A statistical analysis of *Escherichia coli* concentrations in the Pakistani Kabul river

MSc Thesis in Environmental Systems Analysis

Author: N.W. Koelen, 871122452070

Supervisors: Dr. Ir. N. Hofstra, Dr. E.J. Bakker

Examinors: Dr. Ir. N. Hofstra, Prof. Dr. R. Leemans

Course: ESA-80436

Version: 1

Submission date: 29-8-2019

A statistical analysis of *Escherichia coli* concentrations in the Pakistani Kabul river

MSc Thesis in Environmental Systems Analysis

Author: N.W. Koelen, 871122452070

Supervisors: Dr. Ir. N. Hofstra, Dr. E.J. Bakker

Examinors: Dr. Ir. N. Hofstra, Prof. Dr. R. Leemans

Course: ESA-80436

Version: 1

Submission date: 29-8-2019

Table of contents

Summary

Escherichia coli is a commonly used indicator for release of faecal matter into surface waters, associated with illnesses such as diarrhea. Iqbal (2017) sampled the Pakistani stretch of the Kabul river for *E. coli* concentrations and several environmental and hydro-climatic variables from April 2013 until October 2015. His research found *E. coli* concentrations to rise along with river discharge in springtime during the wet season, remain high throughout the summer and decrease along with water levels in autumn. Initial statistical of the data series by Iqbal (2017) using a general linear regression model revealed significant correlations ($p < 0.05$) with water temperature, precipitation and discharge. The resulting model had an adjusted R^2 of 0.61.

Closer examination of the data set revealed an annually repeated secondary peak in *E. coli* in the autumn as discharge and precipitation continued to decrease. This peak was consistent with other data sets of *E. coli* in flowing water systems, which indicates this is an actual phenomenon that occurs beyond coincidence. My research aims to explain this peak, and generate a better understanding of the behaviour of *E. coli* in the environment

An Auto Regressive Integrated Moving Average (ARIMA) model was developed to model the observed fluctuations in *E. coli*. Although the data series was too short for a seasonal ARIMA model needed to analyse the secondary peak, a series of shorter lagged (1-5 days) ARIMA models could be fitted to the data set, but these showed no expected auto correlation. Irregular intervals caused a large fraction of the data set to be discarded.

In addition I performed literature research which aims to find an explanation for the secondary peak. This resulted in a conceptual model driven by rainfall induced processes. The dominant factors are human and livestock sources, and resuspension of sediments with attached *E. coli*.

A lagged regression approach was then applied to find a statistically explain the secondary peak. This goal was broadened to fitting a model to the entire data set due to an insufficient number of observations around the peak. Significant correlations were found with water temperature and conductivity at t-3 and acidity at t-0 ($p < 0.05$). The resulting model has an adjusted R^2 of 0.756.

The observed secondary peak was not modelled statistically and found correlations conflicted with what would be expected from literature. Therefore the question as to what causes the observed *E. coli* behaviour remains open. For future research collecting long and evenly spaced data series is recommended. Researchers should consider the possibility of repetitive behaviour such as the secondary peak and the field conditions that are associated with this behaviour.

1 Introduction

1.1 Background

Faecal contamination of surface waters

Safe sources of fresh water are essential for human health. In many countries, the availability of clean water is unfortunately not a given. This is especially the case in developing countries, where population growth, poor sanitation and increasements in agricultural practices lead to structural faecal contamination of water systems (United Nations, 2015). Waterborne pathogens associated with faecal contamination are a common cause of diseases such as diarrhea, which is one of the main causes of death worldwide (United Nations, 2015; Walker et al., 2013). Among children under the age of 5, diarrhea is the second leading cause of death accounting for one in nine mortalities (Centre for Disease Control and Prevention, 2015).

Faecal matter from both humans and animals ends up in fresh water through various pathways, commonly associated with agricultural practices (livestock source) and poor or non-existent waste water management (human source). Agricultural pathways include direct contamination by disposal of excess manure and defecating livestock as well as indirect contamination by way of (sub)surface runoff from agricultural lands caused by manure application or livestock shedding (Hofstra, 2011). Human faecal matter pathways include continuous discharge of untreated sewage, sewage overflows and discharge from insufficiently functioning waste water treatment plants. Both the overflow of sewages and the runoff from agricultural land are generally caused by heavy precipitation events and/or subsequent flooding (Atherholt, 1998; Gibson, 1998). Nichols et al. (2009) and Hashizume et al. (2007) have indeed connected floods and diseases to waterborne pathogens.

Analysis of faecal bacteria is a commonly used method of assessing faecal contamination of water systems (Odonkor et al., 2013). *Escherichia coli* are such an indicator bacteria (Odonkor et al., 2013). *E. coli's* only natural habitat are the lower intestines of endothermic organisms, which means that presence in the environment is an indication of contamination with either human or livestock faecal matter. A few newly evolved strains of *E. coli* are known to cause illness (Nataro & Kaper, 1998), but the bacteria are usually not pathogenic themselves. They do however indicate an increased risk of contamination with other faeces related bacteria such as salmonella or hepatitis (Brüssow et al., 2004).

As described above, meteorological and environmental circumstances are important to the release of faecal matter into water systems. Numerous studies have highlighted the relations between factors such as temperature, precipitation, discharge, salinity, land use and others and faecal indicator bacteria concentrations in both surface and drinking water. Vermeulen and Hofstra (2013), for example, used data from a 25 year period to find negative correlation with temperature. This indicates that rising temperatures increase the die-off rate of microbial organisms, and positive correlations with precipitation and discharge which indicates the influx of *E. coli* from runoff and overflows. Walters et al. (2011) found a similarly positive correlation between precipitation and faecal indicator bacteria including *E. coli*, and negative correlation with temperature and salinity. Ge et al. (2010) found a positive effect of wave action on *E. coli* concentration through the resuspension of contaminated sediments. Kay et al. (2005) found significant elevations of faecal indicator organism concentrations during high discharge events in a UK river, and developed a model predicting both mean and high flow concentrations based on land use data. The fraction of built-up land was the dominant independent variable in this model, indicating sewage release as an important contributor once more. Whitman & Nevers (2008) performed a study along the shoreline of the North American great lakes, and found positive correlations with wave height and barometric pressure. Islam et al. (2017) found that precipitation and, unexpectedly, water temperature were positively correlated with the concentration of faecal indicator bacteria in Bangladesh,

attributing the coincidence of heavy precipitation and the warm weather season to temperature correlation.

The case of the Kabul river

The Kabul river is one of the largest tributaries to the Indus river. The Indus originates on the Tibetan plateau in the Himalayan Karakorum region and flows down to its estuary in the Arabian sea near Karachi. The Kabul itself originates in the Hindukush mountain range in Afghanistan, stretching over 700 kilometres as it enters Pakistan near Torkham before its confluence with the Indus at Attock Khurd, see Figure 1.1.1. The Pakistani stretch is dammed near the border with Afghanistan, with an irrigation canal running from the reservoir towards the city of Peshawar and then back into the main river at Shah Alam. Downstream of the dam the river divides into three branches which re-join near the city of Charsadda, which is also the confluence point with the Swat river, a major tributary.

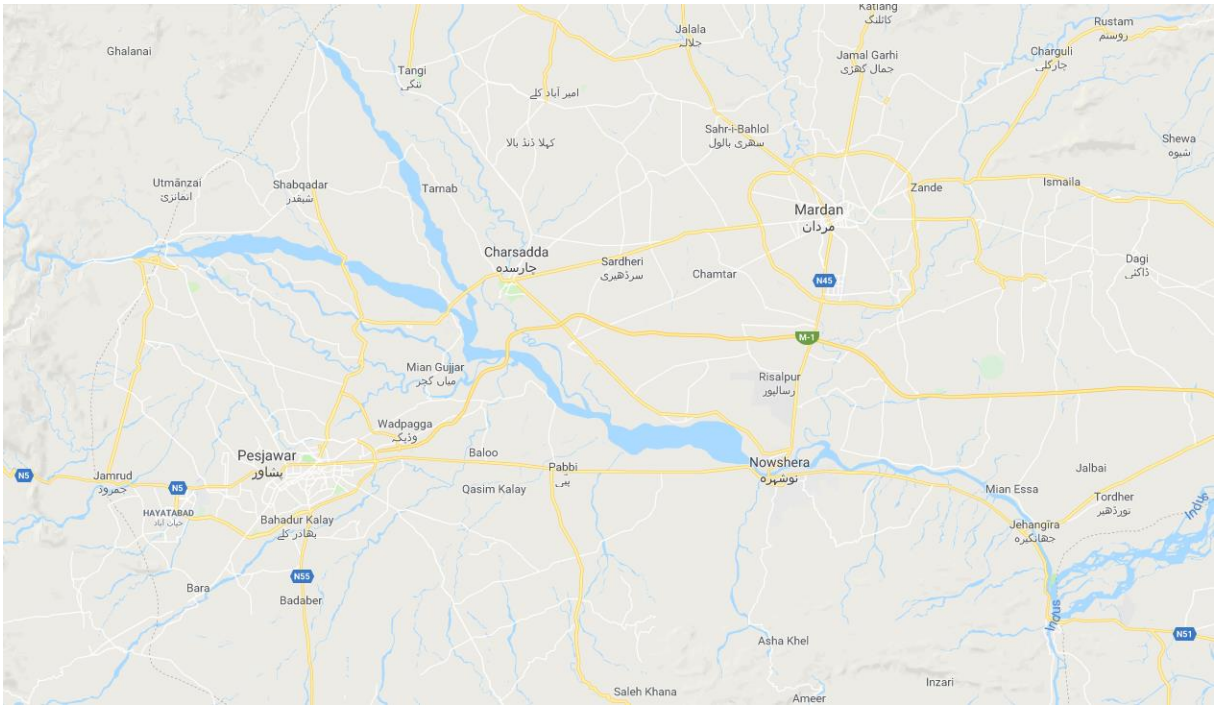

Figure 1.1.1, The Kabul river basin between the border crossing south of Utmanzai and the confluence with the Indus river near Jehangira. Note the division into separate channels between the border and Mian Gujjar. The Swat river tributary is located to the west of Charsadda (Google Maps, 2018)

The Pakistani part of the basin is prone to seasonal flooding due to a combination of heavy monsoon precipitation and meltwater from the Hindukush mountain range. River discharge varies widely throughout the year, see Figure 1.1.2. A large fraction of the area is heavily cultivated for both agricultural purposes as well as urbanization (Iqbal & Hofstra, 2019). The waste water treatment plant near the city of Peshawar was destroyed during extreme floods in 2010 and has been defunct since. As a result surface water quality is generally poor. *E. coli* standards are exceeded year-round, which makes the river unsuitable for bathing or swimming (Iqbal & Hofstra, 2019).

Figure 1.1.2, Differences between minimum discharge in December 2016 and maximum discharge in August 2017 (Sukkar, 2016; Hussain, 2017)

From 2013 to 2015, a large set of data was obtained from the entire Pakistani stretch of the river. Biweekly analysis was done on *E. coli* concentrations, along with measurements of turbidity, conductivity, water temperature and pH. At the same time, data regarding water temperature, precipitation and discharge were collected. Regression analysis showed a positive correlation of *E. coli* with discharge, temperature and precipitation, consistent with earlier studies. Correlation with discharge was especially strong (Iqbal, 2017), see also Figure 1.1.3.

When plotted, the dataset shows some behaviour which cannot directly be attributed to known linear relations. Firstly, *E. coli* concentrations reach their maximum levels before discharge does. A 'first flush' effect from sewages and land runoff could be at work here, as *E. coli's* relations with first flush effects are described in literature, although correlations are generally weak (McCarthy, 2009). Secondly, when water levels subside after the flooding season is over a distinct second peak in *E. coli* concentrations is observed annually. This peak is also seen in other studies (Wilkinson et al., 2006) with very different temporal resolutions.

Wilkinson et al. (2006) studied a hydrograph in a tidal river in the United Kingdom over a course of several hours, where after the major flood wave event a secondary peak, similar to the one seen in Iqbal (2017), can be observed. The authors acknowledge the peak, described as a "pulse", but do not give a proper explanation. A closer analysis of data from Iqbal (2017) shows a slight increase in river discharge prior to the secondary peak, but given the strong correlation with river discharge found in this research it is unlikely to be the sole contributor. Overall, an explanation for the phenomenon is currently absent in literature, and a better understanding of the short term interactions between environmental circumstances and *E. coli* concentrations is needed to help provide answers.

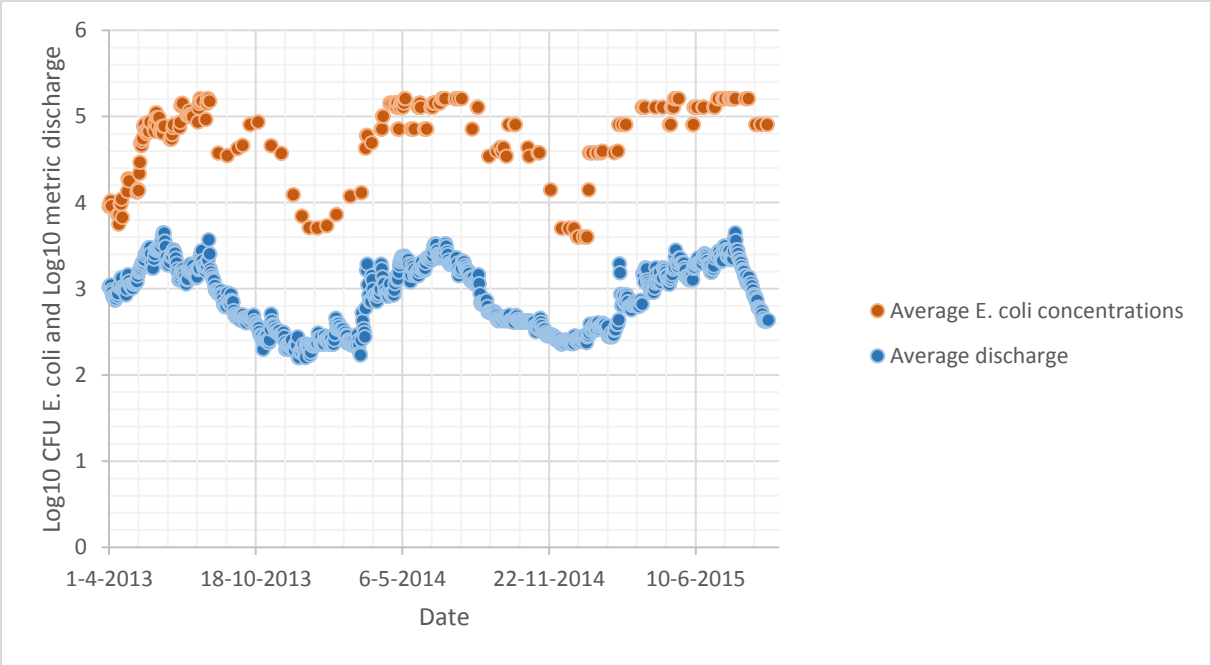

Figure 1.1.3, 30 month graph of E. coli concentrations and discharge in the Kabul river basin. Note the strong relationship between discharge and E. coli concentrations, and also the secondary peak which can be observed in both October '13 and October '14 (Iqbal 2017)

1.2 Problem statement

The presence of *E. coli* in surface waters is a major indicator of contact with faecal matter. *E. coli* concentrations in the Pakistani Kabul river exceed water quality standards year round. Especially during the wet season, when concentrations reach sewage water levels, contact with river water poses a significant risk to human health. *E. coli* concentrations show short term fluctuations, which cannot readily be explained with known linear relations. Exemplary is a secondary peak in concentrations which is observed after major high water events.

Given the risks involved in exposure to surface water that is contaminated with faecal matter, improving the water quality of the Kabul river is urgently needed. Currently, the processes controlling *E. coli* concentrations in the river are insufficiently understood. A better understanding of the relations between environmental conditions and the release of *E. coli* into surface waters is needed as a primary step towards water quality improvement.

1.3 Research objective

Previous studies (*Iqbal, 2017; Wilkinson et al., 2005)* show a distinct secondary peak in *E. coli* concentrations in flowing surface waters as water levels subside after major rainfall, hydrograph or flooding events. This research aims to explain this phenomenon through developing a conceptual model of *E. coli* release into the Kabul river, and through statistical analysis of the Iqbal data set.

The following main research question is addressed in my study:

How can the observed post-high water peak in E. coli concentrations in the Pakistani stretch of the Kabul river basin be explained?

The specific research questions (SRQ's) are:

1. *Are temporal variations in E. coli concentration auto correlated?*

2. *What pathways are likely to contribute to the observed secondary E. coli peak?*

3. *What statistical analysis method is suitable to establish relations between environmental variables and the secondary E. coli peak?*

4. *Can time dependent correlations be shown with:*
   4.1. Physicochemical variables (Water temperature, turbidity, conductivity, acidity)?
   4.2. Meteorological variables (Air temperature, precipitation, discharge)?

The data used in this research were collected by doctor Muhammad Shahid Iqbal in the years 2013 to 2015 (Iqbal, 2017). For his analyses Iqbal used biweekly data, whereas during certain periods far more measurements were taken. This was especially the case during high discharge periods when the phenomena of interest were observed. The full dataset was available for this research.

1.4 Content of report

My report separates the study into three distinctive parts, each with their own introduction>method>results>conclusion structure. I chose this approach since I believe it aides in keeping the diverse subject matter digestible and allow the reader to focus on a specific topic of interest. Firstly the auto correlative statistical analyses are covered, providing some valuable insight into the requirements of auto regressive modelling. Then, a conceptual model is presented that will help gain an understanding of *E. coli* behaviour in the Kabul river system. Finally the Iqbal data set is statistically analysed for time dependent relations, which yields some unexpected results. The reports finishes traditionally with a general discussion and conclusion to summarize the findings of the individual chapters.

2 Seasonal variation in *E. coli* concentrations

2.1 Introduction

This chapter addresses SRQ 1 by statistically modelling the behaviour of *E. coli* through time series analysis. Figure 2.1.1 shows the fluctuations in *E. coli* concentrations over the full two and a half year sampling period, along with discharge data. To the human eye the seasonality is immediately obvious, with concentrations rising in springtime, remaining high throughout the summer before dropping again come autumn. Within this pattern there appears to be additional repetition, with a secondary peak in *E. coli* concentrations after the main event. Repetitive behaviour of a certain variable throughout a time series is called auto correlation. This behaviour can be statistically analysed with an auto regressive statistical model. Auto regression means that a variable is regressed on its own passed values. This chapter aims to show auto correlation in the Iqbal data set by fitting a series of auto regressive models.

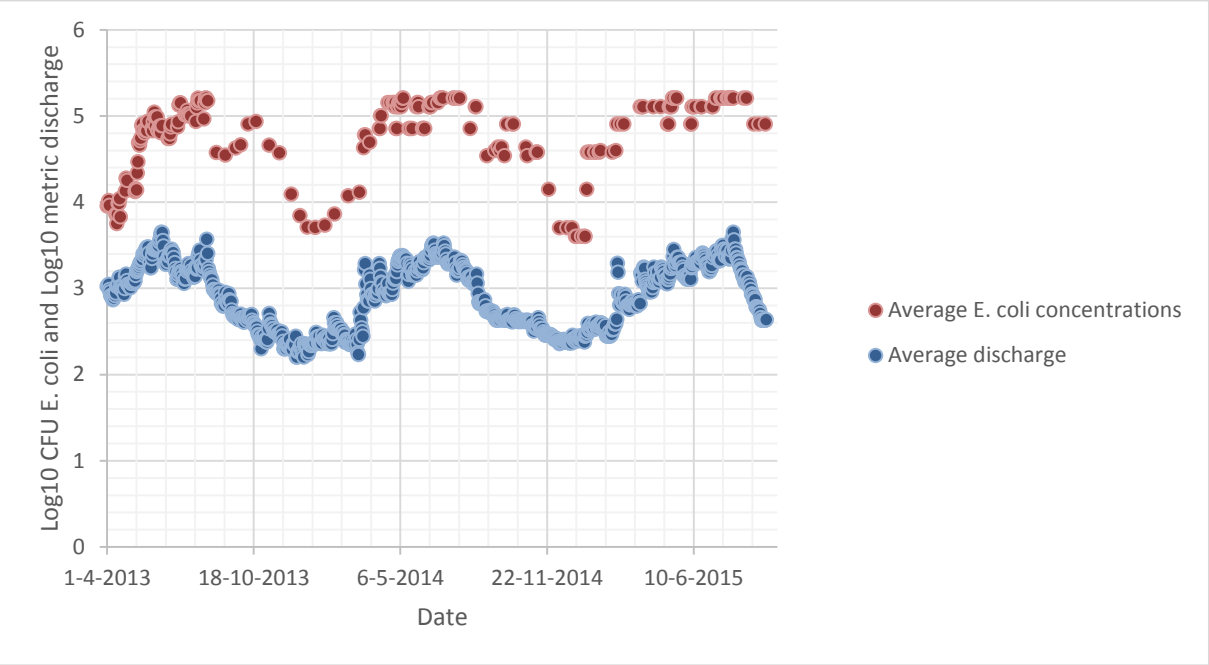

Figure 2.1.1, E coli. and discharge in the Kabul river from April 2013 to September 2015. E coli. can be seen to rise along with discharge in springtime, stay high throughout the summer and then drop when autumn comes, including a secondary 'peak' which occurs while discharge rates continue to decrease.

Within this research several approaches to autocorrelative modelling were tested using the R open source software package. The script that was used can be found in Appendix I, the generated output in Appendix II. The most minimalist option is an auto regressive (AR) type model consisting of (p) terms. An ARMA(*p,q*) type model adds use of a (q) number of moving average terms to the auto regressive part. Finally, ARIMA(*p,d,q*) adds (d) steps of differencing of the time series. This approach will be further explained in the methodology section. It is important to note that auto regressive modelling does not attempt to explain underlying mechanisms that cause the behaviour of the modelled variable and cannot be used as such.

2.2 Methodology

A general ARIMA(*p,d,q*) model can be described with a linear equation as shown below, model variables and parameters are explained in Table 2.2.1.

$$Y_t = c + \Phi_1 Y_{d,t-1} + \Phi_p Y_{d,t-p} + \ldots + \theta_1 \varepsilon_{t-1} + \theta_q \varepsilon_{t-q} + \varepsilon_t \quad \textit{(Equation 1)}$$

The principle of ARIMA is to use lagged values of the dependent variable as regressors (i.e. independent or explanatory variables) for the auto regressive (AR) part, and lagged values of the model error as regressors for the moving average (MA) part. An integrative (I) part is added by differencing the original time series. This is done in order to achieve stationarity, which means the statistic properties do not change depending on which point in the time series is chosen as a starting point.

Table 2.2.1, model variables and parameters of general ARIMA(p,d,q)

Y_t	Dependent variable at time (t)
c	Constant
ε_t	Model error
$\Phi_{1\ldots p}$	Coefficient for independent variable Y_d at time (t-1…p)
$Y_{d,t-1\ldots p}$	Independent variable at time (t-1…p), differenced (d) times
$\theta_{1\ldots q}$	Coefficient for independent variable ε_t at time (t-1…q)
$\varepsilon_{t-1\ldots q}$	Independent variable at (t-1…q)

Measurements in the original research were taken irregularly, but the closest common intervals were consecutive days. With the retention time of water in the research area being roughly half a day (Iqbal, 2017) a maximum lag time of t-5, a factor 10 increase, was set as a limit for auto regression and moving average terms. Differencing was limited to a maximum of 3 times.

With 0-5 possible AR and MA terms and 0-3 orders of integration, in total 6*6*4-4 = 140 different models were fitted to the data set. The models were compared using the Aikake Information Criterion (AIC), which computes goodness of fit as well as parsimony among models (Burnham & Anderson, 2002). Using the AIC, a lower test statistic is considered to be better. This is however a relative number, as the AIC only compares the selected models to one another. A very low score therefore does not have to mean it is a very good model, just that is far better than other models in the tested selection. Because the AIC can only compare models fitted to the same dataset a separate test was performed for each order of integration (0-3). The AIC favours models with a greater number of observations (Burnham & Anderson, 2002). For the given data set this meant that a high number of AR and/or MA terms was a serious drawback, because due to the irregular intervals a lot of data points were then missing lagged regressors and were thus dropped.

If there is auto correlation in a given data set, a properly fitting auto regressive model is expected to reduce the residuals of the outcomes to white noise i.e. only showing random fluctuations. Oppositely, when there is an element of randomness in a data set and an auto regressive model is fitted, the residuals will show auto correlation caused by the model. To see whether the time series of *E. coli* is indeed auto correlated, Partial Auto Correlation Functions (PACF) of the residuals for each of the best scoring models per number of integration were plotted.

2.3 Results

Table 2.3.1 shows the models with the lowest AIC statistic per order of integration. An ARIMA(3,0,0) model has the lowest AIC score of all the models. Note that this is essentially an AR(3) model. The fact that it scores lowest does not necessarily mean it has the best fit, since it was only compared to other models with zero integrations applied to the data set.

Table 2.3.1, AIC statistics of ARIMA(p,d,q) models with 0-3 steps of integration. A lower score indicates a better fitting and more parsimonious model, but only among models applied to the same data set i.e. order of integration.

Model	AIC
ARIMA (3,0,0)	308.20
ARIMA (0,1,5)	311.60
ARIMA (5,2,0)	449.14
ARIMA (0,3,3)	447.79

Figure 2.3.1 to 2.3.4 show respective PACF plots with a lag times up to 10 days for the residuals of each of the models displayed in Table 2.3.1. These are the models with the lowest AIC statistic per order of integration. As mentioned before, a properly fitting ARIMA(p,d,q) model should reduce its residuals to white noise without any auto correlation. This means that the outcome of the auto correlation function should stay within the error bands at each lag time. An outcome that exceeds the error bands shows that there is significant auto correlation in the residuals at the given lag time. Note that the scale and extent of the vertical axis differs for each plot due to large differences in strength of autocorrelation at certain lag times between models.

In each plot the outcomes of the PACF exceed the error bands, in three out of four models this already happens at a lag of one day. The ARIMA(5,2,0) model is the only one where the first two time steps do not show auto correlation. This is, however, offset by very strong relations at lag 9 and 10. Overall, none of the models is able to transform the residuals of the time series to white noise. The ARIMA(3,0,0) model residuals show a good example of behaviour that is typically expected from a poorly fitting auto regressive model, with strong auto correlations at (multiplications of) lag 3.

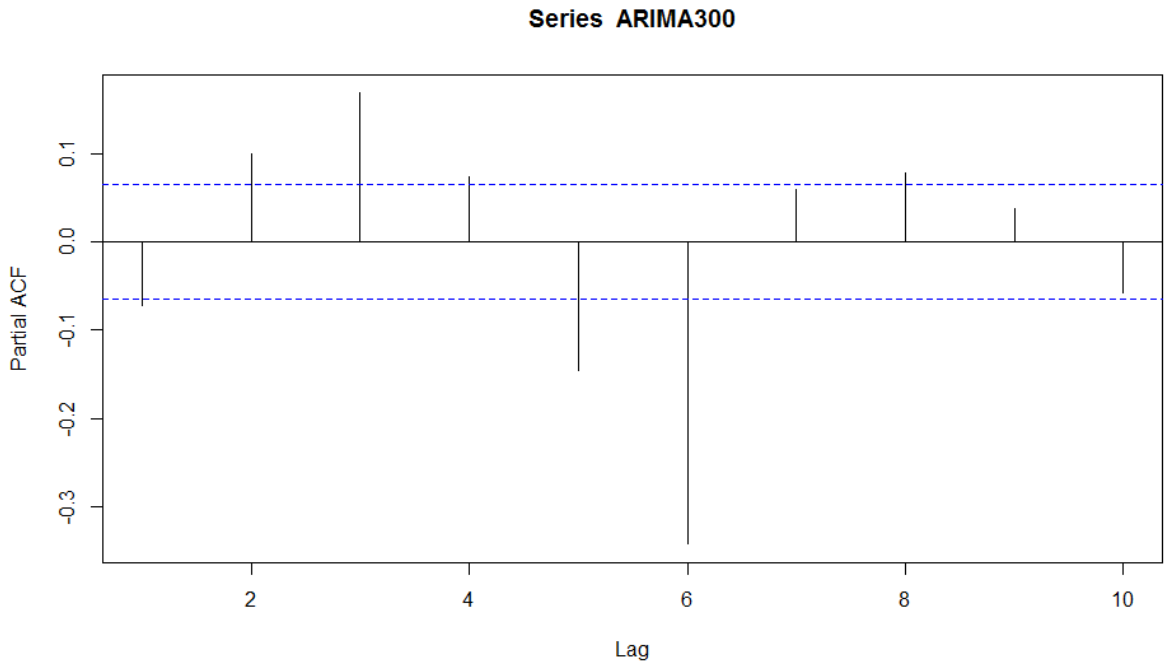

Figure 2.3.1, PACF plot of best fitting model with zero integration steps. Error bands are

displayed in blue. Lag time is given in days on the horizontal axis, the ACF statistic for each day is shown on the vertical axis.

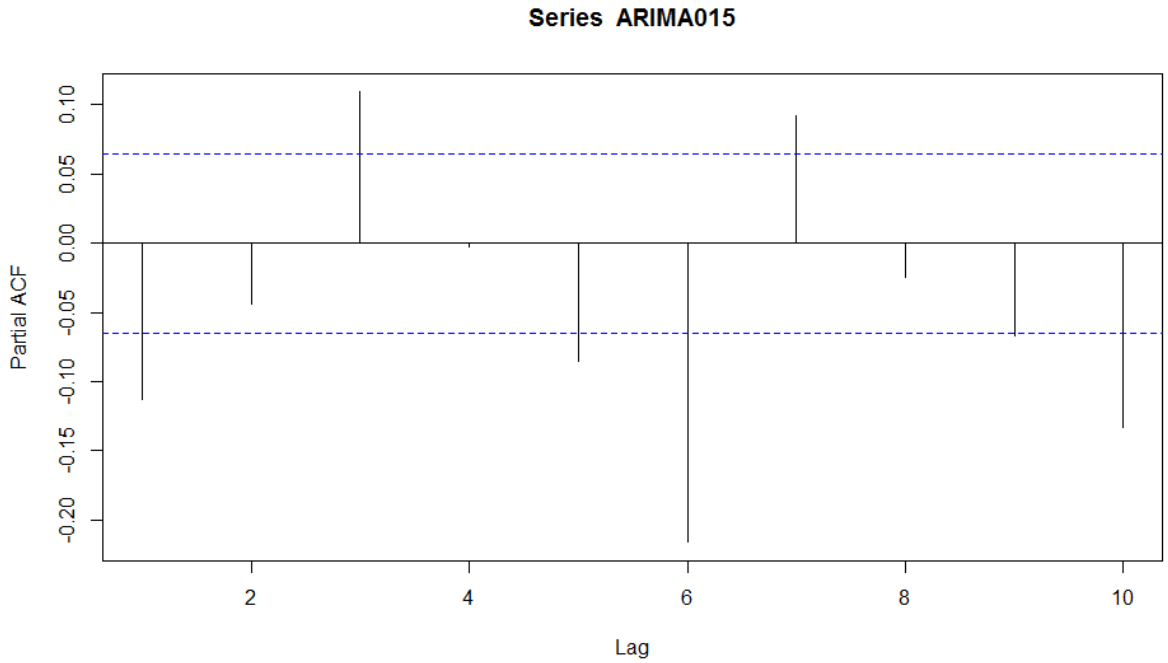

Figure 2.3.2, PACF plot of best fitting model with one integration step.

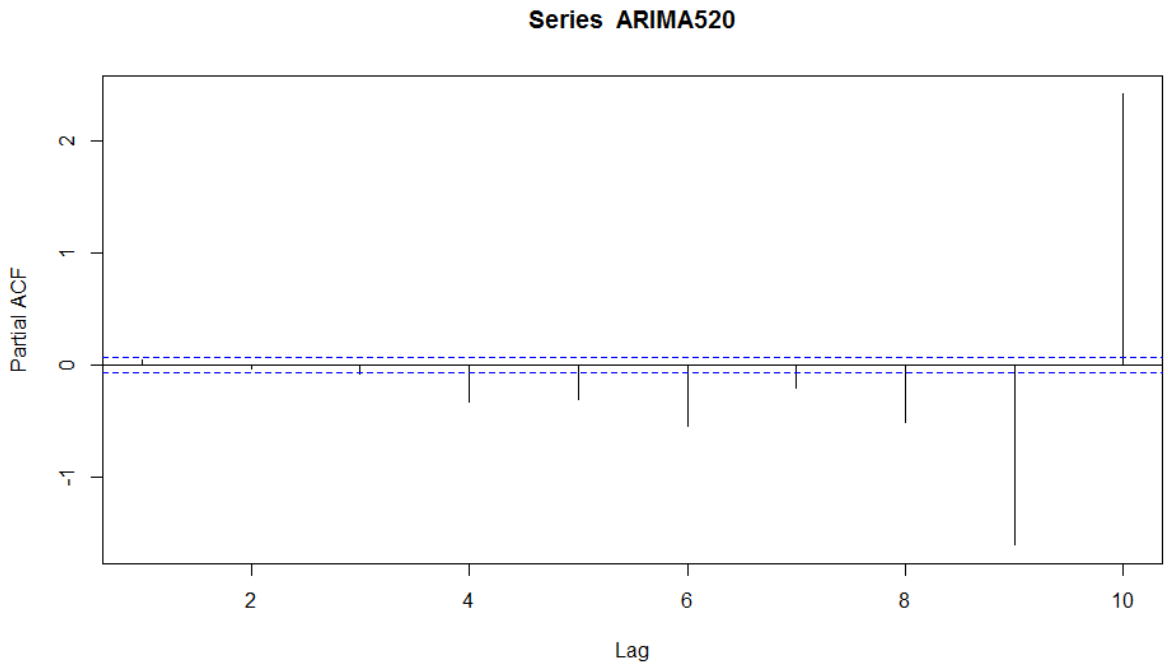

Figure 2.3.3, PACF plot of best fitting model with two integration steps.

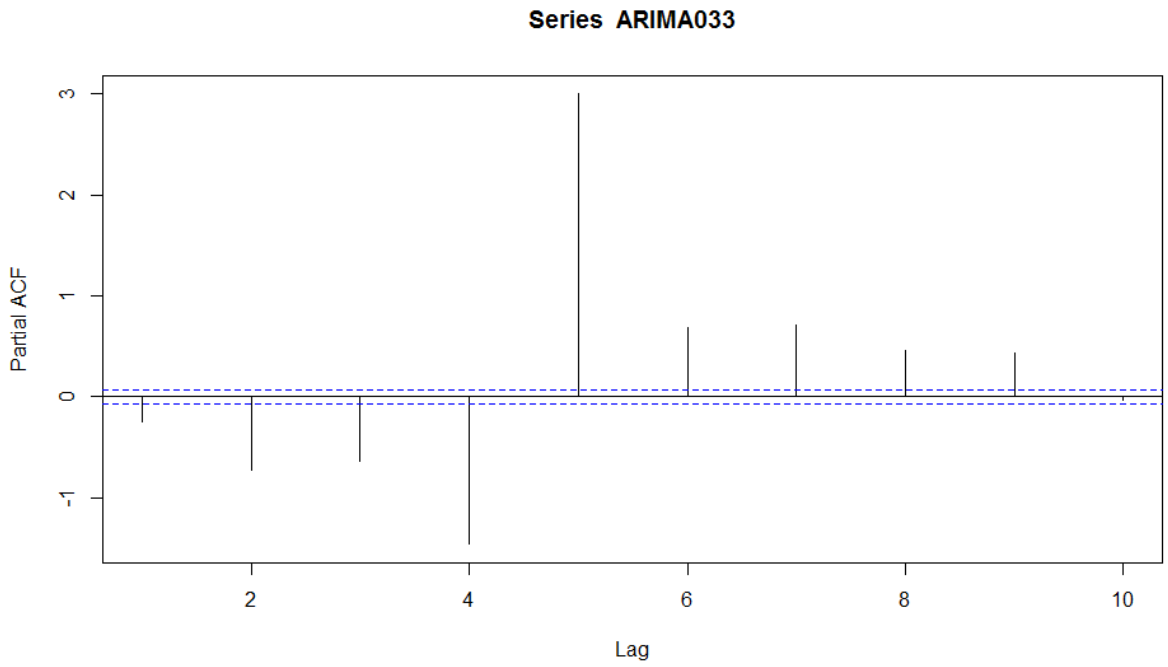

Figure 2.3.4, PACF plot of best fitting model with three integration steps.

2.4 Discussion and conclusion

This chapter aims to find an answer to SQR 1, to see if fluctuations of E. coli in the Kabul river basin are auto correlated. First of all, by choosing days as time steps for the ARIMA models a large proportion of the original data set was excluded as input. Iqbal (2017) only took daily samples throughout the flooding periods in spring and summer, with sampling limited to biweekly intervals for the remainder of time. This means that with the chosen ARIMA models only the flooding events were modelled and not the entire time series. The finding that no auto correlation was shown therefore only applies to periods of high discharge, and only to short term (1-5 day) relations.

Even during periods of daily sampling there sometimes were one or several days without measurements, causing irregularity in the intervals. This is problematic with longer lag times since a 'break' in the string of lagged values causes the data point to be discarded. It is therefore strongly recommended that in future research intervals are kept as regular as possible.

Although a time series stretching 30 months is considered to be very long by microbiological research standards, for the purpose of ARIMA type analyses, especially seasonal ones, it is still relatively short. Box and Tiao (1975) recommend at least 50 observations for an ARIMA model, preferring 100. This implies that to model the yearly occurrence of the secondary peak at least 50 years of evenly spaced data would be required to construct an ARIMA model with a seasonal component. Given the usual length of research projects within the field of environmental systems analysis this is not realistic.

An alternative approach could be to use a variable with a known linear relation with *E. coli* as a substitute. A good example could be the number of registered *E. coli* infections in a given research area. Prerequisite is of course that local institutions have kept records and are willing to make these available. If a strong correlation can be found over the two to three years of sampling that is realistic for research in the field of ESA, one could assume relatively

safely that past fluctuations in infection incidence closely match un-observed fluctuations in presence of *E. coli* in the environment.

In conclusion the question if *E. coli* concentrations in the Kabul river are auto correlated cannot be answered based on the results of my research. None of the best fitting models for each order of integration is able to reduce the residuals of the original time series to white noise. This means that in statistical terms the hypothesis that *E. coli* concentrations are auto correlated would thus be rejected. Given the similarity seen between just two full flooding cycles it is however likely that a longer and more evenly spaced data set, if ever available, will yield a well-fitting ARIMA model.

3 Pathways of *E. coli* in the Kabul river basin

3.1 Introduction

This chapter aims to find an answer to SRQ 2 by identifying the pathways of *E. coli* into the surface water of the Kabul river that are likely to contribute to the observed secondary peak. The goal is to give some direction to the statistical analysis on environmental variables later on. This is done through performing a limited literature research and consequently developing a conceptual model displaying the sources and pathways.

This chapter presents two conceptual models and an attempt at a theoretical explanation for the secondary peak.

3.2 Methodology

A two stage approach was used to develop a conceptual model for *E. coli* specifically for the Kabul river. Firstly a general conceptual state/rate model was developed displaying the known pathways contributing to the release of *E. coli* into surface waters. This model was then used as a basis for constructing a model specifically showcasing the researched part of the Kabul river.

Literature research was performed using an 'oil spill' approach with relevant literature being checked for references and citations to make sure recent findings could be adopted as well. The main source articles for references and citations were review articles by Vermeulen et al. (2015) and Cho et al. (2016).

3.3 Results

General conceptual model

In Figure 3.3.1 a conceptual model for the pathways of *E. coli* into and out of the surface water system is shown. The model is largely adapted from Hofstra (2011) and Iqbal (2017). The model has the form of the widely used state/rate box model with input and output fluxes (rates) in respectively orange and green representing transport or conversion of a chosen variable, in this case *E. coli*. These fluxes determine the concentration or amount (state) of that variable in a chosen spatial body, in this case surface water. Pathways are separated into those that transport *E. coli* into the system from a human source, those that do so from an animal source, and some auxiliary ones. The pathways are described in more detail in the proceedings of this section.

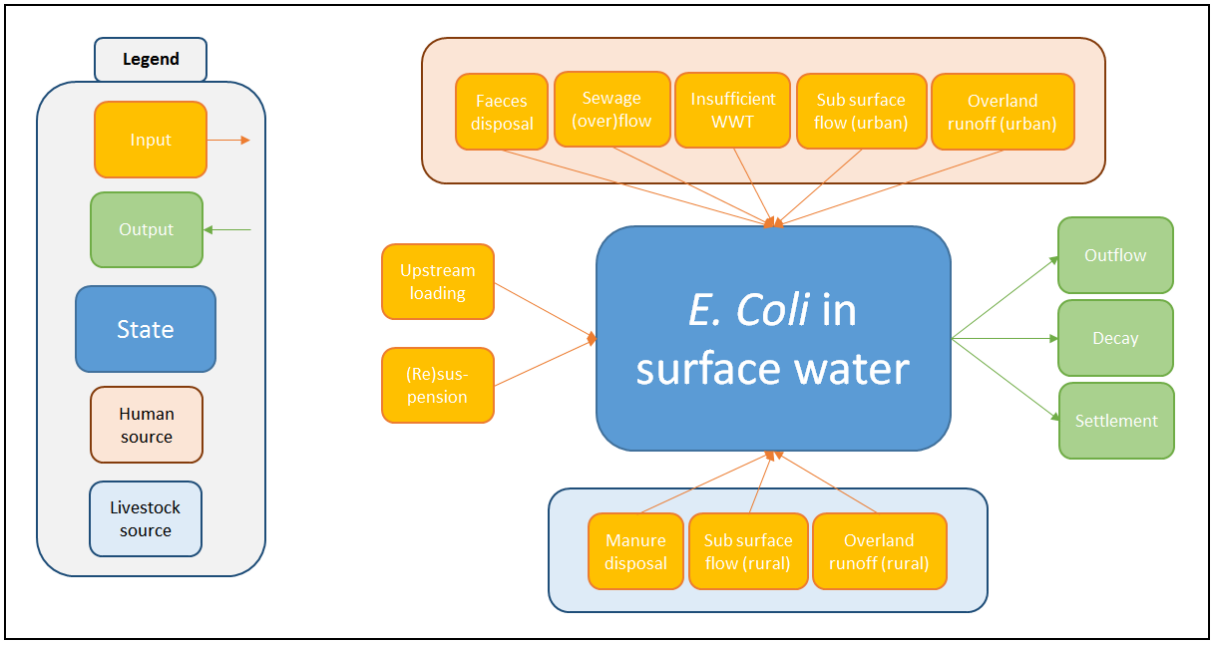

Figure 3.3.1, General conceptual model of E. coli in surface waters.

Upstream loading constitutes the *E. coli* that is fluvially transported into the system from the outside of the specific spatial bounds. Usually this will be an artificial or natural border like a dam, a weir, a waterfall or other type of obstacle. The *E. coli* that enters the system through this pathway did obviously originate from a human or animal source in some way, but since the release and transport into surface water occurred outside the given system these processes are not included in the model. Outflow is the opposite process, *E. coli* leaving the system through the set spatial bound by suspended fluvial transport. Decay is usually a fairly constant rate representing the die-off of *E. coli* in the environment. A higher decay rate is associated with rising water temperatures (Hofstra, 2011).

The settlement and resuspension pathways represent *E. coli* being exchanged between surface water and sediment, with sediment not being included as a 'state' in this model. Settlement occurs when suspend particles to which *E. coli* has attached itself are overcome by gravity and settle into the sediment (Pachepsky & Shelton, 2011). Since *E. coli* has a density similar to water, settlement of free floating organisms or colonies is minimal at best (Pachepsky & Shelton, 2011) and can be expected to be non-existent under turbulent conditions in flowing water. Suspension happens when particles are elevated into the water column from the sediment due to an increase in kinetic energy of the water, for instance a rise in flow velocity (Pachepsky & Shelton, 2011) or wave action (Whitman & Nevers, 2008).

As far as pathways from human sources go direct disposal is not strictly a pathway controlled by environmental variables (except for gravity), but included in the model for sake of completeness. When human waste is collected and transported by sewage systems, release into the environment can occur through overflows. These are associated with heavy rainfall events causing water flow to temporarily exceed the capacity of the system (Rechenburg et al., 2006), but a poorly designed and/or maintained system can also cause a continuous release (Iqbal, 2017). If the sewage system does its job and *E. coli* is transported to a waste water treatment plant (WWTP), release into the environment is dependent on the proper functioning of the WWTP. For instance, the Peshawar WWTP in the research area of this study was heavily damaged during floods in 2010 and never restored, causing it to be defunct for the entire sampling period (Iqbal, 2017). Subsurface transport is dependent on pores and cracks in the soil, which is largely influenced by the sorptive properties of the sediment type (Guzman et al, 2009). Finally, there is overland runoff. Without a sewage system, which can often be the case in rural areas, human waste can build up over time and

a single heavy rainfall event can cause a considerable release of *E. coli* into the surface water (Kistemann et al., 2002)

Regarding pathways from livestock sources, direct disposal plays a more important role due to cattle grazing in the direct vicinity of surface waters (Iqbal, 2017). Subsurface transport can become a mayor pathway if a drainage system is present, because transport towards the drain to cracks and pores only has to occur over a minimal distance for *E. coli* to reach the drain and flow towards the surface water (Guzman et al., 2009). The overland runoff pathway is essentially the same as for human sources but due to the density of livestock generally being higher and off course the absence of sewer systems this pathway can represent very large releases of *E. coli* into the environment (Kistemann et al., 2002).

Conceptual model of Kabul river research area

In Figure 3.3.2 a conceptual state/rate box model of the research area of this study is displayed. Model properties are the same as the general conceptual model discussed previously, with the state boxes representing the different stretches of the Kabul river in the research area and the input and output rates having been made specific to the given stretch.

The first stretch starts right below the Warsak dam and continues towards the point where the river divides into various separate branches. There are no major urban areas connected to this stretch, so pathways are limited to agricultural surface runoff and manure application which goes for all stretches. Iqbal (2017) described the abundance of animal shelters all over the flood plain where great amounts of manure were stored. Upon the commencement of the floods all this manure ends up in the water over a short space of time. Off course the Warsak stretch is the one to receive *E. coli* through upstream loading from above the dam. Among the branches are the Khyali stretch and the Sha Alam stretch. A sampling point was located in both stretched during the research. In addition to the mentioned pathways the Khyali stretch receives water from the Swat river which is connected to the city of Charsada, adding urban runoff and untreated sewage (Iqbal, 2017) to the system. The Shah Alam stretch receives water from an irrigation canal connected to the city of Peshawar, which adds urban runoff and untreated sewage in the same fashion (Iqbal, 2017). The separate branches re-join east of Charsada, forming the Nowshera stretch which is connected to the city of Nowshera, once more adding a flux of *E. coli* through urban runoff and untreated sewage (Iqbal, 2017). Settlement and resuspension can be expected to play a role throughout the system, but the extent to which is subject of speculation on my part. Decay changes with temperature, but with a retention time for *E. coli* of less than half a day in the ‘state’ boxes temperature does not show relevant fluctuations.

Looking at the model it is obvious that rainfall driven processes are dominant in both the human and the livestock sourced pathways, and can therefore be expected to be a strong predictor of variation in the observed concentrations of *E. coli*. The extent of the influence of settlement and (re)suspension is a great unknown in this model, to be elaborated on in the discussion and conclusion.

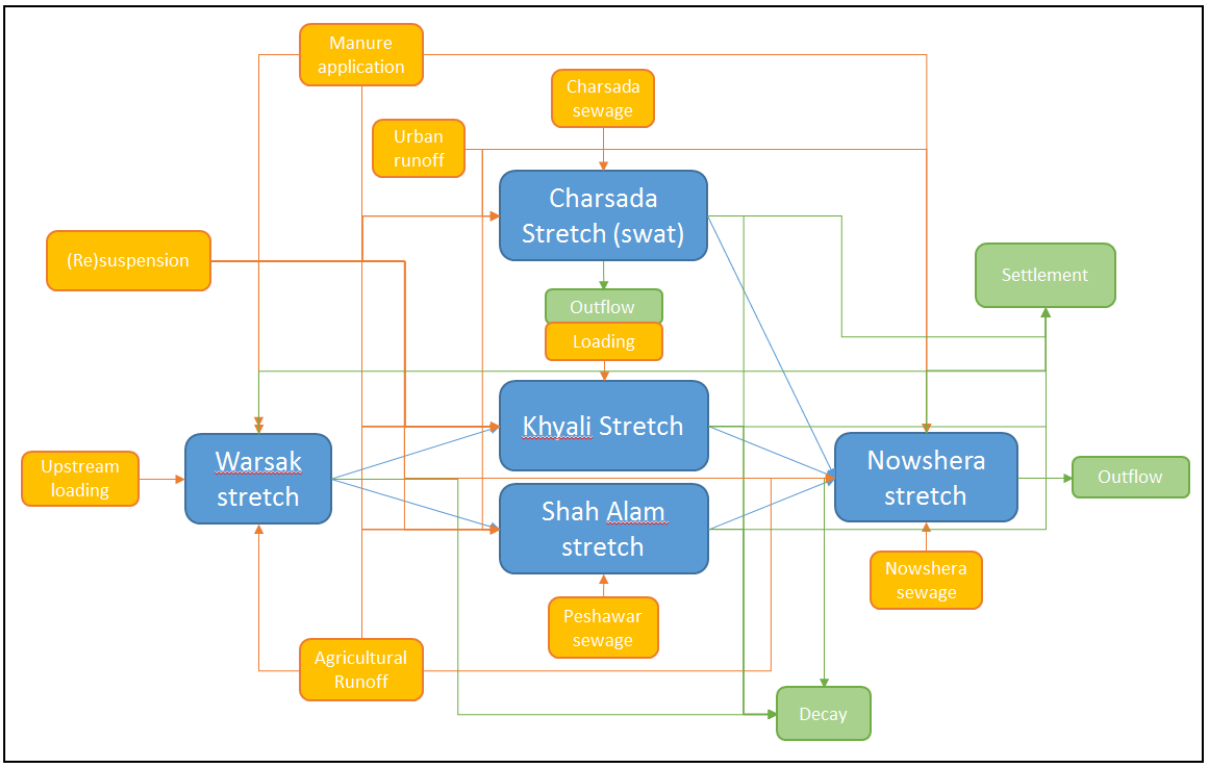

Figure 3.3.2, Conceptual model of E. coli specific to the research area of the Kabul river that was examined in this study.

3.4 Discussion and conclusion

The goal of this chapter was to see what pathways are most likely to contribute to the observed secondary *E. coli* peak. The model shows clearly that most pathways are rainfall driven. It is therefore curious that the secondary peaks occur in a period where major rainfall events have largely subsided, pointing towards other pathways being responsible. The two pathways that are not directly driven by rainfall are (re)suspension in the incoming side, and settlement on the outgoing side. The quantitive role of settlement and resuspension in release of *E. coli* is relatively poorly understood (Iqbal, 2017; Vermeulen & Hofstra, 2013), but might be important in trying to explain the secondary peak, since this phenomenon occurs at a moment where the influence of rainfall and subsequent discharge on *E. coli* release is expected to be dwindling.

Given the known influence of wave action on suspension of sediments (Whitman & Nevers, 2008), and the ability of *E. coli* to sustain itself for a long time under anaerobic conditions (Pandey et al., 2015), a resuspension of sediments could occur as the water resides and slight waves start rolling across the floodplain causing a re-release of *E. coli* into the system. Additionally a decrease in discharge might actually cause an increase in flow velocity due to a proportionally large reduction in frontal surface flow area, promoting further resuspension.

In conclusion the model shows that the majority of the input fluxes are rainfall driven, (re)suspension being the exception. The downward trend of rainfall and associated discharge when the secondary peak occurs points towards the (re)suspension pathway as a likely cause. Even so, it seems unlikely that rainfall driven pathways play no role at all if only for the density of people and livestock in the basin. Rainfall is still expected to be an important explanatory variable in the statistical analysis. Regarding (re)suspension, the closest measure of suspended sediments available in this study is water turbidity, see next chapter, and thus water turbidity is considered to be a likely predictor of the secondary peak.

4 Time dependent relations of *E. coli* with environmental variables

4.1 Introduction

This chapter aims to answer SQR 3 and 4, to find correlations between *E. coli* concentrations in the Kabul river and environmental variables. The original goal was to focus on a model that would explain *E. coli* variation around the peak, but due to an insufficient number of observations the approach was broadened to analysing the entire time series. The same variables available to Iqbal (2017) were used to see if an alternative regression approach could yield a better fitting model. The data set was acquired from Iqbal (2017) and is normally distributed.

Linear relations between *E. coli* and environmental variables have been studied fairly extensively. Temperature can be correlated negatively (Vermeulen & Hofstra, 2013; Walters et al., 2011) as well as positively (Islam et al., 2017). Since a higher temperature increases the decay rate of *E. coli*, positive correlations are considered to be caused by a separate mechanism. Temperature increasements often occur along with rises in precipitation level and river discharge (Islam et al., 2017), and these two variables have widely been found to correlate positively with *E. coli* (Vermeulen & Hofstra, 2013; Walters et al., 2011; Kay et al., 2005). Salinity was found to be negatively correlated (Walters et al, 2011), while barometric pressure showed a positive correlation (Whitman & Nevers, 2008).

In the mentioned publications, along with many others, statistical analyses usually focussed on direct relations in terms of temporal separation, meaning the measurements of both the dependent variable (*E. coli* or other aquatic microbiota) and the independent variables were taken on the same day, often even within the same few minutes.

Following the pathways described in the previous chapter it might take several days for *E. coli* organisms to end up in the surface water from the moment they are released. This means that developing regressions models with lagged variables can be of great help in gaining a better understanding of the influence of these variables on the functioning of those pathways.

4.2 Methodology

The original goal of this part of the research was to try and find lagged correlations between environmental variables and the secondary peak in *E. coli*. Closer examination of the data set however showed that the two observed peaks consisted of just three and two measurements each. Since a total of five observations is too small to fit a statistical model, the goal was changed to performing a lagged regression analysis on the entire data set to see if the statistical model developed by Iqbal (2017) could be improved upon. A lagged regression model is essentially the same as the general linear model that is widely used, the only difference being that the independent variables are lagged on the dependent variable. For statistical analyses, the R open source software package was used. The script that was used can be found in Appendix III, the generated output in Appendix IV.

Six variables were available to be tested as independent variables in a lagged regression model. They are displayed in Table 4.2.1. With the retention time of river water in the research area being half a day again a lag of five days, a factor ten increase, was set as a limit.

Table 4.2.1, independent variables for lagged regression model

Variable	Unit
Discharge	M^3*s^{-1}
Conductivity	$mS*M^{-1}$
Acidity	pH
Precipitation	$mM*d^{-1}$
Temperature	^oC
Turbidity	mM

As a start all independent variables with lag 0-5 were put into a single model. From this point an iterative approach was used, eliminating as many variables as possible while trying to raise the adjusted R^2 statistic, which quantifies the extent to which the model explains the variation in observations, as high as possible. After having arrived at the best fitting model, the variables which showed the greatest explanatory value ($p < 0.05$) were checked for both collinearity and interaction. Presence of collinearity implies that independent variables are not independent from one another, an example is the strong relation between rainfall and discharge. Collinearity was checked by regressing each model variable over another. Interaction means that the predictive strength of one variable is dependent on variation in another. An example is correlation between algae concentration and phosphate levels, but only above a certain water temperature. Note that the relation between phosphate and water temperature is non-causal in this example. Interaction was checked by adding an extra term constituting the product of multiplication of two variables to the original model. Results are given in the next paragraph.

4.3 Results

The lagged regression model with concentration of *E. coli* as the dependent variable and observations of temperature, conductivity and acidity as independent variables is shown below. Model terms are explained in Table 4.3.1, model parameters and values are given in Table 4.3.2.

$$Y_t = \alpha + \beta_1 X_{1,t-3} + \beta_2 X_{2,t-3} + \beta_3 X_3 + u_t \quad (adj.\ R^2 = 0.756) \qquad (Equation\ 2)$$

Table 4.3.1, model terms of lagged regression model.

Y_t	Dependent variable at time (t)
α	Constant
u_t	Model error
$\beta_{1\ldots n}$	Coefficient for independent variable $X_{1\ldots n}$
$X_{1\ldots n,t-1\ldots n}$	Independent variable at time (t-$1\ldots n$)

Table 4.3.2, Parameters and values of lagged regression model. All estimates are found to be significant ($p < 0.05$). The model has an adjusted R^2 of 0.756

Variable	Term	Lag	Coefficient	Estimate	p < 0.05
Intercept	α	*n.a.*	*n.a.*	-9.520	*
Temperature	X_1	3	β_1	0.02822	*
Conductivity	X_2	3	β_2	-0.002134	*
Acidity	X_3	0	β_3	1.683	*

Tables 4.3.3 and 4.3.4 respectively show the results of the collinearity and interaction checks. The lagged observations of temperature and conductivity at t-3 are showing significant collinearity ($p < 0.05$). The adjusted R^2 however remains fairly low, indicating that variables not included in the model play a greater role in explaining variation in temperature and conductivity.

The interaction variables added to the original model are not found to be significant except for interaction between conductivity at t-3 and acidity at t-0. The model with this interaction variable added has an adjusted R^2 of 0.7632 compared to 0.756 for the original model. This is an improvement of less than one percent which indicates that although interaction is present it plays a relatively minor role in explaining the variation in *E. coli* concentrations.

Table 4.3.3, Collinearity check for model variables. Only temperature and conductivity at t-3 for both variables are found to be significantly collinear, but to a relatively small degree.

Variables	p < 0.05	Adj. Rsq
X1 ~ X2	*	0.2754
X1 ~ X3	-	-0.00537
X2 ~ X3	-	-0.01029

Table 4.3.4, Interaction check for model variables. Significant interaction is only shown between conductivity at t-3 and acidity at t-0. Note that the improvement in R^2 over the original model is less than 1%.

Parameter	p < 0.05	Adj. Rsq.	Difference
*X1*X2*	-	0.7525	-0.35%
*X1*X3*	-	0.7568	0.08%
*X2*X3*	*	0.7632	0.72%

4.4 Discussion and conclusion

This chapter's goal was to develop an approach to statistically analysing the secondary peak, and to see what variables can be statistically shown to contribute. Due to a limited number of observations around the peak the goal was altered to finding out whether time dependent relations between *E. coli* and environmental variables can be shown statistically for the full time series.

The general linear model fitted to the biweekly dataset by Iqbal (2017) had an adjusted R^2 of 0.61, which was already fairly high when compared to similar studies that found values between 0.3 and 0.6 (Kay et al, 2005; Whitman & Nevers, 2008; Walters et al,. 2011; Vermeulen & Hofstra, 2013). The value of well over 0.7 that was found in this study is unseen in literature as far as regression of *E. coli* over environmental variables goes. Since correlation does not imply causation the immediate question that these results raise is if temperature, conductivity and acidity are indeed directly or indirectly influencing the release of *E. coli* into surface water, or if completely separate mechanisms are at work.

The expected strong correlation with rainfall when using lagged variables was not found at all. This could be due to the fact that rainfall can have a greater or smaller effect based on other circumstances. A heavy rain shower after a long dry period during which a lot of manure has accumulated on the lands will result in a big run off of *E. coli* towards the surface water. If that very same shower would repeat itself a few days later it would first of all be largely absorbed by the now damp soil and there would be little if any manure present to be

transported towards the water as well. This irregular consequence makes rainfall a hard predictor to model. A purposeful use of interaction variables, in this case for instance soil humidity and manure levels, could improve this considerately.

Out of the variables available for this research I considered turbidity to be the best measure for (re)suspension of sediments. Same as for rainfall, the statistical analysis did not find correlations with turbidity to be significant or adding to the model's explanatory value. A likely reason for this is that the level of turbidity in the column is influenced by many more factors than just (re)suspension of sediments within the system boundaries. Suspension from outside, erosion and possibly bacterial and algae growth or release of biofilm can all have a cumulative role in determining the level of turbidity. The proportionate contribution of the in-system (re)suspension process might be big or small, and this uncertainty is likely to explain the low correlations that were found.

In conclusion, the original goal of this chapter to develop a model that finds correlations for the secondary peak was not reached due to a lack of suitable data. By fitting a lagged regression model tot the data set, an adjusted R^2 of 0.756 is achieved with just three independent variables. This indicates that a very well fitting model is found, and that time dependent relation can be considered existing and strong. Collinearity and interaction between variables was shown but to a relatively minor degree.

5 Overall discussion and conclusion

5.1 Discussion

Auto regression

The major cause for the lack of meaningful results this research offered in terms of auto regressive modelling was that despite the data set stretching a very long time span by the standards of microbiological research, it was still way too short to be used for a time series analysis. Irregular intervals between observations further amplified this, since a single break in the string off lagged variables will cause the data point to be discarded.

The obvious solution to this is continuous sampling at daily intervals for the entire research period, but this might be overly labor-intensive. Alternatively weekly or even biweekly intervals might be used, provided a long enough time series can be constructed from these observations. Given the fact that short term auto correlation was not shown in this research, longer term relations on the level of weeks or months can be an interesting topic for future research.

When looking at seasonal relations, as in annually repetitive behavior, it is not likely that any series of data collected by the ESA group will ever be long enough to properly fit an auto regressive model to. The possibility of substituting *E. coli* with another variable of known linear relations of which longer time series exist was discussed in section 2.4, and this could be an interesting path to explore in future research. Since *E. coli* is not necessarily a pathogen itself and its function is essentially that of an indicator for fecal contamination, another thoughtfully selected variable could be used for that purpose as well.

Pathways of E. coli

As mentioned before, the role of settlement and resuspension of sediment, and the extent to which it influences presence of *E. coli* in the environment is not completely clear at this point in time. Wave action and flow velocity could be important predictors of this but were unfortunately not measured by Iqbal. Turbidity was chosen as a variable indicative of (re)suspension, but even within the system itself this can be caused by other factors like erosion associated with precipitation or even biological factors such as bacterial growth or, to a certain degree, algae growth or release of biofilm. Selection of an appropriate variable related to the (re)suspension of sediments could help shed more light on this matter.

Growth of *E. coli* was not included in the conceptual model since the model is specific to a river system with relatively fast flowing water. There is no scientific evidence of *E. coli* growth under such conditions. (Iqbal, 2017) A process that could be going on however is growth in the puddles and moist soils that persist on the floodplain after water subsides as discharge decreases in autumn. Certain strains of *E. coli* are known to be capable of growth in the environment under still water conditions (van Elsas et al., 2011). If conditions right after the flooding season favor growth near the edges of the low-discharge canals, one could imagine a relatively small rain shower or even a slight increase in water level being able to cause a substantial release of *E. coli* into the surface water. This is off course speculative, but in future research this is a factor that could be kept in mind and justifies extra measurements when conditions that could promote growth occur in a given research area.

Regression on environmental variables

Although regression can be used to develop models with great predictive value, as this study has shown once more, it might not necessarily be a useful tool when it comes to establishing causal relations. In any given water system there may be hundreds, if not thousands of environmental variables acting upon the presence of microbiota like *E. coli* in the surface water and upon each other. To choose a limited set of variables and use correlative values to go looking for causation is a tricky process, as underlying mechanics might be very different.

To end on a positive note, my study found very strong correlations and a high explanatory value when constructing a regressive model with lagged variables. It would be interesting to see how the model holds up when it is fitted to similar data sets *on E. coli* in flowing waters. The data from Islam (2018) that were collected in Bangladesh could be a good starting point for this since they are readily available at the ESA group and stretch a similar time span.

5.2 Conclusion

The main question that my study aimed to answer was how the observed secondary peak in *E. coli* concentrations could be explained. Due to a time series being too short to fit an ARIMA model with a seasonal component, existing literature being inconclusive, and an insufficient number of observations around the peak for statistical regression on environmental variables, this question remains open by the end of this research.

Fortunately not all is bad. My conceptual model indicates that the secondary peak is likely caused by processes not driven by rainfall to a large degree, since it occurs at a moment when both rainfall and subsequent discharge are on the descend. Resuspension of *E. coli* carrying sediments and possibly even a pathway involving environmental growth could well be the causes, more specific research could help validate these findings.

Finally my lagged regression model showed a predictive strength that I have not seen in any other study on the topic of faeces-related microbiota in the aquatic environment. These findings strongly imply that time-dependent relations between *E. coli* and environmental variables exist, and that lagged regression might be a more suitable approach to statistical analysis than the commonly used linear models.

References

Atherholt, T. B. *et al.* (1998) 'Effect of rainfall on Giardia and crypto', *Journal / American Water Works Association*, 90(9), pp. 66–80.

Box, G. E. P. and Tiao, G. C. (1975) 'Intervention analysis with applications to economic and environmental problems', *Journal of the American Statistical Association*, 70(349), pp. 70–79.

Burnham, K. P. and Anderson, D. R. (2002) *Model Selection and Multi-model inference*. Springer, New York.

Van Elsas, J. D. *et al.* (2011) 'Survival of Escherichia coli in the environment: Fundamental and public health aspects', *ISME Journal*. Nature Publishing Group, 5(2), pp. 173–183.

Fischer Walker, C. L. *et al.* (2013) 'Global burden of childhood pneumonia and diarrhoea', *The Lancet*, 381(9875), pp. 1405–1416.

Ge, Z. *et al.* (2010) 'Coastal loading and transport of Escherichia coli at an embayed beach in Lake Michigan', *Environmental Science and Technology*, 44(17), pp. 6731–6737.

Gibson, C. J. *et al.* (1998) 'Combined sewer overflow: A source of cryptosporidium and giardia?' Pittsburgh: Water Science Technology.

Guzman, J. A. *et al.* (2009) 'Escherichia coli transport from surface-applied manure to subsurface drains through artificial biopores', *Journal of Environmental Quality*, 38(6), pp. 2412–2421.

Hashizume, M. *et al.* (2007) 'Association between climate variability and hospital visits for non-cholera diarrhoea in Bangladesh: Effects and vulnerable groups', *International Journal of Epidemiology*, 36(5), pp. 1030–1037.

Hofstra, N. (2011) 'Quantifying the impact of climate change on enteric waterborne pathogen concentrations in surface water', *Current Opinion in Environmental Sustainability*. Elsevier B.V., 3(6), pp. 471–479.

Iqbal, M. S. (2017) *Quantifying the Impact Socioeconomic Development and Climate Change on Escherichia Coli Concentrations in the Pakistani Kabul River*. Wageningen University and Research, Wageningen.

Iqbal, M. S. and Hofstra, N. (2019) 'Modeling Escherichia coli fate and transport in the Kabul River Basin using SWAT', *Human and Ecological Risk Assessment*. Taylor & Francis, 25(5), pp. 1279–1297.

Islam, M. M. M., Hofstra, N. and Islam, M. A. (2017) 'The Impact of Environmental Variables on Faecal Indicator Bacteria in the Betna River Basin, Bangladesh', *Environmental Processes*. Environmental Processes, 4(2), pp. 319–332.

Jorch, G. *et al.* (2010) 'Empfehlungen zur Struktur und Ausstattung von Intensivstationen - Hintergrundtext -', *DIVI Leitlinie*, 68(Divi), pp. 1–40.

Kay, D. *et al.* (2005) 'Predicting faecal indicator fluxes using digital land use data in the UK's sentinel Water Framework Directive catchment: The Ribble study', *Water Research*, 39(16), pp. 3967–3981.

Kistemann, T. *et al.* (2002) 'Microbial load of drinking water reservoir tributaries during extreme rainfall and runoff', *Applied and Environmental Microbiology*, 68(5), pp. 2188–2197.

McCarthy, D. T. (2009) ‘A traditional first flush assessment of E. coli in urban stormwater runoff’, *Water Science and Technology*, 60(11), pp. 2749–2757.

Nataro, J. P. and K. J. B. J. and Kaper, J. (1998) ‘Diarrheagenic Escherichia coli Strains’, *Clinical Microbiology Reviews*, 11(1), pp. 142–201.

Nichols, G. *et al.* (2009) ‘Rainfall and outbreaks of drinking water related disease and in England and Wales’, *Journal of Water and Health*, 7(1), pp. 1–8.

Odonkor, S. T. and Ampofo, J. K. (2013) ‘Escherichia coli as an indicator of bacteriological quality of water: an overview’, *Microbiology Research*, 4(1), p. 2.

Pachepsky, Y. A. and Shelton, D. R. (2011) ‘Escherichia coli and fecal coliforms in freshwater and estuarine sediments’, *Critical Reviews in Environmental Science and Technology*, 41(12), pp. 1067–1110.

Pandey, P. K. *et al.* (2015) ‘Escherichia coli persistence kinetics in dairy manure at moderate, mesophilic, and thermophilic temperatures under aerobic and anaerobic environments’, *Bioprocess and Biosystems Engineering*, 38(3), pp. 457–467.

Rechenburg, A. *et al.* (2006) ‘Impact of sewage treatment plants and combined sewer overflow basins on the microbiological quality of surface water’, *Water Science and Technology*, 54(3), pp. 95–99.

Vermeulen, L. C. and Hofstra, N. (2014) ‘Influence of climate variables on the concentration of Escherichia coli in the Rhine, Meuse, and Drentse Aa during 1985-2010’, *Regional Environmental Change*, 14(1), pp. 307–319.

Walters, S. P., Thebo, A. L. and Boehm, A. B. (2011) ‘Impact of urbanization and agriculture on the occurrence of bacterial pathogens and stx genes in coastal waterbodies of central California’, *Water Research*. Elsevier Ltd, 45(4), pp. 1752–1762.

Whitman, R. L. and Nevers, M. B. (2008) ‘Responses along 23 Chicago Beaches’, *Environmental science & technology*, (219), pp. 9217–9224.

Wilkinson, J. *et al.* (2006) ‘Processes driving the episodic flux of faecal indicator organisms in streams impacting on recreational and shellfish harvesting waters’, *Water Research*, 40(1), pp. 153–161.

Appendix I

```
# ### Clears environment, sets working directory and imports packages
# rm(list=ls())
# setwd("C:/Users/Niek/Dropbox/MES_Afstuderen/R")
#
# # install.packages("ggplot2")
# # install.packages("forecast")
# # install.packages("tseries")
# # install.packages("dplyr")
# #
# library(ggplot2)
# library(forecast)
# library(tseries)
# library(dplyr)
#
# ### Reads dataframe
# D <- read.csv2("Kabuldata3.csv", header=TRUE, stringsAsFactors = FALSE,
dec = ".")
# View(D)

# # ARIMA
#
# ### Turns column of log10 averages into vector
# listA <- (D$Logaverage)
#
# ### Turns column of log10 averages into vector with mean = 0
# listB <- (D$Logaverage)-mean(D$Logaverage, na.rm = TRUE)
# mean(listB, na.rm = TRUE)
#
# ### Plots ACF of Log10 averages
# acf(listB, na.action = na.pass)
#
# ### Generates ARIMA(1,1,1) model, vector of residuals and ACF of
residuals
# mod1 <- arima(listA, order = c(1,1,1), include.mean = TRUE,
transform.pars = TRUE)
# Mod1res <- (mod1$residuals)
# acf(Mod1res, na.action = na.pass)
#
# ### Generates ARIMA(1,1,1) model etc. of mod1 residuals
# mod2 <- arima(Mod1res, order = c(1,1,1), include.mean = TRUE,
transform.pars = TRUE)
# Mod2res <- (mod2$residuals)
# acf(Mod2res, na.action = na.pass)
#
# ### Generates ARIMA(p,d,q) model, vector of residuals and ACF of
residuals
# mod3 <- arima(listA, order = c(1,0,0), include.mean = TRUE,
transform.pars = TRUE)
# summary(mod3)
# Mod3res <- (mod3$residuals)
# tsdisplay(Mod3res)
#
# ### Generates ARIMA model with automatically selected (p,d,q) values,
vector of residuals and ACF of residuals
# mod4 <- auto.arima(listA, stepwise=FALSE, approximation=FALSE)
# Mod4res <- (mod4$residuals)
```

```
# acf(Mod4res, na.action = na.pass)
# plot(Mod4res)
#
# mod5 <- auto.arima(listB, lambda=0, d=0, D=9, max.order=20, seasonal =
TRUE, stepwise=FALSE, trace = TRUE, approximation=FALSE)
# summary(mod5)
# Mod5res <- (mod5$residuals)
# tsdisplay(Mod5res)
#
# mod6 <- auto.arima(listB, lambda=0, d=1, D=9, max.order=20, seasonal =
TRUE, stepwise=FALSE, trace = TRUE, approximation=FALSE)
# summary(mod6)
# Mod6res <- (mod6$residuals)
# tsdisplay(Mod6res)
#
# mod7 <- auto.arima(listB, lambda=0, d=2, D=9, max.order=20, seasonal =
TRUE, stepwise=FALSE, trace = TRUE, approximation=FALSE)
# summary(mod7)
# Mod7res <- (mod7$residuals)
# tsdisplay(Mod7res)
#
# mod8 <- auto.arima(listB, lambda=0, d=3, D=9, max.order=20, seasonal =
TRUE, stepwise=FALSE, trace = TRUE, approximation=FALSE)
# summary(mod8)
# Mod8res <- (mod8$residuals)
# tsdisplay(Mod8res)
#
# mod9 <- auto.arima(listA, lambda=0, d=4, D=9, max.order=20, seasonal =
TRUE, stepwise=FALSE, trace = TRUE, approximation=FALSE)
# summary(mod9)
# Mod9res <- (mod9$residuals)
# tsdisplay(Mod9res)
#
# ### Calculates Aikake Index Criterion for selected models
# AIC(mod5, mod6, mod7, mod8)
#
# ### Simulates selected model
# # arima.sim(mod5,730)
#
# Mod1res <- (mod1$residuals)
# pacf(Mod1res, lag = 15, na.action = na.pass)
#
# ARIMA300 <- (mod5$residuals)
# pacf(ARIMA300, lag = 10, na.action = na.pass)
# ARIMA015 <- (mod6$residuals)
# pacf(ARIMA015, lag = 10, na.action = na.pass)
# ARIMA520 <- (mod7$residuals)
# pacf(ARIMA520, lag = 10, na.action = na.pass)
# ARIMA033 <- (mod8$residuals)
# pacf(ARIMA033, lag = 10, na.action = na.pass)
```

Appendix II

```
> ### Clears environment, sets working directory and imports packages
> rm(list=ls())
> setwd("C:/Users/Niek/Dropbox/MES_Afstuderen/R")
>
> # install.packages("ggplot2")
> # install.packages("forecast")
> # install.packages("tseries")
> # install.packages("dplyr")
> #
> library(ggplot2)
> library(forecast)
> library(tseries)

    'tseries' version: 0.10-45

    'tseries' is a package for time series analysis and computational finan
ce.

    See 'library(help="tseries")' for details.

> library(dplyr)

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union

>
> ### Reads dataframe
> D <- read.csv2("Kabuldata3.csv", header=TRUE, stringsAsFactors = FALSE, d
ec = ".")
> View(D)
> # ARIMA
>
> ### Turns column of log10 averages into vector
> listA <- (D$Logaverage)
>
> ### Turns column of log10 averages into vector with mean = 0
> listB <- (D$Logaverage)-mean(D$Logaverage, na.rm = TRUE)
> mean(listB, na.rm = TRUE)
[1] 1.131366e-16
> #
> # ### Plots ACF of Log10 averages
> # acf(listB, na.action = na.pass)
> #
> ### Generates ARIMA(1,1,1) model, vector of residuals and ACF of residual
s
> mod1 <- arima(listA, order = c(1,1,1), include.mean = TRUE, transform.par
s = TRUE)
> Mod1res <- (mod1$residuals)
> acf(Mod1res, na.action = na.pass)
> #
> # ### Generates ARIMA(1,1,1) model etc. of mod1 residuals
```

```
> # mod2 <- arima(Mod1res, order = c(1,1,1), include.mean = TRUE, transform
.pars = TRUE)
> # Mod2res <- (mod2$residuals)
> # acf(Mod2res, na.action = na.pass)
> #
> # ### Generates ARIMA(p,d,q) model, vector of residuals and ACF of residu
als
> # mod3 <- arima(listA, order = c(1,0,0), include.mean = TRUE, transform.p
ars = TRUE)
> # summary(mod3)
> # Mod3res <- (mod3$residuals)
> # tsdisplay(Mod3res)
> #
> # ### Generates ARIMA model with automatically selected (p,d,q) values, v
ector of residuals and ACF of residuals
> # mod4 <- auto.arima(listA, stepwise=FALSE, approximation=FALSE)
> # Mod4res <- (mod4$residuals)
> # acf(Mod4res, na.action = na.pass)
> # plot(Mod4res)
>
> mod5 <- auto.arima(listB, lambda=0, d=0, D=9, max.order=20, seasonal = TR
UE, stepwise=FALSE, trace = TRUE, approximation=FALSE)

 ARIMA(0,0,0) with zero mean     : 628.59
 ARIMA(0,0,0) with non-zero mean : 400.3824
 ARIMA(0,0,1) with zero mean     : Inf
 ARIMA(0,0,1) with non-zero mean : 344.5966
 ARIMA(0,0,2) with zero mean     : 494.1731
 ARIMA(0,0,2) with non-zero mean : 334.699
 ARIMA(0,0,3) with zero mean     : 478.124
 ARIMA(0,0,3) with non-zero mean : 328.4945
 ARIMA(0,0,4) with zero mean     : Inf
 ARIMA(0,0,4) with non-zero mean : 315.7001
 ARIMA(0,0,5) with zero mean     : Inf
 ARIMA(0,0,5) with non-zero mean : 313.3025
 ARIMA(1,0,0) with zero mean     : 364.5905
 ARIMA(1,0,0) with non-zero mean : 313.8286
 ARIMA(1,0,1) with zero mean     : Inf
 ARIMA(1,0,1) with non-zero mean : 314.5353
 ARIMA(1,0,2) with zero mean     : Inf
 ARIMA(1,0,2) with non-zero mean : 310.2188
 ARIMA(1,0,3) with zero mean     : Inf
 ARIMA(1,0,3) with non-zero mean : 311.3913
 ARIMA(1,0,4) with zero mean     : Inf
 ARIMA(1,0,4) with non-zero mean : 313.0117
 ARIMA(1,0,5) with zero mean     : Inf
 ARIMA(1,0,5) with non-zero mean : 310.2665
 ARIMA(2,0,0) with zero mean     : 358.0599
 ARIMA(2,0,0) with non-zero mean : 315.287
 ARIMA(2,0,1) with zero mean     : Inf
 ARIMA(2,0,1) with non-zero mean : Inf
 ARIMA(2,0,2) with zero mean     : Inf
 ARIMA(2,0,2) with non-zero mean : 311.7353
 ARIMA(2,0,3) with zero mean     : Inf
 ARIMA(2,0,3) with non-zero mean : Inf
 ARIMA(2,0,4) with zero mean     : Inf
 ARIMA(2,0,4) with non-zero mean : 314.9333
 ARIMA(2,0,5) with zero mean     : Inf
 ARIMA(2,0,5) with non-zero mean : Inf
 ARIMA(3,0,0) with zero mean     : 337.1036
 ARIMA(3,0,0) with non-zero mean : 308.2688
 ARIMA(3,0,1) with zero mean     : Inf
```

```
 ARIMA(3,0,1) with non-zero mean : 310.0165
 ARIMA(3,0,2) with zero mean     : Inf
 ARIMA(3,0,2) with non-zero mean : Inf
 ARIMA(3,0,3) with zero mean     : Inf
 ARIMA(3,0,3) with non-zero mean : Inf
 ARIMA(3,0,4) with zero mean     : Inf
 ARIMA(3,0,4) with non-zero mean : Inf
 ARIMA(3,0,5) with zero mean     : Inf
 ARIMA(3,0,5) with non-zero mean : Inf
 ARIMA(4,0,0) with zero mean     : 338.5497
 ARIMA(4,0,0) with non-zero mean : 309.838
 ARIMA(4,0,1) with zero mean     : Inf
 ARIMA(4,0,1) with non-zero mean : 311.7075
 ARIMA(4,0,2) with zero mean     : Inf
 ARIMA(4,0,2) with non-zero mean : Inf
 ARIMA(4,0,3) with zero mean     : Inf
 ARIMA(4,0,3) with non-zero mean : Inf
 ARIMA(4,0,4) with zero mean     : Inf
 ARIMA(4,0,4) with non-zero mean : Inf
 ARIMA(4,0,5) with zero mean     : Inf
 ARIMA(4,0,5) with non-zero mean : Inf
 ARIMA(5,0,0) with zero mean     : 339.8985
 ARIMA(5,0,0) with non-zero mean : 310.967
 ARIMA(5,0,1) with zero mean     : Inf
 ARIMA(5,0,1) with non-zero mean : 312.7615
 ARIMA(5,0,2) with zero mean     : Inf
 ARIMA(5,0,2) with non-zero mean : Inf
 ARIMA(5,0,3) with zero mean     : Inf
 ARIMA(5,0,3) with non-zero mean : Inf
 ARIMA(5,0,4) with zero mean     : Inf
 ARIMA(5,0,4) with non-zero mean : Inf
 ARIMA(5,0,5) with zero mean     : Inf
 ARIMA(5,0,5) with non-zero mean : Inf



 Best model: ARIMA(3,0,0) with non-zero mean

Warning messages:
1: In log(x) : NaNs produced
2: In log(x) : NaNs produced
> summary(mod5)
Series: listB
ARIMA(3,0,0) with non-zero mean
Box Cox transformation: lambda= 0

Coefficients:
         ar1      ar2     ar3     mean
      0.7089  -0.1881  0.2736  -1.6749
s.e.  0.0868   0.1213  0.0868   0.1340

sigma^2 estimated as 0.05385:  log likelihood=-149.1
AIC=308.2   AICc=308.27   BIC=332.3

Training set error measures:
                     ME       RMSE        MAE      MPE     MAPE     MASE
ACF1
Training set 0.03824169 0.08275727 0.06889123 -22.092 56.39965 1.405801 0.2
676559
> Mod5res <- (mod5$residuals)
> tsdisplay(Mod5res)
>
```

```
> mod6 <- auto.arima(listB, lambda=0, d=1, D=9, max.order=20, seasonal = TR
UE, stepwise=FALSE, trace = TRUE, approximation=FALSE)

 ARIMA(0,1,0)                    : 373.6188
 ARIMA(0,1,0) with drift         : 375.6275
 ARIMA(0,1,1)                    : 340.0341
 ARIMA(0,1,1) with drift         : 342.0347
 ARIMA(0,1,2)                    : 314.7577
 ARIMA(0,1,2) with drift         : 316.6321
 ARIMA(0,1,3)                    : 315.8067
 ARIMA(0,1,3) with drift         : 317.6615
 ARIMA(0,1,4)                    : 315.4679
 ARIMA(0,1,4) with drift         : 317.3112
 ARIMA(0,1,5)                    : 311.3966
 ARIMA(0,1,5) with drift         : 313.1709
 ARIMA(1,1,0)                    : 361.872
 ARIMA(1,1,0) with drift         : 363.8851
 ARIMA(1,1,1)                    : 313.3933
 ARIMA(1,1,1) with drift         : Inf
 ARIMA(1,1,2)                    : 314.8522
 ARIMA(1,1,2) with drift         : 316.6702
 ARIMA(1,1,3)                    : Inf
 ARIMA(1,1,3) with drift         : Inf
 ARIMA(1,1,4)                    : Inf
 ARIMA(1,1,4) with drift         : Inf
 ARIMA(1,1,5)                    : 313.1935
 ARIMA(1,1,5) with drift         : Inf
 ARIMA(2,1,0)                    : 336.0249
 ARIMA(2,1,0) with drift         : 338.0409
 ARIMA(2,1,1)                    : 315.3217
 ARIMA(2,1,1) with drift         : 317.0826
 ARIMA(2,1,2)                    : Inf
 ARIMA(2,1,2) with drift         : Inf
 ARIMA(2,1,3)                    : Inf
 ARIMA(2,1,3) with drift         : Inf
 ARIMA(2,1,4)                    : Inf
 ARIMA(2,1,4) with drift         : Inf
 ARIMA(2,1,5)                    : Inf
 ARIMA(2,1,5) with drift         : Inf
 ARIMA(3,1,0)                    : 336.817
 ARIMA(3,1,0) with drift         : 338.8359
 ARIMA(3,1,1)                    : Inf
 ARIMA(3,1,1) with drift         : Inf
 ARIMA(3,1,2)                    : Inf
 ARIMA(3,1,2) with drift         : Inf
 ARIMA(3,1,3)                    : Inf
 ARIMA(3,1,3) with drift         : Inf
 ARIMA(3,1,4)                    : Inf
 ARIMA(3,1,4) with drift         : Inf
 ARIMA(3,1,5)                    : Inf
 ARIMA(3,1,5) with drift         : Inf
 ARIMA(4,1,0)                    : 337.4529
 ARIMA(4,1,0) with drift         : 339.4722
 ARIMA(4,1,1)                    : Inf
 ARIMA(4,1,1) with drift         : Inf
 ARIMA(4,1,2)                    : Inf
 ARIMA(4,1,2) with drift         : Inf
 ARIMA(4,1,3)                    : Inf
 ARIMA(4,1,3) with drift         : Inf
 ARIMA(4,1,4)                    : Inf
 ARIMA(4,1,4) with drift         : Inf
 ARIMA(4,1,5)                    : Inf
```

```
 ARIMA(4,1,5) with drift         : Inf
 ARIMA(5,1,0)                    : 331.4874
 ARIMA(5,1,0) with drift         : 333.5051
 ARIMA(5,1,1)                    : 311.7667
 ARIMA(5,1,1) with drift         : 313.6386
 ARIMA(5,1,2)                    : 313.7385
 ARIMA(5,1,2) with drift         : 315.6066
 ARIMA(5,1,3)                    : Inf
 ARIMA(5,1,3) with drift         : Inf
 ARIMA(5,1,4)                    : Inf
 ARIMA(5,1,4) with drift         : Inf
 ARIMA(5,1,5)                    : Inf
 ARIMA(5,1,5) with drift         : Inf



 Best model: ARIMA(0,1,5)

Warning messages:
1: In log(x) : NaNs produced
2: In log(x) : NaNs produced
> summary(mod6)
Series: listB
ARIMA(0,1,5)
Box Cox transformation: lambda= 0

Coefficients:
          ma1      ma2     ma3     ma4      ma5
      -0.2959  -0.3196  0.0159  0.0055  -0.2813
s.e.   0.0980   0.1122  0.1221  0.1396   0.1183

sigma^2 estimated as 0.05241:  log likelihood=-149.65
AIC=311.3   AICc=311.4   BIC=340.21

Training set error measures:
                     ME       RMSE        MAE       MPE     MAPE     MASE
ACF1
Training set 0.03934389 0.08297339 0.06637039 -10.29936 47.24576 1.35436 0.
217672
> Mod6res <- (mod6$residuals)
> tsdisplay(Mod6res)
>
> mod7 <- auto.arima(listB, lambda=0, d=2, D=9, max.order=20, seasonal = TR
UE, stepwise=FALSE, trace = TRUE, approximation=FALSE)

 ARIMA(0,2,0)                    : 646.8537
 ARIMA(0,2,1)                    : Inf
 ARIMA(0,2,2)                    : Inf
 ARIMA(0,2,3)                    : Inf
 ARIMA(0,2,4)                    : Inf
 ARIMA(0,2,5)                    : Inf
 ARIMA(1,2,0)                    : 574.8638
 ARIMA(1,2,1)                    : Inf
 ARIMA(1,2,2)                    : Inf
 ARIMA(1,2,3)                    : Inf
 ARIMA(1,2,4)                    : Inf
 ARIMA(1,2,5)                    : Inf
 ARIMA(2,2,0)                    : 495.4364
 ARIMA(2,2,1)                    : Inf
 ARIMA(2,2,2)                    : Inf
 ARIMA(2,2,3)                    : Inf
 ARIMA(2,2,4)                    : Inf
```

```
 ARIMA(2,2,5)                    : Inf
 ARIMA(3,2,0)                    : 459.7439
 ARIMA(3,2,1)                    : Inf
 ARIMA(3,2,2)                    : Inf
 ARIMA(3,2,3)                    : Inf
 ARIMA(3,2,4)                    : Inf
 ARIMA(3,2,5)                    : Inf
 ARIMA(4,2,0)                    : 518.2151
 ARIMA(4,2,1)                    : Inf
 ARIMA(4,2,2)                    : Inf
 ARIMA(4,2,3)                    : Inf
 ARIMA(4,2,4)                    : Inf
 ARIMA(4,2,5)                    : Inf
 ARIMA(5,2,0)                    : 449.2359
 ARIMA(5,2,1)                    : Inf
 ARIMA(5,2,2)                    : Inf
 ARIMA(5,2,3)                    : Inf
 ARIMA(5,2,4)                    : Inf
 ARIMA(5,2,5)                    : Inf



 Best model: ARIMA(5,2,0)

Warning messages:
1: In log(x) : NaNs produced
2: In log(x) : NaNs produced
> summary(mod7)
Series: listB
ARIMA(5,2,0)
Box Cox transformation: lambda= 0

Coefficients:
          ar1      ar2      ar3      ar4      ar5
      -1.0305  -1.1858  -0.8956  -0.4916  -0.2674
s.e.   0.1155   0.1143   0.1269   0.1280   0.1137

sigma^2 estimated as 0.06202:  log likelihood=-218.57
AIC=449.14   AICc=449.24   BIC=478.04

Training set error measures:
                      ME        RMSE         MAE       MPE     MAPE     MASE
ACF1
Training set -0.003932565 0.08536613 0.05470083 -12.58154 36.82981 1.11623
0.2691834
> Mod7res <- (mod7$residuals)
> tsdisplay(Mod7res)
>
> mod8 <- auto.arima(listB, lambda=0, d=3, D=9, max.order=20, seasonal = TR
UE, stepwise=FALSE, trace = TRUE, approximation=FALSE)

 ARIMA(0,3,0)                    : 848.7901
 ARIMA(0,3,1)                    : Inf
 ARIMA(0,3,2)                    : Inf
 ARIMA(0,3,3)                    : 447.8347
 ARIMA(0,3,4)                    : 468.8896
 ARIMA(0,3,5)                    : Inf
 ARIMA(1,3,0)                    : 781.7834
 ARIMA(1,3,1)                    : Inf
 ARIMA(1,3,2)                    : 684.5745
 ARIMA(1,3,3)                    : Inf
 ARIMA(1,3,4)                    : 551.9566
```

```
 ARIMA(1,3,5)                    : Inf
 ARIMA(2,3,0)                    : 706.7919
 ARIMA(2,3,1)                    : Inf
 ARIMA(2,3,2)                    : 526.3256
 ARIMA(2,3,3)                    : Inf
 ARIMA(2,3,4)                    : Inf
 ARIMA(2,3,5)                    : Inf
 ARIMA(3,3,0)                    : Inf
 ARIMA(3,3,1)                    : Inf
 ARIMA(3,3,2)                    : 468.0088
 ARIMA(3,3,3)                    : Inf
 ARIMA(3,3,4)                    : Inf
 ARIMA(3,3,5)                    : Inf
 ARIMA(4,3,0)                    : 573.762
 ARIMA(4,3,1)                    : Inf
 ARIMA(4,3,2)                    : 495.2555
 ARIMA(4,3,3)                    : Inf
 ARIMA(4,3,4)                    : Inf
 ARIMA(4,3,5)                    : Inf
 ARIMA(5,3,0)                    : 567.4822
 ARIMA(5,3,1)                    : Inf
 ARIMA(5,3,2)                    : 579.3099
 ARIMA(5,3,3)                    : Inf
 ARIMA(5,3,4)                    : Inf
 ARIMA(5,3,5)                    : Inf



 Best model: ARIMA(0,3,3)

Warning messages:
1: In log(x) : NaNs produced
2: In auto.arima(listB, lambda = 0, d = 3, D = 9, max.order = 20, seasonal
= TRUE,  :
  Having 3 or more differencing operations is not recommended. Please consi
der reducing the total number of differences.
3: In log(x) : NaNs produced
> summary(mod8)
Series: listB
ARIMA(0,3,3)
Box Cox transformation: lambda= 0

Coefficients:
          ma1     ma2      ma3
      -1.8837  0.9227  -0.0346
s.e.   0.0042  0.0101   0.0062

sigma^2 estimated as 0.0645:  log likelihood=-219.9
AIC=447.79   AICc=447.83   BIC=467.05

Training set error measures:
                      ME       RMSE         MAE       MPE     MAPE      MAS
E        ACF1
Training set -0.007743183 0.07806477 0.04818398 -20.55038 39.99469 0.983246
7 -0.07909052
> Mod8res <- (mod8$residuals)
> tsdisplay(Mod8res)
>
> mod9 <- auto.arima(listA, lambda=0, d=4, D=9, max.order=20, seasonal = TR
UE, stepwise=FALSE, trace = TRUE, approximation=FALSE)

 ARIMA(0,4,0)                    : -69.06803
```

```
 ARIMA(0,4,1)                    : Inf
 ARIMA(0,4,2)                    : -596.1111
 ARIMA(0,4,3)                    : -554.0937
 ARIMA(0,4,4)                    : -588.4382
 ARIMA(0,4,5)                    : Inf
 ARIMA(1,4,0)                    : Inf
 ARIMA(1,4,1)                    : Inf
 ARIMA(1,4,2)                    : Inf
 ARIMA(1,4,3)                    : Inf
 ARIMA(1,4,4)                    : Inf
 ARIMA(1,4,5)                    : Inf
 ARIMA(2,4,0)                    : Inf
 ARIMA(2,4,1)                    : Inf
 ARIMA(2,4,2)                    : Inf
 ARIMA(2,4,3)                    : Inf
 ARIMA(2,4,4)                    : Inf
 ARIMA(2,4,5)                    : Inf
 ARIMA(3,4,0)                    : Inf
 ARIMA(3,4,1)                    : Inf
 ARIMA(3,4,2)                    : Inf
 ARIMA(3,4,3)                    : Inf
 ARIMA(3,4,4)                    : Inf
 ARIMA(3,4,5)                    : Inf
 ARIMA(4,4,0)                    : Inf
 ARIMA(4,4,1)                    : Inf
 ARIMA(4,4,2)                    : Inf
 ARIMA(4,4,3)                    : Inf
 ARIMA(4,4,4)                    : Inf
 ARIMA(4,4,5)                    : Inf
 ARIMA(5,4,0)                    : Inf
 ARIMA(5,4,1)                    : Inf
 ARIMA(5,4,2)                    : Inf
 ARIMA(5,4,3)                    : Inf
 ARIMA(5,4,4)                    : Inf
 ARIMA(5,4,5)                    : Inf



 Best model: ARIMA(0,4,2)

Warning message:
In auto.arima(listA, lambda = 0, d = 4, D = 9, max.order = 20, seasonal = T
RUE,  :
  Having 3 or more differencing operations is not recommended. Please consi
der reducing the total number of differences.
> summary(mod9)
Series: listA
ARIMA(0,4,2)
Box Cox transformation: lambda= 0

Coefficients:
          ma1     ma2
      -1.8293  0.9201
s.e.   0.0011  0.0015

sigma^2 estimated as 0.0001193:  log likelihood=301.07
AIC=-596.14   AICc=-596.11   BIC=-581.69

Training set error measures:
                      ME       RMSE         MAE          MPE      MAPE      MAS
E       ACF1
```

```
Training set -0.004592951 0.1039823 0.06613447 -0.09174883 1.383622 1.34954
6 -0.4153647
> Mod9res <- (mod9$residuals)
> tsdisplay(Mod9res)
>
> ### Calculates Aikake Index Criterion for selected models
> AIC(mod5, mod6, mod7, mod8)
     df      AIC
mod5  5 308.2028
mod6  6 311.3040
mod7  6 449.1432
mod8  4 447.7906
Warning message:
In AIC.default(mod5, mod6, mod7, mod8) :
  models are not all fitted to the same number of observations
>
> ### Simulates selected model
> # arima.sim(mod5,730)
>
> Mod1res <- (mod1$residuals)
> pacf(Mod1res, lag = 15, na.action = na.pass)
>
> ARIMA300 <- (mod5$residuals)
> pacf(ARIMA300, lag = 10, na.action = na.pass)
> ARIMA015 <- (mod6$residuals)
> pacf(ARIMA015, lag = 10, na.action = na.pass)
> ARIMA520 <- (mod7$residuals)
> pacf(ARIMA520, lag = 10, na.action = na.pass)
> ARIMA033 <- (mod8$residuals)
> pacf(ARIMA033, lag = 10, na.action = na.pass)
```

Appendix III

```
# ### Clears environment, sets working directory and imports packages
# rm(list=ls())
# setwd("C:/Users/Niek/Dropbox/MES_Afstuderen/R")
#
# # install.packages("ggplot2")
# # install.packages("forecast")
# # install.packages("tseries")
# # install.packages("dplyr")
# #
# library(ggplot2)
# library(forecast)
# library(tseries)
# library(dplyr)
#
# ### Reads dataframe
# D <- read.csv2("Kabuldata3.csv", header=TRUE, stringsAsFactors = FALSE,
dec = ".")
# View(D)

# # # REGRESSION MODELLING
# #
# ### Turns columns from dataframe into lagged strings from t-1 to t-5
# require(dplyr)
#
# Logdischarge1 <- lag(D$Logdischarge, 1)
# Logdischarge2 <- lag(D$Logdischarge, 2)
# Logdischarge3 <- lag(D$Logdischarge, 3)
# Logdischarge4 <- lag(D$Logdischarge, 4)
# Logdischarge5 <- lag(D$Logdischarge, 5)
#
# Avtemp1 <- lag(D$Avtemp, 1)
# Avtemp2 <- lag(D$Avtemp, 2)
# Avtemp3 <- lag(D$Avtemp, 3)
# Avtemp4 <- lag(D$Avtemp, 4)
# Avtemp5 <- lag(D$Avtemp, 5)
#
# Avturb1 <- lag(D$Avturb, 1)
# Avturb2 <- lag(D$Avturb, 2)
# Avturb3 <- lag(D$Avturb, 3)
# Avturb4 <- lag(D$Avturb, 4)
# Avturb5 <- lag(D$Avturb, 5)
#
# Avturb1 <- lag(D$Avturb, 1)
# Avturb2 <- lag(D$Avturb, 2)
# Avturb3 <- lag(D$Avturb, 3)
# Avturb4 <- lag(D$Avturb, 4)
# Avturb5 <- lag(D$Avturb, 5)
#
# Avec1 <- lag(D$Avec, 1)
# Avec2 <- lag(D$Avec, 2)
# Avec3 <- lag(D$Avec, 3)
# Avec4 <- lag(D$Avec, 4)
# Avec5 <- lag(D$Avec, 5)
#
# Avph1 <- lag(D$Avph, 1)
# Avph2 <- lag(D$Avph, 2)
```

```
# Avph3 <- lag(D$Avph, 3)
# Avph4 <- lag(D$Avph, 4)
# Avph5 <- lag(D$Avph, 5)
#
# Avpre1 <- lag(D$Avpre, 1)
# Avpre2 <- lag(D$Avpre, 2)
# Avpre3 <- lag(D$Avpre, 3)
# Avpre4 <- lag(D$Avpre, 4)
# Avpre5 <- lag(D$Avpre, 5)
#
# Totpre1 <- lag(D$Totpre, 1)
# Totpre2 <- lag(D$Totpre, 2)
# Totpre3 <- lag(D$Totpre, 3)
# Totpre4 <- lag(D$Totpre, 4)
# Totpre5 <- lag(D$Totpre, 5)
#
# ### Linear models with lagtimes from 1 to 5 days for discharge,
turbidity, conductivity, acidity and precipitation
# Genlinlag0 <- lm (Logaverage ~ Logdischarge + Avtemp + Avturb  + Avec +
Avph + Avpre , data=D)
# Genlinlag1 <- lm (Logaverage ~ Logdischarge1 + Avtemp1 + Avturb1  + Avec1
+ Avph1 + Avpre1 , data=D)
# Genlinlag2 <- lm (Logaverage ~ Logdischarge2 + Avtemp2 + Avturb2  + Avec2
+ Avph2 + Avpre2 , data=D)
# Genlinlag3 <- lm (Logaverage ~ Logdischarge3 + Avtemp3 + Avturb3  + Avec3
+ Avph3 + Avpre3 , data=D)
# Genlinlag4 <- lm (Logaverage ~ Logdischarge4 + Avtemp4 + Avturb4  + Avec4
+ Avph4 + Avpre4 , data=D)
# Genlinlag5 <- lm (Logaverage ~ Logdischarge5 + Avtemp5 + Avturb5  + Avec5
+ Avph5 + Avpre5 , data=D)
#
# summary(Genlinlag0)
# summary(Genlinlag1)
# summary(Genlinlag2)
# summary(Genlinlag3)
# summary(Genlinlag4)
# summary(Genlinlag5)
#
# ### Attempt at improved model
# Genlinlag0.1 <- lm (Logaverage ~ Logdischarge + Avtemp1 + Avturb5 + Avec3
+ Avph1 + Avpre1, data=D)
# summary(Genlinlag0.1)
#
# ### Regression on individual explanatory variables
# Lagdischarge <- lm (Logaverage ~ Logdischarge, data=D)
# Lagdischarge1 <- lm (Logaverage ~ Logdischarge1, data=D)
# Lagdischarge2 <- lm (Logaverage ~ Logdischarge2, data=D)
# Lagdischarge3 <- lm (Logaverage ~ Logdischarge3, data=D)
# Lagdischarge4 <- lm (Logaverage ~ Logdischarge4, data=D)
# Lagdischarge5 <- lm (Logaverage ~ Logdischarge5, data=D)
#
# Lagtemp <- lm (Logaverage ~ Avtemp, data=D)
# Lagtemp1 <- lm (Logaverage ~ Avtemp1, data=D)
# Lagtemp2 <- lm (Logaverage ~ Avtemp2, data=D)
# Lagtemp3 <- lm (Logaverage ~ Avtemp3, data=D)
# Lagtemp4 <- lm (Logaverage ~ Avtemp4, data=D)
# Lagtemp5 <- lm (Logaverage ~ Avtemp5, data=D)
#
```

```
# Lagturb <- lm (Logaverage ~ Avturb, data=D)
# Lagturb1 <- lm (Logaverage ~ Avturb1, data=D)
# Lagturb2 <- lm (Logaverage ~ Avturb2, data=D)
# Lagturb3 <- lm (Logaverage ~ Avturb3, data=D)
# Lagturb4 <- lm (Logaverage ~ Avturb4, data=D)
# Lagturb5 <- lm (Logaverage ~ Avturb5, data=D)
#
# Lagec <- lm (Logaverage ~ Avec, data=D)
# Lagec1 <- lm (Logaverage ~ Avec1, data=D)
# Lagec2 <- lm (Logaverage ~ Avec2, data=D)
# Lagec3 <- lm (Logaverage ~ Avec3, data=D)
# Lagec4 <- lm (Logaverage ~ Avec4, data=D)
# Lagec5 <- lm (Logaverage ~ Avec5, data=D)
#
# Lagph <- lm (Logaverage ~ Avph, data=D)
# Lagph1 <- lm (Logaverage ~ Avph1, data=D)
# Lagph2 <- lm (Logaverage ~ Avph2, data=D)
# Lagph3 <- lm (Logaverage ~ Avph3, data=D)
# Lagph4 <- lm (Logaverage ~ Avph4, data=D)
# Lagph5 <- lm (Logaverage ~ Avph5, data=D)
#
# Lagpre <- lm (Logaverage ~ Avpre, data=D)
# Lagpre1 <- lm (Logaverage ~ Avpre1, data=D)
# Lagpre2 <- lm (Logaverage ~ Avpre2, data=D)
# Lagpre3 <- lm (Logaverage ~ Avpre3, data=D)
# Lagpre4 <- lm (Logaverage ~ Avpre4, data=D)
# Lagpre5 <- lm (Logaverage ~ Avpre5, data=D)
#
# summary(Lagdischarge)$adj.r.squared
# summary(Lagdischarge1)$adj.r.squared
# summary(Lagdischarge2)$adj.r.squared
# summary(Lagdischarge3)$adj.r.squared
# summary(Lagdischarge4)$adj.r.squared
# summary(Lagdischarge5)$adj.r.squared
#
# summary(Lagtemp)$adj.r.squared
# summary(Lagtemp1)$adj.r.squared
# summary(Lagtemp2)$adj.r.squared
# summary(Lagtemp3)$adj.r.squared
# summary(Lagtemp4)$adj.r.squared
# summary(Lagtemp5)$adj.r.squared
#
# summary(Lagturb)$adj.r.squared
# summary(Lagturb1)$adj.r.squared
# summary(Lagturb2)$adj.r.squared
# summary(Lagturb3)$adj.r.squared
# summary(Lagturb4)$adj.r.squared
# summary(Lagturb5)$adj.r.squared
#
# summary(Lagec)$adj.r.squared
# summary(Lagec1)$adj.r.squared
# summary(Lagec2)$adj.r.squared
# summary(Lagec3)$adj.r.squared
# summary(Lagec4)$adj.r.squared
# summary(Lagec5)$adj.r.squared
#
# summary(Lagph)$adj.r.squared
# summary(Lagph1)$adj.r.squared
```

```
# summary(Lagph2)$adj.r.squared
# summary(Lagph3)$adj.r.squared
# summary(Lagph4)$adj.r.squared
# summary(Lagph5)$adj.r.squared
#
# summary(Lagpre)$adj.r.squared
# summary(Lagpre1)$adj.r.squared
# summary(Lagpre2)$adj.r.squared
# summary(Lagpre3)$adj.r.squared
# summary(Lagpre4)$adj.r.squared
# summary(Lagpre5)$adj.r.squared
#
# ### Second attempt at improved model
# Genlinlag0.2 <- lm (Logaverage ~ Logdischarge + Avtemp + Avturb5 + Avec +
Avph + Avpre2, data=D)
# Genlinlag0.2.1 <- lm (Logaverage ~ Logdischarge + Avtemp + Avturb5 + Avec
+ Avph, data=D)
# Genlinlag0.2.2 <- lm (Logaverage ~ Logdischarge + Avturb5 + Avec + Avph,
data=D)
# Genlinlag0.2.3 <- lm (Logaverage ~ Logdischarge + Avturb5 + Avph, data=D)
# Genlinlag0.2.4 <- lm (Logaverage ~ Logdischarge + Avturb5, data=D)
#
# summary(Genlinlag0.2)$adj.r.squared
# summary(Genlinlag0.2.1)$adj.r.squared
# summary(Genlinlag0.2.2)$adj.r.squared
# summary(Genlinlag0.2.3)$adj.r.squared
# summary(Genlinlag0.2.4)$adj.r.squared
#
# GenlinlagA <- lm (Logaverage ~ Logdischarge + Avtemp + Avtemp1 + Avtemp2
+ Avtemp3 + Avec + Avec1 + Avec2 + Avec3 + Avph + Avph1 + Avph2 + Avpre +
Avpre2, data=D)
# summary(GenlinlagA)
#
# GenlinlagB <- lm (Logaverage ~ Avtemp3 + Avec3 + Avph, data=D)
# summary(GenlinlagB)
#
# GenlinlagC <- lm (Logaverage ~ Logdischarge + Avtemp + Avtemp1 + Avtemp2
+ Avtemp3 + Avtemp4 + Avtemp5 + Avec + Avec1 + Avec2 + Avec3 + Avec4 +
Avec5 + Avph + Avph1 + Avph2 + Avph4 + Avph5 + Avpre + Avpre1 + Avpre2 +
Avpre3 + Avpre4 + Avpre5, data=D)
# summary(GenlinlagC)
#
# GenlinlagD <- lm (Logaverage ~ Avec3 + Avph, data=D)
# summary(GenlinlagD)
#
# GenlinlagE <- lm (Logaverage ~ Avtemp3 + Avec3 + Avph, data=D)
# summary(GenlinlagE)
#
# GenlinlagF <- lm (Logaverage ~ Logdischarge4 + Avec3 + Avph, data=D)
# summary(GenlinlagF)
#
#
# # Testing for interactions
#
# GenlinlagE <- lm (Logaverage ~ Avtemp3 + Avec3 + Avph, data=D)
# summary(GenlinlagE)
#
```

```
# GenlinlagG <- lm (Logaverage ~ Avtemp3 + Avec3 + Avph + Avtemp3*Avec3,
data=D)
# summary(GenlinlagG)
#
# GenlinlagH <- lm (Logaverage ~ Avtemp3 + Avec3 + Avph + Avtemp3*Avph,
data=D)
# summary(GenlinlagH)
#
# GenlinlagJ <- lm (Logaverage ~ Avtemp3 + Avec3 + Avph + Avec3*Avph,
data=D)
# summary(GenlinlagJ)
#
#
# # Testing for collinearity
#
# ColT1 <- lm (Avtemp3 ~ Avec3, data=D)
# summary(ColT1)
#
# ColT2 <- lm (Avtemp3 ~ Avph, data=D)
# summary(ColT2)
#
# ColT3 <- lm (Avec3 ~ Avph, data=D)
# summary(ColT3)
```

Appendix IV

```
> # ### Clears environment, sets working directory and imports packages
> rm(list=ls())
> setwd("C:/Users/Niek/Dropbox/MES_Afstuderen/R")
> #
> # # install.packages("ggplot2")
> # # install.packages("forecast")
> # # install.packages("tseries")
> # # install.packages("dplyr")
> #
> library(ggplot2)
> library(forecast)
> library(tseries)

    'tseries' version: 0.10-45

    'tseries' is a package for time series analysis and computational finan
ce.

    See 'library(help="tseries")' for details.

> library(dplyr)

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union

>
> ### Reads dataframe
> D <- read.csv2("Kabuldata3.csv", header=TRUE, stringsAsFactors = FALSE, d
ec = ".")
> View(D)
> # # REGRESSION MODELLING
> #
> ### Turns columns from dataframe into lagged strings from t-1 to t-5
> require(dplyr)
>
> Logdischarge1 <- lag(D$Logdischarge, 1)
> Logdischarge2 <- lag(D$Logdischarge, 2)
> Logdischarge3 <- lag(D$Logdischarge, 3)
> Logdischarge4 <- lag(D$Logdischarge, 4)
> Logdischarge5 <- lag(D$Logdischarge, 5)
>
> Avtemp1 <- lag(D$Avtemp, 1)
> Avtemp2 <- lag(D$Avtemp, 2)
> Avtemp3 <- lag(D$Avtemp, 3)
> Avtemp4 <- lag(D$Avtemp, 4)
> Avtemp5 <- lag(D$Avtemp, 5)
>
> Avturb1 <- lag(D$Avturb, 1)
> Avturb2 <- lag(D$Avturb, 2)
> Avturb3 <- lag(D$Avturb, 3)
> Avturb4 <- lag(D$Avturb, 4)
> Avturb5 <- lag(D$Avturb, 5)
```

```
>
> Avturb1 <- lag(D$Avturb, 1)
> Avturb2 <- lag(D$Avturb, 2)
> Avturb3 <- lag(D$Avturb, 3)
> Avturb4 <- lag(D$Avturb, 4)
> Avturb5 <- lag(D$Avturb, 5)
>
> Avec1 <- lag(D$Avec, 1)
> Avec2 <- lag(D$Avec, 2)
> Avec3 <- lag(D$Avec, 3)
> Avec4 <- lag(D$Avec, 4)
> Avec5 <- lag(D$Avec, 5)
>
> Avph1 <- lag(D$Avph, 1)
> Avph2 <- lag(D$Avph, 2)
> Avph3 <- lag(D$Avph, 3)
> Avph4 <- lag(D$Avph, 4)
> Avph5 <- lag(D$Avph, 5)
>
> Avpre1 <- lag(D$Avpre, 1)
> Avpre2 <- lag(D$Avpre, 2)
> Avpre3 <- lag(D$Avpre, 3)
> Avpre4 <- lag(D$Avpre, 4)
> Avpre5 <- lag(D$Avpre, 5)
>
> Totpre1 <- lag(D$Totpre, 1)
> Totpre2 <- lag(D$Totpre, 2)
> Totpre3 <- lag(D$Totpre, 3)
> Totpre4 <- lag(D$Totpre, 4)
> Totpre5 <- lag(D$Totpre, 5)
>
> ### Linear models with lagtimes from 1 to 5 days for discharge, turbidity
, conductivity, acidity and precipitation
> Genlinlag0 <- lm (Logaverage ~ Logdischarge + Avtemp + Avturb  + Avec + A
vph + Avpre , data=D)
> Genlinlag1 <- lm (Logaverage ~ Logdischarge1 + Avtemp1 + Avturb1  + Avec1
+ Avph1 + Avpre1 , data=D)
> Genlinlag2 <- lm (Logaverage ~ Logdischarge2 + Avtemp2 + Avturb2  + Avec2
+ Avph2 + Avpre2 , data=D)
> Genlinlag3 <- lm (Logaverage ~ Logdischarge3 + Avtemp3 + Avturb3  + Avec3
+ Avph3 + Avpre3 , data=D)
> Genlinlag4 <- lm (Logaverage ~ Logdischarge4 + Avtemp4 + Avturb4  + Avec4
+ Avph4 + Avpre4 , data=D)
> Genlinlag5 <- lm (Logaverage ~ Logdischarge5 + Avtemp5 + Avturb5  + Avec5
+ Avph5 + Avpre5 , data=D)
>
> summary(Genlinlag0)

Call:
lm(formula = Logaverage ~ Logdischarge + Avtemp + Avturb + Avec +
    Avph + Avpre, data = D)

Residuals:
     Min       1Q   Median       3Q      Max
-0.76558 -0.16605  0.03612  0.14975  0.69670

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.2988468  1.1795732  -5.340 2.79e-07 ***
Logdischarge  0.1673380  0.1396122   1.199  0.23227
Avtemp        0.0265333  0.0088290   3.005  0.00303 **
Avturb        0.0001325  0.0001967   0.674  0.50130
```

```
Avec         -0.0008523  0.0008185  -1.041  0.29914
Avph          1.1962792  0.1286149   9.301  < 2e-16 ***
Avpre         0.0088548  0.0029419   3.010  0.00299 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2572 on 179 degrees of freedom
  (729 observations deleted due to missingness)
Multiple R-squared:  0.6873,   Adjusted R-squared:  0.6768
F-statistic: 65.57 on 6 and 179 DF,  p-value: < 2.2e-16

> summary(Genlinlag1)

Call:
lm(formula = Logaverage ~ Logdischarge1 + Avtemp1 + Avturb1 +
    Avec1 + Avph1 + Avpre1, data = D)

Residuals:
     Min       1Q   Median       3Q      Max
-0.88615 -0.18096  0.04516  0.25037  0.57920

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.6857825  0.8557321   4.307 3.66e-05 ***
Logdischarge1  0.3129381  0.2514463   1.245   0.2160
Avtemp1        0.0015285  0.0136310   0.112   0.9109
Avturb1        0.0005129  0.0003080   1.665   0.0988 .
Avec1         -0.0019382  0.0013301  -1.457   0.1480
Avph1          0.0270447  0.0658696   0.411   0.6822
Avpre1         0.0082567  0.0046462   1.777   0.0784 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3236 on 108 degrees of freedom
  (800 observations deleted due to missingness)
Multiple R-squared:  0.3557,   Adjusted R-squared:  0.3199
F-statistic: 9.936 on 6 and 108 DF,  p-value: 1.023e-08

> summary(Genlinlag2)

Call:
lm(formula = Logaverage ~ Logdischarge2 + Avtemp2 + Avturb2 +
    Avec2 + Avph2 + Avpre2, data = D)

Residuals:
     Min       1Q   Median       3Q      Max
-0.93104 -0.13128  0.02931  0.21847  0.54033

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.6318240  0.9425170   3.853 0.000229 ***
Logdischarge2  0.3433238  0.2793731   1.229 0.222579
Avtemp2        0.0054932  0.0149915   0.366 0.714983
Avturb2        0.0007262  0.0003272   2.219 0.029187 *
Avec2         -0.0016611  0.0014837  -1.120 0.266118
Avph2         -0.0066113  0.0707402  -0.093 0.925764
Avpre2         0.0077601  0.0059347   1.308 0.194622
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.295 on 83 degrees of freedom
  (825 observations deleted due to missingness)
```

```
Multiple R-squared:  0.3557,   Adjusted R-squared:  0.3092
F-statistic: 7.638 on 6 and 83 DF,  p-value: 1.519e-06

> summary(Genlinlag3)

Call:
lm(formula = Logaverage ~ Logdischarge3 + Avtemp3 + Avturb3 +
    Avec3 + Avph3 + Avpre3, data = D)

Residuals:
     Min       1Q   Median       3Q      Max
-0.82620 -0.12079  0.00884  0.15740  0.51757

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.1262575  1.0109996   5.070 3.04e-06 ***
Logdischarge3 -0.1373361  0.2880159  -0.477  0.63494
Avtemp3       -0.0075064  0.0157110  -0.478  0.63428
Avturb3        0.0011854  0.0003554   3.336  0.00136 **
Avec3         -0.0029261  0.0014747  -1.984  0.05110 .
Avph3          0.0418448  0.0703447   0.595  0.55383
Avpre3         0.0025506  0.0075282   0.339  0.73575
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2667 on 71 degrees of freedom
  (837 observations deleted due to missingness)
Multiple R-squared:  0.481,    Adjusted R-squared:  0.4371
F-statistic: 10.97 on 6 and 71 DF,  p-value: 1.302e-08

> summary(Genlinlag4)

Call:
lm(formula = Logaverage ~ Logdischarge4 + Avtemp4 + Avturb4 +
    Avec4 + Avph4 + Avpre4, data = D)

Residuals:
     Min       1Q   Median       3Q      Max
-0.83379 -0.11677  0.00124  0.14631  0.50545

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.571e+00  1.075e+00   4.251 7.67e-05 ***
Logdischarge4  3.119e-02  3.089e-01   0.101  0.91991
Avtemp4       -7.847e-06  1.834e-02   0.000  0.99966
Avturb4        1.102e-03  3.863e-04   2.852  0.00598 **
Avec4         -1.766e-03  1.526e-03  -1.157  0.25176
Avph4          2.701e-03  7.654e-02   0.035  0.97197
Avpre4         7.690e-03  1.306e-02   0.589  0.55815
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2586 on 59 degrees of freedom
  (849 observations deleted due to missingness)
Multiple R-squared:  0.5404,   Adjusted R-squared:  0.4937
F-statistic: 11.56 on 6 and 59 DF,  p-value: 1.624e-08

> summary(Genlinlag5)

Call:
lm(formula = Logaverage ~ Logdischarge5 + Avtemp5 + Avturb5 +
    Avec5 + Avph5 + Avpre5, data = D)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.60168 -0.11671  0.00459  0.11796  0.57959

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.2117263  1.1778384   5.274 3.01e-06 ***
Logdischarge5 -0.4028138  0.3407667  -1.182    0.243
Avtemp5       -0.0171499  0.0174849  -0.981    0.331
Avturb5        0.0017384  0.0004047   4.295 8.24e-05 ***
Avec5         -0.0029773  0.0015526  -1.918    0.061 .
Avph5          0.0071381  0.0714216   0.100    0.921
Avpre5         0.0067661  0.0074121   0.913    0.366
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2307 on 49 degrees of freedom
  (859 observations deleted due to missingness)
Multiple R-squared:  0.6183,   Adjusted R-squared:  0.5716
F-statistic: 13.23 on 6 and 49 DF,  p-value: 7.662e-09

>
> ### Attempt at improved model
> Genlinlag0.1 <- lm (Logaverage ~ Logdischarge + Avtemp1 + Avturb5 + Avec3
+ Avph1 + Avpre1, data=D)
> summary(Genlinlag0.1)

Call:
lm(formula = Logaverage ~ Logdischarge + Avtemp1 + Avturb5 +
    Avec3 + Avph1 + Avpre1, data = D)

Residuals:
     Min       1Q   Median       3Q      Max
-0.67402 -0.09755  0.05441  0.13808  0.47488

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.5187997  1.2802425   1.186  0.24326
Logdischarge  0.8122822  0.3568566   2.276  0.02888 *
Avtemp1       0.0055936  0.0196641   0.284  0.77769
Avturb5       0.0009746  0.0003286   2.966  0.00533 **
Avec3         0.0008497  0.0010952   0.776  0.44290
Avph1        -0.0149238  0.0529796  -0.282  0.77980
Avpre1       -0.0009642  0.0121240  -0.080  0.93705
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2281 on 36 degrees of freedom
  (872 observations deleted due to missingness)
Multiple R-squared:  0.3795,   Adjusted R-squared:  0.276
F-statistic: 3.669 on 6 and 36 DF,  p-value: 0.00604

>
> ### Regression on individual explanatory variables
> Lagdischarge <- lm (Logaverage ~ Logdischarge, data=D)
> Lagdischarge1 <- lm (Logaverage ~ Logdischarge1, data=D)
> Lagdischarge2 <- lm (Logaverage ~ Logdischarge2, data=D)
> Lagdischarge3 <- lm (Logaverage ~ Logdischarge3, data=D)
> Lagdischarge4 <- lm (Logaverage ~ Logdischarge4, data=D)
> Lagdischarge5 <- lm (Logaverage ~ Logdischarge5, data=D)
>
```

```
> Lagtemp <- lm (Logaverage ~ Avtemp, data=D)
> Lagtemp1 <- lm (Logaverage ~ Avtemp1, data=D)
> Lagtemp2 <- lm (Logaverage ~ Avtemp2, data=D)
> Lagtemp3 <- lm (Logaverage ~ Avtemp3, data=D)
> Lagtemp4 <- lm (Logaverage ~ Avtemp4, data=D)
> Lagtemp5 <- lm (Logaverage ~ Avtemp5, data=D)
>
> Lagturb <- lm (Logaverage ~ Avturb, data=D)
> Lagturb1 <- lm (Logaverage ~ Avturb1, data=D)
> Lagturb2 <- lm (Logaverage ~ Avturb2, data=D)
> Lagturb3 <- lm (Logaverage ~ Avturb3, data=D)
> Lagturb4 <- lm (Logaverage ~ Avturb4, data=D)
> Lagturb5 <- lm (Logaverage ~ Avturb5, data=D)
>
> Lagec <- lm (Logaverage ~ Avec, data=D)
> Lagec1 <- lm (Logaverage ~ Avec1, data=D)
> Lagec2 <- lm (Logaverage ~ Avec2, data=D)
> Lagec3 <- lm (Logaverage ~ Avec3, data=D)
> Lagec4 <- lm (Logaverage ~ Avec4, data=D)
> Lagec5 <- lm (Logaverage ~ Avec5, data=D)
>
> Lagph <- lm (Logaverage ~ Avph, data=D)
> Lagph1 <- lm (Logaverage ~ Avph1, data=D)
> Lagph2 <- lm (Logaverage ~ Avph2, data=D)
> Lagph3 <- lm (Logaverage ~ Avph3, data=D)
> Lagph4 <- lm (Logaverage ~ Avph4, data=D)
> Lagph5 <- lm (Logaverage ~ Avph5, data=D)
>
> Lagpre <- lm (Logaverage ~ Avpre, data=D)
> Lagpre1 <- lm (Logaverage ~ Avpre1, data=D)
> Lagpre2 <- lm (Logaverage ~ Avpre2, data=D)
> Lagpre3 <- lm (Logaverage ~ Avpre3, data=D)
> Lagpre4 <- lm (Logaverage ~ Avpre4, data=D)
> Lagpre5 <- lm (Logaverage ~ Avpre5, data=D)
>
> summary(Lagdischarge)$adj.r.squared
[1] 0.466879
> summary(Lagdischarge1)$adj.r.squared
[1] 0.4362613
> summary(Lagdischarge2)$adj.r.squared
[1] 0.4432576
> summary(Lagdischarge3)$adj.r.squared
[1] 0.4546074
> summary(Lagdischarge4)$adj.r.squared
[1] 0.4658926
> summary(Lagdischarge5)$adj.r.squared
[1] 0.4609687
>
> summary(Lagtemp)$adj.r.squared
[1] 0.3179971
> summary(Lagtemp1)$adj.r.squared
[1] 0.08923104
> summary(Lagtemp2)$adj.r.squared
[1] 0.04771418
> summary(Lagtemp3)$adj.r.squared
[1] 0.1359708
> summary(Lagtemp4)$adj.r.squared
[1] 0.1939363
> summary(Lagtemp5)$adj.r.squared
[1] 0.1194126
>
> summary(Lagturb)$adj.r.squared
```

```
[1] 0.45203
> summary(Lagturb1)$adj.r.squared
[1] 0.2160522
> summary(Lagturb2)$adj.r.squared
[1] 0.1864722
> summary(Lagturb3)$adj.r.squared
[1] 0.380731
> summary(Lagturb4)$adj.r.squared
[1] 0.447264
> summary(Lagturb5)$adj.r.squared
[1] 0.4711694
>
> summary(Lagec)$adj.r.squared
[1] 0.4515687
> summary(Lagec1)$adj.r.squared
[1] 0.1810171
> summary(Lagec2)$adj.r.squared
[1] 0.1324311
> summary(Lagec3)$adj.r.squared
[1] 0.2465301
> summary(Lagec4)$adj.r.squared
[1] 0.3170176
> summary(Lagec5)$adj.r.squared
[1] 0.3243358
>
> summary(Lagph)$adj.r.squared
[1] 0.4662325
> summary(Lagph1)$adj.r.squared
[1] -0.002852279
> summary(Lagph2)$adj.r.squared
[1] -0.007014022
> summary(Lagph3)$adj.r.squared
[1] 0.0002710402
> summary(Lagph4)$adj.r.squared
[1] -0.0009746309
> summary(Lagph5)$adj.r.squared
[1] -0.01351399
>
> summary(Lagpre)$adj.r.squared
[1] 0.01987085
> summary(Lagpre1)$adj.r.squared
[1] 0.02508925
> summary(Lagpre2)$adj.r.squared
[1] 0.02547138
> summary(Lagpre3)$adj.r.squared
[1] 0.01463716
> summary(Lagpre4)$adj.r.squared
[1] 0.01872139
> summary(Lagpre5)$adj.r.squared
[1] 0.01908308
>
> ### Second attempt at improved model
> Genlinlag0.2 <- lm (Logaverage ~ Logdischarge + Avtemp + Avturb5 + Avec +
Avph + Avpre2, data=D)
> Genlinlag0.2.1 <- lm (Logaverage ~ Logdischarge + Avtemp + Avturb5 + Avec
+ Avph, data=D)
> Genlinlag0.2.2 <- lm (Logaverage ~ Logdischarge + Avturb5 + Avec + Avph,
data=D)
> Genlinlag0.2.3 <- lm (Logaverage ~ Logdischarge + Avturb5 + Avph, data=D)
> Genlinlag0.2.4 <- lm (Logaverage ~ Logdischarge + Avturb5, data=D)
>
> summary(Genlinlag0.2)$adj.r.squared
```

```
[1] 0.6235025
> summary(Genlinlag0.2.1)$adj.r.squared
[1] 0.6229372
> summary(Genlinlag0.2.2)$adj.r.squared
[1] 0.6289957
> summary(Genlinlag0.2.3)$adj.r.squared
[1] 0.62958
> summary(Genlinlag0.2.4)$adj.r.squared
[1] 0.5338101
>
> GenlinlagA <- lm (Logaverage ~ Logdischarge + Avtemp + Avtemp1 + Avtemp2
+ Avtemp3 + Avec + Avec1 + Avec2 + Avec3 + Avph + Avph1 + Avph2 + Avpre + A
vpre2, data=D)
> summary(GenlinlagA)

Call:
lm(formula = Logaverage ~ Logdischarge + Avtemp + Avtemp1 + Avtemp2 +
    Avtemp3 + Avec + Avec1 + Avec2 + Avec3 + Avph + Avph1 + Avph2 +
    Avpre + Avpre2, data = D)

Residuals:
     Min       1Q   Median       3Q      Max
-0.26673 -0.11216  0.01793  0.08896  0.31375

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -11.355819   1.451037  -7.826 2.12e-10 ***
Logdischarge   0.215767   0.191944   1.124  0.26603
Avtemp        -0.037842   0.025829  -1.465  0.14880
Avtemp1        0.008871   0.025839   0.343  0.73272
Avtemp2        0.012631   0.025142   0.502  0.61747
Avtemp3        0.052602   0.018802   2.798  0.00716 **
Avec          -0.004380   0.003557  -1.231  0.22358
Avec1          0.004467   0.003538   1.263  0.21223
Avec2          0.003065   0.002770   1.106  0.27352
Avec3         -0.005106   0.001943  -2.628  0.01122 *
Avph           2.010152   0.236471   8.501 1.79e-11 ***
Avph1         -0.134338   0.110223  -1.219  0.22832
Avph2         -0.090613   0.083992  -1.079  0.28555
Avpre          0.003572   0.003548   1.007  0.31867
Avpre2         0.007684   0.004054   1.895  0.06349 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1522 on 53 degrees of freedom
  (847 observations deleted due to missingness)
Multiple R-squared:  0.8264,   Adjusted R-squared:  0.7806
F-statistic: 18.03 on 14 and 53 DF,  p-value: 2.346e-15


>
> GenlinlagB <- lm (Logaverage ~ Avtemp3 + Avec3 + Avph, data=D)
> summary(GenlinlagB)

Call:
lm(formula = Logaverage ~ Avtemp3 + Avec3 + Avph, data = D)

Residuals:
     Min       1Q   Median       3Q      Max
-0.45877 -0.12528  0.02858  0.10674  0.43843

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) -9.520410   1.281876  -7.427 1.79e-10 ***
Avtemp3      0.028219   0.005474   5.155 2.14e-06 ***
Avec3       -0.002134   0.000384  -5.556 4.39e-07 ***
Avph         1.682967   0.148621  11.324  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1779 on 72 degrees of freedom
  (839 observations deleted due to missingness)
Multiple R-squared:  0.7657,   Adjusted R-squared:  0.756
F-statistic: 78.45 on 3 and 72 DF,  p-value: < 2.2e-16

>
> GenlinlagC <- lm (Logaverage ~ Logdischarge + Avtemp + Avtemp1 + Avtemp2
+ Avtemp3 + Avtemp4 + Avtemp5 + Avec + Avec1 + Avec2 + Avec3 + Avec4 + Avec
5 + Avph + Avph1 + Avph2 + Avph4 + Avph5 + Avpre + Avpre1 + Avpre2 + Avpre3
+ Avpre4 + Avpre5, data=D)
> summary(GenlinlagC)

Call:
lm(formula = Logaverage ~ Logdischarge + Avtemp + Avtemp1 + Avtemp2 +
    Avtemp3 + Avtemp4 + Avtemp5 + Avec + Avec1 + Avec2 + Avec3 +
    Avec4 + Avec5 + Avph + Avph1 + Avph2 + Avph4 + Avph5 + Avpre +
    Avpre1 + Avpre2 + Avpre3 + Avpre4 + Avpre5, data = D)

Residuals:
      Min        1Q    Median        3Q       Max
-0.129110 -0.035704 -0.002567  0.036983  0.120274

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.5444338  2.3074737  -2.836  0.01140 *
Logdischarge  0.1630338  0.3736068   0.436  0.66805
Avtemp       -0.1225258  0.0364475  -3.362  0.00370 **
Avtemp1      -0.0267245  0.0244076  -1.095  0.28882
Avtemp2      -0.0338663  0.0272837  -1.241  0.23136
Avtemp3      -0.0310135  0.0276889  -1.120  0.27826
Avtemp4       0.1109997  0.0467651   2.374  0.02967 *
Avtemp5       0.0778961  0.0385019   2.023  0.05907 .
Avec         -0.0022528  0.0033294  -0.677  0.50774
Avec1         0.0027791  0.0036828   0.755  0.46081
Avec2         0.0011861  0.0036549   0.325  0.74951
Avec3         0.0043122  0.0029237   1.475  0.15851
Avec4         0.0011779  0.0024154   0.488  0.63201
Avec5        -0.0084766  0.0025176  -3.367  0.00366 **
Avph          1.5885399  0.2748998   5.779 2.23e-05 ***
Avph1        -0.0197030  0.0982850  -0.200  0.84349
Avph2         0.0552974  0.0942941   0.586  0.56529
Avph4        -0.2643007  0.1205344  -2.193  0.04253 *
Avph5         0.0167529  0.1135031   0.148  0.88440
Avpre         0.0005889  0.0071801   0.082  0.93559
Avpre1       -0.0066024  0.0059791  -1.104  0.28488
Avpre2        0.0163227  0.0057630   2.832  0.01150 *
Avpre3        0.0002434  0.0064101   0.038  0.97016
Avpre4        0.0276087  0.0090390   3.054  0.00717 **
Avpre5       -0.0008053  0.0120393  -0.067  0.94745
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0897 on 17 degrees of freedom
  (873 observations deleted due to missingness)
Multiple R-squared:  0.9547,   Adjusted R-squared:  0.8907
```

```
F-statistic: 14.92 on 24 and 17 DF,  p-value: 2.736e-07

>
> GenlinlagD <- lm (Logaverage ~ Avec3 + Avph, data=D)
> summary(GenlinlagD)

Call:
lm(formula = Logaverage ~ Avec3 + Avph, data = D)

Residuals:
     Min       1Q   Median       3Q      Max
-0.55790 -0.12578  0.03252  0.13351  0.35564

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.8716127  1.4823729  -5.985 7.43e-08 ***
Avec3       -0.0024925  0.0004388  -5.680 2.58e-07 ***
Avph         1.6836923  0.1727009   9.749 7.44e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2067 on 73 degrees of freedom
  (839 observations deleted due to missingness)
Multiple R-squared:  0.6793,   Adjusted R-squared:  0.6705
F-statistic:  77.3 on 2 and 73 DF,  p-value: < 2.2e-16

>
> GenlinlagE <- lm (Logaverage ~ Avtemp3 + Avec3 + Avph, data=D)
> summary(GenlinlagE)

Call:
lm(formula = Logaverage ~ Avtemp3 + Avec3 + Avph, data = D)

Residuals:
     Min       1Q   Median       3Q      Max
-0.45877 -0.12528  0.02858  0.10674  0.43843

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.520410   1.281876  -7.427 1.79e-10 ***
Avtemp3      0.028219   0.005474   5.155 2.14e-06 ***
Avec3       -0.002134   0.000384  -5.556 4.39e-07 ***
Avph         1.682967   0.148621  11.324  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1779 on 72 degrees of freedom
  (839 observations deleted due to missingness)
Multiple R-squared:  0.7657,   Adjusted R-squared:  0.756
F-statistic: 78.45 on 3 and 72 DF,  p-value: < 2.2e-16

>
> GenlinlagF <- lm (Logaverage ~ Logdischarge4 + Avec3 + Avph, data=D)
> summary(GenlinlagF)

Call:
lm(formula = Logaverage ~ Logdischarge4 + Avec3 + Avph, data = D)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5611 -0.1215  0.0337  0.1364  0.3514
```

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -8.4524960  1.4958872  -5.650 3.10e-07 ***
Logdischarge4 -0.0357119  0.1716302  -0.208 0.835766
Avec3         -0.0025248  0.0007115  -3.548 0.000692 ***
Avph           1.6489773  0.1896486   8.695 8.61e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2056 on 71 degrees of freedom
  (840 observations deleted due to missingness)
Multiple R-squared:  0.6621,   Adjusted R-squared:  0.6479
F-statistic: 46.38 on 3 and 71 DF,  p-value: < 2.2e-16
> # Testing for interactions
>
> GenlinlagE <- lm (Logaverage ~ Avtemp3 + Avec3 + Avph, data=D)
> summary(GenlinlagE)

Call:
lm(formula = Logaverage ~ Avtemp3 + Avec3 + Avph, data = D)

Residuals:
     Min       1Q   Median       3Q      Max
-0.45877 -0.12528  0.02858  0.10674  0.43843

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.520410   1.281876  -7.427 1.79e-10 ***
Avtemp3      0.028219   0.005474   5.155 2.14e-06 ***
Avec3       -0.002134   0.000384  -5.556 4.39e-07 ***
Avph         1.682967   0.148621  11.324  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1779 on 72 degrees of freedom
  (839 observations deleted due to missingness)
Multiple R-squared:  0.7657,   Adjusted R-squared:  0.756
F-statistic: 78.45 on 3 and 72 DF,  p-value: < 2.2e-16

>
> GenlinlagG <- lm (Logaverage ~ Avtemp3 + Avec3 + Avph + Avtemp3*Avec3, da
ta=D)
> summary(GenlinlagG)

Call:
lm(formula = Logaverage ~ Avtemp3 + Avec3 + Avph + Avtemp3 *
    Avec3, data = D)

Residuals:
     Min       1Q   Median       3Q      Max
-0.45680 -0.12474  0.02898  0.10692  0.43735

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -9.529e+00  1.299e+00  -7.336 2.83e-10 ***
Avtemp3        2.878e-02  1.058e-02   2.720   0.0082 **
Avec3         -2.078e-03  9.639e-04  -2.156   0.0344 *
Avph           1.684e+00  1.499e-01  11.230  < 2e-16 ***
Avtemp3:Avec3 -4.396e-06  7.039e-05  -0.062   0.9504
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 0.1791 on 71 degrees of freedom
  (839 observations deleted due to missingness)
Multiple R-squared:  0.7657,   Adjusted R-squared:  0.7525
F-statistic: 58.02 on 4 and 71 DF,  p-value: < 2.2e-16

>
> GenlinlagH <- lm (Logaverage ~ Avtemp3 + Avec3 + Avph + Avtemp3*Avph, dat
a=D)
> summary(GenlinlagH)

Call:
lm(formula = Logaverage ~ Avtemp3 + Avec3 + Avph + Avtemp3 *
    Avph, data = D)

Residuals:
     Min       1Q   Median       3Q      Max
-0.41105 -0.12106  0.02951  0.10565  0.42992

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.5688341 12.6755718   0.360    0.720
Avtemp3      -0.6961313  0.6483649  -1.074    0.287
Avec3        -0.0024232  0.0004628  -5.236  1.6e-06 ***
Avph          0.0209476  1.4949980   0.014    0.989
Avtemp3:Avph  0.0857849  0.0767832   1.117    0.268
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1776 on 71 degrees of freedom
  (839 observations deleted due to missingness)
Multiple R-squared:  0.7698,   Adjusted R-squared:  0.7568
F-statistic: 59.35 on 4 and 71 DF,  p-value: < 2.2e-16

>
> GenlinlagJ <- lm (Logaverage ~ Avtemp3 + Avec3 + Avph + Avec3*Avph, data=
D)
> summary(GenlinlagJ)

Call:
lm(formula = Logaverage ~ Avtemp3 + Avec3 + Avph + Avec3 * Avph,
    data = D)

Residuals:
     Min       1Q   Median       3Q      Max
-0.47175 -0.12605  0.01769  0.11236  0.43180

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.529457   9.603998   0.784 0.435653
Avtemp3      0.022684   0.006215   3.650 0.000498 ***
Avec3       -0.072643   0.039374  -1.845 0.069217 .
Avph        -0.330663   1.133899  -0.292 0.771431
Avec3:Avph   0.008400   0.004690   1.791 0.077582 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1752 on 71 degrees of freedom
  (839 observations deleted due to missingness)
Multiple R-squared:  0.7759,   Adjusted R-squared:  0.7632
F-statistic: 61.44 on 4 and 71 DF,  p-value: < 2.2e-16

>
```

```
>
> # Testing for collinearity
>
> ColT1 <- lm (Avtemp3 ~ Avec3, data=D)
> summary(ColT1)

Call:
lm(formula = Avtemp3 ~ Avec3, data = D)

Residuals:
     Min       1Q   Median       3Q      Max
-26.9859  -1.5747  -0.0451   1.8957   5.7430

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 26.985907   0.906720  29.762  < 2e-16 ***
Avec3       -0.030835   0.003613  -8.534 4.65e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.789 on 188 degrees of freedom
  (725 observations deleted due to missingness)
Multiple R-squared:  0.2792,   Adjusted R-squared:  0.2754
F-statistic: 72.84 on 1 and 188 DF,  p-value: 4.655e-15

>
> ColT2 <- lm (Avtemp3 ~ Avph, data=D)
> summary(ColT2)

Call:
lm(formula = Avtemp3 ~ Avph, data = D)

Residuals:
     Min       1Q   Median       3Q      Max
-20.5883  -0.8758  -0.3309   0.6780   6.5614

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  22.6654     2.6785   8.462 1.31e-12 ***
Avph         -0.2448     0.3204  -0.764    0.447
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.808 on 77 degrees of freedom
  (836 observations deleted due to missingness)
Multiple R-squared:  0.007524, Adjusted R-squared:  -0.005365
F-statistic: 0.5838 on 1 and 77 DF,  p-value: 0.4472

>
> ColT3 <- lm (Avec3 ~ Avph, data=D)
> summary(ColT3)

Call:
lm(formula = Avec3 ~ Avph, data = D)

Residuals:
    Min      1Q  Median      3Q     Max
-205.42  -34.46   -2.26   18.59  211.21

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  187.659     38.569   4.866 5.93e-06 ***
```

```
Avph            2.092      4.613   0.454    0.651
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 54.83 on 77 degrees of freedom
  (836 observations deleted due to missingness)
Multiple R-squared:  0.002665, Adjusted R-squared:  -0.01029
F-statistic: 0.2058 on 1 and 77 DF,  p-value: 0.6514
```