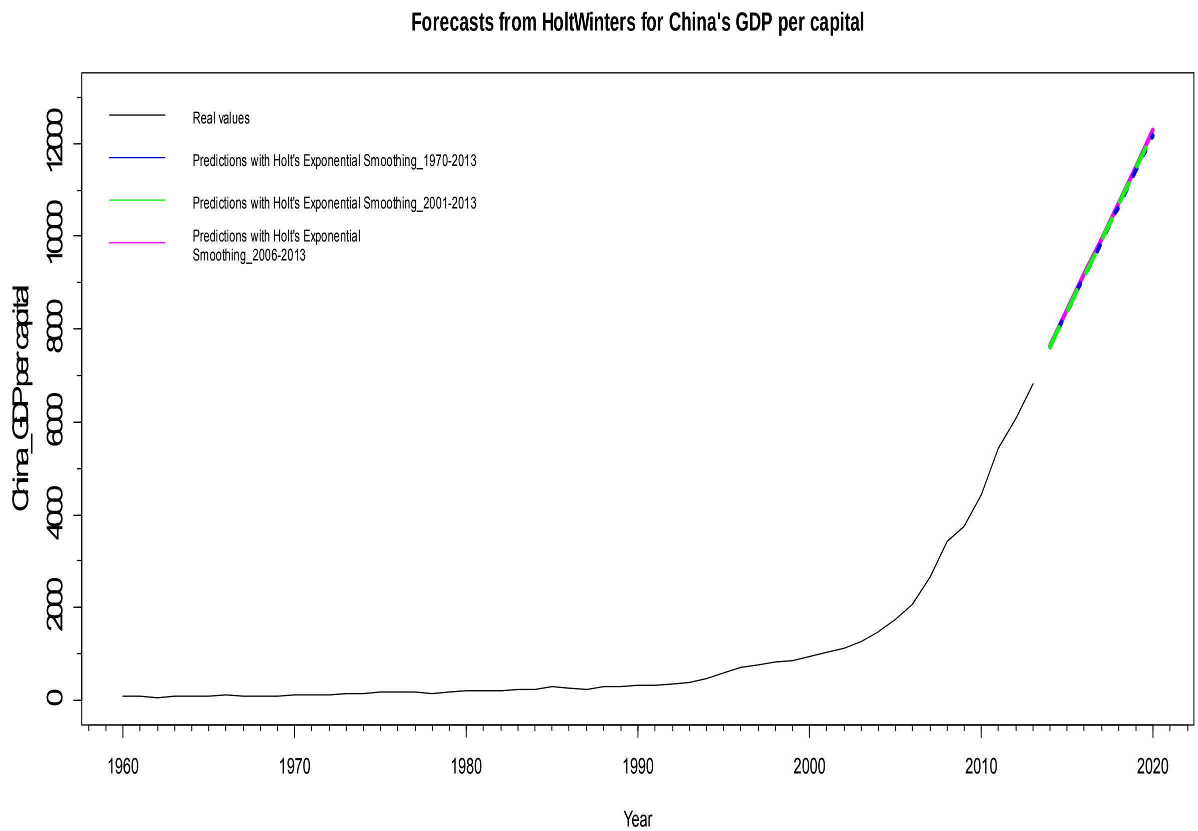
It can be noticed that the Holt's Exponential model gives similar results to the Linear analyses for the smaller subsets. In fact is very close to Linear analysis for subset 2006-2013 and less close, but still with similar slope, to subset 2001-2013. On the contrary, the Linear Regression analysis for the whole dataset gives very different and non-accurate results, possibly because it fails to incorporate the abrupt change in the slope of the China time series after year 2000.

The second analysis was to apply the algorithm to the dataset up to year 2008. The results are shown in Figure 4.2.6, where the blue dashed line gives the forecasts for the dataset up to 2008, the black dashed line the real values form 2009 to 2013 and the blue solid line gives the forecasts from the whole dataset. From the graph we can remark how well the forecasts from 2008 fit with the real values, something that increases the confidence that the forecasts with Holt's Exponential smoothing can be proved satisfactory.



[Figure 4.2.6] – Analysis for China, GDP per capita and the dataset up to 2008

The Final step for the Holt's Exponential smoothing analysis was to apply the model to the two smaller subsets of data from year 2001 to year 2013 and from year 2006 to year 2013. The results are shown in the same graph (Figure 4.2.7), where the green line shows the forecasts for the set 2001-2013 and the purple line for the set 2006-2013. The blue line is the forecasts for the whole dataset.



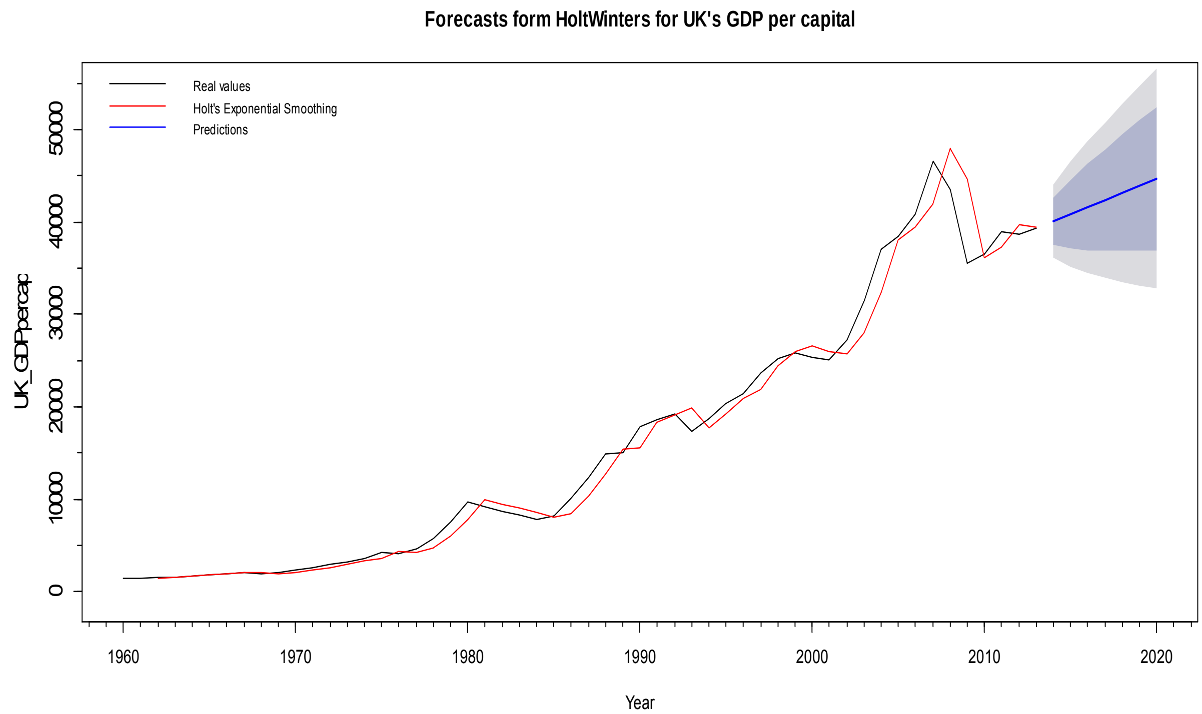
[Figure 4.2.7] – Analyses for China, GDP per capita and the subsets 2001-2013 and 2006-2013

As can be seen from the graph, in contrast with Linear Regression where the different analyses give significantly different results, the Holt's Exponential smoothing model for China gives almost identical results for the three different datasets that were used. This also contributes to the reliability of the algorithm.

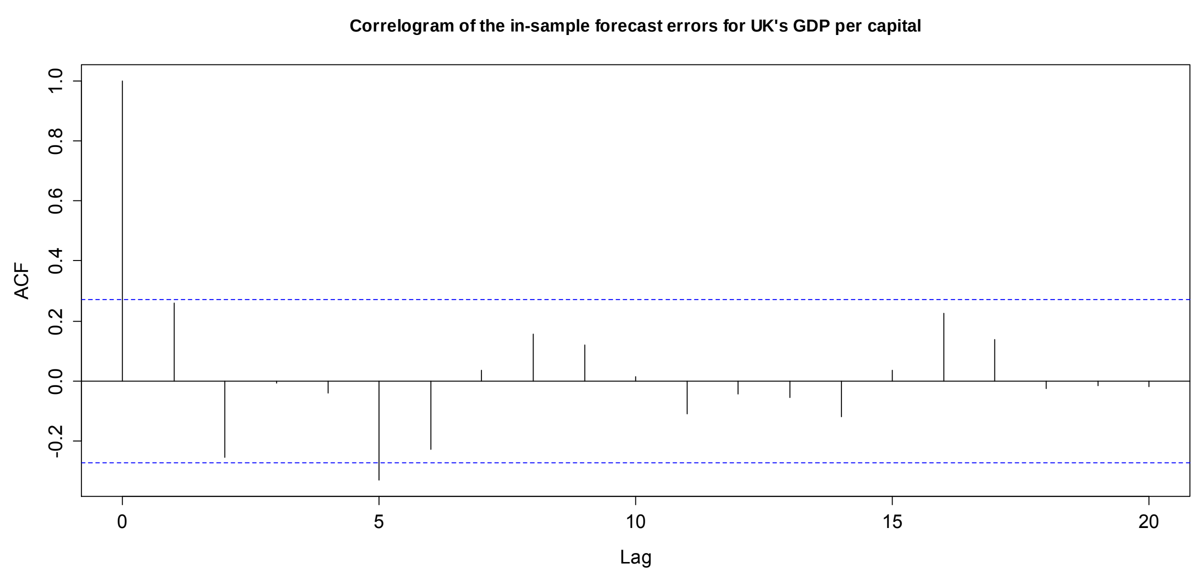
4.2.2.Analysis for UK's GDP per capita

The same analysis process was implemented for the UK time series and GDP per capita Indicator. First was conducted again the analysis for the whole dataset from year 1960 to year 2013. The results can be seen in the Figure 4.2.8, where the real values (black line), the smoothed line (red line), the forecasts (blue line) and the 80% and the 95% prediction intervals (blue shaded areas) are plotted.

The parameters a and b are estimated by R as a=1 and b=0.046. In order to check the effectiveness of the model, as in the analysis for China, the correlogram of the in-sample forecast errors is plotted (Figure 4.2.9).

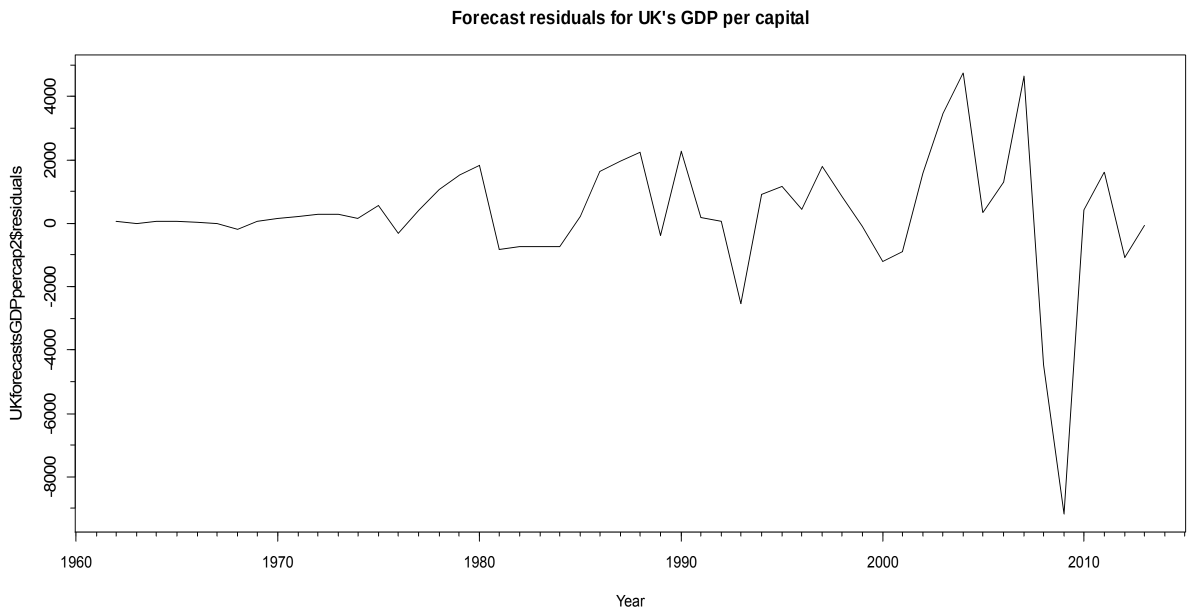


[Figure 4.2.8] – Analysis for UK, GDP per capita and whole dataset

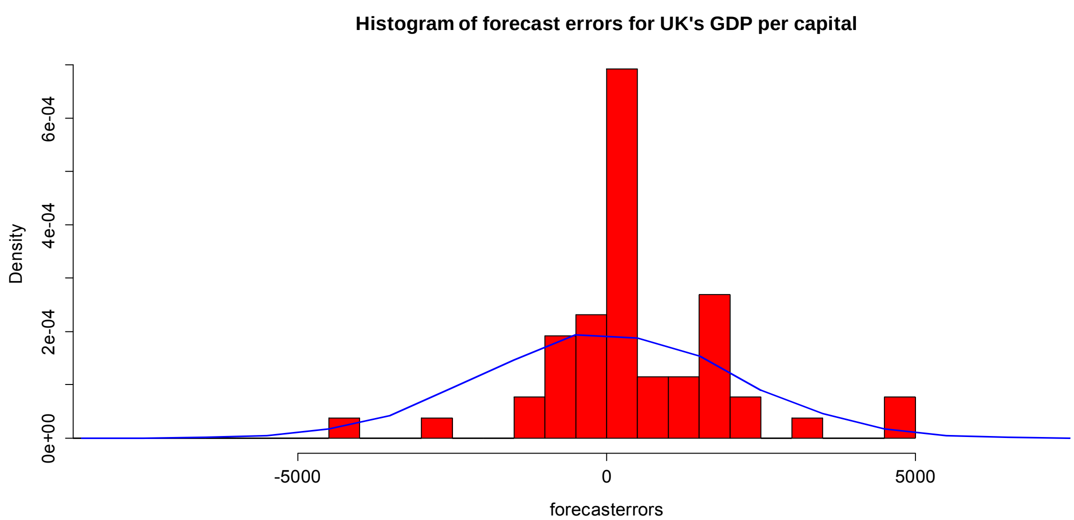


[Figure 4.2.9] – Correlogram of in-sample errors

From this we can see that the autocorrelations after lag 5 are practically zero. Also, with the Box-Ljung test the p-value was calculated to be 0.1175, so there is little evidence that the residuals have non-zero autocorrelations. Finally, the suggested model is also evaluated by checking if the forecast errors have a mean zero and are normally distributed. This can be done by plotting the errors (Figure 4.2.10) and making a histogram of their distribution (Figure 4.2.11).



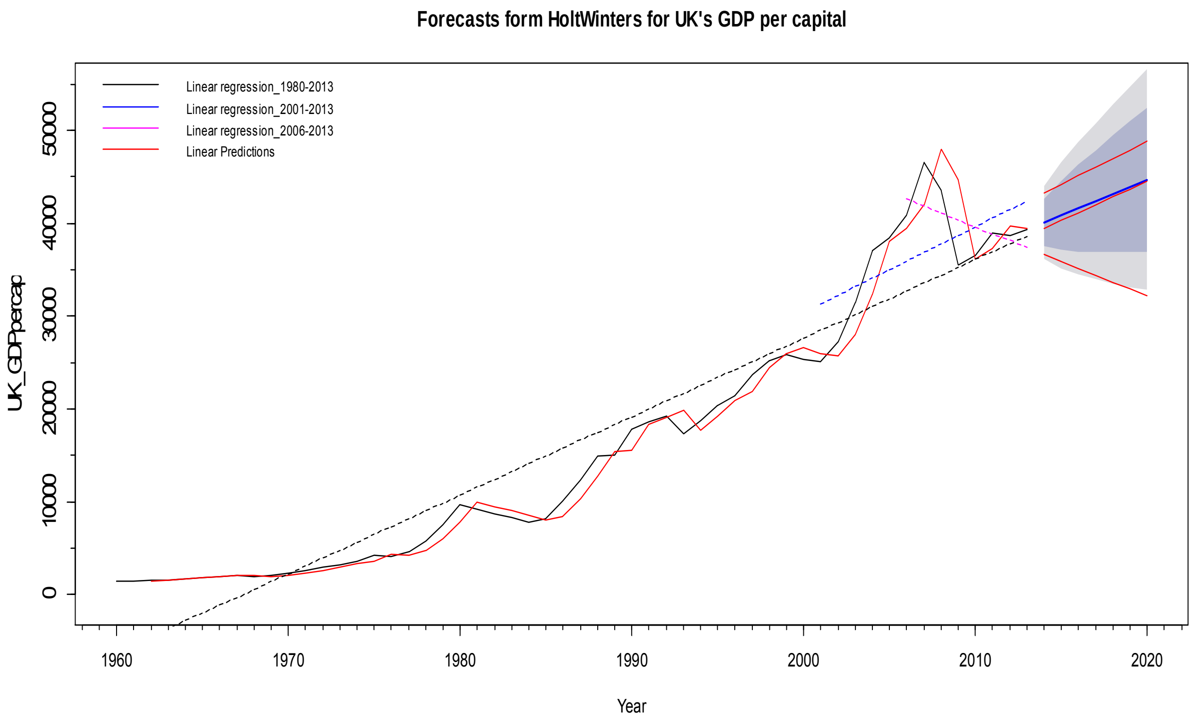
[Figure 4.2.10] – Forecast residuals



[Figure 4.2.11] – Histogram and distribution of forecast residuals

The absence of autocorrelations for the forecast errors, the high p-value and the fact the errors are normally distributed with a mean around zero indicate that the model is adequate and the confidence that the forecasts will be accurate is significant.

At this point, again the three forecasts from Linear Regression analysis (for the whole dataset, the dataset 2001-2013 and the dataset 2006-2013) were added to the plot for comparison with Holt's Exponential smoothing model, as can be seen in the Figure 4.2.12.

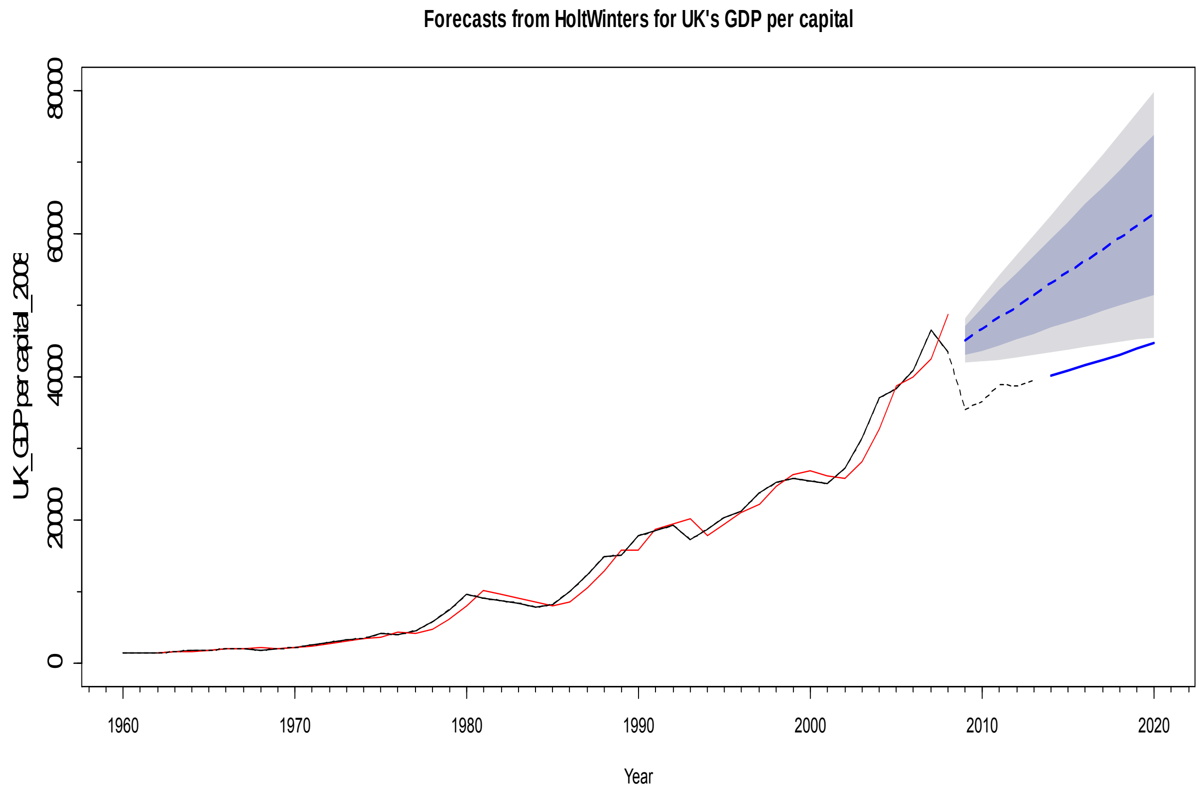


[Figure 4.2.12] – Comparison of Linear Regression and Holt's Exponential smoothing

In this case the Holt's Exponential model gives almost identical results with the Linear forecasts from the whole dataset and similar to those from the dataset 2001-2013. This could be explained by the fact that the UK time series exhibits an almost constant trend through time, on the contrary to China, where due to the abrupt change in the trend the Linear forecasts with the whole dataset were totally underestimated.

On the other hand, for UK, the Linear analysis with only a few observations (from 2006 to 2013) gives a forecast with negative slope, indicating that UK's GDP will continue to fall, contradicting all other analyses.

The project continues with the analysis to the dataset up to year 2008. The results are shown in Figure 4.2.13, where the blue dashed line gives the forecasts for the dataset up to 2008, the black dashed line the real values form 2009 to 2013 and the blue solid line gives the forecasts from the whole dataset.

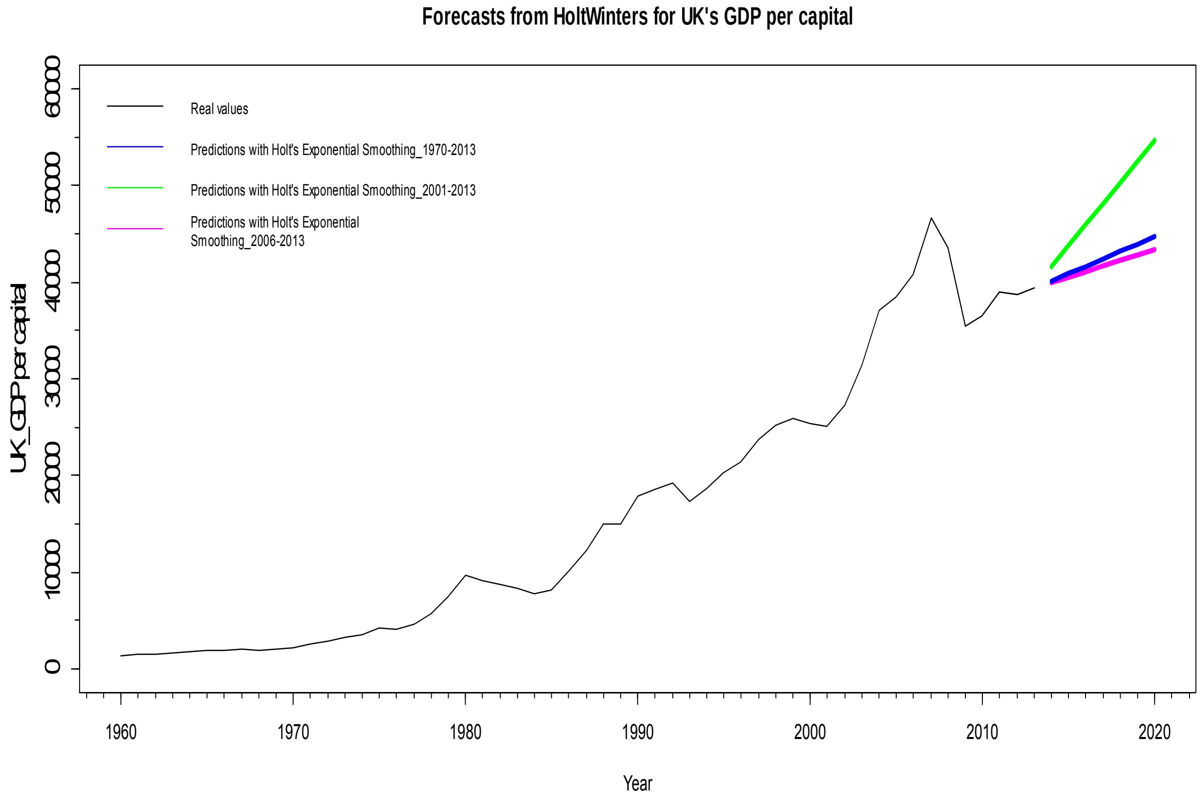


[Figure 4.2.13] – Analysis for UK, GDP per capita and the dataset up to 2008

From the graph we can see that the forecasts from 2008, although they have similar slope with the real values, they overestimate the values of GDP as they could not predict the abrupt drop in GDP that happened immediately after year 2008. This suggests a limitation of the predictive methods, as external political or social factors may influence the financial time series at any time without the ability to be predicted by the models.

The Final step for the Holt's Exponential smoothing analysis was to apply the model to the two smaller subsets of data from year 2001 to year 2013 and from year 2006 to year 2013. The results are shown in Figure 4.2.14, where the green line shows the forecasts for the set 2001-2013 and the purple line for the set 2006-2013. The blue line is the forecasts for the whole dataset.

In this case, the analysis for the set 2006-2013 gives similar results with the one for the whole dataset, while the analysis for the set 2001-2013 seems to overestimate the value of UK's GDP, in comparison with the previous ones.



[Figure 4.2.14] – Analyses for UK, GDP per capita and the subsets 2001-2013 and 2006-2013

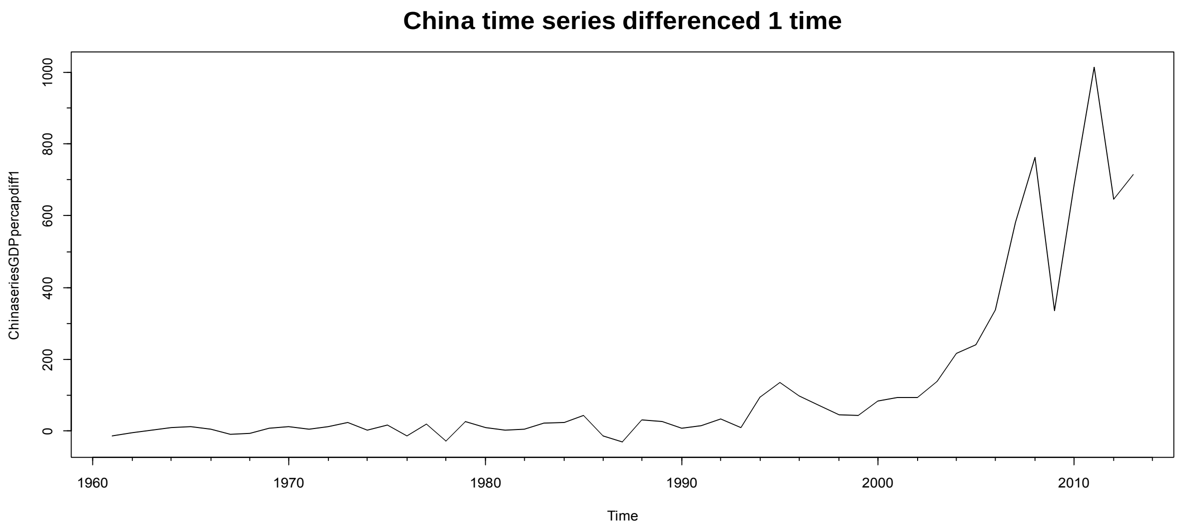
The rest of the graphs from Holt's Exponential smoothing are shown in Appendix C.2. (p.161), while more complete and general conclusions for the Holt's Exponential smoothing analysis, are given in the Comparison of Algorithms and Discussion Chapter (Chapter 5).

4.3. ARIMA

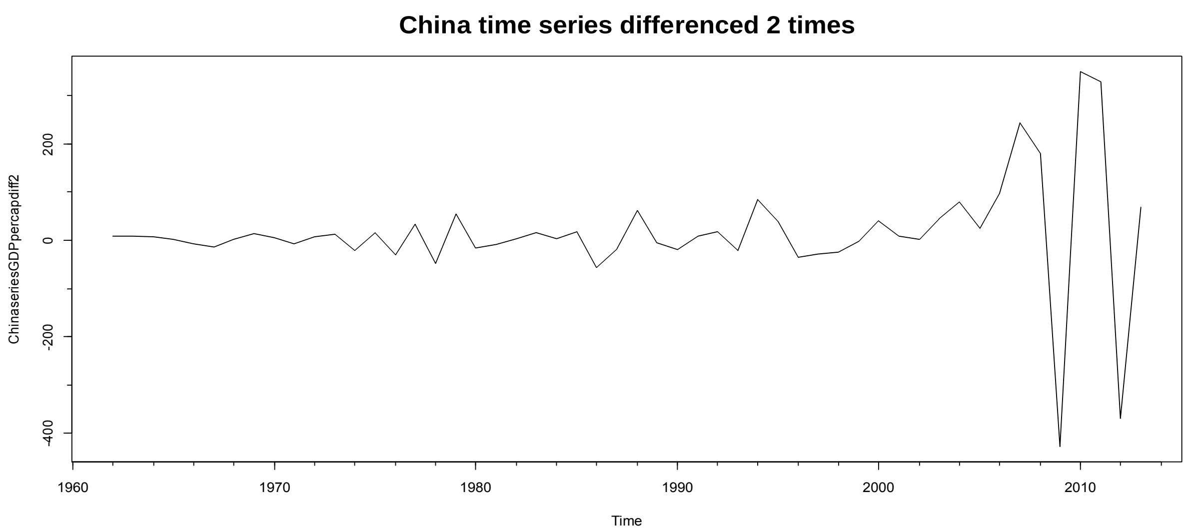
4.3.1.Analysis for China's GDP per capita

As has been mentioned, in order to maintain the consistency of the analysis the same datasets were analysed with ARIMA modeling as with the other two algorithms. The first analysis was again for the whole dataset of the China time series from year 1960 to year 2013 and for the GDP per capita Indicator.

Before the ARIMA model could be applied, the China time series should be differenced successively until it became stationary. Figure 4.3.1 and Figure 4.3.2 show the one and two times differenced time series respectively.

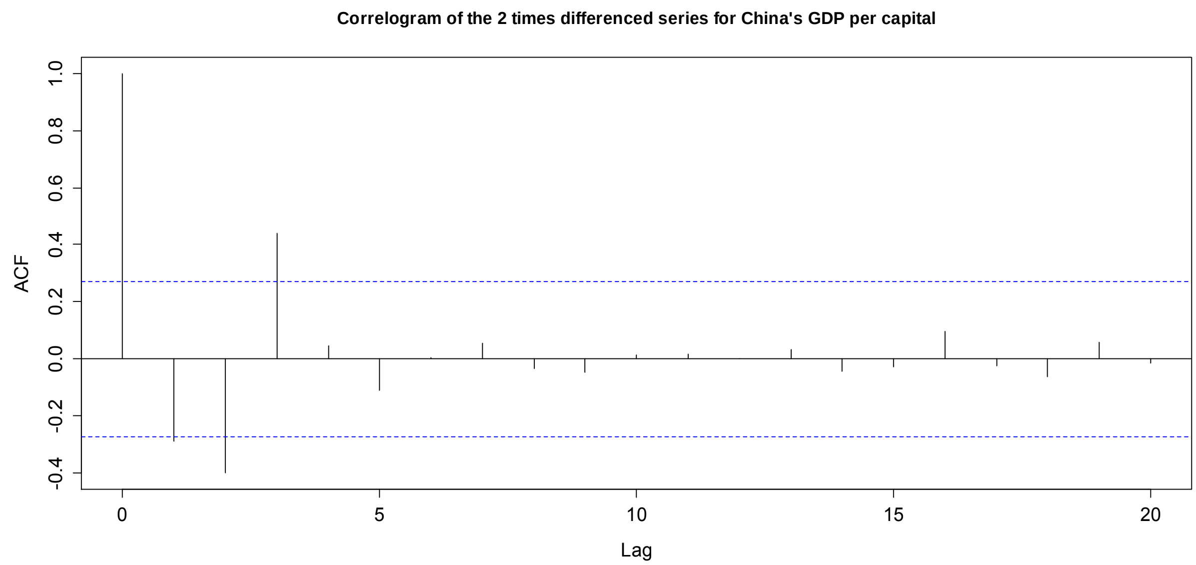


[Figure 4.3.1] – One time differenced China time series

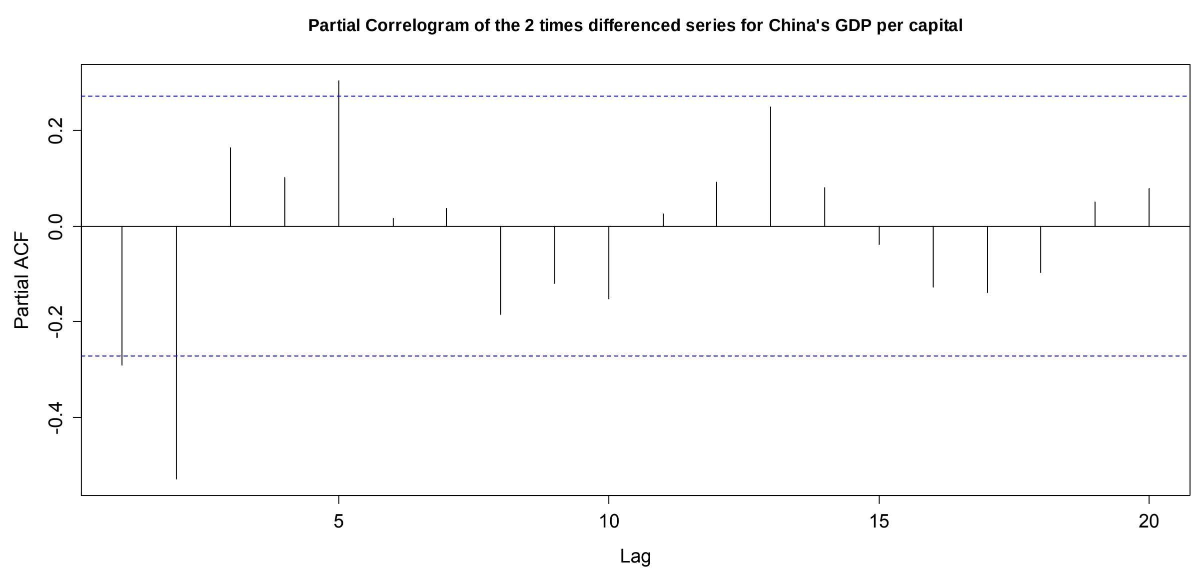


[Figure 4.3.2] – Two times differenced China time series

The one time differenced time series is not stationary, but the twice differenced time series present a constant mean around zero and can be considered stationary. So, the parameter (d) of the ARIMA model was chosen as d=2. The next step was to plot the autocorrelogram (ACF) and partial autocorrelogram (PACF) of the stationary time series (that is, of the twice differenced time series for this case) in order to estimate the parameters (p) and (q).

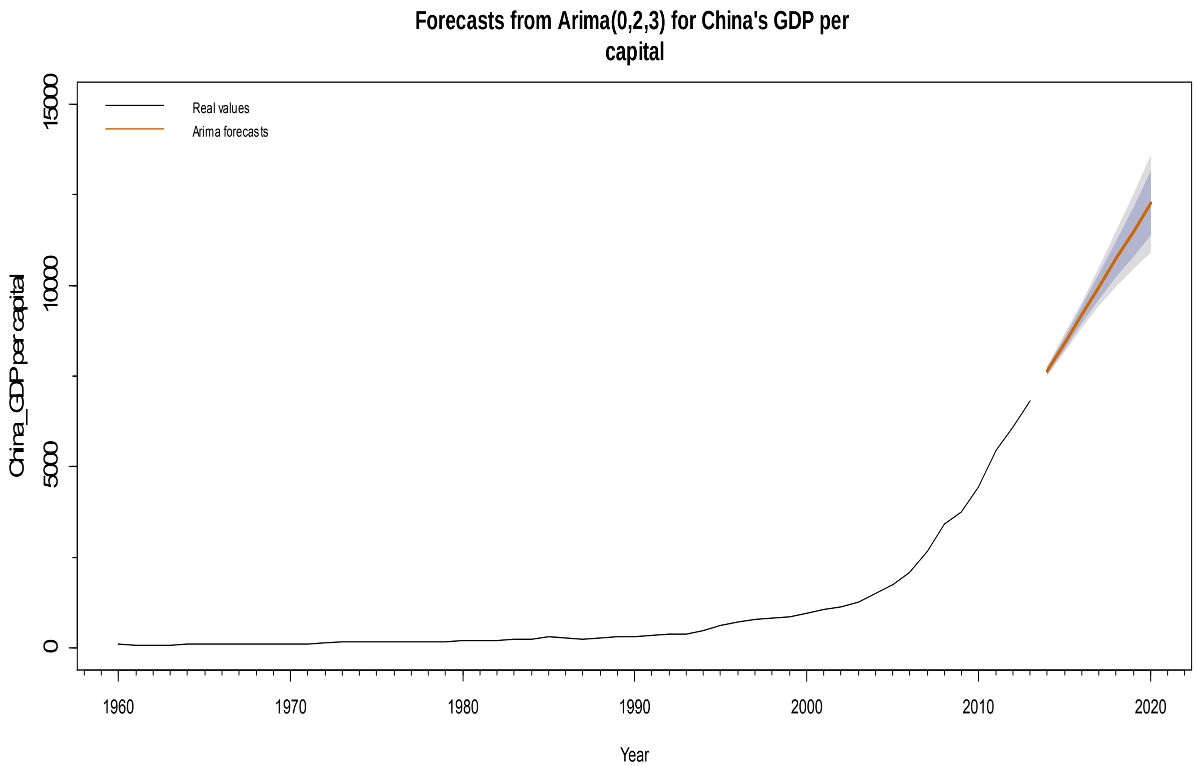


[Figure 4.3.3] – Autocorrelogram (ACF) of the twice differenced China time series



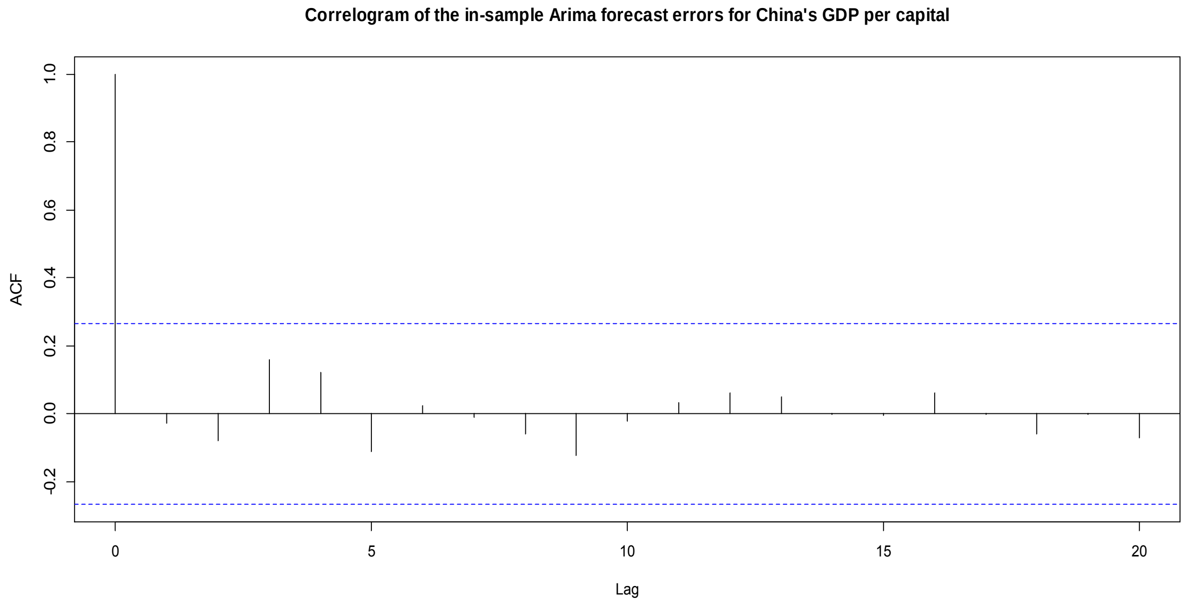
[Figure 4.3.4] – Partial autocorrelogram (PACF) of the twice differenced China time series

The ACF graph shows that the autocorrelations are zero after the 3rd lag, so it is q=3. From the PACF graph we see that the autocorrelations could be considered zero (smaller than the significant boundaries that are represented with the blue lines) after lag 5, so p=5. That means that the ARIMA model could be of order (p,d,q) = (5,2,3), but because this model has too many parameters and in an effort to follow the principle of parsimony, we chose instead to apply the model ARIMA(0,2,3). Applying this model we get the results shown in Figure 4.3.5, where we see the real values (black line), the forecasts (orange line) and the 80% and the 95% prediction intervals (blue shaded areas).



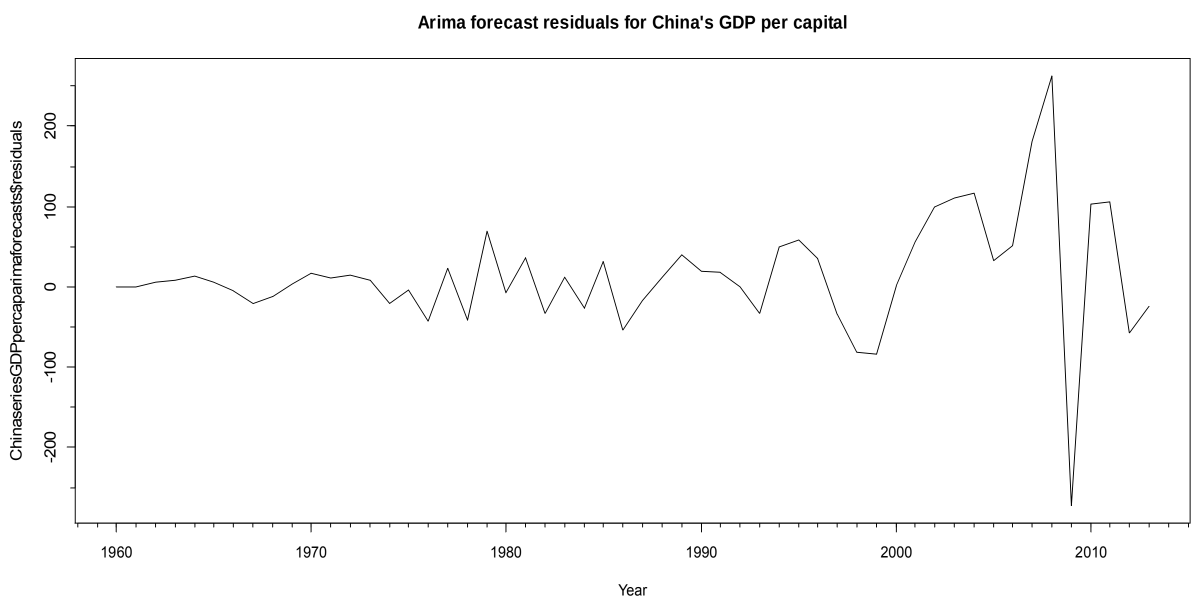
[Figure 4.3.5] – Analysis for China, GDP per capita and whole dataset

In order to evaluate the effectiveness of the model we should examine the residuals (in-sample errors) of the forecasts, where the residual are calculated as the difference between real and fitted values. If the predictive model is adequate, the residuals must have no correlations for successive predictions. In order to check this, the correlogram of the in-sample forecast errors is plotted (Figure 4.3.6).

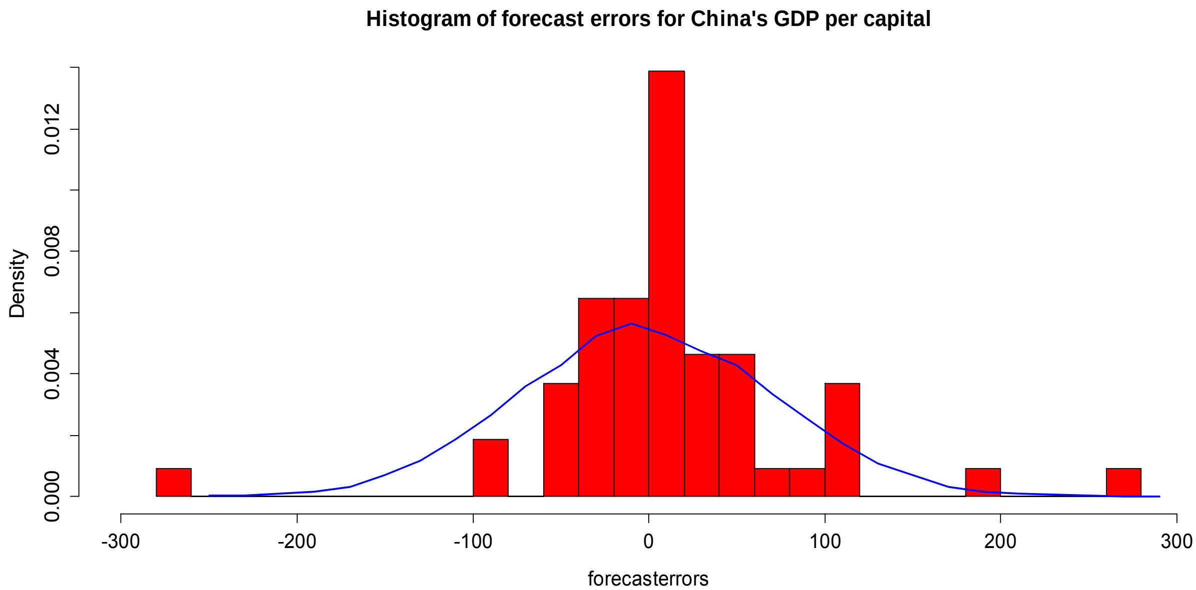


[Figure 4.3.6] – Correlogram of in-sample errors of ARIMA forecasts

From this we can see that the autocorrelations are practically zero as they never exceed the significant boundaries (blue lines). Also, with the Box-Ljung test, the p-value was calculated to be p-value = 0.998, so there is little evidence that the residuals have non-zero autocorrelations. Finally, the suggested model is also evaluated by checking if the forecast errors have a mean zero and are normally distributed. This can be done by plotting the errors (Figure 4.3.7) and making a histogram of their distribution (Figure 4.3.8).



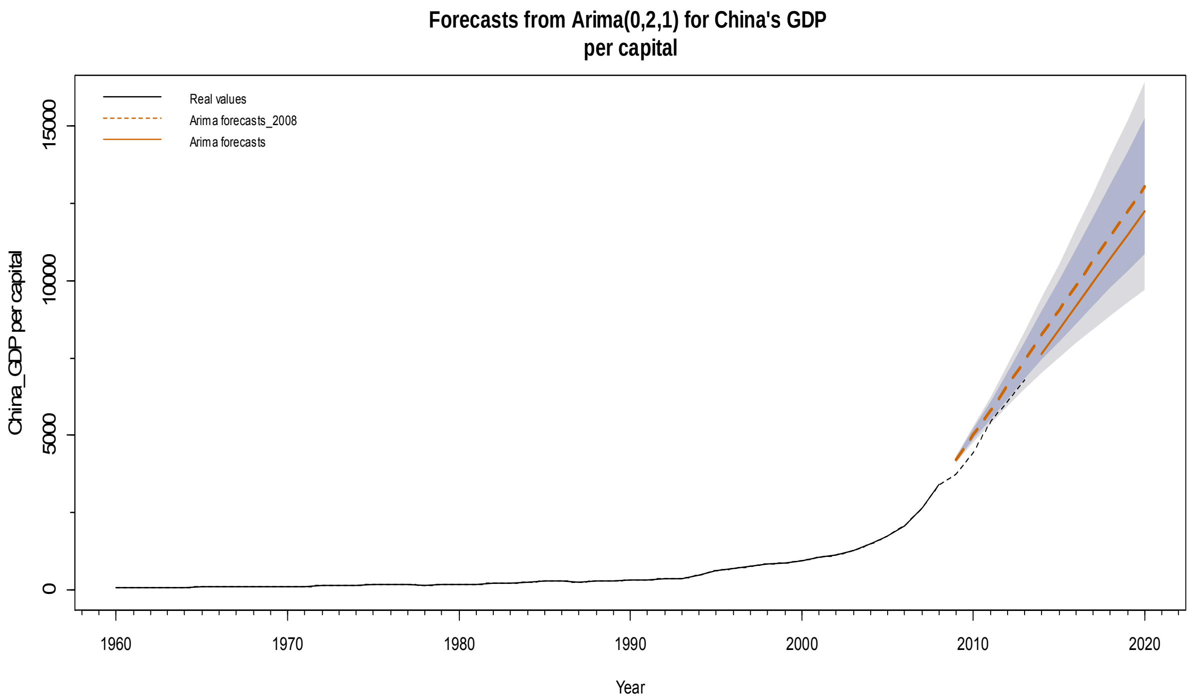
[Figure 4.3.7] –Residuals of ARIMA forecasts



[Figure 4.3.8] – Histogram and distribution of forecast residuals

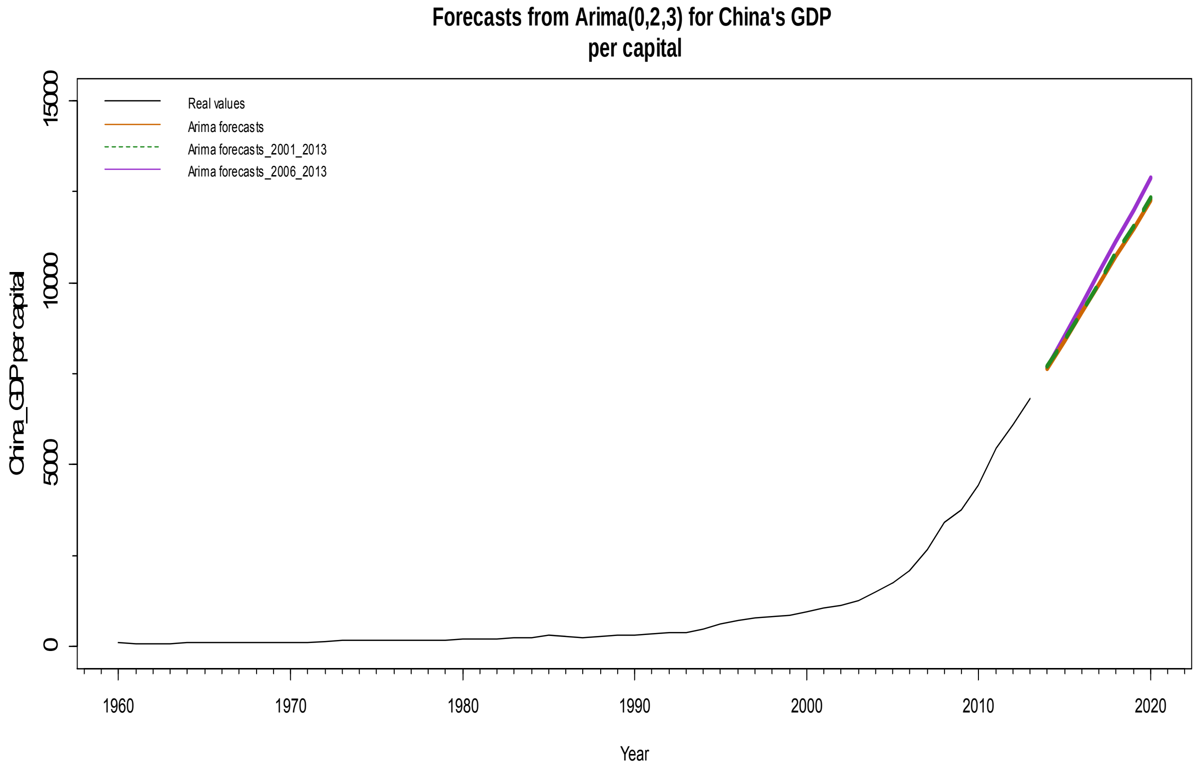
The absence of autocorrelations for the forecast errors, the high p-value and the fact the errors are normally distributed with a mean around zero allow us to assume that the model is adequate and the confidence that the forecasts will be accurate is significant.

The second analysis was to apply the algorithm to the dataset up to year 2008. With the same procedure as before, it was found that the appropriate model for this case was the ARIMA(0,2,1). The results are shown in Figure 4.3.9, where the orange dashed line gives the forecasts for the dataset up to 2008, the black dashed line the real values from 2009 to 2013 and the orange solid line gives the forecasts from the whole dataset. From the graph we can notice how well the forecasts from 2008 fit with the real values, something that increases the confidence that the forecasts with ARIMA model can be proved accurate.



[Figure 4.3.9] – Analysis for China, GDP per capita and the dataset up to 2008

Finally, the ARIMA model was applied to the two smaller subsets of data from year 2001 to year 2013 and from year 2006 to year 2013. In this case the model that was the most appropriate was the ARIMA(0,2,3). The results are shown together in the same graph (Figure 4.3.10), where the dark green line shows the forecasts for the set 2001-2013 and the dark purple line for the set 2006-2013. The orange line is the forecasts for the whole dataset.



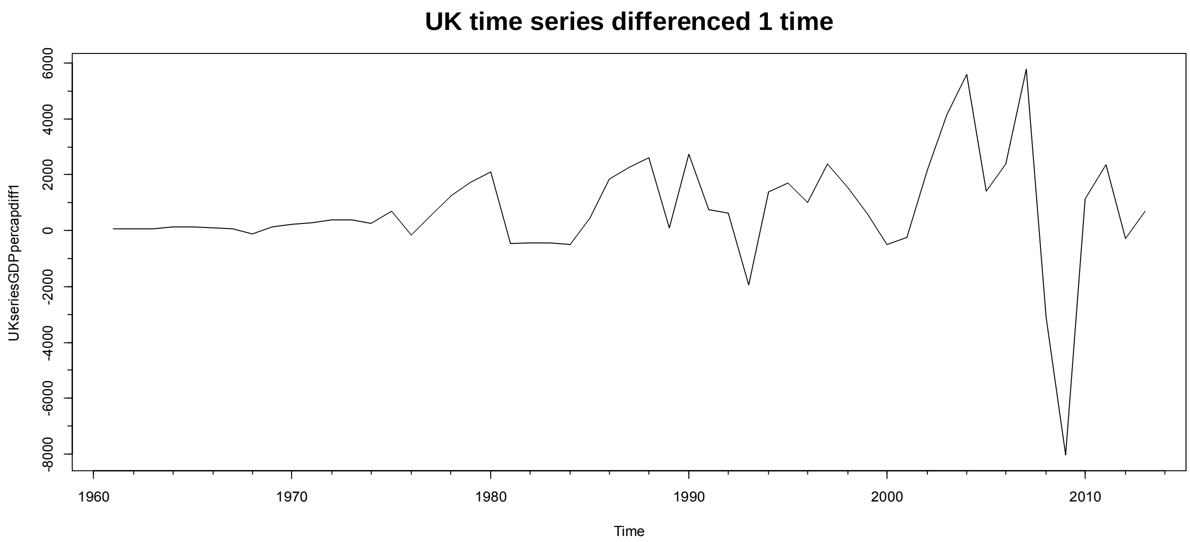
[Figure 4.3.10] – Analyses for China, GDP per capita and the subsets 2001-2013 and 2006-2013

As we can see from the graph, the forecasts from the analyses with the three different datasets are very close to each other, something that increases the confidence that the ARIMA model is reliable and gives accurate forecast without depending so much on the selection of the dataset.

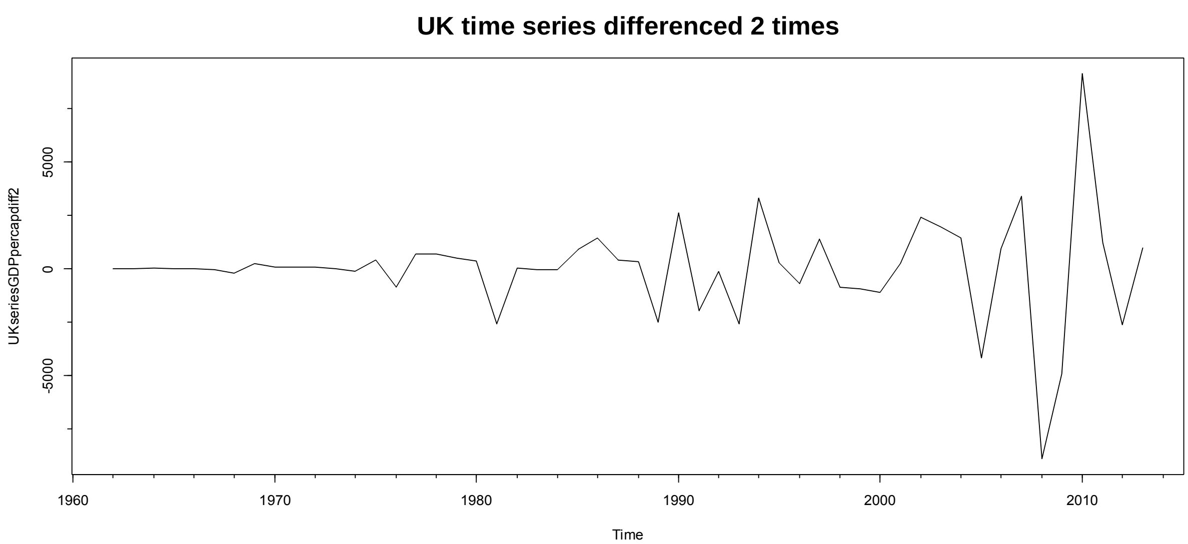
In this point, it is worth to be noted that the selection of the parameters (p,d,q) is very important and crucial for the final result, as slightly different parameters can give totally different forecasts. The experience of the researcher plays a very significant role for the selection of the parameters and in addition a great deal of experimentation with different combinations of (p,d,q) values is proven to be beneficial. Finally, the auto.arima() function of R provides valuable help in the selection of the parameters process, as it can be used complementary in order to check the results that different set of values give.

4.3.2.Analysis for UK’s GDP per capita

The analysis process for the UK time series and GDP per capita Indicator was similar. First was conducted again the analysis for the whole dataset from year 1960 to year 2013. The time series was not stationary, so it had to be differenced. The one and two times differenced UK time series are shown in Figure 4.3.11 and Figure 4.3.12 respectively.

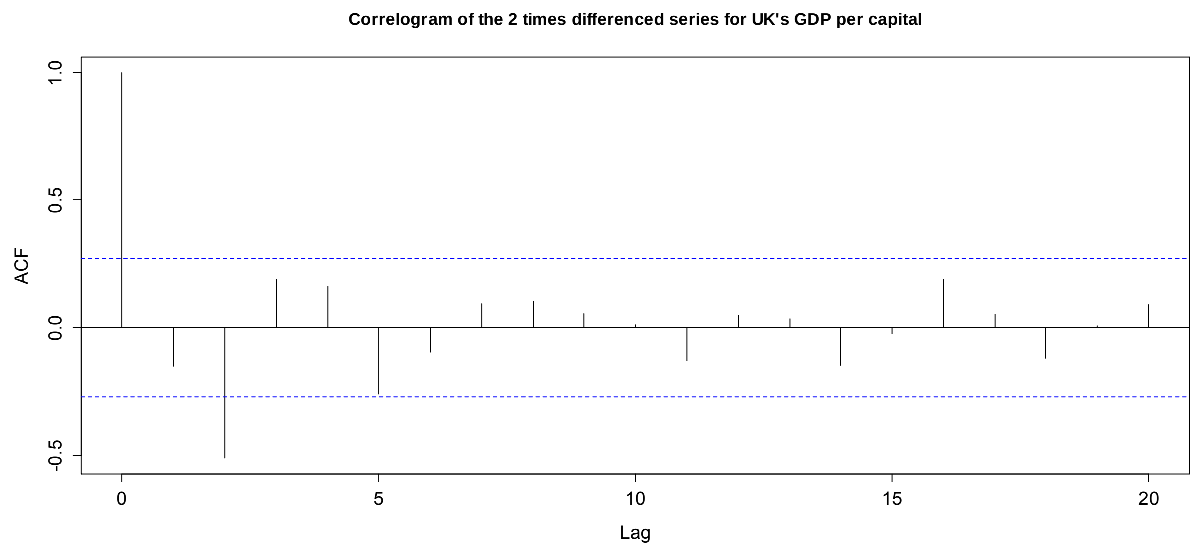


[Figure 4.3.11] – One time differenced UK time series

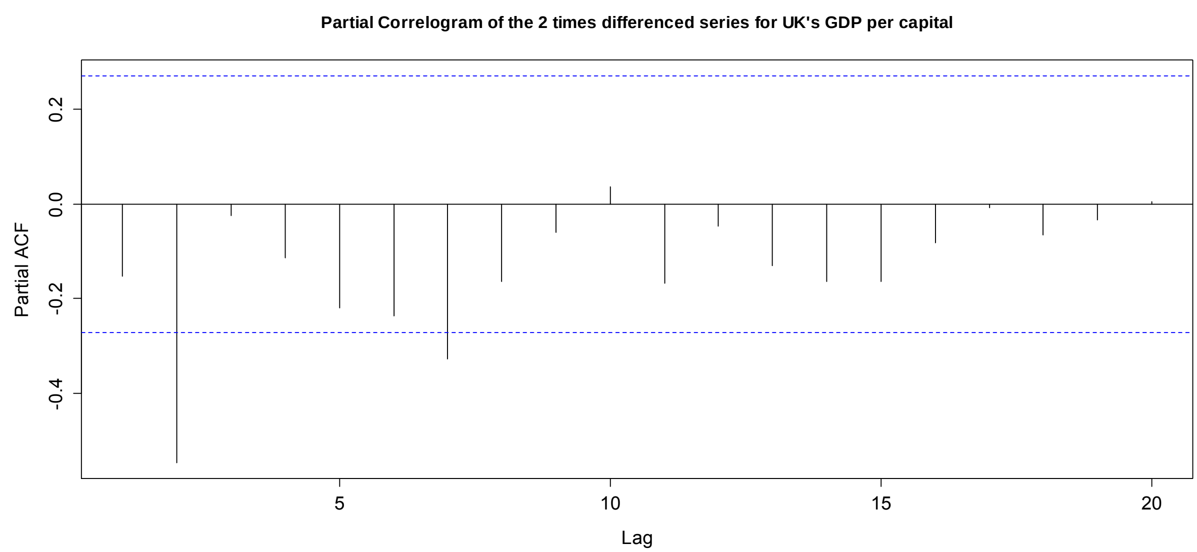


[Figure 4.3.12] – Two times differenced UK time series

The two times differenced times series can be considered stationary in a more satisfactory level, so the parameter (d) was chosen as d=2. The autocorrelogram (ACF) (Figure 4.3.13) and partial autocorrelogram (PACF) (Figure 4.3.14) of the stationary time series (that is, of the twice differenced time series for this case) were then plotted in order to estimate the parameters (p) and (q).

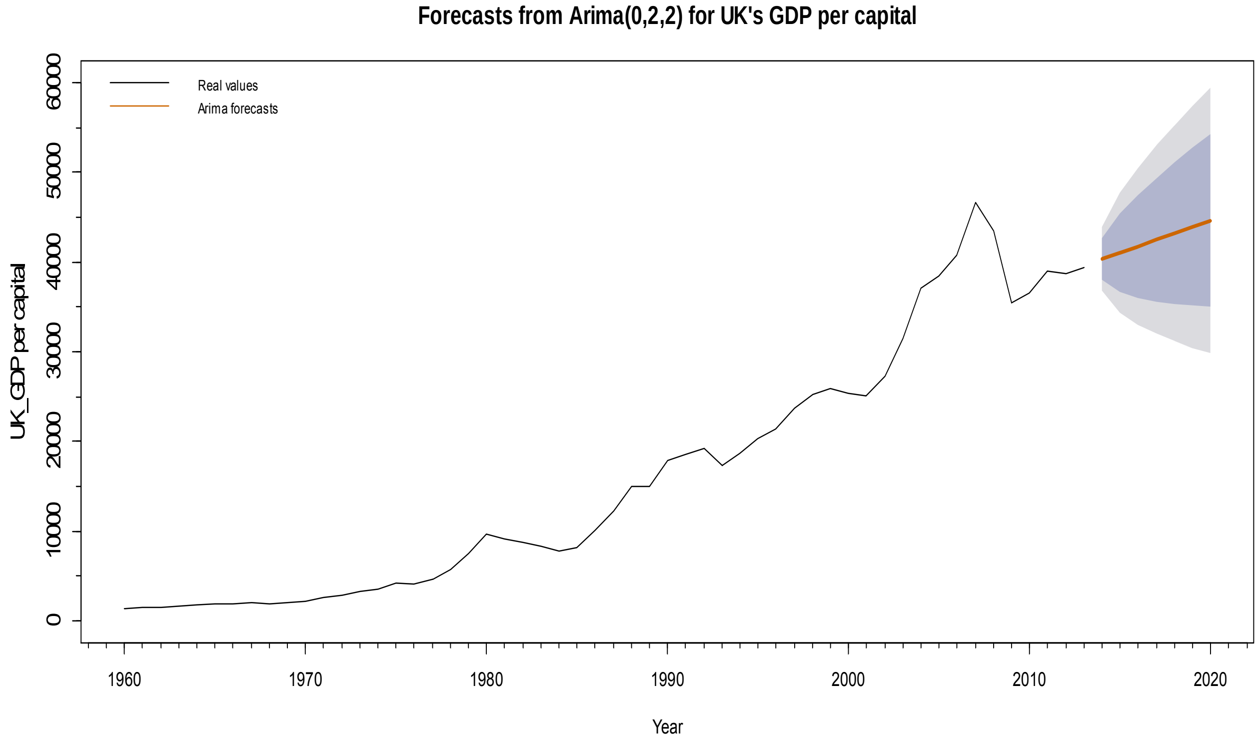


[Figure 4.3.13] – Autocorrelogram (ACF) of the twice differenced UK time series



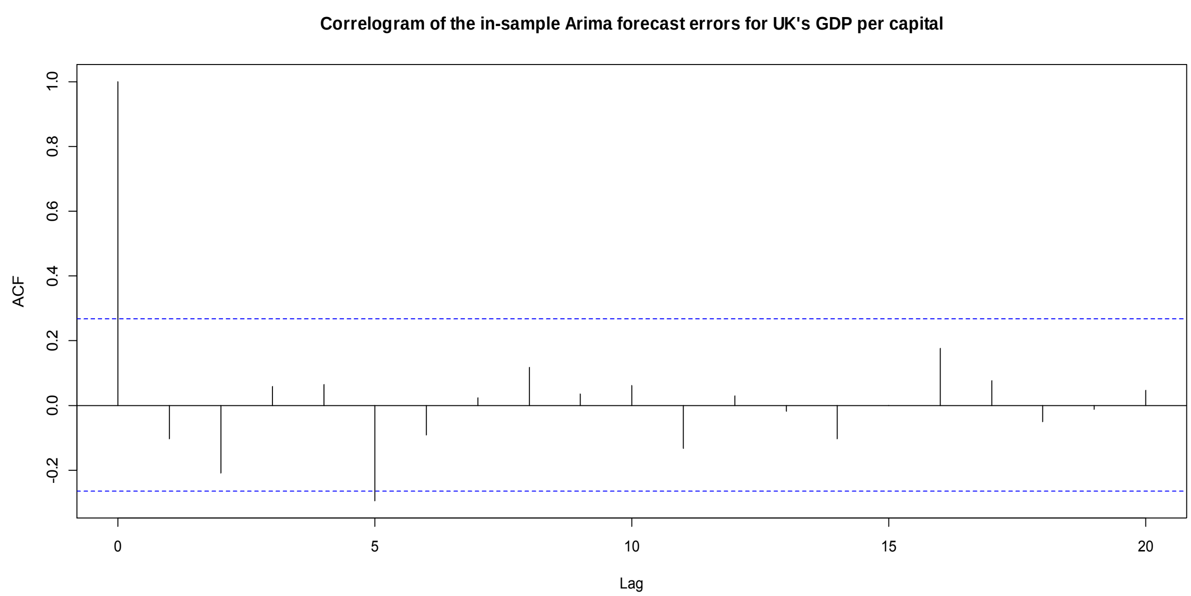
[Figure 4.3.14] – Partial autocorrelogram (PACF) of the twice differenced UK time series

The ACF graph shows that the autocorrelations are zero (practically smaller than the blue lines that represent the significant boundaries) after the 2nd lag, so it is q=2. From the PACF graph we see that the autocorrelations could be considered zero after lag 7, so it would be p=7. That means that the ARIMA model could be of order (p,d,q) = (7,2,2) which is a very complex model. Because this model has too many parameters and in an effort to follow the principle of parsimony, we chose instead to apply the model ARIMA(0,2,2). Applying this model we get the results shown in Figure 4.3.15, where we see the real values (black line), the forecasts (orange line) and the 80% and the 95% prediction intervals (blue shaded areas).



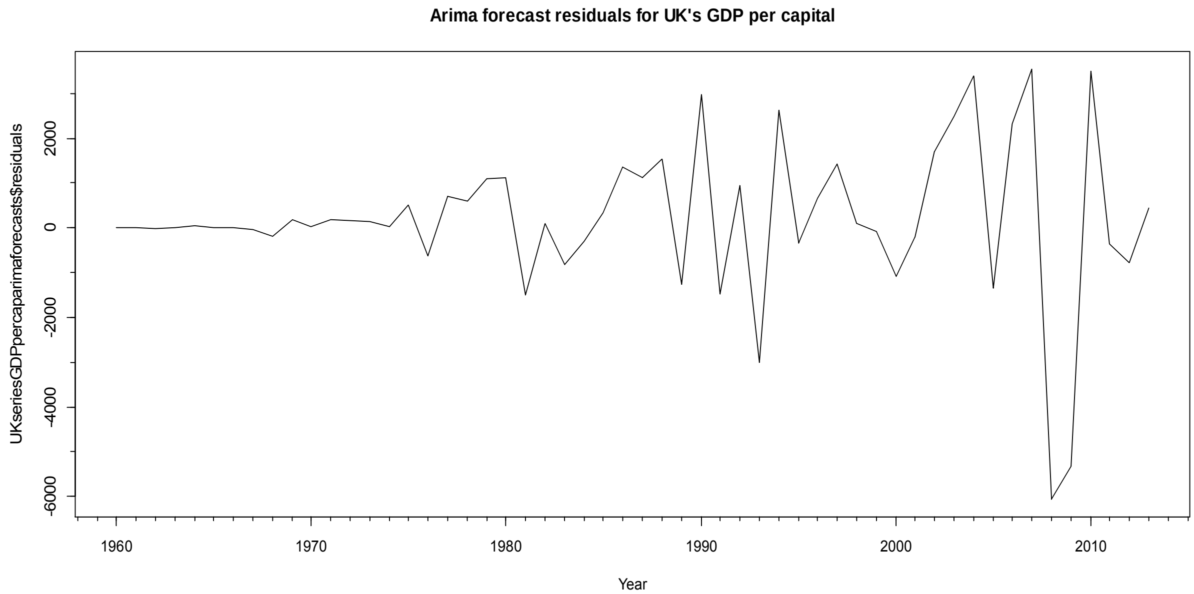
[Figure 4.3.15] – Analysis for UK, GDP per capita and whole dataset

In order to examine the residuals (in-sample errors) of the forecasts the correlogram of the in-sample forecast errors is plotted (Figure 4.3.16).

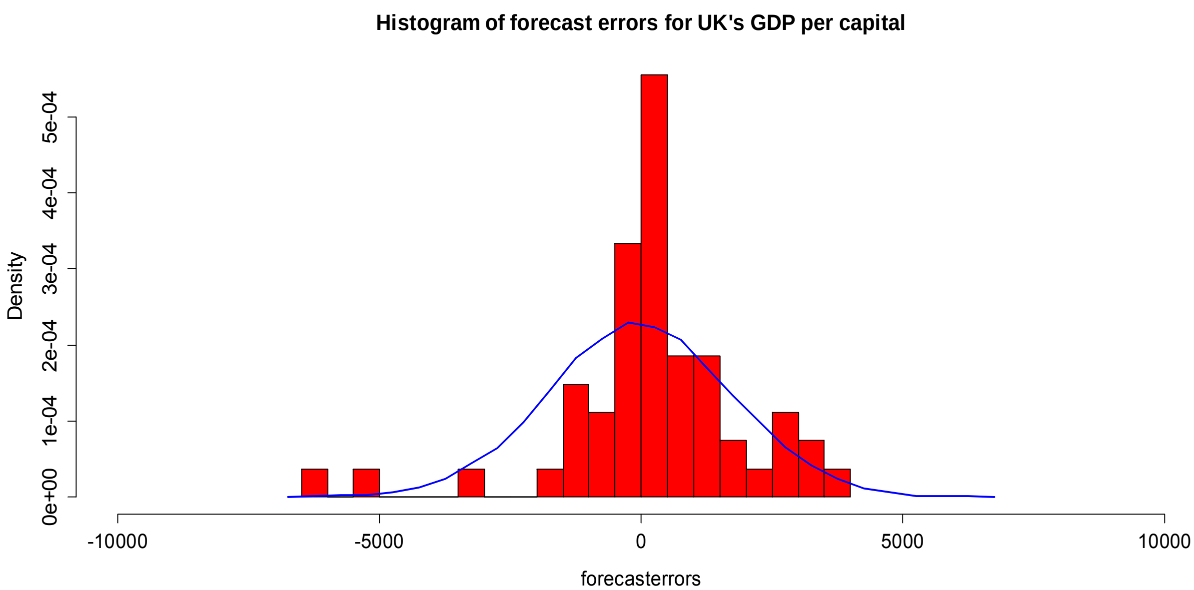


[Figure 4.3.16] – Correlogram of in-sample errors of ARIMA forecasts

From this we can see that the autocorrelations are practically zero as only at lag 5 they slightly exceed the significant boundaries (blue lines). Also, with the Box-Ljung test, the p-value was calculated to be p-value = 0. 7007, so there is little evidence that the residuals have non-zero autocorrelations. Finally, by plotting the errors (Figure 4.3.17) and making a histogram of their distribution (Figure 4.3.18) it is checked if the forecast errors have a mean zero and are normally distributed.



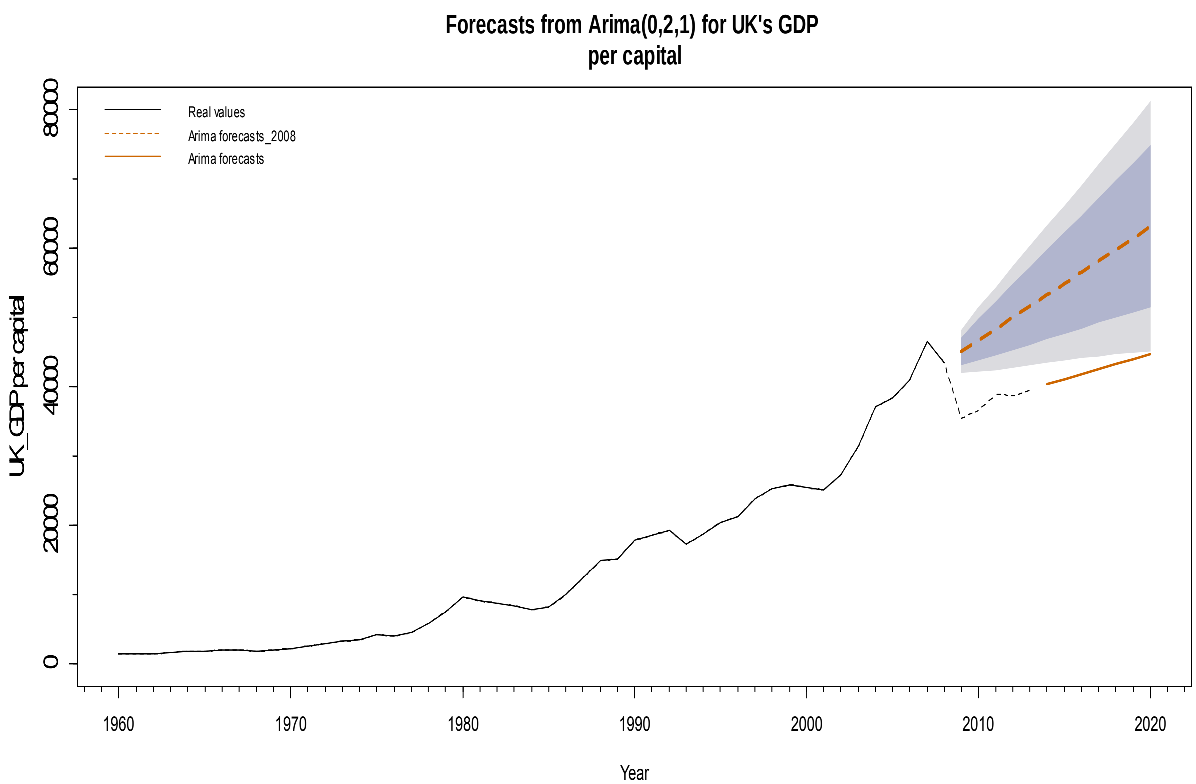
[Figure 4.3.17] –Residuals of ARIMA forecasts



[Figure 4.3.18] – Histogram and distribution of forecast residuals

The absence of autocorrelations for the forecast errors, the high p-value and the fact the errors are normally distributed with a mean around zero allow us to assume that the model is adequate and the confidence that the forecasts will be accurate is significant.

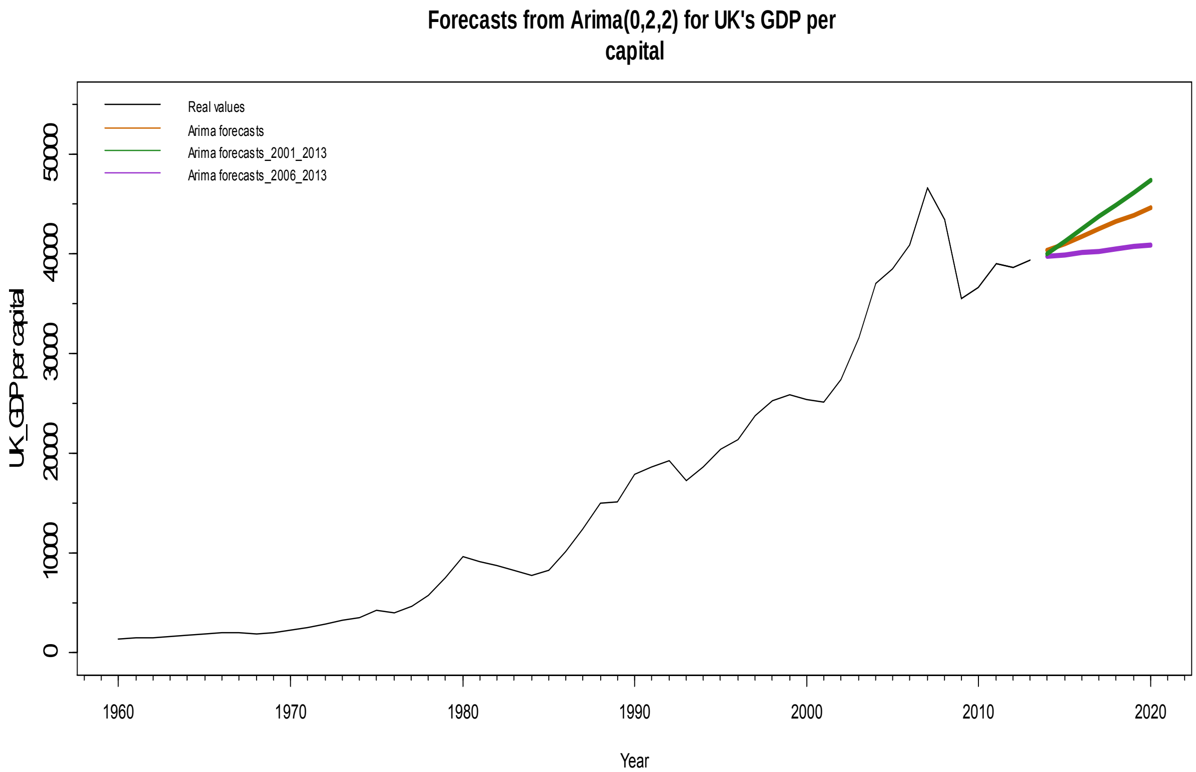
The second analysis was to apply the algorithm to the dataset up to year 2008. With the similar procedure as before, it was found that the appropriate model for this case was the ARIMA(0,2,1). The results are shown in Figure 4.3.19, where the orange dashed line gives the forecasts for the dataset up to 2008, the black dashed line the real values form 2009 to 2013 and the orange solid line gives the forecasts from the whole dataset.



[Figure 4.3.19] – Analysis for UK, GDP per capita and the dataset up to 2008

As we can see the result is similar to the one we got with the Holt's Exponential smoothing analysis. We can see that the forecasts from 2008, although they have similar slope with the real values, they overestimate the values of GDP as they could not predict the abrupt drop in GDP that happened immediately after year 2008. Again, we can conclude that some events can influence the financial data series in a way that is not easy to be predicted by the algorithms, as the change they bring is totally different from the previous existing pattern.

Finally, as in the previous cases, the ARIMA model was applied to the two smaller subsets of data from year 2001 to year 2013 and from year 2006 to year 2013. In this case the model that was the most appropriate was the ARIMA(0,2,2). The results are shown together in the same graph (Figure 4.3.20), where the dark green line shows the forecasts for the set 2001-2013 and the dark purple line for the set 2006-2013. The orange line is the forecasts for the whole dataset.

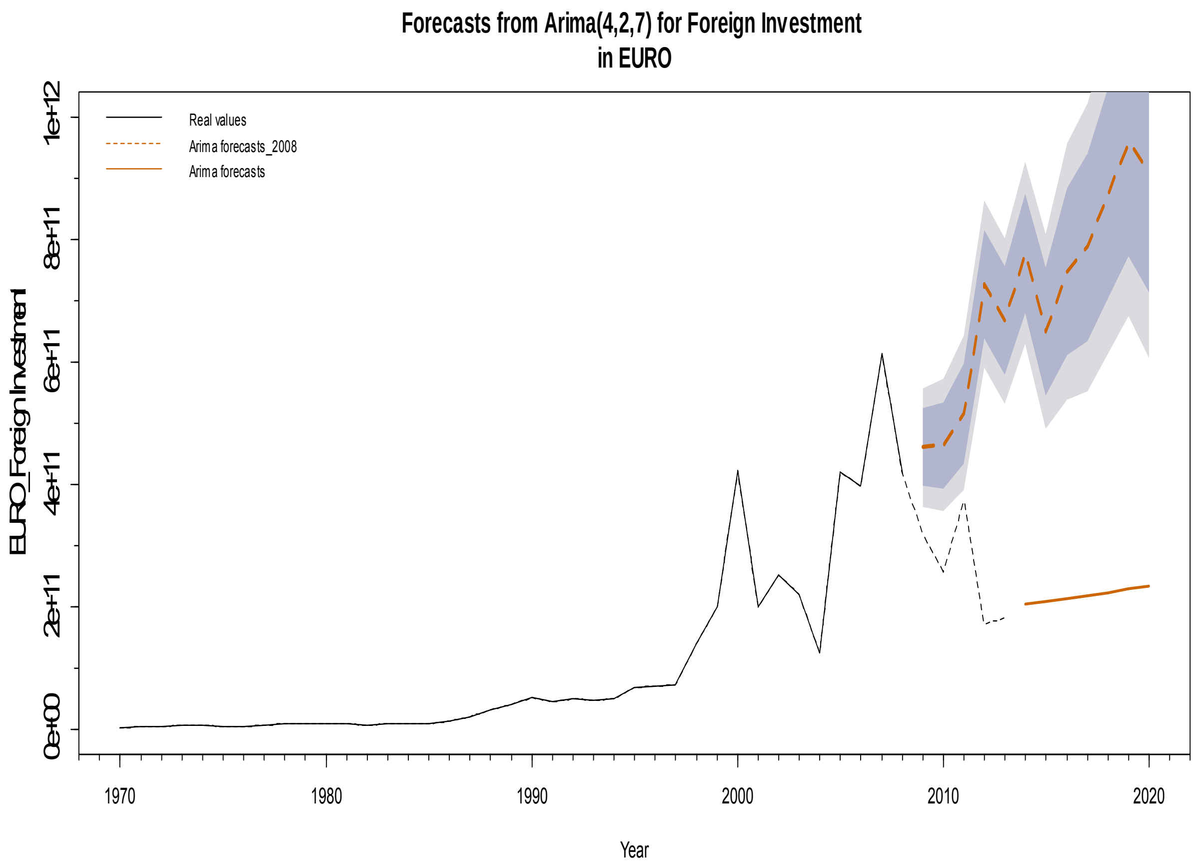


[Figure 4.3.20] – Analyses for UK, GDP per capita and the subsets 2001-2013 and 2006-2013

From the graph we can see that the ARIMA model for the whole dataset and the model for the subset 2001-2013 give similar results, while the analysis for the set 2006-2013 seem to underestimate a little the values of UK's GDP, although still they are not so far from the other two. Overall, it seems that ARIMA gives satisfactory and reliable predictions without depending so much on the selection of the dataset that is used in the analysis.

4.3.3.Analysis for Foreign Investment in EURO zone

At this point, for the ARIMA model, an additional very interesting example is presented in order to emphasize the complexity, but also the flexibility of the model compared with the two previous algorithms. The initial analysis for the Foreign Investment in EURO zone for the dataset until the year 2008, gave a suggested ARIMA(4,2,7) model. The forecasts with this model can be seen in the Figure 4.3.21.



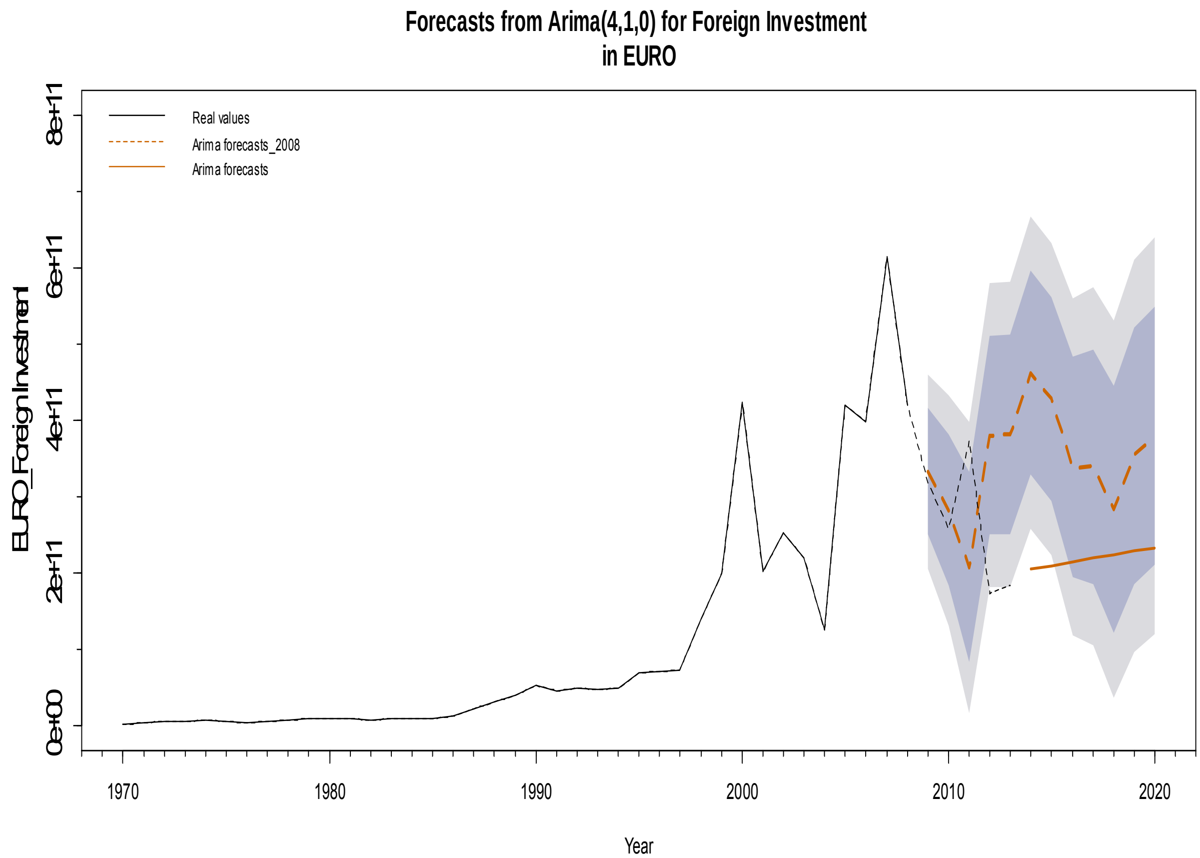
[Figure 4.3.21] – Analysis for EURO zone, Foreign Investment and the dataset up to 2008 with ARIMA(4,2,7)

A first interesting remark is that the predictions from this complex model are not linear as they were with Linear Regression and Holt's Exponential smoothing, but present fluctuations, following the pattern of the real data. The predictions, though, that this model gave are very similar to Holt's Exponential smoothing for the specific case (in terms of their values), but very far from the real values.

In an effort to identify a model that would give better predictions, different combinations of parameters (p,d,q) were applied. One of these, the ARIMA(4,1,0) based on the one time differenced time series gave the forecasts that are shown in Figure 4.3.22. It can be seen that this model gives predictions that are much closer to the real values and have a more similar pattern.

Therefore, it is very important to emphasize the significance of identifying the appropriate parameters for the ARIMA model. Even a small change in the value of a parameter can have significant change in the form and values of the forecasts. For this, great attention, experience and experimentation is necessary during the parameters selection process.

On the other hand, it can be seen that ARIMA is a powerful tool, as it offers not only the opportunity for experimentation but also the chance to identify a combination of parameters that fit the real values in the best way by giving non-linear forecasts.



[Figure 4.3.22] – Analysis for EURO zone, Foreign Investment and the dataset up to 2008 with ARIMA(4,1,0)

Chapter 5 – Comparison of Results and Discussion

As has been mentioned already the aim of the project is twofold. First of all, it aims to compare emerging and developed markets and to identify the real potential of emerging economies. Secondly, it aims to compare the different algorithms used in the analysis, identify their reliability and their limitations and attempt to extract general conclusions that would assist an analyst to choose the best algorithm for his/her predictions. In order to succeed in these aims, it was necessary to compare the results for the different countries and for the different algorithms. For this, a series of diagrams were constructed based on the outcomes of the analyses, as was initially described in the relevant subsection of the Methods Chapter (p. 31) and in more detail in the following sections.

5.1. Comparison of emerging and developed markets

In order to compare the results for the emerging and developed markets, the forecasts that were made for the four countries were plotted in the same graph. Firstly, were designed the forecasts that were made by using the whole dataset of the time series and then by using the dataset from 2001-2013.

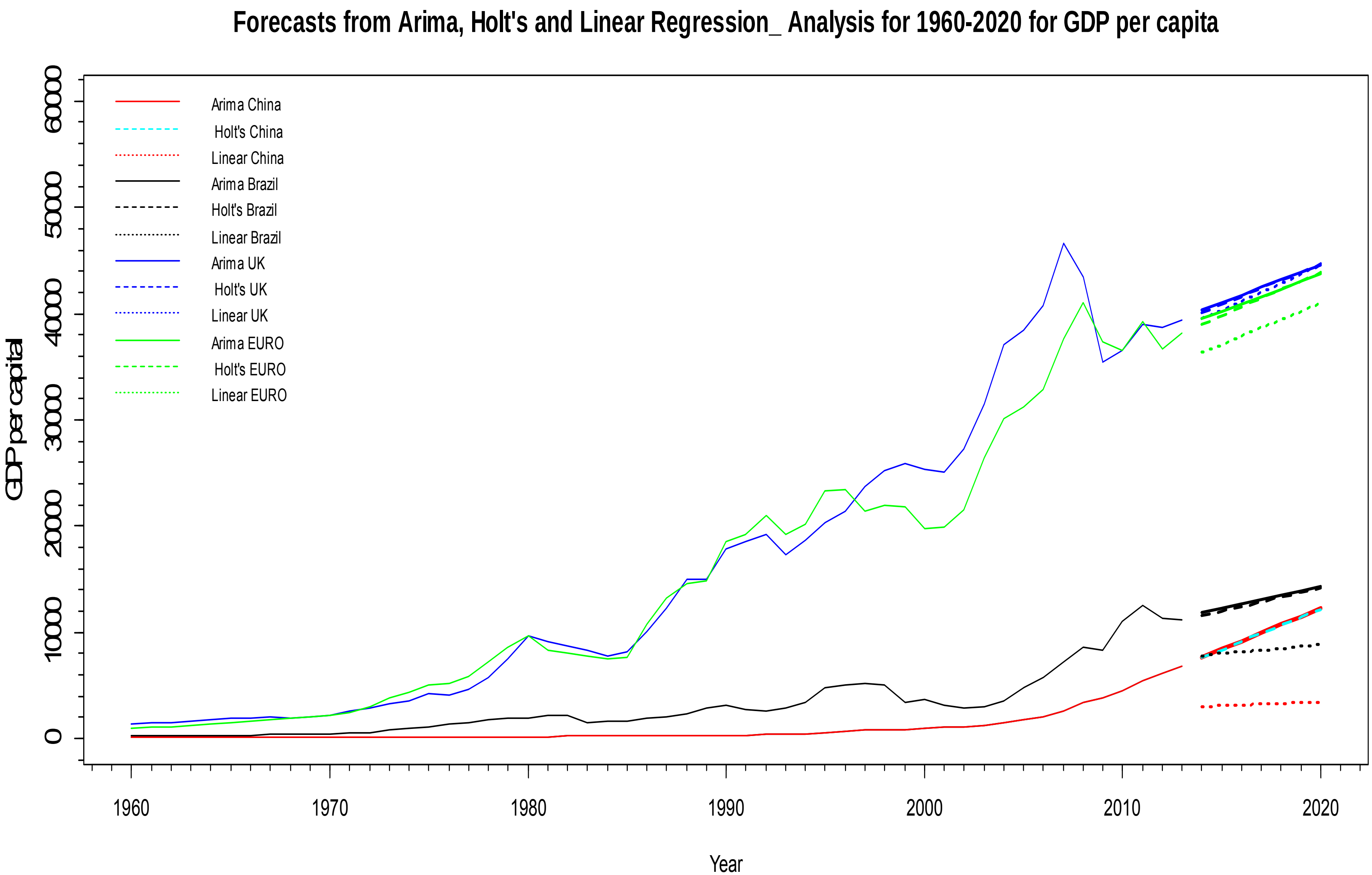
5.1.1. Comparison of different countries' forecasts made using the whole dataset

For each indicator separately, the forecasts that were made for the four countries with the three algorithms using the whole dataset were plotted in the same graph. In total the next four diagrams that are shown in Figures 5.1.1 to 5.1.4 were created.

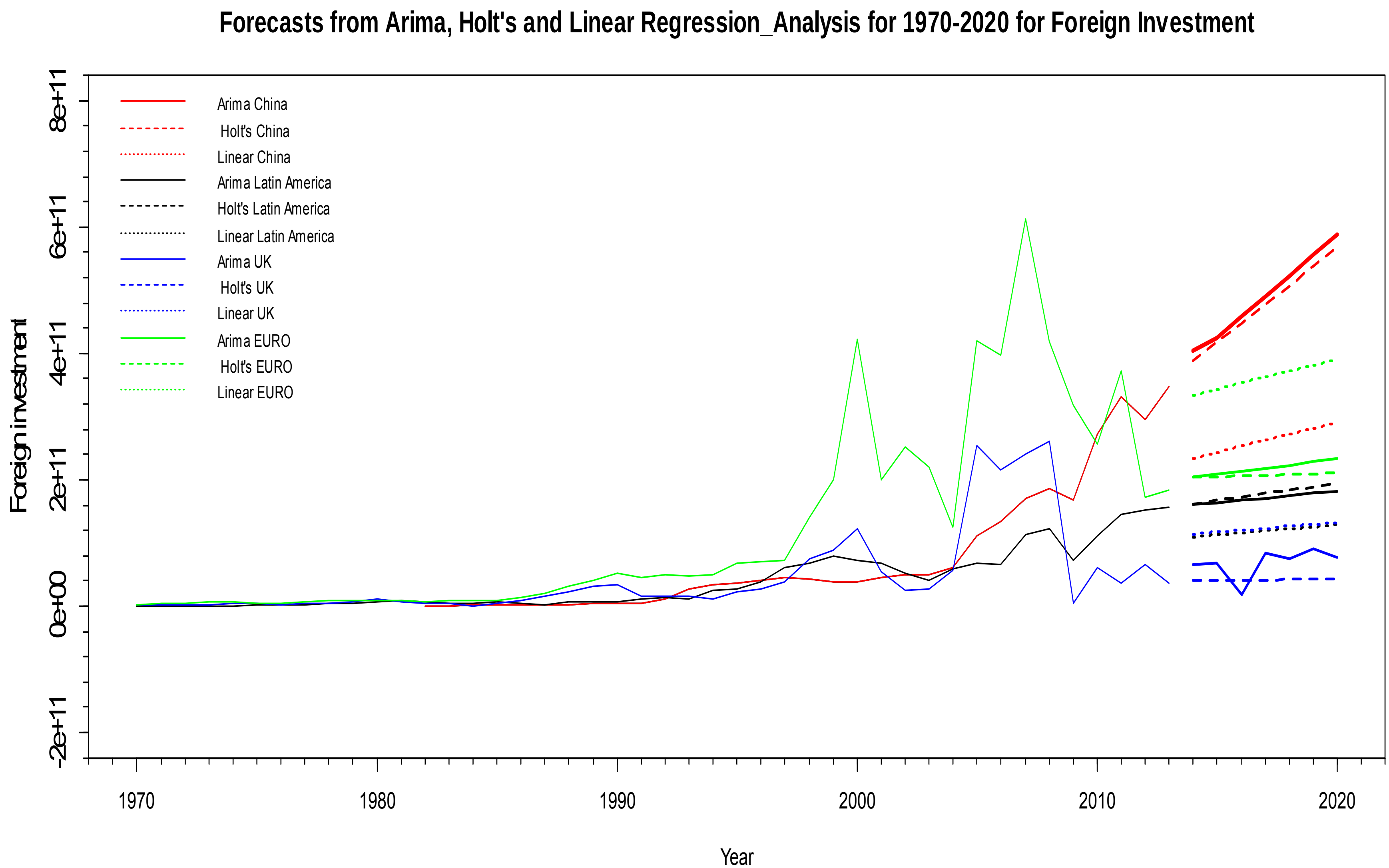
From the first graph we can observe that although the absolute values of GDP per capita for the emerging countries (China and Brazil) are much smaller than those of the developed ones (UK and EURO zone), the slopes of the forecast lines are quite similar for all countries. This means that the rate in which the GDP per capita is expected to grow every year for the next years is similar for the emerging and developed countries.

In the next graph we can see the huge increase in the Foreign Investments in China after 2008 and the simultaneous significant decrease for the UK and EURO zone. The forecasts show that China will gather the biggest amount of foreign investment the next years, both as an absolute value and as a rate of increase (as the forecast line for China is the steepest). At the same time the investments in Latin America seem to approach the ones in EURO zone and be much bigger than those in UK.

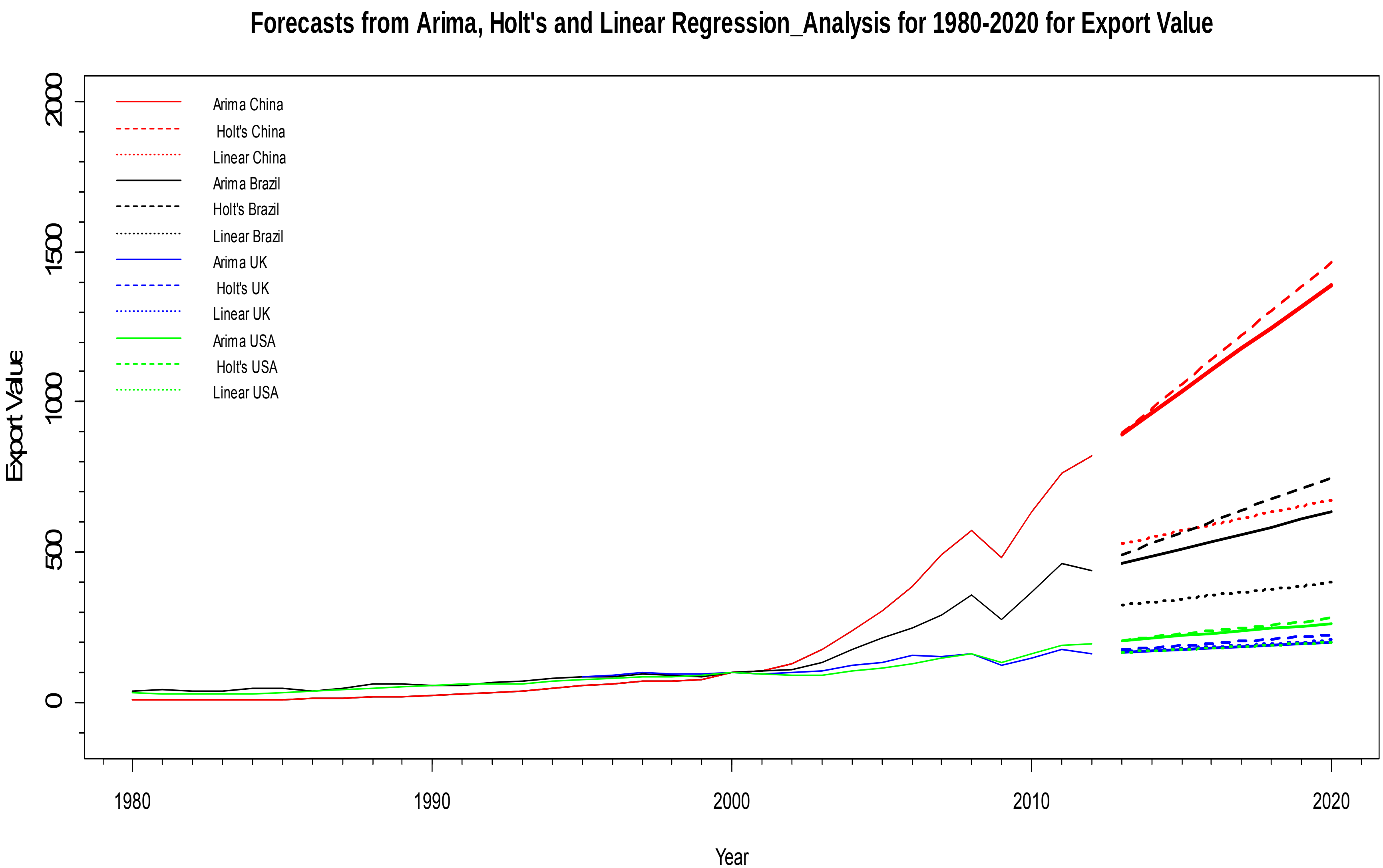
Finally, from the next two graphs we can see that although the Export and Import values were almost similar for all countries for about two decades (from 1980 to 2000), after 2000 the values for China, Brazil and India have known a great increase while for UK and USA remain relatively steady. The forecasts for China, Brazil and India give much bigger absolute values and rates of increase for these two indicators than the relevant for UK and USA.



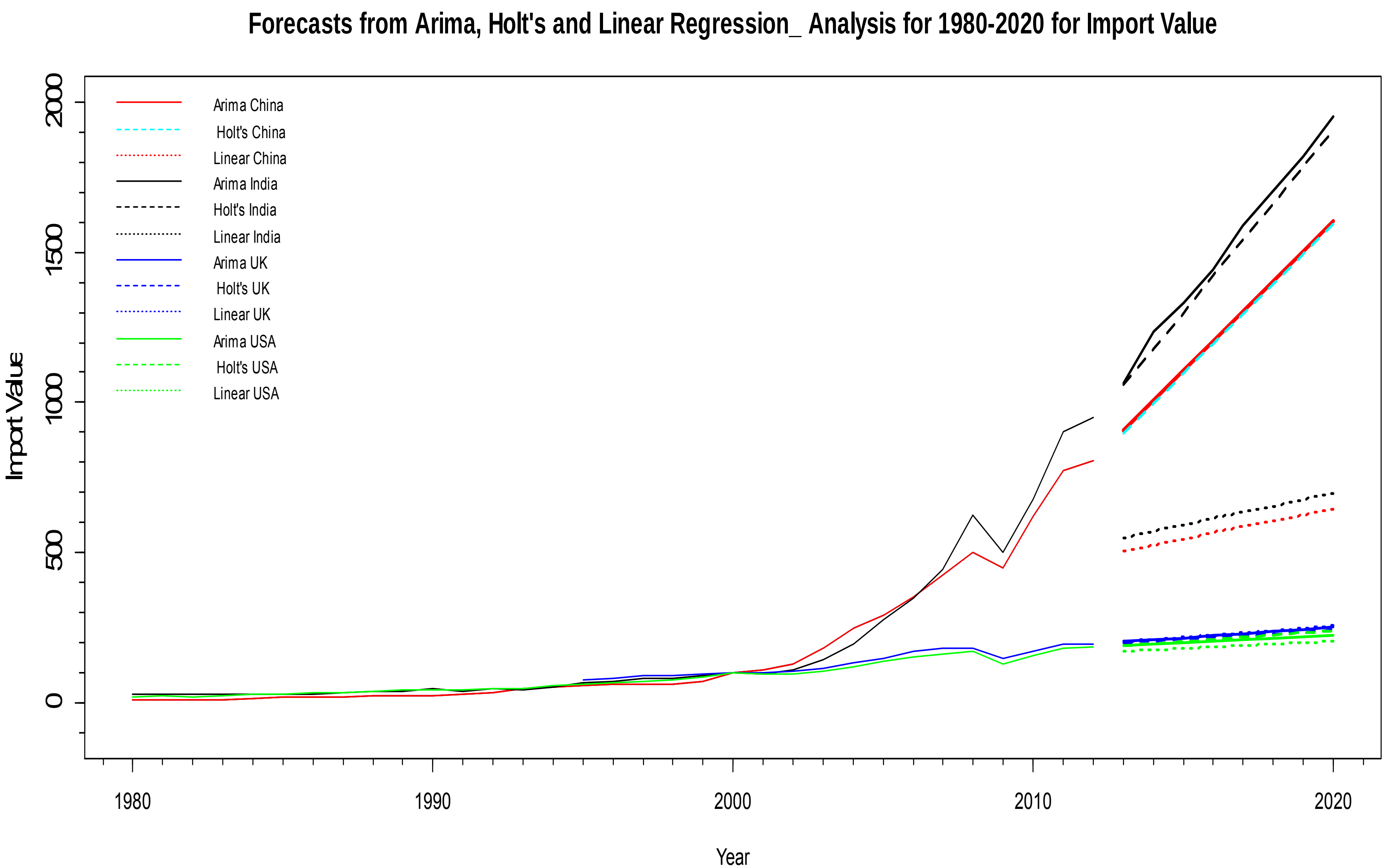
[Figure 5.1.1] – Comparison of different countries' forecasts for GDP per capita by using the whole dataset



[Figure 5.1.2] – Comparison of different countries' forecasts for Foreign Investment by using the whole dataset



[Figure 5.1.3] – Comparison of different countries' forecasts for Export Value by using the whole dataset



[Figure 5.1.4] – Comparison of different countries' forecasts for Import Value by using the whole dataset

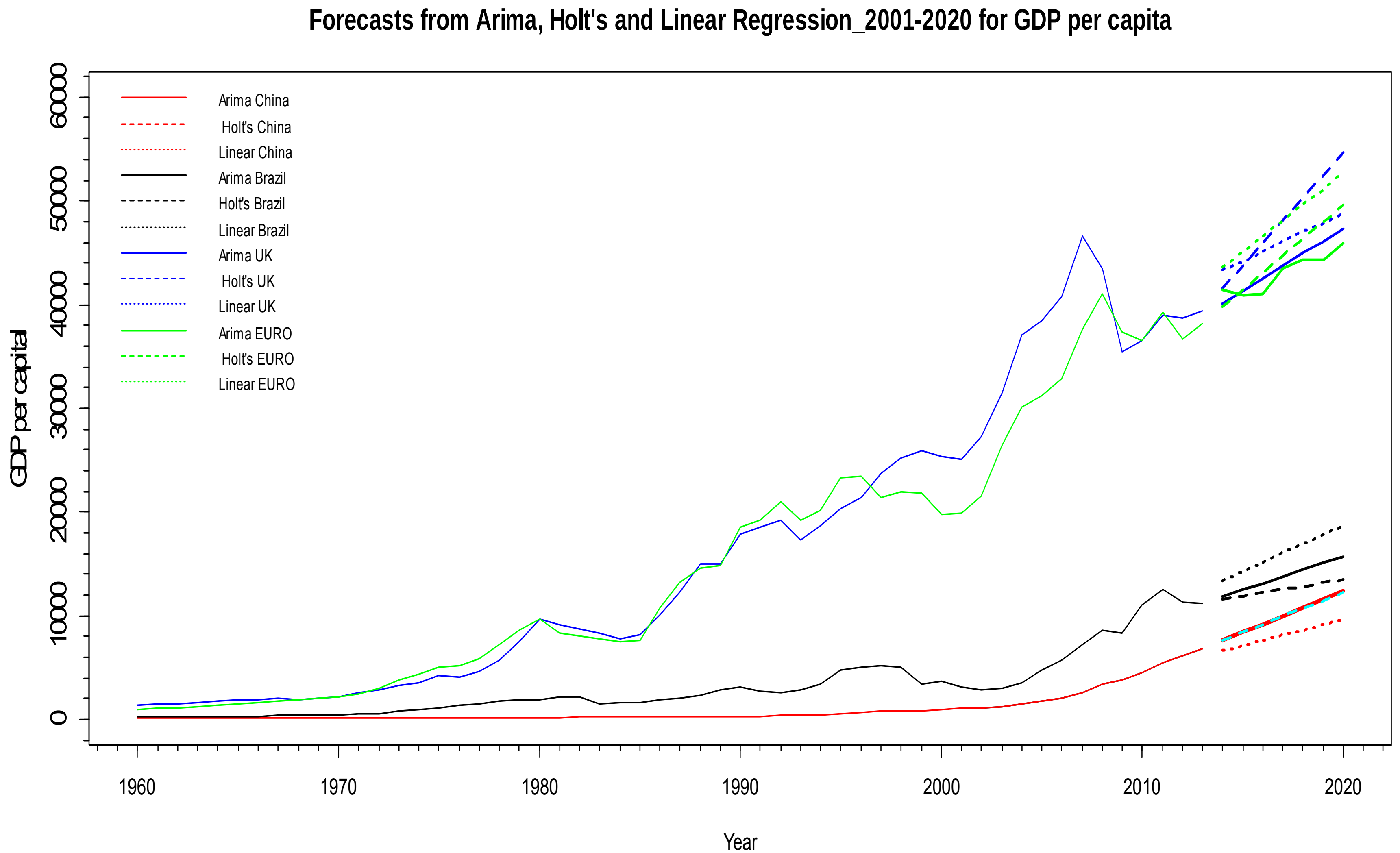
5.1.2. Comparison of different countries' forecasts made using the dataset from year 2001 to year 2013

In this case there were plotted in the same graph for each indicator, the forecasts that were made for the four countries with the three algorithms using the dataset from year 2001 to year 2013. In total the next four diagrams that are shown in Figures 5.1.5 to 5.1.8 were created.

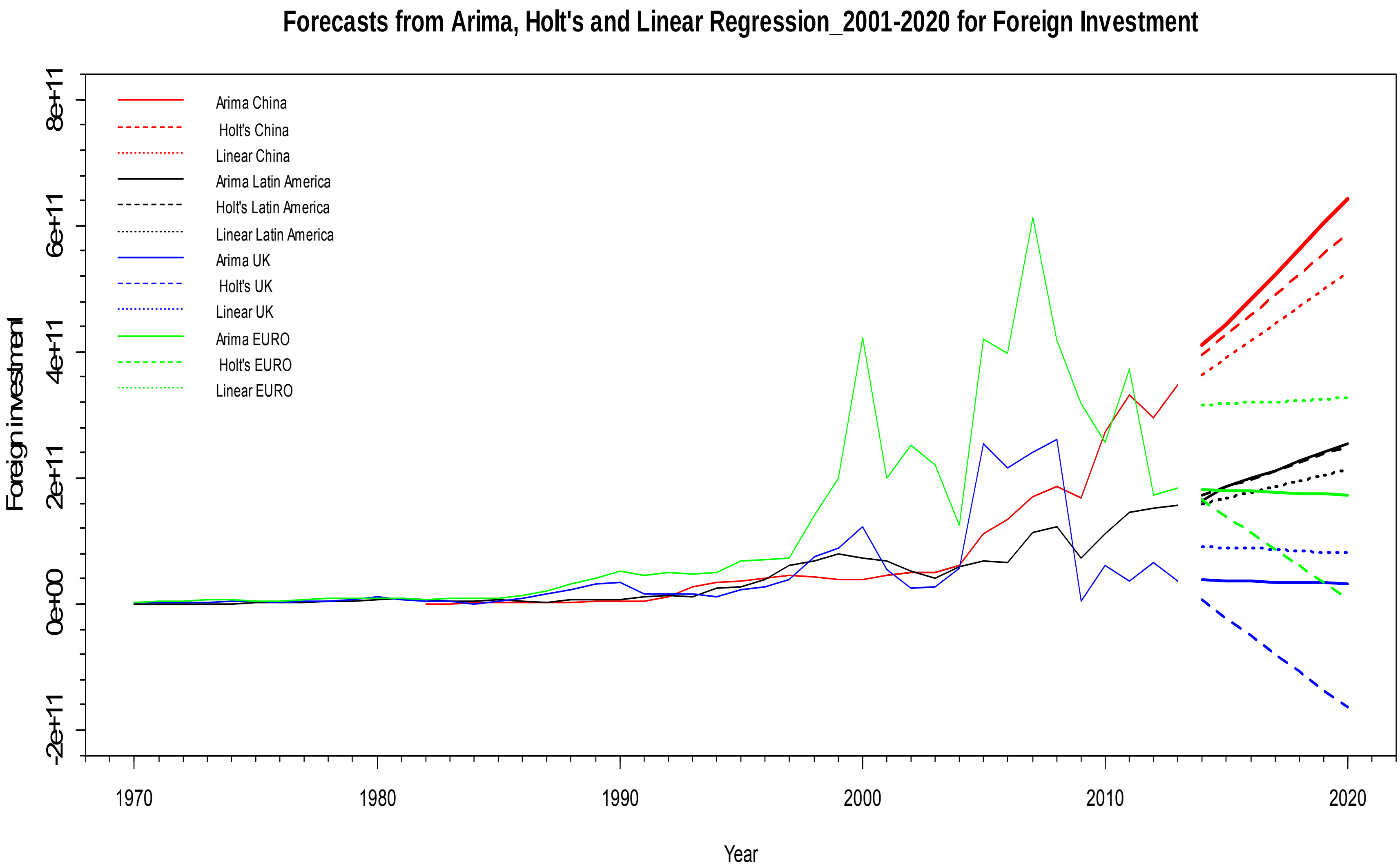
The results are very similar with the ones that the analysis with the whole dataset gave, but in this case the absolute values and the slopes of forecasts lines are even bigger for the emerging markets, especially for the Linear Regression analysis. This time for the Foreign Investment indicator the values for Latin America are bigger even from the EURO zone, as both the EURO zone and the UK present negative slopes.

In the rest of the graphs, for the GDP per capita, the emerging countries (China and Brazil) have again slopes of the forecast lines that are similar to the developed countries. As for the Export and Import value the forecasts for China, Brazil and India give again much bigger absolute values and rates of increase for these two indicators than the relevant for UK and USA.

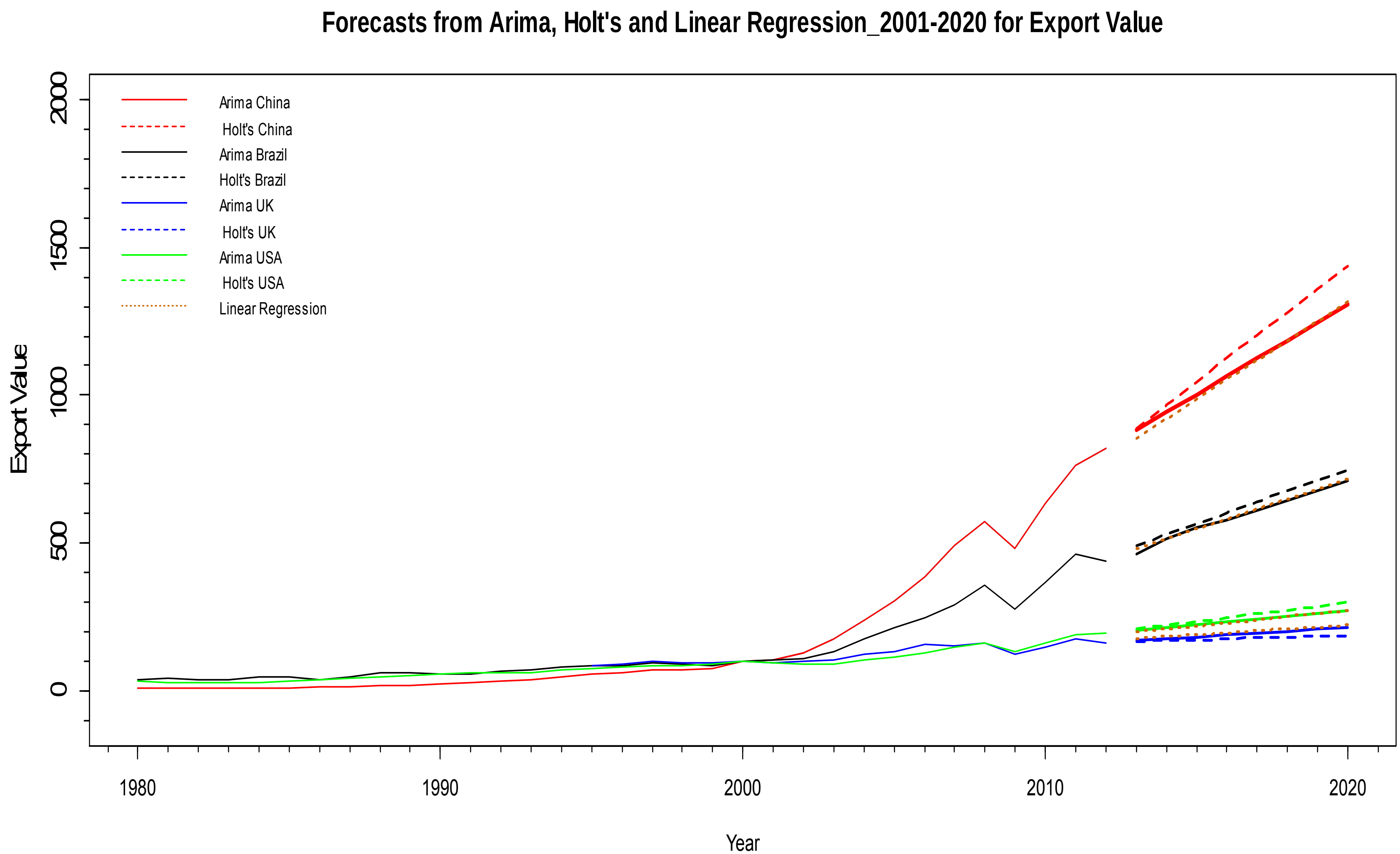
From both analyses, it is obvious that the forecasts for the emerging markets present higher absolute values and rates of growth for the three indicators (Foreign Investment, Export value, Import value) and absolute values that are smaller than those of the developed countries but similar rates of increase for one indicator (GDP per capita). The GDP for the emerging markets though, is logically expected to be smaller from the GDP of the developed countries as the difference is very big. That what is of interest is that the rate of growth now is similar to the developed countries even for the GDP per capita. For the other indicators, as already mentioned, the rates of growth and the potential of the emerging countries are bigger than those of the developed economies in all three algorithms.



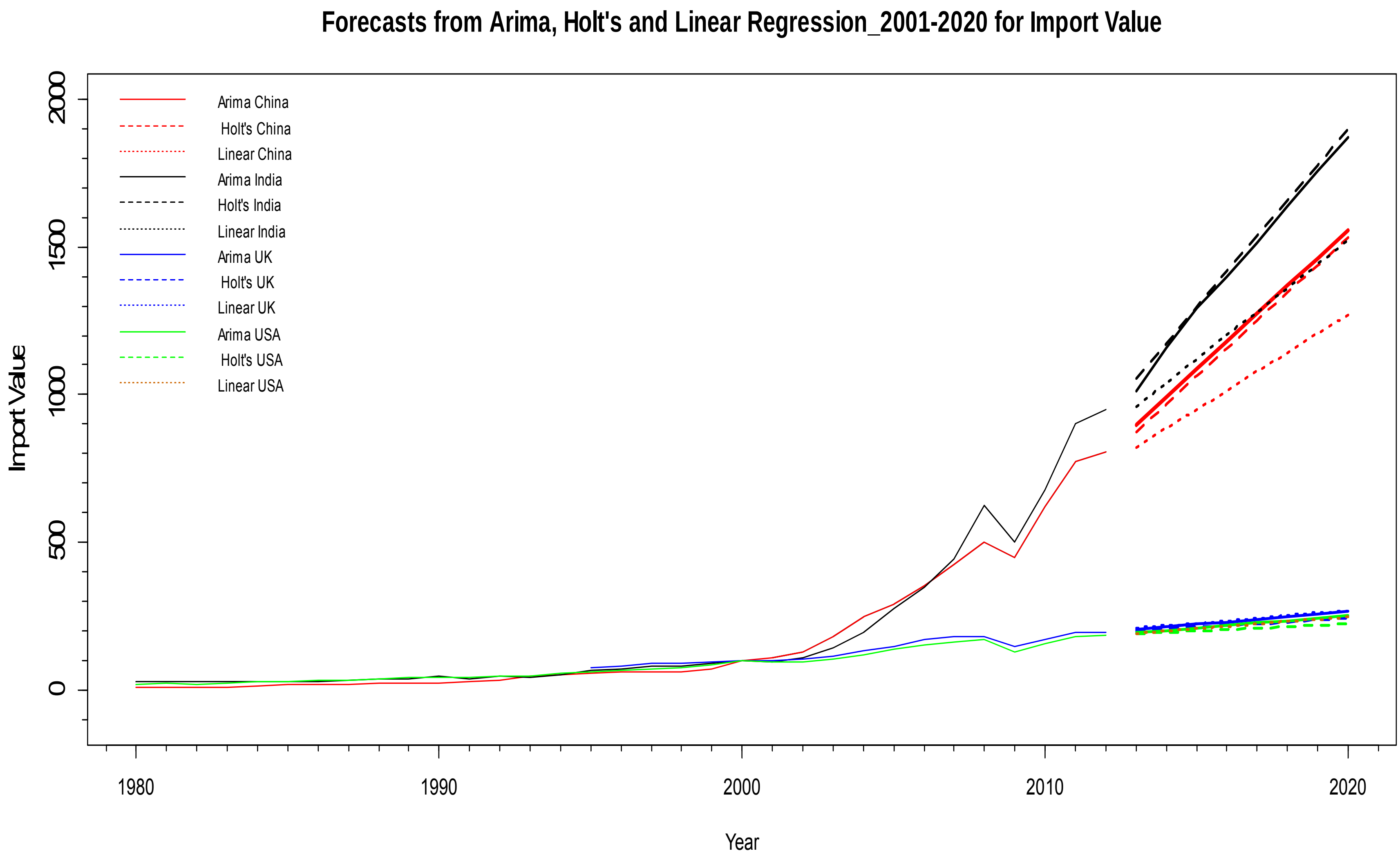
[Figure 5.1.5] – Comparison of different countries' forecasts for GDP per capita by using the dataset from year 2001 to year 2013



[Figure 5.1.6] – Comparison of different countries' forecasts for Foreign Investment by using the dataset from year 2001 to year 2013



[Figure 5.1.7] – Comparison of different countries' forecasts for Export Value by using the dataset from year 2001 to year 2013



[Figure 5.1.8] – Comparison of different countries' forecasts for Import Value by using the dataset from year 2001 to year 2013

5.1.3. Comparison of different countries' coefficient of slopes from Linear Regression

The Table 5.1.1 presents the coefficients of slopes for all analyses that were conducted with Linear Regression for every indicator, country and different dataset. The coefficients are calculated from R with the lm() function for every analysis and are equal to the average increase or decrease of each indicator in one year. In this sense, they represent the rate of growth for a country and are very useful for the comparison between emerging and developed countries.

The same coefficients are visualized for each indicator in graphs in Figures 5.1.9 to 5.1.12 for a more efficient comparison.

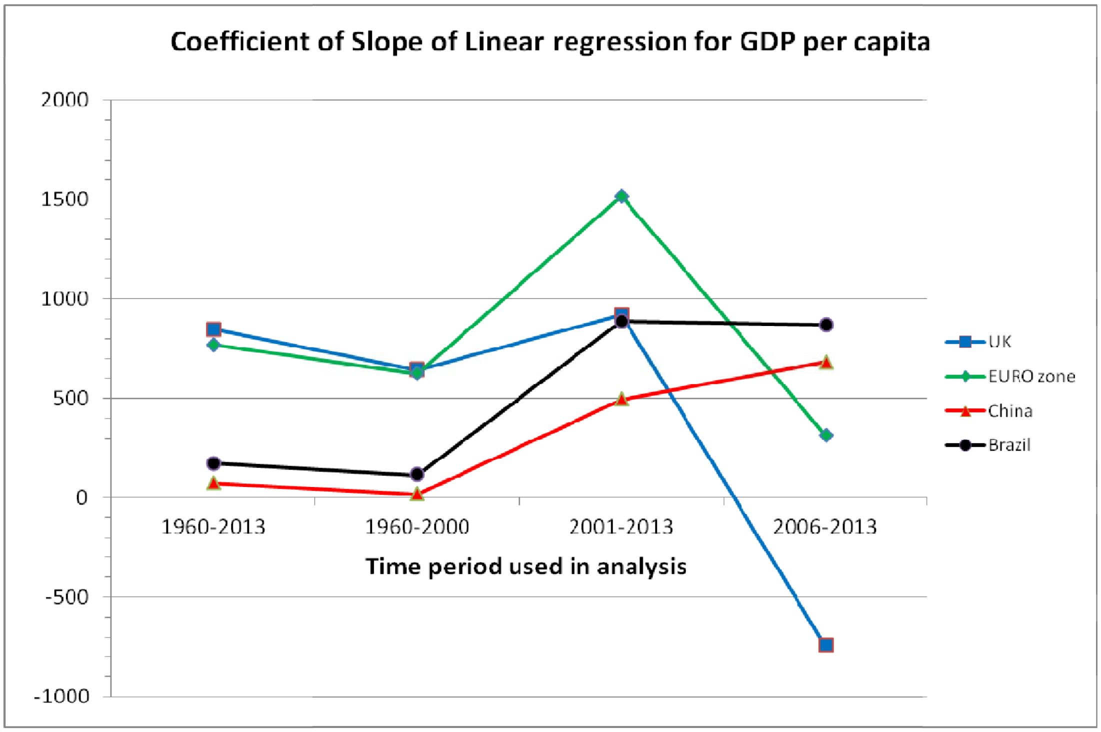
From the Table and the graphs we can see again that for the three indicators (Foreign Investment, Export Value, Import Value) the coefficients of emerging markets are always similar or higher than those of the developed countries and only for GDP per capita and the analysis of the whole dataset the developed economies have much bigger values of slopes. Even in this case though, when the analysis was made with the dataset 2001-2013 the emerging markets have started to balance the difference and in the analysis for subset 2006-2013 they present higher rates of growth.

Another important remark has to do with the analysis for the two subsets, from 1960 to 2000 and from 2001 to 2013. When examining the coefficients for these two cases, it is very obvious the very big increase in the values of the emerging markets in the subset 2001 to 2013. For instance, for the Export and Import Value indicators, the values of the coefficient for the developed countries for the analysis from 2001 to 2013 are 2 to 3 times bigger than the analysis for the subset 1980 to 2013. For the emerging markets, the relevant values are 10 to 25 times bigger, indicating the huge increase in the rate of growth for these countries after the year 2000.

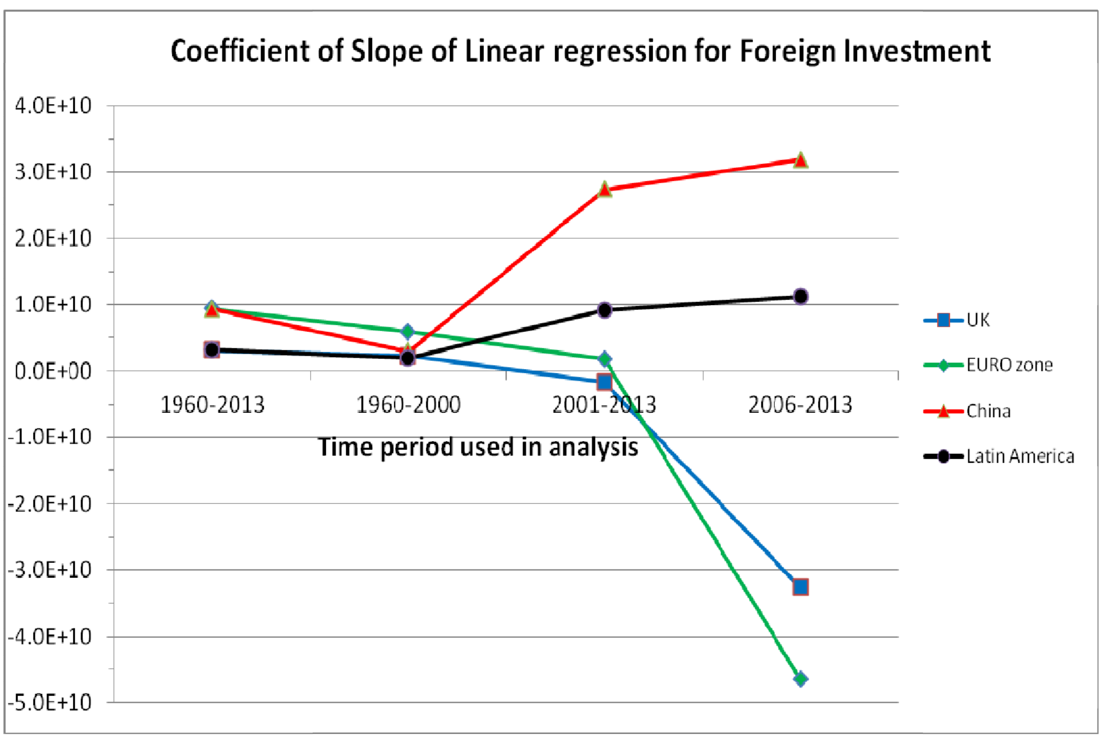
Finally, the analysis for the dataset from year 2006 to year 2013, give even higher rates for the emerging markets, while the developed countries have even smaller values and in some cases they present negative rates. It is therefore clear from all graphs that the developed countries the last years have stable or declining rates of growth, while the emerging markets have known a big increase on their rate of development after the year 2000 and justify in this way their name as emerging economies.

Coefficient of Slope of Linear regression					
Indicator	Country	Time period used in analysis			
		1960-2013	1960-2000	2001-2013	2006-2013
GDP per capita	UK	846.265	641.465	920.076	-743.709
	EURO zone	768.400	622.900	1516.000	312.600
	China	73.219	16.930	496.066	680.479
	Brazil	170.385	113.522	886.748	869.485
		1970-2013	1970-2000	2001-2013	2006-2013
Foreign Investment	UK	3.09E+09	2.20E+09	-1.71E+09	-3.27E+10
	EURO zone	9.43E+09	5.90E+09	1.74E+09	-4.66E+10
	China	9.37E+09	2.94E+09	2.74E+10	3.18E+10
	Latin America	3.15E+09	1.96E+09	9.20E+09	1.12E+10
		1980-2012	1980-2000	2001-2012	2006-2012
Export value	UK	5.161	2.707	6.360	1.784
	USA	4.712	3.681	10.040	10.140
	China	20.700	4.300	66.000	67.970
	Brazil	10.932	3.274	33.200	32.990
		1980-2012	1980-2000	2001-2012	2006-2012
Import value	UK	7.491	4.344	8.514	3.046
	USA	5.219	3.458	7.992	4.375
	China	19.710	3.928	63.780	78.060
	India	21.110	3.455	80.190	99.280

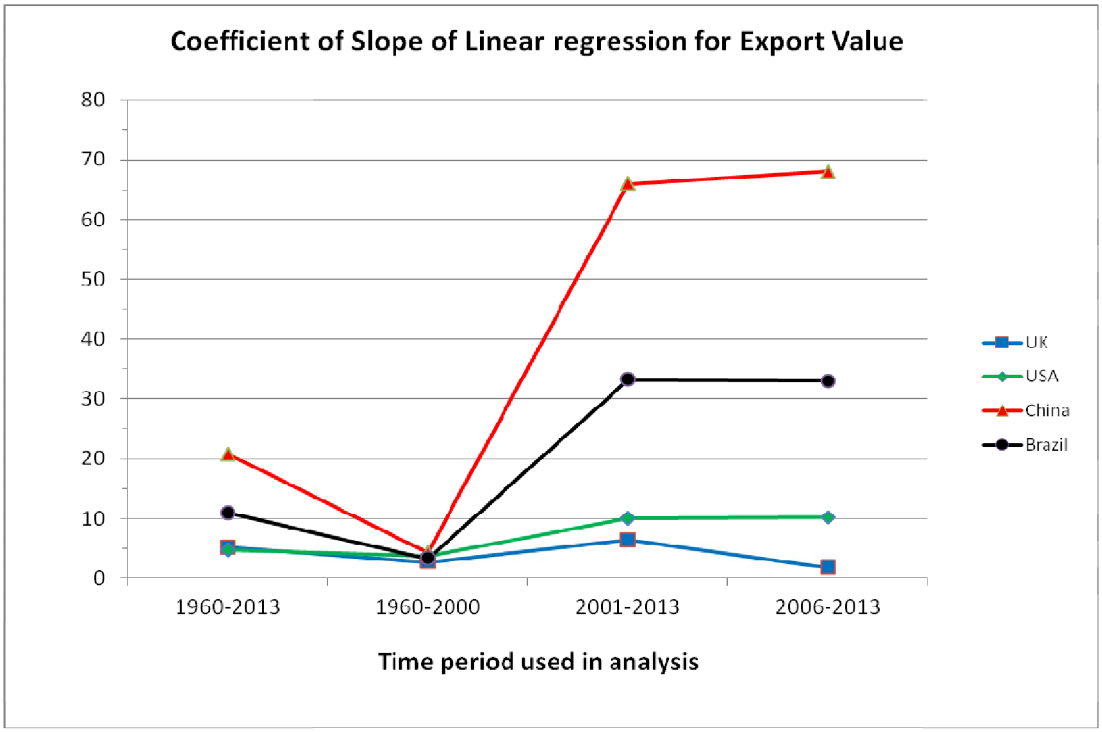
[Table 5.1.1] – Coefficients of slope from Linear Regression analysis



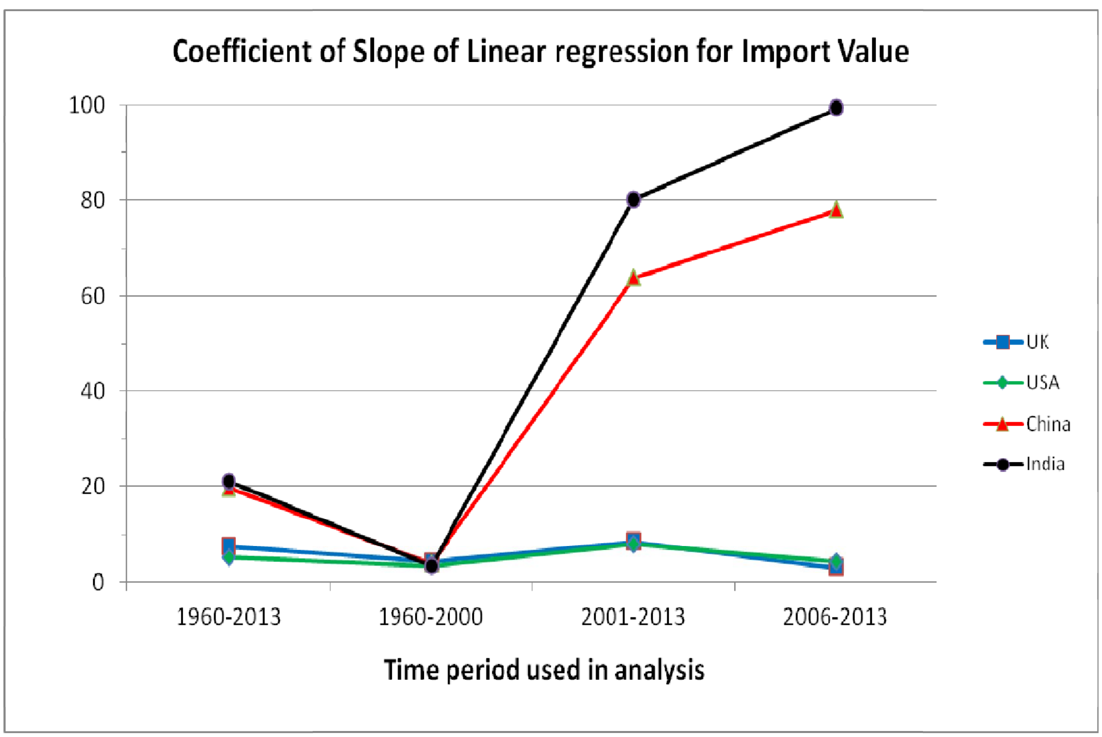
[Figure 5.1.9] – Comparison of coefficients of slope of countries' forecasts for GDP per capita



[Figure 5.1.10] – Comparison of coefficients of slope of countries' forecasts for Foreign Investment



[Figure 5.1.11] – Comparison of coefficients of slope of countries' forecasts for Export Value



[Figure 5.1.12] – Comparison of coefficients of slope of countries' forecasts for Import Value

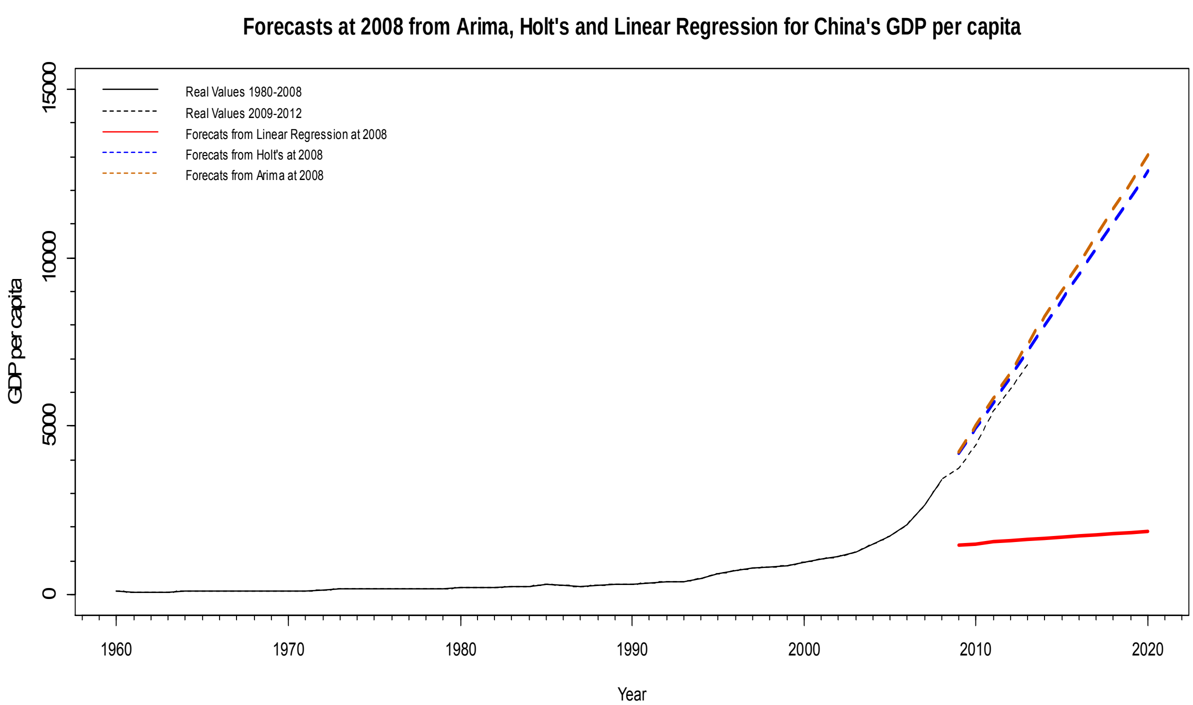
5.2. Comparison of algorithms

The next step in the comparison of the results was to evaluate the algorithms and compare them, in an effort to identify which algorithm provides the best forecasts and the appropriateness and limitations of each method.

5.2.1 Evaluation of algorithms based on the forecasts made at year 2008

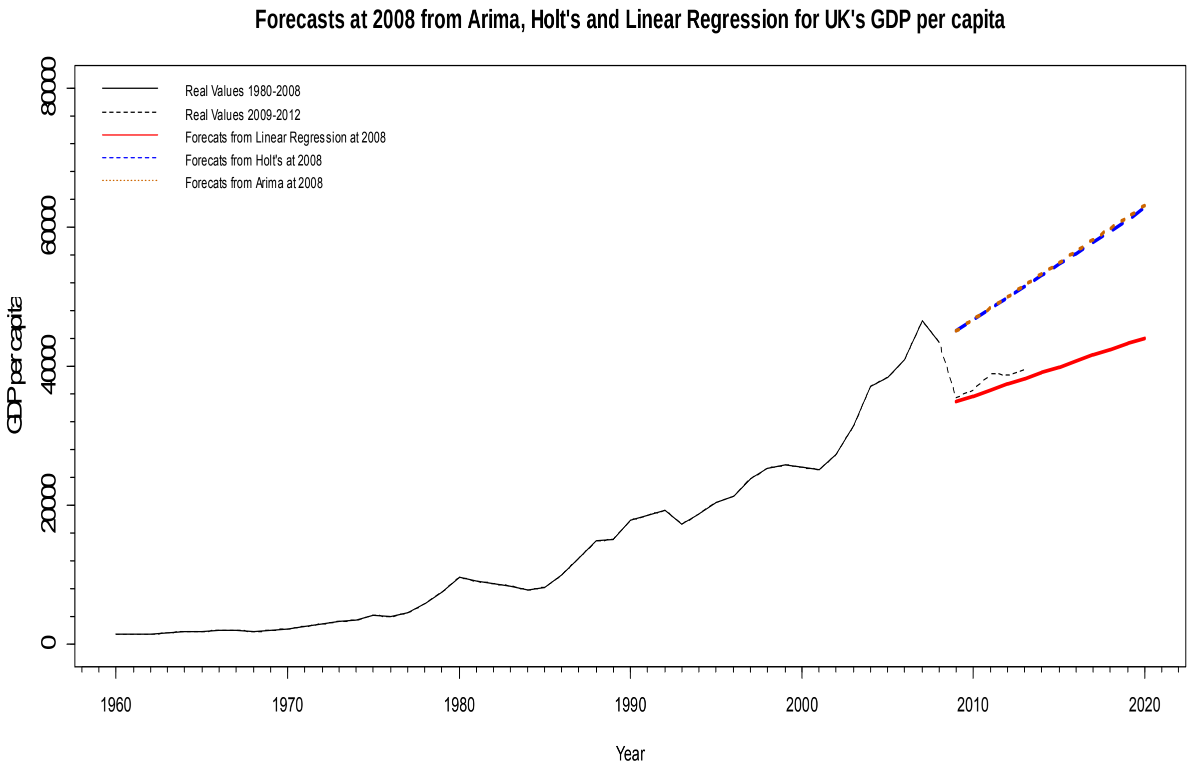
In order to evaluate the effectiveness of each method, for every indicator and every country, each algorithm was applied to the dataset until the year 2008 and the forecasts made with this analysis were compared with the real values for the year 2009 to 2013. In this way, we are able to compare the forecasts that a researcher would have done in 2008 with the real values observed the next years and practically check if the algorithm gives accurate predictions.

For each algorithm 16 analyses took place (4 indicators with 4 countries each one), so in total 3*16=48 analyses were done. For each pair of indicator and country, the analyses of the three algorithms were plotted in the same graph, so 16 graphs were constructed. In Figure 5.2.1 are shown the results for the threes analyses for China's GDP per capita. The red line is the Linear Regression forecasts, the blue line the Holt's Exponential smoothing forecasts, the orange line the predictions from ARIMA and the black dashed line, the real values for years 2009 to 2013. We can remark how well the forecasts from Holt's Exponential smoothing and ARIMA fit the real values, while the Linear Regression predictions are very far.



[Figure 5.2.1] – Comparison of algorithms' predictions at 2008 with the real values of China's GDP per capita

On the contrary, on the UK's GDP per capita graph (Figure 5.2.2) we can see that the Linear Regression predictions fit better to the real values than the other two methods that seem to overestimate the value of GDP and fail to predict the drop that took place after 2008.



[Figure 5.2.2] – Comparison of algorithms' predictions at 2008 with the real values of UK's GDP per capita

The rest 14 graphs for the other combinations of indicators and countries can be found in the Appendix D.1. (p. 290). From these graphs and their visual observation the Table 5.2.1 was constructed. In this Table every case which an algorithm provides good forecasts is shown with a check mark. As we can see the most effective algorithm is the ARIMA with a percentage of 75% successful predictions. Second is the Holt's Exponential smoothing with slightly over one out of two good predictions (56.25%) while Linear Regression gives satisfactory results only in the 31.25% of the cases.

The Table 5.2.2 gives the same results separately for emerging and developed economies. From this some interesting remarks can be made. First of all, the Linear Regression does not give satisfactory results for emerging markets in any occasion. All of its successful cases are for developed countries, where it seems that the specific time series follow a constant pattern through time. For the emerging markets that exhibit an abrupt change in their rate of growth after 2000, the Linear Regression fails to give good predictions.

Comparison of Algorithms' predictions at 2008 with the real values				
		Algorithm		
Indicator	Country	Linear Regr. 1980-2012	Holt's 1980-2012	Arima 1980-2012
GDP per capita	China		√	√
	Brazil			√
	UK	√		
	EURO zone	√		
Foreign Investment	China		√	√
	Latin America			√
	UK			
	EURO			√
Export value	China		√	√
	Brazil		√	√
	UK	√	√	√
	USA		√	√
Import value	China		√	√
	India		√	√
	UK	√	√	√
	USA	√		
	Total:	5	9	12
		31.25%	56.25%	75.00%

A √ mark indicates a good fit with the real values

[Table 5.2.1] – Effectiveness of algorithm's predictions made at 2008

	Algorithm		
	Linear Regr. 1980-2012	Holt's 1980-2012	Arima 1980-2012
Emerging Markets	0	6	8
	0.00%	75.00%	100.00%
Developed Markets	5	3	4
	62.50%	37.50%	50.00%
Total:	5	9	12
	31.25%	56.25%	75.00%

[Table 5.2.2] – Effectiveness of algorithm’s predictions made at 2008 as percentage of total analyses

On the other hand ARIMA gives successful predictions for emerging markets in all cases and also in one out of two cases for developed economies. Holt’s Exponential smoothing, although it gives good results for emerging markets in 3 out of the 4 analyses, for developed countries it is not proved so satisfactory as it has only 37.5%.

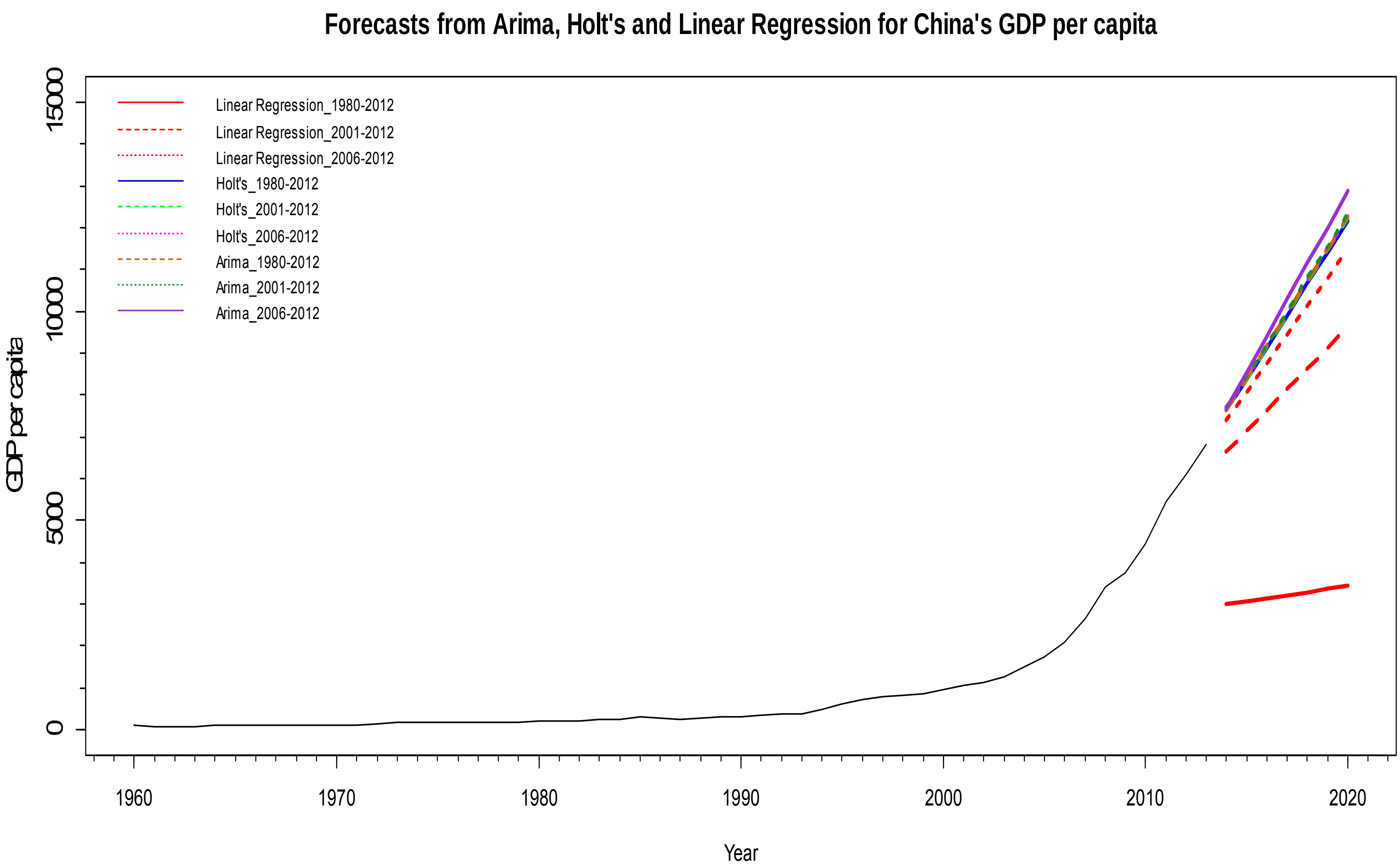
In general it seems that because the emerging markets do not have a constant trend through time, but present a big increase in their trend the last 15 years, the Linear Regression algorithm when using the whole dataset cannot give adequate forecasts. On the contrary, ARIMA and Holt’s Exponential smoothing seem to be much more suitable, as they face well this change in the slope of the emerging markets line and continue to give accurate forecasts.

5.2.2 Comparison of all algorithms' forecasts made for each country individually

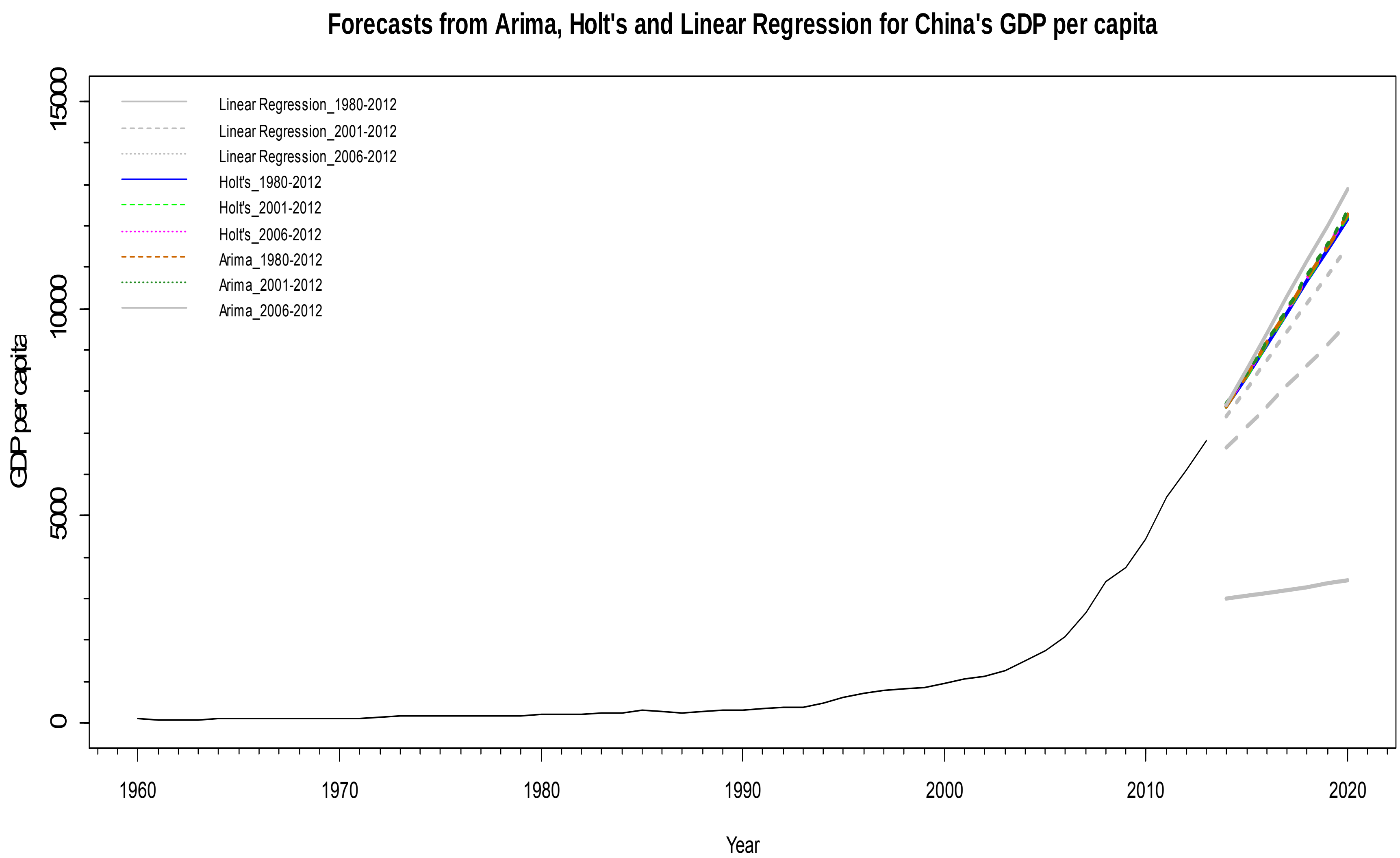
The next step in the comparison process was to plot all the predictions that were made with the three algorithms for every pair of country and indicator. As for every algorithm there were made predictions with three different datasets (the whole dataset, the subset 2001-2013 and the subset 2006-2013), totally nine lines were plotted in the same graph for every pair of country and indicator. This visualization of the forecasts in one graph, allow the easy comparison of the different methods that were used, and enable the identification of their reliability and their limitations.

For instance, in the Figure 5.2.3 there are plotted the nine forecasts for the China's GDP per capita. A first remark is that there is a core group of lines that give very similar forecasts and can be considered reliable, but there are also some dispersed lines away from that core that give forecasts that seem to be totally unreliable. In the next graph (Figure 5.2.4) the core group of reliable forecasts is emphasized while the dispersed forecasts have been put into the background with grey color. With this visualization, we can identify and emphasize the algorithms that gave good results for this case.

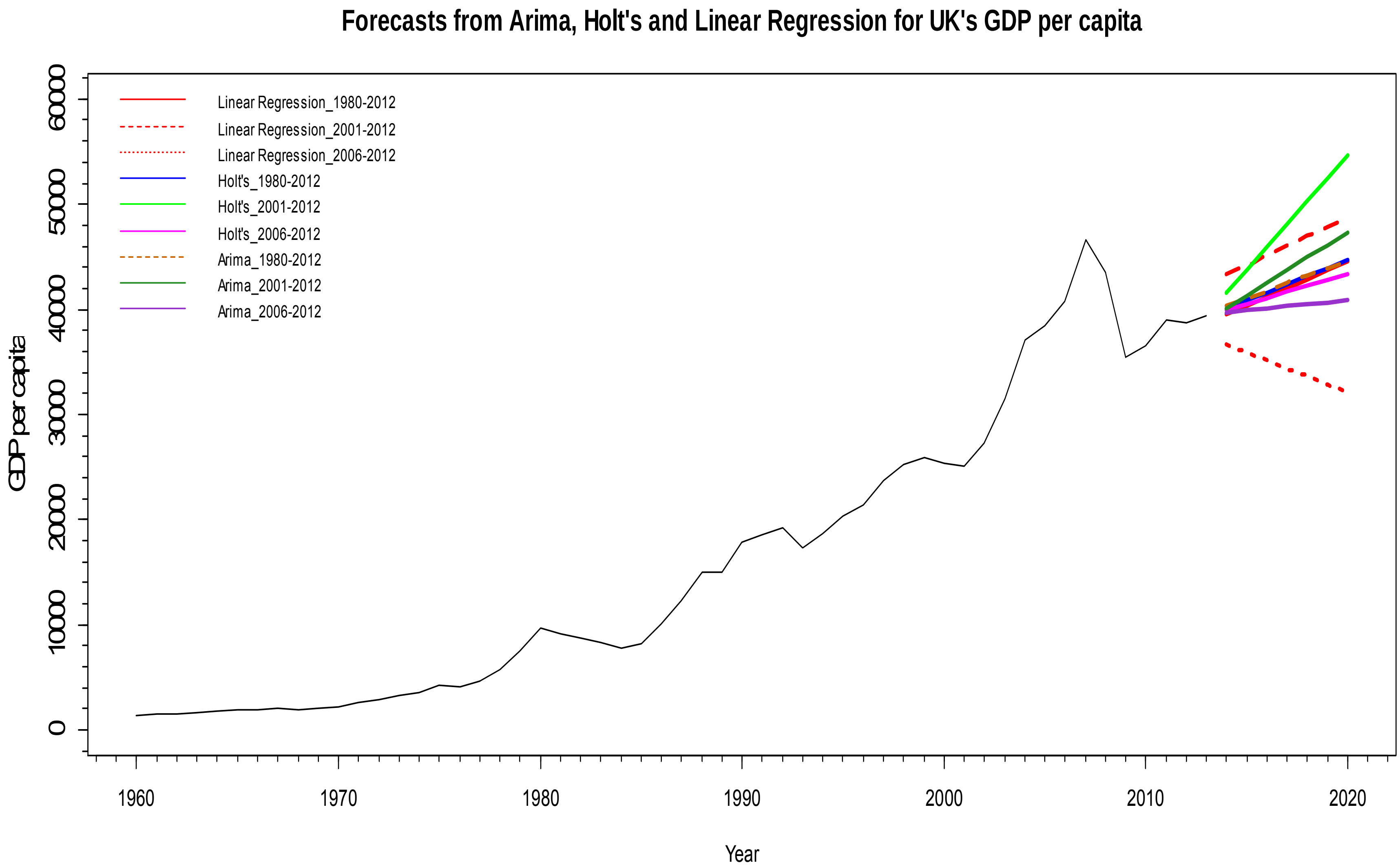
By repeating the same procedure for all pairs of indicators and countries, it is possible to find out which algorithms give the best results across all sixteen cases and identify the most reliable methods. Another example for UK's GDP per capita is given in Figures 5.2.5 and 5.2.6, while the rest of the graphs are shown in the Appendix D.2. (p.297).



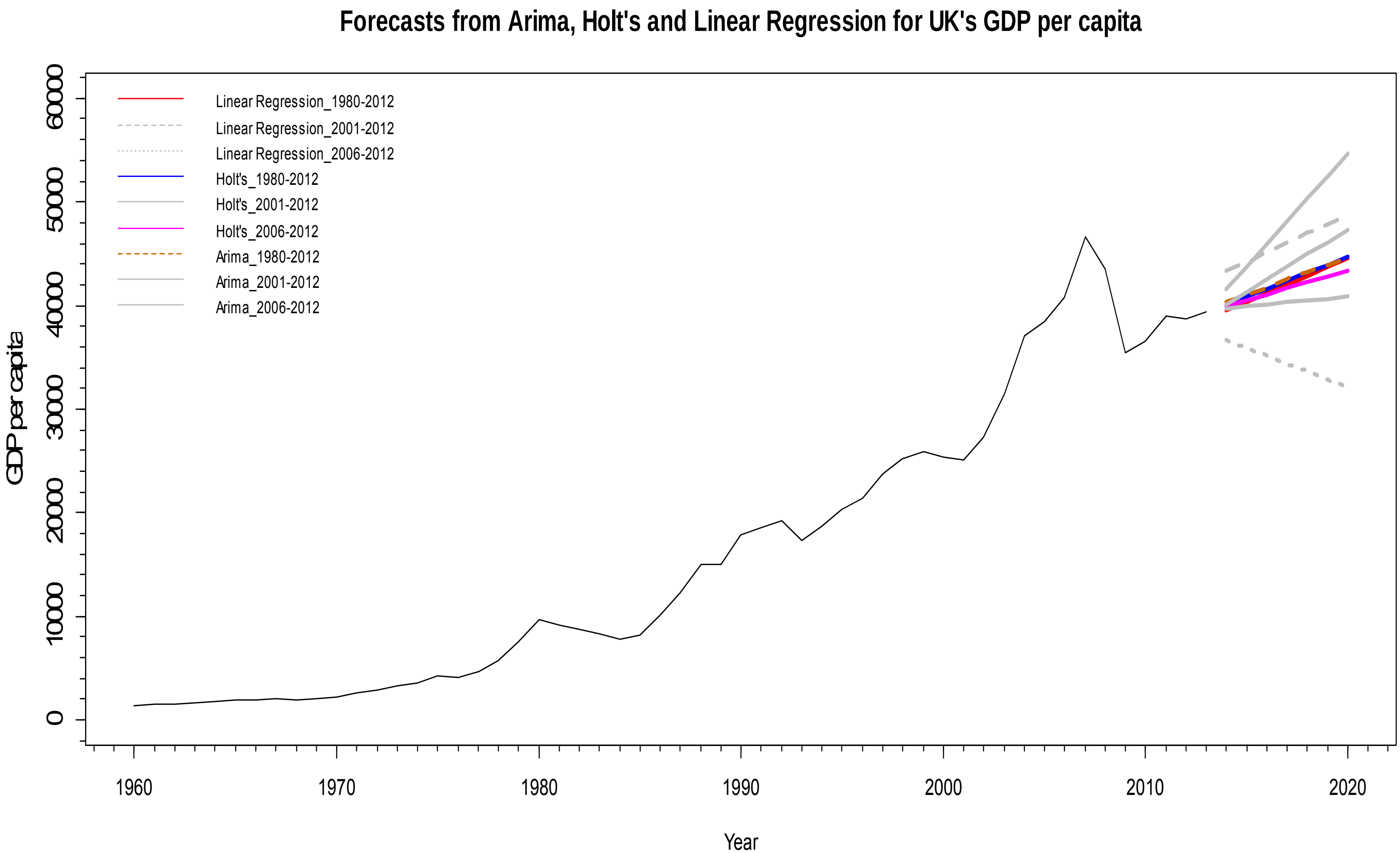
[Figure 5.2.3] – Comparison of all predictions for China’s GDP per capita



[Figure 5.2.4] – Selection of the best predictions for China's GDP per capita



[Figure 5.2.5] – Comparison of all predictions for UK's GDP per capita



[Figure 5.2.6] – Selection of the best predictions for UK's GDP per capita

After creating all the graphs, in an effort to identify the best algorithms, a grade from a rating from 0 to 2 was assigned to them, depending on how well their forecasts were considered to be for each case. A grade of 2 represents very good predictions, a grade of 0, a forecast line that seems to be totally inaccurate and a grade of 1, predictions that lie between these two extremes.

For example, in the China's GDP per capita case that is shown in Figures 5.2.3 and 5.2.4, the forecasts that are on the core group and are shown in color in both graphs would be assigned a value of 2 (the three analyses with Holt's Exponential smoothing and the analyses with ARIMA for the whole dataset and the subset 2001-2013). The two Linear Regression analyses (for the whole dataset and for the subset 2001-2013) get a grade of 0, while the Linear Regression and ARIMA analyses for the dataset 2006-2013, although they don't belong to the core group and are shown in grey color, get a value of 1, as they are quite close to the core group of forecasts.

The results of this rating for all the cases are shown in Table 5.2.3. From this table some useful conclusions can be drawn:

1. The algorithm that gave the best results is ARIMA model for the whole dataset with effectiveness 90.63%, followed closely by the Holt's Exponential smoothing for the whole dataset with 87.50% and the ARIMA for the subset 2001-2013 with 81.25%.
2. The ARIMA algorithm gave acceptable results with all the datasets that were used, so it can be considered the most reliable method.
3. The Holt's Exponential smoothing gave very good results with the whole dataset, but was not so satisfactory with the subsets 2001-2013 and 2006-2013.
4. The Linear Regression method had the worst results of the three algorithms and only when applied for the subset 2001-2013 was quite acceptable, but still with only 50% effectiveness. The Linear Regression analysis for the whole dataset gave the worst forecasts of all the cases that were studied.
5. The subset 2006-2013 gave the worst results comparing with the other two datasets for ARIMA and Holt's Exponential smoothing and in Linear Regression is better only from the whole dataset, but still with low percentage of effectiveness (43.75%). This indicates that so small datasets with only a few recent values are not adequate for accurate predictions.

Suitability of Algorithms for each Indicator and Country											
		Algorithm									
Indicator	Country	Linear Regr. 1960-2013	Linear Regr. 2001-2013	Linear Regr. 2006-2013	Holt's 1960-2013	Holt's 2001-2013	Holt's 2006-2013	Arima 1960-2013	Arima 2001-2013	Arima 2006-2013	
GDP per capita	China	0	0	1	2	2	2	2	2	1	20
	Brazil	0	0	0	2	2	0	2	1	1	
	UK	2	0	0	2	0	2	2	1	1	20
	EURO zone	0	0	1	2	1	0	2	2	2	
Foreign Investment	China	0	1	2	2	2	2	2	1	2	25
	Latin America	0	2	1	2	1	1	2	1	1	
	UK	1	2	0	1	0	0	2	1	1	14
	EURO	0	0	0	2	0	0	2	1	1	
Export value	China	0	2	2	1	1	0	2	2	2	26
	Brazil	0	2	2	2	2	1	1	2	2	
	UK	2	1	1	1	1	0	2	2	1	23
	USA	0	2	2	2	1	0	1	2	2	
Import value	China	0	0	1	2	2	2	2	2	1	24
	India	0	0	1	2	2	2	2	2	1	
	UK	2	2	0	1	1	1	2	2	0	22
	USA	0	2	0	2	1	2	1	2	1	
	Total:	7	16	14	28	19	15	29	26	20	
		21.88%	50.00%	43.75%	87.50%	59.38%	46.88%	90.63%	81.25%	62.50%	

[Table 5.2.3] – Rating of algorithms for all cases

From the previous table the aggregated scores for emerging and developed markets and for each indicator separately were calculated and are shown in the Table 5.2.4.

	Emerging Markets	Developed Markets	Total
GDP per capita	20	20	40
	55.56%	55.56%	55.56%
Foreign Investment	25	14	39
	69.44%	38.89%	54.17%
Export Value	26	23	49
	72.22%	63.89%	68.06%
Import Value	24	22	46
	66.67%	61.11%	63.89%
Total:	95	79	
	65.97%	54.86%	

[Table 5.2.4] – Aggregated scores for markets and indicators

This table shows that the emerging markets received better total score than the developed ones. This means that the predictive models that were applied, gave better forecasts for the emerging markets. We could assume from this that the pattern of the emerging markets time series is easier to interpret and conduct forecasts, as during the last years they follow a constant increasing trend. On the other hand, the developed markets are less predictable, as probably due to the financial crisis, they still appear to make complex and uncertain moves.

In terms of the indicators, it seems that the Export and Import Value indicators present higher probability to receive accurate forecasts independently of the algorithm used. On the other hand, the GDP per capita and Foreign Investment, mainly for the developed markets, seem to be more difficult to give reliable predictions.

5.3. Discussion

From the comparison of emerging and developed markets, we can say with confidence that the emerging markets have a vital role to play in the future economy. For all indicators, the big increase in the values of emerging countries after the year 2000 it is obvious and undoubted. This clearly justifies their title as emerging markets. In addition, the forecasts for the next years show that they will continue to grow with rates that in the worst case are equal and in most cases higher than the relevant of the developed countries. Furthermore, at the moment the emerging markets seem to be more predictable than the developed, which seem to remain unstable and not so easy to predict.

In terms of the algorithms that were used in the analysis, the ARIMA model is definitely the most reliable. Both in the evaluation of the 2008 predictions with the real values (p. 80) and in the graphical comparison of forecast with the other methods (p. 84), the ARIMA model had the better rating with very satisfactory percentage of effectiveness (75% and 90.63% respectively). Apart from that, the Holts Exponential smoothing when applied for the whole dataset also can give reliable forecasts. On the other hand, the Linear Regression analysis can give reliable forecasts only when it is applied in shorter periods of time, where the values of the indicator can be considered to have a relatively constant trend through time. The predictions for each indicator from the three methods are given in Appendix B.3 (p.127).

Finally, as we saw from the analysis of the UK's GDP example (p. 44) even the reliable algorithms like ARIMA and Holt's Exponential smoothing can sometimes fail to predict the future values accurately. This may happen due to factors that cannot be predicted. For example, social or political events can influence the financial environment and change the values of the indicators unexpectedly in a way that does not follow any of their previous patterns. In this case the algorithms of course are not able to provide accurate forecasts.

Chapter 6 – Evaluation, Reflections and Conclusions

6.1. Evaluation of the project

In terms of the initial objectives, as they were defined in the project proposal and are repeated in the Chapter 1 of this report (p.8), the project successfully accomplished all of them. The financial indicators that were used for the analysis and comparison of emerging and developed markets were indicators that are representative of the development of each country. As mentioned in the Data Gathering section of this report (p. 18) the source is a highly reputable institution and the purpose of these indicators is to measure the development of a country.

The World Bank (2013) website, regarding the indicators, states that “The primary World Bank collection of development indicators, compiled from officially-recognized international sources…presents the most current and accurate global development data available”. Of course on the website there are many more financial indicators. For this project, due to the imposed limitations of time, there were selected the four that were considered the most appropriate. In Appendix B.4. (p. 131), though, is provided a list of additional financial indicators that could be used in a similar analysis.

In total three algorithms were applied with the R Project software, as were initially estimated. The selection of the predictive algorithms that were used in the analysis was based on extended literature review for algorithms suitable for financial time series, as described in section 3.2.2 (p.19) of this report. The three algorithms had different levels of complexity and flexibility. The Linear regression method is a simple and quick method that provides only linear forecasts. The Holt’s Exponential smoothing is a little more complex. It involves the determination of two parameters, the parameter (α) that applies weighting on the data and the parameter (b) that expresses the trend of the time series. Finally, the third method is the most complex, with three parameters (p,d,q) and offers also non linear predictions.

The reasoning behind the selection was to test different methods, with different levels of complexity and ease of application and compare their results. Again, of course there are many more algorithms that worth to be tested, but the limitation of time for the project had to limit also the selection of algorithms for this project. For instance, Kalman Filter is considered to be another good algorithm for financial time series that would be tested if the time available for the project was adequate.

The comparison of the results was extended and gave a significant amount of generalised conclusions. The comparison, as fully described in Chapter 5, was conducted between emerging and developed markets (p. 66) and between different algorithms (p. 80). The conclusions from these comparisons are summarized in the next Chapter 6.2 (p. 95) and include important comments about the reliability of each method. The hypothesis that emerging markets have a dominant role to play in the future economy is well underpinned by the results of the analysis.

As for the visualization techniques part of the project, the project devoted deliberately less time than it was initially estimated in the project proposal. This was because it was decided to devote more time in the analysis by analyzing more countries and more indicators, aiming to increase in this way the reliability of the analysis and the ability to generalise the results. Still, some visualization techniques were used in the comparison of the results to enable the comparison. This was done by plotting different forecasts in the same plot, using different colors for each country and method or plotting in the background with grey color the forecasts that were not considered to be accurate. Also, recommendations for appropriate visualization techniques were made in the Chapter 2 and a suggestion for developing a useful visualization tool is provided in the Suggestions for Future Work Chapter 6.3. (p. 97).

The selection of R Project for the analysis is totally justified as it the most widely used statistical software environment for academic and research reasons. Although there are other environments as well, such as SAS, R has the advantage that as an open source software it can be installed and used free. In addition, R provides better plotting capabilities and there is also plenty of supporting documentation and users communities available on the internet that can offer valuable guidance and help.

Finally, the whole project was well organised, as in general it followed the project plan that was created in the project proposal submission. The tasks took place in the same order as was proposed. The only difference was, as already mentioned, that the visualisation techniques part was much smaller, as the analysis phase was expanded from 31 to 54 days to include the analysis for the additional countries and indicators. In addition, the Conclusions stage also lasted longer, as the bigger analysis led to more results to compare and evaluate, and finally took 15 days to be completed, instead of 7 days that was estimated.

6.2. Conclusions

From the analysis of each method and the comparison of their results, a significant number of generalised conclusions were drawn. Although these conclusions have already been discussed throughout this report in different chapters and they have been justified with references to specific figures or tables, in this Chapter they are briefly summarized and put together in thematic sections.

Emerging and developed markets

1. The change in the pattern of emerging markets after the year 2000 is obvious in all indicators.
2. From the comparison diagrams it is obvious that the emerging markets have generally bigger rates of growth than the developed countries. Especially in the predictions with dataset 2001-2013 the difference is even bigger.
3. The emerging markets at the moment are easier to predict, as the last years they follow a constant trend, while the developed appear to be unstable and remain unpredicted.

4. The Export and Import Value Indicators gave the best predictions especially for the emerging markets. While the GDP per capita had the same percentage of effectiveness for both markets, the difference for the Foreign Investment Indicator for the two markets is very big, with the emerging markets being much more reliable.

Linear Regression

5. When the time series does not have a constant trend through the time period of the analysis, the Linear Regression does not give good results.
6. In these cases a good practice is to split the dataset into smaller datasets that exhibit relatively constant trend and apply the Linear Regression to these smaller subsets.
7. The Linear Regression analysis for a time period of 10 to 15 years gave the best results, while smaller subset proved to be inadequate.

Holt's Exponential smoothing

8. In Holt's Exponential smoothing, the best results were taken when all the available data were used. This can be explained by the fact that the method automatically applies weighting to the values, depending on how old or new they are, with the parameter α.
9. For the developed countries, when the Holt's Exponential smoothing was applied in the smaller datasets with only values from the last 10 to 15 years, the analysis gave negative values for indicators that could only be positive. Obviously this is a limitation of applying the method in very small datasets.

ARIMA model

10. Arima is a more complex, but more flexible algorithm. It can be fitted better to the real values curve and give better predictions.
11. Arima is a robust and reliable method, as it gives good results for all datasets used and the best results on the evaluation with real values with high levels of effectiveness.
12. Arima allows for non-linear predictions.
13. Experience and experimentation is very important for the selection of the appropriate ARIMA(p,d,q) parameters. Significant help for that is provided by the auto.arima() function in R software environment.

Selection of Dataset

14. The selection of the dataset that will be used in the analysis is very important and must take into consideration the pattern of the time series and the algorithm that will be used.
15. Very small datasets with only a few observed values are proved inadequate. Very big datasets also can be proved inadequate for Linear Regression, as the possibility for a

very large dataset that the time series will exhibit constant trend and can be represented by a line is relatively small.

General Conclusions

16. Every time series is unique. The first step in the analysis is to plot the time series, examine its shape and identify patterns. This is crucial for the successful selection of the appropriate analysis method.
17. For the selection of appropriate parameters for the analysis, it is very important the analyst to have significant experience, knowledge and critical decision skills, as the results vary significantly for different selection of parameters.
18. The algorithms are useful to identify trends, but cannot always predict fluctuations and the real values accurately.
19. No matter how effective a model is, always there may be specific external events that may lead to abrupt and unpredicted by the algorithms changes. For instance, the financial crisis could not have been easily predicted by the models, as there were political and social factors that could not have been taken into account by the algorithms.
20. For this, the models are better to be used for predictions only for the next 3 to 4 years and should always be reassessed when new observations are become known.

6.3. Suggestions for future work

Predictive analytics is a very exciting and wide field that can find application in many different sectors and is not yet fully explored. This project, in the limited time that was available, examined the ways predictive algorithms can be used in financial time series. There is still space for a lot and interesting work to be done towards this direction. Further work could be undertaken in order to evaluate the results of this project and to expand its application. Different algorithms can be tested, different source of data, different indicators or countries and of course different software environment. In addition, visualization techniques could be explored, that would assist the application of predictive algorithms and the easy and efficient comparison of their results, in order to enhance decision making.

This project hopes to give basic guidance and inspiration to researchers to explore the predictive analytics field even further. Some ideas and proposals for future work are listed below:

1) Test different values of the parameter α and b in Holt's Exponential smoothing and compare the results with the results that R gave.
2) Test more complex ARIMA models and non-linear models for forecasting.
3) Evaluate the results of this project by comparing its forecasts with the real values of the following years (2014-2015).

4) Another analyst could evaluate the reliability and robustness of the algorithms by estimating his/her own parameters for Arima or Holt and compare the results with this project.
5) Other algorithms could be applied on the same data and evaluate the conclusions of this project regarding the emerging markets.
6) The same algorithms could be applied in different data from other source and the conclusions of this project regarding the reliability of the algorithms could be evaluated.
7) Explore Visualization techniques for the better representation of the outcomes of this research and more efficient decision making.
8) Design a tool which will visually enable the easier selection of the best values for the parameters of Arima and Holt's (for instance by visualizing the curves when inputting different parameters).
9) Conduct research on how to predict when the emerging markets will stop having increased rate of growth (for instance by studying previous emerging markets).
10) Identify ways to predict the next generation of emerging markets (by studying the signs of previous emerging markets).

6.4. Reflections

The whole project was a very interesting experience that was full of challenges. Through this project I had the chance to discover new areas and expand my knowledge and skills in the fields of predictive analytics and financial time series, which I consider to be two very promising areas with excellent career perspectives. In that sense, the project successfully equipped me with the foundation and the necessary experience and skills for my future career.

First of all, the literature review that I did, gave me the chance to gain a deep understanding of the current trends in the field and an insight of the algorithms that are used in the financial predictions. In addition, I learnt to use effectively the R Project software environment, in order to apply the algorithms and make predictions. This involved a considerable amount of coding, which was the most challenging part of my work, as I did not have significant previous coding experience. The fact that I was finally able to write code that could be rerunnable and be applied for different countries and different indicators by only changing slightly the number or the name of a column, was something that definitely filled me with satisfaction.

Furthermore, I had the opportunity to explore in depth three algorithms, understand their theoretical background, their differences, advantages and limitations. I realised the importance of experimentation for the selection of appropriate parameters and in many cases I had to apply critical decision making. I also learnt to analyse financial data and was able to create a general framework for the process of making predictions for financial time series. Finally, I gained awareness on the emerging markets and how they can influence the future global economy.

Of course, during the duration of the project there was made some mistakes or omissions that caused some minor delays. If I had to run the project again, I would definitely start the learning of R Project environment from the first day, as it proved more time consuming than expected. There are a lot of function and commands to learn and the sooner someone starts to experiment with the environment the better.

Besides, I would choose the plotting parameters carefully from the beginning, as in the early phases of the project I was plotting the graphs with the default settings. This led inevitably to unclear and confusing diagrams and in some cases I had to plot the graphs later again in order to maintain consistency of colors and types in the data and forecast lines through the whole project.

Lastly, I would start the writing of the report much earlier even from the first phase of the project and I would construct a basic structure. The lack of this, made the report writing a much more difficult process at the later stages, when I have to recall important aspects of the analysis.

Overall, despite the difficulties and the limitation of time, the project was really a very enjoyable and rewarding experience that was successfully accomplished and I will remember it with pleasure and satisfaction.

References

Box, G. E. P., and Jenkins, G. M. (1970). *Time series analysis: Forecasting and control*. San Francisco: Holden Day (revised ed. 1976).

Buono, P., Plaisant, C., Simeone, A., Aris, A., Shneiderman, B., Shmueli, G. and Jank, W. (2007) 'Similarity-based forecasting with simultaneous previews: A river plot interface for time series forecasting'. In *Information Visualization, 2007*. IV'07. 11th International Conference (pp. 191–196). IEEE. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4271981 (Accessed: 09 September 2014).

Dablemont, S., Van Bellegem, S., & Verleysen, M. (n.d.). *Forecasting' High' and' Low' of financial time series by Particle systems and Kalman filters*. Available at: http://perso.uclouvain.be/michel.verleysen/papers/ffm07sd2.pdf (Accessed: 22/4/2014).

Dalgaard, P. (2008) *Introductory statistics with R.* (2nd edn.). New York: Springer.

Date, P. (2009) 'Linear and nonlinear filtering in mathematical finance: an overview'. Proceedings of the *9th WSEAS International Conference on Systems Theory and Scientific Computation* (ISTASC'09), Moscow, Russia, August 20-22, 2009. WSEAS Press.

Dawson, C. W. (2005) *Projects in computing and information systems: a student's guide*. Harlow, England: Addison-Wesley.

De Gooijer, J. G., and Hyndman, R. J. (2006). *25 years of time series forecasting*. International Journal of Forecasting, 22(3), 443–473. doi:10.1016/j.ijforecast.2006.01.001

Firmino, P. R. A., de Mattos Neto, P. S. G. and Ferreira, T. A. E. (2014). *Correcting and combining time series forecasters*. Neural Networks, 50, 1–11. doi:10.1016/j.neunet.2013.10.008

Holt, C. C. (2004) *Forecasting seasonals and trends by exponentially weighted moving averages*. International Journal of Forecasting, 20(1), 5–10.

Ichikawa, Y., Tsunawaki, T., Fujishiro, I., & Yoon, H. (2002). *A Visualization Environment for Multiple Daytime Stock Price Predictions*.

Javed, W., McDonnel, B. and Elmqvist, N. (2010) *Graphical perception of multiple time series. Visualization and Computer Graphics*, IEEE Transactions on, 16(6), 927–934.

Keim, D. A., Mansmann, F., Schneidewind, J., Thomas, J. and Ziegler, H. (2008) *Visual analytics: Scope and challenges*. Springer. Available at: https://crawford.anu.edu.au/public_policy_community/content/doc/2008_Keim_Visual_analytics.pdf (Accessed: 07 September 2014)

Kalekar, P. S. (2004) *Time series forecasting using Holt-Winters exponential smoothing*. Kanwal Rekhi School of Information Technology, 4329008, 1–13.

Kalman, R. E. (1960) *A new approach to linear filtering and prediction problems*. Journal of Basic Engineering, 82(1), 35–45.

King, S. (2011) *The Southern Silk Road.* HSBC Global Research. Available at: http://m.hsbc.de/mobil/downloads/publikationen/studien/TheSouthernSilkRoad.pdf (Accessed: 01 September 2014).

Lacasa, L., Luque, B., Ballesteros, F., Luque, J., and Nuño, J. C. (2008). *From time series to complex networks: The visibility graph*. Proceedings of the National Academy of Sciences, 105(13), 4972–4975.

Lind, Marchal, Wathen, and Waite (2009) 'Time Series and Forecasting' in *Basic Statistics for Business and Economics* (3rd Canadian edn).

Mills, T. C. (2008). *The econometric modelling of financial time series* (3rd ed.). Cambridge, UK□; New York: Cambridge University Press.

Oates, B. J. (2006) *Researching information systems and computing.* London: SAGE Publications.

R-bloggers (2014). Available at: http://www.r-bloggers.com/cheat-sheet-for-prediction-and-classification-models-in-r/ (Accessed: 14 August 2014).

R Core Team (2014) *R: A Language and Environment for Statistical Computing.* Available at: http://www.r-project.org/ (Accessed: 01 July 2014).

R Core Team (2000) *R Data Import/Export*. Available at: ftp://200.236.31.6/CRAN/doc/manuals/R-data.pdf (Accessed: 06 July 2014).

Savikhin, A., Maciejewski, R. and Ebert, D. S. (2008) 'Applied visual analytics for economic decision-making'. In *Visual Analytics Science and Technology*, 2008. VAST'08. IEEE Symposium on (pp. 107–114). IEEE. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4677363 (Accessed: 07 September 2014).

Schorfheide, F. and Wolpin, K. *On the Use of Holdout Samples for Model Selection*. Available at: http://economics.sas.upenn.edu/~schorf/papers/SW-PP.pdf (Accessed: 20/9/2014).

Shmueli, G. and Koppius, O. R. (2011) *Predictive Analytics in Information Systems Research*. MIS Quarterly, 35(3). Available at: http://search.ebscohost.com/login.aspx?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=02767783&AN=63604908&h=h8V7LBJGKWZInFBHhM%2FcRWroz9%2BgFvzh3g6W0f5pa6uMuMN9xazDZmNe6gh0M0V%2FR5PNFcTERy%2Fn2bdQcqoiSQ%3D%3D&crl=c

Shumway, R. H. (2011) *Time Series Analysis and Its Applications: With R Examples* (2010), (3rd edn.). Available at: http://www.stat.pitt.edu/stoffer/tsa3/R_toot.htm (Accessed: 15 July 2014).

Statistics Netherlands (2009) *Checklist Quality of Statistical Output*. Available at: http://www.cbs.nl/NR/rdonlyres/4119715F-7437-4379-9A70-90A0893F949E/0/2009ChecklistQualityofStatisticalOutput.pdf (Accessed: 01 September 2014).

StatSoft, Inc. (2013) 'How To Identify Patterns in Time Series Data: Time Series Analysis' in *Electronic Statistics Textbook.* Available at: https://www.statsoft.com/Textbook/Time-Series-Analysis/button/3#exponential (Accessed: 22 July 2014).

The World Bank (2013) *Terms of Use for Datasets Listed in The World Bank Data Catalog*. Available at:http://web.worldbank.org/WBSITE/EXTERNAL/0,,contentMDK:22547097~pagePK:50016803~piPK:50016805~theSitePK:13,00.html (Accessed: 01 September 2014).

The World Bank (2013) *World Development Indicators*. Available at: http://data.worldbank.org/data-catalog/world-development-indicators (Accessed: 01 September 2014).

Tsay, R. S. (2005) *Analysis of financial time series*. Hoboken, N.J.: Wiley.

Using R for Linear Regression. Available at: http://www.montefiore.ulg.ac.be/~kvansteen/GBIO0009-1/ac20092010/Class8/Using%20R%20for%20linear%20regression.pdf (Accessed: 16 July 2014).

Using R for Time Series Analysis (2010) Available at: http://a-little-book-of-r-for-time-series.readthedocs.org/en/latest/src/timeseries.html (Accessed: 02 August 2014).

Venables, W. N., Smith, D. M. and Team, R. D. C. (2002). *An introduction to R.* Available at: http://www.math.vu.nl/sto/onderwijs/statlearn/R-Binder.pdf (Accessed: 01 July 2014).

Ward, K. (2011) *The world in 2050*. *Quantifying the Shift in the Global Economy.* HSBC. Available at: https://161.113.4.87/1/PA_1_083Q9FJ08A002FBP5S00000000/content/usshared/Premier/Promotions/2012/Jan/pdf/The_World_in_2050.pdf (Accessed: 20/4/2014).

Winters, P. R. (1960) *Forecasting sales by exponentially weighted moving averages*. Management Science, 6(3), 324–342.

Zhuang, E., Small, M., and Feng, G. (2014). 'Time series analysis of the developed financial markets' integration using visibility graphs'. *Physica A: Statistical Mechanics and Its Applications*, 410, 483–495. doi:10.1016/j.physa.2014.05.058

Appendix A – Project Proposal

A.1. Project Proposal

INM373 Research Methods and Professional Issues Coursework

Task 2: Research Proposal

Research Project:

"Evaluating predictive analytics methods for comparing emerging and developed markets"

By:

Germanos Pappas

Supervisor: Dr Cagatay Turkay

Spring term 2014

1. Introduction

Predictive analytics include a variety of statistical models that aim to create predictions about the future. These predictions are based on current and historical data and they play an important role in a variety of fields, such as marketing, banking, financial services or manufacturing and retail industries. Especially in the banking and finance sector, it can be a valuable tool for effective economic decision making and strategic planning. This is even more relevant in the developed markets, where after the financial downturn that they faced the last years, they now need to predict effectively and accurately future investments in an effort to recover. The current estimation is that the global economy is revolutionised, the trade and capital flows related to emerging markets will increase tenfold in the next forty years (King, 2011) and the banks should focus their operations in these markets. As Karen Ward from HSBC notes in The World in 2050 “By 2050, the emerging world will have increased five-fold and will be larger than the developed world…..19 of the top 30 economies by GDP will be countries that we currently describe as emerging” (Ward, 2011).

This is a hypothesis that is worth being investigated. Predictive analytics could be used to evaluate the validity of this hypothesis. But which predictive algorithms could be the most appropriate in order to analyse financial data and whether different algorithms give similar results or not in a fast paced environment like the banking industry, are questions for investigation. Also, which visualization techniques could be used, so these predictions to be presented in a more efficient way to the users is of special interest. Having these in mind, this research project aims to answer the above questions by evaluating the predictive analytics methods that could be used for comparing emerging and developed markets.

2. Purpose

According to the classification of Oates (2006, p. 16–21) except from adding to the existing body of knowledge, this project has two main purposes. Firstly it aims to find evidence to inform practice by comparing different predictive algorithms and recommending the most appropriate for financial analysis. The second purpose is to predict, plan and control, as it aims to provide knowledge and tools to help people make better predictions about economic decisions and plan financial investments. In addition, the project contributes to my personal needs, as it will help me gain experience and knowledge in Business Intelligence in the Finance sector, where I wish to pursue a successful career.

In order to achieve the aim of the proposed research project, specific, measurable, appropriate, realistic and time-related objectives are set by applying the SMART technique (Dawson, 2005, p. 59). These objectives are:

- Identify suitable financial indicators from emerging and developed markets to use in the predictive analysis
- Identify predictive algorithms suitable for financial analysis
- Apply at least 3 predictive algorithms with the statistical computing software R Project to the selected financial data from emerging and developed markets
- Compare the predictions of the different algorithms
- Identify the reliability of algorithms for financial predictions

- Present their predictions with suitable visualization techniques and make recommendations for appropriate techniques of presentation
- Test the validity of the hypothesis that emerging markets will dominate in the future global economy

The products of the research, as they are defined in Oates (2006, p. 21–24), are an exploration of an area or field and a critical analysis of the predictive analytics methods for financial predictions. The outcome will be a report that emphasizes on the comparison and evaluation of different algorithms. Recommendations will be made regarding the suitability of these algorithms for financial predictions and suggestions for better communication of their results to users with appropriate visualization techniques. More specifically, advantages and disadvantages of each algorithm will be presented regarding the type and volume of financial data that is available. Emphasis will be given on financial indicators regarding trade, revenues, exports and imports of goods and services.

As for the resources needed for the accomplishment of the project, they are all free of cost. The data that will be used in the analysis are accessed free from open data websites, the analysis will be made with the free open source software programming language R and the needed IT hardware resources will be personal or University's computers that are accessible free.

The beneficiaries of the research project will be Banks and companies related to the Banking Industry, Insurance companies, Financial and Business Analysts that would be able to make better predictions for their organisations and Traders and Investors that would be able to improve their decision making processes. In addition, other fields that use predictions and forecasts, such as auctions, marketing, manufacturing or retail industry could possibly benefit from the project's outcome as they could use part of the conclusions and recommendations of the research.

3. Critical Context

A plethora of statistical models have been developed that make projections for future values of variables based on existing data. These models are used to analyze time series, which are collections of data that are recorded through a period of time, with the aim to identify patterns and trends that allow predictions for the future. Linear regression models, moving average models like ARMA and ARIMA, and Exponential Smoothing are some of the most well known techniques in predictive analytics.

Even from 1957, Holt used exponentially weighted moving averages to forecast trends in seasonal data and later Winters in 1960 utilized the same technique to forecast seasonal sales. The Holt-Winters Exponential Smoothing, as is known since then, remains still a powerful technique in time series forecasting (Kalekar, 2004). At the same time many others models have been developed with the Kalman Filter algorithm (Kalman, 1960) having a prominent position among them.

Despite, though, the great number of statistical models that have been developed and the potentialities they offer for predicting future trends and values, predictive analytics have not been used as much as it would be expected. A recent report from MIS Quarterly (Shmueli & Koppius, 2011) states that "We address a particularly large gap, namely, the near-absence of predictive analytics in mainstream empirical IS research". Shmueli & Koppius researched the published articles in two major journals, the MIS Quarterly and the Information Systems Research. The outcome of their research was 252 articles that seemed to be related to predictive analytics or forecasting, but only 7 of them had really employed predictive analytics (Shmueli & Koppius, 2011, pp. 560–561).

A similar research on Scopus and Google Scholar for academic papers on predictive analytics did not come up with significant results. Only a few papers are available and these are mainly in fields such as sales, auctions, healthcare or retail industry. In financial predictive analytics the most significant works the last years is the research of Dablemont, Van Bellegem and Verleysen in forecasting High and Low of financial time series with the use of Kalman Filters and an overview of Kalman filters in mathematical finance by Date, P. (2009).

As for the representation of the predictions that are made with the statistical algorithms, a number of different visualization techniques have been developed. Savikhin, Maciejewski and Ebert (2008) present an application of visual analytics for improving economic decision making. Javed, McDonnel and Elmqvist (2010) explore different line graph techniques, such as simple line graph, braided graph, small multiples and horizon graphs, for multiple time series and present advantages and disadvantages of each technique. In addition, Buono et al (2007) created the TimeSearcher, a time series visualization tool that enables users to change parameters and see the corresponding forecast through an interactive interface. But still, as he notes in his introduction, “little attention has been given to the visualization and the user interaction” and definitely this is an area that is yet to be explored. Similarly, Keim, Mansmann, Schneidewind, Thomas and Ziegler (2008) draw attention to the novel analysis capabilities that the emerging field of visual analytics offers, by integrating interactive visualization methods and statistical prediction models.

Summarizing, the literature review reveals that although there have been developed many statistical algorithms for analyzing time series and forecasting, there is limited research on how these algorithms can be used as predictive analytics, especially in the finance sector. Moreover, there is no comparison between these algorithms that would enable the user to select the appropriate model depending on the data that he has access to or the scope of his prediction. In addition, although the way that the forecasts are communicated to the user is of great importance in order to enhance the efficiency of the model, often it is overlooked. Taking all this in account, there is a significant gap and lack of knowledge in the field of predictive analytics in the finance domain.

Considering also that the emerging markets are a new phenomenon that could change the financial balance and relationships in the next years, a research about their future dynamic and a comparison with the developed markets is worth undertaking. Thus, this project will try to fill the above existing gap in the predictive analytics for the financial domain by applying, comparing and presenting the results of different algorithms for predicting financial indicators of emerging and developed markets. For the application of the predictive algorithms, the open source software for statistical computing and graphics, R Project, will be used and so the documentation of the software (Venables, Smith, & R Team, 2002) is a valuable source.

4. Approaches: Methods & Tools for Analysis & Evaluation

The project is clearly of the Evaluation type, according to Dawson's categories of computing projects (Dawson, 2005, p. 6–7), as it focuses on evaluating different predictive analytics methods. The strategy of this project is Design and Creation, as "a research project that investigates, compares and evaluates two or more methods or models" and "the contribution to the knowledge is the critical comparison end evaluation" (Oates, 2006, p. 111). It will involve the five steps of design and creation process: awareness by doing a literature review, suggestion of the main research question and the possible solution, development of the project by applying and comparing the predictive algorithms, evaluation of their comparison and finally the conclusion of the results.

The analysis will be quantitative by analyzing the data with statistical models and comparing their results. The main benefits that the quantitative analysis brings to the project, according to Oates (2006, p. 263), are that the analysis is based on well-established techniques and specific measured quantities, it provides scientific respectability and that large volumes of data can be analysed and processed quickly. On the other hand, the main disadvantage is that the quality of the analysis depends highly on the quality of the data and special caution must be given on the data gathering process.

The data gathering method of this project will be the use of documents. More specifically the data for the analysis will be extracted from open data websites, such as the http://data.worldbank.org or http://pages.stern.nyu.edu/~adamodar/New_Home_Page/data.html, which provide open access to huge amounts of data about development in countries from all over the world. The access to these data is easy, quick and free as they are immediately downloadable from the websites. This is definitely a big advantage as there is no need to gather data with other time-consuming techniques or obtain consent for the use of data, as they are anonymous. Of course the reliability of these data is of big concern. As the data are gathered and stored in the websites for other purposes, it would be reasonable to question their suitability for the specific project and considerable evaluation shall be made in order to identify their appropriateness.

The data gathering will be the first phase of my project. As there are many websites that provide open access to datasets, a selection must be made for which websites will be used as source of data. The decision will be made mainly depending on the reliability of the host site and the purpose that the data are gathered and stored there. The evaluation guide for documents' analysis and evaluation from Oates (2006, p. 241–242) will provide me valuable assistance for this selection. Apart from selecting the source website, a choice of countries and the appropriate financial variables that will be used in the analysis, has to be made. As the number of countries and their respective financial indicators are big, the amount of available data is huge. For this, a suitable selection of data shall be made in order to ensure the feasibility of the project and the accomplishment of the aim. Definitely, a big drawback is that it may not be easy to evaluate the data correctly and there is the danger of using data that are not adequate, accurate or they have been censored and adjusted to the needs of their publisher.

The second important step of the project is to identify predictive algorithms that could be suitable for financial predictions. For this, an extended review of the relevant literature is necessary. The book *Analysis of financial time series* (Tsay, 2005) will offer valuable help in this step as it provides details for a great number of algorithms used in financial analysis. In addition to the literature review, my Supervisor with his personal experience and knowledge will provide me with guidance for the appropriate selection of models. After an initial discussion, it seems that Linear regression models, Box-Jenkins and Holt-Winters Exponential smoothing for time series and Kalman Filters are models that are worth testing. Definitely, the final selection of the algorithms that will be examined will be made after the literature review and also by considering the available time for the project. As there is a big number of available algorithms, the selection must be very cautious and reasonable.

The next stage will be the analysis of the data with the application of the selected algorithms. This will be done with the open source statistical tool R Project and it will be the longer phase of the project. Considering that it is software that I have never used before and might demand more advanced programming skills than those I have, I will need to put significant effort in learning how to use the programme effectively. For this, I have to allocate a significant part of the project for this task in my project planning. The manual “An Introduction to R” (Venables et al., 2002) which can be found on the Documentation part of the R Project website will be the basis of my learning.

The comparison of the results requires a visualization technique for the best representation. Again, a literature review will be necessary in order to be able to select the most suitable solution. There are many options for representing the predictions and a lot of challenges in the way the results could be presented to users in the most effective way. As this is a new field for me a considerable amount of time would be necessary for an in depth approach. Due to the limitations on the time of the project the level of developing or evaluating different visualization techniques cannot be defined precisely at the moment. Definitely, an appropriate visualization technique will be selected for the representation of the results. The horizon graphs seem to be a good option, but this will be confirmed after the literature review. Apart from that, if the time permits it, I will try to compare different visualization techniques and consider more parameters in the presentation of the results.

Overall, it is a very interesting and challenging project and it is important to note that its success depends solely on me and the effort I will put on it. I can arrange my time and the work on hours and dates that are convenient for me, as my work does not depend on external factors, such as waiting responses from participants in data gathering. Besides, the resources are all free and available immediately, so there will not be any delays on the project due to requiring consent to use the data and software. This gives me confidence that the project is feasible and the outcome will be of the desired high quality.

The limitations of the project have to do mainly with the time constraints that an MSc project has. The time is limited to three months, so considering the great number of algorithms and visualization techniques that are available I will not be able to explore and use all of them, but have to restrict the selection in those that seem to be the most suitable. Of course this selection will have limitations as some important algorithms might not be explored. Moreover, the algorithms will be tested only with the use of R Project. If the available period of time was bigger the same algorithms could be tested with a different statistical software package. This would enable more accurate and complete conclusions to be derived. Similarly, it would be very interesting to have the time to use more visualization techniques for the comparison of the results.

Despite these limitations, the project as explained in the Introduction and Critical Context of this proposal will contribute to the predictive analytics field and it can be the basis for further research in the fields of predictive analytics and visualization techniques.

5. Plan of work

The Project Plan and the Work Breakdown Structure can be seen in the Gantt chart in Figure 1, where the main tasks have been identified and have been divided into subtasks with their relevant dependencies. The project will start on 16 of June 2014 and will finish at 16 of September 2014. Meetings with Supervisor have been also identified and are shown as Milestones which are important for the continuation and successful completion of the project.

6. Risks

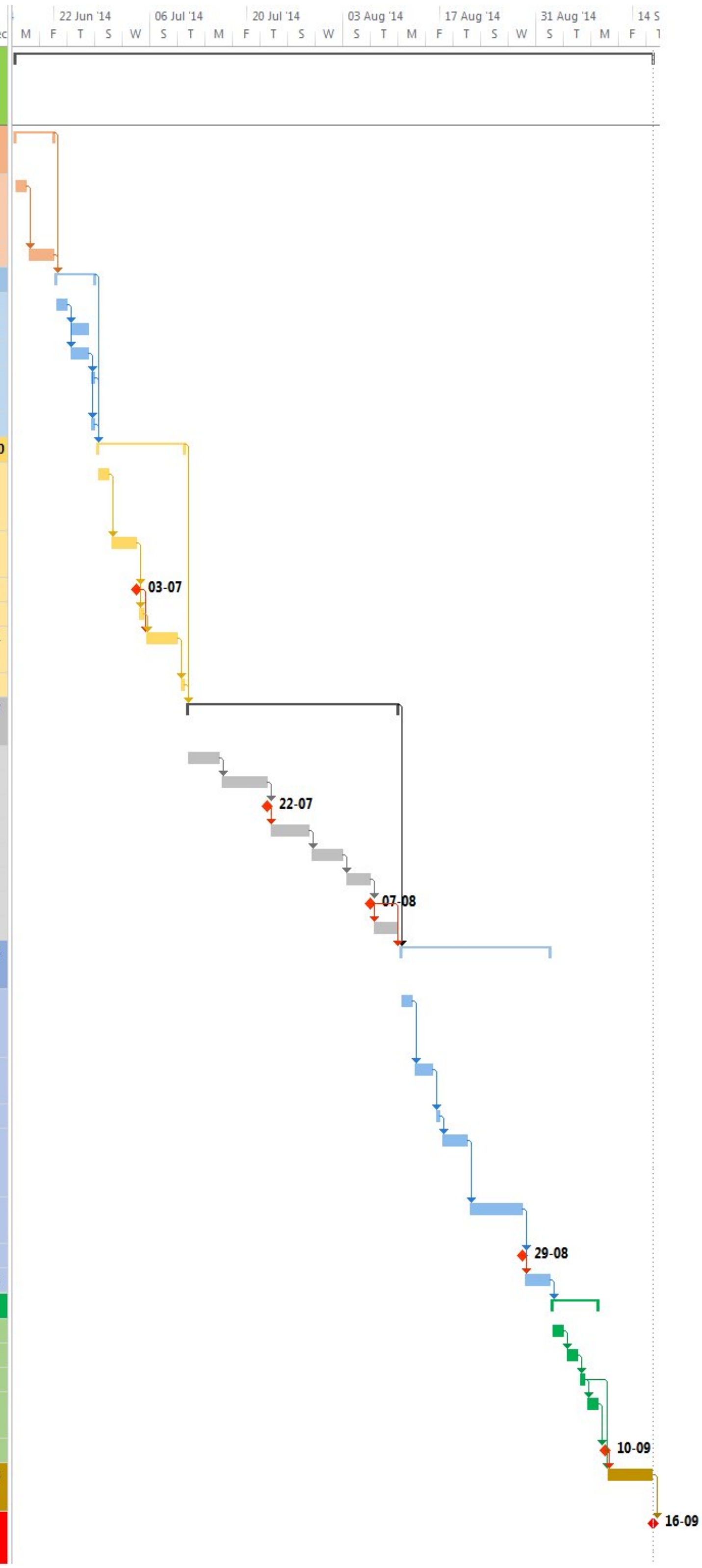
The risks that are relevant to the implementation of this project have been identified and assessed according to the risk management process of Dawson (2005, pp. 75–80) and can be seen in Risk Assessment Matrix (Table 1). The identified risks are classified as either technical or non-technical. The impact of each risk is calculated by multiplying its likelihood by its respective consequence on the project. The illustration of the impact that each risk has on the project is done by using the RAG grading in order to emphasize the criticality of the risks. In addition, for each risk, appropriate measures of mitigation or contingency plans are identified, so as to ensure the successful completion of the project. Since the risks have been identified, they can be controlled by using checkpoints, looking for triggers and revising the Risk Matrix during the project lifecycle.

7. Ethical, Professional & Legal Issues

This project does not involve human participation, as this is defined in the Research Ethics Policy of School of Informatics. The data that will be used in this research are anonymised secondary data, already published in open data websites and as a consequence the research project does not require ethical review and approval from the Informatics Research Ethics Panel or the University Research Ethics Committee. The project poses only minimal and predictable risk to the resercher as can also be seen in the Research Ethics Checklist. No other human is involved in the project except my Supervisor, as it is a clearly scientific research project that does not include participants in any way.

An issue that has to be considered is the reliability of the data that will be used in the research. As mentioned above they are anonymised secondary data published in open data websites and so they have to be treated with caution regarding their accuracy. For this, every effort will be made to evaluate the reliability of the source of data prior to their use. It is worth mentioning though, that a possible inaccuracy in data will affect only the correctness of the prediction that will be made and not the evaluation and comparison of the algorithms.

	Task Name	Duration	Start	Finish	Predec
1	**Project: Evaluating predictive analytics methods for comparing emerging and developed markets**	93 days	Mon 16-06-14	Tue 16-09-14	
2	**1. General Introductory Literature Review**	6 days	Mon 16-06-14	Sat 21-06-14	
3	Search for academic papers, articles and books in the area of Predictive Analytics and Finance	2 days	Mon 16-06-14	Tue 17-06-14	
4	Reading	4 days	Wed 18-06-14	Sat 21-06-14	3
5	**2. Data Gathering**	6 days	Sun 22-06-14	Fri 27-06-14	4,2
6	Evaluate and select sources of data	2 days	Sun 22-06-14	Mon 23-06-14	
7	Decide which countries to compare	3 days	Tue 24-06-14	Thu 26-06-14	6
8	Decide which financial indicators to use	3 days	Tue 24-06-14	Thu 26-06-14	6
9	Get permission if needed, download and store the data	1 day	Fri 27-06-14	Fri 27-06-14	8
10	Summarize and write report	1 day	Fri 27-06-14	Fri 27-06-14	8
11	**3. Selection of algorithms**	13 days	Sat 28-06-14	Thu 10-07-14	5,9,10
12	Literature review (specific for algorithms) and reselection of algorithms	2 days	Sat 28-06-14	Sun 29-06-14	
13	Analytical review of preselected algorithms	4 days	Mon 30-06-14	Thu 03-07-14	12
14	Milestone1: Meeting with Supervisor	0 days	Thu 03-07-14	Thu 03-07-14	13
15	Selection of algorithms to be tested	1 day	Fri 04-07-14	Fri 04-07-14	13
16	Reading, Practice and deep understanding of selected algorithms	5 days	Sat 05-07-14	Wed 09-07-14	15,14
17	Write report	1 day	Thu 10-07-14	Thu 10-07-14	16
18	**4. Analysis of data with statistical tool R**	31 days	Fri 11-07-14	Sun 10-08-14	11,17
19	Read documentation and practice R tool	5 days	Fri 11-07-14	Tue 15-07-14	
20	Predictive analysis with Algorithm 1	7 days	Wed 16-07-14	Tue 22-07-14	19
21	Milestone2: Meeting with Supervisor	0 days	Tue 22-07-14	Tue 22-07-14	20
22	Predictive analysis with Algorithm 2	6 days	Wed 23-07-14	Mon 28-07-14	21
23	Predictive analysis with Algorithm 3	5 days	Tue 29-07-14	Sat 02-08-14	22
24	Comparison of results	4 days	Sun 03-08-14	Wed 06-08-14	23
25	Milestone3: Meeting with Supervisor	0 days	Thu 07-08-14	Thu 07-08-14	24
26	Review, Conclusions and report writing	4 days	Thu 07-08-14	Sun 10-08-14	25
27	**5. Presentation with visualization techniques**	22 days	Mon 11-08-14	Mon 01-09-14	18,25
28	Literature review (specific for visualization) and preselection of techniques	2 days	Mon 11-08-14	Tue 12-08-14	
29	Analytical review of preselected techniques	3 days	Wed 13-08-14	Fri 15-08-14	28
30	Selection of visualization techniques	1 day	Sat 16-08-14	Sat 16-08-14	29
31	Reading, Practice and deep understanding of selected visualization techniques	4 days	Sun 17-08-14	Wed 20-08-14	30
32	Apply visualization techniques for the representation of the results of analysis	8 days	Thu 21-08-14	Thu 28-08-14	31
33	Milestone4: Meeting with Supervisor	0 days	Fri 29-08-14	Fri 29-08-14	32
34	Review, Conclusions and report writing	4 days	Fri 29-08-14	Mon 01-09-14	32,33
35	**6. Project Conclusions**	7 days	Tue 02-09-14	Mon 08-09-14	34
36	Comparison of algorithms	2 days	Tue 02-09-14	Wed 03-09-14	
37	Comparison of visualization techniques	2 days	Thu 04-09-14	Fri 05-09-14	36
38	Comments on R tool	1 day	Sat 06-09-14	Sat 06-09-14	37
39	Comparison of emerging and developed markets	2 days	Sun 07-09-14	Mon 08-09-14	38
40	Milestone5: Meeting with Supervisor	0 days	Wed 10-09-14	Wed 10-09-14	39
41	**7. Review of the whole Project, last amendments and final report writing**	7 days	Wed 10-09-14	Tue 16-09-14	40,38
42	**Submission of the Project**	0 days	Tue 16-09-14	Tue 16-09-14	41



Gantt chart [Figure 1]

Category		Risk Event	Likelihood	Consequence	Risk Impact	Mitigation	Contingency Plan
Non-Technical	1	The data may be inadequate	2	5	10	Avoidance: Research the datasets thoroughly before deciding which financial indicators and data will use	Choose other financila indicators or other countries for which there are adequate data. Identify more website that could act as sources for data
	2	The data may be inaccurate or false	3	4	12	Avoidance: Research the datasets thoroughly and assess their reliability before deciding which website	Chose other open data website that is more reliable and trustworthy
	3	Illness and unavailability to work	2	3	6	Avoidance: Start the project earlier and plan to finish it a few days before the deadline, so even if an illness occurs, there will still be sufficient time to complete the project	Working overtime and rescheduling the plan in order to accommodate for the lost time
	4	Wrong time estimation of a task	3	4	12	Avoidance: Make a very careful and reasonble project plan. Seek advice from supervisor in determining the	Working overtime and rescheduling the plan in order to accommodate for the lost time
	5	Change objective(s) of project due to unexpected findings during literature review	1	5	5	Avoidance: Make a thorough literature review before deciding the final objectives of the project	Be flexible, aware and ready to adapt in changes in the scope or the objectives of the project. Work overtime to cover the lost time
	6	Supervisor unavailable due to excessive workload, illness, trip abroad, etc	2	4	8	Avoidance: Arrange appointments with the supervisor several days in advance and in date and time that is convenient for him	Work on other parts of the project that do not need the consultation of supervisor until he is available again
	7	Incapability to complete particular tasks	2	5	10	Avoidance: Choose methods and techniques that are within my capabilities and knowledge as much as	Seek assistance from supervisor or other expert
	8	May need to travel in my home country for personal unexpected reasons	1	4	4	Avoidance: Start the project earlier and plan to finish it a few days before the deadline, so even if a trip abroad occurs, there will still be sufficient time to complete the project	Work from my home country with a laptop and communicate and seek advice from my supervisor through emails and Skype
	9	The literature review may not offer adequate information as it is a new field with no much research done	2	5	10	Avoidance: Do an exhaustive literature review in order to increse the possibilities of finding adequate information	Seek assistance from supervisor or other expert. May need to modify some of the objectives of the project
	10	The selection of suitable algorithms may be more difficult than expected due to the big number of algorithms that has to be explored	2	4	8	Avoidance: Consult the supervisor in order to make a preselection of algorithms that worth being examined	Decide a deadline until when the algorithms to be examined will be definitely selected. If needed, narrow the scope of the project by examining specific algorithms
	11	The selection of suitable visualization techniques may be more difficult than expected due to the big number of alternatives that has to be explored	2	4	8	Avoidance: Consult the supervisor in order to make a preselection of visualization techniques that worth being examined	Decide a deadline until when the visualization techniques to be examined will be definitely selected. If needed, narrow the scope of the project by applying specific visualization techniques
Technical	12	The R programming language might be more difficult to use than expected	3	4	12	Avoidance: Allocate sufficient time in the project plan for learning the R language in order to cover for possible difficulties that may demand extra time.	Work more hours than planned to cover the lost time and seek advice from an experienced user
	13	Loss of important files	2	5	10	Avoidance: Keep regularly back-up of files	Restore the lost work as soon as possible
	14	Hardware problems, PC is damaged	2	5	10	Avoidance: Do regularly maintainance of computer, conservative use of computer - avoid overuse for non-	Use other IT resources, eg University computers
	15	Virus and malware	1	5	5	Avoidance: Update anti-virus and firewall system frequently	Ensure that unaffected data are transferred to another hard disk, destroy the virus and malware immediately
	16	Problems with Internet connection	2	3	6	Avoidance: Download document and files in a hard disk so as to minimize the work that demands intenet	Work on another part of the project that does not require internet connection until the connection is

Risk Likelihood	
High	3
Medium	2
Low	1

Risk Consequence	
Very High	5
High	4
Medium	3
Low	2
Very Low	1

Risk Impact
11-15
6-10
1 - 5

Risk Assessment Matrix [Table1]

References

Buono, P., Plaisant, C., Simeone, A., Aris, A., Shneiderman, B., Shmueli, G. and Jank, W. (2007) 'Similarity-based forecasting with simultaneous previews: A river plot interface for time series forecasting'. In *Information Visualization*, 2007. IV'07. 11th International Conference (pp. 191–196). IEEE. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4271981

Dablemont, S., Van Bellegem, S., & Verleysen, M. (n.d.). Forecasting' High' and' Low' of financial *time series by Particle systems and Kalman filters*. Available at: http://perso.uclouvain.be/michel.verleysen/papers/ffm07sd2.pdf (Accessed: 22/4/2014).

Date, P. (2009) 'Linear and nonlinear filtering in mathematical finance: an overview'. Proceedings of the *9th WSEAS International Conference on Systems Theory and Scientific Computation* (ISTASC'09), Moscow, Russia, August 20-22, 2009. WSEAS Press

Dawson, C. W. (2005) *Projects in computing and information systems: a student's guide*. Harlow, England: Addison-Wesley.

Holt, C. C. (2004) *Forecasting seasonals and trends by exponentially weighted moving averages*. International Journal of Forecasting, 20(1), 5–10. doi:10.1016/j.ijforecast.2003.09.015.

Javed, W., McDonnel, B. and Elmqvist, N. (2010) *Graphical perception of multiple time series.* Visualization and Computer Graphics, IEEE Transactions on, 16(6), 927–934.

Kalekar, P. S. (2004) Time *series forecasting using Holt-Winters exponential smoothing*. Kanwal Rekhi School of Information Technology, 4329008, 1–13.

Kalman, R. E. (1960) *A new approach to linear filtering and prediction problems*. Journal of Basic Engineering, 82(1), 35–45.

Keim, D. A., Mansmann, F., Schneidewind, J., Thomas, J. and Ziegler, H. (2008) *Visual analytics: Scope and challenges*. Springer. Available at: https://crawford.anu.edu.au/public_policy_community/content/doc/2008_Keim_Visual_analytics.pdf

King, S. (2011) *The Southern Silk Road.* HSBC Global Research. Available at: http://m.hsbc.de/mobil/downloads/publikationen/studien/TheSouthernSilkRoad.pdf (Accessed: 20/4/2014).

Oates, B. J. (2006) *Researching information systems and computing*. London: SAGE Publications.

Savikhin, A., Maciejewski, R. and Ebert, D. S. (2008) 'Applied visual analytics for economic decision-making'. *In Visual Analytics Science and Technology*, 2008. VAST'08. IEEE Symposium on (pp. 107–114). IEEE. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4677363

Shmueli, G., & Koppius, O. R. (2011) *Predictive Analytics in Information Systems Research*. MIS Quarterly, 35(3). Available at: http://search.ebscohost.com/login.aspx?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=02767783&AN=63604908&h=h8V7LBJGKWZInFBHhM%2FcRWroz9%2BgFvzh3g6W0f5pa6uMuMN9xazDZmNe6gh0M0V%2FR5PNFcTERy%2Fn2bdQcqoiSQ%3D%3D&crl=c

Tsay, R. S. (2005) *Analysis of financial time series*. Hoboken, N.J.: Wiley.

Venables, W. N., Smith, D. M. and Team, R. D. C. (2002) *An introduction to R*. Network Theory Ltd. Available at: http://www.math.vu.nl/sto/onderwijs/statlearn/R-Binder.pdf (Accessed: 23/4/2014).

Ward, K. (2011) *The world in 2050.Quantifying the Shift in the Global Economy.* HSBC. Available at: https://161.113.4.87/1/PA_1_083Q9FJ08A002FBP5S00000000/content/usshared/Premier/Promotions/2012/Jan/pdf/The_World_in_2050.pdf (Accessed: 20/4/2014).

Winters, P. R. (1960) *Forecasting sales by exponentially weighted moving averages*. Management Science, 6(3), 324–342.

Research Ethics Checklist

School of Informatics BSc, MSc, MA Projects

If the answer to any of the following questions (1 – 3) is NO, your project needs to be modified.		*Delete as appropriate*
1.	Does your project pose only minimal and predictable risk to you (the student)?	Yes
2.	Does your project pose only minimal and predictable risk to other people affected by or participating in the project?	Yes
3.	Is your project supervised by a member of academic staff of the School of Informatics or another individual approved by the module leaders?	Yes

If the answer to either of the following questions (4 – 5) is YES, you MUST apply to the University Research Ethics Committee for approval. (You should seek advice about this from your project supervisor at an early stage.)		*Delete as appropriate*
4.	Does your project involve animals?	No
5.	Does your project involve pregnant women or women in labour?	No

If the answer to the following question (6) is YES, you MUST complete the remainder of this form (7 – 19). If the answer is NO, you are finished.		*Delete as appropriate*
6.	Does your project involve human participants? For example, as interviewees, respondents to a questionnaire or participants in evaluation or testing?	No

If the answer to any of the following questions (7 – 13) is YES, you MUST apply to the Informatics Research Ethics Panel for approval and your application may be referred to the University Research Ethics Committee. (You should seek advice about this from your project supervisor at an early stage.)		*Delete as appropriate*
7.	Could your project uncover illegal activities?	Yes / No
8.	Could your project cause stress or anxiety in the participants?	Yes / No
9.	Will you be asking questions of a sensitive nature?	Yes / No
10.	Does your project rely on covert observation of the participants?	Yes / No
11.	Does your project involve participants who are under the age of 18?	Yes / No
12.	Does your project involve adults who are vulnerable because of their social, psychological or medical circumstances (vulnerable adults)?	Yes / No

13.	Does your project involve participants who have learning difficulties?	Yes / No

The following questions (14 – 16) must be answered YES, i.e. you MUST COMMIT to satisfy these conditions and have an appropriate plan to ensure they are satisfied.	*Delete as appropriate*

14.	Will you ensure that participants taking part in your project are fully informed about the purpose of the research?	Yes / No
15.	Will you ensure that participants taking part in your project are fully informed about the procedures affecting them or affecting any information collected about them, including information about how the data will be used, to whom it will be disclosed, and how long it will be kept?	Yes / No
16.	When people agree to participate in your project, will it be made clear to them that they may withdraw (i.e. not participate) at any time without any penalty?	Yes / No

The following questions (17 – 19) must be answered and the requested information provided.	*Delete as appropriate*

17. Will consent be obtained from the participants in your project? Yes / No

Consent from participants will be necessary if you plan to gather personal, medical or other sensitive data about them. “Personal data” means data relating to an identifiable living person; e.g. data you collect using questionnaires, observations, interviews, computer logs. The person might be identifiable if you record their name, username, student id, DNA, fingerprint, etc.

*If **YES**, provide the consent request form that you will use and indicate who will obtain the consent, how are you intending to arrange for a copy of the signed consent form for the participants, when will they receive it and how long the participants will have between receiving information about the study and giving consent, and when the filled consent request forms will be available for inspection (**NOTE**: subsequent failure to provide the filled consent request forms will automatically result in withdrawal of any earlier ethical approval of your project):*

18. Have you made arrangements to ensure that material and/or private information obtained from or about the participating individuals will remain confidential? Yes / No

Provide details:

19. Will the research be conducted in the participant's home or other non-University location? Yes / No

*If **YES**, provide details of how your safety will be preserved:*

Templates

The templates available from the links below **must** be adapted according to the needs of your project before they are submitted for consideration. The sample form provided for projects involving children is to be used by the parents/guardians of the children participating in the research project.

Adult information sheet:

http://www.city.ac.uk/__data/assets/word_doc/0018/153441/TEMPLATE-FOR-PARTICIAPNT-INFORMATION-SHEET.doc

Adult consent form:

http://www.city.ac.uk/__data/assets/word_doc/0004/153418/TEMPLATE-FOR-CONSENT-FORM.doc

Child information sheet:

http://www.city.ac.uk/__data/assets/word_doc/0003/153462/Sample-Child-Information-Sheet.doc

Child consent form:

http://www.city.ac.uk/__data/assets/word_doc/0020/153461/Sample-child-consent-1.doc

Appendix B – Data

B.1. Links to Data

A/A	Indicator	Code	Last Update	Downloaded	Link
1	World Development Indicators (General information)	N/A	22-07-14	N/A	http://data.worldbank.org/data-catalog/world-development-indicators
2	GDP per capita (current US$)	NY.GDP.PCAP.CD	22-07-14	26-07-14	http://data.worldbank.org/indicator/NY.GDP.PCAP.CD
3	Foreign direct investment, net inflows (BoP, current US$)	BX.KLT.DINV.CD.WD	22-07-14	26-07-14	http://data.worldbank.org/indicator/BX.KLT.DINV.CD.WD
4	Export value index (2000 = 100)	TX.VAL.MRCH.XD.WD	22-07-14	26-07-14	http://data.worldbank.org/indicator/TX.VAL.MRCH.XD.WD
5	Import value index (2000 = 100)	TM.VAL.MRCH.XD.WD	22-07-14	26-07-14	http://data.worldbank.org/indicator/TM.VAL.MRCH.XD.WD

[Table B.1.1] – Links to the selected Financial Indicators

B.2. Data Tables

<u>GDP per capita</u>

Country Name	1960	1961	1962	1963	1964	1965	1966	1967	1968	1969
Brazil	208.387	203.190	257.823	289.058	258.628	258.239	312.063	343.526	370.725	399.733
China	88.722	75.049	70.122	73.420	84.573	97.470	103.181	95.497	90.371	98.890
India	83.807	87.043	91.676	103.160	117.863	121.697	91.751	98.157	101.695	109.530
Philippines	254.444	267.164	156.689	168.021	175.935	187.118	199.940	207.427	224.624	241.702
Russian Federation	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
United Kingdom	1380.306	1452.545	1513.651	1592.614	1729.400	1850.955	1959.628	2023.628	1896.387	2032.347
United States	2881.100	2934.553	3107.937	3232.208	3423.396	3664.802	3972.123	4152.020	4491.424	4802.642
Middle East & North Africa	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Latin America & Caribbean	355.291	362.836	406.618	395.884	428.848	452.201	482.419	481.687	506.985	551.316
Euro area	934.664	1019.493	1120.872	1245.620	1372.769	1484.635	1605.775	1732.284	1845.119	2025.043
East Asia & Pacific	89.869	79.001	71.480	74.994	84.972	96.782	103.044	97.393	95.837	104.751

1970	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984
440.635	499.904	580.671	767.995	994.726	1143.132	1377.857	1552.781	1728.643	1891.706	1930.538	2115.074	2208.833	1558.419	1567.316
111.823	117.182	130.111	155.078	157.999	175.866	162.921	182.679	154.973	182.284	193.023	195.304	201.443	223.250	248.287
114.404	120.697	125.201	146.442	166.564	161.032	164.109	189.617	209.352	227.916	271.250	275.321	279.221	296.918	282.286
186.768	201.031	211.396	258.357	343.234	360.661	402.654	450.110	506.072	596.381	684.630	731.717	741.796	645.482	594.041
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2241.971	2525.628	2891.981	3257.306	3516.420	4204.922	4041.040	4565.465	5785.462	7511.108	9622.975	9142.467	8718.267	8266.524	7777.644
5246.962	5623.588	6109.692	6741.101	7242.324	7819.959	8611.461	9471.529	10587.416	11695.363	12597.646	13992.923	14439.015	15561.268	17134.316
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
583.531	631.217	697.149	887.392	1152.050	1211.003	1310.769	1399.565	1548.803	1789.247	2069.730	2334.478	2147.599	1868.239	1799.749
2196.184	2475.563	2972.146	3837.131	4317.791	4987.431	5182.716	5873.706	7156.716	8628.120	9623.020	8328.728	8039.710	7827.949	7495.576
113.622	118.692	131.267	161.749	179.017	197.304	197.152	224.466	214.701	245.143	279.670	294.755	302.113	311.177	331.303

1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
1636.601	1928.720	2074.410	2287.237	2893.658	3086.915	2677.150	2526.337	2791.519	3426.102	4749.814	5107.794	5219.098	4979.145	3411.866
291.776	279.185	249.413	280.968	307.490	314.431	329.749	362.808	373.800	469.213	604.228	703.121	774.468	820.866	864.731
302.646	317.110	347.810	361.932	353.820	375.891	310.084	324.495	308.535	354.855	383.551	410.818	427.236	425.445	455.474
565.749	535.164	579.039	643.575	704.766	715.295	715.502	815.065	817.357	941.716	1064.841	1163.848	1131.400	970.614	1091.783
NA	NA	NA	NA	3428.762	3485.112	3427.318	3095.087	2929.303	2663.457	2669.946	2643.906	2737.555	1834.846	1330.748
8209.837	10063.839	12333.137	14951.052	15057.179	17805.249	18571.362	19211.862	17270.116	18664.394	20349.959	21349.435	23734.422	25266.395	25870.989
18269.279	19114.824	20100.789	21483.114	22922.465	23954.523	24404.995	25492.956	26464.783	27776.427	28781.950	30068.227	31572.635	32948.951	34639.120
NA	NA	NA	NA	NA	NA	NA	NA	1168.579	1169.243	1305.974	1473.062	1493.610	1507.728	1539.359
1821.196	1815.505	1863.018	2073.948	2241.253	2555.618	2656.781	2817.366	3235.923	3751.377	3886.247	4130.380	4444.199	4383.632	3884.772
7685.882	10761.048	13271.254	14549.944	14798.445	18522.519	19161.421	21028.265	19180.113	20230.138	23301.967	23499.671	21403.672	21921.272	21740.444
356.536	348.167	337.664	370.984	395.201	417.452	448.111	494.182	525.983	625.949	763.857	866.750	887.655	802.034	871.574

2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
3694.463	3128.145	2810.695	3039.672	3607.192	4739.305	5787.976	7193.925	8622.552	8373.458	10978.260	12576.196	11319.974	11208.083
949.178	1041.638	1135.448	1273.641	1490.380	1731.125	2069.344	2651.260	3413.589	3748.504	4433.341	5447.309	6092.782	6807.431
457.284	466.214	486.640	565.335	649.711	740.114	830.163	1068.679	1042.084	1147.239	1417.074	1539.606	1503.004	1498.872
1043.456	961.717	1004.991	1015.780	1084.765	1200.938	1398.827	1680.551	1920.992	1831.974	2135.918	2357.571	2587.017	2764.585
1771.583	2100.357	2373.393	2974.738	4109.385	5338.412	6947.502	9145.454	11699.679	8615.673	10709.769	13324.288	14090.649	14611.701
25361.940	25121.038	27301.460	31437.006	37021.146	38432.311	40807.561	46591.128	43486.913	35454.949	36572.500	38927.069	38648.901	39350.637
36467.295	37285.816	38175.376	39682.472	41928.886	44313.585	46443.810	48070.385	48407.077	46998.820	48357.674	49854.523	51755.215	53142.890
1576.981	1589.226	1530.525	1676.237	1899.018	2179.018	2495.527	3018.932	3748.297	3538.063	4014.700	4519.496	4697.815	4313.089
4235.706	4065.106	3519.912	3567.249	4034.108	4838.892	5611.751	6573.711	7585.013	7030.359	8611.928	9539.821	9404.301	9616.842
19720.857	19919.415	21562.845	26474.488	30162.858	31145.503	32879.718	37622.636	40993.393	37375.094	36519.032	39313.582	36650.020	38167.840
950.974	1002.921	1098.818	1230.669	1415.349	1623.084	1936.266	2426.869	3045.722	3262.165	3885.293	4699.641	5187.389	5689.975

Foreign Investment

Country Name	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989
Brazil	NA	NA	NA	NA	NA	1.30E+09	1.56E+09	1.83E+09	2.01E+09	2.42E+09	1.91E+09	2.52E+09	2.91E+09	1.61E+09	1.59E+09	1.44E+09	3.45E+08	1.17E+09	2.80E+09	1.13E+09
China	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	4.30E+08	6.36E+08	1.26E+09	1.66E+09	1.88E+09	2.31E+09	3.19E+09	3.39E+09
India	NA	NA	NA	NA	NA	-1.03E+07	-7.71E+06	-3.61E+07	1.81E+07	4.86E+07	7.92E+07	9.19E+07	7.21E+07	5.64E+06	1.92E+07	1.06E+08	1.18E+08	2.12E+08	9.13E+07	2.52E+08
Russian Federation	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Philippines	-1.04E+06	2.16E+07	4.10E+06	5.93E+07	1.31E+08	1.14E+08	1.54E+08	2.10E+08	1.01E+08	7.00E+06	-1.06E+08	1.72E+08	1.60E+07	1.05E+08	9.00E+06	1.20E+07	1.27E+08	3.07E+08	9.36E+08	5.63E+08
United Kingdom	1.49E+09	1.77E+09	1.21E+09	2.72E+09	4.37E+09	3.32E+09	3.01E+09	4.43E+09	3.79E+09	6.47E+09	1.01E+10	5.88E+09	5.41E+09	5.18E+09	-3.47E+08	5.48E+09	8.57E+09	1.59E+10	2.26E+10	3.17E+10
United States	1.26E+09	8.70E+08	1.35E+09	2.12E+09	3.33E+09	2.56E+09	3.25E+09	2.90E+09	5.85E+09	8.70E+09	1.69E+10	2.52E+10	1.25E+10	1.05E+10	2.48E+10	2.00E+10	3.54E+10	5.85E+10	5.77E+10	6.83E+10
Middle East & North Africa	1.27E+08	8.87E+07	1.46E+08	6.14E+08	6.59E+08	5.99E+08	2.27E+08	2.95E+08	8.20E+08	9.03E+08	2.48E+08	5.49E+08	1.94E+08	3.52E+08	9.89E+08	1.43E+09	1.04E+09	7.72E+08	1.56E+09	1.64E+09
Latin America & Caribbean	5.21E+08	7.71E+08	3.31E+07	6.75E+08	3.49E+08	2.52E+09	1.84E+09	2.74E+09	3.48E+09	4.66E+09	5.70E+09	7.92E+09	6.36E+09	5.17E+09	4.02E+09	5.84E+09	3.99E+09	2.71E+09	6.36E+09	6.20E+09
Euro area	2.33E+09	3.44E+09	4.60E+09	5.77E+09	6.56E+09	5.29E+09	4.20E+09	5.97E+09	7.94E+09	9.18E+09	9.47E+09	8.86E+09	7.10E+09	8.77E+09	8.87E+09	9.21E+09	1.32E+10	2.10E+10	3.11E+10	4.02E+10
East Asia & Pacific	2.23E+08	1.71E+08	2.03E+08	3.48E+08	7.78E+08	5.66E+08	6.49E+08	7.49E+08	7.06E+08	6.98E+08	1.14E+09	2.01E+09	2.40E+09	2.82E+09	2.84E+09	2.95E+09	3.12E+09	3.93E+09	6.74E+09	8.96E+09

1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
9.89E+08	1.10E+09	2.06E+09	1.29E+09	3.07E+09	4.86E+09	1.12E+10	1.97E+10	3.19E+10	2.86E+10
3.49E+09	4.37E+09	1.12E+10	2.75E+10	3.38E+10	3.58E+10	4.02E+10	4.42E+10	4.38E+10	3.88E+10
2.37E+08	7.35E+07	2.77E+08	5.50E+08	9.73E+08	2.14E+09	2.43E+09	3.58E+09	2.63E+09	2.17E+09
NA	NA	1.16E+09	1.21E+09	6.90E+08	2.07E+09	2.58E+09	4.86E+09	2.76E+09	3.31E+09
5.30E+08	5.44E+08	2.28E+08	1.24E+09	1.59E+09	1.48E+09	1.52E+09	1.22E+09	2.29E+09	1.25E+09
3.35E+10	1.65E+10	1.66E+10	1.65E+10	1.07E+10	2.17E+10	2.74E+10	3.75E+10	7.47E+10	8.93E+10
4.85E+10	2.32E+10	1.98E+10	5.14E+10	4.61E+10	5.78E+10	8.65E+10	1.06E+11	1.79E+11	2.89E+11
7.20E+08	1.15E+09	2.37E+09	2.79E+09	2.44E+09	9.07E+08	1.36E+09	1.95E+09	2.64E+09	1.91E+09
7.29E+09	1.19E+10	1.36E+10	1.23E+10	2.55E+10	2.70E+10	3.88E+10	6.03E+10	6.89E+10	7.90E+10
5.31E+10	4.56E+10	4.90E+10	4.72E+10	4.94E+10	6.81E+10	7.03E+10	7.23E+10	1.41E+11	1.99E+11
1.05E+10	1.33E+10	2.14E+10	3.89E+10	4.56E+10	5.08E+10	5.86E+10	6.26E+10	5.78E+10	5.05E+10

2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
3.28E+10	2.25E+10	1.66E+10	1.01E+10	1.82E+10	1.55E+10	1.94E+10	4.46E+10	5.07E+10	3.15E+10	5.33E+10	7.15E+10	7.61E+10	8.09E+10
3.84E+10	4.42E+10	4.93E+10	4.95E+10	6.21E+10	1.11E+11	1.33E+11	1.69E+11	1.87E+11	1.67E+11	2.73E+11	3.32E+11	2.96E+11	3.48E+11
3.58E+09	5.47E+09	5.63E+09	4.32E+09	5.77E+09	7.27E+09	2.00E+10	2.52E+10	4.34E+10	3.56E+10	2.74E+10	3.65E+10	2.40E+10	NA
2.71E+09	2.75E+09	3.46E+09	7.96E+09	1.54E+10	1.55E+10	3.76E+10	5.59E+10	7.48E+10	3.66E+10	4.32E+10	5.51E+10	5.06E+10	7.93E+10
2.24E+09	1.95E+08	1.54E+09	4.91E+08	6.88E+08	1.66E+09	2.71E+09	2.92E+09	1.34E+09	2.06E+09	1.07E+09	2.01E+09	3.22E+09	3.86E+09
1.22E+11	5.38E+10	2.55E+10	2.76E+10	5.73E+10	2.54E+11	2.15E+11	2.41E+11	2.62E+11	4.06E+09	6.13E+10	3.62E+10	6.68E+10	3.51E+10
3.21E+11	1.67E+11	8.44E+10	6.38E+10	1.46E+11	1.38E+11	2.94E+11	3.40E+11	3.33E+11	1.54E+11	2.59E+11	2.53E+11	2.04E+11	2.36E+11
3.92E+09	3.72E+09	8.20E+09	1.01E+10	1.04E+10	1.85E+10	2.98E+10	3.36E+10	3.55E+10	2.93E+10	2.56E+10	1.58E+10	2.28E+10	1.80E+09
7.37E+10	6.72E+10	5.26E+10	4.04E+10	5.93E+10	6.73E+10	6.60E+10	1.13E+11	1.22E+11	7.16E+10	1.11E+11	1.46E+11	1.52E+11	1.57E+11
4.23E+11	2.01E+11	2.52E+11	2.20E+11	1.25E+11	4.19E+11	3.97E+11	6.14E+11	4.19E+11	3.18E+11	2.57E+11	3.72E+11	1.72E+11	1.83E+11
4.53E+10	4.92E+10	5.93E+10	5.95E+10	7.78E+10	1.37E+11	1.62E+11	2.09E+11	2.27E+11	1.90E+11	3.20E+11	3.86E+11	3.56E+11	3.85E+11

Export Value

Country Name	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989
Brazil	36.547	42.285	36.625	39.754	49.024	46.544	40.571	47.606	60.804	62.417
China	7.263	8.831	8.957	8.919	10.489	10.975	12.416	15.825	19.067	21.082
France	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Germany	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
India	20.259	19.574	22.081	21.585	22.302	21.566	22.178	26.659	31.226	37.451
Philippines	14.431	14.214	12.490	12.292	13.256	11.591	12.080	14.269	17.650	19.523
Russian Federation	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
United Kingdom	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
United States	31.376	30.850	29.252	28.971	30.831	31.827	36.779	42.854	48.863	52.055

1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
57.027	57.402	64.977	69.991	79.050	84.375	86.625	96.145	92.781	87.108
24.916	28.856	34.085	36.815	48.557	59.702	60.612	73.351	73.720	78.222
NA	NA	NA	NA	NA	92.154	93.484	92.454	98.111	99.609
NA	NA	NA	NA	NA	95.105	95.182	93.182	98.687	98.838
42.401	41.829	46.314	50.901	59.043	72.276	78.116	82.607	78.899	84.161
20.403	22.123	24.509	27.974	33.442	46.354	54.051	65.901	77.905	96.874
NA	NA	NA	NA	NA	77.207	84.352	84.095	70.874	71.929
NA	NA	NA	NA	NA	83.580	90.673	98.489	96.123	95.680
59.509	60.404	64.159	63.133	71.252	74.783	79.941	88.140	87.239	88.986

2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
100.000	105.632	109.512	132.593	175.399	215.042	250.019	291.459	359.119	277.572	366.326	464.521	440.103
100.000	106.780	130.655	175.852	238.089	305.756	388.831	489.744	574.107	482.182	633.120	761.781	822.147
100.000	98.953	101.506	119.963	138.343	141.808	151.734	171.240	188.568	148.342	160.177	182.519	174.132
100.000	103.851	111.878	136.536	165.300	176.386	201.310	240.026	262.727	203.478	228.709	267.780	255.628
100.000	102.316	116.212	139.131	180.863	235.058	287.423	354.321	459.725	389.126	534.105	714.748	691.879
100.000	86.512	93.250	95.960	105.096	109.265	125.568	133.661	129.984	101.799	136.390	127.938	137.711
100.000	96.999	102.156	129.412	174.423	232.109	288.997	337.411	448.995	288.842	381.422	496.983	503.880
100.000	95.783	98.411	107.343	122.047	135.037	157.577	154.218	161.481	124.646	146.094	176.503	164.502
100.000	93.245	88.641	92.691	104.215	115.240	131.212	146.844	164.652	135.058	163.478	189.333	197.883

Import Value

Country Name	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
Brazil	42.573	41.069	35.935	28.656	25.942	24.445	26.534	28.281	27.383	33.900	38.418	39.142	39.345	47.313	61.396	92.317	97.166	109.548	103.425	88.274
China	8.859	9.780	8.568	9.503	12.177	18.771	19.060	19.199	24.553	26.273	23.699	28.340	35.801	46.185	51.363	58.696	61.746	63.188	62.351	73.676
France	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	85.558	87.086	84.267	90.992	93.350
Germany	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	93.535	92.439	89.876	94.969	95.669
India	28.850	29.925	28.698	27.290	29.642	30.914	29.931	32.365	37.074	39.884	45.765	39.687	45.763	44.230	52.099	67.362	73.641	80.415	83.419	91.181
Philippines	22.393	22.896	22.341	21.542	17.372	14.732	14.207	19.411	23.581	30.169	35.119	34.735	41.853	50.472	61.147	76.539	92.165	104.308	85.062	87.957
Russian Federation	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	123.500	139.478	149.175	129.319	88.128
United Kingdom	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	76.979	82.643	88.576	92.430	93.665
United States	21.244	22.829	20.811	22.251	28.200	28.795	31.101	34.375	38.047	40.737	42.566	42.095	45.422	49.364	56.133	61.213	65.276	71.390	74.990	84.129

2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
100.000	99.995	84.778	86.726	113.284	132.373	163.425	215.959	310.995	227.952	326.614	404.078	397.778
100.000	108.234	131.173	183.429	249.409	293.281	351.723	424.895	503.309	447.029	620.488	774.799	807.944
100.000	97.199	97.396	117.980	139.317	149.134	160.325	186.660	212.097	165.969	180.390	213.023	199.328
100.000	98.014	98.853	121.905	144.312	156.677	182.810	212.711	238.939	186.775	212.677	253.013	235.382
100.000	97.805	109.693	140.826	193.653	277.294	346.273	445.180	623.085	499.200	679.764	901.468	949.799
100.000	94.311	110.977	114.985	124.509	133.651	146.049	156.629	163.176	123.902	157.905	172.015	176.518
100.000	108.948	123.543	154.149	197.336	254.182	332.902	452.876	591.432	388.673	503.836	656.216	679.753
100.000	99.017	104.861	115.036	135.552	147.948	173.222	179.423	182.421	149.505	170.247	194.037	195.971
100.000	93.638	95.309	103.474	121.153	137.593	152.313	160.439	172.277	127.475	156.371	179.933	185.450

B.3. Data tables with forecasts

GDP per capita

Year	2014			2015			2016		
Country Name	Linear Regr. 2001-2013	Holt 1960-2013	Arima 1960-2013	Linear Regr. 2001-2013	Holt 1960-2013	Arima 1960-2013	Linear Regr. 2001-2013	Holt 1960-2013	Arima 1960-2013
China	6652.14	7617.21	7636.99	7148.20	8377.59	8424.73	7644.27	9137.96	9191.86
Brazil	13390.73	11624.91	11824.17	14277.48	12041.73	12253.87	15164.23	12458.56	12675.10
UK	43298.42	40112.31	40320.09	44218.50	40873.98	41036.88	45138.57	41635.65	41753.66
EURO zone	43596.04	38977.18	39462.29	45112.11	39786.52	40173.12	46628.17	40595.87	40883.95

2017			2018			2019			2020		
Linear Regr. 2001-2013	Holt 1960-2013	Arima 1960-2013	Linear Regr. 2001-2013	Holt 1960-2013	Arima 1960-2013	Linear Regr. 2001-2013	Holt 1960-2013	Arima 1960-2013	Linear Regr. 2001-2013	Holt 1960-2013	Arima 1960-2013
8140.34	9898.34	9959.00	8636.40	10658.72	10726.14	9132.47	11419.10	11493.28	9628.53	12179.48	12260.42
16050.98	12875.38	13096.32	16937.73	13292.20	13517.55	17824.48	13709.03	13938.78	18711.22	14125.85	14360.00
46058.65	42397.32	42470.45	46978.72	43158.99	43187.24	47898.80	43920.66	43904.02	48818.88	44682.33	44620.81
48144.24	41405.21	41594.79	49660.31	42214.55	42305.62	51176.37	43023.89	43016.45	52692.44	43833.23	43727.29

Foreign Investment

Year	2014			2015			2016		
Country Name	Linear Regr. 2001-2013	Holt 1970-2013	Arima 1970-2013	Linear Regr. 2001-2013	Holt 1970-2013	Arima 1970-2013	Linear Regr. 2001-2013	Holt 1970-2013	Arima 1970-2013
China	3.62E+11	3.88E+11	4.05E+11	3.90E+11	4.18E+11	4.26E+11	4.17E+11	4.48E+11	4.58E+11
Latin America & Caribbean	1.59E+11	1.62E+11	1.60E+11	1.68E+11	1.67E+11	1.64E+11	1.77E+11	1.73E+11	1.68E+11
UK	9.10E+10	4.09E+10	6.57E+10	8.93E+10	4.12E+10	6.72E+10	8.75E+10	4.15E+10	1.72E+10
EURO zone	3.16E+11	2.04E+11	2.05E+11	3.18E+11	2.05E+11	2.10E+11	3.20E+11	2.06E+11	2.14E+11

2017			2018			2019			2020		
Linear Regr. 2001-2013	Holt 1970-2013	Arima 1970-2013	Linear Regr. 2001-2013	Holt 1970-2013	Arima 1970-2013	Linear Regr. 2001-2013	Holt 1970-2013	Arima 1970-2013	Linear Regr. 2001-2013	Holt 1970-2013	Arima 1970-2013
4.44E+11	4.78E+11	4.91E+11	4.72E+11	5.07E+11	5.23E+11	4.99E+11	5.37E+11	5.56E+11	5.27E+11	5.67E+11	5.88E+11
1.86E+11	1.78E+11	1.71E+11	1.95E+11	1.84E+11	1.75E+11	2.05E+11	1.89E+11	1.78E+11	2.14E+11	1.94E+11	1.82E+11
8.58E+10	4.18E+10	8.46E+10	8.41E+10	4.21E+10	7.43E+10	8.24E+10	4.23E+10	9.16E+10	8.07E+10	4.26E+10	7.82E+10
3.21E+11	2.07E+11	2.19E+11	3.23E+11	2.09E+11	2.24E+11	3.25E+11	2.10E+11	2.28E+11	3.27E+11	2.11E+11	2.33E+11

Export Value

Year	2013			2014			2015			2016		
Country Name	Linear Regr. 2001-2012	Holt 1980-2012	Arima 1980-2012	Linear Regr. 2001-2012	Holt 1980-2012	Arima 1980-2012	Linear Regr. 2001-2012	Holt 1980-2012	Arima 1980-2012	Linear Regr. 2001-2012	Holt 1980-2012	Arima 1980-2012
China	854.765	897.810	892.998	920.766	978.537	963.850	986.768	1059.263	1034.701	1052.770	1139.989	1105.553
Brazil	481.402	491.566	464.149	514.601	528.435	488.195	547.800	565.303	512.242	580.999	602.172	536.288
UK	178.310	175.702	169.262	184.670	182.795	174.022	191.030	189.888	178.782	197.390	196.981	183.543
USA	200.454	206.991	206.073	210.492	217.311	214.264	220.530	227.631	222.454	230.568	237.951	230.644

2017			2018			2019			2020		
Linear Regr. 2001-2012	Holt 1980-2012	Arima 1980-2012	Linear Regr. 2001-2012	Holt 1980-2012	Arima 1980-2012	Linear Regr. 2001-2012	Holt 1980-2012	Arima 1980-2012	Linear Regr. 2001-2012	Holt 1980-2012	Arima 1980-2012
1118.771	1220.715	1176.404	1184.773	1301.442	1247.256	1250.775	1382.168	1318.107	1316.776	1462.894	1388.958
614.198	639.041	560.334	647.397	675.910	584.380	680.596	712.778	608.426	713.795	749.647	632.473
203.750	204.074	188.303	210.110	211.167	193.063	216.470	218.260	197.823	222.830	225.353	202.583
240.606	248.271	238.834	250.644	258.591	247.024	260.681	268.912	255.215	270.719	279.232	263.405

Import Value

Year	2013			2014			2015			2016		
Country Name	Linear Regr. 2001-2012	Holt 1980-2012	Arima 1980-2012	Linear Regr. 2001-2012	Holt 1980-2012	Arima 1980-2012	Linear Regr. 2001-2012	Holt 1980-2012	Arima 1980-2012	Linear Regr. 2001-2012	Holt 1980-2012	Arima 1980-2012
China	822.526	896.826	906.231	886.303	996.228	1005.527	950.080	1095.629	1104.823	1013.857	1195.031	1204.119
India	959.908	1058.051	1064.037	1040.099	1178.920	1235.613	1120.290	1299.789	1330.715	1200.480	1420.657	1443.139
UK	209.276	201.635	202.971	217.790	207.299	209.970	226.304	212.963	216.970	234.818	218.627	223.969
USA	192.398	188.930	190.582	200.390	196.291	195.715	208.381	203.653	200.847	216.373	211.015	205.979

2017			2018			2019			2020		
Linear Regr. 2001-2012	Holt 1980-2012	Arima 1980-2012	Linear Regr. 2001-2012	Holt 1980-2012	Arima 1980-2012	Linear Regr. 2001-2012	Holt 1980-2012	Arima 1980-2012	Linear Regr. 2001-2012	Holt 1980-2012	Arima 1980-2012
1077.634	1294.432	1303.415	1141.411	1393.834	1402.710	1205.188	1493.236	1502.006	1268.965	1592.637	1601.302
1280.671	1541.526	1588.651	1360.861	1662.395	1702.107	1441.052	1783.264	1817.546	1521.242	1904.133	1950.358
243.331	224.291	230.969	251.845	229.955	237.968	260.359	235.619	244.968	268.873	241.283	251.967
224.365	218.377	211.112	232.356	225.739	216.244	240.348	233.100	221.376	248.340	240.462	226.508

B.4. Additional Financial Indicators

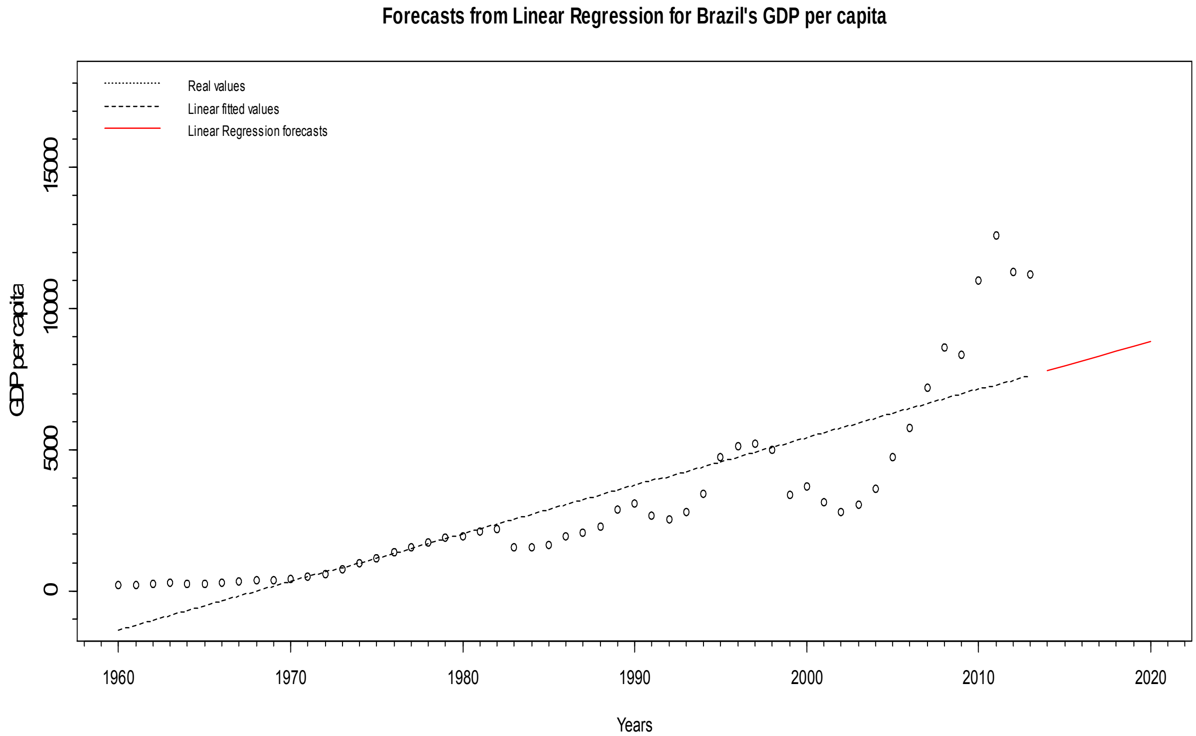
A/A	Indicator	Link
1	GDP (current US$)	http://data.worldbank.org/indicator/NY.GDP.MKTP.CD
2	GDP growth (annual %)	http://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG
3	Inflation, consumer prices (annual %)	http://data.worldbank.org/indicator/FP.CPI.TOTL.ZG
4	Total reserves (includes gold, current US$)	http://data.worldbank.org/indicator/FI.RES.TOTL.CD
5	Unemployment, total (% of total labor force)	http://data.worldbank.org/indicator/SL.UEM.TOTL.ZS
6	Air transport, registered carrier departures worldwide	http://data.worldbank.org/indicator/IS.AIR.DPRT
7	Internet users (per 100 people)	http://data.worldbank.org/indicator/IT.NET.USER.P2
8	Mobile cellular subscriptions (per 100 people)	http://data.worldbank.org/indicator/IT.CEL.SETS.P2
9	Ease of doing business index (1=most business-friendly)	http://data.worldbank.org/indicator/IC.BUS.EASE.XQ
10	Export volume index (2000 = 100)	http://data.worldbank.org/indicator/TX.QTY.MRCH.XD.WD
11	Import volume index (2000 = 100)	http://data.worldbank.org/indicator/TM.QTY.MRCH.XD.WD
12	Research and development expenditure (% of GDP)	http://data.worldbank.org/indicator/GB.XPD.RSDV.GD.ZS

[Table B.4.1] – Links to additional Financial Indicators

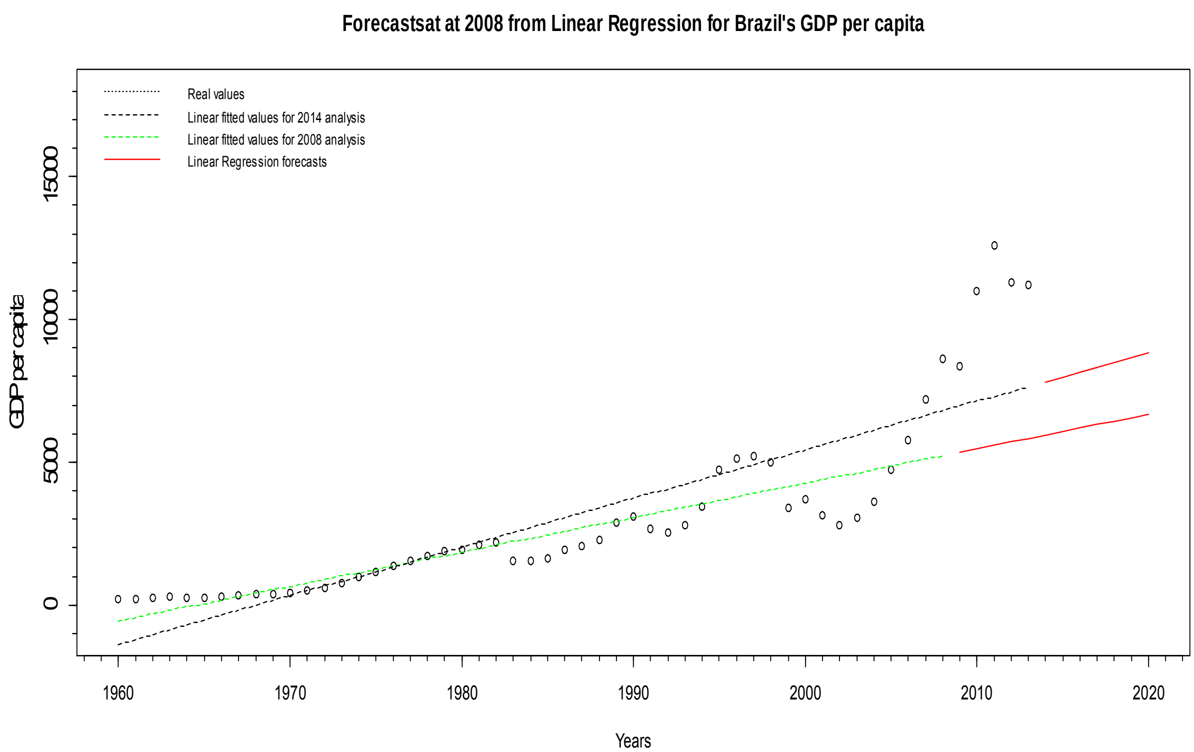
Appendix C – Graphs from analysis

C.1. Linear Regression

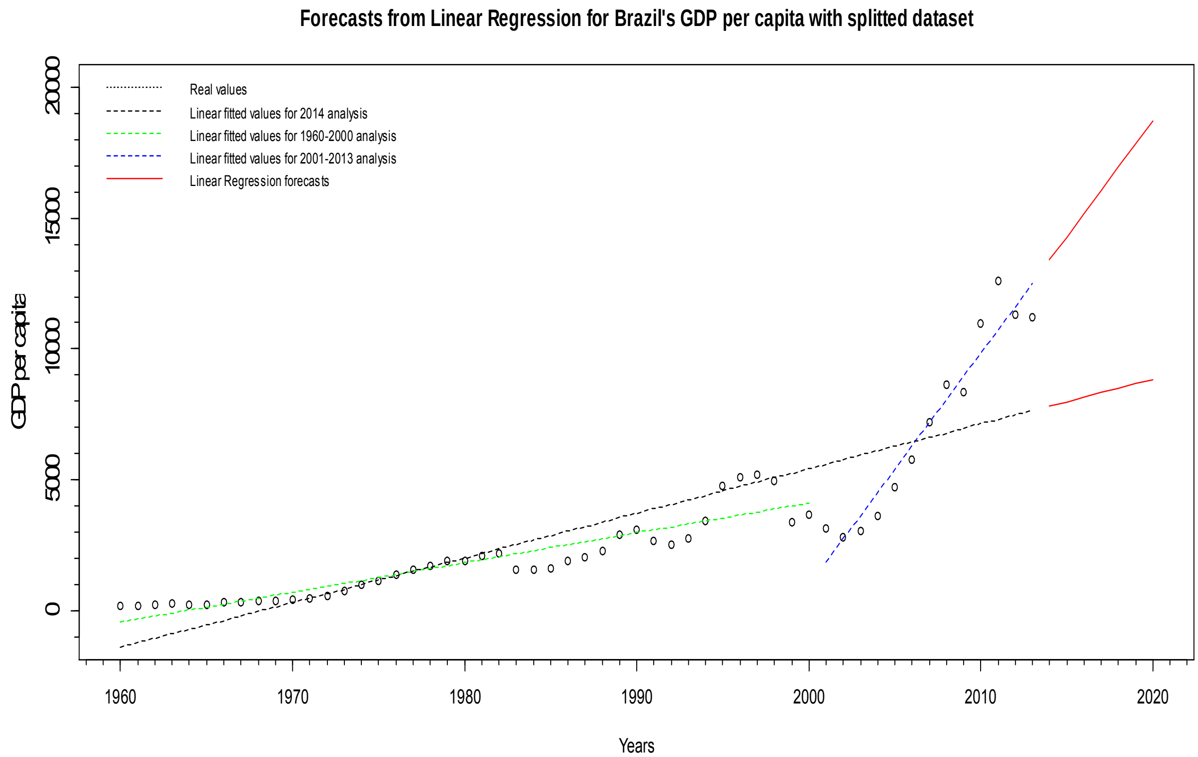
GDP per capita - Brazil



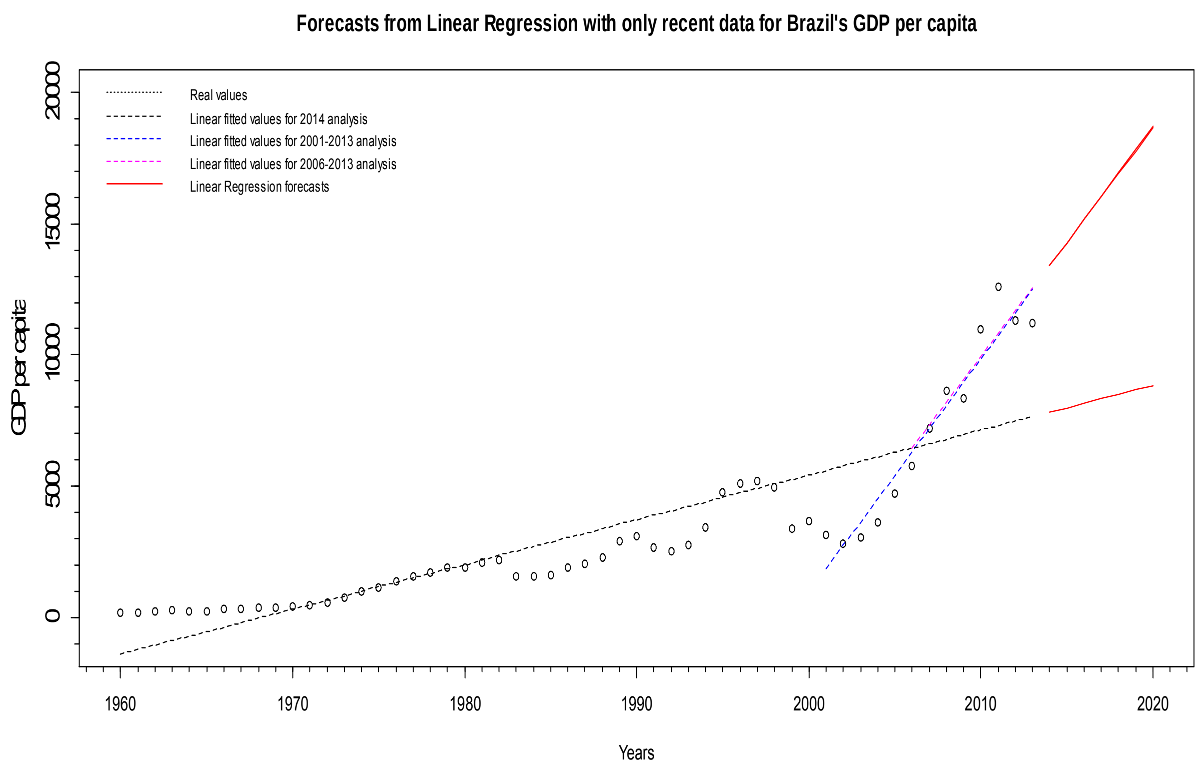
[Figure C.1.1] – Analysis for Brazil, GDP per capita and whole dataset



[Figure C.1.2] – Analysis for Brazil, GDP per capita and the dataset up to 2008

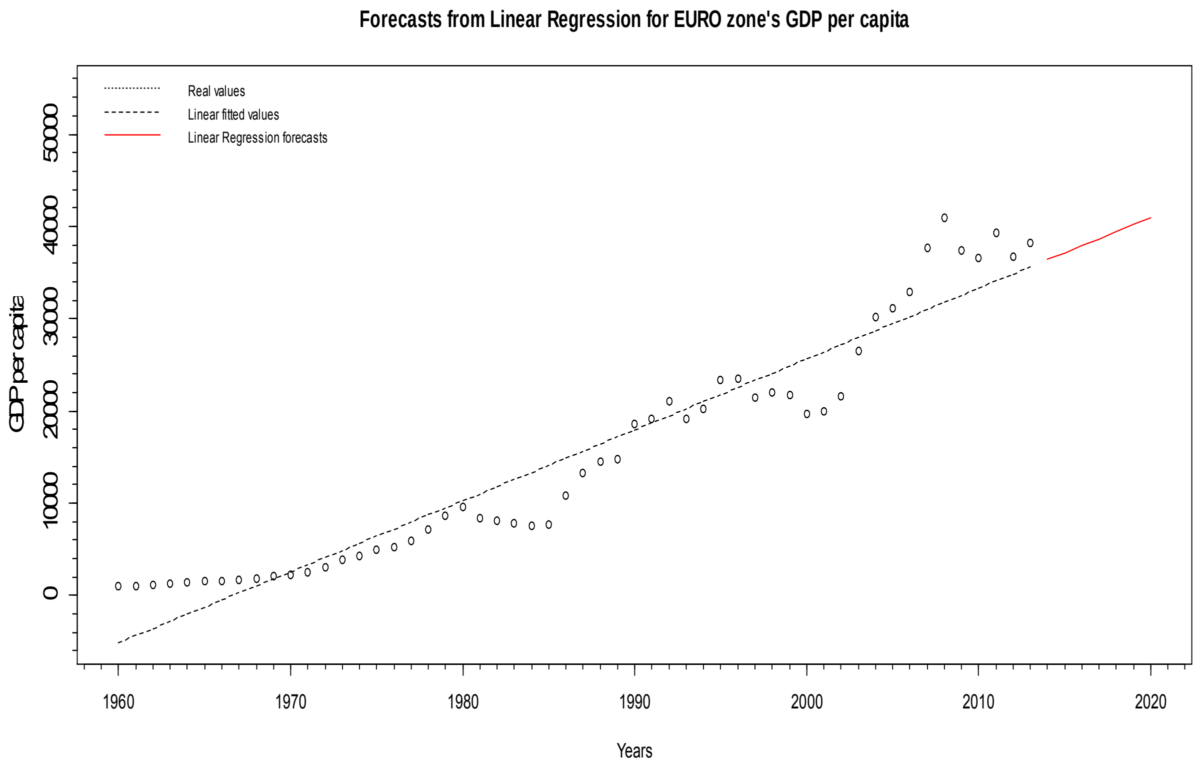


[Figure C.1.3] – Analyses for Brazil, GDP per capita and the subsets up to 2000 and 2001-2013

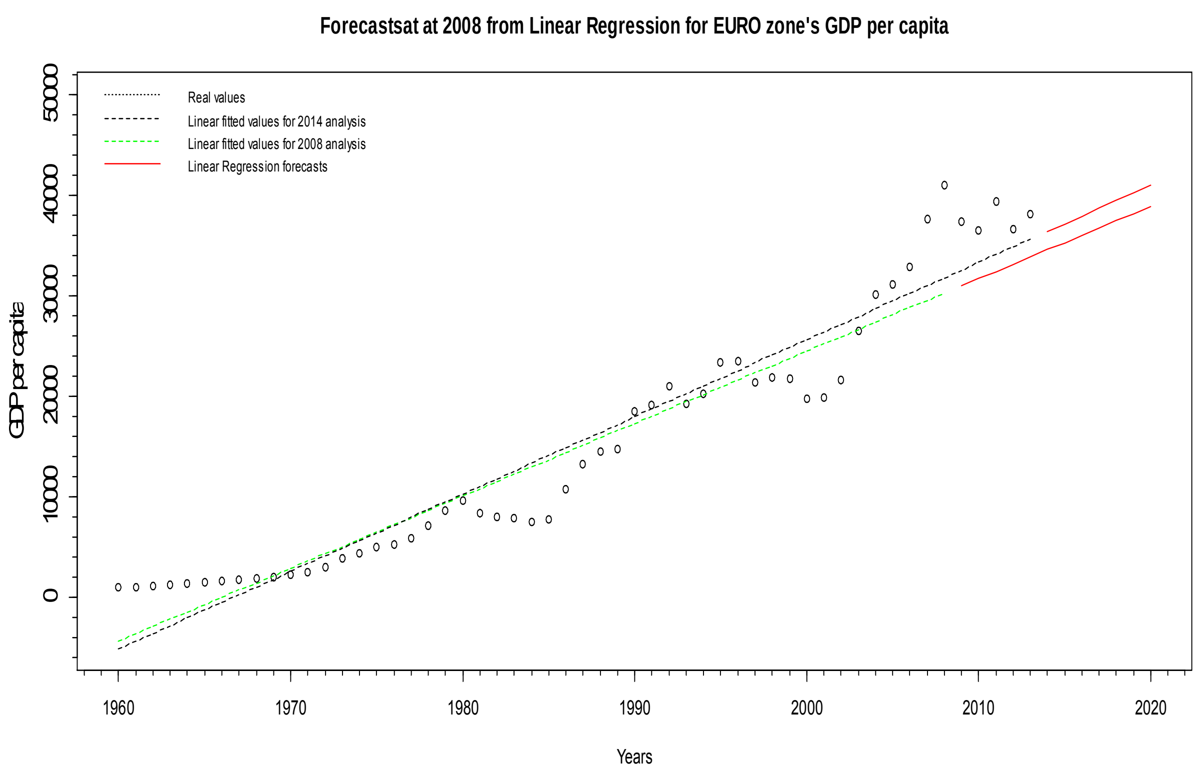


[Figure C.1.4] – Analyses for Brazil, GDP per capita and the subset from 2006-2013

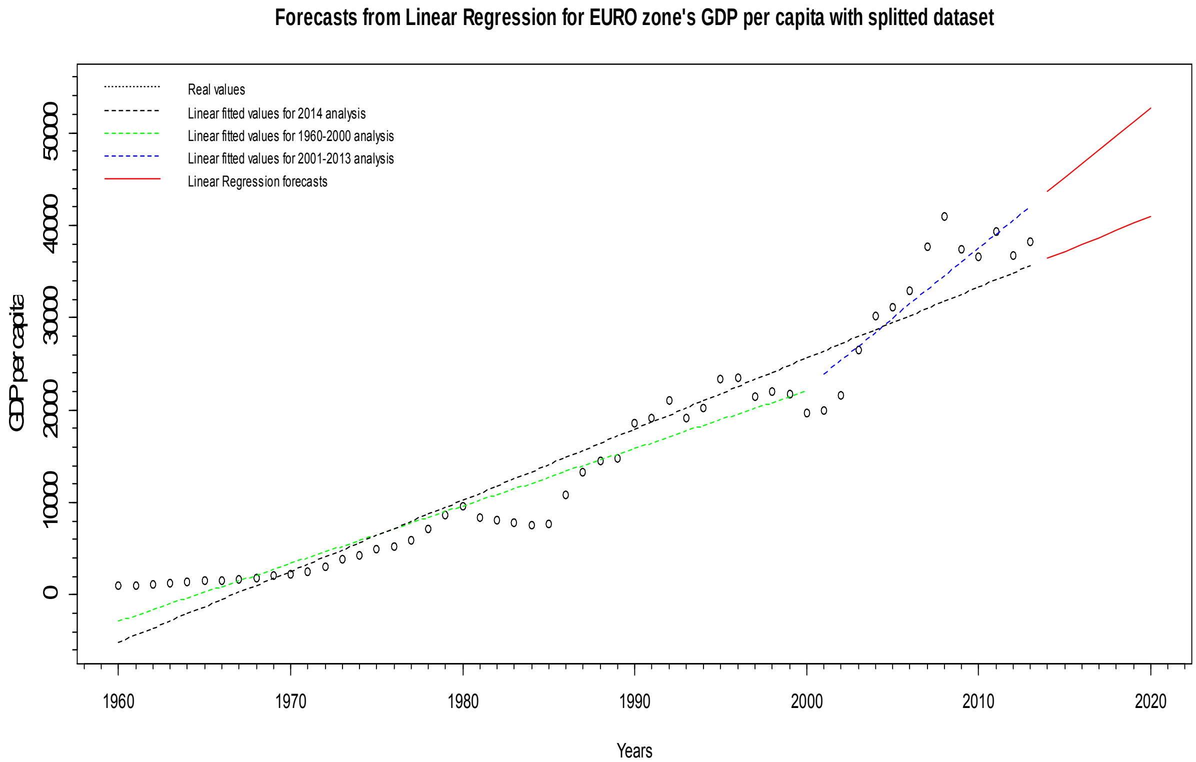
GDP per capita – EURO zone



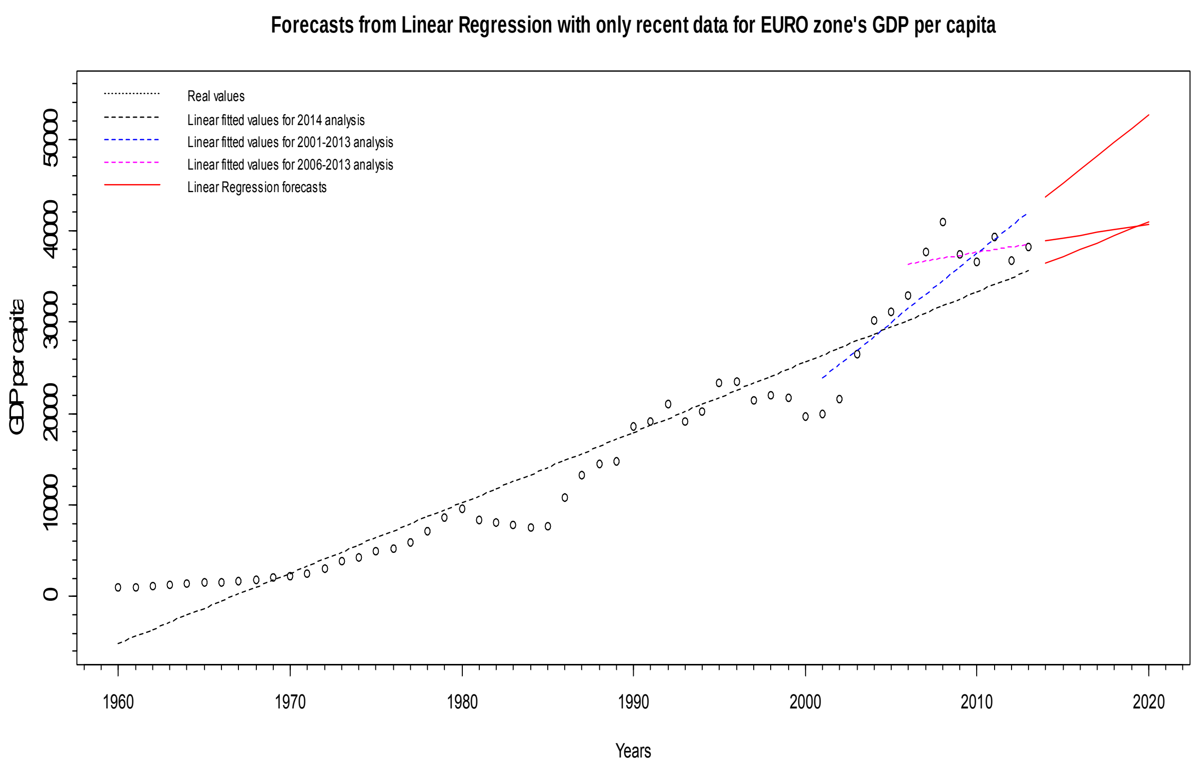
[Figure C.1.5] – Analysis for EURO zone, GDP per capita and whole dataset



[Figure C.1.6] – Analysis for EURO zone, GDP per capita and the dataset up to 2008

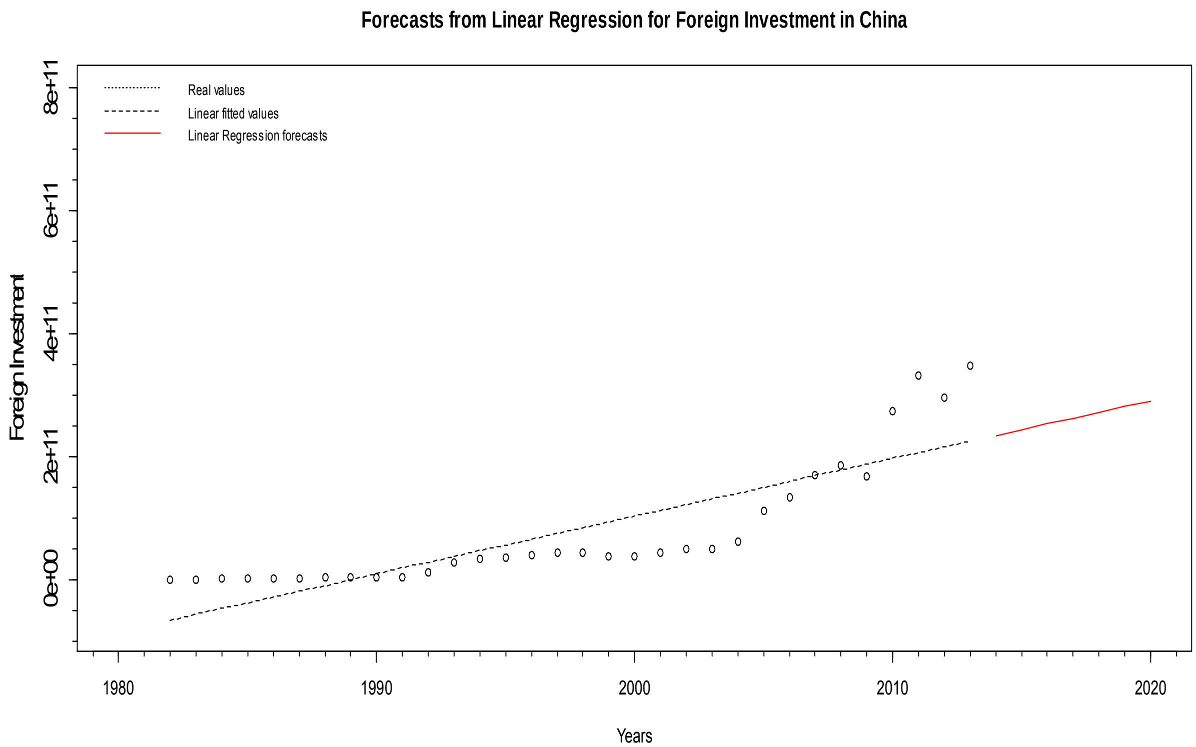


[Figure C.1.7] – Analyses for EURO zone, GDP per capita and the subsets up to 2000 and 2001-2013

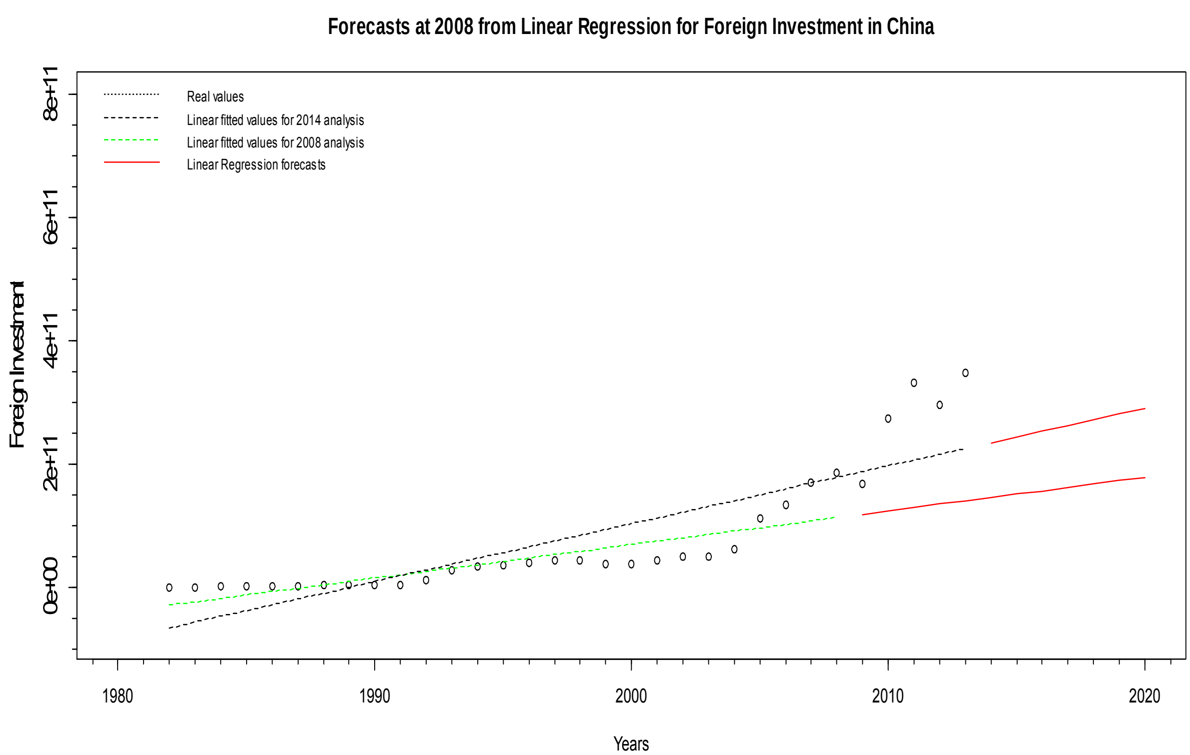


[Figure C.1.8] – Analyses for EURO zone, GDP per capita and the subset from 2006-2013

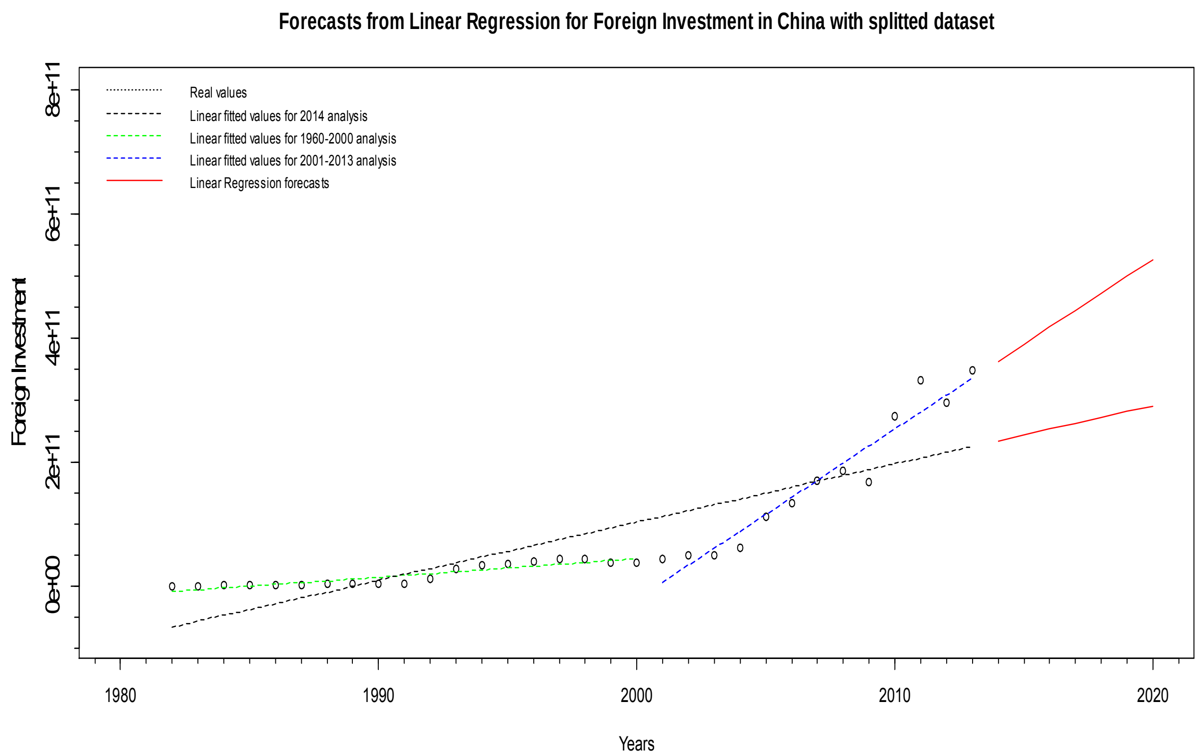
Foreign Investment – China



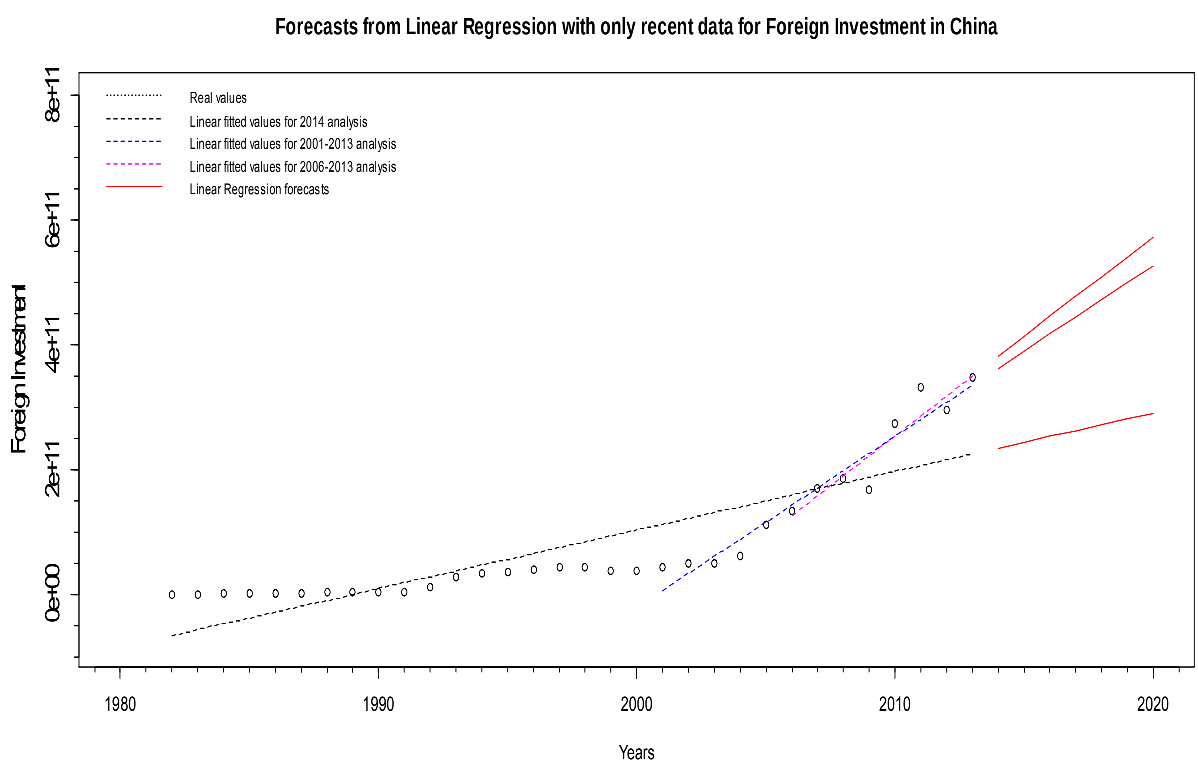
[Figure C.1.9] – Analysis for China, Foreign Investment and whole dataset



[Figure C.1.10] – Analysis for China, Foreign Investment and the dataset up to 2008

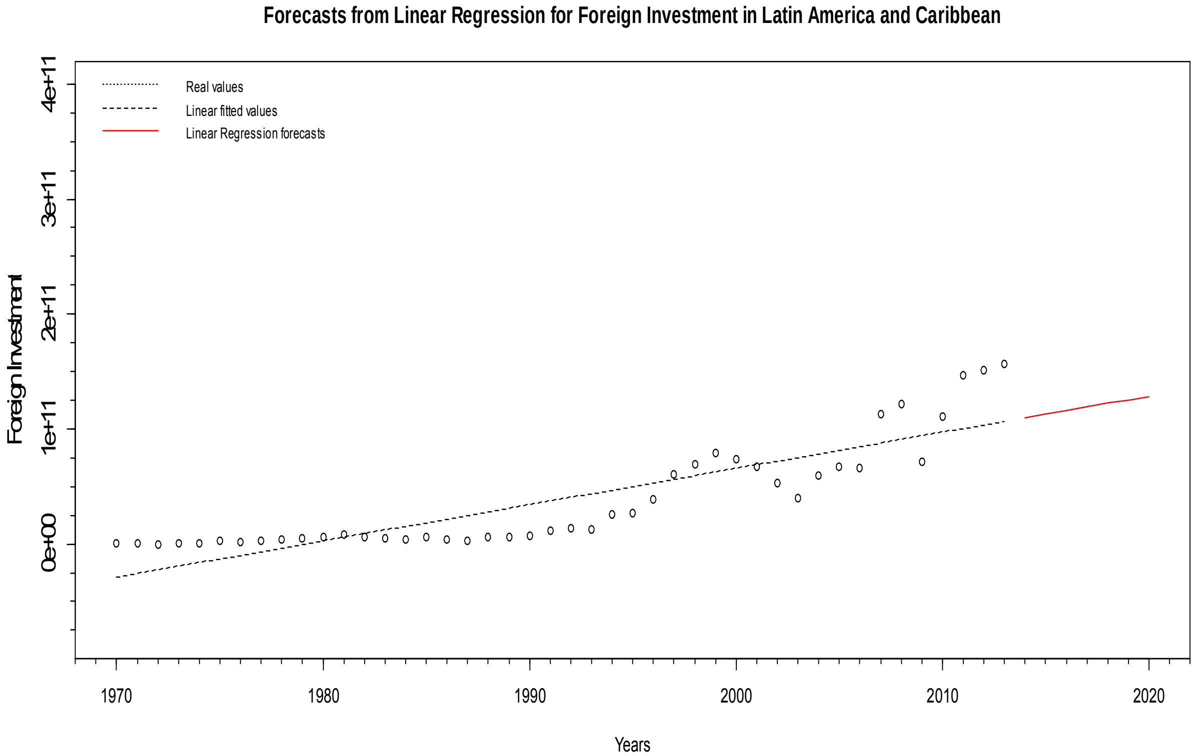


[Figure C.1.11] – Analyses for China, Foreign Investment and subsets up to 2000 and 2001-2013

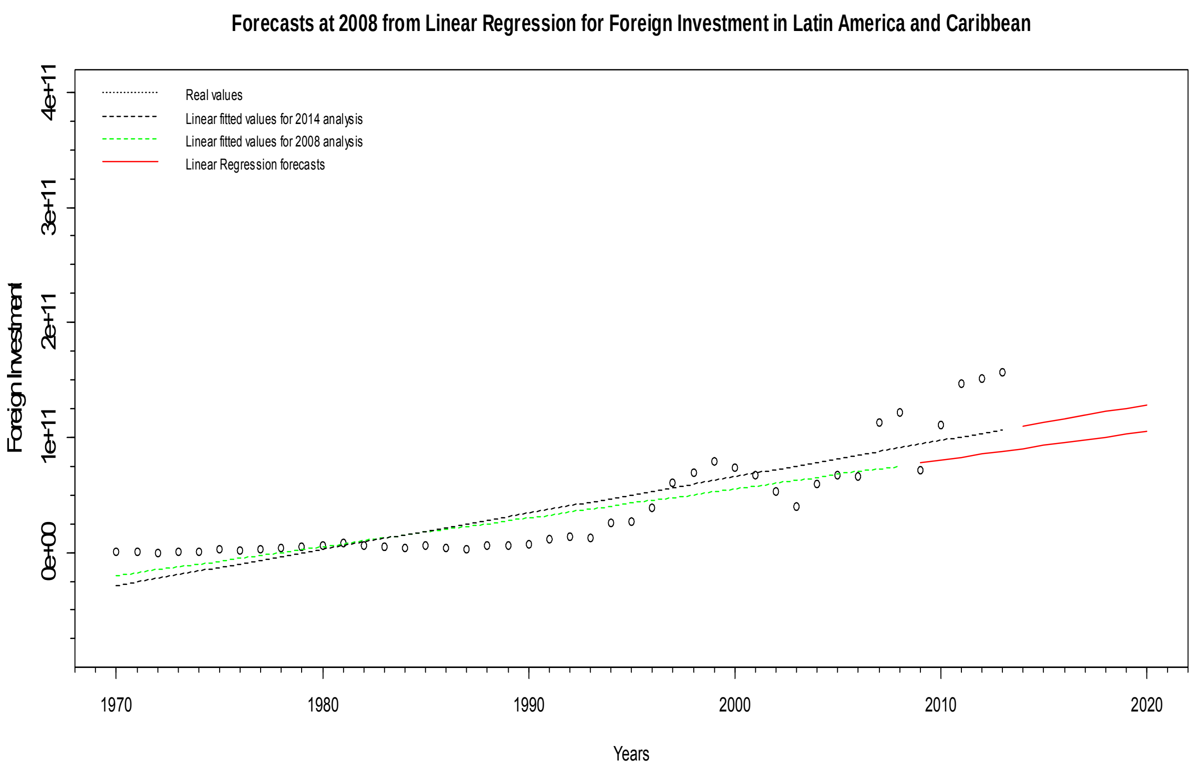


[Figure C.1.12] – Analyses for China, Foreign Investment and the subset from 2006-2013

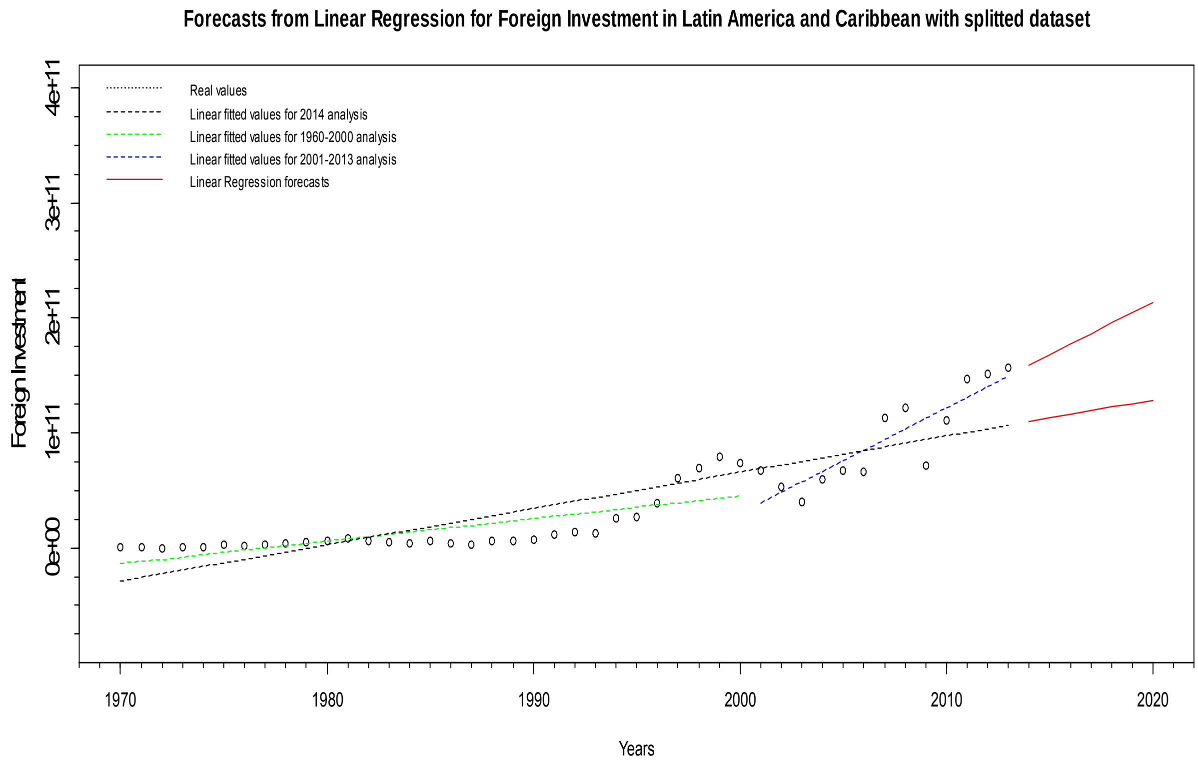
Foreign Investment – Latin America and Caribbean



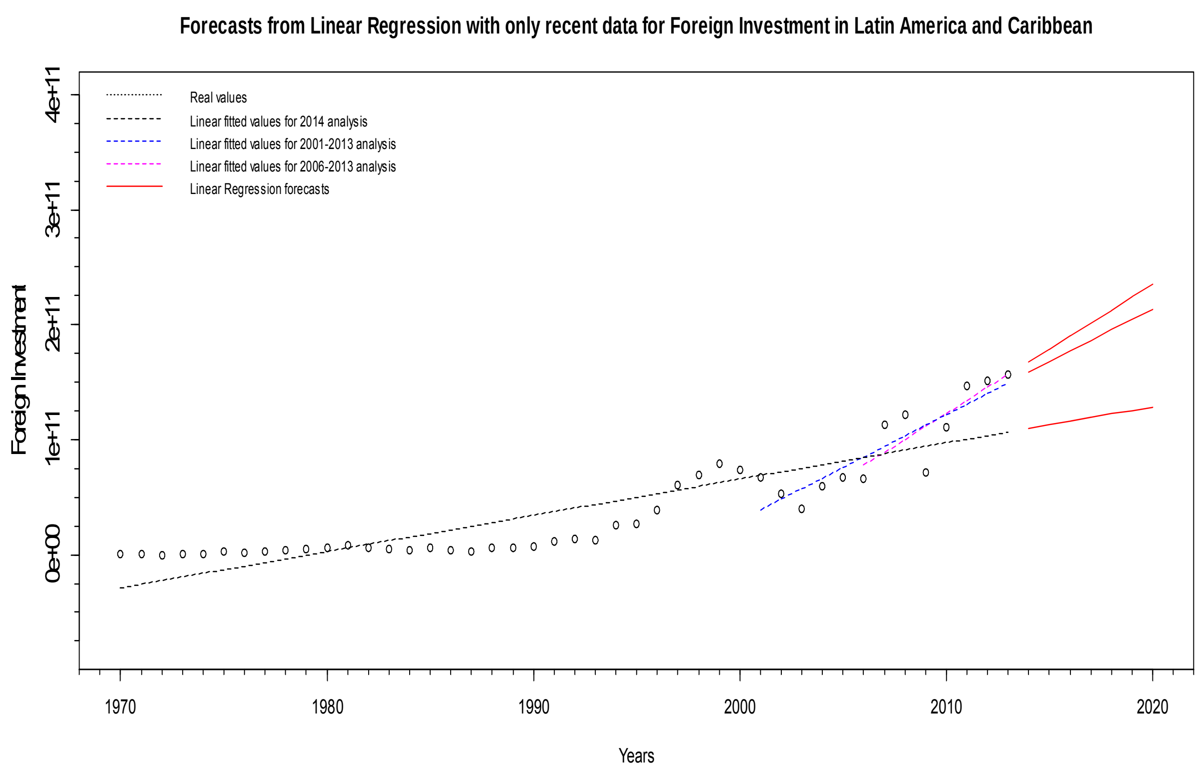
[Figure C.1.13] – Analysis for Latin America and Caribbean, Foreign Investment and whole dataset



[Figure C.1.14] – Analysis for Latin America, Foreign Investment and the dataset up to 2008

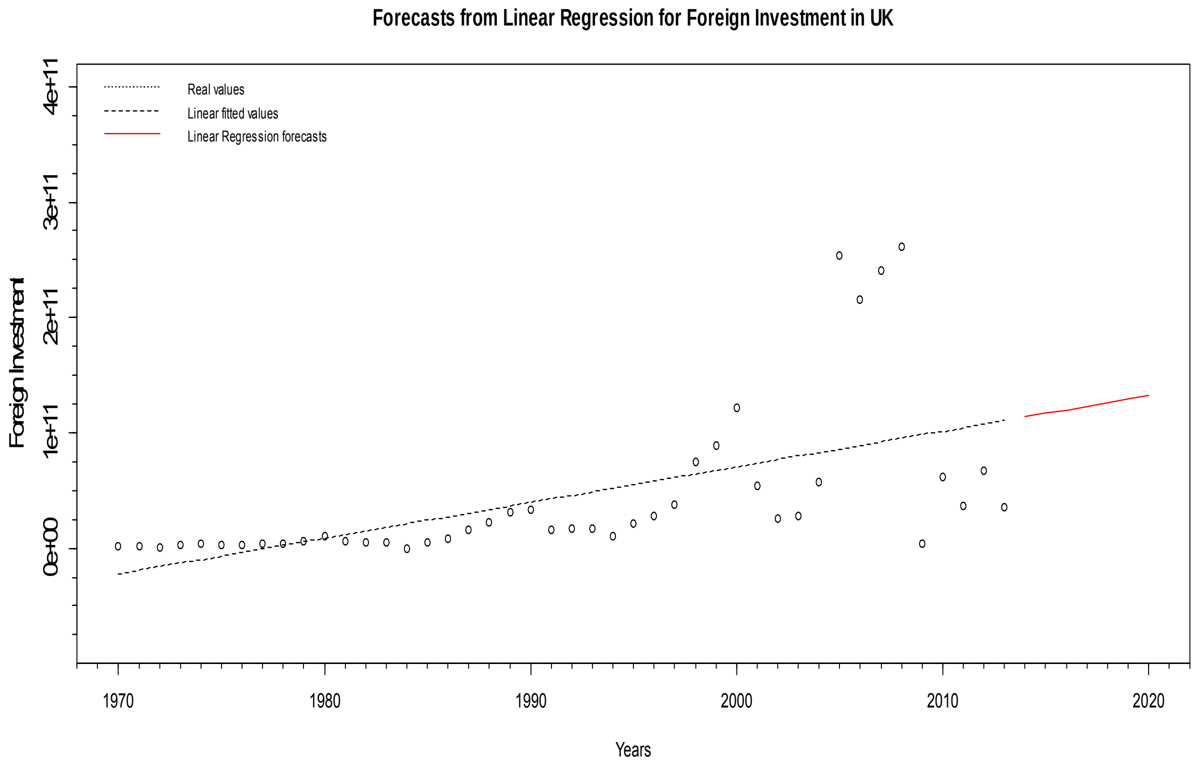


[Figure C.1.15] – Analyses for Latin America and Caribbean, Foreign Investment and the subsets up to 2000 and 2001-2013

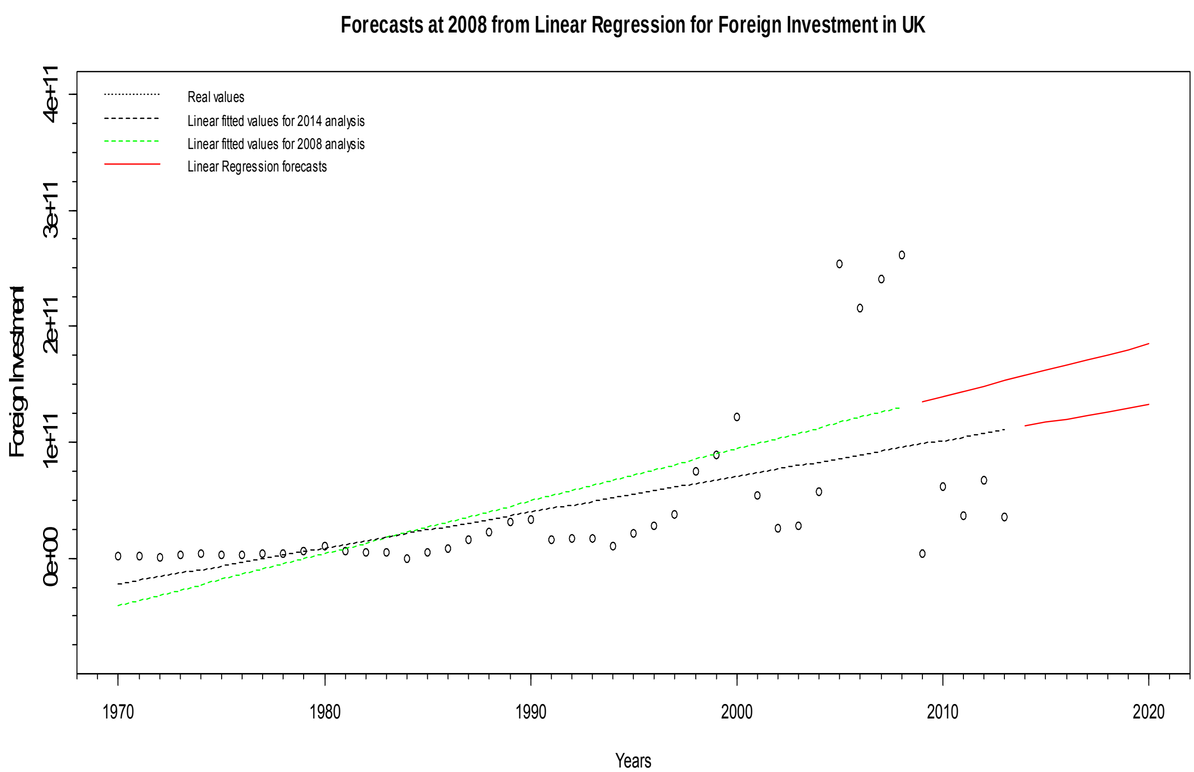


[Figure C.1.16] – Analyses for Latin America, Foreign Investment and the subset from 2006-2013

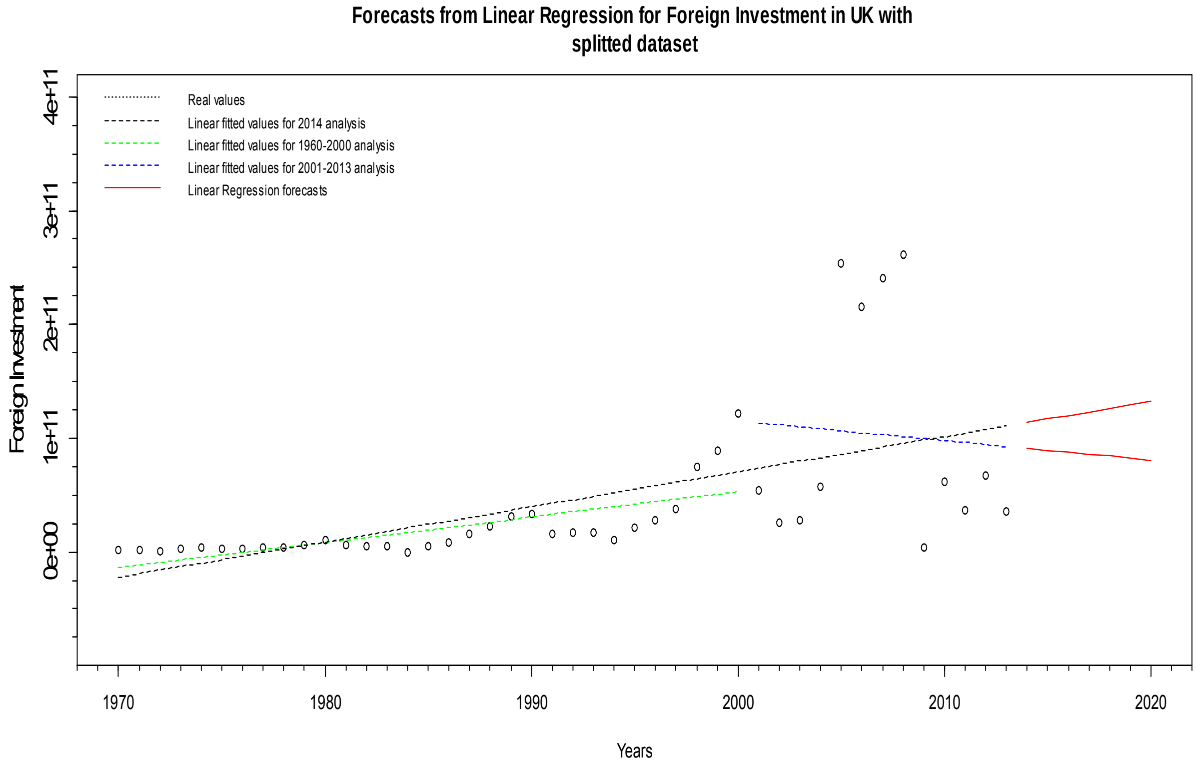
Foreign Investment – UK



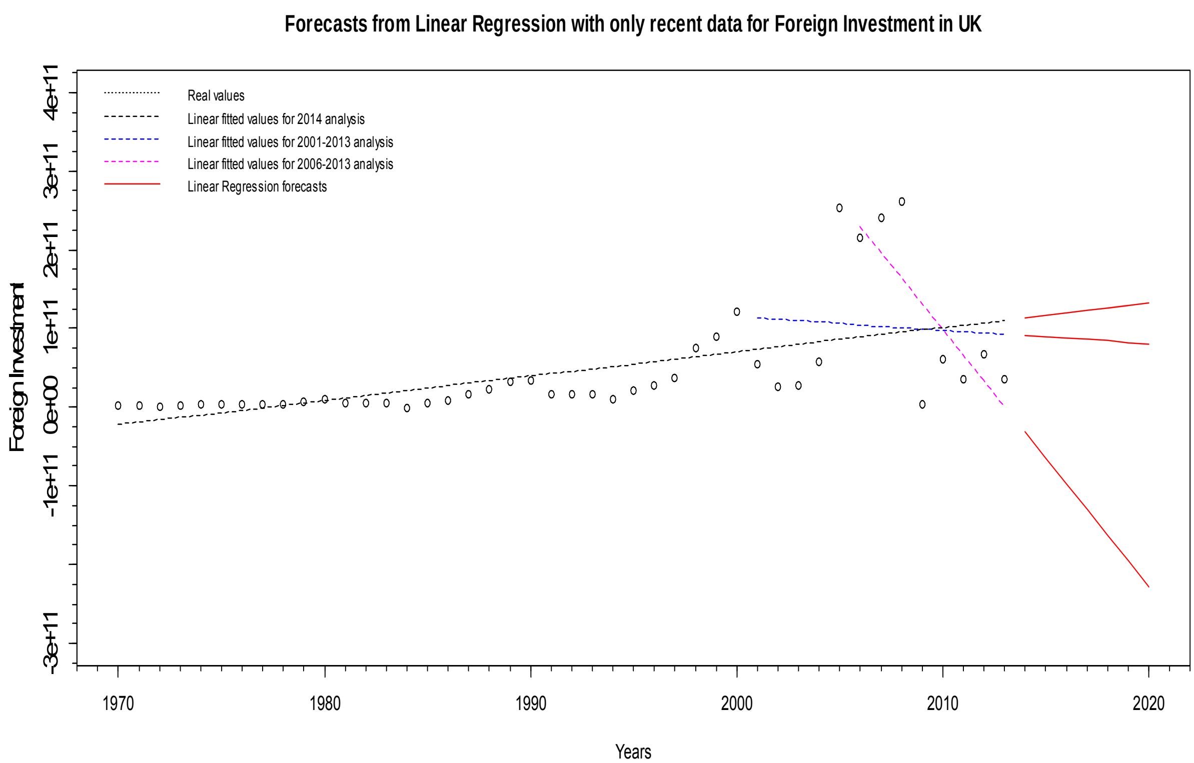
[Figure C.1.17] – Analysis for UK, Foreign Investment and whole dataset



[Figure C.1.18] – Analysis for UK, Foreign Investment and the dataset up to 2008

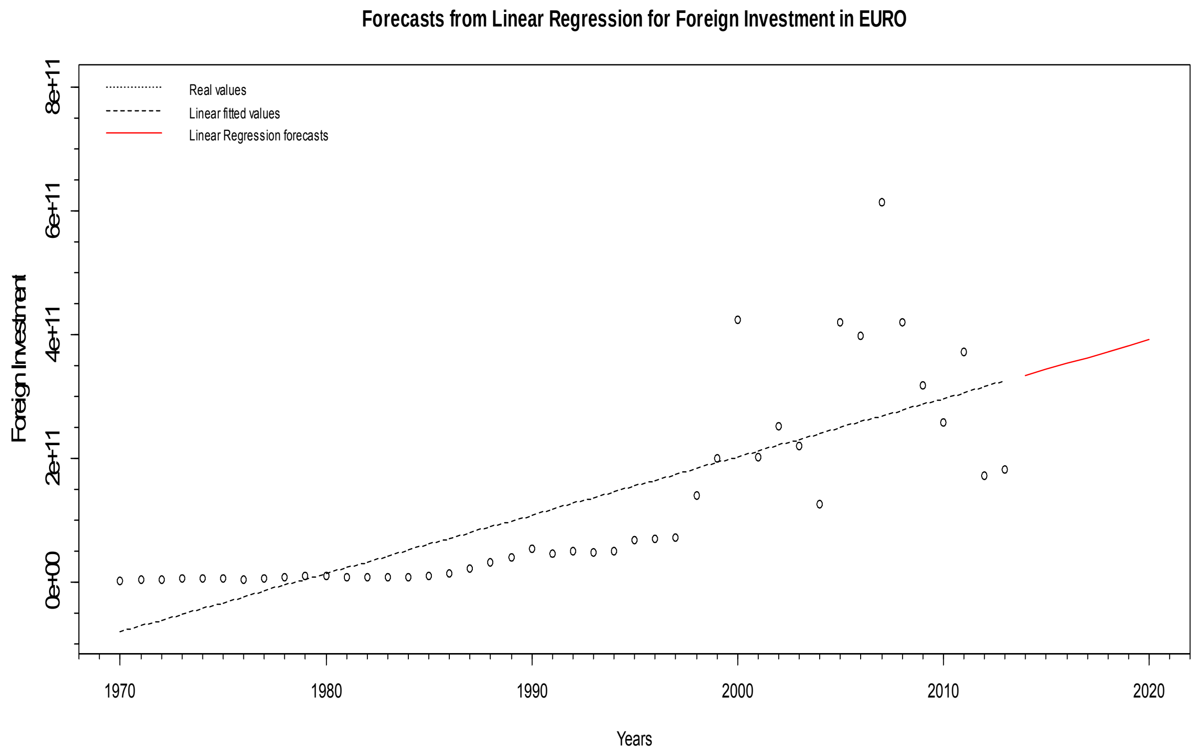


[Figure C.1.19] – Analyses for UK, Foreign Investment and the subsets up to 2000 and 2001-2013

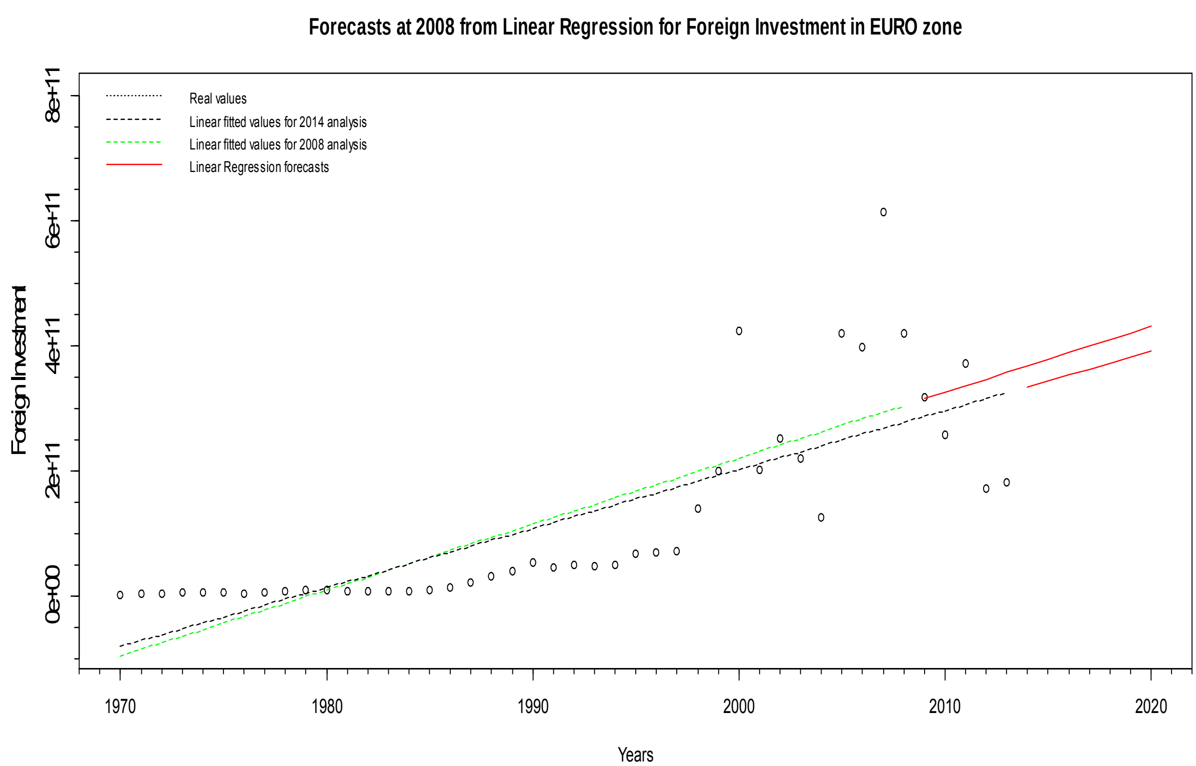


[Figure C.1.20] – Analyses for UK, Foreign Investment and the subset from 2006-2013

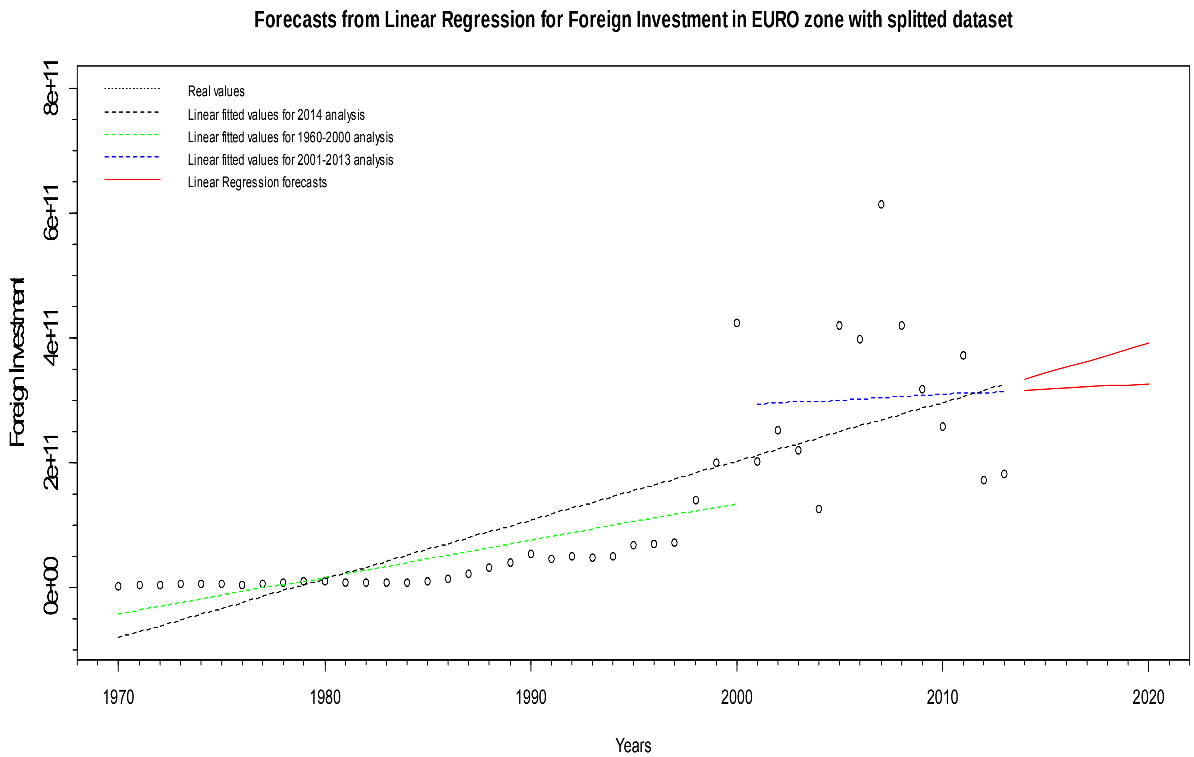
__Foreign Investment – EURO zone__



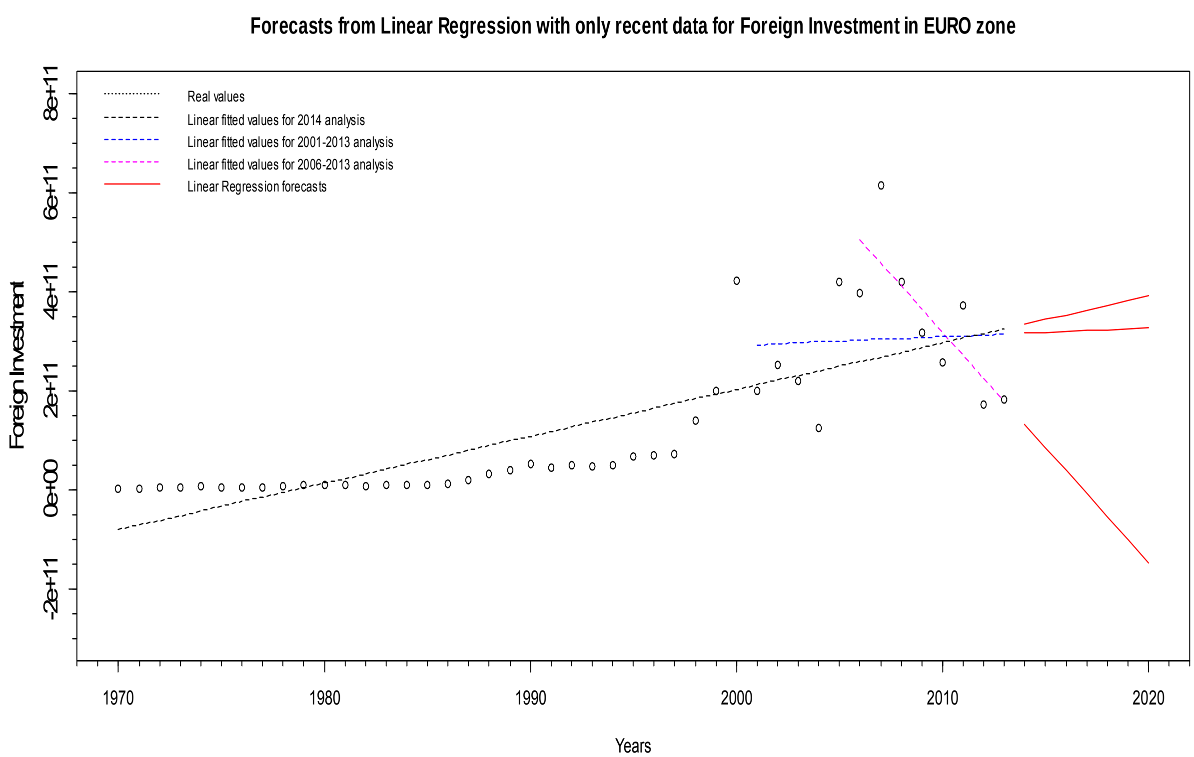
[Figure C.1.21] – Analysis for EURO zone, Foreign Investment and whole dataset



[Figure C.1.22] – Analysis for EURO zone, Foreign Investment and the dataset up to 2008

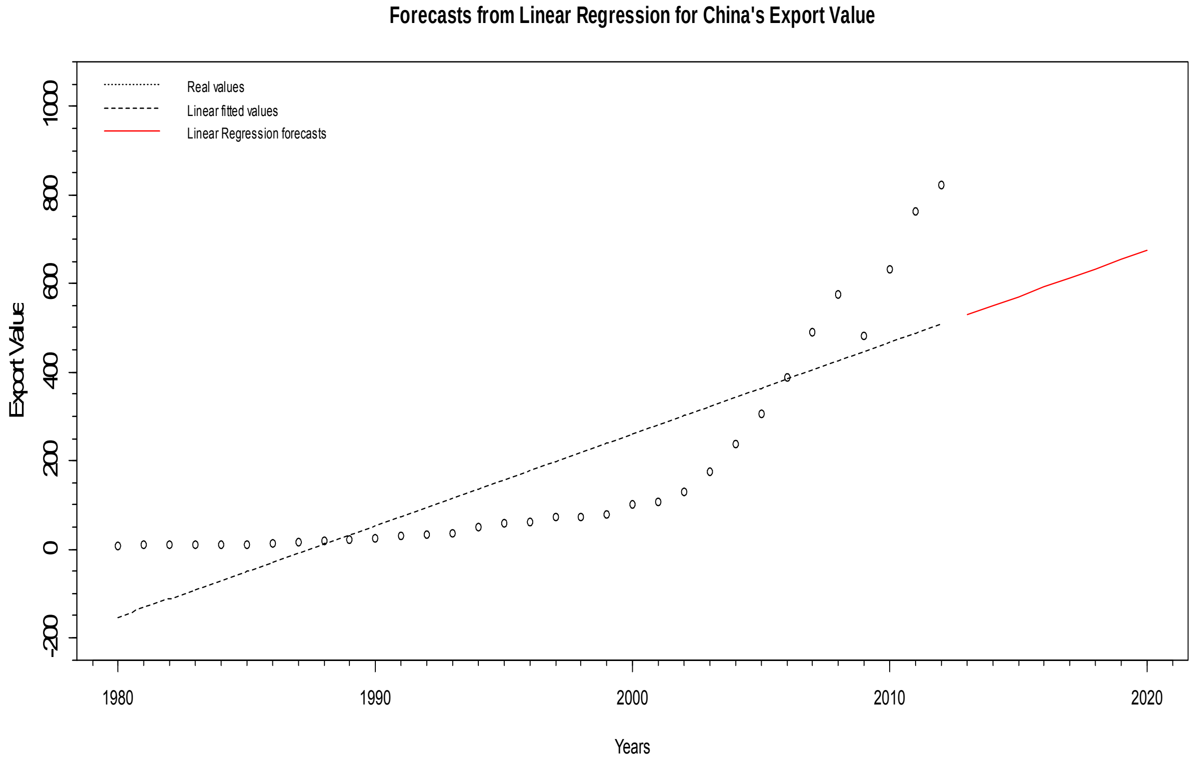


[Figure C.1.23] – Analyses for EURO zone, Foreign Investment and the subsets up to 2000 and 2001-2013

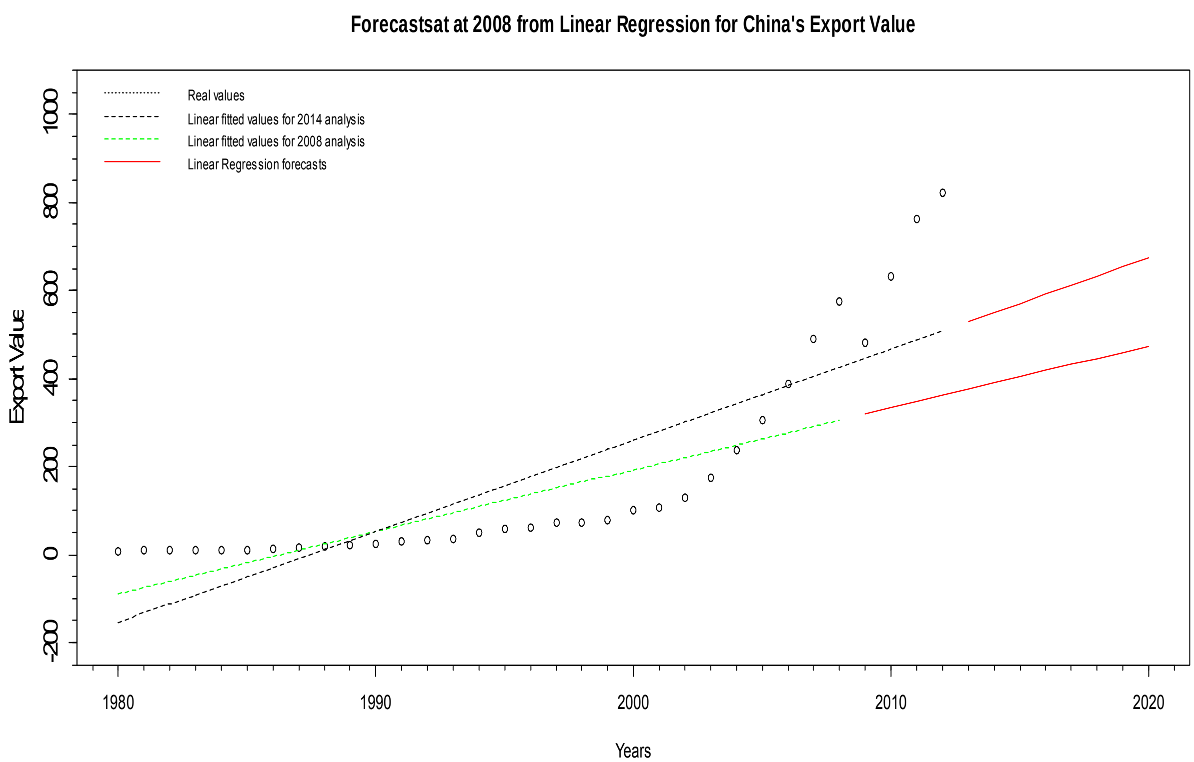


[Figure C.1.24] – Analyses for EURO zone, Foreign Investment and the subset from 2006-2013

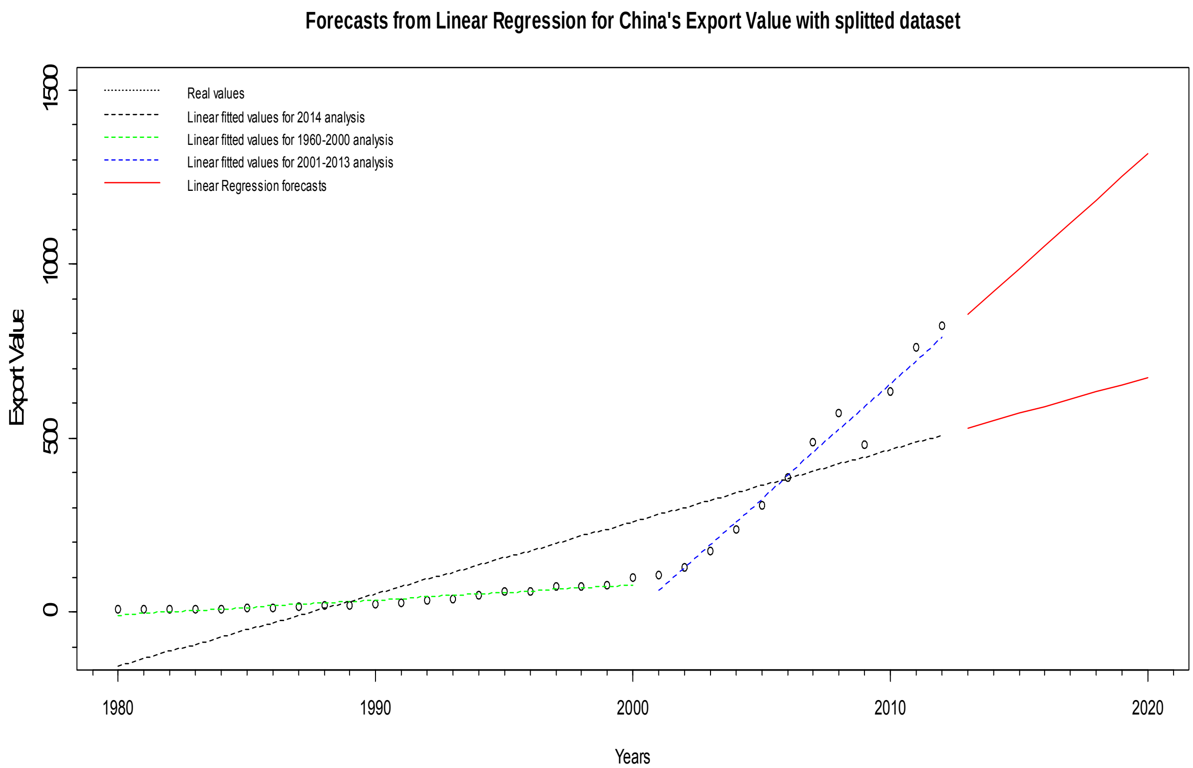
Export Value – China



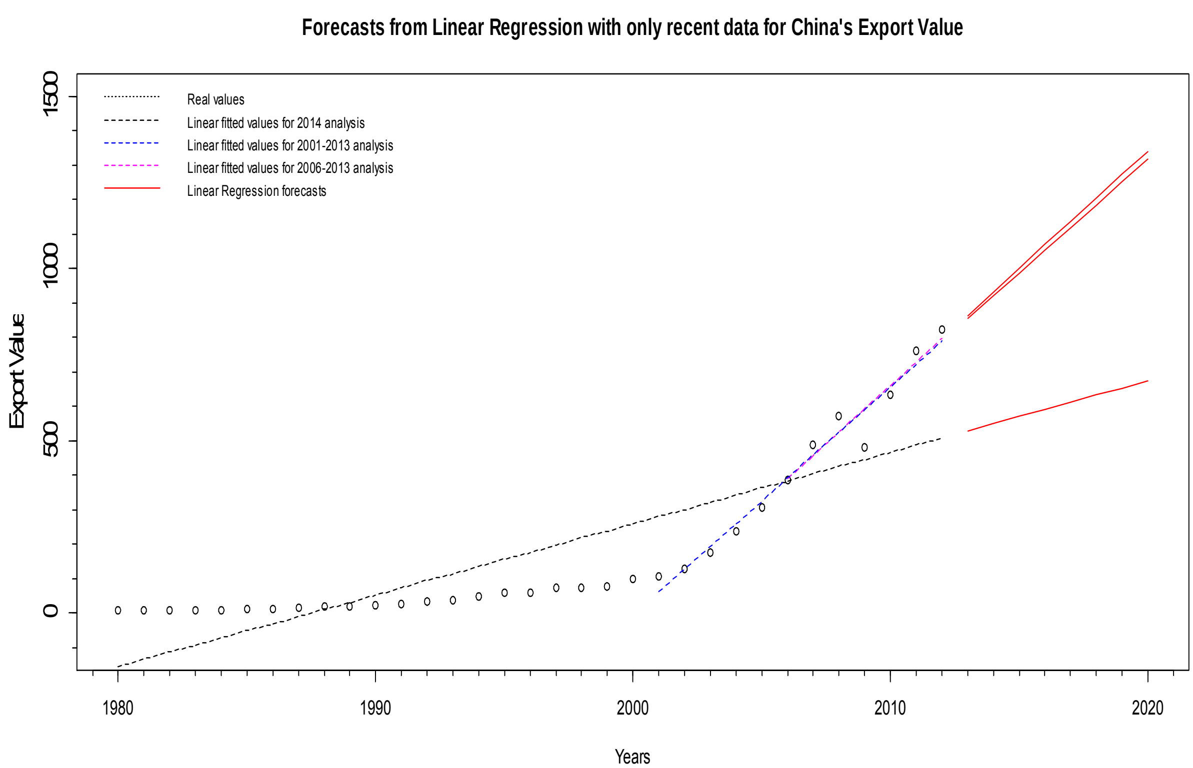
[Figure C.1.25] – Analysis for China, Export Value and whole dataset



[Figure C.1.26] – Analysis for China, Export Value and the dataset up to 2008

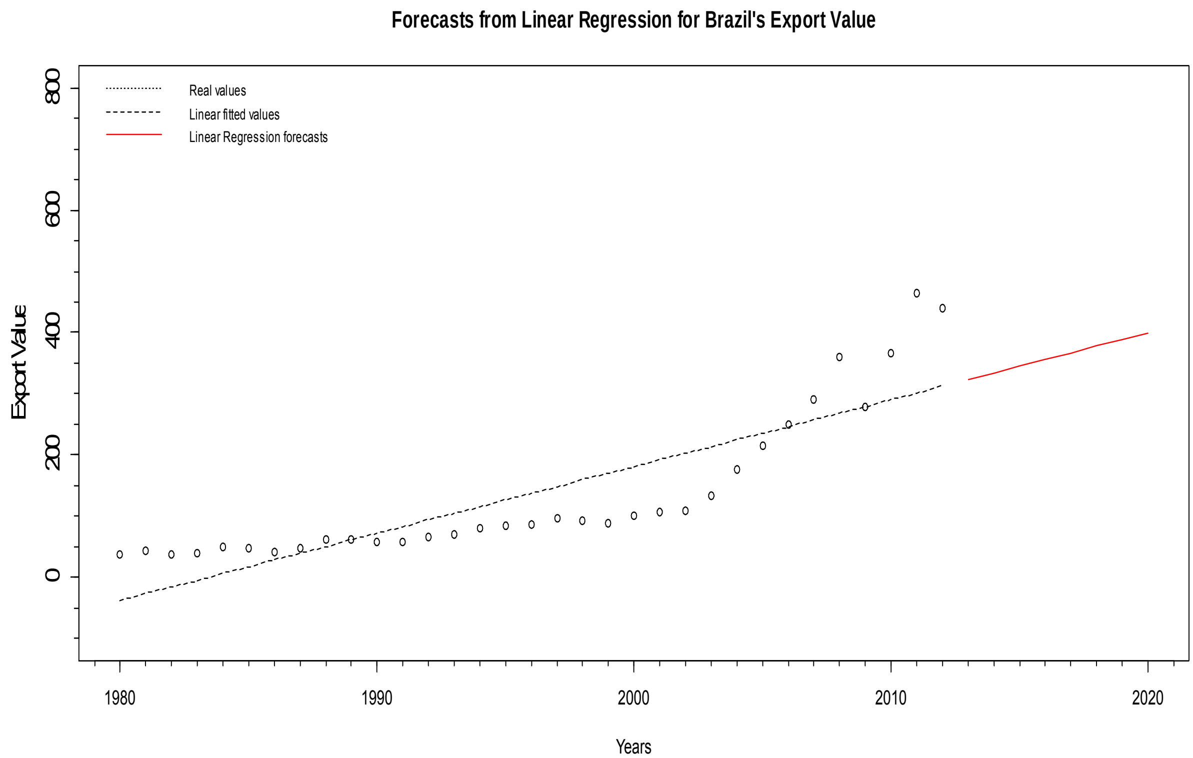


[Figure C.1.27] – Analyses for China, Export Value and the subsets up to 2000 and 2001-2013

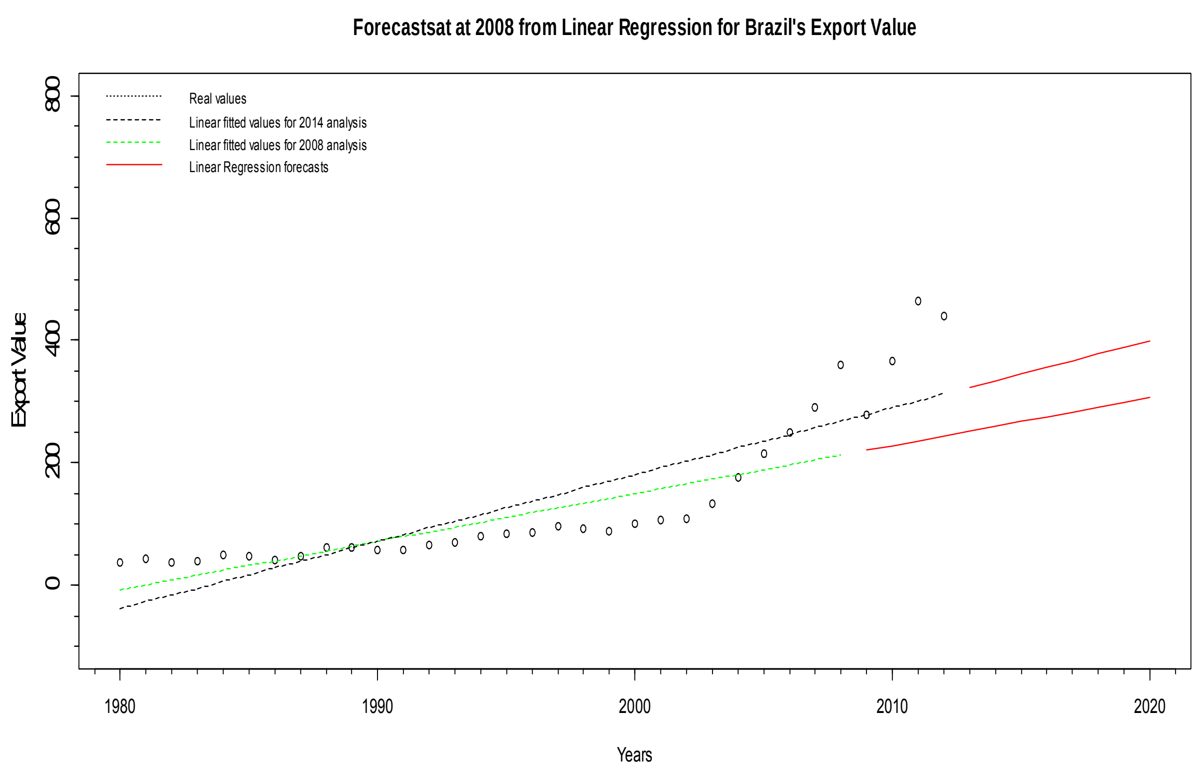


[Figure C.1.28] – Analyses for China, Export Value and the subset from 2006-2013

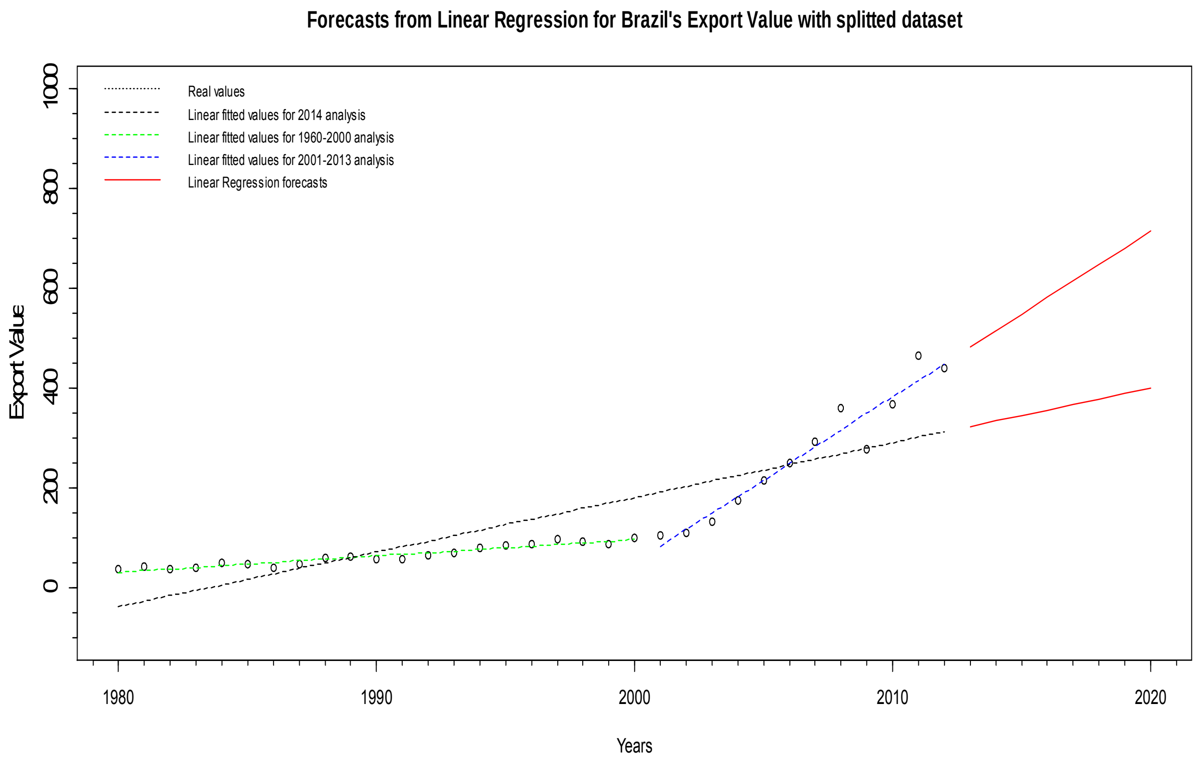
Export Value – Brazil



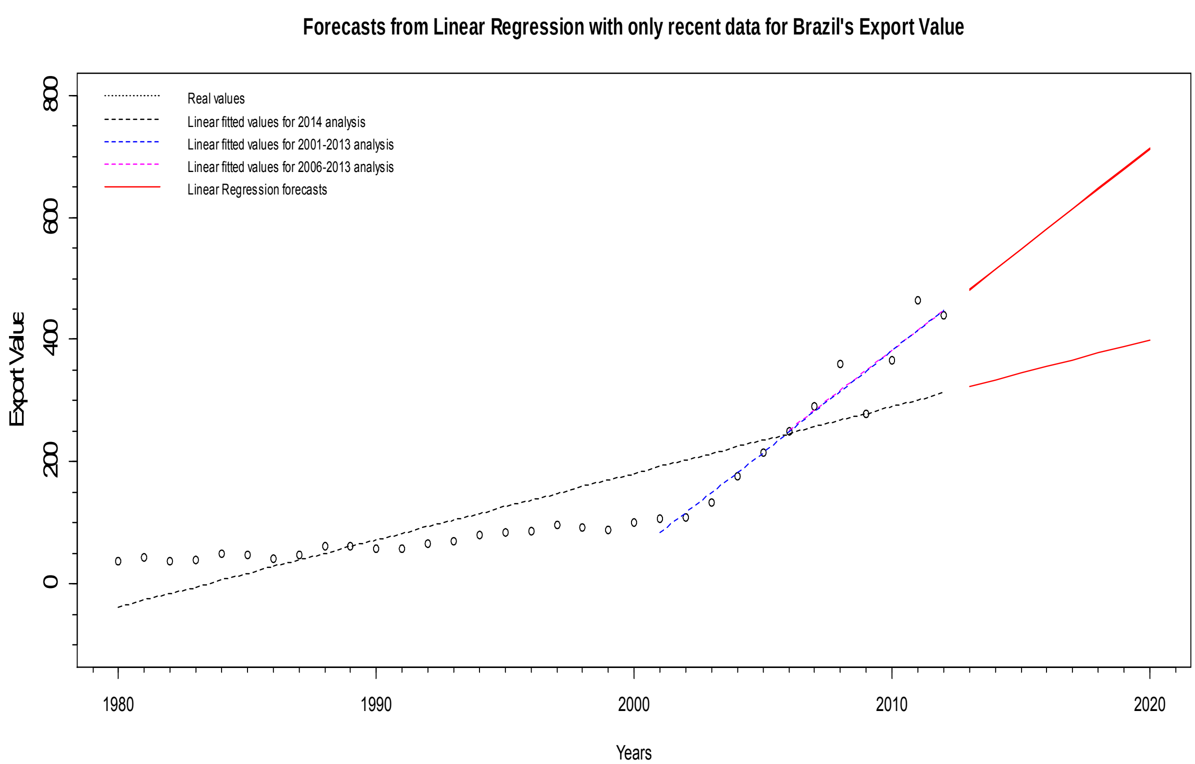
[Figure C.1.29] – Analysis for Brazil, Export Value and whole dataset



[Figure C.1.30] – Analysis for Brazil, Export Value and the dataset up to 2008

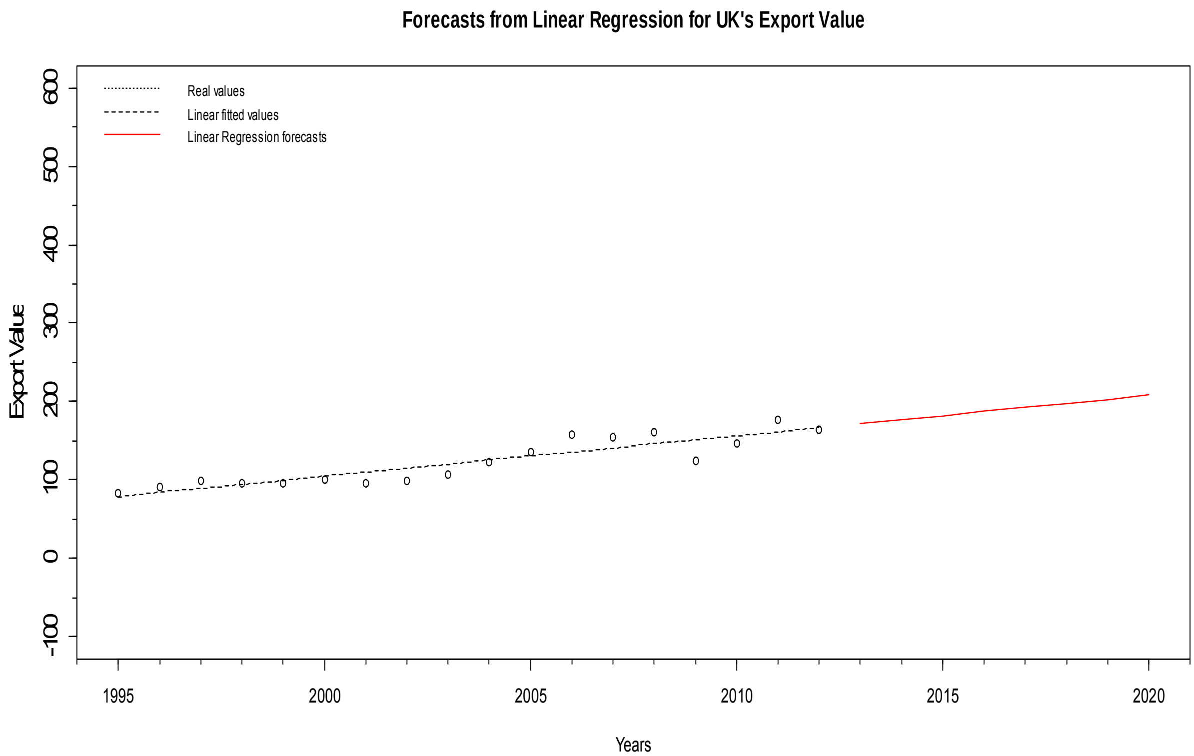


[Figure C.1.31] – Analyses for Brazil, Export Value and the subsets up to 2000 and 2001-2013

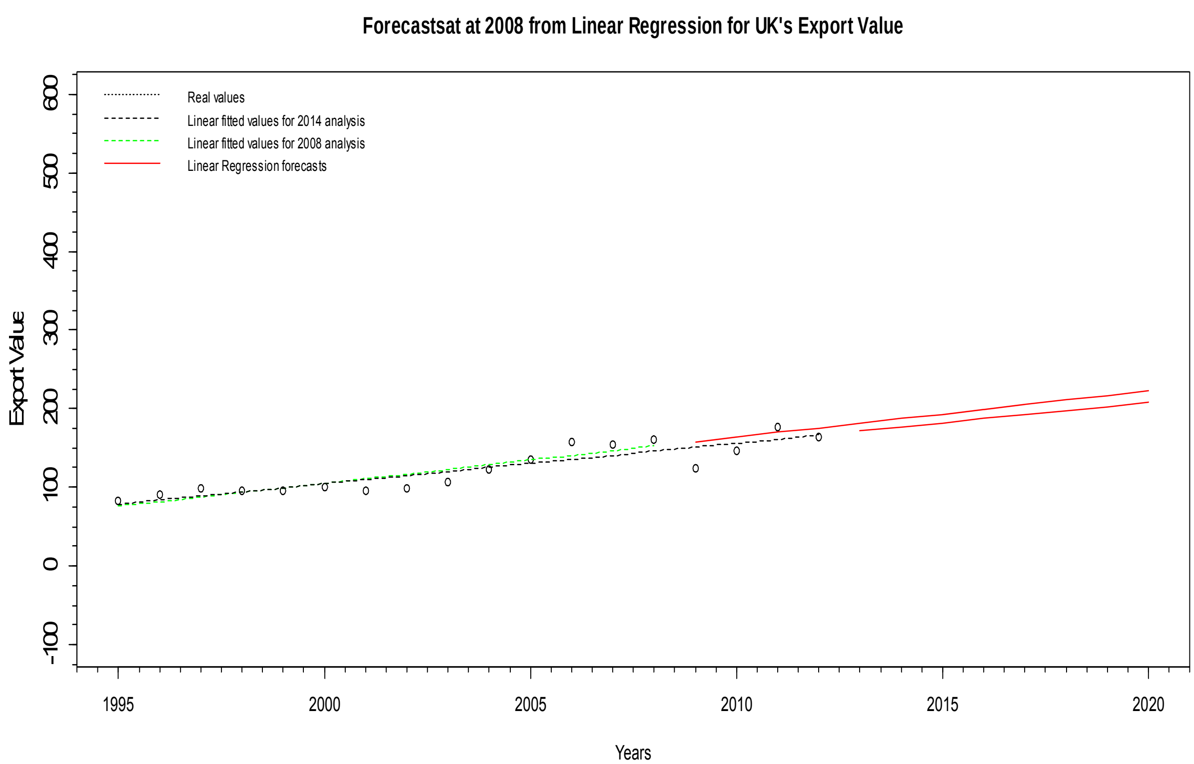


[Figure C.1.32] – Analyses for Brazil, Export Value and the subset from 2006-2013

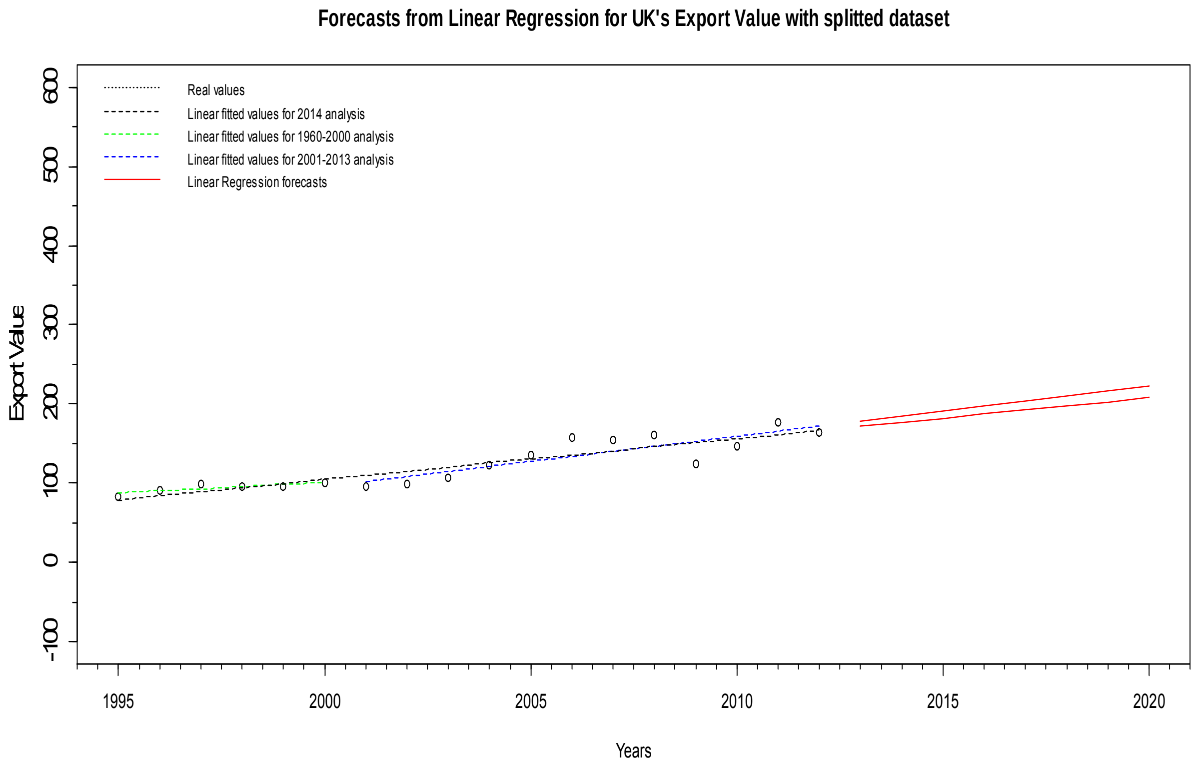
Export Value – UK



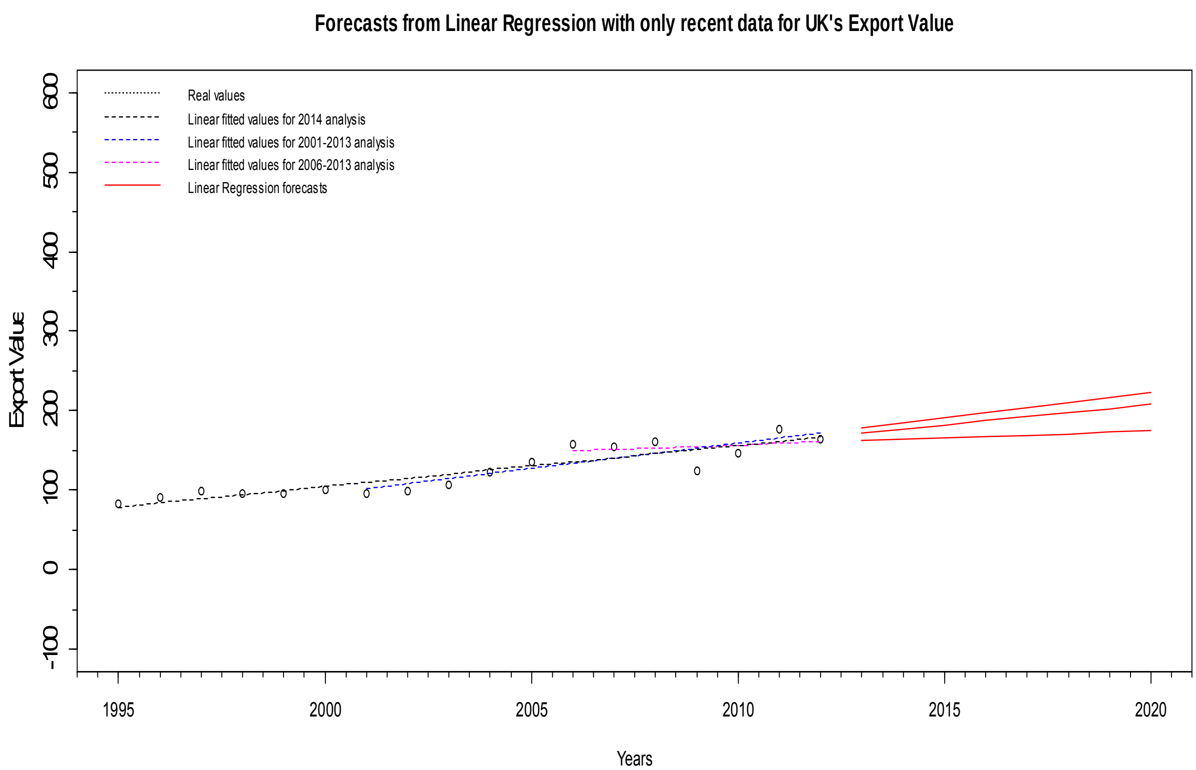
[Figure C.1.33] – Analysis for UK, Export Value and whole dataset



[Figure C.1.34] – Analysis for UK, Export Value and the dataset up to 2008

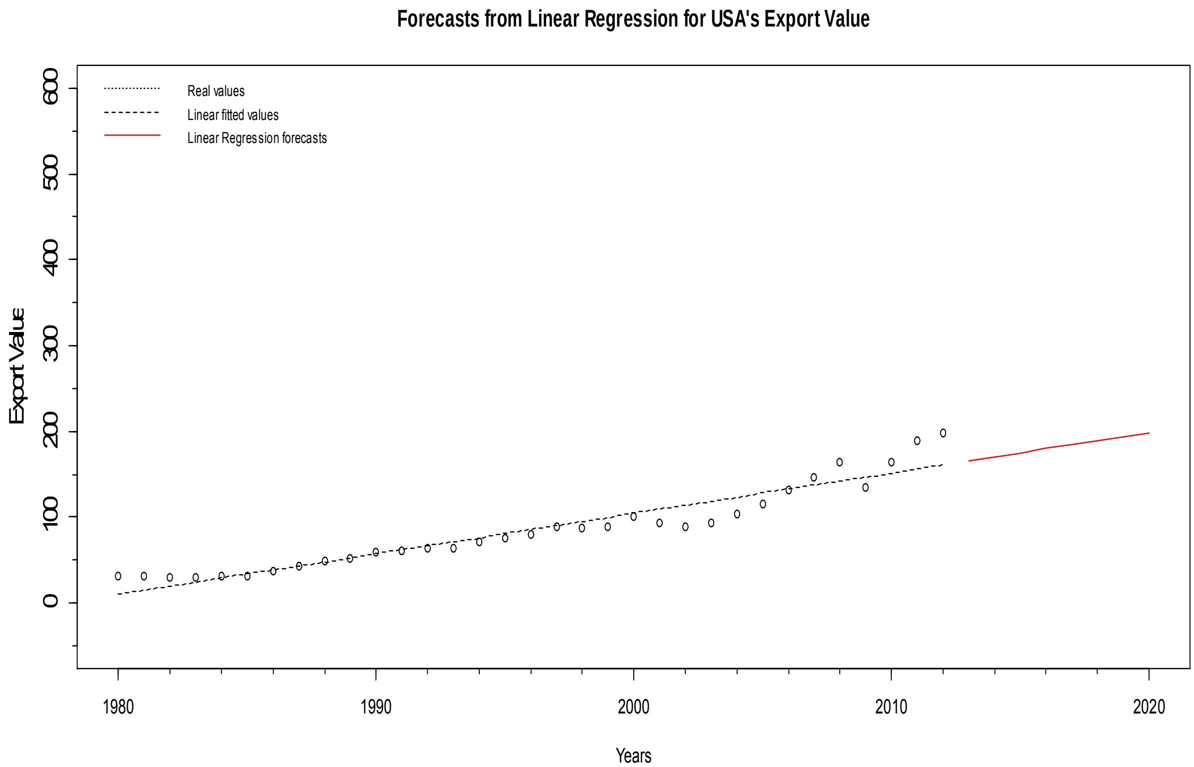


[Figure C.1.35] – Analyses for UK, Export Value and the subsets up to 2000 and 2001-2013

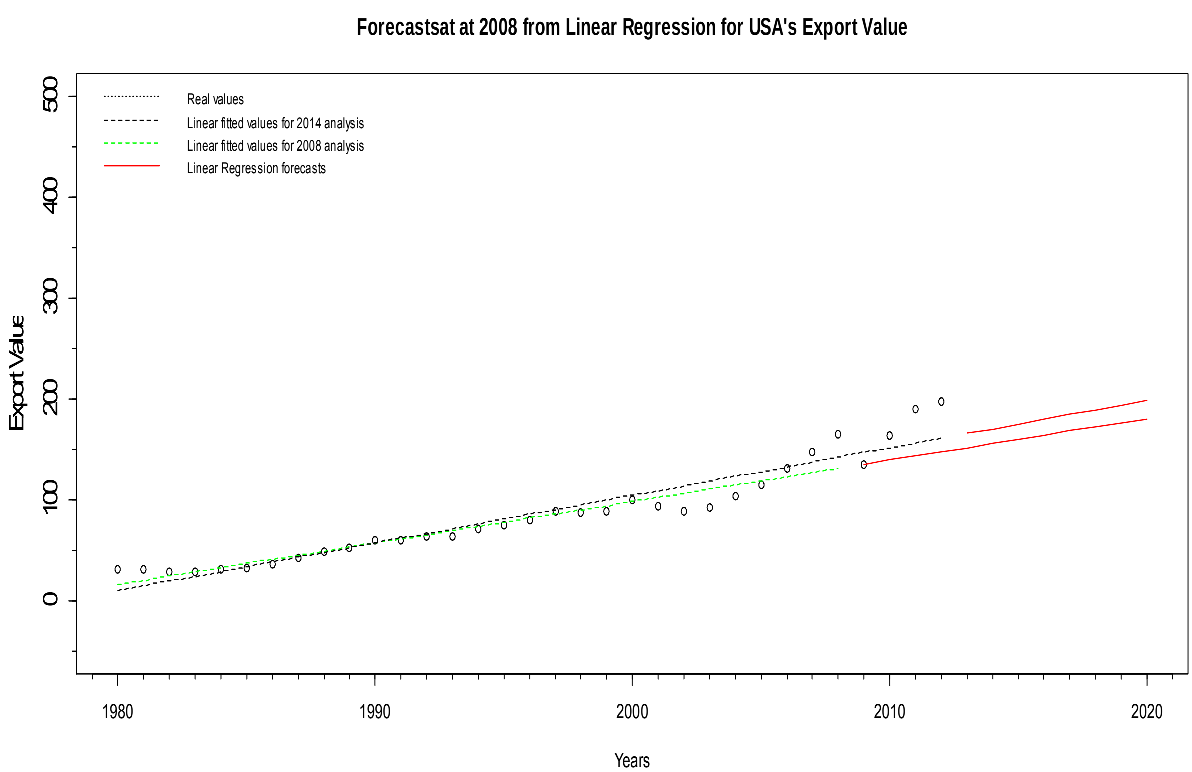


[Figure C.1.36] – Analyses for UK, Export Value and the subset from 2006-2013

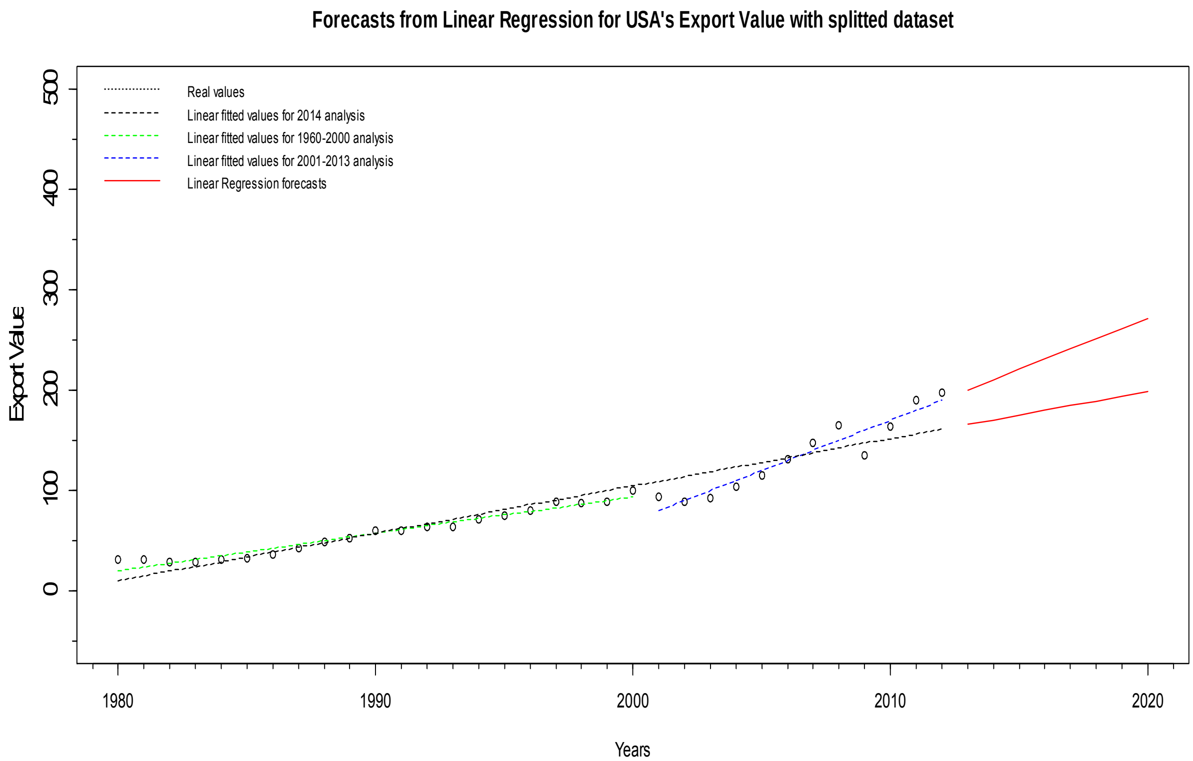
Export Value – USA



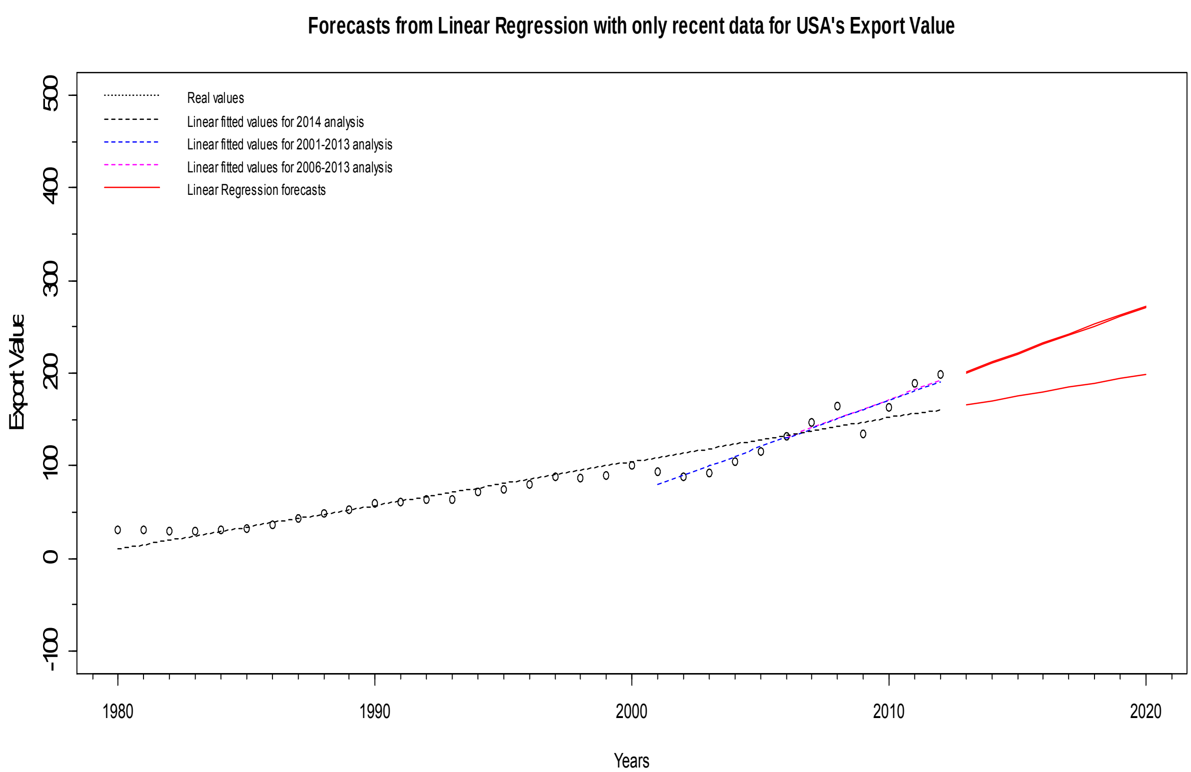
[Figure C.1.37] – Analysis for USA, Export Value and whole dataset



[Figure C.1.38] – Analysis for USA, Export Value and the dataset up to 2008

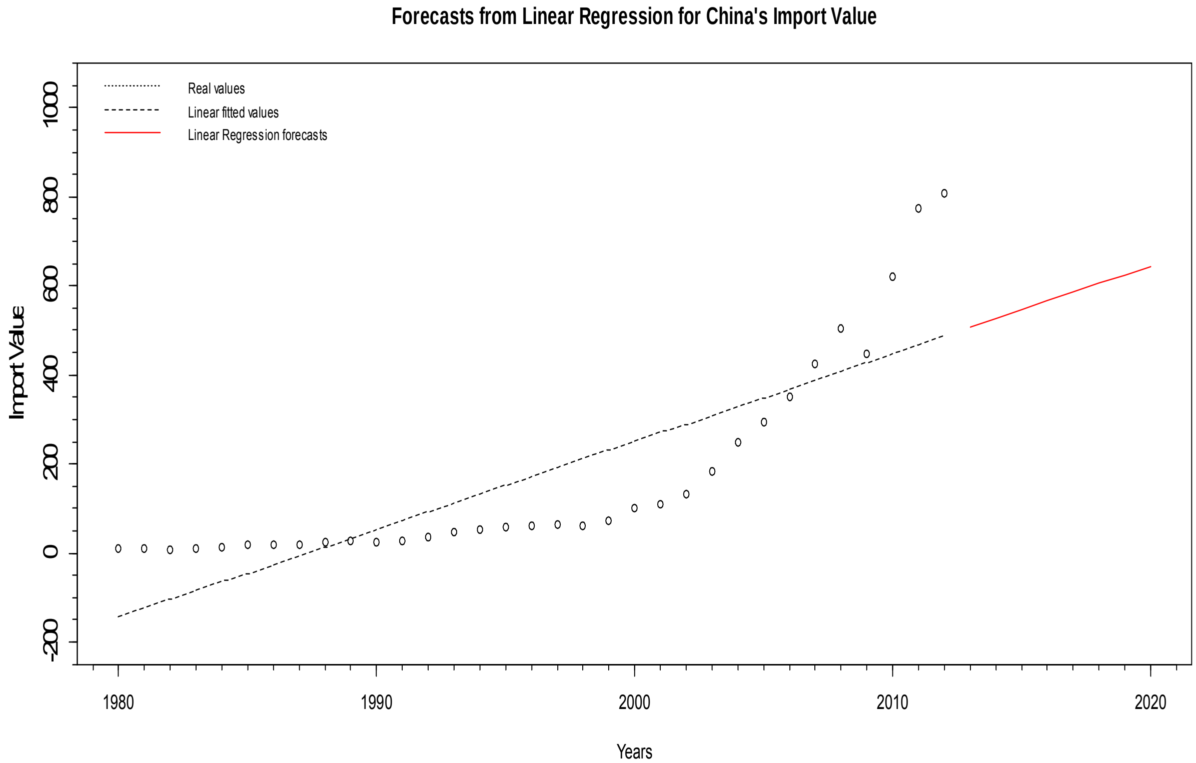


[Figure C.1.39] – Analyses for USA, Export Value and the subsets up to 2000 and 2001-2013

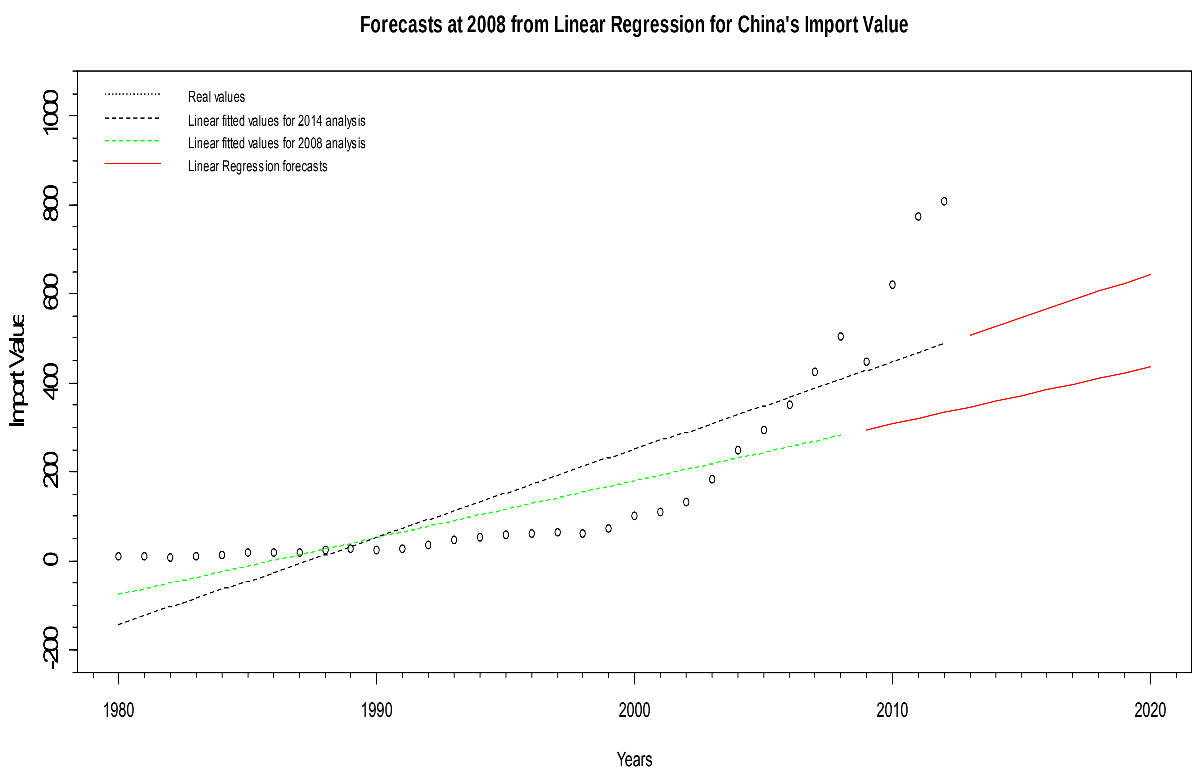


[Figure C.1.40] – Analyses for USA, Export Value and the subset from 2006-2013

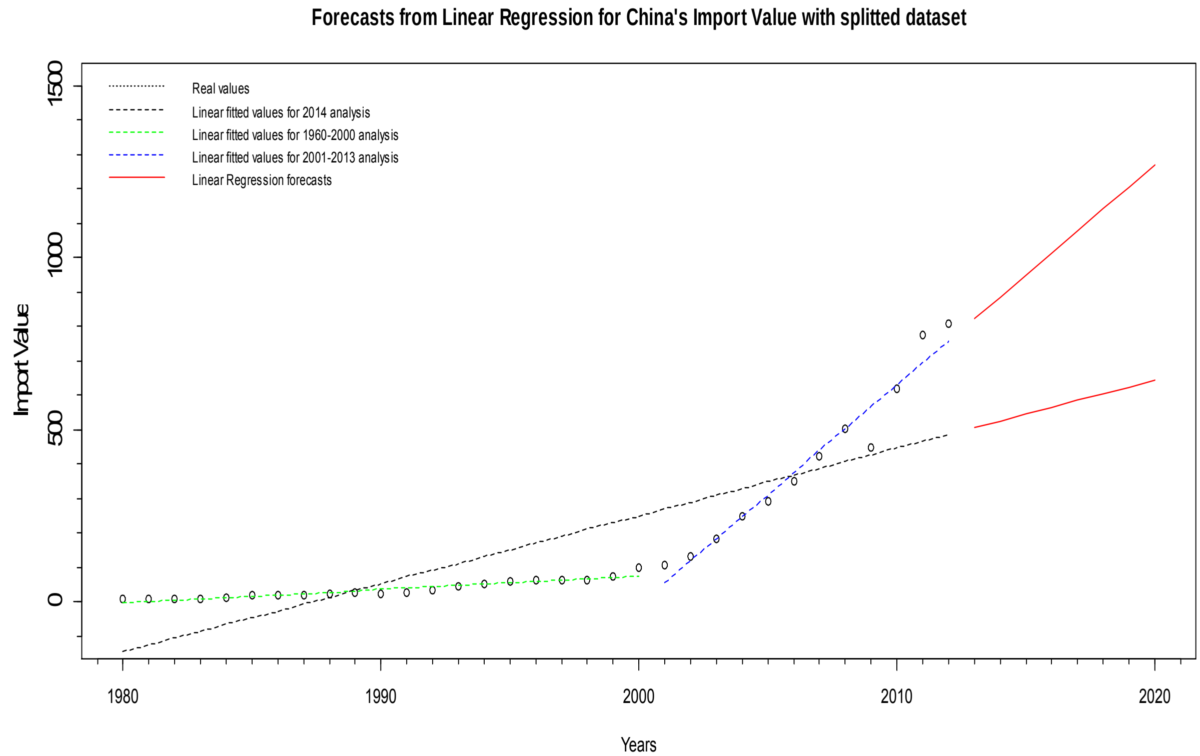
Import Value – China



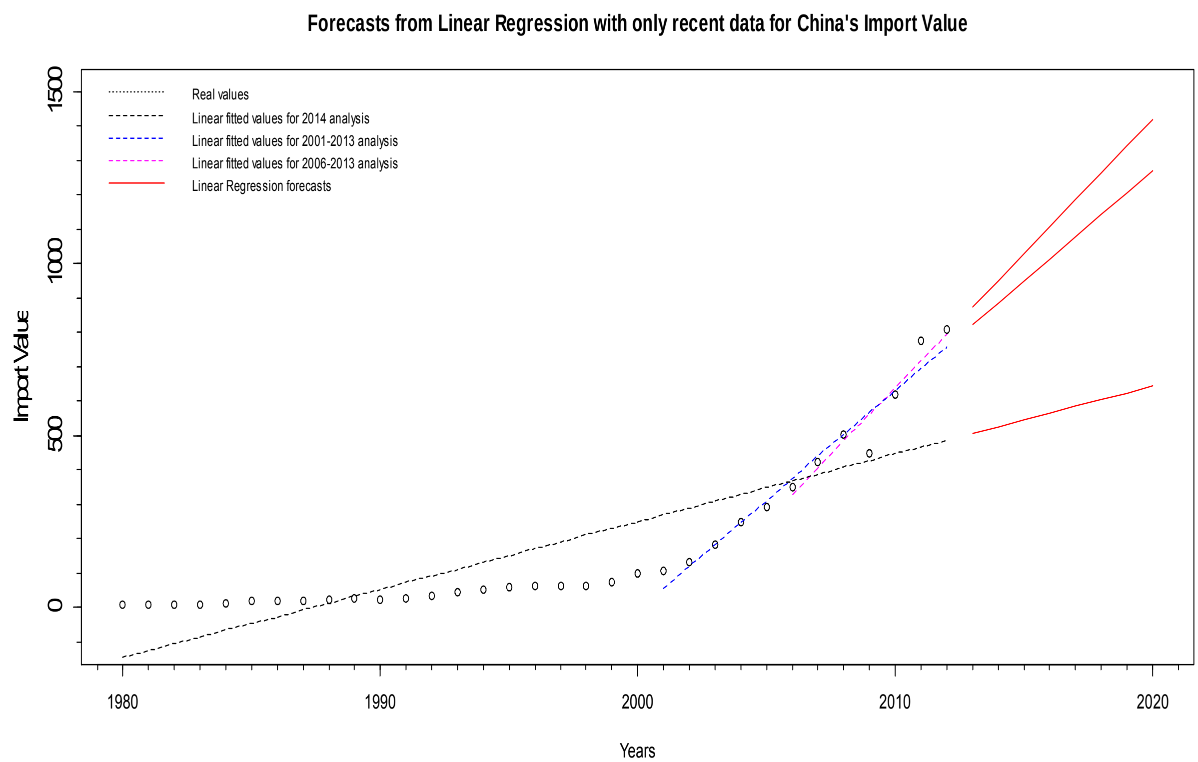
[Figure C.1.41] – Analysis for China, Import Value and whole dataset



[Figure C.1.42] – Analysis for China, Import Value and the dataset up to 2008

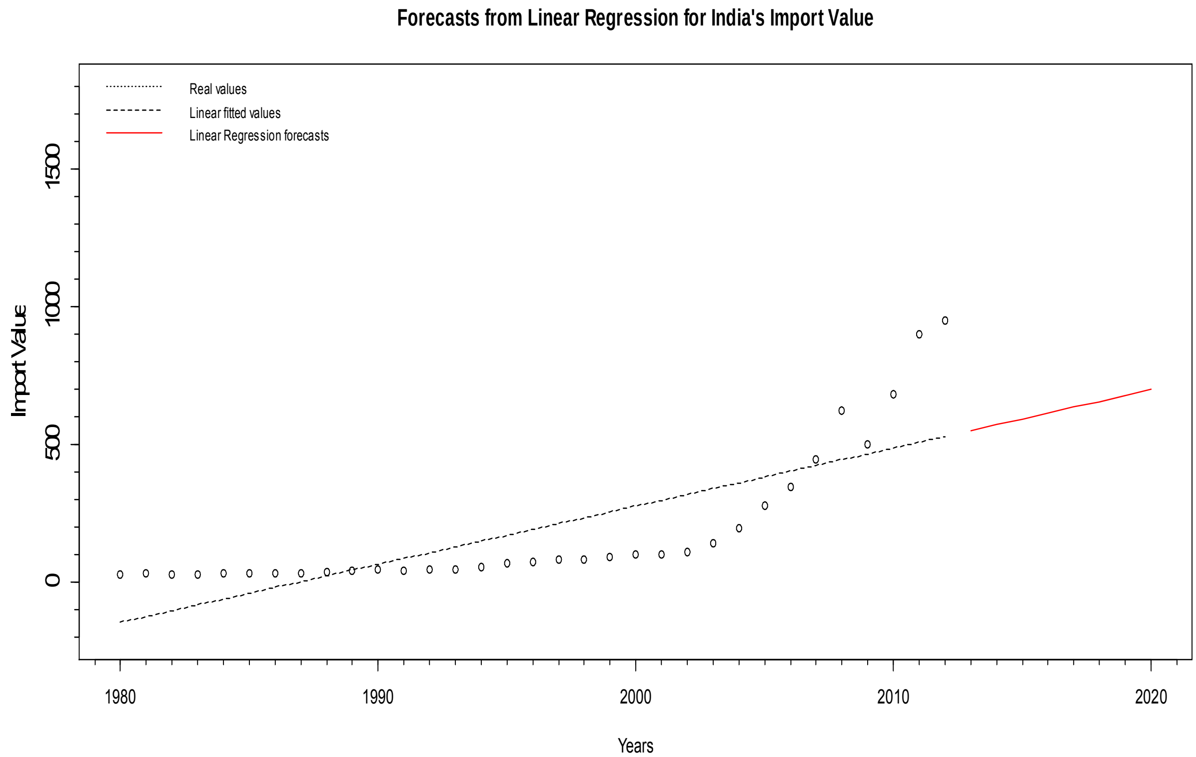


[Figure C.1.43] – Analyses for China, Import Value and the subsets up to 2000 and 2001-2013

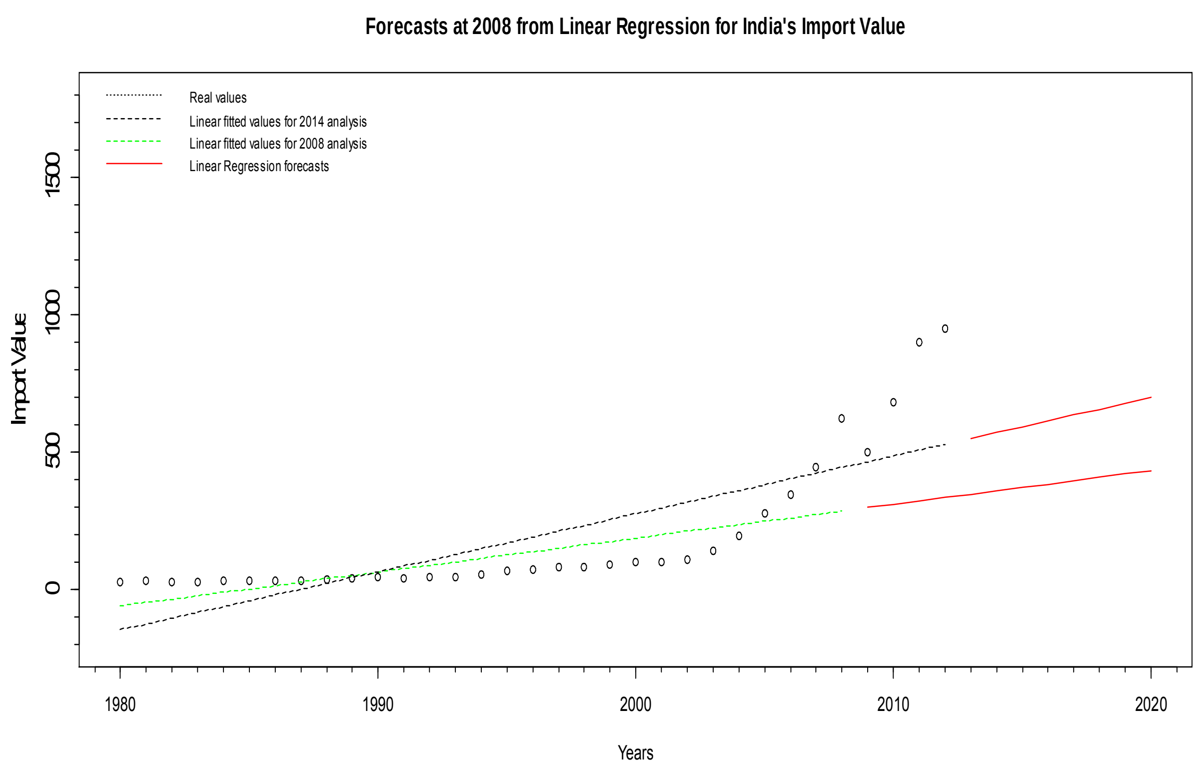


[Figure C.1.44] – Analyses for China, Import Value and the subset from 2006-2013

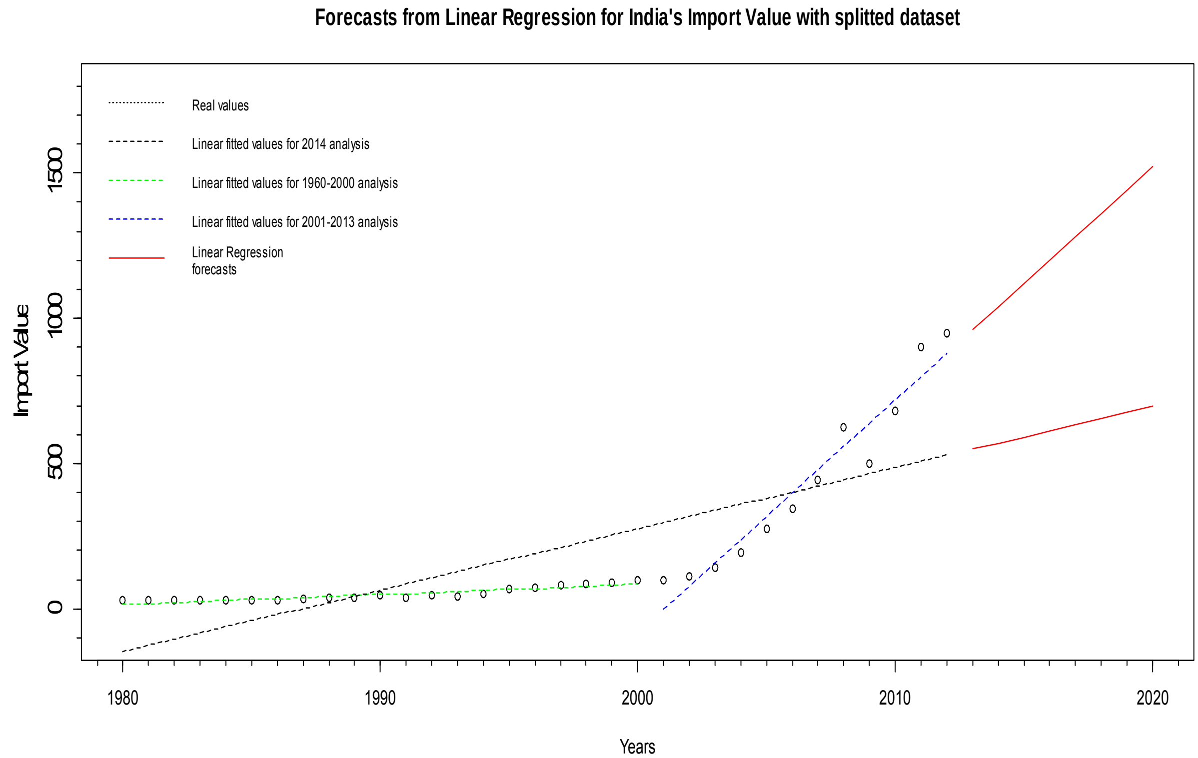
Import Value – India



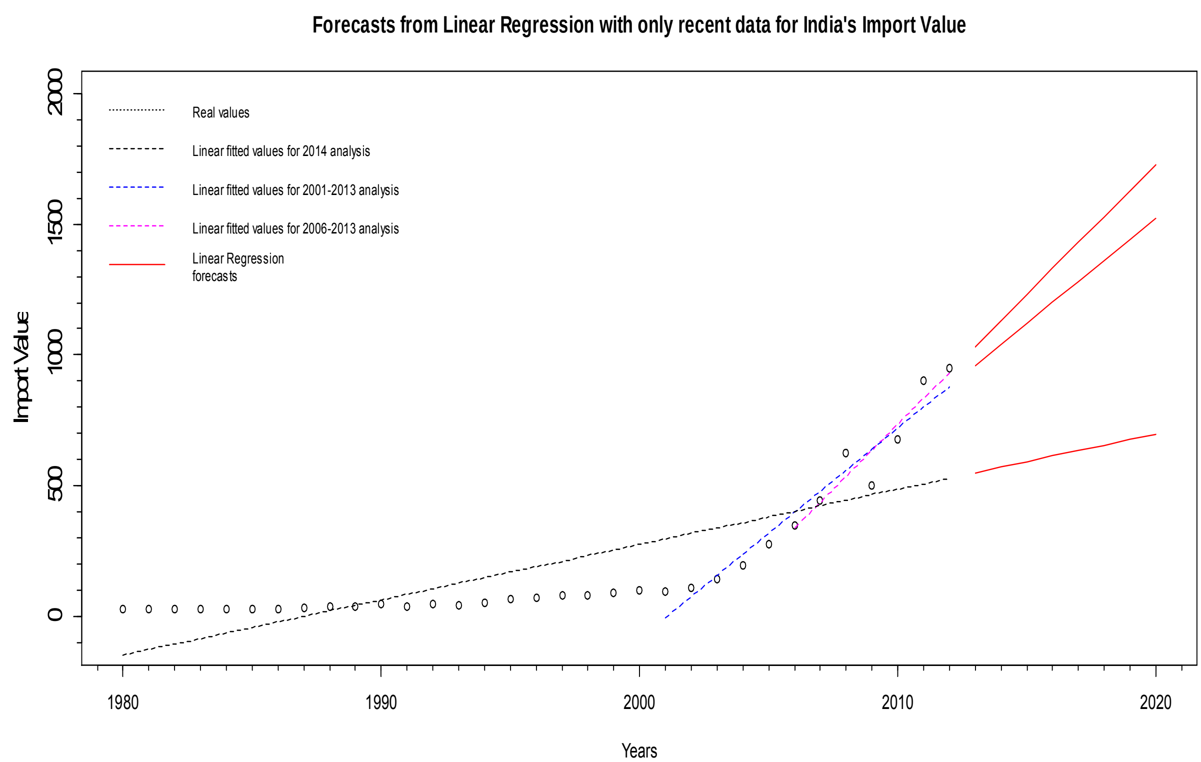
[Figure C.1.45] – Analysis for India, Import Value and whole dataset



[Figure C.1.46] – Analysis for India, Import Value and the dataset up to 2008

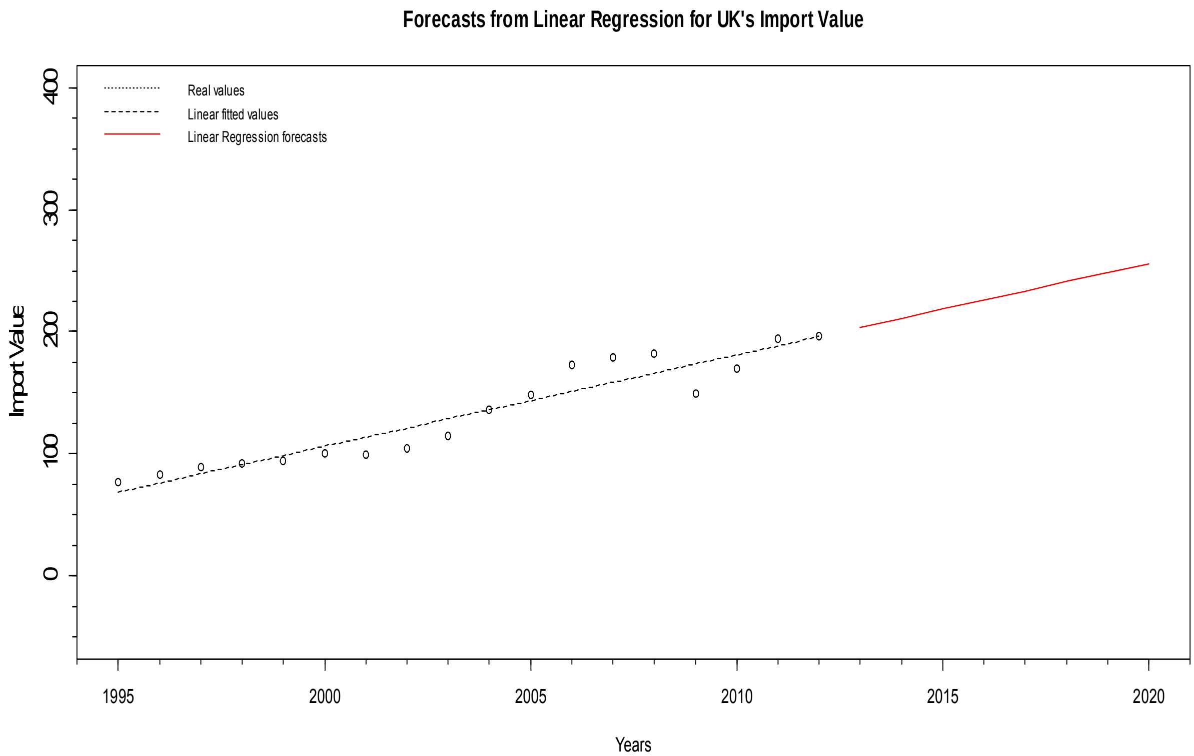


[Figure C.1.47] – Analyses for India, Import Value and the subsets up to 2000 and 2001-2013

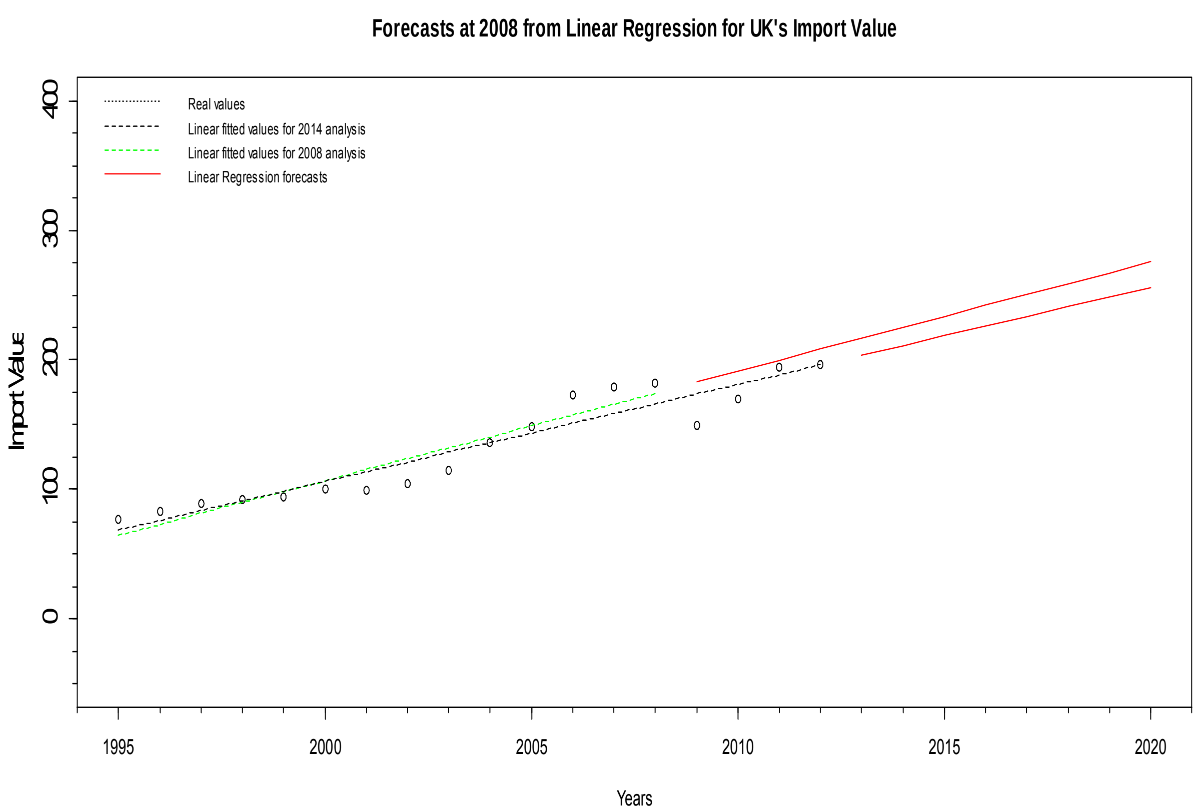


[Figure C.1.48] – Analyses for India, Import Value and the subset from 2006-2013

Import Value – UK

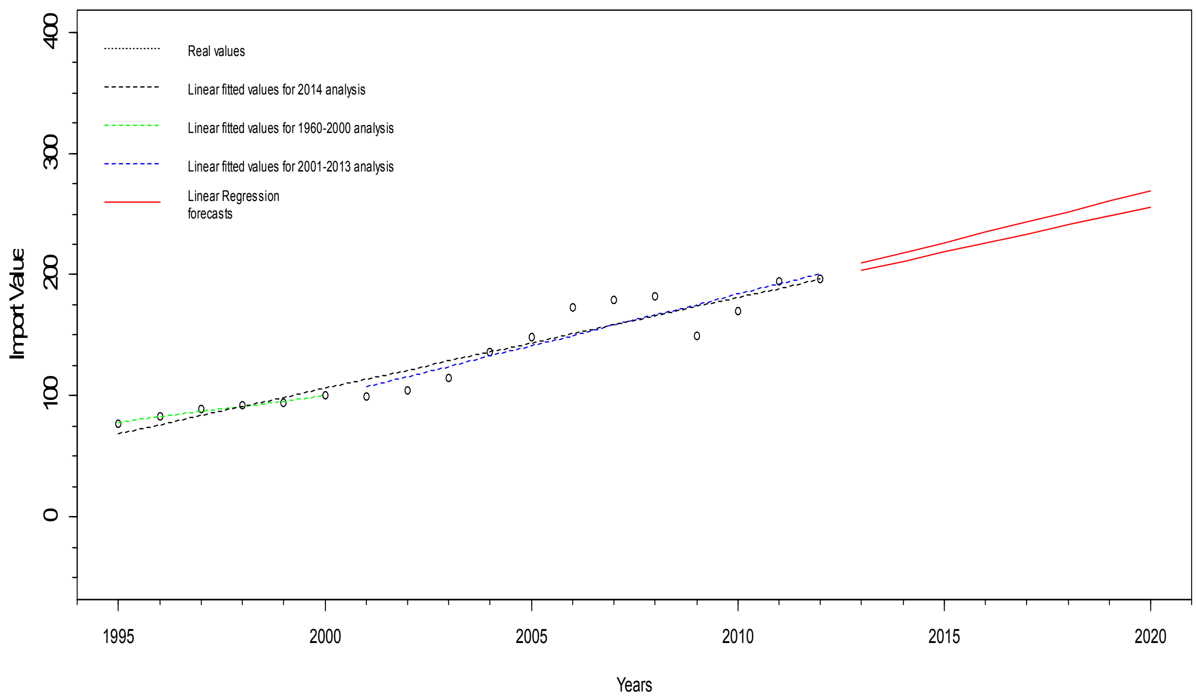


[Figure C.1.49] – Analysis for UK, Import Value and whole dataset



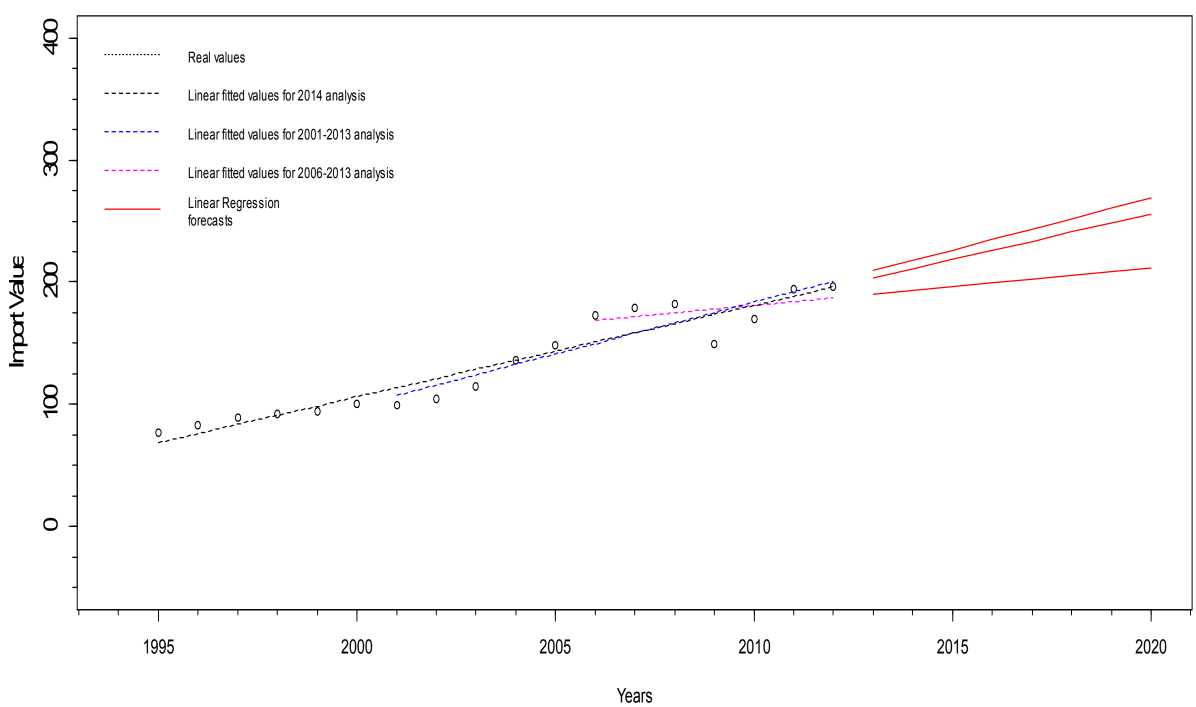
[Figure C.1.50] – Analysis for UK, Import Value and the dataset up to 2008

Forecasts from Linear Regression for UK's Import Value with splitted dataset



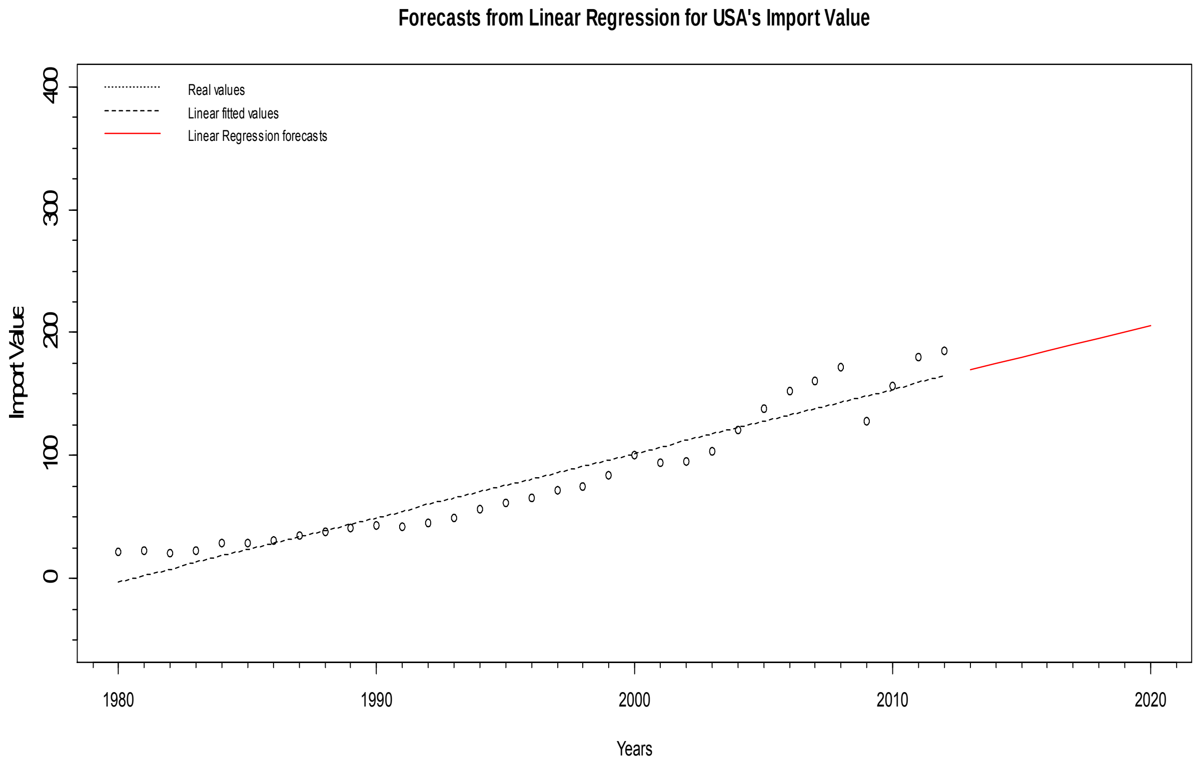
[Figure C.1.51] – Analyses for UK, Import Value and the subsets up to 2000 and 2001-2013

Forecasts from Linear Regression with only recent data for UK's Import Value

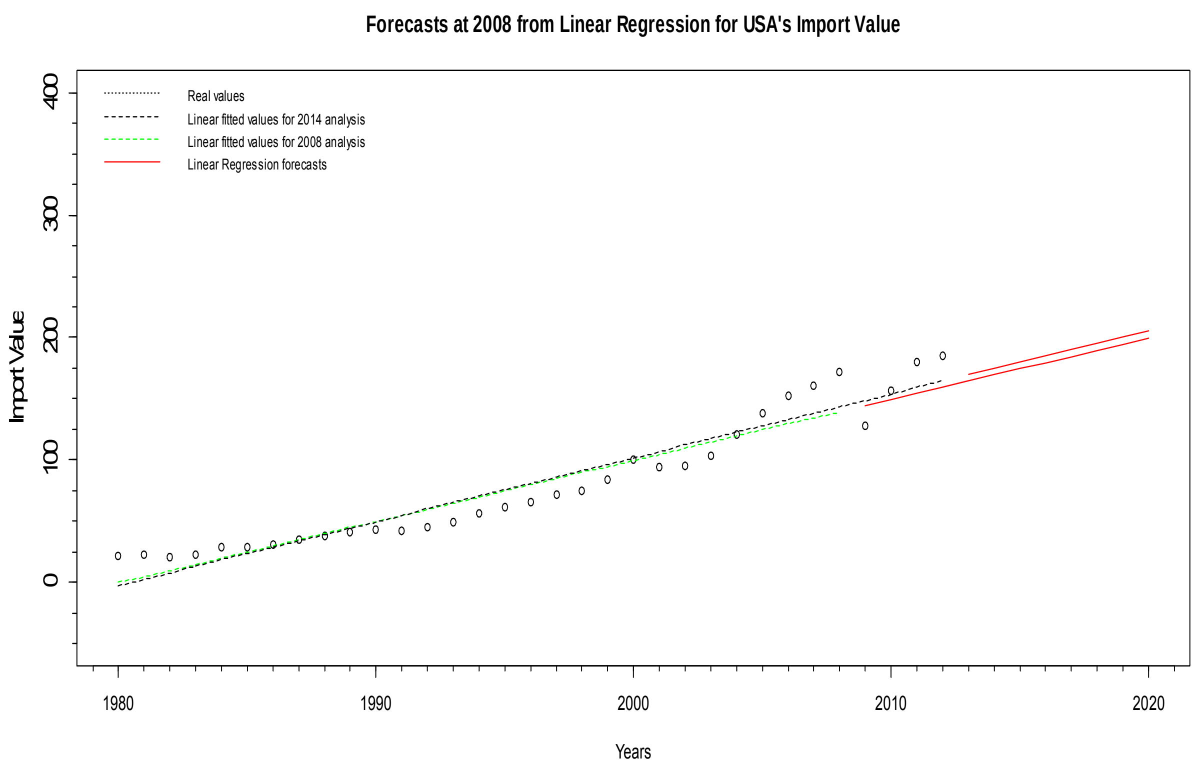


[Figure C.1.52] – Analyses for UK, Import Value and the subset from 2006-2013

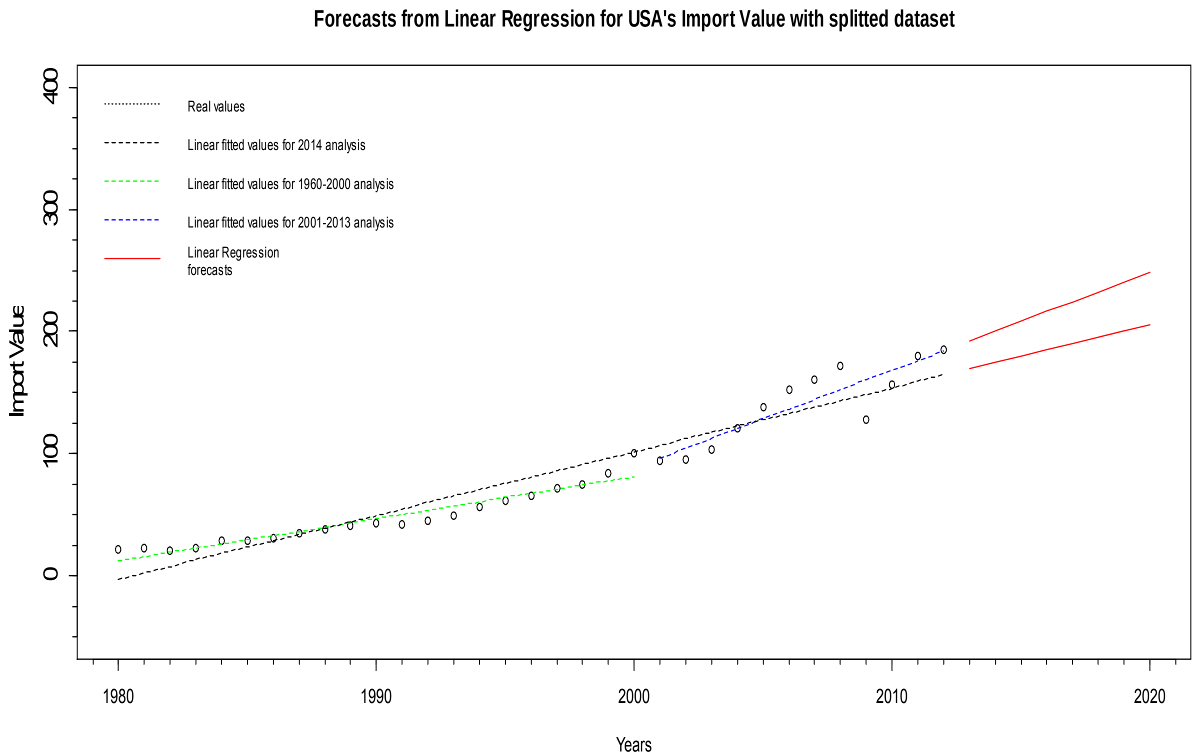
Import Value – USA



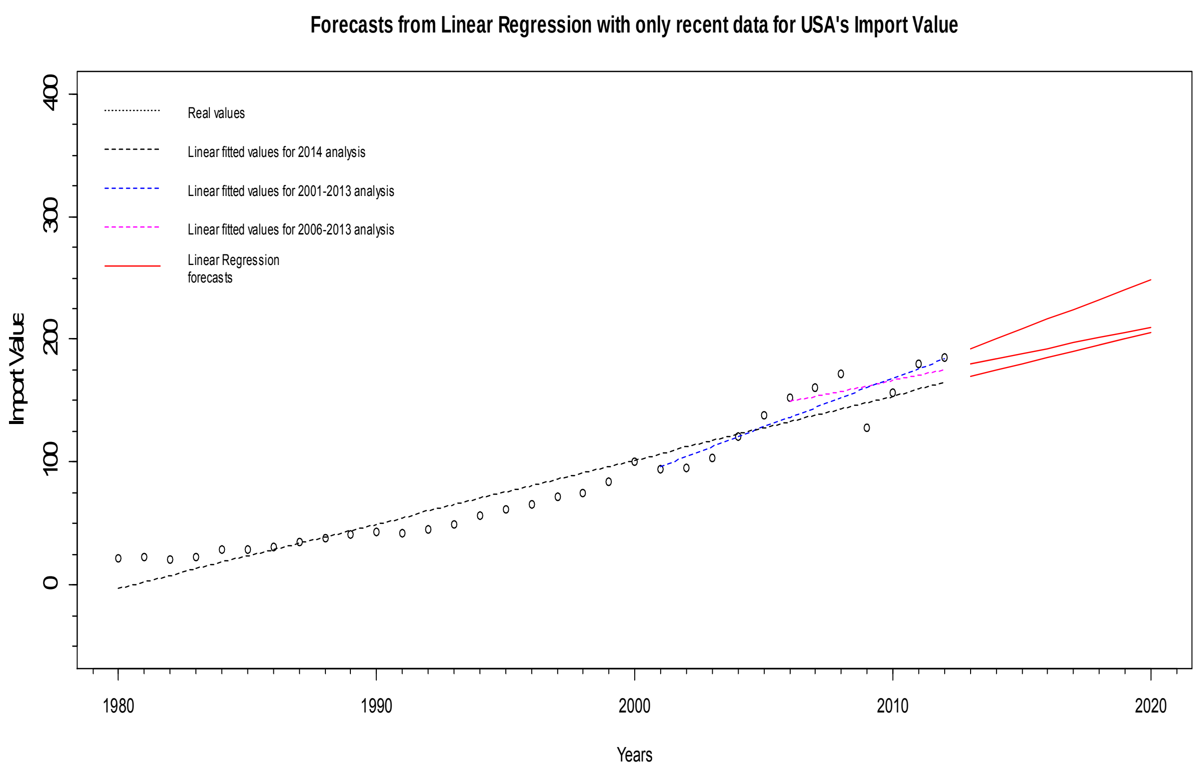
[Figure C.1.53] – Analysis for USA, Import Value and whole dataset



[Figure C.1.54] – Analysis for USA, Import Value and the dataset up to 2008



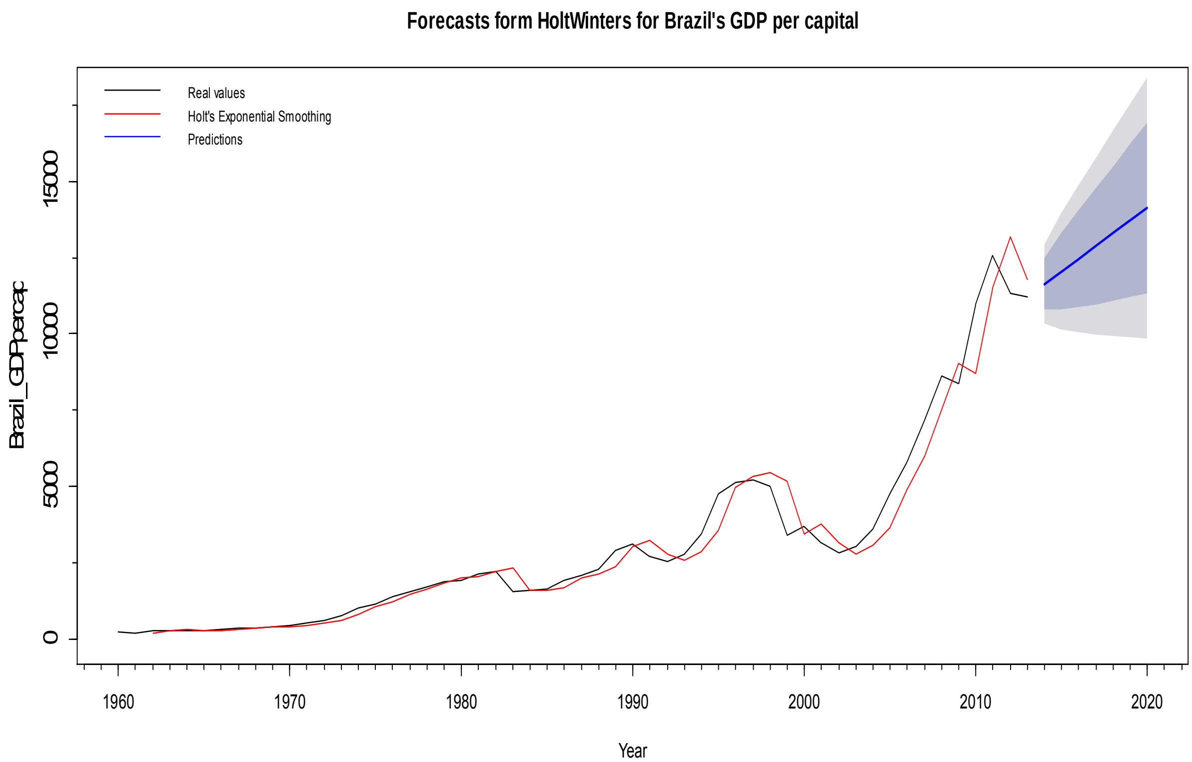
[Figure C.1.55] – Analyses for USA, Import Value and the subsets up to 2000 and 2001-2013



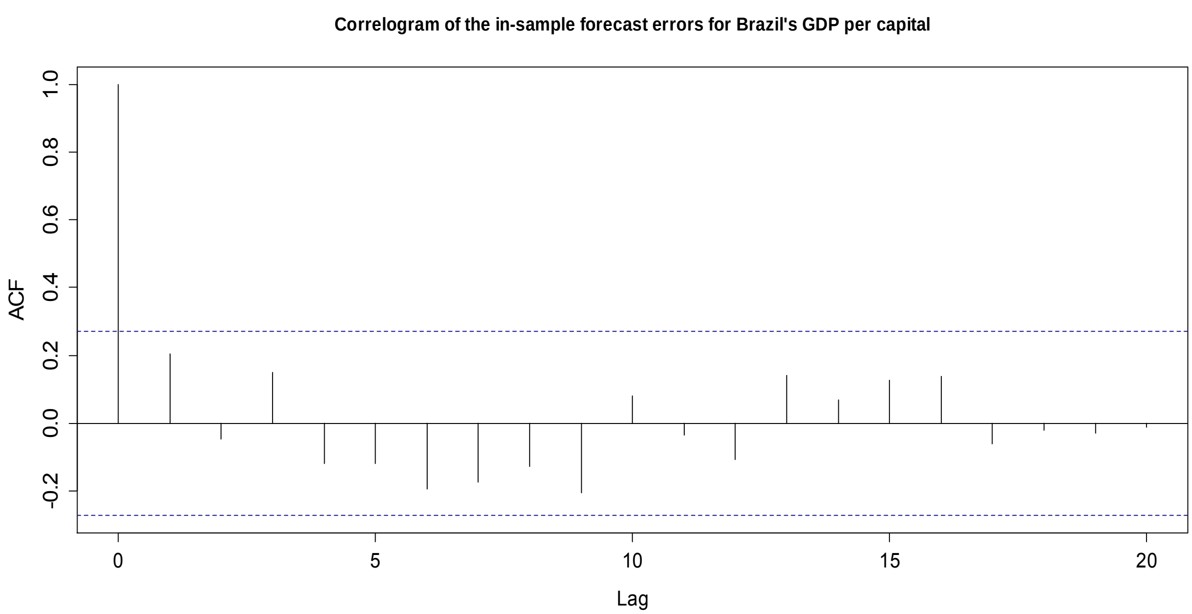
[Figure C.1.56] – Analyses for USA, Import Value and the subset from 2006-2013

C.2. Holt's Exponential Smoothing

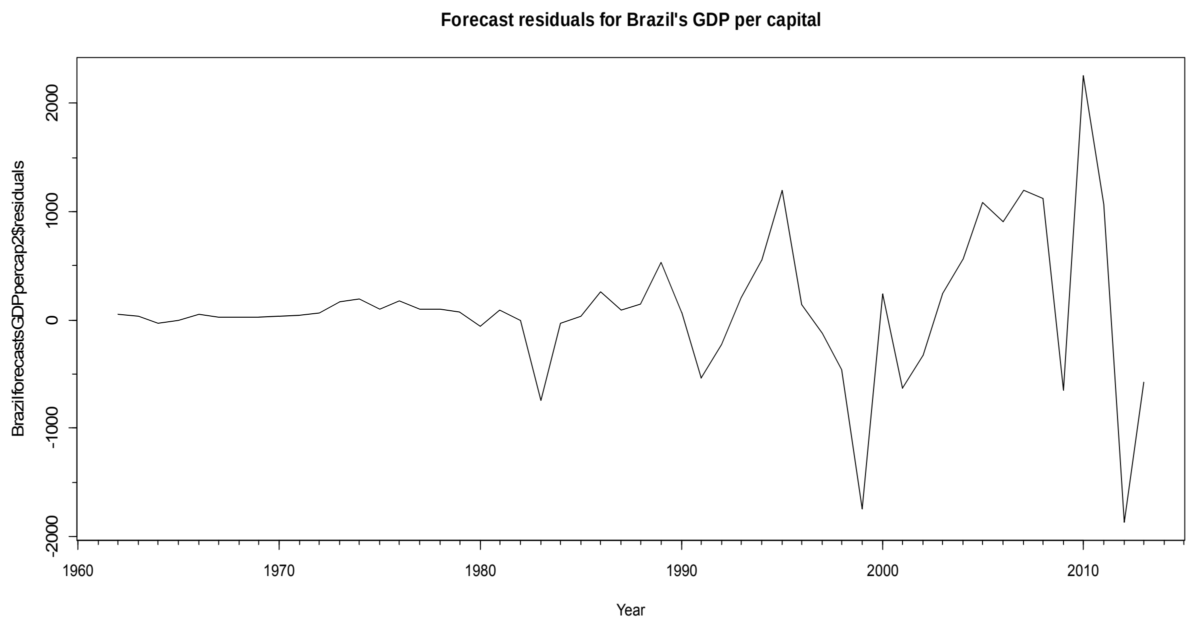
GDP per capita – Brazil



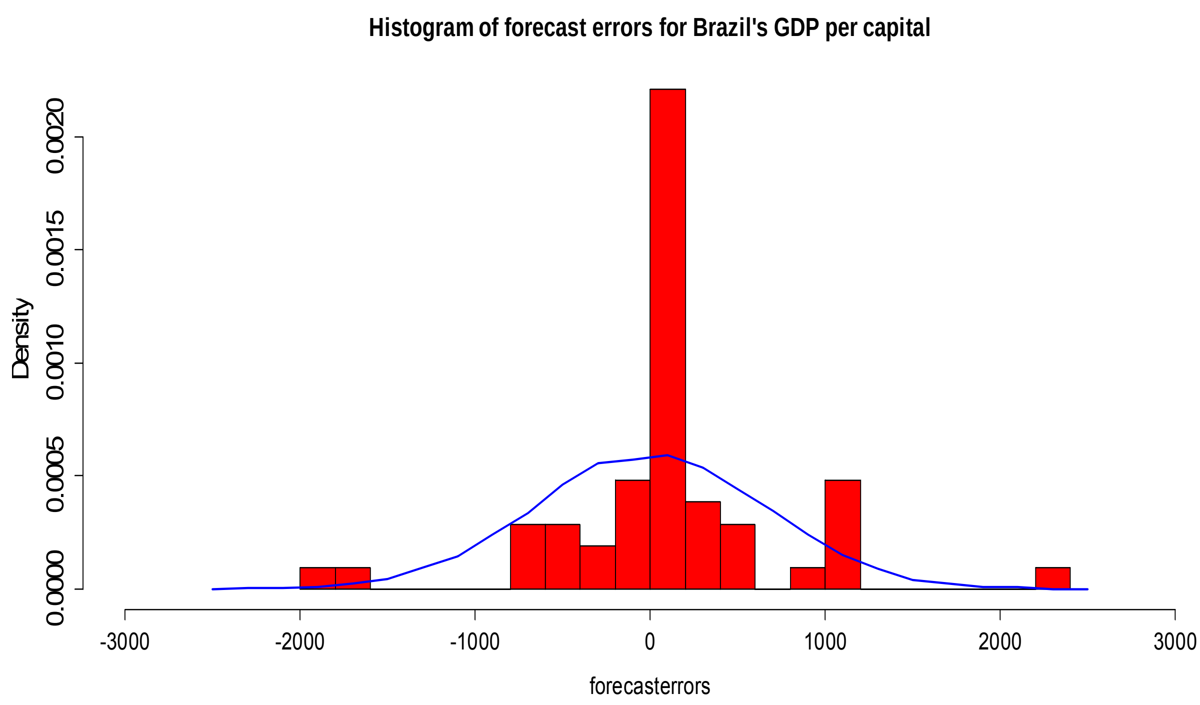
[Figure C.2.1] – Analysis for Brazil, GDP per capita and whole dataset



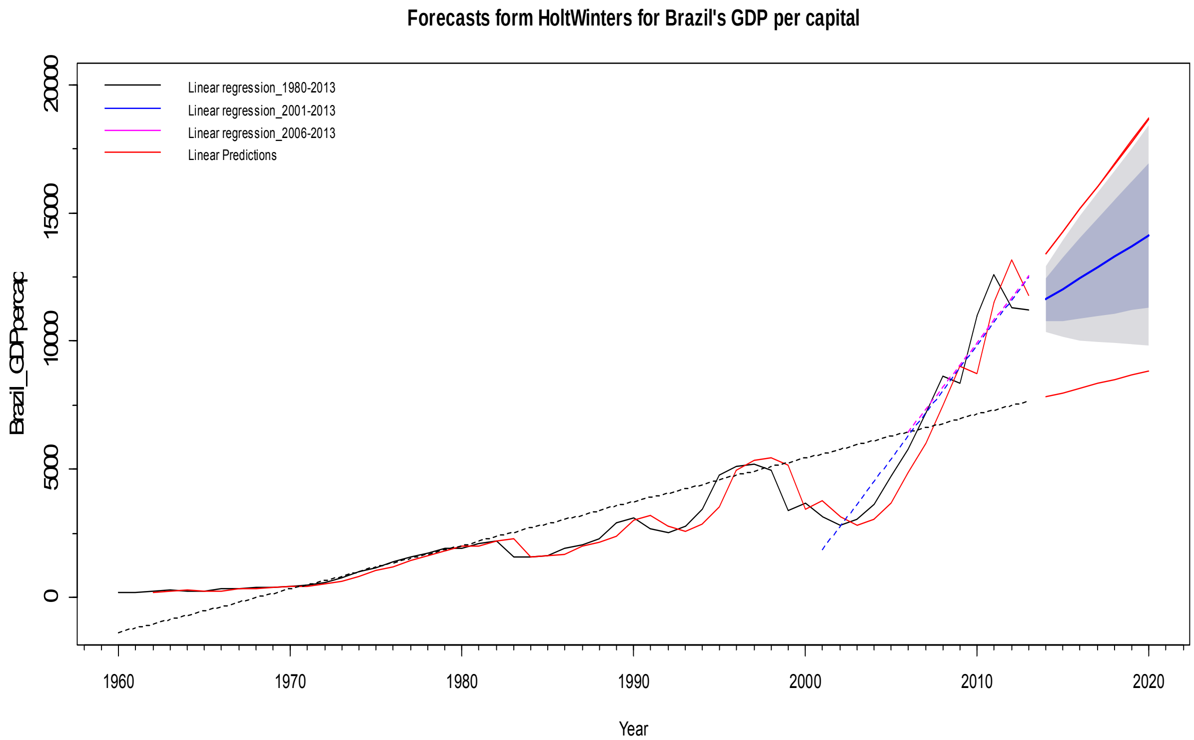
[Figure C.2.2] – Correlogram of in-sample errors



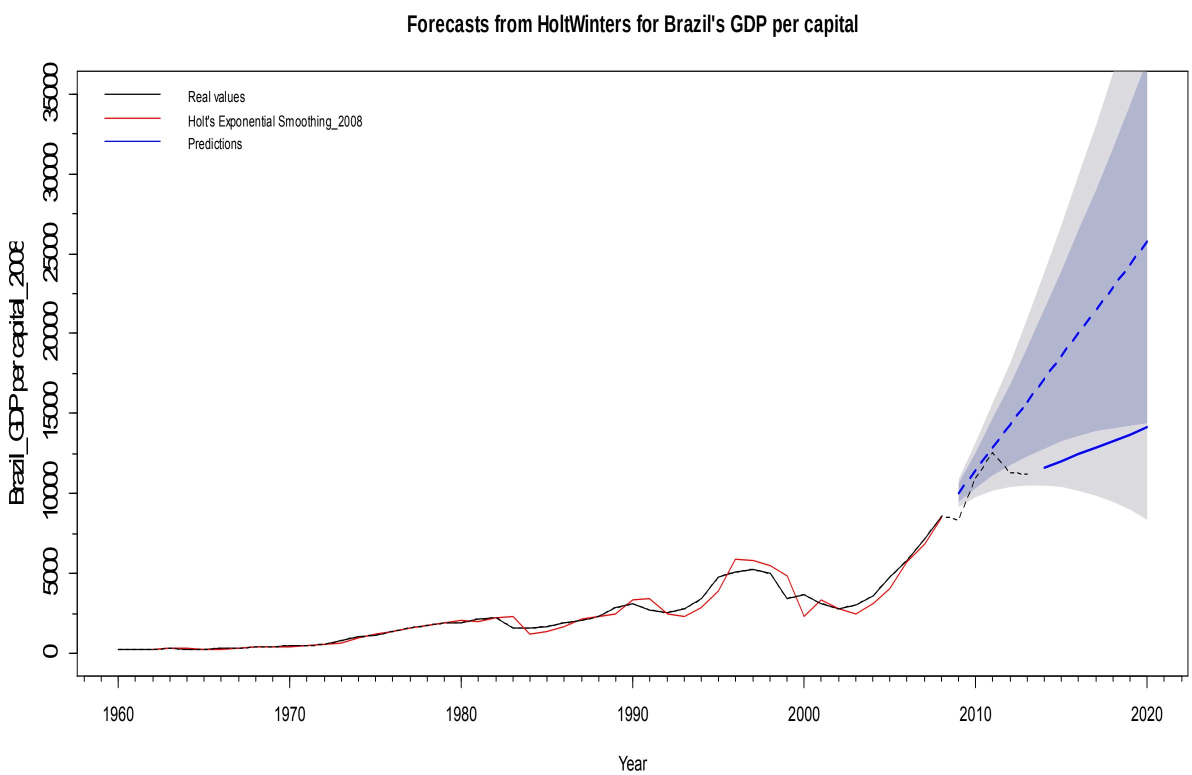
[Figure C.2.3] – Forecast residuals



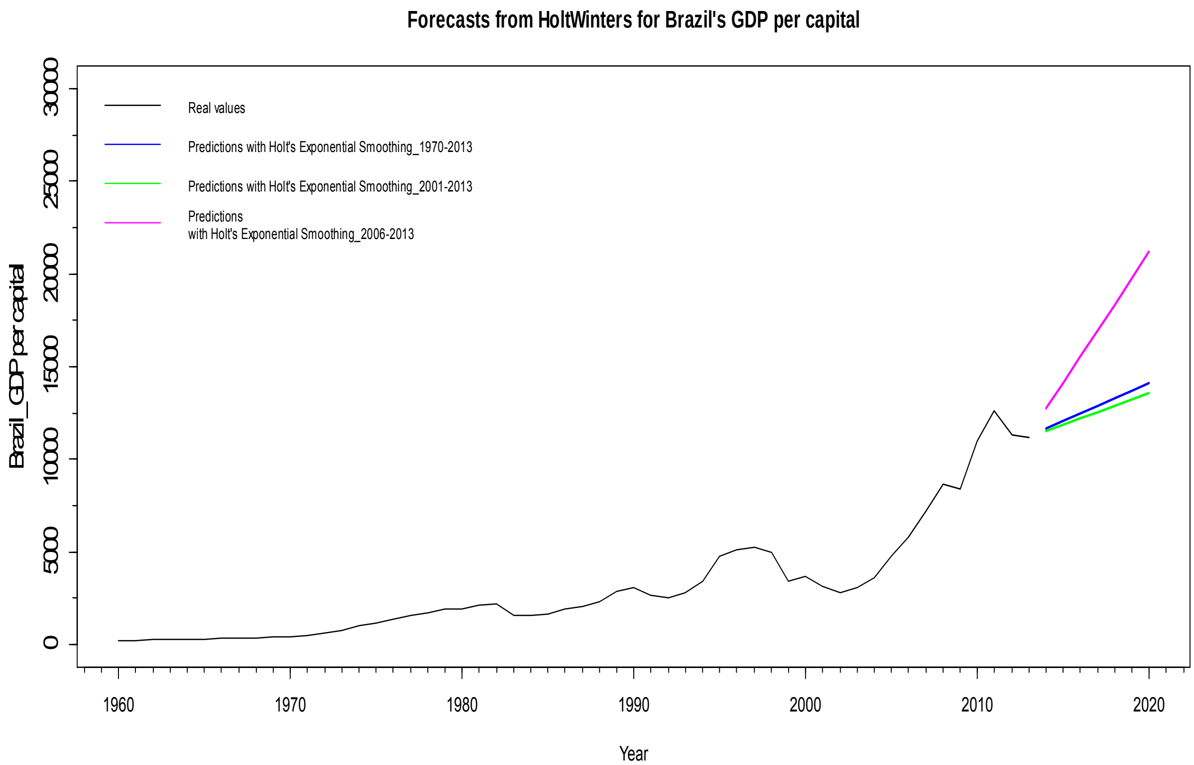
[Figure C.2.4] – Histogram and distribution of forecast residuals



[Figure C.2.5] – Comparison of Linear Regression and Holt's Exponential smoothing

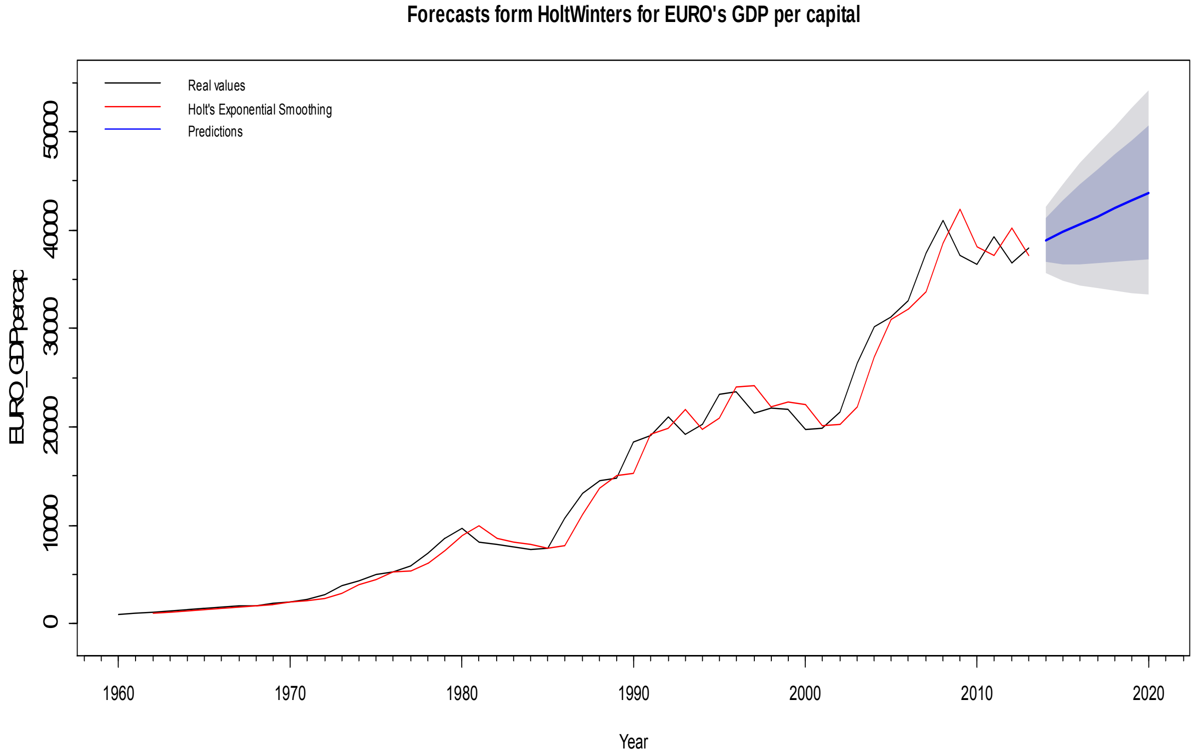


[Figure C.2.6] – Analysis for Brazil, GDP per capita and the dataset up to 2008

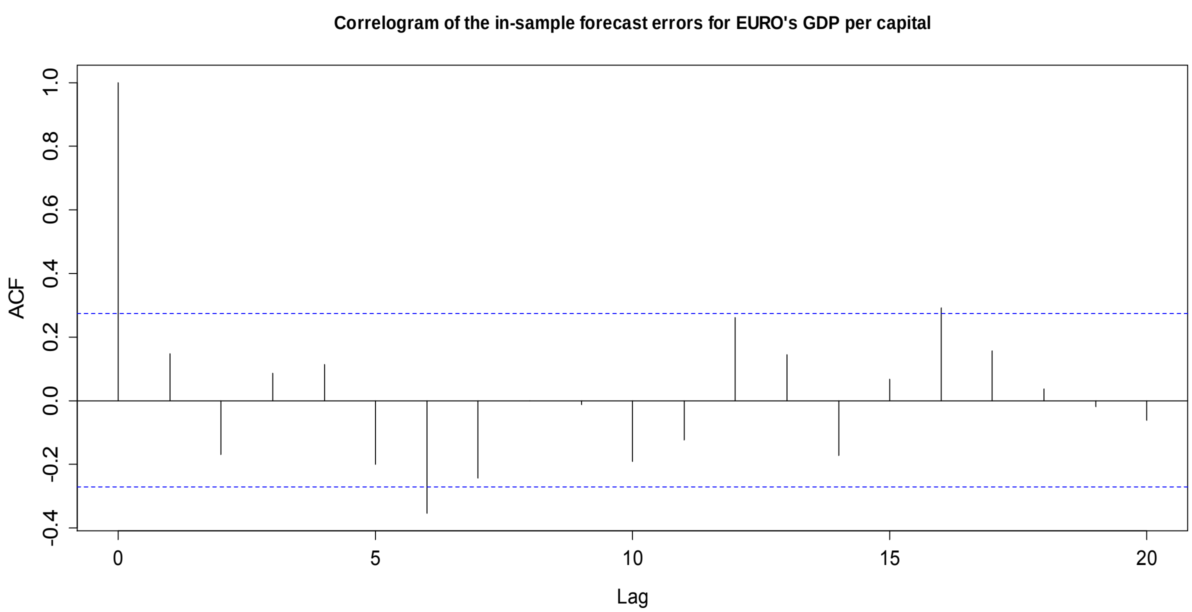


[Figure C.2.7] – Analyses for Brazil, GDP per capita and the subsets 2001-2013 and 2006-2013

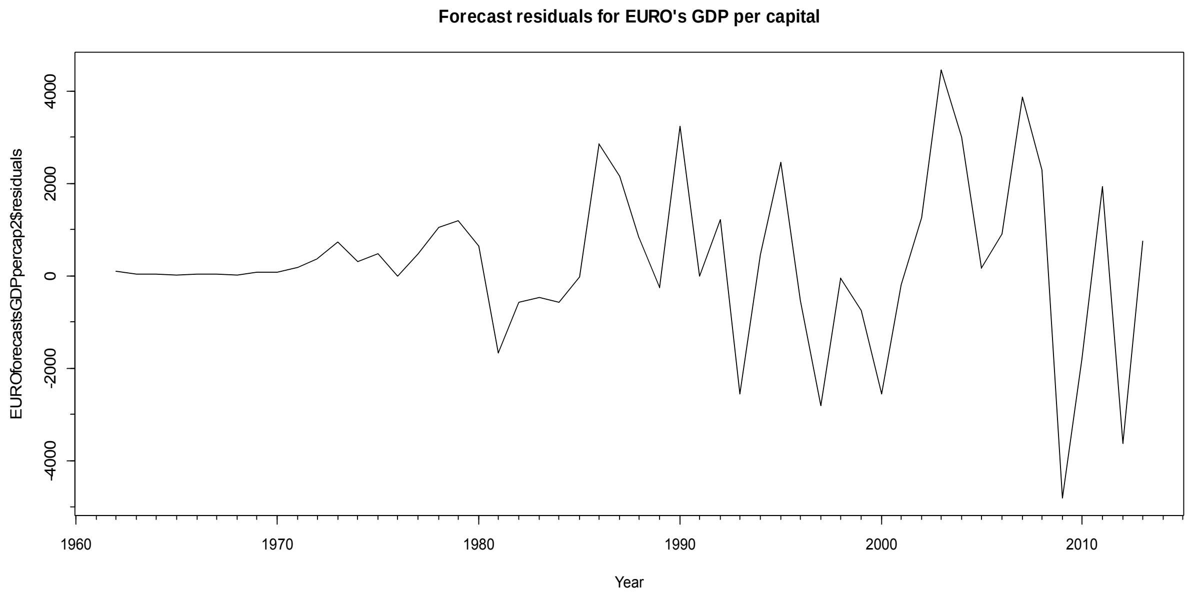
GDP per capita – EURO zone



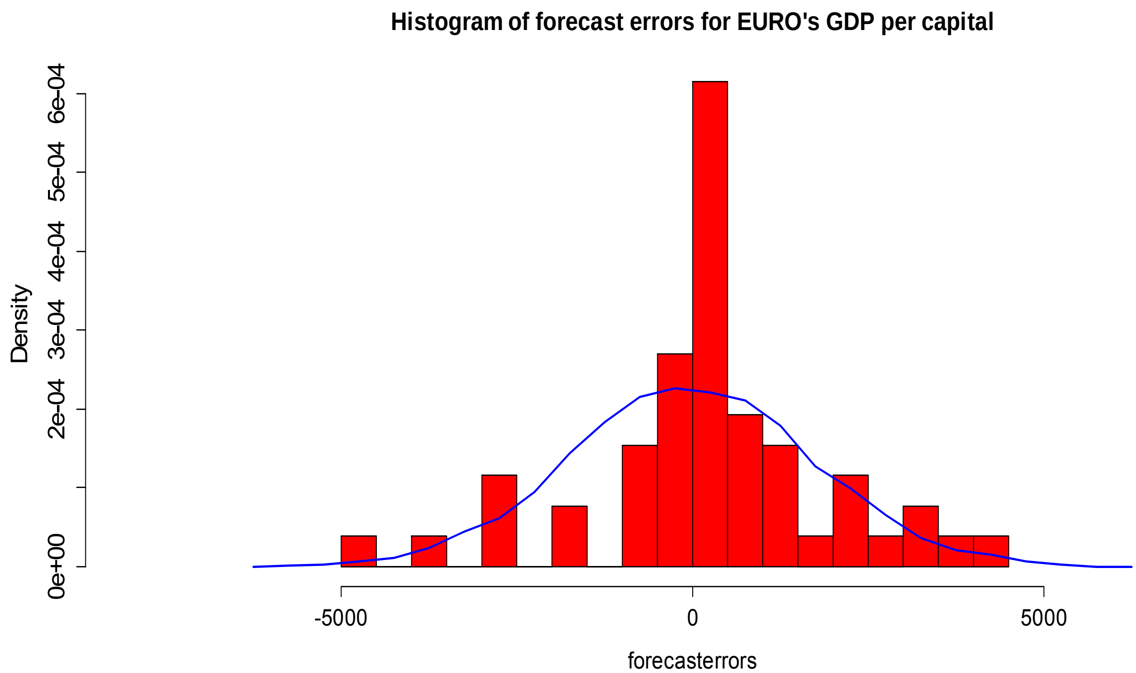
[Figure C.2.8] – Analysis for EURO zone, GDP per capita and whole dataset



[Figure C.2.9] – Correlogram of in-sample errors

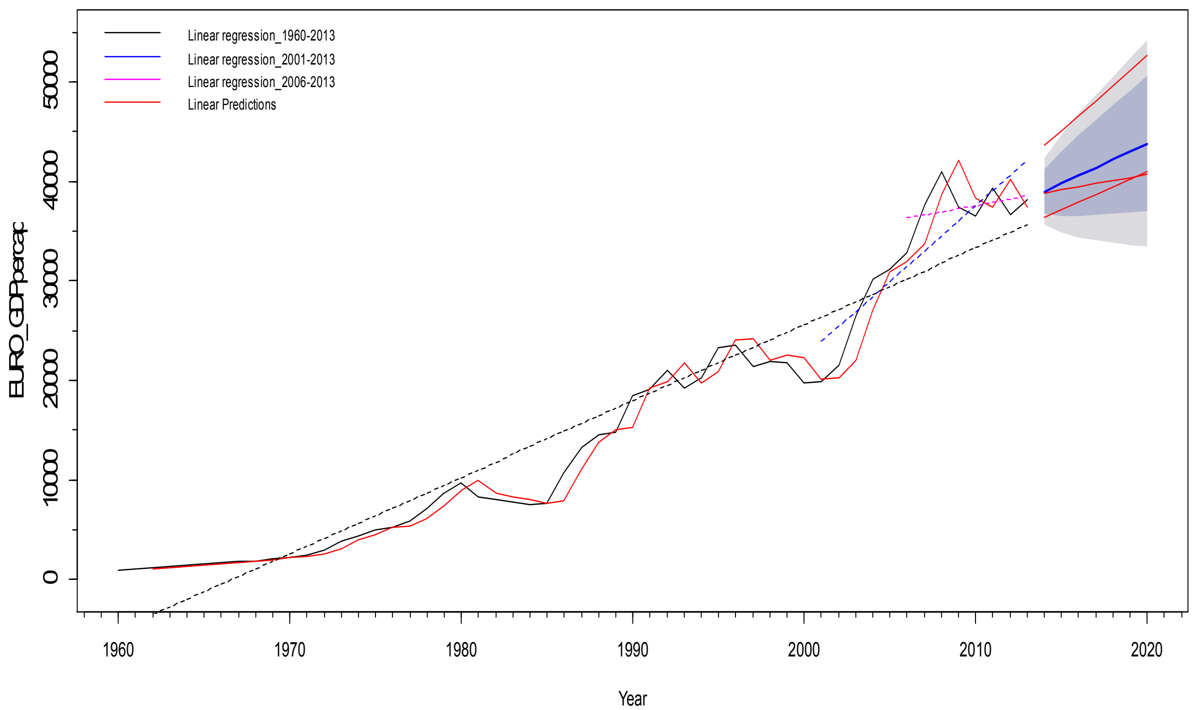


[Figure C.2.10] – Forecast residuals



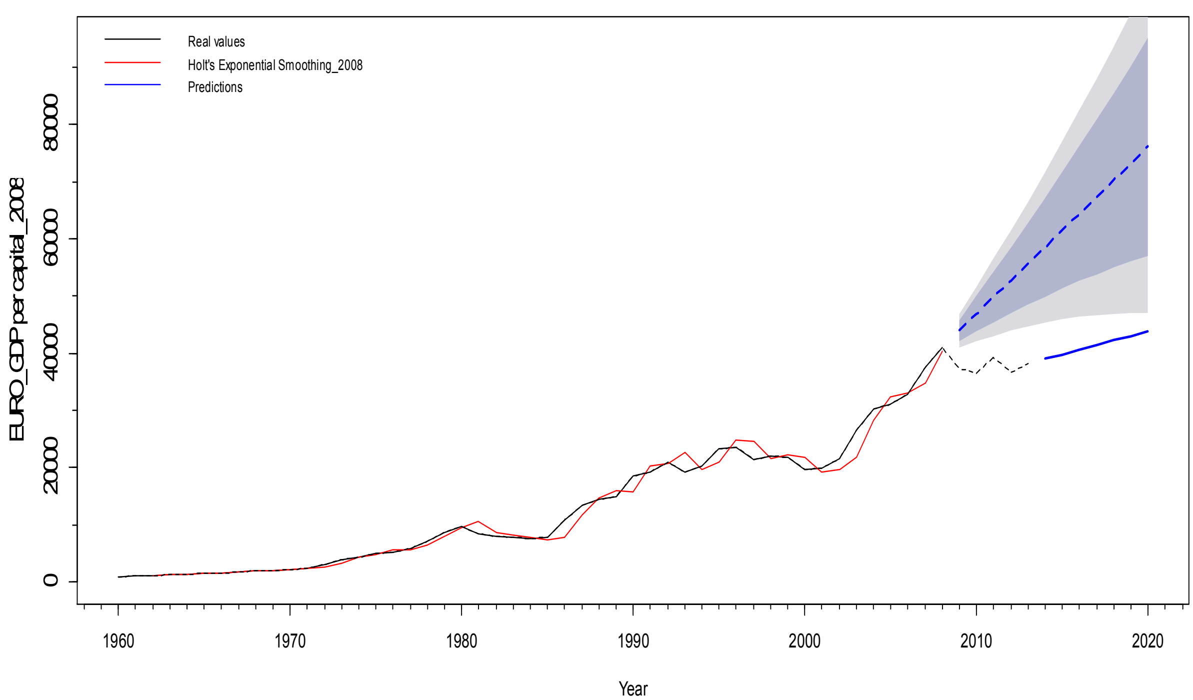
[Figure C.2.11] – Histogram and distribution of forecast residuals

Forecasts form HoltWinters for EURO's GDP per capital

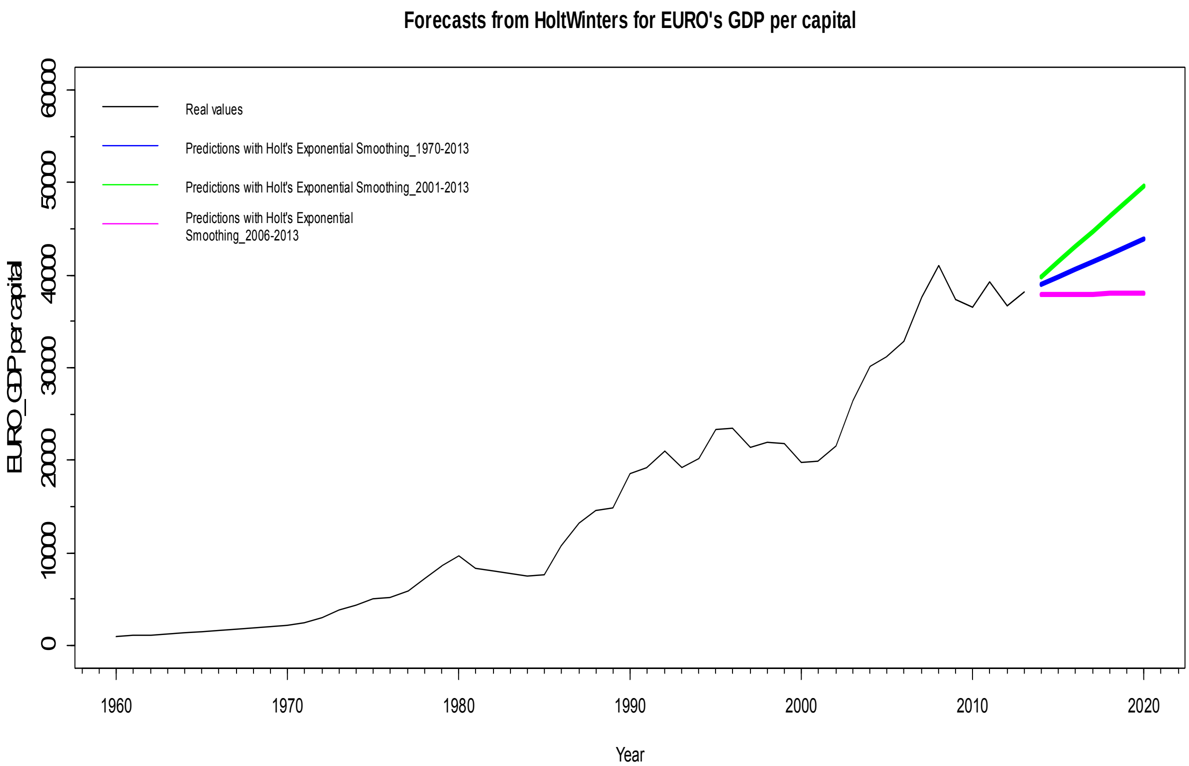


[Figure C.2.12] – Comparison of Linear Regression and Holt's Exponential smoothing

Forecasts from HoltWinters for EURO's GDP per capital

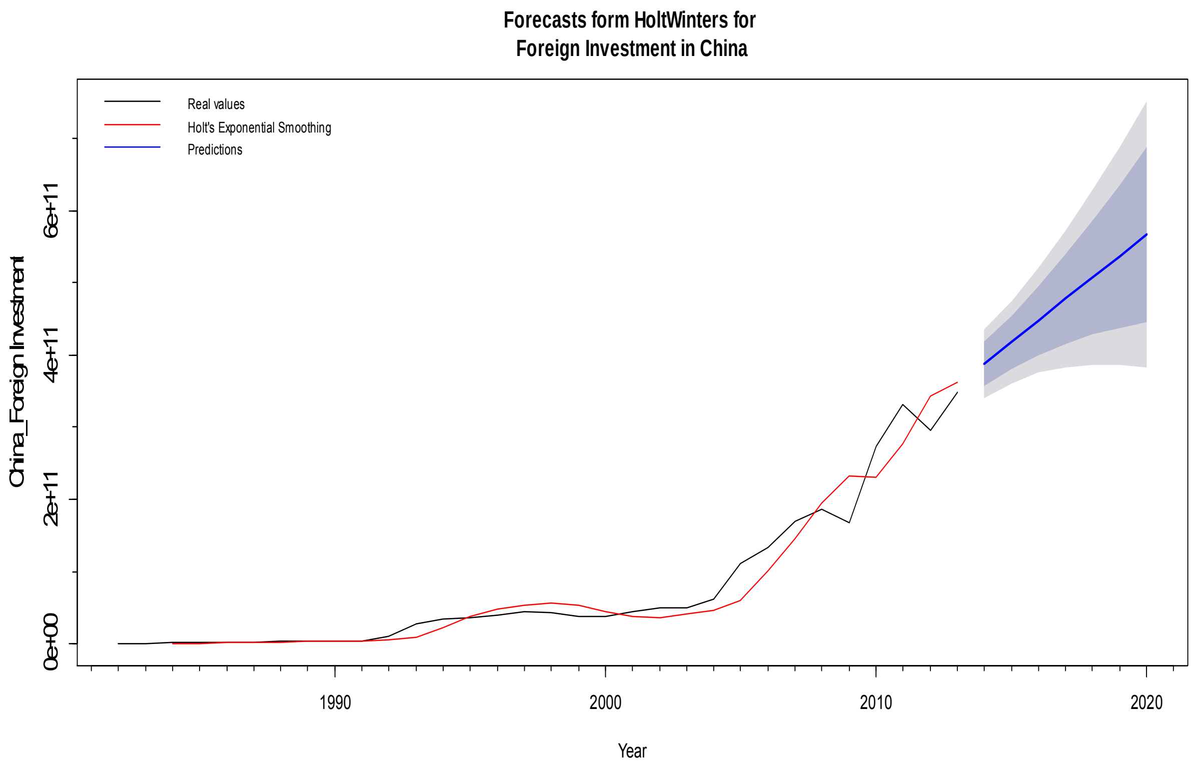


[Figure C.2.13] – Analysis for EURO zone, GDP per capita and the dataset up to 2008

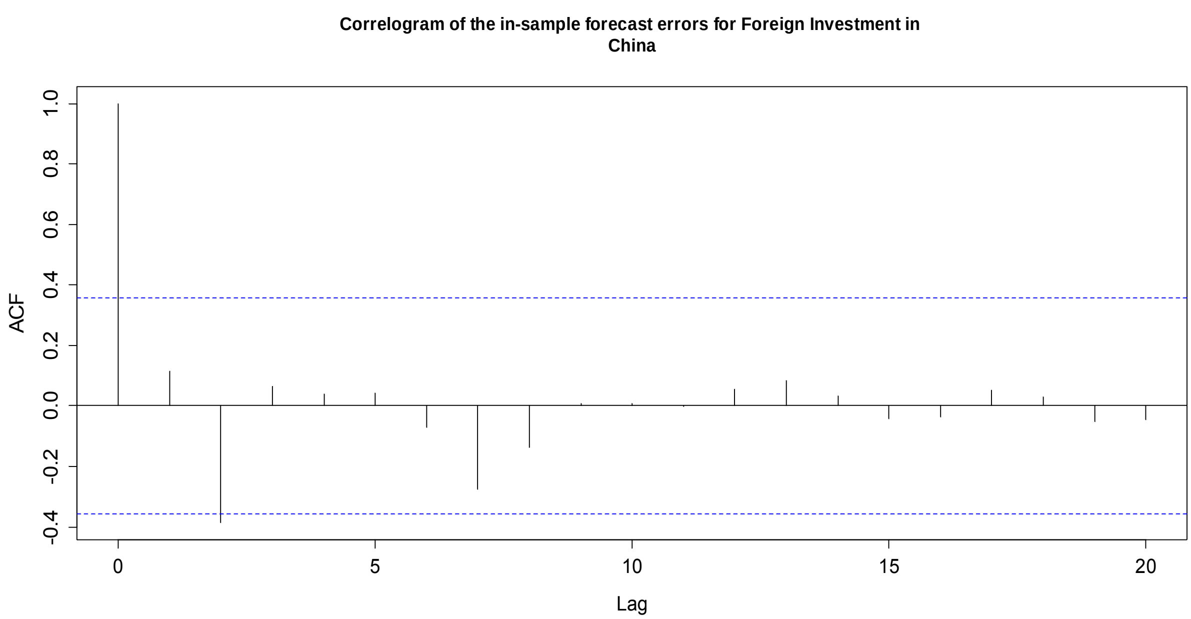


[Figure C.2.14] – Analyses for EURO, GDP per capita and the subsets 2001-2013 and 2006-2013

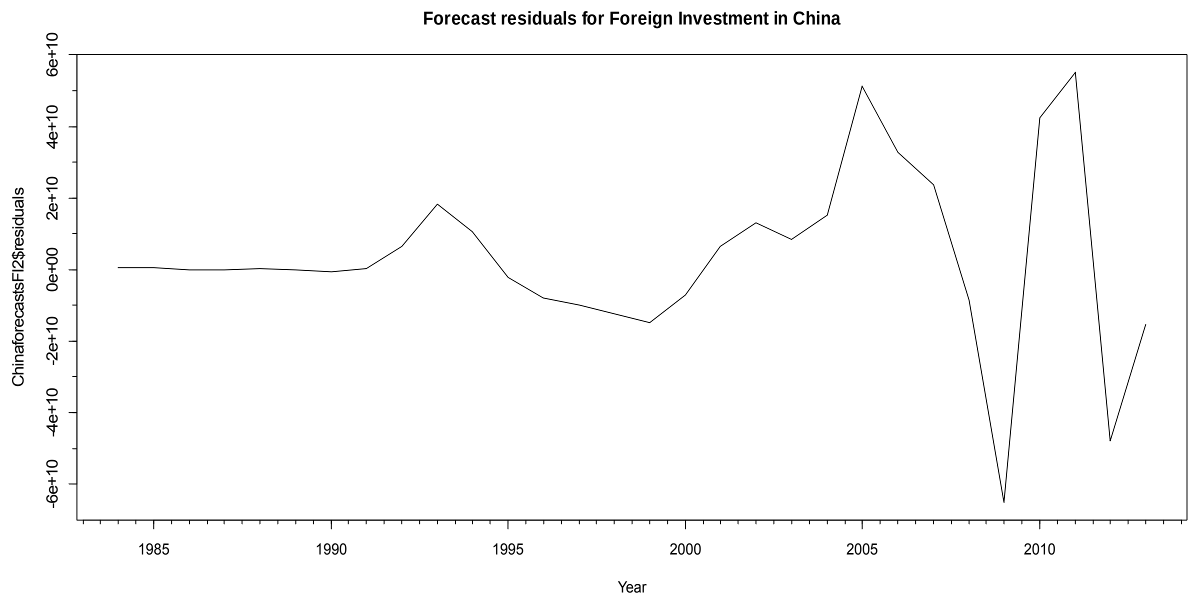
Foreign Investment – China



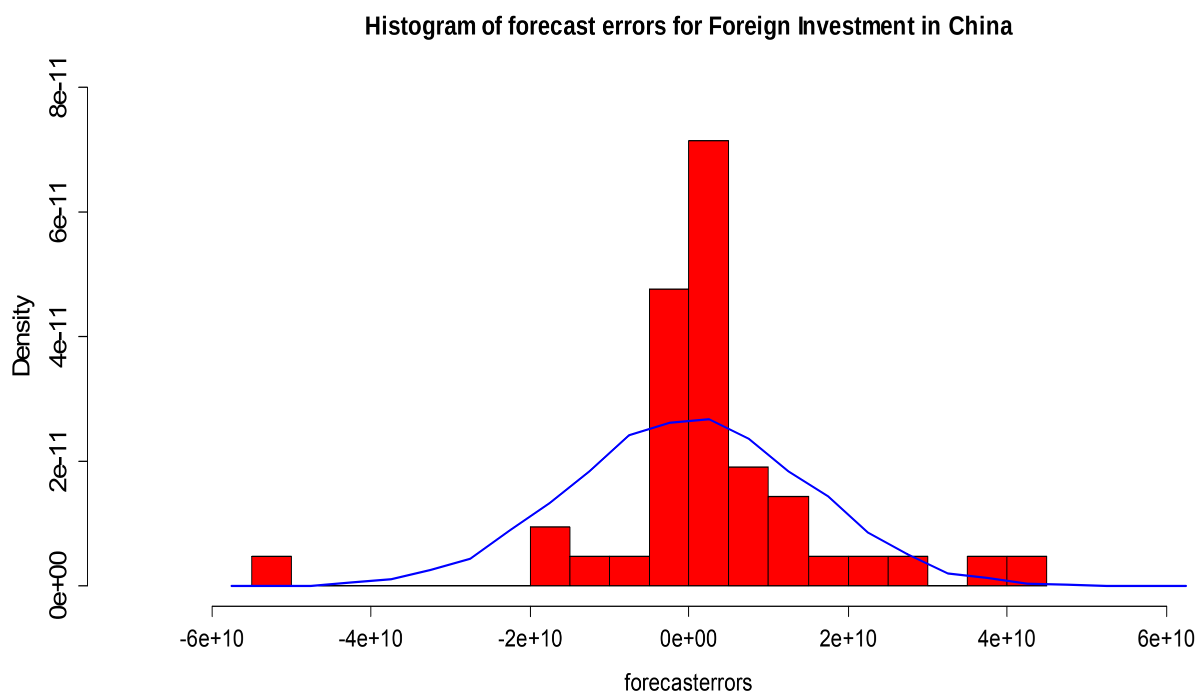
[Figure C.2.15] – Analysis for China, Foreign Investment and whole dataset



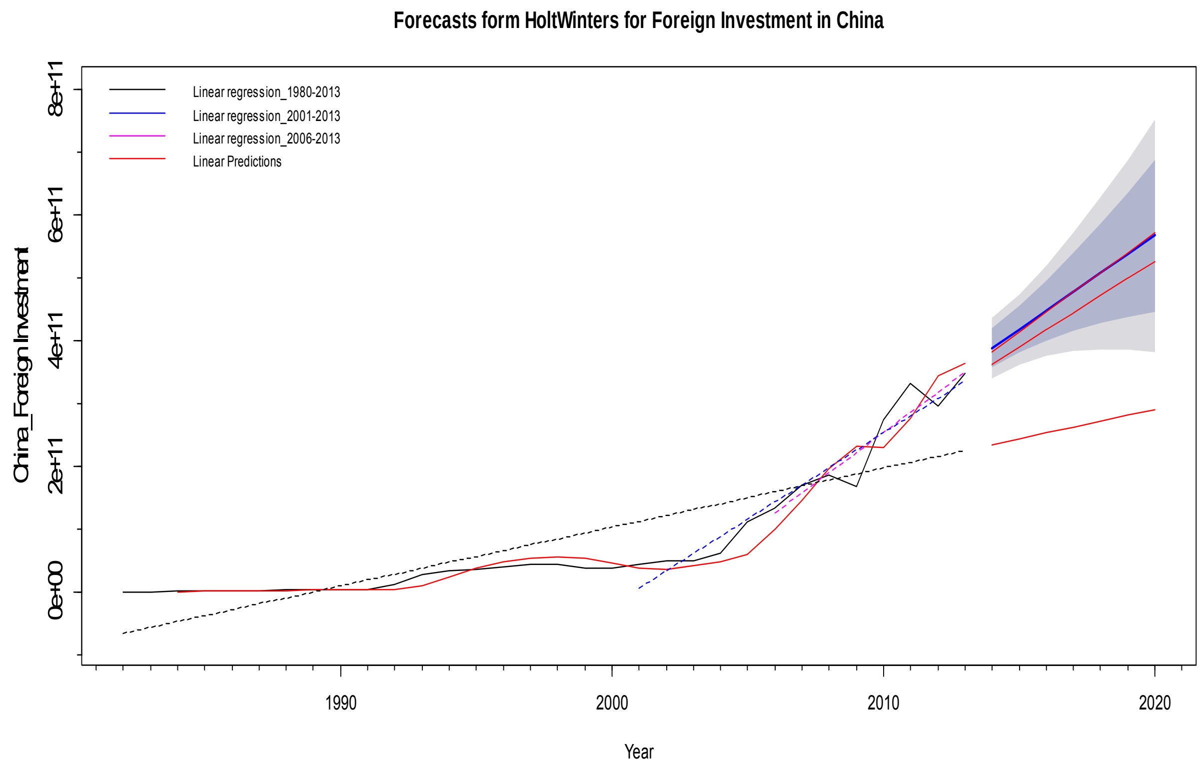
[Figure C.2.16] – Correlogram of in-sample errors



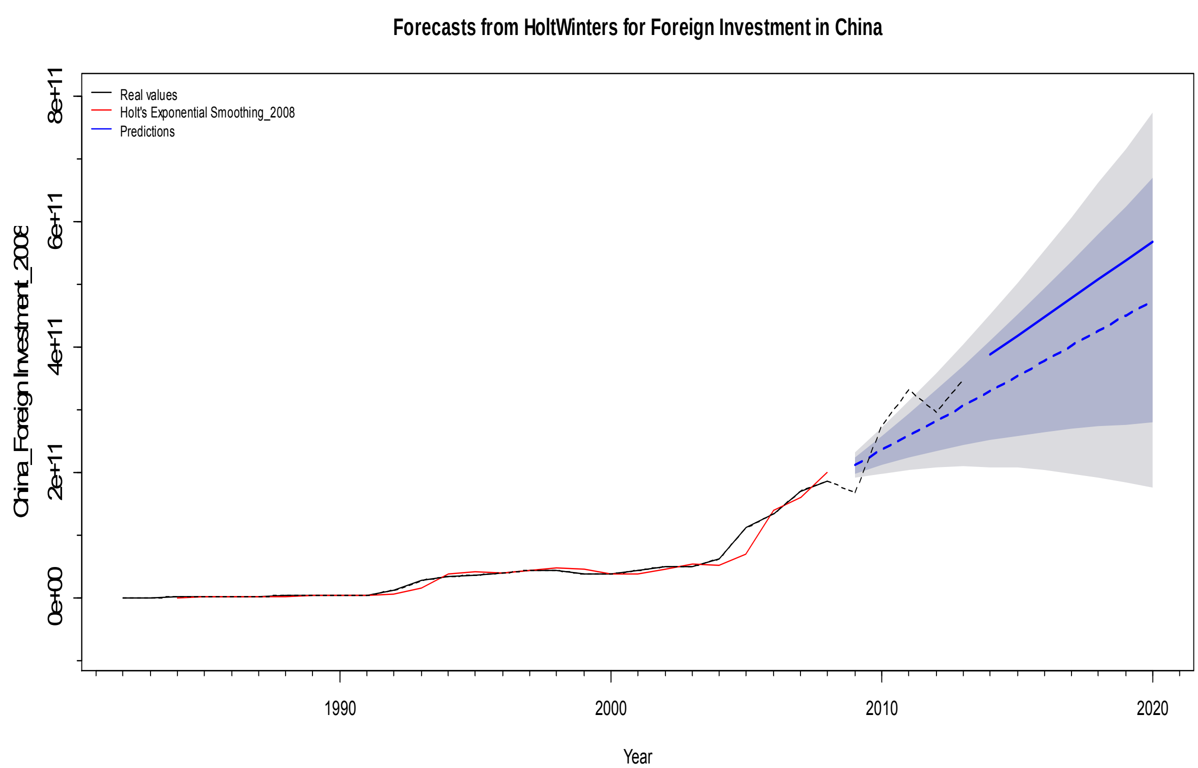
[Figure C.2.17] – Forecast residuals



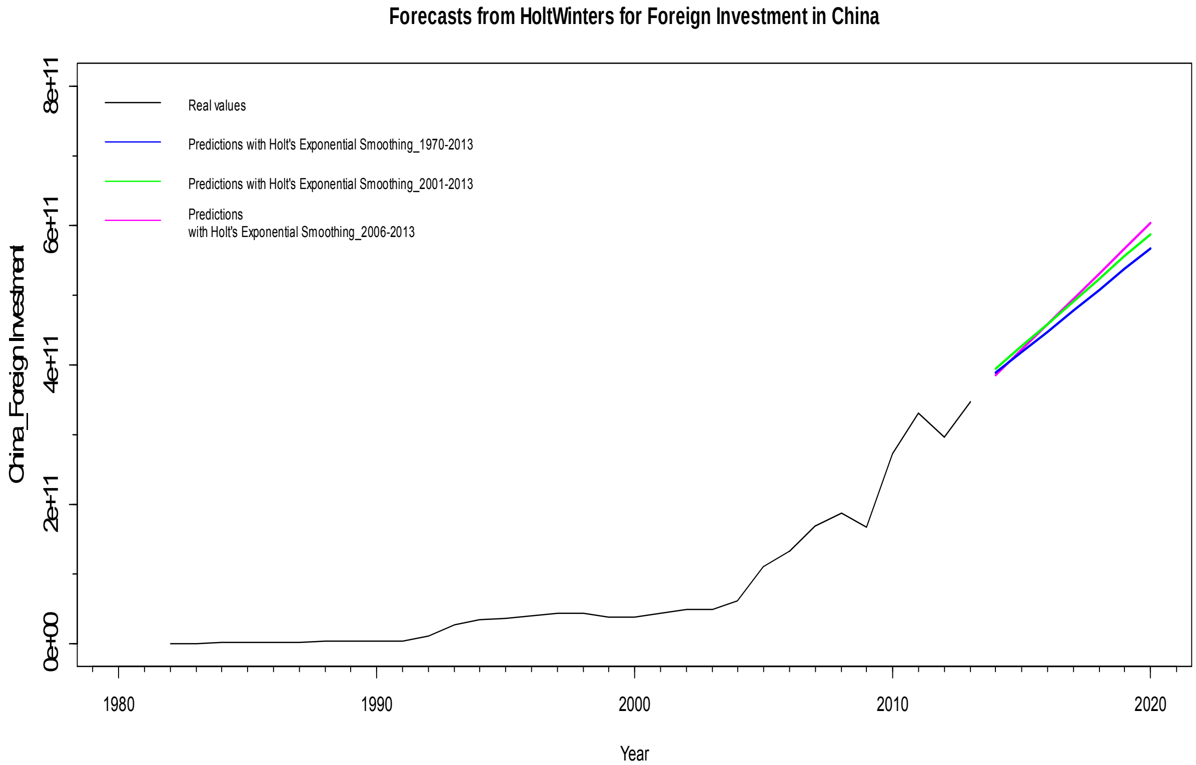
[Figure C.2.18] – Histogram and distribution of forecast residuals



[Figure C.2.19] – Comparison of Linear Regression and Holt's Exponential smoothing

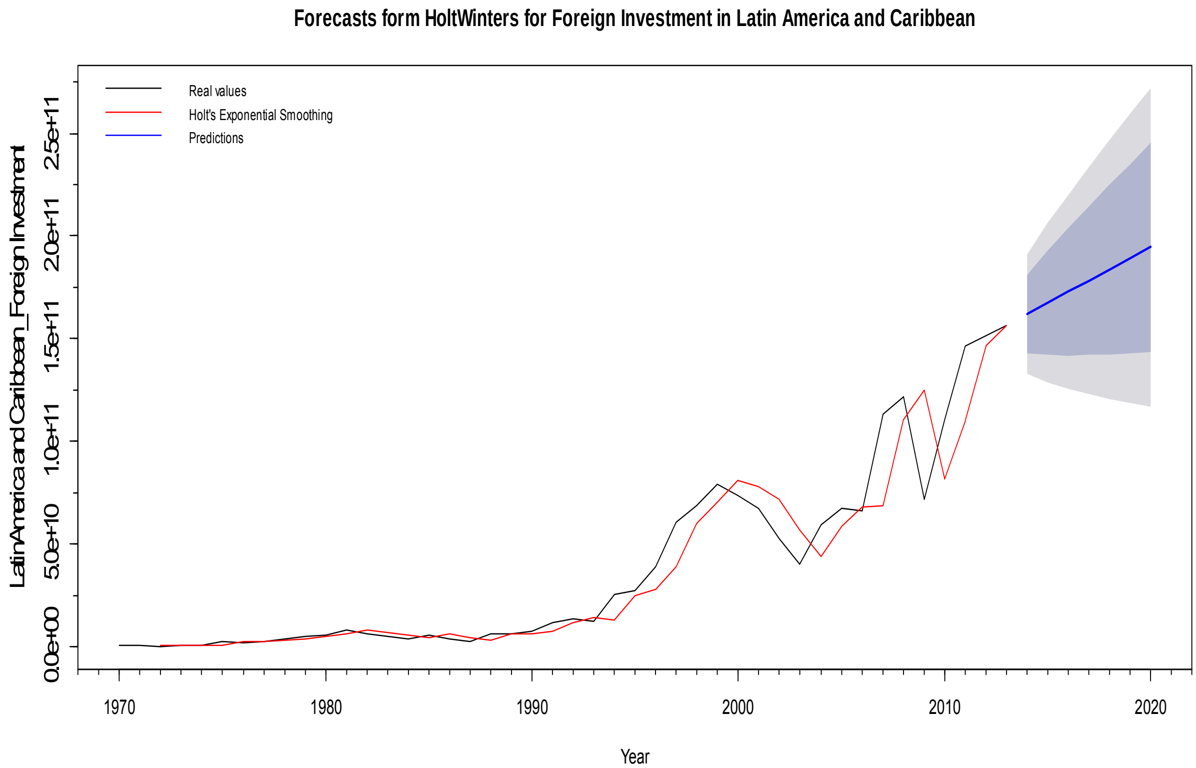


[Figure C.2.20] – Analysis for China, Foreign Investment and the dataset up to 2008

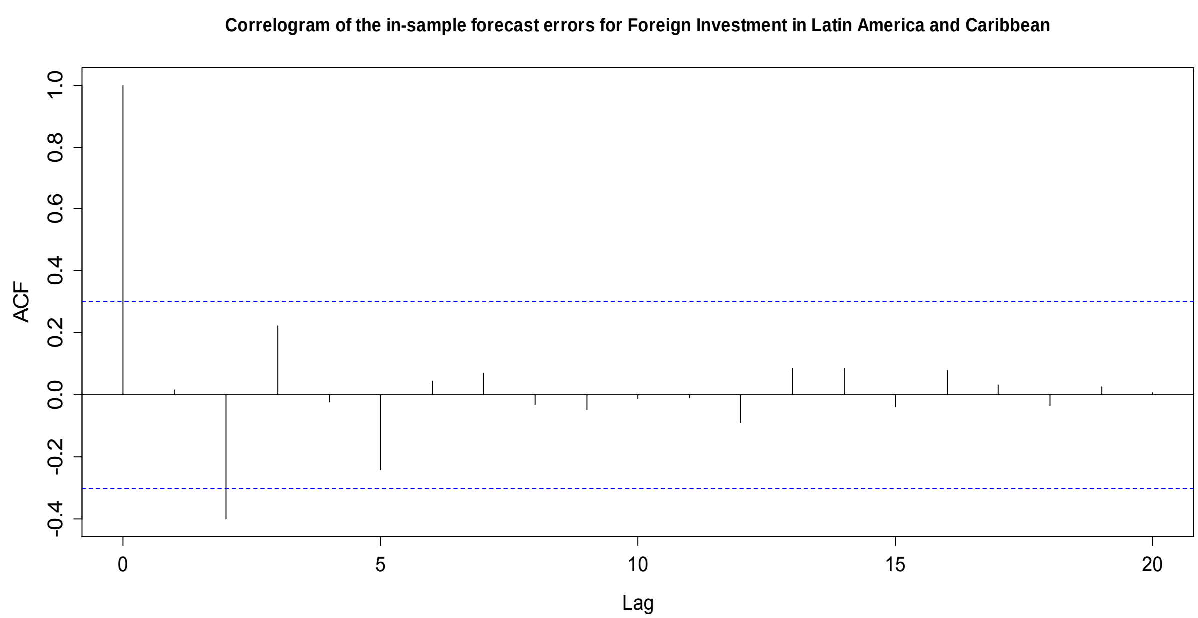


[Figure C.2.21] – Analyses for China, Foreign Investment and the subsets 2001-2013 and 2006-2013

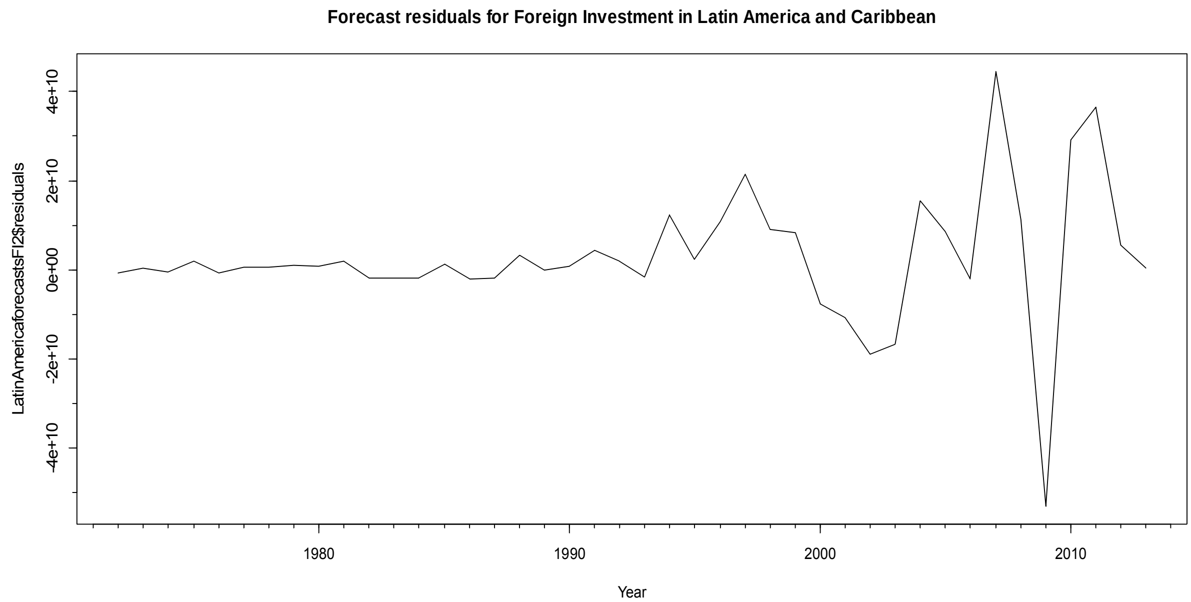
Foreign Investment – Latin America and Caribbean



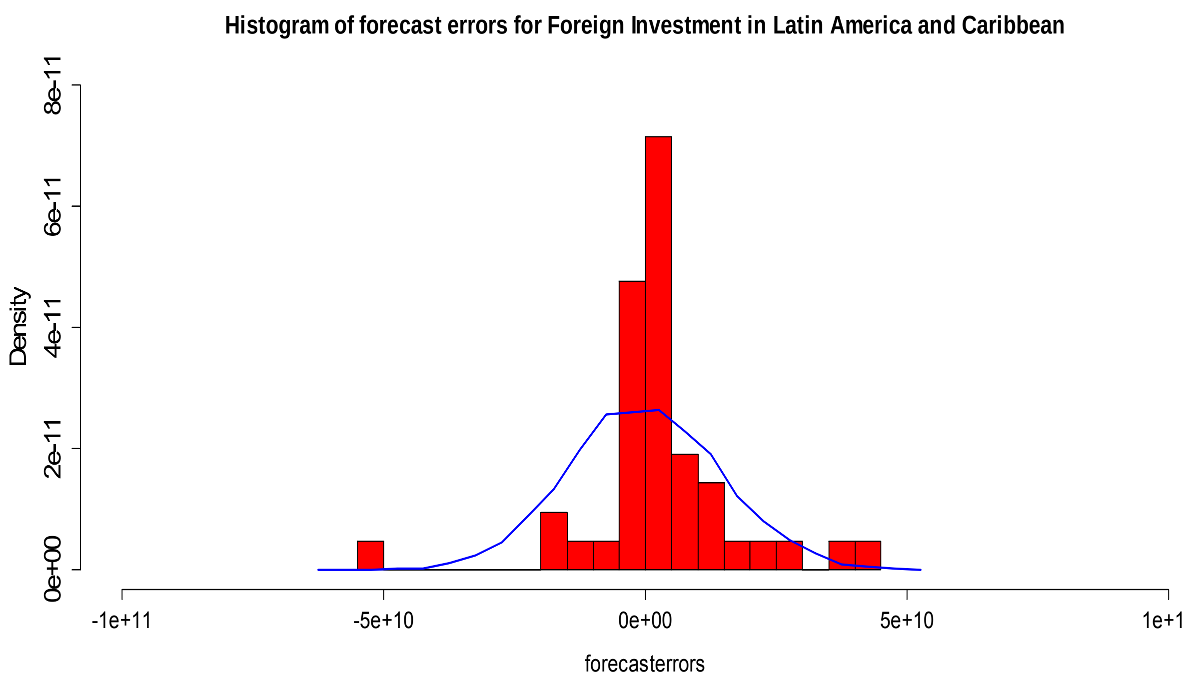
[Figure C.2.22] – Analysis for Latin America and Caribbean, Foreign Investment and whole dataset



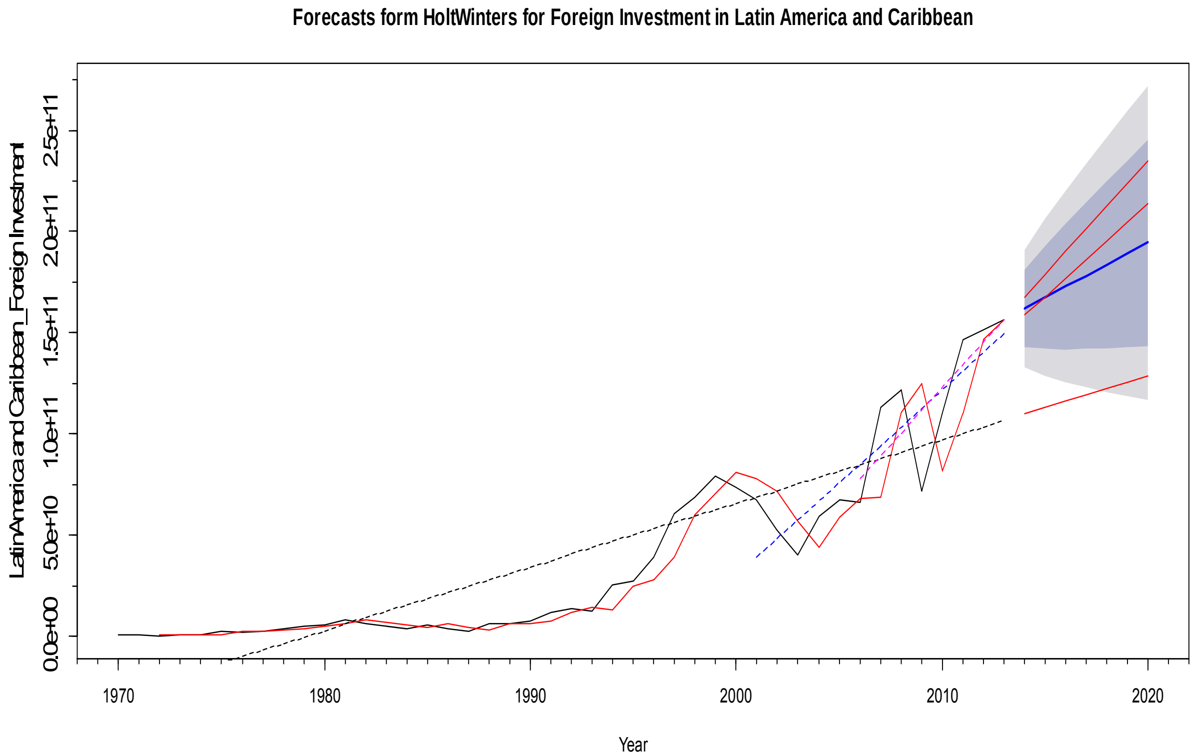
[Figure C.2.23] – Correlogram of in-sample errors



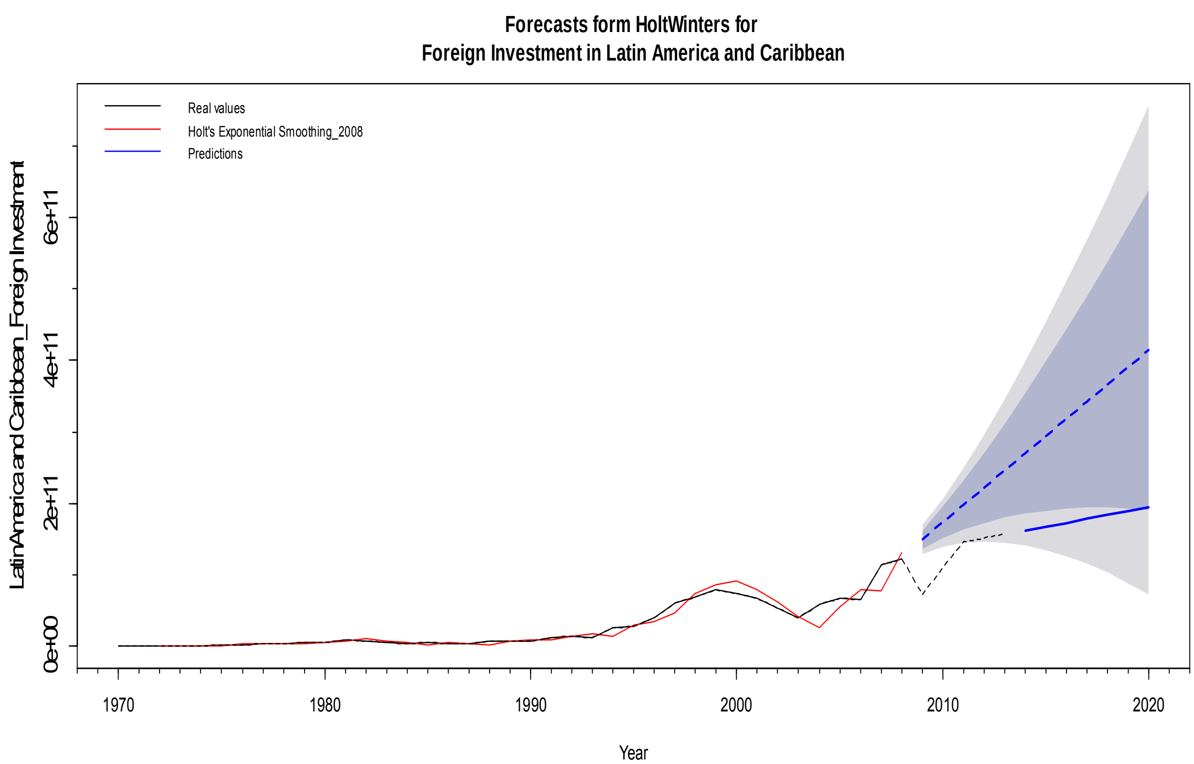
[Figure C.2.24] – Forecast residuals



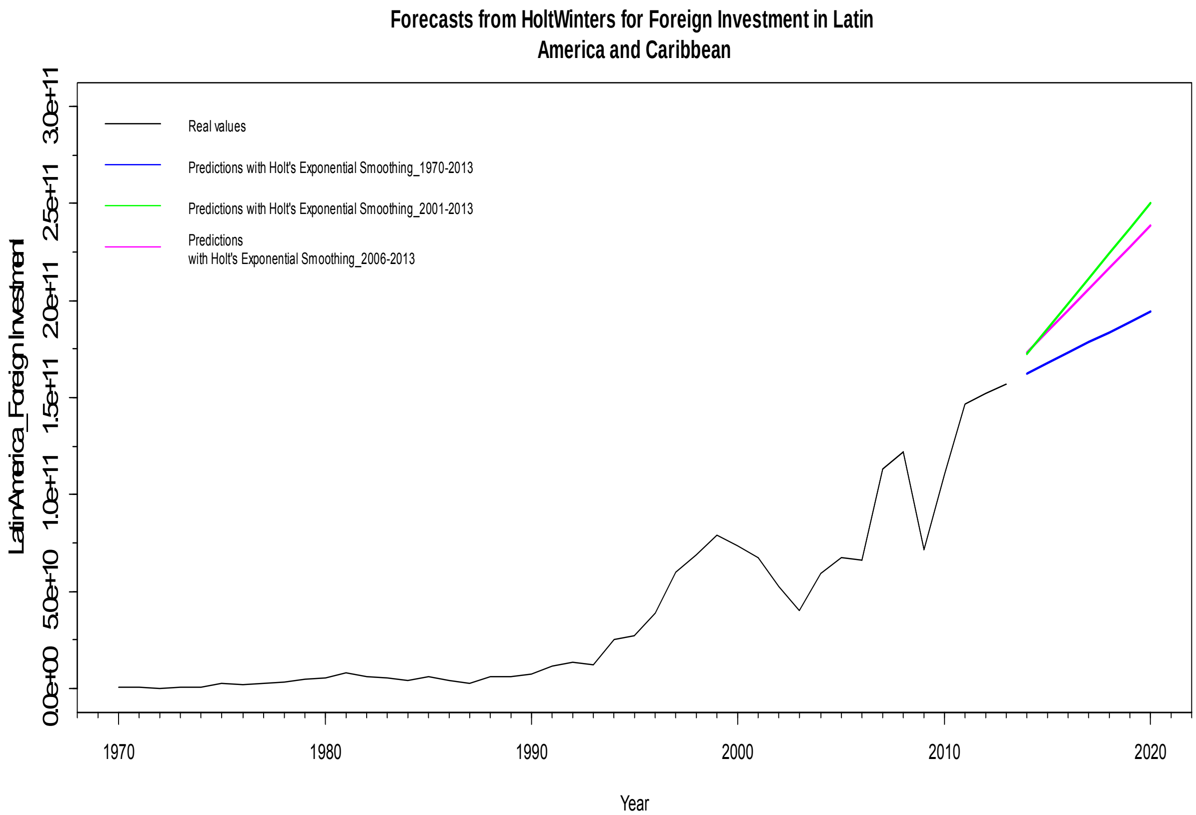
[Figure C.2.25] – Histogram and distribution of forecast residuals



[Figure C.2.26] – Comparison of Linear Regression and Holt's Exponential smoothing

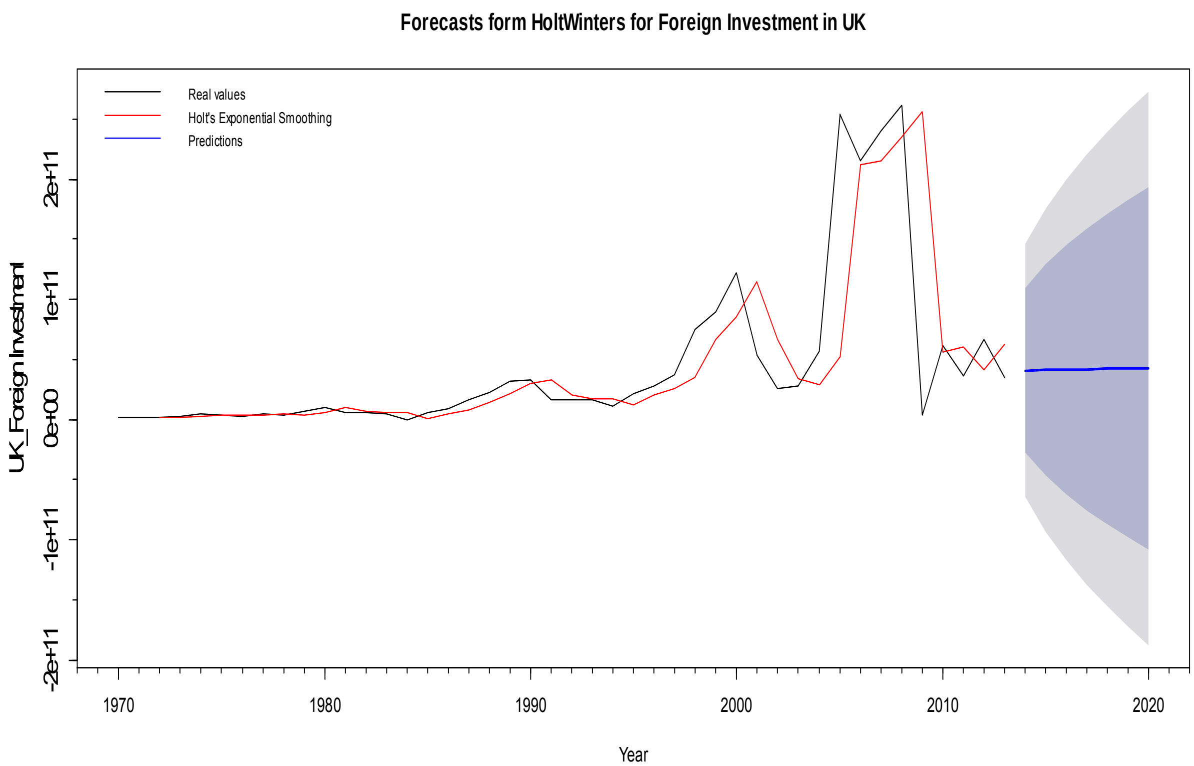


[Figure C.2.27] – Analysis for Latin America, Foreign Investment and the dataset up to 2008

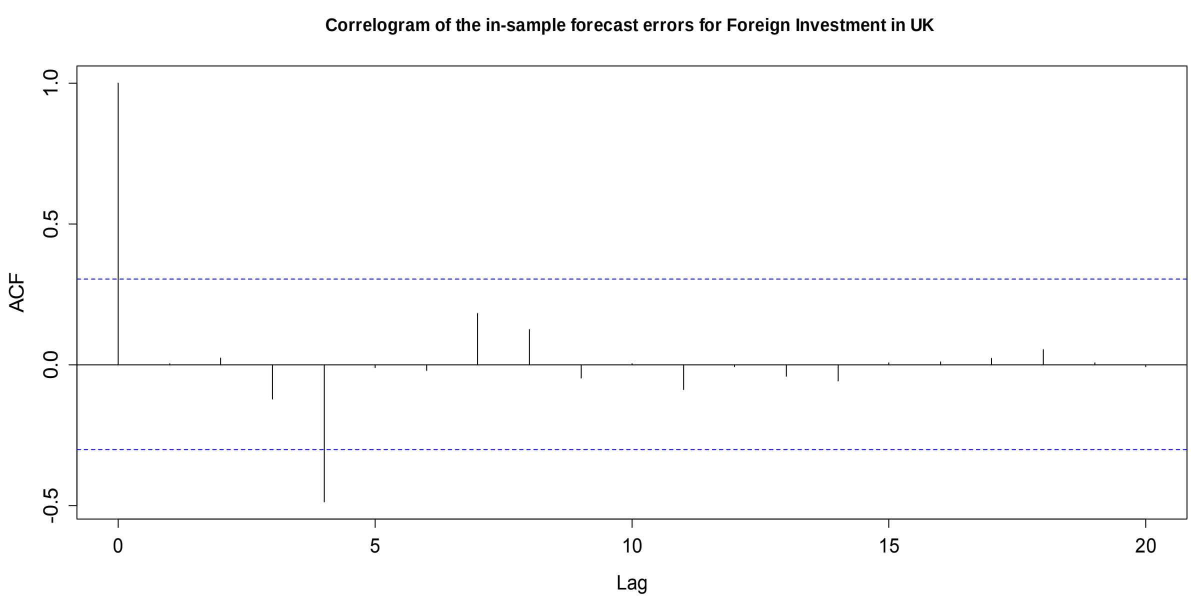


[Figure C.2.28] – Analyses for Latin America, Foreign Investment and the subsets 2001-2013 and 2006-2013

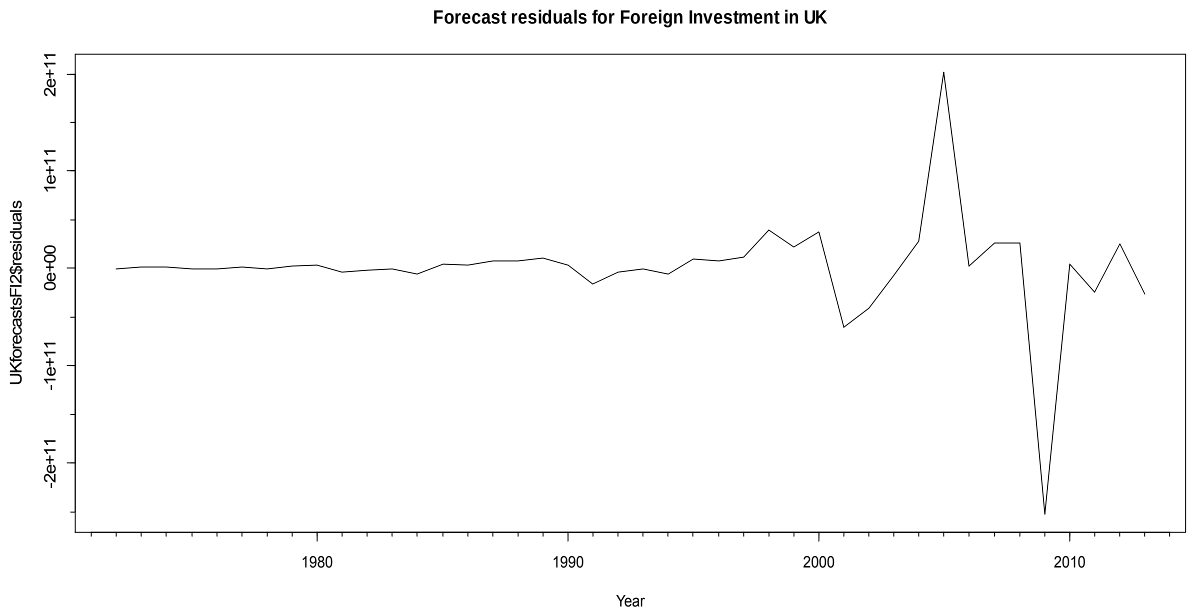
Foreign Investment – UK



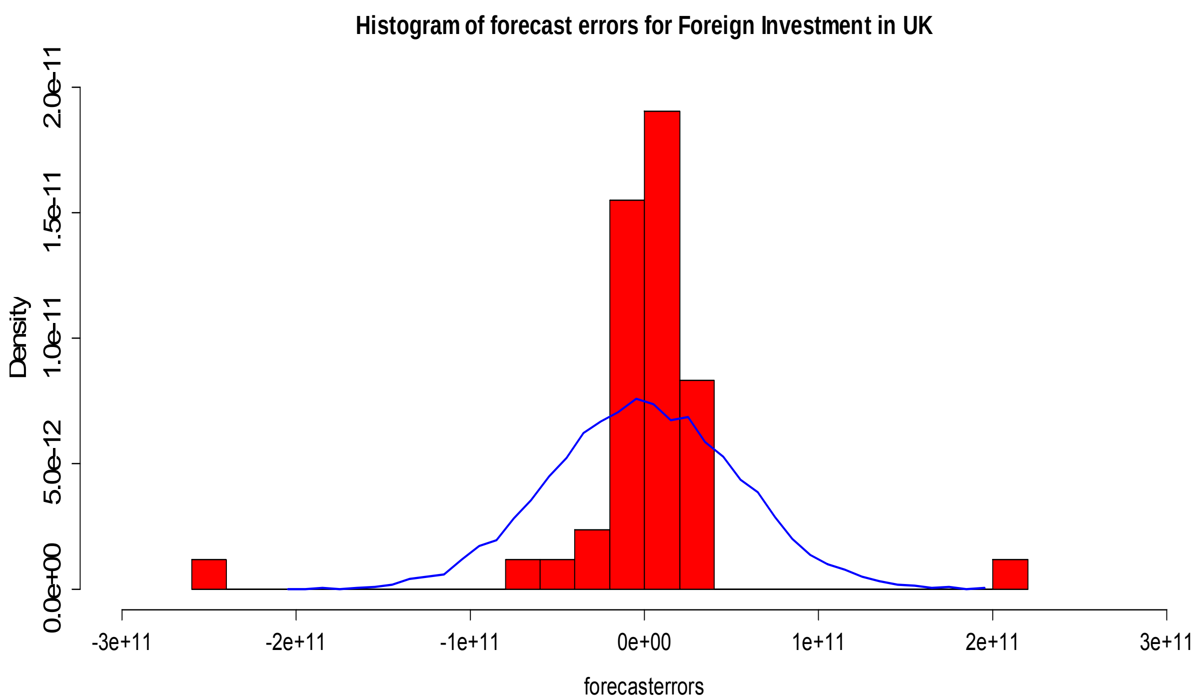
[Figure C.2.29] – Analysis for UK, Foreign Investment and whole dataset



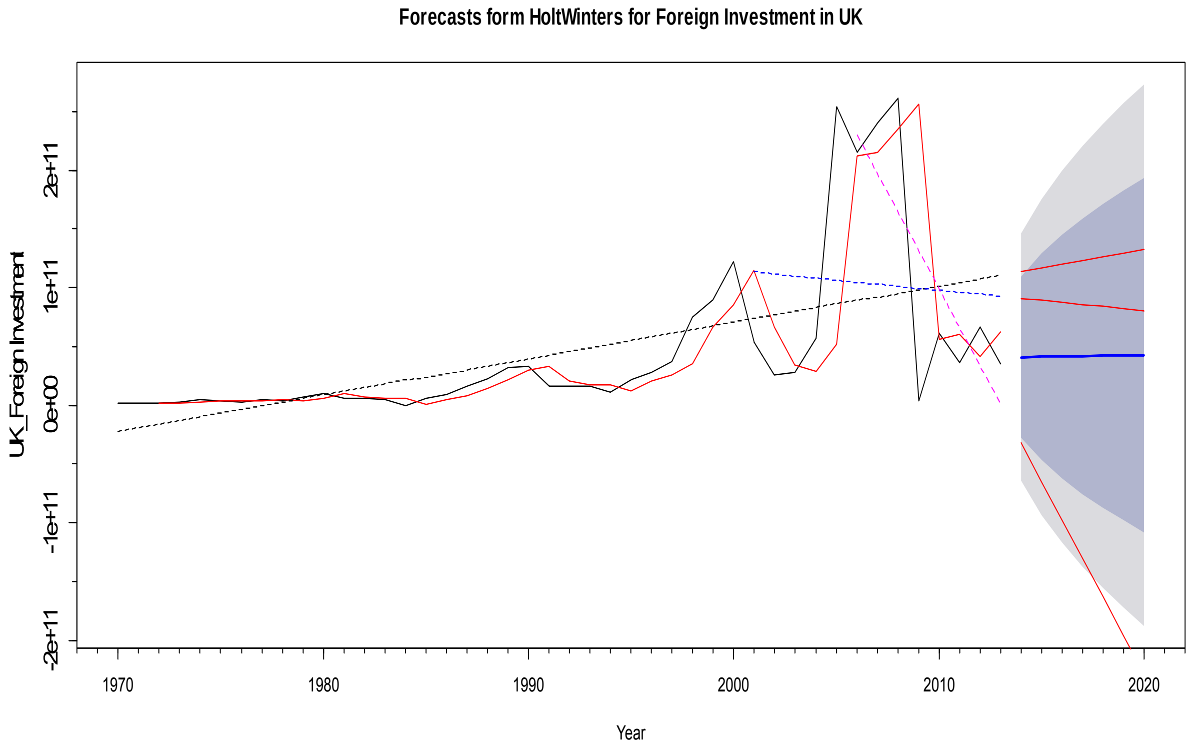
[Figure C.2.30] – Correlogram of in-sample errors



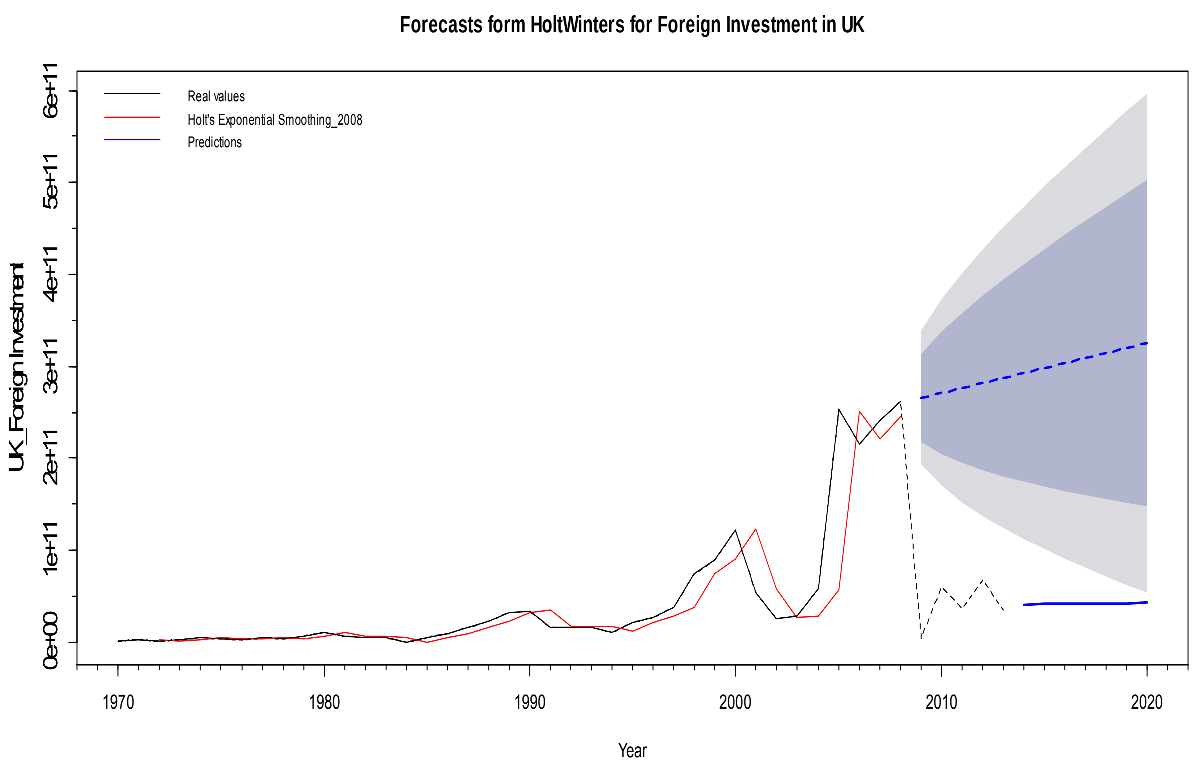
[Figure C.2.31] – Forecast residuals



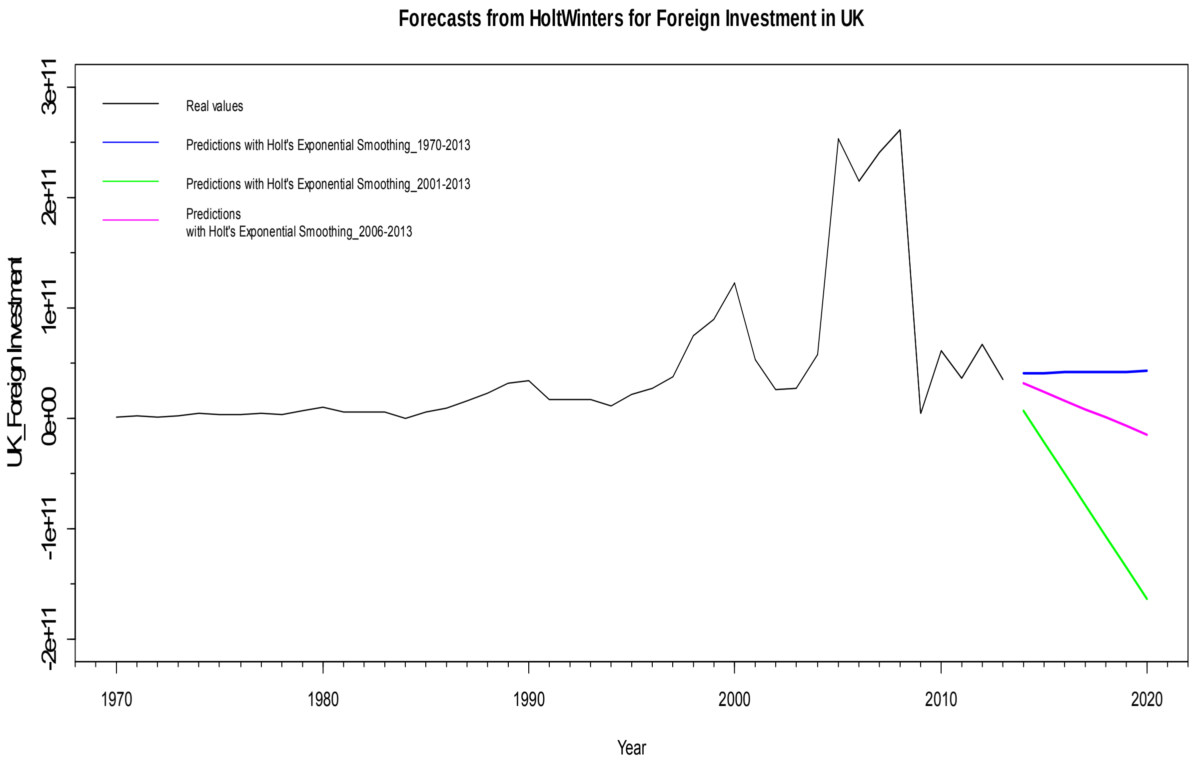
[Figure C.2.32] – Histogram and distribution of forecast residuals



[Figure C.2.33] – Comparison of Linear Regression and Holt's Exponential smoothing

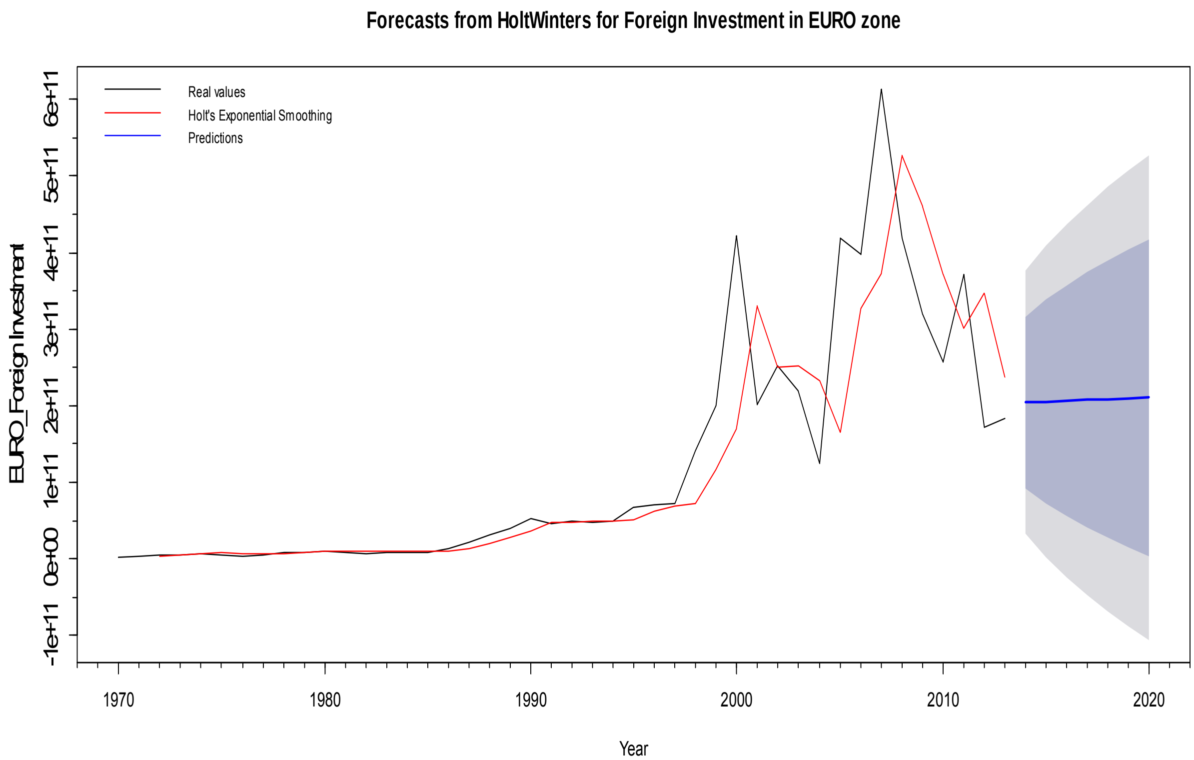


[Figure C.2.34] – Analysis for UK, Foreign Investment and the dataset up to 2008

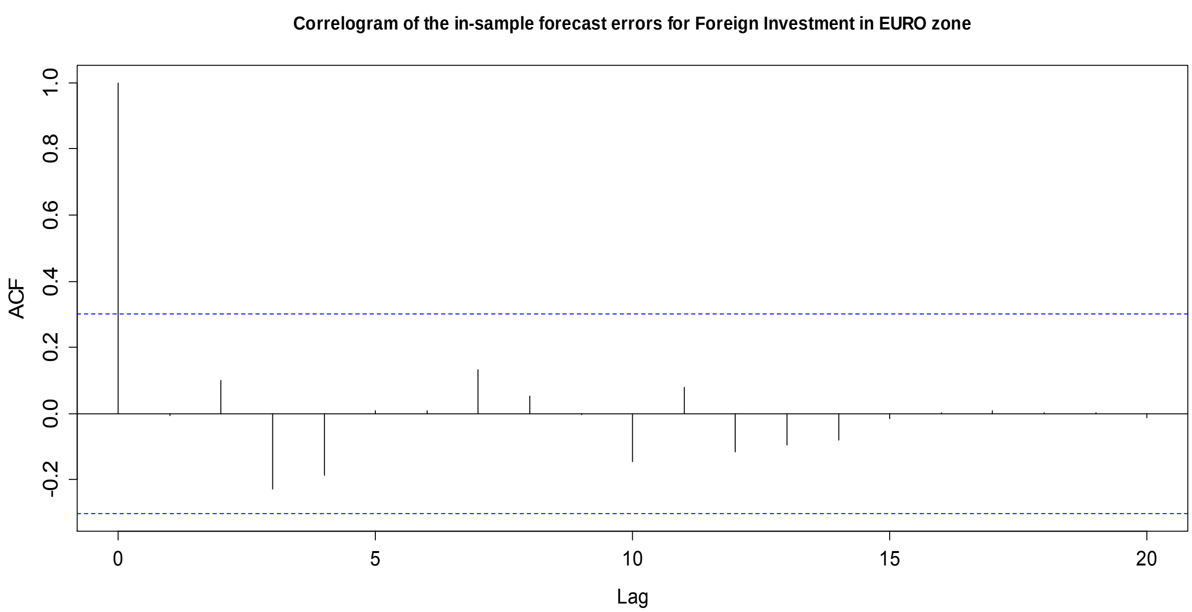


[Figure C.2.35] – Analyses for UK, Foreign Investment and the subsets 2001-2013 and 2006-2013

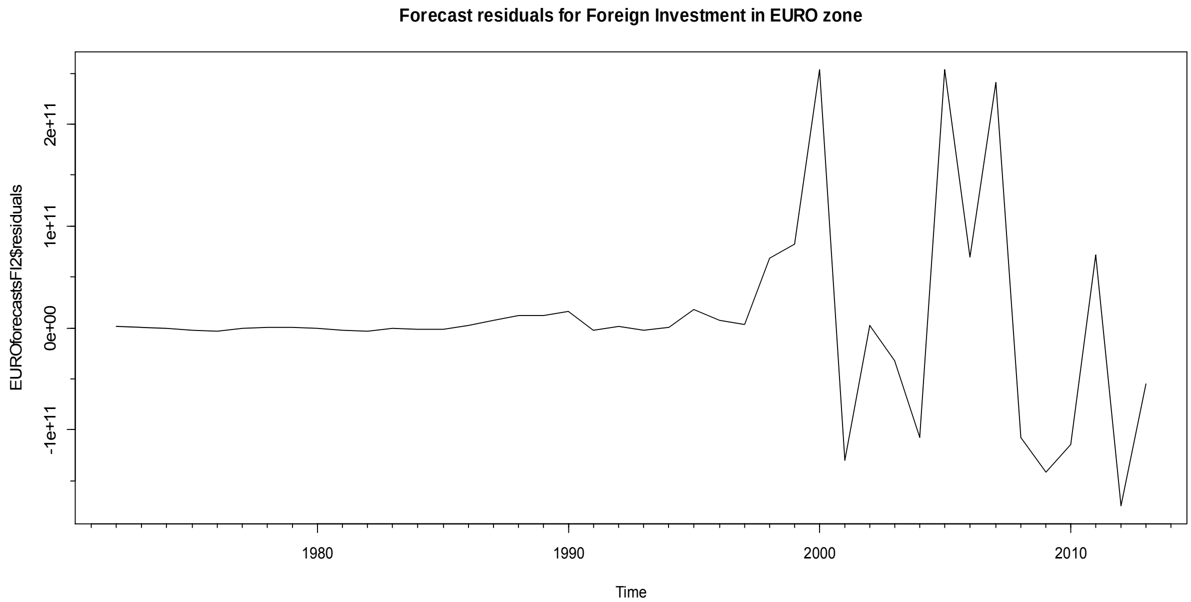
Foreign Investment – EURO zone



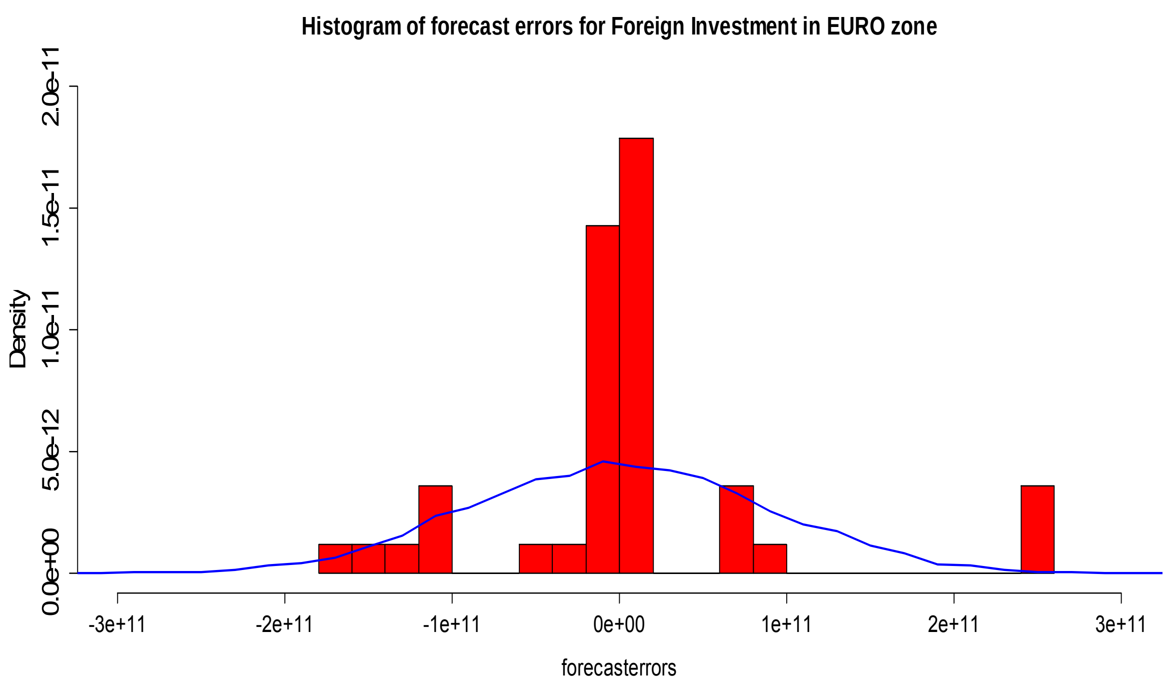
[Figure C.2.36] – Analysis for EURO zone, Foreign Investment and whole dataset



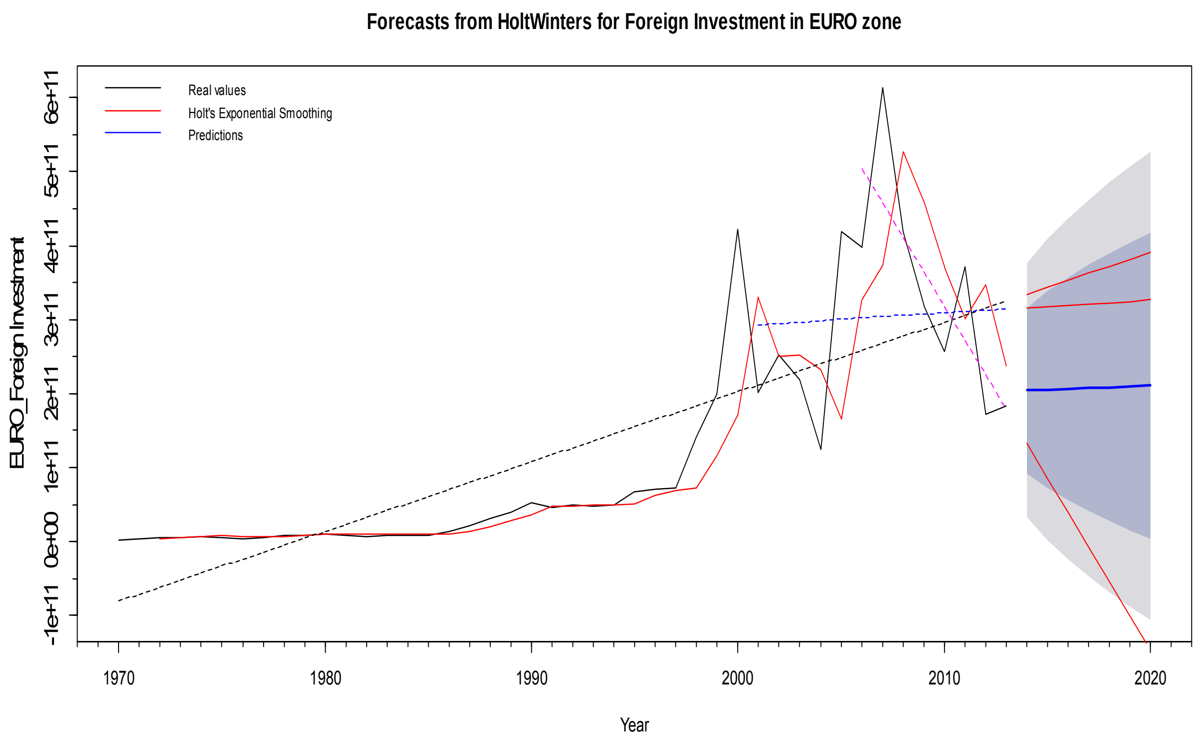
[Figure C.2.37] – Correlogram of in-sample errors



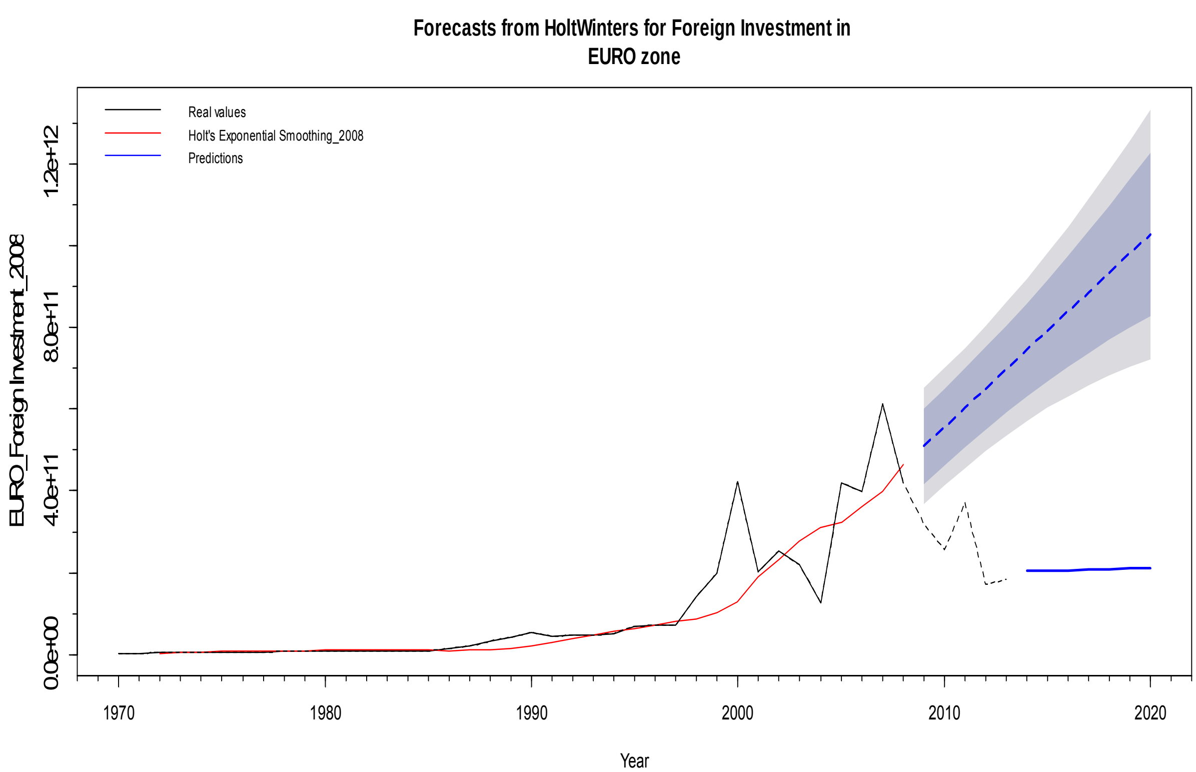
[Figure C.2.38] – Forecast residuals



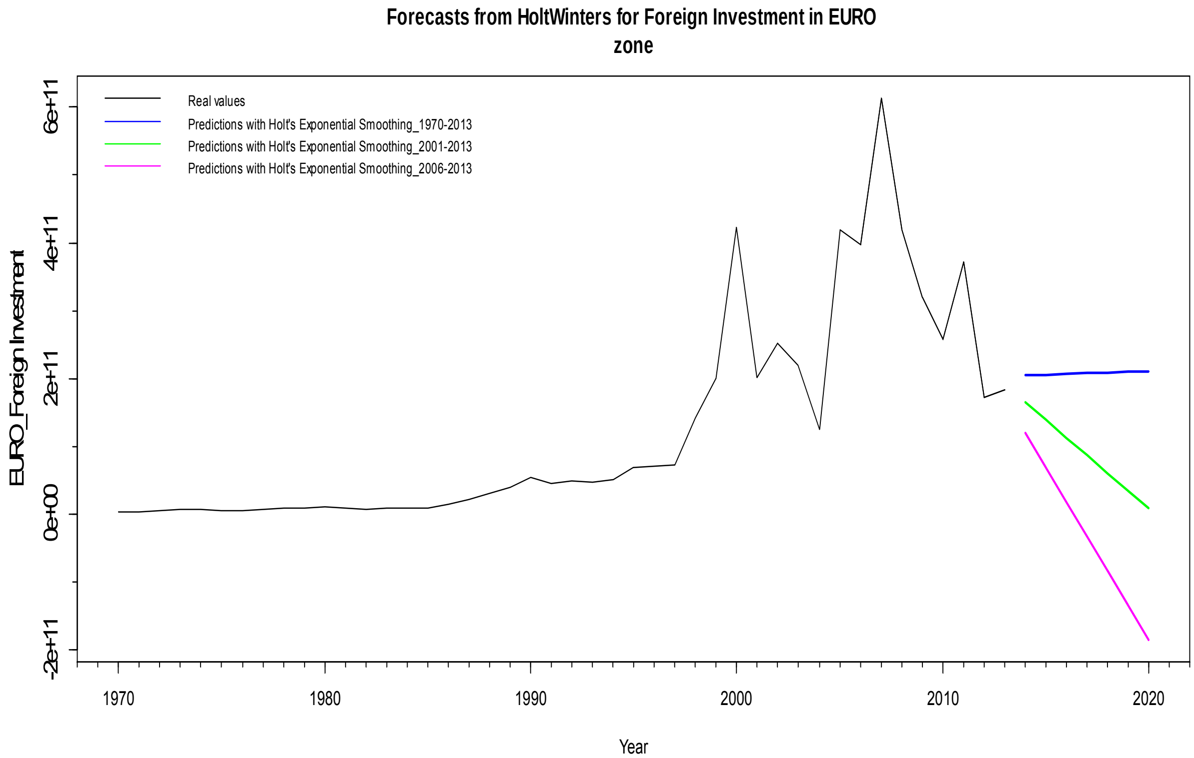
[Figure C.2.39] – Histogram and distribution of forecast residuals



[Figure C.2.40] – Comparison of Linear Regression and Holt's Exponential smoothing

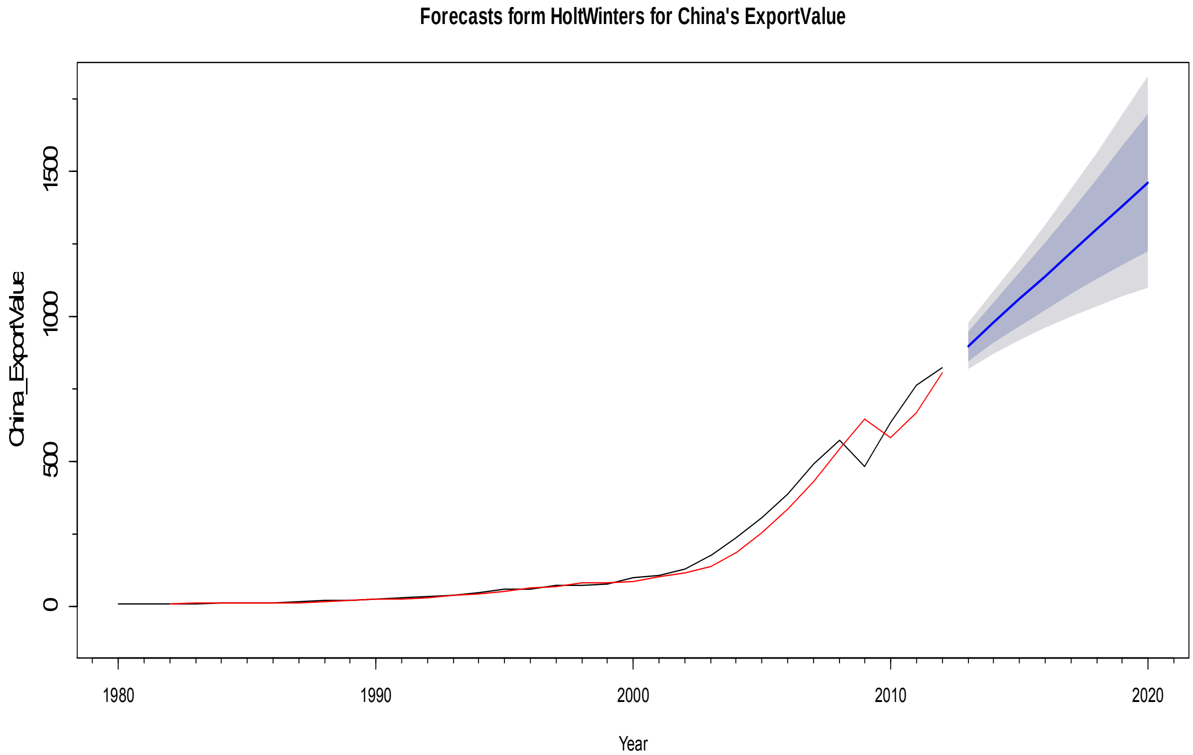


[Figure C.2.41] – Analysis for EURO zone, Foreign Investment and the dataset up to 2008

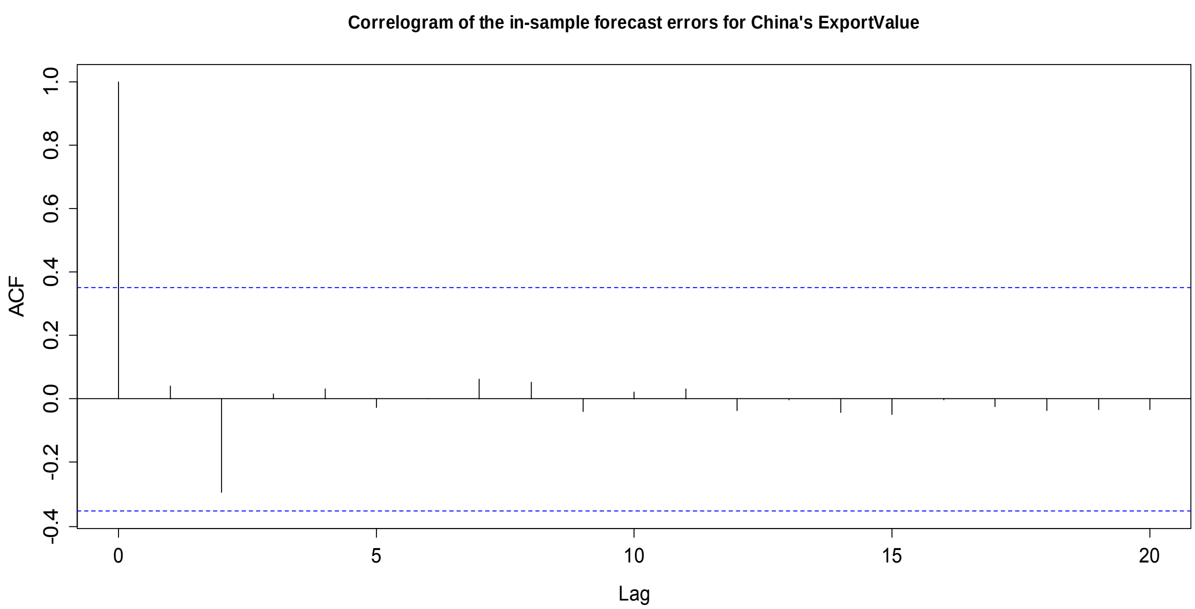


[Figure C.2.42] – Analyses for EURO zone, Foreign Investment and the subsets 2001-2013 and 2006-2013

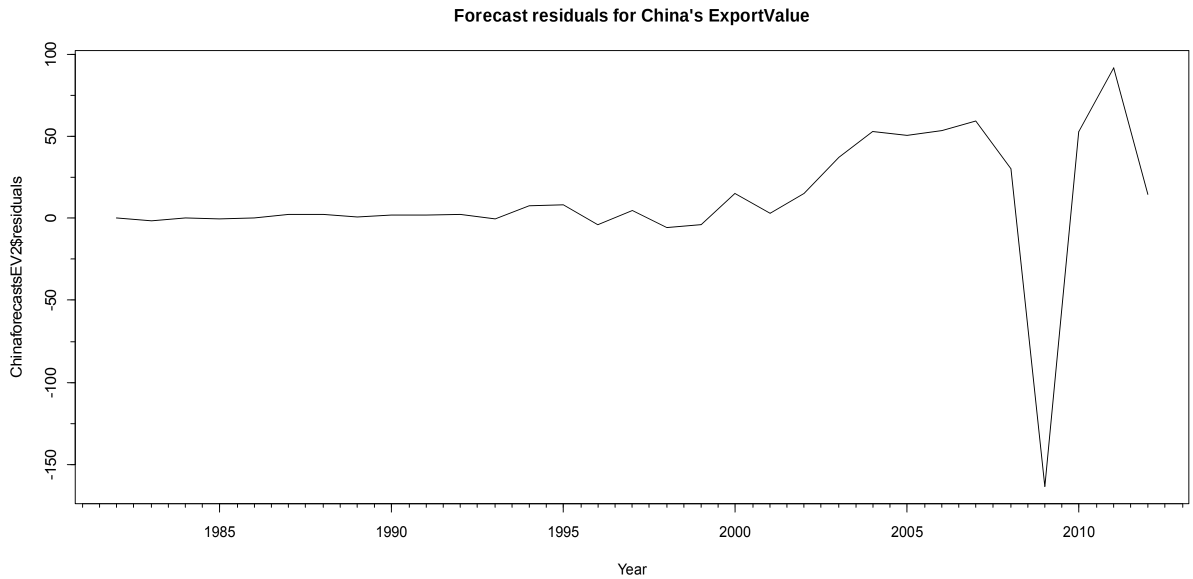
Export Value – China



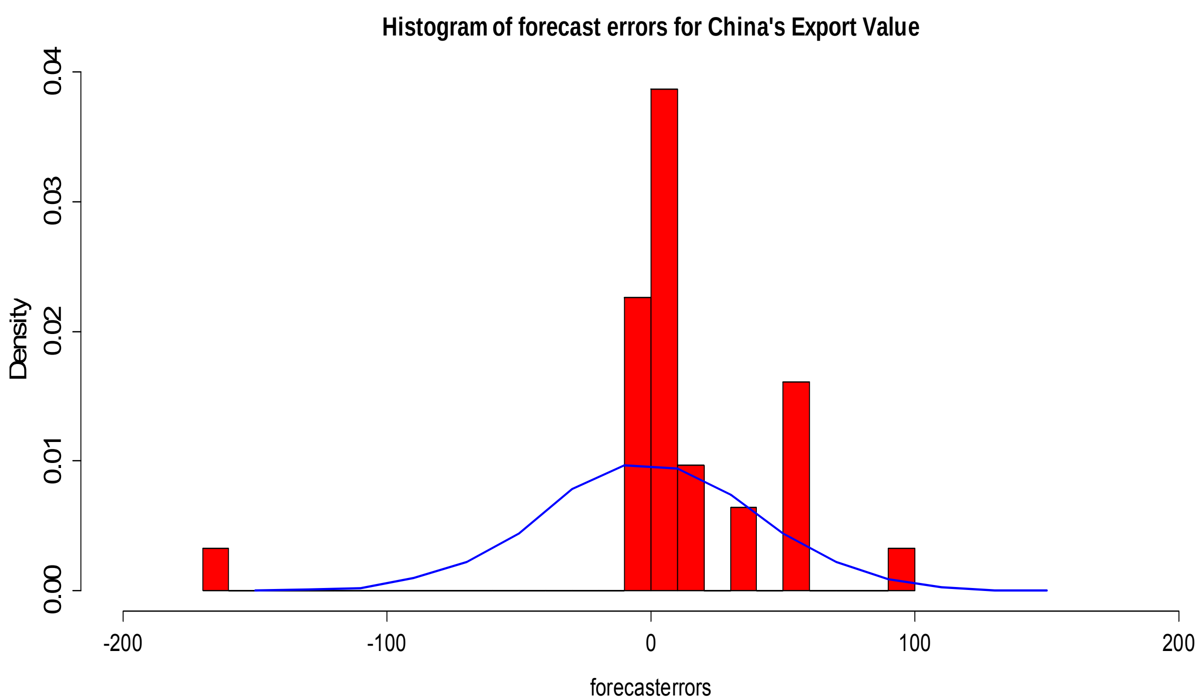
[Figure C.2.43] – Analysis for China, Export Value and whole dataset



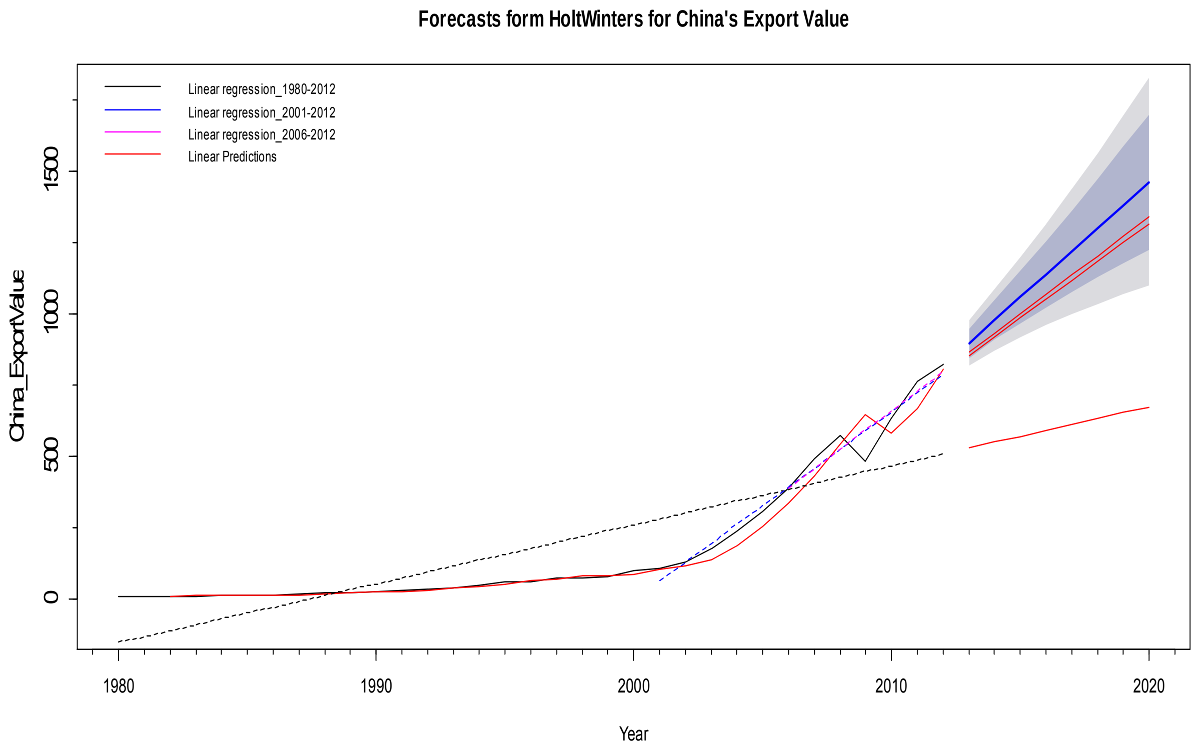
[Figure C.2.44] – Correlogram of in-sample errors



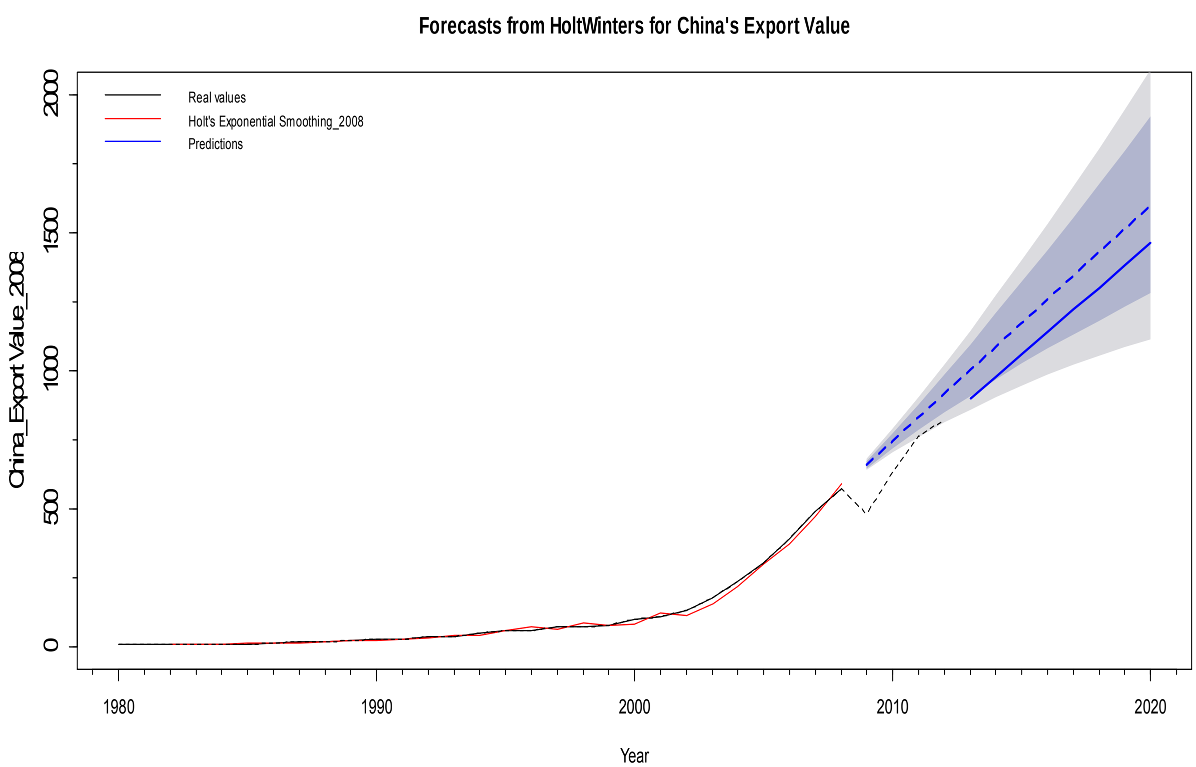
[Figure C.2.45] – Forecast residuals



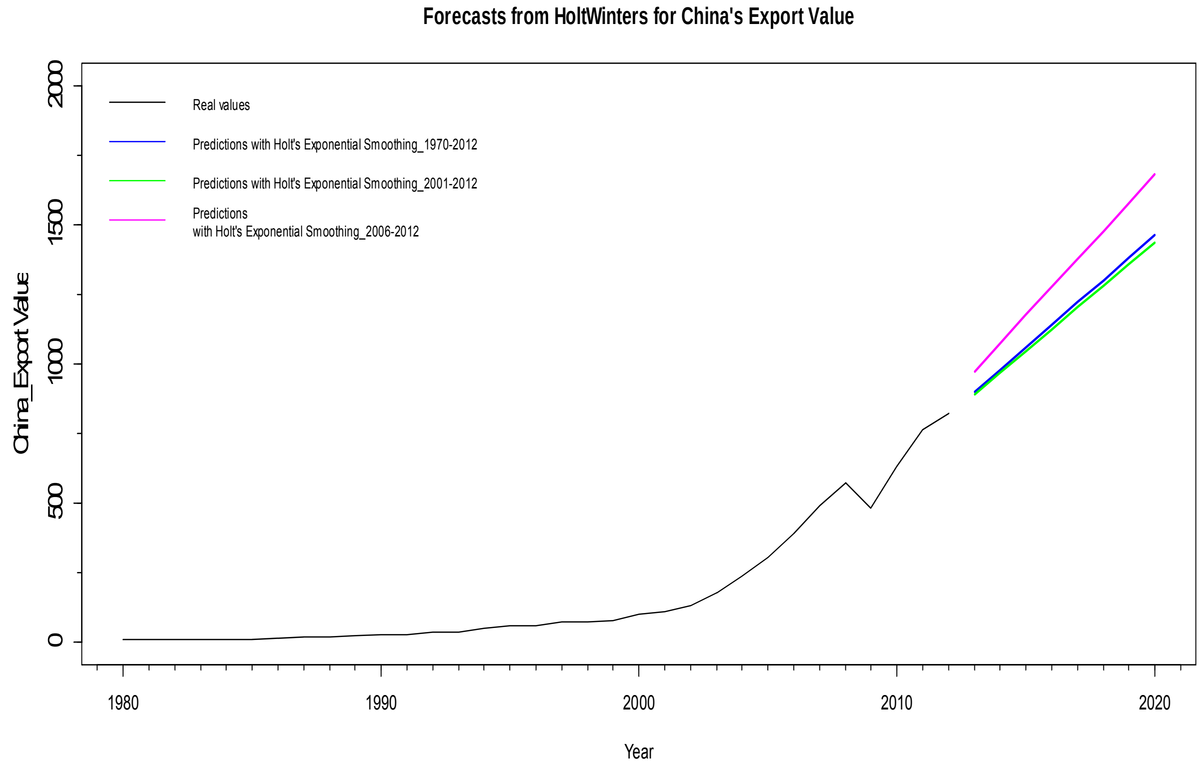
[Figure C.2.46] – Histogram and distribution of forecast residuals



[Figure C.2.47] – Comparison of Linear Regression and Holt's Exponential smoothing

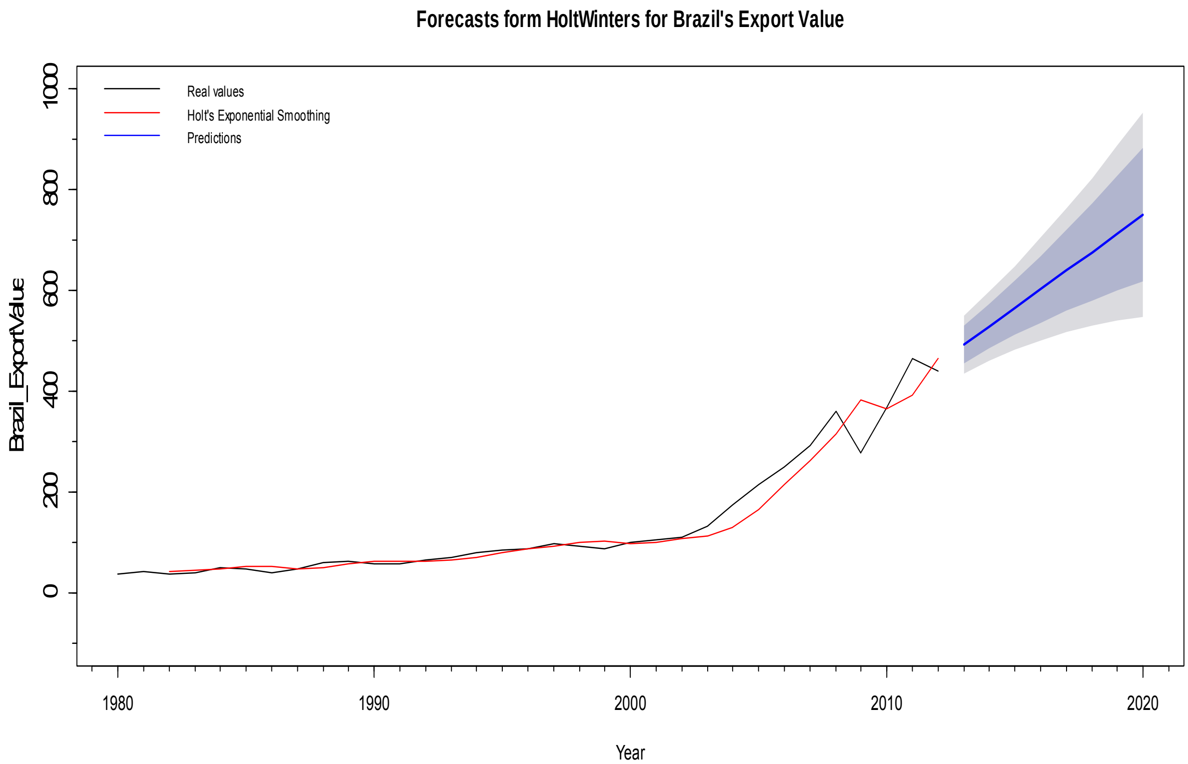


[Figure C.2.48] – Analysis for China, Export Value and the dataset up to 2008

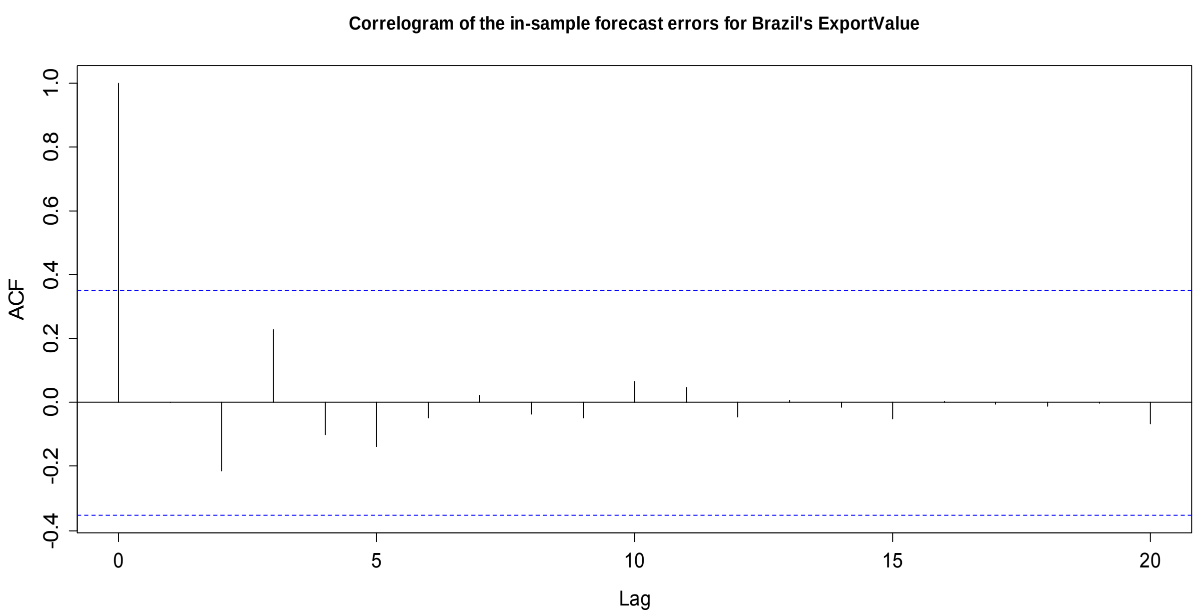


[Figure C.2.49] – Analyses for China, Export Value and the subsets 2001-2013 and 2006-2013

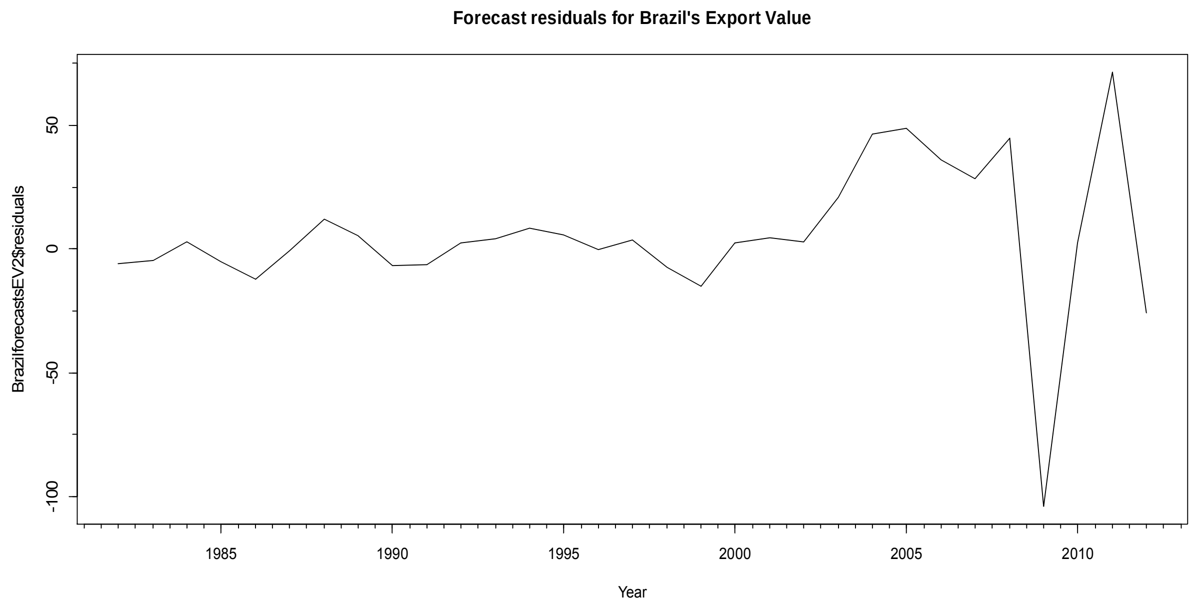
Export Value – Brazil



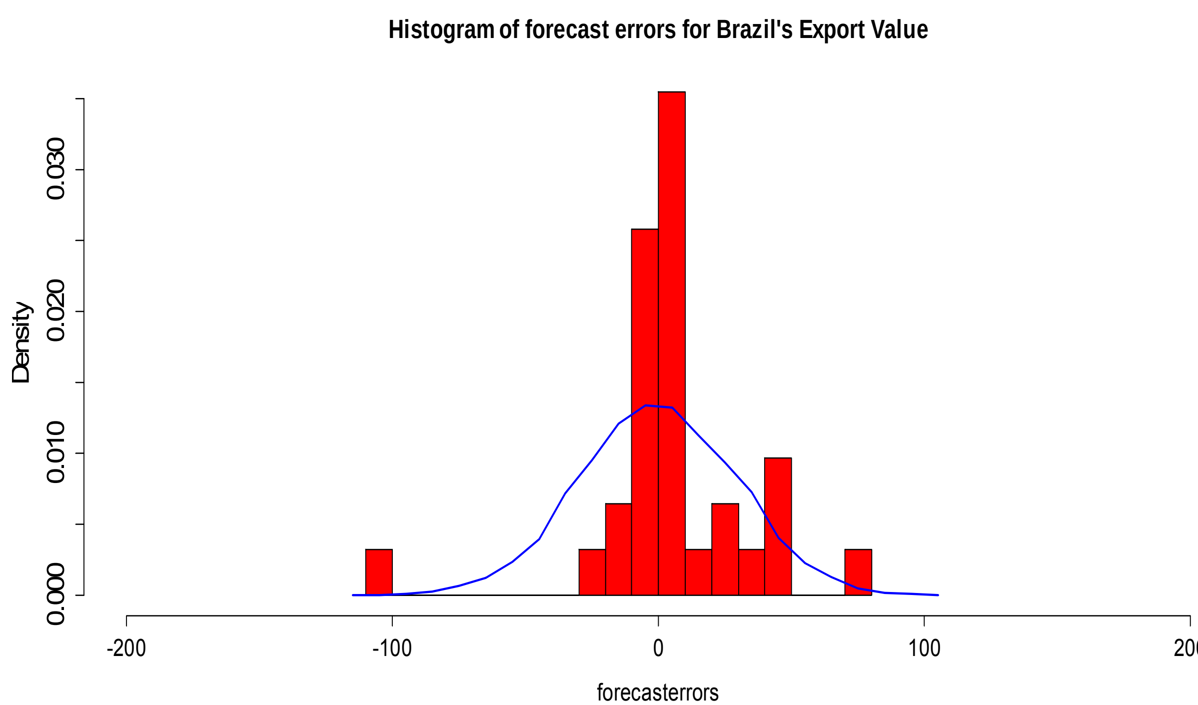
[Figure C.2.50] – Analysis for Brazil, Export Value and whole dataset



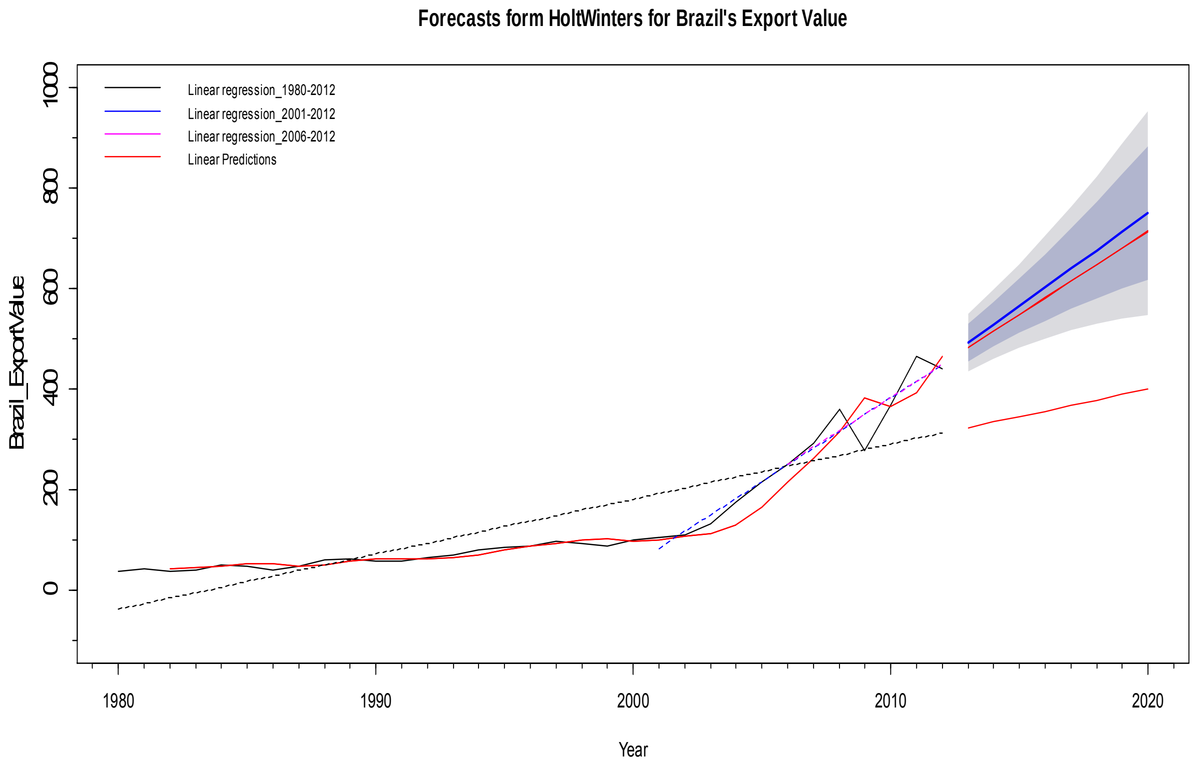
[Figure C.2.51] – Correlogram of in-sample errors



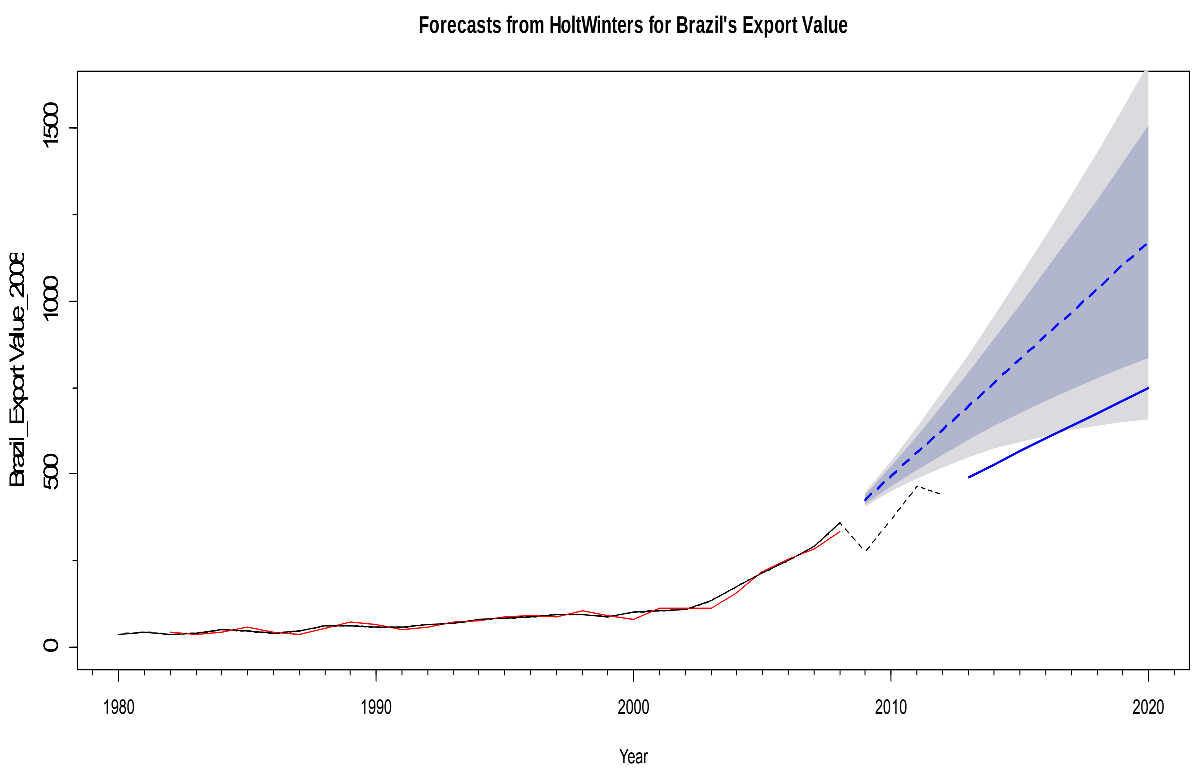
[Figure C.2.52] – Forecast residuals



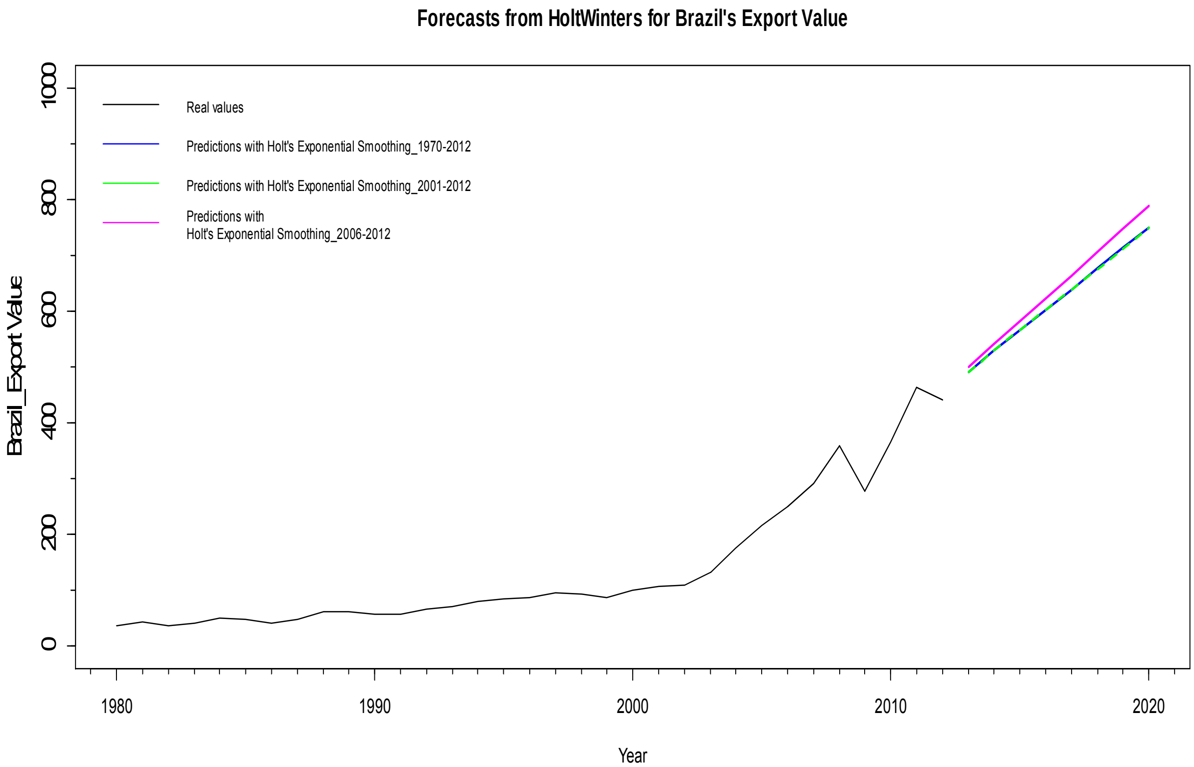
[Figure C.2.53] – Histogram and distribution of forecast residuals



[Figure C.2.54] – Comparison of Linear Regression and Holt's Exponential smoothing

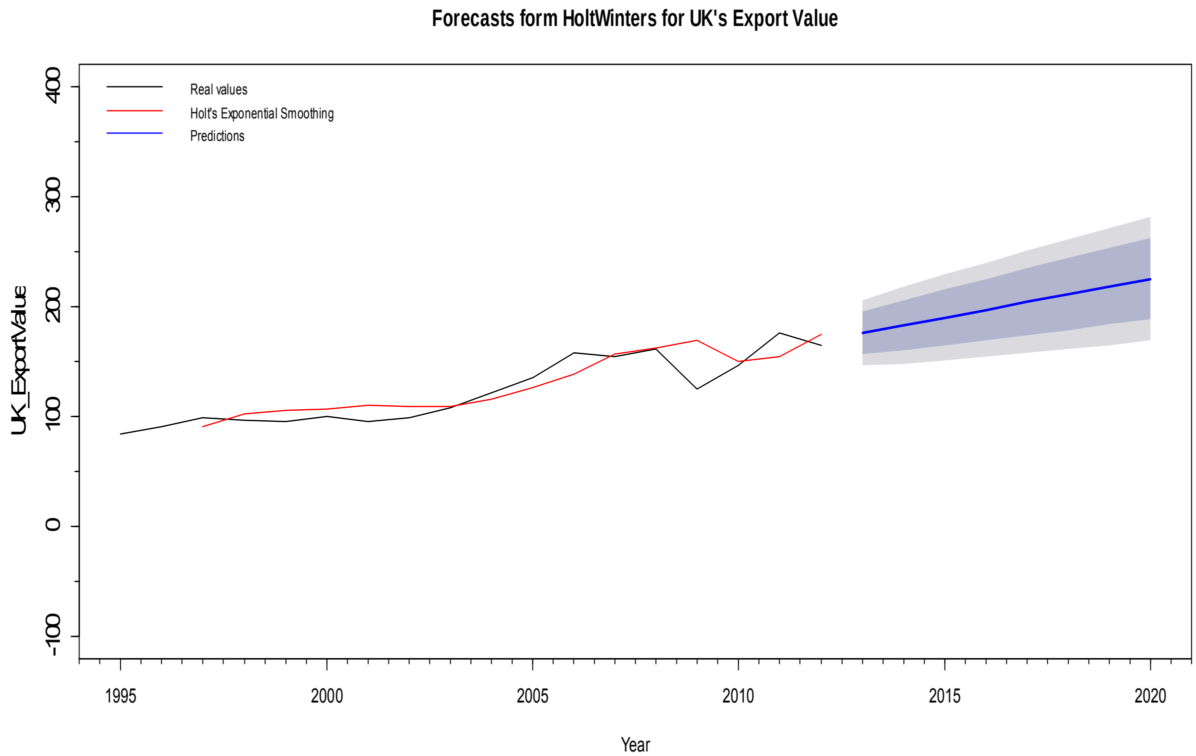


[Figure C.2.55] – Analysis for Brazil, Export Value and the dataset up to 2008

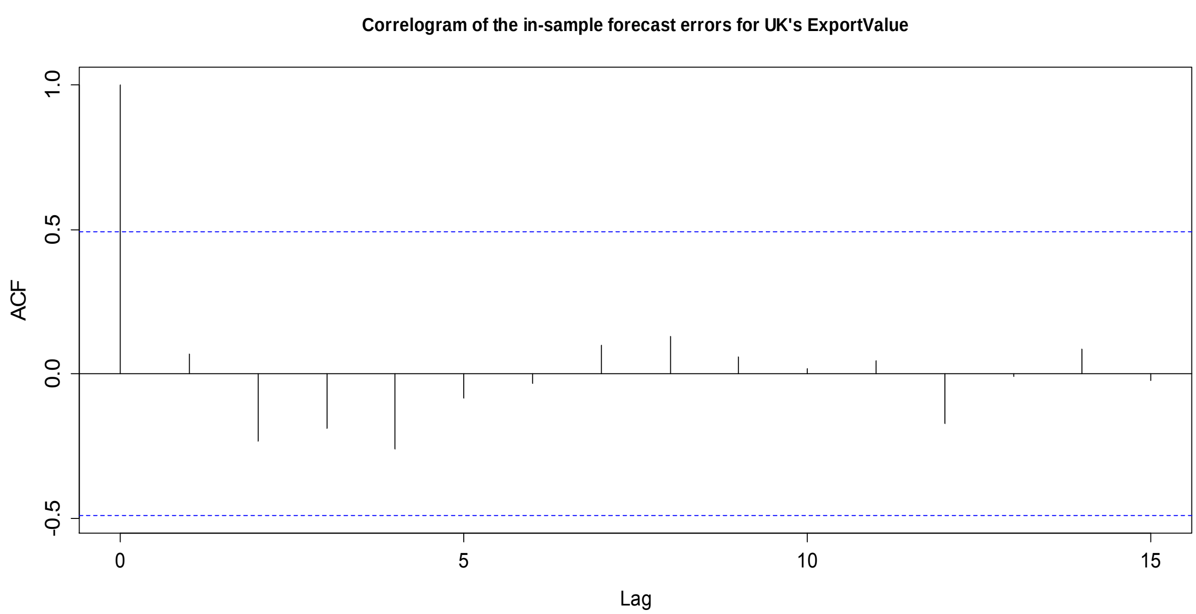


[Figure C.2.56] – Analyses for Brazil, Export Value and the subsets 2001-2013 and 2006-2013

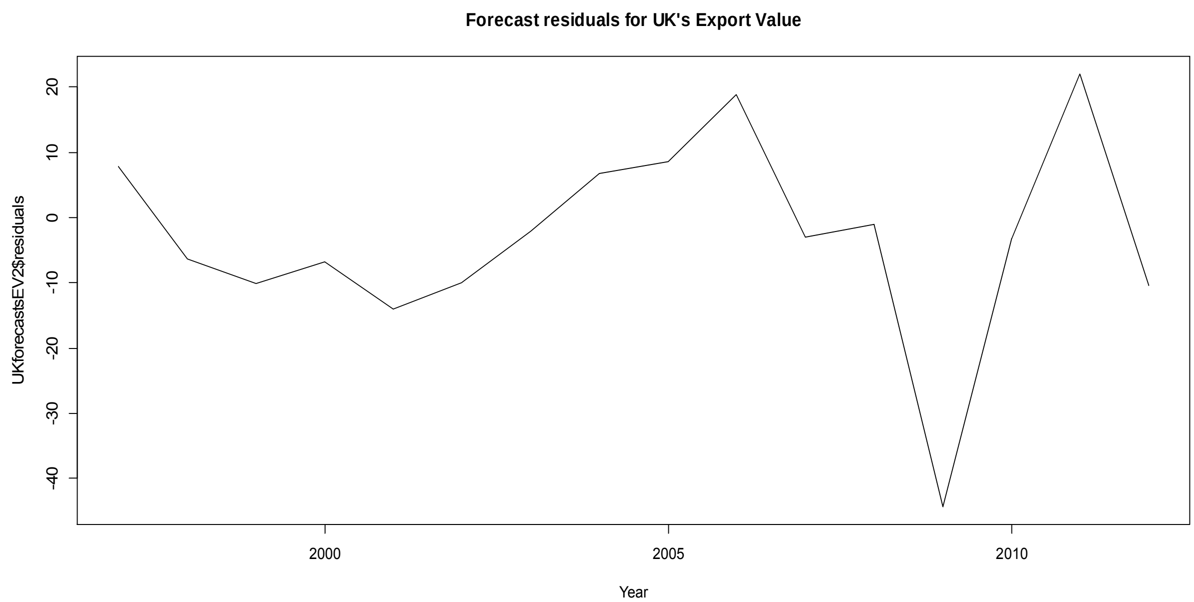
Export Value – UK



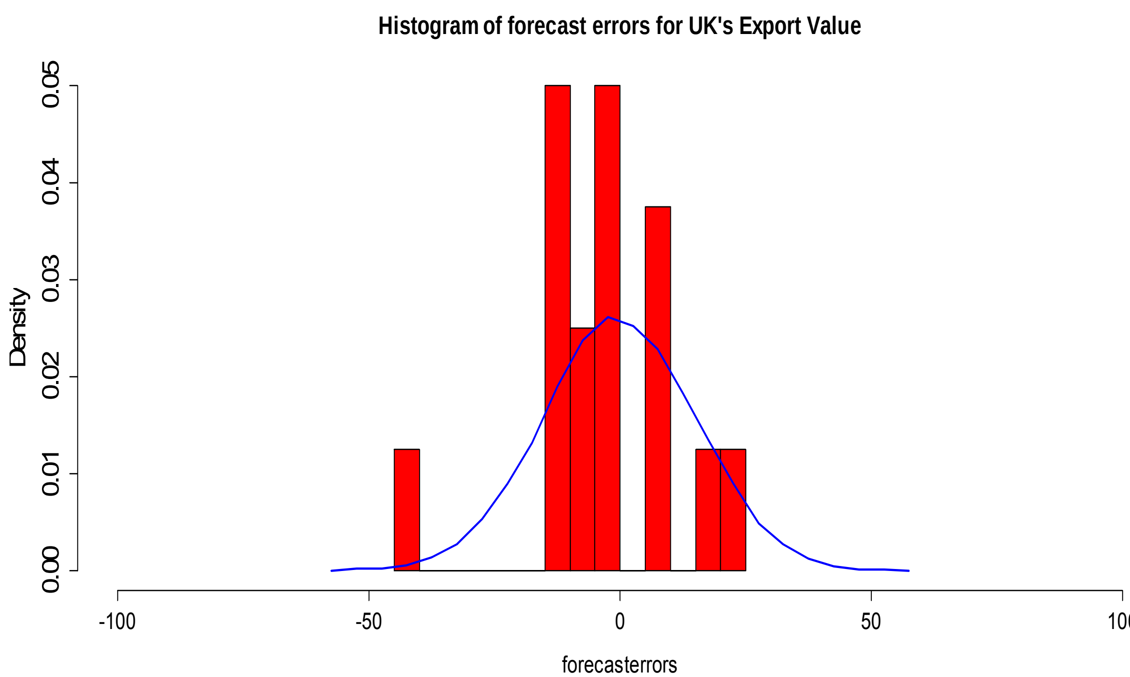
[Figure C.2.57] – Analysis for UK, Export Value and whole dataset



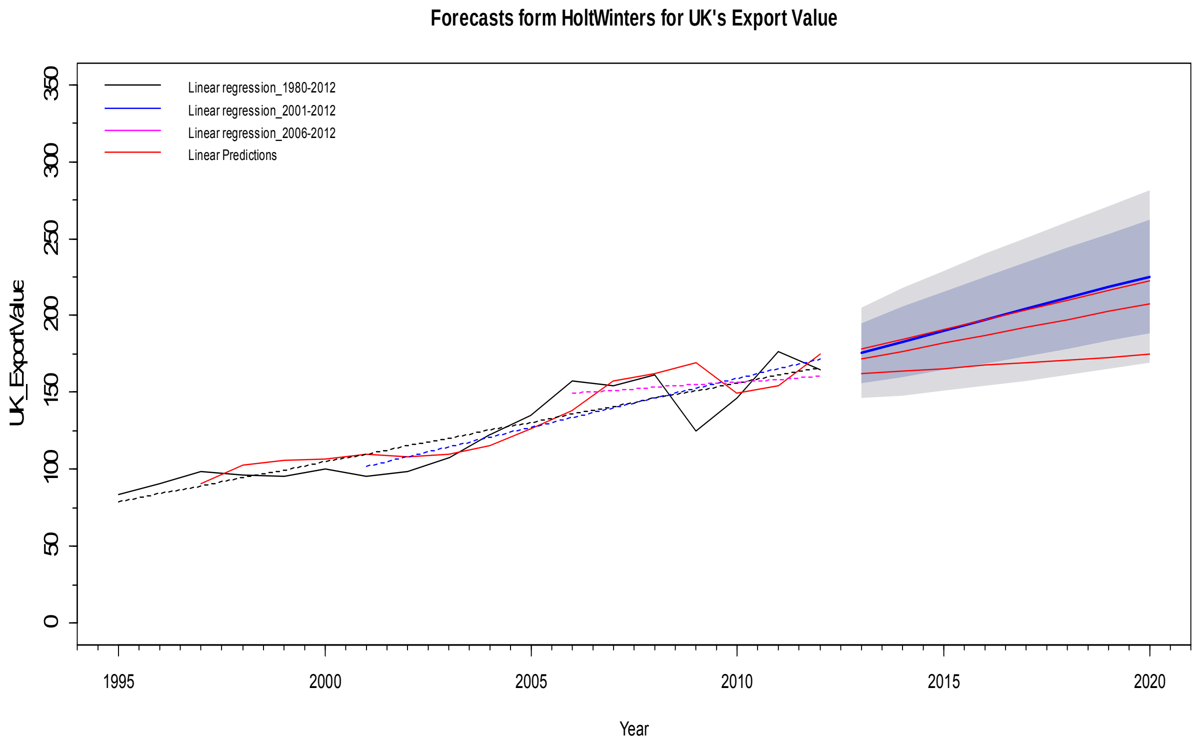
[Figure C.2.58] – Correlogram of in-sample errors



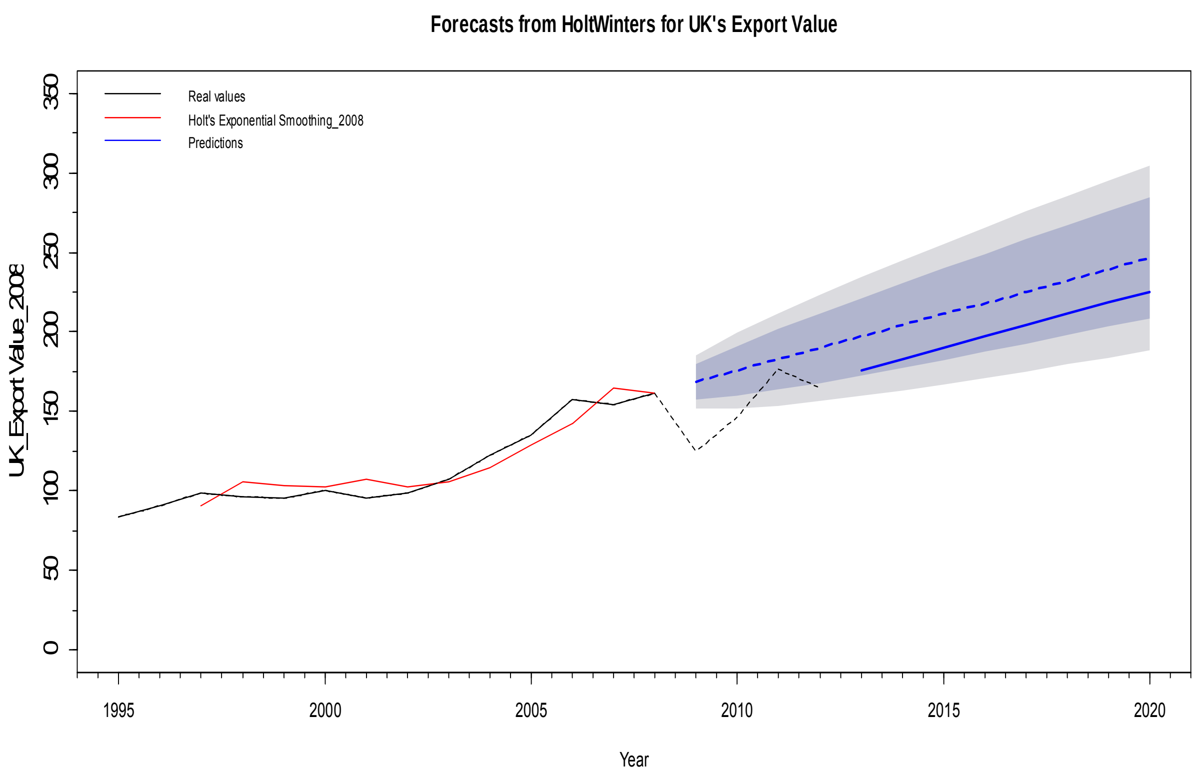
[Figure C.2.59] – Forecast residuals



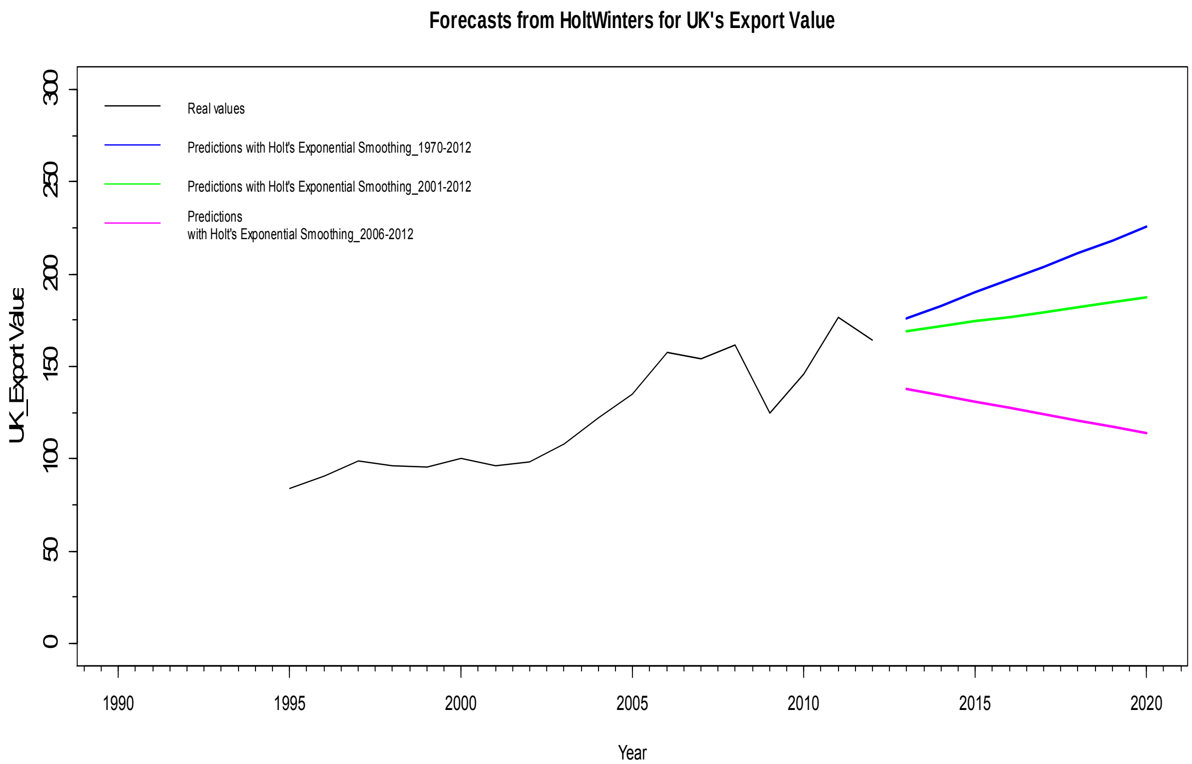
[Figure C.2.60] – Histogram and distribution of forecast residuals



[Figure C.2.61] – Comparison of Linear Regression and Holt's Exponential smoothing

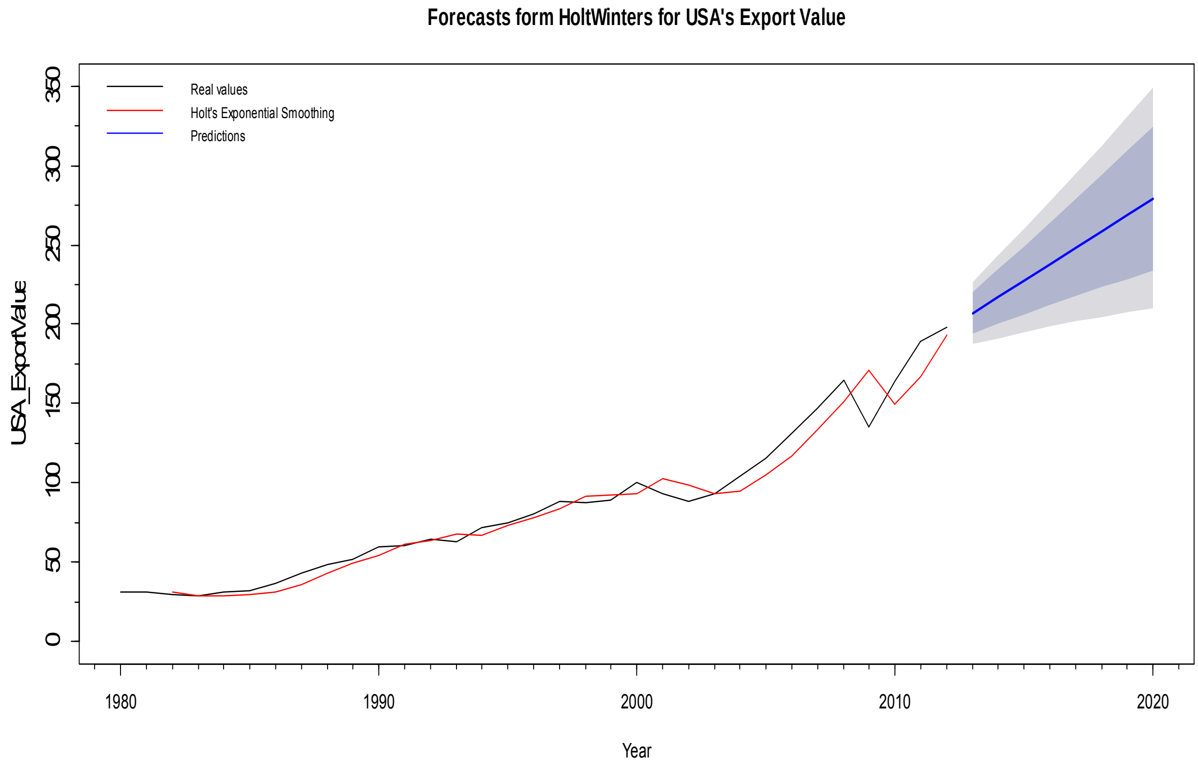


[Figure C.2.62] – Analysis for UK, Export Value and the dataset up to 2008

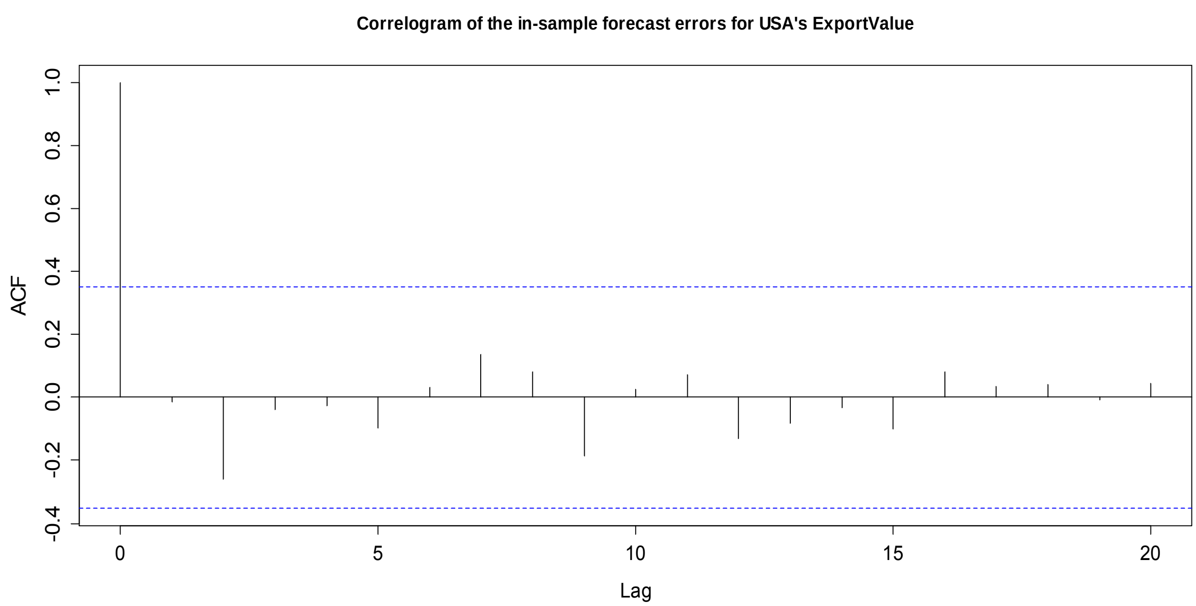


[Figure C.2.63] – Analyses for UK, Export Value and the subsets 2001-2013 and 2006-2013

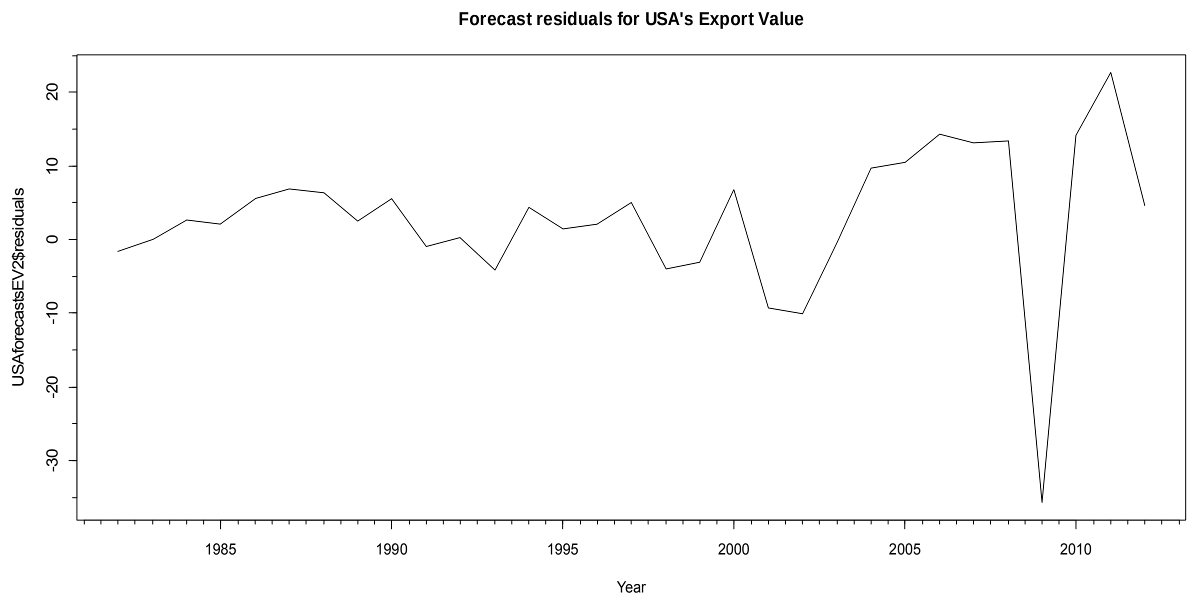
Export Value – USA



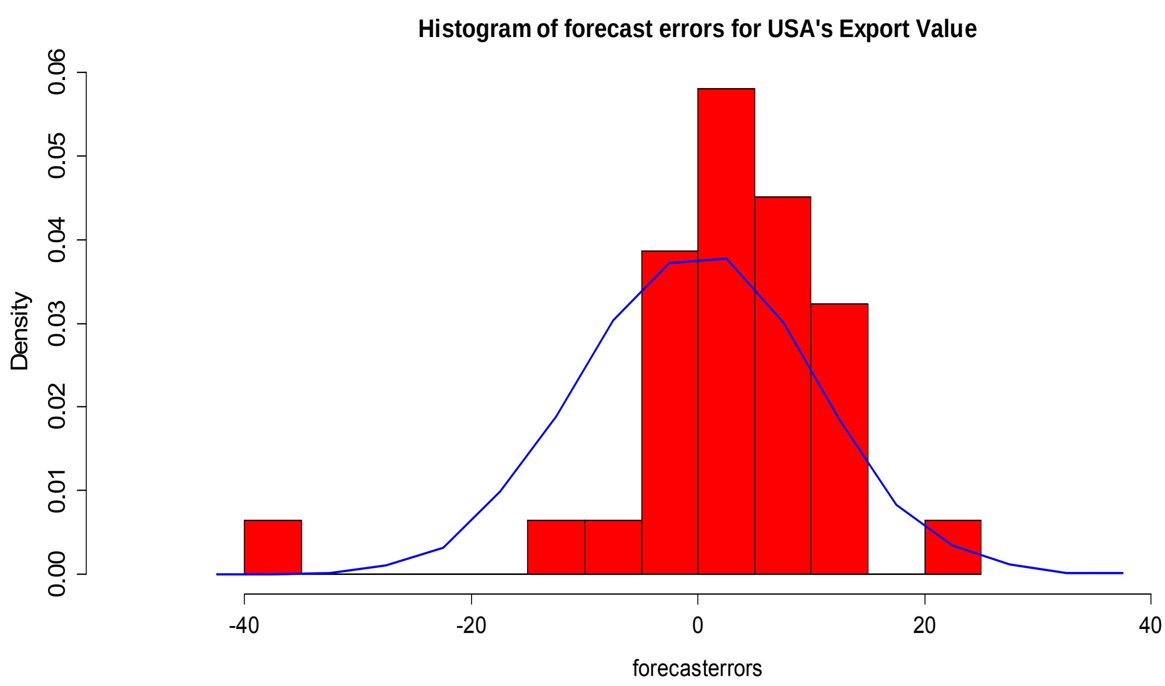
[Figure C.2.64] – Analysis for USA, Export Value and whole dataset



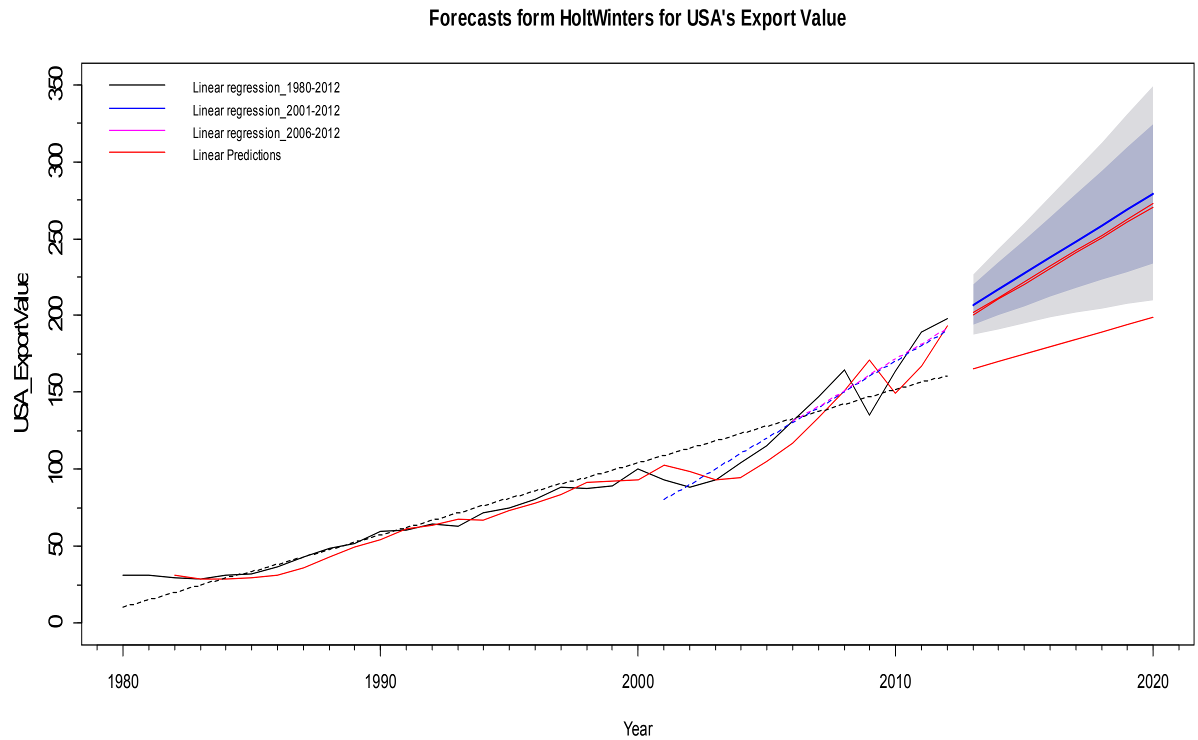
[Figure C.2.65] – Correlogram of in-sample errors



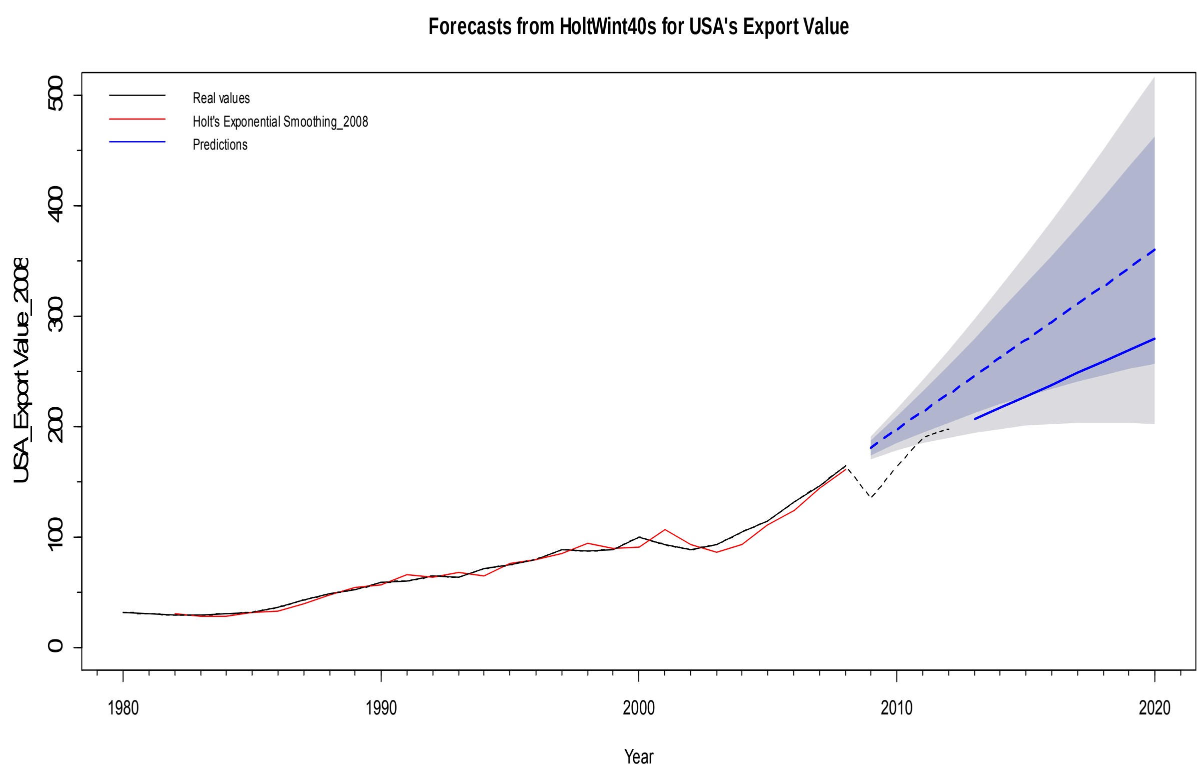
[Figure C.2.66] – Forecast residuals



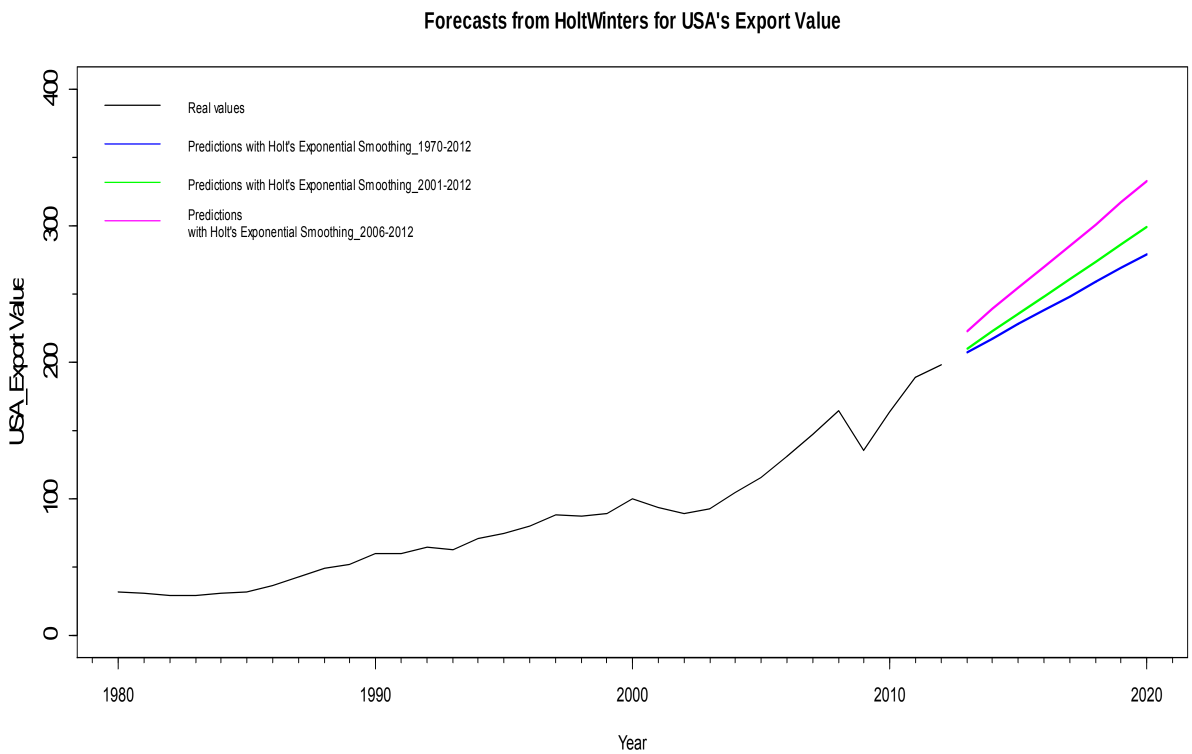
[Figure C.2.67] – Histogram and distribution of forecast residuals



[Figure C.2.68] – Comparison of Linear Regression and Holt's Exponential smoothing

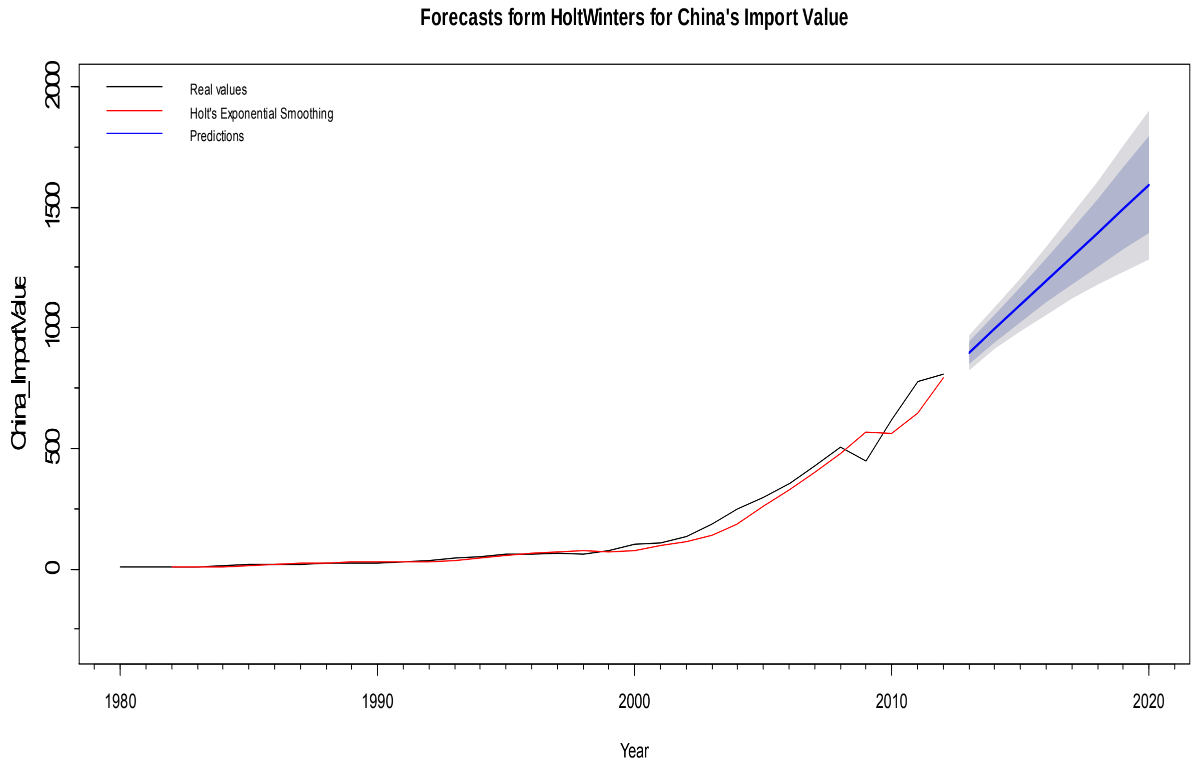


[Figure C.2.69] – Analysis for USA, Export Value and the dataset up to 2008

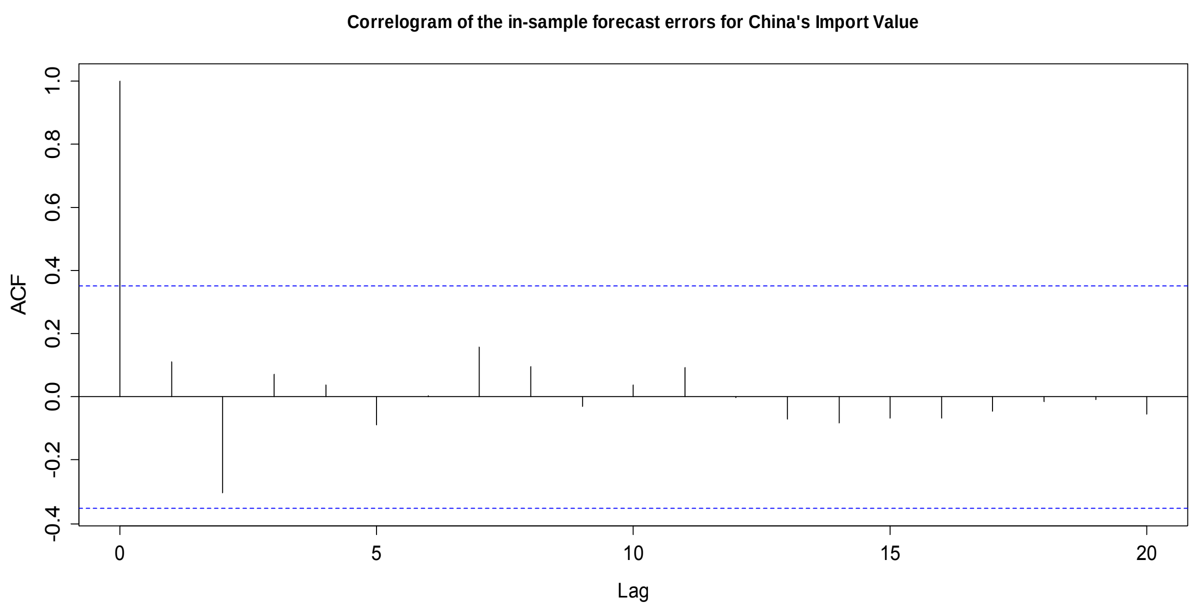


[Figure C.2.70] – Analyses for USA, Export Value and the subsets 2001-2013 and 2006-2013

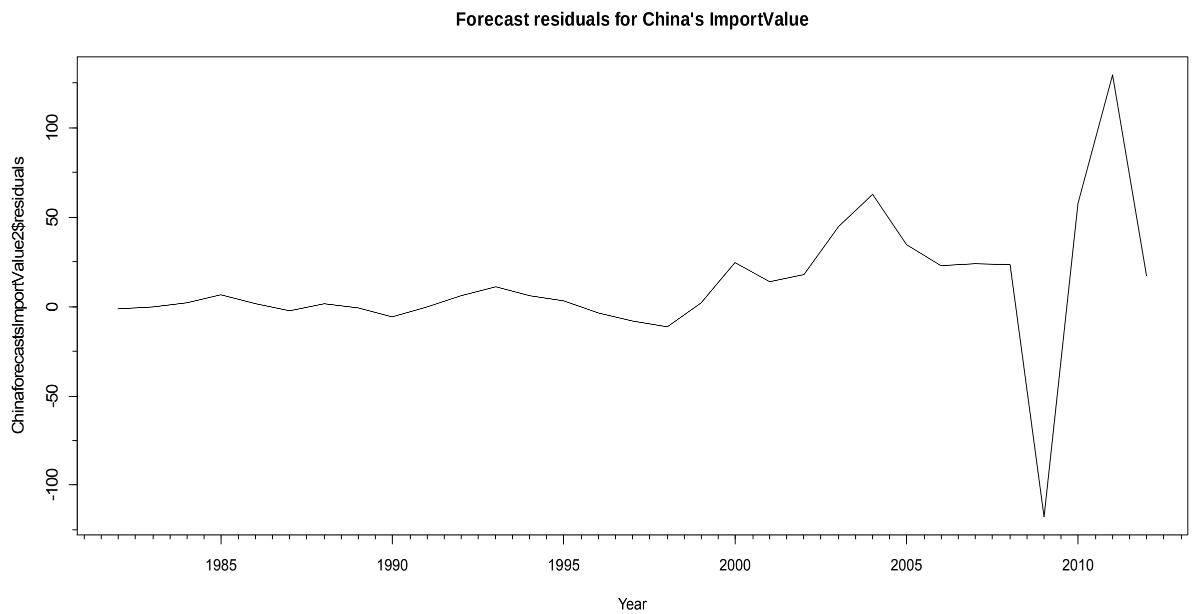
Import Value – China



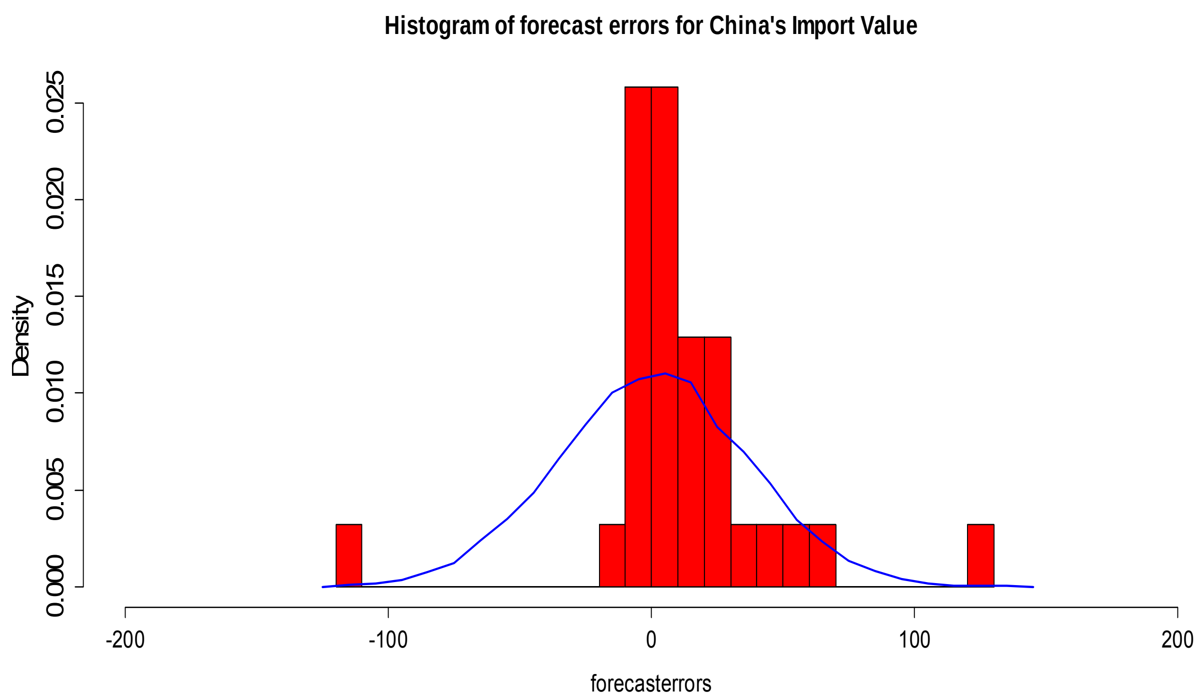
[Figure C.2.71] – Analysis for China, Import Value and whole dataset



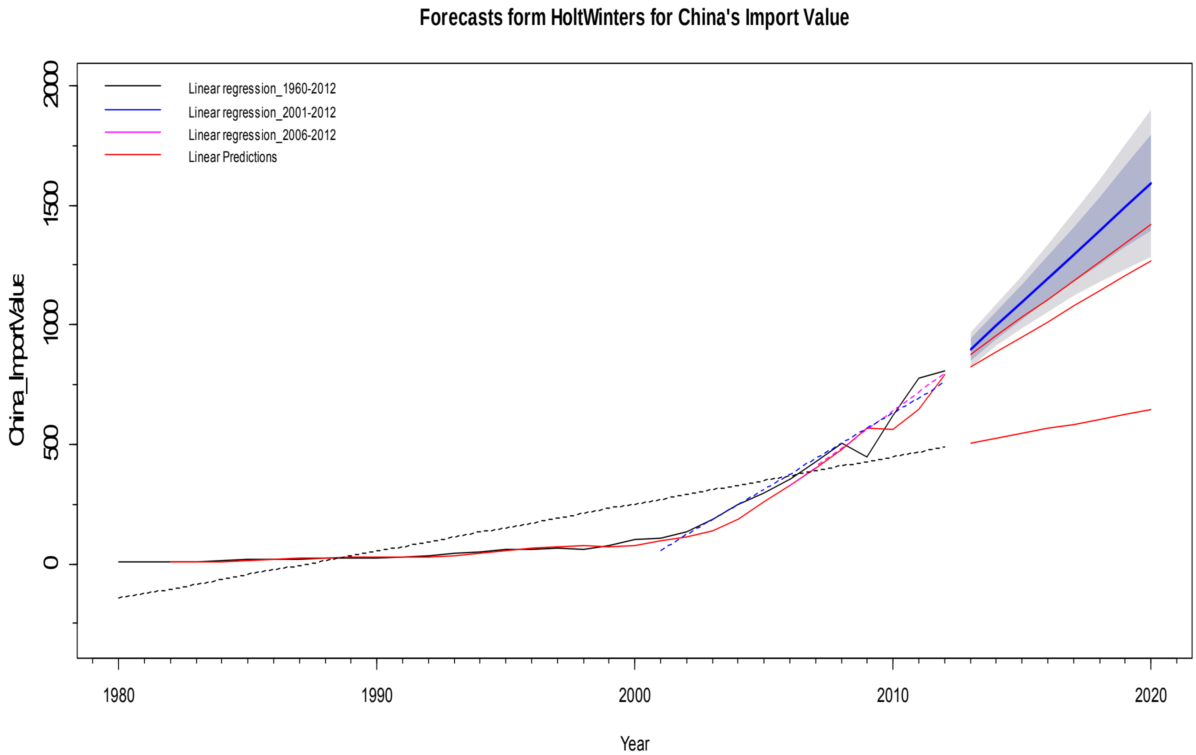
[Figure C.2.72] – Correlogram of in-sample errors



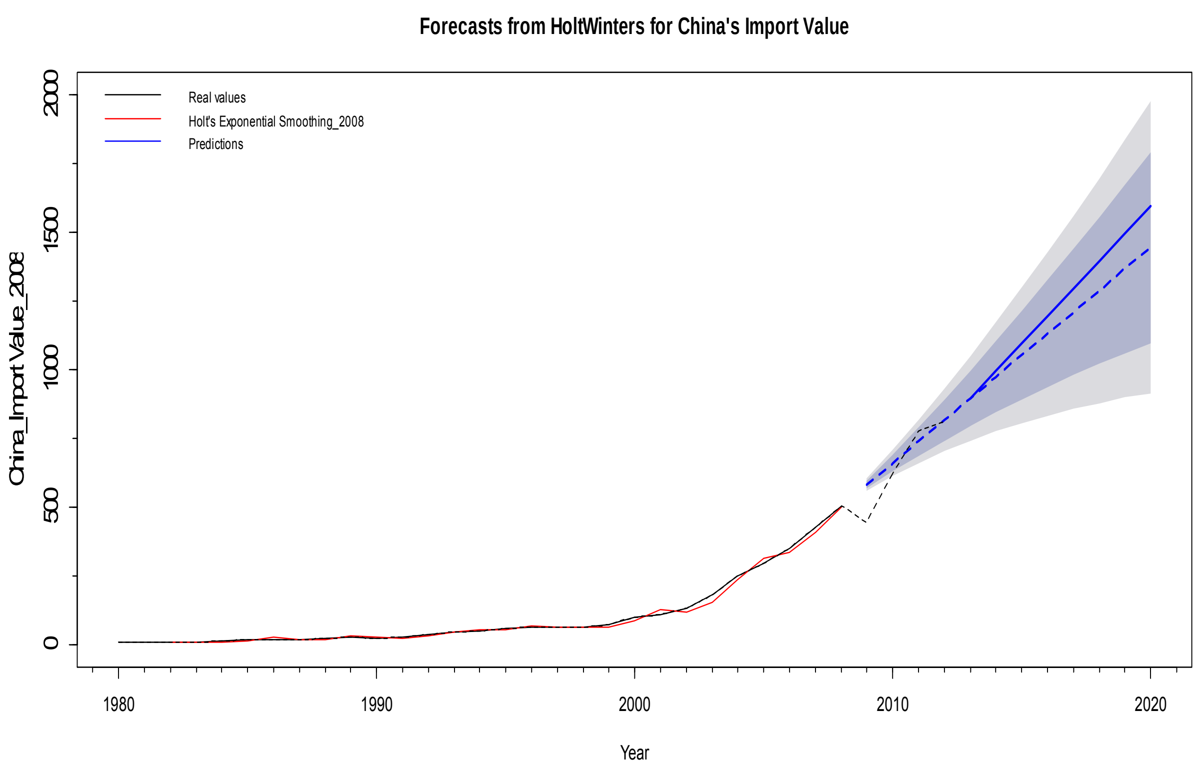
[Figure C.2.73] – Forecast residuals



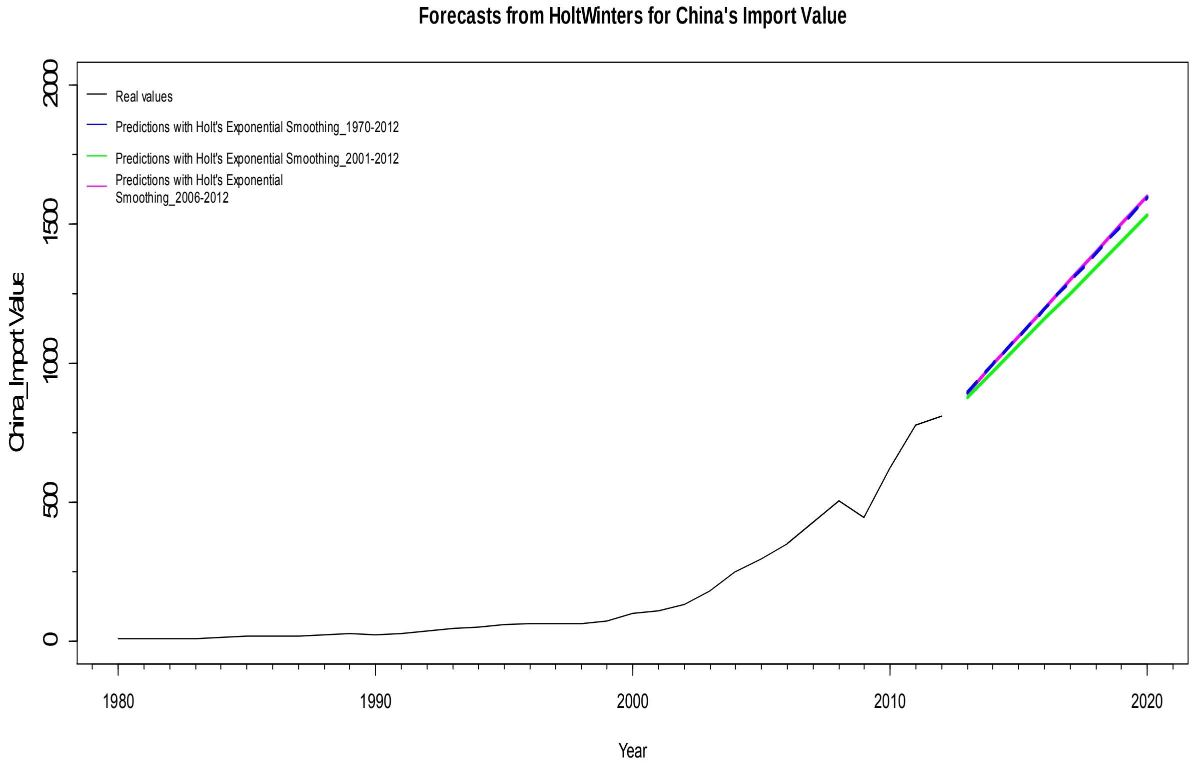
[Figure C.2.74] – Histogram and distribution of forecast residuals



[Figure C.2.75] – Comparison of Linear Regression and Holt’s Exponential smoothing

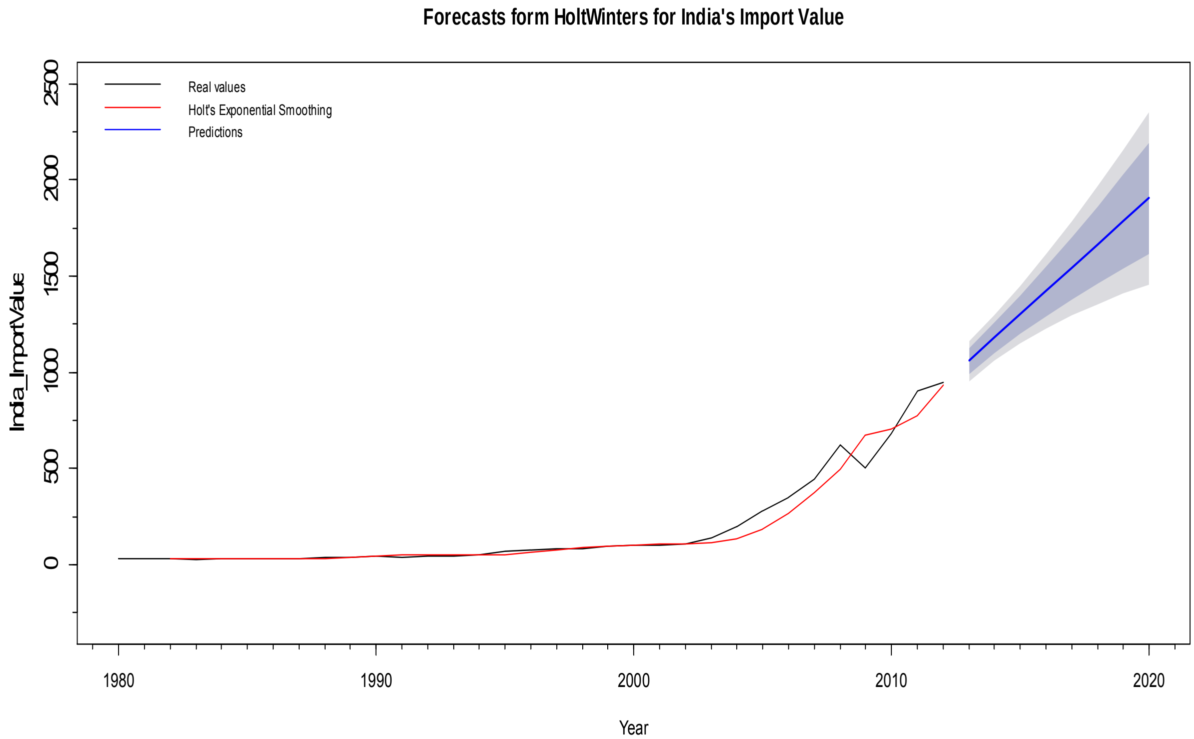


[Figure C.2.76] – Analysis for China, Import Value and the dataset up to 2008

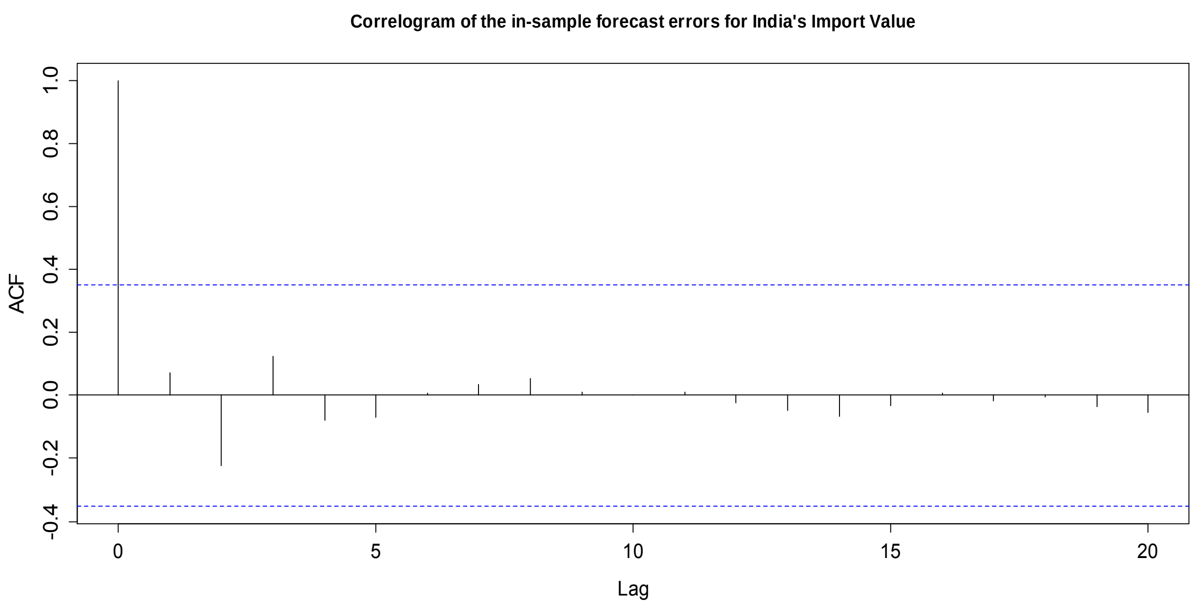


[Figure C.2.77] – Analyses for China, Import Value and the subsets 2001-2013 and 2006-2013

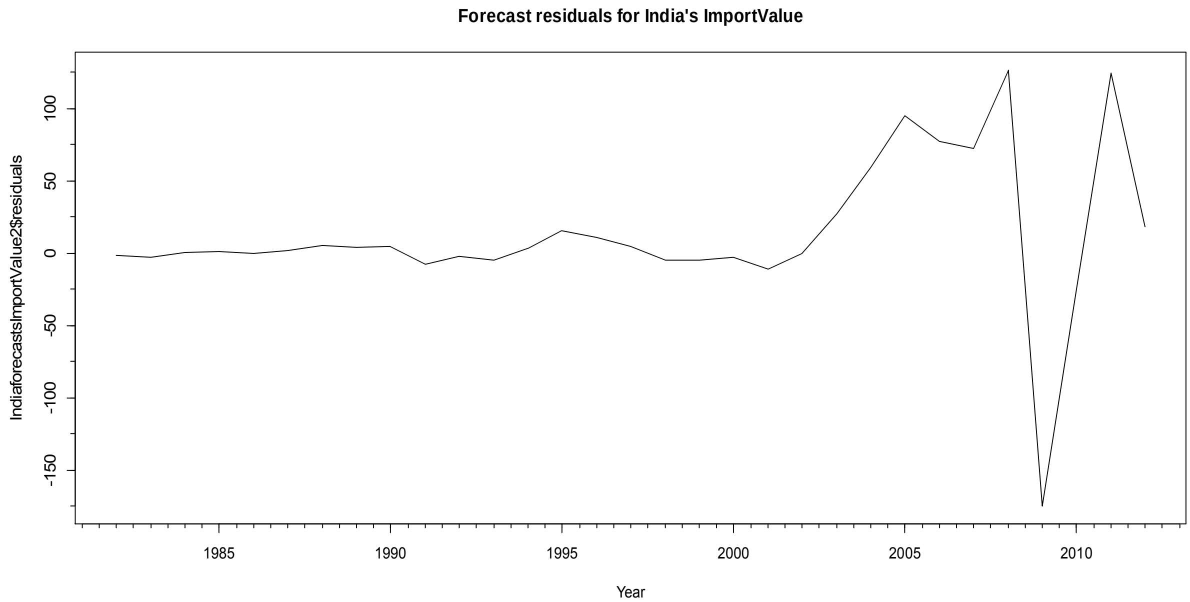
Import Value – India



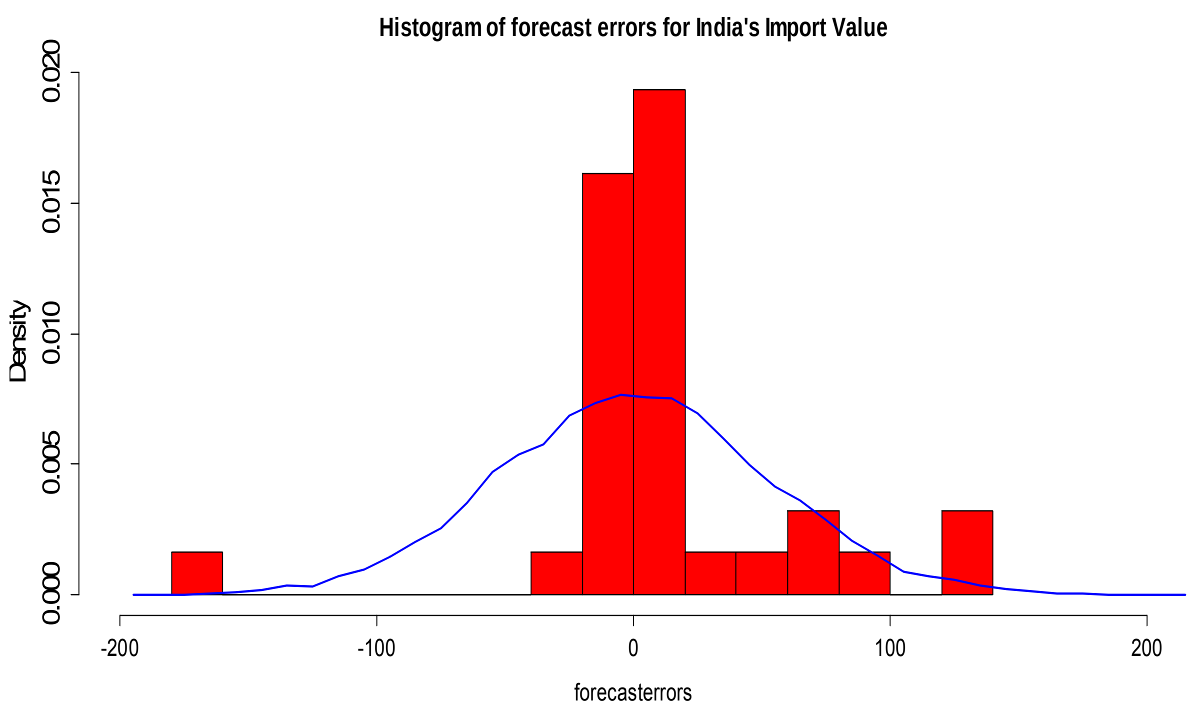
[Figure C.2.78] – Analysis for India, Import Value and whole dataset



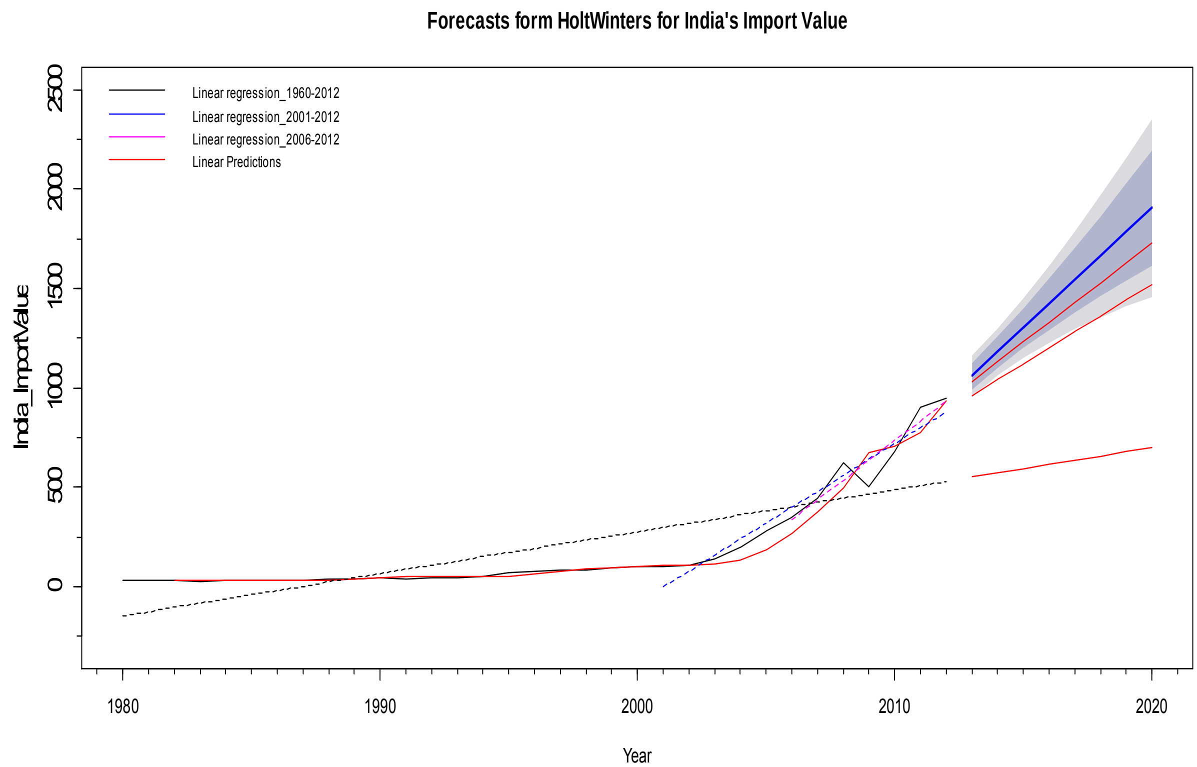
[Figure C.2.79] – Correlogram of in-sample errors



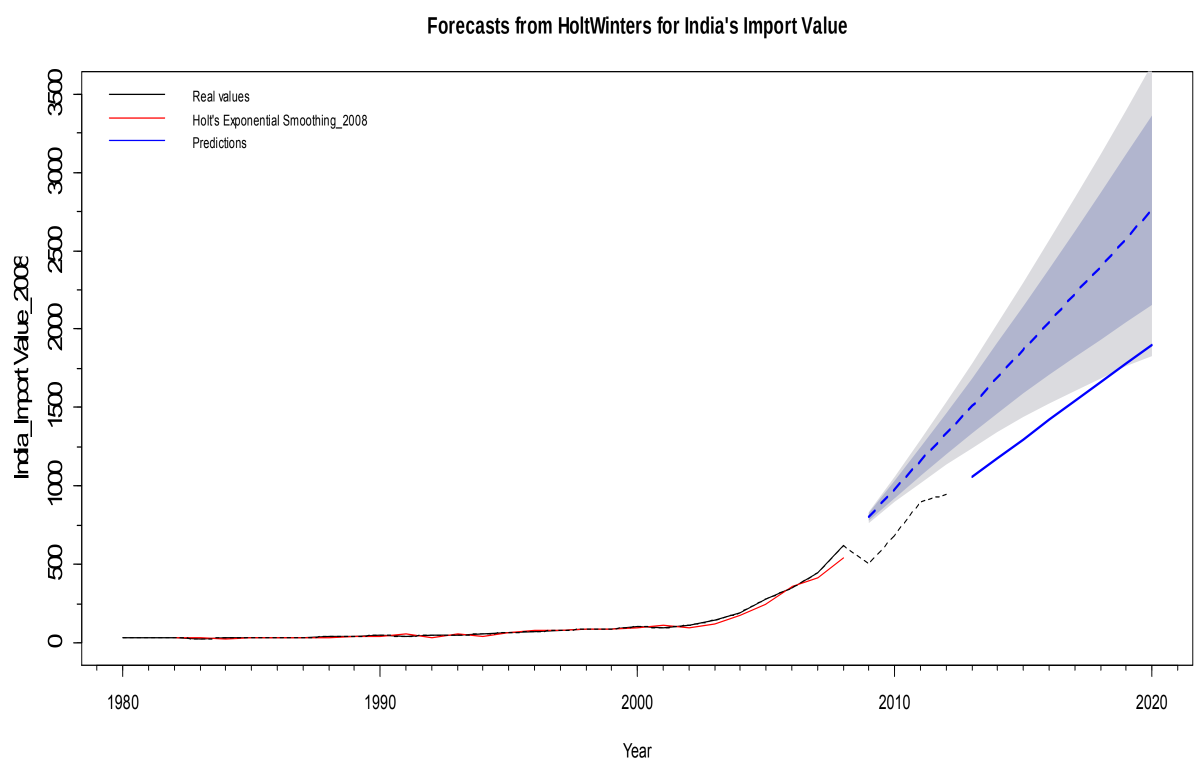
[Figure C.2.80] – Forecast residuals



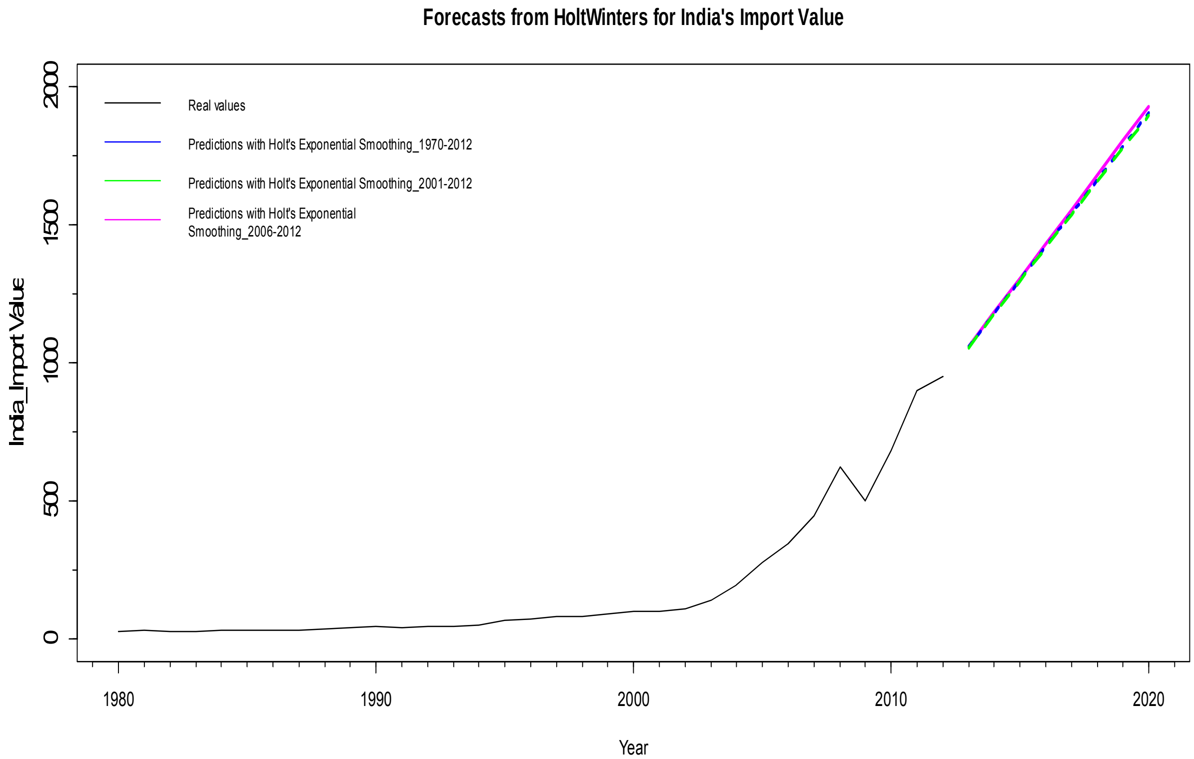
[Figure C.2.81] – Histogram and distribution of forecast residuals



[Figure C.2.82] – Comparison of Linear Regression and Holt’s Exponential smoothing

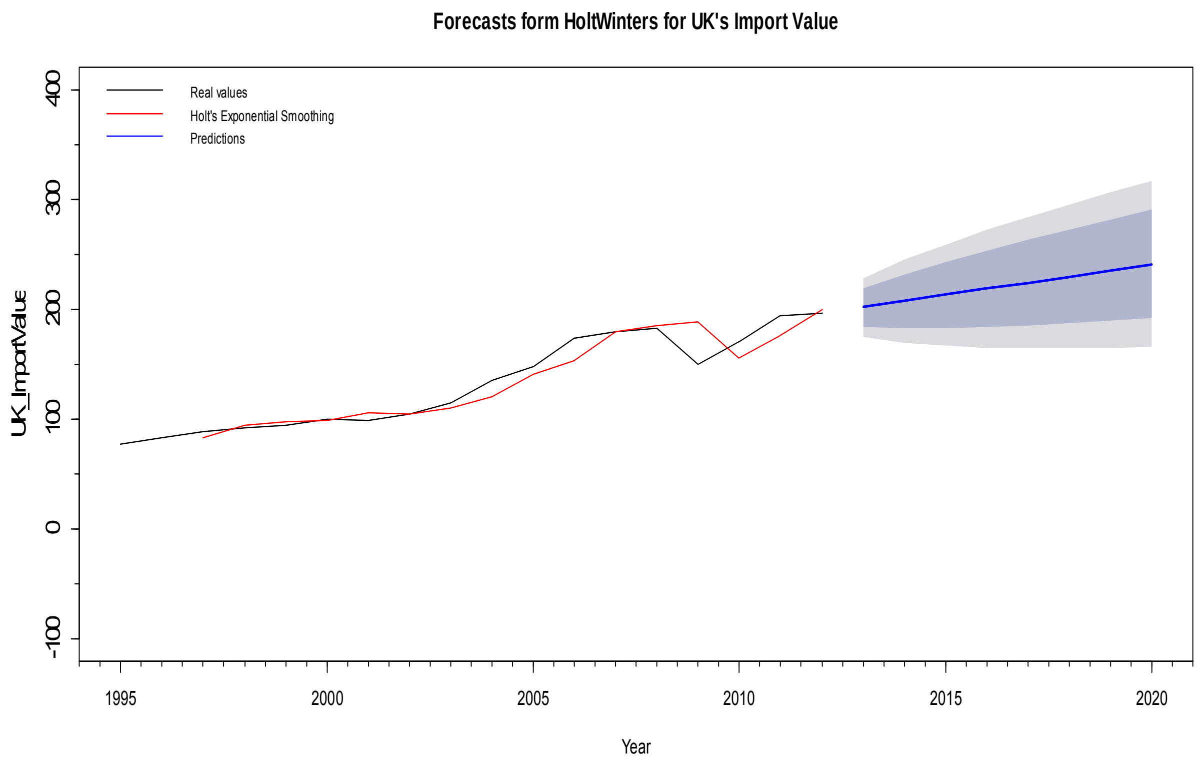


[Figure C.2.83] – Analysis for India, Import Value and the dataset up to 2008

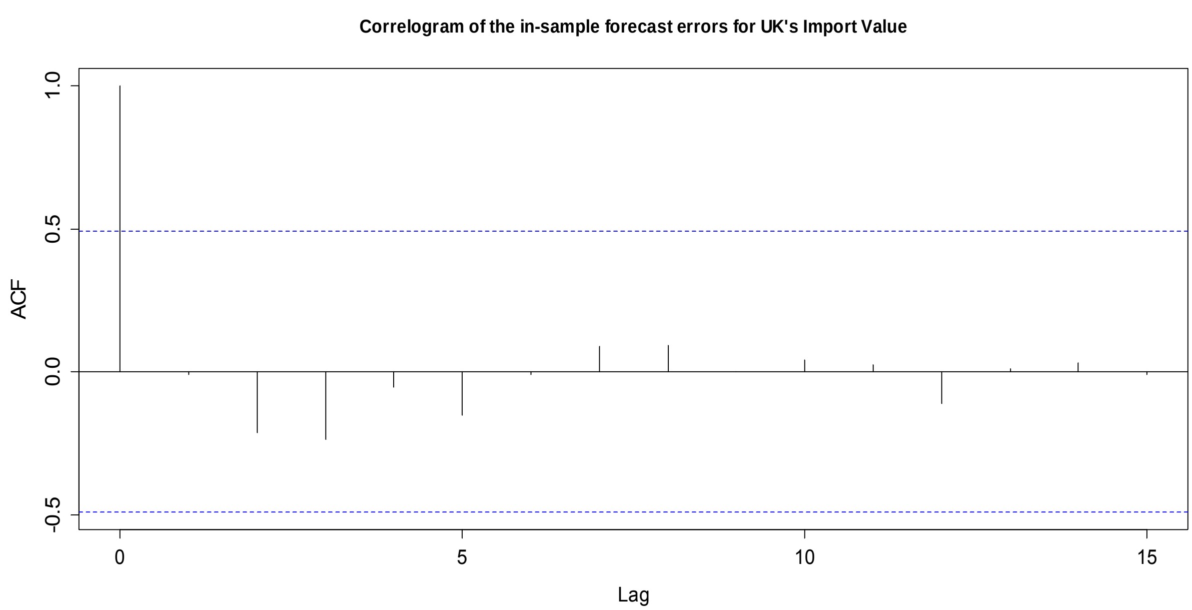


[Figure C.2.84] – Analyses for India, Import Value and the subsets 2001-2013 and 2006-2013

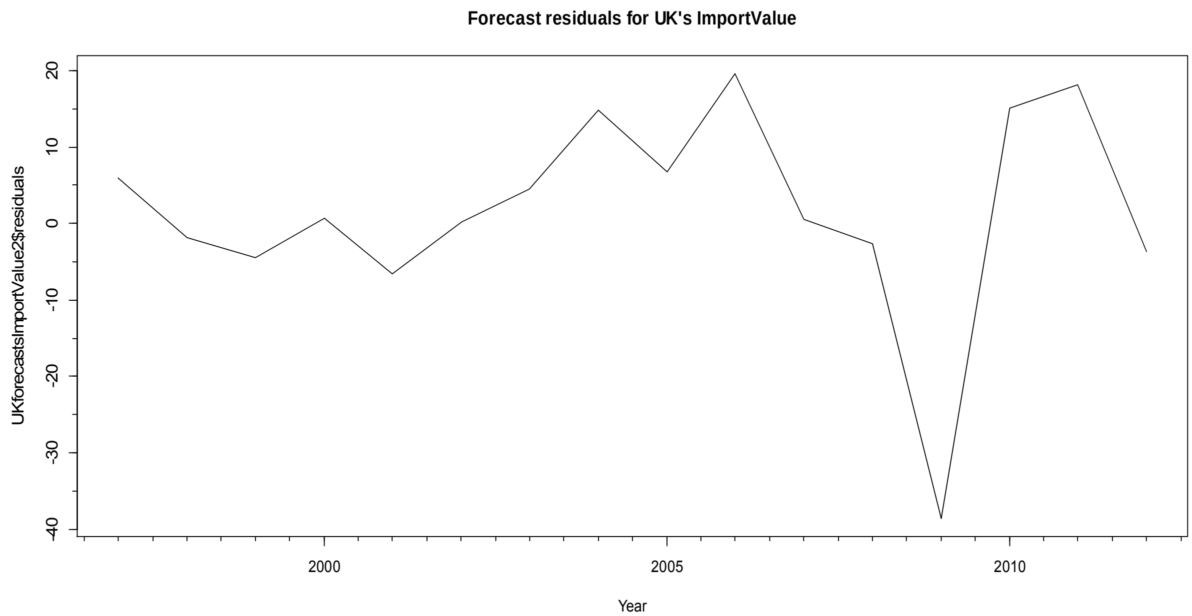
Import Value – UK



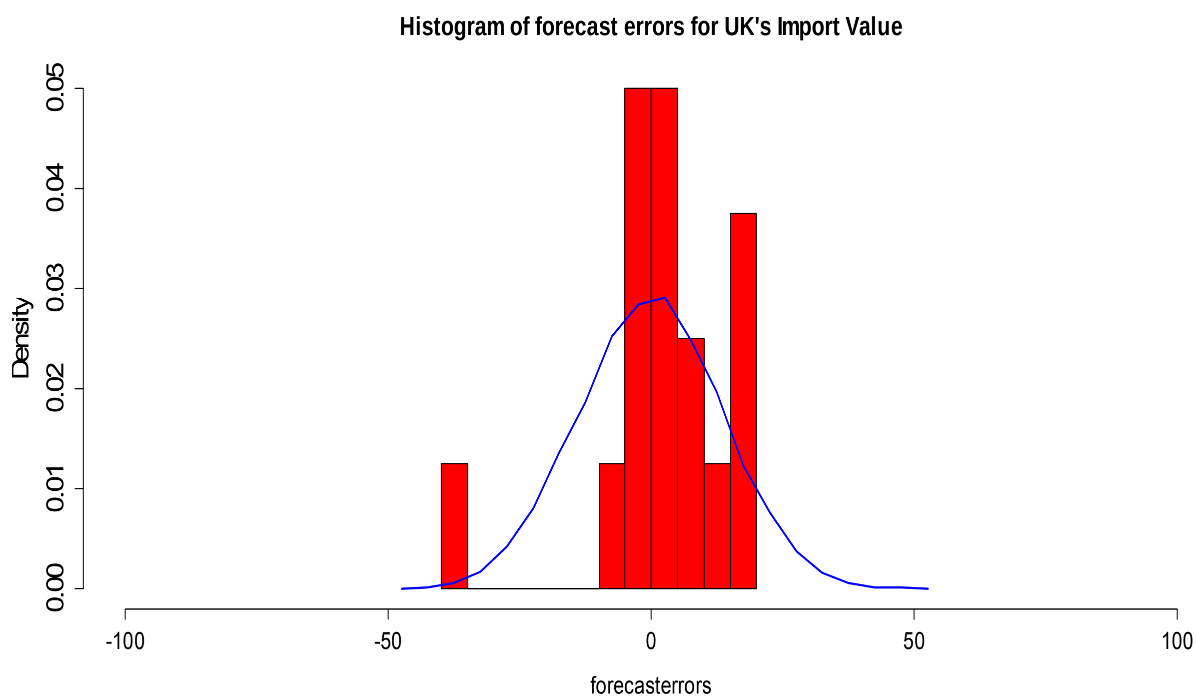
[Figure C.2.85] – Analysis for UK, Import Value and whole dataset



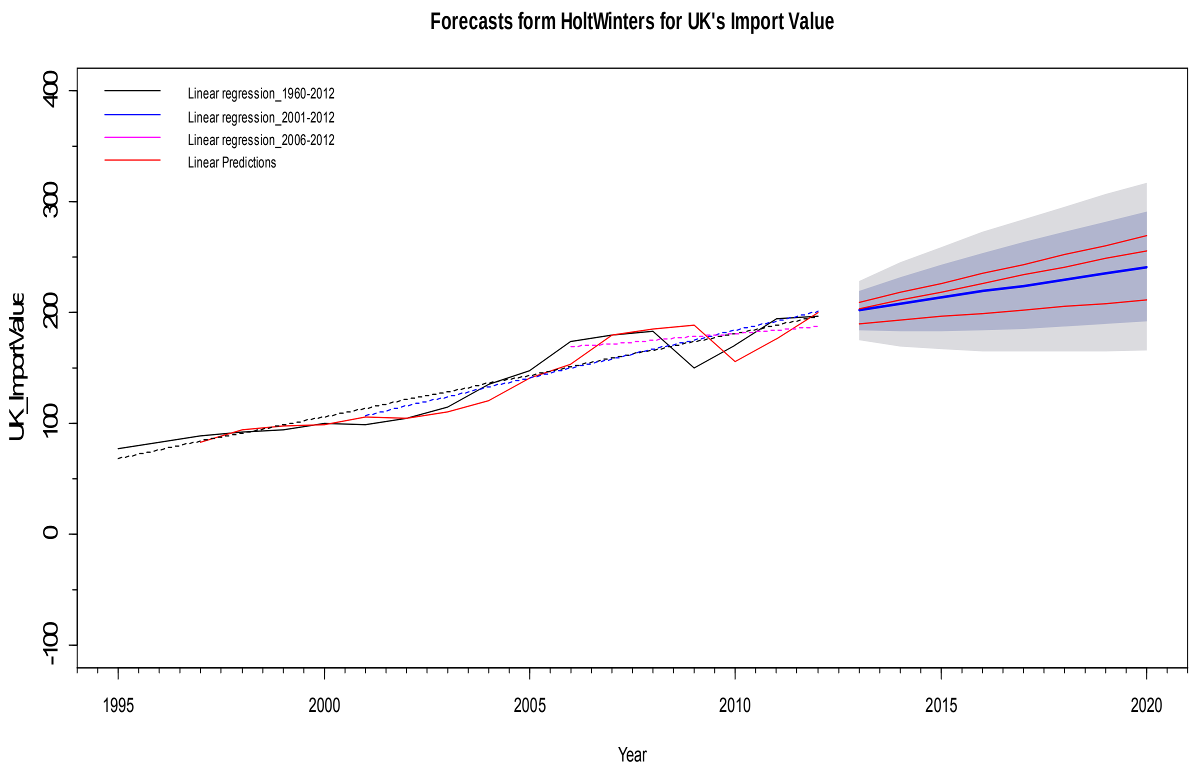
[Figure C.2.86] – Correlogram of in-sample errors



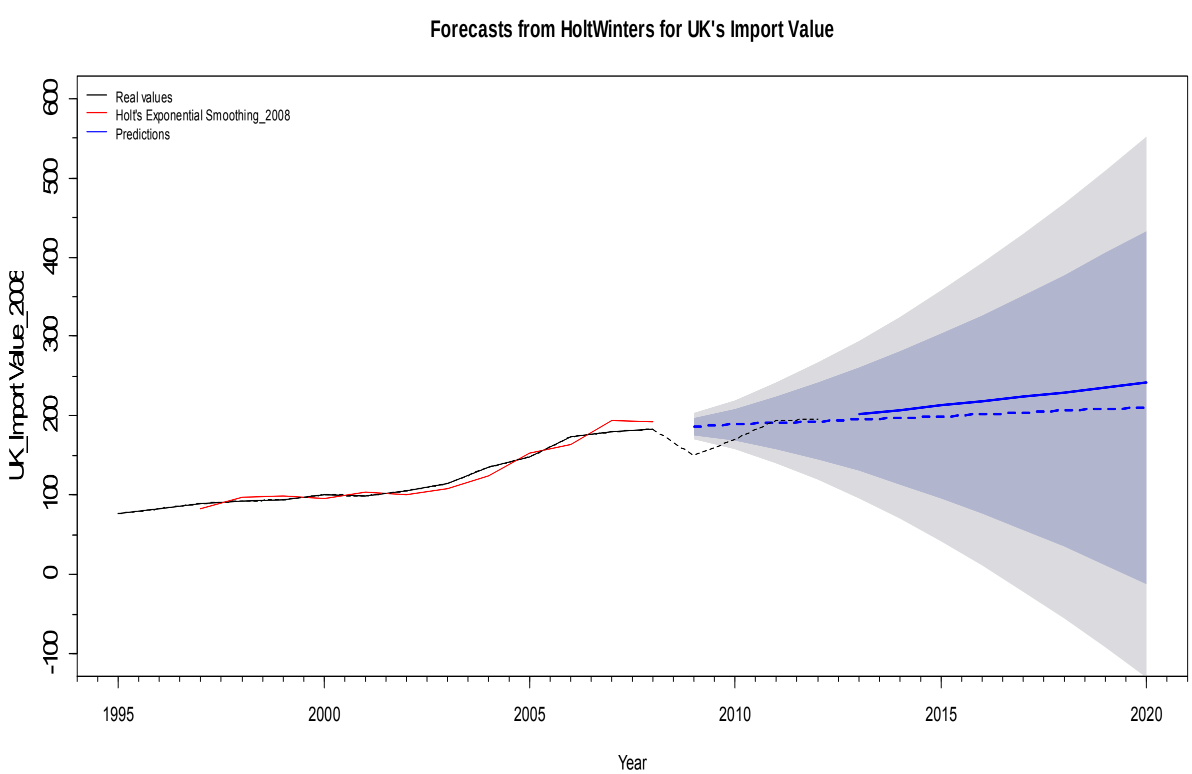
[Figure C.2.87] – Forecast residuals



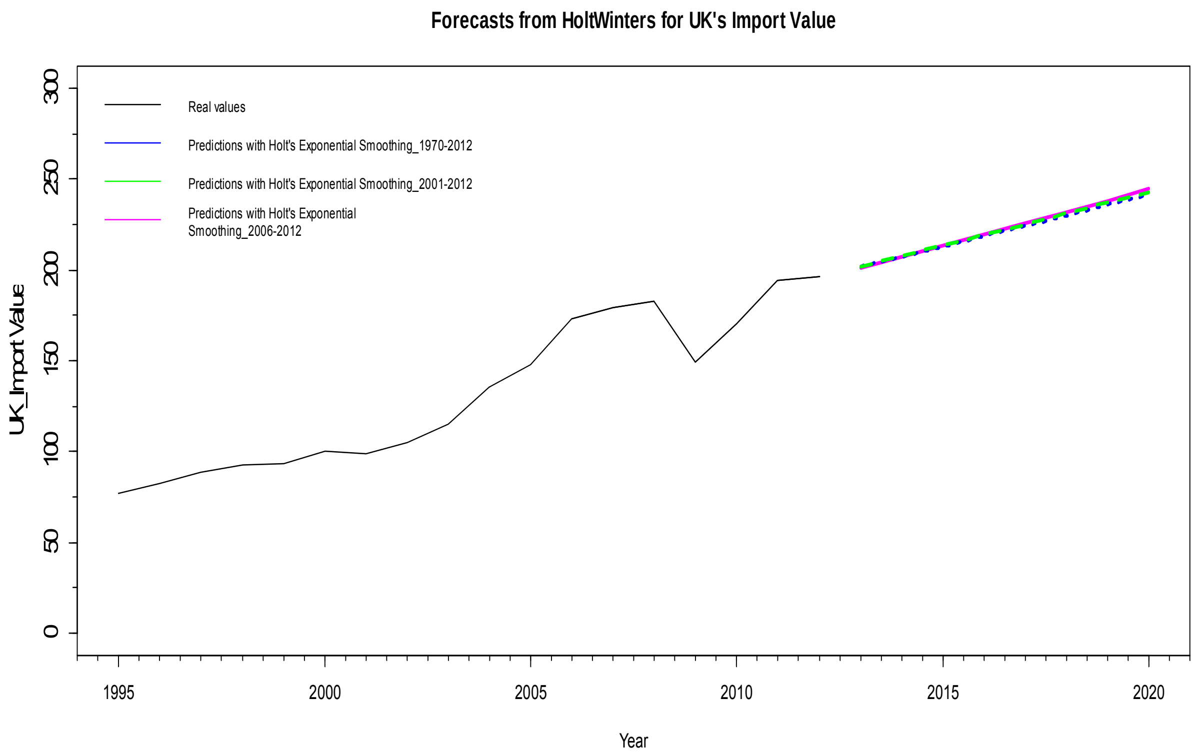
[Figure C.2.88] – Histogram and distribution of forecast residuals



[Figure C.2.89] – Comparison of Linear Regression and Holt's Exponential smoothing

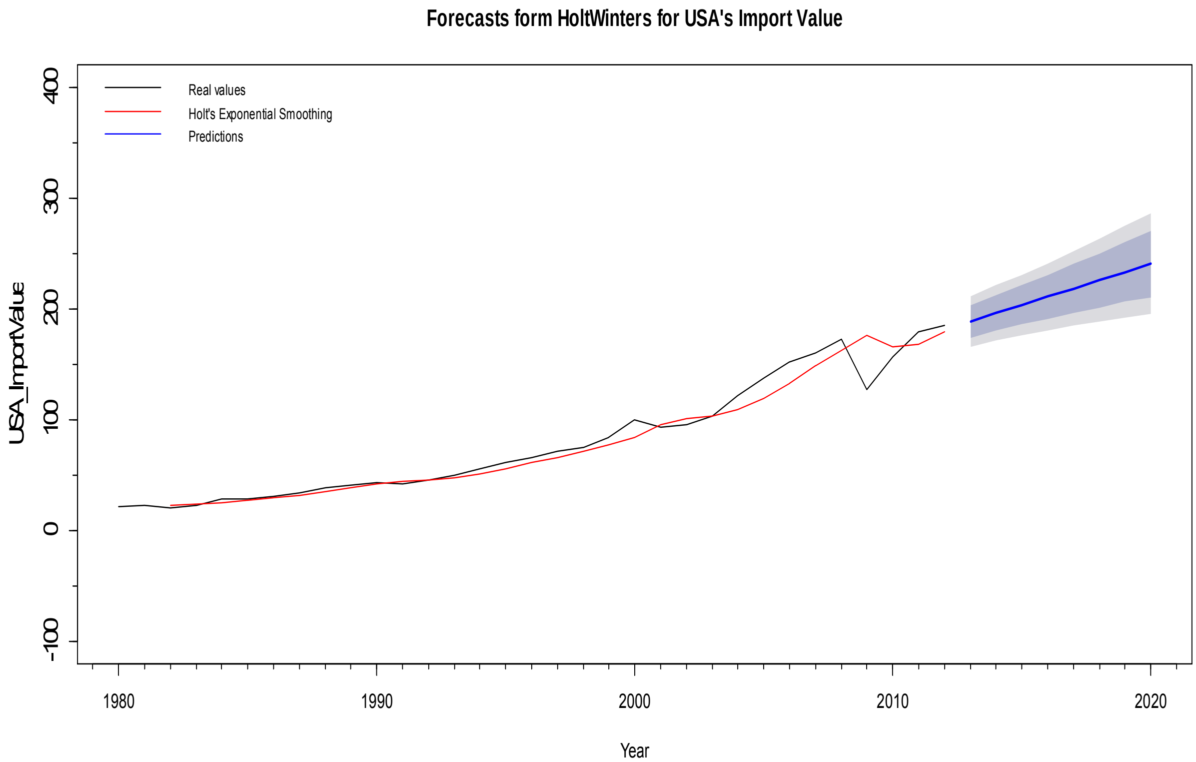


[Figure C.2.90] – Analysis for UK, Import Value and the dataset up to 2008

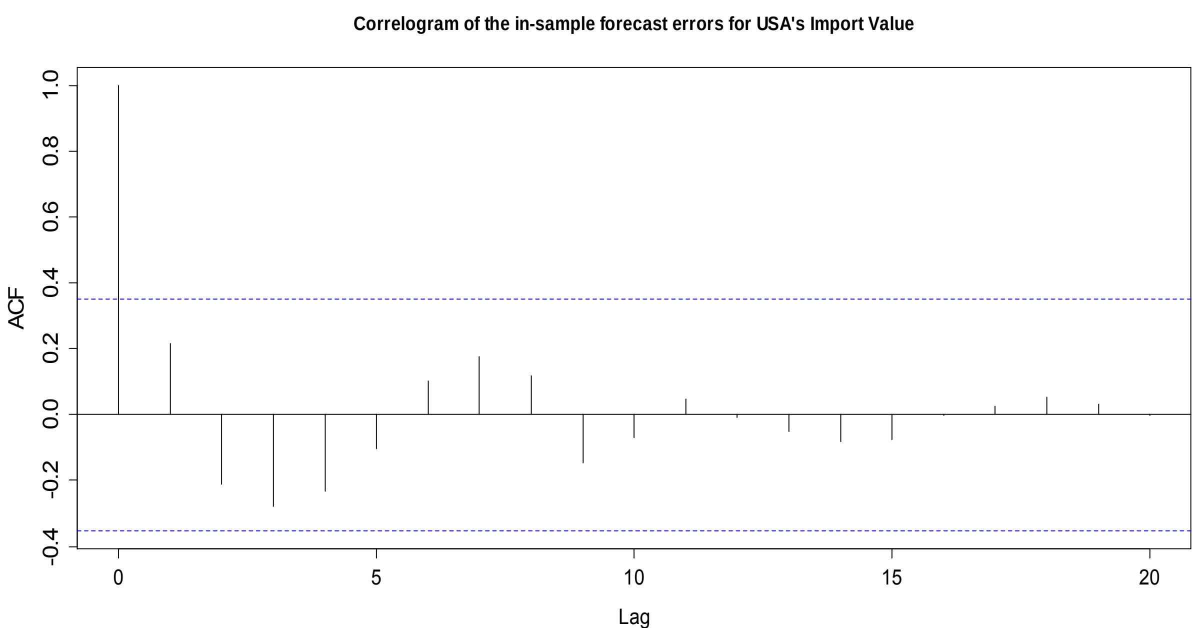


[Figure C.2.91] – Analyses for UK, Import Value and the subsets 2001-2013 and 2006-2013

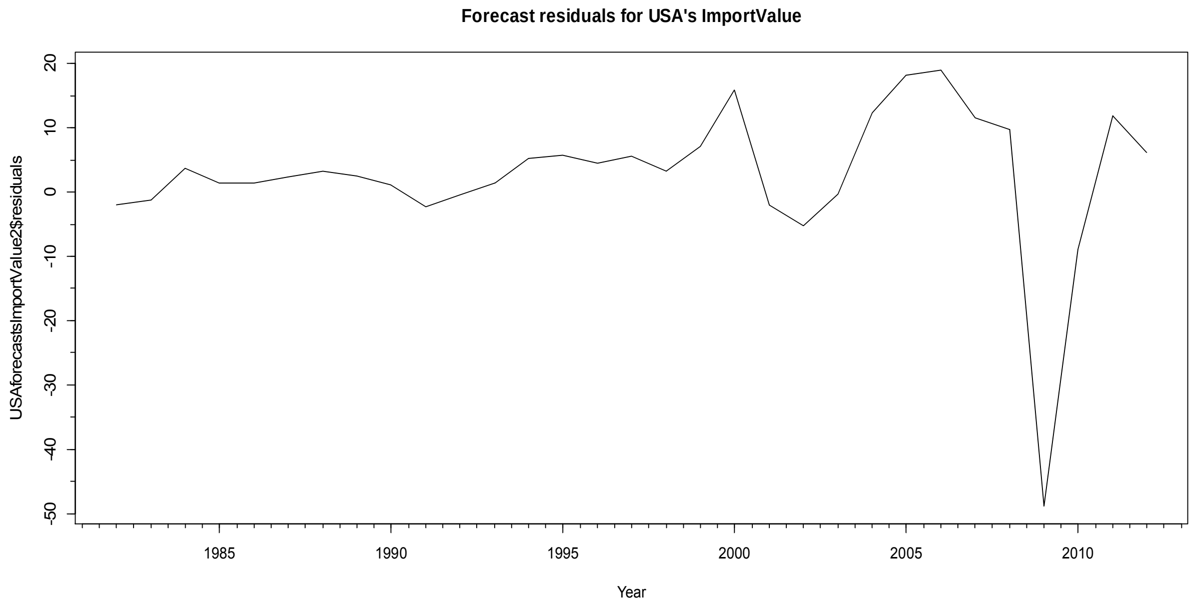
Import Value – USA



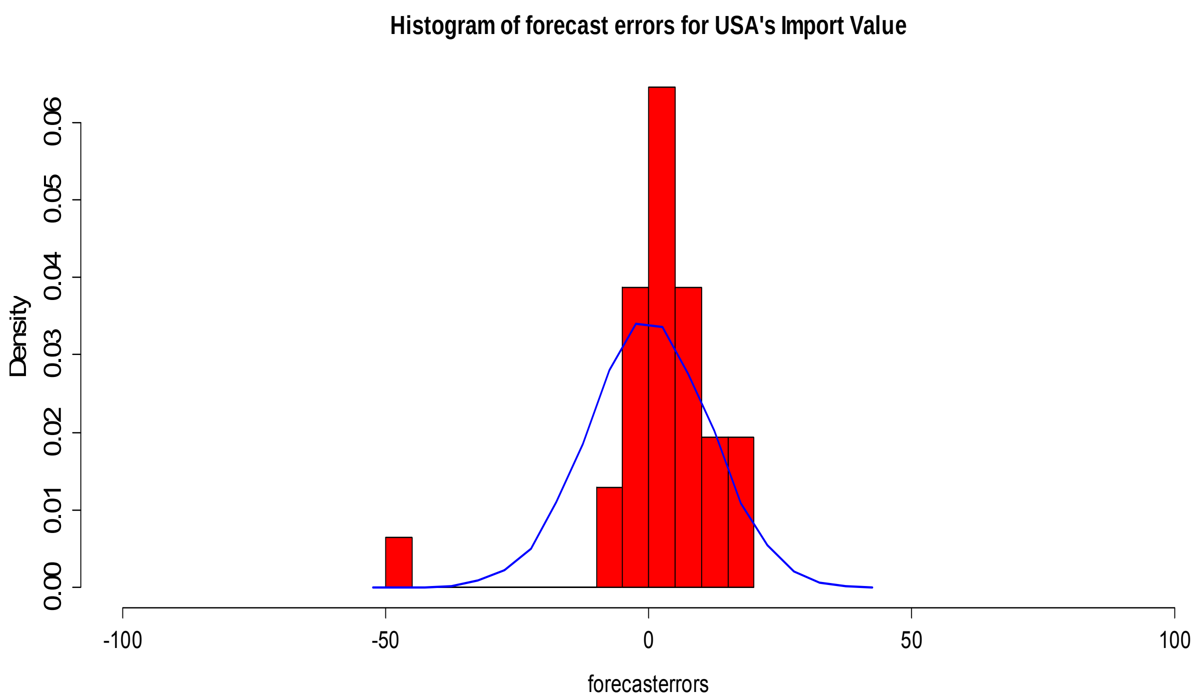
[Figure C.2.92] – Analysis for USA, Import Value and whole dataset



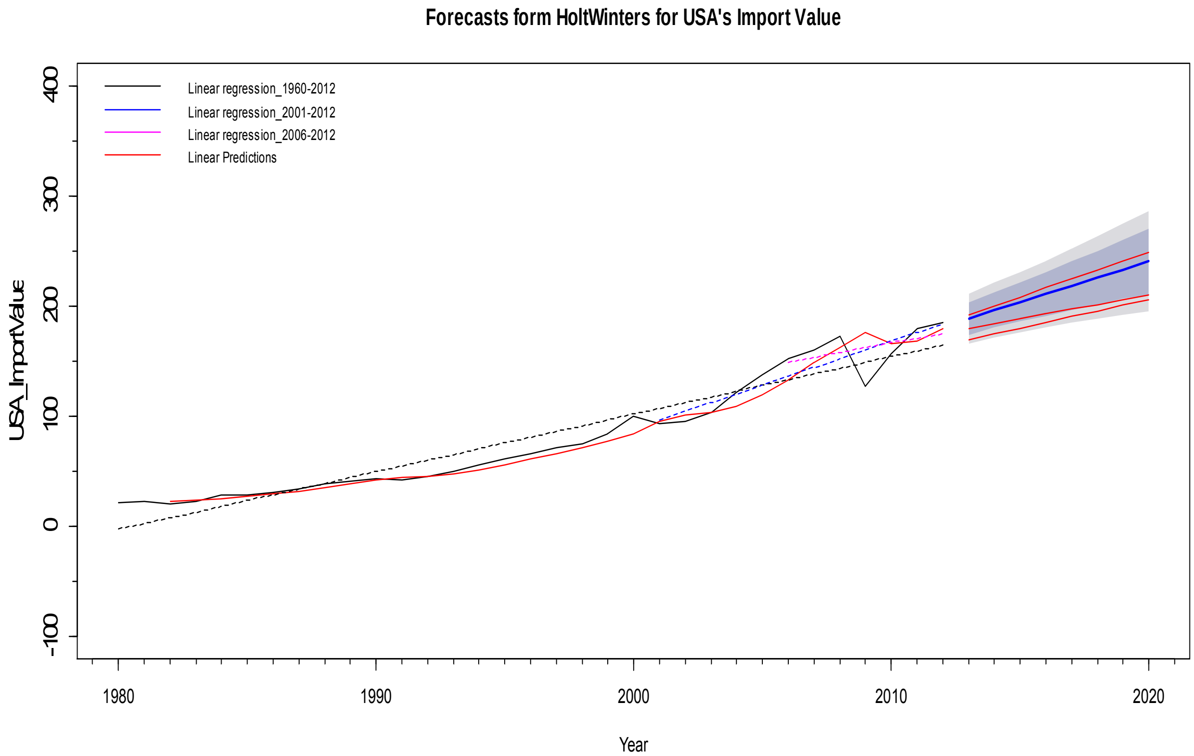
[Figure C.2.93] – Correlogram of in-sample errors



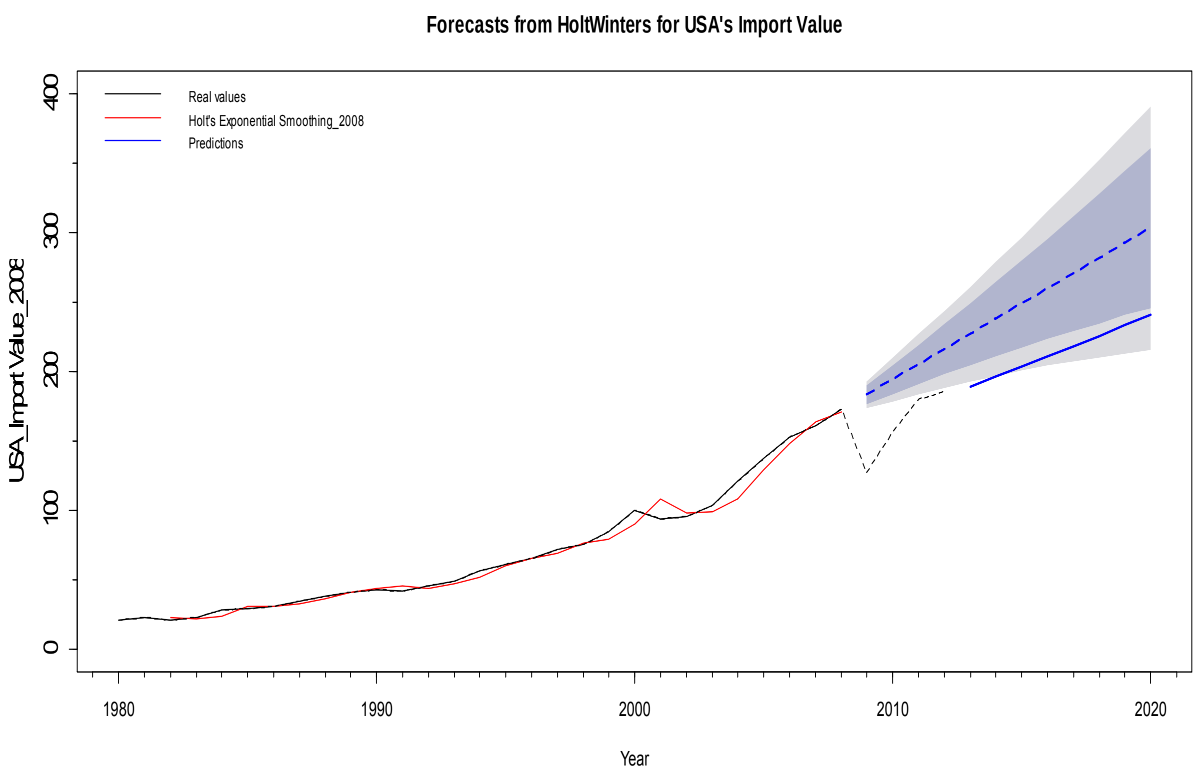
[Figure C.2.94] – Forecast residuals



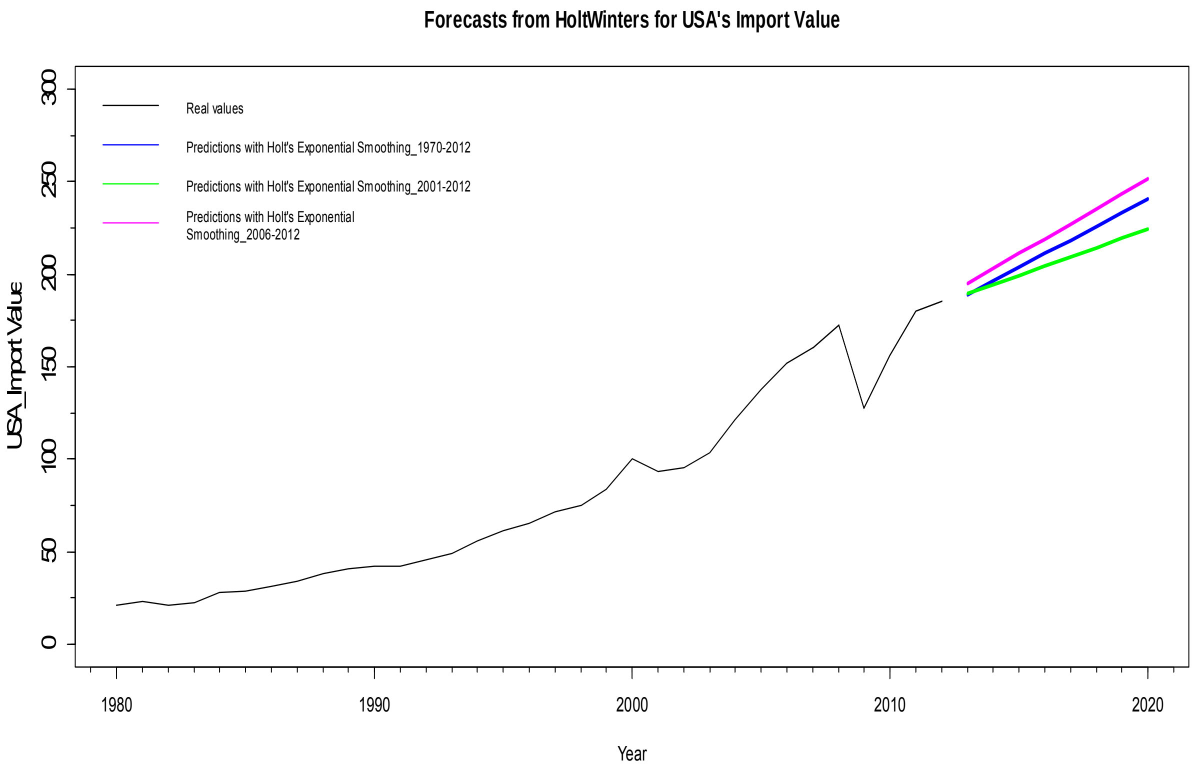
[Figure C.2.95] – Histogram and distribution of forecast residuals



[Figure C.2.96] – Comparison of Linear Regression and Holt's Exponential smoothing



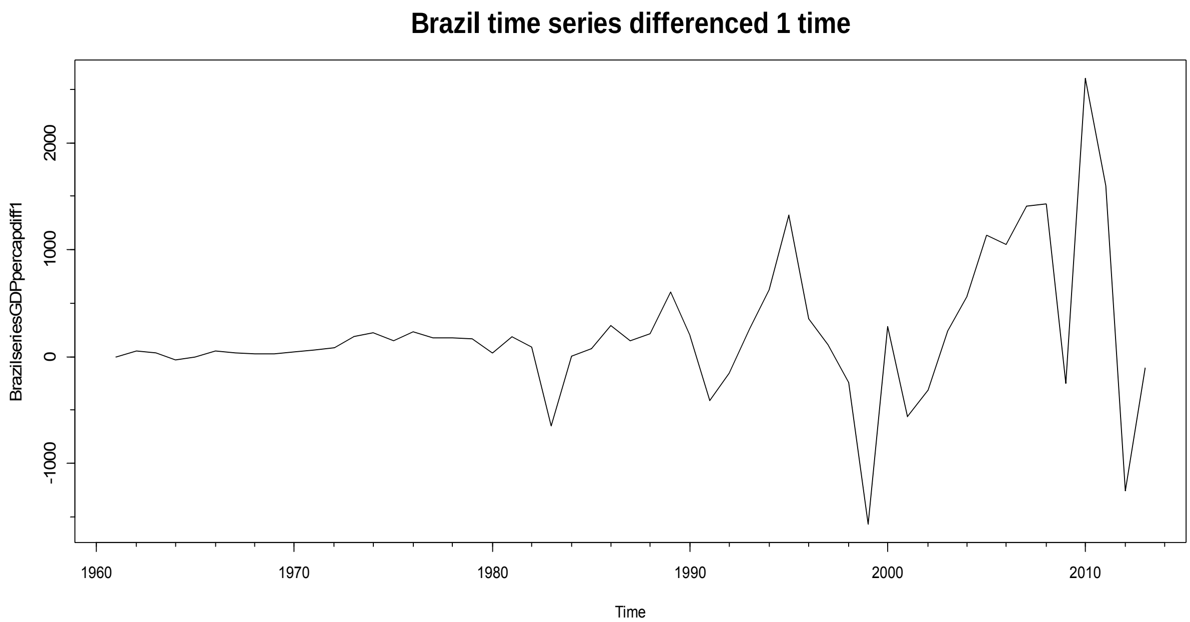
[Figure C.2.97] – Analysis for USA, Import Value and the dataset up to 2008



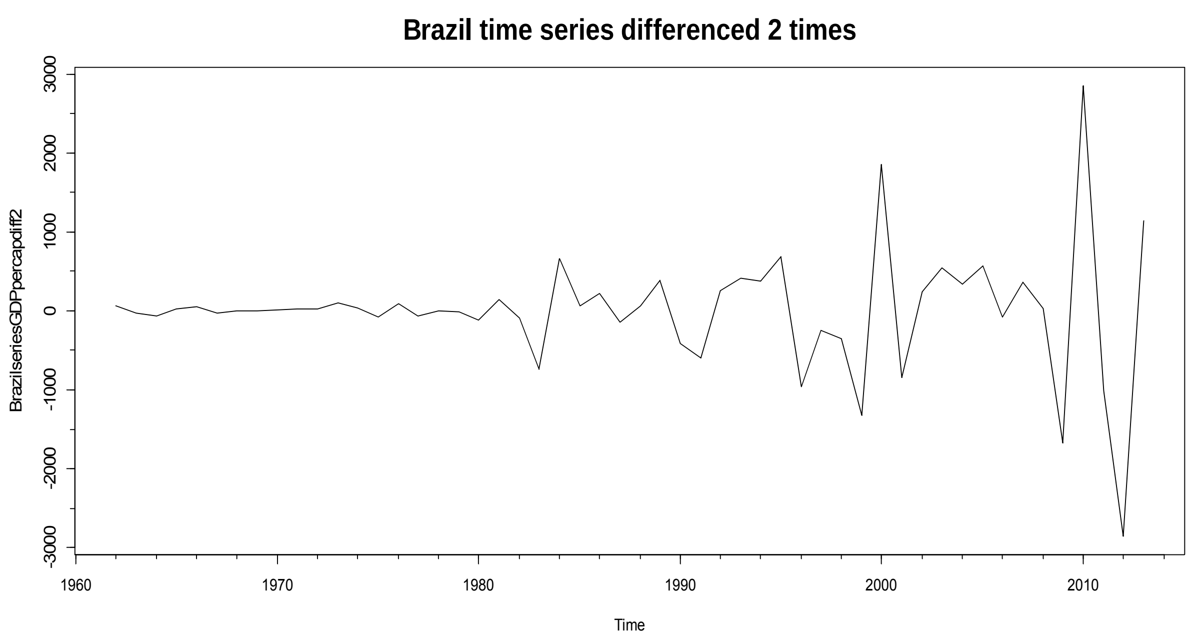
[Figure C.2.98] – Analyses for USA, Import Value and the subsets 2001-2013 and 2006-2013

C.3. ARIMA

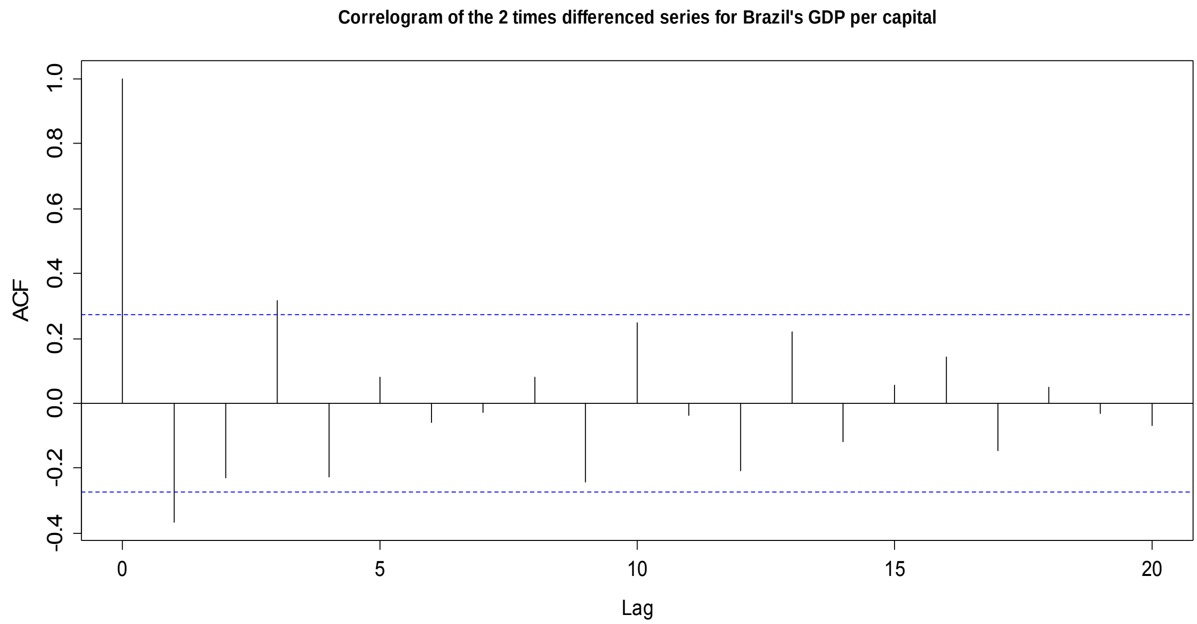
GDP per capita – Brazil



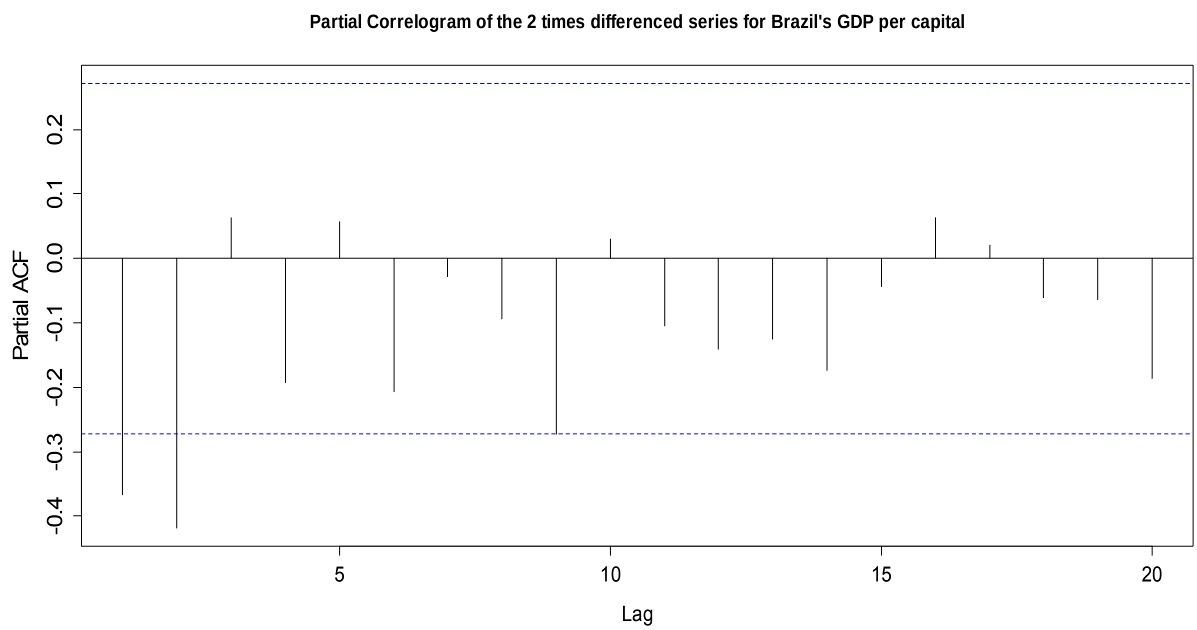
[Figure C.3.1] – One time differenced Brazil time series



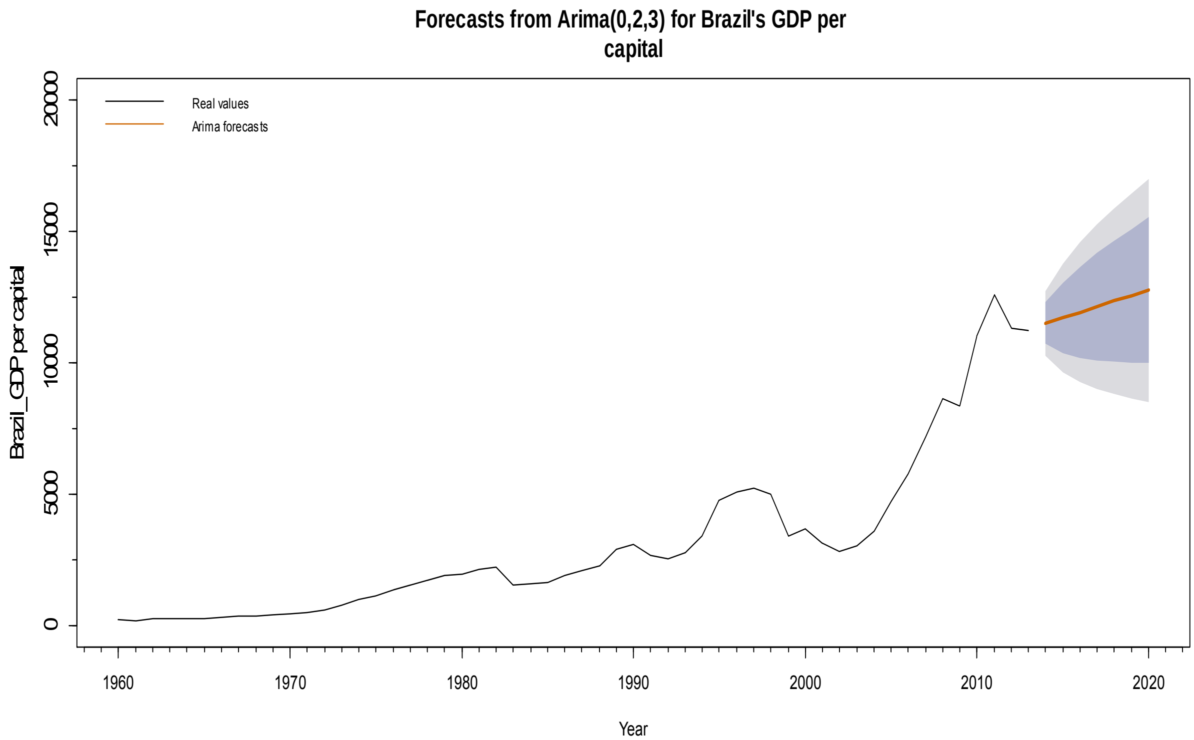
[Figure C.3.2] – Two times differenced Brazil time series



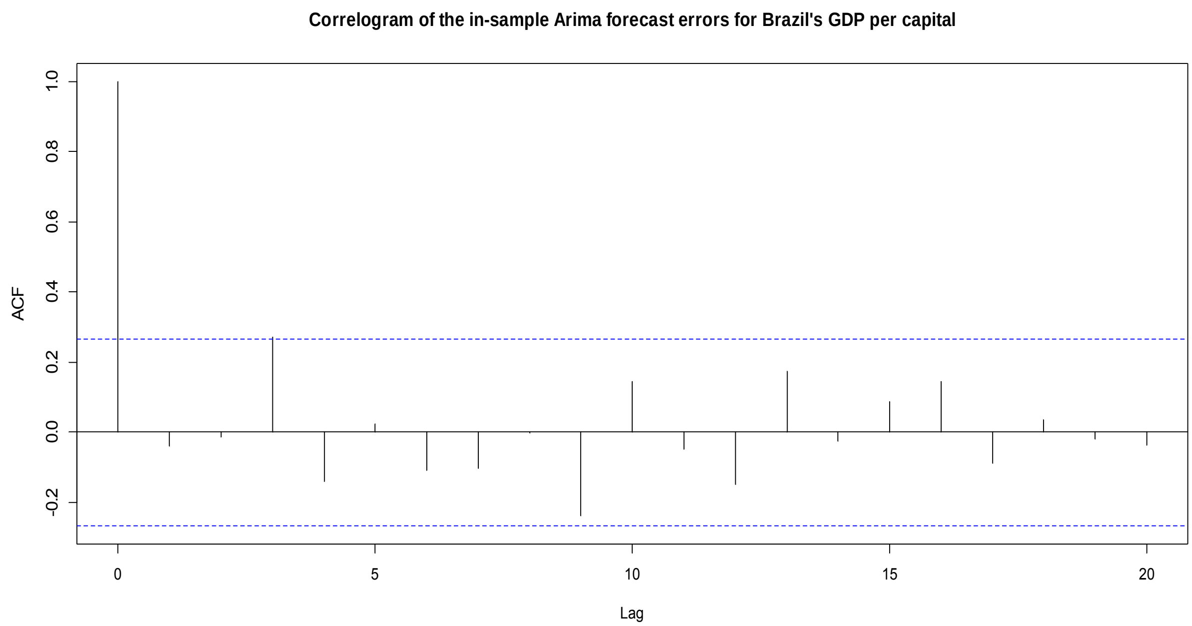
[Figure C.3.3] – Autocorrelogram (ACF) of the twice differenced Brazil time series



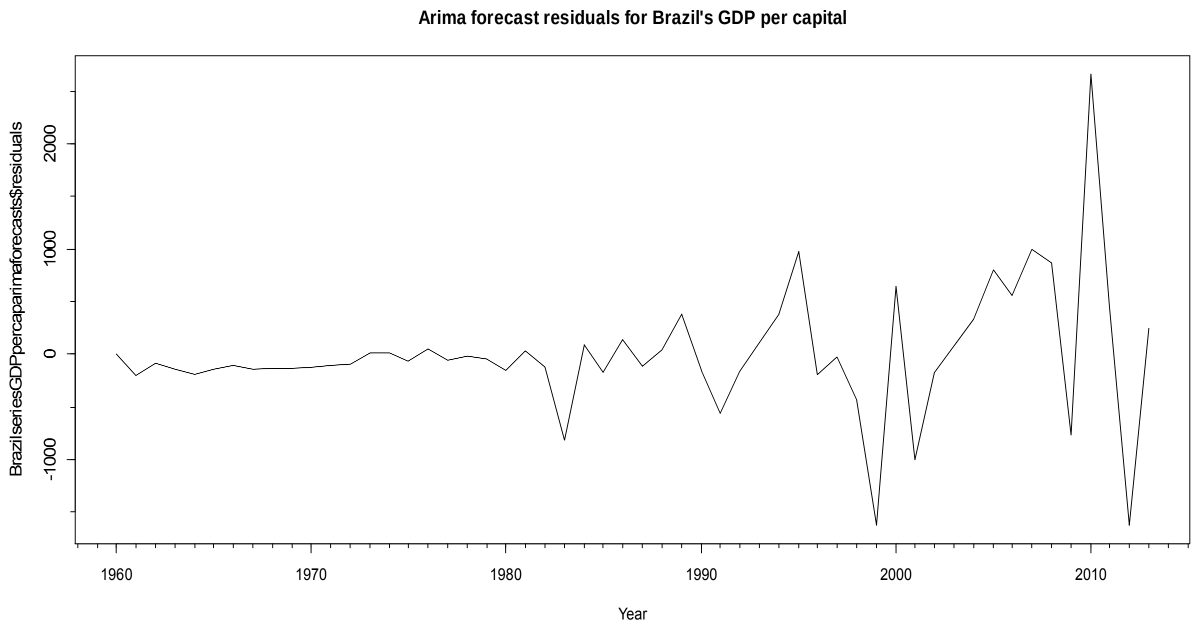
[Figure C.3.4] – Partial autocorrelogram (PACF) of the twice differenced Brazil time series



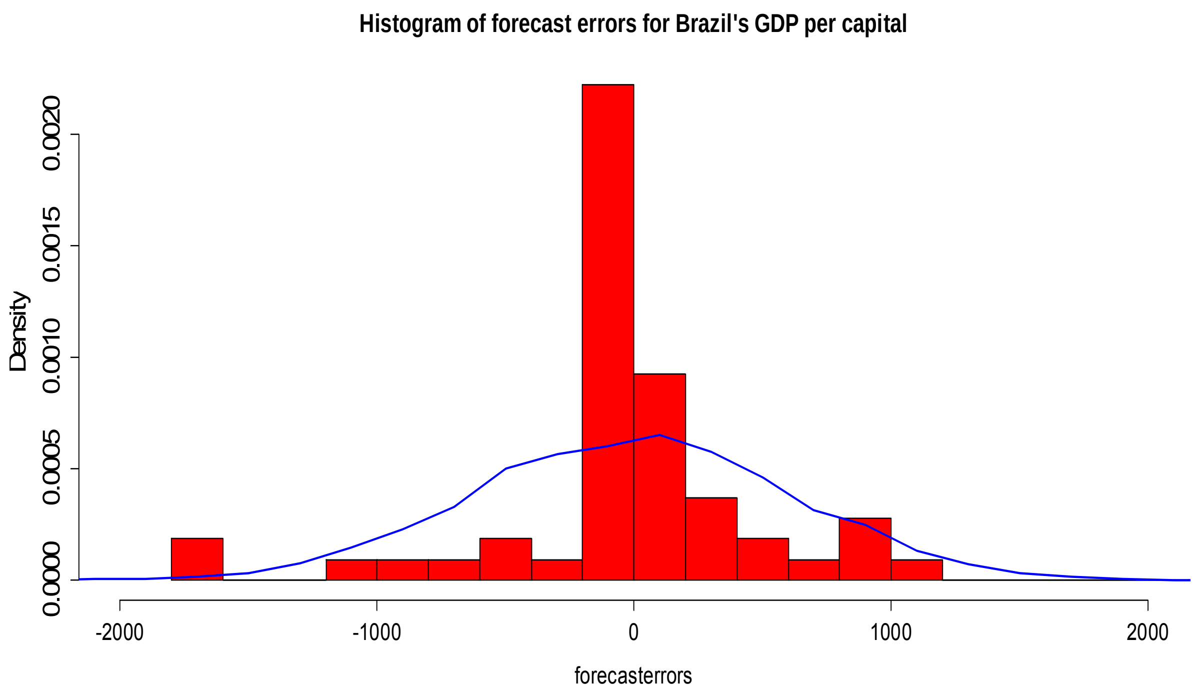
[Figure C.3.5] – Analysis for Brazil, GDP per capita and whole dataset



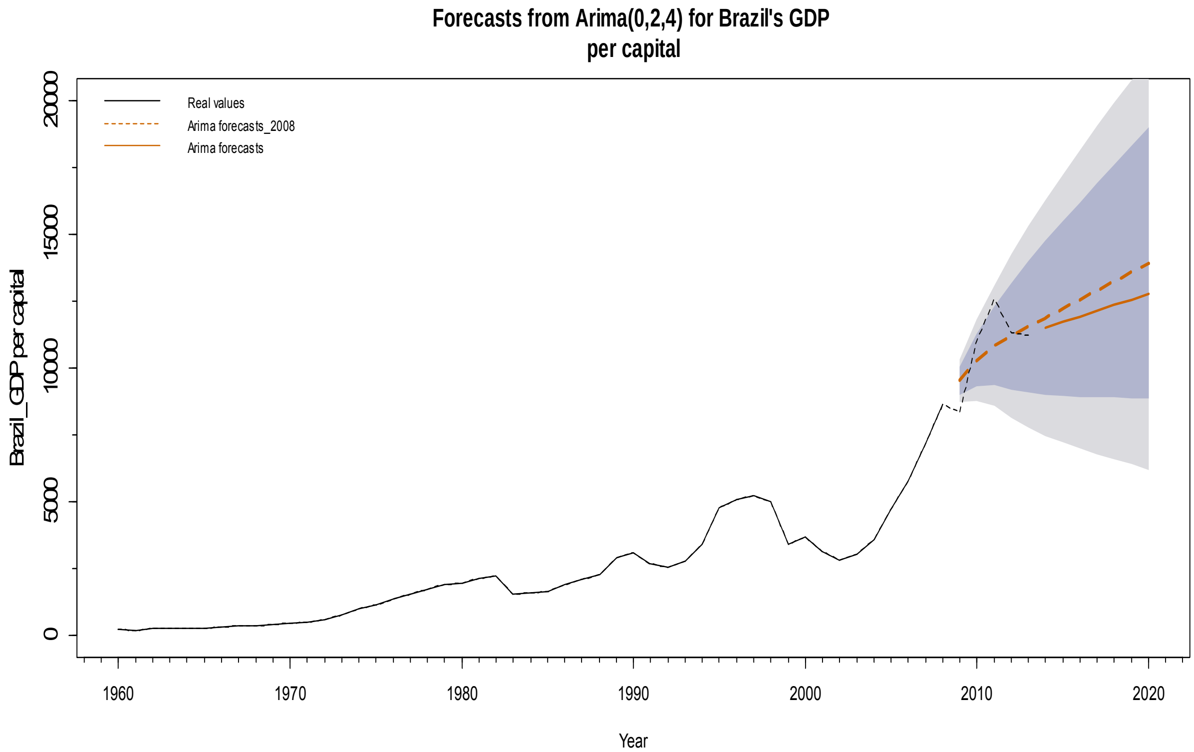
[Figure C.3.6] – Correlogram of in-sample errors of ARIMA forecasts



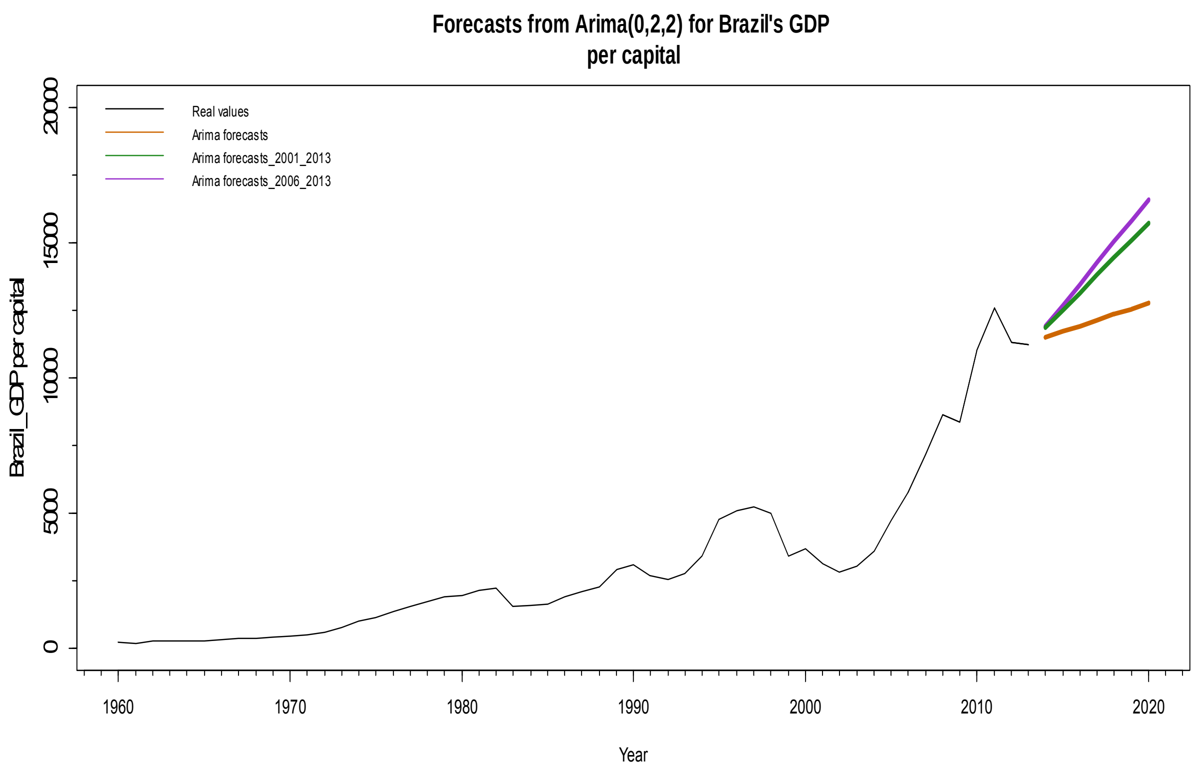
[Figure C.3.7] –Residuals of ARIMA forecasts



[Figure C.3.8] – Histogram and distribution of forecast residuals

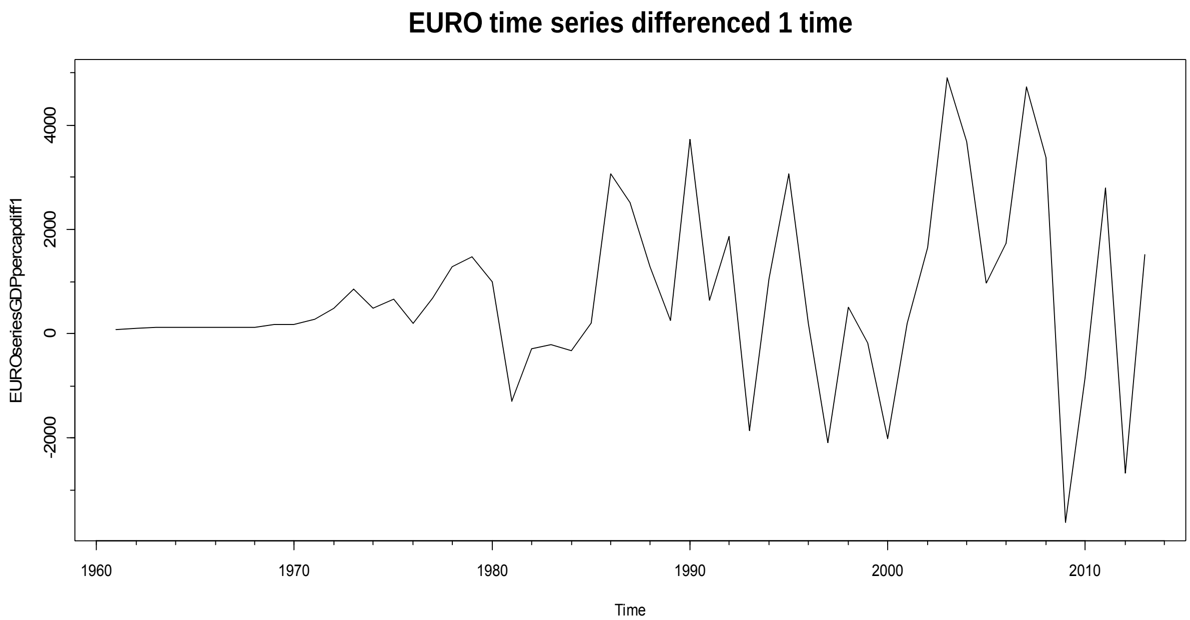


[Figure C.3.9] – Analysis for Brazil, GDP per capita and the dataset up to 2008

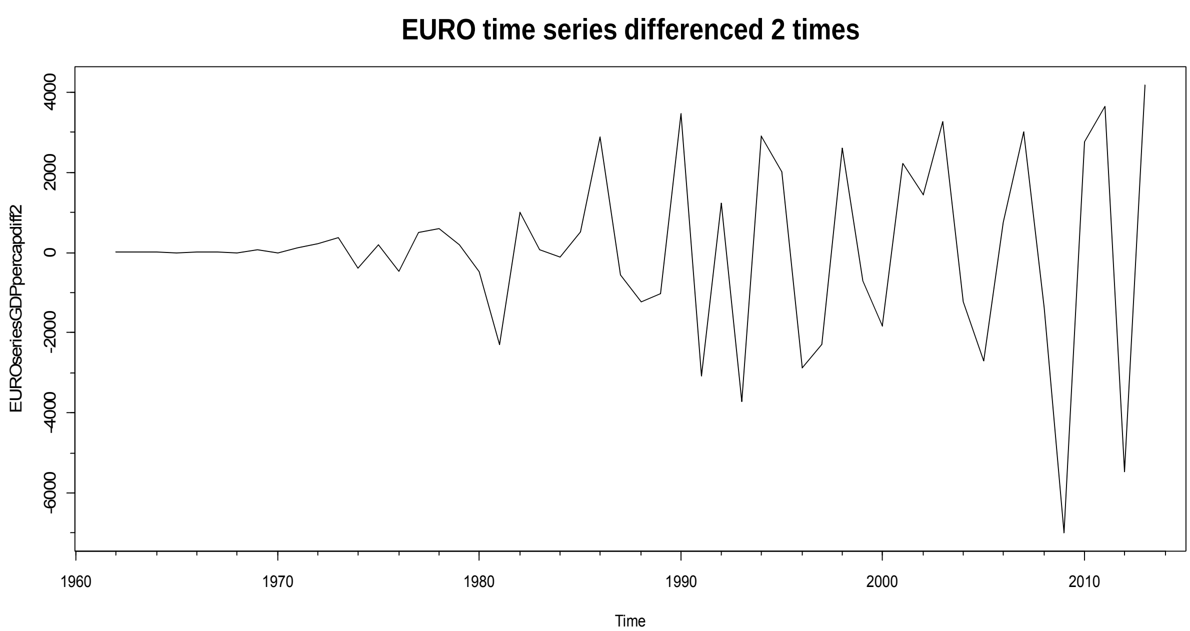


[Figure C.3.10] – Analyses for Brazil, GDP per capita and the subsets 2001-2013 and 2006-2013

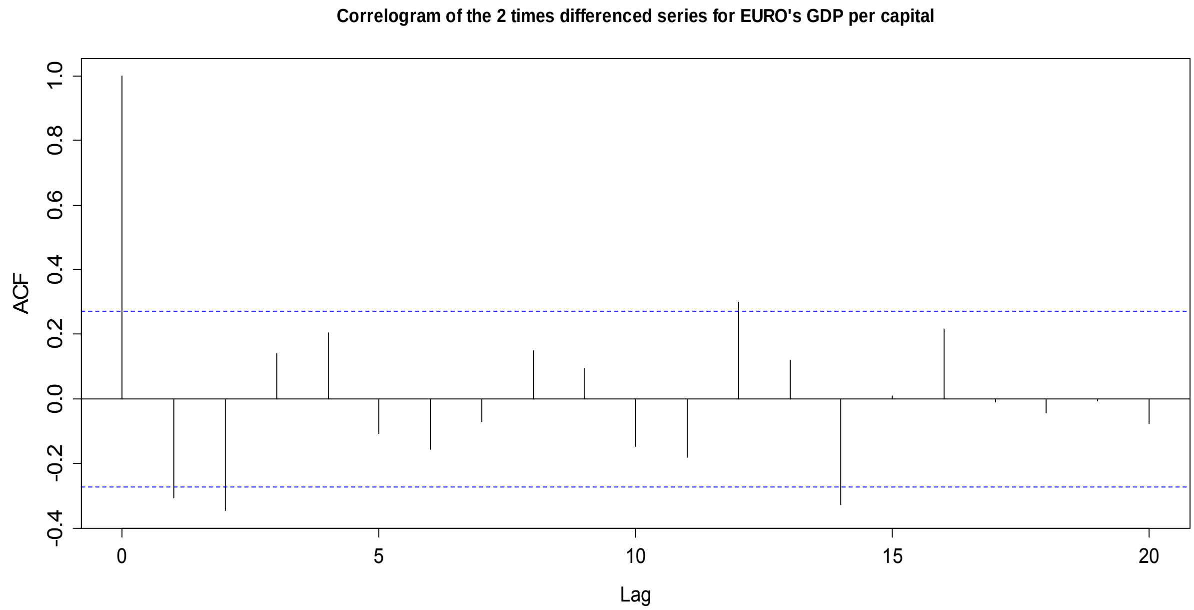
GDP per capita – EURO zone



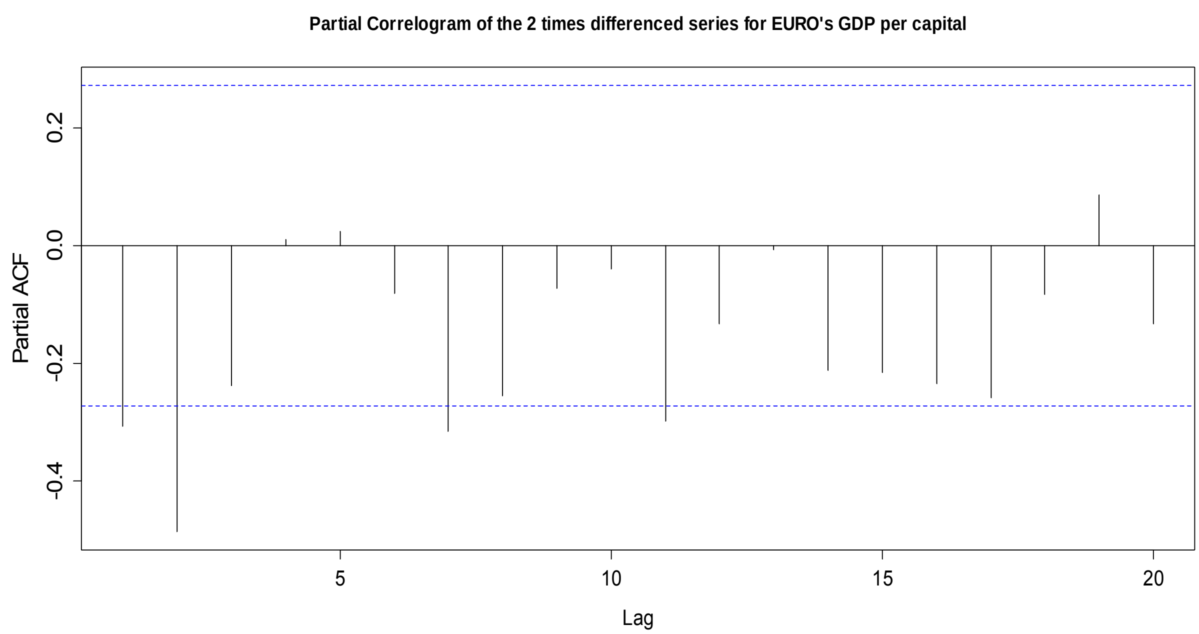
[Figure C.3.11] – One time differenced EURO zone time series



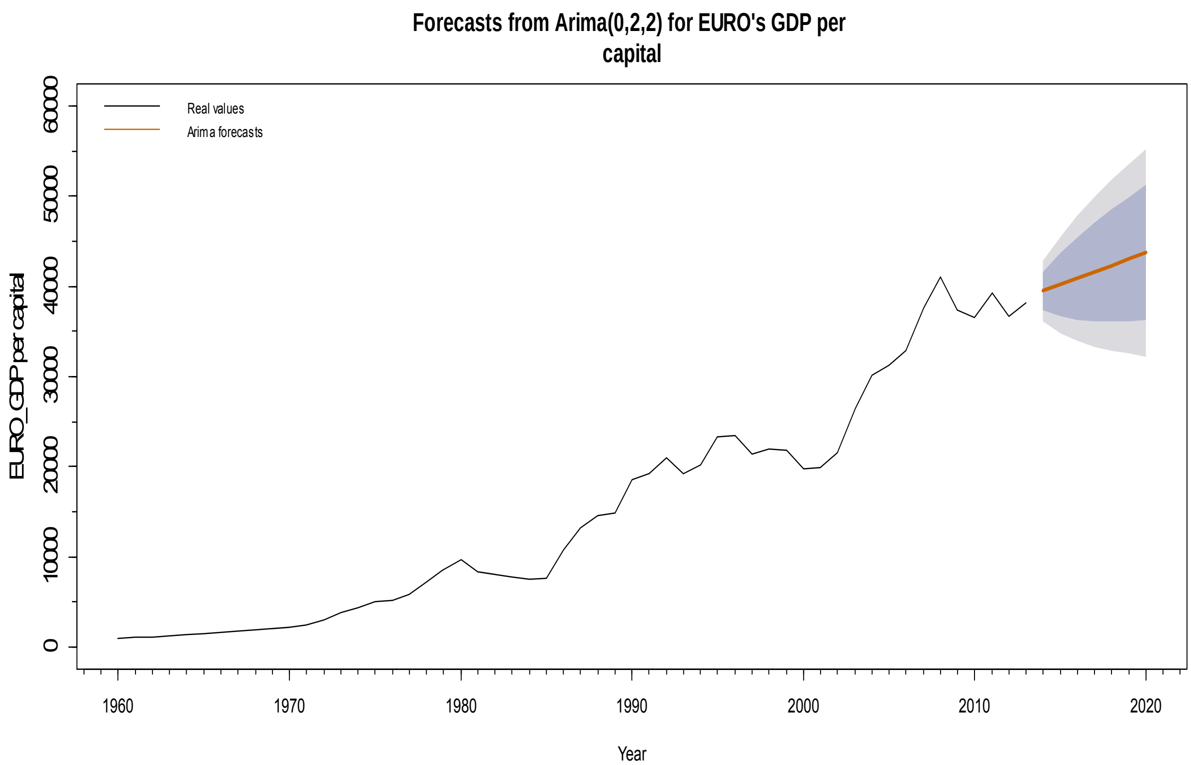
[Figure C.3.12] – Two times differenced EURO zone time series



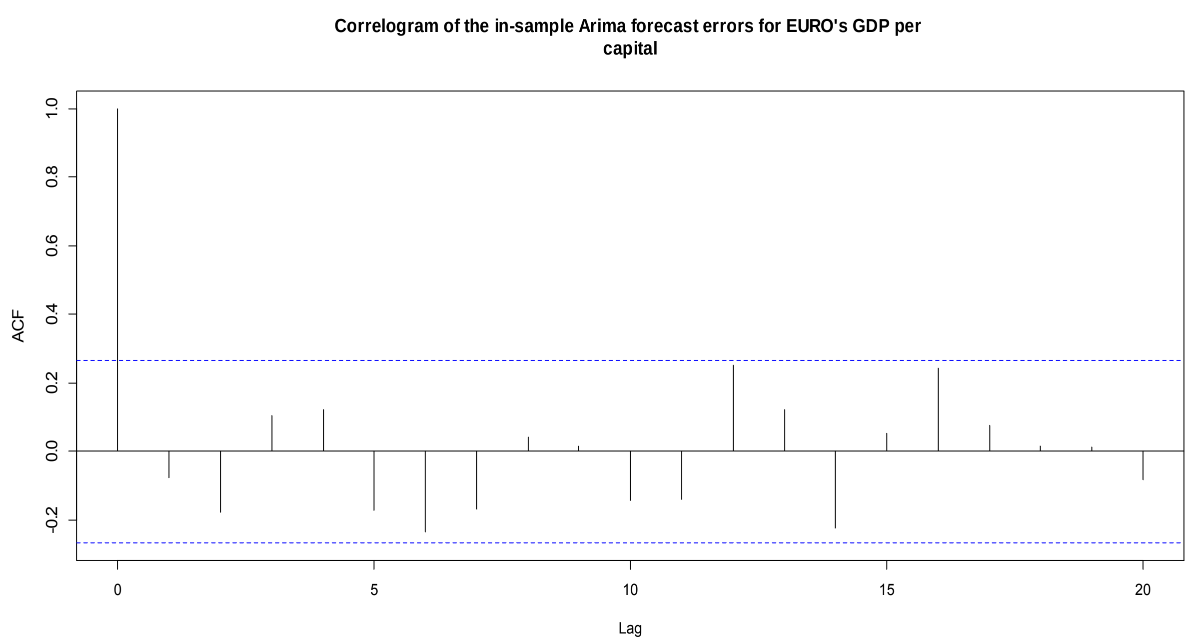
[Figure C.3.13] – Autocorrelogram (ACF) of the twice differenced EURO zone time series



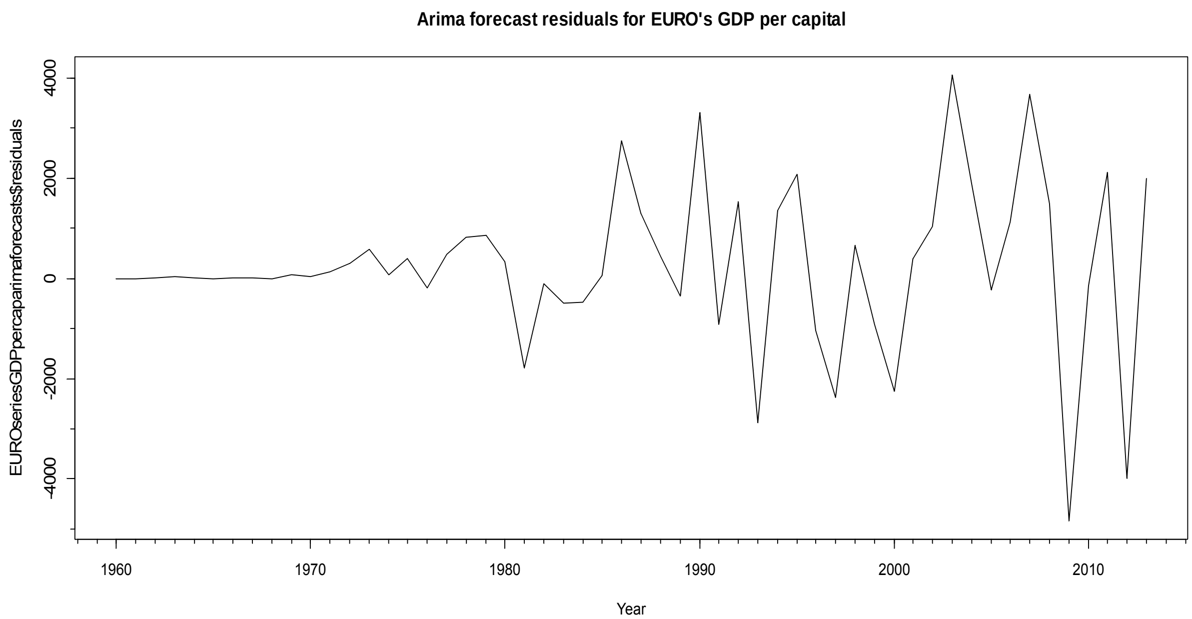
[Figure C.3.14] – Partial autocorrelogram (PACF) of the twice differenced EURO zone time series



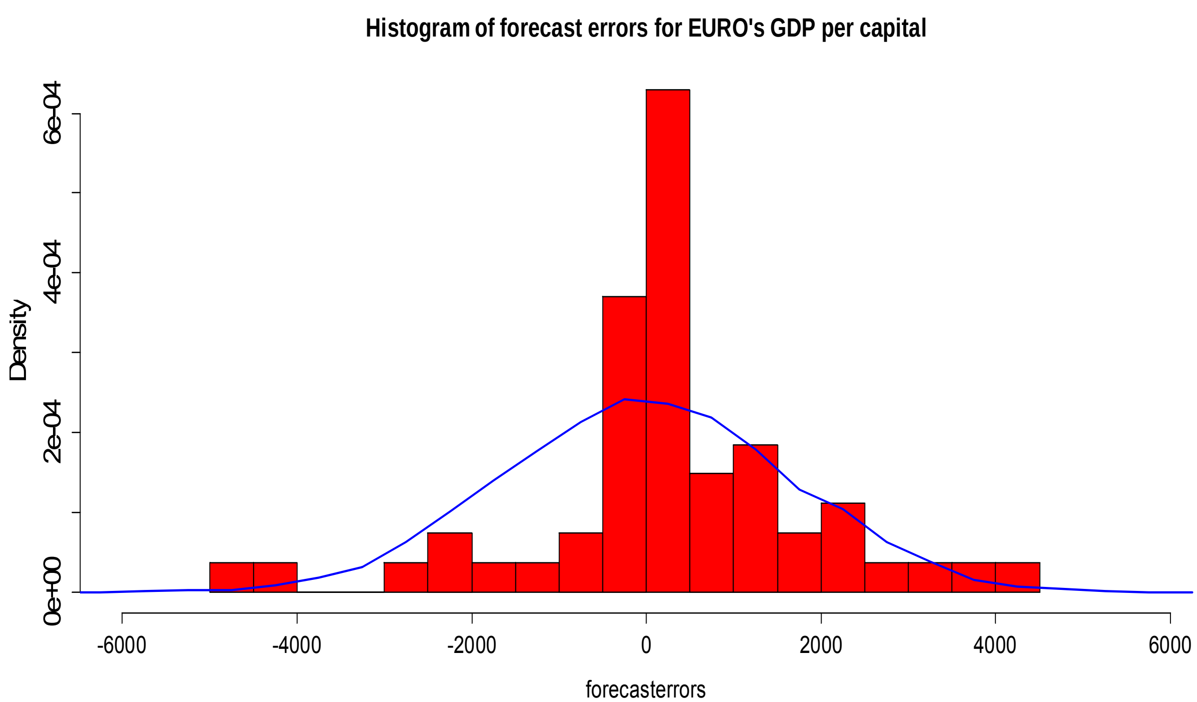
[Figure C.3.15] – Analysis for EURO zone, GDP per capita and whole dataset



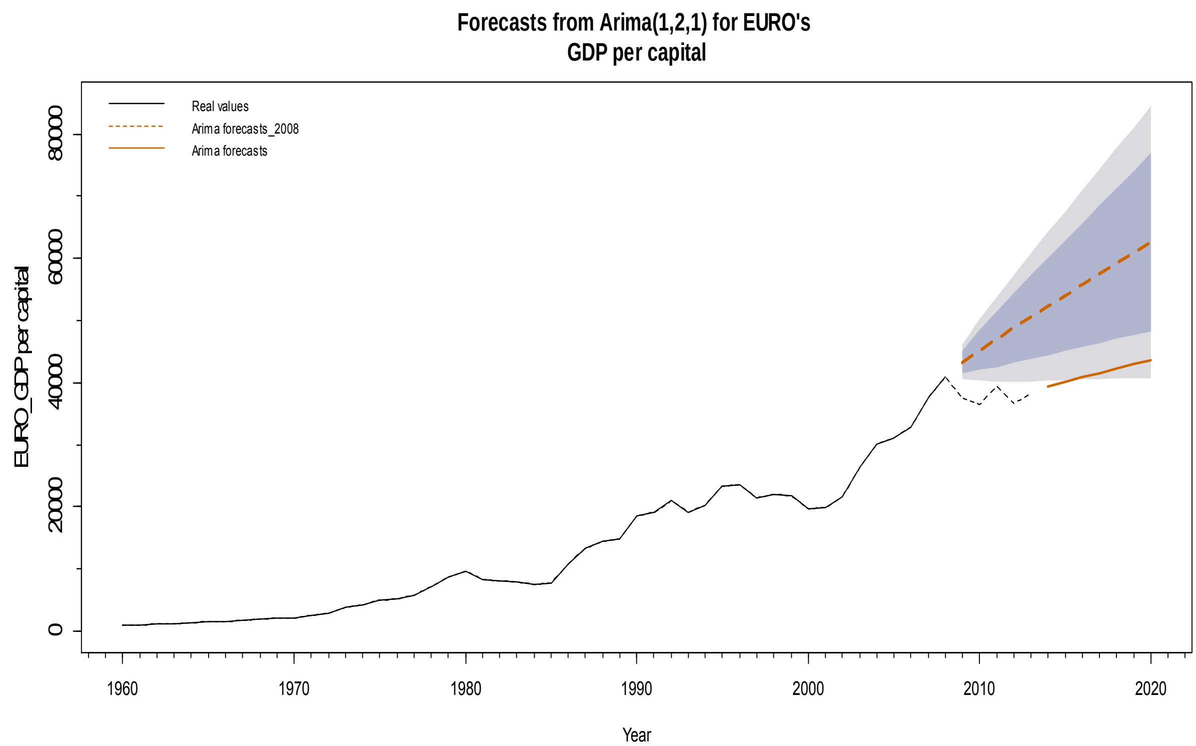
[Figure C.3.16] – Correlogram of in-sample errors of ARIMA forecasts



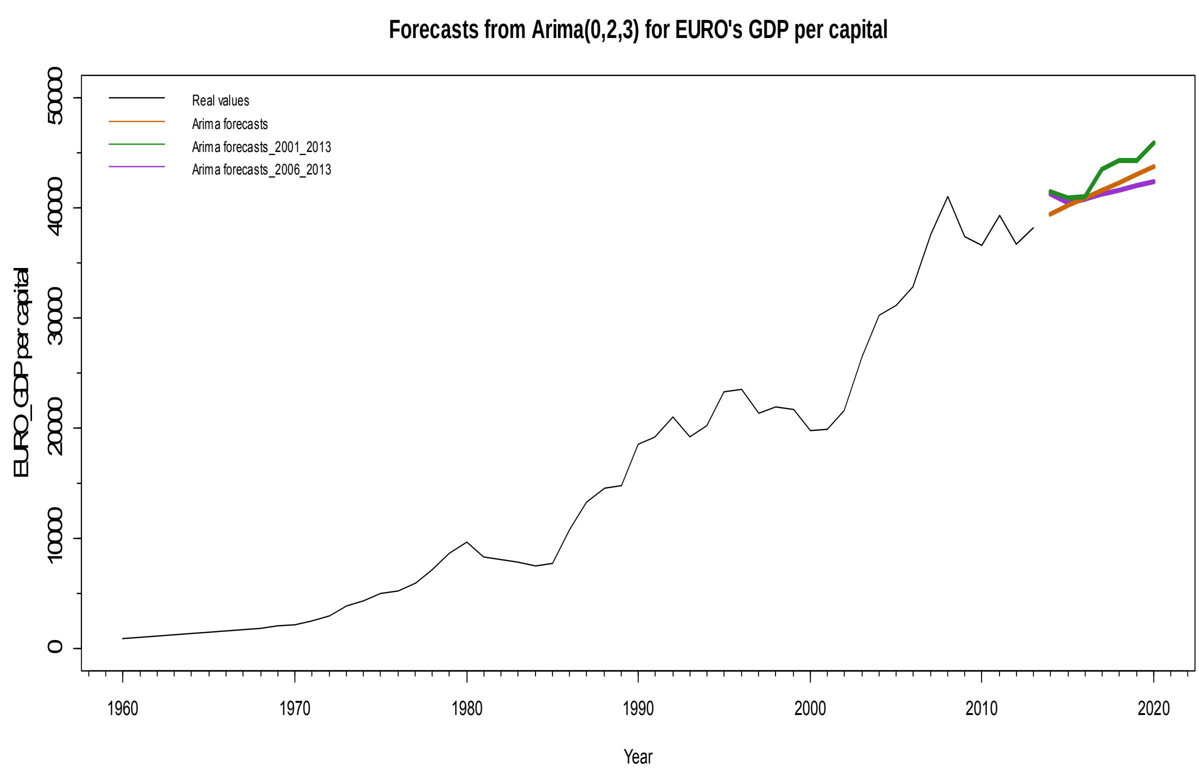
[Figure C.3.17] –Residuals of ARIMA forecasts



[Figure C.3.18] – Histogram and distribution of forecast residuals

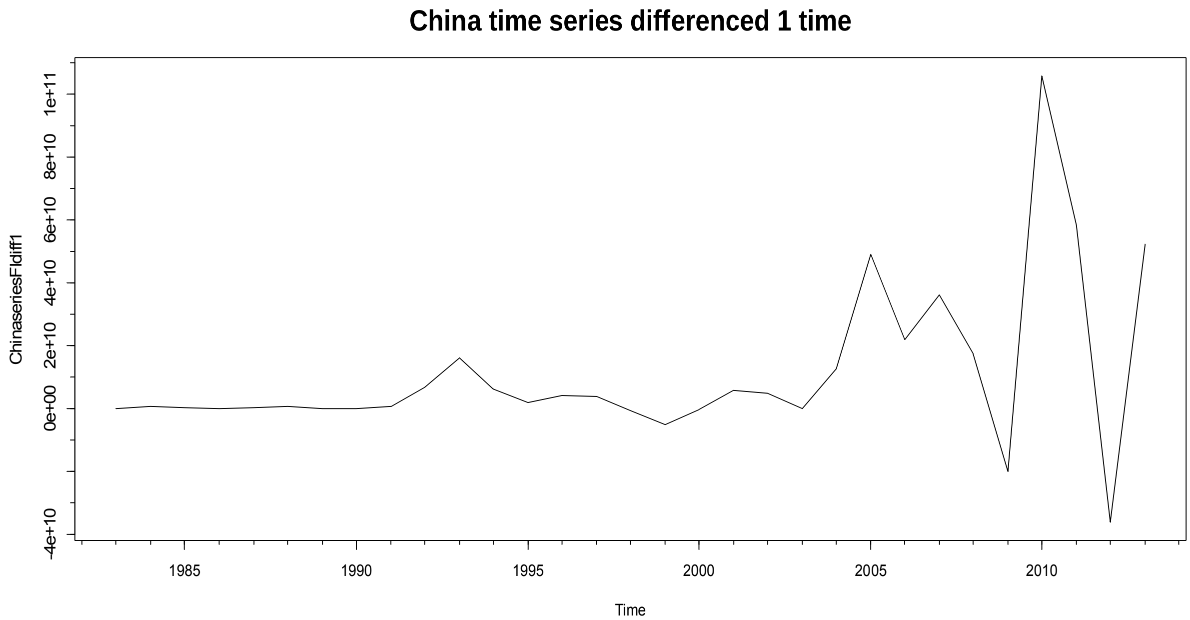


[Figure C.3.19] – Analysis for EURO zone, GDP per capita and the dataset up to 2008

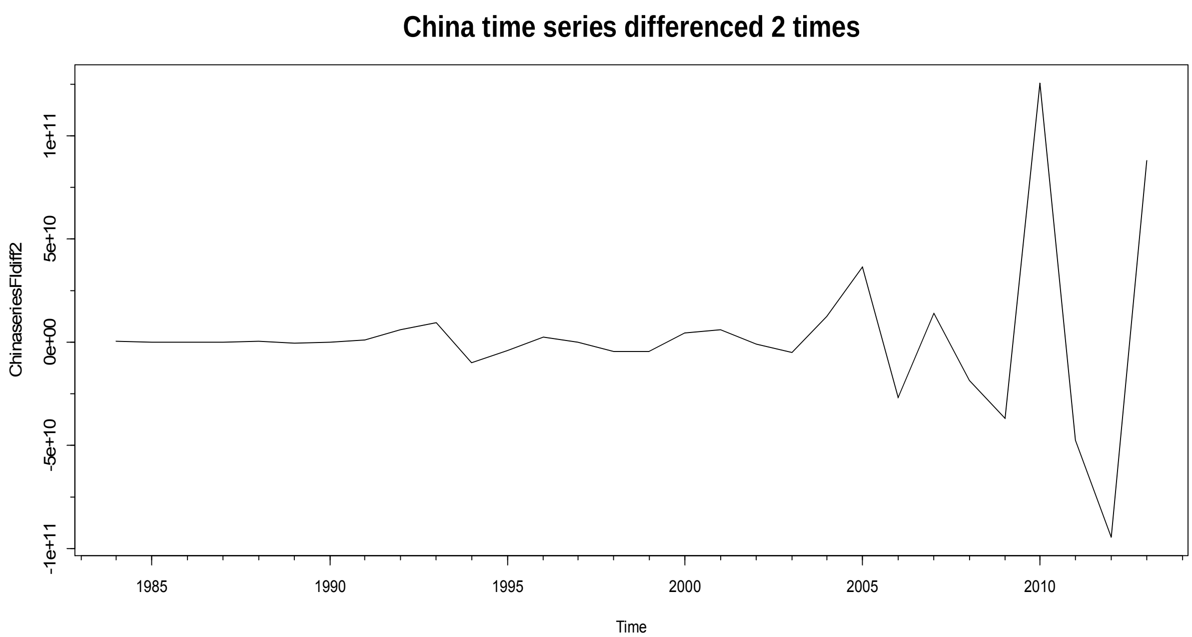


[Figure C.3.20] – Analyses for EURO zone, GDP per capita and the subsets 2001-2013 and 2006-2013

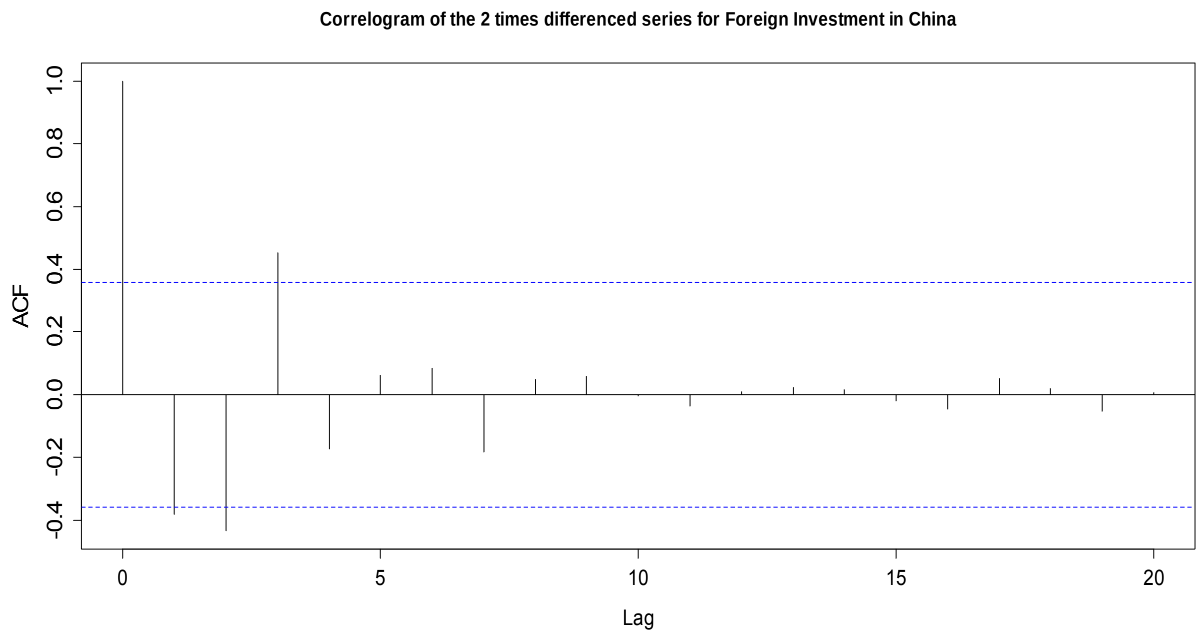
Foreign Investment – China



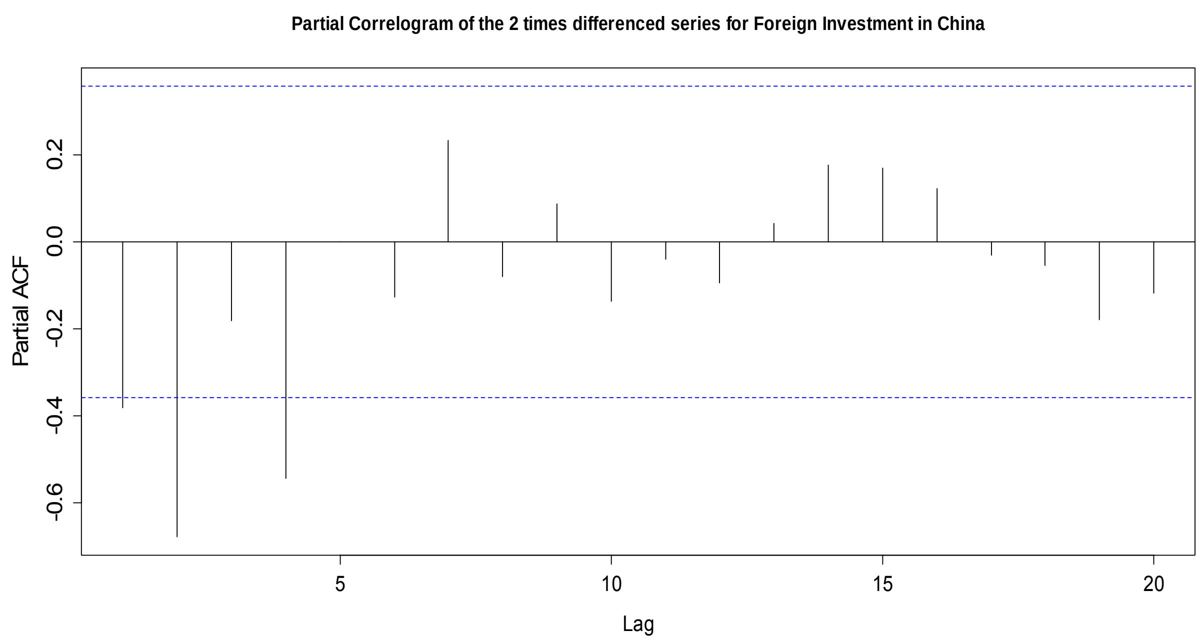
[Figure C.3.21] – One time differenced China time series



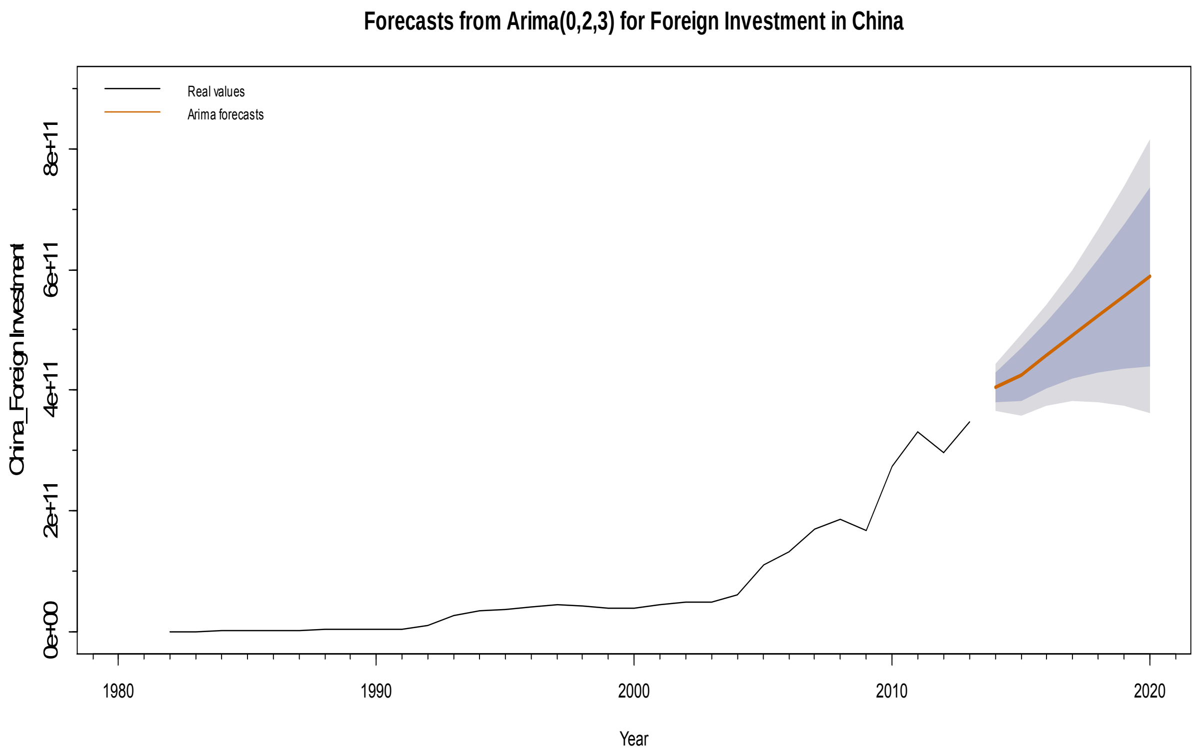
[Figure C.3.22] – Two times differenced China time series



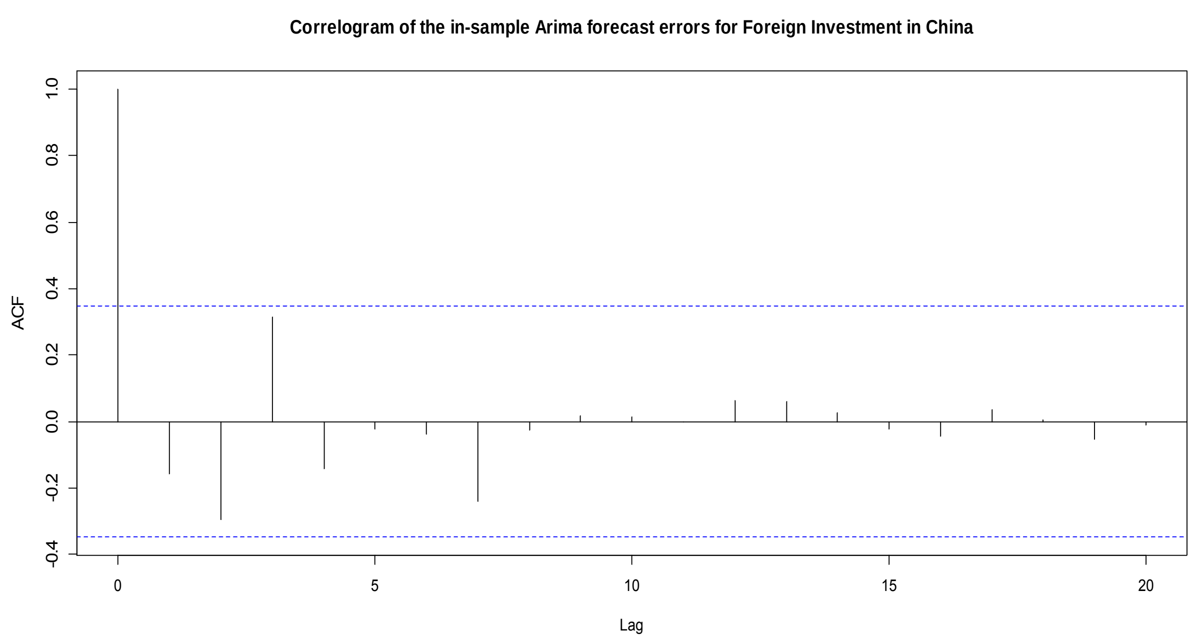
[Figure C.3.23] – Autocorrelogram (ACF) of the twice differenced China time series



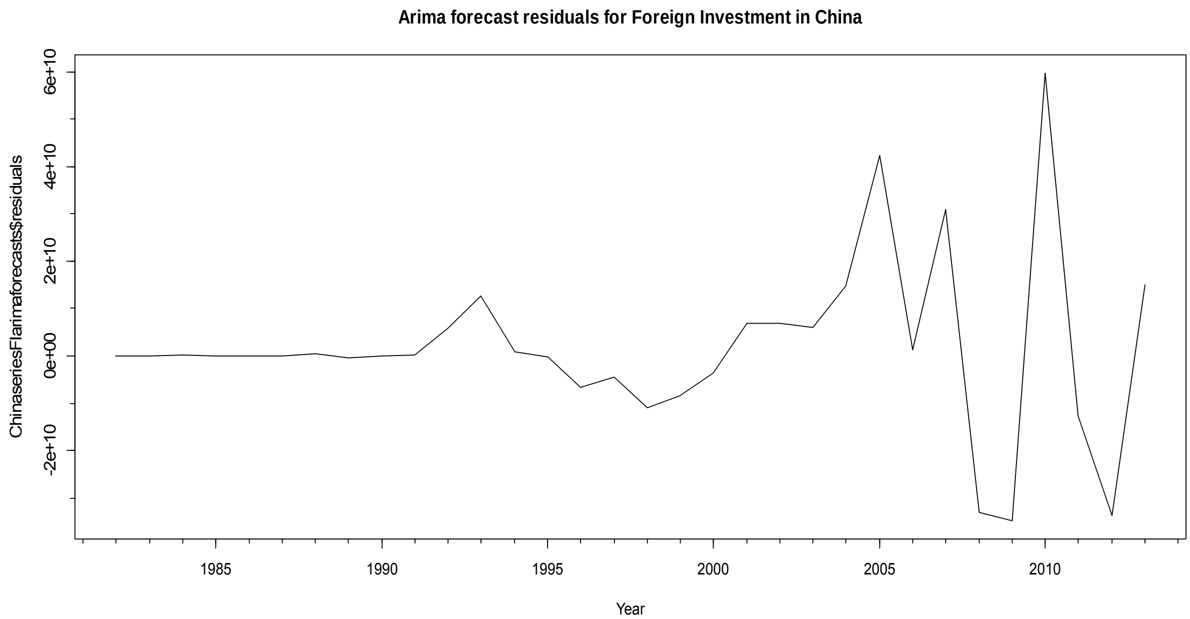
[Figure C.3.24] – Partial autocorrelogram (PACF) of the twice differenced China time series



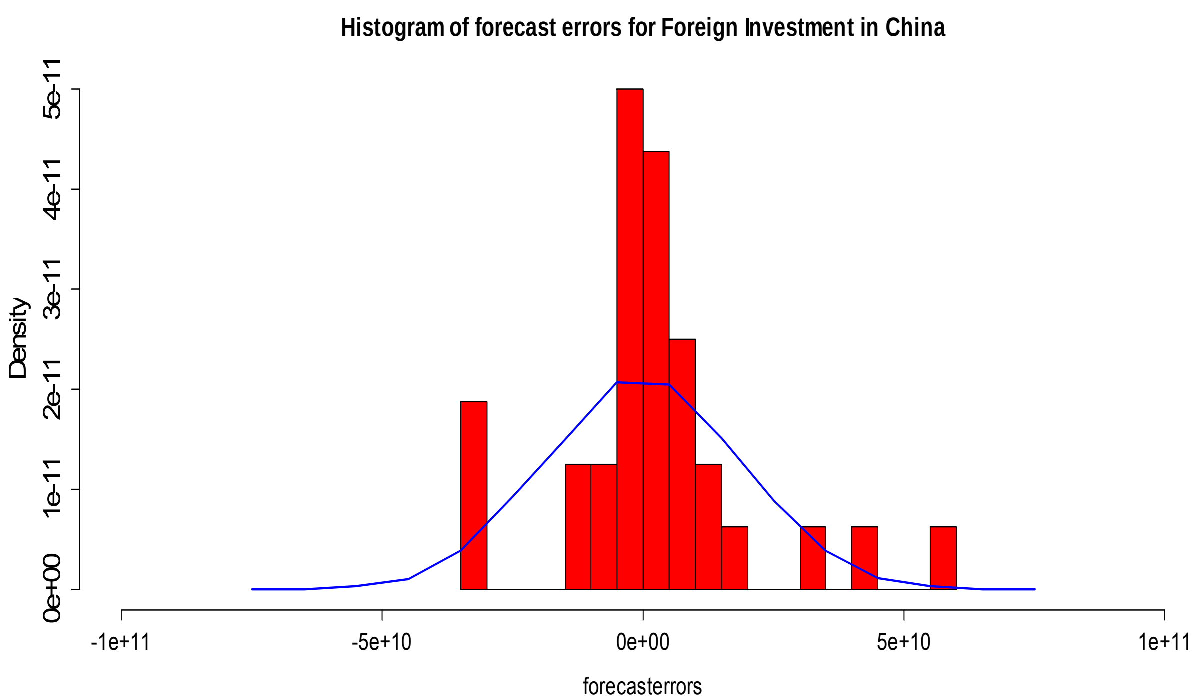
[Figure C.3.25] – Analysis for China, Foreign Investment and whole dataset



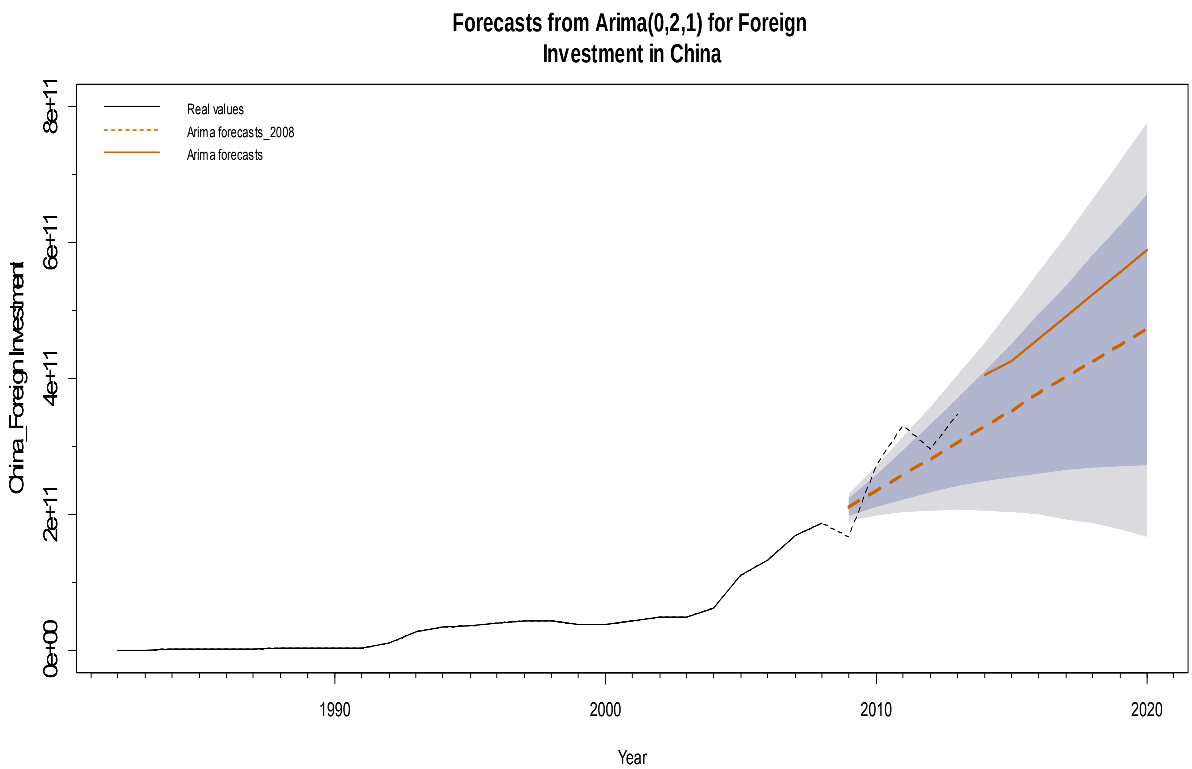
[Figure C.3.26] – Correlogram of in-sample errors of ARIMA forecasts



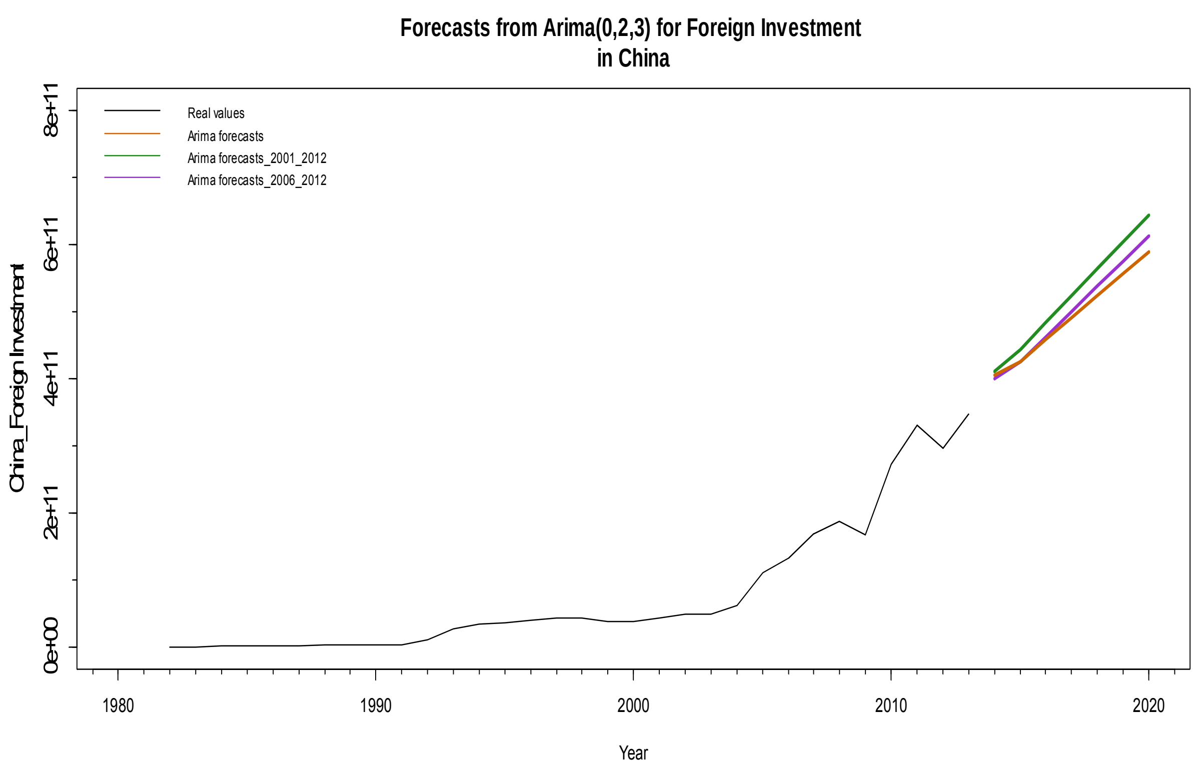
[Figure C.3.27] –Residuals of ARIMA forecasts



[Figure C.3.28] – Histogram and distribution of forecast residuals

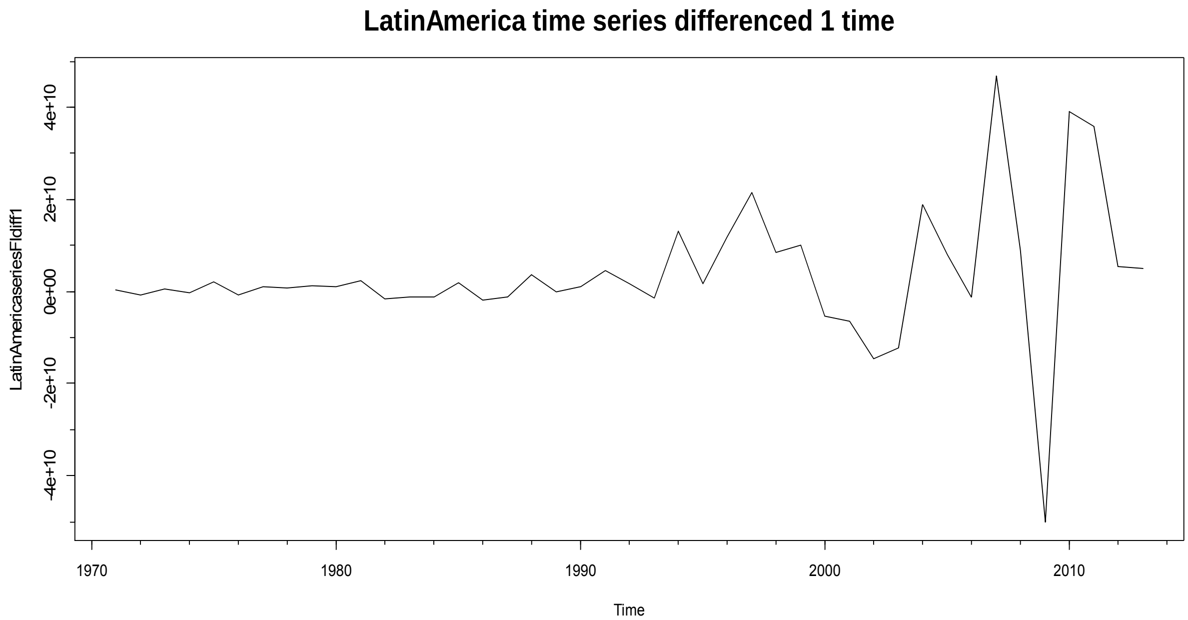


[Figure C.3.29] – Analysis for China, Foreign Investment and the dataset up to 2008

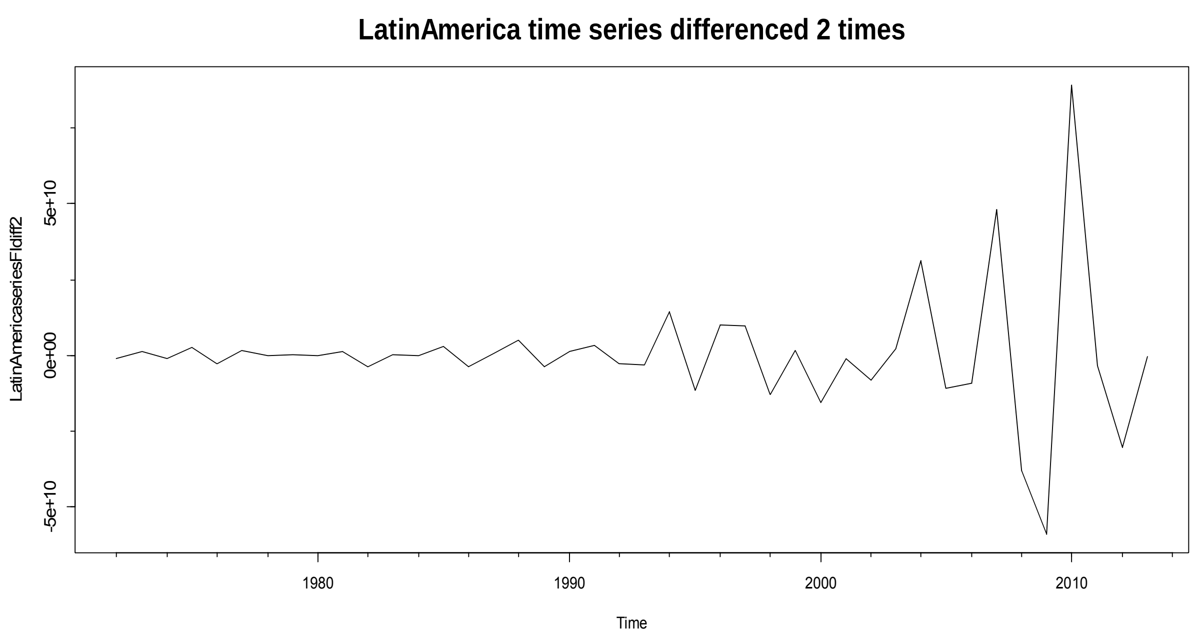


[Figure C.3.30] – Analyses for China, Foreign Investment and the subsets 2001-2013 and 2006-2013

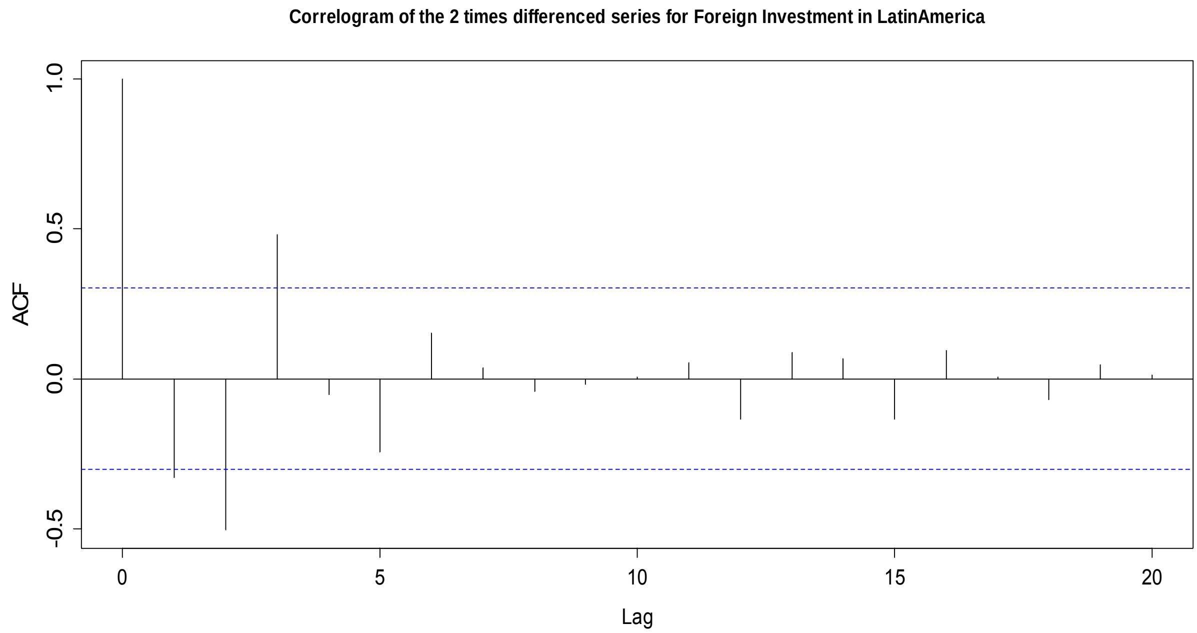
<u>Foreign Investment – Latin America and Caribbean</u>



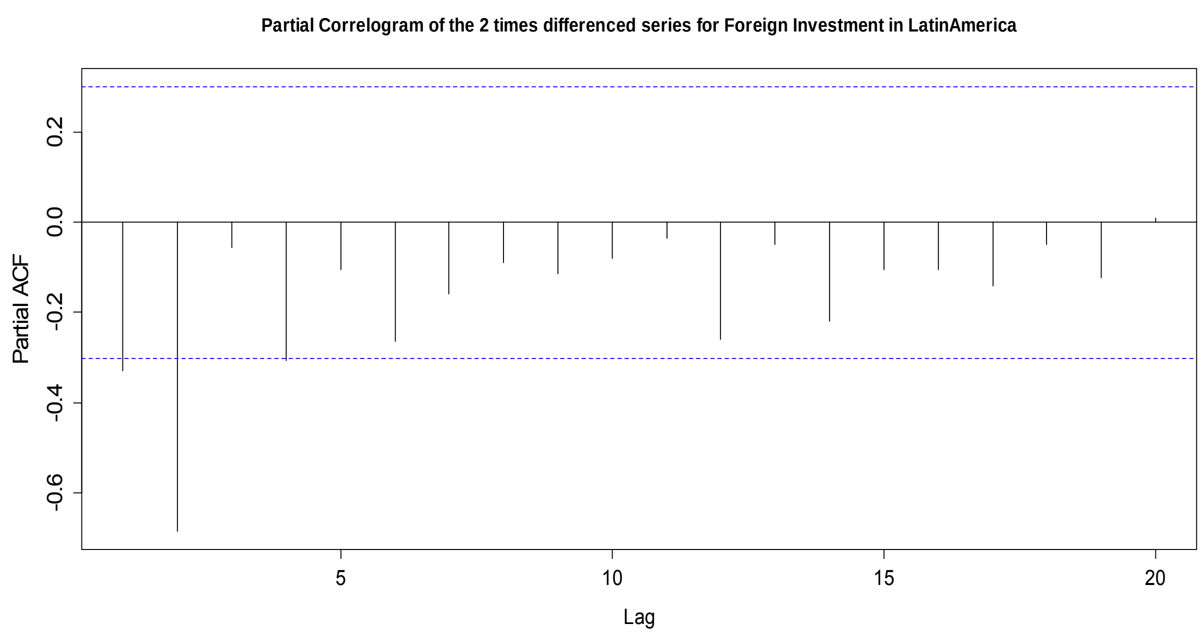
[Figure C.3.31] – One time differenced Latin America and Caribbean time series



[Figure C.3.32] – Two times differenced Latin America and Caribbean time series

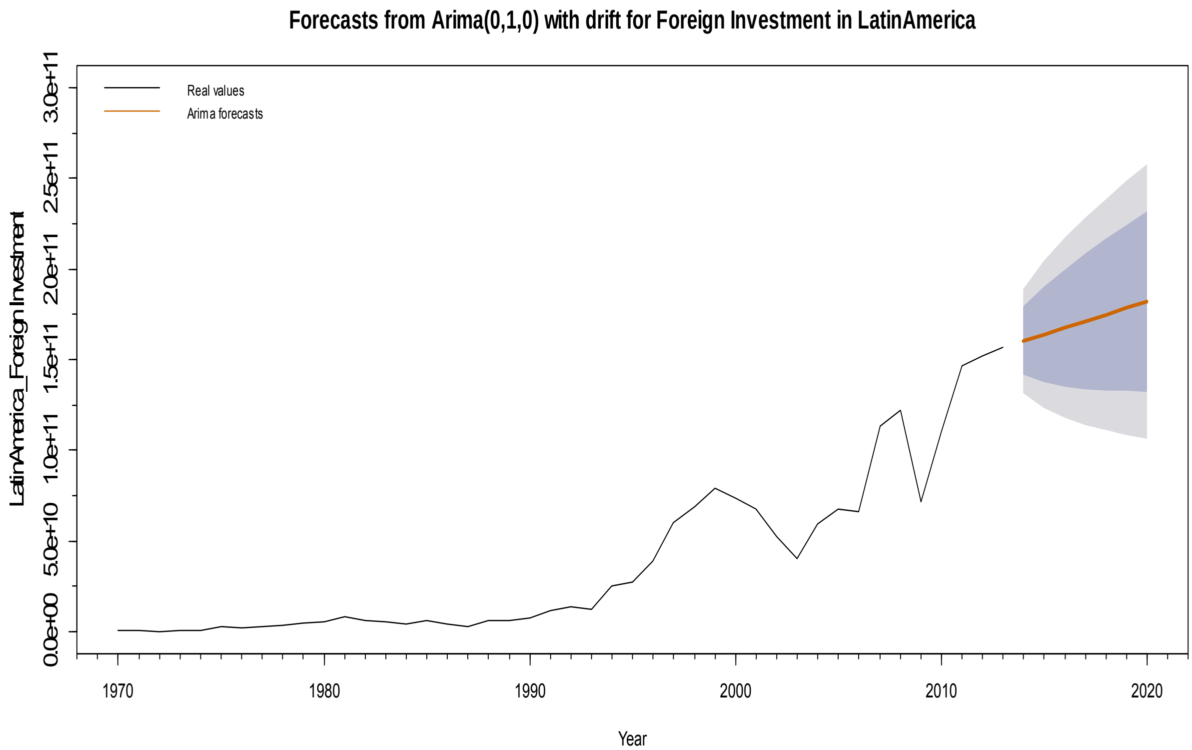


[Figure C.3.33] – Autocorrelogram (ACF) of the twice differenced Latin America and Caribbean time series

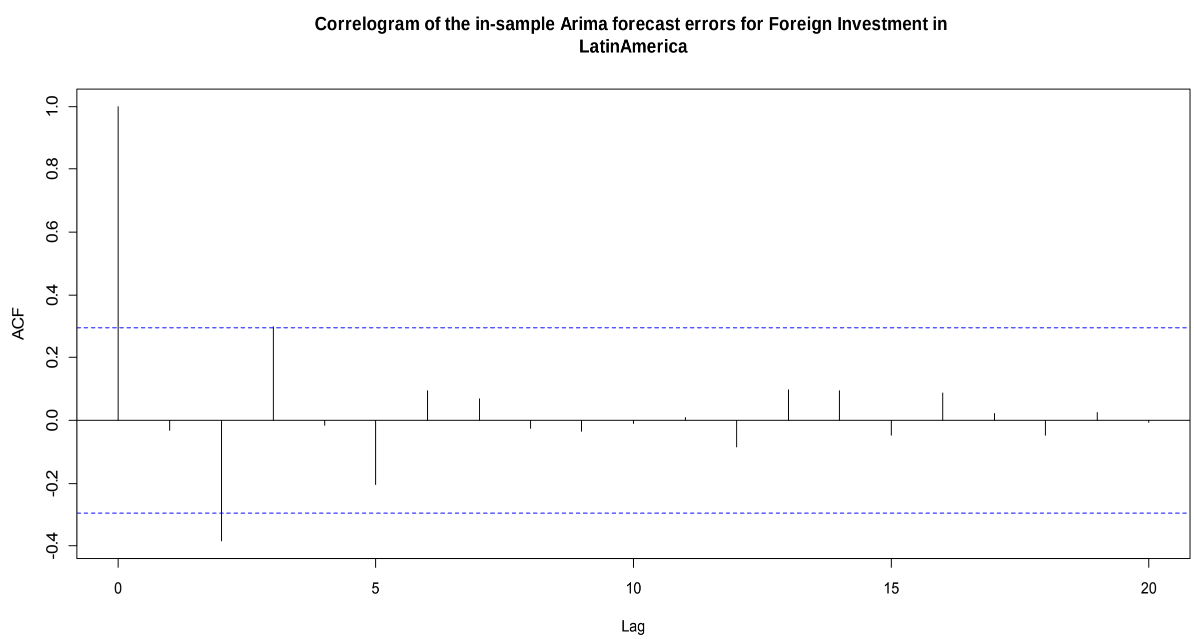


[Figure C.3.34] – Partial autocorrelogram (PACF) of the twice differenced Latin America and Caribbean time series

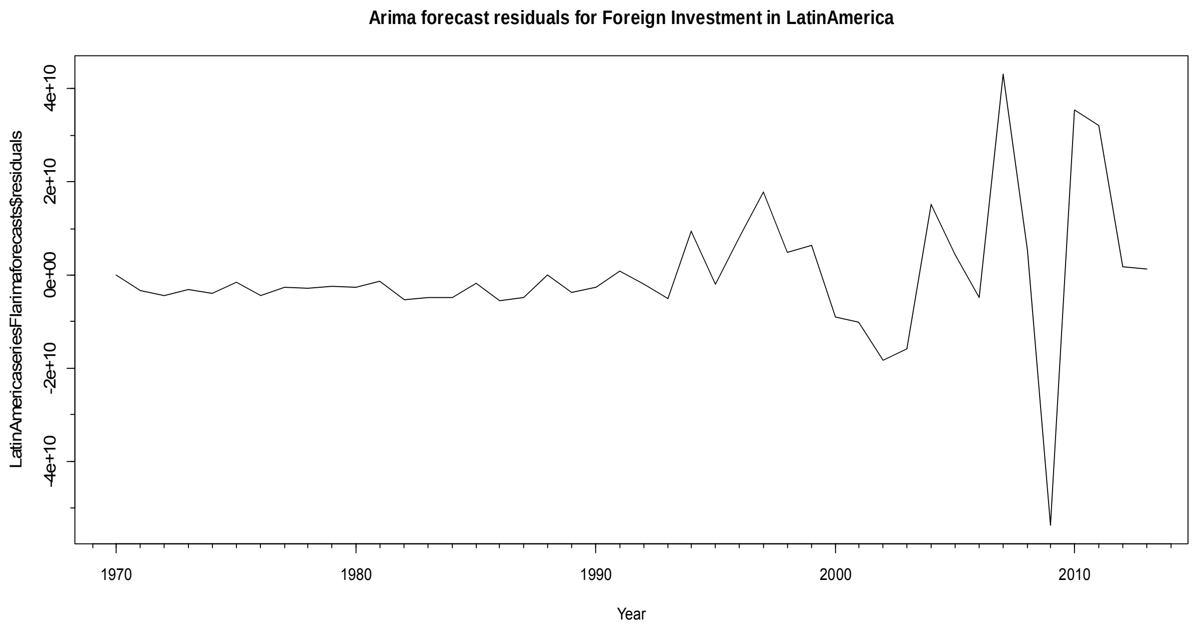
<u>Remark</u>: The suggested methodology gave an ARIMA model (0,2,3), but the results we get with that do not seem reasonable. I chose to continue the analysis with the model '(0,1,0) with drift' as suggested by the auto.arima function, which gives more reasonable results.



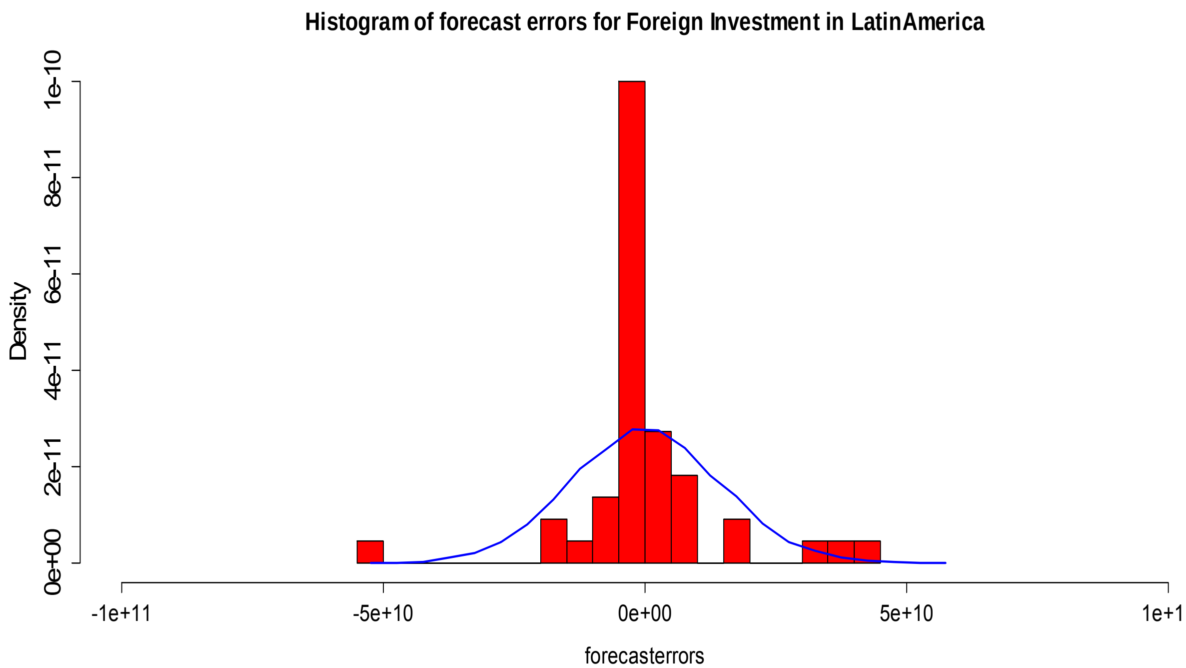
[Figure C.3.35] – Analysis for Latin America and Caribbean, Foreign Investment and whole dataset



[Figure C.3.36] – Correlogram of in-sample errors of ARIMA forecasts

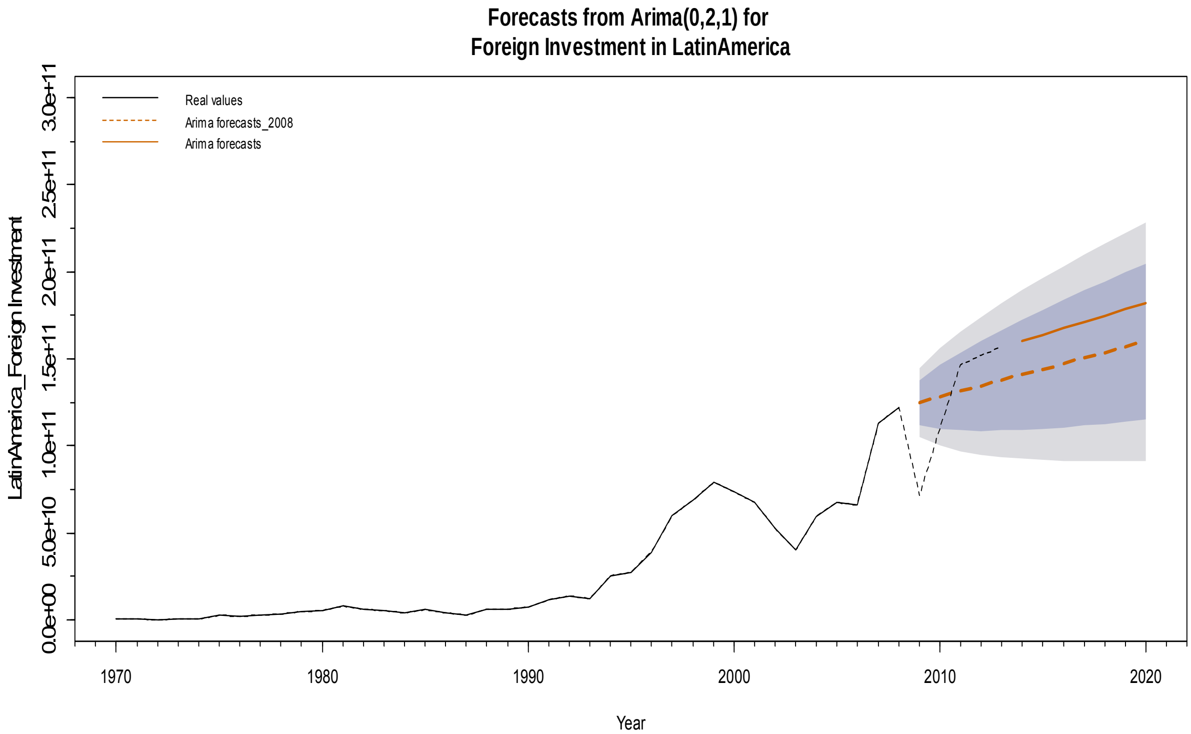


[Figure C.3.37] –Residuals of ARIMA forecasts

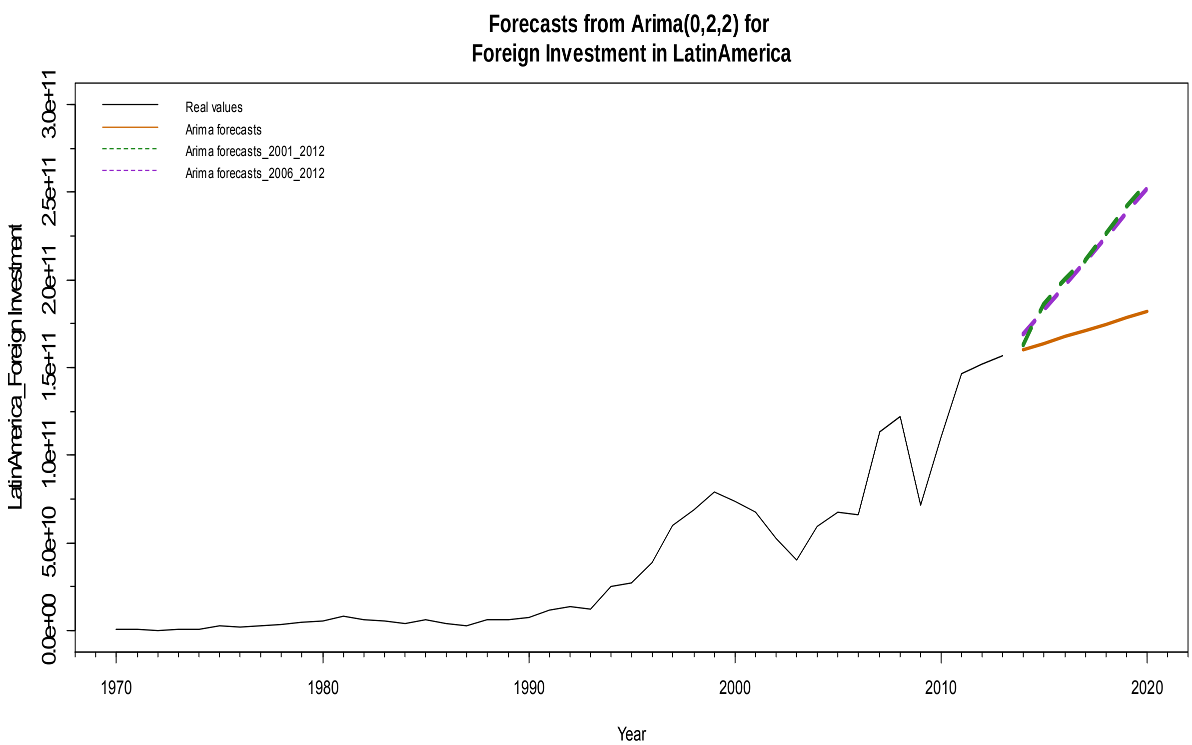


[Figure C.3.38] – Histogram and distribution of forecast residuals

<u>Remark</u>: For the following analysis the methodology gave an ARIMA model (0,2,3) but the results we get with that do not seem appropriate. I chose to continue the analysis with the model '(0,1,0) with drift' as suggested by the auto.arima function, which gives more reasonable results.

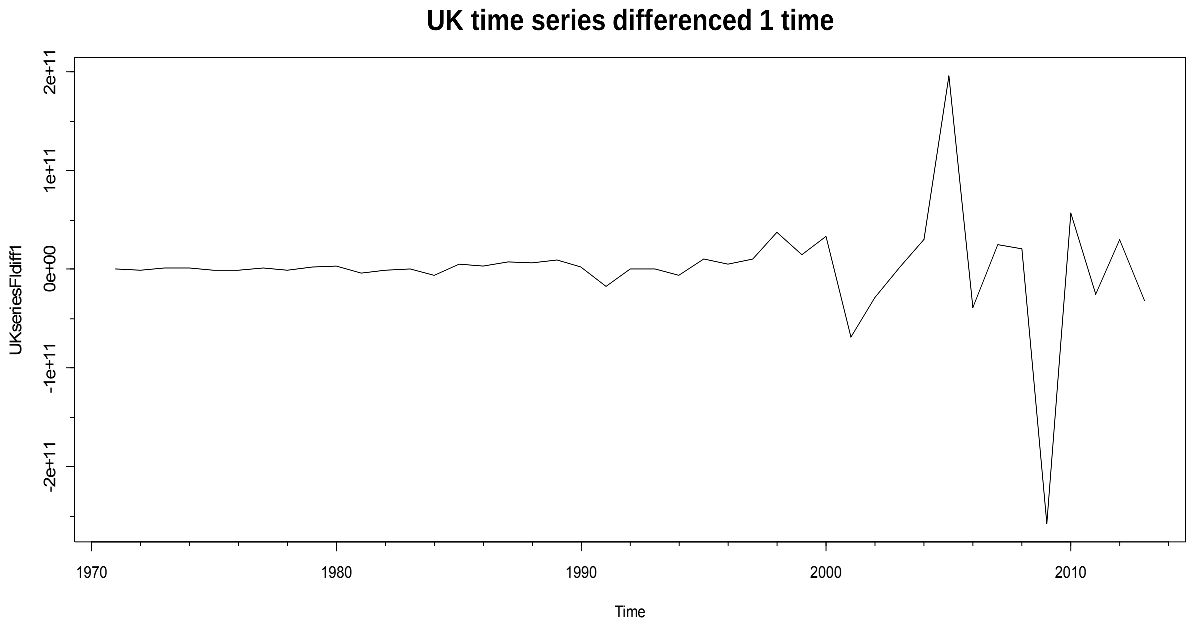


[Figure C.3.39] – Analysis for Latin America and Caribbean, Foreign Investment and the dataset up to 2008

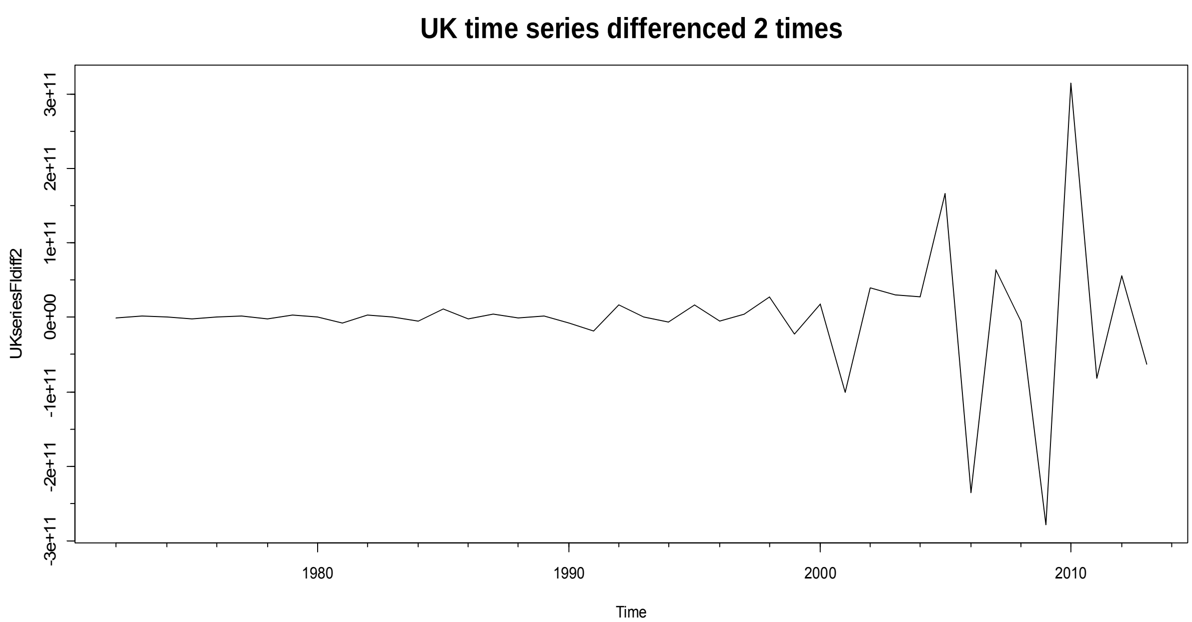


[Figure C.3.40] – Analyses for Latin America and Caribbean, Foreign Investment and the subsets 2001-2013 and 2006-2013

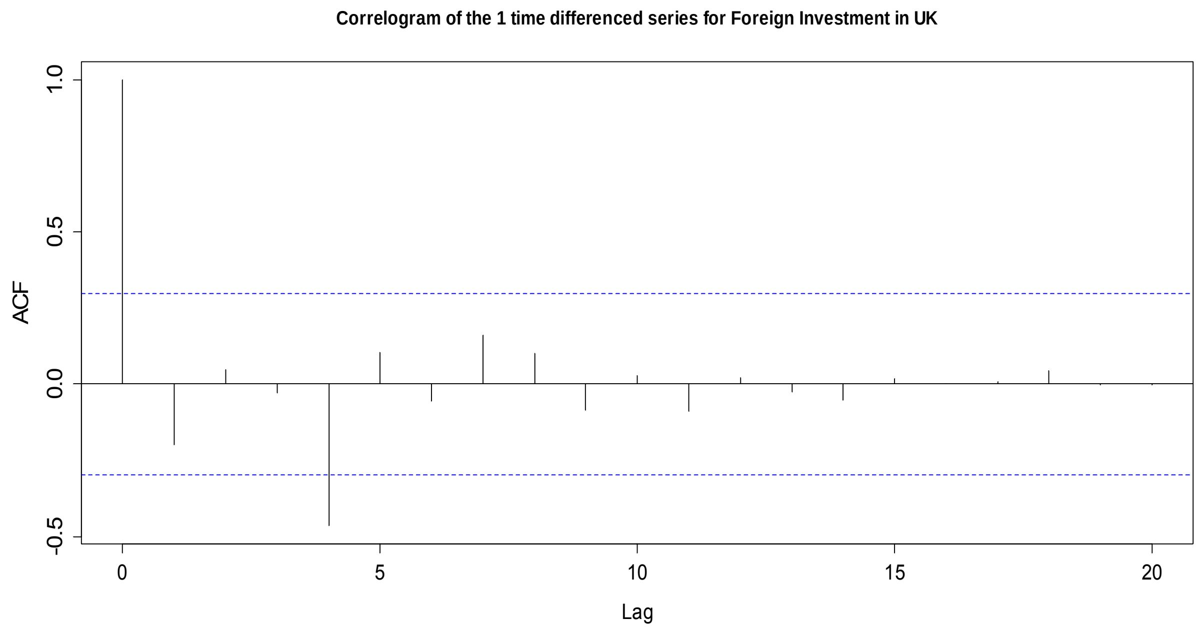
Foreign Investment – UK



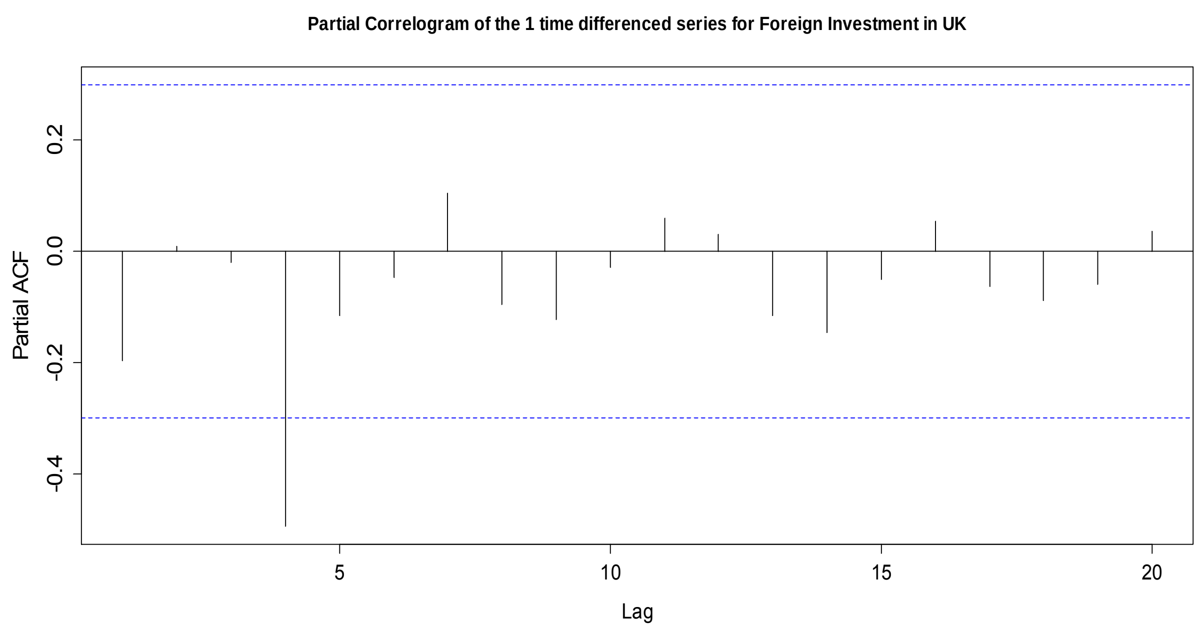
[Figure C.3.41] – One time differenced UK time series



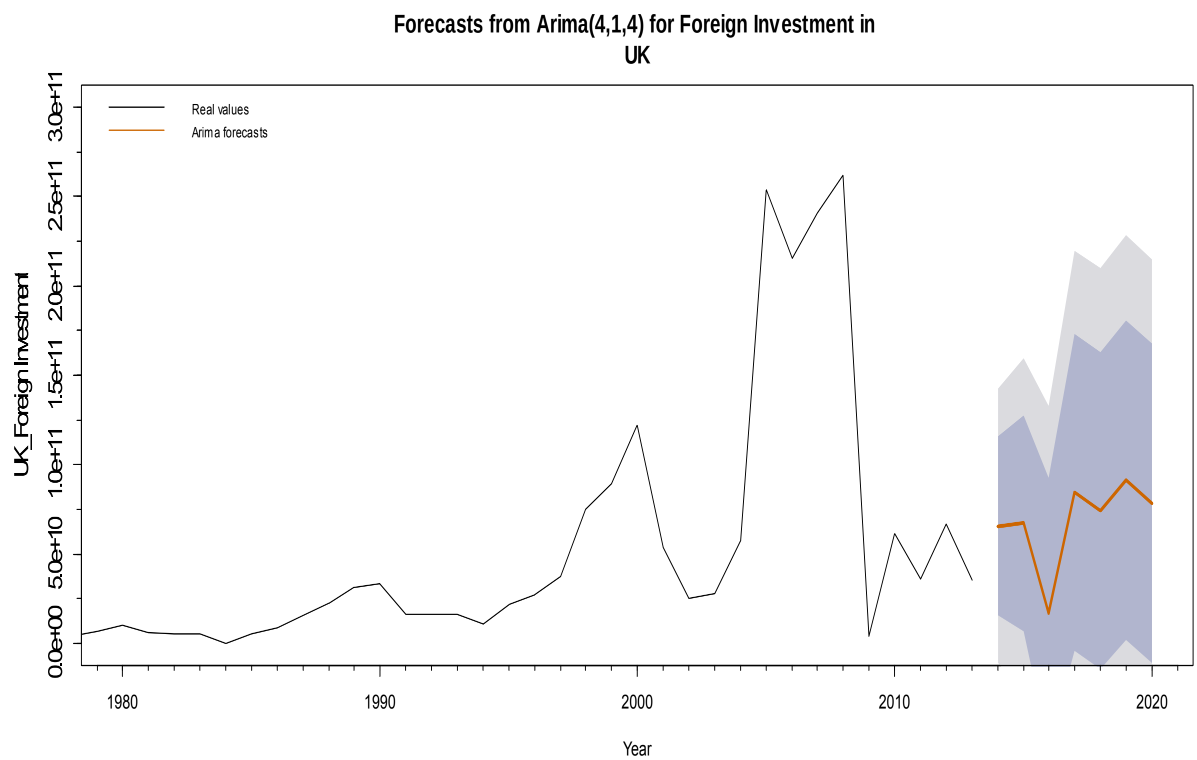
[Figure C.3.42] – Two times differenced UK time series



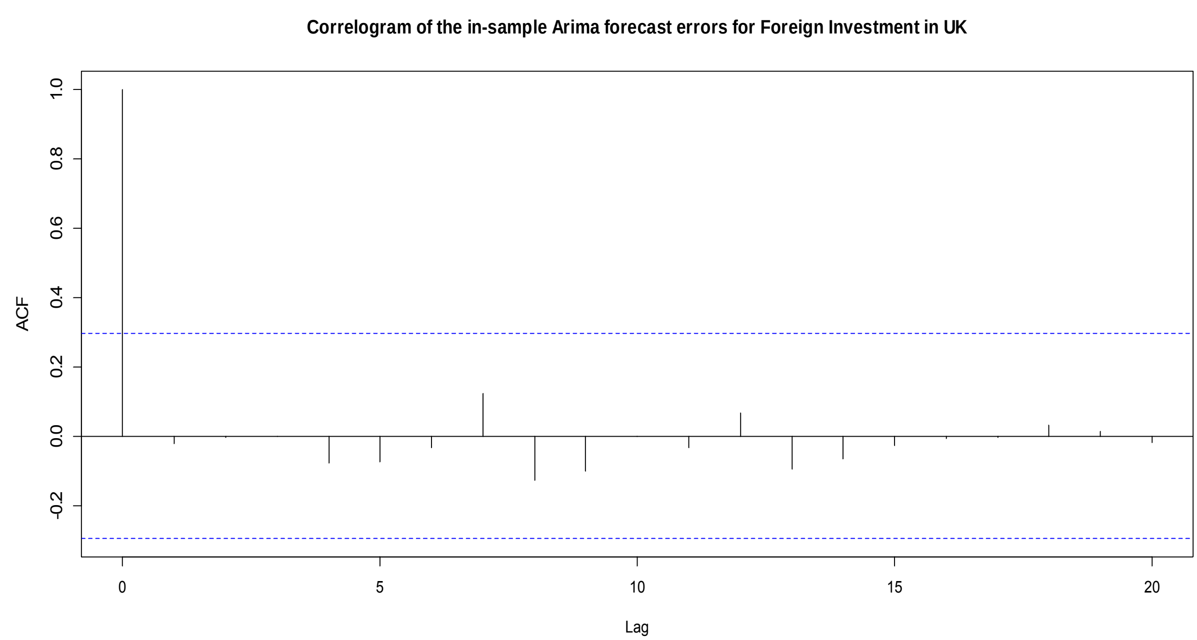
[Figure C.3.43] – Autocorrelogram (ACF) of the twice differenced UK time series



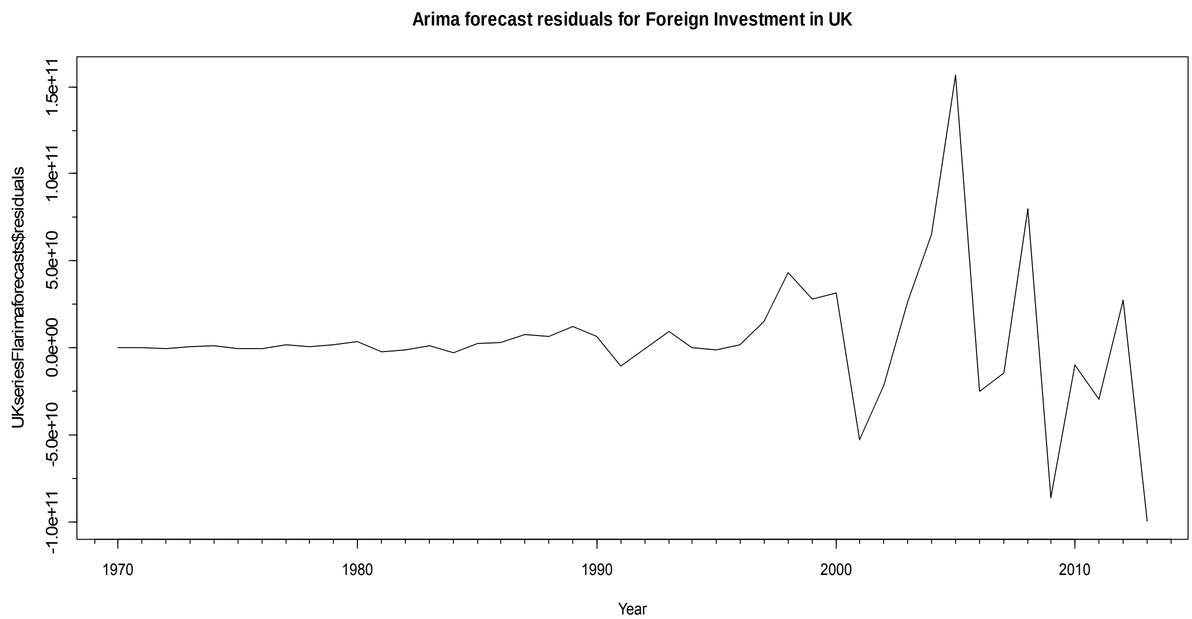
[Figure C.3.44] – Partial autocorrelogram (PACF) of the twice differenced UK time series



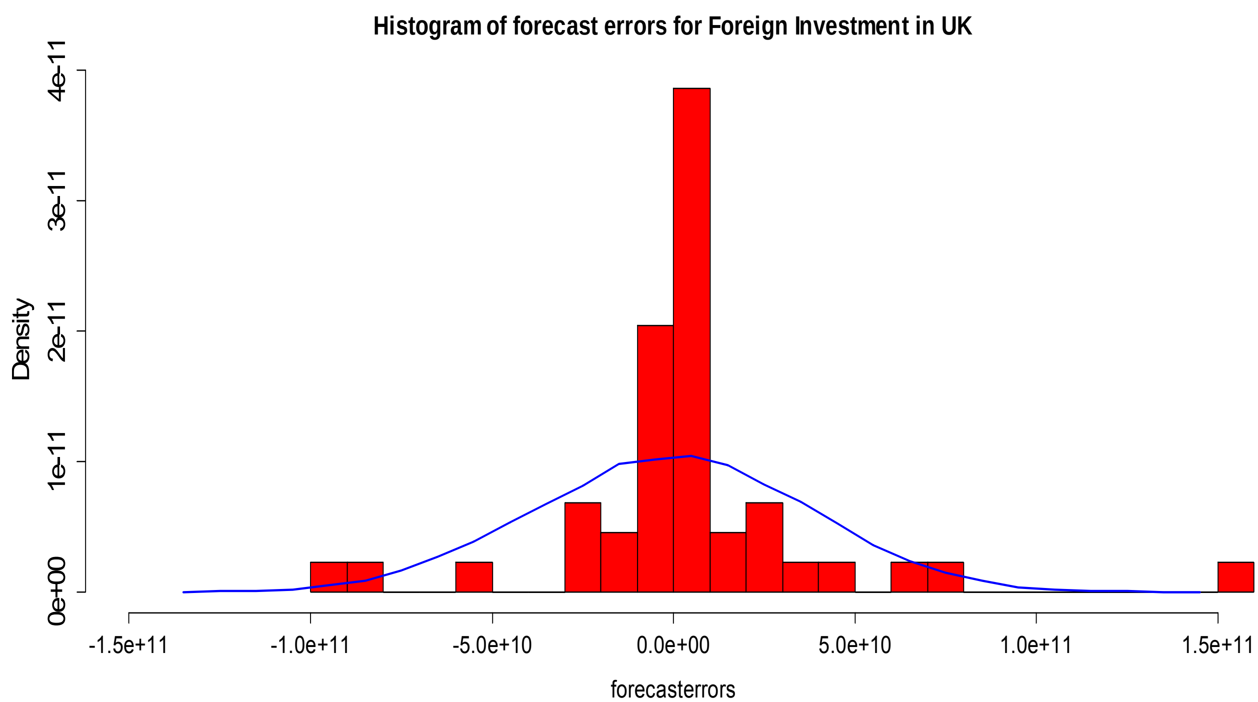
[Figure C.3.45] – Analysis for UK, Foreign Investment and whole dataset



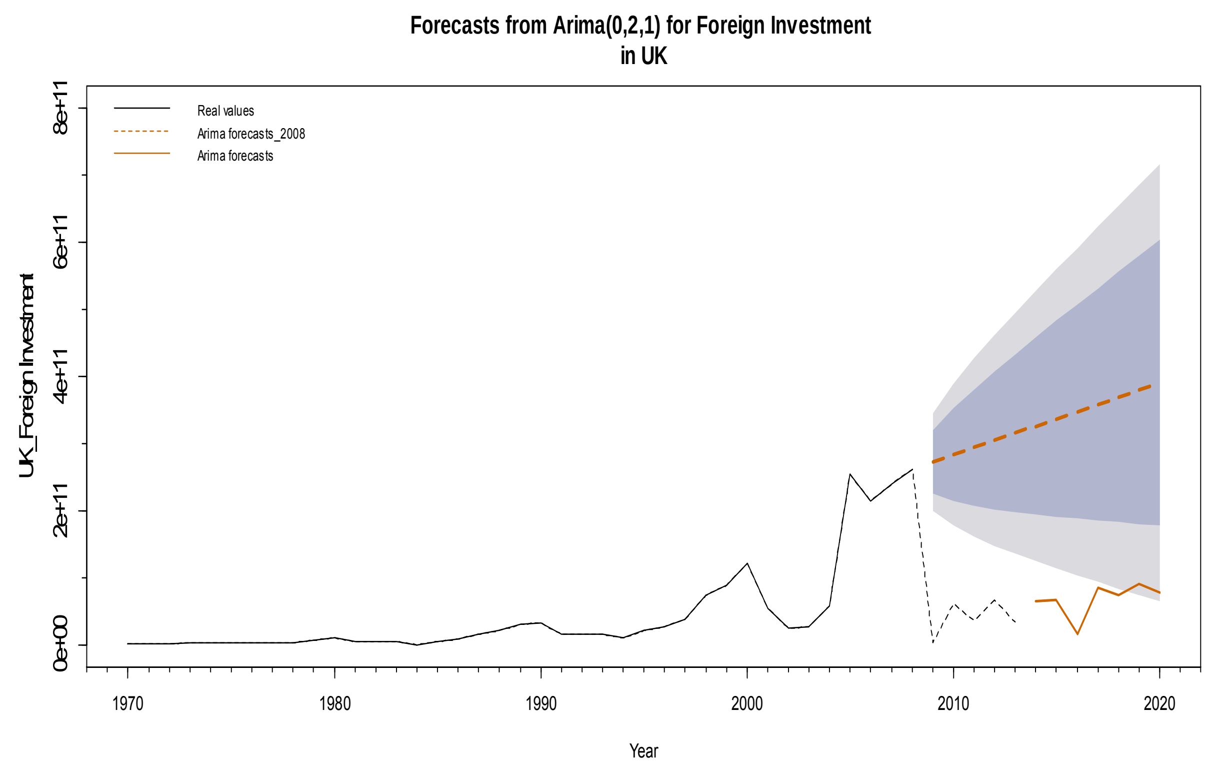
[Figure C.3.46] – Correlogram of in-sample errors of ARIMA forecasts



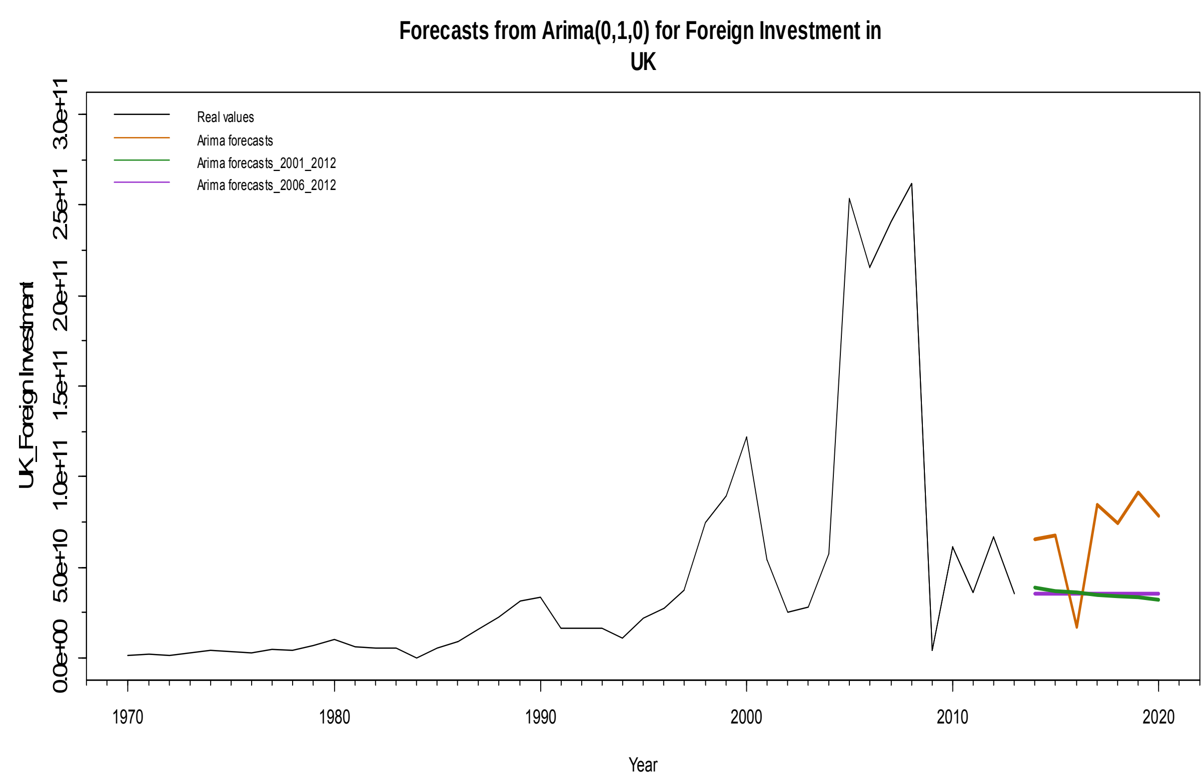
[Figure C.3.47] –Residuals of ARIMA forecasts



[Figure C.3.48] – Histogram and distribution of forecast residuals

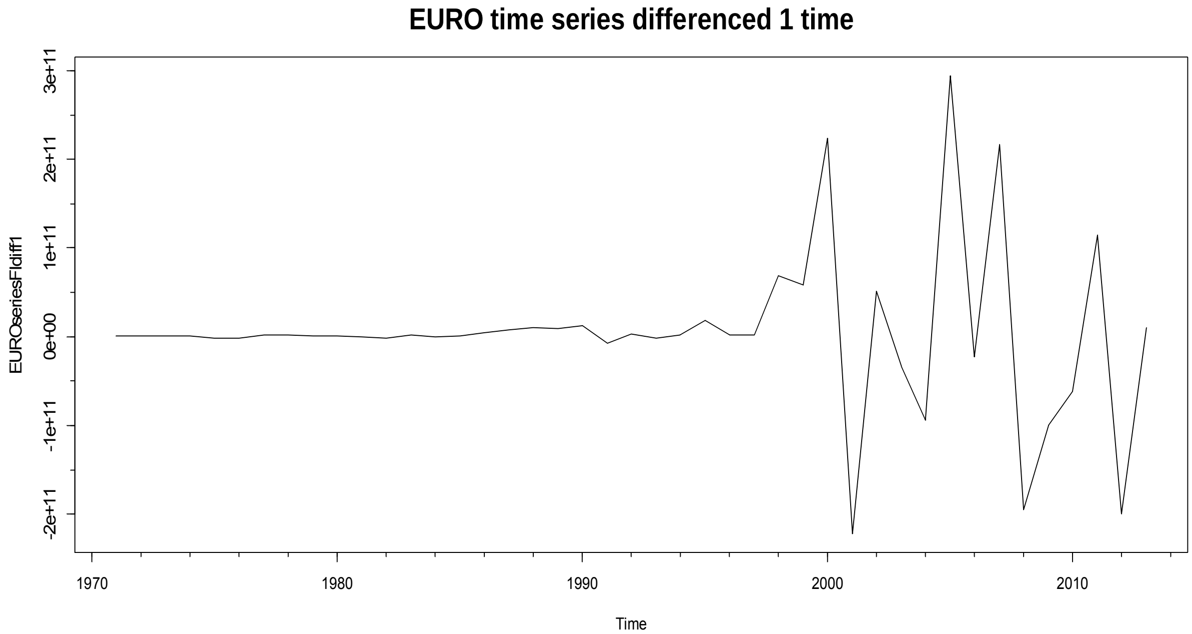


[Figure C.3.49] – Analysis for UK, Foreign Investment and the dataset up to 2008

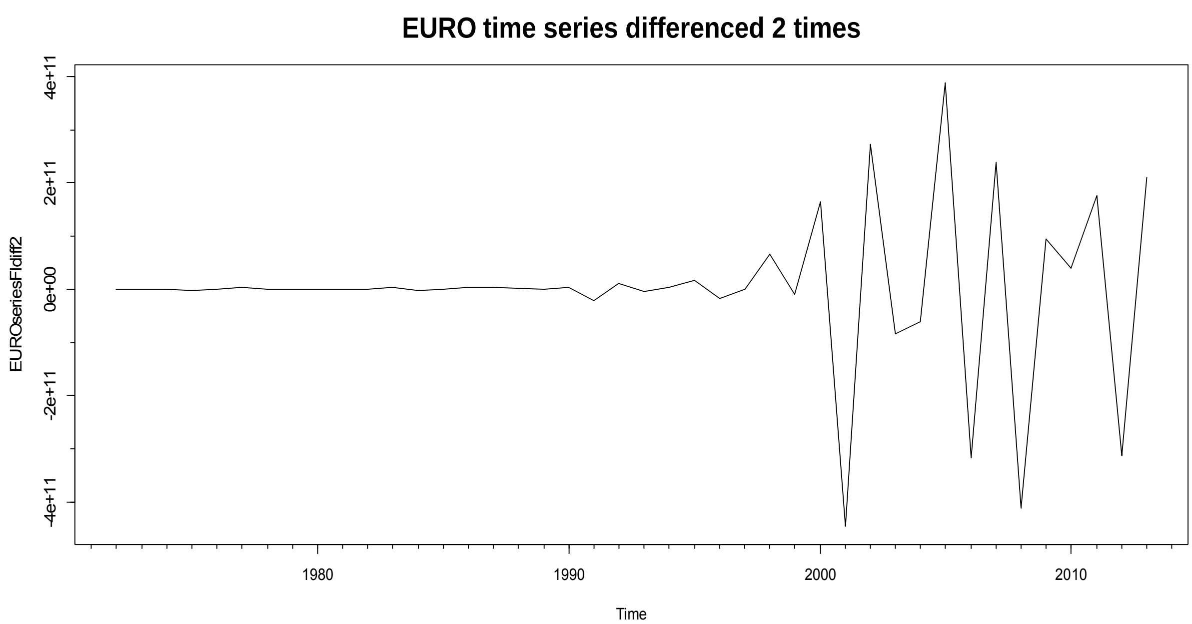


[Figure C.3.50] – Analyses for UK, Foreign Investment and the subsets 2001-2013 and 2006-2013

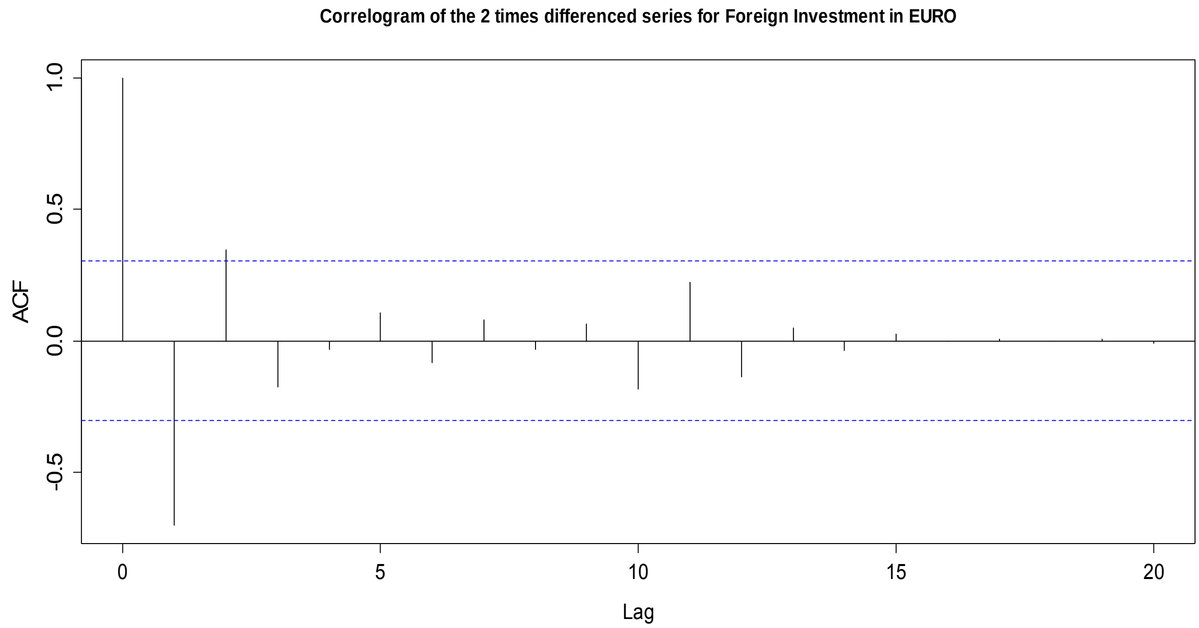
Foreign Investment – EURO zone



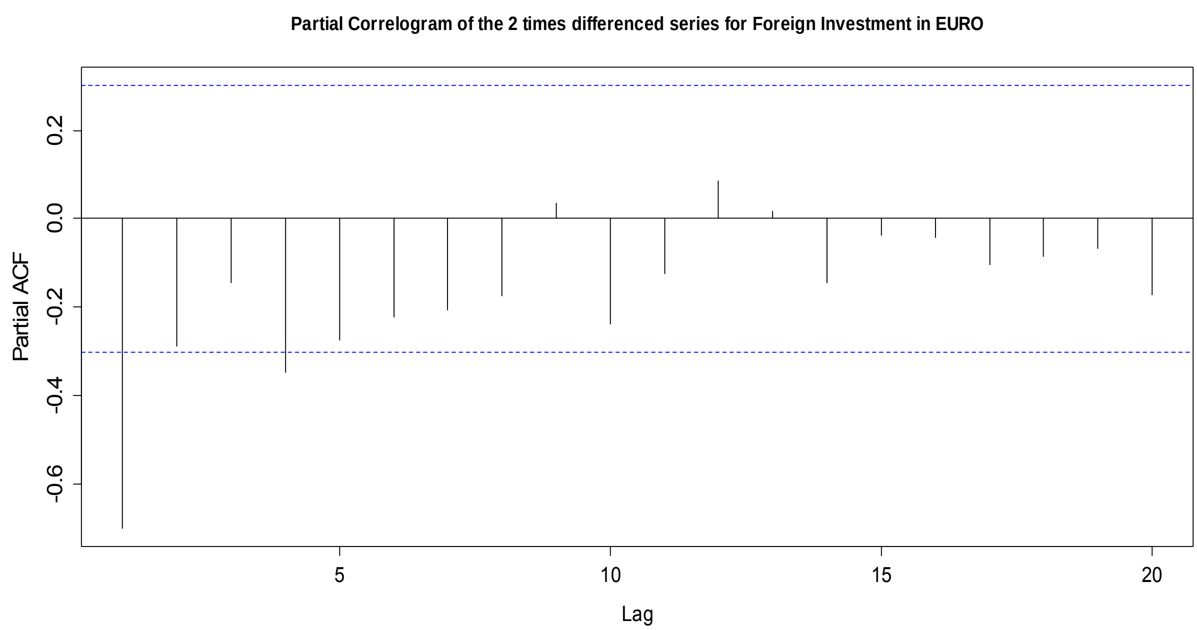
[Figure C.3.51] – One time differenced EURO zone time series



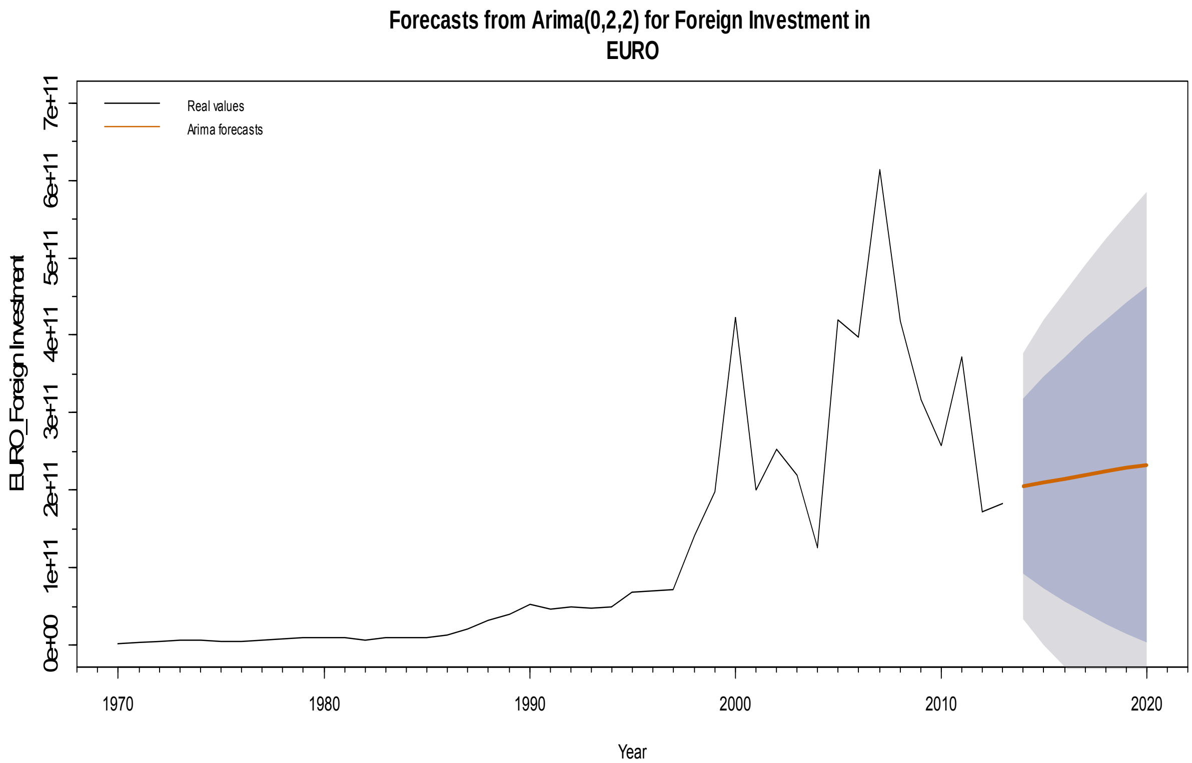
[Figure C.3.52] – Two times differenced EURO zone time series



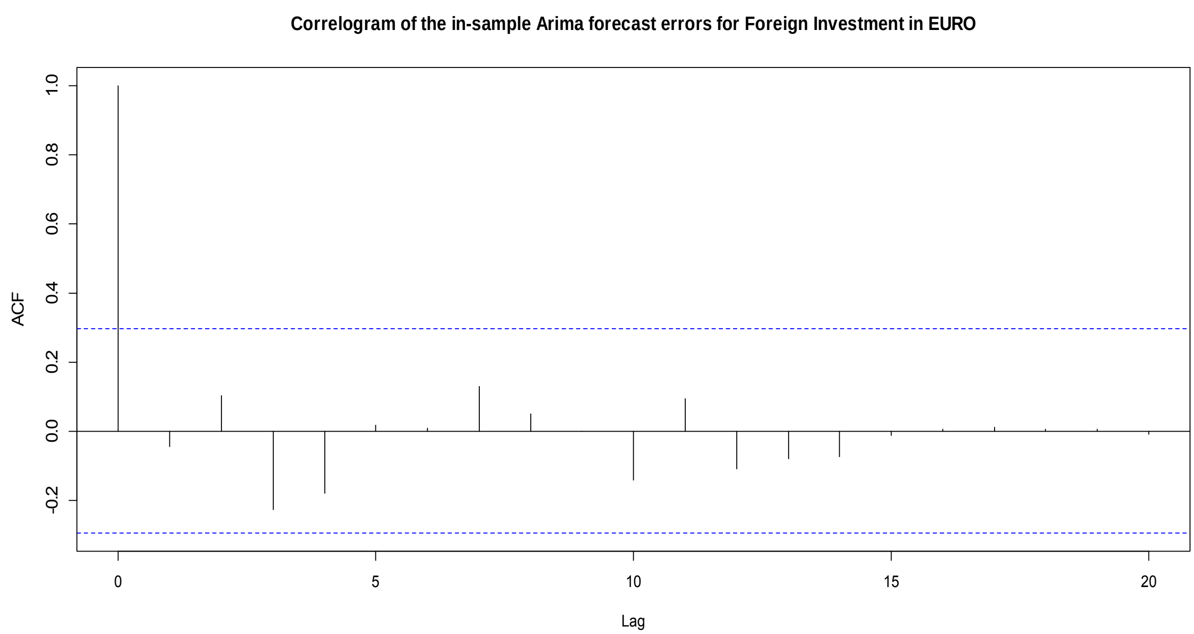
[Figure C.3.53] – Autocorrelogram (ACF) of the twice differenced EURO zone time series



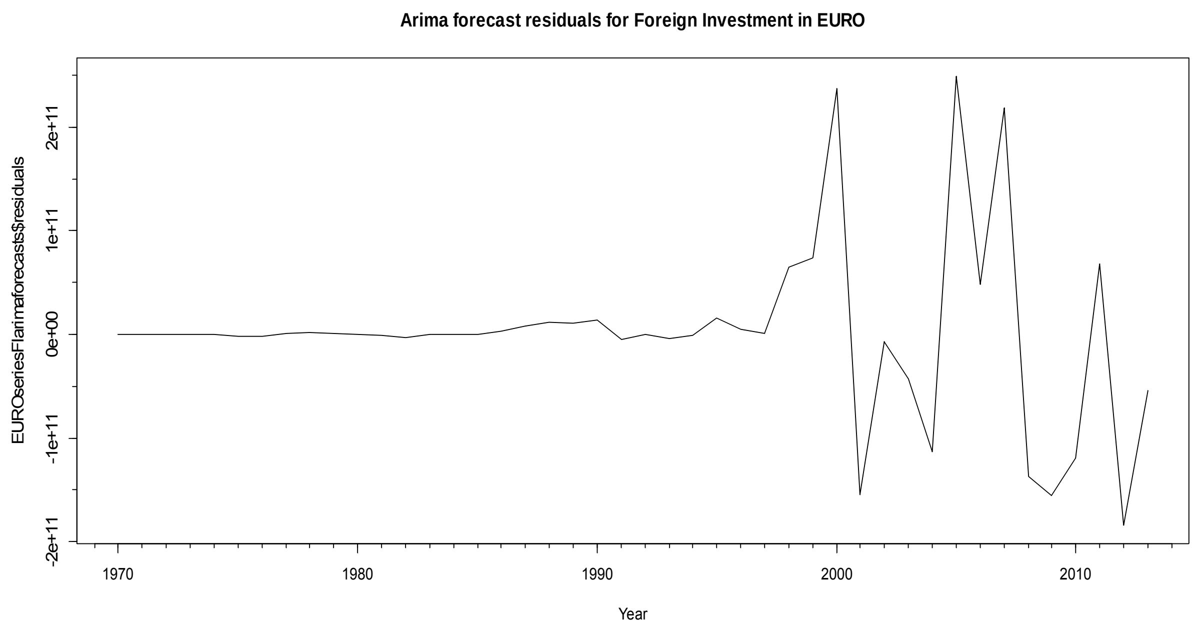
[Figure C.3.54] – Partial autocorrelogram (PACF) of the twice differenced EURO zone time series



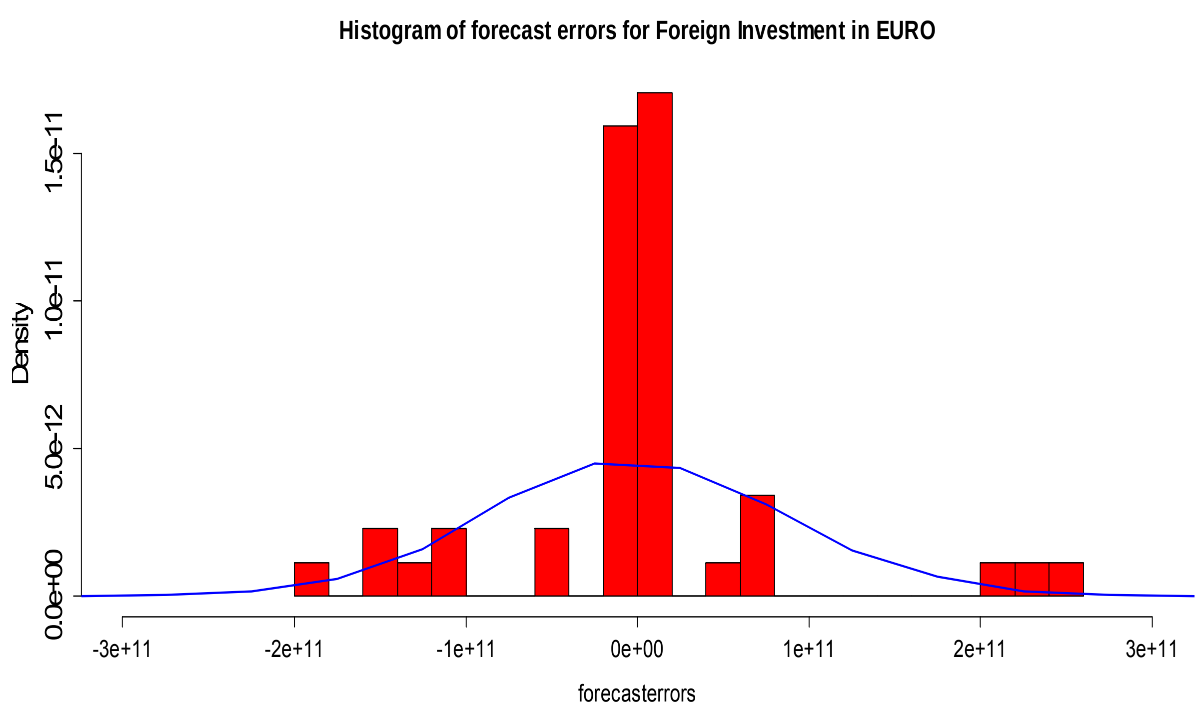
[Figure C.3.55] – Analysis for EURO zone, Foreign Investment and whole dataset



[Figure C.3.56] – Correlogram of in-sample errors of ARIMA forecasts

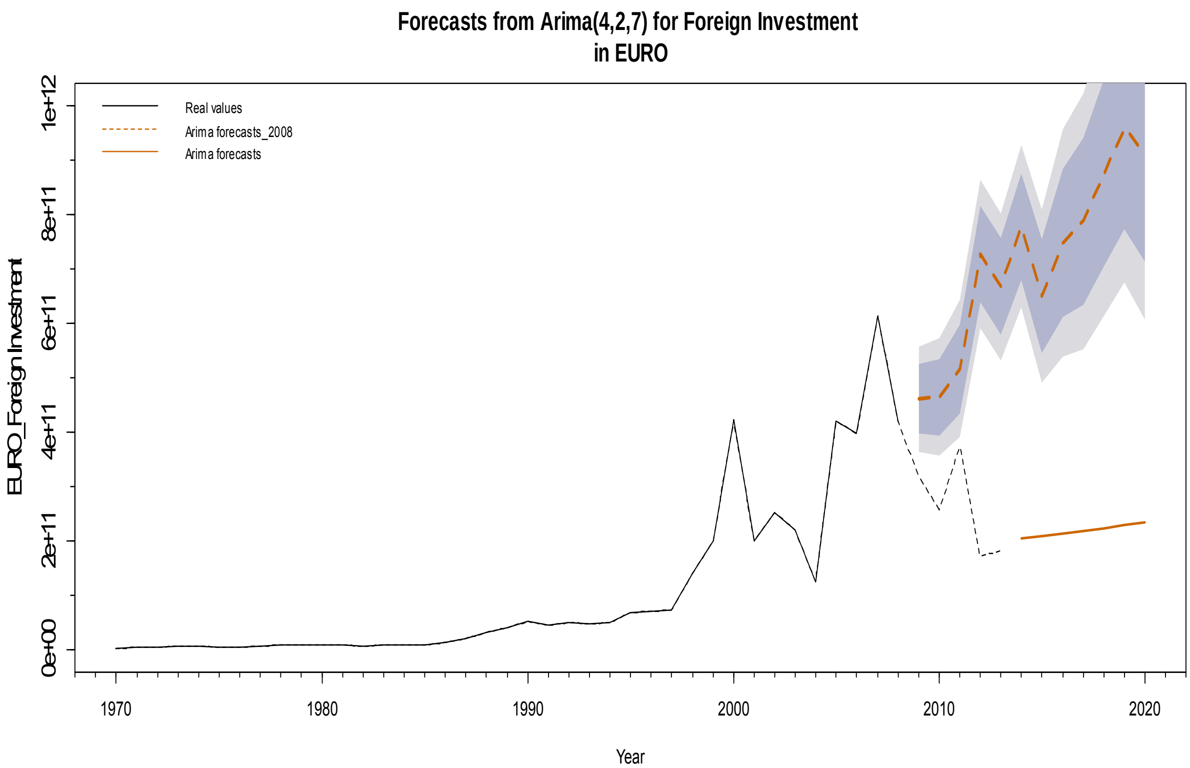


[Figure C.3.57] –Residuals of ARIMA forecasts

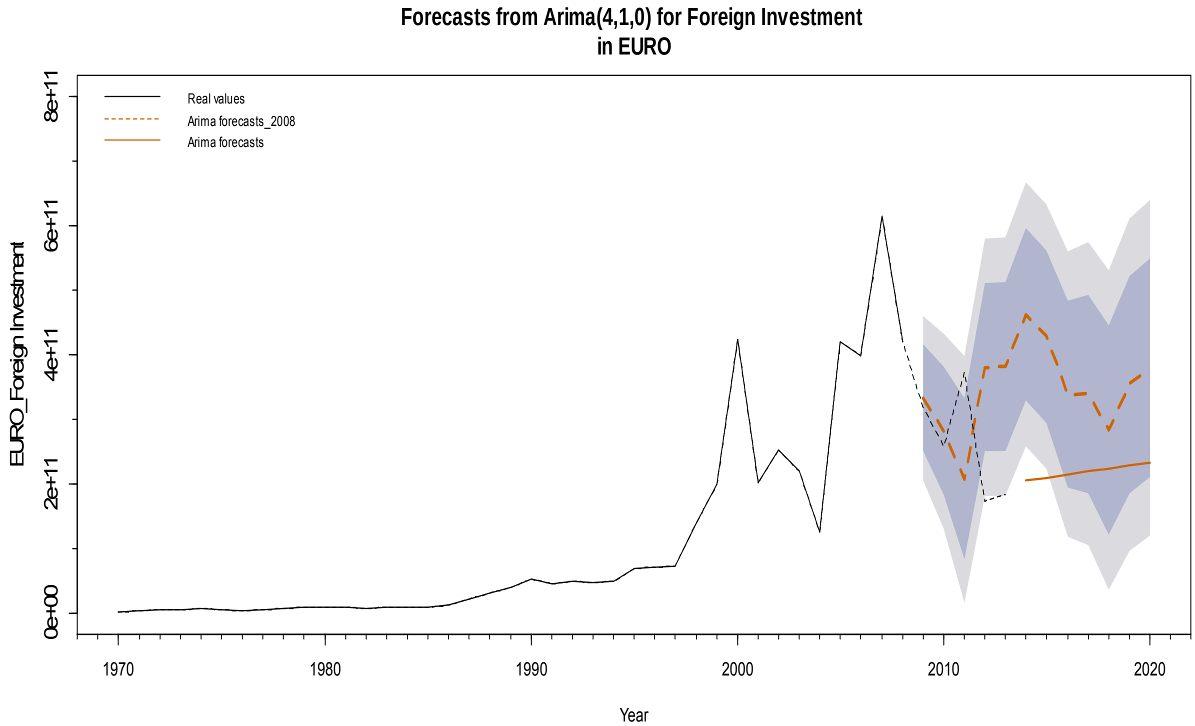


[Figure C.3.58] – Histogram and distribution of forecast residuals

<u>Remark</u>: If we use the 2 times differenced times series the methodology gives a model Arima(4,2,7) which gives predictions similar to Holt's (Figure C.3.59.a), but very far from the real values. But if I apply the methodology in the 1 time differenced time series we get the model Arima(4,1,0) which gives an interesting curve, which follows better the real values (Figure C.3.59.b).

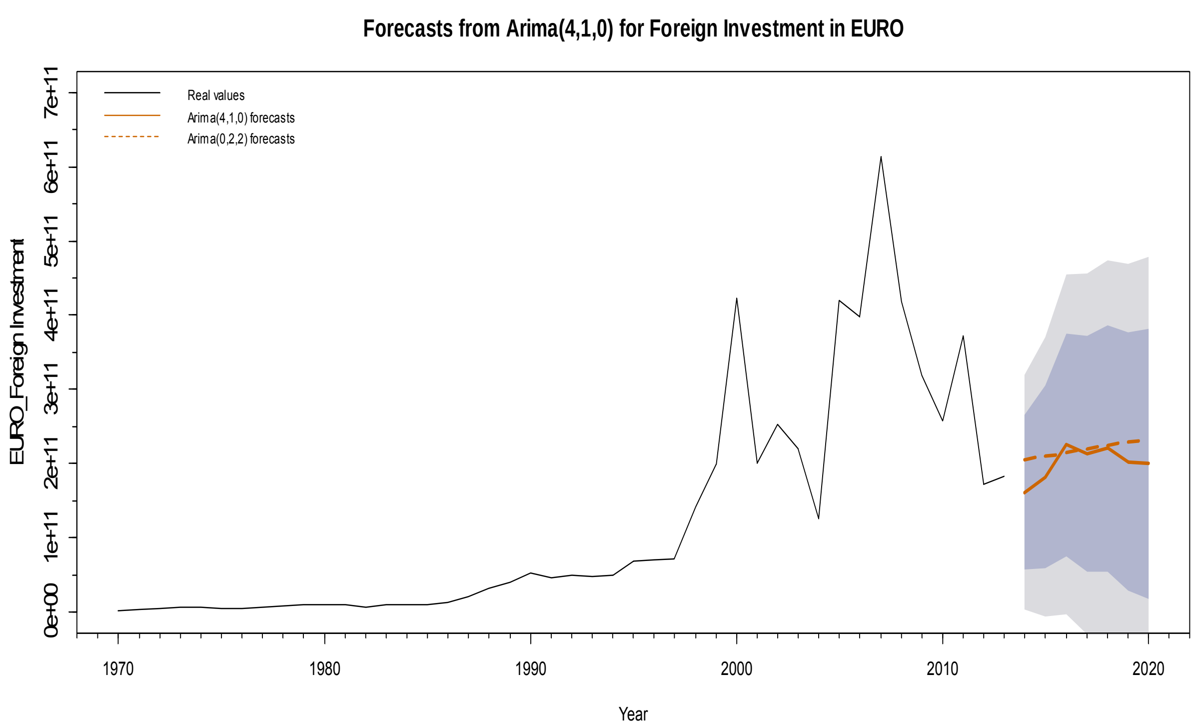


[Figure C.3.59.a] – Analysis for EURO zone, Foreign Investment and the dataset up to 2008

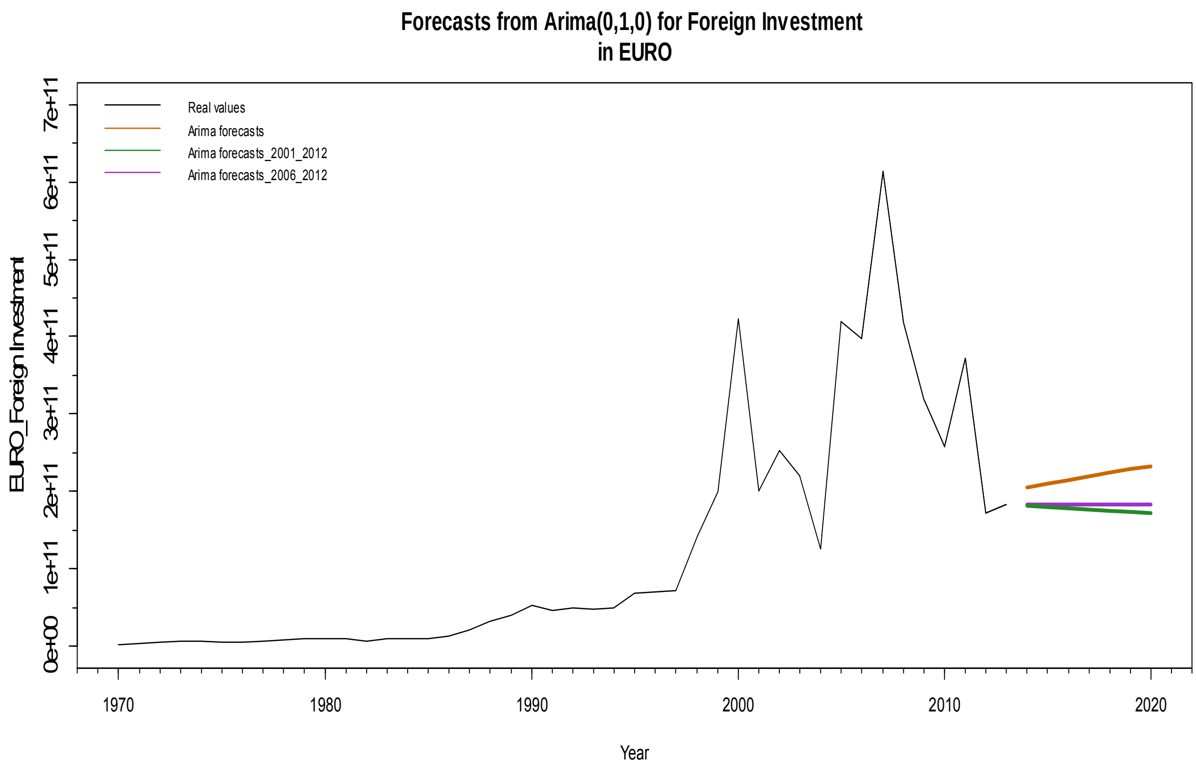


[Figure C.3.59.b] – Analysis for EURO zone, Foreign Investment and the dataset up to 2008

Remark: Then I went back and applied Arima(4,1,0) to whole dataset (Figure C.3.59.c)

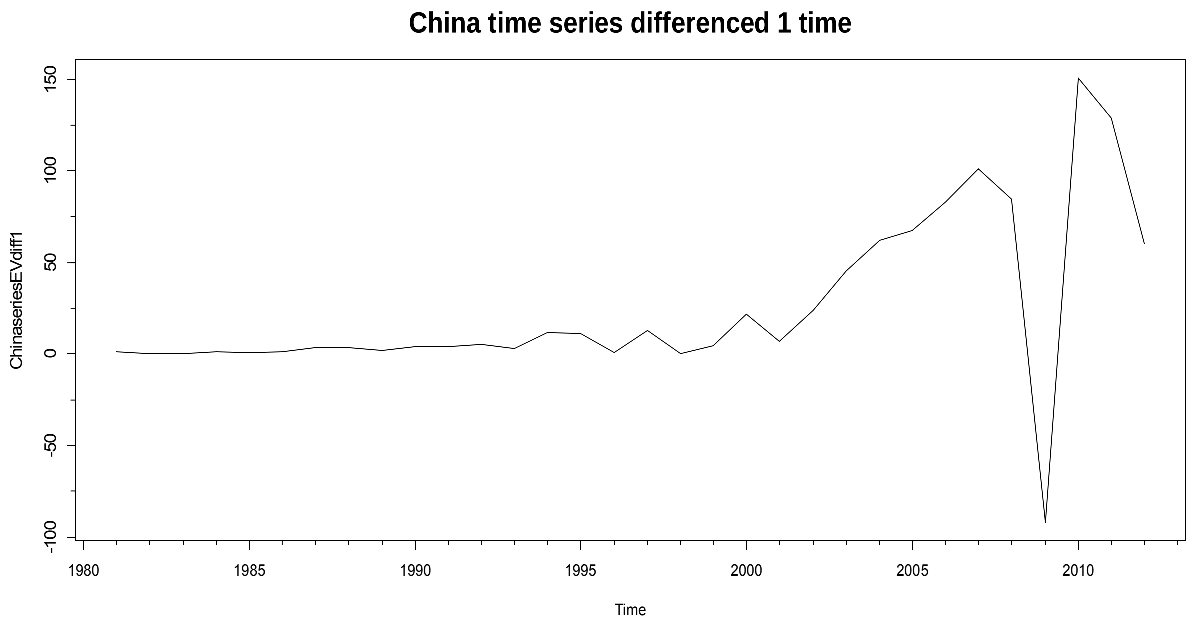


[Figure C.3.59.c] – Analysis for EURO zone, Foreign Investment and the whole dataset with ARIMA(4,1,0)

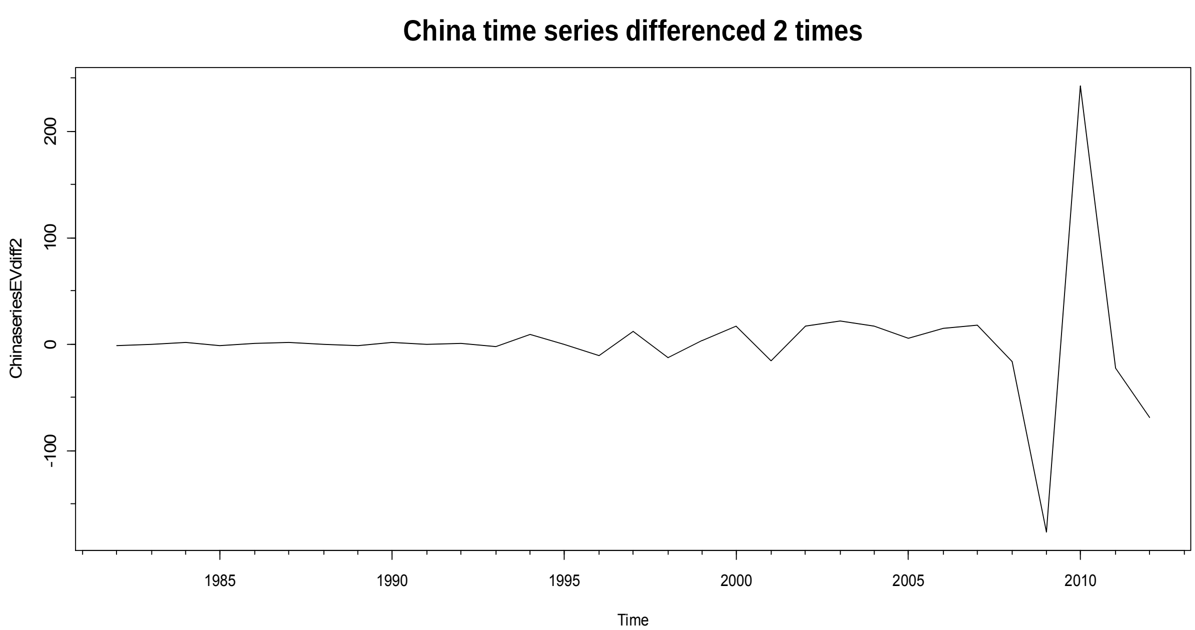


[Figure C.3.60] – Analyses for EURO zone, Foreign Investment and the subsets 2001-2013 and 2006-2013

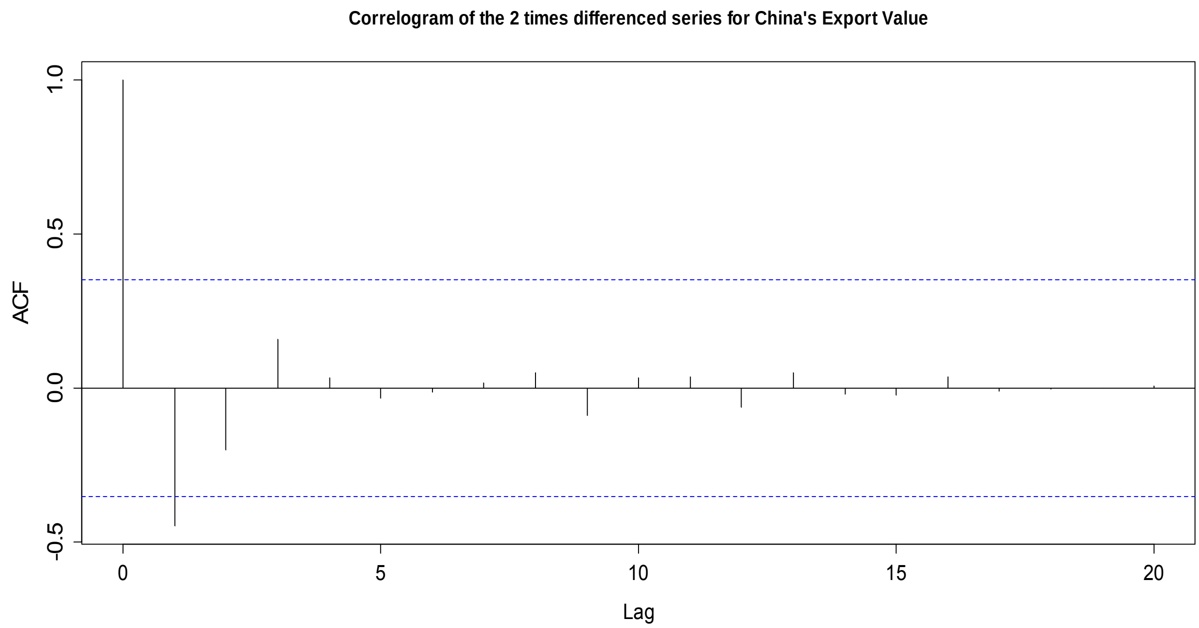
Export Value – China



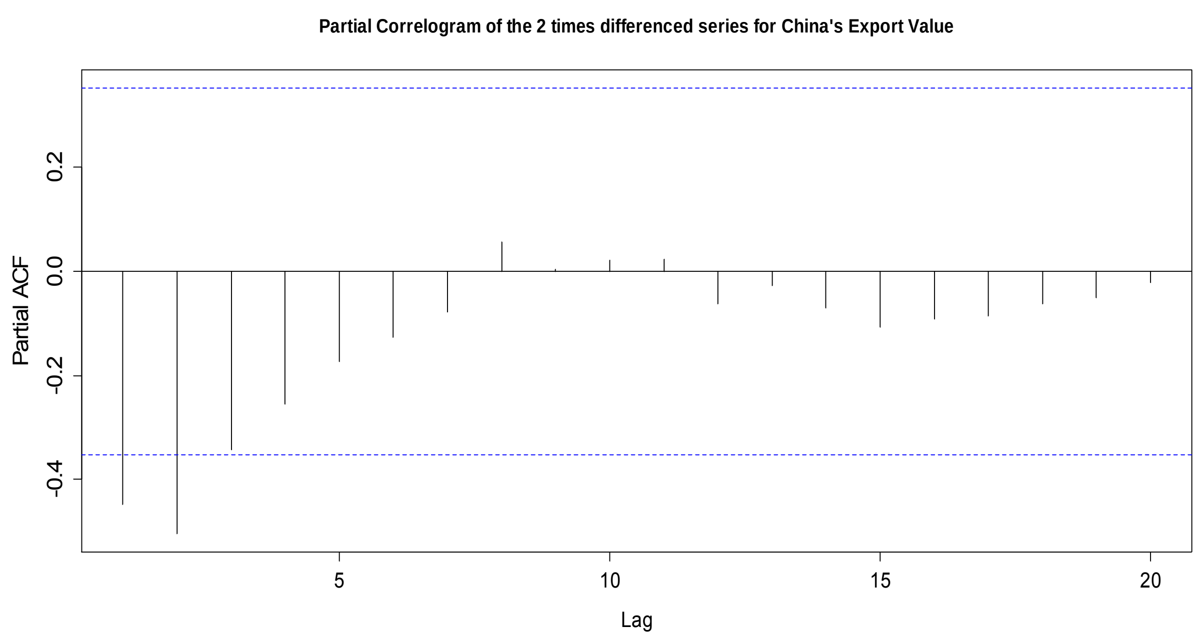
[Figure C.3.61] – One time differenced China time series



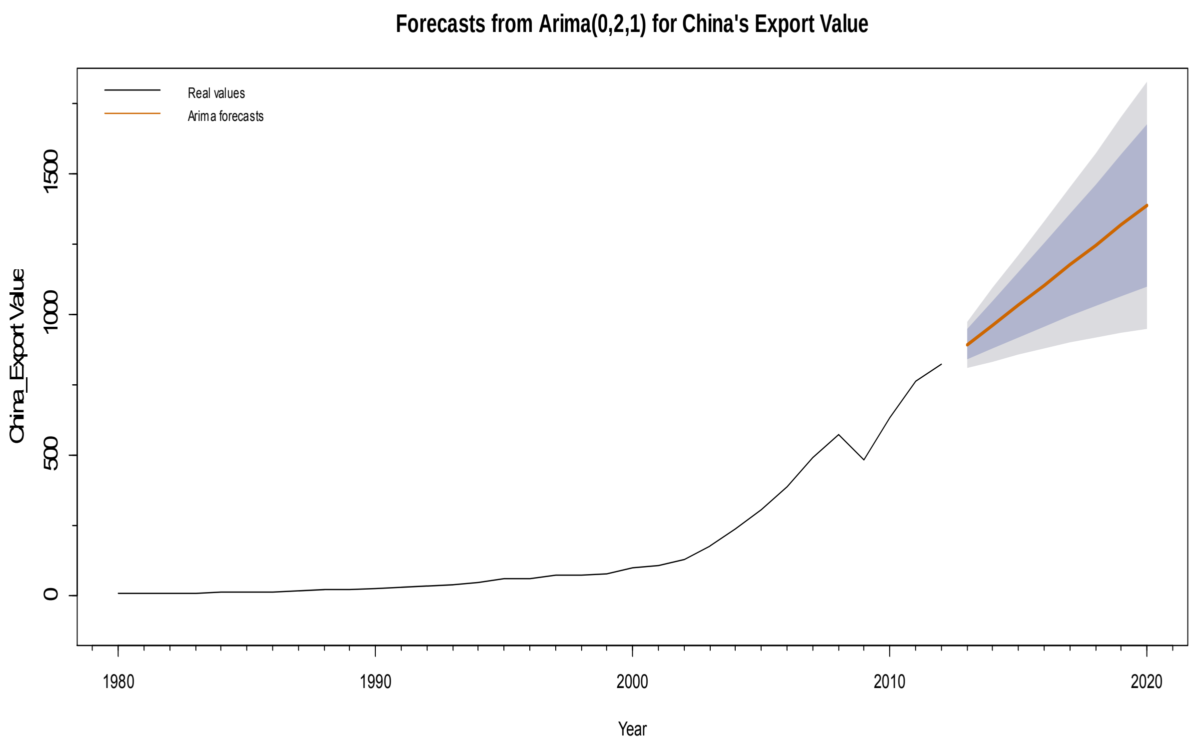
[Figure C.3.62] – Two times differenced China time series



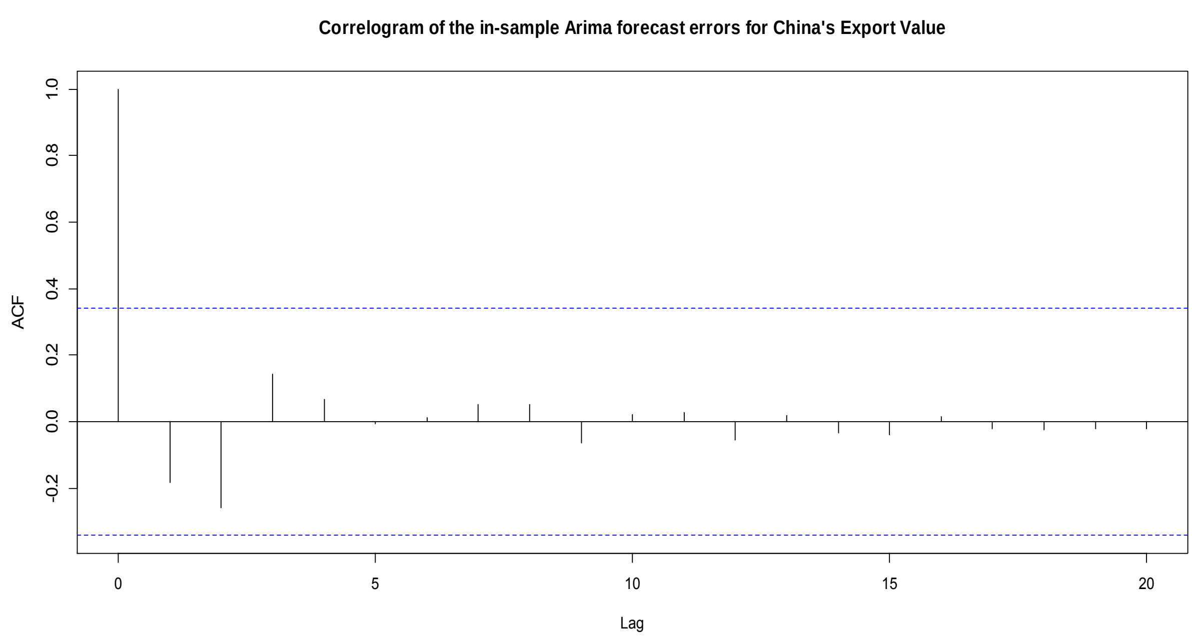
[Figure C.3.63] – Autocorrelogram (ACF) of the twice differenced China time series



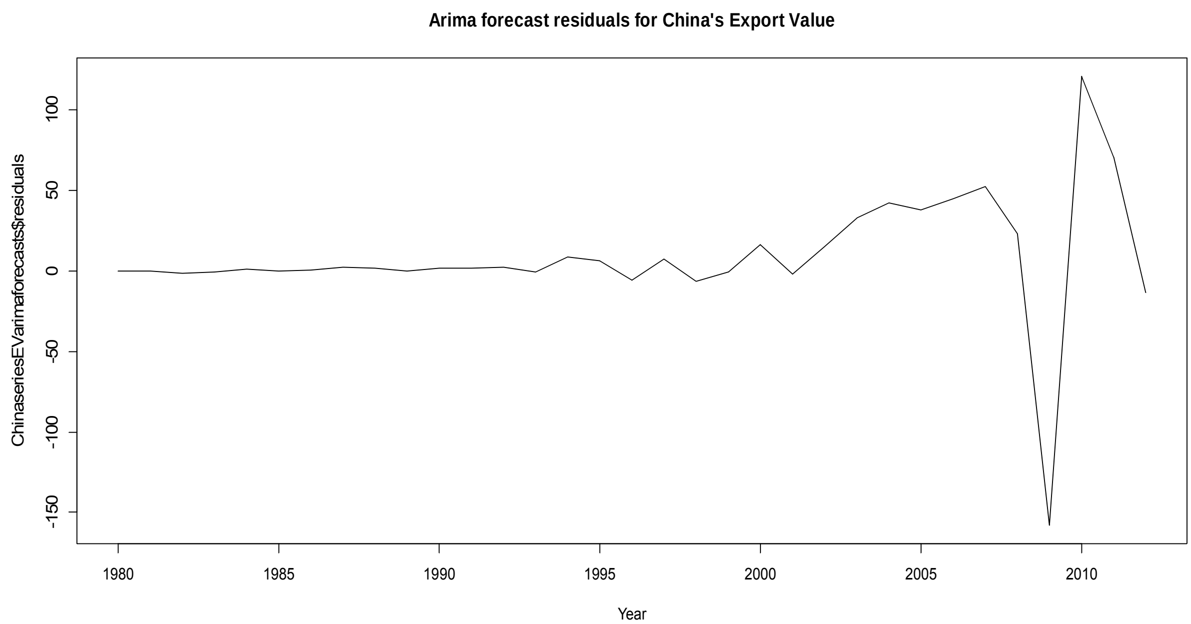
[Figure C.3.64] – Partial autocorrelogram (PACF) of the twice differenced China time series



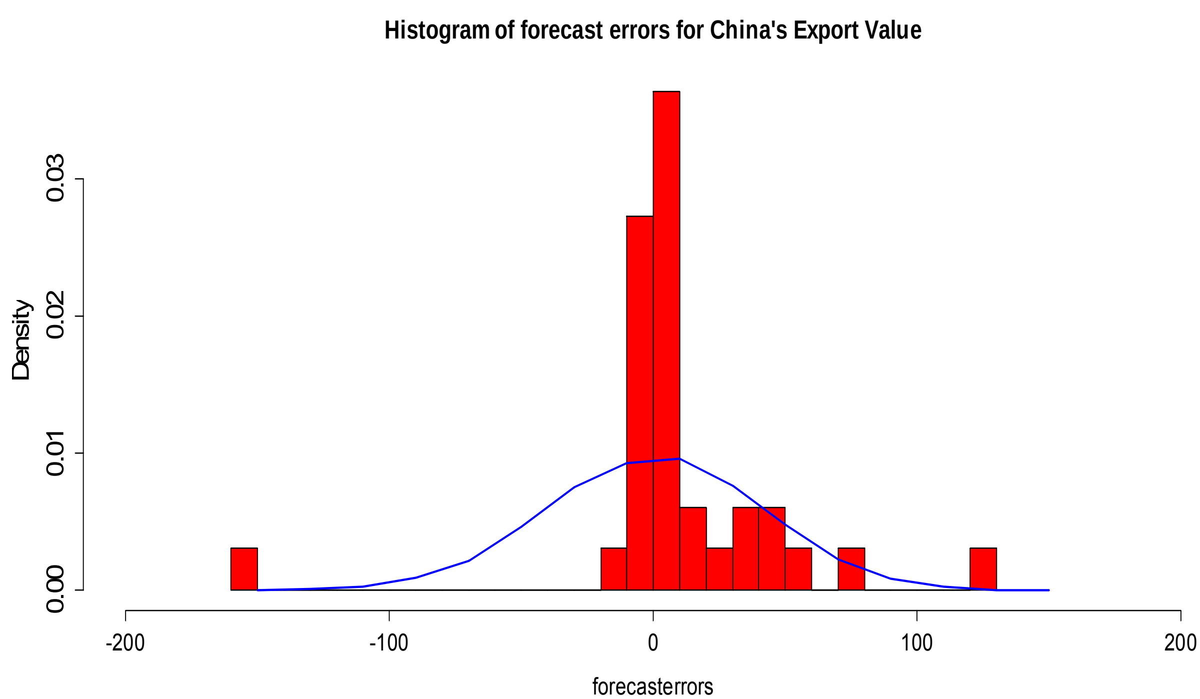
[Figure C.3.65] – Analysis for China, Export Value and whole dataset



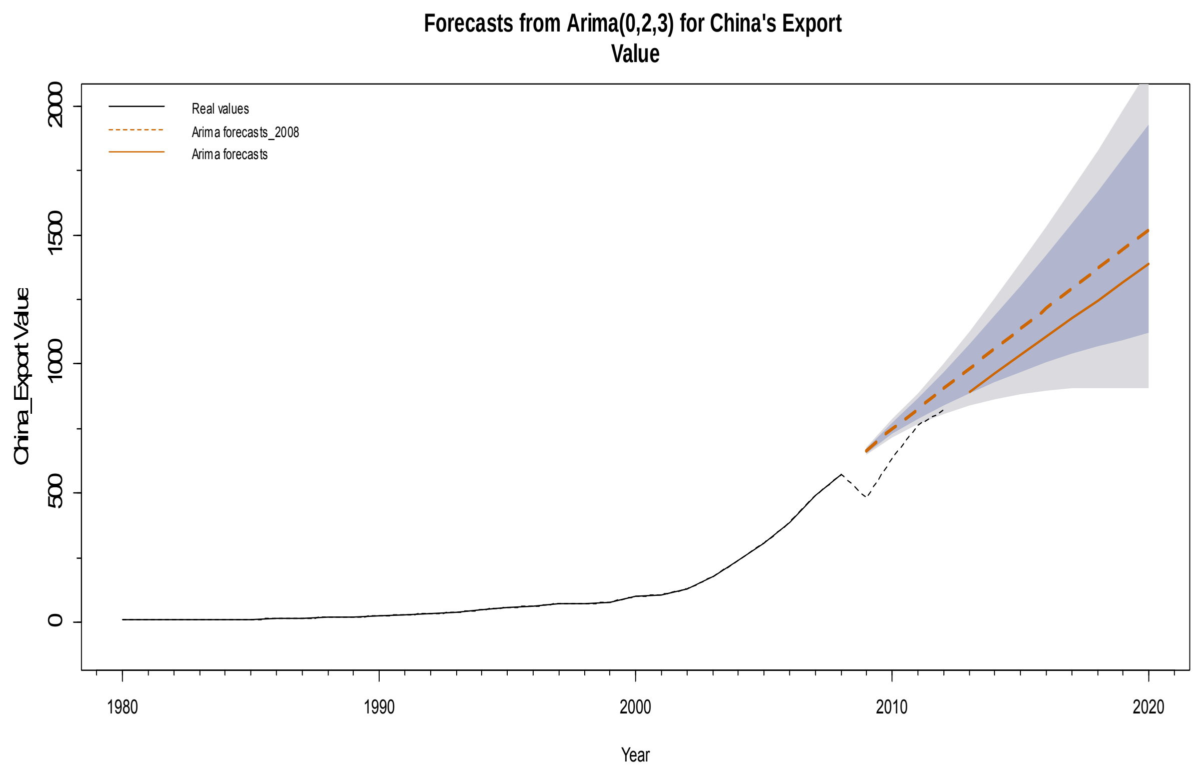
[Figure C.3.66] – Correlogram of in-sample errors of ARIMA forecasts



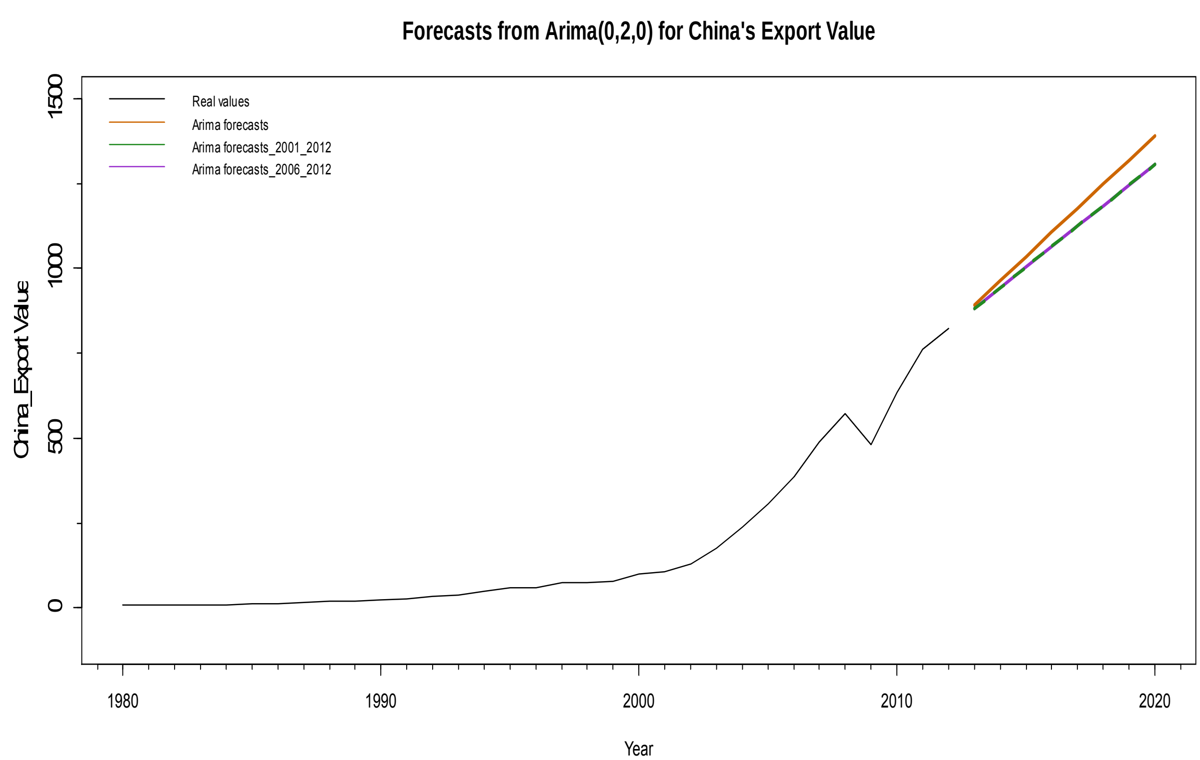
[Figure C.3.67] –Residuals of ARIMA forecasts



[Figure C.3.68] – Histogram and distribution of forecast residuals

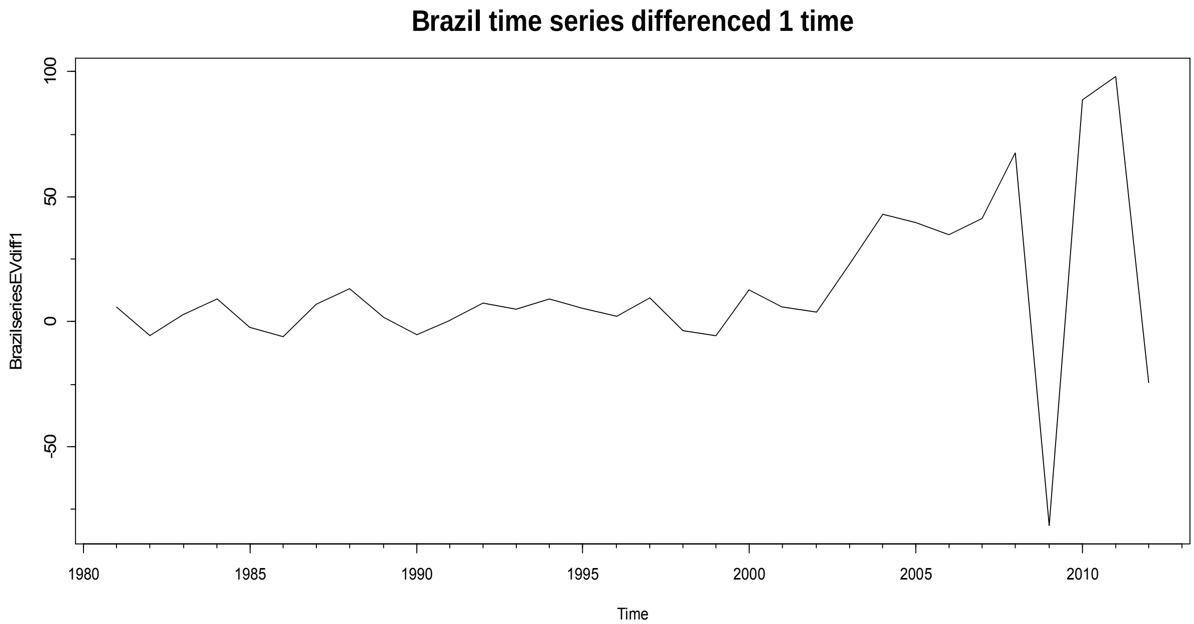


[Figure C.3.69] – Analysis for China, Export Value and the dataset up to 2008

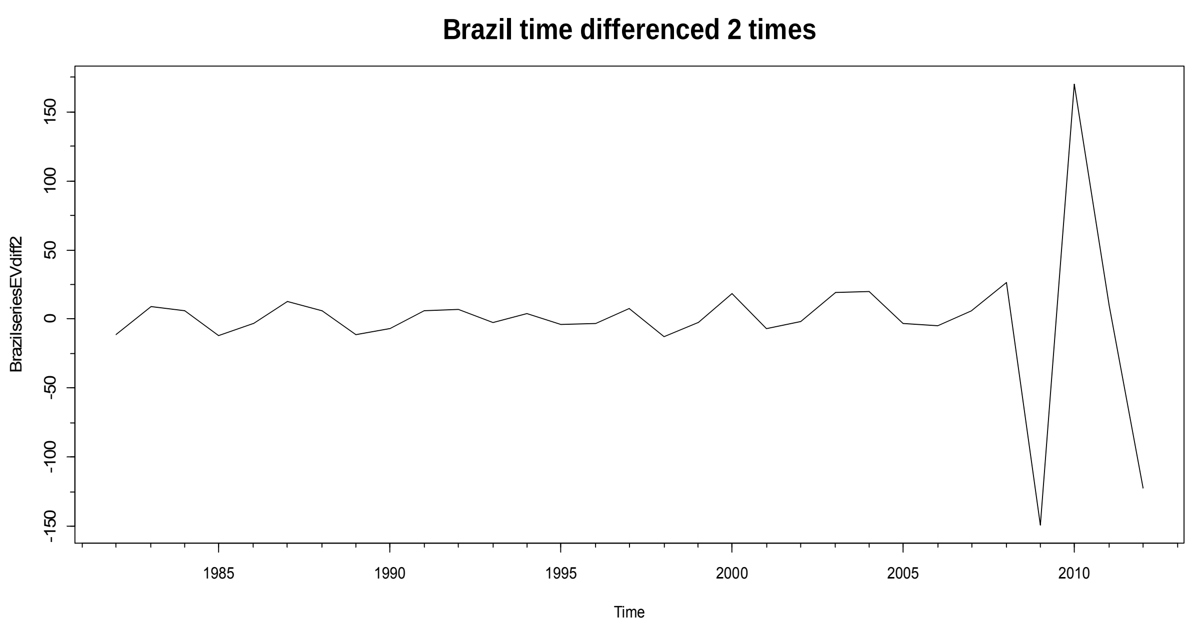


[Figure C.3.70] – Analyses for China, Export Value and the subsets 2001-2013 and 2006-2013

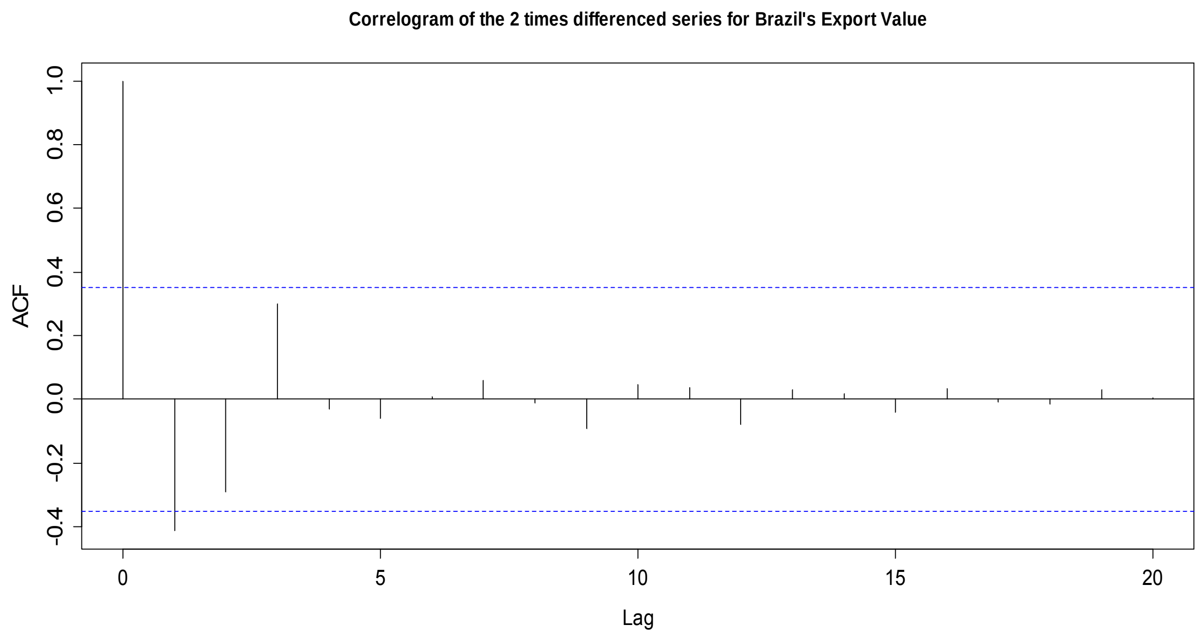
Export Value – Brazil



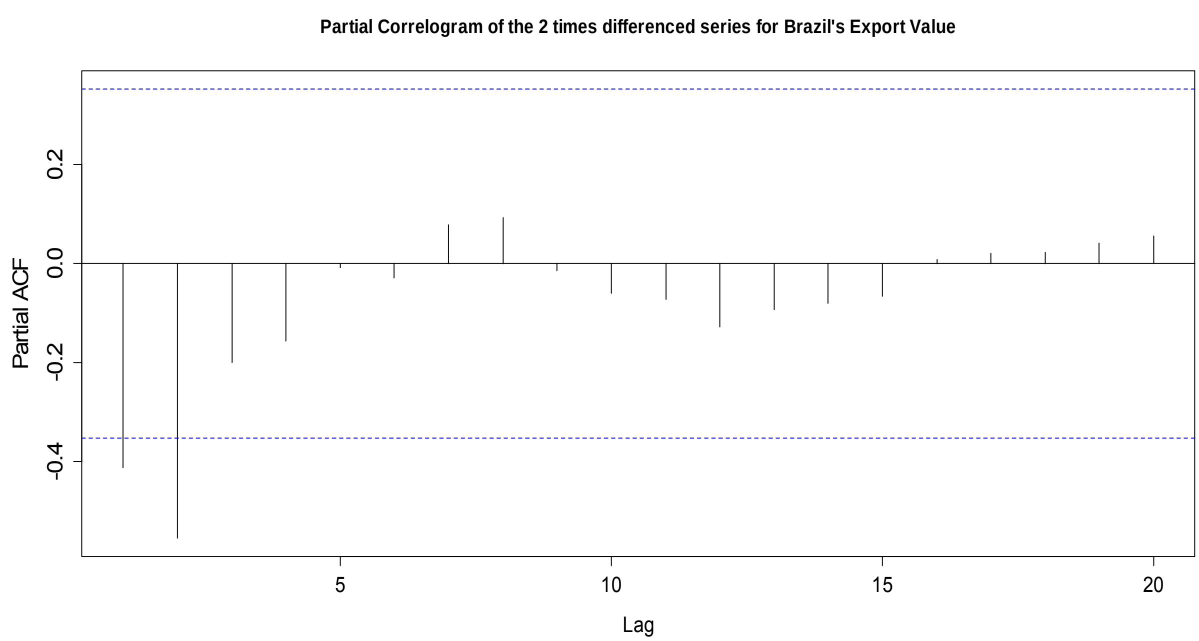
[Figure C.3.71] – One time differenced Brazil time series



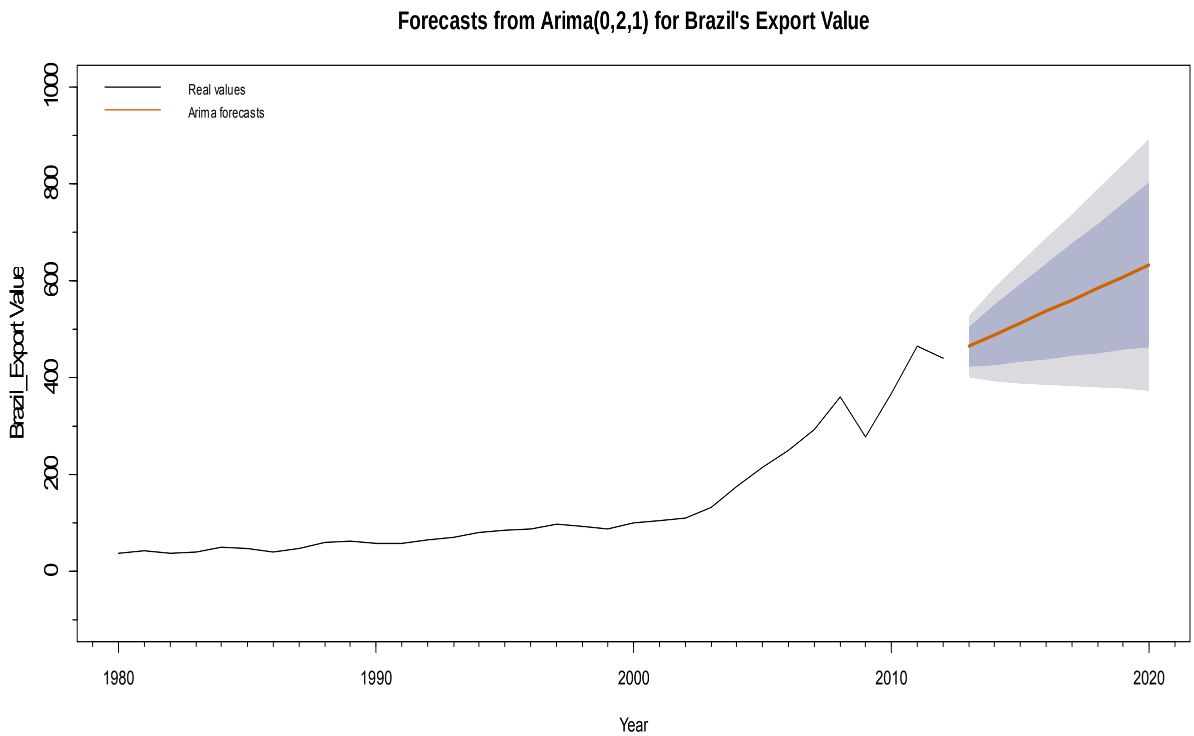
[Figure C.3.72] – Two times differenced Brazil time series



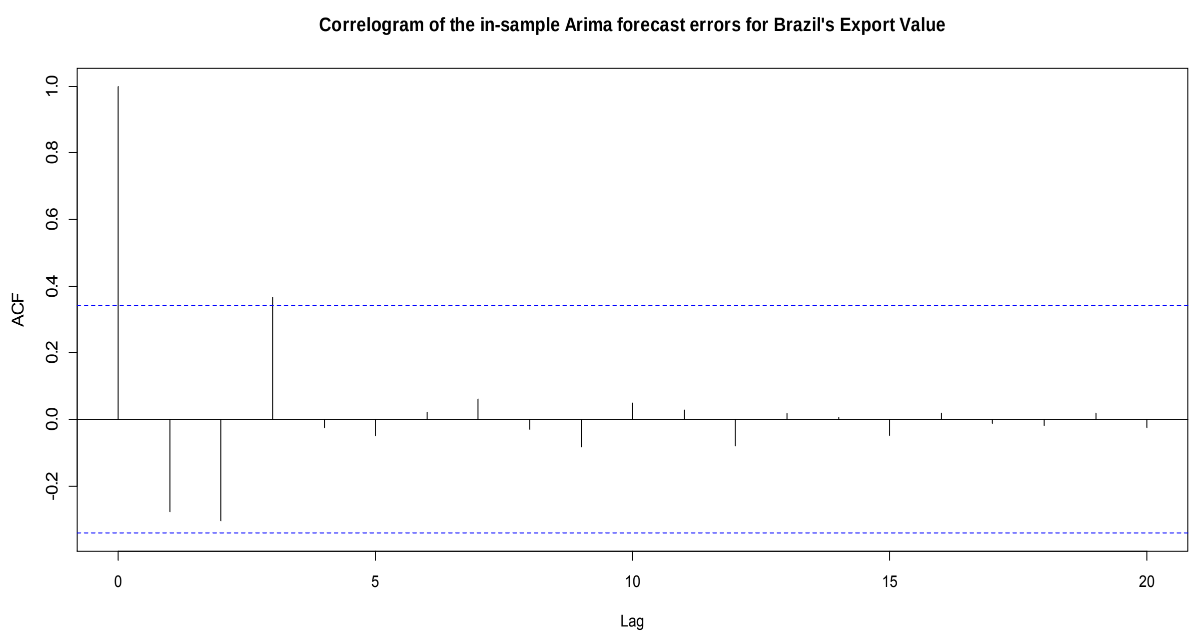
[Figure C.3.73] – Autocorrelogram (ACF) of the twice differenced Brazil time series



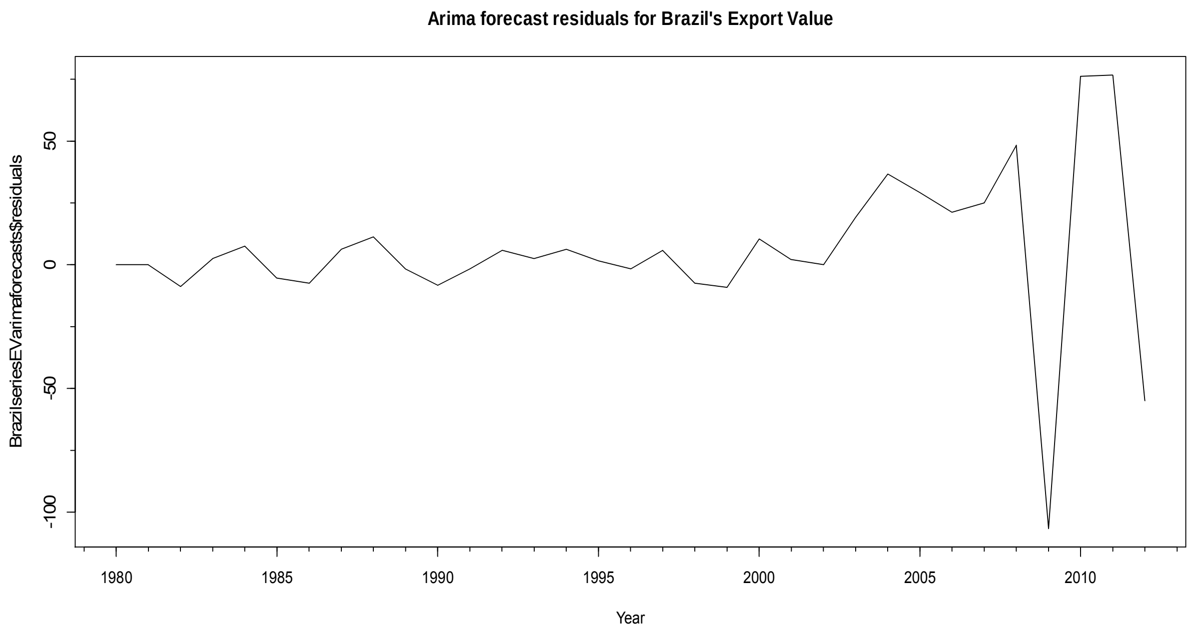
[Figure C.3.74] – Partial autocorrelogram (PACF) of the twice differenced Brazil time series



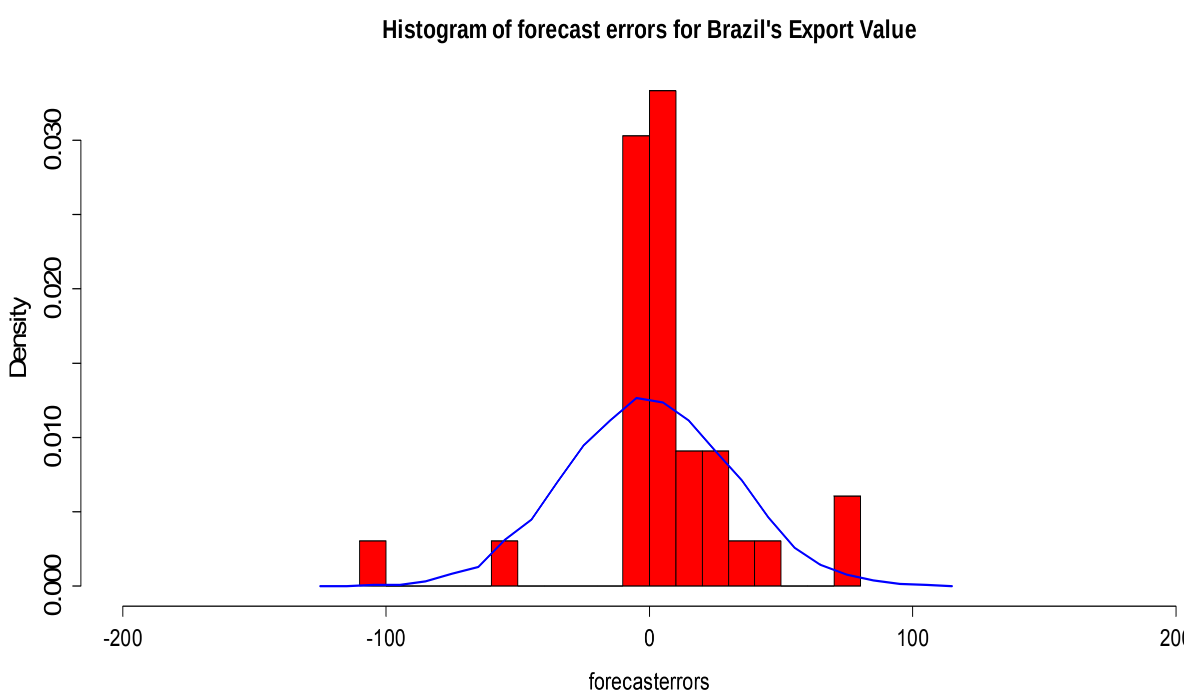
[Figure C.3.75] – Analysis for Brazil, Export Value and whole dataset



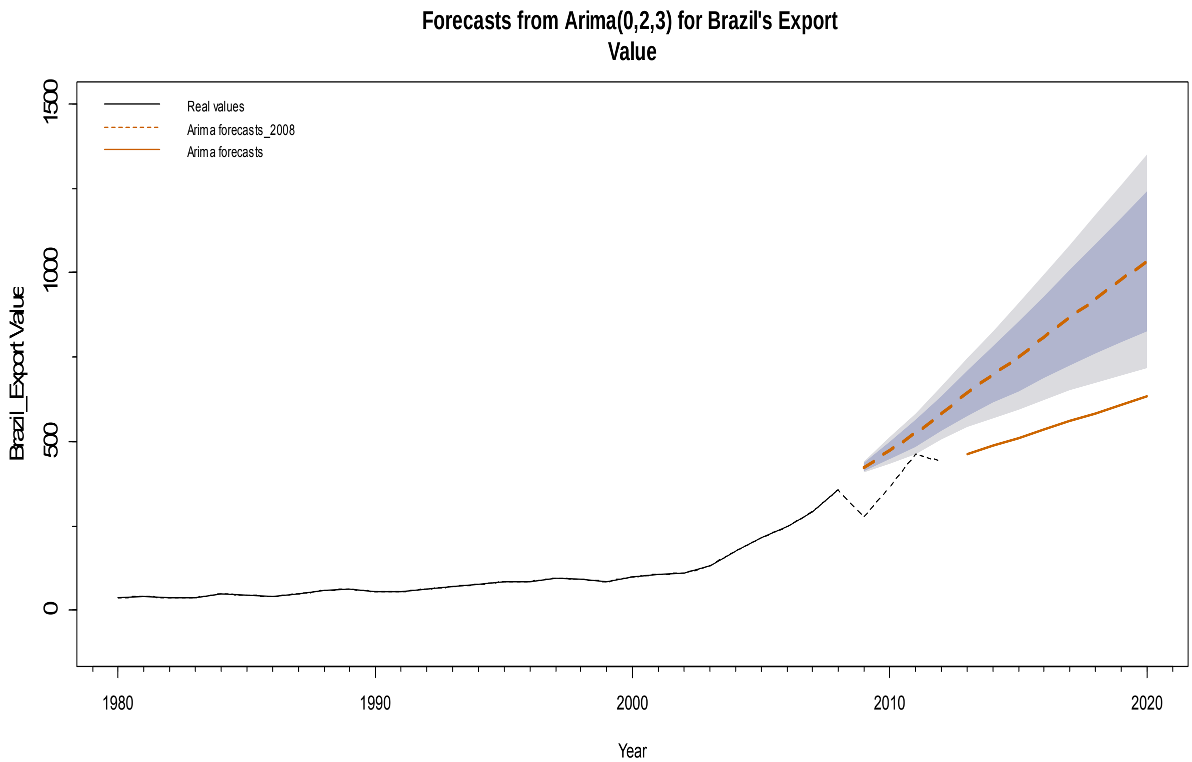
[Figure C.3.76] – Correlogram of in-sample errors of ARIMA forecasts



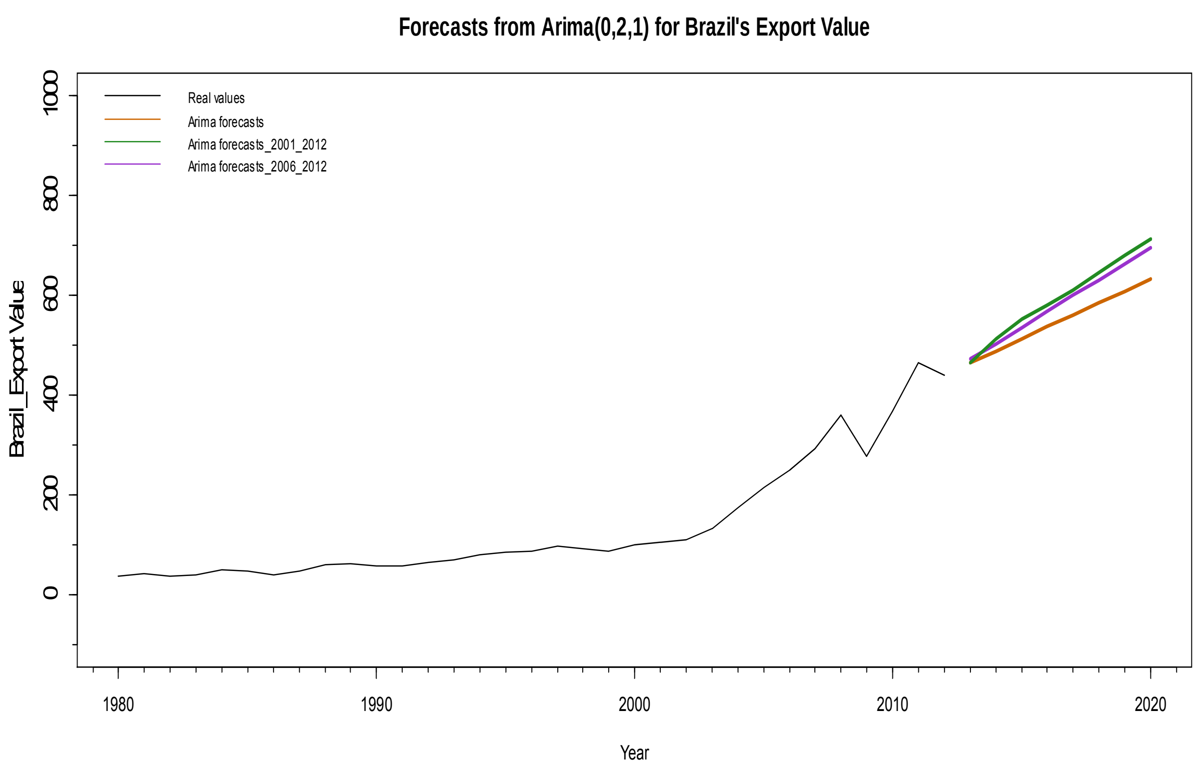
[Figure C.3.77] –Residuals of ARIMA forecasts



[Figure C.3.78] – Histogram and distribution of forecast residuals

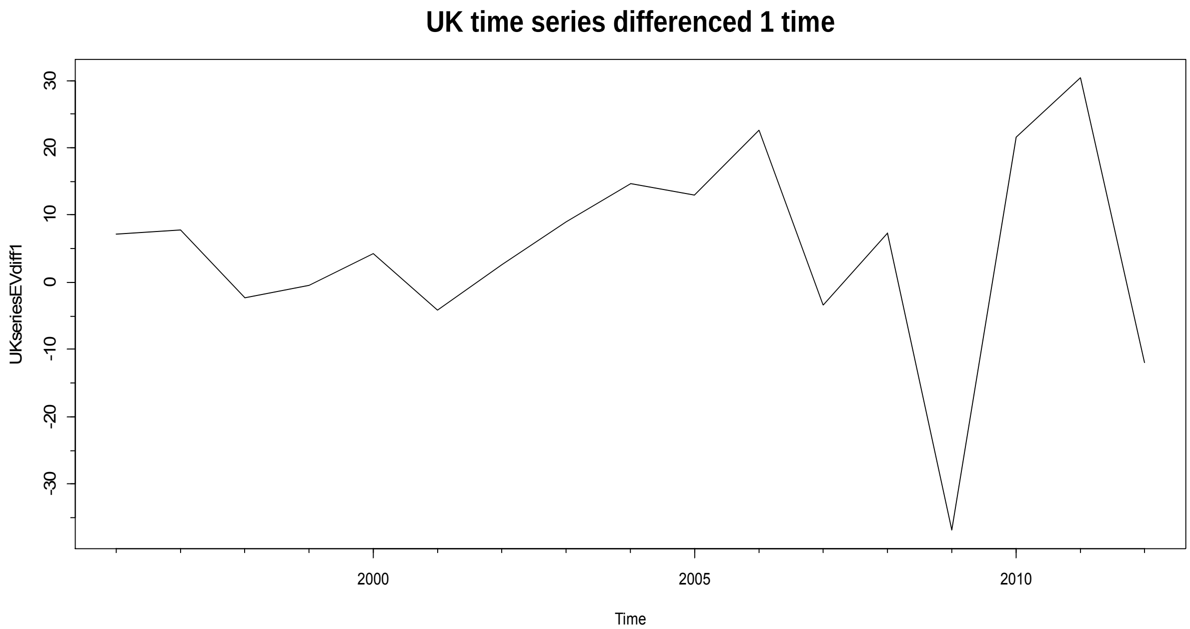


[Figure C.3.79] – Analysis for Brazil, Export Value and the dataset up to 2008

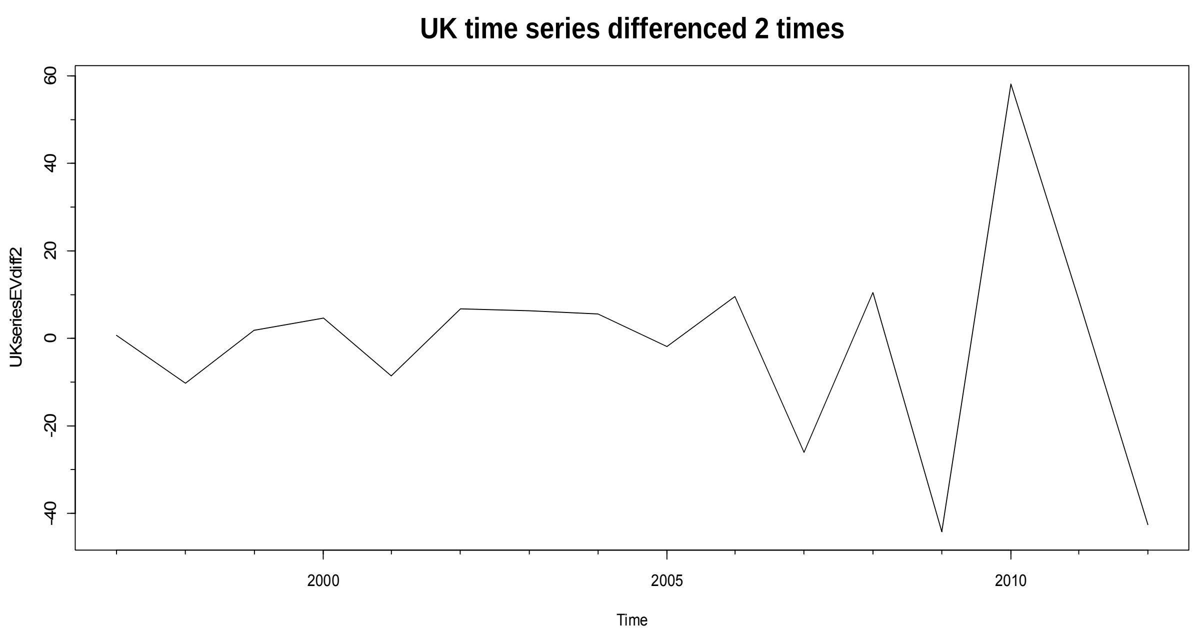


[Figure C.3.80] – Analyses for Brazil, Export Value and the subsets 2001-2013 and 2006-2013

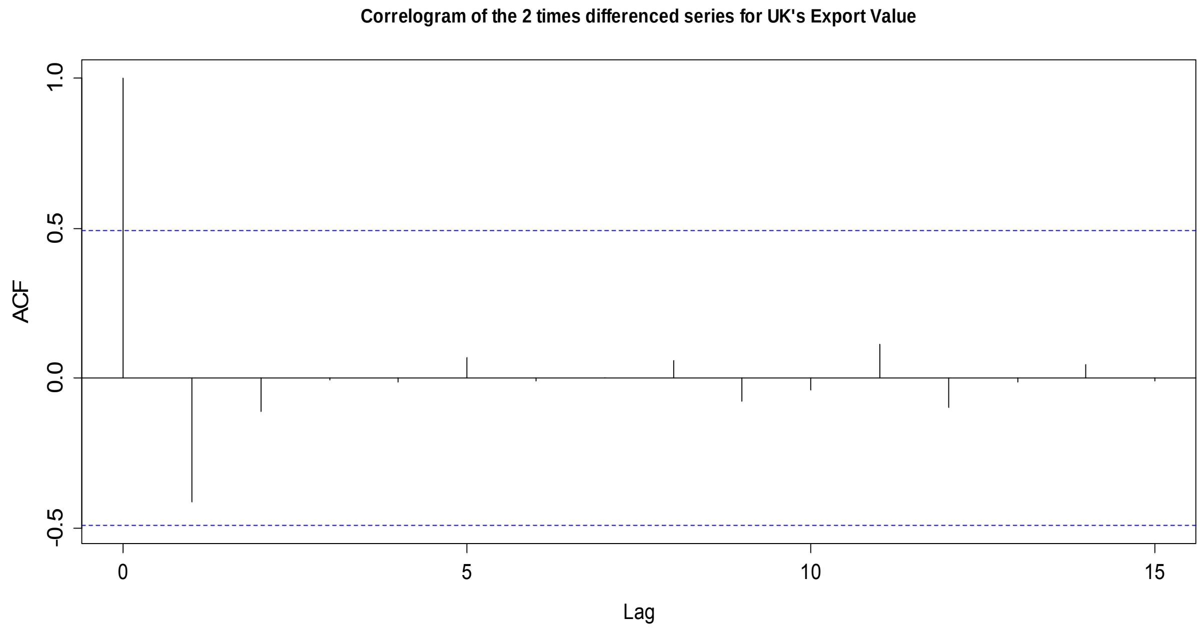
Export Value – UK



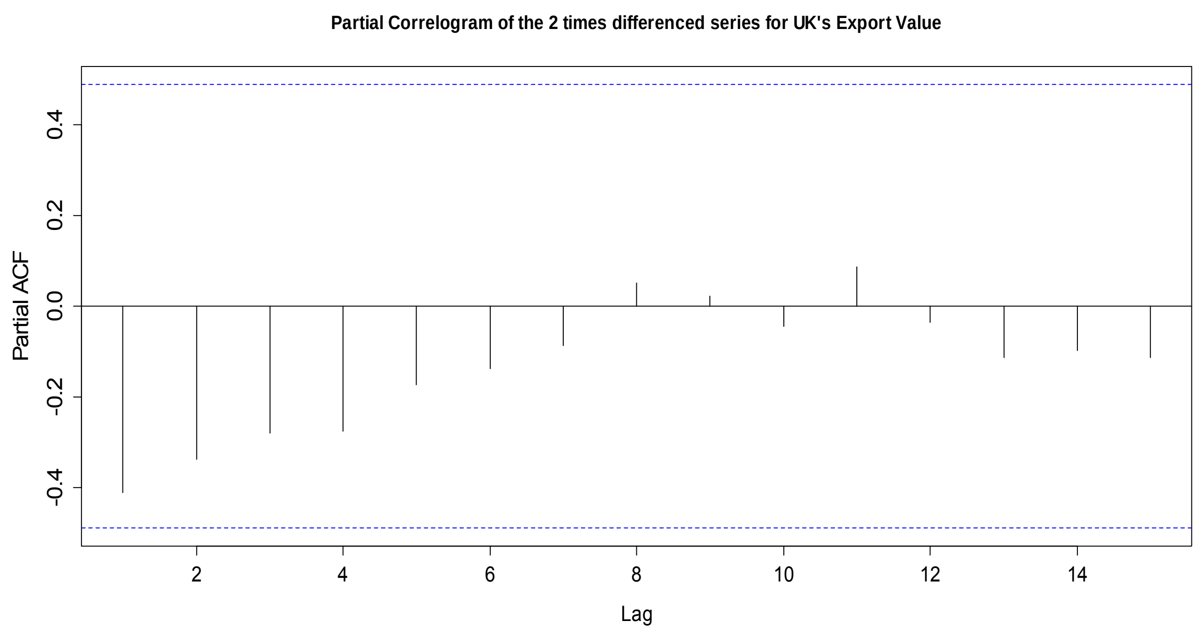
[Figure C.3.81] – One time differenced UK time series



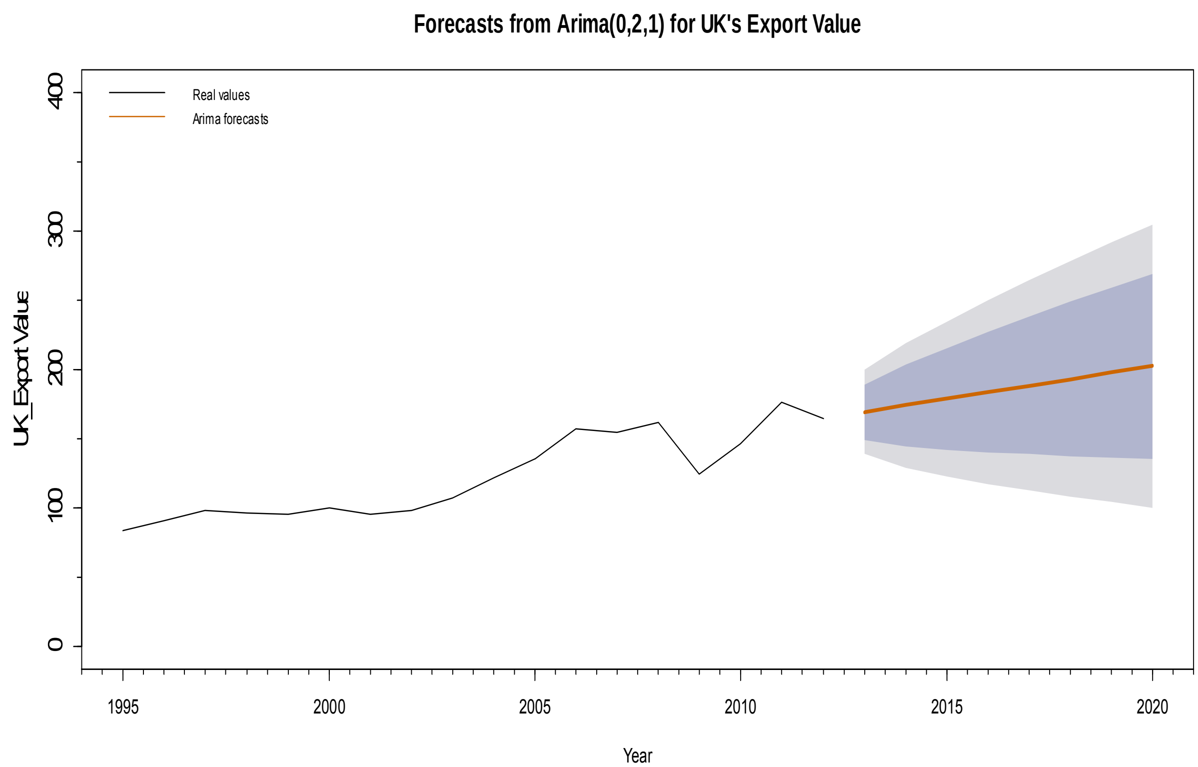
[Figure C.3.82] – Two times differenced UK time series



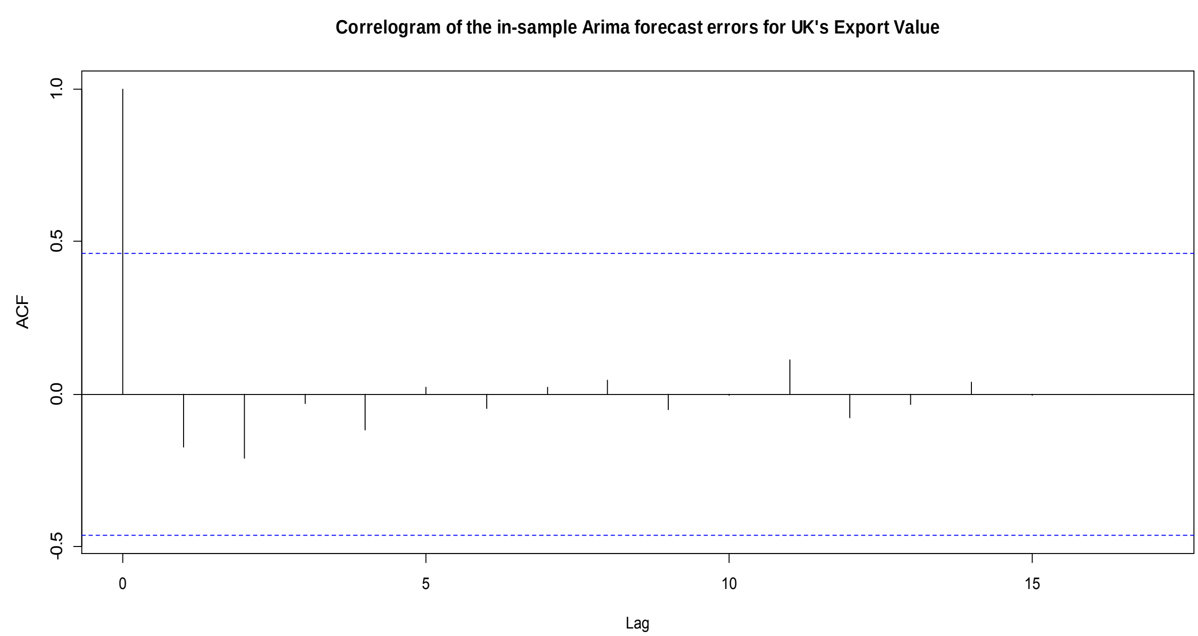
[Figure C.3.83] – Autocorrelogram (ACF) of the twice differenced UK time series



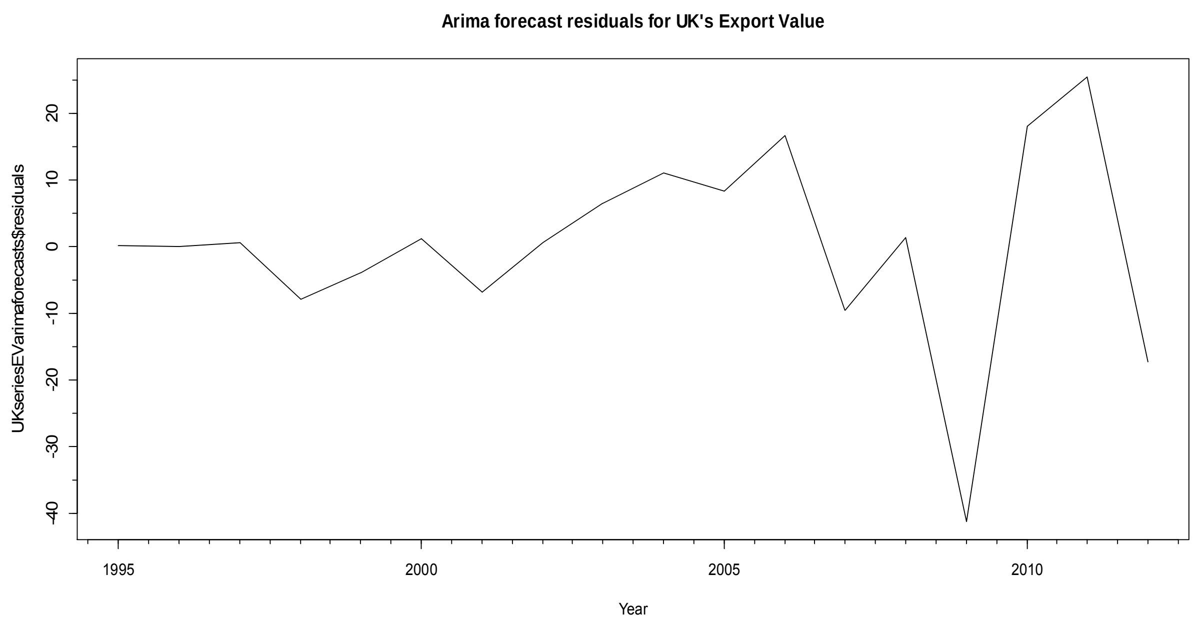
[Figure C.3.84] – Partial autocorrelogram (PACF) of the twice differenced UK time series



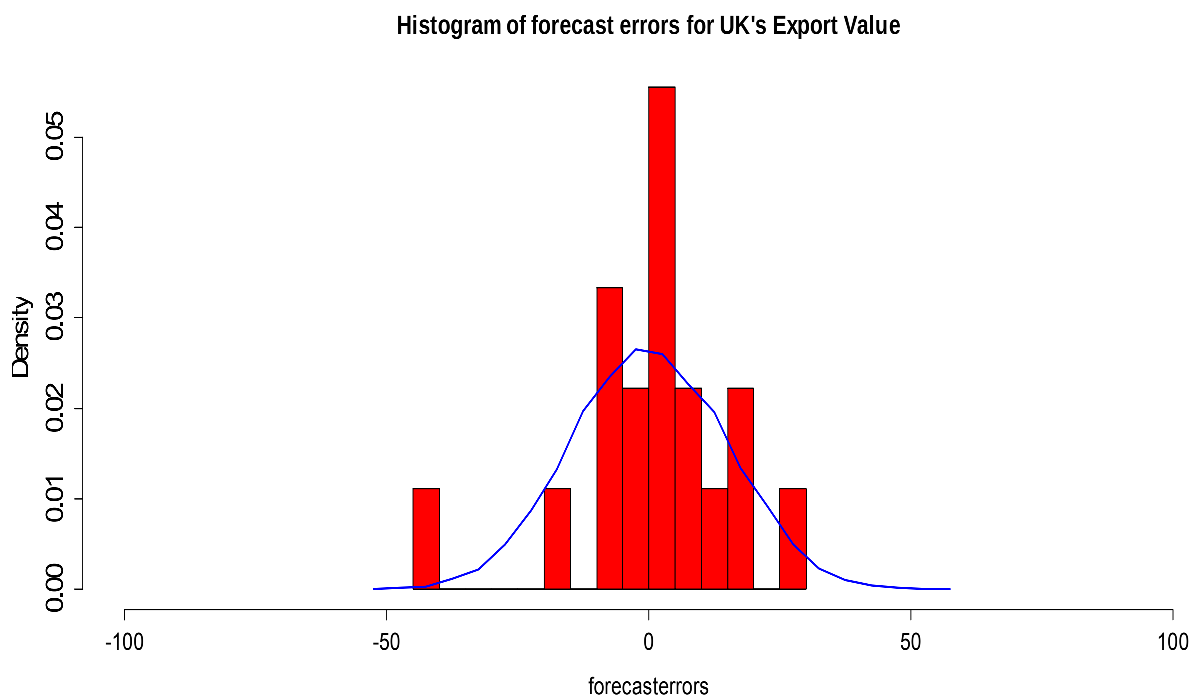
[Figure C.3.85] – Analysis for UK, Export Value and whole dataset



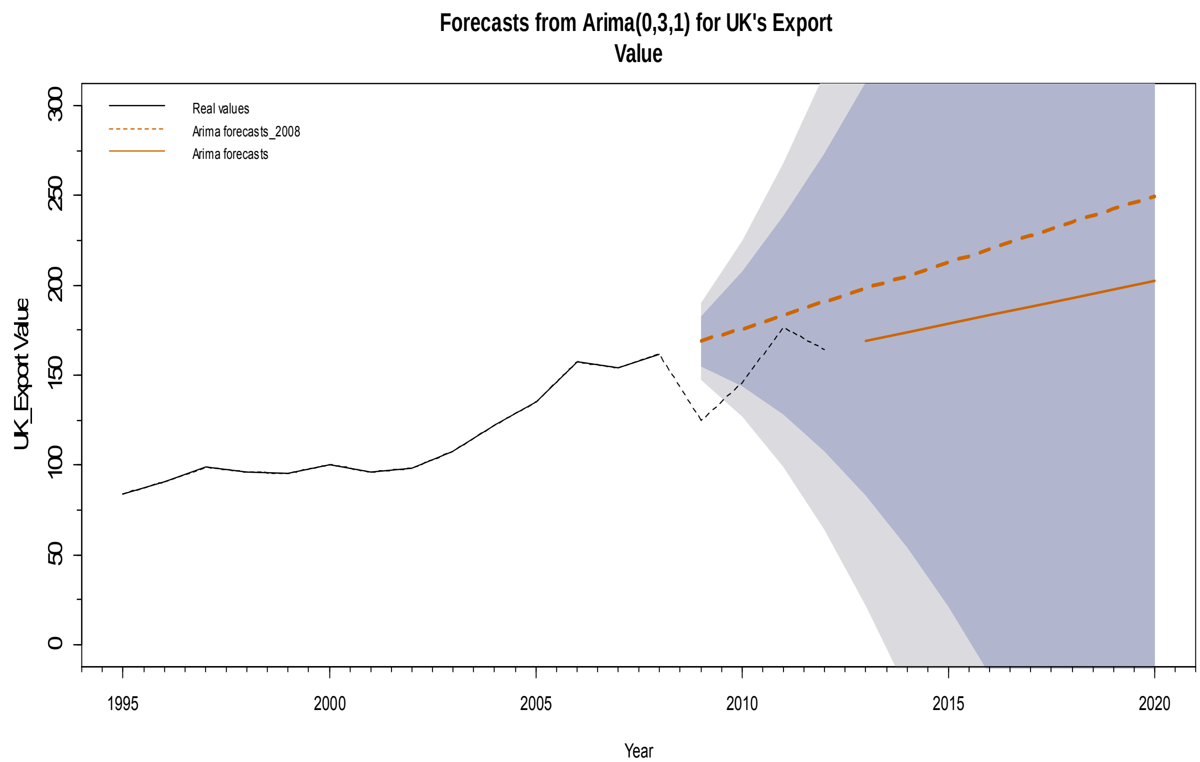
[Figure C.3.86] – Correlogram of in-sample errors of ARIMA forecasts



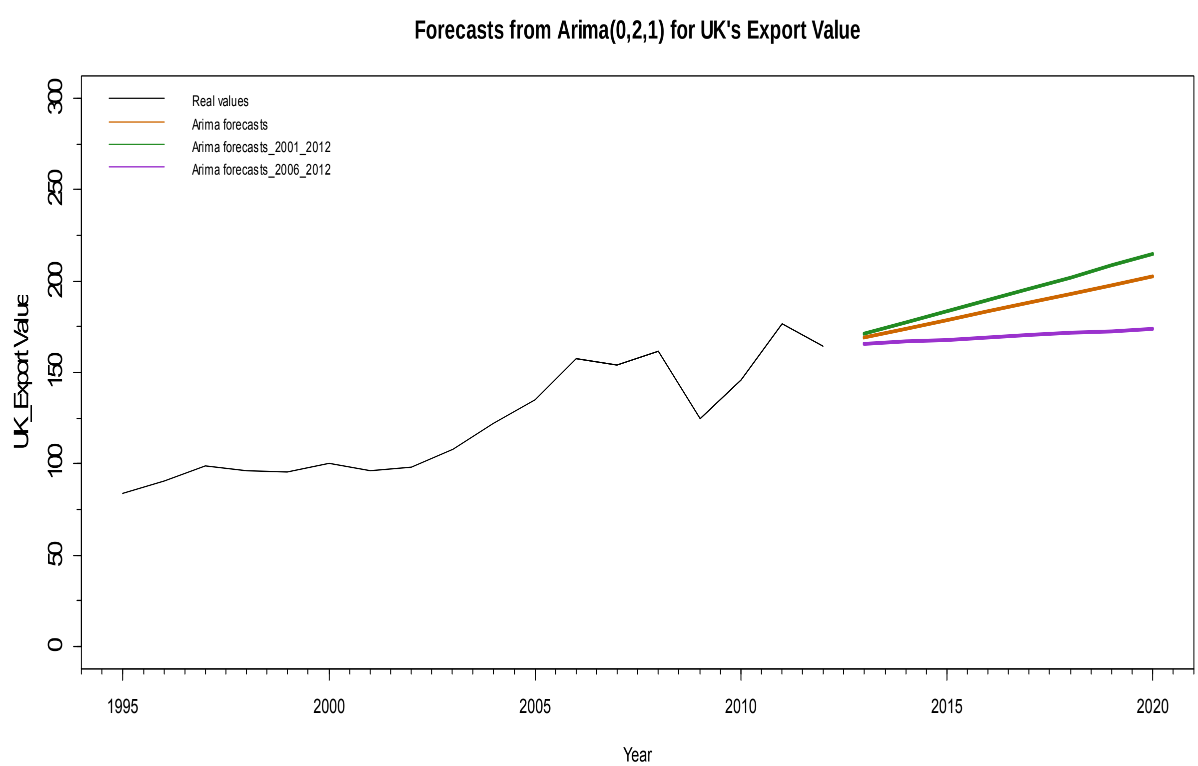
[Figure C.3.87] –Residuals of ARIMA forecasts



[Figure C.3.88] – Histogram and distribution of forecast residuals

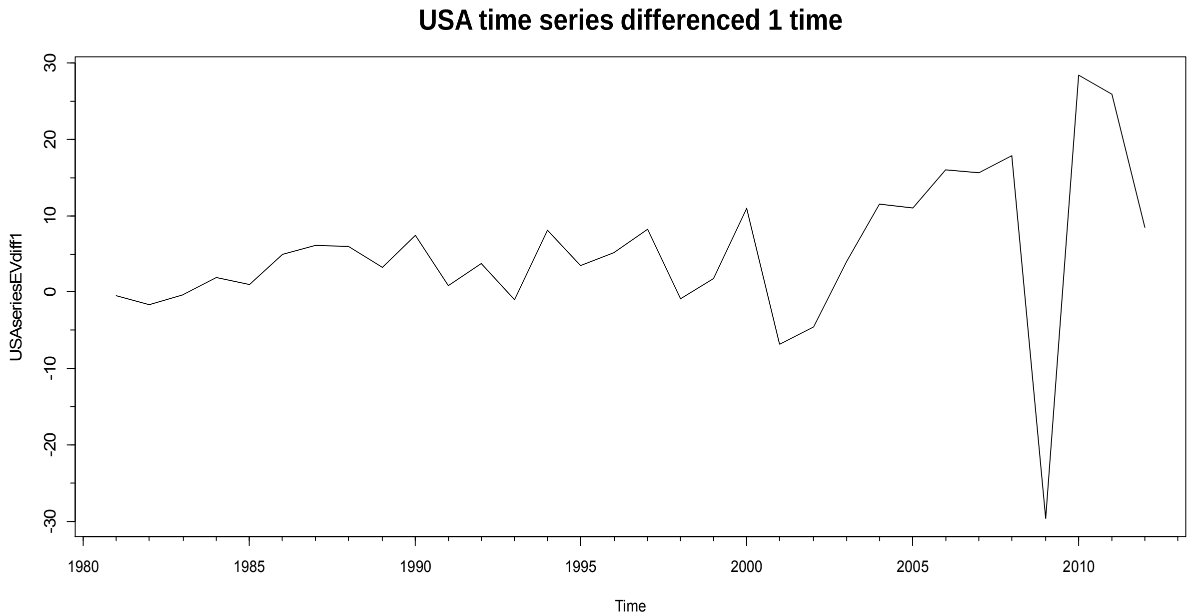


[Figure C.3.89] – Analysis for UK, Export Value and the dataset up to 2008

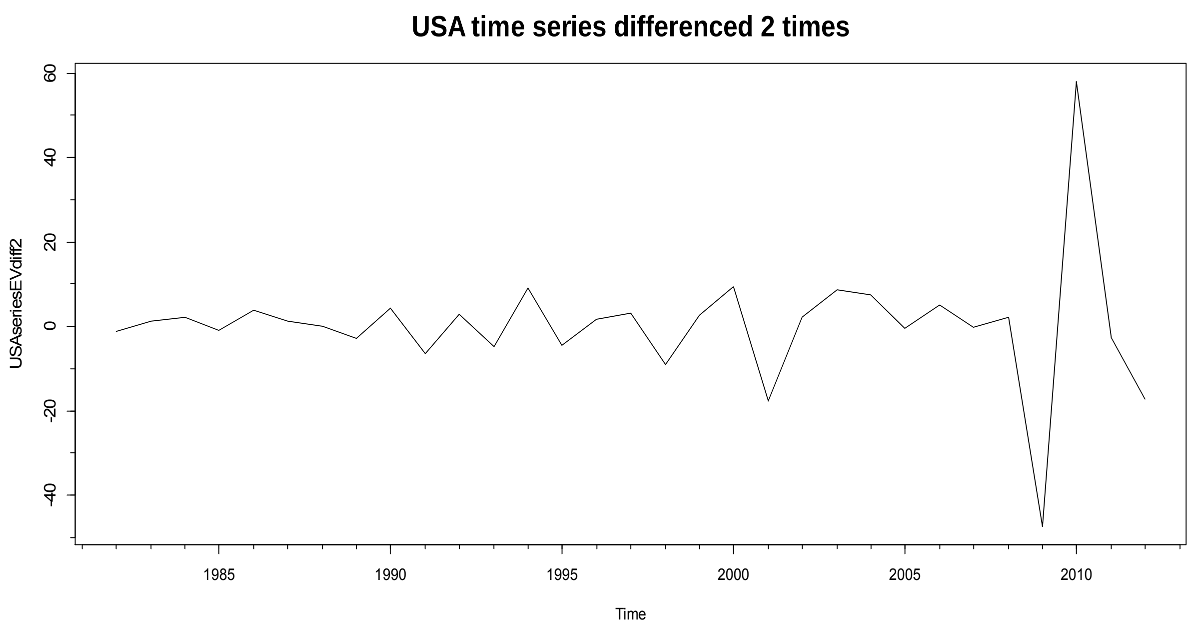


[Figure C.3.90] – Analyses for UK, Export Value and the subsets 2001-2013 and 2006-2013

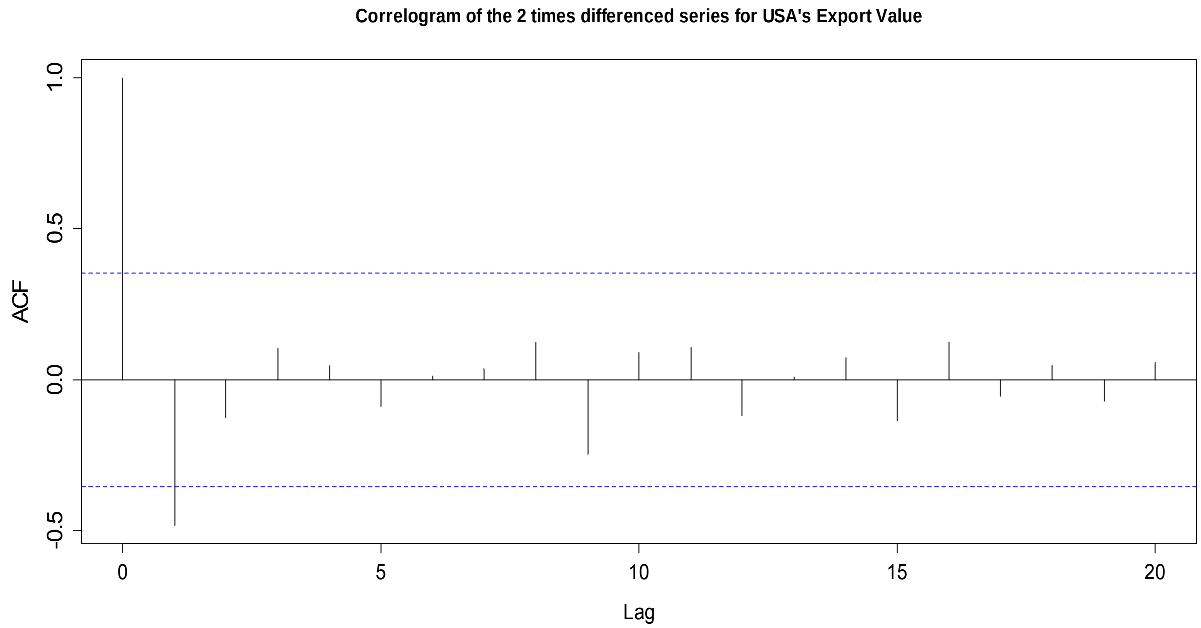
Export Value – USA



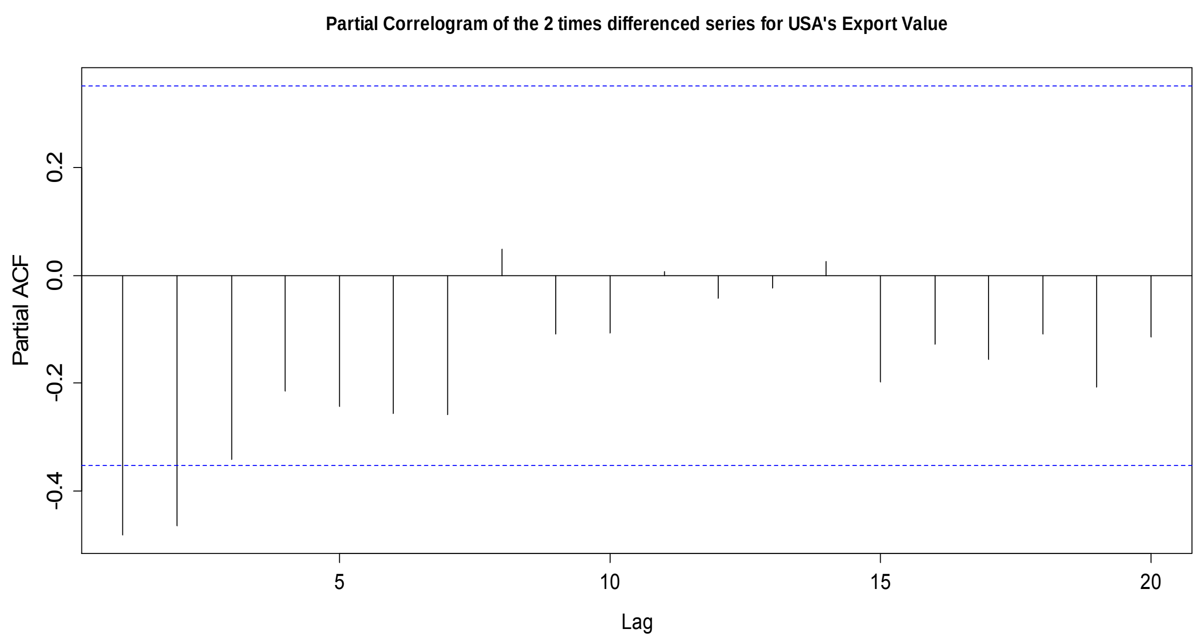
[Figure C.3.91] – One time differenced USA time series



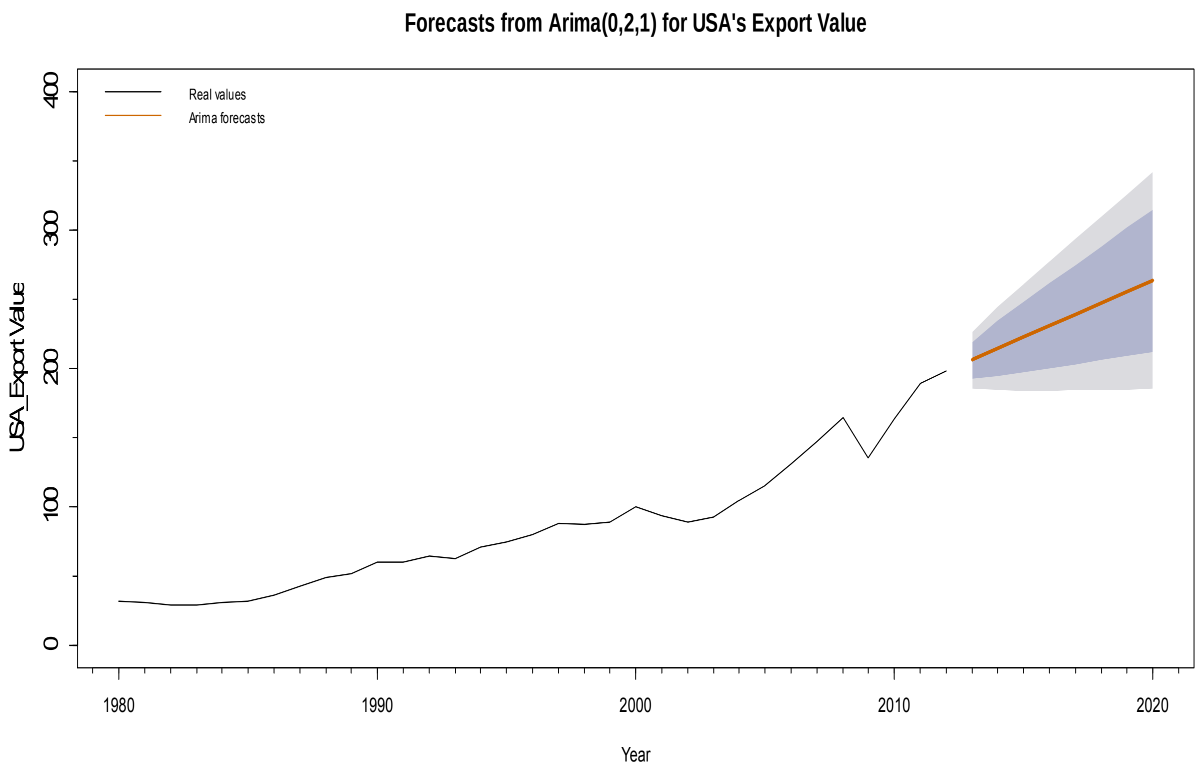
[Figure C.3.92] – Two times differenced USA time series



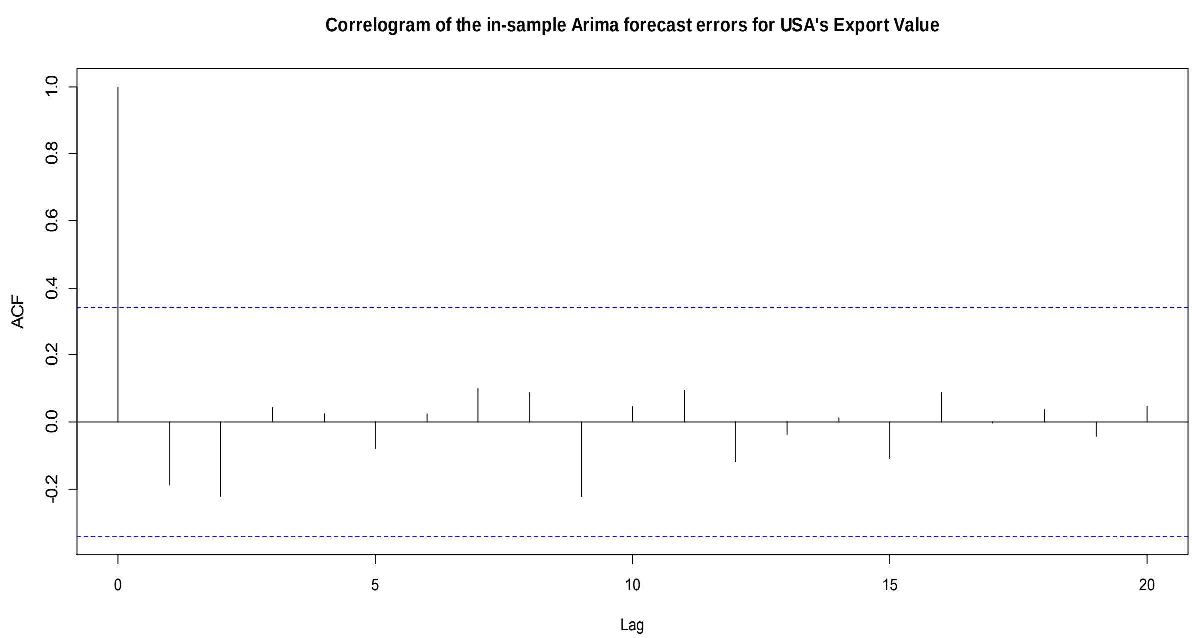
[Figure C.3.93] – Autocorrelogram (ACF) of the twice differenced USA time series



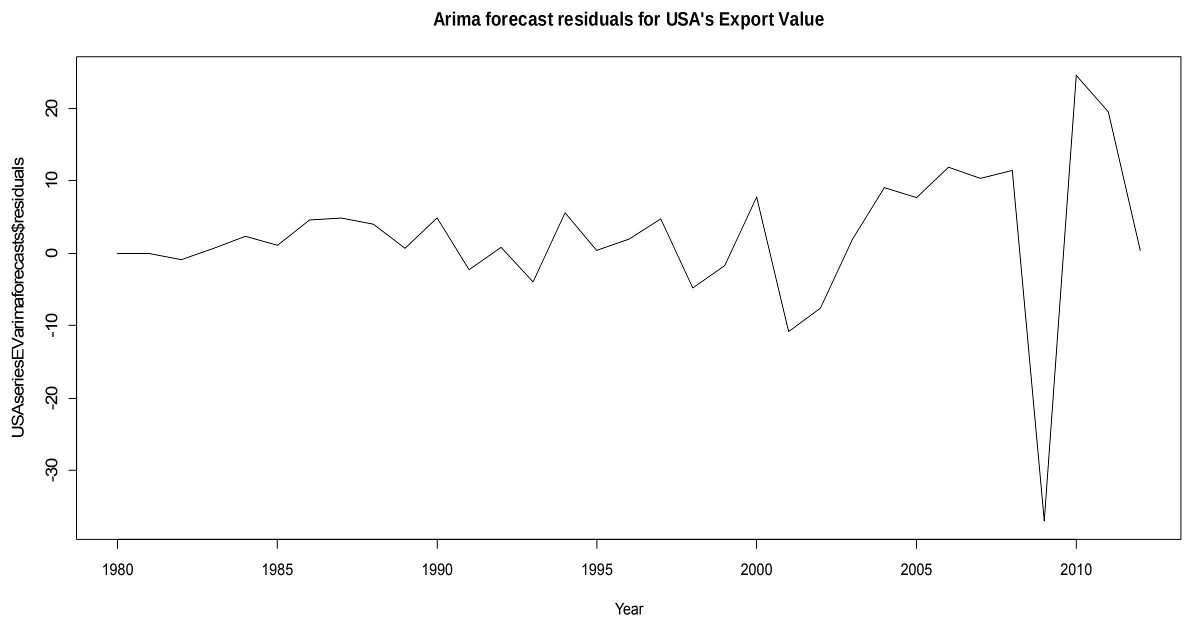
[Figure C.3.94] – Partial autocorrelogram (PACF) of the twice differenced USA time series



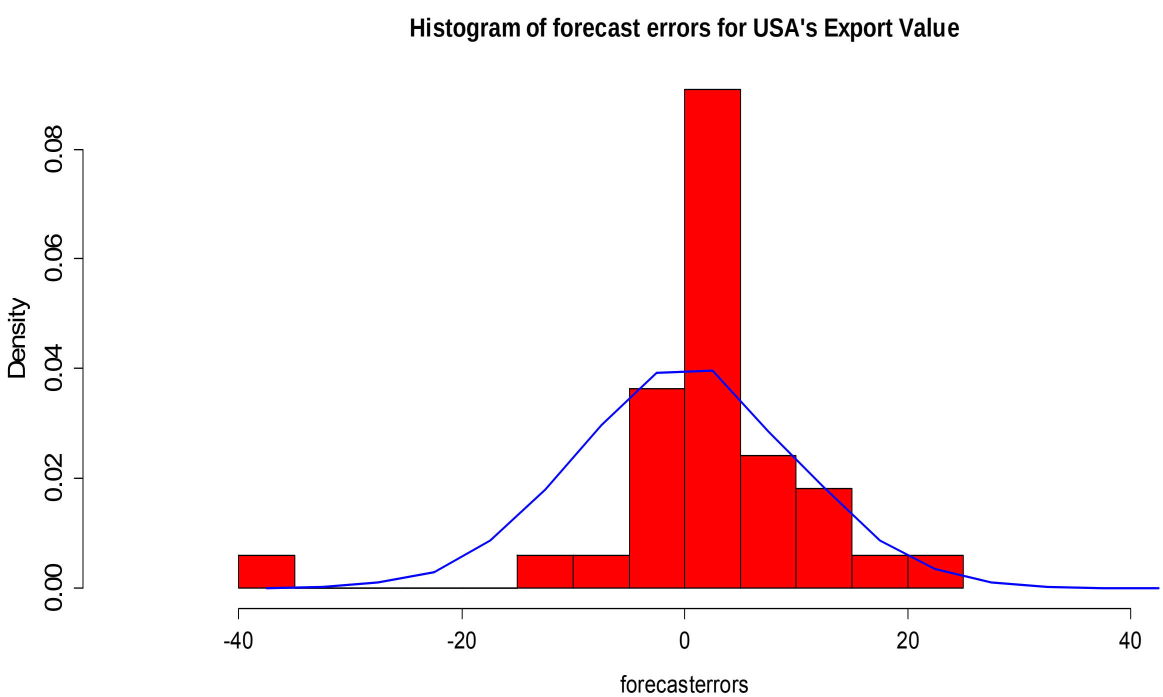
[Figure C.3.95] – Analysis for USA, Export Value and whole dataset



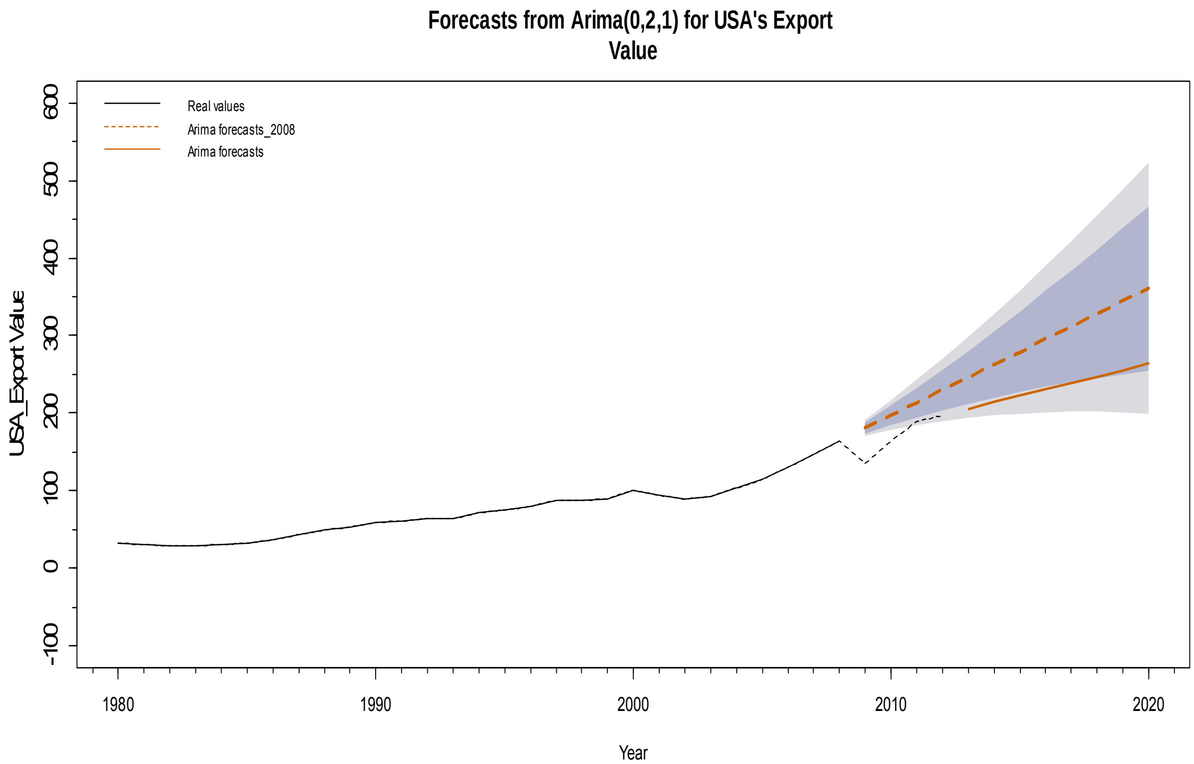
[Figure C.3.96] – Correlogram of in-sample errors of ARIMA forecasts



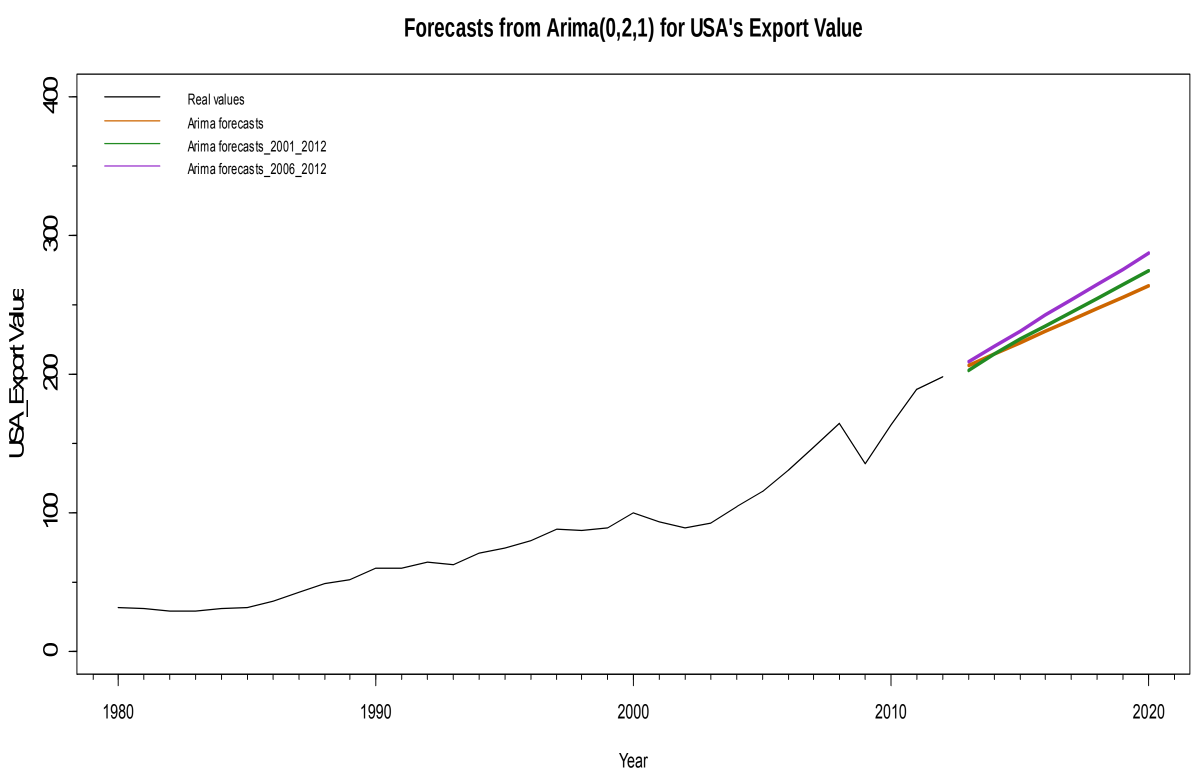
[Figure C.3.97] –Residuals of ARIMA forecasts



[Figure C.3.98] – Histogram and distribution of forecast residuals

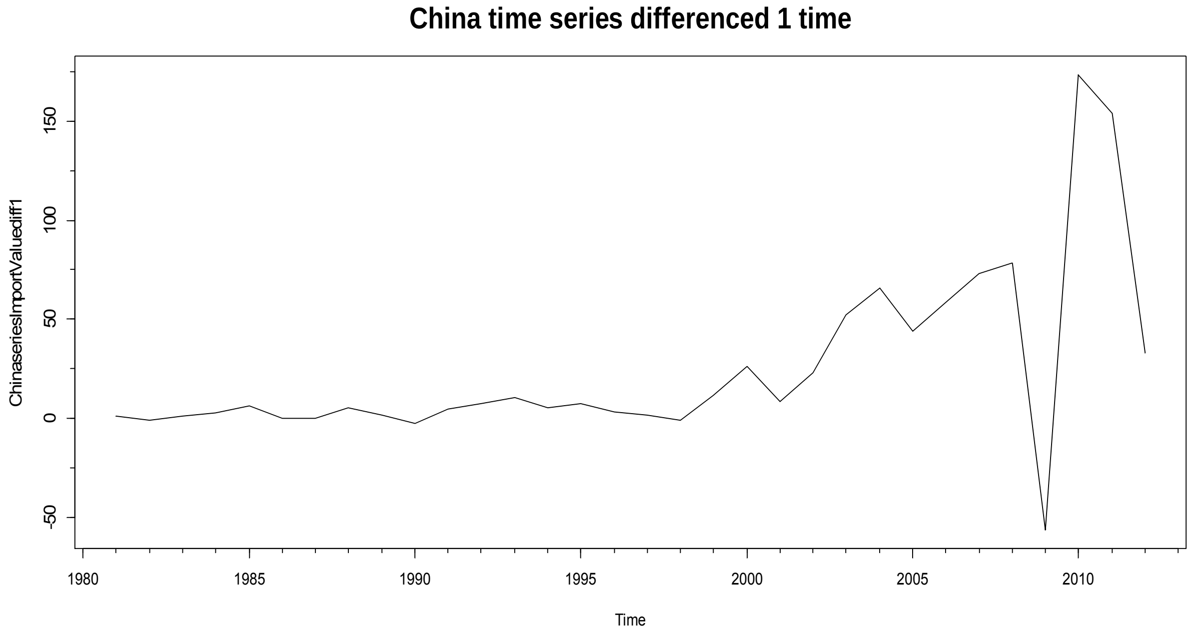


[Figure C.3.99] – Analysis for USA, Export Value and the dataset up to 2008

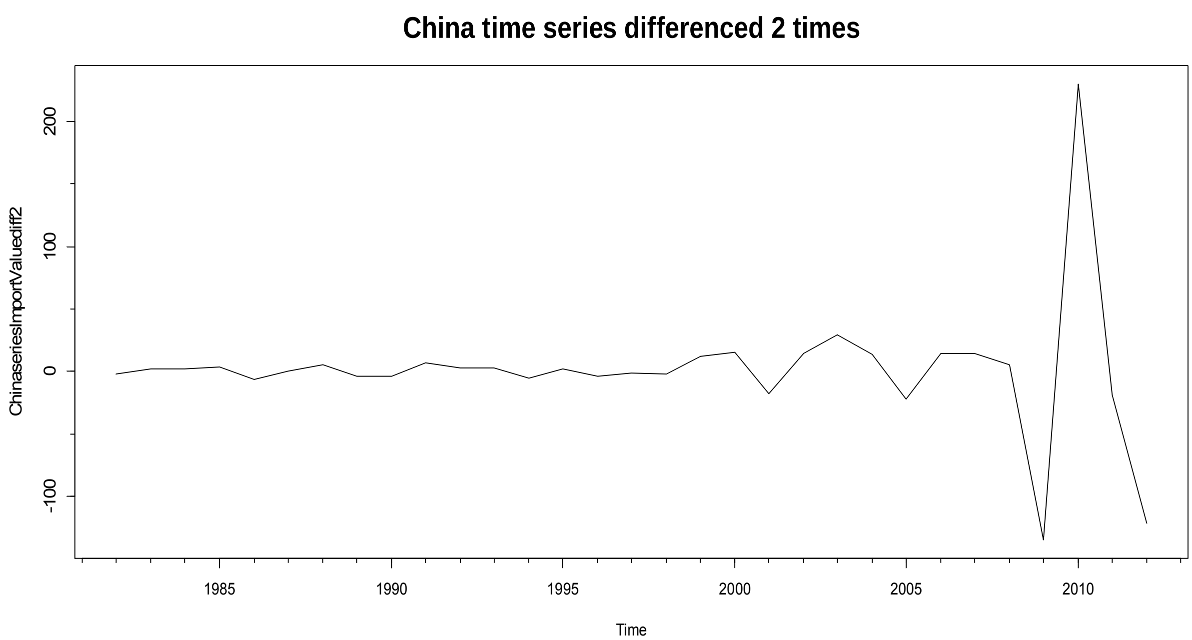


[Figure C.3.100] – Analyses for USA, Export Value and the subsets 2001-2013 and 2006-2013

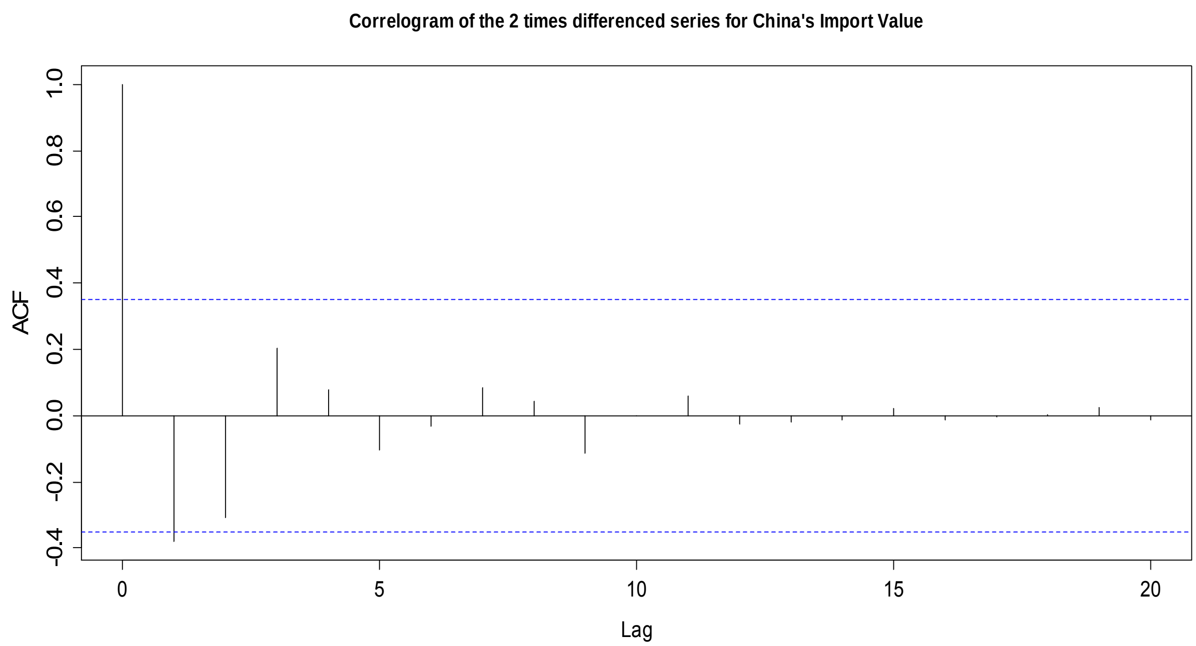
Import Value – China



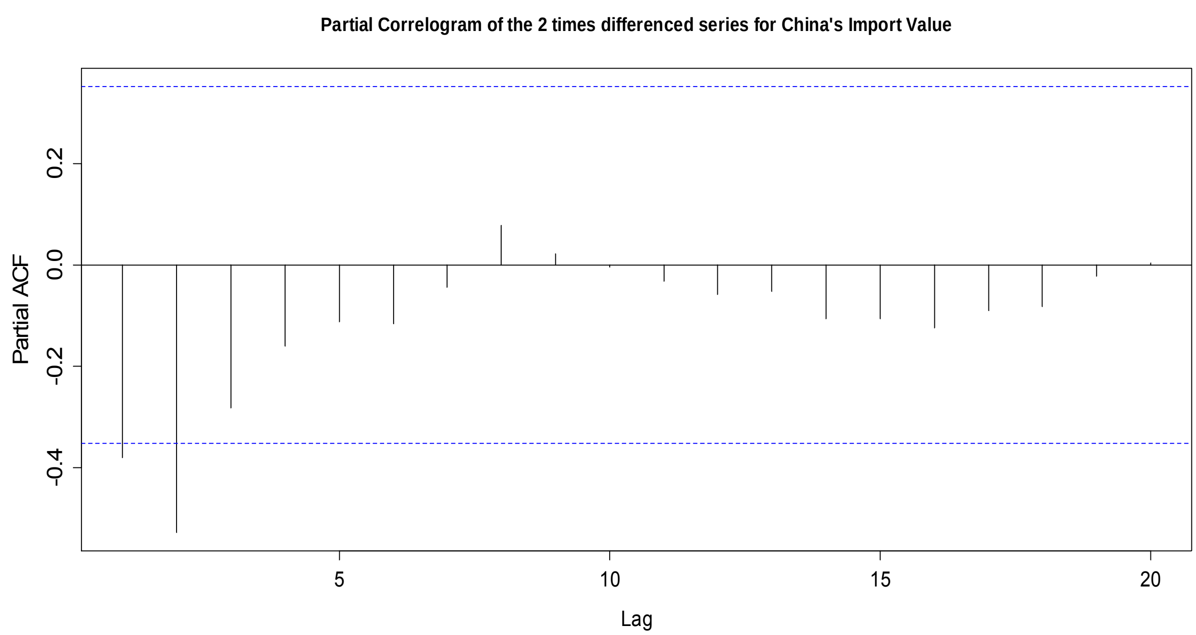
[Figure C.3.101] – One time differenced China time series



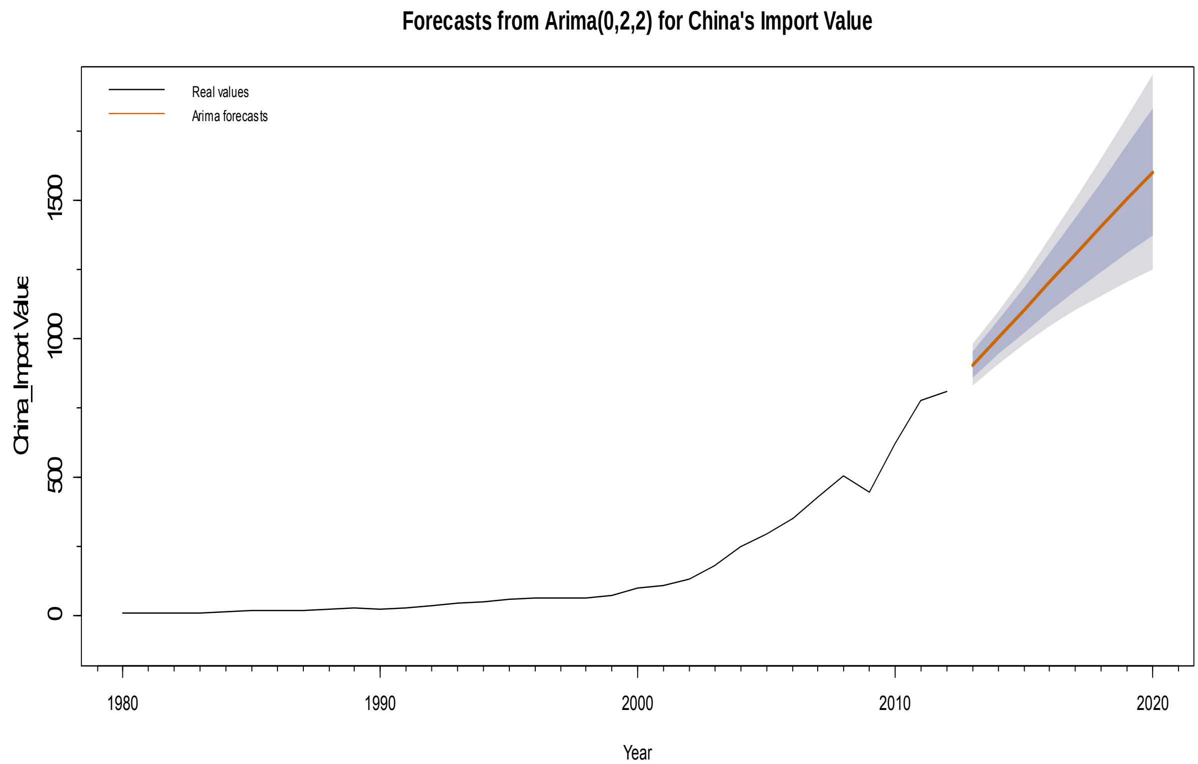
[Figure C.3.102] – Two times differenced China time series



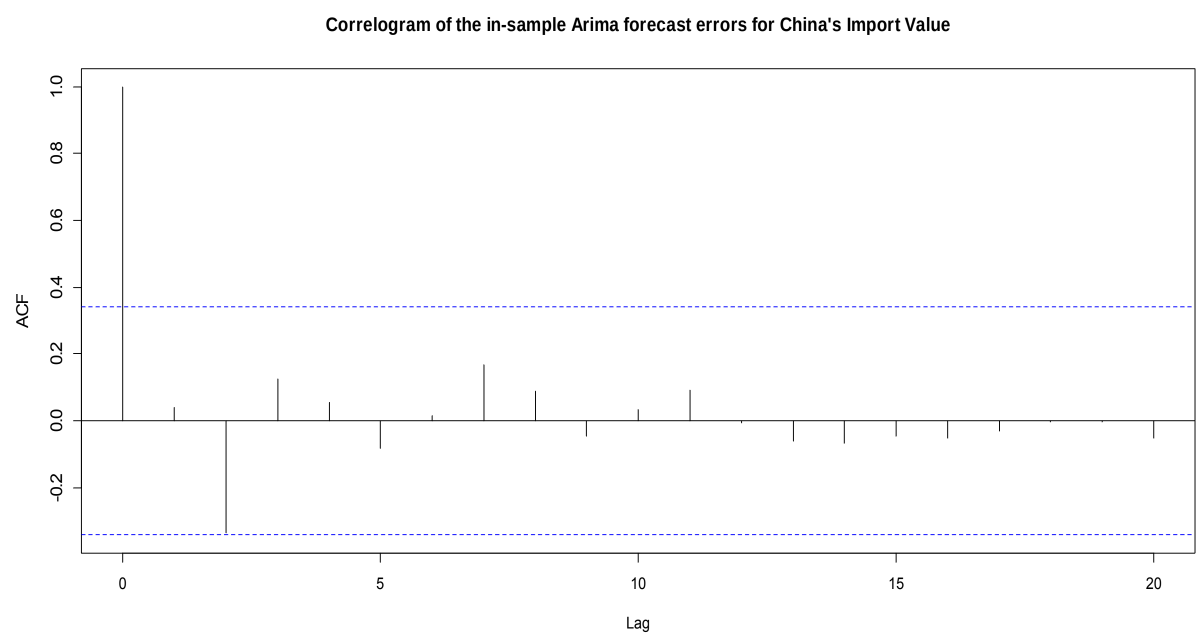
[Figure C.3.103] – Autocorrelogram (ACF) of the twice differenced China time series



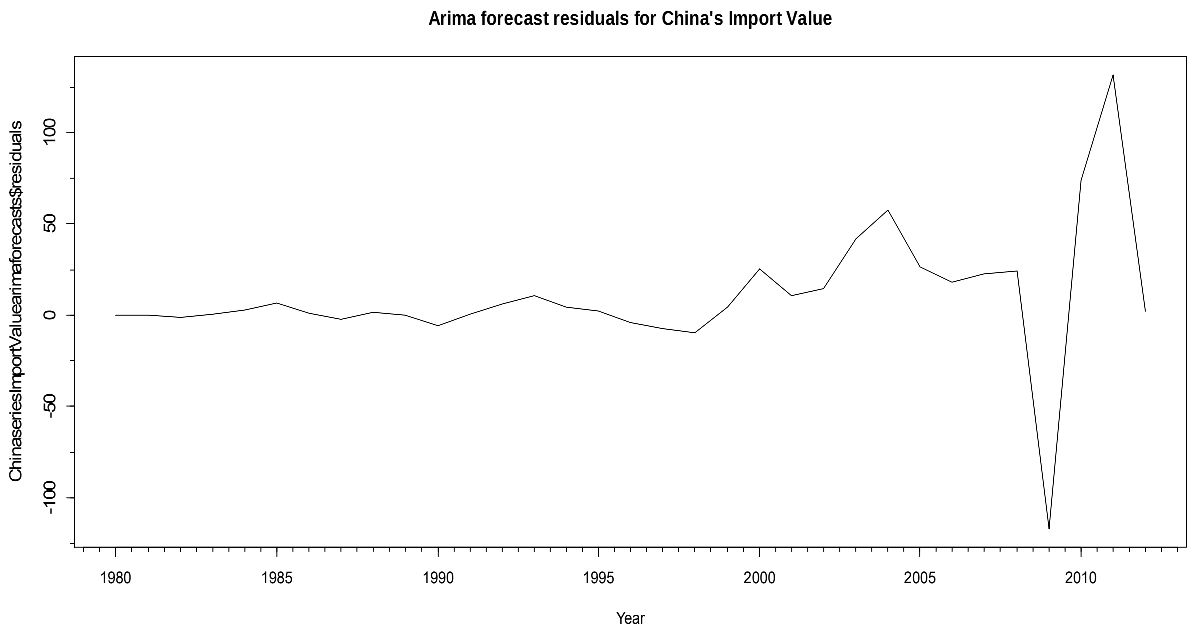
[Figure C.3.104] – Partial autocorrelogram (PACF) of the twice differenced China time series



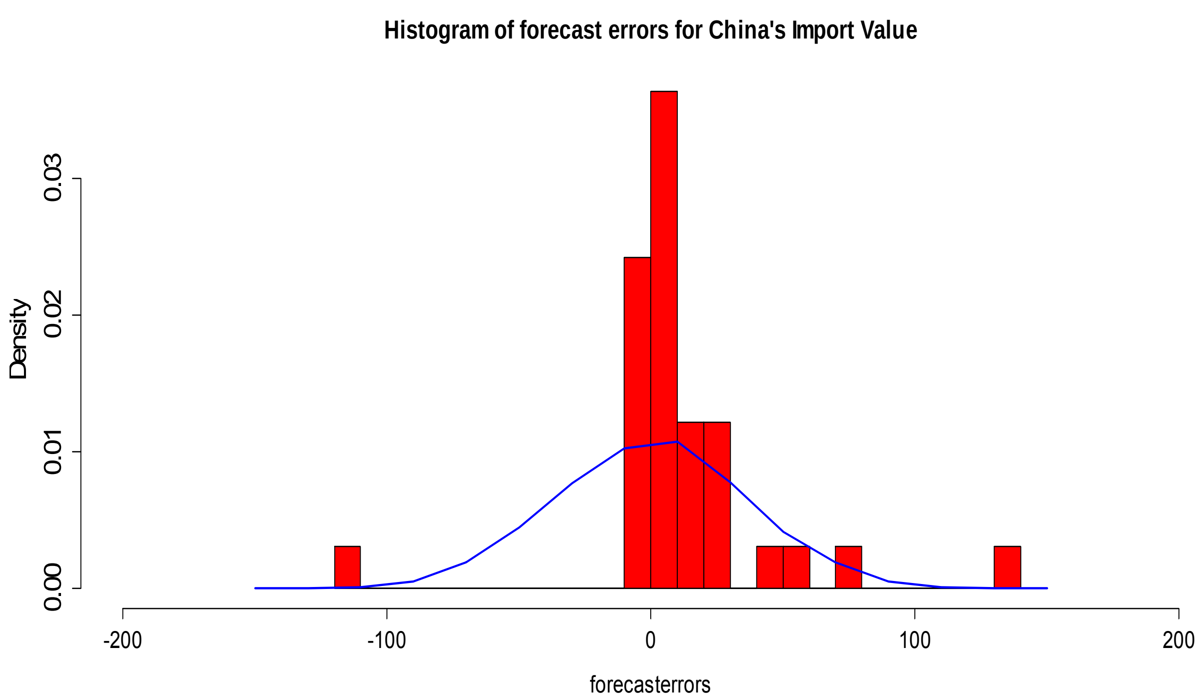
[Figure C.3.105] – Analysis for China, Import Value and whole dataset



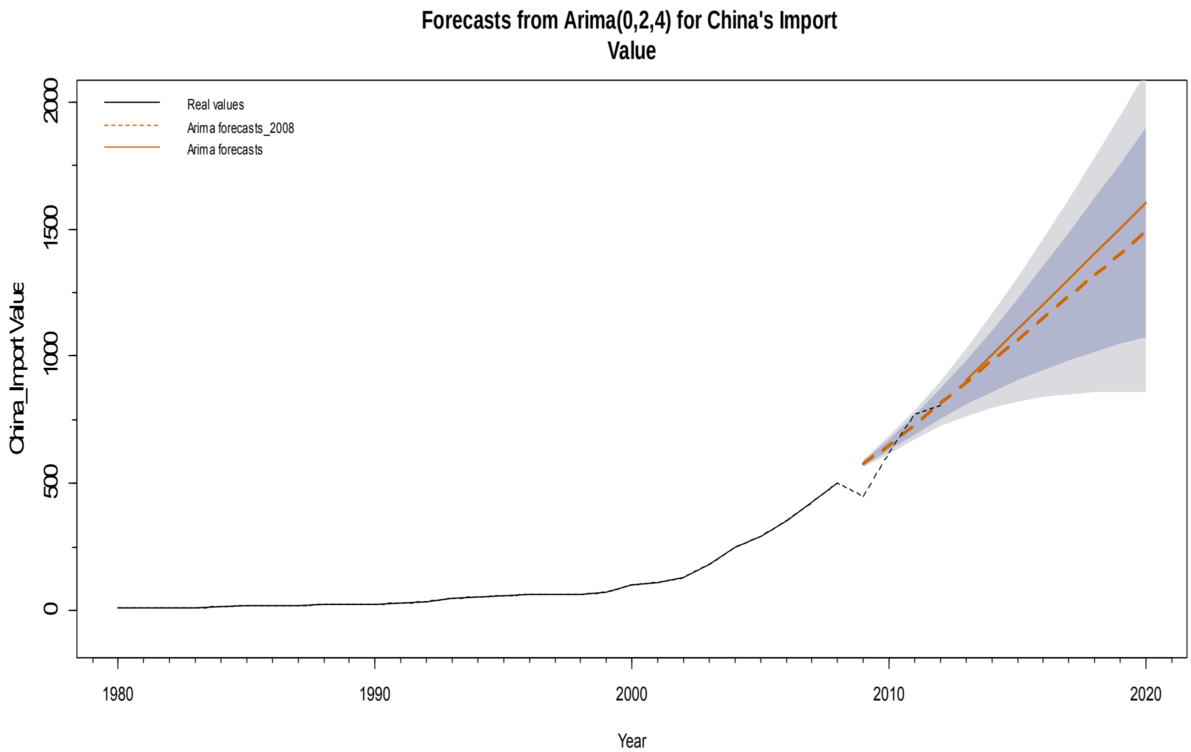
[Figure C.3.106] – Correlogram of in-sample errors of ARIMA forecasts



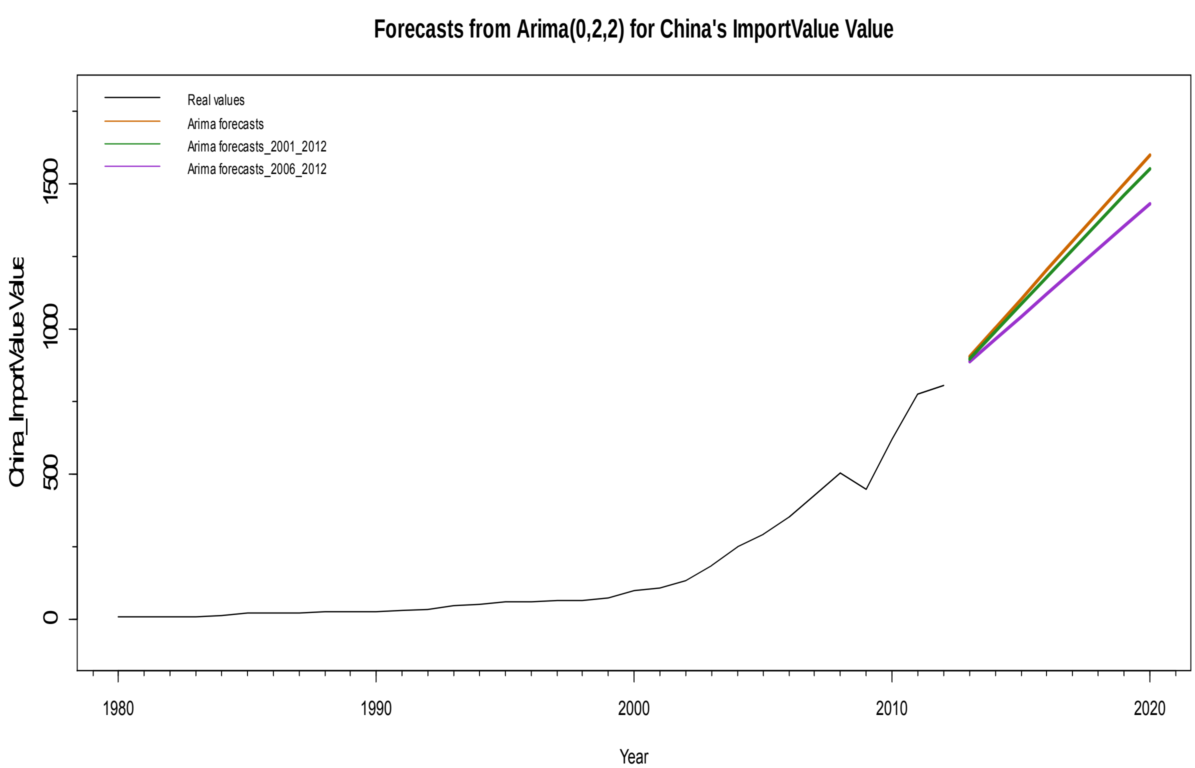
[Figure C.3.107] –Residuals of ARIMA forecasts



[Figure C.3.108] – Histogram and distribution of forecast residuals

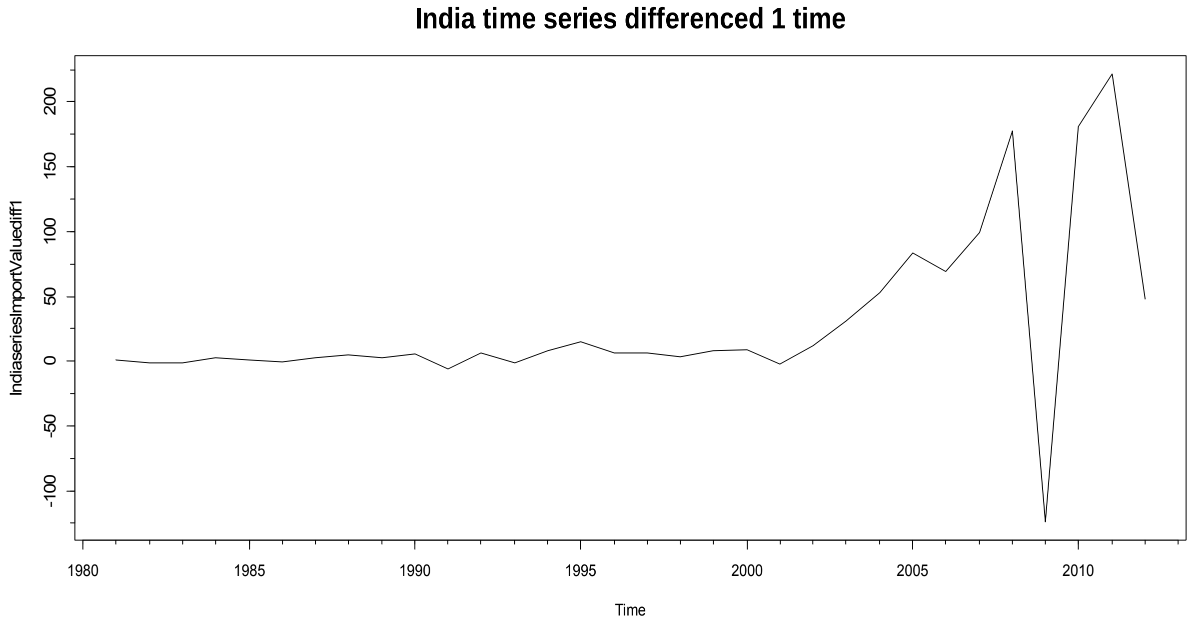


[Figure C.3.109] – Analysis for China, Import Value and the dataset up to 2008

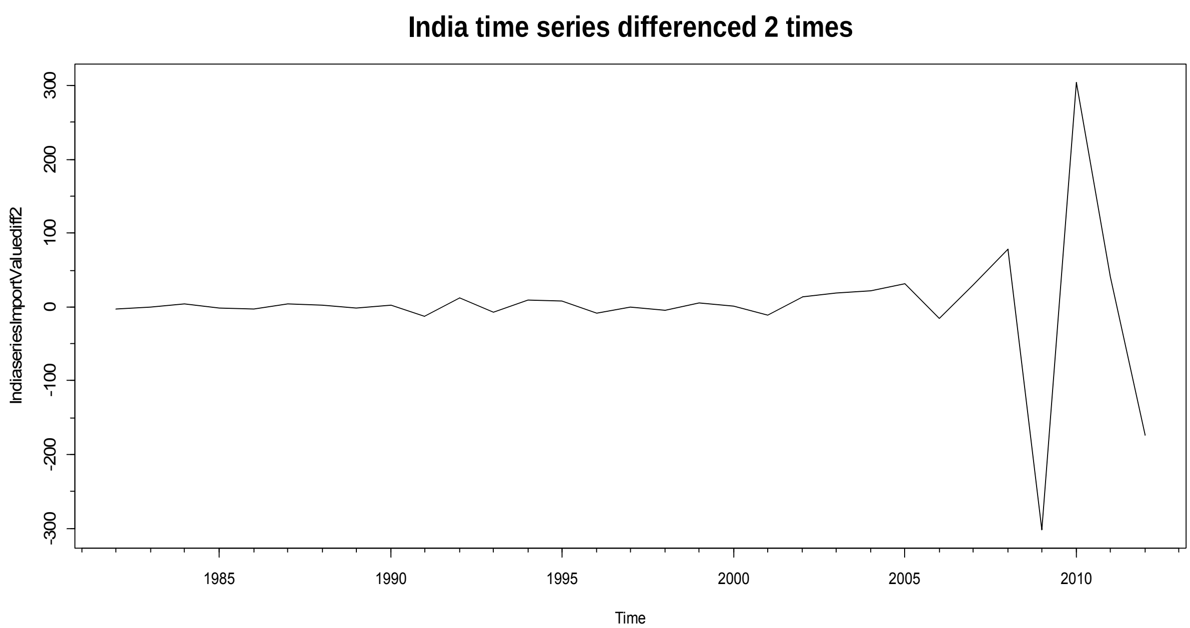


[Figure C.3.110] – Analyses for China, Import Value and the subsets 2001-2013 and 2006-2013

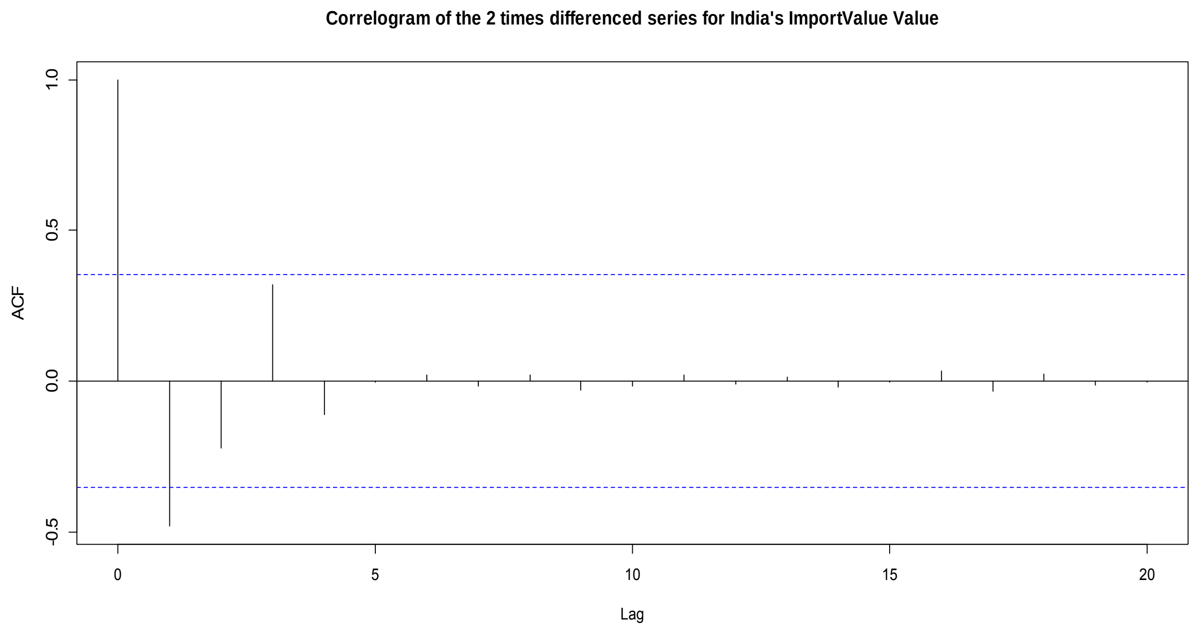
Import Value – India



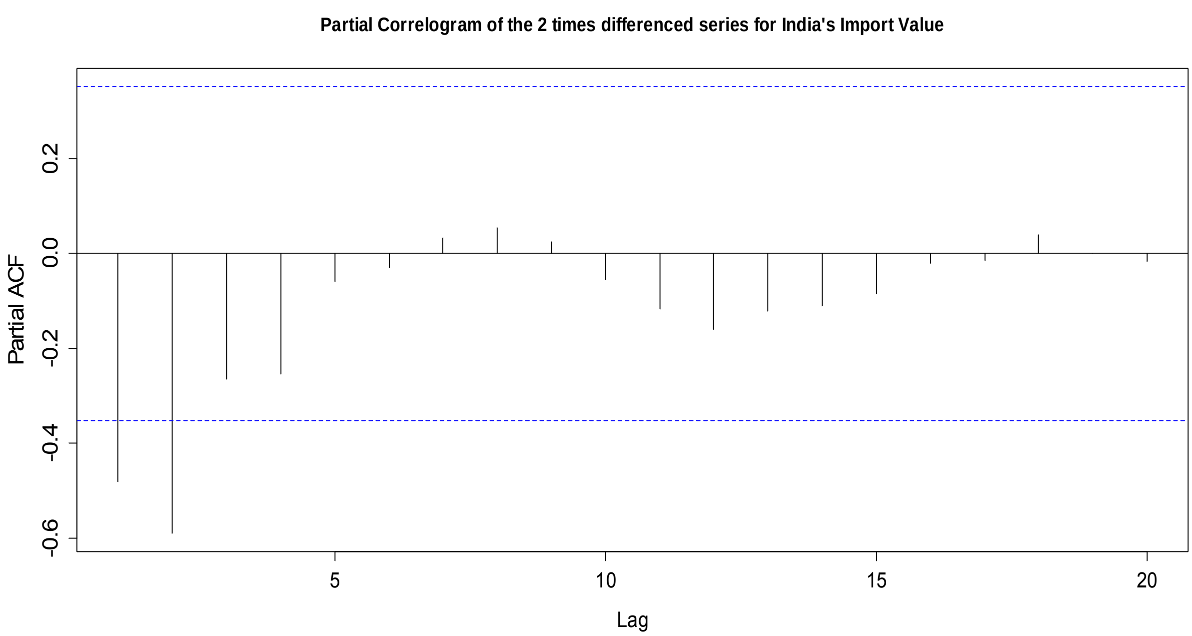
[Figure C.3.111] – One time differenced India time series



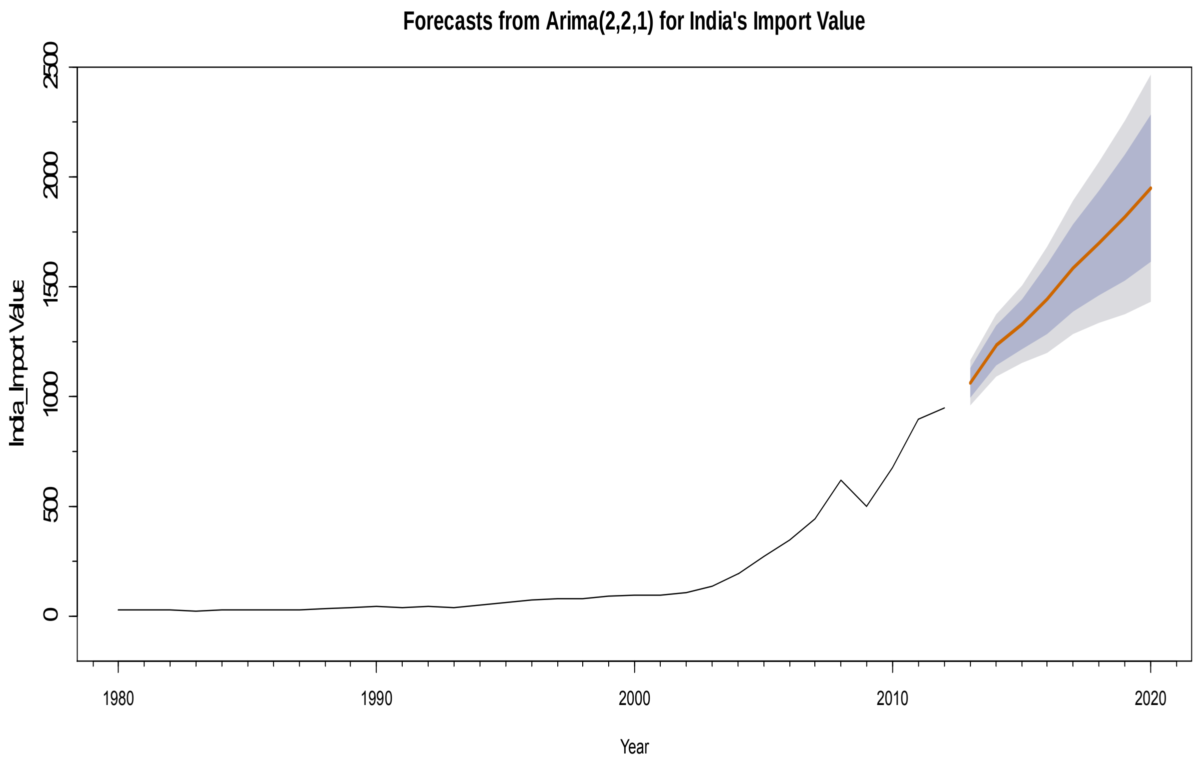
[Figure C.3.112] – Two times differenced India time series



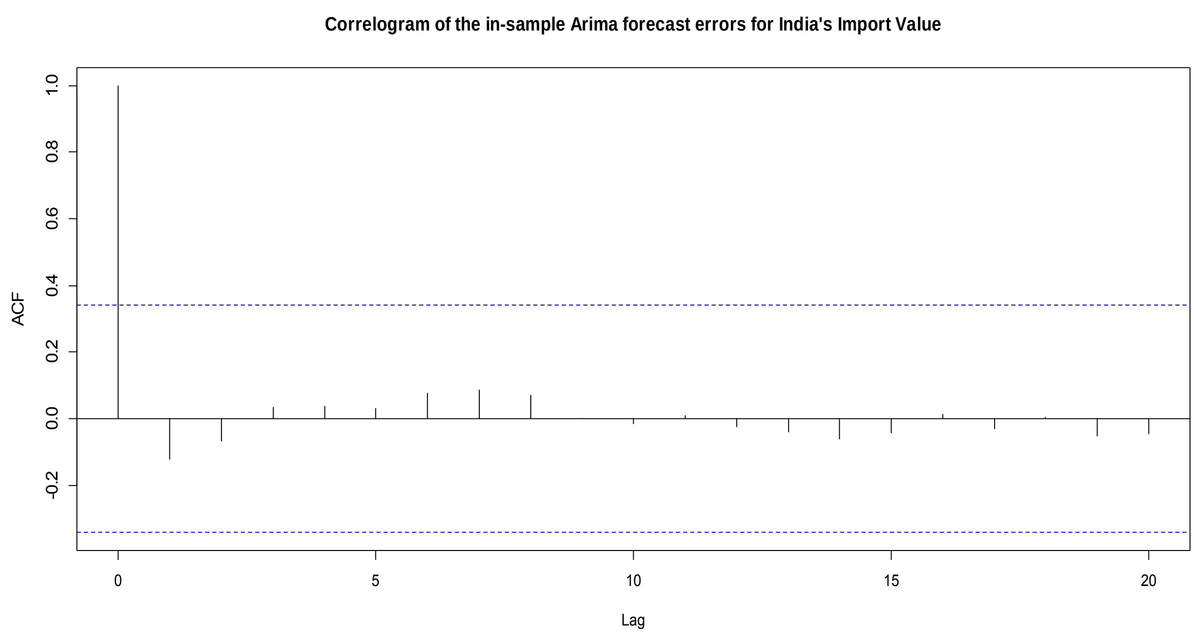
[Figure C.3.113] – Autocorrelogram (ACF) of the twice differenced India time series



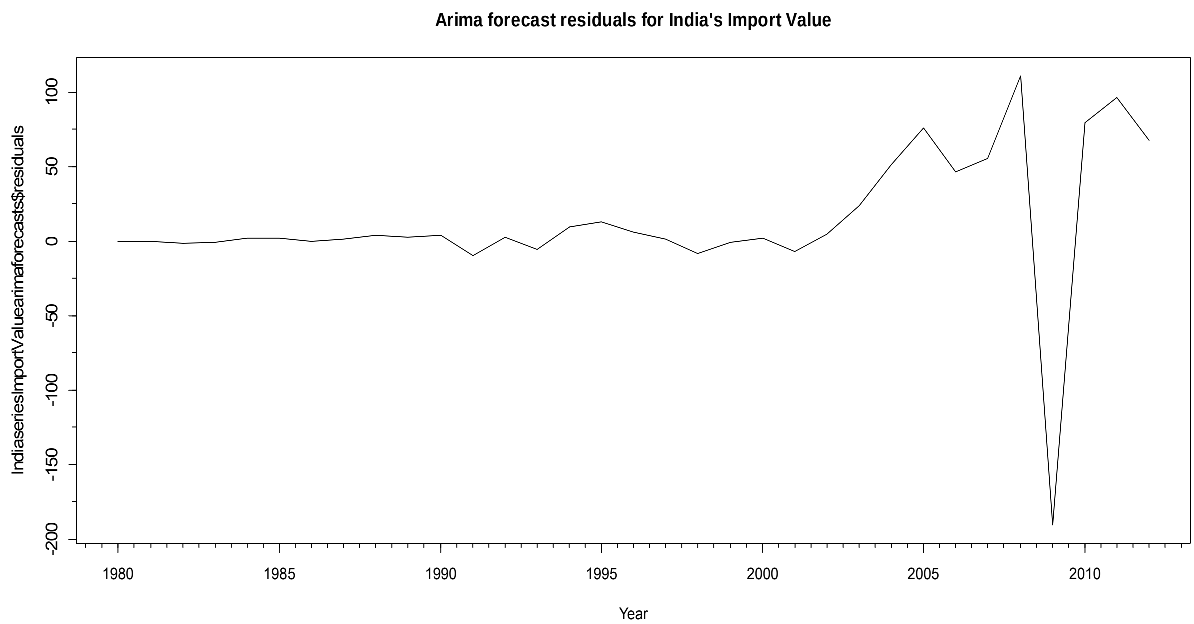
[Figure C.3.114] – Partial autocorrelogram (PACF) of the twice differenced India time series



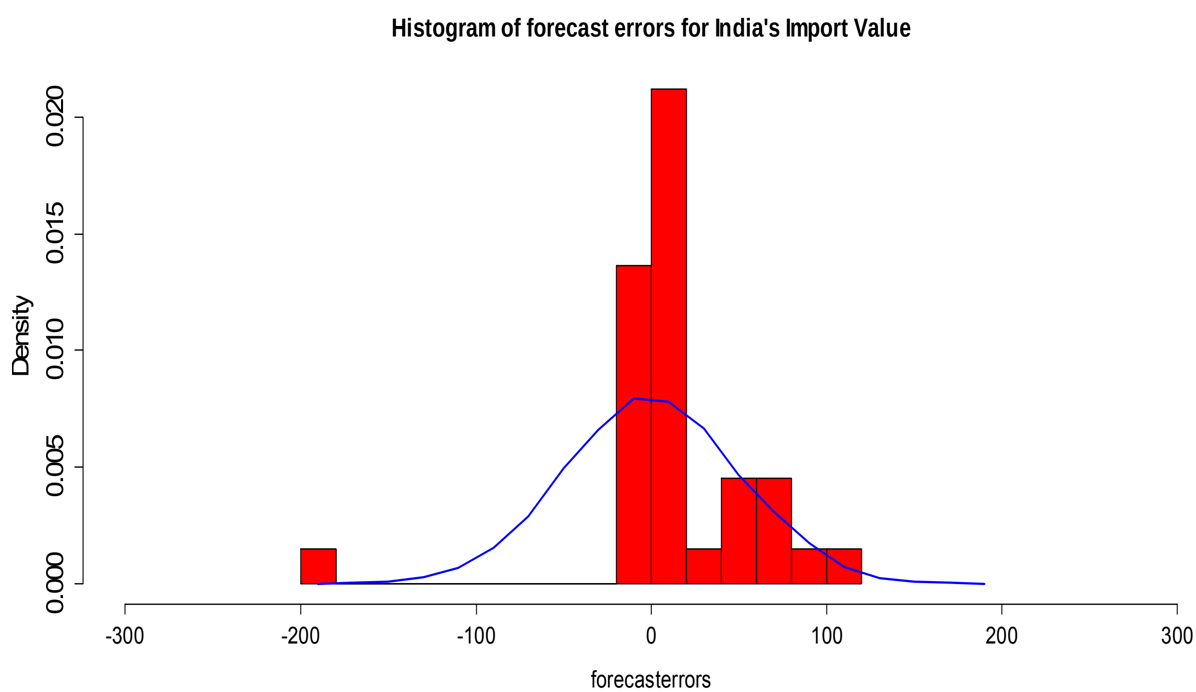
[Figure C.3.115] – Analysis for India, Import Value and whole dataset



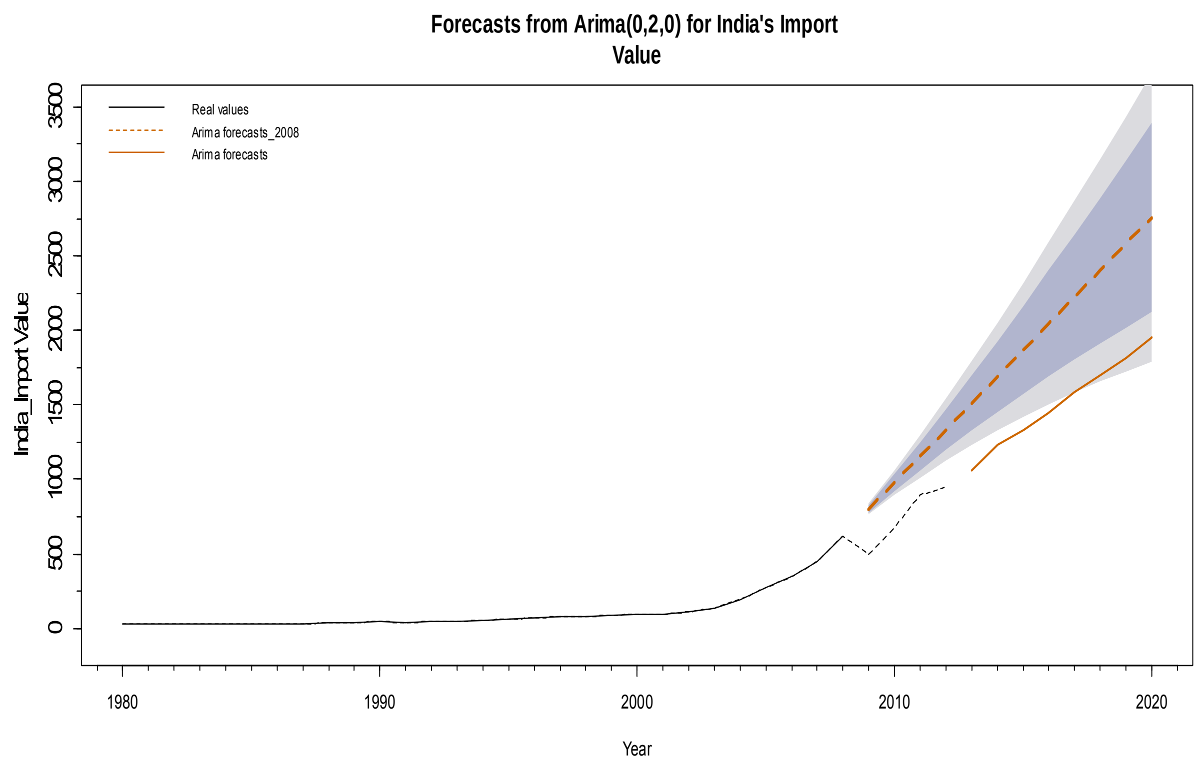
[Figure C.3.116] – Correlogram of in-sample errors of ARIMA forecasts



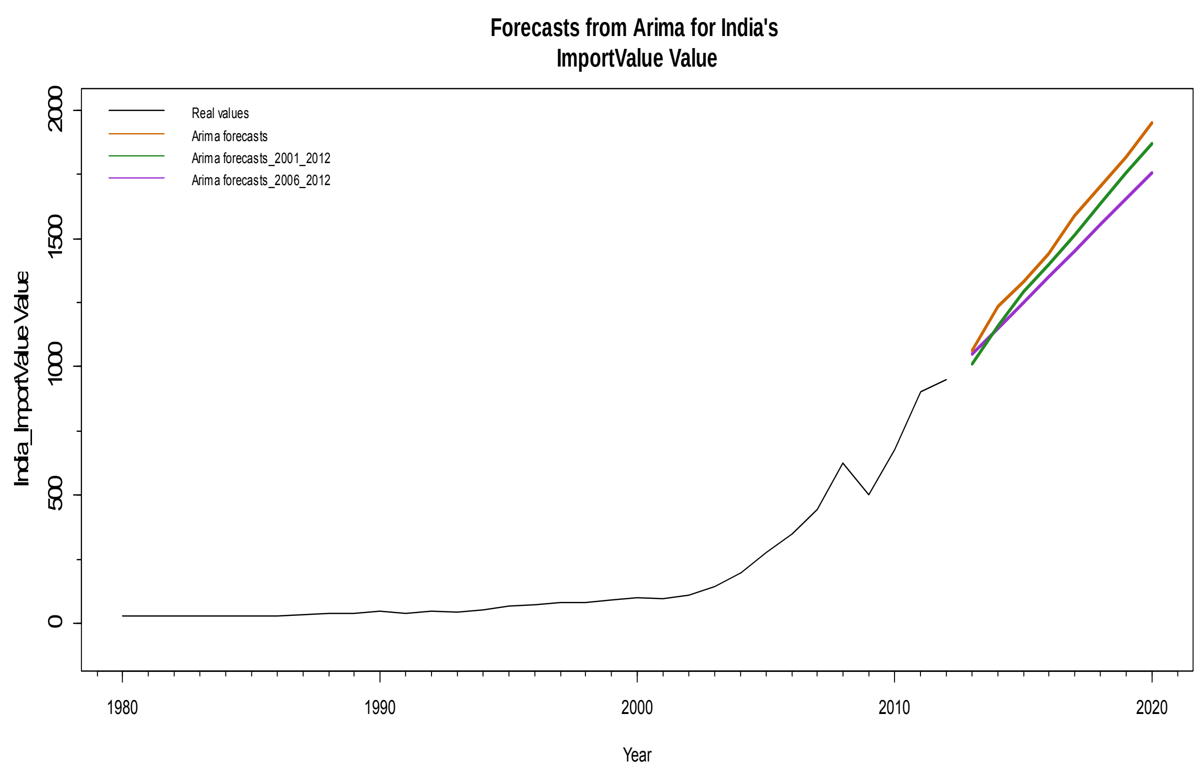
[Figure C.3.117] –Residuals of ARIMA forecasts



[Figure C.3.118] – Histogram and distribution of forecast residuals

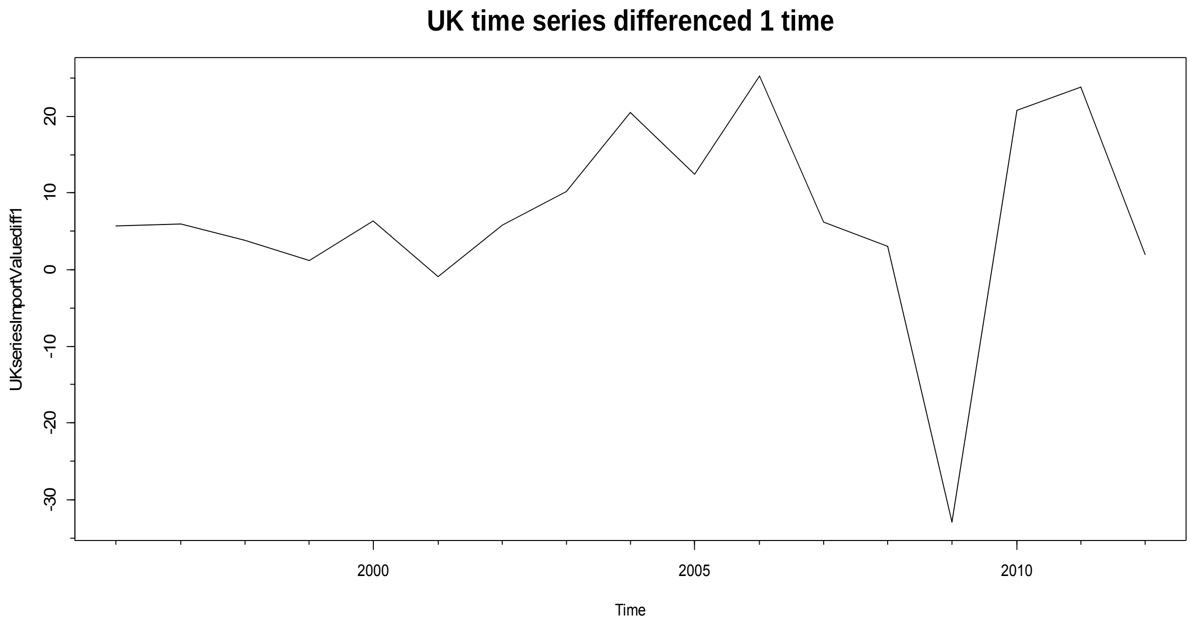


[Figure C.3.119] – Analysis for India, Import Value and the dataset up to 2008

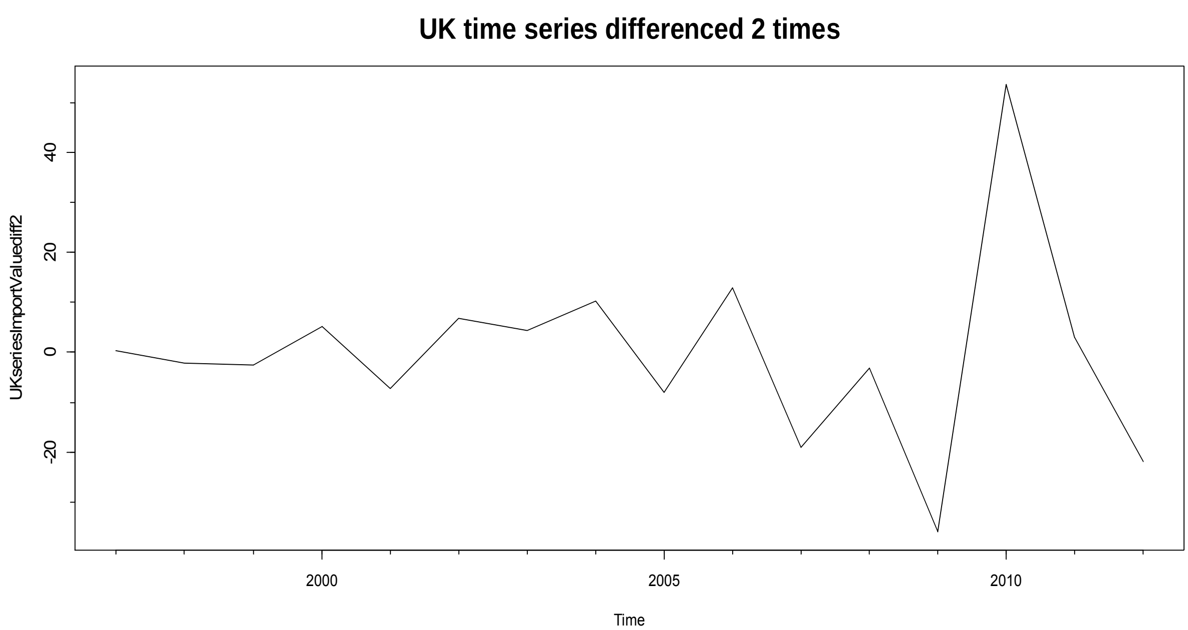


[Figure C.3.120] – Analyses for India, Import Value and the subsets 2001-2013 and 2006-2013

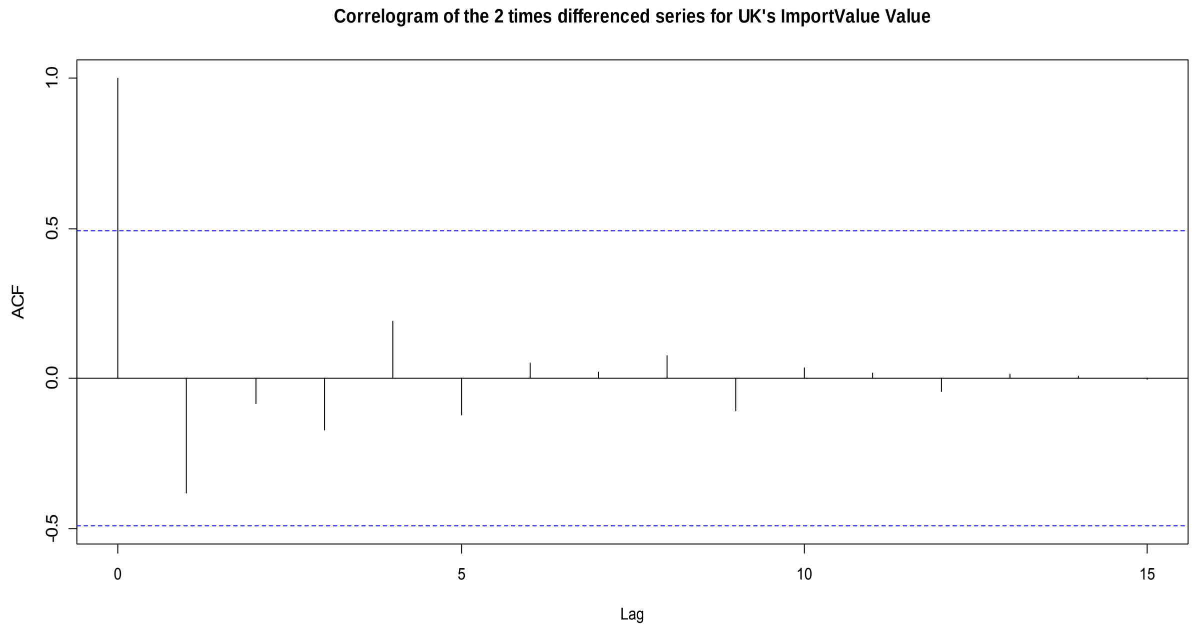
Import Value – UK



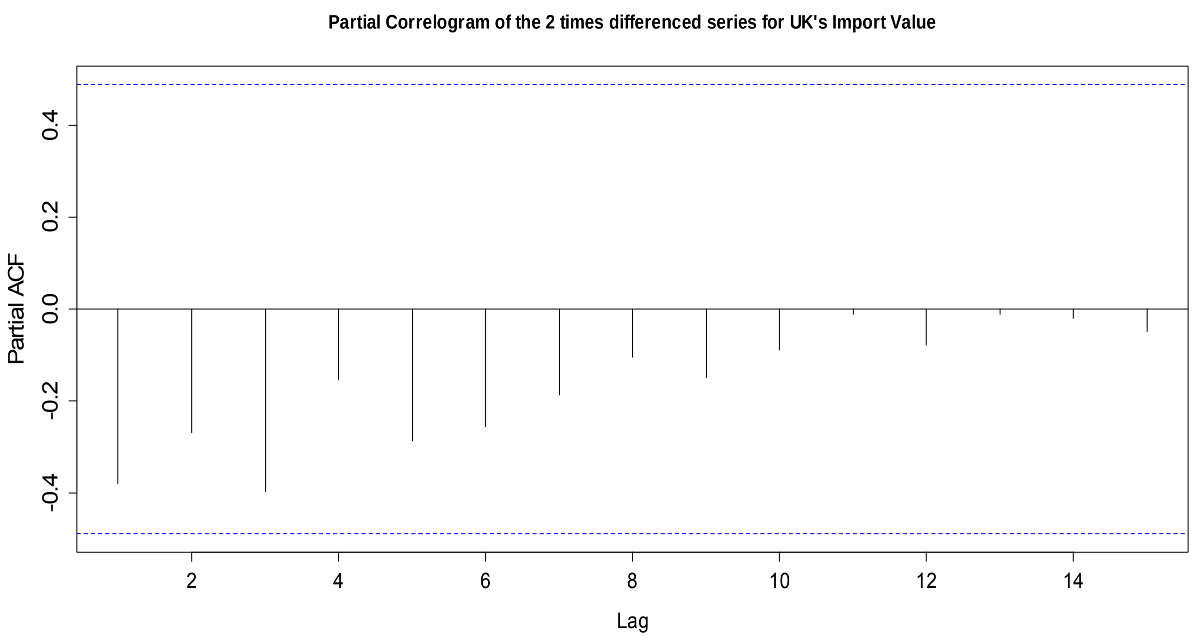
[Figure C.3.121] – One time differenced UK time series



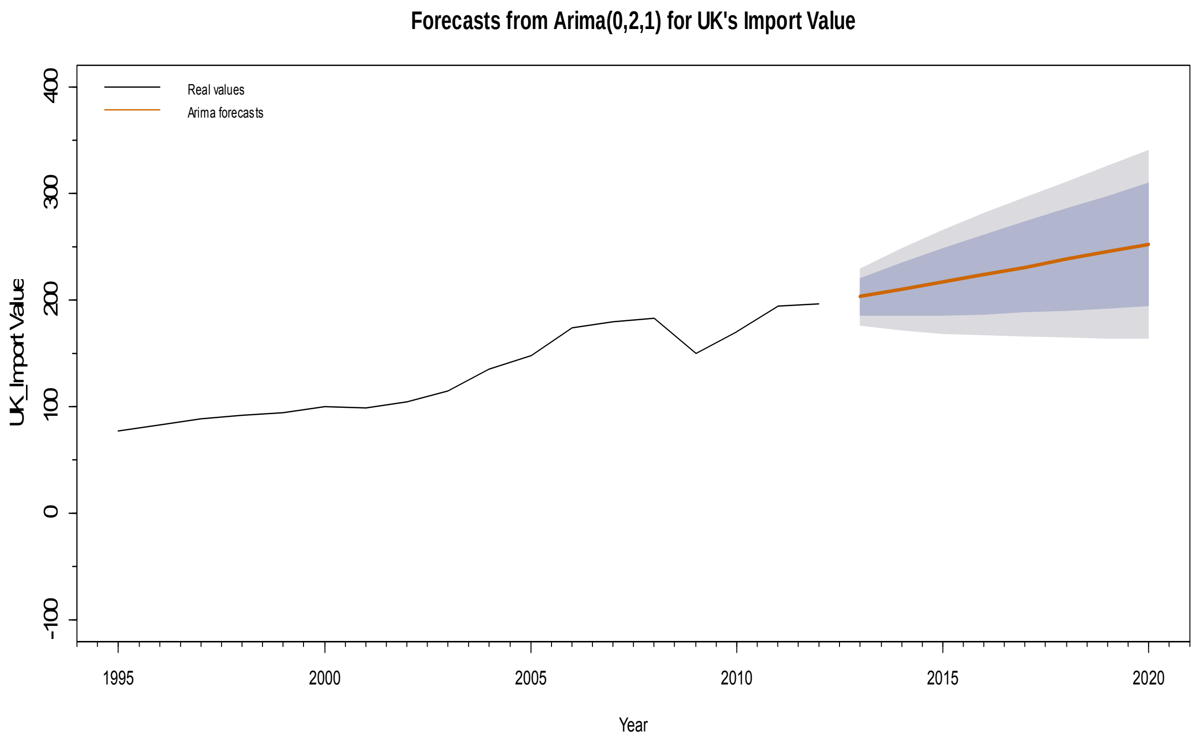
[Figure C.3.122] – Two times differenced UK time series



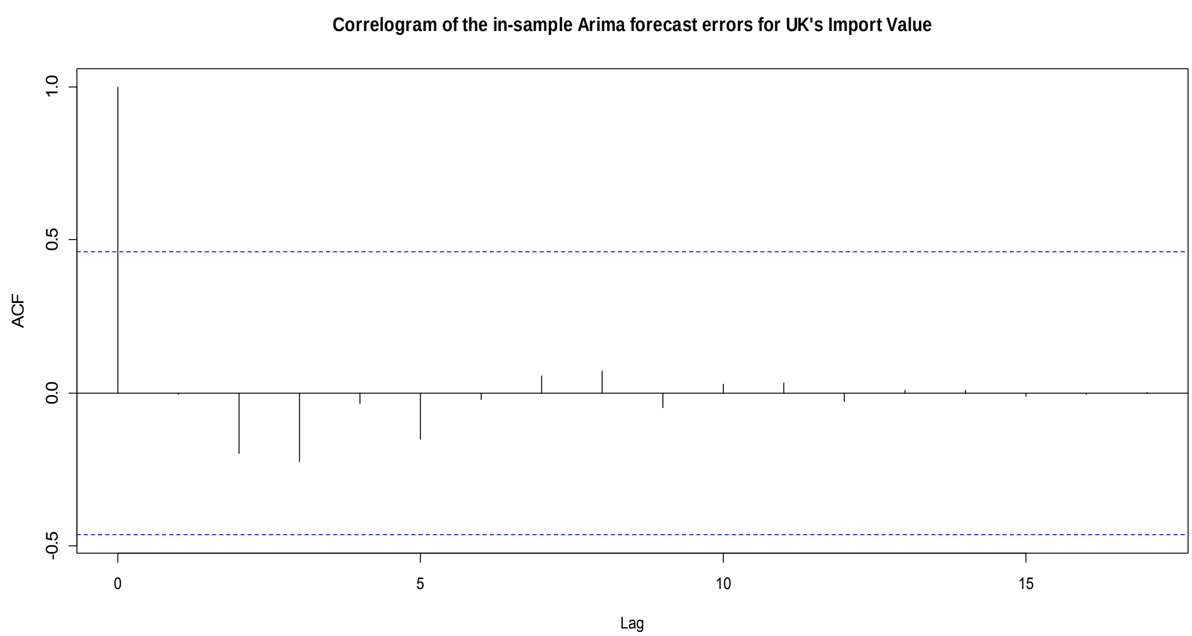
[Figure C.3.123] – Autocorrelogram (ACF) of the twice differenced UK time series



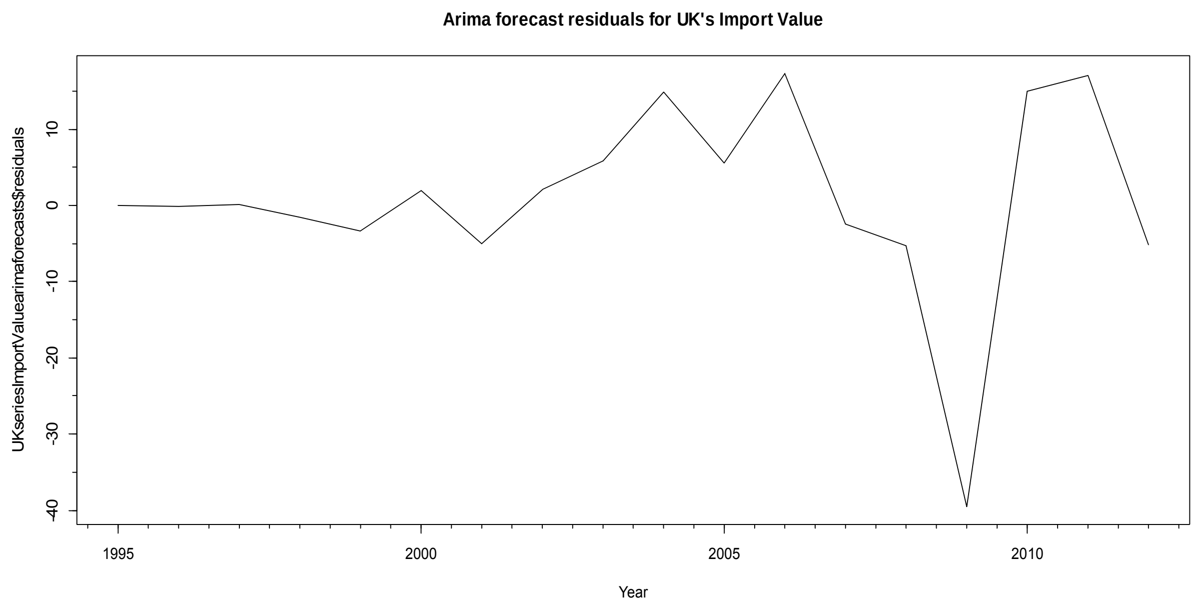
[Figure C.3.124] – Partial autocorrelogram (PACF) of the twice differenced UK time series



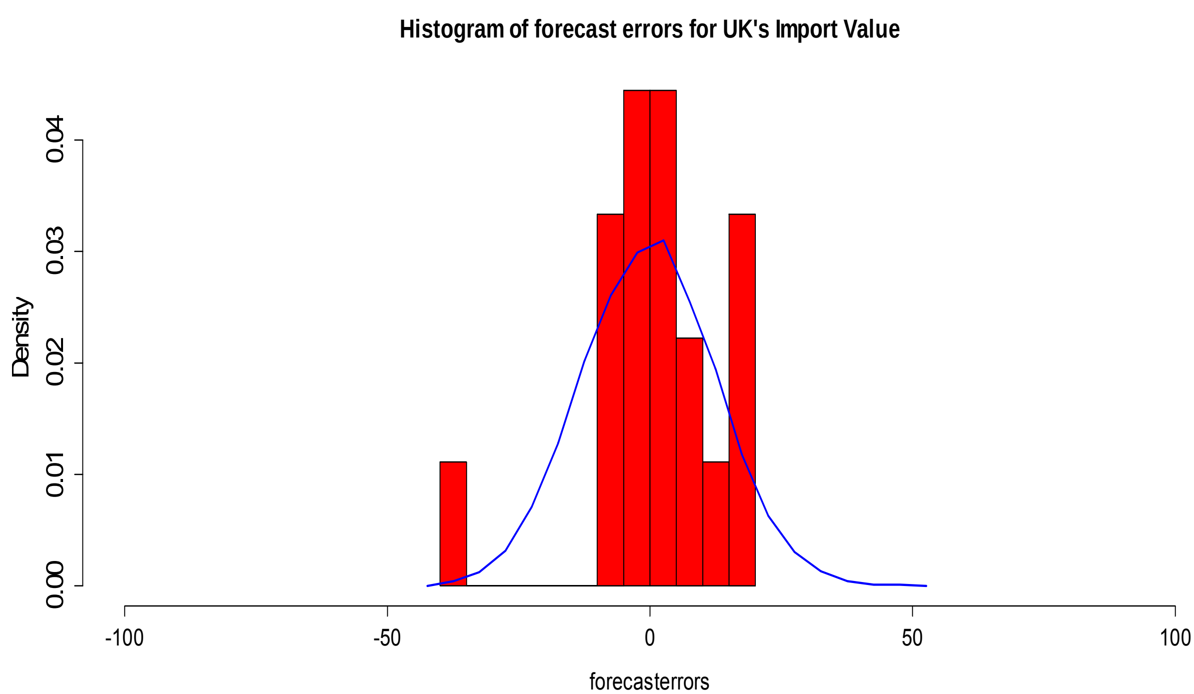
[Figure C.3.125] – Analysis for UK, Import Value and whole dataset



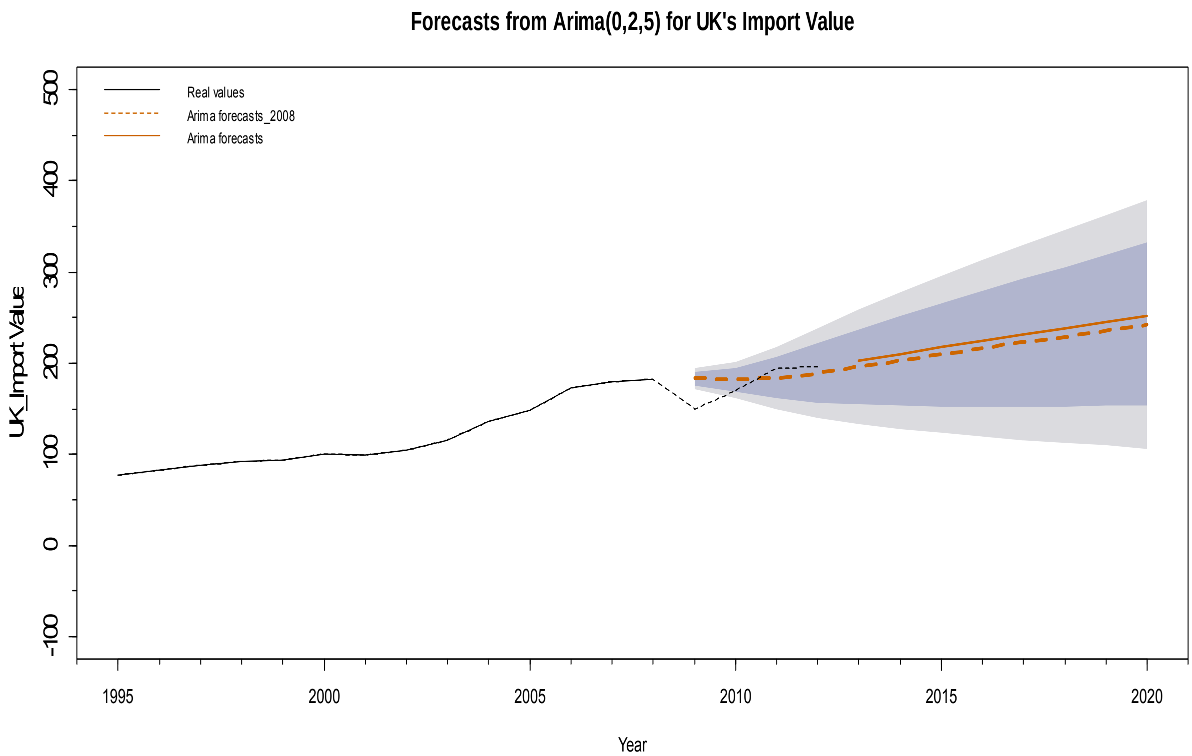
[Figure C.3.126] – Correlogram of in-sample errors of ARIMA forecasts



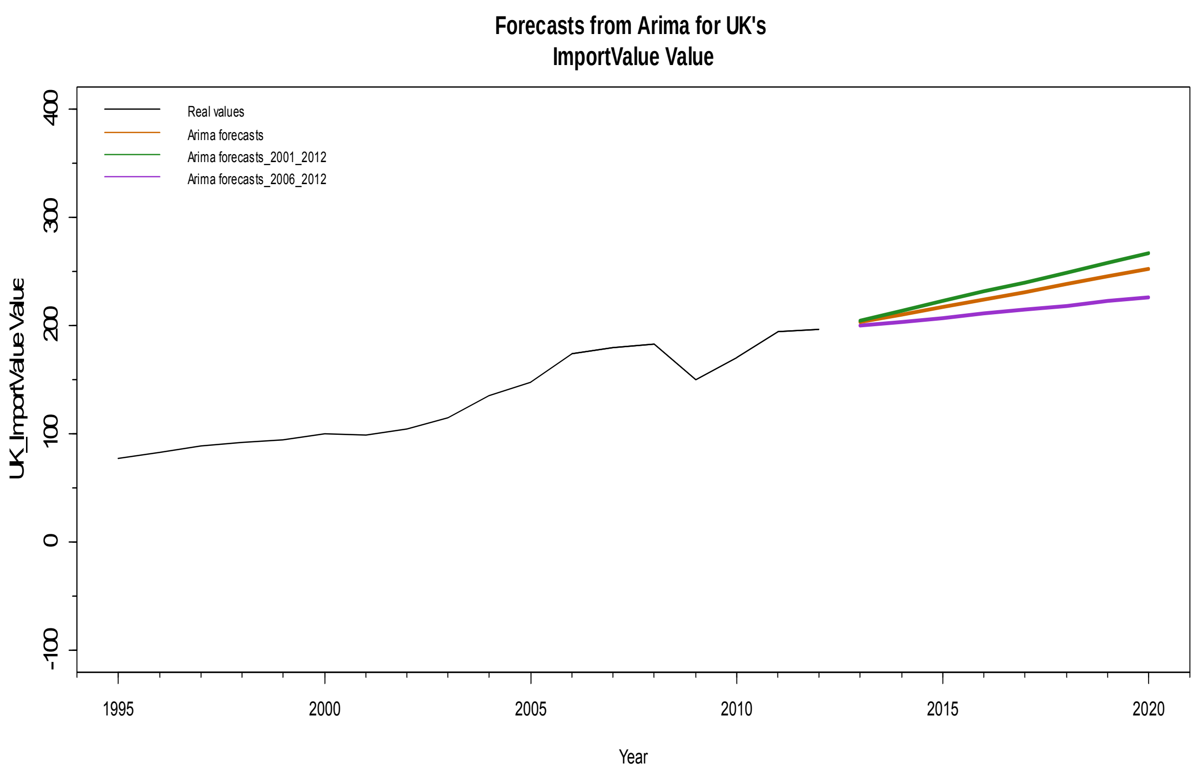
[Figure C.3.127] –Residuals of ARIMA forecasts



[Figure C.3.128] – Histogram and distribution of forecast residuals

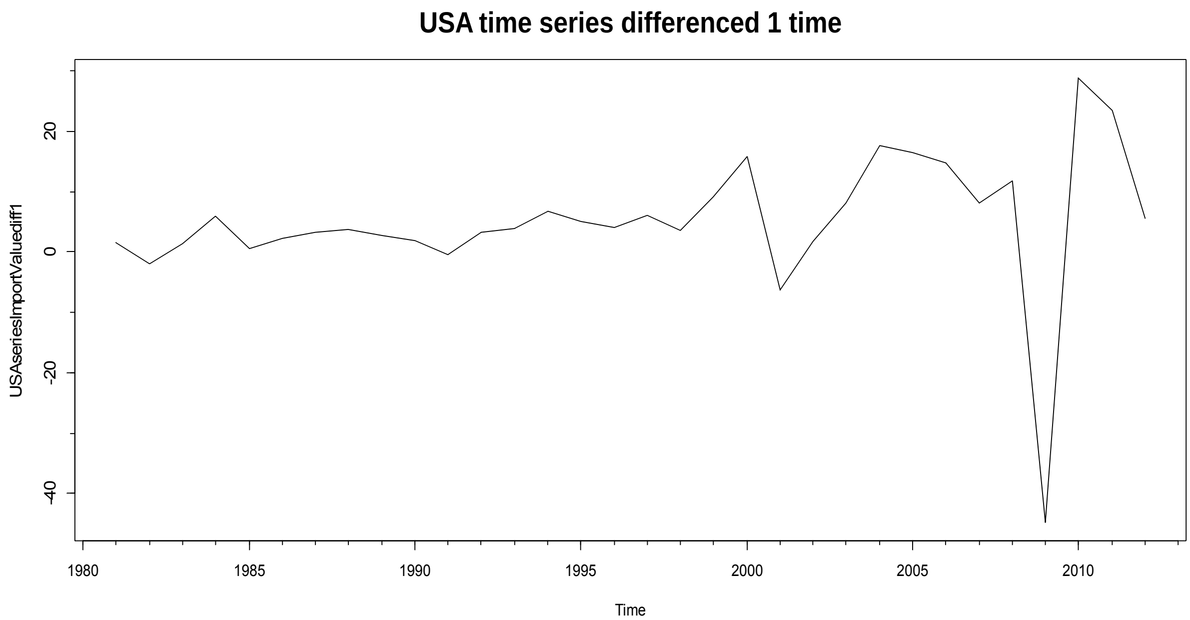


[Figure C.3.129] – Analysis for UK, Import Value and the dataset up to 2008

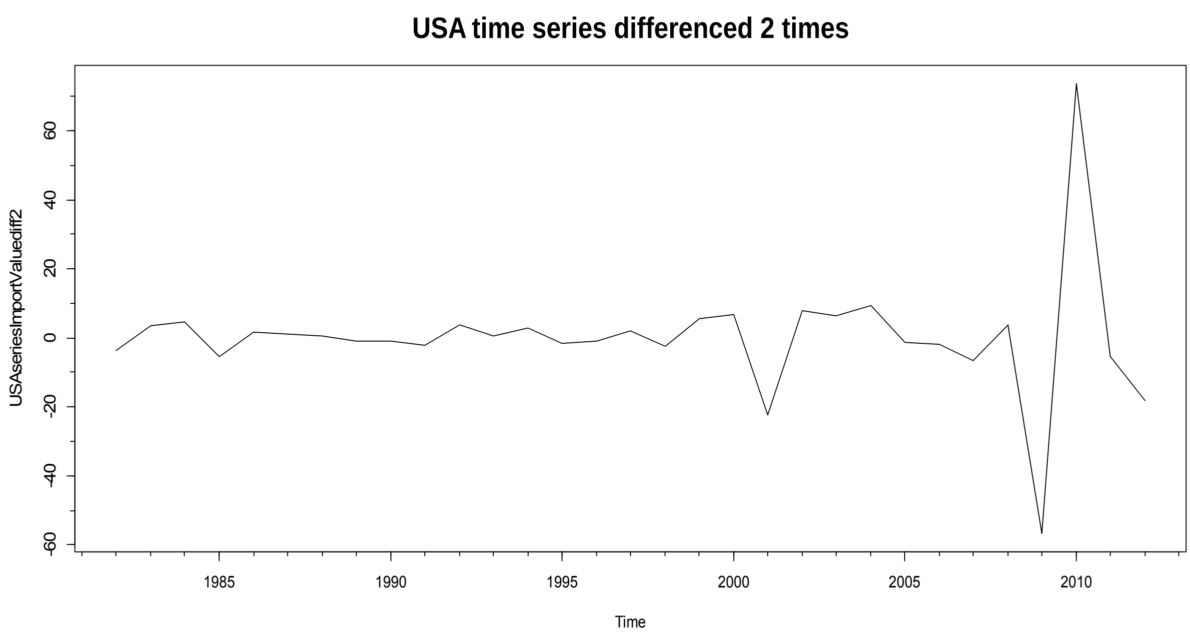


[Figure C.3.130] – Analyses for UK, Import Value and the subsets 2001-2013 and 2006-2013

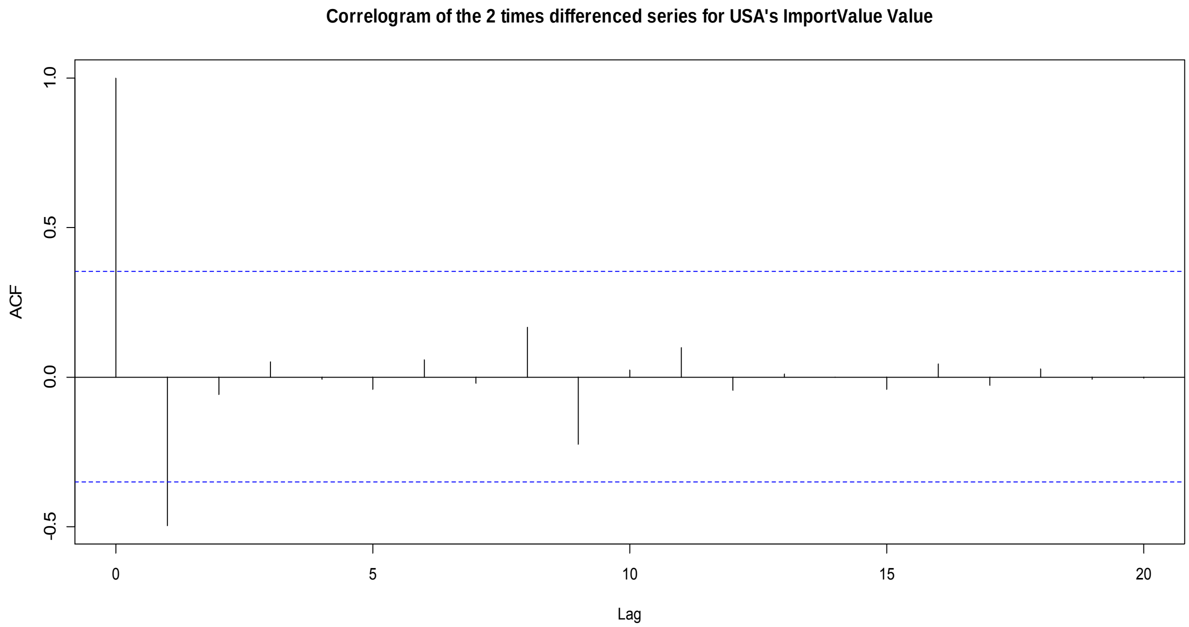
Import Value – USA



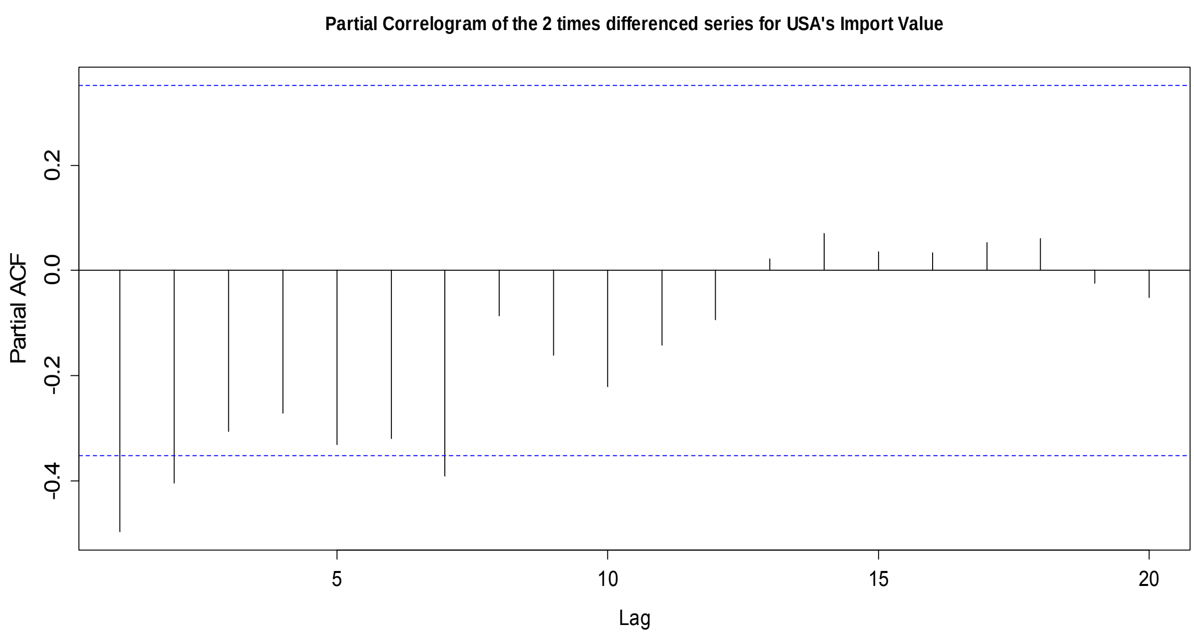
[Figure C.3.131] – One time differenced USA time series



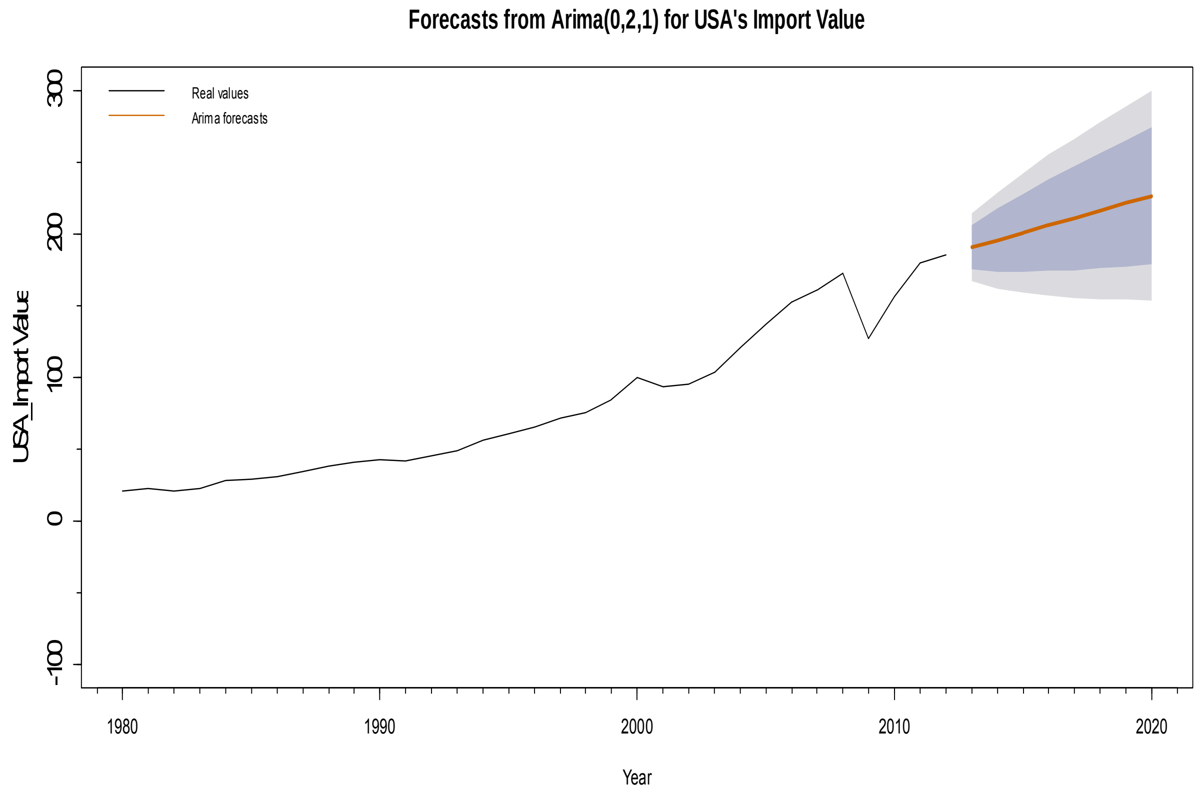
[Figure C.3.132] – Two times differenced USA time series



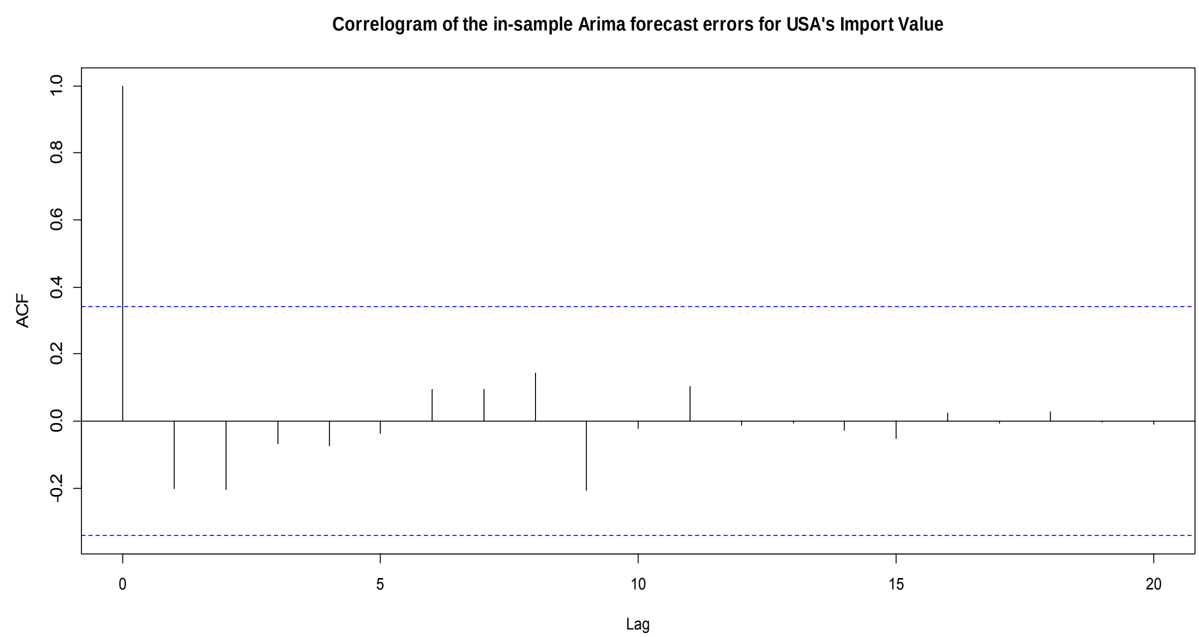
[Figure C.3.133] – Autocorrelogram (ACF) of the twice differenced USA time series



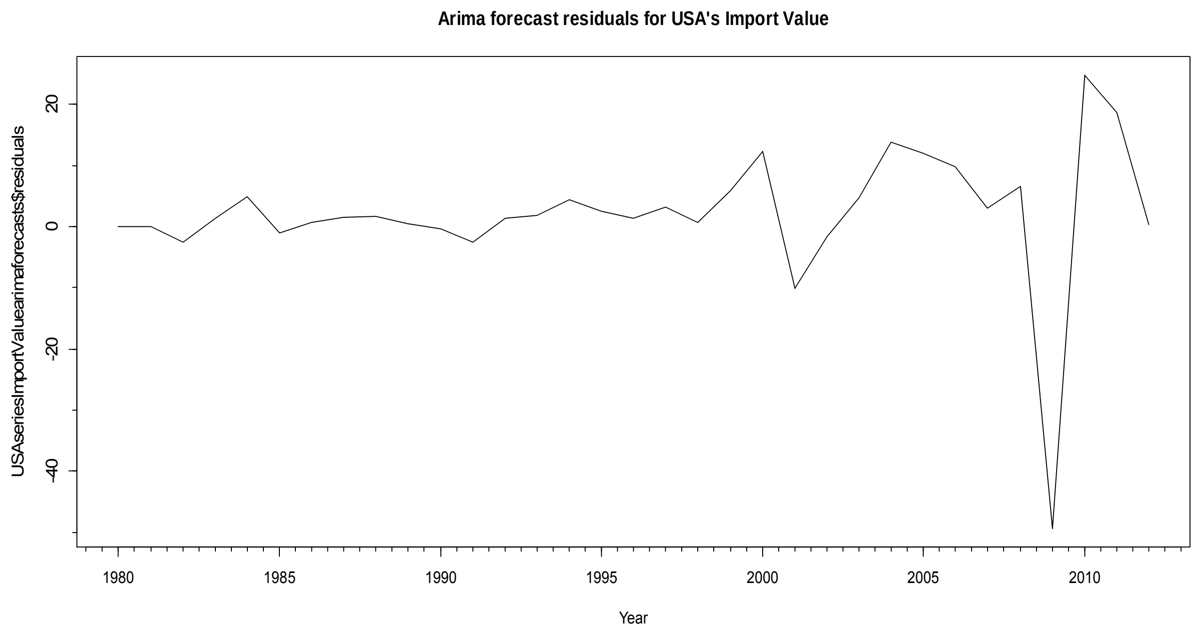
[Figure C.3.134] – Partial autocorrelogram (PACF) of the twice differenced USA time series



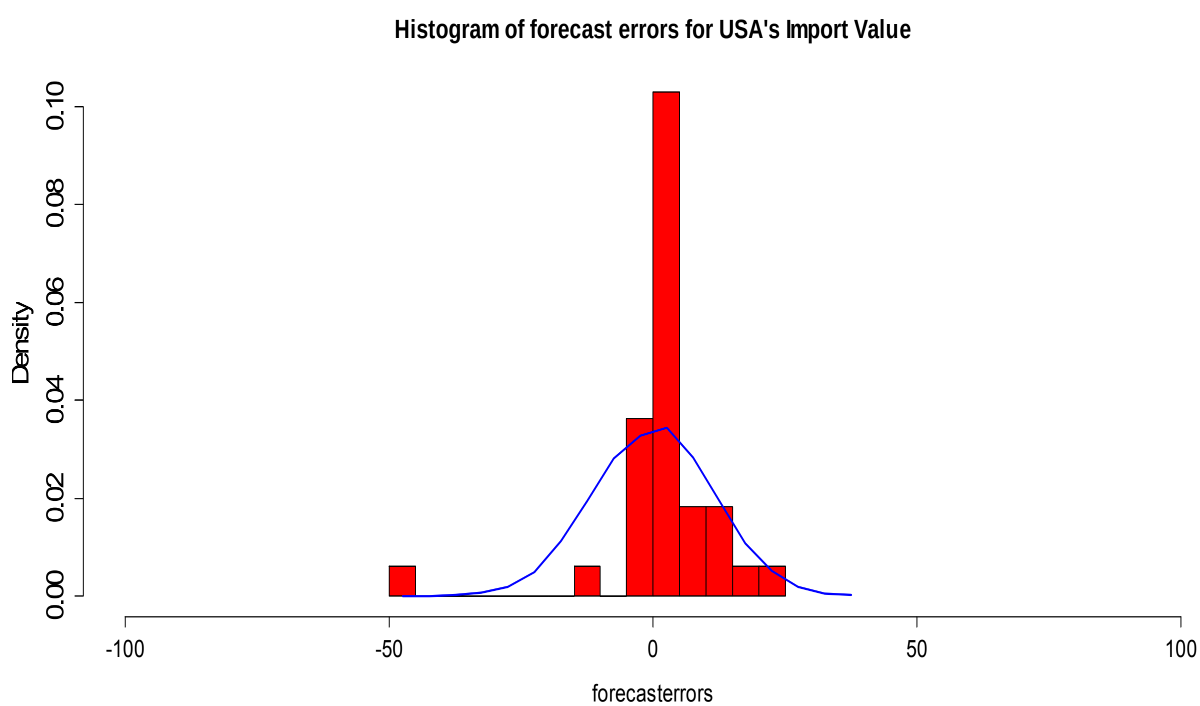
[Figure C.3.135] – Analysis for USA, Import Value and whole dataset



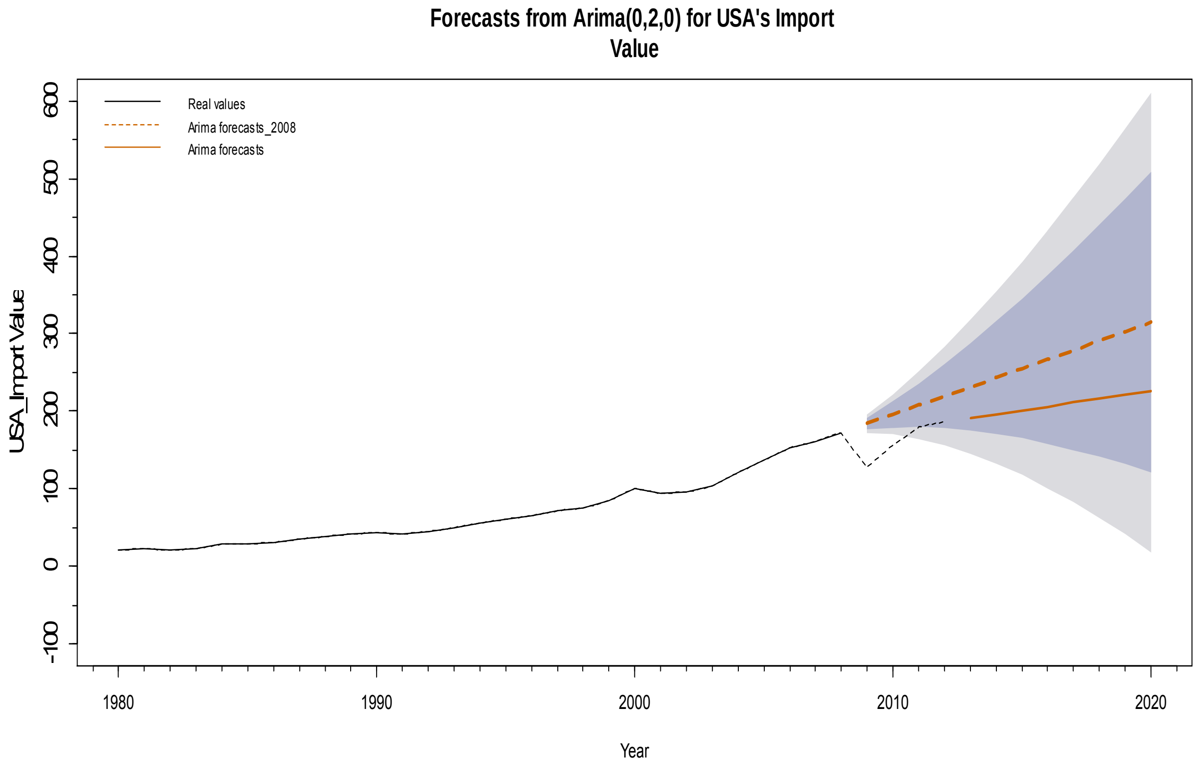
[Figure C.3.136] – Correlogram of in-sample errors of ARIMA forecasts



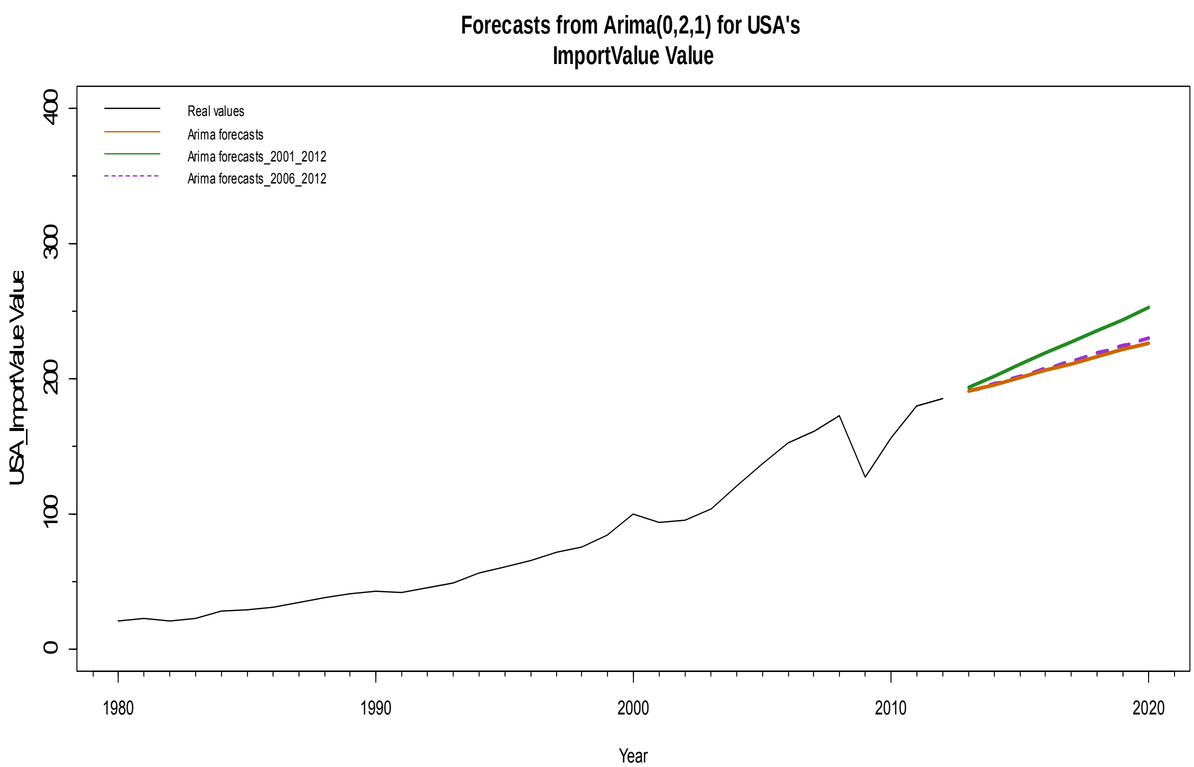
[Figure C.3.137] –Residuals of ARIMA forecasts



[Figure C.3.138] – Histogram and distribution of forecast residuals



[Figure C.3.139] – Analysis for USA, Import Value and the dataset up to 2008

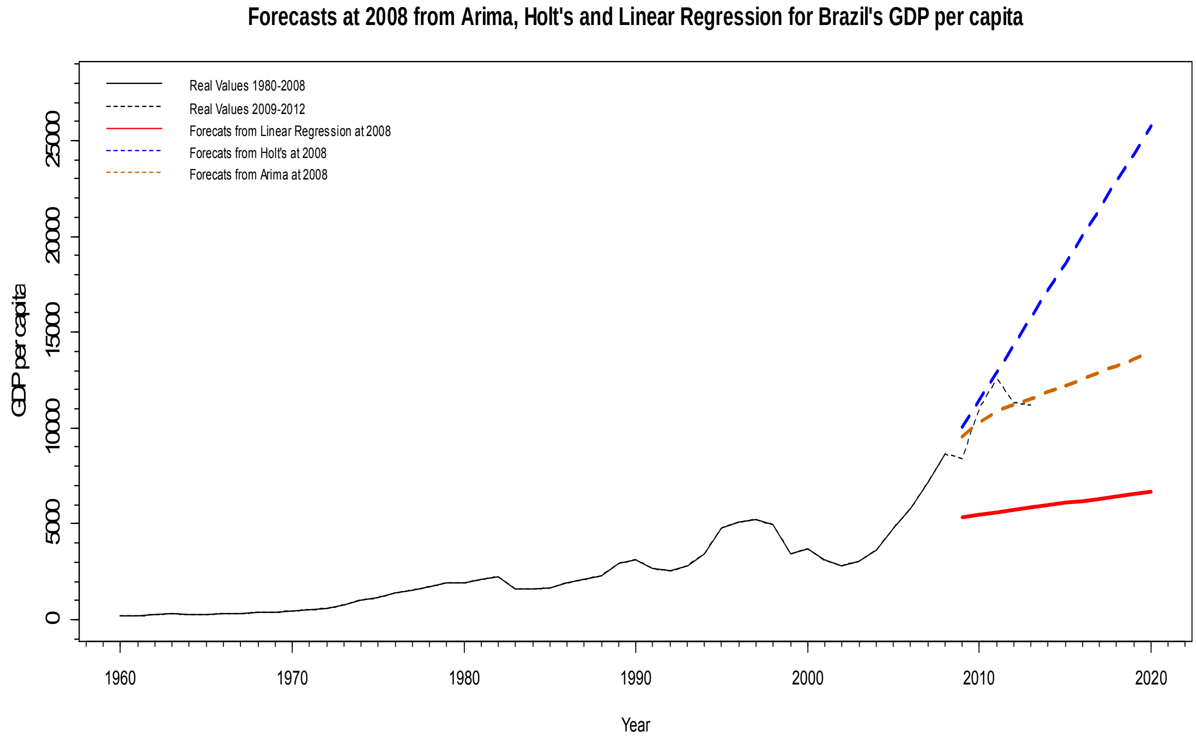


[Figure C.3.140] – Analyses for USA, Import Value and the subsets 2001-2013 and 2006-2013

Appendix D – Graphs from comparison of algorithms

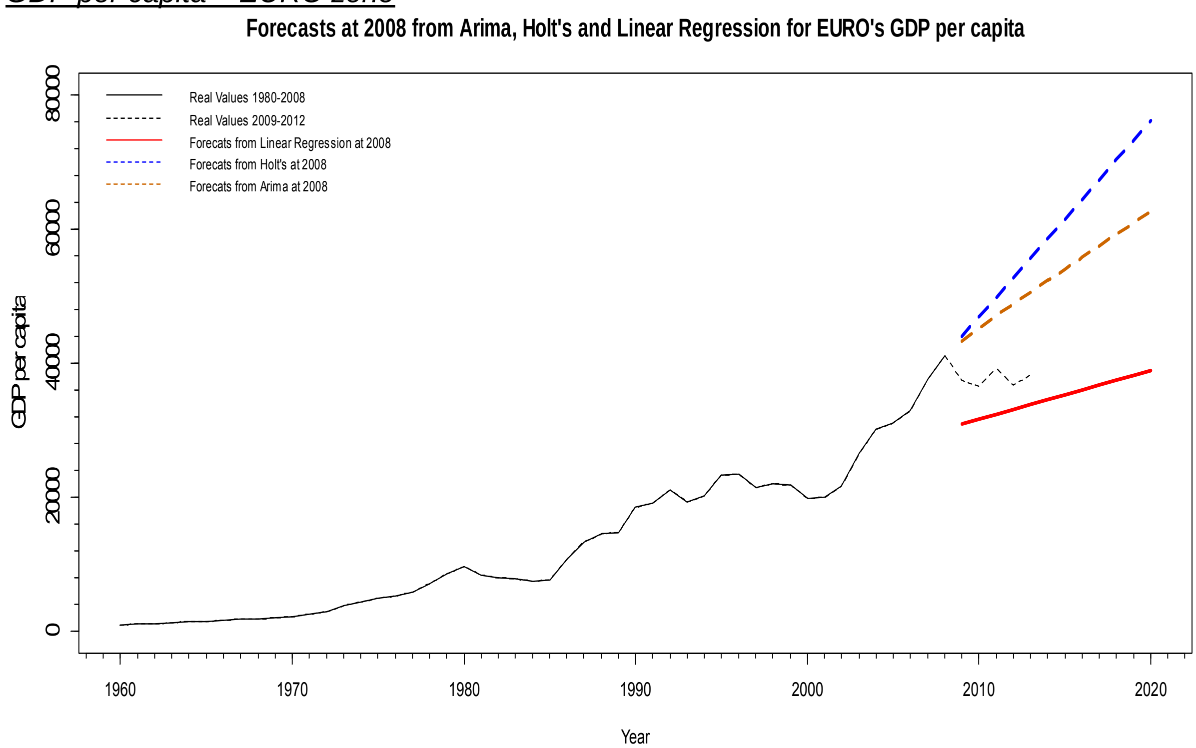
D.1. Graphs from evaluation of algorithms based on the forecasts made at year 2008

GDP per capita – Brazil



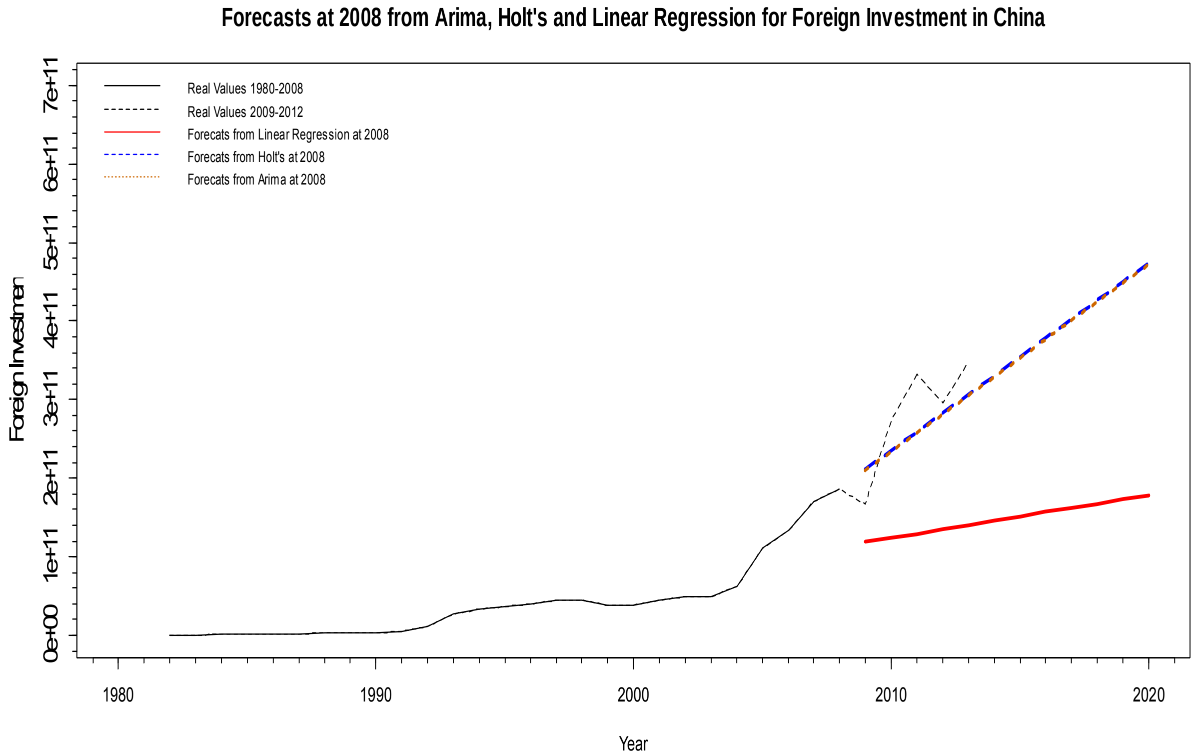
[Figure D.1.1] – Comparison of 2008 forecasts for Brazil and GDP per capita

GDP per capita – EURO zone



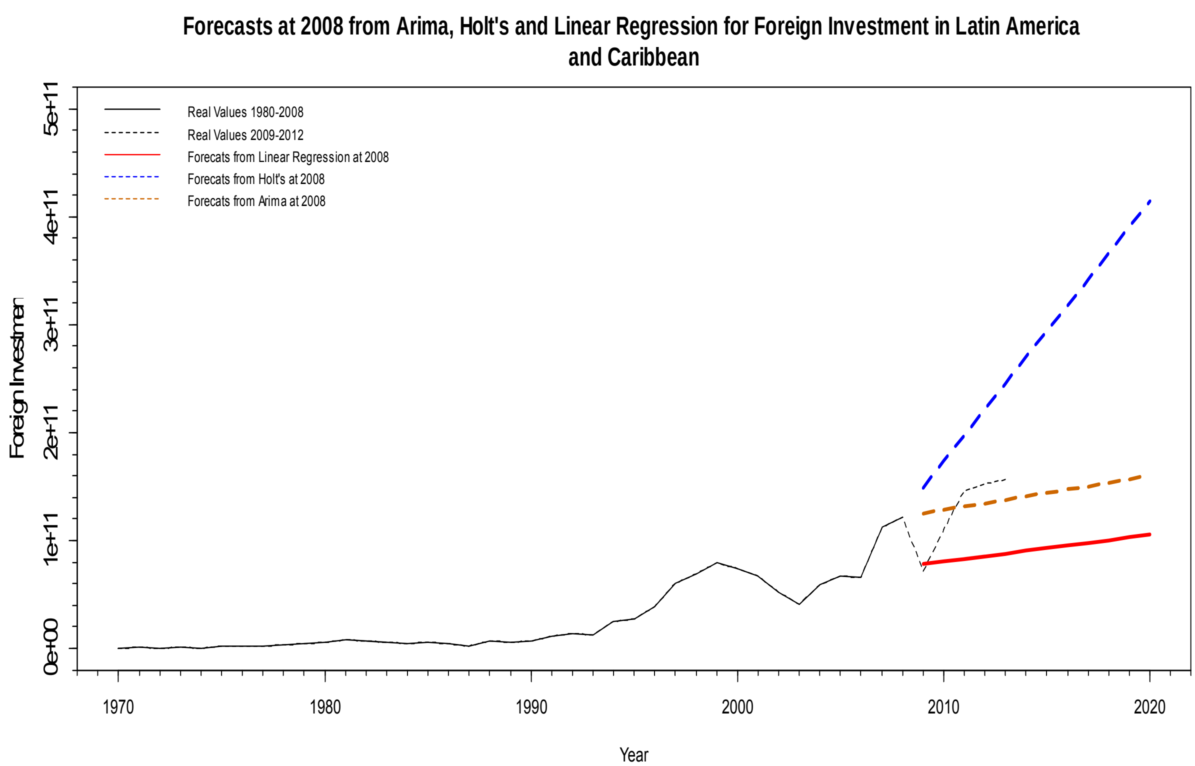
[Figure D.1.2] – Comparison of 2008 forecasts for EURO zone and GDP per capita

Foreign Investment – China



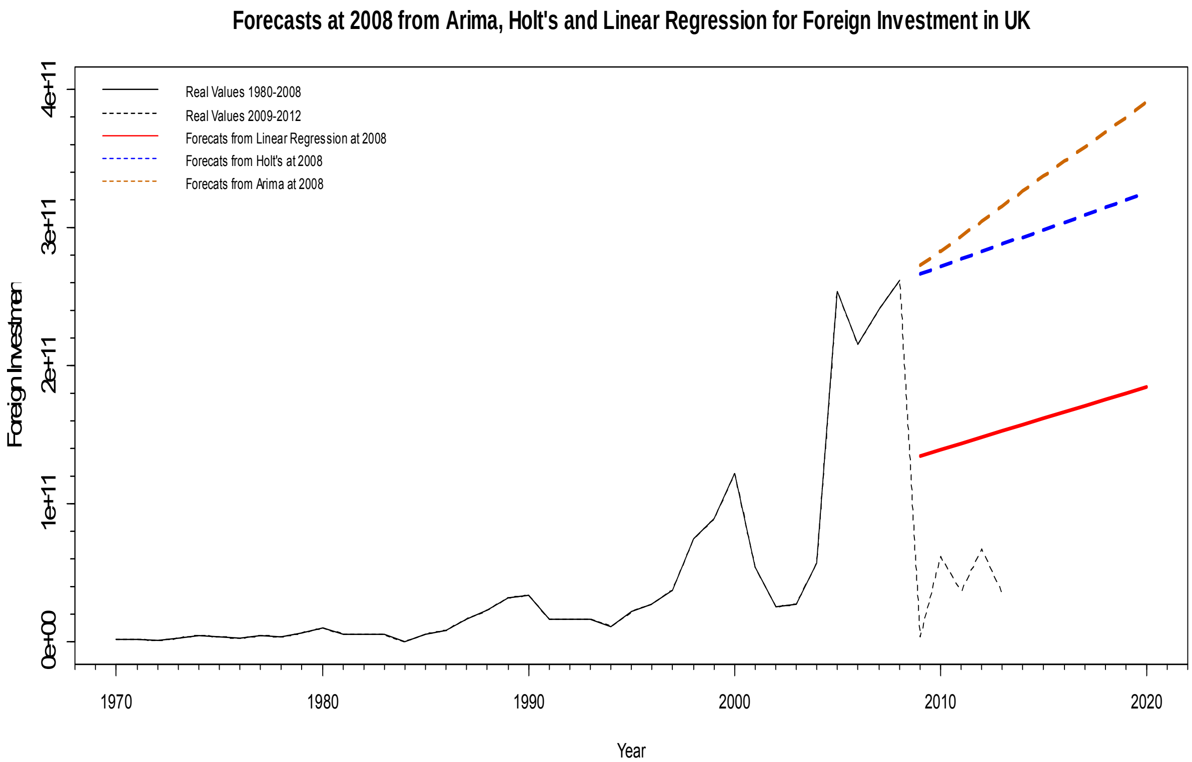
[Figure D.1.3] – Comparison of 2008 forecasts for China and Foreign Investment

Foreign Investment – Latin America and Caribbean



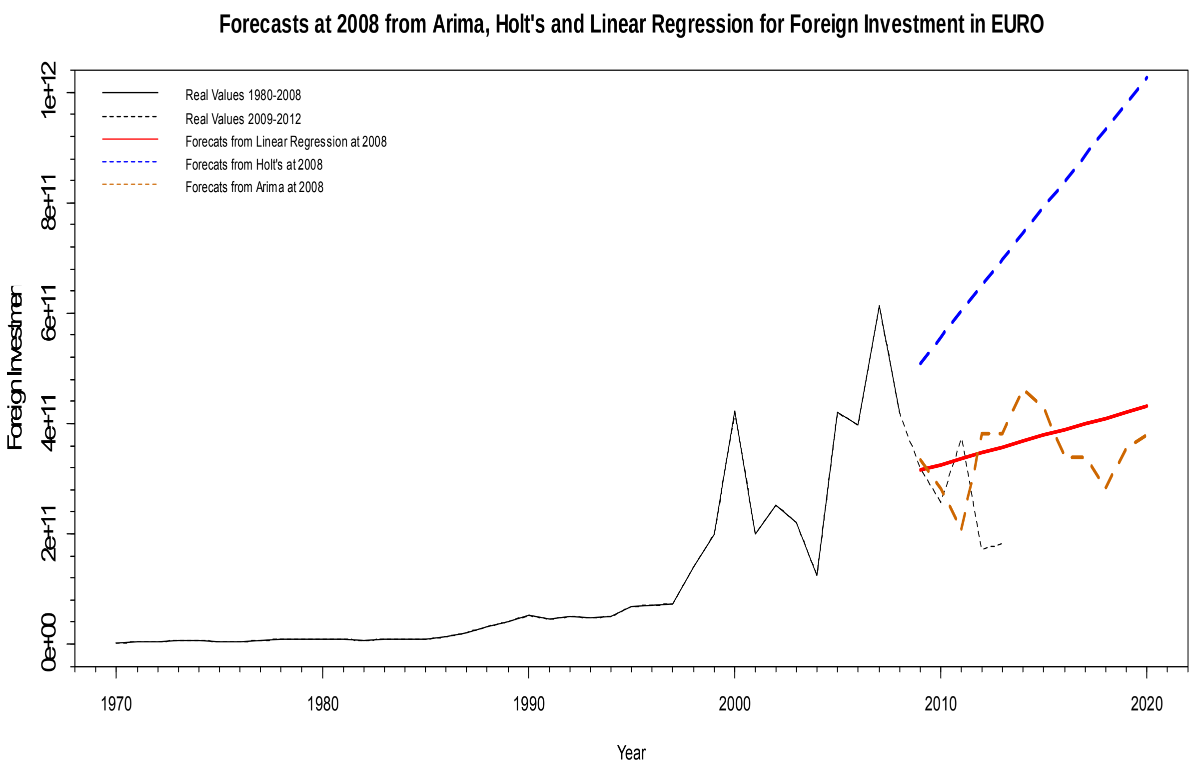
[Figure D.1.4] – Comparison of 2008 forecasts for Latin America and Foreign Investment

Foreign Investment – UK



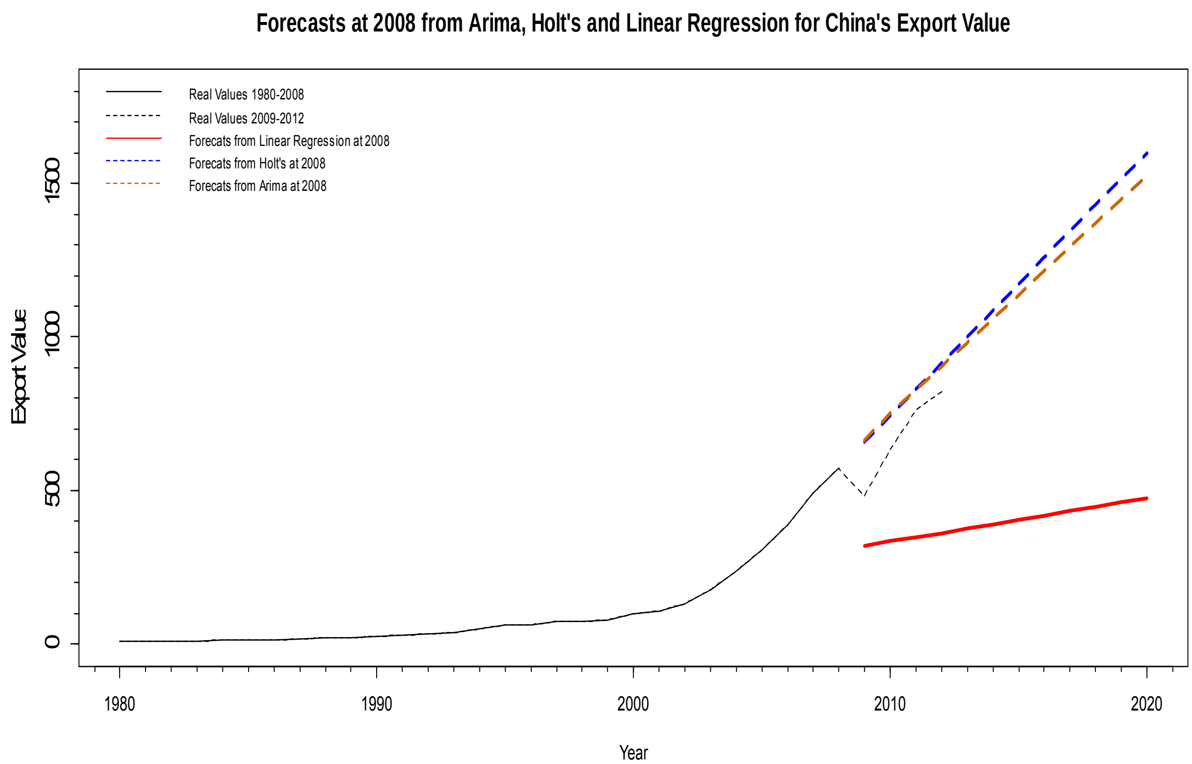
[Figure D.1.5] – Comparison of 2008 forecasts for UK and Foreign Investment

Foreign Investment – EURO zone



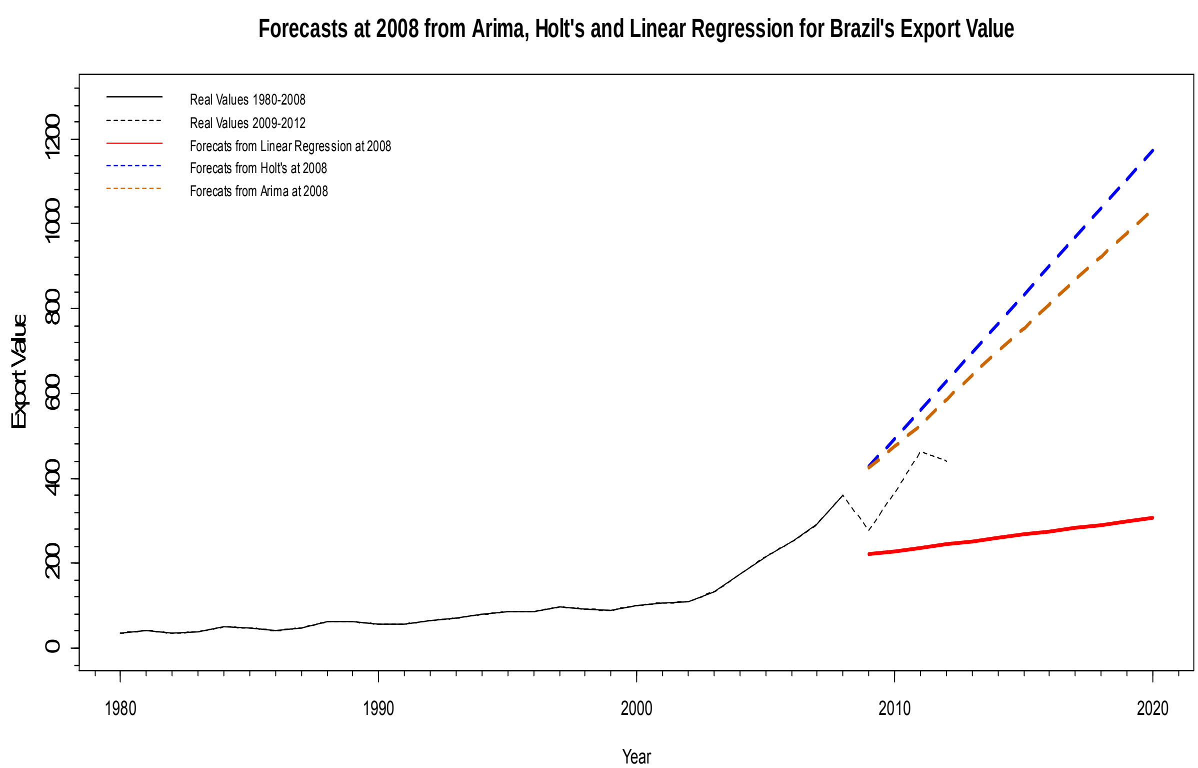
[Figure D.1.6] – Comparison of 2008 forecasts for UK and Foreign Investment

Export Value – China



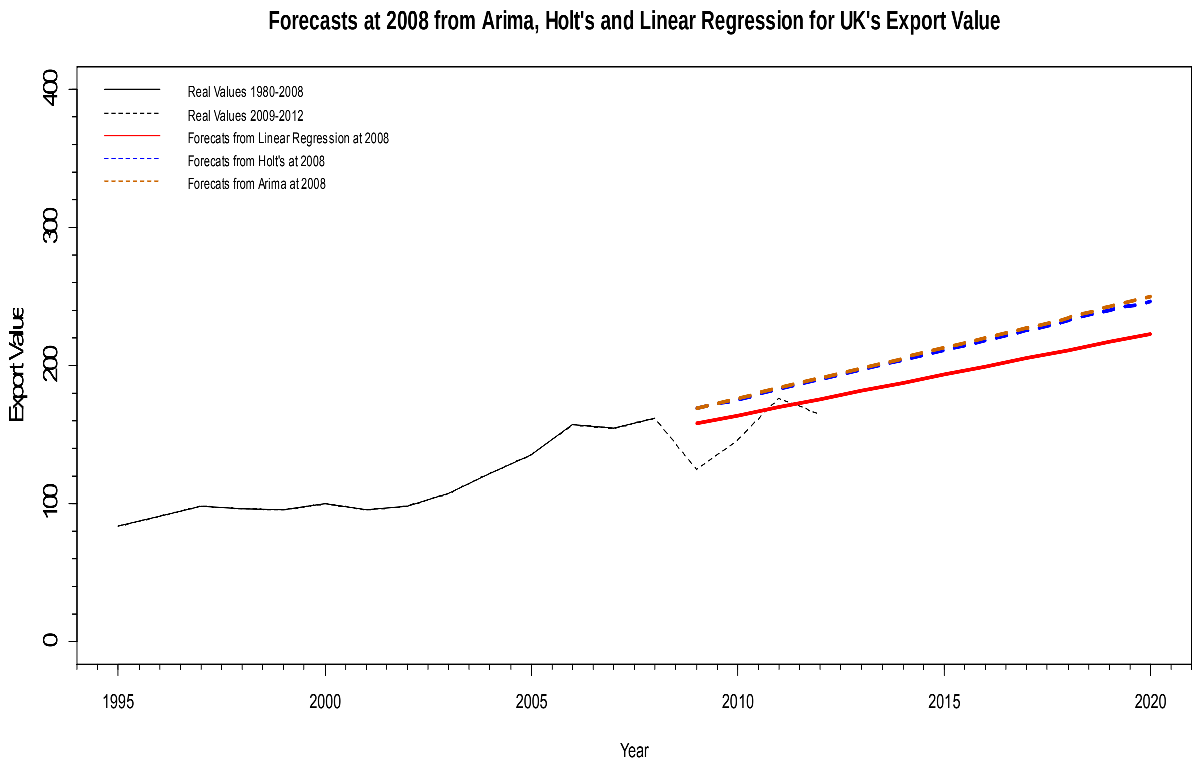
[Figure D.1.7] – Comparison of 2008 forecasts for China and Export Value

Export Value – Brazil



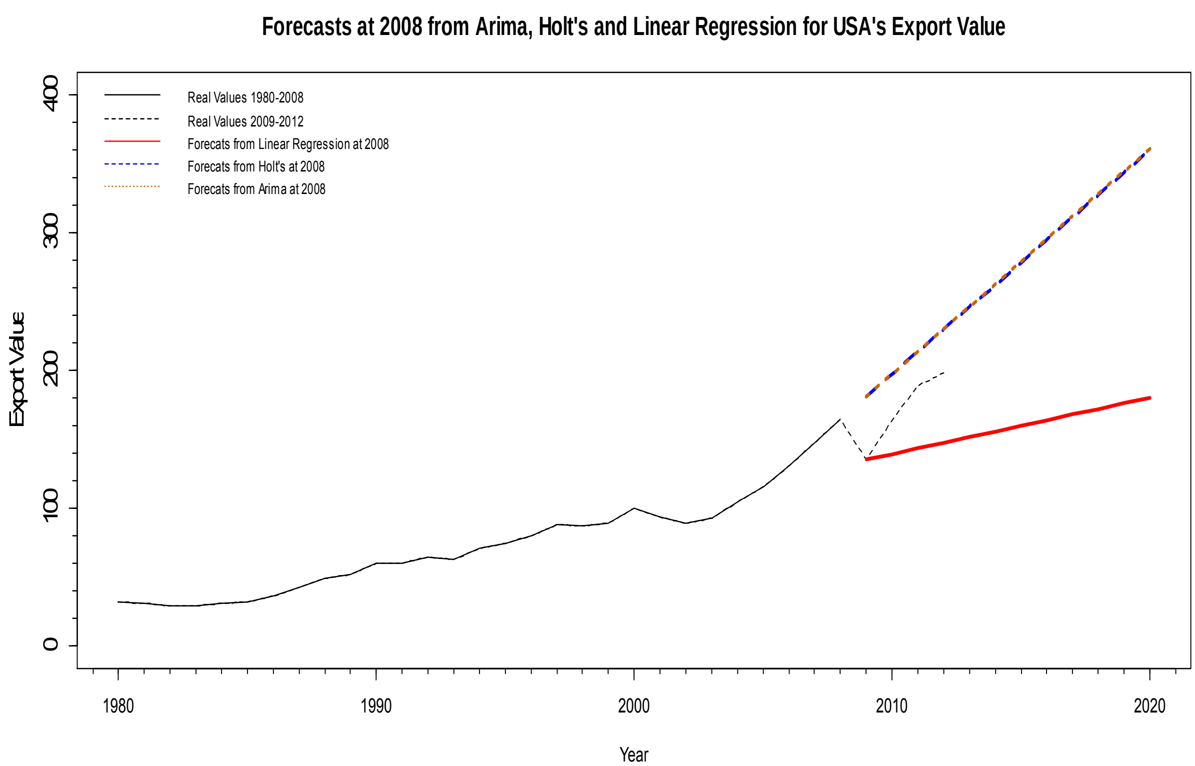
[Figure D.1.8] – Comparison of 2008 forecasts for Brazil and Export Value

Export Value – UK



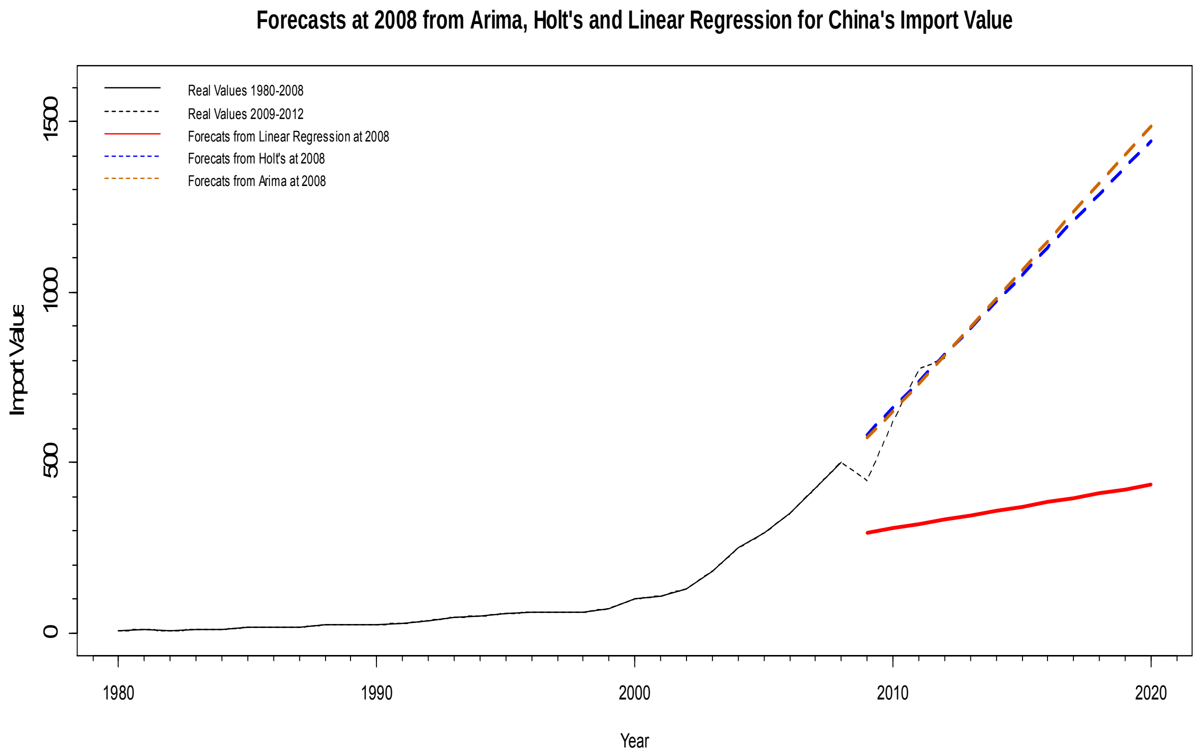
[Figure D.1.9] – Comparison of 2008 forecasts for UK and Export Value

Export Value – USA



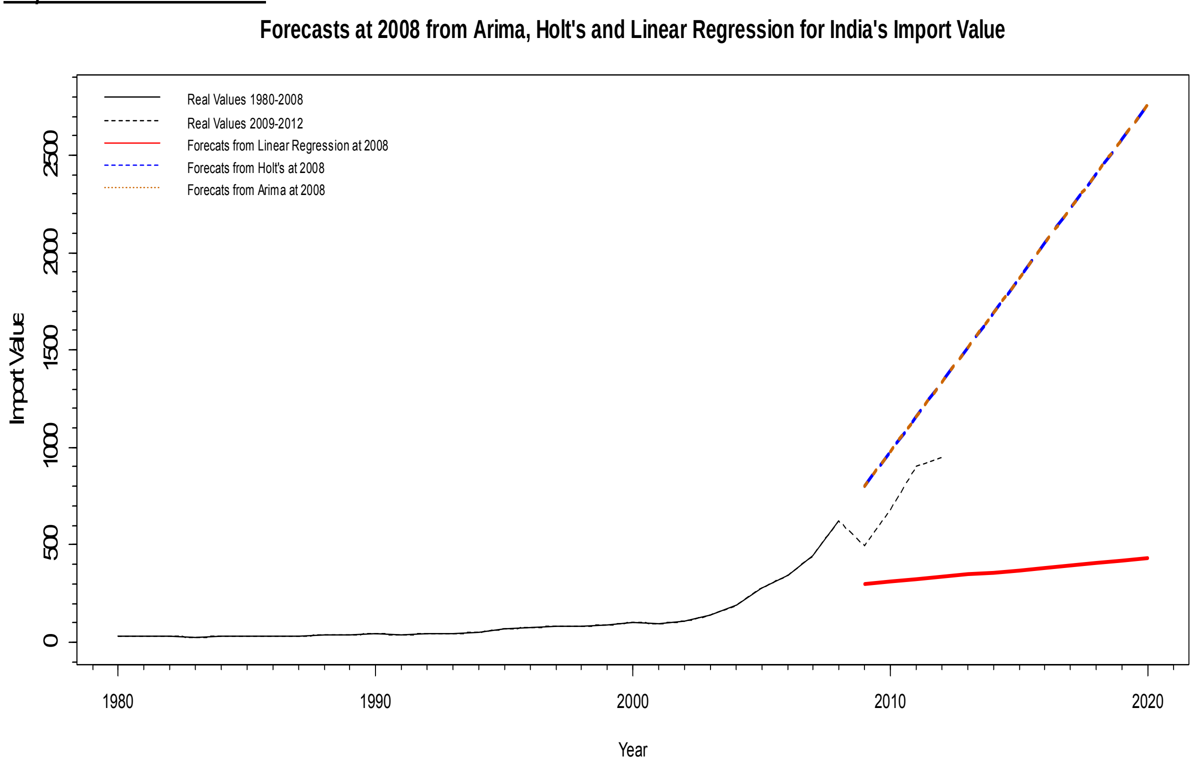
[Figure D.1.10] – Comparison of 2008 forecasts for USA and Export Value

Import Value – China



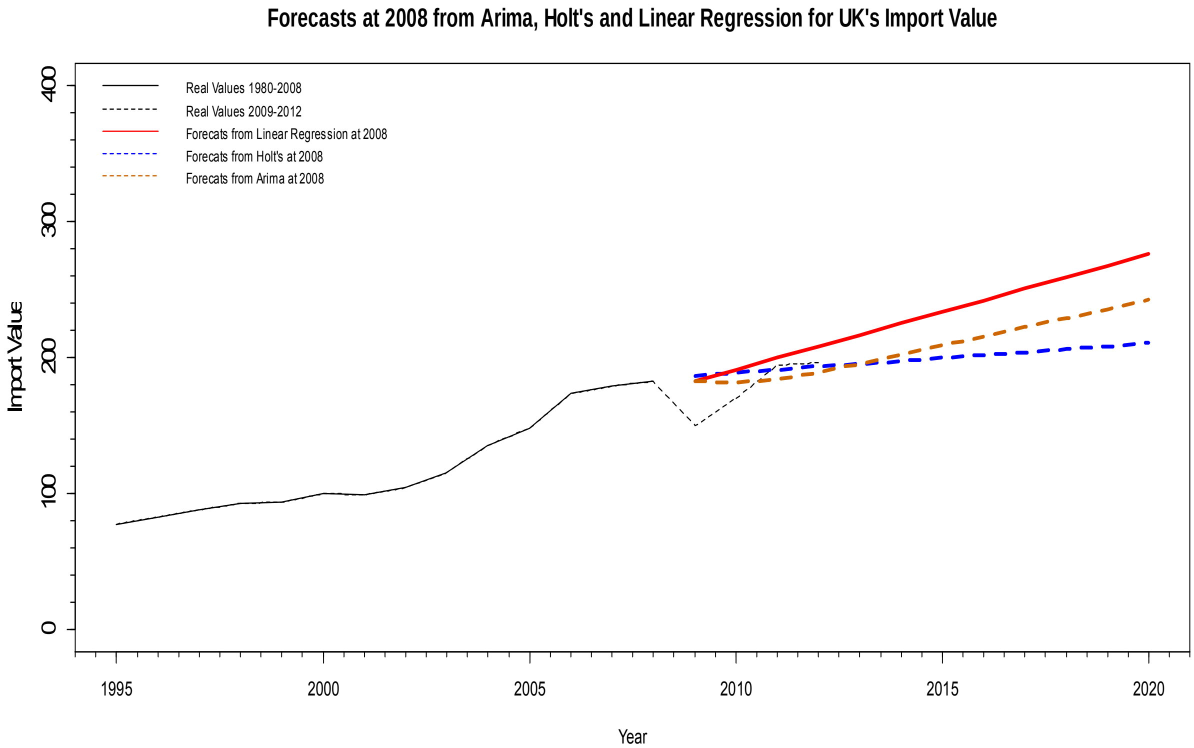
[Figure D.1.11] – Comparison of 2008 forecasts for China and Import Value

Import Value – India



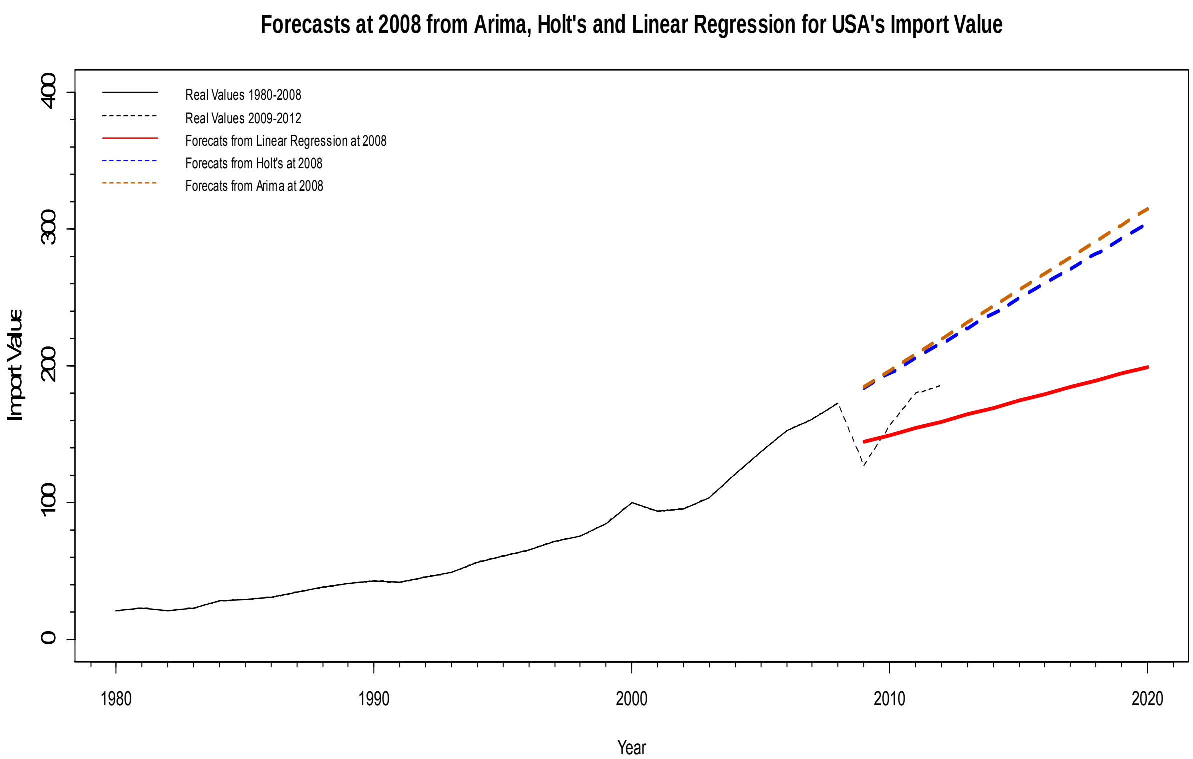
[Figure D.1.12] – Comparison of 2008 forecasts for India and Import Value

Import Value – UK



[Figure D.1.13] – Comparison of 2008 forecasts for UK and Import Value

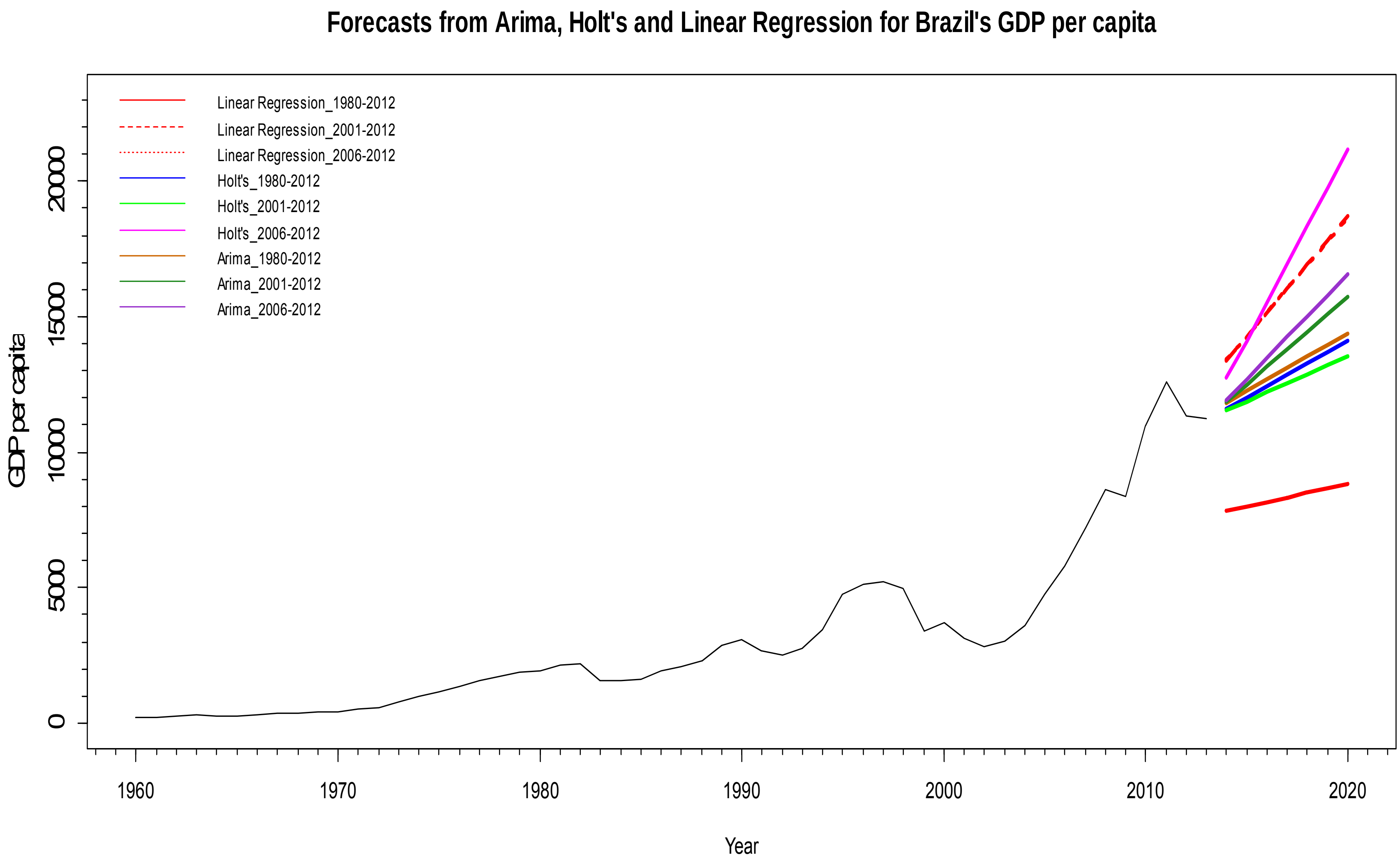
Import Value – USA



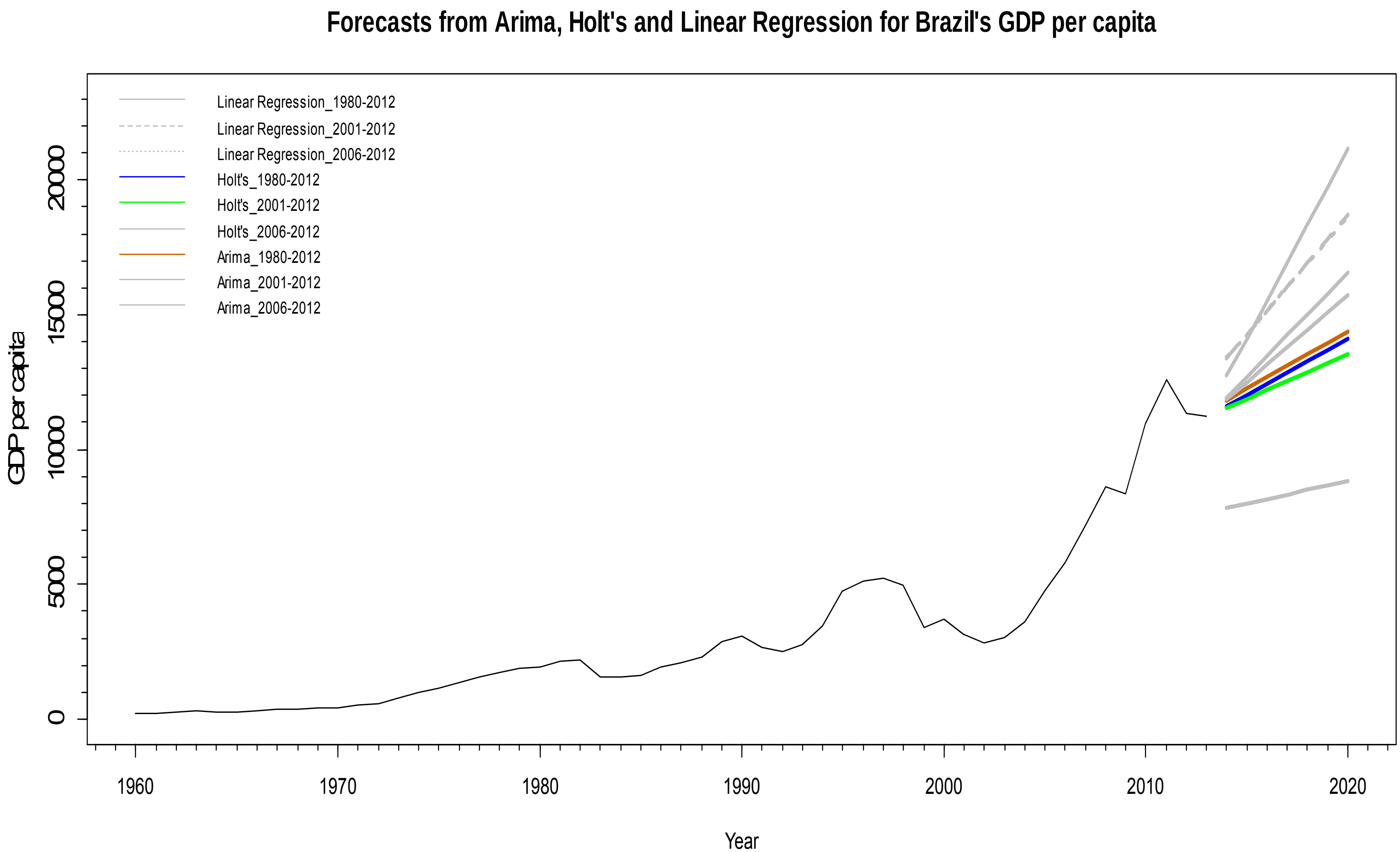
[Figure D.1.14] – Comparison of 2008 forecasts for USA and Import Value

D.2. Graphs comparing all predictions for each country

GDP per capita – Brazil

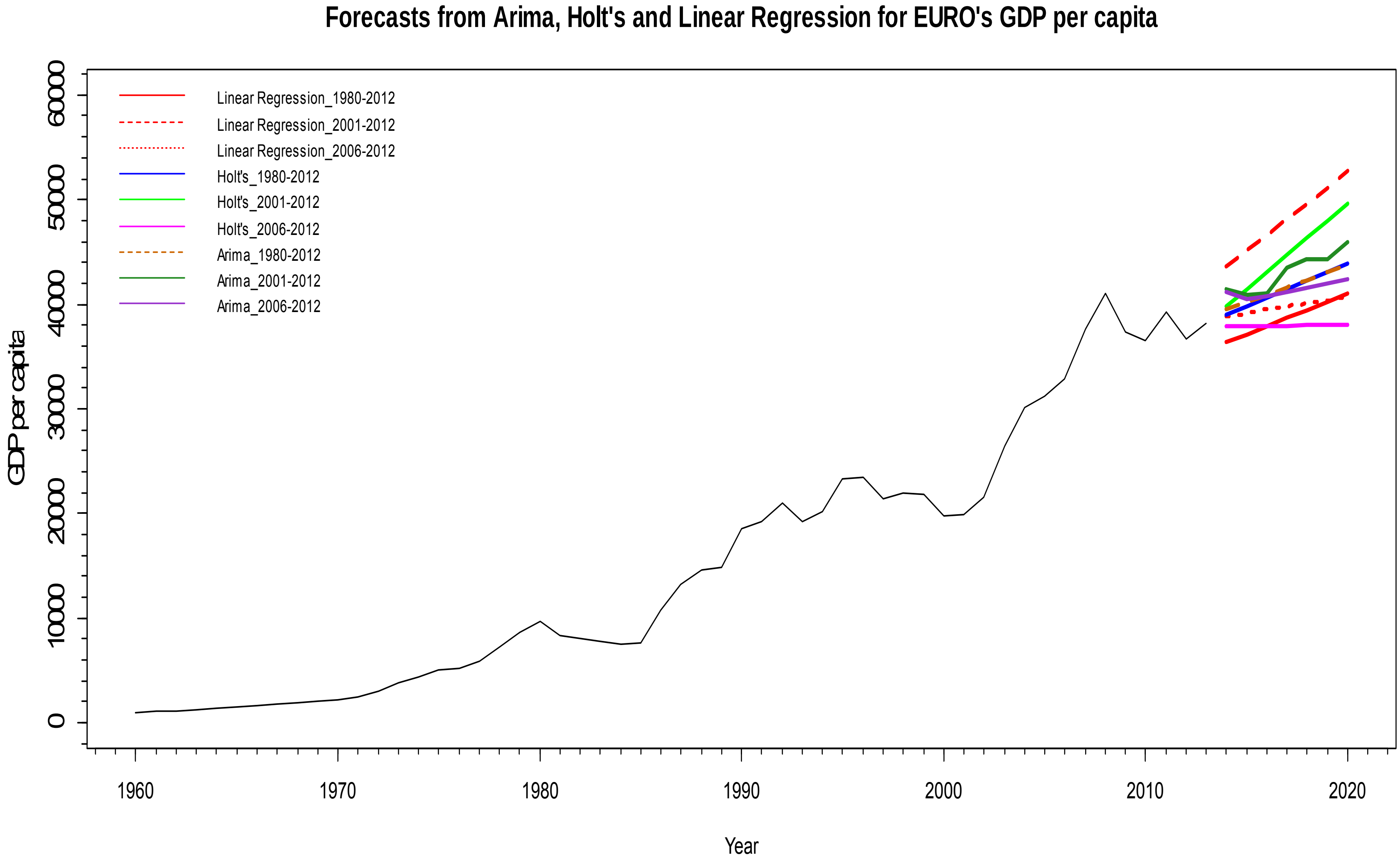


[Figure D.2.1] – Comparison of all predictions for Brazil's GDP per capita

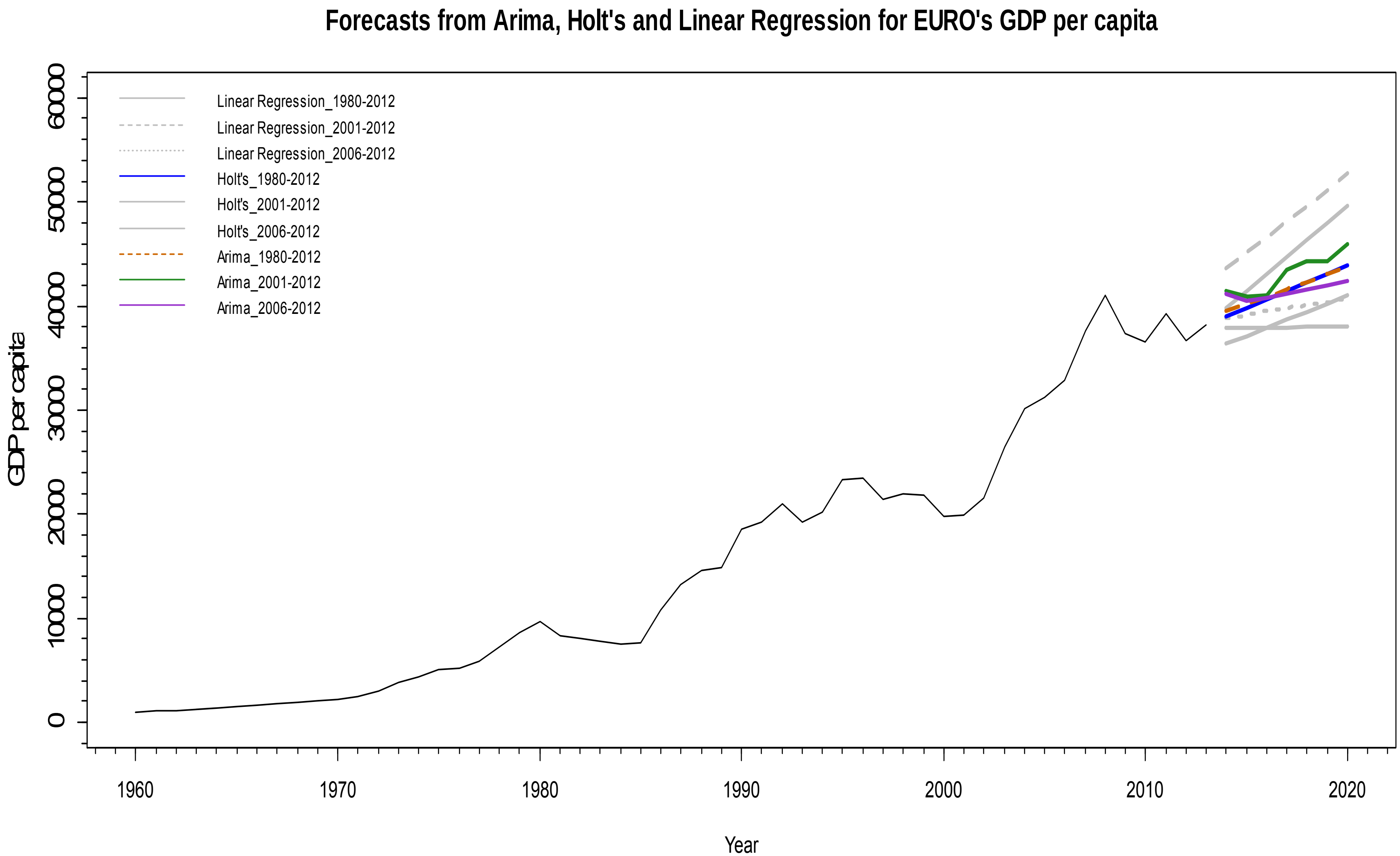


[Figure D.2.2] – Selection of the best predictions for Brazil's GDP per capita

GDP per capita – EURO zone

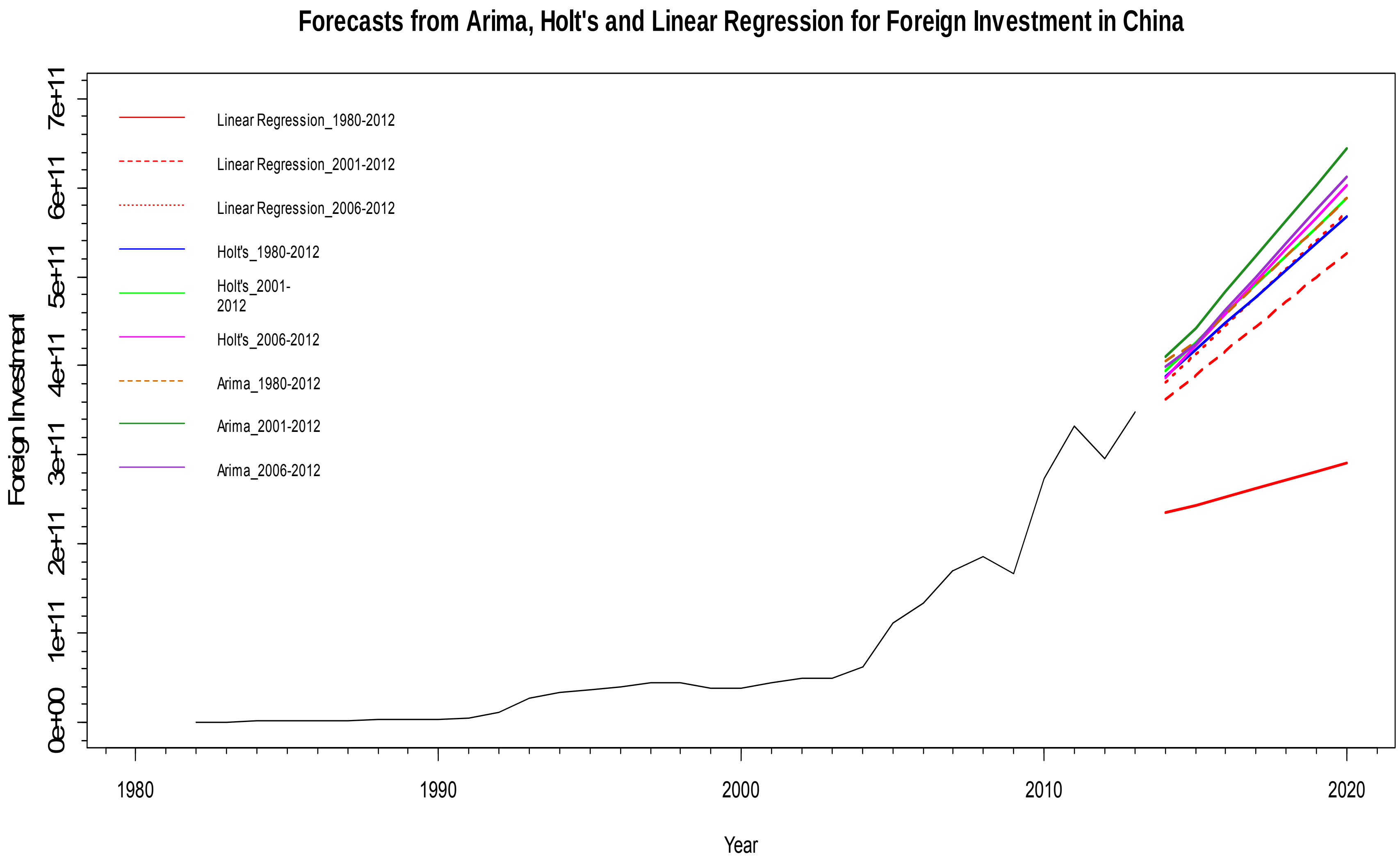


[Figure D.2.3] – Comparison of all predictions for EURO zone's GDP per capita

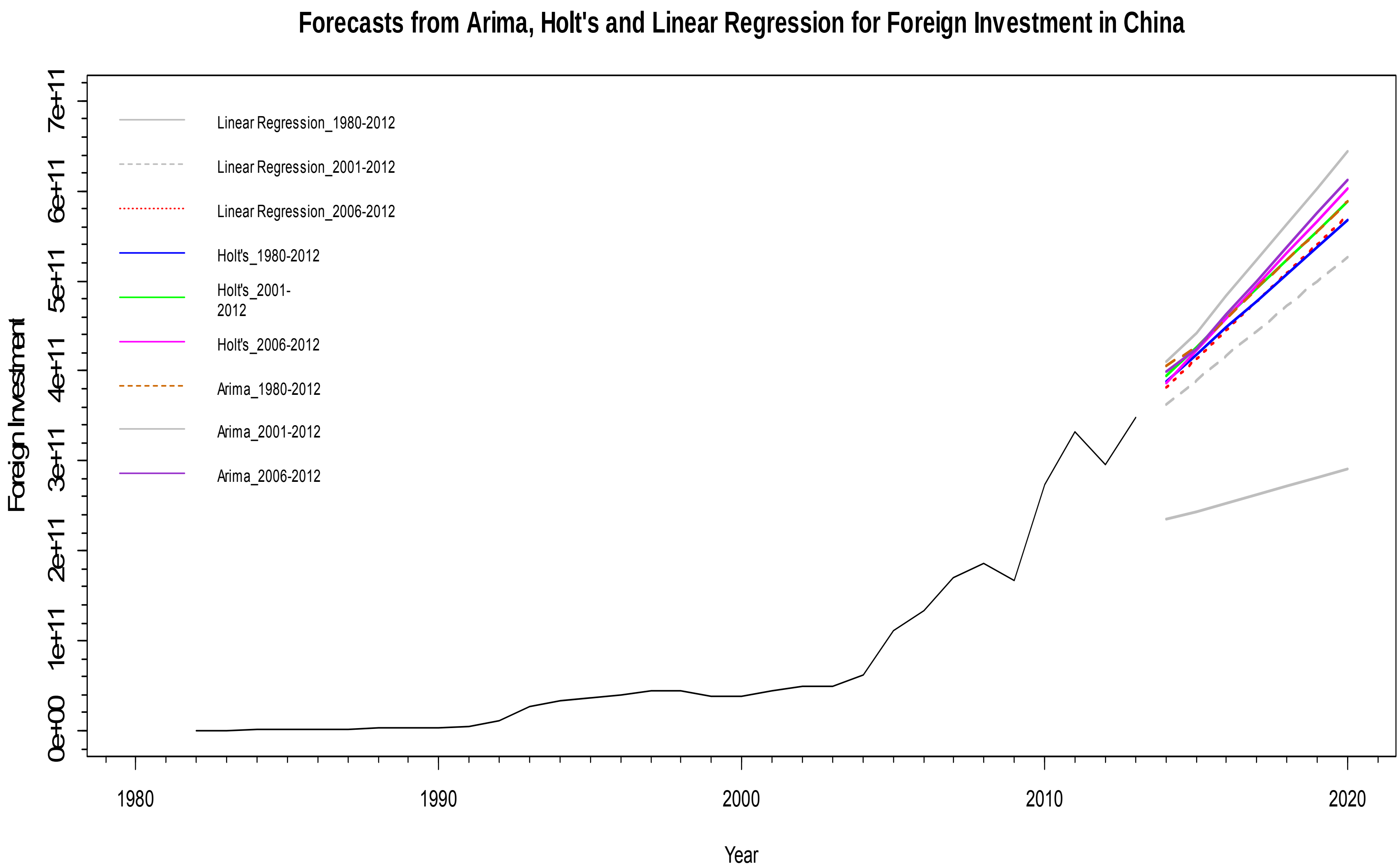


[Figure D.2.4] – Selection of the best predictions for EURO zone's GDP per capita

Foreign Investment – China

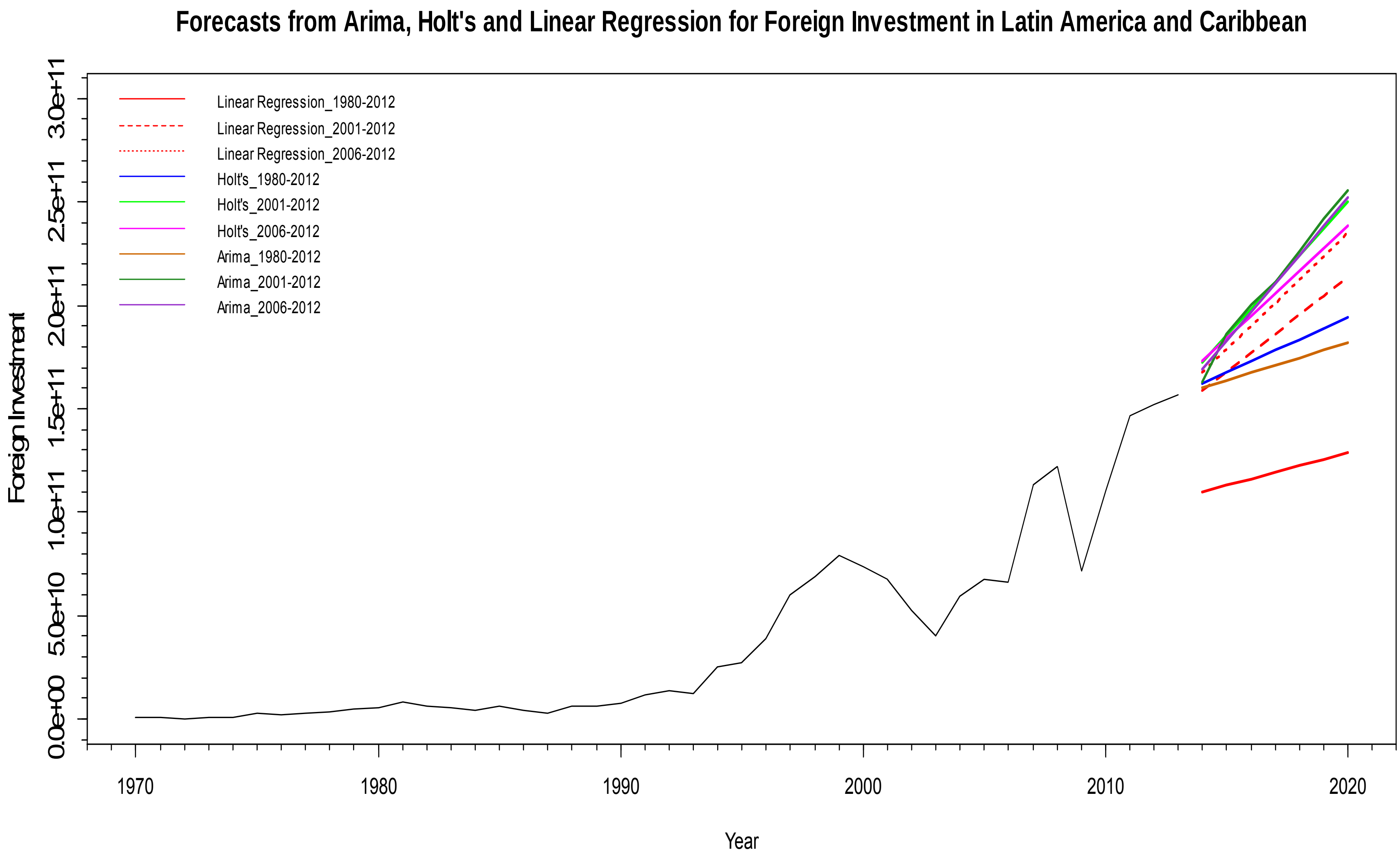


[Figure D.2.5] – Comparison of all predictions for Foreign Investment in China

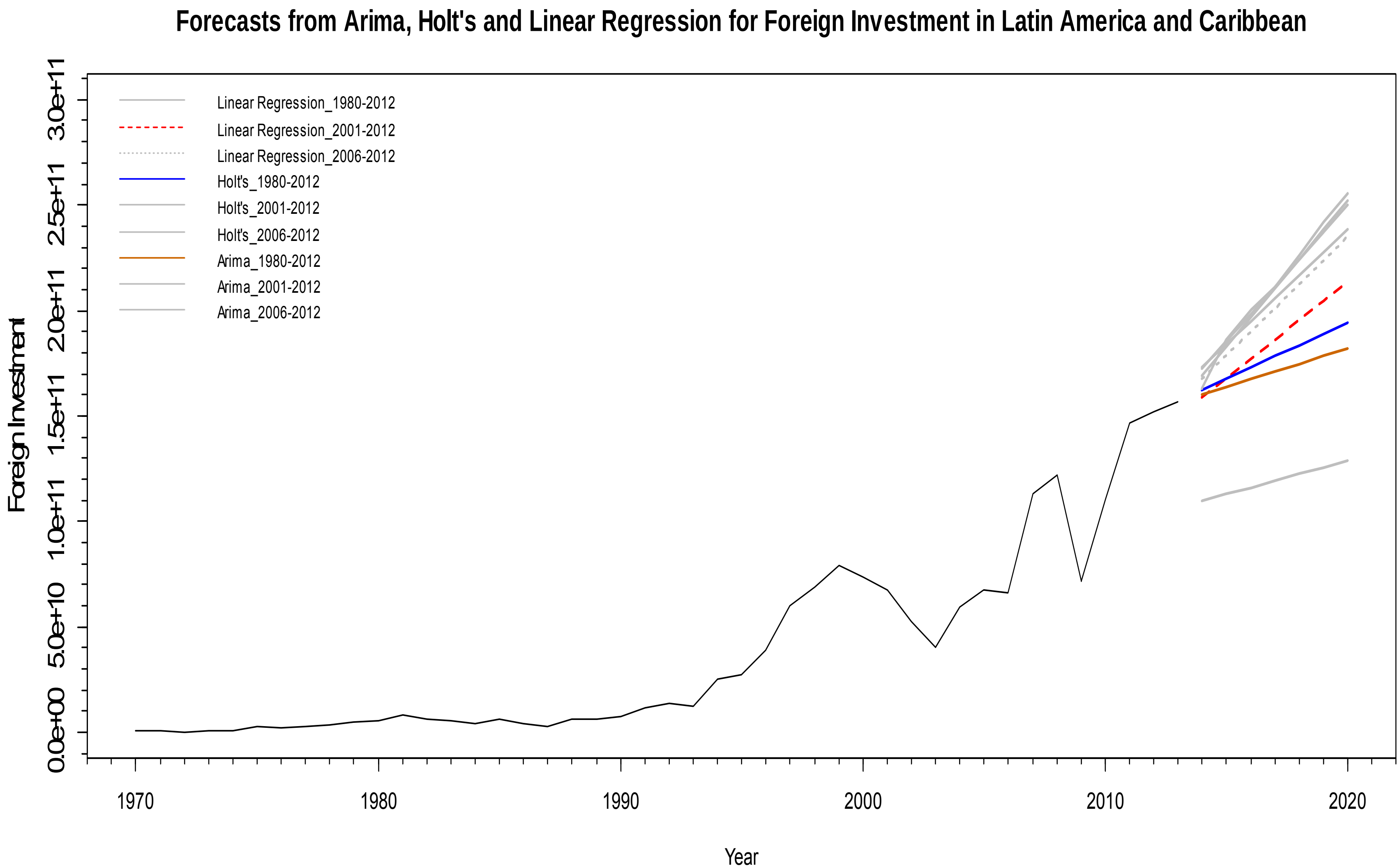


[Figure D.2.6] – Selection of the best predictions for Foreign Investment in China

Foreign Investment – Latin America and Caribbean

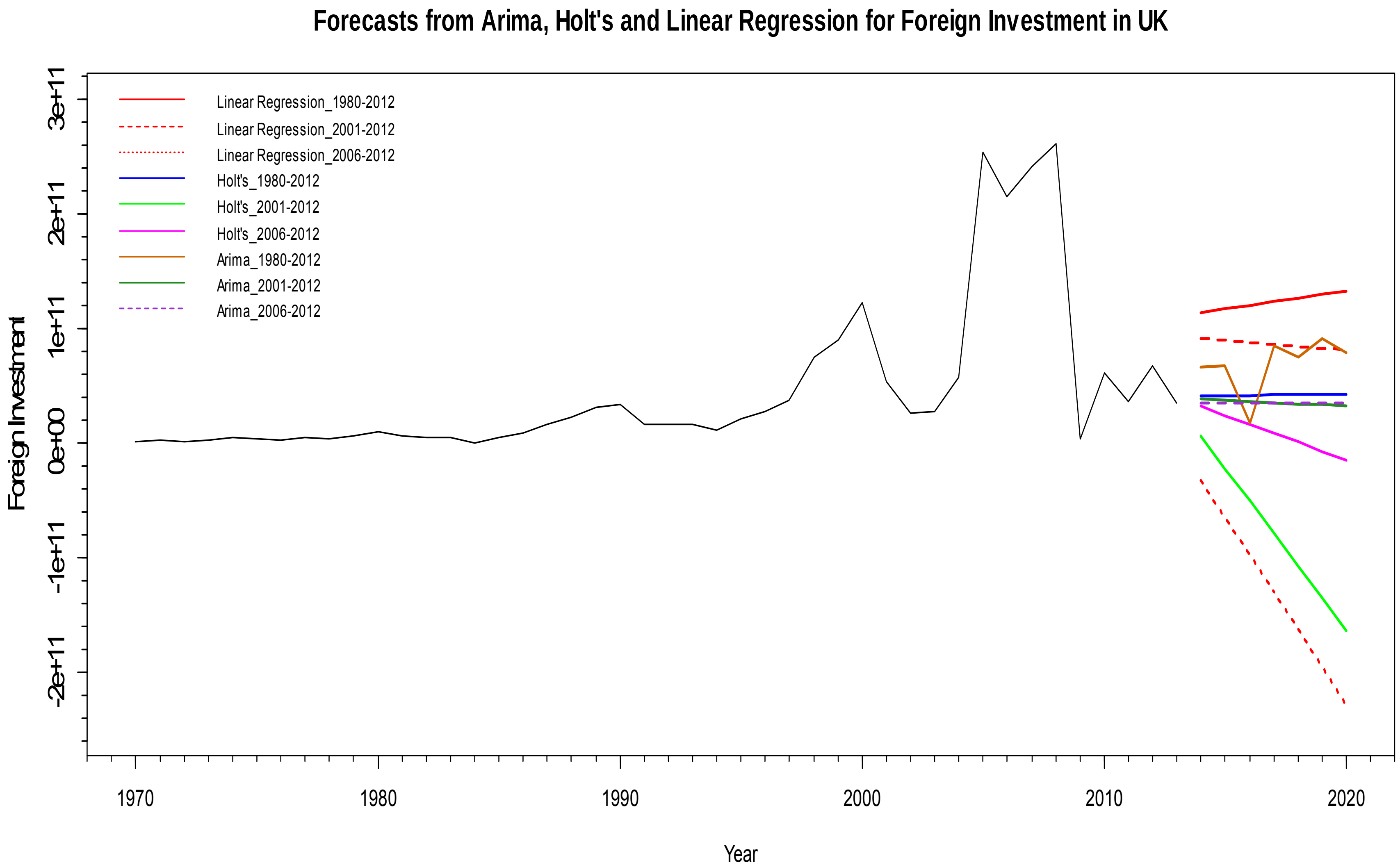


[Figure D.2.7] – Comparison of all predictions for Foreign Investment in Latin America and Caribbean

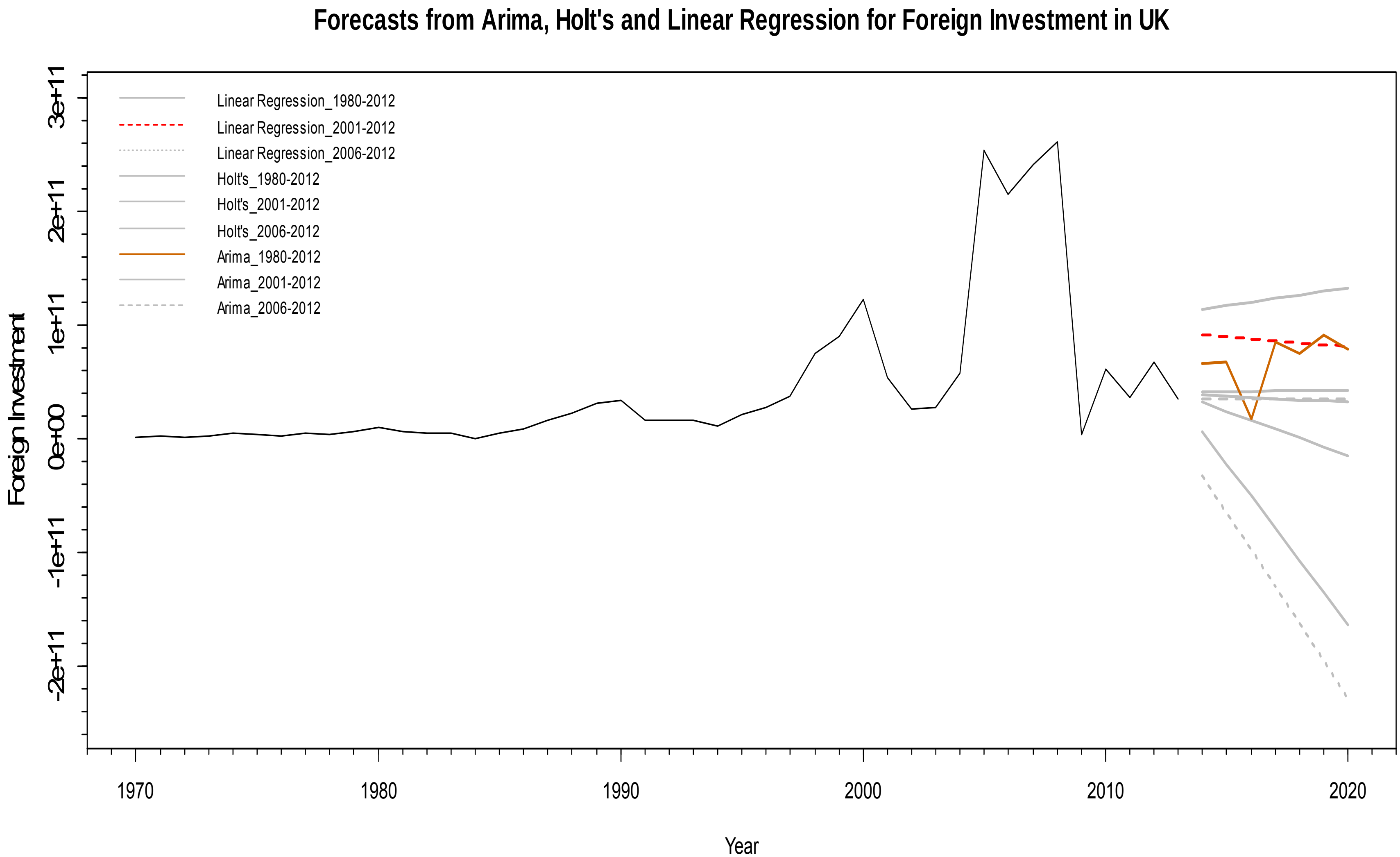


[Figure D.2.8] – Selection of the best predictions for Foreign Investment in Latin America and Caribbean

Foreign Investment – UK

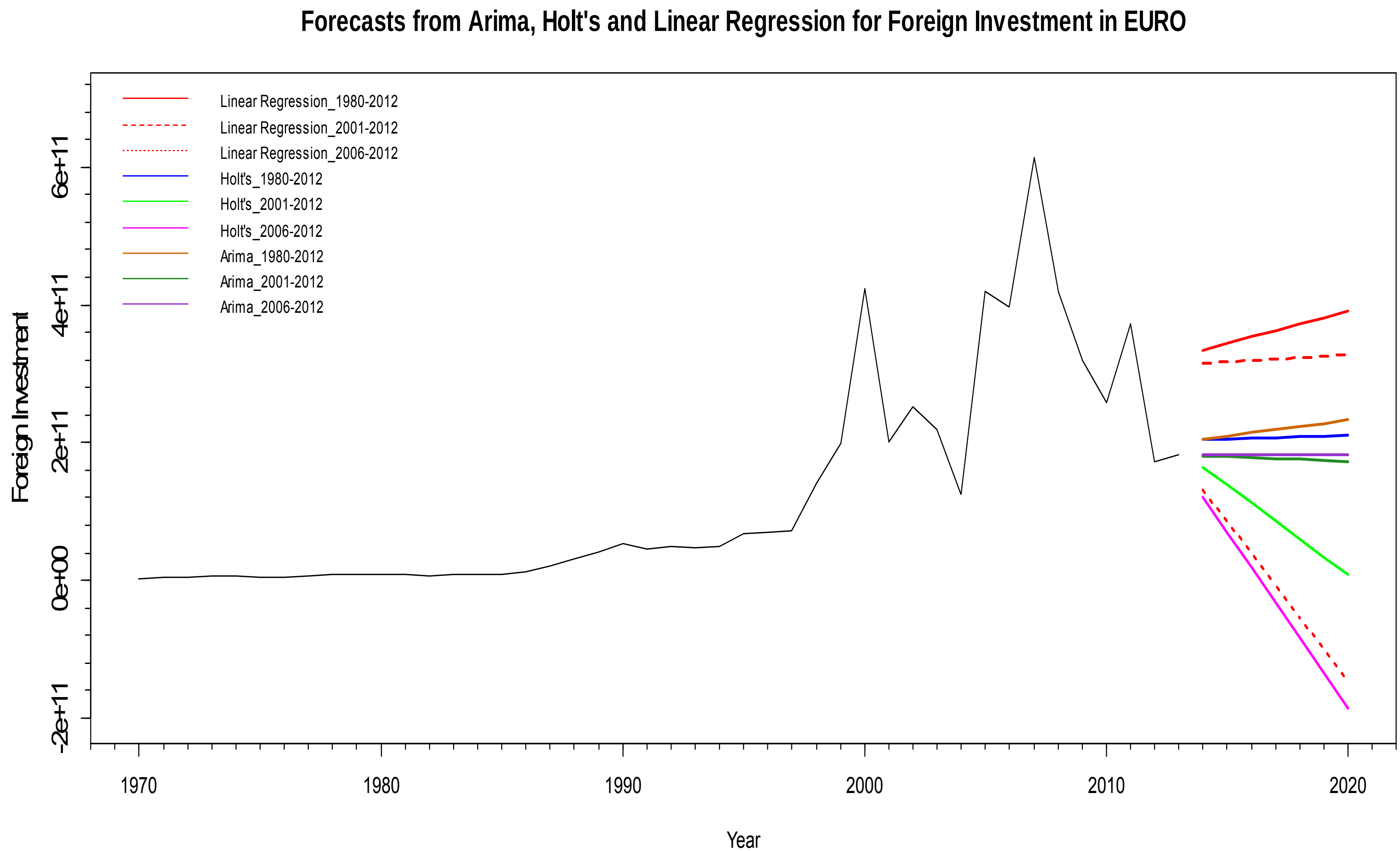


[Figure D.2.9] – Comparison of all predictions for Foreign Investment in UK

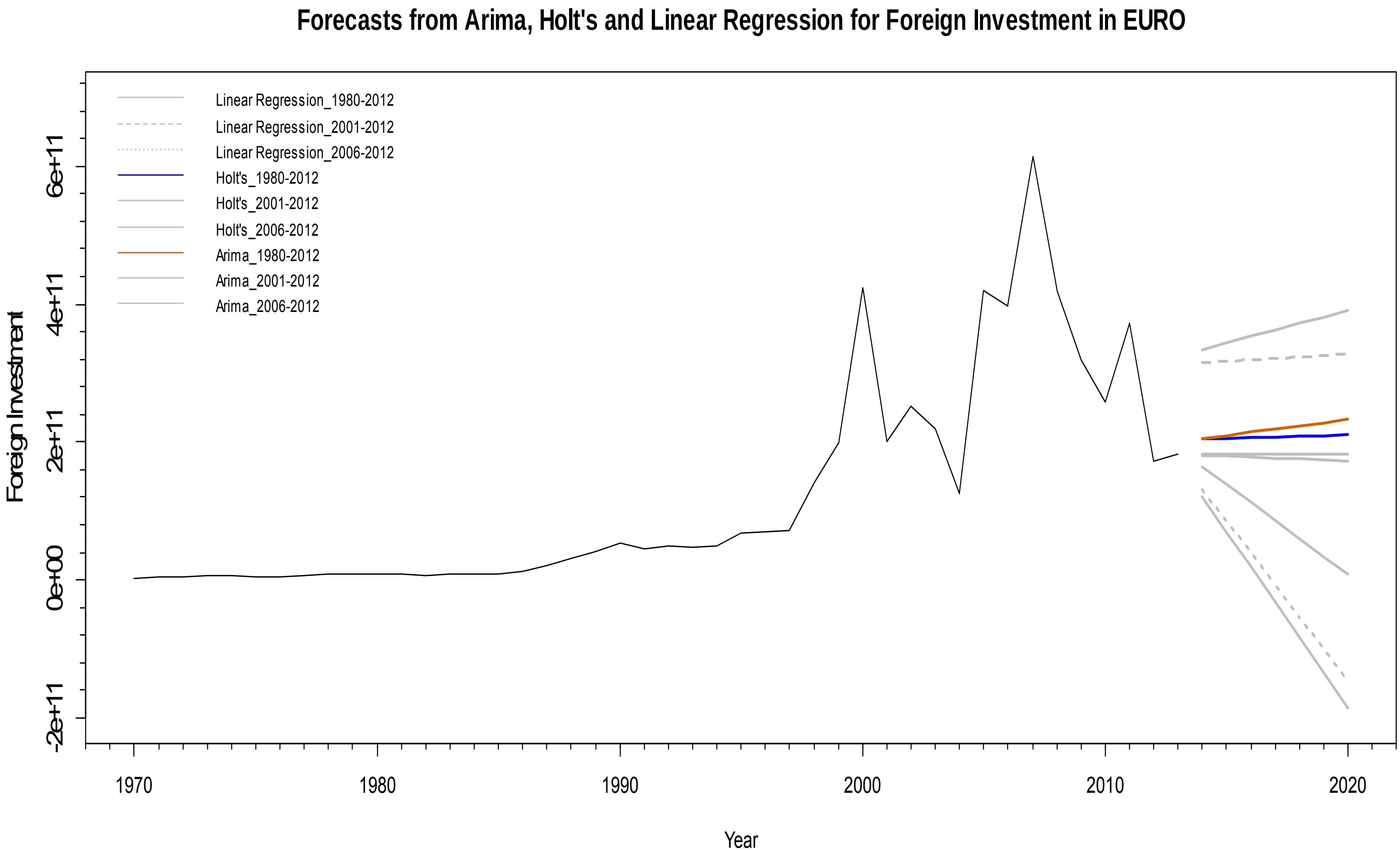


[Figure D.2.10] – Selection of the best predictions for Foreign Investment in UK

Foreign Investment – EURO

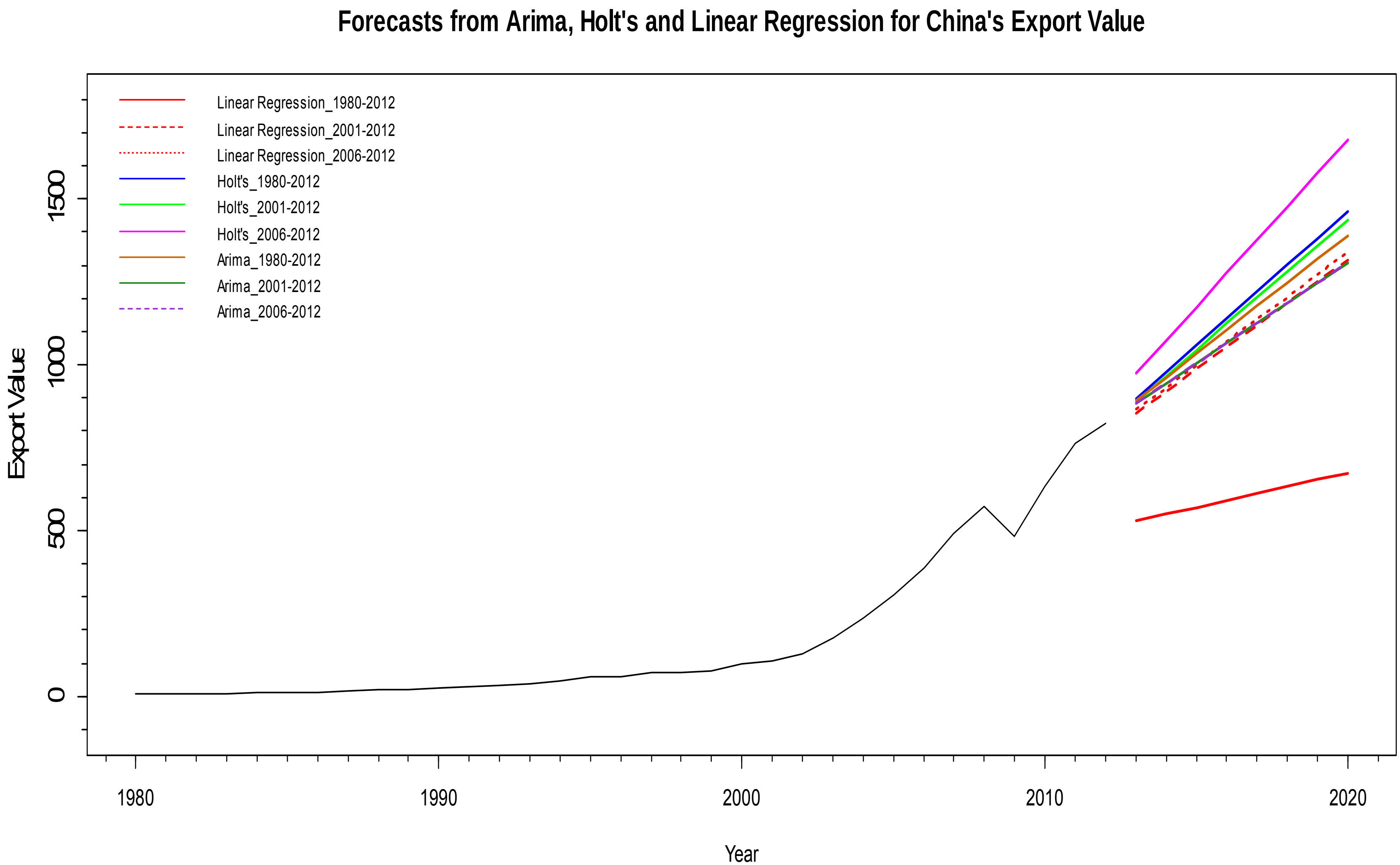


[Figure D.2.11] – Comparison of all predictions for Foreign Investment in EURO zone

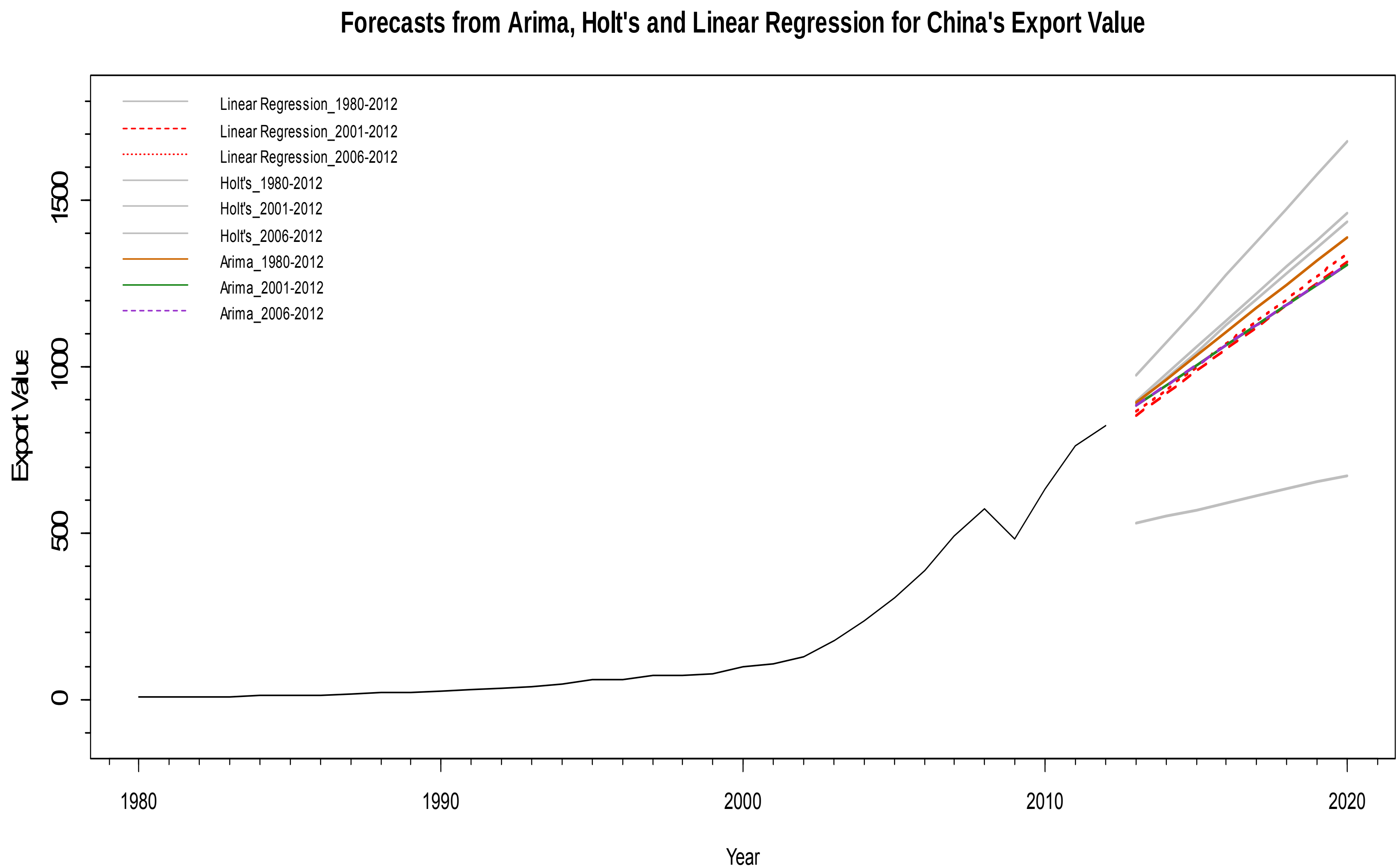


[Figure D.2.12] – Selection of the best predictions for Foreign Investment in EURO zone

Export Value – China

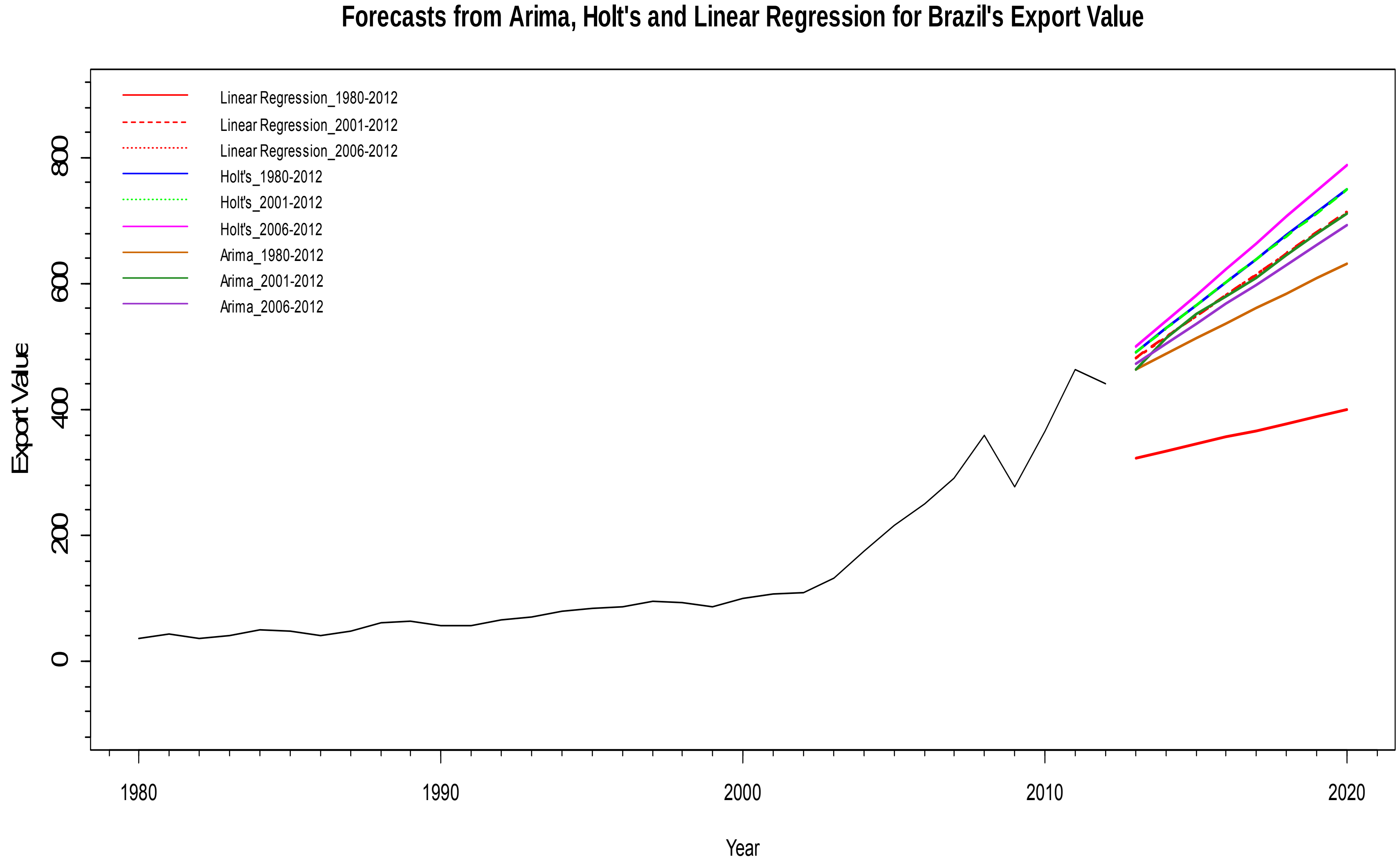


[Figure D.2.13] – Comparison of all predictions for China's Export Value

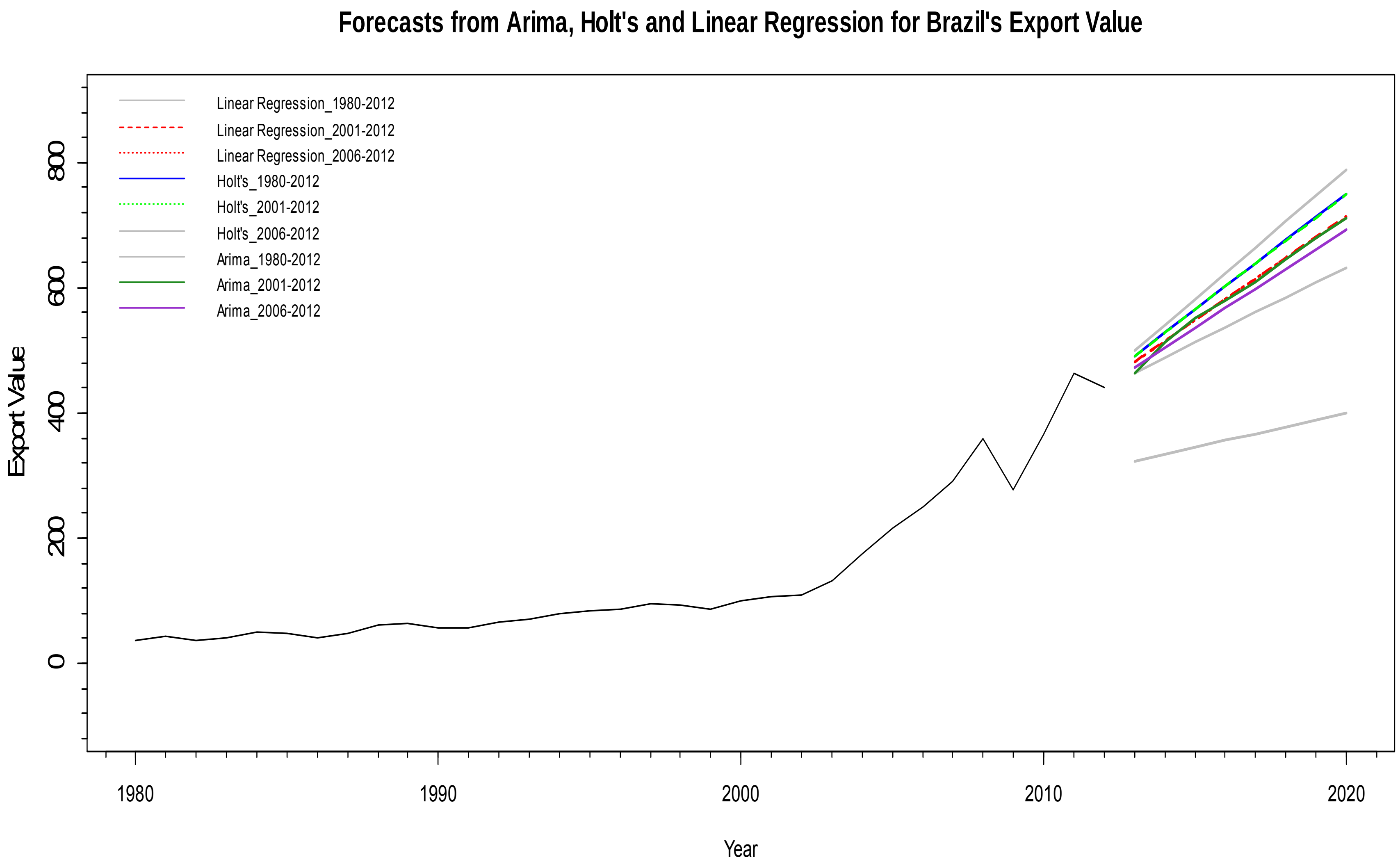


[Figure D.2.14] – Selection of the best predictions for China's Export Value

Export Value – Brazil

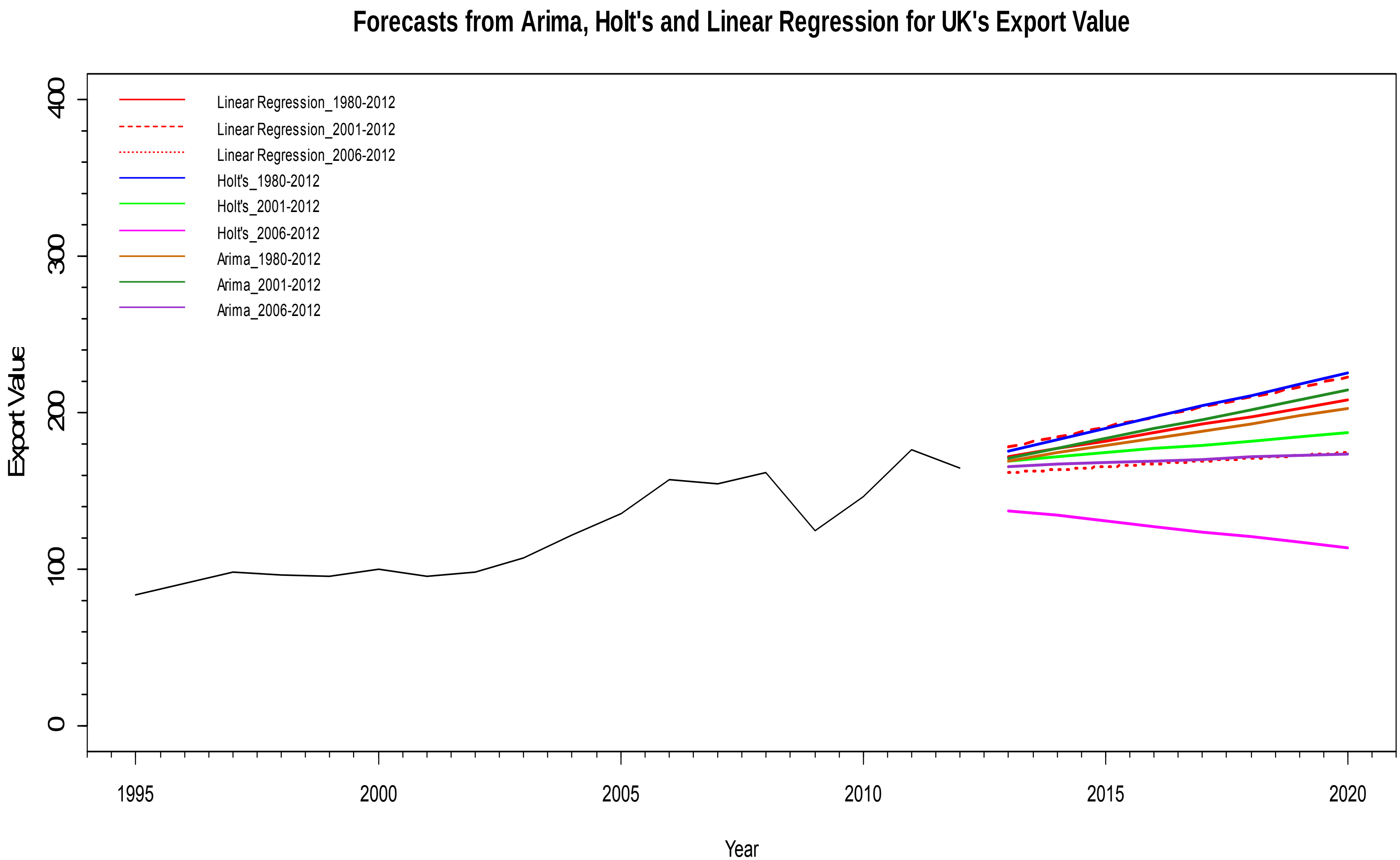


[Figure D.2.15] – Comparison of all predictions for Brazil's Export Value

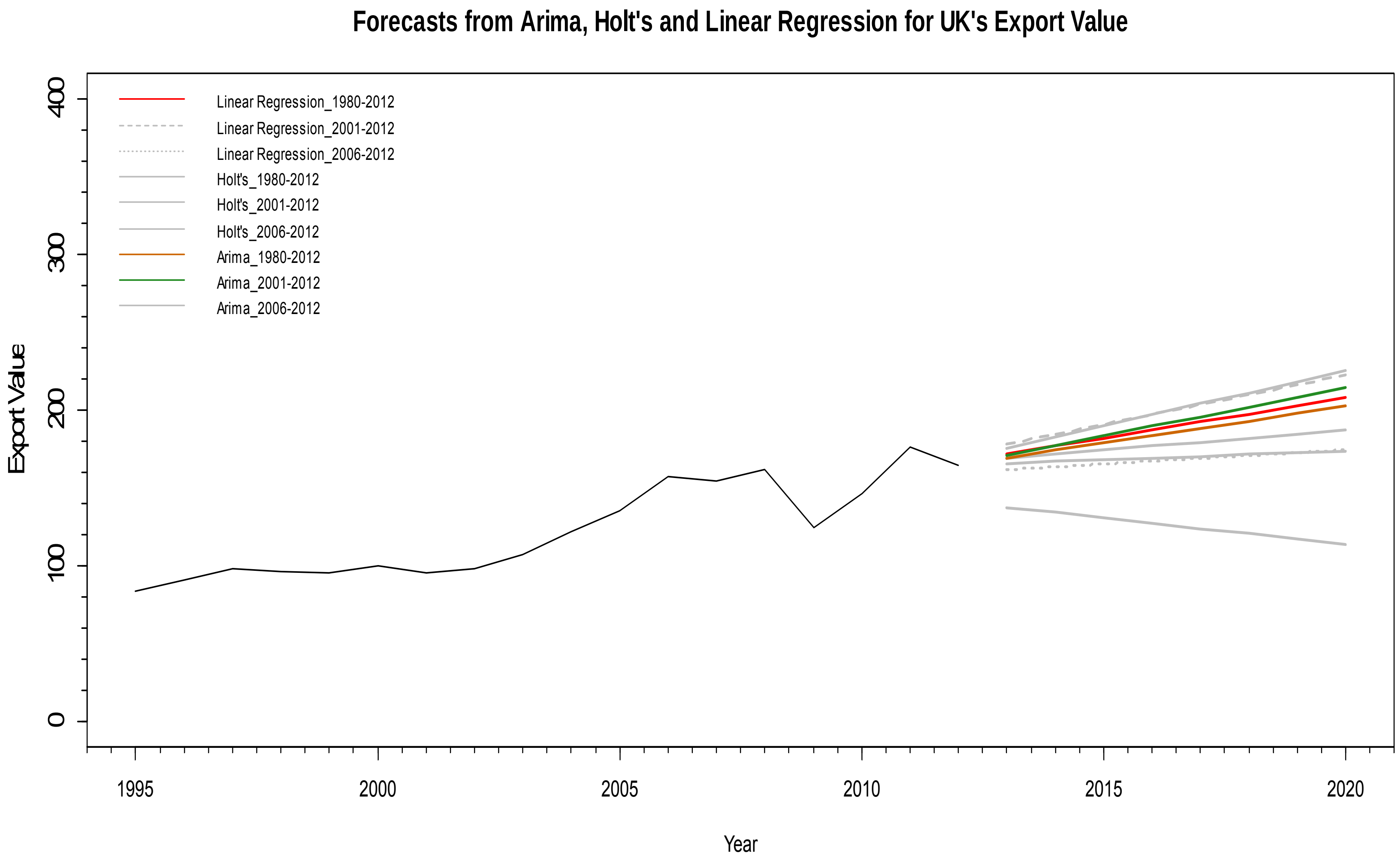


[Figure D.2.16] – Selection of the best predictions for Brazil's Export Value

Export Value – UK

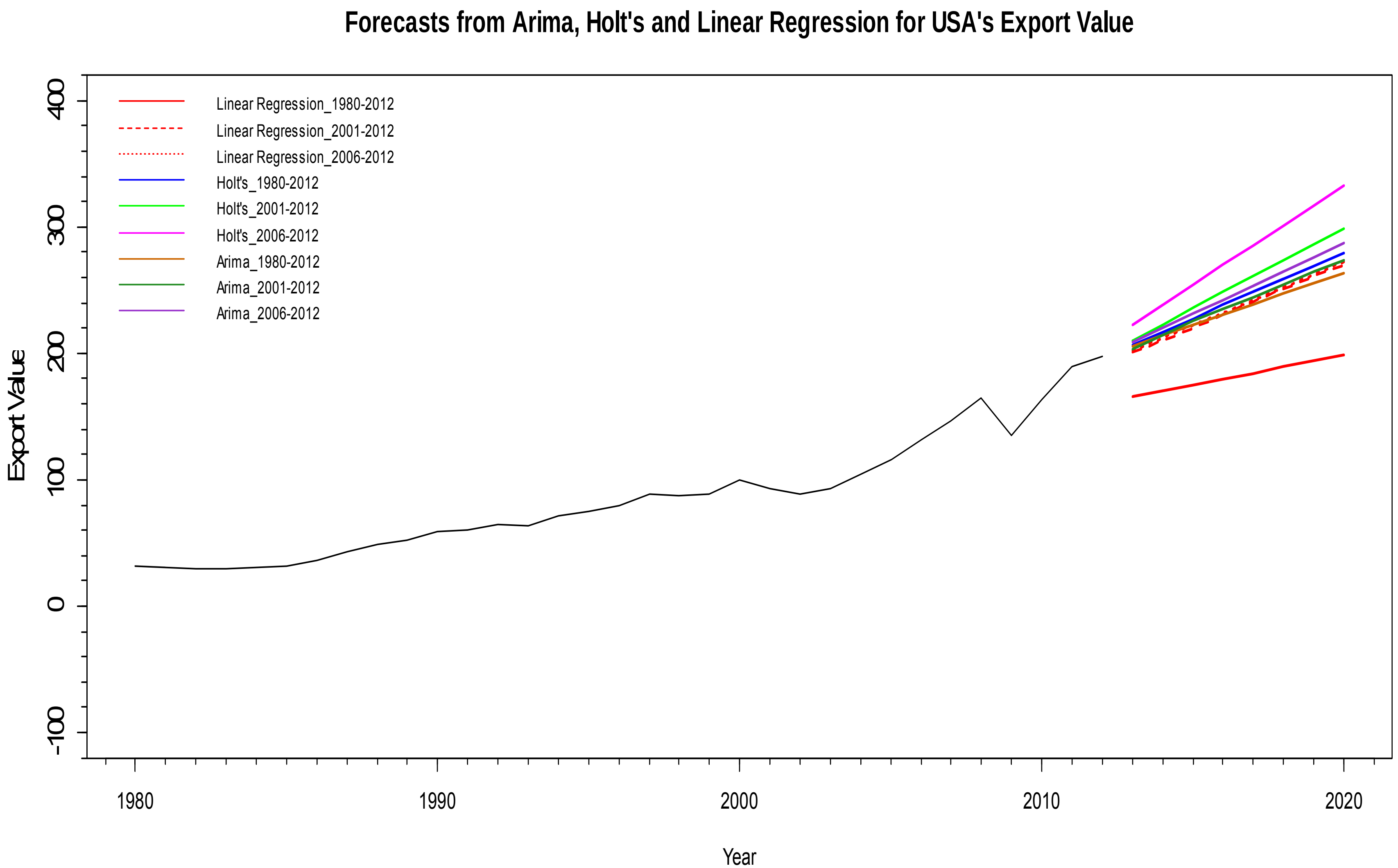


[Figure D.2.17] – Comparison of all predictions for UK's Export Value

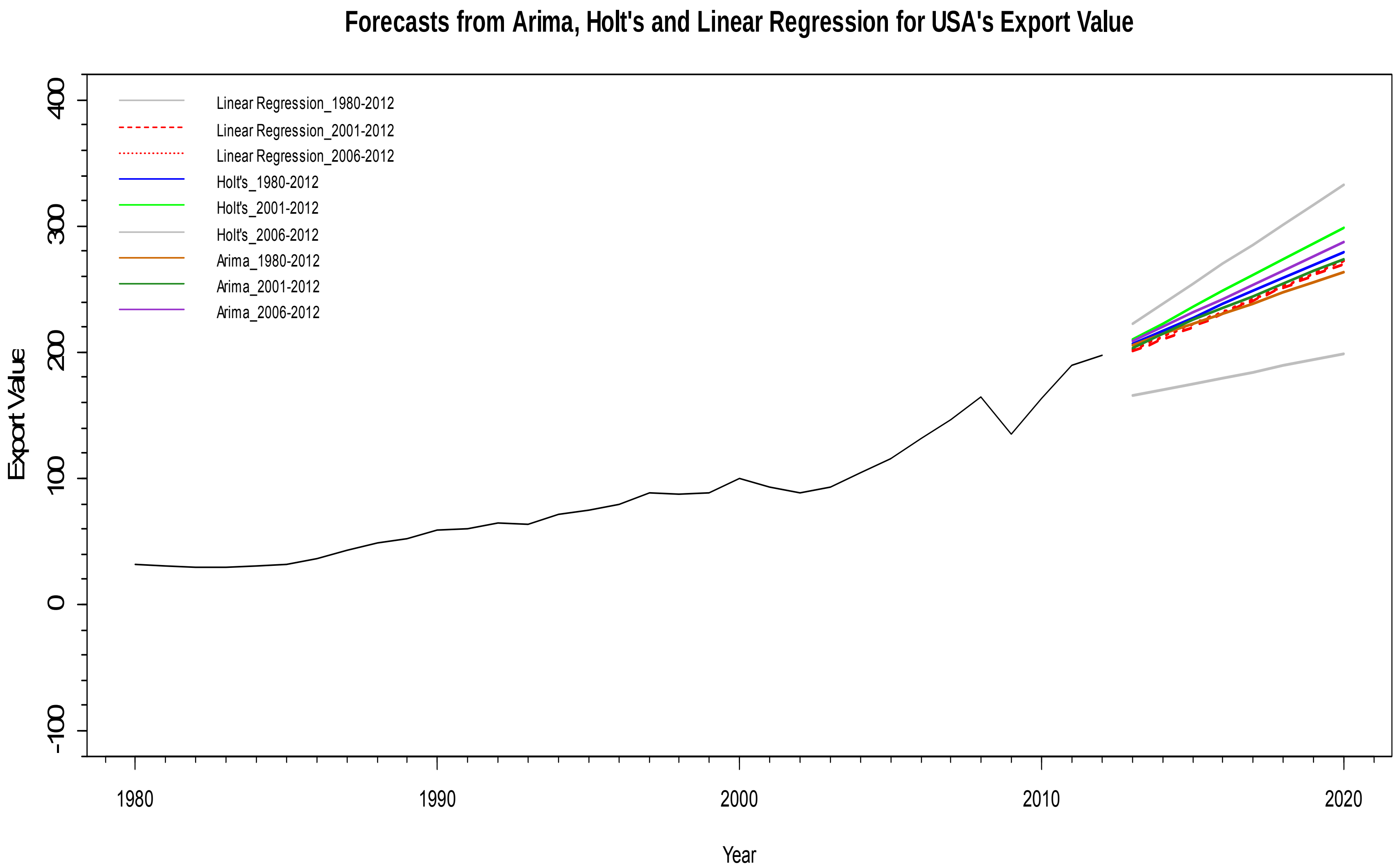


[Figure D.2.18] – Selection of the best predictions for UK's Export Value

Export Value – USA

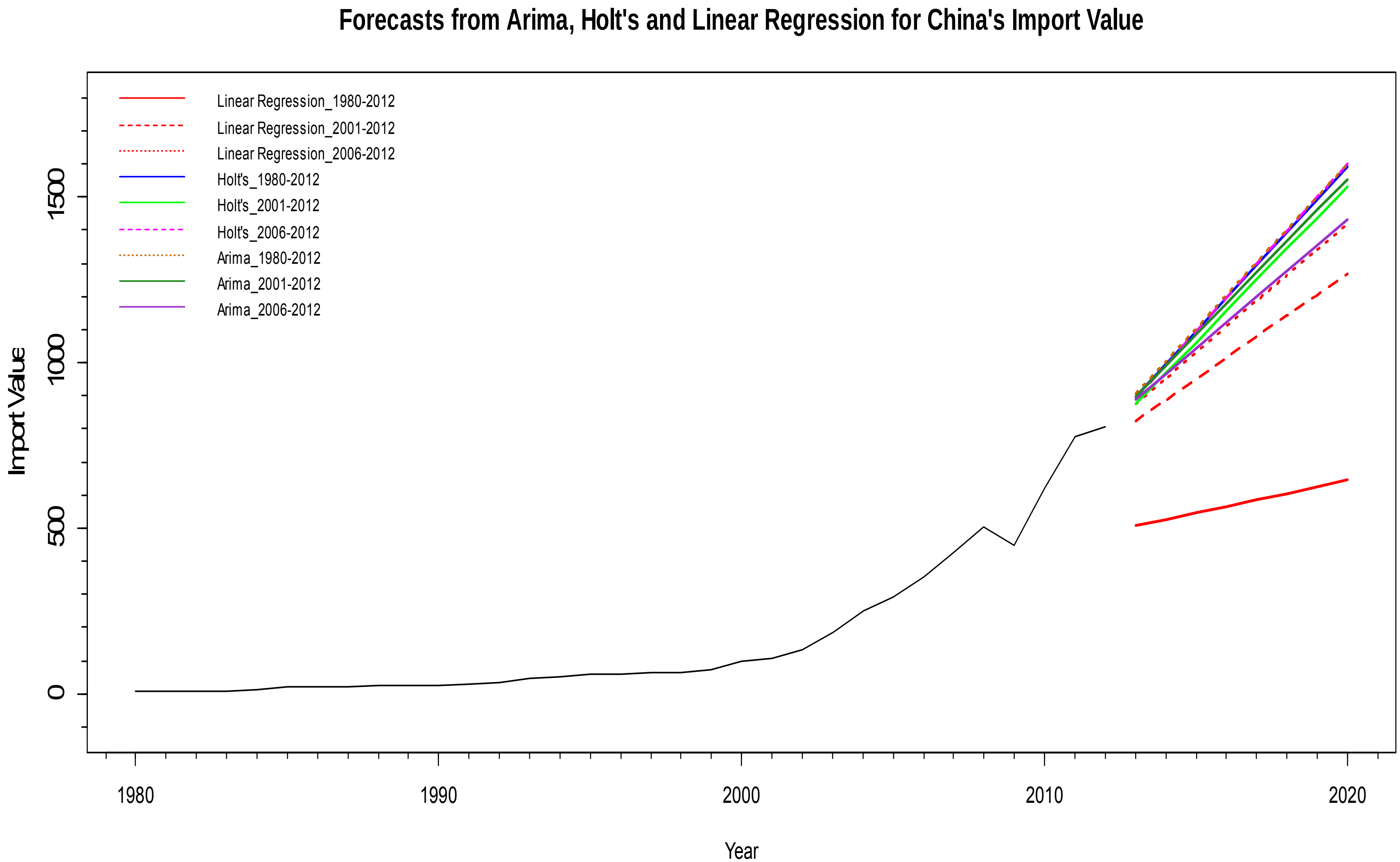


[Figure D.2.19] – Comparison of all predictions for USA's Export Value

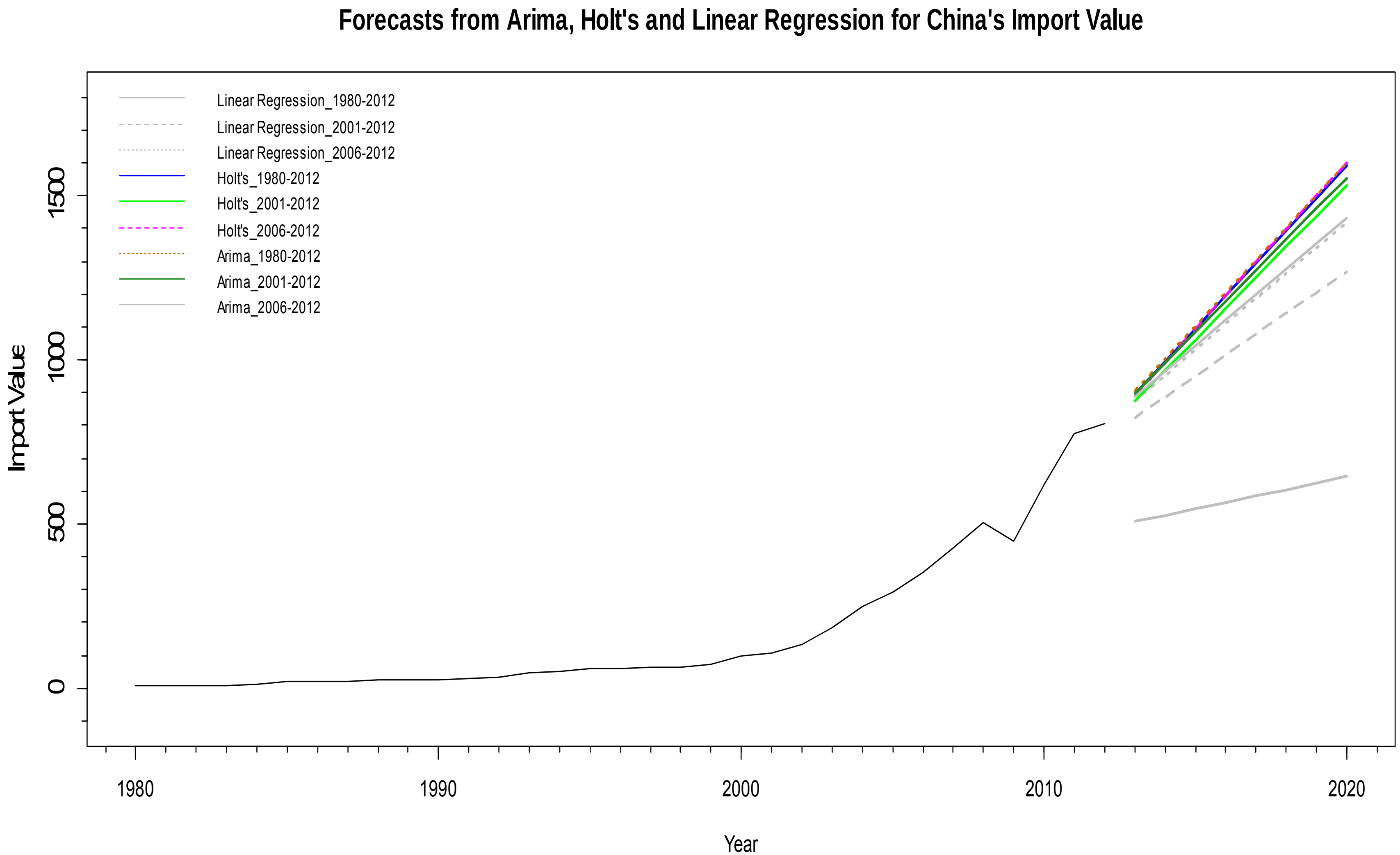


[Figure D.2.20] – Selection of the best predictions for USA's Export Value

Import Value – China

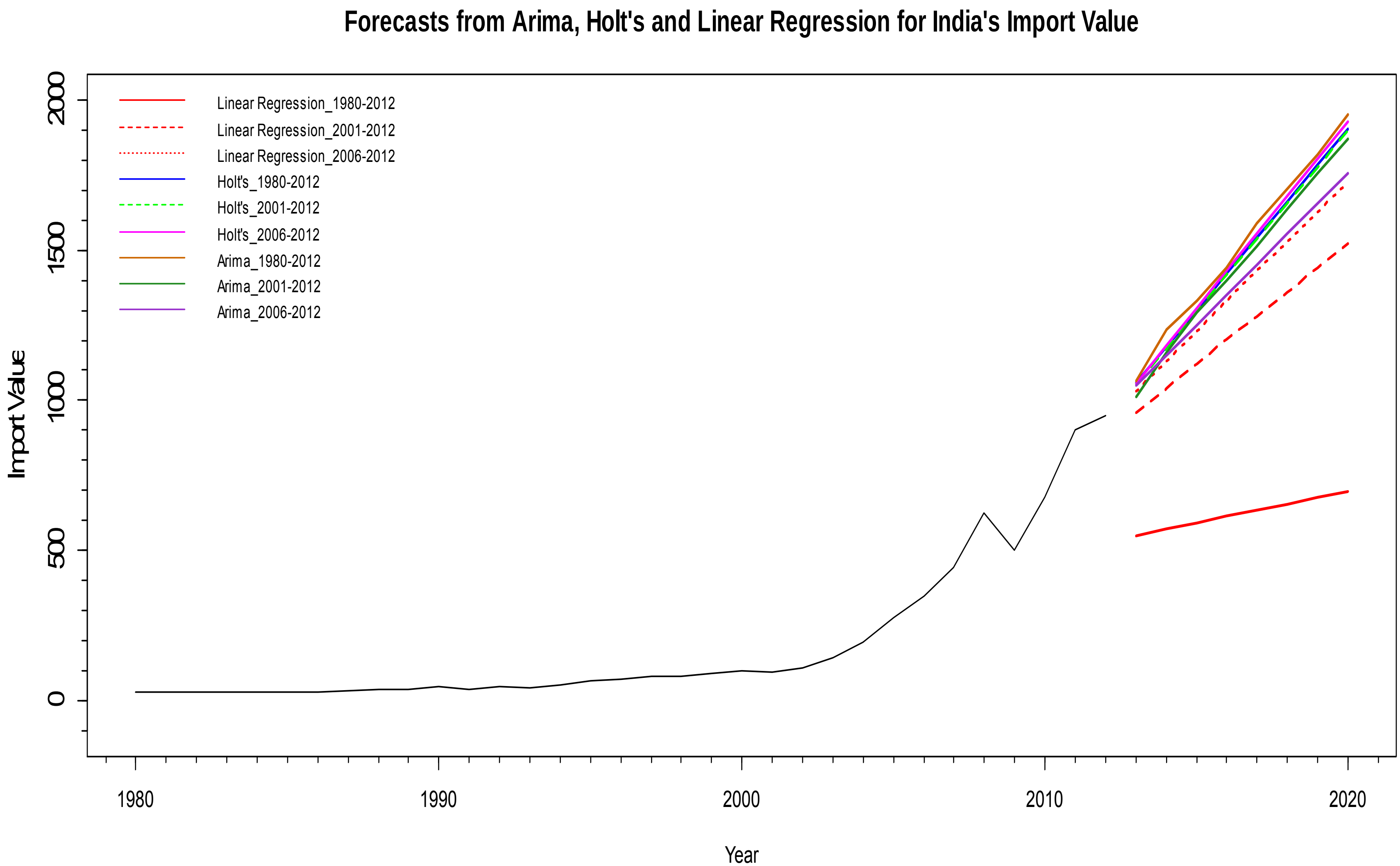


[Figure D.2.21] – Comparison of all predictions for China's Import Value

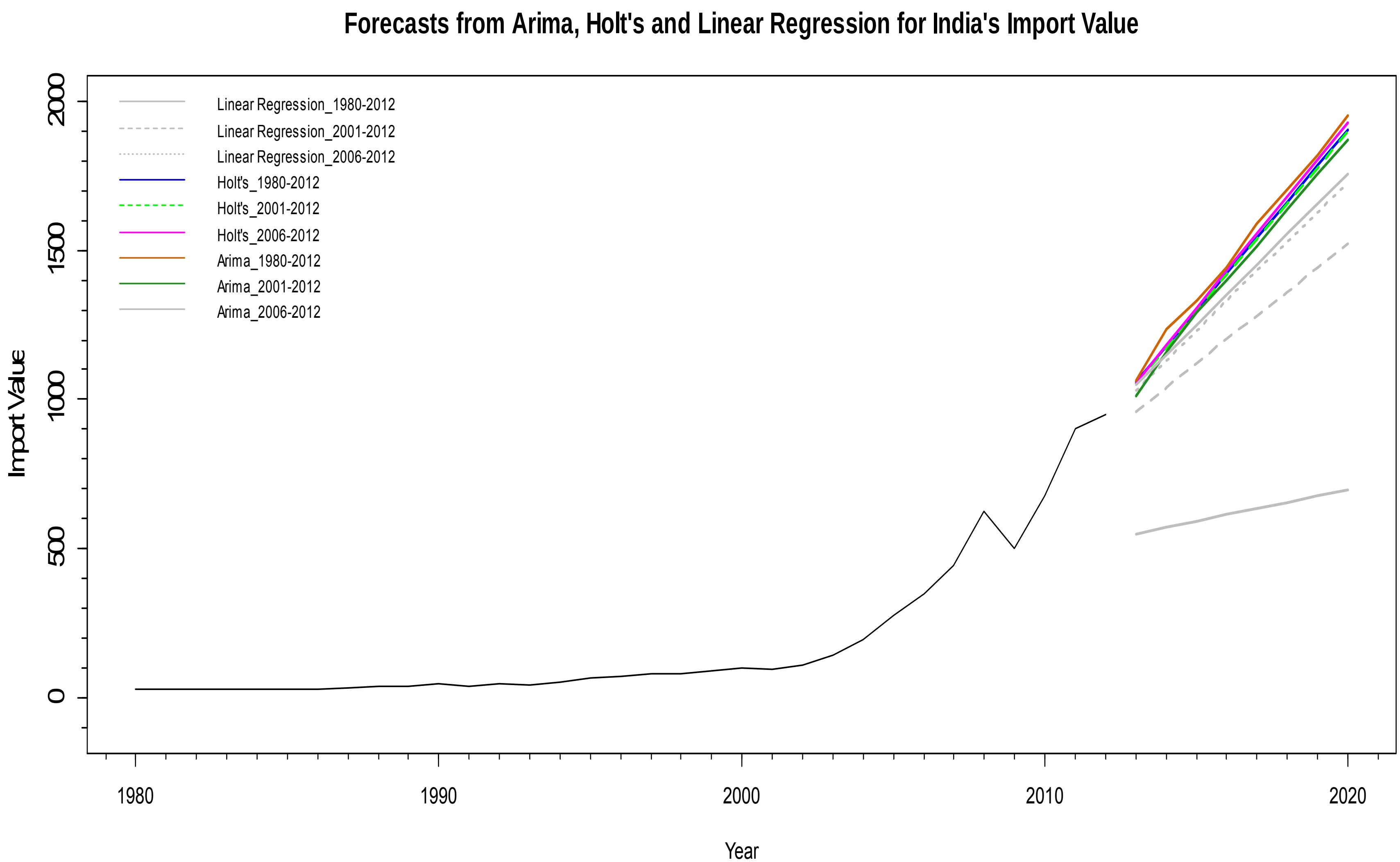


[Figure D.2.22] – Selection of the best predictions for China's Import Value

Import Value – India

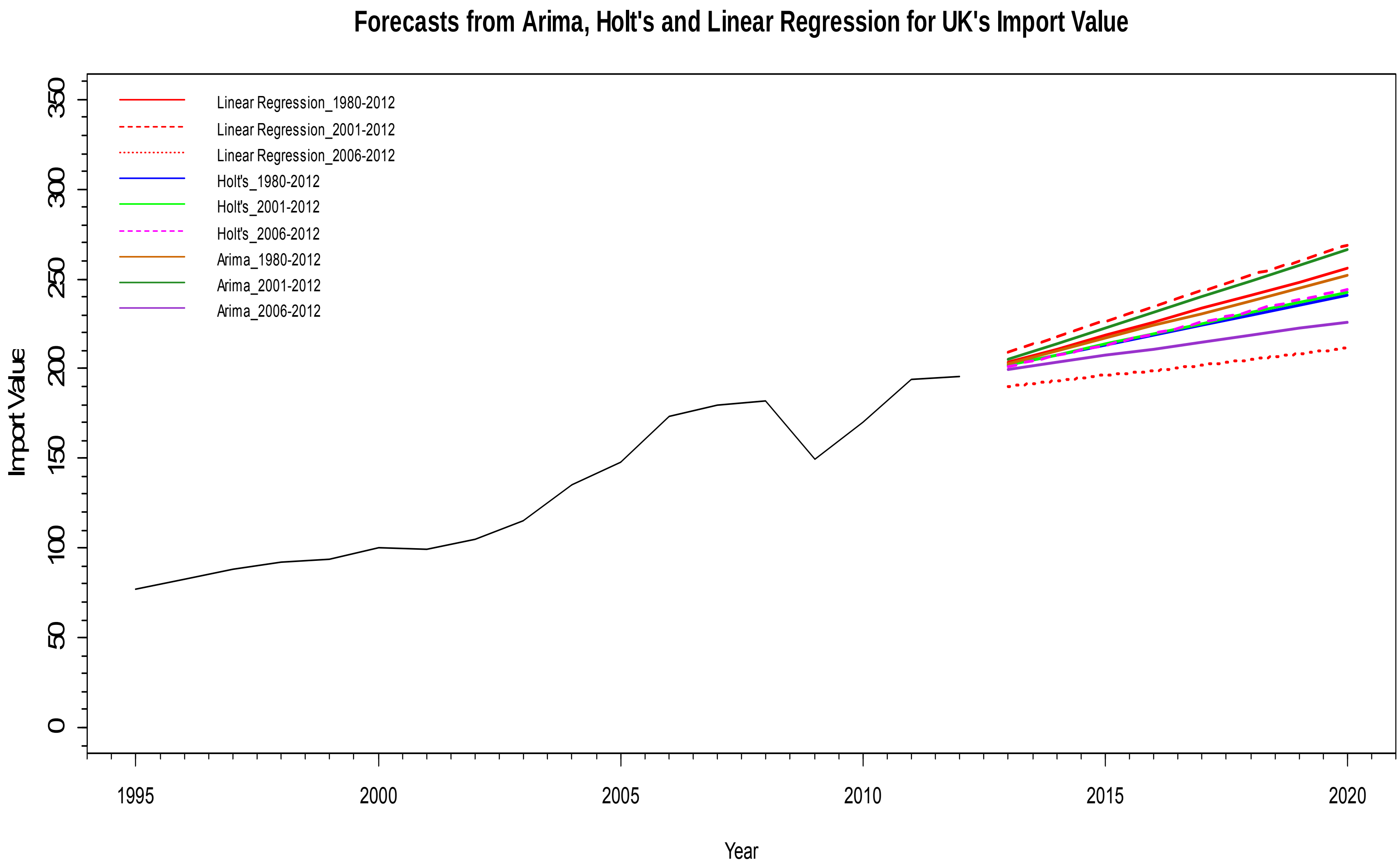


[Figure D.2.23] – Comparison of all predictions for India's Import Value

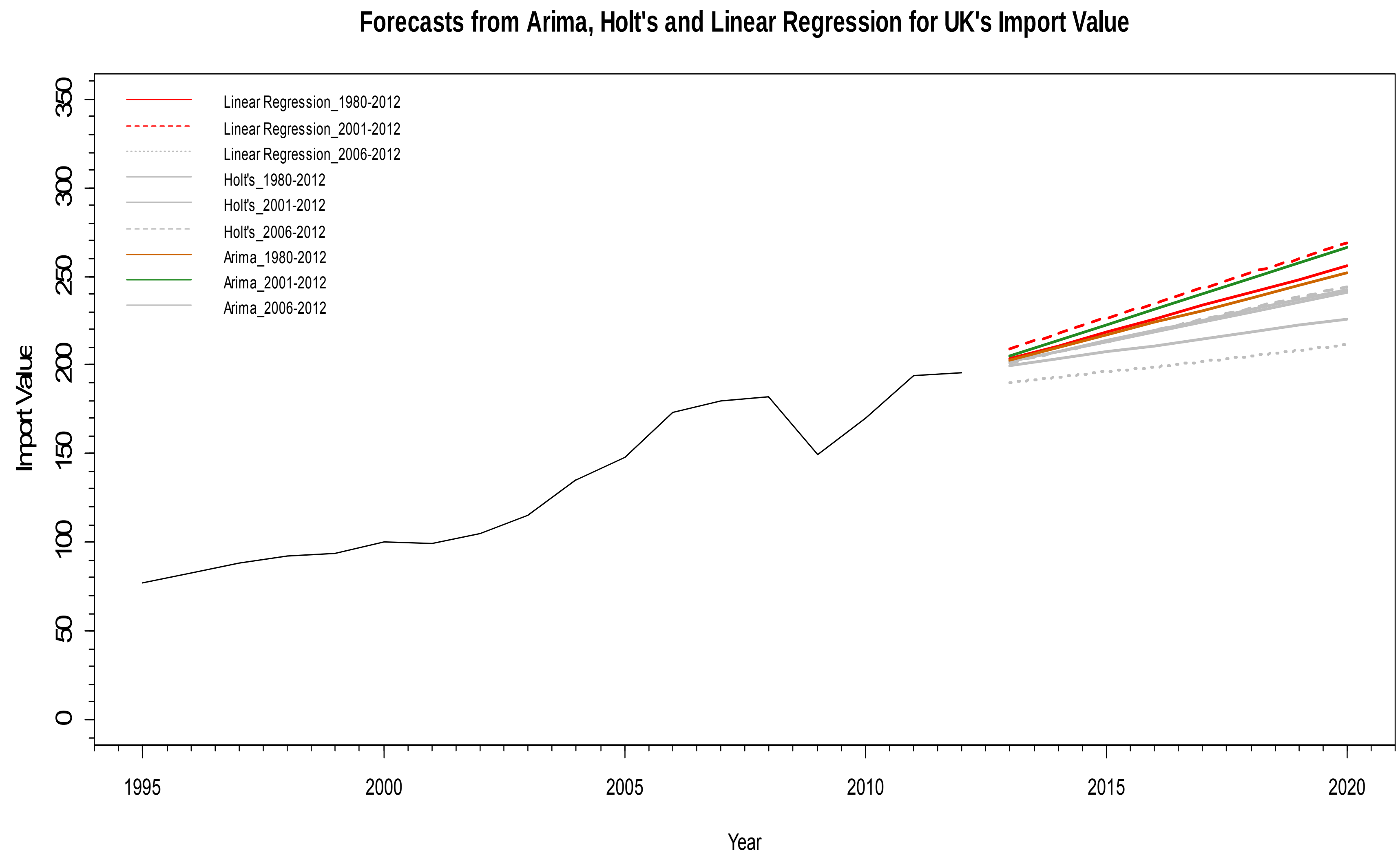


[Figure D.2.24] – Selection of the best predictions for India's Import Value

Import Value – UK

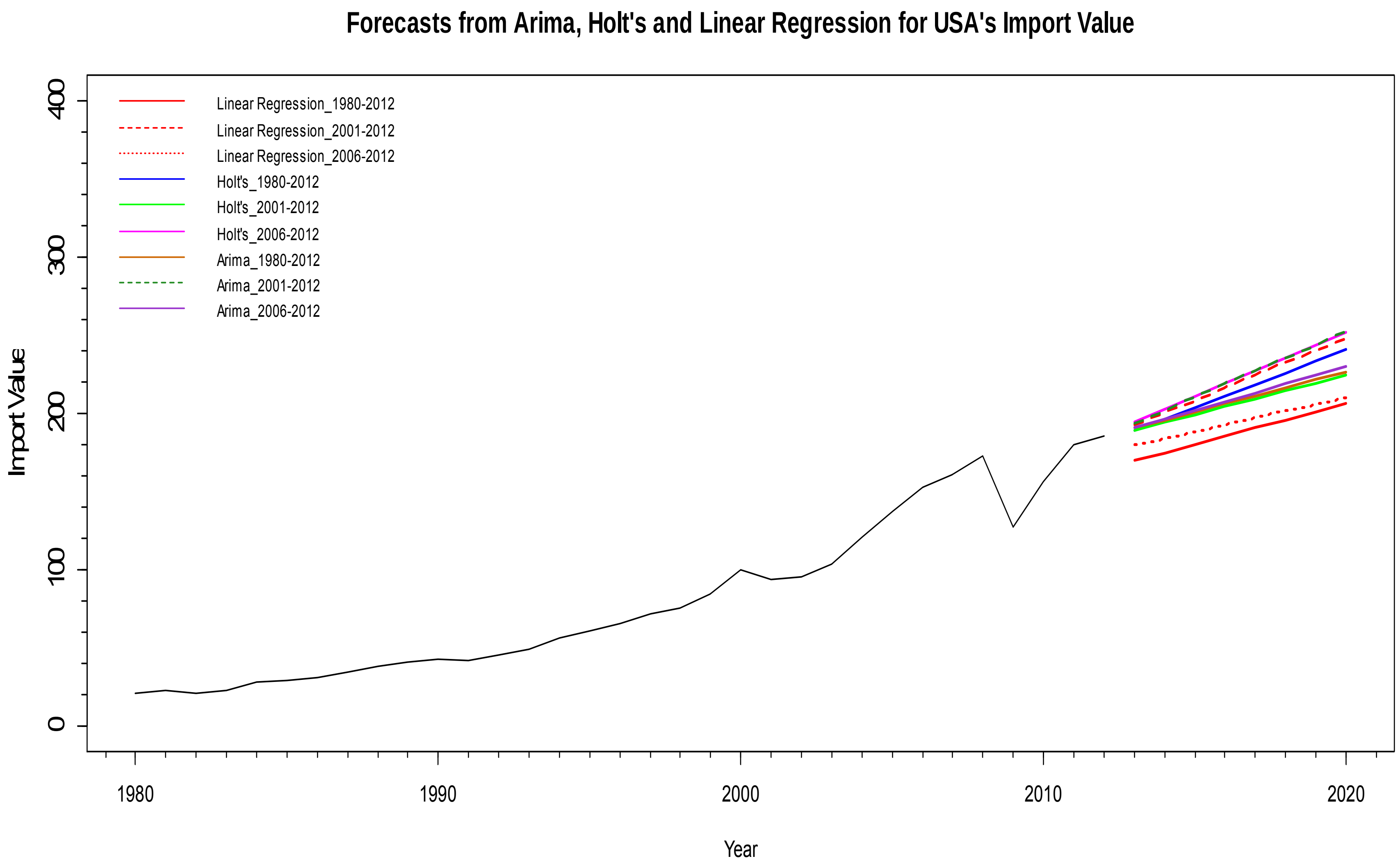


[Figure D.2.25] – Comparison of all predictions for UK's Import Value

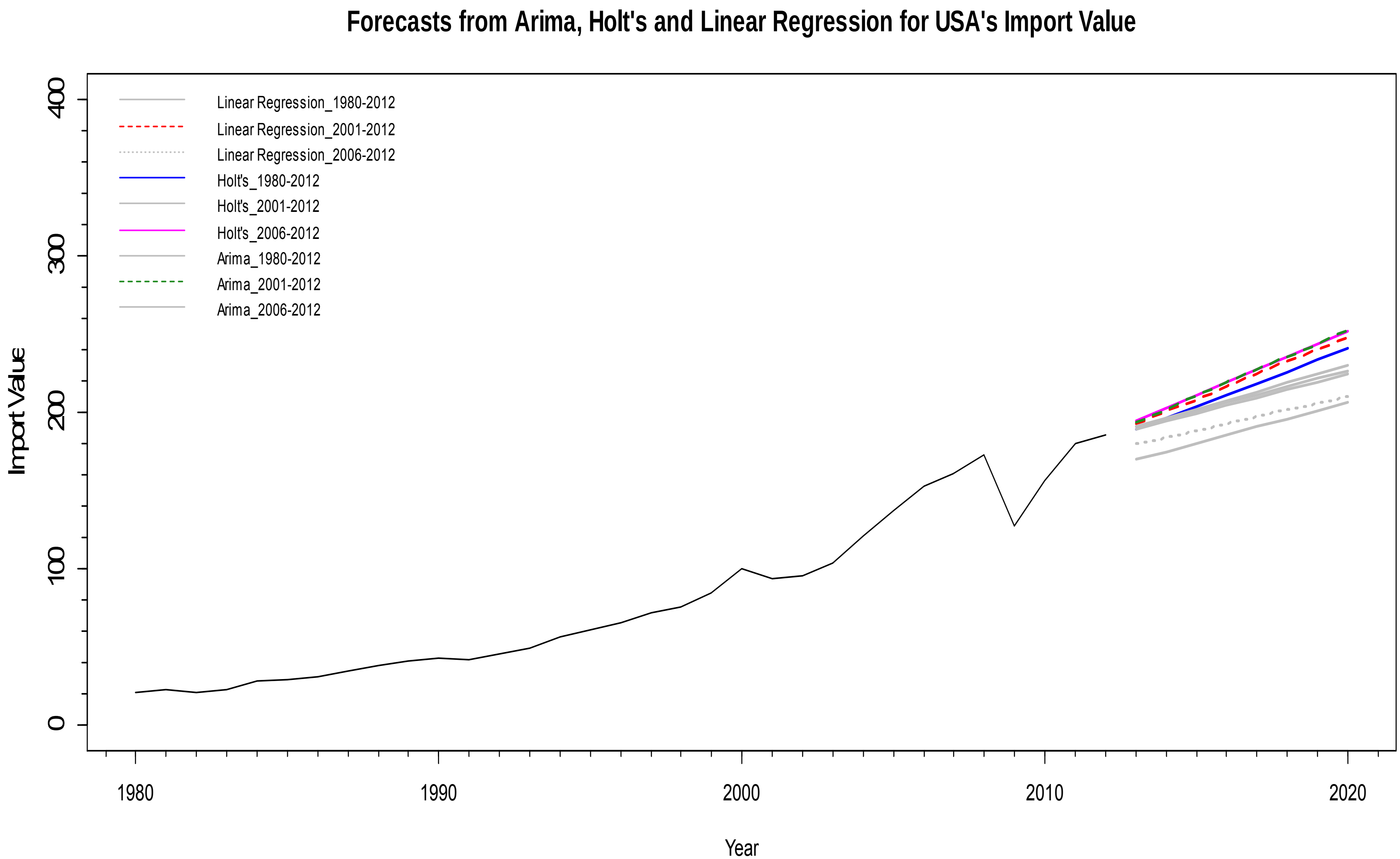


[Figure D.2.26] – Selection of the best predictions for UK's Import Value

Import Value – USA



[Figure D.2.26] – Comparison of all predictions for USA's Import Value



[Figure D.2.28] – Selection of the best predictions for USA's Import Value

Appendix E – Sample Code

E.1. General Code

E.1.1. Code for importing the data for the GDP per capita Indicator into R and transforming them into a suitable form before applying an algorithm

 # before someone applies the following code, has first to download the "GDP per capita_all.csv" file from the following link:
https://www.dropbox.com/s/9j4kgwg515czrdt/GDP%20per%20capita_all.csv?dl=0
and place the file into his/her local working directory of R

```
GDPpercap_all<-read.table("GDP per capita_all.csv", header= TRUE, sep = ",")  # read the data from the file in the working directory and create an object with the name GDPpercap_all

GDPpercap_all<-t(GDPpercap_all)   # interchanging columns and rows

colnames(GDPpercap_all)<-GDPpercap_all[1,] # editing the names of the columns

GDPpercap_all<-GDPpercap_all[-1,-1]  # deleting rows and columns that are not useful

GDPpercap_all<-ts(GDPpercap_all, 1960) #creating time series object starting at 1960

plot(GDPpercap_all, plot.type = c("single"), col= c('black', 'red', 'darkolivegreen','orange','yellow','blue','forestgreen','purple','magenta','green','cyan'), main='GDP per capita', ylab='GDP', xlab='Year', xlim=c(1960,2015), ylim=c(-10000,65000)) # plot all the times eries in one diagram

library(Hmisc)
minor.tick(nx=10, ny=4, tick.ratio=0.5) #add minor tick marks to the plot

legend('topleft', c('Brazil', 'China', 'India', 'Philippines', 'Russia', 'UK', 'USA', 'Middle East & North Africa', 'Latin Ameica', 'EURO zone', 'East Asia'), lty=1, col= c('black', 'red', 'darkolivegreen', 'orange', 'yellow','blue','forestgreen','purple','magenta','green','cyan'), bty='n', cex=.75) # adds legend to the plot
```

E.2. Linear Regression

E.2.1. Code for China, GDP per capita and whole dataset

```
Years<-c(1960:2013)  # create dataframe for the time period

lm.China_GDPpercap<-lm(GDPpercap_all[,2]~Years) # applies linear regression to the second column of GDPpercap_all object (China time series)

summary(lm.China_GDPpercap)

resid(lm.China_GDPpercap)  # calculates the residuals
fitted(lm.China_GDPpercap)  # calculates the fitted values

# layout(matrix(1:4,2,2))
# plot(lm.China_GDPpercap) # plots residuals for the evaluation of the model

plot(Years, GDPpercap_all[,2], xlim=c(1960,2020), ylim=c(-1000,12000), ylab='GDP per capita', main='Forecasts from Linear Regression for China\'s GDP per capita')    # plots the real values of the time series

library(Hmisc)
minor.tick(nx=10, ny=2, tick.ratio=0.5) # add minor tick marks to the plot

lines(Years,fitted(lm.China_GDPpercap), lty='dashed', col='black')
# plots the fitted values

newyears<-c(2014:2020)  # defines the years dataframe for the predictions

China_GDPpercap_pred<-predict(lm.China_GDPpercap,data.frame(Years=newyears), level = 0.95, interval = "confidence") # calculates the predictions for years 2014 to 2020

lines(newyears,China_GDPpercap_pred[,1], col='red') # plots the predictions on the previous plot

legend('topleft', c('Real values','Linear fitted values', 'Linear Regression forecasts'), col= c('black', 'black', 'red'), lty=c(3,2,1), bty='n', cex=.75)  # adds legend to the plot
```

E.2.2. Code for China, GDP per capita and the dataset up to year 2008

Years2<-c(1960:2008) # *create dataframe for the time period*

lm.China2_GDPpercap<-lm(GDPpercap_all[1:49,2]~Years2) # *applies linear regression to the second column of GDPpercap_all object until 2008 (China time series)*

summary(lm.China2_GDPpercap)

resid(lm.China2_GDPpercap) # *calculates the residuals*

fitted(lm.China2_GDPpercap) # *calculates the fitted values*

layout(matrix(1:4,2,2))

plot(lm.China2_GDPpercap) # *plots residuals for the evaluation of the model*

plot(Years, GDPpercap_all[,2], xlim=c(1960,2020), ylim=c(-1000,13000), ylab='GDP per capita', main='Forecastsat at 2008 from Linear Regression for China\'s GDP per capita') # *plots the real values of the time series*

library(Hmisc)
minor.tick(nx=10, ny=2, tick.ratio=0.5) # *add minor tick marks*

lines(Years2,fitted(lm.China2_GDPpercap), lty='dashed', col='green')
# *plots the fitted values*

newyears2<-c(2009:2020) # *defines the years dataframe for the predictions*

China2_GDPpercap_pred<-predict(lm.China2_GDPpercap,data.frame(Years2=newyears2), level = 0.95, interval = "confidence") # *calculates the predictions for years 2009 to 2020*

lines(newyears2,China2_GDPpercap_pred[,1], col='red') # *plots the predictions*

legend('topleft', c('Real values','Linear fitted values for 2014 analysis', 'Linear fitted values for 2008 analysis' ,'Linear Regression forecasts'), col= c('black', 'black', 'green', 'red'), lty=c(3,2,2,1), bty='n', cex=.75) # *adds legend to the plot*

E.2.3. Code for China, GDP per capita and the dataset up to year 2000

Years3<-c(1960:2000) # *create dataframe for the time period*

lm.China3_GDPpercap<-lm(GDPpercap_all[1:41,2]~Years3) # *applies linear regression to the second column of GDPpercap_all object until 2000 (China time series)*

summary(lm.China3_GDPpercap)

resid(lm.China3_GDPpercap) # *calculates the residuals*

fitted(lm.China3_GDPpercap) # *calculates the fitted values*

layout(matrix(1:4,2,2))

plot(lm.China3_GDPpercap) # *plots residuals for the evaluation of the model*

plot(Years, GDPpercap_all[,2], xlim=c(1960,2020), ylim=c(-1000,12000), ylab='GDP per capita', main='Forecasts from Linear Regression for China\'s GDP per capita') # *plots the real values of the time series*

library(Hmisc)
minor.tick(nx=10, ny=2, tick.ratio=0.5) #*add minor tick marks to the plot*

lines(Years3,fitted(lm.China3_GDPpercap), lty='dashed', col='green')
# *plots fitted values*

E.2.4. Code for China, GDP per capita and the dataset from year 2001 to year 2013

Years4<-c(2001:2013) # *create dataframe for the time period*

lm.China4_GDPpercap<-lm(GDPpercap_all[42:54,2]~Years4) # *applies linear regression to the second column of GDPpercap_all object for period 2001 to 2013 (China time series)*

summary(lm.China4_GDPpercap)

resid(lm.China4_GDPpercap) # *calculates the residuals*

fitted(lm.China4_GDPpercap) # *calculates the fitted values*

layout(matrix(1:4,2,2))

plot(lm.China4_GDPpercap) # *plots residuals for the evaluation of the model*

plot(Years, GDPpercap_all[,2], xlim=c(1960,2020), ylim=c(-1000,15000), ylab='GDP per capita', main='Forecasts from Linear Regression for China\'s GDP per capita with splitted dataset') # *plots the real values of the time series*

library(Hmisc)
minor.tick(nx=10, ny=5, tick.ratio=0.5) # add minor tick marks

lines(Years4,fitted(lm.China4_GDPpercap), lty='dashed', col='blue')
# plots the fitted values

newyears4<-c(2014:2020) # *defines the years dataframe for the predictions*

China4_GDPpercap_pred<-predict(lm.China4_GDPpercap,data.frame(Years4= newyears4), level = 0.95, interval = "confidence") # *calculates the predictions for years 2014 to 2020*

lines(newyears4,China4_GDPpercap_pred[,1], col='red') # *plots the predictions*

legend('topleft', c('Real values','Linear fitted values for 2014 analysis', 'Linear fitted values for 1960-2000 analysis' ,'Linear fitted values for 2001-2013 analysis', 'Linear Regression forecasts'), col= c('black', 'black', 'green', 'blue', 'red'), lty=c(3,2,2,2,1), bty='n', cex=.75) *# adds legend to the plot*

E.2.5. Code for China, GDP per capita and the dataset from year 2006 to year 2013

Years6<-c(2006:2013) *# create dataframe for the time period*

lm.China6_GDPpercap<-lm(GDPpercap_all[47:54,2]~Years6) # *applies linear regression to the second column of GDPpercap_all object for period 2006 to 2013 (China time series)*

summary(lm.China6_GDPpercap)

resid(lm.China6_GDPpercap) # *calculates the residuals*

fitted(lm.China6_GDPpercap) # *calculates the fitted values*

layout(matrix(1:4,2,2))

plot(lm.China6_GDPpercap) # *plots residuals for the evaluation of the model*

plot(Years, GDPpercap_all[,2], xlim=c(1960,2020), ylim=c(-1000,15000), ylab='GDP per capita', main='Forecasts from Linear Regression with only recent data for China\'s GDP per capita') # *plots the real values of the time series*

library(Hmisc)
minor.tick(nx=10, ny=5, tick.ratio=0.5) # *add minor tick marks*

lines(Years6,fitted(lm.China6_GDPpercap), lty='dashed', col='magenta')

plots the fitted values

newyears6<-c(2014:2020) # *defines the years dataframe for the predictions*

China6_GDPpercap_pred<-predict(lm.China6_GDPpercap,data.frame(Years6=newyears6), level = 0.95, interval = "confidence") # *calculates the predictions for years 2014 to 2020*

lines(newyears6,China6_GDPpercap_pred[,1], col='red') # *plots the predictions*

legend('topleft', c('Real values','Linear fitted values for 2014 analysis', 'Linear fitted values for 2001-2013 analysis' ,'Linear fitted values for 2006-2013 analysis', 'Linear Regression forecasts'), col= c('black', 'black', 'blue', 'magenta', 'red'), lty=c(3,2,2,2,1), bty='n', cex=.75) # *adds legend to the plot*

E.3. Holt's Exponential Smoothing

E.3.1. Code for China, GDP per capita and whole dataset

ChinaseriesGDPpercap<-GDPpercap_all[,2]

ChinaseriesGDPpercap<-ts(ChinaseriesGDPpercap,1960) # *create a time series object for China from the initial dataset*

mode(ChinaseriesGDPpercap)<-'numeric'

ChinaforecastsGDPpercap<-HoltWinters(ChinaseriesGDPpercap,gamma=FALSE, l.start=ChinaseriesGDPpercap[1],b.start=(ChinaseriesGDPpercap[2]-ChinaseriesGDPpercap[1])) # *applies Holt's Exponential smoothing to the time series object for China*

ChinaforecastsGDPpercap # *gives the a and b parameters of the smoothing*

ChinaforecastsGDPpercap$fitted # *gives the fitted values of smoothing*

ChinaforecastsGDPpercap$SSE # *gives the Sum of Squared Error (SSE)*

library("forecast") # *calls the forecast() function*

ChinaforecastsGDPpercap2<-forecast.HoltWinters(ChinaforecastsGDPpercap, h=7)
ChinaforecastsGDPpercap2 # *calculates and prints the forecasts for the next 7 years*

plot.forecast(ChinaforecastsGDPpercap2,ylab='China_GDPpercap',xlab='Year', main='Forecasts form HoltWinters for China\'s GDP per capital', ylim=c(-1000,15000)) # *plots the real values and the forecasts*

lines(ChinaforecastsGDPpercap$fitted[,1], col='red') # *plots the smoothed line*

library(Hmisc)
minor.tick(nx=10, ny=2, tick.ratio=0.5) # *add minor tick marks*

legend('topleft', c('Real values', 'Holt\'s Exponential Smoothing', 'Predictions'), lty=1, col= c('black', 'red', 'blue'), bty='n', cex=.75) # *adds legend to the plot*

acf(ChinaforecastsGDPpercap2$residuals, lag.max=20, main='Correlogram of the in-sample forecast errors for China\'s GDP per capital', cex.main=1.5, cex.lab=1.3, cex.axis=1.3) # *plots the autocorrelogram (ACF)*

Box.test(ChinaforecastsGDPpercap2$residuals, lag=20, type="Ljung-Box") # *performs the Ljung-Box test and gives the p-value* plot.ts(ChinaforecastsGDPpercap2$residuals, xlab='Year', main='Forecast residuals for China\'s GDP per capital', cex.lab=1.5, cex.axis=1.5)   # *plots the forecast residuals*

adjusted from the Using R for Time Series Analysis web page:

```
plotForecastErrors <- function(forecasterrors)
  {
    # make a histogram of the forecast errors:
    mybinsize <- IQR(forecasterrors)/4
    mysd   <- sd(forecasterrors)
    mymin  <- min(forecasterrors) - mysd*5
    mymax  <- max(forecasterrors) + mysd*3
    # generate normally distributed data with mean 0 and standard deviation mysd
    mynorm <- rnorm(10000, mean=0, sd=mysd)
    mymin2 <- min(mynorm)
    mymax2 <- max(mynorm)
    if (mymin2 < mymin) { mymin <- mymin2 }
    if (mymax2 > mymax) { mymax <- mymax2 }
    # make a red histogram of the forecast errors, with the normally distributed data overlaid:
    mybins <- seq(mymin, mymax, mybinsize)
    hist(forecasterrors, col="red", freq=FALSE, breaks=20, main='Histogram of forecast errors for China\'s GDP per capital', xlim=c(-800,800), cex.main=1.4, cex.lab=1.3, cex.axis=1.3)
    # freq=FALSE ensures the area under the histogram = 1
    # generate normally distributed data with mean 0 and standard deviation mysd
    myhist <- hist(mynorm, plot=FALSE, breaks=20)
    # plot the normal curve as a blue line on top of the histogram of forecast errors:
    points(myhist$mids, myhist$density, type="l", col="blue", lwd=2)
  }
```

plotForecastErrors(ChinaforecastsGDPpercap2$residuals) # *plots a histogram of forecasts residuals with a normal distribution curve*

E.3.2. Code for China, GDP per capita and the dataset up to year 2008

ChinaseriesGDPpercap_2008<-GDPpercap_all[1:49,2]

ChinaseriesGDPpercap_2008<-ts(ChinaseriesGDPpercap_2008,1960) # *create a time series object for China until year 2008 from the initial dataset*

mode(ChinaseriesGDPpercap_2008)<-'numeric'

ChinaforecastsGDPpercap_2008<-HoltWinters(ChinaseriesGDPpercap_2008, gamma=FALSE,l.start=ChinaseriesGDPpercap_2008[1], b.start=(ChinaseriesGDPpercap_2008[2]-ChinaseriesGDPpercap_2008[1]))
# *applies Holt's Exponential smoothing to the time series object for China until year 2008*

```
ChinaforecastsGDPpercap_2008 #gives the a and b parameters of the smoothing

ChinaforecastsGDPpercap_2008$fitted #gives the fitted values of smoothing

ChinaforecastsGDPpercap_2008$SSE # gives the Sum of Squared Error (SSE)

library("forecast")  #  calls the forecast() function

ChinaforecastsGDPpercap2_2008<- forecast.HoltWinters(ChinaforecastsGDPpercap_2008, h=12)
ChinaforecastsGDPpercap2_2008 # calculates and prints the forecasts for next 12 years

plot.forecast(ChinaforecastsGDPpercap2_2008, ylab='China_GDP per capital_2008', xlab='Year', main='Forecasts from HoltWinters for China\'s GDP per capital', ylim=c(0,15000), flty='dashed') # plots the real values and the forecasts

lines(ChinaforecastsGDPpercap_2008$fitted[,1], col='red') # plots the smoothed line
library(Hmisc)
minor.tick(nx=10, ny=2, tick.ratio=0.5) # add minor tick marks

lines(GDPpercap_all[,2], lty='dashed')# plots the real values from 2009 to 2013
lines(ChinaforecastsGDPpercap2$mean, col='blue', lwd=2) # plots the forecasts made with the whole dataset

legend('topleft', c('Real values','Holt\'s Exponential Smoothing_2008', 'Predictions'), lty=1, col=c('black','red','blue'), bty='n', cex=.75)
# adds legend to the plot
```

E.3.3. Code for China, GDP per capita and the dataset from year 2001 to year 2013

```
ChinaseriesGDPpercap_2001<-GDPpercap_all[42:54,2]
ChinaseriesGDPpercap_2001<-ts(ChinaseriesGDPpercap_2001, 2001) # create a time series object for China from year 2001 until year 2013 from the initial dataset

mode(ChinaseriesGDPpercap_2001)<-'numeric'

ChinaforecastsGDPpercap_2001<-HoltWinters(ChinaseriesGDPpercap_2001,
gamma=FALSE,l.start=ChinaseriesGDPpercap_2001[1],    b.start=(ChinaseriesGDPpercap_2001[2]-
ChinaseriesGDPpercap_2001[1]))
# applies Holt's Exponential smoothing to the time series object for China from year 2001 to year 2013

ChinaforecastsGDPpercap_2001 # gives the a and b parameters of the smoothing

ChinaforecastsGDPpercap_2001$fitted #gives the fitted values of smoothing
```

ChinaforecastsGDPpercap_2001$SSE # *gives the Sum of Squared Error (SSE)*

library("forecast") # *calls the forecast() function*

ChinaforecastsGDPpercap2_2001<-forecast.HoltWinters(ChinaforecastsGDPpercap_2001, h=7)
ChinaforecastsGDPpercap2_2001   # *calculates and prints the forecasts for next 7 years*

plot.forecast(ChinaforecastsGDPpercap2_2001, ylab='China_GDP per capital_2001-2013', xlab='Year',main='Forecasts from HoltWinters for China\'s GDP per capital',fcol='green', ylim=c(0,15000)) # *plots the real values and the forecasts*

lines(ChinaforecastsGDPpercap_2001$fitted[,1], col='red') # *plots the smoothed line*
library(Hmisc)
minor.tick(nx=5, ny=2, tick.ratio=0.5) # *add minor tick marks*

lines(ChinaforecastsGDPpercap2$mean, col='blue', lwd=2, lty='dashed') # *plots the forecasts made with the whole dataset*

legend('topleft', c('Real values','Predictions with Holt\'s Exponential Smoothing_1970-2013','Predictions with Holt\'s Exponential Smoothing_2001_2013'), lty=1, col= c('black', 'blue', 'green'), bty='n', cex=.75)   # *adds legend to the plot*

E.3.4. <u>Code for China, GDP per capita and the dataset from year 2006 to year 2013</u>

ChinaseriesGDPpercap_2006<-GDPpercap_all[47:54,2]
ChinaseriesGDPpercap_2006<-ts(ChinaseriesGDPpercap_2006, 2006) # *create a time series object for China from year 2006 until year 2013 from the initial dataset*

mode(ChinaseriesGDPpercap_2006)<-'numeric'
ChinaforecastsGDPpercap_2006<-HoltWinters(ChinaseriesGDPpercap_2006,
gamma=FALSE,l.start=ChinaseriesGDPpercap_2006[1], b.start=(ChinaseriesGDPpercap_2006[2]-ChinaseriesGDPpercap_2006[1]))
# *applies Holt's Exponential smoothing to the time series object for China from year 2006 to year 2013*

ChinaforecastsGDPpercap_2006 # *gives the a and b parameters of the smoothing*

ChinaforecastsGDPpercap_2006$fitted #*gives the fitted values of smoothing*

ChinaforecastsGDPpercap_2006$SSE # *gives the Sum of Squared Error (SSE)*

library("forecast") # *calls the forecast() function*

ChinaforecastsGDPpercap2_2006<- forecast.HoltWinters(ChinaforecastsGDPpercap_2006, h=7)
ChinaforecastsGDPpercap2_2006 # *calculates and prints the forecasts for next 7 years*

plot.forecast(ChinaforecastsGDPpercap2_2006, ylab='China_GDP per capital', xlab='Year',main='Forecasts from HoltWinters for China\'s GDP per capital', fcol='magenta', xlim=c(1960,2020), plot.conf=FALSE, ylim=c(0,13000), flwd=4) # *plots the real values from 2006 to 2013 and the forecasts*

lines(GDPpercap_all[,2]) # *plots the real values*
lines(ChinaforecastsGDPpercap2$mean, col='blue', lwd=4, lty='dotted') # *plots the forecasts made with the whole dataset*
lines(ChinaforecastsGDPpercap2_2001$mean,col='green',lwd=4,lty='dashed')) # *plots the forecasts made with the dataset from year 2001 to year 2013*

legend('topleft', c('Real values','Predictions with Holt\'s Exponential Smoothing_1970-2013', 'Predictions with Holt\'s Exponential Smoothing_2001-2013', 'Predictions with Holt\'s Exponential Smoothing_2006-2013'), lty=1, col= c('black', 'blue', 'green','magenta'), bty='n', cex=.75) # *adds legend to the plot*

E.4. ARIMA

E.4.1. Code for China, GDP per capita and whole dataset

```
ChinaseriesGDPpercap<-GDPpercap_all[,2]
ChinaseriesGDPpercap<-ts(ChinaseriesGDPpercap,1960) # create a time series object for China from the initial dataset

mode(ChinaseriesGDPpercap)<-'numeric'
ChinaseriesGDPpercapdiff1<-diff(ChinaseriesGDPpercap, differences=1)
plot(ChinaseriesGDPpercapdiff1, main='China time series differenced 1 time', cex.lab=1.5, cex.axis=1.5, cex.main=1.8) # difference the time series once and plot it

ChinaseriesGDPpercapdiff2<-diff(ChinaseriesGDPpercap, differences=2)
plot(ChinaseriesGDPpercapdiff2, main='China time series differenced 2 times', cex.lab=1.5, cex.axis=1.5, cex.main=1.8) # difference the time series twice and plot it

acf(ChinaseriesGDPpercapdiff2, lag.max=20, main='Correlogram of the 2 times differenced series for China\'s GDP per capital', cex.main=2, cex.lab=1.3, cex.axis=1.3) # plots the autocorrelogram of the two times differenced time series
acf(ChinaseriesGDPpercapdiff2, lag.max=20, plot=FALSE) # give the autocorrelation values

pacf(ChinaseriesGDPpercapdiff2, lag.max=20, main='Partial Correlogram of the 2 times differenced series for China\'s GDP per capital', cex.main=2, cex.lab=1.3, cex.axis=1.3) # plots the partial autocorrelogram of the two times differenced time series
pacf(ChinaseriesGDPpercapdiff2, lag.max=20, plot=FALSE) # give the partial autocorrelation values

ChinaseriesGDPpercaparima<-arima(ChinaseriesGDPpercap,order=c(0,2,3))
# applies ARIMA(0,2,3) model to the time series

library("forecast") # calls the forecast() function
ChinaseriesGDPpercaparimaforecasts<-forecast.Arima(ChinaseriesGDPpercaparima, h=7)
ChinaseriesGDPpercaparimaforecasts # calculates and prints the forecasts for next 7 years

plot.forecast(ChinaseriesGDPpercaparimaforecasts, ylab='China_GDP per capital', xlab='Year', main='Forecasts from Arima(0,2,3) for China\'s GDP per capital', ylim=c(0,15000), fcol='darkorange3', flwd=3, cex.main=1.3) ) # plots the real values and the forecasts
legend('topleft', c('Real values','Arima forecasts'), lty=c(1,1), col= c('black', 'darkorange3'), bty='n', cex=.75) # adds legend to the plot

acf(ChinaseriesGDPpercaparimaforecasts$residuals,lag.max=20,main='Correlogram of the in-sample Arima forecast errors for China\'s GDP per capital') ) # plots the autocorrelogram (ACF)
Box.test(ChinaseriesGDPpercaparimaforecasts$residuals,lag=20,type="Ljung-Box") # performs the Ljung-Box test and gives the p-value
```

```
plot.ts(ChinaseriesGDPpercaparimaforecasts$residuals,xlab='Year',main='Arima forecast residuals
for China\'s GDP per capital', cex.lab=1.5, cex.axis=1.5)  # plots the forecast residuals

# adjusted from the Using R for Time Series Analysis web page:
plotForecastErrors <- function(forecasterrors)
  {    # make a histogram of the forecast errors:
    mybinsize <- IQR(forecasterrors)/4
    mysd   <- sd(forecasterrors)
    mymin  <- min(forecasterrors) - mysd*5
    mymax  <- max(forecasterrors) + mysd*3
    # generate normally distributed data with mean 0 and standard deviation mysd
    mynorm <- rnorm(10000, mean=0, sd=mysd)
    mymin2 <- min(mynorm)
    mymax2 <- max(mynorm)
    if (mymin2 < mymin) { mymin <- mymin2 }
    if (mymax2 > mymax) { mymax <- mymax2 }
    # make a red histogram of the forecast errors, with the normally distributed data overlaid:
    mybins <- seq(mymin, mymax, mybinsize)
    hist(forecasterrors, col="red", freq=FALSE, breaks=20, main='Histogram of forecast errors for
China\'s GDP per capital', xlim=c(-300,300), cex.main=1.4, cex.lab=1.3, cex.axis=1.3)
    # freq=FALSE ensures the area under the histogram = 1
    # generate normally distributed data with mean 0 and standard deviation mysd
    myhist <- hist(mynorm, plot=FALSE, breaks=20)
    # plot the normal curve as a blue line on top of the histogram of forecast errors:
    points(myhist$mids, myhist$density, type="l", col="blue", lwd=2)
  }
plotForecastErrors(ChinaseriesGDPpercaparimaforecasts$residuals)
# plots a histogram of forecasts residuals with a normal distribution curve
```

E.4.2. Code for China, GDP per capita and the dataset up to year 2008

```
ChinaseriesGDPpercap_2008<-GDPpercap_all[1:49,2]
ChinaseriesGDPpercap_2008<-ts(ChinaseriesGDPpercap_2008,1960) # create a time series object for China until year 2008 from the initial dataset

mode(ChinaseriesGDPpercap_2008)<-'numeric'
ChinaseriesGDPpercap_2008diff1<-diff(ChinaseriesGDPpercap_2008, differences=1)
plot(ChinaseriesGDPpercap_2008diff1, main='China time series differenced 1 time', cex.lab=1.5, cex.axis=1.5, cex.main=1.8)  # difference the time series once and plot it

ChinaseriesGDPpercap_2008diff2<-diff(ChinaseriesGDPpercap_2008, differences=2)
plot(ChinaseriesGDPpercap_2008diff2, main='China time series differenced 2 times', cex.lab=1.5, cex.axis=1.5, cex.main=1.8) # difference the time series twice and plot it

acf(ChinaseriesGDPpercap_2008diff2, lag.max=20, main='Correlogram of the 2 times differenced series for China\'s GDP per capital')  # plots the autocorrelogram of the two times differenced time series
acf(ChinaseriesGDPpercap_2008diff2, lag.max=20, plot=FALSE

pacf(ChinaseriesGDPpercap_2008diff2,lag.max=20,main='Partial Correlogram of the 2 times differenced series for China\'s GDP per capital')
# plots the partial autocorrelogram of two times differenced time series
pacf(ChinaseriesGDPpercap_2008diff2, lag.max=20, plot=FALSE

ChinaseriesGDPpercap_2008arima<-arima(ChinaseriesGDPpercap_2008, order=c(0,2,1)) # applies ARIMA(0,2,1) model to the time series

library("forecast")  #  calls the forecast() function
ChinaseriesGDPpercap_2008arimaforecasts<-forecast.Arima(ChinaseriesGDPpercap_2008arima,
h=12)
ChinaseriesGDPpercap_2008arimaforecasts # calculates the forecasts for next 12 years

plot.forecast(ChinaseriesGDPpercap_2008arimaforecasts,ylab='China_GDP per capital', xlab='Year', main='Forecasts from Arima(0,2,1) for China\'s GDP per capital', ylim=c(0,16000), fcol='darkorange3', flwd=3, cex.main=1.3, flty='dashed')  # plots the real values until 2008 and the forecasts

lines(GDPpercap_all[,2], lty='dashed') # plots the real values from 2009 to 2013
lines(ChinaseriesGDPpercaparimaforecasts$mean, col='darkorange3', lwd=2)# plots the forecasts made with the whole dataset
legend('topleft', c('Real values','Arima forecasts_2008','Arima forecasts'),  lty=c(1,2,1), col= c('black', 'darkorange3', 'darkorange3'), bty='n', cex=.75)  # adds legend to the plot
```

E.4.3. Code for China, GDP per capita and the dataset from year 2001 to year 2013

ChinaseriesGDPpercap_2001<-GDPpercap_all[42:54,2]
ChinaseriesGDPpercap_2001<-ts(ChinaseriesGDPpercap_2001, 2001) *# create a time series object for China from year 2001 until year 2013 from the initial dataset*

mode(ChinaseriesGDPpercap_2001)<-'numeric'
ChinaseriesGDPpercap_2001diff1<-diff(ChinaseriesGDPpercap_2001, differences=1)
plot(ChinaseriesGDPpercap_2001diff1, main='China time series differenced 1 time', cex.lab=1.5, cex.axis=1.5, cex.main=1.8) *# difference the time series once and plot it*

ChinaseriesGDPpercap_2001diff2<-diff(ChinaseriesGDPpercap_2001, differences=2)
plot(ChinaseriesGDPpercap_2001diff2, main='China time series differenced 2 times', cex.lab=1.5, cex.axis=1.5, cex.main=1.8) *# difference the time series twice and plot it*

acf(ChinaseriesGDPpercap_2001diff2, lag.max=20, main='Correlogram of the 2 times differenced series for China\'s GDP per capital') *# plots the autocorrelogram of the two times differenced time series*
acf(ChinaseriesGDPpercap_2001diff2, lag.max=20, plot=FALSE

pacf(ChinaseriesGDPpercap_2001diff2, lag.max=20, main='Partial Correlogram of the 2 times differenced series for China\'s GDP per capital')
 # plots the autocorrelogram of the two times differenced time series
pacf(ChinaseriesGDPpercap_2001diff2, lag.max=20, plot=FALSE

ChinaseriesGDPpercap_2001arima<-arima(ChinaseriesGDPpercap_2001, order=c(2,2,0)) *# applies ARIMA(0,2,2) model to the time series*

library("forecast") *# calls the forecast() function*
ChinaseriesGDPpercap_2001arimaforecasts<-forecast.Arima (ChinaseriesGDPpercap_2001arima, h=7)
ChinaseriesGDPpercap_2001arimaforecasts *# calculates the forecasts for next 7 years*

plot.forecast(ChinaseriesGDPpercap_2001arimaforecasts, ylab='China_GDP per capital', xlab='Year', main='Forecasts from Arima(2,2,0) for China\'s GDP per capital', ylim=c(0,20000), fcol='forestgreen', flwd=3, cex.main=1.3) *# plots the real values from 2001 to 2013 and the forecasts*

lines(GDPpercap_all[,2]) *# plots the real values*
lines(ChinaseriesGDPpercaparimaforecasts$mean, col='darkorange3', lwd=2) *# plots the forecasts made with the whole dataset*

legend('topleft', c('Real values','Arima forecasts','Arima forecasts_2001_2013'),lty=1,col=c('black','darkorange3','darkolivegreen4'), bty='n', cex=.75) *# adds legend to the plot*

E.4.4. Code for China, GDP per capita and the dataset from year 2006 to year 2013

ChinaseriesGDPpercap_2006<-GDPpercap_all[47:54,2]
ChinaseriesGDPpercap_2006<-ts(ChinaseriesGDPpercap_2006, 2006) # *create a time series object for China from year 2006 until year 2013 from the initial dataset*

mode(ChinaseriesGDPpercap_2006)<-'numeric'
ChinaseriesGDPpercap_2006diff1<-diff(ChinaseriesGDPpercap_2006, differences=1)
plot(ChinaseriesGDPpercap_2006diff1, main='China time series differenced 1 time', cex.lab=1.5, cex.axis=1.5, cex.main=1.8) # *difference the time series once and plot it*

ChinaseriesGDPpercap_2006diff2<-diff(ChinaseriesGDPpercap_2006, differences=2)
plot(ChinaseriesGDPpercap_2006diff2, main='China time series differenced 2 times', cex.lab=1.5, cex.axis=1.5, cex.main=1.8) # *difference the time series twice and*

acf(ChinaseriesGDPpercap_2006diff2, lag.max=20, main='Correlogram of the 2 times differenced series for China\'s GDP per capital') # *plots the autocorrelogram of the two times differenced time series*
acf(ChinaseriesGDPpercap_2006diff2, lag.max=20, plot=FALSE)

pacf(ChinaseriesGDPpercap_2006diff2,lag.max=20, main='Partial Correlogram of the 2 times differenced series for China\'s GDP per capital') # *plots the autocorrelogram of the two times differenced time series*
pacf(ChinaseriesGDPpercap_2006diff2, lag.max=20, plot=FALSE)

ChinaseriesGDPpercap_2006arima<-arima(ChinaseriesGDPpercap_2006, order=c(0,2,3))
# *applies ARIMA(0,2,3) model to the time*

library("forecast") # *calls the forecast() function*

ChinaseriesGDPpercap_2006arimaforecasts<-forecast.Arima(ChinaseriesGDPpercap_2006arima, h=7)
ChinaseriesGDPpercap_2006arimaforecasts # *calculates the forecasts for next 7 years*

plot.forecast(ChinaseriesGDPpercap_2006arimaforecasts, ylab='China_GDP per capital', xlab='Year', main='Forecasts from Arima(0,2,3) for China\'s GDP per capital', xlim=c(1960,2020), ylim=c(0,15000), fcol='darkorchid3', flwd=4, cex.main=1.3, plot.conf=FALSE) # *plots the real values from 2006 to 2013 and the forecasts*

lines(GDPpercap_all[,2]) # *plots the real values*
lines(ChinaseriesGDPpercaparimaforecasts$mean, col='darkorange3', lwd=4)
lines(ChinaseriesGDPpercap_2001arimaforecasts$mean, col='forestgreen', lwd=4, lty='dashed') # *plots the forecasts made with the dataset from year 2001 to year 2013*