UDC 330.115

T. ERMOLIEVA
International Institute for Applied Systems Analysis, Laxenburg, Austria, e-mail: *ermol@iiasa.ac.at*.

Y. ERMOLIEV
International Institute for Applied Systems Analysis, Laxenburg, Austria; V.M. Glushkov Institute of Cybernetics, National Academy of Sciences of Ukraine, Kyiv, Ukraine, e-mail: *ermoliev@iiasa.ac.at*.

P. HAVLIK
International Institute for Applied Systems Analysis, Laxenburg, Austria, e-mail: *havlik.petr@gmail.com*.

A. LESSA-DERCI-AUGUSTYNCZIK
International Institute for Applied Systems Analysis, Laxenburg, Austria, e-mail: *augustynczik@iiasa.ac.at*.

N. KOMENDANTOVA
International Institute for Applied Systems Analysis, Laxenburg, Austria, e-mail: *komendan@iiasa.ac.at.*

T. KAHIL
International Institute for Applied Systems Analysis, Laxenburg, Austria, e-mail: *kahil@iiasa.ac.at*.

J. BALKOVIC
International Institute for Applied Systems Analysis, Laxenburg, Austria, e-mail: *balkovic@iiasa.ac.at*.

R. SKALSKY
International Institute for Applied Systems Analysis, Laxenburg, Austria, e-mail: *skalsky@iiasa.ac.at*.

C. FOLBERTH
International Institute for Applied Systems Analysis, Laxenburg, Austria, e-mail: *folberth@iiasa.ac.at*.

P.S. KNOPOV
V.M. Glushkov Institute of Cybernetics, National Academy of Sciences of Ukraine, Kyiv, Ukraine, e-mail: *knopov1@yahoo.com*; *knopov1@gmail.com*.

G. WANG
China Agricultural University (CAU), Beijing, China, e-mail: *gangwang@cau.edu.cn*.

CONNECTIONS BETWEEN ROBUST STATISTICAL ESTIMATION, ROBUST DECISION=MAKING WITH TWO-STAGE STOCHASTIC OPTIMIZATION, AND ROBUST MACHINE LEARNING PROBLEMS[1]

Abstract. The paper discusses connections between the problems of two-stage stochastic programming, robust decision-making, robust statistical estimation, and machine learning. In the conditions of uncertainty, possible extreme events and outliers, these problems require quantile-based criteria, constraints, and "goodness-of-fit" indicators. The two-stage STO problems with quantile-based criteria can be effectively solved with the iterative stochastic quasigradient (SQG) solution algorithms. The SQG methods provide a new type of machine learning algorithms that can be effectively used for general-type nonsmooth, possibly discontinuous, and nonconvex problems, including quantile regression and neural network training. In general problems of decision-making, feasible solutions and concepts of optimality and robustness are characterized from the context of decision-making situations. Robust ML

[1]The development of robust decision-making, statistical estimation, machine learning and Big Data analysis problems, respective solution procedures and case studies, is supported by the joint project between the International Institute for Applied Systems Analysis (IIASA) and National Academy of Sciences of Ukraine (NASU) on "Integrated robust modeling and management of food-energy-water-land use nexus for sustainable development". The work has received partial support from the Ukrainian National Fund for Strategic Research, grant No. 2020.02/0121, and project CPEA-LT-2016/10003 jointly with Norwegian University for Science and technology. The paper contributes to EU PARATUS (CL3-2021-DRS-01-03, SEP-210784020) project on "Promoting disaster preparedness and resilience by co-developing stakeholder support tools for managing systemic risk of compounding disasters".

approaches can be integrated with disciplinary or interdisciplinary decision-making models, e.g., land use, agricultural, energy, etc., for robust decision-making in the conditions of uncertainty, increasing systemic interdependencies, and “unknown risks.”

Keywords: two-stage STO, robust decision-making and statistical estimation, robust quantile regression, machine learning, general problems of robust decision making, systemic risks, uncertainties.

INTRODUCTION

Various problems of decision-making under uncertainty, statistics, big data analysis, artificial intelligence (AI) can be formulated or can be reduced to the two-stage stochastic optimization (STO) problems. For example, these are problems inherent to engineering, economics, finance, operations research, etc., that involve minimization or maximization of an objective or a goal function when randomness is present in model’s data and parameters, e.g., observations, costs, prices, returns, crop yields, temperature, precipitation, soil characteristics, water availability, emissions, return periods of natural disasters, etc. Uncertain parameters can be interpreted as environment-determining variables [1], that conditions the performance of the system under investigation. Randomness can enter the problems in several ways:

1) through stochastic (exogenous or/and endogenous) parameters, e.g., costs, prices, returns, crop yields;

2) stochastic resources, e.g., water, land, biomass, investments;

3) random occurrence of exogenous natural disasters depleting resources and assets;

4) stochastic endogenous events (systemic risks) induced by decisions of various agents.

Non-normal probability distributions of stochastic parameters and percentile-based criteria functions. Stochastic variables can be characterized by means of a probability distribution (parametric or nonparametric) function or can be represented by probabilistic scenarios. Probability distributions of stochastic parameters are often non-normal, heavy tailed and even multimodal. For example, Fig. 1 depicts probability distribution of wheat yields for several countries — major grain producers. Horizontal axis denotes yield (in kilograms per hectare of harvested land) and vertical axis shows the number of years (frequency) the corresponding yield occurred in the 1960–2012 period. Cumulative distribution refers to the percentage of total of the yield occurrences at or below the value on the horizontal axis. Low wheat yields on the left-hand side visualized in all four panels can cause imbalances in grain supply-demand chains and thus lead to prices increase, market disturbances, trade bans, etc. These low values correspond to about 20 % percentiles of the crop yield observations and can correspond to production years characterized by bad weather conditions, e.g., low precipitation or/and high temperature in important grain growth periods/months (e.g., grain filling period).

If probability distribution functions of stochastic parameters are non-normal, non-symmetrical, or even heavy tailed, such decision-making or/and parameter estimation criteria as Mean-Variance, Ordinary Least Square (OLS) or Root Mean Squared Error (RMSE) are not appropriate. They rely only on the first two moments, i.e., mean and variance, which characterize normal distributions. Thus, the information about the tails of the distributions is not accounted for. Extreme values (outliers) can distort the results, i.e., the criteria are not robust. In statistics and machine learning, the OLS and RMSE estimates can be misleading, and the effects can be different for different subsets of data sample.