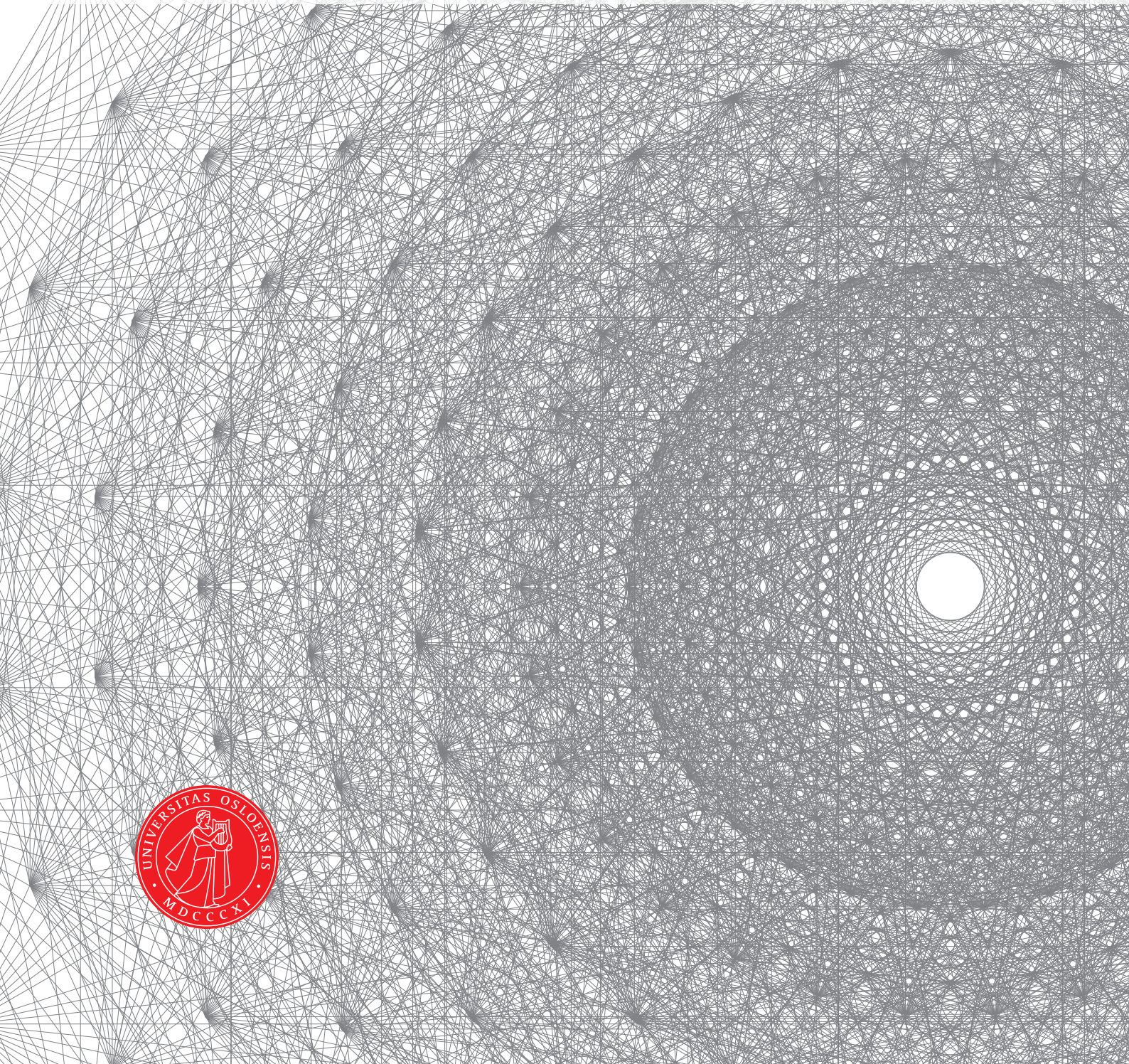
UiO : Department of Mathematics
University of Oslo

Total least squares estimation of maritime battery capacity

Anna Kejvalova

Master's Thesis, Spring 2022



This master's thesis is submitted under the master's programme *Stochastic Modelling, Statistics and Risk Analysis*, with programme option *Statistics*, at the Department of Mathematics, University of Oslo. The scope of the thesis is 60 credits.

The front page depicts a section of the root system of the exceptional Lie group E_8, projected into the plane. Lie groups were invented by the Norwegian mathematician Sophus Lie (1842–1899) to express symmetries in differential equations and today they play a central role in various parts of mathematics.

Abstract

As the shift to environmentally friendly electric and hybrid vessels powered by lithium-ion batteries is taking place globally, there is a growing need for reliable digital tools to monitor battery health and enable safe operation at sea. The total capacity is a measure of battery's maximum energy storage capability and it degrades over time due to several factors. It can be estimated from ordinary linear regression of integrated current measurements on differences in battery's state of charge. Since measurements always have some uncertainty associated with them, the ordinary linear regression gives a capacity estimate biased towards zero by not taking these uncertainties into account. The total least squares approach proposed by Plett (2011) is aimed at correcting for errors in the observed measurements and has other advantages such as recursive implementation and low computational costs. We implement this method and apply it to real battery sensor data. The obtained results are sensitive towards the assumption on the magnitudes of the measurement uncertainties but a goodness of fit criterion helps in better understanding which values are reasonable if additional information is not available. Finally, we compare the results to one annual test and discuss possible improvements of the method's performance.

Acknowledgements

I would like to thank my supervisors Ingrid K. Glad, Morten Stakkeland and Erik Vanem for your guidance through each stage of the process. Thank you for introducing me to the highly interesting topics of this thesis. Your insightful feedback pushed me to sharpen my thinking and brought my work to a higher level, and your immense knowledge and plentiful experience have inspired me and steered me in the right direction. I could not have asked for a better team of supervisors, and I hope another lucky student will get the same opportunity in the future.

I would further like to thank my friends and fellow students at the University of Oslo, and finally, I must express my very profound gratitude to my family for providing me with unfailing support for everything I do and continuous encouragement throughout my years of study.

Oslo, 2022
Anna Kejvalova

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction and Outline

The transition to renewable energies in the maritime transportation sector is an important contribution to the green shift. The Norwegian Government's ambition as outlined in the *Norwegian National Action Plan for Green Shipping* is to reduce emissions from domestic shipping and fishing vessels by half by 2030 and promote the development of zero- and low-emission solutions for all vessel categories (Departementene 2022). Although Norway is a world leader within green shipping, the pace of this shift must be increased in order to achieve these climate goals. Electric or hybrid ships using batteries are emerging as an attractive alternative to fossil fuels due to their significant environmental benefits.

Currently lithium-ion batteries are the leading energy storage technology and thanks to their development, transportation including maritime is becoming increasingly battery driven. A key safety aspect of battery-powered ships is ensuring that the available energy stored in the batteries is sufficient to cover the required power demand at all times, as loss of propulsion power in critical situations can lead to serious accidents (Vanem et al. 2021). Lithium-ion batteries are subject to ageing processes and these affect both the amount of charge that can be stored as well as the performance of the power delivery. Therefore, reliable estimation and prediction of actual available energy of a battery is crucial for safe and sustainable operation of battery-powered ships. Most maritime systems are designed with an expected lifetime of 10 years with the end of life typically defined as SOH = 70-80%, where SOH stands for the ratio of remaining capacity to the initial capacity (Vanem et al. 2021). Annual capacity tests can be conducted to measure battery capacity but they are very time-consuming and typically require the ship to be taken out of operation for one full day per year. Hence new data-driven approaches to SOH monitoring and prediction utilizing sensor data need to be explored to estimate the effect of degradation of the batteries.

The most common tools for assessing battery health are calculations of "State of Health" and "State of Charge". In order to obtain these, it is important to estimate the total capacity of the batteries well and in real time. In this thesis, we will work with improved estimation of total capacity. It has been shown that error-in variables models are well suited for this problem and yield the most precise estimates of total capacity (Plett 2011). The main goal of this thesis is to construct methods for estimating total capacity of maritime

batteries while taking into account the uncertainty of measurements. We will present measurement error regression and compare it to the total least squares methods derived by Plett (2011) which we will implement and try out on real battery sensor data from the Norwegian battery producer Corvus Energy.

We now give an outline of the thesis. Firstly, in Chapter 2 we illustrate how a lithium-ion battery works and introduce some related terminology. In Chapter 3 we describe measurement error models and the consequences of ignoring measurement error for naive statistical analysis. In Chapter 4 we present the total least squares methods introduced by Plett (2011). In Chapter 5, we try to come up with a model for the measurement uncertainties for the variables we will be using in our analysis. In Chapter 6, we describe the different measures that had to be taken in the data preparation stage. In Chapter 7, we apply the methods to our data. We then compare the estimates of these methods to one annual test and describe a goodness-of-fit criterion. Finally, we summarize and discuss our findings, addressing the strengths and weaknesses of our proposed methods as well as possible improvements for future research.

CHAPTER 2

Lithium-ion battery terminology

2.1 Overview

This chapter gives general information on batteries that will be required as background knowledge for the rest of the thesis. A lithium-ion battery is a widely used energy storage system that converts chemical energy into electrical energy and vice versa, by means of an electrochemical reaction. As the world is becoming electrified, rechargeable lithium-ion batteries have gained a dominant position in the energy storage market and have replaced other batteries in many application areas due to their high energy density which means that they can store more energy in a given volume (Mikolajczak et al. 2012, chapter 2). They typically consist of many battery cells which again can be divided into a few main components. In this chapter, we will describe how a lithium-ion battery works, define other battery related quantities as well as mechanisms that cause battery degeneration.

2.2 Main components of a lithium-ion battery cell

The main function of a lithium-ion battery is to store and then release energy by converting chemical energy into electric energy. The basic unit of a lithium-ion battery is a battery cell that exerts electric energy by charging and discharging (Mikolajczak et al. 2012, chapter 1). A single cell can be sufficient for many portable electronic devices whereas for large scale applications, many cells integrated into packs or modules are required to meet the energy and power demands (Zhang and Lee 2011, chapter 2). A lithium-ion battery cell consists of the following main components: the positive and negative electrodes, often referred to as the *cathode* and *anode*, respectively, *the electrolyte*, a *separator* and *current collectors* (Korthauer 2018, chapter 2).

The cell's active materials reside in the electrodes, where the oxidation, loss of electrons, and reduction, gain of electrons, processes take place in order to liberate or bind lithium ions Li^+ and electrons e^- (Vanem et al. 2021). These reversible redox reactions between the cathode and anode are the basis for the rechargeability of the battery cell (Zhang and Lee 2011, chapter 2). The liberated lithium ions are allowed to diffuse between the electrodes through the electrolyte, and the electrons as electricity carriers can be transported by the current collectors to generate a potential between the battery terminals and hence drive a current in an outer circuit. The main function of the electrolyte

is therefore to transport lithium ions between the electrodes while the separator electrically isolates both electrodes to prevent self-discharge from the cell (Vanem et al. 2021).

A rechargeable battery cell operates in two modes, that is *charging* and *discharging*. A schematic view of a typical lithium-ion cell in these two modes is shown in Figure 2.1 below. During discharging, the lithium ions migrate from the negative electrode through the electrolyte and the separator to the positive electrode, while at the same time, the electrons are transported from the negative electrode via an outer electrical connection to the positive electrode (Korthauer 2018, chapter 2). During charging, this process is reversed and the lithium ions migrate from the positive to the negative electrode through the electrolyte and the separator. When the battery is fully charged, the active lithium ions reside in the anode, and when it is fully discharged, they reside in the cathode.

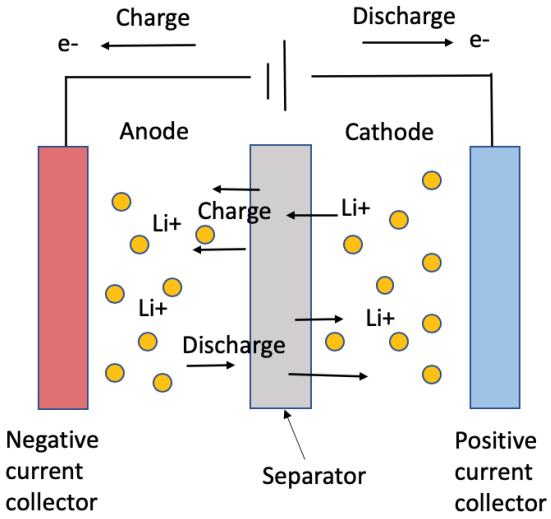


Figure 2.1: Illustration of a typical lithium-ion battery during charge and discharge

For a rechargeable battery, these two processes at the positive and negative electrodes can be repeated many times in a sequence of charge-discharge cycles. The cycle life of a rechargeable battery refers to the number of full discharge-charge cycles the battery can experience before its end of life and it is influenced by many different factors such as rate and depth of the cycles, temperature and humidity (Vanem et al. 2021).

2.3 Battery terminology

Battery systems are typically equipped with a battery management system, *BMS*, whose main function is to protect the battery cells from overcharging and extreme temperatures, and thereby increase their lifetime. It uses sensor technology to monitor and control the battery to ensure it is always operating

within safe limits and that any risk of damage to the battery is prevented. It measures the battery cell control parameters such as current, voltages and temperatures, and enables the switching on and off of the battery system (Korthauer 2018, chapter 2). Besides, it monitors and controls the battery's charging and discharging process and the condition of the battery in terms of its available capacity for energy storage (Vanem et al. 2021).

Many battery cells are necessary to generate the amount of power needed to propel an electric vessel. The composition of an electric vessel battery might vary slightly but generally they are are composed of *cells, modules* and *packs*. Eventually, packs can be integrated into *arrays*. Cells, modules, packs and arrays are units of clustered batteries. In simple terms, a cluster of assembled cells forms a module, and a cluster of modules forms a pack. Finally, packs bundled into an array are the final form of the battery installed in the electric vessel together with a battery management system and a cooling system that controls and manages for example the battery's temperature and voltage (Mikolajczak et al. 2012, chapter 1). The purpose of connecting several battery cells in a system is to increase the energy and power of the battery and moreover, to facilitate easy replacement of faulty parts of larger battery packs (Mikolajczak et al. 2012, chapter 1).

The cells and modules can be connected *in series* or in *parallel* as Figure 2.2 shows, or a mixture of both. This impacts the voltage and capacity of the battery system. A battery system wired in series will have the voltages of individual cells added together, so that the overall voltage is increased, but the capacity stays the same. In contrast, a battery system wired in parallel will have the individual capacities added together while the battery voltage remains the same (Mikolajczak et al. 2012, chapter 1). Altogether, connecting cells in series increases the potential of the battery system while connecting cells in parallel increases the capacity of the system.

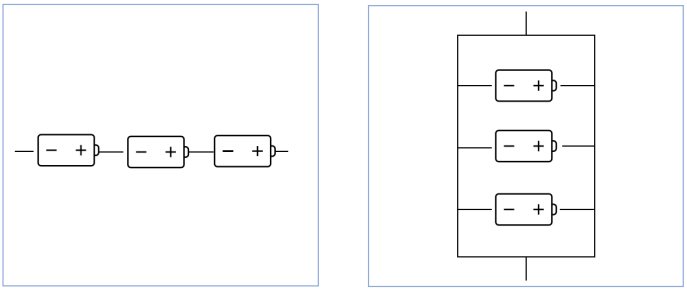


Figure 2.2: Illustration of battery cells connected in series (left) and in parallel (right)

In the following we will define various quantities which describe the present condition of a battery and may be monitored by a condition monitoring system through collected sensor data.

- **Total capacity**

  The total capacity of a battery cell, denoted by Q, indicates the maximum electrical charge that the battery cell is capable of holding (Plett 2011). The value of battery capacity is commonly expressed in ampere-hours (Ah) or ampere-seconds (As). It is important to note that the total capacity is not a fixed quantity but it decays over time as the battery wears out.

- **Nominal capacity**

  The nominal capacity Q_{nom} is a constant quantity specified by the battery manufacturer and it describes the capacity of a new battery cell (Plett 2011).

- **State of Health (SOH)**

  The capacity of a battery to store energy will typically degrade over time, and the state of health, SOH, is measure of such degradation. There are alternative definitions of SOH, some related to capacity, other related to internal resistance or power. In this thesis we will define SOH as a measure of the battery's total capacity relative to its nominal capacity (Plett 2011). Its value is typically given in % by

  $$SOH = \frac{Q}{Q_{nom}} \cdot 100(\%). \tag{2.1}$$

  In other words, SOH indicates which point the battery has reached in its life cycle and how well it performs compared to a fresh battery.

- **State of Charge (SOC)**

  With a rechargeable batter system, the amount of energy available at all times will vary continuously as the battery is repeatedly charged and discharged. SOC is a unitless value between 0% and 100 % that indicates the relative level of charge presently held by the battery cell (Plett 2011). Hence, SOC of 100% corresponds to a fully charged cell while SOC of 0% corresponds to an empty discharged cell. SOH and SOC depend on each other and influence the battery performance. Moreover, SOC is not to be confused with the battery cell total capacity but they are related through an equation which we will state later.

- **C-rate**

  The C-rate is a measure of the rate at which a battery is being charged or discharged related to its nominal capacity. It is defined as the current through the battery divided by the current draw under which the battery would theoretically deliver its nominal capacity in one hour (Team 2008). As follows, a C-rate of 1C means that the battery is fully charged, or fully discharged within one hour, so a battery of capacity 150 Ah would provide 150 A of current for one hour at 1C.

Usually, SOC cannot be measured directly but it can be inferred based on other measured variables. An accurate SOC and SOH estimation method will help improve a battery's performance and reliability, and ultimately prolong its

lifetime. Moreover, estimation of battery capacity is closely related to that of SOC, as SOC is usually defined as the ratio between available capacity and the total cell capacity (Zhang and Lee 2011). The main methods for SOC estimation of lithium-ion batteries can be divided into three: a current-based method, a voltage-based method, and a fusion of these two approaches which seeks to combine the information obtained from voltage and current measurements using non-linear filters, such as the extended *Kalman filter* (Movassagh et al. 2021). Kalman filter uses a recursive algorithm that continuously predicts the future state of charge and corrects it using measurements performed on the system, including current, voltage and temperature. An accurate dynamic model including dependencies on operating and environmental conditions is required for the Kalman filter to function correctly (Movassagh et al. 2021). We will look at the current-based method, commonly known as *Coulomb counting*, in more detail.

Electric current I is a measure of the flow of electric charge q and it is given in amperes. A flow is a rate, meaning an amount over an elapsed time t, so current can be expressed as the partial derivative

$$I(t) = \frac{\partial q(t)}{\partial t}. \tag{2.2}$$

When we integrate on both sides, we obtain the formula

$$q = \int_{t_1}^{t_2} I(\tau)\, d\tau. \tag{2.3}$$

which represents the amount of charge flowing between times t_1 and t_2. We define current $I(t)$ to be positive when charge is being transported into the battery, i.e. when battery is being charged, and a negative value when it is being discharged. Accordingly, $q > 0$ corresponds to charging and $q < 0$ corresponds to discharging.

The Coulomb counting method, also known as ampere hour counting or current integration, is the most common technique for calculating the SOC (Lu et al. 2013). It works by integrating the current flowing over time to derive the total sum of energy entering or leaving the battery. During a full charging cycle, when the battery goes from being fully discharged to fully charged, this method integrates the current flowing to or from the battery to estimate the total battery capacity Q directly, according to the basic relation

$$Q = \int_{t_0}^{t_1} I(\tau)\, d\tau, \tag{2.4}$$

where $I(\tau)$ is the current at time τ, and t_0 and t_1 refer to times where SOC = 0% and SOC=100%, respectively (Vanem et al. 2021).

This equation can be modified to include the Coulomb efficiency factor η which is equal to 1 while discharging and is smaller than 1 while charging (Lu et al. 2013). Using this approach, a full cycle is required to be able to estimate the total capacity which is rarely the case in actual operations, and also the measurements need to be performed under controlled conditions, with

constant, typically low, C-rate and a specific temperature. Moreover, any errors in current measurements will accumulate and subjecting the battery to full cycles may contribute to accelerated degradation and shortening of the battery lifetime (Vanem et al. 2021).

However, total capacity estimation can be based on Coulomb counting of not necessarily full cycles. By definition, SOC can be considered as the rate of the integral of the current flowing in and out of the cell of a battery over the total capacity. Starting from a fully discharged battery with SOC = 0% at time t_0, SOC at time t can be computed as the ratio of integrated current from t_0 until t by

$$SOC(t) = \frac{1}{Q} \int_{t_0}^{t} \eta I(\tau)\, d\tau. \tag{2.5}$$

The relationship between the total capacity Q and SOC at times t_1 and t_2 is then as follows

$$SOC(t_2) = SOC(t_1) + \frac{1}{Q} \int_{t_1}^{t_2} \eta I(\tau)\, d\tau \tag{2.6}$$

where $I(\tau)$ is the battery cell current at time τ measured in amperes, which is positive when charging and negative when discharging, and η is a unitless Coulomb efficiency factor (Lu et al. 2013). If the time is measured in seconds, then the unit of total capacity Q will be ampere-seconds. In order to obtain a total capacity Q given in ampere-hours, the factor 3600 is required to convert the time measured in seconds to hours and the resulting formula is as follows

$$SOC(t_2) = SOC(t_1) + \frac{1}{Q} \int_{t_1}^{t_2} \frac{\eta I(\tau)}{3600}\, d\tau \tag{2.7}$$

The formula above is the mathematical basis for most battery capacity estimation methods and by rewriting it, we obtain the following equation

$$\underbrace{\int_{t_1}^{t_2} \frac{\eta I(\tau)}{3600}\, d\tau}_{y} = Q \underbrace{\left(SOC(t_2) - SOC(t_1)\right)}_{x}. \tag{2.8}$$

The linear structure $y = Qx$ of this equation allows us to compute an estimate of Q by using a regression technique (Plett 2011). However, both the integrated current values y and the difference between the SOC values x have sensor noise or estimation noise associated with them and ordinary least squares regression which does not take into account measurement errors may lead to inaccurate and biased estimate of the total capacity.
Besides, the Coulomb counting formulas above illustrate the mutual dependence between the SOC and total capacity estimates. An accurate estimate of Q is needed to give accurate estimates of SOC and vice versa. Therefore, when the goal is to estimate the total capacity Q, the Coulomb counting method for SOC estimation may be inappropriate since it leads to such circular dependencies and hence unstable estimates (Plett 2011). Plett recommends the use of Kalman filter based methods for SOC estimation which seem to be sensitive to errors in the capacity estimates as it corrects for the estimates using voltage measurements.

2.4 Battery ageing

The performance of Li-ion batteries deteriorates with time and usage due to the degradation of their electrochemical components, which leads to a capacity and power fade. This is called *battery ageing* and is a consequence of various ageing mechanisms influenced by different factors such as battery chemistry and manufacturing, as well as environmental and operating conditions. Accurate health diagnostic and prognostic tools are crucial to ensure the safety and reliability of batteries despite ageing. These are then implemented in the battery management system for online condition monitoring and enable the users to keep track of the performance of the battery and schedule maintenance and repairs in advance.

Various degradation processes contribute to aging of lithium-ion batteries and they affect different elements of a battery. Battery aging can be divided into two modes, *calendar* and *cyclic aging*. Calendar aging comprises all processes that occur regardless of battery's charge-discharge cycling, therefore even when the battery is not in use (Li et al. 2019). In contrast, cyclic aging refers to the ageing from the continuous battery charge and discharge cycles and is affected by additional factors such as overcharge and -discharge, current rate and cycling depths. Very deep cycles, that is larger variations in SOC, typically increase the rate of battery degradation, compared to shallow cycles. For example, high SOC implies low Li content in the active material of the cathode which then increases its tendency to chemically decompose the electrolyte components (Lu et al. 2013). Furthermore, higher levels of current, that is charging/discharging the battery at higher C-rates, will accelerate the degradation (Vanem et al. 2021). Overcharging the cell can generate significant heat and this can trigger a series of side reactions at both electrodes. Cells are also exposed to other stress factors, such as damage caused during manufacturing, or electrode material expansion during operation (Li et al. 2019). Mechanical loads might form cracks within the active materials where the lithium ions are intercalated, causing them to no longer be electrically connected (Korthauer 2018, chapter 2).

The lifetime of a battery depends on the operating conditions, the applied materials, the electrolyte composition, and the quality of the production process (Korthauer 2018, chapter 2). The literature on maritime battery systems specifically is rather scarce. Nevertheless, many parallels can be drawn from other battery application areas such as electric vehicles, and the overall degradation mechanisms are believed to be very similar (Vanem et al. 2021). The differences between maritime batteries and others are mainly related to battery size and designs, different operational environments and loading profiles, and different safety aspects (Vanem et al. 2021). The cycles of maritime batteries will vary according to the type of operation and also the environmental conditions under which it is operated. It has not yet been fully explored to what extent exposure to factors such as humid and saline environments or ship motions may influence battery degradation. Furthermore, the temperatures and loads may not be evenly distributed within a battery system consisting of several modules and battery cells. The overall degradation will likely depend on the battery design and different cells may experience different degradation trends (Vanem et al. 2021).

CHAPTER 3

Measurement Error Modelling

3.1 Overview

Every measurement has some random noise associated with it. This uncertainty in the measurement may arise for different reasons such as limitations of various measuring devices, environmental factors or carelessness of the experimenter. The consequences of ignoring measurement error can range from negligible to more extensive. If measurement error is large, estimations of coefficients and the variable selection of a statistical model may be greatly affected. This chapter presents different types of measurement errors, summarizes some of the known results about the effects of measurement error in linear regression and describes a way of modelling measurement error as well as some of the statistical methods used to correct for its effects. Measurement error in explanatory variables has many effects ranging from attenuation of the slope estimates to a loss of power for detecting interesting relationships among variables and masking of features of the data in nonlinear models (Carroll, Ruppert et al. 2006, chapter 1). In simple linear regression in particular, measurement errors cause bias of the slope estimate in the direction towards zero, meaning underestimation of its absolute value. Such a bias is commonly referred to as *attenuation* or *attenuation to the null* (Carroll, Ruppert et al. 2006, chapter 3). The main objectives of this chapter are to find out what are the consequences of measurement error for naive statistical analyses and how one can correct for it.

3.2 General concept and notation

Firstly, we broadly look at different types of errors and the motivation behind modelling measurement errors. Measurement error occurs when one cannot measure exactly a variable of interest that enters into a model. There are many reasons measurement errors occur, and they can be classified into *random* and *systematic error* depending on how the measurements are obtained. Random errors are fluctuations which may vary from observation to observation due to uncontrolled factors such as limited precision of the measurement instruments while systematic errors are reproducible inaccuracies caused by imprecise instruments and faulty equipment (Bevington and Robinson 1992, chapter 1).

There are different ways of expressing measurement error and characterizing the accuracy of measurements. One way of quantifying error is the *absolute error*,

defined as the absolute difference between the true and measured value. Since we usually do not know the the actual true value, we use the maximum possible error. The *relative error* is the error proportional to the true value and is defined by absolute error divided by the observed measurement (Bevington and Robinson 1992, chapter 1). Relative error is often written as a percentage. In our thesis, when defining measurement error, we will speak of the absolute error.

Statistical regression models define a mathematical relationship between one or more independent explanatory variables X and a dependent response variable Y. When the true value of one of these variables is not observable for some reason, it is common to substitute them with observed variables which contain measurement error. This substitution complicates the statistical analysis of the observed data when the purpose is statistical inference about a model defined in terms of the true values. Measurement error modelling seeks to fit a model described in terms of the true values given the observed error-prone data (Stefanski 2000).

The key components of measurement error modelling are a statistical *model for the true values*, a *measurement error model* which specifies the relationship between the true and observed values, and *additional assumptions* or *extra data* such as replicate values or standard errors of the error-prone variables needed to correct for the error (Buonaccorsi 2010, chapter 1).

Many different notations are being used for measurement errors which make it rather difficult to read different literature on this topic. Carroll, Ruppert et al. (2006), for example, write $\mathbf{X}$ and $\mathbf{W}$, Fuller (1987) uses x and X, and lastly, Buonaccorsi (2010) uses x and $\mathbf{W}$, for the true value and its error-prone measurement, respectively. Throughout the thesis, we will use our own notation consistently, where we denote the true values by x^* and y^*, and the observed measurements by x and y. We will denote the measurement errors in x and y, by Δx and Δy, respectively. Further we will follow the common notation of upper case letters for random variables and lower case for observed values.
One important distinction has to be made between two ways of modelling the true variables, namely the *functional case*, where the true values x^* are regarded as fixed unknown constants, and the *structural case*, where the true x^* are treated as random variables, hence written as X^*, which are usually assumed to be independent and identically distributed with mean μ_{X^*} and a covariance Σ_{X^*} (Buonaccorsi 2010, chapter 1).

3.3 Models for true values

The general regression model for the true values $\mathbf{y}^*$ and $\mathbf{x}^*$ is commonly defined by

$$Y_i^* | \mathbf{x}_i^* = f(\mathbf{x}_i^*, \beta) + \epsilon_i \tag{3.1}$$

where the function $f(\mathbf{x}_i^*, \beta)$ defines the type of regression model and ϵ_i is a random independent error term with $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = v(\mathbf{x}_i^*, \beta, \sigma)$ for some variance function v (Buonaccorsi 2010, chapter 6). In case of constant variance it is common to write $v(\mathbf{x}_i^*, \beta, \sigma) = \sigma^2$. Specific classes of models include for example linear models, nonlinear models and generalized linear

models such as logistic, probit and Poisson models. Linear models are linear in the parameters and the most simple one is linear regression which we will explore in more detail in the rest of the chapter. Statistical analysis that is conducted by ignoring measurement error is called a *naive approach* (Buonaccorsi 2010). The effects of measurement error are determined by its distribution and the appropriate methods used for correcting for these effects will vary depending on the model.

3.4 Measurement error models

There is a wide variety of ways of modelling measurement errors. First differentiation can be made between how the true and observed variables are related to each other, namely whether one makes an assumption about the distribution of observed values $\mathbf{x}$ given the true values $\mathbf{x}^*$ or vice versa. If the true values $\mathbf{x}^*$ are fixed or we condition on them, then the *classical measurement error model* applies (Buonaccorsi 2010, chapter 1). It models the conditional distribution of $\mathbf{x}$ given the true $\mathbf{x}^*$. On the contrary, when for example an experimenter is trying to achieve a target value $\mathbf{x}$ but the true value achieved is $\mathbf{x}^*$ which varies in repeated sampling, then the *Berkson model* applies. It models the conditional distribution of $\mathbf{x}^*$ given the target $\mathbf{x}$ (Buonaccorsi 2010, chapter 1). In our thesis, we will further focus on the classical measurement error model.

Measurement error can affect covariates in many different ways. If we consider measurement error in $\mathbf{x}$, the simple general definition of the *linear measurement error model* states, written in our notation, that

$$\mathbf{X}|\mathbf{x}^* = \theta_0 + \Theta_1\mathbf{x}^* + \Delta\mathbf{x} \tag{3.2}$$

where $\Delta\mathbf{x}$ is a random measurement error with $E(\Delta\mathbf{x}|\mathbf{x}^*) = 0$. (Buonaccorsi 2010, chapter 6)
Hence it follows that

$$E(\mathbf{X}|\mathbf{x}^*) = \theta_0 + \Theta_1\mathbf{x}^*. \tag{3.3}$$

Nonlinear measurement error includes the types of error where

$$E(\mathbf{X}|\mathbf{x}^*) = g(\theta, \mathbf{x}^*) \tag{3.4}$$

where $g(\theta, \mathbf{x}^*)$ is nonlinear in the θ's (Buonaccorsi 2010, chapter 6).

Much of the literature is based around *classical additive measurement error*, in which the truth is measured with additive error, usually with constant variance. The classical additive measurement error is a special case of the linear measurement error with $\theta_0 = 0$ and $\Theta_1 = I$ where I is the identity matrix (Buonaccorsi 2010, chapter 6).
Therefore, the classical additive measurement error is defined as

$$\mathbf{X}|\mathbf{x}^* = \mathbf{x}^* + \Delta\mathbf{x}. \tag{3.5}$$

Since here again we assume zero-mean measurement error $\Delta\mathbf{x}$ with $E(\Delta\mathbf{x}|\mathbf{x}^*) = 0$, it follows that

$$E(\mathbf{X}|\mathbf{x}^*) = \mathbf{x}^*, \tag{3.6}$$

so $\mathbf{X}$ is unbiased for the unobserved $\mathbf{x}^*$.

Further differentiation can be made about the form of measurement error variances, denoted $\sigma^2_{\Delta\mathbf{x}}$, and possibly covariances. In a *homoscedastic model*, the variance of $\mathbf{X}$ given $\mathbf{x}^*$ is constant while in the *heteroscedastic model* the variance varies across observations (Buonaccorsi 2010, chapter 6).

As mentioned before, in this thesis we will focus on classical measurement error, specifically additive error. Firstly, we will allow for measurement error both in predictor and response and possibly varying measurement error variances, then we will narrow it down to special cases and present well-known results for the additive measurement error model with no error in the response or uncorrelated measurement errors. With continuously measured variables, as in our case, the classical additive error model where the measurement error structure is approximately normal with constant variance is often assumed. How one could check this assumption is an open question with many possible solutions. It has been suggested, in order to determine whether the normality assumption holds and whether measurement errors have constant variance, to assess the normality of the differences in replicates of a single measured value by plotting the intra-individual standard deviation against the mean and seeing if there are any obvious trends indicating otherwise, or by forming a qq-plot of the differences across individuals (Carroll, Ruppert et al. 2006, chapter 1). However, replicates required for these kinds of checks are not always available.

3.5 Simple linear regression and additive measurement error

Consider a simple linear regression model which assumes that there is no measurement error. Suppose we observe n data pairs $\{(y_i^*, x_i^*), i = 1, ..n\}$ which denote the true values without measurement error. Then the model is defined by

$$Y_i^*|x_i^* = \beta_0 + \beta_1 x_i^* + \epsilon_i, \quad i = 1, ...n \tag{3.7}$$

where Y^* is the response variable, x^* the predictor and ϵ_i are independent random variables with mean 0 and variance σ^2_ϵ. The error term ϵ_i denotes the error in equation, also called residual, which accounts for the lack of fit when the model does not fully represent the actual relationship between the independent and the dependent variables, and it has to be distinguished from potential measurement error in the response variable (Buonaccorsi 2010, chapter 4).

Regression models are most commonly analyzed using the least squares and maximum likelihood approach where maximum likelihood requires an assumption about the distribution of $Y^*|x^*$. The least squares method may be either unweighted or weighted. Unweighted least squares minimizes the sum of squared residuals to estimate the coefficients while weighted least squares minimizes the weighted squared residuals. In the case of a heteroscedastic model with nonconstant variance, the weighting often leads to iteratively reweighted least squares (Buonaccorsi 2010, chapter 6).

Using the ordinary, unweighted, least squares approach, we minimize the

sum of squared residuals

$$\sum_{i=1}^{n}(Y_i^* - \beta_0 - \beta_1 x_i^*)^2 \tag{3.8}$$

to obtain the following estimators

$$\hat{\beta}_0 = \bar{Y}^* - \hat{\beta}_1 \bar{x}^* \tag{3.9}$$

and

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i^* - \bar{x}^*)(y_i^* - \bar{y}^*)}{\sum_{i=1}^{n}(x_i^* - \bar{x}^*)^2}. \tag{3.10}$$

When there is no measurement error in any variables, these estimators are unbiased (Buonaccorsi 2010, chapter 3).

The *classical measurement error model*, also called *additive measurement error model*, assumes that one is unable to observe the true values of the predictor and/or response variable directly but rather with some additive error. For observation i, we define X_i as the error-prone measurement of X_i^*, and if there is error in the response, we define Y_i as the error-prone measurement of Y_i^*.

Now we substitute the error-prone measurements for the true values and similarly, using the least squares approach, we obtain the following *naive estimators* of the coefficients and the error variance

$$\hat{\beta}_{0naive} = \bar{Y} - \hat{\beta}_{1naive}\bar{X} \tag{3.11}$$

and

$$\hat{\beta}_{1naive} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{3.12}$$

$$\hat{\sigma}^2_{naive} = \sum_{i=1}^{n}(Y_i - (\hat{\beta}_{0naive} + \hat{\beta}_{1naive}X_i))^2/(n-2) \tag{3.13}$$

They are called *naive* because they ignore the effects of measurement error. Without any measurement errors, these are again unbiased estimators.

Given the true x_i^* and y_i^*, the classical measurement error model specifies the joint behavior of X_i and Y_i as the sums of the true value and measurement error (Buonaccorsi 2010, chapter 4). We define that given x_i^* and y_i^*,

$$X_i = x_i^* + \Delta x_i \tag{3.14}$$

$$Y_i = y_i^* + \Delta y_i \tag{3.15}$$

with

$$\begin{aligned} E(\Delta x_i | x_i^*) &= 0 \\ E(\Delta y_i | y_i^*) &= 0 \\ Var(\Delta x_i | x_i^*) &= \sigma^2_{\Delta x_i} \\ Var(\Delta y_i | y_i^*) &= \sigma^2_{\Delta y_i} \\ Cov(\Delta x_i, \Delta y_i | x_i^*, y_i^*) &= \rho_i. \end{aligned}$$

In the above, Δx_i and Δy_i are the measurement errors in X_i and Y_i, respectively. In addition, we assume that the measurement errors $(\Delta x_i, \Delta y_i)$ are independent over i and uncorrelated with ϵ_i (Buonaccorsi 2010, chapter 4).

If either of the variables is observed exactly without any error, then the corresponding measurement error as well as its variance and covariance is set equal to 0.

We allowed in our definition above for a heteroscedastic measurement error model in which the measurement error variances and covariances vary with observation i. This could occur in practice due to many reasons such as a change in sampling effort, change in variability in the measuring instrument or the fact that the variance may be related to the true value. (Buonaccorsi 2010, chapter 4)

However, since it is difficult to state any general results on the exact behavior of naive analyses in this case, the majority of the literature has focused on special cases of measurement error models where there is no measurement error in y or the measurement variances do not change with i (Buonaccorsi 2010, chapter 4). Let us look at at the special case of *normal structural model with normal additive measurement error with constant measurement error variances* as defined in Buonaccorsi (2010, chapter 4).

Normal structural model with normal additive measurement error and constant measurement error variances/covariance

We will here consider a structural model. Assume that $X_1^*, ..X_n^*$ are independent normally distributed random variables with mean μ_{X^*} and variance $\sigma^2_{X^*} > 0$, and that the error term ϵ_i is normally distributed with mean 0 and variance σ^2_ϵ. Further we assume that given the true values (y_i^*, x_i^*), the measurement errors Δx_i and Δy_i are bivariate normal and their variances, denoted by $\sigma^2_{\Delta x}$ and $\sigma^2_{\Delta y}$, respectively, are constant for all i, as well as their covariance ρ.

As Buonaccorsi (2010) in chapter 4 shows, if the true values (Y_i^*, X_i^*) are distributed normally with

$$\begin{pmatrix} Y_i^* \\ X_i^* \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_{Y^*} \\ \mu_{X^*} \end{pmatrix}, \quad \begin{pmatrix} \sigma^2_{Y^*} & \sigma_{X^*Y^*} \\ \sigma_{X^*Y^*} & \sigma^2_{X^*} \end{pmatrix}\right] \tag{3.16}$$

where N stands for normal distribution, and if given y_i^*, x_i^*,

$$\begin{pmatrix} Y_i \\ X_i \end{pmatrix} \sim N\left[\begin{pmatrix} y_i^* \\ x_i^* \end{pmatrix}, \quad \begin{pmatrix} \sigma^2_{\Delta y} & \rho \\ \rho & \sigma^2_{\Delta x} \end{pmatrix}\right], \tag{3.17}$$

it follows that (Y_i, X_i), as defined in Equations 3.14 and 3.15, are bivariate normal with

$$\begin{pmatrix} Y_i \\ X_i \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_{Y^*} \\ \mu_{X^*} \end{pmatrix}, \quad \begin{pmatrix} \sigma^2_{Y^*} + \sigma^2_{\Delta y} & \sigma_{X^*Y^*} + \rho \\ \sigma_{X^*Y^*} + \rho & \sigma^2_{X^*} + \sigma^2_{\Delta x} \end{pmatrix}\right] \tag{3.18}$$

Using the assumptions from linear regression

$$E[Y^*|X^* = x^*] = \beta_0 + \beta_1 x^* \quad \text{and} \quad Var(Y^*|X^* = x^*) = \sigma^2_\epsilon \tag{3.19}$$

which imply that

$$\mu_{Y^*} = \beta_0 + \beta_1 \mu_{X^*}$$
$$\sigma^2_{Y^*} = Var(\beta_0 + \beta_1 x^* + \epsilon) = \beta_1^2 \sigma^2_{X^*} + \sigma^2_\epsilon \quad \text{and}$$
$$\beta_1 = \sigma_{X^*Y^*} / \sigma^2_{X^*},$$

we can rewrite Equation (3.18) as

$$\begin{pmatrix} Y_i \\ X_i \end{pmatrix} \sim N \left[\begin{pmatrix} \beta_0 + \beta_1 \mu_{X^*} \\ \mu_{X^*} \end{pmatrix}, \quad \begin{pmatrix} \beta_1^2 \sigma^2_{X^*} + \sigma^2_\epsilon + \sigma^2_{\Delta y} & \beta_1 \sigma^2_{X^*} + \rho \\ \beta_1 \sigma^2_{X^*} + \rho & \sigma^2_{X^*} + \sigma^2_{\Delta x} \end{pmatrix} \right]$$

Now applying the well-known results about conditional distribution from a bivariate normal distribution, we have that the conditional distribution of $Y_i | X_i = x_i$ is normal with mean

$$\begin{aligned} \mu_{Y|X} &= \beta_0 + \beta_1 \mu_{X^*} + (\beta_1 \sigma^2_{X^*} + \rho)(\sigma^2_{X^*} + \sigma^2_{\Delta_x})^{-1}(x_i - \mu_{X^*}) \\ &= \beta_0 + \beta_1 \mu_{X^*} - \frac{\mu_{X^*}(\beta_1 \sigma^2_{X^*} + \rho)}{\sigma^2_{X^*} + \sigma^2_{\Delta x}} + \frac{x_i^*(\beta_1 \sigma^2_{X^*} + \rho)}{\sigma^2_{X^*} + \sigma^2_{\Delta x}} \\ &= \beta_0 + \frac{\beta_1 \mu_{X^*} \sigma^2_{X^*} + \beta_1 \mu_{X^*} \sigma^2_{\Delta_x} - \mu_{X^*} \beta_1 \sigma^2_{X^*} - \mu_{X^*} \rho}{\sigma^2_{X^*} + \sigma^2_{\Delta_x}} + \frac{x_i(\beta_1 \sigma^2_{X^*} + \rho)}{\sigma^2_{X^*} + \sigma^2_{\Delta x}} \\ &= \underbrace{\beta_0 + \frac{\mu_{X^*}}{\sigma^2_{X^*} + \sigma^2_{\Delta x}}(\beta_1 \sigma^2_{\Delta x} - \rho)}_{\gamma_0} + x_i \underbrace{\frac{\beta_1 \sigma^2_{X^*} + \rho}{\sigma^2_{X^*} + \sigma^2_{\Delta x}}}_{\gamma_1} \end{aligned} \tag{3.20}$$

and variance

$$\sigma^2_\xi = \sigma^2_\epsilon + \beta_1^2 \sigma^2_{X^*} - \frac{(\beta_1 \sigma^2_{X^*} + \rho)^2}{\sigma^2_{X^*} + \sigma^2_{\Delta x}} \tag{3.21}$$

This leads to the model

$$Y_i | x_i = \gamma_0 + \gamma_1 x_i + \xi_i \tag{3.22}$$

where the error term ξ_i is normally distributed with mean 0 and variance σ^2_ξ, and γ_0 and γ_1 are bias expressions for the naive estimators in that

$$E(\hat{\beta}_{0naive}) = \gamma_0, \quad E(\hat{\beta}_{1naive}) = \gamma_1 \quad \text{and} \quad E(\hat{\sigma}^2_{naive}) = \sigma^2_\delta \tag{3.23}$$

under the given assumptions (Buonaccorsi 2010, chapter 4).

Special case: no error in the response or uncorrelated measurement errors

We assume that there is no correlation between measurement error Δx_i and Δy_i, so that $\rho = 0$. Then the naive slope becomes

$$\gamma_1 = \left(\frac{\sigma^2_{X^*}}{\sigma^2_{X^*} + \sigma^2_{\Delta x}}\right)\beta_1 = \kappa \beta_1 \tag{3.24}$$

where κ is called the *reliability ratio* (Buonaccorsi 2010). Since $\kappa < 1$, we see that the naive simple linear regression of Y on X gives an estimate which is attenuated, meaning biased, towards zero with κ defining the degree of attenuation. The larger the measurement error, meaning the larger its variance $\sigma^2_{\Delta x}$, the stronger the attenuation of the regression slope (Buonaccorsi 2010, chapter 4).

One expects that because X is an error-prone predictor, it has a weaker relationship with the response than x^*. This can be seen by the attenuation as well as by the residual variance of this regression of Y on X which is

$$\sigma^2_\xi = \sigma^2_\epsilon + \beta_1^2\sigma^2_{X^*} - \frac{(\beta_1\sigma^2_{X^*})^2}{\sigma^2_{X^*} + \sigma^2_{\Delta x}} = \sigma^2_\epsilon + \frac{\beta_1^2\sigma^4_{X^*} + \beta_1^2\sigma^2_{X^*}\sigma^2_{\Delta x} - \beta_1^2\sigma^4_{X^*}}{\sigma^2_{X^*} + \sigma^2_{\Delta x}} =$$
$$= \sigma^2_\epsilon + \frac{\beta_1^2\sigma^2_{X^*}\sigma^2_{\Delta x}}{\sigma^2_{X^*} + \sigma^2_{\Delta x}} = \sigma^2_\epsilon + \kappa\beta_1^2\sigma^2_{\Delta x} > \sigma^2_\epsilon.$$

It is not surprising to see that measurement error, as an additional source of error, increases the variability about the line and makes the data more noisy.

Illustration of the bias

We illustrate the bias in the coefficients by looking at a simple model similar to the example presented by Carroll, Ruppert et al. 2006 in chapter 3. We consider a simple regression model

$$Y^*|x^* = x^* + \epsilon \tag{3.25}$$

where the true values x^* are standard normally distributed with $\mu_x = 0$, $\sigma^2_x = 1$, and the error term ϵ is independent of x^* and has mean zero and variance $\sigma^2_\epsilon = 1$. We assume that there is only measurement error in the single covariate x^* and it has mean zero and measurement error variance $\sigma^2_{\Delta x} = 0.25$.

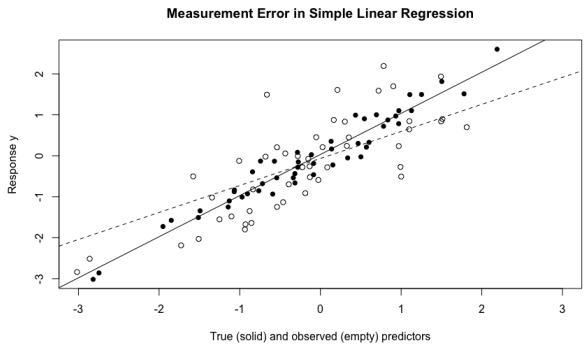


Figure 3.1: Illustration of bias in simple linear regression with measurement error in x

In the Figure 3.1 we plotted both true and error-prone datasets together. The solid circles and solid line display the true data (y^*, x^*) and their ordinary

regression line while the empty circles and the dashed line display the error-contaminated data $(y^*, x^* + \Delta x)$ and their regression line. We can clearly see the slope attenuation and the bias in the regression line due to the additive measurement error in the predictor since the least squares regression of y^* on $x^* + \Delta x$ gives an estimate that is attenuated to zero. We can also see that the error-prone data has much more variability about the line than the true data due to the additional variance from measurement error. This shows that the error-prone predictor has a weaker relationship with the response than the true predictor (Carroll, Ruppert et al. 2006). As an effect of measurement error, the slope is attenuated and the data are more noisy.

3.6 Correcting for measurement error

In order to correct for measurement error, additional data or knowledge of some of the measurement error variances or their reliability ratio is required. We have shown in Equation 3.24 that the expected value of the naive estimator of the slope β_1 in the special case of uncorrelated measurement errors is $\gamma_1 = \kappa\beta_1$, that is the true β_1 multiplied by the reliability ratio κ.
If the reliability ratio κ is known or can be estimated, then one will obtain an unbiased estimate of β_1 simply by dividing it by κ. The corrected estimator of the slope becomes

$$\hat{\beta}_{MM} = \hat{\kappa}^{-1}\hat{\beta}_{1naive} \tag{3.26}$$

where $\hat{\kappa}$ is the estimated reliability ratio (Buonaccorsi 2010, chapter 4). It is also called the method of moments estimator and it requires knowledge or at least estimability of the measurement error variance in x, $\sigma^2_{\Delta x}$ (Carroll, Ruppert et al. 2006, chapter 3).

The price for reducing bias is an increased variance. This is commonly referred to as *bias-variance-tradeoff*. Let us assume that the naive slope estimate $\hat{\beta}_{1naive}$ from ordinary least squares regression of the observed variables has mean

$$E(\hat{\beta}_{1naive}) = \kappa\beta_1 \tag{3.27}$$

for a known reliability ratio κ, and variance

$$Var(\hat{\beta}_{1naive}) = \sigma^2_{\beta_{1naive}}. \tag{3.28}$$

The method of moments estimator of β_1 which corrects for attenuation bias is given as defined above by $\hat{\beta}_{MM} = \kappa^{-1}\hat{\beta}_{1naive}$ with mean

$$E(\hat{\beta}_{MM}) = \beta_1 \tag{3.29}$$

and variance

$$Var(\hat{\beta}_{MM}) = \kappa^{-2}\sigma^2_{\beta_{1naive}}. \tag{3.30}$$

Hence we see that the bias was reduced to zero but the variance increased relative to the variance of the uncorrected estimate (Carroll, Ruppert et al. 2006, chapter 3).

Alternative methods for correcting bias are regression calibration or SIMEX, which is short for simulation extrapolation. Regression calibration replaces

the true x^* by approximation of the regression of x^* on the observed x, which requires accurate estimates of x^* (Carroll, Ruppert et al. 2006, chapter 4). The basic idea of the SIMEX method is to add simulated measurement error with increasing variance to the original data, run the statistical models with these increasingly error-prone data, identify a trend of the model parameter estimates versus the variance of the added measurement error, and extrapolate the trend back to the point with no measurement error (Shang 2012). This approach is computationally intensive as it is based on simulations but it can be applied to measurement error models of various forms (Shang 2012).
Another method for correcting bias due to measurement error is orthogonal regression, also known as total least squares, which we will present in the next chapter.

CHAPTER 4

Total Least Squares

4.1 Overview

This chapter introduces several methods proposed by Gregory L. Plett specifically for estimation of battery cell total capacity, from traditional weighted total least squares to approximate weighted total least squares (Plett 2011). Total least squares is a generalization of the least squares approach in ordinary regression which allows for measurement error in both explanatory and response variables and is hence well suited in situations when the data are corrupted by noise which is very often the case in engineering (Van Huffel and Lemmerling 2002, chapter 1).

4.2 Total least squares

Total least squares is a modelling technique in which measurement errors in both dependent and independent variables are taken into account. It is closely related to the concept of measurement error modelling discussed in the previous chapter, which is also known as error-in-variables regression in the field of statistics. While measurement error modelling provides statistical analysis, total least squares is more application-oriented and it is widely used in engineering fields like signal processing, system identification where the data modification idea is explained from a geometric point of view independent from its statistical interpretation (Van Huffel and Lemmerling 2002, chapter 1). In the following section, we will present a formal definition of the total least squares method which is the basis for Plett's methods.

Generally speaking, total least squares is a numerical tool for finding an approximate solution to an overdetermined system of equations $\mathbf{X}\beta \approx \mathbf{y}$ where both the vector $\mathbf{y}$ and the matrix $\mathbf{X}$ are assumed to be measured with error (Markovsky and Van Huffel 2007).

Assume we have a multivariate model described by the linear equation

$$\mathbf{x}_1^*\beta_1 + ... + \mathbf{x}_p^*\beta_p = \mathbf{y}^* \tag{4.1}$$

where the goal is to estimate the p-dimensional vector of unknown parameters $\beta = [\beta_1, ...\beta_p]^T$. When we measure n observations of all the variables such that

$n > p$, then this model gives rise to an *overdetermined* set of n linear equations

$$\mathbf{X}^*\beta \approx \mathbf{y}^* \tag{4.2}$$

where $\mathbf{X}^*$ is an $n \times p$ data matrix and $\mathbf{y}^*$ is an n-dimensional vector (Pešta 2018). It is called overdetermined because the number of equations exceeds the number of unknown parameters. These systems are in most of total least squares literature on purpose not formulated as an equation because in many cases the exact solution does not exist. Therefore only an approximation can be found and the solution yields the "best fit" of the overdetermined system (Pešta 2018).

Ordinary east squares is the traditional approach to approximating such an overdetermined system. It assumes that the data matrix X^* is measured exactly while the vector $\mathbf{y}^*$ contains measurement errors $\Delta\mathbf{y}$, so instead of $\mathbf{y}^*$ we observe the sum $\mathbf{y} = \mathbf{y}^* + \Delta\mathbf{y}$.
The misfit in the dependent variable is minimized by

$$\min_{\beta\in\mathbb{R}^p, \Delta\mathbf{y}\in\mathbb{R}^n} \|\Delta\mathbf{y}\|_2 \quad \text{such that} \quad \mathbf{y}^* + \boldsymbol{\Delta}\mathbf{y} = \mathbf{X}^*\beta. \tag{4.3}$$

In the total least squares approach, we assume that both the explanatory and response variables are observed with measurement errors $\Delta\mathbf{X}$ and $\Delta\mathbf{y}$ such that

$$\mathbf{X} = \mathbf{X}^* + \Delta\mathbf{X} \quad \text{and} \quad \mathbf{y} = \mathbf{y}^* + \Delta\mathbf{y}. \tag{4.4}$$

Since instead of the true $\mathbf{X}^*$ and $\mathbf{y}^*$, we observe the error-prone $\mathbf{X}$ and $\mathbf{y}$, we can rewrite the model as

$$\mathbf{y}^* + \Delta\mathbf{y} = [\mathbf{X}^* + \Delta\mathbf{X}]\beta. \tag{4.5}$$

Total least squares seeks to minimize the squares of errors in both the dependent and independent variables by

$$\min_{[\Delta\mathbf{X}\Delta\mathbf{y}]\in\mathbb{R}^{n\times(p+1)}, \beta\in\mathbb{R}^p} \|[\Delta\mathbf{X}\Delta\mathbf{y}]\|_F \tag{4.6}$$

such that they satisfy the Equation (4.5) above (Pešta 2018).
The $\mathbf{F}$ stands for Frobenius norm which is commonly used to minimize the measurement errors to construct the estimators. From a geometric point of view, it tries to minimize the orthogonal distance between the observations and a fitted hyperplane (Pešta 2013). It is defined for an $n \times m$ matrix $\mathbf{A}$ as

$$\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}^2}. \tag{4.7}$$

After a minimizing $[\Delta\hat{\mathbf{X}}\Delta\hat{\mathbf{y}}]$ has been found, any β satisfying the model (4.5) is a *TLS* solution (Pešta 2018). A closed-form solution has been derived to be

$$\hat{\beta} = ((\mathbf{X}^*)^\top\mathbf{X}^* - \sigma_{p+1}^2\mathbf{I}_p)^{-1}(\mathbf{X}^*)^\top\mathbf{y}^* \tag{4.8}$$

where σ_{p+1}^2 is the smallest singular value of $[\mathbf{X}^*\mathbf{y}^*]$ (Markovsky and Van Huffel 2007).
Comparing this expression to the well-known ordinary least squares solution

$$\tilde{\beta} = ((\mathbf{X}^*)^\top\mathbf{X}^*)^{-1}(\mathbf{X}^*)^\top\mathbf{y}^*, \tag{4.9}$$

we see they are similar except for the term containing σ_{p+1}^2. In the presence of independently and identically distributed equally sized measurement errors, TLS removes the bias in the OLS estimator due to measurement error by subtracting the error covariance matrix estimated by $\sigma_{p+1}^2 \mathbf{I}_p$ from the data covariance matrix $(\mathbf{X}^*)^\top \mathbf{X}^*$ (Markovsky and Van Huffel 2007).

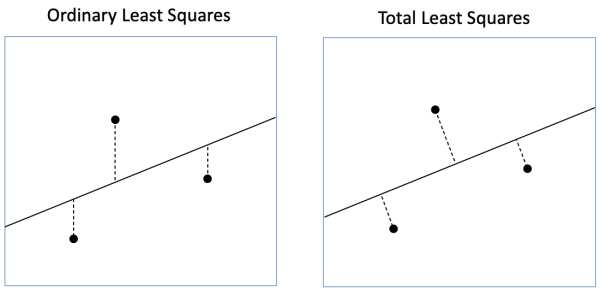


Figure 4.1: Illustration of ordinary and total least squares methods

A graphical illustration of ordinary and total least squares can be seen in Figure 4.1. As we can see in the OLS approach on the left side of the figure, the independent variable is assumed to be measured without error, so a residual represents the vertical distance between the observed point and the fitted regression line. TLS on the other hand accounts for errors in observations on both the x-axis and the y-axis. Hence a residual represents the shortest distance between the data point and the fitted line, that is, the residual is perpendicular to the line (Markovsky and Van Huffel 2007). For this reason TLS is often called *orthogonal regression*.

Total least squares can be applied to both linear and non-linear models and allows for any number of predictors and complicated error structures. It has been extended to solve weighted, structured and regularized total least squares problems (Markovsky and Van Huffel 2007). *Weighted total least squares* which we will investigate in more detail later, should be considered when the measurement errors are independent and have unequally sized variances. They use appropriate scaling matrices in order to maintain consistency (Markovsky and Van Huffel 2007). A special case of TLS for a two-dimensional dataset (y_i, x_i) and independent errors is called *Deming regression* (Cornbleet and Gochman 1979).

4.3 Total least squares for battery capacity estimation

Throughout this chapter we will seek to estimate the total cell capacity Q by using the linear relationship between the explanatory variables x_i, defined as the estimated change in SOC over a time interval $i = [t_1, t_2]$

$$x_i = SOC(t_2) - SOC(t_1), \tag{4.10}$$

and response y_i, defined as the accumulated ampere hours passing through the cell over the same time interval,

$$y_i = \int_{t_1}^{t_2} \frac{\eta i(\tau)}{3600}\, d\tau, \tag{4.11}$$

as mentioned in Equation 2.8. Plett (2011) assumes the Coulomb efficiency factor η to be approximately 1 and hence does not take its effect into consideration. We will also assume $\eta = 1$ in our analysis but will later discuss how this factor may be implemented to give more precise results.

We assume that both dependent and independent variables are measured with error such that instead of the true values x_i^* and y_i^* we observe the sums

$$x_i = x_i^* + \Delta x_i \quad \text{and} \quad y_i = y_i^* + \Delta y_i. \tag{4.12}$$

The goal is to find the coefficient Q in

$$\mathbf{y}^* = Q\mathbf{x}^* \tag{4.13}$$

which can be written as

$$(\mathbf{y} - \Delta\mathbf{y}) = Q(\mathbf{x} - \Delta\mathbf{x}) \tag{4.14}$$

where the n-dimensional vectors of measurement errors $\Delta\mathbf{y}$ and $\Delta\mathbf{x}$ consist of zero-mean Gaussian random variables with known variances $\sigma^2_{\Delta y_i}$ and $\sigma^2_{\Delta x_i}$ (Plett 2011). The derivation of the linear relationship in Equation 4.13 was presented in chapter 2 based on the Coulomb counting formula. Note that Equation 4.13 states that if all variables are measured without any error, then they are exactly linearly related, hence there is no error in equation. This is typically considered to be the case in physics, where the variables are related by certain physical laws. Plett assumes for his models that the total capacity Q is an electrochemical property of the battery cell that is independent from both temperature and C-rate (Plett 2011). This is a necessary assumption when the model is defined without any error in equation. In practice, only in rare circumstances does the real data fall exactly on the straight line in absence of measurement error, and there is always equation error present (Carroll and Ruppert 1996).

Plett (2011) formulates this as an optimization problem which seeks to minimize the so-called *merit function*, denoted by χ^2. The merit function is a loss function commonly used in engineering and it measures the agreement between data and the fitting model for a particular choice of parameters (Press 1992). Hence smaller merit function of a model means better fit for the data. The parameters are usually adjusted based on the value of the merit function until a smallest value is obtained, producing the "best fit" for the data.
Using the total least squares approach, Plett (2011) defines the merit function as the weighted sum of squared errors Δx_i and Δy_i, therefore

$$\chi^2 = \sum_{i=1}^{n} \frac{(\Delta x)^2}{\sigma^2_{\Delta x_i}} + \frac{(\Delta y_i)^2}{\sigma^2_{\Delta y_i}}. \tag{4.15}$$

In the following sections we will present three possible approaches suggested by Plett on how to minimize this merit function. First, he applies numerical methods to iteratively find the optimal estimate $\hat{Q}$ since the merit function in the form above does not have a closed-form analytical solution. Second, he simplifies the merit function by assuming that the measurement error variances $\sigma^2_{\Delta x_i}$ and $\sigma^2_{\Delta x_i}$ are proportional. Third, he derives an approximate solution to this total least squares problem motivated by geometric relationships between the uncertainties. We will refer to these methods by using Plett's acronyms, *WTLS* for numerical weighted total least squares, *TLS* for total least squares with proportional uncertainties and *AWTLS* for approximate weighted total least squares.
In addition, Plett adapts two of these methods, *TLS* and *AWTLS*, to include a *fading memory* update by modifying the merit function with a forgetting factor in such a way that it puts more emphasis on more recent rather than earlier measurements. Furthermore, *TLS* and *AWTLS* approaches make recursive implementation possible which may reduce storage and computational costs when the number of measurements n becomes large. Without a recursive update, the entire vectors $\mathbf{x}$ and $\mathbf{y}$ must be stored and the growing number of computations becomes costly, making this method not well suited for real-time application with limited storage and computational capabilities (Plett 2011).

Weighted total least squares - WTLS

We find an estimate of the cell total capacity $\hat{Q}$ by minimizing the weighted sum of squared errors Δx_i plus the weighted sum of squared errors Δy_i, that is we want to find a coefficient $\hat{Q}$ that minimizes the following merit function, written in our notation as

$$\chi^2_{WTLS} = \sum_{i=1}^{n} \frac{(\Delta x)^2}{\sigma^2_{\Delta x_i}} + \frac{(\Delta y_i)^2}{\sigma^2_{\Delta y_i}} = \sum_{i=1}^{n} \frac{(x_i - x_i^*)^2}{\sigma^2_{\Delta x_i}} + \frac{(y_i - y_i^*)^2}{\sigma^2_{\Delta y_i}}, \tag{4.16}$$

where x_i^* and y_i^* are the unknown true values, and x_i and y_i are the noisy measured data points (Plett 2011).

Plett's approach is to use a Lagrange multiplier λ_i to augment the merit function with the constraint that $y_i^* = \hat{Q}x_i^*$ which gives

$$\chi^2_{WTLS,\lambda} = \sum_{i=1}^{n} \frac{(x_i - x_i^*)^2}{\sigma^2_{\Delta x_i}} + \frac{(y_i - y_i^*)^2}{\sigma^2_{\Delta y_i}} - \lambda_i(y_i^* - \hat{Q}x_i^*) \tag{4.17}$$

By setting the partial derivatives of this augmented merit function equal to zero

$$\frac{\partial \chi^2_{WTLS,\lambda}}{\partial x_i^*} = \frac{\partial \chi^2_{WTLS,\lambda}}{\partial y_i^*} = \frac{\partial \chi^2_{WTLS,\lambda}}{\partial \lambda_i} = 0, \tag{4.18}$$

Plett obtains the equations

$$x_i^* = \frac{x_i\sigma^2_{\Delta y_i} + \hat{Q}y_i\sigma^2_{\Delta x_i}}{\sigma^2_{\Delta y_i} + \hat{Q}^2\sigma^2_{\Delta x_i}} \quad \text{and} \quad y_i^* = \hat{Q}x_i^*, \tag{4.19}$$

and consequently, the merit function in Equation 4.16 can be written as

$$\chi^2_{WTLS} = \sum_{i=1}^{n} \frac{(y_i - \hat{Q}x_i)^2}{\hat{Q}^2\sigma^2_{\Delta x_i} + \sigma^2_{\Delta y_i}} \tag{4.20}$$

We minimize the merit function by setting its partial derivative

$$\frac{\partial \chi^2_{WTLS}}{\partial \hat{Q}} = \sum_{i=1}^{n} \frac{2(\hat{Q}x_i - y_i)(\hat{Q}y_i\sigma^2_{\Delta x_i} + x_i\sigma^2_{\Delta x_i})}{(\hat{Q}^2\sigma^2_{\Delta x_i} + \sigma^2_{\Delta y_i})^2} \tag{4.21}$$

equal to 0. Unfortunately deriving with respect to $\hat{Q}$ does not give an expression that would offer a closed-form solution and therefore a numerical method has to be applied (Plett 2011).
Plett (2011) suggests to use the *Newton-Raphson search* to find the $\hat{Q}$. It is a numerical algorithm which iteratively approximates the roots of a differentiable function f, meaning the solutions to $f(x) = 0$. When this method is applied to the derivative $f^{'}$ of a twice-differentiable function f, it becomes an optimization problem which numerically finds the solutions to $f'(x) = 0$, hence the minima or maxima of the function f. Starting with an initial guess x_0, this optimization method approximates at each step the function f by a second-order Taylor expansion around the current value x_t

$$f(x) \approx f(x_t) + (x - x_t)f'(x_t) + \frac{1}{2}(x - x_t)^2 f''(x_t) \tag{4.22}$$

and constructs a sequence of updates

$$x_{t+1} = x_t + \frac{f'(x_t)}{f''(x_t)}. \tag{4.23}$$

As the number of iterations goes towards infinity, the sequence should converge to a stationary point of the function f.

In our case, this sequence is constructed by iterating the equation

$$\hat{Q}_{t+1} = \hat{Q}_t - \frac{\frac{\partial \chi^2_{WTLS}}{\partial \hat{Q}}}{\frac{\partial^2 \chi^2_{WTLS}}{\partial \hat{Q}^2}} \tag{4.24}$$

where the numerator is the Jacobian and the denominator the Hessian of the merit function (Plett 2011). The iterations stop when a criterion is met, such as the absolute convergence criterion, when the absolute difference between the previous and current approximations is less than a predefined threshold, $|\hat{Q}_{t+1} - \hat{Q}_t| < \epsilon$. The Newton-Raphson search can for example be initialized with the ordinary least squares estimate of Q. Since the merit function is convex, this method is guaranteed to converge to the global minimum (Plett 2011).

Since the Newton-Raphson search has to be performed on the entire data every time new measurements are added to the vectors $\mathbf{x}$ and $\mathbf{y}$, this method does not allow a recursive update, which can have significant storage and computational implications.
Plett adds the fading memory update by reformulating the merit function as

$$\chi^2_{FMWTLS} = \sum_{i=1}^{n} \gamma^{n-i} \frac{(y_i - \hat{Q}x_i)^2}{\hat{Q}^2\sigma^2_{\Delta x_i} + \sigma^2_{\Delta y_i}}$$

where γ is the forgetting factor in the range $0 \ll \gamma \leq 1$. Hence early measurements will contribute less to the merit function compared to more recent ones.

Since the weighted total least squares method does not provide a recursive solution, its application in real-time has limited storage capabilities. By imposing additional assumptions on the form of the uncertainties in $\mathbf{x}$ and $\mathbf{y}$, Plett derives a simplified method to estimate battery cell total capacity which we will present in the next section.

Simplified method with proportional uncertainties - TLS

We assume that the uncertainties on x_i and y_i are *proportional*, meaning that there exists a constant k such that

$$\sigma_{\Delta x_i} = k\sigma_{\Delta y_i} \tag{4.25}$$

where $\sigma_{\Delta x_i}$ and $\sigma_{\Delta y_i}$ denote the standard deviations of the random measurement errors Δx_i and Δy_i for $i = 1, ...n$ (Plett 2011). Therefore in this case, the uncertainties must be proportionally related by a scaling factor k and cannot be chosen arbitrarily.
The *WTLS* merit function in Equation 4.16 simplifies, as Plett (2011) writes, to

$$\chi^2_{TLS} = \sum_{i=1}^{n} \frac{(x_i - x_i^*)^2}{k^2\sigma^2_{\Delta y_i}} + \frac{(y_i - y_i^*)^2}{\sigma^2_{\Delta y_i}} = \sum_{i=1}^{n} \frac{(y_i - \hat{Q}x_i)^2}{(\hat{Q}^2k^2 + 1)\sigma^2_{\Delta y_i}} \tag{4.26}$$

This merit function can again be minimized by setting the partial derivative equal to zero, and consequently, Plett (2011) finds the exact solution to $\hat{Q}$ to be one of the roots

$$\hat{Q} = \frac{-\left(\sum_{i=1}^{n}(x_i^2 - k^2y_i^2)/(\sigma^2_{\Delta y_i})\right) \pm a}{2\sum_{i=1}^{n} k^2(x_iy_i)/\sigma^2_{\Delta y_i}}$$
$$\text{with} \quad a = \sqrt{\left(\sum_{i=1}^{n}(x_i^2 - k^2y_i^2)/(\sigma^2_{\Delta y_i})\right)^2 + 4k^2\left(\sum_{i=1}^{n}(x_iy_i)/(\sigma^2_{\delta y_i})\right)^2}. \tag{4.27}$$

Plett shows that this quadratic equation always has one positive root and one negative root and argues that due to the form of the Routh array, the larger, thus the positive root is the correct solution to $\hat{Q}$ (Plett 2011).

This method can be implemented in a recursive way by defining running sums which get updated whenever new measurements are obtained. Fading memory can also be added to the merit function with the forgetting factor γ in the same way as previously.

This method offers a closed-form solution without any numerical computations. However, it puts a constraint on the relationship between the uncertainties in x and y. In the next section, we will introduce an approximation to total least squares which both gives an exact analytical solution and allows for arbitrary uncertainties.

Approximate weighted total least squares - AWTLS

Figure 4.2 shows the geometric relationship between the noisy measurement point (x_i, y_i) and its true value (x_i^*, y_i^*) on the line $y^* = \hat{Q}x^*$ which motivates this approximate solution derived by Plett (2011). We force the line connecting the points (x_i, y_i) and (x_i^*, y_i^*), denoted by R_i, to be perpendicular to the line $y^* = \hat{Q}x^*$, as in TLS (Plett 2011).

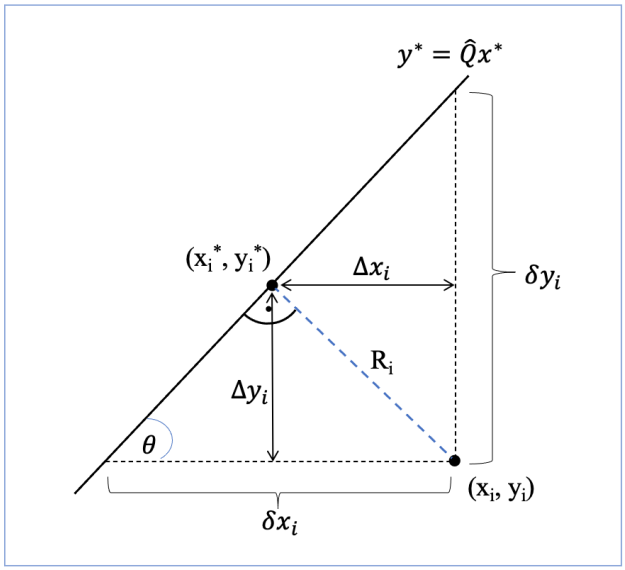


Figure 4.2: Illustration of derivation of approximate WTLS

We define δx_i to be the distance along the x-axis, and δy_i to be the distance along the y-axis between a data point $i = (x_i, y_i)$ and the line, for all $i = 1, ...n$. Then the slope of the line is

$$\hat{Q} = \frac{\delta y_i}{\delta x_i} \quad \forall i = 1, ...n. \tag{4.28}$$

The angle of the line is then

$$\theta = \tan^{-1} \hat{Q}. \tag{4.29}$$

The shortest distance between the line and a given data point (x_i, y_i) can be computed as Plett (2011) writes, in our notation, by

$$R_i = \delta y_i \cos\theta = \delta y_i / \sqrt{1 + \tan^2\theta} = \delta y_i / \sqrt{1 + \hat{Q}^2}. \tag{4.30}$$

Let measurement errors in x_i and y_i be denoted as before by Δx_i and Δy_i. From Figure 4.2 we see that they are the x- and y-components of the perpendicular distance between the noisy measurement points and the true values on the line (Plett 2011). Hence they can be defined by

$$\Delta x_i = R_i \sin\theta \quad \text{and} \quad \Delta y_i = R_i \cos\theta. \tag{4.31}$$

Using that

$$\sin^2\theta = 1 - \cos^2\theta = \frac{\hat{Q}^2}{1+\hat{Q}^2}, \tag{4.32}$$

we can write

$$\Delta x_i^2 = \left(\frac{\delta y_i^2}{1+\hat{Q}^2}\right)\left(\frac{\hat{Q}^2}{1+\hat{Q}^2}\right) \tag{4.33}$$

and

$$\Delta y_i^2 = \left(\frac{\delta y_i^2}{1+\hat{Q}^2}\right)\left(\frac{1}{1+\hat{Q}^2}\right). \tag{4.34}$$

In addition, it holds that $\delta y_i = y_i - \hat{Q}x_i$ (Plett 2011). By combining all these equations above, the initial merit function in Equation 4.16 can be rewritten, according to Plett (2011), as

$$\chi^2_{AWTLS} = \sum_{i=1}^{n} \frac{(\Delta x)^2}{\sigma^2_{\Delta x_i}} + \frac{(\Delta y_i)^2}{\sigma^2_{\Delta y_i}} = \sum_{i=1}^{n} \frac{(y_i - \hat{Q}x_i)^2}{(1+\hat{Q}^2)^2}\left(\frac{\hat{Q}^2}{\sigma^2_{\Delta x_i}} + \frac{1}{\sigma^2_{\Delta y_i}}\right). \tag{4.35}$$

The partial derivative with respect to $\hat{Q}$ becomes a quartic equation whose roots are candidate solutions for $\hat{Q}$. Plett proposes to use the Ferari method to find the four roots and selects the optimum to be the one that gives the lowest value of the merit function (Plett 2011).

This method allows a simplification of the merit function using trigonometric identities and it has the same advantageous properties as TLS, such as a closed-form solution, a recursive implementation and a fading memory update. Most importantly, it allows for individual weighting on the data points.

CHAPTER 5

Model for measurement uncertainties

5.1 Overview

In this chapter, we present one possible way of modelling the measurement uncertainties in x and y applicable to our data. For the variable x we are able to derive bounds for the measurement error variance given information from our data provider. For the response variable y we differentiate between regular and irregular sampling intervals, and while the uncertainty in y is more complicated to quantify, we are able to come to the conclusion that the measurement error variance increases with the length of the time interval we integrate over. We will show that our conclusion holds by creating simulations of current.

5.2 Error Quantification and Propagation

We obtained a large sample of observations of current and state of charge which were all subject to error or uncertainty. However, the response and predictor variables in our linear model are not these measurements directly, but functions of them. When we want to determine a variable that is a function of one or more different measured variables, we must carry over the uncertainties in these individual measurements to determine the uncertainty in the target variable. This is called *propagation of errors*. (Bevington and Robinson 1992, chapter 3). Suppose we have two measured variables, u and v, which have variances σ_u^2, σ_v^2 and covariance σ_{uv}. We can express the variance of a dependent variable $x = f(u, v)$ for a function f in terms of the variances of u and v according to Bevington and Robinson (1992) by

$$\sigma_x^2 = \sigma_u^2 \left(\frac{\partial x}{\partial u}\right)^2 + \sigma_v^2 \left(\frac{\partial x}{\partial v}\right)^2 + 2\sigma_{uv} \left(\frac{\partial x}{\partial u}\right)\left(\frac{\partial x}{\partial v}\right) \tag{5.1}$$

If x is the weighted sum of u and v, defined as $x = au + bv$ for constants a, b, then we obtain the following uncertainty in x

$$\sigma_x^2 = a^2\sigma_u^2 + b^2\sigma_v^2 + 2ab\sigma_{uv}. \tag{5.2}$$

The expression (5.1) is also known as the *error propagation equation* and we will apply it in the next sections to approximate uncertainties in our x and y variables.

5.3 Measurement uncertainty in x

Assume that we have a set of n observations of state of charge, SOC, at sampling times $t = \{t_1, ...t_n\}$ and let SOC_k denote the observed state of charge at time t_k for all $k = 1, ...n$.
We previously defined the predictor x as the difference between two SOC values in a time interval. Assume that we are unable to observe the true SOC measurement but rather observe the sum

$$SOC_k = SOC_k^* + \delta_k \quad \forall k = 1, ...n \tag{5.3}$$

with SOC_k^* as the true state of charge at time t_k and δ_k as the measurement error. Here we are modelling the functional case, as defined in chapter 3, where the true values SOC_k^* are fixed and not random variables. We were informed by our data provider that δ_k is approximately a zero-mean Gaussian variable with standard deviation $\sigma_\delta = 2.5$, hence it has variance $\sigma_\delta^2 = 6.25$.
For a chosen time interval $i = [t_m, t_l]$ where $t_m, t_l \in t, 1 \leq m < l \leq n$, we then define our input variable as

$$x_i = SOC_l - SOC_m = (SOC_l^* + \delta_l) - (SOC_m^* + \delta_m) = \underbrace{(SOC_l^* - SOC_m^*)}_{x_i^*} + \underbrace{(\delta_l - \delta_m)}_{\Delta x_i} \tag{5.4}$$

where x_i^* is the difference between the true SOC values and Δx_i is the measurement error in x_i. Thus we see that the equation above has the same form as equation (3.14) in chapter 3.
Since δ_k as a zero-mean Gaussian variable has $E[\delta_k] = 0$, then

$$E(SOC_k) = SOC_k^* \quad \forall k \in \{1, ...n\}, \tag{5.5}$$

and

$$E[x_i] = SOC_l^* - SOC_m^* \quad \forall l, m \in \{1, ..n\}, 1 \leq m < l \leq n. \tag{5.6}$$

The variance of x_i is defined as

$$\sigma_{x_i}^2 = Var(x_i) = Var((SOC_l^* - SOC_m^*) + (\delta_l - \delta_m)) \tag{5.7}$$

$$= Var(\delta_l - \delta_m) = E[(\delta_l - \delta_m)^2] - E[(\delta_l - \delta_m)]^2 = E[(\delta_l - \delta_m)^2] \tag{5.8}$$

since SOC_k^* are fixed values for all $k \in \{1, ...n\}$, and therefore have variance 0.

The variance of measurement error Δx_i can be calculated by

$$\sigma_{\Delta x_i}^2 = E[(\delta_l - \delta_m)^2] - E[\delta_l - \delta_m]^2 = E[\delta_l^2 - 2\delta_l\delta_m + \delta_m^2] = E[\delta_l^2] - 2E[\delta_l\delta_m] + E[\delta_m^2] = Var(\delta_l) - 2E[\delta_l\delta_m] + Var(\delta_m) = 2\sigma_\delta^2 - 2E[\delta_l\delta_m] \tag{5.9}$$

which is in accordance with the *error propagation equation* in (5.1) by which

$$\sigma_{\Delta x_i}^2 = 2\sigma_\delta^2 - 2Cov(\delta_l, \delta_m) = 2\sigma_\delta^2 - 2E[\delta_l\delta_m] \tag{5.10}$$

If the individual measurement errors δ_k are independent for all $k \in \{1, ...n\}$, then $2E[\delta_l\delta_m] = 0 \quad \forall l, m \in \{1, ..n\}, 1 \leq m < l \leq n$ and the variance of Δx_i becomes the sum of the two variances

$$\sigma_{\Delta x_i}^2 = 2\sigma_\delta^2. \tag{5.11}$$

On the contrary, if they are fully dependent and the measurement error can be modelled as a constant bias over the time span where we are generating the x_i's, then $2E[\delta_l \delta_m] = 2\sigma_\delta^2$ and the measurement error variance becomes

$$\sigma^2_{\Delta x_i} = 0. \tag{5.12}$$

While there may be some dependency which we are unable to quantify between SOC values measured close in time to each other, it is reasonable to assume that the measurements become independent as the time interval between them increases and the variance of their difference reaches the upper bound of $2\sigma_\delta^2$ in (5.11). We conclude that the actual variance of the difference between any two SOC measurements is somewhere in the range from 0 to $2\sigma_\delta^2$, therefore $\sigma^2_{\Delta x_i} \in [0, 12.5]$.

5.4 Measurement uncertainty in y

The response variable y is generated by integrating the current signal $I(t)$ over a given time interval. Assume that we have a set of n observations of current at sampling times $t = \{t_1, ...t_n\}$ and let I_k denote the current value at time t_k for all $k = 1, ...n$.
For a chosen time interval $i = [t_m, t_l]$ where $t_m, t_l \in t, 1 \leq m < l \leq n$, we then define our response variable as

$$y_i = \int_{t_m}^{t_l} I(\tau)\, d\tau \tag{5.13}$$

Since we only know the function $I(t)$ at isolated measurement points, we cannot determine the exact integral analytically but will have to use numerical integration methods to approximate the integral with our observations. We will use the simplest numerical integration method which calculates the area under a function by partitioning the region into rectangles and then adding all of their areas together. We will approximate the integral above using observed values of current, I_k, obtained at sampling times t_k within the integration interval i, such that $t_m \leq t_k \leq t_l$. The length of the integration interval is denoted as $\Delta_i = t_l - t_m$.

Again, we assume that the values of current we observe are not the true values but they are are measured with a corresponding additive measurement error. In the next sections, we will differentiate between two cases depending on the sampling technique.

Current measurements obtained at regular intervals

We consider the case where the integral is calculated from measurements obtained at regular intervals. Assume as before that we have a sequence of n measurements of current, denoted I_k, obtained at sampling times t_k, $k \in 1, ...n$. Since we assumed that the sampling times are regular, the time steps t_k form a uniform partition of the sampling interval $[t_1, t_n]$ and the step length between any two consecutive measurements is constant and denoted by $\theta = t_{k+1} - t_k = (t_n - t_1)/n \quad \forall k \in \{1, ...n-1\}$. Let N denote the number of measurements observed in the integration interval $i = [t_m, t_l]$, so that

$N = l - m + 1$. Note that the length of the integration interval can be written as $\Delta_i = (l - m) \cdot \theta = (N - 1) \cdot \theta$.
We are not able to observe the true current values I_k^* but rather noisy measurements which can be written as the sum

$$I_k = I_k^* + \xi_k \tag{5.14}$$

where I_k^* is the fixed true current value and ξ_k is the random measurement error. It is reasonable to assume that these measurement errors are independent zero-mean random variables with

$$E[\xi_k] = 0 \tag{5.15}$$

and

$$E[\xi_k \xi_l] = 0 \quad \forall k, l \in \{1, ...n\}, k \neq l. \tag{5.16}$$

The variance of the measurement errors is constant and it is denoted by σ_ξ^2.

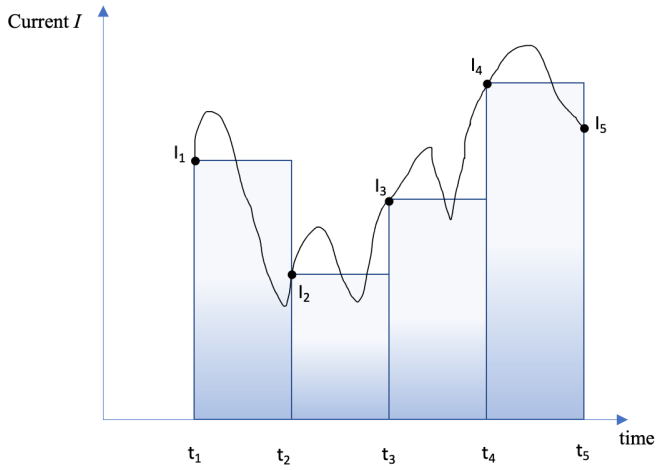


Figure 5.1: Illustration of regular current measurements with the corresponding numerical integral

The integral y_i can be estimated using a first order numerical approximation

$$y_i = \int_{t_m}^{t_l} I(\tau)\, d\tau \approx \sum_{k=m}^{l-1} I_k \cdot \theta = \sum_{k=m}^{l-1} (I_k^* + \xi_k) \cdot \theta = y_i'. \tag{5.17}$$

Figure 5.1 shows regular current measurements $I_1, ...I_5$ and the blue filled areas of the rectangles illustrate the numerical integral as we calculate it. We can bring the expression above on the same form as (3.15) by writing

$$y_i' = \sum_{k=m}^{l-1} (I_k^* + \xi_k) \cdot \theta = \underbrace{\sum_{k=m}^{l-1} I_k^* \cdot \theta}_{y_i'^*} + \underbrace{\sum_{k=m}^{l-1} \xi_k \cdot \theta}_{\Delta y_i'} \tag{5.18}$$

where the first sum describes the numerical approximation of the true y_i^* and the second sum is the aggregated measurement error. We have that the expected value of y_i' is

$$E\left[\sum_{k=m}^{l-1}(I_k^*+\xi_k)\cdot\theta\right]=\sum_{k=m}^{l-1}E[(I_k^*+\xi_k)\cdot\theta]=\sum_{k=m}^{l-1}\theta(E[I_k^*]+E[\xi_k])=\sum_{k=m}^{l-1}\theta\cdot I_k^* \tag{5.19}$$

and the variance of y_i' is given by

$$\begin{aligned}\sigma_{y_i'}^2 &= Var(\sum_{k=m}^{l-1}I_k^*\cdot\theta+\sum_{k=m}^{l-1}\xi_k\cdot\theta)=Var(\sum_{k=m}^{l-1}\xi_k\cdot\theta)\\ &=\theta^2\sum_{k=m}^{l-1}Var(\xi_k)+2\theta^2\sum_{m\le i<j\le l-1}Cov(\xi_i,\xi_j)=\theta^2\sum_{k=m}^{l-1}Var(\xi_k)\\ &=(l-m)\cdot\theta^2\cdot\sigma_\xi^2=(N-1)\cdot\theta^2\cdot\sigma_\xi^2=\Delta_i\cdot\theta\cdot\sigma_\xi^2=\frac{\Delta_i^2}{N-1}\cdot\sigma_\xi^2\end{aligned} \tag{5.20}$$

The last two equalities hold since we have $\theta = \Delta_i/(N-1)$. From this expression we see that on one hand, for a fixed length of integration intervals, Δ_i, the variance $\sigma_{y_i'}^2$ of the numerical approximation decreases with a shorter sampling interval θ, hence with an increasing number of samples N. Therefore a higher frequency of samples in a given fixed integration interval Δ_i gives lower measurement uncertainty in the numerical approximation y_i'. On the other hand, if we want to analyze the effect of varying integration interval length Δ_i, we assume that the sampling interval θ is fixed, such that the number of samples per second is fixed. The variance $\sigma_{y_i'}^2$ which can be written as above $\Delta_i\cdot\theta\cdot\sigma_\xi^2$ is then a linear function of Δ_i and hence a linear function of the duration of the integration interval. We can conclude that when integrating current over time, the measurement uncertainty linearly increases with the integration interval length if the sampling interval length θ is fixed. If the number of sampling intervals is kept fixed at N, then the measurement uncertainty will increase quadratically with the sampling interval length θ according to Equation 5.20.

When estimating the uncertainty in the integral y_i, it is not sufficient to only consider the propagated measurement error in the finite sum approximation. The error introduced by the numerical integration also needs to be taken into account. Thus the total error consists of two components, firstly the measurement error, and secondly, the error in approximating the integral by a finite sum. We call this the *numerical integration error*. This error depends on the relationship between the dynamics and smoothness of the current as well as the length of the sampling interval, and decreases when the sampling intervals become shorter. If the sampling intervals were very small relative to the change in the signal within this interval length, with a sampling rate of a few milliseconds at most, one could assume that the numerical integration error is negligible compared to the measurement error and that our numerical approximation in (5.18) converges to the integral in (5.13). This is not the case in our data, and without a proper lab setup to accurately quantify the integration error, the numerical integration error remains unknown. We can conclude that when the measurements are obtained at regular intervals, the

measurement error variance is linear with integration time, and we will verify the results we derived above with simulations in the next subsection.

Simulations of regular current measurements

We want to show that with a constant sampling rate θ, the measurement error variance grows linearly with increasing integration interval length. We assume that the true current signal I_k^* has value zero over a time period. Then it follows that the integrated current y_i is also 0 over any integration interval i. We generate 10 000 noisy zero-mean measurements of current, denoted I_k, with variance $\sigma_\xi^2 = 1$ at regular timesteps t_k as can be seen in 5.2, so that the interval length between two measurements, θ, is constant and equal to 1 second. We calculate the noisy numerical integrals over for instance 1000 seconds each,

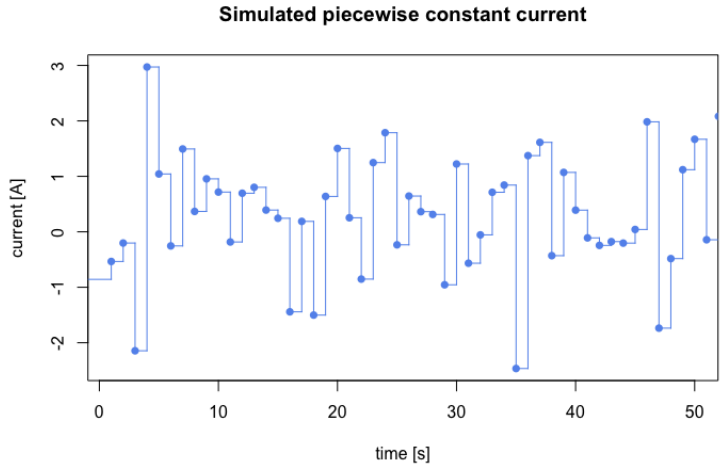


Figure 5.2: Illustration of piecewise constant current measurements at regular sampling frequency of 1 second

so that the integration interval length is fixed at $\Delta_i = 1000$, and compute the variance of these. By resampling 100 times and averaging over these variances, we will verify that our derivation of the measurement error variance $\sigma^2_{y_i'}$ in (5.20) is correct. Since true current I_k^* is 0, the numerical integral we compute is just the measurement error $\Delta y_i'$ in (5.18).

The results of our simulations can be seen in figure 5.3. The blue circles in the plot illustrate the measurement error variance computed numerically for varying integration lengths from 60 s to 1200s, meaning for integration lengths Δ_i in the range $[1, 20]$ min on the x-axis. The black line illustrates the analytically computed measurement error variances as in (5.20) and it corresponds well with the blue points, confirming our previous derivations. These simulations verified that the measurement error variance grows linearly with increasing Δ_i.

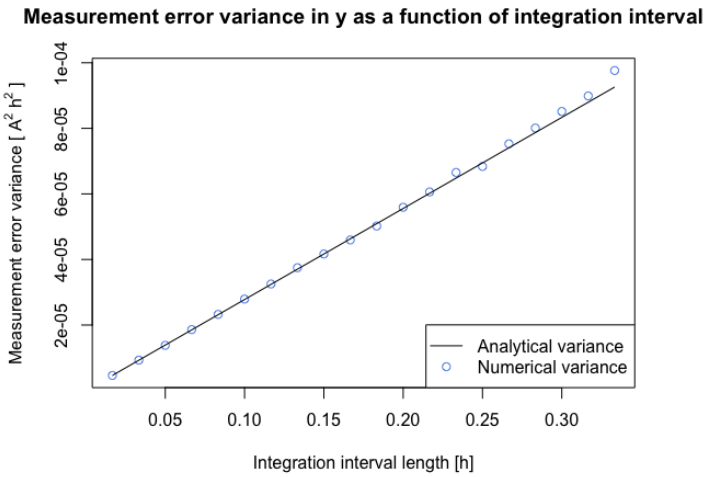


Figure 5.3: Illustration of measurement error variances computed analytically and numerically for varying integration interval lengths Δ_i at a constant sampling frequency θ

Current measurements obtained at irregular intervals

Now we consider the case when we have irregular sampling frequency, meaning that the current measurements are obtained at irregular intervals. The data appears to be stored based on both *asynchronous* and *synchronous sampling*. The former refers to a technique where samples are only stored when the current value exceeds a certain threshold T. The latter describes a technique where a new data point is stored at fixed time steps even when the value does not change from the previous data point. Therefore the current is sampled regularly with an underlying sampling frequency but no samples are stored unless the measured current has changed more than T since the previous stored value. This causes irregular sampling times.
Assume we have a sequence of n measurements of current, denoted I_k, obtained at sampling times t_k, $k \in \{1, ...n\}$. Further assume that a current measurement I_{k-1} is obtained at time t_{k-1}. We know that until the next current measurement is obtained at time t_k, the actual current signal is somewhere inside the band $I_{k-1} \pm T$ as can be seen in figure 5.4.

Using the same numerical integration method as before, we integrate the current between any two consecutive time steps t_{k-1} and t_k by computing the area of the rectangle

$$z_k = I_{k-1} \cdot (t_k - t_{k-1}) = I_{k-1} \cdot \theta_k \quad \forall k \in \{2, ...n\} \tag{5.21}$$

with initial value $z_1 = 0$ since at time t_1 the integral is 0. Here θ_k denotes the time interval length between two measurements I_k and I_{k-1}. The error due to numerical integration can for this single z_k be bounded by

$$e_k = T \cdot \theta_k \quad \forall k \in \{2, ..n\}. \tag{5.22}$$

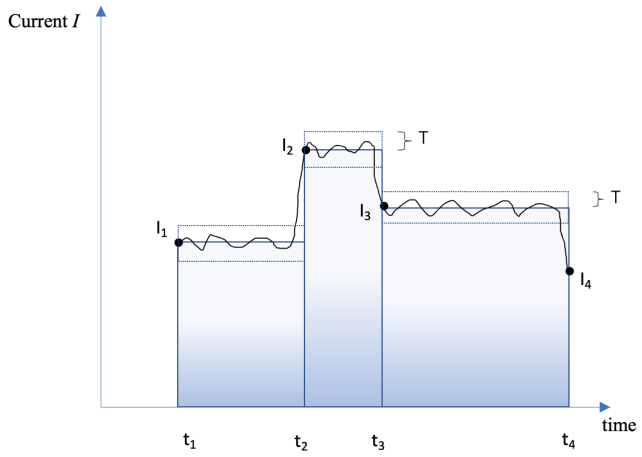


Figure 5.4: Illustration of irregular current measurements with threshold T and the corresponding numerical integral

This error is proportional to the duration of the interval θ_k and the threshold T.

When we calculate our response y_i, we integrate over many of these time intervals θ_k included in the integration interval $i = [t_m, t_l]$. Similarly as in the case with regular measurements, we define

$$y_i = \int_{t_m}^{t_l} I(\tau)\, d\tau \approx \sum_{k=m}^{l-1} z_{k+1} = \sum_{k=m}^{l-1} I_k \cdot \theta_{k+1} = \sum_{k=m}^{l-1} (I_k^* + \xi_k) \cdot \theta_{k+1} = y_i'. \quad (5.23)$$

which can be written as (3.15) by writing

$$y_i' = \sum_{k=m}^{l-1} (I_k^* + \xi_k) \cdot \theta_{k+1} = \underbrace{\sum_{k=m}^{l-1} I_k^* \cdot \theta_{k+1}}_{y_i'^*} + \underbrace{\sum_{k=m}^{l-1} \xi_k \cdot \theta_{k+1}}_{\Delta y_i'} \quad (5.24)$$

Assuming as before that the measurement errors ξ_k are independent mean-zero random variables with variance σ_ξ^2, we can derive the expected value of y_i'

$$\begin{aligned} E\left[\sum_{k=m}^{l-1} (I_k^* + \xi_k) \cdot \theta_{k+1}\right] &= \sum_{k=m}^{l-1} E[(I_k^* + \xi_k) \cdot \theta_{k+1}] = \sum_{k=m}^{l-1} \theta_{k+1}(E[I_k^*] + E[\xi_k]) \\ &= \sum_{k=m}^{l-1} \theta_{k+1} \cdot I_k^* \end{aligned} \quad (5.25)$$

and the variance by

$$\sigma_{y'}^2 = Var\left(\sum_{k=m}^{l-1} \xi_k \cdot \theta_{k+1}\right) = \sum_{k=m}^{l-1} \theta_{k+1}^2 Var(\xi_k) = \sum_{k=m}^{l-1} \theta_{k+1}^2 \sigma_\xi^2 = \sigma_\xi^2 \cdot \sum_{k=m}^{l-1} \theta_{k+1}^2 \tag{5.26}$$

due to the independence of ξ_k.

In addition to the propagated measurement errors in the finite sum, we need to take into account the numerical integration error. We can approximate its upper bound by adding e_k defined in (5.22) over the time stamps included in the integration interval by

$$\sum_{k=m}^{l-1} e_{k+1} = \sum_{k=m}^{l-1} T \cdot \theta_{k+1}. \tag{5.27}$$

Simulations of irregular current measurements

Here we will again only simulate measurement error, and not the additional numerical integration error due to the asynchronous sampling. We assume that the true current signal I_k^* is constant zero over a time period. Then it follows that the integrated current y_i is also 0 over any integration interval i. The current is measured at the start of each interval, with a measurement error of variance $\sigma_\xi^2 = 1$. We generate 10 000 noisy zero-mean measurements of current, denoted I_k, with variance $\sigma_\xi^2 = 1$ at irregular timesteps t_k as can be seen in 5.4, so that the interval lengths between two measurements, θ_k, are integers drawn uniformly from $\theta_k \in [1, 60]$ seconds.

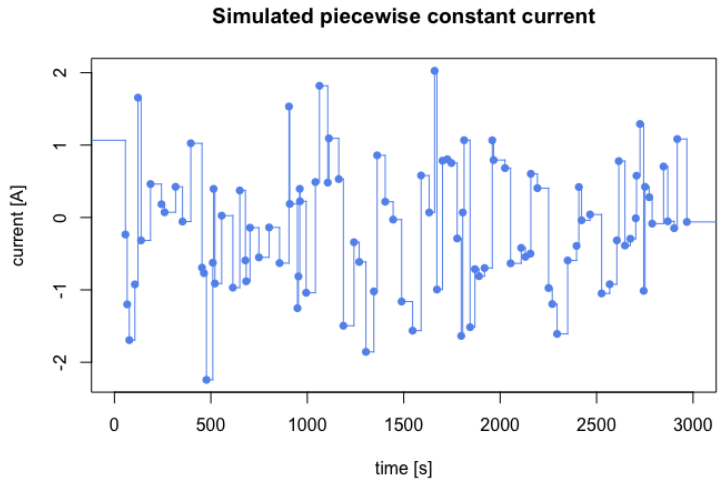


Figure 5.5: Illustration of simulated piecewise constant current measurements obtained at irregular time steps

For each partition of the entire sampling interval, we compute the measurement error variance by taking the differences between the true and error-prone integrals and then computing the variance of these. We resample this process 100 times and take averages of the variances to get more consistent estimates. The following plot shows the measurement error variance as a function of increasing integration interval lengths Δ_i on the x-axis, and motivates us to believe that also for this irregular sampling strategy, the measurement error variance is increasing with integration time. In fact, it increases faster than in the case of regular sampling.

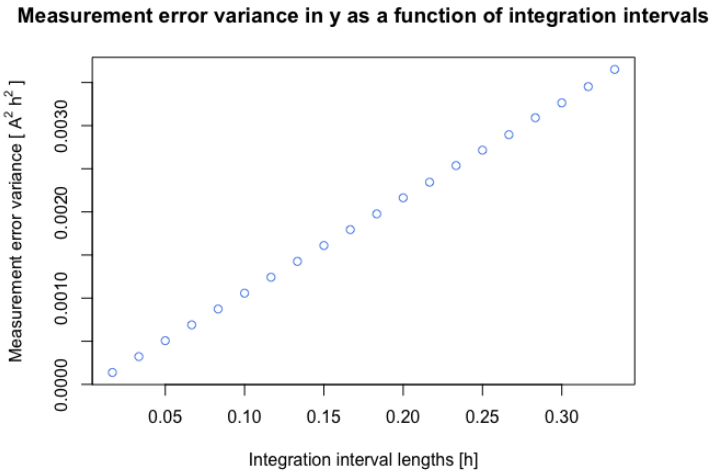


Figure 5.6: Illustration of measurement error variance for varying integration interval lengths Δ_i

Altogether, we can conclude that the uncertainty in y is a function of measurement error and numerical integration error, and although we have to assume it to be unknown due to lack of any additional information, we showed that it increases with integration time. In the next chapter we will present our datasets and the data cleaning steps we took.

CHAPTER 6

Data Preparation

6.1 Overview

Data cleaning is the process of preparing the data for statistical analysis and it is a fundamental part of data science. To ensure that the data is of high quality, it needs to pass a set of criteria. Incorrect, irrelevant, incomplete or duplicated data is not very helpful in the analysis and may in fact provide inaccurate results. There are various methods for cleaning data and in this chapter, we will present our datasets and describe all measures taken in the data preparation stage, such as the usage of a resampling strategy, handling of outliers and data gaps, in order to maximize the data accuracy. One of the main challenges in the data management and analysis has been the enormous quantity of data available.

6.2 About datasets

We are working with real battery sensor data from the Norwegian battery producer Corvus Energy. This data set is from four years of operation of an electric vessel. The operational dataset we will use pertains to a hybrid vessel, that is, a vessel that can combine electric and diesel propulsion system. The battery system is composed of two arrays where each array has 9 packs in parallel. The structure of a pack is illustrated in Figure 6.1. Each pack comprises 21 modules connected in series and each module has a 12s2p configuration of 75 Ah cells, meaning that it is a series of 12 elements, each consisting of 2 parallel cells. Then the capacity per module is 150 Ah. Since the pack is composed of modules in series, the nominal capacity of each pack is also 150 Ah. For every year from 2015 to 2019 and for each of the nine battery packs we obtained two separate large datasets of current and State of Charge (SOC) values. The datasets of one pack in one year for both current and SOC values combined contain typically around 10-15 million observations. The current measurements are given in amperes while SOC values are given as integers expressing percentages. Data of such a significant quantity present big challenges in terms of computational time when loading, handling and modifying it.

Moreover, the datasets are not complete and the amount of collected measurements is irregular and seems to vary depending on the time of the year as can be seen in the histograms in Figure 6.2. Since the operating schedule of a vessel may vary throughout the year, the frequency of the samples

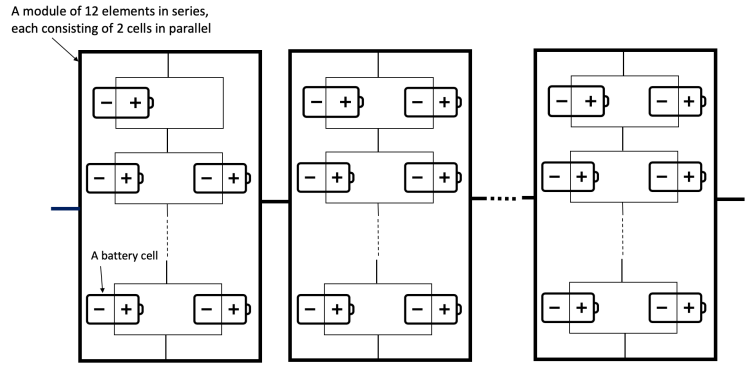


Figure 6.1: Illustration of a pack of modules connected in series

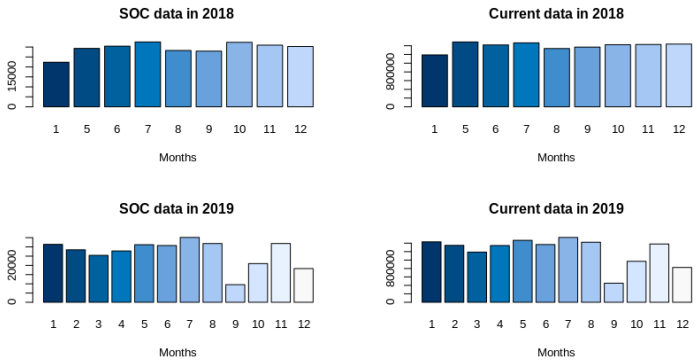


Figure 6.2: Histograms illustrating the amount of SOC and current data points collected for each month in years 2018 and 2019 for battery pack 1

is lower during some months or there are longer time windows when the vessel is not operating at all. The completeness of the datasets should be considered during the analysis and later we will present how we handled such data gaps.

The initial step in the cleaning process was the removal of duplicates and invalid data. By duplicates we mean multiple measurements which were recorded at the exact same timestamp, and by invalid data we refer to measurements which were generated incorrectly and do not conform to some range constraints. For example, SOC values which are outside of the range $[0, 100]\%$ cannot possibly be measured by the definition of SOC and should be removed if they are found in the datasets.
Furthermore, we arranged the measurements chronologically based on the time stamps at which they were collected and created subsets of datasets for each year so that the SOC and current measurements cover the same time range, that is they start and end at the same time stamp of the particular year.

Figures 6.3 and 6.4 show plots of SOC and current measurements during

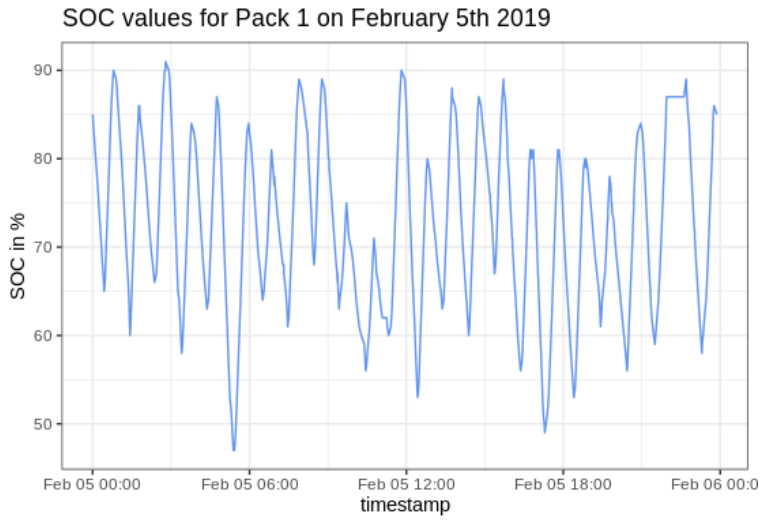


Figure 6.3: Example illustration of SOC values on February 5th 2019

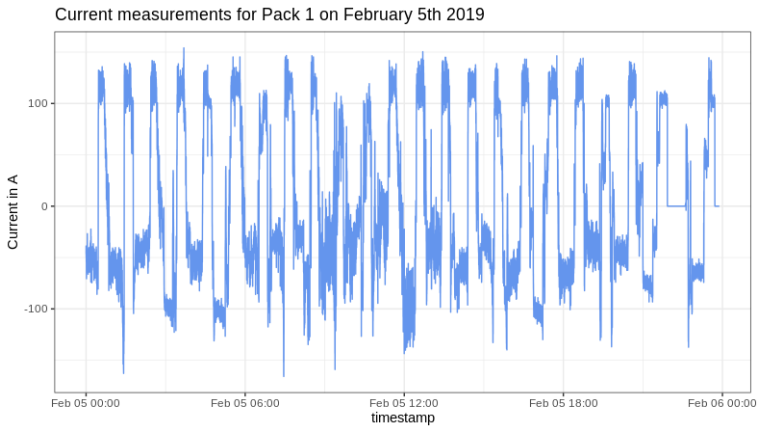


Figure 6.4: Example illustration of current measurements on February 5th 2019

one full day of vessel operation. We can see how the cycles vary throughout the day. Current measurements remain mostly within the range $[-150, 150]$A while the SOC values vary between 50% and 90%. Current of 150 A corresponds to a C-rate of 1 since the nominal capacity of the battery pack is 150 Ah. Next we will present the steps we took to modify the raw SOC and current measurements to bring them on the final form of the x and y variables required for our statistical modelling.

6.3 Resampling strategy

The time intervals between individual measurements of current in the raw data vary a lot which may lead us to believe that the data follows the asynchronuous storage technique described in chapter 5. According to this technique, data

is not immediately backed up if the new value is within a fixed band T of the previously stored measurement. A simple and effective solution for such irregularly sampled time series data is to apply a resampling strategy which we will describe in the following section through various steps.

For our statistical model we defined the response variable y as integrated current over a time interval and the explanatory variable x as the difference in SOC values at the start and end of the same time interval. For measurement data, it is not possible to determine the integral exactly since the function of the current signal is only sampled at discrete time stamps. Therefore, we need to use numerical integration to approximate the integral of the function, as described in the previous chapter. If we assume that the measurements in our data are only recorded if they exceed a certain band around the previous observation, it is reasonable to approximate the integral with the left Riemann sum. Left Riemann sum is computed by summing up areas of rectangles at each measurement point where the height of each rectangle is determined by the value of the current measured at the left endpoint of the interval.

We implemented the numerical integration in the following way. Let the current measurements obtained at $t = \{t_1, t_2, ...t_n\}$ timestamps be denoted $I = \{I_1, I_2, ...I_n\}$. At each timestamp t_i for $i = 1, ...n$ we compute a numerical integral from t_1 until t_i using left Riemann sum, resulting in a vector of cumulative sums $R_1, ..R_n$, as described in the algorithm below.

Algorithm 1 Cumulative numerical integral sum

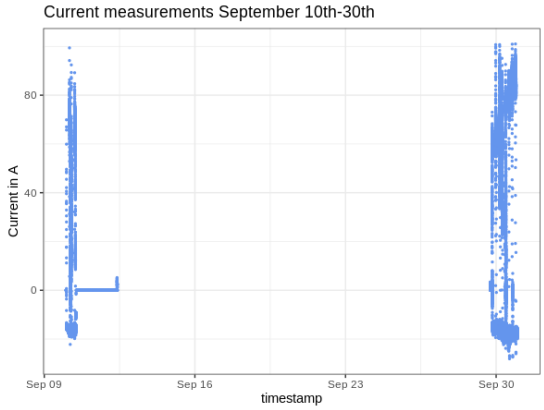
$R_1 \leftarrow 0$
for $k \leftarrow 2$ to n **do**
    $R_k \leftarrow R_{k-1} + I_{k-1} \cdot (t_k - t_{k-1})$
end for

Next we linearly interpolate the cumulatively integrated current $R_1, ..R_n$ at time stamps which split the entire time range $[t_1, t_n]$ uniformly into time intervals of a chosen length. These time intervals are the integration intervals. Finally, we generate the response variable y by taking successive differences of the interpolated values in these newly constructed regular intervals.

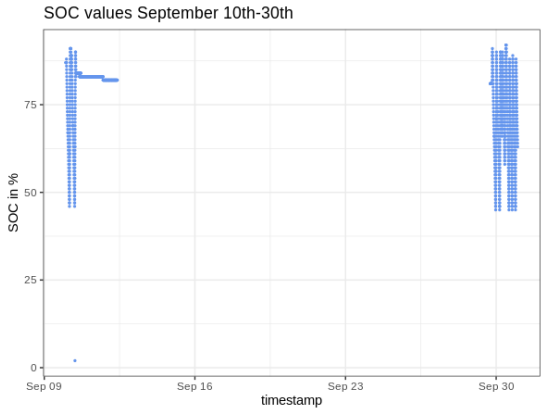
We apply a similar procedure to the SOC and interpolate the values at the same time stamps as we used for current. We get the target explanatory variable x by taking successive differences between the interpolated values.

6.4 Data gaps

There are time windows of different lengths in the raw data in which no measurements at all were recorded. An example of such a time period can be seen in Figures 6.5a and 6.5b where no measurements were stored for 20 days. After the linear interpolation of the datasets over an entire year described in the previous step, artificial data points in these time windows are being constructed which do not represent actual measurements and have no statistical meaning. It is therefore reasonable to remove all variables x and y which fall inside these datagaps. Given that one cycle is approximately one hour long, as can be seen



(a) Current



(b) SOC

Figure 6.5: Illustration of a data gap between September 10th-30th 2019

in Figure 6.4, we decide to use a fixed limit of 15 minutes and remove all y and x variables that were calculated over time intervals falling fully or just partly into any such time gaps without any current measurements of minimum length 15 minutes.

6.5 Outliers

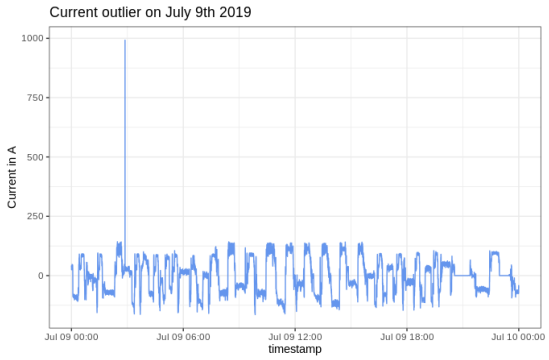
The datasets contain several bit errors due to an unknown cause. Some of the bit errors appear in signal names and timestamps, and can easily be detected and discarded. However, other errors appeared in the numerical values and thus appeared as outliers in the measurements. Therefore, the raw datasets for both current and SOC contain a few outliers which deviate significantly from the overall pattern of the other observations. Outliers should be investigated since they can contain valuable information about the data and its collection. We find that the outliers in our current data represent actual measurements that have been corrupted. Keeping them in the raw data may cause unusual behavior in the integrated current.

In order to remove outliers from both current and SOC values, we wrote an outlier detection algorithm that finds spikes in the data in the following way. For each measurement of current I_j we compute the differences $\Delta_{I_j} = I_j - I_{j-1}$ and $\Delta_{I_{j+1}} = I_{j+1} - I_j$, where I_{j-1} and I_{j+1} are the previous and following measurement, respectively. If both of the Δ's have opposite signs and their absolute values exceed a certain fixed limit θ_I, this indicates that I_j is an outlier. The question is how to choose such threshold θ_I. We find that a $\theta_I = 200$ A is sufficient to filter out the largest outliers in current which are incorrect with high probability and may influence our analysis. Spikes in the current signal can occur from a physical point of view due to for example sudden turning on and off the load, and therefore we choose this limit high to make sure no real observations are filtered away.
We apply the same algorithm to detect sudden spikes in SOC, filtering out values using threshold $\theta_{SOC} = 30\%$. Furthermore, one needs to consider the time steps between the measurements when computing the differences between their values. For example, we saw that our current measurements as seen in Figure 6.4 vary approximately from -150 A to 150 A, so the maximum C-rate achieved is 1. The battery pack of capacity 150 Ah goes from fully charged to fully discharged in one hour at a C-rate of 1, which corresponds to a $\Delta_{SOC_j} = 100/60 \approx 1.67$ in one minute. So an SOC change within a minute which deviates greatly from this value seems unrealistic. Since we remove time windows containing no measurements which exceed 15 minutes as discussed in the previous chapter, the largest time step between two SOC values is 15 minutes. Hence, since the same C-rate of 1 would give a difference of 25 over a 15 min interval, a limit θ_{SOC} of 30 for SOC seems rather reasonable.
After finding all outliers in both current and SOC data, we remove all x and y variables which were computed over time intervals including any outliers. By discarding periods where an outlier occurs, not just the measurement itself, the outlier removal will reduce the size of the dataset, but it should not affect the estimates in other ways.

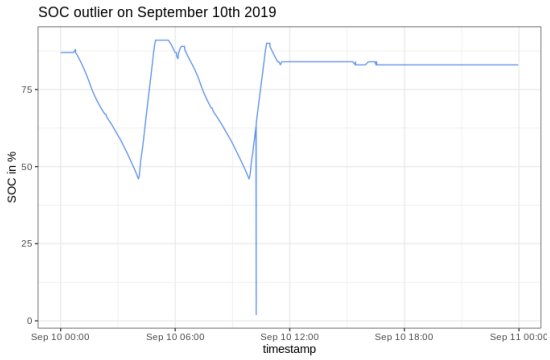
SOC value in %	timestamp
61	2019-09-10 10:11:21.34
62	2019-09-10 10:12:31.44
63	2019-09-10 10:13:39.54
2	2019-09-10 10:14:31.38
63	2019-09-10 10:14:32.46
64	2019-09-10 10:14:47.35
65	2019-09-10 10:15:54.39

Table 6.1: Example of an outlier in the SOC

Tables 6.1 and 6.2 show examples of outliers found in the SOC and current data for pack 1 from year 2019. SOC cannot realistically go from 63 to 2 percent in less than a minute since that would correspond to a C-rate of 36.6, which is clearly outside the possible range. Besides, current of 992 A would correspond to a C-rate of 6.6 which may be allowed for a few seconds but seems rather unrealistic from a physical point of view. Both of these outliers are also illustrated in Figures 6.6a and 6.6b.



(a) Current outlier



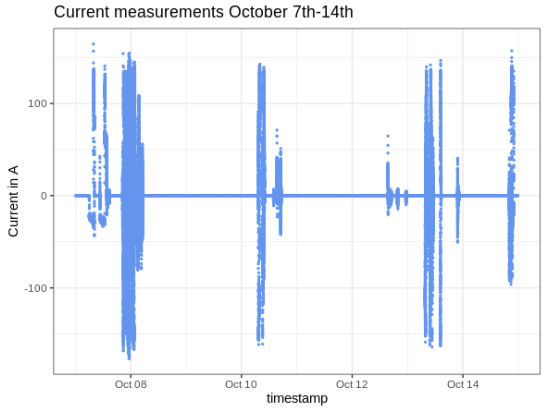
(b) SOC outlier

Figure 6.6: Outliers in current and SOC

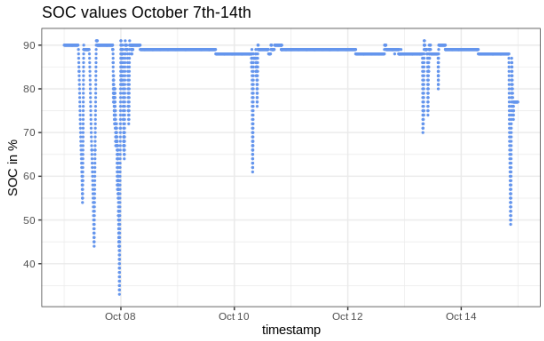
Current value in A	timestamp
34	2019-07-09 02:52:20.22
34.9	2019-07-09 02:52:21.24
35.9	2019-07-09 02:52:23.21
992.1	2019-07-09 02:52:25.25
37.3	2019-07-09 02:52:27.18
36.8	2019-07-09 02:52:28.19
34.3	2019-07-09 02:52:30.18

Table 6.2: Example of an outlier in the current

An alternative way of removing outliers which we did not apply to our data is to simply filter out unrealistic measurements which exceed the allowed range of C-rates specified by the battery manufacturer. For the battery packs we are working with, the maximum C-rate during discharge should not exceed 4, while during charge, the maximum C-rate allowed is 3C, which corresponds to maximum current of 600 A and 450 A, correspondingly. Over short periods of a few seconds even higher C-rates are allowed as sudden changes can occur



(a) Current



(b) SOC

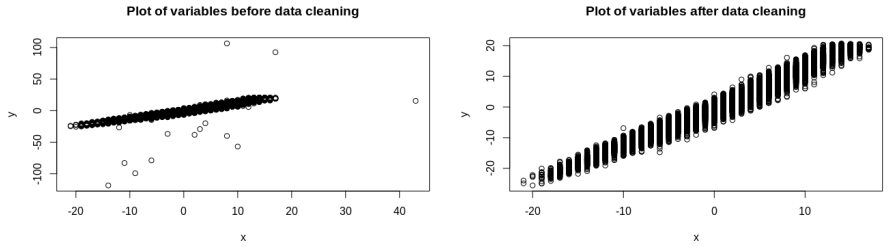
Figure 6.7: Illustration of zero values in current measurements with corresponding SOC values

when for example turning on and off the load. This approach provides a straight-forward solution which is easy to implement.

6.6 Zero values in current

We decide to remove intervals where the batteries are not in use and therefore, the current is not flowing and has value 0 over many consecutive measurements. When there is no current flow, then in principle the SOC does not change, if we disregard any self-discharge. Periods where the battery is not used, are not valuable when estimating its state, and may skew the results of the analysis.

We remove them in the following way. In the integrated and interpolated current data resulting in the final y variables, we find all values equal to 0 and check that the original current measurements are also all equal to 0 in the corresponding time interval, thereby making sure that the zero is not just a result of summing up positive and negative charge. For all y with value equal to 0, if the raw current measurements collected over the time interval used for numerical integration of the y variable are also all zero, then we remove the



(a) Before (b) After

Figure 6.8: Illustration of variables before and after data cleaning from pack 1 in 2019 with regular 10 minute integration intervals

value from the final dataset. Finally, we also remove all x values obtained over the same time interval. Figures 6.7a and 6.7b show such periods of several days in the raw data where the current was measured to be equal zero while the SOC stayed constant.

raw current	14728368
raw SOC	322679

	y
After resampling	50600
After gap removal	47949
After irrelevant zeroes removal	45812
After outliers removal	45810

Table 6.3: Number of datapoints for pack 1 2019 after each data cleaning process

Table 6.3 shows the amount of data points left after each data cleaning process for battery pack 1 in year 2019. Starting from very large samples of both current and SOC, we resampled using integration intervals of 10 minutes and obtained 50600 points of variable x and y each. For x and y each, approximately 5.2% of these data points were removed during data gap removal, other 4.2% were removed during the removal of zero values in current and only two outliers were found in the dataset, hence less than 10% of the data were removed in total during data cleaning, which some might consider a negligible amount.

The Figure 6.8 illustrates the difference between the variables x and y as obtained from the above described resampling strategy from battery pack 1 in 2019, and after we removed all data gaps, outliers and irrelevant zero current values. Fitting a simple linear regression of y on x in both cases gives estimated capacity $\hat{Q}_{before} = 125.9$ Ah and $\hat{Q}_{after} = 125.8$ Ah, so we see that data cleaning affects the estimates to only a small extent.

CHAPTER 7

Results

7.1 Overview

In this chapter we will present the results from the WTLS methods and compare these to OLS. We will discuss how to approximate the measurement uncertainties and evaluate which ones give a good fit based on a goodness of fit criterion. Lastly we will compute confidence bounds to give an idea about the uncertainty of our capacity estimates and compare our results with one annual SOH test.

7.2 Initial results using OLS

First we will ignore the presence of measurement errors and look at the naive capacity estimates we obtain from ordinary linear regression. In Figure 7.1

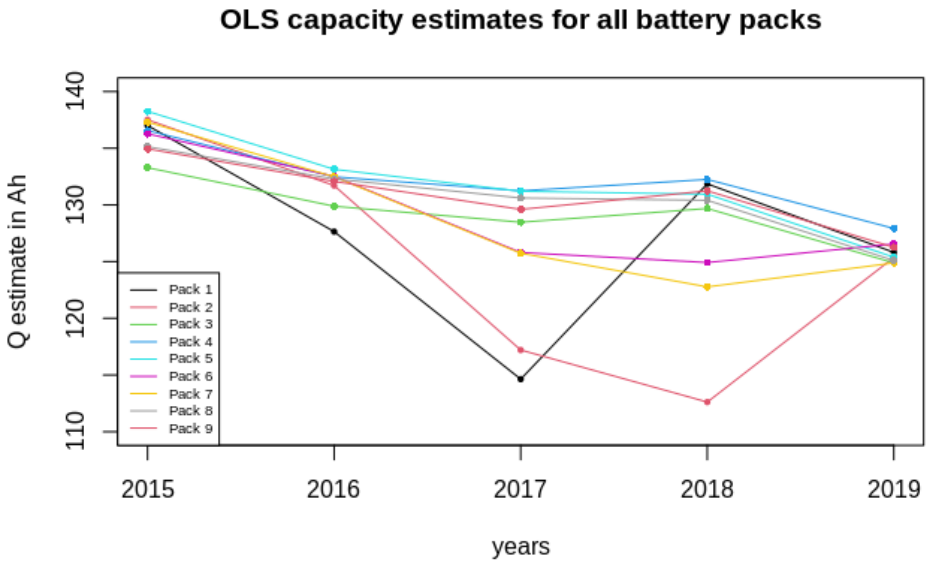


Figure 7.1: Illustration of yearly OLS estimates of capacity Q for all battery packs with 10 min integration intervals

we computed an estimate for each year from 2015 to 2019 for each of the 9 packs, illustrated with different colors. Most of the packs show a moderate degradation trend as expected. We will closer investigate battery pack 5.

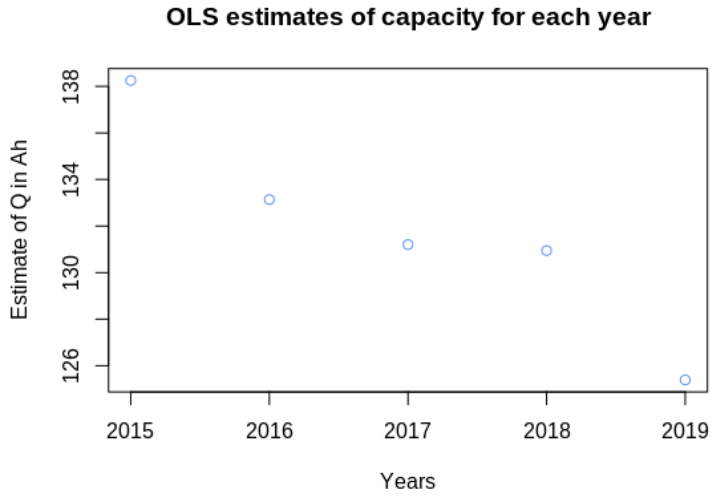


Figure 7.2: Illustration of yearly OLS estimates of capacity Q for pack 5 with 10 minute integration intervals

Figure 7.2 illustrates capacity estimates computed for each year separately for pack 5. We see that the capacity decreases from 138.25 Ah to 125.39 Ah, which corresponds to an SOH degradation by approximately 9%, considering the battery pack has nominal capacity of 150 Ah.

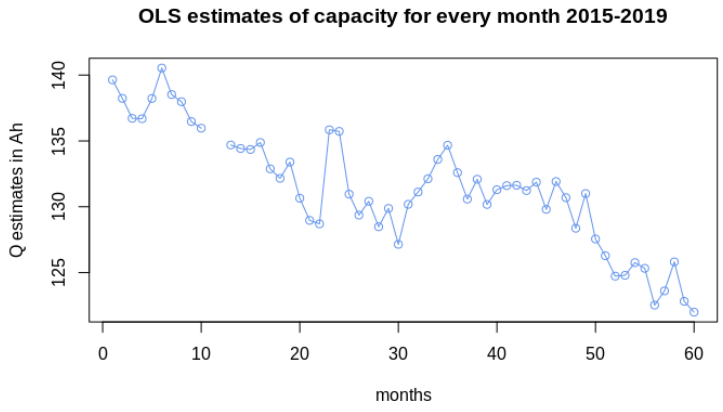


Figure 7.3: Illustration of monthly OLS estimates of capacity Q for pack 5 with 10 minute integration intervals

Figure 7.3 illustrates capacity estimates computed for each month separately instead of year. In 2015, no estimates were obtained in November and December due to missing data. The overall trend is that the capacity is decreasing but the estimates vary a lot. This variation may explained by variations in the operation of the battery, seasonal variations in the temperature or other parameters not taken into account in this work.

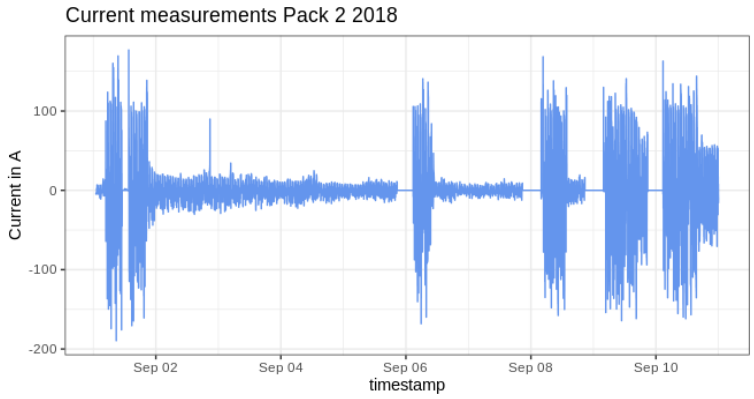


Figure 7.4: Current data for pack 2 in september 2018

From Figure 7.1 we notice that the results for packs 1 and 2 are not as convincing as the others. The capacity estimates decrease in years 2017 and 2018, respectively, and then increase again. Such increases in the capacity indicate that there is some unexpected behaviour taking place in the data since battery capacity typically degrades over time. It may be caused by repairs but we do not have sufficient information to conclude so.

We observed one possible cause of these dubious estimates in the raw current

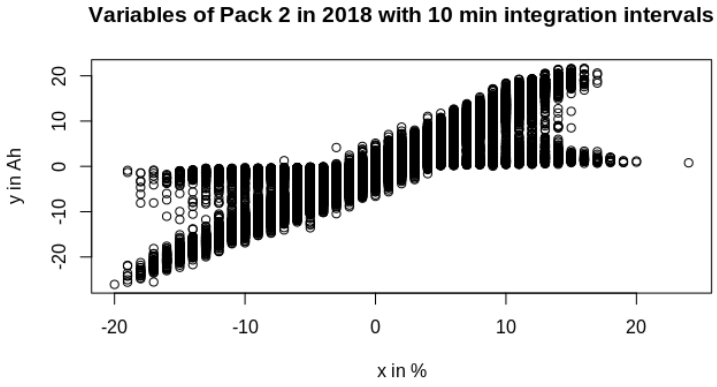


Figure 7.5: Variables x and y of pack 2 in 2018 with integration length of 10 min

data and were informed by the data provider that there were faulty current

sensors in some battery packs. Figure 7.4 illustrates such time periods in battery pack 2 where the current measurements are too close to zero and after integration, they do not correspond to the changes in SOC, as can be seen in Figure 7.5 along the x-axis. For future research, this issue could be solved by constructing an algorithm which filters out these peculiar data. We will continue our analysis disregarding data from battery packs which show signs of such faulty sensor behaviour.

In the following, we will return our focus to data from battery pack 5 in 2019 with integration length of 10 minutes and discuss the results from ordinary least squares regression. Ignoring the measurement errors and applying OLS to the error-prone (x_i, y_i)'s, we obtain a naive capacity estimate $\hat{Q}_{OLS}$ of 125.39 Ah which corresponds to SOH of 83.6%, given that the nominal capacity is 150 Ah. We can see the plot of the variables including the regression line in Figure 7.6. This model gives a good fit with adjusted R^2 of 97.9% but we know from theory that due to measurement errors, this estimate is attenuated towards zero, and that the statistical properties are disturbed.

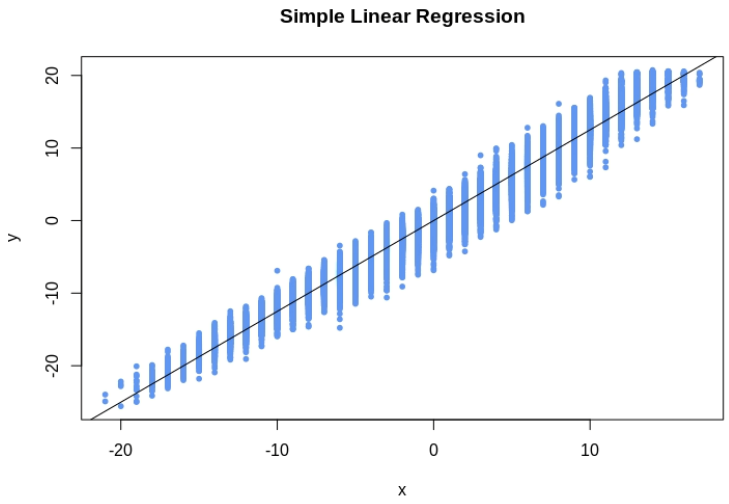


Figure 7.6: Illustration of OLS regression line for pack 5 in 2019

In fact, the QQ plot of residuals from ordinary regression in Figure 7.7 can be used to see that the normality assumption is violated. Ideally the residuals will follow the straight dashed line. In our model the residuals tend to be larger in the lower tails than what one should expect if they were normally distributed, that is, they have heavier lower tails than they should, and the distribution is left-skewed.

The sample variance for the ordinary linear regression can be computed by

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \hat{Q}_{OLS} \cdot x_i)^2, \tag{7.1}$$

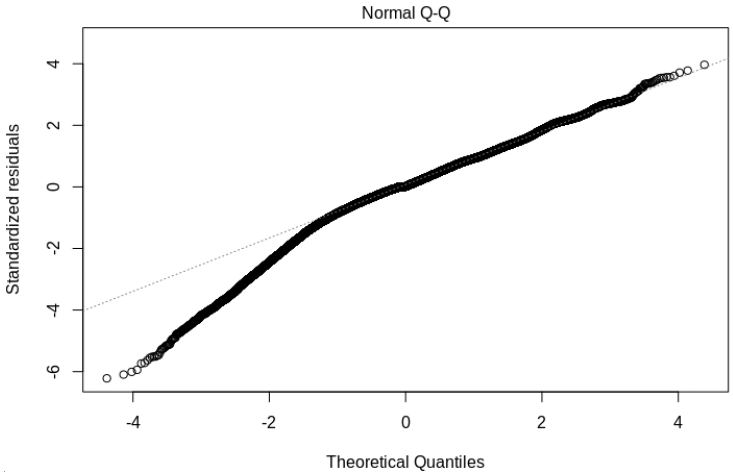


Figure 7.7: Illustration of QQ plot

and for our data set for pack 5 in 2019 we obtain a sample variance $s^2 = 1.77$. Fitting a naive linear regression model of the error-prone y on x which does not consider measurement error in x causes an increased variance of residuals which are the vertical distances from the observed points to the naive regression line. This variance estimate might be considered an approximation to the upper bound of the measurement error variance in y. Later we will use $\hat{Q}_{OLS}$ as a rough guess of Q, in order to search for more information about the measurement error variances $\sigma^2_{\Delta x}$ and $\sigma^2_{\Delta y}$.

We have shown before that ordinary linear regression underestimates the slope estimates when there is measurement error in the explanatory variables. In the following section we will see if weighted total least squares yields more precise capacity estimates.

7.3 WTLS

We will compare results from 2019, mainly focusing on the comparison between OLS and WTLS since TLS is a special case of WTLS with proportional uncertainties, and AWTLS is an approximation which should give the same results as WTLS. The main advantage of these two alternative approaches is their recursive implementation.

To start, in chapter 5 we derived a range for the uncertainties in x. We will put these as parameters into our models and compute WTLS capacity estimates on a grid of $\sigma^2_{\Delta y}$ and $\sigma^2_{\Delta x}$. Since we look only to the year 2019, we include pack 1 and 2 here, since the current sensor did not show faulty behavior as in 2017 and 2018. The results for pack 1 are shown in Table 7.1. The OLS estimate is 125.81 Ah. The WTLS estimates are higher and vary between 125.85 Ah and 128.88 Ah. For fixed $\sigma^2_{\Delta y}$, the estimates become larger with increasing

	$\sigma^2_{\Delta x}$						
$\sigma^2_{\Delta y}$	0.1	0.5	1	2	5	10	12
0.1	127.71	128.55	128.71	128.80	128.85	128.87	128.88
0.5	126.56	127.71	128.16	128.48	128.71	128.80	128.82
1	126.23	127.18	127.71	128.16	128.55	128.71	128.74
2	126.04	126.69	127.18	127.71	128.28	128.55	128.60
5	125.91	126.23	126.56	127.01	127.71	128.16	128.26
10	125.86	126.04	126.23	126.56	127.18	127.71	127.84
12	125.85	126.00	126.17	126.46	127.04	127.57	127.71
	$\hat{Q}_{OLS} = 125.81$ Ah						

Table 7.1: WTLS results for pack 1 in 2019 for varying uncertainties in x and y

$\sigma^2_{\Delta x}$, while for a fixed $\sigma^2_{\Delta x}$ they decrease with an increasing $\sigma^2_{\Delta y}$. The lowest estimate in the table was obtained in the bottom left corner for $\sigma^2_{\Delta x} = 0.1$ and $\sigma^2_{\Delta y} = 12$. Since the ratio between the two error variances $\phi = \frac{\sigma^2_{\Delta x}}{\sigma^2_{\Delta y}}$ here is very close to 0, meaning that the measurement error in x is very small compared to measurement error in y, the model is close to ordinary linear regression which assumes no measurement error in x at all. On the other hand, for the inverse ratio with $\sigma^2_{\Delta x} = 12$ and $\sigma^2_{\Delta y} = 0.1$ we obtain a capacity estimate of 128.88 Ah in the top right corner. Hence the larger this ratio ϕ is, the higher the estimates become. Along the diagonal, we see that when the ratio is equal to 1, the capacity estimates are equal to 127.71 Ah.

The OLS estimate for pack 2 for 2019 is 125.37 Ah as shown in Table 7.2.

	$\sigma^2_{\Delta x}$						
$\sigma^2_{\Delta y}$	0.1	0.5	1	2	5	10	12
0.1	126.85	127.51	127.63	127.70	127.75	127.76	127.76
0.5	125.94	126.85	127.20	127.45	127.63	127.70	127.71
1	125.69	126.43	126.85	127.20	127.51	127.63	127.66
2	125.54	126.05	126.43	126.85	127.29	127.51	127.55
5	125.44	125.69	125.94	126.30	126.85	127.20	127.28
10	125.40	125.54	125.69	125.94	126.43	126.85	126.95
12	125.40	125.51	125.65	125.87	126.32	126.74	126.85
	$\hat{Q}_{OLS} = 125.37$ Ah						

Table 7.2: WTLS results for pack 2 in 2019 for varying uncertainties in x and y

The WTLS estimates are again higher and vary between 125.40 Ah and 127.76 Ah depending on the size of the uncertainties while along the diagonal, the capacity estimates are all equal to 126.85 Ah.

The results for pack 5 in 2019 are shown in Table 7.3 and they are similar to pack 1 and pack 2. The WTLS estimates increase with an increasing ratio ϕ and are within the range [125.42,128.06] Ah, while OLS estimates a capacity of 125.39 Ah. The WTLS method showed similar results for the rest of the battery packs in year 2019 and they can be found in Tables A.1, A.2, A.3, A.4, A.5 and A.6 in appendix A.

	$\sigma^2_{\Delta x}$						
$\sigma^2_{\Delta y}$	0.1	0.5	1	2	5	10	12
0.1	127.04	127.78	127.92	127.99	128.04	128.06	128.06
0.5	126.03	127.04	127.43	127.71	127.92	127.99	128.01
1	125.75	126.58	127.04	127.43	127.78	127.92	127.94
2	125.58	126.15	126.58	127.04	127.54	127.78	127.82
5	125.47	125.75	126.03	126.43	127.04	127.43	127.52
10	125.43	125.58	125.75	126.03	126.58	127.04	127.15
12	125.42	125.55	125.70	125.95	126.46	126.92	127.04
	$\hat{Q}_{OLS} = 125.39$ Ah						

Table 7.3: WTLS results for pack 5 in 2019 for varying uncertainties in x and y

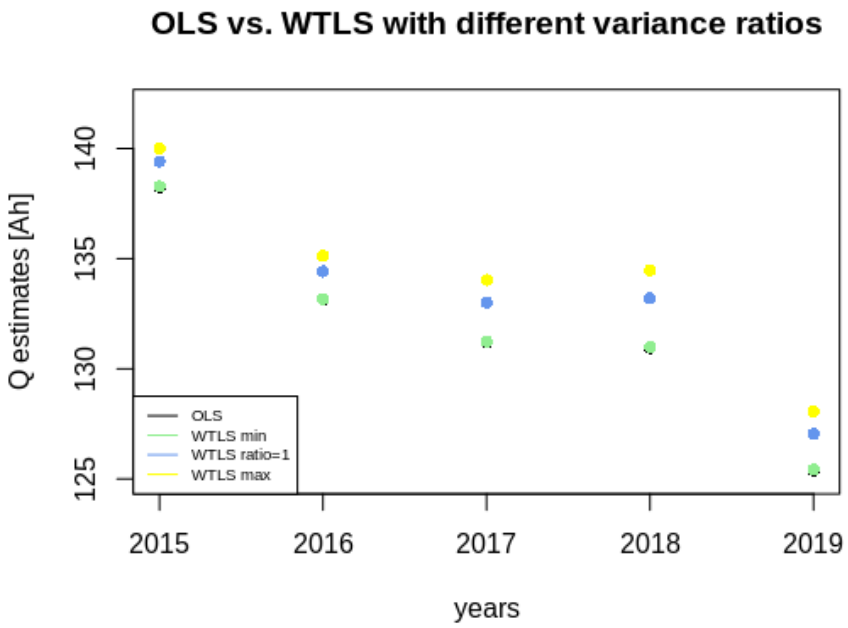


Figure 7.8: Illustration of OLS and WTLS capacity estimates for three different ratios of measurement error variances for pack 5 2015-2019

Figure 7.8 illustrates WTLS capacity estimates for pack 5 from year 2015 to 2019 with three different values for ratios of measurement error variances as well as one estimate per year from ordinary linear regression. These results are also given in Table 7.4 where WTLS min corresponds to a capacity estimate obtained for the lowest ratio in our tables $\phi = \frac{\sigma^2_{\Delta x}}{\sigma^2_{\Delta y}} = 0.1/12 \approx 0.0083$, and WTLS max was obtained for ratio $\phi = 12/0.1 = 120$. We see in the figure that the black OLS and the green WTLS estimates with ratio close to 0

are almost indistinguishable. This is expected since model where the measurement error in y is increasingly dominating should become very similar to OLS.

	2015	2016	2017	2018	2019
OLS	138.25	133.14	131.20	130.95	125.39
WTLS min	138.28	133.17	131.24	130.99	125.42
WTLS ratio $\phi = 1$	139.42	134.42	133.01	133.19	127.04
WTLS max	140.01	135.13	134.03	134.47	128.06

Table 7.4: OLS vs. WTLS estimates of capacity Q in Ah for pack 5

These results are already interesting and demonstrate the usefulness of the WTLS approach for handling measurement error regression. Using WTLS estimates with ratio $\phi = 1$ as reference points, we see that ignoring measurement error and fitting OLS regression gives an underestimate of capacity approximately between 1% and 1.5% which is of some practical importance for the battery provider. We would, however like to understand more about the magnitudes of the uncertainties in x and y, and will in the following pursue some attempts.

Previously, we approximated an upper bound for $\sigma^2_{\Delta y}$ with the sample variance. Another way of approximating the measurement error variances in x and y from the naive linear regression model can be constructed from Figure 4.2. The measurement errors in each data point can be approximated by

$$\Delta x_i = R_i \sin\theta = \delta y_i \cos\theta \sin\theta \tag{7.2}$$

and

$$\Delta y_i = R_i \cos\theta = \delta y_i \cos\theta \cos\theta \tag{7.3}$$

where $\theta = \tan^{-1} \hat{Q}_{OLS}$ and δy_i is the vertical distance between the observed (x_i, y_i) and the regression line $\hat{Q}_{OLS} \cdot \mathbf{x}$ so that $\delta y_i = |\hat{Q}_{OLS} \cdot x_i - y_i|$. We can easily compute the two variances of these approximations using our data and the fitted regression model, and for the dataset for pack 5 in 2019 we obtain the estimated measurement error variances $\hat{\sigma}^2_{\Delta x} = 0.197$ and $\hat{\sigma}_{\Delta y} = 0.125$, which are much smaller than the previously assumed range of variances $[0, 12.5]$. However, we need to be aware that ordinary linear regression underestimates the slope when there is measurement error in x and y, and hence these crude estimates only serve as indications. In the next section we will describe a goodness of fit criterion which can also give us an idea about the magnitude of the uncertainties.

7.4 Chi-Square test for goodness of fit

There are a few methods to compute goodness of fit and confidence intervals for total least squares but none of them are widely accepted and it is still an area of research. The methods commonly used for ordinary linear regression may not make any sense for errors-in-variables regression because of the different model assumptions. Press (1992) argues that for a fitting

procedure to be useful, it should provide parameters, error estimates on the parameters, and a statistical measure of goodness-of-fit. When the goodness of fit criterion suggests that the model gives a bad fit to the data, then the parameters and their error estimates are probably worthless (Press 1992, chapter 15).

For the WTLS model Plett (2011) proposes to use a chi-square test for evaluating the goodness of fit. When we assume that the measurements errors in x and y are uncorrelated and Gaussian, then the merit function χ^2 in Equation 4.16 is a chi-squared random variable since it is a sum of n squares of normally distributed quantities normalized to unit variance (Press 1992, chapter 15). This knowledge of the distribution and number of degrees of freedom can be used to determine from the optimized values of the merit functions how reliable the model fit is. If we have chosen the estimate of capacity and the measurement error variances correctly, we may expect that the value of the merit function will be approximately equal to the number of degrees of freedom ν and we can conclude that the data are well described by the hypothesized merit function.

The incomplete gamma function $P(\chi^2|\nu)$ is defined as the probability that the observed chi-square for a correct model should be less than a value χ^2 for ν degrees of freedom. Its complement $Q(\chi^2|\nu)$ is the probability that the observed chi-square will exceed the value χ^2 by chance even for a correct model (Press 1992, chapter 6). The complementary incomplete gamma function is given by the formula

$$Q(\chi^2|\nu) = \frac{1}{\Gamma(\nu/2)} \int_{\chi^2/2}^{\infty} \exp^2 t^{(\nu/2-1)}\, dt, \tag{7.4}$$

and gives a measure of the goodness of fit of a model (Plett 2011).

If on the one hand $Q(\chi^2|\nu)$ gives a small probability, then either the model is wrong or the measurement error variances are larger than assumed. Besides, $Q(\chi^2|\nu)$ does not measure the credibility of the assumption of normally distributed measurement errors. Nonnormal errors may create outliers which decrease the probability $Q(\chi^2|\nu)$. Therefore, other reasons for a low probability could be that measurement errors are not normally distributed or they have been underestimated (Press 1992, chapter 15). Since this is a fairly common case, experimenters often accept low probabilities $Q(\chi^2|\nu) > 0.001$ (Plett 2011). If on the other hand $Q(\chi^2|\nu)$ is very close to 1, the data seems too good to be true. According to Press (1992), the cause of too good a chi-square fit is almost always that the experimenter has overestimated the measurement errors.
The χ^2 statistic has a mean ν and a standard deviation $\sqrt{2\nu}$, and, for large ν, it approaches the normal distribution by the Central Limit Theorem. Therefore, a general rule of thumb is that a typical value of χ^2 for a moderately good fit is $\chi^2 \approx \nu$ (Press 1992, chapter 15). When the uncertainties associated with a set of measurements are unknown, considerations related to χ^2 fitting can be used to derive an approximation to their value. This approach prohibits an independent assessment of goodness of fit but it allows us to obtain some kind of error bar to the observations.

We will compare the value of the merit function χ^2 with a critical chi-square value at significance level α, for which a common value is 5%, hence $\alpha = 0.05$.

If $\chi^2 > \chi^2_{\nu,\alpha}$ for ν degrees of freedom, then the model does not describe the data well or the measurement uncertainties have been underestimated. If $\chi^2 < \chi^2_{\nu,1-\alpha}$, then either the data is "too good to be true" or the measurement uncertainties have been too conservatively overestimated.

The merit function we minimize to obtain a WTLS capacity estimate in Equation 4.16 is a chi-squared random variable with $2n-1$ degrees of freedom since n observations x_i and n observations y_i were used to construct it and one degree of freedom was lost when fitting $\hat{Q}$ (Plett 2011). From the optimal estimates we found and the uncertainties we assumed, the merit function can easily be computed. For data of Pack 1 in 2019 there are $n = 45810$ observations of x and y each, meaning that the merit function χ^2 has $2n-1 = 91619$ degrees of freedom. Critical chi-square values at significance level $\alpha = 0.05$ are $\chi^2_{91619,0.05} = 92324.24$ and $\chi^2_{91619,0.95} = 90916.04$.

On the same grid of measurement error variances $\sigma^2_{\Delta x}$ and $\sigma^2_{\Delta y}$ as before, we compute the values of merit functions χ^2 using the estimated capacities $\hat{Q}$ from Table 7.1. We see in Table 7.5 that for $\sigma^2_{\Delta x}, \sigma^2_{\Delta y} > 1$, much smaller values of χ^2 compared to the critical value of 90916 were obtained, leading us to believe that the measurement uncertainties were overestimated. This confirms our assumption from the previous section that the uncertainties are much smaller. When we do the same computations for a smaller range of measurement error variances, we see in Table 7.6 that the values of χ^2 are closer to the critical value.

		$\sigma^2_{\Delta x}$						
		0.1	0.5	1	2	5	10	12
$\sigma^2_{\Delta y}$	0.1	340968	97815	51695	26604	10832	5448	4545
	0.5	134828	68194	42075	23811	10339	5321	4456
	1	76700	49377	34097	21038	9781	5169	4349
	2	41176	31778	24689	17048	8827	4891	4150
	5	17230	15340	13483	10845	6819	4208	3648
	10	8750	8235	7670	6741	4938	3410	3033
	12	7310	6948	6541	5854	4446	3169	2841

Table 7.5: Observed χ^2 values for estimated $\hat{Q}$ from pack 1 in 2019 for varying uncertainties in x and y

By narrowing down the measurement error variances, we can also narrow down the range of possible capacity estimates $\hat{Q}$, from $[125.85, 128.88]$Ah as computed in Table 7.1 to $[126.23, 128.71]$Ah in Table 7.7. In fact, we could narrow down the measurement error variances even more but this would only have a negligible effect on the capacity estimates.

We will do the same for pack 2 in 2019 to see if we obtain similar results. We have $n = 43596$ observations of x and y each, hence the merit function χ^2 should have value approximately around the corresponding number of degrees of freedom to give a good fit, which is $2n-1 = 87191$. Critical chi-square values at significance level $\alpha = 0.05$ are $\chi^2_{87191,0.05} = 87879.01$ and $\chi^2_{87191,0.95} = 86505.26$.

	$\sigma^2_{\Delta x}$						
$\sigma^2_{\Delta y}$	0.1	0.2	0.4	0.5	0.6	0.8	1
0.1	340968	210377	119053	97815	83005	63711	51695
0.2	246887	170484	105188	88267	76033	59527	48907
0.4	158890	123444	85242	73799	65059	52594	44134
0.5	134828	108446	77839	68194	60669	49693	42075
0.6	117088	96688	71612	63373	56828	47091	40198
0.8	92688	79445	61722	55512	50432	42621	36900
1.0	76700	67414	54223	49377	45322	38920	34097

Table 7.6: Observed χ^2 values for estimated $\hat{Q}$ from pack 1 in 2019 for smaller range of uncertainties in x and y

	$\sigma^2_{\Delta x}$						
$\sigma^2_{\Delta y}$	0.1	0.2	0.4	0.5	0.6	0.8	1
0.1	127.71	128.16	128.48	128.55	128.6	128.67	128.71
0.2	127.18	127.71	128.16	128.28	128.37	128.48	128.55
0.4	126.69	127.18	127.71	127.87	127.99	128.16	128.28
0.5	126.56	127.01	127.54	127.71	127.84	128.03	128.16
0.6	126.46	126.88	127.4	127.57	127.71	127.91	128.06
0.8	126.32	126.69	127.18	127.35	127.49	127.71	127.87
1.0	126.23	126.56	127.01	127.18	127.32	127.54	127.71
	$\hat{Q}_{OLS} = 125.81$ Ah						

Table 7.7: Estimated capacity $\hat{Q}$ in Ah from pack 1 in 2019 for smaller range of uncertainties in x and y

	$\sigma^2_{\Delta x}$						
$\sigma^2_{\Delta y}$	0.1	0.5	1	2	5	10	12
0.1	261092	75205	39779	20481	8341	4196	3500
0.5	102796	52218	32286	18300	7956	4096	3431
1	58422	37728	26109	16143	7520	3978	3347
2	31347	24239	18864	13055	6777	3760	3192
5	13113	11684	10280	8281	5222	3229	2801
10	6658	6269	5842	5140	3773	2611	2324
12	5563	5289	4982	4462	3395	2425	2176

Table 7.8: Observed χ^2 values for estimated $\hat{Q}$ from pack 2 in 2019 for varying uncertainties in x and y

In Table 7.8 we see again that for $\sigma^2_{\Delta x}, \sigma^2_{\Delta y} > 1$ we obtained very small values of χ^2 compared to the critical chi-square values, indicating overestimation of the uncertainties. The chi-square goodness of fit criterion shows that for smaller measurement error variances, the model describes the data better, as can be seen in Table 7.9.

Again we can narrow down the range of possible capacity estimates $\hat{Q}$ from $[125.4, 127.76]$ Ah as in Table 7.2 to $[125.69, 127.63]$ Ah. We see from these results that assuming smaller measurement error variances based on the chi-

	$\sigma^2_{\Delta x}$						
$\sigma^2_{\Delta y}$	0.1	0.2	0.4	0.5	0.6	0.8	1
0.1	261092	161430	91499	75205	63836	49015	39779
0.2	188641	130546	80715	67770	58401	45750	37602
0.4	121193	94321	65273	56551	49882	40358	33885
0.5	102796	82809	59562	52218	46483	38107	32286
0.6	89244	73796	54765	48499	43515	36092	30830
0.8	70618	60597	47160	42444	38582	32636	28275
1.0	58422	51398	41405	37728	34649	29781	26109

Table 7.9: Observed χ^2 values for estimated $\hat{Q}$ from pack 2 in 2019 for smaller range of uncertainties in x and y

	$\sigma^2_{\Delta x}$						
$\sigma^2_{\Delta y}$	0.1	0.2	0.4	0.5	0.6	0.8	1
0.1	126.85	127.2	127.45	127.51	127.55	127.6	127.63
0.2	126.43	126.85	127.2	127.29	127.36	127.45	127.51
0.4	126.05	126.43	126.85	126.97	127.07	127.2	127.29
0.5	125.94	126.3	126.71	126.85	126.95	127.1	127.2
0.6	125.87	126.2	126.61	126.74	126.85	127	127.12
0.8	125.76	126.05	126.43	126.57	126.68	126.85	126.97
1.0	125.69	125.94	126.3	126.43	126.54	126.71	126.85
	$\hat{Q}_{OLS} = 125.37$ Ah						

Table 7.10: Estimated capacity $\hat{Q}$ in Ah from pack 2 in 2019 for smaller range of uncertainties in x and y

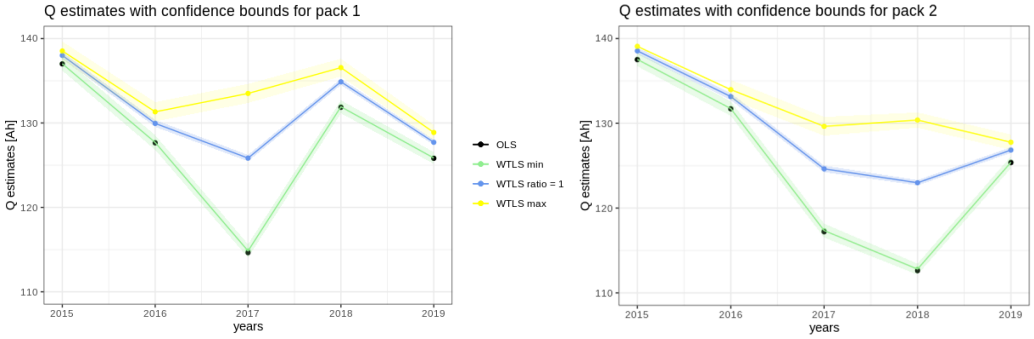
square goodness of fit criterion only has a small effect on the estimates. However, it gives us a better understanding of the magnitude of the uncertainties.

7.5 Confidence limits

After finding a capacity estimate $\hat{Q}$ using one of the weighted total least squares methods, it is also important to estimate its uncertainty and confidence bounds. If we assume that all errors are normally distributed, we can redefine the least squares problem as maximum likelihood optimization. As Plett (2011) suggests, we can form a vector $\mathbf{d}$ by concatenating the response vector $\mathbf{y}$ and the predictor vector $\mathbf{x}$ and a diagonal matrix $\mathbf{\Sigma_d}$ formed by measurement error variances $\sigma^2_{\Delta y}$ followed by $\sigma^2_{\Delta x}$ (Plett 2011). Then minimizing the merit function χ^2 is equivalent to maximizing the likelihood function

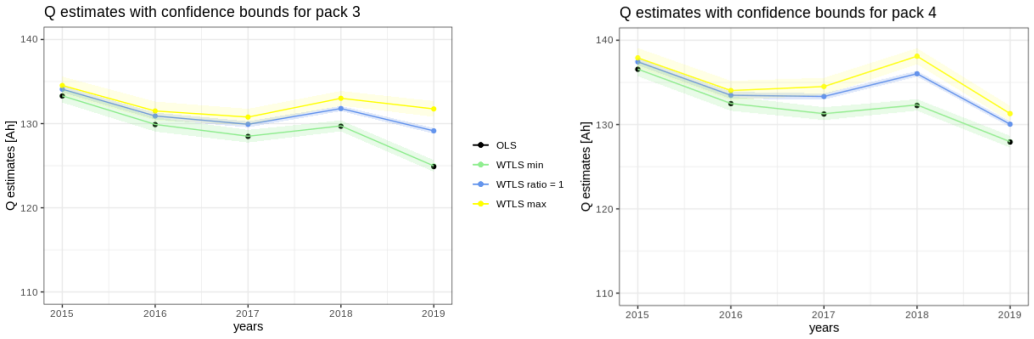
$$ML_{WTLS} = \frac{1}{(2\pi)^n |\mathbf{\Sigma_d}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{d}-\hat{\mathbf{d}})^T \mathbf{\Sigma_d}^{-1}(\mathbf{d}-\hat{\mathbf{d}})\right) \tag{7.5}$$

According to Cramer-Rao theorem, the lower bound to the variance of $\hat{Q}$, denoted by $\sigma^2_{\hat{Q}}$, is given by the negative inverse of the second derivative of the argument of the exponential function, evaluated at the $\hat{Q}$ that minimizes χ^2 or maximizes ML_{WTLS} (Plett 2011).



(a) Pack 1 (b) Pack 2

Figure 7.9: Capacity estimates for packs 1 and 2 from OLS and WTLS with varying measurement error variance ratios including confidence bounds



(a) Pack 3 (b) Pack 4

Figure 7.10: Capacity estimates for packs 3 and 4 from OLS and WTLS with varying measurement error variance ratios including confidence bounds

Hence,

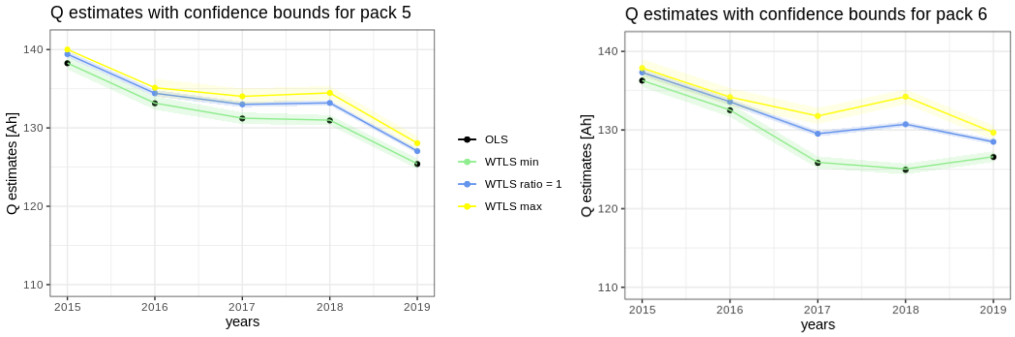
$$\sigma^2_{\hat{Q}} \geq 2\left(\frac{\partial \chi^2_{WTLS}}{\partial \hat{Q}^2}\right). \tag{7.6}$$

Then we can compute confidence intervals as three-sigma bounds

$$(\hat{Q} - 3\sigma_{\hat{Q}}, \hat{Q} + 3\sigma_{\hat{Q}}) \tag{7.7}$$

which with a high probability contain the true total capacity Q.

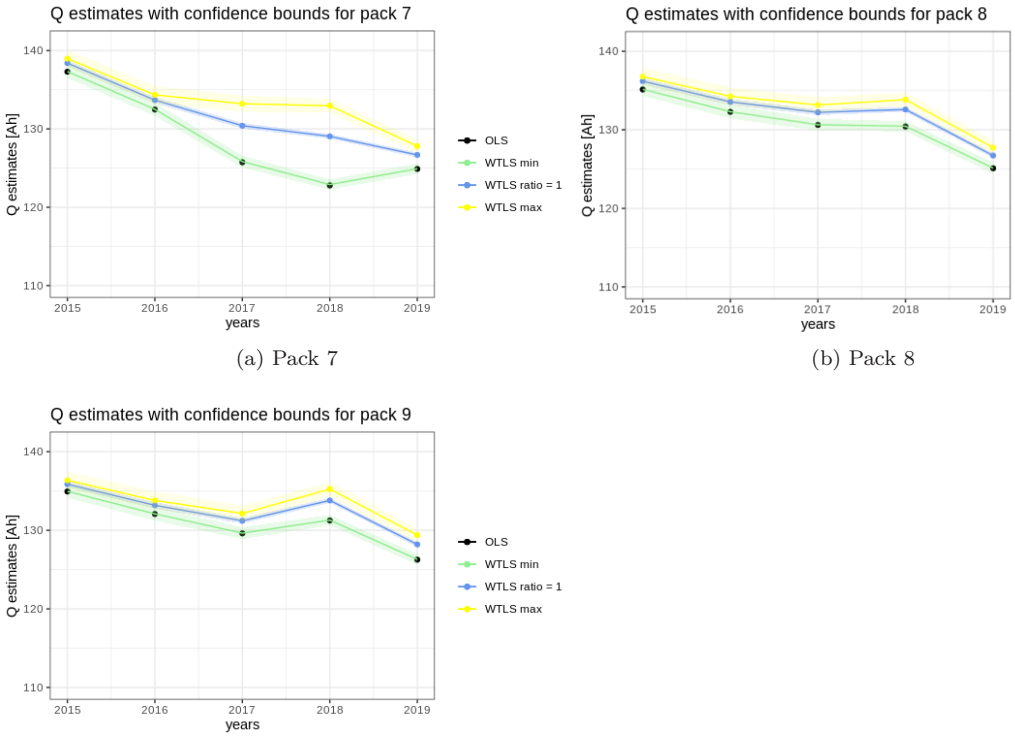
Figures 7.9, 7.10, 7.11 and 7.12 illustrate the OLS estimates of capacity Q for all packs between 2015 and 2019 as well as WTLS estimates for the same three measurement error variance ratios ϕ as in Figure 7.8. We have plotted the results together with their approximate confidence intervals using Equations 7.6 and 7.7, but these intervals are likely not very informative, as we do not know if the measurement errors in the real data are normally distributed. Plett (2011) has presented his results with such confidence intervals, but that is on



(a) Pack 5 (b) Pack 6

Figure 7.11: Capacity estimates for packs 5 and 6 from OLS and WTLS with varying measurement error variance ratios including confidence bounds

simulated data, where this is controlled. We see that the battery Packs 3, 4, 5, 8 and 9 show similar decreasing trends for both OLS and WTLS estimates.



(a) Pack 7 (b) Pack 8

(c) Pack 9

Figure 7.12: Capacity estimates for packs 7-9 from OLS and WTLS with varying measurement error variance ratios including confidence bounds

In the years 2015 and 2016 the OLS and WTLS estimates are relatively close for all packs, but the gap between them widens over time and hence their difference becomes more pronounced. The other packs show unexpected behavior as the OLS estimates for these packs decrease greatly in years 2017 and 2018, and then increase again. This is similar to the anomaly we described in the beginning of this chapter. We suspect it may again have been caused by faulty sensors or a replacement of battery components. The effect of WTLS compared to OLS is even greater here as WTLS is able to correct for the steep decrease if we assume a higher measurement error variance ratio.

Additional information is required to assess the reliability of our estimates, such as yearly independent capacity tests. We received results from one such annual test and will check our estimates against these measured values in the next section.

7.6 SOH annual test comparison

We have one annual test from January 6th 2017 which we can use to assess how well our methods estimate the total capacity. The annual test in Table 7.11 shows SOH values in percentage for each of the 9 packs. In order to get comparable estimates, we will apply WTLS method to data from December 2016 with the measurement error variances $\sigma^2_{\Delta x}, \sigma^2_{\Delta y}$ within the range $[0.1, 1]$ as we showed before that they give better goodness of fit and smaller range of possible estimates $\hat{Q}$. We will then transform the obtained capacity estimates $\hat{Q}$ to obtain SOH by diving them by nominal capacity 150 Ah.

Pack	SOH test
1	92.7
2	92.0
3	91.5
4	92.1
5	92.0
6	92.4
7	92.0
8	91.7
9	91.9

Table 7.11: SOH annual test from January 6th 2017

For pack 1 our results do not agree with the annual test, which measured an SOH value of 92.7%. OLS gave an estimate of SOH equal to 81.7% and WTLS estimates spun over a wide range from 83% to 90.9%, all below the test value. This may again be due to the faulty current sensors discussed in the beginning of this chapter.

For pack 2 the annual test measured SOH of 92% while the OLS gave an estimate of 90.6%, underestimating the test value by 1.4 %. WTLS results for pack 2 are shown in Table 7.13. Again they are all higher than OLS,

	$\sigma^2_{\Delta x}$						
$\sigma^2_{\Delta y}$	0.1	0.2	0.4	0.5	0.6	0.8	1
0.1	87.7	89.2	90.2	90.4	90.6	90.8	90.9
0.2	86	87.7	89.2	89.6	89.8	90.2	90.4
0.4	84.4	86	87.7	88.2	88.6	89.2	89.6
0.5	84	85.5	87.2	87.7	88.1	88.8	89.2
0.6	83.7	85	86.7	87.3	87.7	88.4	88.8
0.8	83.2	84.4	86	86.6	87	87.7	88.2
1.0	83	84	85.5	86	86.5	87.2	87.7
	$SOH_{OLS} = 81.7\%$						
	$SOH_{test} = 92.7\ \%$						

Table 7.12: WTLS estimates of SOH in % from pack 1 in December 2016 for varying uncertainties in x and y

	$\sigma^2_{\Delta x}$						
$\sigma^2_{\Delta y}$	0.1	0.2	0.4	0.5	0.6	0.8	1
0.1	92.7	93.2	93.5	93.6	93.6	93.7	93.7
0.2	92.2	92.7	93.2	93.3	93.4	93.5	93.6
0.4	91.6	92.2	92.7	92.9	93	93.2	93.3
0.5	91.5	92	92.6	92.7	92.9	93.1	93.2
0.6	91.3	91.8	92.4	92.6	92.7	92.9	93.1
0.8	91.2	91.6	92.2	92.4	92.5	92.7	92.9
1.0	91.1	91.5	92	92.2	92.3	92.6	92.7
	$SOH_{OLS} = 90.6\%$						
	$SOH_{test} = 92.0\ \%$						

Table 7.13: WTLS estimates of SOH in % from pack 2 in December 2016 for varying uncertainties in x and y

approximately in the range from 91.1% to 93.7%, covering the test result value of 92%. Depending on the measurement error variances, WTLS can both underestimate and overestimate the test value which is approximately in the middle of this range of possible SOH estimates.

	$\sigma^2_{\Delta x}$						
$\sigma^2_{\Delta y}$	0.1	0.2	0.4	0.5	0.6	0.8	1
0.1	90.8	91.2	91.5	91.5	91.6	91.6	91.7
0.2	90.3	90.8	91.2	91.3	91.4	91.5	91.5
0.4	89.8	90.3	90.8	90.9	91	91.2	91.3
0.5	89.7	90.1	90.6	90.8	90.9	91.1	91.2
0.6	89.6	90	90.5	90.7	90.8	91	91.1
0.8	89.5	89.8	90.3	90.5	90.6	90.8	90.9
1.0	89.4	89.7	90.1	90.3	90.4	90.6	90.8
	$SOH_{OLS} = 88.9\ \%$						
	$SOH_{test} = 91.5\ \%$						

Table 7.14: WTLS estimates of SOH in % from pack 3 in December 2016 for varying uncertainties in x and y

	$\sigma^2_{\Delta x}$						
$\sigma^2_{\Delta y}$	0.1	0.2	0.4	0.5	0.6	0.8	1
0.1	93	93.3	93.5	93.6	93.6	93.7	93.7
0.2	92.5	93	93.3	93.4	93.5	93.5	93.6
0.4	92.1	92.5	93	93.1	93.2	93.3	93.4
0.5	92	92.4	92.8	93	93.1	93.2	93.3
0.6	91.9	92.3	92.7	92.8	93	93.1	93.2
0.8	91.8	92.1	92.5	92.7	92.8	93	93.1
1.0	91.7	92	92.4	92.5	92.6	92.8	93
$SOH_{OLS} = 91.3$ %							
$SOH_{test} = 92.1$ %							

Table 7.15: WTLS estimates of SOH in % from pack 4 in December 2016 for varying uncertainties in x and y

Pack 3 has SOH of 91.5% according to the annual test and OLS estimates it to be 88.9%. Table 7.14 shows WTLS results, ranging from 89.4% up to 91.7% for varying uncertainties, which again include the measured test value. Similar results were obtained for pack 4 with SOH value of 92.1% from the annual test. OLS estimates the SOH to be 91.3%, underestimating it by 0.8%. Table 7.15 shows that WTLS gives higher estimates within the range [91.7,93.7]% for varying uncertainties. Similar results were obtained for the rest of the packs and can again be found in Tables A.7, A.8, A.9, A.10 and A.11 in the appendix.

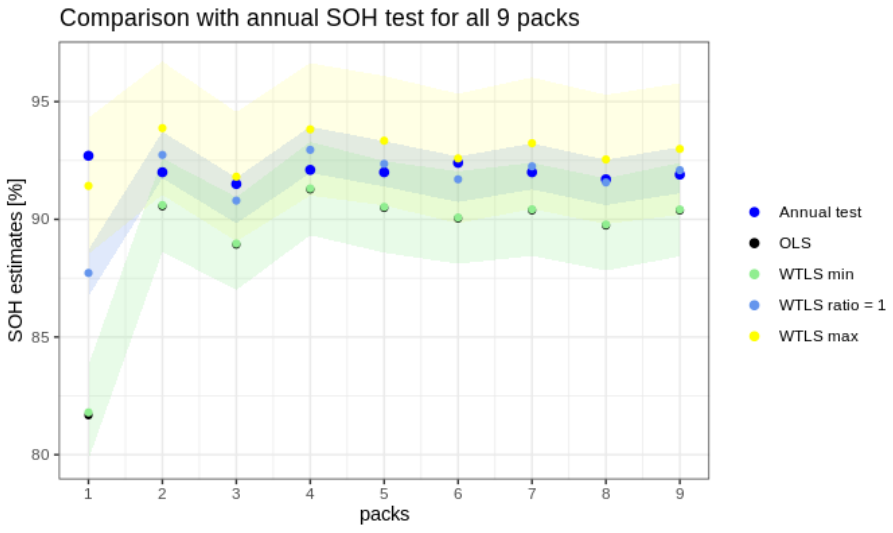


Figure 7.13: Illustration of SOH estimates from OLS and WTLS with confidence intervals compared to the annual test value for all 9 packs

Figure 7.13 summarizes the comparison of the annual test with the OLS and WTLS estimates of SOH for all nine battery packs including the confidence bounds. It shows that OLS underestimates SOH by approximately 2% on average, while the WTLS method, which takes into account meas-

urement errors, gives estimates closer to the test value. With the exception of pack 1, we see that the test value of SOH is included in the confidence interval of estimate obtained for ratio of measurement error variances equal to 1.

From this comparison we see that WTLS has the potential of giving more precise capacity estimates than OLS if we define the measurement uncertainties correctly. However, one annual test is not sufficient evidence to prove good performance of this method. In addition, we need to be aware that the SOH test is not necessarily reliable since it is limited to a few sample points. In particular, there may be variations in for example how the test is performed, the environmental conditions or how well the crew adhere to the test specifications.

CHAPTER 8

Discussion

In this chapter we will discuss the advantages and pitfalls of the total least squares methods as well as its limitations on the SOC estimation algorithm and other interesting future points of research.

8.1 General discussion

The total least squares method we implemented and applied to real battery sensor data offers a simple solution for handling measurement errors and estimating total battery capacity. The WTLS approach which utilizes Newton's method typically requires 4 to 5 iterations to find an optimal estimate which is very computationally efficient. The total least squares method can also be implemented recursively based on the AWTLS approximation and can then be used for real-time monitoring of battery health, which makes it an attractive tool for maritime battery systems. Besides, smaller batches of data are sufficient to get good capacity estimates, which is helpful in maritime settings with limited ship-to-shore connectivity. Although we applied it to battery data on pack level, it can easily be adapted to data on cell or module level.
We came up with algorithms that handle missing data and remove large outliers, but other ways of data cleaning could have been explored to filter away periods of faulty data such as in Figure 7.4.
As we have seen in the results, while naive ordinary linear regression underestimates the true coefficient, total least squares corrects for the attenuation bias and gives more accurate estimates, given one has reliable information on the measurement error variances or is able to estimate them. This method needs the uncertainties in x and y as input and it is not straightforward how to estimate them. We tried to come up with innovative models for these uncertainties in chapter 5 which we illustrated by simulations. However, without some additional information on the SOC estimation algorithm, a high accuracy lab set up to estimate the uncertainties with replicate measurements, or ideally knowledge of the uncertainties themselves, it still remains an open question how to best estimate them.
We have had one annual test to verify our results against, which of course is not enough comparison for a reliable validation of this method. Spot tests such as these annual tests also have measurement errors and may for example not catch seasonal variations in the data. Longer time-series data with several annual

tests are required in order to truly evaluate how well this method performs.

8.2 Error in equation

The statistical literature about measurement error models lists total least squares, also called orthogonal regression, as one alternative way of correcting for measurement error (Carroll, Ruppert et al. 2006, chapter 3). However, they warn about its misuse in the errors-in-variables linear regression because it is modelled from a pure measurement error perspective and does not account for equation error which is an important component of variability (Carroll, Ruppert et al. 2006, chapter 3). Total least squares assumes there is no additional variability about the line in addition to the measurement error, which is rarely the case with real data. Due to this ignorance of equation error, total least squares typically overcorrects for attenuation due to measurement error by exaggerating its effects, and hence overestimates the regression slope (Carroll and Ruppert 1996). It is therefore important to not only estimate the measurement error variances but also to carefully assess the equation error. If the ratio of the measurement error variances in x and y is incorrectly specified, it can result in unacceptably large under- or overcorrection as we have seen in our results in the previous chapter. The statistical literature suggests to run additional studies to observe replicates and use them to estimate the measurement error variances. The method of moments estimator defined in Equation 3.26 may be preferable since it only requires measurement error variance in x, and no equation error or measurement error variance in y (Carroll and Ruppert 1996). Nevertheless, here as well the measurement error variance can be incorrectly estimated when one is not using truly replicate measurements (Carroll and Ruppert 1996).

Our linear model defines capacity as a function of integrated current and change in SOC but it actually depends on other factors such as variations in C-rate and temperature which are not accounted for in our model. These are some of the possible sources of equation error in our data, and by including them in the model, one may be able to obtain more accurate results. Every battery module has several temperature sensors and data collected on these sensors can be included in the models in different ways to get better estimates of the capacity. The variable for temperature could be for example implemented by applying a filter and only considering data within a narrow temperature and C-rate band. When the temperature or C-rate exceed a certain limit, we can remove the measured values and not use them to fit the models. Narrowing down the window for used values and imposing such limits to keep the temperature consistent will reduce the noise and will allow us to get more reliable capacity estimates from the method. One could also try adding other factors as additional regression terms in the regression model, for example as additive terms, multiplicative terms or non-linear terms, and seeing how it affects the results.

8.3 Limitations of the methods on the SOC estimation

As discussed in chapter 2 the estimation of battery capacity with total least squares is very sensitive to how SOC is calculated. The SOC of a battery

cannot be directly measured but has to be derived from externally measurable variables such as voltage, charge and discharge current, which can easily be affected by factors such as temperature, cycle times, discharge rate and voltage (Zhou et al. 2021). This makes it difficult to accurately estimate the SOC in real time. Although our model does not consider how the SOC estimates used to calculate the variables x have been derived, some estimation techniques are more appropriate than others (Plett 2011). For example, the traditional Coulomb counting method based on ampere-hour integration is not suitable because it requires an accurate estimate of total capacity, which is ultimately what we are trying to get with our model. This way it creates a circular dependence between y and x. Plett (2011) prefers to use sigma-point Kalman filters which combine voltage and current data and are insensitive to errors in the capacity value used when making the SOC estimates.

8.4 Coulomb efficiency factor

Another interesting point of future research would be to include the Coulomb efficiency (CE) factor η in Equation 2.8. Coulomb efficiency measures the charge efficiency by which electrons are transferred in batteries (Buchmann 2016, chapter 8). It refers to the ratio of the discharge capacity after a full charge cycle and the charge capacity of the same cycle, and it is usually a fraction of less than 1 (Wang et al. 2021). As lithium-ions move between the anode and cathode during charge and discharge, some are lost to side reactions and prevent the efficiency from reaching 100 percent (Buchmann 2016, chapter 8). This means that the energy retrieved after charging a battery is always less than what had been put in. Lithium-ion batteries as one of the most efficient batteries can have CE of $\eta = 0.99$ or higher, meaning that when the battery is being charged, 99% of charge is actually going in and 1% of charge is lost (Geantil 2020). Overcharging, deep cycling, and extreme temperatures speed up the aging process of a lithium-ion battery and decrease its efficiency further (Geantil 2020). In our model we assumed the Coulomb efficiency factor to be approximately 1, following Plett (2011), and omitting it from the current integration. Yet when we ignore the Coulomb efficiency, the response variables y_i we obtain by integrating current measurements are not directly comparable since the factor η can have different values depending on whether the battery is being charged or discharged. Hence the actual charge or discharge may be different from what we compute. It would be beneficial for the modelling to differentiate between the two processes of charging and discharging. One way of taking into account this factor and getting more precise capacity estimates would be to look at charging and discharging periods separately. Then one would obtain two capacity estimates, for charging and discharging each. In our work, we assume that their difference is negligible and the capacity estimate we obtain can then be regarded as the average between the charge and discharge capacity. It would be interesting to extend the model in future work by looking at charge and discharge intervals separately and making a distinction between these two quantities.

CHAPTER 9

Conclusion

In this thesis, we have studied improved methods of estimating total capacity of maritime battery systems based on a linear model between integrated current and the change in state of charge. When there is measurement error in both response and explanatory variables, the traditional methods such as ordinary linear regression, which only take into account measurement error in response variable, are biased and underestimate the true regression coefficient.

The total least squares approach is aimed at handling this issue and correcting for bias due to measurement errors in both x and y. We have seen that total least squares is a simple and computationally efficient method which is able to give more precise estimates of capacity than ordinary linear regression.

An important part of measurement error modelling is having good estimates of the measurement error variances. These can for example be obtained through replicates generated in a high accuracy lab. We have shown that the measurement error variance in x can be bounded by an upper limit, and that the measurement error variance in y increases with the length of the integration interval. We note that numerical integration error is an additional component of error in the variable y which we are unable to quantify without a proper lab setup.

In case of insufficient information about the exact values of the measurement uncertainties, a rule of thumb based on the chi-square goodness of fit criterion can help narrow down the range of reasonable values for these uncertainties, and hence of the capacity estimates. Finally, we note that longer time-series data including annual tests is necessary for reliable evaluation of this method's performance.

Altogether, we find this method to provide an attractive solution to battery capacity estimation in maritime settings since it can be implemented recursively for real time application and used for online battery health monitoring. Further adaptations can make this method even more precise by for example incorporating additional factors such as temperature or Coulomb efficiency.

APPENDIX A

Tables

	$\sigma^2_{\Delta x}$						
$\sigma^2_{\Delta y}$	0.1	0.5	1	2	5	10	12
0.1	129.14	131.03	131.39	131.58	131.70	131.74	131.75
0.5	126.54	129.14	130.16	130.87	131.39	131.58	131.61
1	125.82	127.95	129.14	130.16	131.03	131.39	131.45
2	125.39	126.84	127.95	129.14	130.42	131.03	131.15
5	125.10	125.82	126.54	127.57	129.14	130.16	130.38
10	124.99	125.39	125.82	126.54	127.95	129.14	129.44
12	124.97	125.31	125.68	126.32	127.64	128.84	129.14
	$\hat{Q}_{OLS} = 124.87$ Ah						

Table A.1: WTLS results for pack 3 in 2019 for varying uncertainties in x and y

	$\sigma^2_{\Delta x}$						
$\sigma^2_{\Delta y}$	0.1	0.5	1	2	5	10	12
0.1	130.04	130.95	131.12	131.22	131.28	131.29	131.30
0.5	128.76	130.04	130.53	130.87	131.12	131.22	131.23
1	128.4	129.46	130.04	130.53	130.95	131.12	131.15
2	128.18	128.91	129.46	130.04	130.66	130.95	131.01
5	128.03	128.4	128.76	129.27	130.04	130.53	130.64
10	127.98	128.18	128.4	128.76	129.46	130.04	130.18
12	127.97	128.14	128.33	128.65	129.31	129.89	130.04
	$\hat{Q}_{OLS} = 127.92$ Ah						

Table A.2: WTLS results for pack 4 in 2019 for varying uncertainties in x and y

	$\sigma^2_{\Delta x}$						
$\sigma^2_{\Delta y}$	0.1	0.5	1	2	5	10	12
0.1	128.49	129.34	129.5	129.59	129.65	129.66	129.67
0.5	127.32	128.49	128.95	129.27	129.5	129.59	129.61
1	126.99	127.96	128.49	128.95	129.34	129.5	129.53
2	126.79	127.46	127.96	128.49	129.07	129.34	129.39
5	126.66	126.99	127.32	127.79	128.49	128.95	129.05
10	126.61	126.79	126.99	127.32	127.96	128.49	128.62
12	126.6	126.75	126.93	127.22	127.82	128.36	128.49
	$\hat{Q}_{OLS} = 126.56$ Ah						

Table A.3: WTLS results for pack 6 in 2019 for varying uncertainties in x and y

	$\sigma^2_{\Delta x}$						
$\sigma^2_{\Delta y}$	0.1	0.5	1	2	5	10	12
0.1	126.68	127.49	127.64	127.72	127.78	127.79	127.8
0.5	125.58	126.68	127.11	127.41	127.64	127.72	127.74
1	125.28	126.18	126.68	127.11	127.49	127.64	127.67
2	125.1	125.71	126.18	126.68	127.22	127.49	127.53
5	124.97	125.28	125.58	126.02	126.68	127.11	127.20
10	124.93	125.1	125.28	125.58	126.18	126.68	126.55
12	124.92	125.06	125.22	125.49	126.05	126.55	126.68
	$\hat{Q}_{OLS} = 124.87$ Ah						

Table A.4: WTLS results for pack 7 in 2019 for varying uncertainties in x and y

	$\sigma^2_{\Delta x}$						
$\sigma^2_{\Delta y}$	0.1	0.5	1	2	5	10	12
0.1	126.72	127.46	127.59	127.67	127.72	127.73	127.74
0.5	125.73	126.72	127.12	127.39	127.59	127.67	127.68
1	125.46	126.27	126.72	127.12	127.46	127.59	127.62
2	125.29	125.85	126.27	126.72	127.22	127.46	127.50
5	125.18	125.46	125.73	126.12	126.72	127.12	127.20
10	125.14	125.29	125.46	125.73	126.27	126.72	126.84
12	125.13	125.26	125.4	125.65	126.15	126.61	126.72
	$\hat{Q}_{OLS} = 125.09$ Ah						

Table A.5: WTLS results for pack 8 in 2019 for varying uncertainties in x and y

	$\sigma^2_{\Delta x}$						
$\sigma^2_{\Delta y}$	0.1	0.5	1	2	5	10	12
0.1	128.2	129.06	129.22	129.31	129.36	129.38	129.38
0.5	127.02	128.2	128.66	128.98	129.22	129.31	129.32
1	126.69	127.66	128.2	128.66	129.06	129.22	129.25
2	126.49	127.16	127.66	128.2	128.78	129.06	129.11
5	126.35	126.69	127.02	127.49	128.2	128.66	128.76
10	126.31	126.49	126.69	127.02	127.66	128.2	128.33
12	126.3	126.45	126.63	126.92	127.52	128.06	128.2
$\hat{Q}_{OLS} = 126.26$ Ah							

Table A.6: WTLS results for pack 9 in 2019 for varying uncertainties in x and y

	$\sigma^2_{\Delta x}$						
$\sigma^2_{\Delta y}$	0.1	0.2	0.4	0.5	0.6	0.8	1
0.1	92.4	92.8	93	93.1	93.1	93.2	93.2
0.2	91.9	92.4	92.8	92.9	92.9	93	93.1
0.4	91.4	91.9	92.4	92.5	92.6	92.8	92.9
0.5	91.3	91.7	92.2	92.4	92.5	92.6	92.8
0.6	91.2	91.6	92.1	92.2	92.4	92.5	92.7
0.8	91	91.4	91.9	92	92.2	92.4	92.5
1.0	90.9	91.3	91.7	91.9	92	92.2	92.4
$SOH_{OLS} = 90.5$ %							
$SOH_{test} = 92.0$ %							

Table A.7: WTLS estimates of SOH in % from pack 5 in December 2016 for varying uncertainties in x and y

	$\sigma^2_{\Delta x}$						
$\sigma^2_{\Delta y}$	0.1	0.2	0.4	0.5	0.6	0.8	1
0.1	91.7	92.1	92.3	92.3	92.4	92.4	92.5
0.2	91.3	91.7	92.1	92.1	92.2	92.3	92.3
0.4	90.8	91.3	91.7	91.8	91.9	92.1	92.1
0.5	90.7	91.1	91.6	91.7	91.8	92	92.1
0.6	90.6	91	91.5	91.6	91.7	91.9	92
0.8	90.5	90.8	91.3	91.4	91.5	91.7	91.8
1.0	90.4	90.7	91.1	91.3	91.4	91.6	91.7
$SOH_{OLS} = 90.0$ %							
$SOH_{test} = 92.4$ %							

Table A.8: WTLS estimates of SOH in % from pack 6 in December 2016 for varying uncertainties in x and y

	$\sigma^2_{\Delta x}$						
$\sigma^2_{\Delta y}$	0.1	0.2	0.4	0.5	0.6	0.8	1
0.1	92.3	92.6	92.9	93	93	93.1	93.1
0.2	91.8	92.3	92.6	92.7	92.8	92.9	93
0.4	91.3	91.8	92.3	92.4	92.5	92.6	92.7
0.5	91.2	91.6	92.1	92.3	92.4	92.5	92.6
0.6	91.1	91.5	92	92.1	92.3	92.4	92.6
0.8	90.9	91.3	91.8	91.9	92.1	92.3	92.4
1.0	90.8	91.2	91.6	91.8	91.9	92.1	92.3
	$SOH_{OLS} = 90.4$ %						
	$SOH_{test} = 92.0$ %						

Table A.9: WTLS estimates of SOH in % from pack 7 in December 2016 for varying uncertainties in x and y

	$\sigma^2_{\Delta x}$						
$\sigma^2_{\Delta y}$	0.1	0.2	0.4	0.5	0.6	0.8	1
0.1	91.6	92	92.2	92.3	92.3	92.4	92.4
0.2	91.1	91.6	92	92.1	92.1	92.2	92.3
0.4	90.6	91.1	91.6	91.7	91.8	92	92.1
0.5	90.5	90.9	91.4	91.6	91.7	91.8	92
0.6	90.4	90.8	91.3	91.4	91.6	91.7	91.9
0.8	90.3	90.6	91.1	91.2	91.4	91.6	91.7
1.0	90.2	90.5	90.9	91.1	91.2	91.4	91.6
	$SOH_{OLS} = 89.7$ %						
	$SOH_{test} = 91.7$ %						

Table A.10: WTLS estimates of SOH in % from pack 8 in December 2016 for varying uncertainties in x and y

	$\sigma^2_{\Delta x}$						
$\sigma^2_{\Delta y}$	0.1	0.2	0.4	0.5	0.6	0.8	1
0.1	92.1	92.4	92.7	92.7	92.8	92.8	92.9
0.2	91.6	92.1	92.4	92.5	92.6	92.7	92.7
0.4	91.2	91.6	92.1	92.2	92.3	92.4	92.5
0.5	91.1	91.5	91.9	92.1	92.2	92.3	92.4
0.6	91	91.4	91.8	92	92.1	92.3	92.4
0.8	90.9	91.2	91.6	91.8	91.9	92.1	92.2
1.0	90.8	91.1	91.5	91.6	91.8	91.9	92.1
	$SOH_{OLS} = 90.4$ %						
	$SOH_{test} = 91.9$ %						

Table A.11: WTLS estimates of SOH in % from pack 9 in December 2016 for varying uncertainties in x and y

APPENDIX B

Code

This chapter includes the most important R code used in this thesis. The total amount of code created for the purpose of this thesis is naturally larger, but for practical purposes only the code for cleaning the data and the total least squares methods we applied in chapter 7 is included.

B.1 Data preparation

First we provide the code required to perform the data cleaning steps described in Chapter 6. The function create_variables takes the raw SOC and current data and returns the x and y data points ready for the analysis. Internally it calls a few helping functions which are also included in the code below. The R packages which were used in this code and which enable efficient handling of large datasets are *dplyr* and *data.table*.

```
create_variables <- function(min,pack,year){
  #Transforms raw SOC and current data into variables x and y and cleans the
     data using help functions
  #Arguments:
    #min: chosen length of integration interval in minutes
    #pack: number of battery pack
    #year: year to read data from
  #Returns:
    #data: data frame containing final x and y variables with the initial
     timestamp of the interval they were computed over in both unix and date
     format
  #reading files from working directory
  soc_filename <- sprintf("soc_pack%d_%d.csv",pack,year)
  current_filename <- sprintf("current_pack%d_%d.csv",pack,year)
  soc <- fread(soc_filename)
  current<- fread(current_filename)
  #converting time stamps to unix format (numeric)
  soc$numtime <- as.numeric(soc$timestamp)
  current$numtime <- as.numeric(current$timestamp)

  x <- create_x(soc,min)
  y <- create_y(current,min)

  #finding time gaps in data longer than 15 minutes
  x_timegaps <- find_soc_timegaps(soc,x,min)
  y_timegaps <- find_current_timegaps(current,y,min)
  timegaps <- merge_timegaps(y_timegaps,x_timegaps)
  #removing variables containing time gaps
  x <- filter(x,!(x$interval_starts %in% timegaps))
```

```
27    y <- filter(y,!(y$interval_starts %in% timegaps))
28
29    #finding current measurements with value equal to 0 and removing the
        corresponding variables
30    zero_intervals <- find_zeroes(y,current,min)
31    zeroes <- get_zero_timestamps(y,zero_intervals,min)
32    x <- filter(x,!(x$interval_starts %in% zeroes))
33    y <- filter(y,!(y$interval_starts %in% zeroes))
34
35    #detecting and removing outliers
36    outliers <- find_outliers(soc,current,y,min)
37    x <- filter(x,!(x$interval_starts %in% outliers))
38    y <- filter(y,!(y$interval_starts %in% outliers))
39
40    data <- data.frame(x$x_val,y$y_val,x$interval_starts,x$timestamp)
41    colnames(data) <- c("x","y","interval_start","timestamp")
42    return(data)
43  }
44
45  #------------------------------------------------------------------------
46  create_x <- function(soc,min){
47    #Creates x variable by linearly interpolating SOC data and computing
        differences in SOC over chosen interval length
48    #Arguments:
49      #soc: SOC data with unix timestamps
50      #min: chosen length of integration interval in minutes (same as for
         current)
51    #Returns:
52      #x: data frame with computed variable x and their initial time stamps in
         numeric and date format
53    time <- soc$numtime[nrow(soc)]-soc$numtime[1]
54    sec <- min*60
55    n <- round(time/sec)
56    #interpolate soc at regular intervals
57    lin.int_soc <-data.frame(approx(soc$numtime, soc$value,method="constant",f
        =0, ties=mean,n=n))
58    x_val <- diff(lin.int_soc$y)
59    interval_starts <- soc$numtime[1]+sec*c(0:(length(x_val)-1))
60    #transforming unix timestamps to date format
61    timestamp <- as.POSIXct(interval_starts, origin ="1970-01-01",tz ="UTC")
62    x <- data.frame(x_val,interval_starts,timestamp)
63    return(x)
64  }
65
66  create_y <- function(current,min){
67    #Numerically integrates the current measurements over regular intervals
68    #Arguments:
69      #current: data frame containing raw current measurements with unix
         timestamps
70      #min: chosen length of integration interval in minutes
71    #Returns:
72      #y: data frame with computed responses y and their initial time stamps in
         numeric and date format
73    #Calculating time intervals between consecutive current measurements
74    intervals <- diff(current$numtime)
75    intervals <- intervals/3600 #converting secs to hours
76
77    #Computing the cumulative integral as left Riemann sum
78    product <- current$value[-nrow(current)]*intervals
79    product <- append(0,product)
80    current$integral <- cumsum(product)
81
```

```
82   #number of intervals of chosen length to split the entire time frame into
83   time <- current$numtime[nrow(current)]-current$numtime[1]
84   sec <- min*60
85   n <- round(time/sec)
86
87   #interpolate current values
88   lin.int_current <-data.frame(approx(current$numtime,current$integral,method=
       "constant",f=0, ties=mean,n=n))
89   y_val <- diff(lin.int_current$y)
90   interval_starts <- current$numtime[1]+sec*c(0:(length(y_val)-1))
91   #transforming unix timestamps to date format
92   timestamp <- as.POSIXct(interval_starts, origin ="1970-01-01",tz ="UTC")
93   y <- data.frame(y_val,interval_starts,timestamp)
94   return(y)
95 }
96
97 #-----------------------------------------------------------------------
98 find_current_timegaps <- function(current,y,min){
99   #Finds time windows in the SOC data longer than 15 minutes
100  #Arguments:
101    #current: raw current data
102    #y: data frame of variable y including initial timestamps
103    #min: integration interval length
104  #Returns:
105    #y_gaps_list: a list of initial timestamps for y variables which are
       inside any timegaps found in current data
106  intervals <- diff(current$numtime)
107  intervals <- intervals/3600
108  time_gap_start <- which(intervals>0.4)
109  time_gap_end <- time_gap_start + 1
110  start <- current$numtime[time_gap_start]
111  end <- current$numtime[time_gap_end]
112  current_gaps <- data.frame(start,end)
113  if (nrow(current_gaps)>0){
114    y_gaps_list <- NULL
115    for (i in (1:nrow(current_gaps))){
116      y_timegaps_middle <- y$interval_starts[current_gaps$start[i]<y$interval_
       starts & current_gaps$end[i]>(y$interval_starts+60*min)]
117      y_timegaps_left <- y$interval_starts[current_gaps$start[i]>=y$interval_
       starts & current_gaps$start[i]<=(y$interval_starts+60*min)]
118      y_timegaps_right <- y$interval_starts[current_gaps$end[i]>=y$interval_
       starts & current_gaps$end[i]<=(y$interval_starts+60*min)]
119      y_timegaps <- c(y_timegaps_left,y_timegaps_middle,y_timegaps_right)
120      y_timegaps <- unique(y_timegaps)
121      y_gaps_list <- c(y_gaps_list, y_timegaps)
122    }
123    return(y_gaps_list)
124  }
125  else return("No timegaps found")
126 }
127
128 find_soc_timegaps <- function(soc,x,min){
129   #Finds time windows in the SOC data longer than 15 minutes
130   #Arguments:
131     #soc: raw SOC data
132     #x: data frame of variable x including initial timestamps
133     #min: integration interval length
134   #Returns:
135     #x_gaps_list: a list of initial timestamps for x variables which are
        inside any timegaps found in SOC data
136   intervals <- diff(soc$numtime)
137   intervals <- intervals/3600
```

```
138   time_gap_start <- which(intervals>0.25)
139   time_gap_end <- time_gap_start + 1
140   start <- soc$numtime[time_gap_start]
141   end <- soc$numtime[time_gap_end]
142   soc_gaps <- data.frame(start,end)
143   if (nrow(soc_gaps)>0){
144     x_gaps_list <- NULL
145     for (i in (1:nrow(soc_gaps))){
146       x_timegaps_middle <- x$interval_starts[soc_gaps$start[i]<x$interval_
      starts & soc_gaps$end[i]>(x$interval_starts+60*min)]
147       x_timegaps_left <- x$interval_starts[soc_gaps$start[i]>=x$interval_
      starts & soc_gaps$start[i]<=(x$interval_starts+60*min)]
148       x_timegaps_right <- x$interval_starts[soc_gaps$end[i]>=x$interval_starts
       & soc_gaps$end[i]<=(x$interval_starts+60*min)]
149       x_timegaps <- c(x_timegaps_left,x_timegaps_middle,x_timegaps_right)
150       x_timegaps <- unique(x_timegaps)
151       x_gaps_list <- c(x_gaps_list, x_timegaps)
152     }
153     return(x_gaps_list)
154   }
155   else return ("No timegaps found")
156 }
157 
158 merge_timegaps <- function(y_timegaps, x_timegaps){
159   #Merges the timegaps found in SOC and current data and returns a list of
        initial timestamps for variables x and y inside these timegaps
160   timegaps <- c(y_timegaps,x_timegaps)
161   timegaps <- unique(timegaps)
162   return(timegaps)
163 }
164 
165 #-------------------------------------------------------------------------
166 
167 find_zeroes <- function(y,current,min){
168   #Finds intervals where the current has value 0 over a longer period
169   #Arguments:
170     #y: y variable with initial time stamps
171     #current: raw current data
172     #min: integration interval length in minutes
173   #Returns:
174     #time_ind: data frame containing first and last timestamp of the period
       with current equal to 0
175   time_ind <- data.frame(from = numeric(0),to = numeric(0))
176   y_zero <- y[y$y_val == 0,]
177   n <- nrow(y_zero)
178   print(n)
179   more <- TRUE
180   i <- 1
181   while(more){
182     from <- y_zero$interval_starts[i]
183     while(i+1<= n & y_zero$interval_starts[i+1] - y_zero$interval_starts[i] ==
        (60*min)){
184       i <- i+1
185       if (i%%1000 == 0){
186         print(i)
187       }
188     }
189     to <-  y_zero$interval_starts[i]+ min*60
190     curr <- current[current$numtime >= from & current$numtime<to,]
191     if (all(curr$value == 0)){
192       time_ind <- rbind(time_ind,c(from,to))
193     }
```

```
194     i <- i+1
195     more <- (i<= n)
196 
197   }
198   colnames(time_ind) <- c("from","to")
199   return(time_ind)
200 }
201 
202 get_zero_timestamps <- function(y,zero_intervals,min){
203   #Finds initial timestamps of x and y variables which were computed over
        longer periods with current equal to 0
204   zero_timestamps <- NULL
205   for (i in (1:nrow(zero_intervals))){
206     timestamps <- y$interval_starts[zero_intervals$from[i]<= y$interval_starts
        & zero_intervals$to[i]>(y$interval_starts+60*min)]
207     zero_timestamps <- c(zero_timestamps,timestamps)
208   }
209   return(zero_timestamps)
210 }
211 
212 #-----------------------------------------------------------------------
213 find_outliers <- function(soc,current,y,min){
214   #Detects outliers in the raw current and SOC data and returns timestamps of
        x and y variables which were computed over time periods with these
        outliers
215 
216   curr_deltas <- diff(current$value)
217   soc_deltas <- diff(soc$value)
218   soc_outliers <- which(diff(sign(soc_deltas))!=0  & abs(soc_deltas[-1])>30 &
        abs(soc_deltas[-length(soc_deltas)])>30)
219   curr_outliers <- which(diff(sign(curr_deltas))!=0 & abs(curr_deltas[-1])>200
        & abs(curr_deltas[-length(curr_deltas)])>200)
220 
221   soc_outliers <- soc_outliers + 1
222   curr_outliers <- curr_outliers + 1
223 
224   soc_out_timestamps <- soc$numtime[soc_outliers]
225   curr_out_timestamps <- current$numtime[curr_outliers]
226 
227   outliers <- c(soc_out_timestamps,curr_out_timestamps)
228   outliers <- unique(outliers)
229 
230   if (length(outliers)>0){
231     out_list <- NULL
232     for (i in (1:length(outliers))){
233       out <- y$interval_starts[outliers[i]>= y$interval_starts & outliers[i]<(
        y$interval_starts+60*min)]
234       out_list <- c(out_list,out)
235     }
236     return(out_list)
237   }
238   else return("No outliers found")
239 }
```

Listing B.1: Data cleaning and creation of variables x and y

B.2 WTLS,TLS and AWTLS functions

We also provide functions for the three methods proposed by Plett which we described in detail in chapter 4. WTLS, TLS and AWTLS take the variables x

and y and standard deviations of measurement errors as input and return the optimal estimate of capacity $\hat{Q}$ as well as upper and lower confidence bounds of the estimate.

```
wtls <- function(x,y,sigma_x,sigma_y) {
    #Performs TLS to find a capacity estimate Q
    #Arguments:
        #x: vector of explanatory variable x
        #y: vector of response variable y
        #sigma_x <-  a numeric vector of measurement error standard deviations in x
        #sigma_y <- a numeric vector of measurement error standard deviations in y
    #Returns:
        #Q_est: optimal capacity estimate Q
        #upper_CI: upper confidence bound for the capacity estimate
        #lower_CI: lower confidence bound for the capacity estimate
    #initial value for capacity Q from ordinary linear regression
    Q_init <-  lm(y ~ x-1)$coefficients
    Q_list <- Q_init
    Numit <- 0 #number of iterations
    eps <- 1e-10 #absolute stopping criterion
    more = TRUE
    while(more)
    {
      #first derivative of the merit function
      first_der <- sum(2*(Q*x-y)*(Q*y*sigma_x^2+x*sigma_y^2)/((Q^2*sigma_x^2 + sigma_y^2)^2))
      #second derivative of the merit function
      second_der <- 2*sum((x^2*sigma_y^4 + sigma_x^4*(3*Q^2*y^2 - 2*Q^3*x*y) - sigma_x^2*sigma_y^2*(3*Q^2*x^2 - 6*Q*x*y + y^2))/((Q^2*sigma_x^2 + sigma_y^2)^3))
      #iteration based on Newton's method
      Q_new <- Q - first_der/second_der
      more = abs(Q-Q_new) > eps
      Q <- Q_new
      Numit <- Numit+1
      if (more == TRUE) {
        Q_list <- c(Q_list, Q_new)
      }

    }
  Q_est <- Q_new
  #computing merit function with the obtained capacity estimate
  merit <- sum((y-Q_est*x)^2/(Q_est^2*sigma_x^2+sigma_y^2))
  #Confidence intervals as 3-sigma bounds
  Hessian <-2*(sum((x^2*sigma_y^4 + sigma_x^4*(3*Q_est^2*y^2 -
  2*Q_est^3*x*y) - sigma_x^2*sigma_y^2*(3*Q_est^2*x^2 -
  6*Q_est*x*y + y^2))/(Q_est^2*sigma_x^2 + sigma_y^2)^3))
  sigma_Q <- sqrt(2/Hessian)
  lower_CI <- Q_est - 3*sigma_Q
  upper_CI <- Q_est + 3*sigma_Q
  res <- c(lower_CI, Q_est,upper_CI,merit)
  return(res)
}
```

Listing B.2: Weighted total least squares method using Newton-Raphson search

```
tls <- function(x,y,sigma_x,sigma_y) {
  #Performs TLS to find a capacity estimate Q
  #Arguments:
    #x: vector of explanatory variable x
```

```
    #y: vector of response variable y
    #sigma_x <-  a numeric vector of measurement error standard deviations in
     x
    #sigma_y <- a numeric vector of measurement error standard deviations in y
  #Returns:
    #Q_est: optimal capacity estimate Q
    #upper_CI: upper confidence bound for the capacity estimate
    #lower_CI: lower confidence bound for the capacity estimate
  k <- sigma_x/sigma_y #ratio of measurement error st.deviations
  c1 <- sum(x^2/sigma_y^2)
  c2 <- sum(x*y/sigma_y^2)
  c3 <- sum(y^2/sigma_y^2)
  #roots of derivative of merit function are candidate solutions for Q
  first_root <- (-(c1-k^2*c3)+ sqrt((c1-k^2*c3)^2 + 4*k^2*c2^2))/(2*k^2*c2)
  second_root <- (-(c1-k^2*c3)- sqrt((c1-k^2*c3)^2 + 4*k^2*c2^2))/(2*k^2*c2)
  #choosing the positive root as the capacity estimate Q
  if (first_root>0) {
    Q_est <- first_root
  } else {
    Q_est <- second_root
  }
  #Confidence intervals as 3-sigma bounds
  Hessian <- (-4*k^4*c2*Q_est^3+6*k^4*c3*Q_est^2+(-6*c1+12*c1)*k^2*Q_est +2*(
     c1-k^2*c3)/(Q_est^2*k^2+1)^3)
  sigma_Q <- sqrt(2/Hessian)
  lower_CI <- Q_est - 3*sigma_Q
  upper_CI <- Q_est + 3*sigma_Q
  res <- c(lower_CI, Q_est,upper_CI)
  return(res)
}
```

Listing B.3: Total least squares method with proportional uncertainties

```
awtls <- function(x,y,sigma_x,sigma_y) {
  #Performs AWTLS method to find a capacity estimate Q
  #Arguments:
    #x: vector of explanatory variable x
    #y: vector of response variable y
    #sigma_x <-  a numeric vector of measurement error standard deviations in
     x
    #sigma_y <- a numeric vector of measurement error standard deviations in y
  #Returns:
    #Q_est: optimal capacity estimate Q
  #defining running sums to simplify the formulas
  c1 <- sum(x^2/sigma_y^2)
  c2 <- sum(x*y/sigma_y^2)
  c3 <- sum(y^2/sigma_y^2)
  c4 <- sum(x^2/sigma_x^2)
  c5 <- sum(x*y/sigma_x^2)
  c6 <- sum(y^2/sigma_x^2)
  #computing the roots of the derivative of the merit function which is a
     quartic polynomial
  poly_coeff <- c(-c2, c1-2*c3+c6, 3*c2-3*c5, 2*c4-c1-c6, c5)
  roots <- polyroot(poly_coeff)
  #the positive roots are candidate solutions for Q
  roots <- roots[which(Re(roots)>0)]
  roots <- Re(roots)
  merit <- NULL
  #computing the value of the merit function for all candidates Q
  for (Q in roots){
    merit <- c(merit, sum(((y-Q*x)^2/(1+Q^2)^2)*((Q^2/sigma_x^2)+(1/sigma_y^2)
     )))
```

```
    # alternatively computing merit function using the recursive sums
    #merit <- c(merit, (1/(Q^2+1)^2)*(c4*Q^4 - 2*c5*Q^3+(c1+c6)*Q^2-2*c2*Q+c3)
     )
  }
  #choosing the root which minimizes the merit function
  Q_est <- roots[which.min(merit)]
  return(Q_est)
}
```

Listing B.4: Approximate weighted total least squares method

Bibliography

Bevington, P. R. and Robinson, D. K. (1992). *Data reduction and error analysis for the physical sciences*. Second. With 1 IBM-PC floppy disk (5.25 inch; HD). McGraw-Hill, Inc., New York, pp. xx+328.

Buchmann, I. (2016). *Batteries in a Portable World - A Handbook on Rechargeable Batteries for Non-Engineers*. Cadex Electronics Inc., p. 360.

Buonaccorsi, J. P. (2010). *Measurement error*. Interdisciplinary Statistics. Models, methods, and applications. CRC Press, Boca Raton, FL, pp. xxvi+437.

Carroll, R. J. and Ruppert, D. (1996). 'The Use and Misuse of Orthogonal Regression in Linear Errors-in-Variables Models'. In: *The American statistician* vol. 50 (1), pp. 1–6.

Carroll, R. J., Ruppert, D. et al. (2006). *Measurement error in nonlinear models*. Second. Vol. 105. Monographs on Statistics and Applied Probability. A modern perspective. Chapman & Hall/CRC, Boca Raton, FL, pp. xxviii+455.

Cornbleet, J. and Gochman, N. (1979). 'Incorrect Least-Squares Regression Coefficients in Method-Comparison Analysis'. In: *Clinical Chemistry* vol. 25, pp. 432–438.

Departementene (2022). *Handlingsplan for grønn skipsfart*. URL: https://www.regjeringen.no/no/dokumenter/handlingsplan-for-gronn-skipsfart/id2660877/ (visited on 08/05/2022).

Geantil, P. (2020). *Top 5 Factors That Affect Industrial Battery Efficiency*. URL: https://www.fluxpower.com/blog/top-5-factors-that-affect-industrial-battery-efficiency (visited on 05/05/2022).

Korthauer, R. (2018). *Lithium-Ion Batteries: Basics and Applications*. Springer Berlin Heidelberg, pp. xx + 413.

Li, Y. et al. (2019). 'Data-driven health estimation and lifetime prediction of lithium-ion batteries: A review'. In: *Renewable sustainable energy reviews* vol. 113, p. 109254.

Lu, L. et al. (2013). 'A review on the key issues for lithium-ion battery management in electric vehicles'. In: *Journal of Power Sources* vol. 226, pp. 272–288.

Markovsky, I. and Van Huffel, S. (2007). 'Overview of total least-squares methods'. In: *Signal processing* vol. 87, pp. 2283–2302.

Mikolajczak, C. et al. (2012). *Lithium-Ion Batteries Hazard and Use Assessment*. SpringerBriefs in fire. New York, NY: Springer-Verlag, pp. xii+115.

Movassagh, K. et al. (2021). 'A critical look at coulomb counting approach for state of charge estimation in batteries'. In: *Energies* vol. 14, p. 4074.

Pešta, M. (2018). 'Total Least Squares Approach in Regression Methods'. In: *WDS'08 Proceedings of Contributed Papers, Part I*, pp. 88–93.

Pešta, M. (2013). 'Total least squares and bootstrapping with applications in calibration'. In: *Statistics* vol. 47, no. 5, pp. 966–991.

Plett, G. L. (2011). 'Recursive approximate weighted total least squares estimation of battery cell total capacity'. In: *Journal of Power Sources* vol. 196, pp. 2319–2331.

Press, W. H. (1992). *Numerical recipes in C: the art of scientific computing*. Cambridge University Press, pp. xxvi+ 994.

Shang, Y. (2012). 'Measurement Error Adjustment Using the SIMEX Method: An Application to Student Growth Percentiles: Measurement Error Adjustment Using the SIMEX Method'. In: *Journal of educational measurement* vol. 49, pp. 446–465.

Stefanski, L. A. (2000). 'Measurement error models'. In: *J. Amer. Statist. Assoc.* vol. 95, no. 452, pp. 1353–1358.

Team, M. E. V. (2008). *A Guide to Understanding Battery Specifications*. URL: http://mit.edu/evt/summary_battery_specifications.pdf (visited on 26/04/2022).

Van Huffel, S. and Lemmerling, P. (2002). *Total Least Squares and Errors-in-Variables: Modeling Analysis, Algorithms and Applications*. Springer Science+Business Media Dordrecht, pp. x+397.

Vanem, E. et al. (2021). 'Data-driven state of health modelling—A review of state of the art and reflections on applications formaritime battery systems'. In: *Journal of Energy Storage* vol. 43, p. 103158.

Wang, S. et al. (2021). *Battery System Modeling*. Elsevier, p. 354.

Zhang, J. and Lee, J. (2011). 'A review on prognostics and health monitoring of Li-ion battery'. In: *Journal of Power Sources* vol. 196, pp. 6007–6014.

Zhou, W. et al. (2021). 'Review on the battery model and SOC estimation method'. In: *Processes* vol. 9, p. 1685.