

# EVALUATION OF STATISTICAL ANALYSES FOR THE IDENTIFICATION OF SURROGATES AND INDICATORS USING HISTORICAL PLANT DATA FROM A WATER RECLAMATION PLANT

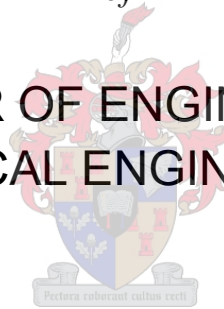
*by*

Cornelius Johannes Coomans

Thesis presented in partial fulfilment  
of the requirements for the Degree

*of*

MASTER OF ENGINEERING  
(CHEMICAL ENGINEERING)



in the Faculty of Engineering  
at Stellenbosch University

*Supervisor*

Dr L Auret

*Co-Supervisors*

Prof AJ Burger and Mr CD Swartz

March 2017

## ***Declaration***

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: March 2017

*Copyright © 2017 Stellenbosch University*

*All rights reserved*

## ABSTRACT

The lag time associated with water quality monitoring at water reclamation plants (WRPs) is a major hurdle in the way of implementing potable water reclamation in areas suffering from water shortages. The application of advanced monitoring techniques, which rely in part on surrogate and indicator variables, are one way of reducing the lag time associated with water quality monitoring.

The aim of this study was to evaluate statistical analyses that could be used to identify variable relationships, which in turn could be used for the development of surrogate and indicator variables, following the data-driven approach. The plant data used in this study were obtained from an existing WRP that has been operational for more than five years without undergoing any major changes to the treatment and operational procedures.

An initial assessment of the data found that the data contained large amounts of missing values. The assessment also identified the data periods during which the plant was operating under 'normal' conditions. Several time periods were removed since abnormal events occurred during these time periods.

Pre-processing the data consisted of outlier removal (three sigma rule and Hampel filter), noise reduction (moving average filter) and missing data replacement (linear interpolation). The statistical analyses, Pearson's and Spearman's correlation, principal component analysis (PCA), linear discriminant analysis (LDA) and partial least squares (PLS) regression, were then incorporated into models for identifying variable relationships. The performance of the different statistical analyses were measured using statistical metrics such as  $R^2$  for correlation, visualisation of separation for PCA, classification error for LDA and both  $R^2$  and mean squared error (MSE) for the PLS models.

The bivariate correlations provided the most concise results, whilst the LDA models could not be effectively assessed due to a change in the behaviour of the training and testing data. The PLS models performed poorly and did not produce any significant results. Expert process knowledge was also used to determine which variable relationships, identified by the models, could be regarded as valuable contributions, and which ought to be regarded as trivial.

Overall it was found that the bivariate correlations were effective for detecting relationships between variables. PCA was a valuable tool that provided insight into the potential use of multivariate analyses. LDA and PLS regression may require further testing before a definitive ruling can be made regarding their usefulness for identifying variable relationships from unprocessed historical plant data.

Although historical data could be used to identify variable relationships using bivariate correlations, it is not recommended for multivariate statistical analyses. A planned sampling campaign could be much more effective for data collection than using historical data, although the cost associated with a planned sampling campaign must be taken into consideration.

## OPSOMMING

Die tydsverloop wat verband hou met watergehaltemonitoring by waterherwinningswerke (WHW's) is 'n groot hindernis vir die implementering van drinkbarewaterherwinning in gebiede wat onder watertekorte gebuk gaan. Die toepassing van gevorderde monitoringstegnieke wat gedeeltelik staatmaak op surrogaat- en aanwyserveranderlikes is een manier om hierdie tydsverloop te verminder.

Die doel van hierdie studie was om statistiese ontledings te evalueer wat gebruik kan word om veranderlike verhoudings, wat aangewend kan word vir die ontwikkeling van surrogaat- en aanwyserveranderlikes, op grond van die data-gedrewe benadering te identifiseer. Die aanlegdata wat vir hierdie navorsing gebruik is, verkry vanaf 'n bestaande WHW wat reeds vir vyf jaar werksaam is sonder dat enige groot veranderinge aan behandelings en bedryfsprosedures ondergaan is.

Deur 'n aanvanklike assessering van die data is bevind dat die data groot hoeveelhede ontbrekende waardes bevat. Met die assessering is datatyperke ook geïdentifiseer waartydens die aanleg onder 'normale' omstandighede bedryf is. Verskeie tydperke is verwyder aangesien abnormale gebeure daartydens plaasgevind het.

Voorverwerking van die data het begin met uitskieterverwydering (driesigma-reël en Hampel-filter), geraasvermindering (bewegendegemiddelde-filter) en ontbrekende data-vervanging (lineêre interpolasie). Die statistiese ontledings, Pearson en Spearman se korrelasie, hoofkomponentontleding (PCA), lineêre diskriminantontleding (LDA) en gedeeltelike kleinste kwadrate- (PLS-)regressie is in modelle gebruik vir die identifisering van veranderlike verhoudings. Die prestasie van die statistiese ontledings is gemeet met behulp van statistiese maatstawwe soos  $R^2$  vir korrelasie, visualisering van skeiding vir PCA, klassifikasiefout vir LDA en sowel  $R^2$  as gemiddelde kwadraatfout vir die PLS-modelle.

Die tweeveranderlike korrelasies het die bondigste resultate getoon, terwyl die LDA-modelle nie doeltreffend beoordeel kon word nie as gevolg van 'n verandering in die gedrag van die opleiding- en toetsdata. Die PLS-modelle het swak presteer en het nie enige noemenswaardige resultate gelewer nie. Deskundige proses kennis is ook gebruik om te bepaal watter veranderlike verhoudings, wat deur die modelle geïdentifiseer is, as waardevolle bydraes beskou kon word, en watter as onbeduidend beskou behoort te word.

In die algemeen is bevind dat die tweeveranderlike korrelasies doeltreffend was vir die identifisering van verwantskappe tussen veranderlikes. PCA was 'n waardevolle instrument wat insig verskaf het in die potensiële gebruik van meerveranderlike ontledingstegnieke. LDA- en PLS-regressie vereis moontlik verdere toetsing voordat 'n finale beslissing gemaak kan word met betrekking tot die nut daarvan vir die identifisering van veranderlike verhoudings deur gebruik te maak van onverwerkte historiese aanlegdata.

Hoewel historiese data gebruik kon word om veranderlike verhoudings met behulp van tweeveranderlike korrelasies te identifiseer, word dit nie aanbeveel vir meerveranderlike statistiese ontledings nie. 'n Beplande steekproefnemingsveldtog kan baie doeltreffender wees vir data-insameling

as die gebruik van historiese data, hoewel die koste wat verband hou met 'n beplande steekproefnemingsveldtog in ag geneem moet word.

## **ACKNOWLEDGEMENTS**

Dr L Auret, University of Stellenbosch

Prof AJ Burger, University of Stellenbosch

Mr CD Swartz, Chris Swartz Water Utilization Engineers

Dr JP Barnard, University of Stellenbosch

Mr JG Menge

Mr J Esterhuizen

Mr S Muller

Mr R Mertens

Mr K Nikodemus

Mr R Munyandi

Water Research Commission

My wife, Reinette, for her love, support and encouragement

My parents, Johan and Alet, for raising me in God's ways

God the Father, my creator

Jesus Christ, my saviour

Holy Spirit, my comforter

## TABLE OF CONTENT

<b>ABSTRACT</b> .....	<b>III</b>
<b>OPSOMMING</b> .....	<b>IV</b>
<b>ACKNOWLEDGEMENTS</b> .....	<b>VI</b>
<b>TABLE OF CONTENT</b> .....	<b>VII</b>
<b>LIST OF FIGURES</b> .....	<b>X</b>
<b>LIST OF TABLES</b> .....	<b>XIII</b>
<b>ACRONYMS AND ABBREVIATIONS</b> .....	<b>XIV</b>
<b>1 INTRODUCTION</b> .....	<b>1</b>
1.1 IMPORTANT DEFINITIONS.....	1
1.1.1 Mathematical notation .....	1
1.1.2 Water reclamation plant monitoring and soft sensors .....	1
1.1.3 Parameter and variable definitions regarding surrogates and indicators .....	3
1.1.4 Treatment technologies, processes and units: Performance and testing .....	5
1.2 BACKGROUND .....	7
1.3 AIM AND APPROACH .....	11
<b>2 WATER RECLAMATION PLANT</b> .....	<b>13</b>
2.1 WATER RECLAMATION OVERVIEW .....	13
2.1.1 Types of water reclamation .....	13
2.2 PROCESS OVERVIEW .....	16
2.3 DESCRIPTION OF TREATMENT UNITS .....	17
2.3.1 Dissolved Air Flotation.....	17
2.3.2 Sand Filter .....	18
2.3.3 Ozonation .....	19
2.3.4 Biologically Activated Carbon.....	20
2.3.5 Granular Activated Carbon.....	20
2.3.6 Ultrafiltration .....	21
2.3.7 Chlorination .....	22
<b>3 LITERATURE REVIEW</b> .....	<b>23</b>
3.1 PLANT MONITORING.....	23
3.1.1 Risk management .....	23

3.1.2	Plant performance .....	25
3.2	THE KNOWLEDGE BASED APPROACH FOR SURROGATE AND INDICATOR DEVELOPMENT.....	27
3.2.1	Selection of relevant variables .....	28
3.2.2	Experimental setup.....	29
3.2.3	Sampling and analytical procedures .....	30
3.2.4	Results typically obtained.....	31
3.3	THE DATA-DRIVEN APPROACH FOR SURROGATE AND INDICATOR DEVELOPMENT.....	35
3.3.1	Data validation techniques .....	36
3.3.2	Data preparation and pre-processing.....	38
3.3.3	Exploratory data analysis .....	43
3.3.4	Univariate techniques.....	44
3.3.5	Multivariate techniques.....	45
3.4	SUMMARY OF LITERATURE REVIEW .....	52
<b>4</b>	<b>METHODOLOGY .....</b>	<b>54</b>
4.1	OBTAINING, ORGANISING AND ASSESSING HISTORICAL PLANT DATA .....	54
4.1.1	Organising and initial selection of plant data.....	54
4.1.2	Data validation and final selection of plant data.....	55
4.1.3	Data pre-processing .....	57
4.2	DEVELOPING STATISTICAL MODELS FOR THE IDENTIFICATION OF SURROGATES AND INDICATORS.....	61
4.2.1	Bivariate statistical analyses .....	61
4.2.2	Multivariate statistical analyses.....	62
4.3	ASSESSING THE PERFORMANCE OF THE DEVELOPED STATISTICAL MODELS..	66
4.4	IDENTIFYING AND EVALUATING RELEVANT VARIABLE RELATIONSHIPS USING EXPERT PROCESS KNOWLEDGE .....	68
<b>5</b>	<b>RESULTS AND DISCUSSION.....</b>	<b>70</b>
5.1	DATA SELECTION AND VALIDATION .....	70
5.1.1	Data selection.....	70
5.1.2	Data validation.....	71
5.2	DATA PRE-PROCESSING .....	77



5.2.1	Outlier analysis .....	77
5.2.2	Noise attenuation .....	80
5.2.3	Missing data replacement .....	83
5.3	DEVELOPMENT AND PERFORMANCE OF THE STATISTICAL MODELS .....	84
5.3.1	Bivariate analyses .....	84
5.3.2	Multivariate analyses .....	89
5.4	SUMMARY AND EVALUATION OF IDENTIFIED RELATIONSHIPS USING EXPERT PROCESS KNOWLEDGE .....	109
5.4.1	Correlation analysis .....	110
5.4.2	Linear Discriminant Analysis .....	115
5.4.3	Partial Least Squares regression .....	118
<b>6</b>	<b>CONCLUSIONS .....</b>	<b>121</b>
6.1	OBTAINING, ORGANISING AND ASSESSING HISTORICAL PLANT DATA .....	121
6.2	DEVELOPMENT AND PERFORMANCE OF THE STATISTICAL MODELS .....	121
6.2.1	Data pre-processing .....	121
6.2.2	Bivariate correlation analysis .....	122
6.2.3	Principal Component Analysis .....	122
6.2.4	Linear Discriminant Analysis .....	122
6.2.5	Partial Least Squares regression .....	123
6.3	EVALUATION OF IDENTIFIED RELATIONSHIPS USING EXPERT PROCESS KNOWLEDGE .....	123
6.4	OVERALL CONCLUSIONS .....	124
<b>7</b>	<b>RECOMMENDATIONS AND FUTURE WORK .....</b>	<b>125</b>
	<b>REFERENCES .....</b>	<b>XIV</b>
	<b>APPENDIX A: DETAILED INFORMATION FOR ALL VARIABLES INCLUDED IN THE STUDY ...</b>	<b>XXI</b>
	<b>APPENDIX B: SUFFICIENT REPRESENTATION DATA VALIDATION RESULTS .....</b>	<b>XXVIII</b>
	<b>APPENDIX C: LIST OF SIGNIFICANT CORRELATIONS FROM PEARSON AND SPEARMAN CORRELATION ANALYSES .....</b>	<b>XLII</b>

## LIST OF FIGURES

<b>Figure 2.1:</b> Different types of water reclamation .....	14
<b>Figure 3.1:</b> Illustration of classical and exploratory data analysis philosophies.....	44
<b>Figure 4.1:</b> Illustration of training and testing data separation .....	57
<b>Figure 4.2:</b> Block diagram illustrating outlier removal algorithm .....	58
<b>Figure 4.3:</b> Block diagram of noise removal algorithm.....	60
<b>Figure 4.4:</b> Illustration of the creation of the <b>X</b> and <b>Y</b> data sets .....	62
<b>Figure 4.5:</b> Box diagram of numeric to categorical transformation algorithm .....	64
<b>Figure 4.6:</b> Block diagram of PLS workflow .....	65
<b>Figure 4.7:</b> Confusion matrix plot example .....	67
<b>Figure 5.1:</b> Number of missing values per variable (Swartz, et al., 2016) .....	71
<b>Figure 5.2:</b> Percentage missing values per variable (Swartz, et al., 2016) .....	72
<b>Figure 5.3:</b> Number of dirty data points per variable (Swartz, et al., 2016) .....	73
<b>Figure 5.4:</b> Number of unique values per variable (Swartz, et al., 2016) .....	74
<b>Figure 5.5:</b> Percentage of unique values per variable (Swartz, et al., 2016).....	74
<b>Figure 5.6:</b> Overall mass balance (accumulated volume) (Swartz, et al., 2016) .....	75
<b>Figure 5.7:</b> Run length plot for all variables in the data (Swartz, et al., 2016) .....	76
<b>Figure 5.8:</b> Outlier results for UV <sub>254</sub> in the WWTP Final Effluent (empirical) .....	77
<b>Figure 5.9:</b> Outlier results for Nitrate in the WWTP Final Effluent (normal).....	78
<b>Figure 5.10:</b> Outlier results for DOC in the WWTP Final Effluent (lognormal).....	78
<b>Figure 5.11:</b> Outlier results with Hampel filter for UV <sub>254</sub> in the WWTP Final Effluent (empirical) .....	79
<b>Figure 5.12:</b> Outlier results with Hampel filter for Nitrate in the WWTP Final Effluent (normal) .....	79
<b>Figure 5.13:</b> Outlier results with Hampel filter for DOC in the WWTP Final Effluent (lognormal).....	79
<b>Figure 5.14:</b> Variable removed due to lack of variance .....	80
<b>Figure 5.15:</b> Noise reduction results for pH from the DAF (window size = 5).....	80
<b>Figure 5.16:</b> Noise reduction results for pH from the DAF (window size = 11).....	81
<b>Figure 5.17:</b> Noise reduction results for pH from the DAF (window size = 25).....	81
<b>Figure 5.18:</b> Noise reduction results for Total Alkalinity from the DAF (window size = 5) .....	81
<b>Figure 5.19:</b> Noise reduction results for Total Alkalinity from the DAF effluent (window size = 11) ....	82
<b>Figure 5.20:</b> Noise reduction results for Total Alkalinity from the DAF effluent (window size = 25) ....	82
<b>Figure 5.21:</b> SF D runtime with spline interpolation showing overfitting .....	83
<b>Figure 5.22:</b> SF D runtime with linear interpolation.....	84
<b>Figure 5.23:</b> Pearson's correlation coefficient before separating the MVR .....	85
<b>Figure 5.24:</b> Spearman's correlation coefficient before separating the MVR .....	85
<b>Figure 5.25:</b> Pearson's correlation coefficient after separating the MVR .....	87
<b>Figure 5.26:</b> Spearman's correlation coefficient after separating the MVR .....	87
<b>Figure 5.27:</b> Variance explained by principal components .....	89
<b>Figure 5.28:</b> Multi-plot of scatter plots for the first four principal components .....	90
<b>Figure 5.29:</b> Scatter plot of 1 <sup>st</sup> and 2 <sup>nd</sup> PC coloured for Nitrate from the WWTP Clarifier .....	91

<b>Figure 5.30:</b> Scatter plot of 1 <sup>st</sup> and 2 <sup>nd</sup> PC coloured for Nitrite from the WWTP Clarifier .....	91
<b>Figure 5.31:</b> Scatter plot of 1 <sup>st</sup> and 2 <sup>nd</sup> PC coloured for EC from the UF .....	92
<b>Figure 5.32:</b> Scatter plot of 3 <sup>rd</sup> and 2 <sup>nd</sup> PC coloured for EC from the UF .....	92
<b>Figure 5.33:</b> Scatter plot of 4 <sup>th</sup> and 2 <sup>nd</sup> PC coloured for EC from the UF .....	93
<b>Figure 5.34:</b> Scatter plot of 1 <sup>st</sup> and 2 <sup>nd</sup> PC coloured for TDS (Calc) from the UF .....	93
<b>Figure 5.35:</b> Scatter plot of 3 <sup>rd</sup> and 2 <sup>nd</sup> PC coloured for TDS (Calc) from the UF .....	94
<b>Figure 5.36:</b> Scatter plot of 4 <sup>th</sup> and 2 <sup>nd</sup> PC coloured for TDS (Calc) from the UF .....	94
<b>Figure 5.37:</b> Scatter plot of 1 <sup>st</sup> and 2 <sup>nd</sup> PC coloured for TDS (Calc) from the UF (testing data).....	95
<b>Figure 5.38:</b> Confusion plot for EC from the UF .....	96
<b>Figure 5.39:</b> Confusion plot for Nitrate from the WWTP Clarifier .....	96
<b>Figure 5.40:</b> Confusion plot for Nitrite from the WWTP Clarifier .....	97
<b>Figure 5.41:</b> Confusion plot for UV <sub>254</sub> in the WWTP Final Effluent .....	97
<b>Figure 5.42:</b> Confusion plot for Residual O <sub>3</sub> from the Ozone Contact B.....	98
<b>Figure 5.43:</b> Confusion plot for pH from the WWTP Clarifier .....	99
<b>Figure 5.44:</b> MSEP per component for Temperature from the WWTP Clarifier .....	100
<b>Figure 5.45:</b> Variance explained per component for Temperature from the WWTP Clarifier .....	100
<b>Figure 5.46:</b> MSEP per component for Nitrite from the WWTP Clarifier .....	101
<b>Figure 5.47:</b> Variance explained per component for Nitrite from the WWTP Clarifier .....	101
<b>Figure 5.48:</b> MSEP per component for Chlorophyll A in the WRP Influent.....	102
<b>Figure 5.49:</b> MSEP per component for Calcium Hardness in the WWTP Final Effluent .....	102
<b>Figure 5.50:</b> Variance explained per component for Chlorophyll A in the WRP Influent .....	103
<b>Figure 5.51:</b> Variance explained per component for Calcium Hardness in the WWTP Final Effluent .....	103
<b>Figure 5.52:</b> MSEP per component for Clostridium Spores in the WWTP Final Effluent .....	104
<b>Figure 5.53:</b> MSEP per component for Faecal Coliform in the WWTP Final Effluent.....	104
<b>Figure 5.54:</b> Variance explained per component for Clostridium Spores in the WWTP Final Effluent .....	105
<b>Figure 5.55:</b> Variance explained per component for Faecal Coliform in the WWTP Final Effluent ...	105
<b>Figure 5.56:</b> Predicted vs Observed scatter for Temperature from the WWTP Clarifier .....	106
<b>Figure 5.57:</b> Predicted vs Observed scatter for Nitrite from the WWTP Clarifier.....	107
<b>Figure 5.58:</b> Predicted vs Observed scatter for Chlorophyll A in the WRP Influent.....	107
<b>Figure 5.59:</b> Predicted vs Observed scatter for Calcium Hardness in the WWTP Final Effluent .....	108
<b>Figure 5.60:</b> Predicted vs Observed scatter for Clostridium Spores in the WWTP Final Effluent .....	109
<b>Figure 5.61:</b> Predicted vs Observed scatter for Faecal Coliform in the WWTP Final Effluent .....	109
<b>Figure 5.62:</b> Summary of Pearson correlations for each sampling point.....	111
<b>Figure 5.63:</b> Summary of Spearman correlations for each sampling point.....	112
<b>Figure 5.64:</b> Total Alkalinity from the Sand Filter against time .....	116
<b>Figure 5.65:</b> Potassium in the WRP Influent against time .....	116
<b>Figure 5.66:</b> Potassium in the Final Water against time .....	117
<b>Figure 5.67:</b> Summary of LDA results .....	118

**Figure 5.68:** Summary of PLS results ..... 119

## LIST OF TABLES

<b>Table 2.1:</b> Process design parameters for each unit from the studied plant.....	17
<b>Table 2.2:</b> Data obtained for DAF.....	18
<b>Table 2.3:</b> Data obtained for sand filter .....	19
<b>Table 2.4:</b> Data obtained for Ozone Contact.....	20
<b>Table 2.5:</b> Data obtained for BAC .....	20
<b>Table 2.6:</b> Data obtained for GAC .....	21
<b>Table 2.7:</b> Data obtained for UF .....	22
<b>Table 2.8:</b> Data obtained for final water .....	22
<b>Table 3.1:</b> Surrogates and indicators for treatment performance via the knowledge based approach	32
<b>Table 3.2:</b> Variables suggested for treatment process monitoring.....	32
<b>Table 3.3:</b> Surrogates used for treatment process performance validation .....	34
<b>Table 4.1:</b> Description of data IDs .....	55
<b>Table 4.2:</b> Variable distribution and limits used for outlier detection .....	58
<b>Table 4.3:</b> LDA limits used to categorise variables .....	64
<b>Table 5.1:</b> FEM variables according to expert process knowledge.....	110
<b>Table 5.2:</b> Top five variables with significant Pearson correlations per sampling location .....	113
<b>Table 5.3:</b> Pearson's correlation coefficients of interest from before separating the MVR .....	114
<b>Table 5.4:</b> Spearman's correlation coefficients of interest from before separating the MVR.....	114

## ACRONYMS AND ABBREVIATIONS

ADWT	advanced drinking water treatment
AGWR	Australian guidelines for water recycling
AR	autoregressive
BAC	biological activated carbon
BDOC	biodegradable dissolved organic carbon
BNR	biological nutrient removal
BOD	biological oxygen demand
CA	cluster analysis
CCP	critical control point
CDPH	California Department of Public Health
CEB	chemically enhanced backwash
CEC	contaminant of emerging concern
CIP	clean in place
COD	chemical oxygen demand
DAF	dissolved air flotation
DBP	disinfection by-product
DEET	diethyltoluamide
DMF	dual media filtration
DOC	dissolved organic carbon
DPR	direct potable reuse
EBCT	empty bed contact time
EDA	exploratory data analysis
EfOM	effluent organic matter
ES	effective size
FEM	fast and easy to measure
FIR	finite impulse responses
GAC	granular activated carbon
IIR	infinite impulse responses
IPR	indirect potable reuse
LDA	linear discriminant analysis
LIMS	laboratory information management system
LnO	leave-n-out
LOO	leave-one-out
LRV	log removal value
MA	moving average
MAD	median absolute deviation
MAX	moving average adaptive exponential
MBR	membrane bio-reactor
MF	microfiltration
MLR	multiple linear regression
MSE	mean square error
MSEP	mean square error of prediction
MVR	multivariate record
MW	molecular weight

NDMA	n-nitrosodimethylamine
NF	nanofiltration
NIPALS	non-linear interactive partial least squares
NOM	natural organic matter
NPR	non-potable reuse
NTU	nephelometric turbidity unit
NWRI	National Water Research Institute
NWS	noise window size
PC	principal component
PCA	principal component analysis
PCR	principal component regression
PLS	partial least squares
PSA	pressure swing absorption
P-SDM	pseudo slow and difficult to measure
QA	quality assurance
QC	quality control
RMSE	root mean square error
RO	reverse osmosis
SAT	soil aquifer treatment
SCADA	supervisory computer and data acquisition
SDM	slow and difficult to measure
SF	sand filter
SIMPLS	simple partial least squares
SLR	solid loading rate
SMP	soluble microbial product
SWRO	salt water reverse osmosis
TOC	total organic carbon
TOX	total organic halogen
TrOC	trace organic compounds
UF	ultrafiltration
UV	ultra violet
UVA	ultra violet absorbance
VOD	vent ozone destruction
WRP	water reclamation plant
WSP	water safety plan
WTP	water treatment plant
WWTP	wastewater treatment plant

# 1 INTRODUCTION

This chapter is aimed at providing the reader with the necessary information required to understand the background (history and current state) of the field of research that will be presented. This chapter will also discuss the aim and approach followed. Unfortunately, these discussions require the use of specific terms that may not be known to the reader.

This chapter therefore starts with a sub-section explaining the relevant terms that will be used throughout. Should the reader be familiar with this field of research, this sub-section may be disregarded, or simply used retrospectively.

The intention of this research was to study the operation and performance of water reclamation plants (WRPs) in an attempt to unveil certain patterns or variable behaviours that could be exploited in order to simplify and reduce the cost of monitoring programmes. These monitoring programmes are tasked with ensuring the safety of the final water produced by WRPs, as well as prolonging the lifetime of the equipment and treatment technologies used on the plant whilst also optimising the operation of the plant in order to reduce costs.

## 1.1 IMPORTANT DEFINITIONS

### 1.1.1 Mathematical notation

Throughout the report, mathematical equations and notations are used in order to simplify or elaborate (in mathematical terms) what is being discussed, as well as the actual equations used during the analyses. It should be noted that the data will always be referred to in terms of linear algebra, which is to say that the data are considered a matrix composed of a number of column vectors containing an equal amount of records (although some records may be empty due to missing data).

The following notation style will be used throughout the paper, unless for some specific illustrative purpose. It is recommended that the reader refers back to this section, should any mathematical terms appear ambiguous.

Term	Mathematical notation	Short hand	Note
Matrix with size	$\mathbf{X} [n, m]$ or $\mathbf{X}_{n,m}$	$\mathbf{X}$	Upper case, bold, italic
Variable (column vector)	$\mathbf{x}_j$	$\mathbf{x}$	Lower case, bold, italic
Sample (data element)	$\mathbf{x}_{i,j}$	$\mathbf{x}_{i,j}$	Lower case, bold, italic, with column and row identifiers

### 1.1.2 Water reclamation plant monitoring and soft sensors

The various monitoring practises used at WRPs, types of data obtained from WRPs and methods used to obtain data at WRPs are defined here.



- Plant monitoring:** In order to ensure that the plant is operating under optimal conditions (the design conditions of the plant) it is necessary to continually measure the various parameters and variables that pertain to the design conditions of the plant. The process of continually measuring plant conditions can be referred to as plant monitoring. Plant monitoring typically consists of operational monitoring and compliance monitoring, each concerning the following groups of data, namely operational data or quality data, respectively.
- Operational data:** Also referred to as operational control data; this group of data relates to the functioning and operation of the treatment units that make up the treatment system (e.g. flow rates, differential pressure, dosing rates of chemicals).
- Quality data:** Also referred to as quality control data or water quality data; this group of data relates to the quality of the water produced by the various treatment units that make up the treatment system. If standards are available, this data will form part of compliance monitoring (e.g. pH, EC, Turbidity, COD, *E. coli*).
- Advanced monitoring:** Unlike traditional monitoring - which only makes use of direct parameter or variable measurements - advanced monitoring, or advanced monitoring systems, make use of certain measurements in order to infer the status or level of other parameters or variables that were not measured directly.
- Monitoring tool:** A monitoring tool is a technology, or process, that is used to provide variable measurements and may include the taking of samples and analysing samples to obtain measured values for a given variable. Conventional monitoring tools rely on direct variable measurements and may consist of sensors or probes, as well as analytical laboratory equipment and even automatic sampling devices. Advanced monitoring tools, on the other hand, does not make use of direct variable measurements and are an essential part of any advanced monitoring system, or programme. They rely on external data (easy to measure variables) coupled with computational algorithms in order to produce variable measurements in an indirect, or intuitive manner.
- Soft sensors:** A soft sensor is an advanced monitoring tool typically employed by advanced monitoring systems. The purpose of a soft sensor is to provide information of the quality or quantity of a given parameter or variable, just like a normal sensor. The difference is that a soft sensor is not a physical device, instead it is an equation used by an algorithm run by a computer, which then calculates the most probable quality or quantity of a parameter or variable based on the information received by the computer (typically measurements from traditional sensors).

### 1.1.3 Parameter and variable definitions regarding surrogates and indicators

First and foremost, the definition and distinction between the terms ‘parameter’ and ‘variable’ require clarity.

**Parameter:** A parameter is a descriptive characteristic of a population. Any parameter is fixed for a given population and will only change if the population changes. Therefore, the mean score for a test will be a finite value, until another test is written, or the marks of some tests are altered.

**Variable:** A variable is any measured value that may vary during observation. In terms of the above example, the score of a given test is a variable that changes from one student to the next. In combination it can therefore be seen that variables make up a population and parameters measure characteristics of the population.

For instance, a treatment unit like a sand filter will have certain dimensions and fixed operational conditions. These can be used to describe the parameters of the unit (volume, retention time, etc.). But the temperature, or the pH, of the water in the unit cannot be referred to as parameters. Instead, they are variables.

Furthermore, if one regards the temperature or pH measurements for a given month, then the mean, standard deviation and variance of those measurements can be called parameters for the unit, for that month, but again, the temperature and pH measurements themselves are variables.

The following definitions and the remainder of the thesis will therefore make use of the term ‘variable’ in the correct, statistical, manner. The following terms, or acronyms rather, are of particular importance to the work done in this study:

**FEM variable:** Fast and easy to measure (FEM) variable are variables that can be accurately measured at a fast rate (less than a minute) and without much effort with regard to analytical equipment, procedures and personnel requirements.

**SDM variable:** Slow and difficult to measure (SDM) variables are variables that can only be accurately measured at a slow rate (more than an hour) and typically require a substantial amount of analytical equipment, procedures and personnel in order to obtain accurate results.

**P-SDM variable:** Pseudo slow and difficult to measure (P-SDM) variables are variables that have been categorised as SDM based on the historical data that are available for them (reflecting a slow measurement rate), but are in reality FEM variables. This misclassification of these variables are primarily due to decisions, or mistakes, made during data capturing.

The following three terms are often used together, or in conjunction with one another, but are by no means synonymous to one another.

**Target variable:** Any monitoring programme will identify certain target variables, which are of importance to the monitoring programme. Target variables are therefore the variables that are desired to be measured and quantified.

**Surrogate variable:** A surrogate variable is a measured variable that provides information about a target variable, but is not the target variable itself. Surrogate variables are practically easier to measure than the target variable and have a statistically quantifiable relationship with the target variable. E.g. UV254 for Ammonia

**Indicator variable:** A contaminant (chemical or pathogen) that has similar behavioural (pathway and/or removal) characteristics and/or physical properties (functional group) to another contaminant, or group thereof, that is preferably easier to detect and quantify. In most cases indicators are used in a qualitative manner, in order to indicate the presence or absence of certain contaminants.

The main difference between surrogates and indicators, therefore, is that surrogates are much easier to measure than the target variable, whereas indicator variables may also be difficult to measure, but can be used to represent a group of contaminants. Although the majority of surrogates are quantitative variables and the majority of indicators are qualitative variables, this is not the true distinction between surrogates and indicators since some indicators can be used quantitatively (Genthe and Kfir, 1992).

It may be easier to understand the differences of these variables by using mathematical notation. Given a target variable  $y$ , a surrogate or indicator for that variable can be described with the aid of the following equations:

$$\text{Equation 1:} \quad y = f(x, L, S) + e$$

$$\text{Equation 2:} \quad \hat{y} = \hat{f}(X)$$

Equation 1 shows a generic formulation for the behaviour of a target variable where  $x$  represents a variable accounting for water quality ( $x$  may also be multivariate, but for this example was kept univariate),  $L$  represents unmeasured operational process variations (including time of day),  $S$  indicates any unknown factors, and  $e$  indicates random error.

If Equation 1 can be reduced to Equation 2, then it may be assumed that the variable  $x$  (or one of the  $x$  variables in the  $X$  matrix) is either a surrogate or indicator for  $y$ . Whether  $x$  is an indicator or a surrogate depends on the characteristics of the variable. If  $x$  is a FEM variable, then it is a surrogate, and if  $x$  is a SDM variable, then  $x$  is an indicator variable. If a target variable  $y$  has both a surrogate variable and an indicator variable; only the surrogate variable will be used, since surrogates are much faster and easier to measure. It is also more likely that the surrogate variable will be able to produce quantitative information. The definition for qualitative and quantitative variables are:

**Quantitative variable:** A quantitative variable is a variable that provides quantitative (continuous or discrete) information that can be used to distinguish whether or not a variable level is within a given, or desired, range or not. These types of variables are especially helpful and necessary in cases where variables pertain to treatment unit performance, as well as water safety regarding contaminants with threshold effects.

**Qualitative variable:** Qualitative variables are variables that indicate status such as a simple yes/no answer. Qualitative variables are used to determine whether or not a certain contaminant is present, or detectable, in the water. In which case a qualitative variable will provide a present/absent answer.

It is therefore reasonable to assume that in most cases, indicator variables are qualitative and surrogate variables are quantitative. Although there are exceptions in both cases, this is generally a good assumption.

Surrogates and indicators can be developed in different ways. This report is interested in two of the methods typically used in identifying and developing these variables.

**Knowledge based:** These methods rely on expert process knowledge and experience. With this method, target- and surrogate/indicator variables are linked based on scientific knowledge of the relationship that exists between these variables. Statistical analyses performed on data (typically from a designed experiment) are then used to confirm these relationships.

**Data-driven methods:** Data-driven methods make use of various statistical methods, in order to identify any meaningful relationships between the variables from a data set. The data can be acquired using any approach (computer simulation, historical plant data, designed experiments, etc.).

In most cases a combination of methods will be applied. Knowledge based methods may use experiments to confirm expectations, and experimental methods will include expert process knowledge of the variables during the planning phase of the experimental setup and methodology.

#### **1.1.4 Treatment technologies, processes and units: Performance and testing**

Throughout literature there are many occasions where either treatment units, treatment systems, treatment technologies or treatment processes are mentioned, or referred to. It is therefore important to understand what exactly is meant by these different terms.

**Treatment system:** This refers to the combination of various treatment units, all working together towards a similar, or common purpose. A large treatment system can therefore comprise of multiple smaller treatment systems, each making use of different units, processes and technologies.

- Treatment unit:** This refers to the physical structure where the treatment occurs. May also refer to several of the same structures that are connected in parallel.
- Treatment process:** This refers to the scientific process (chemical, physical, etc.) that is responsible for the treatment within a treatment unit.
- Treatment technology:** This is the technology that is used to implement the treatment process and is housed within the treatment unit.

These terms are very closely related and are in many cases used synonymously. Using an example, the difference between these terms may become clearer:

A **treatment system**, responsible for treating river water to potable standards for a nearby town, may be said to make use of the following **treatment units**: coagulation/flocculation, sedimentation, filtration and disinfection. As an example, consider a rapid gravity sand filtration treatment unit. It consists of a concrete basin containing a network of pipes at the bottom covered with a thick layer of sand. This particular treatment unit makes use of a physical separation **treatment process**, in the form of filtration (via entrapment) taking place in the sand layer. The treatment process is achieved through the use of treatment technology. In this case the **treatment technology** is the sand used in the filter. If activated carbon (a different **treatment technology**) were used instead of sand, then the **treatment process** and **treatment unit** would be different.

In summary, a treatment system is a combination of treatment units. A treatment unit is a physical structure or device that holds a specific treatment technology. And the treatment technology is implements a specific treatment process. All four of these terms are related and if any one of them are changed or altered, the remainder will also be influenced.

The next step is to test the performance of these treatment units, processes and systems, which can be described by the following terms:

- Plant performance:** The performance of the plant is determined by the performance of each of the individual units comprising the treatment system. Various tests and monitoring programmes can be used to assess the performance of the individual treatment units, as well as the plant as a whole.
- Functionality tests:** These tests are performed to determine the functionality of a treatment unit-, or technology. A functional treatment unit will also be a well performing treatment unit and therefore functionality testing is an essential part of performance monitoring. One of many examples is the integrity testing performed on membrane technologies.

Apart from functionality tests, the performance of a treatment unit will also have to be measured by the quality of the water produced by such a unit. In many cases, a simple variable measurement can be

compared to a target level stipulated by the design manual of the unit. With regard to microbial contaminants, however, the majority of the target levels will be zero, and therefore very difficult to quantify and guarantee. It is for this reason that log removal values (LRVs) have been developed.

**Log removal value:** A performance measure for treatment units based on the amount of contaminant (typically pathogens) removed during the treatment process expressed in logarithms. For instance, a log 4 removal means that 99.99% of a contaminant has been removed. A log 7 removal means that 99.99999% of a contaminant has been removed.

## 1.2 BACKGROUND

The growing water demand in arid and semi-arid countries makes the reuse of secondary treated municipal wastewater for potable purposes all the more attractive (Lazarova et al., 2013). At this point in time, however, water reclamation for potable purposes is strictly performed as a last resort (Marais, 2012). The main driver against water reclamation for potable purposes is a general lack of confidence (from the public as well as water professionals) in the ability of reclamation plants to produce safe water (Haddad, et al., 2009). This is primarily due to the complexity and cost of plant operation, control and monitoring systems that are required to produce safe potable water from secondary treated wastewater.

The majority of secondary treated wastewater that is being reclaimed, is therefore used for non-potable purposes (industrial and agricultural use), which alleviates the pressure on conventional sources for potable water. Unfortunately, some regions still require potable water reclamation, especially during extreme drought periods (Asano et al., 2007). This led to the development of advanced water treatment processes that can be implemented in WRPs in order to produce safe potable water from secondary treated wastewater. WRPs carry an abnormally high amount of risk since the quality of the feed water to WRPs are much worse than that of a conventional water treatment plant (WTP).

There are practically only three ways in which secondary treated wastewater can be used for potable purposes. The first is called '*de facto*' reuse and involves the unintentional use of secondary treated wastewater for potable purposes (Rice et al., 2013). It should be noted that since this form of reuse is unintentional, no planning with regard to treatment, monitoring and safety is performed. This form of reuse is simply mentioned for the sake of completeness and should not be regarded or compared alongside second and third forms of potable reuse.

The second form of potable reuse is called indirect potable reuse (IPR) and involves the treatment of secondary treated effluent after being stored in a natural water body (river, aquifer or dam) for a certain period of time. The third is called direct potable reuse (DPR) and involves the treatment of secondary treated wastewater directly from the wastewater treatment plant (WWTP). Depending on the type of potable reuse (IPR or DPR), several different plant configurations may be employed in order to ensure effective and consistent treatment of the secondary treated wastewater. The different reuse types and plant configurations carry different risks and involve a variety of socio-economic, political and environmental implications.

The monitoring protocols implemented at WRPs also impact the risk associated with potable reuse. This is primarily due to the time required to detect ineffective treatment (lag time) in comparison to the time it takes for the produced water to reach the first end user (member of the public). This lag time in the monitoring system should be as little as possible, or the produced water should be stored in an engineered buffer. An engineered buffer is used for storing water and is designed to provide a retention time greater than the lag time of the monitoring system (Gerrity, et al., 2013). It should be noted, however, that the use of engineered buffers are not regulated and does not form part of any legislation, but it is considered good practice as a means of reducing the risk of the plant.

Engineered buffers are large, expensive structures that require an upgrade (increase in volume) whenever the plant achieves, or wishes to achieve, a new production capacity (increase in flow rate of final water) since the retention time in the buffer must remain greater than the lag time of the monitoring system. It will therefore be a great aid to the plant if the monitoring system could be capable of testing the water fast enough to avoid the need of an engineered buffer (Leverenz, et al., 2011). This means that the monitoring system should be able to guarantee the safety of the water relying only on the retention time of the final treatment unit and the distribution network piping, rather than an engineered buffer.

In order to reduce the lag time of the monitoring system, most plants make use of online measurement technologies. When it comes to operational variables, online sensors have been developed that can produce an accurate result at a semi-continuous rate (less than a second lag time), depending on the variable (Storey et al., 2011). Unfortunately, the same cannot be said of the water quality variables. The sensors that are designed to measure water quality variables are either too expensive to be afforded by operating companies, or they are of poor quality, requiring frequent down times for calibration and maintenance, defeating the purpose of having the technology in the first place.

Researchers are currently addressing this issue from two sides. The first approach is to improve online measuring technology, especially focusing on microbial water quality variables (Miles et al., 2011). The other approach is to improve the value of the existing online technology through computational models and statistical algorithms. Both approaches are of equal importance since a breakthrough in either one may result in a new way of monitoring WRPs that will not only allow reliable production of safe drinking water in water scarce areas, but will also make the practise of potable reuse much more affordable and feasible to communities.

In the case of the latter approach, surrogate- and indicator variables can be used to improve the value of online sensors and also reduce the amount of laboratory analyses that are required by the monitoring system. Since there are not enough sensors available to measure every variable of importance, surrogates and indicators can be used instead, which would allow a few sensors with proper data capturing- and analysis to be capable of measuring and inferring enough variable data to ensure that the treatment system is operational and the final water is safe.

The development of surrogates and indicators (seen in Equation 2) rely on statistical analyses that are used to find correlations between variables. A plant that is capable of monitoring several operational control variables at high frequency (FEM variables) with the results of that monitoring available, for instance, as a data matrix  $\mathbf{X}$ , would like to find a water quality variable ( $\mathbf{Y}$ ), which can only be measured at a low frequency (SDM variables), for which the relationship between  $\mathbf{X}$  and  $\mathbf{Y}$  can be described by any of the following equations:

$$\text{Equation 3:} \quad \mathbf{Y} = f(\mathbf{X})$$

$$\text{Equation 4:} \quad \mathbf{Y} = f(\mathbf{X}) + e$$

$$\text{Equation 5:} \quad \mathbf{Y} = f(\mathbf{X}, \mathbf{L}, \mathbf{S}) + e$$

$$\text{Equation 6:} \quad \mathbf{Y} \propto \mathbf{X}$$

The above equations represent a spectrum of situations, ranging from most ideal in Equation 3 to least ideal in Equation 6. In Equation 3,  $\mathbf{Y}$  is completely dependent on  $\mathbf{X}$  only. In Equation 4,  $\mathbf{Y}$  is dependent on  $\mathbf{X}$ , but a random error is present. In Equation 5,  $\mathbf{Y}$  is dependent on  $\mathbf{X}$  as well as  $\mathbf{L}$  and  $\mathbf{S}$ , and a random error is also present. In Equation 6 there is only a vague similarity between  $\mathbf{Y}$  and  $\mathbf{X}$ . All of the above equations can be used to form a model that functions on surrogates and indicators. With such a model, the SDM variables ( $\mathbf{Y}$ ) can be estimated based on the values of the FEM variables ( $\mathbf{X}$ ).

Further statistical analyses can then be used to express this relationship (using regression), as well as the statistical certainty to which the relationship can be expected to be accurate (using statistical significance tests).

$$\text{Equation 7:} \quad \hat{\mathbf{Y}} = \mathbf{Y}$$

$$\text{Equation 8:} \quad \hat{\mathbf{Y}} \cong \mathbf{Y}$$

The ideal situation, which was represented in Equation 3, will result in a perfect model, Equation 7, where the estimated values of the target variables are equal to the actual values of the target variables. Unfortunately, in reality there are several factors that are responsible for non-ideal situations (seen in Equation 4 to Equation 6), which result in an imperfect model, Equation 8, where the estimated values of the target variables are only approximately equal to the actual values of the target variables. The purpose of the statistical significance test is to ensure that only models with a sufficiently accurate estimation capability be regarded.

Since this approach completely depends on statistical analyses performed on measurement data, it is important that the data be collected correctly. The difference between Equation 3 and Equation 4 is a random error,  $e$ , which could be caused by inconsistent analytical procedures or measurement errors.

$$\text{Equation 4:} \quad \mathbf{Y} = f(\mathbf{X}) + e$$

The operating conditions during which the data is collected should also be varied as much as possible (within the normal operating conditions of the plant) in order to determine the relationship between the variables under different operating conditions, granted that sufficient time is allowed between changes



in operating conditions in order to ensure that the plant/experiment is running under steady state conditions.

Equation 5: 
$$Y = f(X, L, S) + e$$

In Equation 5,  $L$  contains several variables that describe the operational state of the plant, similar to  $X$ , but  $X$  contains variables describing the state of the water (primarily in terms of water quality variables). If the plant conditions are not varied over the course of the model development,  $L$ , would be a constant value and should therefore be removed from Equation 5. However, in reality variations in the operating conditions of the plant are inevitable, which means that a model built on the exclusion of  $L$  will be inaccurate, and possibly obsolete when applied in a real-world situation.

The variable  $S$  in Equation 5 refers to an unknown underlying effect that influences the value of  $Y$ , but is not measured or even identified as a relevant factor. There are thus three main sources of inaccuracy when it comes to developing models: natural variations, gross error and unexplained variations, of which only two can be accounted for. Natural variations and gross error can be detected and, to some extent, removed or corrected using data analyses.

There are mainly three data capturing approaches that can be followed in order to obtain the variable data that are required for the statistical analyses. The most basic data capturing approach is to design and build an experimental setup specifically to meet the needs of the variables that will be tested (Her et al., 2002). The advantage of this approach is that the researcher has complete control over every aspect of the experiment. The operating conditions, sampling locations, sampling frequency and analytical methods are completely adjustable by the researcher in order to gain the precise data that will be required for the statistical analyses. The disadvantage is that this approach has high costs in terms of physical equipment, laboratory space (depends on the scale of the experiment, but ought to be as large as possible) and time (to design, commission and run the experimental setup, as well as analysing the samples) (Babcock et al., 2001).

The second approach is to make use of an existing, operational plant, in order to perform full-scale experiments (Naismith et al., 2005). In this case the main focus is to operate the plant correctly whilst following a very strict and well-designed monitoring programme in order to get the right samples at the right time. This data capturing approach relies heavily on the operators of the plant, who will have to grant permission to the researcher to gain access to the plant, take samples of the plant and potentially change the operating conditions of the plant. The advantage of this approach is the size of the experiment and the low experimental costs. The disadvantage is that the experiment is limited with regard to the operating conditions that are available for testing and may require a long time to collect sufficient data (Shrestah and Kazama, 2007).

The third approach makes use of historical plant data, from an existing and operational plant. The advantage of this approach is that there is no experimental costs in terms of physical equipment, laboratory space or time. The operation of the plant, sampling of the plant and analysing of the samples have all been performed regardless of the researcher's interest in the data. The disadvantage to this

approach is apparent, the researcher has no control over the data whatsoever. The operating conditions of the plant, the sampling locations and frequencies, as well as the variables being analysed and measured cannot be manipulated or varied by the researcher at all. In terms of testing the performance of a soft-sensor such data are ideal, since it is real-world data, but as far as developing a soft-sensor is concerned, such data are the least ideal.

### **1.3 AIM AND APPROACH**

This study made use of historical plant data, from an existing WRP, instead of data generated through experimentation. It was mentioned earlier that historical plant data is ideal for testing the performance of advanced monitoring tools, but not ideal for developing these tools in the first place. Therefore, the aim of this study is to determine whether or not historical plant data from an existing WRP can be used for identifying surrogate and indicator variables following the data-driven approach.

In order to achieve this aim, several objectives that are in line with the data-driven approach were established. The importance and scope of each of the objectives of the study will be discussed in this section of the report. The study consisted of four objectives:

- Obtain, organize and assess historical plant data from an operational WRP
- Develop statistical models for the identification of surrogates and indicators
- Assess the performance of the developed statistical models
- Identify and evaluate relevant variable relationships using expert process knowledge

#### Obtain, organize and assess historical plant data from an operational WRP

In order to obtain maximum value from the statistical analyses, it is important that the data be collected at a WRP that has been operational for a long time. During this time, the plant should ideally not have undergone any changes regarding the treatment units and operational procedures. This is important since any variation in the treatment units or operational procedures will result in different plant characteristics, which means that the data from before and after the changes may not be comparable. Each of the statistical analyses require at least some minimal amount of viable data points in order to establish with statistical certainty that the observations and conclusions made from the analyses will remain consistent throughout the future of the plant (unless the plant undergoes changes), and therefore should test well on a section of the data that has been kept separate for this purpose.

#### Develop statistical models for the identification of surrogates and indicators

The development of statistical models requires the pre-processing of the data. This is of major importance since the performance of the models depend on the quality of the data that are used by the models. The models that were developed functioned in one of two ways. Either the models tested the correlation between the different variables directly, or they predicted the values of certain variables, using independent variables as inputs. In the case of the latter, the models will then evaluate the accuracy of the predictions in order to determine whether or not surrogates or indicators are present in the data.

### Assess the performance of the developed statistical models

After the model has been developed, it can be tested. The data used to test the model on should be independent, which is to say that the data should not have been used during the development of the model in the first place. The data does not have to come from a different plant, or laboratory, but simply from a time period that was not included during the development of the model.

The most appropriate time period is the most recent time period. In this way the model can be built and tested in a manner that will most closely resemble the actual real-world function that it will perform once completed.

### Identify and evaluate relevant variable relationships using expert process knowledge

Once the models have identified potential surrogates and indicators, expert process knowledge was used to scrutinise the variables in order to remove variables that are already known to be surrogates or indicators, as well as surrogates or indicators of little value.

It is important to remember that the aim of developing advanced monitoring tools is to reduce the lag time present in conventional monitoring tools. Therefore, if a correlation exists between two variables that are both equally easy to measure accurately at a high frequency, then that correlation is of little value since it will not result in a faster monitoring tool.

## **2 WATER RECLAMATION PLANT**

This chapter is devoted to understanding the treatment system which is at the centre of this research. Understanding the different treatment units, retention times, number of streams entering and exiting each treatment unit, as well as some of the basic design principals behind the treatment units is of great importance for interpreting results and literature.

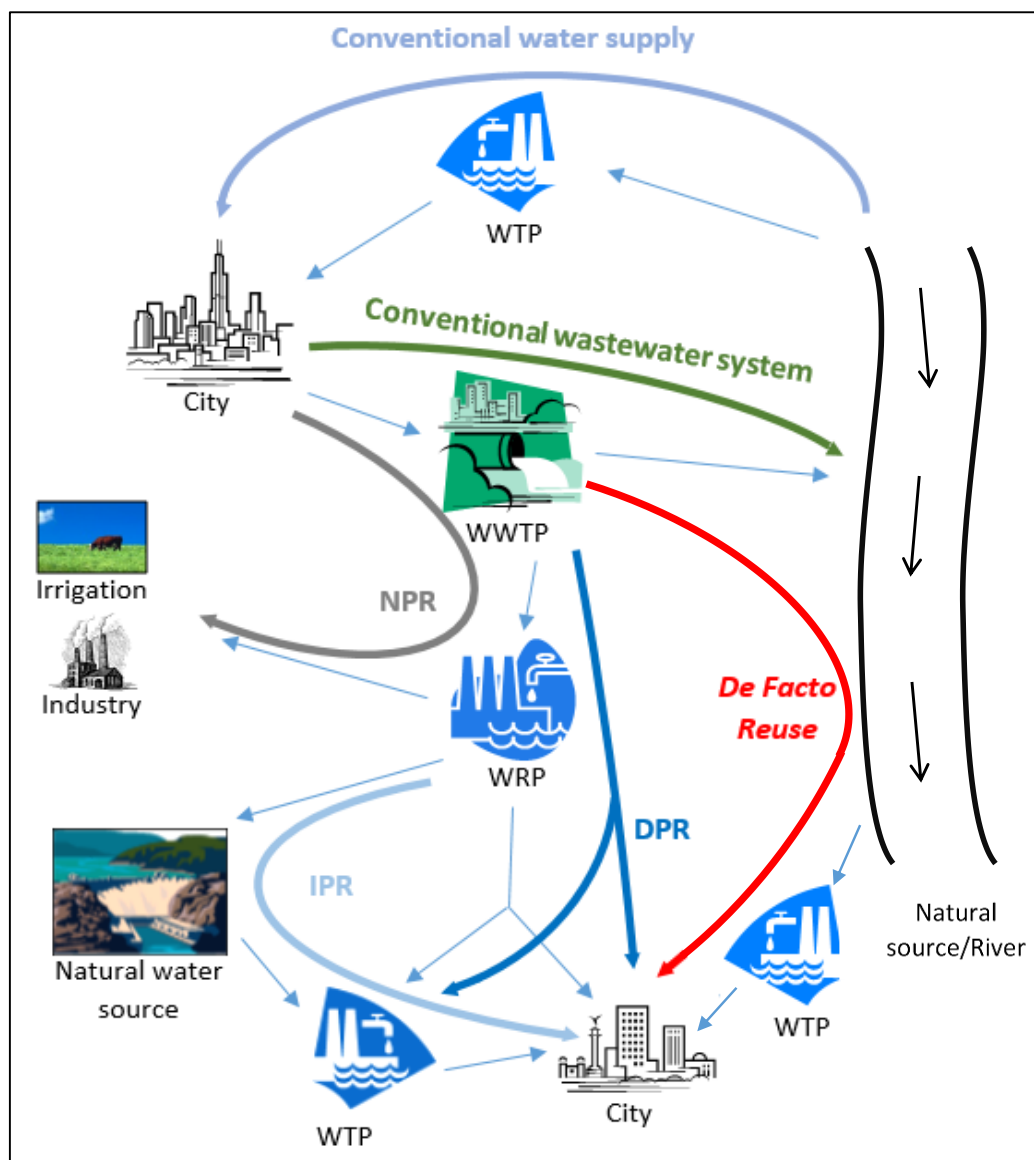
This chapter will first discuss the plant in a general overview after which each of the different treatment units will be explained individually in more detail.

### **2.1 WATER RECLAMATION OVERVIEW**

Water reclamation is the process by which wastewater is treated to a standard that is suitable for beneficial use. Various wastewater treatment processes have been used throughout history with some sources tracing back all the way to ~ 3 000 years B.C. (Levine, et al., 2014). The beneficial use of wastewater that can be considered water reuse, however, was first observed around 1890 when farmers around Mexico City made use of canals in order to irrigate agricultural areas with untreated or minimally treated wastewater (Levine, et al., 2014).

#### **2.1.1 Types of water reclamation**

In the illustration of the types of reclamation (Figure 2.1), thin arrows represent actual flow paths and thick arrows represent the general water cycle that incorporates all the adjacent flow paths and treatment processes.



**Figure 2.1:** Different types of water reclamation

Figure 2.1 identifies four types of water reclamation namely: Non-potable reuse (NPR), indirect potable reuse (IPR), direct potable reuse (DPR) and a fourth known as *de facto* reclamation or unintentional IPR.

### Non-potable reuse

In the case of non-potable reclamation, wastewater is treated to be used by industries, agricultural irrigation, landscape irrigation, recreational lakes and even to replenish wetlands or to build synthetic wetlands (National Academy of Sciences, 2012). The standards used to treat the wastewater are determined by the purpose of the reclaimed water. For instance, the water produced for industrial cooling towers will not have the same quality as the water produced for a recreational lake.

Wastewater reclaimed for non-potable use typically does not comply with drinking water standards. It is therefore required that the use of non-potable reclaimed water is clearly indicated at the point of use to warn people in the area that the water should not be used for drinking. The use of non-potable

reclaimed water also requires that a separate distribution system is used to eliminate the risk of cross-contamination between potable and non-potable water.

### Indirect potable reuse

Indirect potable reclamation is similar to non-potable reuse since the water produced during IPR is not of a drinking standard and can therefore not be distributed using a potable water distribution system. What separates IPR from NPR is the fact that with IPR the water can be discharged into a drinking water source, surface or ground, with the aim of increasing drinking water availability (Rodriguez, et al., 2009).

Because the water produced during IPR is typically destined for only one or two sources, the distribution of IPR water is much simpler to manage, however, in most cases the WWTP of a town is far away from the drinking water source of the town. This means that despite the simplicity of the discharge system, it will not necessarily be cost effective due to long piping distances and large pumping requirements (Khan, 2013).

*De facto* reclamation is very common in South Africa and many other countries, including the USA. *De facto* reclamation occurs when natural water sources, usually rivers or dams, are used for both water abstraction and wastewater discharge (in the form of return flows). *De facto* reuse is similar to IPR with one exception; *de facto* reuse does not have a WRP or advanced drinking water treatment (ADWT) process in its water cycle. *De facto* reclamation has been studied in the US and has been steadily increasing since the 1980's (Rice, et al., 2013). This form of reclamation will not be explored any further in this document since it does not make use of a WRP.

### Direct potable reuse

Direct potable reclamation is when wastewater is treated to potable standards. The water can then either be blended with potable water from a WTP or be blended with the source water for a WTP directly before entering the WTP and distributed using the existing distribution network. Since the water from a DPR plant can come from a WWTP and not a large surface water body or groundwater source, the effort of transporting water to the distribution network is much smaller (Khan, 2013). In comparison with groundwater abstraction, water from a DPR plant is easier to pump since the plant is at surface level (Leverenz, et al., 2011).

### Comparing the different types of reuse

There are several advantages and disadvantages to the different types of reclamation. It should be understood that in most cases the 'best' type of water reclamation will depend on various factors, therefore any type of reclamation cannot simply be labelled as the best, cheapest, simplest, most efficient or safest for all applications. The factors that play a role in selecting the most appropriate form of reclamation for a given application include the relative locations of WWTP, WTP and conventional water sources, the consistency of conventional water source quality, the quality of the wastewater to be reclaimed, the required quality of the treated water, the capital cost of the system, the operational cost of the system and the maintenance requirements of the system (IWA, 2014).

In most cases, the advantages of DPR over IPR and NPR are mainly from a financial and managerial point of view (Chalmers, et al., 2011). This is especially the case where conventional water sources are far away from WTP and WWTP are near WTP. The variability in water quality of conventional water sources is very detrimental to IPR systems since this would complicate the quality control operations of the final water produced. There is also the added risk that the WTP may choose to start using water from an alternative source, in this case the IPR plant would have to either rebuild their discharge system in order to discharge into the new water source or the plant should be upgraded to a DPR plant.

The following should be kept in mind concerning the different methods of water reclamation. Non-potable water reclamation carries a very small risk when the water use is managed correctly. IPR makes use of natural barriers or buffers to reduce the risk of contamination to the end user and makes public acceptance easier, but can complicate quality control procedures. DPR carries the largest amount of risk since a failure at a plant may lead to direct contamination of water consumers. There are several different process designs for a given type of reclamation, and each design has its own advantages and disadvantages in terms of risk, cost efficiency, resource management and public perception.

It is exactly for that reason that most DPR plants incorporate a multiple barrier design in order to reduce the risk of contamination to the water consumer. DPR plants make use of ADWT processes which involves a tertiary treatment process that removes trace constituents from secondary effluent (WWTP effluent) followed by the removal of dissolved constituents and finally conditioning and disinfection (Cain, 2011).

## **2.2 PROCESS OVERVIEW**

The WRP of concern for this study receives secondary treated wastewater from a WWTP that makes use of the following treatment processes:

- Mechanical screens
- Degritting
- Primary settling
- Activated sludge and trickling bio-filters in parallel
- Secondary settling
- Chlorination

The WWTP is a biological nutrient removal (BNR) plant which makes use of eight maturation ponds before discharging the treated wastewater to a river. This is also the point (the outlet of the final maturation pond) where the feed to the WRP is taken from.

Table 2.1 shows a summary of the different treatment units and some design parameters for these units.

**Table 2.1:** Process design parameters for each unit from the studied plant

Treatment process	Process Design Parameters
Upstream WWTP	BNR (long sludge retention time)
Pre Ozone and Coagulation / Dissolved Air Flotation (DAF)	Contact time: 3 min Coagulant: FeCl <sub>3</sub> , HCl, polyelectrolyte DAF SLR: 4 m <sup>3</sup> /m <sup>2</sup> /hr
Dual Media Filtration (DMF) / a.k.a. Sand Filtration (SF)	Rate: 6 m/hr Anthracite: 0.7 m (ES 1.3) Sand: 0.7 m (ES 0.7)
Ozonation	Dose: 17 mg/L mg O <sub>3</sub> /mg DOC: 1.1 Contact time: 24 minutes Contact time: 12 mg/L/min
Biological activated carbon (BAC)	EBCT: 10 minutes minimum Bed depth: 1.5 m
Granular activated carbon (GAC)	EBCT: 20 minutes minimum Number of stages: 2 Bed depth: 1.5 m
Ultrafiltration	Flux: 70 L/m <sup>2</sup> /hr Recovery: 92%
Chlorination	Contact time: 1 hour Contact time: 27 mg/L/min pH: 7.8 - 8.2 Temperature: 15 - 20 °C

## 2.3 DESCRIPTION OF TREATMENT UNITS

### 2.3.1 Dissolved Air Flotation

The DAF treatment unit in question was designed to be operated for two hours, after which a froth removing cycle shall commence. The froth is removed by increasing the water level inside the DAF tank. The froth can then flow into the froth collection channel via an overflow. Treated water is continuously drawn from the tank using a manifold (perforated collection pipe). The DAF serves as a barrier against suspended solids as well as organic matter (NOM and EfOM).

The DAF unit is completely controlled using a supervisory data and acquisition (SCADA) system that regulates the intervals between de-scumming (froth removal) and de-sludging cycles. The sludge and scum is sent to the wastewater sump which drains to a nearby WWTP.

The data that were obtained for the DAF unit can be seen in Table 2.2 where the count column indicates the number of data points for each of the variables that were included in the study.



**Table 2.2:** Data obtained for DAF

ID	Analysis	Count	ID	Analysis	Count
107	Turbidity	860	115	Calcium hardness	126
108	pH	849	116	Total Alkalinity	123
109	Temperature	851	117	Magnesium hardness	124
110	EC	250	118	Total hardness	124
111	COD	245	119	Chlorophyll A	118
112	DOC	248	120	Iron (Fe)	93
113	TDS (Calc)	245	121	Manganese (Mn)	94
114	UV <sub>254</sub>	244	122	Total Algal Count	47

### 2.3.2 Sand Filter

The sand filter in question is a dual media rapid gravity sand filter that makes use of silicate sand (750 mm deep) and Anthracite (700 mm deep) as a filter media. A 200 mm deep layer of grit is on the bottom of the filters and acts as a support layer for the filter bed. There are 5 filters in total, each with a length of 12.6 m and a width of 3.0 m resulting in 37.8 m<sup>2</sup> filtration area per filter and 189 m<sup>2</sup> in total. The filters are operated at a filtration rate of 6 m/h.

The main treatment goal of the filters is the removal of iron and manganese as well as suspended solids. The target level for assessing the performance of the filters is a turbidity of below 0.5 nephelometric turbidity unit (NTU). An online turbidity sensor is installed on the outlet of the sand filters in order to monitor the performance of the filters in real time.

Backwashing the filters consists of three stages:

- Air scour only (55 m/h) for 30 seconds
- Air scour (55 m/h) and slow water rinse (12 m/h) for 4 minutes
- High rinse with water (24 m/h) for 4 minutes

The backwash water is obtained from a reservoir that is supplemented using a bleed stream from the GAC outlet. The first filtrate (first 10 minutes of water filtered after a backwash cycle) is automatically recycled to the plant inlet.

The data that were obtained for the SF unit can be seen in Table 2.3 where the count column indicates the number of data points for each of the variables that were included in the study.

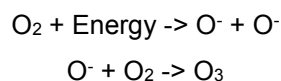
**Table 2.3:** Data obtained for sand filter

ID	Analysis	Count	ID	Analysis	Count
124	Turbidity	848	137	Total hardness	124
125	pH	851	138	Clostridium Spores	124
126	Temperature	850	139	Clostridium Viable	124
127	COD	245	140	Faecal coliform	124
128	EC	244	141	Faecal streptococci	124
129	DOC	248	142	<i>Pseudomonas aeruginosa</i>	124
130	TDS (Calc)	244	143	Total coliform	121
131	UV <sub>254</sub>	245	144	Chlorophyll A	123
132	Iron (Fe)	194	145	HPC	124
134	Total Alkalinity	125	146	<i>E Coli</i> (confirmed) Tryptone	123
135	Calcium hardness	124	147	Somatic coliphage	110
136	Magnesium hardness	124	148	Total Algal Count	51

### 2.3.3 Ozonation

The ozone used in this process is generated on site. This requires several auxiliary processes. First of all, oxygen should be produced with a very high purity (93%). This purity is achieved using several treatment steps, the main step is a pressure swing absorption (PSA) process. The PSA process starts by compressing air to 750 kPa (gauge), the compressed air then enters refrigeration dryers in order to remove any moisture from the air. Once the moisture is removed, the air is pumped through three different filters (activated carbon filter, high efficiency filter and a coalescing filter) in order to remove any dust and oil from the air. The air then enters the PSA pressure chambers which are used to absorb the other gasses (not oxygen) contained in the air, resulting in a high purity oxygen gas.

The oxygen is then pumped to the ozone generation plant where electric discharge is used to break the bonds between diatomic oxygen molecules. The resulting oxygen radicals then bind with remaining diatomic oxygen molecules to form ozone. This two-step reaction can be seen below:



The ozone is contacted with the water in a three stages. Ozone is dosed at the start of each of the stages using diffusers. The water then flows to the next stage through baffled sections for optimal contact. The ozone dosage for stage one, two and three are 9 mg/L, 4 mg/L and 4 mg/L respectively, resulting in a maximum ozone dosage of up to 17 mg/L. The off-gas from the ozonation treatment step is split into two lines, one going to the pre-ozonation treatment step, and the other to the vent ozone destruction (VOD) where the ozone is released into the atmosphere after being destroyed almost instantly. The residual ozone that remains in the water is destroyed using hydrogen peroxide, which is dosed at the outlet of the ozonation treatment step, prior to entering the BAC.

The data that were obtained for the Ozone Contact unit can be seen in Table 2.4 where the count column indicates the number of data points for each of the variables that were included in the study.

**Table 2.4:** Data obtained for Ozone Contact

ID	Analysis	Count	ID	Analysis	Count
156	pH	652	162	DOC	245
157	Temperature	653	163	UV <sub>254</sub>	242
158	Turbidity	652	164	Clostridium Viable	230
159	Residual O <sub>3</sub>	203	170	EC	124
160	HPC	250	171	TDS (Calc)	124
161	COD	244			

### 2.3.4 Biologically Activated Carbon

Seven BAC filters are installed, although the system is designed to deliver maximum output even if only five of the seven filters are operational. The filters are operated in an upflow direction with a normal filtration rate of 6.4 m/h and a maximum filtration rate of 8.90 m/h. The filter beds are 9.1 m long and 2.5 m wide resulting in a filtration area of 23 m<sup>2</sup> per filter or 159 m<sup>2</sup> in total. The empty bed contact time (EBCT) of the filters are 14.2 minutes under normal operating conditions and 10.1 minutes under maximum throughput conditions.

Backwashing consists of water only, at a linear velocity of 30 m/h for 20 minutes. After each backwash the first filtrate (first 10 minutes of normal operation after backwashing) is recycled to the inlet of the plant.

The data that were obtained for the BAC unit can be seen in Table 2.5 where the count column indicates the number of data points for each of the variables that were included in the study.

**Table 2.5:** Data obtained for BAC

ID	Analysis	Count	ID	Analysis	Count
173	DOC	180	178	TDS (Calc)	125
174	COD	127	179	Turbidity	126
175	EC	125	180	Iron (Fe)	94
176	pH	125	181	Manganese (Mn)	93
177	Temperature	127			

### 2.3.5 Granular Activated Carbon

The GAC filters in question consist of double bed or duel-stage filters which are essentially two filter beds in series. There are a total of 7 double bed GAC units operated in parallel. The first bed is operated in an upflow manner, similar to the BAC filters, and the second bed is operated in a downflow manner. All 14 beds are the same size, 9.1 m long by 2.5 m wide, the filtration area is therefore 45.5 m<sup>2</sup> per stage or 318.5 m<sup>2</sup> in total. Also similar to the BAC filters, the GAC filters are designed to produce the

maximum output even if only 5 of the 7 GAC units are operational. The filtration rate under normal operating conditions is 6.35 m/h with a maximum filtration rate of 8.90 m/h. The EBCT under normal operating conditions is 14 minutes with a minimum of 10 minutes.

The filters are backwashed consecutively (the upflow filter is backwashed once per week and the downflow filter once every two weeks), each bed is backwashed for 20 minutes with water only at a linear velocity of 30 m/h. The first filtrate is also recycled to the inlet of the plant. The two filter stages are exchangeable, meaning either one of the beds can be used as the first or the second stage, which is necessary since the activated carbon in the beds will not be exhausted at the same rate. When the first stage carbon is exhausted, new carbon will be placed in the bed and that bed will then become the second stage for that filtration unit. This means that the first stage of each of the GAC filtration units will always contain the oldest carbon.

The data that were obtained for the GAC unit can be seen in Table 2.6 where the count column indicates the number of data points for each of the variables that were included in the study.

**Table 2.6:** Data obtained for GAC

ID	Analysis	Count	ID	Analysis	Count
182	Turbidity	851	188	EC	244
183	pH	850	189	DOC	244
184	Temperature	867	190	TDS (Calc)	244
187	COD	248			

### 2.3.6 Ultrafiltration

Ultrafiltration (UF) is one of the most common membrane filtration processes in the water industry. In wastewater treatment systems UF membranes are used in membrane bioreactors (MBRs); in water treatment UF is often the membrane of choice for pre-treatment during salt water reverse osmosis (SWRO). UF provides a good middle ground between feed pressures and permeate quality. UF is a good barrier for most pathogens, organics and suspended solids, but comes at a fraction of the electricity consumption and sensitivity of reverse osmosis (RO) membranes.

The ultrafiltration membrane plant consists of six UF racks that are grouped into two skids with three racks each. Each rack has a maximum capacity of 200 m<sup>3</sup>/h and a design capacity of 85 – 187.5 m<sup>3</sup>/h.

UF membranes are susceptible to fouling (agglomeration of impurities on the membrane surface that cannot be effectively removed by backwashing). Fouled membranes can be cleaned using chemical products during a backwashing cycle, this is called chemically enhanced backwash (CEB). The dosing of anti-scalents is also prescribed to reduce the rate of fouling on the membranes pro-actively. Fouling can also be combatted using clean in place (CIP) cycles which entails dosing high concentrations of chemicals into the feed of the membrane. Depending on the membrane, the cleaning chemicals can be strong acids, bases or oxidants. During a CIP cycle, the UF permeate is wasted, since the strong chemicals are not suitable to enter the product water supply.

The data that were obtained for the UF unit can be seen in Table 2.7 where the count column indicates the number of data points for each of the variables that were included in the study.

**Table 2.7:** Data obtained for UF

ID	Analysis	Count	ID	Analysis	Count
192	pH	655	196	EC	579
193	Temperature	652	199	HPC	577
194	Turbidity	662	202	<i>Pseudomonas aeruginosa</i>	125
195	TDS (Calc)	579			

### 2.3.7 Chlorination

Chlorine is a strong oxidant and is commonly used as a disinfectant at water treatment plants. WRP A makes use of chlorine gas which is dosed into the final water where mixing allows the chlorine to be rapidly dispersed in the water. Similarly to ozonation, the concentration of the chlorine, the mixing of the chlorine in the water and the contact time with the water, all plays a role in the effectiveness of the disinfection achieved.

One of the main advantages of using chlorination for disinfection, as opposed to ultra violet (UV) light, H<sub>2</sub>O<sub>2</sub> or ozone, is the fact that the chlorine remains in the dosed water for a long period of time. This residual chlorine level is of great importance and is commonly used as a surrogate for disinfection effectiveness with regard to bacteria and viruses. The target residual chlorine level at WRP A is between 0.9 and 1.2 mg/L.

The data that were obtained for the final water (after chlorination) can be seen in Table 2.8 where the count column indicates the number of data points for each of the variables that were included in the study.

**Table 2.8:** Data obtained for final water

ID	Analysis	Count	ID	Analysis	Count
204	pH	855	221	Total hardness	242
205	Temperature	855	223	Orthophosphate	126
207	Free chlorine	853	224	TKN	126
210	EC	583	225	Nitrate	125
211	TDS (Calc)	583	226	Nitrite	125
214	HPC	577	227	Chlorophyll A	123
215	COD	247	235	Chloride (Cl)	31
216	DOC	245	236	Sulphate (SO <sub>4</sub> )	31
217	Total Alkalinity	246	237	TOC	31
219	Calcium hardness	243	238	Potassium (K)	28
220	Magnesium hardness	242	239	Sodium (Na)	28

## **3 LITERATURE REVIEW**

### **3.1 PLANT MONITORING**

When it comes to water reuse, the health of the environment and end-users (where potable reuse is considered) that receive the reclaimed water, is of great importance. It is for this reason that most risk assessments regarding water reuse, focus on the human health effects and the environmental impacts caused by water reuse systems.

Since water reclamation is so closely related to drinking water and wastewater treatment, it is understandable that most of the monitoring protocols for WRPs are similar to the protocols followed at WWTPs and WTPs. In most cases, the monitoring practices performed at WRPs have been created by adapting the practices performed at WTPs and WWTPs since the treatment technologies and water quality standards are similar (USEPA, 2012).

The plant monitoring systems in place at WRPs form part of larger management systems, such as water safety plans (WSPs) and quality control and quality assurance (QC and QA) systems used to govern the operation and performance of the plant (USEPA, 2012). The purpose of any monitoring system is to evaluate and ensure the performance of the treatment technology and management tools implemented at the plant according to the targets set by the governing systems in place (USEPA, 2012). Plant monitoring systems not only ensures the optimal performance of the plant but also forms a large part of the risk management of the plant.

The focus of plant monitoring is therefore equally weighted on risk management (which forms part of the QA of the plant) and plant performance (which forms part of the QC of the plant). The risk management aspect of the monitoring system is solely concerned with the safety of the final water being produced. This requires a thorough understanding of the incoming water quality as well as the removal capacity of each of the treatment units, how they affect one another and how they will respond should a failure occur at any point in the treatment system. The plant performance aspect of the monitoring system is responsible for ensuring that the plant is operated at optimal conditions, conducive for maximum treatment performance and efficiency of all the treatment units in the plant (Tchobanoglous et al., 2011).

It should be noted that the extent of the monitoring performed at WRPs does not depend on the treatment technology used at the plant, instead it only depends on the type of reuse being performed (TCEQ, 1997). Another factor that affects the extent of a WRP's monitoring programme is the authority that has jurisdiction over the plant. In the USA and many other countries, there may be several authorities that stipulate treatment and final water quality targets (Mancha, 2013).

#### **3.1.1 Risk management**

Researchers have identified several thousands of natural and industrial pollutants that may be encountered in the water reuse system. In many cases, health studies have been conducted in order to determine at what levels of contamination these pollutants can become hazardous to the environment

and consumers (Ivarsson, 2011). It is, however, not expected that a monitoring system be able to test for all these pollutants on a daily basis. Instead, relationships between the different contaminants and plant performance may be exploited using surrogate variables, if sufficient data generated from robust statistical experiments show that the surrogate variables are reliable (USEPA, 2012).

With the large number of contaminants found in reclaimed water there is a risk that consumers may be exposed to contaminants that can cause acute or chronic health problems. The risk management aspect of the monitoring system is therefore responsible for reducing the health risk to the community below the tolerable levels, as prescribed by legislation (WHO, 2011). There are two monitoring systems that can be used to aid in this process. The first is compliance monitoring, which is a standard legislative requirement for plants that produce potable water to a community. The second system is called raw, or feed, water monitoring and is responsible for monitoring the catchment area (sources that feed the WRP). Unfortunately, feed water monitoring is not a standard requirement and as such is not commonly practiced. There are, however, guidelines that strongly recommend the use of feed water monitoring systems (Swartz, et al., 2015).

### Compliance monitoring

Compliance monitoring includes any monitoring done by the plant in order to keep record of the final water quality leaving the plant (Leverenz, et al., 2011). Compliance monitoring usually includes monitoring the end of the treatment process on-site and continues through the entire distribution system (random locations within the distribution system) in order to verify the quality of the water delivered to the end user (Chen, et al., 2013).

Compliance monitoring systems consists of both in-house and external monitoring programmes. The external monitoring is regulated by a governing authority and may be done by an agent of the government or an independent, impartial agent. The in-house monitoring programmes are established by the operators of the plant. In some cases the owner of the plant, as well as the customers of the plant (in case the water is sold to a third party), will also have a say in the monitoring requirements. In such cases legal contracts will be drawn up, in order to establish the minimum monitoring requirements (Du Pisani, 2006).

In any event, the absolute minimum monitoring requirements for any WRP is established by the governing authority that has jurisdiction over the plant. Any other agreements concerning the monitoring requirements can only be more conservative, never less conservative, than the requirements established by the governing authority. The water quality standards used to regulate WRPs can vary depending on the type of reuse being performed by the system (USEPA, 2012).

Depending on the standard, there may be various ways in which samples should be taken during compliance monitoring. Most legislative standards or guidelines will provide the number of locations (in the distribution system) as well as the sampling frequency (often a function of the population size being served) at which samples should be taken (WHO, 1997). Compliance monitoring typically requires that samples be taken at the final treatment process step (disinfection) and at several points

throughout the distribution system. The samples are then analysed for a large variety of variables including physical, macro-chemical, micro-chemical and microbiological contaminants. The frequency of the sampling is typically measured on a monthly scale.

The quality of the monitoring system and the laboratories responsible for analysing samples and generating results, plays a major role in public acceptance. Penalties for non-compliance also ensures that WRPs keep the focus of their work on the quality of the final water produced.

### Feed water monitoring

Feed water monitoring at WTPs has been performed long before water reclamation schemes were considered as an alternative for public drinking water supply. The protocols and principles used by WTPs to monitor their raw water made it possible for WRPs to do the same with the wastewater that feeds the WWTP, which in turn, feeds the WRP.

The monitoring of treated wastewater entering the plant can form part of the operational control monitoring and will be discussed later. In the case of feed water monitoring, the aim is to monitor the entire source (or catchment area) in order to understand the different sources of wastewater that can potentially enter the WRP (Anderson, et al., 2010). This forms part of the risk management of the plant since the contaminants in the sources can serve as a baseline of the hazard that should be removed by the WRP (Swartz, et al., 2015).

From a legislative point of view the monitoring of raw wastewater streams and sources that feed a WRP will only be done as part of the feasibility study required to show that water reclamation is an option in a given area. According to the USEPA (2012) there should be a larger focus on feed water monitoring as a means to document the different contaminants and concentrations that can be expected to enter a WRP, as well as establishing effective diversion alternatives.

As is the case with compliance monitoring, feed water monitoring makes use of low frequency samples, taken at sampling points that cover a large geographical area and are analysed for a large number of water quality variables (Swartz, et al., 2015).

### **3.1.2 Plant performance**

There are primarily two motivations, or needs, behind monitoring systems used for plant performance, the first is for the evaluation of the treatment system performance, and the second is for ensuring and optimising the consistency and reliability of the treatment units (NRC, 2012). Operational control monitoring can be used to obtain the data required for satisfying both of the above mentioned needs.

With the development of online sensor technology, monitoring systems have become more capable than before (USEPA, 2012). Using online sensors in combination with surrogate variables provide WRPs with a better view of the plant's performance as well as allowing for automated control systems.



### Operational control monitoring

Where feed water monitoring is concerned with the water entering the WRP and compliance monitoring is concerned with the water leaving the WRP; operational control monitoring is concerned with the water inside the WRP, entering and exiting the various treatment units that make up the plant.

Unlike feed water monitoring and compliance monitoring, operational control monitoring requires a very high measuring frequency of variables measured throughout the plant. It is therefore important to select the correct variables to be monitored. Ideally the variables should have the following features (NRMMC, 2008):

- Measurable using online (real-time) methods/technologies
- Relate with hazardous water constituent removals
- Indicative of treatment process performance
- Respond quickly (faster than the retention time of the unit being monitored) to corrective actions

The variables typically measured during operational control monitoring are therefore different to the variables measured during feed water and compliance monitoring. This is mainly due to the high sampling frequency required for operational control monitoring (Leverenz, et al., 2011). Many of the variables measured during feed water and compliance monitoring simply take too long to measure and are therefore not suitable to be included in the operational control monitoring programme.

As was the case with feed water monitoring, it is important to monitor and understand the behaviour of the WWTP that feeds the WRP, however, with operational control monitoring, this is done in order to improve the performance of the plant by providing an early warning that will allow the plant operators to make the necessary adjustments to the plant, which will either ensure the correct treatment of the poor quality incoming water, or divert it away from the plant in order to protect the treatment units (NRMMC, 2008).

Operational control monitoring can also be viewed as an assessment tool that is used to confirm and control the performance of the different treatment units by following a schedule of observations and measurements (NRMMC, 2008). It is therefore important that the critical control points (CCPs) identified by the QA system be incorporated into the operational control monitoring system. The data obtained through operational control monitoring can also be used to improve the operational control monitoring system by developing, or continually improving early warning and alarm systems that prompt corrective actions in order to prevent the deterioration of the water quality, and in turn, protect the health of the end users (NRMMC, 2008).

As is the case with feed water monitoring, there are currently very little external standards that regulate the operational control monitoring protocols of WRPs. This, again, does not mean that no operational control monitoring takes place, it simply means that the standards and guidelines used

by the WRPs are created in-house and are not being regulated by an external authority. There are, however, some states in the USA that require a degree of operational control monitoring, or what they call facility reliability regulations or guidelines, from WRPs (USEPA, 2012).

These regulations primarily consist of alarm systems that act as warning for power failures or treatment process failures. It is also recommended that each of the different barriers (key treatment units in the plant), used to reduce pathogens, at WRPs should be evaluated individually (USEPA, 2012). The evaluation of treatment barriers entails much more than simply measuring water quality variables, but also functionality variables for the different treatment processes.

The use of surrogates and indicators are of great importance to operational control monitoring systems. Since these systems require high (semi-continuous) measurement frequencies, it is likely that a plant will make use of several surrogates and indicators in order to evaluate the different treatment units employed by the treatment system. Unfortunately, some surrogates and indicators are site-specific, and should be evaluated on a case-by-case basis (Drewes, et al., 2010).

### **3.2 THE KNOWLEDGE BASED APPROACH FOR SURROGATE AND INDICATOR DEVELOPMENT**

When it comes to the development of surrogates and indicators, one school of thought was primarily applied up to the 21<sup>st</sup> century. Although there may have been several small variations in the methods applied, from the perspective of today (recent 15-20 years), all of the previous methods applied for surrogate and indicator development can be grouped under one category, namely the knowledge based approach.

The methods follow the same basic structure or flow of information governed by hypotheses, assumptions and/or previous research. The knowledge based approach for surrogate and indicator development primarily consists of the following steps:

- 1) Selection of relevant variables
- 2) Experimental setup
- 3) Sampling and analytical procedures
- 4) Results typically obtained

Despite the many variations of techniques and methods that can be applied during each of the steps mentioned above, the basic structure and premise of these studies remain the same and can therefore be grouped together. The following subsections will illustrate, from literature, the variety of procedures that can be followed during the development of surrogates and indicators using the knowledge based approach. The results typically obtained by these studies are discussed in Section 3.2.4 and are illustrated in Table 3.1, Table 3.2 and Table 3.3.

### 3.2.1 Selection of relevant variables

The first step in developing surrogates and indicators, according to the knowledge based approach, consists of determining the relevant variables, which will be the surrogates or indicators as well as the intended target variables. The relevant variables can be selected based on no information (hypothesis), little information (assumption) or well-founded information (facts). Depending on this selection, the study may or may not have an increased focus in first confirming, or accepting, the selection that was made. In other words, if a hypothesis or assumption was used to select the relevant variables, the first step of the study will typically be either to accept, or reject the hypothesis or assumptions that were made.

In many cases, however, researchers will make use of well-founded information (facts), which they have found in literature or in previous studies. In that case, the selection of the relevant variables is simply stated and the study moves on to the next step.

In a study by the NWRI (2012), the California Department of Public Health (CDPH) requested that the National Water Research Institute (NWRI), of the US, perform a study to determine whether biodegradable dissolved organic carbon (BDOC) may be a suitable alternative to total organic carbon (TOC) as a surrogate for measuring the removal of organics in groundwater recharge using soil aquifer treatment (SAT). In this case, the selection of the relevant variables was based on the hypothesis that BDOC will perform similarly to TOC during the removal of organics via SAT.

Tu et al. (2013) performed a study that was focused only on a single treatment unit, rather than a treatment system as a whole. In the study, it was determined that boron is a suitable surrogate for n-nitrosodimethylamine (NDMA) during reverse osmosis membrane treatment. The selection of boron and NDMA as the relevant variables of the study was based on the similarity of the physical and chemical properties of the molecules. The similarity between these two molecules are well-founded, and can be considered factual.

Dickenson et al. (2009) performed a study in order to identify surrogates and indicators for the removal of trace organic compounds (TrOCs), in this case chemicals, from wastewater using oxidation treatment processes (specifically UV/H<sub>2</sub>O<sub>2</sub> and O<sub>3</sub>/H<sub>2</sub>O<sub>2</sub>). The study performed a selection of relevant variables based on a previous study by Dickenson et al. (2011).

The selection of the relevant variables in a study performed by Drewes et al. (2008) was based on an assumption that a combination of surrogates (in this case bulk measurements) and a limited list of organic contaminants, can be used to assess the removal of all of the organic contaminants of concern. The selected surrogate variables were proven for certain treatment processes, but what Drewes et al. (2008) did was to use a combination of surrogates for specific treatment processes in order to derive the performance of an entire treatment system.

Snyder et al. (2007) performed a study on the removal of EDCs and pharmaceuticals in drinking water sources. The selection of relevant variables was based on the physicochemical properties of the target pollutants that were observed in earlier studies. The physicochemical properties included the molecular

size, polarity, aromaticity, volatility, natural functional groups and acidity of the compounds (Snyder, et al., 2007).

From the above studies it is clear that several methods and motivations can be used for selecting the relevant variables for a given study, but in all cases these selections formed part of the initial steps of the research which adheres to the knowledge based approach for developing surrogates and indicators.

### **3.2.2 Experimental setup**

A large variety of experimental setups may be used during the development of indicators and surrogates when following the knowledge based approach. The experimental setup may vary in scale and complexity with scales ranging from desktop/laboratory to full-size/scale studies and complexity ranging from as simple as a single treatment unit to a complete treatment system consisting of multiple treatment units integrated in series and parallel. Depending on the scale, the experiment could also make use of various water sources in order to duplicate or simulate the raw water characteristics that can be expected in the real world.

Snyder et al. (2007) performed bench-scale tests using water from natural sources and then spiking them with finite amounts of known TrOCs and allowing different contact times in order to accurately simulate the pollution of water as it occurs in the environment. The bench-scale experiments were conducted batch-wise in jars, while a pilot-scale experiment was also done under continuous operating conditions. The bench-scale experiments were validated by investigating full-scale plants where similar processes are being used.

In contrast, the work done by Byrne et al. (2011) consisted of installing an online UV absorbance spectrophotometer at a water treatment plant. The data recorded by the spectrophotometer was then compared to the data obtained from water sample laboratory analyses in order to determine correlations between the UV absorbance of the water and the concentrations and nature of various constituents found in the water. In the case of Byrne et al. (2011), the readings from the spectrophotometer was compared to the required chlorine dosage in order to determine if the UV absorbance can be a suitable surrogate parameter for chlorine demand in natural surface waters.

The work done by Aull (2005), consisted of taking samples from natural water bodies where several peripheral land activities take place, in order to identify surrogates and indicators of pollution that can be used to simplify the detection of non-point pollution in natural water systems. The aim of the experiment was to determine if normal water quality parameters, measured over several seasons, can serve as surrogates for compounds and organisms associated with polluting activities.

In the case of Aull (2005), the experimental setup was virtually non-existent and merely consisted of developing and executing a sampling campaign at the existing sub-basins. This work is similar to the work done by Byrne et al. (2011) in the sense that the experimental setup was not a creation or design by the project itself, but simply made use of an existing treatment or natural system from where data could be obtained in a meaningful way. Aull (2005) simply takes it a step further by using a system

where there is practically no control over the 'operating conditions' of the system (i.e. pollution levels, flow rates, characteristic of pollutants, etc.).

Gerrity et al. (2012) collected unfiltered secondary treated wastewater from over a dozen different WWTPs located in the USA, Switzerland and Australia. The study was aimed at identifying surrogates for the performance of ozone treatment of secondary treated wastewater. The collected wastewater was used to perform bench-scale experiments testing various forms of oxidation under different ozone dosing protocols. The results of the bench-scale tests were then compared to pilot-scale studies (one from literature and three from the study itself) as well as full-scale study (using samples collected at an independent WWTP with operational data from the ozonation treatment applied) in order to verify their findings.

From all the studies mentioned above, it can be seen that a large variety of experimental setups and protocols can be applied in order to develop or identify surrogates and indicators. Despite this large variety, all of these studies can still be regarded as following the knowledge based approach for surrogate and indicator development.

### **3.2.3 Sampling and analytical procedures**

The sampling procedures followed by studies that make use of the knowledge based approach may also vary from one study to another. In most cases, however, the sampling procedure is much more related to the experimental setup, rather than simply the approach being followed.

Swartz et al. (2003) conducted a study wherein the colour levels of water were correlated with various forms of natural organic matter (NOM) and organic carbon (TOC and DOC). During the study, ten different locations (WTPs) were sampled over a three week sampling campaign that was conducted four times. The sampling campaign had to take place over a three week period due to the distance between the sampling locations and the availability of analytical laboratories. The sampling procedure turned out to be one of the largest expenditures, both in terms of time and money of the project. This was also the case in a similar study by McKnight et al. (2001).

The work done by Aull (2005), consisted of taking samples from natural water bodies where several peripheral land activities take place, in order to identify surrogates and indicators of pollution that can be used to simplify the detection of non-point pollution in natural water systems. Much like the work done by Swartz et al. (2003), several samples were taken throughout multiple seasons which meant that the research took much longer to complete.

On the other hand, Byrne et al. (2011) did not take any samples, instead they installed a high performance UVA sensor (S::CAN) at the WTP of interest. The sensor technology, which was capable of performing accurate analyses at a high rate, was used in order to collect the required data on-site. This allowed them to collect the experimental data required for the study without having to take conventional samples. The sensor was installed and gathered data during the relevant periods when information was required. The sampling, or data capturing portion was therefore minimal.

In several situations, especially where laboratory and pilot scale experiments were conducted, researchers will take large samples from relevant sources to perform their experiments on. The choice of source for these initial samples then also depends on the aims of the research. Snyder et al. (2007) made use of natural water which they then spiked with controlled amounts of pollutants in order to simulate water samples instead of collecting the samples from the actual sources. A similar method was also followed by Gerrity et al. (2012), although in that case several large initial samples were taken at different locations in order to add more variety to the research.

In terms of analytical methods, all of the researchers made use of standard methods, where possible. The work done by Anumol and Snyder (2013), however, strived to develop new analytical methods that have not been adopted into the standard methods. In many cases where surrogates and indicators were being developed for CECs, custom analytical methods had to be used since analytical methods for those compounds, at the concentrations they are present in, are still considered experimental and have not been adopted as a standardised analytical method (Dickenson, et al., 2011).

### **3.2.4 Results typically obtained**

Some results for a few of the above mentioned studies will be shown in this section, although only to indicate that both knowledge based data-driven approaches are successful at developing surrogates and indicators. It is, however, the aim of this research to determine if the data-driven approach is capable of revealing or identifying surrogates and indicators, especially variables that have not been anticipated, and therefore are seldom considered for research projects following the knowledge based approach.

It should be noted that the results are not always of the same nature. In the case of Drewes et al. (2008) the surrogates that have been identified are mainly qualitative indicators of treatment performance, whereas Swartz et al. (2003) provided quantitative results in the form of equations depicting the relationship between target and surrogate variables as well as the statistical significance of the correlations that were observed.

The data-driven approach is no different, in some cases only qualitative results are achieved resulting in performance and removal indicators with a Yes/No, or Absent/Present function. In other cases it is also possible to identify and quantify relationships between variables that are statistically sound and mathematically quantifiable, therefore, rejecting the null hypothesis and so indicating a significant correlation between the variables.

Some of the surrogates and indicators developed by Drewes et al. (2008) with respect to treatment performance, were arranged into three groups of variables and can be seen in Table 3.1.

**Table 3.1:** Surrogates and indicators for treatment performance via the knowledge based approach

Physical and chemical variables	Organic variables	Indicator organisms
Alkalinity	Biological Oxygen Demand	<i>E. coli</i>
Boron	Biodegradable Dissolved Organic Carbon	Faecal coliforms
Conductivity	Chemical Oxygen Demand	Cytopathogenic viruses
Hardness	Colour	Protozoan parasites
Nitrogen	Fluorescence Spectrometry	Total coliforms
Phosphorus	Total Organic Carbon	Heterotrophic Plate Count
Total Dissolved Solids	Molecular Weight	Somatic Coliphages
Turbidity	Total Organic Halogen	
	UV Spectrometry	

For more detail regarding the different variables, see Drewes et al. (2008). In all of the cases listed in Table 3.1, the surrogates and indicators are only qualitative, and cannot be used for quantitative prediction.

Table 3.2 shows different treatment process units typically used in WRPs. The table also indicates the hazardous variables that are targeted for removal or deactivation by each of the given treatment processes. Furthermore, Table 3.2 shows several surrogate and indicator variables that have been associated with each of the treatment processes as well as functionality tests and variables that can be used additionally to the surrogate and indicator variables in order to derive information about the performance of the different treatment processes.

**Table 3.2:** Variables suggested for treatment process monitoring

Treatment process	Hazard/Target variable	Functionality test/variable	Surrogate variable	Indicator variable
<b>Membrane filtration</b>	<ul style="list-style-type: none"> <li>• Enteric bacteria</li> <li>• Viruses</li> <li>• Protozoa</li> <li>• Helminths</li> </ul>	<ul style="list-style-type: none"> <li>• Transmembrane pressure</li> <li>• Pressure-based tests</li> </ul>	<ul style="list-style-type: none"> <li>• Total organic carbon</li> <li>• Turbidity</li> <li>• Particle counts</li> </ul>	
<b>Reverse osmosis</b>	<ul style="list-style-type: none"> <li>• Chemical hazards</li> <li>• Enteric bacteria</li> <li>• Viruses</li> <li>• Protozoa</li> <li>• Helminths</li> </ul>	<ul style="list-style-type: none"> <li>• Transmembrane pressure</li> <li>• Permeate and brine flow rate</li> <li>• Conductivity in permeate and brine</li> </ul>	<ul style="list-style-type: none"> <li>• Conductivity</li> <li>• Total organic carbon</li> </ul>	<ul style="list-style-type: none"> <li>• Boron</li> <li>• NDMA</li> <li>• Chloroform</li> </ul>

**Table 3.2:** Variables suggested for treatment process monitoring (Continued)

Treatment process	Hazard/Target variable	Functionality test/variable	Surrogate variable	Indicator variable
<b>Advanced oxidation</b>	<ul style="list-style-type: none"> <li>Organic chemicals</li> <li>Enteric bacteria</li> <li>Viruses</li> <li>Protozoa</li> <li>Helminths</li> </ul>	<ul style="list-style-type: none"> <li>Ultraviolet light dose and transmissivity</li> <li>Hydrogen peroxide dose</li> <li>Oxidation reduction potential</li> </ul>		<ul style="list-style-type: none"> <li>DEET</li> <li>Caffeine</li> <li>Meprobamate</li> </ul>
<b>Powdered activated carbon</b>	<ul style="list-style-type: none"> <li>Organic chemicals</li> </ul>	<ul style="list-style-type: none"> <li>Dose rate</li> <li>Contact time</li> </ul>	<ul style="list-style-type: none"> <li>Total organic carbon</li> </ul>	<ul style="list-style-type: none"> <li>Estrone</li> <li>Caffeine</li> <li>DEET</li> </ul>
<b>Soil aquifer treatment</b>	<ul style="list-style-type: none"> <li>Organic chemicals</li> </ul>		<ul style="list-style-type: none"> <li>Total organic carbon</li> </ul>	<ul style="list-style-type: none"> <li>Meprobamate</li> </ul>
<b>Chlorination</b>	<ul style="list-style-type: none"> <li>Enteric bacteria</li> <li>Viruses</li> </ul>	<ul style="list-style-type: none"> <li>Dose rate</li> <li>Contact time</li> <li>Temperature</li> <li>pH</li> <li>Residual Cl<sub>2</sub></li> </ul>		<ul style="list-style-type: none"> <li>Hetrotrophic plate count</li> <li>Bacteriophage</li> </ul>

[Adapted from the NRMCC (2008)]

The information shown in Table 3.2 was obtained from various published sources, but was organised and tabulated by the NRMCC (2008). The data were obtained through experiments that were conducted on different scales; from laboratory bench and pilot-plant scale experiments like the work done by Snyder et al. (2007) and Snyder et al. (2003), to full-scale experiments like the work done by Byrne et al. (2011).

In terms of treatment performance for the removal of various contaminants, the Australian Guidelines for Water Recycling (AGWR) proposed a method for validating treatment units based on their ability to remove certain key variables for which they are designed. In many cases, including Naismith et al. (2005), surrogates and/or indicators along with integrity testing techniques are used to validate the removal capacity of certain compounds based on the design purpose of the treatment unit (NRMCC, 2006). Table 3.3 shows several treatment processes that can be applied for removing certain contaminants from the water. The table also indicates several surrogates and/or indicators that can be used to validate the performance of the given treatment units.



**Table 3.3:** Surrogates used for treatment process performance validation

Treatment Process	Treatment Unit	Treatment Purpose	Performance Validation Variable/Technique
Adsorption	BAC, GAC, multi-media filters	Pathogens, organic compounds and inorganic compounds	Multi-media filters: various micro-organisms, turbidity, suspended solids, BOD, pH and chlorine
			GAC and BAC: nutrients and organic compounds
			Ozone and BAC: DOC, BOD, suspended solids, nitrogen, phosphorus and micro-organisms
Biological	Activated sludge, trickling filters, oxidation/aeration basins, anaerobic systems, BAC	Solid particles, nutrients, biodegradable organic matter, pathogens	Computer modelling, automated image analysis, pathogen level prediction and trending tools
Chemical and photochemical oxidation	UV, UV/H <sub>2</sub> O <sub>2</sub> , ozone, ozone/H <sub>2</sub> O <sub>2</sub> , free chlorine, chloramine and chlorine dioxide	Bacteria, viruses, parasites, taste, odour, colour, organic chemicals	For UV: computational fluid dynamics, dyed microspheres, iodate (actinometer), online spectrophotometry (transmissivity for dose control)
			For ozone: contact time (online)
			Chlorine: no validation required, as long as contact time/turbidity, pH and temperature are correctly accounted for
			Generally (for disinfection): log removal of pathogens
Membranes	Microfiltration (MF), ultrafiltration (UF), nanofiltration (NF) and reverse osmosis (RO)	Bacteria, viruses, protozoa, small particles, chemicals, dissolved solids	Currently used: vacuum hold test, pulse integrity test, MS2 challenge test, pressure decay integrity test, Rhodamine WT tracer dye
	Spiral-wound (RO and NF), shell and tube/hollow fibre (MF and UF), plate and frame (RO and UF)		

[Adapted from the AWRCE (2013)]

### 3.3 THE DATA-DRIVEN APPROACH FOR SURROGATE AND INDICATOR DEVELOPMENT

The only major alternative to the knowledge based approach for developing surrogates and indicators, is the data-driven approach. The knowledge based approach is much more traditional, following strict scientific protocols for each step of the scientific process in order to plan, build and execute an experiment for obtaining data which is then directly used to prove or disprove the original assumption.

The data-driven approach, on the other hand, makes use of data, actual plant data or from computer simulations, in combination with statistical analyses in order to detect, define and quantify relationships between the variables contained in the data. The data-driven approach can also be illustrated in the same stepwise fashion as the knowledge based approach:

- 1) Obtain data (real-world, experimental or artificial)
- 2) Perform statistical analyses to clean and prepare the data
- 3) Perform statistical analyses/build statistical models
- 4) Test statistical models
- 5) Evaluate results

This approach is closely related to the approach followed for the development of soft sensors, since soft sensors are simply an extension, or application, of the identification process. Haimi et al. (2013), Kadlec et al. (2008) and Lin et al. (2007) proposed methods for developing soft sensors that consisted of the same objectives as is listed here. One major difference is that soft sensors operate continually and therefore make use of dynamic statistical techniques and emphasise model maintenance, which is not of importance to this study (Jing et al., 2012).

If the data comes from an existing, operational plant, the data can be collected through on-site experimentation or conventional monitoring practices (historical data) that was collected and stored in some form of a database.

The main difference between collecting data via experiments and using historical data, is the level of control. Obtaining data via experiments allows the researcher to control where and when the samples are taken, what variables are analysed for and most importantly, the data can be collected more than once in case an error occurred at some point during the experiment.

With the knowledge based approach, the emphasis of the research is placed on the scientific procedure for obtaining the data that will be used to prove or disprove the hypothesis. In contrast, the data-driven approach does not place an emphasis on obtaining the data, but rather on analysing the data in a more complete way.

The focus of the data-driven approach is therefore on the statistical analysis of plant data, irrespective of the origin of the data. This section will therefore discuss the various statistical analyses that can be used during the different stages of developing surrogates and indicators following the data-driven approach.

### 3.3.1 Data validation techniques

Data validation (not to be confused with model validation) is a process that is used to determine the suitability of data for analysis. Depending on the desired analyses that are to be performed, different data validation techniques should be applied to determine the suitability of the given data for each of the aspects that will be analysed.

For the analyses that were performed in this project, the following data validation techniques were identified as potentially useful (Swartz, et al., 2016):

- Missing data characterisation
- Dirty data characterisation
- Data reconciliation
- Testing for sufficient representation

These techniques will now be discussed in more detail.

#### Missing data characterisation

The first technique is very straight forward. For each of the variables contained in the data, a count is made of the number of missing values. This number can then be represented as a ratio or percentage of the total number of data records that are contained in the data. Due to the low sampling rate of the data, the data did not show elements of a time series. The data were therefore considered to be stationary (steady-state), which allowed for the simple missing data characterisation.

The missing data characterisation is a useful determinant of the quantity of the data and can be used to determine the need for interpolation. On the other hand, it can also be used to disqualify certain analytical and pre-processing techniques, especially interpolation, which should not be applied on data containing more than 20% missing values.

Once the characterisation is complete, the information can be displayed in a simple plot, or histogram in order to view the percentage of missing values for each of the variables contained in the data, at a glance.

#### Dirty data characterisation

Depending on the data acquisition process, there are several errors, or mistakes, that can lead to invalid, data elements (specific record of a specific variable). These errors, or mistakes, can have several effects on the data that should be identified and removed during the pre-processing of the data. There is, therefore, more than one way to quantify how much dirty data is present.

The first characterisation counts the number of impossible measurements, or sensor readings (e.g. a pH of 15, or a turbidity of -3) for each of the variables contained in the data. These counts can then be expressed directly, or as a percentage of the total number of data records, and can be effectively displayed in a plot, or graph showing the number of dirty data elements per variable.

This characterisation can also be used to identify data elements that contain some other error due to an error, or mistake, in the data capturing process (e.g. data containing characters or symbols, in-between or instead of, numeric values).

The second characterisation counts the number of unique values that is contained in the variables. Since it is highly unlikely for any variable to maintain an absolutely constant value, variables with a low number of unique values are suspicious and should be excluded from the analyses. Again, the counts can be expressed as the actual count or the percentage of unique values, for each variable. A simple plot or graph can then be used to display the information for all the variables.

### Data reconciliation

This technique makes use of a mass balance over each of the treatment units that can be used to determine whether any of the measurements are impossible due to unwarranted accumulation or disappearance of stream flows, or constituents. Data reconciliation should not be confused with dirty data characterisation. In the case of data reconciliation, the invalid data points may be completely realistic and valid when viewed in isolation.

Data reconciliation attempts to reconcile all the measurements contained in the data, whilst making use of the actual plant design (process flow diagram) in order to determine whether the data makes sense. Data reconciliation is very effective at identifying where measurement errors occur, which can aid operating companies in identifying faulty sensors or analytical equipment.

Data reconciliation requires accurate flow measurements for all streams in the plant as well as the variance of each of the measuring devices (sensors or analytical equipment). The retention time of the processes should also be taken into consideration, especially where online sensors are applied.

### Testing for sufficient representation

All the previously described techniques are used to determine whether the data is an accurate, or realistic representation of the plant, but they do not determine whether or not there is a sufficient representation of the behaviour and character of the plant.

Testing for sufficient representation comprises of several techniques, as illustrated by Swartz, et al., (2016); Napier-Munn, (2014), which determine whether there are small variations and different operating conditions contained within the data. In particular whether or not these variations are a sufficient representation of the actual plant. Since it is highly unlikely that any plant will be operating under the exact same conditions for every hour of the day and every day of the week, the data obtained from a plant should show signs of changes in the operating conditions of the plant.

A model built on data that is not a sufficient representation of the plant, may perform very well during the periods when the plant is being operated in the same way that was captured by the data, but when a period arises during which the plant operation is varied, the model will no longer function correctly and becomes obsolete.

### 3.3.2 Data preparation and pre-processing

Although the previous section (data validation) can be considered part of the pre-processing of the data, it is only done in order to assess the data, but not in order to manipulate and change the data in order to improve the quality of the data, which is discussed in this section.

Before the plant data can be used, the data must be pre-treated in order to remove unwanted or unusable portions of the data. This is done in order to improve the function of methods that will be applied (Rosén, 1998). Fortunately computerised methods can be used to pre-treat the data, however, the quality of the data is of great importance when statistical analyses are considered (Lin, et al., 2007). It is therefore possible that the data obtained for analyses will still be unusable if the data is irreversibly corrupted.

Depending on the source of the data, i.e. laboratory analysis or online sensor, there may be several different disturbances affecting the validity of the data. Online sensor measurements are susceptible to a plethora of different disturbances and faults, ranging from electromagnetic interference, infrequent calibration, incorrect calibration, lack of maintenance, exhaustion of consumables and even incorrect installation, all of which can corrupt data beyond any point of remediation.

Measurements from laboratory analyses, on the other hand, are more reliable although not entirely infallible. Disturbances and faults can include any number of analytical mistakes, incorrect sampling protocols and human error when capturing the results onto an electronic database or archive.

No form of measurement or data capturing is flawless and therefore any set of data should be screened in order to identify corrupt measurements and employ appropriate corrective measures. The most common forms of corrupt data are noise, missing values, outliers and measurement drift. These flaws can be identified and remedied using logical algorithms. These algorithms test the reasonableness of the measurement and then removes the effect of the corrupt data, or the corrupted data itself.

#### Outliers

Outliers are data points with values that are extreme, or radical, in comparison to the values of the data points in their near vicinity. Unfortunately when it comes to outliers, it is not only a matter of replacing, substituting or adjusting their values, but it is also an issue to detect outliers. Outliers can be caused by many factors and the problem is determining whether or not the extreme value is a correct measure of reality, or is caused by a disturbance or error in the measuring and/or data capturing process.

Before outliers can be removed or replaced, they must first be detected. In most cases the use of algorithms based on the statistical characteristics of the data are suggested. These methods can be considered reactive, but there are also pro-active methods. Redundant measurements using independent sensors and/or samples may be used to verify whether or not an outlier is in fact faulty or not.

A more primitive method for outlier detection makes use of time series and a simple visual inspection. Unfortunately this manual method can only be performed off-line and becomes very strenuous and time consuming when the amount of data is very large. In the case of large data sets, an automatic online algorithm is a suitable substitute for manual inspection since it is much more effective and objective.

Aggarwal (2013) has done extensive work in the field of outlier analysis, which include both the detection and replacement or substitution of outliers in data. The criteria identified as most important in deciding what method to use for outlier detection includes the type of data, the size of the data, the availability of outlier examples and also the level of interpretability required.

The most common methods for identifying outliers are based on one dimensional extreme value analysis. These methods target specific outliers: only outliers with values that are either too big or too small. These methods are therefore only effective in certain applications. Fortunately, most applications require outlier detection of this nature (Aggarwal, 2013).

With extreme value analysis the data is analysed using statistical tests. For a normal distributed data set, the z-test can be used to identify extreme values. Depending on the distribution of the data, the statistical tails may differ, therefore the scores used to discern whether values are outliers or not may differ. This is the biggest concern regarding extreme value analysis; if the data is not normally distributed, but the statistical test assumes the data is normally distributed, some values may incorrectly be identified as outliers, or vice versa.

Fortunately, most statistical tests are able to provide a reasonable explanation regarding the outlier scores of data. Many statistical tests, like the z-test, are easy to analyse since they can easily be interpreted in terms of probabilities of significance.

Lin et al. (2007) proposed the use of the three-sigma-rule (Equation 9) since it is a popular univariate method for identifying outliers.

$$\text{Equation 9:} \quad |x_i - \bar{x}| > 3 \cdot \sigma$$

Where  $x_i$  is the data point in question,  $\bar{x}$  is the mean of the variable and  $\sigma$  is the standard deviation of the variable. Unfortunately, both the mean and the standard deviation of a variable is sensitive to outliers (outliers have a large effect on the mean and standard deviation). This results in a large number of outliers going undetected since the presence of the outliers affect the mean and standard deviation.

In order to avoid this, an outlier-resistant method should be used that does not rely on the sample mean and standard deviation. One such method is the Hampel identifier, also known as the Hampel filter (Lin, et al., 2007). This method (Equation 10) makes use of the sample median and median absolute deviation (MAD) in order to identify outliers.

$$\text{Equation 10:} \quad |x_i - x^*| > h \cdot x_{MAD}$$

Again,  $x_i$  is the data point being evaluated, but in this case  $x^*$  is the median of the variable,  $x_{MAD}$  is the MAD of the variable and  $h$  is a multiplication factor that can be any constant value (typically 3).

Since the Hampel identifier makes use of estimates that are robust with regard to outliers, it can be considered an effective method for detecting outliers (Kadlec, et al., 2008).

### Noise

The most common problem found in data is known as noise. Noise is prevalent in almost all types of data measurement and storage systems. Noise can typically be categorised as either measurement noise or process noise (Rosén, 1998).

Process noise is caused by any form of variance within the process being measured. The variance can depend on ambient conditions, operational protocols, maintenance plans and random variation in treatment performance.

Measurement noise, on the other hand, occurs during the sampling and data acquisition process and includes variations due to sampling location, sampling method, sampling technology and electromagnetic disturbances (Xiong, et al., 2006). In most cases, measurement noise has a much higher frequency than process noise and can therefore be remedied using different methods (Rosén, 1998).

Generally, filters are used to reduce or remove noise from measurement data. These filters can range from simple analogue filters as described by Georgakopoulos and Yang (2001) and simple digital filters as described by Rosén (1998) to complex novel digital filters as described by Tani et al. (2015) and Rajput and Rajput (2006).

Noise reduction will largely make use of preceding and/or succeeding data values in order to reconstruct or estimate new values for data points in the same series. When it comes to noise reduction, good results can be achieved using simple methods that are easy to apply. Better results may be achieved using more complex methods, but for the purposes of this study, the improvement in data quality using simple noise reduction methods is sufficient.

For noise reduction of WWTP data, it is suggested to make use of a moving average adaptive exponential (MAX) filter (Berg, 1996). Unfortunately, the MAX filter does not preserve the effects of discontinuities in the data very well. For data containing discontinuities, it is recommended to make use of median filters (Piovoso, et al., 1992). Since noise caused by process variations are of value to this project, low-pass filters are most appropriate. This is because low-pass filters allow low frequency noise to pass (process variations), but high frequency noise is removed (measurement drift), and noise caused by process variations are typically of a low frequency, whilst noise caused by measurement and instrument variations result in high frequency noise.

Since the algorithm being developed will be used in real time, any data manipulations will only have current and historical data available. It is for this reason that only linear causal digital filters will be considered for this research. The general form of such a filter can be seen below.

Equation 11: 
$$\hat{y}_{(k)} = -a_1\hat{y}_{(k-1)} \dots -a_n\hat{y}_{(k-n)} + b_0y_{(k)} + \dots + b_my_{(k-m)}$$

In Equation 11,  $\hat{y}$  is the filtered signal and  $y$  is the measured signal. Depending on the choice of the filter parameters and order ( $a_1 \dots a_n, b_0 \dots b_m, n$  and  $m$ ), the filter can have different forms, each with different features. If at least one of the  $a$  coefficients are non-zero and only the  $b_0$  coefficient is non-zero, an autoregressive (AR) filter is formed. Autoregressive filters have infinite impulse responses (IIR), which means that all of the data points preceding the current data point being evaluated are used for determining the new value of the data point being evaluated. It can also be referred to as having an infinite “memory”. On the other hand, if all of the  $a$  coefficients are zero, the filter becomes a moving average (MA) filter. A MA filter has finite impulse responses (FIR) and is sometimes referred to as a FIR filter. MA filters are typically used as low-pass filters.

MA filters can be expressed by Equation 12:

Equation 12: 
$$\hat{y}_{(k)} = b_0y_{(k)} + \dots + b_my_{(k-m)}$$

The variable  $m$  determines the number of previous data points used to determine the new value of the data point being evaluated, or in other words the “memory” of the filter. The coefficient  $b$  determines the weighting of each of the previous data points, therefore, if the weight is to be distributed equally over all the data points, then all the values of  $b$  must be the same, and the sum of all the  $b$  coefficients must be equal to one. An equally weighted MA filter can therefore be written as seen in Equation 13.

Equation 13: 
$$\hat{y}_{(k)} = \frac{1}{m+1} [y_{(k)} + \dots + y_{(k-m)}]$$

It may not seem apparent, but the size of  $m$  is also responsible for the delay of the filter. Since  $m$  number of data points are required before the current data point can be evaluated. The value of  $m$  should therefore not be considered hastily.

An alternative filter that is commonly used as a low-pass filter, is the exponential filter. A first order exponential filter can be expressed in the equation seen in Equation 14:

Equation 14: 
$$\hat{y}_{(k)} = \alpha\hat{y}_{(k-1)} + (1 - \alpha)y_{(k)}$$

In Equation 14,  $\alpha$  can have any value between 0 – 1. As the value of  $\alpha$  increases, the filter becomes less sensitive, and unfortunately, the time delay also increases.

These filters are not ideal for preserving discontinuities, like step changes, which are common during standard operation of water and wastewater treatment plants and also water reclamation plants making use of filter and membrane processes that undergo periodical backwashing and chemical dosing cycles.

Using median filters is more appropriate in order to conserve the effects of discontinuities in the data. Median filters (see Equation 15) are completely separate from moving average filters and can be



described as a moving window, wherein the median of the data in that window is used to determine the output of the filter.

Equation 15: 
$$\hat{y}_{(k)} = \text{median}[y_{(k)}, y_{(k-1)}, \dots, y_{(k-l)}]$$

The variable  $l$  is known as the filter length, and can be used to indicate the “memory” of the filter. Again, it should be noted that a larger value of  $l$  will lead to a larger delay in the filter. Median filters can also be used as a fast and coarse method for removing outliers if the duration of the outliers are less than half of the filter length.

### Missing values

Missing values are typically caused by errors and malfunctions in the measuring and data capturing processes (Kadlec, et al., 2008). Alternatively, many missing values are produced when the data is cleaned by testing the reasonability of the data. In this case it is likely that values are removed without being replaced or substituted. There is unfortunately no way of avoiding this problem. If a data set contains, for instance, outliers or nonsensical data (pH levels higher than 14), then there is no reasonable way to replace that value with another value, but rather remove the value entirely, thus resulting in a missing value.

Missing data can result in major problems when it comes to multivariate and dynamic data analyses. Noise reduction, for instance, will be much less accurate or beneficial if the data contains frequent or prolonged missing values. It is therefore critical to reduce the number of missing values by any means necessary. In cases where static data analyses will be performed, a small amount of missing data (less than 20%) will not necessarily be problematic, but more than that may. The amount of artificial data can be compared with the threshold used in the hypothesis tests ( $p = 0.05$ ) in which case one in twenty values may be erroneous. If more than 20% of the data is artificial, then more than four out of twenty values may be erroneous, which is considerably more.

The first option for removing missing values is to simply exclude the questionnaire, containing the missing value, from the study (Pigott, 2001; Soley-Bori, 2013). This is, however, not a viable solution in cases where the missing values are distributed over a large portion of the data. If the missing data is considerable, it is often impossible to reconstruct the data in a meaningful way (Rosén, 1998).

### Replacing missing values and outliers

Once the data have been processed and all the outliers have been removed, the data will contain missing values. Missing values are replaced whether they are caused by faulty data acquisition or data cleaning (removing outliers, etc.). The most basic form of replacing missing values is interpolation. Interpolation makes use of data preceding and succeeding the missing value; it is therefore only applicable to off-line applications. For online applications extrapolation is used, although only in cases where the missing values are not consecutively prolonged.

Linear interpolation is the simplest form of interpolation and consists of replacing a missing value with the average of the closest preceding and succeeding values. Linear interpolation can be performed using Equation 16:

Equation 16: 
$$\hat{y}_{(k)} = \frac{1}{2} [y_{(k-1)} + y_{(k+1)}]$$

In Equation 16,  $\hat{y}$  is the replaced value at time  $k$ . Linear interpolation is an acceptable method for replacing missing data in the case where the missing data is not consecutively prolonged. In cases where missing data is consecutive and prolonged, other methods should be explored. Spline techniques are more sophisticated than linear interpolation, using cubic equations to provide smooth curves that result in more continuous transitions between known and predicted data values (De Boor, 1978).

### 3.3.3 Exploratory data analysis

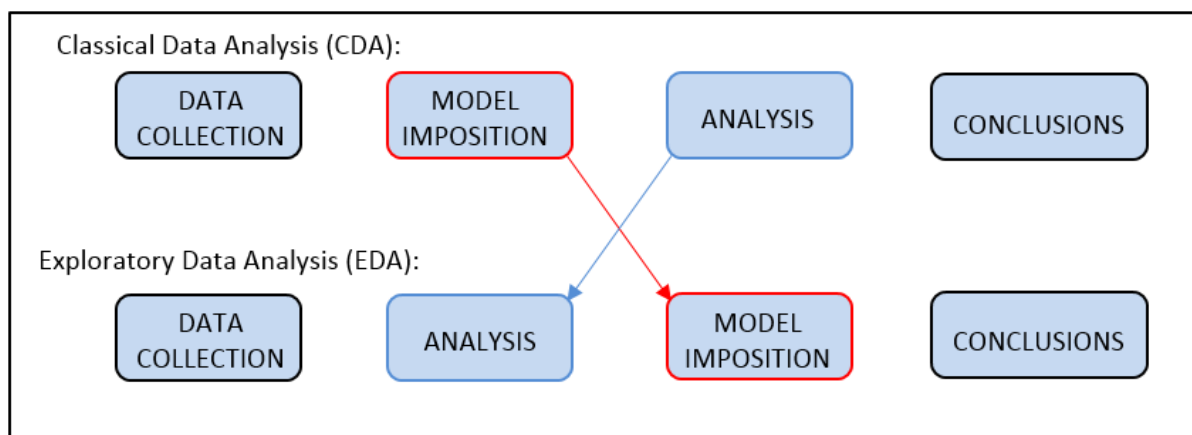
The concept of exploratory data analysis (EDA) was defined by John W Tukey in the late 1970's (Brillinger, 2011). There are numerous data analysis techniques that can be employed during EDA, most of which existed long before Tukey's development of EDA. The goal of EDA can be summarised in the following points (Tukey, 1977):

- Propose hypotheses for the causes of phenomena observed
- Evaluate assumptions on which statistical inference can be based
- Provide support for the selection of suitable statistical tools and techniques
- Emphasise further data gathering using surveys or experiments

EDA can also be used as an initial analyses of the data in order to aid with the following (Seltman, 2015):

- Mistake detection
- Assumption checking
- Identification of appropriate models
- Identifying relationships between variables
- Quantifying the relationships between explanatory and outcome variables

Achieving these goals requires not only the use of data analyses, but also the correct application of these analyses with regards to the greater philosophy being followed. The typical way these analyses were approached was shaped by classical data analysis. EDA set out to develop a new approach guided by the abovementioned objectives, by not only introducing new analyses, but also a new philosophy regarding data analysis. Despite being different from classical data analysis, the different analyses used are not mutually exclusive (Ge, 2011). The difference in approaches applied by each philosophy is illustrated in Figure 3.1.



**Figure 3.1:** Illustration of classical and exploratory data analysis philosophies

Since EDA prefers to perform analyses on the data, rather than the model, the analyses largely consist of graphical techniques that can allow the user to ‘see’ the data more clearly and therefore determine what model would naturally fit the data better. Unfortunately, like any technique that depends on human interpretation, there are some limitations with regard to the number of analyses that can be performed at a given time, as well as implications regarding subjectivity and bias. There are, however, some quantitative techniques that can also be applied when performing EDA in order to reduce the effects and limitations imposed by human interpretation.

Graphical, or data visualisation, techniques can be implemented during EDA in order to discover ‘hidden’ relations, trends and biases in the data (Ge, 2011). There are several key questions that can be asked about any data set, the first step of any analysis is to identify which questions are relevant for the study at hand, and which questions need not be asked or answered (NIST-Sematech, 2012).

Visual and graphical EDA techniques may become less desirable when large volumes of data (with regard to the number of variables, not the number of records) require analysis, thus, becoming too user intensive and timely. It is proposed to make use of computational EDA techniques in cases where graphical EDA techniques become impractical. These techniques range in simplicity from basic univariate methods to advanced multivariate techniques and can be described as quantitative, rather than graphical EDA techniques.

Since the non-graphical EDA techniques are closely related to classical data analysis techniques, and can also be sorted into two major groups namely univariate and multivariate techniques, it is easier to discuss these techniques on their own in the following sections.

### 3.3.4 Univariate techniques

This section as well as the following section will discuss the different statistical techniques applicable to this work with regard to the complexity of the techniques, namely univariate and multivariate.

Since the aim of this study is to identify relationships between variables, the very most basic statistical analysis that can contribute to achieving the aim of this project will have to be at least bivariate. With

this being said, there are some univariate techniques that can aid in the data analysis process via pre-processing of the data. The importance of data preparation was discussed earlier and should not be underestimated. Any future models will inherently be flawed if the initial data contains abnormalities (Lin, et al., 2007). The statistical methods recommended for pre-processing the data primarily consisted of univariate techniques.

### 3.3.5 Multivariate techniques

Multivariate techniques typically refer to techniques where three or more variables are being analysed together. The following multivariate techniques can be used for the identification and quantification of relationships between variables.

#### Bivariate techniques

There is a multitude of bivariate statistical data analysis methods commonly used in classical and exploratory data analysis. This thesis, however, is only concerned with the detection and quantification of relationships between variables, and therefore will only make use of various forms of correlation and regression techniques. Two different bivariate correlation methods will be used and compared in this study, the first is Pearson's correlation coefficient and the second is Spearman's rank coefficient.

Both of these methods will be used since each of them have different strengths and weaknesses regarding the type of distribution, quality and quantity of the data.

#### *Pearson's Correlation Coefficient*

This correlation method is a bivariate analysis that measures the strength of similarity between two variables. The coefficient (denoted as  $r$ ) can only have a value between -1 and 1 where 1 indicates a strong positive correlation, -1 indicates a strong negative correlation and 0 indicates no correlation.

The correlation can be calculated with Equation 17:

$$\text{Equation 17: } r = \frac{N \sum xy - \sum x \sum y}{\sqrt{[N(\sum x)^2 - \sum(x^2)][N(\sum y)^2 - \sum(y^2)]}}$$

Where:

$r$	=	Pearson's correlation coefficient
$N$	=	number of values in the data set
$x$	=	first variable
$y$	=	second variable

Pearson's correlation assumes that both variables are normally distributed as well as having a linear relationship and that both variables are homoscedastic (have the same variance).

#### *Spearman's Rank Correlation*

The Spearman rank correlation (denoted as  $\rho$ ) is also a measure of the similarity between two variables resulting in a value between -1 and 1. But the major difference between the Pearson correlation coefficient and the Spearman rank correlation is that the Spearman rank correlation makes no assumptions regarding the distribution of the data. It is therefore a much more applicable

correlation to use when the variables being analysed cannot be considered normally distributed. Depending on the data, there are different equations that can be used to calculate the Spearman correlation. The first step is to rank each of the scores in the variables according to size, then, if the ranked data does not contain any duplicates (scores that have the same rank, for either of the variables) Equation 18 can be used:

Equation 18: 
$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Where:  $\rho$  = Spearman's rank correlation  
 $d_i$  = the difference between the ranks of corresponding scores  
 $n$  = the number of values in the variables

With large data sets, it is likely that many values will have the same rank and in that case, Equation 19 should be used:

Equation 19: 
$$\rho = r_{rgX,rgY}$$

$$= \frac{N \sum rgX. rgY - \sum rgX \sum rgY}{\sqrt{[N(\sum rgX)^2 - \sum (rgX^2)][N(\sum rgY)^2 - \sum (rgY^2)]}}$$

Where:  $\rho$  = Spearman's rank correlation  
 $r$  = Pearson's correlation coefficient  
 $N$  = number of values in the variables  
 $rgX$  = first variable ranked values  
 $rgY$  = second variable ranked values

Equation 19 shows that, in the case where duplicate ranked values exist, the Pearson correlation is simply applied to the ranked variables in order to produce the Spearman rank correlation.

Since the Spearman correlation does not assume anything regarding the distribution of the variables, this method of correlation is much more valuable than the Pearson correlation when the data contain non-normal variables. The Spearman correlation, does, however assume that both variables are monotonically related as well as being ordinal (can be ordered on an arbitrary numerical scale).

### Multivariate techniques

Multivariate statistical analyses are a more complex form of data analysis, but can also be more valuable than simpler statistical techniques. Where bivariate analysis compares one variable to another, multivariate analysis is capable of considering multiple variables. This means that the dependent or independent variable is no longer a single variable (or vector), but instead becomes a collection of several variables in an array, or matrix (Rosén, 1998).

As was the case with the bivariate techniques, the multivariate techniques that are of interest to this research should aid in identifying correlations, or likeness, between variables (single or multiple variables). There are several multivariate techniques that can be used to aid in this process, the

most common are cluster analysis (CA), principal component analysis (PCA) and partial least squares (PLS) (Shrestah and Kazama, 2007).

#### *Principal component analysis*

PCA is one of the most basic multivariate statistical techniques found in literature and is used to reduce the dimensionality of a data set (Rosén, 1998). This reduction takes place by transforming the original variable space into principal component (PC) latent variable space. The PCs have two key features; firstly they are all orthogonal, and secondly they lie in the direction of maximum variance, in a decreasing order (Rosén, 2001).

Suppose the data is arranged in a matrix,  $\mathbf{X} [n,m]$ , which have been normalized (using z-scores). The matrix contains  $m$  variables that have been measured  $n$  number of times. The dimension, or rank, of this matrix will then be  $r$ , where  $r \leq m$ . The goal of PCA is to reduce the dimensionality of the variable space by separating the original variable space into two new sub-spaces, namely the principal component space,  $\mathbf{M}$ , and the noise (or residual) space,  $\mathbf{E}$ . This is done by assuming that the data matrix can be expressed as follows:

$$\text{Equation 20:} \quad \mathbf{X} = \mathbf{m}_1 + \mathbf{m}_2 + \mathbf{m}_3 + \dots + \mathbf{m}_a + \mathbf{E}$$

Where,  $\mathbf{m}_1$  to  $\mathbf{m}_a$  are column vectors (matrices of size  $[n, 1]$ ) and  $\mathbf{E}$  is error. Ideally, the new variable subspace,  $\mathbf{M}$  (which contains the column vectors  $\mathbf{m}_1$  to  $\mathbf{m}_a$ ) will contain the majority of the variance of the original data, but a lower dimension, e.g.  $a < r$ . If however  $a = r$ , then  $\mathbf{E}$  must be zero, which means that all the variability of the original data is contained in the matrix  $\mathbf{M}$ .

Therefore, if  $a < r$  then the principal component space retained less variables than the original variable space, thereby reducing the dimensionality of the system. The only remaining task is in calculating the matrix,  $\mathbf{M}$ .

The matrix,  $\mathbf{M}$ , can be written as the product of two other matrices as seen in Equation 21:

$$\text{Equation 21:} \quad \mathbf{M} = \mathbf{T}\mathbf{P}^T$$

Which means that the original data can be expressed as follows:

$$\text{Equation 22:} \quad \mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E}$$

And in turn:

$$\text{Equation 23:} \quad \mathbf{X}_r = \mathbf{t}_1\mathbf{p}_1^T + \mathbf{t}_2\mathbf{p}_2^T + \mathbf{t}_3\mathbf{p}_3^T + \dots + \mathbf{t}_a\mathbf{p}_a^T + \mathbf{E}$$

The matrices  $\mathbf{T}$  and  $\mathbf{P}$  are called the scores (containing the score vectors,  $\mathbf{t}$ ) and loadings (containing the loading vectors,  $\mathbf{p}$ ), respectively. The loadings can be calculated using the single value decomposition (SVD) of the covariance of the original data matrix, as follows:

$$\text{Equation 24:} \quad \text{cov}(\mathbf{X}) = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^T$$

Where  $\Lambda$  is a diagonal matrix containing the eigenvalues  $\lambda$ . The matrix,  $\mathbf{P}$ , consists of eigenvectors as its column vectors in order that:

Equation 25: 
$$\text{cov}(\mathbf{X})\mathbf{p}_i = \lambda_i\mathbf{p}_i$$

$\lambda_i$  is therefore the eigenvalue associated with  $\mathbf{p}_i$ . The covariance matrix of the original data,  $\mathbf{X}$ , can then be estimated as follows:

Equation 26: 
$$\text{cov}(\mathbf{X}) = \frac{\mathbf{X}^T\mathbf{X}}{n-1}$$

Where  $n$  is still the number of samples of the original data matrix. It should also be noted that the matrix  $\mathbf{P}$  is a unitary matrix, which means that:

$$\mathbf{P}\mathbf{P}^T = \mathbf{I} \quad \text{and} \quad \mathbf{P}^T\mathbf{P} = \mathbf{I} \quad \text{and} \quad \mathbf{P}^T = \mathbf{P}^{-1}$$

Where  $\mathbf{I}$  is the identity matrix.

The scores matrix,  $\mathbf{T}$ , containing the score vectors,  $\mathbf{t}_i$ , is simply the original data projected into the principal component subspace, which is a coordinate system defined by the principal components, rather than the original vectors.

PCA is a useful tool that is typically applied for multivariate EDA. In many cases PCA is applied as a graphical method where new data is plotted in the PC space in order to identify visual changes, in the structure and grouping, from the plot of the training data. This is done by simply calculating the scores of the new data, using the model loadings, as seen in Equation 27:

Equation 27: 
$$\hat{\mathbf{T}} = \mathbf{X}_{new}\mathbf{P}_{model}$$

### *Linear discriminant analysis*

Linear discriminant analysis (LDA), also referred to as Fisher's discriminant analysis, is a technique that performs dimensionality reduction and classification (similar to PCA) but in this case it is done in order to find the direction for which the classes in the data are optimally separated (Nor, et al., 2015).

In short, this is done by maximising the between-class variance, whilst minimising the within-class variance (Balakrishnama & Ganapathiraju, 1998). In order to perform LDA, the data are arranged in two matrices, one containing the multivariate record (MVR) and the other the response variables, which have been transformed to categorical variables according to the classes of the data.

The MVR can be illustrated as follows:

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,m} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,m} \end{bmatrix}$$

Where  $n$  is the number of records and  $m$  the number of variables in the MVR.

There are two approaches to LDA, namely class-dependent transformation and class-independent transformation. This study will follow the latter approach, since it is slightly simpler (only has one optimising criterion). LDA starts by computing the mean for each of the classes, as well as the mean of the complete data set in the MVR using Equation 28.

Equation 28: 
$$\mu_{MVR} = p_1 \cdot \mu_1 + p_2 \cdot \mu_2 \cdots p_j \cdot \mu_j$$

Where  $p$  is the *a priori* probabilities of the classes,  $\mu$  is the mean of each of the classes and  $j$  is the number of classes. Typically  $j < m$  and in the case of this study, only two classes were used ( $j = 2$ ).

The within-class and between-class scatter is then used by LDA to formulate a criteria for the separability of the classes. The within-class scatter is determined using the expected covariance of each of the classes (Equation 29) in order to calculate the scatter measures (Equation 30).

Equation 29: 
$$cov_j = (x_j - \mu_j)(x_j - \mu_j)^T$$

Equation 30: 
$$S_w = \sum_j p_j \cdot (cov_j)$$

The between class scatter can be calculated using Equation 31.

Equation 31: 
$$S_b = \sum_j (\mu_j - \mu_{MVR})(\mu_j - \mu_{MVR})^T$$

LDA then becomes an optimising problem which strives to maximise  $S_b$  whilst minimising  $S_w$ , or simply to maximise  $S_b/S_w$ . Mathematically, this is done by using an optimising criterion, as shown in Equation 32. Since the class-independent transformation approach was used in this study, only one optimising criterion is required, instead of one optimising criterion per class.

Equation 32: 
$$criterion = inv(S_w) \cdot S_b$$

By maximising this criterion, the axis of the transformed space is defined. The eigenvector matrix of the criteria, as seen in Equation 32, is then used to perform a linear transformation of the data (Equation 33). The best transformation is typically along the largest eigenvector axis.

Equation 33: 
$$x_{trans} = trans^T \cdot x^T$$

Where  $x_{trans}$  is the transformed variable,  $trans^T$  is the transposed transformation matrix (LDA transform) and  $x^T$  is the transposed original variable. The same transformation matrix can then be used to also transform the variables in the testing data set (Equation 34).

Equation 34: 
$$y_{trans} = trans^T \cdot y^T$$

Once the testing variable is transformed, the data points can be classified using the Euclidean distance, as seen in Equation 35.



Equation 35: 
$$dist_j = y_{trans,i} - \mu_{trans,j}$$

Where  $j$  is an index value for each of the classes,  $\mu_{trans,j}$  is the mean of the transformed training class and  $y_{trans,i}$  is the  $i^{\text{th}}$  data point in the transformed testing variable. The equation is applied once per class, for each data point in the testing variable. This results in a distance for each class, for each data point in the testing variable. Each data point is then classified according to the class at which the shortest distance was calculated.

#### *Partial least squares regression*

Partial least squares (PLS) regression is a statistical method that combines multiple linear regression (MLR) and PCA. This is done by performing 'pseudo PCA' (since the components are not necessarily orthogonal) on both the dependent and independent matrices (multivariate records) whilst information is passed between the two matrices, iteratively, in order to ensure that the principal components of  $\mathbf{X}$  are not only responsible for describing a maximum variance in  $\mathbf{X}$ , but also a maximum correlation with  $\mathbf{Y}$  (in other words, maximising the covariance between  $\mathbf{X}$  and  $\mathbf{Y}$ ) (Hervè, 2007). The form of a basic MLR model can be seen below:

Equation 36: 
$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$$

In Equation 36, both  $\mathbf{Y}$  and  $\mathbf{X}$  are multivariate records of the dependent and independent variables respectively,  $\mathbf{B}$  is called the regression matrix and  $\mathbf{E}$  is the residual matrix.

PLS is used to determine the regression matrix,  $\mathbf{B}$ , by minimising the residual (Rosén, 1998). This algorithm makes use of three basic equations, as seen below:

Equation 37: 
$$\mathbf{X} = \mathbf{TP}^T + \mathbf{E}$$

Equation 38: 
$$\mathbf{Y} = \mathbf{UQ}^T + \mathbf{F}$$

Equation 39: 
$$\mathbf{X} = \mathbf{TBQ}^T$$

Equation 37 and Equation 38 simply shows the 'pseudo PCA' form of the independent and dependent variables, where  $\mathbf{U}$  and  $\mathbf{Q}$  are the scores and loadings, respectively, of the independent variable.  $\mathbf{E}$  and  $\mathbf{F}$  are the residuals of the independent and dependent variables, respectively. Equation 37 and Equation 38 can also be referred to as the outer relations of the independent and dependent variables, respectively. In PLS the outer relations of the variables are built separately (Rosén, 1998).

Equation 39 contains the inner relation that exists between the independent and dependent variables. The main feature of the inner relation is the regression matrix,  $\mathbf{B}$ . Depending on the relationship between the  $\mathbf{X}$  and  $\mathbf{Y}$ , the inner relation can be formulated differently. The simplest expression of the internal relation is a linear one:

Equation 40: 
$$\mathbf{u} = \mathbf{bt}$$

Where  $\mathbf{u}$  and  $\mathbf{t}$  are column vectors of  $\mathbf{U}$  and  $\mathbf{T}$  respectively. The least squares method can be used in order to estimate  $b$ .

Equation 41: 
$$b = \frac{\mathbf{t}^T \mathbf{u}}{\mathbf{t}^T \mathbf{t}}$$

Equation 41 shows how the least squares method can be applied in order to obtain a regression coefficient for a given combination of  $\mathbf{t}$  and  $\mathbf{u}$  variables.

The final equation, which contains both the inner and outer relation, can be seen in Equation 42:

Equation 42: 
$$\mathbf{Y} = \mathbf{T} \mathbf{B} \mathbf{Q}^T + \mathbf{F}$$

Where  $\mathbf{B}$ , is a diagonal matrix of  $\mathbf{b}$ , the regression vector.

The non-linear interactive partial least squares (NIPALS) algorithm is a common algorithm that iteratively solves the above system of equations (Rosén, 2001). Another algorithm which was used in this study, is the SIMPLS algorithm that produced a  $\mathbf{b}$  vector called the regression vector. This vector can then directly be used to perform the desired predictions as illustrated in Equation 43.

Equation 43: 
$$\hat{\mathbf{Y}} = \mathbf{b} \mathbf{X}_{test}$$

The importance of PLS regression is that it overcomes several issues that arise when large data sets are used for MLR. The main issue arising from MLR is multicollinearity, which is likely when  $\mathbf{X}$  contains a large number of variables in comparison to the number of records in  $\mathbf{X}$  (Hervè, 2007).

One solution is to reduce the number of variables in  $\mathbf{X}$ . This can typically be performed using stepwise regression, or PCA (which reduces the dimensionality of  $\mathbf{X}$ ). In the case of PCA, the scores of  $\mathbf{X}$  are then used in the MLR in order to predict values of  $\mathbf{Y}$ . Since the scores are orthogonal, this eliminates the issue of collinearity, which is essentially what principal component regression (PCR) entails (Rosén, 1998).

In PLS, the issue of multicollinearity is resolved by performing “PCA” on both the dependent and independent variables (Hervè, 2007). An issue that arises with PLS is deciding the number of PCs (or latent variables) to retain. Since the  $\mathbf{X}$  variable is independent of the  $\mathbf{Y}$  variable, there is no sense in choosing the number of latent variables using the same methods that are used to select the optimal number of PCs from  $\mathbf{X}$  in PCA, since those PCs are used to explain  $\mathbf{X}$ , not  $\mathbf{Y}$  (Hervè, 2007).

It is only after performing PLS that the number of latent variables that were retained can be analysed in order to determine how much of the variance in  $\mathbf{X}$  and  $\mathbf{Y}$  are explained. The most common method for determining the optimal number of latent variables to retain is a form of error-estimation like cross-validation (Kadlec, et al., 2008). This entails the selection of two data sets, a training set and a testing set. The models are trained using the training data and then applied to the testing data. Typically this procedure is repeated several times, once for every different number of retained latent

variables in order to compare the performance of each run using some form of numeric performance measure (RMSE [root mean squared error],  $R^2$ , p-value, etc.) (Kadlec, et al., 2008).

Different cross-validation techniques have been developed, but only two techniques will be discussed here. The first is the k-fold cross-validation technique and the second is the leave-n-out (LnO) cross-validation technique (Buydens, 2013).

With the k-fold technique, the data is divided into 'k' number of folds (segments of equal size). Then the PLS model is applied to all but one of the folds, which means that 'k-1' folds represent the training data, and the remaining fold is the testing data. This procedure is repeated 'k' number of times in order that each of the folds have a turn at being the testing data set. The performance for each run is the averaged for all 'k' number of runs.

The second method, LnO, divides the data into a training and a testing set using only one point of division, 'n', which is typically between 10% and 20% of the total number of records in the original data set (Buydens, 2013). The PLS model is then applied to both data sets and the performance is recorded. This method is much simpler than the k-fold technique, but it has some disadvantages.

If a data set contains some abnormality that is observed only once (or in a small range within the data records), then the model will show poor performance each time that abnormality lies within the testing data set of the cross-validation method.

With the k-fold technique, this will likely only happen with one of the 'k' folds, which means that the poor performance will be improved when the performance measures of each of the 'k' runs are averaged.

With the LnO method, on the other hand, if the data abnormality happens to fall within the 'n' range of records containing the test set, then the performance of the model will be poor without any indication that the performance may have been better if only another range of 'n' could be selected. Thus, using the k-fold technique ensures that the results are unbiased since the test data set varies over the complete range of data records that are available.

### **3.4 SUMMARY OF LITERATURE REVIEW**

From the literature review, it is evident that WRPs rely on monitoring systems for risk management and plant performance. In both cases, there is a need to reduce the lag time associated with sampling and obtaining analytical results. The knowledge based approach is a well-established approach for identifying surrogates and indicators and makes use of expert process knowledge in order to simplify and streamline the process.

The data-driven approach, on the other hand, is relatively new and can benefit from the correct application of expert process knowledge, especially in the field of water science and treatment systems. The statistical analyses that have been found in literature (despite not being applied at WRPs) was useful for this study also. Although the results of this study may not be directly comparable to that found

in literature, good information was obtained with regards to building the statistical models and what to expect in terms of outputs. Data pre-processing was highlighted as a crucial component of the data-driven approach and a good understanding of training and testing the statistical models have been obtained.

Ultimately, the procedures that were proposed by Haimi et al. (2013), Kadlec et al. (2008) and Lin et al. (2007) were summarised as follows:

- 1) Data acquisition, selection and inspection
- 2) Data pre-processing
- 3) Model design (selection and training)
- 4) Model maintenance (testing and validation)

This approach was followed to some extent and played a role in the methods that were applied in this study. With regard to the methods that were included in this study, the following were selected based on their success in similar studies:

- Bivariate analyses
  - Pearson's Correlation Coefficient (parametric method)
  - Spearman's Rank Correlation (non-parametric method)
- Multivariate analyses
  - PCA (for dimensional reduction of the MVR)
  - LDA (for categorical/qualitative prediction)
  - PLS regression (for quantitative prediction)

These methods and their application to the data are discussed in more detail in the next section.

## 4 METHODOLOGY

From the literature review it is known that the objectives identified in order to achieve the aim of this study, are in accordance with that of similar studies. The practical implementation of methods that were used in order to complete each of the objectives of this study are discussed in this chapter.

### 4.1 OBTAINING, ORGANISING AND ASSESSING HISTORICAL PLANT DATA

The input data which was used in this study was obtained from an operational WRP that had been operational for more than five years at the time the data were obtained. The data were obtained from several sources including the laboratory information management system (LIMS) from the laboratory that performed the majority of the off-site analyses on samples, data logs that recorded the results of analyses performed at the on-site laboratory and recordings that were made from the SCADA system.

This data contained numerous imperfections such as outliers, missing values and noise which necessitate pre-processing of the data; but before this could be done, the larger scale imperfections should be addressed first. These large scale imperfections are major events that effect the entire plant such as power failures, treatment failures and on-site emergencies. These events are typically logged in an event log book, or registry, which was also made available to this study.

#### 4.1.1 Organising and initial selection of plant data

The plant data were collected in several different formats and were immediately consolidated into a single database listing the time, sampling location and analysis name for each of the data elements that were obtained. During these processes several issues were addressed, most notable: incorrect data labels and inconsistent sampling and measurement rates.

In cases where incorrect data labels were used (due to typos), the labels were corrected (if sufficient information was available). The inconsistent sampling and measurement rates were addressed by averaging all samples that were taken more than once per day in order to provide a daily average. Where samples were taken at a rate slower than once per day, the un-sampled dates were considered missing values. A sample-and-hold strategy may be applicable in some studies, but it was not conceivable to assume that variables on the plant would remain constant for more than a day, let alone a week or two.

The organised data could then be transformed into a matrix containing different variables as columns, and different dates as rows (records). Each variable was provided with an identification (ID) number and each ID corresponded to a single sampling location and analysis. The complete list of IDs (also referred to as variables) with their descriptions can be seen in Appendix A, but Table 4.1 shows an excerpt from the complete list, for illustrative purposes.

**Table 4.1:** Description of data IDs

ID	Sampling Point	Analysis
1	WWTP Clarifier	Nitrate
27	WWTP Final Eff.	Turbidity
90	WRP Inf.	Chlorophyll A
121	DAF	Manganese (Mn)
124	SF	Turbidity
159	Ozone Contact Final	Residual O <sub>3</sub>
173	BAC	DOC
187	GAC	COD
196	UF	EC
216	Final Water	DOC
240	Operational	Pre-Ozone (mg/l)
246	Operational	Raw Water (m <sup>3</sup> /h)

A total of 270 IDs were created in treatment process order. The last 30 variables are operational variables that were recorded from the SCADA of the plant and were not listed in treatment process order, but were kept under the same 'sampling point' in order to distinguish the SCADA data from the remaining data. This was done since the SCADA data contained much less missing values and were also in many cases duplicate measurements from already existing variables.

#### **4.1.2 Data validation and final selection of plant data**

A validation of the plant data were performed as part of another research project (Swartz, et al., 2016). Since the aim of this study was to detect correlations and not process performance, the validation method will not be identical.

The validation was performed to help identify sections of data that will be most suitable for detecting relationships between the variables, which can then be exploited to aid the monitoring of the plant during normal operating conditions. It is therefore important to identify sections of data that cannot be considered normal operating conditions. From there on it is only a matter of having sufficient amount of data in order to add a degree of certainty to the results of the analyses.

##### Normal operating conditions

The data received for this study included a large amount of water quality data, a smaller amount of operational data and also the log books used on-site to log events. The events typically logged include dates and duration periods of downtime caused by power failure, treatment failure and out-of-spec raw water. It was therefore easy to identify the periods where the plant was not running under normal operating conditions.

Dates that experienced critical events (events resulting in a plant shut-down) were used to create a mask. The mask was then applied to the plant data in order to remove the data records that were

measured whilst a critical event was taking place. The remaining data were considered to have been captured during normal operating conditions.

It should be noted that small and large fluctuations in the plant performance and operational procedures are still included in the data. It is important that the data include these periods since they will also occur in the real world. It is therefore also necessary to test the data in order to ensure that the data is a sufficient representation of the actual plant behaviour.

### Minimum data set size

After the data had been reduced to only data that were measured during normal operating conditions, it was important to determine if the remaining data contained enough records that will ensure that the statistical analyses yield results that are both reliable (repeatable under similar, but not identical conditions) as well as robust (remained functional under similar, but not identical conditions).

For the univariate and bivariate analyses the minimum data set size can be set to 30 data points. This is a general rule of thumb typically used in simple statistical analyses, such as bivariate correlations like Pearson's correlation coefficient. Any pair of variables that have more than 30 data records in common can therefore be used, since the data can be considered a stationary time range that was taken over a long period of time (5 years). Thus this assumption is valid when also considering the low sampling rate which results in the data being stationary.

For multivariate analyses, the minimum data set size is dependent on the type of analysis that will be performed which can be grouped according to the number of latent variables that are retained. The first group considers analyses where all the latent variables are retained; the other group considers analyses where a finite number of latent variables, which are less than the total number of latent variables, are retained.

Variables with too few data points were immediately removed from the data set that was used for the remainder of the tests and analyses, including pre-treatment.

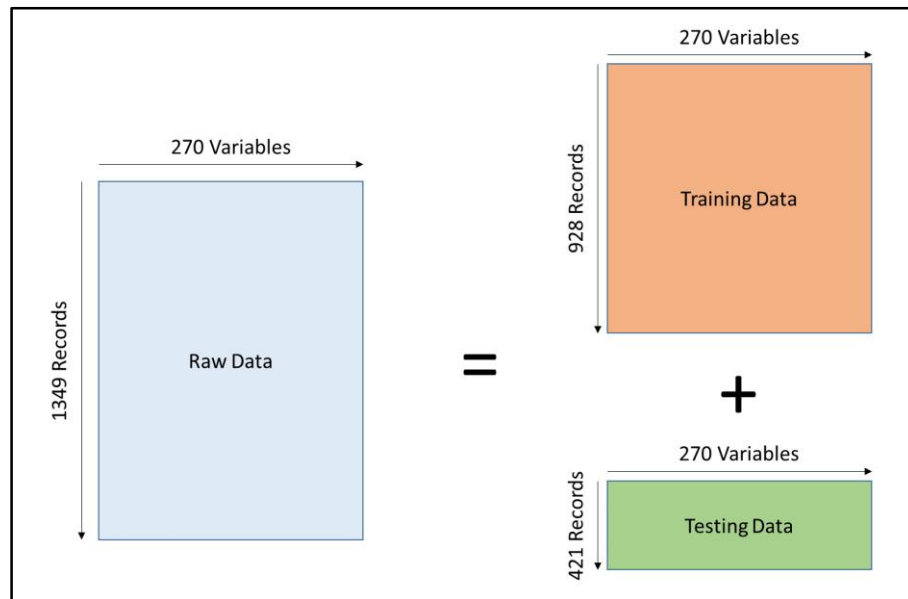
### Processing ability

Data that contain large amounts of missing and dirty data (data with erroneous record information), hinders the performance of statistical models. It is therefore important to quantify the amount of missing values and dirty data in the data set. Interpolation can be used to replace missing values (to a certain extent) and outlier analysis can be used to remove dirty data.

It was therefore important to determine the degree to which the data had been corrupted in order to determine whether or not pre-processing would actually improve the quality of the data - without adding too much artificial data, which would result in data that is a poor representation of the plant behaviour.

### 4.1.3 Data pre-processing

The remaining data, after validation and selection, were divided into two data sets: the first for training, and the second for testing the different models (see Figure 4.1).



**Figure 4.1:** Illustration of training and testing data separation

Before the models could be trained, however, it was necessary to pre-process the training data, in order to ensure that the models will function properly and provide accurate results.

#### Outliers

Due to the nature of the data, outliers can be expected and should be tested for and removed when found. Statistical estimates were used for each of the variables in order to determine whether or not a data point truly belongs to the variable, or not. In order to obtain the statistical estimates required for this task, the distribution type of each of the variables had to be identified first.

The variables were classified as one of three possible distributions:

- Normal distribution
- Lognormal distribution
- Empirical distribution

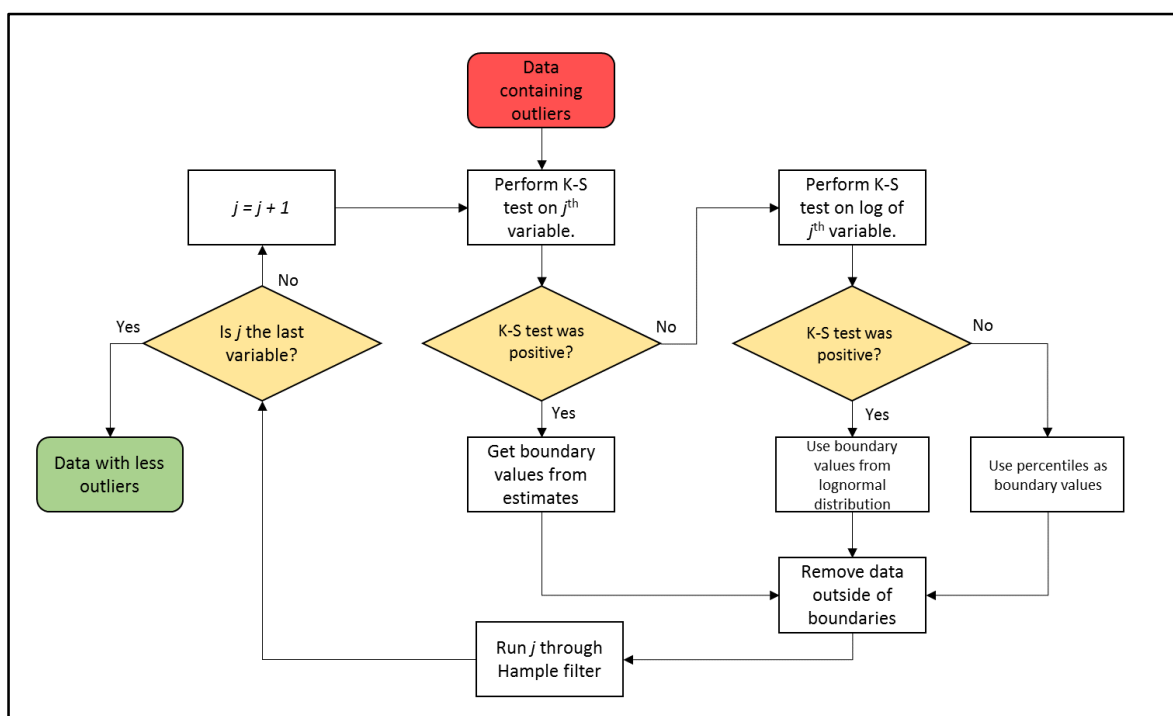
The distributions as well as the upper and lower limits that were used to identify outliers for each of the variables can be seen in Appendix A, but an excerpt of this can also be seen in Table 4.2.



**Table 4.2:** Variable distribution and limits used for outlier detection

ID	Sampling Point	Analysis	Distribution	Lower limit	Upper limit
1	WWTP Clarifier	Nitrate	Normal	-4.4	21.5
23	WWTP MP B8	Ammonia	Lognormal	0.1	20.2
29	WWTP Final Eff.	COD	Empirical	11.9	86.0
73	WRP Inf.	DOC	Lognormal	4.4	12.1
107	DAF	Turbidity	Lognormal	0.4	2.0
132	SF	Iron (Fe)	Empirical	0.0	0.2
159	Ozone Contact Final	Residual O <sub>3</sub>	Lognormal	0.1	0.5
173	BAC	DOC	Lognormal	1.2	5.5
187	GAC	COD	Lognormal	1.8	27.5
194	UF	Turbidity	Lognormal	0.0	0.2
207	Final Water	Free chlorine	Empirical	0.7	3.6
246	Operational	Raw Water (m <sup>3</sup> /h)	Empirical	400.0	900.0
270	Operational	GAC runtime (h)	Empirical	0.0	368.2

A block diagram of the algorithm used to remove outliers from the data is shown in Figure 4.2.

**Figure 4.2:** Block diagram illustrating outlier removal algorithm

Once the distributions for all of the variables had been determined, the relevant statistical estimates (mean and standard deviation) were determined. The outliers could then very easily be detected by setting an upper and lower threshold using the statistical estimates. In the case of the normal and lognormal distributions, outliers were identified using Equation 9.

Equation 9:  $|x_i - \bar{x}| > 3 \cdot \sigma$

Where  $x_i$  is any data point in the variable  $x$  whilst  $\bar{x}$  and  $\sigma$  are the mean and standard deviation of the variable  $x$  respectively (or of the transformed variable in the case of lognormal variables).

In the case where a variable could not be classified as either having a normal or lognormal distribution, it was classified empirical. In the case of empirical variables, outliers were identified using the following equations:

Equation 44:  $x_i > p_{99}$

Equation 45:  $x_i < p_1$

Where  $x_i$  is any data point in the variable  $x$  whilst  $p_{99}$  and  $p_1$  are the 99<sup>th</sup> and 1<sup>st</sup> percentile of the variable  $x$  respectively. All of the variables (irrespective of distribution) were then further treated for outliers using a Hampel filter, since Hampel filters are resistant to outliers. The Hampel filter used Equation 10 to determine outliers whilst moving through each variable with a moving window the size of 23 data points.

Equation 10:  $|x_i - x^*| > h \cdot x_{MAD}$

In this case,  $x_i$  is the data point being evaluated,  $x^*$  is the median of the data points in the window and  $x_{MAD}$  is the MAD of the data points in the window. The value of  $h$  was set to 3.

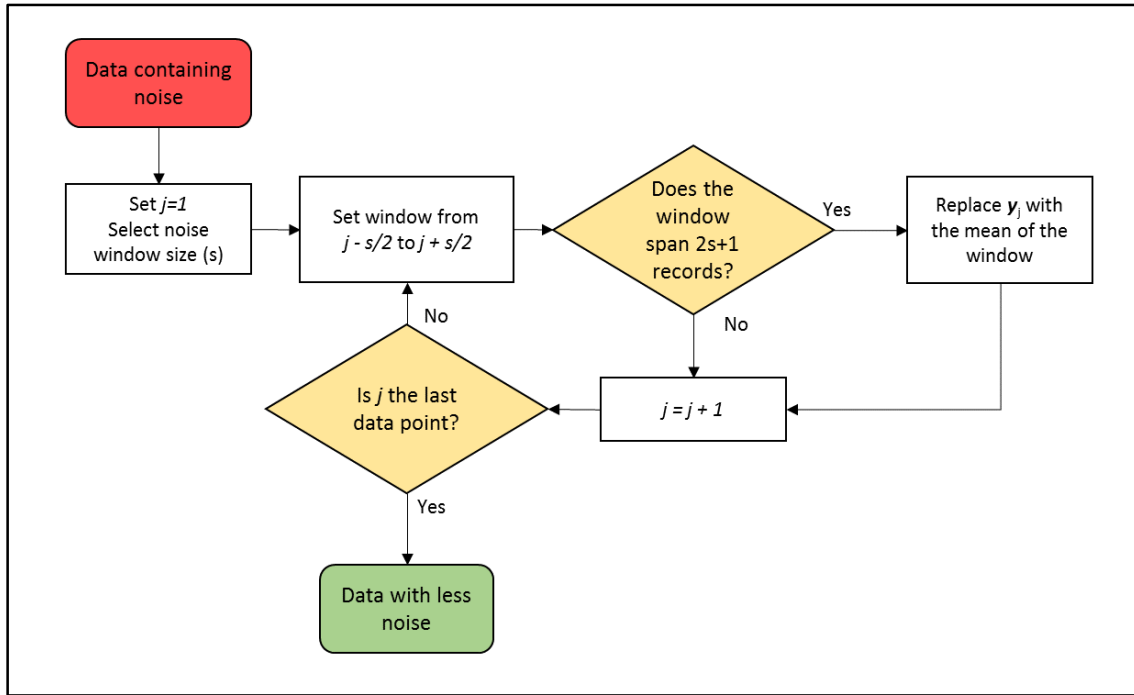
The parameters of the Hampel filter were not chosen at random, several tests were conducted at three different values for the window size (11, 23 and 51) and  $h$  (2, 3 and 5). Using expert process knowledge, the values of 23 and 3 for the window size and  $h$ , respectively, were chosen.

Since the removal of outliers has a large effect on the variance of a variable, variables with zero or insignificant variance-to-mean ratio (or that could not produce unique values for the 1<sup>st</sup> and 99<sup>th</sup> percentiles) were removed as a whole from the data since they cannot contribute to the identification or verification of any form of relationship with other variables. This is because the methods used to determine relationships between variables typically compares the similarity between the variances (and changes within the variance) of the variables.

### Noise

After the outliers were removed, the remaining data were considered as relevant and true measurements of the different variables they represent, this did not, however, mean that they were 100% accurate and completely reliable. Noise could still be present, and responsible for measurement variations that are not truly present on the plant.

An equally weighted, centred (or symmetric), moving average filter (see Figure 4.3) was applied to the data.



**Figure 4.3:** Block diagram of noise removal algorithm

Equation 46 shows the equation for the equally weighted, centred, MA filter that was used to perform the noise reduction illustrated in Figure 4.3.

Equation 46: 
$$\hat{y}_{(j)} = \frac{1}{2 \cdot s + 1} [y_{(j-s)} + \dots + y_j + \dots + y_{(j+s)}]$$

In this case  $\hat{y}_{(j)}$  represents the new data point and  $s$  is a sizing factor used to determine the size of the moving window. For this study a value of  $s = 5$  was used and therefore the size of the moving window was 11 data records.

Similarly to the parameters of the Hampel filter, the value of  $s$  was chosen after performing several tests using a variety of values for  $s$  (2, 5 and 12).

### Replacing missing values

Data containing large amounts of missing values can be problematic when it comes to building statistical models, especially when multivariate analyses are considered. It is therefore important that missing values be replaced, but with caution. If too many data points are replaced, the data will contain a high degree of artificial data, which means that the resulting models will, in part, not function in the real-world since the real-world data will not contain any artificial data. The simplest way of replacing missing values is by interpolation.

It was also decided to make use of linear interpolation, but only on the MVR, which consisted of the FEM variables, all of which contained less than 20% missing values. It was, however, still expected that the interpolated data may end up reducing the accuracy of the models by removing, or hiding, some of the characteristics of the actual historical data, but this was considered an acceptable risk.

## 4.2 DEVELOPING STATISTICAL MODELS FOR THE IDENTIFICATION OF SURROGATES AND INDICATORS

At this stage, the data were in a format and condition where the statistical analyses could begin. This was done in two phases. The first phase was responsible for developing and training statistical models that were applied to the data in order to either quantify the relationships between the variables directly (bivariate analyses); or make predictive models of the SDM variables based on the FEM variables (multivariate analyses). In the case of the multivariate analyses, surrogates and indicators can be identified through further work if the prediction of the SDM variables were accurate. This was done during the second phase, responsible for testing the models, by applying them to a new data set that is completely independent from the training data set and has not been analysed by any of the models before.

### 4.2.1 Bivariate statistical analyses

For the bivariate analyses, the training data were arranged in a matrix according to the ID values of each of the variables (which is in the same order as the treatment processes of the studied WRP). The different variables within each group (of the same sampling point) were arranged in order from least to most missing values in the variable. This was simply done in order to make it easier to interpret the results from the bivariate analyses (which will be discussed in more detail at a later stage).

Two bivariate analyses were performed in order to determine the Pearson correlation coefficient and the Spearman correlation coefficient. In both cases, the p-values for each correlation were also calculated.

The p-values refer to the result of a hypothesis test, wherein the null hypothesis ( $H_0$ ) states that there is no correlation between the variables and the alternative hypothesis ( $H_1$ ) states that there is a correlation between the variables. In both cases, a p-value less than or equal to 0.05 was considered sufficient in order to reject the null hypothesis, meaning that there is sufficient statistical certainty that there is some correlation between the variables.

A correlation matrix was then built wherein the position of each value represents the correlation between the variables according to the row and column number where the value is located. This can also be represented as follows:

$$\begin{bmatrix} r_{1,1} & \cdots & r_{1,j} \\ \vdots & \ddots & \vdots \\ r_{i,1} & \cdots & r_{i,j} \end{bmatrix}$$

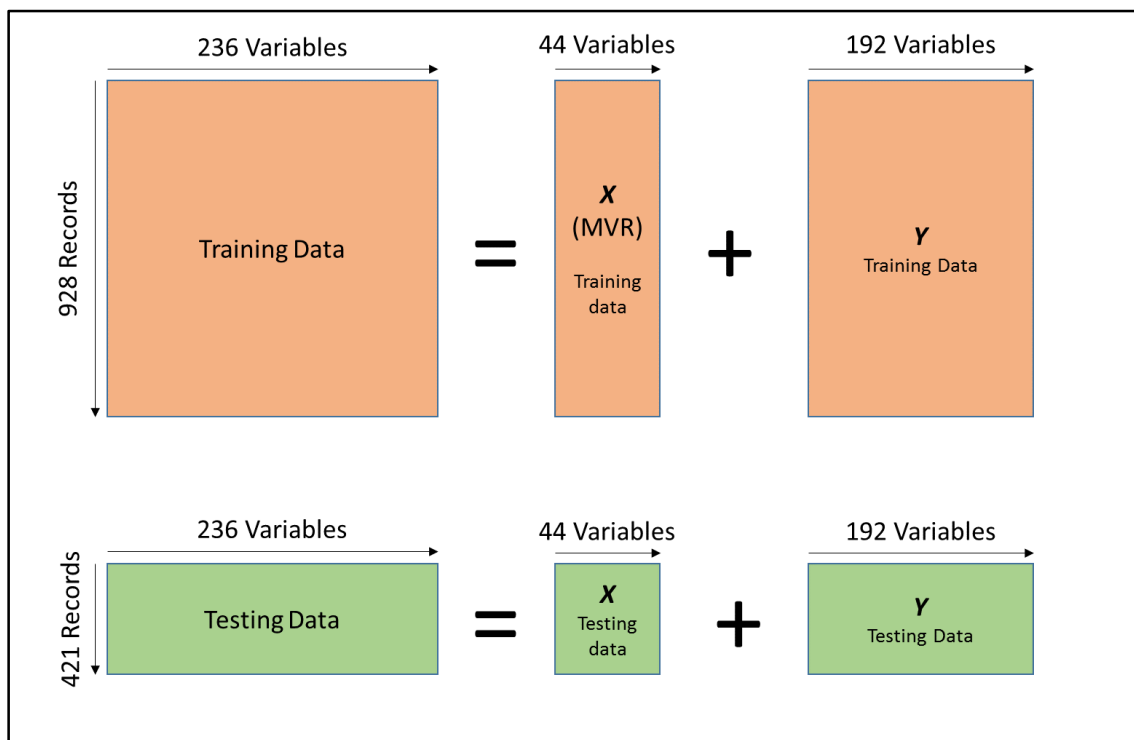
Where  $r$  is the correlation coefficient,  $i$  represents the row number and  $j$  represents the column number. Therefore,  $r_{3,45}$  represents the correlation coefficient between the 3<sup>rd</sup> and 45<sup>th</sup> variables. In order to simplify the visualisation of the results, the correlations that had insufficient certainty for rejecting the null hypothesis (where the p-value was above 0.05) were removed from the correlation matrix.

The resulting matrix (consisting of only statistically significant correlations) was then plotted on to a graph showing a colour map (where correlations higher than 0.7 are coloured red, higher than 0.8 are coloured green and higher than 0.9 are coloured blue).

#### 4.2.2 Multivariate statistical analyses

For the multivariate analyses, the data structure had to be changed even more (see Figure 4.4). Apart from having a training and testing data set, the multivariate analyses necessitate the creation of an **X** variable, known as the predictor variable, which is a MVR (contains more than one variable). By definition, the **X** variable contains the FEM variables, the remaining variables are grouped together to form the **Y** variable, known as the response variable.

From Figure 4.4 it should be noted that the training and testing data sets no longer contain 270 variables, but only 236. This is due to the variables that were removed from the study for having insignificant variances.



**Figure 4.4:** Illustration of the creation of the **X** and **Y** data sets

The multivariate analyses were then performed or trained on the training data using the **X** and **Y** variables. In all cases, the **X** variable was considered a single entity whilst the **Y** variable was seen as a collection of **y** vectors in order to build and train models that will relate the entire **X** variable to a single **y** vector. Therefore, each of the multivariate analyses produced 192 results, which in turn means that for the supervised analyses, 192 different models were built and trained.

### Principal component analysis

PCA was performed on the MVR in order to reduce the dimensionality of the data. Several plots were made of the resulting principal components and component scores. The percentage of variance explained by each of the principal components were plotted in order to identify whether the majority of the variance in the MVR data could be explained by only a few principal components. If this is the case then there is a high degree of correlation between the different variables. The remaining components that explain only a small amount of the variance in the data may be capturing noise, it is therefore an option to exclude those components from future analyses.

A multi-plot consisting of six scatter plots were made of the first four principal components against each other, since these capture the highest amount of variance and are least likely to represent noise in the data. This was done in order to identify whether or not clusters or distinguishing features can be identified in the scatter plots. This can be done by colouring the scatter plots according to the values of the variables contained in the response variable (those not included in the MVR). By doing this, it is possible to identify a potential relationship between the MVR and each of the response variables individually.

The aim of using PCA is to reduce the  $\mathbf{X}$  variable in dimension in order to better visualise the variance in the variable, and after colouring with the  $\mathbf{y}$  variable, to see if any clusters in the PC scatter plots correspond to certain levels in the  $\mathbf{y}$  variable. Although this is much more useful for dynamic continuous soft sensors, it can still be indicative of a relationship between the  $\mathbf{X}$  and  $\mathbf{y}$  variables.

### Linear discriminant analysis

The first step in performing LDA is to make a categorical response variable ( $\mathbf{C}$ ) which is the same size as the numeric training response variable ( $\mathbf{Y}$ ). In order to transform the numeric values into categorical values, boundary values (or limits) ought to be established for each cluster (group of similar categorical values). The numerical values can then be categorised using the limits of the clusters in order that all the numeric values that fall within the limits of a certain cluster, will also have the same category.

Initially it was planned to group the data into three clusters based on the limits provided by water quality standards and guidelines. Unfortunately, only 19 variables had published limits, which meant that only 10% of the  $\mathbf{Y}$  variable could be used for LDA models. It was therefore decided to rather use the mean and standard deviation (see Equation 47) of each  $\mathbf{y}$  variable to make an 'LDA limit' for that  $\mathbf{y}$  variable. The data were then also grouped into two clusters: 'Above' and 'Below', representing numeric values of a given variable that were either above or below the LDA limit for that variable.

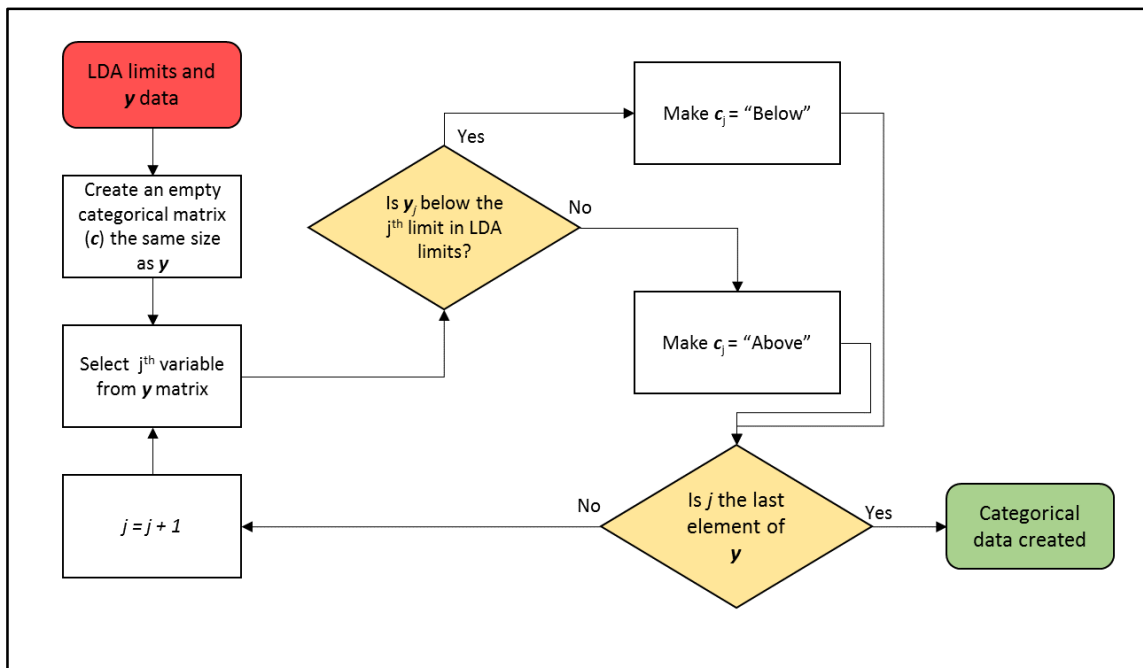
Equation 47: 
$$LDA\ limit_j = \bar{y}_j + 1 \cdot \sigma_j$$

Where  $\bar{y}_j$  is the mean of the  $j^{\text{th}}$   $\mathbf{y}$  vector and  $\sigma_j$  is the standard deviation of the  $j^{\text{th}}$   $\mathbf{y}$  vector. The LDA limits for all of the variables can be seen in Appendix A as well as the excerpt shown in Table 4.3.

**Table 4.3:** LDA limits used to categorise variables

ID	Sampling Point	Analysis	LDA limit
1	WWTP Clarifier	Nitrate	11.8
23	WWTP MP B8	Ammonia	3.3
29	WWTP Final Eff.	COD	54.1
73	WRP Inf.	DOC	8.5
107	DAF	Turbidity	1.4
132	SF	Iron (Fe)	0.2
159	Ozone Contact Final	Residual O <sub>3</sub>	0.3
173	BAC	DOC	3.3
187	GAC	COD	14.9
194	UF	Turbidity	0.1
207	Final Water	Free chlorine	7.7
246	Operational	Raw Water (m <sup>3</sup> /h)	1952.9
270	Operational	GAC runtime (h)	251.4

These limits were then used to transform the **Y** variable into the **C** variable, containing the categories of each of the **y** variables (see Figure 4.5). This same algorithm was used for both the training and the testing data.

**Figure 4.5:** Box diagram of numeric to categorical transformation algorithm

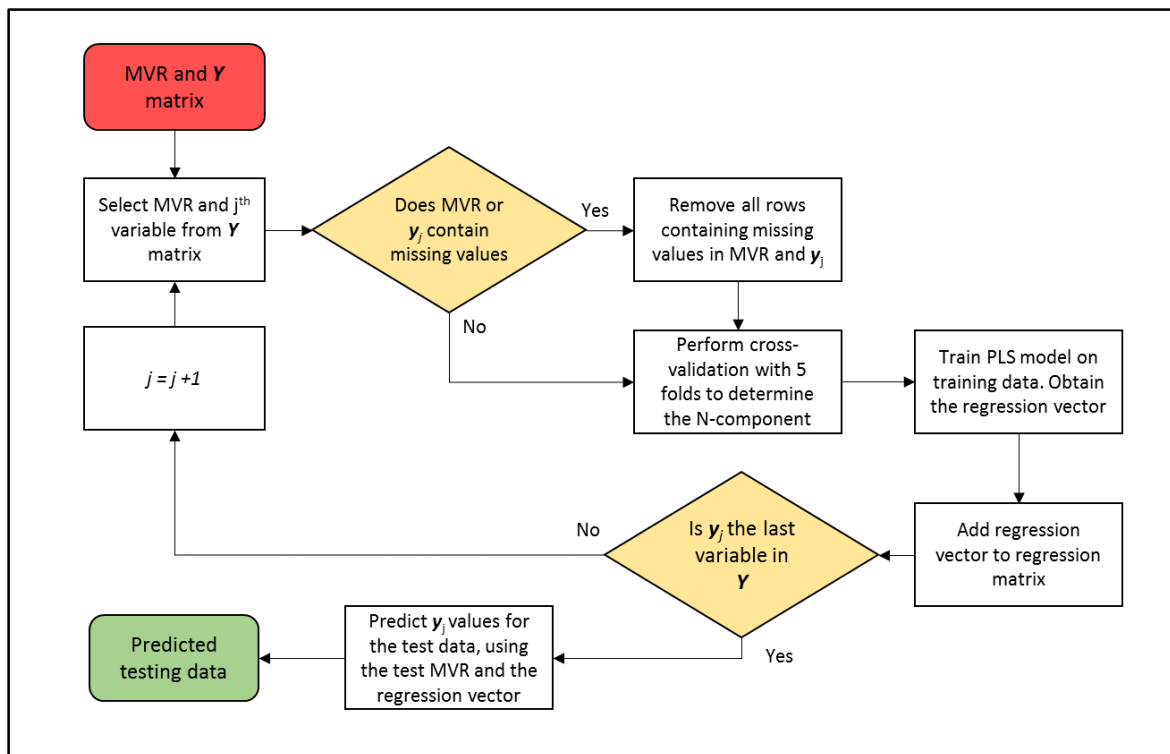
The categorical data were then used to build and train LDA models that make use of the **X** MVR in order to predict the category of each specific **y** variable for each specific record contained in the **y** variable. The accuracy of these predictions can then be calculated in order to determine whether or not the **X** MVR can in fact be used to predict the categories of the **Y** variable, which will be indicative of a potential relationship between the **X** and **Y** variables.

### Partial least squares regression

PLS regression is another supervised statistical analysis, although it provides much more detail than LDA. In LDA, a model is trained to place data points of the response variable in groups based on a linear combination of the MVR. In PLS regression however, latent variables of the MVR is used to predict actual values of the response variables. The information provided by PLS regression is therefore not simply a category containing a wide range of values, but rather a single predicted value for each record corresponding to the records of the MVR.

In this study, PLS1 (only a single  $y$  variable is predicted) was performed using the SIMPLS algorithm to perform the iterations. The model was trained on the training data using k-fold cross validation. Initially 10 folds were used, but this resulted in certain folds containing too few data points (in cases where the response variable had very little data points). It was then decided to make use of only 5 folds in order to perform the cross validation.

Figure 4.6 shows a block diagram of the PLS algorithm that was used to train and test the PLS models with. The mathematical equations used in this process are discussed in Section 3.3.5.



**Figure 4.6:** Block diagram of PLS workflow

Cross-validation was performed in order to determine the ideal number of components (N-component). This was done by plotting the average mean standard error of prediction (MSEP) over the total number of folds and for each number of PLS components used by the model. From the plots, the N-component was chosen where the lowest MSEP was found. After the N-component was selected, the PLS model was trained in order to produce the required regression vector that corresponds to each  $y$  vector.



Once this process was completed for all of the vectors in the  $Y$  variable, the regression matrix ( $B$ ) was used with the  $X$  MVR of the testing data, in order to predict values for the  $Y$  variable of the testing data. The predicted and actual  $Y$  variable for the testing data could then be compared in order to determine the prediction accuracy of the models. If the predictions are accurate, it may indicate that there is a relationship between the  $X$  MVR and the  $Y$  variable.

### 4.3 ASSESSING THE PERFORMANCE OF THE DEVELOPED STATISTICAL MODELS

Of the different statistical methods that were used in this study, different outcomes are expected and therefore, different performance measures will be used to assess the models that were based on these statistical methods. This section will discuss the different tests and criteria that were used to assess the performance of each of the different types of statistical models that were used in this study.

#### Bivariate correlations

The performance of the bivariate correlations were assessed by inspecting the correlation coefficient that was calculated between a variable pair and then checking to see if the null hypothesis can be rejected ( $p\text{-value} \leq 0.05$ ). This assessment was made even simpler by the fact that results were graphically illustrated after removing the results that were not relevant (where the null hypothesis could not be rejected).

Using the colour map that was discussed earlier, the bivariate correlations could be assessed by simply inspecting (visually) how many blue, green and red areas are indicated. Interpreting these results, however, require knowledge of the variables and processes from which they originated. This will be discussed at a later stage.

#### Principal Component Analysis

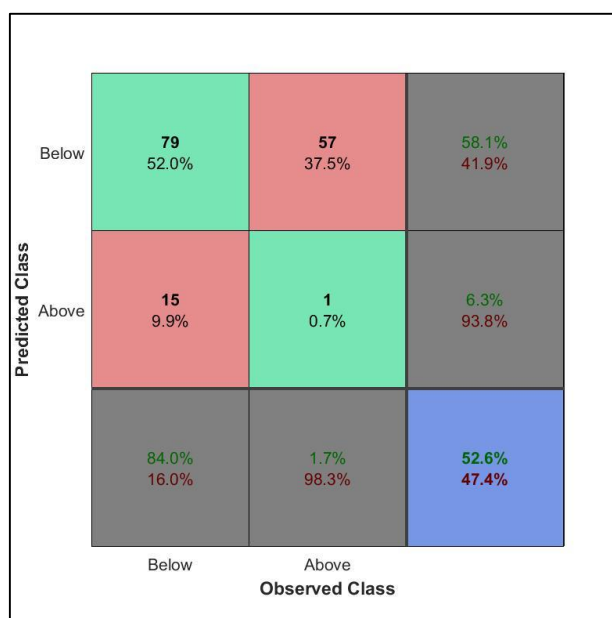
Assessing the performance of the PCA models consisted of a purely visual inspection of the plots that were generated. These plots show the first four principal components of the  $X$  MVR plotted against each other, whilst also being coloured according to each of the response variables.

In cases where a high degree of separation exists between the data points, as well as a similar separation between the colours of those data points, it is probable that a relationship exists between the  $X$  and  $y$  variables.

#### Linear Discriminant Analysis

Assessing the performance of the LDA models were done using the classification error of the predicted classes. This was done by comparing the predicted categorical data (that were predicted using the testing MVR) to the observed categorical data of the testing  $Y$  variable. The classification errors were displayed using confusion matrix plots in order to easily determine the overall classification error as well as the number and nature of each of the errors that were made (false positive or false negative).

It is important to understand the information provided by the confusion matrix plots, therefore, an example will be discussed with the help of Figure 4.7. The plot shows a matrix that is  $(k + 1)$  by  $(k + 1)$  in size, where  $k$  is the number of classes into which the data were categorised. Therefore, the plot actually shows the predictions of the actual classes in the first  $k$  rows and  $k$  columns of the matrix. The last row and the last column of the matrix serve as summaries whilst the very last block in the matrix (bottom right) shows an overall summary of the predictions that were made.



**Figure 4.7:** Confusion matrix plot example

In the  $k$ -by- $k$  portion of the matrix, the actual counts of predicted and observed categories are shown as well as the percentage of the count in relation to the total number of entries made. From Figure 4.7 it can therefore be concluded that the LDA model correctly classified 79 data points as 'Below' which corresponds to 52% of the data points that are included in the model. From the right hand summary column it is known that those 79 predictions constitutes 58.1% of all the predictions that were of the category 'Below'. The other 41.9% of the predictions that were 'Below' were incorrect, thus 57 data points were incorrectly classified as 'Below', this constitutes 37.7% of the data points that were included in the model.

The same applies for the data that were classified as 'Above'. 15 data points (or 9.9% of the data) were incorrectly classified as 'Above', whilst only one data point (0.7% of the data) were correctly classified as 'Above'. In the summary of the 'Above' predictions it is indicated that only 6.3% of the data points that were classified as 'Above' were in fact correctly classified.

The overall summary (bottom right block) indicates that of all the predictions made by the model 52.6% were correct, and 47.4% were incorrect. Thus the accuracy of the model is 52.6%, although this is not the only factor that ought to be considered. As far as water quality predictions are

concerned, it is always better to err on the side of caution. Therefore, it is much more acceptable for a model to make a false-positive error (wrongfully classifying a 'Below' value as 'Above') than a false-negative error. The rate of false-negative errors will therefore also play a role in determining the performance of the LDA models.

#### Partial Least Squares Regression

The PLS models were responsible for predicting numeric values of the response variable of the testing data by using the MVR of the testing data. The two sets of response variables for the testing data (one with observed values and one with predicted values) should therefore be compared to one another in order to determine the similarity between the predicted and observed response variables.

The similarity between these two response variables were tested using  $R^2$  values. The observed and predicted response variables were also plotted against each other on a scatter plot in order to visually assess the accuracy of the model predictions. The plots also showed a line through the origin at a 45 degree angle, representing a perfect prediction, in order to simplify the visual assessments.

#### **4.4 IDENTIFYING AND EVALUATING RELEVANT VARIABLE RELATIONSHIPS USING EXPERT PROCESS KNOWLEDGE**

In order to perform multivariate statistical analyses, the pre-processed data had to be separated into two unique data sets. This separation was done with the statistical analyses in mind. The first data set, called the predictor variable/matrix (or the  $X$  matrix), was to serve as the MVR for the multivariate analyses. Since it is important that the MVR contains no missing values, only variables with less than 20% missing values were included, since those are the only variables that could undergo missing value replacement. The MVR, therefore, contain the FEM variables.

The second data set was called the response variable/matrix (or the  $Y$  matrix) for the statistical models. This group contained two types of variables, the SDM variables and the P-SDM variables.

The P-SDM variables are simply variables that are FEM (based on expert process knowledge), but were grouped in the response variable/matrix since they contained too many missing values to be included in the MVR.

Due to the above mentioned grouping of the different types of variables, the results of the statistical analyses could only be one of the following for each of the models:

**FEM → P-SDM** (accurate prediction of P-SDM variables using FEM variables)

**FEM → SDM** (accurate prediction of SDM variables using FEM variables)

The usefulness or value of these results could be interpreted using expert process knowledge:

### FEM → P-SDM

This result is mainly trivial from the perspective of monitoring and controlling the WRP. But from the perspective of building and evaluating statistical models, this result is still of value since the predicted variable, from a data perspective, was SDM.

### FEM → SDM

This result can be considered the most ideal outcome of the research. It is indicative that a statistical model could be used to predict SDM variables using only FEM variables. This is an essential element of soft-sensors and can potentially be used to great advantage with regards to the monitoring and control of WRPs.

The results from each of the statistical analyses (correlations and models) were reviewed in order to evaluate the different statistical analyses in two critical aspects (evaluation criteria):

- 1) Did the statistical analysis successfully identify relationships that are already known (from expert process knowledge) to exist between certain variables?
- 2) Did the statistical analysis identify any new relationships between FEM and SDM variables that may be of value to the potable water recycling community?

These two questions served as the final evaluation of the statistical analyses that were included in this study. The first criterion is much more important than the second. If any of the statistical analyses failed in this aspect, that statistical analyses should be considered inappropriate for the task of identifying surrogates and indicators.

Models that failed in the second aspect were not considered failures. Although they do not contribute to solving the issue experienced by the potable water reclamation community, they do fulfil the objectives and ultimately the aim of this study.

Models that did not fail in any of the above mentioned aspects, were considered an ideal outcome of this study in that the aims of the study were met, and a contribution to the potable water reclamation community was made.

It should be noted that if the models in this study perform poorly, it is not a definitive proof that the models relied on inferior statistical analyses. Rather, it reflects on the choice of data included in the study, which should be considered as only a single case study.

## 5 RESULTS AND DISCUSSION

This chapter shows the results of the study and in particular the statistical models that were built in order to obtain these results. Since there was a large variety of statistical analyses that were applied in the study, this chapter discusses each of the type of analyses separately.

The structure of this chapter is in line with the objectives of the study, therefore the results will be shown and discussed as follows:

- data validation and pre-processing
- development of the statistical models
- performance assessment of the statistical models
- evaluation of the identified relationships using expert process knowledge

The results for each of these objectives are shown in separate sub-sections, with the exception of the development and performance assessment of the statistical models, which are shown together in the same sub-section. The last subsection for the evaluation of the identified relationships will also serve as a summary of the results.

### 5.1 DATA SELECTION AND VALIDATION

It was mentioned in the Methodology that the historical plant data that were obtained for this study, came from three different sources and had many different formats. The process of selecting, validating and pre-processing this raw data into a usable format suitable for the statistical analyses is shown here.

#### 5.1.1 Data selection

The raw data that were received for this study originally consisted of five years (more than 12 000 records) of data measured for more than 900 variables (unique sampling points and type of analyses). Upon further investigation, it was found that the majority of the data labels were either duplicates (different name used for the same sample point or type of analysis) or redundant (sample points of individual streams before mixing).

The initial selection and organising of the data therefore entailed the amalgamation of variables that are the same, but are labelled differently; and the deletion of redundant variables. The redundant variables were variables measured on individual streams coming from individual treatment units before the individual streams are mixed together to produce the outflow of the treatment process. One example of this is with the sand filters. The plant makes use of 14 gravity sand filtration units. Each of these units are monitored individually, but the outflow streams from each of the individual units are mixed together before entering the next treatment process in the treatment system.

After this initial organising and selection processes, the data were reduced to 270 variables measured over a five year period (more than 12 000 records). During the data collection period (the five year period) the WRP experienced several events, some of which were critical events, resulting in a plant

production halt. All of these events were logged in an event registry. From the event registry, a mask was created in order to remove data records that were created during a critical event (since the plant was not actually operational when those records were made).

The application of this mask and the subsequent removal of records that were created during non-operational times was considered the second selection step. The final resulting data set thus contained 270 variables measured over a five year period, but with only 1349 records. This data set was then considered the raw data from which all future analyses (validation and pre-processing included) were done.

### 5.1.2 Data validation

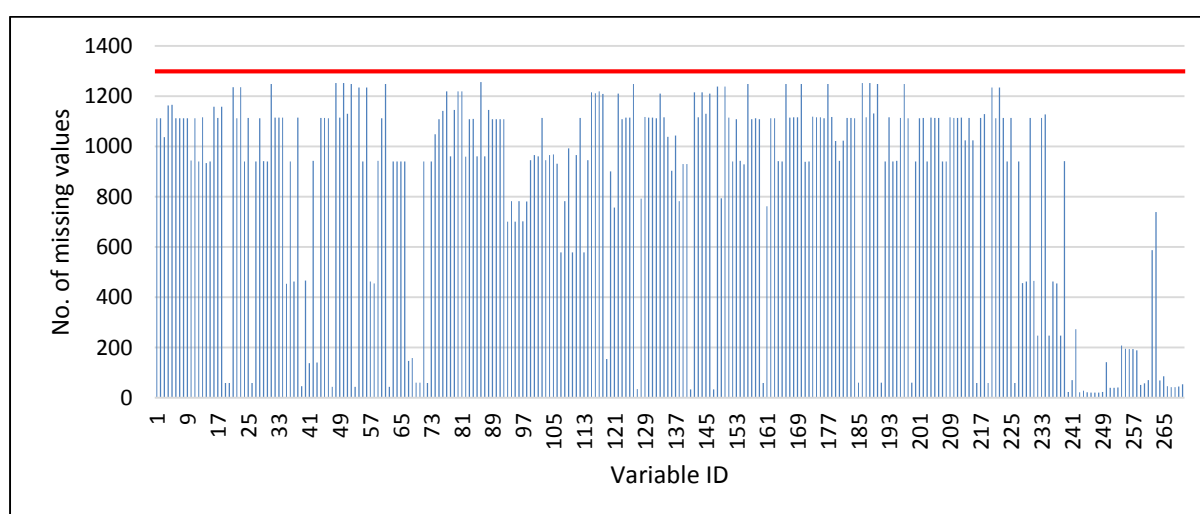
Data validation was performed on the data of this study as part of a research project by Swartz et al. (2016). This validation entailed the following:

- Missing data characterisation
- Dirty data characterisation
- Data reconciliation
- Minimum data set size requirement
- Testing for sufficient representation

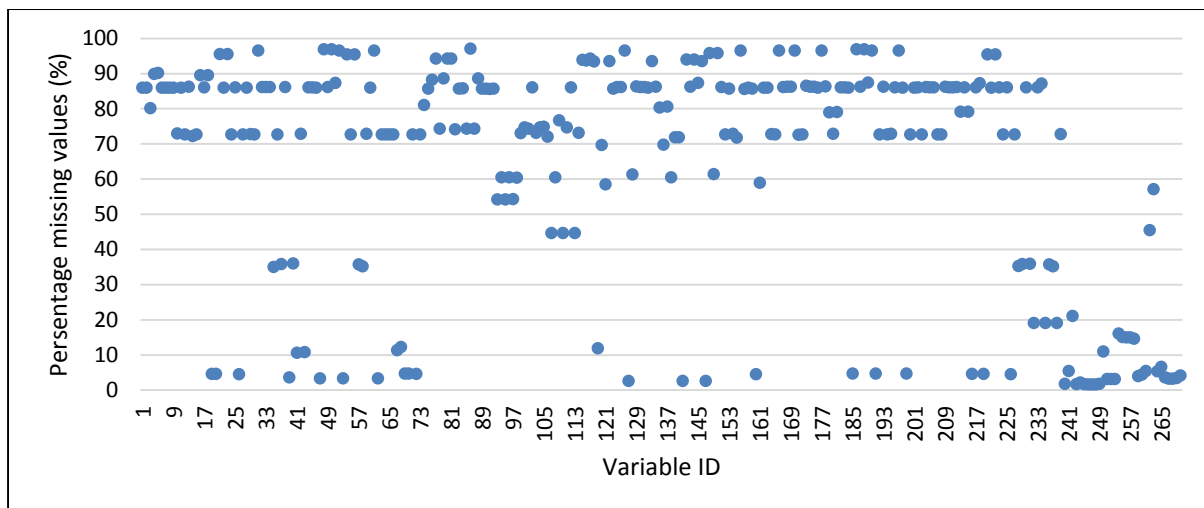
The results of that data validation will only be summarised here, since the detailed results can be found in the original publication of the work by Swartz et al. (2016).

#### Missing data characterisation

From the missing data characterisation, two graphs were generated: one showing the counts of missing data per variable, and the other showing the percentage of missing data for each of the variables. The data are illustrated in Figure 5.1 and Figure 5.2 respectively.



**Figure 5.1:** Number of missing values per variable (Swartz, et al., 2016)



**Figure 5.2:** Percentage missing values per variable (Swartz, et al., 2016)

Figure 5.1 and Figure 5.2 show that the majority of variables contain more than 50% missing values. The red line in Figure 5.1 indicates the number of records (1 349), which is the maximum number of values that could be contained in any given variable of this data set. In terms of the data set as a whole, the maximum number of data points for the whole data set is 364 230 (270 variables with 1 349 records each), however, the actual data set only contains 127 497, which means that the average percentage of missing data for the entire data record is 65%.

This should be expected if it is taken into consideration that the variables included in the data set contains both FEM variables as well as SDM variables. From variable ID 241 to 270 (all the operational variables) it can be seen that the percentage missing values are much less, since these variables are FEM variables.

This is, however, still a major concern for any multivariate statistical analyses which require a complete multivariate record. In the light of the above results, it is expected that multivariate analyses will not perform well, if at all, with the current data. Interpolation should be performed during the pre-processing of the data, although it will be unwise to perform interpolation on data containing more than 20% missing values.

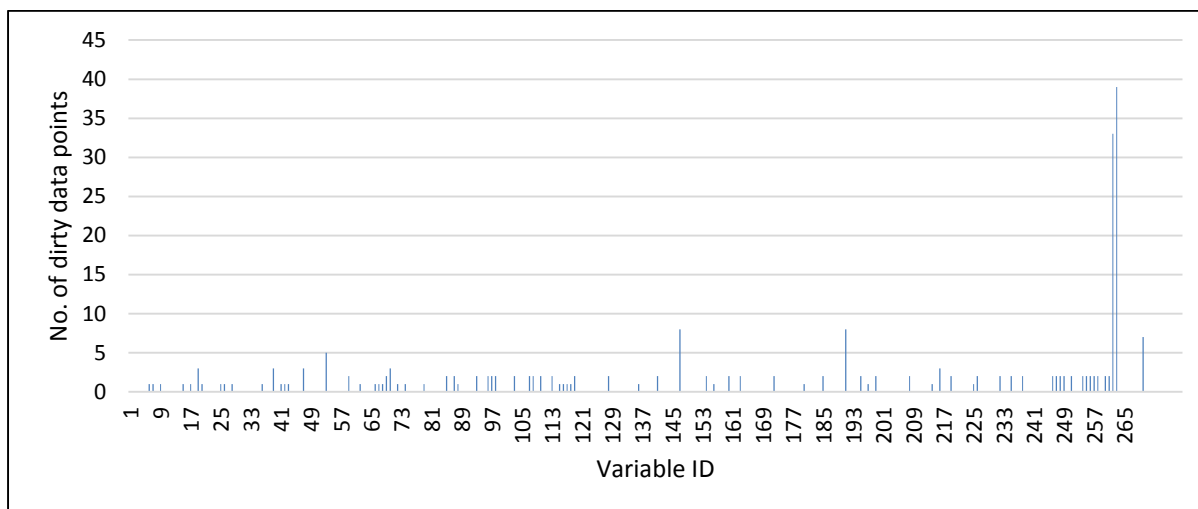
#### Dirty data characterisation

The dirty data characterisation observed two aspects of the data, the first is the number of dirty data points, which have been identified as impossible values (outside the range of logically definable values for a given variable).

#### *Extreme value count*

Figure 5.3 shows the number of dirty data points per variable. It can be seen that there is a low number of dirty data points per variable, which is a desired result. The dirty data points will be

removed during the pre-processing of the data, which means that the number of missing values will increase.



**Figure 5.3:** Number of dirty data points per variable (Swartz, *et al.*, 2016)

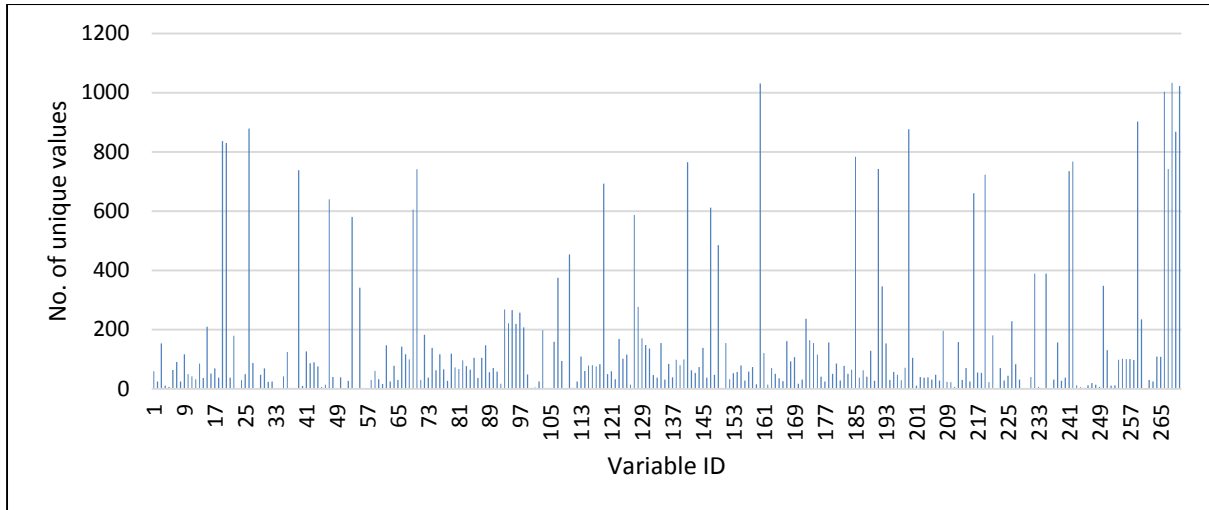
A low number of dirty data points also indicate that the monitoring equipment is operated and maintained correctly. There are two variables (ID 262 and 263) with a noticeable number of dirty data points. These are operational variables indicating the runtime (time of operation in between backwashing cycles) of the BAC and GAC respectively. The BAC and GAC filters should not be operated for more than 7 days between backwashing, the range indicating dirty data for the BAC and GAC runtimes were both [0 300] hours, which corresponds to about 12 days of continuous operation.

One explanation (also the most likely to be correct) for the high BAC and GAC runtimes are that there have been power failures, which interrupt the operation of the filters. The measured time between backwashing cycles could therefore be responsible for counting hours of down time in addition to hours of operation. This all depends on the technique used to measure the runtime of the filters. Since it is highly unlikely that these filters were continuously operational for 12 days, those values can be considered erroneous.

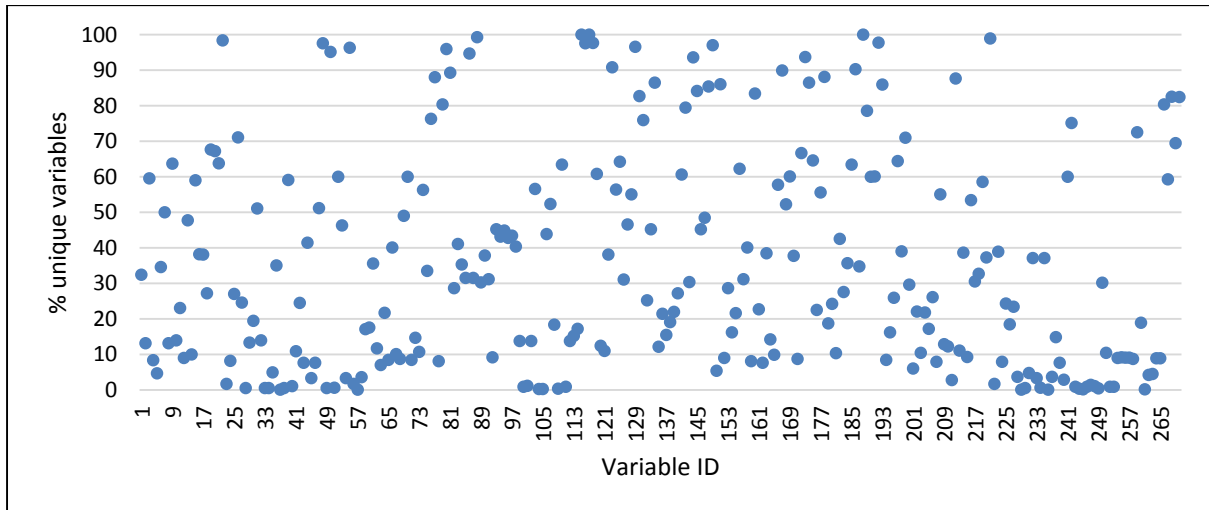
#### *Count of unique values*

The number and percentage of unique values per variable are shown in Figure 5.4 and Figure 5.5 respectively. In these figures, a high value is desired. Again, the maximum that could be achieved is 1 349, but none of the variables come close to this (Figure 5.4). This is largely due to the number of missing values.





**Figure 5.4:** Number of unique values per variable (Swartz, et al., 2016)



**Figure 5.5:** Percentage of unique values per variable (Swartz, et al., 2016)

Figure 5.5 is a much better representation of the actual state of the data. The percentage of unique values were calculated based on the number of values for each variable, and not the number of records in the data set. Based on Figure 5.5, the data contains a high amount of unique values, which again, indicates that the data is of a good analytical quality.

### Data reconciliation

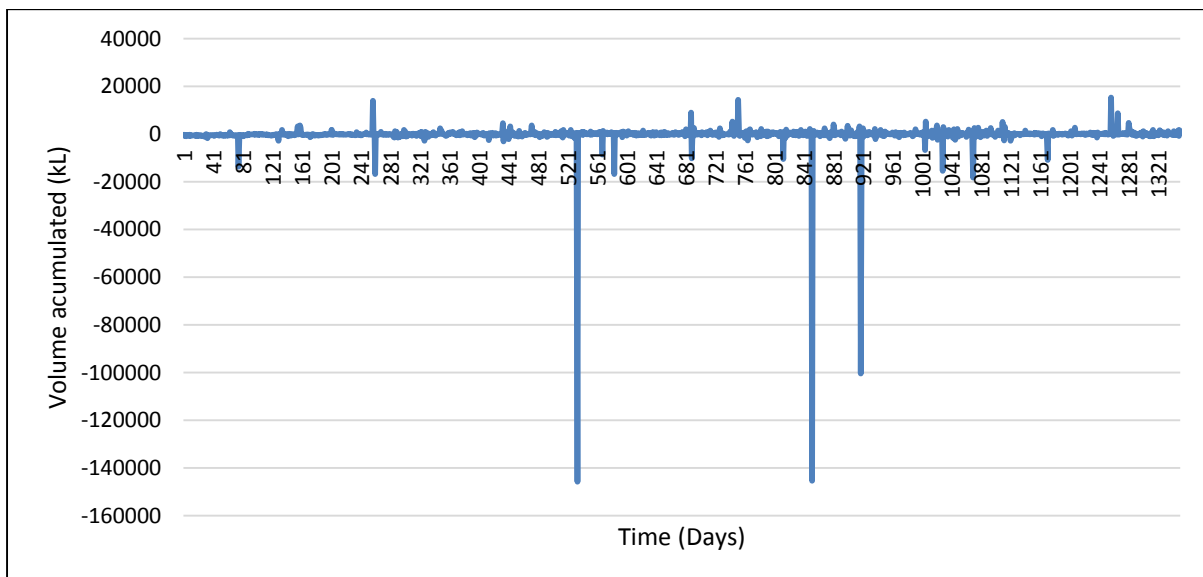
Insufficient data were available to perform data reconciliation (no flow rates of streams in between treatment units were available). From the information that was available, an overall mass balance was constructed. The mass balance made use of the total volume of water that entered the plant, minus the total volume of final water produced by the plant, minus the total volume of wastewater produced by the plant.

Figure 5.6 shows the total volume of accumulated water in the plant based on daily measurements (which should be close to zero). The calculations were performed as follows:

$$\text{Equation 48: } V_{\text{accumulated water}} = V_{\text{incoming water}} - V_{\text{final water}} - V_{\text{wastewater}}$$

The results indicate several discrepancies (Figure 5.6), where large volumes of water were either accumulated, or were unaccounted for. Unfortunately, this information cannot be used to base any sensible conclusions on, since there are several holding tanks that may or may not have stored or discharged a volume of water. These discharge and storage data could have been used to explain the small discrepancies in Figure 5.6, but were not included in the data that was made available to the project.

The three major spikes (-145 749 kL, -145 370 kL and -100 342 kL) should be regarded as erroneous (dirty data) since it is completely impossible for the plant to be able to produce or discharge such a massive volume of water.



**Figure 5.6:** Overall mass balance (accumulated volume) (Swartz, et al., 2016)

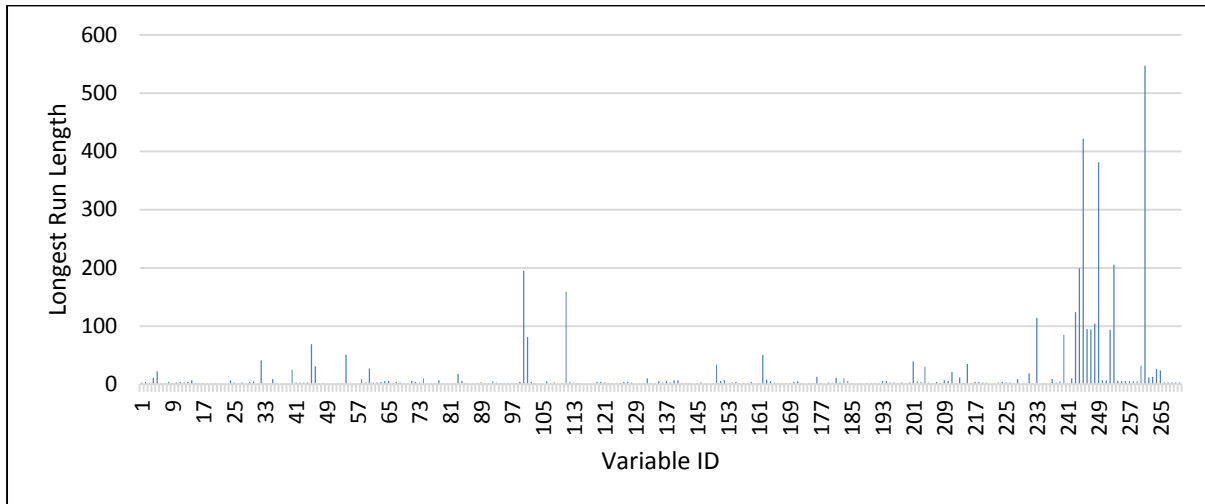
The lesser spikes in Figure 5.6 are less than the rated daily production volume of the plant (21 000 kL). These discrepancies may be cases where storage and holding tanks were filled or emptied, which cannot be accounted for with the current data available to the project. Ultimately the mass balance provides little information and adds little value with regard to data reconciliation.

#### Testing for sufficient representation

Several tests were performed in order to determine whether or not the data provides a sufficient representation of the expected plant behaviour (variability, operational conditions and performance).

*Run length test*

The first test makes use of a run length plot. This plot (Figure 5.7) indicates the longest number of data points that had a constant value for each of the variables in the data set. In this test, long run lengths are undesirable since it indicates either a fault in the measurement process, or very stable and constant plant conditions.



**Figure 5.7:** Run length plot for all variables in the data (Swartz, *et al.*, 2016)

From Figure 5.7 it can be seen that the majority of the variables have small run lengths. Several operational variables (ID 241 to 270) have run lengths longer than 100, which is slightly less than 10% of the data contained in the variable (Figure 5.7). This may indicate stable operating conditions, but since the remainder of the variables does not reflect this, it is more likely that the measurement and data acquisition of the operational variables are non-ideal.

The long run lengths of the operational variables may be explained by the source of the data. The operational data that were made available to the study, consisted of a schedule that is completed by the first shift (08h00) process controllers every day. In the schedule, the process controller writes down several values that are currently displayed on the SCADA of the plant.

The operational data were therefore not live data, but rather a snap shot of the operational conditions that were taken at the same time every day. It can therefore be expected that several of the operational variables are constant, since this may be the values stipulated by the start-up procedure of the plant. What it comes down to is the following: the operational variables may have changed during the day and night, but the data that was captured only represents the state of the plant at roughly 08h00 in the morning.

Other tests for sufficient representation were also performed on the data, these tests consisted of moving average plots, standard error plots and CUSUM plots for key variables of each of the treatment units in the plant. A summary of these results can be seen in Appendix B, but for now the

only relevant conclusion is that in all three cases, the data showed a sufficient degree of variability to be accepted as a good representation of the actual plant conditions.

## 5.2 DATA PRE-PROCESSING

The data pre-processing consisted of three steps, done in the following order:

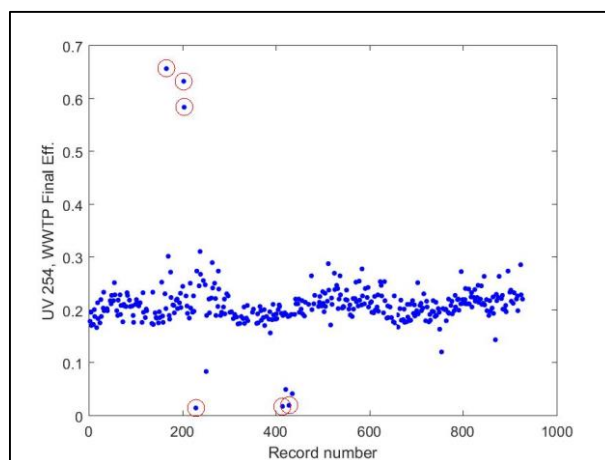
- 1) Outlier analysis (detection and removal)
- 2) Noise attenuation
- 3) Missing data replacement

The results for each of these steps will be shown and discussed in this section of the report. It should be noted, however, that for each of these analyses approximately 270 plots were created. These plots can be provided by the author on request, but it is too impractical to show them here or in any appendix.

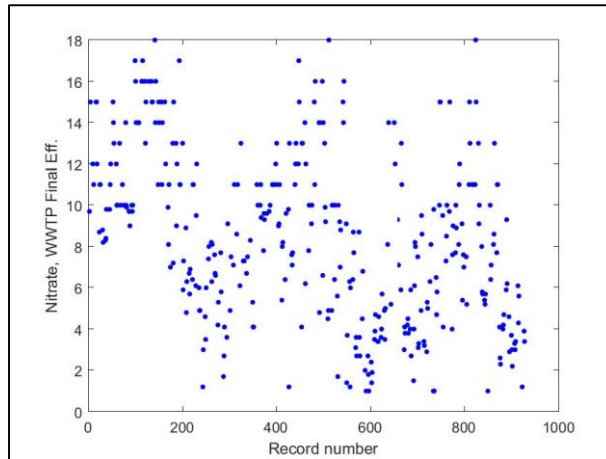
### 5.2.1 Outlier analysis

It was mentioned in the methodology that each variable was first classified into a statistical distribution in order to determine the upper and lower limits by which outliers should be detected. The variables with normal and lognormal distributions were analysed according to the statistical estimates obtained from the distributions, but for those variables that could not be categorised into any distribution, an empirical approach was used. The upper and lower boundaries were set by the 99<sup>th</sup> and 1<sup>st</sup> percentiles respectively. After removing the outliers using the upper and lower limits, the data were further analysed using a Hampel filter.

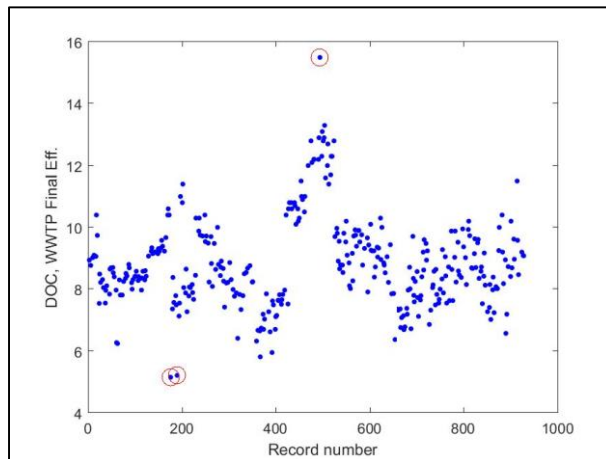
Three examples were chosen to illustrate the outlier removal performance. The three examples all contain approximately the same number of data points and include a case for each distribution type: empirical (Figure 5.8), normal (Figure 5.9) and lognormal (Figure 5.10). At this point the Hampel filter has not been applied yet.



**Figure 5.8:** Outlier results for UV<sub>254</sub> in the WWTP Final Effluent (empirical)



**Figure 5.9:** Outlier results for Nitrate in the WWTP Final Effluent (normal)

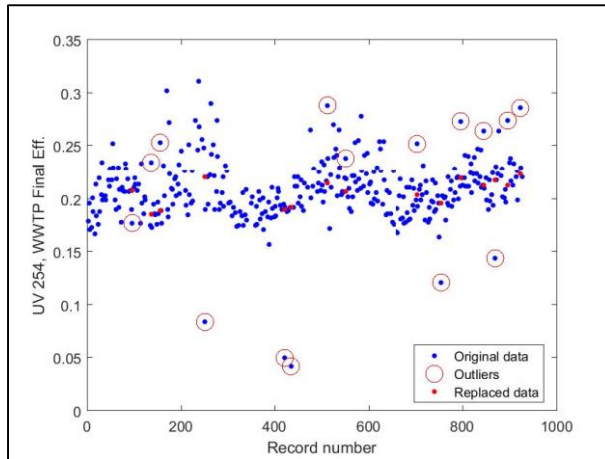


**Figure 5.10:** Outlier results for DOC in the WWTP Final Effluent (lognormal)

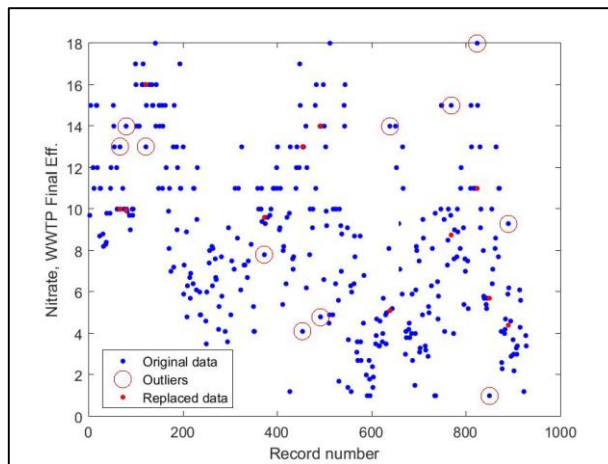
Applying the Hampel filter (Figure 5.11, Figure 5.12 and Figure 5.13) was necessary to remove outliers that were undetected by the three-sigma outlier analysis. The improved performance of the Hampel filter can be ascribed to two main features that is lacking from the three-sigma analysis:

- 1) Resistance to outliers
- 2) Localised analysis as opposed to global analysis

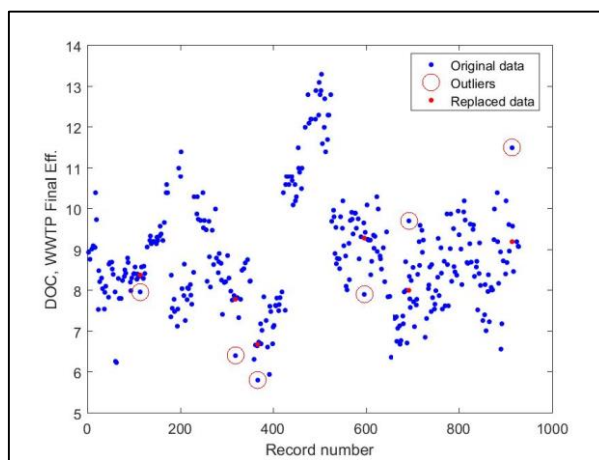
The fact that the Hampel filter is resistant to outliers is important since a variable containing many outliers (Figure 5.8) will overestimate the boundaries used for removing the outliers. Secondly, since the Hampel filter makes use of a moving window, any trends within the variable will be taken into consideration and will not be able to mask outliers (Figure 5.9).



**Figure 5.11:** Outlier results with Hampel filter for UV<sub>254</sub> in the WWTP Final Effluent (empirical)

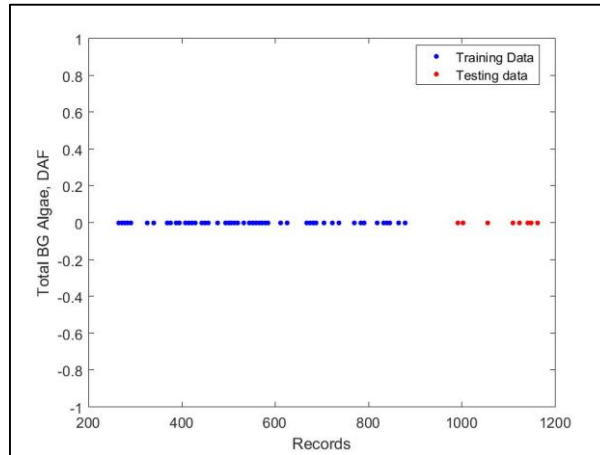


**Figure 5.12:** Outlier results with Hampel filter for Nitrate in the WWTP Final Effluent (normal)



**Figure 5.13:** Outlier results with Hampel filter for DOC in the WWTP Final Effluent (lognormal)

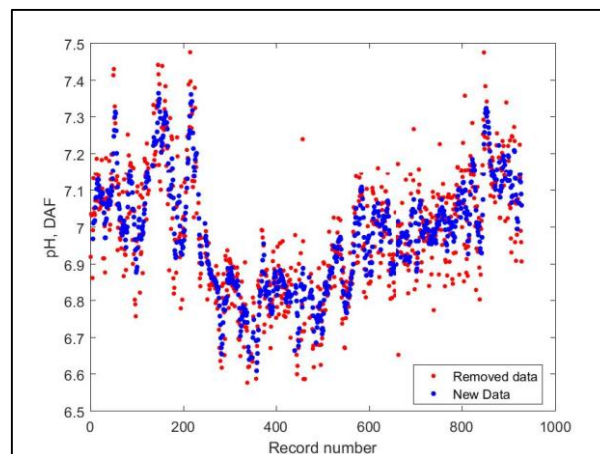
The outlier analysis algorithm was also used to remove variables with very small variances (Figure 5.14). Of the original 270 variables, only 238 variables remained.



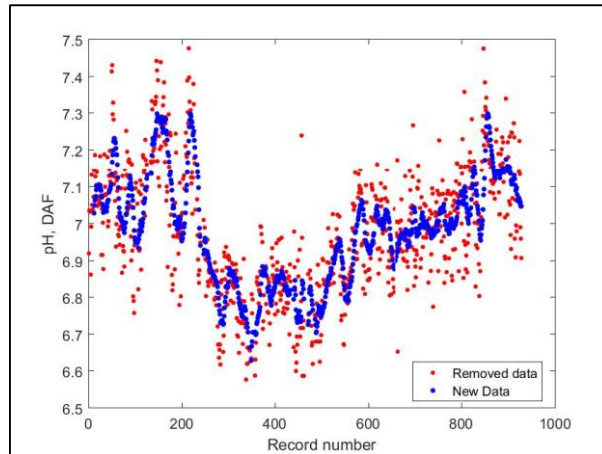
**Figure 5.14:** Variable removed due to lack of variance

### 5.2.2 Noise attenuation

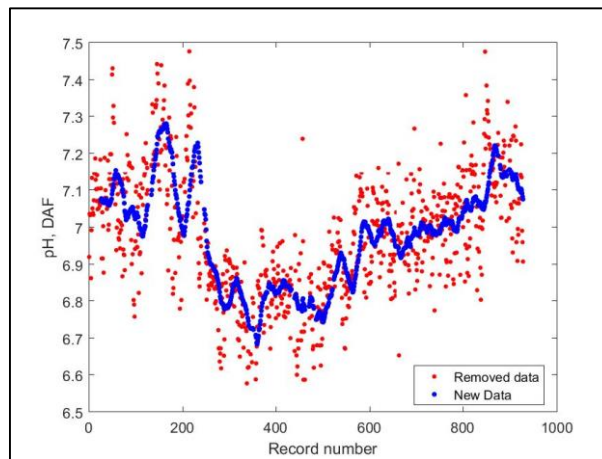
A moving window algorithm was used to remove and replace noisy data points. The appropriate size of the moving window was difficult to determine since its impact is heavily dependent on the frequency of the data. Three noise filters were tested on all of the variables using a window size of 5, 11 and 25. By comparing the results of the noise filters on high frequency variables (Figure 5.15, Figure 5.16 and Figure 5.17) to the results of the noise filters on low frequency variables (Figure 5.18, Figure 5.19 and Figure 5.20) it was possible to select the most appropriate window size.



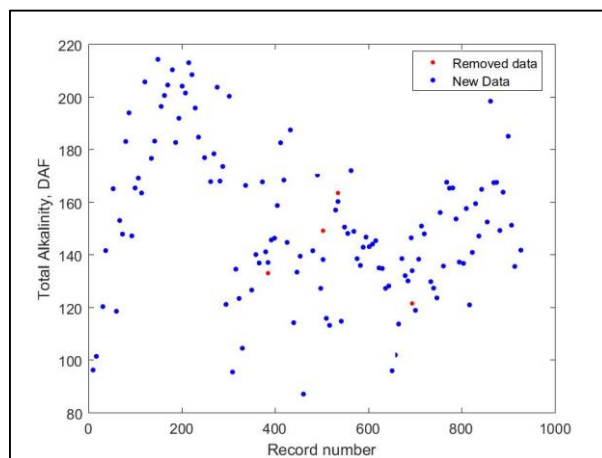
**Figure 5.15:** Noise reduction results for pH from the DAF (window size = 5)



**Figure 5.16:** Noise reduction results for pH from the DAF (window size = 11)

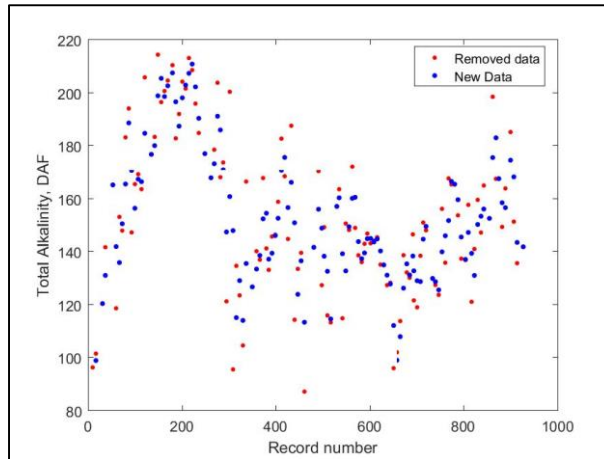


**Figure 5.17:** Noise reduction results for pH from the DAF (window size = 25)

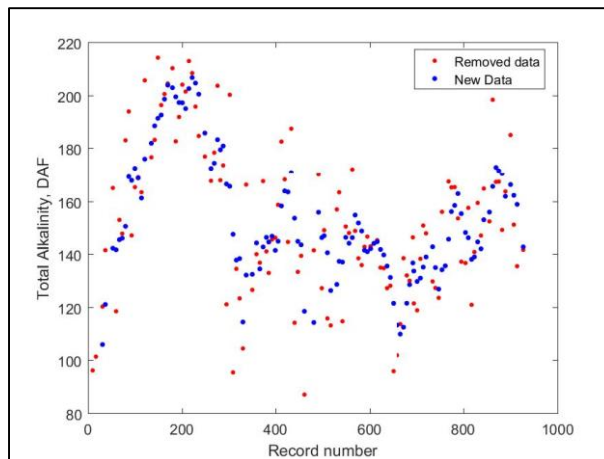


**Figure 5.18:** Noise reduction results for Total Alkalinity from the DAF (window size = 5)





**Figure 5.19:** Noise reduction results for Total Alkalinity from the DAF effluent (window size = 11)



**Figure 5.20:** Noise reduction results for Total Alkalinity from the DAF effluent (window size = 25)

At a window size of 5 it was found that the high frequency variables still contained some noise (Figure 5.15), whilst the low frequency variables showed hardly any improvement (Figure 5.18). With a window size of 25, on the other hand, it was found that the high frequency variables contained an unrealistically small amount of noise (Figure 5.17), whilst the low frequency variables still contained a realistic amount of noise (Figure 5.20). Quantifying a reasonable amount of noise is somewhat subjective, but in this case the most appropriate filter was chosen if it removed the fast time variations without affecting the slow time variations in the data.

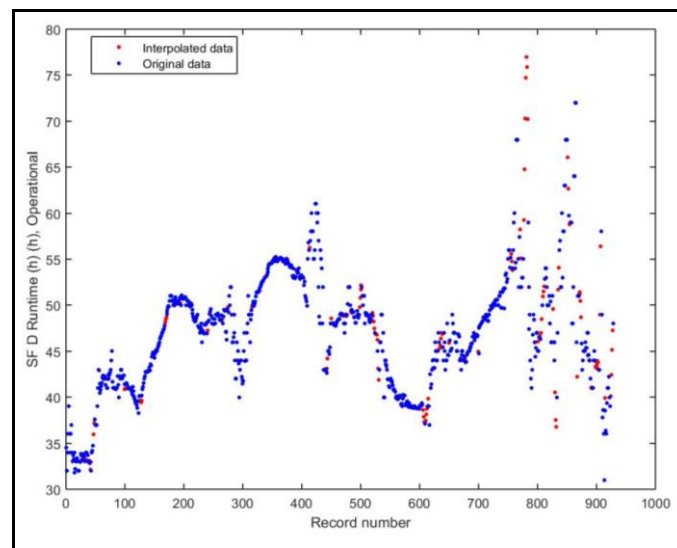
Ultimately the noise filter with a window size of 11 was chosen as the most appropriate noise filter. This resulted in high frequency variables still containing a realistic amount of noise (Figure 5.16), whilst the low frequency variables remained slightly noisy (Figure 5.19). Despite the relatively large amount of noise that remained in the low frequency variables, the data produced using a window size of 11 was

used for all of the future analyses. This is in part also because it is more sensible to average a value using the average over a two week period, rather than the average over an entire month.

### 5.2.3 Missing data replacement

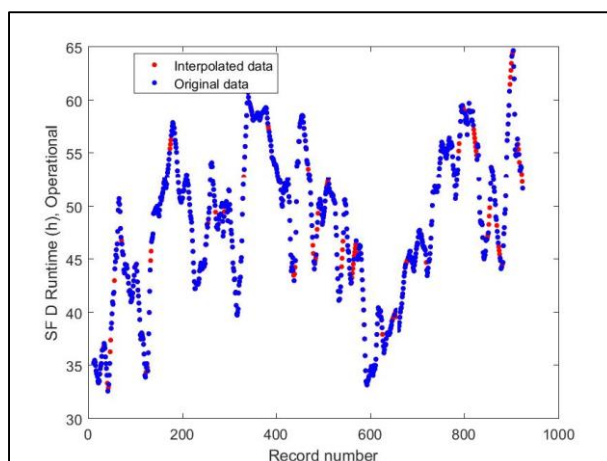
At this stage, the training data set had already been separated into two new data sets. The one contained the predictor variables (FEM) and is referred to as the  $X$  matrix, or the MVR. The other data set contains the response variables and is referred to as the  $Y$  matrix.

The variables contained in the MVR were the only variables suitable for any form of interpolation, since they contained less than 20% missing values. Linear interpolation was applied to the MVR in order to replace any missing values. Spline interpolation was also attempted, but there were cases of overfitting (Figure 5.21), and it was therefore discarded from the study.



**Figure 5.21:** SF D runtime with spline interpolation showing overfitting

The linear interpolation provided good results and was capable of removing most of the missing values from the MVR (Figure 5.22). Since linear interpolation cannot extrapolate (in cases where the missing values were at the start or end of the data record), some missing values still remained in the MVR.



**Figure 5.22:** SF D runtime with linear interpolation

### 5.3 DEVELOPMENT AND PERFORMANCE OF THE STATISTICAL MODELS

The results of the analyses performed on the pre-treated plant data are shown and discussed in this section. The results have been separated into two subsections, one for the bivariate analyses and one for the multivariate analyses. In all cases, only a portion of the results are shown since it would require too many pages to show all the results and would not be an effective method of communicating the results.

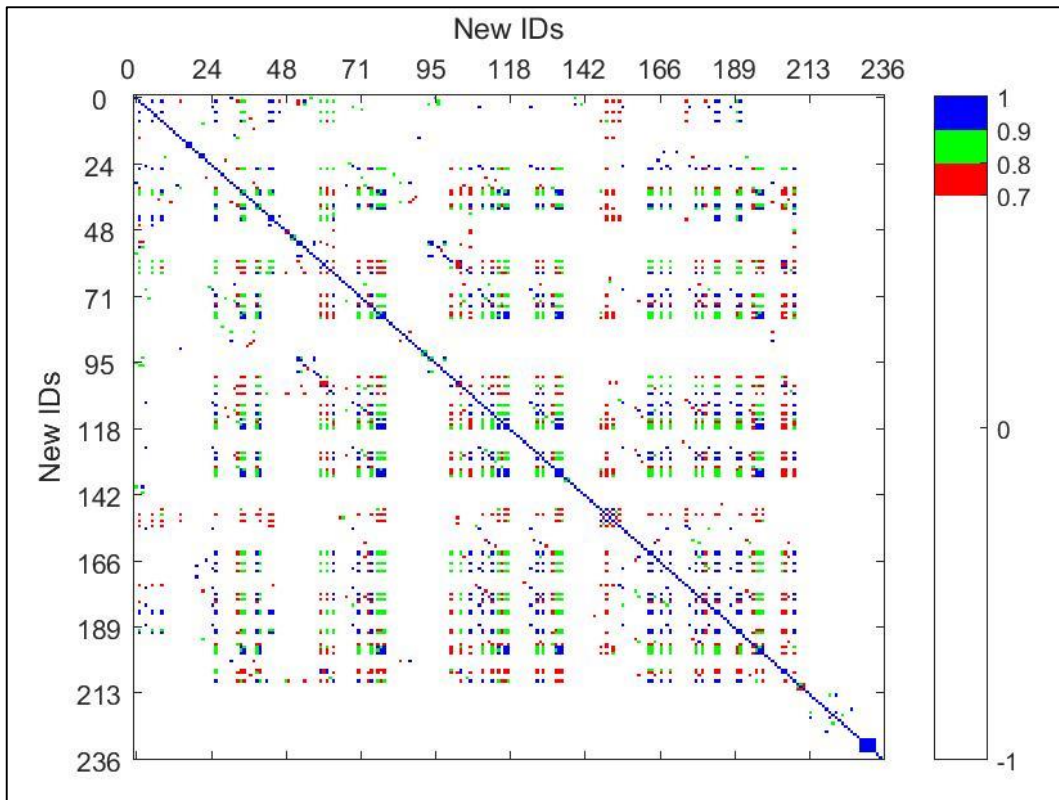
#### 5.3.1 Bivariate analyses

The bivariate analyses consisted of two types of correlation calculations: the first for Pearson's correlation coefficient and the second for Spearman's correlation coefficient. Correlations were performed on the complete training data set before separating the FEM variables from the SDM variables, as well as after. The correlation results were used to form correlation matrices which were then coloured in order to illustrate the results graphically. It should be noted that correlations that did not lead to a rejection of the null hypothesis ( $p$ -value greater than 0.05) were removed from the results prior to creating the plots.

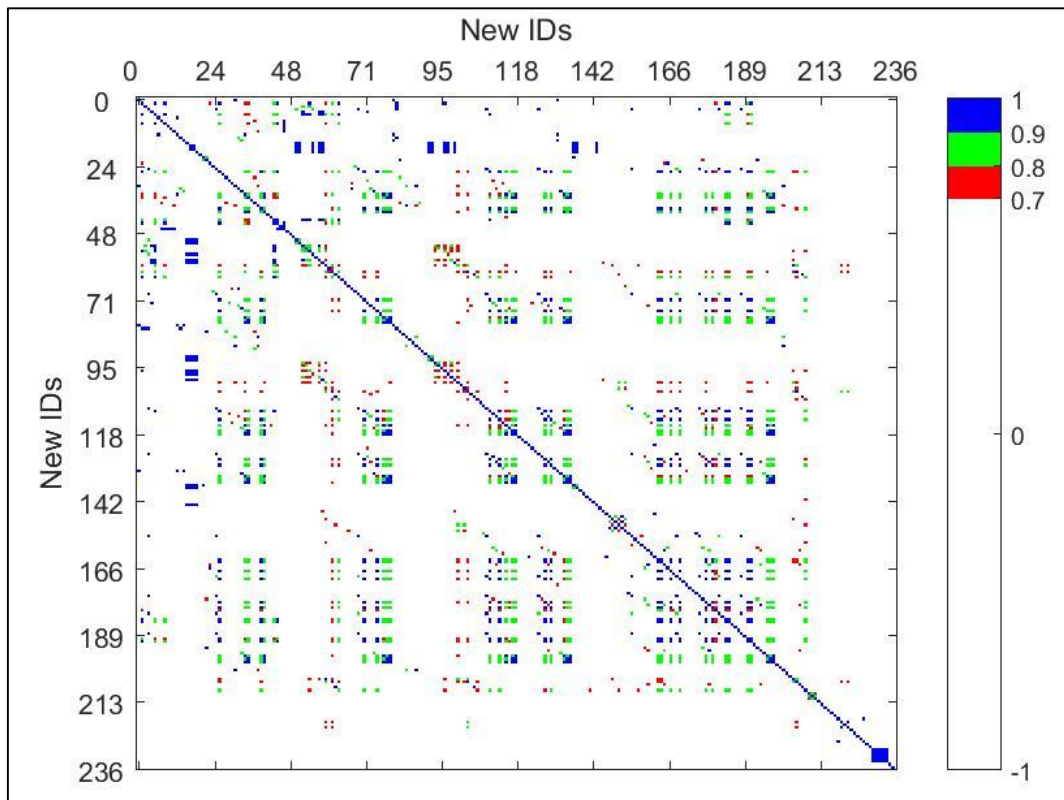
##### Before separating MVR

The results from before the separation indicate that several correlations above 0.9, using both the Pearson (Figure 5.23) and Spearman (Figure 5.24) correlation coefficient, were detected. The colours red, green and blue seen in the plots represent the magnitude of the correlation, greater than 0.7, 0.8 and 0.9, respectively.

Since the data were contained in a single matrix, the correlation matrix is symmetrical with the diagonal showing values of 1. This is normal since the values on the diagonal represent correlations between the same variable. In some cases, there is a missing value on the diagonal - this was due to variables that did not have sufficient variability to perform a correlation test on.



**Figure 5.23:** Pearson's correlation coefficient before separating the MVR



**Figure 5.24:** Spearman's correlation coefficient before separating the MVR

Of all the Pearson correlation coefficients that were calculated, 1 598 coefficients were equal or greater than 0.7 as well as having a p-value less than 0.05 (Figure 5.23). The number of Spearman correlation coefficients, that were equal or greater than 0.7 as well as having a p-value less than 0.05, were 1 294 (Figure 5.24).

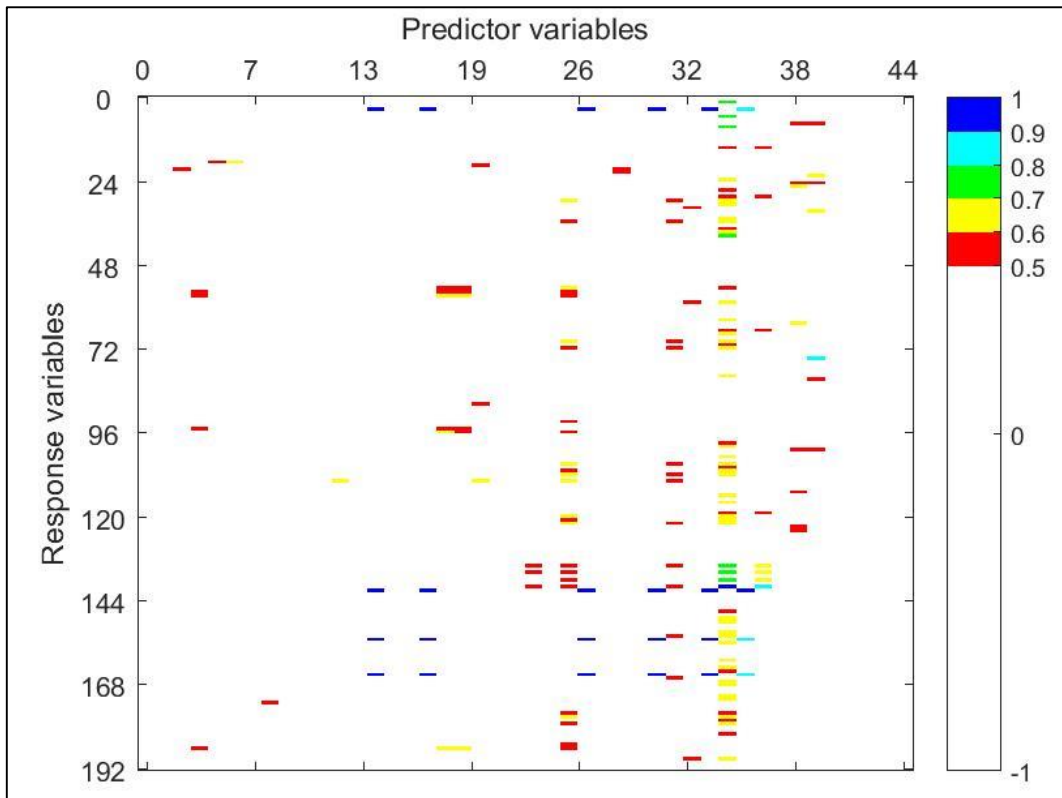
In both cases, the number of significant correlations (above 0.7 with a p-value less than 0.05) were much less than the total number of correlations that were calculated, which was 55 696 correlations. Considering that the 44 FEM variables are also included in the correlations (which could produce 1 936 correlations) the counts of significant correlations of 1 598 and 1 294 for the Pearson and Spearman correlations, respectively, are low.

Since both these counts are below 1 936, it is possible that all of the correlations are in fact between FEM variables. For this reason, the correlation analyses were repeated using the separated predictor and response variables.

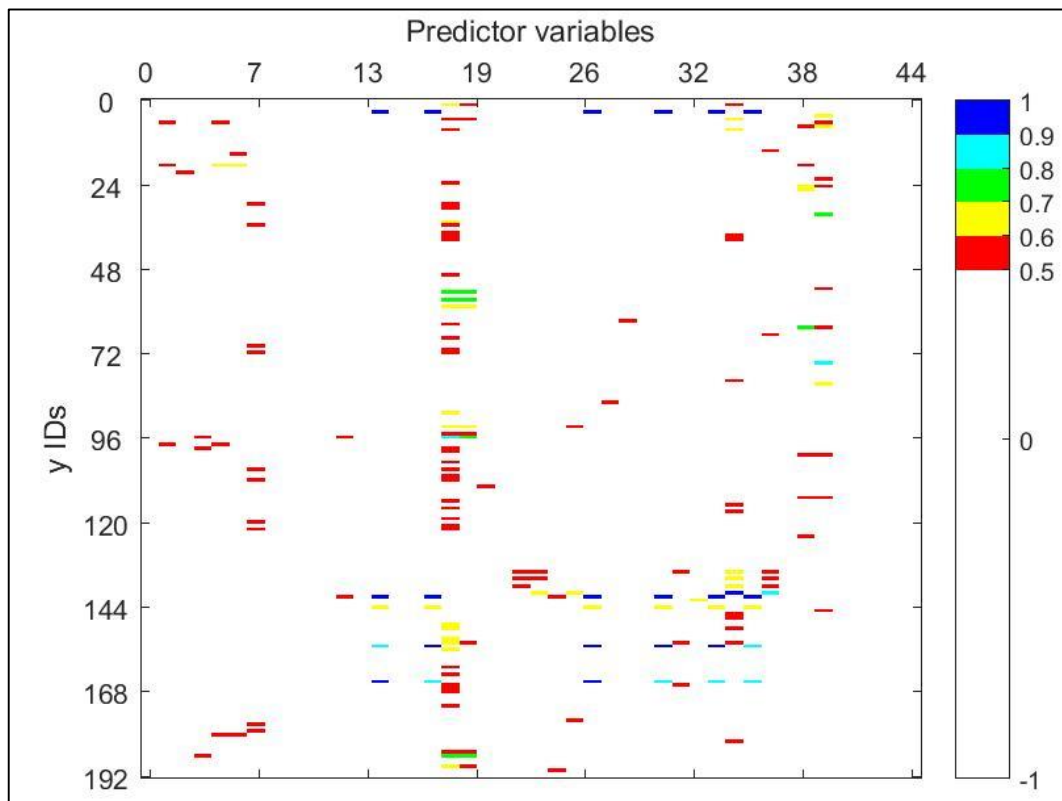
After careful inspection of the results, it was found that the majority of the correlations were in fact trivial (between FEM variables). These correlations were removed in order to create a list of correlations that were significant as well as relevant to the study. These correlations that are of interest (Table 5.3 and Table 5.4) will be further discussed when applying expert process knowledge.

#### After separating the MVR

The results from after the separation indicate that few correlations were above 0.9 for both the Pearson (Figure 5.25) and Spearman (Figure 5.26) correlation coefficient. Since there were so little correlations of significance, two colours were added to the plot. The colours red, yellow, green, cyan and blue seen in the plots represent the magnitude of the correlation greater than 0.5, 0.6, 0.7, 0.8 and 0.9, respectively.



**Figure 5.25:** Pearson's correlation coefficient after separating the MVR



**Figure 5.26:** Spearman's correlation coefficient after separating the MVR

Unlike the previous set of correlation matrices, the correlation matrices that were generated after separating the MVR are not symmetrical. There is therefore not a series of perfect correlations on the diagonal of these matrices and the x and y axis of these matrices are not the same. The x axis indicates the predictor variables and the y axis indicates the response variables.

Of all the Pearson correlation coefficients that were calculated, 34 coefficients were significant (equal or greater than 0.7 as well as having a p-value less than 0.05) (Figure 5.25). The number of Spearman correlation coefficients, that were significant were 37 (Figure 5.26).

In both cases, the number of significant correlations were much less than the total number of correlations that were calculated, which was 8 448 correlations. Since there were so few significant correlations, all of the significant correlations were tabulated for both the Pearson and Spearman correlations (Appendix C).

Upon further investigation of these results, it was found that no correlations of interest can be observed in either the Pearson or Spearman correlation results. This was unexpected, since the results from before the separation of the MVR showed several correlations of interest. The fact that the MVR was removed from the data for the second round of correlation tests affected the results in two ways:

- 1) There are no FEM against FEM correlations.
- 2) There are no SDM against SDM correlations.

This means that the only possible correlations to observe in the second round of correlations should be between FEM and SDM variables only. Although this may seem obvious, the results indicate otherwise. The main reason for this discrepancy is the nature of the pseudo-SDM variables. The results of the second correlation test seem trivial, but are in fact correlations between various pairs of pseudo-SDM variables.

This brings into question the approach of grouping the pseudo-SDM variables with the SDM variables, rather than the FEM variables. The ultimate finding of the second round of correlation tests are of no interest with regards to the aim of this study, but it does highlight something to take note of.

The decision to group variables based on the number of data points was effective in identifying FEM variables. It also aided in the data analyses since the selected FEM variables were all relatively complete (containing less than 20% missing values). Unfortunately, the disadvantage of this decision was the misclassification of poorly measured FEM variables as SDM variables (called pseudo-SDM variables).

Despite this misclassification, the analysis should still have been able to identify correlations between FEM and SDM variables, should there exist any correlation. The fact that no correlations of interest were observed in the second round of correlation tests should therefore not be considered the result of the misclassification alone.

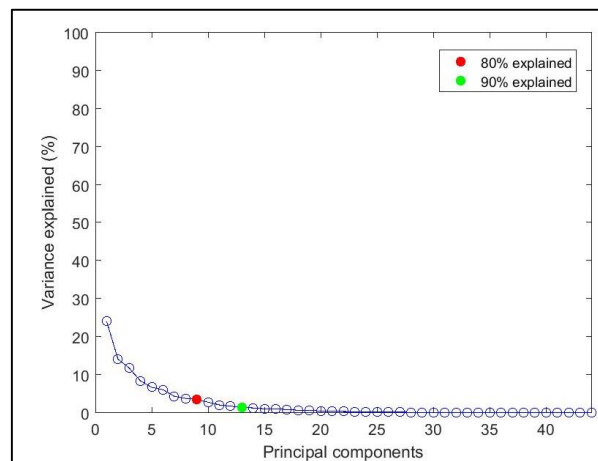
### 5.3.2 Multivariate analyses

The multivariate analyses performed during this study consisted of PCA, LDA and PLS regression. The results for each of these analyses will be shown and discussed here. Where applicable, the results from training the models will be discussed first.

#### Principal Component Analysis results

PCA is a visualisation technique used to reduce the dimensionality of data with the aim of visualising highly dimensional data. PCA was only performed on the MVR, due to the high number of missing values in the **Y** variable.

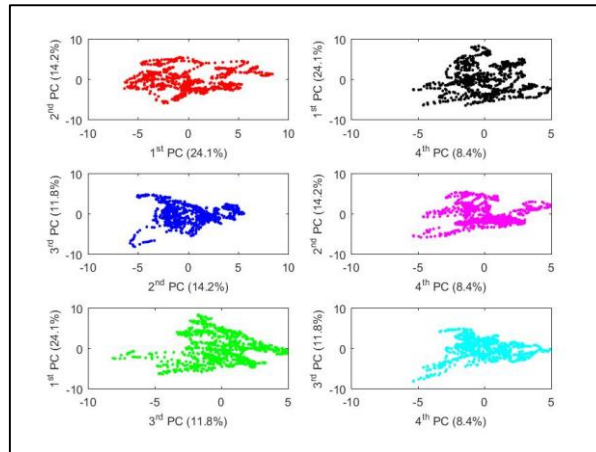
The extent to which the dimensionality of the data can be reduced was tested using a plot, which indicated the amount of variance explained by each of the PCs (Figure 5.27). In this case, it was found that 80% and 90% of the variance in the MVR could be explained by 9 and 13 PCs, respectively.



**Figure 5.27:** Variance explained by principal components

By plotting the different PCs against each other in a scatter plot (Figure 5.28) it is possible to determine visually whether the data in any of the scatter plots form groups or clusters.





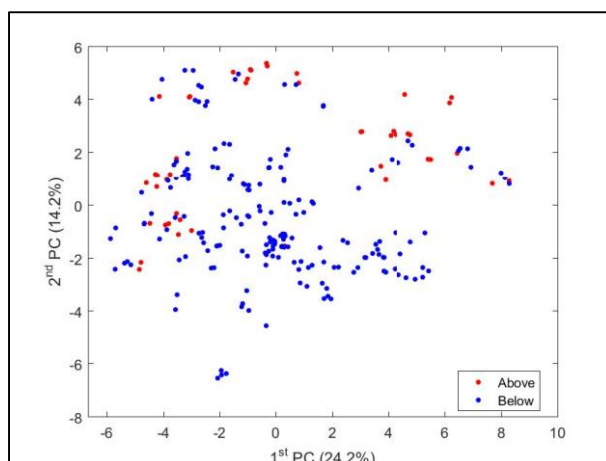
**Figure 5.28:** Multi-plot of scatter plots for the first four principal components

These clusters may point to specific events or operational characteristics, which may then be exploited in order to determine whether or not significant changes in the operational procedures of the plant have taken place, or whether or not a significant change in certain water quality variables have taken place.

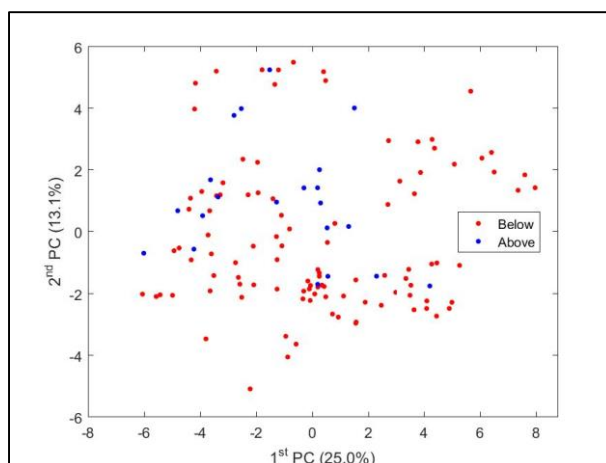
In order to further utilise the scatter plots of the different PCs, it is possible to make use of a colour map to colour the data points in the plots according to their numeric values. Since there is a large variety of possible numeric values for any given data point, it is more practical to rather make use of a colour map that colours the different data plots according to a category or group classifier.

Since LDA was performed, it was possible to colour the data points according to the classes that were used during LDA. Multi-plots of scatter plots for the first four PCs were coloured according to the LDA groups of the data. These plots were made after removing the records from the MVR that corresponded to missing values in the respective *c* categorical variable of each of the response variables. The scatter plots were also graphed individually since the multi-plot reduced the clarity of the plots significantly. The percentage of variance explained by each PC is placed in parentheses next to the number of the PC on the axes of the plots.

For random data it is expected that the data in the scatter plots, corresponding to different classes (different colours), will be evenly distributed. This was the case for several variables (Figure 5.29 and Figure 5.30), indicating that there is probably no relationship between the MVR and the classification of said variable.

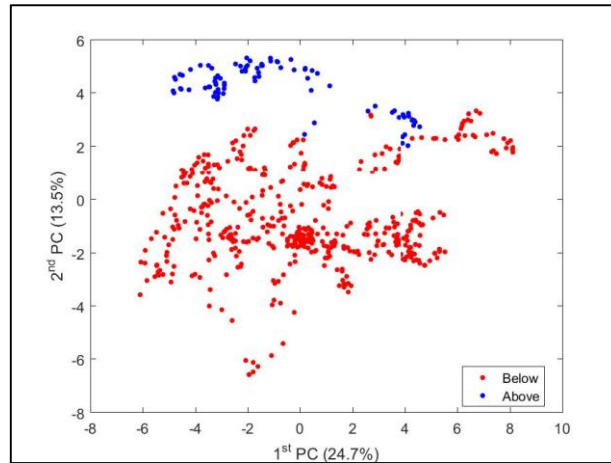


**Figure 5.29:** Scatter plot of 1<sup>st</sup> and 2<sup>nd</sup> PC coloured for Nitrate from the WWTP Clarifier



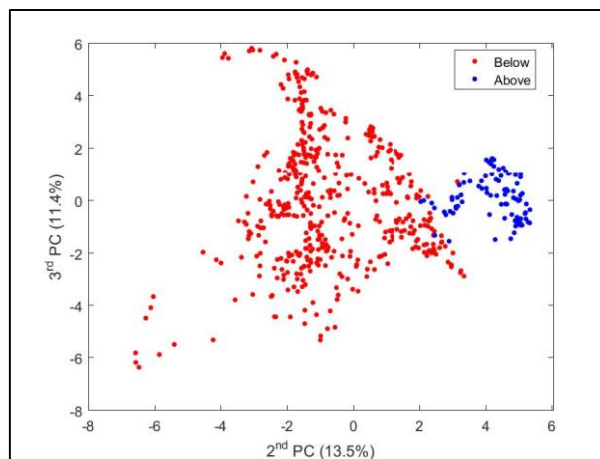
**Figure 5.30:** Scatter plot of 1<sup>st</sup> and 2<sup>nd</sup> PC coloured for Nitrite from the WWTP Clarifier

In cases where the data corresponding to different classes (different colours), are not evenly distributed but clustered together or lying within the same characteristic feature (Figure 5.31), it may be more probable that a relationship exists between the MVR and the classification of corresponding  $y$  variable.

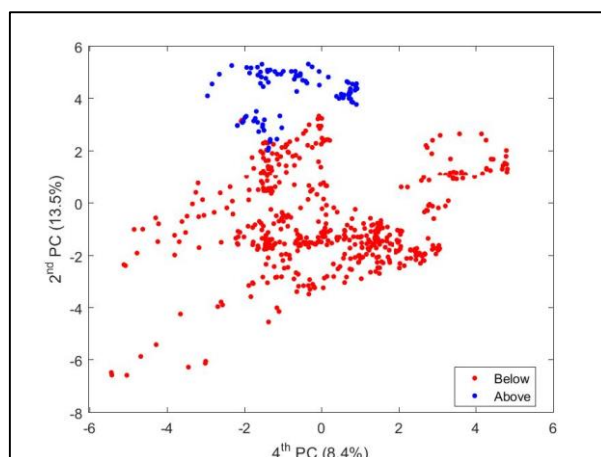


**Figure 5.31:** Scatter plot of 1<sup>st</sup> and 2<sup>nd</sup> PC coloured for EC from the UF

It can therefore be expected that the classes of these variables may, although it is not guaranteed, be more accurately predicted by the LDA models. It was also found that when clustering occurred, the clusters were most noticeable on the plots for the 1<sup>st</sup> and 2<sup>nd</sup> PCs, 2<sup>nd</sup> and 3<sup>rd</sup> PCs and 2<sup>nd</sup> and 4<sup>th</sup> PCs (Figure 5.31, Figure 5.32 and Figure 5.33).

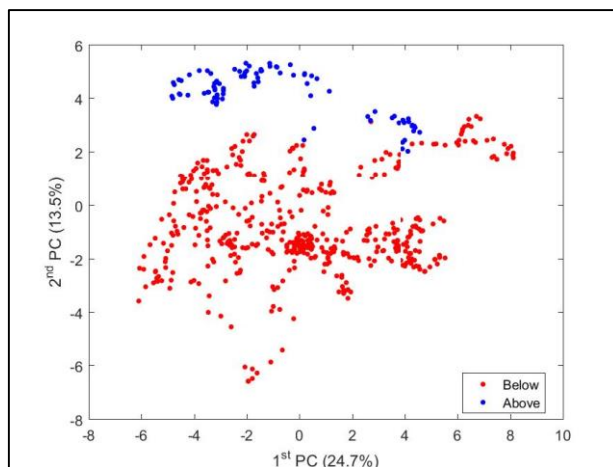


**Figure 5.32:** Scatter plot of 3<sup>rd</sup> and 2<sup>nd</sup> PC coloured for EC from the UF

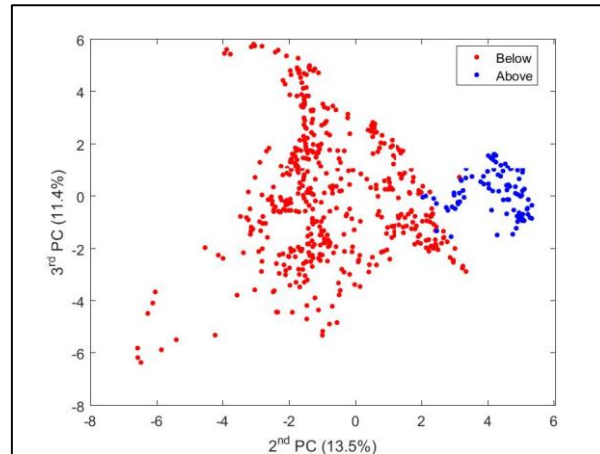


**Figure 5.33:** Scatter plot of 4<sup>th</sup> and 2<sup>nd</sup> PC coloured for EC from the UF

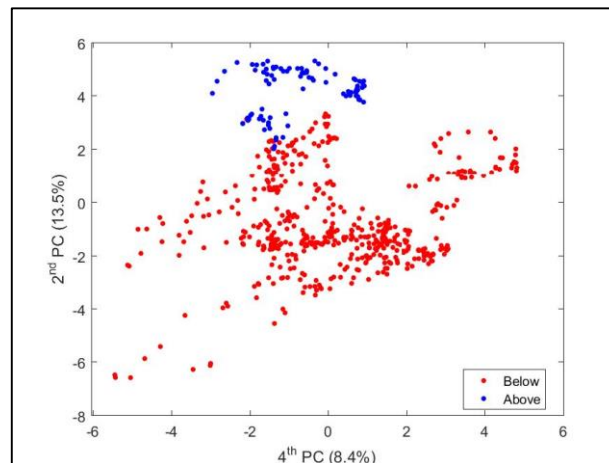
It was also found that variables with a high correlation (expressed by the Pearson's correlation coefficient) also had very similar PC scatter plots. TDS (Calc) is a variable used to express TDS, but in fact it is calculated from EC, thus it has a perfect correlation (except when there is data corruption). The PC plots for the TDS (Calc) variable showed the same distribution of classification than the EC (Figure 5.34, Figure 5.35 and Figure 5.36).



**Figure 5.34:** Scatter plot of 1<sup>st</sup> and 2<sup>nd</sup> PC coloured for TDS (Calc) from the UF



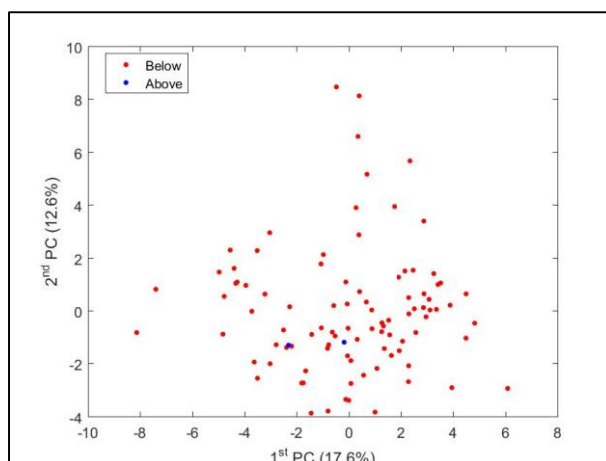
**Figure 5.35:** Scatter plot of 3<sup>rd</sup> and 2<sup>nd</sup> PC coloured for TDS (Calc) from the UF



**Figure 5.36:** Scatter plot of 4<sup>th</sup> and 2<sup>nd</sup> PC coloured for TDS (Calc) from the UF

This same procedure was repeated on the testing data, but did not provide any useful information. This is primarily because of two reasons:

- The testing data were not pre-processed (contains outliers, noise and missing values)
- The testing data consists of much fewer data records



**Figure 5.37:** Scatter plot of 1<sup>st</sup> and 2<sup>nd</sup> PC coloured for TDS (Calc) from the UF (testing data)

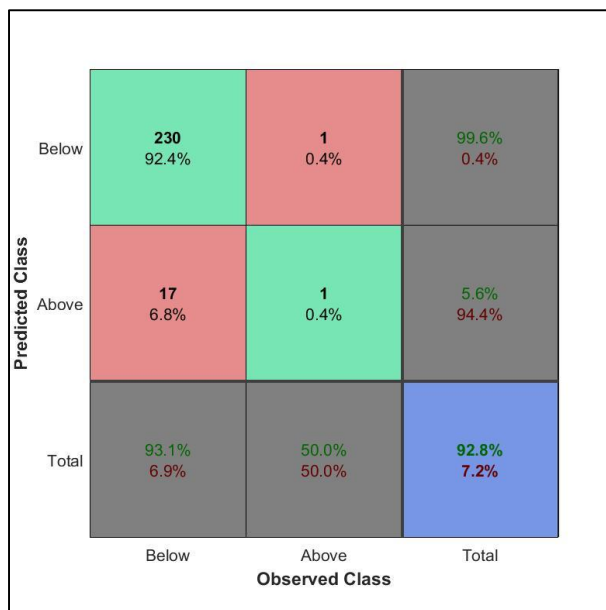
Since the PCA functioned poorly on the testing data, it was expected that the LDA would also suffer from these problems.

### Linear Discriminant Analysis

LDA creates a linear combination, using the variables in the MVR, in order to predict the class to which the values in the response variables will fall (see section 4.2.2 for the classification of the response variables). The LDA models were built on the training data and then applied to the testing data. Confusion matrices were created for the results of each of the LDA models in order to view the accuracy of the predictions made by the models.

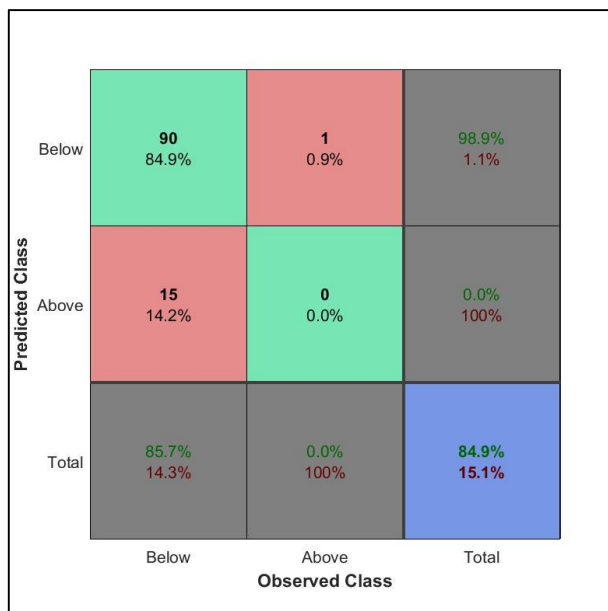
During the discussion of the PCA results, several variables showed clear separations between the different LDA classes and it was expected that this will reflect in the performance of the LDA models, but this was not always the case. In most cases, the LDA models performed better than expected (after reviewing the PCA results for those variables).

The PCA results for EC in the UF effluent (Figure 5.32) indicated that the majority of the data points within the 'Above' class were located in a cluster that was also separated from the data that lay within the 'Below' class. It was therefore expected that the LDA model for the EC in the UF effluent will perform well. Confusion matrices were created of the LDA models in order to view the classification precision of the models from these plots it was easy to determine which LDA models functioned well. As expected, the LDA model for the EC in the UF effluent performed well, with a classification error of only 7.2% (Figure 5.38).

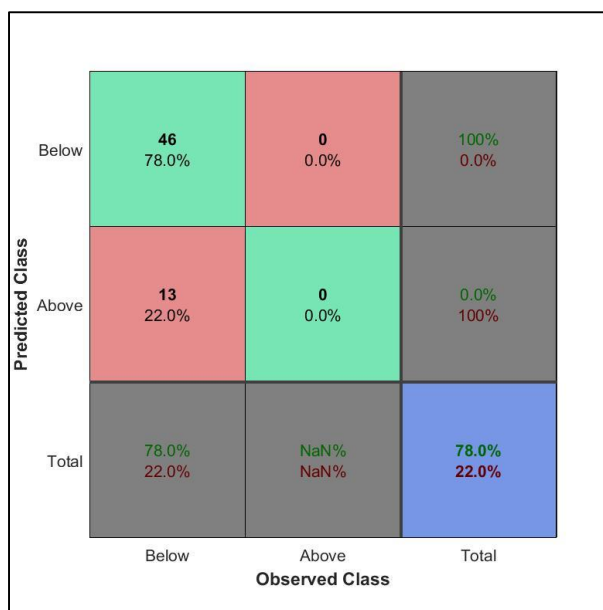


**Figure 5.38:** Confusion plot for EC from the UF

From the PCA analyses it was expected that the LDA models for the Nitrate and Nitrite in the WWTP Clarifier effluent will not perform well, however, the models had classification errors of 15.1% and 22.0% respectively (Figure 5.39 and Figure 5.40).

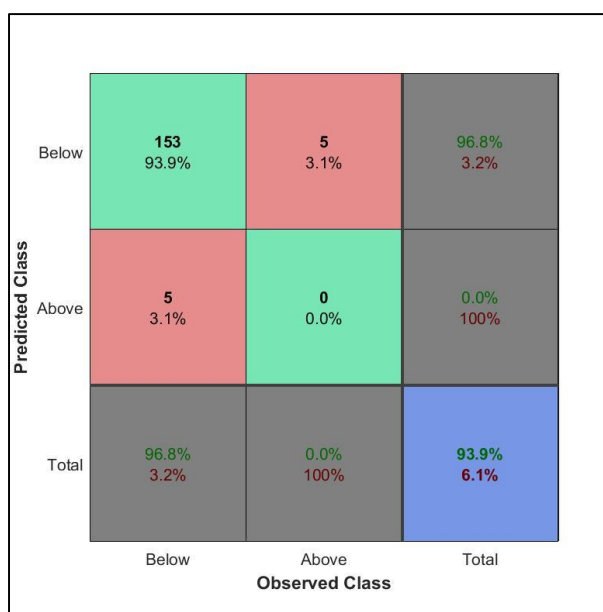


**Figure 5.39:** Confusion plot for Nitrate from the WWTP Clarifier



**Figure 5.40:** Confusion plot for Nitrite from the WWTP Clarifier

Unfortunately, despite the good results of the LDA models, it was unclear whether or not these models were actually performing well or not. With most of the LDA models, that had classification errors less than 10%, it was found that there were very few (less than 20) observed cases falling within the ‘Above’ class. It is therefore not sure whether the LDA models are predicting accurately or not, since the overwhelming majority of the observed data points lay within the ‘Below’ class (Figure 5.41).

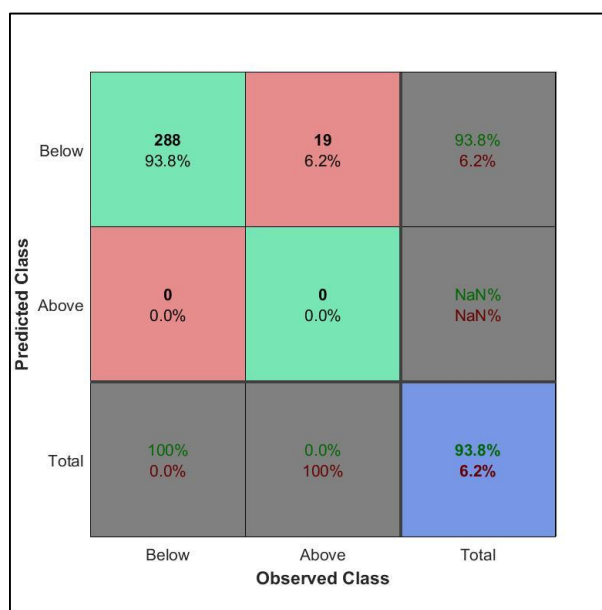


**Figure 5.41:** Confusion plot for UV<sub>254</sub> in the WWTP Final Effluent



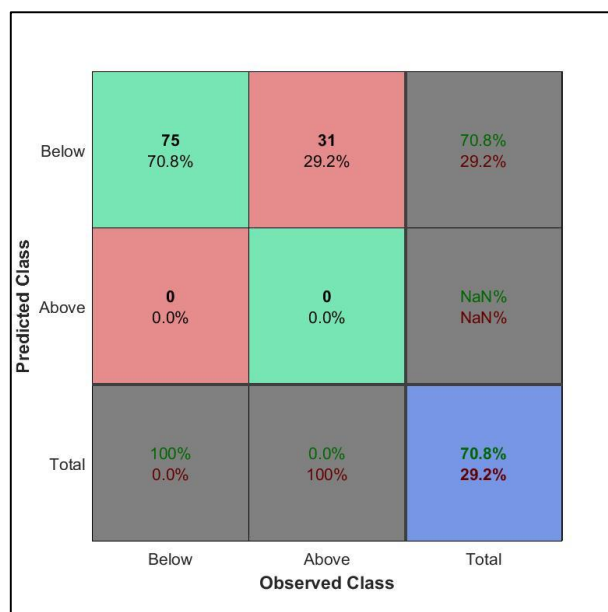
This tendency was also observed in the distribution of the types of classification errors that were made. In general, the majority of the classification errors that were made were false-positive errors, which should be considered a good thing since it is much better for a WRP to unnecessarily re-treat water that was wrongfully classified as unsafe, than to wrongfully not re-treat water that was wrongfully classified as safe (in the case of a false-negative error).

However, in the cases where data points were observed in the 'Above' class, the number of false-negative classification errors were much higher than the number of false-positive classification errors (Figure 5.42).



**Figure 5.42:** Confusion plot for Residual O<sub>3</sub> from the Ozone Contact B

Upon further investigation, it was found that the majority of the LDA models were completely unsuccessful in predicting when a data point should fall within the 'Above' class (Figure 5.42 and Figure 5.43).



**Figure 5.43:** Confusion plot for pH from the WWTP Clarifier

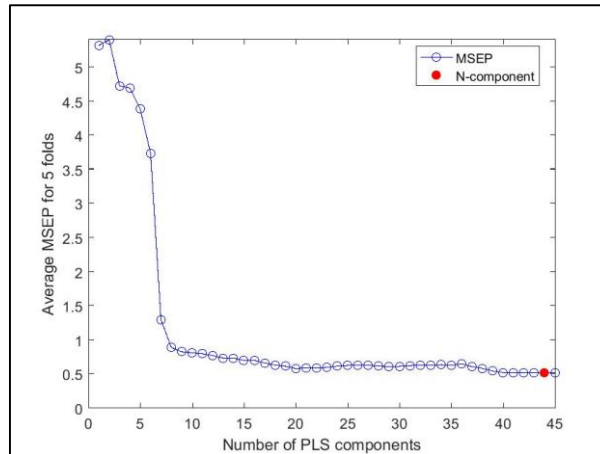
### Partial least squares regression

The PLS models were trained and cross-validated using 5 k-folds. From the cross-validation, plots were made of the MSEV for each component (latent variable) included in the model. This was also how the optimal number of components (N-component) was selected for inclusion in the final models that were applied on the testing data.

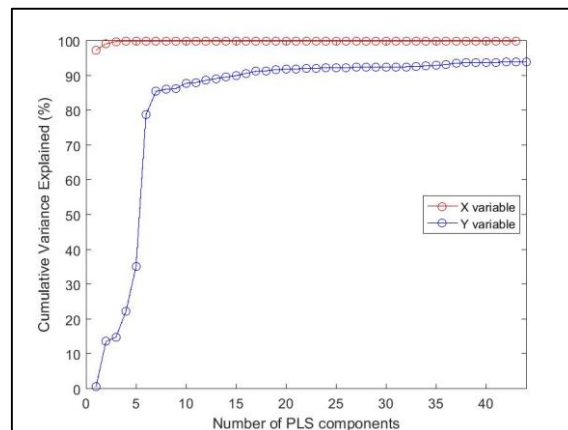
The average MSEV was plotted for each of the cross-validation runs in order to select the N-component, where the lowest MSEV was obtained. These plots (Figure 5.44, Figure 5.46, Figure 5.48, Figure 5.49, Figure 5.52 and Figure 5.53) illustrate the average MSEV against the number of components (latent variables). The purpose of the cross-validation is to determine the number of components at which the average MSEV is a minimum. Plots were also made of the cumulative variance explained in both the **X** and **Y** variables for each of the components up to and including the N-component (Figure 5.45).

For illustrative purposes, the models for the following variables will be discussed here:

- Temperature in the WWTP Clarifier effluent
- Nitrite in the WWTP Clarifier effluent
- Chlorophyll A in the WRP Influent
- Calcium Hardness in the WWTP Final Effluent
- Clostridium Spores in the WWTP Final Effluent
- Faecal Coliform in the WWTP Final Effluent



**Figure 5.44:** MSEP per component for Temperature from the WWTP Clarifier



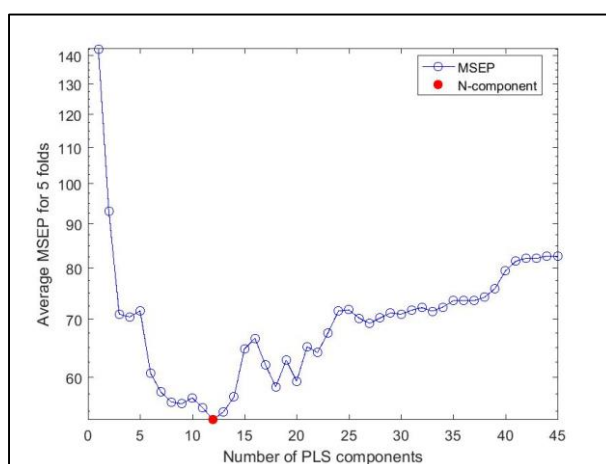
**Figure 5.45:** Variance explained per component for Temperature from the WWTP Clarifier

Of the various MSEP plots that were made, many different shapes and forms were observed. These shapes could be explained, to some extent, with regard to the expected quality of the models that would be built for the variables. Five main shapes were identified of which the first two are the expected shape that may produce a model that performs well.

The shapes of the explained variance plots were not interpreted. Since these plots indicate the cumulative variance explained, all of them increase monotonically and could only be classified as either having an obvious elbow, or not. In either case, the value of these plots were for determining the expected variance that would be explained by the models. Depending on the noise in the data, the ideal percentage of variance explained by the models will differ. Since it is of no value if the model reproduces the noise in the training data, it is therefore possible for a model to perform well, even if it does not explain 100% of the variance in the data.

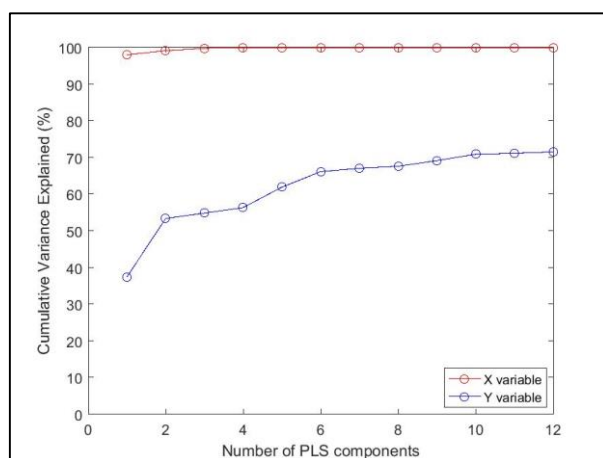
Of the shapes observed in the MSEP plots, the first shape indicated an elbow (Figure 5.44). In these cases, there is a large gain in adding more components until a certain point is reached whereafter there is little to gain. In the cases where this shape was observed, it was expected that the models would perform well since the ideal number of components to include in the model was clear. This was also reflected in the plots of the cumulative variance explained, which indicated a clear elbow (Figure 5.45).

In other cases, a minimum is produced near the start of the plot (Figure 5.46). In these cases each additional component adds value to the model, until a certain point is reached whereafter the addition of more components is detrimental to the model (most likely these components represent noise in the data). Although these models may not be able to explain all the variance in the data (Figure 5.47), the predictions may be more accurate since the noise containing components were, presumably, not selected for the model.



**Figure 5.46:** MSEP per component for Nitrite from the WWTP Clarifier

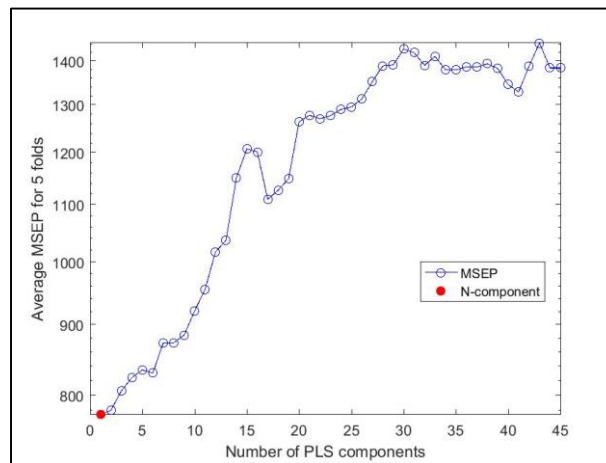
Again, it was expected that the models would perform well since the ideal number of components to include in the model was clear. The explained variance by the model is just over 70% (Figure 5.47), which is certainly enough to produce a well performing PLS model.



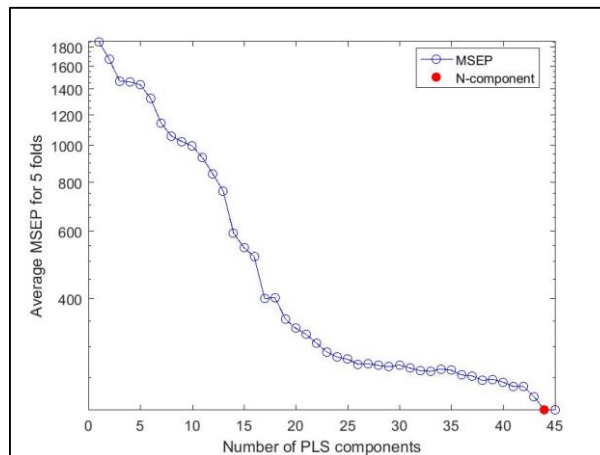
**Figure 5.47:** Variance explained per component for Nitrite from the WWTP Clarifier

The first two shapes that were discussed earlier, are considered the most ideal situation since there is a clear point at which the model no longer benefits from additional components. The remaining shapes that will be discussed are less ideal and may represent cases where PLS models will not perform well. In these cases, there is not a clear point at which the PLS models will no longer benefit from additional components.

The third and fourth shape, which may indicate unideal model performance, was a monotonic increase (Figure 5.48) or decrease (Figure 5.49) in the MSEP plot without any inflections or local minima or maxima, respectively.



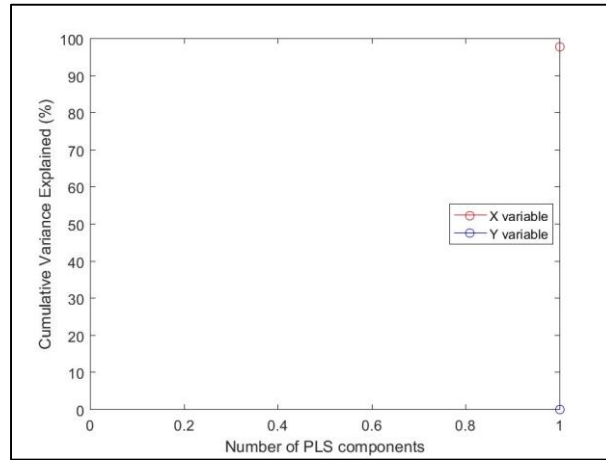
**Figure 5.48:** MSEP per component for Chlorophyll A in the WRP Influent



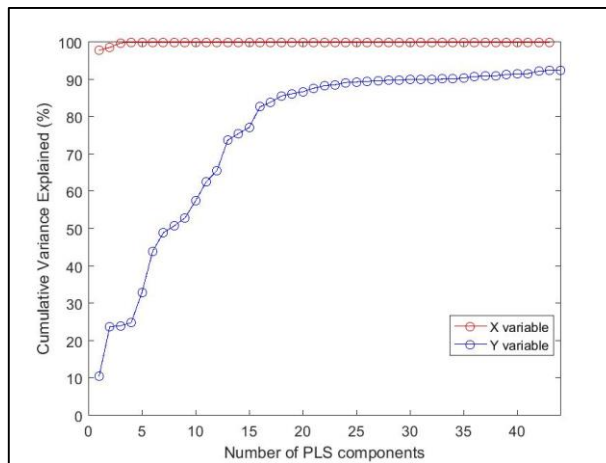
**Figure 5.49:** MSEP per component for Calcium Hardness in the WWTP Final Effluent

In the case of a monotonic increase in MSEP the selected N-component for the PLS models will be low (for Chlorophyll A in the WRP Influent, only one component was selected), which could result in a model that is not capable of explaining a large amount of variance in the data (Figure 5.50). On the other hand, in the case of a monotonic decrease in the MSEP, it is likely that too many

components will be included in the model (Figure 5.51), which could introduce noise and therefore inaccurate predictions.

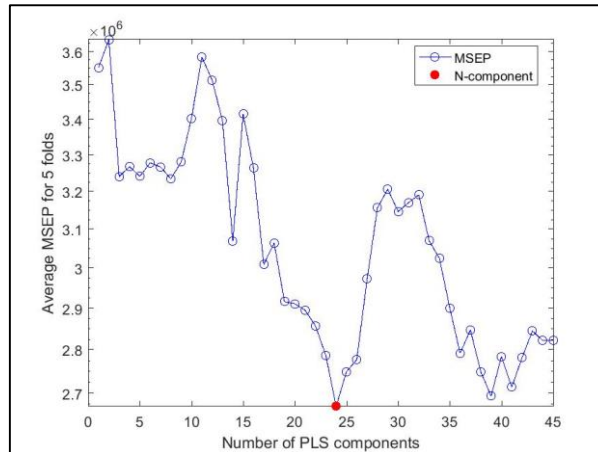


**Figure 5.50:** Variance explained per component for Chlorophyll A in the WRP Influent



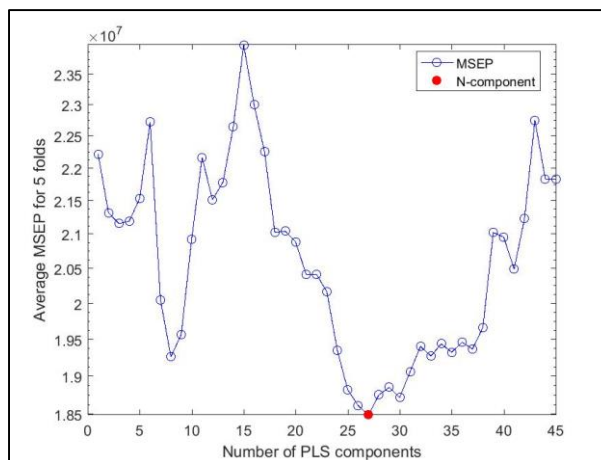
**Figure 5.51:** Variance explained per component for Calcium Hardness in the WWTP Final Effluent

The fifth shape, which may also indicate unideal model performance, was characterised by random, or erratic, behaviour. Initially it was suspected that these plots (Figure 5.52 and Figure 5.53) were the result of having too few data points within a k-fold, but after comparing the counts of data points in these plots to others, there was no link found between the number of data points and the shape of the plot. Instead, it was found that these plots generally belong to variables that indicate biological constituents as a number, or count.

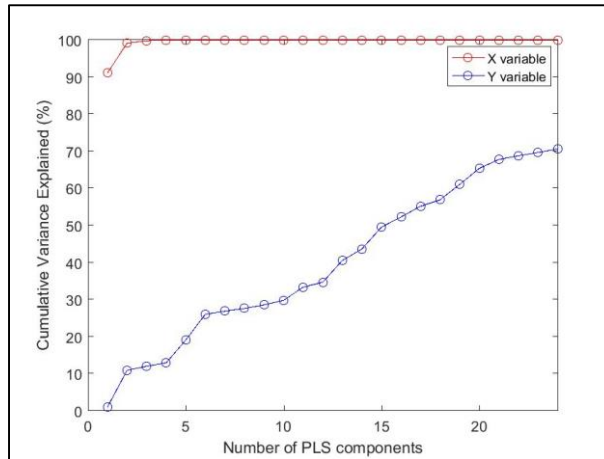


**Figure 5.52:** MSEP per component for Clostridium Spores in the WWTP Final Effluent

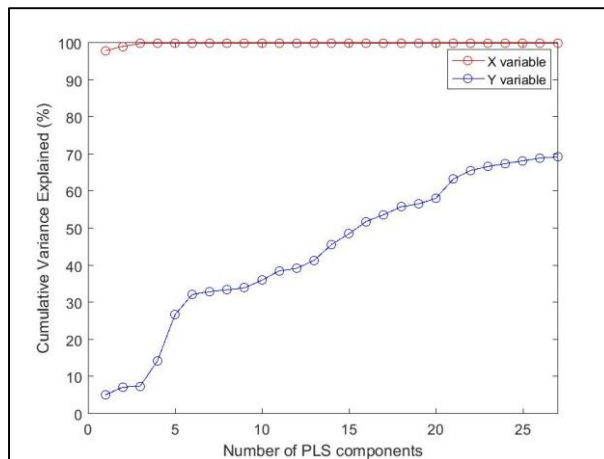
The erratic nature of these plots (Figure 5.52 and Figure 5.53) may, therefore, be due to the high variance of the variables, where the majority of the values are low (less than 10) but certain values (that were not considered outliers) are as high as  $10^{12}$ . It is therefore expected that the different cross-validation runs contained an unequal distribution of high values in the various k-folds, which led to the erratic behaviour observed in the plots.



**Figure 5.53:** MSEP per component for Faecal Coliform in the WWTP Final Effluent



**Figure 5.54:** Variance explained per component for Clostridium Spores in the WWTP Final Effluent



**Figure 5.55:** Variance explained per component for Faecal Coliform in the WWTP Final Effluent

The performance of the PLS models was assessed using the  $R^2$  value produced after plotting the observed and predicted testing data of the  $y$  variables against each other in a scatter plot. These plots also indicate a straight line going through the origin at a  $45^\circ$  slope, which represents a perfect fit ( $R^2 = 1$ ) between the predicted and observed data.

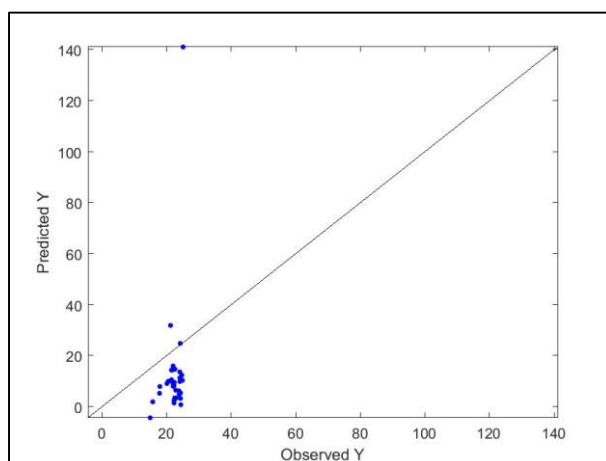
A summary table with the  $R^2$  values for all the PLS models will be showed at a later stage, but first, the scatter plots for each of the variables that were illustrated with regard to their MSE and variance explained, will be illustrated and discussed.

The observed and predicted values for the Temperature in the WWTP Clarifier effluent were expected to be accurate, as discussed earlier, but this was not the case. The predicted Temperature values included several data points that could be considered outliers, or even impossible values, including a temperature of  $140^\circ\text{C}$  (Figure 5.56).



In the scatter plot, the expectation is that the data will lie in a straight line which indicates that the value for each given record for both variables are the same. This will also mean that the mean, standard deviation and variance of the predicted data are the same as the observed data. One reason for the poor performance can definitely be the selection of the N-component during the training of the model. It was shown that the MSEV made a clear elbow, yet the selected N-component was not at the elbow, but rather to global minimum from all of the cross-validation runs.

From Figure 5.44 it would have been expected that the number of components that were to be included in the model would be seven or eight, but instead 44 were selected. This was a result of using an algorithm that could automatically select the N-component during the development of the PLS models. It is assumed that several of the components included in the model represented excess variation (noise) in the data, thus resulting in the inaccurate prediction.



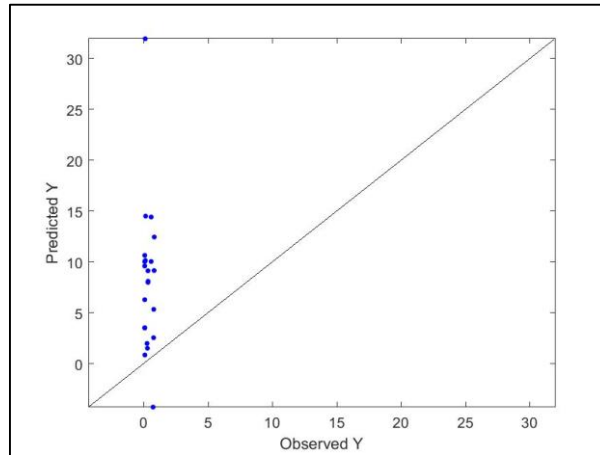
**Figure 5.56:** Predicted vs Observed scatter for Temperature from the WWTP Clarifier

The performance of the PLS model that was used to predict the Nitrite in the WWTP Clarifier effluent, also performed poorly (Figure 5.57). This was an unexpected outcome considering that the most appropriate N-component was in fact selected in this case (Figure 5.46).

Again it was found that the model predicted a few values that could be considered outliers (including a negative value), but also that the overall predictions had too much variance. Unlike the model for Temperature in the WWTP Clarifier effluent that predicted the values around a lower than observed mean value, the model for Nitrate in the WWTP Clarifier effluent predicted the values around a higher than observed mean value (Figure 5.57).

Since it can be assumed that the ideal number of components was selected for the model, the second most likely cause for the inaccurate prediction, namely poor quality input data, is the most likely. The testing data that were used by the models did not undergo any pre-processing, and

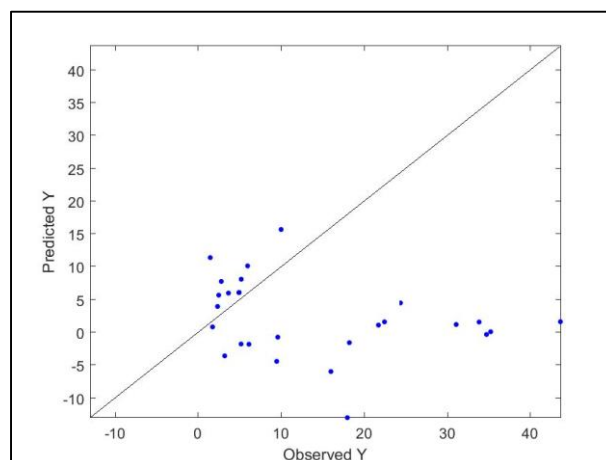
therefore contained many outliers as well as missing values. It is most likely the outliers in the input data that are responsible for the high variance in the predicted  $y$  variables.



**Figure 5.57:** Predicted vs Observed scatter for Nitrite from the WWTP Clarifier

The results for the Chlorophyll A in the WRP Influent indicate that the model did not perform well (Figure 5.58). The predicted data had a variance slightly smaller than the observed values, although a large portion of the values were impossible (less than zero). Unlike the other models that predicted outliers and generally larger than observed variances, this model predicted very few obvious outliers and had a variance smaller than that of the observed data.

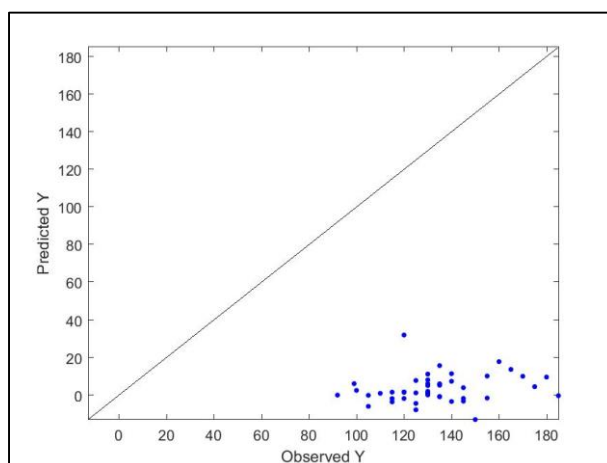
From the training of the model (Figure 5.48) it was found that only one component was included in the model, this resulted in an almost zero percent of variance in the  $y$  variable being explained by the model (Figure 5.50). This model, however, was never expected to perform well since every additional component included in the model would have potentially increased the MSE of the predictions made by the model, and selecting only one component made it impossible for the model to completely explain the variance in the data.



**Figure 5.58:** Predicted vs Observed scatter for Chlorophyll A in the WRP Influent

Similar to the model for Chlorophyll A in the WRP Influent, the model for Calcium hardness in the WWTP Final Effluent was not expected to perform well. The MSEP plot (Figure 5.49) had a monotonic decrease which resulted in 44 components being included in the model. As expected, the model did not perform well (Figure 5.59). Although, unlike the first two models (of which the one also made use of 44 components) the predicted data from this model had no obvious outliers and a smaller variance than the observed data. The mean of the predicted data were also much lower than the observed data and it was also found that a few data points were impossible (below zero).

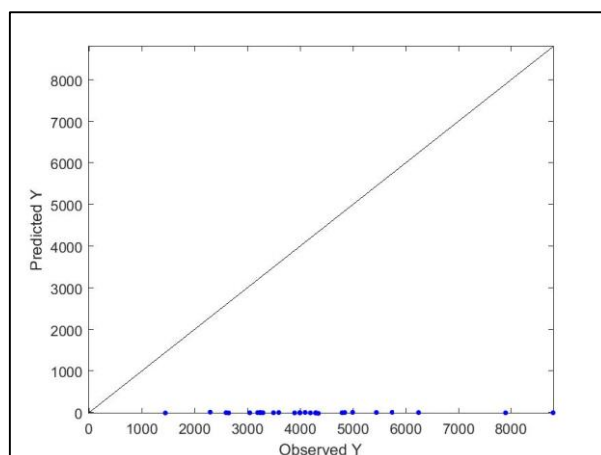
This result was quite similar to that of the Chlorophyll A in the WRP model, although there are practically no other similarities between these models. As was the case with the Temperature in the WWTP Clarifier effluent model, it is assumed that the large number of components increased the likelihood for the model to produce noise, or rather, that the noise in the training data may be affecting the performance of the final model.



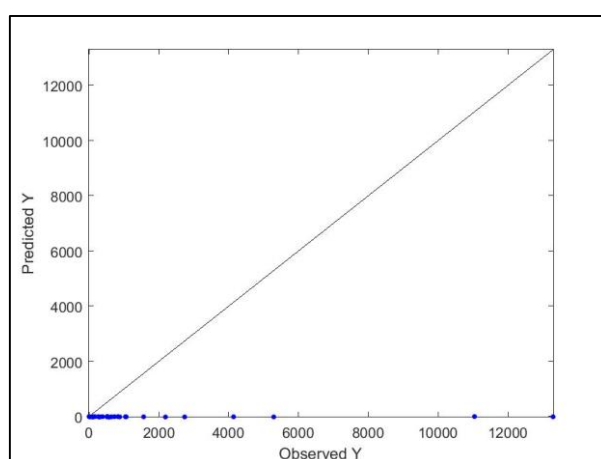
**Figure 5.59:** Predicted vs Observed scatter for Calcium Hardness in the WWTP Final Effluent

The models for the Clostridium Spores and Faecal Coliform in the WWTP Final Effluent showed similar erratic behaviour during the training of these models (Figure 5.52 and Figure 5.53). Due to this erratic behaviour, it was difficult to determine the optimal number of components to include in these models, and therefore it was also expected that these models would not perform well.

This assumption proved to be correct since both models performed poorly (Figure 5.60 and Figure 5.61). In both cases, the predicted values from the models were virtually zero. This also resulted in the mean and variance of the predicted values being zero. Unlike the previous models, the fact that these models produced data with such a small amount of variance, was unexpected.



**Figure 5.60:** Predicted vs Observed scatter for Clostridium Spores in the WWTP Final Effluent



**Figure 5.61:** Predicted vs Observed scatter for Faecal Coliform in the WWTP Final Effluent

Both models were expected to be able to explain a fair amount (70%) of the variability in the data (Figure 5.54 and Figure 5.55), and yet both models produced data with virtually no variability. The models also included an average amount of components (24 and 27) and were not suspected to suffer from noise captured in the components.

#### **5.4 SUMMARY AND EVALUATION OF IDENTIFIED RELATIONSHIPS USING EXPERT PROCESS KNOWLEDGE**

In this section the results of the developed models, and specifically the relationships that were identified, are summarised and evaluated according to expert process knowledge in order to determine the value and relevance of the identified relationships. This will be done for each of the different analyses that were performed (with the exception of PCA): correlation, LDA and PLS.

In order to evaluate the results, it is important to understand what correlations, or relationships, are of interest or importance to the suppliers of reclaimed water. An extensive list of monitoring equipment

was compiled in order to determine the availability and capability of sensors in the water and wastewater sector (Benten Water Solutions, 2016). From this list, a table was constructed (Table 5.1) indicating which variables are considered FEM within the plant operation and monitoring community. This expert process knowledge was used to evaluate whether or not the relationships identified in the study, would also be of value in their real-world applications.

**Table 5.1:** FEM variables according to expert process knowledge

Variable	Speed	Variable	Speed
Ammonia (Absorbance)	1 – 60 seconds	Hydrogen Sulphide (H <sub>2</sub> S)	Continuous
Ammonia (Ion selective electrode )	As low as 1 second	MnO <sub>4</sub> (spectrophotometric)	1 – 60 seconds
BOD	1 – 10 seconds	Nitrate	As low as 1 second
CDOM/FDOM	1 – 10 seconds	Nitrite (Ion selective electrode)	As low as 1 second
COD (Optical absorbance)	1 – 60 seconds	Nitrite (spectrophotometric absorbance)	1 – 60 seconds
Colour	1 – 60 seconds	ORP (Amperometric)	As low as 1 second
DO (Amperometric)	15 – 30 seconds	pH	As low as 1 second
DO (Luminescence)	< 60 seconds	Refractive Index (RI)	As low as 1 second
DOC (Optical absorbance)	1 – 60 seconds	Salinity	As low as 1 second
EC (Electrode)	As low as 1 second	Spectral absorption coefficient (SAC)	As low as 1 second
EC (Inductive)	As low as 1 second	Streaming Current	Continuous
Free Chlorine (Electrochemical)	Continuous	TOC (Optical absorbance)	1 – 60 seconds
Hardness (Capillary electrophoresis)	<1 minute	Turbidity	< 60 seconds
Hardness (Ion selective electrode )	As low as 1 second	UV254/SAC254	1 – 60 seconds

[Information obtained from Benten Water Solutions (2016)]

The shaded sections in Table 5.1 indicate equipment that are not commonly used due to cost or complexity. Therefore, the main FEM variables from expert process knowledge consists of: DO, EC, Free Chlorine, pH, Salinity and Turbidity.

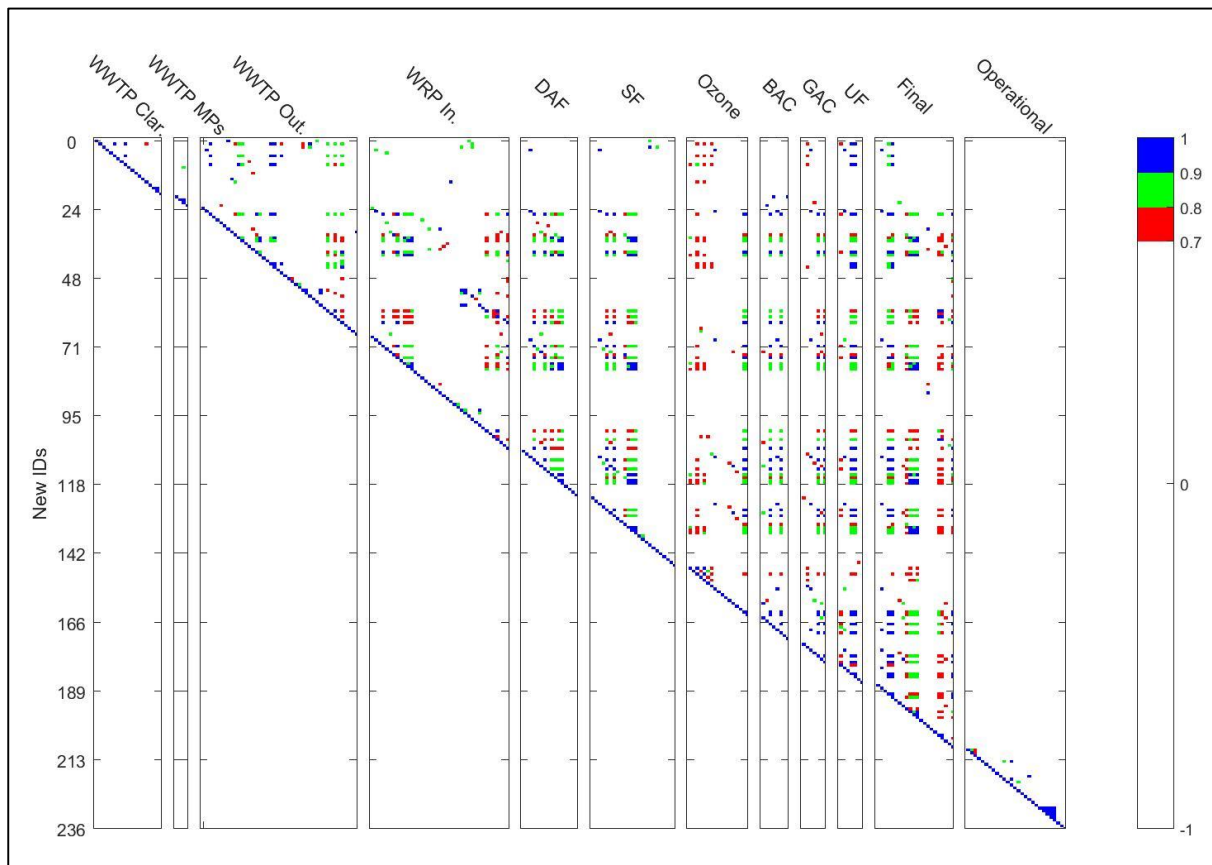
For detailed information regarding the sensor technologies, see Storey et al. (2011).

#### 5.4.1 Correlation analysis

The results of the correlation analyses before the separation of the MVR were visually illustrated in Figure 5.23 and Figure 5.24 for the Pearson and Spearman correlation coefficients, respectively; and the results of the correlation analyses after the separation of the MVR were visually illustrated in Figure 5.25 and Figure 5.26 for the Pearson and Spearman correlation coefficients, respectively.

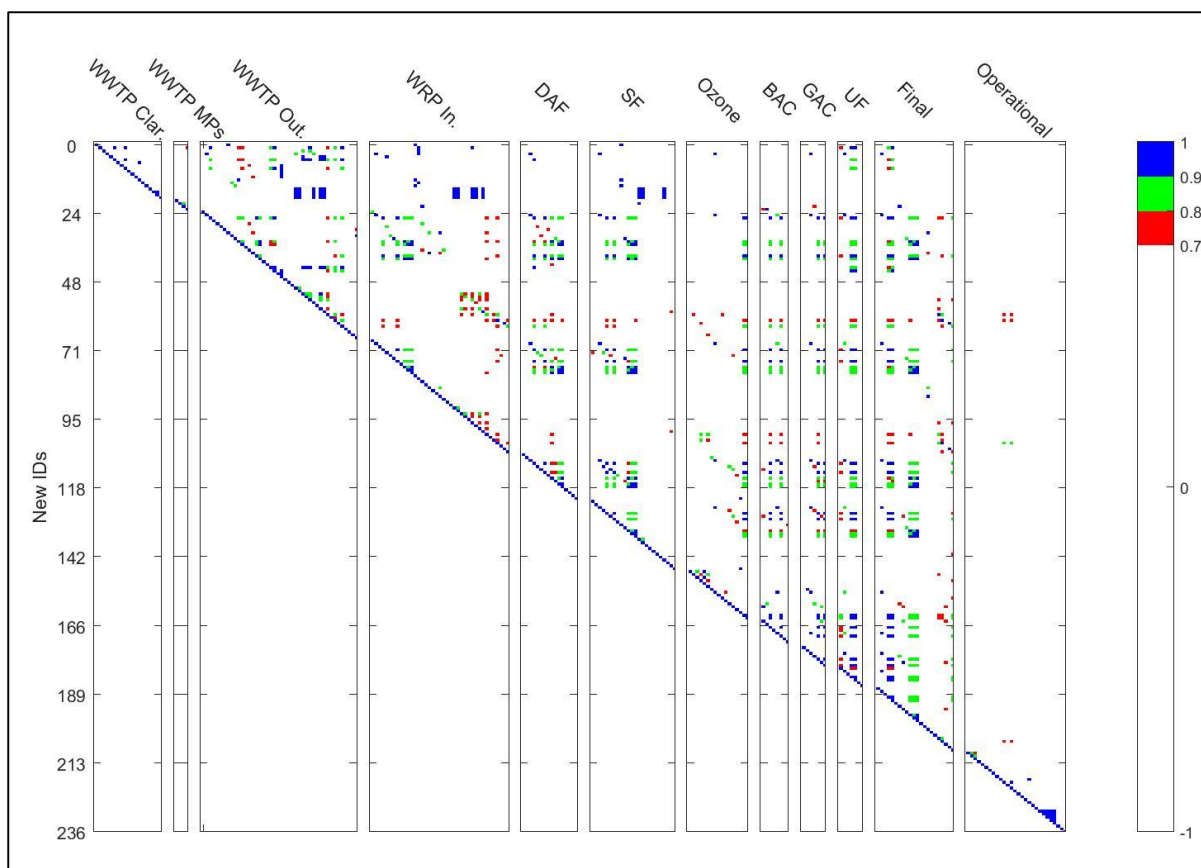
Although these plots provide a good overview of all the correlations that were performed, they did not provide a clear indication of the variables that actually correlated, or the actual correlation coefficients. In order to improve the clarity and interpretability of those results, new plots were created in order to illustrate the correlations for the variables at each of the different sampling points, or treatment units.

The new plots (Figure 5.62 and Figure 5.63) show the Pearson and Spearman correlations before separating the MVR. Since the correlation matrices are symmetrical, the bottom half of the data were removed in order to get a clearer indication of the number of significant correlations that were detected for the variables of each of the sampling locations.



**Figure 5.62:** Summary of Pearson correlations for each sampling point

The variables from the WWTP Clarifier (WWTP Clar. in Figure 5.62 and Figure 5.63), the WWTP maturation ponds (WWTP MPs in Figure 5.62 and Figure 5.63) and the operational variables, show very few significant correlations. Each of the other sampling points on the other hand contain variables that have a high, medium and low number of significant correlations.



**Figure 5.63:** Summary of Spearman correlations for each sampling point

Since there were so few significant correlations from the correlation analyses that were performed after the separation of the MVR, it was assumed that the variables with a high number of significant correlations before separation of the MVR (Figure 5.62 and Figure 5.63) primarily consist of correlations between variables of the same nature (FEM-FEM or SDM-SDM). In order to verify this assumption, the top five variables with regard to the number of significant correlations they were part of, were tabulated for each of the sampling locations (Table 5.2).

It was found that the majority of the variables with a high number of significant correlations were in fact FEM variables (Table 5.1). Since these high numbers of significant correlations were not observed after separating the MVR, it can be assumed that the majority of the correlations formed with these FEM variables, were with other FEM variables.

It was also found that the majority of the top five variables were very similar in nature (Table 5.2). These variables [EC, TDS (Calc), Total Hardness, Calcium Hardness, Magnesium Hardness pH and TS] are primarily measures of ion concentrations, and specifically ions originating from similar sources (leachate from soil and stone).

**Table 5.2:** Top five variables with significant Pearson correlations per sampling location

WWTP Clarifier	Analysis	EC	Total solids (TS:105°)	TDS (TDS:180°C)	Temperature	pH
	Count	18	14	11	9	3
WWTP MP	Analysis	Ammonia	EC	Total solids (TS:105°C)	TDS (TDS:180°)	Temperature
	Count	8	0	0	0	0
WWTP Final Eff.	Analysis	EC	Calcium hardness	TDS (Calc)	Magnesium hardness	Total hardness
	Count	44	38	36	33	30
WRP Inf.	Analysis	EC	TDS (Calc)	Calcium hardness	Magnesium hardness	Total hardness
	Count	31	30	26	25	25
DAF	Analysis	EC	TDS (Calc)	Calcium hardness	Magnesium hardness	Total hardness
	Count	26	25	21	20	19
SF	Analysis	EC	TDS (Calc)	Calcium hardness	Magnesium hardness	Total hardness
	Count	20	19	16	15	14
Ozone Contact	Analysis	EC	TDS (Calc)	Residual O3	Free chlorine	Temperature
	Count	17	16	7	6	5
BAC	Analysis	EC	TDS (Calc)	Temperature	DOC	pH
	Count	13	12	4	4	2
GAC	Analysis	EC	TDS (Calc)	Temperature	DOC	COD
	Count	11	10	3	3	2
UF	Analysis	TDS (Calc)	EC	pH	Temperature	Turbidity
	Count	8	7	5	2	1
Final Water	Analysis	EC	TDS (Calc)	Calcium hardness	DOC	Magnesium hardness
	Count	6	5	3	2	2
Operational	Analysis	SF E Runtime (h)	SF C Runtime (h)	Pre-Ozone (mg/l)	SF D Runtime (h)	Main Ozone A (mg/l)
	Count	5	4	3	3	2

It is, however, still not clear whether or not the correlation analyses revealed any valuable correlations. The correlation results were therefore tabulated (Appendix C) in order to display the information for the two variables with regard to their sampling locations and type of analysis, as well as the correlation coefficient that was calculated. The tables were then also reduced in order to only indicate significant correlations (above 0.7 with a p-value less than, or equal to, 0.05).

Finally, after applying expert process knowledge (by comparing the variables listed in Appendix C with those listed in Table 5.1), a short list of variable correlations were determined that could add value to the water reclamation community (Table 5.3 and Table 5.4).



**Table 5.3:** Pearson's correlation coefficients of interest from before separating the MVR

ID 1	ID 2	Sample Point 1	Analysis 1	Sample Point 2	Analysis 2	R
6	46	WWTP Clarifier	Total solids (105°C)	WWTP Final Eff.	Chlorophyll A	0.998
20	171	WWTP MP B1	Ammonia	BAC	Iron (Fe)	0.988
23	165	WWTP MP B8	Ammonia	BAC	COD	0.983
3	139	WWTP Clarifier	pH	SF	Faecal streptococci	0.938
2	96	WWTP Clarifier	EC	WRP Inf.	Total coliform	0.883
3	93	WWTP Clarifier	pH	WRP Inf.	Faecal coliform	0.859
3	96	WWTP Clarifier	pH	WRP Inf.	Total coliform	0.822
3	141	WWTP Clarifier	pH	SF	Total coliform	0.821

The correlation between Total solids and Chlorophyll A may have been disregarded since Total solids is not considered a FEM variable, but since it is much faster and cheaper to measure than Chlorophyll A, it was retained in the short list. Likewise, the correlations with Ammonia have been retained since it can be used as an early warning for the performance of the BAC. The remaining correlations are self-explanatory and links the two common FEM variables (EC and pH) to some of the most problematic SDM variables (microbiological pathogens).

These correlations definitely warrant further attention. From expert process knowledge it is known that pH is an important factor when disinfecting wastewater using chlorine (which is the case in this plant). These correlations, therefore, reflect and emphasise the importance of pH control during chlorination.

The same procedure was followed with the Spearman correlation results (Table 5.4).

**Table 5.4:** Spearman's correlation coefficients of interest from before separating the MVR

ID 1	ID 2	Sample Point 1	Analysis 1	Sample Point 2	Analysis 2	R
6	52	WWTP Clarifier	Total solids (105°C)	WWTP Final Eff.	Faecal coliform	1.000
6	58	WWTP Clarifier	Total solids (105°C)	WWTP Final Eff.	Somatic coliphage	1.000
9	46	WWTP Clarifier	TDS (180°C)	WWTP Final Eff.	Chlorophyll A	1.000
13	131	WWTP Clarifier	Chlorophyll A	SF	Iron (Fe)	1.000
4	50	WWTP Clarifier	Temperature	WWTP Final Eff.	Clostridium Spores	0.857
4	55	WWTP Clarifier	Temperature	WWTP Final Eff.	HPC	0.857
3	52	WWTP Clarifier	pH	WWTP Final Eff.	Faecal coliform	0.821
3	54	WWTP Clarifier	pH	WWTP Final Eff.	Total coliform	0.821
2	23	WWTP Clarifier	EC	WWTP MP B8	Ammonia	0.791
30	67	WWTP Final Eff.	UV 254	WWTP Final Eff.	TOC	0.728
154	157	Ozone Contact Final	Temperature	Ozone Contact Final	HPC	0.704

The short list of the Spearman results (Table 5.4) show many similarities to that of the Pearson results (Table 5.3). In this case, however, a correlation between TDS and Chlorophyll A was observed, as well as Total solids. Again, technically these correlations are between two SDM variables, but since the one SDM variable is much slower and more difficult to measure, they are still considered valuable. Several correlations were also observed between Temperature and microbiological pathogens, which could be of great value.

As was the case with the Pearson results (Table 5.3) the Spearman results (Table 5.4) also show correlations pH and microbial variables, which was expected due to the importance of pH during disinfection. Another correlation that can be explained by expert process knowledge, is between total solids and microbial pathogens (Faecal coliforms and Somatic coliphage) since any solids in the wastewater act as a shield against the disinfectant. Therefore it can be expected that more pathogens survive the disinfection when more solids are present in the wastewater during disinfection.

The correlation between  $UV_{254}$  and TOC is also well known and the majority of online TOC/DOC sensors make use of multiple wavelength UVA measurements. The correlation between Chlorophyll A and Iron is due to the operational practices on-site. Coagulant (in this case  $FeCl_3$ ) is dosed in the DAF in proportion to certain contaminants that are measured upstream of the DAF. Ammonia (Table 5.3) and chlorophyll A (Table 5.4) are some of the variables that may require more coagulant during the DAF treatment, thus resulting in more iron in the downstream units like the SF and BAC.

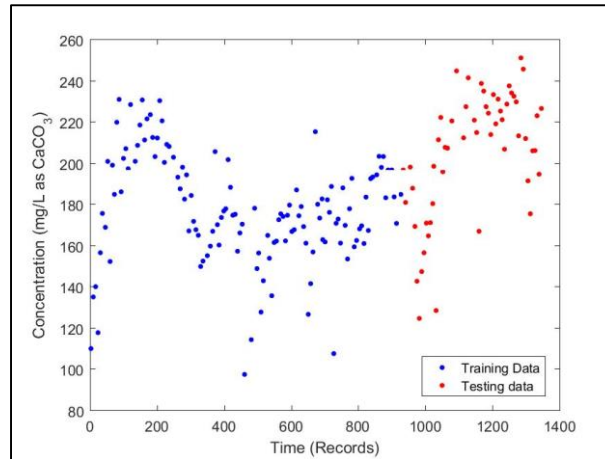
Overall, from the multitude of correlations that were calculated, a great number were rejected. Of the significant correlations, only a small number of correlations were considered valuable. Although there were few valuable correlations found in this study, some of these correlations could result in surrogates that may have a large impact on the way WRPs are monitored in the future.

#### **5.4.2 Linear Discriminant Analysis**

During the discussion of the LDA models' results, it was speculated that the good performance of the LDA models may be misrepresentative since the testing data contained few data points classified as 'Above'. Therefore, it was uncertain whether or not the good performance of the LDA models reflect accurate prediction, or simply a bias in the testing data. If this is in fact the case, it would be expected that LDA models for variables that contained many testing data points that were classified as 'Above', would perform poorly.

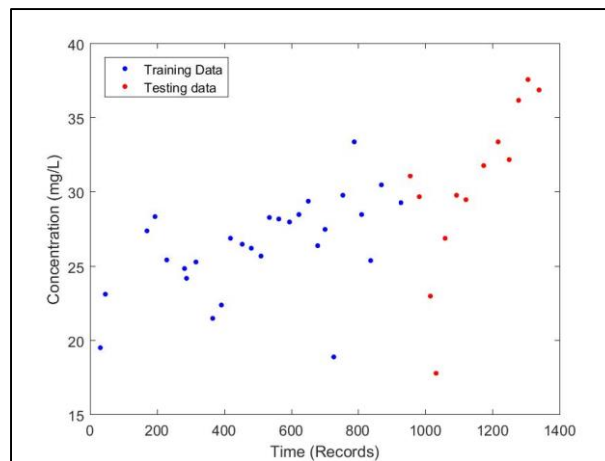
Since the LDA limits that were used to classify the data into categories were constants (remained the same over the entire period of the data), it is expected that variables with a large number of data points classified as 'Above' will have a positive trend in a plot of the variable over time. Of all the LDA models, only three had an accuracy less than 50%, namely: Total Alkalinity in the SF, Potassium in the WRP Influent and Potassium in the Final Water.

The raw data of these three variables (which were used to establish the LDA limits) were plotted against time (record number) in order to determine if a positive trend was present in the data for these variables (Figure 5.64, Figure 5.65 and Figure 5.66).

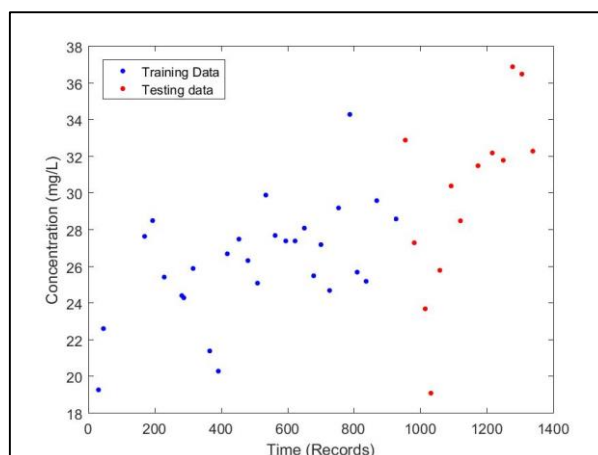


**Figure 5.64:** Total Alkalinity from the Sand Filter against time

In the case of the Total Alkalinity in the sand filter effluent, it was found that a positive trend in the data is visible, although only in the last two-thirds of the data. Despite not being a global trend, the majority of the values of the testing data are greater than the majority of the values of the training data.



**Figure 5.65:** Potassium in the WRP Influent against time



**Figure 5.66:** Potassium in the Final Water against time

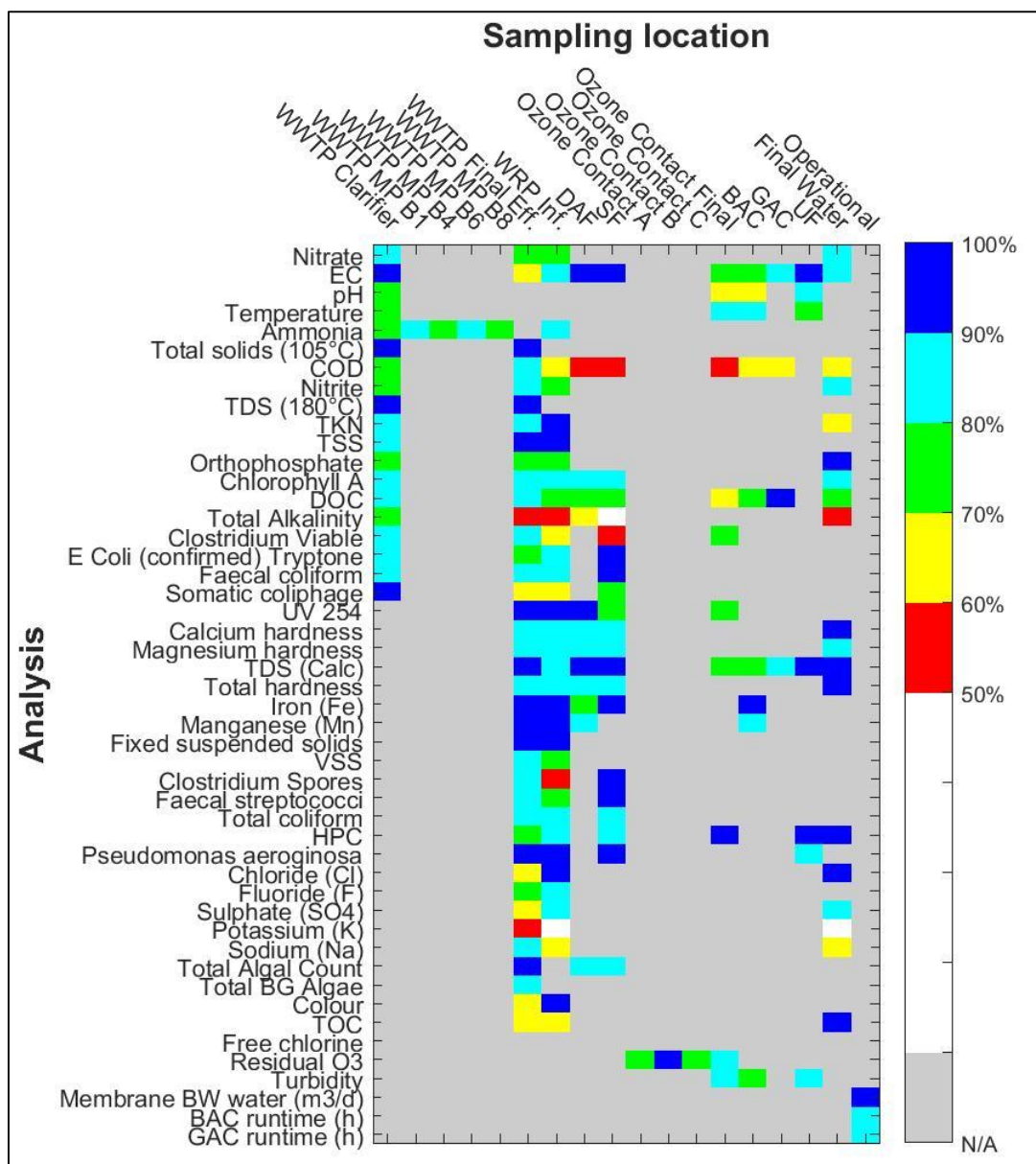
For the Potassium in the WRP Influent and Final Water, a positive trend was detected over the entire range of the data (training and testing). Therefore, the majority of the values of the testing data were in fact larger than the majority of the values of the training data.

These three variables can therefore be expected to produce a large number of data points in the testing data that were categorised as 'Above', and therefore the LDA models for these variables were expected to perform poorly, which was confirmed.

The mismatch in the ratio of values being classified as Above/Below between the training and testing data is also indicative of a change in the process, which in itself could serve as an indicator to plant operators. Unfortunately there was not enough data in the testing data set to definitively prove that there was a change from the training to the testing data. Also, in order to keep the models as robust as possible, it was decided not to adjust, or correct, the models in the cases where a mismatch did occur.

In order to have an overall view of all the LDA models, a graphic was generated (Figure 5.67) that illustrates the model accuracy using different colours. The same colours were used in some of the correlation matrices and indicate an accuracy of above 50%, 60%, 70%, 80% and 90% using the colours red, yellow, green, cyan and blue, respectively. An accuracy of less than 50% remained white and areas where no models were constructed were coloured grey.

The accuracy of the models were calculated as the percentage of correct predictions from the total number of predictions that were made.



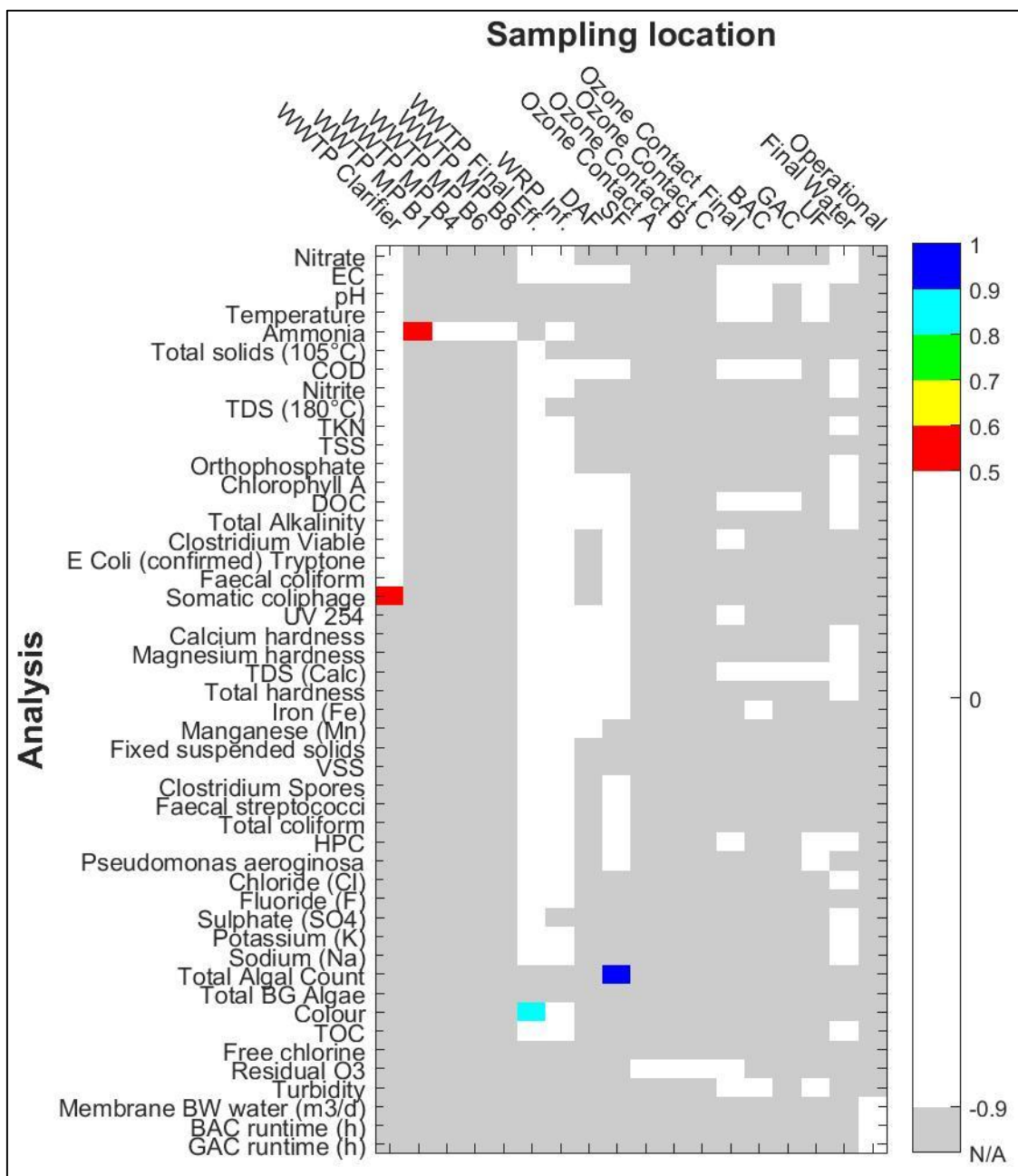
**Figure 5.67:** Summary of LDA results

Of all the LDA models, only six models had an accuracy below 50% (Figure 5.67). By comparing Figure 5.67 with Table 5.1 it was possible to determine which of the LDA models not only performed well, but are also of importance to the water reclamation community. Unlike the correlation analyses, a large portion of the LDA models could be of value to the water reclamation community; including models for variables like Ammonia, DOC, COD, TOC as well as microbiological pathogens and several heavy metals (iron and manganese).

### 5.4.3 Partial Least Squares regression

The results for the PLS regression models were illustrated using scatter plots of the observed and predicted variable values (Section 5.3.2). As was the case with the LDA models, the results for all of the PLS models were illustrated on a single graph using different colours (Figure 5.68). The same colour scheme was used, however, since the  $R^2$  values were used to determine the accuracy of the models,

the different colours (red, yellow, green, cyan and blue) correspond to  $R^2$  values above 0.5, 0.6, 0.7, 0.8 and 0.9, respectively.



**Figure 5.68:** Summary of PLS results

The PLS models performed noticeably poorer than the LDA models, since the majority of the spaces are white (except for the grey spaces where no model was built), indicating an  $R^2$  value less than 0.5. The significance test that was used to determine the significance of the correlations in the correlation analyses, was also used to determine the significance of the  $R^2$ -values for the PLS models. Thus, if the hypothesis test resulted in a p-value higher than 0.05, the  $R^2$  value should be rejected.

However, after performing the significant test and rejecting the appropriate  $R^2$  values, zero PLS models were observed with an  $R^2$  value larger than 0.5. This means that none of the PLS models were capable

of identifying statistically significant relationships between variables. There is therefore no sensible application of expert process knowledge required. Figure 5.68 illustrates the  $R^2$  values before removing those values that corresponded to p-values larger than 0.05.

Although the PLS models did not identify variable relationships that could be of value to the water reclamation industry, the process of developing, testing and evaluating the PLS models could still be of value to researchers that plan on performing similar studies. With the large variety of PLS models that were build, the fact that all of the PLS models performed poorly indicates that the problem most probably lies with the testing data. Since the testing data did not undergo any form of pre-processing, outliers, noise and missing values were present in the testing data.

However, this will always be the case. Since it is not an option to pre-process the testing data, and since the data originates from a real-world plant, there is no solution to this problem. It may be beneficial to add lag times or time delays between the time series of the various variables based on the retention time of the treatment units where the variables were measured, but in this case, the retention time of those units were insignificant in comparison to the rate at which the variables were measured.

## 6 CONCLUSIONS

Several conclusions can be made from this study and will be discussed in this chapter. The structure of this chapter is in line with the objectives of the study. The conclusions relevant to each of these objectives will be discussed in the subsequent sub-sections.

### 6.1 OBTAINING, ORGANISING AND ASSESSING HISTORICAL PLANT DATA

Most of the data validation tests were performed successfully, only the data reconciliation was lacking since there were not enough data for the flow rates of the plant. From the missing data and dirty data characterisations there were no clear reason to include or exclude certain data periods. Eventually it was decided to make use of the event logs in order to exclude data periods during which critical events were logged. The data from these periods were then removed from the research data.

Minimum data set size requirements were met on the new data set and the new data set also proved to be a sufficient representation of the actual plant. Despite having many missing values, the selected data were considered appropriate for the analyses that would follow.

### 6.2 DEVELOPMENT AND PERFORMANCE OF THE STATISTICAL MODELS

Data validation was effective in removing data portions that did not represent the actual normal operation of the plant. The data, however, still required pre-processing before the statistical models could be built on the data.

#### 6.2.1 Data pre-processing

Data pre-processing was an essential step in the development of the statistical models. The three sigma outlier removal algorithm worked well and effectively removed outliers from variables with normal, lognormal and empirical distributions. The addition of the Hampel filter was critical since it was capable of removing and replacing outliers that remained due to the local variations in the data and the Hampel filter's resistance to outliers.

Noise removal proved most effective with a window size of 11 data points. Due to the large variety of sampling rates for the different variables, the noise filter was slightly over sensitive for high sample rate variables, and slightly insensitive for low sample rate variables. Nonetheless, the filter with a window size of 11 performed the best overall.

Linear interpolation was used to replace the missing values in the MVR. Only the MVR underwent missing value replacement, since the other variables contained more than 20% missing values.

The pre-processing of the data were considered successful and of great importance for the development of the statistical analyses that was required.



## 6.2.2 Bivariate correlation analysis

Both Pearson and Spearman correlations were calculated in order to determine if any bivariate relationships could be identified between the variables. The Spearman correlation was included in the study since there were several variables that did not have normal distributions. Both correlations were calculated on the data before and after separating the MVR from the response variables.

The results of the correlations were evaluated using a hypothesis test. Only correlations above 0.7 that resulted in a rejection of the null-hypothesis (at a p-value less than, or equal to, 0.05) were considered significant correlations. The Spearman correlation analysis produced a larger number of significant correlations than the Pearson correlation analysis. This was expected since the Pearson correlation assumes a normal distribution in the variables, whilst the data contained a significant amount of non-normally distributed variables.

In both cases, however, a large number of significant correlations was observed between the different variables that were analysed.

## 6.2.3 Principal Component Analysis

From the PCA it was found that variables that had a high bivariate correlation, also had similar coloured scatter plots. It was also found that some variables contained clusters of the same colour on certain features that were observed in the scatter plots.

The expectation was that predictive multivariate models would perform better for variables that showed either separable clusters in the scatter plots of the principal components, or a clustering of colours in the scatter plots. However, this was not the case.

PCA was therefore also performed on the testing data in order to determine whether there was a significant difference between the training and the testing data, and this was found to be true. It was therefore expected that the LDA and PLS models would perform poorly when applied on the testing data.

## 6.2.4 Linear Discriminant Analysis

When all of the LDA models were evaluated, it was found that the majority of the models had an accuracy of 50% or higher, with only six models having an accuracy less than 50%. In total, 49 of the 192 LDA models had an accuracy of 90% or higher. This was unexpected since the PCA indicated that the LDA and PLS models were unlikely to perform well.

After further investigation, however, it was found that the LDA models did not perform well in cases where there were a significant (~10%) number of data points in the 'Above' class of the observed data. It was also found that the majority of the categorical response variables in the observed and predicted data contained a very small fraction of data points belonging to the 'Above' class. The LDA models therefore appear to perform well, but it was impossible to confirm whether or not this was due to the accuracy of the models, or a bias toward 'Below' classifications in the testing data.

### **6.2.5 Partial Least Squares regression**

When the results of the models were observed, it was found that all of the models performed poorly, irrespective of the findings during the training of the models.

As with the LDA models, the performance of all of the PLS models were graphed on a single graph in order to provide a fast overview of the models. The performance of the models were measured by correlating the observed data with the predicted data for each of the variables. These  $R^2$  values were then plotted using different colours to represent different performance levels (accuracy based on  $R^2$  values).

Again, a hypothesis test was used to determine if there is a significant correlation between the observed and predicted values for each of the models. It was found that only 6 of the 192 models could reject the null-hypothesis ( $p$ -value less than, or equal to 0.05), and of all 192 models, only 2 had a  $R^2$  greater than 0.7. This was most likely due to the poor quality of the testing data which have also been highlighted by the PCA analyses.

## **6.3 EVALUATION OF IDENTIFIED RELATIONSHIPS USING EXPERT PROCESS KNOWLEDGE**

Using expert knowledge proved very useful for determining the value of the results that were generated by the models, and most importantly, highlighting results that could potentially impact the potable water recycling community in a positive way.

The results from the correlation analyses contained a large amount of trivial correlations and it was only after applying expert process knowledge that a short list of correlations were obtained. This short list contained statistically significant correlations that could also be of value to the potable water recycling community. Thus, bivariate correlation was considered a success with regard to both of the evaluation criteria.

The most common relationship that was detected, between EC and TDS (Calc), was found at each of the treatment units. This is exactly in line with the expectation, since it is known from expert process knowledge that the TDS (Calc) is not a measured value, but in fact a calculated value using a linear relationship between EC and TDS. In the cases where TDS (180°C) were actually measured, these variables also correlated with EC, which was also expected from expert process knowledge.

The majority of the relationships that were detected throughout the treatment system consisted of variables that measure ionic concentrations in the streams (EC, TDS, Total Hardness, Calcium Hardness, Magnesium Hardness, pH and TS). From expert process knowledge it is known that all of these variables are FEM, and are expected to correlate since they measure ions that have similar pathways. These relationships were, therefore, also considered trivial, since they do not add new knowledge to the water reclamation community.

Of the relationships that were identified, those that were considered valuable to the water reclamation community (Table 5.3 and Table 5.4) were primarily within the WWTP upstream of the WRP. These relationships, despite not being for variables within the WRP, were still considered valuable since they can serve as warnings and indicators to the WRP. Many of these relationships can be used to great advantage to the WRP, since they require a thorough understanding of the influent (final effluent from the WWTP) in order to operate the WRP correctly.

The valuable relationships identified by the correlation analysis are characterised by having SDM variables as the response variable, and the response variables being of great importance to the WRP. From Table 5.3 and Table 5.4 it can be seen that the majority of the response variables are microbiological or Chlorophyll A. Both of which are not typically important for WWTPs, but play a major role in the performance and operation of WRPs.

There was not any reason, or scope, to apply expert knowledge to the PCA results and therefore the LDA and PLS models were the only multivariate results that were evaluated using expert process knowledge. From the majority of the LDA results, it was found that LDA also succeeded in both of the evaluation criteria. Unfortunately, since the results of the LDA were considered flawed, it was not possible to determine if LDA definitively adheres to the evaluation criteria, or not.

Since the PLS models performed poorly, there were no results that could be evaluated using expert process knowledge. This does not rule PLS regression out as a statistical method for identifying surrogates and indicators, but as far as this study goes, PLS could not be proven to have a significant contribution for the identification of surrogates and indicators using historical plant data.

## **6.4 OVERALL CONCLUSIONS**

Overall, the study was successful in applying statistical analyses to historical plant data, which were capable of identifying relationships between variables. These relationships were both statistically significant, as well as being of value to the potable water recycling community.

However, not all of the statistical analyses performed well. In general, the evaluation revealed that the multivariate analyses performed poorly and this was assumed to be due to the quality of the testing data. For the multivariate analyses, the testing data that were included in the MVR contained missing values, which was detrimental to the performance of the models that were tested on that MVR.

## 7 RECOMMENDATIONS AND FUTURE WORK

This study showed that historical plant data can be used for the application of statistical analysis with the aim of identifying relationships between the measured variables. There is, however, a high likelihood that statistical data will contain many errors, or flaws, which will require correction before performing the analyses. It is recommended that the data be pre-processed with careful consideration of the types of techniques and the factors that influence the operation of these techniques.

Simple bivariate correlations may prove very useful in identifying relationships, despite being less sophisticated, since they are more robust. The use of multivariate analyses must be considered based on the data that is available. If the data contains many missing values in the testing data, it is likely that the multivariate models will not function properly since the multivariate record will contain too many missing values.

It is therefore recommended that WRP operators or owners that are interested in developing surrogates and indicators for their WRP, make use of simple techniques if historical data is the only data available. If, however, they wish to make use of multivariate analyses, it will be better to collect data by means of a planned sampling campaign and paying careful attention to the time at which the samples are taken, keeping in mind the residence time between each of the units.

The cost of such a sampling campaign should also be carefully considered and weighed against the potential cost saving of implementing advanced monitoring techniques at the plant.

## REFERENCES

- Aggarwal, C., 2013. Outlier Analysis. 1st ed. New York: IBM T. J. Watson Research Center.
- Anderson, P., Denslow, N., Drewes, J.E., Olivieri, A., Schlenk, D. and Snyder, S., 2010. Monitoring Strategies for Chemicals of Emerging Concern (CECs) in Recycled Water, Sacramento, California: State Water Resources Control Board.
- Anumol, T. & Snyder, S., 2013. Sensitive LC/MS quantitation of trace organic contaminants in water with online SPE enrichment, Tucson, AZ, USA: Agilent Technologies, Inc.
- Asano, T., Burton, F.L., Leverenz, H.L., Tsuchihashi, R. And Tchobanoglous, G., 2007. Water Reuse: Issues, Technologies, and Applications. 1st Edition ed. New York: McGraw-Hill.
- Aull, E.M., 2005. Water quality indicators in watershed subbasins with multiple land uses, Massachusetts: Worcester Polytechnic Institute.
- AWRCE, Australian Water Recycling Centre of Excellence., 2013. Australian water recycling. [Online]. Available at: <http://www.australianwaterrecycling.com.au/factsheets.html> [Accessed 20 October 2015].
- Babcock, R., King, S., Khan, E. & Stenstrom, M., 2001. Use of Biodegradable Dissolved Organic Carbon to Assess Treatment Process Performance in Relation to Solids Retention Time. Water Environment Research, 73(5), pp. 517-525.
- Balakrishnama, S. & Ganapathiraju, A., 1998. LINEAR DISCRIMINANT ANALYSIS - A BRIEF TUTORIAL, s.l.: Institute for Signal and Information Processing, Department of Electrical and Computer Engineering, Mississippi State University.
- Benten Water Solutions, 2016. Online Water Quality Sensors and Monitors Compendium. [Online] Available at: <http://www.wqsmc.org/> [Accessed 27 July 2016].
- Berg, S., 1996. Diagnosis Problems in Wastewater Settling, Sweden: Lund Institute of Technology, Department of Industrial Electrical Engineering and Automation.
- Brillinger, D., 2011. Data analysis, exploratory. In: International encyclopedia of political science. Thousand Oaks, CA, USA: SAGE Publications, Inc., pp. 531-538.
- Buydens, L., 2013. Partial Least Squares: A Tutorial. Lecture notes, Radboud University Nijmegen.
- Byrne, A., Chow, C., Trolio, R., Lethorn, A., Lucas, J. and Korshin, G.V., 2011. Development and Validation of Online Surrogate Parameters for Water Quality Monitoring at a Conventional Water Treatment Plant Using a UV Absorbance Spectrolyser. Intelligent Sensors, Sensor Networks and Information Processing, Volume 1, pp. 200-204.

Cain, C.R., 2011. An analysis of direct potable water reuse acceptance in the United States: Obstacles and Opportunities, Baltimore: Johns Hopkins Bloomberg School of Public Health.

Chalmers, B., Yoder, D. & Patel, M., 2011. Indirect Potable Reuse versus Direct Potable Reuse What's the Difference. Chandler, Arizona, WaterReuse Association.

Chen, Z., Ngo, H. & Guo, W., 2013. Risk Control in Recycled Water Schemes. *Environmental Science and Technology*, 43(22), pp. 2439-2510.

De Boor, C., 1978. *A Practical Guide to Splines*, Springer-Verlag, 1978

Dickenson, E.R.V., Drewes, J.E., Sedlak, D.L., Wert, E.C. and Snyder, S.A., 2009. Applying Surrogates and Indicators to Assess Removal Efficiency of Trace Organic Chemicals during Chemical Oxidation of Wastewaters. *Environmental Science and Technology*, 43(16), pp. 6242-6247.

Dickenson, E., Drewes, J. E., Snyder, S. A. & Sedlak, D., 2011. Indicator Compounds: An Approach for Using Monitoring Data to Quantify the Occurrence and Fate of Wastewater-Derived Contaminants in Surface Waters. *Water Research*, Volume 43, pp. 1199-1212.

Drewes, J., Sedlak, D., Snyder, S. & Dickenson, E., 2008. *Development of Indicators and Surrogates for Chemical Contaminant Removal during Wastewater Treatment and Reclamation*, Alexandria, VA, USA: WaterReuse Foundation.

Drewes, J., Anderson, P., Denslow, N., Olivieri, A., Schlenk, D., Snyder, S., 2010. *Final Report Monitoring Strategies for Chemicals of Emerging Concern (CECs) in Recycled Water – Recommendations of a Science Advisory Panel.*, Sacramento, California: California State Water Resources Control Board.

Du Pisani, P., 2006. Direct reclamation of potable water at Windhoek's Goreangab reclamation plant. *Elsevier*, Volume 188, pp. 79-88.

Genthe, B. & Kfir, R., 1992. *Studies on Microbiological Drinking Water Quality Guidelines*, Pretoria, South Africa: Water Research Commission.

Georgakopoulos, D. & Yang, W., 2001. Circuit noise reduction by analogue lowpass filtering and data averaging. *Electronics Letters*, 37(19), pp. 1147-1148.

Gerrity, D., Gamage, S., Jones, D., Korshin, G.V., Lee, Y., Pisarenko, A., Trenholm, R.A., von Gunter, U., Wert, E.C. and Snyder, S.A., 2012. Development of surrogate correlation models to predict trace organic contaminant oxidation and microbial inactivation during oxonation. *Water Research*, Volume 47, pp. 6257-6272.

Gerrity, D., Pecson, B., Trussell, R. & Trussell, R., 2013. Potable Reuse Treatment Trains throughout the World. *Water Supply: Research and Technology-AQUA*, 62(6), pp. 321-338.

Ge, X., 2011. Exploratory Data Analysis, San Diego: Department of Mathematics and Statistics, San Diego State University.

Haddad, B., Rozin, P., Nemerhoff, C. & Slovic, P., 2009. The Psychology of Water Reclamation and Reuse, Alexandria, Virginia: WateReuse Foundation.

Haimi, H., Mulas, M., Corona, F. & Vahala, R., 2013. Data-derived soft-sensors for biological wastewater treatment plants: An Overview. Environmental Modelling & Software, Volume 47, pp. 88-107.

Her, N., Amy, G., Foss, D. & Cho, J., 2002. Variations of molecular weight estimation by HP-size exclusion chromatography with UVA versus online DOC detection. Environmental Science and Technology, 36(15), pp. 3393-3399.

Hervè, A., 2007. Partial Least Square Regression, Richardson, TX: The University of Texas at Dallas.

Ivarsson, O.O.A., 2011. Risk assessment for South Africa's first direct wastewater reclamation system for drinking water production, Göteborg, Sweden: Chalmers University of Technology.

IWA, International Water Association., 2014. Water Recycling and Reuse: Potential, Safety and Best Practices. [Online]

Available at:

<http://www.iwahq.org/contentsuite/upload/iwa/Document/Water%20Reuse%20IWA%20%20SG%20flyer.pdf>

[Accessed 26 March 2014].

Jing, T., Shuyin, Z., Guangxin, Z., Dibo, H., Pingjie, H. and Jian, Z., 2012. New Design for Water Quality Early Warning Systems, Hangzhou, China: Department of Control Science & Engineering, Zhejiang University.

Kadlec, P., Gabrys, B. & Strandt, S., 2008. Data-driven Soft Sensors in the Process Industry. Computers and Chemical Engineering, p. doi:10.1016/j.compchemeng.2008.12.012.

Khan, S., 2013. Drinking water through recycling: The benefits and costs of supplying direct to the distribution system, Melbourne: Australian academy of technological sciences and engineering.

Lazarova, V., Asano, T., Bahri, A. & Anderson, J., 2013. Milestones in Water Reuse. 1st ed. London, UK: IWA Publishing.

Leverenz, H. L., Tchobanoglous, G. & Asano, T., 2011. Direct potable reuse: a future imperative. Journal of Water Reuse and Desalination, 1(1), pp. 2-10.

Levine, A., Leverenz, H. & Asano, T., 2014. Encyclopedia of Life Support Systems. [Online]

Available at: <http://www.eolss.net/sample-chapters/c03/e2-20a-06-00.pdf>

[Accessed 02 April 2014].

Lin, B., Recke, B., Knudsen, J. K. & Jørgensen, S. B., 2007. A systematic approach for soft sensor development. *Computers and Chemical Engineering*, Volume 31, pp. 419-425.

Mancha, E., 2013. Texas Water Development Board. [Online]  
Available at: <http://www.twdb.texas.gov/publications/shells/WaterReuse.pdf>  
[Accessed 08 July 2014].

Marais, P., 2012. Beaufort West water reclamation plant. Umhlanga, KwaZulu-Natal, Water Institute of South Africa.

McKnight, D.M., Boyer, E.W., Westerhoff, P.K., Doran, P.R., Kuble, T. And Andersen, D.T., 2001. Spectrofluorometric characterization of dissolved organic matter for indication of precursor organic material and aromaticity. *Limnology and Oceanography*, 46(1), pp. 38-48.

Naismith, J., 2005. Membrane integrity - Direct turbidity measurement of filtrate from MF membrane modules at an operating potable water treatment plant. *Desalination*, 179(1), pp. 25-30.

Miles, S.L., Sinclair, R.G., Riley, M.E. and Pepper, I.L., 2011, Evaluation of Select Sensors for Real-Time Monitoring of *Escherichia coli* in Water Distribution Systems. *Applied and Environmental Microbiology*, Volume 77, pp. 2813 – 2816. American Society for Microbiology

Napier-Munn, T.J., 2014, *Statistical Methods for Mineral Engineers: How to design experiments and analyse data*. Julius Kruttschnitt Mineral Research Centre, Indooroopilly, Queensland

National Academy of Sciences, 2012. National Academy of Sciences. [Online]  
Available at: <http://nas-sites.org/waterreuse/what-is-water-reuse/types-of-water-reuse/>  
[Accessed 31 July 2013].

NIST-Sematech, 2012. National Institute of Standards and Technology. [Online]  
Available at: <http://www.itl.nist.gov/div898/handbook/>  
[Accessed 14 November 2015].

Nor, N., Hussain, M. & Hassan, C., 2015. Process monitoring and fault detection in non-linear chemical process based on multi-scale Kernel Fisher Discriminant Analysis. Copenhagen, Elsevier B. V.

NRC, National Research Council., 2012. *Water Reuse: Potential for Expanding the Nation's Water Supply Through Reuse of Municipal Wastewater*, Washington DC: National Academy of Sciences.

NRMMC, Natural Resource Management Ministerial Council., 2006. *Australian Guidelines for Water Recycling: Managing Health and Environmental Risks (Phase 1)*, Canberra, Australia: Environment Protection and Heritage Council, the Natural Resource Management Ministerial Council and the Australian Health Ministers' Conference.

NRMMC, Natural Resource Management Ministerial Council., 2008. *Australian Guidelines for Water Recycling: Augmentation of Drinking Water Supplies*, Canberra, Australia: Environment Protection and



Heritage Council, the National Health and Medical Research Council and the Natural Resource Management Ministerial Council.

NWRI, National Water Research Institute., 2012. BDOC as a Performance Measure for Organic Removal in Groundwater Recharge of Recycled Water, Fountain Valley, California: National Water Research Institute.

Pigott, T., 2001. A Review of Methods for Missing Data. *Educational Research and Evaluation*, 7(4), pp. 353-383.

Piovoso, M., Kosanovich, K. & Pearson, R., 1992. *Monitoring Process Performance in Real-Time*. Chicago, E.I. Du Pont de Nemours & Company (Inc.).

Rajput, S. & Rajput, S., 2006. Signal preserving seismic interference noise attenuation on 3D marine seismic data. New Orleans, Society of Exploration Geophysicists.

Rice, J., Wutich, A. & Westerhoff, P., 2013. Assessment of De Facto Wastewater Reuse across the U.S.: Trends between 1980 and 2008. *Environmental Science & Technology*, Volume 47, pp. 11099-11105.

Rodriguez, C., Van Buynder, P., Lugg, R., Blair, B., Devine, B., Cook, A. and Weinstein, P., 2009. Indirect Potable Reuse: A Sustainable Water Supply Alternative. *International Journal of Environmental Research and Public Health*, Volume 6, pp. 1174-1209.

Rosén, C., 1998. *Monitoring Wastewater Treatment Systems*, Sweden: Universitetsstryckeriet, Lund University.

Rosén, C., 2001. *A chemometric approach to process monitoring and control*, Lund, Sweden: Department of Industrial Electrical Engineering, Lund University.

Seltman, H., 2015. *Experimental Design and Analysis*. Pittsburg: Carnegie Mellon University.

Shrestha, S. & Kazama, F., 2007. Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji river basin, Japan. *Environmental Modelling & Software*, Volume 22, pp. 464-475.

Snyder, S., Westerhoff, P., Yoon, Y. & Sedlak, D., 2003. Pharmaceuticals, Personal Care Products, and Endocrine Disruptors in Water: Implications for the Water Industry. *Environmental Engineering Science*, 20(5), pp. 449-469.

Snyder, S.A., Wert, E.C., Lei, H., Westerhoff, P. and Yoon, Y., 2007. *Removal of EDC's and Pharmaceuticals in Drinking and Reuse Treatment Processes*, Denver: American Water Works Association Research Foundation (AwwaRF).

Soley-Bori, M., 2013. *Dealing with missing data: Key assumptions and methods for applied analysis*, Boston, USA: Boston University School of Public Health.

Storey, M. V., van der Gaag, B. & Burns, B. P., 2011. Advances in on-line drinking water quality monitoring and early warning systems. *Water Research*, 45(2), pp. 741-747.

Swartz, C.D., Morrison, I.R., Thebe, T., Engelbrecht, W.J., Cloete, V.B., Knott, M., Loewenthal, R.E. and Krüger, P., 2003. Characterisation and chemical removal of organic matter in South African coloured surface waters, Pretoria, South Africa: Water Research Commission.

Swartz, C.D., Genthe, B., Menge, J.G., Coomans, C.J., Offringa, G., 2015. Guidelines for monitoring, management and communication of water quality in the direct reclamation of municipal wastewater for drinking purposes, Pretoria, South Africa: Water Research Commission.

Swartz, C.D., Genthe, B., Petrik, L.F., Tijani, J.O., Adeleye, A., Coomans, C.J., Ohlin, A., Falk, D. and Menge, J.G., 2016. Emerging contaminants in wastewater treated for direct potable re-use: The human health risk priorities in South Africa, Pretoria: Water Research Commission.

Tani, W., Suehiro, E., Nishii, T., Kono, A., Negi, N., Takahashi, S., Kawamitsu, H., Sugimura, K. and Kobe, J.P., 2015. Electronic Presentation Online System. [Online]  
Available at: <http://dx.doi.org/10.1594/ecr2015/C-0856>  
[Accessed 30 October 2015].

TCEQ, Texas Commission on Environmental Quality., 1997. Texas Administrative Code. [Online]  
Available at: [http://info.sos.state.tx.us/pls/pub/readtac\\$ext.ViewTAC?tac\\_view=3&ti=30&pt=1](http://info.sos.state.tx.us/pls/pub/readtac$ext.ViewTAC?tac_view=3&ti=30&pt=1)  
[Accessed 08 July 2014].

Tchobanoglous, G., Leverenz, H., Nellor, M. & Crook, J., 2011. Direct Potable Reuse: A Path Forward, Alexandria: WaterReuse Research Foundation and WaterReuse California.

Tukey, J., 1977. *Exploratory Data Analysis*. 1st ed. Reading, Massachusetts: Addison-Wesley.

Tu, K.L., Fujioka, T., Khan, S.J., Poussade, Y., Roux, A., Drewes, J.D., Chivas, A.R. and Nghiem, L.D., 2013. Boron as a Surrogate for N Nitrosodimethylamine Rejection by Reverse Osmosis Membranes in Potable Water Reuse Applications. *Environmental Science and Technology*, Volume 47, pp. 6425-6430.

USEPA, United States Environmental Protection Agency., 2012. Guidelines for Water Reuse, Washington, D.C.: EPA: United States Environmental Protection Agency.

USEPA, United States Environmental Protection Agency, 2015. Review of Coliphages as Possible Indicators of Faecal Contamination for Ambient Water Quality, Washington, DC: USEPA, United States Environmental Protection Agency.

WHO, World Health Organization, 1997. *Guidelines for Drinking-water Quality*. 2<sup>nd</sup> edition, Geneva, Switzerland.

WHO, World Health Organization, 2011. Guidelines for Drinking-water Quality. 4<sup>th</sup> edition, Geneva, Switzerland.

Xiong, H., Pandey, G., Steinbach, M. & Kumar, V., 2006. Enhancing data analysis with noise removal. Knowledge and Data Engineering, 18(3), pp. 304-319.

## APPENDIX A: DETAILED INFORMATION FOR ALL VARIABLES INCLUDED IN THE STUDY

Table A1: Details for all variables included in statistical analyses

ID	Sampling Point	Analysis	Distribution	Lower limit	Upper limit	LDA limit
1	WWTP Clarifier	Nitrate	Normal	-4.4	21.5	11.8
2	WWTP Clarifier	EC	Empirical	89.1	210.0	160.6
3	WWTP Clarifier	pH	Empirical	7.2	8.4	8.1
4	WWTP Clarifier	Temperature	Empirical	11.2	25.4	23.9
5	WWTP Clarifier	Ammonia	Empirical	0.0	9.5	2.5
6	WWTP Clarifier	Total solids (TS:105°)	Empirical	537.5	1363.0	1007.0
7	WWTP Clarifier	COD	Lognormal	10.8	128.2	72.5
8	WWTP Clarifier	Nitrite	Normal	-0.9	2.2	1.0
9	WWTP Clarifier	TDS (TDS:180°)	Empirical	500.0	1316.6	968.9
10	WWTP Clarifier	TKN	Lognormal	0.3	22.5	5.6
11	WWTP Clarifier	TSS	Empirical	10.0	169.6	90.0
12	WWTP Clarifier	Orthophosphate	Normal	-1.7	5.4	2.8
13	WWTP Clarifier	Chlorophyll A	Lognormal	0.0	89.0	10.9
14	WWTP Clarifier	DOC	Lognormal	5.3	14.1	10.0
15	WWTP Clarifier	Total Alkalinity	Lognormal	133.4	347.0	268.7
16	WWTP Clarifier	Clostridium Viable	Lognormal	1824.3	212698.2	52865.8
17	WWTP Clarifier	E Coli (confirmed) Tryptone	Lognormal	1050.4	2580821.2	604937.3
18	WWTP Clarifier	Faecal coliform	Lognormal	1497.3	2538135.5	629970.9
19	WWTP Clarifier	Somatic coliphage	Lognormal	569.4	239045.2	38339.1
20	WWTP MP B1	Ammonia	Lognormal	0.2	22.9	4.6
21	WWTP MP B4	Ammonia	Lognormal	0.2	18.3	3.6
22	WWTP MP B6	Ammonia	Lognormal	0.3	8.7	3.1
23	WWTP MP B8	Ammonia	Lognormal	0.1	20.2	3.3
24	WWTP Final Eff.	pH	Empirical	7.3	8.8	8.2
25	WWTP Final Eff.	Temperature	Empirical	9.6	27.5	41.8
26	WWTP Final Eff.	EC	Empirical	63.4	225.0	1311.8
27	WWTP Final Eff.	Turbidity	Empirical	0.9	23.7	10.5
28	WWTP Final Eff.	Ammonia	Lognormal	0.1	4.7	1.2
29	WWTP Final Eff.	COD	Empirical	11.9	86.0	54.1
30	WWTP Final Eff.	UV 254	Empirical	0.0	0.3	0.3
31	WWTP Final Eff.	Nitrate	Normal	-3.7	20.5	11.5
32	WWTP Final Eff.	DOC	Lognormal	5.5	13.8	10.0
33	WWTP Final Eff.	Total Alkalinity	Normal	110.5	309.6	257.2
34	WWTP Final Eff.	Calcium hardness	Empirical	80.8	305.0	178.7
35	WWTP Final Eff.	Magnesium hardness	Empirical	29.0	223.5	131.4
36	WWTP Final Eff.	TKN	Lognormal	0.3	16.0	4.6
37	WWTP Final Eff.	Nitrite	Empirical	0.1	1.7	0.6
38	WWTP Final Eff.	Orthophosphate	Empirical	0.2	4.2	2.6

ID	Sampling Point	Analysis	Distribution	Lower limit	Upper limit	LDA limit
39	WWTP Final Eff.	TDS (Calc)	Empirical	541.6	1407.0	1069.2
40	WWTP Final Eff.	Total hardness	Empirical	129.2	583.9	302.2
41	WWTP Final Eff.	Iron (Fe)	Lognormal	0.0	0.5	0.2
42	WWTP Final Eff.	Manganese (Mn)	Empirical	0.0	2.0	0.4
43	WWTP Final Eff.	Total solids (TS:105°)	Empirical	583.5	1366.0	1023.9
44	WWTP Final Eff.	TDS (TDS:180°)	Empirical	563.6	1306.4	980.2
45	WWTP Final Eff.	TSS	Empirical	10.0	404.8	89.4
46	WWTP Final Eff.	Chlorophyll A	Lognormal	0.1	294.5	44.7
47	WWTP Final Eff.	Fixed suspended solids	Empirical	0.0	313.2	37.8
48	WWTP Final Eff.	TSS	Lognormal	0.2	130.1	46.2
49	WWTP Final Eff.	VSS	Empirical	10.0	90.8	20.6
50	WWTP Final Eff.	Clostridium Spores	Lognormal	377.0	25319.2	5947.3
51	WWTP Final Eff.	Clostridium Viable	Lognormal	1002.6	28181.0	8953.5
52	WWTP Final Eff.	Faecal coliform	Lognormal	35.8	78414.0	7771.6
53	WWTP Final Eff.	Faecal streptococci	Lognormal	1.7	10626.9	896.7
54	WWTP Final Eff.	Total coliform	Lognormal	117.8	1106268.6	69040.1
55	WWTP Final Eff.	HPC	Lognormal	4049.1	1196792.2	207818.5
56	WWTP Final Eff.	Pseudomonas aeruginosa	Lognormal	8.9	27762.9	10757.7
57	WWTP Final Eff.	E Coli (confirmed) Tryptone	Lognormal	27.5	76680.5	7068.2
58	WWTP Final Eff.	Somatic coliphage	Lognormal	37.5	79716.3	8062.2
59	WWTP Final Eff.	Chloride (Cl)	Lognormal	61.4	335.6	188.8
60	WWTP Final Eff.	Fluoride (F)	Lognormal	0.2	1.1	0.6
61	WWTP Final Eff.	Sulphate (SO <sub>4</sub> )	Lognormal	49.3	546.1	230.0
62	WWTP Final Eff.	Potassium (K)	Normal	12.2	40.8	33.1
63	WWTP Final Eff.	Sodium (Na)	Empirical	NaN	NaN	219.8
64	WWTP Final Eff.	Total Algal Count	Lognormal	227.2	9188.3	3628.9
65	WWTP Final Eff.	Total BG Algae	Empirical	NaN	NaN	36.6
66	WWTP Final Eff.	Colour	Lognormal	21.6	151.5	81.2
67	WWTP Final Eff.	TOC	Lognormal	3.1	22.4	10.2
68	WRP Inf.	pH	Empirical	7.4	9.0	8.4
69	WRP Inf.	Temperature	Empirical	9.0	28.1	42.9
70	WRP Inf.	Turbidity	Empirical	0.8	6.9	4.0
71	WRP Inf.	EC	Empirical	81.5	211.8	161.4
72	WRP Inf.	COD	Lognormal	12.9	51.3	33.6
73	WRP Inf.	DOC	Lognormal	4.4	12.1	8.5
74	WRP Inf.	Total Alkalinity	Lognormal	131.3	326.4	252.8
75	WRP Inf.	TDS (Calc)	Empirical	583.8	1419.6	1079.5
76	WRP Inf.	UV 254	Empirical	0.1	0.4	0.2
77	WRP Inf.	Calcium hardness	Empirical	84.7	292.0	176.3
78	WRP Inf.	Magnesium hardness	Empirical	24.2	296.1	131.5
79	WRP Inf.	Total hardness	Empirical	124.2	577.0	297.0
80	WRP Inf.	Iron (Fe)	Empirical	0.0	637.4	70.6
81	WRP Inf.	Manganese (Mn)	Empirical	0.0	0.3	0.1

ID	Sampling Point	Analysis	Distribution	Lower limit	Upper limit	LDA limit
82	WRP Inf.	Ammonia	Empirical	0.2	1.8	0.7
83	WRP Inf.	Fixed suspended solids	Empirical	0.0	300.7	37.1
84	WRP Inf.	Orthophosphate	Lognormal	0.1	11.9	2.0
85	WRP Inf.	TSS	Empirical	0.0	310.1	41.9
86	WRP Inf.	VSS	Empirical	10.0	29.0	15.7
87	WRP Inf.	Nitrate	Lognormal	1.5	37.5	11.4
88	WRP Inf.	Nitrite	Empirical	0.1	1.7	0.5
89	WRP Inf.	TKN	Lognormal	0.2	14.5	18.8
90	WRP Inf.	Chlorophyll A	Empirical	0.0	287.9	44.7
91	WRP Inf.	Clostridium Spores	Normal	-1045.1	6059.7	4132.8
92	WRP Inf.	Clostridium Viable	Normal	-1122.9	9098.0	6582.3
93	WRP Inf.	Faecal coliform	Lognormal	3.5	189240.4	5341.2
94	WRP Inf.	Faecal streptococci	Lognormal	0.4	12026.4	579.8
95	WRP Inf.	Pseudomonas aeruginosa	Lognormal	4.4	23247.7	4737.2
96	WRP Inf.	Total coliform	Lognormal	13.4	3298120.2	65420.3
97	WRP Inf.	HPC	Empirical	64.2	663900.0	206724.4
98	WRP Inf.	E Coli (confirmed) Tryptone	Lognormal	2.1	232400.2	4810.7
99	WRP Inf.	Somatic coliphage	Lognormal	2.3	251006.4	5641.1
100	WRP Inf.	Chloride (Cl)	Lognormal	67.3	353.5	202.8
101	WRP Inf.	Colour	Lognormal	10.5	193.5	75.1
102	WRP Inf.	Fluoride (F)	Lognormal	0.1	1.3	0.6
103	WRP Inf.	Sulphate (SO4)	Lognormal	54.2	442.7	217.8
104	WRP Inf.	TOC	Lognormal	3.8	13.4	8.6
105	WRP Inf.	Potassium (K)	Lognormal	18.0	38.9	32.2
106	WRP Inf.	Sodium (Na)	Lognormal	115.1	305.3	224.0
107	DAF	Turbidity	Lognormal	0.4	2.0	1.4
108	DAF	pH	Empirical	6.6	7.6	10.4
109	DAF	Temperature	Empirical	11.6	27.4	24.8
110	DAF	EC	Empirical	69.4	1413.4	288.9
111	DAF	COD	Empirical	8.6	32.4	24.8
112	DAF	DOC	Lognormal	2.3	8.2	5.3
113	DAF	TDS (Calc)	Empirical	640.0	1474.0	1147.0
114	DAF	UV 254	Empirical	0.1	0.3	0.2
115	DAF	Calcium hardness	Empirical	54.1	270.0	168.9
116	DAF	Total Alkalinity	Lognormal	73.2	289.5	189.3
117	DAF	Magnesium hardness	Empirical	24.4	551.3	142.0
118	DAF	Total hardness	Empirical	103.4	786.4	298.1
119	DAF	Chlorophyll A	Lognormal	0.0	18.4	2.2
120	DAF	Iron (Fe)	Empirical	0.0	29.3	6.6
121	DAF	Manganese (Mn)	Empirical	0.0	1.9	0.4
122	DAF	Total Algal Count	Lognormal	31.3	781.5	331.2
123	DAF	Total BG Algae	Lognormal	NaN	NaN	0.0
124	SF	Turbidity	Empirical	0.1	0.3	0.3

ID	Sampling Point	Analysis	Distribution	Lower limit	Upper limit	LDA limit
125	SF	pH	Empirical	7.3	8.8	11.3
126	SF	Temperature	Empirical	10.3	27.4	24.9
127	SF	COD	Empirical	7.5	28.4	22.1
128	SF	EC	Empirical	98.5	221.9	175.7
129	SF	DOC	Lognormal	1.8	7.5	4.6
130	SF	TDS (Calc)	Empirical	660.0	1486.6	1177.4
131	SF	UV 254	Empirical	0.0	0.2	0.1
132	SF	Iron (Fe)	Empirical	0.0	0.2	0.2
133	SF	Manganese (Mn)	Empirical	0.0	0.5	0.1
134	SF	Total Alkalinity	Lognormal	109.0	281.3	216.4
135	SF	Calcium hardness	Empirical	83.0	266.6	167.1
136	SF	Magnesium hardness	Empirical	27.3	531.5	138.1
137	SF	Total hardness	Empirical	120.1	759.8	294.0
138	SF	Clostridium Spores	Empirical	0.0	744.6	96.0
139	SF	Clostridium Viable	Lognormal	0.9	80.7	30.5
140	SF	Faecal coliform	Empirical	0.0	170.7	23.5
141	SF	Faecal streptococci	Empirical	0.0	513.5	61.5
142	SF	Pseudomonas aeruginosa	Empirical	0.0	3354.6	423.6
143	SF	Total coliform	Lognormal	0.2	1501.8	340.5
144	SF	Chlorophyll A	Empirical	0.0	2.4	0.5
145	SF	HPC	Lognormal	79.0	115309.6	17497.8
146	SF	E Coli (confirmed) Tryptone	Empirical	1.0	172.8	26.9
147	SF	Somatic coliphage	Empirical	0.0	678.8	95.6
148	SF	Total Algal Count	Empirical	NaN	NaN	170.1
149	SF	Total BG Algae	Lognormal	NaN	NaN	0.0
150	Ozone Contact A	Free chlorine	Empirical	0.1	2.0	1.3
151	Ozone Contact A	Residual O3	Lognormal	0.1	2.8	0.7
152	Ozone Contact B	Free chlorine	Empirical	0.1	1.9	1.2
153	Ozone Contact B	Residual O3	Lognormal	0.1	2.7	0.8
154	Ozone Contact C	Free chlorine	Lognormal	0.1	4.0	1.1
155	Ozone Contact C	Residual O3	Lognormal	0.0	2.3	0.6
156	Ozone Contact Final	pH	Empirical	6.9	8.0	7.7
157	Ozone Contact Final	Temperature	Empirical	13.9	27.6	25.2
158	Ozone Contact Final	Turbidity	Empirical	0.1	0.3	0.2
159	Ozone Contact Final	Residual O3	Lognormal	0.1	0.5	0.3
160	Ozone Contact Final	HPC	Empirical	0.0	3863.2	1968.3
161	Ozone Contact Final	COD	Lognormal	5.5	36.3	22.0

ID	Sampling Point	Analysis	Distribution	Lower limit	Upper limit	LDA limit
162	Ozone Contact Final	DOC	Lognormal	1.8	6.8	4.4
163	Ozone Contact Final	UV 254	Lognormal	0.0	0.1	0.1
164	Ozone Contact Final	Clostridium Viable	Empirical	0.0	1.5	0.6
165	Ozone Contact Final	Clostridium Spores	Empirical	0.0	0.6	0.2
166	Ozone Contact Final	Faecal coliform	Lognormal	NaN	NaN	0.0
167	Ozone Contact Final	Total coliform	Empirical	0.0	0.6	0.1
168	Ozone Contact Final	Faecal streptococci	Lognormal	NaN	NaN	0.0
169	Ozone Contact Final	Somatic coliphage	Lognormal	NaN	NaN	0.0
170	Ozone Contact Final	EC	Empirical	91.4	223.3	176.9
171	Ozone Contact Final	TDS (Calc)	Empirical	612.8	1496.4	1185.5
172	BAC	UV 254	Lognormal	0.0	0.1	0.0
173	BAC	DOC	Lognormal	1.2	5.5	3.3
174	BAC	COD	Lognormal	3.2	35.0	18.1
175	BAC	EC	Empirical	92.5	226.5	176.2
176	BAC	pH	Lognormal	6.9	8.0	7.6
177	BAC	Temperature	Normal	11.3	29.5	23.7
178	BAC	TDS (Calc)	Empirical	619.5	1517.6	1181.0
179	BAC	Turbidity	Lognormal	0.0	0.5	0.3
180	BAC	Iron (Fe)	Empirical	0.0	0.1	0.1
181	BAC	Manganese (Mn)	Empirical	0.0	0.2	0.0
182	GAC	Turbidity	Empirical	0.1	0.2	0.2
183	GAC	pH	Empirical	6.9	7.8	7.6
184	GAC	Temperature	Normal	9.3	32.9	25.0
185	GAC	Iron (Fe)	Empirical	0.0	0.1	0.0
186	GAC	Manganese (Mn)	Empirical	0.0	0.1	0.0
187	GAC	COD	Lognormal	1.8	27.5	14.9
188	GAC	EC	Empirical	98.6	221.9	176.0
189	GAC	DOC	Empirical	0.8	3.4	2.8
190	GAC	TDS (Calc)	Empirical	660.8	1486.6	1179.4
191	GAC	UV 254	Empirical	0.0	0.1	0.1
192	UF	pH	Empirical	7.0	7.8	7.6
193	UF	Temperature	Empirical	12.8	27.1	24.4
194	UF	Turbidity	Lognormal	0.0	0.2	0.1
195	UF	TDS (Calc)	Empirical	670.0	1499.8	1182.2
196	UF	EC	Empirical	100.0	223.8	176.4



ID	Sampling Point	Analysis	Distribution	Lower limit	Upper limit	LDA limit
197	UF	Total coliform	Lognormal	NaN	NaN	0.0
198	UF	Faecal coliform	Lognormal	NaN	NaN	0.0
199	UF	HPC	Empirical	0.0	315.6	323.0
200	UF	UV 254	Lognormal	0.0	0.0	0.0
201	UF	Faecal streptococci	Lognormal	NaN	NaN	0.0
202	UF	Pseudomonas aeruginosa	Empirical	0.0	20.0	2.4
203	UF	Somatic coliphage	Lognormal	NaN	NaN	0.0
204	Final Water	pH	Empirical	7.0	8.4	7.9
205	Final Water	Temperature	Empirical	11.9	27.5	42.3
206	Final Water	Turbidity	Empirical	0.0	0.1	0.2
207	Final Water	Free chlorine	Empirical	0.7	3.6	7.7
208	Final Water	Iron (Fe)	Empirical	0.0	0.2	24.9
209	Final Water	Manganese (Mn)	Empirical	0.0	0.4	0.1
210	Final Water	EC	Empirical	20.5	225.0	178.8
211	Final Water	TDS (Calc)	Empirical	704.0	1508.0	1189.8
212	Final Water	Total coliform	Lognormal	NaN	NaN	0.0
213	Final Water	Faecal coliform	Lognormal	NaN	NaN	0.0
214	Final Water	HPC	Empirical	0.0	4.9	611.2
215	Final Water	COD	Empirical	4.0	19.0	12.6
216	Final Water	DOC	Empirical	0.8	2.5	1.8
217	Final Water	Total Alkalinity	Normal	79.8	280.4	223.5
218	Final Water	UV 254	Empirical	0.0	0.0	0.0
219	Final Water	Calcium hardness	Empirical	63.2	265.7	171.3
220	Final Water	Magnesium hardness	Empirical	26.5	265.8	129.7
221	Final Water	Total hardness	Empirical	111.8	528.0	289.7
222	Final Water	Ammonia	Empirical	0.2	0.2	0.2
223	Final Water	Orthophosphate	Empirical	0.0	2.4	0.5
224	Final Water	TKN	Empirical	0.1	2.2	0.7
225	Final Water	Nitrate	Lognormal	2.0	35.0	12.2
226	Final Water	Nitrite	Empirical	0.1	1.7	0.3
227	Final Water	Chlorophyll A	Empirical	0.0	1.4	0.2
228	Final Water	Clostridium Spores	Lognormal	NaN	NaN	0.0
229	Final Water	Clostridium Viable	Lognormal	NaN	NaN	0.0
230	Final Water	Faecal streptococci	Lognormal	NaN	NaN	0.0
231	Final Water	Pseudomonas aeruginosa	Lognormal	NaN	NaN	0.0
232	Final Water	Somatic coliphage	Lognormal	NaN	NaN	0.0
233	Final Water	Total Algal Count	Lognormal	NaN	NaN	94.0
234	Final Water	Total BG Algae	Lognormal	NaN	NaN	0.0
235	Final Water	Chloride (Cl)	Lognormal	92.7	449.4	260.8
236	Final Water	Sulphate (SO4)	Lognormal	28.5	741.0	222.7
237	Final Water	TOC	Lognormal	0.4	4.6	2.8
238	Final Water	Potassium (K)	Lognormal	18.3	38.0	31.5
239	Final Water	Sodium (Na)	Empirical	NaN	NaN	244.6

ID	Sampling Point	Analysis	Distribution	Lower limit	Upper limit	LDA limit
240	Operational	Pre-Ozone (mg/l)	Empirical	2.0	5.0	5.0
241	Operational	Main Ozone A (mg/l)	Empirical	7.5	15.0	13.8
242	Operational	Main Ozone B (mg/l)	Empirical	3.0	6.0	5.2
243	Operational	FeCl3 as FeCl3 (mg/l)	Empirical	75.0	90.0	91.8
244	Operational	KMnO4 (mg/l)	Empirical	0.1	0.1	0.1
245	Operational	Main Ozone C (mg/l)	Empirical	0.0	0.5	0.4
246	Operational	Raw Water (m <sup>3</sup> /h)	Empirical	400.0	900.0	1952.9
247	Operational	Polymer (mg/l)	Empirical	0.0	0.1	0.0
248	Operational	DAF Descum Delay (min)	Empirical	1.2	120.0	128.5
249	Operational	Avg. Free Chlorine (mg/l)	Empirical	0.5	2.3	1.7
250	Operational	DAF Descum (min)	Empirical	1.2	5.0	5.1
251	Operational	Total Recycled Water (m <sup>3</sup> )	Empirical	100.9	2405.1	1862.9
252	Operational	Final Water Pumped (m <sup>3</sup> )	Empirical	1556.5	17236.9	21856.2
253	Operational	Total Water wasted (m <sup>3</sup> )	Empirical	173.2	2656.9	1981.4
254	Operational	Total Raw Water In (m <sup>3</sup> )	Empirical	1658.9	18346.6	19325.8
255	Operational	Total Oxygen production (Nm <sup>3</sup> /h)	Empirical	38.1	95.0	91.3
256	Operational	Waste vs Raw (%)	Empirical	2.4	26.7	22.6
257	Operational	Oxygen concentration: (%)	Empirical	90.1	95.0	100.8
258	Operational	Ozone concentration: (% w/w)	Empirical	12.0	12.0	12.7
259	Operational	Cl gas A (kg)	Empirical	0.0	56.9	36.2
260	Operational	Recycle water (m <sup>3</sup> /d)	Empirical	112.9	2375.4	1396.3
261	Operational	Cl gas B (kg)	Empirical	0.0	53.5	35.3
262	Operational	Avg. H2O2 (l/h)	Empirical	0.1	17.2	10.6
263	Operational	SF E Runtime (h)	Empirical	27.0	67.8	58.0
264	Operational	SF C Runtime (h)	Empirical	26.2	68.0	57.4
265	Operational	SF D Runtime (h)	Empirical	26.2	68.0	57.4
266	Operational	SF B Runtime (h)	Empirical	15.8	65.8	57.1
267	Operational	SF A Runtime (h)	Empirical	27.0	64.9	56.7
268	Operational	Membrane BW water (m <sup>3</sup> /d)	Empirical	0.0	3139.6	3068.9
269	Operational	BAC runtime (h)	Empirical	0.0	359.0	235.5
270	Operational	GAC runtime (h)	Empirical	0.0	368.2	251.4

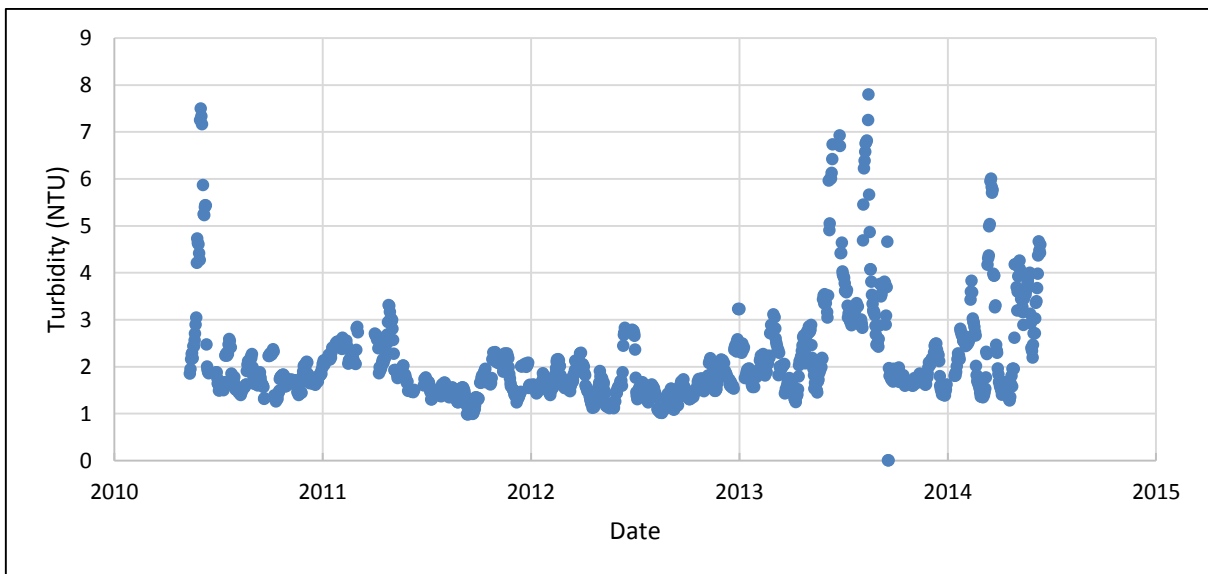
## APPENDIX B: SUFFICIENT REPRESENTATION DATA VALIDATION RESULTS

The three tests for sufficient representation (that are shown here) were performed on individual variables. Since there are so many variables, only one key variable for each of the treatment units in the plant were selected to be illustrated here.

### Moving average

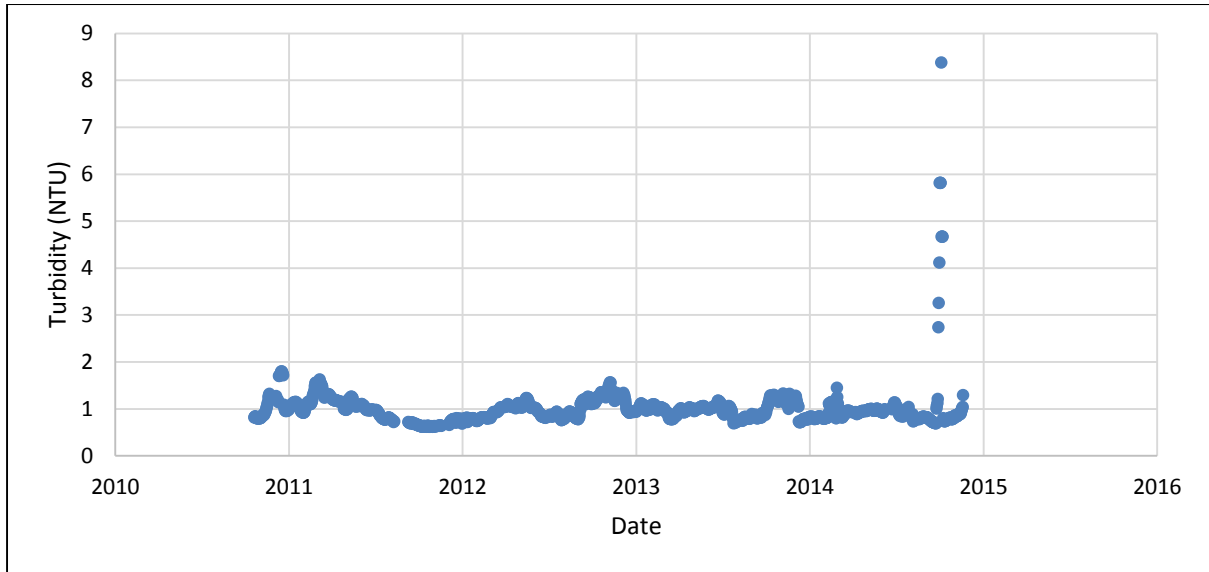
The test that will be discussed here is a simple moving average plot. From a moving average plot it is expected to see small temporal variations throughout the entire period of the data. If the plot is too flat (showing too little variation) it may be indicative of plant down times, measurement failures or erroneous sampling procedures.

On the other hand, if the moving average plot indicates a period of too great variation it is possible that there was a major upset in the treatment processes upstream of the observed process, or again, that measurements or sampling procedures were done incorrectly. In all of the cases that will be viewed here, the moving average was calculated using the average of ten data points per window.



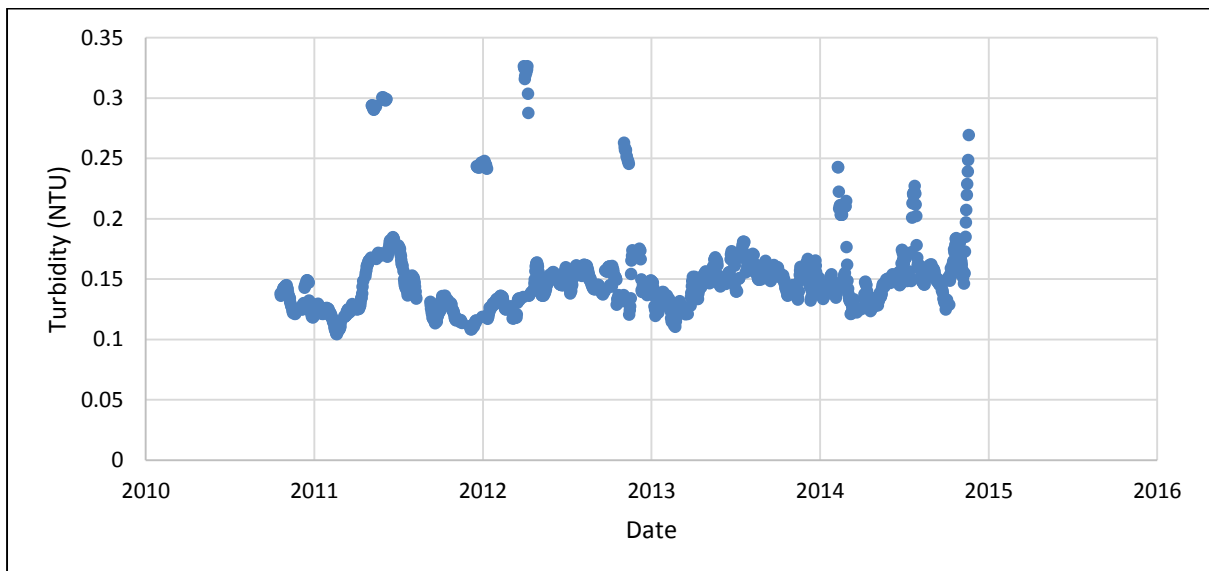
**Figure B1:** Moving average Turbidity in the WRP Influent (Swartz, *et al.*, 2016)

Figure B1 shows the moving average plot for turbidity in the influent of the WRP. The average turbidity is mainly between 1-3 NTU with an upset right at the beginning of the data period as well as two significant periods of upset near the end of the data period. Although these upsets seem extreme, they do not go above a turbidity of 8 NTU, which is still within reason considering that it is the influent to the plant.



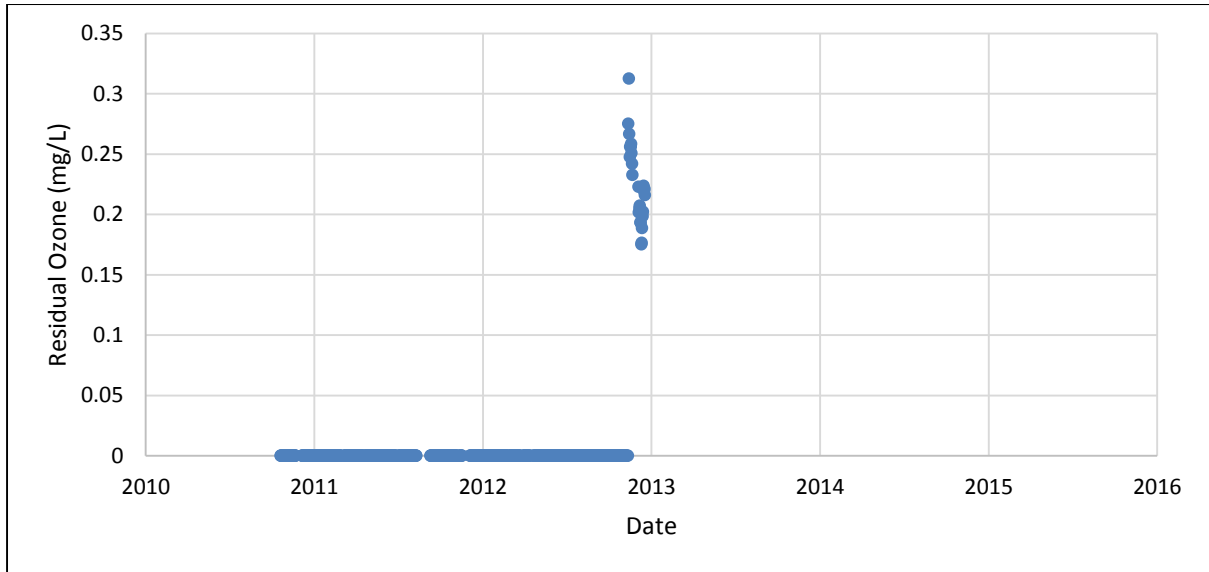
**Figure B2:** Moving average Turbidity from the DAF (Swartz, *et al.*, 2016)

From the moving average of the turbidity in the DAF effluent, seen in Figure B2, it is clear that the average turbidity have come down to between 1-2 NTU. The upsets seen in Figure B1 have clearly been dealt with correctly since they are no longer visible after the first treatment unit. There is, however, one upset near the end of the data period, reaching a turbidity higher than 8 NTU. This may be due to the prolonged upsets in the plant effluent, or a random event which is independent of the upstream units. But generally there is a sufficient constant variation in the data (Figure B2).



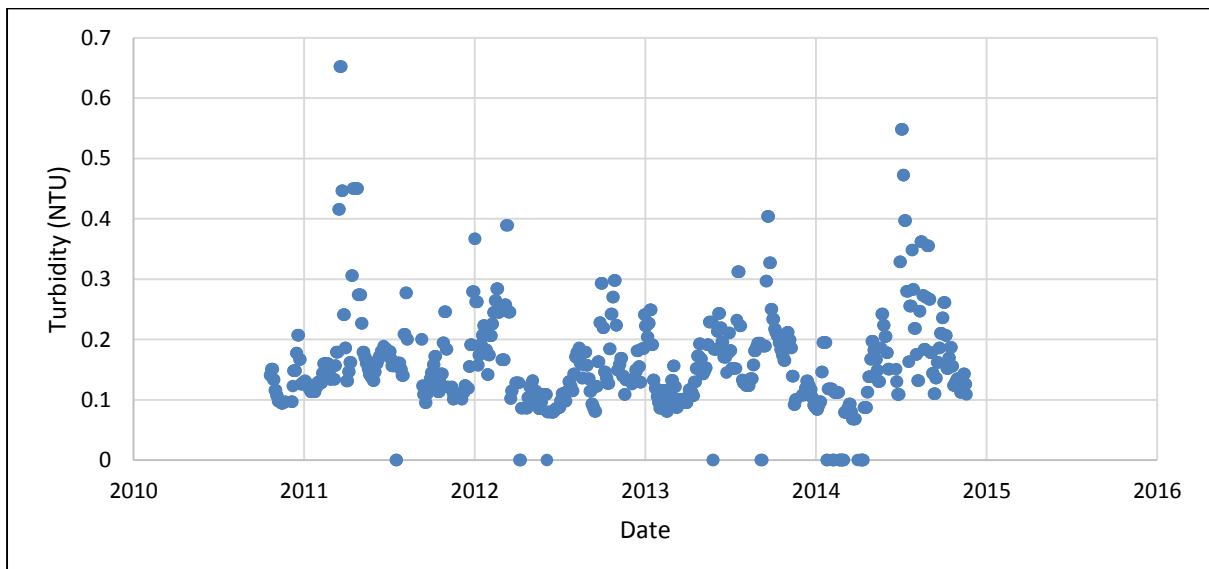
**Figure B3:** Moving average Turbidity from the SF (Swartz, *et al.*, 2016)

From Figure B3, a new order of magnitude in the turbidity levels were being achieved. The moving average remain mostly between 0.1-0.2 NTU with small upsets reaching values that are still less than 0.35 NTU. Considering that most guidelines require an NTU of less than 1 NTU, these results are satisfying and indicative of being a good representation of the actual process unit operation.



**Figure B4:** Moving average Residual Ozone from the Final Ozone Contact (Swartz, *et al.*, 2016)

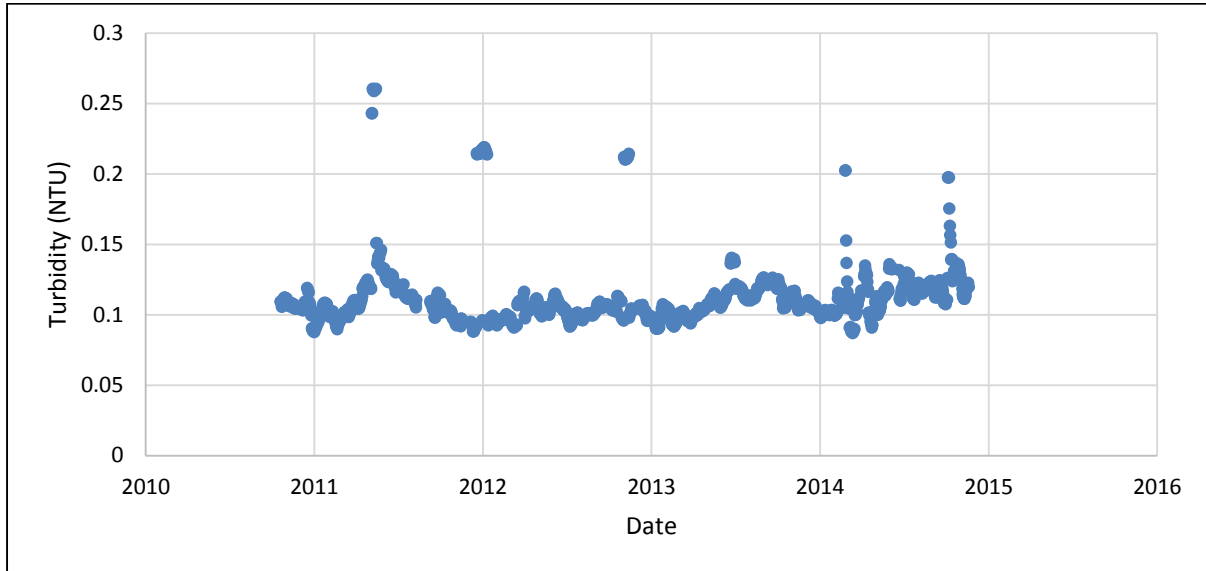
For the ozonation unit it was decided to observe the moving average of the residual ozone in the effluent of the ozonation treatment unit. From Figure B4 it can be seen that the data is not complete over the entire period that was observed in the other data. Upon further investigation it was found that the plant only started measuring residual ozone in the middle of the year 2012. Before that time, residual chlorine was measured. Since it is always better to favour the more recent data, it was decided to make use of the residual ozone rather than the residual chlorine data. There were no major upsets during the period of the data, but sufficient variation between 0.1-0.4 mg/L, which is a good representation of the actual unit operation (Figure B4).



**Figure B5:** Moving average Turbidity from the BAC (Swartz, *et al.*, 2016)

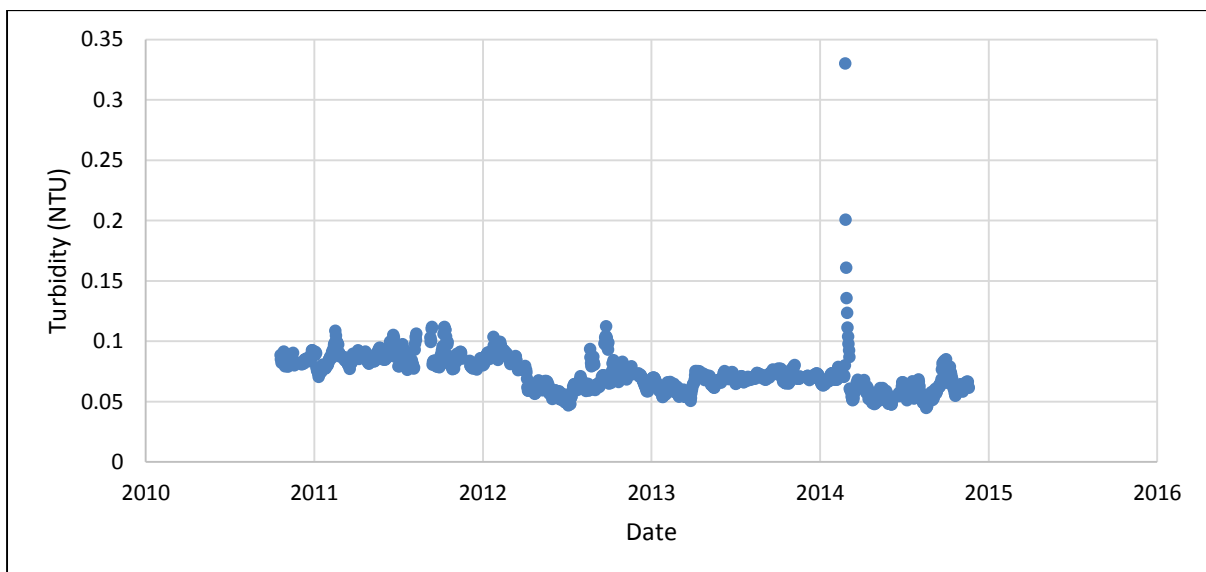
Figure B5 shows the moving average of the turbidity in the effluent of the BAC filters. The turbidity generally varies between 0.1-0.2 NTU with consistent peaks during the winter periods reaching levels as high as 0.6 NTU. The seasonal variation seen in the data can be expected from biological

treatment processes since the biological activity responsible for the performance of the process is dependent on the temperature of the water. Therefore most biological processes will perform poorer in the winter periods than the summer periods. Figure B5 is therefore indicative that the data is a good representation of the actual process operation.



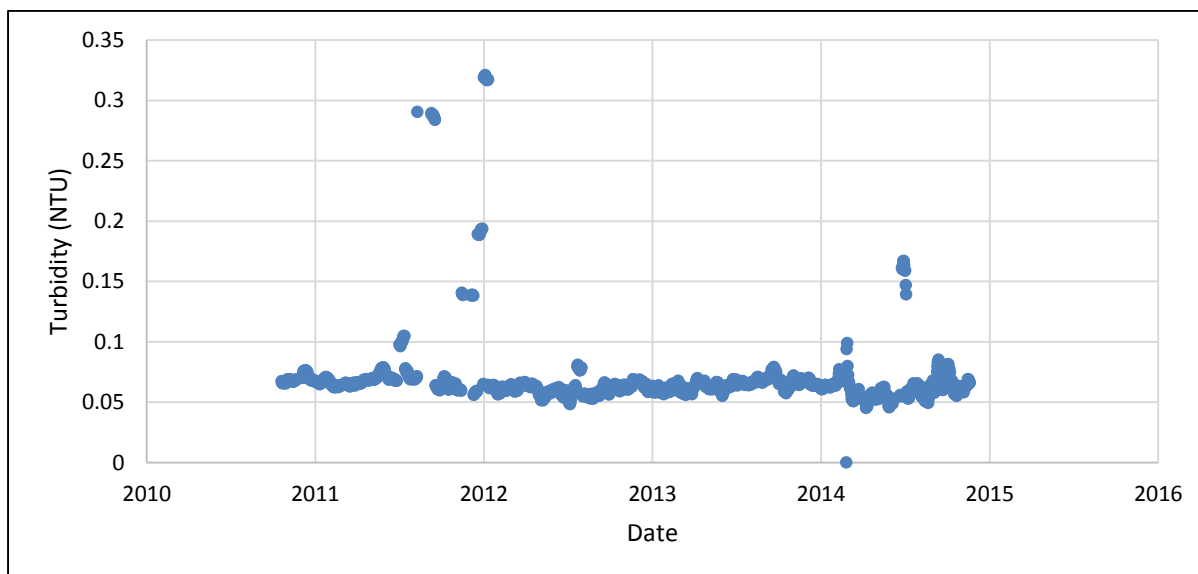
**Figure B6:** Moving average Turbidity from the GAC (Swartz, *et al.*, 2016)

The BAC is immediately followed by the GAC, but unlike the BAC, the GAC is not a biological process and is therefore not as dependent on temperature. There is no seasonality to the variation of the data (Figure B6). There are, however, a small number of data points that jump from the prevailing average of between 0.1-0.15 NTU to a value of between 0.2-0.25 NTU, but this may be due to an annual cleaning or replacing practice, rather than seasonal variation in the water temperature.



**Figure B7:** Moving average Turbidity in the UF (Swartz, *et al.*, 2016)

The UF unit is the last treatment step responsible for removing the last few contaminants that may still be present in the water. Figure B7 shows that the UF was capable of reducing the turbidity to an average below 0.1 NTU. There was one upset at the end of the year 2013 which resulted in a turbidity spike above 0.3 NTU, but this is clearly an isolated incidence. Otherwise, small variations in the data are present throughout the entire period of the data which indicates that the data is a good representation of the actual treatment process operation.

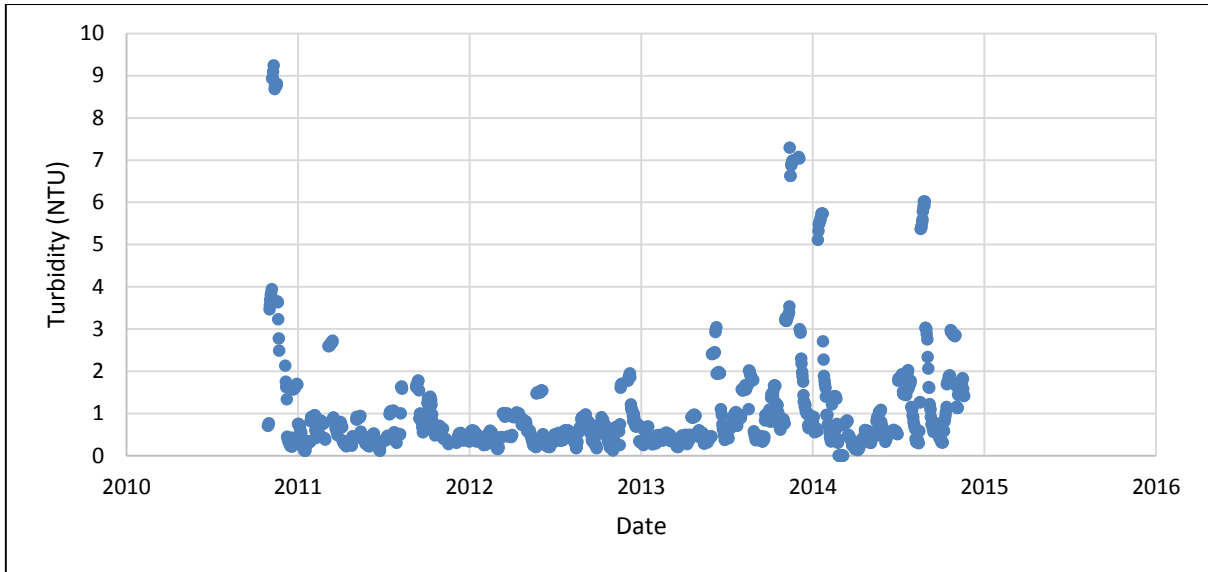


**Figure B8:** Moving average Turbidity in the Final Water (Swartz, *et al.*, 2016)

Figure B8 indicates the moving average of the turbidity in the final water produced by the WRP. The turbidity generally remains below 0.1 NTU with a few upsets resulting in a turbidity higher than 0.15 NTU. These upsets are in no way alarming since they can be caused by overdosing stabiliser during the final chemical stabilising of the water. The variation of the data is much less than the other treatment units, which is also expected and it can be accepted that the data is a good representation of the actual process operation.

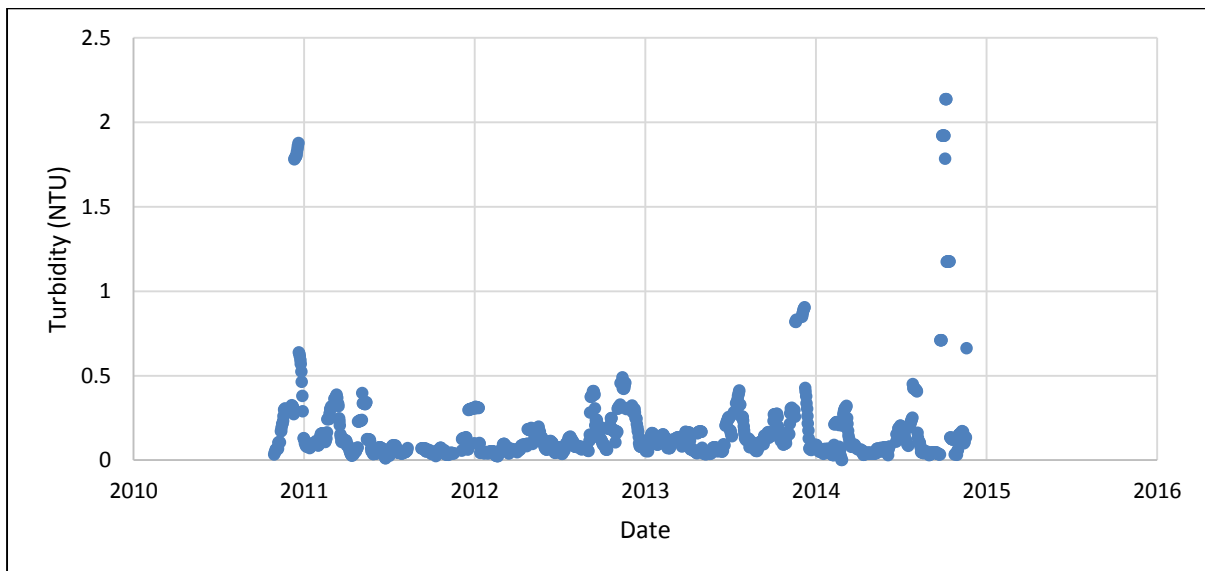
### Standard error

A standard error plot consists of plotting the standard deviation for a variable within a moving window. The closer the plot is to zero, the more constant the variable behaved. For sufficient representation it is ideal if the standard error is consistently low, but not zero. Spikes in the standard error typically indicate process upsets where the variable values show a sudden increase or decrease.



**Figure B9:** Standard error for Turbidity in the WRP Influent (Swartz, et al., 2016)

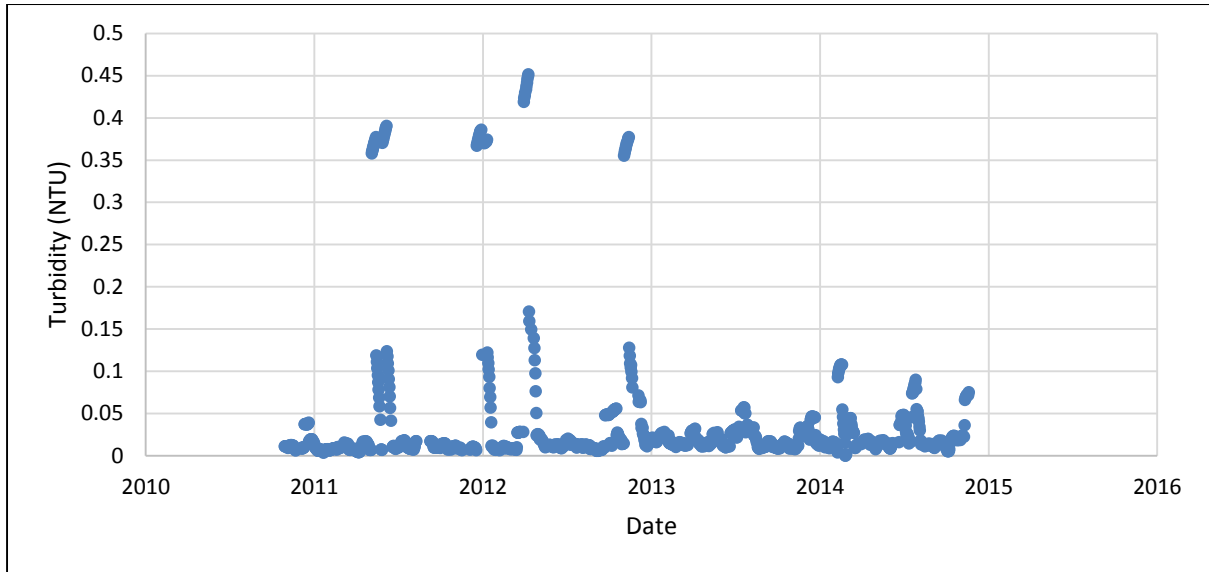
Figure B9 shows the standard error plot for turbidity in the influent of the WRP. From the plot it can be seen that there was an upset right at the beginning of the data period as well as two significant periods of upset near the end of the data period. Despite the few plant updates, the remainder of the plot indicates a standard error greater than zero, but less than one.



**Figure B10:** Standard error for Turbidity from the DAF (Swartz, et al., 2016)

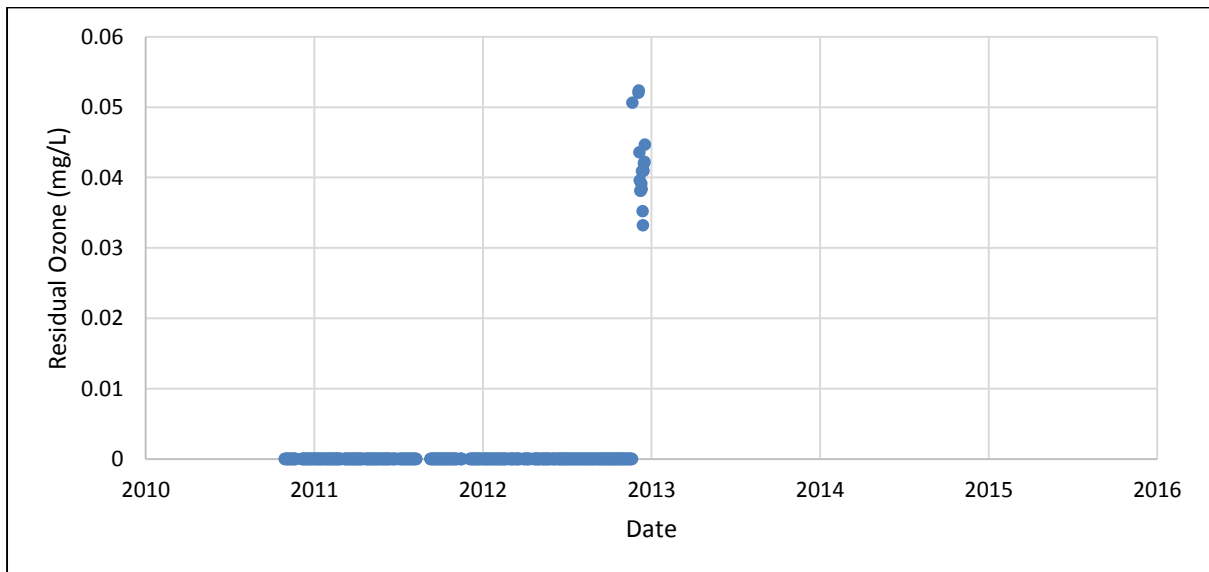
From the standard error of the turbidity in the DAF effluent (Figure B10) it can be seen that the standard error is generally below 0.5 and higher than zero. The upsets seen in Figure B9 can also be seen in this plot, although to a lesser extent. There is, however, two upsets near the beginning and end of the data period, that may indicate abnormal plant behaviour. This may be due to the prolonged upsets in the plant effluent, or a random event which is independent of the upstream units. But generally there is a sufficient constant variation in the data seen in Figure B10.





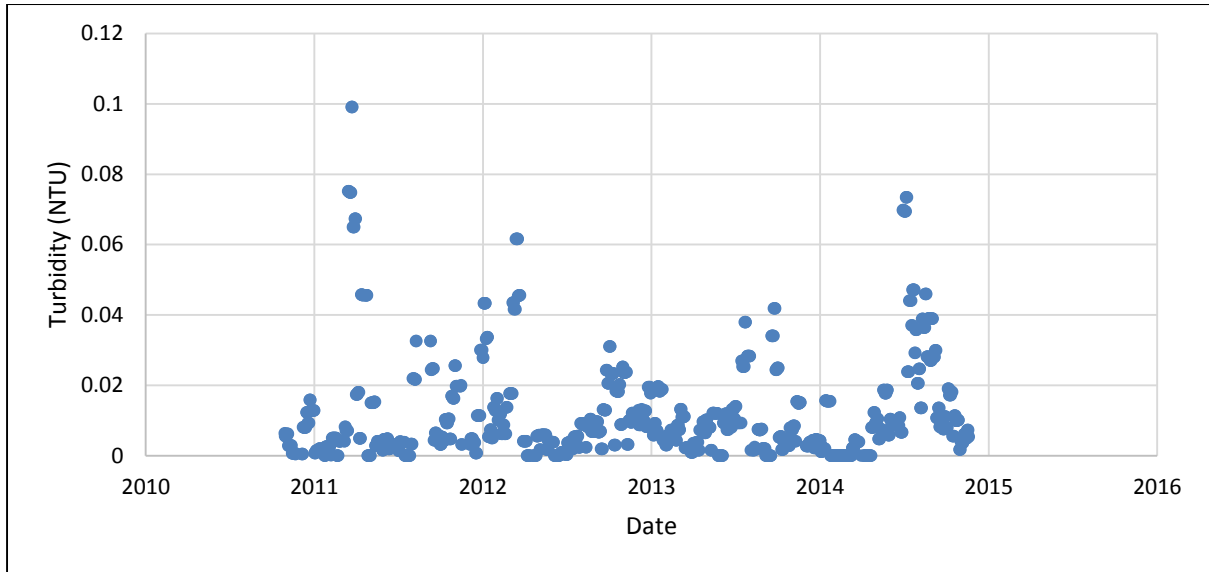
**Figure B11:** Standard error for Turbidity from the SF (Swartz, *et al.*, 2016)

Figure B11 shows that the standard error has several spikes. These increases are indicative of abnormal behaviour, but they also seem evenly spaced over the data period. It may be that the spikes are caused by some form of periodic build-up or cleaning being performed on the treatment unit. The data should undergo pre-treatment in order to determine if the spikes in the standard error plot were caused by on-site events or errors in the data itself.



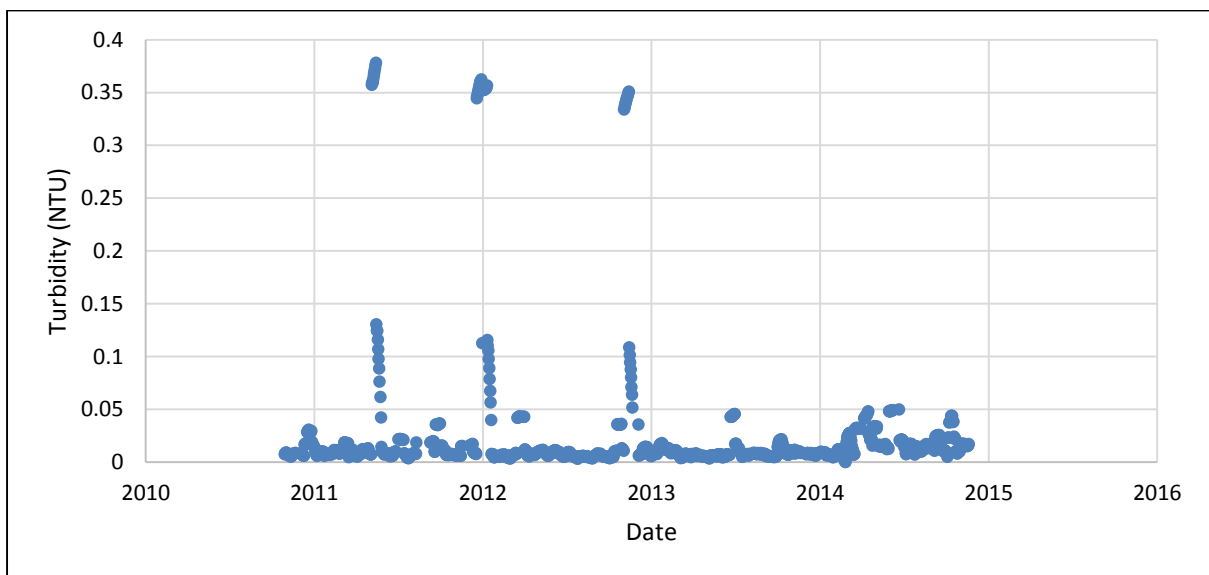
**Figure B12:** Standard error for Residual Ozone from the Final Ozone Contact (Swartz, *et al.*, 2016)

As was the case with the moving average, the data for the residual ozone in the ozonation effluent is not complete over the entire period that was observed in the other data (Figure B12). The standard error is much more variable here (Figure B12) than was the case for the previous treatment units. There was also a large spike early in the year 2013, which may indicate an upset in the behaviour of the treatment unit.



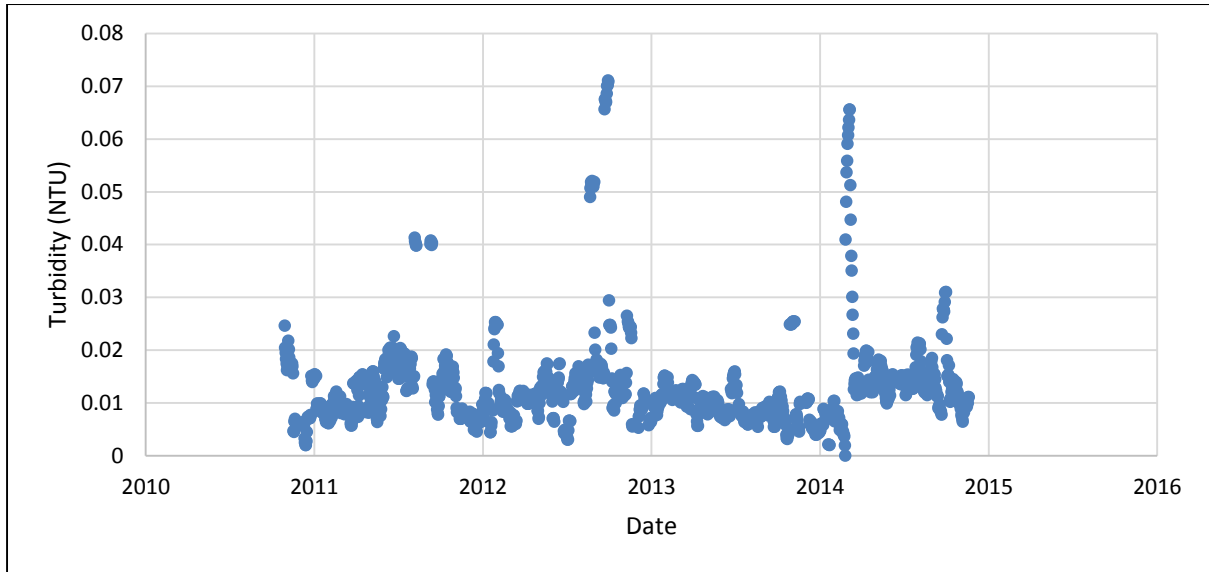
**Figure B13:** Standard error for Turbidity from the BAC (Swartz, et al., 2016)

Figure B13 shows the standard error for the turbidity in the BAC effluent. As was the case with the moving average, the data seems to have seasonal spikes during every winter period. Since it is known that the biological processes perform less efficient in colder temperatures, these spikes are not a cause for concern. Figure B13 is therefore indicative that the data is a good representation of the actual process operation.



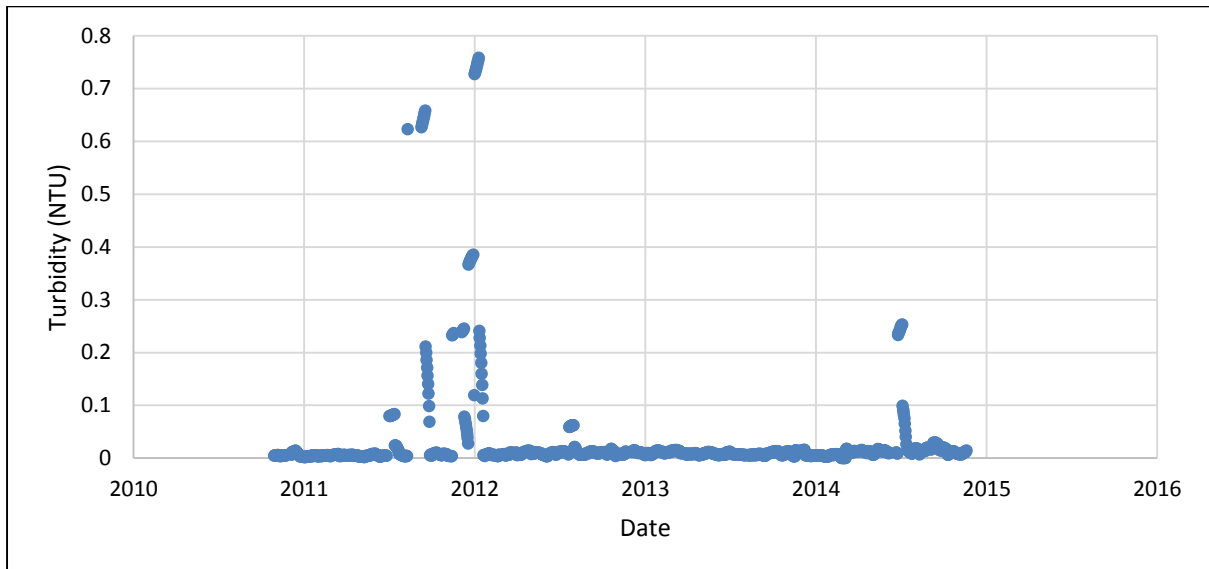
**Figure B14:** Standard error for Turbidity from the GAC (Swartz, et al., 2016)

From Figure B14 there is no seasonality to the variation of the data. There are, however, three spikes in consecutive summer periods, but not in all the summer periods. The remainder of the standard errors are not zero and show slight variations over time, which suggests that the data is a good representation of the actual treatment process.



**Figure B15:** Standard error for Turbidity from the UF (Swartz, *et al.*, 2016)

The UF unit is much more stable with its standard errors being generally less than 0.02 (Figure B15). There seems to be several spikes in the plot, which may indicate abnormal behaviour or events at the plant. For the remainder of the data, however, it is shown (Figure B15) that the standard error is above zero and has a small degree of variation over the entire data period.



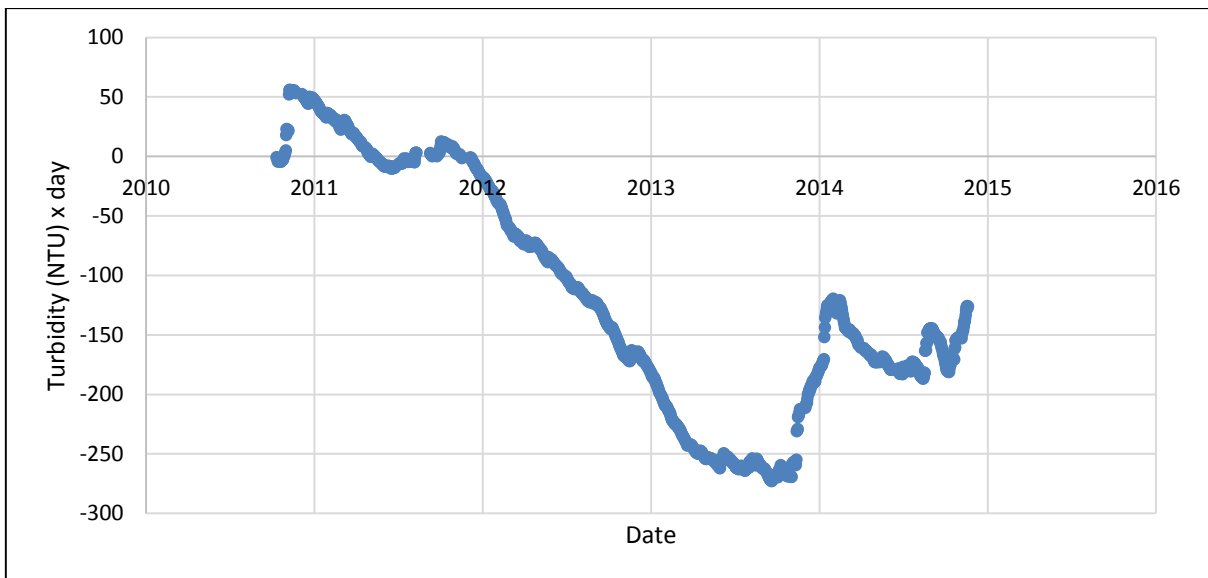
**Figure B16:** Standard error for Turbidity in the Final Water (Swartz, *et al.*, 2016)

From Figure B16, there are a few spikes during the year 2011 as well as a smaller spike near the end of 2014. It was mentioned earlier that these spikes may be caused by overdosing chemical stabilisers during the final processing of the water. The rest of the data show a small degree of variation in the standard error which is to be expected, and therefore a sign that the data is a good representation of the actual process behaviour.

### CUSUM chart

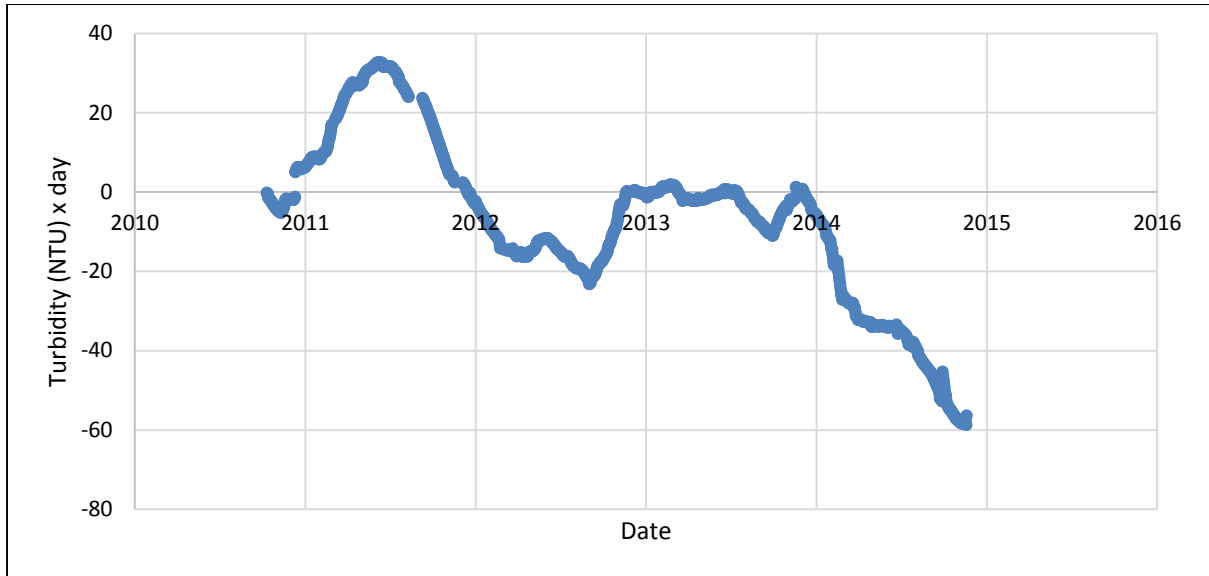
Cusum charts can be used to easily identify process upsets. The chart plots a variable such that the slope of the plot is equal to the average value of the variable. A variable that remains at a stable, constant level will produce a cusum chart with a constant slope. The chart is therefore useful for detecting changes in the average variable values.

Cusum charts require some initiation or target value (T) from which to start the plot, this value is typically equal to the average of the variable, but can be changed in order to increase or reduce the detail seen during certain periods.



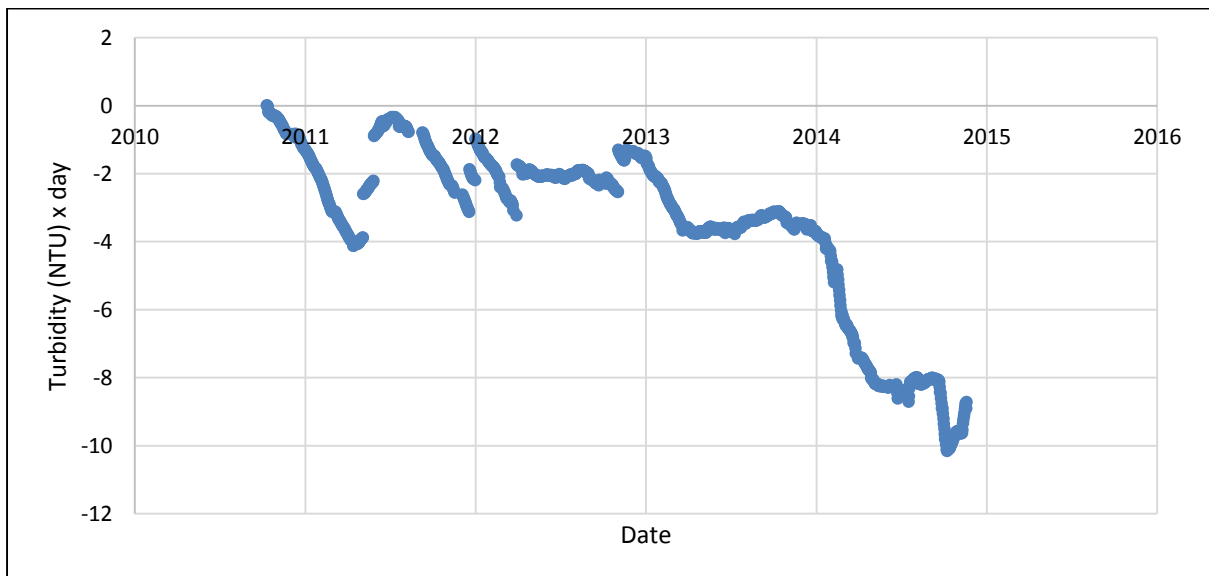
**Figure B17:** Cusum chart ( $T = 2.1$ ) for Turbidity in the WRP Influent (Swartz, *et al.*, 2016)

Figure B17 is a cusum chart for turbidity in the influent to the WRP. There were primarily two events, a small one at the beginning of the year 2011 and a major one at the end of the year 2013. In-between these two events, the slope of the plot remained fairly constant with small variations, but no upsets. This is indicative of consistent plant behaviour which is a good representation of the actual process.



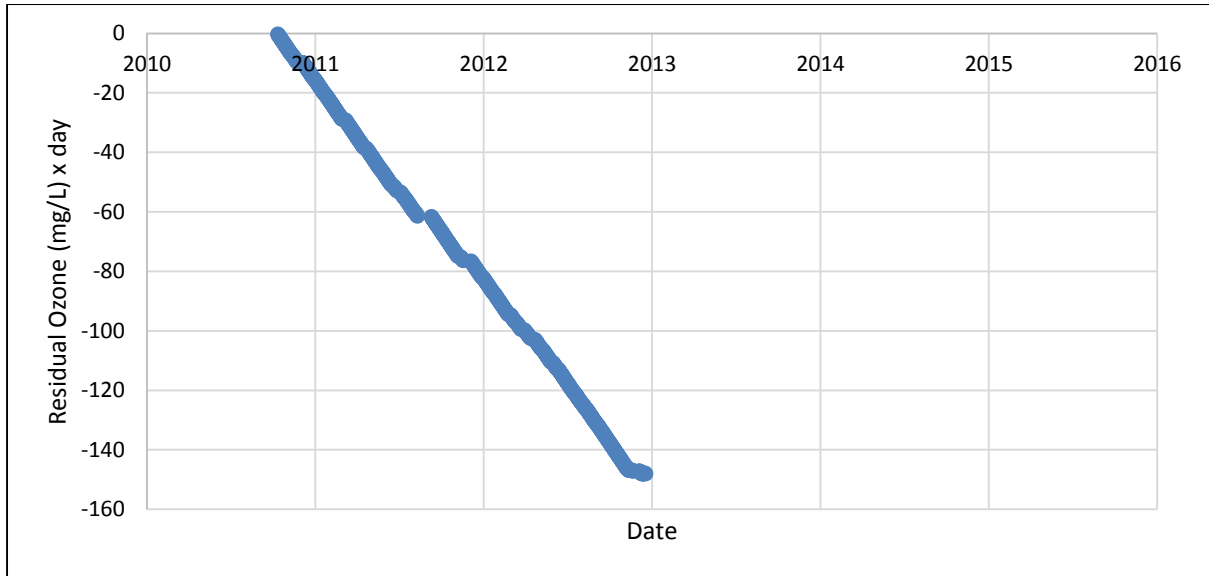
**Figure B18:** Cusum chart ( $T = 0.97$ ) for Turbidity from the DAF (Swartz, *et al.*, 2016)

A cusum chart for turbidity in the DAF effluent can be seen in Figure B18. It is clear that the average for the turbidity underwent several changes. These changes may seem extreme, but since the target value for the cusum charts are the averages of the actual variables, the variation in the slope simply indicates a variation in the average of the moving window relative to that of the whole variable.



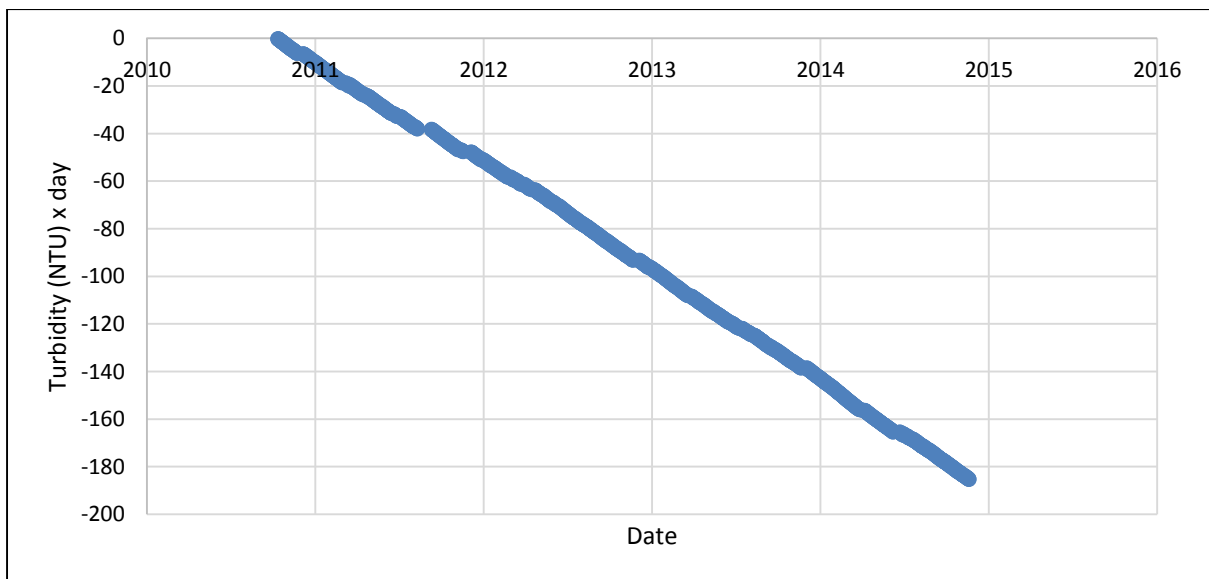
**Figure B19:** Cusum chart ( $T = 0.15$ ) for Turbidity from the SF (Swartz, *et al.*, 2016)

Figure B19 shows that there were several upsets in the turbidity of the sand filter during the period starting in the middle of the year 2011 until the start of the year 2013. The drastic changes in the slope of the plots indicate a major upset and this data may not be entirely suitable for further analyses, unless pre-treatment of the data could be effectively applied in order to increase the quality of the data. With that being said, this data is still a good representation of the actual process behaviour.



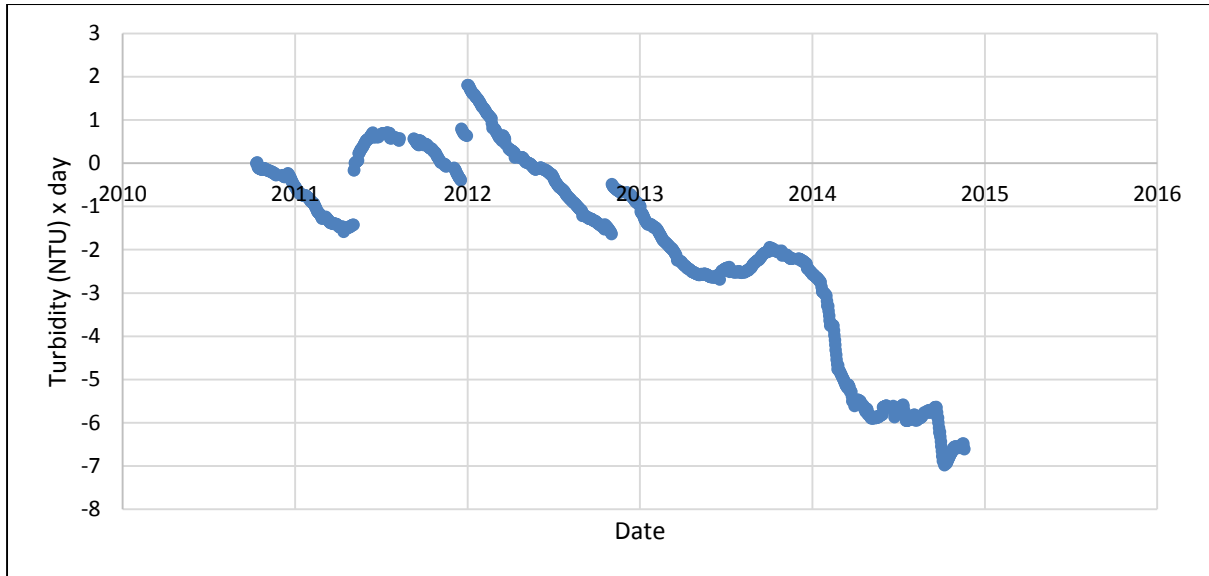
**Figure B20:** Cusum chart ( $T = 0.23$ ) for Residual Ozone from the Final Ozone Contact (Swartz, *et al.*, 2016)

The cusum chart (Figure B20) for the residual ozone in the effluent of the ozonation treatment step shows that the average residual ozone have remained more or less constant, with just enough variation to indicate that the data is a good representation of the actual process behaviour.



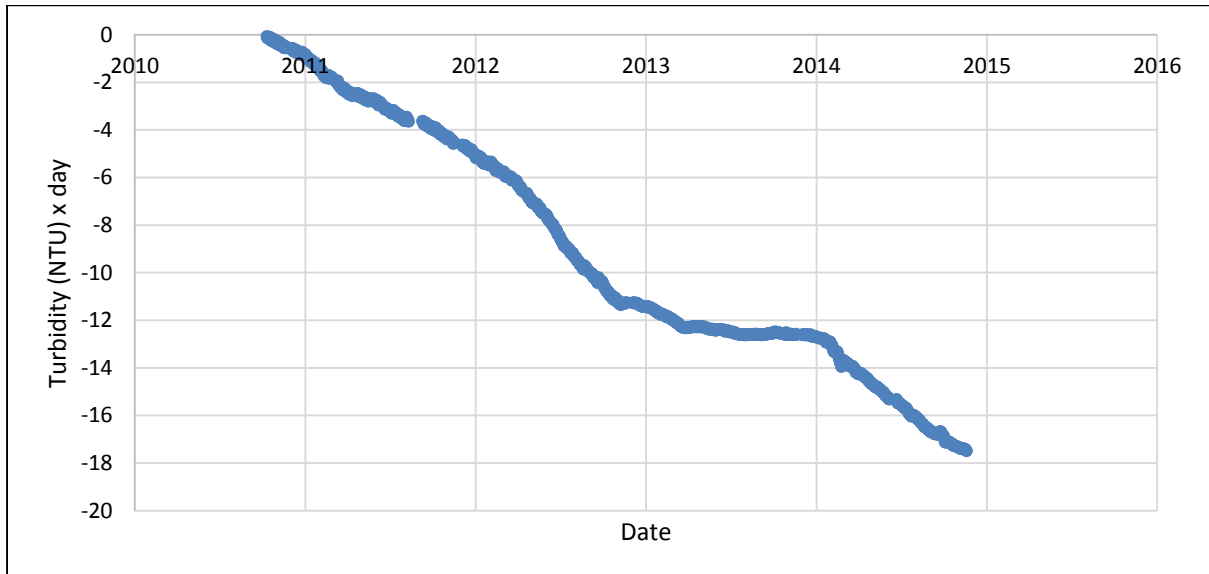
**Figure B21:** Cusum chart ( $T = 0.036$ ) for Turbidity from the BAC (Swartz, *et al.*, 2016)

Figure B21 shows the cusum chart of the turbidity in the effluent of the BAC filters. It can be clearly seen that the average turbidity in the BAC effluent remains constant, which is unexpected since both the moving average and standard error plots indicated a strong seasonality in the BAC data. It is, however, possible that these seasonal variations were too subtle to show clearly on the cusum chart.



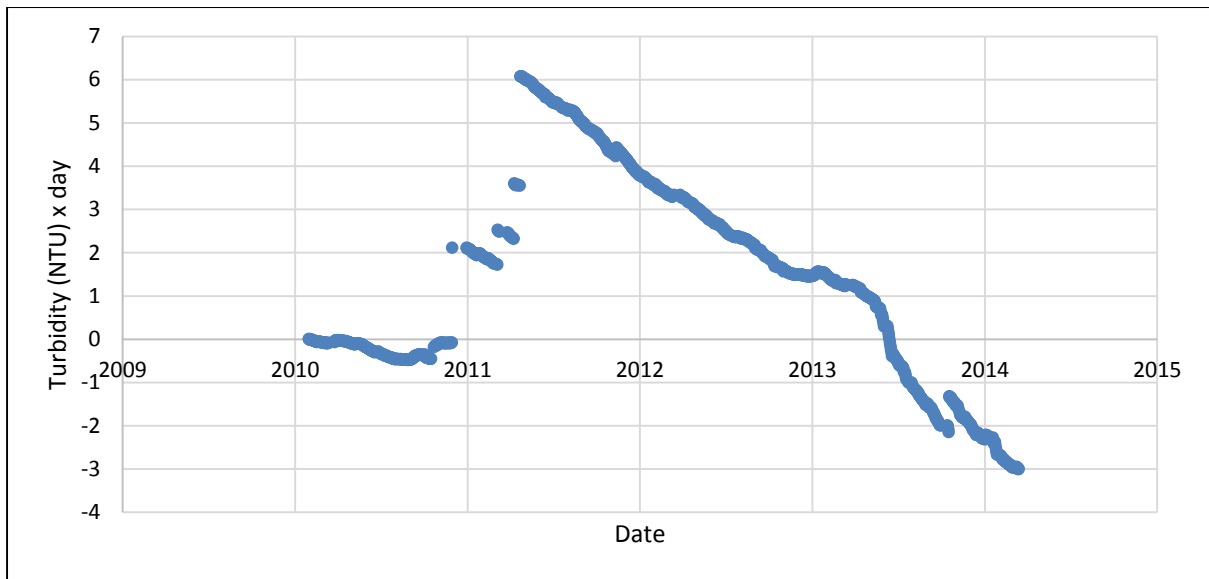
**Figure B22:** Cusum chart ( $T = 0.11$ ) for Turbidity from the GAC (Swartz, *et al.*, 2016)

Figure B22 shows that there is no seasonality to the variation of the data, but there are several disturbances between the years 2010 to 2012. From the year 2013, however, these disturbances have been reduced and can only be seen as a small change in the average turbidity. There is also a more-or-less constant slope from the middle of the year 2011 up to the end of the data period. The small upsets in the data may be due to operational practices which have been improved over time.



**Figure B23:** Cusum chart ( $T = 0.071$ ) for Turbidity from the UF (Swartz, *et al.*, 2016)

From Figure B23, the UF performed very consistently showing no dramatic changes in the cusum chart. Slight variations can be seen in the slope, but nothing indicative of a process upset. The data is therefore a good representation of the actual treatment process operation.



**Figure B24:** Cusum chart ( $T = 0.071$ ) for Turbidity in the Final Water (Swartz, *et al.*, 2016)

Figure B24 indicates the cusum chart for the turbidity in the final water produced by the WRP. It can be seen that the average turbidity remained constant from the year 2010 until the year 2011, then there was a major upset and from the period past the upset, early in the year 2011 until the end of the data period, the average turbidity was fairly constant again. Near the end of the year 2013 there was another upset, though much smaller.

Overall it appears as if the data showed a sufficient amount of variation to be considered representative of the actual processes that occur on the plant.



## APPENDIX C: LIST OF SIGNIFICANT CORRELATIONS FROM PEARSON AND SPEARMAN CORRELATION ANALYSES

Table C1: Significant Pearson correlation coefficients before separating the MVR

ID 1	ID 2	Sample Point 1	Sample Point 2	Analysis 1	Analysis 2	R
6	46	WWTP Clarifier	WWTP Final Eff.	Total solids (TS:105°)	Chlorophyll A	0.998
15	90	WWTP Clarifier	WRP Inf.	Total Alkalinity	Chlorophyll A	0.998
20	171	WWTP MP B1	BAC	Ammonia	Iron (Fe)	0.988
23	165	WWTP MP B8	BAC	Ammonia	COD	0.983
20	167	WWTP MP B1	BAC	Ammonia	pH	0.980
3	139	WWTP Clarifier	SF	pH	Faecal streptococci	0.938
2	96	WWTP Clarifier	WRP Inf.	EC	Total coliform	0.883
1	95	WWTP Clarifier	WRP Inf.	Nitrate	Pseudomonas aeruginosa	0.860
3	93	WWTP Clarifier	WRP Inf.	pH	Faecal coliform	0.859
1	139	WWTP Clarifier	SF	Nitrate	Faecal streptococci	0.853
5	72	WWTP Clarifier	WRP Inf.	Ammonia	COD	0.836
3	96	WWTP Clarifier	WRP Inf.	pH	Total coliform	0.822
3	141	WWTP Clarifier	SF	pH	Total coliform	0.821
22	193	WWTP MP B6	Final Water	Ammonia	COD	0.820
10	22	WWTP Clarifier	WWTP MP B6	TKN	Ammonia	0.813
66	150	WWTP Final Eff.	Ozone Contact B	Colour	Residual O3	0.811
3	54	WWTP Clarifier	WWTP Final Eff.	pH	Total coliform	0.810
31	84	WWTP Final Eff.	WRP Inf.	Nitrate	Orthophosphate	0.807
54	208	WWTP Final Eff.	Final Water	Total coliform	Sodium (Na)	0.797
2	52	WWTP Clarifier	WWTP Final Eff.	EC	Faecal coliform	0.789
2	46	WWTP Clarifier	WWTP Final Eff.	EC	Chlorophyll A	0.750
37	206	WWTP Final Eff.	Final Water	Nitrite	TOC	0.736
53	103	WWTP Final Eff.	WRP Inf.	Faecal streptococci	Sulphate (SO4)	0.731
65	150	WWTP Final Eff.	Ozone Contact B	Total BG Algae	Residual O3	0.719
52	59	WWTP Final Eff.	WWTP Final Eff.	Faecal coliform	Chloride (Cl)	0.712

**Table C2:** Significant Spearman correlation coefficients before separating the MVR

ID 1	ID 2	Sample Point 1	Sample Point 2	Analysis 1	Analysis 2	R
6	52	WWTP Clarifier	WWTP Final Eff.	Total solids (TS:105°)	Faecal coliform	1.000
6	53	WWTP Clarifier	WWTP Final Eff.	Total solids (TS:105°)	Faecal streptococci	1.000
6	58	WWTP Clarifier	WWTP Final Eff.	Total solids (TS:105°)	Somatic coliphage	1.000
8	46	WWTP Clarifier	WWTP Final Eff.	Nitrite	Chlorophyll A	1.000
9	46	WWTP Clarifier	WWTP Final Eff.	TDS (TDS:180°)	Chlorophyll A	1.000
10	46	WWTP Clarifier	WWTP Final Eff.	TKN	Chlorophyll A	1.000
11	46	WWTP Clarifier	WWTP Final Eff.	TSS	Chlorophyll A	1.000
12	46	WWTP Clarifier	WWTP Final Eff.	Orthophosphate	Chlorophyll A	1.000
13	131	WWTP Clarifier	SF	Chlorophyll A	Iron (Fe)	1.000
14	81	WWTP Clarifier	WRP Inf.	DOC	Manganese (Mn)	1.000
122	161	DAF	Ozone Contact Final	Total Algal Count	Clostridium Viable	1.000
146	161	SF	Ozone Contact Final	Total Algal Count	Clostridium Viable	1.000
102	150	WRP Inf.	Ozone Contact B	Fluoride (F)	Residual O3	0.881
4	50	WWTP Clarifier	WWTP Final Eff.	Temperature	Clostridium Spores	0.857
4	55	WWTP Clarifier	WWTP Final Eff.	Temperature	HPC	0.857
3	52	WWTP Clarifier	WWTP Final Eff.	pH	Faecal coliform	0.821
3	54	WWTP Clarifier	WWTP Final Eff.	pH	Total coliform	0.821
54	204	WWTP Final Eff.	Final Water	Total coliform	Chloride (Cl)	0.792
2	23	WWTP Clarifier	WWTP MP B8	EC	Ammonia	0.791
62	150	WWTP Final Eff.	Ozone Contact B	Potassium (K)	Residual O3	0.776
59	93	WWTP Final Eff.	WRP Inf.	Chloride (Cl)	Faecal coliform	0.732
66	152	WWTP Final Eff.	Ozone Contact C	Colour	Residual O3	0.731
30	67	WWTP Final Eff.	WWTP Final Eff.	UV 254	TOC	0.728
52	100	WWTP Final Eff.	WRP Inf.	Faecal coliform	Chloride (Cl)	0.727
53	100	WWTP Final Eff.	WRP Inf.	Faecal streptococci	Chloride (Cl)	0.723
72	123	WRP Inf.	SF	COD	Turbidity	0.722
54	208	WWTP Final Eff.	Final Water	Total coliform	Sodium (Na)	0.715
57	204	WWTP Final Eff.	Final Water	E Coli (confirmed) Tryptone	Chloride (Cl)	0.710
154	157	Ozone Contact Final	Ozone Contact Final	Temperature	HPC	0.704

**Table C3:** Significant Pearson correlation coefficients after separating the MVR

xID	yID	x Sample Point	y Sample Point	x Analysis	y Analysis	R
17	141	GAC	Ozone Contact Final	Temperature	Temperature	0.985
33	141	SF	Ozone Contact Final	Temperature	Temperature	0.984
30	141	DAF	Ozone Contact Final	Temperature	Temperature	0.983
26	141	Final Water	Ozone Contact Final	Temperature	Temperature	0.974
35	141	WRP Inf.	Ozone Contact Final	Temperature	Temperature	0.973
14	141	WWTP Final Eff.	Ozone Contact Final	Temperature	Temperature	0.958
26	165	Final Water	UF	Temperature	Temperature	0.930
17	155	GAC	BAC	Temperature	Temperature	0.929
34	140	GAC	Ozone Contact Final	pH	pH	0.926
14	165	WWTP Final Eff.	UF	Temperature	Temperature	0.921
26	155	Final Water	BAC	Temperature	Temperature	0.919
30	155	DAF	BAC	Temperature	Temperature	0.918
33	155	SF	BAC	Temperature	Temperature	0.916
14	4	WWTP Final Eff.	WWTP Clarifier	Temperature	Temperature	0.913
30	165	DAF	UF	Temperature	Temperature	0.911
14	155	WWTP Final Eff.	BAC	Temperature	Temperature	0.909
26	4	Final Water	WWTP Clarifier	Temperature	Temperature	0.908
17	165	GAC	UF	Temperature	Temperature	0.907
17	4	GAC	WWTP Clarifier	Temperature	Temperature	0.907
33	165	SF	UF	Temperature	Temperature	0.906
30	4	DAF	WWTP Clarifier	Temperature	Temperature	0.905
33	4	SF	WWTP Clarifier	Temperature	Temperature	0.901
35	155	WRP Inf.	BAC	Temperature	Temperature	0.900
35	165	WRP Inf.	UF	Temperature	Temperature	0.895
35	4	WRP Inf.	WWTP Clarifier	Temperature	Temperature	0.890
36	140	DAF	Ozone Contact Final	pH	pH	0.846
39	75	WWTP Final Eff.	WRP Inf.	Ammonia	Ammonia	0.845
34	9	GAC	WWTP Clarifier	pH	TDS (TDS:180°)	0.764
34	6	GAC	WWTP Clarifier	pH	Total solids (TS:105°)	0.753
34	134	GAC	Ozone Contact A	pH	Free chlorine	0.744
34	138	GAC	Ozone Contact C	pH	Free chlorine	0.739
34	136	GAC	Ozone Contact B	pH	Free chlorine	0.730
34	40	GAC	WWTP Final Eff.	pH	TDS (TDS:180°)	0.727
34	2	GAC	WWTP Clarifier	pH	EC	0.717

**Table C4:** Significant Spearman correlation coefficients after separating the MVR

xID	yID	x Sample Point	y Sample Point	x Analysis	y Analysis	R
17	141	GAC	Ozone Contact Final	Temperature	Temperature	0.981
33	141	SF	Ozone Contact Final	Temperature	Temperature	0.978
30	141	DAF	Ozone Contact Final	Temperature	Temperature	0.976
35	141	WRP Inf.	Ozone Contact Final	Temperature	Temperature	0.961
26	141	Final Water	Ozone Contact Final	Temperature	Temperature	0.959
14	141	WWTP Final Eff.	Ozone Contact Final	Temperature	Temperature	0.941
34	140	GAC	Ozone Contact Final	pH	pH	0.939
14	4	WWTP Final Eff.	WWTP Clarifier	Temperature	Temperature	0.932
17	155	GAC	BAC	Temperature	Temperature	0.924
30	4	DAF	WWTP Clarifier	Temperature	Temperature	0.923
33	4	SF	WWTP Clarifier	Temperature	Temperature	0.921
17	4	GAC	WWTP Clarifier	Temperature	Temperature	0.921
14	165	WWTP Final Eff.	UF	Temperature	Temperature	0.919
26	4	Final Water	WWTP Clarifier	Temperature	Temperature	0.918
26	165	Final Water	UF	Temperature	Temperature	0.917
33	155	SF	BAC	Temperature	Temperature	0.915
35	4	WRP Inf.	WWTP Clarifier	Temperature	Temperature	0.913
30	155	DAF	BAC	Temperature	Temperature	0.911
26	155	Final Water	BAC	Temperature	Temperature	0.907
30	165	DAF	UF	Temperature	Temperature	0.900
17	165	GAC	UF	Temperature	Temperature	0.899
33	165	SF	UF	Temperature	Temperature	0.897
35	155	WRP Inf.	BAC	Temperature	Temperature	0.896
35	165	WRP Inf.	UF	Temperature	Temperature	0.896
14	155	WWTP Final Eff.	BAC	Temperature	Temperature	0.890
39	75	WWTP Final Eff.	WRP Inf.	Ammonia	Ammonia	0.861
36	140	DAF	Ozone Contact Final	pH	pH	0.816
18	96	Operational	WRP Inf.	Final Water Pumped	Sulphate (SO4)	0.805
19	96	Operational	WRP Inf.	Total Raw Water	Sulphate (SO4)	0.782
18	55	Operational	WWTP Final Eff.	Final Water Pumped	Chloride (Cl)	0.779
19	55	Operational	WWTP Final Eff.	Total Raw Water	Chloride (Cl)	0.765
18	57	Operational	WWTP Final Eff.	Final Water Pumped	Sulphate (SO4)	0.758
18	186	Operational	Final Water	Final Water Pumped	Sulphate (SO4)	0.749
19	57	Operational	WWTP Final Eff.	Total Raw Water	Sulphate (SO4)	0.733
38	65	SF	WRP Inf.	Turbidity	COD	0.724
19	186	Operational	Final Water	Total Raw Water	Sulphate (SO4)	0.720
39	33	WWTP Final Eff.	WWTP Final Eff.	Ammonia	Nitrite	0.712