Prescriptive and descriptive quality metrics for the quality assessment of operational data

Quality assessment for data-driven and hybrid models in the process industry

Isabell Viedt[1], Jonathan Mädler[1], Valentin Khaydarov[2] and Leon Urbas[1]

Abstract: In the process industry data-driven and hybrid modeling approaches are increasingly popular in regards to process monitoring, optimization and control. The major problem with process data is that the data collected in process plants during operation, even though available in vast amounts, might generally be low in information content. The collected data usually represents certain operating points while anomalies, ramp-up and shut-down are rare occurrences and therefore only seldom covered. Due to its possibly low quality, the use of such data might lead to an inadequate model coverage and overall low model performance. Data quality assessment prior to modeling is crucial to allow an estimation of model quality prior to the model development. Therefore, the following paper discusses prescriptive and descriptive assessment metrics for the quality assessment of process data and their potential application in the quality assurance of data-driven and hybrid models. This approach will in later application support the user in their choice of modeling approach.

Keywords: Data quality assessment, prescriptive data quality metrics, hybrid modeling.

1 Introduction

In the process industry most process cannot sufficiently be described via a first principles approach because most times there is a lack of knowledge about the underlying biochemical, chemical and biological process [CR19]. Therefore, data-driven and hybrid modeling approaches are increasingly popular in the process industry in regards to process monitoring, optimization and control [St14, Gär21]. Data collected during plant operation usually only covers certain operation point and can therefore be low in information content. Special events as anomalies, ramp-up and shut down are only seldom covered [Me19]. The use of low quality process data might lead to an inadequate model coverage and overall low model performance. This is especially important the more data is eventually used for the implementation of these simulation models. Hybrid models will generally require less over all data and lower quality data for the same performance as solely data-driven models [St14]. Even though the required data quality varies for the individual modeling application, for both hybrid and data-driven models the resulting

[1] Technische Universität Dresden, Professur für Prozessleittechnik, Helmholtzstr. 10, 01069 Dresden, isabell.viedt@tu-dresden.de, jonathan.maedler@tu-dresden.de, leon.urbas@tu-dresden.de
[2] Technische Universität Dresden, Process-to-Order-Lab (P2O-Lab), Helmholtzstraße 16, 01069 Dresden, valentin.khaydarov@tu-dresden.de

doi:10.18420/inf2022_88

model performance correlates with the utilized data for modeling. As data quality of process data is generally lower, it is important to consider data quality assessment prior to model development. With this, unsuitable or insufficient data can either be flagged for pre-processing to improve the overall quality or discarded all together because the data utilization would lead to insufficient model performance [St14]. To address this challenge, the following paper discusses how prescriptive assessment metrics can be utilized in the assessment of process data to support the user in decision making and their potential application in the quality assurance of data-driven and especially hybrid models [VMU22].

2 Data quality assessment

In order to enable informed decision making about the quality of data, it is first necessary to define what good data quality dimensions are [MVU21]. Therefore, the authors utilize already existing data quality frameworks that define attributes that can be allocated to data quality. One example for a data quality model to describe data quality dimensions is the quality model of ISO/IEC 25012 [ISO08]. This quality model is commonly used for the data quality assessment in software. The standard differentiates between two types of data quality: inherent data quality and system-dependent data quality.

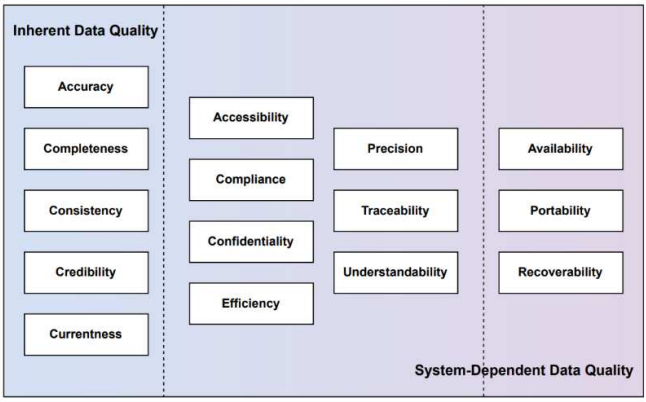

Fig. 1: Data quality model according to ISO/IEC 25012

Inherent data quality refers to the degree to which quality characteristics of data have the intrinsic potential to satisfy stated and implied needs. Therefore, data quality refers to data itself in this case. Whereas system-dependent data quality refers to the degree to which data quality is reached and preserved within a computer system when data is used under specified conditions and therefore quality depends on the used technological domain.

Besides dimensions for data quality, in literature already exist numerous frameworks for the assessment of data quality [CR19, Ja20]. Most only consider the data quality

dimensions or characteristics and disregard actually assessment measures and metrics in favor of being generic and widely applicable [CR19, He18]. To allow qualitative and quantitative assessment of the data quality dimensions, specific measures or metrics must be allocated to them in form of statistical measures and assessment specifications [He18, CR19, Ja20]. The SQuaRE standard already provides an initial assessment approach with the accompanying standard ISO/IEC 25024 [ISO15] which defines data quality measures (DQM) for quantitatively measuring the data quality in terms of characteristics defined in ISO/IEC 25012. In the following, the term data quality metrics is used to define the DQM.

3 Descriptive and prescriptive data quality metrics

With the quality process data being poor in most cases, strategies for pre-assessing the quality of the data before model implementation are necessary [MVU21]. For this purpose, the authors propose the structuring of the data quality assessment measures or metrics into the categories prescriptive data quality metrics and descriptive data quality metrics similar to prescriptive and descriptive data analytics [Me19]. Prescriptive data quality metrics allow the assessment of the data quality before implementation into a model or algorithm, descriptive metrics will assess the quality of the implemented software or modeling product. Descriptive data quality assessment (cf. descriptive analytics [Me19]) will give insights into the overall data quality. With this definition most quality metrics can at some point be defined as descriptive metrics. When analyzing ISO/IEC 25024 in regards to the defined data quality measures, it can be concluded that most system-dependent and inherent quality metrics can be classified as descriptive metrics. The problem with a descriptive approach is that there is no recommended action and the fault detection only occurs after model implementation. Therefore, model improvement through an additional iteration loop is required, which utilizes more resources, time, and money.

The concept of prescriptive data quality assessment has the purpose of assessing data quality prior to model implementation and also includes recommendations of action in regard to type of suitable pre-processing steps and modeling approach. With this, the modeling options are narrowed down to the most suitable methods. To achieve prescriptive data quality assessment, specific data quality metrics and measures must be allocated. One example of a prescriptive data quality measure is the measure ‘data value completeness’ allocated to the inherent quality dimension completeness. The purpose of the measure and metric is to verify if the data set for example represents all necessary features or operating points. Further ‘Data model accuracy’ for the quality dimension accuracy coupled with a statistical approach can be utilized to predict model quality via data quality [He18]. Prescriptive data quality assessment is an interesting addition to the quality- and test-driven modeling approach presented in Mädler et al. (2021) [MVU21]. This approach also requires the quality and scope of the necessary context data to assess the inherent data quality prescriptively [VMU22]. Therefore, qualitative measures like ‘Metadata completeness’ for the quality dimension completeness [ISO15] can also be utilized in prescriptive data quality assessment.

4 Conclusion and outlook

This paper presented the concept of structuring data quality metrics into a prescriptive and descriptive category. With this, the first step into integrating this data quality assessment approach into a general model quality assessment framework (cf. [VMU22, MVU21]) has been taken. Since data quality is important for the quality assessment of hybrid models, future assessment framework can be improved by conducting prescriptive data quality assessment before the start of model building to exclude unfit data or to select the most suitable data set. This will also allow a supported decision between different data-driven and hybrid modeling approaches based on the overall data quality. The next step is the validation via case studies. For this quality models for the specific modeling goal are derived and the data is then tested against those quality models, metrics and reference values to verify if the presented approach supports the user with model selection.

Literaturverzeichnis

[CR19] Cichy, C.; Rass, S.: An overview of data quality frameworks. IEEE Access 7, 24634-24648, 2019.

[Gär21] Gärtler, M. et al.: The Machine Learning Life Cycle in Chemical Operations–Status and Open Challenges. Chemie Ingenieur Technik, 93(12), 2063-2080, 2021.

[He18] Heinrich, B. et.al.: Requirements for data quality metrics. Journal of Data and Information Quality (JDIQ) 9(2), 1-32, 2018.

[ISO08] ISO/IEC 25012:2008, Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Data quality model, 2008.

[ISO15] ISO/IEC 25024:2015, Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Measurement of Data Quality, 2015.

[Ja20] Jain, A. et al.: Overview and importance of data quality for machine learning tasks. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 3561-3562, 2020, August.

[Me19] Menezes, B. C. et al.: Predictive, prescriptive and detective analytics for smart manufacturing in the information age. IFAC-PapersOnLine, 52(1), 568-573, 2019.

[MVU21] Mädler, J., Viedt, I., Urbas, L.: Applying quality assurance concepts from software development to simulation model assessment in smart equipment. In Computer Aided Chemical Engineering. Elsevier. 50, 813-818, 2021.

[St14] Von Stosch, M., Oliveira, R., Peres, J., de Azevedo, S. F.: Hybrid semi-parametric modeling in process systems engineering: Past, present and future. Computers & Chemical Engineering, 60, 86-101, 2014.

[VMU22] Viedt, I., Mädler, J., Urbas, L.: Quality assessment for dynamic, hybrid semi-parametric state observers. In Computer Aided Chemical Engineering. Elsevier. 51, 813-818, 2022.