DATA-DRIVEN MODELING FOR WASTEWATER TREATMENT PROCESSES

APPLICATION OF DATA-DRIVEN MODELING TECHNIQUES TO WASTEWATER TREATMENT PROCESSES

by

EMMA HERMONAT, B.Eng

A Thesis
Submitted to the School of Graduate Studies
in Partial Fulfillment of the Requirements for
the Degree Master of Applied Science

MASTER OF APPLIED SCIENCE (2022) McMaster University
(Chemical Engineering) Hamilton, Ontario, Canada

TITLE: Application of Data-Driven Modeling Techniques to Wastewater Treatment Processes

AUTHOR: Emma Hermonat, B.Eng
(McMaster University, Hamilton, ON)

SUPERVISOR: Dr. Prashant Mhaskar

NUMBER OF PAGES: xi, 104

ABSTRACT

Wastewater treatment plants (WWTPs) face increasingly stringent effluent quality constraints as a result of rising environmental concerns. Efficient operation of the secondary clarification process is essential to be able to meet these strict regulations. Treatment plants can benefit greatly from making better use of available resources through improved automation and implementing more process systems engineering techniques to enhance plant performance. As such, the primary objective of this research is to utilize data-driven modeling techniques to obtain a representative model of a simplified secondary clarification unit in a WWTP.

First, a deterministic subspace-based identification approach is used to estimate a linear state-space model of the secondary clarification process that can accurately predict process dynamics, with the ultimate objective of motivating the use of the subspace model in a model predictive control (MPC) framework for closed-loop control of the clarification process. To this end, a low-order subspace model which relates a set of typical measured outputs from a secondary clarifier to a set of typical inputs is identified and subsequently validated on simulated data obtained via Hydromantis's WWTP simulation software, GPS-X. Results illustrate that the subspace model is able to approximate the nonlinear process behaviour well and can effectively predict the dynamic output trajectory for various candidate input profiles, thus establishing its candidacy for use in MPC.

Subsequently, a framework for forecasting the occurrence of sludge bulking—and consequently clarification failure—based on an engineered interaction variable that aims to capture the relationship between key input variables is proposed. Partial least squares discriminant analysis (PLS-DA) is used to discriminate between process conditions associated with clarification failure versus effective clarification. Preliminary results show that PLS-DA models augmented with the interaction variable demonstrate improved predictions and higher classification accuracy.

ACKNOWLEDGEMENTS

First and foremost, I must extend my sincerest gratitude to my supervisor, Dr. Prashant Mhaskar, who has provided me with invaluable advice and guidance, continuous support and unending patience throughout my time as his student. I am extremely lucky to have such a caring supervisor whose dedication to their students is inspiring. Thank you, Dr. Mhaskar, for being an exemplary supervisor and, above all, an admirable person.

I would also like to thank both Dr. Jake Nease and Dr. Vladimir Mahalec for their willingness to sit on my thesis committee. A special shoutout to Dr. Jake Nease who has been somewhat of a mentor for me during my years as both an undergraduate and graduate student. Thank you for letting me be a "long-time friend" of ChE 4G03 and for showing me that anything is possible — even an almost hole-in-one for a first-time golfer.

Thank you to McMaster University, the Department of Chemical Engineering, the McMaster Advanced Control Consortium (MACC) and Hydromantis Environmental Software Solutions (ESS), Inc. for the support—both personal and financial—that I have received during my graduate degree.

A huge thank you to my friends and family for their endless love, support and encouragement. Thank you to my big sister, Megan, for always keeping me grounded and being the perfect example of a strong woman growing up. Thank you to my sweet sweet niece, Ellie, for being a constant source of smiles, belly laughs and love.

Thank you to my wonderful partner, Matt, who inspires me every day to be my best self. I apologize for being grumpier than normal while working on this thesis and I promise that the best is yet to come.

Last but certainly not least, I am forever indebted to my incredible parents, Mark and Yvonne, without whom none of this would be possible. I attribute my love of STEM to all the fun, sneakily academic activities they set up for me as a child and, for this reason, I dedicate this milestone to them.

Without such a team behind me, I would certainly not be where I am today.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Wastewater treatment plants (WWTPs) are facing increasingly stringent requirements and effluent quality constraints due to ever increasing environmental protection concerns, which has in turn brought increased attention to understanding WWTP operation through mathematical models. A good quantitative understanding of the wastewater treatment process (a model) is essential for the implementation of reliable and efficient monitoring and control methods that can enable plants to better manage resources and meet certain specs. In practice, there are two main modeling techniques: first-principles (mechanistic) models and data-driven models. First-principles models are built using explicit knowledge of the process mechanisms and invoke fundamental physical and chemical laws that describe the system, often utilizing algebraic, ordinary, or partial differential equations. Conversely, data-driven models exclusively utilize available measured process data to identify parameters for a model structure chosen *a priori*.

A fairly comprehensive library of first-principles models exist in the general area of wastewater treatment, such as the activated sludge model (ASM) family [Henze *et al.*, 2006; Von Sperling, 2007a] and models which describe the sedimentation processes [Takács *et al.*, 1991; Jeppsson, 1996; Jeppsson and Diehl, 1996]. While they have proven valuable in capturing the general trends of the process, the high complexity and nonlinearity of the wastewater treatment process prevents first-principles models from truly predicting variable behaviour precisely. Moreover, the task of building, maintaining and calibrating these first-principles models to specific process units (and to the treatment process as a whole) remains challenging due to the large number of model parameters that must be identified [Hauduc *et al.*, 2009; Dürrenmatt and Gujer, 2012]. The applicability of first-principles models is further limited by the very specific conditions under which these models are constructed (and therefore valid for) [Sánchez, 2004]. These limitations imply that first-principles models cannot be used to reliably describe or tightly control the wastewater treatment process, in turn limiting the plant's ability to meet the tight specifications imposed.

The limitations of existing first-principles models in combination with increasing availability of data and continually improving computational capabilities have motivated the use of increasingly popular data-driven modeling techniques. Data-driven approaches aim to construct a simpler—often linear—model from measured process data and, as such, do not require a detailed prior knowledge of the system itself. Many data-driven modeling approaches exist which largely differ in model structure and intrinsic computation, among which include latent variable-based methods and subspace-based identification methods.

Subspace identification is a two-step statistical-based approach that utilizes measured input-output data to first estimate a state trajectory and then compute the model parameters for a linear time-invariant (LTI) state-space model of the process [Moonen *et al.*, 1989; van Overschee and de Moor, 1995; de Moor *et al.*, 1999]. The key advantage of subspace identification is its conceptual and computational simplicity; subspace methods are fast (noniterative), numerically stable algorithms that obtain models useful for predicting process variable behaviour and subsequently as the basis for feedback control [van Overschee and de Moor, 1996]. Subspace identification methods have proven a useful tool in many engineering applications and have historically been utilized to model and control various industrial processes [Favoreel *et al.*, 2000; Meidanshahi *et al.*, 2017; Misra and Nikolaou, 2003; Bastogne *et al.*, 1997] and for designing effective fault detection and isolation systems [Basseville *et al.*, 2000; Ding *et al.*, 2009; Wei *et al.*, 2010; Shahnazari *et al.*, 2018].

Partial least squares (PLS; also referred to as projection to latent spaces)—another statistical-based method—is classified as a latent variable multivariate regression (LVMR) method in which data of high dimensionality is projected into lower-dimensional latent variable subspaces to obtain a static linear multivariate model. PLS is an extremely useful tool for understanding and identifying important relationships between process inputs and outputs, as well as for predicting a set of response variables from a set of predictor variables [Corbett and Mhaskar, 2016]. Arguably one of the most recognized and widely used data-driven approaches, PLS has been applied in practice extensively for purposes such as predicting variable information, monitoring and controlling industrial processes, optimizing process operation and product quality, and also as a general dimensionality reduction technique [Dunn, 2010; Dong and Qin, 2015]. PLS is most commonly used in the area of chemometrics – largely for analytical instrument calibration and for deriving quantitative structure-property relations (QSPR), quantitative structure-activity relations (QSAR) and comparative molecular field analysis (CoMFA) models from molecular data [Wold *et al.*, 2001]. More generally, PLS has also been applied for steady-state process modeling, dynamic process modeling, and process monitoring [Qin, 1993; Lakshminarayanan *et al.*, 1997; Kresta *et al.*, 1991; Kresta, 1992]. Partial least squares discriminant analysis (PLS-DA) is a variant of PLS regression that is used as a discrimination method when the response (output) data is categorical. PLS-DA is a versatile modeling approach but is most often used to discriminate between two or more predefined class labels and predict the class membership of observations based on corresponding measured input data. Like PLS regression, the use of PLS-DA is quite prominent in the field of chemometrics and has been used in fields such as multivariate image analysis, medical diagnostics, soil science, food analysis, metabolomics and other general *omics* data analyses, and much more [Ruiz-Perez *et al.*, 2018; Chevallier *et al.*, 2006; Ballabio and Consonni, 2013; Tan *et al.*, 2004; Worley and Powers, 2013].

Owing to the fact that wastewater treatment plants can benefit from making better use of available resources through improved automation and the incorporation of process systems engineering techniques, the integration of data-driven modeling tools has garnered much interest in the wastewater treatment field. In particular, data-driven techniques have been explored for the purposes of process monitoring and fault detection, variable prediction, and advanced automated control of the treatment process [Corominas *et al.*, 2018; Newhart *et al.*, 2019].

The use of subspace methods—which are by design suited for model-based control—in the general area of wastewater treatment is limited to date. Lindberg [1997] utilizes subspace identification to obtain a model relating the concentrations of nitrate and ammonium in the last aerated zone of an activated sludge process to three process inputs and three measured disturbances. The identified subspace model is subsequently used in the design of a multivariable feedforward controller with the aim of maintaining nitrate and ammonium concentrations as close to their desired set-points as possible. This study demonstrates the ability of the subspace model to effectively predict nitrate and ammonium levels and highlights the importance of equipping the controller with a reliable model capable of accurate predictions. Similarly, Sánchez and Katebi [2003] utilize a model estimated via subspace identification to design a model predictive controller to control dissolved oxygen (DO) levels in an activated sludge WWTP. Sotomayor *et al.* [2003] investigates the performance of various subspace identification methods for modeling a complex activated sludge process with two inputs, four disturbances and two outputs. The described studies are excellent examples illustrating potential value of using subspace identification methods to model and control the wastewater treatment process and prompt further exploration of applying subspace-based approaches to additional treatment units, quality variables and control implementations.

In the field of wastewater treatment, PLS is often used to build a predictive model relating a set of input variables (e.g, initial water quality, plant operational information and design characteristics) to a set of output variables (often effluent water quality properties). Specifically, PLS has been used as a successful variable prediction and process monitoring tool in WWTPs [Rosen and Olsson, 1998; Choi and Lee, 2005; Aguado *et al.*, 2006; Woo *et al.*, 2009], and for water quality estimation and monitoring via soft-sensors developed from measured ultraviolet-visible (UV-Vis) spectroscopy data [Langergraber *et al.*, 2003; Lourenço *et al.*, 2008; Platikanov *et al.*, 2014]. On the other hand, the use of PLS-DA in WWTPs seems to be quite narrow, with most results aimed at characterizing and discriminating between samples taken at either different locations within a WWTP or between different WWTPs entirely [Singh *et al.*, 2005; Perez-Lopez *et al.*, 2021; Yotova *et al.*, 2019]. Perez-Lopez *et al.* [2021] apply PLS-DA to evaluate and characterize the wastewater microbiome at different stages in the treatment process. More specifically, this study utilizes PLS-DA to discriminate among water samples taken at three different points in the WWTP—the influent, the anoxic reactor (denitrification), and the final effluent—based on changes in the concentrations of various peptides and proteins detected at the sample locations. The PLS-DA model is able to identify the most relevant peptides and proteins present at the three location and subsequently classify the samples according to their point in the treatment process. The use of PLS-DA here shows promise as a tool to highlight composition differentiation and help indicate WWTP performance based on bacterial activity and proteomics variability over the course of the treatment process. Wang *et al.* [2017] shows that PLS-DA can also be used to discriminate between dry and wet climate conditions based on data describing the seasonal variation of WWTP influent characteristics in cold climate regions. In this study, PLS-DA is used to classify WWTP influent as deriving from either wet or dry climate conditions based on measured data for the influent flow rate, water temperature and chemical composition of the wastewater during both the warm and cold seasons. The model is able to attain a classification accuracy of 91%, which prompts the authors to explore the development of a scenario-based soft sensor to survey and control WWTPs as a subject of future work. Existing studies which employ PLS-DA to characterize the treatment process and discriminate various operating conditions and wastewater compositions highlight the utility of PLS-DA and motivates further exploration of potential uses for discriminatory models in WWTPs. One potential application of classification in WWTPs is *failure analysis and maintenance management*, the goal of which is to construct a model that can classify and predict process failure modes (either for a specific treatment unit or the overall WWTP). Classification models can be a valuable condition monitoring and fault detection tool for WWTPs [Bertolini *et al.*, 2021].

Data is continuously being collected at WWTPs but has yet to be used to its full potential. As such, data-driven approaches which exploit the available data are extremely attractive to WWTPs, providing opportunities to improve and optimize treatment performance by building better models and designing more efficient data-driven monitoring and control systems.

1.2 Research problem statement

This research explores the application of data-driven modeling techniques—both dynamic and static—to the secondary clarification unit within a wastewater treatment process. The aim of this work is twofold:

1. **Dynamic modeling**

   Utilize subspace-based model identification to estimate a representative linear time-invariant state-space model of the secondary clarification process from simulated data. The objective of this piece is to construct a suitable linear model that can describe the nonlinear process dynamics of the secondary clarifier and accurately predict the output variable trajectories for multiple candidate input profiles.

2. **Static modeling**

   Utilize partial least squares discriminant analysis to identify a linear static model from simulated steady-state data that can be used to predict the occurrence of clarification failure based on key input variables and parameters in the wastewater treatment process. The objective of this piece is to discriminate various clarifier operational conditions which result in failure of the clarification unit, with a focus on defining a relevant interaction variable that can improve predictions and enhance classifier performance.

1.3 Outline of the thesis

The thesis is comprised of five chapters covering the application of both dynamic and static data-driven modeling techniques to a simplified secondary clarification system. The thesis is organized as follows:

Chapter 2 provides an overview of the conventional municipal wastewater treatment process with a focus on the importance of the secondary clarification unit to the overall treatment performance and effluent quality. A review of the key data-driven modeling approaches used in this work—subspace-based model identification (dynamic modeling methodology) and partial least squares discriminant analysis (static-based classification methodology)—is also given.

Chapter 3 explores the application of subspace identification to the secondary clarification unit. The objective of this chapter is to identify an appropriate model which captures the dynamic behaviour of the clarification process and can effectively predict variable information, ultimately motivating candidacy of the subspace model for use in a model predictive control scheme. It is in this chapter that a description of the simplified secondary clarification unit is given and the data collection and simulation processes used throughout the thesis are detailed.

Chapter 4 focuses on the identification and classification of input variable conditions which lead to clarification failure using static modeling techniques. To this end, PLS-DA is used to discriminate between potential clarification failure and normal effective clarifier operation. In this chapter, we explore the effect of augmenting the **X**-space with an engineered interaction variable on discrimination results. A third-order interaction term that aims to explain the relation between the input variables and how they ultimately affect response classification is defined and incorporated as an additional predictor in the PLS-DA model structure. The value of the interaction term is established via comparison of classifier performance for the base-case PLS-DA model (without interaction) and the PLS-DA model which includes the third-order interaction term.

Finally, a summary of the key contributions of this work along with some recommendations for potential future directions are presented in Chapter 5.

Chapter 2

Preliminaries

2.1 Overview of wastewater treatment process

Wastewater can be defined as any water that has been contaminated via human activity. More specifically, wastewater is contaminated water sourced from a combination of domestic, industrial, commercial and agricultural activities, as well as any leachate, stormwater, surface run-off and sewer infiltration [Volcke *et al.*, 2020]. The overall objective of the wastewater treatment process is to remove and dispose of any contaminants present in the influent wastewater so that the resultant water is clean enough to be discharged out to a surrounding body of water, such as a nearby lake or river. Influent wastewater enters the WWTP as a mixture of liquid and solid wastes, contaminated with varying concentrations of dissolved and suspended organic and inorganic solids, nutrients, and microorganisms such as pathogenic viruses, bacteria, algae and fungi [Von Sperling, 2007b]. Though these contaminants comprise only roughly 0.1% of raw wastewater, failure to remove them before discharging the water back to the environment would have detrimental effects on existing ecosystems and aquatic life and pose serious health risks to humans [Von Sperling, 2007b].

Figure 2.1 shows a schematic of the conventional municipal WWTP. The treatment process is typically comprised of four levels of treatment: preliminary treatment, primary treatment, secondary treatment and tertiary (or advanced) treatment. Each successive level of treatment targets the removal of various pollutants, further purifying the water in order to reach the required effluent quality discharge standards.

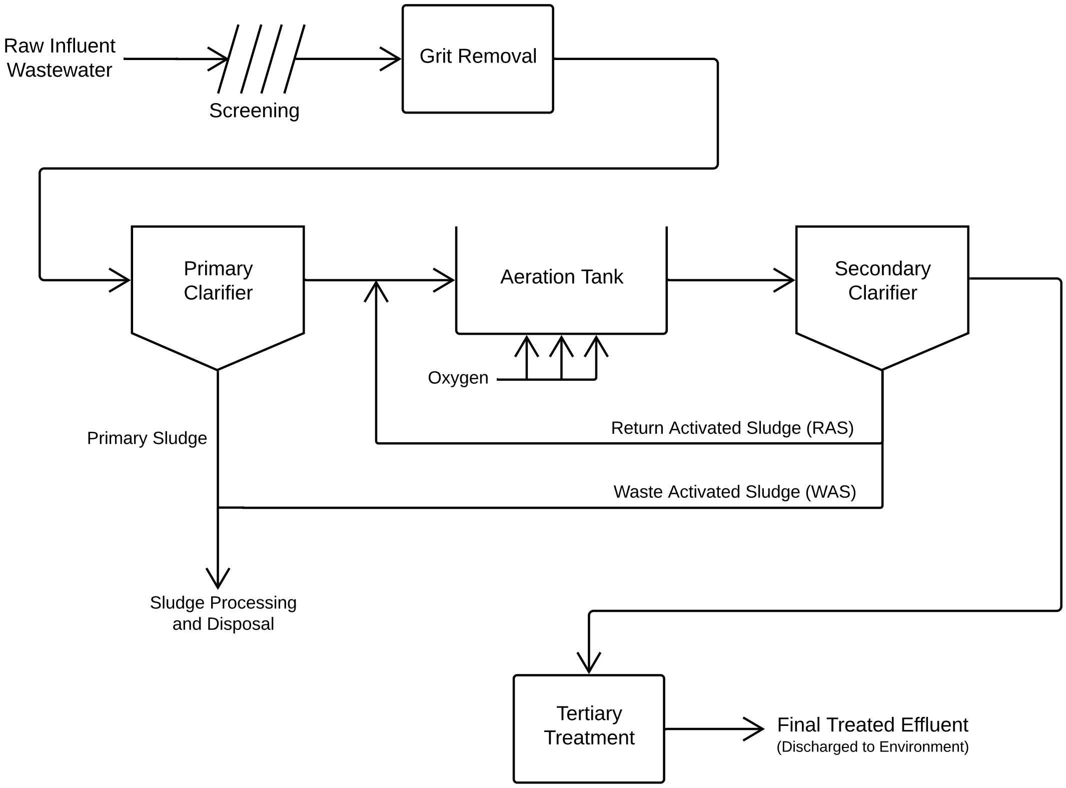


Figure 2.1: Schematic of conventional municipal wastewater treatment plant.

Preliminary treatment is the first phase of the treatment process. Preliminary treatment focuses on the removal of coarse solids from influent wastewater in preparation for subsequent stages and to avoid consequent damage or clogging of downstream equipment. As such, influent wastewater first goes through a screening process to remove any large objects such as plastics, metals, paper and wood, before being subject to a grit removal process which allows sand, small stones and food waste to settle out of the water based on density differences.

The objective of the primary treatment stage is to reduce the solids load for downstream units via the physical removal of organic materials that either float (scum) or readily settle out by gravity (sludge) [MetCalf & Eddy, Inc. *et al.*, 2014]. Gravitational separation of solid matter from liquid is carried out in large primary clarifiers (also referred to as sedimentation or settling tanks) to partially remove settleable suspended solid matter from preliminarily treated wastewater. It is typical for WWTPs to have at least two primary clarifiers, with the exact number of tanks depending largely on plant influent flow, influent wastewater characterizations and size limitations. Clarification tanks reduce the velocity of the incoming wastewater and maintain an average hydraulic detention time of about 1–2 hours [MetCalf & Eddy, Inc. *et al.*, 2014], thereby allowing heavier, more dense solids to settle to the bottom of the tank and other less-dense materials, such as oil and grease, to rise and form a layer of scum at the surface of the tank. The accumulated solids, called *primary sludge*, is pumped from the tank for subsequent processing and disposal, while scum is removed via mechanical skimmers and disposed of. In general, the primary treatment stage is able to remove about 50–70% of settleable suspended solids and 25–40% of biodegradable organic matter (as measured by biochemical oxygen demand; BOD) from the water [Von Sperling, 2007b]. The remaining partially treated wastewater exits the clarifier as *primary effluent* and continues on the secondary treatment stage.

Secondary treatment often employs biological treatment processes and further gravitational sedimentation to remove most of the residual solid and organic matter present in primary effluent. The first step of this treatment stage involves the biological degradation of pollutants by microorganisms through a suspended growth process known as the *activated sludge process*. In contrast to the predominantly physical mechanisms employed in the preliminary and primary treatment levels, the activated sludge process uses millions of single-celled microorganisms—mostly aerobic and heterotrophic—to metabolize organic matter in the wastewater and convert it into biomass, carbon dioxide, water and other inorganic end-products, thus effectively reproducing [Von Sperling, 2007a]. In large aeration tanks, wastewater is mixed with bacteria-rich activated sludge and loaded with oxygen (or air) to facilitate and accelerate the digestion of organic matter by the bacteria. The mixture of water and activated sludge within the aeration tank is referred to as a *mixed liquor*. The mixed liquor is pumped to a secondary clarification tank where the microbial suspension settles and is separated from the clarified water by gravity in a manner similar to primary sedimentation. The biomass that settles and accumulates at the bottom of the tank is known as *activated sludge*. A portion of this activated sludge is recycled to the aeration tank as *return activated sludge* (RAS) to maintain a high microbe concentration in the aeration tanks and further aid biodegradation. Due to this recirculation, aeration tanks have a relatively short hydraulic detention time around 6–8 hours, but a much longer solids retention time (i.e., the average time biomass remains in the system) anywhere from 1–18 days [MetCalf & Eddy, Inc. *et al.*, 2014]. The remaining sludge leaves the clarifier as *waste activated sludge* (WAS) and is stabilized before being mixed with primary sludge for further processing and disposal. Clarified water exits the secondary clarifier either as final effluent for discharge to the receiving body of water or as secondary effluent for further, more rigorous purification in the tertiary treatment stage. In general, the secondary treatment stage can remove roughly 80–95% of settleable solids and up to 99% of biodegradable organic matter [Von Sperling, 2007a; Henze, 2002].

Tertiary (advanced) treatment is considered to be any additional treatment process beyond secondary treatment and is often used to enhance effluent quality when secondary effluent is not of sufficient quality for discharge to a sensitive environment. Tertiary treatment can also be utilized to remove specific (often toxic or non-biodegradable) pollutants, such as nitrogen, phosphorus, heavy metals, and pathogenic bacteria. Common tertiary treatment methods include: filtration; chemical, UV, and ozone disinfection; carbon adsorption; chemical precipitation; distillation; reverse osmosis; and electrodialysis [U.S. EPA, 2004; MetCalf & Eddy, Inc. *et al.*, 2014].

2.1.1 Importance of secondary clarification unit

As previously discussed, the two main functions of the secondary clarifier are clarification and thickening. The physical process of separating solids from liquid via gravitational sedimentation combined with mechanical mechanisms such as sludge scrapers and scum skimmmers employed by clarifiers allow for the continuous removal of solids. Figure 2.2 shows a cross-sectional diagram of a typical circular secondary clarification tank. In this configuration, influent wastewater enters the clarifier via an inlet feedwell located in the center of the tank and is distributed radially throughout. As suspended solids particles settle, sludge is collected via a slow-moving mechanical scraper rotating around the tank's central axis and subsequently removed through a separate pipe. Scum that accumulates at the surface of the tank is similarly removed by a rotating skimmer arm. Finally, clarified water exits the top of the tank overtop overflow weirs and is collected for discharge. Weirs are often also fitted with a baffle to prevent any floating solids particles from exiting the clarifier in the effluent.

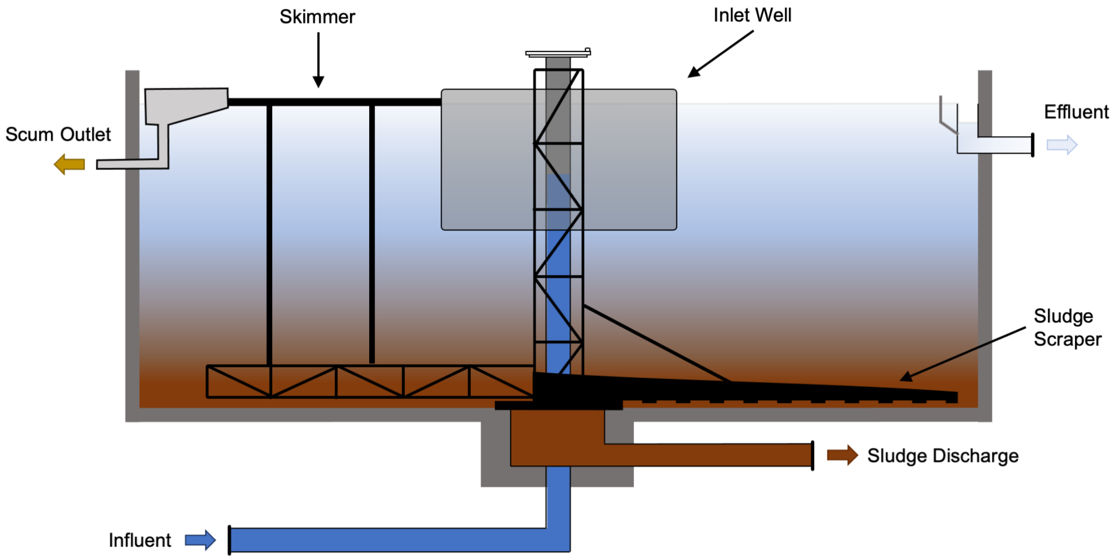


Figure 2.2: Cross-sectional diagram of a circular flat-bottom clarifier.

One key indication of clarification performance is the concentration of total suspended solids (TSS) remaining in the final effluent. By definition, effluent TSS quantifies the amount of non-filterable particles suspended in the water that are larger than two microns in size and do not settle out by gravity during the treatment process. Effective operation of the secondary clarifier ensures that secondary effluent maintains an acceptable total suspended solids (TSS) concentration, as outlined by effluent quality standards set by government-imposed regulations [Environment Canada, 2012]. It is important to measure and regulate effluent TSS concentrations. A high concentration of suspended solids in the effluent can negatively impact the receiving body of water—harming water quality, the aquatic ecosystem and even affecting human health. Beyond just decreasing water clarity, high suspended solids concentrations can block sunlight from penetrating the water and reduce (or in extreme cases, entirely inhibit) photosynthesis in submerged vegetation [Sorensen *et al.*, 1977]. A reduction in plant and algae productivity directly corresponds to decreased dissolved oxygen (DO) levels. Suspended solids particles can also absorb heat from the sunlight, thereby increasing surface water temperatures and further depleting DO. Aquatic ecosystems are quite sensitive to changes in DO and a sustained decrease in DO levels can cause oxygen-dependent species, such as fish and other aquatic biota, to die off due to critical oxygen shortage [Sorensen *et al.*, 1977; Bilotta and Brazier, 2008]. In addition, high TSS concentrations can facilitate the transport of harmful contaminants and toxins sorbed to suspended solids particles—such as heavy metals, halogenated organic compounds, inorganic nutrients (i.e., nitrogen and phosphorus), and pathogenic microorganisms (i.e., bacteria, viruses, parasites, etc.)—through the water system, which can cause ecological damage and be hazardous to human health [Sorensen *et al.*, 1977; Akpor and Muchie, 2011].

The potential deleterious effects of high TSS effluent are manifold. For this reason, secondary clarification is widely considered the most critical and sensitive step of the treatment process and is often referred to as the bottleneck of treatment performance as it commonly serves as the final process unit in a WWTP [Ji *et al.*, 1996; Griborio, 2004]. A successful model of the secondary clarification process can be used to predict effluent TSS levels and construct control applications which optimize clarification efficiency and ensure the solids concentration is maintained within the prescribed range. As such, the secondary clarifier is chosen as the representative process unit for this research.

2.2 Overview of subspace identification methods

Subspace identification refers to a class of model identification methods that aim to construct low-order state-space models from measured input-output data. The basic idea of these methods is to utilize linear regression to estimate a subspace from which a state variable sequence can be extracted and subsequently fit to the state-space model structure [Ljung and McKevey, 1996; Huang *et al.*, 2005; Qin, 2006]. Subspace identification methods rely heavily on matrix factorization techniques such as singular value decomposition (SVD) and QR-decomposition to estimate the state variable sequence. The three most prominent subspace-based approaches are: canonical variate analysis (CVA) [Larimore, 1990], numerical subspace state-space system identification (N4SID) [van Overschee and de Moor, 1994], and multivariable output-error state-space (MOESP) [Verhaegen and Dewilde, 1992].

In general, given a process with p inputs and q outputs, subspace identification estimates an n^{th}-order discrete-time LTI state-space model of the system, represented mathematically by the following set of difference equations:

$$\begin{aligned} \boldsymbol{x}_{k+1} &= \mathbf{A}\boldsymbol{x}_k + \mathbf{B}\boldsymbol{u}_k + \boldsymbol{w}_k \\ \boldsymbol{y}_k &= \mathbf{C}\boldsymbol{x}_k + \mathbf{D}\boldsymbol{u}_k + \boldsymbol{v}_k \end{aligned} \tag{2.1}$$

where k is the current sampling time, $\boldsymbol{x}_k \in \mathbb{R}^n$ is the process state vector, $\boldsymbol{u}_k \in \mathbb{R}^p$ and $\boldsymbol{y}_k \in \mathbb{R}^q$ represent the process inputs and outputs respectively, matrices $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times p}$, $\mathbf{C} \in \mathbb{R}^{q\times n}$ and $\mathbf{D} \in \mathbb{R}^{q\times p}$ are the system matrices, and vectors $\boldsymbol{w}_k \in \mathbb{R}^n$ and $\boldsymbol{v}_k \in \mathbb{R}^q$ are unobserved, stationary white noise vector sequences of zero mean and respectively denote process and output measurement noise.

Subspace identification methods can be broadly categorized as deterministic, stochastic and combined deterministic-stochastic methods, with the key distinction between the three being how process noise and disturbances are handled. For simplicity, the present work utilizes the deterministic subspace identification approach presented in Moonen *et al.* [Moonen *et al.*, 1989], in which neither process nor measurement noise are considered by the model (i.e., $\boldsymbol{w}_k \equiv \mathbf{0}$, $\boldsymbol{v}_k \equiv \mathbf{0}$). The algorithm utilized to determine the state vector sequence and system matrices is briefly summarized next.

The algorithm first establishes a valid state vector sequence as the intersection of the row spaces of two block Hankel matrices constructed from the input-output data. The output block Hankel matrices (i block rows, j columns) are defined as:

$$\left[\begin{array}{c} \mathbf{Y}_p \\ \hline \mathbf{Y}_f \end{array}\right] = \left[\begin{array}{cccc} \boldsymbol{y}_k & \boldsymbol{y}_{k+1} & \cdots & \boldsymbol{y}_{k+j-1} \\ \boldsymbol{y}_{k+1} & \boldsymbol{y}_{k+2} & \cdots & \boldsymbol{y}_{k+j} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{y}_{k+i-1} & \boldsymbol{y}_{k+i} & \cdots & \boldsymbol{y}_{k+j+i-2} \\ \hline \boldsymbol{y}_{k+i} & \boldsymbol{y}_{k+i+1} & \cdots & \boldsymbol{y}_{k+j+i-1} \\ \boldsymbol{y}_{k+i+1} & \boldsymbol{y}_{k+i+2} & \cdots & \boldsymbol{y}_{k+j+i} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{y}_{k+2i-1} & \boldsymbol{y}_{k+2i} & \cdots & \boldsymbol{y}_{k+j+2i-2} \end{array}\right] \tag{2.2}$$

where $\mathbf{Y}_p$ represents past outputs and $\mathbf{Y}f$ represents future outputs, both with respect to the current time step k. Hankel matrices $\mathbf{U}_p$ and $\mathbf{U}_f$ are similarly constructed for the past and future system inputs respectively. Matrices $\mathbf{H}_1$ and $\mathbf{H}_2$ are thereby defined as the concatenation of the past output and input blocks and the future output and input blocks, respectively, as described by the following expression.

$$\mathbf{H}_1 = \begin{bmatrix} \mathbf{Y}_p \\ \mathbf{U}_p \end{bmatrix}, \qquad \mathbf{H}_2 = \begin{bmatrix} \mathbf{Y}_f \\ \mathbf{U}_f \end{bmatrix} \tag{2.3}$$

The overall Hankel matrix, $\mathbf{H}$, which captures both the past and future inputs and outputs is thus given by the concatenation of $\mathbf{H}_1$ and $\mathbf{H}_2$.

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_1 \\ \mathbf{H}_2 \end{bmatrix} \tag{2.4}$$

A valid state vector sequence, $\mathbf{X}$, is identified as the intersection of the row spaces of $\mathbf{H}_1$ and $\mathbf{H}_2$ and can be obtained via successive SVDs performed on matrix $\mathbf{H}$. Mathematically, the identified state trajectory is given as

$$\mathbf{X} = \begin{bmatrix} \boldsymbol{x}_{k+i} & \boldsymbol{x}_{k+i+1} & \cdots & \boldsymbol{x}_{k+i+j-1} \end{bmatrix} \tag{2.5}$$

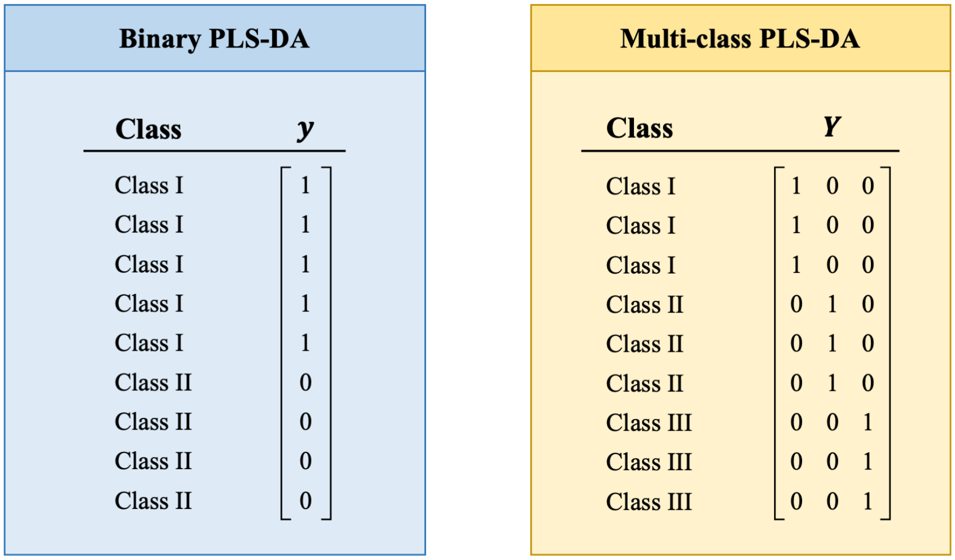
Finally, once the state trajectory $\mathbf{X}$ is determined, the system matrices (i.e., $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{D}$) are easily obtained as the least-squares solution to the below overdetermined system of linear equations describing the process.

$$\begin{bmatrix} \boldsymbol{x}_{k+i+1} & \boldsymbol{x}_{k+i+2} & \cdots & \boldsymbol{x}_{k+i+j-1} \\ \boldsymbol{y}_{k+i} & \boldsymbol{y}_{k+i+1} & \cdots & \boldsymbol{y}_{k+i+j-2} \end{bmatrix} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_{k+i} & \boldsymbol{x}_{k+i+1} & \cdots & \boldsymbol{x}_{k+i+j-2} \\ \boldsymbol{u}_{k+i} & \boldsymbol{u}_{k+i+1} & \cdots & \boldsymbol{u}_{k+i+j-2} \end{bmatrix} \tag{2.6}$$

2.3 Overview of partial least squares discriminant analysis

Discriminant analysis is a statistical tool used to classify observations into two or more discrete, mutually exclusive groups or categories which are known *a priori* [Morrison, 1969]. A classification boundary based on a linear combination of the independent (predictor) variables that best discriminates between the groups is identified with the general goal of maximizing the between-group variance relative to within-group variance [Morrison, 1969; Huberty, 1975]. The classification boundary is determined from a training data set for which group membership is known (i.e., the model is calibrated), typically for the purpose of predicting the group membership of new observations.

Partial least squares discriminant analysis (PLS-DA) is a supervised linear classification method based on PLS regression, combining both dimensionality reduction (i.e., PLS component construction) and discriminant analysis in a single algorithm [Lee *et al.*, 2018]. Similar to standard PLS, PLS-DA relates a set of response variables ($\mathbf{Y}$) to a set of predictor variables ($\mathbf{X}$); however, in PLS-DA, the $\mathbf{Y}$ block is instead comprised of dummy variables containing the class membership information for each class G in binary form (i.e., +1 if an observation belongs to group G, 0 if it does not), as illustrated in Figure 2.3 [Lee *et al.*, 2018]. When only two classes are considered ($G = 2$), class membership information is encoded within a single dummy vector $\mathbf{y}$ in which observations belonging to the first class (i.e., class $G = 1$; often called the `POSITIVE` class) are given a y-value of 1 and remaining observations are given a value of 0 (i.e., class $G = 2$; often called the `NEGATIVE` class). Multi-class problems ($G > 2$) utilize a dummy matrix $\mathbf{Y}$ with G columns which correspond to each class label. In this work, we focus only on PLS-DA for binary classification problems.



Binary PLS-DA

Class	***y***
Class I	1
Class I	1
Class I	1
Class I	1
Class I	1
Class II	0
Class II	0
Class II	0
Class II	0

Multi-class PLS-DA

Class	***Y***		
Class I	1	0	0
Class I	1	0	0
Class I	1	0	0
Class II	0	1	0
Class II	0	1	0
Class II	0	1	0
Class III	0	0	1
Class III	0	0	1
Class III	0	0	1

Figure 2.3: Response data encoding for binary PLS-DA (single dummy vector **y**) and multi-class (dummy matrix **Y**) PLS-DA problems. Adapted from Lee *et al.* [2018].

2.3.1 Model building

At its core, PLS-DA is essentially PLS regression applied to categorical response data. Mathematically, PLS-DA employs the same algorithm as PLS and, as such, the objective of PLS-DA is also to optimally explain the variation in both the **X** and **Y**-spaces individually, as well as provide the strongest possible relationship between **X** and **Y** [Dunn, 2010]. For a binary PLS-DA problem with M predictor variables and N observations, the two fundamental model equations are:

$$\begin{aligned} \mathbf{X} &= \mathbf{T}\mathbf{P}' + \mathbf{E} \\ \mathbf{y} &= \mathbf{U}\mathbf{q}' + \mathbf{f} \end{aligned} \tag{2.7}$$

where $\mathbf{X}$ $(N \times M)$ is the input data matrix, $\mathbf{y}$ $(N \times 1)$ is a vector containing class membership information, $\mathbf{T}$ $(N \times A)$ and $\mathbf{U}$ $(N \times A)$ are the scores of $\mathbf{X}$ and $\mathbf{y}$, $\mathbf{P}$ $(M \times A)$ and $\mathbf{q}$ $(1 \times A)$ are the loadings of $\mathbf{X}$ and $\mathbf{y}$, and $\mathbf{E}$ $(N \times M)$ and $\mathbf{f}$ $(N \times 1)$

are the residuals (error) associated with $\mathbf{X}$ and $\mathbf{y}$, respectively. The above equations represent the joint decomposition of $\mathbf{X}$ and $\mathbf{y}$. A schematic of the vectors and matrices defined for PLS-DA is given in Figure 2.4.

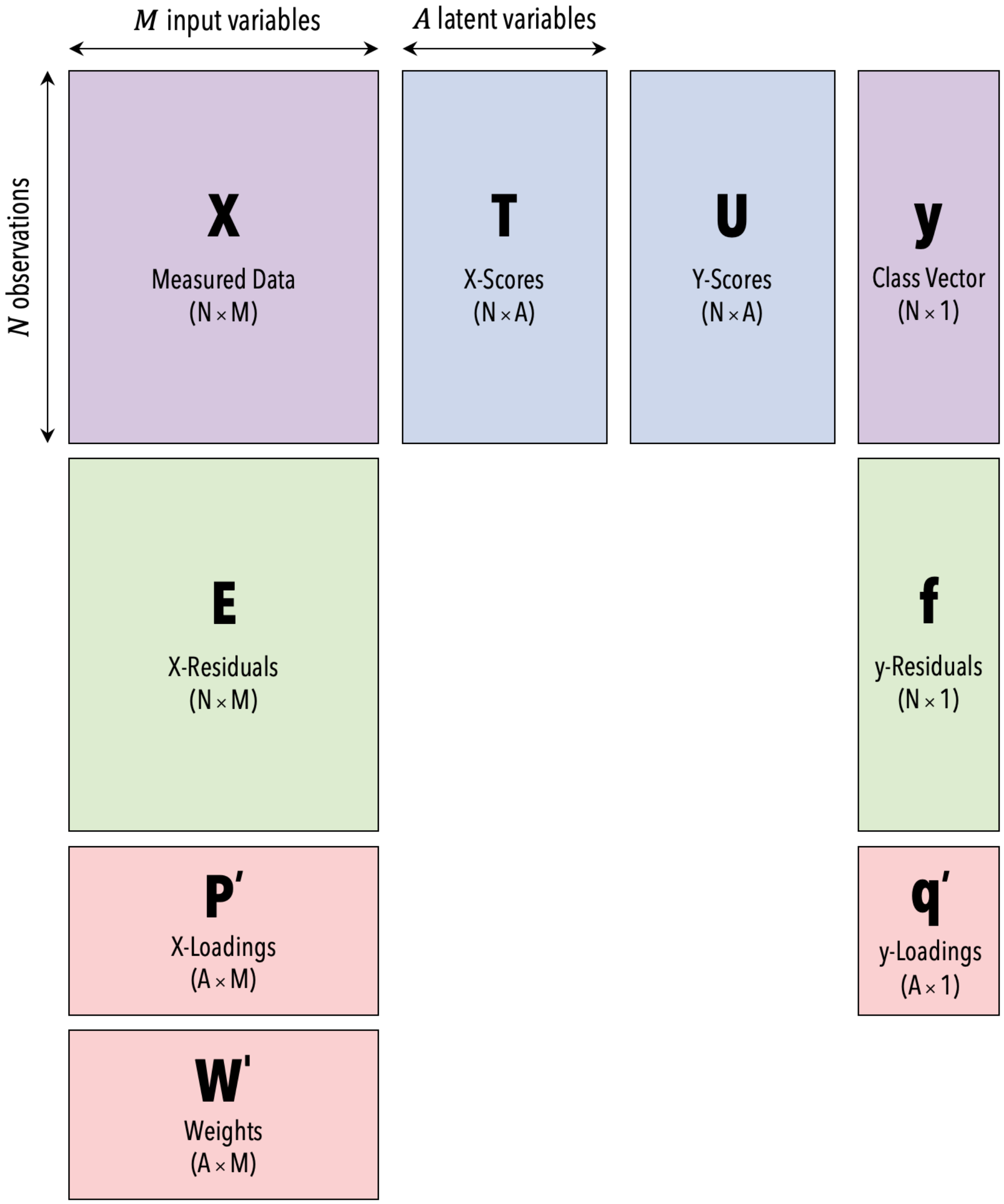


Figure 2.4: Matrices and vectors used in the PLS-DA algorithm for binary classification problems. Adapted from Geladi and Kowalski [1986].

A training dataset for which both the input data $\mathbf{X}$ and corresponding class membership vector $\mathbf{y}$ are known *a priori* is used to calibrate the PLS-DA classification model. The PLS-DA algorithm used in this work — which follows the nonlinear iterative partial least-squares (NIPALS) algorithm developed by H. Wold [1975] for computing sequential principal components — is described next [Jackson, 2005; Dunn, 2010]. Note that $\mathbf{X}$ is assumed to be appropriately preprocessed (i.e., mean-centered and scaled). We also select the response vector $\mathbf{y}$ to be the initial estimate for vector $\mathbf{u}$ during the calculation of the first component.

(1) Compute weight vector $\mathbf{w}$, which represents the slope coefficients from regressing each column m of $\mathbf{X}$ onto the score vector $\mathbf{u}$.

$$\mathbf{w} = \frac{1}{\mathbf{u}'\mathbf{u}} \cdot \mathbf{X}'\mathbf{u} \tag{2.8}$$

The weight vector is typically normalized to unit length so that we can maximize the covariance of $\mathbf{y}$ and the projection of $\mathbf{X}$ onto the *direction* $\mathbf{w}$.

$$\mathbf{w} = \frac{\mathbf{w}}{\sqrt{\mathbf{w}'\mathbf{w}}}$$

(2) Compute $\mathbf{X}$-scores, $\mathbf{t}$, which represent the slope coefficients from regressing each row n of $\mathbf{X}$ onto $\mathbf{w}$.

$$\mathbf{t} = \frac{1}{\mathbf{w}'\mathbf{w}} \cdot \mathbf{X}\mathbf{w} \tag{2.9}$$

(3) Compute $\mathbf{y}$-space loadings vector $\mathbf{q}$, which represents the slope coefficients from regressing vector $\mathbf{y}$ onto $\mathbf{w}$.

$$\mathbf{q} = \frac{1}{\mathbf{t}'\mathbf{t}} \cdot \mathbf{y}'\mathbf{t} \tag{2.10}$$

(4) Compute $\mathbf{y}$-space score vector $\mathbf{u}$, which represents the slope coefficients from regressing each row n of $\mathbf{y}$ onto $\mathbf{q}$.

$$\mathbf{u} = \frac{1}{\mathbf{q}'\mathbf{q}} \cdot \mathbf{y}\mathbf{q} \tag{2.11}$$

These four steps are iterated until $\mathbf{u}$ has reasonably converged, at which point the values of $\mathbf{w}$, $\mathbf{t}$, $\mathbf{q}$ and $\mathbf{u}$ define the a^{th} component. With this, we can compute the $\mathbf{X}$-space loadings $\mathbf{p}$ for the a^{th} component as the slope coefficients from the regression of each column m of $\mathbf{X}$ onto $\mathbf{t}$.

$$\mathbf{p} = \frac{1}{\mathbf{t}'\mathbf{t}} \cdot \mathbf{X}'\mathbf{t} \tag{2.12}$$

From here, $\mathbf{X}$ must be deflated to remove variability explained by the computed component [Dunn, 2010]. The variation in $\mathbf{X}$ captured by the a^{th} component is quantified by computing the predicted $\hat{\mathbf{X}}$ matrix as the product of $\mathbf{X}$-scores, $\mathbf{t}$, and newly calculated $\mathbf{X}$-loadings, $\mathbf{p}$. The residuals matrix $\mathbf{E}$ for the $\mathbf{X}$-space can then be determined as the difference between the $\mathbf{X}$ matrix used to calculate the component and the predicted $\hat{\mathbf{X}}$ matrix. Mathematically,

$$\hat{\mathbf{X}} = \mathbf{t}\mathbf{p}' \tag{2.13}$$

$$\mathbf{E} = \mathbf{X} - \hat{\mathbf{X}} \tag{2.14}$$

Matrix $\mathbf{E}$ represents the remaining variance in the $\mathbf{X}$-space left to be explained by additional components. Deflation of $\mathbf{y}$ and computation of the $\mathbf{y}$-space residuals vector, $\mathbf{f}$, is carried out in the same manner. As such, $\mathbf{E}$ and $\mathbf{f}$ are used in place of $\mathbf{X}$ and $\mathbf{y}$, respectively, to calculate the next component (i.e., $\mathbf{X}_{a+1} = \mathbf{E}$ and $\mathbf{y}_{a+1} = \mathbf{f}$).

$$\hat{\mathbf{y}} = \mathbf{t}\mathbf{q}' \tag{2.15}$$

$$\mathbf{f} = \mathbf{y} - \hat{\mathbf{y}} \tag{2.16}$$

Cross-validation (CV) can be employed to avoid overfitting and determine the optimal number of components to retain in the final PLS-DA model. This work employs a k-fold cross-validation methodology that randomly divides the training data into k groups and iterates through calibrating a PLS-DA model on $k-1$ groups and then testing the model on the remaining group, until each group has acted as the test set once [Wold, 1978]. A visual representation of a simple cross-validation procedure using $k = 3$ CV groups is shown in Figure 2.5. The optimal number of model components is chosen as that which yields the highest average R^2 value (i.e., fraction of variation modeled by the component) while also minimizing the predicted residual error sum of squares (PRESS) for each of the k prediction groups. The PRESS statistic is used to quantify the predictive ability of a candidate PLS-DA model structure and is mathematically expressed as:

$$\mathrm{PRESS} = \sum_{i=1}^{k}(\mathbf{y}_i - \widehat{\mathbf{y}}_i)^2 \tag{2.17}$$

The Q^2 statistic, which represents the predictive quality of the model on a test set or cross-validation group, should also be taken into account when selecting the number of components. Unlike R^2, Q^2 is not inflationary—rather, Q^2 will initially increase before either stabilizing or decreasing beyond a certain number of components. The point where Q^2 starts to decrease thus defines a threshold for the number of components, after which the model's predictive power will no longer improve [Dunn, 2010]. In general, there exists a trade-off between a model's fit (i.e., R^2) and its predictive abilities (i.e., Q^2); cross-validation helps determine the optimal balance between the two [Dunn, 2010; Esbensen and Geladi, 2010].

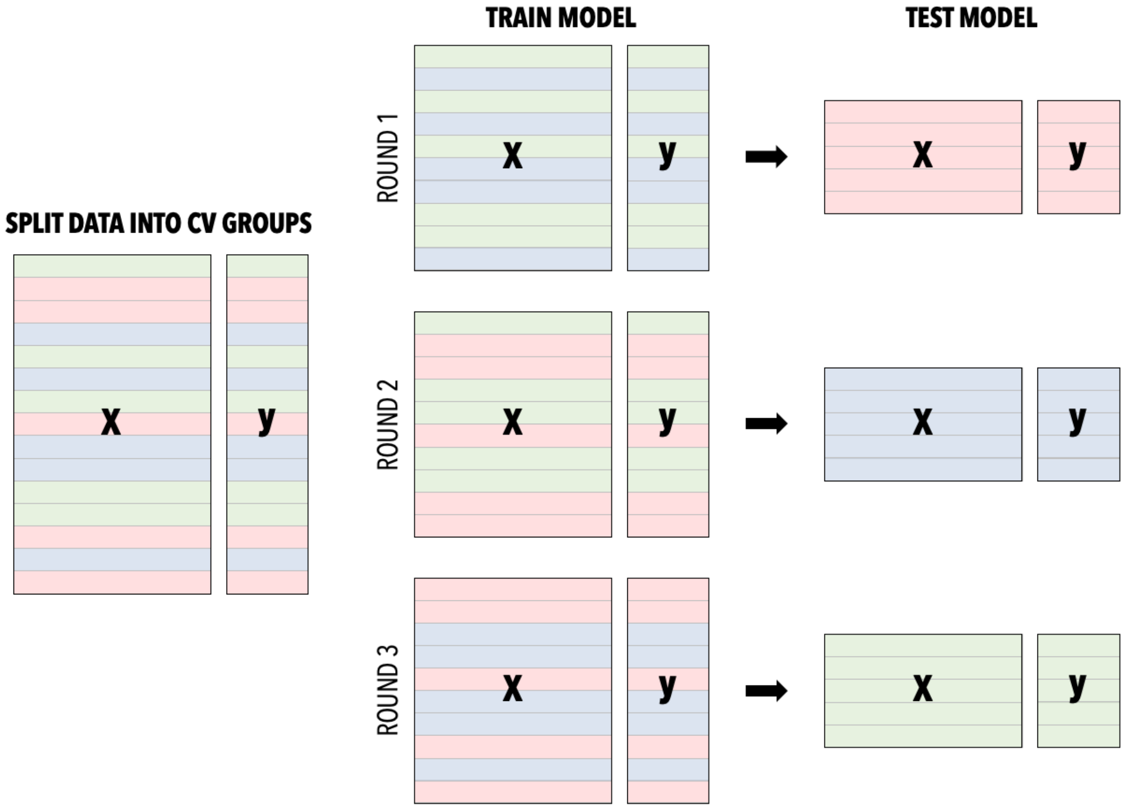


Figure 2.5: Graphical representation of a simple k-fold cross-validation procedure with $k = 3$ cross-validation groups. Adapted from Dunn [2010].

2.3.2 Predicting class membership

Once a PLS-DA model has been calibrated, the class membership for a new set of observations can be estimated via the following prediction equation:

$$\hat{\mathbf{y}} = \mathbf{X}_{new}\boldsymbol{\beta} \tag{2.18}$$

where $\hat{\mathbf{y}}$ is a vector containing the predicted class labels, $\mathbf{X}_{new}$ is the new input data matrix, and $\boldsymbol{\beta}$ is the regression coefficient vector containing information on how each input variable contributes to class membership [Lee *et al.*, 2018]. Mathematically, $\boldsymbol{\beta}$ is defined as:

$$\boldsymbol{\beta} = \mathbf{W}(\mathbf{P}'\mathbf{W})^{-1}\mathbf{q}' \tag{2.19}$$

Though class information is encoded in a binary manner during calibration, the PLS-DA model will not produce perfect binary predictions. Thus, a classification threshold (σ) must be defined such that observations with a predicted $\hat{y}$-value greater than σ are assigned to class $G = 1$ and observations with predicted values less than σ are therefore assigned to class $G = 2$ [Ballabio and Consonni, 2013]. The classification threshold is particularly sensitive to datasets with skewed class distributions. As such, σ is often strategically selected to minimize the number of false positives (i.e., observations incorrectly assigned to class $G = 1$) and false negatives (i.e., observations incorrectly assigned to class $G = 2$) [Ballabio and Consonni, 2013].

2.3.3 Assessing model performance

A confusion matrix can be used to visualize the performance of a classification model. Figure 2.6 depicts a standard confusion matrix for a binary classification problem and summarizes the four potential outcomes of classification:

- **True Positive (TP)** quantifies the number of observations correctly predicted as `POSITIVE` by the model.

- **False Positive (FP)** quantifies the number of observations incorrectly predicted as `POSITIVE` by the model (i.e., observations that actually belong to `NEGATIVE` but were predicted as `POSITIVE`).

- **False Negative (FN)** quantifies the number of observations incorrectly predicted as `NEGATIVE` by the model (i.e., observations that actually belong to `POSITIVE` but were predicted as `NEGATIVE`).

- **True Negative (TN)** quantifies the number of observations correctly predicted as `NEGATIVE` by the model.

		Actual Condition	
		POSITIVE	**NEGATIVE**
Predicted Condition	**POSITIVE**	True Positive (*TP*)	False Positive (*FP*)
	NEGATIVE	False Negative (*FN*)	True Negative (*TN*)

Figure 2.6: Standard confusion matrix for a binary classification model.

A variety of classification metrics can be readily determined from the confusion matrix including: *accuracy*, *balanced accuracy*, *balanced error rate* (BER), *sensitivity*, *specificity*, and the *Matthews Correlation Coefficient* (MCC) [Tharwat, 2020; Chicco and Jurman, 2020]. These metrics serve as model performance indicators to assess the prediction accuracy of the classification model.

In practice, *accuracy* is the most commonly used metric to evaluate classification model performance and is defined as the proportion of all observations that have been correctly classified according to their known label.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{TP + TN}{P + N} \tag{2.20}$$

where P and N represent the total number of `POSITIVE` and `NEGATIVE` observations, respectively. *Accuracy* values span the interval $[0, 1]$, with 0 indicating no correct classification and 1 indicating perfect classification. The popularity of the accuracy metric is largely due to its computational simplicity and easy interpretation; however, *accuracy* is sensitive to data with skewed class distributions and may be unreliable and misleading when applied to imbalanced class data [Chicco and Jurman, 2020].

As an alternative to the standard *accuracy* measure, *balanced accuracy* (BA) can be used evaluate true classification accuracy when dealing with class imbalance. *Balanced accuracy* is defined as the average classification accuracy attained for either class and assumes an equal cost of misclassification between the classes [Brodersen *et al.*, 2010]. Like *accuracy*, BA values also span $[0, 1]$ and can be interpreted in the same manner.

$$BA = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) \tag{2.21}$$

Derived from *balanced accuracy*, the *balanced error rate* (BER) is a measure of the proportion of observations that are misclassified by the PLS-DA model. As such, the *balanced error rate* also spans the interval $[0, 1]$ and can be interpreted inversely to *balanced accuracy*, with a low BER indicating more accurate classification.

$$BER = \frac{1}{2}\left(\frac{FP}{FP+TN} + \frac{FN}{FN+TP}\right) = 1 - BA \tag{2.22}$$

Sensitivity—also known as the *true positive rate* (TPR)—and *specificity*—also known as the *true negative rate* (TNR)—are essentially measures of accuracy for the `POSITIVE` and `NEGATIVE` classes, respectively; *sensitivity* evaluates the ability of the model to correctly predict `POSITIVE` observations, while *specificity* evaluates the ability of the model to correctly predict `NEGATIVE` observations [Tharwat, 2020].

$$Sensitivity\,(TPR) = \frac{TP}{TP+FN} \tag{2.23}$$

$$Specificity\,(TNR) = \frac{TN}{TN+FP} \tag{2.24}$$

Sensitivity and *specificity* values both range from 0 to 1 and can be interpreted similar to the *accuracy*; however, *sensitivity* and *specificity* are not sensitive to imbalanced class distributions [Tharwat, 2020].

The *false positive rate* (FPR) and *false negative rate* (FNR) can be derived from *sensitivity* and *specificity* and are defined in Equations 2.25 and 2.26, respectively. FPR complements *specificity* and represents the proportion `NEGATIVE` observations incorrectly classified as `POSITIVE`. Similarly, FNR complements *sensitivity*, representing the proportion `POSITIVE` observations incorrectly classified as `NEGATIVE`. Both the FPR and FNR also range between 0 and 1; however, contrary to *sensitivity* and *specificity*, low values are desired as they indicate a low rate of misclassification.

$$FPR = 1 - TNR = \frac{FP}{FP + TN} \tag{2.25}$$

$$FNR = 1 - TPR = \frac{FN}{FN + TP} \tag{2.26}$$

Finally, *Matthews correlation coefficient* (MMC) is a reliable measure of the correlation between actual and predicted class values and is comparable to the Pearson's correlation coefficient between two variables [Matthews, 1975; Chicco and Jurman, 2020]. The MCC accounts for both true and false positive and negatives, summarizing the entire confusion matrix in a single parameter via the following expression:

$$\text{MMC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{2.27}$$

MCC values range between -1 and +1, with -1 indicating perfect misclassification, 0 indicating completely random prediction, and +1 indicating perfect classification [Chicco and Jurman, 2020]. *Matthews correlation coefficient* is especially useful because it is one of the only metrics that accounts for not only the entire confusion matrix, but also the ratio between each element in the confusion matrix [Chicco, 2017]. MCC is therefore considered to be unaffected by skewed class distributions. In other words, MCC will only produce a high score if the classification model is predicting both `POSITIVE` and `NEGATIVE` class observations exceptionally well [Chicco and Jurman, 2020].

Chapter 3

Application of subspace identification to secondary clarifier

3.1 Description of the simulation model

This work focuses on a simplified version of the secondary clarification unit within the wastewater treatment process, as presented in Figure 3.1. Recall that the secondary clarifier serves two main functions: clarification and solids thickening. As such, influent wastewater at a certain flow rate Q_{INF} and TSS concentration X_{INF} is fed into the clarifier with the objective of removing solid particulates and producing a clear effluent. Once clarified, effluent exits the tank for discharge at a flow rate Q_{EFF} and TSS concentration X_{EFF}, while settled sludge is pumped from the bottom for disposal at a flow rate Q_{WAS} and TSS concentration X_{WAS}.

Dynamic simulations of the described clarification system are carried out in GPS-X, a wastewater treatment plant simulator developed by Hydromantis ESS, Inc. The process employs a ten-layer (each of equal thickness) circular flat-bottom settling tank with a surface area of 100 m^2 and a water depth height of 3.0 m. Influent wastewater enters at a feed point height of 1.0 m from the bottom of the tank. A simple one-dimensional (Simple-1D) nonreactive sedimentation model based on the solids flux concept is used to simulate the clarification process, in which biological reactions in the settler are ignored and the only numerically integrated variable is the suspended solids concentration. The sedimentation model employed considers only vertical flow in the settler and assumes that incoming solids are distributed instantaneously and uniformly across the entire cross-sectional area of the feed layer, as depicted in Figure 3.2. Simulations utilize the comprehensive MANTIS2LIB model library in which fifty-two state variables are available. Greater detail on the model library and sedimentation models employed by GPS-X are available in the Hydromantis ESS, Inc. [2019]. Note that the simulator is built using state-of-the-art dynamic modeling tools developed by Hydromantis and therefore captures the process nonlinearities and complexities fairly realistically. Thus, this simulator is being used as the test bed to demonstrate the applicability of subspace-based modeling techniques to the problem of wastewater treatment. The specific goal of this work is to build a model that is able to provide accurate predictions for output variables X_{EFF} and X_{WAS} based on measured data for input variables Q_{INF}, X_{INF} and Q_{WAS}.

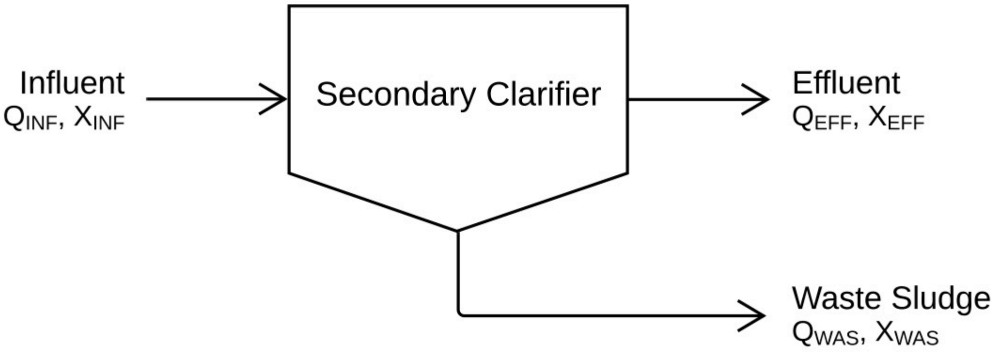


Figure 3.1: Schematic of secondary clarification unit in a WWTP.

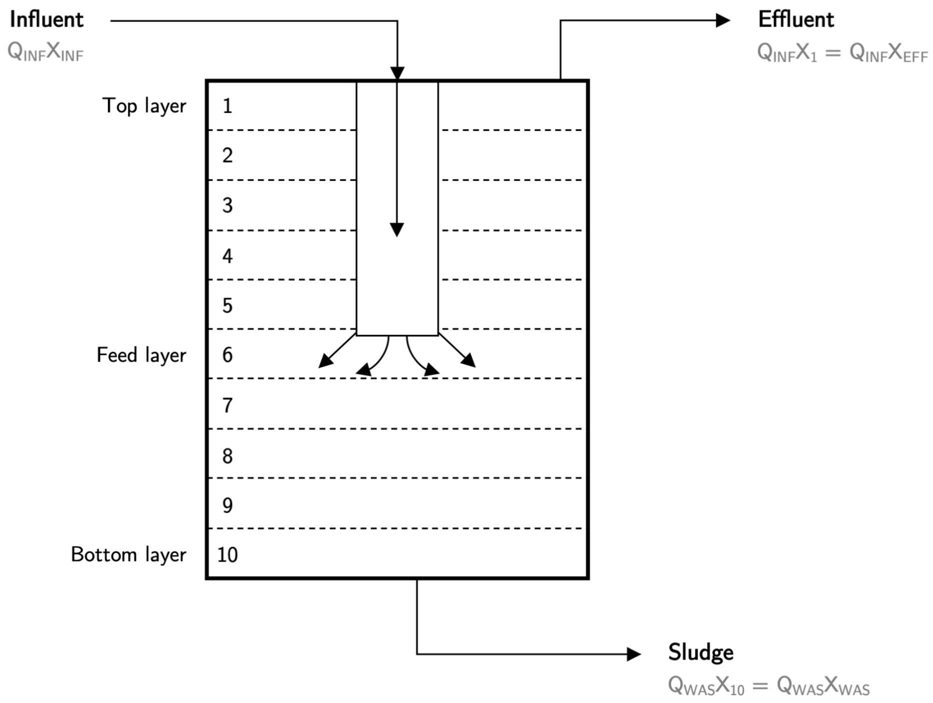


Figure 3.2: Diagram of one-dimensional sedimentation model depicting stream flows. Figure adapted from Hydromantis ESS, Inc. [2019].

3.2 Database generation

The three inputs considered for this system are: the influent flow rate of wastewater to the secondary clarifier (Q_{INF}), the influent TSS concentration (X_{INF}), and the flow rate of secondary sludge removed from the clarifier (Q_{WAS}). Input data is generated using a pseudo-random binary sequence (PRBS) and is subsequently used as the inputs to dynamic process simulations in GPS-X to collect corresponding output data for the TSS concentrations in both the clarified effluent (X_{EFF}) and the secondary sludge (X_{WAS}). The inputs are constrained within the considered ranges listed in Table 3.1.

Table 3.1: Typical input variable operating ranges compared to reduced ranges considered during data generation for subspace model identification.

Input Variable	**Typical Range**	**Considered Range**	**Unit**
Q_{INF}	240 - 4800	1500 - 3500	m^3/d
X_{INF}	1500 - 3500	1500 - 3500	mg/L
Q_{WAS}	240 - 4800	1500 - 3500	m^3/d

A PRBS signal is used to randomly generate input data for Q_{INF} and Q_{WAS} by perturbing the system various distances away from the midpoint of their valid ranges (computed from the considered ranges of the corresponding input variable). There are nine possible perturbation "levels" for the PRBS to take:

$$Levels = \begin{bmatrix} -1 & -0.75 & -0.50 & -0.25 & 0 & 0.25 & 0.50 & 0.75 & 1 \end{bmatrix}$$

where -1 represents the lower bound, 0 is the midpoint and 1 is the upper bound of the variable's valid range. Values of Q_{INF} and Q_{WAS} are constrained such that

$$Q_{INF} = Q_{EFF} + Q_{WAS}$$

where Q_{EFF} denotes the effluent flow rate from the secondary clarifier.

Remark 3.1. *The validity range of the subspace identification technique was explored and it was observed that utilizing larger input variable ranges generally resulted in a less effective predictive subspace model. This is to be expected as subspace identification inherently builds a linear model and thus the validity of the resultant model should always be checked via cross-validation to assess how well the model is able to generalize to a new dataset. If the entire range of operation is of interest, multiple subspace models can be built and connected* *such that each individual model is valid within a certain range of input variable values and is employed when the plant is operating in said range. Moreover, it is possible to identify certain key operating points via 'indicator variables' that allow you to split the data into various bins based on the value of the indicator variable and build separate working models for each bin. It is also possible that the connection between these individual models be automated, thereby enabling the switch between models to be done online. The development and connection of multiple models remains the subject of future work.*

Input data for X_{INF} is generated such that the corresponding solids loading rate (SLR) falls between 50–100 kg/m^{2}d (typical range is approximately 10–150 kg/m^{2}d). The SLR represents the amount of solids that can be removed per unit of clarifier surface area per day and is considered when generating input data for X_{INF} to ensure normal clarifier operation during process simulations. X_{INF} is related nonlinearly to SLR via the following relation:

$$SLR = \frac{Q_{INF} \cdot X_{INF}}{A} \tag{3.1}$$

where A represents the surface area of the clarifier. Thus, for every value of Q_{INF} determined by the PRBS signal, it is ensured that X_{INF} is chosen such that the SLR in turn falls within the prescribed range. To achieve this, values for X_{INF} are generated in a manner similar to Q_{INF} and Q_{WAS}; however, the allowable range for X_{INF} is updated for each observation based on the value of Q_{INF} at that point. It is ensured that a unique combination of values for all three input variables is generated over the dataset to avoid repetition between training observations. Note that the SLR expression presented here is valid only for the simplified secondary clarification process we are considering; systems including aeration must also account for the RAS recycle flow [MetCalf & Eddy, Inc. *et al.*, 2014].

The input and output datasets generated are considered to be corrupted by measurement noise to reflect the presence of sensor-induced noise inherent to real-life processes. The noise is assumed to be Gaussian white noise with zero mean and a distribution spread according to the standard deviation values given in Tables 3.2 and 3.3, which are chosen based on the variable ranges.

Table 3.2: Noise parameters for input variables.

Input Variable	$\boldsymbol{\sigma}_{\textbf{noise}}$	**Unit**
Q_{INF}	5.0	m^3/d
X_{INF}	5.0	mg/L
Q_{WAS}	5.0	m^3/d

Table 3.3: Noise parameters for output variables.

Output Variable	$\boldsymbol{\sigma}_{\textbf{noise}}$	**Unit**
X_{EFF}	0.075	mg/L
X_{WAS}	10.0	mg/L

The input profiles constructed for both the training and validation datasets are presented in Figure 3.3. Input data is generated over seven days (10,080 minutes) for the training dataset, which is used to identify the subspace model, and over two days (2880 minutes) for the validation dataset, which is used to evaluate the quality of the model and its predictions. Both the training and validation input sequences are generated using a sampling time of $\Delta t = 5$ minutes and hold the inputs constant for 8-hour time periods. The corresponding training and validation output trajectories obtained via GPS-X simulations are shown in Figure 3.4.

Prior to model identification, a standard moving average filter with a window size of five observations is employed to mitigate the effect of the noisy signals on the estimated model. The effect of both the added measurement noise and the implemented filter is demonstrated on the validation data in Figures 3.5 and 3.6.

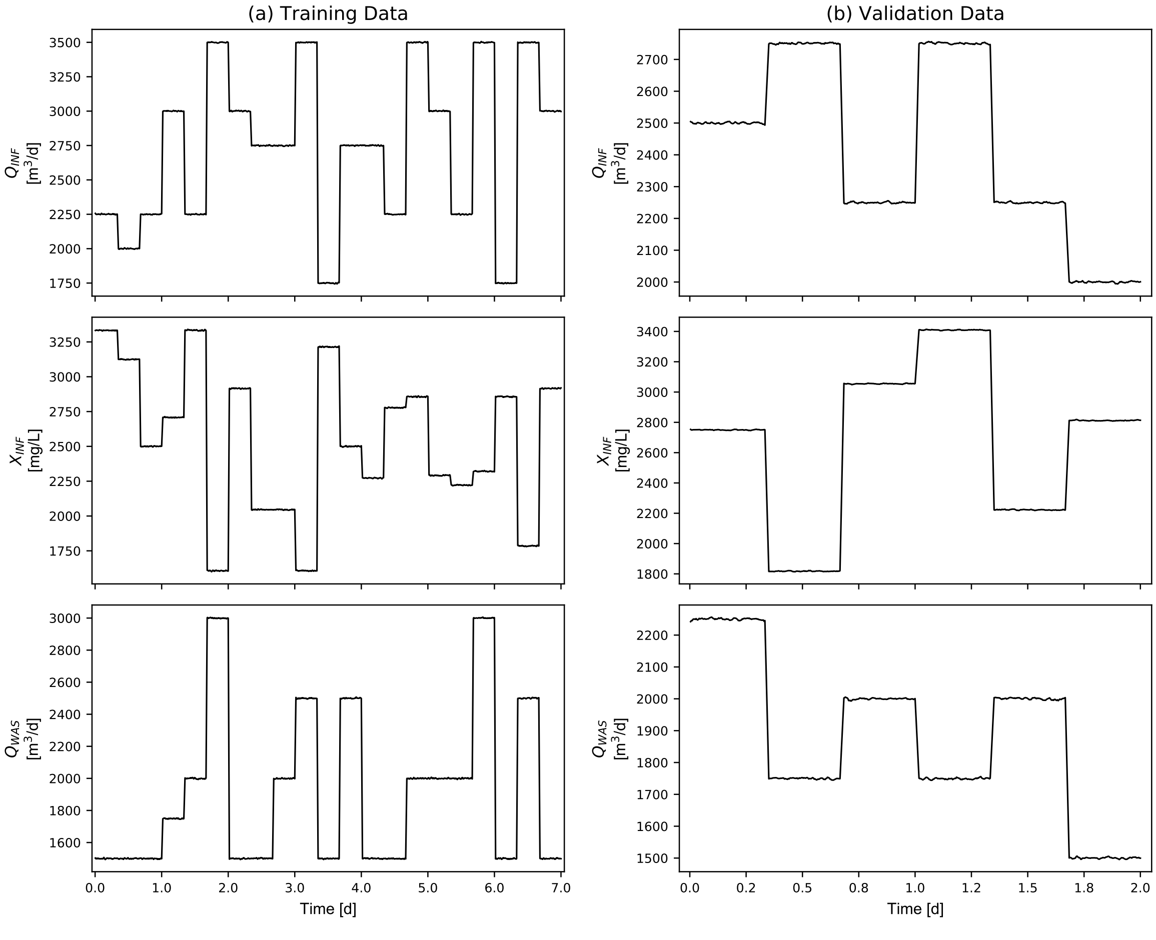


Figure 3.3: Input variable profiles for (a) training and (b) validation datasets.

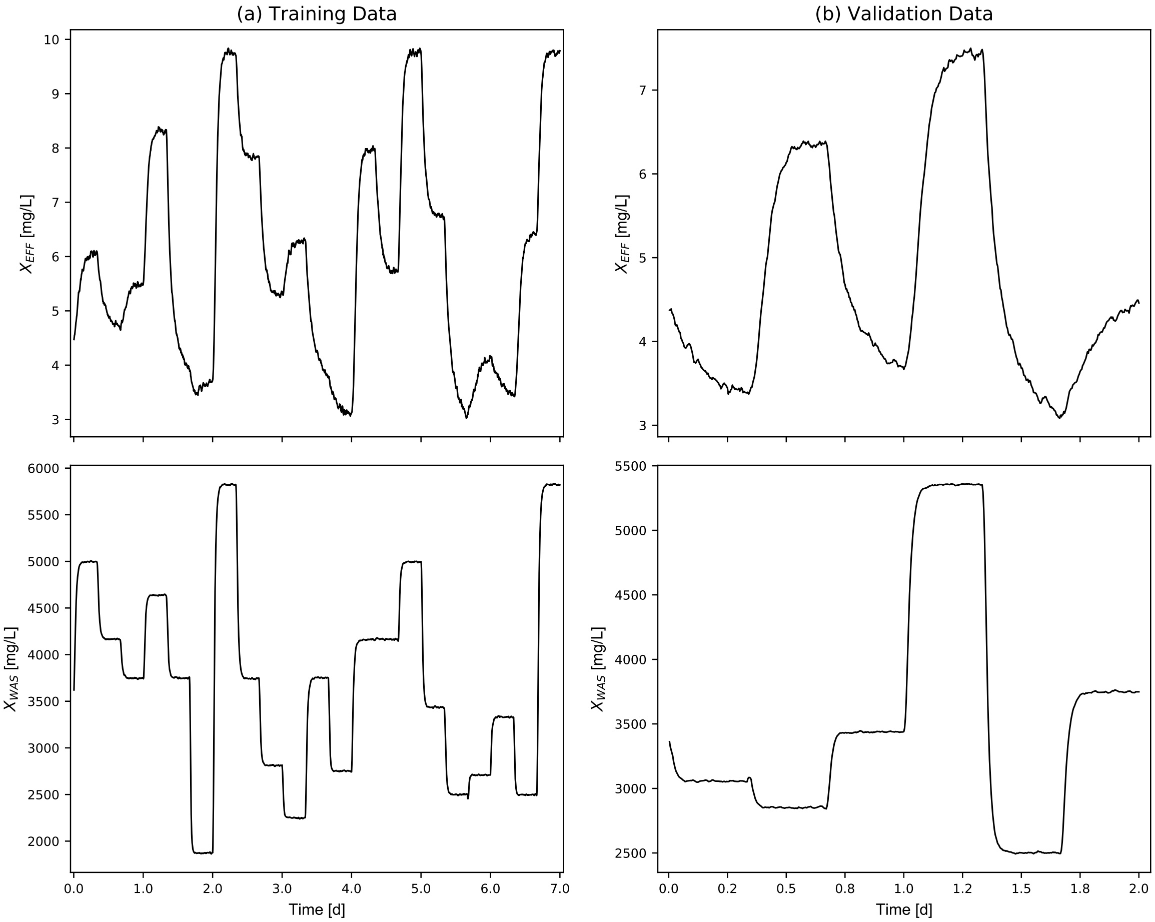


Figure 3.4: Output variable profiles for (a) training and (b) validation datasets.

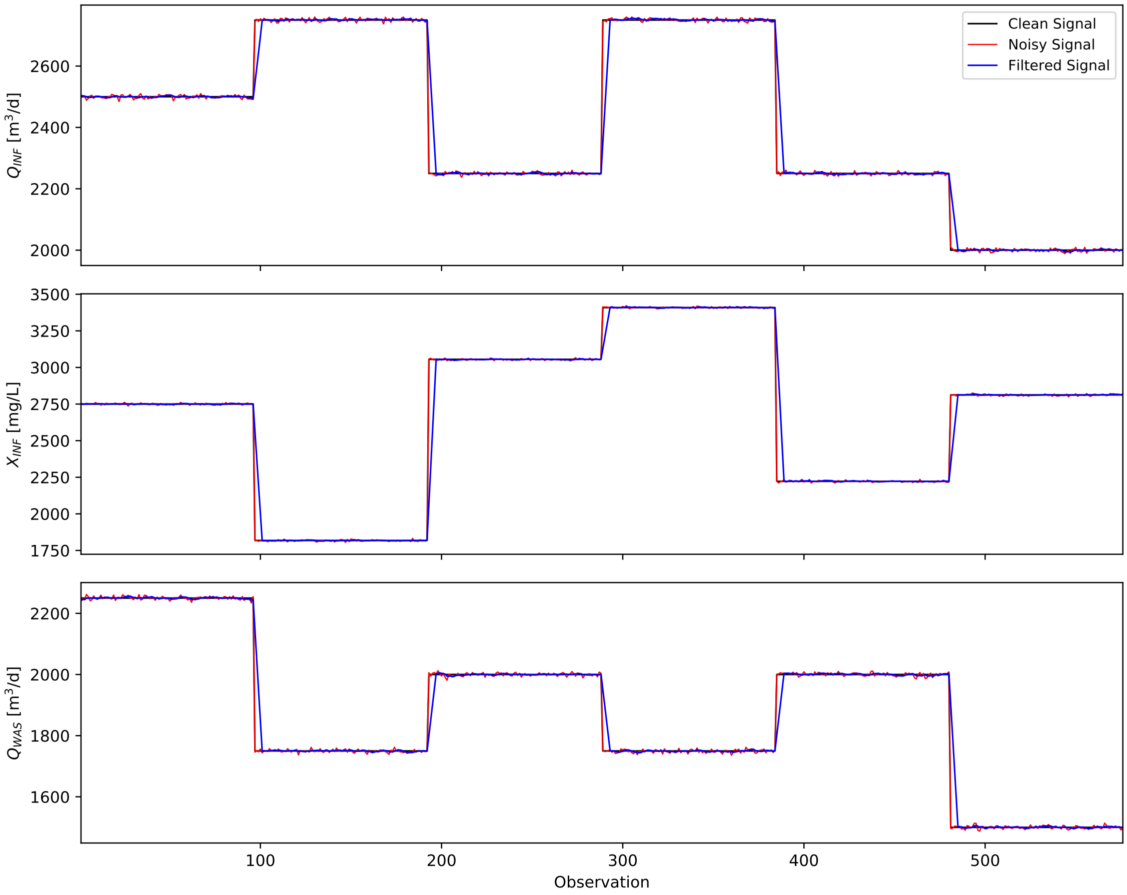


Figure 3.5: Effect of added measurement noise (red) and moving average filter (blue) on generated validation input profiles (black).

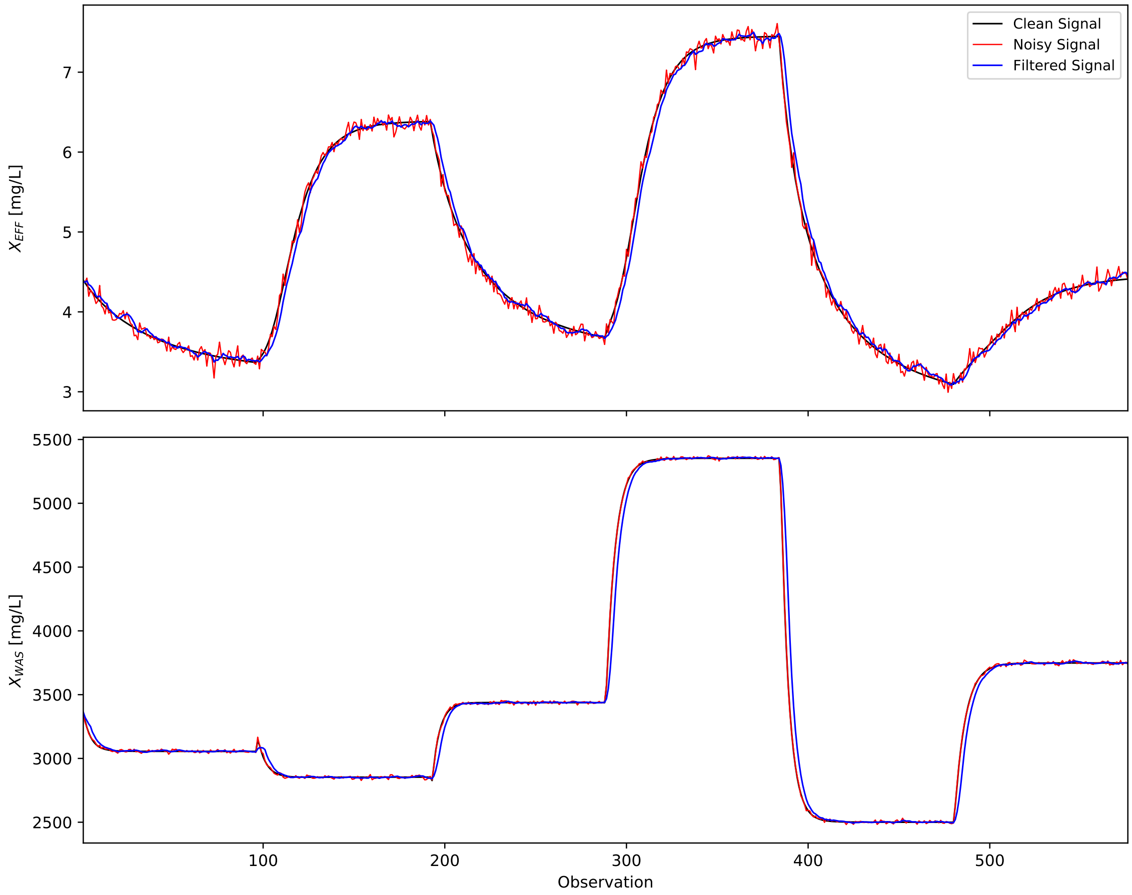


Figure 3.6: Effect of added measurement noise (red) and moving average filter (blue) on simulated validation output profiles (black).

3.3 Identification of subspace model

This work employs the deterministic subspace identification method proposed in Moonen *et al.* [1989], as outlined in Section 2.2. This algorithm makes heavy use of Hankel matrices and SVD to establish a state trajectory from input-output data before calculating the system matrices as the least-squares solution to the set of linear system equations describing the process (Equation 2.1). The number of block rows (i) and columns (j) in the Hankel matrices are user-defined parameters which are chosen to be relatively large such that $j \gg i$ to capture sufficient system information. In practice, it is common to set $i = n + 1$ so that the number of block rows is at least larger than the number of states we want to identify. We also choose $j = N - 2i + 1$ to ensure that all N training observations are used [van Overschee and de Moor, 1996].

A linear subspace model of the described secondary clarification system is identified from the training dataset, which is preprocessed to zero mean and unit variance. A system order of $n = 5$ is selected based on a five-fold cross-validation procedure implemented on the training data in combination the model's prediction accuracy given a separate unique test set. Cross-validation results report similar cumulative prediction error values anywhere from five to nine subspace states, with only a small difference in error between each. As such, the lowest number of states with an acceptably small cumulative error value is chosen to avoid the possibility of overfitting. With this, a unique dataset is used to test the model predictions to further confirm a fifth-order model is acceptable for this system.

> **Remark 3.2.** *The choice of system order is particularly important for fitting a relevant LTI to process data. For linear systems in particular, overfitting during model construction can be avoided by examining the singular values of the Hankel matrix H, which can be obtained via SVD. Arranged in descending order, model order can be approximated by computing and comparing the ratio*

> *between consecutive singular values and then setting the model order equal to the index number corresponding to the maximum ratio value. The singular values of H provide a clear metric for determining the appropriate number of states for linear processes—even before cross-validation—and help to visualize the value that additional states add to the process model. The wastewater treatment process considered in this work, however, is significantly nonlinear and, consequently, there is no drastic "drop-off" observed in the singular values. As such, this approach was used as a starting point to find a reasonable range for the model order before performing cross-validation and selecting the number of states that minimizes prediction error.*

Model quality is evaluated by comparing the model response to the observed process outputs for the same input profile. To this end, the subspace model outputs are computed accordingly for each observation k by evaluating the linear output equation $\hat{y}_k = Cx_k + Du_k$ utilizing the identified C and D matrices. Figure 3.7 shows how the subspace model (orange) fits to the process data (black) used to train the model.

Identification error is quantified by calculating the root mean squared error (RMSE) between the process outputs y and model responses $\hat{y}$ for both output variables. RMSE is calculated via the following equation:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (y_i \;-\; \hat{y}_i)^2}{N}} \tag{3.2}$$

where N represents the number of measurements in the dataset. RMSE values measuring the error between the observed process outputs and the identified model are presented in Table 3.4. Note that the reported error values are dimensionless as the RMSE is quantified using the standardized values for both the process data and predictions, prior to returning the data to its original scale and units.

Table 3.4: RMSE values for identification results.

Output Variable	**RMSE**
X_{EFF}	0.0149
X_{WAS}	0.0070

Remark 3.3. *It is important to note that the present system does not directly consider purely manipulated inputs—rather, the ability to build an effective subspace model that can be utilized in future work to model the wastewater treatment process in the context where manipulated inputs are considered. In practice, current input variables Q_{INF} and X_{INF} frequently fluctuate based on the load and composition of influent wastewater to the plant from the surrounding community. Q_{WAS}, however, may possibly be treated as a manipulable variable and is often determined as a percentage of the wastewater flow rate entering the clarifier such that the sludge accumulated at the bottom of the tank does not exceed a certain critical level. Future work will look at building subspace models with additional manipulated inputs and instead treating the current inputs as disturbances to still exploit the important information and relationships captured by these variables.*

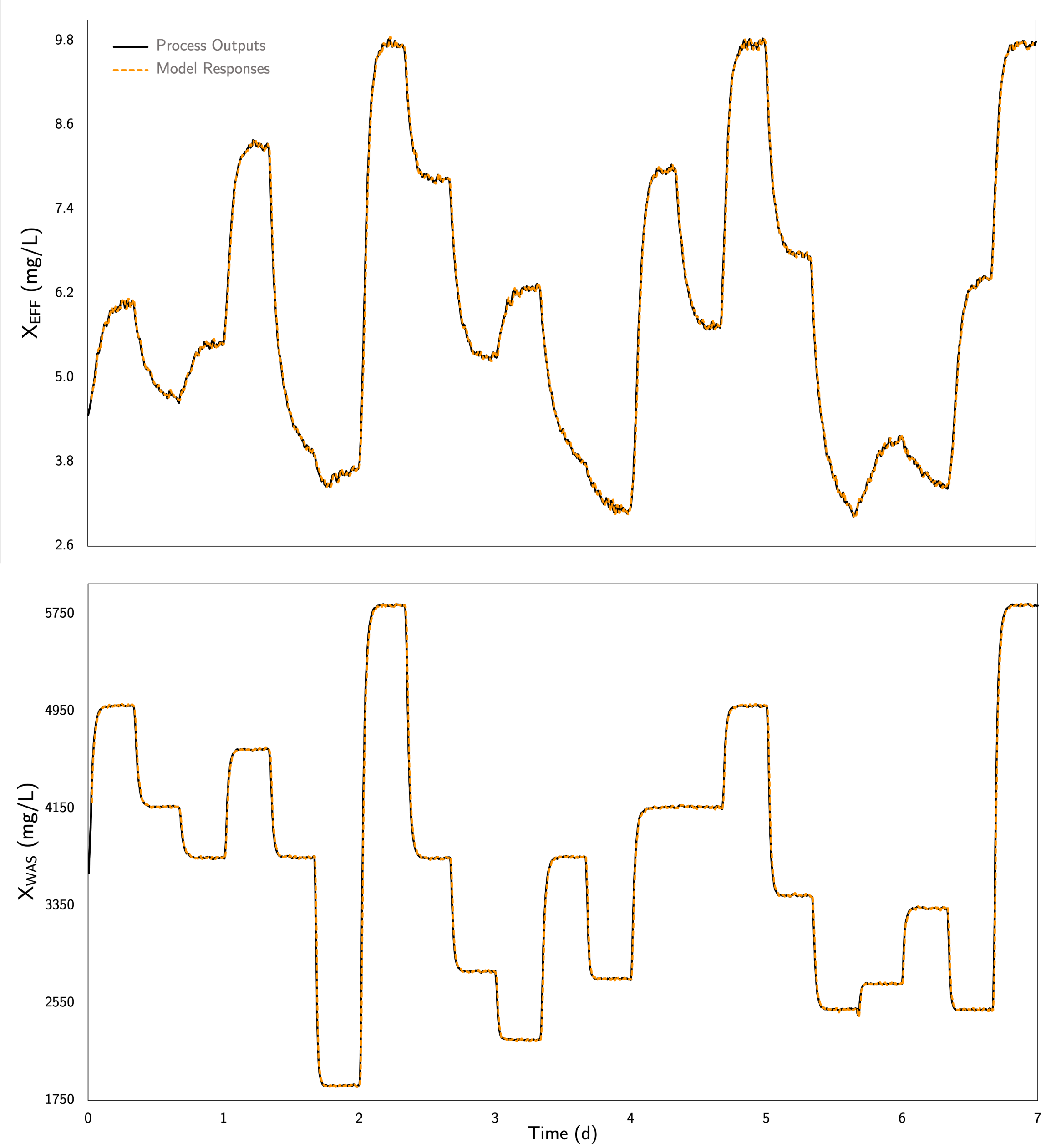


Figure 3.7: Model identification results; process outputs are in black and subspace model responses are given by the orange dashed line.

3.4 Model validation and prediction results

The identified model must be tested on a new, unique dataset to evaluate how accurately the subspace model is able to reproduce the dynamic system behaviour. Note that the identified model is a state-space model, where the number of states is typically higher than the number of measured outputs and, as such, the initial value of the states cannot be directly calculated from the measured outputs alone. Thus, the method requires some initial observations from a new dataset in order to estimate the states of the subspace model before being able to predict further.

Prior to prediction, the model must be initially calibrated to measured process data using a state observer. The use of a state observer essentially allows the subspace model to be continuously calibrated as new measurement data becomes available. This calibration period continues until the model's predicted outputs have reasonably converged to the measured outputs, after which the model is allowed to predict the output variable behaviour for a candidate input trajectory. In this work, a standard Luenberger observer is employed as the state observer in order to estimate the subspace state trajectory. During the calibration period, the state estimation equation is corrected with feedback from the estimation error between the measured outputs y_k and predicted outputs $\hat{y}_k$, as described by the following modified system equations.

$$\begin{aligned} \hat{x}_{k+1} &= A\hat{x}_k + Bu_k + L(y_k - \hat{y}_k) \\ \hat{y}_k &= C\hat{x}_k + Du_k \end{aligned} \tag{3.3}$$

where L is the observer gain matrix and is designed such that $(A - LC)$ is stable and the observer poles are placed appropriately in the unit circle. Present work considers the Luenberger observer to have converged if model predictions are within 2.5% of the process, after which the observer is "deactivated" and the subspace model is allowed to predict the output trajectories for the remainder of the validation input profile.

Remark 3.4. *The need for a state estimation period is not a limitation of the current subspace identification approach used in this work. While it is entirely possible to identify an ARX (Auto-Regressive with eXogenous variables) or RNN (Recurrent Neural Network) model from the available data, neither model structure offers an express advantage over the current subspace approach. ARX models are simple input-output models used to describe process dynamics by relating current output observations to past input and output data via a least-squares regression. Bypassing the need for state variables entirely, the ARX model structure necessitates the collection of past input and output measurements to ensure prediction accuracy and thus still requires a calibration period similar to subspace models prior to prediction. RNN models, on the other hand, use a set of internal states to process a sequence of inputs in order to recursively update the states and predict the output trajectory. Similar to linear subspace identification, the states in an RNN model have no physical, interpretable meaning and thus require state initialization in order to predict effectively. Therefore, neither ARX nor RNN models have a definitive advantage over subspace identification regarding the necessary calibration period prior to predicting.*

The predictive ability of the subspace model is evaluated by comparing the model predictions to the observed process outputs for the same input signal. The state observer estimations are found to converge to the process outputs reasonably after roughly 36 time-steps (around 3.0 hours). Figure 3.8 shows the predicted output trajectories using the identified subspace model from three different instances: (a) after 3.0 hours at time-step $k = 36$ (blue), (b) after 13.0 hours at $k = 156$ (green) and (c) after 23.0 hours at $k = 276$ (red). One of the primary goals of this work is to construct a subspace model capable of producing accurate multi-step predictions. A good predictive model is imperative for building successful model predictive control

(MPC) schemes. An effective MPC formulation internally utilizes the model and a candidate future input profile to predict the future process outputs over a specified prediction horizon and subsequently computes and implements an appropriate control action to drive the process toward a desired objective or operating point. As such, the state estimation loop is opened at three different instances to demonstrate the ability of the subspace model to predict dynamic process behaviour over a specified prediction horizon — ten hours (120 time-steps) — from various points in the process. Figure 3.8 demonstrates the success of the subspace model in predicting future output trajectories for multiple candidate input profiles throughout the validation dataset and motivates its candidacy for use in MPC applications.

Model effectiveness—reflected via prediction error—is quantified by calculating the RMSE between the process outputs y and the subspace model predictions $\hat{y}$ for both output variables. Prediction error is calculated from the end of the calibration period onward and is scaled to reflect the number of observations in the prediction set. The RMSE values for the validation results are presented in Table 3.5. Note that the reported error values are once again dimensionless as the RMSE is calculated using standardized data.

Table 3.5: RMSE values for validation results.

Calibration Period	**RMSE**	
	$\boldsymbol{X_{EFF}}$	$\boldsymbol{X_{WAS}}$
3.0 hours	0.143	0.098
13.0 hours	0.144	0.073
23.0 hours	0.159	0.071

The validation RMSE values are sufficiently low for both output variables, indicating that the subspace model is able to predict the dynamic variable behaviour effectively.

The validation results presented here demonstrate the viability of using the developed subspace model for feedback control. The ability of the subspace model to predict the process behaviour appropriately each time the state observer loop is opened adequately illustrates the potential utility of this model within an MPC framework.

Remark 3.5. *Improving subspace model predictions via hybrid modeling strategies is an attractive approach to account for the available process knowledge captured by existing first-principles models. First-principles models can either be used in series with data-driven models [Anderson* et al.*, 2000]—in which a data-driven model is used to determine the parameters of a first-principles model—or in parallel [Ghosh* et al.*, 2019]—in which a dynamic data-driven model is built around the residual error between measured process outputs and the first-principles model outputs. Hybrid modeling allows for the nonlinear process behaviour to be captured through first-principles relations, with the remaining dynamics described by the subspace model. In addition, the incorporation of first-principles models via hybrid models can help ensure that any physical process constraints are taking into consideration. Hybrid models offer many advantages over purely data-driven approaches and will be explored in future work.*

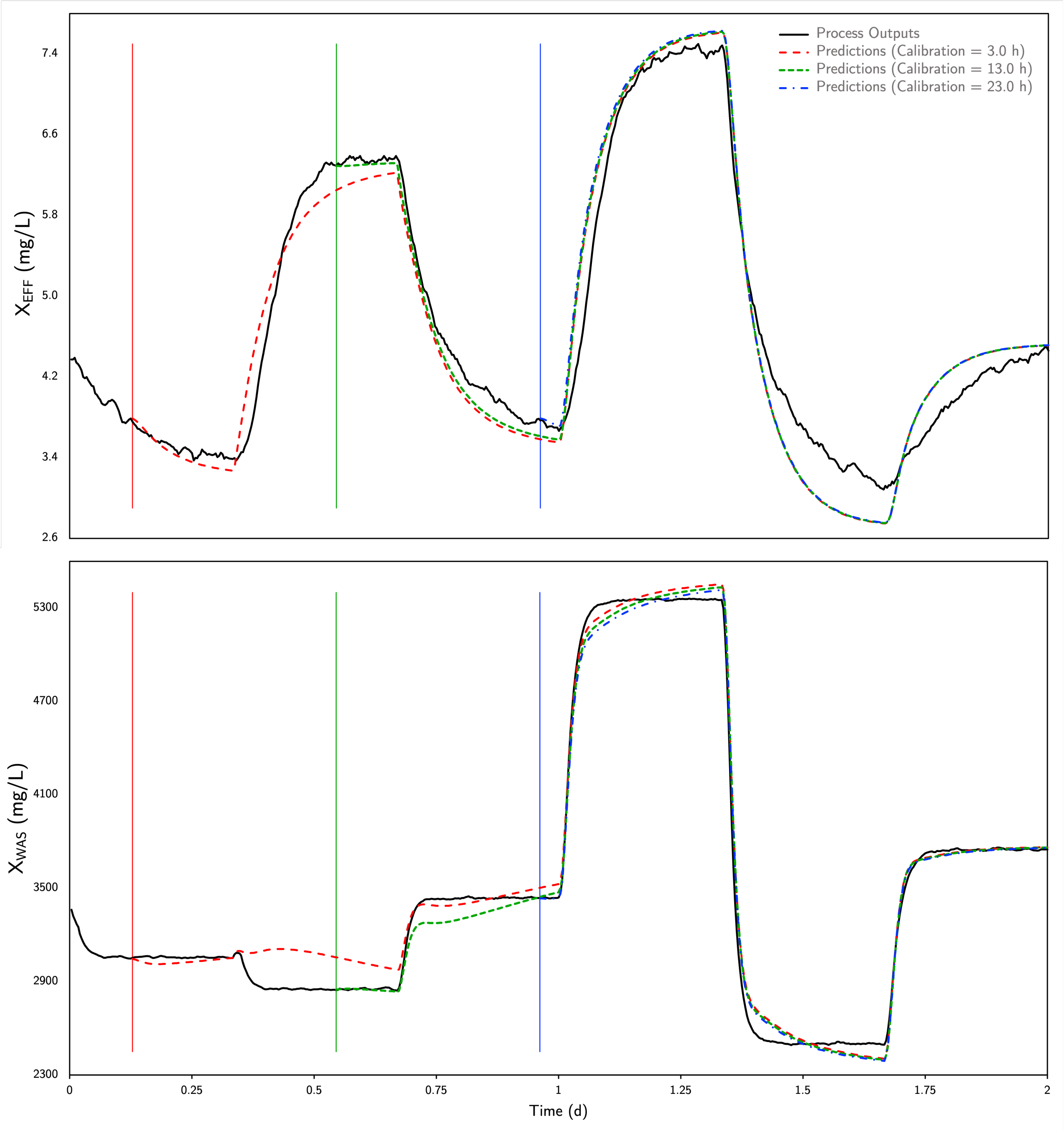


Figure 3.8: Predicted output trajectories by subspace model with three different calibration period durations: (a) 3.0 hours (red), (b) 13.0 hours (green), and (c) 23.0 hours (blue). Process outputs are given in black.

3.5 Summary of contribution

The application of subspace identification for estimating an appropriate dynamic model of the secondary clarification unit in a wastewater treatment plant was explored in this chapter. It has been shown that it is possible to estimate a relevant linear time-invariant state-space model of the secondary clarification process that is able to effectively predict the total suspended solids concentrations in both the final effluent and the waste sludge streams from the secondary clarifier with minimal prediction error, as illustrated by the validation results. Results also highlight the ability of the subspace model to produce accurate multi-step output trajectory predictions for multiple candidate input profiles and, as a result, establish the candidacy of the identified subspace model for use in a model predictive control framework.

Chapter 4

Application of PLS-DA for predicting clarifier failure

4.1 Failure of clarification unit

Clarifier failure occurs when an excessive amount of suspended solids are present in the clarifier effluent. The occurrence of high effluent suspended solids (ESS)—and therefore clarifier failure—can often be linked to poor settling performance of secondary sludge [Torfs *et al.*, 2016].

There are typically four distinct types of settling that take place within the secondary clarifier: discrete (Type I), flocculent (Type II), hindered (Type III) and compression (Type IV); each differing in sedimentation rate based on the size, density, concentration, and degree of interaction between suspended solids particles [Water Environment Federation, 2005]. Figure 4.1 depicts the four settling zones that occur during the sedimentation process and shows how the height of the sludge interface progresses with time during a standard batch column settling test.

Type I (discrete) and Type II (flocculent) settling describe the actual separation of solids particles from the water and therefore largely contribute to the clarification function of the secondary clarifier. Type I settling describes the tendency of solids

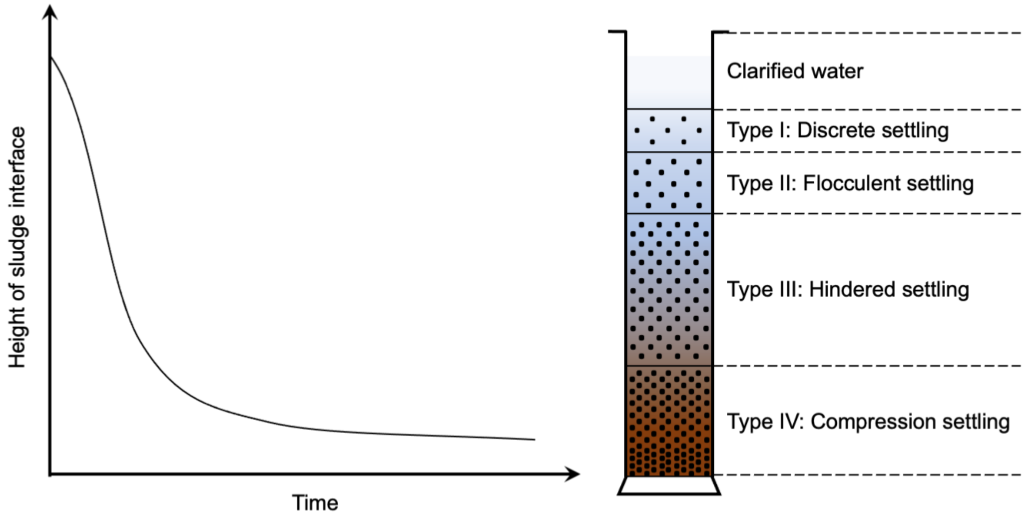

Height of sludge interface
Time
Clarified water
Type I: Discrete settling
Type II: Flocculent settling
Type III: Hindered settling
Type IV: Compression settling

Figure 4.1: Sedimentation profile showing evolution of sludge interface height with time and corresponding observable settling regions during a column settling test.

particles to settle independently at their individual terminal velocity, free from physical interaction with other particles [Water Environment Federation, 2005]. Type I settling predominantly occurs in the upper region of the clarifier due to the considerably low solids concentration present here. Type II (flocculent) settling occurs in the region just below Type I near the influent well as particles start to interact and coalesce to form larger flocs, thus increasing the settling velocity of the aggregate [Water Environment Federation, 2005]. Type III (hindered) settling takes place in somewhat of a transitional region, in which particles start to settle collectively as a matrix at the same velocity. It is here that a distinct solids-liquid interface starts to form, below which characterizes the sludge blanket [Torfs *et al.*, 2016]. Finally, Type IV (compression) settling takes place at the bottom of the clarifier, where TSS concentrations are so high that the sludge blanket is forced to compact and thicken by compression under the weight of the above solids particles [Torfs *et al.*, 2016]. As such, Type III and Type IV settling dominates the thickening function of the clarifier.

Settling failure can be attributed to factors such as a high sludge blanket height, flocculation problems, and poor hydrodynamics [Water Environment Federation, 2005; Torfs *et al.*, 2016]. If the rate of sludge removal from the clarifier is too low with respect to the influent flow rate, the sludge blanket will propagate to the surface of the tank and cause solids to exit through the effluent. Moreover, a high sludge blanket in combination with a low solids removal rate will promote the formation of anaerobic conditions and ultimately lead to denitrification and sludge flotation [Water Environment Federation, 2005]. Similarly, dispersed solids particles that fail to flocculate during sedimentation lack the critical mass necessary to settle and subsequently thicken, consequently escaping with the effluent. Hydraulic instability due to excessive turbulence (often caused by high flows) can also impact the settling process by breaking up existing flocs and cause solids particles to redistribute and resuspend throughout the clarifier.

The sludge volume index (SVI) is one of the most commonly used means of monitoring and characterizing sludge settleability in the secondary clarifier. SVI is is a physical parameter of the sludge that is determined via simple laboratory tests and is defined as the volume occupied by one gram of sludge after allowing a one-litre sample of wastewater to settle in a settleometer (typically a 1 or 2-L graduated cylinder) for 30 minutes [Dick and Vesilind, 1969]. Mathematically, SVI is calculated as:

$$\text{SVI}\left(\frac{\text{mL}}{\text{g}}\right) = \frac{\text{Volume of Settled Sludge } (\frac{\text{mL}}{\text{L}})}{\text{Suspended Solids Concentration } (\frac{\text{mg}}{\text{L}})} \times 1000\,\frac{\text{mg}}{\text{g}} \tag{4.1}$$

In general, lower SVI values indicate good sludge settleability, with an SVI range of 80–150 mL/g typically used as the benchmark to produce high quality effluent [Henze, 2002; Medora Corporation, 2016]. SVI values lower than 80 mL/g usually indicates a dense secondary sludge (often old and over-oxidized). Sludge at this low SVI tends to initially settle quite rapidly; however, the dense solids particles often cannot form large flocs as they settle, thereby preventing the formation of a uniform sludge blanket and resulting in excessive turbidity in the supernatant water [Medora Corporation, 2016]. Conversely, sludge at SVI values greater than 150 mL/g is typically much less dense and appears to be light and fluffy. This sludge is able to form some flocs during settling; however, they tend to settle very slowly (or not at all) and compacts poorly at the water-solids interface [Medora Corporation, 2016]. It is important to note that while SVI is more an intrinsic parameter that largely depends on the properties of the sludge itself, it can be controlled by manipulable variables such as the waste sludge rate or the rate of sludge recycle in a conventional activated sludge process.

Clarifier performance as a whole is most often described in terms of effluent solids concentration. The solids removal efficiency of a clarifier can be calculated as

$$\text{Solids Removal Efficiency (\%)} = \frac{X_{INF} - X_{EFF}}{X_{INF}} \times 100\% \tag{4.2}$$

where X_{INF} and X_{EFF} are the influent and effluent TSS concentrations respectively. For the purpose of this work, clarifier failure is defined in terms of effluent TSS. More specifically, the clarifier is considered to have failed if process simulations return an effluent solids concentration greater than or equal to 50 mg/L (i.e., a separation efficiency less than 97%).

4.2 Database generation

This work utilizes the same secondary clarifier layout as presented in Figure 3.1. Process data for this system is also generated in a manner similar to the approach previously introduced in Section 3.2, albeit with a few modifications to reflect the new problem. In particular, we are no longer generating dynamic data, rather static points which represent steady-state conditions.

This system considers four process inputs — Q_{INF}, X_{INF}, Q_{WAS} and SVI — and a single measured output — X_{EFF}. A PRBS signal is once again used to randomly generate input data for Q_{INF} and Q_{WAS} by perturbing the system various distances away from the midpoint of their valid ranges, as outlined in Table 4.1. Compared to the previous chapter, larger variable ranges are utilized here to include conditions that are sure to effect clarifier failure.

Table 4.1: Typical input variable ranges compared to the reduced ranges considered during data generation for the PLS-DA classification model.

Input Variable	**Typical Range**	**Considered Range**	**Unit**
Q_{INF}	240 - 4800	240 - 4800	m^3/d
X_{INF}	1500 - 3500	1500 - 3500	mg/L
Q_{WAS}	240 - 4800	240 - 4800	m^3/d
SVI	50 - 400 [a]	100 - 200	mL/g

[a] Sourced from Torfs *et al.* [2016]

There are 21 possible perturbation "levels" for the PRBS to take in the range $[-1, 1]$, where -1 represents the lower bound, 0 is the midpoint and 1 is the upper bound of the variable ranges.

$$Levels = \begin{bmatrix} -1 & -0.9 & -0.8 & \cdots & 0 & \cdots & 0.8 & 0.9 & 1 \end{bmatrix}$$

Note that the values of Q_{INF} and Q_{WAS} remain constrained such that the influent flow rate is equal to the sum of the effluent and waste sludge flows. Input data for X_{INF} is also generated using a PRBS; however, due to its nonlinear relation to Q_{INF} through the SLR (recall Equation 3.1), the valid range for X_{INF} is updated for each observation based on the chosen Q_{INF} values such that the SLR falls between 10–150 kg/m^2d.

A unique combination of Q_{INF}, X_{INF} and Q_{WAS} are generated for each SVI value considered between 100–200 mL/g, increasing in increments of 10 mL/g. 1200 data points are created at each individual SVI level (1000 for training, 200 for testing), resulting in a total of 13,200 data points generated.

Steady-state simulations of the clarifier are carried out in GPS-X utilizing the same process conditions and clarifier parameters described in Section 3.1 to obtain output data for X_{EFF} which corresponds to the generated inputs. Each individual data point specifies the input conditions for a discrete GPS-X simulation which is allowed to reach steady-state over a span of eight days; thus, essentially 13,200 unique simulations are executed. A four-dimensional scatter plot is presented in Figure 4.2 to visualize the effects of input variables Q_{INF} (x-axis), X_{INF} (y-axis) and Q_{WAS} (z-axis) on the effluent TSS concentration (colour) for a representative SVI value of 150 mL/g.

Using the obtained X_{EFF} data, each observation is classified either as `FAILED` or `NORMAL` operation based on whether or not X_{EFF} exceeds 50 mg/L. Figure 4.3 complements Figure 4.2 and presents a four-dimensional scatter plot showing how the effects of Q_{INF} (x-axis), X_{INF} (y-axis) and Q_{WAS} (z-axis) on an observation's class label (colour), again at a representative SVI value of 150 mL/g.

Table 4.2 summarizes the number of observations classified as `FAILED` (i.e., clarification failure) and `NORMAL` (i.e., effective clarification) in the datasets used to train and test the PLS-DA classification model at each of the considered SVI values between 100–200 mL/g.

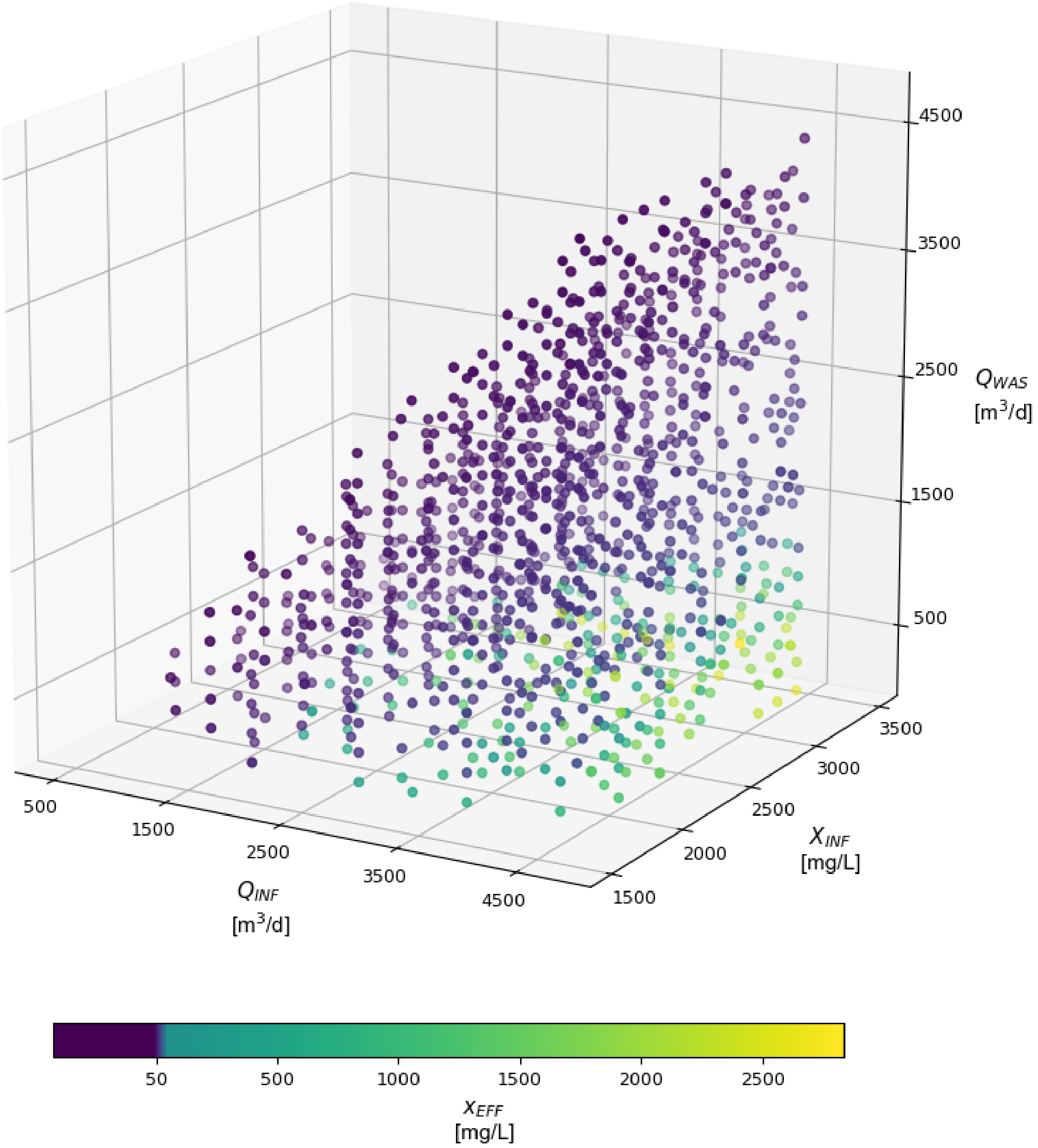


Figure 4.2: 4D scatter plot showing effect of inputs Q_{INF} (x-axis), X_{INF} (y-axis), and Q_{WAS} (z-axis) on X_{EFF} (colour) at a representative SVI value of 150 mL/g.

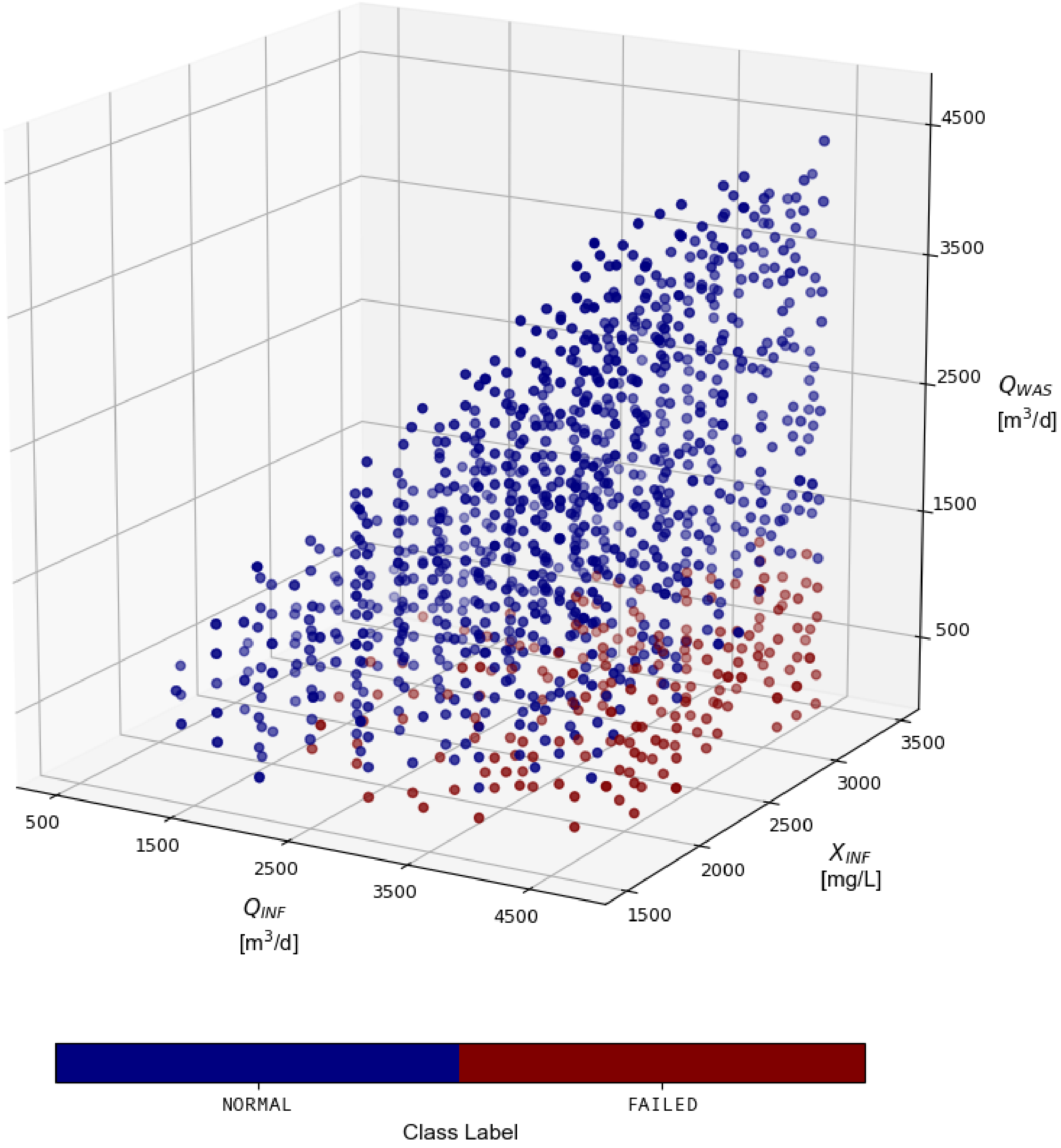


Figure 4.3: 4D scatter plot showing effect of inputs Q_{INF} (x-axis), X_{INF} (y-axis), and Q_{WAS} (z-axis) on the classification of observations as being either `FAILED` (red) or `NORMAL` (blue) at a representative SVI value of 150 mL/g.

Table 4.2: Summary of the number of observations classified as `FAILED` and `NORMAL` in the training and test sets at each SVI value considered.

SVI	**Data Set**	**Clarifier Failure** (`FAILED`)	**Effective Operation** (`NORMAL`)
100	Train	96	904
	Test	20	180
110	Train	108	892
	Test	17	183
120	Train	126	874
	Test	31	169
130	Train	140	860
	Test	33	167
140	Train	171	829
	Test	26	174
150	Train	168	832
	Test	38	162
160	Train	188	812
	Test	38	162
170	Train	189	811
	Test	38	162
180	Train	220	780
	Test	35	165
190	Train	243	757
	Test	55	145
200	Train	284	716
	Test	65	135

4.3 Defining an interaction term

The secondary clarification process is a complicated system in which many variables interact and are interdependent. In order to fully understand and capture the effect of these interactions on the system, *interaction terms* (sometimes referred to as *moderators*) can be constructed from the original set of measured input variables. In general, it is common for multivariate latent variable regression models to expand the **X**-space by augmenting the predictor matrix with engineered variables derived from the raw data [Dunn, 2010]. The process of transforming predictor variables is known as *feature engineering.* Feature engineering is of particular importance in predictive modeling applications as the nonlinear transformation of predictors can help linear models better capture the effects of predictor interactions on the response [Kuhn and Johnson, 2019]. Pairwise-interaction is the most common type of engineered variables considered in statistical modeling and is defined as the pairwise product or pairwise division of each main predictor variable [Kuhn and Johnson, 2019].

Interaction plots are a useful tool to evaluate whether the effect of a certain predictor variable on the system response is also dependent on the value of another predictor. Figure 4.4 presents the interaction plots showing the pairwise interaction effects between each possible pair of main predictor variables. In general, variables that do not interact are characterized by parallel lines on an interaction plot. The absence of any parallel lines in any of the interaction plots suggest that notable interaction effects exist between each predictor pair. Therefore, it can be said that the effect of any one input variable on X_{EFF} is also dependent on the other predictors. Moreover, we can see that Q_{WAS} is the only predictor variable with an inverse relationship to X_{EFF} in that low Q_{WAS} values are associated with high X_{EFF} values (and promptly clarifier failure) seemingly regardless of the corresponding Q_{INF} and X_{INF} levels. This contrasts the interaction effects seen in Q_{INF} and X_{INF}.

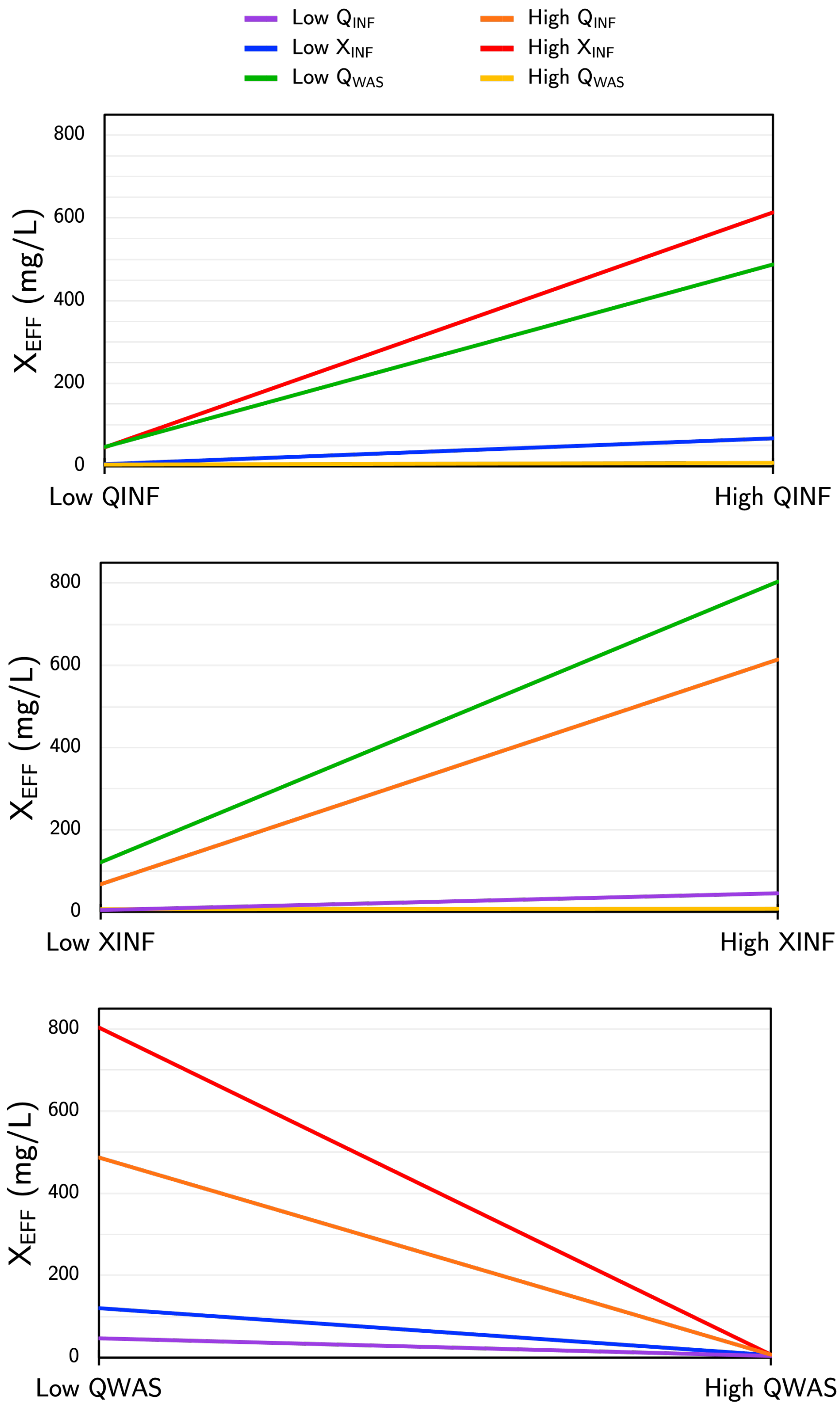


Figure 4.4: Interaction plots for effluent TSS (X_{EFF}) showing pairwise interactions between main predictors (Q_{INF}, X_{INF} and Q_{WAS}) for a representative SVI of 150 mL/g.

In order to effectively capture the interaction between all three of the main predictors, we define the interaction term (denoted R) as

$$R = \frac{Q_{INF} \cdot X_{INF}}{Q_{WAS}} \tag{4.3}$$

Physically, this variable represents the ratio of the TSS flow into the clarifier to the flow rate of sludge exiting the clarifier bottoms. A high value of R is therefore indicative of secondary clarifier failure, attributable to clarifier overload due to either excessive inlet solids or an extremely low rate of sludge removal. The goal of this new feature is thus to adequately describe the interaction between Q_{INF} and X_{INF} (i.e., the influent solids flow, represented by the product of Q_{INF} and X_{INF}) and also how these variables interact with Q_{WAS} (accounted for via the division by Q_{WAS}). Moreover, because the interaction plots show that X_{EFF} (and therefore the prevalence of clarifier failure) increases for low values of Q_{WAS}, we hope that in dividing by Q_{WAS}, a greater discrepancy between observations associated with effective clarification and those associated with clarification failure can be achieved.

The following section explores the effect of the interaction term R on classification performance and provides a comparison between the base-case PLS-DA model (i.e., without any interaction terms) and the PLS-DA model with interaction.

Remark 4.1. *It should be noted that effect of standard pairwise-product and pairwise-division interactions on PLS-DA performance was also evaluated and compared to the interaction variable R defined above. It was found that standard pairwise-interactions had a less significant impact on model performance and resulted in higher degrees of misclassification in the test set.*

4.4 Prediction of clarifier failure

4.4.1 Model calibration

PLS-DA model calibration with a seven-fold cross-validation procedure is performed in Aspen ProMV®, a multivariate analysis software developed by AspenTech, Inc. Two PLS-DA models are constructed for each SVI value—a base-case model which simply uses Q_{INF}, X_{INF}, and Q_{WAS} as input variables (henceforth referred to as *Model A*) and a second model which considers the previously described interaction term, R, as an additional input variable (*Model B*), thereby resulting in a total of 22 PLS-DA models. In general, $M - 1$ components (where M is the number of input variables) are retained in both *Model A* (2 components) and *Model B* (3 components). Recall that each PLS-DA model is trained on 1000 of the 1200 data points collected at each SVI value, with the remaining 200 data points reserved for testing.

The classification threshold (σ) for each PLS-DA model is determined such that both sensitivity (true positive rate) and specificity (true negative rate) are maximized, thereby also maximizing the balanced accuracy score and minimizing the false positive and false negative rates. In general, the value of σ at which the sensitivity and specificity scores are equal gives an acceptable threshold value. In the case of identifying secondary clarification failure, it is particularly important to minimize the number of false negatives, as classifying a `FAILED` observation as `NORMAL` has more detrimental repercussions than classifying a `NORMAL` observation as `FAILED`. As such, a moderate preference for maintaining a sensitivity score slightly greater than or equal to specificity was implemented to ensure a low number of false negative classifications.

A threshold plot can be used to visualize the effect of σ on the sensitivity and specificity scores and determine the optimal threshold value. Figure 4.5 presents example threshold plots for both PLS-DA *Models A* and *B* at a representative SVI value of 150 mL/g, showing the optimal threshold values which occur at approximately $\sigma = 0.191$ and $\sigma = 0.165$, respectively. It should be noted that the threshold value is determined during model calibration and is thus computed based on the predicted **y**-values obtained from applying the calibrated PLS-DA model to the same data used to train the model. Therefore, the threshold determined during model calibration is the same threshold used for discrimination when the PLS-DA model is applied to new data. For the sake of conciseness, the results presented in this section focus only on the PLS-DA models constructed at SVI = 150. Similar results are seen at each SVI value; therefore, the results given here are considered to be representative of the entire system. Additional classification results for the remaining SVI values can be found in the Supplementary Materials.

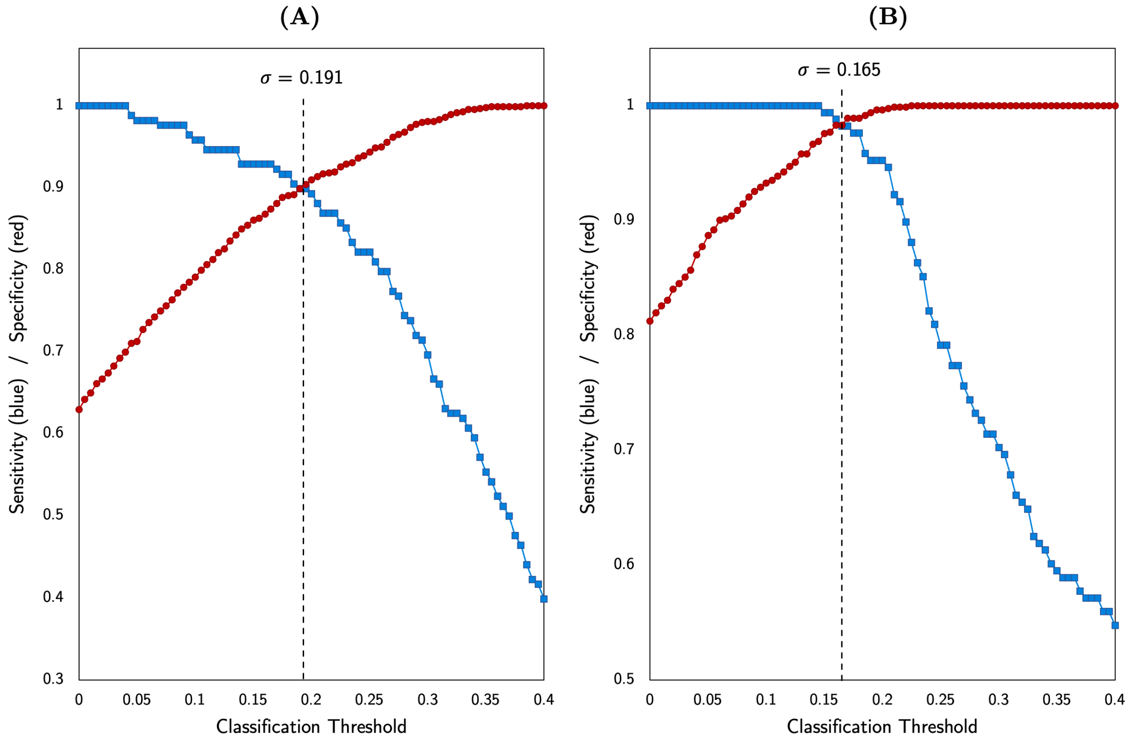


Figure 4.5: Threshold plots showing the effect of threshold on sensitivity (red) and specificity (blue) for (A) *Model A* and (B) *Model B* (both at SVI = 150 mL/g). The optimal threshold value (- - -) occurs at the point where sensitivity equals specificity.

Figures 4.6 and 4.7 respectively present the score and loadings plots for the first two components in both PLS-DA models. The first and second components (also referred to as LV1 and LV2) respectively explain roughly 36.7% and 30.4% of the total variation in the **X**-space for *Model A* and about 40.8% and 29.4% for *Model B*. Score plots present the scores from the first two model components as a scatter plot and are an excellent means of visualizing the separation between the classes. We can see from the score plots in Figure 4.6 that the separation between `FAILED` and `NORMAL` observations in the scores presents itself quite differently between *Model A* and *Model B*; `FAILED` observations generally have negative score values on LV1 in *Model A* but larger positive scores on LV1 in *Model B*. Moreover, the spread of the scores in *Model A* is very narrow in comparison to *Model B* and exhibit a greater degree of overlap between the `FAILED` and `NORMAL` classes. Although it is clear that the `FAILED` class concentrates at lower negative LV1 scores, there is no distinct line of separation between the two classes. In contrast, we can see that *Model B* shows a compact clustering of `NORMAL` observations at lower (mostly negative) score values on LV1 and a more loose, wide spread of `FAILED` observations at large positive score values on LV1.

Loading plots (Figure 4.7) closely relate to score plots, presenting the loadings (i.e. direction of projection) that define the first two components as a scatter plot to portray the significance of each predictor variable in discriminating between the classes [Dunn, 2010]. In general, **X**-space loadings (green) located close to the origin indicate that a predictor has a weak influence on the corresponding component, while loadings located near a **y**-space loading (orange) are positively correlated with the corresponding class label. The loading plots suggest that `NORMAL` observations are predominantly characterized by higher values of Q_{WAS} given its proximity to the `NORMAL` loading on both LV1 and LV2 in *Model A* and its proximity to `NORMAL` on LV1 in *Model B*. Similarly, we can see that R positively correlates with clarifier failure in *Model B* and that `FAILED` observations are thus characterized by high R values. These findings are

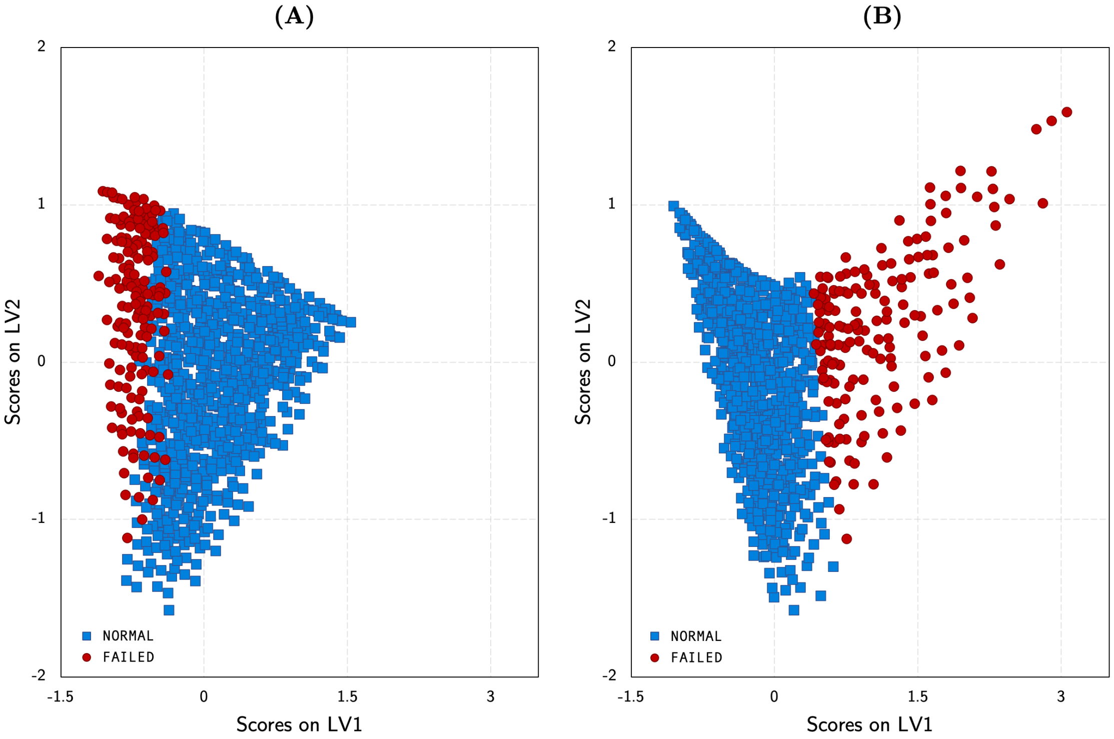


Figure 4.6: PLS-DA score plots showing the separation between `FAILED` (red) and `NORMAL` observations (blue) along the first two principal components for (A) *Model A* and (B) *Model B* (both at SVI = 150 mL/g).

further corroborated by the regression coefficients for the `FAILED` class.

Figure 4.8 presents the regression coefficient values as a bar plot sorted by descending contribution to the `FAILED` class. The negative coefficients for Q_{WAS} in both models confirm that Q_{WAS} is the main variable contributing to the classification of `NORMAL` observations. The positive regression coefficients for Q_{INF} and X_{INF} in *Model A* suggest contribution to `FAILED` classifications; however, the small magnitude of these coefficients indicates that the influence of these variables on classification is quite weak. The influence of Q_{INF} and X_{INF} on model predictions is even weaker in *Model B* with coefficient values both less than 0.1. Moreover, it is evident that R is extremely

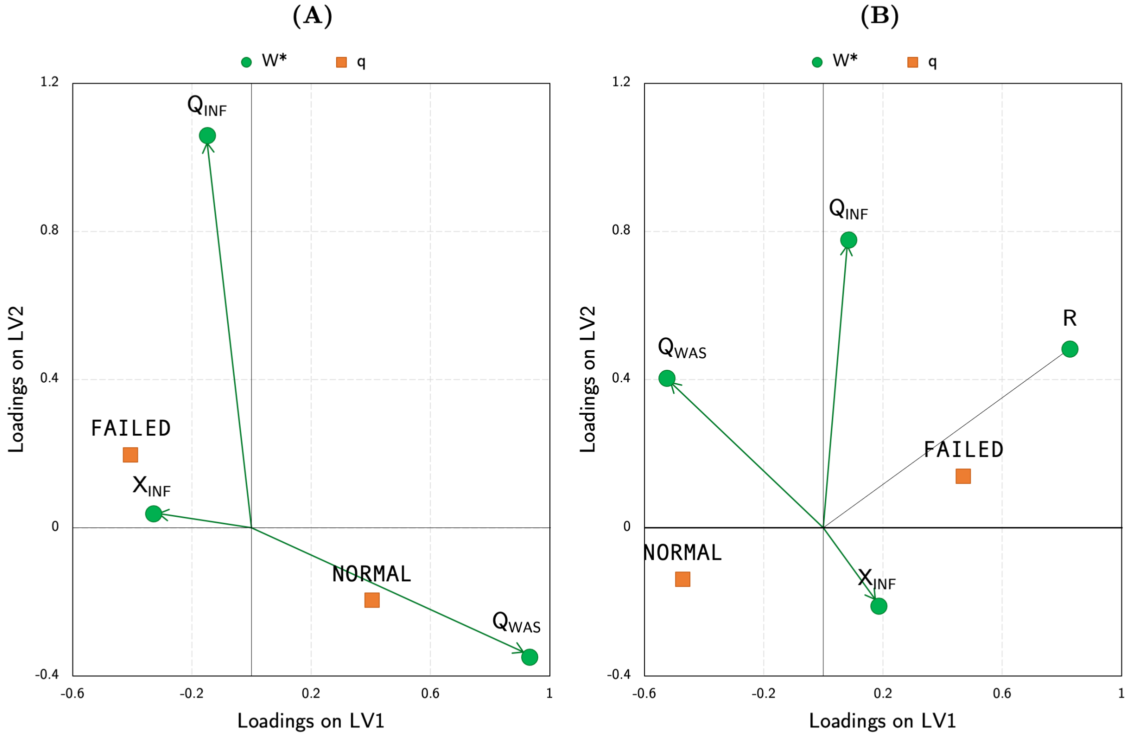


Figure 4.7: PLS-DA loading plots showing how inputs (green) relate to each other and to the response (orange) along the first two principal components for (A) *Model A* and (B) *Model B* (both at SVI = 150 mL/g).

important to classification in *Model B*, with a large positive coefficient signifying a significant contribution to identifying observations that belong to the `FAILED` class.

The discriminatory power of the predictor variables can be measured using *Variable Importance to Projection* (VIP). A VIP score is calculated for each individual variable in the **X**-space and essentially quantifies the contribution of each variable to the overall PLS-DA model. In general, higher VIP scores indicate higher importance, particularly when also associated with a large regression coefficient (absolute) value [Akarachantachote *et al.*, 2014]. *Variable Importance to Projection* is often utilized as a variable selection method when dealing with high-dimensional data, in which only predictors with a VIP greater than or equal to one are retained by the model (although Chong and Jun [2005] and Akarachantachote *et al.* [2014] have presented frameworks

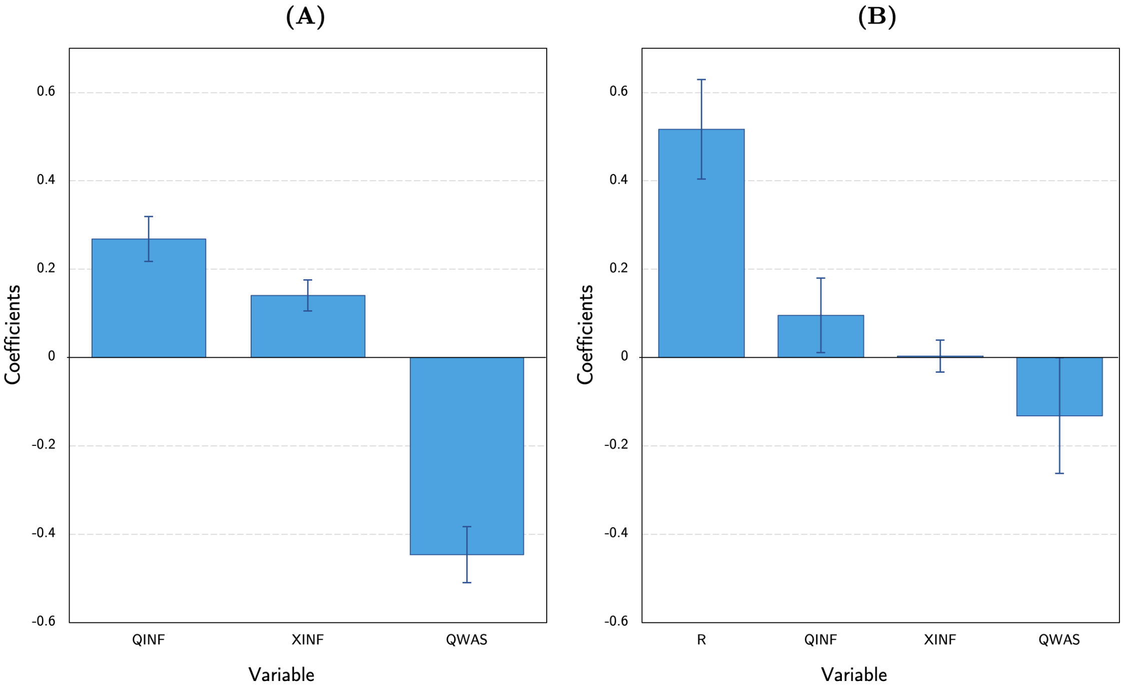


Figure 4.8: Regression coefficients for the `FAILED` class for (A) *Model A* and (B) *Model B* (both at SVI = 150 mL/g), at a 95% confidence level.

for selecting proper VIP threshold values as an alternative to the strict ‘greater than one’ rule); however, because the goal of current work is to demonstrate the validity and utility of PLS-DA modeling to identify conditions that result in clarifier failure, all input variables are retained regardless of VIP score. Instead, VIP is simply used to analyze the contribution of each variable and better understand how the interaction term R improves classification in *Model B*.

Figure 4.9 presents a bar plot of the VIP scores for each predictor variable in both PLS-DA models, sorted by descending importance. It can be seen that PLS-DA *Model A* heavily relies on Q_{WAS} for discrimination, with a VIP = 1.453. While Q_{WAS} is still considered a variable of importance in *Model B* (VIP = 1.048), the additional interaction term (R) has a much larger VIP score of 1.604 and indicates that R has the greatest contribution to discrimination in *Model B*. Recalling the score and loading plots presented in Figures 4.6 and 4.7, we know that Q_{WAS} largely contributes to the classification of `NORMAL` observations in both *Models A* and *B* (further confirmed by the negative regression coefficient values in Figure 4.8). In contrast, variables Q_{INF} and X_{INF} both contribute positively to the classification of `FAILED` observations in *Model A*; however, with VIP scores less than the conventional cut-off value, they would typically be considered unimportant to the model and discarded. The high VIP score associated with R in *Model B*, in combination with its large positive regression coefficient and its location in the loading plot relative to the `FAILED` loading, indicates that R contributes significantly to discrimination—particularly improving upon the classification of `FAILED` observations.

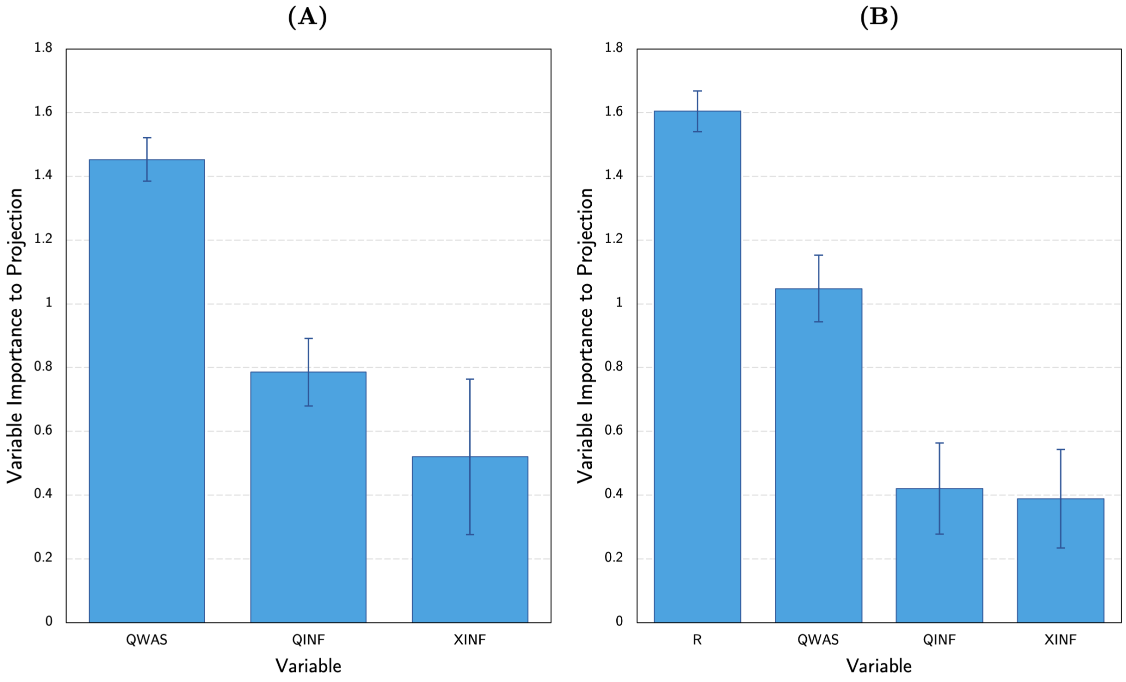


Figure 4.9: VIP bar plots sorted by descending importance for (A) *Model A* and (B) *Model B* (both at SVI = 150 mL/g), at a 95% confidence level.

The above analyses demonstrate the feasibility of the constructed PLS-DA models for predicting the occurrence of clarifier failure for new data.

4.4.2 PLS-DA predictions

Following calibration, each PLS-DA model is subsequently tested on the remaining 200 data points to evaluate the true predictive capability of the classification model. Table 4.3 presents the confusion matrices for *Models A* and *B*, summarizing the performance of both models based on predicted class membership for the test set.

Table 4.3: Confusion matrices reporting test set classification results for (A) *Model A* (base-case PLS-DA) and (B) *Model B* (PLS-DA with R), both at SVI = 150 mL/g.

(A)

		Actual Class		
		FAILED	**NORMAL**	
Predicted Class	**FAILED**	33	15	48
	NORMAL	5	147	152
		38	162	

(B)

		Actual Class		
		FAILED	**NORMAL**	
Predicted Class	**FAILED**	38	2	40
	NORMAL	0	160	160
		38	162	

Looking at the confusion matrices, we can see that *Model B* shows a significant improvement in the number of correct classifications compared to *Model A*, with the number of false positive (FP) and false negative (FN) classifications decreasing dramatically. Minimizing the number of false negatives is of particular importance in the application of PLS-DA to clarification failure classification—it is better raise alarms for potential secondary clarifier failure than to miss the failure entirely. As such, *Model B* is able to eliminate false negative classifications entirely while also decreasing the number of false positives by 80%. A similar reduction in the number of FP and FN classifications between PLS-DA *Models A* and *B* is also seen at the remaining SVI levels (see Supplementary Materials for additional classification results). Notably, *Model B* is associated with a 100% decrease in the number of false negatives for all but one SVI level—a 64% decrease in FN is seen in *Model B* at SVI = 200—and a decrease in the number of false positives anywhere from 70–96%.

Figure 4.10 complements the confusion matrices, presenting a visualization of the predicted **y**-values and how the determined classification threshold (recall the threshold plots in Figure 4.5) facilitates discrimination. The discrimination plots highlight the difference in how the two PLS-DA models generate predictions, showing that the predicted **y**-values from *Model A* span a much smaller range (-0.47 to 0.58) than *Model B* (-0.28 to 1.80), thereby resulting in a less defined separation between the two classes and ultimately increasing the potential for misclassification. In contrast, *Model B* is

associated with a much narrower spread in the predicted values for `NORMAL` observations and a wider spread in the predicted values for `FAILED` observations, similar to the trends observed in the score plots. The difference in the range and distribution of predicted values between the two models as illustrated by the discrimination plots suggests that the incorporation of the interaction term in *Model B* enables the PLS-DA model to produce more confident predictions that are able to better discriminate between the classes.

Classification parameters such as the *balanced accuracy* (BA), *balanced error rate* (BER), *sensitivity* (i.e., true positive rate), *specificity* (i.e., true negative rate) and the *Matthews correlation coefficient* (MCC) are derived from the test set and used to quantify PLS-DA model performance. Table 4.4 reports these metrics for both the training and test set classification results for the PLS-DA model at SVI = 150. It can be seen that model performance is comparable for both model fitting and validation; thus, we can reasonably confirm that both PLS-DA *Models A* and *B* are stable and

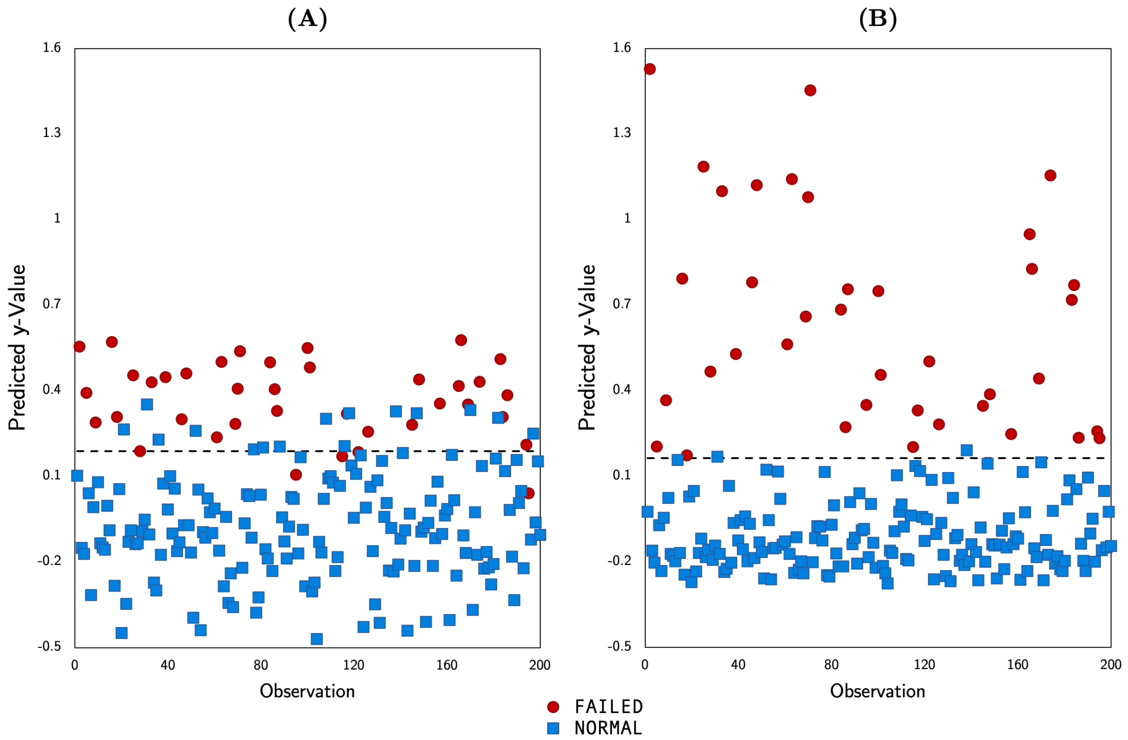


Figure 4.10: PLS-DA discrimination plots showing the predicted **y**-values and how the classification threshold (- - -) discriminates between `FAILED` (red) and `NORMAL` (blue) observations for (A) *Model A* and (B) *Model B* (both at SVI = 150 mL/g).

reliable for use on new data. A consistent improvement is seen in *Model B* across all performance metrics. Note that values for the *false positive* and *false negative rates* are not presented here as these metrics can be easily calculated from *sensitivity* and *specificity.*

The following tables present a summary of the *balanced accuracy* (Table 4.5), *sensitivity* (Table 4.6), *specificity* (Table 4.7) and *Matthews correlation coefficient* (Table 4.8) values obtained from test set classification results for both PLS-DA *Models A* and *B* at all considered SVI values. The change in each metric due to *Model B*'s interaction term is also given, as well as the average metric values over all SVIs. A similar summary of performance metrics obtained during model calibration can be found in Tables S2 to S5 in the Supplementary Materials. A consistent improvement in the performance of *Model B* over *Model A* is observed for all classification metrics at all SVI values, indicating that the defined interaction term is both significant and valuable in facilitating effective discrimination by PLS-DA.

Table 4.4: PLS-DA model performance metrics for fitting and test set classification results for both *Model A* and *Model B* at a representative SVI value of 150 mL/g.

	BA	**BER**	**TPR**	**TNR**	**MCC**
Model A					
Fitting	0.898	0.101	0.899	0.899	0.703
Test Set	0.888	0.112	0.868	0.907	0.713
Model B					
Fitting	0.983	0.017	0.982	0.983	0.941
Test Set	0.994	0.006	1.000	0.988	0.969

Remark 4.2. *It was observed that discrimination based solely on the value of the defined interaction variable R also generates acceptable classification results. The classification threshold is determined in a similar manner to the PLS-DA models (i.e., find the value of R that maximizes the number of correct classification) with R values greater than the threshold indicating clarifier failure and R values below the threshold represent effective clarifier operation. In this case, the computed sensitivity scores are comparable to the sensitivities for PLS-DA Model B; however, an general decrease in specificity is seen, with an average specificity of 0.951 across all SVIs when predicting clarifier failure using just R compared to 0.983 for PLS-DA Model B.*

Table 4.5: Comparison of *balanced accuracy* (BA) scores for *Model A* (PLS-DA *without R*) and *Model B* (PLS-DA *with R*) test set class membership predictions at each SVI value considered.

SVI	**Balanced Accuracy (BA)**		**Change in BA**
	Model A	*Model B*	
100	0.894	0.983	0.089
110	0.892	0.992	0.100
120	0.926	0.994	0.068
130	0.937	0.997	0.060
140	0.868	0.986	0.117
150	0.888	0.994	0.106
160	0.867	0.997	0.130
170	0.893	0.997	0.104
180	0.880	0.982	0.102
190	0.885	0.990	0.105
200	0.886	0.954	0.069
AVERAGE	0.890	0.988	0.098

Table 4.6: Comparison of *sensitivity* scores for *Model A* and *Model B* test set class membership predictions at each SVI value considered.

SVI	**Sensitivity**		**Change in TPR**
	Model A	*Model B*	
100	0.900	1.000	0.100
110	0.882	1.000	0.118
120	0.935	1.000	0.065
130	0.970	1.000	0.030
140	0.846	1.000	0.154
150	0.868	1.000	0.132
160	0.895	1.000	0.105
170	0.842	1.000	0.158
180	0.857	1.000	0.143
190	0.873	1.000	0.127
200	0.831	0.938	0.108
AVERAGE	0.882	0.994	0.113

Table 4.7: Comparison of *specificity* scores for *Model A* and *Model B* test set class membership predictions at each SVI value considered.

SVI	**Specificity**		**Change in TNR**
	Model A	*Model B*	
100	0.889	0.967	0.078
110	0.902	0.984	0.082
120	0.917	0.988	0.071
130	0.904	0.994	0.090
140	0.891	0.971	0.080
150	0.907	0.988	0.080
160	0.840	0.994	0.154
170	0.944	0.994	0.049
180	0.903	0.982	0.079
190	0.897	0.979	0.083
200	0.881	0.970	0.089
AVERAGE	0.898	0.983	0.085

Table 4.8: Comparison of *Matthews correlation coefficient* (MCC) values for *Model A* (PLS-DA *without R*) and *Model B* (PLS-DA *with R*) test set class membership predictions at each SVI value considered.

SVI	**Matthews Correlation Coefficient**		**Change in MCC**
	Model A	*Model B*	
100	0.603	0.862	0.259
110	0.589	0.914	0.325
120	0.751	0.963	0.212
130	0.760	0.982	0.223
140	0.614	0.903	0.289
150	0.713	0.969	0.256
160	0.629	0.984	0.355
170	0.764	0.984	0.220
180	0.686	0.907	0.221
190	0.739	0.964	0.224
200	0.699	0.909	0.209
AVERAGE	0.686	0.940	0.254

4.5 Summary of contribution

The utility of partial least squares discriminant analysis for identifying static conditions that result in clarification failure was established in this chapter. In particular, a novel third-order interaction term that aims to capture the relation between the three input variables was engineered and incorporated as an additional predictor in the **X**-space and the impact of the interaction term on PLS-DA prediction and discrimination was explored. Discrimination results highlight the efficacy of the engineered interaction term in improving PLS-DA predictions and reducing the prevalence of misclassification, as demonstrated by a consistent improvement in all classification metrics and PLS-DA model performance parameters considered in comparison to the base-case PLS-DA model. PLS-DA with interaction shows potential as a useful process monitoring tool to assist wastewater treatment plant operators in forecasting clarification failure and responding accordingly.

Chapter 5

Conclusions and recommendations

5.1 Conclusions

This thesis investigates the utility of both dynamic and static data-driven modeling techniques for estimating a relevant model of the secondary clarification process in a wastewater treatment plant.

The first part of the thesis explored the dynamic modeling piece—covering the application of subspace model identification to a secondary clarification unit. The objective of this work was to investigate the suitability of subspace identification for identifying an appropriate LTI state-space model of the clarification process. To this end, a subspace model was trained and subsequently validated on simulation data obtained via dynamic GPS-X simulations and subsequently corrupted with a random white noise signal. Validation results were successful in demonstrating that a linear deterministic subspace methodology was able to approximate the nonlinear process behaviour reasonably well can therefore be accepted as a suitable model for the current system and range of operation considered. It should be noted, however, that the continued use of a linear deterministic model may need to be re-evaluated as process complexity is increased and additional dynamics are considered in future work. Results showed that the subspace model was able to effectively predict the dynamic behaviour of the two output variables—the effluent and waste sludge TSS concentrations—with minimal prediction error. Moreover, the subspace model was able to produce accurate multi-step predictions for multiple candidate input profiles throughout the validation dataset and thus established its candidacy for use in a model predictive control framework.

The second part of the thesis focused on the static modeling piece—exploring the application of discriminative modeling to the secondary clarification unit for the purpose of failure prediction. A framework for forecasting the occurrence of sludge bulking (and therefore clarification failure) based on PLS-DA and an engineered interaction variable was described and subsequently utilized to predict and discriminate process conditions associated with clarifier failure from those associated with effective clarification. Results showed that the interaction term—which nonlinearly relates all three input variables—sufficiently captured the interaction between the predictors and consistently reduced the rate of misclassification. A classification accuracy of 98.8% was averaged across all PLS-DA models that augmented the predictor matrix with the interaction term—an almost 10% improvement over the base-case PLS-DA models.

5.2 Topics for future research

The secondary clarification system considered in this work is of course a substantial simplification of a very complex and notably nonlinear process. In the current work, simplification was necessary to establish the utility and validity of the considered data-driven modeling techniques at a base-case level and consequently motivate their application to a more representative secondary clarifier layout and even to the wastewater treatment process in general. As such, future work should focus on increasing process complexity via the incorporation of additional key input and output variables that affect clarification performance, eventually using measured data (in place of simulated data) to model real-world secondary clarification processes. Additional variables that could be considered include: temperature, percentage of sludge recycle to secondary clarifier (i.e., RAS), pH, dissolved oxygen (DO) concentrations, nitrogen and phosphorus concentrations, biological oxygen demand (BOD), and total organic carbon (TOC), among many others.

Subspace model of secondary clarifier

Future directions for this work include: (*i*) increasing process complexity, (*ii*) incorporating the subspace model in a model predictive control framework, and (*iii*) utilizing hybrid modeling strategies.

Process complexity

As previously mentioned, additional input and/or output variables should be considered to better represent real-life process conditions. Process complexity can also be increased via the incorporation of additional wastewater treatment units. If data is available from all of the various treatment units within the WWTP, a single model can be built which relates a set of raw influent variables directly to key effluent quality variables, thus describing the entire wastewater treatment process as a whole. Alternatively, it is also possible to construct multiple separate models for individual treatment units that can be subsequently connected such that the output from one process unit is used as the input to the next successive unit in the WWTP layout.

Model predictive control

The application of the identified subspace model for feedback control should be explored in future work. Current work establishes the ability of the subspace model to produce quality multi-step predictions for multiple candidate input profiles, thereby highlighting its viability for use in a model predictive control scheme. MPC implementations for the secondary clarification process should focus on controlling the secondary clarifier effluent quality such that effluent quality standards are maintained at a minimum cost.

Hybrid modeling

Future work should explore the use of hybrid modeling strategies to account for existing first-principles models and ultimately improve upon subspace model predictions. Hybrid models offer many advantages—mainly the ability to account for already available process knowledge via well-known and established first-principles relations while also still utilizing data-driven techniques to capture the remaining dynamics present in the measured data. Moreover, subspace identification—an intrinsically data-driven approach—may lose physical insight of the process and, as a result, identify a model that does not necessarily respect physical constraints inherent to the process. Hybrid modeling strategies can resolve this limitation, ensuring that physical process constraints are considered.

Predicting and classifying secondary clarification failure

Future directions for this work include: (*i*) increasing process complexity, (*ii*) employing alternative discrimination methods and (*iii*) fault classification and diagnosis.

Process complexity

As previously mentioned, additional input and/or output variables as well as additional treatment units can be considered in order to better represent real-life process conditions. The incorporation of more predictor variables can enable a larger portion of variation in both the $\mathbf{X}$– and $\mathbf{y}$–spaces to be explained by the PLS-DA model; however, additional predictors will necessitate feature selection to retain only variables that are useful in predicting the response and thus classifying conditions of clarification failure.

Alternative discrimination methods

PLS-DA with the same number of principal components as predictor variables shows only a slight improvement in model performance with regard to prediction and discrimination as a result of the associated increase in the amount of **X**-space variation that is captured by the PLS-DA model with the additional component. This therefore prompts the exploration of other discrimination methods that could better suit the dataset such as linear discriminant analysis (LDA), k-nearest neighbours (KNN), logistic regression and support vector machine (SVM).

Fault diagnosis

It is possible that the problem presented in this work be reformulated to instead focus on fault-specific classification. Faults arise when a process variable deviates from its acceptable operating range, potentially leading to unsatisfactory performance or even process failure. To this end, we can construct a multi-class PLS-DA model aimed at diagnosing and classifying common faults associated with clarification failure, consequently also identifying important fault-specific features and patterns in the data. The preliminary results presented in this thesis suggest that PLS-DA—particularly PLS-DA with interaction—shows promise as a process monitoring tool to advise decision-making in WWTPs. As such, the expansion of this work to target fault classification can help operators respond appropriately upon fault detection.

References

Aguado, D., Ferrer, A., Seco, A., and Ferrer, J. (2006). Comparison of different predictive models for nutrient estimation in a sequencing batch reactor for wastewater treatment. *Chemometrics and Intelligent Laboratory Systems*, **84**(1-2), 75–81.

Akarachantachote, N., Chadcham, S., and Saithanu, K. (2014). Cutoff Threshold of Variable Importance in Projection for Variable Selection. *International Journal of Pure and Applied Mathematics*, **94**(3), 307–322.

Akpor, O. B. and Muchie, M. (2011). Environmental and public health implications of wastewater quality. *African Journal of Biotechnology*, **10**(13), 2379–2387.

Anderson, J. S., McAvoy, T. J., and Hao, O. J. (2000). Use of Hybrid Models in Wastewater Systems. *Industrial and Engineering Chemistry Research*, **39**(6), 1694–1704.

Ballabio, D. and Consonni, V. (2013). Classification tools in chemistry. Part 1: linear models. PLS-DA. *Analytical Methods*, **5**(6), 3790–3798.

Basseville, M., Abdelghani, M., and Benveniste, A. (2000). Subspace-based fault detection algorithms for vibration monitoring. *Automatica*, **36**(1), 101–109.

Bastogne, T., Noura, H., Richard, A., and Hittinger, J.-M. (1997). Application of subspace methods to the identification of a winding process. In *1997 European Control Conference (ECC)*, pp. 2168–2173. IEEE.

Bertolini, M., Mezzogori, D., Neroni, M., and Zammori, F. (2021). Machine Learning for industrial applications: A comprehensive literature review. *Expert Systems with Applications*, **175**, 114820.

Bilotta, G. S. and Brazier, R. E. (2008). Understanding the influence of suspended solids on water quality and aquatic biota. *Water Research*, **42**(12), 2849–2861.

Brodersen, K. H., Ong, C. S., Stephan, K. E., and Buhmann, J. M. (2010). The Balanced Accuracy and Its Posterior Distribution. In *2010 20th International Conference on Pattern Recognition*, pp. 3121–3124, Istanbul, Turkey. IEEE.

Chevallier, S., Bertrand, D., Kohler, A., and Courcoux, P. (2006). Application of PLS-DA in multivariate image analysis. *Journal of Chemometrics*, **20**(5), 221–229.

Chicco, D. (2017). Ten quick tips for machine learning in computational biology. *BioData Mining*, **10**(35).

Chicco, D. and Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, **21**(6).

Choi, S. W. and Lee, I.-B. (2005). Multiblock PLS-based localized process diagnosis. *Journal of Process Control*, **15**(3), 295–306.

Chong, I.-G. and Jun, C.-H. (2005). Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems*, **78**(1-2), 103–112.

Corbett, B. and Mhaskar, P. (2016). Subspace Identification for Data-Driven Modeling and Quality Control of Batch Processes. *AIChE Journal*, **62**(5), 1581–1601.

Corominas, L., Garrido-Baserba, M., Villez, K., Olsson, G., Cortés, U., and Poch, M. (2018). Transforming data into knowledge for improved wastewater treatment operation: A critical review of techniques. *Environmental Modelling and Software*, **106**, 89–103.

de Moor, B., van Overschee, P., and Favoreel, W. (1999). Algorithms for Subspace State-Space System Identification: An Overview. In *Applied and Computational Control, Signals, and Circuits*, Vol. 1, chapter 6, pp. 247–311. Birkhäuser, Boston, MA, USA.

Dick, R. I. and Vesilind, A. (1969). The Sludge Volume Index: What Is It? *Water Pollution Control Federation*, **41**(7), 1285–1291.

Ding, S. X., Zhang, P., Naik, A., Ding, E. L., and Huang, B. (2009). Subspace method aided data-driven design of fault detection and isolation systems. *Journal of Process Control*, **19**(9), 1496–1510.

Dong, Y. and Qin, S. J. (2015). Dynamic-Inner Partial Least Squared for Dynamic Data Modeling. *IFAC-PapersOnLine*, **48**(8), 117–122. 9th IFAC Symposium on Advanced Control of Chemical Processes ADCHEM 2015.

Dunn, K. (2010). Latent Variable Modelling. In *Process Improvement Using Data*, chapter 6, pp. 315–402. learnche.org, 6ec14c edition. *Available at* https://learnche.org/pid/. Last updated on November 13, 2021.

Dürrenmatt, D. J. and Gujer, W. (2012). Data-driven modeling approaches to support wastewater treatment plant operation. *Environmental Modelling and Software*, **30**, 47–56.

Environment Canada (2012). Wastewater Systems Effluent Regulations. SOR/2012-139.

Esbensen, K. H. and Geladi, P. (2010). Principles of Proper Validation: use and abuse of re-sampling for validation. *Journal of Chemometrics*, **24**(3-4), 168–187.

Favoreel, W., de Moor, B., and van Overschee, P. (2000). Subspace state space system identification for industrial processes. *Journal of Process Control*, **10**(2), 149–155.

Geladi, P. and Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, **185**, 1–17.

Ghosh, D., Hermonat, E., Mhaskar, P., Snowling, S., and Goel, R. (2019). Hybrid Modeling Approach Integrating First-Principles Models with Subspace Identification. *Industrial and Engineering Chemistry Research*, **58**(30), 13533–13543.

Griborio, A. (2004). *Secondary Clarifier Modeling: A Multi-Process Approach*. PhD Thesis, University of New Orleans, New Orleans, Louisiana, USA.

Hauduc, H., Gillot, S., Rieger, L., Ohtsuke, T., Shaw, A., Takács, I., and Winkler, S. (2009). Activated sludge modelling in practice: an international survey. *Water Science and Technology*, **60**(8), 1943–1951.

Henze, M. (2002). Activated Sludge Treatment Plants. In *Wastewater treatment: Biological and chemical processes*, chapter 4, pp. 113–142. Springer-Verlag, 2nd edition.

Henze, M., Gujer, W., Mino, T., and van Loosedrecht, M. (2006). *Activated Sludge Models ASM1, ASM2, ASM2d and ASM3*, Vol. 5. IWA Publishing.

Huang, B., Ding, S. X., and Qin, S. J. (2005). Closed-loop subspace identification: an orthogonal projection approach. *Journal of Process Control*, **15**(1), 53–66.

Huberty, C. J. (1975). Discriminant Analysis. *Review of Educational Research*, **45**(4), 543–598.

Hydromantis ESS, Inc. (2019). *GPS-X Technical Reference v8.0.* Hydromantis Environmental Software Solutions, Inc. *Available at* https://www.hydromantis.com/help/GPS-X/docs/8.0/Technical/.

Jackson, J. E. (2005). *A User's Guide to Principal Components.* Wiley Series in Probability and Statistics. John Wiley & Sons.

Jeppsson, U. (1996). *Modelling Aspects of Wastewater Treatment Processes.* PhD Thesis, Lund Institute of Technology, Lund, Sweden.

Jeppsson, U. and Diehl, S. (1996). An evaluation of a dynamic model of the secondary clarifier. *Water Science and Technology*, **34**(5), 19–26.

Ji, Z., McCorquodale, J. A., Zhou, S., and Vitasovic, Z. (1996). A Dynamic Solids Inventory Model for Activated Sludge Systems. *Water Environment Research*, **68**(3), 329–337.

Kresta, J. V. (1992). *The applications of partial least squares to problems in chemical engineering.* PhD Thesis, McMaster University, Hamilton, ON, Canada.

Kresta, J. V., MacGregor, J. F., and Marlin, T. E. (1991). Multivariate statistical monitoring of process operating performance. *Canadian Journal of Chemical Engineering*, **69**(1), 35–47.

Kuhn, M. and Johnson, K. (2019). *Feature Engineering and Selection: A Practical Approach for Predictive Models.* Chapman and Hall/CRC Press, New York, NY, 1st edition.

Lakshminarayanan, S., Shah, S. L., and Nandakumar, K. (1997). Modeling and control of multivariable processes: Dynamic PLS approach. *AIChE Journal*, **43**(9), 2307–2322.

Langergraber, G., Fleischmann, N., and Hofstädter, F. (2003). A multivariate calibration procedure for UV/VIS spectrometric quantification of organic matter and nitrate in wastewater. *Water Science and Technology*, **47**(2), 63–71.

Larimore, W. E. (1990). Canonical variate analysis in identification, filtering, and adaptive control. In *29th IEEE Conference on Decision and Control*, Vol. 2, pp. 596–604. IEEE.

Lee, L. C., Liong, C.-Y., and Jemain, A. A. (2018). Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: A review of contemporary practice strategies and knowledge gaps. *The Analyst*, **143**(15), 3526–3539.

Lindberg, C.-F. (1997). *Control and Estimation Strategies Applied to the Activated Sludge Process.* PhD Thesis, Uppsala University, Uppsala, Sweden.

Ljung, L. and McKevey, T. (1996). Subspace Identification from Closed Loop Data. *Signal Processing*, **52**(2), 209–215.

Lourenço, N. D., Menezes, J. C., Pinheiro, H. M., and Diniz, D. (2008). Development of PLS calibration models from UV-Vis spectro for TOC estimation at the outlet of a fuel park wastewater treatment plant. *Environmental Technology*, **29**(8), 891–898.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, **405**(2), 442–451.

Medora Corporation (2016). Predicting Mixing Requirements Using the Sludge Volume Index (SVI). Informational Bulletin ID1700.

Meidanshahi, V., Corbett, B., Adams II, T. A., and Mhaskar, P. (2017). Subspace model identification and model predictive control based cost analysis of a semicontinuous distillation process. *Computers and Chemical Engineering*, **103**, 39–57.

MetCalf & Eddy, Inc., Tchobanoglous, G., Tsuchihashi, R., Burton, F. L., and Stensel, H. D. (2014). *Wastewater Engineering: Treatment and Resource Recovery*. McGraw-Hill Higher Education, New York, NY, 5th edition.

Misra, P. and Nikolaou, M. (2003). Input design for model order determination in subspace identification. *AIChE Journal*, **49**(8), 2124–2132.

Moonen, M., de Moor, B., Vandenberghe, L., and Vandewalle, J. (1989). On- and off-line identification of linear state-space models. *International Journal of Control*, **49**(1), 219–232.

Morrison, D. G. (1969). On the Interpretation of Discriminant Analysis. *Journal of Marketing Research*, **6**(2), 156–163.

Newhart, K. B., Holloway, R. W., Hering, A. S., and Cath, T. Y. (2019). Data-driven performance analyses of wastewater treatment plants: A review. *Water Research*, **157**, 498–513.

Perez-Lopez, C., Ginebreda, A., Carrascal, M., Barcelò, D., Abia, J., and Tauler, R. (2021). Non-target protein analysis of samples from wastewater treatment plants using the regions of interest-multivariate curve resolution (ROIMCR) chemometrics method. *Journal of Environmental Chemical Engineering*, **9**(4), 105752.

Platikanov, S., Rodriguez-Mozaz, S., Huerta, B., Barceló, D., Cros, J., Batle, M., Poch, G., and Tauler, R. (2014). Chemometrics quality assessment of wastewater treatment plant effluents using physicochemical parameters and UV absorption measurements. *Journal of Environmental Management*, **140**, 33–44.

Qin, S. J. (1993). Partial least squares regression for recursive system identification. In *Proceedings of the 32nd IEEE Conference on Decision and Control*, Vol. 3, pp. 2617–2622. IEEE.

Qin, S. J. (2006). An overview of subspace identification. *Computers & Chemical Engineering*, **30**(10), 1502–1513.

Rosen, C. and Olsson, G. (1998). Disturbance detection in wastewater treatment plants. *Water Science and Technology*, **37**(12), 197–205.

Ruiz-Perez, D., Guan, H., Madhivanan, P., Mathee, K., and Narasimhan, G. (2018). So you think you can PLS-DA? In *2018 IEEE 8th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS)*, pp. 1–1, Los Alamitos, CA, USA. IEEE Computer Society.

Shahnazari, H., Mhaskar, P., House, J. M., and Salsbury, T. I. (2018). Heating, ventilation and air conditioning systems: Fault detection and isolation and safe parking. *Computers and Chemical Engineering*, **108**, 139–151.

Singh, K. P., Malik, A., Mhan, D., Sinha, S., and Singh, V. K. (2005). Chemometric data analysis of pollutants is wastewater—a case study. *Analytica Chimica Acta*, **532**(1), 15–25.

Sorensen, D. L., McCarthy, M. M., Middlebrooks, J. E., and Porcella, D. B. (1977). Suspended and Dissolved Solids Effects on Freshwater Biota: A Review. Technical report, U.S. Environmental Protection Agency.

Sotomayor, O. A. Z., Park, S. W., and Garcia, C. (2003). Multivariable identification of an activated sludge process with subspace-based algorithms. *Control Engineering Practice*, **11**(8), 961–969.

Sánchez, A. (2004). *Data-Driven Control Design of Wastewater Treatment Systems.* PhD Thesis, University of Strathclyde, Glasgow, UK.

Sánchez, A. and Katebi, M. R. (2003). Predictive control of dissolved oxygen in an activated sludge wastewater treatment plant. In *2003 European Control Conference (ECC)*, pp. 2424–2429, Cambridge, UK. IEEE.

Takács, I., Patry, G. G., and Nolasco, D. (1991). A dynamic model of the clarification-thickening process. *Water Research*, **25**(10), 1263–1271.

Tan, Y., Shi, L., Tong, W., Gene Hwang, G. T., and Wang, C. (2004). Multi-class tumor classification by discriminant partial least squares using microarray gene expression data and assessment of classification models. *Computational Biology and Chemistry*, **28**(3), 235–243.

Tharwat, A. (2020). Classification assessment methods. *Applied Computing and Informatics*, **17**(1), 168–192.

Torfs, E., Nopens, I., Winkler, M. K. H., Vanrolleghem, P. A., Balemans, S., and Smets, I. Y. (2016). Settling Tests. In *Experimental Methods in Wastewater Treatment*, chapter 6, pp. 235–262. IWA Publishing.

U.S. EPA (2004). Primer for Municipal Wastewater Treatment Systems. Technical report, United States Environmental Protection Agency.

van Overschee, P. and de Moor, B. (1994). N4SID: Subspace algorithms for the identification of combined deterministic-stochastic systems. *Automatica*, **30**(1), 74–93.

van Overschee, P. and de Moor, B. (1995). A Unifying Theorem for Three Subspace System Identification Algorithms. *Automatica*, **31**(12), 1853–1864.

van Overschee, P. and de Moor, B. (1996). *Subspace Identification for Linear Systems: Theory, Implementation, Applications.* Kluwer Academic Publishers, Boston, MA, USA.

Verhaegen, M. and Dewilde, P. (1992). Subspace model identification Part 1. The output-error state-space model identification class of algorithms. *International Journal of Control*, **56**(5), 1187–1210.

Volcke, E. I. P., Solon, K., Comeau, Y., and Henze, M. (2020). Wastewater Characteristics. In *Biological Wastewater Treatment: Principles, Modeling and Design*, chapter 3, pp. 77–103. IWA Publishing, 2nd edition.

Von Sperling, M. (2007a). *Activated Sludge and Aerobic Biofilm Reactors*, Vol. 5 of *Biological Wastewater Treatment Series*. IWA Publishing.

Von Sperling, M. (2007b). *Wastewater Characteristics, Treatment and Disposal*, Vol. 1 of *Biological Wastewater Treatment Series*. IWA Publishing.

Wang, X., Kvaal, K., and Ratnaweera, H. (2017). Characterization of influent wastewater with periodic variation and snow melting effect in cold climate area. *Computers and Chemical Engineering*, **106**, 202–211.

Water Environment Federation (2005). *Clarifier Design: WEF Manual of Practice No. FD-8.* McGraw-Hill Education, 2nd edition.

Wei, X., Verhaegen, M., and van Engelen, T. (2010). Sensor fault detection and isolation for wind turbines based on subspace identification and Kalman filter techniques. *International Journal of Adaptive Control and Signal Processing*, **24**(8), 687–707.

Wold, H. (1975). Path Models with Latent Variables: The NIPALS Approach. In *Quantitative Sociology*, pp. 307–357. Academic Press.

Wold, S. (1978). Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models. *Technometrics*, **20**(4), 397–405.

Wold, S., Sjöström, M., and Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, **58**(2), 109–130.

Woo, S. H., Jeon, C. O., Yun, Y.-S., Choi, H., Lee, C.-S., and Lee, D. S. (2009). On-line estimation of key process variables based on kernel partial least squares in an industrial cokes wastewater treatment plant. *Journal of Hazardous Materials*, **161**(1), 538–544.

Worley, B. and Powers, R. (2013). Multivariate Analysis in Metabolomics. *Current Metabolomics*, **1**(1), 92–107.

Yotova, G., Lazarova, S., Kudłak, B., Zlateva, B., Mihaylova, V., Wieczerzak, M., Venelinov, T., and Tsakovski, S. (2019). Assessment of the Bulgarian Wastewater Treatment Plants' Impact on the Receiving Water Bodies. *Molecules*, **24**(12), 2274.

Supplementary materials

Additional PLS-DA results

Table S1: Summary of classification threshold values for *Model A* (base-case PLS-DA) and *Model B* (PLS-DA with R) at each SVI value considered.

SVI	**Classification Threshold (σ)**	
	Model A	*Model B*
100	0.147	0.211
110	0.155	0.201
120	0.174	0.171
130	0.175	0.168
140	0.192	0.167
150	0.191	0.165
160	0.185	0.145
170	0.181	0.152
180	0.186	0.120
190	0.182	0.127
200	0.173	0.118

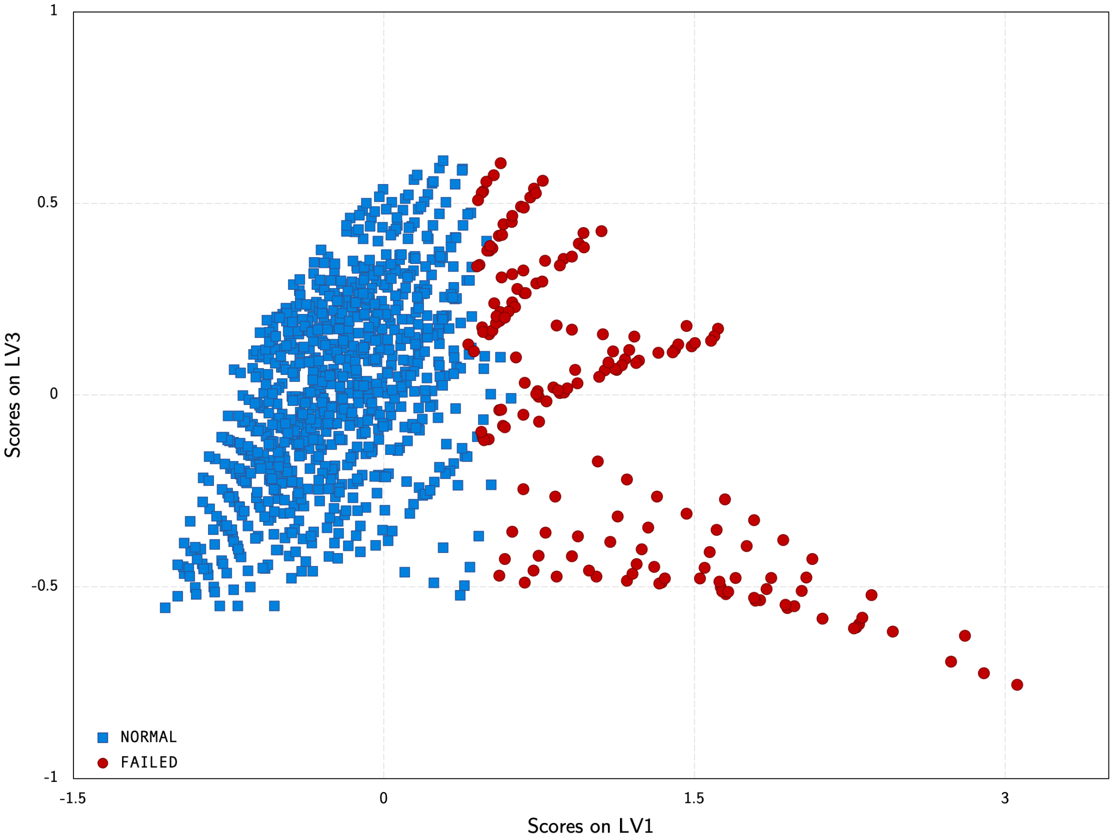


Figure S1: Score plot showing the separation between `FAILED` (red) and `NORMAL` observations (blue) along the first and third principal components of PLS-DA *Model B* (at SVI = 150 mL/g).

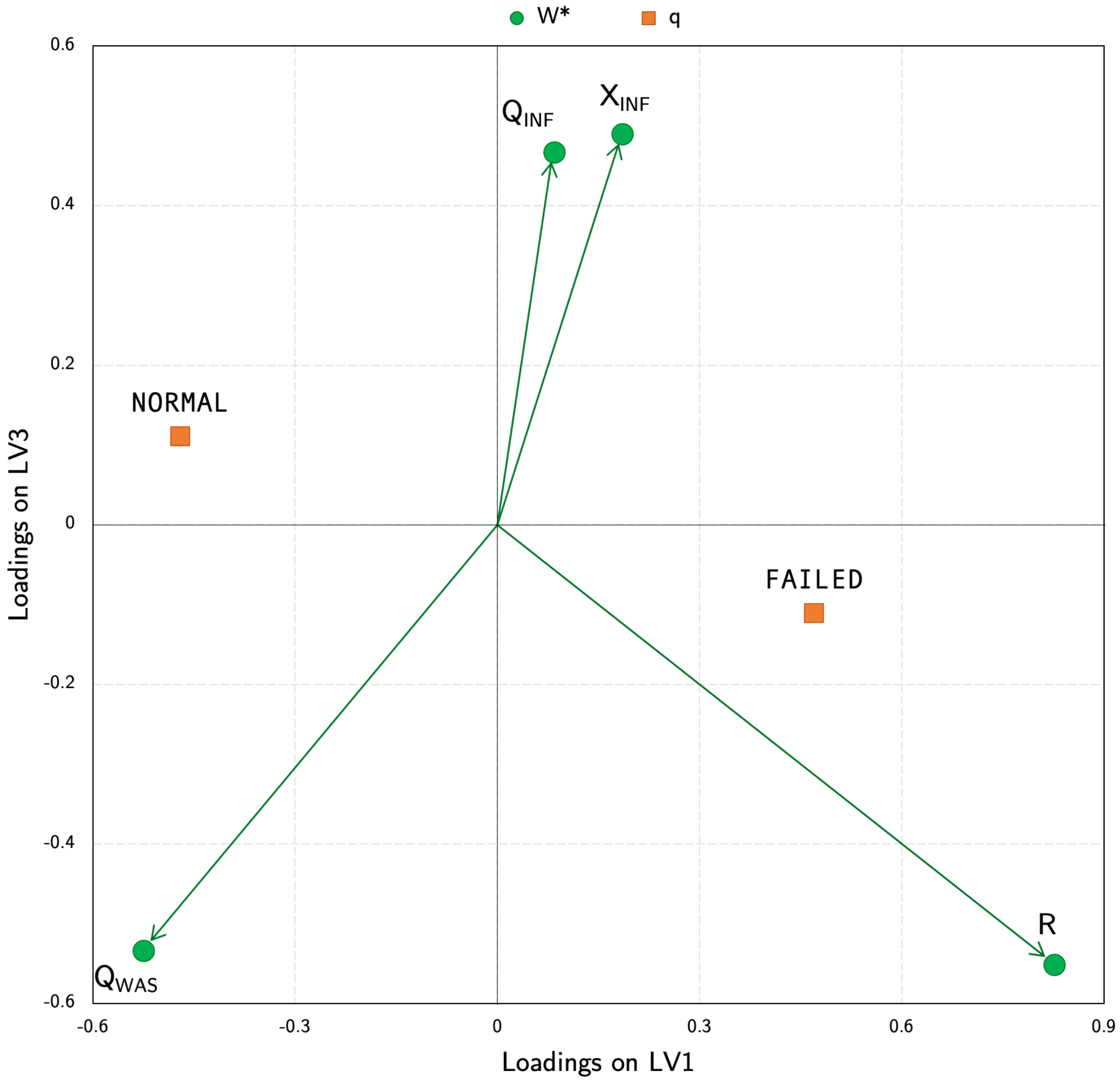


Figure S2: Loading plot showing how the inputs (green) relate to each other and to each class (orange) along the first and third principal components of PLS-DA *Model B* (at SVI = 150 mL/g).

Classification metrics for PLS-DA model calibration

Table S2: Comparison of *balanced accuracy* (BA) scores obtained during calibration for *Model A* (PLS-DA *without R*) and *Model B* (PLS-DA *with R*) at each SVI value considered.

SVI	**Balanced Accuracy (BA)**		**Change in BA**
	Model A	*Model B*	
100	0.882	0.976	0.094
110	0.876	0.980	0.104
120	0.888	0.976	0.088
130	0.897	0.980	0.083
140	0.894	0.986	0.092
150	0.898	0.983	0.084
160	0.903	0.979	0.075
170	0.893	0.989	0.096
180	0.900	0.985	0.085
190	0.916	0.980	0.064
200	0.918	0.963	0.045

Table S3: Comparison of *sensitivity* (i.e., *true positive rate*) scores obtained during calibration for *Model A* (PLS-DA *without R*) and *Model B* (PLS-DA *with R*) test set class membership predictions at each SVI value considered.

SVI	**Sensitivity (TPR)**		**Change in TPR**
	Model A	*Model B*	
100	0.885	0.979	0.094
110	0.880	0.981	0.102
120	0.889	0.976	0.087
130	0.900	0.979	0.079
140	0.895	0.988	0.094
150	0.899	0.982	0.083
160	0.904	0.979	0.074
170	0.894	0.989	0.095
180	0.900	1.000	0.100
190	0.918	0.979	0.062
200	0.919	0.968	0.049

Table S4: Comparison of *specificity* (i.e., *true negative rate*) scores obtained during calibration for *Model A* (PLS-DA *without R*) and *Model B* (PLS-DA *with R*) test set class membership predictions at each SVI value considered.

SVI	**Specificity (TNR)**		**Change in TNR**
	Model A	*Model B*	
100	0.879	0.973	0.094
110	0.873	0.979	0.105
120	0.888	0.976	0.088
130	0.893	0.981	0.088
140	0.894	0.984	0.090
150	0.899	0.983	0.084
160	0.903	0.979	0.076
170	0.893	0.989	0.096
180	0.900	0.969	0.069
190	0.914	0.980	0.066
200	0.918	0.958	0.041

Table S5: Comparison of *Matthews correlation coefficient* (MCC) values obtained during calibration for *Model A* (PLS-DA *without R*) and *Model B* (PLS-DA *with R*) test set class membership predictions at each SVI value considered.

SVI	**Matthews Correlation Coefficient (MCC)**		**Change in MCC**
	Model A	*Model B*	
100	0.570	0.870	0.300
110	0.576	0.901	0.325
120	0.633	0.900	0.267
130	0.666	0.925	0.259
140	0.694	0.949	0.255
150	0.703	0.941	0.238
160	0.729	0.934	0.205
170	0.706	0.965	0.259
180	0.741	0.935	0.193
190	0.788	0.947	0.159
200	0.809	0.907	0.099

Confusion matrices

Table S6: Confusion matrices summarizing test set classification results for *Model A* and *Model B* (both at SVI = 100 mL/g).

Model A

		Actual		
		FAILED	NORMAL	
Predicted	FAILED	18	20	38
	NORMAL	2	160	162
		20	180	

Model B

		Actual		
		FAILED	NORMAL	
Predicted	FAILED	20	6	26
	NORMAL	0	174	174
		20	180	

Table S7: Confusion matrices summarizing test set classification results for *Model A* and *Model B* (both at SVI = 110 mL/g).

Model A

		Actual		
		FAILED	NORMAL	
Predicted	FAILED	15	18	33
	NORMAL	2	165	167
		17	183	

Model B

		Actual		
		FAILED	NORMAL	
Predicted	FAILED	17	3	20
	NORMAL	0	180	180
		17	183	

Table S8: Confusion matrices summarizing test set classification results for *Model A* and *Model B* (both at SVI = 120 mL/g).

Model A

		Actual		
		FAILED	NORMAL	
Predicted	FAILED	29	14	43
	NORMAL	2	155	157
		31	169	

Model B

		Actual		
		FAILED	NORMAL	
Predicted	FAILED	31	2	33
	NORMAL	0	167	167
		31	169	

Table S9: Confusion matrices summarizing test set classification results for *Model A* and *Model B* (both at SVI = 130 mL/g).

Model A

		Actual		
		FAILED	NORMAL	
Predicted	FAILED	32	16	48
	NORMAL	1	151	152
		33	167	

Model B

		Actual		
		FAILED	NORMAL	
Predicted	FAILED	33	1	34
	NORMAL	0	166	166
		33	167	

Table S10: Confusion matrices summarizing test set classification results for *Model A* and *Model B* (both at SVI = 140 mL/g).

Model A

		Actual		
		FAILED	NORMAL	
Predicted	FAILED	22	19	41
	NORMAL	4	155	159
		26	174	

Model B

		Actual		
		FAILED	NORMAL	
Predicted	FAILED	26	5	31
	NORMAL	0	169	169
		26	174	

Table S11: Confusion matrices summarizing test set classification results for *Model A* and *Model B* (both at SVI = 160 mL/g).

Model A

		Actual		
		FAILED	NORMAL	
Predicted	FAILED	34	26	60
	NORMAL	4	136	140
		38	162	

Model B

		Actual		
		FAILED	NORMAL	
Predicted	FAILED	38	1	39
	NORMAL	0	161	161
		38	162	

Table S12: Confusion matrices summarizing test set classification results for *Model A* and *Model B* (both at SVI = 170 mL/g).

Model A

		Actual		
		FAILED	NORMAL	
Predicted	FAILED	32	9	41
	NORMAL	6	153	159
		38	162	

Model B

		Actual		
		FAILED	NORMAL	
Predicted	FAILED	38	1	39
	NORMAL	0	161	161
		38	162	

Table S13: Confusion matrices summarizing test set classification results for *Model A* and *Model B* (both at SVI = 180 mL/g).

Model A

		Actual		
		FAILED	NORMAL	
Predicted	FAILED	30	16	46
	NORMAL	5	149	154
		35	165	

Model B

		Actual		
		FAILED	NORMAL	
Predicted	FAILED	35	3	38
	NORMAL	0	162	162
		35	165	

Table S14: Confusion matrices summarizing test set classification results for *Model A* and *Model B* (both at SVI = 190 mL/g).

Model A

		Actual		
		FAILED	NORMAL	
Predicted	FAILED	48	15	63
	NORMAL	7	130	137
		55	145	

Model B

		Actual		
		FAILED	NORMAL	
Predicted	FAILED	55	3	58
	NORMAL	0	142	142
		55	145	

Table S15: Confusion matrices summarizing test set classification results for *Model A* and *Model B* (both at SVI = 200 mL/g).

Model A

		Actual		
		FAILED	NORMAL	
Predicted	FAILED	54	16	70
	NORMAL	11	119	130
		65	135	

Model B

		Actual		
		FAILED	NORMAL	
Predicted	FAILED	61	4	65
	NORMAL	4	131	135
		65	135	