# Comparative analysis of two-dimensional data-driven efficient frontier estimation algorithms

Ilya Yuskevich
Skoltech Space Center
Skolkovo Institute of Science
and Technology
Moscow, Russia
Email: ilia.iuskevich@skolkovotech.ru

Rob Vingerhoeds
Dept. Engineering Complex Systems
ISAE SUPAERO
University of Toulouse
Toulouse, France
Email: rob.vingerhoeds@isae.fr

Alessandro Golkar
Skoltech Space Center
Skolkovo Institute of Science
and Technology
Moscow, Russia
Email: a.golkar@skoltech.ru

*Abstract*—**In this paper we show how the mathematical apparatus developed originally in the field of econometrics and portfolio optimization can be utilized for purposes of conceptual design, requirements engineering and technology roadmapping. We compare popular frontier estimation models and propose an efficient and robust nonparametric estimation algorithm for two-dimensional frontier approximation. The proposed model allows to relax the convexity assumptions and thus enable estimating a broader range of possible technology frontier shapes compared to the state of the art. Using simulated datasets we show how the accuracy and the robustness of alternative methods such as Data Envelopment Analysis and nonparametric and parametric statistical models depend on the size of the dataset and on the shape of the frontier.**

*Index Terms*—**Pareto frontiers, efficient frontiers, technology planning, requirements engineering, decision support systems**

## I. Introduction

Technology roadmapping requires forecasting the evolution of technology in order to address its impact in future system developments. This is particularly important when conceiving new developments for engineering products which development time spans years and involve multi-billion investments, such as in the case of aircraft and spacecraft systems. Technology evolution forecasting is traditionally performed by pure expert assessments, or expert assessment aided by computational approaches. Two classes of computational approaches in forecasting are the so-called model-based (physics driven) and data-driven (statistics driven) approaches.

The goal of model-based forecasting approaches is to identify potential physical limits affecting the evolution of the tradespace, and to assess the impact of the aforesaid limits on the tradespace of possible system evolutions. Model-based approaches explore the tradespace of possible system development options by accounting for all engineering disciplines involved, any applicable physical limit and potentially including other constraints coming from program developments, such as risk, schedule, and cost. An example of physical limit is the amount of speed increase that a single stage rocket can deliver, for a given stage mass fraction and propellant specific impulse, given by the vertical asymptote of the Tsiolkovski's rocket equation [1]. A second example of physical limit is the flight envelope in terms of payload and range for a given aircraft configuration, that is given by the Breguet equation [2]. The complexity of model-based forecasting approaches may vary from simple algebraic equations to more sophisticated multi-disciplinary simulations, such as for example Finite Element Models (FEM) or computational fluid dynamic simulations.

Notwithstanding this complexity, the general goal behind those approaches, however, remains the same. Tradespace exploration models estimate system performance in terms of a set of Figures of Merit (FoMs) using the model equations as 'transfer functions' translating design inputs into FoMs. System model is the function transforming design variables to FoMs. Current system developments are mapped within the modeled tradespaces using FoMs, and the distance between current systems and the Pareto (efficient) frontier gives a measure of the incremental technology evolution designers can expect for a given technology set. While model-based approaches set the limits for incremental evolution, they do not provide time references as to when a given evolution can be expected. Other drawbacks of model-based approaches include potential oversimplification of the models, leading to approximations and potentially under or overestimation of technology performance, and the complexity and time required to develop and validate such models in engineering practice.

Data-driven approaches take the forecasting problem from a different angle. Rather than estimating system performance using physics-based modeling, they consider real data of past system developments. Applying statistical approaches such as multi-variable regression and others, they then estimate technology performance as a function of prior observations. The advantage of these approaches lies in their relative simplicity, and in the abundance of available methods coming from the fields of statistics, econometrics, and engineering design. An example of this approach application to the problem of forecasting the automobile performances is discussed in [3].

The drawbacks of these approaches are the neglectance of hard limits dictated by physics, and the potential issue of statistical significance of the results achieved in cases with reduced sample sizes (which is often the case in engineering design).

In our research we are addressing two questions. What are the most frequent frontier shapes are common for the requirements engineering of a complex system, and, consequently, which statistical estimation model should be chosen. Our goal in this paper is to compare the accuracy of different approaches for data-driven technology forecasting, namely Data Envelopment Analysis (DEA), nonparametric and parametric frontier estimation models. As a result, we obtain initial insights on the validity of different computational methods.

The remainder of this paper is structured as follows. Section 2 provides a classification of efficient frontier estimation algorithms, further developing the characterization between model-based and data-based approaches. Section 3 describes the data-driven frontier estimation algorithms which accuracy is compared in this paper. The section as well defines the benchmark against which the comparison is made. Section 4 provides a numerical comparison of the accuracy of the algorithms and provides a discussion of the results. Section 5 draws conclusion from the paper and outlines directions for future research.

## II. Classification of the efficient frontier estimation algorithms

An integrating concept for these approaches is an efficient frontier which can be defined as a surface in the n-dimensional FoMs space which contains all efficient points we potentially could achieve. An efficient frontier is referred to as a Pareto frontier when for each point on the frontier it is impossible to improve one figure of merit (FoMs) without worsening the other FoMs.

Standard DEA models (Charnes, Cooper and Rhodes (CCR) [4] and Banker, Charnes and Cooper (BCC) [5]) are able to estimate only convex upward Pareto efficient frontiers. It is sufficient for estimating production-possibility frontiers which are the tradeoffs of producing the combination of competing goods with the limited resources. However, the curves representing possible tradeoffs between performances of a technical system are in a broader range of shapes due to the often nonlinear physical nature of underlying processes.

For example, on fig. 1 a is depicted convex-downward Pareto efficient frontier which can be found in analysing petrol engine car power vs miles per gallon fuel consumption measure [3].

U-shape frontiers (fig. 1 d, e) are common for aircraft fuel efficiency analyses. If FoM 1 is an aircraft velocity then the total airdrag (FoM 2) will have a U-shape as a compound of increasing with velocity parasitic drag and decreasing with velocity lift-induced drag [6].

The simplest way to obtain frontier with concavity-convexity changes (non-convex) is to analyse FoM composed from frontier of U-shape and convex downward Pareto efficient frontier (fig. 1 c).

In practice this happens when we are trying to estimate some broad technology field with complex physics behind. As an example, the power conversion efficiency depending on wavelength in solar cells exhibit such behaviour.

Let's now show that U-shape frontier is not Pareto efficient. Suppose ideally we want to maximize both FoMs. Then for low values of FoM 1 frontier behaves as usual Pareto-efficient frontier. For FoM 1 high values we observe simultaneous improvement of both FoMs (which is contradicting with definition of Pareto efficiency). This does not mean that optimal solutions for all applications are existing only in this part of frontier. Suppose FoM 1 in our example is the capacity of some transportation system. In general, we want to maximize capacity. However, capacity is often associated with size, which may be constrained for some applications.

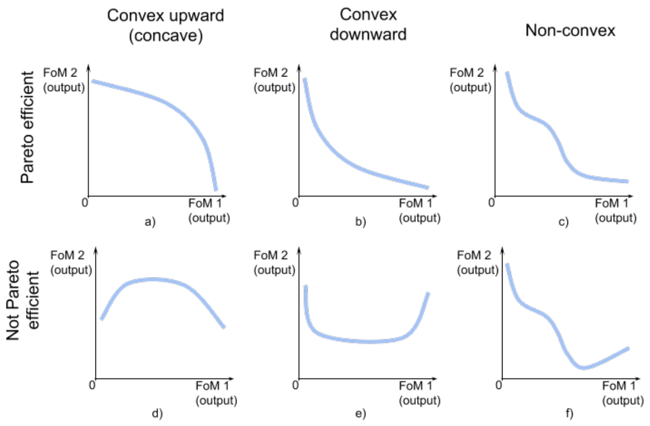

Fig. 1: All possible in a two-dimensional space classes of efficient frontier shapes (only upper left is DEA compatible).

In technology planning process it often makes sense, however, to explore all options regardless the application so to see whole picture.

It is therefore important to obtain a good estimation on frontiers of any shape, so to know what is technically achievable. The main goal of frontier estimation algorithms is to construct the frontier accurately and efficiently. The mathematical setting for model-based and data-driven cases is however completely different.

For model-based estimation, we have a model allowing us to generate as many points as needed. If the frontier is a continuous function then the number of points that could be generated theoretically unlimited. If the setting is discrete we can theoretically construct a full frontier as the number of possible values is limited.

In the case of data-driven estimation we have the points already generated in the form of market data and the number of points is therefore always limited; in this case we can never achieve a perfect information about the frontier.

In model-based estimation, the calculation of FoMs for each design point may take a significant amount of time (e.g. to calculate gain of the microwave antenna with specific configuration one may need to solve a set of electrodynamic differential equations using FEM). That is why many efforts in this field were devoted to the development of the special methods whose goal it is to generate solutions evenly distributed in the design space with as objective to reduce the computational time to solve the multiobjective optimization problem [7]–[9]. Some methods are indicating as an important sub-goal the ability to overcome the frontier concavity-convexity changes [7].

The classification of efficient frontier estimation algorithms is shown on figure 2.

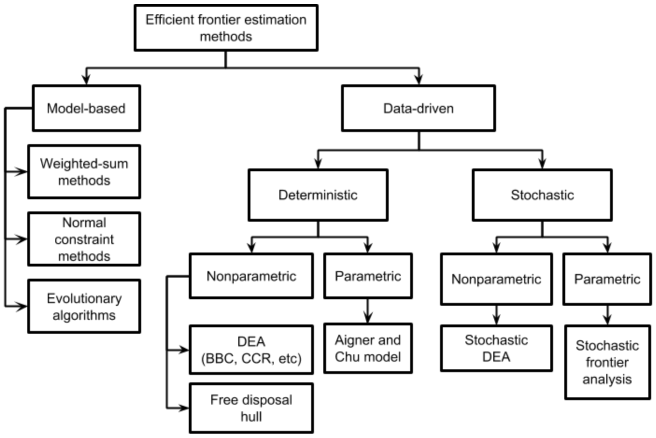

Fig. 2: Efficient frontier estimation algorithms classification.

## III. Data-driven efficient frontier estimation algorithms

Data-driven approaches have been considered mostly in the econometry rather than in engineering sciences. For purposes of conceptual design we are interested in the following properties of the estimation algorithm: proximity of the estimate to the real frontier and the ability to estimate all possible shapes of technology frontiers.

The most general statistical model of the efficient frontier was formulated within the framework of the Stochastic Frontier analysis (SFA) in [10]. For each i-th design point $(x_i, y_i)$ of the given set in n-dimensional FoM space we can write the following statistical model:

$$f(x) = f(x_i, \beta) + u_i + v_i \tag{1}$$

The function $f(x, \beta)$ is the efficient technology frontier that represents current best practice. $\beta$ is a vector of frontier's parameters to estimate. According to SFA, there are two types of random processes influencing performances of a design point (or decision making unit, DMU). The first one ($u_i \leq 0$) represents an internal inefficiency of each design point. It is called slack (see fig. 3) and has the one-sided distribution (any design point could lie in front of frontier, so slacks are always negative). The second random component $v_i$ is distributed symmetrically and represents factors that are out of control of a design team. This random component introduces the concept of the stochastic frontier. Authors of [10] have pointed out that best achievable performances could vary across DMUs due to external factors such as climate, policy, socio-economic situation, etc. Accordingly, if $\sigma(v) = 0$ the stochastic frontier model (1) turns into deterministic frontier model. In this paper we are discussing the latter class of frontier estimation models.

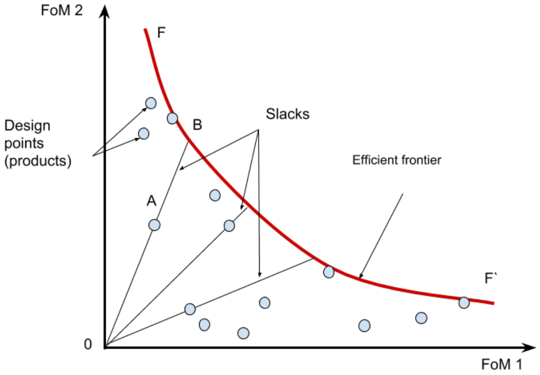

Fig. 3: Efficient frontier and finite number of design points.

In our numerical experiment for parametric approaches testing we will use the deterministic model proposed by Aigner and Chu [11]. The model of a design point with two conflicting FoMs $(x_i, y_i)$ can be represented in the following form:

$$y_i = x_i^{\beta_1} e^{\beta_0} e^{u_i}, \; u_i \leq 0 \tag{2}$$

Or in the logarithmic form:

$$\ln y_i = \beta_1 \ln x_i + \beta_0 + u_i, \; u_i \leq 0 \tag{3}$$

A frontier now can be estimated by solving the following quadratic optimization problem:

$$\begin{gathered}\min_{\beta_0, \beta_1} \Big( \sum_i \ln y_i - \beta_1 \ln x_i - \beta_0 \Big)^2 \\ \text{subject to } \beta_1 \ln x_i + \beta_0 \geq \ln y_i\end{gathered} \tag{4}$$

The model (3) estimates only convex in $(x, y)$ space frontiers. To estimate concave frontiers we propose to use the following modification of this model:

$$\begin{gathered}\frac{y_i}{y_{max}} = (1 - x_i^{\beta_1} e^{\beta_0}) e^{u_i} \rightarrow \\ \beta_1 \ln x_i + \beta_0 = \ln \left(1 - \frac{y_i}{y_{max}} e^{-u_i}\right)\end{gathered} \tag{5}$$

$$\begin{gathered}\min_{\beta_0, \beta_1} \Big( \sum_i \ln \left(1 - \frac{y_i}{y_{max}}\right) - \beta_1 \ln x_i - \beta_0 \Big)^2 \\ \text{subject to } \beta_1 \ln x_i + \beta_0 \geq \ln \left(1 - \frac{y_i}{y_{max}}\right)\end{gathered} \tag{6}$$

The main disadvantage of parametric models is that one need to define the frontier model $f(x, \beta)$. In case of conceptual design of a complex system an analytical model of a performance limiting curve has an unknown form. That is why nonparametric models are of the particular interest to us.

The most popular nonparametric model in econometrics is DEA. The idea of DEA is to calculate the relative efficiency of decision making units with multiple FoMs with the specific linear optimization procedure [4], [5].

$$\begin{aligned} &\min_{\lambda} \theta \\ &\text{subject to } \theta x_i - X\lambda \geq 0 \\ &Y\lambda \geq y_i \\ &e\lambda = 1 \\ &\lambda \geq 0 \end{aligned} \tag{7}$$

where $X$ and $Y$ are the vectors of input and output FoMs. $\lambda$ is the vector of weighting coefficients. $\theta$ is the output-oriented efficiency score of the j-th design point. The process is repetitive for each design point $(x_j, y_j)$.

CCR [4] and BCC [5] DEA models are able to estimate only concave frontiers. This drawback is major for our purposes. In engineering two or more conflicting FoMs could have different units of measure and therefore possess distinct physical natures (e.g. one FoM may represent mass and a second one cost). As result, frontiers may be both convex and concave.

To relax the concavity constraint, some other approaches are available.

The most trivial approach is to transform output-oriented FoMs to input oriented by the nonlinear transformation $(X' = Y^{-1})$ and solve the inputs minimization problem instead of outputs maximization with the same DEA estimation model. However, this is not the best option since the nonlinear transformation itself is a source of error.

Also, so called free disposal hull approach was presented in [12]. This model is free from any assumptions of a frontier's shape. Consequently, a set of nondominated points is the result of such procedure, which means the big estimation error, especially for small sets.

The existing models hence impose either redundant constraints on the frontier's shape, or lead to the poor approximation. Therefore, we propose our own nonparametric approach. We assume that in most cases technology frontiers are either concave or convex on the entire domain. Thus, the only constraint we want to put on the arbitrary frontier is the prohibition of convexity and concavity changes:

$$f(x)'' \geq 0 \; or \; f(x)'' \leq 0 \tag{8}$$

For sorted by $x$ finite set $(x_i, y_i)$ of nondominated design points we can formulate the following quadratic optimization problem:

$$\begin{aligned} &\min_{f_i} \sum_i (f_i - y_i)^2 \\ &\text{concave: } \frac{f_{i+1} - f_i}{x_{i+1} - x_i} \geq \frac{f_{i+2} - f_{i+1}}{x_{i+2} - x_{i+1}} \\ &\text{convex: } \frac{f_{i+1} - f_i}{x_{i+1} - x_i} \leq \frac{f_{i+2} - f_{i+1}}{x_{i+2} - x_{i+1}} \\ &f_i \geq y_i \end{aligned} \tag{9}$$

The relative distance from a design point to a corresponding efficient frontier with respect to the zero point is called radial efficiency (ratio between AB and OB on fig. 3). There are two ways to calculate the accuracy (or goodness of fit). It can be defined as an average distance between the estimated and the true efficient frontier or as an average absolute deviation between the radial efficiencies calculated for points with respect to the true frontier and to its estimate. We will use the latter in our numerical experiment in the following form:

$$\Delta = \frac{1}{N} \sum_{i=1}^{N} |\theta_{i,true} - \theta_{i,estimated}| \tag{10}$$

In real life, the true frontier is unknown but in this paper to perform the methods comparison we will generate the true frontier and the slacks to demonstrate the methods performances.

For illustrative purposes we will use three real technology frontiers with different curvature for generation of simulated datasets.

The first one is the Tsiolkovsky's rocket equation:

$$f\left(\frac{m_f}{m_0}\right) = -I_{sp}\, g_0 \,\ln \frac{m_f}{m_0} \tag{11}$$

where $\frac{m_f}{m_0}$ ratio of the final mass to the initial mass, $g_0$ – standard gravity, $I_{sp}$ – specific impulse (we will take $I_{sp}$ = 250 s).

The second one is the piecewise linear payload-range diagram for an airplane with the parameters: 50 ton of maximum payload, 5000 km range with maximum payload, 25 ton of payload with maximum fuel, 9000 km with maximum fuel, 11000 km range with zero payload.

And the third frontier is derived from the equation for a power:

$$P = \frac{W}{t} \tag{12}$$

where $P$ is power, $W$ is work (or energy), $t$ is time. This equation is generic and therefore very common for many conceptual design studies.

For all the frontiers we will generate random datasets consisted of 15 and 30 points with slacks distributed exponentially (with $\beta = 0.15$). Expected values and variances are calculated for 25 trials. The visualizations of the simulated datasets are shown on the figures 4-9. The numerical results are listed in the table 1.

## IV. Conclusions

In this paper we have conducted the numerical study in which the following empirical findings were obtained:

- In the field of conceptual design studies for the maximization problem of two conflicting FoMs both convex (Tsiolkovsky rocket equation and power equation) and concave (payload-range diagram) frontiers may be encountered.
- Nonlinear transformation of the outputs maximization problem to the inputs minimization problem in some cases does not preserve the form of frontier and therefore leads to transformation errors.

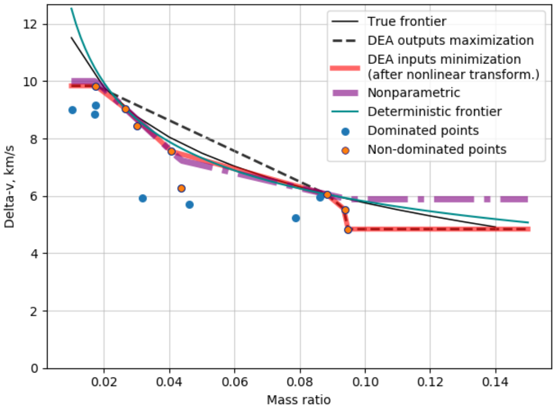

Fig. 4: Tsiolkovsky rocket equation, 15 points.

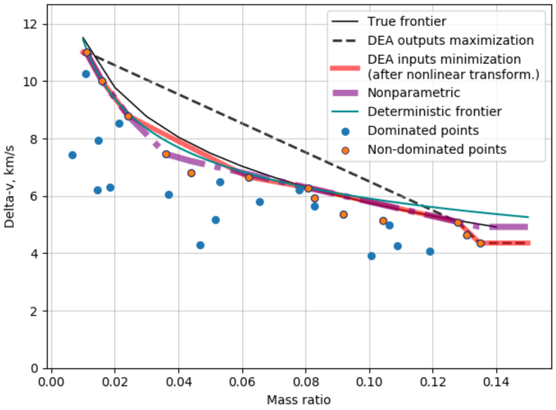

Fig. 7: Tsiolkovsky rocket equation, 30 points.

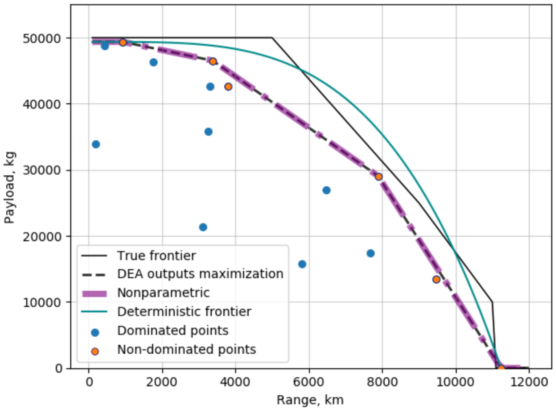

Fig. 5: Payload-range diagram, 15 points.

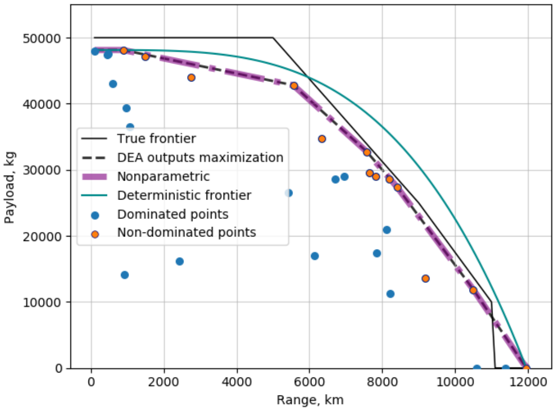

Fig. 8: Payload-range diagram, 30 points.

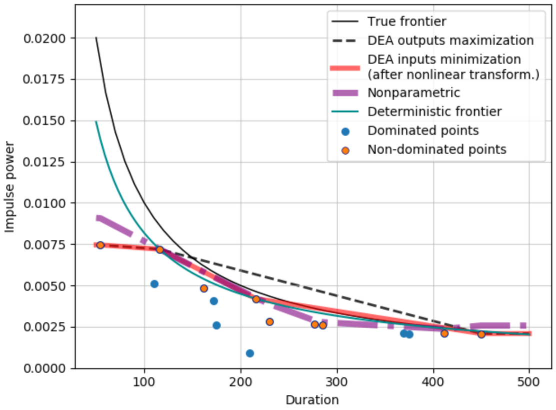

Fig. 6: Reciprocal function, 15 points.

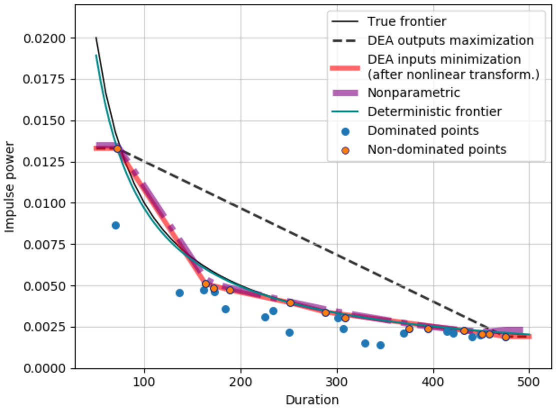

Fig. 9: Reciprocal function, 30 points.

TABLE I: Results of numerical experiment

| | **Expected value and variance of errors ($\mu$, $\sigma^2$), %** | | | |
|---|---|---|---|---|
| ***Model, number of points*** | ***DEA VRS*** | ***DEA Inputs transf.*** | ***Nonparam.*** | ***Parametric (Aigner and Chu)*** |
| Payload-range, 15 | (3.7, 1.3) | - | (3.7, 1.3) | (3.8, 1.2) |
| Tsiolkovsky, 15 | (4.3, 1.1) | (4.8, 1.7) | (3.6, 1.3) | (3.4, 1.9) |
| Reciprocal, 15 | (12.2, 3.1) | (5.6, 1.8) | (4.1, 1.8) | (2.4, 1.4) |
| Payload-range, 30 | (2.6, 1.1) | - | (2.6, 1.1) | (3.5, 1.1) |
| Tsiolkovsky, 30 | (5.2, 1.2) | (2.9, 0.7) | (2.1, 0.5) | (2.4, 1.0) |
| Reciprocal, 30 | (15.2, 2.9) | (3.7, 0.8) | (2.3, 0.6) | (1.0, 0.6) |

- With respect to DEA our proposed approach is equal in the case of concave frontiers estimation and more accurate in the case of convex frontiers.
- With respect to the parametric approach our proposed algorithm provides less accurate and robust estimates. However, in the case of inappropriate function form assumptions (see results for the Tsiolkovsky equation where the exponential assumption is not valid for the logarithmic function) our nonparametric approach provides better results. And since in systems engineering frontiers often take very complex forms and it is impossible to find an analytic form or an appropriate approximation curve for most of tradespace exploration problems, nonparametric approaches are preferable.

The importance of the stochastic frontier paradigm adoption for conceptual design is one of the direction for the further research. Undoubtedly, quite often the technical efficiency is not the only decisive factor due to protectionist measures. In this case a goal of a design team may be formulated as the achieving of the mediocre performances. On the other hand, the world-class performances may not be achievable for some regional manufacturers. However, such analysis requires the strong emphasis on the model's assumptions and the existence of more extensive datasets and therefore may be applied only to mass production industries.

Another direction for the further research is the possible extension of our nonparametric approach to the general n-dimensional form, which was left out of the current study scope.

## References

[1] M. J. Turner, *Rocket and spacecraft propulsion: principles, practice and new developments*. Springer Science & Business Media, 2008.

[2] H. Curtis, A. Filippone, M. V. Cook, T. Megson, M. Tooley, D. Wyatt, L. R. Jenkinson, J. Marchman, F. De Florio, and J. Watkinson, *Aerospace engineering desk reference*. Elsevier, 2009.

[3] I. Yuskevich, R. Vingerhoeds, and A. Golkar, "Two-dimensional Pareto frontier forecasting for technology planning and roadmapping," in *Systems Conference (SysCon), 2018 Annual IEEE International*, pp. 1–7, IEEE, 2018.

[4] A. Charnes, W. W. Cooper, and E. Rhodes, "Measuring the efficiency of decision making units," *European journal of operational research*, vol. 2, no. 6, pp. 429–444, 1978.

[5] R. D. Banker, A. Charnes, and W. W. Cooper, "Some models for estimating technical and scale inefficiencies in data envelopment analysis," *Management science*, vol. 30, no. 9, pp. 1078–1092, 1984.

[6] J. D. Anderson, *Aircraft performance and design*, vol. 1. WCB/McGraw-Hill Boston, MA, 1999.

[7] I. Y. Kim and O. L. de Weck, "Adaptive weighted-sum method for bi-objective optimization: Pareto front generation," *Structural and multidisciplinary optimization*, vol. 29, no. 2, pp. 149–158, 2005.

[8] A. Messac, A. Ismail-Yahaya, and C. A. Mattson, "The normalized normal constraint method for generating the Pareto frontier," *Structural and multidisciplinary optimization*, vol. 25, no. 2, pp. 86–98, 2003.

[9] E. Zitzler and L. Thiele, "Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach," *IEEE transactions on Evolutionary Computation*, vol. 3, no. 4, pp. 257–271, 1999.

[10] D. Aigner, C. K. Lovell, and P. Schmidt, "Formulation and estimation of stochastic frontier production function models," *Journal of econometrics*, vol. 6, no. 1, pp. 21–37, 1977.

[11] D. J. Aigner and S.-F. Chu, "On estimating the industry production function," *The American Economic Review*, pp. 826–839, 1968.

[12] D. Deprins, L. Simar, and H. Tulkens, "Measuring labor-efficiency in post offices," in *Public goods, environmental externalities and fiscal competition*, pp. 285–309, Springer, 2006.