Delft University of Technology

Additional Thesis

Reducing the Measurement Frequency of Chloride and Ammonium Concentrations in Wieringermeer Landfill Leachate via SARIMAX Model

Rong Hu
4789911

Supervisors: Prof. Dr. Ir. Timo Heimovaara
Co-supervisor: Dr. Ir. F.C. Femke Vossepoel

March 9, 2020

ABSTRACT

Research on sustainable landfill management has been studied since 30 years ago in the Netherlands, the principle of which is to reduce the emission of harmful substances from the landfill to the surrounding soil and groundwater. As for this purpose, the active treatment is applied on Wieringermeer landfill, meanwhile, the long-term monitoring of substance concentration is of great importance. The measurement frequency of chemical concentration is twice per month, which costs around 48000 €/ year. To save money by reducing the measurement frequency, SARIMAX model is studied as a tool of data interpolation. For this analysis, we currently focus on the concentration measurements of chloride and ammonium. By comparing the SARIMAX interpolated data and the data with reduced size, the results indicate that directly dropping half of the measurements can be regarded as an acceptable way to reduce the measurement frequency, as the data properties are well preserved and the errors in estimating the mass of substances leaching out are in the acceptable range. However, interpolating using SARIMAX model doesn't have significant improvements in preserving the data properties. Further quartering the data can lead to large deviations in data properties.

ACKNOWLEDGEMENTS

When I first began with this topic, the feeling was mostly about fighting against the stress and motivating myself to get to know a brand new field. Now when it comes to an end of the thesis, a new world of data science has open its door to me.

First of all, I'd like to express my special thanks of gratitude to my supervisor Prof. Timo Heimovaara for providing me with the possibility to complete this report and all the valuable advice on scientific research. He is always able to clearly explain complex topics and nuances. I would like to say thanks to Dr. Femke Vossepoel for her valuable time and her generous encouragement. I want to express my love and thanks to my boy friend Jing Huang who gave me so much support and help. Thanks to the advice on the thesis from my classmate Yi Dai. It is impossible to finish the work without the help from them.

ACRONYMS

ACF Auto-correlation Function.
AIC Akaike Information Criterion.
AR Auto Regressive.
ARIMA Autoregressive Integrated Moving Average.

EC Electrical Conductivity.
ECDF Empirical Cumulative Distribution Function.

MA Moving Average.

PACF Partial Auto-correlation Function.
pEV Potential Evapotranspiration.

RMSE Root Mean Squared Error.

SARIMAX Seasonal Autoregressive Integrated Moving Average with Exogenous Regressors.
STL Seasonal and Trend Decomposition Using Loess.

CONTENTS

1 INTRODUCTION

1.1 PROJECT OVERVIEW

The Wieringermeer landfill leachate contains a series of chemical compounds that are potentially harmful to soil, groundwater and surface water under and next to the landfill. In order to protect the environment, the emission potential of harmful substances needs to be reduced to an environmentally protective level. This is done by active treatment on the waste body [Rohwerder, 2017]. Long term monitoring of substance concentrations in the leachate is of great importance. At present, the frequency of lab measurements is about twice per month. In this research, the possibility of reducing the measurement frequency is investigated. We currently focus on the concentration measurements of chloride and ammonium.

1.2 STATISTICAL TIME SERIES MODELLING

The concentration data obtained from the monitoring system can be considered as time series containing a series of data points indexed in time order. Time series analysis comprises methods for analyzing time series data, which are to extract meaningful statistics of the data. Based on the information extracted, there are several types of time series modelling. For example, explanatory analysis [Tukey, 1993], curve fitting [Kolb, 1984] and forecasting [Rob and George, 2018].

Forecasting is about predicting the future with all of the given information, including historical data and knowledge of any future events that might impact the forecasts. Proper forecasting of future values can be used as interpolation.

Exponential smoothing and ARIMA (autoregressive integrated moving average) models are the two most widely used approaches of time series forecasting [Rob and George, 2018]. For the exponential smoothing framework, forecasts are the weighted averages of past observation values, with the weights decaying exponentially as the observations get older. In the ARIMA model, the AR part indicates that the evolving variable of interest is regressed on its own lagged (i.e., prior) values. In the MA part, the regression error is a linear combination of error terms whose values occurred at various times in the past [Wikipedia, 2019b].

As an extension to ARIMA, SARIMAX (seasonal autoregressive integrated moving average with exogenous regressors) supports the direct modelling of the seasonal component, and it also makes use of the information from other related time series. As the concentration data is

highly seasonal related, meanwhile, some time series, such as rainfall and outflow data, are available to provide extra information, SARIMAX is chosen to do the forecasting.

1.3 RESEARCH QUESTIONS

To reduce the measurement frequency, one strategy is to directly reduce the size of the data sets (less measurements). Another strategy is by applying SARIMAX to interpolate in the data set with a low frequency and to estimate measurement points at a higher frequency. Two methods are investigated and compared in this research.

The intention of this research is to find out how the SARIMAX model performs on modelling the missing measurements. To accomplish that, the error generated by the modelling process must be quantified. The changes in data properties after modelling should be analyzed. All the results need to be clearly interpreted in order to provide support for policy decisions by the landfill manager. For this reason, the present study must give clear-cut answers to the questions listed below:

- Are the SARIMAX model simulations accurate enough (are the quantified error of the model results in an acceptable range)?
- Are the data with reduced size and the raw data from the same distribution?
- Are the SARIMAX modelled data and the raw data from the same distribution ?
- Does the application of SARIMAX model have significant improvement in preserving the data properties compared to directly reducing the size of the data?

2 METHODOLOGY

The flow chart of the main steps in methodology is presented in Figure 2.1. There are mainly five parts. The first part is data regularization which aims at preparing equidistant input data for SARIMAX model. The regularized time series are analyzed to extract information for the model. After parameter selection, model training and diagnostics, the model is ready for simulating. The results are analyzed in three aspects: error generated by the model, changes in data properties and estimation on mass of substances. In the end, sensitivity analysis is performed.

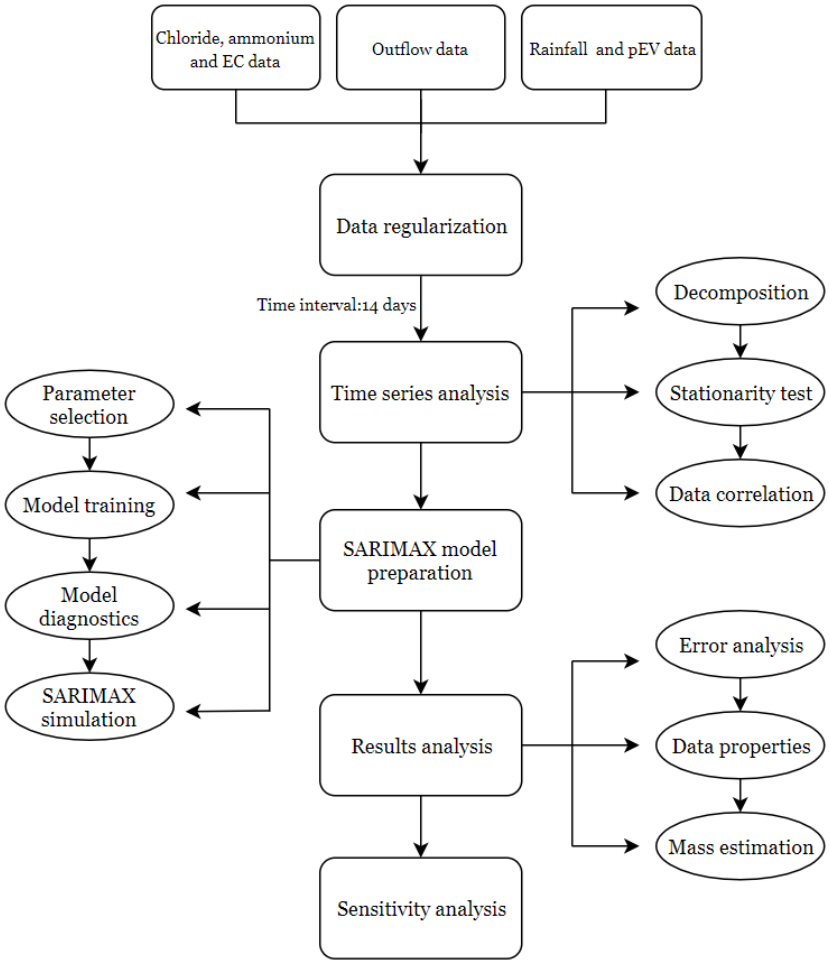


Figure 2.1: Flow chart of the methodology

2.1 DATA REGULARIZATION

Data sets available for this study are: chloride concentrations (mg/L), ammonium concentrations(mg/L) and electrical conductivity (EC) values (ms/cm) which are sampled from landfill leachate roughly every 14 days; cumulative outflow (m^3) which is measured by sensor every 15 minutes in the drainage system; rainfall and evaporation measurements that are obtained from the Berkhout Weather Station of the Royal Netherlands Meteorological Institute (KNMI).

The concentrations and EC data are not collected strictly equidistant in time. The collection frequencies are different between some data sets. However, one precondition for conducting SARIMAX modelling is that all data sets should have equal time interval, thus, data regularization is needed. In this study, the interval is set to be 14 days corresponding to the measurement frequency of twice a month. There are different approaches to achieve equidistant time interval for different data sets:

Chloride, ammonium concentrations and EC data:
For chloride and ammonium concentrations data, the cubic-spline interpolation (*interp 1d (kind = cubic))* is used to generate a continuous time series. Afterwards, a new data frame with an equidistant time interval of 14 days is generated by picking out the data of the designated dates.

Outflow data:
The cumulative outflow data (m^3) are measured every 15 minutes by sensors. First, *resample ()* function is used to turn the data frequency into 1 day, the data of every 14 days are picked out and formed a new DataFrame. Because the data are given in the cumulative form, *diff ()* function is applied to obtain the outflow volume of every 14-days-period.

Rainfall data and potential evapotranspiration (pEV) data:
The rainfall and pEV(m) are given in daily frequency, the first step is to cumulate it up using the *cumsum ()* function, after whichthe data of every 14 days are picked out and formed a new DataFrame. In order to keep the consistency of the units between exogenous data, the rainfall and pEV should be multiplied by the landfill area to change the unit from m to m^3. The bottom layer area of the landfill is 28355 m^2.

What should be noticed is that interpolating should not change the properties of the data sets, therefore, the mean values and density plots are used to check if the underlying properties of the data sets have been changed after interpolating.

2.2 TIME SERIES ANALYSIS

The methods of time series analysis used in this research consist of time series decomposition in Section 2.2.1, stationarity test in Section 2.2.2 and correlation analysis between endogenous and exogenous data in Section 2.2.3.

2.2.1 Time series decomposition

The long-term increase or decrease in the data set is trend. Seasonality is the variation occurred regularly with a specific time interval that is less than a year, such as monthly and weekly. Identifying the trend and seasonal pattern can help us make a better decision on selecting the model parameters. One tool to identify trend and seasonality in the time series is decomposition. Decomposition is the primary step for studying time series data, the information extracted out consists of three components: trend, seasonality and residual [Rob and George, 2018]. In this research, a robust and versatile method for time series decomposing is used, which is called seasonal and trend decomposition using Loess (STL) decomposition.

2.2.2 Stationarity of the time series

Stationarity means the property of the time series does not depend on the time it is observed [Rob and George, 2018]. Stationarity is an assumption underlying the SARIMAX modelling procedures [Palachy, 2019]. Two types of statistical test are used in the stationarity analysis: the Augmented Dickey-Fuller Test and the KPSS Test.

The Augmented Dickey-Fuller Test:

The time series can be described by a characteristic equation which consists of a series of monomials, each monomial has a root. If the characteristic equation of a time series has a root of 1 (unit root), such a time series is non-stationary [Stephanie, 2016]. The Augmented Dickey-Fuller is one of the most widely used unit root tests [Brownlee, 2016]. The test allows one to calculate a quantitative value (p-value) which is based on the probability that one of the following two hypotheses is true.

Null Hypothesis (H_0): It suggests the time series has a unit root, meaning the time series is non-stationary.

Alternate Hypothesis (H_1): It suggests the time series does not have a unit root, meaning the time series is stationary.

If the returned p-value $<$ 0.05, the null hypothesis (H_0) can be rejected and the alternative hypothesis (H_1) can be supported at a 95% confidence limit, which means the data does not have a unit root and is statistically stationary. Conversely, if the returned p-value $>=$ 0.05, it means we cannot reject the null hypothesis (H_0), the data has a unit root and is statistically non-stationary.

The KPSS Test:

The KPSS test can also be used to check the presence of a unit root [Kwiatkowski et al., 1992]. Contrary to the Dickey-Fuller tests, the null hypothesis assumes the data is stationary around a linear trend or a mean, while the alternative is the presence of a unit root (non-stationary). The KPSS test is often used to complement Dickey-Fuller-type tests. The two hypotheses are:

Null hypothesis(H_o): It suggests the time series does not have a unit root, meaning the time series is statistically stationary.

Alternate Hypothesis (H_1): It suggests the time series has a unit root, meaning the time series is statistically non-stationary.

If the returned p-value $<$ 0.05, we can reject the null hypothesis (H_o) and support the alternative hypothesis (H_1) at a 95% confidence limit, which means the data has a unit root and is statistically non-stationary. Conversely, if the returned p-value $>=$ 0.05, null hypothesis (H_o) cannot be rejected, which indicates the data does not have a unit root and is statistically stationary.

2.2.3 Correlation between endogenous data and exogenous data

There are two types of data involved in the SARIMAX model: endogenous and exogenous data. Endogenous data is the measured data set that we fit the model and make future predictions on. Exogenous data can be input to the model as extra information that might help the model estimation and plays the role as the regressor in the SARIMAX model. In this case, the endogenous data are chloride and ammonium concentrations. The exogenous data are EC, outflow, rainfall and pEV data.

In a simplified example, we can write an AR(1) with an exogenous regressor as:

$$y(t) = ay(t-1) + bx(t) + e(t) \tag{2.1}$$

Where y(t) is the output, y(t-1) is the lagged values, x(t) is the exogenous variable and e(t) is the random error term. This function reveals the idea of how the exogenous data is functioning in the model. By investigating the correlation between exogenous data and endogenous data, one can better make use of the exogenous information and make it a helpful source to increase the model accuracy. Two types of correlation coefficients - Pearson and Spearman are used to reveal the relation between exogenous and endogenous data. A larger Pearson coefficient corresponding to a stronger linear correlation. A linear relationship means that the variables move in the same direction at a constant rate. The Spearman coefficient corresponds to monotonic relationship, in which the variables tend to move in the same direction, however it is not necessary to move at a constant rate [Minitab, 2019].

2.3 MODELLING PROCESS

The SARIMAX modelling process consists of parameter selection in Section 2.3.1, model training in Section 2.3.2, model diagnostics in Section 2.3.3 and SARIMAX simulation in Section 2.3.4.

2.3.1 Parameter selection

After preparing the data, the next step is to determine the model parameters. The proper selection of parameter plays a decisive role in the model performance.

The parameters in the SARIMAX model are in the form of **(p, d, q) × (P, D, Q)$_{\mathbf{m}}$**. The seven parameters: **p, d, q, P, D, Q, m**, are all non-negative integers. **p** is the order (number of time lags) of the AR model, **q** is the order of the MA model. **d** is the degree of differencing (the number of times to take difference on the data set to make it stationary).

The lowercase letters **(p, d, q)** are the short term parameters that operate on the adjacent lags. Accordingly, the capital letters **(P, D, Q)** hold the same meanings as corresponding lowercase letters, but they are the long term seasonal parameters that operate on a seasonal time scale. **m** is the number of time steps for a single seasonal period.

Autocorrelation function(ACF) and partial autocorrelation function(PACF) to determine d,D

The first and the most important step in parameter selection is the determination of the differencing orders (**d, D**) needed to stationarize the time series. Usually, the correct amount of differencing(**d, D**) is the lowest order of differencing that yields a time series which fluctuates around a well-defined mean value. ACF and PACF plots can help us to guess the reasonable values for the order of differencing[Robert Nau, 2019]. Some rules should be mentioned:

- The ACF will drop to zero relatively quickly for a stationary time series. Meanwhile, the ACF decreases slowly for a non-stationary data .
- It probably needs a higher order of differencing on the series if positive autocorrelations are present out to a high number of lags.
- A higher order of differencing is not needed if the autocorrelations are all small and patternless or the lag-1 autocorrelation is zero or negative.
- The series is possibly over-differenced if the lag-1 autocorrelation is -0.5 or more negative.

According to the rules above, by plotting out the ACF and PACF curves of the time series, the proper **d, D** parameters can be determined.

Grid search to determine p,q

After differencing, the next step in fitting a SARIMAX model is to determine whether AR or MA terms (**p,q**) are needed to correct any autocorrelation that remains in the differenced series. According to Robert Nau [2019], adding an AR term corrects for mild under-differencing, while adding an MA term corrects for mild over-differencing.

The literature indicates that,in some cases, the ACF plot and the PACF plot can be used to determine appropriate **p** and **q** values, however, there might be a degree of subjectivity in selecting which values to apply [Rob and George, 2018]. Beyond that, the Akaike information criterion (AIC) can also be used in parameter selection. AIC is used to estimate the relative

amount of information lost by a given model. The lower the AIC value, the less information loss of the model, so the higher the quality of it[Wikipedia, 2019a].

Therefore, in this study, a grid search over different combinations of **p,q** (limited to a maximum order of 2 for simplicity) is used. The configuration that leads to the lowest AIC value will be used in the following modelling process.

2.3.2 Model training

When training the models, it is a common practice to separate the available data into two portions, training and testing data. The training set is what we will use to fit the model for each possible value of the manually-set parameters, and the test set is what we use to evaluate our results for reporting and get a sense for how well our model will do on new data in the real world [Ramesh Sridharan, 2011].

The time series in this study contains data points from 2012-6-28 to 2019-1-1 with a time span of roughly 6.5 years, the first 4.5 years (122 points) are used as training data and the last 2 years (48 points) are used as testing data.

The *SARIMAX ()* function from the *statsmodels.tsa.statespace* package is used to fit the model. The concentration data of chloride and ammonium in the first 4.5 years are used as endogenous training data. Meanwhile, rainfall, pEV, outflow and EC in the first 4.5 years are used as exogenous data. The optimal parameters identified in the previous step are applied. After that, *model.fit ()* function is used to fit the model by maximum likelihood via Kalman filter. By executing the command of *print (model _ fit.summary())*, the model results can be presented as a table.

2.3.3 Model diagnostics

Definition of model diagnostics

The property of the residuals (the difference between the initial data and fitted values) should be evaluated after fitting the model, this step is called model diagnostics. A good model should yield residuals with the following properties [Rob and George, 2018]:

1. The mean of the residuals is zero. If the mean of the residuals is not zero, then the fitting is biased.
2. The residuals are uncorrelated. If there is information left which is not used in the model estimation, correlations would be remained between residuals.
3. The variance of the residuals is constant.
4. The residuals are normally distributed.

Checking these properties is important in order to see whether all of the available information is used in the model. Any results that do not satisfy the first and the second properties can

be improved. In addition to the essential properties(the first and the second), it is useful (but not necessary) for the residuals to also have the third and the fourth properties.

Methods of model diagnostics

The first step in model diagnostics is to print the mean of the residuals and see if it meets the first requirement. The second step is to test the diagnostics by applying *plot_diagnostics ()* function to check if the residuals meet the other three requirements, this function returns four plots:

1. The standardized residuals (residuals divided by the standard deviation of residuals) are plotted over time.
2. The histogram and the estimated density of standardized residuals are plotted in the same graph with a normal(0,1) density as reference.
3. A normal Q-Q plot are plotted with a normal reference line.
4. Correlogram plots of the autocorrelation are plotted versus time lags (the autocorrelations should be near zero for all time lag).

If the residuals have a mean m, by simply subtracting m from all fitted values, the bias problem will be solved. Besides, the Box-Cox transformation is a way to transform non-normal variables into a normal shape [Box and Cox, 1964]). By conducting the transformation on endogenous data, the model performance might be improved [Rob and George, 2018].

2.3.4 SARIMAX simulation

To generate the 48 points in the two testing years, all the odd positions in that 48 points are SARIMAX simulations and the even positions are the real measured data. The idea behind this is to reduce the measurement frequency from twice a month to once a month by interpolating one of the measurements in a month by SARIMAX simulation. The programming idea is described below:

```
HD = [Historical data]
While t < t_end:
   P = Model(HD)
   HD = HD.append(P)
   t = new(t)
   HM = next_measurement(t)
   HD = HD.append(HM)
```

HD is the historical data available for modelling. Starting from a specific point of time (t), SARIMAX is applied to make one-step ahead forecasting. P is the value obtained and is appended to the historical data with t as index. Going on to the next point of time (t), HM is the measured data at this specific time (t) and is appended to the historical data with a index of t. The loop will be repeated until t_{end} is reached.

2.4 RESULTS EVALUATION

After generating the data sets, the results evaluation section consists of error analysis in Section 2.4.1, change in data properties after modelling in Section 2.4.2 and estimation on mass of the substances leached out in Section 2.4.3.

2.4.1 Error analysis

Absolute error and relative error

The difference between the real observed value and the model simulation is the simulation error, which can be described as:

$$e_t = y_t - \hat{y}_t \tag{2.2}$$

Where y_t is the testing data (measured value), $\hat{y}_t$ is the modelled value.

The relative error is calculated as:

$$Re_t = (y_t - \hat{y}_t)/y_t \tag{2.3}$$

RMSE

Root mean squared error (RMSE) is the standard deviation of the model errors, which can be described as:

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\left(y_t - \hat{y}_t\right)^2} \tag{2.4}$$

What should be mentioned is that RMSE is a scale-dependent error, the errors are on the same scale as the original data. therefore, it can only be used to make comparisons between data sets that have the same units. By comparing the RMSE, we can get an impression of model performs in different scenarios.

2.4.2 Changes in data properties

By looking at the change of the mean values after modelling, one can tell how well the model performs in different cases. Moreover, the difference in standard deviation can also offer us the information on how the model behaves. Besides this, the empirical cumulative distribution function (ECDF) plots provide us with insight into the data distribution. Further, the Mann-Whitney U test[Rosie Shier, 2004] is used to test whether the two data sets can be considered originating from the same distribution. The null hypothesis for this test is that the two input groups have the same distribution. One fails to reject the null hypothesis if the returned p-value is larger than 0.05, in another word, the two groups can be considered originating from the same distribution [Deborah J. Rumsey, 2016].

Three scenarios are formulated to do the comparisons:

Scenario 1: complete raw measurements in the two testing years (48 points) & complete modelled data in the two testing years (halved measurements + SARIMAX simulations) (48 points).

Scenario 2: complete raw measurements in the two testing years (48 points) & halved measurements in the two testing years (24 points).

Scenario 3: complete raw measurements in the two testing years (48 points) & quartered measurements in the two testing years (12 points).

2.4.3 Estimation on mass (kg) of the substances

Another important aspect for the evaluating system is the mass of substances leaching out from the landfill. The cumulative mass leaching out in the two testing years is calculated for the four scenarios below:

Scenario 1: complete raw measurements (48 points).

Scenario 2: halved measurements (24 points) + SARIMAX simulations (24 points).

Scenario 3: halved measurements (24 points).

Scenario 4: quartered measurements (12 points).

The equation to calculate cumulative mass is:

$$M = \int_{n=1}^{N} (q_{\text{out,t}} \times C_t) dt \tag{2.5}$$

Sampling frequency of the concentration is once every 14 days, we assume the measured concentration to represent the average concentration over that 14 days.

2.5 SENSITIVITY ANALYSIS

Exogenous data act as regressors in the model estimating process. SARIMAX is a linear model [Fulton, 2018], thus, exogenous regressors can only enter in linearly. For this reason, exogenous data sets that have weak linear correlations with endogenous data should be discarded from the model.

Larger Pearson correlation coefficient indicates stronger linear relationship between two data sets. By looking at the Pearson coefficients, one can decide on which exogenous data to be kept and which to be discarded (as all of the four exogenous data sets are used in the base case scenario). By comparing the results of different scenarios (with different exogenous data), one can better improve the model performance.

2.6 SARIMAX INTERPOLATION ON HISTORY DATA IN 2014-2016

The regularization step on the history data (EC, chloride and ammonium data) before being used in the SARIMAX model, is in principle an interpolation step. It is noticed that the measurement frequencies in 2012-2013 and 2017-2019 are roughly twice a month. However, the measurement frequency and the number of the data points are halved in 2014-2016 at the middle stage of the data set. In this period, the cubic-spline interpolation used in the base case scenario may generate large errors during interpolating. Thus, the SARIMAX model is considered as an alternative to do the interpolation in 2014-2016, as the contributions from the related exogenous data are also involved.

For EC data, there is no highly correlated exogenous data available (as it will be used later as exogenous data for chloride and ammonium concentrations interpolating). Therefore, the SARIMAX interpolating on EC data is only based on autocorrelation. The cubic-spline interpolation is first applied to make the whole data set equidistant in time. In the middle three years (2014-2016), the SARIMAX model is used to do the interpolation, in another word, SARIMAX model simulates the missing measurements, the programming idea is the same as what is mentioned in Section 2.3.4.

For chloride and ammonium data, cubic-spline interpolation is first applied on the whole data sets in the middle three years (2014-2016), then, the SARIMAX model is used to do the interpolation. Additionally, the SARIMAX interpolated EC data is used as exogenous data in the model. The results will be discussed in Chapter 3, Section 3.7.

3 RESULTS AND DISCUSSIONS

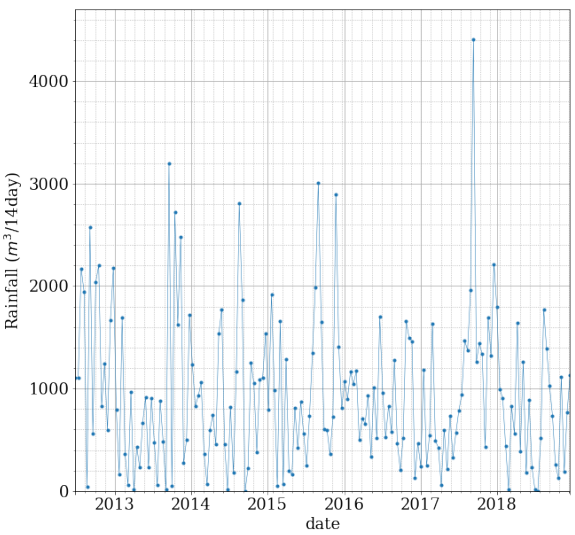
In this chapters, all the results are presented. The discussions are from seven aspects: time series analysis in Section 3.1; error quantification in Section 3.2; comparisons on data properties in Section 3.3; estimation on mass of the substances leaching out in Section 3.4; the change of correlation between endogenous and exogenous data after modelling in Section 3.5; sensitivity analysis in Section 3.6 and SARIMAX interpolation on history data in Section 3.7.

3.1 TIME SERIES ANALYSIS

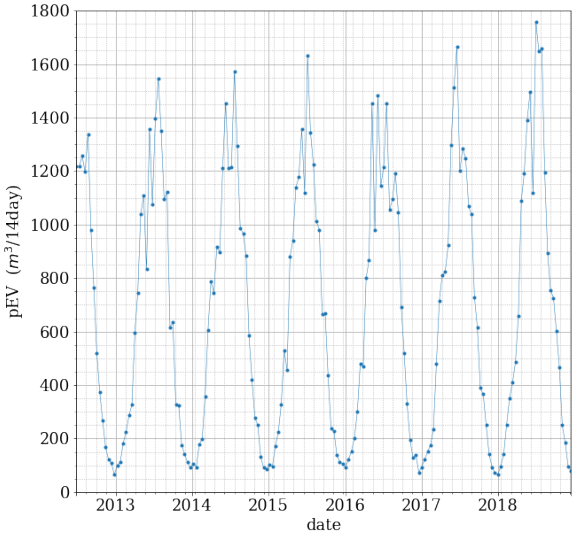
3.1.1 Data visualization

Figure 3.1 shows the rainfall, pEV and outflow data after regularization. The EC, chloride and ammonium data are the raw data without interpolation. From Figure 3.1(b),(c) and (d), potential evaporation, cumulative outflow and bi-weekly outflow data show clear seasonal patterns. The potential evaporation is higher in the summer and lower in the winter, consequently, the outflow is the opposite. The seasonal dynamics are mainly controlled by evaporation.

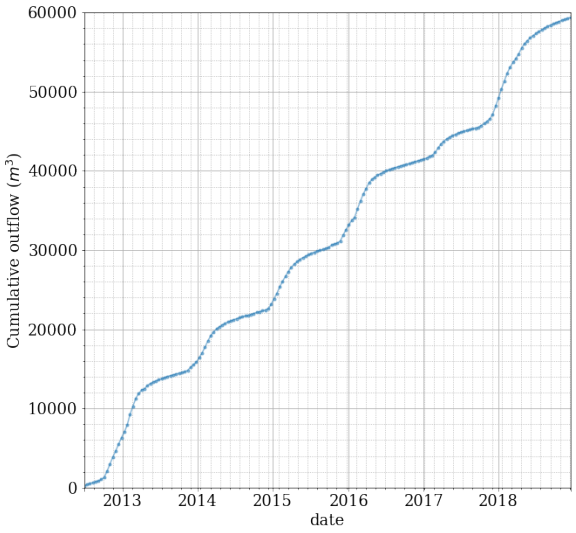
From Figure 3.1(e), (f) and (g), in EC, chloride and ammonium data, it can be observed that there are more noisy signals in 2012-2013 and 2018-2019, and less noisy signals in 2014-2016. This is because of the lower measurement frequency in the middle three years compared to the other years. Contrary to the outflow, the seasonal patterns in EC, chloride and ammonium concentrations have higher values in summer and lower values in winter. Higher evaporation and less outflow in the summer period might lead to the peak values of the concentrations. The seasonal dynamic of EC is similar to that of the concentration.

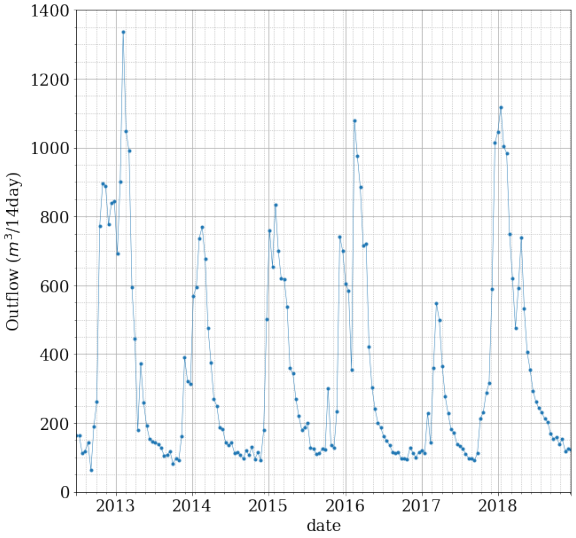


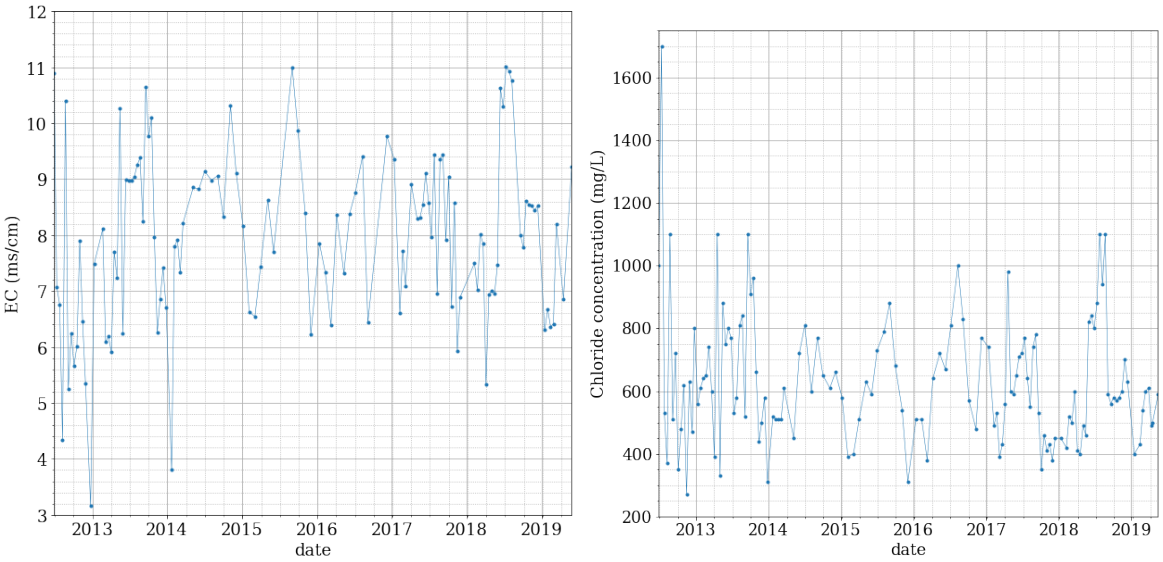
(a) Rainfall

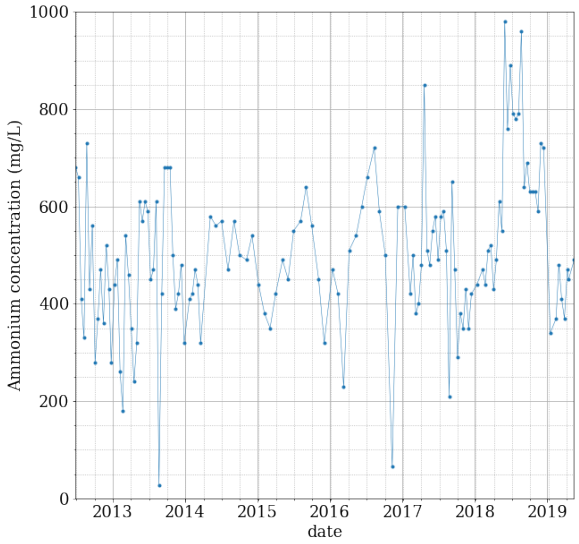


(b) pEV







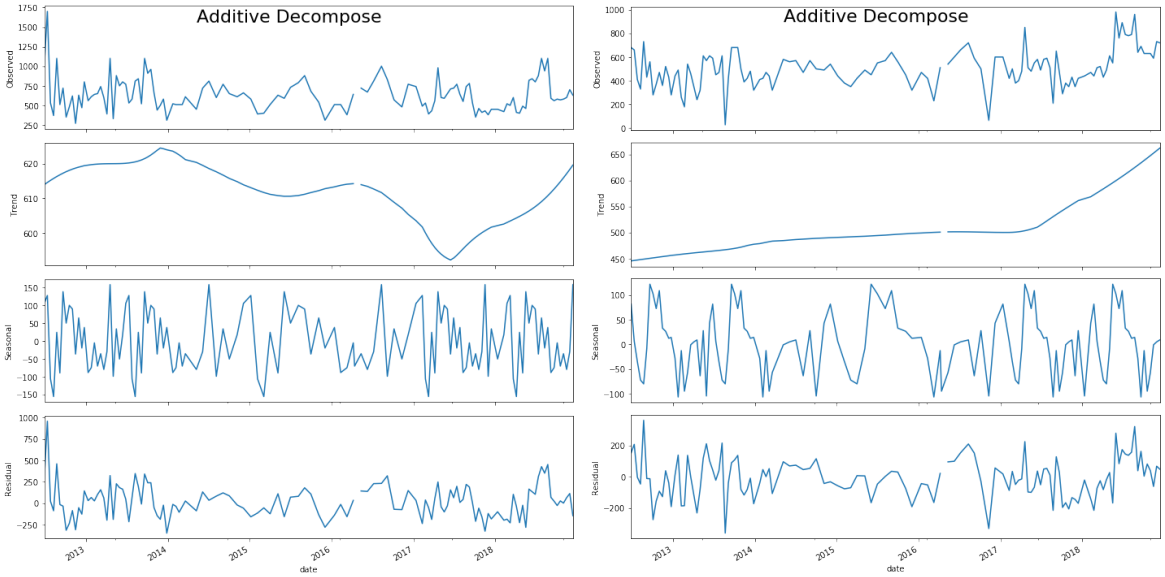


(c) Ammonium concentration

Figure 3.1: Data visualization

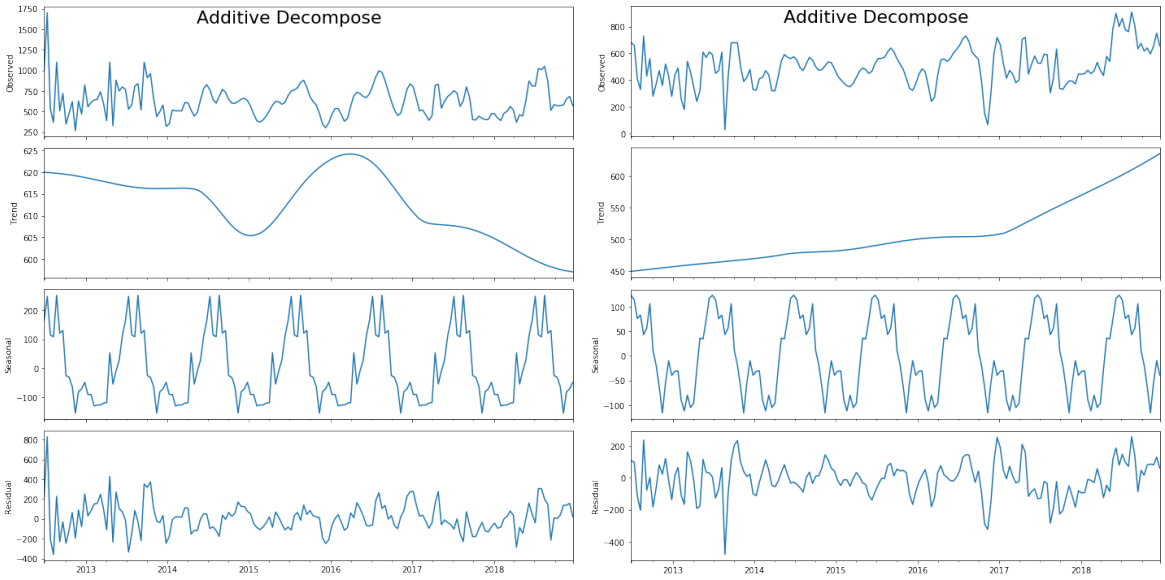
3.1.2 Decomposition and stationarity

STL decomposition is a robust and versatile method for decomposing time series[Rob and George, 2018]. The time series is split into three components, trend, seasonality and residual. The concentration data (raw and interpolated data) are decomposed and displayed in Figure 3.2 and Figure 3.3.



(a) Chloride **(b)** Ammonium

Figure 3.2: STL decomposition on raw data(without interpolation)



(a) Chloride **(b)** Ammonium

Figure 3.3: STL decomposition on data after interpolation

As Figure 3.2 (a) shows, the STL decomposed trend in chloride data varies in the range of 595-625 mg/L. The variation amplitude is around 30 mg/L, which is relatively a small number compared to the magnitude of the'aw data. The ADF and KPSS test both indicate that the chloride time series is statistically stationary. The seasonal pattern of the raw data (Figure 3.2 (a)) is less distinct than in the interpolated cases (Figure 3.3 (a)), it is dense at both ends and sparse in the middle. The decomposition after interpolation shows a clear yearly seasonal pattern, where chloride concentration is high in summer and low in winter.

As for ammonium, Figure 3.2 (b) seems to indicate a slow ascending trend in the data set from 450 to 650 mg/L, where the slope of the trend becomes steeper after 2017. Despite this observed trend in the decomposition, it is statistically insignificant as the ADF and KPSS test both indicate that it is a stationary series. The yearly seasonal pattern of the ammonium data is similar to that of the chloride data.

From what has been discussed above, some conclusions can be drawn. As the data frequency is changed by the interpolation, one can infer that the data frequency has impacts on the estimation of the seasonal pattern. Moreover, both two concentration data sets are statistically stationary.

3.1.3 Correlation between endogenous data and exogenous data

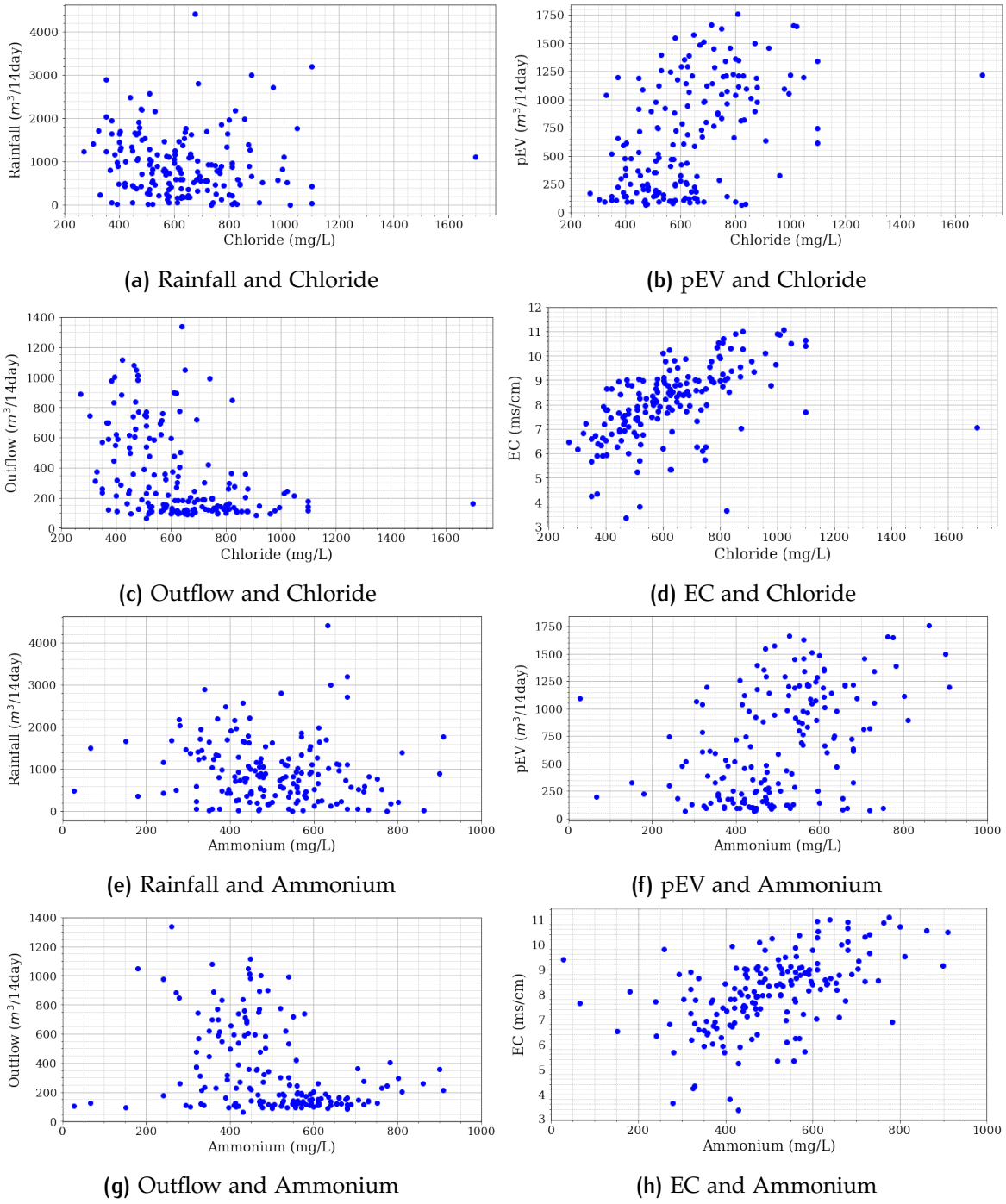
To show the correlation visually, the exogenous data are plotted versus the endogenous data, the results are displayed in Figure 3.4. The Pearson and the Spearman correlation coefficients are calculated between data sets and displayed in Table 3.1.

Table 3.1: Correlation coefficients

	Chloride		Ammonium	
	Pearson	Spearman	Pearson	Spearman
Outflow	-0.450	-0.440	-0.302	-0.197
EC	0.775	0.690	0.576	0.510
Rainfall	-0.134	-0.261	-0.216	-0.357
pEV	0.665	0.625	0.497	0.448

In both cases of chloride and ammonium, the Pearson coefficients of outflow, EC and pEV are larger than their Spearman coefficients, which means they are more likely to be linearly correlated to the concentration. On the contrary, the Pearson coefficient of rainfall is smaller than its spearman coefficient, which indicates that the rainfall is more likely to be in a monotonic relationship with the concentration.

In addition, the relatively large Pearson coefficient of EC reveals a strong linear correlation between EC and substance concentration. What else should be noticed is that all coefficients related to chloride are larger than ammonium, and the strength of correlation might have impacts on the model performance, which will be discussed further in Section 3.2.



(a) Rainfall and Chloride

(b) pEV and Chloride

(c) Outflow and Chloride

(d) EC and Chloride

(e) Rainfall and Ammonium

(f) pEV and Ammonium

(g) Outflow and Ammonium

(h) EC and Ammonium

Figure 3.4: Correlation between exogenous data and endogenous data

3.2 ERROR QUANTIFICATION

The modelled and raw data are plotted in the same figure (Figure 3.5). The RMSEs of the SARIMAX simulation based on corresponding measurements are calculated for both chloride and ammonium (according with Equation 2.4). The relative error is calculated according with

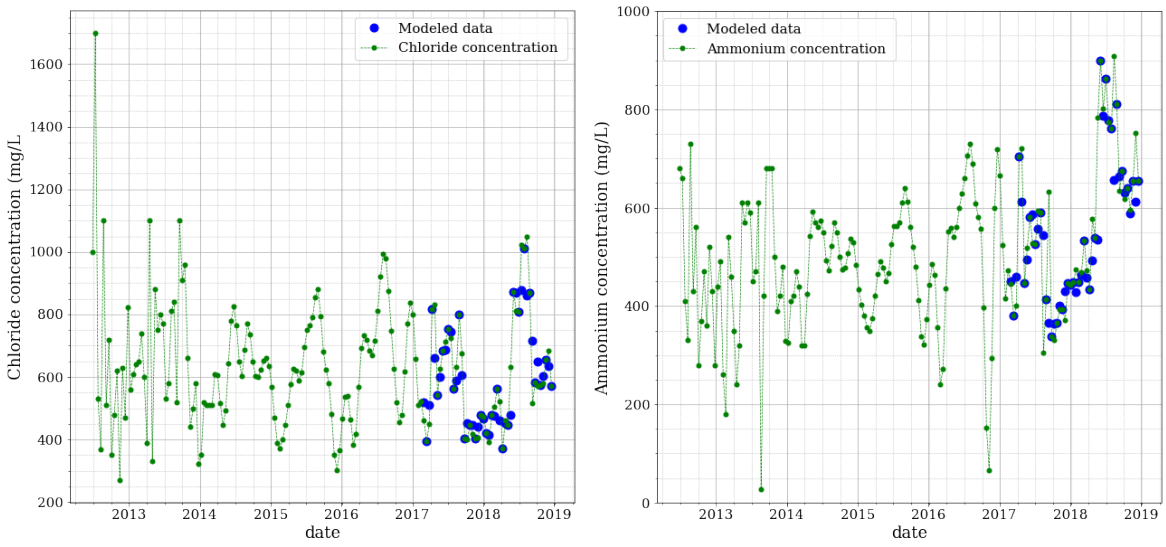
Equation 2.3, the mean of the relative errors for all simulations is shown in Table 3.2. The absolute error and the relative error are plotted in Figure 3.6 and Figure 3.7 with respect to time.

According to Table 3.2, the RMSE of the ammonium model is higher than that of the chloride model. The reason for this might come from the effect of the exogenous data. Back to Table 3.1, one can see that the correlation coefficients between ammonium and EC/outflow/pEV are all lower than that of chloride. However, the exogenous regressors play important roles in the model estimation based on their correlation with the data to be simulated. Therefore, stronger correlation might lead to better results. Another reason for this could be that there is a sudden increasing trend in ammonium data since 2017, and this emerging trend could have added uncertainty to the model output.

Figure 3.5 shows good fitness of the model, as the blue dots match well with the green dots. Despite some large spikes of absolute error in Figure 3.7, the relative error against the magnitude of the real measurements are small. As shown in Table 3.2, the mean of the relative errors are only -0.251% and -1.568% for chloride and ammonium, respectively. Thus, the error generated by the modelling process can be regarded as being in a acceptable range.

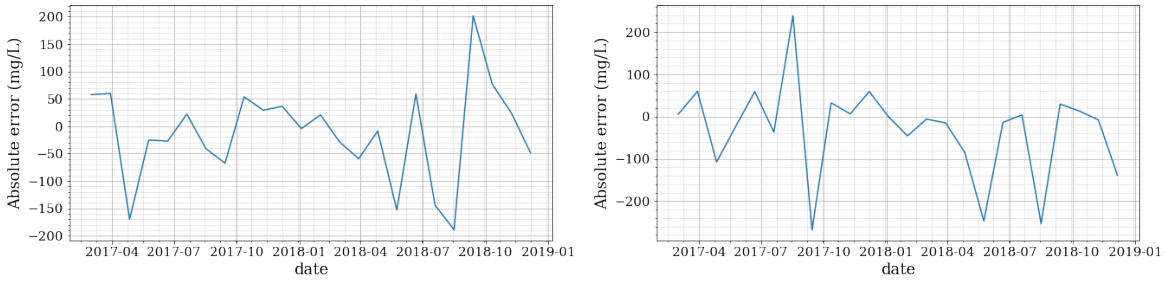
Table 3.2: Quantified error

	Chloride	Ammonium
RMSE of simulated data(mg/L)	88.212	113.397
Mean of relative error	-0.251%	-1.568%



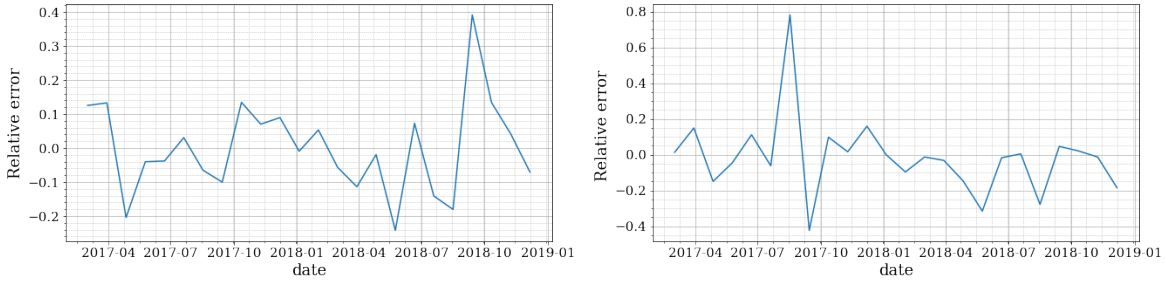
(a) Chloride **(b)** Ammonium

Figure 3.5: Raw data and modelled data



(a) Chloride **(b)** Ammonium

Figure 3.6: Absolute simulation error



(a) Chloride **(b)** Ammonium

Figure 3.7: Relative simulation error

3.3 COMPARISON ON DATA PROPERTIES

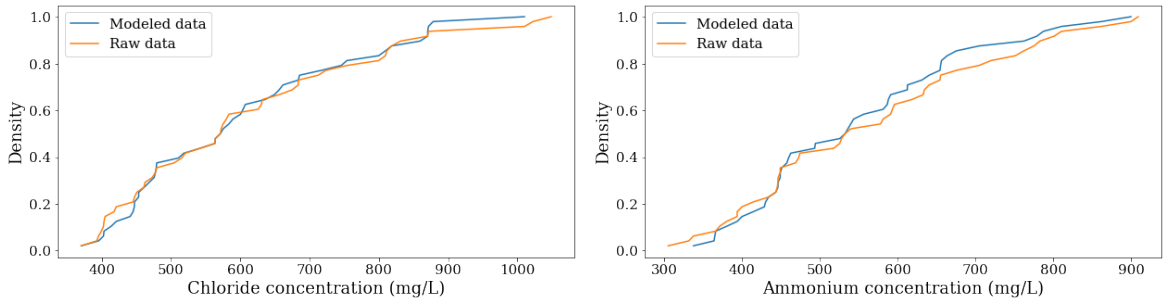
In order to check whether the simulated values for the missing points and the measured values can be considered to come from the same distribution, the mean and standard deviation of the data, together with the Mann-Whitney U test are checked for both. The ECDF plots are presented to give the visual results on the data distributions.

Scenario 1: complete raw measurements in the two testing years & modelled data in the two testing years (half simulated and half measured)

The data properties for scenario 1 are shown in Table 3.3. The density distributions of the modelled and measured data are plotted together in Figure 3.8 in order to make comparisons.

Table 3.3: Data properties scenario1

	Chloride	Ammonium
Mean of raw measurements (mg/l)	605.615	564.321
Mean of modelled data (mg/l)	598.839	548.885
STD of raw measurements (mg/l)	181.267	161.151
STD of modelled data (mg/l)	162.768	141.016
P-value of Mann-Whitney U Test	0.474	0.368



(a) Chloride ECDF plot

(b) Ammonium ECDF plot

Figure 3.8: Scenario1 Density plots of raw data and modelled data

For the modelled chloride data, the change of mean is 598.839 - 605.615 = -6.776 mg/L. The relative change is 6.776/605.615 = 1.12 %, which is small compared to the measurement error in reality (assumed to be 10 %). For the modelled ammonium data, the change of mean is 548.885 – 564.321 = -15.44 mg/L, with a relative change of 15.44/564.321 = 2.73 %, which is larger than that in the chloride model but still smaller than the assumed measurement error of 10%. Therefore, the mean values of the data do not change significantly after modelling. The standard deviation of the data both decrease after modelling, which illustrates that the SARIMAX interpolated data are less volatile compared to the raw data.

By looking at the ECDF plots for chloride (Figure 3.8 (a)), the two curves overlap to a large extent, which means the data density distribution is well retained in the chloride case. However, in the ammonium case, it shows larger deviation between the two curves (Figure 3.8 (b)), which means the density distribution changes in a perceptible range after modelling.

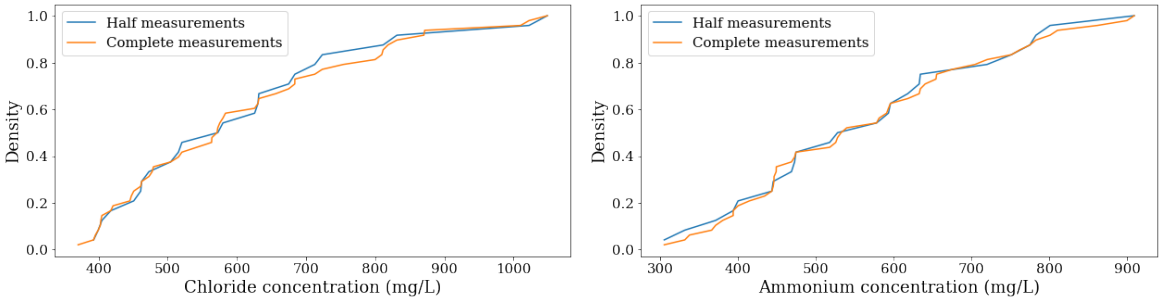
A large p-value (> 0.05) indicates that the evidence against the null hypothesis is relatively weak, hence one cannot reject it [Deborah J. Rumsey, 2016]. The null hypothesis in this scenario is that the two groups have the same distribution, the p-values for chloride and ammonium are both larger than 0.05 from Table 3.3, hence we conclude that the two distributions are similar. Despite the perceptible deviation on the ECDF plot in ammonium case, from the perspective of the statistical test, one can still say that the modelled data and raw data are from the same distributions.

Scenario 2: complete raw measurements in the two testing years & halved raw data in the two testing years

In this scenario, the data properties before and after halving the number of measurements are compared and the results are shown in Table 3.4. The Mann-Whitney U test can not be used as the numbers of data points are different.

Table 3.4: Data property scenario2

	Chloride	Ammonium
Mean of raw measurements (mg/l)	605.615	564.321
Mean of halved raw data (mg/l)	606.370	564.152
STD of raw measurements (mg/l)	181.267	161.151
STD of halved raw data (mg/l)	183.616	163.348



(a) Chloride ECDF plot **(b)** Ammonium ECDF plot

Figure 3.9: Scenario2 Density plots of raw data and modelled data

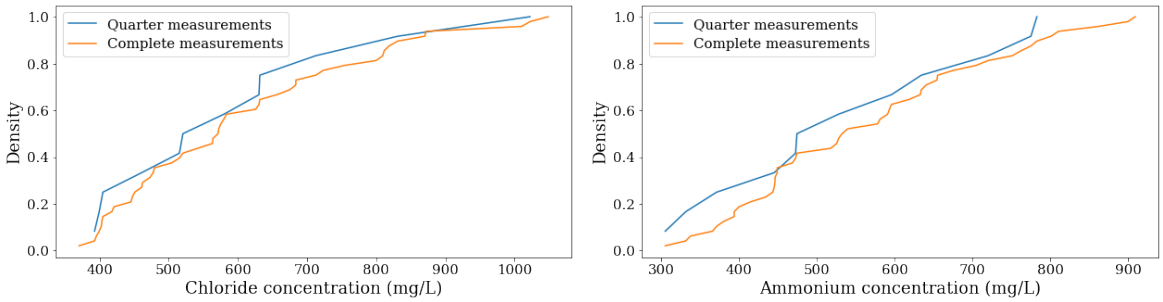
Comparing the results in Table 3.3 and Table 3.4, one can see that the changes in mean and standard deviation are smaller in scenario 2. The deviations between the orange curves and the blue curves in the ECDF plots are also less obvious than that in scenario 1, especially in the ammonium case. Thus, the halved data and raw data can be regarded as from the same distribution.

Scenario 3: complete raw measurements in the two testing years & quartered raw data in the two testing years

Further on, the mean value and the standard deviation are compared after quartering the data, the results are shown in Table 3.5. The density distributions of the two data sets are plotted together in Figure 3.10.

Table 3.5: Data property scenario3

	Chloride	Ammonium
Mean of raw measurements (mg/l)	605.615	564.321
Mean of quartered raw data (mg/l)	591.948	536.189
STD of raw measurements (mg/l)	181.267	161.151
STD of quartered raw data (mg/l)	190.771	166.153



(a) Chloride ECDF plot **(b)** Ammonium ECDF plot

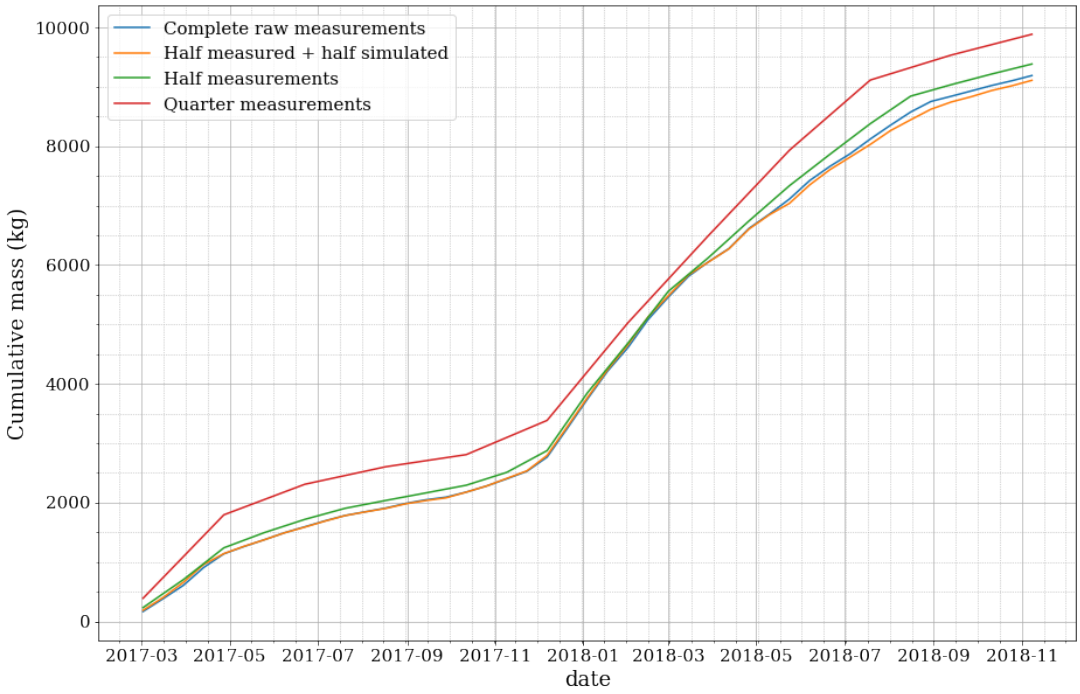
Figure 3.10: Scenario3 Density plots of raw data and modelled data

By comparing the results in Table 3.3, Table 3.4 and Table 3.5, the change of mean in scenario 3 is the largest among the three scenarios. The change of standard deviation in scenario 3 is larger than that in scenario 2 but smaller than that in scenario 1. It shows obvious deviations between the blue curves and the orange curves, which means the density distributions are significantly changed after quartering the data. Thus, the quartered data and raw data can no longer be regarded as coming from the same distribution.

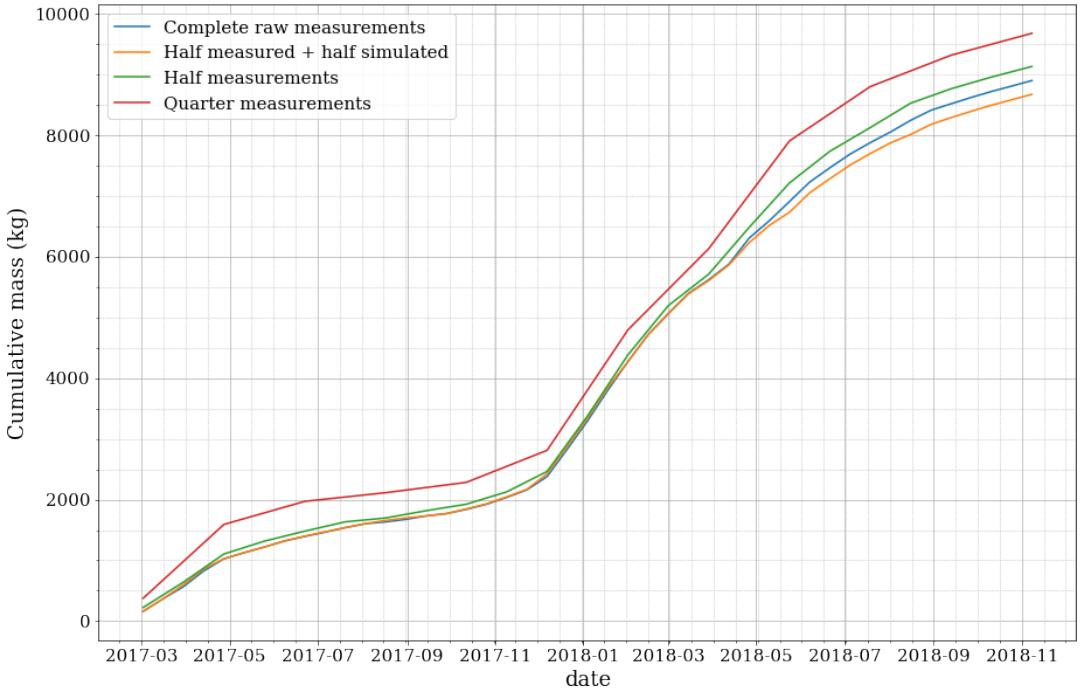
In conclusion, by comparing the three scenarios, it can be inferred that halving the measurement points better preserved the data properties compared to SARIMAX interpolating, however, further decreasing the measurement frequency by quartering the data is inadvisable.

3.4 ESTIMATION ON MASS (KG) OF THE SUBSTANCES

The cumulative mass of substances leaching out in the two testing years are calculated in four scenarios according to Equation 2.5. The results are shown in Figure 3.11.



(a) Chloride



(b) Ammonium

Figure 3.11: Cumulative mass leached out in the last 2 years (kg)

In the case of chloride (Figure 3.11 (a)), the blue curve(raw data) and the orange curve(modelled data) overlap to a large extent except for the slight deviation after May 2018, which means the model gives a result that is close to the real situation in the perspective of cumulative mass. In

the case of ammonium (Figure 3.11(b)), the blue curve and the orange curve overlap to a large extent until an obvious deviation shows up in May 2018. This deviation is due to the relatively large simulation error occurs in the later stage and is amplified by the cumulative effect.

The green curves (halved measurements) show relatively large deviation from the blue curves compared to the orange curves, the deviations between the red curves(quartered data) and blue curves are the most obvious one, which means the decreasing of the measurement frequency leads to a miscalculation on the cumulative mass.

To further quantify the error, the RMSEs on the mass estimation are calculated for the three experimental datasets (using the results of raw measurements as reference), the values are displayed in Table 3.6. According to the estimation based on raw measurements, the total mass leaching out in the testing two years is 9422.821 kg for chloride, and 9154.657 kg for ammonium. By normalizing the RMSEs to the total mass ($RMSE/mass_{tot}$), the results are shown in Table 3.7.

Table 3.6: RMSE of cumulative mass

	Chloride	Ammonium
Modelled data (kg)	54.726	125.248
Halved raw data(kg)	153.316	169.638
Quartered raw data(kg)	661.669	641.974

Table 3.7: Normalized RMSE of cumulative mass

	Chloride	Ammonium
Modelled data	0.581%	1.368%
Halved raw data	1.627%	1.853%
Quartered raw data	7.021%	6.813%

As the results in Table 3.6 show, by applying the model, the accuracy of mass estimation is improved compared to decreasing the number of measurements. However, in case 2 with halved data, the errors are small with respect to the total mass in that two years according to Table 3.7, thus, halved data sets can still be used to estimate the mass with a relatively high accuracy.

3.5 THE CHANGE OF CORRELATION BETWEEN ENDOGENOUS AND EXOGENOUS DATA

One important factor considered in SARIMAX is the correlation between its endogenous and exogenous data. By figuring out how do the correlations change after modelling, one can better understand SARIMAX model. The Pearson, Spearman coefficients before and after modelling are calculated respectively and shown in Table 3.8 and Table 3.9.

Table 3.8: Correlation coefficients changes of chloride model

	Chloride raw measurements		Chloride model	
	Pearson	Spearman	Pearson	Spearman
Outflow	-0.450	-0.440	-0.487	-0.467
EC	0.775	0.690	0.815	0.740
Rainfall	-0.134	-0.261	-0.160	-0.241
pEV	0.665	0.625	0.637	0.602

Table 3.9: Correlation coefficients changes of ammonium model

	Ammonium raw measurements		Ammonium model	
	Pearson	Spearman	Pearson	Spearman
Outflow	-0.302	-0.197	-0.332	-0.242
EC	0.576	0.510	0.622	0.534
Rainfall	-0.216	-0.357	-0.391	-0.408
pEV	0.497	0.448	0.528	0.470

By looking at Table 3.8 and Table 3.9 , the Pearson and Spearman coefficients are increased after modelling in all the cases, it reveals that the SARIMAX process may artificially create correlations that do not exist in the raw data.

3.6 SENSITIVITY ANALYSIS

For chloride, the coefficients in Table 3.1 indicate that the correlation between rainfall and chloride is apparently lower than that of the other three exogenous data, thus, it's of interest to find out the impact of removing rainfall from the exogenous data.

For ammonium, Table 3.1 tells us that only EC has a strong correlation with ammonium concentration, so the model of only considering EC as exogenous data will be investigated. All the scenarios are defined in Table 3.10 and the model results are displayed in Table 3.11.

Table 3.10: Sensitivity analysis scenarios

	Endogenous data	Exogenous data
Scenario 1	Chloride	Outflow, EC, rainfall, pEV
Scenario 2	Chloride	Outflow, EC, pEV
Scenario 3	Chloride	none
Scenario 4	Ammonium	Outflow, EC, rainfall, pEV
Scenario 5	Ammonium	EC
Scenario 6	Ammonium	none

Table 3.11: Model results of sensitivity analysis on exo-data

	RMSE of Concentration (mg/L)	RMSE of mass (kg)
Scenario 1	61.904	48.055
Scenario 2	60.901	54.276
Scenario 3	95.009	201.815
Scenario 4	79.697	122.761
Scenario 5	76.331	119.941
Scenario 6	77.196	137.202

For chloride, the RMSE of concentration simulation in scenario 2 is lower than that in scenario 1, however, the RMSE of the mass estimation is larger. From these results, one can conclude that the rainfall data doesn't have much effects on the model performance. The reason for this might be: the accuracy of the model results largely depends on the strongly correlated exogenous data, for example, EC and pEV. The weakly correlated one might not have many effects on the results, either being considered or not. What's more, the two RMSEs of scenario 3 are both larger than that in scenario 1 and 2, which indicates that the involving of strongly correlated exogenous data indeed improves the model performance.

As for ammonium, comparing the results of scenario 4 and scenario 5, one can see that the discarding of less correlated data indeed lead to better results for both concentration and mass estimation. In the case where all the exogenous data are weakly correlated, it's better to only involve the one with the strongest correlation.

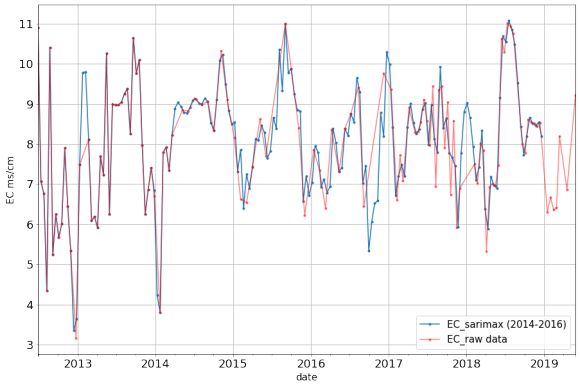
3.7 SARIMAX INTERPOLATION ON HISTORY DATA IN 2014–2016

Before conducting SARIMAX simulation in the two testing years, SARIMAX model is used to interpolate the historical data in the period with low-frequency (2014-2016), the results are compared to that of cubic spline interpolation.

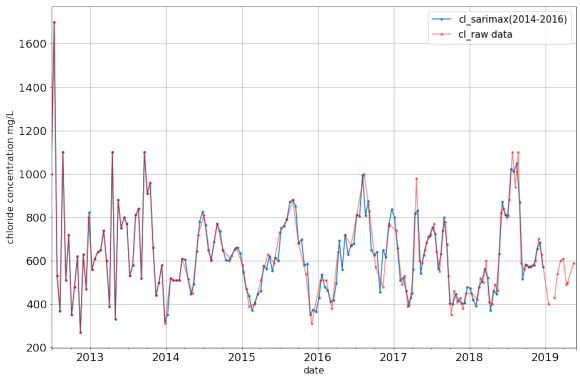
3.7.1 Data visualization and decomposition

The blue lines in Figure 3.12 are the SARIMAX interpolated (in 2014-2016) data, the red lines are the raw data without any interpolation.

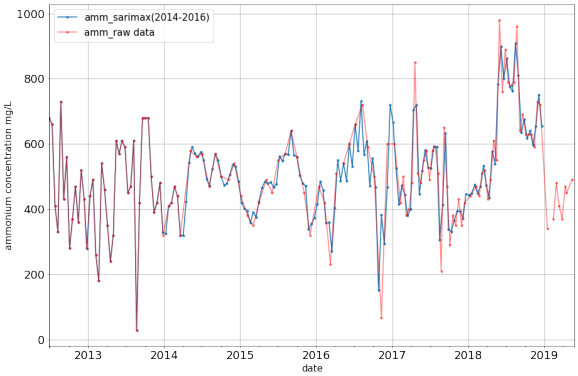
In 2014-2016, the blue line adds some small fluctuations in the intermediate points based on the red line, but the two lines substantially follow the same variation trend. When performing SARIMAX interpolation in 2014-2016, the available history data for training the model are all the data points before 2014, which are from June 2012 to December 2013. However, only 1.5 years of data available might not be sufficient for training the model, therefore, the accuracy of the prediction in 2014-2016 is reduced. Once the deviation exists in the historical data, eventually, the accuracy of the later modelling process (SARIMAX modelling on the two testing years) will be affected.



(a) EC



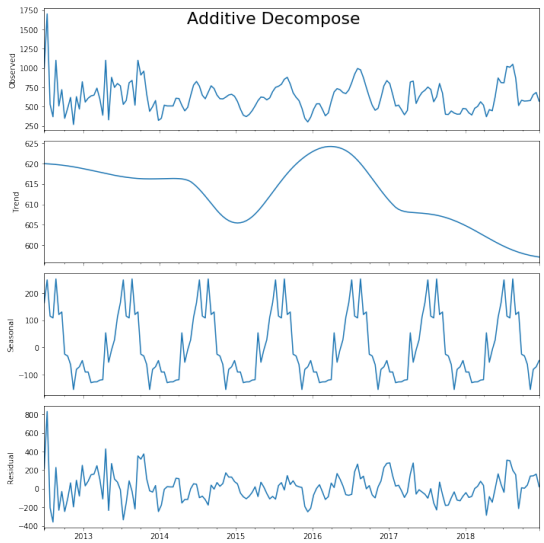
(b) Chloride



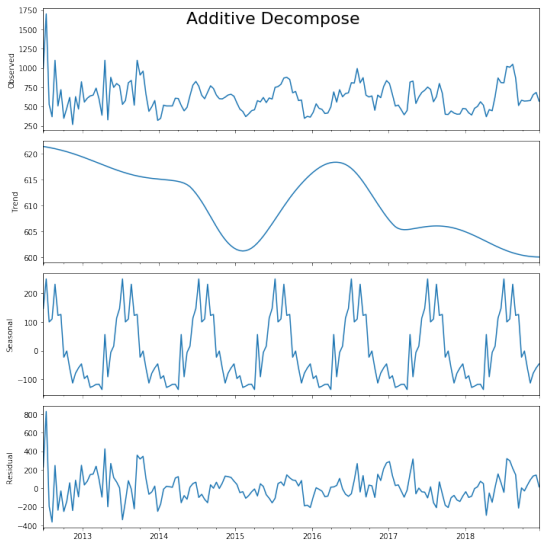
(c) Ammonium

Figure 3.12: Raw data and SARIMAX interpolation

The STL decomposition is applied on cubic-spline interpolated and SARIMAX interpolated data. The results are displayed in Figure 3.13 and Figure 3.14.

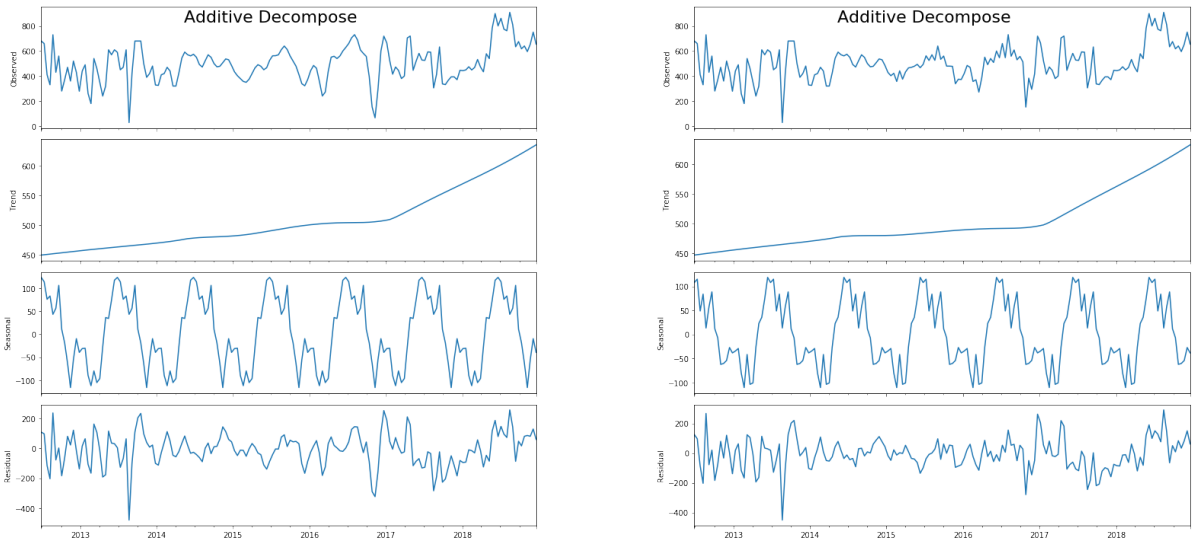


(a) Cubic spline interpolation



(b) SARIMAX interpolation

Figure 3.13: STL decomposition of chloride data



(a) Cubic spline interpolation

(b) SARIMAX interpolation

Figure 3.14: STL decomposition of ammonium data

From Figure 3.13 and Figure 3.14, one can see that the trend and the seasonality are similar in the two interpolation cases.

3.7.2 Error generated

For the two cases, the RMSEs of the model simulation in the two testing years (2017-2018) are listed in Table 3.12.

Table 3.12: RMSE of concentration estimation

	Chloride	Ammonium
SARIMAX interpolating on the history data (2014-2016)(mg/L)	94.518	120.926
Cubic interpolating on the whole history data(mg/L)	87.546	112.708

From the table, the simulation errors in the two testing years are larger in the case of SARIMAX interpolation than in the case of cubic spline interpolation.

3.7.3 Mass estimation

Table 3.13 shows the RMSEs of the cumulative mass estimation in the two testing years (2017-2018).

Table 3.13: RMSE of mass estimation

	Chloride	Ammonium
SARIMAX interpolating on the history data (2014-2016)(kg)	55.460	205.334
Cubic interpolating on the whole history data(kg)	48.055	122.761

Both in the case of chloride and ammonium, the SARIMAX interpolation gives higher RMSEs on the total mass estimation.

What should be mentioned is that, as for the equidistance of the time series is necessary when using SARIMAX model, the cubic spline interpolation is still needed in the first step to regularize the data set. When conducting the iterative interpolation during 2014-2016, the appended measurements are the corresponding cubic spline interpolated results rather than the real raw data.

To sum up, the cubic spline interpolation on the whole historical data set might be a better choice compared to SARIMAX interpolation in 2014-2016. Because to be used as history data in the later modelling process, the cubic spline interpolated data gives smaller errors both in concentration and mass simulation.

4 CONCLUSIONS

In this chapter, the most important findings in the thesis will be summarized.

From the decomposition results, it is noticed that the trend in chloride data is not obvious. A slowly ascending trend in the ammonium data is observed and the slope of the trend becomes steeper after 2017. However, despite the observed trend of ammonium data in decomposition, the statistical tests indicate that both chloride and ammonium data are statistical stationary.

The 2 years' SARIMAX simulation generated an RMSE of 88.212 mg/L for the chloride model and an RMSE of 113.397 mg/L for the ammonium model. Because of the stronger correlation with exogenous data and the more stable data structure, the accuracy of chloride model is higher than that of ammonium model. The mean of the relative simulation errors are both small and negligible in two cases, also, the data dynamics are well preserved after modelling.

The sensitivity analysis of the SARIMAX model indicates that the strongly linear correlated exogenous data is helpful to the model simulation. In the case where all available exogenous data are of weak linear correlation, discarding of the data with weaker correlation may improve the model performance.

In both chloride and ammonium cases, the modelled data and raw data can be regarded as from the same distribution according to the Mann-Whitney U test. However, the data properties (mean, std, density distribution) are better preserved with the data of halved measurements, and the halved data and raw data are more likely to come from the same distribution. After further decreasing the number of measurements, the quartered data shows obvious deviations on data properties.

By looking at the estimation of the cumulative mass leaching out in the last two years, the modelled data gave the result that is closer to the real case compared to the data with lower frequency. However, in case with halved data, the errors with respect to the total mass leached out in that two years (which are 1.627% for chloride case and 1.853% for ammonium case) are small and can be neglected.

In general, directly dropping half of the measurements can be regarded as an acceptable way to reduce the measurement frequency, with well preserved data properties and accurate estimation on mass of substances leached out. However, interpolating using the SARIMAX model doesn't have significant improvement in preserving the data properties. Further decreasing the measurement frequency by quartering the data is inadvisable.

For the future study, repeating the analysis with EC data obtained with a 15-minutes interval may give more insight in the error generated by this approach.

BIBLIOGRAPHY

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*,, pages 211–252.

Brownlee, J. (2016). How to check if time series data is stationary with python. Retrieved from https://machinelearningmastery.com/time-series-data-stationary-python/.

Deborah J. Rumsey (2016). *Statistics For Dummies, 2nd Edition*. Retrieved from https://www.dummies.com/education/math/statistics/what-a-p-value-tells-you-about-statistical-data/.

Fulton, C. (2018). Sarimax issue #4456. Retrieved from https://github.com/statsmodels/statsmodels/issues/4456. Accessed: 2019-09-25.

Kolb, W. M. (1984). *Curve fitting for programmable calculators*. Imtec.

Kwiatkowski, D., Phillips, P. C., Schmidt, P., and Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, pages 159–178.

Minitab (2019). Linear, nonlinear, and monotonic relationships. Retrieved from https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/supporting-topics/basics/linear-nonlinear-and-monotonic-relationships/.

Palachy, S. (2019). Stationarity in time series analysis. Retrieved from https://towardsdatascience.com/stationarity-in-time-series-analysis-90c94f27322.

Ramesh Sridharan (2011). Lecture notes from mit: Statistics for research projects, chapter 5. http://www.mit.edu/~6.s085/notes/lecture5.pdf. [Online; accessed 10-September-2019].

Rob, J. H. and George, A. (2018). *Forecasting: principles and practice, 2nd edition*. Monash University, Australia. Retrieved from https://otexts.com/fpp2/.

Robert Nau (2019). Statistical forecasting: notes on regression and time series analysis. Retrieved from https://people.duke.edu/~rnau/411home.htm.

Rohwerder, B. (2017). Solid waste and faecal sludge management in situations of rapid, mass displacement. (k4d helpdesk report 228). Brighton, UK: Institute of Development Studies.

Rosie Shier (2004). Statistics:2.3 the mann-whitney u test. *Mathematics Learning support center*. Retrieved from https://www.lboro.ac.uk/media/wwwlboroacuk/content/mlsc/downloads/2.3_mann_whitney.pdf.

Stephanie (2016). Unit root: Simple definition, unit root tests. Retrieved from https://www.statisticshowto.datasciencecentral.com/unit-root/.

Tukey, J. W. (1993). Exploratory data analysis: past, present and future. Technical report, PRINCETON UNIV NJ DEPT OF STATISTICS.

Wikipedia (2019a). Akaike information criterion — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Akaike_information_criterion&oldid=918068894. [Online; accessed 29-September-2019].

Wikipedia (2019b). Autoregressive integrated moving average — Wikipedia, the free encyclopedia. Retrieved from https://en.wikipedia.org/w/index.php?title=Autoregressive_integrated_moving_average&oldid=914714596.

A APPENDIX A

In this appendix, a complete example going through all steps of the SARIMAX model on chloride data will be described.

Step 1

Following the method described in Section 2.1, all data sets are made equidistant in time. The ADF and KPSS results show that the data can be considered to be stationary.

Step 2

The autocorrelation and partial autocorrelation plots are used to determine the order of differencing.

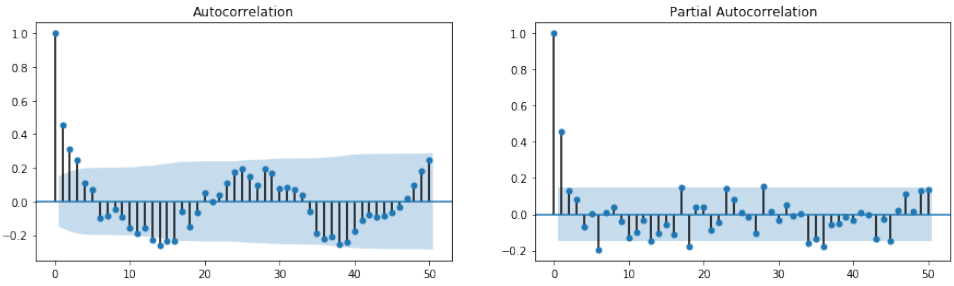


Figure A.1: ACF and PACF on raw chloride data

As the ACF graph showes, the raw data is clearly not stationary enough giving the repeating pattern and not dropping to zero after 50 lags. By conducting a seasonal differencing and then a first order differencing, the ACF and PACF give the results as following.

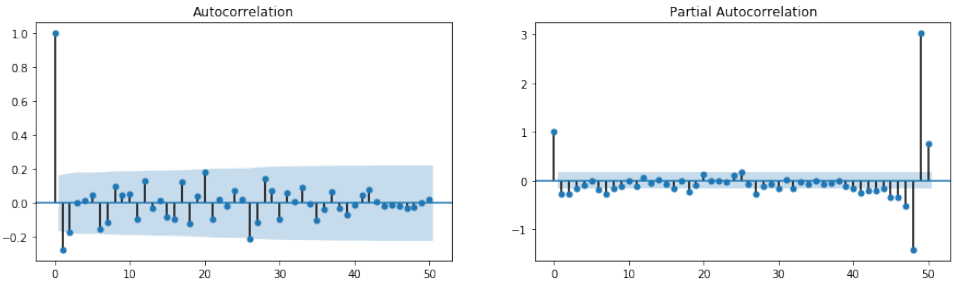


Figure A.2: ACF and PACF on chloride data after differencing

As the ACF graph showes, the lag-1 autocorrelation is negative and the ACF drops to zero relatively quickly, therefore, it can be regarded as stationary after differencing twice.

Seasonal differencing: a 26-lag difference of the series. This corresponds to a seasonal difference order of 1(d).

First order differencing: a 1-lag difference of the 26-lag differenced series. This corresponds to a non-seasonal difference order of 1(D).

From the above, the parameter can determined as SARIMA(?, 1, ?)x(?, 1, ?).

Step 3

By conducting a grid search for p, q, P, Q values, the combination give the lowest AIC is ARIMA(0, 1, 1)x(1, 1, 0, 26) - AIC:1283.8095231963698.

Step 4

Fitting the model on chloride data using the selected parameter. Then conducting the model diagnostics.

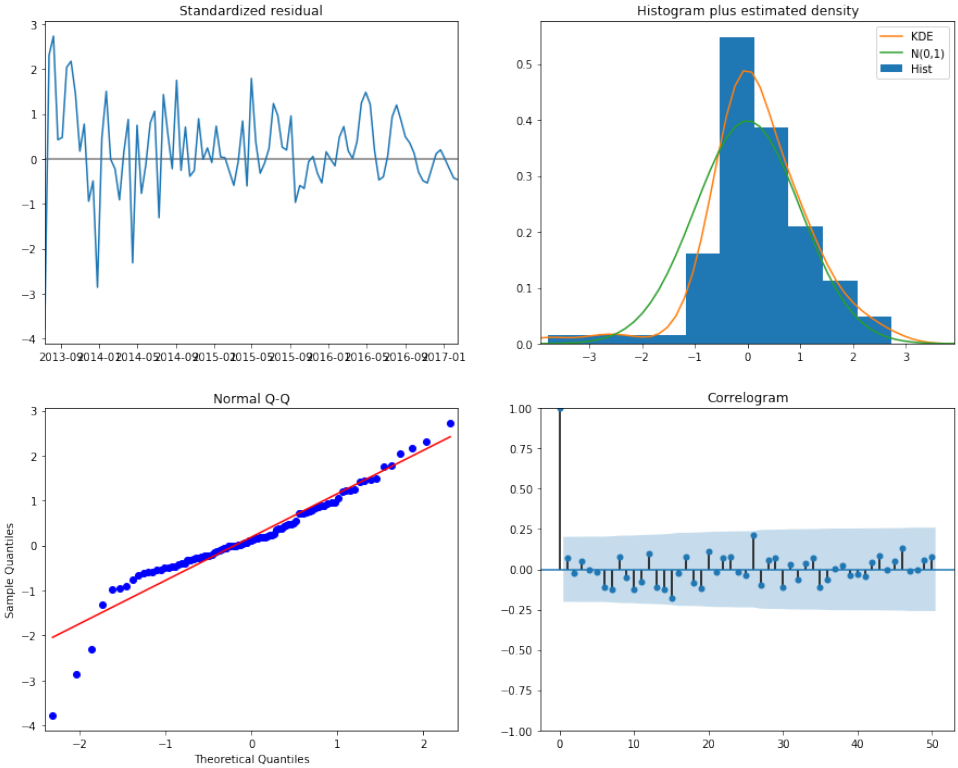


Figure A.3: Model diagnostics

Correlogram lookes quite well as no significant spikes appeared. However, the mean of the residual is 29.520 which is not close to zero, thus, adjustment on the data is needed.

Step 5

The Box-Cox transformation is used to fix this problem. After transforming the data, the proper parameters are identified using the same methods as before. Afterwards, we fitted a new model on the transformed data. As expected, the model diagnostics give the better results that the mean of residuals is 0.0362. This step finalizd the forward model.

Step 6

Conducting the iterative forecasting described in Section 2.3.4, predicted one point at a time, adding the predicted value to the training data set for fitting the next model.

Step 7

The simulated data is now in the Box-Cox transformed scale, it should be transformed back to the original scale using reverse Box-Cox transformation to obtain the modelled data.