Small Dataset Acquisition for Machine Learning Analysis of Industrial Processes with Possible Uncertainties

Xukuan Xu
Technische Hochschule Aschaffenburg
Aschaffenburg, Germany
e-mail: xukuan.xu@th-ab.de

Felix Conrad
Technische Universität Dresden
Dresden, Germany
e-mail: felix.conrad@tu-dresden.de

Andreas Gronbach
Fraunhofer-ISC
Würzburg, Germany
e-mail: andreas.gronbach@isc.fraunhofer.de

Michael Möckel
Technische Hochschule Aschaffenburg
Aschaffenburg, Germany
e-mail: michael.moeckel@th-ab.de

***Abstract*—As the algorithms mature, the bottleneck in applying Machine Learning (ML) to process analysis, monitoring and control is often caused by the availability of suitable data and the cost of data acquisition. For many ML projects, datasets have been collected independently of subsequent analysis. In industrial production, data acquisition and coverage of possible process uncertainties pose challenges to the preparation of suitable datasets. This article discusses dataset generation for ML from scratch under the constraint of limited resources with process uncertainties. A new approach towards an adapted Design Of Experiments (DOE) is proposed with the aim of sampling data more efficiently. In this way, we contribute to the challenge of preparing datasets for ML applications.**

Keywords-Small-data; Process uncertainty; Design Of Experiments(DOE); Machine learning.

I. INTRODUCTION

ML makes it possible to efficiently excavate valuable information from data with its powerful data analysis capabilities. With the prosperous advancement of algorithm research, model building is no longer a challenge limiting ML applications [1]. In fact, according to a survey from Crowdflower in 2016 [2], the efforts of data scientists are mainly (60%) consumed by data organizing and data cleaning. After this, 19% of the time is spent collecting datasets. This shows that data preparation is the bottleneck of ML applications in the current stage. However, this difficulty is often overlooked by the informatics community. In most cases, the datasets are unthinkingly pre-existing. With this standpoint, they simply optimate the algorithm at the software side for data analysis. However, the dataset's quality determines the upper limit of data analysis. Therefore, in some cases, it may be unfeasible to look at a solution only from the ML model side.

It is both a challenge and an advantage to look at data preparation from the perspective of a production engineer. Collecting a single element of the dataset requires that a product is physically produced and the relevant data is measured during the manufacturing process. In practice, an extra number of products is required to account for deficient outcomes. This limits the amount of usable data for ML analysis. The overall amount of data is often constrained by cost considerations. However, pre-existing knowledge, experience or even intuition of the process often allows an engineer to focus the data generation on particularly relevant subsets of an overly complex parameter space.

Purpose-built datasets for ML modeling may address two possible directions [3]:

I. Finding the control variables and their optimal values that give rise to an optimal response
II. Exploring the neighborhood around the optimal values to generate knowledge for monitoring, anomaly detection and control

We investigate the latter under the constraint of limited resources (e.g., time, budget) for data acquisition and fixed overall statistical process uncertainty. Based on the data obtained from the Lithium Ion Battery (LIB) production line in the KIproBatt project [4], we describe the practical difficulties in preparing datasets for industrial production in Section 2. In Section 3, existing DOE approaches are described. A set of experimental design schemes suitable for ML modeling is proposed. In addition, we propose a new Small-Data DOE (SD-DOE) suitable for ML modeling with process uncertainty.

II. DESCRIPTION OF SMALL-DATA CONTEXT

A. Small data problem

Small-batch production is often unavoidable in laboratory research, on a pilot production stage prior to upscaling, or in customer-specific (individualized) manufacturing [5]. Often, data acquisition is limited by budget or time constraints to datasets with less than one thousand elements. The particular choice of selected data points affects the outcomes of subsequent analysis. For illustration, we consider the project KIproBatt as an example of a typical small-scale data generation: a total of ca. 500 Li-ion battery cells is to be produced with a semi-automatic production line in a laboratory environment. Research questions include the imp-

 ISBN: 978-1-68558-041-4

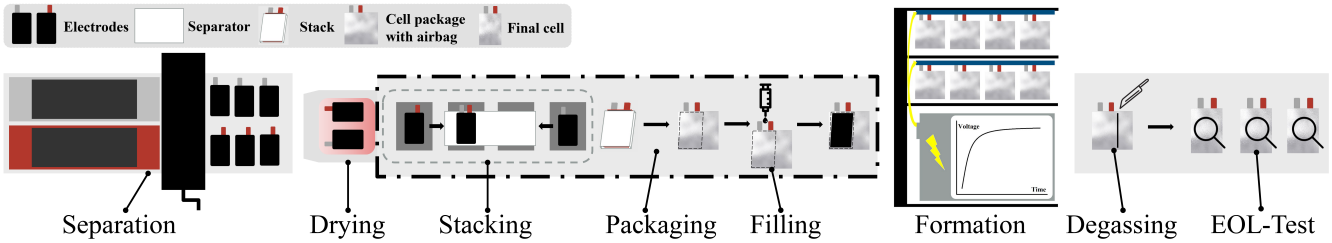

Figure 1. LIB cell assembly process from separation to EOL-test

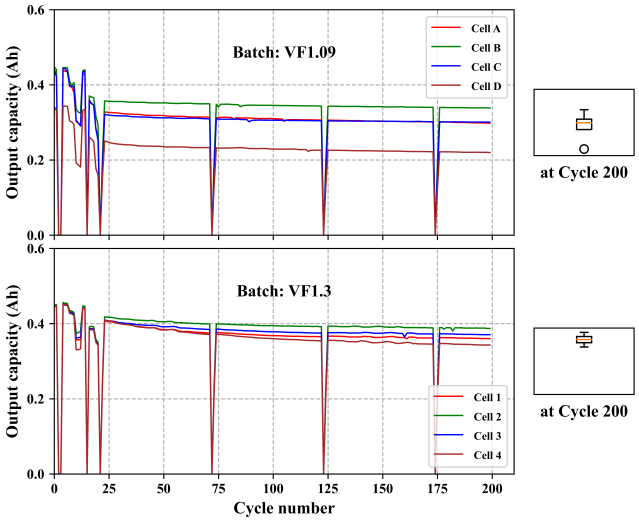

Figure 2. Cell output capacity related to cycle number in a cycling test

act of process deviations on the quality of final cells as well as the exploration of complex correlations among process parameters. Note that one cannot define the "small-data problem" by sole reference to a fixed amount of data. Instead, the characteristics and complexity both of the research objectives and the applied ML methods have to be considered.

B. *Lack of process knowledge & complexity of the production process*

The number of required data depends on the complexity of the process. A large number of features, non-linear relationships and interactions between features increase the complexity of the process and thus the number of data points required. These conditions are often found in industrial production processes [6]. The assembly process of a LIB pouch cell is an example of such a complex process and is depicted in Figure 1: cell assembly starts with electrode separation. Then, the anodes and cathodes are dried and fed into a glove box with a controlled atmosphere. Next, a stacking machine assembles the electrodes with a separator into cell stacks (Z-fold stacking). After the packaging, sealing and electrolyte filling, the cell is activated by the first charge and discharge (formation). The gas generated in this procedure is removed and the cell is finally sealed.

The complexity of this multi-step process leads to manifold variable interdependencies. Hence, an effective analysis should be based on an ML approach. However, it is challenged by limited data, which may lead to undersampling of the parameter space and a lack of convergence of the ML models. We define this as the fundamental characteristic of small-data context.

C. *Process uncertainty*

Complex processes are normally investigated for a limited set of process parameters only. While the remaining parameters are, in theory, assumed to remain constant, their unavoidable fluctuations contribute to statistical uncertainty in all measured data. Other sources for uncertainties lie, for instance, in the measurement uncertainties of the used sensors. This uncertainty is manifested in the data as identical input parameters will lead to a statistical spreading in the target responses.

In the KIproBatt project, using the injected electrolyte volume as the only tunable factor with two levels, we produced four cells at each level while ensuring that the rest of the process parameters were consistent. Each cell was then tested according to the same cycling protocol to evaluate its performance. The cycling protocol also includes non-cycling tests such as pulse, c-rate, and quick charge tests. As reflected in Figure 2, the troughs that occur every 50 cycles indicate the pulse test. The results, using Output Capacity (OC) as an indicator, are shown in Figure 2. It can be seen that the performance of the battery cells within each batch varies. As the box plot illustrates, the process uncertainty is so evident in batch VF1.09 that cell D is judged to be an outlier (box plot).

The reasons for this might be processing errors due to human operations, a lack of process understanding that leaves some potential variables uncontrolled, or measurement errors in the hardware. However, in the end, what emerges is the uncertainty of the OC.

No direct conclusion can be derived when the process uncertainty exceeds the variation imposed on control variables.

Usually, uncertainty reduction could be achieved either by optimizing hardware or by repeated measurement and averaging. However, for fixed measurement capacity, the latter implies a reduced ability for parameter space exploration. Therefore, DOE strategies can be developed further to find new compromises between resource allocation for uncertainty reduction and for parameter space sampling.

III. DOE STRATEGY

A. *Existing DOE strategies*

DOE is an established approach to systematically collect

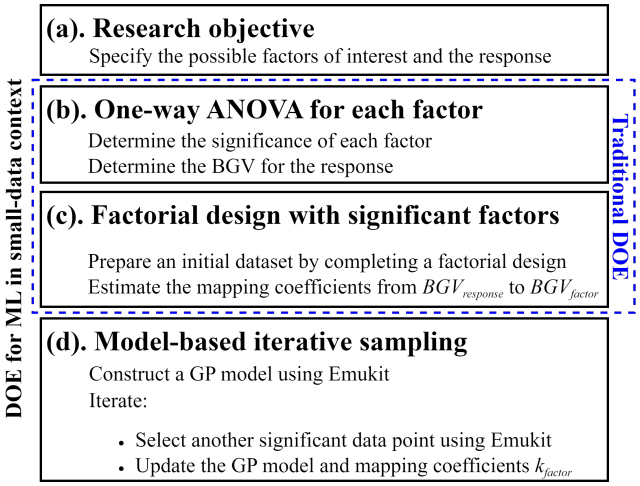

Figure 3. Proposed DOE workflow in small-data context

TABLE I. ANOVA: OC VERSUS EV

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Electrolyte	1	0.011542	0.011542	8.16	0.029
Error	6	0.008482	0.001414		
Total	7	0.020024			

information about a system or process. It aims at delivering the most relevant experimental data for addressing a given research objective. The origin of classical DOE can be traced back to the Analysis Of Variance (ANOVA) proposed by FISHER in the 1920s [7]. Traditional DOE has a set of proven paradigms: screening design for identifying relevant parameters and response surface design for detailed investigation of optimal parameter configurations. With the development of data science and easier access to data, ML tools have been successfully applied to many data analysis problems. ML has unparalleled efficiency advantages in analyzing big data (compared to the volume of data in traditional DOE) with complex interdependencies.

However, little attention has been paid to the interplay of data set generation and ML-based data analysis. A series of studies have conducted the generation of datasets for ML based on traditional DOEs in the past five years [8][9]. In addition, motivated by some ML algorithm developments, iterative data acquisition schemes have been discussed.

Emukit [9] provides such a model-based iterative DOE scheme within a Bayesian optimization framework. The Emukit DOE tool starts from a set of given initial data points and iterates the following three steps to generate sample points in a given input space:

- fit a prediction model to the existing data
- find the next point with the highest marginal predictive variance as predicted by the prediction model
- add this new data point to the existing dataset

Such iteration allows for the most efficient allocation of a limited number of data points based on certain metrics, such as marginal predictive variance of the model. This model-based scheme works well with ML data analysis since a prediction model (e.g., gaussian process model, GP model) is used to predict the target response and calculate the variance during each iteration of data acquisition.

The Emukit approach has shown excellent performance in relevant papers and simulation experiments [10][11] but would profit from further practical validation. In addition, uncertainties are not taken into account for the collected data points. Therefore, we use the Emukit method within the framework of traditional DOE and extend its iterative sampling strategy to account for data uncertainties. The resulting approach is particularly suited for the small-data context with comparatively large uncertainties.

B. Iterative DOE for small-data context

The proposed workflow in a small-data context is shown in Figure 3. We first present the first two steps (a) and (b):

First, factors of interest and their ranges are specified w.r.t. the research objective. In the second step, the range of each factor can be divided into at least two levels. Then, we perform a one-way ANOVA for each factor. ANOVA is performed on adjacent pairs with comparable variance to evaluate each pair's significance. Depending on the upper limit of the data volume, at least two replicate trials at each level are required to determine the significance of the factors (p-value). A level of significance is fixed, e.g., $p_0 = 0.05$ or $p_0 = 0.1$. If $p > p_0$, the considered factor is not significant within the interval defined by the adjacent levels.

We illustrate the subsequent DOE procedure for the example case of process analysis in battery cell production, as performed by the project KIproBatt.

A large number of factors may influence the battery cell performance. Initially, we determine the Electrolyte Volume (EV) as the only varying factor of interest and specify its range between 1.09 gram and 1.3 gram. We have produced 4 battery cells (data points) at each considered level (EV = 1.09 gram, 1.3 gram). For simplicity, we only illustrate the use of analysis of variance for this single factor. We specify the response as the lithium battery cell's OC at cycle 200. We perform ANOVA and obtain the following results (cf Table I).

For a given level of significance of $p_0 = 0.05$, we identify the electrolyte volume as a significant factor for the response ($0.029 < 0.05$) in this interval. Adjusted Mean Squares (Adj MS) are calculated by dividing the adjusted sum of squares by the Degrees Of Freedom (DF). From the Adjusted Mean Square Error (Adj MSE), we obtain the within variance as an estimate for the data uncertainties Δ_{OC}. Root mean square error allows this estimate to have the same units as the response. Thus, we have:

$$Adj\ MSE = {\Delta_{OC}}^2 = 0.001414 \quad (1)$$

With the existing within variance in this example, our requirement for significance can be relaxed until $p = p_0$. Assuming such an extreme $p = 0.05$, we can calculate the minimum required Between-group Variance (BGV) of the factor on the response such that the factor can still be determined as a significant factor. The corresponding F value can be taken from the tabulated F-distribution with group = 2, number of observations = 8 ($F\ value_{2-1,8-2,\alpha=0.05}$).

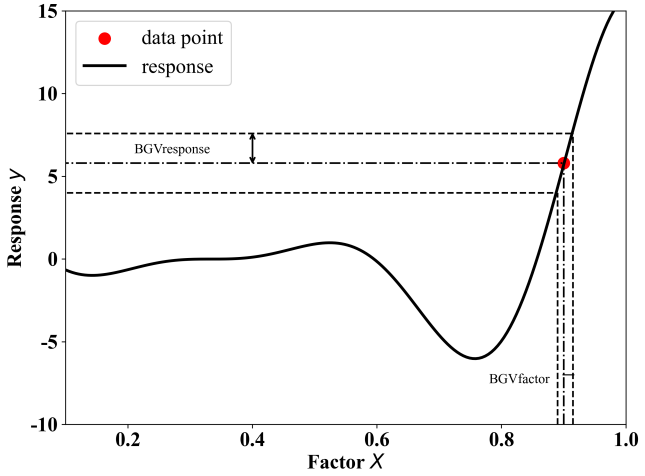

Figure 4. $\boldsymbol{BGV_{factor}}$ defines the regions unsuitable for sampling

$$BGV_{response,min} = \sqrt{F\ value_{1,6,a=0.05} * Adj\ MSE} \quad (2)$$

With the existing Δ_{OC} the critical choice of factor levels in a significance test is the limiting choice for level setting in data generation. Assuming that the levels we set for the EV are too close to each other, then the statistical spreading due to within variance may limit the distinguishability between neighboring levels. If this principle is applied to select the next data point, it can be determined whether this data point represents an additional significant level for the considered factor.

So far, we have only identified an $BGV_{response,min}$. We still need to map this $BGV_{response,min}$ to the corresponding factor $BGV_{factor,min}$, e.g., the electrolyte volume. This will be addressed in step (c) in Figure 3.

The response depends on multiple factors. For the battery cell production, next to the electrolyte volume, experts believe [6] that Drying Time (DT), Wetting Time (WT) after filling, Coating Defects (CD) on electrodes, and Stacking Accuracy (SA) also have considerable impact on the output capacity. A preliminary predictive model for the response $OC_{cycle\ 200}(X_{EV}, X_{DT}, X_{WT}, X_{CD}, X_{SA})$ can be built with the data collected in this factorial design. This model allows calculating the derivative of the response w.r.t each factor

$$k_{factor} = \frac{\partial OC_{cycle\ 200}}{\partial X_{factor}} \quad (3)$$

for local inversion. By using a simple multilinear regression model, e.g.:

$$OC_{cycle\ 200} = \sum_{i=1}^{5} k_i X_i + b \quad (4)$$

the coefficients k_i for each factor X_i in (4) are the mapping coefficients to linear order. Thereby, we can map the $BGV_{response,min}$ (on the responses) to the $BGV_{factor,min}$ (on the factors) and thus determine the minimum required between-group variance $BGV_{factor,min}$ for each factor.

$$BGV_{factor,min} = BGV_{response,min}/k_{factor} \quad (5)$$

As reflected in Figure 4, the $BGV_{factor,min}$ defines an environment around each factor value where no significant data points can be chosen. Each new data point will be used to update the model and the mapping coefficients to determine a more accurate estimate for $BGV_{factor,min}$. Under this framework, we can proceed the iterative sampling in step (d) until all data points for a machine learning dataset have been collected.

IV. CONCLUSION

This article discussed the characteristics of small data problems with process uncertainties. A new approach towards an adapted DOE is proposed with the aim of sampling data more efficiently under such circumstances. This DOE approach is applied to the battery cell production for the project KIprobatt and we are looking forward to presenting our following results.

REFERENCES

[1] F. Conrad, M. Mälzer, M. Schwarzenberger, H. Wiemer, and S. Ihlenfeldt, "Benchmarking AutoML for regression tasks on small tabular data in materials design", Sci Rep, vol. 12, no. 1, Art. no. 1, pp. 19350, Nov. 2022, doi: 10.1038/s41598-022-23327-1.

[2] Figure Eight. *CrowdFlower: Data science report*. [Online]. Available from: https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf [retrieved: 02, 2023].

[3] A. Dean, D. Voss and D. Draguljić, Design and Analysis of Expermients, 2nd Edition. New York, NY: Springer, 2017.

[4] X. Xu et al., *KIproBatt: Exploring smart battery cell production based on a generic system architecture and an AI-enhanced process monitoring*. [Online]. Available from: https://doi.org/10.13140/RG.2.2.11573.76006 2021.11.07

[5] J. Fleischer, G. Lanza and K. Peter, "Quantified Interdependencies between Lean Methods and Production Figures in the Small Series Production," Manufacturing Systems and Technologies for the New Frontier, pp. 89–92, 2008, doi: 10.1007/978-1-84800-267-8_17.

[6] M. Westermeier, *Qualitätsorientierte Analyse komplexer Prozessketten am Beispiel der Herstellung von Batteriezellen.* [online]. Available from: https://www.mec.ed.tum.de/fileadmin/w00cbp/iwb/Institut/Dissertationen/322_Westermeier_Markus.pdf [retrieved: 02, 2023].

[7] R.A. Fisher, The Arrangement of Field Experiments in Breakthroughs in Statistics. New York, NY: Springer, 1992.

[8] L. Salmaso et al., "Design of experiments and machine learning to improve robustness of predictive maintenance with application to a real case study", Communications in Statistics - Simulation and Computation, vol. 51, no. 2, pp. 570–582, Feb. 2022, doi: 10.1080/03610918.2019.1656740.

[9] A. Paleyes et al., "Emulation of physical processes with Emukit". arXiv, Oct. 25, 2021. doi: 10.48550/arXiv.2110.13293.

[10] M. Zhang, A. Parnell, D. Brabazon, and A. Benavoli, "Bayesian Optimisation for Sequential Experimental Design with Applications in Additive Manufacturing". arXiv, Nov. 23, 2021. doi: 10.48550/arXiv.2107.12809.

[11] Z. Liu et al., "Machine learning with knowledge constraints for process optimization of open-air perovskite solar cell manufacturing", Joule, vol. 6, no. 4, pp. 834–849, Apr. 2022, doi: 10.1016/j.joule.2022.03.003.

 ISBN: 978-1-68558-041-4