Estimation and Forecasting Team Strength Dynamics in Football: Investigation into Structural Breaks

José Sellitti Rangel Junior

A thesis submitted to the University of Salford in partial fulfilment of the requirements for the degree of Doctor of Philosophy (PhD).

Salford Business School

University of Salford

December 2018

Table of Contents

List of Figures ..... **iii**

List of Tables ..... **v**

Acknowledgements ..... **vi**

Declaration ..... **vii**

Abbreviations ..... **viii**

Abstract ..... **ix**

1. Introduction **1**

Contribution **2**

Outline **3**

Publications **3**

2. Literature Review **4**

Direct match outcome models **4**

Indirect outcome models **7**

Team strength dynamics **12**

Structural breaks and forecasting **16**

3. Data Analysis and Results **20**

Methodology **20**

Estimation of time decay **24**

Team strength dynamics **26**

Testing for the presence of structural breaks **35**

4. Forecasting **65**

Model 1 – Random walk **65**

Model 2 – Random walk with drift **65**

Model 3 – Auto-ARIMA **66**

Model 4 – Time-varying parameter model **66**

Model 5 – Equal weights model averaging **68**

One-step ahead forecast of team strengths **68**

Mean Square Error (MSE) comparison **74**

Forecasting match outcome probabilities **77**

Reliability plots for out-of-sample match outcome predictions 77

Average RPS for out-of-sample match outcome predictions 80

Discussion **81**

5. Conclusion **84**

References **87**

Appendix A – Forecasts from the simple random walk model **93**

Appendix B – Forecasts from the random walk with drift model **117**

Appendix C – Forecasts from the auto-arima model **141**

Appendix D – Forecasts from the time-varying parameter model **165**

Appendix E – Smoothed State Estimates – Kalman Smoothing **189**

List of Figures

Figure 1 – Mean Ranked Probability Score versus ξ, the decay factor 25

Figure 2 – Predicted Log-Likelihood versus ξ, the decay factor 25

Figure 3 – Attack Strength Parameter Estimates over 9 seasons (until 2016-17) 28

Figure 4 - Defence Strength Parameter Estimates over 9 seasons (until 2016-17) 31

Figure 5 – Time series for home advantage over the 266 rounds 34

Figure 6 - Time series for ρ over the 266 rounds 35

Figure 7 - Attack Breakpoint Estimates over 266 rounds, from 2010-11 season to 2016-17 season 37

Figure 8 - Defence Breakpoint Estimates over 266 rounds, from 2010-11 season to 2016-17 season 49

Figure 9 - (Model 1) Simple Random Walk Forecasts for Arsenal – Attack 70

Figure 10 - (Model 1) Simple Random Walk Forecasts for Arsenal – Defence 70

Figure 11 - (Model 2) Random Walk With Drift Forecasts for Arsenal – Attack 71

Figure 12 - (Model 2) Random Walk With Drift Forecasts for Arsenal - Defence 71

Figure 13 - (Model 3) Auto-Arima Forecasts for Arsenal – Attack 72

Figure 14 - (Model 3) Auto-Arima Forecasts for Arsenal – Defence 72

Figure 15 - (Model 4) TVP Forecasts using Kalman Filter for Arsenal – Attack 73

Figure 16 - (Model 4) TVP Forecasts using Kalman Filter for Arsenal – Defence 73

Figure 17 - Reliability Plots and Histogram of Model Forecasts 78

Figure 18 - (Model 1) Simple Random Walk Forecasts – Attack 93

Figure 19 - (Model 1) Simple Random Walk Forecasts – Defence 105

Figure 20 - (Model 2) Random Walk With Drift Forecasts – Attack 117

Figure 21 - (Model 2) Random Walk With Drift Forecasts - Defence 129

Figure 22 - (Model 3) Auto-Arima Forecasts – Attack 141

Figure 23 - (Model 3) Auto-Arima Forecasts – Defence 153

Figure 24 - (Model 4) Time-varying Parameter Forecasts using Kalman Filter – Attack 165

Figure 25 - (Model 4) Time-varying Parameter Forecasts using Kalman Filter – Defence .. 177

Figure 26 - Kalman Smoothing State Estimates (model 4) – Attack 189

Figure 27 - Kalman Smoothing State Estimates (model 4) – Defence 197

List of Tables

Table 1 - Mean attack and defence strengths of teams and the standard errors from the mean strength parameter over 266 match rounds 27

Table 2 - Parameter Breakpoint Estimates Attack 61

Table 3 - Parameter Breakpoint Estimates Defence 63

Table 4 - Mean Square Error Forecast - Attack 75

Table 5 - Mean Square Error Forecast - Defence 76

Table 6 - Mean RPS Out-of-Sample Forecast 81

Acknowledgements

First and foremost, I would like to thank God to have had the opportunity to embark in this journey. It has not been an easy one. I have thought of giving up several times over the course of this thesis (as every PhD student does), but trusting in His Divine Providence, I believe I was in the right place at the right time for reasons unbeknown to me. Therefore I dedicate it first and foremost to Him and to His Blessed Mother, Who I believe to be interceding for me at all times. I want to thank all those members of the Church who have given me the encouragement and support, in particular Fr Shaun K. Braiden from St James and All Souls Catholic Church (Salford Precinct) and Roman Catholic Chaplain of the University of Salford, along with Sister Teresa Wild and Fr James Cadman. I will remember my time at the Chaplaincy fondly and cherish the great memories that were made there.

I also thank all my friends: those I have made along the way and those who have already been in my life before I began my thesis at the University of Salford in September, 2013. They have been incredibly supportive.

I want to thank Professor Ian McHale, who was my supervisor during my first year, and Professor Philip Scarf, my supervisor from the second year onward. Thank you for your ideas, inputs, motivation, guidance, help and support. I have learned a lot from both of you.

I wish to thank my girlfriend Maja for her love and support. She has been extremely caring, patient and encouraging with me, particularly in the final stretch of the thesis. Thank you for everything.

Last, but not least, I want to thank my whole family, in particular my parents and sister. Mum and Dad, thank you for all the support and encouragement, from start to finish. Not only financially throughout the thesis, but emotionally and psychologically too. Liz, thank you for being the wonderful sister you are.

Declaration

I declare that the research contained in this thesis was carried out by me. It has not been previously submitted to this or any other Institution for the award of a degree or any other qualification.

I declare that this thesis:

1) embodies the results of my own course of study and research
2) has been composed by myself
3) has been seen by my supervisor before presentation
4) has been granted the appropriate level of ethics approval.

Abbreviations

2SLS – Two-Stage Least Squares.

ADF – Augmented Dickey-Fuller.

AIC – Akaike Information Criterion

AR(p) – Autoregressive process of order p.

ARFIMA – Autoregressive Fractionally Integrated Moving Average.

ARIMA(p,d,q) – Autoregressive Integrated Moving Average.

BIC – Bayes Information Criterion.

CDF – Cumulative Distribution Function.

DF-GLS ERS test – Dickey-Fuller Generalised Least Squares Elliot-Rothenberg and Stock test.

DGLM – Dynamic Generalised Linear Modelling.

DGP – Data Generating Process.

EM – Expectation-Maximisation.

EWMA – Exponential Weighted Moving Average.

GK – Giordani and Kohn (2008) model.

I(d) process – Integrated of order d.

IV – Instrumental Variable.

KM – Kaplan Meier.

KP – Koop and Potter (2007) model.

LRT – Likelihood Ratio Test.

MCMC – Markov Chain Monte Carlo.

MSE – Mean Square Error.

OLS – Ordinary Least Squares.

PLL – Predicted Log-Likelihood.

PPT – Pesaran et al. (2006) model.

RPS – Ranked Probability Score.

RSS – Residual Sum of Squares.

TVP – Time-varying Parameter.

TVP-VAR – Time-Varying Parameter Vector Autoregression.

Abstract

This PhD thesis studies the dynamics of team strengths in football. It investigates the presence of structural breaks, which occur when there is a change in parameters that govern dynamics in a time series. In football, such structural breaks occur because of events such as squad changes during transfer markets as well as managerial or ownership changes.

Team strengths are estimated across seven seasons of the Premiership and Championship football leagues and then analysed through a time series perspective, based on the double Poisson model with an added dependence parameter for lower scores and an exponential decay factor that adds more weight to more recent matches. This weighting scheme means that a pseudo-likelihood is used to estimate strength parameters. A rolling window approach is used to obtain a time series for the attack and defence strengths of teams in order to investigate the presence of structural breaks. We show that structural breaks are present in the majority of the time series. These present a challenge for the prediction of match outcomes. By not taking parameter discontinuity into account, one is in essence forecasting team strengths for the next match using incorrect parameter values.

We then carry out a forecasting exercise. This involves comparing the mean square error of the one-step ahead forecast of team strengths for all teams, using the two most recent seasons as the out-of-sample forecasting period. We find that different models have a smaller mean square error for different teams, but in particular two models stand out as the best ones: a simple random walk and forecasts made by model averaging. Even though the time-varying parameter model performs quite poorly according to the mean square error, it provides the best match predictions for one of our sub-samples. We conclude that different forecasting models that account for structural breaks can certainly improve forecast accuracy, although our findings are consistent with the econometrics literature that no one model forecasts best all the time. Given the prevalence of structural breaks in determining the dynamics of team strengths, this research has important implications for bookmakers and punters in the betting industry to take these matters into consideration when modelling football match outcomes.

1. Introduction

Football (association football or soccer) is the most popular sport in the world, with an estimated 3.5 billion fans and 250 million players worldwide (Sporty Desk, 2015). Betting on outcomes in football matches is also very popular, unsurprisingly, and the value of the football betting market in 2012 is estimated to be between £500 billion and £700 billion (Keogh and Rose, 2013). Consequently, statistical modelling of outcomes in football matches is popular among researchers, both in academia and industry, not only for the potential for financial returns but also for the challenges that such modelling presents. This is not to say that betting drives all research in statistical modelling in football and many interesting problems relating to tactical questions (e.g. Wright and Hirotsu, 2003; Hirotsu and Wright, 2006; Brillinger, 2007; Tenga et al., 2010; Titman et al., 2015), team, player and manager rating (e.g. Knorr-Held, 2000; Bruinshoofd and Weel, 2003; Schryver and Eisinga, 2011; Baker and McHale, 2015), competitive balance and outcome uncertainty (e.g. Koning, 2000; Buraimo and Simmons, 2015), match importance (e.g. Scarf and Shi, 2008; Goossens et al, 2012), tournament outcome prediction (e.g. Koning et al., 2002; Groll et al., 2015), and tournament design and scheduling (e.g. Scarf et. al, 2009; Scarf and Yusof, 2011; Goossens and Spieksma, 2011; Lenten et al., 2013) have been studied. Nonetheless, modelling results and scores, and other in-match outcomes, both straightforward (e.g. first player to score) and unusual (e.g. number of player cautions), motivated by the search for betting market inefficiency, has been a major motivational factor in the development of state-of-the-art models.

Statistical models have important implications for both sides of the market. On one hand, bookmakers use them to help set their odds competitively, so that they are both interesting to punters and profitable for the bookmaker. On the other hand, punters may use them to seek opportunities for profit. Given this importance, one could argue that research in modelling football match outcomes is to bookmakers and punters as financial mathematics and macroeconometrics is to stock markets.

Following the book (later turned movie) *Moneyball* (Lewis, 2003) that discusses the use of statistical analysis by Oakland Athletics MLB team for player valuation, which allowed Athletics to be competitive with teams with much greater payrolls, football clubs have begun to use statistical modelling, particularly in player analytics, and several top teams around the world now have an analytics team. There are however, significant challenges posed by the continuous nature of the "beautiful game", the significant interactions between players, and the

multi-faceted and dynamic roles that individual players assume as they take on more or less attacking and defensive responsibilities (McHale et al., 2012). This is an important strand of modelling research in football but it is not considered in this PhD thesis.

Knorr-Held (2000) and Rue and Salvesen (2000) were the firsts to investigate team strength dynamics in football. Between then and Koopman and Lit (2015), team strength dynamics has not received much attention from researchers. This thesis offers an opportunity to revisit the topic of dynamics in team strengths by investigating the presence of structural breaks in these dynamic models, something that has not been the centre of attention in the sports forecasting literature but is a fiercely researched and debated topic in financial and macroeconometrics.

Contribution

The main contribution of this thesis is that we demonstrate structural breaks are present in the majority of time series of team strengths. Therefore it has important implications for forecasting, as demonstrated in Clements and Hendry (1998). If we consider that a structural break has occurred, the dynamics governing team strengths will have changed. If one wishes to forecast the team strength one-step ahead (i.e. for the next match), one would be in essence forecasting one step ahead with the wrong parameters and consequently, this will lead to poor predictions of match outcomes. In practice, this is a problem that has important implications for bookmakers in setting their odds accordingly as well as punters trying to obtain financial returns from their bets.

This contribution opens up to two significant questions: i) given a break exists in the time series, is there a way to predict where the next break will occur out-of-sample and ii) how can we predict what the next parameters will be?

This thesis does not provide the answers for those questions. In fact, no research in time series forecasting has the definite answer to them: different models and forecasting techniques provide different results when applied to a variety of datasets.

We believe that the contributions of this thesis opens up opportunities for researchers to further develop models in this area. To date, this kind of modelling has been confined within a macroeconomic or a financial econometric context. There is thus significant opportunity to apply it not only football but in any other sport in which dynamics are involved.

Outline

This thesis is structured as follows: Chapter 2 describes the literature review in terms of modelling match outcomes, both direct and indirect, with more focus on the latter. Team strength dynamics are explained as well as the implications of structural breaks for forecasting. Chapter 3 describes the methodology used to obtain parameter estimates for team strengths over time, on which subsequently a time series analysis is carried out. This is followed by the methodology used to detect structural breaks and the results are presented. Chapter 4 presents a forecasting exercise of five different models used: three of which address structural breaks in a simple fashion, and two which do not. The forecasting performance of the models is compared by using the mean square error. Those forecasts are then utilised in deriving match outcome probabilities and the ranked probability score is used to evaluate which model forecasts match outcomes the best. Chapter 4 concludes with a discussion about the problems associated with structural break modelling, forecasting techniques and potential extensions for future research. Chapter 5 concludes this thesis.

Publications

There are two publications that have resulted from this thesis:

1) Rangel Jr (2015) – presented at the 5th International Conference on Mathematics in Sports, which is not discussed in this thesis.
2) Scarf and Rangel Jr (2017) – book chapter published in "The Handbook of Statistical Methods and Analyses in Sports. Most of the literature review of this thesis has contributed to the publication.

2. Literature Review

Match outcome models for football can be split into two classes: direct and indirect models. The former directly predict the outcome of a match, as a win, draw or loss, and ordered probit and logit models are generally used (Koning, 2000; Goddard, 2005). The latter indirectly model the bivariate score (goals-for and goals-against), to which a bivariate probability distribution is attributed. Given that the number of goals is classified as a non-negative integer, count data models such as the Poisson or negative binomial distributions are implemented and the seminal papers are Maher (1982) and Dixon and Coles (1997). Other approaches have also been investigated such as models for goal differences through the Skellam distribution (Karlis and Ntzoufras, 2009), models which predict shots for and against (McHale and Scarf, 2007) and ratings model such as ELO (Hvattum and Arntzen, 2010) and pi-rating models (Constantinou and Fenton, 2013). This chapter will describe direct and indirect models, with emphasis on the latter, especially because these are both richer in both the modelling questions that they pose and the applications they underpin.

It appears to be the case that indirect models have taken over from direct, not least because the most recent paper published in the statistical literature that uses the direct approach is Goddard (2005). This shift in interest is probably as a result of the many challenges associated with indirect models that open the way for further research. These challenges include: the issue of dependence within the bivariate distribution of goals scored in a match; the modelling of the dynamics of attacking and defensive strengths; the contribution of players' strengths to team strengths; count data model selection. One of the main challenges ahead of the literature lies in modelling player and team strength dynamics. The pioneering work here is Knorr-Held (2000) and Rue and Salvesen (2000). The way this literature review is structured reflects this view: the direct models are presented followed by indirect models. Then the literature on team strength dynamics is described, followed by the implications of structural breaks for forecasting.

Direct match outcome models

Direct measurement of the result as win, draw, or loss (from the point of view of the reference team) leads to the class of generalised linear models suitable for a categorical response variable. To model a dependent variable of this type, it is common to use an ordered

response model with an appropriate link function, which links the probabilities for each category to the linear combination of model parameters and covariates. For an ordered probit, the probit link function is used, and for ordered logistic regression the logit link function is used. The idea is that a latent continuous variable exists (but is not observed) which underlies the ordered responses (which are observed) and cut-points divide the real line into a series of intervals corresponding to the categories of the ordered response. For football match outcomes, we have three categories (win, draw or loss) and two cut-points. Thus, for the probit model let y^* be the latent variable that is some linear combination of explanatory variables X and their coefficients β plus error, denoted

$$y^* = X'\beta + \varepsilon, \tag{1}$$

where $\varepsilon \sim N(0,1)$. The latent variable y^* takes values Y as follows

$$Y = \begin{cases} 1\ (\text{win}) & \text{if } \delta_1 \le y^* < +\infty \\ 0\ (\text{draw}) & \text{if } \delta_{-1} \le y^* < \delta_1 \\ -1\ (\text{loss}) & \text{if } -\infty < y^* < \delta_{-1} \end{cases} \tag{2}$$

where δ_{-1} and δ_1 are the cut-points. Thus Y has a trinomial distribution with probabilities p_{-1}, p_0 and p_1 ($p_{-1} + p_0 + p_1 = 1$) obtained by substituting (1) into (2):

$$\begin{aligned} p_{-1} = \Pr(Y = -1) = \Pr(\text{loss}) &= \Pr(-\infty < X'\beta + \varepsilon < \delta_{-1}) = \Pr(X'\beta + \varepsilon < \delta_{-1}) \\ &= \Pr(\varepsilon < \delta_{-1} - X'\beta) = \Phi(\delta_{-1} - X'\beta) \end{aligned}$$

$$\begin{aligned} p_0 = \Pr(Y = 0) = \Pr(\text{draw}) &= \Pr(\delta_{-1} \le X'\beta + \varepsilon < \delta_1) \\ &= \Pr(\delta_{-1} - X'\beta \le \varepsilon < \delta_1 - X'\beta) = \Phi(\delta_1 - X'\beta) - \Phi(\delta_{-1} - X'\beta) \end{aligned}$$

$$p_1 = \Pr(Y = 1) = \Pr(\text{win}) = \Phi(+\infty - X'\beta) - \Phi(\delta_1 - X'\beta) = 1 - \Phi(\delta_1 - X'\beta),$$

where $\Phi(.)$ is the cumulative distribution function (CDF) of the standard normal distribution. The parameters of the model are typically estimated by maximising the log-likelihood function. The vast majority of econometric and statistical software packages have in-built functions for this purpose.

In the ordered logistic (or logit) model, often called the proportional odds model, the log odds of the ordered outcomes are assumed to be linearly related to the covariates and parameters:

$$\log\left\{\frac{\Pr(\text{win})}{1 - \Pr(\text{win})}\right\} = \log\left\{\frac{p_1}{1 - p_1}\right\} = \alpha_1 + X'\beta \tag{3}$$

$$\log\left\{\frac{\Pr\,(\text{win or draw})}{1-\Pr\,(\text{win or draw})}\right\}=\log\left\{\frac{p_1+p_0}{1-p_1-p_0}\right\}=\alpha_0+X'\beta \qquad (4)$$

Thus

$$\Pr(\text{win})=\ \text{logit}^{-1}(\alpha_1+X'\beta),$$

$$\Pr(\text{win or draw})=\ \text{logit}^{-1}(\alpha_0+X'\beta),$$

where $\text{logit}^{-1}\colon \Re \to [0,1]$ such that $\text{logit}^{-1}(x)=\frac{\exp(x)}{1+\exp\,(x)}$. Although this model is very similar to the ordered probit model, it is not generally used for modelling match outcomes in football. Nonetheless, their model-fit diagnostics are often very similar to the ordered probit.

The motivations for the use of direct models vary. Forrest and Simmons (2000a), for example, compare tipster performance in predicting match outcomes, whereas Forrest and Simmons (2000b) compare Pools Panel football match decisions with actual results. Audas et al. (2002) analyse the impact of within-season managerial change on team performance in English football. Forrest and Simmons (2002) study the impact of competitive balance on football attendance. Koning (2000) estimates team strengths and home advantage for the purpose of analysing competitive balance in Dutch football. Dobson and Goddard (2003) study short-term persistence in sequences of match results to conclude that negative persistence effects exist, implying that series of consecutive wins and consecutive losses tend to end sooner than expected if there were no persistence. Goddard and Asimakopoulos (2004) examine the predictability of English league football results, and utilise several explanatory variables including: match significance for championship, promotion or relegation issues and geographical distance between home and away teams. They test for weak-form inefficiency in the betting market and conclude that their forecasting model contains additional information not included in the bookmakers' odds. Finally, Goddard (2005) compares the forecasting performance of four different direct models using a pseudo-likelihood statistic and concludes that the best performing model is a "hybrid" model – an ordered probit regression with results-based independent variables and goals-based lagged performance covariates. This model outperforms the other models in 6 out of 10 seasons. Nevertheless, they did not implement the model to produce *ex ante* forecasts or test the percentage return on bookmakers' odds, such as Goddard and Asimakopoulos (2004), which would have been more informative.

Regarding the interpretation of the parameter estimates, since the probit and logistic models both apply a non-linear link function, the coefficients in β do not correspond with the linear interpretation as in an ordinary least squares (OLS) regression. One has to compute marginal effects using

$$\frac{\partial E[y^*|x]}{\partial x} = \left\{\frac{dF(x'\beta)}{d(x'\beta)}\right\}\beta = f(x'\beta)\beta \quad (5)$$

For the probit model, $f(x'\beta)\beta = \phi(x'\beta)\beta$, so that a marginal change in one of the explanatory variables depends on the values that the explanatory variable takes.

Research into direct outcome models for forecasting football matches have, since Goddard (2005), not really developed. This is because indirect match outcomes have presented larger modelling challenges. Additionally, modelling exact scores and goal differences has wider applications in betting, not least to spread betting.

Indirect outcome models

Moroney (1956) observed that football scores were fitted better by a negative binomial distribution than a Poisson distribution. Nonetheless, Maher (1982) proposed a double-Poisson model that is considered the backbone for all indirect outcome models for modelling football scores. In this seminal paper, Maher estimates the attack and defence parameters of teams and a home advantage parameter. The model is as follows: consider team *i* playing at home against team j and let X_{ij} and Y_{ij} denote the goals scored by the home and away teams respectively. Then

$$X_{ij} \sim Poisson(\alpha_i\beta_j\gamma) \quad (6)$$

$$Y_{ij} \sim Poisson(\alpha_j\beta_i) \quad (7)$$

with X_{ij} and Y_{ij} independent. In the first Poisson mean in (6) the parameter α_i is interpreted as the attacking strength of the home team *i* (tendency to score goals), β_j is the defensive weakness of the away team *j* (tendency to concede goals), and γ is the home advantage effect. These three parameters determine the goal-scoring rate of the home team (X_{ij}). The parameters that determine the goal-scoring rate of the away team (Y_{ij}) are α_j, the attack strength of the away team j and β_i, the defensive weakness of the home team i. In the simplest variation of this model, the home advantage effect is the same for all teams. Maher used the method of maximum

likelihood to estimate the parameters, and noted that the model underestimated the number of draws. This motivated Dixon and Coles (1997) to enhance the model by adding a dependence parameter:

$$P(X_{ij} = x, Y_{ij} = y) = \tau_{(\lambda,\mu)}(x, y)\frac{\lambda^x \exp(-\lambda)}{x!}\frac{\mu^y \exp(-\mu)}{y!} \quad (8)$$

where

$$\ln \lambda = \alpha_i + \beta_j + \gamma \quad (9)$$

$$\ln \mu = \alpha_j + \beta_i \quad (10)$$

and

$$\tau_{(\lambda,\mu)}(x, y) = \begin{cases} 1 - \lambda\mu\rho & \text{if } x = y = 0 \\ 1 + \lambda\rho & \text{if } x = 0, y = 1 \\ 1 + \mu\rho & \text{if } x = 1, y = 0 \\ 1 - \rho & \text{if } x = y = 1 \\ 1 & \text{otherwise} \end{cases} \quad (11)$$

so that ρ is the dependence parameter that modifies the independence assumption for low scoring matches ($x \leq 1, y \leq 1$).

With N teams, there are $2N + 2$ parameters to estimate in total. The model is overparameterised because, under the specification, strength can only be measured relatively (that is, one can only know for example whether A is stronger than B and one cannot know how strong is A). Therefore the parameter space must be constrained, and Dixon and Coles (1997) use the constraint

$$N^{-1}\sum_{i=1}^{N} \alpha_i = 1 \quad (12)$$

The authors acknowledged that the strength parameters are only locally constant, developing a pseudo-likelihood function for estimation in which the outcomes of more recent matches are given greater weight. Thus

$$L(\alpha_i, \beta_i, \rho, \gamma : i = 1, \dots, N)$$

$$= \prod_{k=1}^{J} \left[\tau_{(\lambda,\mu)}(x_k, y_k) \frac{\exp(-\lambda)\lambda_k^{x_k}}{x!} \frac{\exp(-\mu)\mu_k^{y_k}}{y!} \right]^{\phi(t-t_k)} \tag{13}$$

where (x_k, y_k) is the result of match k played at time $t_k < t$, J is the number of matches in the sample, and ϕ is a non-increasing function of time. In fact, $\phi(t) = \exp(-\xi t)$, where ξ was chosen to maximise the prediction of match outcomes and $\hat{\xi} = 0.0065$. A large ξ puts large weight on recent matches and $\xi = 0$ is the static model, with all matches given equal weight.

Dixon and Robinson (1998) developed a model in which scoring rates depend on the current match state, and proposed a bivariate non-homogeneous Poisson process for goals scored, with the intensities for home and away goals varying linearly (at rates ξ_1 and ξ_2 for home and away teams respectively) with match time t, and strength parameters contingent on the current score (x, y):

$$\begin{aligned} \lambda_i(x, y, t) &= \alpha_i(x, y)\beta_j(x, y)\gamma + \xi_1 t \\ \mu_j(x, y, t) &= \alpha_j(x, y)\beta_i(x, y) + \xi_2 t \end{aligned} \tag{14}$$

They conclude that goal-scoring rates increase for both teams throughout a match and when a goal is scored by the opposition. This model however is really only useful for in-play betting, but they demonstrated an improvement in match outcome prediction compared to the models of Maher (1982) and Dixon and Coles (1997).

The next interesting development is due to Karlis and Ntzoufras (2003). They describe a bivariate Poisson distribution that captures positive dependence in home and away goals. Consider the (pairwise) independent Poisson random variables $Z_k, k = 1, 2, 3$ with means $\lambda_1, \lambda_2, \lambda_3$ and form the bivariate pairs $X = Z_1 + Z_3$ and $Y = Z_2 + Z_3$. Then

$$P_{X,Y}(x, y) = P(X = x, Y = y)$$

$$= \exp\{-(\lambda_1 + \lambda_2 + \lambda_3)\} \sum_{k=0}^{\min(x,y)} \frac{\lambda_1^{x-k}\lambda_2^{y-k}\lambda_3^{k}}{(x-k)!\,(y-k)!\,k!} \tag{15}$$

where

$$E(X) = \lambda_1 + \lambda_3$$
$$E(Y) = \lambda_2 + \lambda_3 \quad (16)$$
$$\text{cov}(X, Y) = \lambda_3$$

Karlis and Ntzoufras (2003) also "diagonally inflate" to increase the probability of observing draws. This is specified by

$$P_D(x, y) = \begin{cases} (1-p)BP(x, y|\lambda_1, \lambda_2, \lambda_3) & \text{if } x \neq y \\ (1-p)BP(x, y|\lambda_1, \lambda_2, \lambda_3) + pD(x, \theta) & \text{if } x = y \end{cases} \quad (17)$$

where $D(x, \theta)$ is a discrete distribution with parameter vector θ. They recommend to use a simple discrete distribution for *D.* The Poisson marginals then become

$$E(X) = (1-p)(\lambda_{1i} + \lambda_{3i}) + p\theta_1$$
$$E(Y) = (1-p)(\lambda_{2i} + \lambda_{3i}) + p\theta_1 \quad (18)$$

where θ_1 is the parameter in the Bernoulli distribution and

$$\lambda_{1i} = \mu + \text{home} + \text{att}_{hi} + \text{def}_{gi}$$
$$\lambda_{2i} = \mu + \text{att}_{gi} + \text{def}_{hi} \quad (19)$$
$$\lambda_{3i} = \beta^{con} + \gamma_1 \beta_{hi}^{home} + \gamma_2 \beta_{gi}^{away}$$

where λ_{1i} (home score rate) depends on a constant μ, the home advantage parameter home, attack parameter of the home team, att_{hi}, and defence parameter of the away team, def_{gi}; λ_{2i} (away score rate) depends on constant μ, away attack parameter, att_{gi} and home defence parameter, def_{hi}; and λ_{3i} (covariance parameter) depends on a constant term β^{con}, γ_1 and γ_2 are dummy variables, and β_{hi}^{home} and β_{gi}^{away} are parameters that depend on the home and away team respectively.

Karlis and Ntzoufras (2003) applied the diagonally inflated bivariate Poisson model to the Italian Serie A league to capture the underpredicted number of draws observed in the dataset, particularly the 1 – 1 scores. A possible explanation for the tendency for many draws in this league (Italian Serie A, 1991-1992 season) is that the scoring system of 2-1-0 for wins-draws-losses was still in place. They compared the fit between the double-Poisson, the bivariate Poisson and the diagonally inflated models for a variety of diagonal distributions. The best

fitting model according to the Akaike Information Criterion (AIC), Bayes Information Criterion (BIC) and Likelihood Ratio Test (LRT) was the bivariate Poisson model with an extra parameter for 1 – 1 draws.

An issue of concern regarding these bivariate models is that they only permit non-negative correlation. Since λ_3 is a Poisson mean, it cannot be negative. A solution was offered by McHale and Scarf (2011) who used particular copula functions that allow negative dependence. In application to international football, these authors found that goals scored in a match (X_{it}, Y_{jt}) are negatively dependent. In general, a copula provides a flexible means of joining marginal distributions to form multivariate distributions with interesting dependence structures, although some care is required with discrete distributions. The copula function itself is a multivariate distribution with all univariate marginal distributions as $U(0,1)$. Hence C is the distribution of a multivariate uniform random vector. For a bivariate distribution F with margins F_1 and F_2, the copula associated with F is a distribution function $C: [0,1]^2 \rightarrow [0,1]$ that satisfies

$$F(x,y) = C\{F_1(x), F_2(y)\}, (x,y) \in \Re^2 \qquad (20)$$

McHale and Scarf (2011) make use of extendable Archimedean copulas that can model both positive and negative dependence. They fit three different copulas and showed that Frank's copula provides the best fit to the international football data set, in which

$$C(u,v) = -\kappa^{-1} \log\{1 - (1 - e^{-\kappa u})(1 - e^{-\kappa v})/(1 - e^{\kappa})\}, (\kappa \in \Re) \qquad (21)$$

where u and v are specified with marginal distributions to give the full joint distribution. In the case of McHale and Scarf (2001) they experimented with both Poisson and negative binomial marginal distributions to model football match results. While copulas can capture general dependence structure, often at the expense of estimating only one additional parameter, it is unclear whether their use in model specification is preferred to the direct dependence models of Dixon and Coles (1997) and Karlis and Ntzoufras (2003).

What is clear however is that one of the most investigated topics in the literature is strength dynamics, and although it is some decades since Dixon and Coles (1997) acknowledged strength dynamics but did not really model them (instead they proposed rolling, fitting and forecasting approach in which recent matches were given more weight in strength estimation), the important contributions to this topic have been much more recent.

Team strength dynamics

Team strength dynamics can be modelled either deterministically or stochastically. Focusing first on the former, Baker and McHale (2015) use a smoothly varying function (barycentric rational interpolation) to specify attack and defence strength for all time, and while their model can be used for short term forecasting, their motivation is backward looking, focusing on the identification of the strongest team. In the aforementioned paper they discuss in particular the historically strongest team. These authors have had success with their approach in identifying the all time best not only in football but also in tennis and golf (Baker and McHale, 2014). The advantage of the deterministic approach is that the likelihood function is both easier to specify and to maximise.

Modelling team strength dynamics stochastically has received considerably more attention in the literature and Knorr-Held (2000) and Rue and Salvesen (2000) and later Crowder et al. (2002) pioneered this research. Several different stochastic processes have been used to model dynamics in team strengths, and these models can estimated using either classical or Bayesian methodologies. A natural place to start when modelling dynamics stochastically is by some type of autoregressive process, where a team's strength at time *t* is related to its strength at some time $t - s, s > 0$. Knorr-Held (2000) considered that time-dependent abilities α_i followed a random walk $\alpha_{i,t} \sim N\left(\alpha_{i,t-1}, \sigma^2\right)$, where σ^2 was estimated in order to maximise the in-sample predictability of the model with respect to observed final team rankings. The model was implemented on the 1996-1997 season of the German Bundesliga. As no forecasts were made and no comparison with a benchmark model carried out, there is no indication as to how well the model can forecast results.

In Rue and Salvesen (2000), strength parameters for attack and defence follow a Brownian motion, a stochastic process in continuous time. These authors also include in the Poisson marginal mean, as a new development, a covariate that they call the psychological effect (denoted γ) of the superior team underestimating the strength of the weaker one. Then, letting t' and $t'' > t'$ denote two different points in time, the attack strength dynamics are such that

$$\alpha_i^{t''} - \alpha_i^{t'} = \left\{B_{\alpha,i}\left(\frac{t''}{\tau}\right) - B_{\alpha,i}\left(\frac{t'}{\tau}\right)\right\}\frac{\sigma_{\alpha,i}}{\sqrt{\left\{1-\gamma(1-\frac{\gamma}{2})\right\}}} \qquad (22)$$

where $B_{\cdot,\cdot}(t), t \geq 0$ is a standard Brownian motion, the subscript marks denoting attack (subscript β for defence) and team. The time parameter τ is a scaling factor identical for all teams and specifies the inverse loss of memory rate for α_i^t, var $(\alpha_i^{t''} - \alpha_i^{t'}) \propto \sigma_{\alpha,i}^2/\tau$. The model was estimated through Bayesian Markov chain Monte Carlo (MCMC) methods. The model used data from the first half of the 1997-1998 Premier League and Division 1 English league and was shown to perform as well as the bookmaker in a simulated betting experiment. In this experiment, they placed bets to maximise a particular utility function: betting on matches which would give a positive expected profit, placing the bets with low variance of profit. They obtained a final return of 39.6%, winning 15 bets out of 48 placed in the English Premier League and 54.0% in Division 1 (27 out of 64). However, the lower bounds of the 95% confidence interval of the betting returns were negative, indicating that there was still a risk of losing money. They also showed that combination bets gave lower returns.

There then followed something of a lull in developments until Owen (2011) presented a dynamic generalised linear modelling (DGLM) framework that allows some or all parameters to be time dependent. He identified the challenge of estimating the "evolution" variance σ^2 (a volatility parameter in the strength dynamics whereby a higher evolution variance implies more volatile strength dynamics). He estimates this parameter by maximising the one-step ahead predictive probability of the model:

$$P_1 = \exp\left\{\frac{1}{N}\sum_{k=1}^{N} \log e[P(O_k)]\right\} \tag{23}$$

which is equivalent to the geometric mean of the one-step ahead match predictive probabilities that were actually observed, $P(O_k)$ being the one-match ahead predictive probability that match k would end with the outcome O_k (home win, draw or away win). However, Owen made use of a cumulative measure of P_1 : $P_1(t)$ which included all matches played up to and including round *t*, and as a result, allowed the parameter σ^2 to be updated as more information becomes available. He also used another measure of predictive performance, P_2, a quadratic loss function or a measurement error, as well as its cumulative counterpart $P_2(t)$, which incidentally was also used in Knorr-Held (2000). Owen used three seasons of the Scottish Premier League (2003-2004 to 2005-2006) to consider the efficacy of his model and concluded that the dynamic model provided a better fit to these results than a non-dynamic one. Owen's approach has significant advantages over the continuous-time model of Rue and Salvesen (2000), particularly

in the choices of priors for the attack and defence parameters, which are derived from previous seasons' data. Estimating σ^2 directly in the model also allows greater flexibility, permitting the strength parameters to vary at different times in the season – there appeared to be higher volatility in the strength parameters earlier in the season than later, which seems reasonable. However, he does not apply a betting strategy to investigate possible returns of the model.

Koopman and Lit (2015) use a stationary *AR(1)* process for team strength parameters, including an intercept term, in the form $z_t = \mu + \Phi z_{t-1} + \eta_t$, where z_t are state vectors of the α_{it} and β_{it} elements (attack and defence strengths for team *i* at time *t*), μ is a vector of constants, Φ is a square matrix with the autoregressive coefficients in the diagonal and η_t are the error vectors which are normally independently distributed with mean vector 0 and variance matrix H. Nine seasons of the English Premier League were used in the analysis (2003-2004 to 2011-2012), with the first seven seasons used to provide out-of-sample forecasts for the last two seasons. The authors also implement a betting strategy, betting on matches where the expected value of a unit bet exceeds some benchmark, denoted by $\tau > 0$. They use average odds across 40 different bookmakers, giving a total of 760 betting opportunities (all matches in 2 seasons) for $\tau = 0$. As the value of τ increases, the number of betting opportunities decreases and positive mean returns are obtained when $\tau > 0.12$ (around 270 "value" bets). When $\tau = 0.4$, one unit bet in each of 50 matches and generated a return of 75 units (an expected profit of 50% on the stake).

The autoregressive parameters in Koopman and Lit (2015) were estimated to be $\hat{\phi}_\alpha = 0.9985$ for the attack dynamics, and $\hat{\phi}_\beta = 0.9992$, which are both very close to one so that the strength dynamics, although their model constrained the dynamics to ensure stationarity by setting $0 < \phi_K < 1, K = \alpha, \beta$.

The introduction of dynamics in the modelling of attack and defence parameters has added a layer of depth in football match outcome models that has improved forecasts overall. However, authors have made rather large assumptions as to how attack and defence of teams evolve, and we believe that the questioning of these assumptions is the next step for research in football match outcome modelling. Knorr-Held (2000), Rue and Salvesen (2000) and Owen (2011) all use a type of non-stationary processes to capture dynamics, whereas Koopman and Lit (2015) implement a stationary autoregressive model. Certainly there are advantages to each, however, a non-stationary and a stationary process has different properties, which as a result affect forecasting performance if the model is mis-specified. Below is a brief description of

stationary processes as well as two different non-stationary processes, and subsequently an explanation of how mis-specifying the dynamics could lead to misleading forecasts.

A (covariance) stationary process is one with the following properties:

i) $E(y_t) = \mu,\ -\infty < \mu < \infty \ \ \forall t$

ii) $V(y_t) = \sigma^2 < \infty \ \ \forall t$

iii) $C(y_t, y_{t-s}) = \ \gamma(s) \ \ \forall\, t, s$ (the autocovariance function)

where the mean, variance and covariance do not depend on the time *t*. If we observe team strengths over time in several papers (e.g. Dixon and Coles, 1997; Knorr-Held, 2000; Koopman and Lit, 2015), it appears that the time series plots of strengths do not exhibit the properties attributed to a stationary process. Two classic examples of nonstationary time series that violate at least one of the properties above are trend-stationary and difference-stationary processes.

A simple trend stationary process that could be applied to modelling attack strength dynamics could be formulated as:

$$\alpha_t = \alpha_0 + dt + \varepsilon_t \tag{24}$$

where dt is the trend component of the time series. The mean is time-dependent, as $E(\alpha_t) = \alpha_0 + dt$. This series can be made stationary by detrending. A simple example of a difference-stationary series is a random walk, (used in Knorr-Held (2000) and Owen (2011) where $\alpha_t = \alpha_{t-1} + \varepsilon_t$), for which the second and third criteria are violated: $V(\alpha_t) = t\sigma^2$ and $C(\alpha_t, \alpha_{t-s}) = (t-s)\sigma^2$, so that both the variance and autocovariance functions depend on time *t*. These types of time series processes are often called *I(1)* processes (integrated processes of order 1), because one can achieve stationarity by taking first differences once. Sometimes a time series can have a trend and also be a difference-stationary process, so that it is said to have a stochastic trend. Despite the fact that trend-stationary and difference-stationary processes have very different properties, realisations and diagnostics can look rather similar. The implications for forecasting are that the one step ahead forecast $\hat{\alpha}_{t+1}$ is different for all three processes:

i) For a trend stationary process, $\hat{\alpha}_{T+1|T} = \alpha_0 + d(T+1)$

ii) For a random walk, $\hat{\alpha}_{T+1|T} = \ \alpha_T$

iii) For a stochastic trend, $\hat{\alpha}_{T+1|T} = \alpha_0 + \alpha_T + d(T+1)$

Diebold and Kilian (2000) reinforce this point: that it is very important to decide which dynamic model to use because different models imply different predictions. They recommend pre-testing for a unit root of the autoregressive parameter. There are several tests that can be used to test for the presence of a unit root, most notably the Augmented Dickey-Fuller (ADF) test (Said and Dickey, 1984) and the DF-GLS ERS test (Elliot et al., 1996). For a comprehensive review of these tests and test alternatives, refer to Maddala (2001).

Structural breaks and forecasting

A further issue in modelling dynamics is the possible presence of structural breaks[1]. This occurs when the assumption of parameter "continuity" fails. A break occurs in the trend stationary process (24) when there is a discontinuity in either α_0 or d or both. Such discontinuities may arise naturally in football when teams buy and sell players during transfer windows, and when management and team ownership changes. Clements and Hendry (1998) provide Monte Carlo evidence for *ex-ante* forecast failure in the presence of structural breaks in time series. This is quite common in macroeconomic and financial time series, but it has never been considered when modelling team strength dynamics in football. This adds a further challenge to modelling match outcomes. The presence of structural breaks has important implications for time series analysis. Perron (1989) shows that traditional unit root tests lose power when a structural break is present in the time series. Furthermore, Pesaran and Timmermann (2004) show that ignoring these breaks leads to inconsistent and biased forecasts.

Testing for breakpoints is a huge topic in the econometrics literature that has had incredible developments. From the pioneering work of Chow (1960), which tests for a known single breakpoint, this topic has attracted a considerable amount of attention for econometricians. Zivot and Andrews (1992) provide a test for a unit root null hypothesis against the alternative of a trend stationary process with one endogenous breakpoint (determined by the data itself). Since then, tests for multiple breakpoints have been developed, such as Bai and Perron (1998, 2003). A more comprehensive review of structural change detection can be found in Perron (2006).

[1] This can also be known as structural change or change-point analysis.

Even if these types of structural break tests are powerful in detecting breakpoints and their confidence intervals, they are problematic for forecasting because as new data becomes available, the detection of a new break is incredibly difficult and can take several observations. Other tests such as Zeileis et al. (2005) deal with this detection delay better, but on the other hand struggle to detect smaller breaks.

Structural break testing has not been considered in the sports literature. In fact, one of its only applications is from Palacios-Huerta (2004) as an empirical test of whether exogenous factors such as political and economic events as well as changes in the rules affect the outcomes of matches (in particular the average and the variability of goals per match), but with no application to forecasting.

Solutions to forecasting under parameter instability are generally split into two categories. The first involves modelling the changes parametrically. This means defining a functional form to model the data generating process (DGP): both in terms of the break process itself and how the new parameters post-break are estimated. Two pioneering models include the Rosenberg (1973) and Hamilton (1989) models. Rosenberg (1973) specifies that the parameters change at every period under a mean-reversing mechanism where the speed of reversion is estimated. The Markov-switching model in Hamilton (1989) allows for less frequent changes and the parameters to switch between states. In one way or another, the models in the literature are variations and extensions of them.

Some sophisticated models seem to have had some relative success in forecasting macroeconomic series using Bayesian methodologies. Some examples include Pesaran et al. (2006) (henceforth, PPT) , Koop and Potter (2007) (KP) and Giordani and Kohn (2008) (GK). These models use data to learn where the breaks have occurred in order to sample new parameters. The manner in which the new parameters are sampled is the main difference between these models. Suppose a break has occurred and the parameters, denoted in a vector β_j, have changed to β_{j+1}. PPT specifies the form $\beta_j = \beta_0 + u_j$, $u_j \sim$ i.i.d $N_m(0, B_0)$, which essentially draws β_{j+1} from the conditional mean. PPT also varies the prior and allows a random walk evolution of the coefficients $\beta_j = \beta_{j-1} + u_j$ thereby directly linking the next regime with the current one. The KP model follows the same random walk prior specification for drawing the next set of post-break coefficients. In fact, this is inspired by the literature of time-varying parameter (TVP) models (Stock and Watson, 2007) which specifies that the coefficients β_j change at every observation, so $\beta_j = \beta_{j-1} + u_j$ is in fact $\beta_t = \beta_{t-1} + u_t$. The

break process in PPT is modelled using a Markov process developed in Chib (1998), whereas KP models the duration between breaks as a Poisson distribution.

Giordani and Kohn (2008) specify a dynamic mixture approach. If we have a similar specification of the state transition equation $\beta_t = \beta_{t-1} + u_t$, the GK model includes a Bernoulli random variable k, where $k = 1$ with probability p and $k = 0$ with probability $1 - p$. Thus the evolution of the coefficients becomes $\beta_t = \beta_{t-1} + ku_t$. Rather than sampling the states using the Chib (1998) algorithm, GK rely on the efficient methods of Gerlach et al. (2000) to estimate k.

Xu and Perron (2017) develop a dynamic mixture model, but estimate it in a frequentist framework through the use of a mixture Kalman filter and an Expectation-Maximisation (EM) algorithm. The mean-reverting mechanism developed in their paper seems to forecast the post-break parameters consistently well across different macroeconomic time series.

Bauwens et al. (2015) compares several sophisticated models in an extensive out-of-sample forecasting exercise on 60 different macroeconomic series. The conclusion is that no one model is consistently the best in the presence of structural breaks. This seems to indicate that there is still some research to be carried out on this topic.

Pesaran and Timmermann (2007) consider an additional problem: if breaks occur near the end of the sample, you have a small sample on which to train your new parameter estimates post-break. They recommend using some pre-break data to attempt to remedy the trade-off between bias and variance. This technique is shown to minimise the expected loss function of the forecast (like the mean square error). This awareness leads us to an alternative remedy: rather than modelling breaks parametrically, different forecasting techniques and methods that are robust to structural change could be implemented. These techniques bypass the "necessity" of modelling breaks explicitly. One of the most common techniques is that of model averaging, which pools forecasts from many different models by applying some type of weighting mechanism. Rossi (2013) finds that equally weighted average forecasts seem to be consistently good. Bayesian model averaging and alternative weighting schemes such as maximising the weighted log score (Geweke and Amisano, 2011) also work reasonably well. The choice of window size also has an impact on forecasting performance with the choice of either rolling or recursive windows being most popular. However, averaging across window sizes (Pesaran and Timmermann, 2007) seems to be an alternative solution to this problem of window size and

type selection. Rossi (2013) provides a thorough review of these forecasting techniques and their performances.

It would be interesting to see either of these solutions being applied to a topic outside of macro- or financial economics. A good starting point would be team strength dynamics in football.

3. Data Analysis and Results

Methodology

Match results are available from http://www.football-data.co.uk for the past 22 seasons of the top 4 divisions in UK football, although for the purpose of this thesis, only the last 9 seasons of the Premiership and Championship are used. The data are freely available on the website to download as a csv file.

The model used to obtain and evaluate the time series for team strengths was the Dixon and Coles (1997) model, which was explained in section 2 (equations 8 – 13). However, any other model specification could have been used to investigate dynamics such as Karlis and Ntzoufras (2003) and McHale and Scarf (2011). The software used for fitting the model was R (R Development Core Team, 2008), writing the pseudo-likelihood in equation 13 and maximising it subject to the constraint in equation 12[2].

The methodology is quite unorthodox. Given that team strengths are not observed, a state-space framework would usually be used to estimate the team strengths (state latent variables). However, we choose not to follow that approach for a few reasons. Firstly, the dynamics have to be specified and pre-imposed in the state transition equation. This would not be feasible if the number of lags of either the autoregressive or moving average parameters of the ARIMA process changes over time, allowing more flexibility in the time series analysis. Additionally, it is not feasible to estimate the time decay factor within a state space framework. In fact, after Dixon and Coles (1997), no article seems to consider this issue, particularly because of the preference for modelling in a state-space framework. We consider therefore that this methodology is really quantitatively no different than estimating a two-stage-least-squares (2SLS) or instrumental variables (IV) estimation, common in dealing with the endogeneity problem in econometrics.

The parameter ξ was estimated by minimising the Ranked Probability Score (RPS) as recommended in Constantinou and Fenton (2012) as well as maximising the predicted log-

[2] We modify the constraint of equation 12 slightly, so that the sum of the log of the attack parameters is equal to zero instead of one.

likelihood (PLL) in Dixon and Coles (1997). A higher PLL indicates that the actual match outcomes are more probable according to the model. The RPS is an alternative as it is a measure of the prediction error for the forecast of categorical ordered variables. The generic formula is

$$RPS = \frac{1}{r-1}\sum_{i=1}^{r-1}\left(\sum_{j=1}^{i}\left(p_j - e_j\right)\right)^2 \quad (25)$$

where r is the number of potential outcomes, p_j and e_j represents the difference between the cumulative distributions of forecasts with the observation. Applying this equation to a more concrete example: $p_j$could be the probability forecasts of home win and e_j is 1 if the outcome was a home win and 0 otherwise. This is done across all three outcomes of a football match (home win, draw and away win).

As the RPS approaches 0, it predicts the outcome better as there is a smaller difference between p_j and e_j. Therefore the implied benefit of using RPS is that it is also a measure of distance; given we observe a home win, a larger predicted probability of a draw is considered a smaller error than a higher predicted probability of an away win. The value for RPS is calculated for every match for a given value of ξ. In order to choose the optimal decay factor, the average RPS for the data set was calculated.

It is necessary to use at least 2 seasons of Premiership and Championship matches to obtain estimates of the team strengths. Team strengths are estimated for every round using a rolling window of width two seasons: seasons 2008-09 and 2009-10 are used to obtain the attack, defence and home advantage parameter estimates for $t = 1$, which is the time of the first playing round of season 2010-11. Then, we moved the estimation window by one playing round, meaning that matches of the 2nd round of season 2008-09 until 1st round of 2010-11 is used to estimate the same parameters at $t = 2$. This rolling window methodology is used until we have attack and defence parameter estimates for all teams, from $t = 1, \dots, 266$. There is the inconvenience that the Premier League does not run all 38 rounds of 10 games in "nice" blocks (for instance, some matches must be rearranged because of fixture clashes with other competitions) and the Championship actually has 552 matches per season because there are more teams (46 blocks of 12 games, but also not in "nice" blocks). To facilitate data manipulation, we ordered the data by date and before estimating the team strengths parameter,

we moved each window start and end by 25 matches in our sample[3]. In the end, we obtain a time series of team strengths of 266 observations[4] (until the end of the 2016-17 season)[5].

The dynamic programming algorithm developed in Bai and Perron (2003) was used to calculate the optimal breakpoints, using the "strucchange" package in R (Zeileis et al., 2003). This algorithm is used to test the hypotheses of 0 vs. ℓ and ℓ vs. $\ell + 1$, where ℓ is the number of breakpoints in the total time series.

The "breakpoints" function allows the user to set the maximum value that ℓ can take, which is calculated through a trimming parameter h that has a default value of $h = 0.15$. This is the minimal size (given as a fraction) relative to the sample size that each segment (before a breakpoint) can take. In this case, since we have a time series of 266 measures of team strengths, the minimum segment length before a break can happen is 39 observations.

The Bai and Perron (2003) algorithm is a development that reduces the computational intensity of Bai and Perron (1998). Both are concerned with minimising the residual sum of squares (RSS). Imagine we can split our time series (of team strengths for a specific team) from $t = 1, \dots, T$ $(T = 266)$ with m breakpoints into $m + 1$ segments. With $h = 0.15$ we have $m \leq 6$. In a general framework, the dynamics governing this time series can be written as:

$$y_t = x_t^T \beta_j + u_t$$

where y_t is the observation of the dependent variable at time t, x_t is the vector of regressors at time t (which can take form as trend, constant and lagged values of y_t), x_t^T being its transpose, and β_j the regression coefficients of segment index j and u_t the error term. The set of breakpoints can be written as $\mathcal{I}_{m,n} = \{t_1, \dots, t_m\}$, where t_1 and t_m denote the time of the first

---

[3] Across the Premiership and Championship season, there are a total of 552 + 380 = 932 matches. To obtain the team strengths parameter estimates of 38 games in a season, we would move the estimation window 37 times by 25 matches (925) games and then once by 7 matches.

[4] A time series of 266 observations is a result of 38 games for 7 seasons.

[5] Some teams only compete in the Championship and later Premiership after a certain number of rounds and therefore not all 266 team strength estimates are available. These are: Bournemouth, Brighton and Huddersfield. Southampton competed in League One during the estimation, so the missing parameter estimates were estimated through spline interpolation.

and last breakpoints respectively. Therefore, given a $m+1$ segment partition, the RSS for each segment can be obtained:

$$RSS(t_1, \dots, t_{m+1}) = \sum_{j=1}^{m+1} rss(t_{j-1}+1, t_j) \tag{26}$$

where $rss(t_{j-1}+1, t_j)$ is the minimum residual sum of squares of that jth segment. The goal is to find the breakpoints $(\widehat{t_1}, \dots, \widehat{t_{m+1}})$ that minimise the objective function over all possible segment partitions of the set of breakpoint combinations:

$$(\widehat{t_1}, \dots, \widehat{t_{m+1}}) = \arg \min_{t_1,\dots,t_m} [RSS(t_1, \dots, t_m)] \tag{27}$$

If we used a grid search to obtain the global minimizer of the RSS, we would require least squares operations of order $O(T^{m+1})$. Bai and Perron (2003) develop a dynamic programming algorithm based on the Bellman's principle that reduces the computational intensity to least squares operations of order $O(T^2)$, for any number of segment partitions. Thus the optimal segmentation is the one that satisfies the recursion

$$RSS(\mathcal{J}_{m,n}) = \min_{(m+1)hT<t<T-hT} RSS(\mathcal{J}_{m-1,t}) + rss(t+1, n) \tag{28}$$

More detail on this algorithm can be found in Bai and Perron (2003).

The "breakpoints" function of the "strucchange" package in R runs the Bai and Perron (2003) algorithm and we set the trimming parameter to the default $h = 0.15$ and therefore the maximum number of breaks $m = 16$. The "summary" function outputs shows the estimated parameters of the regression model, whereas the "coef" function using the object of the breakpoints function as a parameter outputs the different coefficients β_j for each section. The optimal breakpoints are selected according to minimum BIC partition.

The greatest advantages that these tests possess is that they are powerful in detecting structural breaks and they do so endogenously, without making any underlying assumptions about the data. They provide us with a great starting point in detecting parameter instability in-sample. The challenge with it is that it is not an appropriate tool to carry out-of-sample forecasts. This is because it is difficult to detect delays as new observations become available. Thus for us, it becomes a tool for in-sample analysis only.

This chapter is structured as follows: firstly the result of the decay factor estimation is presented, followed a description of the time series for the team strengths. Then, the estimated breakpoints for the attack and defence strengths are presented.

Estimation of time decay

Dixon and Coles (1997) estimated $\hat{\xi} = 0.0065$, although their measure of time was half weeks (3.5 days). However, we used days as a measure of time to estimate the decay factor, which would correspond to $\hat{\xi} = 0.0065 \div 3.5 = 0.0018$.

To estimate the decay factor, we used a dataset of three seasons (2014-15 to 2016-17), comprising of both the Premiership and the Championship. Dixon and Coles (1997) in contrast included the 3rd and 4th tiers of English football over 114 half-weeks (or just over one year). The average RPS indicates the minimum value (optimal) to be at $\hat{\xi} = 4 \times 10^{-4}$, whereas if we use the same predicted log-likelihood (PLL) as Dixon and Coles (1997), the optimal value would be $\hat{\xi} = 5 \times 10^{-4}$. This value is one order of magnitude lower than that of Dixon and Coles, which seems to indicate that ξ is susceptible to change over time, and/or the size of the dataset as well as the leagues included in the dataset. This should therefore be an area for future study. The plots of the average RPS and PLL against different values of ξ can be seen in Figure 1 and Figure 2 respectively.

For the rest of the analysis, we have opted to select $\xi = 4 \times 10^{-4}$, which is the optimal value according to the average RPS and this has been treated as fixed in order to estimate the team strength parameters back over the previous 6 seasons.

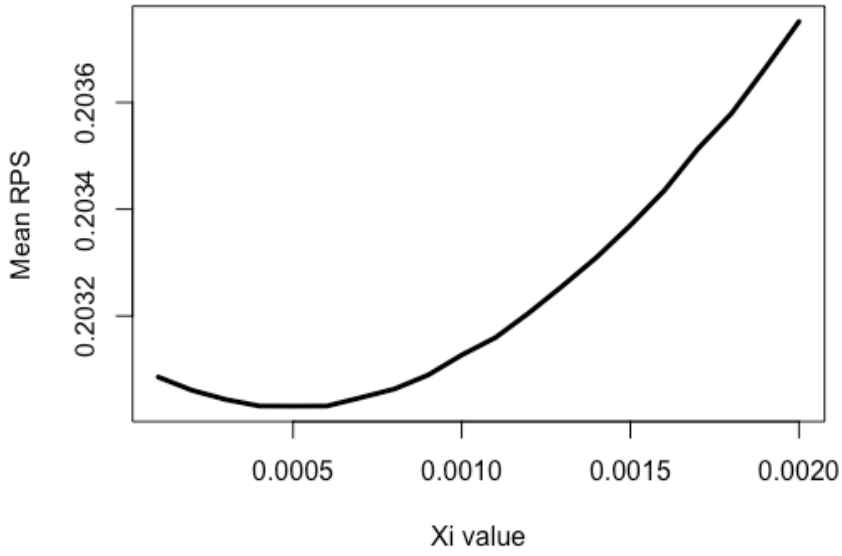


Figure 1 – Mean Ranked Probability Score versus ξ, the decay factor

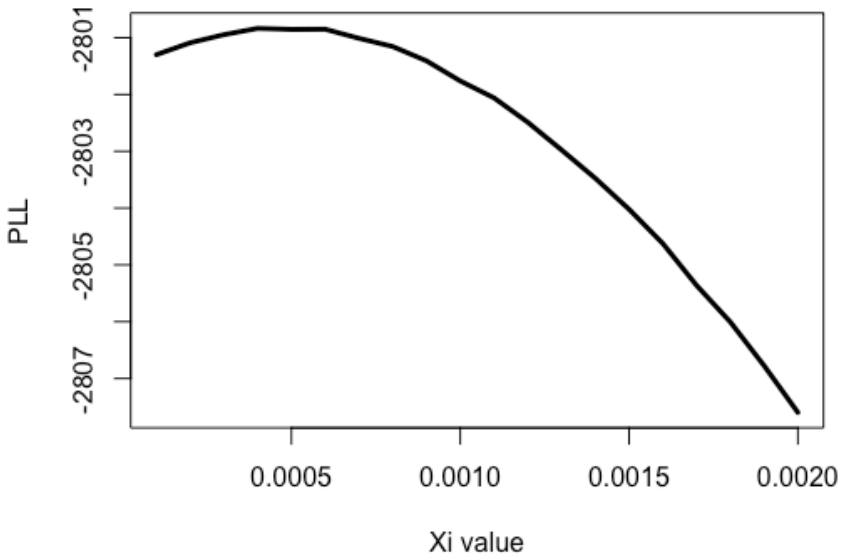


Figure 2 – Predicted Log-Likelihood versus ξ, the decay factor

Team strength dynamics

Table 1 shows the average attack and defence strength estimates for the 23 teams that competed in the Premiership either in the 2015-16 or 2016-17 seasons and the standard error from the mean attack and defence, rounded to 3 significant figures.

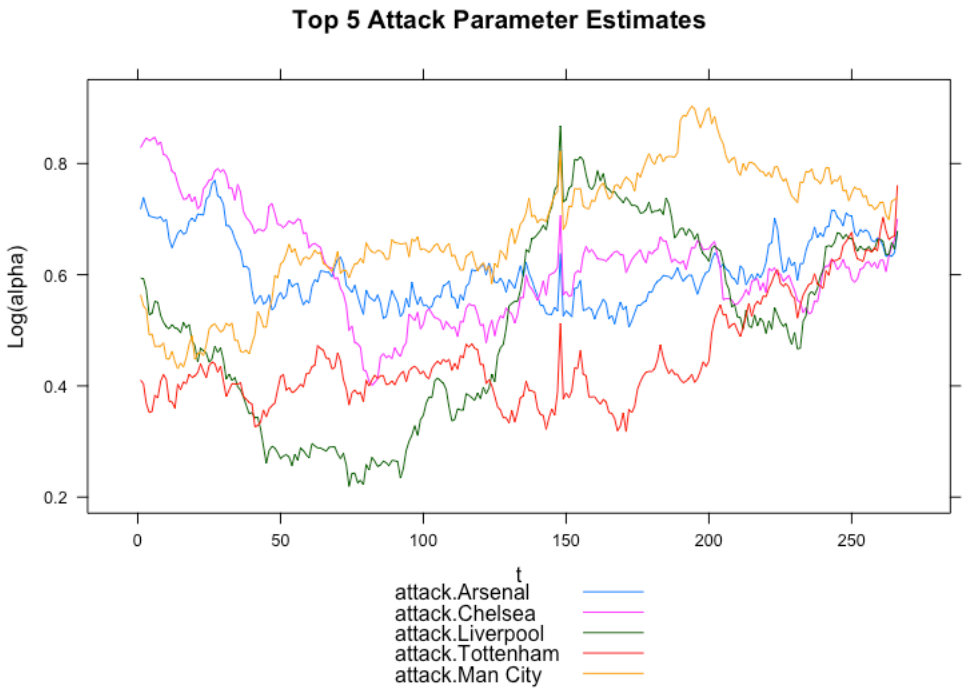
From the teams listed, we can see that Manchester City had the largest average attack strength, followed by Chelsea, Arsenal and Manchester United. Crystal Palace, Southampton, and Swansea however are the teams with the largest standard errors, reflecting that their attack changed significantly over the course of the 7 seasons being analysed. Other teams worth mentioning in this regard are Leicester City and Bournemouth. These are all teams which appear to have experienced significant improvements in their attack strength. The standard errors were measured as the deviation from the average strength parameter of all 266 "match rounds" that have been estimated.

In comparison, Manchester United had the best average defence, followed by Chelsea, Manchester City, Arsenal and Everton, with Southampton, Middlesbrough, and Bournemouth having the highest standard errors.

Figure 3 and Figure 4 illustrate the plots of attack and defence parameter estimates respectively over time. This data will form the basis for our time series analysis in the rest of Chapter 3. The teams shows below were split into 5 different categories for illustration purposes, as they share a similar range of parameter estimates on the y-axis: top 5, teams competing for an Europa League spot, mid-table teams, relegation candidates, and teams that were promoted from the Championship to the Premier League for the 2016-17 season. The time scale $t = 1, \dots, 266$ represents the 38 games per season over the course of 7 seasons. The attack and defence parameter estimates for the top 5 teams can be seen below. It is interesting to note that Tottenham has the strongest attack and defence at $t = 266$, whereas it had the weakest at $t = 1$ out of the "top 5" teams category. Manchester United's decline in attack strength is also noteworthy. Bournemouth's attack has improved substantially compared to other mid-table teams.

Table 1 - Mean attack and defence strengths of teams and the standard errors from the mean strength parameter over 266 match rounds

Team	$\log \overline{\alpha_i}$	$se(\log \alpha_i)$	$\log \overline{\beta_i}$	$se(\log \beta_i)$
Arsenal	0.605	0.0589	-0.341	0.111
Bournemouth	0.154	0.179	0.233	0.224
Brighton	-0.175	0.104	-0.0117	0.0889
Burnley	-0.0361	0.0617	0.0816	0.206
Chelsea	0.614	0.0937	-0.449	0.134
Crystal Palace	-0.108	0.235	0.0515	0.178
Everton	0.342	0.0865	-0.283	0.0677
Huddersfield	-0.187	0.0676	0.387	0.122
Hull City	-0.159	0.118	-0.00903	0.0984
Leicester City	0.145	0.194	0.0264	0.178
Liverpool	0.511	0.170	-0.275	0.119
Man City	0.678	0.116	-0.441	0.0807
Man United	0.601	0.129	-0.454	0.134
Middlesbrough	-0.123	0.0647	0.0540	0.226
Newcastle	0.201	0.0876	0.00216	0.0955
Southampton	0.221	0.218	-0.112	0.282
Stoke	0.0387	0.136	-0.172	0.0517
Sunderland	0.0627	0.0622	-0.0522	0.0614
Swansea	0.0746	0.200	-0.117	0.0579
Tottenham	0.450	0.0919	-0.270	0.131
Watford	0.0925	0.114	0.216	0.169
West Brom	0.160	0.0792	-0.0316	0.119
West Ham	0.187	0.113	-0.0472	0.118



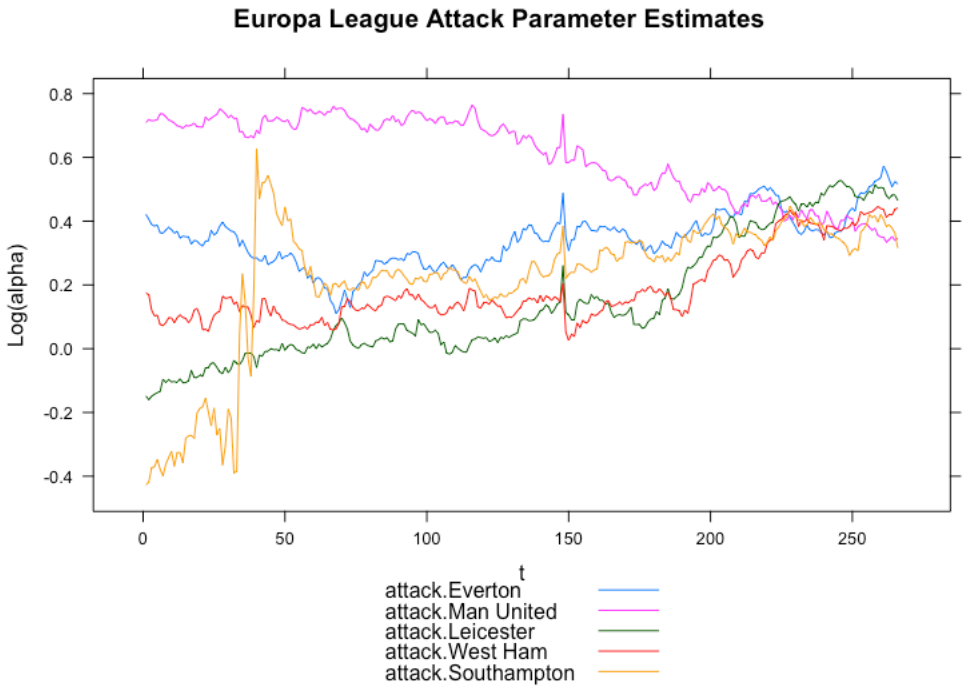
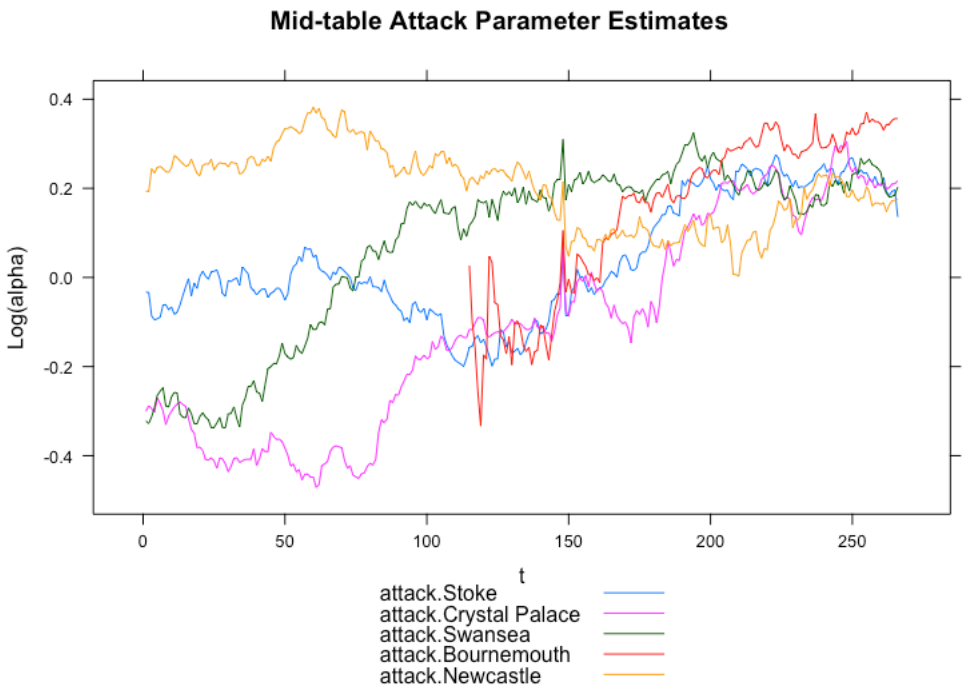


Figure 3 – Attack Strength Parameter Estimates over 9 seasons (until 2016-17)



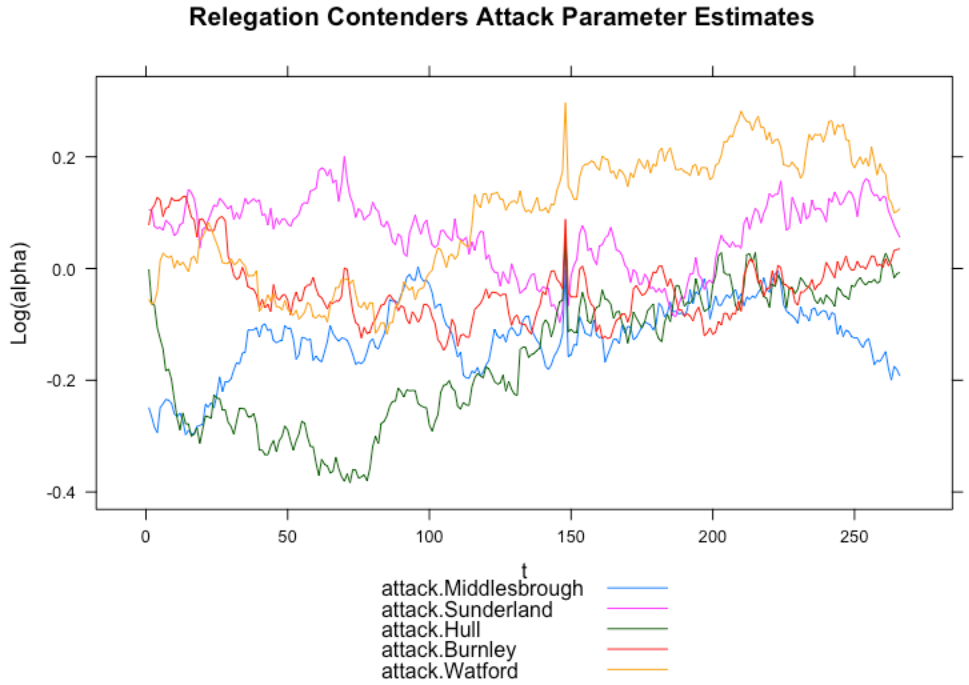


Figure 3 Continued

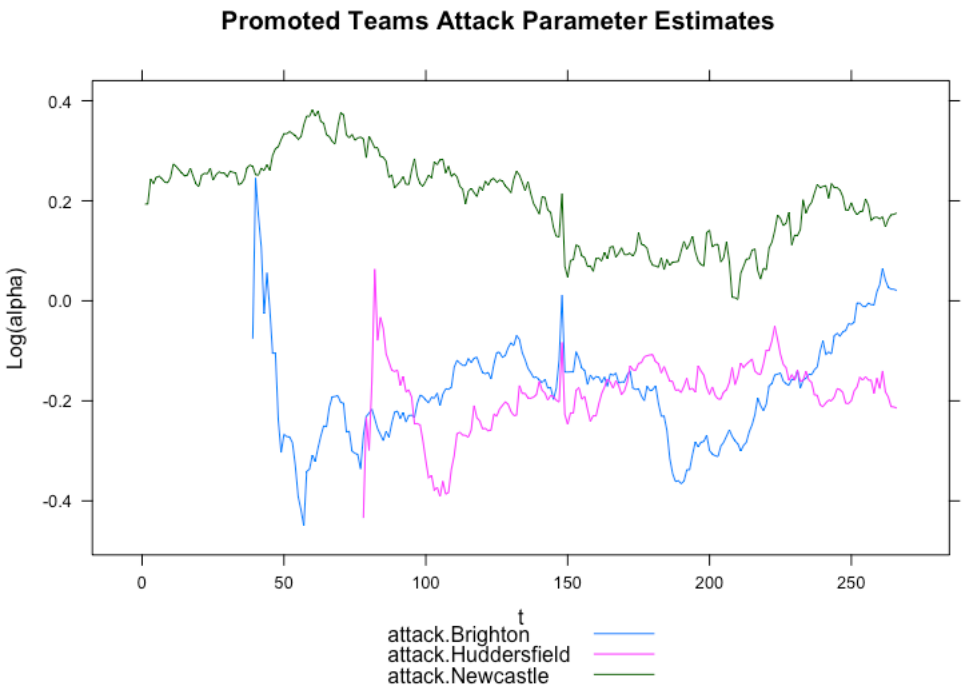
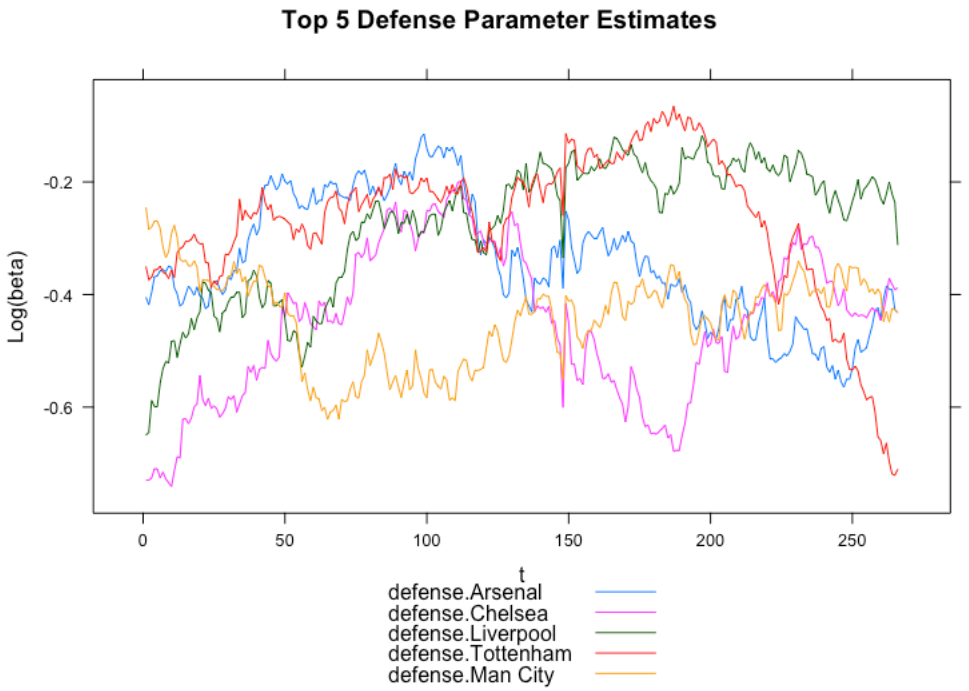


Figure 3 Continued



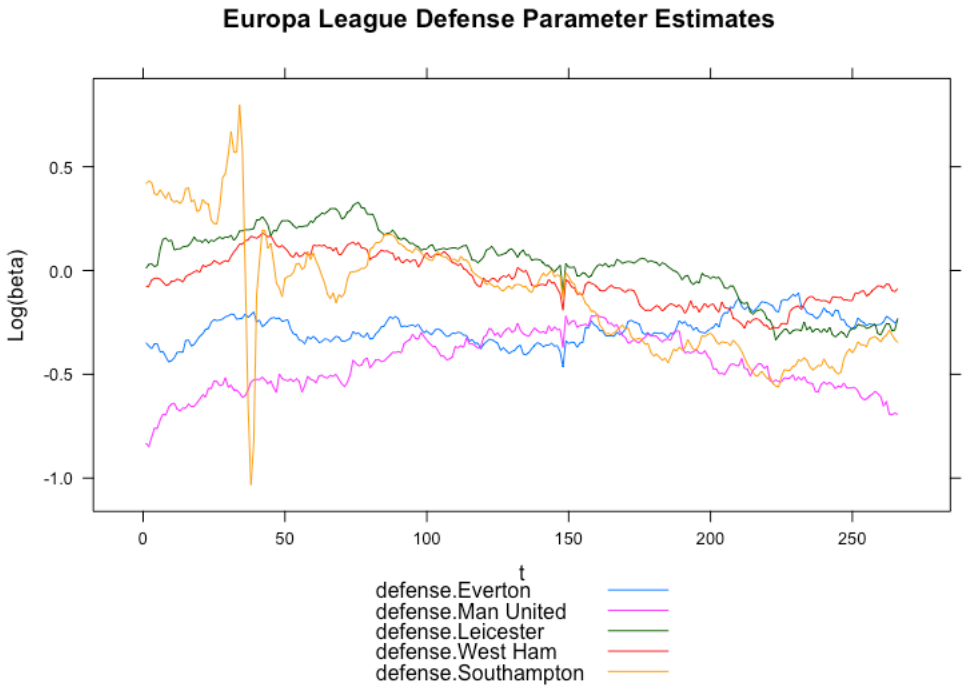
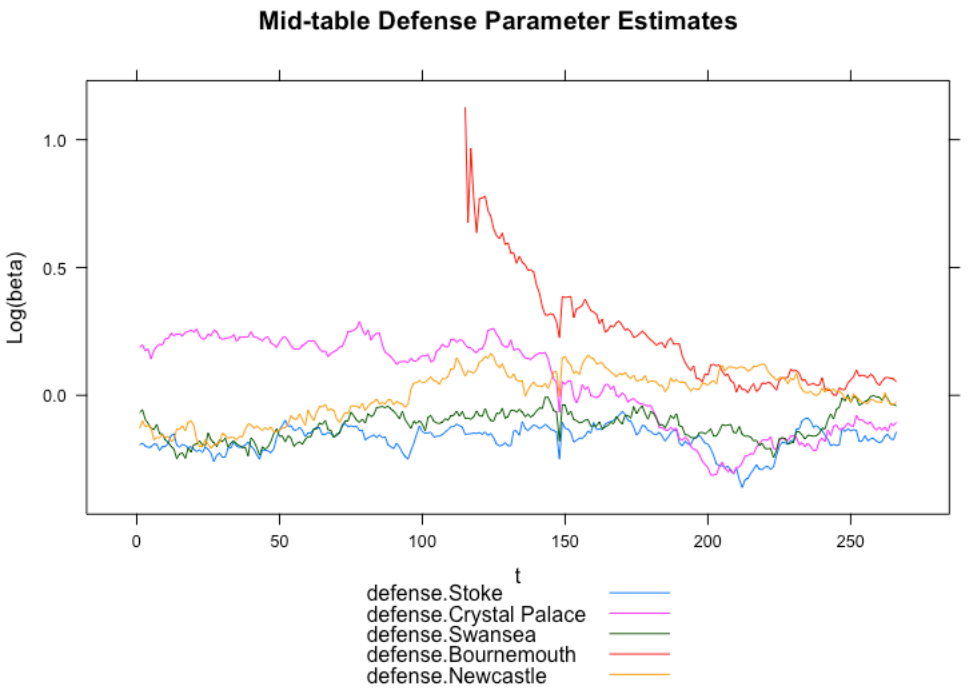


Figure 4 - Defence Strength Parameter Estimates over 9 seasons (until 2016-17)



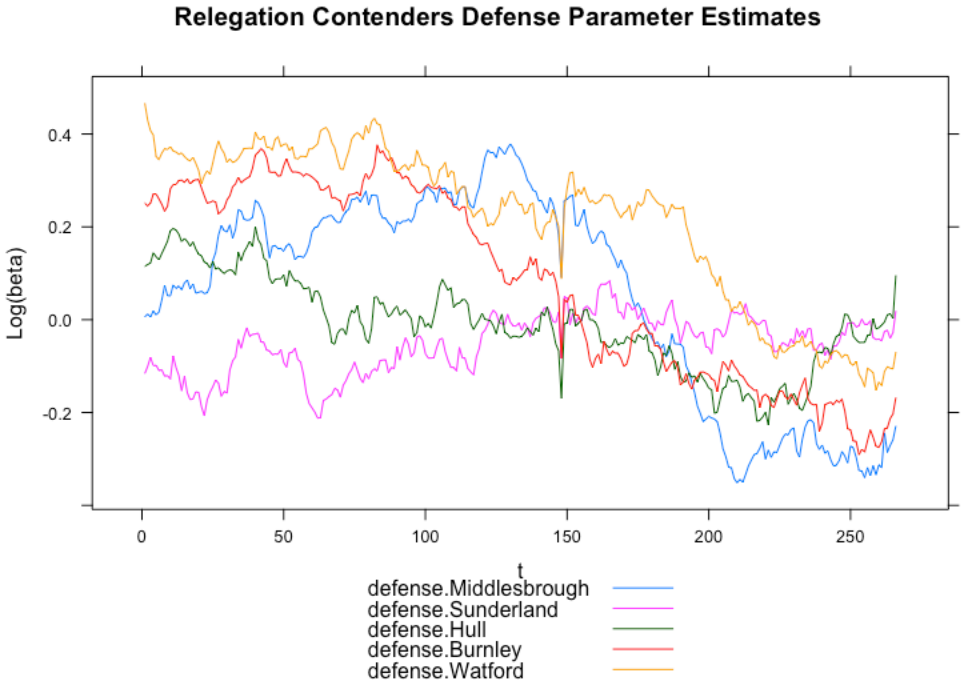


Figure 4 Continued

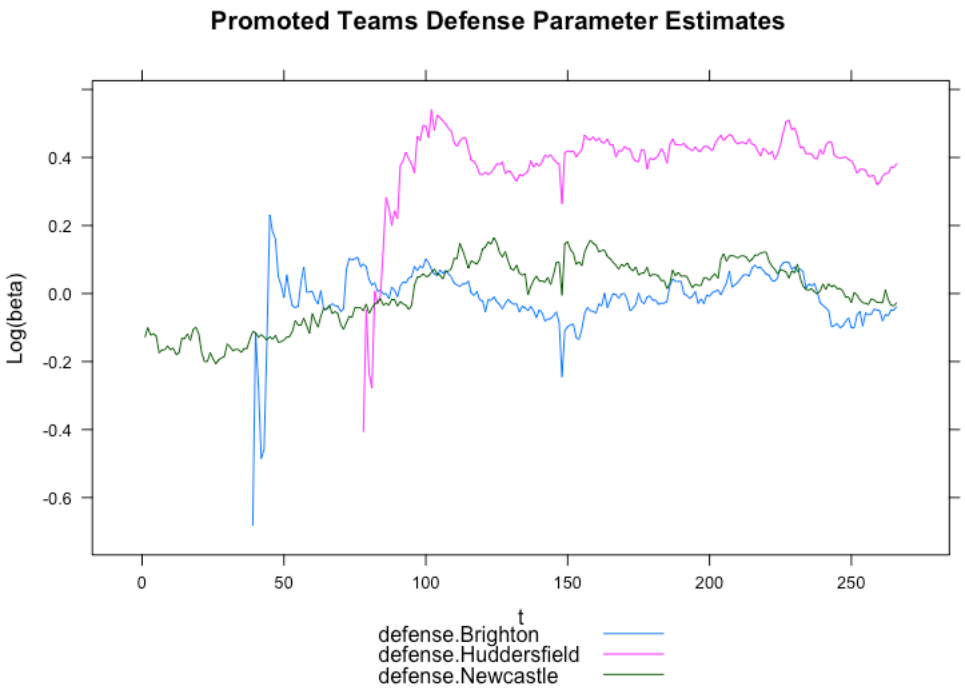


Figure 4 Continued

There is a particular difficulty in estimating the strength of new teams as they join the Championship, and this can be seen with the increased volatility in teams like Huddersfield, Brighton, Bournemouth and Southampton. The way of estimating team strengths when teams have only played a few matches is not a topic of research of this thesis, but could be the subject of future research. Nevertheless, as more matches become available during the rolling window, the volatilities decrease and stabilise and adequate forecasts can be carried out for the last 2 seasons.

It is interesting to note how the log home advantage parameter γ also seems to have declined over the course of 7 seasons. This seems to be in agreement with the findings in Owen (2016).

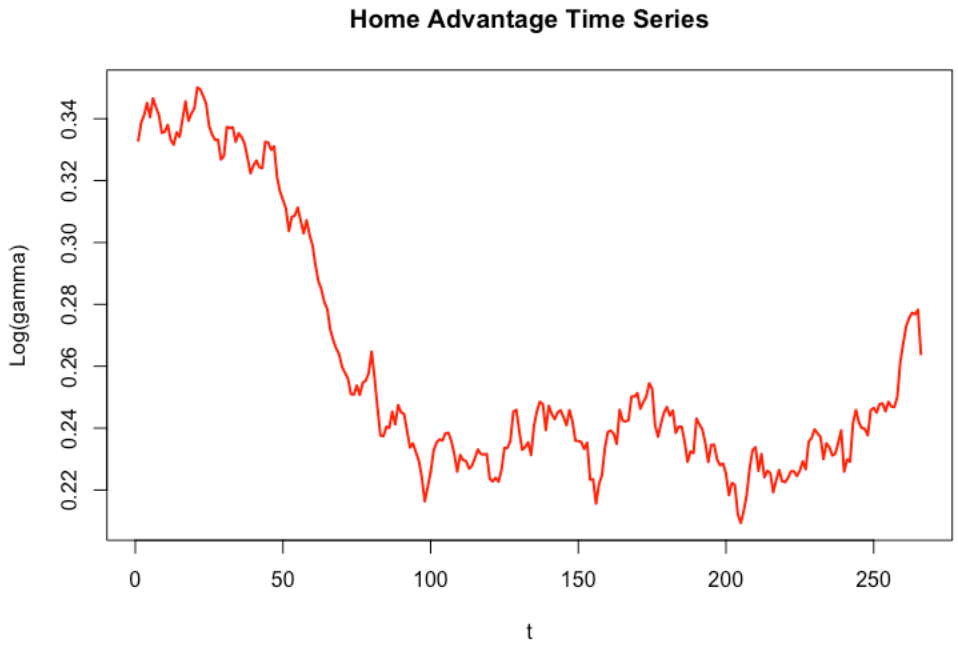


Figure 5 – Time series for home advantage over the 266 rounds

The parameter estimate $\hat{\rho}$ in equation 11 (which inflates or deflates the probability of low scores) was estimated and can be seen in Figure 6. When $\hat{\rho} < 0$ as is the case during most of our sample, the probability of 0-0 and 1-1 draws increase and the probability of 1-0 and 0-1 wins decrease. When $\hat{\rho} > 0$, the opposite occurs. In the case where $\hat{\rho} = 0$, it means that $\tau = 1$ in equation 11 and equation 8 can be simplified to an independent bivariate Poisson model. Our parameter estimate of $\hat{\rho}$ shows that in most of our time series, $\hat{\rho} < 0$.

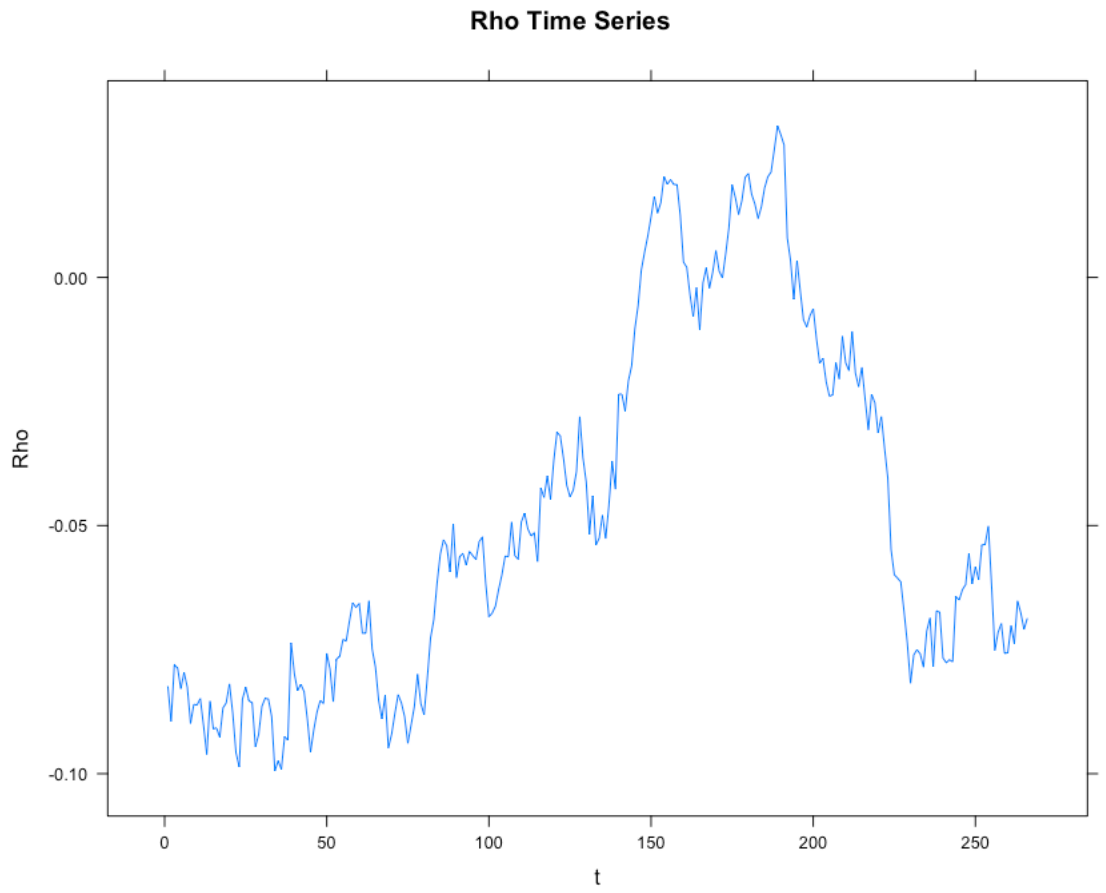


Figure 6 - Time series for ρ over the 266 rounds

Testing for the presence of structural breaks

The Bai and Perron (2003) dynamic programming algorithm is implemented to calculate the number of breakpoints in each time series and the position of these breakpoints. There are only 9 time series out of 46 for which no breakpoints occur: the attacks of Brighton, Burnley, Crystal Palace, Everton and West Brom; and the defences of Arsenal, Manchester City, Southampton and Stoke City. This part of the analysis was done using functions of the "strucchange" package in R.

Figure 7 and Figure 8 illustrate the breakpoints in some of the time series of team strengths (with 95% confidence intervals) and their fitted values of an *AR(1)* model with intercept and a trend component. The black plots represent the time series of attack and defence strengths for each team (same plots as Figure 3 and Figure 4), whereas the blue plots are the fitted estimates. Red plots are the fitted estimates of time series which do not have a structural break. The vertical

dotted lines show the estimated breakpoint in the time series, and the red horizontal line at either side are the 95% confidence intervals of that breakpoint.

The parameter estimates for each segment can be seen in Table 2 and Table 3 for attack and defence time series respectively; where $\hat{c}$ is the intercept, $\hat{\delta}$ the slope of the linear trend component and $\hat{\phi}$ the autoregressive parameter estimate. Most time series have at two breakpoints (split into three segments). When that is the case, the middle segment usually has a smaller value of $\hat{\phi}$, therefore the fitted estimates do not seem to fit particularly well and do not fluctuate around the mean as much as the observed estimates (only the intercept and the trend components are significant in explaining the series). This can be seen graphically in Figure 7 and Figure 8.

Most estimates of the autoregressive parameter $\hat{\phi}$ are in the range $0 < \hat{\phi} < 1$, but there are some notable exceptions. In Liverpool's attack series, $\hat{\phi} > 1$, which is indicative of an explosive process. This explosive behaviour can also be seen in the first segment of Chelsea's defence series.

Regarding other teams' defences, Bournemouth's autoregressive parameter in their first segment is negative, which indicates mean-reversion with alternating sign: this would indicate that their defence will be stronger than the mean next round if it was below the mean this period. Manchester United's autoregressive parameter for their defence time series is very close to zero in their second segment which means that there is extremely little fluctuation around the mean. Similar results can be found for Leicester's and Sunderland's defences.

Given the range of values that the dependent variable can take, it is not surprising that we observe small parameter values for the trend component $\hat{\delta}$. However, the Bai and Perron (2003) procedure still seems to pick up changes in the trend component between segments, particularly when the order of magnitude of the parameter $\hat{\delta}$ changes by at least one, which according to the results, seems to be rather common.

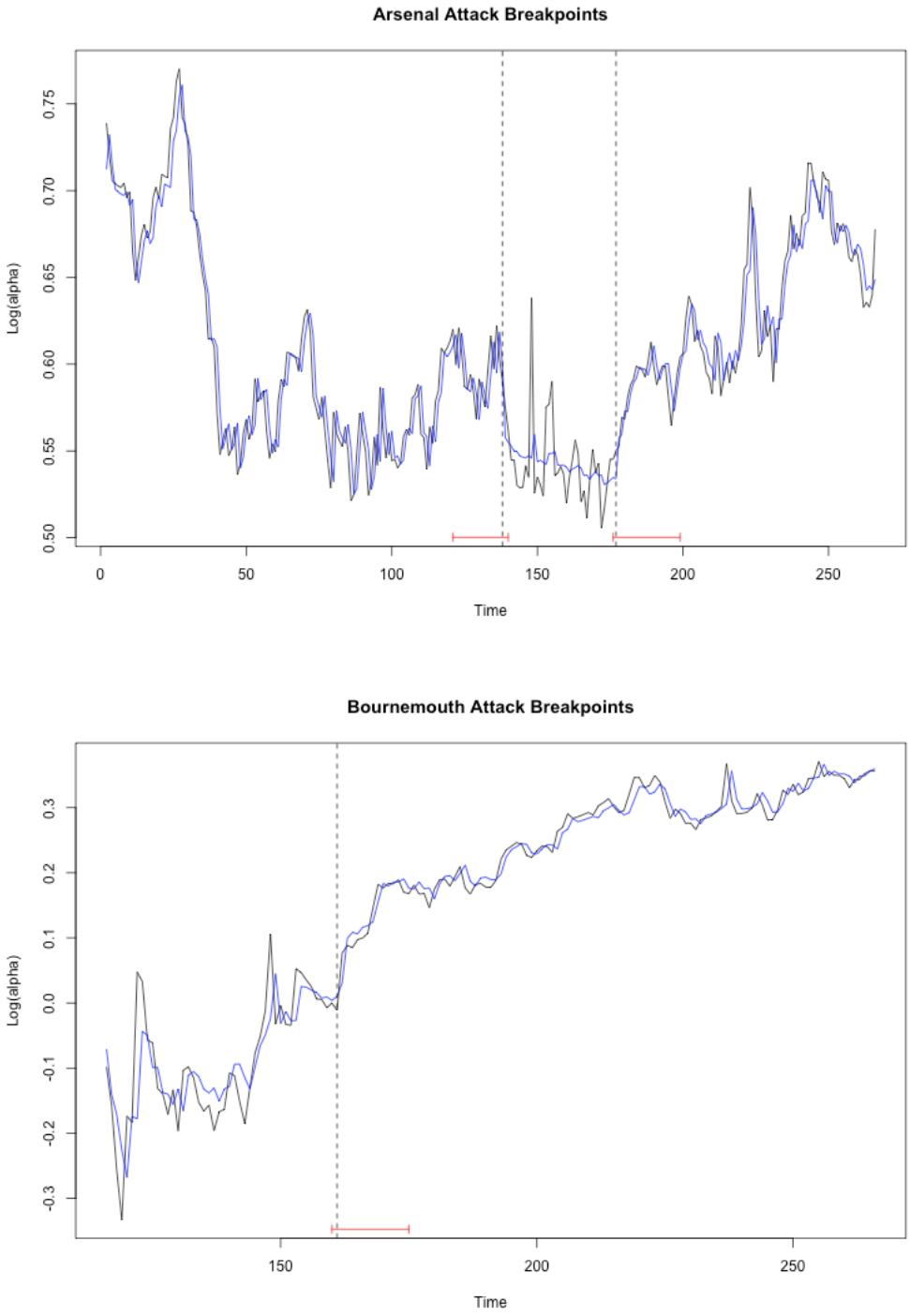
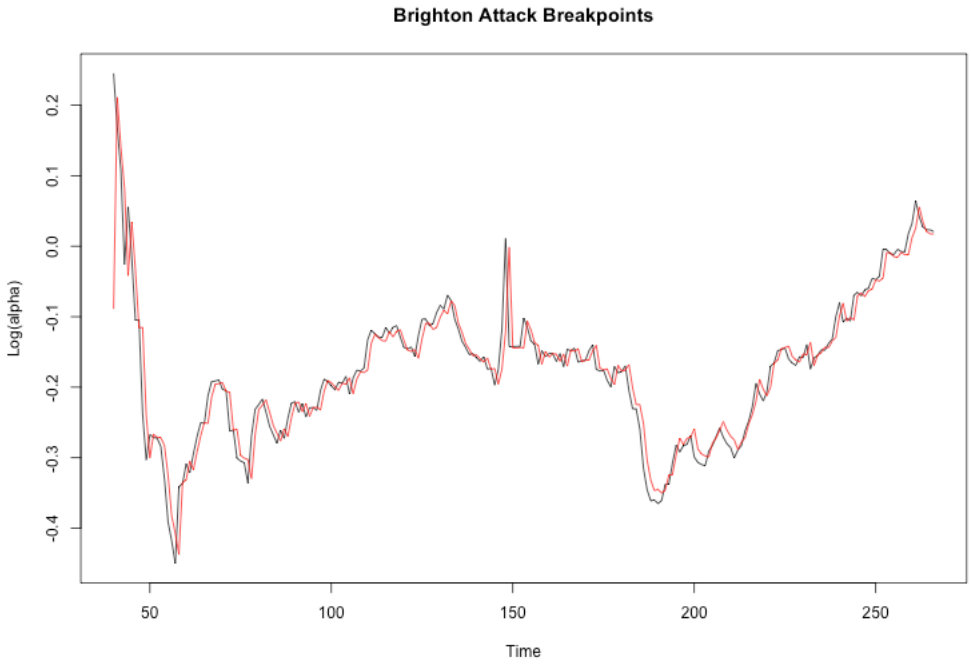


Figure 7 - Attack Breakpoint Estimates over 266 rounds, from 2010-11 season to 2016-17 season



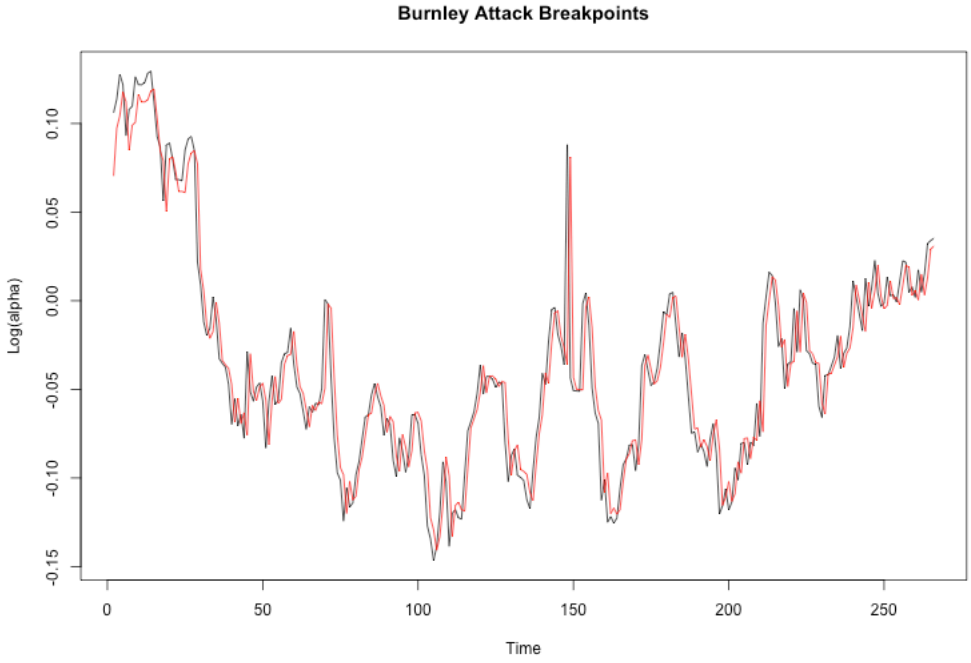
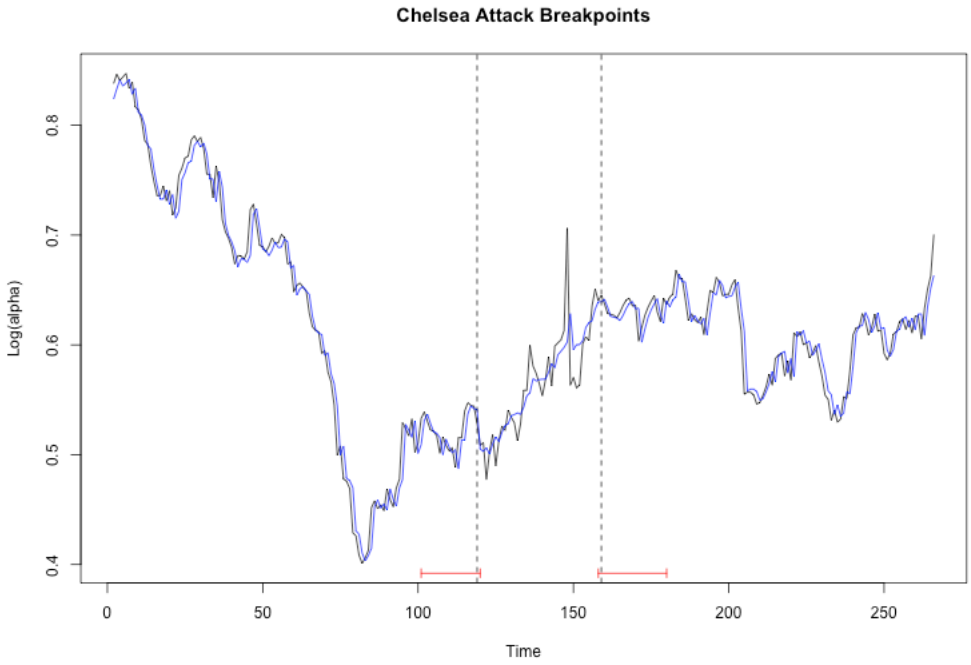


Figure 7 Continued



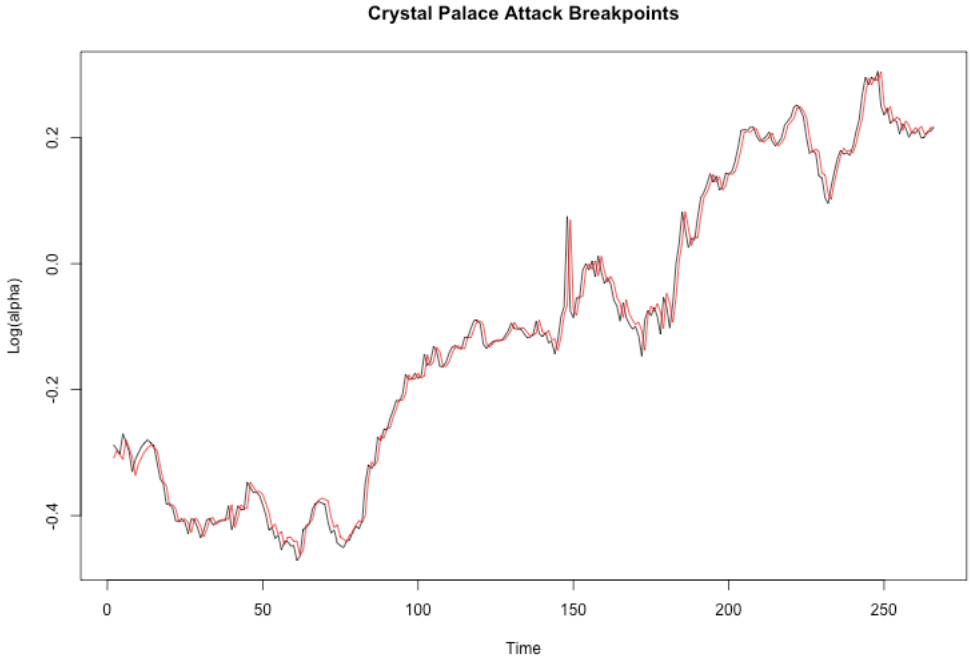
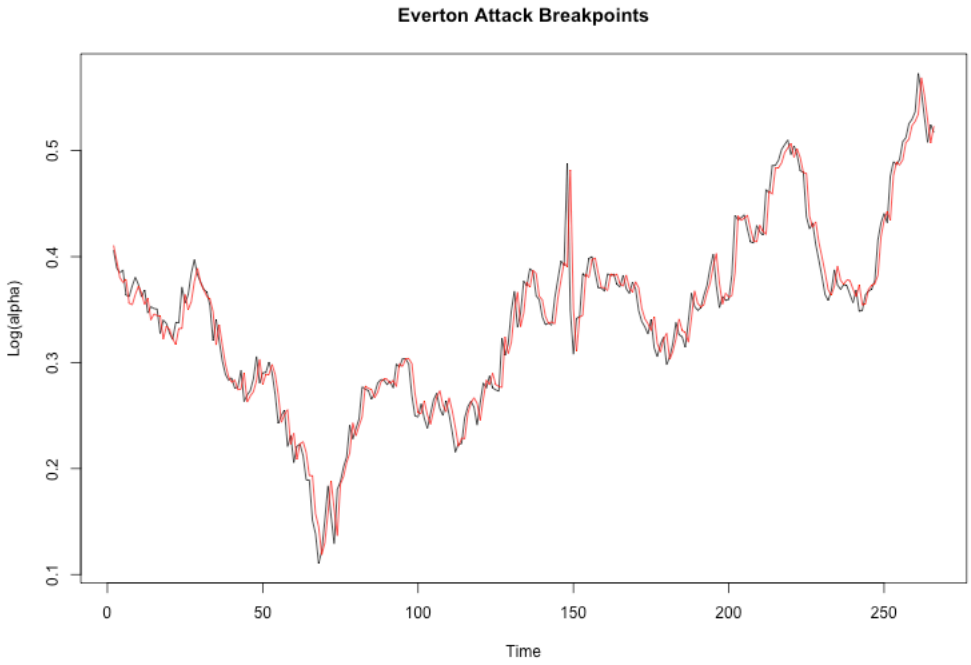


Figure 7 Continued



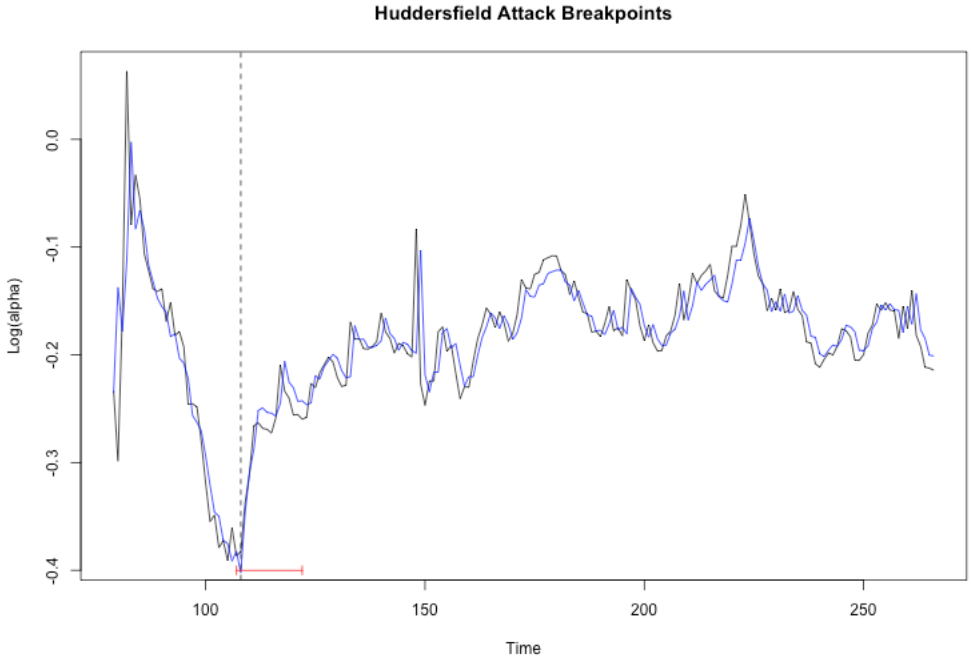
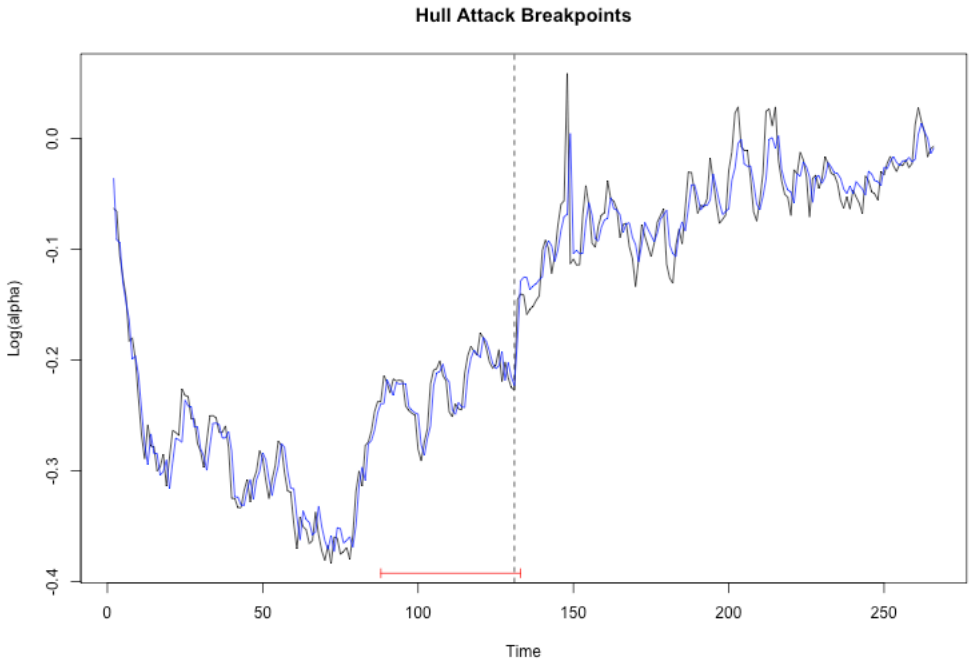


Figure 7 Continued



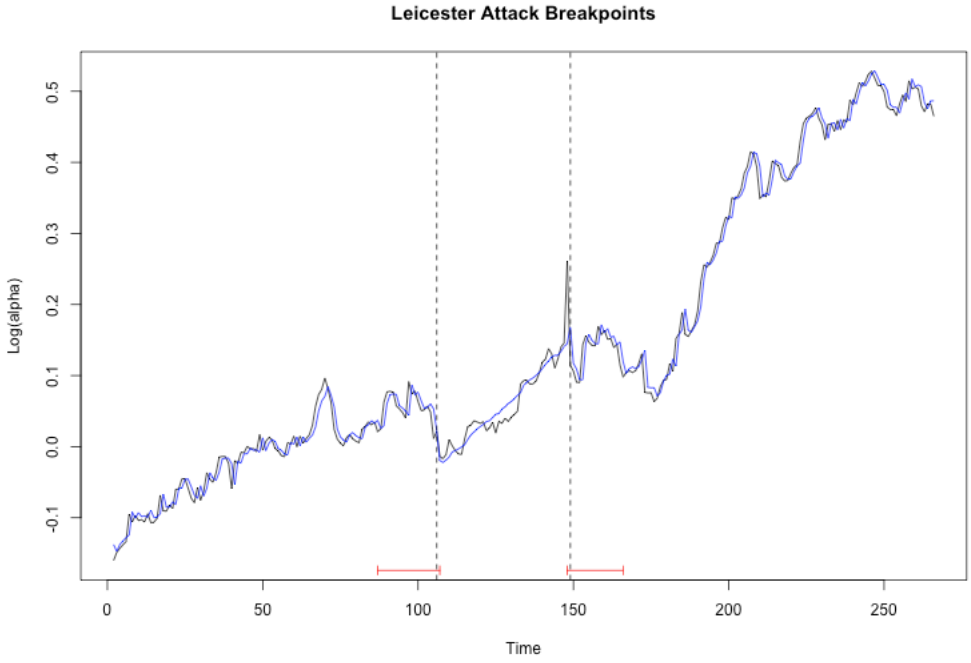
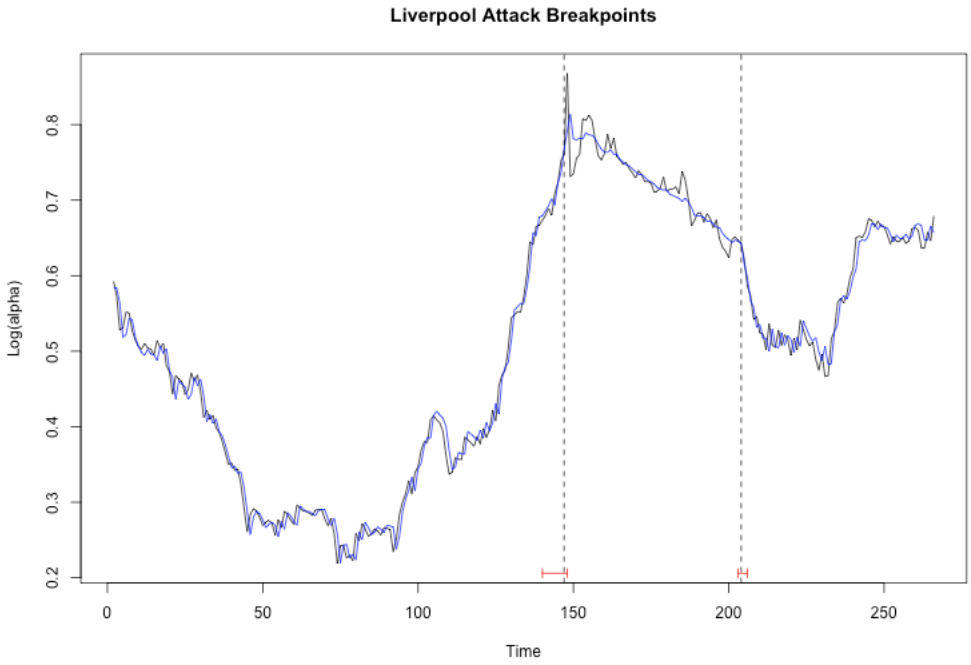


Figure 7 Continued



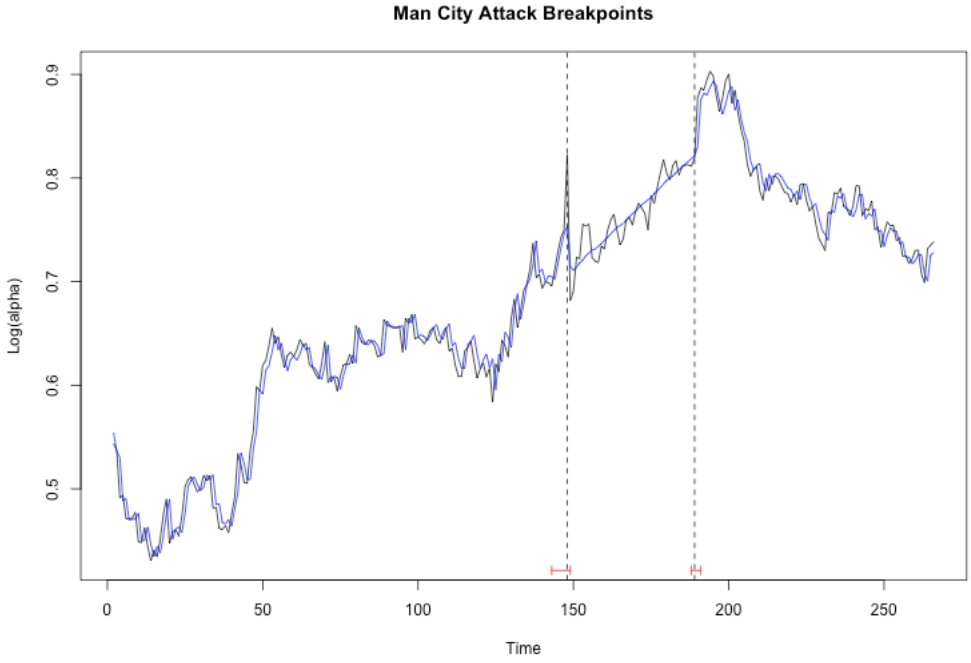
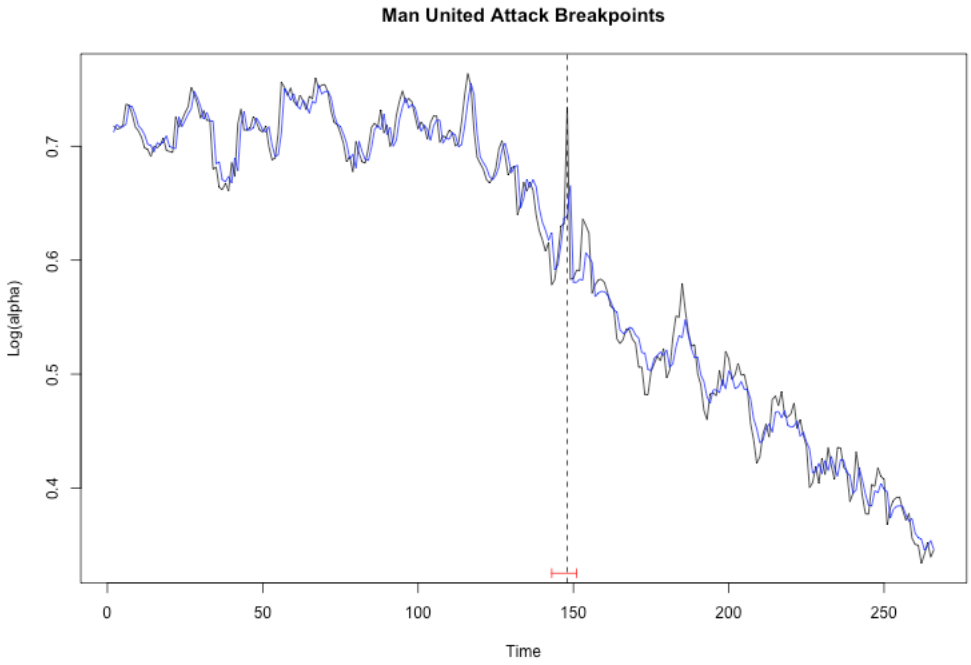


Figure 7 Continued



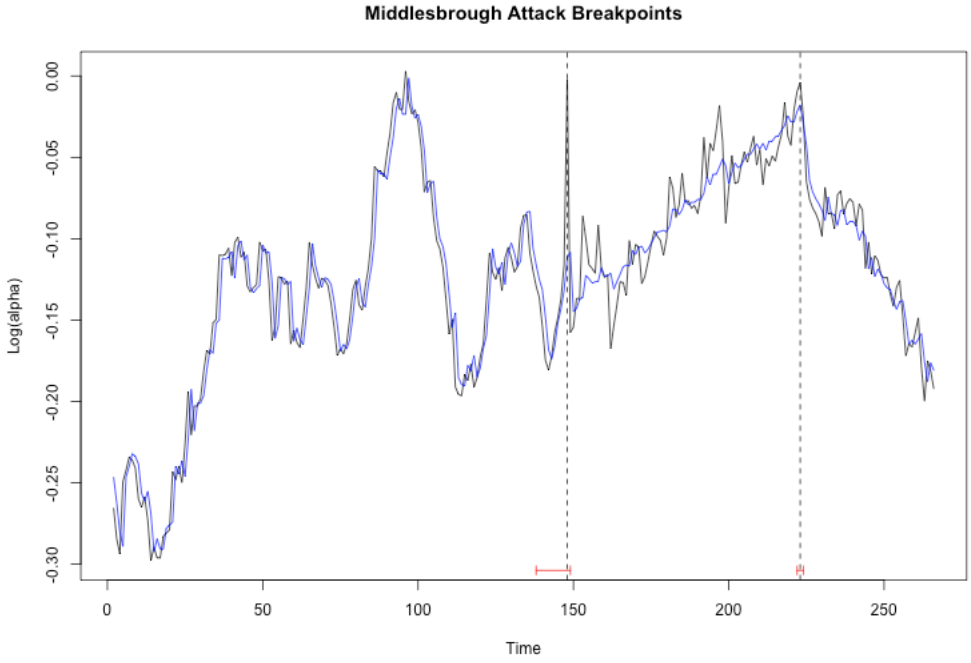
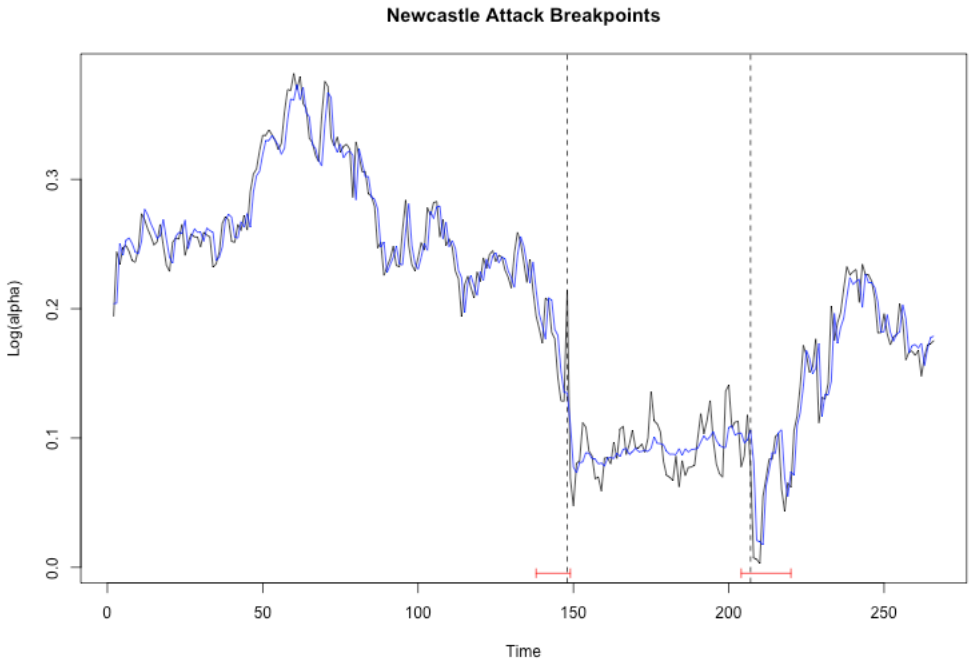


Figure 7 Continued



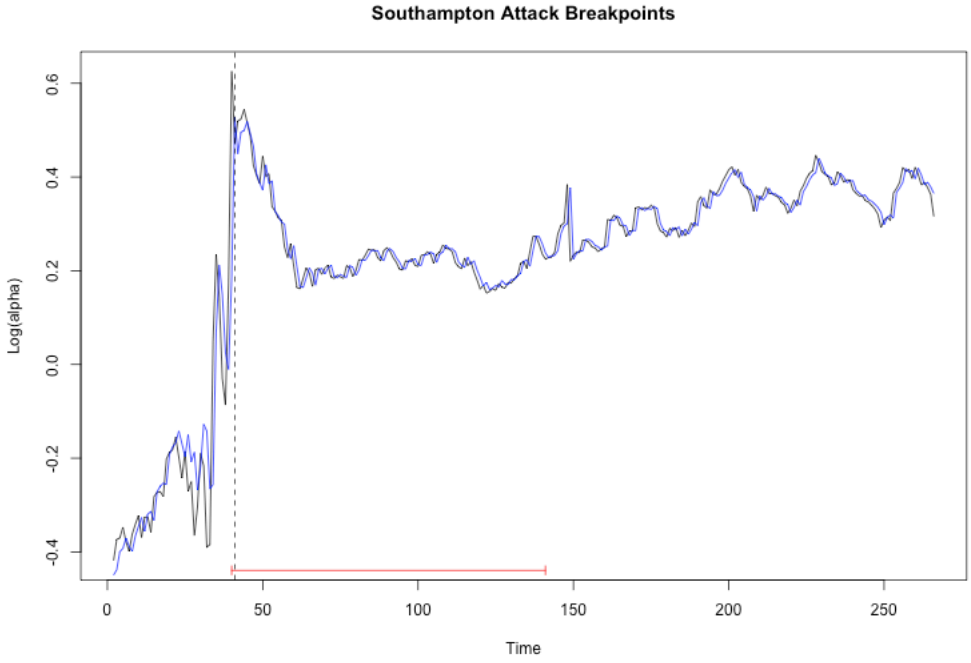
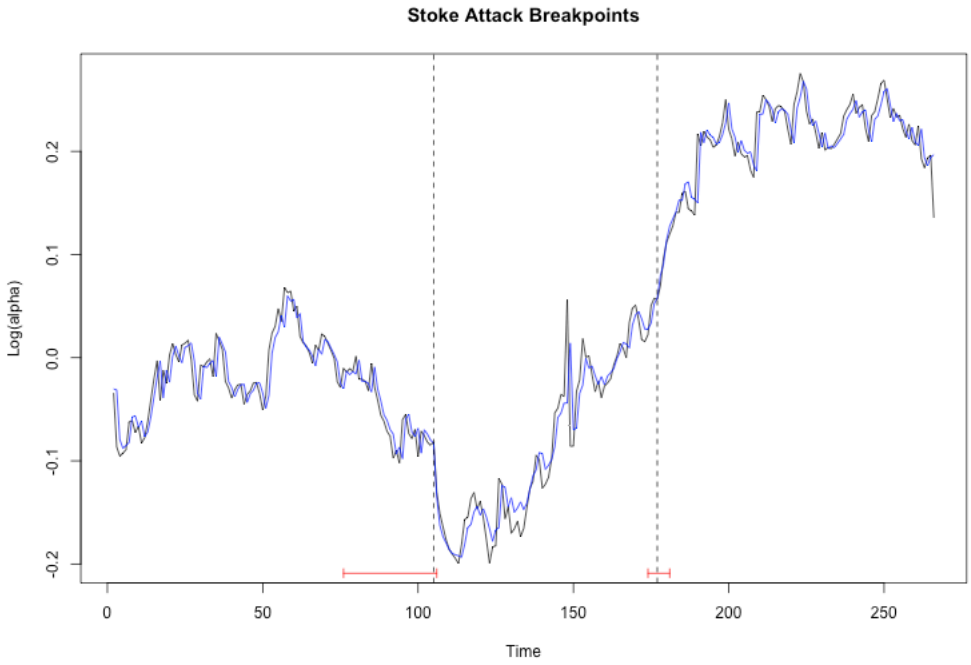


Figure 7 Continued



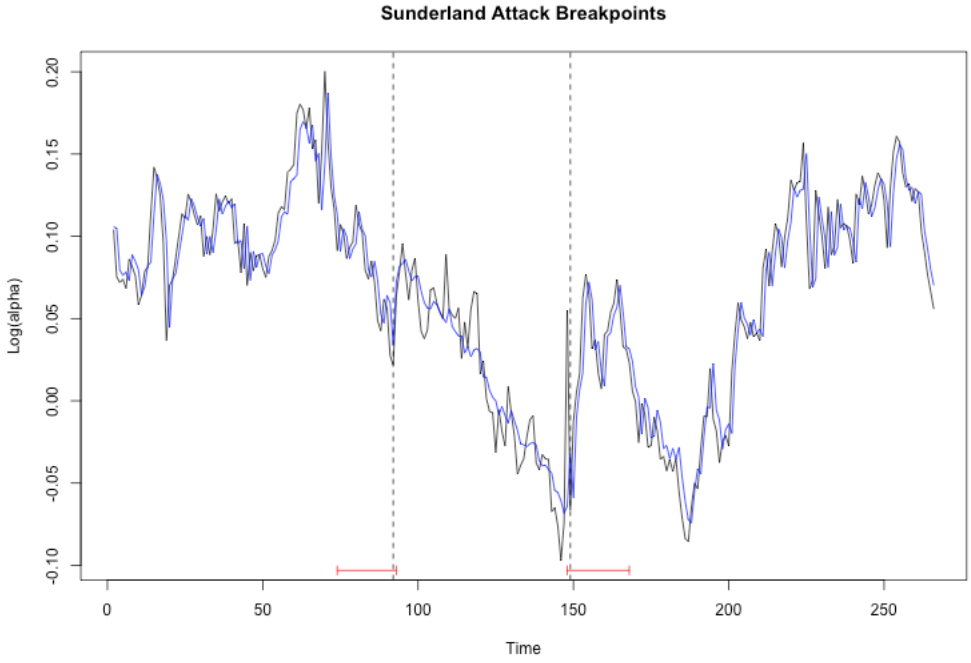
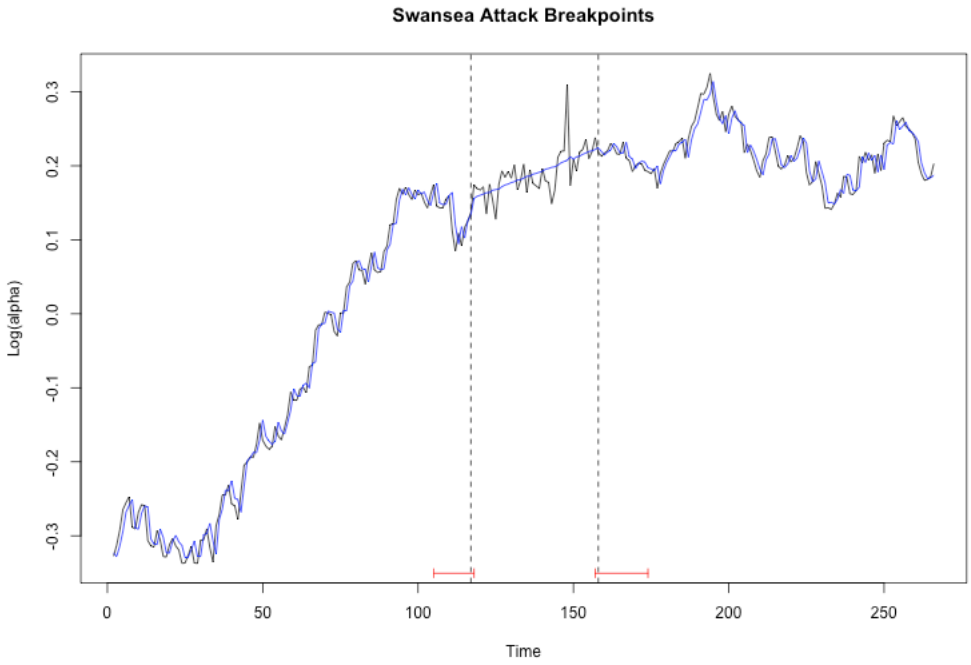


Figure 7 Continued



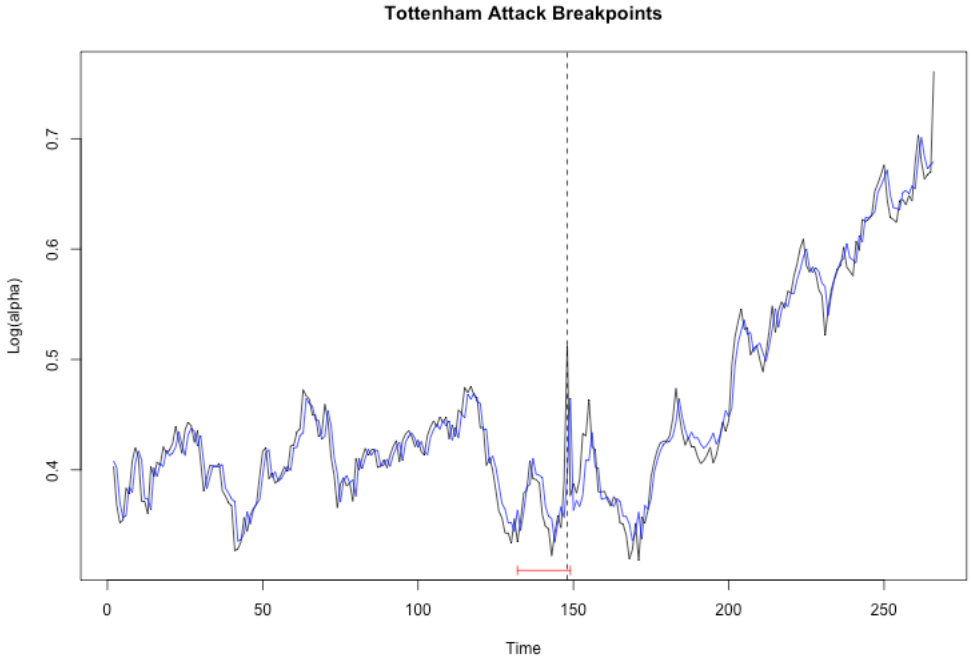
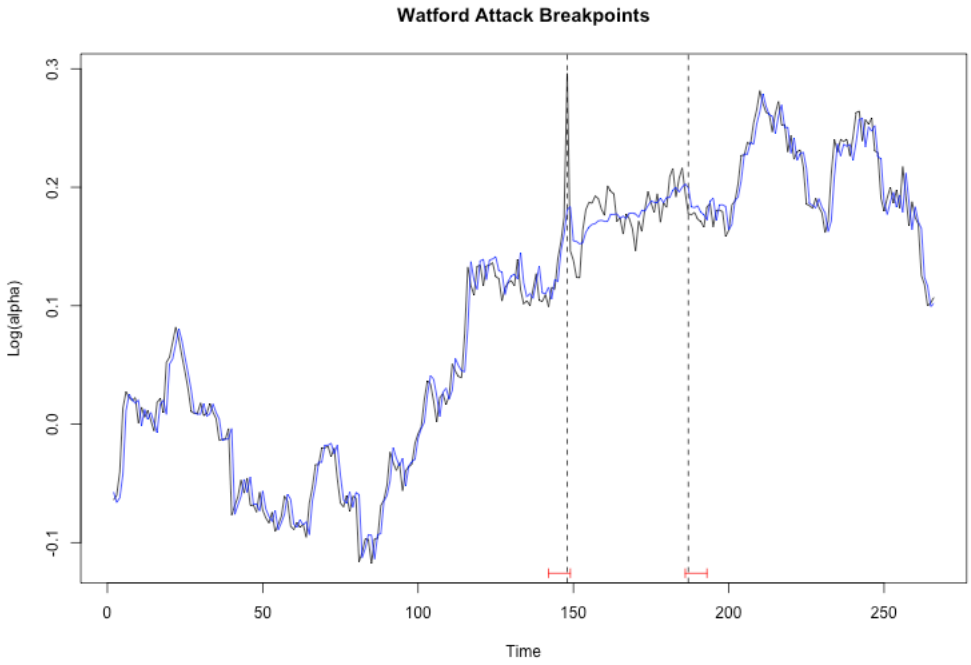


Figure 7 Continued



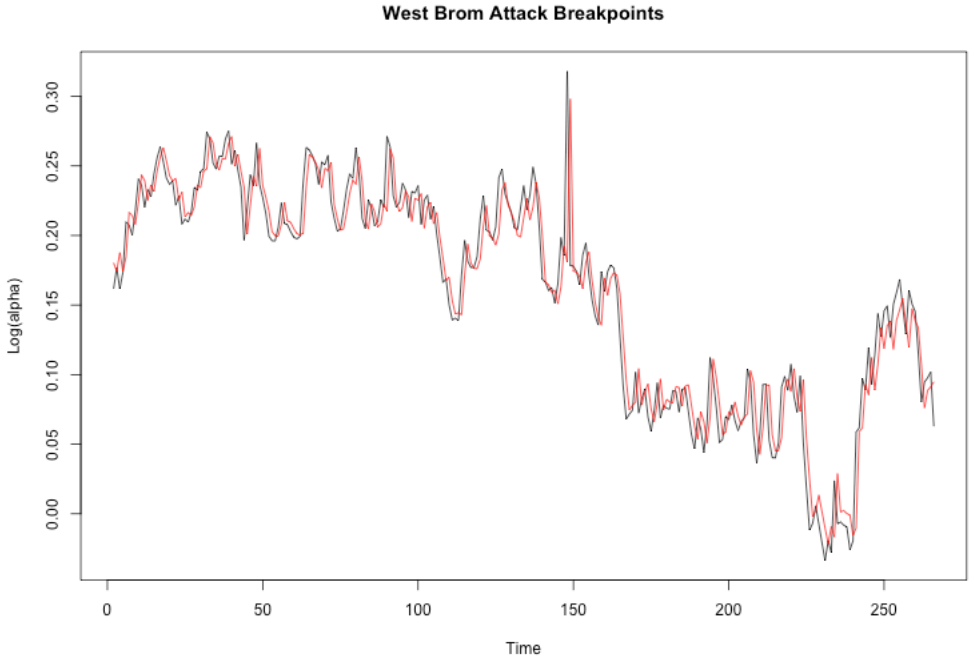


Figure 7 Continued

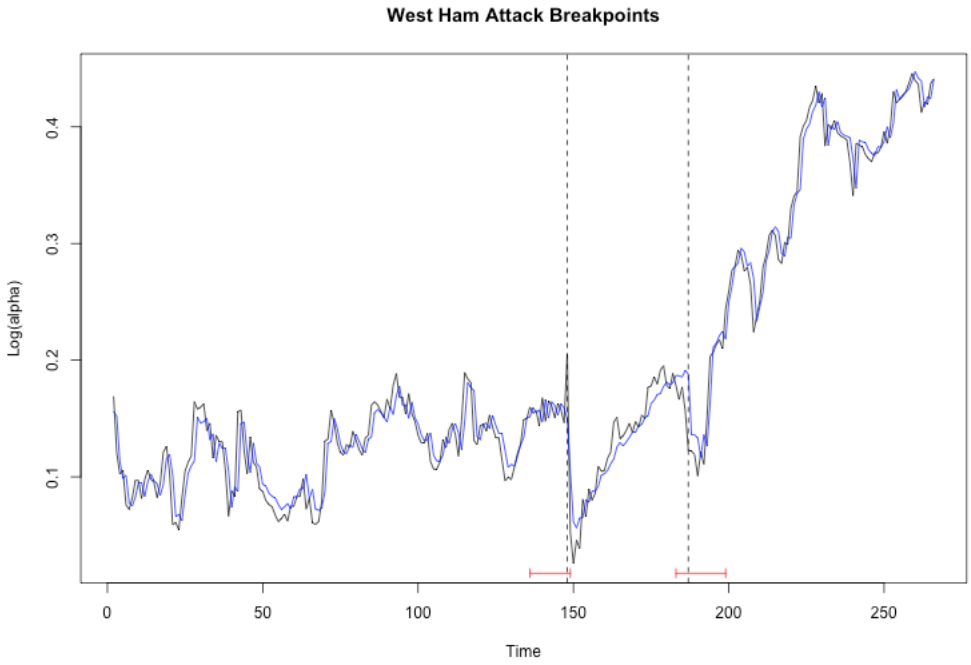


Figure 7 Continued

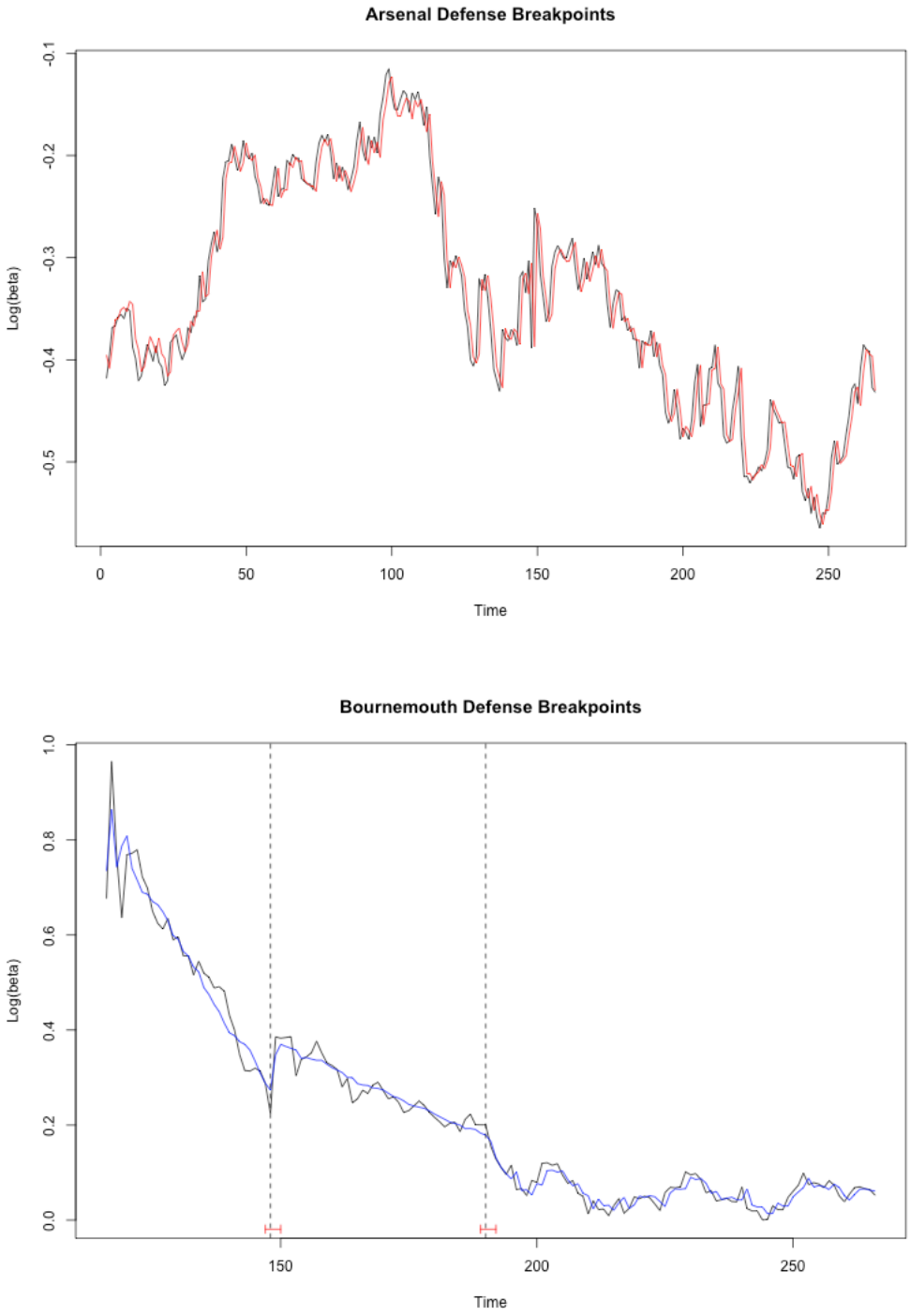
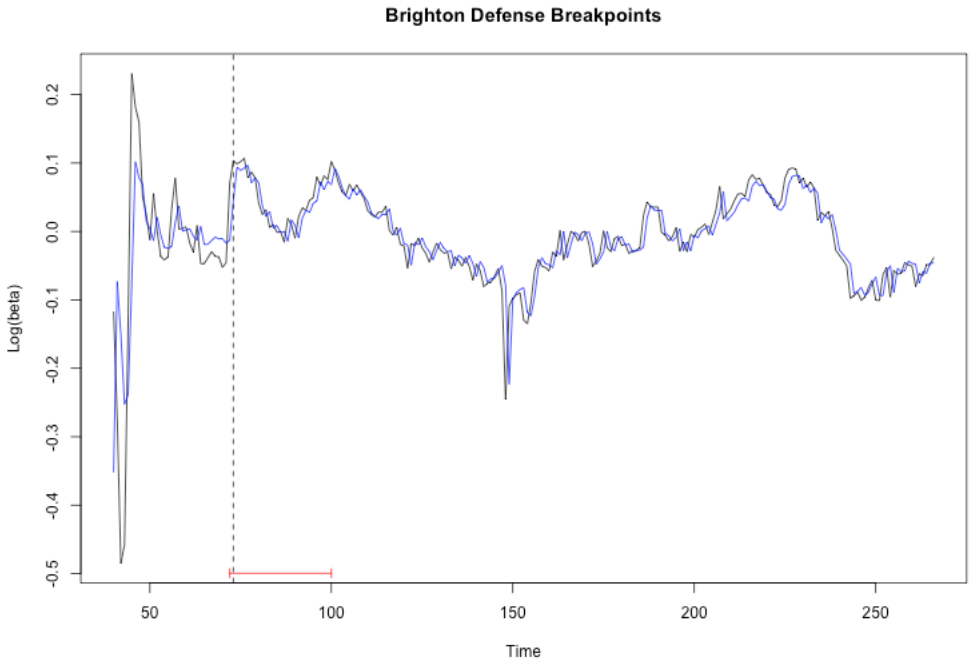


Figure 8 - Defence Breakpoint Estimates over 266 rounds, from 2010-11 season to 2016-17 season



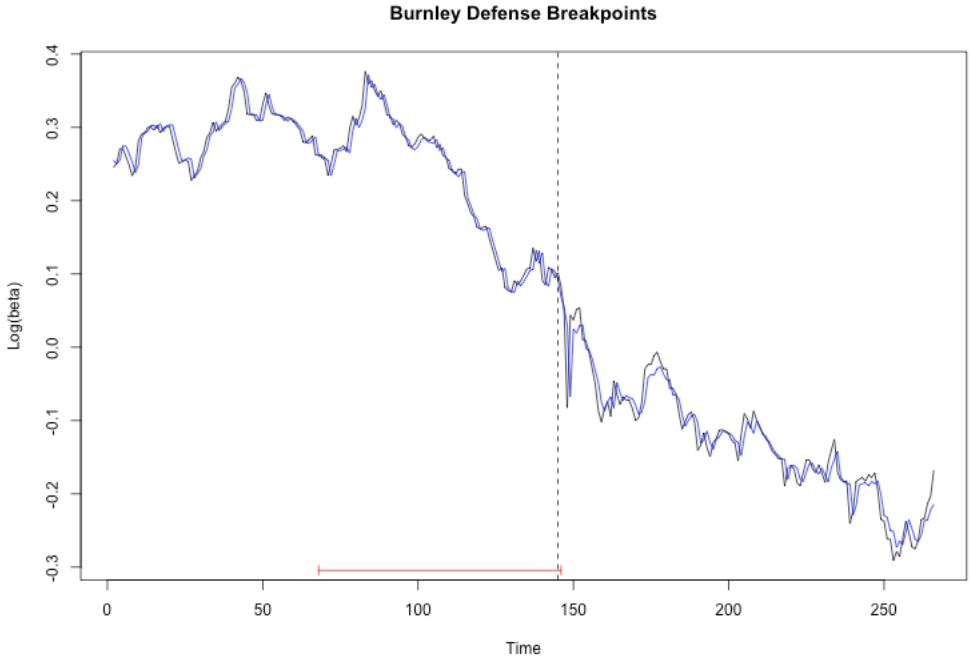
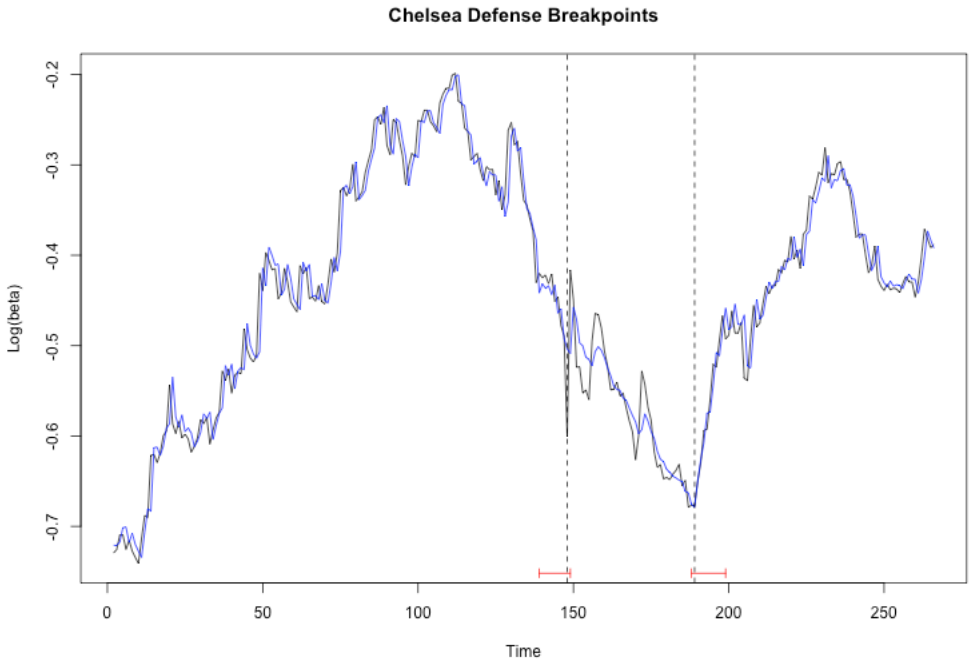


Figure 8 Continued



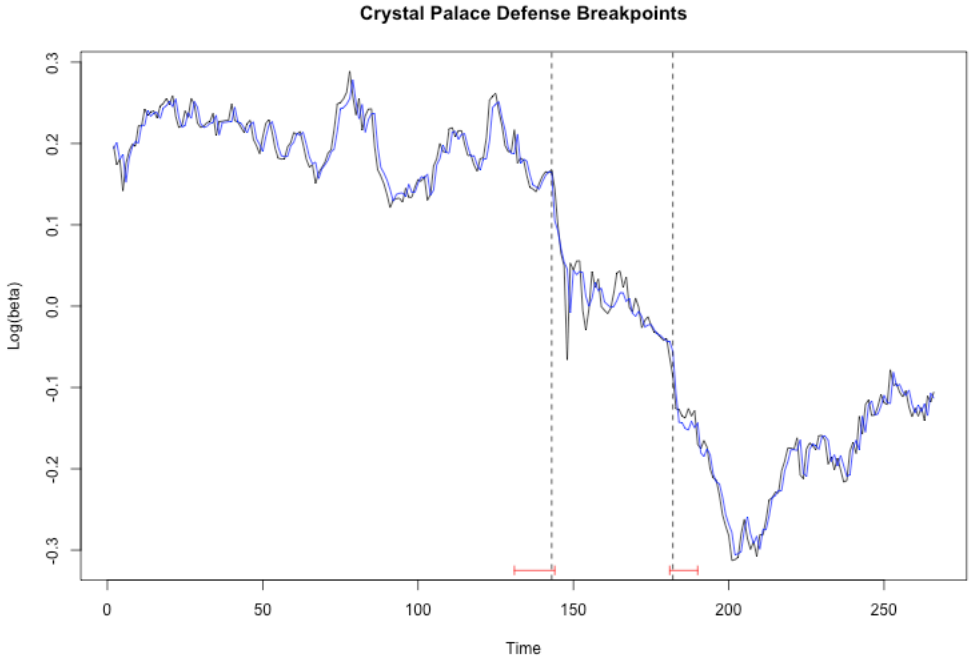
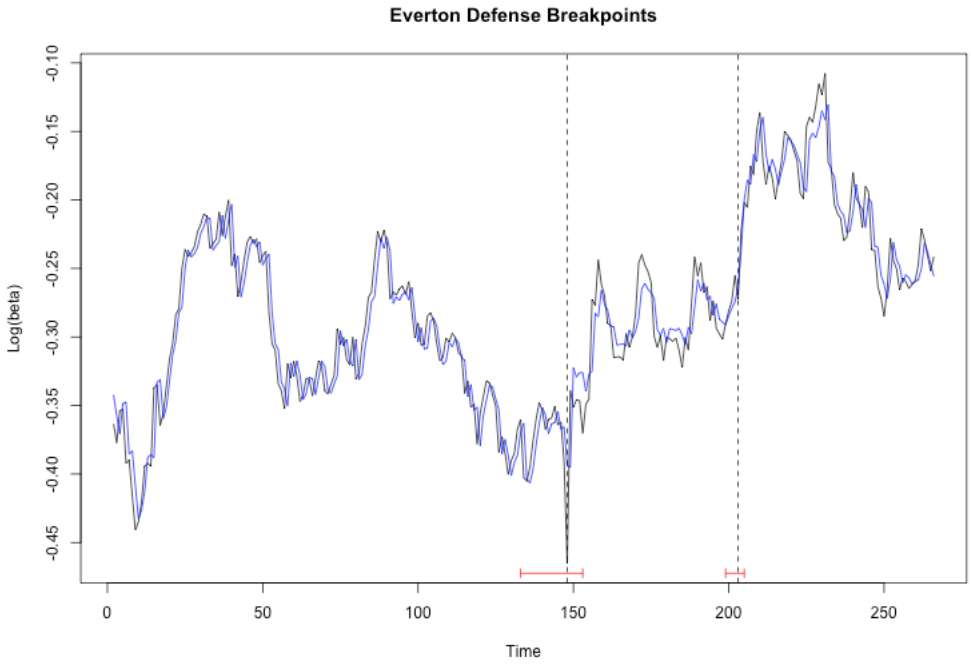


Figure 8 Continued



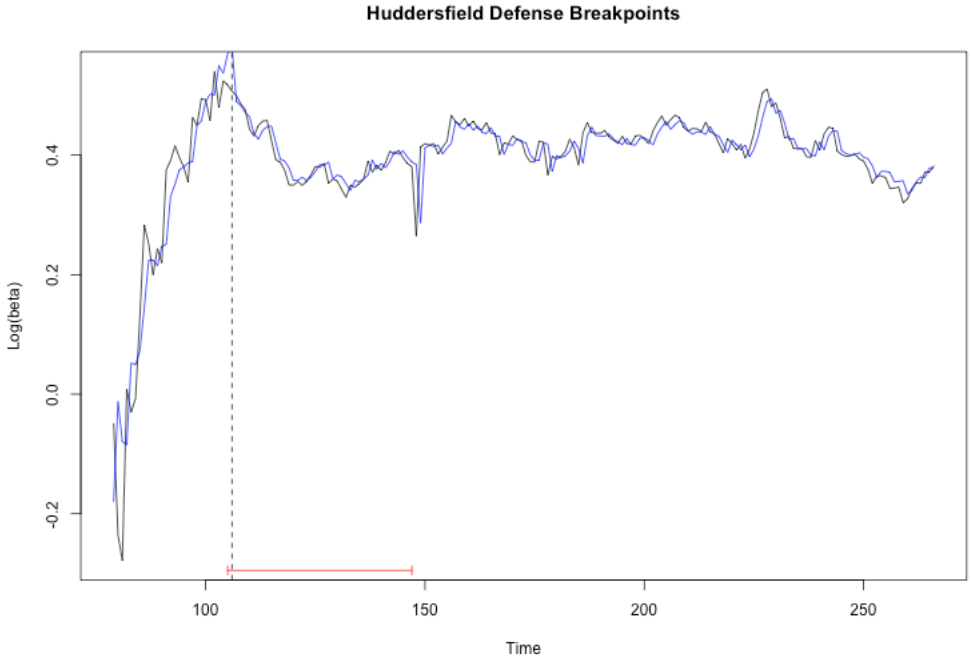
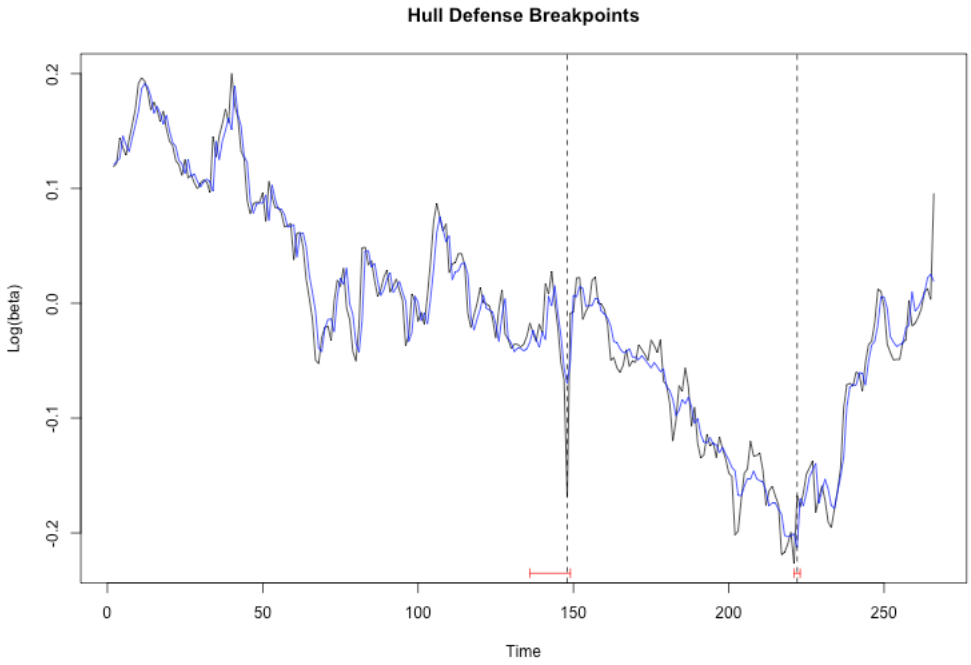


Figure 8 Continued



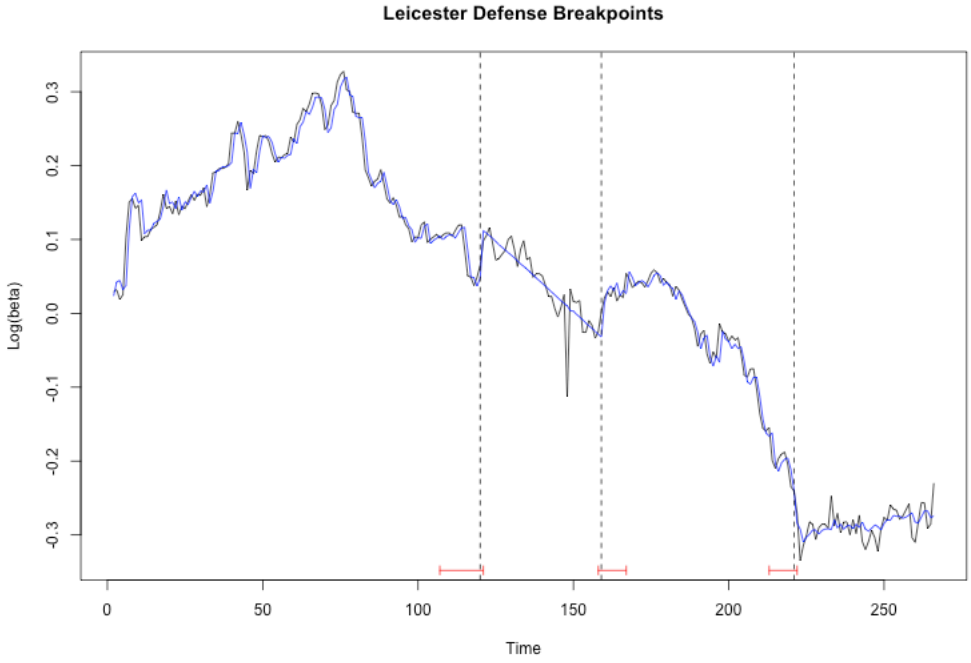
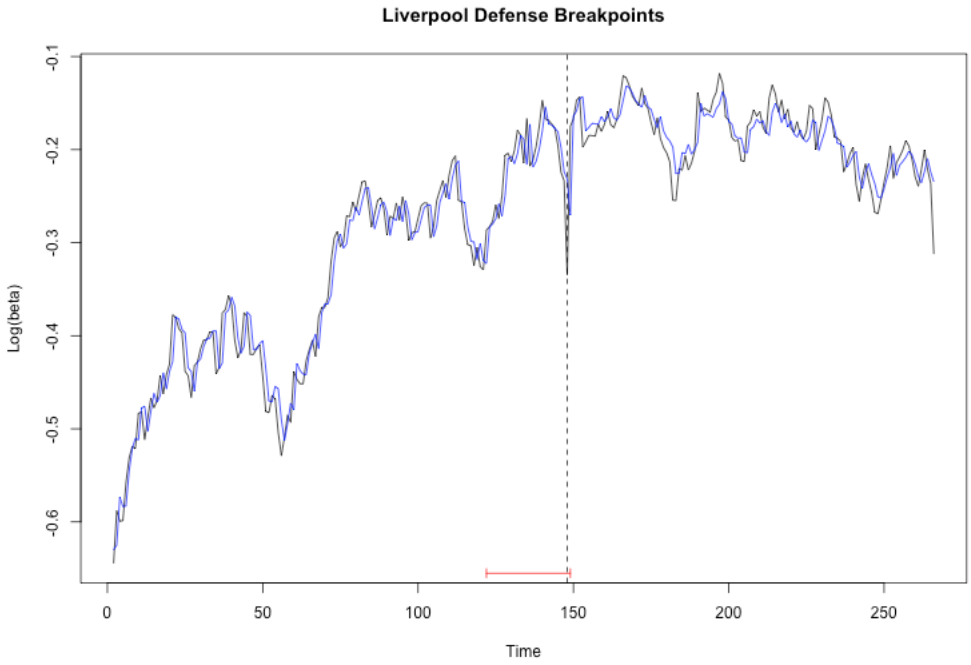


Figure 8 Continued



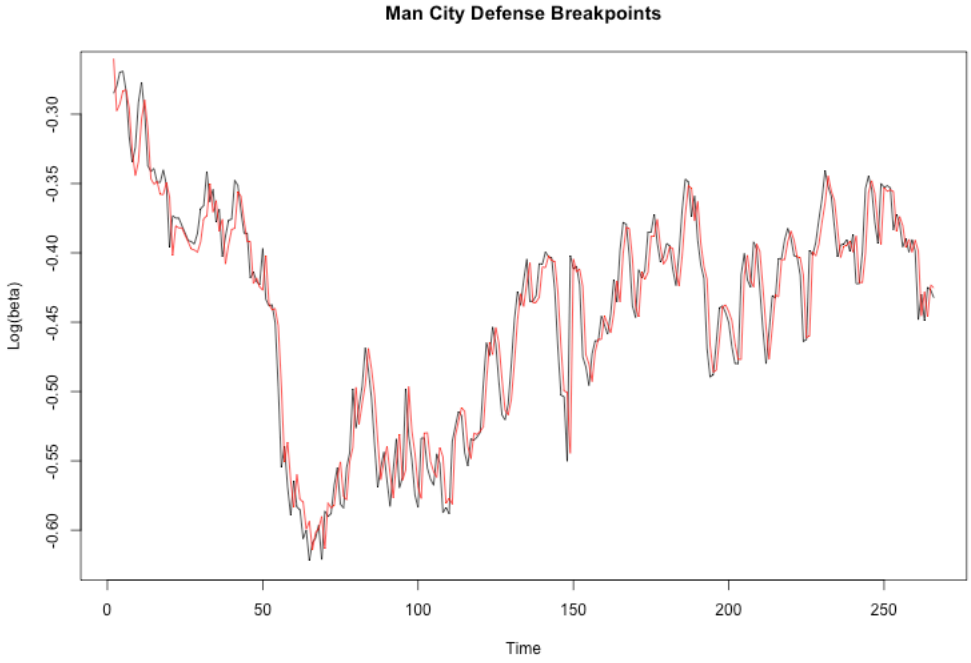
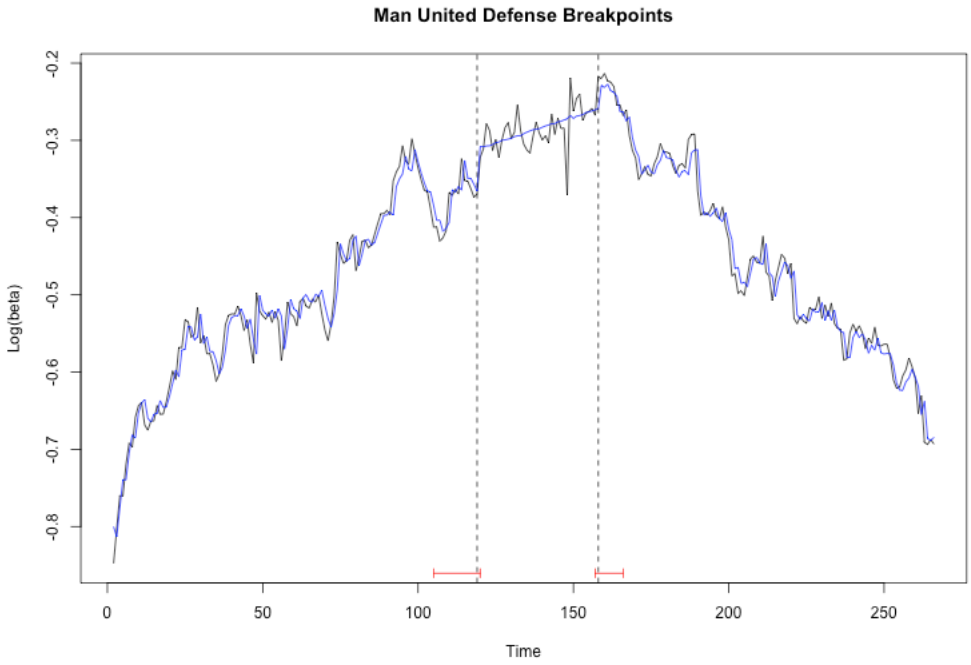


Figure 8 Continued



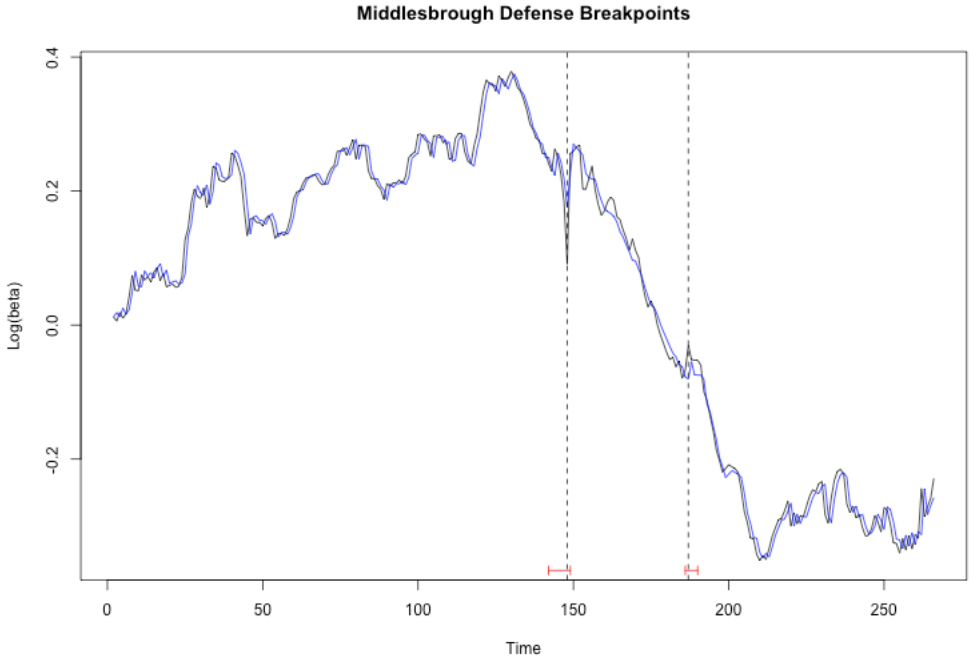
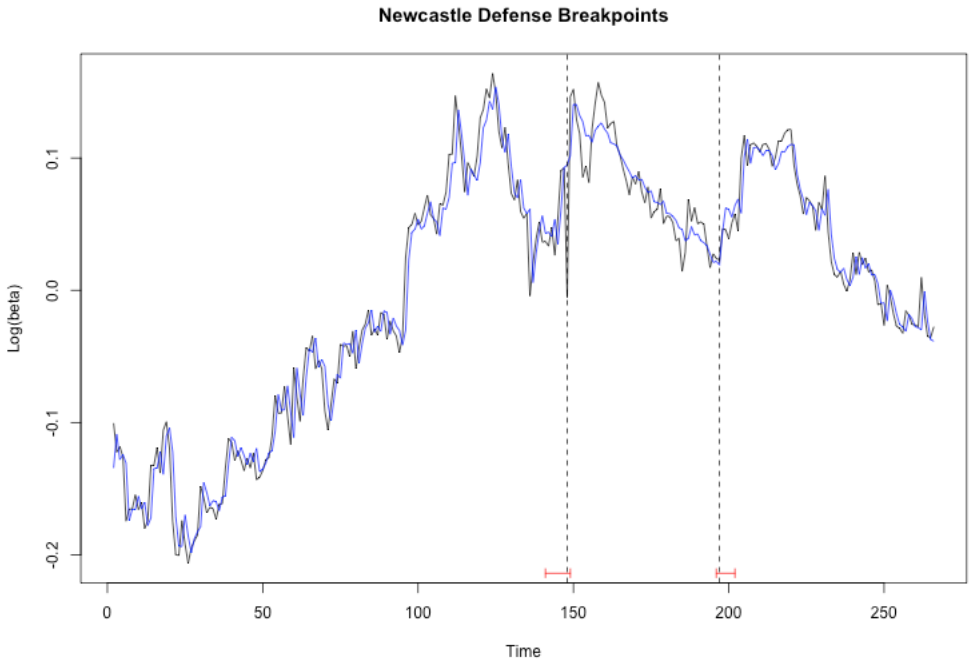


Figure 8 Continued



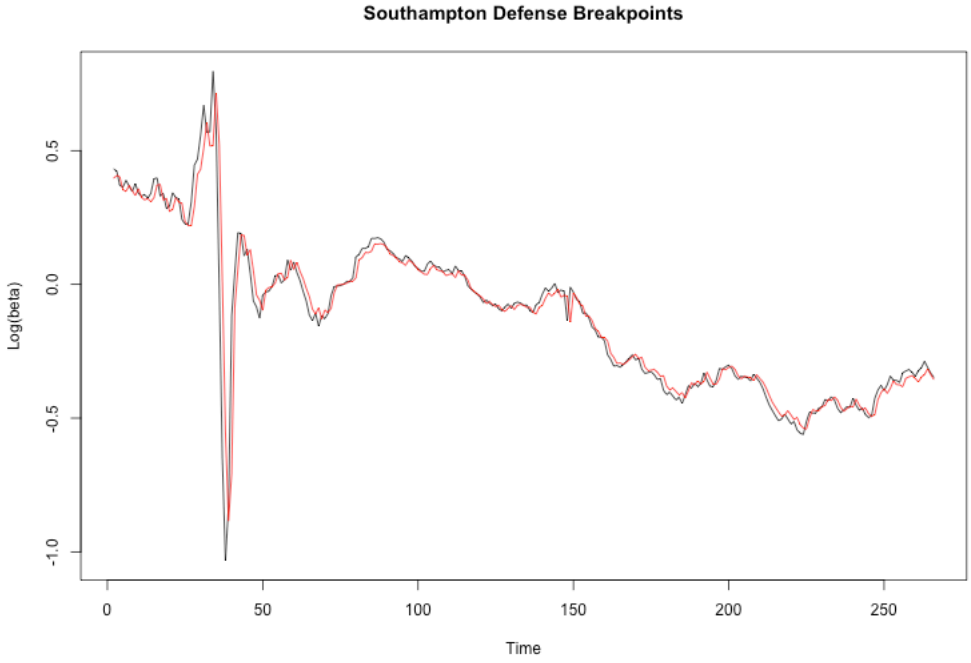
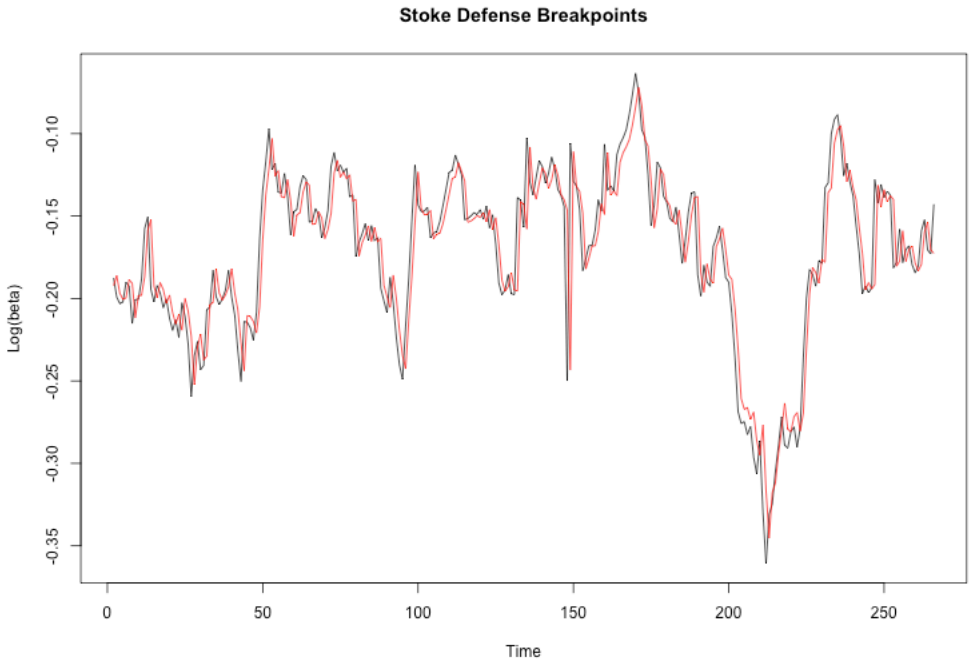


Figure 8 Continued



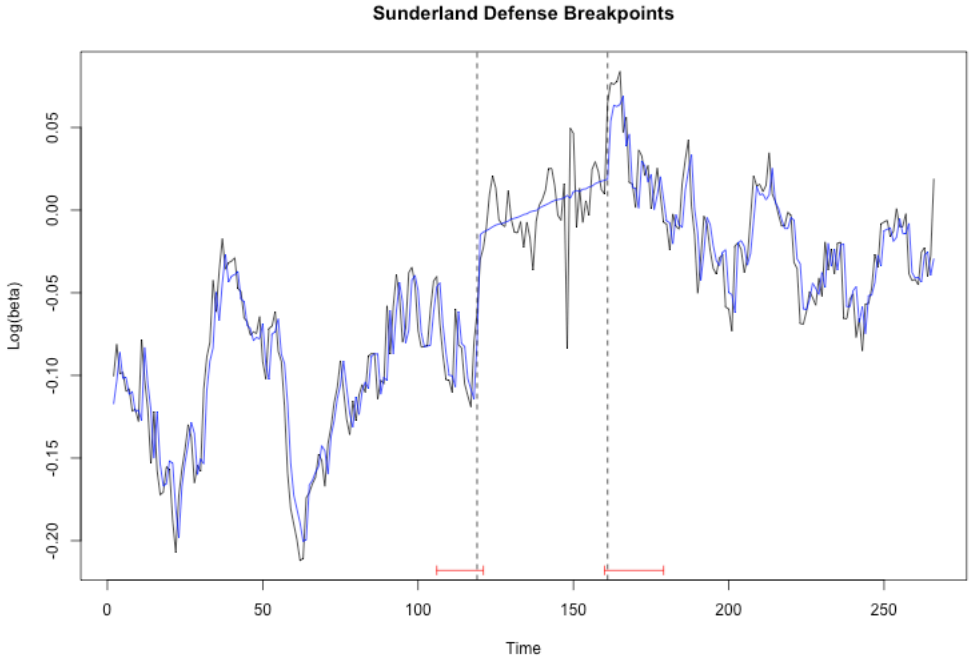
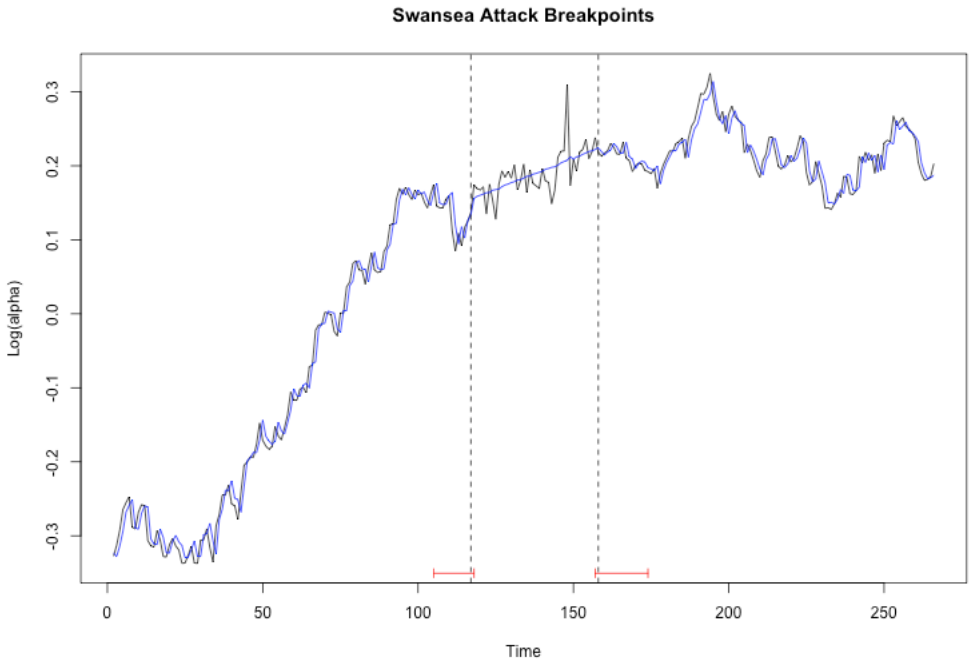


Figure 8 Continued



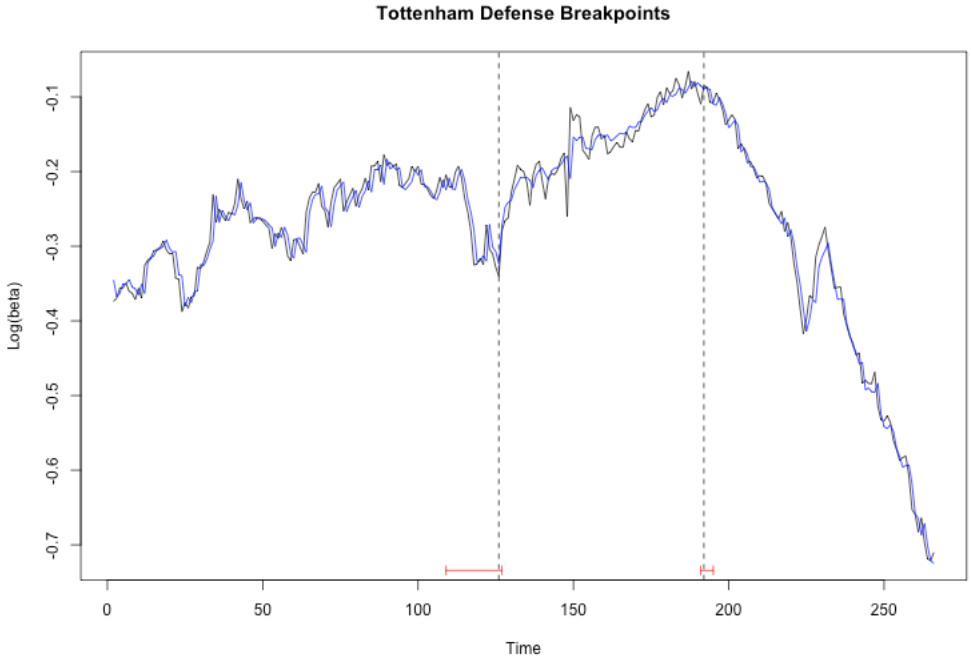
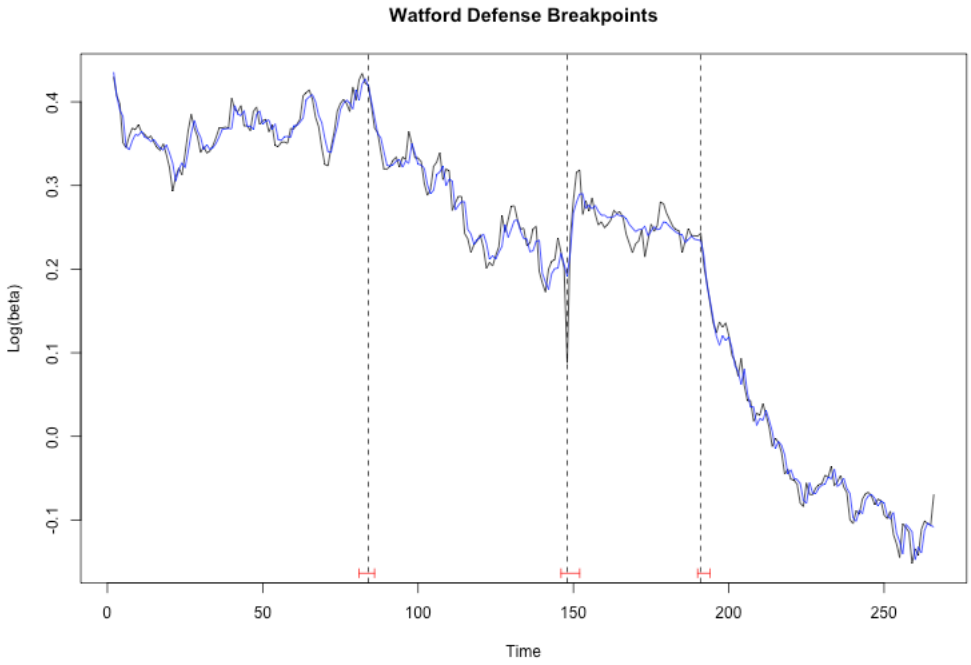


Figure 8 Continued



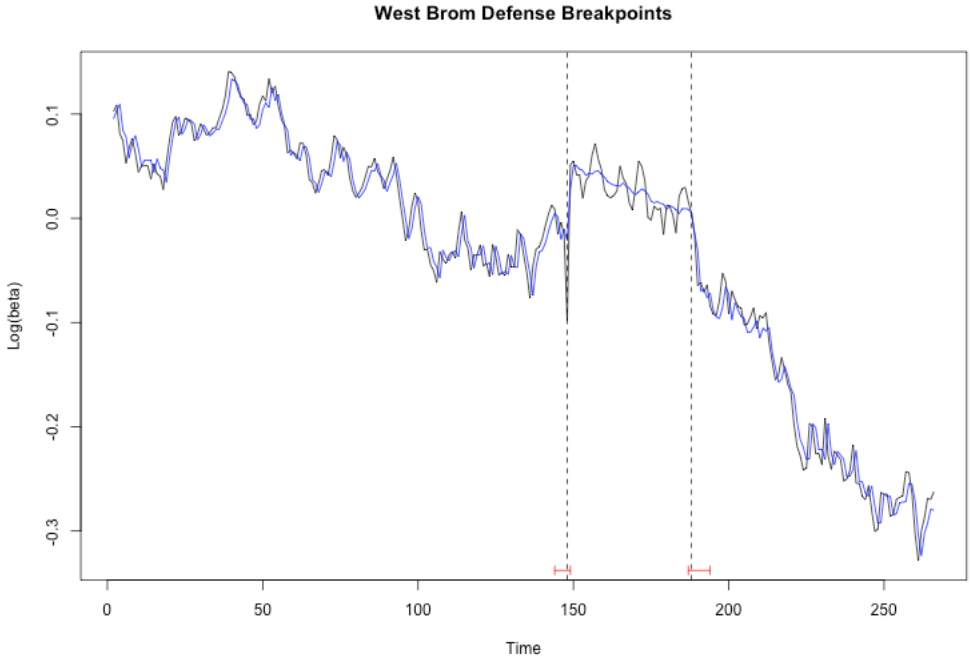


Figure 8 Continued

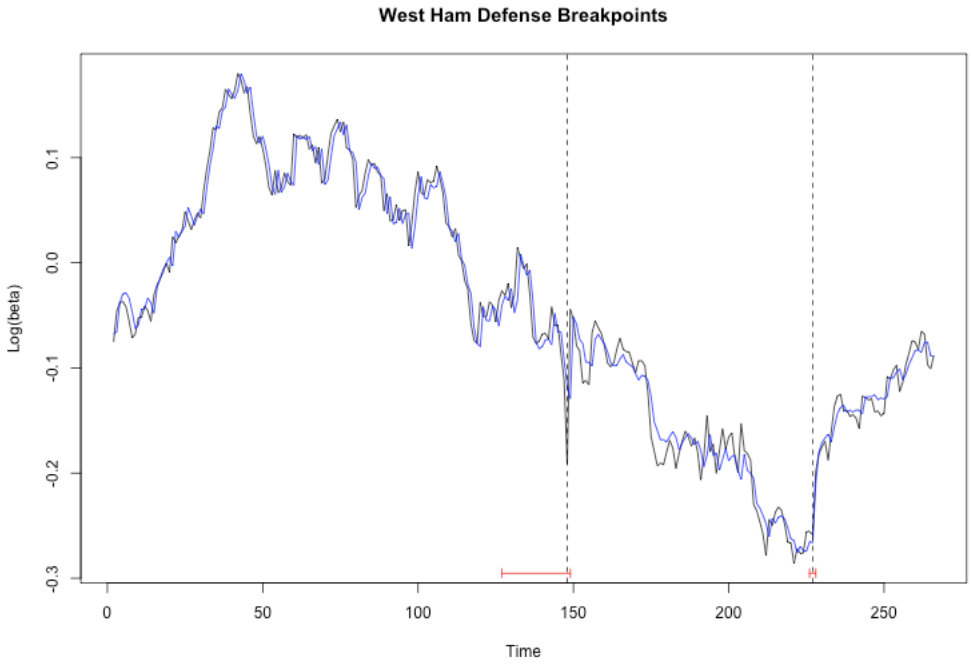


Figure 8 Continued

Table 2 - Parameter Breakpoint Estimates Attack

Team	Segment	$\hat{c}$	$\hat{\delta}$	$\hat{\phi}$
Arsenal	2 - 138	0.0369	-0.0000293	0.941
	139 - 177	0.540	-0.000455	0.137
	178 - 266	0.0719	0.000214	0.813
Bournemouth	116 - 161	-0.333	0.00214	0.572
	162 - 266	-0.0300	0.000434	0.772
Brighton	40 - 266	-0.0205	0.0000609	0.934
Burnley	2 - 266	-0.00300	0.00000615	0.944
Chelsea	2 - 119	0.0160	-0.0000451	0.975
	120 - 159	0.0468	0.00273	0.249
	160 - 266	0.0201	0.0000459	0.952
Crystal Palace	2 - 266	-0.0242	0.00015437	0.950
Everton	2 - 266	0.00986	0.0000552	0.951
Huddersfield	79 - 108	0.514	-0.00661	0.522
	109 - 266	-0.0432	0.0000489	0.805
Hull	2 - 131	-0.0350	0.000115	0.903
	132 - 266	-0.0706	0.000254	0.632
Leicester	2 - 106	-0.0148	0.000261	0.833
	107 - 149	-0.398	0.00351	0.168
	150 - 266	-0.0121	0.000125	0.966
Liverpool	2 - 147	-0.0120	0.000156	1.00
	148 - 204	0.954	-0.00223	0.220
	205 - 266	-0.0648	0.000641	0.853
Manchester City	2 - 148	0.0362	0.000189	0.918
	149 - 189	0.276	0.00267	0.0498
	190 - 266	0.280	-0.000479	0.783

Table 2 Continued

Team	Segment	$\hat{c}$	$\hat{\delta}$	$\hat{\phi}$
Manchester United	2 - 148	**0.0903**	**-0.0000371**	**0.876**
	149 - 266	0.386	-0.000860	0.554
Middlesbrough	2 - 148	-0.00918	0.0000536	0.954
	149 - 223	-0.294	0.00125	0.243
	224 - 266	0.288	-0.00141	0.519
Newcastle	2 - 148	0.0264	-0.0000646	0.919
	149 - 207	0.0168	0.000301	0.228
	208 - 266	-0.0365	0.000243	0.871
Southampton	2 - 41	-0.143	0.00505	0.741
	42 - 266	0.00977	0.0000685	0.929
Stoke	2 - 105	-0.0000774	-0.0000453	0.917
	106 - 177	-0.247	0.00153	0.603
	178 - 266	0.0395	-0.0000517	0.871
Sunderland	2 - 92	0.0132	-0.0000384	0.881
	93 - 149	0.254	-0.00202	0.267
	150 - 266	-0.00841	0.0000674	0.915
Swansea	2 - 117	-0.0224	0.000335	0.936
	118 - 158	-0.0407	0.00163	0.0338
	159 - 266	0.0291	-0.0000300	0.895
Tottenham	2 - 148	0.0451	0.0000306	0.884
	149 - 266	-0.0430	0.000816	0.754
Watford	2 - 148	-0.00260	0.0000669	0.995
	149 - 187	-0.0163	0.000947	0.198
	188 - 266	0.0411	-0.000137	0.947
West Brom	2 - 266	0.0316	-0.000101	0.885
West Ham	2 - 148	0.0164	0.000112	0.799
	149 - 187	-0.342	0.00259	0.295

Table 3 - Parameter Breakpoint Estimates Defence

Team	Segment	$\hat{c}$	$\hat{\delta}$	$\hat{\phi}$
Arsenal	2 - 266	-0.00783	-0.0000502	0.958
Bournemouth	116 - 148	3.83	-0.0234	-0.341
	149 - 190	0.910	-0.00402	0.163
	191 - 266	0.0165	-0.0000140	0.757
Brighton	40 - 73	-0.0488	0.000810	0.491
	74 - 266	0.000930	-0.0000123	0.906
Burnley	2 - 145	0.0116	-0.0000771	0.973
	146 - 266	0.0688	-0.000508	0.731
Chelsea	2 - 148	0.0241	-0.000192	1.02
	149 - 189	0.219	-0.00369	0.298
	190 - 266	-0.0265	-0.0000287	0.914
Crystal Palace	2 - 143	0.0272	-0.0000552	0.884
	144 - 182	0.243	-0.00148	0.451
	183 - 266	-0.0791	0.000269	0.899
Everton	2 - 148	-0.00821	-0.0000643	0.961
	149 - 203	-0.158	0.000200	0.574
	204 - 266	0.0783	-0.000557	0.738
Huddersfield	79 - 106	-1.063	0.0134	0.431
	107 - 266	0.0624	0.0000160	0.839
Hull	2 - 148	0.0180	-0.000191	0.893
	149 - 222	0.298	-0.00192	0.374
	223 - 266	-0.304	0.00120	0.801
Leicester	2 - 120	0.0144	-0.000114	0.959
	121 - 159	0.562	-0.00373	0.019
	160 - 221	0.111	-0.000618	0.911
	222 - 266	-0.328	0.000503	0.282
Liverpool	2 - 148	-0.0401	0.000138	0.908
	149 - 266	-0.00808	-0.000256	0.671
Man City	2 - 266	-0.0287	0.0000241	0.944

Table 3 Continued

Team	Segment	$\hat{c}$	$\hat{\delta}$	$\hat{\phi}$
Manchester United	**2 - 119**	**-0.110**	**0.000462**	**0.831**
	120 - 158	-0.487	0.00139	-0.034
	159 - 266	0.0730	-0.000830	0.781
Middlesbrough	2 - 148	0.00675	-0.000103	1.007
	149 - 187	1.20	-0.00676	0.311
	188 - 266	-0.0432	0.0000877	0.919
Newcastle	2 - 148	-0.0195	0.000213	0.895
	149 - 197	0.385	-0.00189	0.268
	198 - 266	0.132	-0.000534	0.790
Southampton	2 - 266	0.0322	-0.000368	0.874
Stoke	2 - 266	-0.0139	0.000000602	0.919
Sunderland	2 - 119	-0.0133	0.0000498	0.897
	120 - 161	-0.107	0.000779	0.023
	162 - 266	0.0123	-0.0000717	0.816
Swansea	2 - 93	-0.0403	0.000318	0.832
	94 - 152	-0.162	0.000789	0.181
	153 - 266	-0.0223	0.0000760	0.940
Tottenham	2 - 126	-0.0185	0.00000825	0.933
	127 - 192	-0.333	0.00148	0.374
	193 - 266	0.191	-0.00105	0.880
Watford	2 - 84	0.0731	0.000214	0.776
	85 - 148	0.169	-0.000822	0.718
	149 - 191	0.314	-0.000813	0.316
	192 - 266	0.0405	-0.000207	0.884
West Brom	2 - 148	0.0120	-0.000128	0.898
	149 - 188	0.199	-0.00104	0.163
	189 - 266	0.107	-0.000658	0.786
West Ham	2 - 148	0.00802	-0.000106	0.978
	149 - 227	0.169	-0.00131	0.537
	228 - 266	-0.363	0.00121	0.476

4. Forecasting

This forecasting chapter is split into four parts. The first part describes the five different models in this exercise and how the one-step ahead forecasts are calculated. The second part shows for each model the graphical representation of the point estimate of one-step ahead forecasts of team strengths for every team. In the third part we present a table with the mean-square error (MSE) to see which model performed best in this out-of-sample forecasting experiment. In the final part of this chapter, the models are evaluated in their abilities to generate match outcome probabilities according to the average RPS, which is the same measure used to calculate the optimal decay factor in Chapter 3.

For the purpose of this analysis, we assume that the estimates for attack and defence parameters derived from estimating the Dixon and Coles model in the previous Chapter are the "true" values of attack and defence and the forecasts for each model are compared to them. The procedure is as follows: a one-step ahead forecast is carried out on the training set window of 5 seasons in order to forecast the first match of the 6th season. Thus, we use $t = 1, \dots, 190$ in order to generate $\widehat{\alpha_{191}}$ and $\widehat{\beta_{191}}$. Then, the window is moved by 1 observation and $t = 2, \dots, 191$ is used to generate $\widehat{\alpha_{192}}$ and $\widehat{\beta_{192}}$ up until $\widehat{\alpha_{266}}$ and $\widehat{\beta_{266}}$. In all four models, the dependent variable y_t refers to the attack strength α_t and the defence strength β_t.

Model 1 – Random walk

$$y_t = y_{t-1} + \varepsilon_t \tag{29}$$

where $\varepsilon_t \sim N(0, \sigma^2)$. For this model the forecast is simply the last observation:

$$\hat{y}_{T+1|T} = y_T \tag{30}$$

Model 2 – Random walk with drift

$$y_t = c + y_{t-1} + \varepsilon_t \tag{31}$$

where $\varepsilon_t \sim N(0, \sigma^2)$ as in the random walk model, and c is the drift parameter. The forecast is very similar to that for the random walk model:

$$\hat{y}_{T+1|T} = \hat{c} + y_T \tag{32}$$

Model 3 – Auto-ARIMA

An *ARIMA(p,d,q)* process (Box and Jenkins, 1970) for time series y_t can be represented as

$$\left(1 - \sum_{i=1}^{p} \phi_i L^i\right)\Delta^d y_t = \left(1 - \sum_{i=1}^{q} \theta_i L^i\right)\varepsilon_t \quad (33)$$

where L is the lag operator, ϕ_i the parameters from the autoregressive component of the model for $i = 1, \dots, p$, θ_i the moving average parameters $i = 1, \dots, q$. The order of integration is represented by d, with Δ^d being the number of times the process needs to be differenced to make it stationary.

A rolling window for forecasting allows for the lags of the autoregressive and moving average parts of the model to vary and therefore addresses the problem of structural breaks. Different *ARIMA(p,d,q)* models were estimated and the optimal values of *p, d* and *q* chosen according to the BIC before generating the one-step ahead forecasts. The BIC is used instead of the AIC since it penalises the number of parameters more heavily. The "forecast" package in R was used here.

Model 4 – Time-varying parameter model

The time-varying parameter (TVP) model (Durbin and Koopman, 2001) specifies that the state variables (the parameters that determine the team strengths) follow a random walk. Thus, it is as if a break is present at every observation. The one-step ahead forecast is carried out through the Kalman Filter for the best estimate of the team strength. The parameters are estimated through the Kalman Smoothing procedure.

In a general state-space framework:

$$y_t = F_t\theta_t + v_t$$

$$\theta_t = G\theta_{t-1} + w_t$$

where $v_t \sim N(0, \sigma_v^2)$, $w_t \sim N(0, \sigma_w^2)$. The F_t and G matrices specify the transitions between state and observed variable, and the dynamics of the state variables respectively. In this scenario, we have 2 states and one observed variable:

$$y_t = c_t + \phi_t y_{t-1} + \varepsilon_t \quad (34)$$

This means that the state dynamics are

$$c_t = c_{t-1} + \eta_{ct}$$
$$\phi_t = \phi_{t-1} + \eta_{\phi t} \tag{35}$$

where $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$ and $(\eta_{ct}, \eta_{\delta t}, \eta_{\phi t}) \sim \mathcal{N}(0, diag(\sigma_c^2, \sigma_\delta^2, \sigma_\phi^2))$. Thus

$$F_t\theta_t + v_t = (1 \quad y_{t-1}) \begin{pmatrix} c_t \\ \phi_t \end{pmatrix} + v_t$$

$$\theta_t = \begin{pmatrix} c_t \\ \phi_t \end{pmatrix}, G = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \theta_{t-1} = \begin{pmatrix} c_{t-1} \\ \phi_{t-1} \end{pmatrix}, w_t = \begin{pmatrix} \eta_{ct} \\ \eta_{\phi t} \end{pmatrix}$$

The Kalman Filter is an algorithm that calculates the moments of the state vector θ_{t+1}, assumed to be normally distributed, conditioned on the observed data $Y_t = (y_1, \dots, y_t)$ and the state model parameters. The updating equations compute the prediction error and the predicted error variance. The Kalman Smoothing is a backward recursion based on the full set of observations that calculates the mean and variance of specific conditional distributions to thus provide estimates of the state vector and its variance matrix. The model was estimated using the "dlm" package in R, using in-built functions such as "dlmMLE" to obtain the maximum likelihood estimates of the parameters, "dlmFilter" for the Kalman Filter based on the optimal estimates from "dlmMLE", and "dlmSmooth" for the Kalman Smoothing estimates of the state vector.

The TVP model allows for extra flexibility when accounting for structural breaks. Depending on the smoothed estimates of the state variables and the sizes of the variances of the errors, they can capture a structural break at every observation (in which case, the state variables follow a random walk process), or they can have a break at particular point in time, or the state variable estimates could remain constant throughout the duration of the sample.

The forecast $\hat{y}_{T+1|T}$ is carried out by applying the updating equations of the Kalman Filter to obtain the optimal estimator of the state vector $\hat{\theta}_{T+1|T}$.

$$\hat{\theta}_{T+1|T} = G\theta_T$$
$$\hat{y}_{T+1|T} = F_t\hat{\theta}_{T+1|T} \tag{36}$$

Model 5 – Equal weights model averaging

Equal weights model averaging is a very simple method of addressing the problem of forecasting structural breaks out-of-sample. It gives equal weights for all the different models we consider and takes an average of their respective forecasts. Let our vector of weights be $\boldsymbol{w} = (w_1, \dots, w_N)$, where N is the number of models we are averaging, and our vector of one-step ahead forecast estimates be $\hat{\boldsymbol{y}}_{\boldsymbol{T+1|T}} = (\hat{y}_{1\,T+1|T}, \dots, \hat{y}_{N\,T+1|T})$ where $\hat{y}_{1\,T+1|T}$ is the one-step ahead forecast of model 1 and $\hat{y}_{N\,T+1|T}$. In our case, we have four models that we are averaging, thus $w_1, \dots, w_N = 0.25$. Therefore the one-step ahead forecast of model 5 is:

$$\hat{y}_{T+1|T} = \sum_{i=1}^{4} w_i \, \hat{y}_{i\,T+1|T} \tag{37}$$

Rossi (2013) advocates the use of equal weights model averaging saying it can sometimes improve its forecasting performance.

One-step ahead forecast of team strengths

Figure 9 to Figure 16 show the one-step ahead forecasts for the first four models for Arsenal's attack and defence. The rest of the teams can be seen in Appendix A (Figure 18 and Figure 19) to Appendix D (Figure 24 and Figure 25). The black plots represent the actual parameter estimates from the Dixon and Coles model estimated in the previous chapter. The blue lines represent the one-step ahead point forecasts, and the two red lines are the upper and lower 95% confidence intervals. Regarding Figure 24 and Figure 25 (the TVP model), the red dotted line overlaid on the black plots represent the one-step ahead forecast from the Kalman Filter in-sample. Forecast for model 5 have not been shown in this thesis because the upper and lower confidence intervals are very large. Since it is model average, the average of the variances of all four models also need the covariance to be taken into account. This in turn inflates the 95% confidence intervals and for several plots, they lie outside of the graph range (we could have increased the graph range in order to show how large the confidence intervals were, but we wished to keep the ranges uniform for all forecast graphs in order to ease comparison between different figures for the reader).

Graphically, there doesn't seem to be much difference between the simple random walk (model 1) and the random walk with drift (model 2) forecasts. The 95% confidence intervals of the auto-arima (model 3) forecasts are slightly narrower than the model 1 and 2 counterparts. This is to be expected though: given the parameters estimated from the breakpoint analysis in Table 2 and Table 3, only a couple of segments have an autoregressive parameter $\hat{\phi} \cong 1$ (which would indicate model 1 or model 2). However, the graphs for the TVP (model 4) forecasts look like the best fit, with the 95% confidence intervals much narrower than any of the other models.

The graphs for the estimated smoothed states for each team (attack and defence) were calculated through the Kalman Smoothing algorithm and can be seen for all teams in Appendix E (Figure 26 and Figure 27). They illustrate, from left to right on the graphs, how the estimates of the intercept $\hat{c}_t$, and the autoregressive parameter $\hat{\phi}_t$ behave over time. After discarding the initial 5 observations as a burn-in period, we can have a better look at the fluctuations for the estimates. The black plots show the estimates and the red dotted lines are their upper and lower 95% confidence intervals.

For almost every team, $\hat{c}_t$ fluctuates considerably and looks like a random walk. The only exceptions are Arsenal's and West Ham's attack series and Liverpool's defence series. Meanwhile, for the majority of teams, $\hat{\phi}_t$ seems to fluctuate too. For the defence series, however, only Crystal Palace and Watford do not experience fluctuations in the estimates of $\hat{\phi}_t$. Additionally, almost every team has $\hat{\phi}_t < 1$ for all values of t, with Brighton's and Southampton's defences being the only exceptions. This suggests that both teams have experienced considerable growth in those aspects of their game.

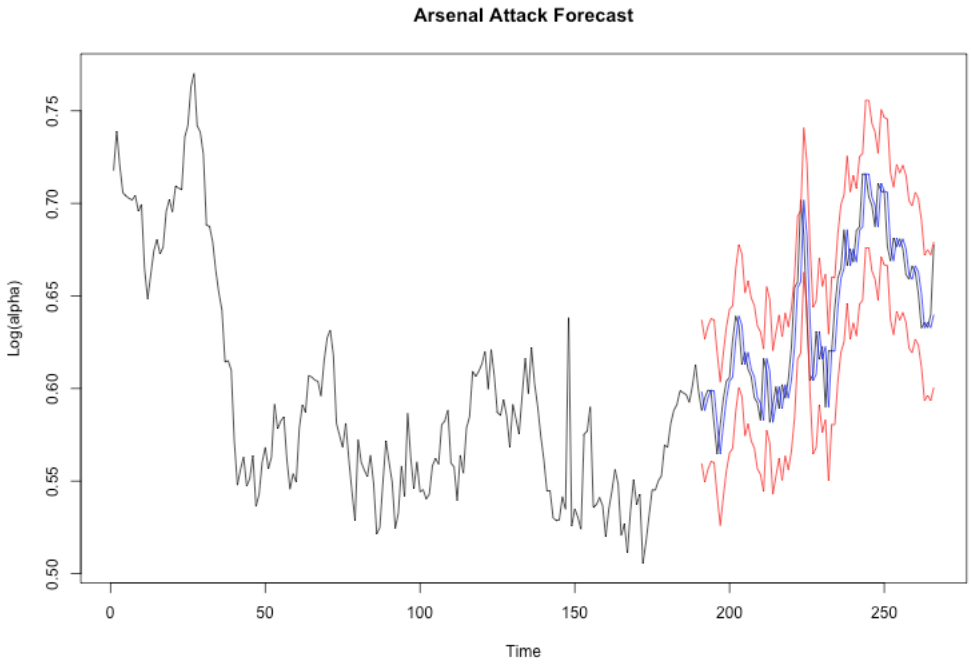


Figure 9 - (Model 1) Simple Random Walk Forecasts for Arsenal – Attack

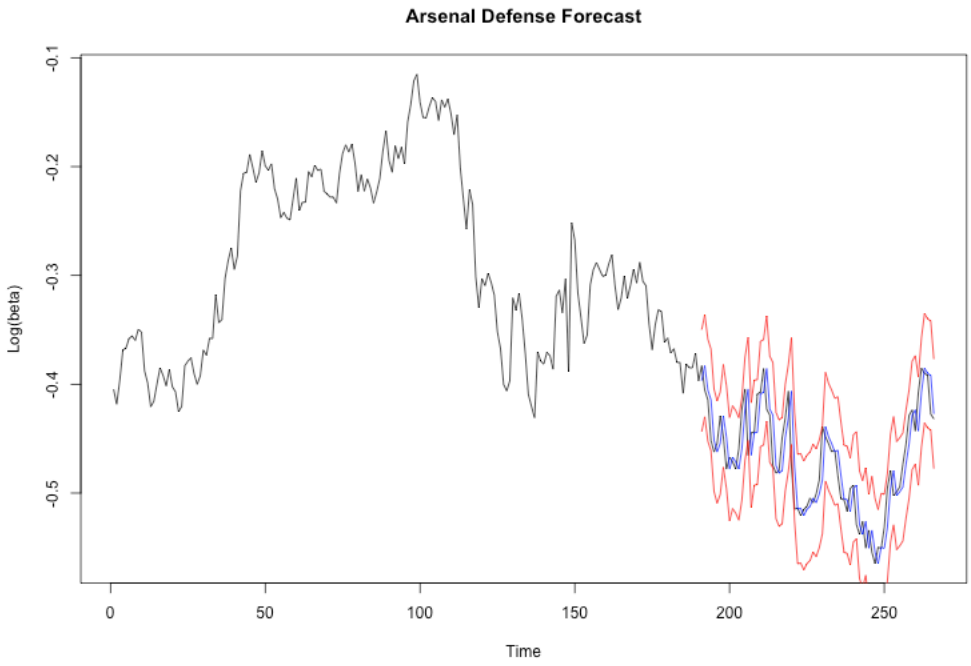


Figure 10 - (Model 1) Simple Random Walk Forecasts for Arsenal – Defence

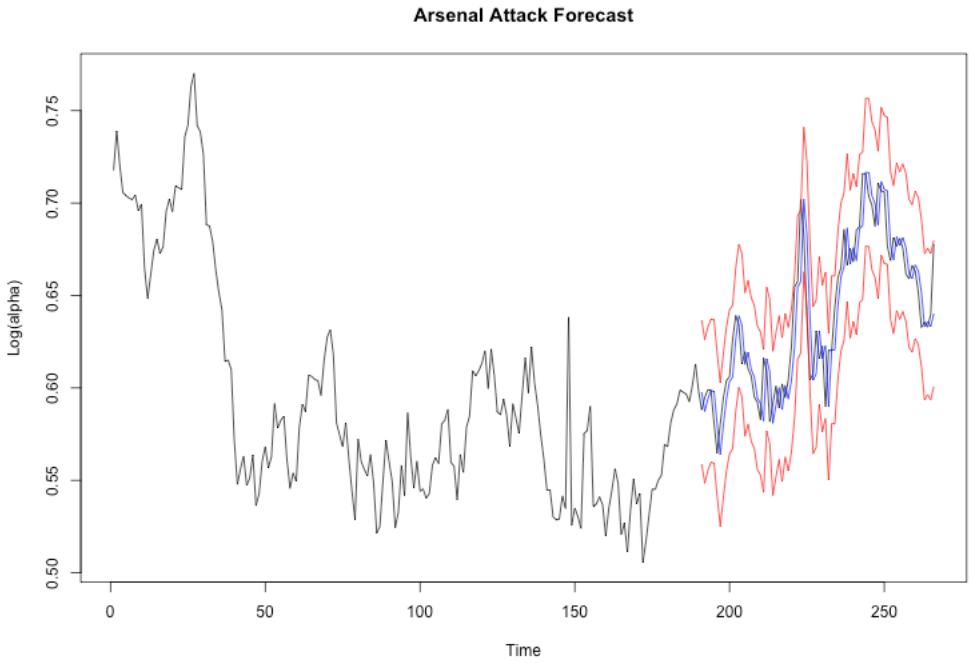


Figure 11 - (Model 2) Random Walk With Drift Forecasts for Arsenal – Attack

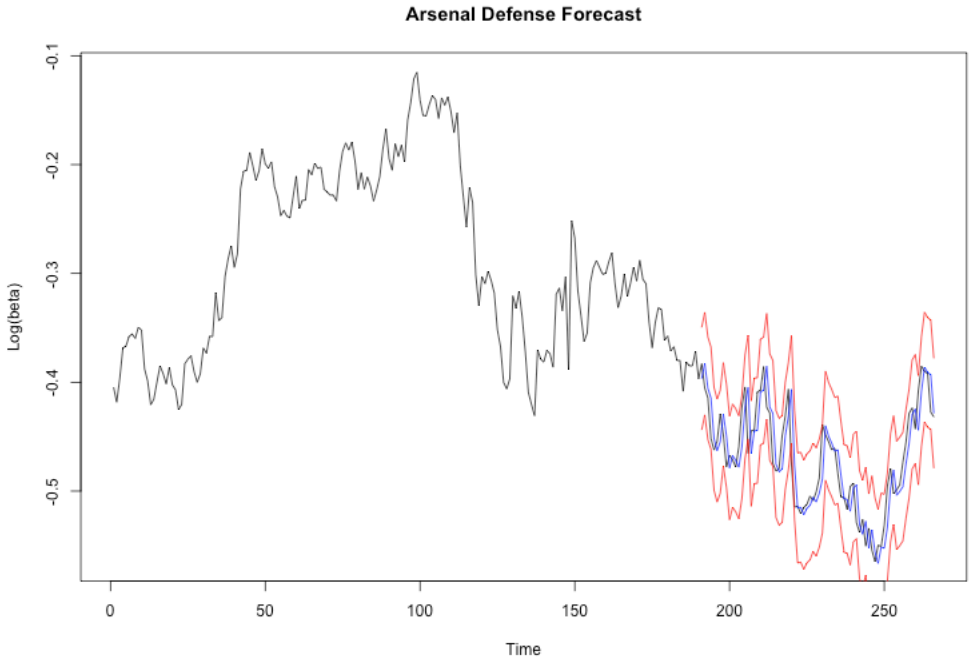


Figure 12 - (Model 2) Random Walk With Drift Forecasts for Arsenal - Defence

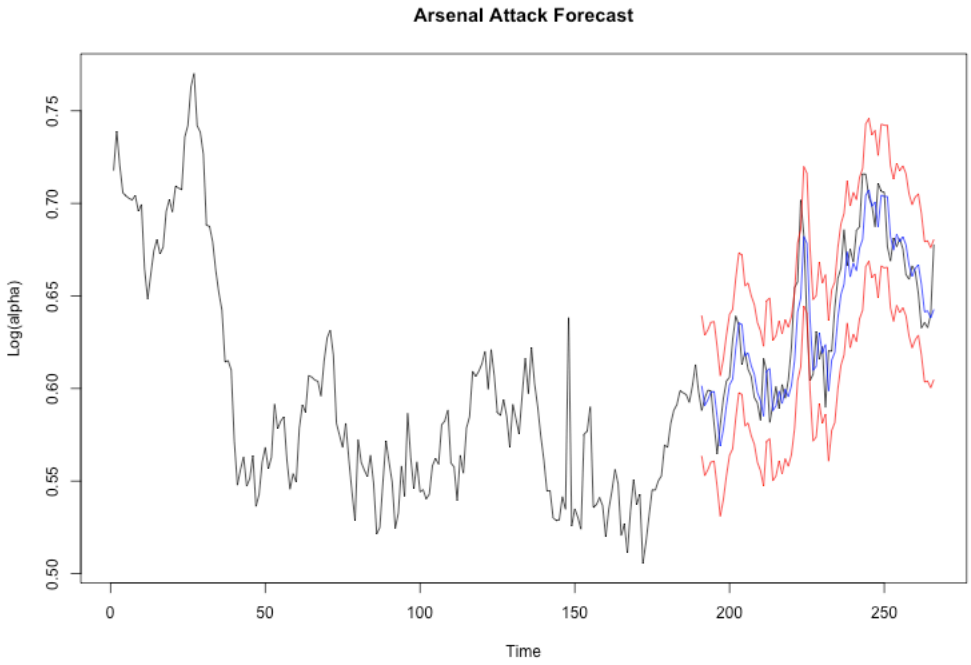


Figure 13 - (Model 3) Auto-Arima Forecasts for Arsenal – Attack

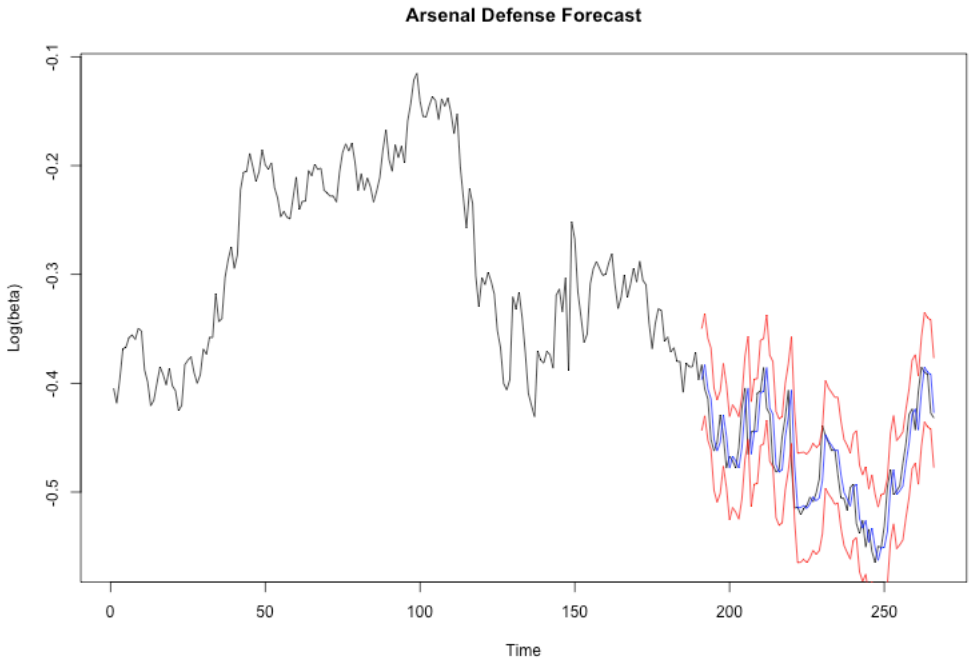


Figure 14 - (Model 3) Auto-Arima Forecasts for Arsenal – Defence

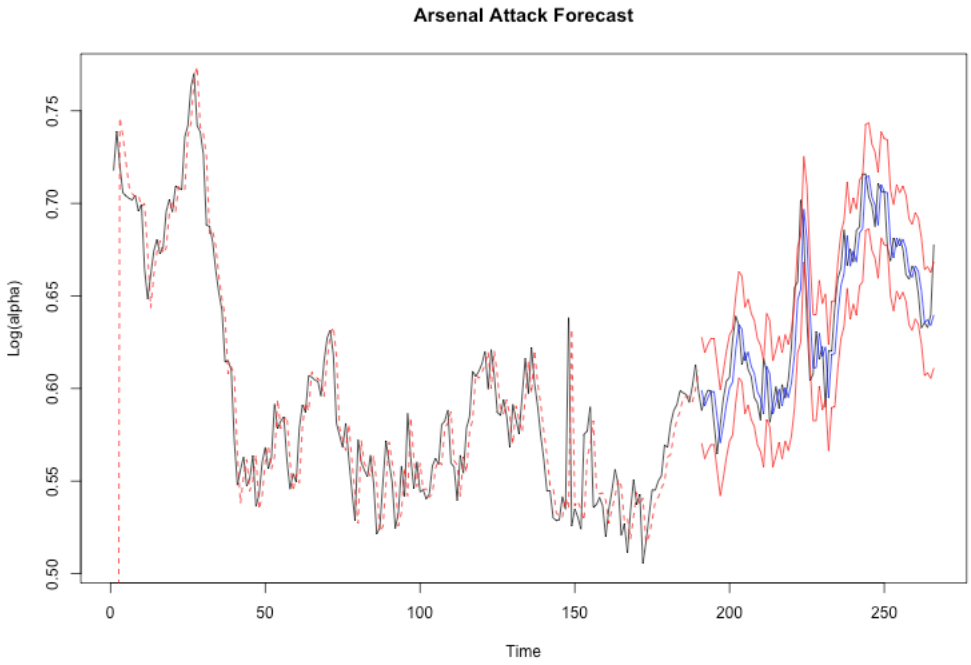


Figure 15 - (Model 4) TVP Forecasts using Kalman Filter for Arsenal – Attack

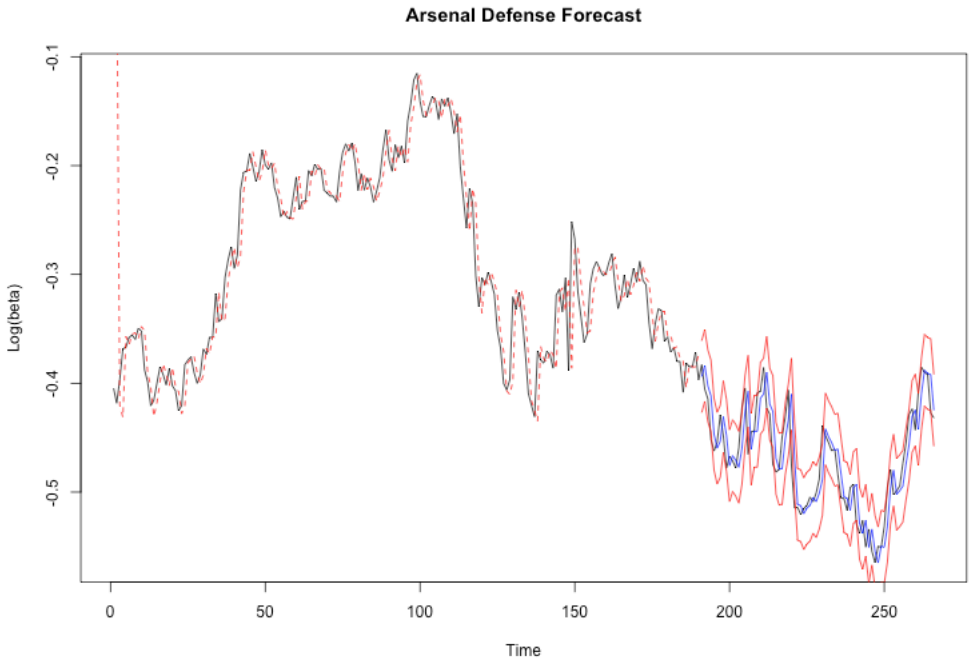


Figure 16 - (Model 4) TVP Forecasts using Kalman Filter for Arsenal – Defence

Mean Square Error (MSE) comparison

The MSE is calculated by the average of the squared errors between the forecast and the actual observations. Its formula is given by

$$MSE = \frac{1}{T}\sum_{t}^{T}(Y_t - \widehat{Y_t})^2$$

where T is the total number of forecasts in the sample, t is the first index of the out-of-sample forecast, Y_j is the actual observation and $\widehat{Y_t}$ is our estimate.

Table 4 and Table 5 demonstrate the values for the out-of-sample forecast MSE for all five models and we come up with interesting results. The best performing models according to MSE seems to be the simple random walk forecast and the model averaging. For the team attack strengths, the simple random walk forecasts has the lowest MSE for 7 teams, compared with 6 teams for the model averaging forecasts. The TVP model forecasts attack strengths the worst according to this criterion. The results for the defence time series are similar for the top two forecasting models, with model 1 having the lowest MSE for 8 teams, and model 5 for 7 teams. Model 4 performs better in forecasting team defence than attack, having the lowest MSE for 4 teams. This is surprising given the fact that the simple random walk model is actually a special case of the TVP model, where $c_t = 0$ and $\phi_t = 0$ for all values of t.

The values in those two tables have been rounded to 6 decimal places and the model with the lowest MSE for every team has been highlighted in bold.

Table 4 - Mean Square Error Forecast - Attack

Team	Model 1	Model 2	Model 3	Model 4	Model 5
Arsenal	**0.000290**	0.000292	0.000318	0.000294	0.000293
Bournemouth	0.000284	**0.000278**	0.000305	0.000285	0.000278
Brighton	0.000259	0.000263	**0.000258**	0.000264	0.000260
Burnley	0.000271	0.000273	**0.000263**	0.000271	0.000267
Chelsea	**0.000249**	0.000249	0.000260	0.000261	0.000252
Crystal Palace	**0.000279**	0.000282	0.000309	0.000294	0.000287
Everton	0.000291	**0.000290**	0.000292	0.000303	0.000293
Huddersfield	0.000323	0.000327	0.000350	0.000331	**0.000322**
Hull	0.000365	0.000367	**0.000349**	0.000358	0.000349
Leicester	0.000242	**0.000233**	0.000265	0.000266	0.000246
Liverpool	0.000316	0.000314	0.000324	0.000314	**0.000313**
Man City	**0.000191**	0.000201	0.000220	0.000198	0.000199
Man United	0.000267	0.000265	**0.000241**	0.000265	0.000254
Middlesbrough	0.000273	0.000274	0.000267	0.000269	**0.000264**
Newcastle	0.000559	0.000563	0.000575	0.000558	**0.000554**
Southampton	**0.000313**	0.000314	0.000366	0.000314	0.000319
Stoke	0.000281	0.000286	0.000283	**0.000278**	0.000280
Sunderland	0.000404	0.000407	0.000415	0.000406	**0.000401**
Swansea	**0.000318**	0.000331	0.000366	0.000335	0.000332
Tottenham	0.000371	**0.000365**	0.000435	0.000383	0.000382
Watford	**0.000269**	0.000273	0.000283	0.000285	0.000274
West Brom	0.000557	0.000560	0.000579	0.000571	**0.000556**
West Ham	0.000317	**0.000313**	0.000322	0.000330	0.000318

Table 5 - Mean Square Error Forecast - Defence

Team	Model 1	Model 2	Model 3	Model 4	Model 5
Arsenal	**0.000513**	0.000517	0.000517	0.000521	0.000515
Bournemouth	0.000376	0.000368	0.000370	0.000429	**0.000366**
Brighton	**0.000351**	0.000360	0.000404	0.000450	0.000366
Burnley	**0.000343**	0.000350	0.000367	0.000348	0.000349
Chelsea	0.000435	**0.000431**	0.000462	0.000437	0.000439
Crystal Palace	0.000332	0.000342	0.000340	**0.000330**	0.000335
Everton	0.000387	0.000388	**0.000377**	0.000397	0.000385
Huddersfield	**0.000232**	0.000239	0.000376	0.000255	0.000259
Hull	0.000471	0.000476	0.000504	0.000490	0.000482
Leicester	**0.000479**	0.000482	0.000493	0.000501	0.000481
Liverpool	0.000390	0.000398	0.000399	**0.000381**	0.000385
Man City	0.000485	0.000489	0.000484	0.000486	**0.000483**
Man United	**0.000453**	0.000459	0.000465	0.000480	0.000459
Middlesbrough	0.000446	0.000452	0.000455	0.000449	**0.000444**
Newcastle	0.000231	0.000235	0.000230	0.000240	**0.000230**
Southampton	0.000461	0.000469	0.000673	**0.000438**	0.000467
Stoke	**0.000379**	0.000380	0.000410	0.000385	0.000383
Sunderland	0.000307	0.000308	0.000303	0.000307	**0.000302**
Swansea	**0.000317**	0.000317	0.000356	0.000355	0.000326
Tottenham	0.000410	0.000404	0.000431	0.000413	**0.000399**
Watford	0.000280	**0.000274**	0.000307	0.000300	0.000287
West Brom	0.000323	0.000320	0.000322	0.000335	**0.000319**
West Ham	0.000348	0.000357	0.000375	**0.000347**	0.000355

Forecasting match outcome probabilities

In this subsection we derive match outcome probabilities from the one-step ahead out-of-sample forecasts of team strengths from the five models. This means we generated 5 different sets of match outcome probabilities (one for each model) for 760 matches in seasons 2015-16 and 2016-17. Firstly, the model predictions are evaluated graphically in Figure 17 through the use of reliability plots. Subsequently, we evaluate their performances numerically using the average RPS (Constantinou and Fenton, 2012), broken down into different out-of-sample forecast segments. This can be seen in Table 6.

Reliability plots for out-of-sample match outcome predictions

We created reliability plots in order to examine how close our generated probabilities are to the empirical probabilities of observing each match outcome: home win, draw, or away win. These can be seen in Figure 17. A perfectly calibrated curve would lie on the $y = x$ line, which means that each predicted probability "bins" of the model coincide with the empirical frequency with which we observe those events. When the curve lies below the $y = x$, we have under-estimated the probability according to the dataset observed, whilst the opposite is true if the curve lies above.

All models seem to slightly overestimate the probability of the home team winning the match, but underestimate draws and home loss. However, the curves in model 5 seems to lie the closest to the $y = x$ line.

Below each reliability plot is a histogram of the predicted probabilities which shows the frequency of each probability band for our forecasts. They are all very similar for all the models. Frequencies above a predicted probability above 0.4 for a draw are negligible, so the reliability plots above that predicted probability value do not matter too much. The same can be said about the high probability bin in the "home lose" graph of all the models. However, the fact that all models under-predict the probability of the draw is indicative that the Poisson distribution together with the Dixon and Coles (1997) low-score adjustments are arguably not an adequate distribution for modelling scores and further research should be carried out in this area concerning count distributions.

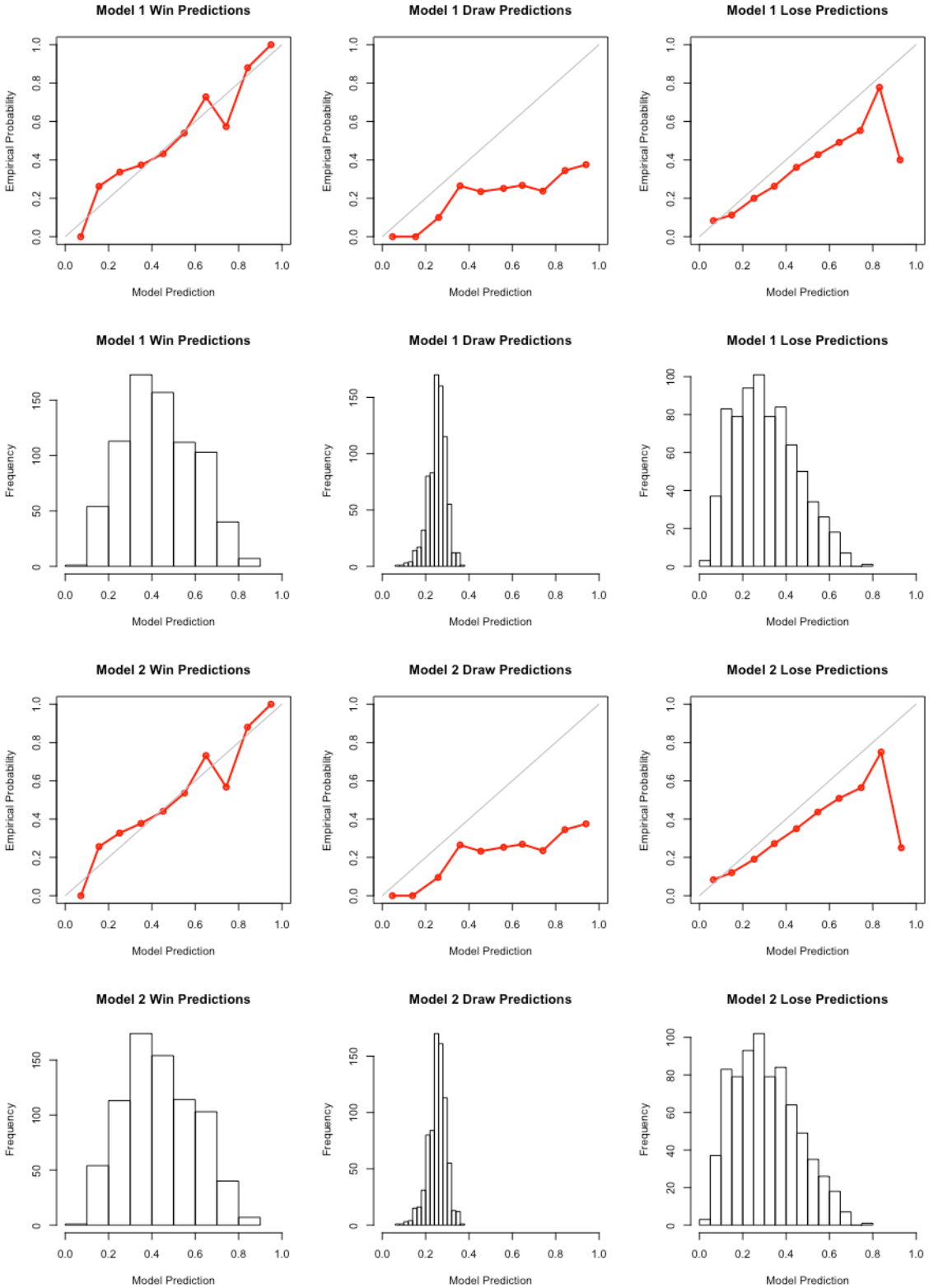


Figure 17 - Reliability Plots and Histogram of Model Forecasts

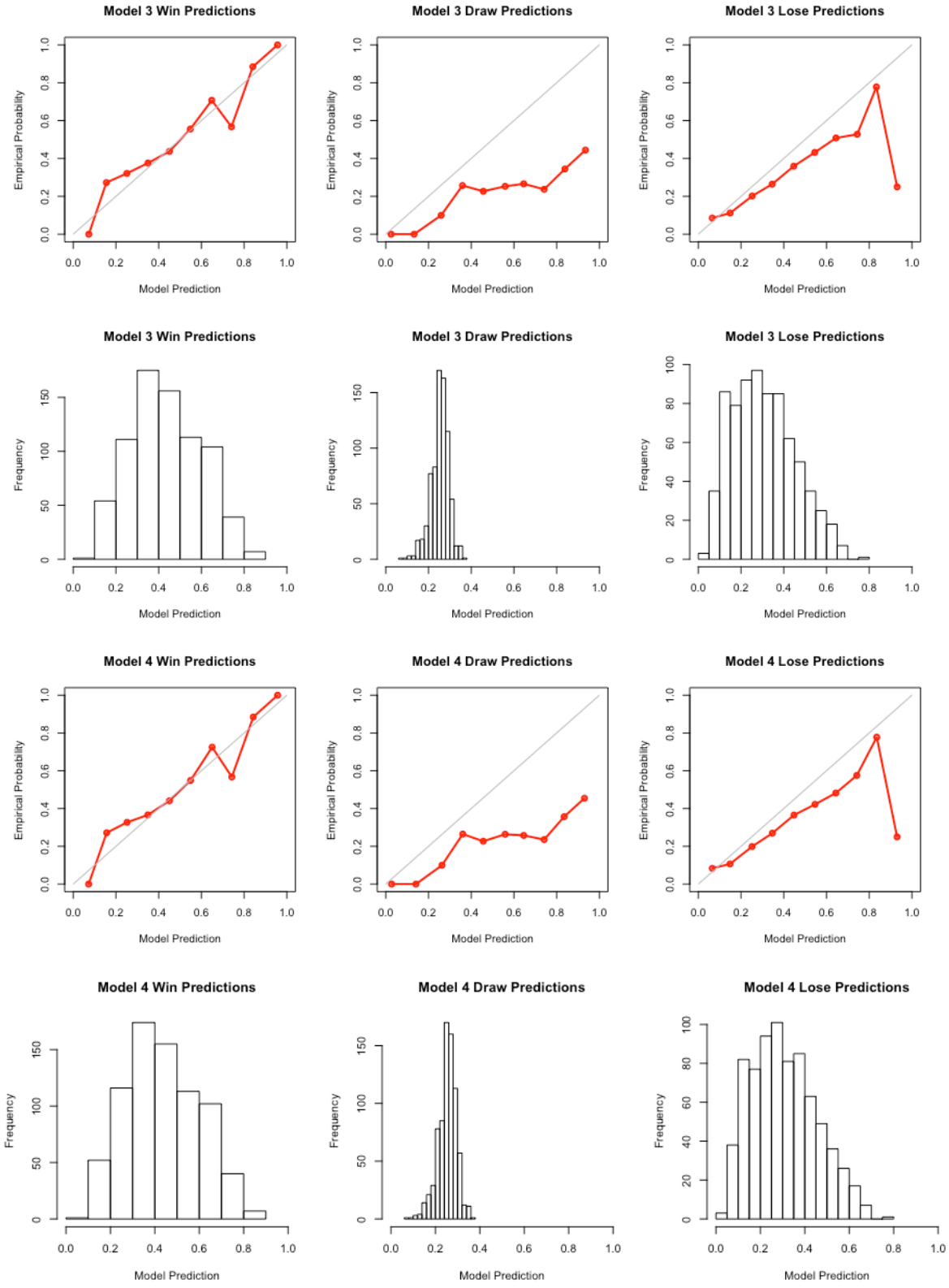


Figure 17 Continued

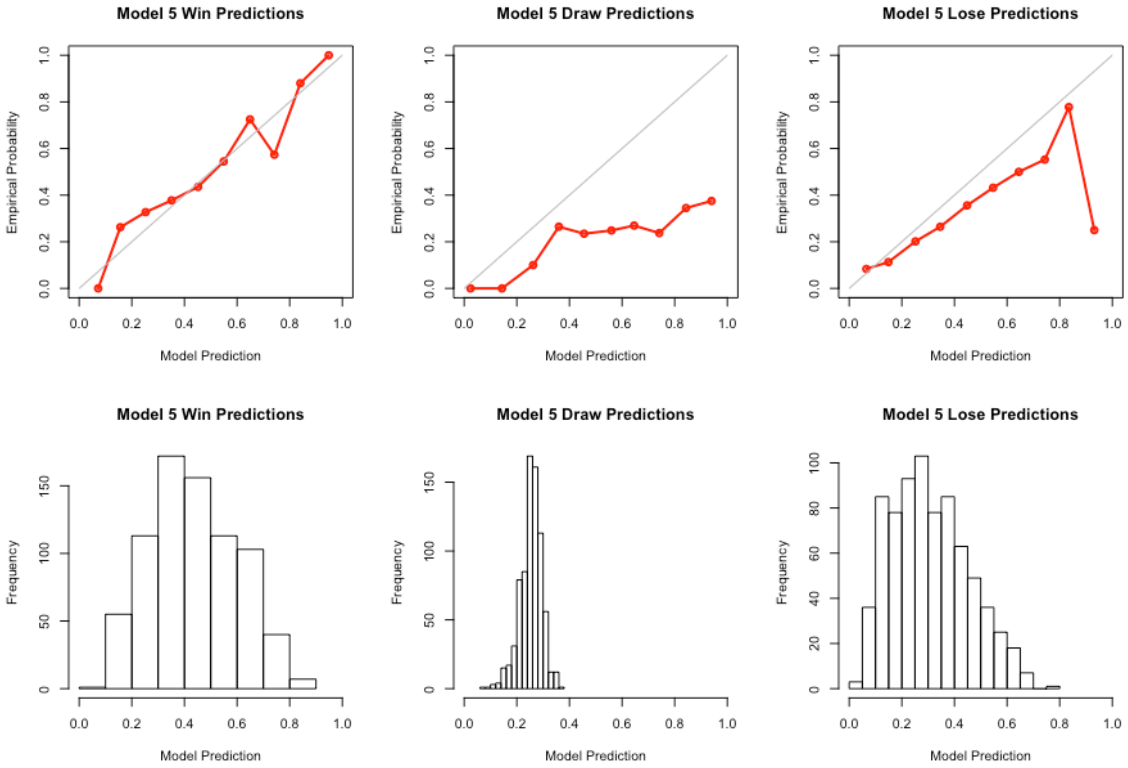


Figure 17 Continued

Average RPS for out-of-sample match outcome predictions

In order to evaluate the out-of-sample performance for each model, we calculated how close our predictions were to the actual match result using the RPS. We filtered the out-of-sample dataset into different segments to investigate whether some models predict better in a particular subsample than others. This seems to be the case.

The motivation for splitting the sample into first and second halves of the season is to test the hypothesis that the first half is difficult to predict. This certainly seems to be the case with our results. For every model, the mean RPS value is lower in the second half of the season than the full season (a smaller value of the RPS indicates better predictions).

For the full out-of-sample forecast dataset and the full 2016-17 season, the simple random walk performs the best. Model 2 (random walk with drift) is the best model in the 2015-16 season sub-sample. The simple random walk model predicts the second half of the 2015-16 the best, but the TVP model has the lowest RPS value in the second half of the 2016-17 season.

Table 6 - Mean RPS Out-of-Sample Forecast

	Seasons 2015-16 2016-17	2015-16	2016-17	2nd Half 2015-16	2nd Half 2016-17
Model 1	**0.20210**	0.21141	**0.19279**	**0.20034**	0.19141
Model 2	0.20213	**0.21138**	0.19289	0.20035	0.19149
Model 3	0.20225	0.21158	0.19292	0.20046	0.19150
Model 4	0.20216	0.21149	0.19283	0.20037	**0.19139**
Model 5	0.20216	0.21146	0.19285	0.20038	0.19145

Discussion

The naïve random walk forecasts performed only slightly better than the model averaging forecasts in terms of having the lowest MSE for most of the time series. However, in terms of predicting match outcome probabilities, the competition lies between the simple random walk forecast and the TVP model. It is not surprising that it did not outperform all other models for all teams. This is consistent with the econometrics literature of modelling and forecasting breaks, as no model seems to forecast best all the time.

This forecasting exercise is intended only as a preliminary one to see how some models in the literature address the structural break problem in forecasting team strengths. Only the 1-step ahead forecast has been considered because of its immediate applicability to sports betting (finding out the team strengths for the upcoming fixture in order to generate predictions for the match outcome). A 19-step ahead or 38-step ahead forecast could be of interest in order to predict table positions mid-season and especially at the end of the season. In terms of the latter, it can provide answers for punters and bookmakers as to which teams are most likely to be relegated, which ones could qualify or miss out on continental competitions, and who is the most likely team to win the Premiership. Along with the 19-step ahead and 38-step ahead forecasts, density forecasting could be used to compare variability between models.

We believe another entire thesis could be dedicated to exploring the forecasting performance of several additional models as well as a plethora of forecasting techniques. This could include other types of model averaging as well as other methods such as exponentially weighted moving average (EWMA), window type and size selection and averaging across windows.

An issue that should be raised is that the team strength time series estimates are most likely to be sensitive to the model specification and the exponential decay factor. In our case, we chose the Dixon and Coles model. Future studies should include comparisons with other bivariate models of team strengths, such as McHale and Scarf (2011) for instance. As pointed out, given the large difference in the decay factor estimate from Dixon and Coles (1997), this area should also be considered a topic for future research, both in terms of how to estimate it (PLL vs. RPS vs. other criteria), the functional form, and differences across leagues and seasons.

The Bai-Perron test was implemented using a specified functional form for the dynamics: an AR(1) with an intercept and a trend. Changing the functional form is likely to affect the results of the test: as more parameters are estimated (for instance, by adding more lags of the autoregressive component, it is likely that the test will fail to reject the null hypothesis (as the test will lose power and more often than not favour the hypothesis with fewer breakpoints).

One methodology which could also be implemented is to model the duration and probability of breaks by using duration analysis, although this does not seem to be very popular in the literature. Even though these types of models may provide some underlying structure that could explain the duration of breaks, the issue of forecasting still remains an important one. The duration analysis can include some covariates such as managerial, ownership change, sponsorship deals, and net player transfers which can affect the hazard function and the probability of the next breakpoint.

In theory, the MSE is a good enough measure of model performance for out-of-sample forecasting exercises, but in terms of its applicability to sports betting, it is purely theoretical. This is the reason we used the average RPS to analyse in practice how good each forecasting model generates match outcome probabilities. Extended work could include a betting performance analysis, although this raises other issues such as money management techniques and bookmakers prices. In practice, the model with the lowest MSE or the lowest RPS might not actually provide the best returns.

We wish to mention one last point, which is an issue that was not addressed in this thesis, but remains a large topic of debate among econometricians. Long memory processes can be easily mistaken with structural change and confusion between both types of time series can also lead to forecasting failure. This area of econometrics could also be an area of future study applied to sports forecasting. Granger and Hyung (2004) show that a model with infrequent

breaks can perform better than an autoregressive fractionally integrated moving average (ARFIMA[6]) process in terms of in-sample fit, but can result in poorer forecasts. Additionally, part of the long memory could be influenced by the presence of breaks in the series. They show that, as the number or size of the breaks in a process increase, the value of the sample autocorrelation function tends to increase. Thus, the process has properties closer to a random walk. This could be the explanation as to why our random walk process has a lower MSE across most of the time series in our forecasting exercise.

[6] ARFIMA processes are a generalised version of ARIMA(p,d,q) processes, except that the differencing parameter d is allowed to take non-integer values.

5. Conclusion

In the course of the last ten to fifteen years, the literature on modelling scores in football matches has moved into researching dynamics to explain how team strengths change over time. In this way, forecasting can take account of the evolution of strength over time. At the very least, this will increase the forecast error relative to forecasts that suppose strengths are not evolving. The basic idea of a forecast is to extrapolate a statistical relationship into the future. Modelling dynamics of strengths allows the strength evolution over time to be extrapolated thus forecasting where team strengths will be at time $t + 1$ in order to generate predictions for scores at time $t + 1$.

Strength dynamics can be considered using time series models. This is the approach we take in this thesis. In this context, a few questions arise, such as what kind of time series process to choose for the strength dynamics. This has important implications in terms of forecasting accuracy and it may be detrimental to choose the wrong model. Another issue that needs to be considered is the one of structural breaks. This is the phenomenon in econometrics where a time series experiences an unexpected change in the regression coefficients which govern the data generating process. When these regression coefficients are time invariant, large forecasting errors may occur and this is an important problem in time series econometrics (Clements and Hendry, 1998). Since then, there have been several attempts in developing forecasting techniques and models to improve forecasting performance of time series under parameter instability, particularly because of the importance of forecasting time series models in society. Several macroeconomic and financial data are modelled using these types of models.

To date, this is the first thesis that addresses these two problems. Firstly it provides the contribution to knowledge that structural breaks are present in team strength dynamics. Through the use of the Bai and Perron (2003) algorithm, which endogenously selects the optimum breakpoints by minimising the global sum of squared residuals in all possible breakpoint partition combinations, we demonstrated that the vast majority of teams strength dynamics have at least one breakpoint.

The second contribution is a natural result of the first one: if structural breaks have been shown in macro- and financial econometrics to result in large forecast errors, why could this not be the case in the football literature that has moved into researching dynamics? We carried out a forecasting exercise to test if this was the case. We compared different dynamic

specifications of team strength dynamics: the simple random walk, random walk with drift, auto-arima process and a time-varying parameter model (as well as a model averaging forecasting technique) in their abilities to firstly: forecast where team strengths would be in the next playing round of the league (the one-step ahead forecast) and secondly: using those forecasts, generating match outcome predictions.

Our out-of-sample test dataset was two seasons of the English Premier League (2015-16 and 2016-17), using team strengths dynamics from the 2010-11 to 2014-15 seasons as a training set. As a measure of how accurate our predicted match outcome probabilities were with the actual match result, we used the ranked probability score (Constantinou and Fenton, 2013), which is a good measure of evaluating the performance of predictions of ordered outcomes. In this forecasting exercise, we demonstrated that, in one of our out-of-sample subsamples (in the second half of the 2016-17 season), the forecasts of team strengths from the time-varying parameter model, which addresses this problem of forecasting under the presence of structural breaks, provided the closest match predictions to the real outcome the predictions of match outcomes according to the ranked probability score. Our forecasting exercise also shows that a simple random walk forecasts of team strengths performs very well in generating match outcomes predictions too, particularly in the first half of the season. This is probably because as the number of breaks increase, the sample autocorrelation function tends to increase and the process has properties that closely resemble a random walk model.

We believe that the findings of this thesis has opened the door to a plethora of research questions to be investigated. Firstly, our estimate of the time decay factor is smaller than the one in Dixon and Coles (1997) by one order of magnitude. This means that we discount past matches less than Dixon and Coles. This actually gives further credibility to our structural breaks finding: a smaller decay factor means historical data contributes more to the current parameter estimates, we would expect the time series of parameter estimates to be more smooth and thus more difficult to reject the null hypothesis of "no structural breaks in the series". Nevertheless, given the difference, the sensitivity of the decay factor estimate to different datasets should be the subject of further study.

The framework for utilising the team strengths parameter estimates over time and modelling them through time series analysis is quite flexible and can be easily implemented in other sports. Essentially one could make use of any probability distribution to derive team strength parameter estimates over time. In the case of football, we could have obtained the parameter estimates over time through the inflated bivariate Poisson of Karlis and Ntzoufras

(2003) or the use of copulas (McHale and Scarf, 2011) to join different marginal count data distributions.

The forecasting exercise can also be enhanced by comparing different types of forecasting models and techniques which address out-of-sample forecasting under the presence of structural breaks. Bayesian methodologies that have been extensively evaluated in Bauwens et al. (2015) could be implemented in a similar out-of-sample forecasting exercise of team strengths to predict match outcomes as well as different types of model averaging (such as Bayesian model averaging).

Finally, we hope we have opened the doors for future researchers to engage in this interesting area of research, which is forecasting sports results. Macro and financial econometricians have a myriad of modelling techniques at their disposal that can be used in other disciplines, such as sports statistics. I hope to see more sports statisticians making use of these techniques not only in football, but in other sports as well.

References

Audas R, Dobson S and Goddard J (2002) The impact of managerial change on a team performance in professional sports. *Journal of Economics and Business* 54, 633–650.

Bai J and Perron P (1998) Estimation and testing linear models with multiple structural changes. *Econometrica* 66, 47 – 78.

Bai J and Perron P (2003) Computation and analysis of multiple structural change models. *Journal of Applied Econometrics* 18, 1 – 22.

Baker R and McHale G (2014) A dynamic paired comparisons model: Who is the greatest tennis player? *European Journal of Operational Research* 236, 677-684.

Baker R and McHale I (2015) Time varying ratings in association football: the all-time greatest team is… *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 178, 481–492.

Bauwens L, Koop G, Korobilis D and Rombouts J (2015) The contribution of structural break models to forecasting macroeconomic series. *Journal of Applied Econometrics* 30, 596–620.

Box G and Jenkins G (1970) *Time Series Analysis:Forecasting and Control*. Holden-Day, San Francisco, CA.

Brillinger D (2007) A potential function approach to the flow of play in soccer. *Journal of Quantitative Analysis in Sports* 3, issue 1, article 3.

Bruinshoofd A and Weel B (2003) Manager to go? Performance dips reconsidered with evidence from Dutch football. *European Journal of Operational Research* 148, 233-246.

Buraimo B and Simmons R (2015) Uncertainty of outcome or star quality? Television audience demand for English premier league football. *International Journal of the Economics of Business* 22, 449-469.

Chib S (1998) Estimation and comparison of multiple change-point models. *Journal of Econometrics* 86, 221 – 241.

Chow G (1960) Tests of equality between sets of coefficients in two linear regressions. *Econometrica* 28, 591 – 605.

Clements M and Hendry D (1998) Forecasting economic processes. *International Journal of Forecasting* 14, 111–131.

Constantinou A and Fenton N (2012) Solving the problem of inadequate Scoring Rules for assessing probabilistic football forecast models. *Journal of Quantitative Analysis in Sports* 8 (1) Article 1.

Constantinou A and Fenton N (2013) Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries. *Journal of Quantitative Analysis in Sports* 9, 37–50.

Crowder M, Dixon MJ, Ledford L, Robinson M (2002). Dynamic modelling and prediction of English Football League matches for betting. *Journal of the Royal Statistical Society: Series D (The Statistician)* 51, 157-168.

Diebold F and Kilian L (2000) Unit root tests are useful for selecting forecasting models. *Journal of Business and Economic Statistics* 18, 265–273.

Dixon M and Coles S (1997) Modelling association football scores and inefficiencies in the football betting market. *Applied Statistics* 46, 265–280.

Dixon M and Robinson M (1998) A birth process model for association football matches. *Journal of the Royal Statistical Society: Series D (The Statistician)* 47, 523–538.

Dobson S and Goddard J (2003) Persistence in sequences of football match results: a Monte Carlo analysis. *European Journal of Operational Research* 148, 247–256.

Durban J and Koopman S (2001) *Time Series Analysis by State Space Methods*. Oxford University Press, Oxford.

Elliot G, Rothenburg T and Stock J (1996) Efficient tests for an autoregressive unit root. *Econometrica* 64, 813–836.

Forrest D and Simmons R (2000a) Forecasting sport: The behaviours and performance of football tipsters. *International Journal of Forecasting* 16, 317–331.

Forrest D and Simmons R (2000b) Making up the results: The work of the football pools panel, 1963 – 1997. *Journal of the Royal Statistical Society: Series D (The Statistician)* 49, 253–260.

Forrest D and Simmons R (2002) Outcome uncertainty and attendance demand in sport: The case of English soccer. *Journal of the Royal Statistical Society: Series D (The Statistician)* 51, 229–241.

Gerlach R, Carter C and Kohn R (2000) Efficient Bayesian inference for dynamic mixture models. *Journal of the American Statistical Association* 95, 819 – 828.

Geweke J and Amisano G (2011) Optimal prediction pools. *Journal of Econometrics* 164, 130 – 141.

Giordani P and Kohn R (2008) Effiicient Bayesian inference for multiple change-point and mixture innovation models. *Journal of Business and Economic Statistics* 26, 66 – 77.

Goddard J (2005) Regression models for forecasting goals and match results in association football. *International Journal of Forecasting* 21, 331–340.

Goddard J and Asimakopoulos I (2004) Forecasting football results and the efficiency of fixed-odds betting. *Journal of Forecasting* 23, 51–66.

Goossens D and Spieksma F (2011) Soccer schedules in Europe: an overview. *Journal of Scheduling* 15, 1099-1425.

Goossens D, Belien J and Spieksma F (2012) Comparing league formats with respect to match importance in Belgian football. *Annals of Operations Research* 194, 223-240.

Granger C W J and Hyung N (2004) Occasional structural breaks and long memory with an application to the S&P 500 absolute stock returns. *Journal of Empirical Finance* 11, 399 – 421.

Groll A, Schauberger G and Tutz G (2015) Prediction of major international soccer tournaments based on team-specific regularized Poisson regression: An application to the FIFA World Cup 2014. *Journal of Quantitative Analysis in Sports* 11, 97–115.

Hamilton J (1989) A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* 57, 357 – 384.

Hirotsu N and Wright M (2006) Modelling tactical changes of formation in association football as a zero-sum game. *Journal of Quantitative Analysis in Sports* 2, issue 2.

Hvattum L and Arntzen H (2010) Using ELO ratings for match result prediction in association football. *International Journal of Forecasting* 26, 460–470.

Karlis D and Ntzoufras I (2003) Analysis of sports data using bivariate Poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician)* 52, 381–393.

Karlis D and Ntzoufras I (2009) Bayesian modelling of football outcomes: using the Skellam's distribution for the goal difference. *IMA Journal of Management Mathematics* 20, 133–145.

Keogh F and Rose G (2013) Football-betting – the global gambling industry worth billions. http://www.bbc.co.uk/sport/football/24354124, accessed on April 22, 2016.

Knorr-Held L (2000) Dynamic rating of sports teams. *Journal of the Royal Statistical Society: Series D (The Statistician)* 49, 261 – 276.

Koning R (2000) Balance in competition in Dutch soccer. *Journal of the Royal Statistical Society: Series D (The Statistician)* 49, 419–431.

Koning R, Koolhaas M, Renes G and Ridder G (2002) A simulation model for football championships. *European Journal of Operational Research* 148, 268–276.

Koop G and Potter S (2007) Estimation and forecasting in models with multiple breaks. *Review of Economic Studies* 74, 763–789.

Koopman S and Lit R (2015) A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 178, 167–186.

Lenten L, Libich J and Stehlık P (2013) Policy timing and footballers' incentives: penalties before or after extra time? *Journal of Sports Economics* 14, 629-655.

Lewis, M. (2003). *Moneyball: The art of winning an unfair game.* New York: W.W. Norton

Maddala G (2001) *Introduction to Econometrics*. 3rd edition, Chichester, West Sussex: Wiley.

Maher M (1982) Modelling association football scores. *Statistica Neerlandica* 36, 109–118.

McHale I and Scarf P (2007) Modelling soccer matches using bivariate discrete distributions with general dependence structure. *Statistica Neerlandica* 61, 432-445.

McHale I and Scarf P (2011) Modelling the dependence of goals scored by opposing teams in international soccer matches. *Statistical Modelling* 11, 219–236.

McHale I, Scarf P, and Folker D (2012) On the development of a soccer player performance rating system for the English Premier League. *Interfaces* 42, 339 – 351.

Moroney M (1956). *Facts from Figures*. Penguin, London, 3rd edition.

Owen A (2011) Dynamic Bayesian forecasting models of football match outcomes with estimation of the evolution variance parameter. *IMA Journal of Management Mathematics* 22, 99–113.

Owen A (2016). The decline in home advantage in the English Premier League in 2015/16: is it random fluctuation or a systematic change? *STN Journal of Sports Modelling and Trading* 1 (1) Article 4.

Palacios-Huerta I (2004) Structural changes during a century of the world's most popular sport. *Statistical Methods and Applications* 13, 241 – 258.

Perron P (1989) The Great Crash, the Oil Price Shock, and the unit root hypothesis. *Econometrica* 57, 1361–1401.

Perron P (2006) Dealing with structural breaks. In: Mills, T C and Patterson K (eds.) *Palgrave Handbook of Econometrics*, Vol 1, Basingstoke: Palgrave-Macmillan 278 – 352.

Pesaran M and Timmermann A (2004) How costly is it to ignore breaks when forecasting the direction of a time series? *International Journal of Forecasting* 20, 411–425.

Pesaran M and Timmermann A (2007) Selection of estimation window in the presence of breaks. *Journal of Econometrics* 137, 134 – 161.

Pesaran M, Pettenuzzo D, Timmermann A (2006) Forecasting time series subject to multiple structural breaks. *Review of Economic Studies* 73, 1057 – 1084.

R Development Core Team (2008) *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Rosenberg B (1973) Random coefficients models: the analysis of a cross section of time series by stochastically convergent parameter regression. In: Berg SV (ed) *Annals of Economic and Social Measurement,* Vol 2, No. 4. Cambridge, MA: NBER, 399 – 428.

Rossi B (2013). Advances in forecasting under instability. In: Elliot G and Timmermann A (eds) *Handbook of Economic Forecasting*, Vol 2. Amsterdam: North-Holland, 1203 – 1324.

Rue H and Salvesen O (2000) Prediction and retrospective analysis of soccer matches in a league. *Journal of the Royal Statistical Society: Series D (The Statistician)* 49, 399–418.

Said S and Dickey D (1984) Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika* 71, 599–607.

Scarf P, Yusof M and Bilbao M (2009) A numerical study of designs for sporting contests. *European Journal of Operational Research* 198, 190–198.

Scarf P and Rangel Jr J S (2017) Models for outcomes of soccer matches. In: Albert J, Glickman M, Swartz T and Koning R (eds) *Handbook of Statistical Methods and Analyses in Sports* Boca Raton, FL: Chapman and Hall/CRC, 341 – 354.

Scarf P and Shi X (2008) The importance of a match in a tournament. *Computers and Operations Research* 35, 2406-2418.

Scarf P and Yusof M (2011) A numerical study of tournament structure and seeding policy for the soccer World Cup Finals. *Statistica Neerlandica* 65, 43-57.

Schryver T and Eisinga R (2011) Piecewise linear regression techniques to analyse the timing of head coach dismissals in Dutch soccer clubs. *IMA Journal of Management Mathematics* 22, 129-137.

Sporty Desk (2015) Top 10 most popular sports in the world. http://www.sportyghost.com/top-10-most-popular-sports-in-the-world/ accessed April 22, 2016.

Stock J and Watson M (2007) Why has U.S inflation become harder to forecast. *Journal of Money, Credit and Banking* 39, 3 – 33.

Tenga A, Holme I, Ronglan L and Bahr R (2010) Effect of playing tactics on goal scoring in Norwegian professional soccer. *Journal of Sports Sciences* 28, 237-244.

Titman A, Costain D, Ridall P and Gregory K (2015) Joint modelling of goals and bookings in association football. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 178, 659-683.

Wright M and Hirotsu N (2003) The professional foul in football: tactics and deterrents. *Journal of the Operational Research Society* 54, 213-221.

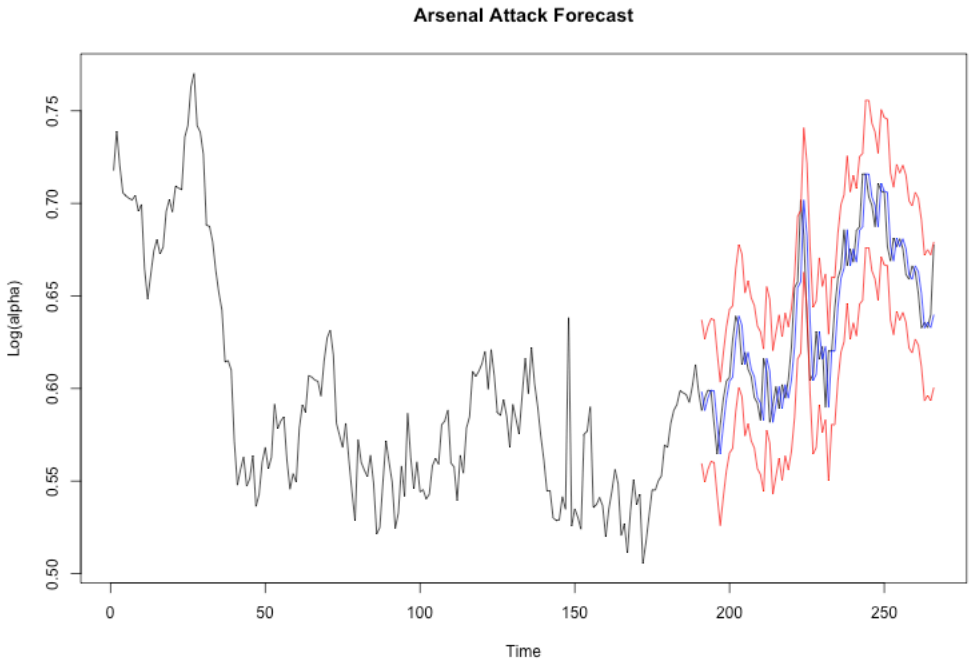
Xu J and Perron P (2017) Forecasting in the presence of in and out of sample breaks. *Boston University – Department of Economics – Working Paper Series wp2015-012* Boston University – Department of Economics.

Zeileis A, Kleiber C, Kramer W and Hornik K (2003) Testing and dating of structural changes in practice. *Computational Statistics and Data Analysis* 44, 109 – 123.

Zeileis A, Leisch F, Kleiber C and Hornik K (2005) Monitoring structural change in dynamic econometric models. *Journal of Applied Econometrics* 20, 99 – 121.

Zivot E and Andrews K (1992) Further evidence on The Great Crash, The Oil Price Shock, and the unit root hypothesis. *Journal of Business and Economic Statistics* 10, 251 – 270.

Appendix A – Forecasts from the simple random walk model



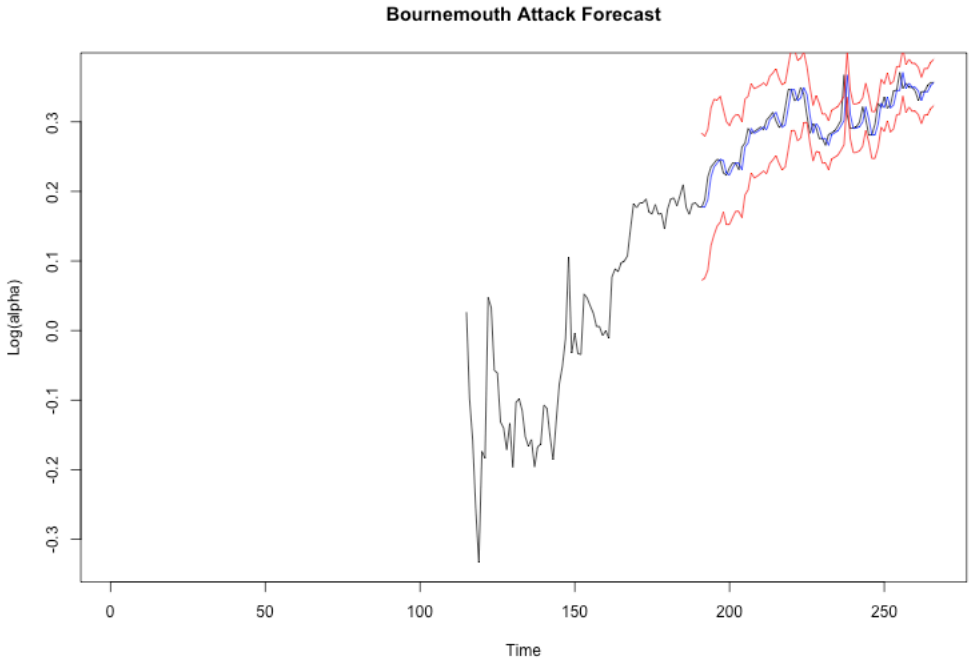
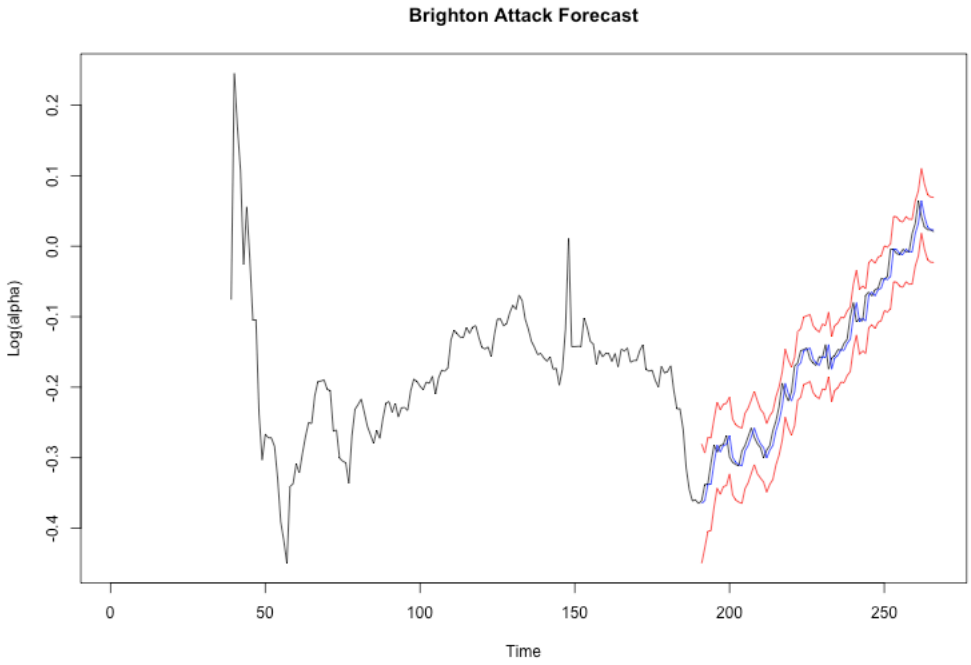


Figure 18 - (Model 1) Simple Random Walk Forecasts – Attack



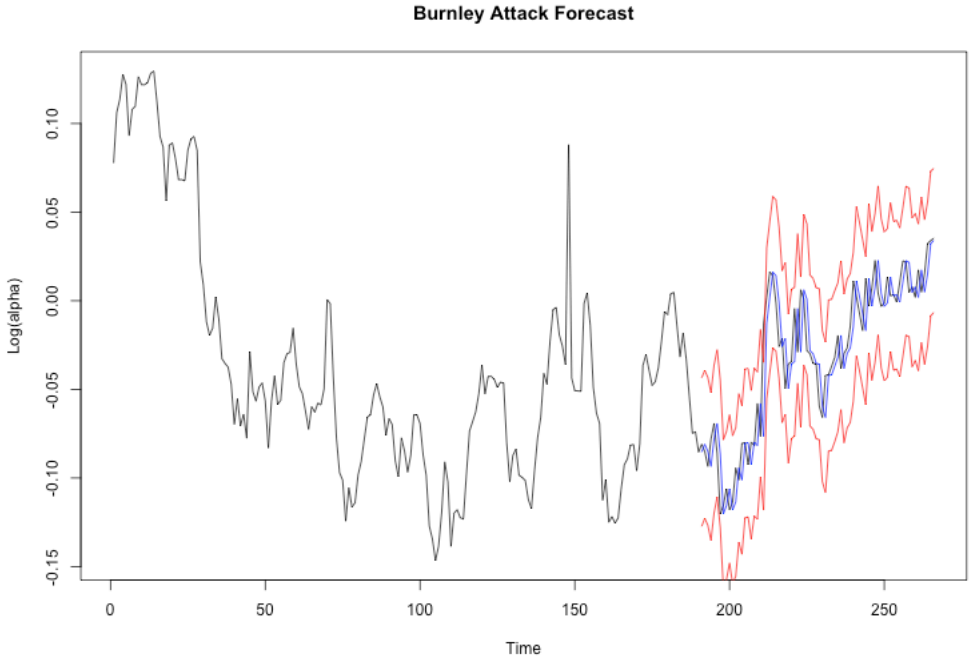
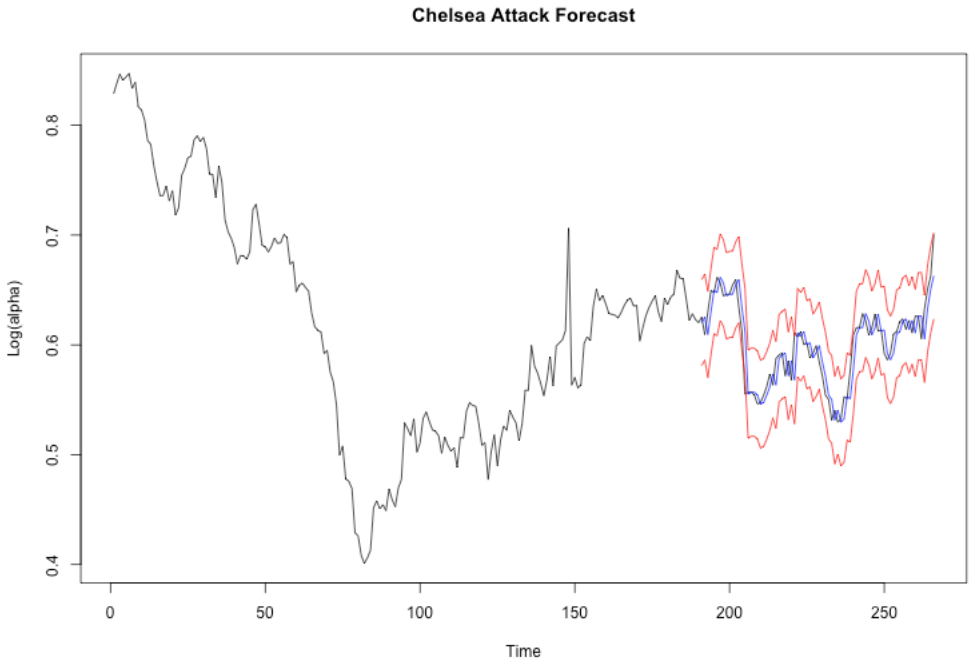


Figure 18 Continued



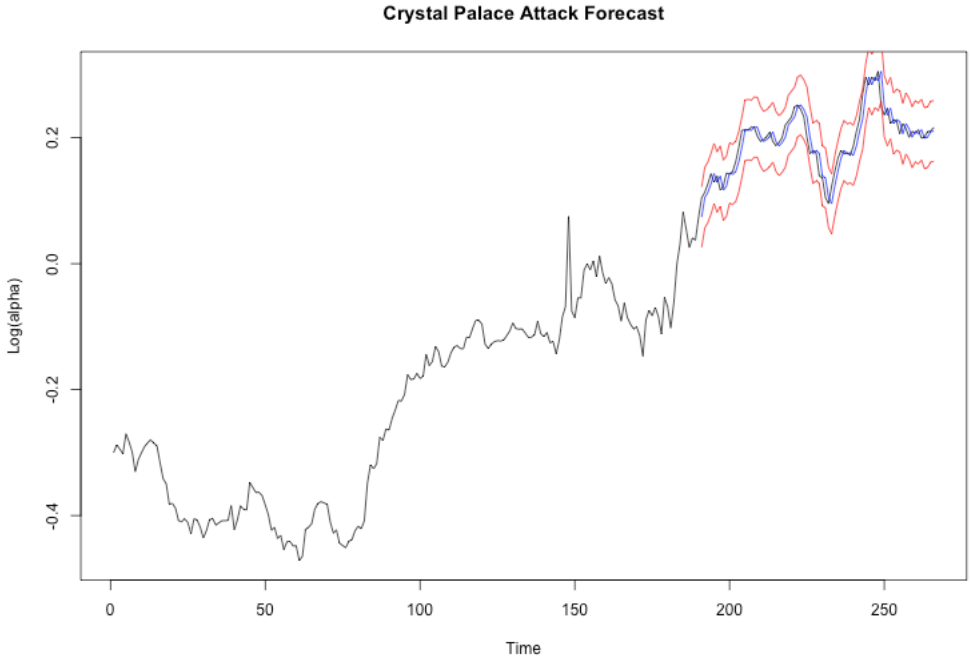
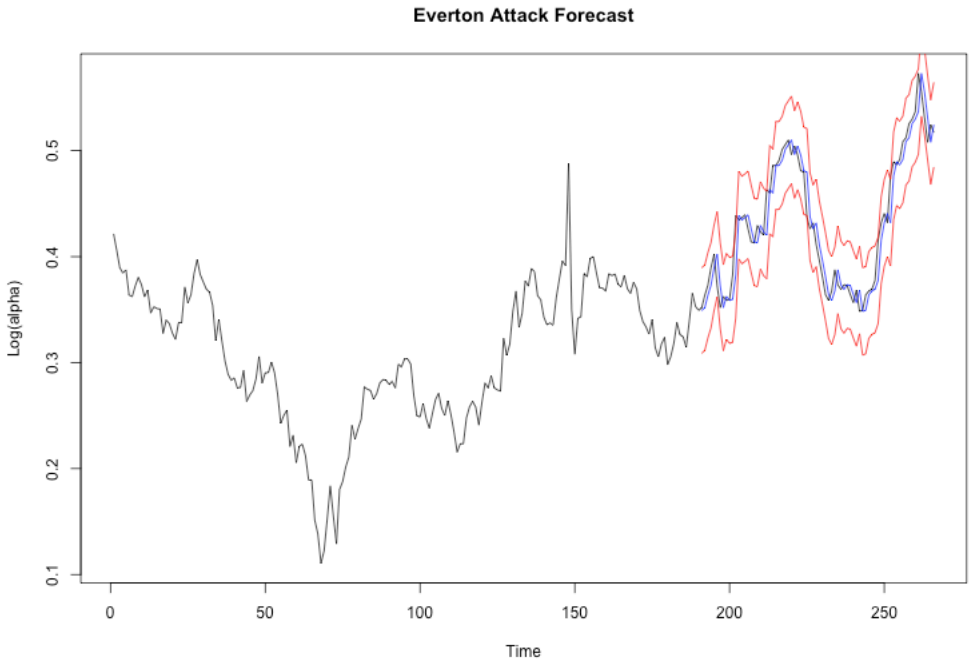


Figure 18 Continued



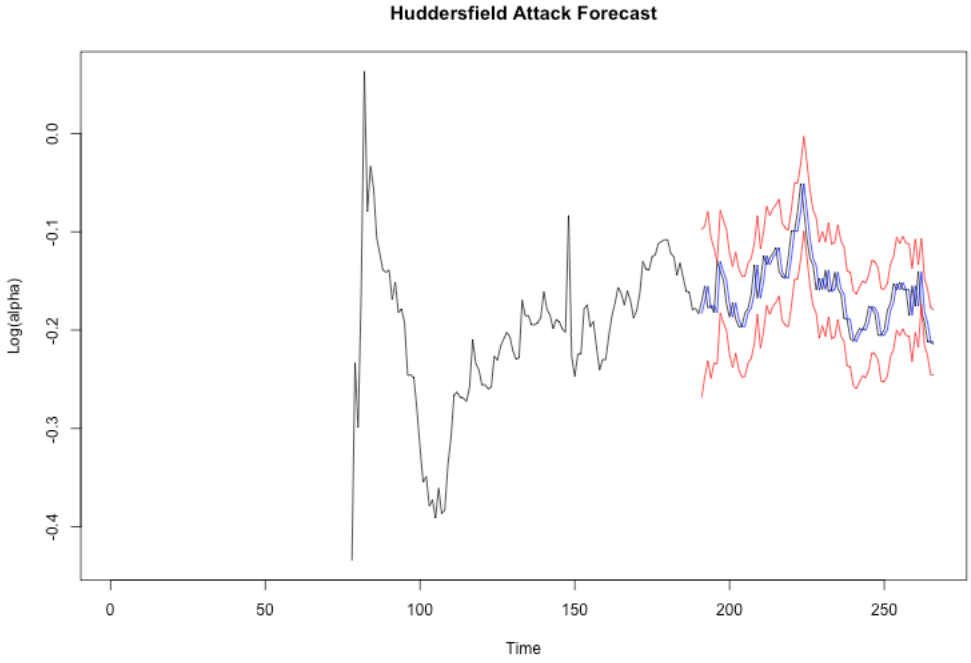
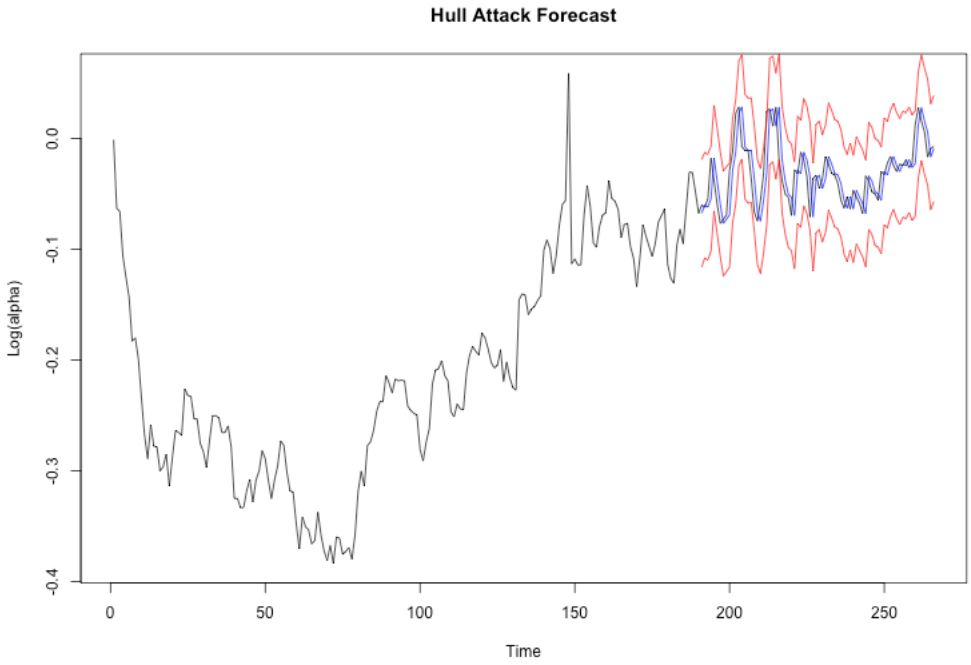


Figure 18 Continued



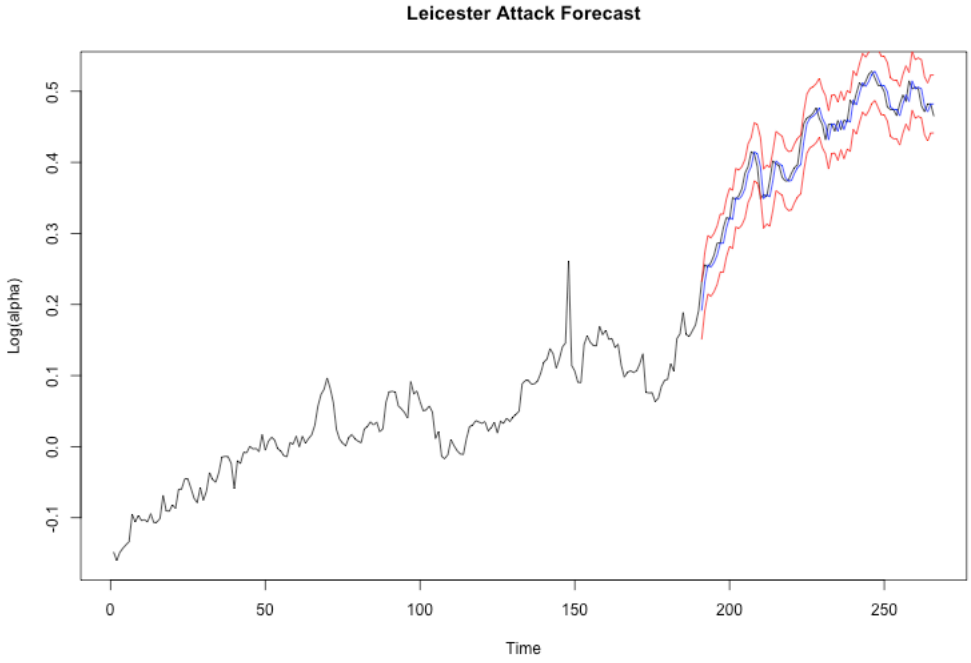
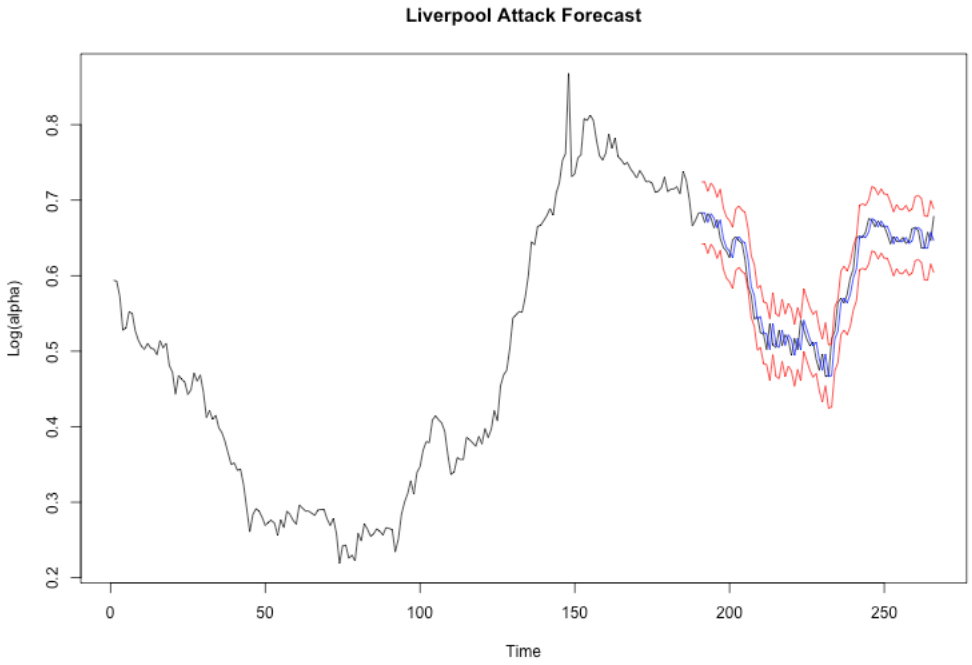


Figure 18 Continued



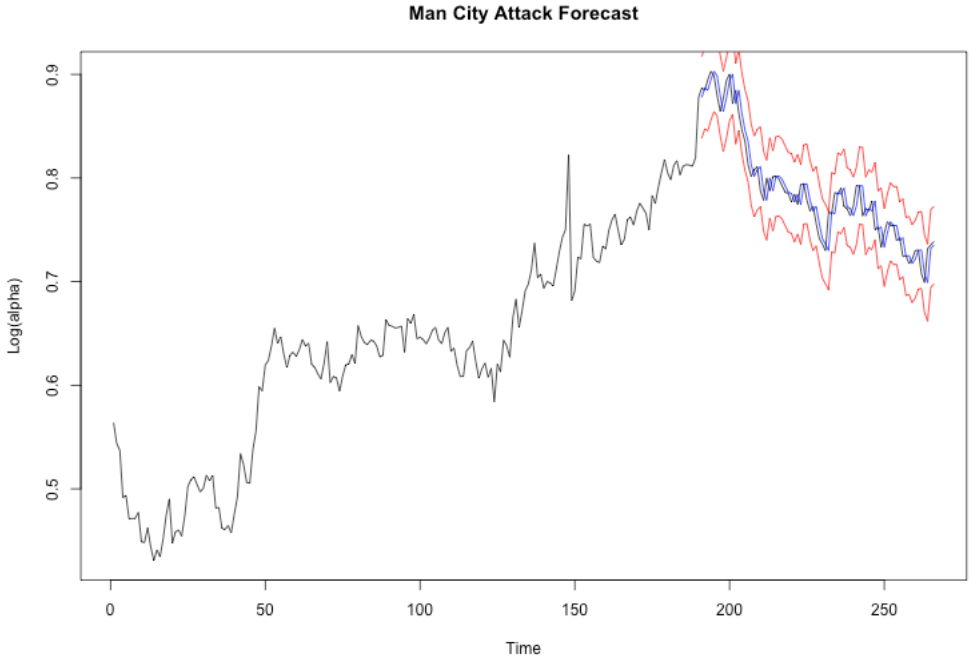
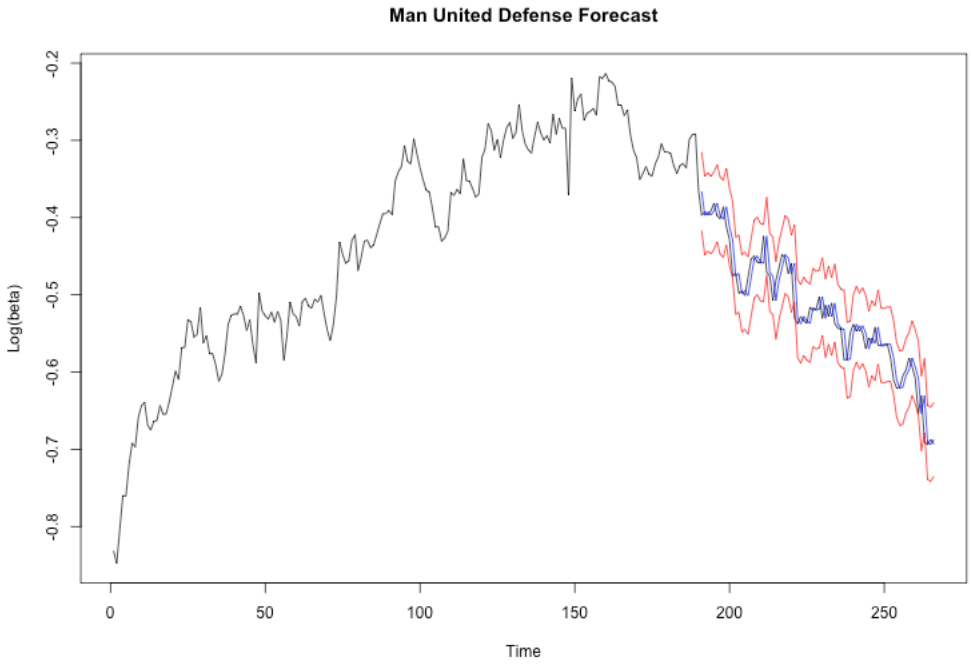


Figure 18 Continued



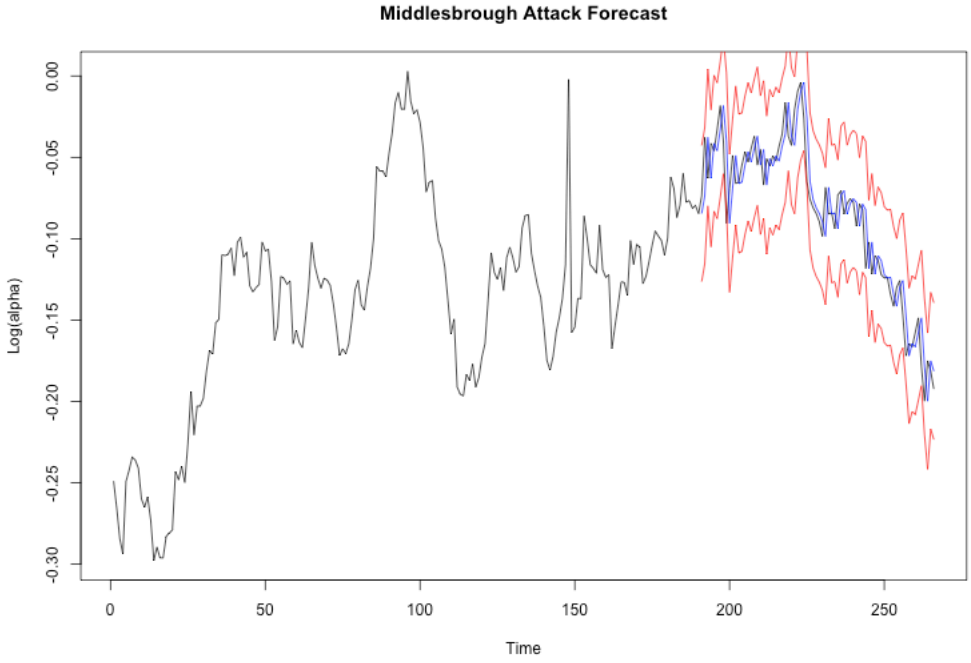
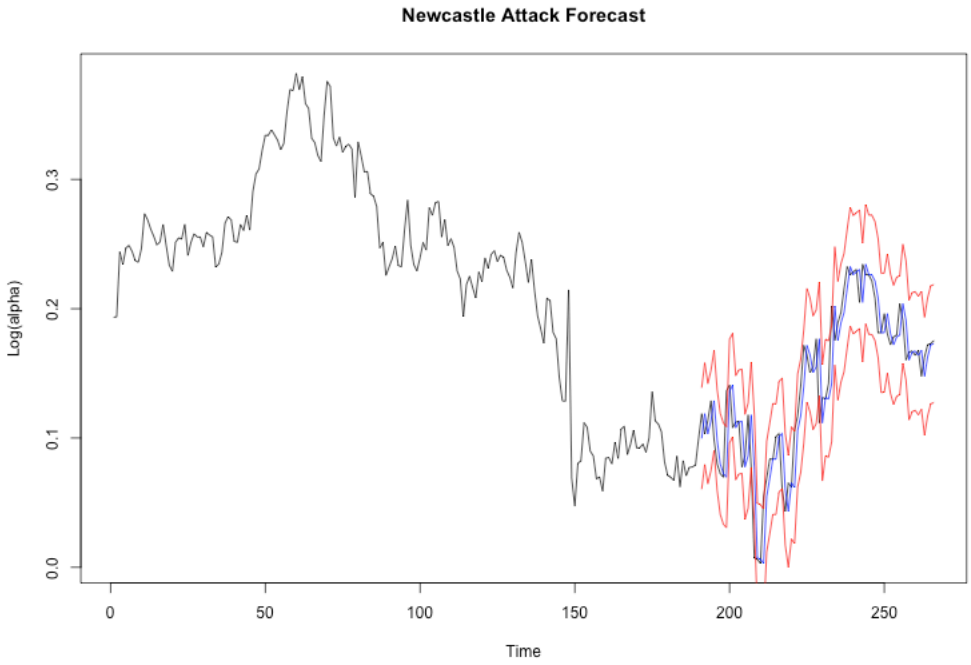


Figure 18 Continued



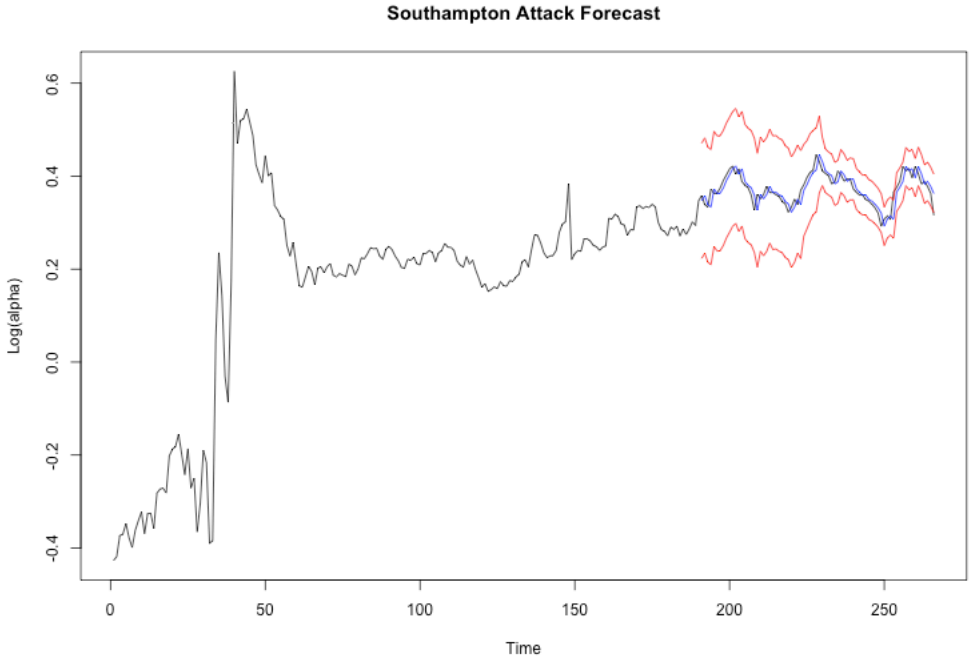
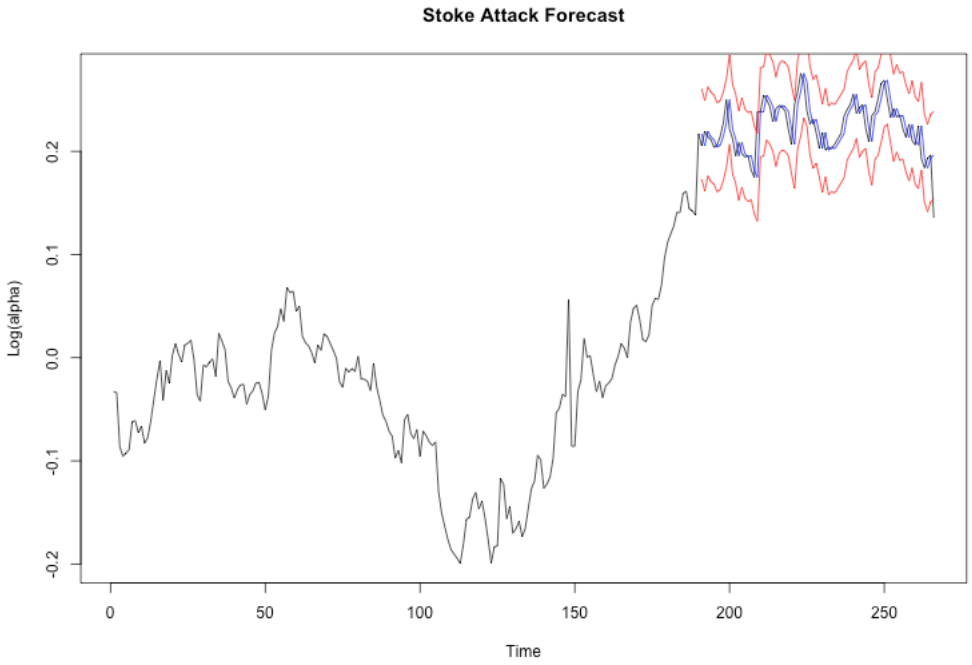


Figure 18 Continued



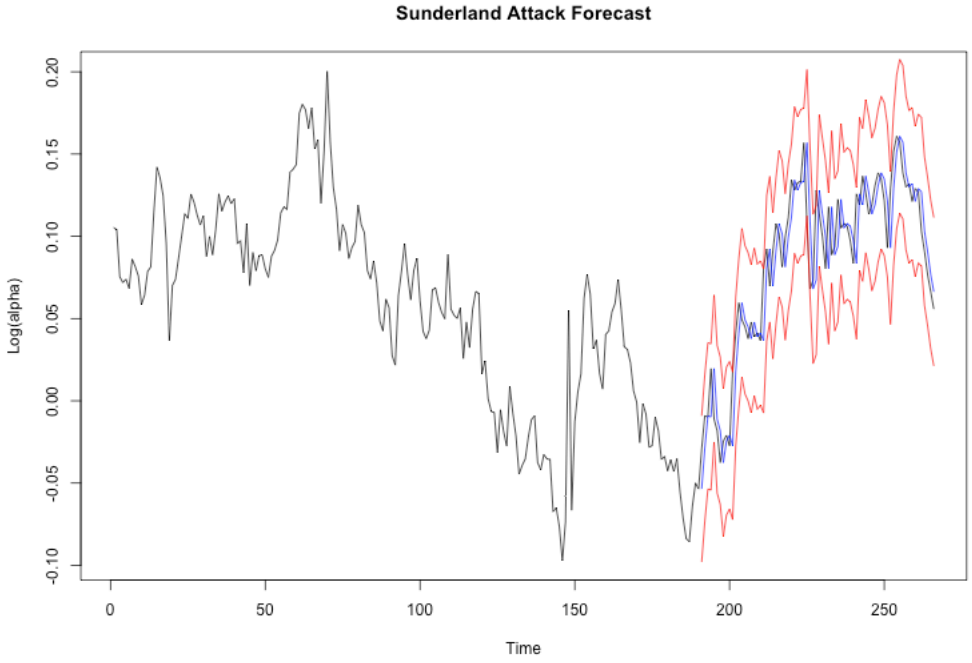
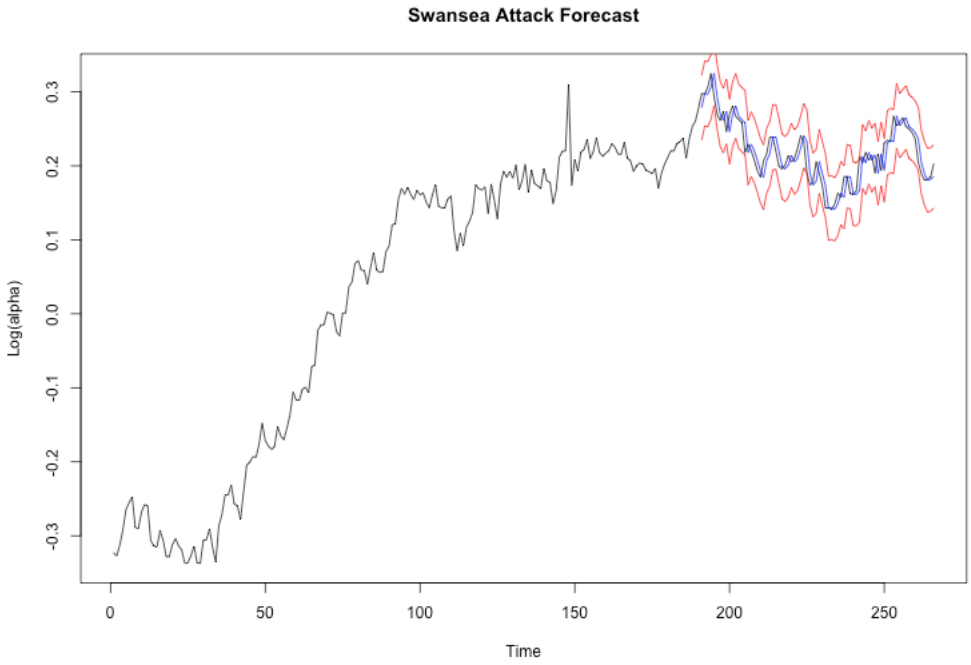


Figure 18 Continued



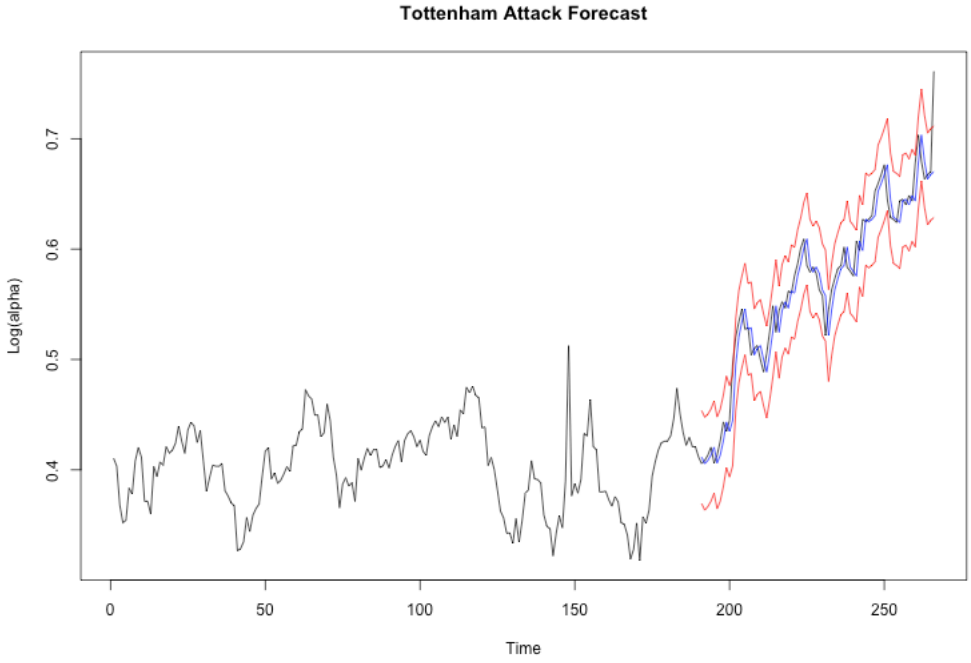
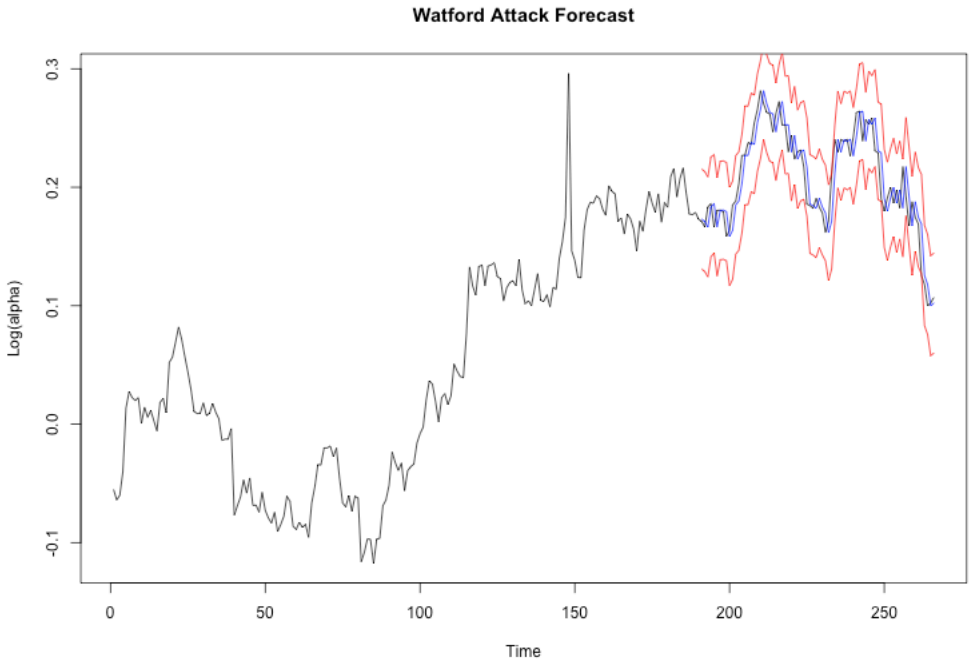


Figure 18 Continued



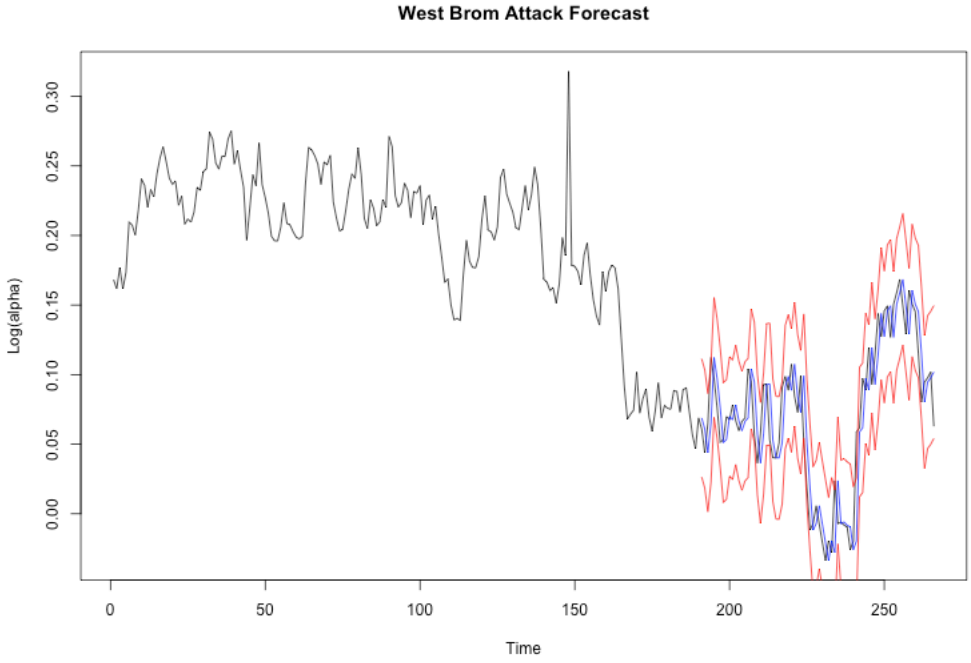


Figure 18 Continued

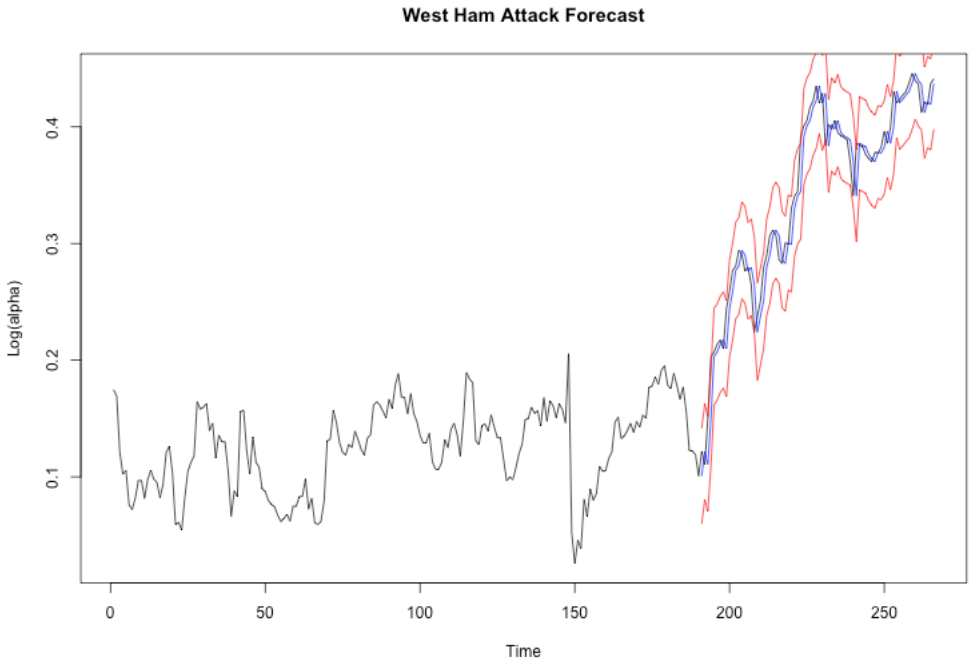


Figure 18 Continued

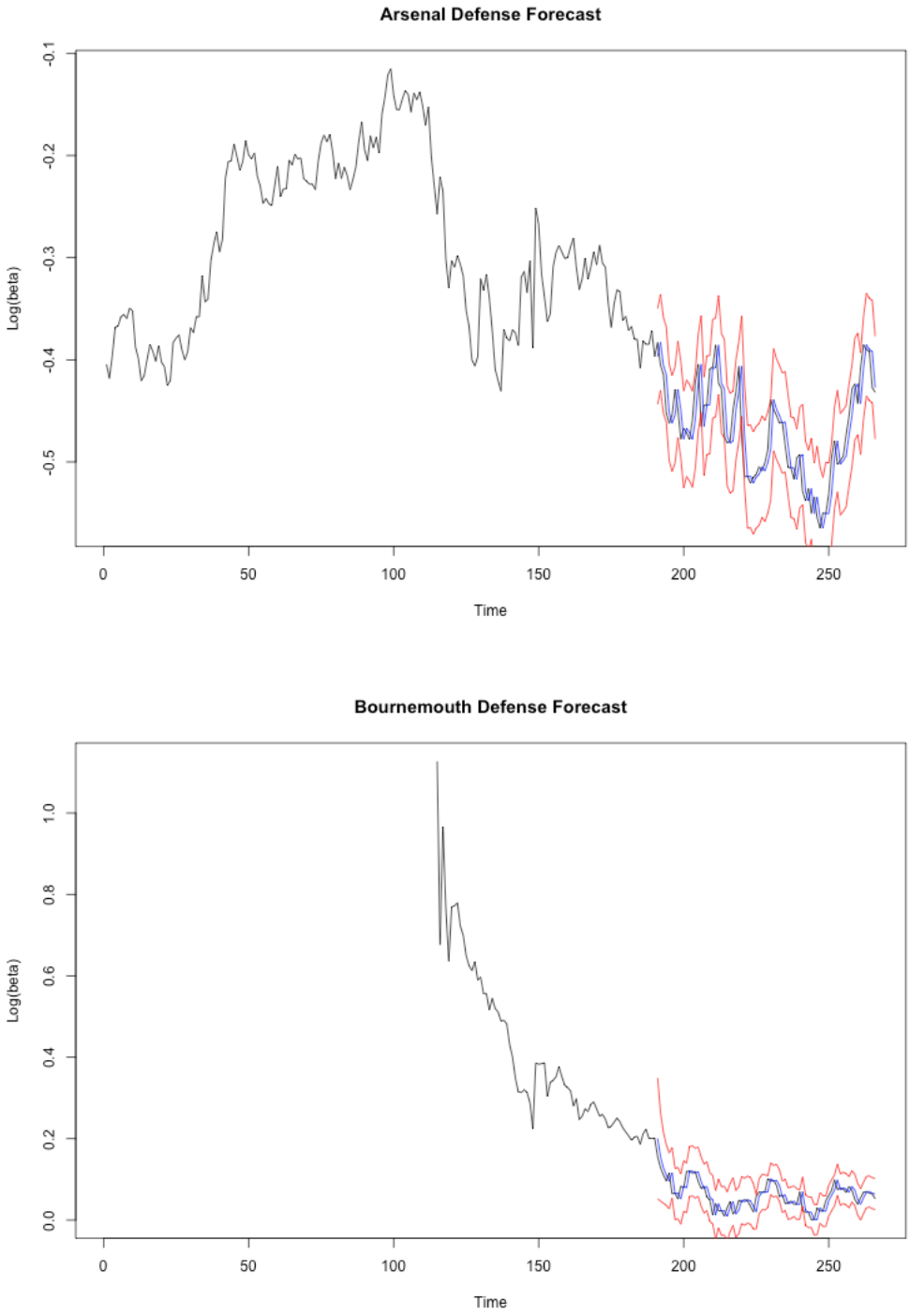
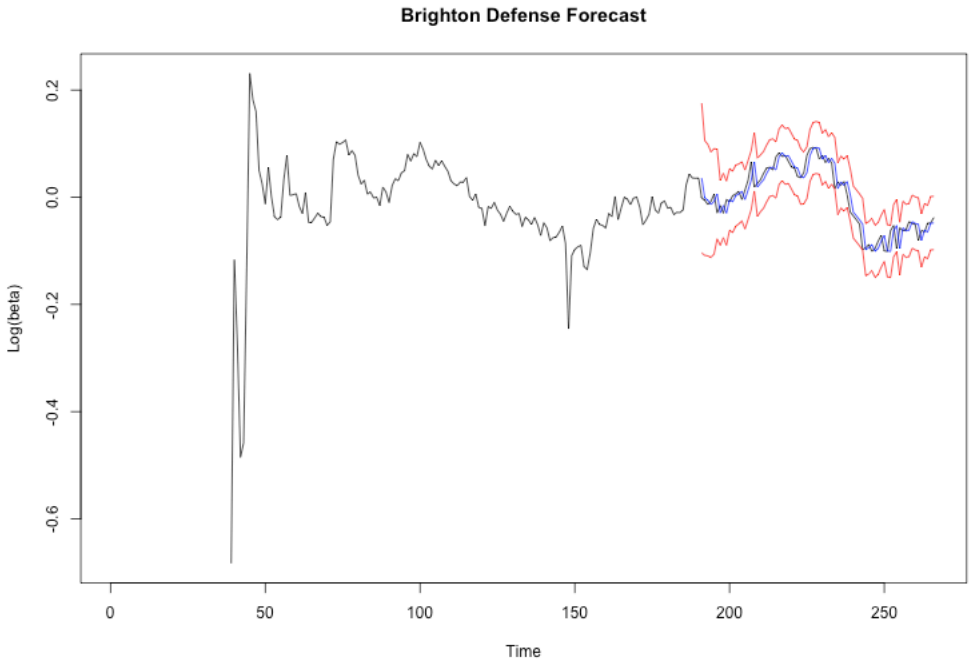


Figure 19 - (Model 1) Simple Random Walk Forecasts – Defence



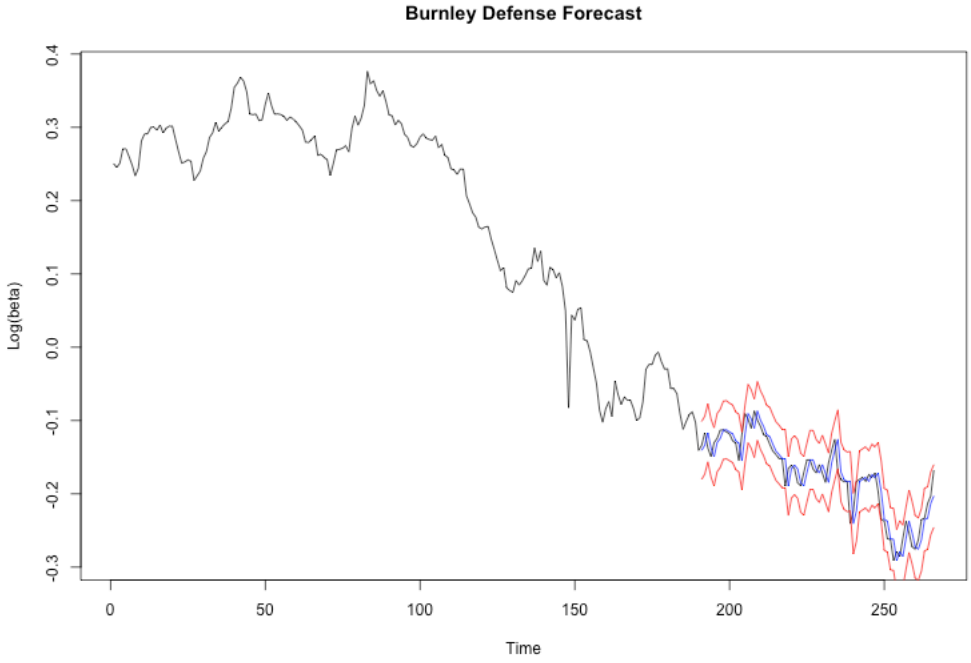
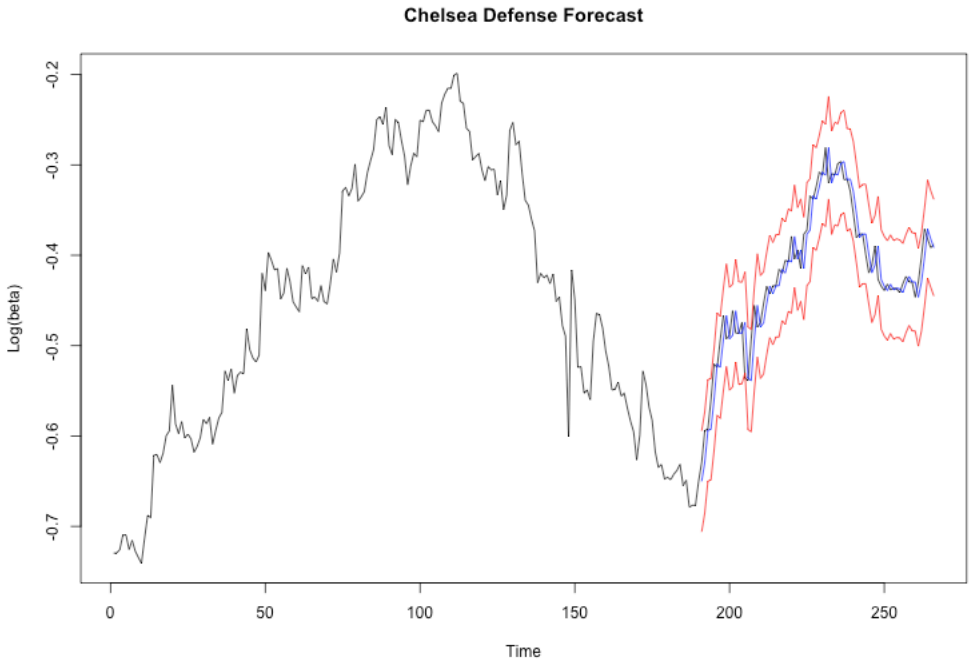


Figure 19 Continued



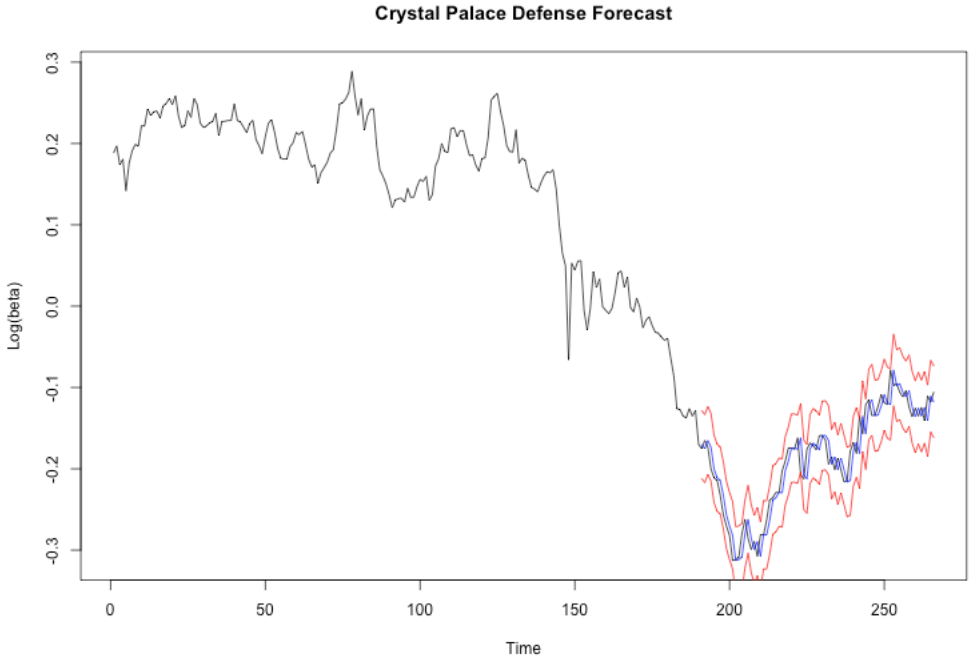
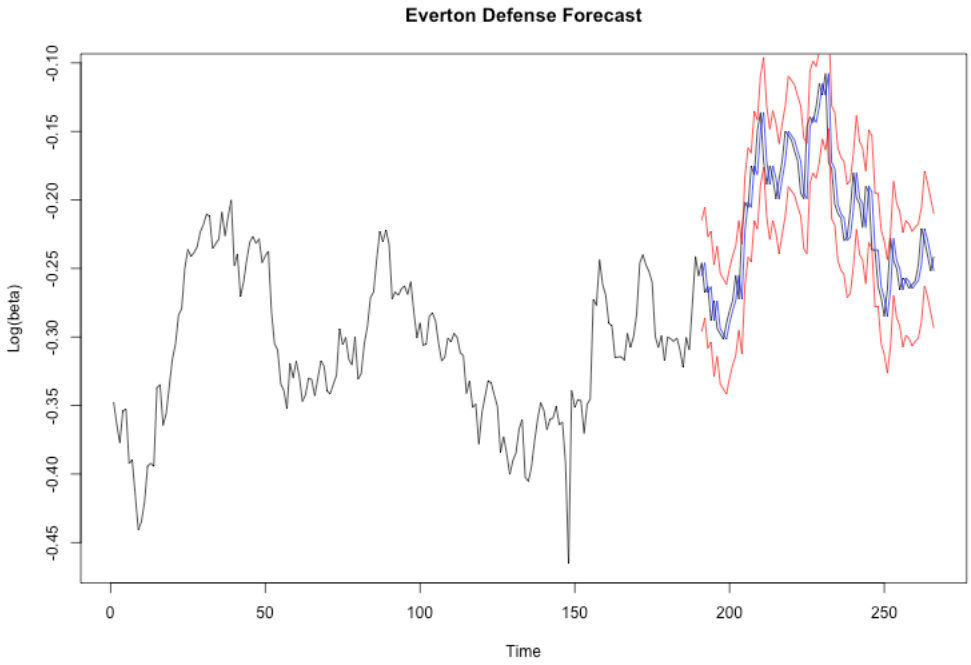


Figure 19 Continued



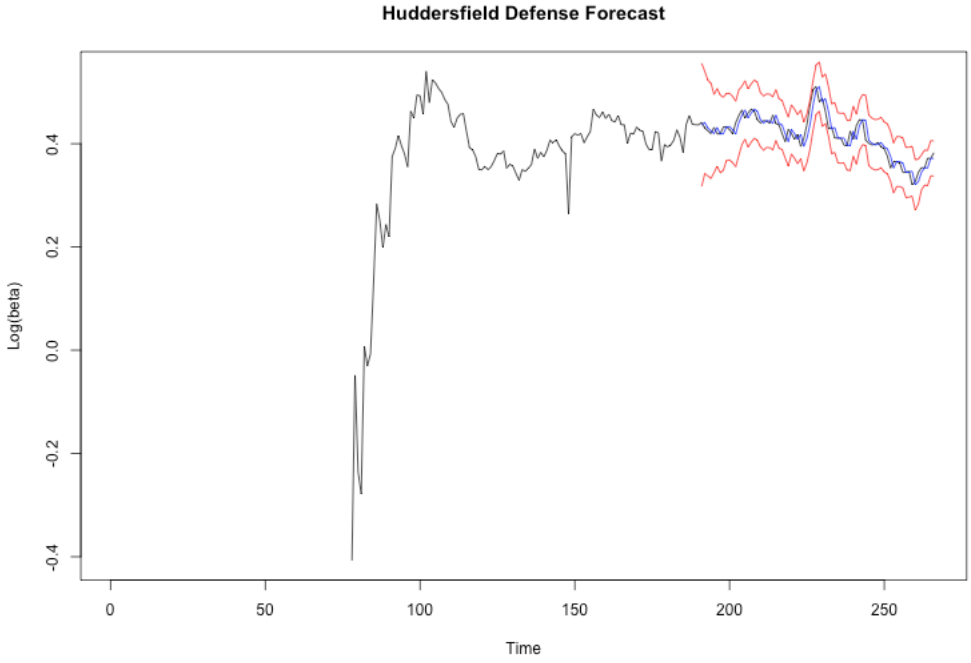
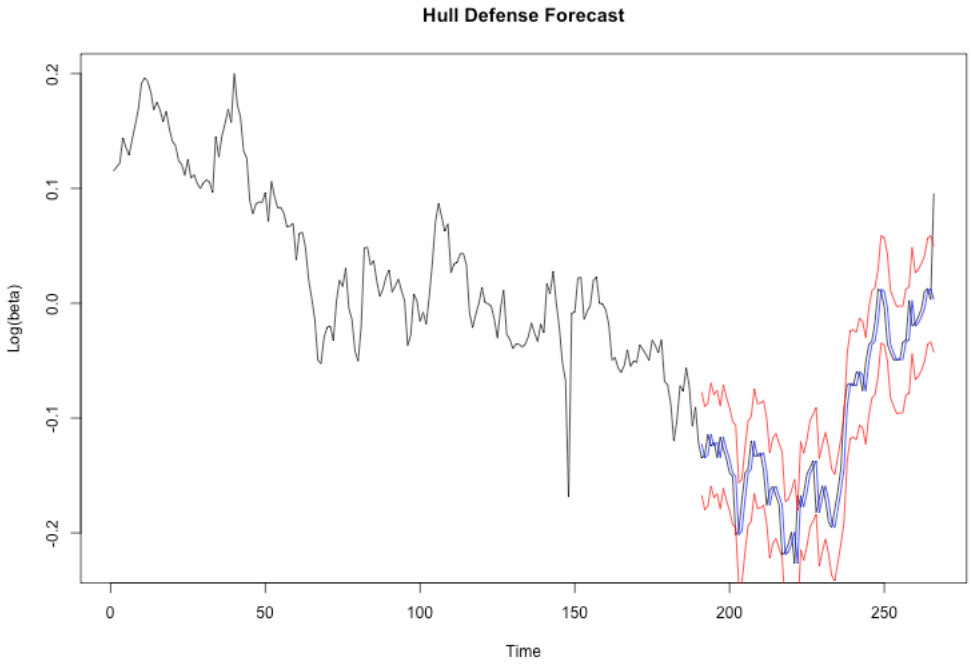


Figure 19 Continued



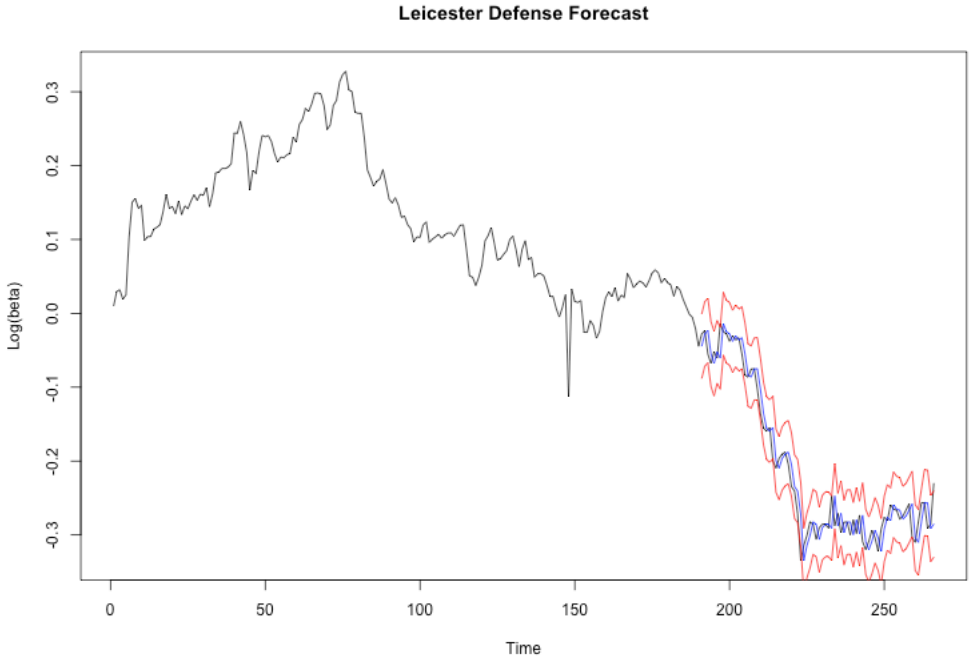
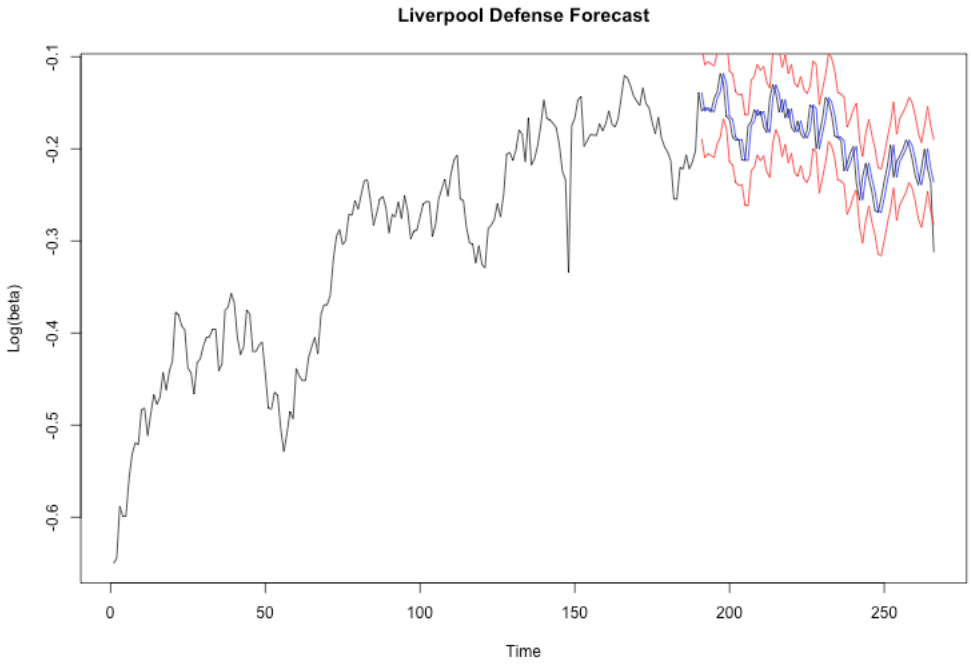


Figure 19 Continued



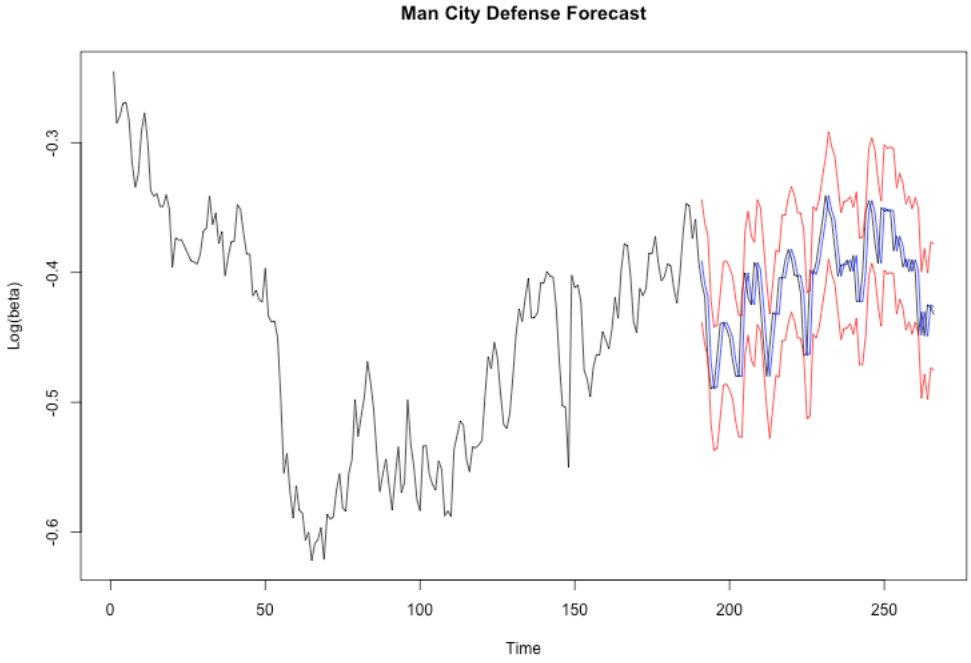
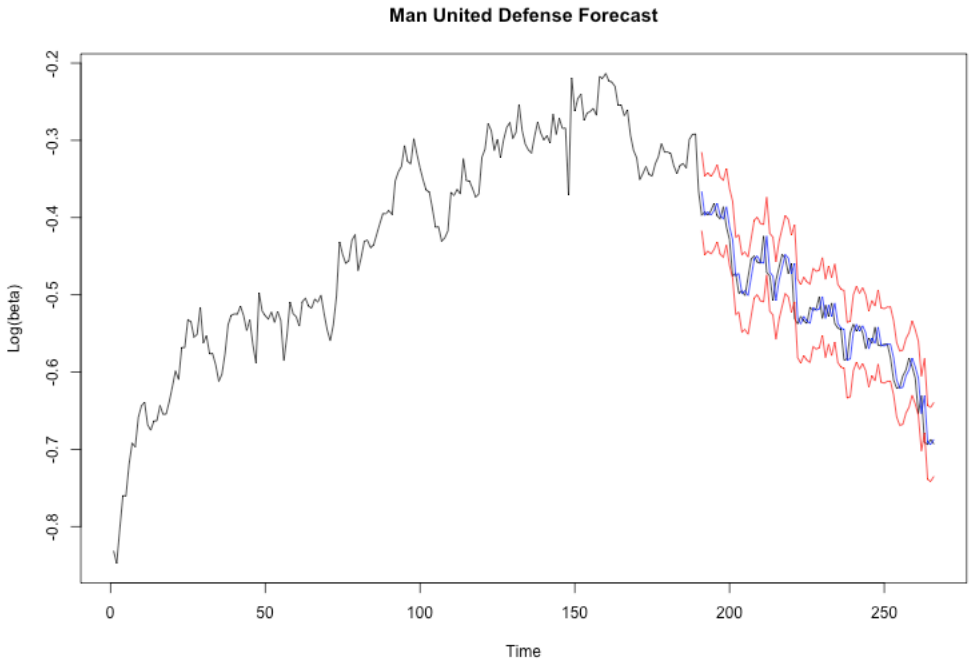


Figure 19 Continued



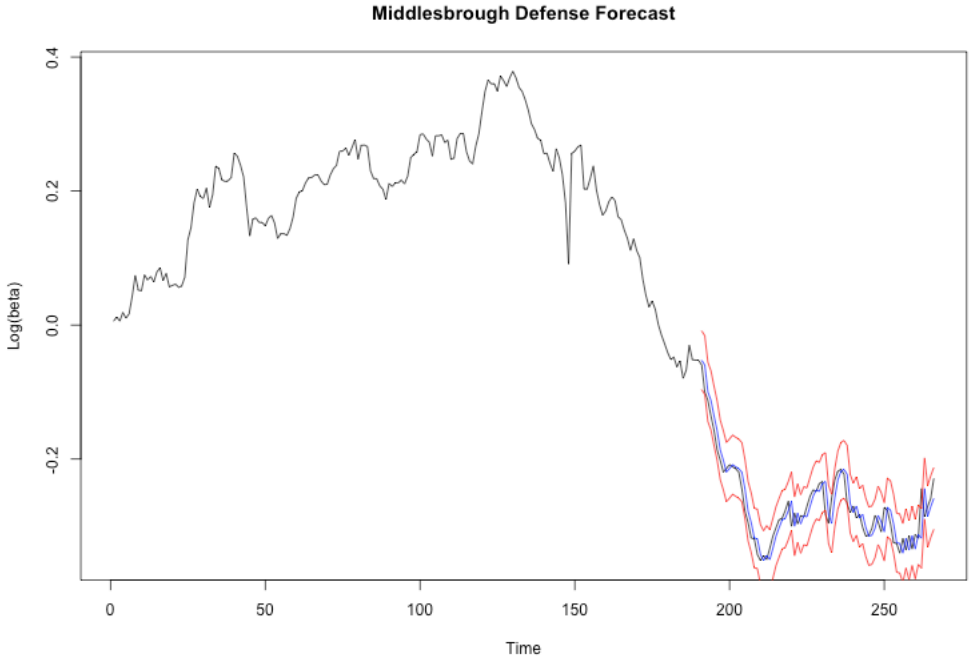
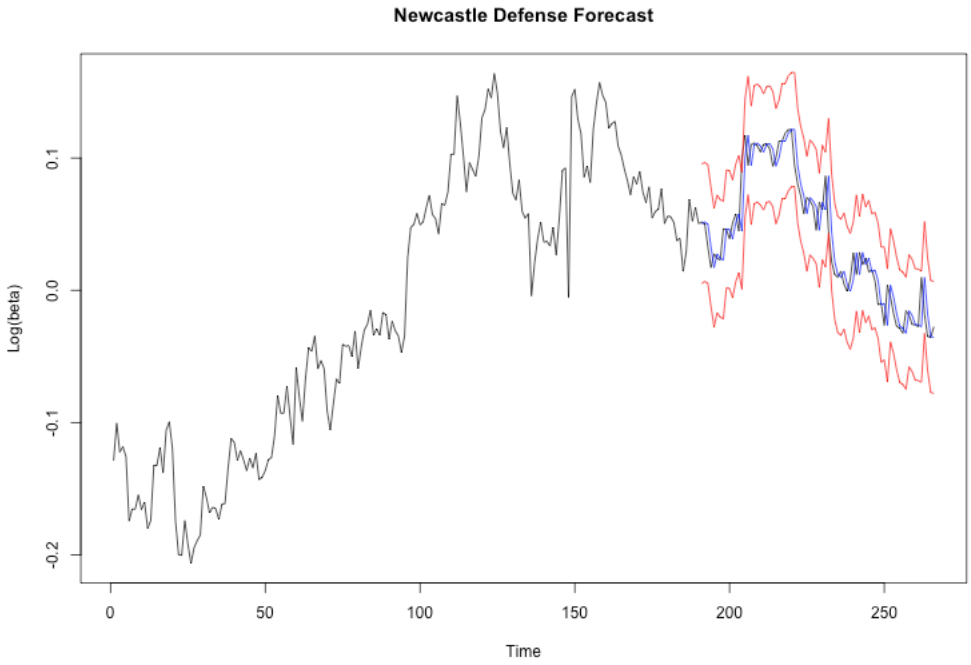


Figure 19 Continued



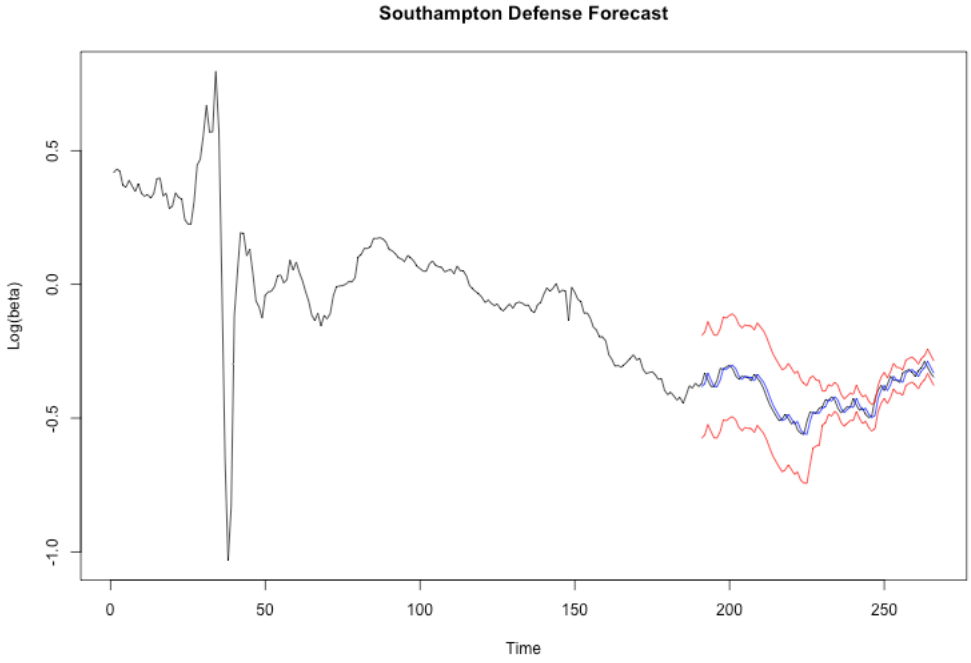
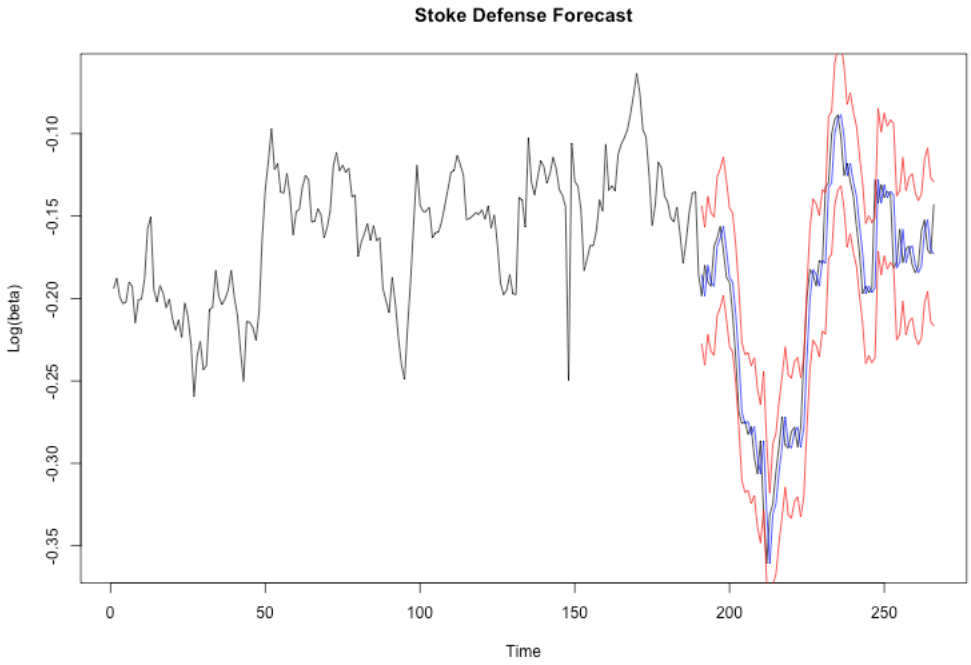


Figure 19 Continued



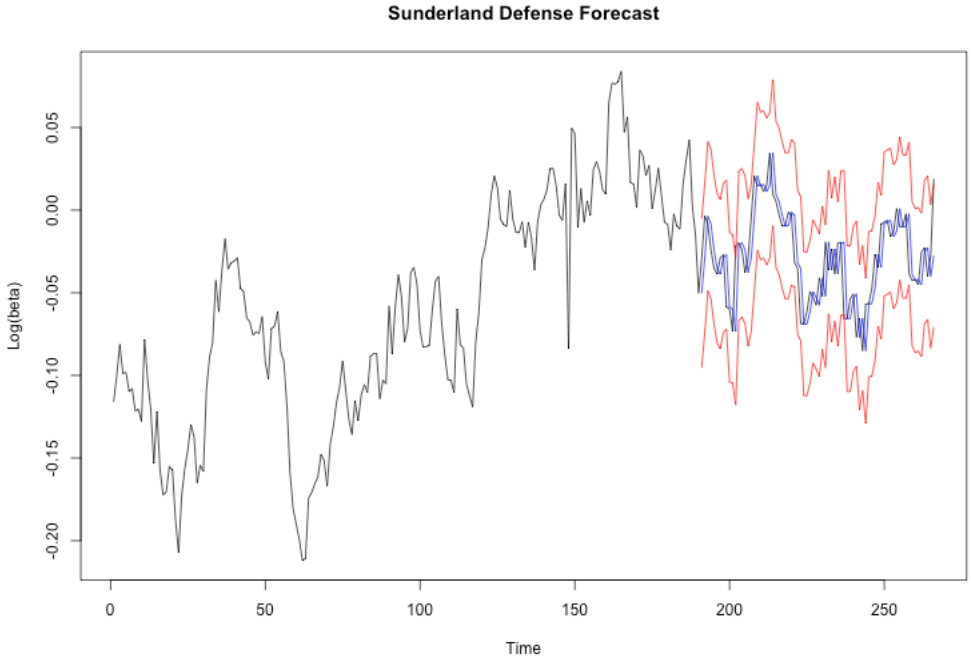
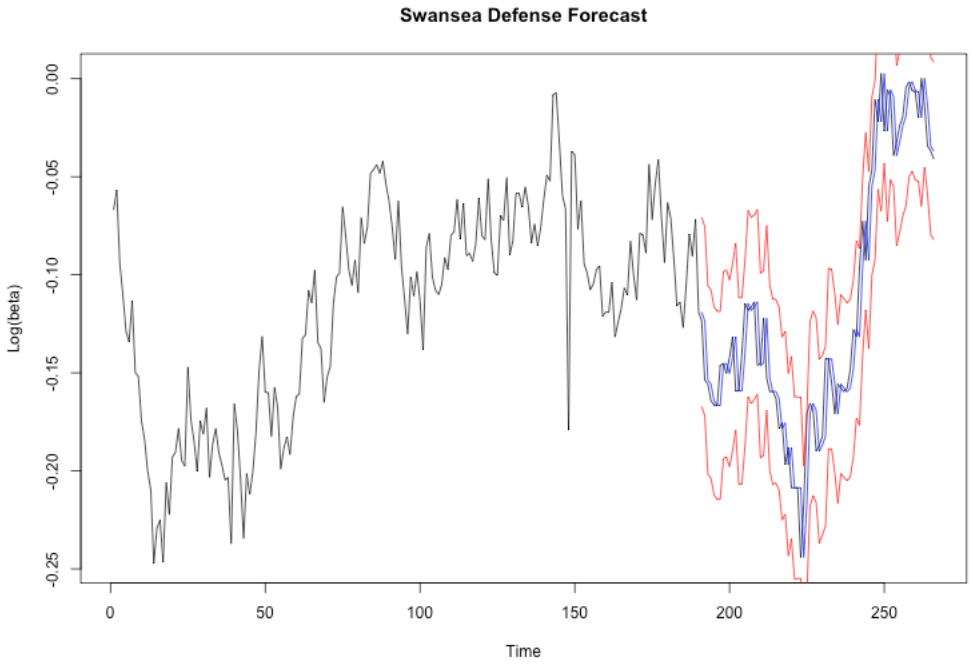


Figure 19 Continued



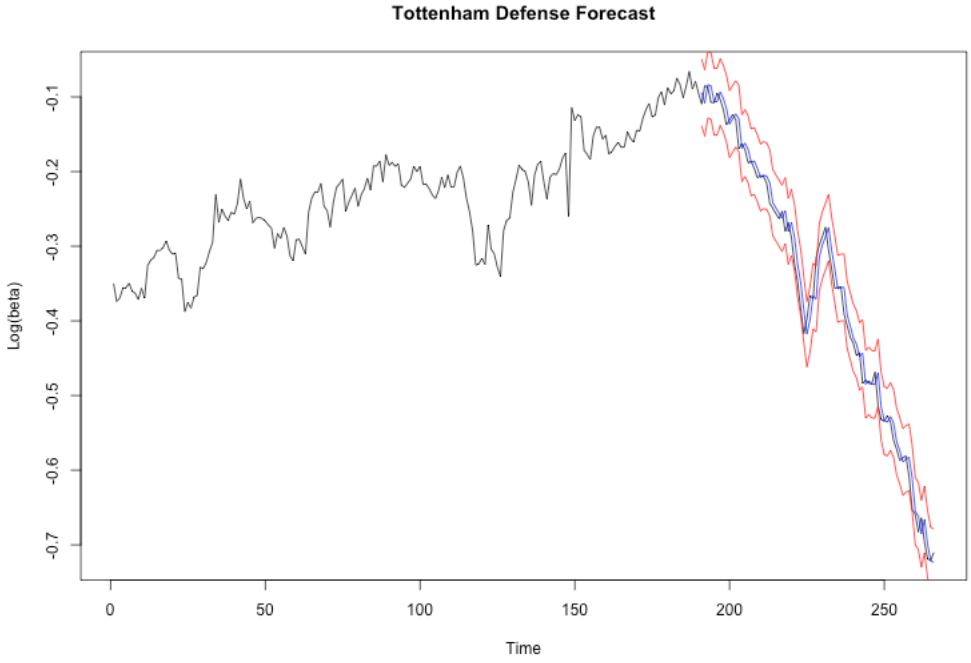
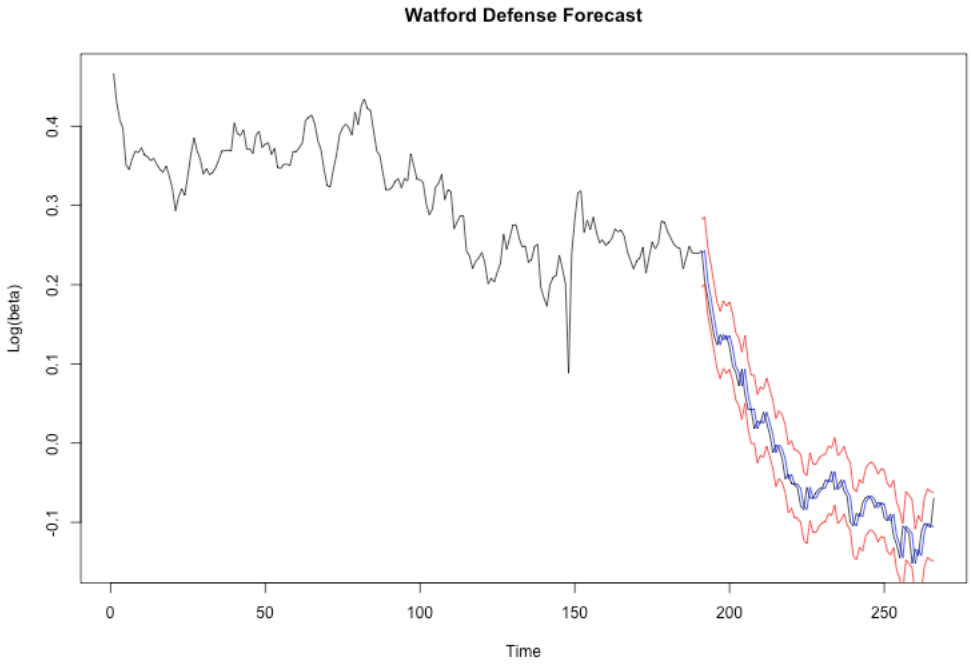


Figure 19 Continued



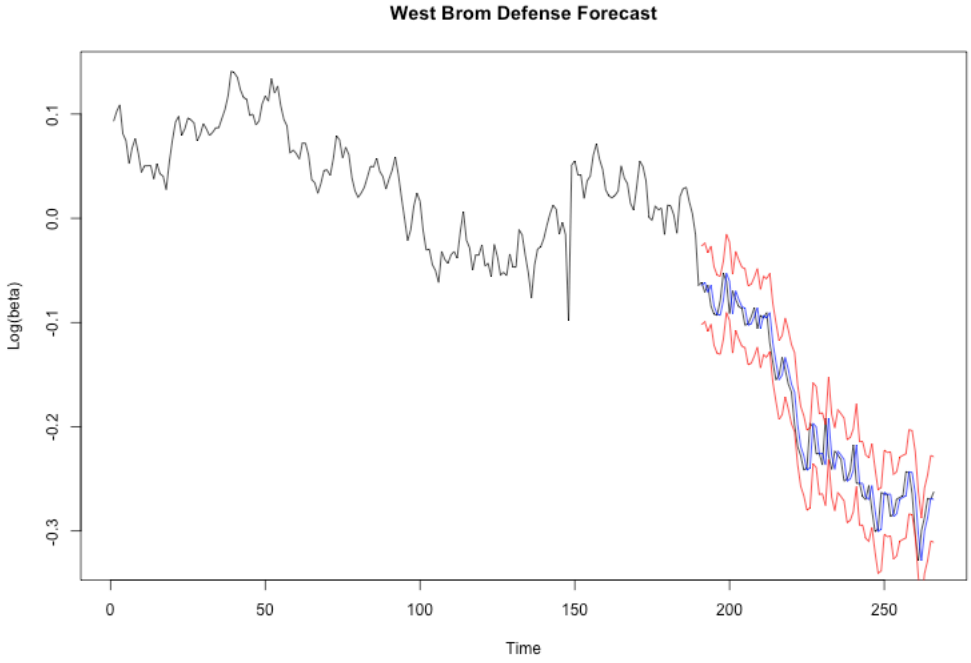


Figure 19 Continued

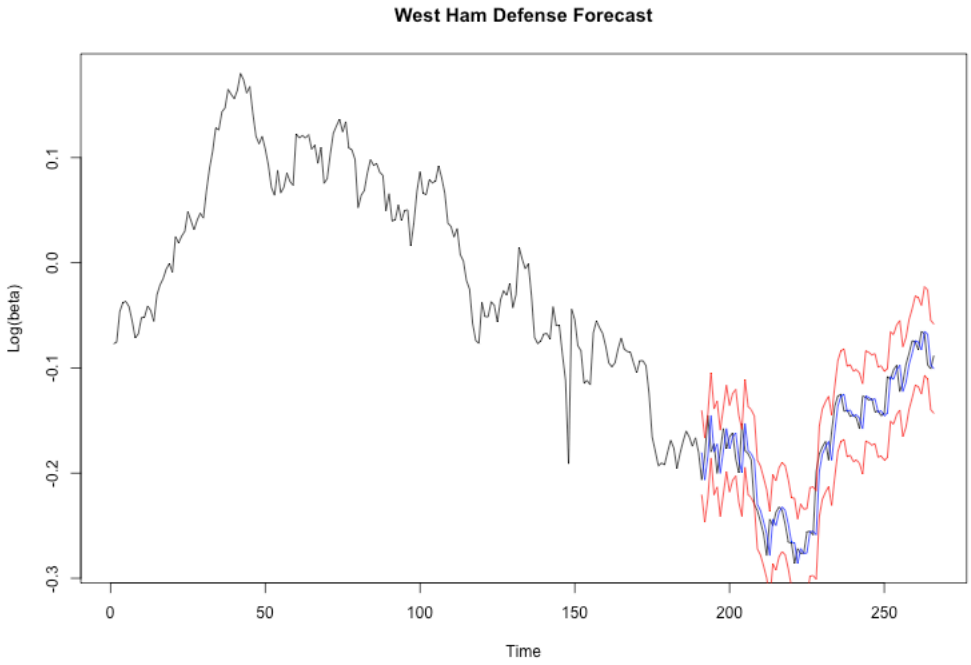


Figure 19 Continued

Appendix B – Forecasts from the random walk with drift model

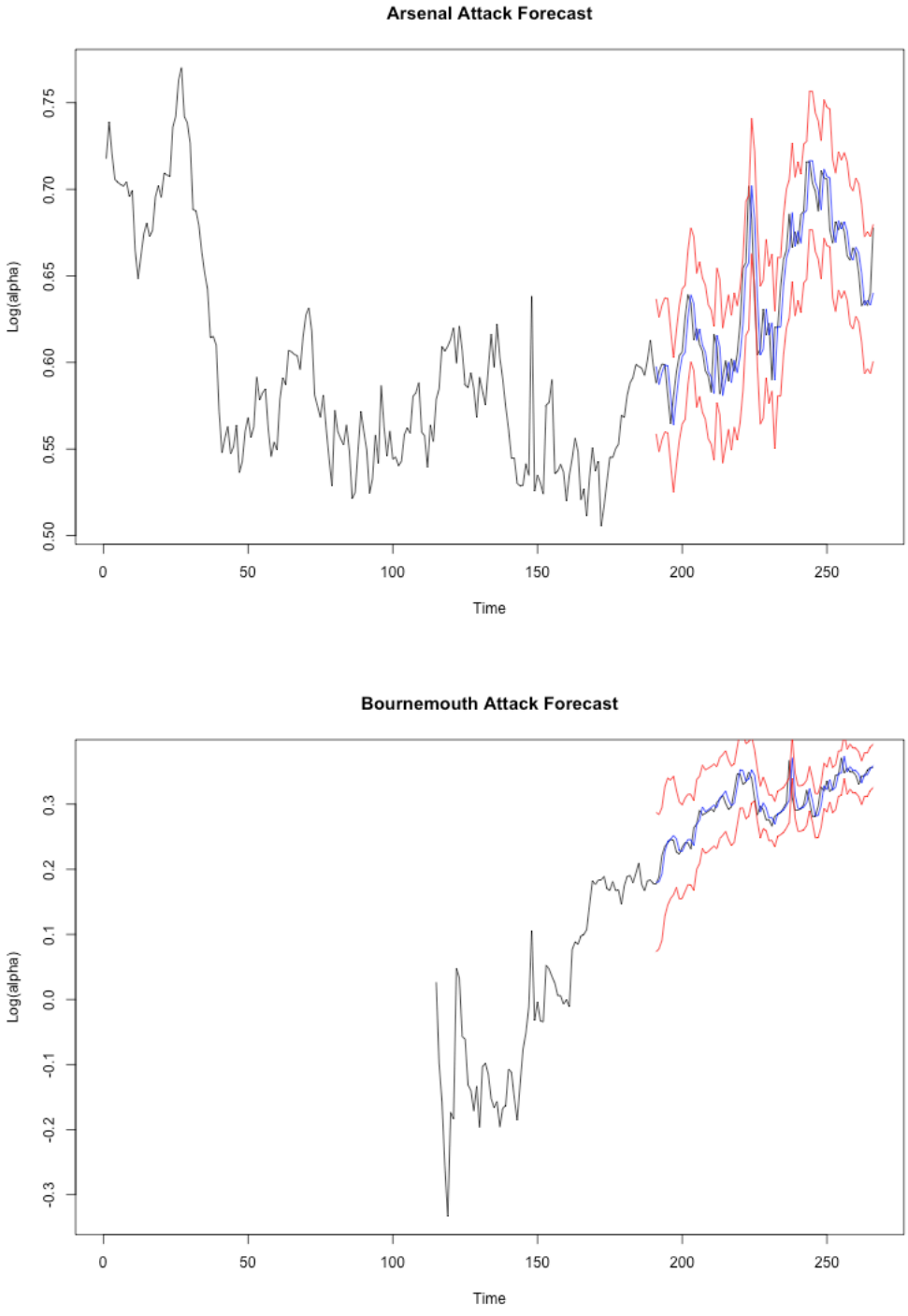
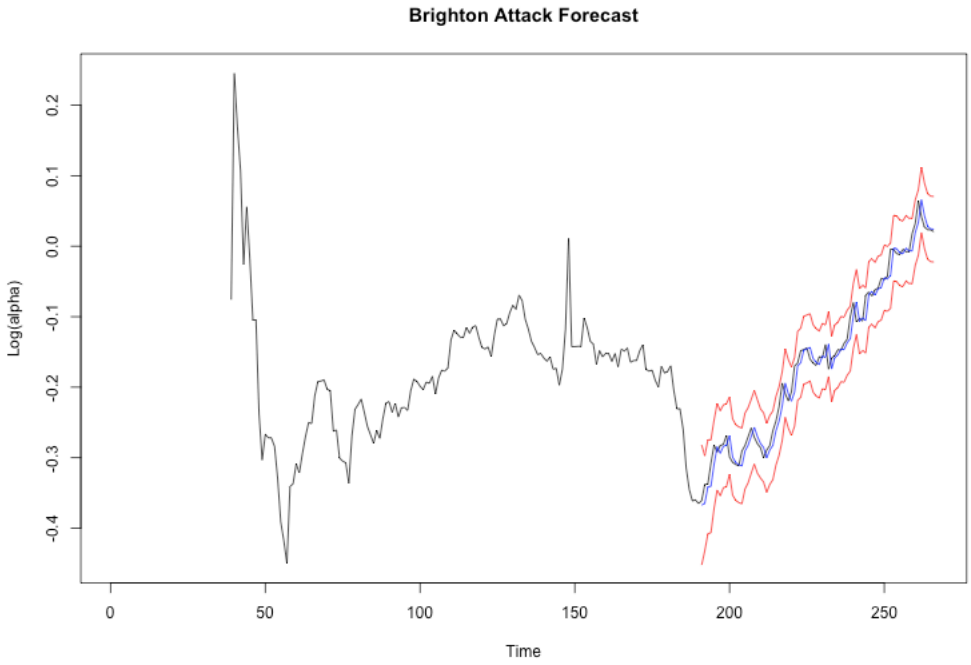


Figure 20 - (Model 2) Random Walk With Drift Forecasts – Attack



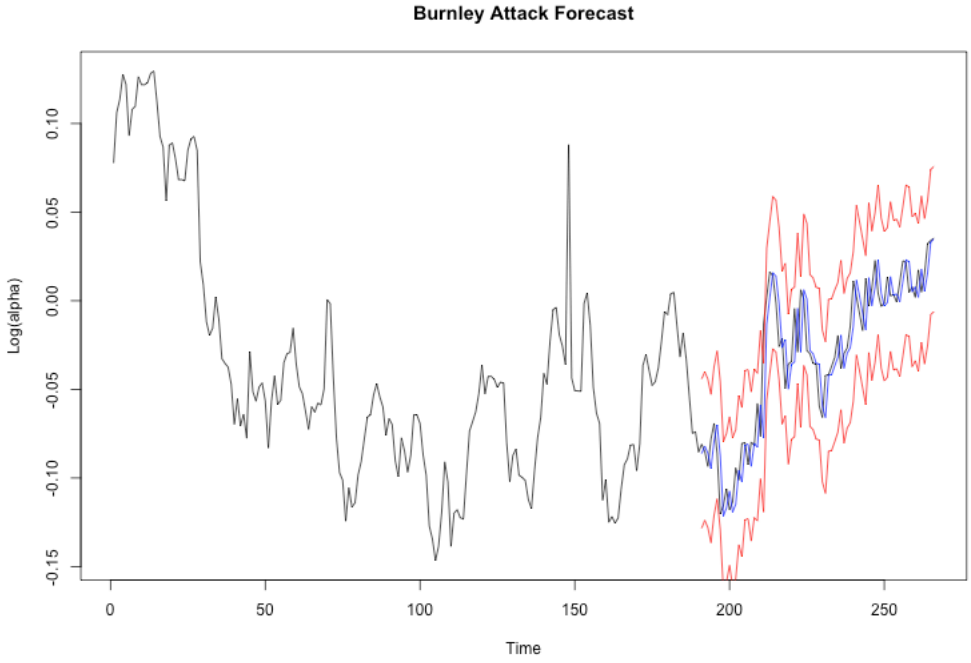
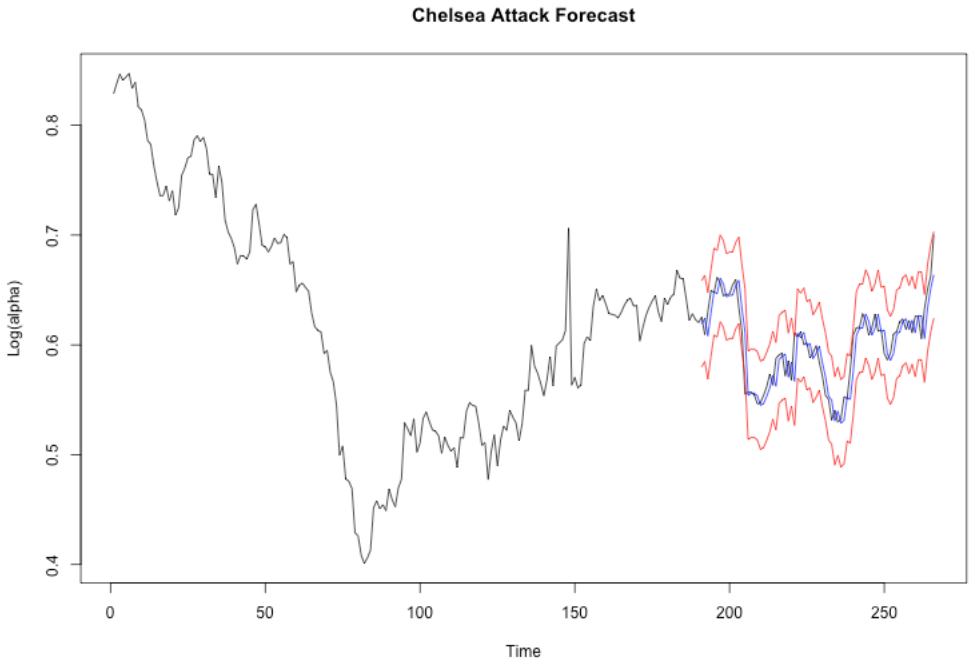


Figure 20 Continued



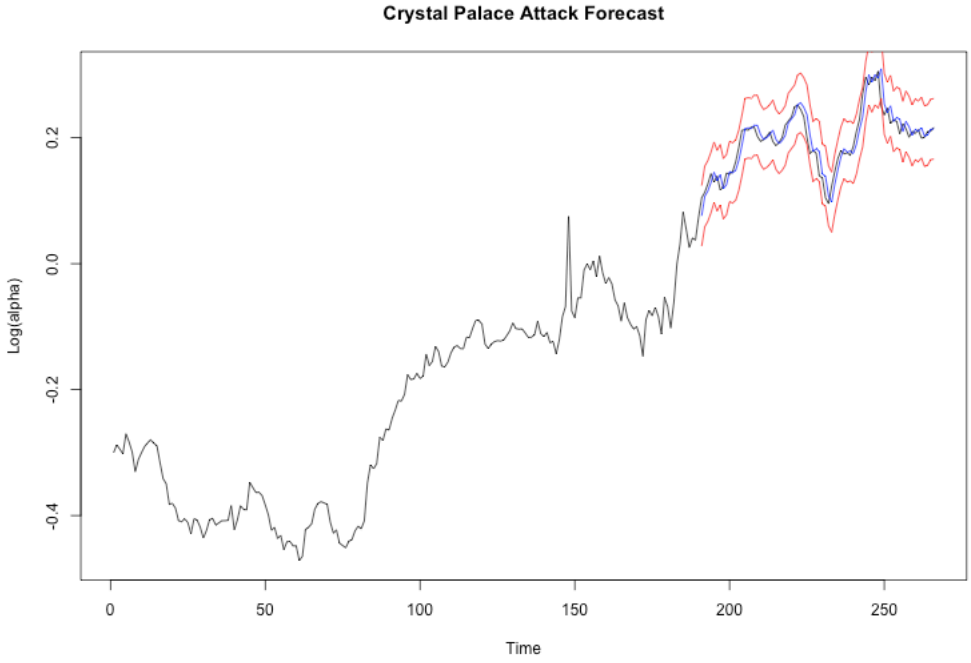
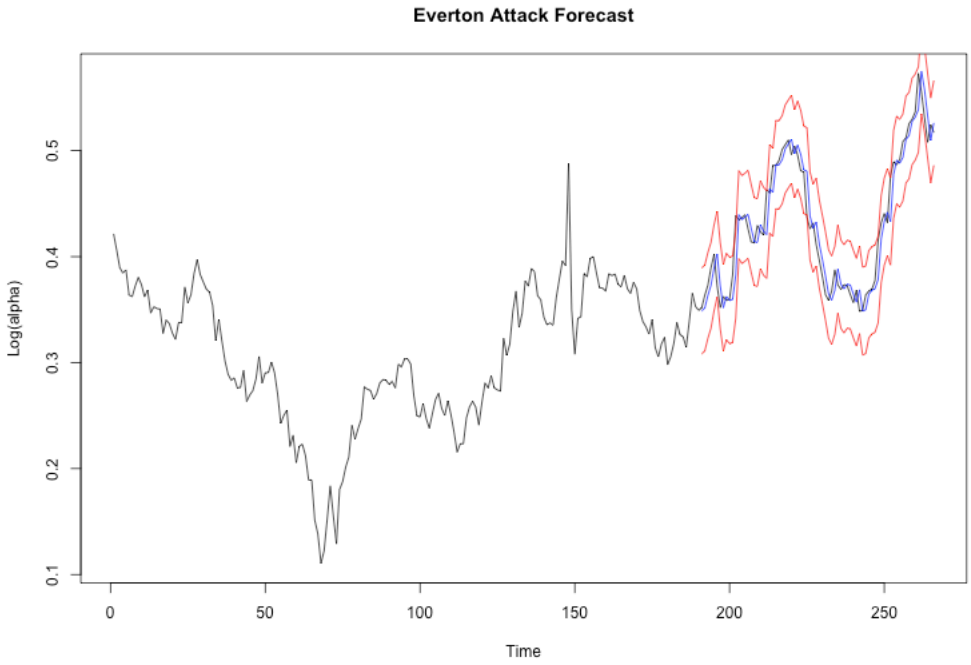


Figure 20 Continued



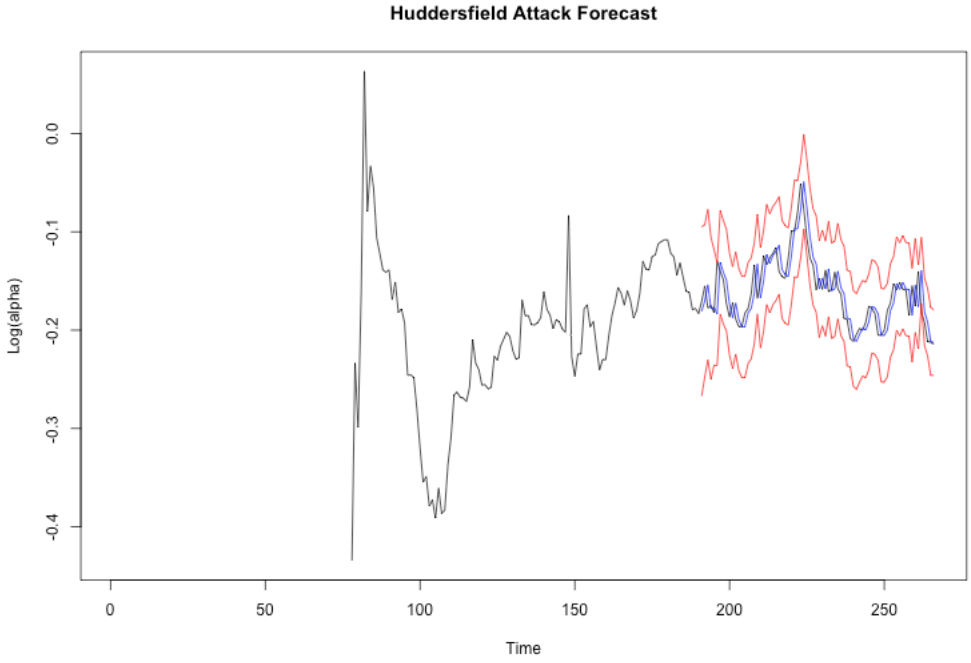
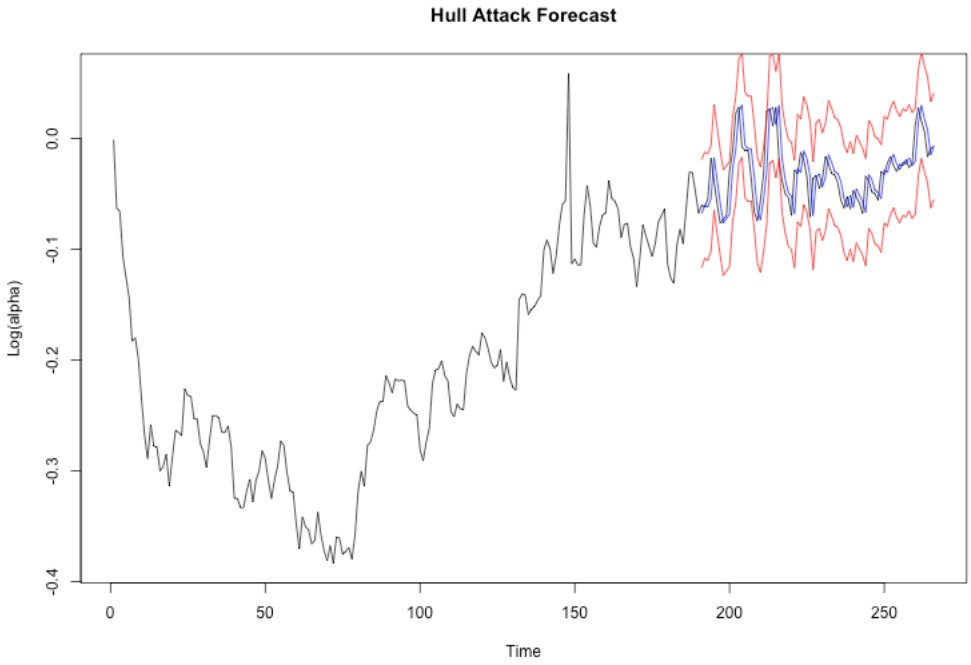


Figure 20 Continued



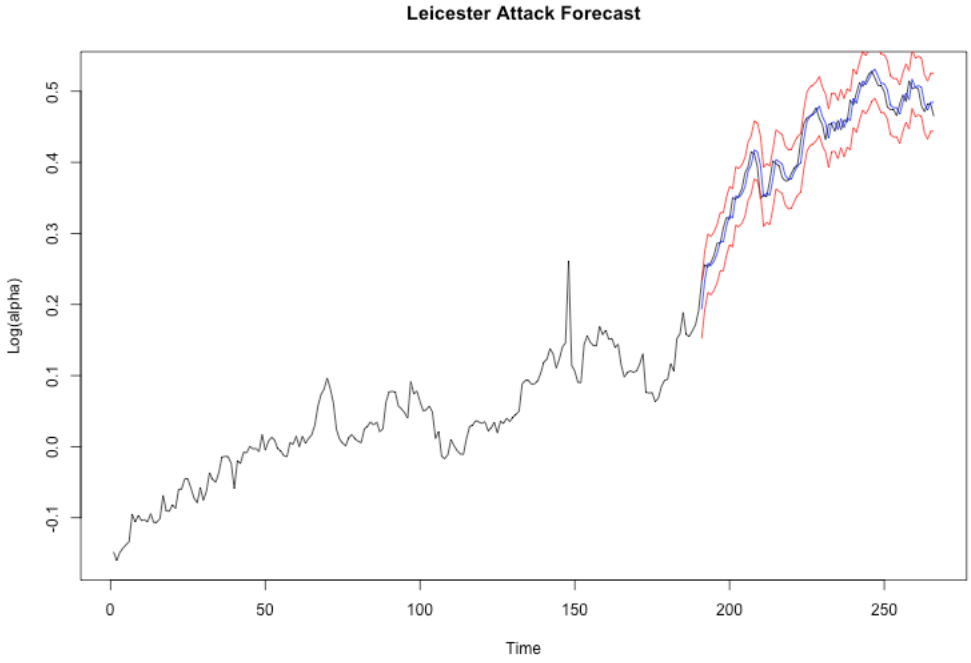
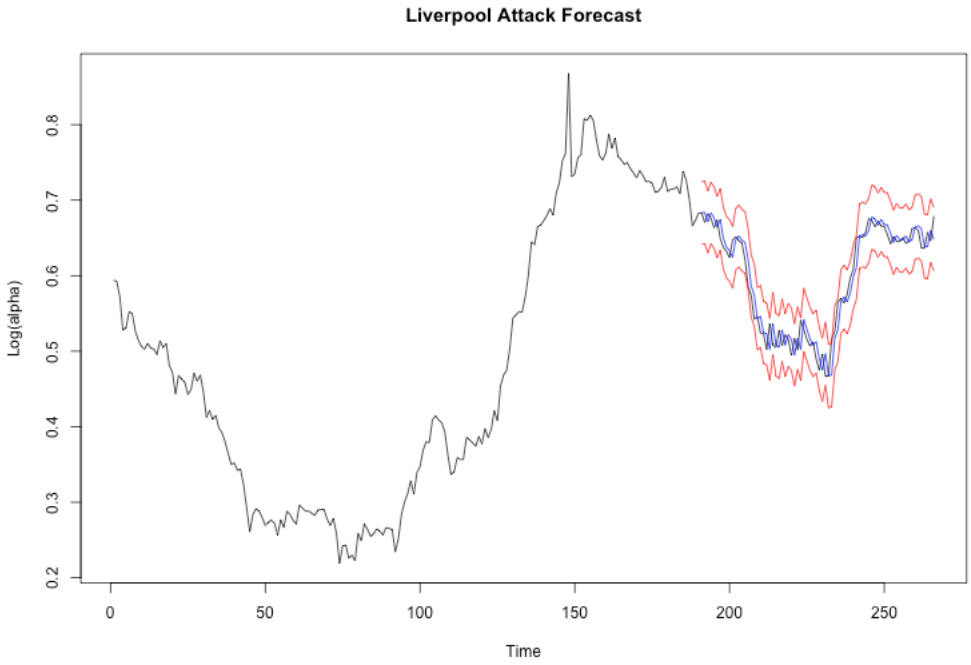


Figure 20 Continued



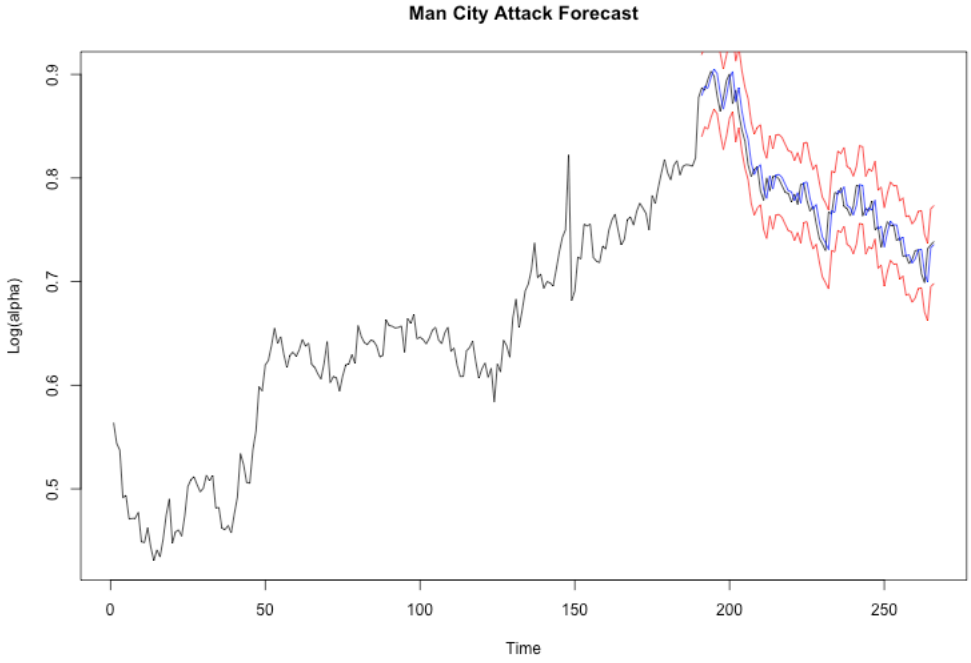
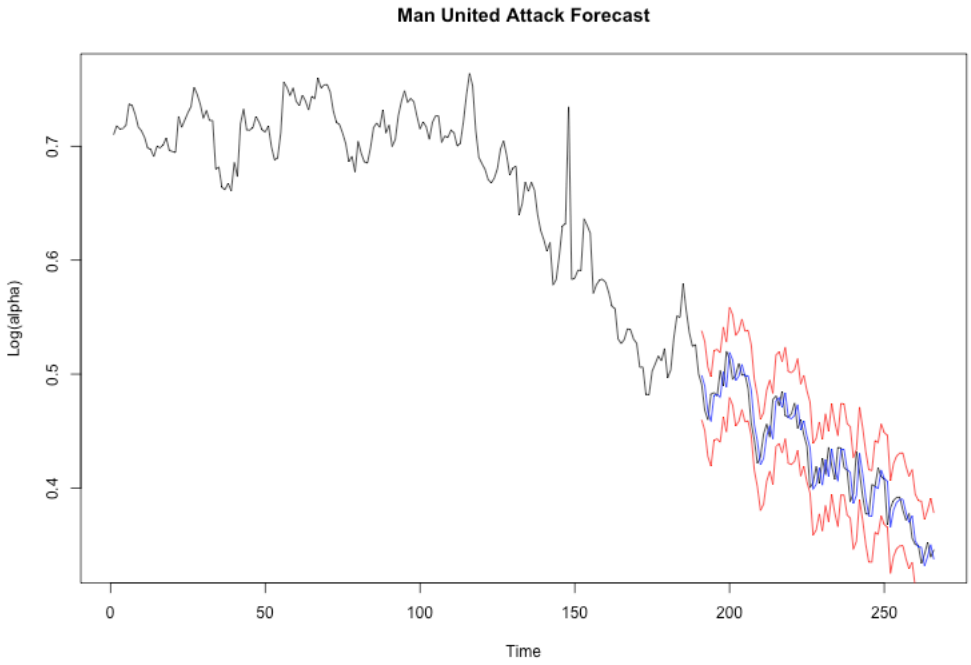


Figure 20 Continued



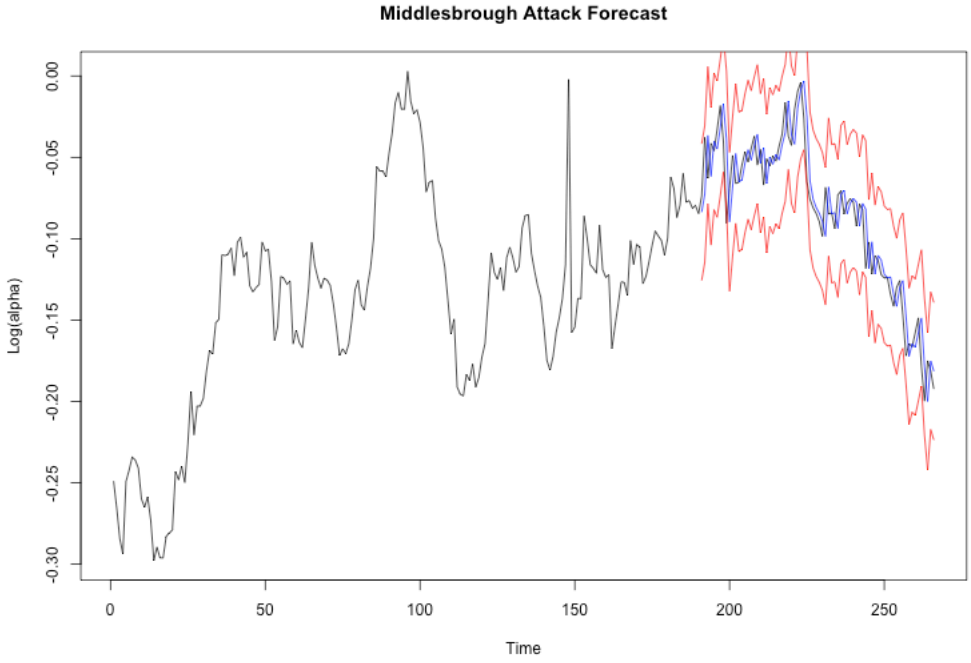
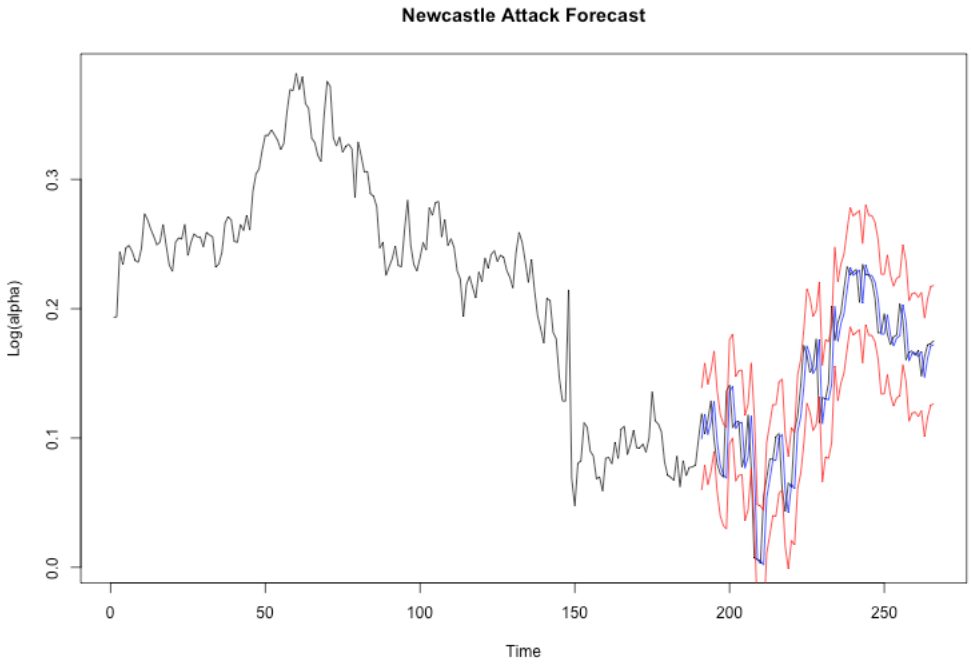


Figure 20 Continued



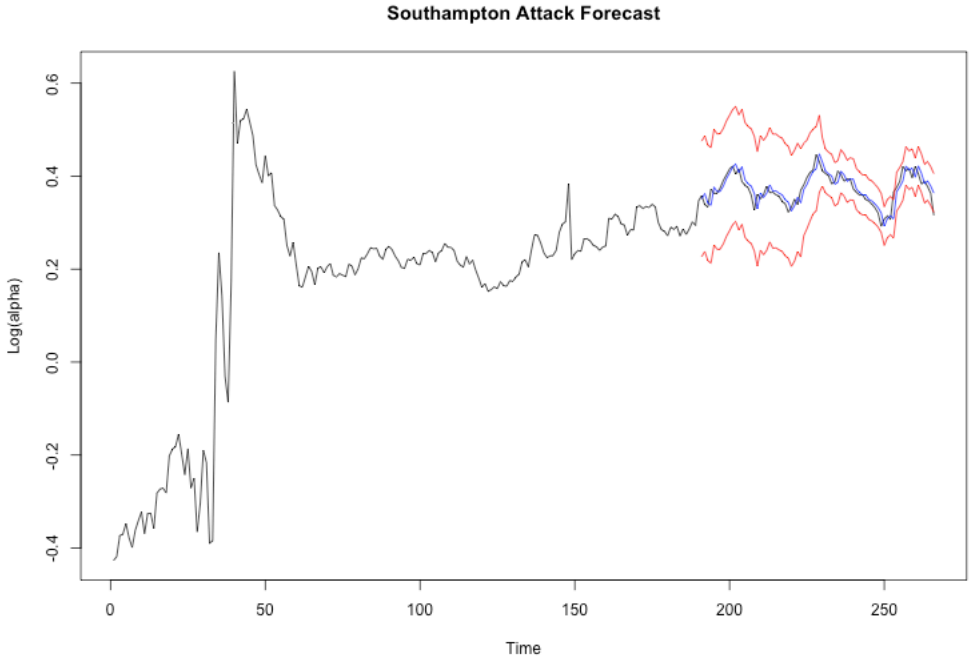
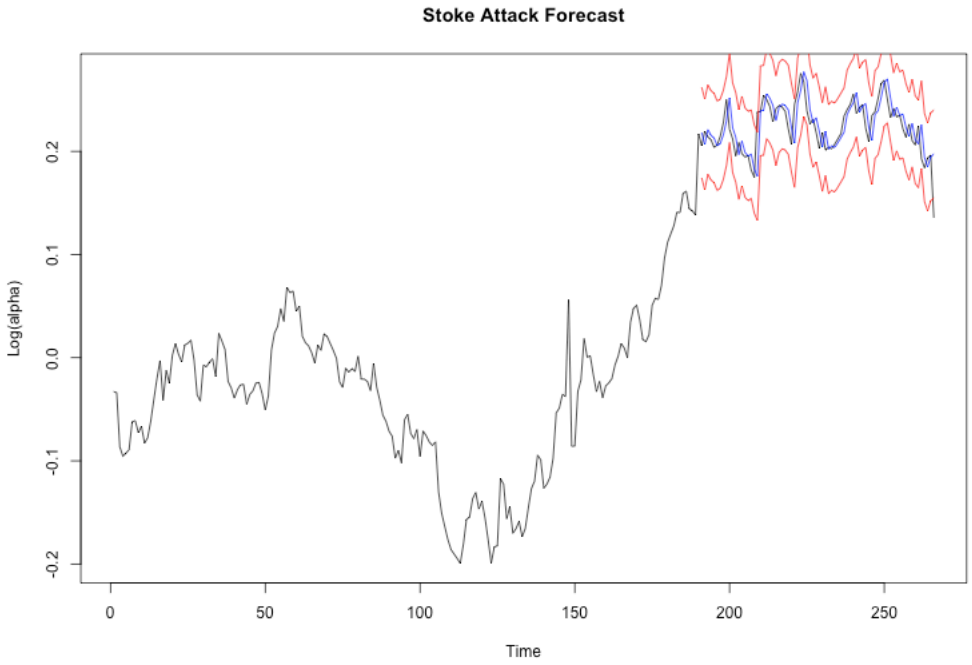


Figure 20 Continued



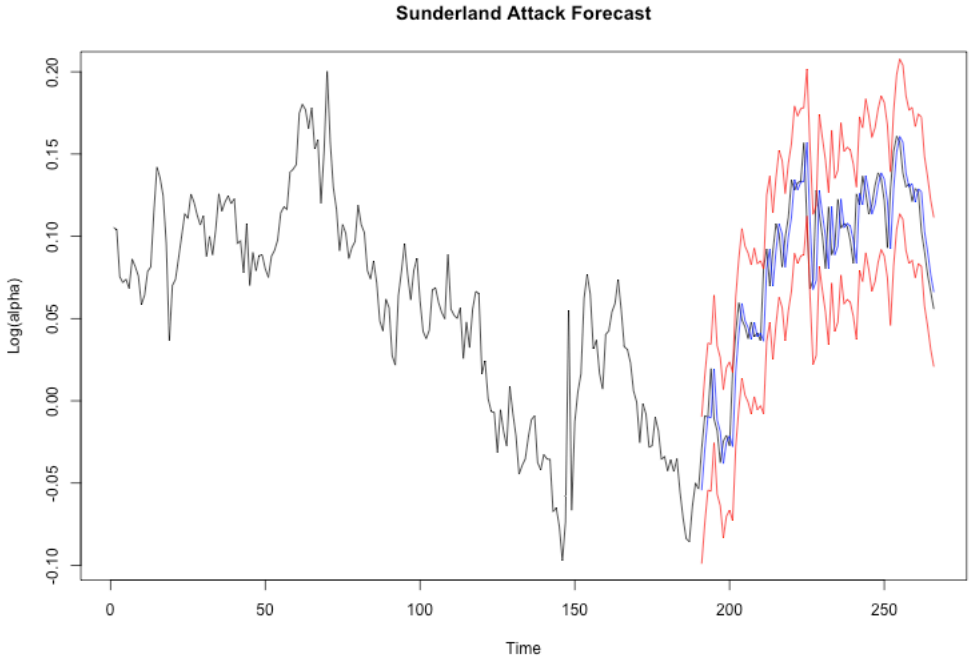
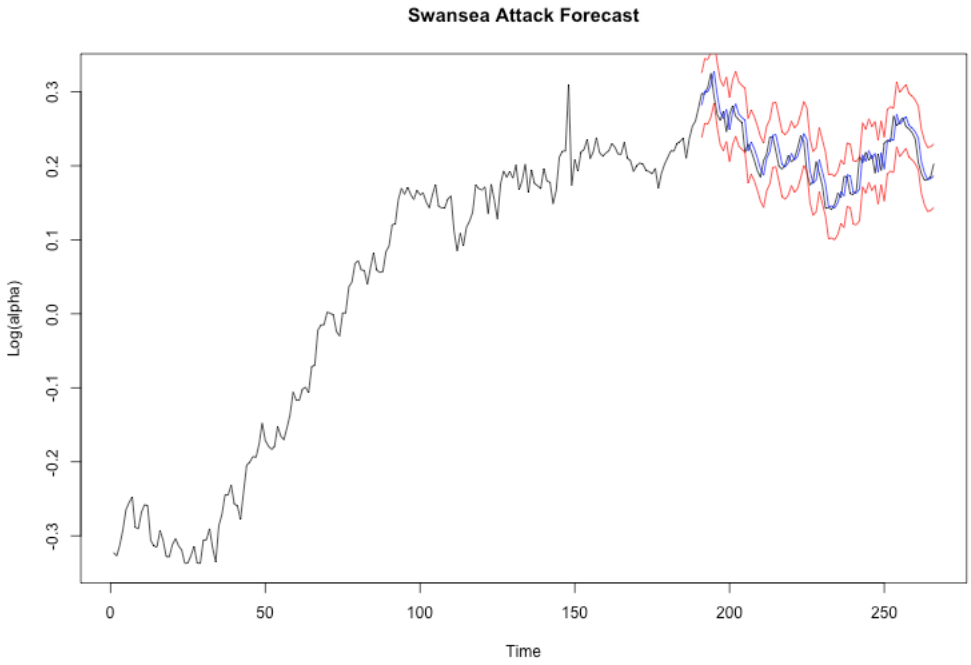


Figure 20 Continued



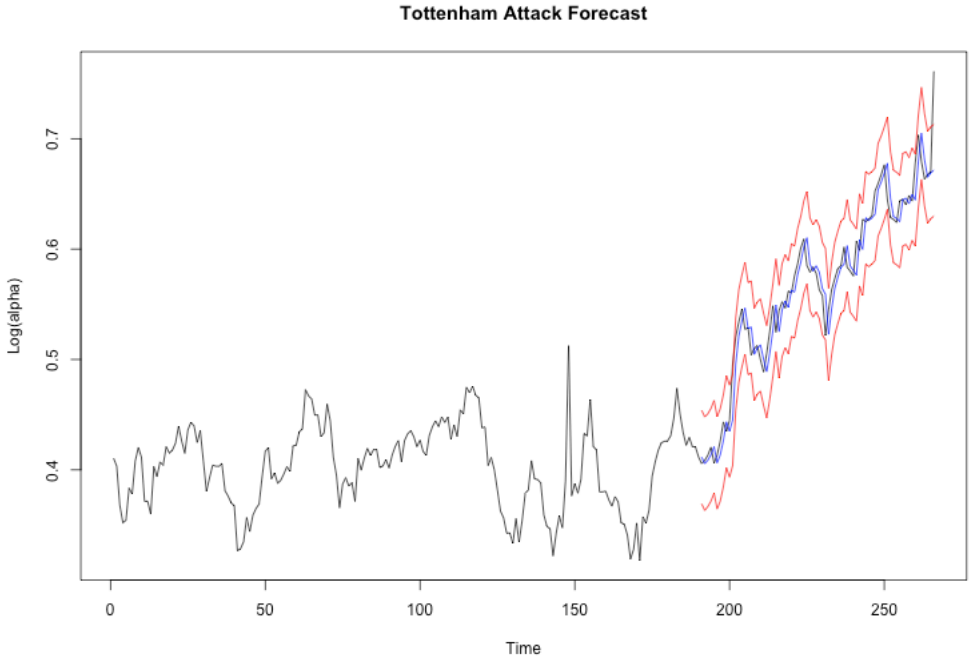
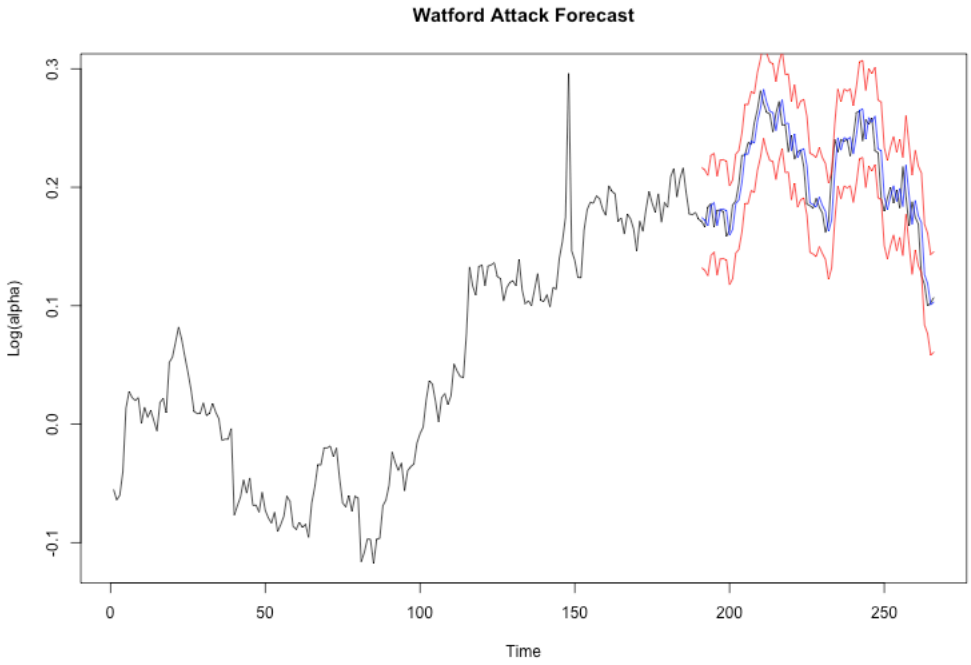


Figure 20 Continued



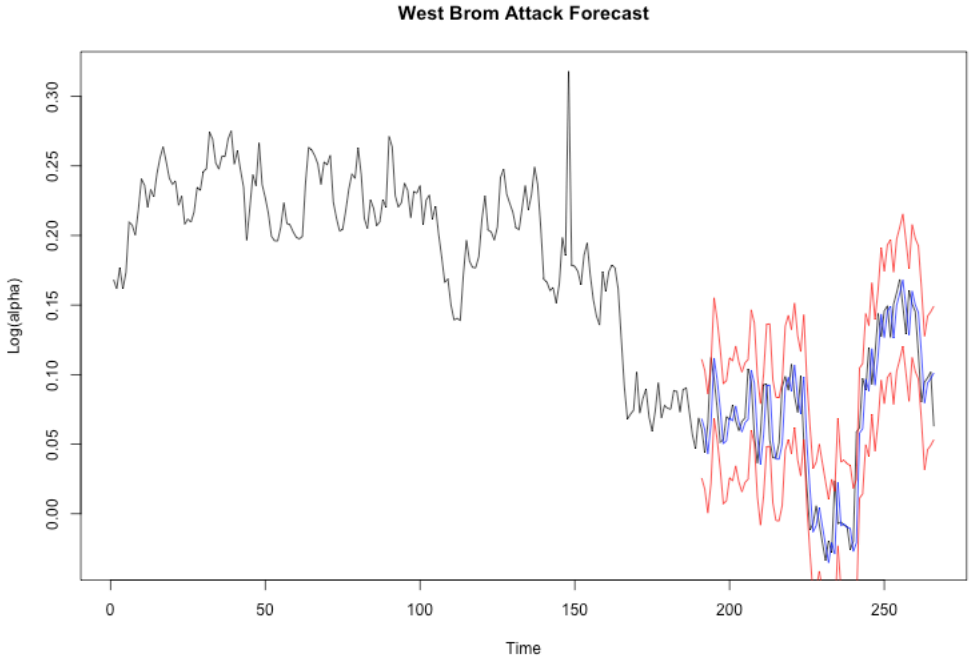


Figure 20 Continued

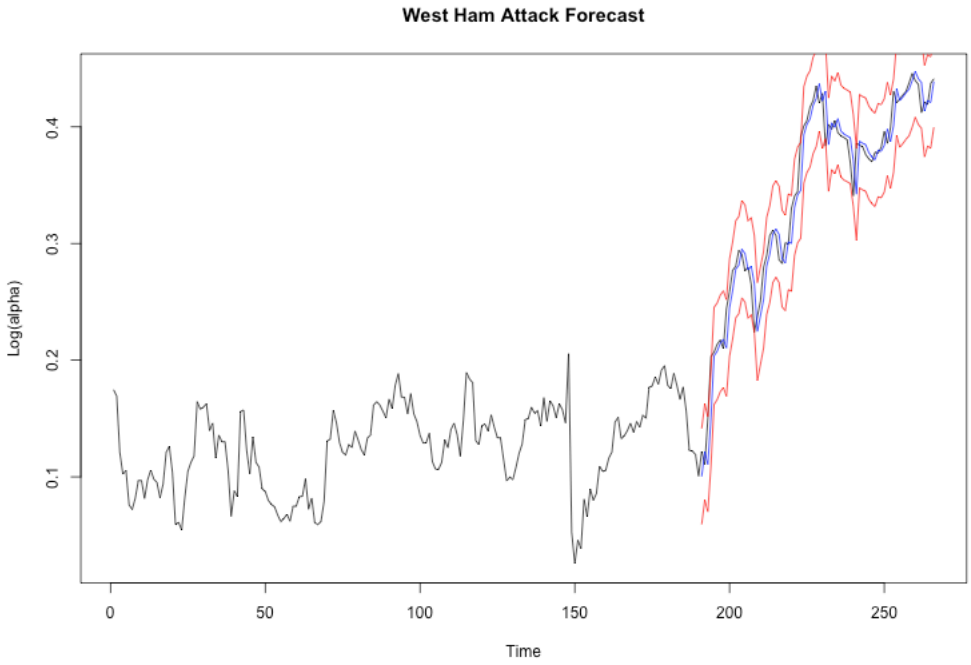


Figure 20 Continued

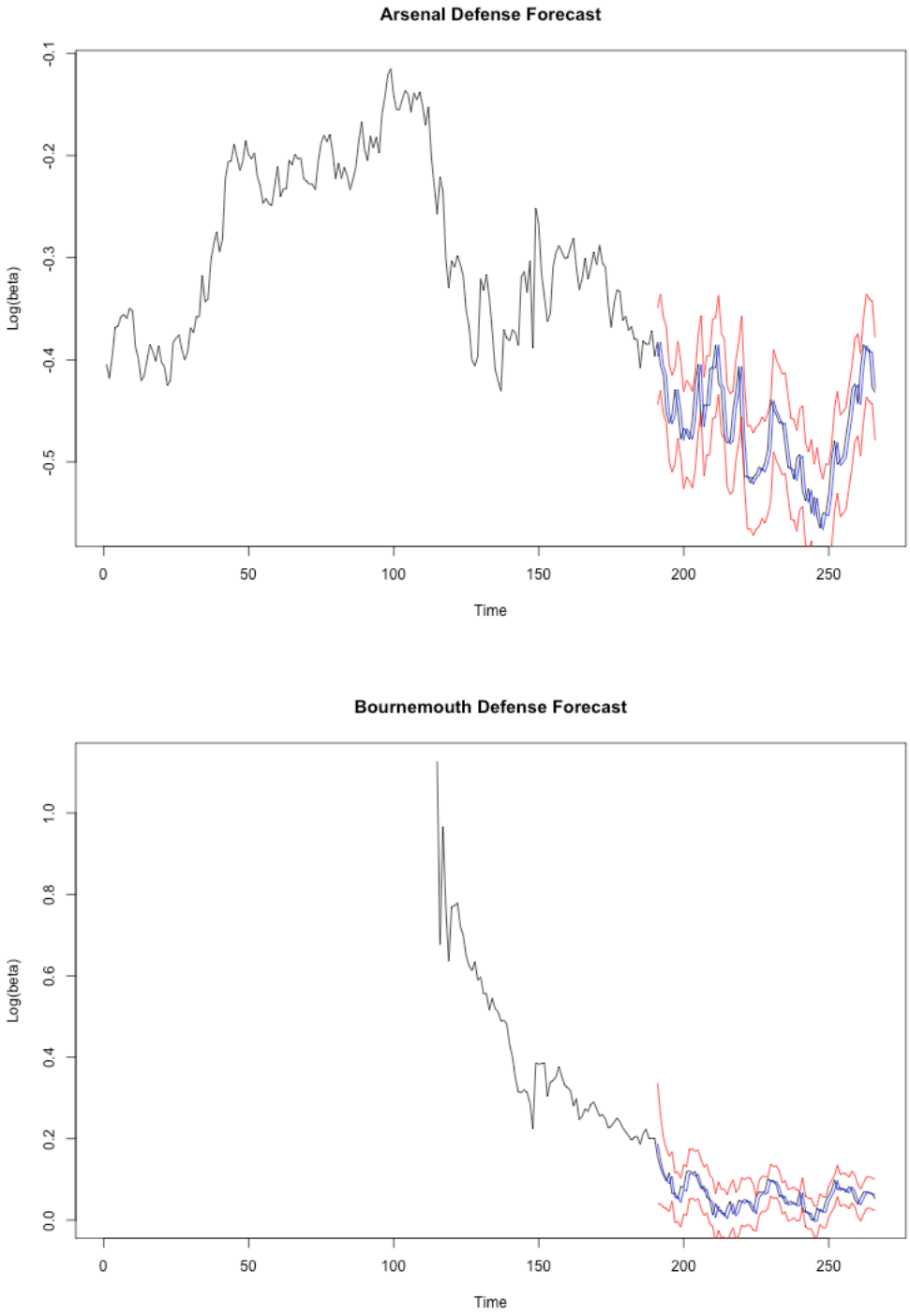
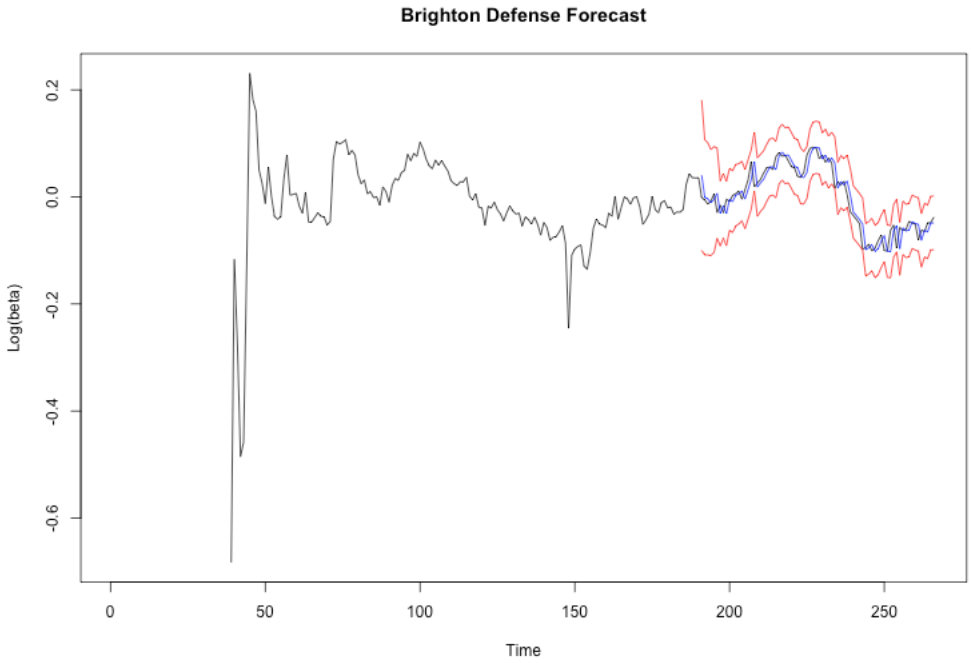


Figure 21 - (Model 2) Random Walk With Drift Forecasts - Defence



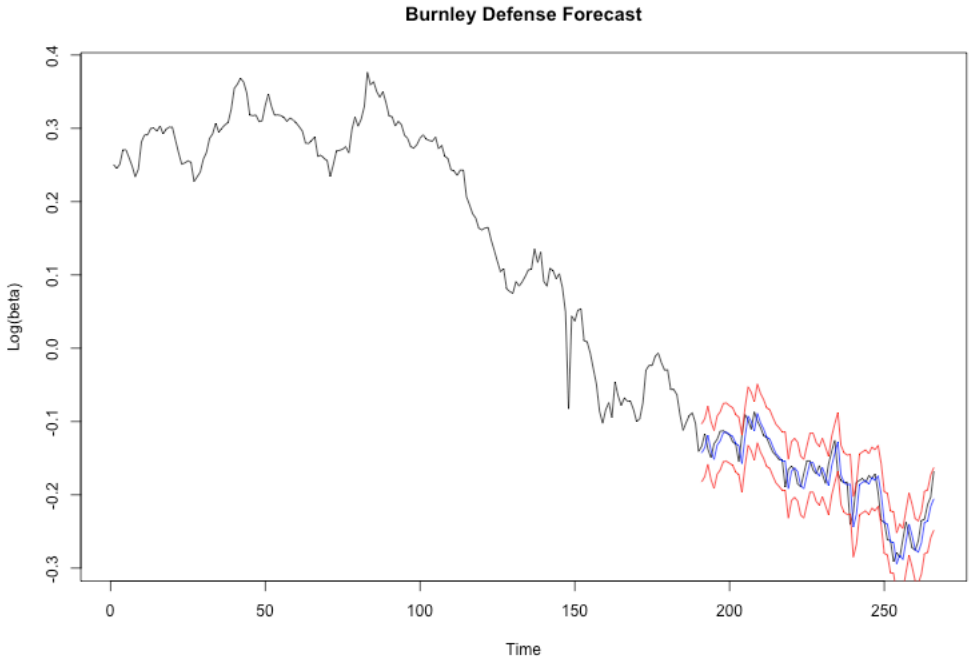
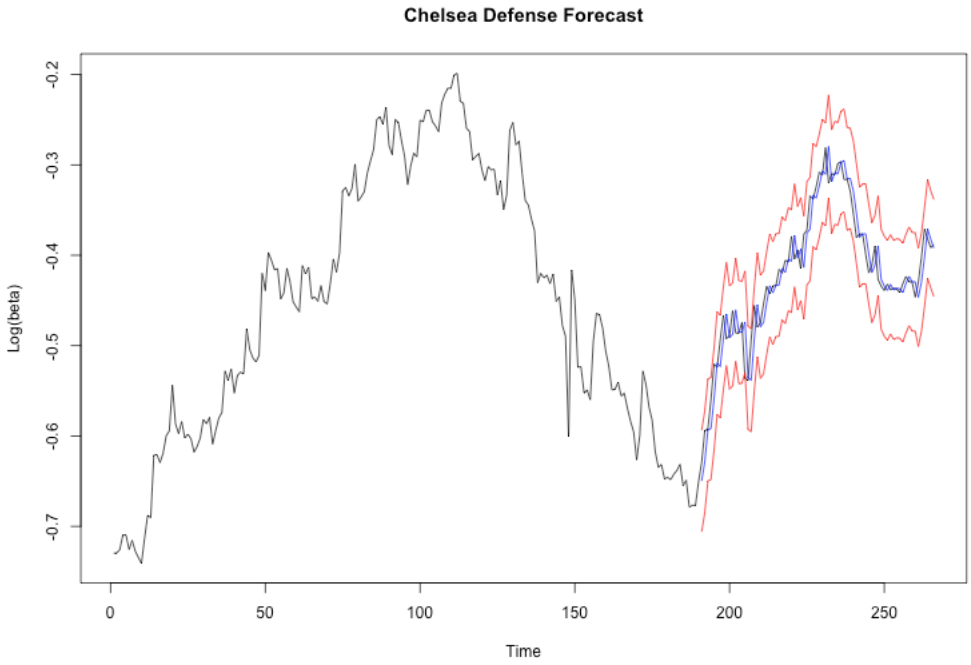


Figure 21 Continued



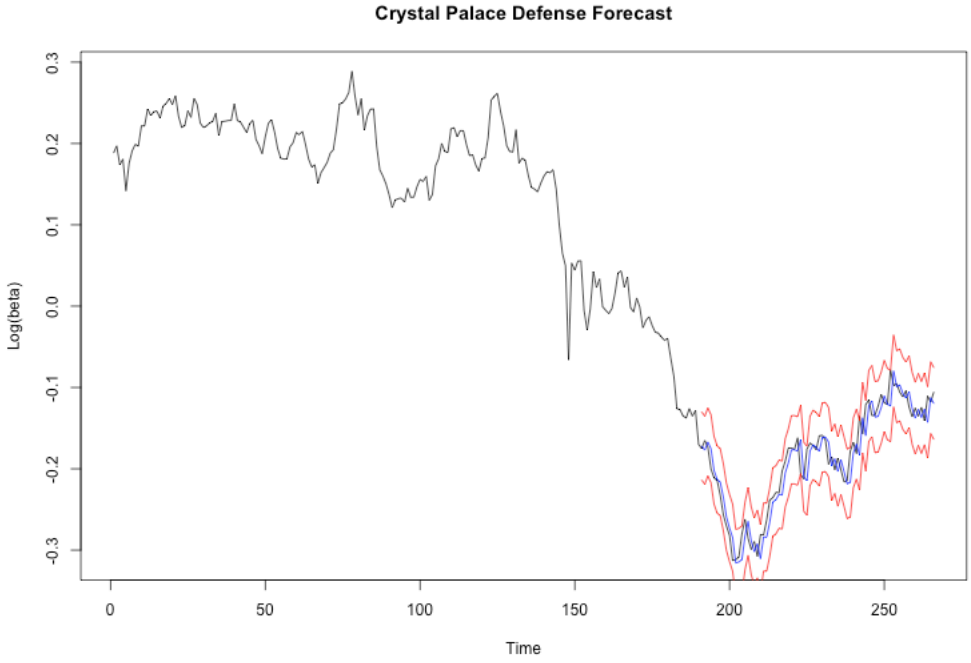
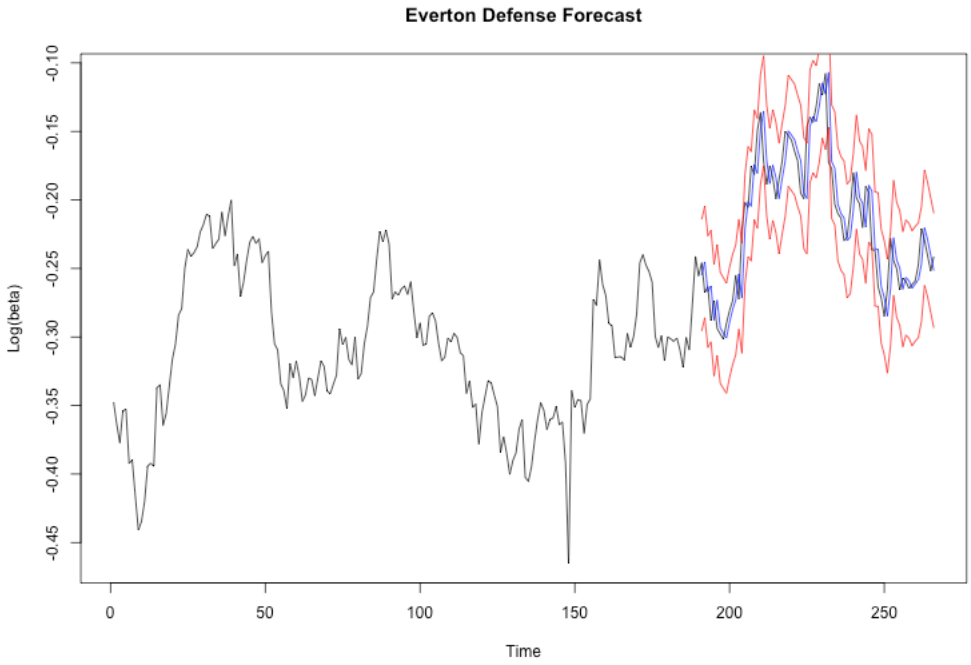


Figure 21 Continued



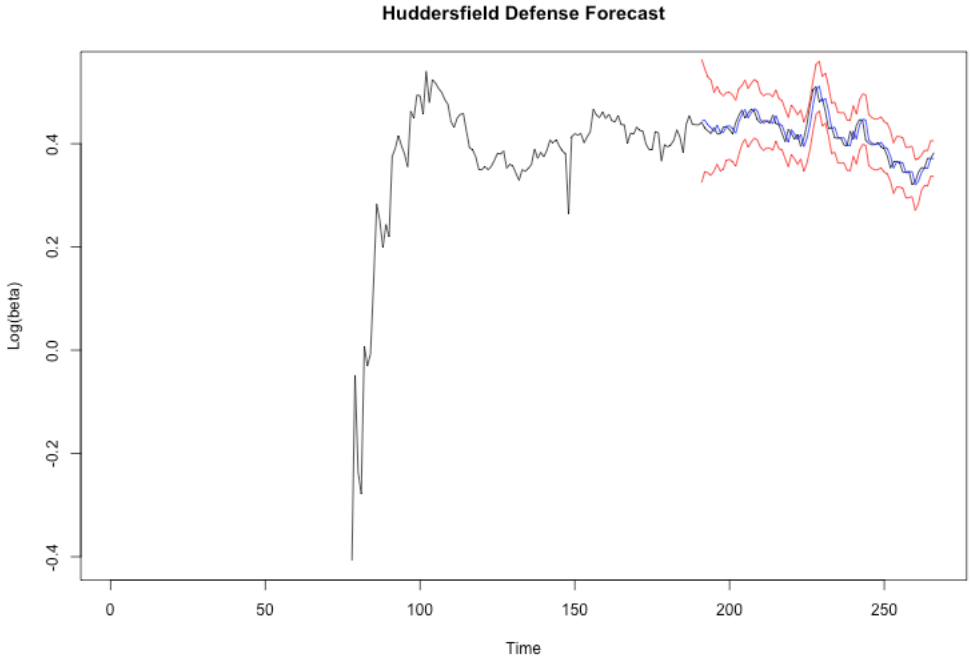
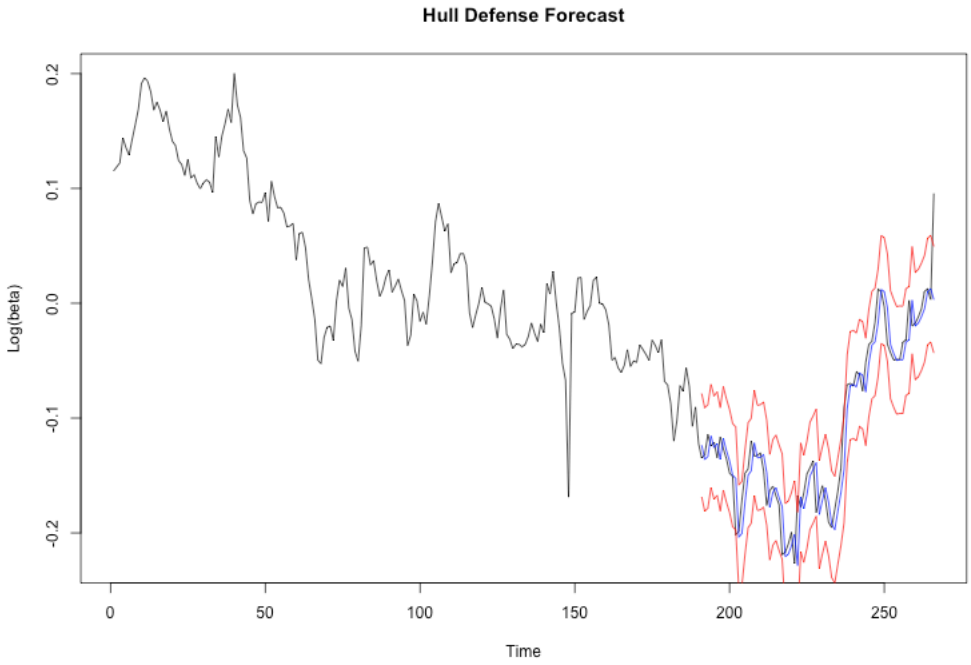


Figure 21 Continued



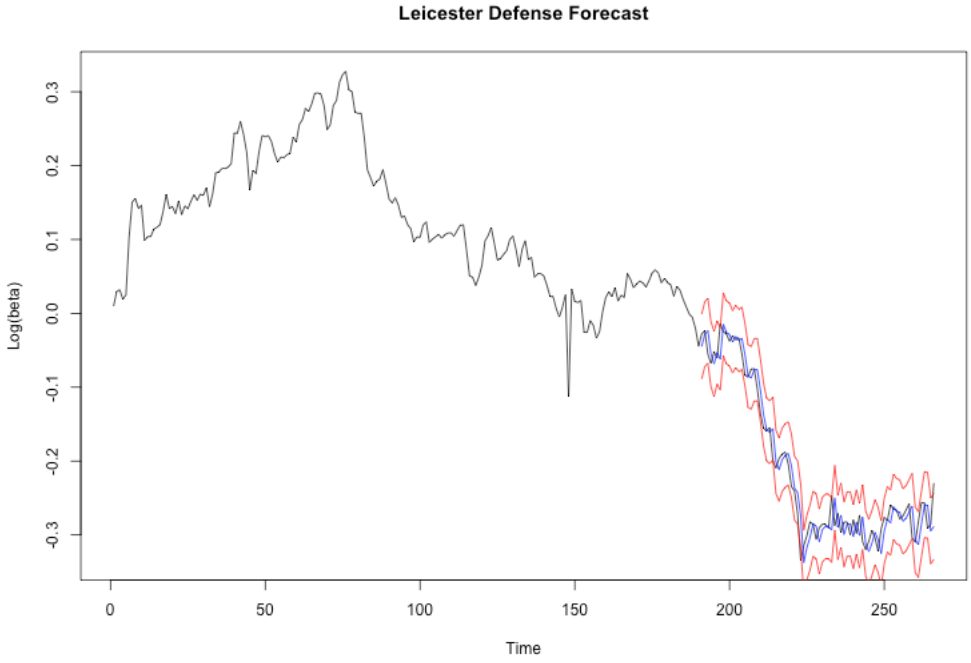
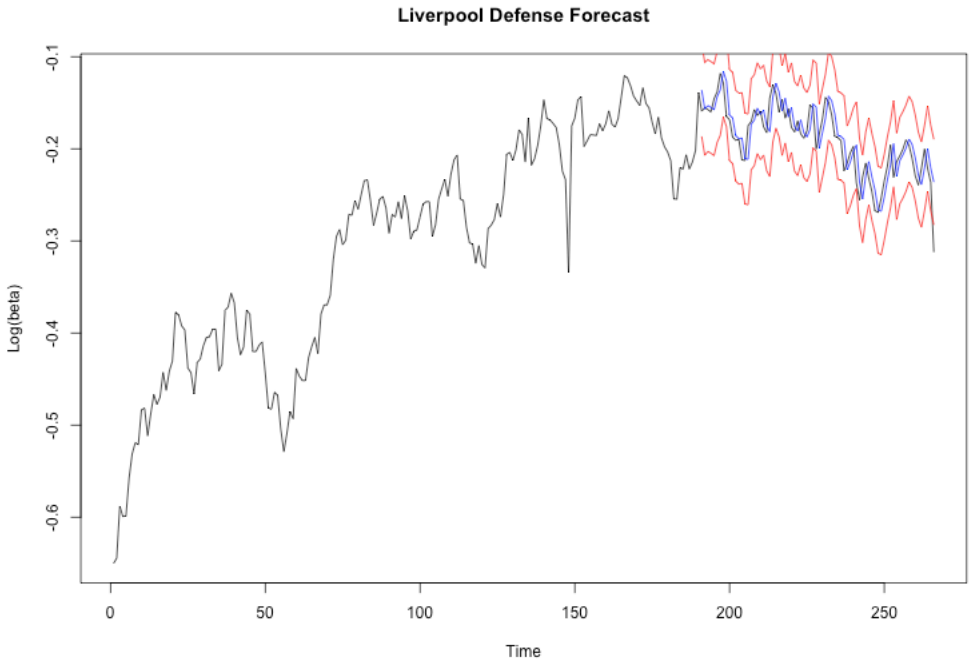


Figure 21 Continued



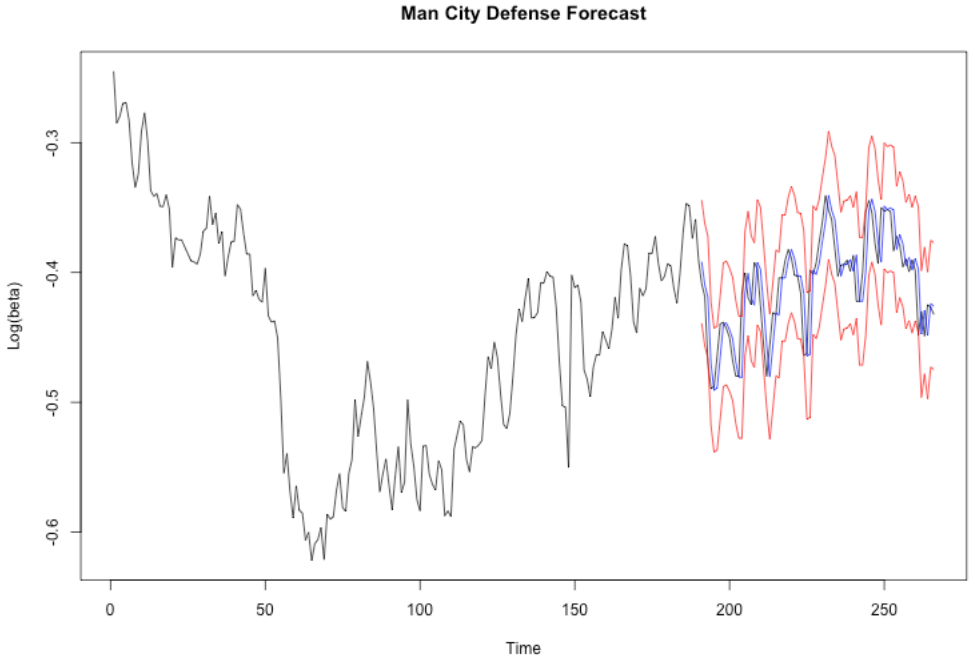
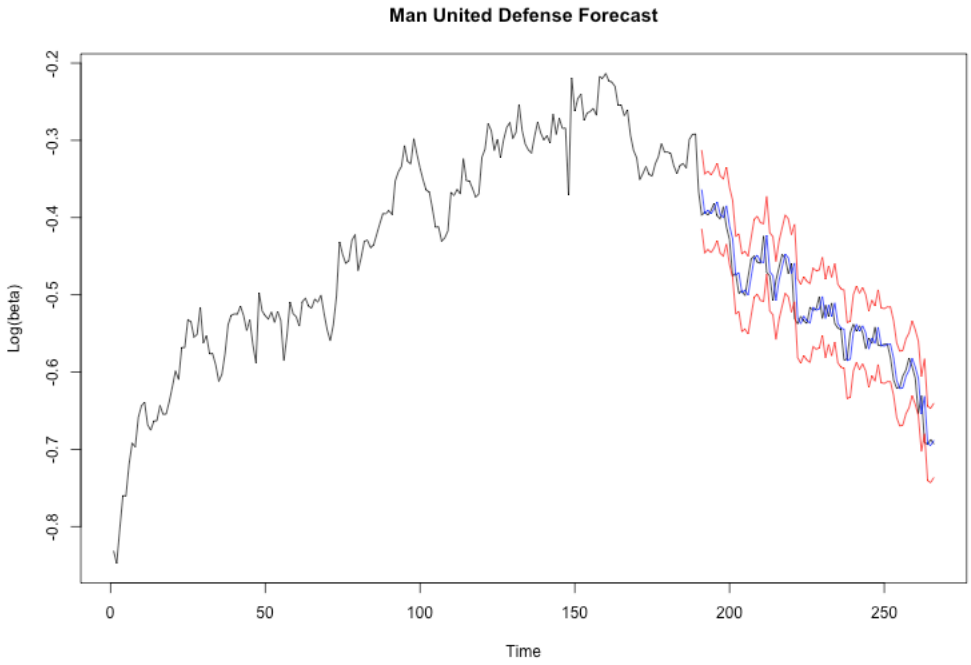


Figure 21 Continued



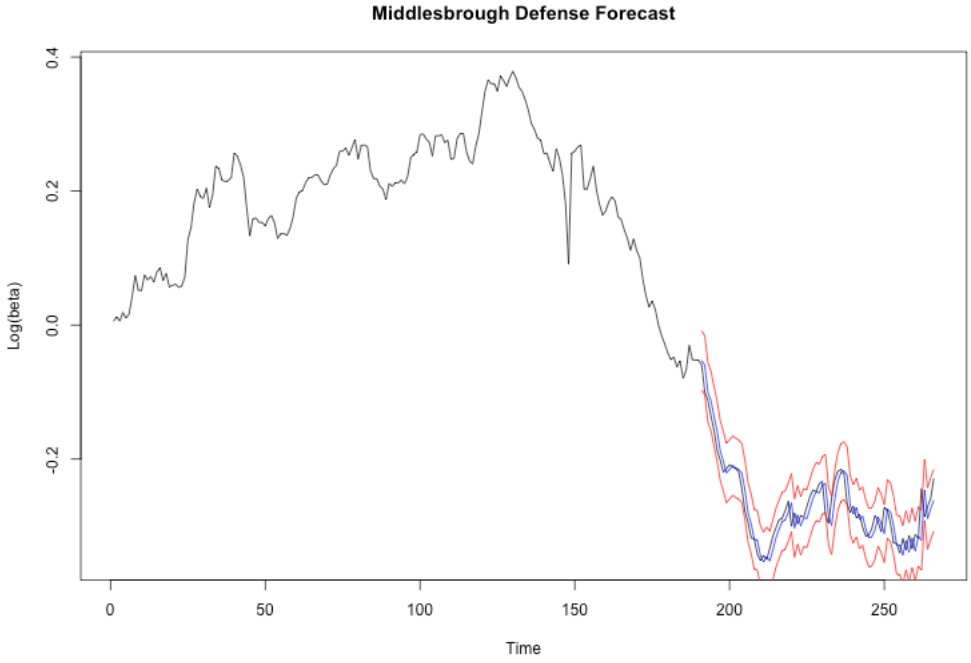
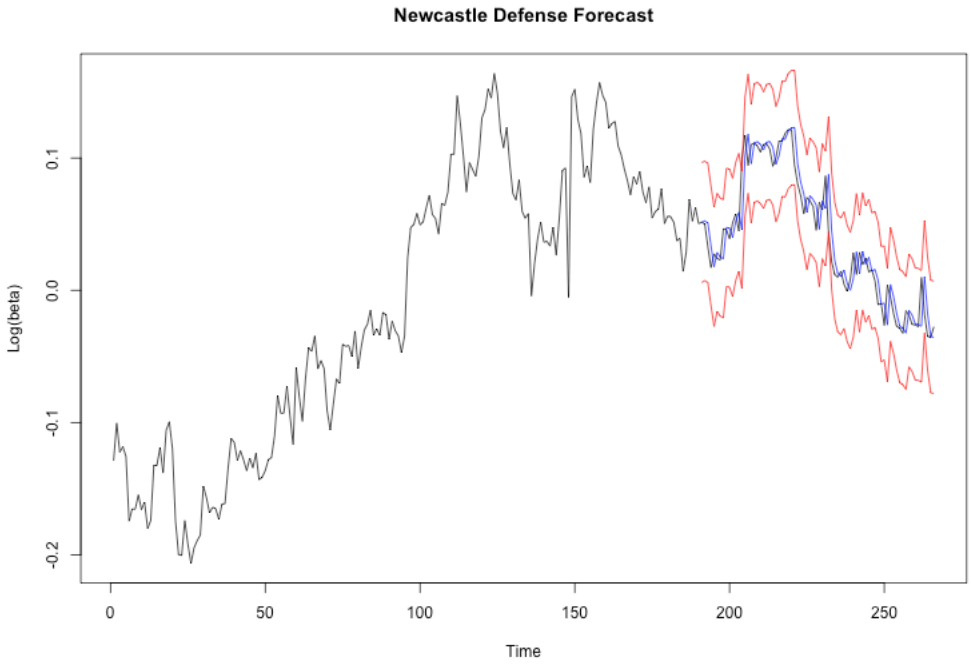


Figure 21 Continued



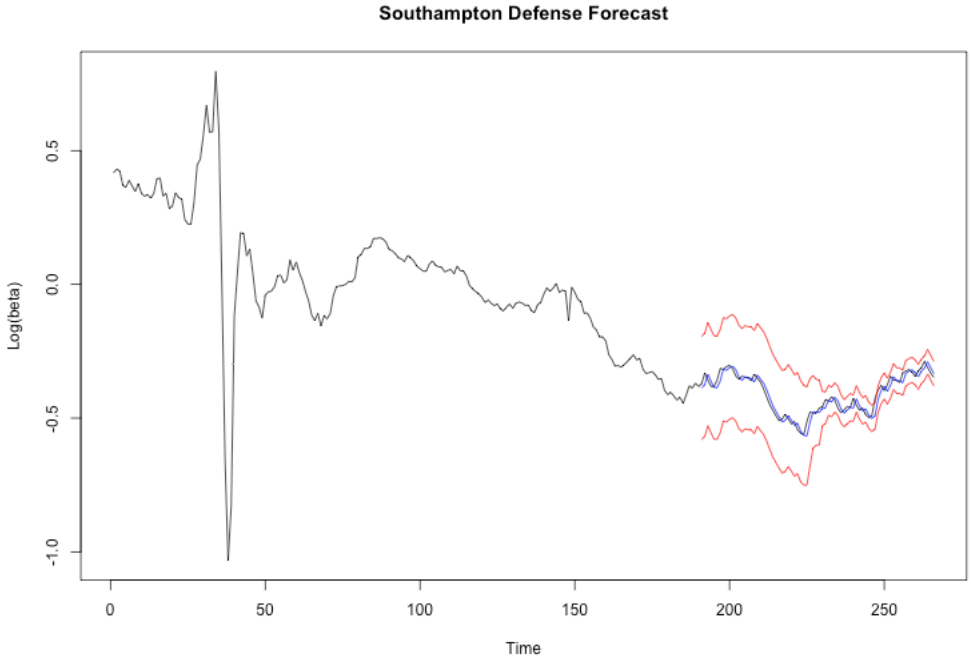
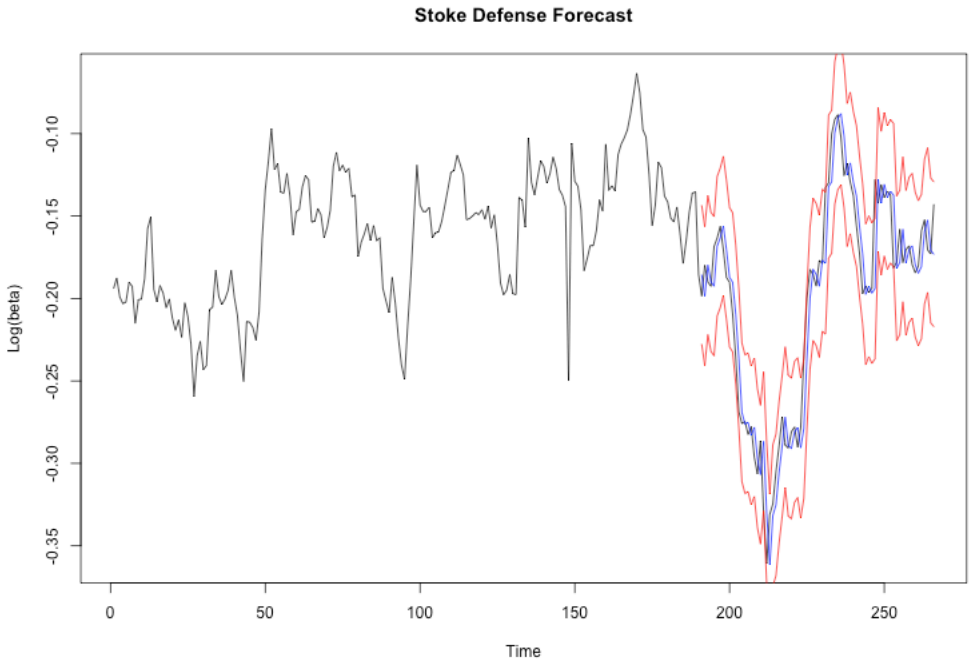


Figure 21 Continued



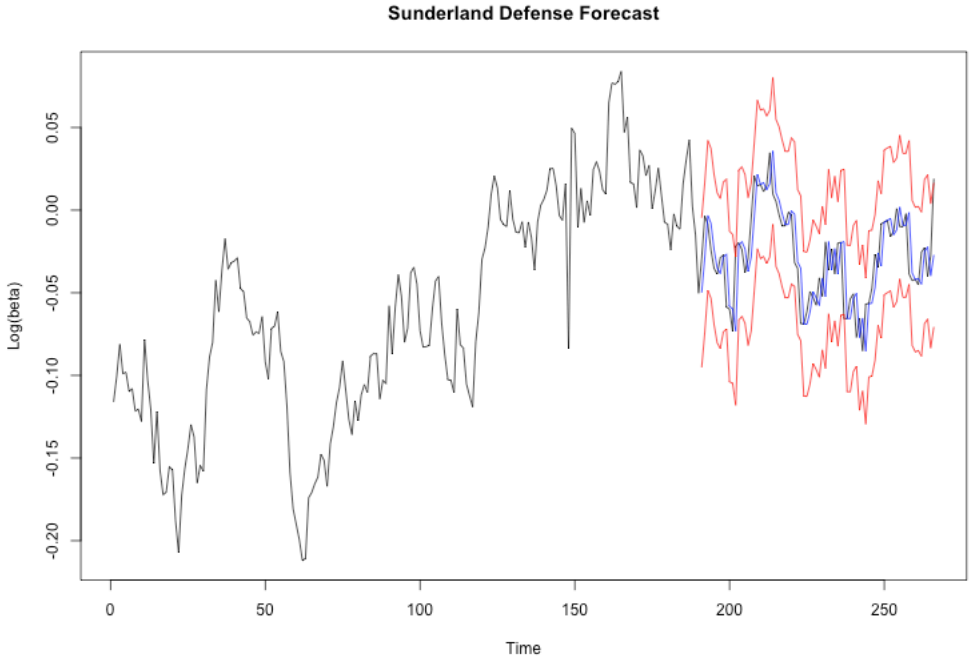
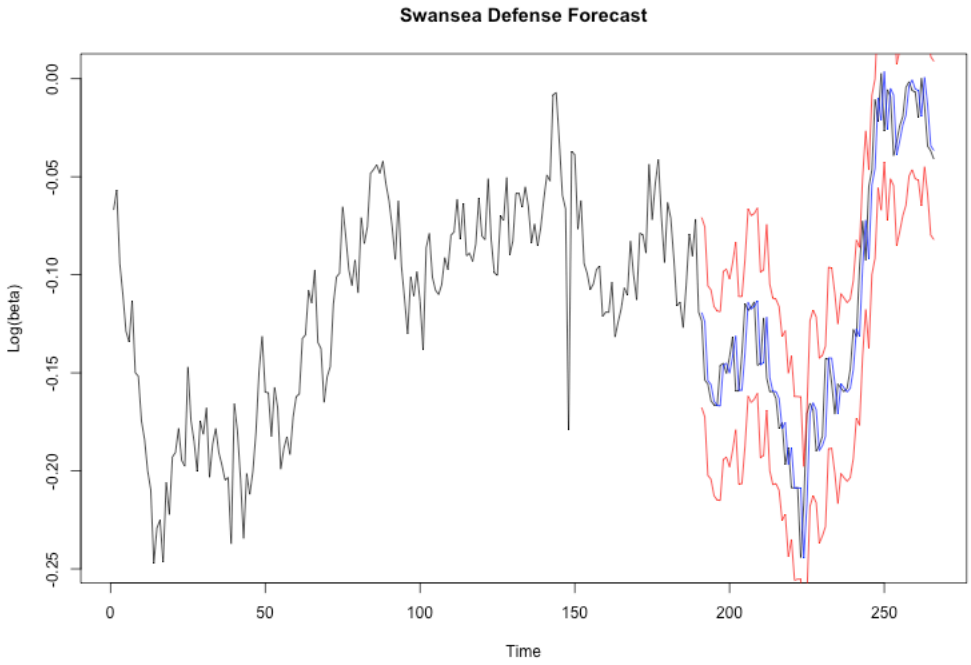


Figure 21 Continued



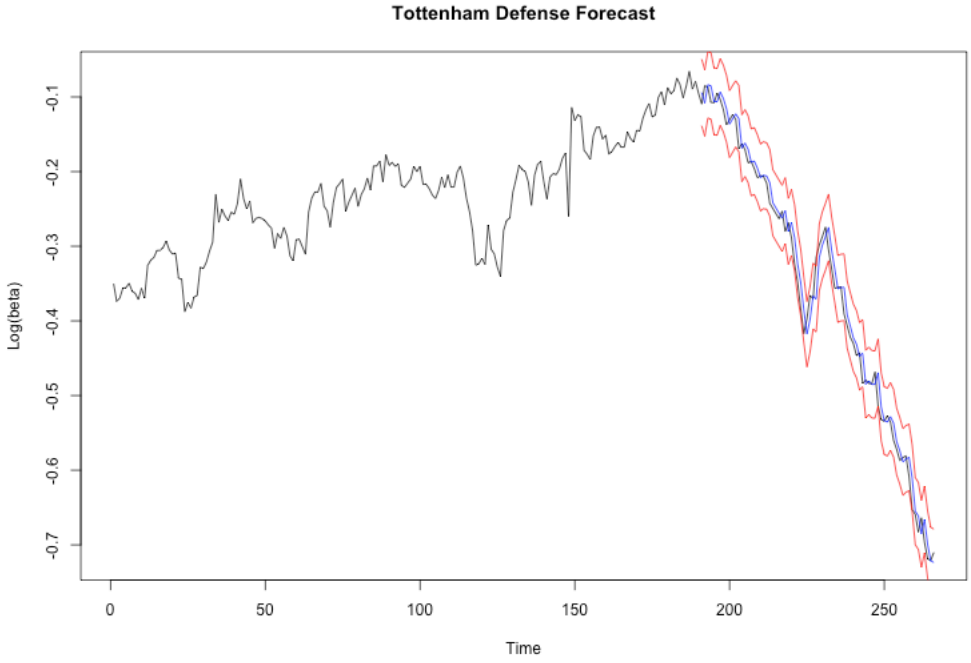
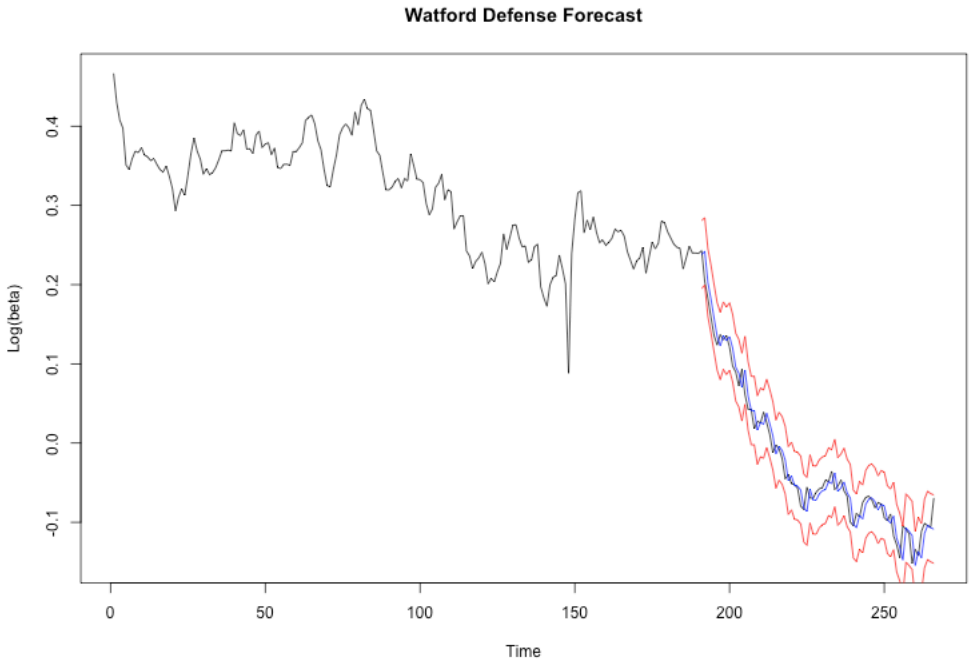


Figure 21 Continued



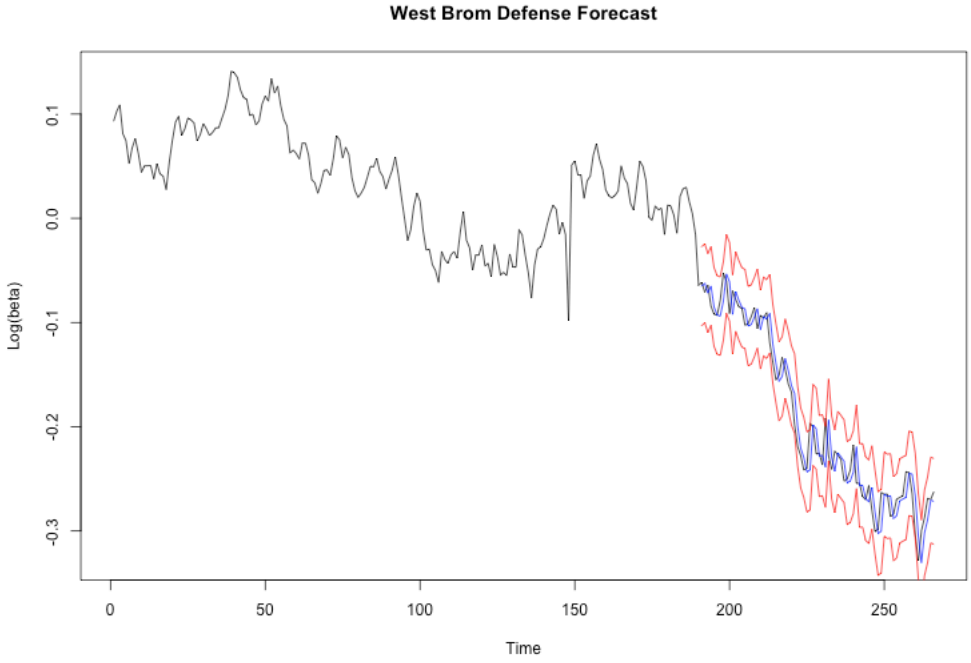


Figure 21 Continued

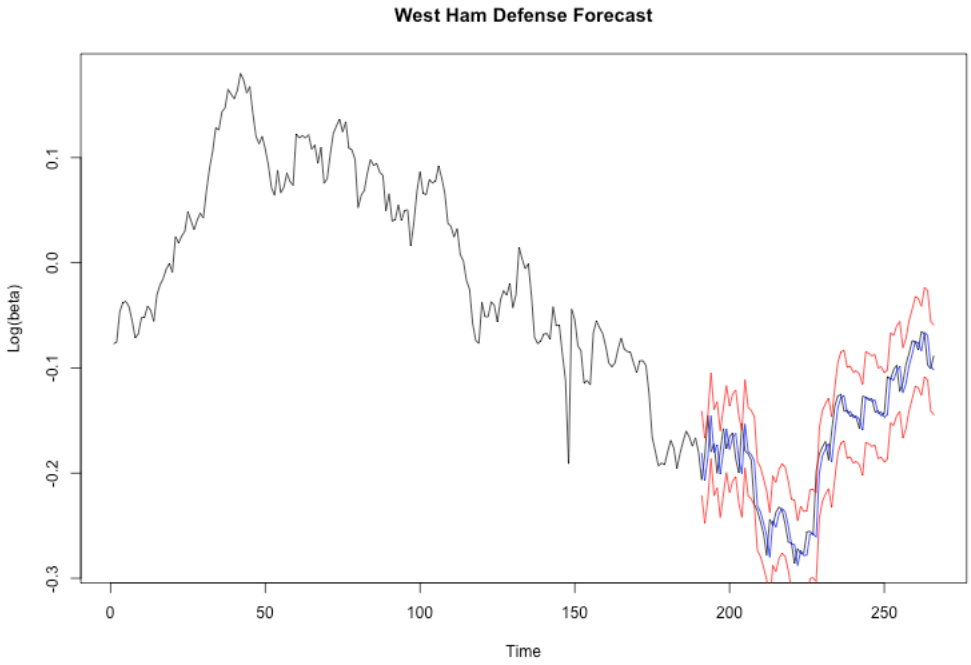
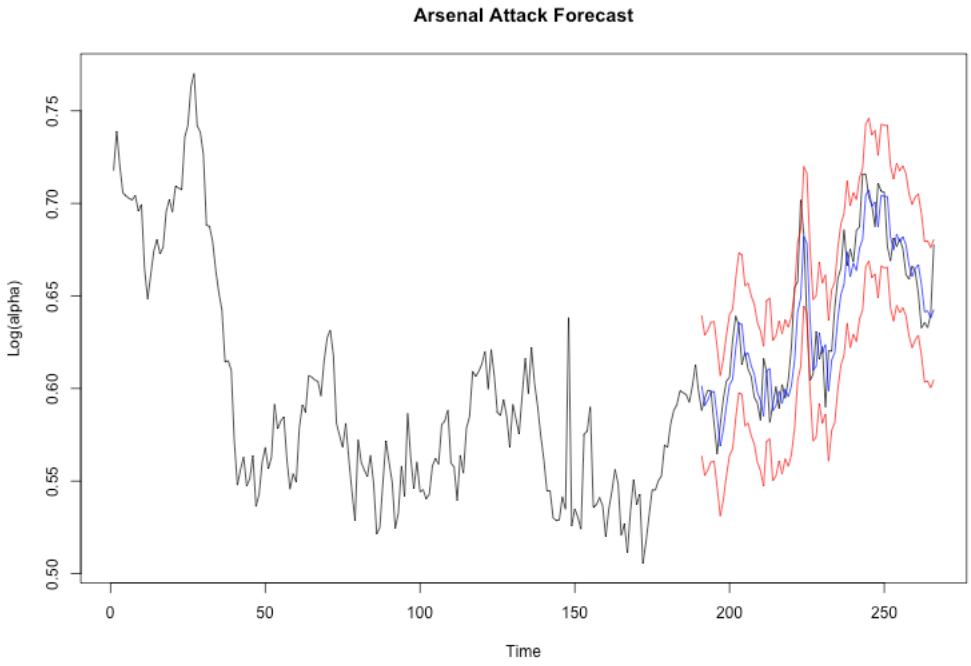


Figure 21 Continued

Appendix C – Forecasts from the auto-arima model



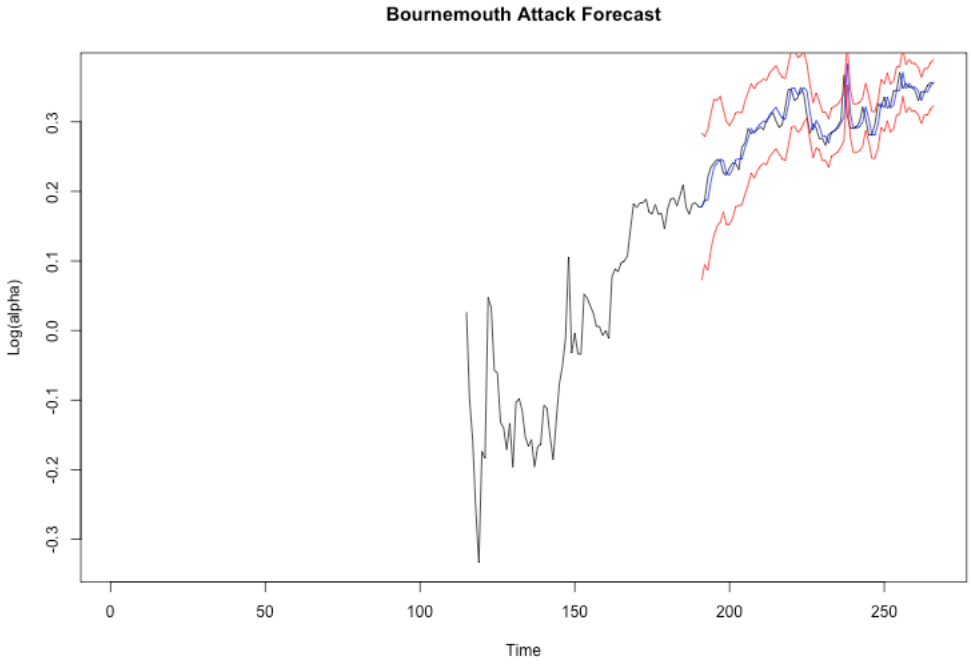
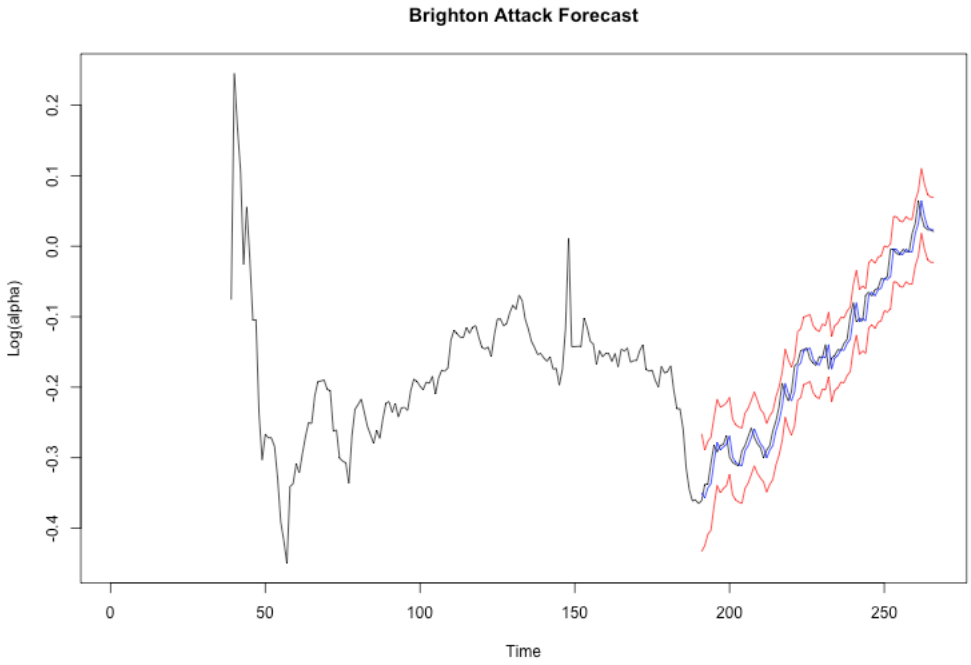


Figure 22 - (Model 3) Auto-Arima Forecasts – Attack



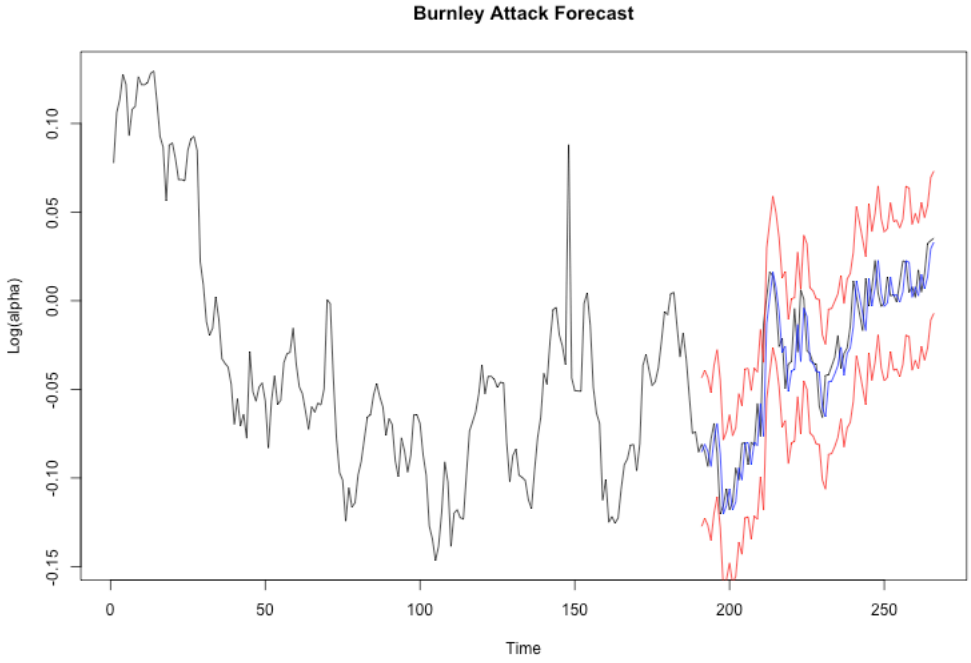
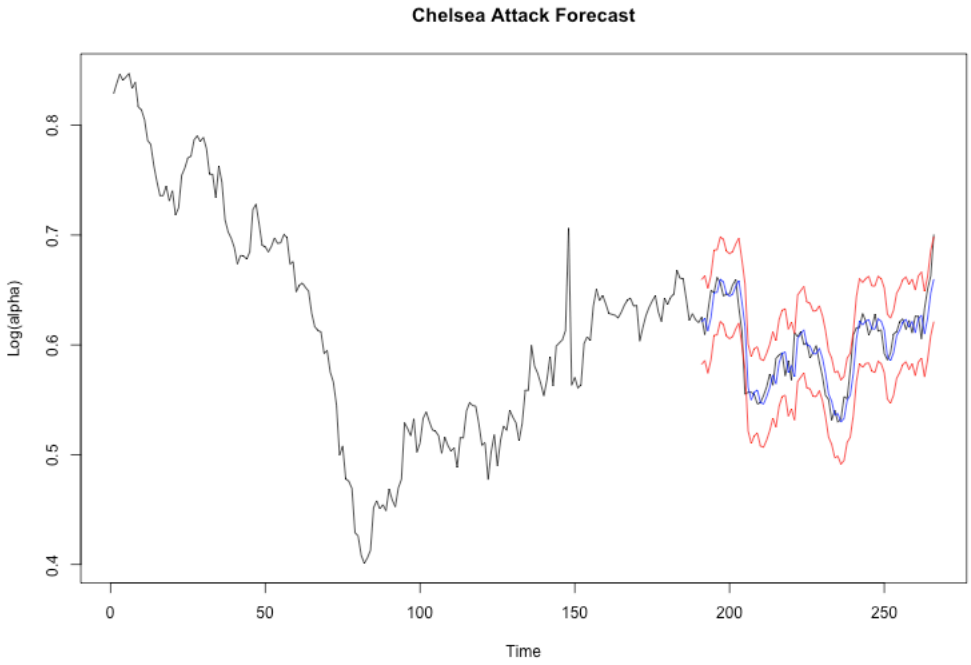


Figure 22 Continued



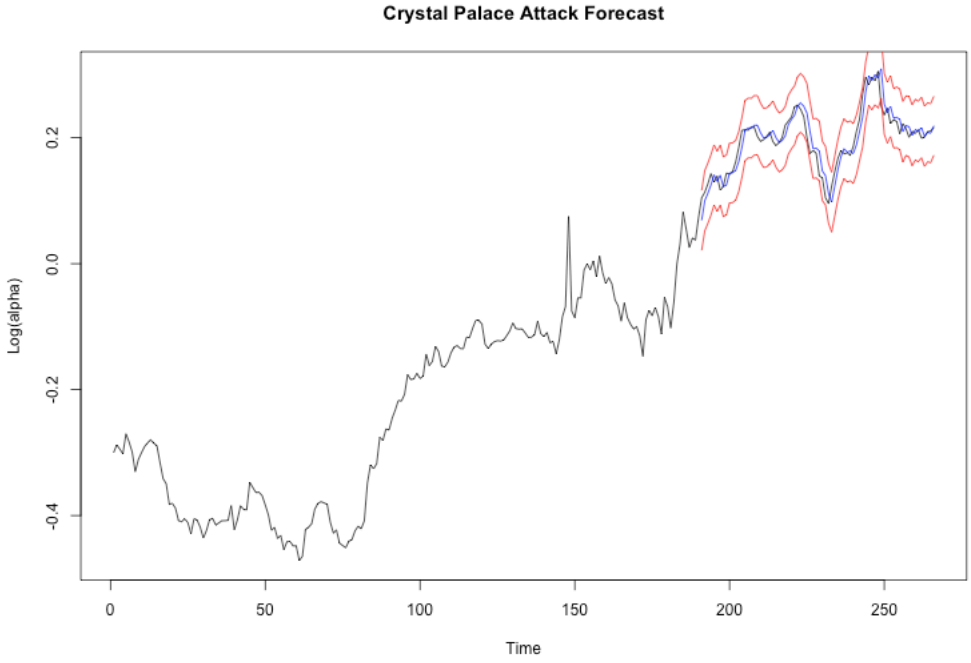
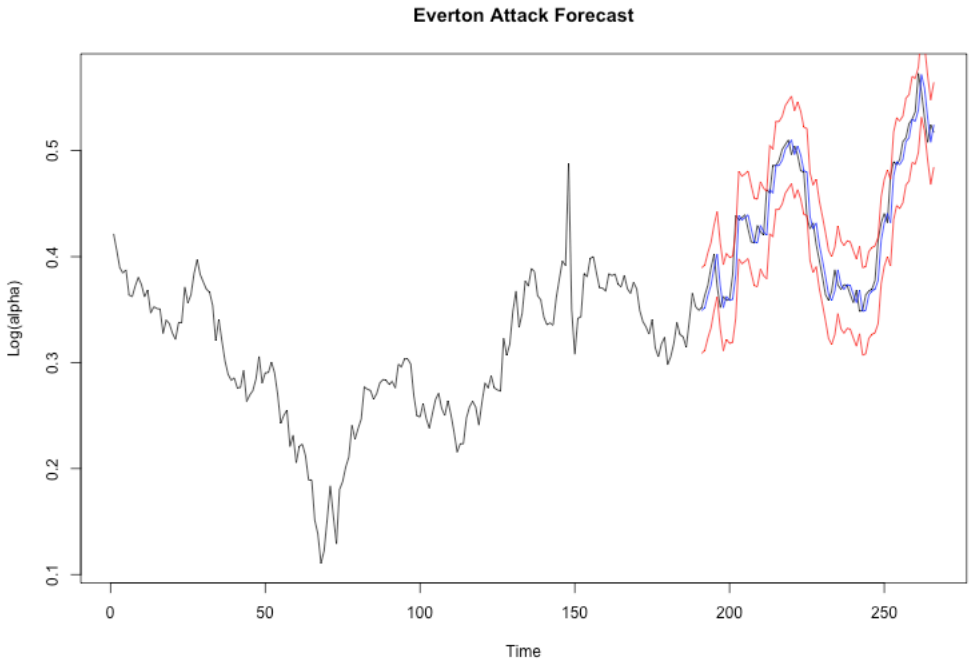


Figure 22 Continued



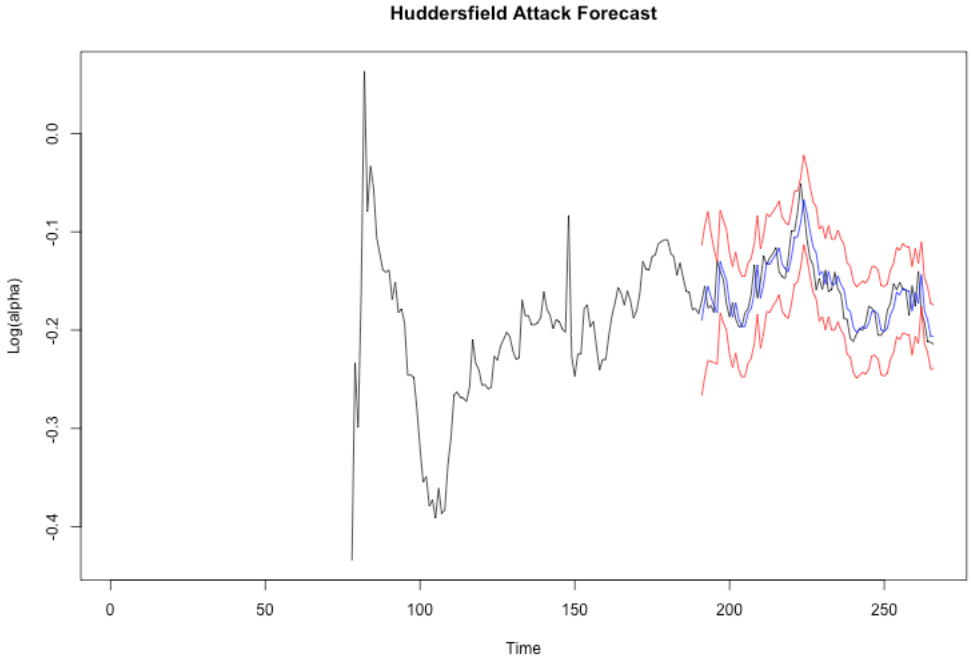
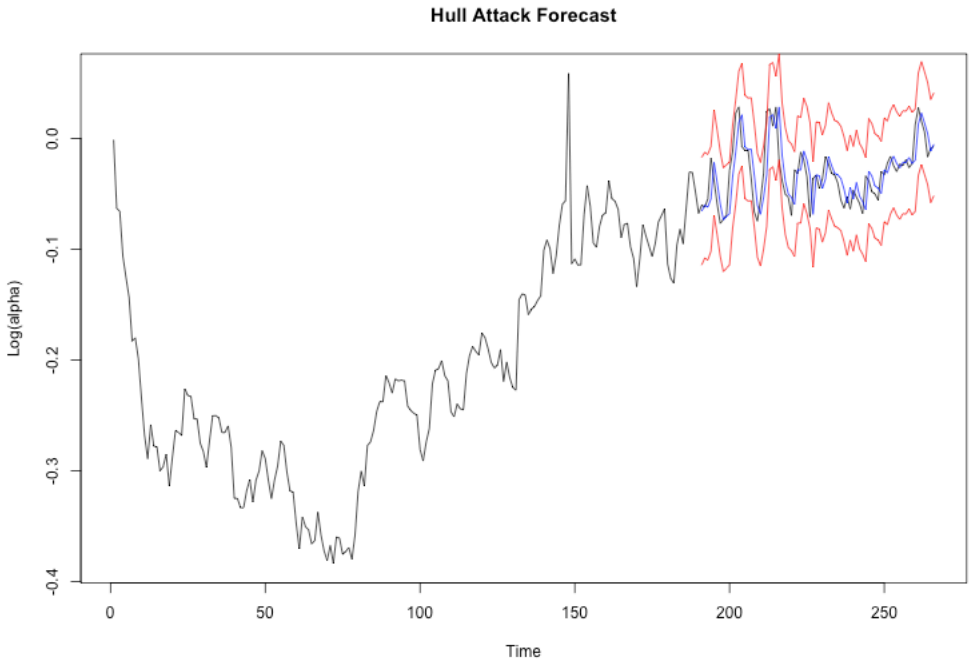


Figure 22 Continued



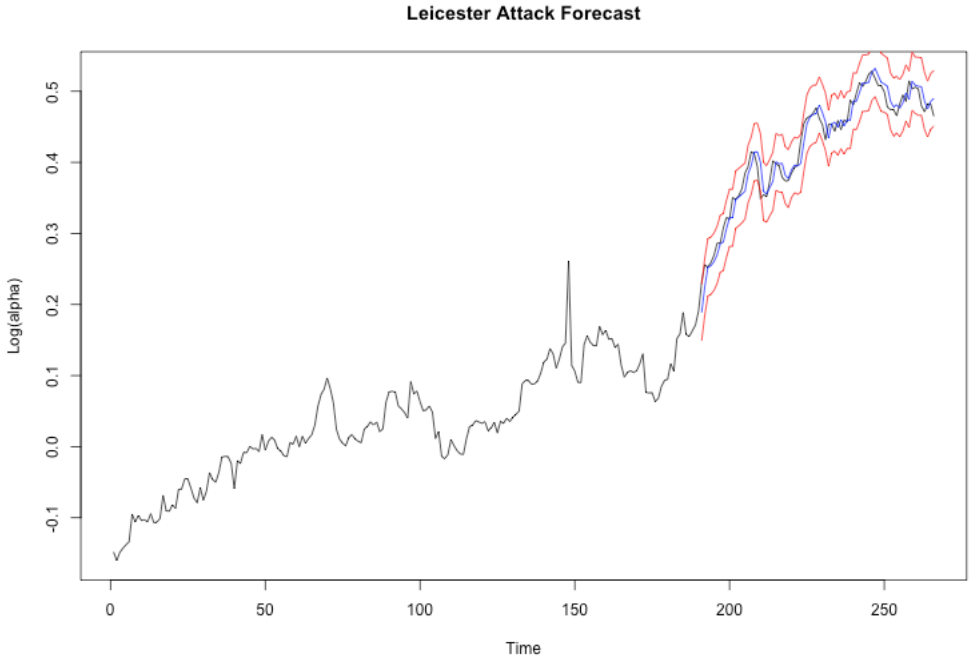
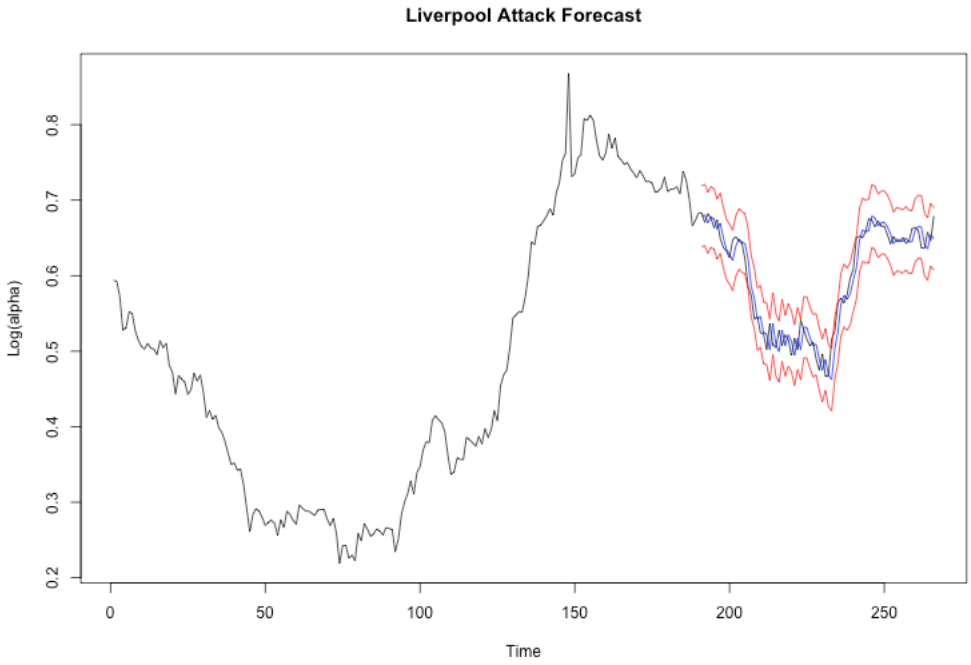


Figure 22 Continued



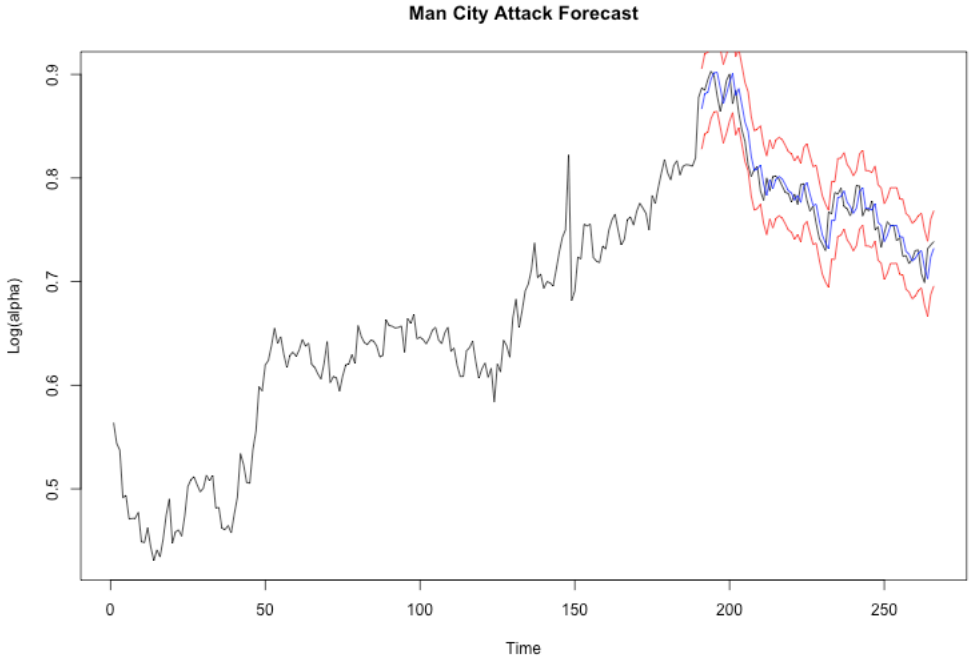
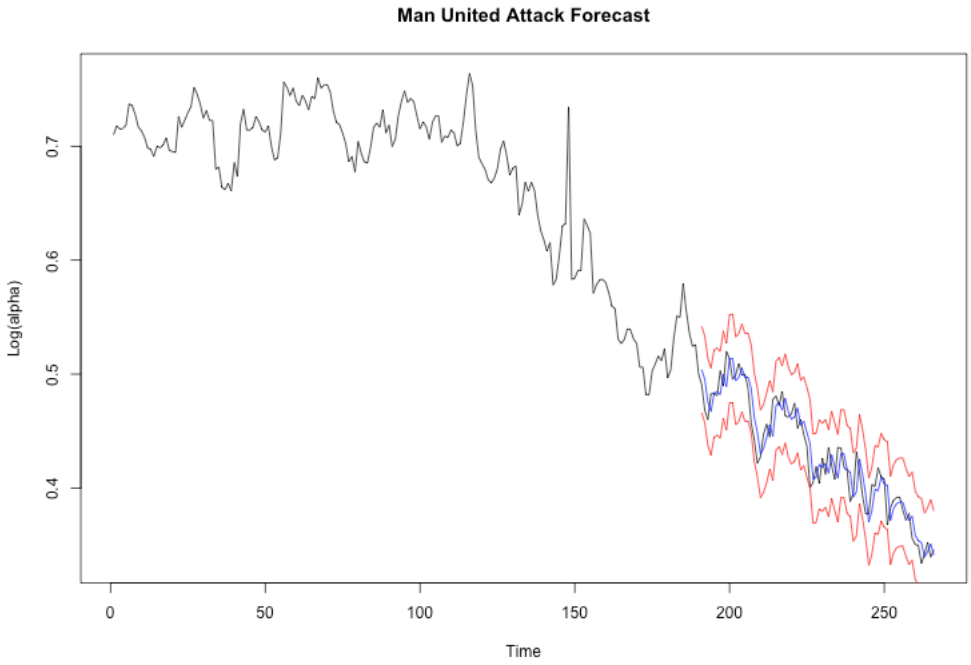


Figure 22 Continued



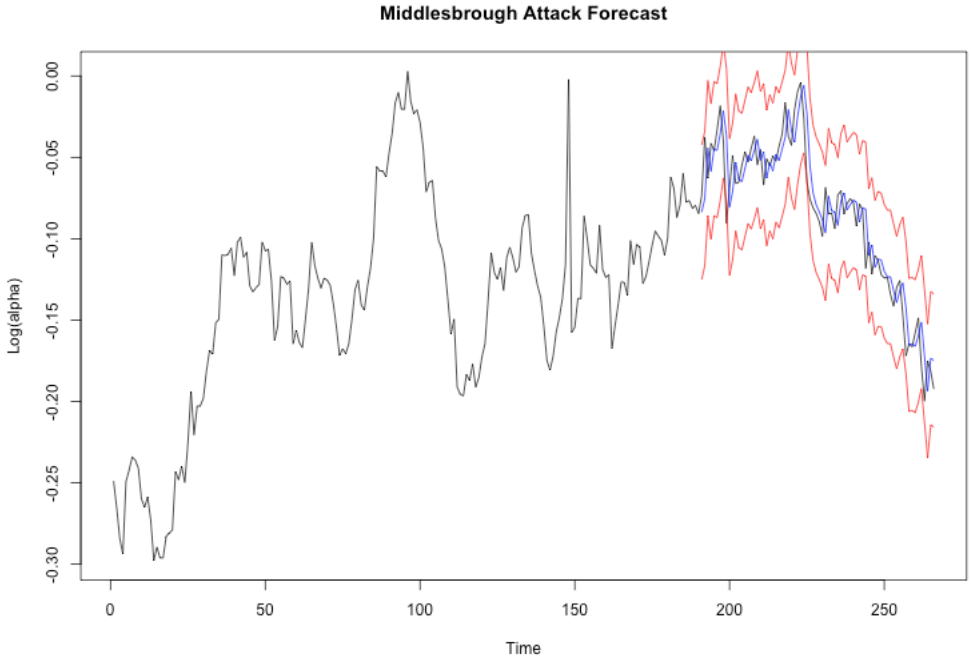
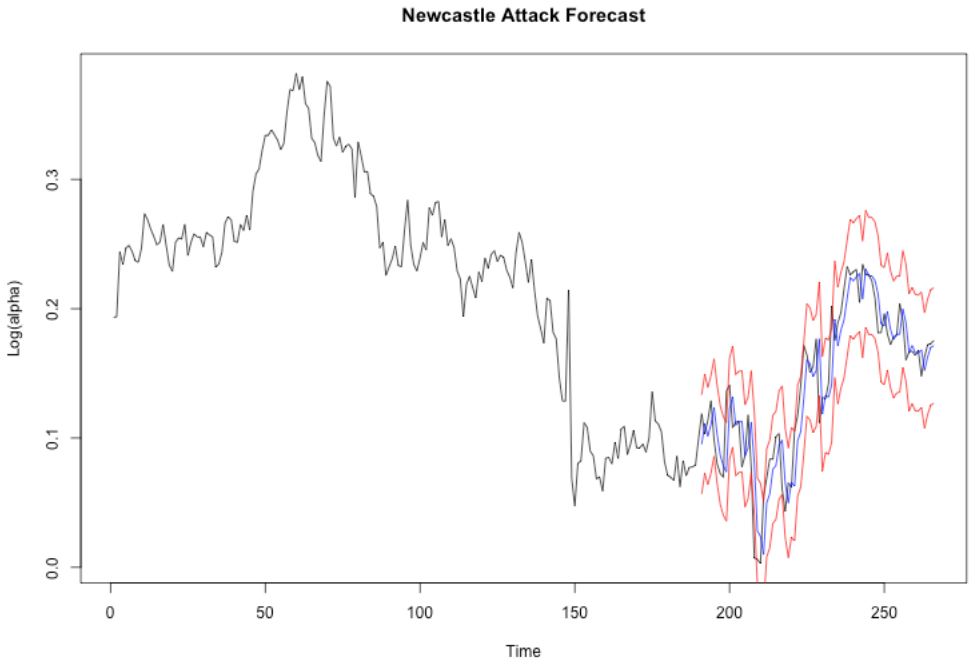


Figure 22 Continued



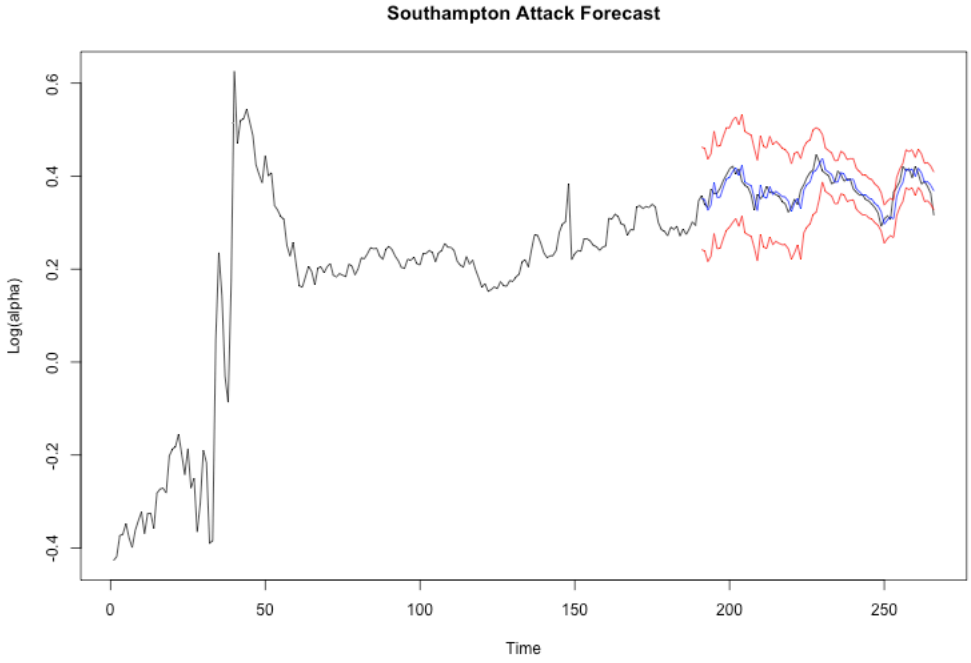
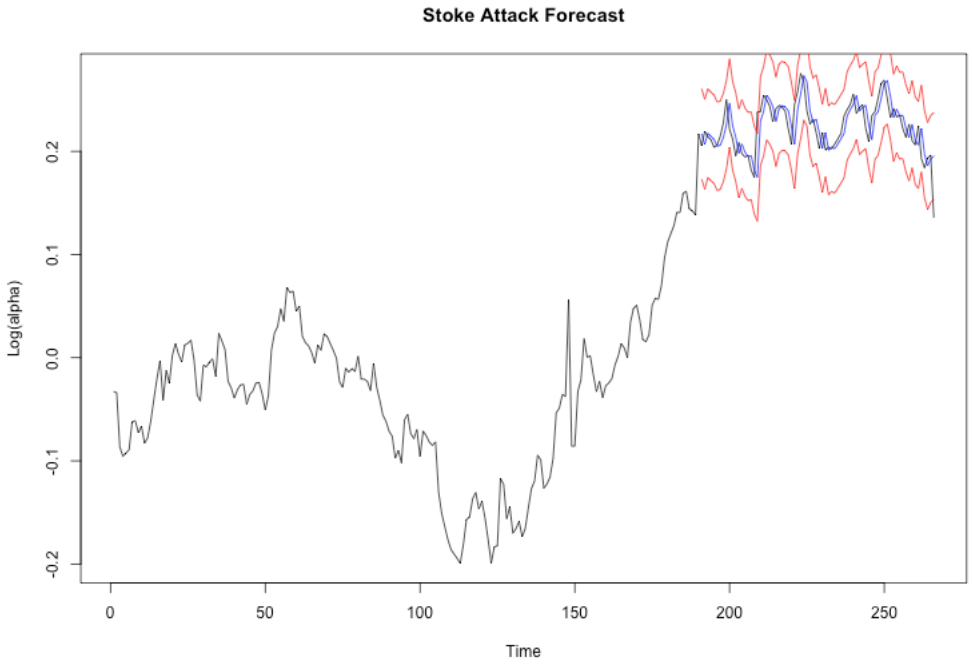


Figure 22 Continued



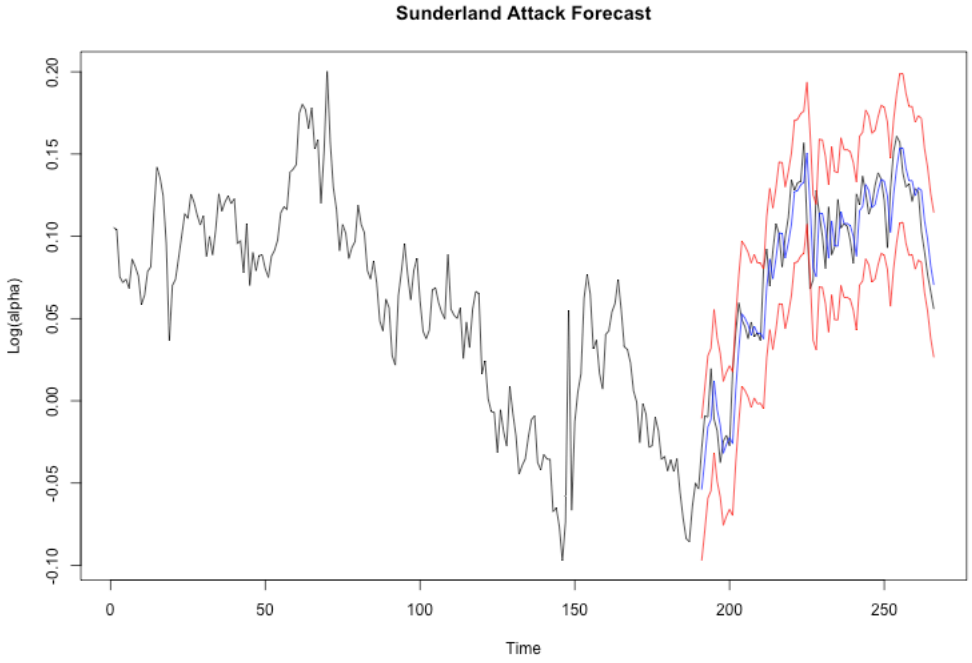
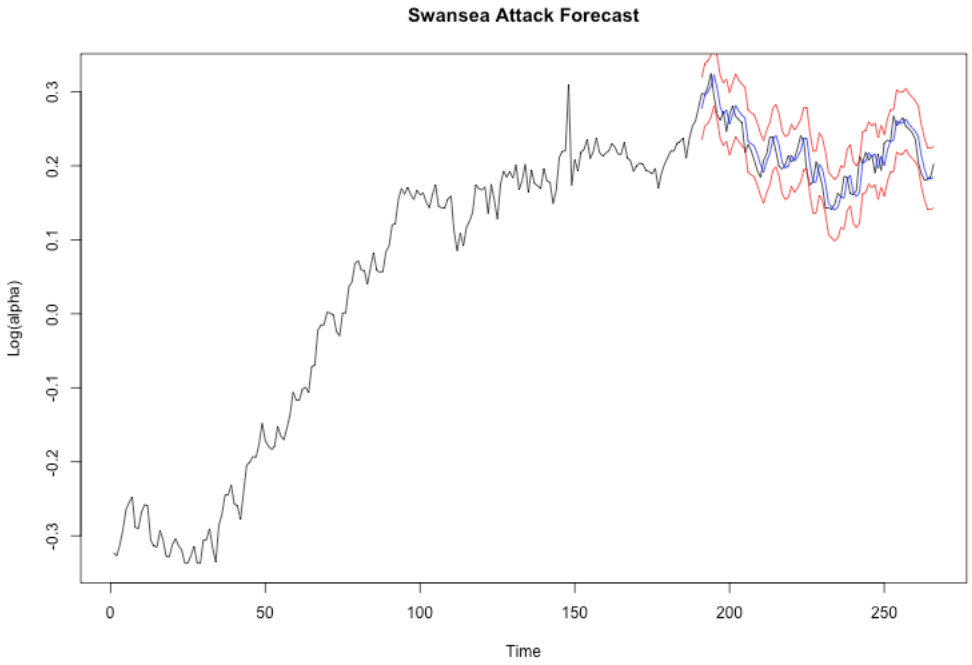


Figure 22 Continued



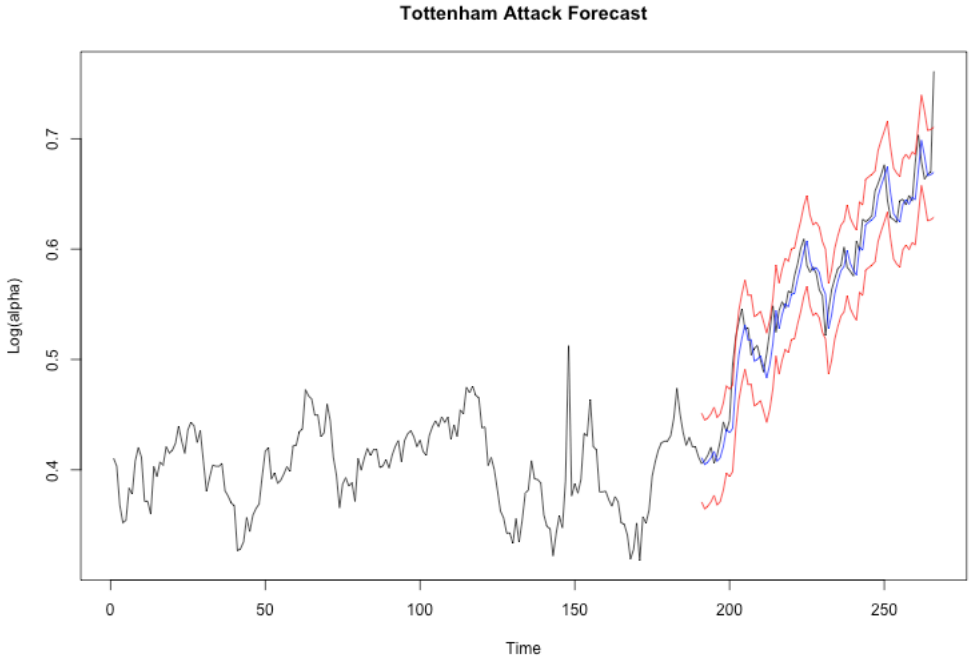
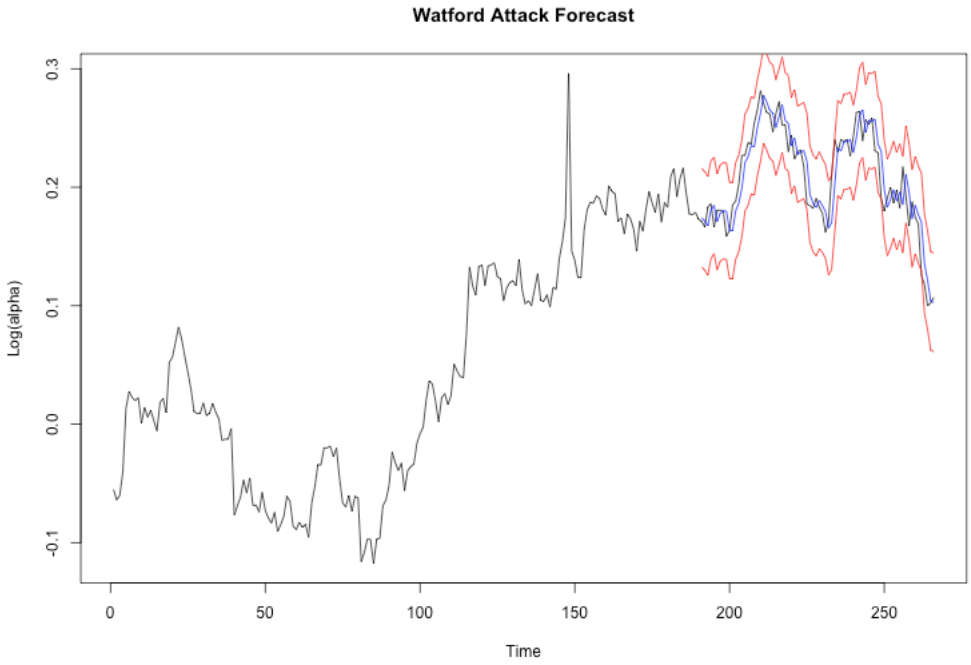


Figure 22 Continued



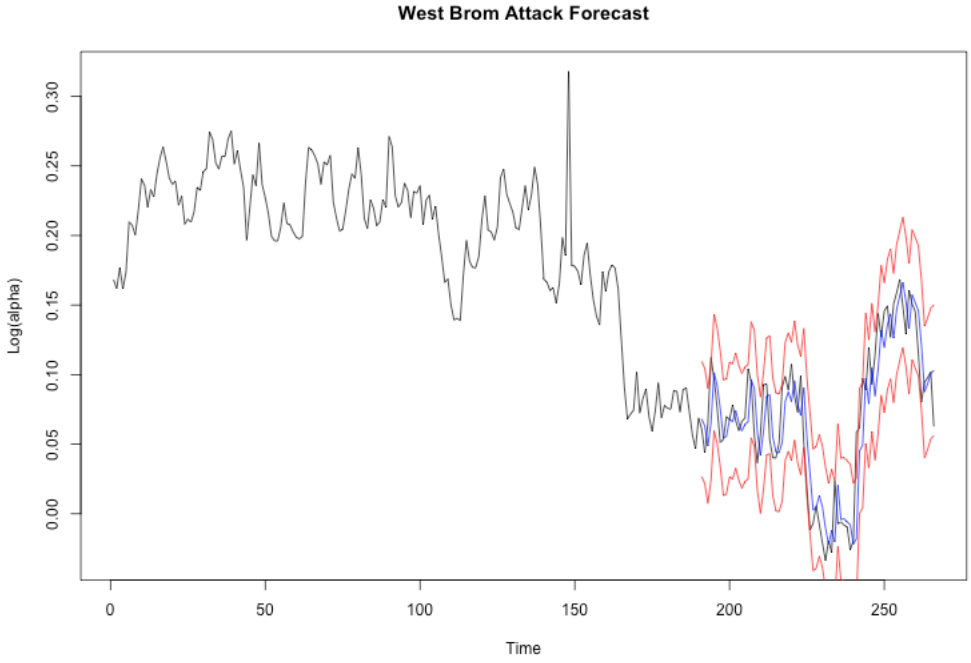


Figure 22 Continued

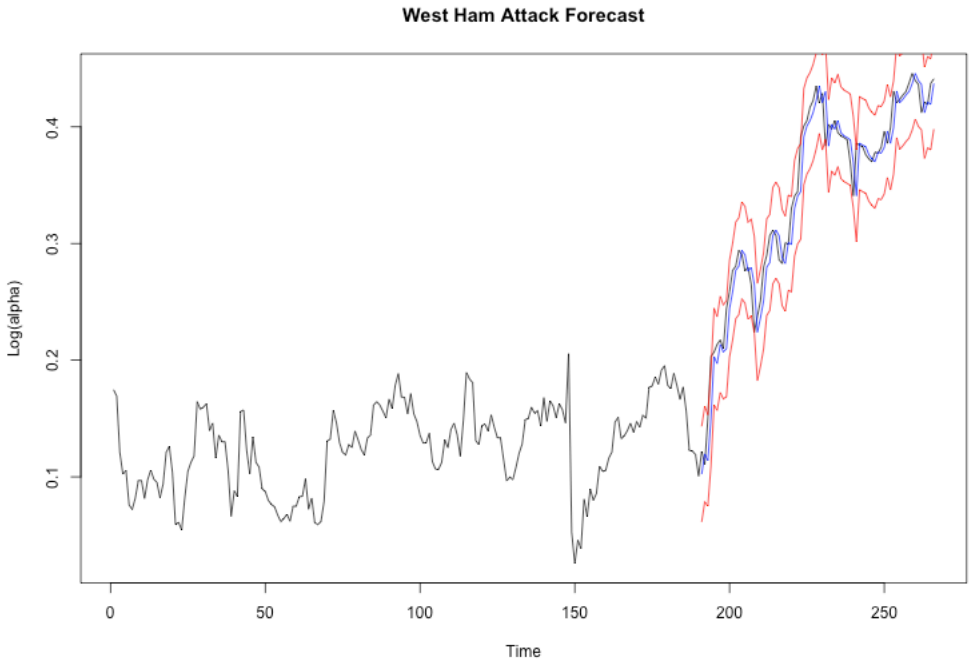
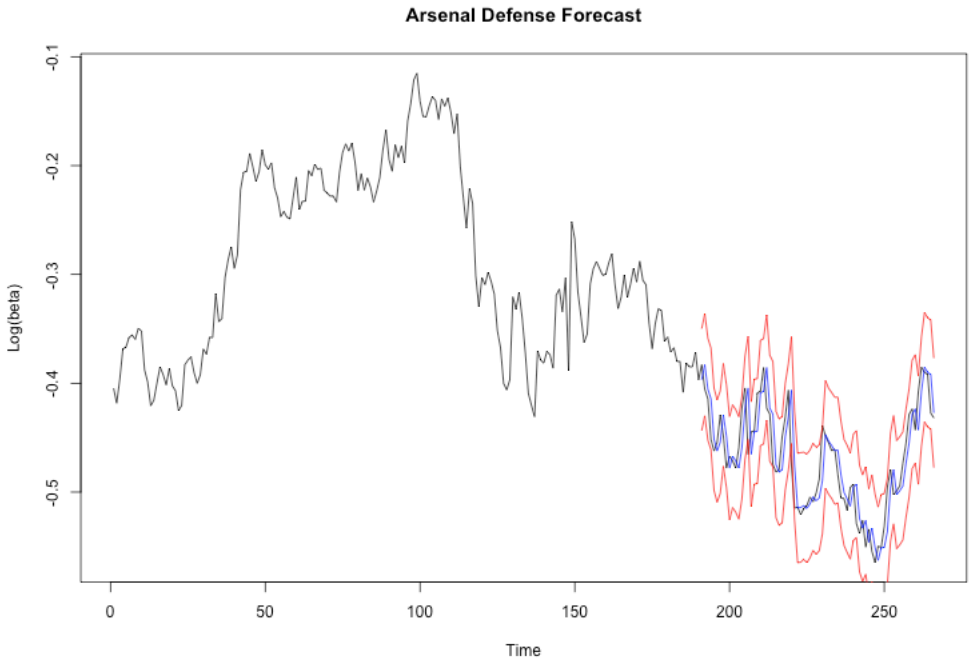


Figure 22 Continued



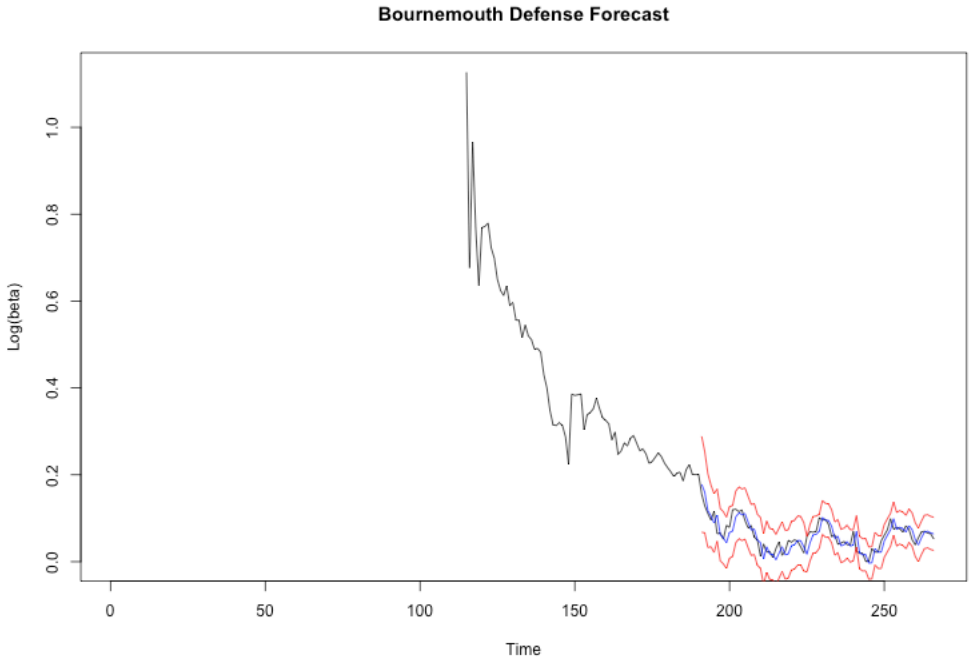
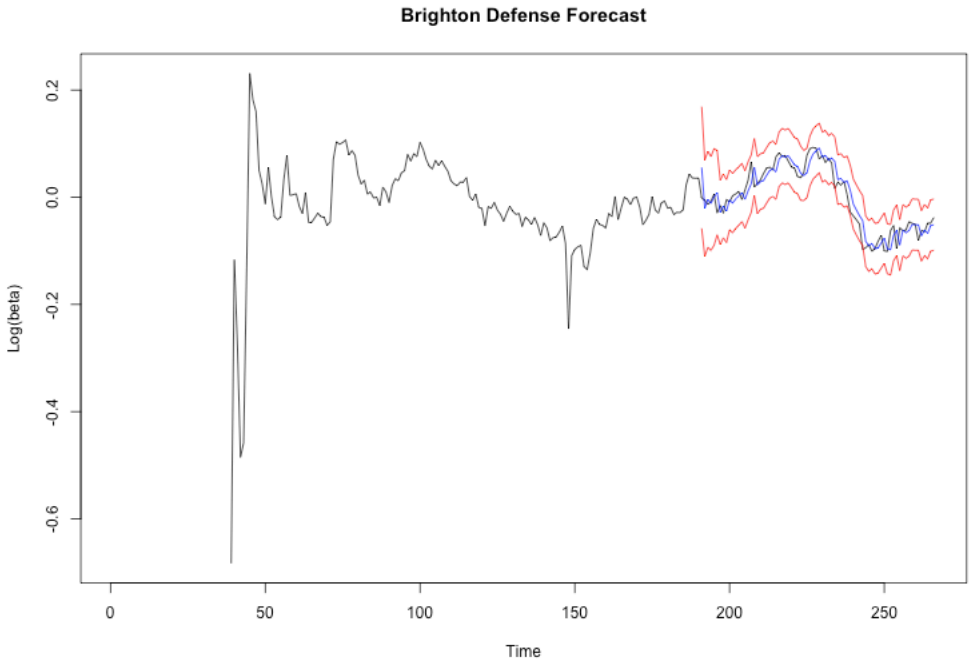


Figure 23 - (Model 3) Auto-Arima Forecasts – Defence



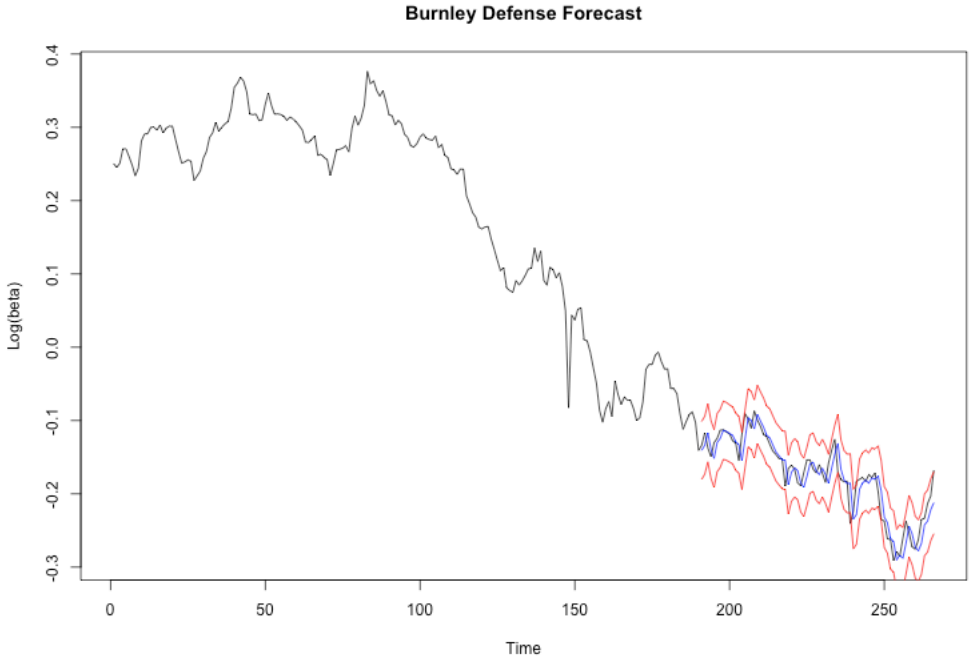
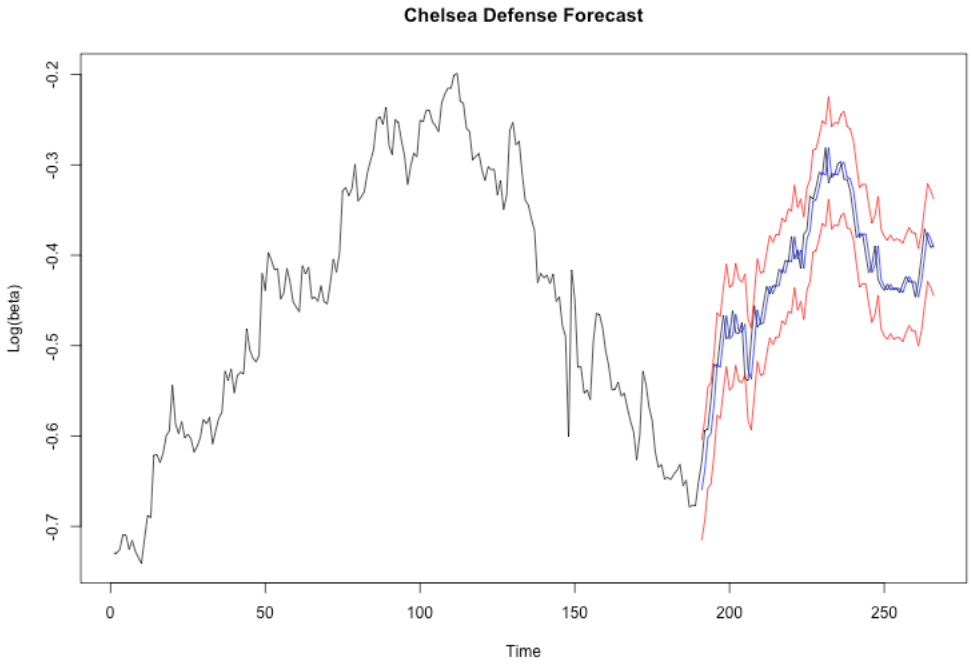


Figure 23 Continued



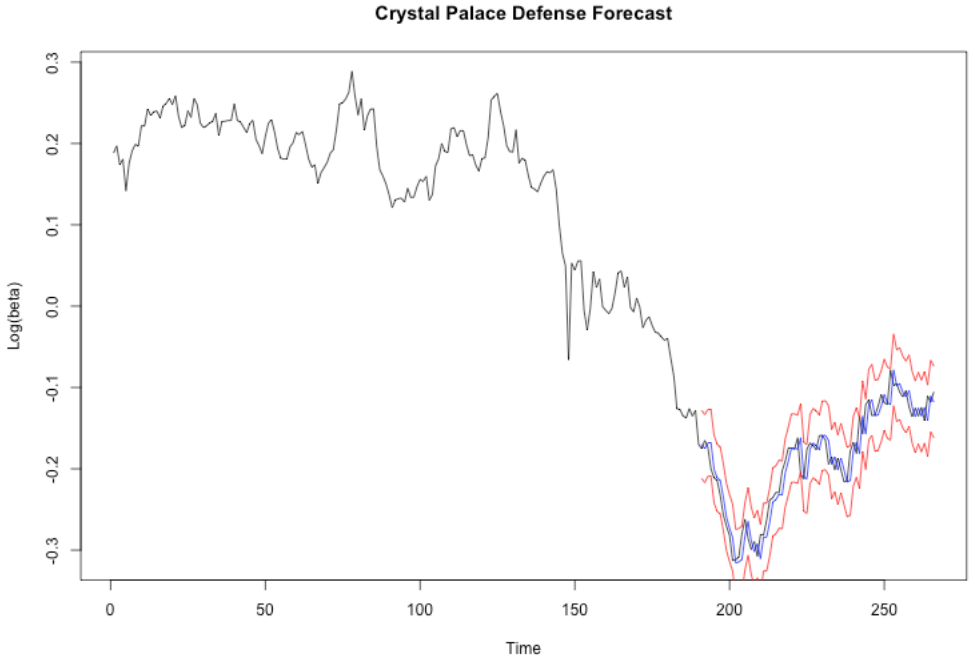
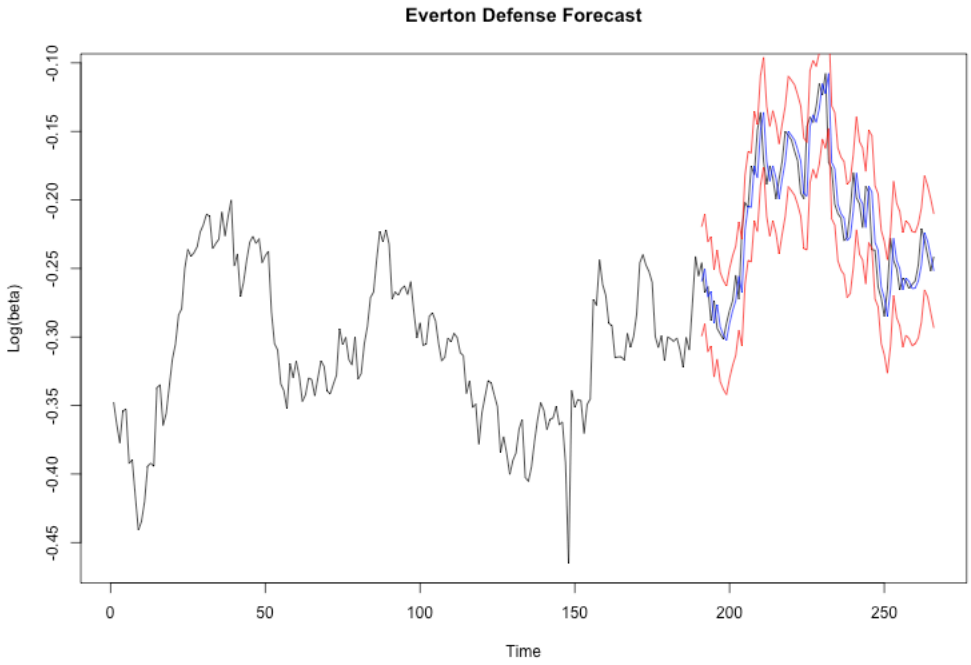


Figure 23 Continued



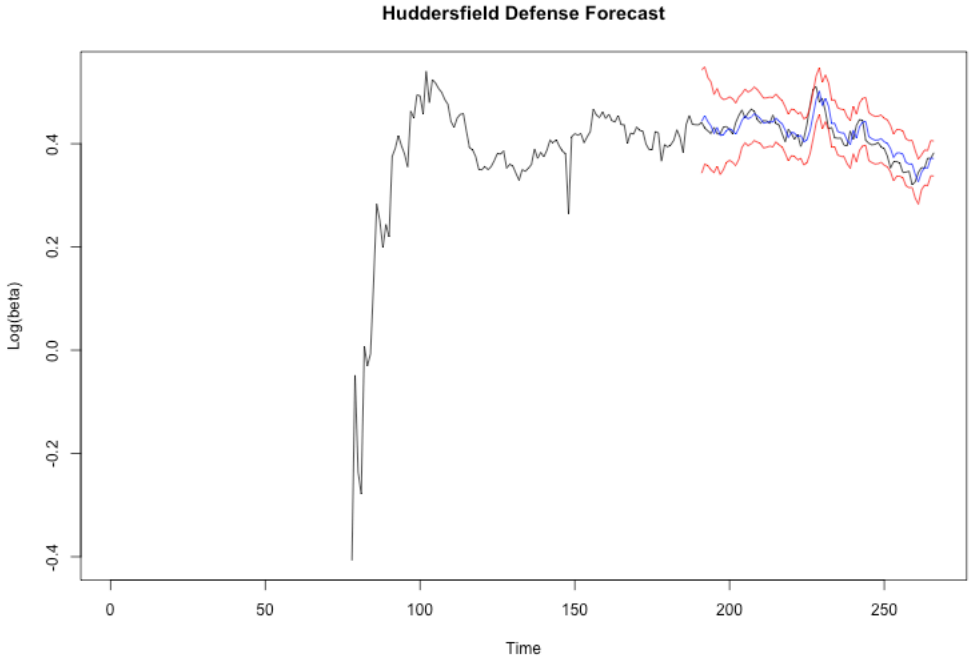
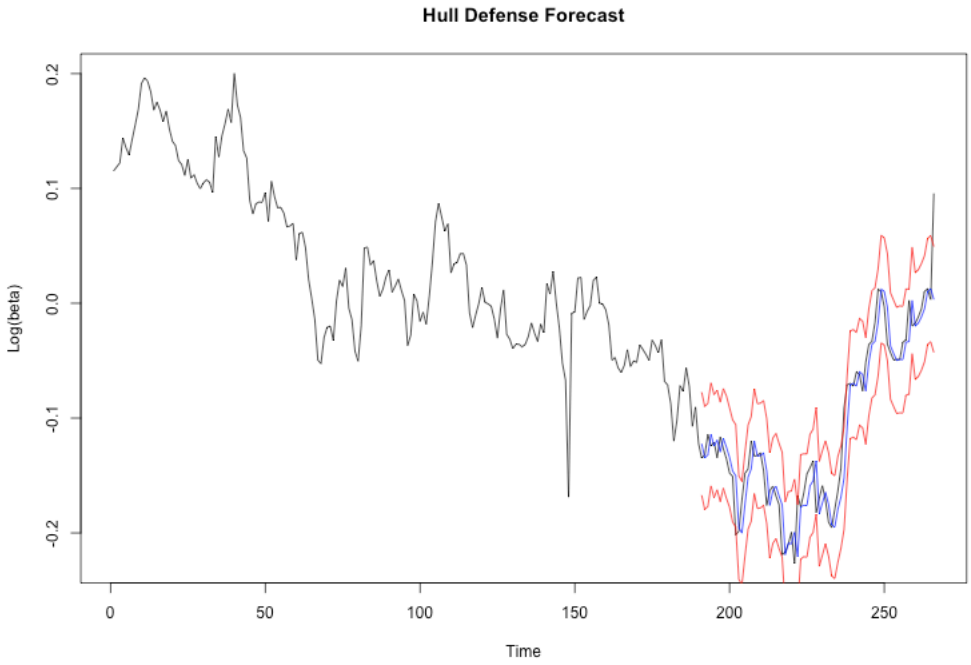


Figure 23 Continued



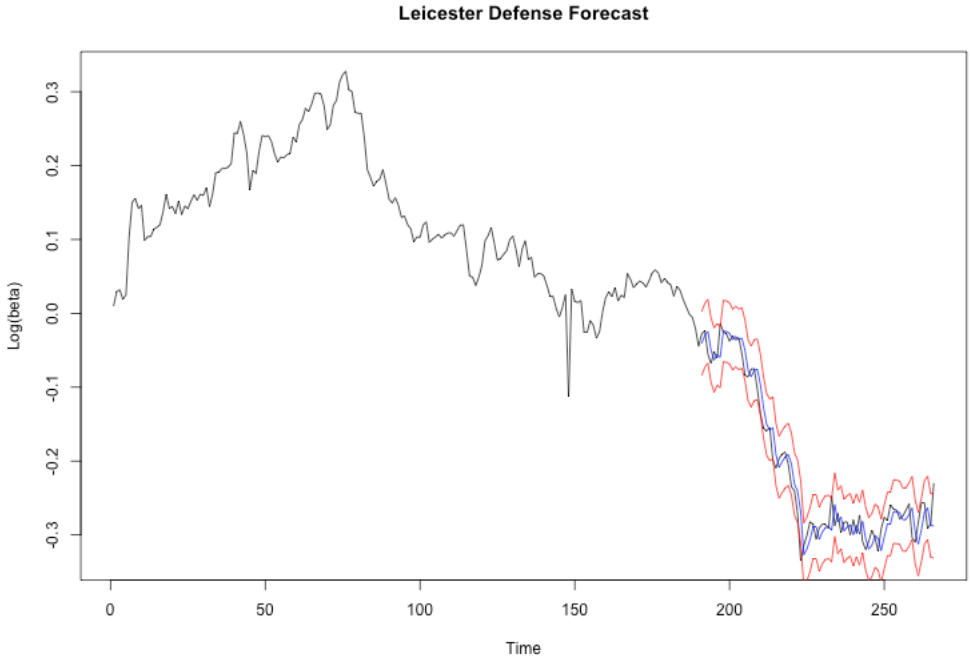
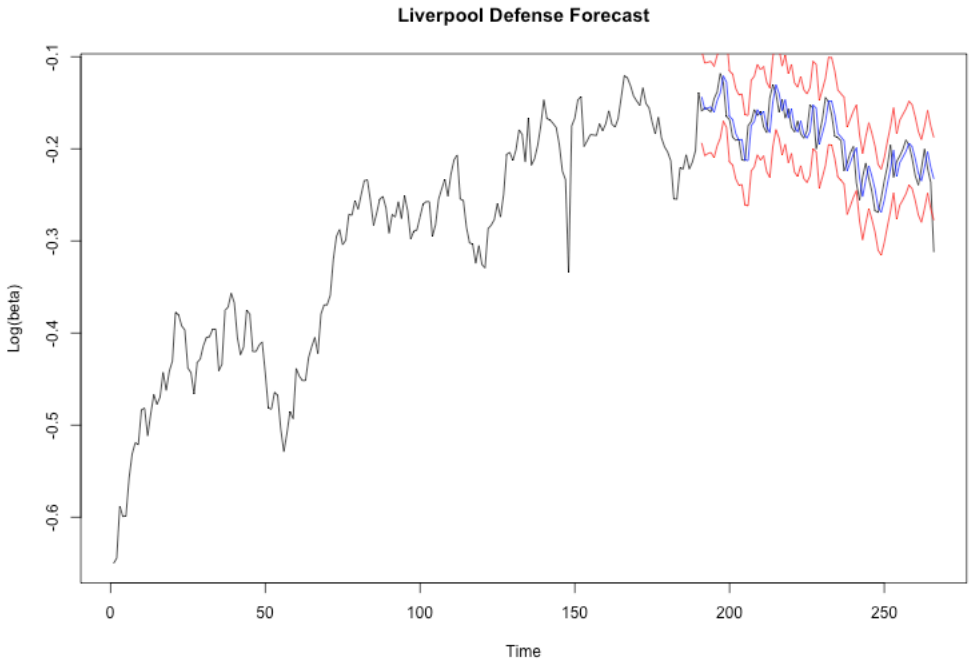


Figure 23 Continued



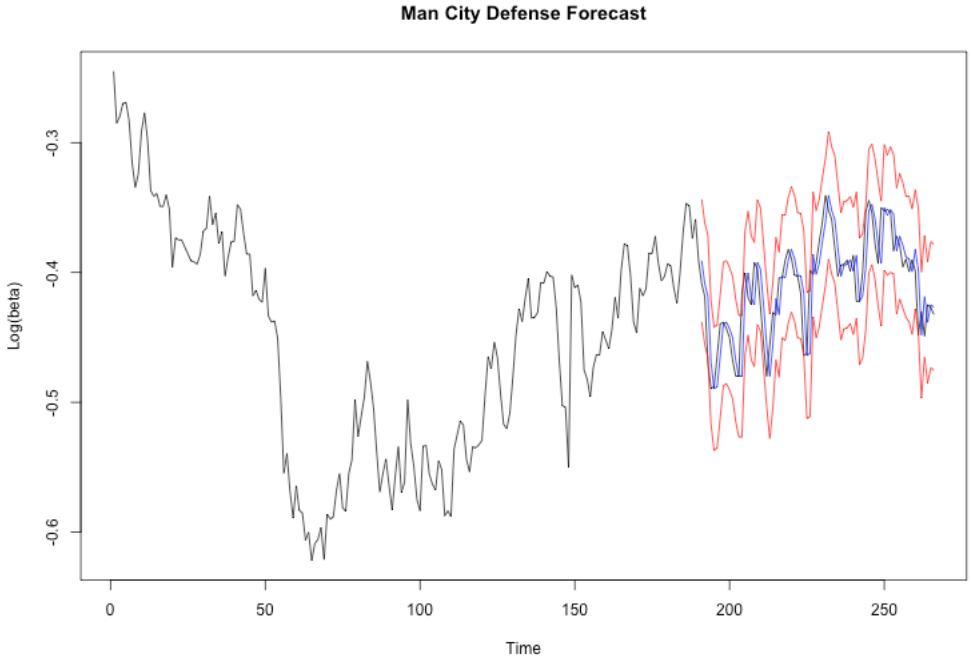
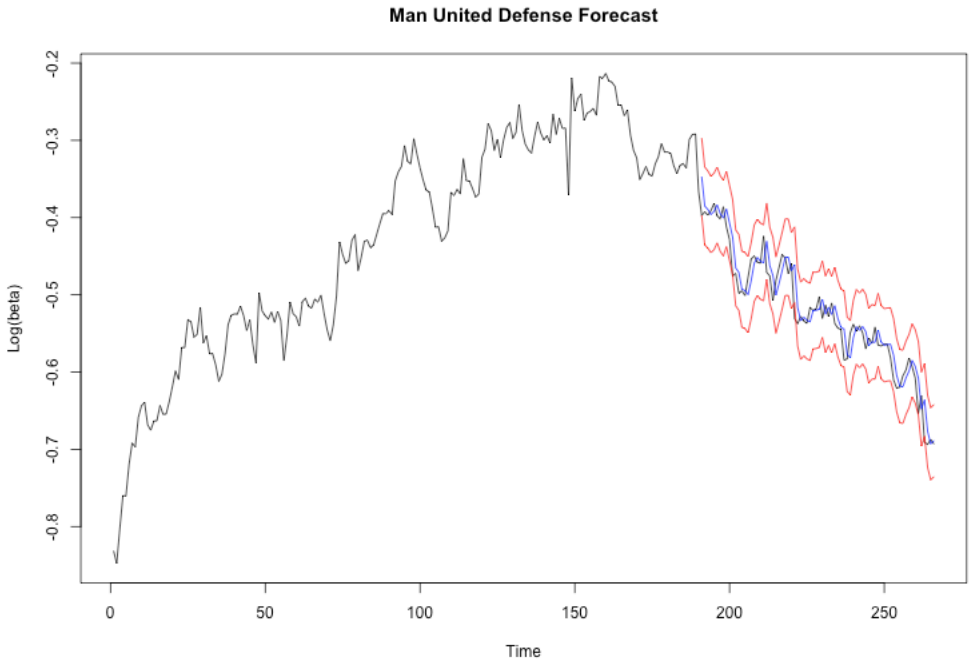


Figure 23 Continued



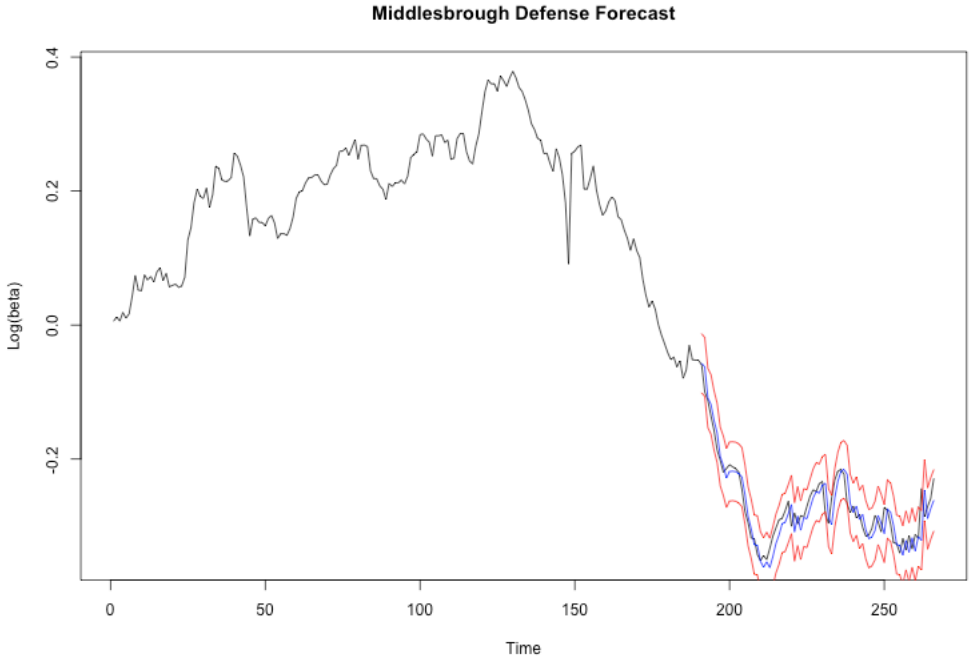
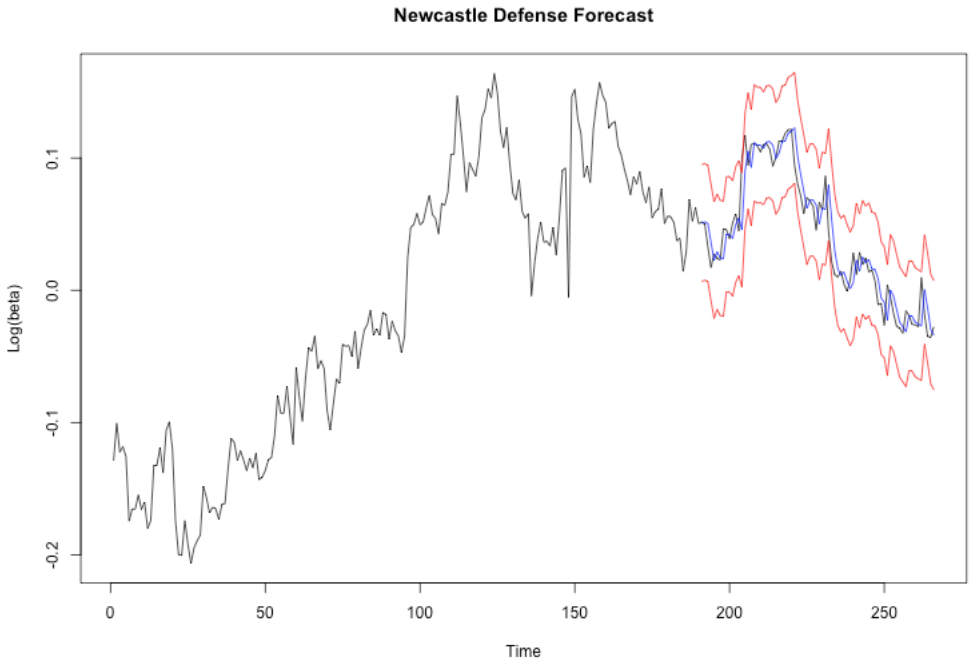


Figure 23 Continued



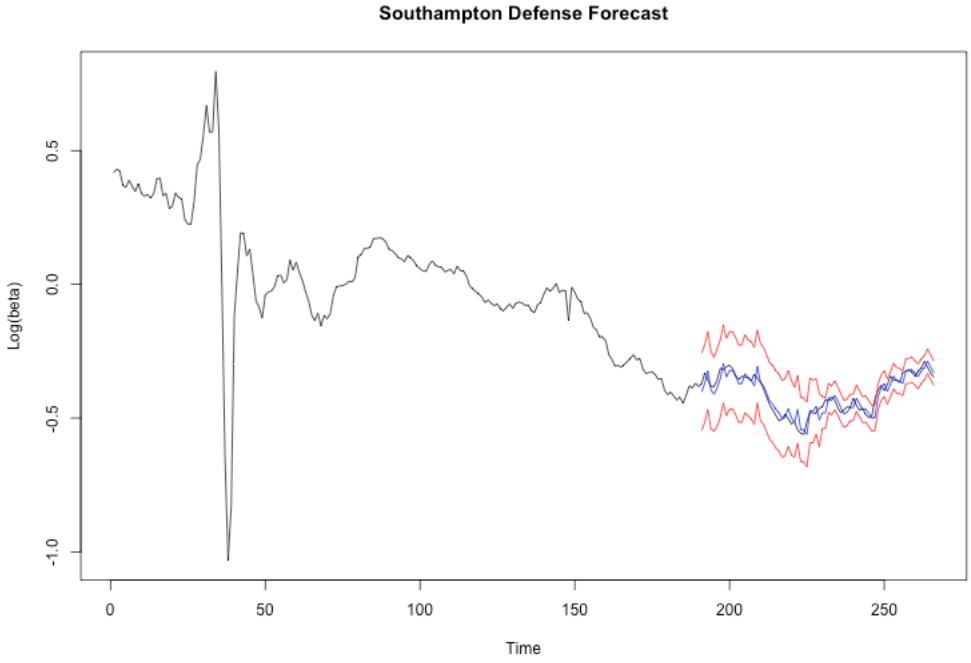
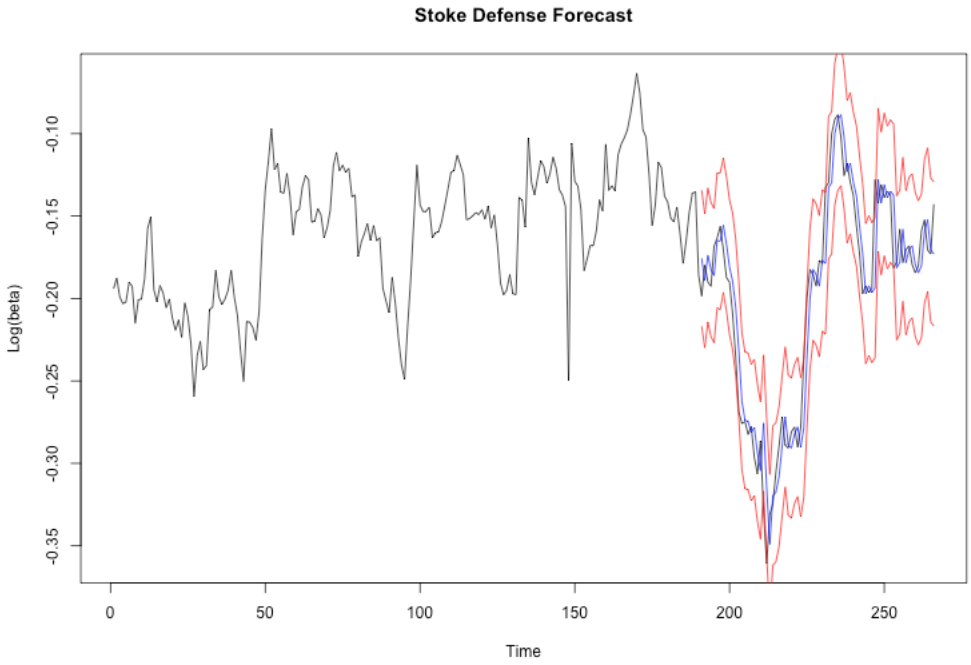


Figure 23 Continued



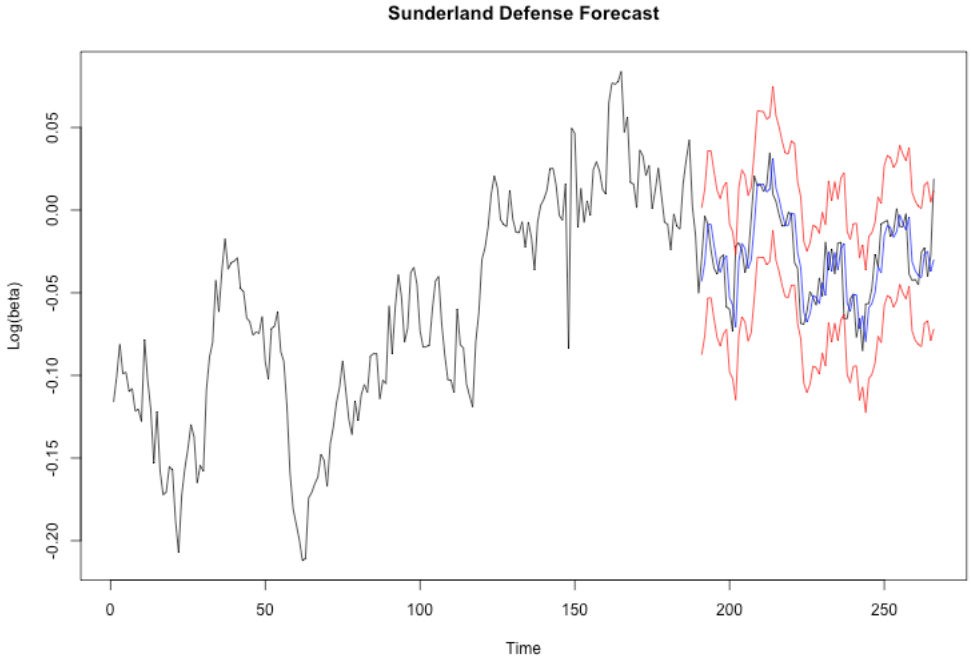
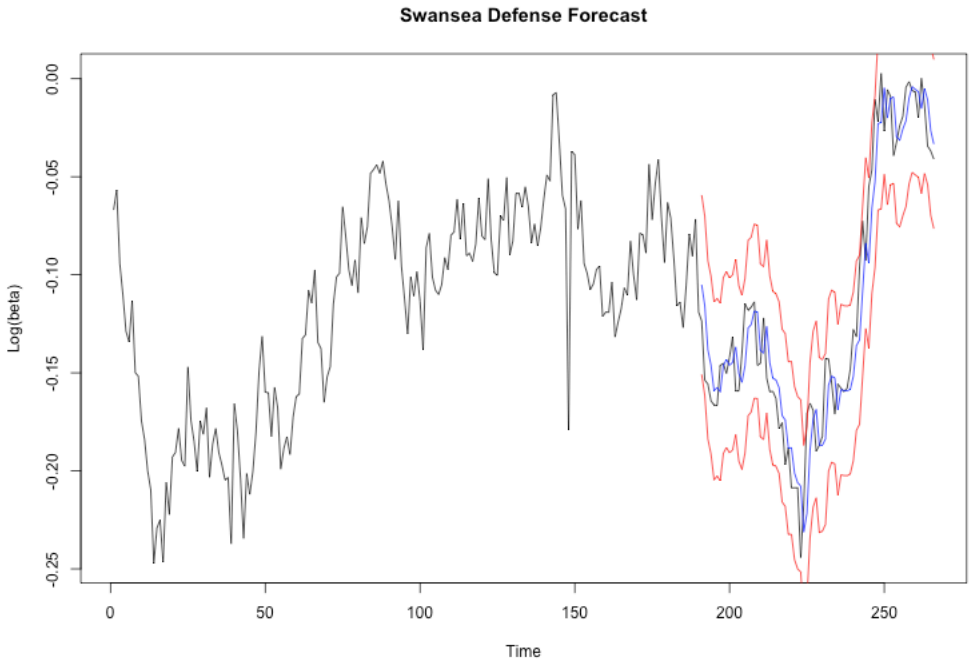


Figure 23 Continued



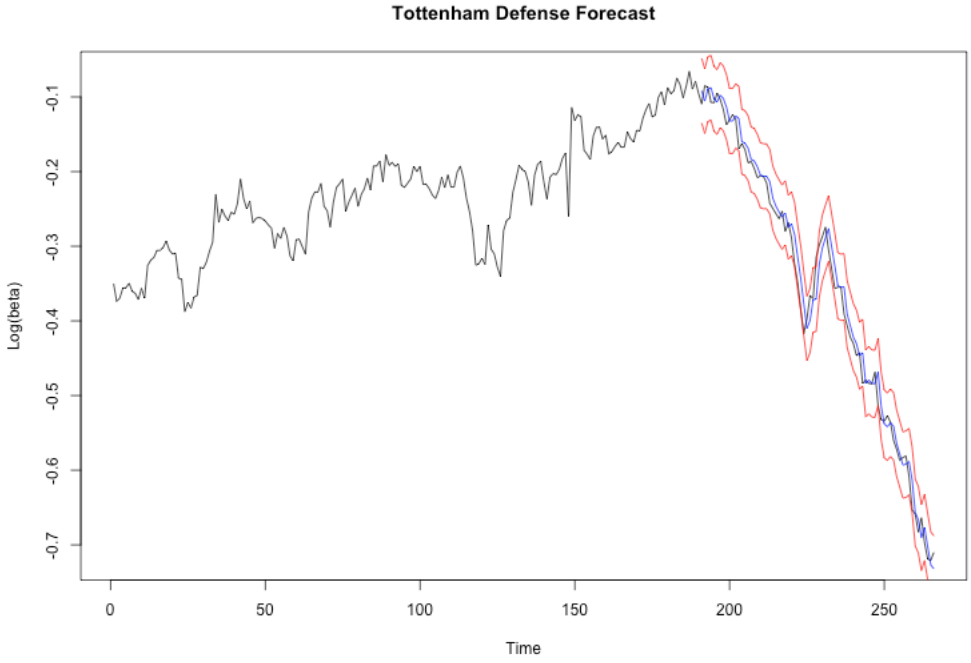
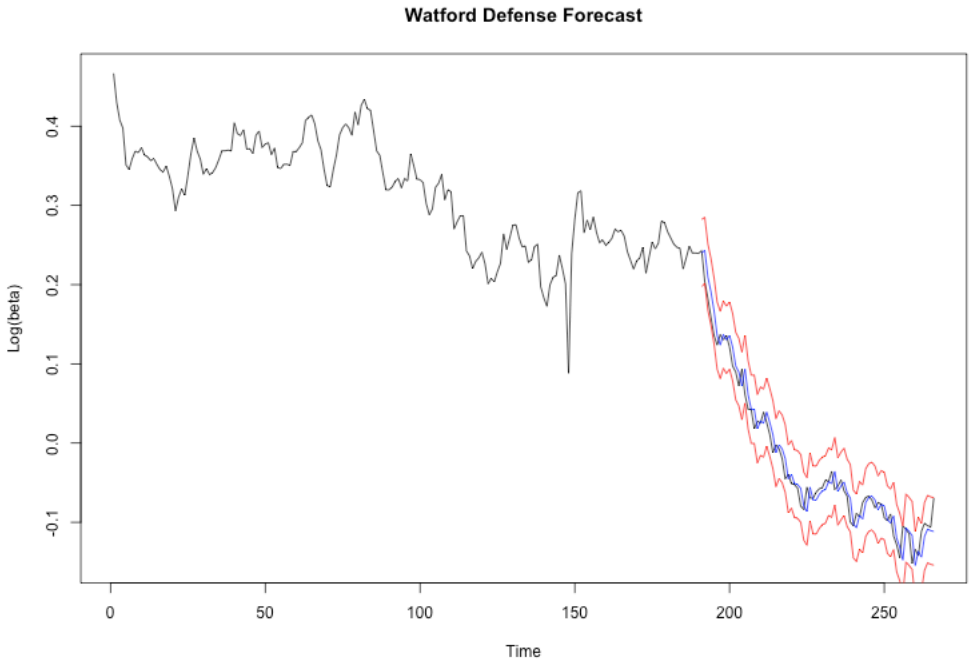


Figure 23 Continued



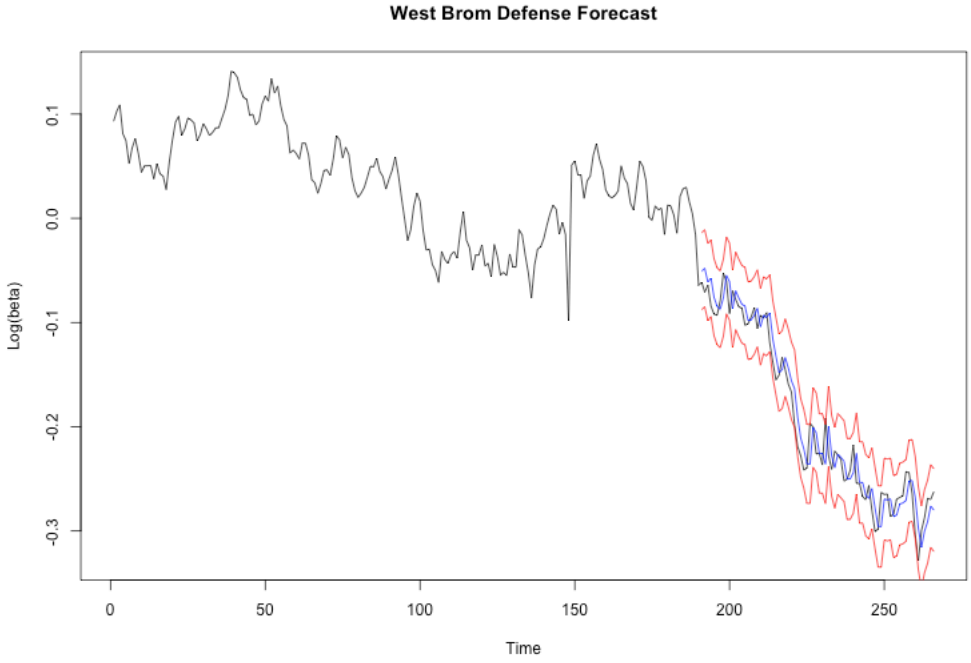


Figure 23 Continued

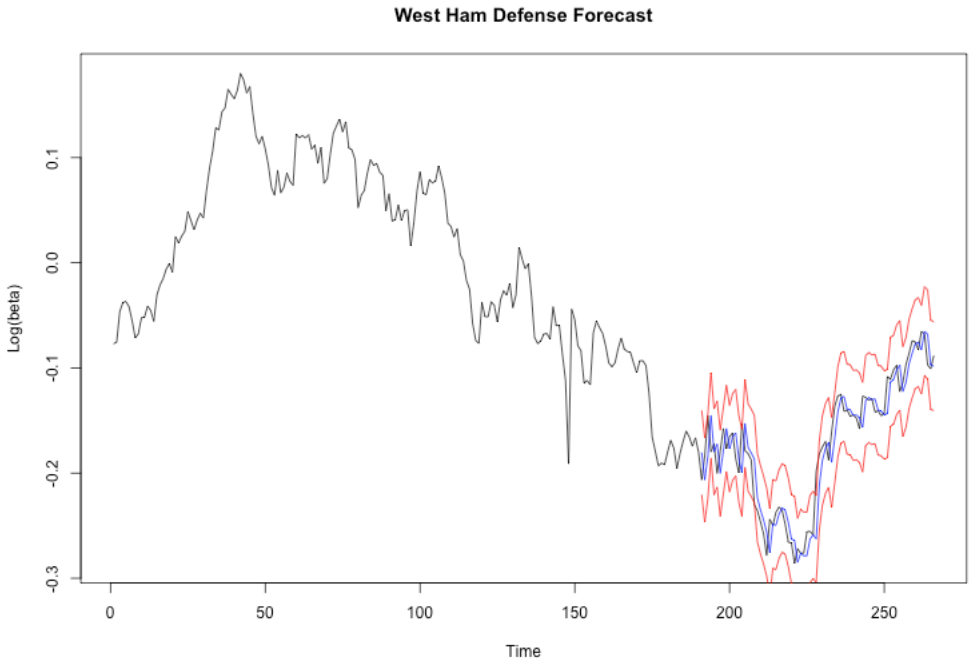


Figure 23 Continued

Appendix D – Forecasts from the time-varying parameter model

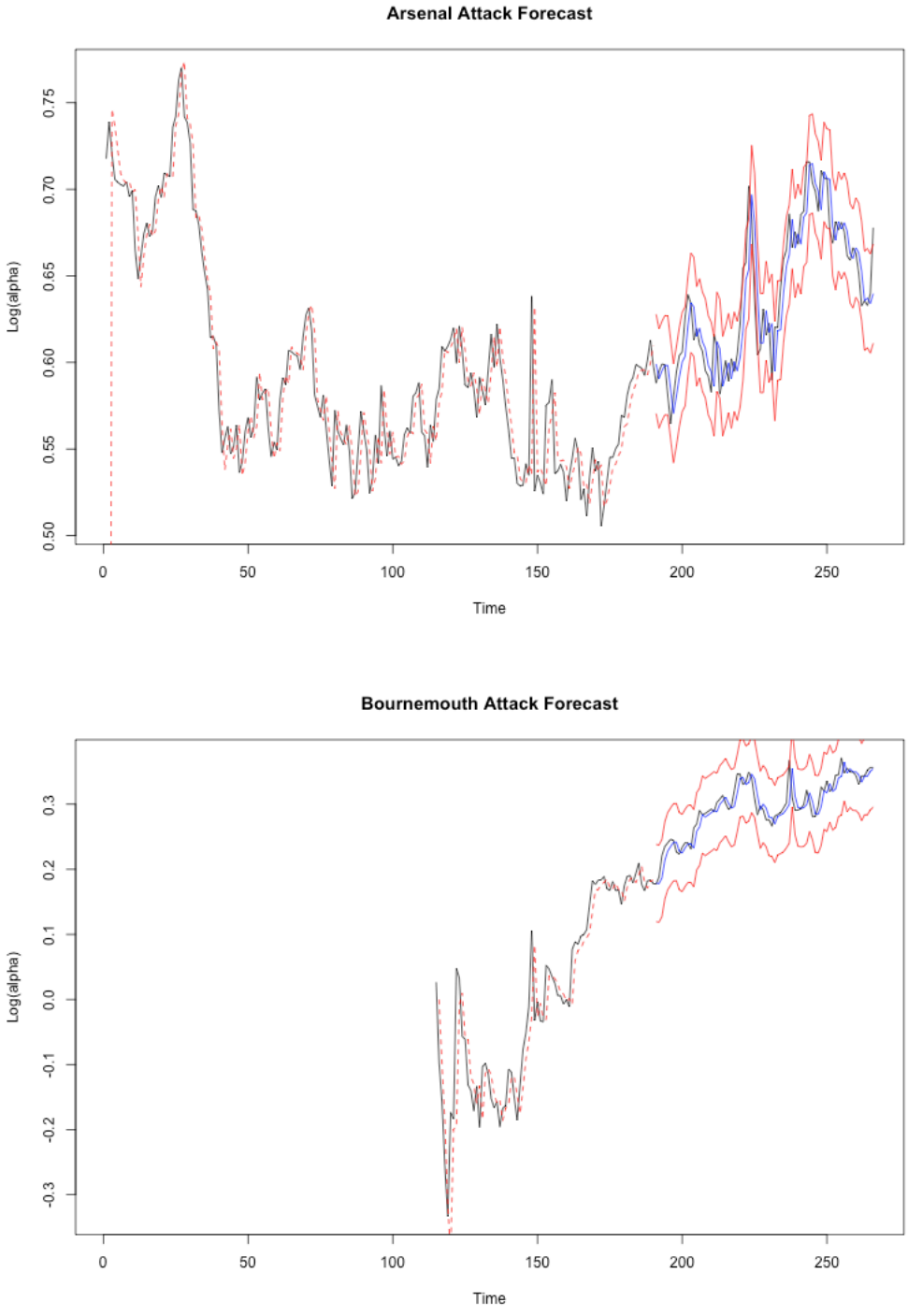
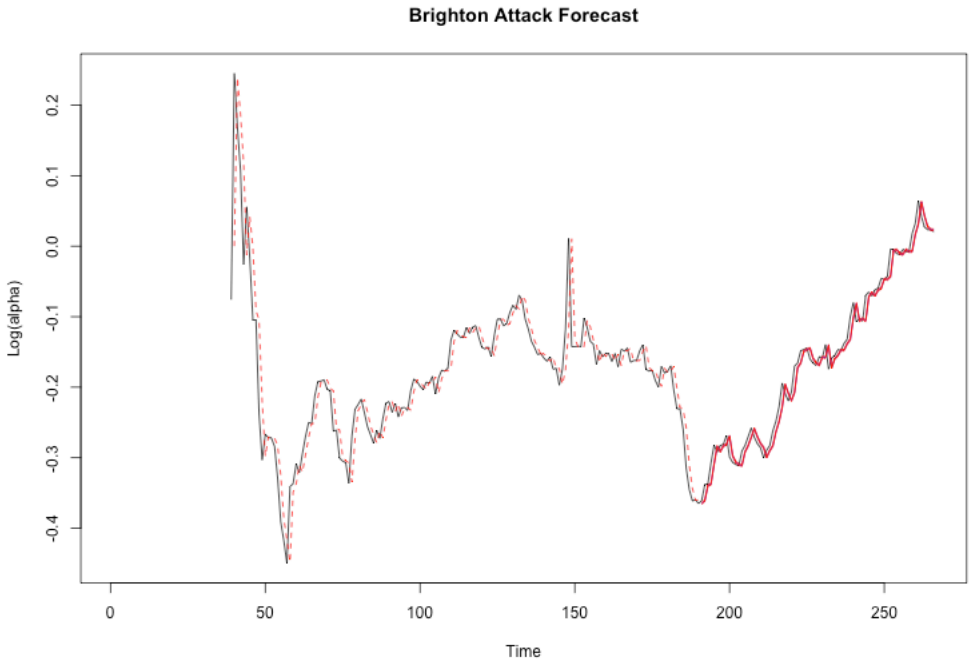


Figure 24 - (Model 4) Time-varying Parameter Forecasts using Kalman Filter – Attack



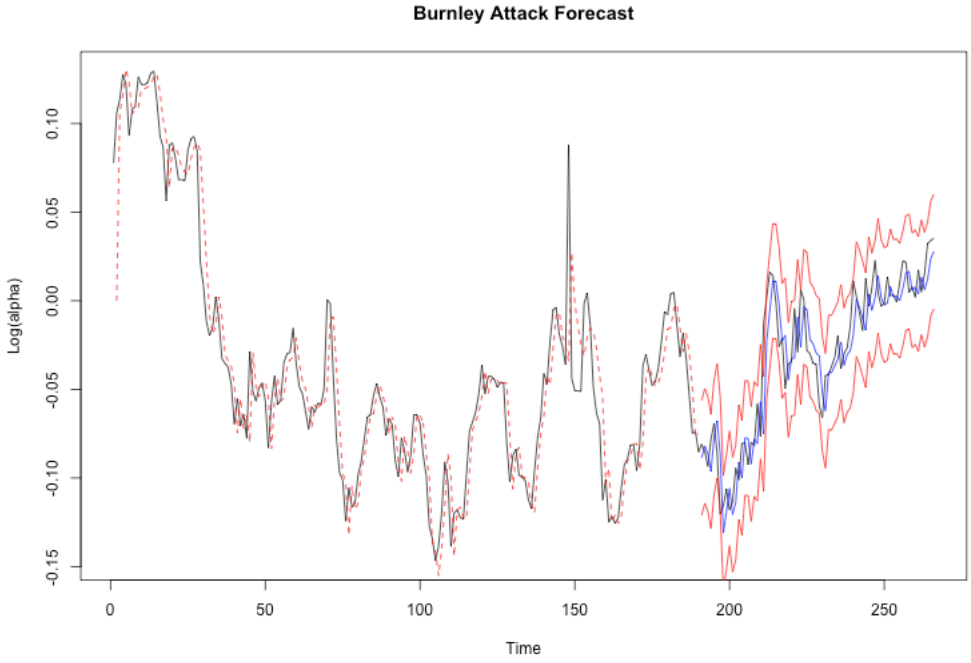
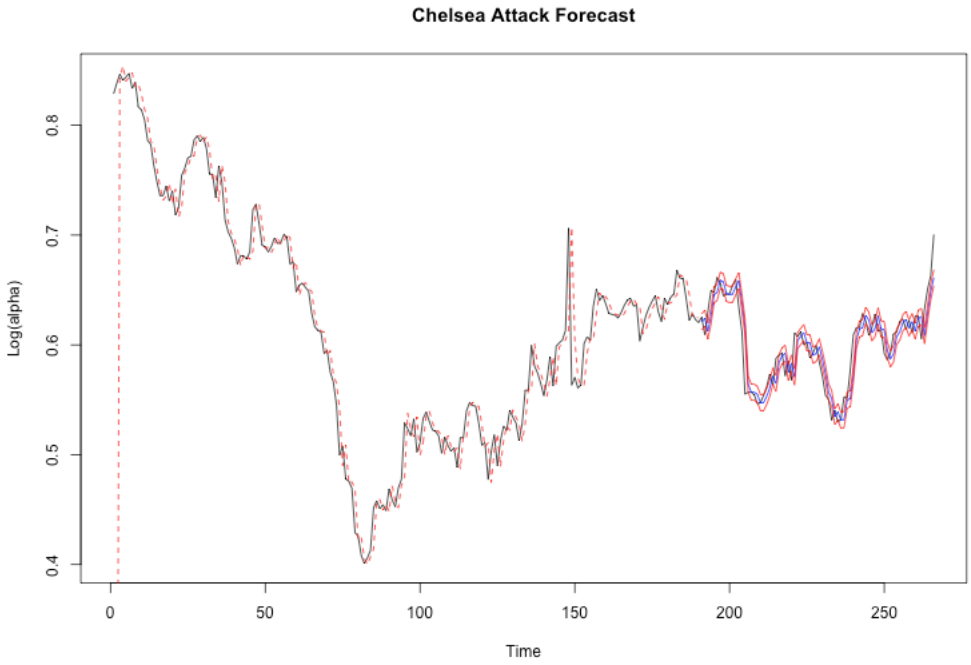


Figure 24 Continued



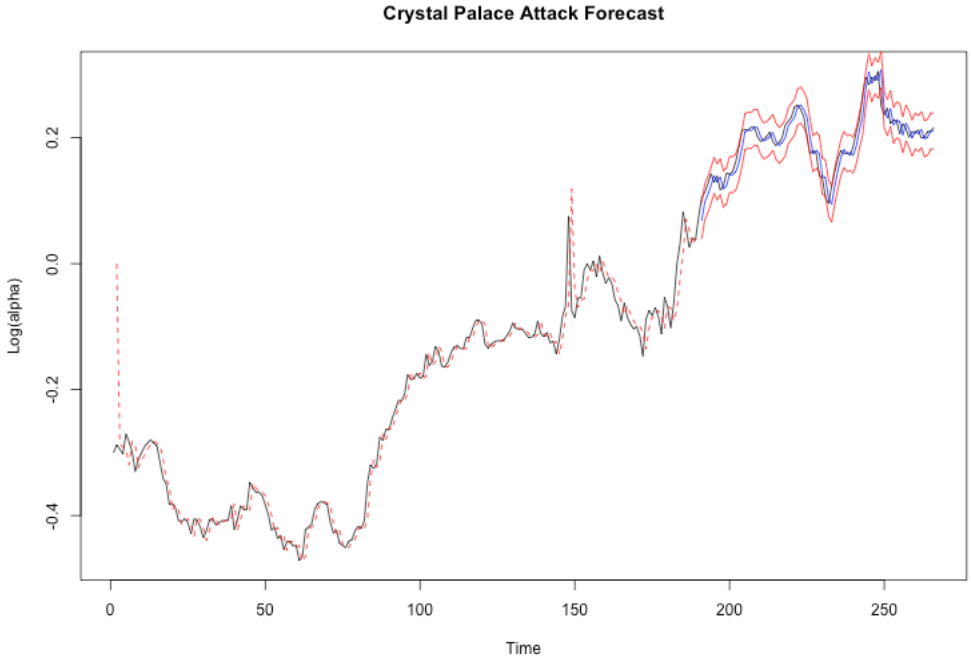
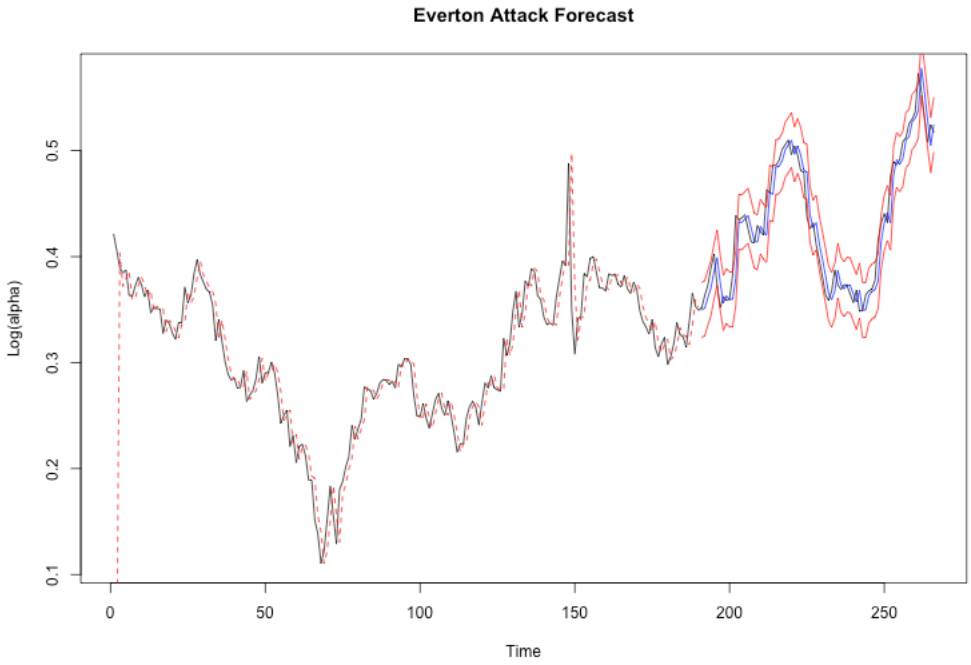


Figure 24 Continued



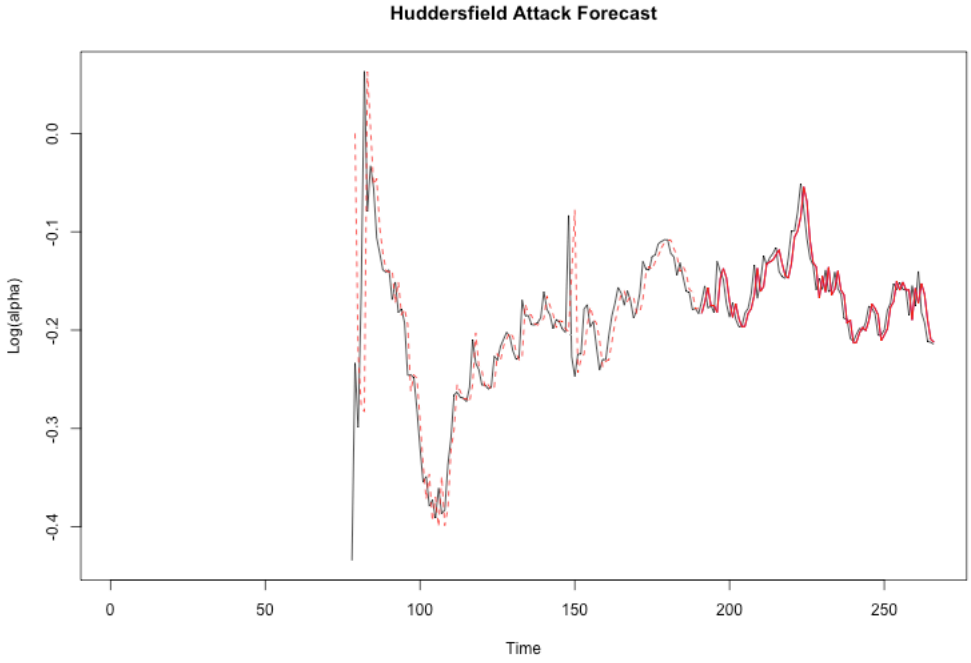
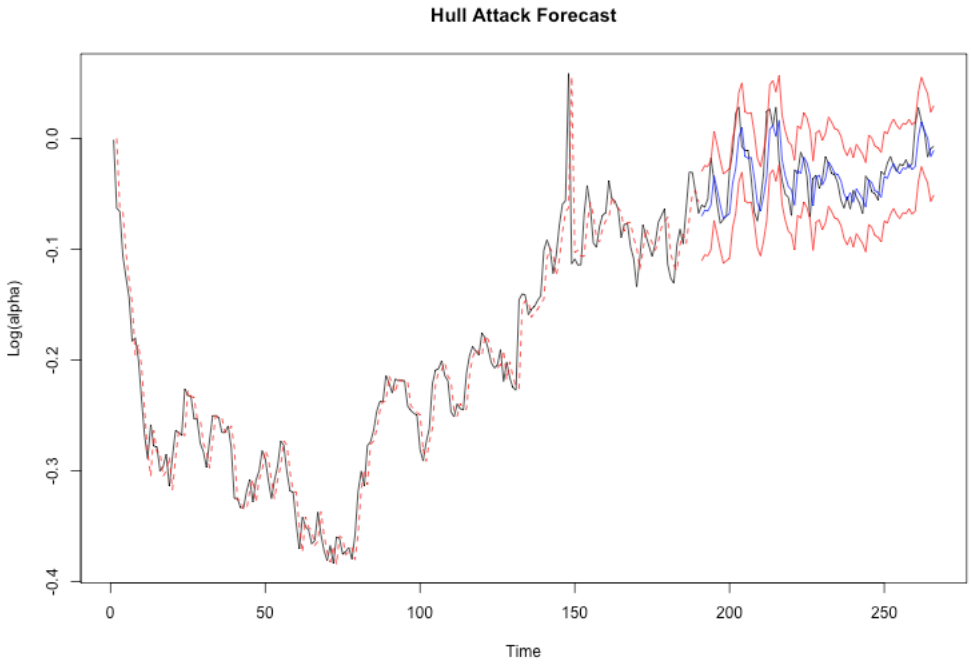


Figure 24 Continued



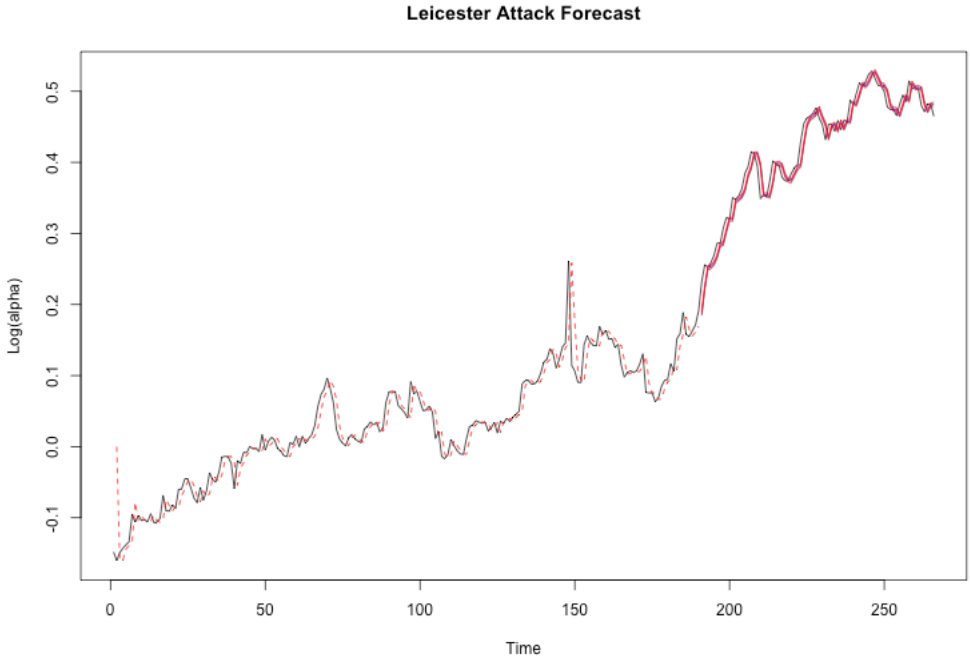
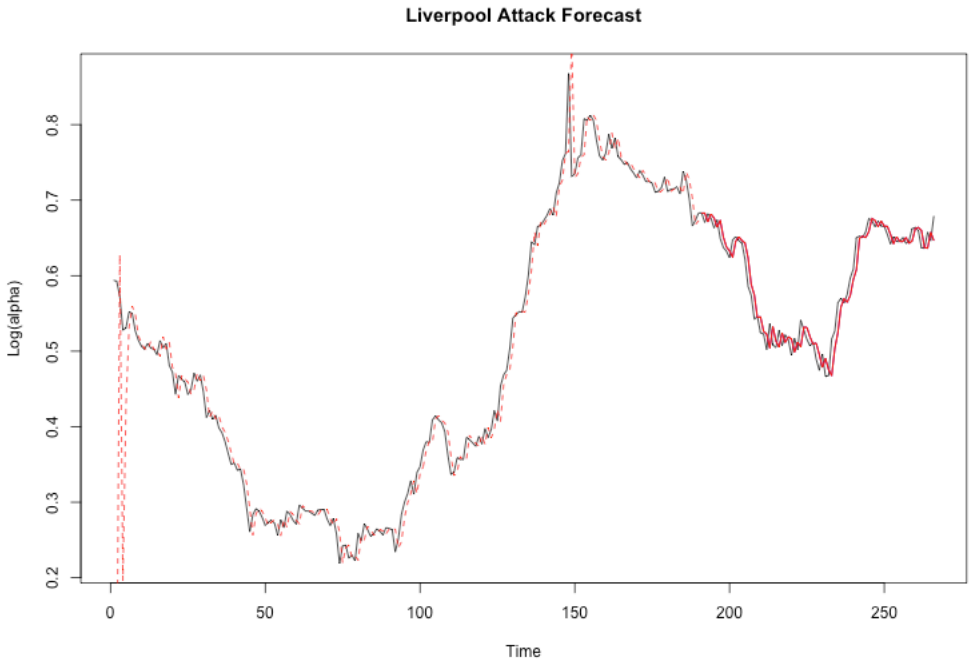


Figure 24 Continued



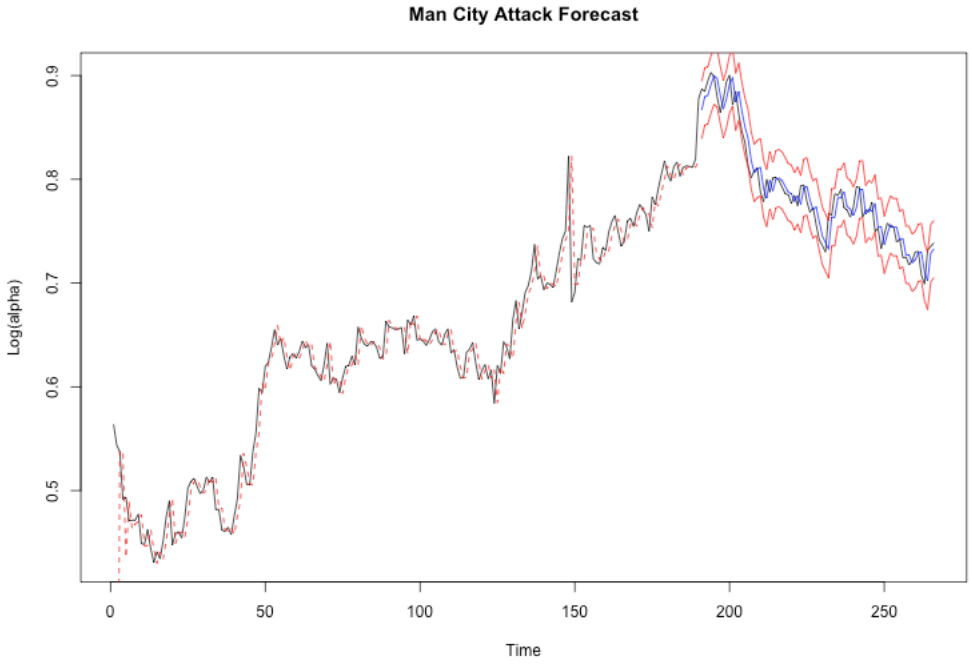
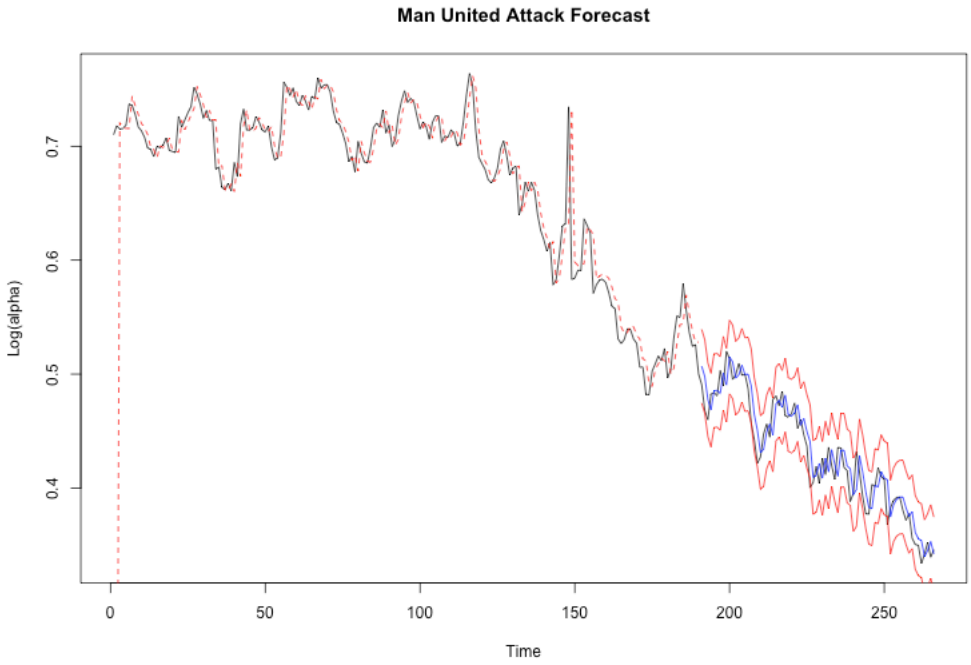


Figure 24 Continued



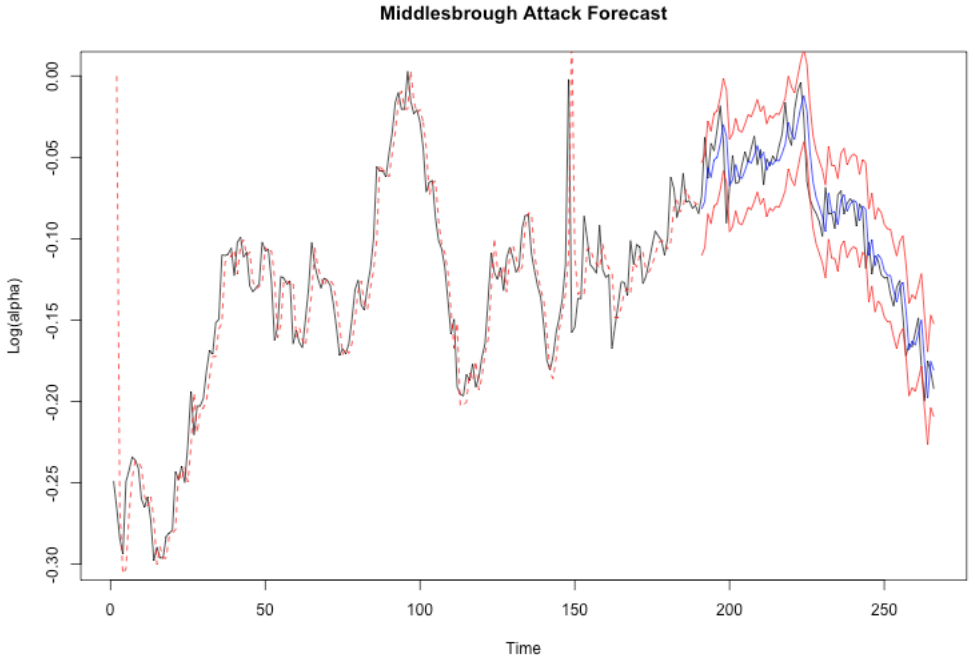
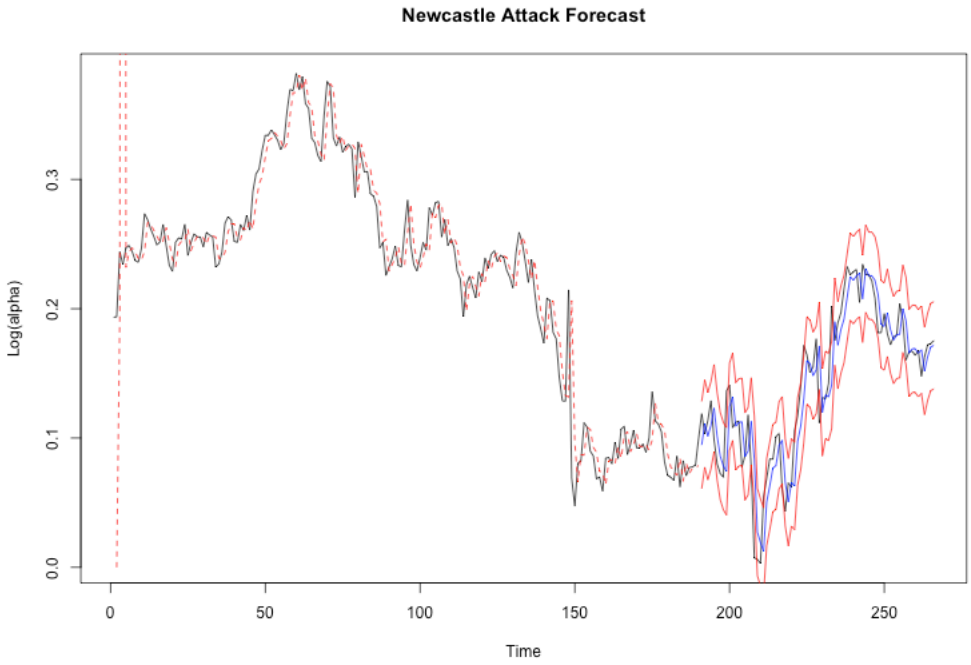


Figure 24 Continued



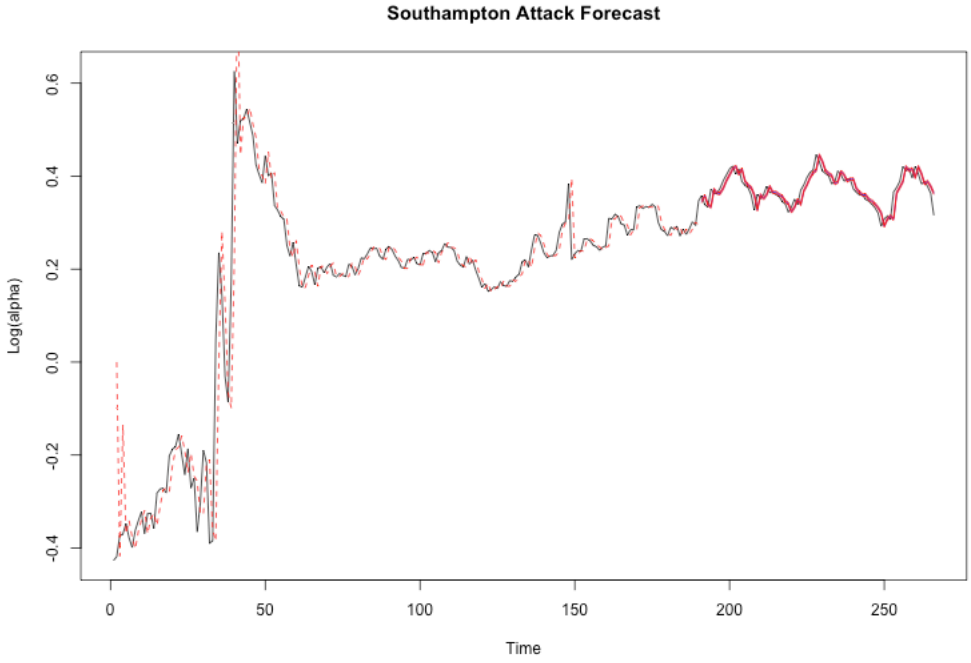
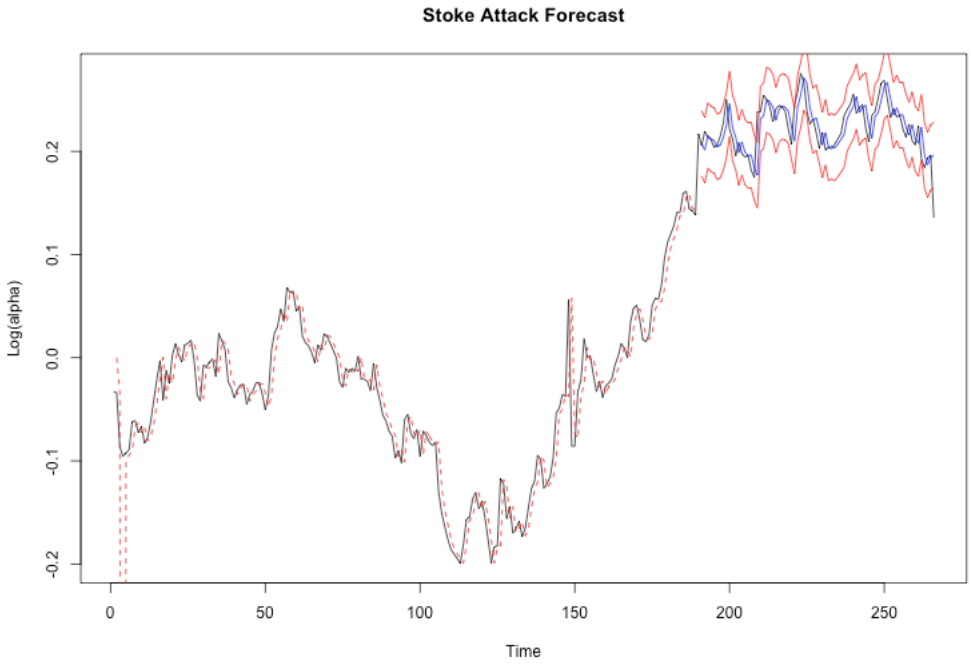


Figure 24 Continued



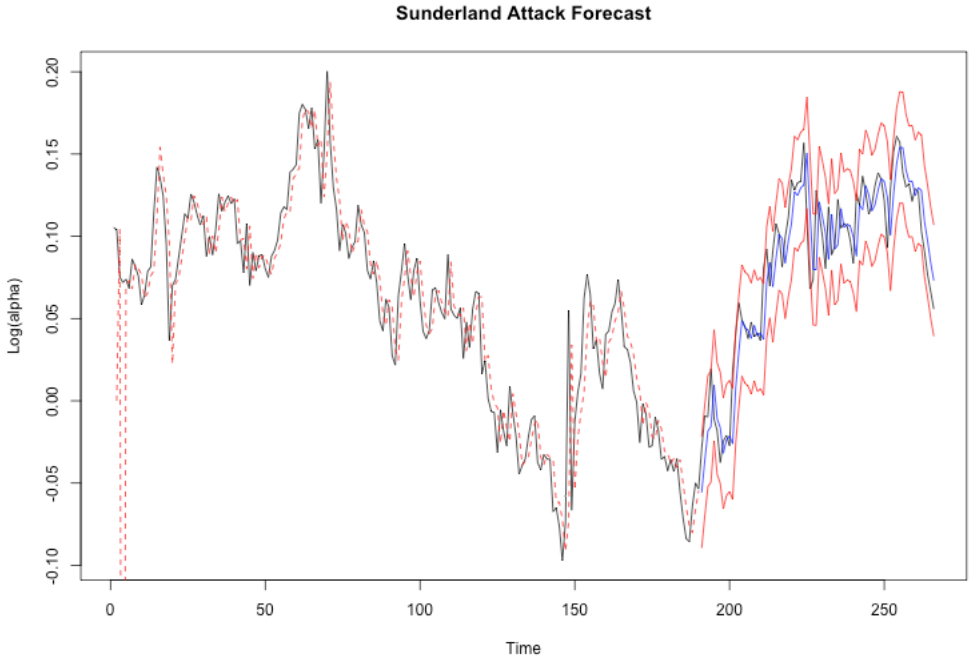
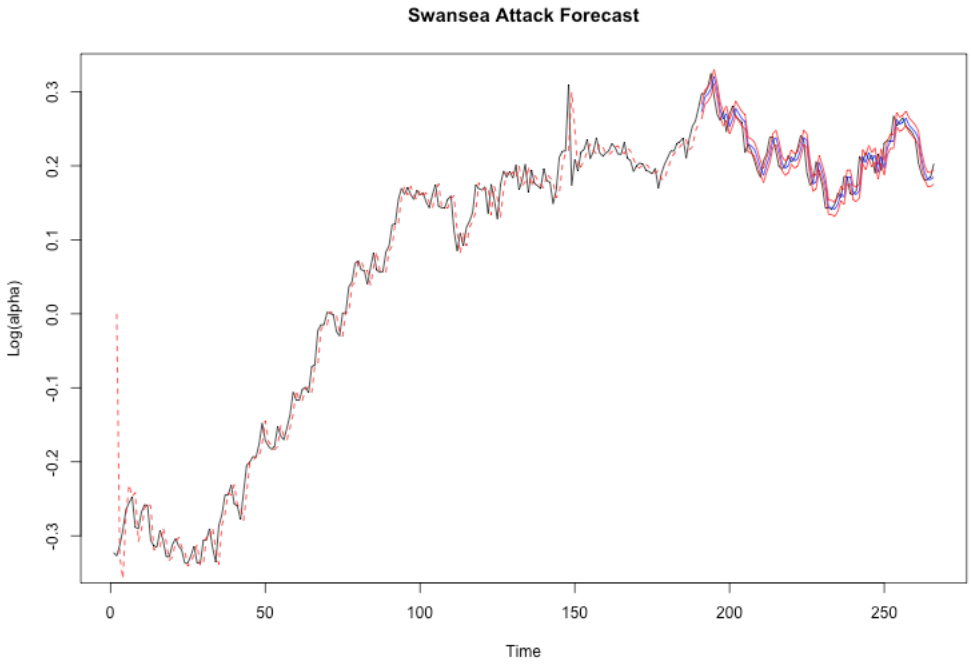


Figure 24 Continued



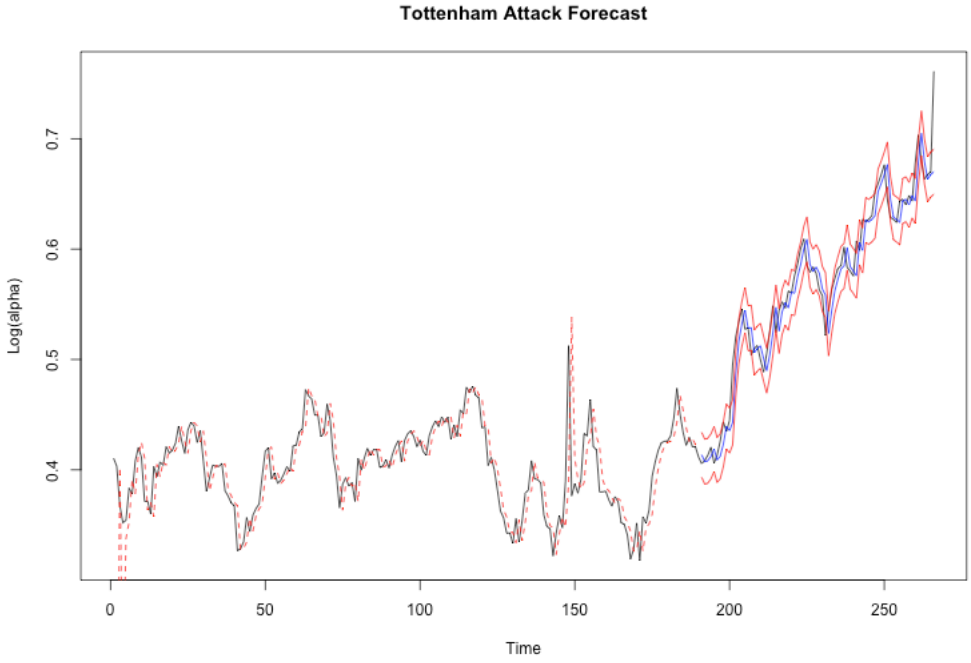
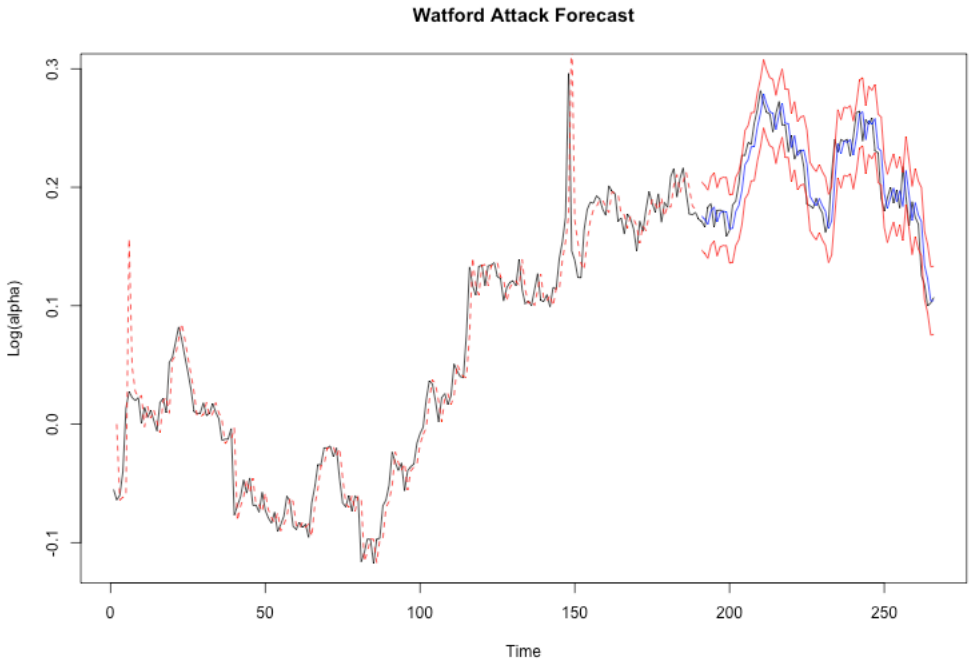


Figure 24 Continued



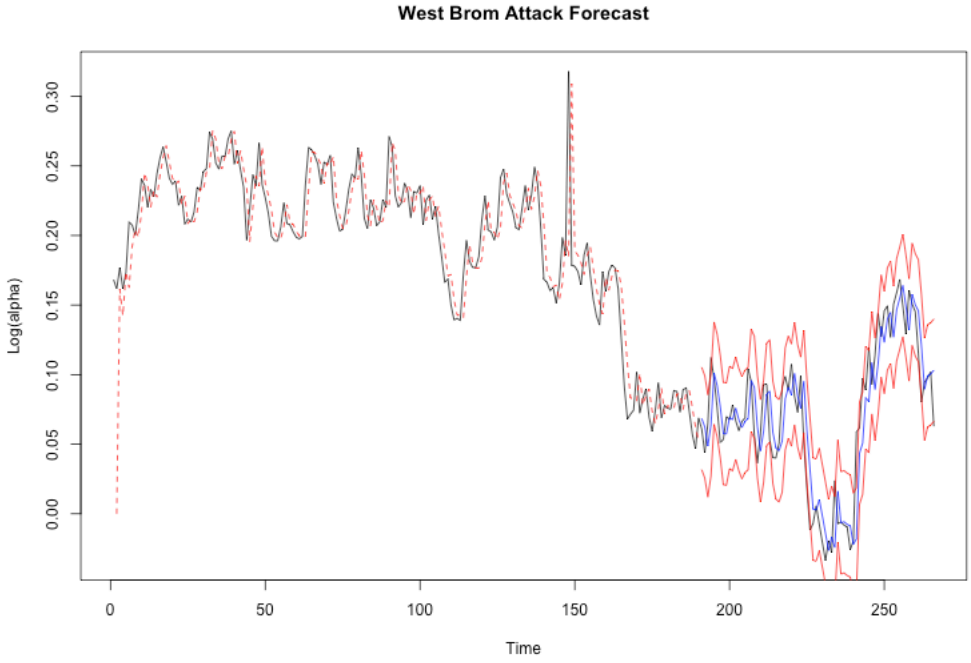


Figure 24 Continued

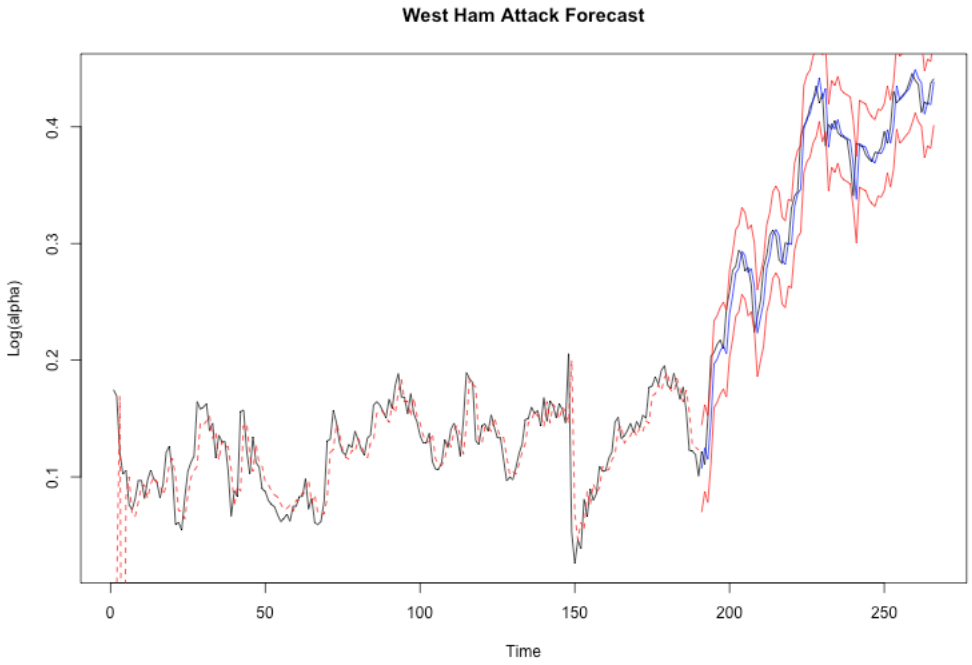
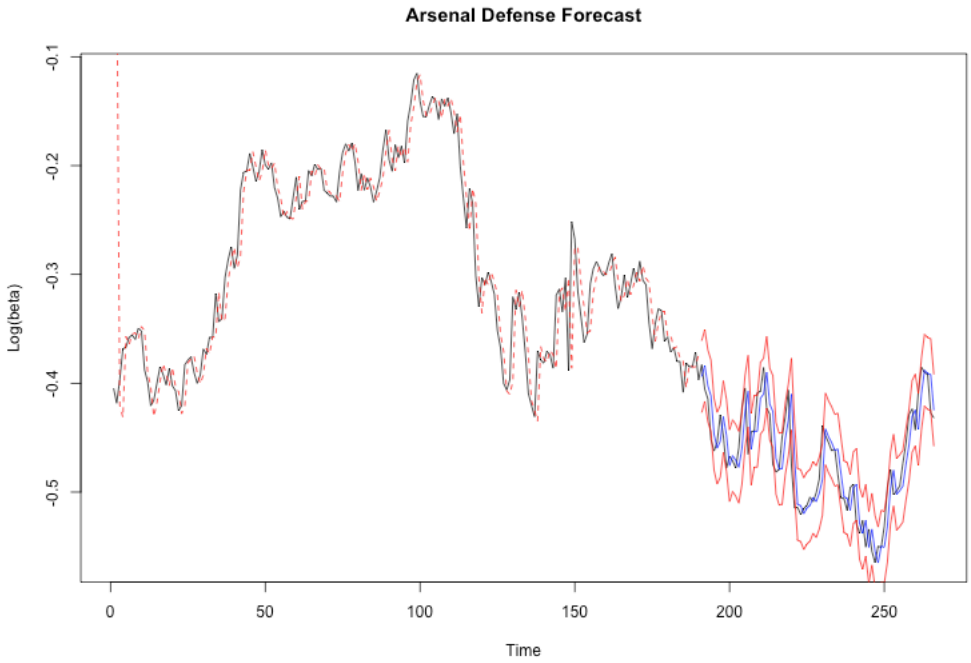


Figure 24 Continued



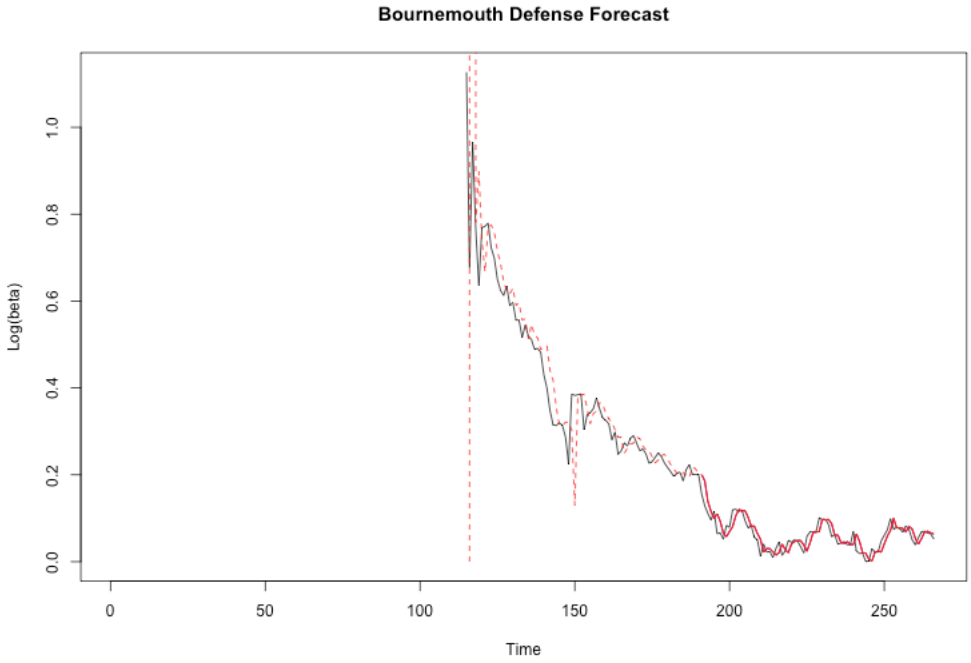
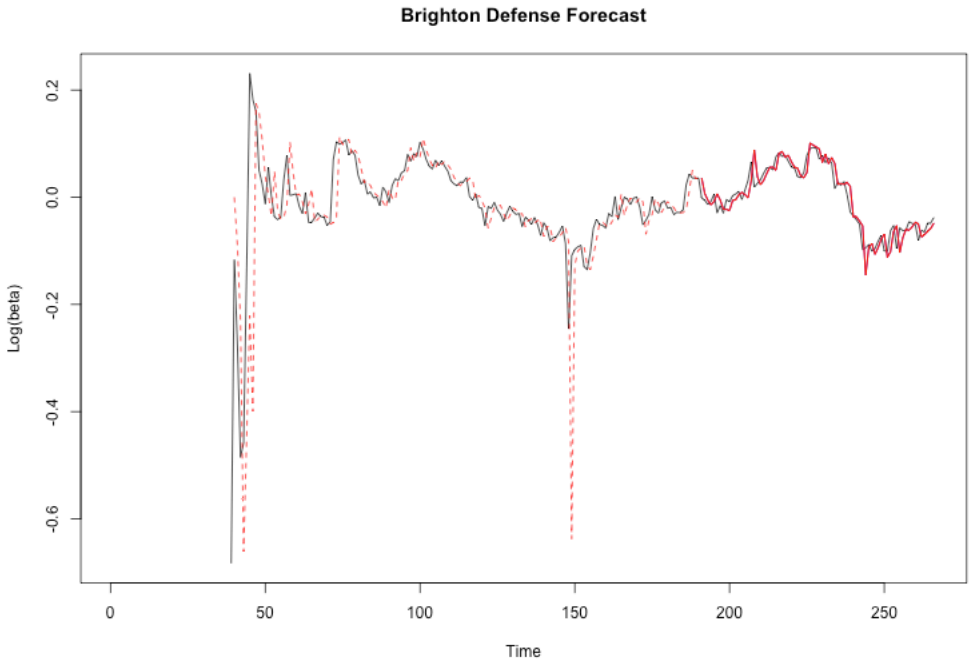


Figure 25 - (Model 4) Time-varying Parameter Forecasts using Kalman Filter – Defence



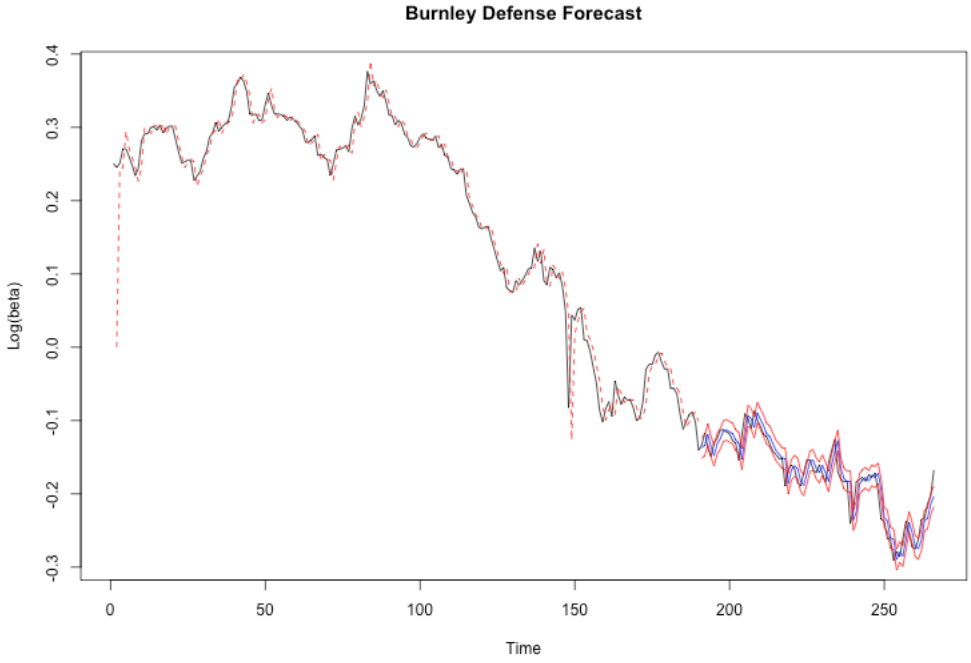
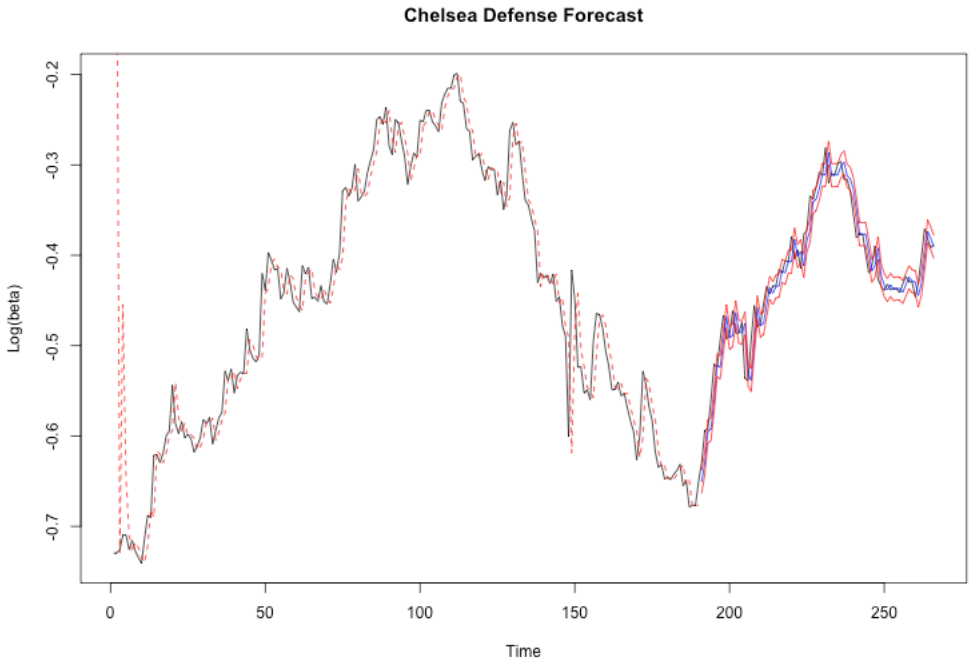


Figure 25 Continued



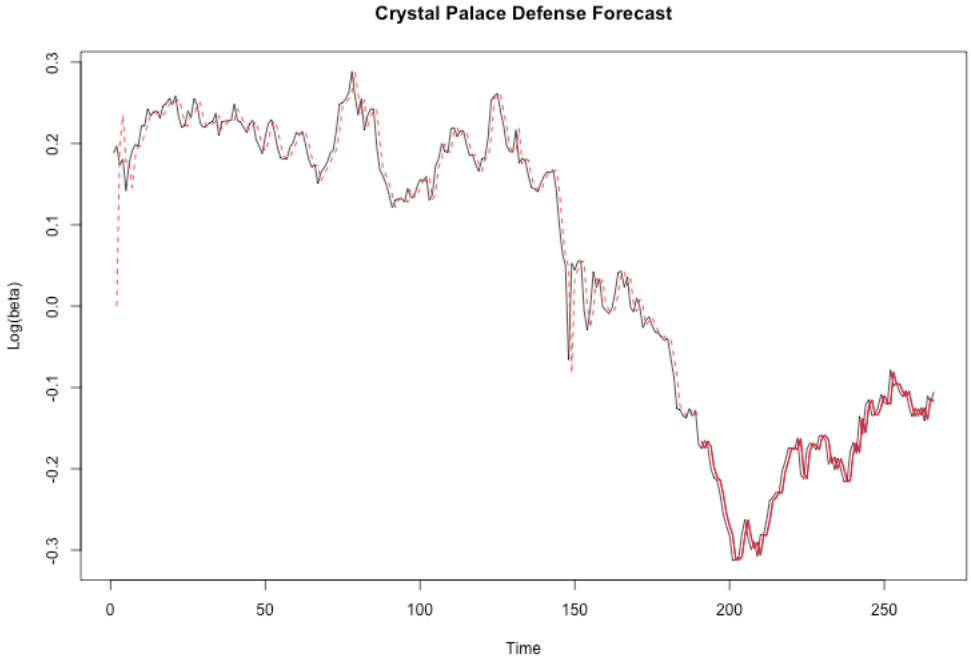
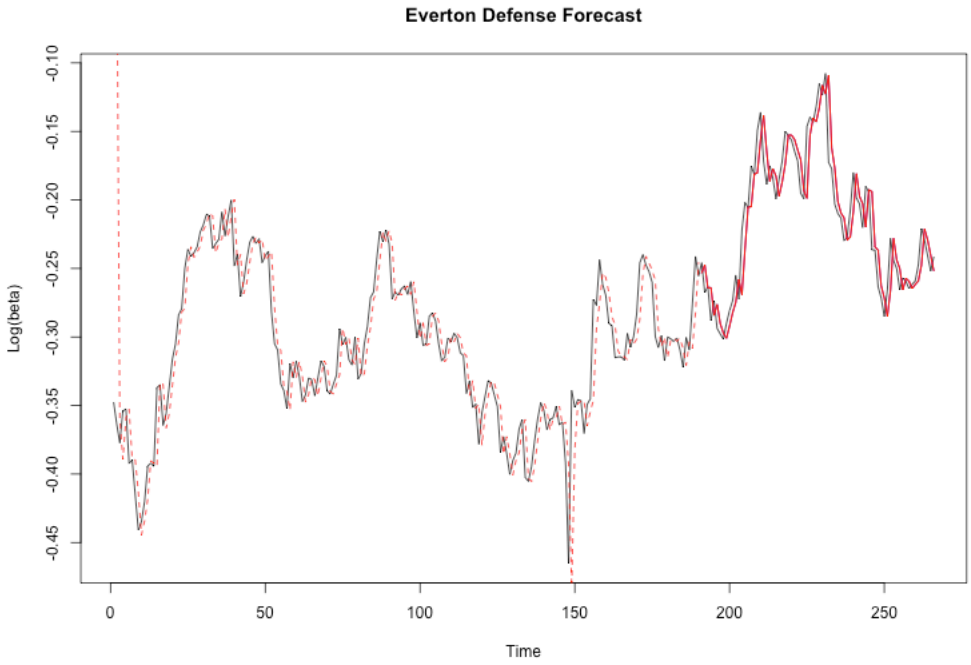


Figure 25 Continued



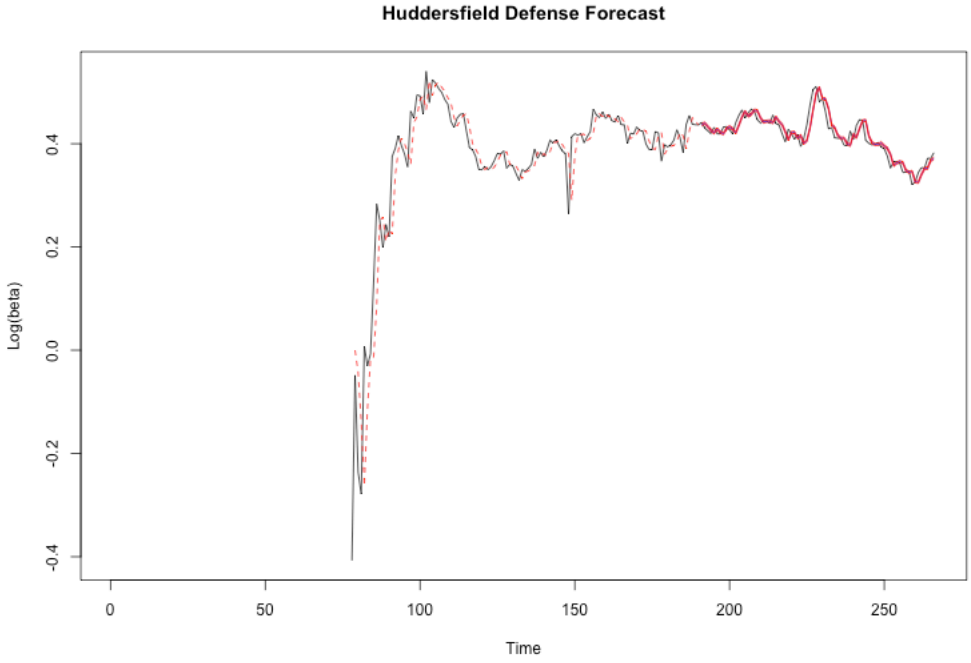
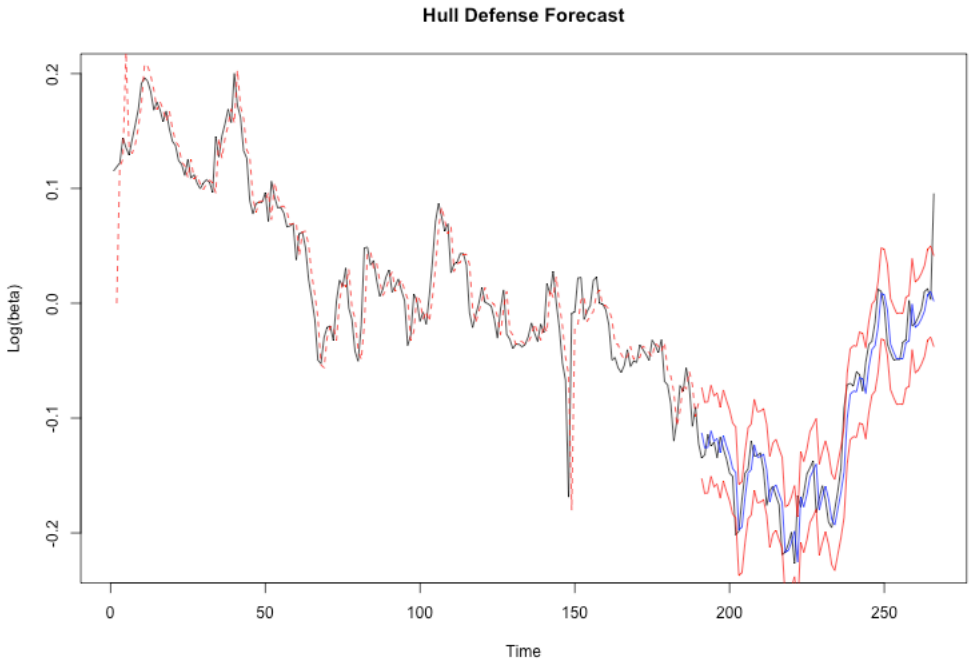


Figure 25 Continued



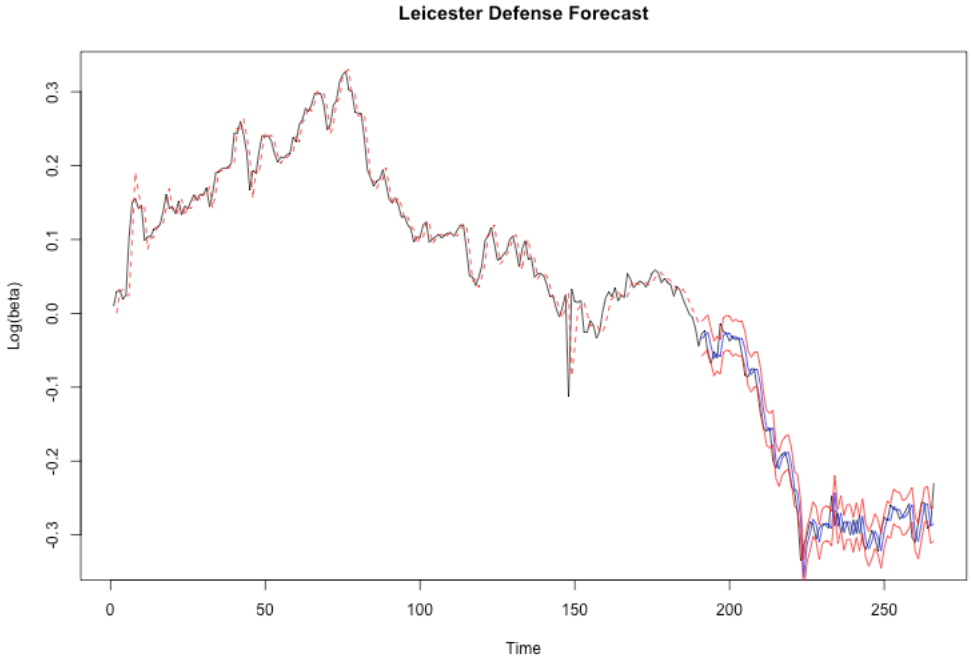
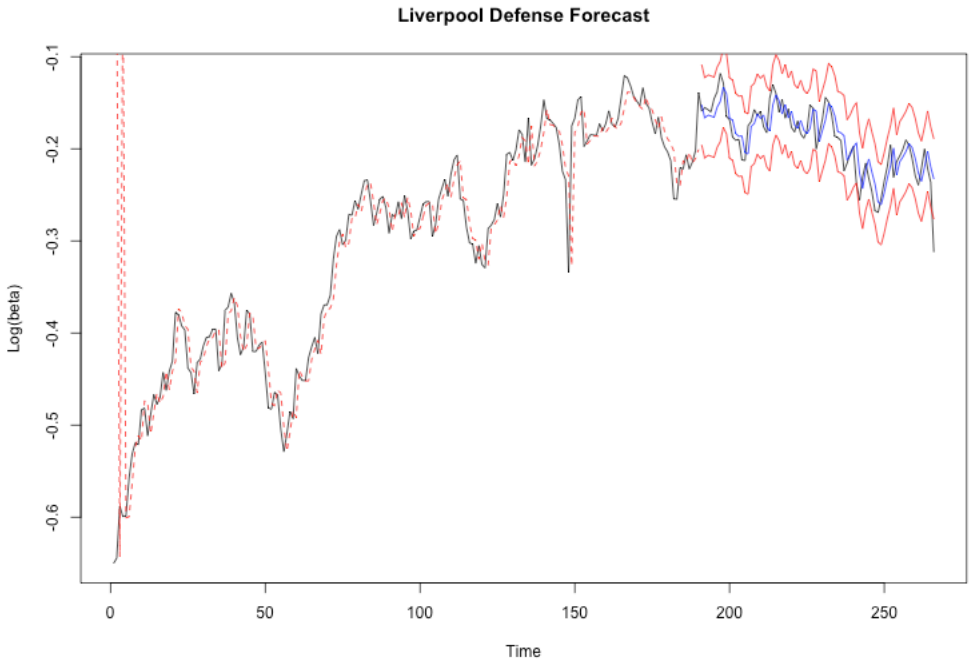


Figure 25 Continued



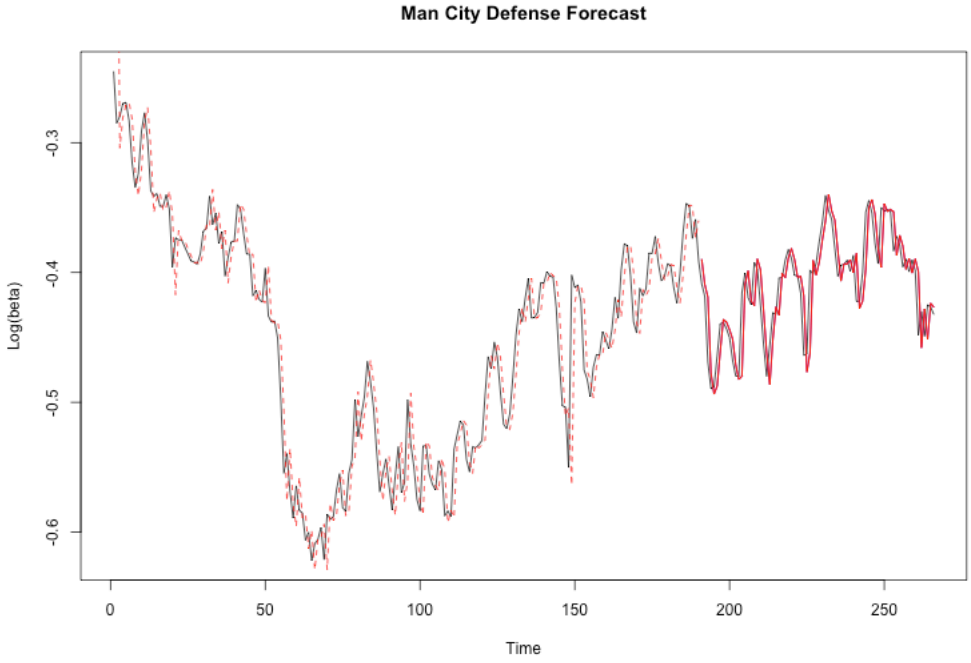
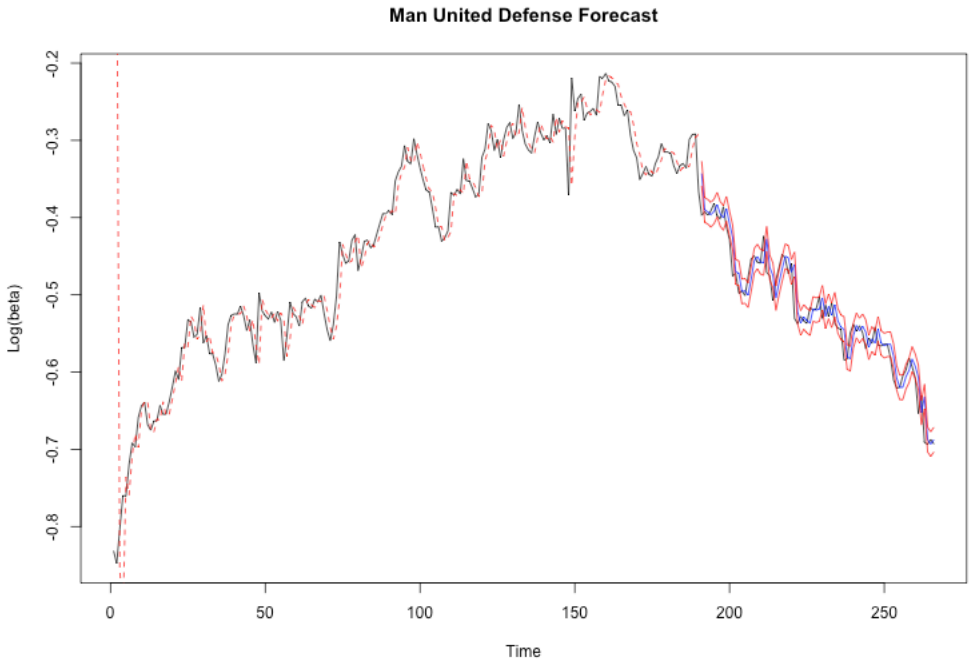


Figure 25 Continued



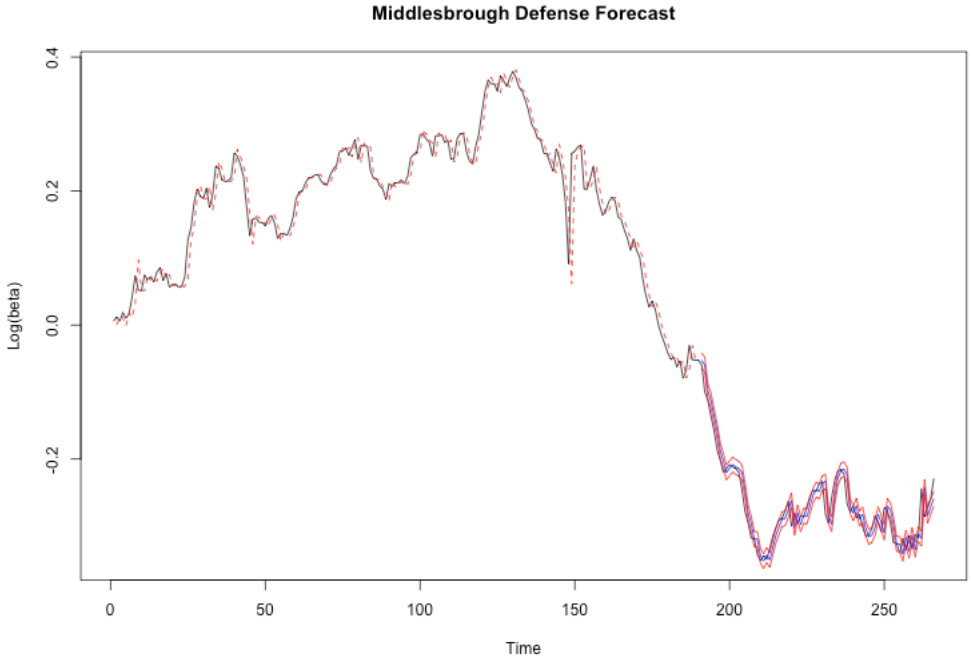
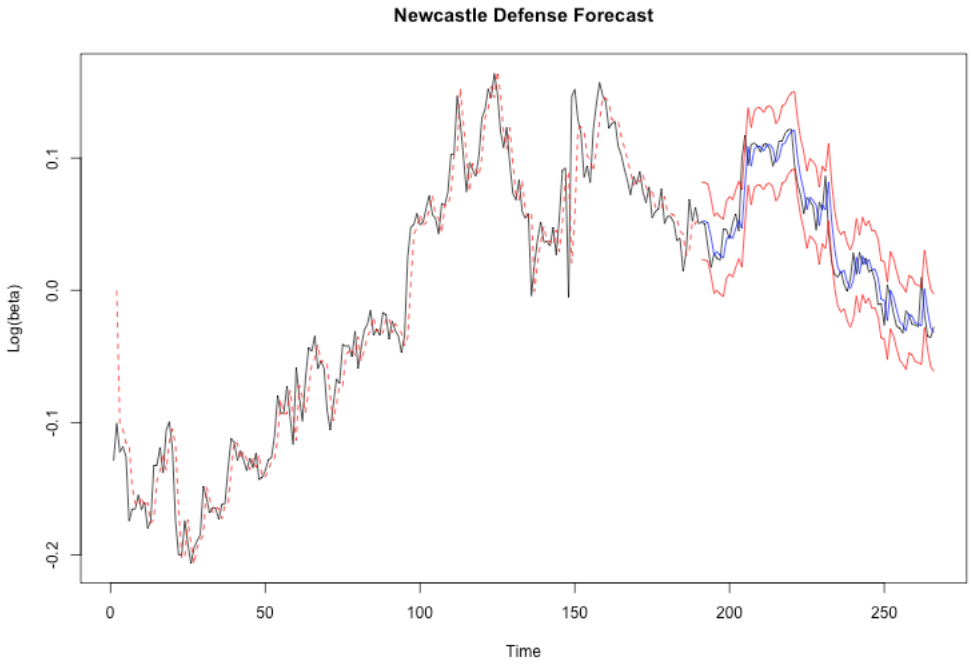


Figure 25 Continued



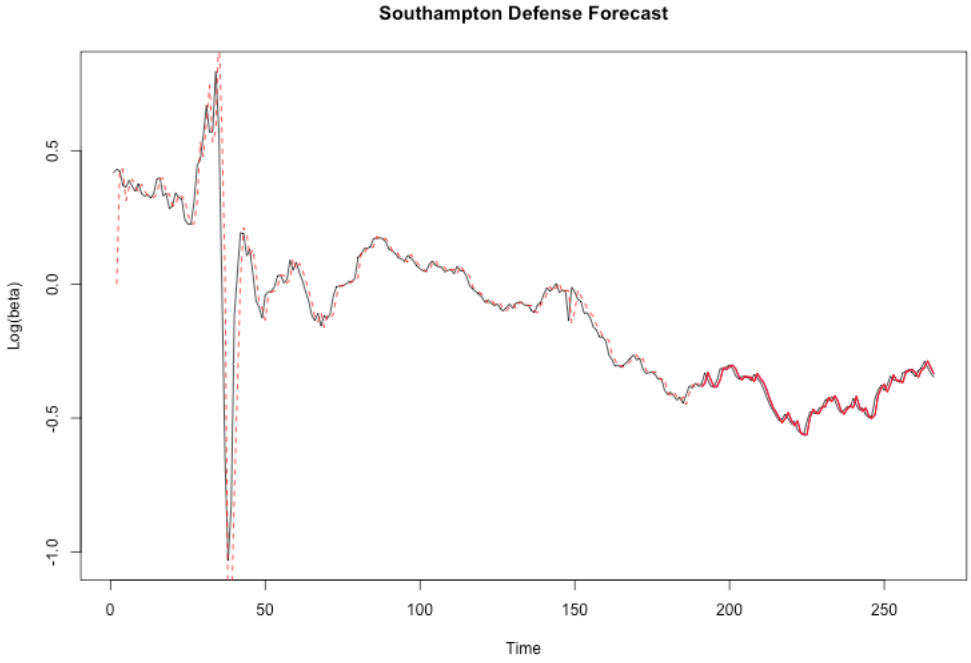
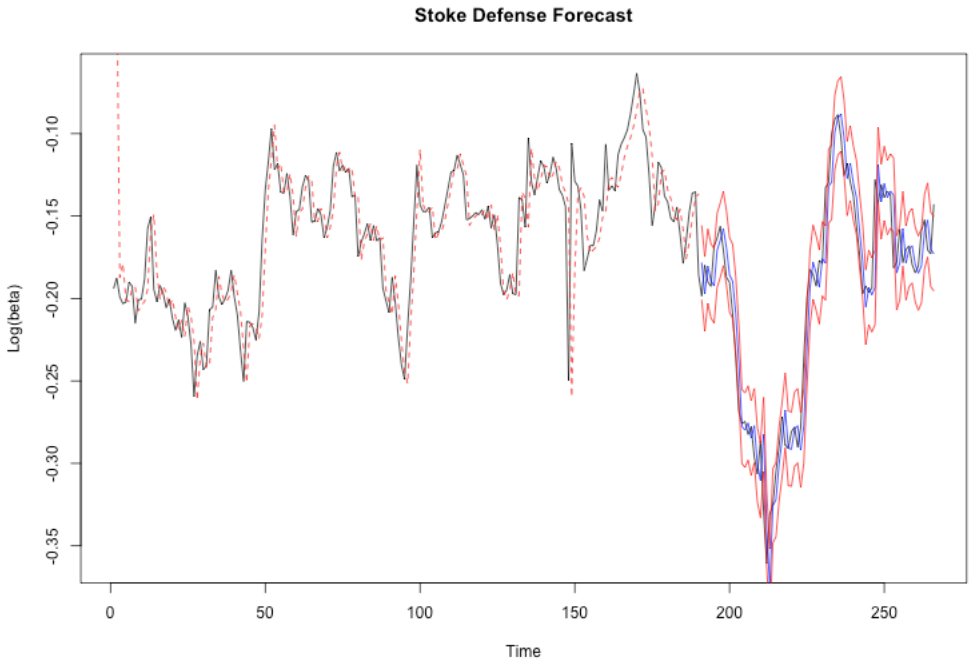


Figure 25 Continued



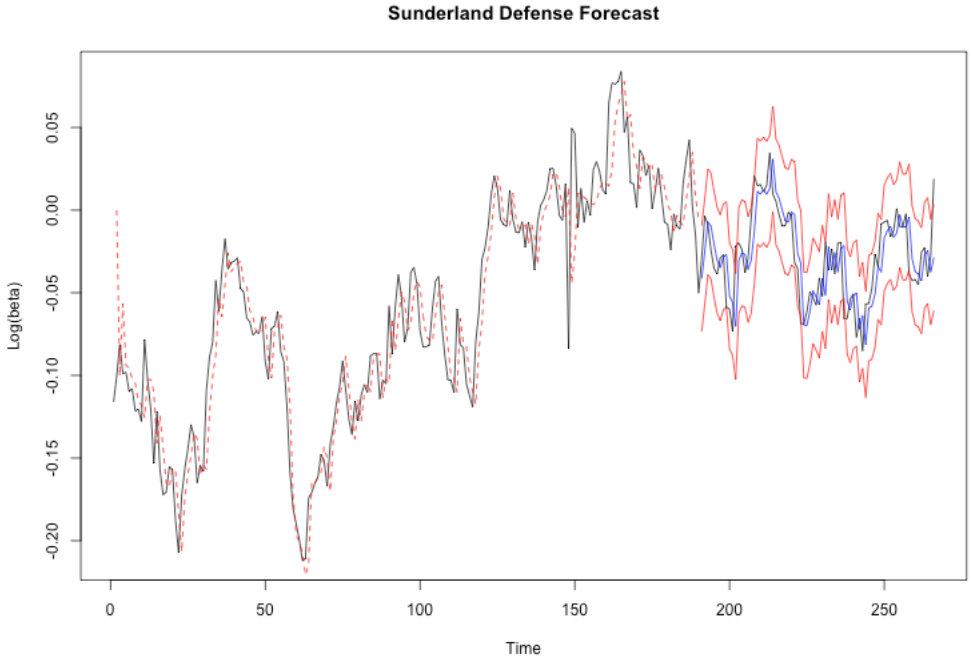
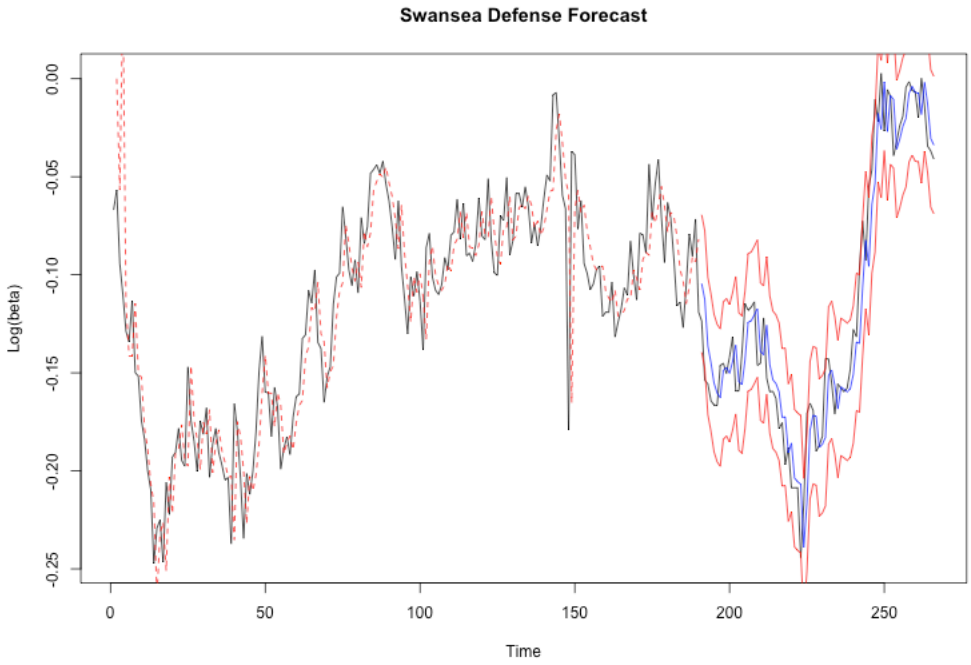


Figure 25 Continued



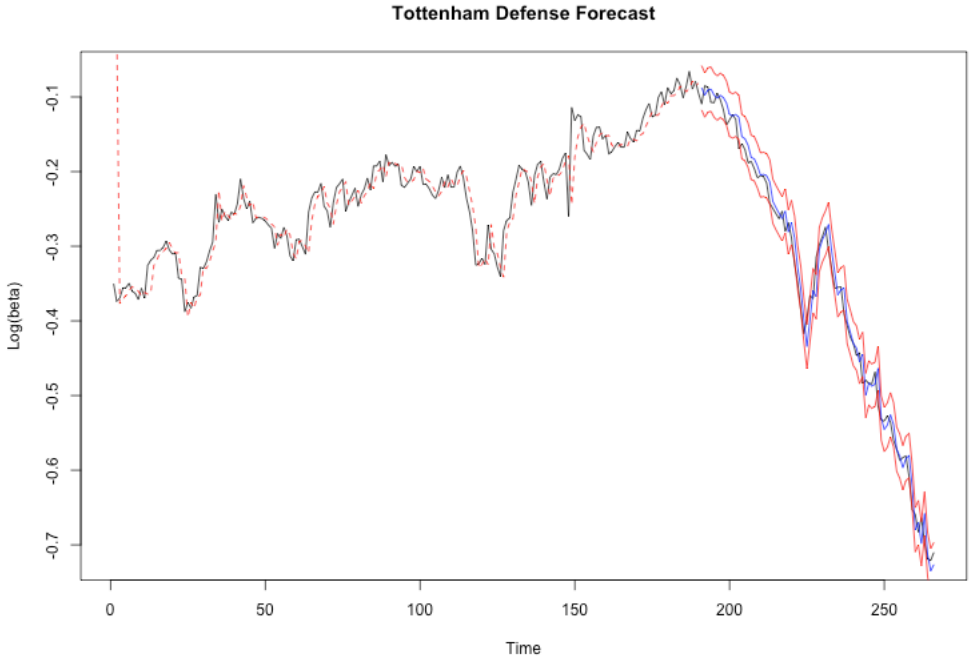
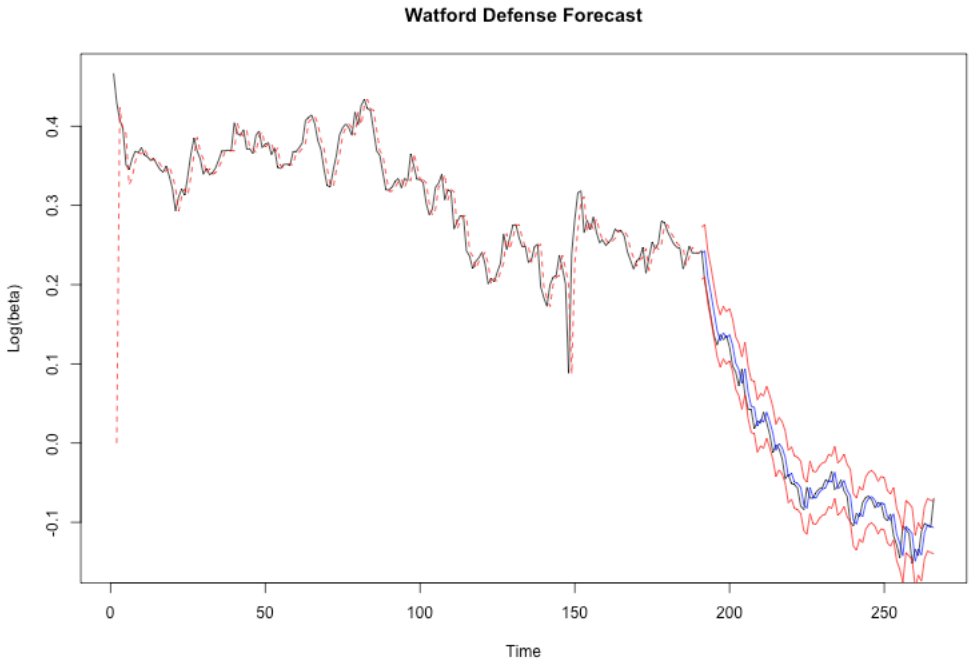


Figure 25 Continued



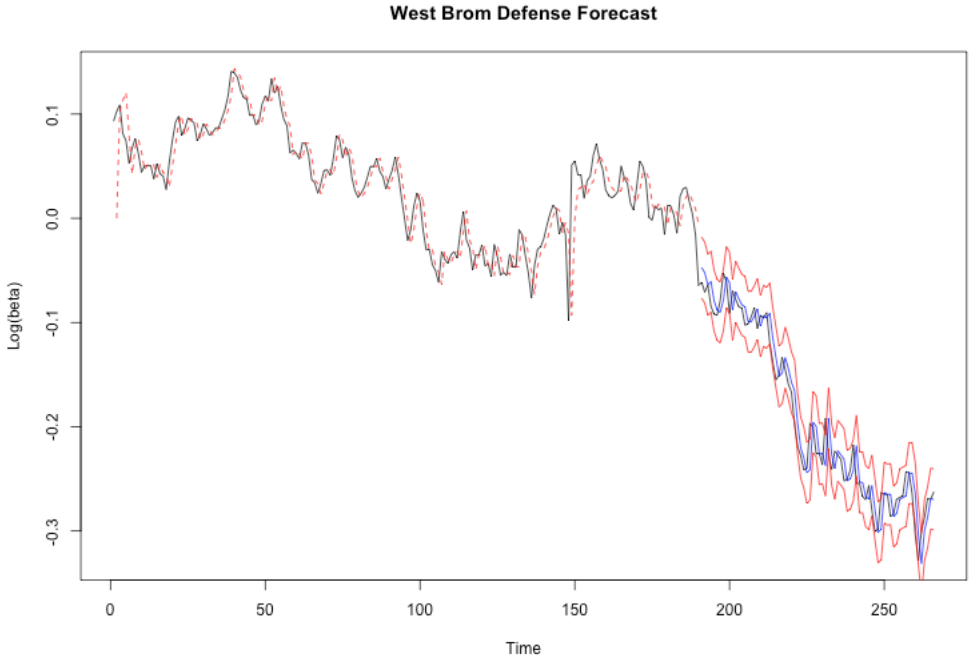


Figure 25 Continued

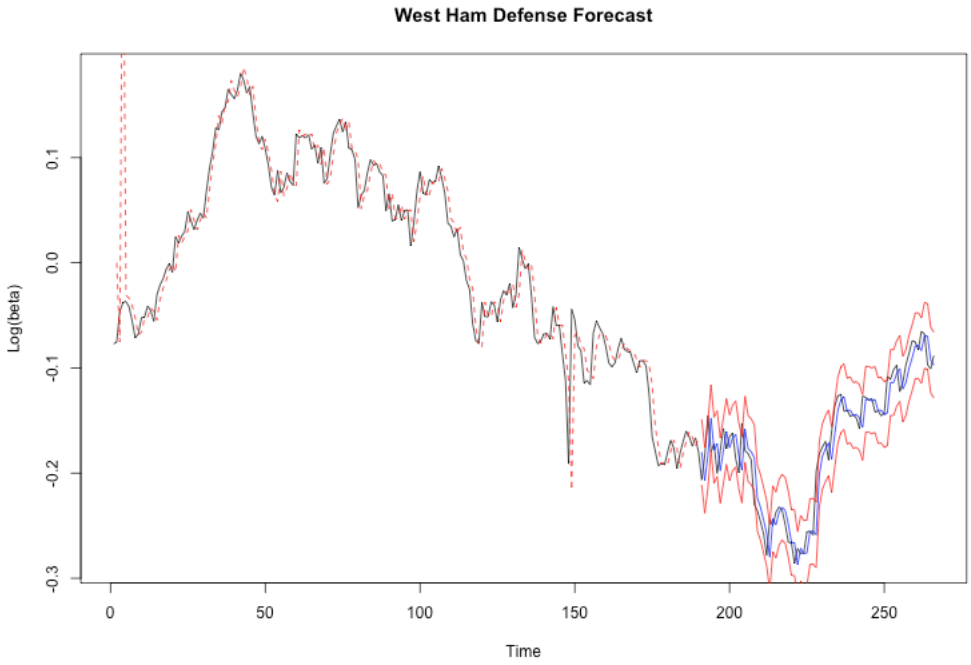
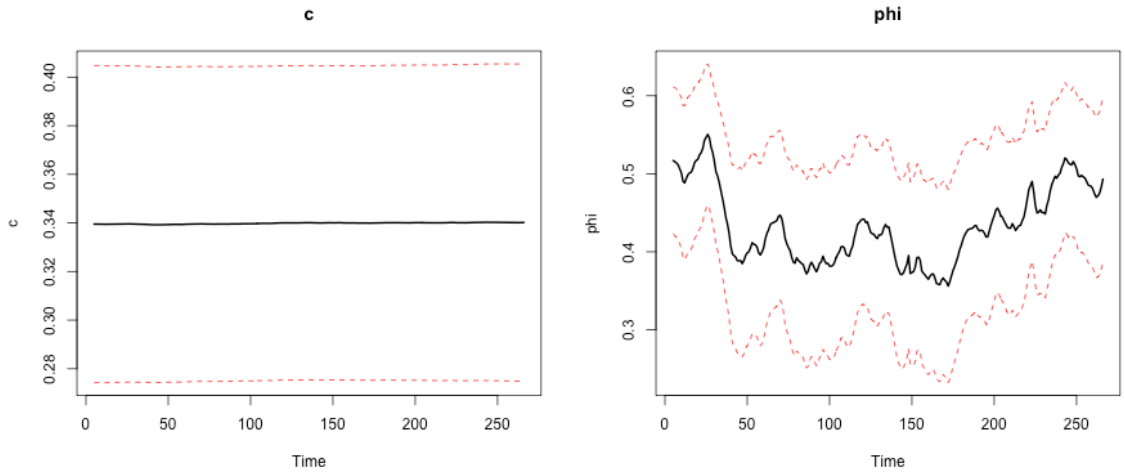


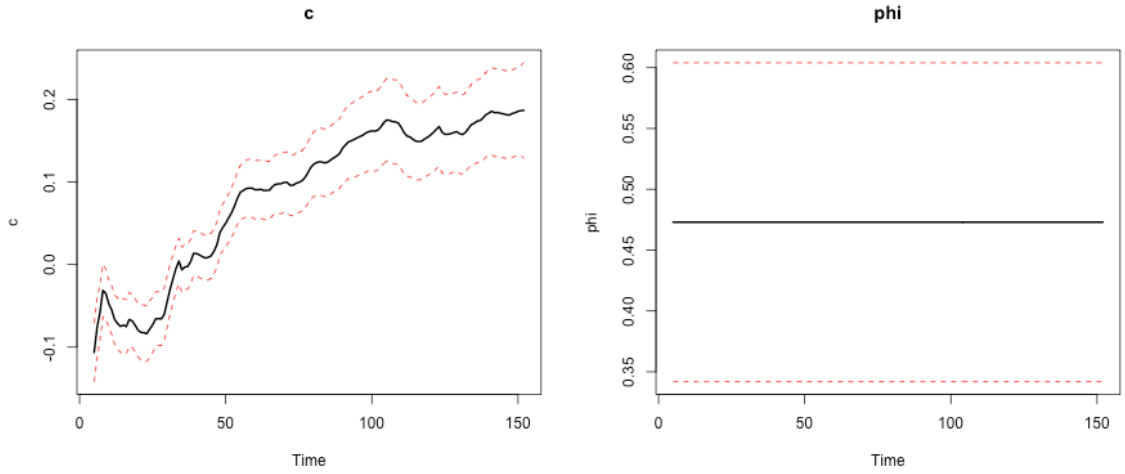
Figure 25 Continued

Appendix E – Smoothed State Estimates – Kalman Smoothing

Arsenal



Bournemouth



Brighton

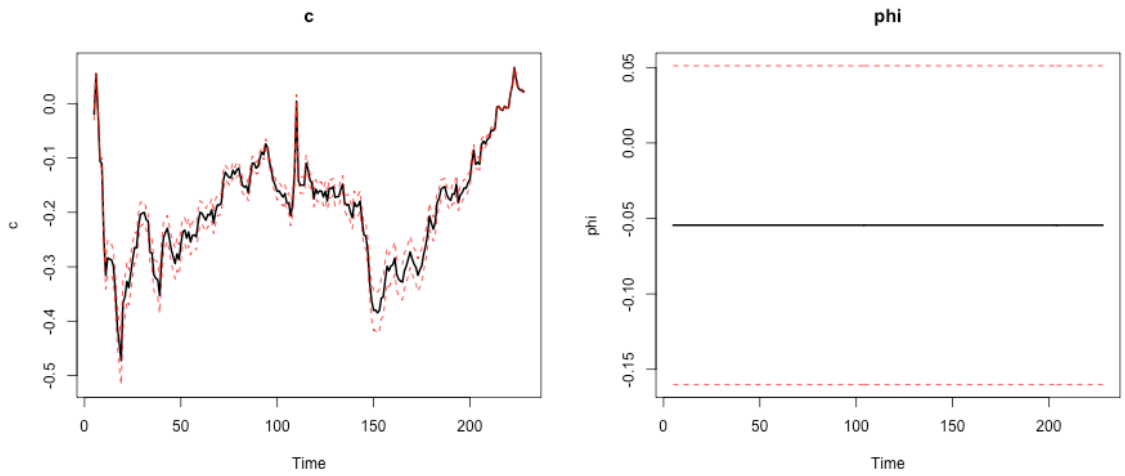
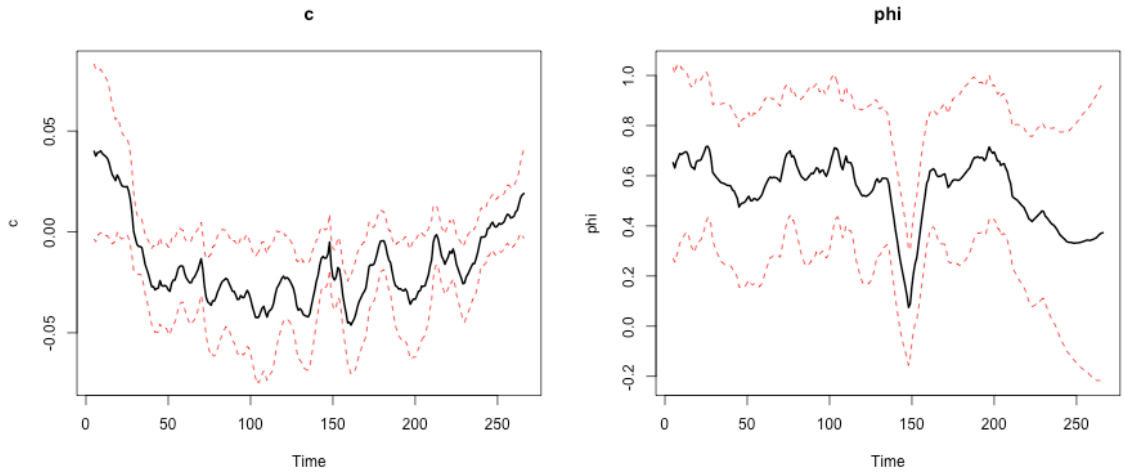
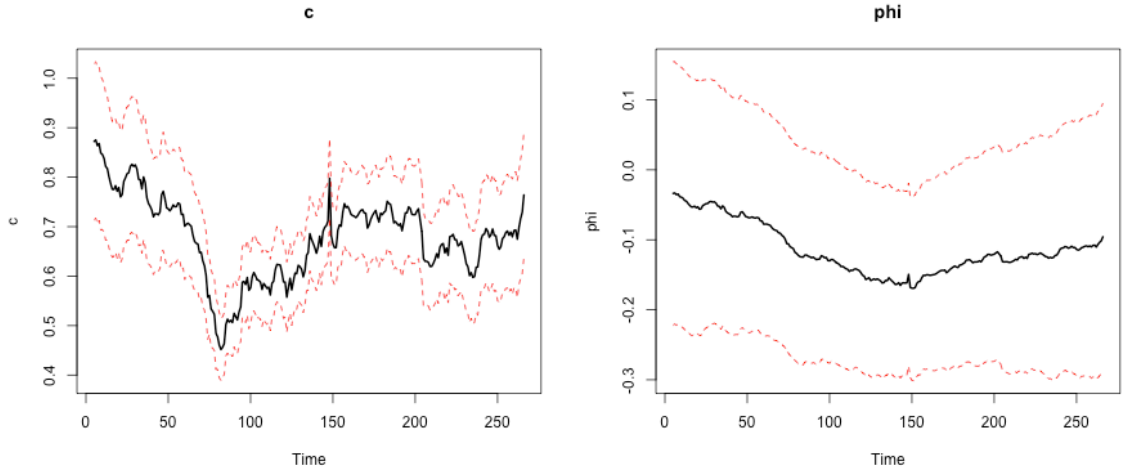


Figure 26 - Kalman Smoothing State Estimates (model 4) – Attack

Burnley



Chelsea



Crystal Palace

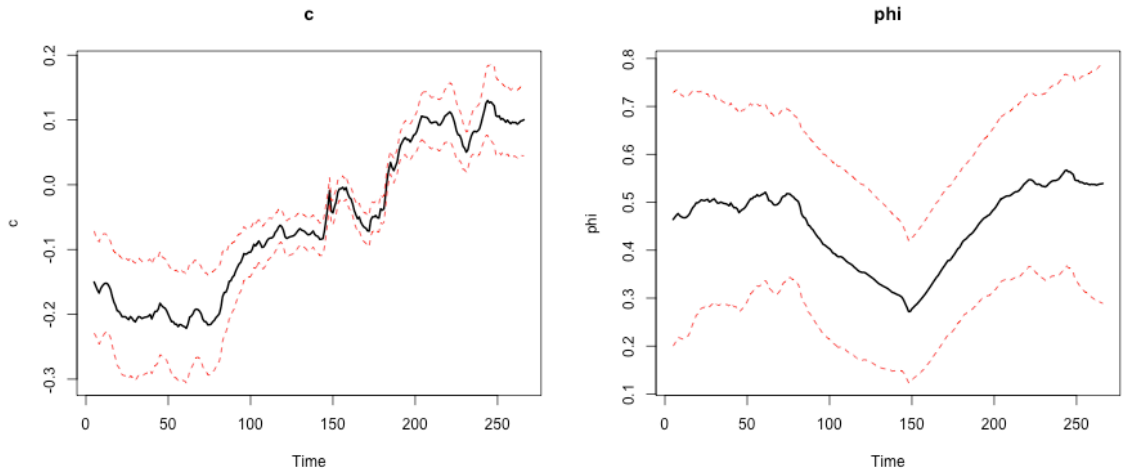
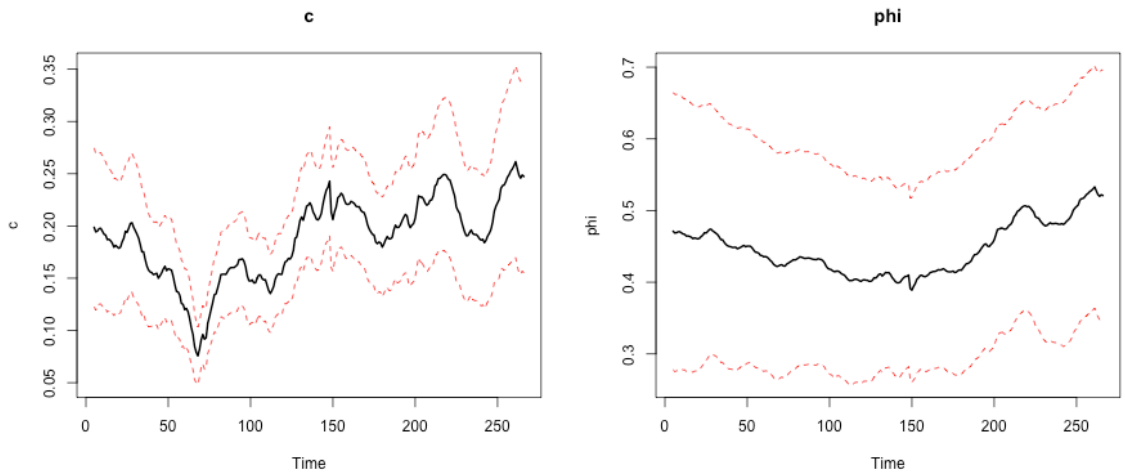
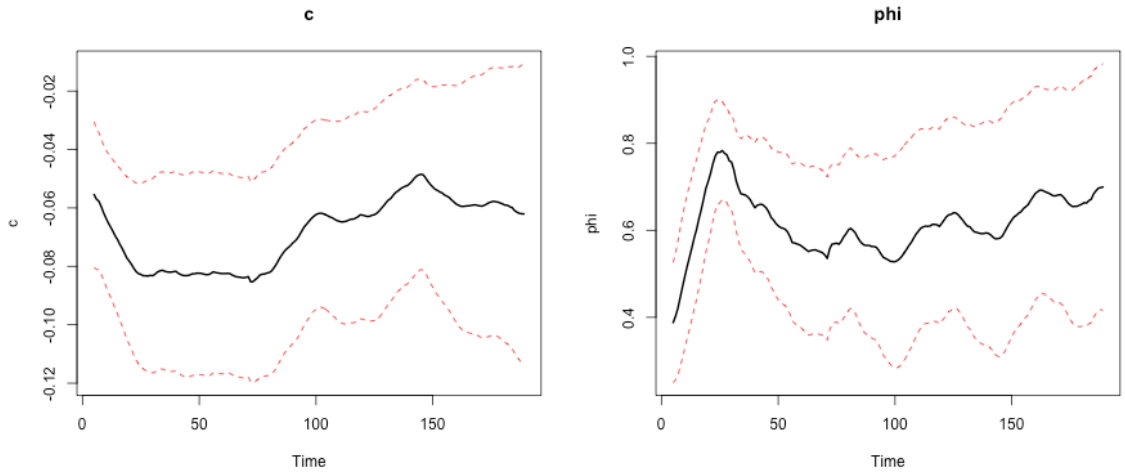


Figure 26 Continued

Everton



Huddersfield



Hull

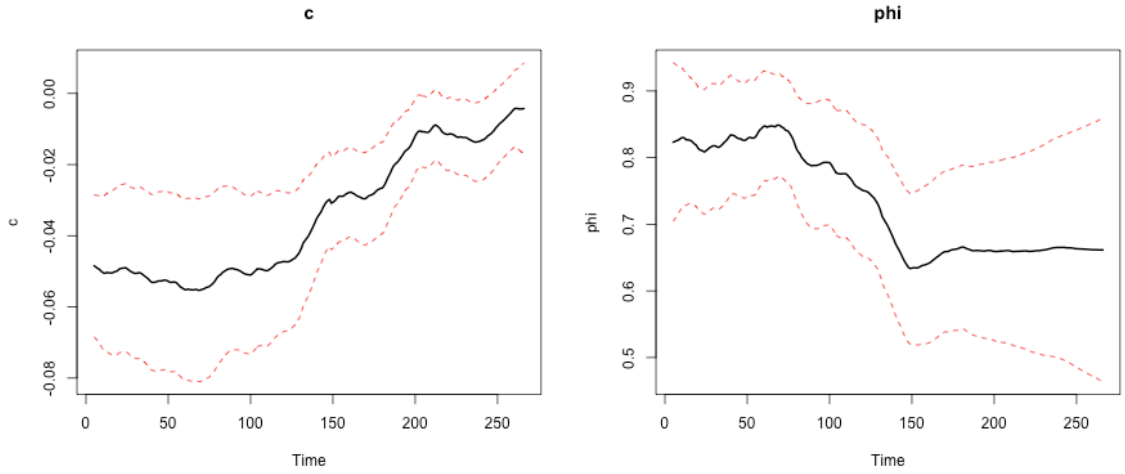
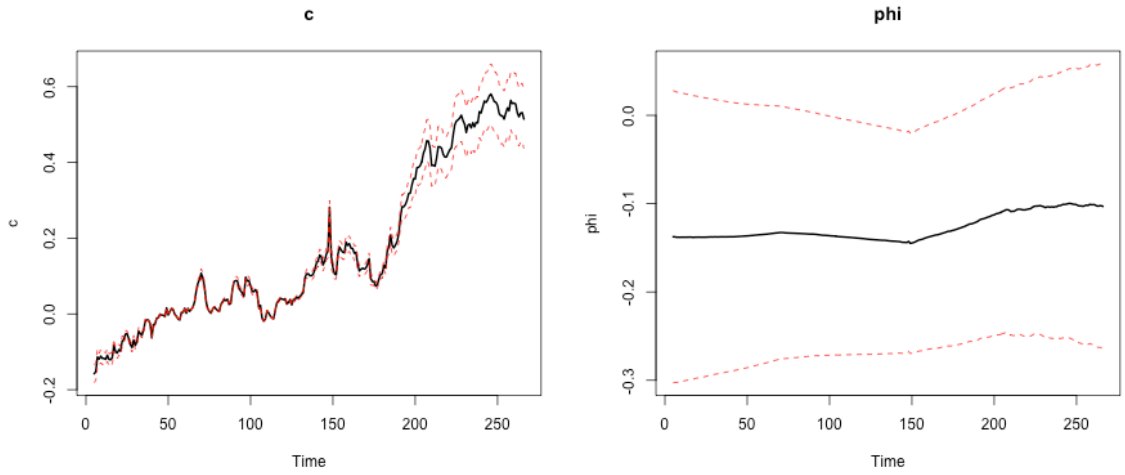
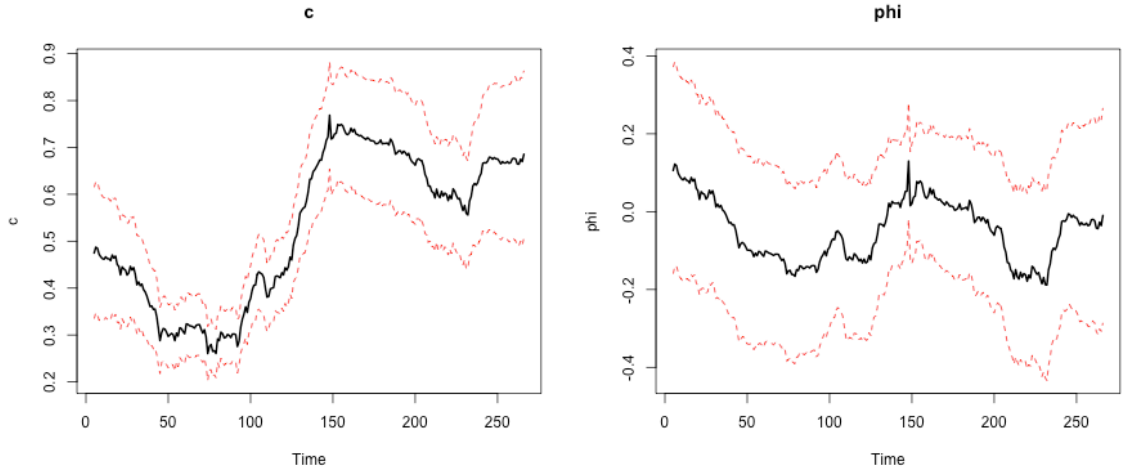


Figure 26 Continued

Leicester



Liverpool



Manchester City

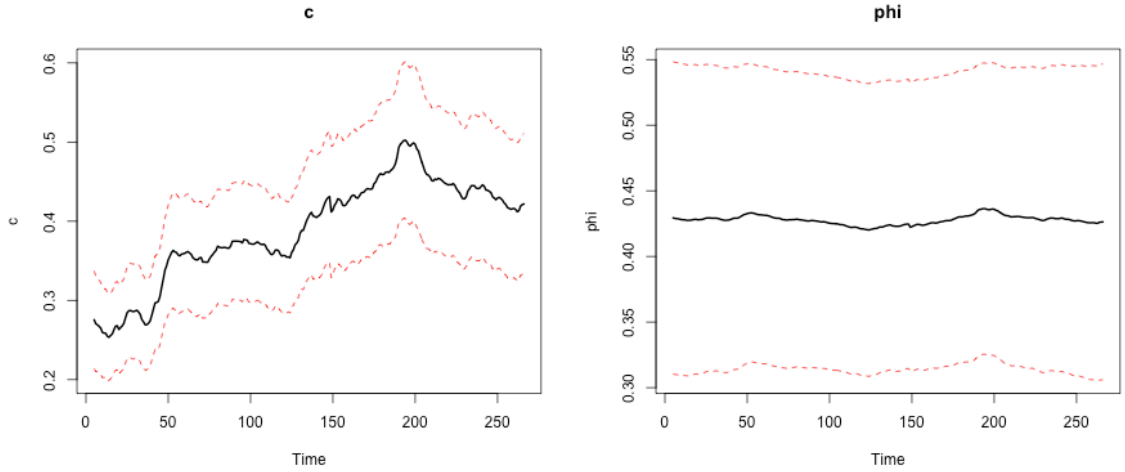
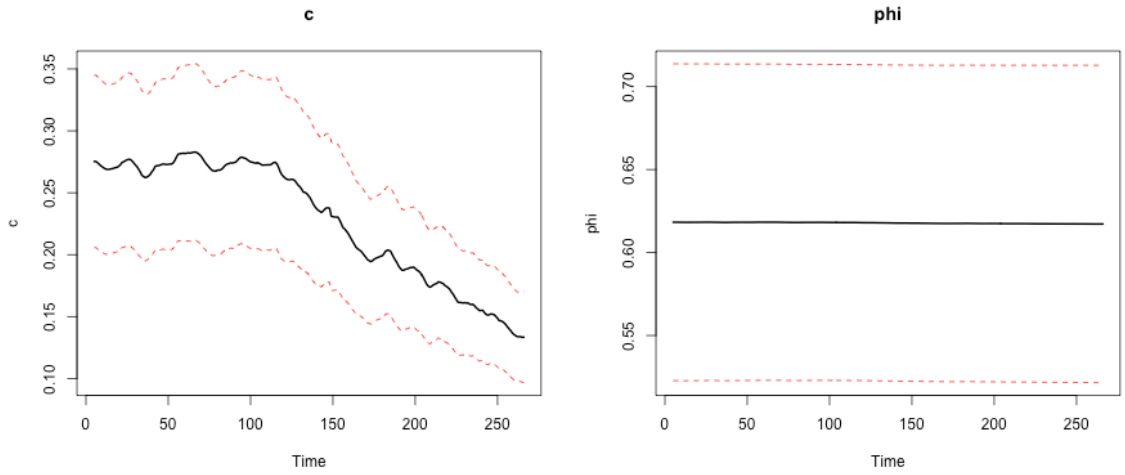
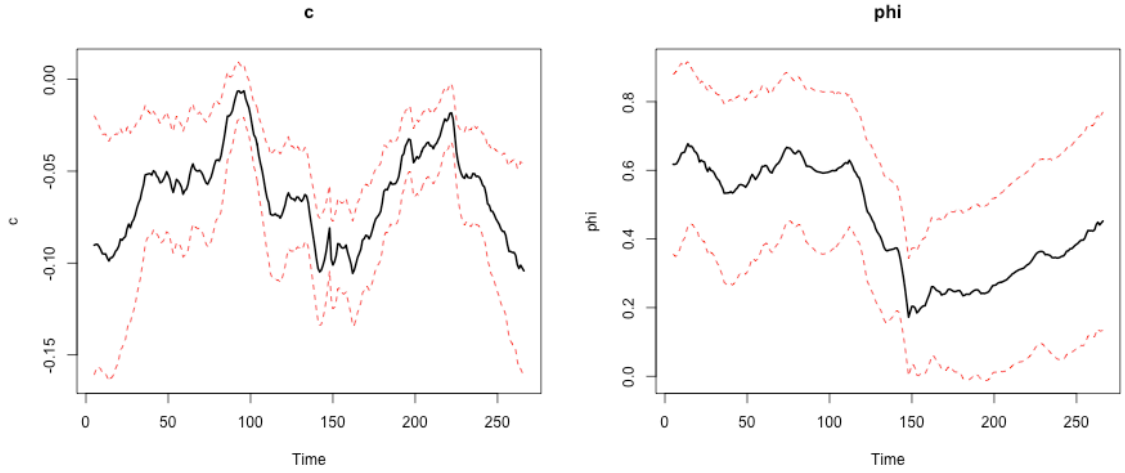


Figure 26 Continued

Manchester United



Middlesbrough



Newcastle

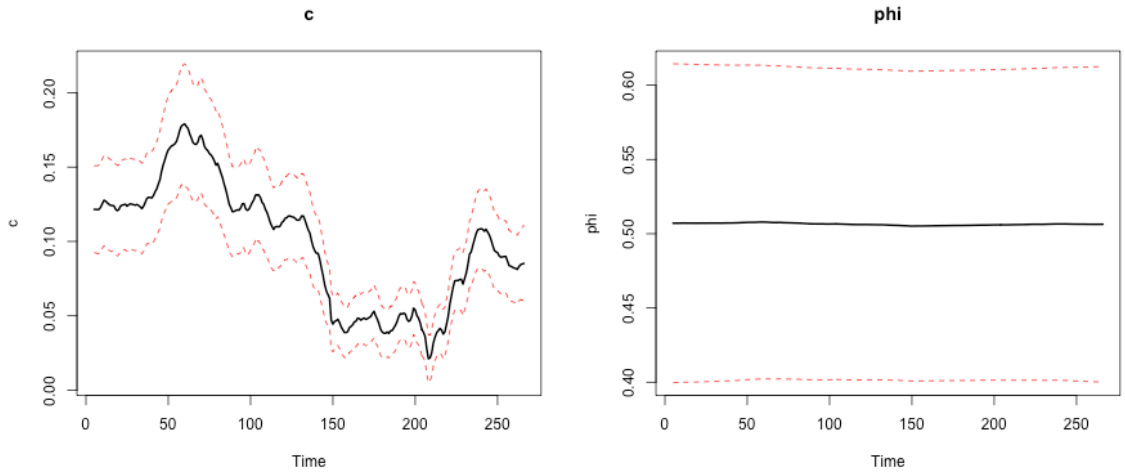
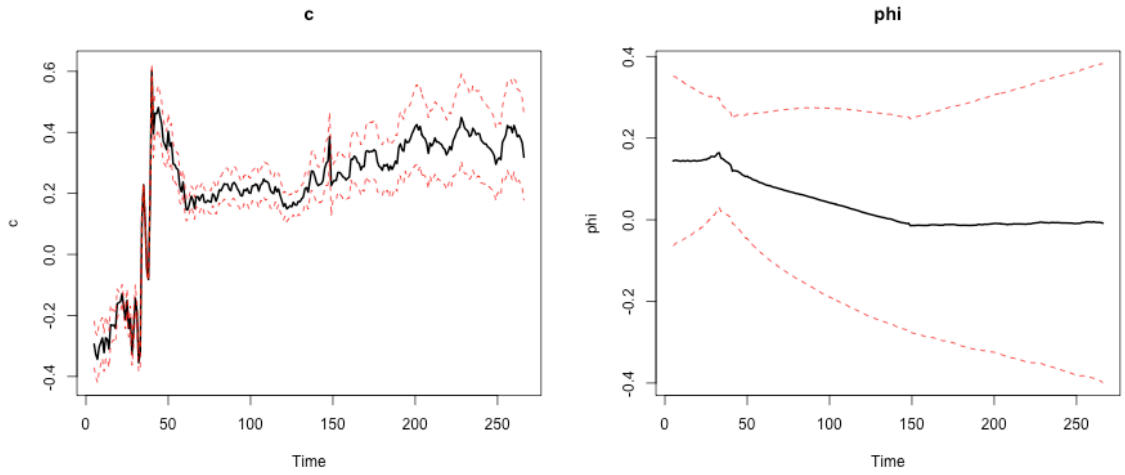
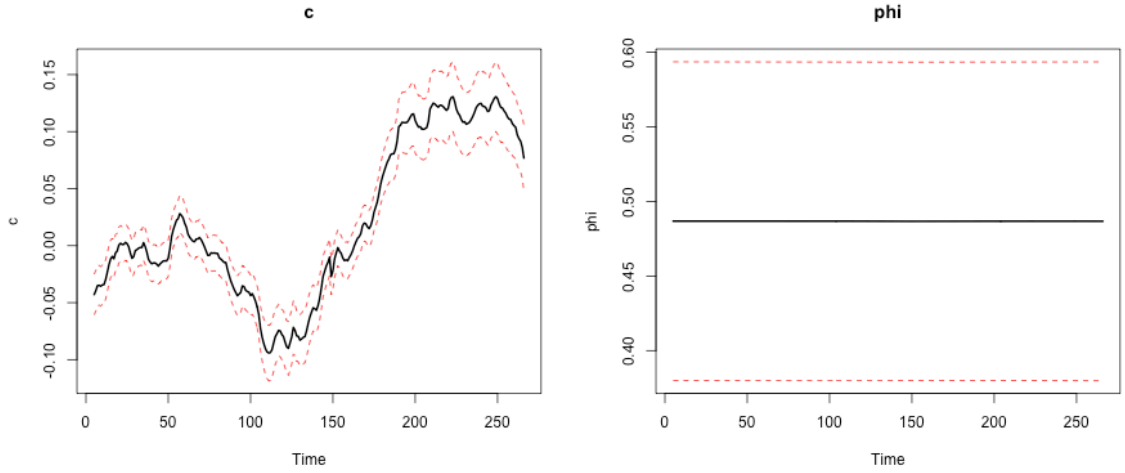


Figure 26 Continued

Southampton



Stoke



Sunderland

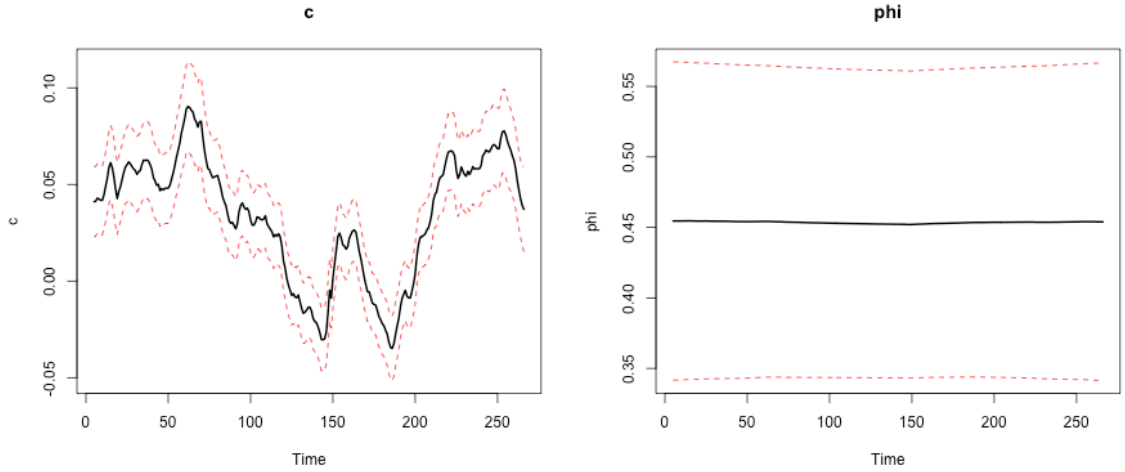
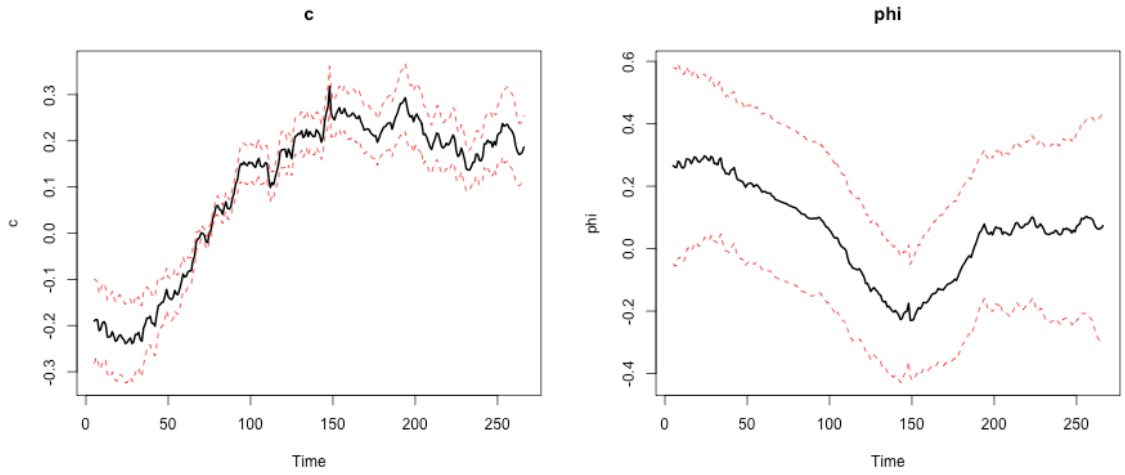
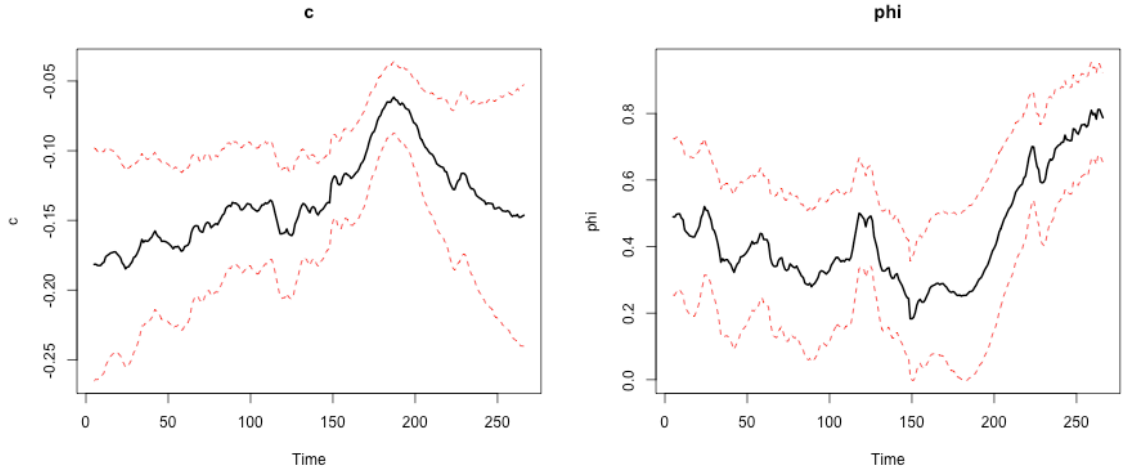


Figure 26 Continued

Swansea



Tottenham



Watford

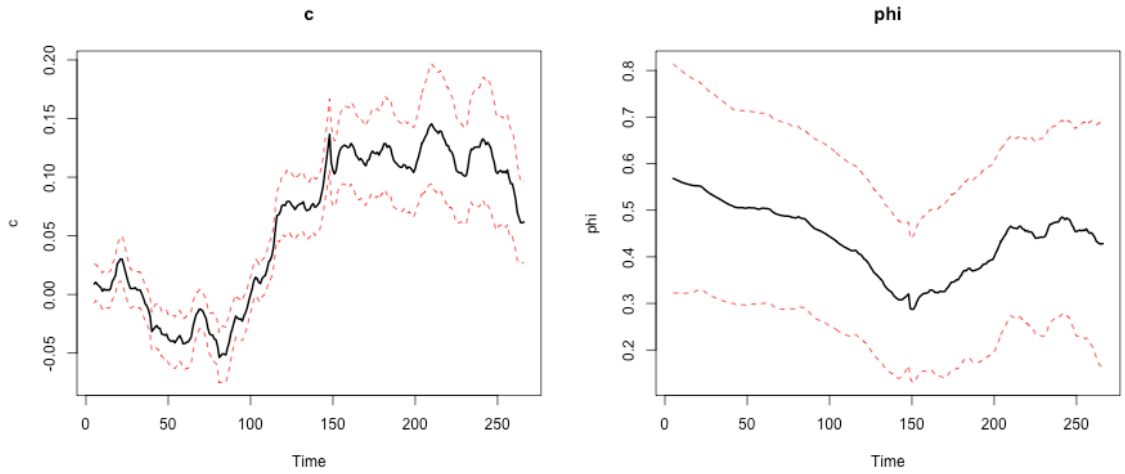
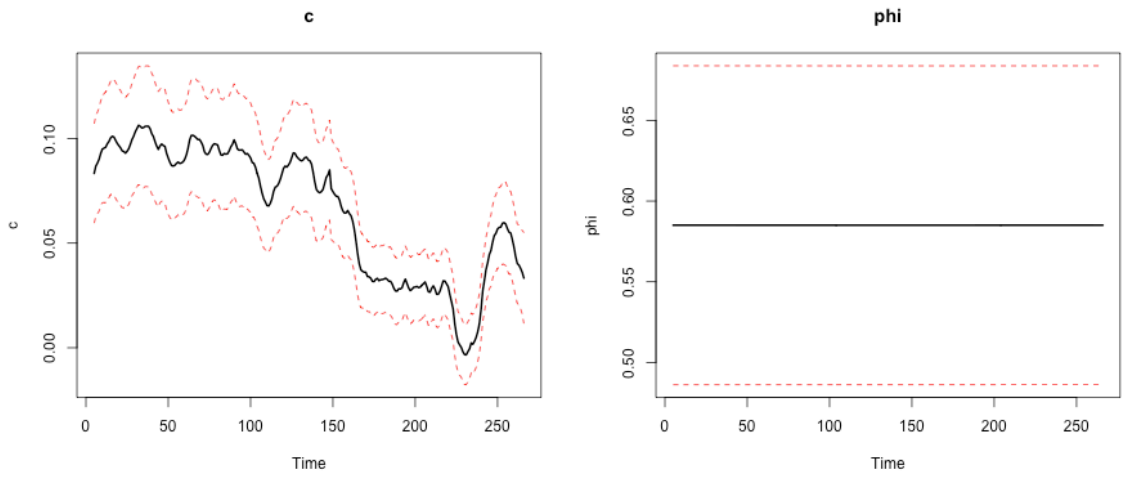


Figure 26 Continued

West Brom



West Ham

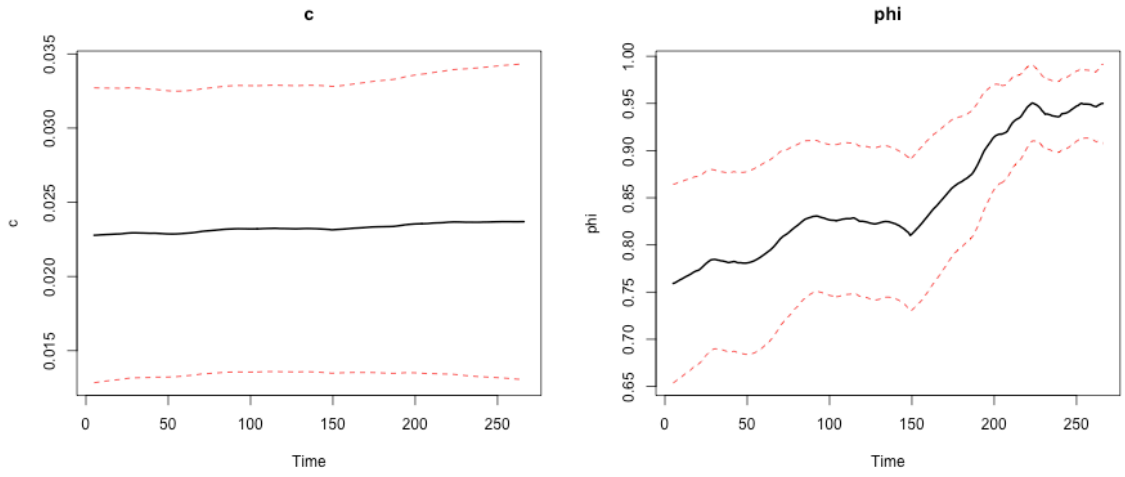
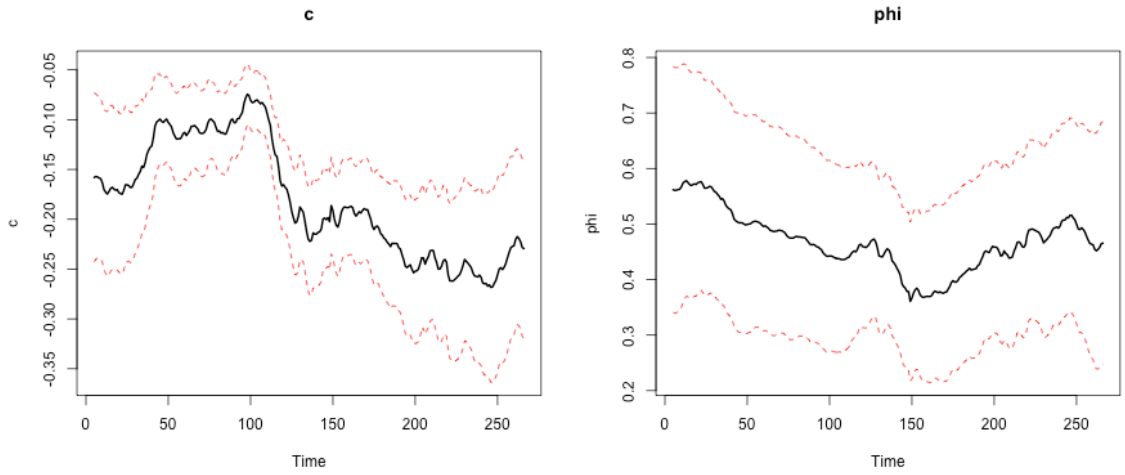
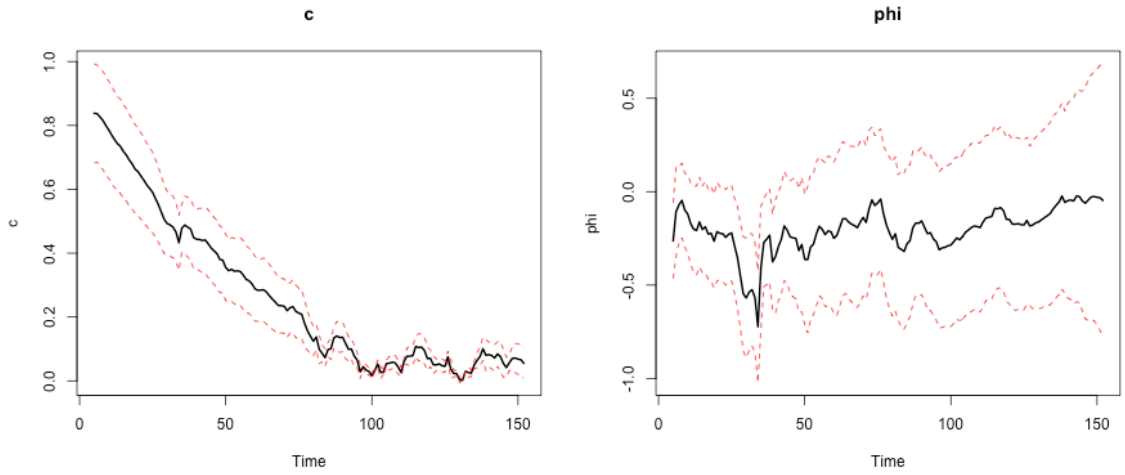


Figure 26 Continued

Arsenal



Bournemouth



Brighton

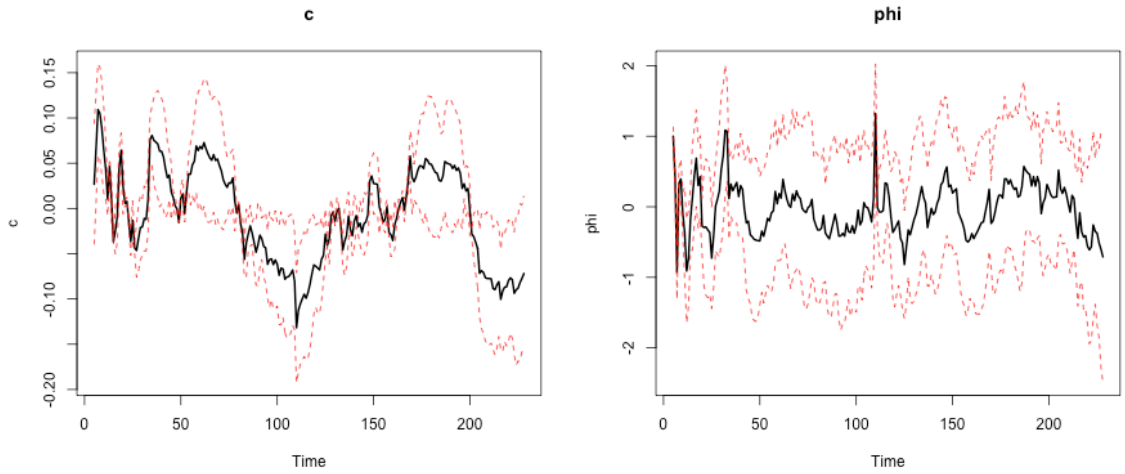
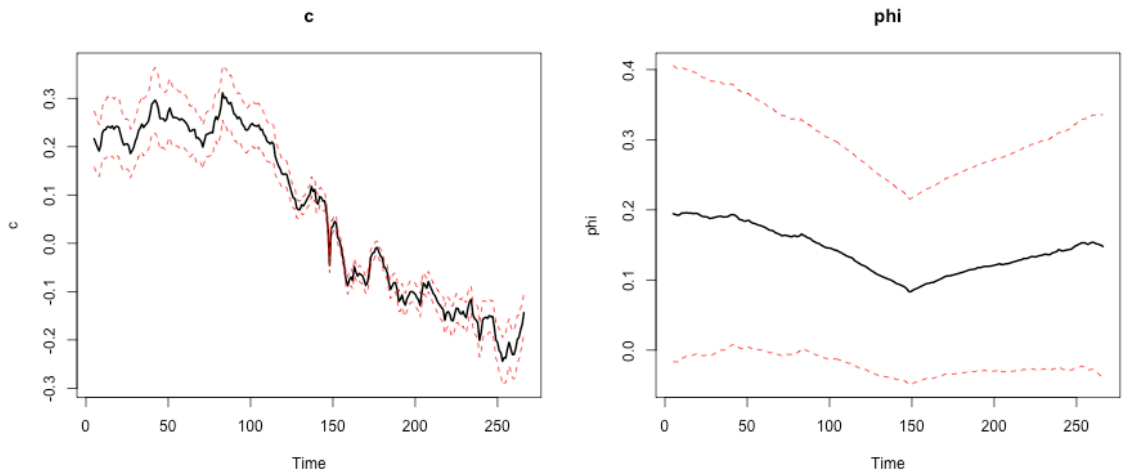
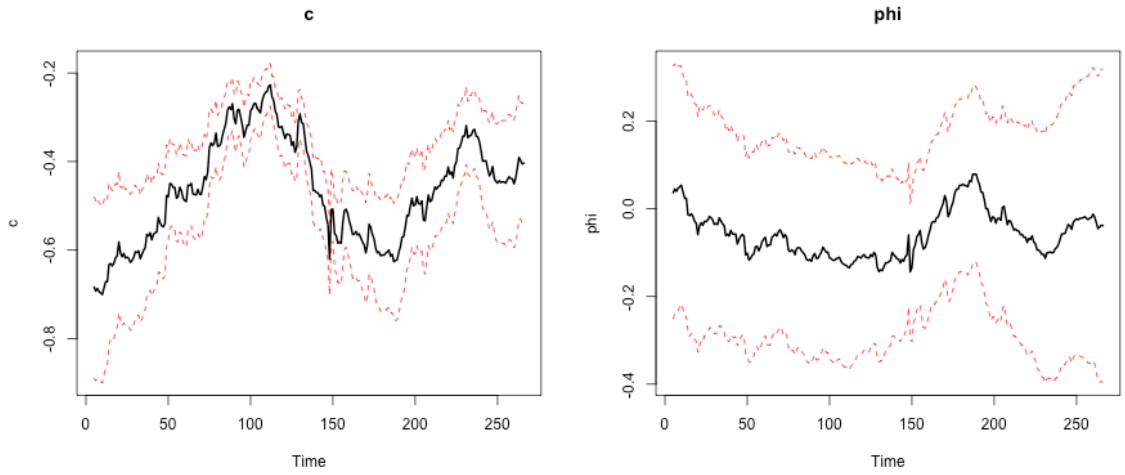


Figure 27 - Kalman Smoothing State Estimates (model 4) – Defence

Burnley



Chelsea



Crystal Palace

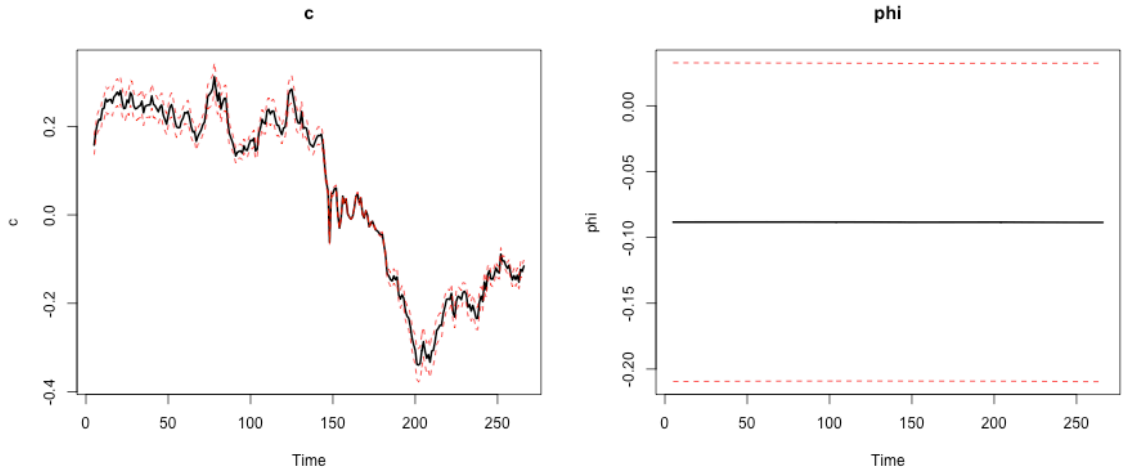
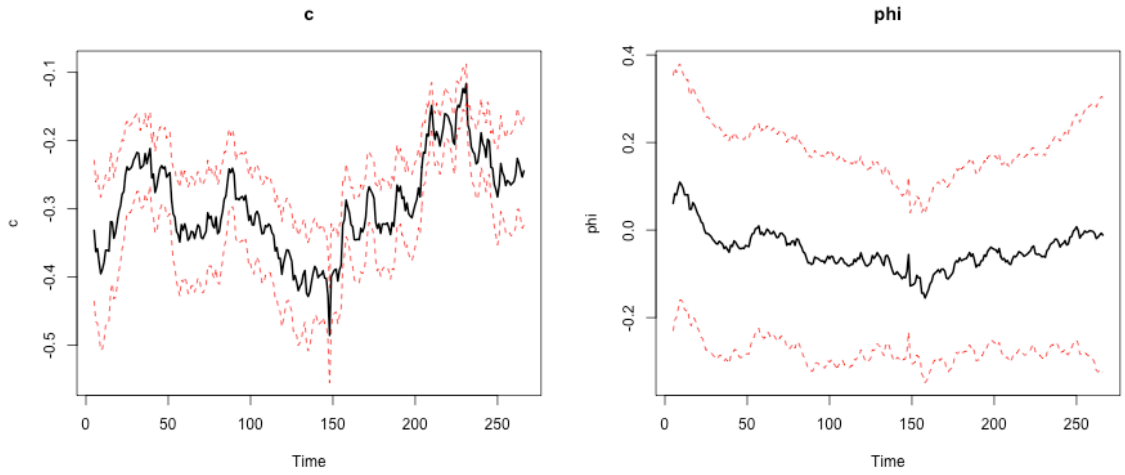
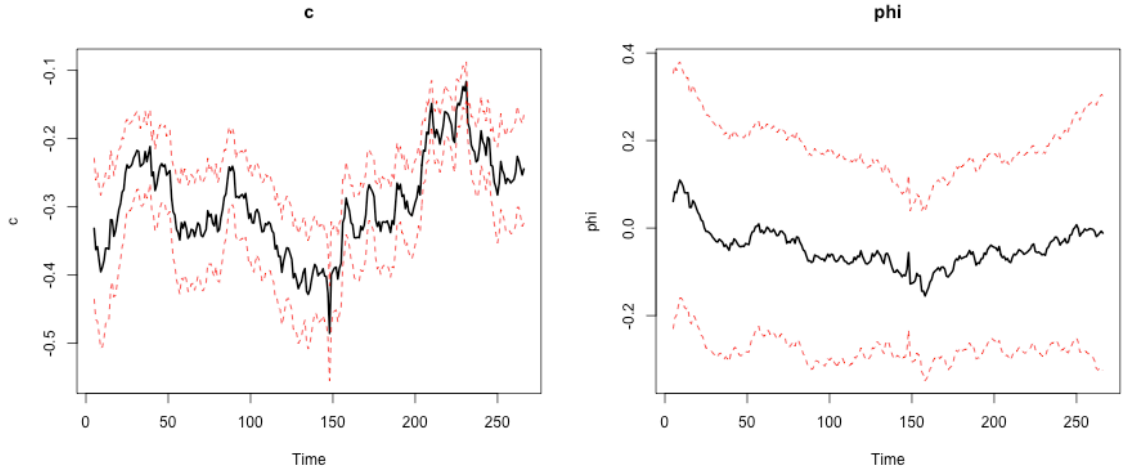


Figure 27 Continued

Everton



Huddersfield



Hull

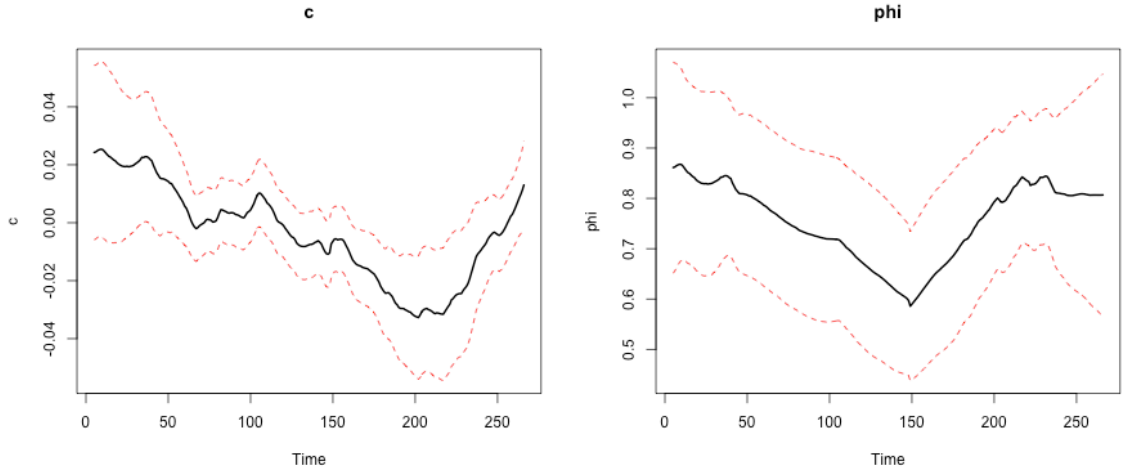
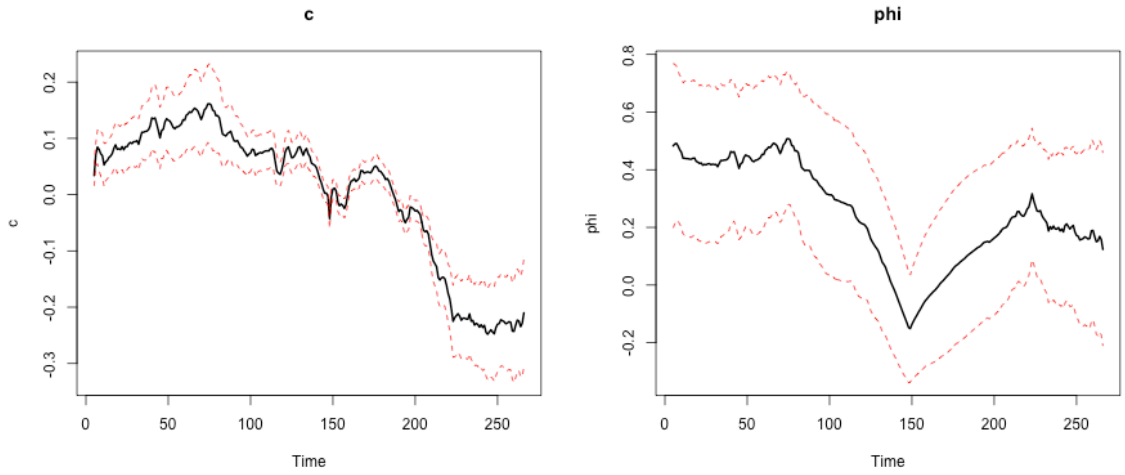
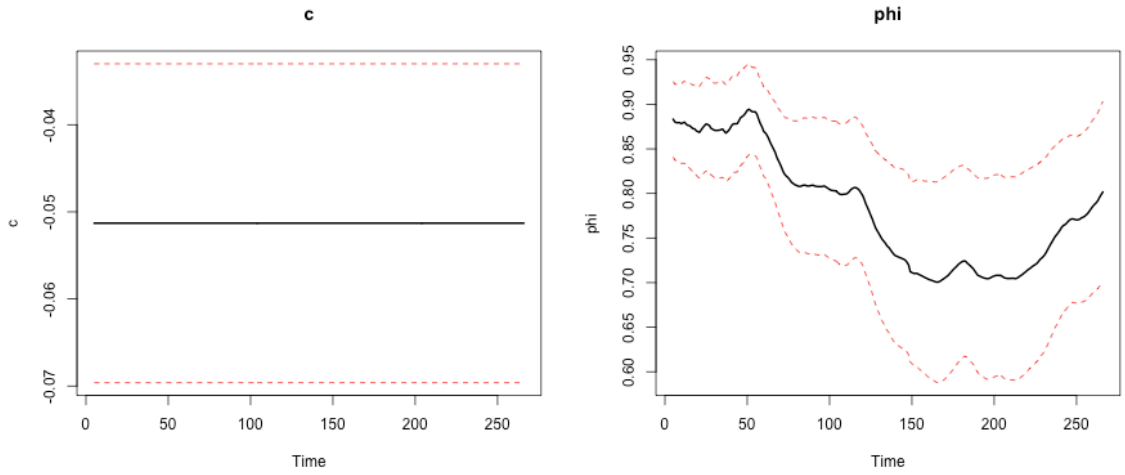


Figure 27 Continued

Leicester



Liverpool



Manchester City

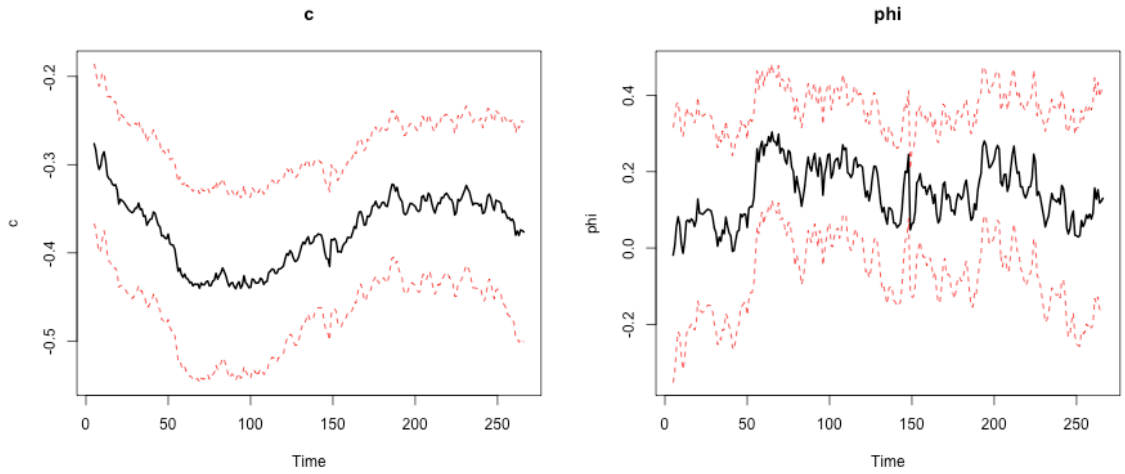
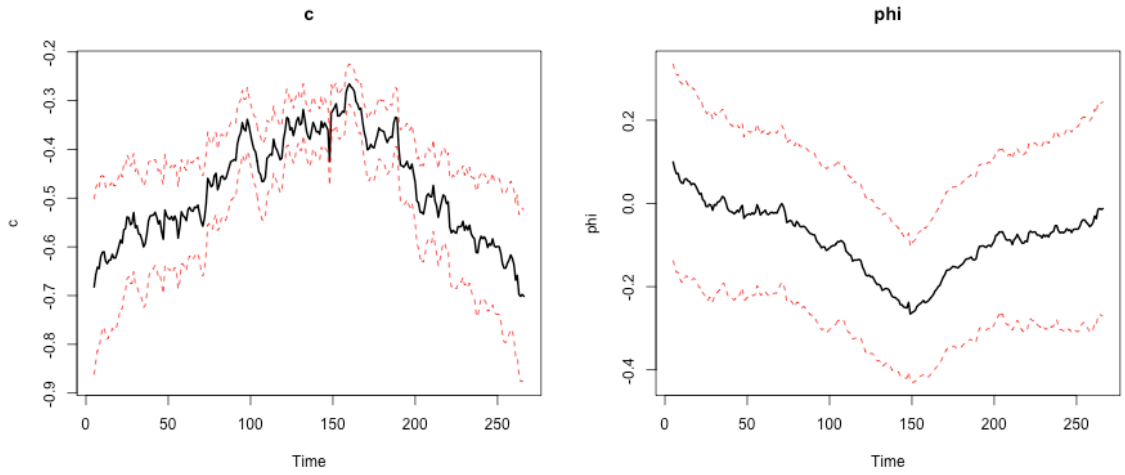
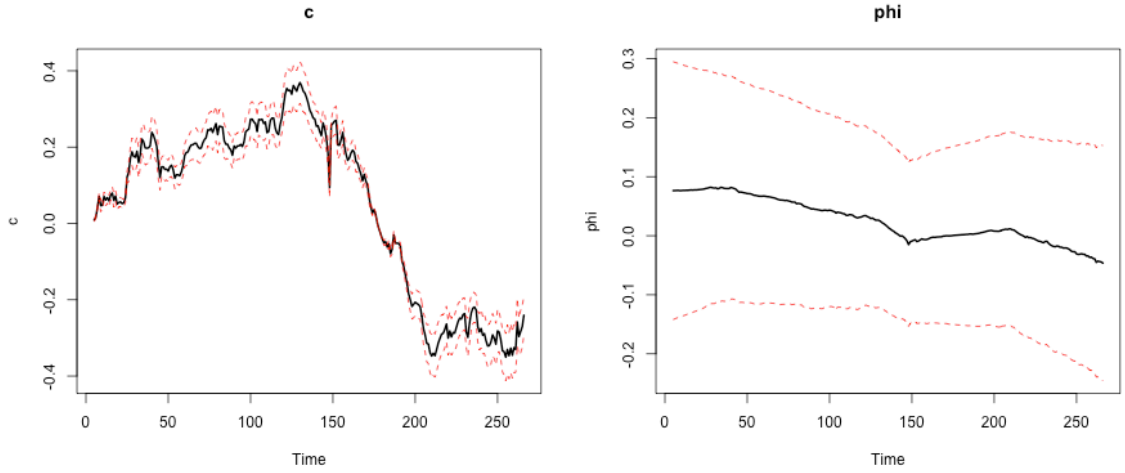


Figure 27 Continued

Manchester United



Middlesbrough



Newcastle

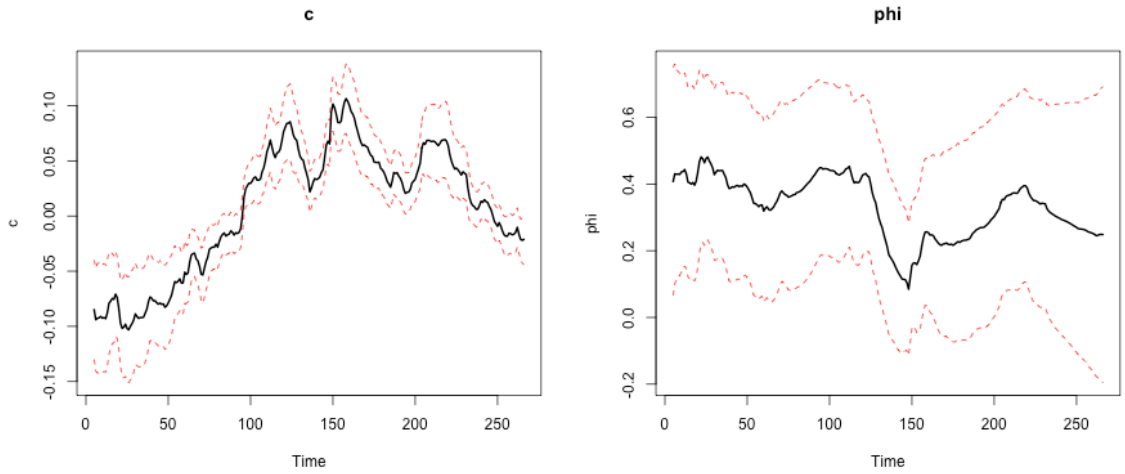
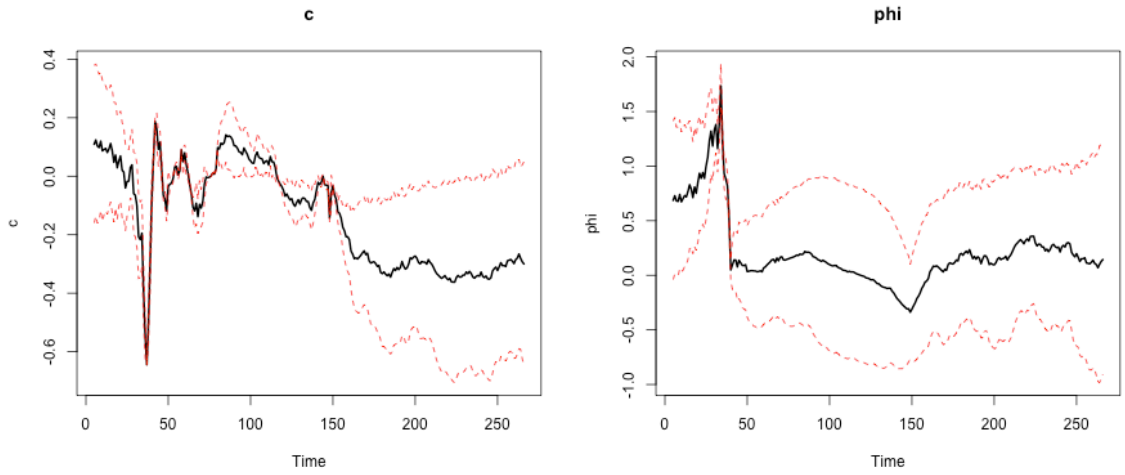
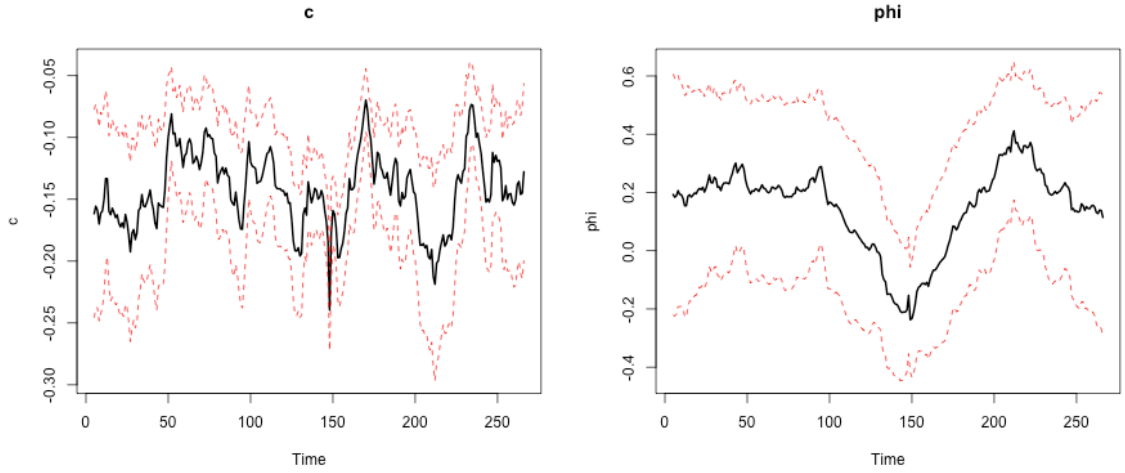


Figure 27 Continued

Southampton



Stoke



Sunderland

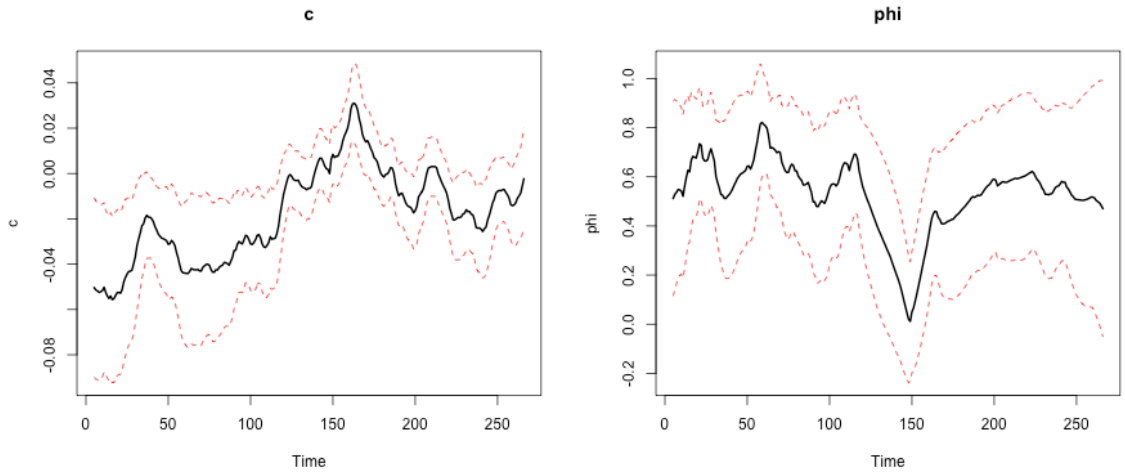
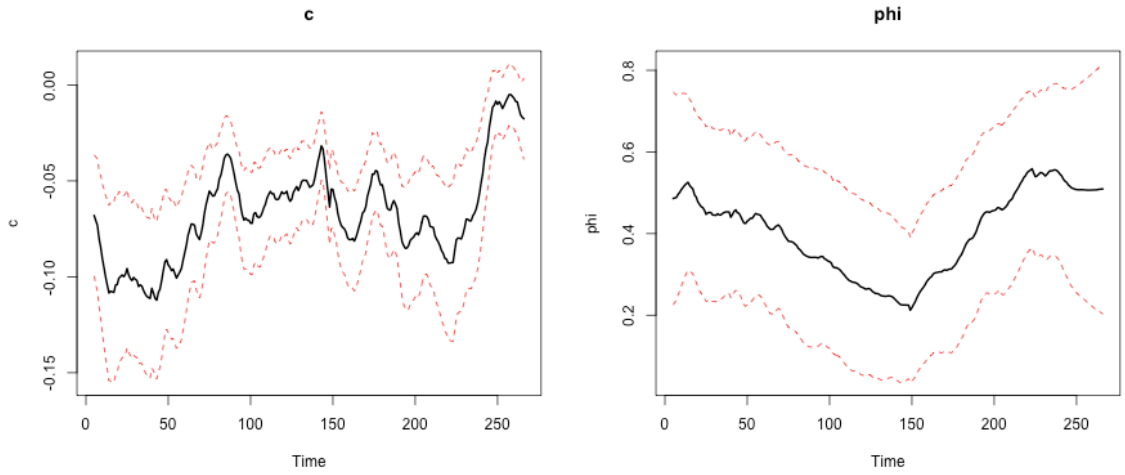
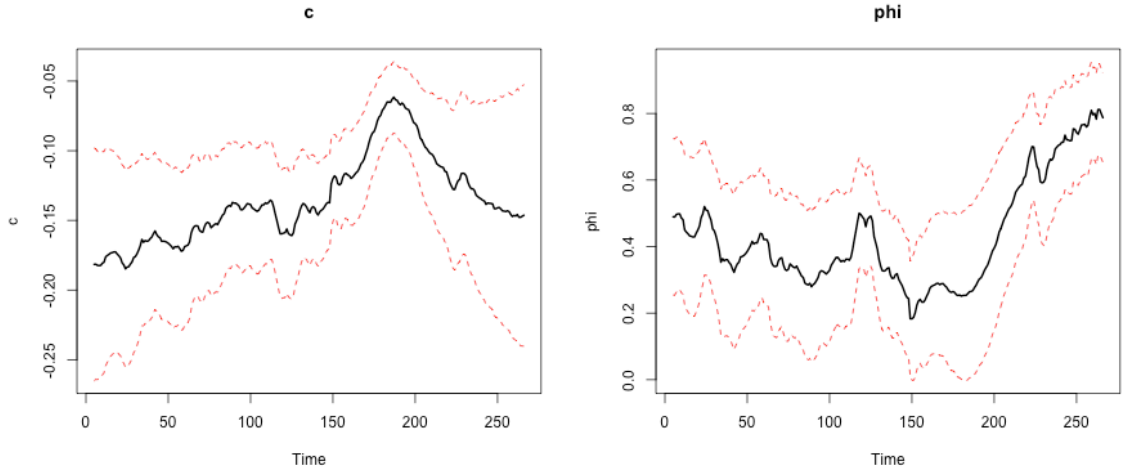


Figure 27 Continued

Swansea



Tottenham



Watford

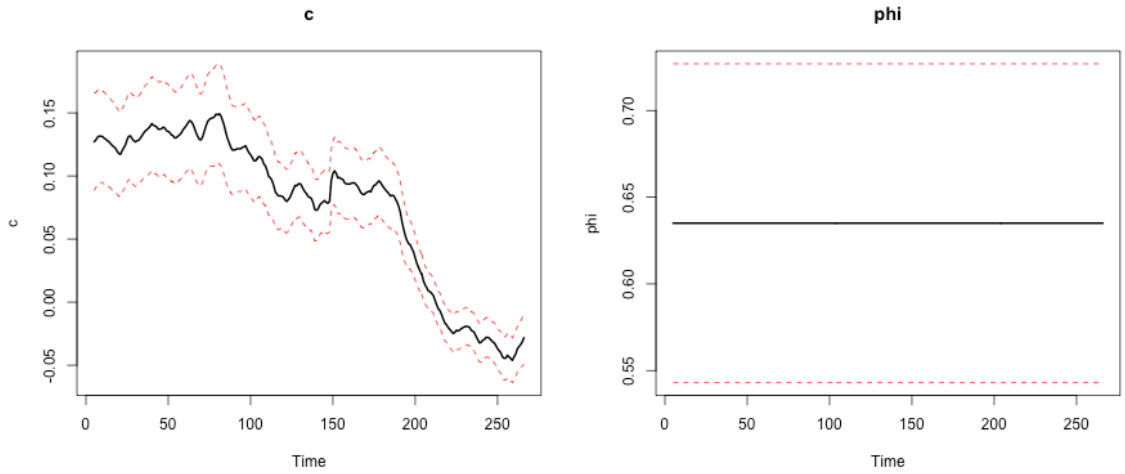
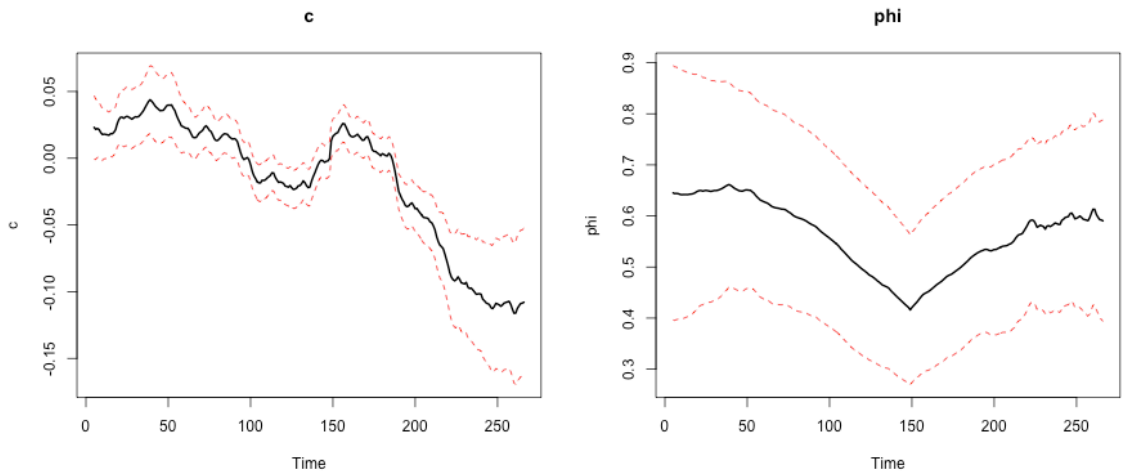


Figure 27 Continued

West Brom



West Ham

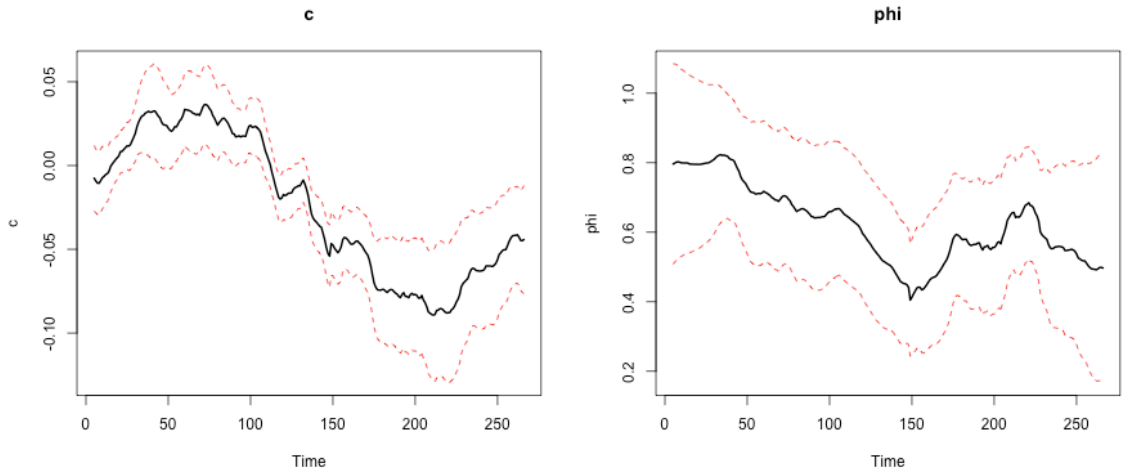


Figure 27 Continued