Automatic method for the estimation of li-ion degradation test sample sizes required to understand cell-to-cell variability

Automatic method for the estimation of li-ion degradation test sample sizes required to understand cell-to-cell variability

Calum Strange[a], Michael Allerhand[a], Philipp Dechent[b], Gonçalo dos Reis[a,c,*]

[a] *School of Mathematics, University of Edinburgh, The Kings buildings, Edinburgh, EH9 3JF, Scotland*
[b] *Institute for Power Electronics and Electrical Drives (ISEA), RWTH Aachen University, Aachen, Germany*
[c] *Centro de Matematica e Aplicacoes (CMA), Faculdade de Ciencias e Tecnologia, Campus da Caparica, Caparica, 2829-516, Portugal*

Abstract

The testing of battery cells is an expensive and long process, and hence understanding how large a test set needs to be is very useful. This work proposes an automated methodology to estimate the smallest sample size of cells required to capture the cell-to-cell variability seen in a larger population. We define cell-to-cell variation based on the slopes of a linear regression model applied to capacity fade curves. Our methodology determines a sample size which estimates this variability within user specified requirements on precision and confidence. The sample size is found using the distributional properties of the slopes under a normality assumption. The implementation is available on GitHub.

For the five datasets in the study, we find that a sample size of 8-10 cells (at a prespecified precision and confidence) captures the cell-to-cell variability of the larger datasets. We show that prior testing knowledge can be leveraged with machine learning models to operationally optimise the design of new cell-testing leading up to a 75% reduction in experimental costs.

Keywords: battery, testing, lithium-ion, degradation, statistics, manufacturing, machine learning

*Corresponding author

1. Introduction

Current lithium-ion cells do not yet meet some applications' required power and energy densities. Therefore, research on new materials is dynamic, and ageing behaviour is an essential part of the evaluation. Furthermore, new applications for batteries with particular requirements - such as the electrification of ships, trucks [1], aircraft [2], tractors and construction machinery - often bring new load profiles, which will most likely induce a different ageing behaviour. Therefore, adapted test and evaluation methods are necessary for a reliable lifetime prediction. Batteries are highly complex systems with physical, chemical and electrical effects taking place simultaneously requiring a lot more effort to accurately model these effects themselves. And thus, in the short and medium term data-driven and empirical models will be used. These methods include diagnostic methods such as deep learning or neural networks based on systematically generated data from accelerated ageing tests in the laboratory [3].

The ageing of lithium-ion batteries depends on the complex interaction of numerous stress factors such as current rate and temperature, which necessitates an extensive test matrix. In addition, the transfer of test results to new batteries with varying materials and dimensions to create models for new cells is very limited. Currently, complex testing is carried out on a small scale on random samples due to the lack of testing resources. This limits the scope of a test regarding the number of different stress factors, the resolution of the influence, and the statistical aspects of cell-to-cell variation.

The ageing is primarily noticeable to the user as lower capacity and thus shorter operating time [4]. Many different stress factors need to be considered for ageing prediction and testing. These factors are, e.g. temperature, storage voltages for calendar ageing as well as cycle depth, state of charge (SoC) range, mechanical pressure, current rate and charge throughput for (charge/discharge) cycle ageing [5, 6].

Ageing takes place in all components of a battery, not only in the electrodes and electrolyte, but also in the casing and separator [7]. The mentioned stress factors influence ageing in electrodes and electrolytes. For example, the dissolution of electrolyte and binder as well as the reduction of the active surface in anodes are accelerated by high temperature and high state-of-charge. These ageing effects lead to capacity and power losses. In contrast, low temperature and high current accelerate the deposition of metallic lithium on the anode surface [8]. Fast development cycles only

allow short testing periods, but the longevity under multiple scenarios must also be guaranteed. Therefore, ageing prediction with accelerated ageing is possible and necessary [9]. For a meaningful acceleration of the lifetime tests, the intensified ageing conditions should not trigger additional ageing mechanisms (e.g. lithium deposition) and the share of irreversible ageing and reversible capacity effects from the inhomogeneity of the lithium distribution and the anode overhang must be separated [10]. Studies combining ageing tests, cell-to-cell variation and post-mortem analyses to investigate the effects of ageing in the material over time require a high number of cells to be aged under comparable conditions and investigated in post-mortem analyses.

The fundamental patterns of how these factors influence ageing are known, and there are already investigations on this subject [4]. The challenge, however, is that each cell type is different, and thus the impact of those factors varies. This means that every new cell has to be investigated in extensive tests to be able to estimate the long-term behaviour. All these measures require very large test capacities. Furthermore, massive parallel tests are necessary to obtain results in a shorter time compared to sequential tests. Naturally, the goal is to test as efficiently and as little as possible and still achieve solid predictability. To uphold high utilisation, channels should be used in succession, testing new stress factors on a new cell when channels become available. In addition, batteries already start to age at the time of production [11]. Therefore, cells that enter a test at different times may already behave differently. Cells, therefore, need to be stored with minimal degradation at low temperatures and medium-low state-of-charge levels.

Furthermore, there are variations between individual cells of the same cell type [12]. They can be attributed to the tolerances in the production and cannot be avoided. Thus, it is not sufficient to test one cell, but all tests must be repeated with multiple cells. In a recent study by the 3rd author, it was shown that more cells should be tested to accurately capture variability than what is typically done today [13]. So far, publications of Design-of-Experiment include only either stress factors [14, 15] or cell-to-cell variation [16, 17]. And, feedback-based experiments include only a minimal design space of stress factors and extremely accelerated ageing (around 30 days of testing per cell) [18]. Therefore, the published posterior methods cannot be used on ageing tests aimed at predicting lifetime at up to 10-15 years of operation.

2. Design of a sequential analysis

All of the above discussed aspects render the testing of batteries very costly. Therefore, it is crucial to consider which stress factors of the measuring matrix and what number of cells are necessary for the intended purpose and how to adjust the design of the experiment during the test phase to incorporate knowledge gained on the fly and in a feedback loop for additional tests.

Battery degradation prediction is also limited by the amount of data available for either creating empirical models or parameterising physics based or data-based models. Furthermore, due to the vast parameter space of stress factors influencing battery degradation, tests can only provide meaningful data when those stress factors are consistently considered.

When a test is finished, and the end-of-life of the cell is reached, the testing equipment is freed for further use, and a question arises: Should you do more of the same testing or consider different stress factors, in order to get the maximum information in a given time? While conducting an ageing study, the result of the study can change with additional tests.

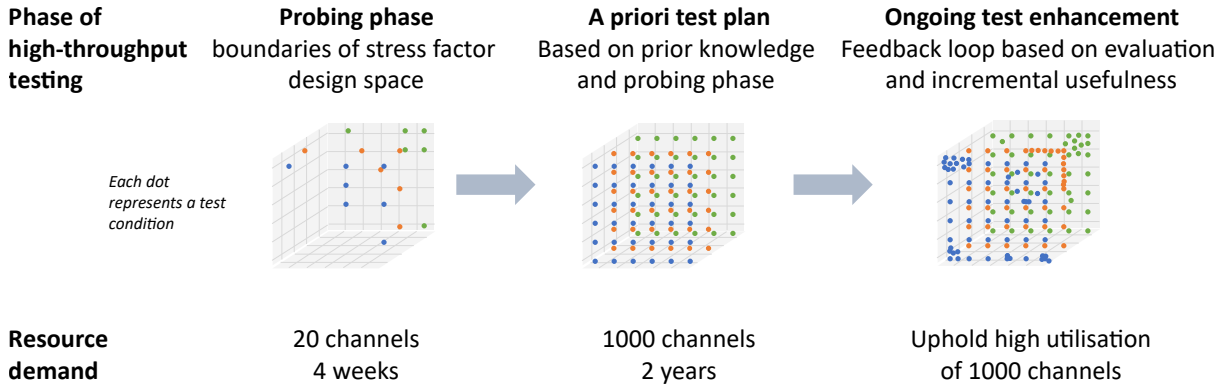

Figure 1: Sketch of high-throughput test example using 1000 channels and different phases of the design of experiment and continuous test enhancement.

Figure 1 shows the idea of the underlying testing concept as an example in this work. Each dot represents a test condition with the stress factors as the axis of the design space. First, within a probing phase, only a few cells (20 in this example) are tested to identify the boundaries of the stress factors under investigation for the given cell. The aim is to collect this data within a short time – for example 4 weeks. Then, based on this data and additional prior knowledge transfer from previous ageing tests, an a priori test plan is created and rolled out on a massive test infrastructure. In this second phase, individual stress factors are tested on 1000 channels parallel.

Finally, in the last phase, channels become available due to cells degrading faster with some stress factors, additional cells are tested to increase the data available at areas of interest with the most amount of information gained at those conditions. The second and third stages overlap and will continue for as long as the equipment is available or a sufficiently high accuracy and diversity of the measuring data is reached. This can be up to 2-3 years of testing.

Figure 2 shows a number of channels and their usage over time. Images a) and b) show example decisions made after one year based on an automatic usefulness calculation: in a) different testing is given priority, while in b) the calculation showed more tests were necessary for the same testing condition since the desired level of confidence was not yet reached.

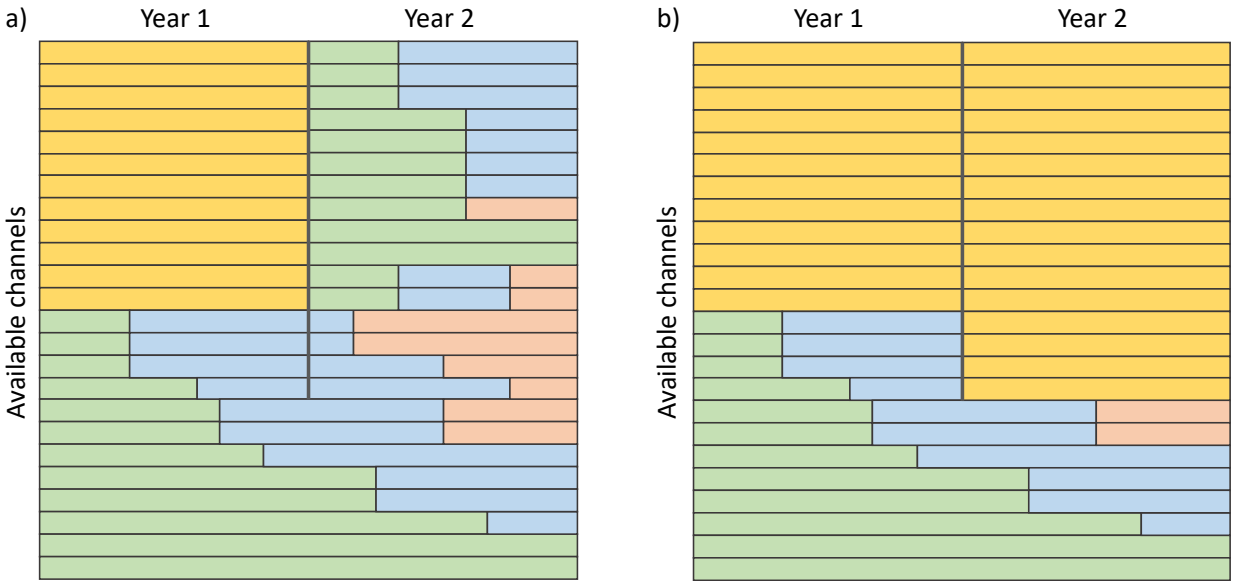

Figure 2: Example for an ageing test plan for 2 years with 24 channels available. Each row represents a channel and a) and b) describe two scenarios with a change of the test plan after one year depending of having tested "enough" to capture variability. Yellow denotes channels used to capture the cell-to-cell variability, green shows the initial run of other stress factors, and blue and red show sequential tests on the same channels. In a) after 1 year enough cells have been tested under the same conditions, so additional test conditions can be tested with the available channels. In b) the amount of data collected is not yet enough and additional tests are started with the same ageing conditions.

3. Datasets overview

In terms of data for this work, we use the same data as [13, Section 2.1] but with novel techniques. For ease of comparison we adopt their nomenclatures. The description next follows closely that in [13, Section 2.1]. For a general overview of publicly available battery data, see [19]. The datasets were chosen for study based on the necessity of testing as many cells as possible within each dataset. All datasets are open source. Each dataset features a single type of commercially available Li-ion cells, however the manufacturers, chemistries, and cell sizes vary from one dataset to the next. Although the methods outlined below can be applied to different form factors, all datasets used 18650 cylindrical cells. Some datasets had identical experimental settings, meaning that each cell was tested in the same manner, whereas others changed the stress factors somewhat beyond the expected uncontrollable experimental variability. The datasets are as follows, and notation-wise we reserve the letter N to denote the total amount of cells in a dataset.

Baumhöfer 2014	48 cells, Sanyo/Panasonic UR18650E, NMC/graphite, 1.85 Ah
Dechent-2020	22 cells, Samsung INR18650-35E, NCA/graphite, 3.5 Ah
Dechent-2017	21 cells, Samsung NR18650-15 L1, NMC/graphite, 1.5 Ah
Severson-2019	67 out of 124 cells, A123 APR18650 M1 A, LFP/graphite, 1.1 Ah
Attia-2020	45 cells, A123 APR18650 M1 A, LFP/graphite, 1.1 Ah
Attia-predicted	45 cells, Predicted data for Attia-2020 using model proposed in [20].

The capacity fade curves in Baumhöfer-2014, Severson-2019 and Attia-2020 (also Attia-predicted) exhibit the so-called *Knee* phenomena of rapid non-linear degradation [21, 22]. The Dechent-2017 and Dechent-2020 contain linear capacity fade trajectories over time. The capacity fade trajectories (y-axis) plotted against time (x-axis) can be found in Figure 3 below. For all the datasets, the capacity is normalised to the nominal capacity, and hence, expressed as a percentage – we work with state of health (SOH).

The Attia-predicted dataset is data generated considering the first 20 cycles of the Attia-2020 dataset and using the one-cycle predictor model proposed in [20].

A critical remark on cell-to-cell variability across datasets.

It should be noted that the datasets considered do not all share the same driving factor of variability. For Baumhöfer-2014 and Dechent-2020 (within each dataset), the cells were cycled

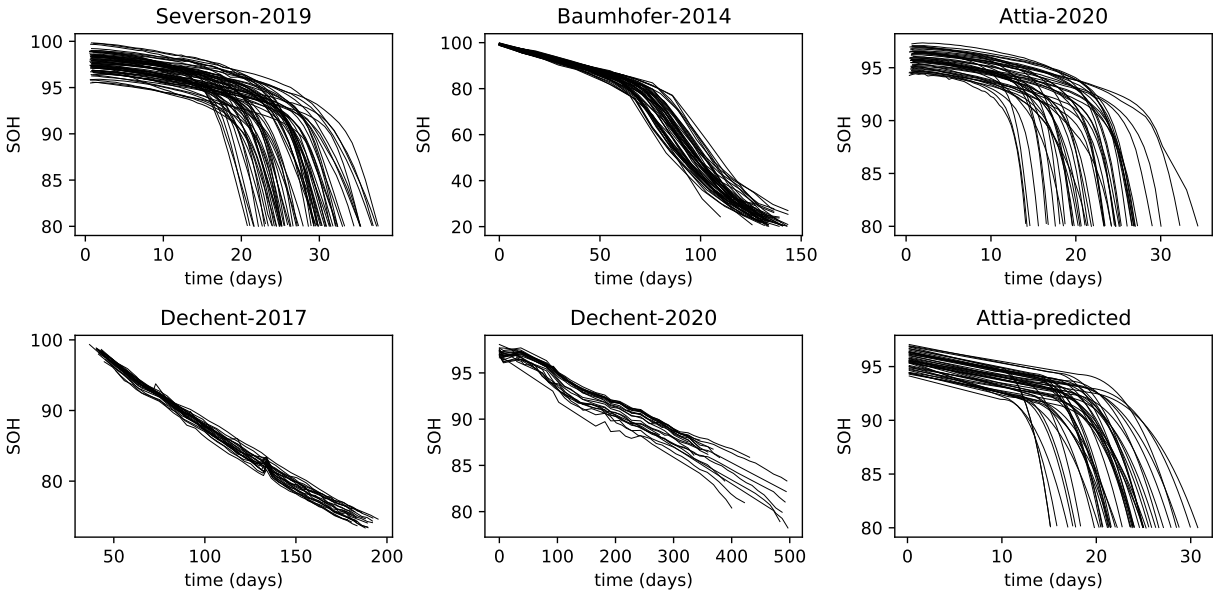

Figure 3: Capacity fade trajectories (y-axis) over time (x-axis) for the six datasets mentioned in Section 3

identically in the same environment. Therefore, the variability observed in the data is essentially the intrinsic *manufacturing variability* of the cells. In contrast, Severson-2019 and Attia-2020 consider a wide range of charge protocols, and this is an additional driving factor for cell-to-cell variability in the datasets. Thus, the variability observed in these datasets is driven by intrinsic and extrinsic factors. However, as in [13], this paper works with a restricted subset of these datasets where the variability of extrinsic factors is lower – in practise, this translated into selecting only cells with a life cycle of between 23 and 40 days, and excluding *Batch 2* of Severson's [23] original dataset. Dechent-2017 also shows extrinsic and intrinsic factors, but with small differences ($< 15\%$ difference of charge current) between the tested cells. The ability to observe solely intrinsic cell-to-cell variability is, of course, experimentally dependent.

It should be emphasised that the methodology proposed in this paper is built on a certain assumption of normality (see next section). Thus it is best suited for experiments whose main source of variability is intrinsic or where the extrinsic variability is lesser. Experiments designed with large levels of extrinsic variability (as with Batch 2 of Severson's [23] dataset compared to Batches 1 and 3) may be multi-modal in nature (e.g., 50 cells tested at $-10°C$ and 50 cells at $40°C$). In such cases, for the purpose of estimating variability, one would either require a multidimensional methodology accounting for extrinsic factors or to cluster the experiment into

distinct datasets where this in less of a factor. The methodology proposed in this manuscript follows this latter framework. At the end of the next section we discuss the current difficulty with the multidimensional methodology.

4. Methodology and estimation

The measure of cell variation for a dataset

Following Dechent et al. [13], we measure variation between cells as variation in the slopes of straight lines (1) fitted through the cell's repeated capacity measures,

$$\text{Model } \textit{Linear-2}: \qquad y(t) = \alpha + \beta t + \varepsilon, \tag{1}$$

where t is time, y is capacity, ε is a normal random variable with zero mean and finite (unknown) variance denoting the errors/residuals. The slope β and intercept α are fitted to the data (via standard least squares). Each slope β represents a cell's rate of capacity fade over successive cycles (the parameter α is discarded). This manuscript focuses on a one-parameter model for variability and it will be shown below that the number of cells necessary to capture variability suggested by this method is already high (e.g., half the total number of cells of the Dechent datasets). More complex models could be explored with a greater availability of data. In general, our methodology can be applied to other normally distributed summary statistics. For clarification, this work improves the statistical methodology deployed by [13] for this problem and does not propose a new measure of variability. This is left for future research.

For a sample of n slopes $\{\beta_i\}_{i=1}^n$ we define the *sample mean* (denoted $\bar{\beta}_n$) and the *sample standard deviation* (denoted $\hat{\sigma}_n$) as

$$\bar{\beta}_n = \frac{1}{n}\sum_{i=1}^{n}\beta_i \quad \text{and} \quad \hat{\sigma}_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\beta_i - \bar{\beta}_n\right)^2}. \tag{2}$$

The sample standard deviation of the slopes β is the measure of cell-to-cell variation chosen for this work.

We make the following working assumption.

Assumption: For each given dataset, the population of slopes β is normally distributed, i.e., slopes $\beta \sim \mathcal{N}(\mu_\beta, \sigma_\beta^2)$, where both the *population mean* μ_β and the *population standard deviation* σ_β are unknown parameters (that differ dataset to dataset).

Under this assumption, $\hat{\sigma}_n$ follows a Chi-distribution with $n-1$ degrees of freedom ([24]), namely

$$\frac{\sqrt{n-1}}{\sigma_\beta}\hat{\sigma}_n \sim \chi_{n-1} \ .$$

In order to specify a confidence level that $\hat{\sigma}_n$ is close to σ_β as a function of n, working with the χ_{n-1} distribution is inconvenient. Nonetheless, it turns out that as n increases the chi-distribution is well approximated by a normal distribution ([24], [25], [26]). We thus further assume that the distribution of $\hat{\sigma}_n$ in (2) can be approximated by a normal distribution with mean σ_β and *standard error* s_n. That is, we assume $\hat{\sigma}_n \sim \mathcal{N}(\sigma_\beta, s_n)$.

Capturing representative cell-to-cell variation for a dataset

The closeness of the estimate $\hat{\sigma}_n$ to the true value σ_β is quantified by the standard error s_n. The number of cells required to capture cell-to-cell variation is thus given by the value of n for which s_n is small enough to ensure a given precision with a given level of confidence. However, the standard error is scale dependent, so it is instead more convenient to work with the *relative standard error* (`RSE`) defined as the percentage ratio of the standard error to the standard deviation

$$\texttt{RSE} := 100\frac{s_n}{\sigma_\beta}. \tag{3}$$

For a concrete prospective: with a normal distribution, roughly 68% of samples are expected to fall within one standard deviation of the mean. Assuming that the sampling distribution of standard deviations is approximately normal, the `RSE` can thus be viewed as an upper bound on how far the sample standard deviation $\hat{\sigma}_n$ is expected to differ from the population standard deviation σ_β, with a confidence of 68% that the bound will not be exceeded.

As the `RSE` is defined in relation to s_n, it is quite easy to obtain confidence levels that our estimate $\hat{\sigma}_n$ will differ from σ_β by no more than any percentage level k%. Simply dividing k by the measured `RSE` will give the number of standard errors that a deviation of k% would correspond to. And then, the number $q := -k/\texttt{RSE}$ can be compared with the CDF of a standard normal to yield the confidence level that it will not be exceeded.

For a given sample size n the `RSE` can be obtained in two ways: theoretically and empirically. From the theoretical perspective, under the asymptotic regime of $n > 10$ ([24], [26]) the `RSE` is

given by a *deterministic expression*:

$$\texttt{RSE} = 100\frac{1}{\sqrt{2(n-1)}} \quad \Rightarrow \quad n = 1 + \frac{1}{2}\frac{1}{\texttt{RSE}^2}. \tag{4}$$

The inversion shown above gives a sample size n which (for the reasons given above) corresponds to a confidence level of approximately 68%. The reader can compare the results this equation gives with Table 1.

To measure the RSE empirically, the quantities in formula (3) must be replaced with estimates: σ_β can be approximated by taking the empirical standard deviation of the largest available sample (the whole dataset), and, s_n by using a bootstrapping procedure to construct a distribution of sample standard deviations and then taking its standard deviation. Concretely, for a given sample size n and a number of bootstrap samples b (say $b = 1000$), sample (with replacement) b sets of n slopes. Taking the standard deviation for each set of slopes produces a distribution of b standard deviations; taking the standard deviation of this distribution gives an estimate for s_n.

Making use of these results in practice

We can now describe concretely our approach to calculate the required number of cells to maintain an accurate picture of cell-to-cell variability. Two elements must be prespecified: a maximum acceptable deviation $k\%$ for the estimate $\hat{\sigma}_n$ of σ_β and the level of confidence required that this k will not be exceeded. Firstly, the linear regression (1) is fitted to the capacity data giving a list of slopes. Then, using this list of slopes, for sample sizes from $n = 2$ up to the full size N of the dataset the RSE is calculated as described above - examples of the resulting values can be seen in Figure 4. For the acceptable deviation level ($k\%$) the probability that it will not be exceeded is then calculated for each sample size - this is presented in Figure 5 for $k = 25\%$. The required sample size is then the smallest sample size providing the required confidence level. The theoretical and empirical results will not always agree and (after checking for outliers and normality as described below in relation to Table 4) we recommend selecting the larger of the two sample sizes.

In Section 5 we compare the empirical and theoretical sample sizes our methodology recommends for the datasets selected for this work. In Table 1 we present the theoretical required number of samples for a range of maximum acceptable deviations (relative to σ_β) and confidence levels that this maximum will not be exceeded.

Confidence (%) \ Maximum acceptable deviation (%)	5	10	15	20	25	30	35	40	50
50	92	24	12	7	5	4	3	3	2
60	143	37	17	10	7	5	4	4	3
68	199	51	23	14	$\boxed{9}$	7	6	5	3
75	266	68	31	18	12	9	7	6	4
80	330	84	38	22	15	11	8	7	5
85	416	105	48	27	18	13	10	8	6
90	543	137	62	35	23	17	13	10	7
95	770	194	87	50	32	23	17	14	9
99.7	1763	442	197	112	72	50	37	29	19

Table 1: Theoretical number of samples required to estimate the population standard deviation for a range of maximum acceptable deviations (relative to σ_β) and confidence levels that this maximum will not be exceeded. The 68% confidence row corresponds to the asymptotic `RSE` result of Equation (4) with the boxed $n = 9$ stands for `RSE` $= 0.25$.

For example, to obtain an estimate of standard deviation s_n that deviates from σ_β by not more than 25% at a confidence level of 68%, Table 1 indicates a sample of at least $n = 9$ cells (see also [26, p120] or [27, p103]).

Comparison to prior art

The main contribution of this work, in comparison to [13], is a statistically quantified choice of the required sample size n. While the definitions chosen for variability are of the same form as there, the methodology developed for choosing the sample size is different. The methodology[1] of [13] requires the selection of a manual threshold limit for each dataset and the exact statistical

[1] From a bird's eye perspective, both here and in [13], the starting point are models like (1) and a variation metric is build from their parameters. To work with the subsampled distributions, we use bootstrapping while [13] uses a hierarchical Bayesian approach. The final aspect, and the main difference of approach, requires a technical explanation. The method explained in the final paragraph of [13, Section 3] is to linearise the relationship between variation and sample size by taking logs – this tacitly assumes an unstated power curve relationship between variation and sample size – then identify a "stable region" of the linearised relationship by extrapolation from *manually* chosen points. It is not clear how this could be automated or which statistical interpretation it has. Finally they threshold deviations from the line to find the smallest sample size n. We compare sampling distributions using the `RSE` (3) as

meaning of this is unclear. In contrast, the parameters of maximum acceptable deviation and level of confidence used by our approach have clear statistical interpretations. The methodology here focuses solely on a standard linear regression model and models capturing non-linear degradation are not included (e.g., the *line-exponential* model highlighted in [13] or the Bacon-Watts model [22]). The three parameters of the line-exponential model in [13] are not easily interpretable and the model suffers from a lack of robustness. Additionally, the line-exponential model requires the full longitudinal data to work well (see [21, 28]) which limits its usability in online applications as is discussed below (see Section 6.2).

This project's implementation under CC BY 4.0 copyright license is publicly available on GitHub (see Additional Information section) as an open invitation for further testing and experimenting.

5. Results

For sake of exposition, this section is presented under the choice of a maximum acceptable deviation of at most 25% and a confidence level of 68%. Figure 4 shows the relative standard error of sample standard deviation as a function of the cell sample size, (starting with the smallest sample with any variation $n = 2$ until the total number N of samples available). The empirical estimates (shown as open circles) and the theoretical estimates (shown as lines) are obtained as described in Section 4.

Figure 5 shows the corresponding confidence levels for a threshold on maximum acceptable deviation specified at $k = 25\%$. These confidence levels are obtained by comparison with a normal distribution as described in Section 4. The figure shows that the theoretical estimates are generally a close fit to the empirical estimates. Comparison with empirical data, such as the Severson-2019 data shown in Figure 5, shows a good agreement in that sample size $n = 9$ is the smallest sample where the percentage `RSE` is not more than 25% with a confidence level of 68%.

Table 2 shows empirical estimates of the sample size needed to estimate standard deviation that deviates from σ_β by not more than $k = 25\%$ with a confidence of 68%. The theoretical estimate is $n = 9$. Where there are differences between the theoretical and empirical results, (for example the

a general scale-invariant measure. The user then specifies a statistically interpretable and justified threshold on the percentage of `RSE` at a confidence level and the sample size is found without any further (manual) choice.

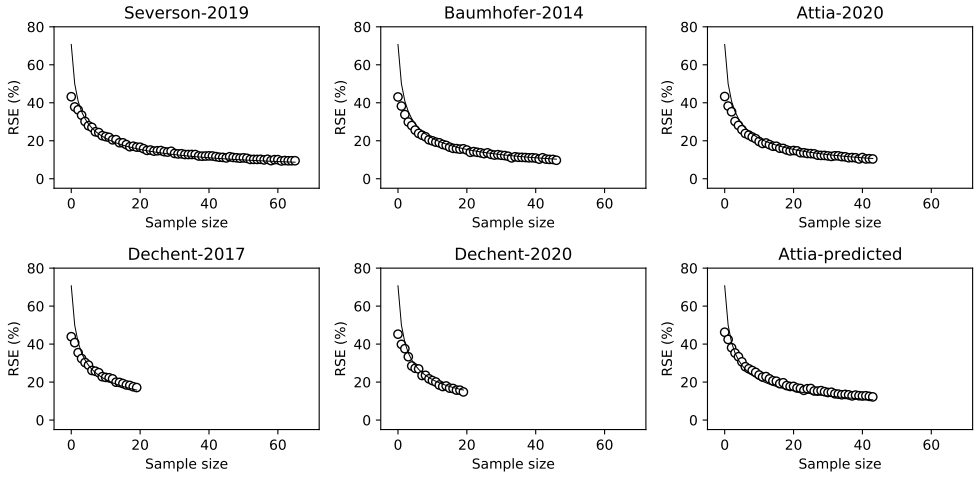

Figure 4: Relative standard error (RSE) of sample standard deviation as a function of sample size. Black continuous line given by deterministic RSE asymptotic approximation of (4).

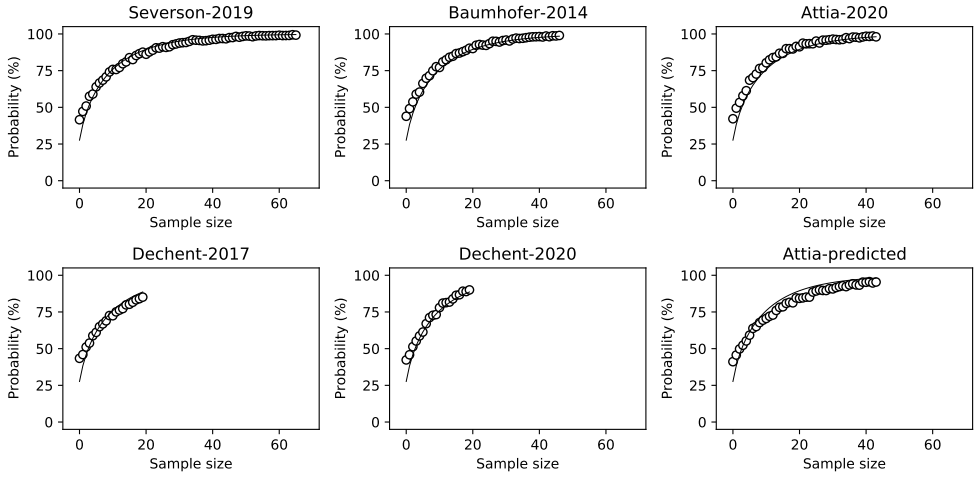

Figure 5: Probability of a random sample estimate with relative error not more than 25%.

Dechent-2017 dataset), it is most probably because the assumption of normal sampling was not met.

		Required Sample Size	
	N	Empirical	Theoretical
Baumhöfer-2014	48	8	
Severson-2019	67	9	
Attia-2020	45	9	9
Dechent-2017	21	10	
Dechent-2020	21	10	

Table 2: Sample sizes for the datasets of this study (at 25% maximum acceptable deviation and 68% confidence) per dataset, and theoretical sample size estimate (see Table 1). N is the total number of cells tested.

Figure 6 shows a Q-Q-plot graphical assessment of distribution normality of cells with Linear-2 (1) slopes in each dataset. There could be several reasons for departures from normality in the Dechent-2017 dataset. One possibility is simply that the dataset has too few cells, for example, both Dechent-2017 and Dechent-2020 have just 21 cells. As a general evaluation, for Severson-2019, Baumhöfer-2014 and Attia-2020, there is a very good agreement of the quantiles (large majority of samples) but there is evident left- and right-skew hinting at a non-symmetric distribution. Figure 3 shows that the Dechent datasets do not display capacity fade curves with knee-points.

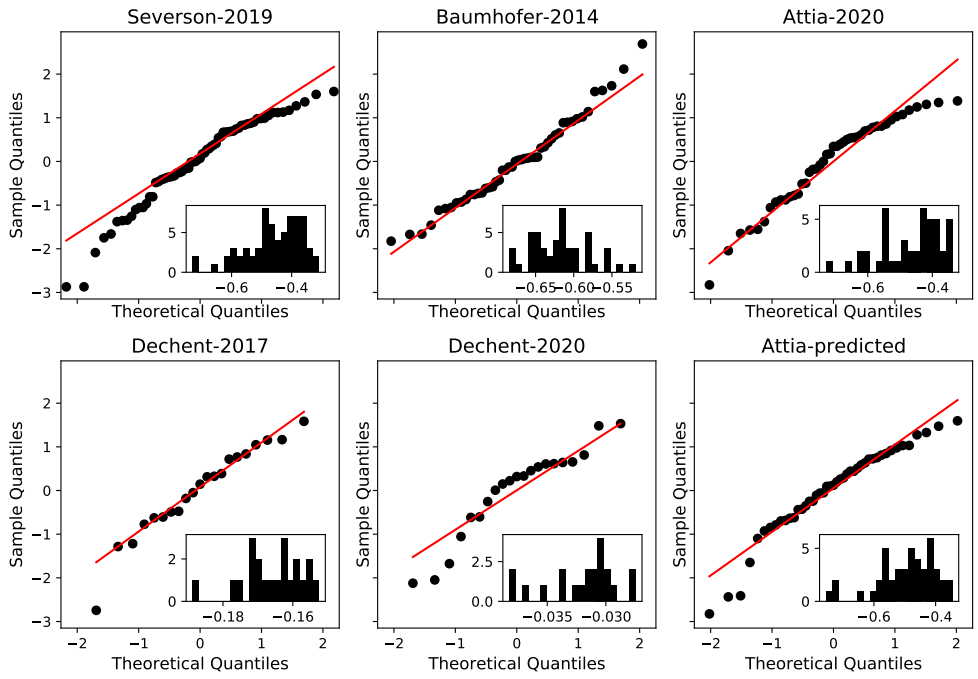

Figure 6: Q-Q plots for the standardised distribution of cell slopes β. The inset histogram plots show the true (non-standardised) distribution of slopes for each dataset.

6. Two applied examples

6.1. Using prior knowledge to inform new testing

From an historical perspective, Attia-2020 [18] appears in the literature one-year after Severson-2019 [23]. Both datasets report cycling data from similar battery cells, and we argue that knowledge gleaned from Severson-2019 could have been used to inform testing for Attia-2020. This example explores this idea.

Imagine an experiment as follows: take from the public sphere the existing Severson-2019 [23] dataset and train on it the machine learning *one-cycle predictor model* of [20] (the one-cycle model is a model designed to predict the remaining capacity degradation trajectory of a Li-ion cell from any single input cycle). Then, start the cycling experiment on the cells of Attia-2020 over a short amount of time (the first 20-cycles) and, on that information, apply the [23] trained one-cycle model to build predicted trajectories for all the cells of Attia-2020. Finally, let the rest of the Attia-2020 experiment take its course. The paths of the capacity fade curves for the three datasets can be found in Figure 3. The methodology of the previous section is then used.

To the obtained linear model slopes of the Attia-predicted dataset apply the sample size methodology developed in Section 4. Table 3 reports the estimated sample sizes for a 25% maximum acceptable deviation and a confidence level of 68% at which a theoretical sample size is computed to be $n = 9$ (see Table 1). The sample size estimate for the Attia-2020 experiment is $n = 9$ while for the Attia-predicted, computed using only very early life data, is $n = 11$.

		Required Sample Size	
	N	Empirical	Theoretical
Severson-2019	67	9	
Attia-2020	45	9	9
Attia-predicted	45	11	

Table 3: Sample size for Attia-predicted (at 25% maximum acceptable deviation and 68% confidence) per dataset, and sample size theoretical estimate (see Table 1). Information on Severson-2019 and Attia-2020 kept for comparison.

We argue that this example validates the idea of using prior information to inform the design of a future experiment. From the calculations, having used $n \approx 10$ cells in the Attia-2020 experiment instead of 45 cells would have sufficed to create a representative sample of cells to capture cell-to-cell variability for that dataset. This reduction in sample size for the testing equates to a $\approx 75\%$ reduction in experimental costs and, in view of Figure 2, would free about 35 cell-cycler channels after just 20-cycles (time-equivalent) of testing (see the concept of Figure 2).

6.2. Generating test cases in an online format to inform larger and longer experiments

In this example, we employ the estimation procedure of Section 4 under a segmentation of input longitudinally across time (recall Figure 1 and 3).

Imagine an experiment as follows: a cell cycling experiment having N cells is allocated to a cell-cycler and it is to last a T amount of time (say 10 weeks). All cycling data is collected[2]. Once the experiment runs through 20% of its allocated time (two weeks), the procedure described in Section 4 is applied to the data available and the representative sample size, say $n_{20\%}$, is determined. Once the experiment runs through 40% of its allocated time the procedure is applied again (to all data

[2]We ignore the possibility of having prior knowledge of the cells, otherwise one can easily leverage the ideas of Section 6.1 by applying, e.g., the one-cycle model at further judiciously chosen time points.

since the beginning of the experiment) and $n_{40\%}$ is estimated. This is then repeated at increments of 20% time until the end of the experiment is reached yielding the estimates $n_{60\%}$, $n_{80\%}$ and $n_{100\%}$.

The estimated samples sizes $n_{x\%}$ per dataset can be seen in Table 4 and these values need to be understood in partnership with a verification of the normality assumption underlying our methodology. This latter element is given in Figure 7 in the form of Q-Q plots at each stage of our theoretical experiment.

		Sample size for percentage of input				
Dataset	N	20%	40%	60%	80%	100%
Baumhöfer-2014	48	7	12	7	7	8
Severson-2019	67	45	17	16	6	9
Attia-2020	45	11	27	6	10	9
Dechent-2017	21	-	7	11	8	10
Dechent-2020	21	9	7	7	9	10
Severson-2019*	66	7	12	18	6	10

Table 4: Required sample size given the first $20, 40, \ldots, 100\%$ of input data (at 25% maximum acceptable deviation and 68% confidence). It should be noted that the theoretical number of required cells is 9 regardless of input size (see Table 1). Severson-2019* denotes the results after an outlier cell is removed from the Severson-2019 dataset (as justified below).

For both Dechent-datasets, which exhibit linear degradation fade curves (Figure 3), the estimated sample sizes $n_{x\%}$ are stable across the longitudinal increments in time of the curves and deviate slightly from the theoretical estimate ($n = 9$). The empty $n_{20\%}$-entry for Dechent-2017 is due to insufficient datapoints on the capacity fade curve over that time interval (see also Figure7). We thus suggest that data is recorded at a higher frequency.

For Attia-2020 and Severson-2019 (see Figure 7), there is a high variability of the data across the 20% to 60% input marks and, prominent, are the few but heavy outliers (on the left tail) that strongly influence the estimate for the sample size. Thus, the results in Table 4 at 20%-60% percentages of input are not inline with the theoretical result. It is also important to note the strong non-normal nature of the slope distributions around 60% for Severson-2019 and 40% for Attia-2020. For this range, cells are transitioning through the inflection point of their capacity fade curve, i.e., some cells have passed their knees (experiencing rapid capacity loss) and others are still

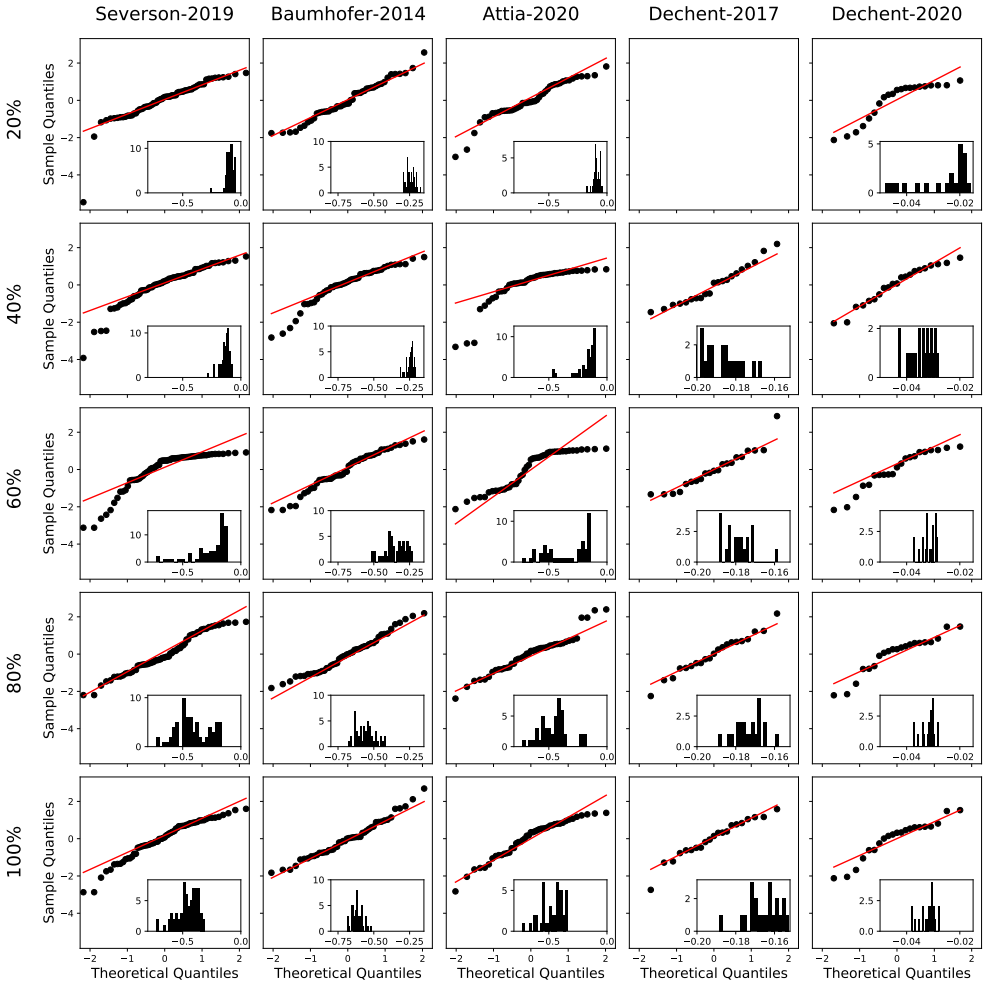

Figure 7: Q-Q plots for the standardised distribution of cell slopes β across the percent input of time data as according to Table 4. Datasets are left-to-right and percent input from top-to-bottom starting at 20% until 100%. The inset histogram plots show the true (non-standardised) distribution of slopes for each dataset and percentage of input data.

maintaining a linear decay. The data thus display a left skew at these percent levels, violating the normality assumption. At the 80% and 100% marks, the data conforms to normality and this is reflected in the estimated n in Table 4 being closer to the theoretical one.

For Severson-2019 there is a cell which decays notably faster during the early life (easily identifiable in Figure 7 at the 20% mark on the left tail). This results in a large variability in estimates of the standard deviation. For this reason, our methodology recommends keeping a large percentage of the cells at this stage. In Table 4, the row Severson-2019* displays the results of our methodol-

ogy after removing altogether the outlier cell (hence $N = 66$ instead of $N = 67$). The estimated sample sizes then conform to those observed for the other datasets (Attia-2020 in particular). This removal can be justified in practice as one cell degrading much faster than all others is likely to be faulty (accounting for significant differences in testing protocol).

For Baumhöfer-2014, the results follow the trend of the two Dechent datasets even though the capacity fade curves exhibit knees. This is explained by the less extreme (more gradual) nature of the knees displayed in the Baumhöfer-2014 dataset: there is no abrupt cliff (as in Severson-2019 and Attia-2020; Figure 3) and thus no large break from normality. We do notice some effect at the 40% level (see Figure 7), where there is a noticeable left skew in the data which accounts for the larger estimated value of n in Table 4. This effect is small in comparison to that observed for Severson-2019 and Attia-2019.

Recognisably, this example is not as conclusive as the previous one, nonetheless, it entails the critical conclusion that the underlying modelling assumption need to be verified for conclusions to be drawn. We strongly believe that this idea warrants further exploration given its potential and hope to revisit it in future research. Lastly, it is very unclear if the line-exponential model would yield better results (see discussion in Section 4) – improving upon this is left for future research.

7. Conclusions and outlook

The goal of this work was to propose a methodology to determine the smallest sample size that captures in a justified and automated way the cell-to-cell variation seen in a larger population. This manuscript improves upon the contribution of [13] by studying anew and re-thinking the underpinning statistical methodology. Under a normality assumption, that needs to be validated as part of the usage, our automatic methodology recommends a choice of $n = 9$ for a maximum acceptable deviation of 25% with a confidence level of 68%.

In future work it would be helpful to model and better explain a representative sub-population able to capture the cell-to-cell variation via the shape of the cell capacity fade trajectory. For clarity of ideas, the manuscript's focus was placed on a linear-regression method and not on models able to capture the non-linear degradation (the reason for this is argued above). As an outlook, with new larger datasets becoming available, this analysis could be performed with more complex health indicators in mind for example derived from OCV, DVA or ICA [29].

One idea for future exploration is that Internal Resistance profiles data can be included as follows: find the β^{Q} for the capacity (Q) curves according to the linear model (1); find the β^{IR} from the Internal Resistance (IR) curves; assume both sets β^{Q} and β^{IR} are normally distributed. Then *sum both*. I.e., define $\beta := \beta^{\mathrm{Q}} + \beta^{\mathrm{IR}}$; since the sum of Normal random variables is a normal random variable then the analysis carries through. This is contingent on Internal Resistance being included in the datasets which is often not the case [19].

Many laboratories have at their disposal large datasets across a rich test matrix where each entry has at most 3 battery cells [19, Section 2.7]. How to incorporate the findings of this work on such small datasets is still an open question – one possibility is to clump together entries of the test matrix to increase the number of available cells. If one has several of these datasets available (created at different timelines, institutions, testing machines), then how to combine them is also unknown. Critically, the message of [30] needs to be emphasised here: the metadata of test sets needs to be sufficiently complete (for instance, adding cell weights be useful for variation analysis). Otherwise, it will be difficult to credibly state that such datasets are sufficiently alike that they can be seen as an independent sample from the same statistical distribution.

Lastly and for perspective, the quantification of cell-to-cell variability is an open research topic and this work joins hands with [13] as educated first steps towards a general solution.

References

[1] W. L. Fredericks, S. Sripad, G. C. Bower, V. Viswanathan, Performance metrics required of next-generation batteries to electrify vertical takeoff and landing (VTOL) aircraft, ACS Energy Letters 3 (12) (2018) 2989–2994. doi:10.1021/acsenergylett.8b02195.
URL https://doi.org/10.1021/acsenergylett.8b02195

[2] S. Sripad, V. Viswanathan, Performance metrics required of next-generation batteries to make a practical electric semi truck, ACS Energy Letters 2 (7) (2017) 1669–1673. doi:10.1021/acsenergylett.7b00432.
URL https://doi.org/10.1021/acsenergylett.7b00432

[3] W. Li, N. Sengupta, P. Dechent, D. Howey, A. Annaswamy, D. U. Sauer, One-shot battery degradation trajectory prediction with deep learning, Journal of Power Sources 506 (2021) 230024. doi:10.1016/j.jpowsour.2021.230024.
URL https://doi.org/10.1016/j.jpowsour.2021.230024

[4] C. R. Birkl, M. R. Roberts, E. McTurk, P. G. Bruce, D. A. Howey, Degradation diagnostics for lithium ion cells, Journal of Power Sources 341 (2017) 373–386. doi:10.1016/j.jpowsour.2016.12.011.
URL https://doi.org/10.1016/j.jpowsour.2016.12.011

[5] J. Schmalstieg, S. Käbitz, M. Ecker, D. U. Sauer, A holistic aging model for li(NiMnCo)o2 based 18650 lithium-ion batteries, Journal of Power Sources 257 (2014) 325–334. doi:10.1016/j.jpowsour.2014.02.012.
URL https://doi.org/10.1016/j.jpowsour.2014.02.012

[6] M. Ecker, J. B. Gerschler, J. Vogel, S. Käbitz, F. Hust, P. Dechent, D. U. Sauer, Development of a lifetime prediction model for lithium-ion batteries based on extended accelerated aging test data, Journal of Power Sources 215 (2012) 248–257. doi:10.1016/j.jpowsour.2012.05.012.
URL https://doi.org/10.1016/j.jpowsour.2012.05.012

[7] S. Atalay, M. Sheikh, A. Mariani, Y. Merla, E. Bower, W. D. Widanage, Theory of battery ageing in a lithium-ion battery: Capacity fade, nonlinear ageing and lifetime prediction, Journal of Power Sources 478 (2020) 229026. doi:10.1016/j.jpowsour.2020.229026.
URL https://doi.org/10.1016/j.jpowsour.2020.229026

[8] T. Waldmann, B.-I. Hogg, M. Wohlfahrt-Mehrens, Li plating as unwanted side reaction in commercial li-ion cells – a review, Journal of Power Sources 384 (2018) 107–124. doi:10.1016/j.jpowsour.2018.02.063.
URL https://doi.org/10.1016/j.jpowsour.2018.02.063

[9] M. Dubarry, B. Y. Liaw, M.-S. Chen, S.-S. Chyan, K.-C. Han, W.-T. Sie, S.-H. Wu, Identifying battery aging mechanisms in large format li ion cells, Journal of Power Sources 196 (7) (2011) 3420–3425. doi:10.1016/j.jpowsour.2010.07.029.
URL https://doi.org/10.1016/j.jpowsour.2010.07.029

[10] M. Lewerenz, A. Marongiu, A. Warnecke, D. U. Sauer, Differential voltage analysis as a tool for analyzing inhomogeneous aging: A case study for LiFePO4|graphite cylindrical cells, Journal of Power Sources 368 (2017) 57–67. doi:10.1016/j.jpowsour.2017.09.059.
URL https://doi.org/10.1016/j.jpowsour.2017.09.059

[11] I. Bloom, B. Cole, J. Sohn, S. Jones, E. Polzin, V. Battaglia, G. Henriksen, C. Motloch, R. Richardson, T. Unkelhaeuser, D. Ingersoll, H. Case, An accelerated calendar and cycle life study of li-ion cells, Journal of Power Sources 101 (2) (2001) 238–247. doi:10.1016/s0378-7753(01)00783-2.
URL https://doi.org/10.1016/s0378-7753(01)00783-2

[12] D. Beck, P. Dechent, M. Junker, D. U. Sauer, M. Dubarry, Inhomogeneities and cell-to-cell variations in lithium-ion batteries, a review, Energies 14 (11) (2021) 3276. doi:10.3390/en14113276.
URL https://doi.org/10.3390/en14113276

[13] P. Dechent, S. Greenbank, F. Hildenbrand, S. Jbabdi, D. U. Sauer, D. A. Howey, Estimation of li-ion degradation test sample sizes required to understand cell-to-cell variability, Batteries & Supercaps 4 (12) (2021) 1821–1829. doi:10.1002/batt.202100148.
URL https://doi.org/10.1002/batt.202100148

[14] W. Prochazka, G. Pregartner, M. Cifrain, Design-of-experiment and statistical modeling of a large scale aging experiment for two popular lithium ion cell chemistries, Journal of The Electrochemical Society 160 (8) (2013) A1039–A1051. doi:10.1149/2.003308jes.
URL https://doi.org/10.1149/2.003308jes

[15] S. Schindler, M. Bauer, H. Cheetamun, M. A. Danzer, Fast charging of lithium-ion cells: Identification of aging-

minimal current profiles using a design of experiment approach and a mechanistic degradation analysis, Journal of Energy Storage 19 (2018) 364–378. doi:10.1016/j.est.2018.08.002.
URL https://doi.org/10.1016/j.est.2018.08.002

[16] S. Santhanagopalan, R. E. White, Quantifying cell-to-cell variations in lithium ion batteries, International Journal of Electrochemistry 2012 (2012) 1–10. doi:10.1155/2012/395838.
URL https://doi.org/10.1155/2012/395838

[17] A. Devie, G. Baure, M. Dubarry, Intrinsic variability in the degradation of a batch of commercial 18650 lithium-ion cells, Energies 11 (5) (2018) 1031. doi:10.3390/en11051031.
URL https://doi.org/10.3390/en11051031

[18] P. M. Attia, A. Grover, N. Jin, K. A. Severson, T. M. Markov, Y.-H. Liao, M. H. Chen, B. Cheong, N. Perkins, Z. Yang, P. K. Herring, M. Aykol, S. J. Harris, R. D. Braatz, S. Ermon, W. C. Chueh, Closed-loop optimization of fast-charging protocols for batteries with machine learning, Nature 578 (7795) (2020) 397–402. doi:10.1038/s41586-020-1994-5.
URL https://doi.org/10.1038/s41586-020-1994-5

[19] G. dos Reis, C. Strange, M. Yadav, S. Li, Lithium-ion battery data and where to find it, Energy and AI 5 (2021) 100081. doi:https://doi.org/10.1016/j.egyai.2021.100081.
URL https://www.sciencedirect.com/science/article/pii/S2666546821000355

[20] C. Strange, G. dos Reis, Prediction of future capacity and internal resistance of li-ion cells from one cycle of input data, Energy and AI 5 (2021) 100097. doi:10.1016/j.egyai.2021.100097.
URL https://doi.org/10.1016/j.egyai.2021.100097

[21] P. M. Attia, A. A. Bills, F. Brosa Planella, P. Dechent, G. dos Reis, M. Dubarry, P. Gasper, R. Gilchrist, S. Greenbank, D. Howey, O. LIU, E. Khoo, Y. Preger, A. SONI, S. Sripad, A. Stefanopoulou, V. Sulzer, Review—"knees" in lithium-ion battery aging trajectories, Journal of The Electrochemical Society (2022).
URL http://iopscience.iop.org/article/10.1149/1945-7111/ac6d13

[22] P. Fermín-Cueto, E. McTurk, M. Allerhand, E. Medina-Lopez, M. F. Anjos, J. Sylvester, G. dos Reis, Identification and machine learning prediction of knee-point and knee-onset in capacity degradation curves of lithium-ion cells, Energy and AI 1 (2020) 100006. doi:10.1016/j.egyai.2020.100006.

[23] K. A. Severson, P. M. Attia, N. Jin, N. Perkins, B. Jiang, Z. Yang, M. H. Chen, M. Aykol, P. K. Herring, D. Fraggedakis, et al., Data-driven prediction of battery cycle life before capacity degradation, Nature Energy 4 (5) (2019) 383–391.

[24] S. Ahn, J. A. Fessler, Standard errors of mean, variance, and standard deviation estimators, EECS Department, The University of Michigan (2003) 1–2https://web.eecs.umich.edu/~fessler/papers/files/tr/stderr.pdf.

[25] I. McLeod, Sampling distribution of the mean and standard deviation in various populations, http://demonstrations.wolfram.com/SamplingDistributionOfTheMeanAndStandardDeviationInVariousPo/, wolfram Demonstrations Project, published: March 7 2011, Accessed: 2010-09-30.

[26] G. E. P. Box, W. G. Hunter, J. S. Hunter, Statistics for experimenters, Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, New York-Chichester-Brisbane, 1978, an introduction to design, data analysis, and model building.

[27] G. E. P. Box, J. S. Hunter, W. G. Hunter, Statistics for experimenters, 2nd Edition, Wiley Series in Probability and Statistics, Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2005, design, innovation, and discovery.

[28] C. Strange, S. Li, R. Gilchrist, G. dos Reis, Elbows of internal resistance rise curves in li-ion cells, Energies 14 (4) (2021). doi:10.3390/en14041206.
URL https://www.mdpi.com/1996-1073/14/4/1206

[29] M. Dubarry, G. Baure, Perspective on commercial li-ion battery testing, best practices for simple and effective protocols, Electronics 9 (1) (2020) 152. doi:10.3390/electronics9010152.
URL https://doi.org/10.3390/electronics9010152

[30] L. Ward, S. Babinec, E. J. Dufek, D. A. Howey, V. Viswanathan, M. Aykol, D. A. Beck, B. Blaiszik, B.-R. Chen, G. Crabtree, et al., Principles of the battery data genome, arXiv preprint arXiv:2109.07278 (2021).

Funding

This project was funded by an industry-academia collaborative grant *EPSRC EP/R511687/1* awarded by *EPSRC & University of Edinburgh* program *Impact Acceleration Account (IAA)*.

G. dos Reis acknowledges support from the *Fundaçao para a Ciencia e a Tecnologia* (Portuguese Foundation for Science and Technology) through the project UIDB/00297/2020 and UIDP/00297/2020 (Center for Mathematics and Applications, CMA/FCT/UNL).

P. Dechent was supported by Bundesministerium für Bildung und Forschung (BMBF 03XP0302C).

Acknowledgements

All authors thank N. Augustin, V. Inacio (both at University of Edinburgh) and S. Greenbank (University of Oxford) for the helpful comments.

Author contributions

All authors provided domain expertise, edited and reviewed the manuscript. M.A.: methodology, software, visualisation; writing – technical report. C.S..: methodology, software, data curation, visualisation; writing, review and editing. P.D.: conceptualisation, visualisation; writing, review and editing. G.d.R.: conceptualisation, supervision, funding acquisition; writing, review and editing.

Competing interest declaration

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Additional information

This work has no supplementary information file.

The code used to produce this research is available under CC BY 4.0 at

https://github.com/calum-strange/sample_sizes_for_batteries .