Challenges for the DOE methodology related to the introduction of Industry 4.0

Jacek Pietraszek[1]* , Norbert Radek[2] , Andrii V. Goroshko[3]

[1] Cracow University of Technology, 31-864 Kraków, Al. Jana Pawła II 37, Poland,
[2] Kielce University of Technology, 25-314 Kielce, Al. Tysiąclecia Państwa Polskiego 7, Poland,
[3] Khmelnytsky National University, 29000 Khmelnytskyi, Instytutska Str. 11, Ukraine,
 *Corresponding author e-mail: jacek.pietraszek@mech.pk.edu.pl

Abstract
The introduction of solutions conventionally called Industry 4.0 to the industry resulted in the need to make many changes in the traditional procedures of industrial data analysis based on the DOE (Design of Experiments) methodology. The increase in the number of controlled and observed factors considered, the intensity of the data stream and the size of the analyzed datasets revealed the shortcomings of the existing procedures. Modifying procedures by adapting Big Data solutions and data-driven methods is becoming an increasingly pressing need. The article presents the current methods of DOE, considers the existing problems caused by the introduction of mass automation and data integration under Industry 4.0, and indicates the most promising areas in which to look for possible problem solutions.

1. Introduction

Historians distinguish three industrial revolutions related to radical changes in production techniques and their effects: a geographic and social change in the distribution of wealth, a change in social relations, and even a change, considered typical in a given era, in family structure. These are the era of steam that began at the end of the 18th century, the era of electricity dating back to the turn of the 19th and 20th centuries, and the era of computers, the conventional beginning of which was established in the 1970s.

Industry 4.0 is a name referring to the promoted "fourth industrial revolution", hence the number, consisting in the integration of manufacturing and automation techniques, data processing and exchange. A characteristic feature of these successive changes is the delay, from several to several dozen years, between an invention characteristic of a given revolution and a massive change in the organizational and production paradigm. In the case of the postulated fourth industrial revolution, the Internet is such a characteristic invention, which created a universal and uniform platform for connecting various devices at any distance.

Each of the parts included in the information-integrated production team has been designed and is associated with an appropriate model of functioning. The complexity of production processes, and in particular the disruptive effects of uncontrolled environmental and raw material factors, make it impossible to stabilize and optimize the process in its entirety. It is necessary to create empirical models of functioning focused on the local aspect of optimization and stabilization. The obtained predictive models are then the basis for making decisions about the necessary corrective actions.

The appropriate tool to carry out these activities is the DOE (Design of Experiment) methodology, known since the 1930s. Its origins are marked by the publications of Fisher (Fisher, 1921; Fisher, 1925), referring to the ANOVA method and Latin squares derived from it, and Yates (Yates, 1935), who developed factorial experimental designs and methods for their analysis. The DOE methodology allowed the Allies to achieve significant successes in industrial production and logistics during World War II. In the next decade, the methodology was adapted to the needs of the chemical industry (Box and Wilson, 1951; Scheffe, 1958) by introducing the response surface methodology (RSM) and mixture plans. The systematization of related formalism is the achievement of two pairs

of mathematicians: Robbins and Monroe (Robbins and Monroe, 1951) and Kiefer and Wolfowitz (Kiefer and Wolfowitz, 1952; Kiefer and Wolfowitz, 1959). The most spectacular successes were achieved during the time-strenuous Apollo program, when new materials and new devices had to be developed in less than a decade and ensured that they achieved both high and stable operating parameters. In turn, the construction of the lunar lander was qualitatively supervised at Grumman by Dorian Shainin (Bhote and Bhote, 2000), who became the author of the well-known Red-X™ methodology for effective process stabilization. The effectiveness of his methods is proved by the fact that during any flight to the moon the lander's devices did not fail.

The traditional DOE methodology was and still is used with the assumption that the data set necessary to build a predictive model and arrive at a design or technological decision should be as small as possible, since each set of measurements is associated with an experimental test that is costly and/or time consuming. Hence, the method of constructing the experiment was defined as designed to achieve the assumed informational and statistical features with a minimized number of tests.

The Industry 4.0 production environment is of a different nature, as the mass sensing of automated production lines has resulted in the current data stream being very high, and the created archival datasets are huge (Karpisz and Kiełbus, 2018). Automated production lines and the processes carried out on them react badly to forcing parameter settings significantly deviating from stable process settings, and this is an element of the tests carried out in the DOE methodology, where the response of the process to such a "jerk" with parameters is observed. Among the methods of analyzing the obtained big data set, the dominant methods are either correlational or based on machine learning. With all their advantages, they have one major disadvantage: the lack of showing and justifying possible cause-effect relationships. An additional drawback is the requirement to use, in most cases, very high computing power obtained only in cloud infrastructure.

In this situation, it is desirable to modify and extend the existing DOE methodology in such a way that, while maintaining its current advantages, it can be used for data analysis and creating predictive models in the Industry 4.0 environment. This need is rationale why this article deals with the challenges facing the DOE methodology in relation to the expansion of the areas covered by Industry 4.0.

The starting point for this consideration is the compilation, characterization and comparison of four basic DOE methodologies, which are widely used and which have become the analytical basis for more complex approaches, e.g. data-driven DMAIC (Define, Measure, Analyze, Improve and Control) used as a core part of SixSigma (Montgomery, 2020). The next necessary step is to consider what major change, from the data analysis point of view, is introduced by the Industry 4.0 context. Finally, the suggested changes that should be made to maintain the usefulness of the DOE methodology should be identified. The following considerations are currently only a starting point to identify those areas of applied mathematics, the adaptation of which to the needs of DOE seems to be the

most promising and effective. Only narrow-scope studies focused on specific fragments of the methodology will provide the necessary solutions, but without a general overview of the situation, it would not be known which of these areas to explore.

2. Four traditional DOE approaches

In the area of DOE methodology, four separate approaches to the problem of designing an experiment and building predictive models can be distinguished (Figure 1).
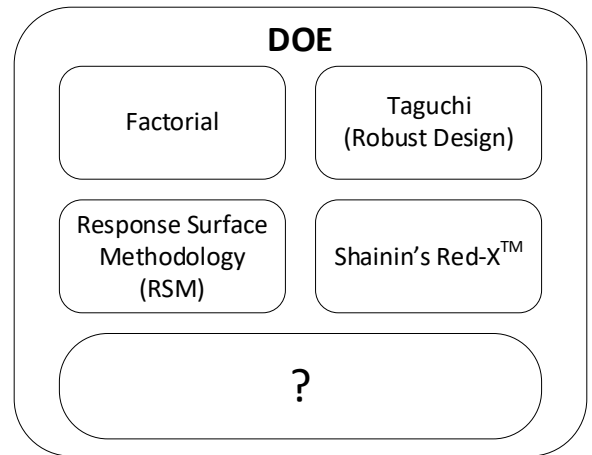

Fig. 1. Four components of the current DOE methodology

The most common, especially in the engineering industry, is the factor`ial approach developed by Yates (Yates, 1935). It is characterized, especially in the case of two-level designs, by a very simple development, application and analysis. The factorial approach enables both an in-depth analysis of the influence of controlled factors and their possible interactions of any order (full factorial variant), and the limitation of the number of experimental tests without determining interactions of higher orders (fractional factorial variant). In the first case, a rich set of information is obtained, but at high economic costs, in the second case, these costs are limited, but so is the set of information. The Yates's factorial approach is also the starting point for the Evolutionary Operation (EVOP) method proposed by Box (Box, 1957), which is a solution for continuous manufacturing processes.

The direct competitors of the factorial approach are the Fisher's Latin squares and the Taguchi method (Robust Design). The first case, historically the oldest (Fisher, 1925), is directed specifically at the analysis of exactly three controlled factors, of which usually one is of interest to the experimenter, and the other two are the dominant environmental factors masking the influence of the former. The method allows the application of multiple levels of factor control, but also imposes limitations as it does not allow for the study of interactions and all factors must have the same number of levels. The Taguchi method (Phadke, 1989), developed more than thirty years later in the 1950s and 1960s, removed some of the limitations of Latin squares while introducing the whole concept of robust design

in which process stability takes precedence over local optimization of the process response. Taguchi's approach assumed the decomposition of the experimental design into two distinct ones: the internal array, controlling highly controlled factors, and the external array, determining factors poorly controlled or simulated environmental disturbing factors.

At the same time, mainly in the chemical industry, but also in metal industry (Lipiński, 2015; Lipiński and Wach, 2015), in anti-corrosion protection (Włodarczyk et al., 2011; Wrońska and Dudek, 2014; Lipiński, 2017), biotechnology (Skrzypczak-Pietraszek et al., 1993) and medical research (Wojnar et al., 2019), the response surface methodology (RSM) initiated by Box and Wilson (Box and Wilson, 1951), and later modified by Scheffe (Scheffe, 1958) for testing mixtures was developed. The main difference from the previously mentioned methods was the introduction of controlled factors with continuous settings and the imposition of the structure of the predictive model in the form of a predetermined function with unknown parameters. Additionally, in the case of mixtures, there was the so-called the condition of summability, which meant that the settings of the controlled factors (shares of the mixture components) could not be selected freely in the experiment, but had to add up to a constant value, usually up to 100%.

A characteristic feature of the methods mentioned so far was the so-called static approach, meaning that all previously planned measurements had to be performed before starting the analysis. A different solution was proposed by Dorian Shainin in his methodology Red-X™ (Bhote and Bhote, 2000; Pacana et al., 2014; Pacana et al., 2018), which he perfected for several decades while working at Grumman Aircraft and General Motors, and after retiring at the Shainin Institute he founded in 1990s. The purpose of the Shainin method is solely to stabilize the process by identifying those controlled factors that are the main causes of process instability. Shainin's method is oriented towards the intensive use of local engineering knowledge of the studied process and the sequential interleaving of tests and subsequent analyzes. This approach allows you to discontinue testing after the major confounders have been identified, thus saving significant resources over typical static experimental testing schemes. Shainin's approach is not a single analytical or computational method, but a complete scheme of conduct, also organizational, and in this respect it is very similar to Six-Sigma, the structure of which also includes building a predictive model based on DOE.

Summarizing, it can be concluded that all the above methods of proceeding assume relatively small data sets and a low data stream intensity. This feature is beneficial in traditional industry, but becomes a strong limitation in the context of Industry 4.0.

3. Data feed changes in Industry 4.0

Industry 4.0 is primarily the automation of production lines and the integration of data transmission within them. This means that the source of data are numerous sensors that are supposed to provide information to the controllers of the production line. Thus, the intensity of the data stream is large and the collected data sets are huge. Consequently, the number of controlled factors is large or very large. Production databases are fed with constant streams of information that can be either processed on-line or processed off-line. Process runs, nominally running according to the settings of controlled factors, are constantly subjected to disturbances with variable characteristics, which is immediately reflected in the transmitted measurement data.

Due to the fact that they are automated lines with reduced staff, they achieve the highest economic efficiency in a continuous operation pattern. Therefore, the traditional DOE approach to building predictive models, assuming the performance of experimental tests in combination with a large deviation from typical production settings, is very reluctant to be accepted by process engineers because it causes instability of the process. The management board is also against this, as the production obtained during such experiments does not meet the requirements of formal quality assurance systems and is therefore only suitable for scrapping and thus is a large additional cost.

Hence, more and more requests are addressed that the analytical methods should reduce the scope of the designed experiments, and start to rely more on passive observations that can be obtained from production databases or, at most, to use design experiments with small ranges of deviations from the nominal process settings, so as not to violate the formal requirements of the system quality assurance.

4. Suggested changes

Taking into account the above considerations, specific comments can be formulated. Certainly, the methods must be less invasive and more observational. This means a departure from the paradigm of a designed experiment towards passive methods that have so far been more used in biology or medicine. In analytical methods, it will be necessary to take into account the unfavorable correlations, which so far have been reduced or even zeroed through the appropriate selection of orthogonal designs.

In continuous processes, it will be appropriate to implement the EVOP method or its appropriate modifications, as its scheme of operation allows to maintain the formal requirements of the quality assurance system despite the experiment being conducted at the moment. Thus, losses related to the scrapping of the current production are avoided. It is purposeful to develop the existing and implement new non-destructive testing methods (Patek et al., 2014; Trzewiczek et al., 2014), as this minimizes economic losses. Image analysis methods (Szczotok and Roskosz, 2005; Szczotok and Sozańska, 2009) seem to be particularly promising here, as the theoretical and analytical background is very developed, and the current development of both optical devices and processing power allows for using these methods online.

As far as analytical methods are concerned, it is necessary to implement analyzes within the DOE that are appropriate for the time series (Pedrycz and Chen, 2013; Shumway and Stoffer, 2017), as currently production environments are charac-

terized by high time variability. This approach, partially already implemented in Shainin's method through Multi-Vari ™ plots (Bhote and Bhote, 2000), allows the identification of unknown confounders by identifying temporal instability patterns.

In research, design and diagnostics, unlike in a typical manufacturing industry, multivariate statistics and stochastic fields are increasingly used, in particular directional statistics (Mardia and Jupp, 2000), multivariate and non-linear approaches (Izenman, 2008), and spatio-temporal statistics (Sherman, 2011). These highly advanced analytical methods, which differ greatly in mathematical apparatus from relatively simple industrial methods, are particularly useful for analyzing the properties and behavior of materials with complex microstructures (Capriz et al., 2002), non-isotropic materials, and materials whose properties change rapidly over time, e.g. metal foams (Rajak and Gupta, 2020), composites (Panasenko, 2005) and nanomaterials (Korzekwa et al., 2018; Korzekwa et al., 2020).

It is desirable to replace the existing parametric predictive models with a non-parametric, data driven model (Heinz, 2011). Such an approach, however, means the necessity to change the methods of estimating uncertainty, as the existing ones were based on numerous assumptions, e.g. model parametricity, linearity of parameters, use of the least squares method, normality of the distribution of disturbing factors, etc.

A convenient method of estimating uncertainty in such cases is the bootstrap method, based on resampling the data set (Pietraszek et al., 2016; Pietraszek et al., 2017). Obviously, the different methods of data analysis lead to differences in the selection of appropriate experimental plans and possible preprocessing of measurement data (Pietraszek and Goroshko, 2014). To avoid the need to analyze very large data sets, it is advisable to use dimensional reduction methods (Pietraszek and Skrzypczak-Pietraszek, 2015). This is highly desirable for both parametric and machine learning approaches. There may also be difficult issues where, for example, the data has been damaged by noise whose distribution has tails heavier than Gaussian (Echeverria and Green, 2019). Similar problems related to the estimation of uncertainty may also appear in the area of management, therefore the discussion presented in this article may be interesting for managers of education (Ulewicz, 2014) and trade (Ingaldi and Ulewicz, 2018).

5. Conclusion

The article discusses the difficulties that the DOE methodology encounters in the context of Industry 4.0. The four DOE methods – factorial, response surface, Taguchi and Shainin Red-X™ – and their characteristics were characterized. The main sources of difficulties that DOE methods encounter when used in the Industry 4.0 environment have been identified. These are large data streams, huge data sets, large data dimensions, non-Gaussian data distributions, non-linear relationships between controlled factors. The proposed directions of modifications, considered to be promising, include data-driven preprocessing combined with dimensional reduction, the use of non-parametric models taking into account highly

nonlinear relationships, the introduction of coupled multivariate descriptions using appropriate multivariate statistics, empirical identification of distributions using Monte Carlo methods.

Further investigations will include the collection of appropriate empirical datasets together with descriptions of use cases to verify the validity of the use for specific data analysis methods and determine their applicability and/or limitations, and if necessary, make necessary adaptations and develop detailed guidelines for use, especially when applying them inside other approaches e.g. SixSigma.

References

Bhote, K.R., Bhote, A.K., 2000. *World Class Quality. Using Design of Experiments to Make It Happen*, AMACOM, New York.

Box, G.E.P.,1957. *Evolutionary Operation: A Method for Increasing Industrial Productivity*, Journal of the Royal Statistical Society Series C-Applied Statistics, 6(2), 81-101, DOI: 10.2307/2985505

Box, G. E. P., Wilson, K. B., 1951. *On the Experimental Attainment of Optimum Conditions*, Journal of the Royal Statistical Society Series B-Statistical Methodology, 13(1), 1-45.

Capriz, G., Ghionna, V.N., Giovine, P., 2002. *Modeling and Mechanics of Granular and Porous Materials*, Boston, Birkhäuser. DOI: 10.1007/978-1-4612-0079-6

Echeverria-Rios, D. Green, P.L., 2019. *Gaussian processes for regression and classificationtasks using non-gaussian likelihoods*, UNCECOMP 2019 3rd ECCOMAS Thematic Conference on Uncertainty Quantification in Computational Sciences and Engineering, 555-568. DOI: 10.7712/120219.6359.18465

Fisher, R.A., 1921. *Studies in Crop Variation. I. An examination of the yield of dressed grain from Broadbalk*, Journal of Agricultural Science 11(2), 107-135, DOI: 10.1017/S0021859600003750

Fisher, R.A., 1925. *Statistical Methods for Research Workers*. Edinburgh, Oliver and Boyd Press.

Heinz, S., 2011. *Mathematical Modeling*, Heidelberg, Springer. DOI: 10.1007/978-3-642-20311-4

Ingaldi, M., Ulewicz, R., 2018. *Evaluation of quality of the e-commerce service*, International Journal of Ambient Computing and Intelligence 9(2), 55-66, DOI: 10.4018/IJACI.2018040105

Izenman, A.J., 2008. *Modern Multivariate Statistical Techniques. Regression, Classification, and Manifold Learning*, New York, Springer Science+Business Media.s

Karpisz, D., Kiełbus, A., 2018. *Selected problems of designing modern industrial databases*, MATEC Web of Conferences, 183, art. 01017, DOI: 10.1051/matecconf/201818301017

Kiefer, J., Wolfowitz, J., 1952. *Stochastic Estimation of the Maximum of a Regression Function*, The Annals of Mathematical Statistics, 23(3), 462-466. DOI: 10.1214/aoms/1177729392

Kiefer, J., Wolfowitz, J., 1959. *Optimum Designs in Regression Problems*, Annals of Mathematical Statistics, 30(2), 271-294, DOI: 10.1214/aoms/1177706252

Korzekwa, J., Bociąga, E., Bochenek, D., 2020. *Investigation of selected polymer composite-aluminum oxide coating tribological systems*, Materials, 13(23), art. 5491, DOI: 10.3390/ma13235491

Korzekwa, J., Gądek-Moszczak, A., Zubko, M., 2018. *Influence of the Size of Nanoparticles on the Microstructure of Oxide Coatings*, Materials Science, 53(5), 709-716, DOI: 10.1007/s11003-018-0127-x

Lipiński, T., 2015. *Modification of Al-11% Si alloy with Cl-based modifier*, Manufacturing Technology, 15(4), 581-587.

Lipiński, T., 2017. *Corrosion effect of 20 % NaCl solution on basic carbon structural S235JR steel*, Engineering for Rural Development, 16, 2017, 1069-1074, DOI: 10.22616/ERDev2017.16.N225

Lipiński, T., Wach, A., 2015. *Dimensional structure of non-metallic inclusions in high-grade medium carbon steel melted in an electric furnace and subjected to desulfurization*, Solid State Phenomena, 223, 46-53, DOI: 10.4028/www.scientific.net/SSP.223.46

Mardia, K.V., Jupp, P.E., 2000. *Directional Statistics*, Hoboken, Wiley.

Montgomery, D.C., 2020. *Introduction to Statistical Quality Control, 8th Edition*, Hoboken, Wiley.

Pacana, A., Czerwińska, K., Bednárová, L., 2018. *Discrepancies analysis of casts of diesel engine piston*, Metalurgija, 57(4), 324-326.

Pacana, A., Gazda, A., Malindzak, D., Stefko, R., 2014. *Study on improving the quality of stretch film by Shainin method*, Przemysł Chemiczny, 93(2), 243-245, DOI: 10.12916/przemchem.2014.243

Panasenko, G., 2005. *Multi-scale Modelling for Structure and Composites*, Dordrecht, Springer.

Patek, M., Konar, R., Sladek, A., Radek, N., 2014. *Non-destructive testing of split sleeve welds by the ultrasonic TOFD method*, Manufacturing Technology, 14(3), 403-407.

Pedrycz, W., Chen, S.-M. (eds.), 2013. *Time Series Analysis, Modeling and Applications*, Springer, Heidelberg, DOI: 10.1007/978-3-642-33439-9

Phadke, M.S., 1989. *Quality Engineering Using Robust Design*, London, Prentice Hall International, Inc.

Pietraszek, J., Dwornicka, R., Krawczyk, M., Kołomycki, M., 2017. *The nonparametric approach to the quantification of the uncertainty in the design of experiments modelling*, UNCECOMP 2017 - Proceedings of the 2nd International Conference on Uncertainty Quantification in Computational Sciences and Engineering, 598-604, DOI: 10.7712/120217.5395.17225

Pietraszek, J., Dwornicka, R., Szczotok, A., 2016. *The bootstrap approach to the statistical significance of parameters in the fixed effects model*, ECCOMAS Congress 2016 - Proceedings of the 7th European Congress on Computational Methods in Applied Sciences and Engineering, 3, 6061-6068, DOI: 10.7712/100016.2240.9206

Pietraszek, J., Goroshko, A., 2014. *The heuristic approach to the selection of experimental design, model and valid pre-processing transformation of DoE outcome*, Advanced Materials Research, 874, 145-149, DOI: 10.4028/www.scientific.net/AMR.874.145

Pietraszek, J., Skrzypczak-Pietraszek, E., 2015. *The uncertainty and robustness of the principal component analysis as a tool for the dimensionality reduction*, Solid State Phenomena, 235, 1-8, DOI: 10.4028/www.scientific.net/SSP.235.1

Rajak, D.K., Gupta, M., 2020. *An Insight Into Metal Based Foams*, Singapore, Springer, DOI: 10.1007/978-981-15-9069-6

Robbins, H., Monro, S., 1951. *A Stochastic Approximation Method*, The Annals of Mathematical Statistics, 22(3): 400-407, DOI: 10.1214/aoms/1177729586

Scheffé, H., 1958. *Experiments with Mixtures*, Journal of the Royal Statistical Society Series B-Statistical Methodology, 20(2), 344-360, DOE: 10.1111/j.2517-6161.1958.tb00299.x

Sherman, M., 2011. *Spatial Statistics and Spatio-Temporal Data. Covariance Functions and Directional Properties,* Hoboken, Wiley.

Shumway, R.H., Stoffer, D.S., 2017. *Time Series Analysis and Its Applications, With R Examples. Fourth Edition*, Springer, Cham, DOI: 10.1007/978-3-319-52452-8

Skrzypczak-Pietraszek, E., Skrzypczak, L., Wesolowska, M., 1993. *Gentiana tibetica King from tissue culture*, Scientia Pharmaceutica, 61(4), 287-296.

Szczotok, A., Roskosz, S., 2005. *New possibilities of light microscopy research resulting from digital recording of images*, Materials Science-Poland, 23(2), 559-565.

Szczotok, A., Sozańska, M., 2009. *A comparison of grain quantitative evaluation performed with standard method of imaging with light microscopy and EBSD analysis*, Praktische Metallographie/Practical Metallography, 46(9), 454-468, DOI: 10.3139/147.110043

Trzewiczek, K., Szczotok, A., Gadek-Moszczak, A., 2014. *Evaluation of the state for the material of the live steam superheater pipe coils of V degree*, Advanced Materials Research, 874, 35-42, DOI: 10.4028/www.scientific.net/AMR.874.35

Ulewicz, R., 2014. *Application of servqual method for evaluation of quality of educational services at the university of higher education*, Polish Journal of Management, 9, 254-264.

Włodarczyk, R., Dudek, A., Nitkiewicz, Z., 2011. *Corrosion analysis of sintered material used for low-temperature fuel cell plates*, Archives of Metallurgy and Materials, 56(1), 181-186, DOI: 10.2478/v10172-011-0021-0

Wojnar, L., Gadek-Moszczak, A. Pietraszek, J., 2019. *On the role of histomorphometric (stereological) microstructure parameters inthe prediction of vertebrae compression strength*, Image Analysis and Stereology, 38(1), 63-73, DOI: 10.5566/ias.2028

Wrońska, A., Dudek, A., 2014. *Characteristics of surface layer of sintered stainless steels after remelting using GTAW method*, Archives of Civil and Mechanical Engineering, 14(3), 425-432, DOI: 10.1016/j.acme.2013.10.011

Yates, F., 1935. *Complex experiments*, Supplement to the Journal of the Royal Statistical Society, 2(2), 181-247, DOI: 10.2307/2983638

DOE方法论与工业4.0引入相关的挑战

關鍵詞

美国能源部
工业4.0
数据驱动的方法
不确定

摘要

将通常称为"工业4.0"的解决方案引入该行业导致需要对基于DOE（实验设计）方法的传统工业数据分析程序进行许多更改。 所考虑的受控因素和观察因素数量的增加，数据流的强度以及所分析数据集的大小揭示了现有程序的缺点。 通过适应大数据解决方案和数据驱动方法来修改程序变得越来越迫切。 本文介绍了DOE的当前方法，考虑了工业4.0下引入大规模自动化和数据集成引起的现有问题，并指出了寻找可能的问题解决方案的最有希望的领域。