3

Robust statistical estimation and two-stage stochastic optimization: quantile regression EPIC meta-model of Soil Organic Carbon for robust decision-making with GLOBIOM

T.Y. Ermolieva, P. Havlik, A. Lessa-Derci-Augustynczik, S. Frank,
A. Deppermann, A. Nakhavali J. Balkovic, R. Skalsky, N. Komendantova

Tatiana Ermolieva ermol@iiasa.ac.at
Petr Havlik havlik.petr@gmail.com
Andrey Lessa-Derci-Augustynczik augustynczik@iiasa.ac.at
Stefan Frank frank@iiasa.ac.at
Andre Deppermann depperma@iiasa.ac.at
Andrè (Mahdi) Nakhavali nakhavali@iiasa.ac.at
Juraj Balkovic balkovic@iiasa.ac.at
Rastislav Skalsky skalsky@iiasa.ac.at
Nadejda Komendantova komendan@iiasa.ac.at

[1]International Institute for Applied Systems Analysis, Laxenburg, Austria

Corresponding author: T.Y. Ermolieva, ermol@iiasa.ac.at

Abstract

The paper discusses the connections between two-stage stochastic optimization and robust statistical estimation. Main question related to statistical predictions is how to use the predictions to optimize the overall decisions and how current decisions can affect predictions. In general problems of decision-making, feasible solutions, concepts of optimality and robustness are characterized from the context of decision-making situations, i.e., systems structure, goals, security constraints, safety norms, supply-demand relationships, thresholds. Robust statistical approaches can be effectively combined with disciplinary or interdisciplinary models, e.g., land use model GLOBIOM, for effective decision-making in the conditions of uncertainty, increasing interdependencies and systemic risks. We discuss a quantile- regression EPIC meta-model for tracking dynamics and uncertainties of Soil Organic Carbon (SOC), which is an important agri-environmental indicator. SOC levels (quantiles) can be controlled with GLOBIOM, to analyze costs and robust land management practices to sequester SOC and fulfill food-energy-water-environmental NEXUS security goals. Quantiles identify critical SOC levels signaling how close is a threshold or a targeted level. The SOC-EPIC meta-model is developed using multisource data from historical observations and results of the bio-physical model EPIC. It enables the analysis of SOC content and respective probabilities as a function of exogenous parameters such as monthly temperature and precipitation and endogenous, decision-dependent parameters, which can be altered by the land management decisions computed with GLOBIOM.

Key words: two-stage STO, robust decision-making and statistical estimation, quantile regression, uncertainties, agri-environmental indicators, soil organic carbon, GLOBIOM

3.1. Introduction

Interdependencies among food, energy, water, environmental systems are increasing. Proper management (control) of such interdependent systems becomes a challenging multidisciplinary problem. Ensuring robust and sustainable performance of the systems in the face of uncertainty and risks is equivalent to equipping the systems with measures that prepare them in advance and facilitate their proper adaptive (operational) responses to changing conditions, minimizing chances of

critical imbalances, thresholds' exceedance, and thus systemic failures due to combinations of various possible uncertain conditions/scenarios.

The robust interdependent ex-ate and ex-post decisions account for various risk-adjusted goals and constraints of the involved systems and agents (Ermoliev and Hordijk 2003; Ermoliev and von Winterfeldt 2012; Ermolieva *et al*. 2016; Ermolieva *et al.* 2022). The risk-adjusted safety (or security) constraints are imposed as critical levels of vital socio-economic, resource, environmental indicators analyzed by robust statistical and machine learning methods (Ermolieva *et al.* 2021). For example, quantiles identify different air and soil pollution levels and their respective probabilities signaling how close is the environmental pollution threshold. Quantile-based analysis of socio-economic, environmental, and demographic indicators allows, e.g., to distinguish population by their socio-economic and environmental vulnerability levels, agricultural production by climatic factors and soil types. The quantile regression addresses the questions regarding what percentage of population leaves in highly polluted areas or in water scarce regions, or the relation between the high air/water pollution and health indicators, or levels of crop yields and soil and climate characteristics (Ermolieva *et al*. 2023). In interdependent natural and anthropogenic systems, it may be possible to controlled critical indicators by decisions. For example, agri-environmental indicators reflecting soil health characteristics depend on land use practices and, therefore, can be improved by robust and sustainable land use management.

Thus, in general problems of robust decision-making, feasible solutions, concepts of optimality and robustness are characterized from the context of decision-making situations, i.e., systems structure, goals, constraints, safety norms, supply-demand relationships, thresholds. In statistics and machine learning, robustness property means that the solution is insensitive to outliers, i.e., additional observations/data cannot significantly affect the solution/estimate (Huber 1981; Vapnik 1995; Knopov 2002). 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, 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 (Ermoliev 1976; Ermoliev and Wets 1988; Ermoliev and Hordijk 2003; Borodina *et al.* 2020; Ermolieva *et al*. 2016; Gorbachuk *et al.* 2019; Ermolieva *et al.* 2022), that condition the performance of the system under investigation.

Stochastic variables can be characterized by means of probability distribution (parametric or nonparametric) functions or can be represented by probabilistic scenarios. Typically, in problems of managing environmental pollution, catastrophic risks, food-water-energy-environmental nexus security, probability distributions of stochastic parameters are non-normal, heavy tailed and even multimodal. For decision-making, statistical estimation and machine learning problems in the presence of non-normal, heavy tailed and possibly multimodal probability distributions it is appropriate to use quantile-based criteria instead of mathematical expectations. The problems can be formulated in the form of two-stage STOs with discontinuous chance (probabilistic or quantile-based) constraints defining vital threshold levels, e.g., food-energy-water-environmental security indicators, used for robust decision-making under uncertainty and risks and for FEWE nexus security management (Ermolieva *et al.* 2016; Ren *et al.* 2018; Ermolieva *et al.* 2021; Gao et al. 2021; Ermolieva *et al.* 2022).

3.1.1. Two-stage decisions

Often, decisions (actions) or parameter estimation have to be performed ex-ante before the values (realizations) of the uncertain parameters become known or observed (Ermoliev 2009a; Ermolieva *et al.* 2022). Sometimes, the observations can be only partial or incomplete, i.e., incomplete "learning". These situations happen, for example, in the process of agricultural production planning under weather variability and market risks (Ermolieva *et al.* 2021, 2022), water reservoir management (Ortiz-Partida *et al.* 2019), investments in irrigation and crop storage facilities (Ermolieva *et al.* 2022), energy technologies investments planning (Ermoliev *et al.* 2023), and in many other application problems.

The ex-ante decisions may require revisions and corrections after receiving additional information (i.e., after "learning" or partial "learning" of uncertain parameters values). Therefore, the ex-ante decisions/solutions can incur costs for their correction, revision, or reversion. Thus, there are two types (two-stage) of decisions. The ex-ante decisions x in the face of uncertainty may be a "here-and-now" decision whereas ex-post decisions y correspond to all future actions to be taken in different time periods in response to the environment created by the chosen x and the observed value of the uncertain parameter ω in that specific time period. The x and y solutions may represent sequences of interdependent ex-ante and ex-post control actions over a given time horizon. In the case of dynamical systems, there may also be an additional group of variables z characterizing states of the system in different time periods. These problems often emerge in operations research models, economics and system analysis, in the theory of optimal control and its applications in engineering, inventory control, etc. Specific applications include: deriving parameters of a statistical model

(parametric or nonparametric) or training a machine learning model that maps an input to an output based on examples of input-output pair through minimizing/maximizing a quantile-based "goodness-of-fit" criteria; deciding on optimal dynamic investments allocation into new technologies (irrigation, energy, agricultural, water management) to minimize costs or/and maximize profits accounting for various norms and constraints; deciding when to release water from a multipurpose reservoir for hydroelectric power generation, agricultural and industrial production, household requirements, environmental constraints, food protection; defining insurance coverage and premiums to minimize risk of bankruptcy of insurers and risk of overpayments of individuals.

3.1.2. Basic model of a two-stage decision-making and parameter estimation.

Let us illustrate the concept of the two-stage decision making and robust statistical estimation problem with an example of a simplest two-stage STO model. The idea of this model is implemented in stochastic GLOBIOM (Ermolieva *et al.* 2016, 2021). Assume, there are observations of an uncertain variable ω , which can be associated with a stochastic parameter, e.g., demand for a certain product or resource (water level). The stochastic variable ω can define the uncertain level of pollution or catastrophe losses to be mittigated ex-ante.

The choice of the decision $\bar{x} \geq x \geq 0$, to match the stochastic variable ω can be associated with a function $f(x, \omega)$ reflecting costs of overestimation and underestimation of ω. In the simplest case, $f(x, \omega)$ is a random piecewise linear function $f(x, \omega) = \max\{\alpha(x - \omega), \beta((\omega - x)\}$, where α defines the unit overestimation/surplus cost and β is the unit underestimation/shortage cost (associated e.g., with implementing measures ex-post, which also can be interpreted as costs of imports, borrowing, or costs of ex-post emergency actions and recovery). The problem is to find the level x that is "optimal", in a sense, for all foreseeable random scenarios/observations ω.

The expected cost criterion leads to the minimization of the following function:

$$F(x) = \max\{\alpha(x - \omega), \beta((\omega - x)\}$$

subject to $\bar{x} \geq x \geq 0$ for a given upper bound x. This stochastic minimax problem is also reformulated as a two-stage stochastic programming.

The optimal solution minimizing $F(x)$ and more general stochastic minimax problems defines quantile type characteristics of solutions (Ermoliev 2009a,b; Ermolieva *et al.* 2022), e.g., CVaR risk measures. For example, if the distribution of ω has a density, $\alpha, \beta > 0$, then the optimal solution x minimizing $F(x)$ is the quantile defined as $Pr\{\omega \leq x\} = \beta/(\alpha + \beta)$.

The problem of minimizing $F(x)$ illustrates the essential difference between the so-called scenario analysis aiming at the straightforward calculation of $x(\omega)$ for various scenarios of ω and the STO optimization approach. The STO model produces one solution that is optimal ("robust") against all possible stochastic scenarios ω).

3.2. Concept of robustness in statistical and general decision-making problems

3.2.1. *Stochastic optimization and safety quantile-based constraints*

A rather general STO problem is formulated as the maximization (minimization) of the expectation function

$$F_0(x) = Ef_0(x, \omega) = \int f_0(x, \omega)P(d\theta), \quad [3.1]$$

subject to constraints

$$F_i(x) = Ef_i(x, \omega) = \int f_i(x, \omega)P(d\theta) \geq 0, i = 1, \dots, m. \quad [3.2]$$

The choice of goal function $f_0(x, \omega)$ and indicators $f_i(x, \omega)$ is essential for the robustness of x. By choosing appropriate functions $f_0(x, \omega)$ and $f_i(x, \omega)$, STO models allow in a natural and flexible way to represent various risks, spatial, social, and temporal heterogeneities, and the sequential resolution of uncertainty in time. Often, as in Example below, constraints $f_i(x, \omega), i = 0, \dots, m$ are analytically intractable, nonsmooth, and even discontinuous functions, and probability measure P is unknown, or only partially known, and may depend on decisions x.

Moreover, decisions x according to a two-stage STO can be composed of anticipative precautionary (mitigation) ex-ante and adaptive (operational) coping ex-post components, which allows to model dynamic decision making processes with flexible adaptive adjustments of anticipative decisions when new information is revealed. The main challenge confronted by STO theory is that it may be practically impossible to evaluate exact values of $F_i(x)$, $i = 0,1, \dots, m$, see, e.g., Example. As

"deterministic" is a degenerated case of "stochastic", STO methods allow to deal with problems which are not solved by standard deterministic methods.

Example: Pollution control and nutrients balance accounting. A common feature of most models used in designing pollution-control policies and nutrients balance accounting (Balkovič *et al.* 2014; Ermolieva *et al.* 2024 and references therein) is the use of transfer coefficients a_{ij} that link the amount of pollution (nutrient) x_j released by source j to the pollution/nutrient concentrations $g_i(x,\omega)$ at the receptor location i as $g_i(x,\omega) = \sum_{j=1}^{n} a_{ij}x_j, \quad i = 0,1,\dots,m.$ The coefficients often depend on meterorological conditions, soil properties, etc. In complex problems, the transfer coefficients a_{ij} are also stochastic values with intractable decision-depedent probability distribution.

The deterministic models ascertain cost-effective pollution/nutrients control strategies x_j, $j = 1,\dots,n$ subject to achieving exogenously specified environmental targets, such as standard/norms b_i at receptors $i = 1,\dots,m$. These models can be improved by the inclusion of safety constraints that account for the random nature of coefficients a_{ij} and ambient standards b_i to reduce impacts of extreme events or exceedance of certain environmnetal thresholds:

$$F_i(x) = Prob[\sum_{j=1}^{n} a_{ij}x_j \le b_i] \ge p_i, i = 1,\dots m, \qquad [3.3]$$

namely, the probability that the accumulated nutrients level in each receptor (region, grid, or country) i will not exceed uncertain critical load b_i at a given probability (acceptable safety level) p_i. The critical load b_i can be identified by experts or through the quantile-based robust statistical or machine learning approaches, as it it discussed in section 3.3.

The constraints [3.3] are known as chance constraints (Ermoliev and Wets 1988; Ermoliev and Hordijk 2003; Ermolieva *et al.* 2016, 2021, 2022, 2013). They can be written in the form of the standard STO model with discontinuous functions: $f_j(x,\omega) = 1 - p_i$ if $\sum_{j=1}^{n} a_{ij}x_j - b_i \le 0$ and $f_j(x,\omega) = -p_i$, otherwise. If $p_i = 1$, $i = 1,\dots,m$, the constraints [3.3] are reduced to constraints of deterministic robustness.

The main computational complexity confronted by STO methods is the lack of explicit analytical formulas for goal functions $F_i(x)$, $i = 0,1,\dots,m$. For example, consider constraints [3.3]. If there is a finite number of possible scenarios $\omega = (a_{ij}, b_i, i = \overline{1,m}, j = \overline{1,n})$ reflecting, say, prevailing weather conditions, then $F_i(x)$

are piecewise constant functions, i.e., gradients of $F_i(x)$ are 0 almost everywhere. Hence, the straightforward conventional optimization methods cannot be used.

Ignorance of risks defined by constraints [3.3] may cause irreversible catastrophic events. Although an average daily concentration of a toxicant in a lake is far below a vital threshold, real concentrations may exceed this threshold for only a few minutes and yet be enough to kill off fish. Constraints of the type [3.3] are important for the regulation of stability in the insurance industry, known as the insolvency constraints. The safety regulation of nuclear reactors requires $p_i = 1 - 10^{-7}$, i.e., a major failure occurs on average only once in 10^7 years. Stochastic models do not, however, exclude the possibility that a disaster may occur next year.

2.2.2. *New problems of statistics and stochastic optimization*

Standard statistical problems are formulated as the minimization of the type [3.1] functionals in the case when the probability measure P is unknown but the sample ω^1, …, ω^N of observations drawn randomly according to P is available. It is assumed that P does not depend on X. In general problems of robust decision maaking, the exact evaluation of $F(x)$ can be practically impossible due to various reasons: probability measure $P(x, d\omega)$ is unknown or only partially known, random function $f(x, \omega)$ is analytically intractable, or the evaluation of $F(x)$ is analytically intractable despite well-defined $f(x, \omega)$ and P.

Statistics (statistical decision theory) deals with situations in which the model of uncertainty and the optimal solution are defined by unknown sampling model P. The main issue is to recover P by using available samples. In other words, the desirable optimal solutions $x = x^*$ is associated with P (or its parameters), the performance of x^* can be observed from available random data on its performance.

STO models were introduced for decision making problems under uncertainty arising in operation research and systems analysis which are typically described by a large number of decision variables and uncertainties. These models deal with fundamentally different situations. The uncertainty, feasible solutions, and performance of the optimal solution are not given by the sampling model. All of these have to be characterized from the context of the decision-making situation. As a consequence, multiple performance indicators, constraints, and dependencies among decisions and uncertainties play a key role. Thus, in STO, which in fact arose as an extension of linear and non-linear programming with their sophisticated computation techniques, the accent is on solving problems (1), (2) with large number of decisions variables, random parameters and constraints.

The classical statistics has been developed on the basis of asymptotic analysis requiring large samples of historical data. New important problems in statistics and STO have to confront situations with small data samples, cases of missing observations or absence of direct observations. These new problems require explicit joint treatment of all relevant interdependent observable, partially observable and non-observable variables by using various prior information in the form of additional constraints describing these interdependencies. These leads to high dimensions. The key issue is the representation of interdependencies enabling to organize pseudo-sampling based on proper characterization of probability measure P by using all available information.

Consequently, these new problems are formulated as general constrained STO problems where estimation of unknown probability measure P is directly associated with goals of overall decision-making problem. Since only specific data are essential for desirable decisions, the combined consideration of statistical estimation within overall decision-making problem can considerably reduce the quality and quantity of estimated information, e.g., the accuracy of the true parameters of P including even requirements on the uniqueness of P.

Consider some important statistical estimation problems which can be formulated as STO model [3.1-3.2]. Instead of asymptotic analysis, this provides the natural criterion of efficiency which can be used to evaluate the convergence to optimal solutions with respect to increasing number of real observations, resampling schemes, and pseudo sampling procedures. This section characterizes also loss functions which are typical for statistics.

2.2.3. *Regression estimation*

Assume that a random function $u(v)$ for each element v from a set V corresponds a random element $u(v)$ of the set U. Assume that $V \subset R^l$, $U \subset R^1$. Let P is a joint probability measure defined on pairs $\theta = (u, v)$. The regression function is defined as the conditional mathematical expectation

$$r(v) = E(u|v) = \int uP(U|v). \qquad [3.4]$$

It is easy to see that $r(v)$ minimizes the functional (providing it is well defined)

$$F(x(v)) = E(u(v) - x(v))^2, \qquad [3.5]$$

where $Eu^2(v) < \infty$, $Ex^2(v) < \infty$.

It follows from the fact that

$$\min_{x(v)} F(x(v)) = E \min_{x} E[(u(v) - x)^2|v],$$

$\frac{d}{dx} = E[(u(v) - x)^2|v] = -2(E(u|v) - x) = 0.$

The estimation of $r(v)$ is usually considered in the set of functions given in a parametric form $\{r(x, v), x \in X\}$. In this case, the criterion (5 4) can be rewritten as

$F(x) = E(u(v) - r(x, v))^2 = E(r(v) - r(x, v))^2 + E(u(v) - r(x, v))^2,$

i.e., the minimum of $F(x)$ is attained at the function $r(x, v)$ which is close to $r(v)$ in the metric $L_2(P)$ defined as $\sqrt{E(r(v) - r(x, v))^2}$

2.2.3. *Quantile based regression*

The conditional expectation $r(v)$ provides a satisfactory representation of stochastic dependencies $u(v)$ when they are well approximated by two first moments, e.g., for normal distributions. For general (possibly, multimodal) distributions it is more natural to use the median or other quantiles instead of the expectation. Let us define the quantile regression function $r_\rho(v)$ as the maximal value y satisfying equation

$P(u(v) \geq y|v) = \rho(v),$ [3.6]

where $0 < \rho(v) < 1$. It can be shown that function $r_\rho(v)$ minimizes the functional

$F(x(v)) = E(\rho(v)x(v) + max\{0, u(v) - x(v)\})$ [3.7]

This is due to the following. First of all, we have

$\min\limits_{x(v)} F(x(v)) = E\min\limits_{x} E[\rho(v)x + max\{0, u(v) - x(v)|v\}].$

Assume that probability $P(d\theta)$ has continuous density function $p(\theta)$, $P(d\theta) = p(\theta)d\theta$. Then from the optimality condition for internal stochastic minimax problem follows that optimal solution x satisfies the equation:

$\rho(v) - \int_x^\infty P(d\theta|v) = 0,$ [3.8]

i.e., indeed, it satisfies (6). Let us note, that the minimization of more general at the first glance functional

$F(x(v)) = E(a(v)x(v) + max\{\alpha(v)(u(v) - x(u)), \beta(v)(x(v) - u(v))\}),$

is reduced to the minimization of (7) with $\rho(v) = (a + \beta)(\alpha + \beta)^{-1}$. The median corresponds to the case when $a \equiv 0$, $\alpha = \beta$. The existence of optimal solution requires $a < \alpha$.

3.3. Quantile based machine learning regression model for tracking the dynamics and uncertainties of soil organic carbon in agricultural soils using multisource data

In this section we introduce a quantile-based statistical (or machine learning) regression model for estimating and predicting annual Soil Organic Carbon (SOC) stocks and stock changes at plow depth under the variability and changing seasonal patterns of temperature and precipitation. Soil Organic Carbon and Soil Organic Matter are important agri-environmental indicators characterizing soil health. The quantile-based linear regression is a two-stage STO model similar to the described in section 2.1.2. It enables estimation of quantiles, in particular, critical environmental loads and thresholds as it is discussed in Example of pollution/nutrients accumulation processes.

2.3.1. *Motivation*

The monitoring, modelling, and mapping of agri-environmental indicators, in particular, SOC, is important for many reasons. SOC is an indicator for soil organic matter (SOM) content, which is a major determinant of soil quality and fertility for food production. Soils with higher SOC can better filter, degrade organic molecules and purify water. SOC accumulation can substantially contribute to climate change mitigation (see discussion and references in Ermolieva *et al*. 2024). SOC stock is a Land Degradation Neutrality indicator used by the United Nations Convention to Combat Desertification (UNCCD). The EU Soil Strategy for 2030 contributes to the objectives of the EU Green Deal and is a part of the Biodiversity Strategy.

Soils have recently become part of the global carbon agenda for climate-change mitigation and adaptation. The “4p1000 initiative” was launched at COP21 by UNFCC under the framework of the Lima-Paris Action Plan (LPAP) in Paris on December 1, 2015. The name of the initiative reflects that a comparatively small proportional increase (4%) of the global SOC stocks in the top-soil of all non-permafrost soils would be similar in magnitude to the annual global net carbon dioxide (CO_2) growth.

The new strategy updates the 2006 EU Soil Thematic Strategy and intends to address land degradation trends. The EU Mission Board for Soil Health and Food proposed a series of quantitative targets to make soils of Europe healthier. Among

them, the aim is to reverse the current SOC concentration losses in croplands (0.5%/yr on average at 20 cm depth) to an increase of 0.1–0.4%/yr by 2030.

Thus, the quantiles (e.g., low and high) of SOC in agricultural soils can serve as targets in a land use model GLOBIOM for estimating optimal land use practices to increase SOC to required levels.

3.3.2. *SOC quantile regression meta model*

The SOC meta-model operates at different spatial scales and provides an effective means for scaling biophysical and land use models' results to required resolutions. By introducing SOC constraints, e.g., equal to the 50th or 75th quantile as estimated from the meta-model, GLOBIOM model (Havlik *et al.*, 2011; Ermolieva *et al.*, 2016, 2021, 2022) can derive an optimal combination of land use practices increasing SOC to the desired level. SOC and other food-energy-water-environmental security constraints identify the overall costs of achieving the food-water-energy-environmental NEXUS security. In fact, in our research, for each EU NUTS2 regions, a separate quantile-based meta model is estimated from historical data and results generated by bio-physical process-based model EPIC (Balkovič *et al.* 2014), at the level of grid cells. The NUTS (Nomenclature of territorial units for statistics) is the EU division of each EU country into 3 levels. NUTS2 is the level of country-specific basic regions.

NUTS2-level SOC quantiles are approximated by fitting separate quantile-based regression models. In classical LR approaches, the regression coefficients (β coefficients) represent the mean increase in the response variable produced by one unit increase in the associated explanatory variables. The β-coefficients obtained from QR represent the change in a specific quantile of the response variable produced by a one unit increase in the associated driver. In this way, QR allows to study how certain drivers affect median (quantile τ=0.5) or extremely low (e.g., τ=0.05) or high (e.g., τ=0.95) SOC stock values. Therefore, it gives a more comprehensive description of the effect of SOC predictors on the whole SOC stock level and the probability distribution (i.e., not just the mean) and may be used to analyze differential SOC stock responses to land practices.

For a random sample $X_1, X_2, \dots, X_n, \dots$ with empirical distribution function $\hat{F}_X(x)$, the τth empirical quantile function can be defined as $\hat{Q}(p) = \hat{F}_X^{-1}(x) = \inf\{x: \hat{F}_X(x) \geq \tau\}$. The τth empirical quantile can be determined by solving the minimization problem

$\hat{Q}(p) = argmin_x\{\sum_{i|X_i \geq x} \tau|X_i - x| + (1-\tau)\sum_{i|X_i<x}|X_i - x|\}.$

For the linear quantile regression, we make an assumption that the τth quantile is given as a linear function of the explanatory variables. For quantile regression, it is possible to calculate any quantile (percentage) for particular values of the dependent variables. Solving the problem for all $\tau \in [0,1]$, it is possible to recover the entire conditional quantile function, i.e., the conditional distribution function, of Y. Coefficients $\beta_m(\tau)$ are functions of the required quantile τ. They are defined as

$\beta(\tau) = argmin_{\beta \in R^m}\{\sum_{i|Y_i \geq \beta'(\tau)X_i} \tau|Y_i - \beta'(\tau)X_i| + (1-\tau)\sum_{i|Y_i<\beta'(\tau)X_i}|Y_i - \beta'(\tau)X_i|\},$

where Y_i are observations of dependent variables, X_i is a vector of independent variables $X_i = (x_{i1}, \dots, x_{im})$, and $\beta(\tau)$ is a vector of coefficients $\beta(\tau) = (\beta_1(\tau), \dots, \beta_m(\tau))$, and m is a number of observations. The QR models give much deeper insights into the complete conditional distribution of SOC stock values as a function of spatial and temporal predictors. By focusing on low (or high) quantiles, regression coefficients inform us about predictors that mainly influence the absence (or presence) of high/low SOC stock over space. By considering independent QR models for different values of τ, this allows for the possibility that the importance of certain predictors may change according to SOC level.

The estimates of the SOC quantile level $Q_\tau(y_i)$ in each SimUs within all NUTS2 regions and EU countries have a probability of

$$Prob\{Q_\tau(y_i) \leq \beta_0(\tau) + \beta_1(\tau)x_{i1} + \beta_2(\tau)x_{i2} + \beta_3(\tau)x_{i3} + \cdots + \beta_m(\tau)x_{im}\} = \tau.$$

The equation means that 100τ percent of the SOC observations/data are less than the value of the τ-quantile.

3.3.3. *Selected results*

The linear regression (LR) analysis of the relationship between the response variable (SOC) and the set of covariate variables was carried out to establish the reference scenario for comparing the SOC quantiles with the mean value predictions.

Trained on EPIC model inputs and results, the estimated NUTS2-specific LR meta-models have an R^2 of about 0.9 to 0.98 for all NUTS2 regions. The estimated QR trends identify the ranges and the respective probabilities of possible SOC content in different years. Figures 1–4 display the SOC content change between the consequent years for NUTS2 regions in the period from 1980 to 2020 in mean change,

the percentage difference between the 50th quantile and the mean value, the 75th quantile change, and the 25th quantile change, respectively, in t/ha.

In Figure 1, more brownish colors indicate the decrease in SOC between the years, and the greenish point to the NUTS2 regions with positive changes between the consequent years. In the upper-left panel of Figure 1, the mean changes in SOC are positive in Central Europe, i.e., the SOC stocks increased. However, the decreasing accumulation of SOC stocks can be observed already in the period from 1985 to 1995; the upper-right panel has less green color when compared to the upper-left one. More of a rapid decumulation of SOC stocks is observed in the southern countries of Europe such as Spain and Portugal. The SOC loss slows down in the north, especially in Sweden, perhaps because of increasing ley farming and subsidies introduced in the early 1990s. This can reveal the strong impact of rather local socio-economic policies on soil carbon storage, which can be captured by the QR meta-model at the resolution of the NUTS2 regions characterized by region- and country-specific characteristics. The policy-driven context needs to be considered in the models' design and applications. The slowing down of SOC decumulation in Sweden and Finland persists as time goes on, as it is shown in the panels of Figure 1. Figure 2 visualizes the percentage difference between the 50th quantile and the mean value of the SOC content change for NUTS2 regions from 1980 to 2020. Figure 2 shows that the mean value of the SOC content change, as estimated by the LR model, can differ from the most likely one, i.e., the 50th quantile. The brownish colors in Figure 2 correspond to the locations (NUTS2 regions), where the mean value is lower than the 50th quantile and the greenish colors correspond to where it is higher. Thus, the brownish colors identify the NUTS2 with underestimated and the greenish with overestimated SOC changes by the traditional LR (using symmetrical or least square goodness-of-fit criteria) models as they cannot properly address the non-normality and the variability of the covariates.

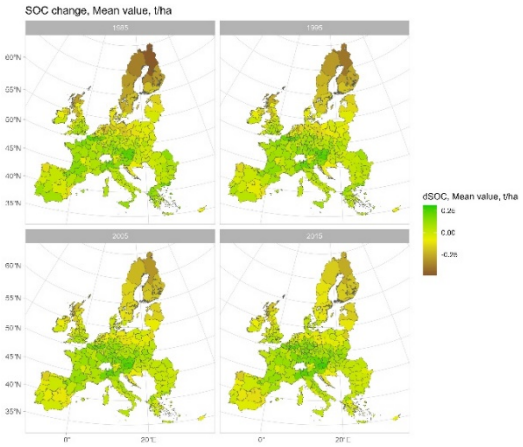

Figure 3.1. *Mean value of the SOC content change for 1980–2000*

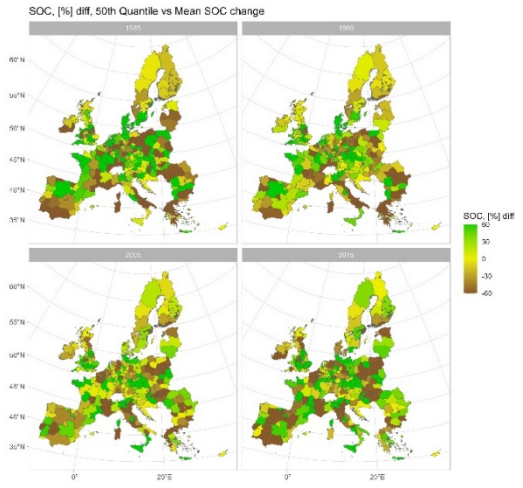

Figure 3.2. *Percentage difference between the 50th quantile and the mean value of the SOC content change for NUTS2 regions from 1980 to 2020*

The discrepancies between the 50th percentile and the mean value of the SOC content changes indicate that the interannual changes in the SOC content are non-

normally distributed. The non-normality can be explained by the variability of the monthly precipitation and temperature patterns affecting components of SOC differently for different soil characteristics and land management practicies. SOC meta-models have been estimated at the NUTS2 level, and, therefore, the discrepancies between the LR and the quantile estimates point to heterogeneities across SimUs within respective NUTS2 regions.

The 50th quantile of the SOC content changes identifies the dominating response of the SOC labile fraction to the interannual variability of climatic indicators including the response to possible extreme weather conditions as well as the response to prevailing land practices. The effects of precipitation on different SOC fractions can be opposite at wet and dry sites. Both the soil DOC (Dissolved Organic Carbon) and MBC (Microbial Organic Carbon) concentrations can decrease at the wet sites but increase at the dry sites under increased precipitation conditions.

SOC accumulation is also influenced by interannual N response to changing climatic conditions in different soils under alternative land use practices. This determines the C:N ratio and, therefore, can significantly influence DOC degradability and leaching and, thus, affect SOC content. The combined effects of precipitation and temperature patterns and their variability on SOC content changes indicate the differing response mechanisms in different soils under alternative land use practices, which can be addressed by the quantile-based SOC meta-models.

Figures 3 and 4 show the 75th and the 25th quantiles of the SOC content changes, thus estimating the ranges and the respective probabilities of how slow and how fast the SOC can change under varying exogenous drivers and local economic and policy conditions. Figure 3, displaying the 75th quantile value, tells that the SOC changes can be “better” than the 75th quantile value exhibited in the figure, however, only with a probability of 0.25. Correspondingly, the 25th quantile value in Figure 4 tells that the SOC changes with the probability of 0.25 can drop below the 25th quantile value exhibited in Figure 4, i.e., below 0.5 t/ha.

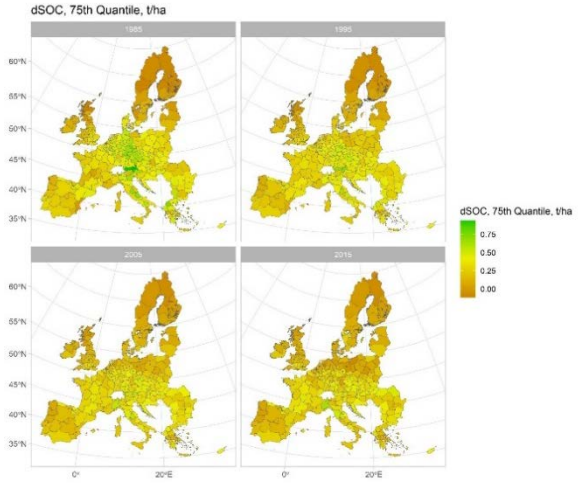

Figure 3.3. *The 75th quantile of the SOC content changes between the consequent years for NUTS2 regions from 1980 to 2020*

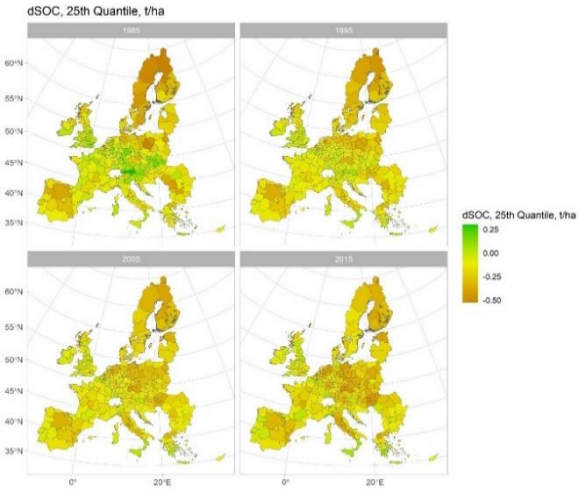

Figure 3.4. *The 25th quantile of the SOC content changes between the consequent years for NUTS2 regions from 1980 to 2020*

2.4. Conclusions

The paper discusses the connections between the two-stage stochastic optimization and quantile based robust statistical and machine learning models. In general problems of decision-making, feasible solutions, concepts of optimality and robustness are characterized from the context of decision-making situations, i.e., systems structure, goals, constraints, safety norms, supply-demand relationships, thresholds. Main question related to statistical estimation and predictions is how to use these predictions to optimize the overall decisions and how current decisions can affect predictions?

In the paper we discuss quantile-based machine learning regression meta-model for tracking dynamics and uncertainties of Soil Organic Carbon (SOC) in agricultural soils as a function of exogenous parameters such as monthly temperature and precipitation and endogenous, decision-dependent parameters, which can be altered by land use decisions derived with GLOBIOM model. The SOC meta-model is developed using multisource data from historical observations and results of a bio-physical model EPIC. Thus, it emulates the EPIC model and can be explicitly linked with GLOBIOM providing an effective means to analyse responces of environmental indicators to land management decisions computed by GLOBIOM

Incorporated as environmental targets into GLOBIOM, the quantiles of SOC meta-models enable the analysis of robust land management practices and the respective costs to increase SOC content to targeted levels.

Acknowledgements

The development of the methodologies and models is supported by 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”, by a joint project between the International Institute for Applied Systems Analysis (IIASA) and the 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 support from the National Research Foundation of Ukraine, grant No. 2020.02/0121. This research has been funded by the European Union’s H2020 Projects ENGAGE (Grant Agreement No. 821471) and COACCH (Proposal ID 776479), European Union's Horizon Europe research and innovation action under grant agreement No. 101086179 (AI4SoilHealth).

References

Balkovič, J., van der Velde, M., Skalský, R., Xiong, W., Folberth, C., Khabarov, N., Smirnov, A., Mueller, N.D., Obersteiner, M. (2014). Global wheat production potentials and management flexibility under the representative concentration pathways. *Global and Planetary Change*, 122, 107-121, https://doi.org/10.1016/j.gloplacha.2014.08.010.

Borodina, O.M., Kyryziuk, S.V., Fraier, O V., Ermoliev, Y.M., Ermolieva, T.Y., Knopov, P.S., Horbachuk, V.M. (2020). Mathematical Modeling of Agricultural Crop Diversification in Ukraine: Scientific Approaches and Empirical Results. *Cybernetics and Systems Analysis*, 56(2), 213-222, 10.1007/s10559-020-00237-6.

Ermoliev, Y., Wets, RJ-B. (1988). *Numerical techniques for stochastic optimization.* Springer Verlag, Heidelberg, Germany.

Ermoliev,Y., Norkin,V. (1997). On nonsmooth and discontinuous problems of stochastic systems optimization. *Europ. J. Oper. Res.*, 101(2), 230-243.

Ermoliev, Y., Hordijk, L. (2003). Global changes: Facets of robust decisions. In *Coping with uncertainty: Modeling and policy issue*, Marti, K., Ermoliev, Y., Makowski, M., Pflug, G. (eds.). Springer Verlag, Berlin, Germany.

Ermoliev, Y. (1976). *Methods of Stochastic Programming*. Nauka, Moscow.

Ermoliev, Y. (2009a). Two-stage stochastic programming: Quasigradient method. In Pardalos, P.M. (ed.), *Encyclopedia of optimization*. Springer Verlag, New York, USA, 3955–3959.

Ermoliev Y. (2009b). Stochastic quasigradient methods in minimax problems. In *Encyclopedia of optimization*, Pardalos, P.M. (ed.). Springer Verlag, New York, USA, 3813-3818.

Ermoliev, Y., Zagorodny, A.G., Bogdanov, V. L., Ermolieva, T., Havlik, P., Rovenskaya, E., Komendantova, N., Obersteiner, M. (2022). Robust Food–Energy–Water–Environmental Security Management: Stochastic Quasigradient Procedure for Linkage of Distributed Optimization Models under Asymmetric Information and Uncertainty. *Cybernetics and Systems Analysis*, 58(1), 45-57, 10.1007/s10559-022-00434-5.

Ermoliev, Y., von Winterfeldt, D. (2012). Systemic risk and security management. In *Managing safety of heterogeneous systems: Lecture notes in economics and mathematical systems*, Springer Verlag, Berlin, Heidelberg, Germany, 19-49.

Ermolieva, T., Havlik, P., Frank, S., Kahil, T., Balkovič, J., Skalský, R., Ermoliev, Y., Knopov, P.S., et al. (2022). A Risk-Informed Decision-Making Framework for Climate Change Adaptation through Robust Land Use and Irrigation Planning. *Sustainability*, 14(3), 1430, 10.3390/su14031430.

Ermolieva, T., Havlík, P., Ermoliev, Y., Mosnier, A., Obersteiner, M., Leclere, D., Khabarov, N., Valin, H., Reuter, W. (2016). Integrated management of land use systems under systemic risks and security targets: A Stochastic Global Biosphere Management Model. *Journal of Agricultural Economics*, 67(3), 584-601.

Ermolieva, T., Ermoliev, Y., Obersteiner, M., Rovenskaya, E. (2021). Two-Stage Nonsmooth Stochastic Optimization and Iterative Stochastic Quasigradient Procedure for Robust Estimation, Machine Learning and Decision Making. In *Resilience in the Digital Age.* Springer. ISBN 978-3-030-70369-1 10.1007/978-3-030-70370-7_4, 45-74.

Ermolieva, T., Havlik, P., Derci Augustynczik, A.L., Boere, E., Frank, S., Kahil, T., Wang, G., Balkovič, J., et al. (2023). A Novel Robust Meta-Model Framework for Predicting Crop Yield Probability Distributions Using Multisource Data. *Cybernetics and Systems Analysis*, doi: 10.1007/s10559-023-00620-z.

Ermolieva, T., Ermoliev, Y., Havlik, P., Derci Augustynczik, A.L., Komendantova, N., Kahil, T., Balkovič, J., Skalský, R., et al. (2023). Connections between robust statistical estimation,

robust decision making with two-stage stochastic optimization, and robust machine learning problems. *Cybernetics and systems analysis*, 59 (3), 33-47.

Ermoliev, Z., Komendantova, N., Ermolieva, T. (2023). Energy Production and Storage Investments and Operation Planning Involving Variable Renewable Energy Sources A Two-stage Stochastic Optimization Model with Rolling Time Horizon and Random Stopping Time. In *Modern Optimization Methods for Decision Making Under Risk and Uncertainty*, Gaivoronski, A., Knopov, P., & Zaslavskyi, V. (eds.), Taylor & Francis, pp. 15-50, ISBN 9781003260196, doi: 10.1201/9781003260196-2.

Ermolieva, T., Havlik, P., Derci Augustynczik, A.L., Frank, S., Balkovič, J., Skalský, R., Deppermann, A., Nakhavali, A., et al. (2024). Tracking the Dynamics and Uncertainties of Soil Organic Carbon in Agricultural Soils Based on a Novel Robust Meta-Model Framework Using Multisource Data. *Sustainability*, 16(16), e6849, doi: 10.3390/su16166849.

Gao, J., Xu, X., Cao, G.-Y., Ermoliev, Y., Ermolieva, T., & Rovenskaya, E. (2021). Strategic decision-support modeling for robust management of the food–energy–water nexus under uncertainty. *Journal of Cleaner Production*, 292 e125995, 10.1016/j.jclepro.2021.125995.

Gorbachuk, V.M., Ermoliev, Y., Ermolieva, T., Dunajevskij, M.S., 2019. Quantile-based regression for the assessment of economic and ecological risks. In *Proceedings of the 5th International scienti_c conference on Computational Intelligence*, 15-20 April, 2019. pp. 188-190 Uzgorod, Ukraine: Ministry of Education and Science of Ukraine.

Huber, P. (1981). *Robust statistics*. John Wiley & Sons, New York, Toronto, Singapore.

Knopov, P. (2002). *Empirical estimates in stochastic optimization and identification*. Springer Verlag, Berlin, Germany.

Ortiz-Partida JP, Kahil T , Ermolieva T, Ermoliev Y, Lane B, Sandoval-Solis S, Wada, Y., 2019. A Two-Stage stochastic optimization for robust operation of multipurpose reservoirs. *Water Resources Management*, 33(11), 3815-3830, doi:10.1007/s11269-019-02337-1.

Ren, M., Xu, X., Ermolieva, T., Cao, G.-Y., Yermoliev, Y. (2018). The optimal technological development path to reduce pollution and restructure iron and steel industry for sustainable transition. *International Journal of Science and Engineering Investigations*, 7(73), 100-105.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag. ISBN: 0-387-98-780-0.