International Institute for
Applied Systems Analysis
Schlossplatz 1
A-2361 Laxenburg, Austria

Tel: +43 2236 807 342
Fax: +43 2236 71313
E-mail: publications@iiasa.ac.at
Web: www.iiasa.ac.at

Interim Report IR-06-002

Sequential Downscaling Methods for Estimation from Aggregate Data

G. Fischer, T. Ermolieva, Y. Ermoliev, and H. Van Velthuizen

Approved by

Leen Hordijk (hordijk@iiasa.ac.at)
Director, IIASA

January 2006

Abstract

Global change processes raise new estimation problems challenging the conventional statistical methods. These methods are based on the ability to obtain observations from unknown true probability distributions, whereas the new problems require recovering information from only partially observable or even unobservable variables. For instance, aggregate data exist at global and national level regarding agricultural production, occurrence of natural disasters, on incomes, etc. without providing any clue as to possibly alarming diversity of conditions at local level. "Downscaling" methods in this case should achieve plausible estimation of local implications emerging from global tendencies by using all available evidences.

The aim of this paper is to develop a sequential downscaling method, which can be used in a variety of practical situations. Our main motivation for this was the estimation of spatially distributed crop production, i.e., on a regular grid, consistent with known national-level statistics and in accordance with geographical datasets and agronomic knowledge. We prove convergence of the method to a generalized cross-entropy maximizing solution. We also show that for specific cases this method is reduced to known procedures for estimating transportation flows and doubly stochastic matrices.

Keywords: Cross-entropy, minimax likelihood, downscaling, spatial estimation.

Acknowledgments

The authors are grateful to the participants of the IFIP/IIASA/GAMM workshop on Coping with Uncertainty, December 13-16, 2004, IIASA, and to the IIASA colleagues for critical suggestions that led to improvements of this paper.

Contents

Sequential Downscaling Methods for Estimation from Aggregate Data

G. Fischer, T. Ermolieva, Y. Ermoliev, and H. Van Velthuizen

1 Introduction

The analysis of global change processes requires the development of methods, which allow for dealing in a consistent manner with data on a multitude of spatial and temporal scales. Although GIS provides detailed geo- physical information, the socio-economic data often exist only at aggregate level. Integrated analysis of economic and environmental impacts of global changes raises a number of new estimation problems for downscaling and upscaling of available data to ensure consistency of biophysical and economic models. For example, aggregate data on national income does not reveal possibly alarming heterogeneity of its concentration among a small fraction of population or within, say, risk-prone regions of a country. We often need to derive information about the occurrence of disasters and induced potential losses in particular locations from information of their occurrence at global or regional levels. Aggregate regional annual concentrations of pollutants may be well within norms, whereas concentrations in some locations may exceed vital levels for a short time and cause irreversible damages.

The estimation of global processes consistent with local data and, conversely, long-term local implications emerging from global tendencies challenge the traditional statistical estimation methods. These methods are based on the ability to obtain observations from unknown true probability distributions. For the new estimation problems, which can be termed downscaling and upscaling estimation problems (see also [2] discussing other downscaling and upscaling problems), we often have only very restricted samples of real observations. Additional experiments to obtain more observations may be expensive, time consuming, dangerous or simply impossible. For example, although we can estimate total "departures" or "arrivals" of passengers in transportation systems, the estimation of passenger flows between different locations requires expensive origin-destination surveys and in many cases the data does not exist [5]. Similar situations occur with projections of migration flows, estimation of flows in communication systems, and trade flows. The paucity or lack of historical data is especially limiting for regions, which are subject to rapid changes (new developments, shocks, instabilities).

The aim of this article is to develop a recursive sequential downscaling method, which can be used for a large variety of practical situations. Our main motivation for this is the spatially explicit estimation of agricultural production, which is outlined in Section 2.1 and Section 5. In this problem we deal with "downscaling", i.e. attribution of known aggregate national or sub-national crop production and land use to particular locations (grid cell; pixel). Sections 2.2, 2.3 outline also the main idea of the sequential downscaling method of Section 3 by using simple known procedures for estimating transportation flows (e.g., migration flows, combining purely probabilistic prediction with available data on total demands and capacities of locations) and transition probabilities.

Section 3 develops a sequential downscaling method for iterative rebalancing estimates to satisfy general balance equations connecting unobservable and observable variables. We prove the convergence of this method to the solution maximizing a cross-entropy function. For specific transportation constraints this method reduces to the procedure proposed in the 1930s by the Leningrad architect G.V. Sheleikhovskii for estimating passenger flows. The convergence of Sheleikhovskiis method to the solution maximizing a cross-entropy function was established in [1] by complex and lengthy analysis of specific mappings arising in the case of the transportation constraints. Our analysis for general constraints is based on duality theory, which significantly simplifies proofs and clarifies the convergence properties. This opens up a way for various modifications and extensions, e.g., to situations with uncertainties when the available higher- level information is imprecise or involves stochastic elements.

Section 4 outlines connections between the maximum entropy principle, widely used (see e.g., [3], [11]) for the new estimation (downscaling) problems and the fundamental maximum likelihood principle of statistical estimation theory. We show that the maximum entropy principle can be viewed as an extension of the maximum likelihood principle, the so-called minimax log likelihood principle. Therefore, the convergence of downscaling methods to solutions maximizing a cross-entropy function can be considered as an analog of the asymptotic consistency [14] analysis in traditional statistical estimation theory.

Section 5 describes a practical application and results of numerical calculations, with a fast convergence of the proposed basic procedure and its possible modifications. Section 6 concludes. As an important topic for future research, it emphasizes the need for incorporating the downscaling methods within the overall decision making problems, i.e., similar to the existing stochastic optimization theory.

2 Downscaling Problems: Motivating Examples

Let us consider situations, very common in regional studies, when direct observations of uncertain parameters on local levels are practically impossible and the estimation of their spatially explicit representation requires a downscaling procedure making use of information at a higher, more aggregate level. The problem of Section 2.1, in fact, motivated the development of discussed in Section 3 sequential downscaling procedure. Sections 2.2, 2.3 outline the main idea of this procedure by using simpler special cases of the problem.

2.1 Spatial Estimation of Agricultural Production

In general, the available information can be summarized as follows (see also Section 5). Extent of arable land a_i, in a pixel i, $i = \overline{1,m}$, is estimated from land cover satellite images. The degree and extents of suitable area for different crops in a pixel comes from FAO/IIASA crop suitability studies [9], [10]. There is also (computer- simulated) spatial information on the attainable yield d_{ij} of crop j, $j = \overline{1,n}$ in pixel i. From statistics, the price p_j of crop j, the value V_j of crop production j in a country, i.e. the total production of crop j multiplied by its price, the crop-wise sown area and production are available. Let x_{ij} be desirable estimates of crop j production in pixel i. This leads to the following estimate $v_{ij} = p_j d_{ij} x_{ij}$ of crop production value j in pixel i. Since production value V_j of crop j in the country is known from statistics, $\sum_{i=1}^{m} v_{ij} = V_j$, we have equations

$$\sum_{j=1}^{n} x_{ij} = a_i, i = \overline{1,m}. \tag{1}$$

$$\sum_{i=1}^{m} d_{ij} x_{ij} = b_j, j = \overline{1, n}, \tag{2}$$

where $b_j = V_j / p_j$.

By introducing new variable y_{ij} characterizing area shares by crop j in pixel i, i.e., $x_{ij} = a_i y_{ij}$, constraints (1), (2) can be written as the following

$$\sum_{j=1}^{n} y_{ij} = 1, i = \overline{1, m}, \tag{3}$$

$$\sum_{i=1}^{m} a_{ij} y_{ij} = e_j, j = \overline{1, n}, \tag{4}$$

where $a_{ij} = d_{ij} a_i$. This modification of constraint (1), (2) allows the use of entropy-like arguments.

There will usually be an infinite number of feasible solutions x_{ij}, $i = \overline{1, m}$, $j = \overline{1, n}$ satisfying equations (3) and (4). Therefore, to find a unique solution requires application of some additional principles. A key idea is to use some additional prior information on crop-specific area distribution, i.e., a prior distribution q_{ij} of crop j in pixel i. This prior can be based upon available crop distribution maps and other ancillary information, such as agro-climatic, biophysical, terrain and soil, demographic and farming systems characteristics (see discussion in Section 5). In any case, regardless of availability and detail of ancillary information, the prior can even be a (least informative) uniform distribution [16]. If a prior distribution $q_{ij} > 0$, $i = \overline{1, m}$, $j = \overline{1, n}$ is available, then a rather natural way to derive the estimates is from the minimization of the function

$$\sum_{j=1}^{n} \sum_{i=1}^{m} y_{ij} ln \frac{y_{ij}}{q_{ij}}, \tag{5}$$

subject to (3), (4), where (5) defines the so-called Kullback-Leibler distance [12] between distributions y_{ij} and q_{ij}.

Remark 1. Function $-\sum_{i,j} y_{ij} ln \frac{y_{ij}}{q_{ij}}$ is termed the cross-entropy, i.e., the minimization of (5) defines the cross-entropy maximizing estimates. Since $\sum_{i,j} x_{ij} ln \frac{x_{ij}}{q_{ij}} = \sum_{i,j} a_i y_{ij} ln \frac{y_{ij}}{q_{ij}} + \sum_i a_i ln a_i$, therefore the minimization of function $\sum_{i,j} x_{ij} ln \frac{x_{ij}}{q_{ij}}$ subject to equations (1), (2) is equivalent to minimization of a generalized (a weighted) cross-entropy $\sum_{i,j} a_i y_{ij} ln \frac{y_{ij}}{q_{ij}}$.

An alternative approach, which we take in this paper, is to derive a sequence of estimates y_{ij}^0, y_{ij}^1, y_{ij}^2, ... from an appropriate behavioral principle and to prove their convergence to a cross-entropy maximizing solution. For instance, a general tendency in farming is to allocate a crop j to pixels with maximum production values $p_j d_{ij}$ (or similar, such as maximum net revenue or maximum net present value in case of perennial crops or forestry activities). However, the straightforward application of such a rule to equations (1), (2) will, in general, lead to an overestimation or underestimation of aggregate known production values V_j, $j = \overline{1, n}$, i.e., situations when condition (2) is not fulfilled. Thus, these rule-based initial estimates require a sequential balancing procedure, which is developed in Section 3. Let us illustrate the main idea of the procedure by using two important special cases.

2.2 Estimation of Interzonal Flows

There can be different types of flows requiring estimation or/and projection procedures. It may be immigration or trade flows between different regions, flows of passengers or goods in transportation systems, or flows of messages in communication systems. Purely statistical projections often require expensive and time consuming origin-destination surveys; the necessary historical information may not exist [5]. In particular, this is a key issue in situations when land use patterns are changing, e.g., due to new development or "shocks" in some locations. In addition, standard statistical procedures often do not take into account such available information as "demands" for departures from locations i, $i = \overline{1,m}$, and "capacities" of locations j, $j = \overline{1,n}$, to accommodate inflows. As a result, they may overestimate or underestimate the actual movements between locations.

The downscaling methods attempt to estimate flows among given locations in a way consistent with available data on the expected total number of "departures" a_i from locations i and arrivals b_j in location j. For unknown (to be estimated) flows x_{ij} clearly $\sum_{j=1}^{n} x_{ij} = a_i$, $i = \overline{1,m}$, $\sum_{i=1}^{m} x_{ij} = b_j$, $j = \overline{1,n}$, i.e., we have a particular case of constraints (1), (2) with $d_{ij} = 1$, $i = \overline{1,m}$, $j = \overline{1,n}$. Assume also that there is a prior probability q_{ij} for a passenger from i to choose the destination j. For example, some behavioral models (see, e.g., [7], p. 414) define q_{ij} proportionally to a "distance" r_{ij} from i to j, $q_{ij} = r_{ij} / \sum_j r_{ij}$.

Consider the following iterative estimation procedure:
(i). If a passenger from location i chooses the destination j with a prior probability q_{ij}, $\sum_j q_{ij} = 1$, then the expected flow from i to j is $x_{ij}^0 = a_i q_{ij}$. Clearly $\sum_j x_{ij}^0 = a_i$, $i = \overline{1,m}$, but there may be overestimation $\sum_i x_{ij}^0 > b_j$ or underestimation $\sum_i x_{ij}^0 < b_j$ of the available b_j.
(ii). Calculate relative imbalances $\beta_j^0 = b_j / \sum_i x_{ij}^0$ and $z_{ij}^0 = x_{ij}^0 \beta_j^0$, $i = \overline{1,m}$, $j = \overline{1,n}$.
(iii). Clearly, $\sum_i z_{ij}^0 = b_j$, $j = \overline{1,n}$, but the estimate z_{ij}^0 may overestimate or underestimate the known demand for departures a_i from i. Therefore, calculate $\alpha_i^0 = a_i / \sum_j z_{ij}^0$, $x_{ij}^1 = z_{ij}^0 \alpha_i^0$, and so on.

This balancing procedure can be summarized also as the following. We can represent x_{ij}^1 as $x_{ij}^1 = a_i q_{ij}^1$, and $q_{ij}^1 = (q_{ij}\beta_j^0)/(\sum_j q_{ij}\beta_j^0)$, $i = \overline{1,m}$, $j = \overline{1,n}$. Assume $x^k = \left\{ x_{ij}^k \right\}$ has been calculated. Then find $\beta_j^k = b_j / \sum_i x_{ij}^k$ and calculate $x_{ij}^{k+1} = a_i q_{ij}^{k+1}$, $q_{ij}^{k+1} = (q_{ij}\beta_j^k / \sum_j q_{ij}\beta_j^k)$, $i = \overline{1,m}$, $j = \overline{1,n}$, and so on. In this form the procedure can be viewed as a sequential redistribution of demands a_i from locations $i = \overline{1,m}$ among locations $j = \overline{1,n}$ by using a Bayesian type of rule for updating the prior distribution q_{ij}: $q_{ij}^{k+1} = q_{ij}\beta_j^k / \sum_j q_{ij}\beta_j^k$, $q_{ij}^0 = q_{ij}$.

Initially this method was proposed by the architect Sheleikhovskii for estimating passenger flows between districts of a city (including possible new districts). Proof of convergence to the solution maximizing $\sum_{ij} x_{ij} ln(x_{ij}/q_{ij})$ was given in [1] using extremely lengthy and complex arguments essentially relying on specific mappings associated with the transportation constraints. In Section 3 we propose a similar method for general constraints (2). We apply duality theory, which allows us to significantly simplify and clarify the analysis (Proposition 1). This opens up an opportunity for various modifications, in particular, to situations with uncertain parameters a_i, b_j, and d_{ij}.

2.3 Estimation of Stochastic Matrices

It is interesting to note that a similar procedure is used in the conventional statistical theory for estimating doubly stochastic matrices (see discussion in [13], [18]). Suppose we can observe transitions of a Markov chain with n states and stochastic matrix $\{P_{ij}\}$.

The usual estimate of P_{ij} is $x_{ij} = \alpha_{ij}/a_i$ where α_{ij} is the number of transitions from i to j, which are observed, and $a_i = \sum_j \alpha_{ij}$. This amounts to a normalization of the rows of matrix $\{\alpha_{ij}\}$. If it was known that $\{P_{ij}\}$ is in fact a doubly stochastic matrix, i.e., $\sum_i P_{ij} = 1$, then it was proposed to alternately normalize (as in Section 2.2) the rows and columns of $\{\alpha_{ij}\}$ in the belief that this iterative process would converge to an estimate of $\{P_{ij}\}$. Proof of convergence of this procedure to a doubly stochastic matrix for rather special cases was given in [13]. From the results in [1] follows the convergence for general doubly stochastic matrixes and the optimality of the resulting estimates as the cross-entropy maximizing solution.

3 Sequential Downscaling Methods

Consider the following problem: minimize

$$\sum_{j=1}^{n}\sum_{i=1}^{m} x_{ij} ln(x_{ij}/q_{ij}), \tag{6}$$

subject to constraints (1), (2), where $q_{ij} > 0$, $d_{ij} > 0$ are given, $a_i > 0$, $b_j > 0$, $i = \overline{1,m}$, $j = \overline{1,n}$. Values $x_{ij} = 0$ are also possible when $q_{ij} = 0$ or $d_{ij} = 0$. Without loss of generality, we assume $x_{ij} > 0$, $q_{ij} > 0$, $\sum_{j=1}^{n} q_{ij} = 1$, $i = \overline{1,m}$, and the set of feasible solutions defined by (1), (2) is not empty.

Consider the following sequential procedure.

Step 1: Compute an initial estimate $x_{ij}^0 = a_i q_{ij}$. Clearly, x_{ij}^0 satisfies (1), $\sum_j x_{ij}^0 = a_i$, since $\sum_j q_{ij} = 1$ but, in general, constraints (2) are violated.

Step 2: For given $x^k = x_{ij}^k$, find β_j^{k+1} satisfying equations

$$\sum_{i=1}^{m} d_{ij} x_{ij}^k e^{d_{ij}\beta_j} = b_j, j = \overline{1,n}. \tag{7}$$

The left hand side of this equality is a strictly monotonic function and β_j^{k+1} can be easily calculated.

Step 3: Calculate $z_{ij}^{k+1} = x_{ij}^k e^{d_{ij}\beta_j^{k+1}}$, and

$$\alpha_i^{k+1} = a_i / \sum_j z_{ij}^{k+1}, i = \overline{1,m}, j = \overline{1,n}. \tag{8}$$

Step 4: Update x_{ij}^k to

$$x_{ij}^{k+1} = \alpha_i^{k+1} z_{ij}^{k+1}, i = \overline{1,m}, j = \overline{1,n}. \tag{9}$$

and so on with Steps 2 - 4, until desirable convergence is reached, e.g., constraints (1), (2) are satisfied with a given accuracy.

In summary, this procedure, similar to Sections 2.2, 2.3 involves a sequential updating of a priori probability distribution q_{ij} by using a Bayesian type of rule: $x_{ij}^{k+1} = a_i q_{ij}^{k+1}$, $q_{ij}^{k+1} = q_{ij}\gamma_j^k / \sum_j q_{ij}\gamma_j^k$, $\gamma_j^k = e^{d_{ij}\beta_j^k}$, where values γ_j^k are calculated using observations of imbalances rather than using observations of real random variables.

Proposition 1. The sequence $x^k = \left\{x_{ij}^k, i = \overline{1,m}, j = \overline{1,n}\right\}$, $k = 0, 1, ...,$ generated by iteration (7)-(9) converges to the solution x^* of constraints (1), (2) minimizing the function (6).

Lemma. There exist such $\alpha_i > 0$, β_j, $i = \overline{1,m}$, $j = \overline{1,n}$, that the optimal solution x^*_{ij} minimizing (6) subject to constraints (1), (2) satisfies the following optimality conditions: $x^*_{ij} = x_{ij}(\alpha, \beta)$,

$$\begin{array}{l} a_i - \sum_j x_{ij}(\alpha,\beta) = 0; i = \overline{1,m}; \\ b_j - \sum_i d_{ij} x_{ij}(\alpha,\beta) = 0; j = \overline{1,n}; \\ x_{ij}(\alpha,\beta) = q_{ij}\alpha_i e^{d_{ij}\beta_j}, i = \overline{1,m}, j = \overline{1,n}. \end{array}$$

Proof. For a continuous, strictly convex function (6) on a non-empty compact set of an Euclidian space there is a unique optimal solution to the minimization problem. Consider the Lagrangian function:

$$L(x,\lambda,\mu) = \sum_{i,j} x_{ij} ln(x_{ij}/q_{ij}) + \sum_{i=1}^{m} \lambda_i (a_i - \sum_{j=1}^{n} x_{ij}) + \sum_{j=1}^{n} \mu_j (b_j - \sum_{i=1}^{m} d_{ij} x_{ij})$$

Since the optimal solution is positive, the optimality conditions lead to

$$\frac{\partial L}{\partial x_{ij}} = ln\frac{x_{ij}}{q_{ij}} + 1 - \lambda_i - d_{ij}\mu_j = 0,$$

$i = \overline{1,m}$, $j = \overline{1,n}$, i.e., the optimal solution can be represented analytically as $x_{ij}(\lambda,\mu) = q_{ij} e^{\lambda_i - 1} e^{d_{ij}\mu_j}$, $i = \overline{1,m}$, $j = \overline{1,n}$. The dual problem reads: find Lagrange multipliers (λ_i, μ_j), $i = \overline{1,m}$, $j = \overline{1,n}$, maximizing function

$$\varphi(\lambda,\mu) = min_x L(x,\lambda,\mu) = L(x(\lambda,\mu),\lambda,\mu).$$

From basic results of convex analysis it follows that $\varphi(\lambda,\mu)$ is a strictly concave continuously differentiable function and the optimality condition can be written as

$$\frac{\partial \varphi}{d\lambda_i} = a_i - \sum_{j=1}^{n} x_{ij}(\lambda,\mu) = 0, i = \overline{1,m},$$

$$\frac{\partial \varphi}{d\mu_j} = b_j - \sum_{i=1}^{m} d_{ij} x_{ij}(\lambda,\mu) = 0, j = \overline{1,n}.$$

By using new notations $\alpha_i = e^{(\lambda_i - 1)}$, $\beta_j = \mu_j$, and the same notations $x_{ij}(\alpha,\beta)$ for corresponding $x_{ij}(\lambda,\mu)$, $\lambda_i = \lambda_i(\alpha_i) = ln\alpha_i + 1$, $\mu_j = \beta_j$ we obtain the proof due to the strict monotonicity of $e^{(\lambda_i - 1)}$.

Proof of Proposition 1. It is easy to see that the sequential method (7)- (9) updates variables $\alpha = (\alpha_1, ..., \alpha_m)$, $\beta = (\beta_1, ..., \beta_n)$ and $x = \{x_{ij}\}$ to satisfy the optimality conditions of **Lemma**. Namely, equations (7) require that the gradient of the strictly concave function of the dual problem $\varphi_\mu(\lambda^k, \mu^{k+1}) = 0$, whereas equations (8) require that the gradient $\varphi_\lambda(\lambda^{k+1}, \mu^{k+1}) = 0$, for some λ^k, μ^k, $k = 0, 1, ...$

Indeed, let us illustrate just a few steps of the method. Solution x^0_{ij} can be represented as $x^0_{ij} = \alpha^0_i q_{ij} e^{d_{ij}\beta^0_j}$, $\alpha^0_i = a_i$, $\beta^0_j = 0$. Clearly, that $\sum_j x^0_{ij} = a_{ij}$, i.e., $\varphi_{\lambda_i}(\lambda^0, \mu^0) = 0$, $\lambda^0_i = \lambda_i(\tilde{\alpha}^0_i)$, $\tilde{\alpha}^0_i = \alpha^0_i$, $\mu^0_j = \beta^0_j$. At **Step 2** values β^1_j modify x^0_{ij} to $y^1_{ij} = \alpha^0_i q_{ij} e^{d_{ij}(\beta^0_j + \beta^1_j)}$, $\sum_i d_{ij} y^1_{ij} = b_j$, i.e., $\varphi_{\mu_j}(\lambda^0, \mu^1) = 0$, $\mu^1_j = \beta^0_j + \beta^1_j$. At **Step 3** values α^1_i modify y^1_{ij} to $x^1_{ij} = \alpha^0_i \alpha^1_i q_{ij} e^{d_{ij}(\beta^0_j + \beta^1_j)}$, $\sum_j x^1_{ij} = a_i$, i.e., $\varphi_{\lambda_i}(\lambda^1, \mu^1) = 0$, $\lambda^1_i = \lambda_i(\tilde{\alpha}^1_i)$, $\tilde{\alpha}^1_i = \alpha^0_i \alpha^1_i$ and so on.

Therefore, the convergence of vectors λ^k, μ^k and $\{x^k_{ij}\}$ to the optimal solutions of the dual and the primal problems follows from the convergence of the cyclic ascent method [17].

Remark 2. It follows from the above that for transportation constraints, i.e., for $d_{ij} = 1$, $i = \overline{1,m}$, $j = \overline{1,n}$, the proposed method is reduced to Sheleikovskii's method. In this case, it also follows that the optimal solution is represented as $x_{ij}(\alpha, \beta) = q_{ij}\alpha_i\beta_j$, $\alpha_i > 0$, $\beta_j > 0$, $i = \overline{1,m}$, $j = \overline{1,n}$, what is typical for the so-called gravity models [4].

4 Minimax Likelihood and Maximum Entropy

Definitely that besides a cross-entropy maximization there exists a vast variety of optimization principles to single out a solution of equations (3), (4). Let us show that minimization of (5) is a natural generalization of the fundamental maximum likelihood principle of statistical theory.

The standard statistical estimation theory deals with the situation when the information on unknown distribution can be derived from observations of underlying random variables. In such a case, the most natural principle for selecting an estimate from a given sample of observations is the maximum likelihood proposed by Fisher [8]. This principle requires that the estimate has to maximize the probability that a given sample is observed.

A downscaling problem deals with the estimation of often unobservable variables. Yet, the uncertainty can also be characterized or interpreted in probabilistic terms. For example, in the estimation of crop production values defined by equations (3), (4), we can think of values $y_{ij} > 0$, $\sum_{j=1}^{n} y_{ij} = 1$ as the probability (the degree of our belief) that a unit area of pixel i is allocated to crop j. It is easy to see that the maximum entropy principle can be viewed as an extension of the maximum likelihood principle.

Consider a situation similar to problems posed in Section 2. Namely, let us assume that there is an underlying random variable ξ with a finite number of possible values $\xi_1, ..., \xi_r$ and the unknown true probability distribution of ξ is concentrated at these points with associated probabilities $p_1^*, ..., p_r^*$, $Prob[\xi = \xi_j] = p_j^*$.

In the statistical estimation the available information is given by a random sample $\xi^1, ..., \xi^N$ of N independent observations of ξ on $(p_1^*, ..., p_r^*)$. A maximum likelihood estimate of the unknown probabilities $(p_1^*, ..., p_r^*)$ is obtained by maximizing the probability (likelihood) of observing $\xi^1, ..., \xi^N$

$$\prod_{k=1}^{N} Prob[\xi = \xi^k] = \prod_{j=1}^{r} p_j^{v_j} \tag{10}$$

subject to constraints $\sum_{j=1}^{r} p_j = 1$, $p_j > 0$, $j = \overline{1,r}$, where v_j is the number of times the value ξ_j has been observed, $\sum_{j=1}^{r} v_j = N$. Since lny is a monotonously increasing function of y, the maximization of (10) is equivalent to maximization of the log likelihood function $ln \prod_{j=1}^{r} p_j^{v_j} = \sum_{j=1}^{r} v_j lnp_j$ or normalized by the number of observations N, $\sum_{j=1}^{r} v_j = N$, the sample mean function

$$\frac{1}{N}\sum_{j=1}^{n} v_j lnp_j. \tag{11}$$

This is a continuous, strictly concave function on the set of R^n determined by linear constraints. By using the Lagrangian function (or the more general fact of Proposition 2 below) we can derive the well known result (see, e.g., [15]) that the unique solution maximizing (11) is the empirical probability function

$$p_j^N = v_j/N, j = \overline{1,r}. \tag{12}$$

Let us consider this differently. The log likelihood function (11) is the sample mean approximation of the expectation

$$Elnp_\xi = \sum_{j=1}^{r} p_j^* lnp_j, \tag{13}$$

where the unknown probability distribution p_j^* is approximated by the frequencies v_j/N derived from an available sample of observations $\xi^1, ..., \xi^N$. In downscaling problems the available information about the unknown probability distribution p_j^*, $j = \overline{1, r}$ is given not by a sample of observations, but by a number of constraints (3) and (4), i.e., $p^* \in P$, where P is the set of all feasible distributions. If $y = (y_1, ..., y_r) \in P$, then we can consider

$$\sum_{j=1}^{r} y_j lnp_j, \tag{14}$$

as an approximation of the expectation function (13) similar to the sample function (11). The log likelihood function (14) is defined for any feasible probability distribution $y \in P$. The worst-case estimate from P leads to minimization of the function

$$V(y) = \max_{p \in P} \sum_{j=1}^{r} y_j lnp_j. \tag{15}$$

w.r.t. $y \in P$. Therefore, the minimization of (15) w.r.t. $p \in P$ is a counterpart to the minimization of (11)

Proposition 2.

$$\min_{y \in P} \max_{p \in P} \sum_{j=1}^{r} y_j lnp_j = \min_{y \in P} \sum_{j=1}^{r} y_j lny_j. \tag{16}$$

Proof. It follows from analogous to (12) fact: if $y \in P$, then $V(y) = \sum_{j=1}^{r} y_j lny_j$. Indeed, for a given $y = (y_1, ..., y_r) \in P$ and $p \in P$ we have $\sum_{j=1}^{r} y_j lnp_j - \sum_{j=1}^{r} y_j lny_j = \sum_{j=1}^{r} y_j ln\frac{p_j}{y_j} < \sum_{j=1}^{r} p_j - \sum_{j=1}^{r} y_j = 0$ since $lnz < z - 1$ for $z > 0$.

Remark 3. In other words, the worst-case estimate leads to the principle of maximizing entropy $-\sum_{j=1}^{r} y_j lny_j$. In the case of a given prior distribution $q = (q_1, ..., q_r)$, we may require the minimization of the difference between the function (14) for $p \in P$ and $\sum_{j=1}^{r} q_j lnp_j$ for the given prior q from P:

$$\min_{y \in P}[\max_{p \in P} \sum_{j=1}^{r} y_j lnp_j - \sum_{j=1}^{r} y_j lnq_j] = \min_{y \in P} \sum y_j ln\frac{y_j}{q_j}, \tag{17}$$

i.e., the maximization of cross-entropy function $-\sum_{j=1}^{r} y_j ln\frac{y_j}{q_j}$ or the Kullback-Leibler distance between distributions y and q. Clearly, instead of selecting a worst-case distribution $y \in P$ in (15) we can take other distributions, which may lead to different downscaling principles. Since the estimation is usually used to support decision making processes, these more general principles may be specific to different types of problems, i.e., explicitly connected with the goals of a decision making problem.

5 Practical Applications

The proposed method has been applied for downscaling aggregate national and subnational data on crop production and land use (Section 2.1) for all main countries of the world. The downscaling was performed country-by-country. For this, the territory of each country was subdivided into grid cells with cultivation share, each cell with spatial resolution of 5 by 5 min latitude-longitude, i.e., urban areas, infrastructure, and water bodies were excluded from the analysis. To illustrate the dimensionality, the number of grid cells with cultivation in France equaled 9042, in Germany 6510, and in Austria 1165. For larger countries, such as United States and Russia, the number of grid cells with active agricultural use reached approximately 95 thousand, for Brazil 80 thousand and about 75 thousand for China. The data on aggregate country-specific agricultural production was obtained from FAO. The list of crops comprised 28 major crops such as wheat, rice, maize, potato, soybean, pulses, oil crops, coffee, tea, tobacco, cotton, etc. Figure 1 shows spatial distribution of downscaled total crop production value for Europe in terms of international prices (Geary-Khamis (GK) dollars of 2000 – 2001 per spatial land unit (grid cell).

Calculation of the prior included important spatial information on percentage of cultivated, rainfed and irrigated land in each grid cell derived using satellite images of land cover classes as well as aggregate statistics of arable land used for annual and perennial crops in each country. For example, Figure 2 shows cultivated land share by grid cell. In addition, the calculation of prior included information on multicropping index, i.e., how many harvests may be obtained per year from a piece of land, derived with AEZ methodology [9], [10], crop suitability (including climate, soil, and terrain conditions) and attainable yields in each spatial land unit, as well as information on characteristics of prevailing farming systems and population distribution.

Versions of the algorithm were written in FORTRAN and MATLAB programming languages and performed on PC. They showed fast convergence and, thus, efficient performance in dealing with large and spatially detailed data. Clearly, the time performance of the algorithm depends on the size of a study region, i.e., number of grid cells or land units considered and the number of crops that can be grown in each location. Applications of the algorithms to global studies showed that to attain high precision (10^{-7}) solution time increased roughly linearly with the increase of problem dimensionality. The performance is often remarkably fast, which is explained by the quality of the prior and the corresponding initial approximation. Thus, for Austria, 7 iterations were needed, for Germany and France about 20 to 30 and for China about 60, which indicates that the algorithm can be efficiently used in large real- world downscaling problems.

Remark 4. The proposed method can easily be modified to reflect problem-specific peculiarities of constraints (1) and (2). An important special case is the transportation constraints, i.e., $d_{ij} = 1$, $i = \overline{1,m}$, $j = \overline{1,n}$. If coefficients d_{ij} are reasonably well approximated by a product of some parameters δ_i, $i = \overline{1,m}$, σ_j, $j = \overline{1,n}$, for instance $d_{ij} = \delta_i \sigma_j$, $i = \overline{1,m}$, $j = \overline{1,n}$, then (1), (2) can be reduced to the transportation constraints by introducing new variables $y_{ij} = \delta_j x_{ij}$ and substituting b_j by b_j/σ_j and a_i by a_i/δ_i, i.e., simply by rescaling. Another simplifying situation occurs when function $e^{d_{ij}\beta_j}$ is approximated by a function $A_{ij} f_j^{\beta_j}$, $i = \overline{1,m}$, $j = \overline{1,n}$, for some parameters $A_{ij} > 0$, $f_j > 0$, $i = \overline{1,m}$, $j = \overline{1,n}$, and β_j varying within the range of plausible solutions of (7).

6 Concluding Remarks

In this paper we analyze numerical downscaling procedures only for situations when aggregate observed information is available and used as constraints on average values. For many practical situations this assumption may be insufficient and the procedures may need to be extended into more rigorous treatment of uncertainty regarding a prior probability q_{ij} and parameters of constraints (1), (2).

For practical applications, the choice of appropriate "priors", their inherent uncertain-

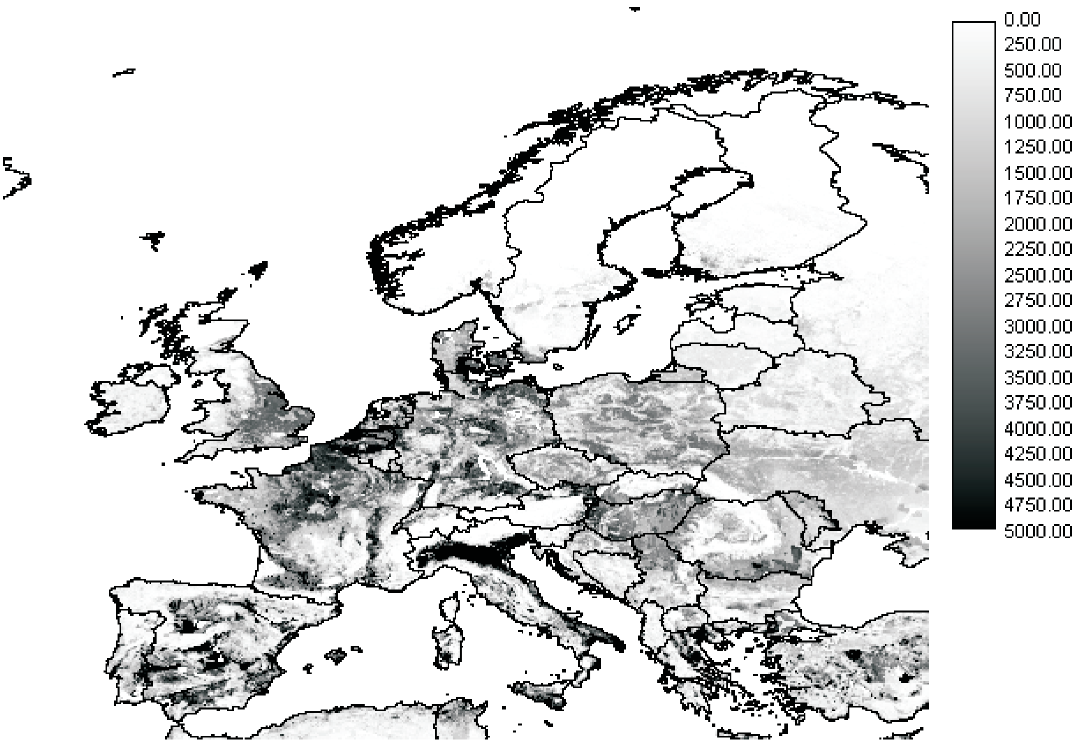

Figure 1: Total crop production value, GK dollars per grid cell.

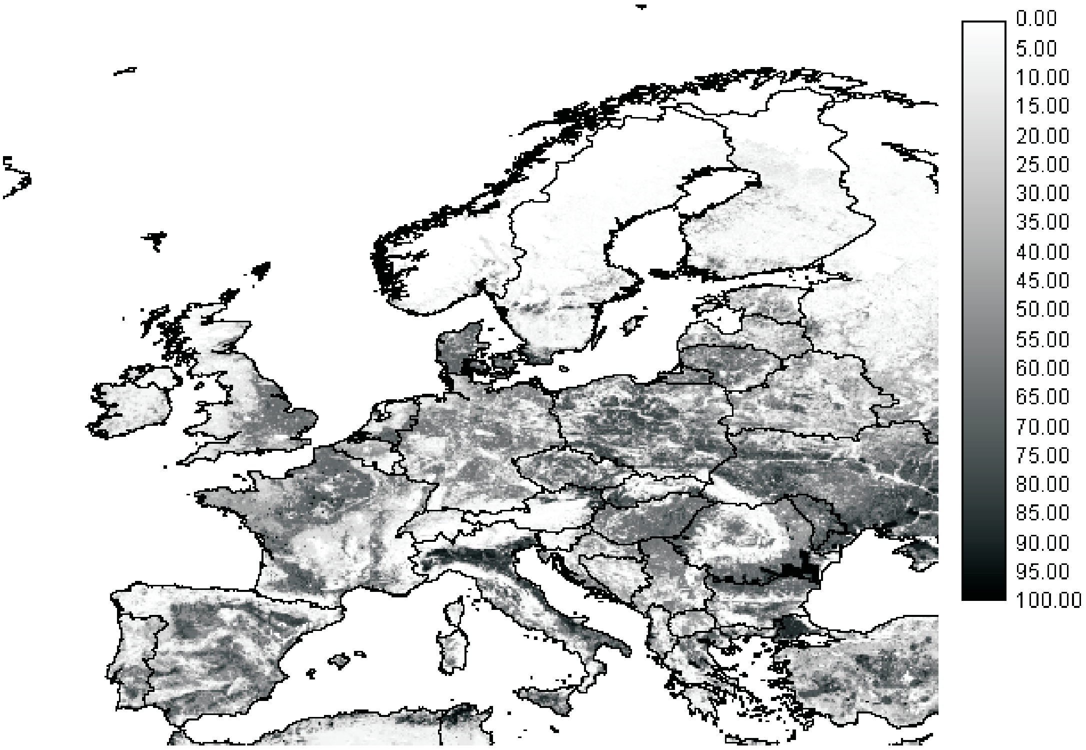

Figure 2: Share of cultivated land per grid cell.

ties and imprecision, are among the major challenges of the downscaling methodology, ultimately determining the success of these procedures.

An important issue for future research, besides the uncertainty of "priors" and other parameters, is concerned with the incorporation of downscaling methods within the overall decision making problems, i.e., similar to the stochastic optimization theory.

References

[1] Bregman, L.M. (1967): Proof of the Convergence of Sheleikhovskii's Method for a Problem with Transportation Constraints. Journal of Computational Mathematics and Mathematical Physics 7/1, 191-204 (Zhournal Vychislitel'noi Matematiki, USSR, Leningrad, 1967).

[2] Bierkens, M.F.P., Finke, P.A., de Willigen, P. (2000): Upscaling and Downscaling Methods for Environmental Research. Kluwer, Dordrecht, The Netherlands.

[3] Borwein, J.M., Lewis, A.S., Nussbaum, R.D. (1994): Entropy Minimization, DAD Problems, and Doubly Stochastic Kernels. Journal of Functional Analysis 123, 264-307.

[4] Carruthers, G.A.P. (1956): An Historical Review of the Gravity and Potential Concepts of Human Interaction. Journal of American Institute of Planners 22.

[5] Esopo D.A., Lefkowitz, B. (1963): An Algorithm for Computing Interzonal Transfers Using the Gravity Model. Operation Research 11/6, 901-907.

[6] Ermoliev, Y (1976): Stochastic Programming Methods. Nauka, Moscow (In Russian).

[7] Ermoliev, Y. and Wets R. (Eds.) (1988): Numerical Techniques of Stochastic Optimization: Computational Mathematics. Springer-Verlag, Berlin, Germany.

[8] Fisher, R.A. (1922). On the Mathematical Foundations of Theoretical Statistics. Philosophical Transactions of the Royal Society of London, Series A222, 309-368.

[9] Fischer, G., van Velthuizen, H.T., Nachtergaele, F.O., and Medow, S. (2000): Global Agro- Ecological Zones 2000. International Institute for Applied Systems Analysis and Food and Agriculture Organization of the United Nations, Laxenburg, Austria. CD-ROM and web-site http://www.iiasa.ac.at/Research/LUC/GAEZ/index.htm.

[10] Fischer, G., H.T. van Velthuizen, M.M. Shah, and F.O. Nachtergaele (2002): Global Agro- ecological Assessment for Agriculture in the 21st Century: Methodology and Results. Research Report RR-02- 02. International Institute for Applied Systems Analysis, Laxenburg, Austria.

[11] Golan, A., Judge, G., Miller, D. (1996): Maximum Entropy Econometrics: Robust Estimation with Limited Data. Series in Financial Economics and Quantitative Analysis, John Wiley and Sons Ltd, Baffins Lane, Chichester, West Sussex PO19 1UD, England.

[12] Kullback, J. (1959): Information Theory and Statistics. John Wiley and Sons, New York. Analysis. Princeton University Press, Princeton. Theory of Communication. 27, 379-423.

[13] Sinkhorn, R. (1964): A Relationship between Arbitrary Positive Matrices and Doubly Stochastic Matrices. Annals of Mathematical Statistics 35/2, 876-879.

[14] Wald, A. (1949): Note on the Consistency of the Maximum Likelihood Estimate. Annals of Mathematical Statistics 20, 595-601.

[15] Wets, R. (1999): Statistical Estimation from An Optimization Viewpoint. Annals of Operation Research 85, 79-101.

[16] Wood, S., Sebastian, K., Scherr, S. (2000): Pilot Analysis of Global Ecosystems: Agroecosystem. A Joint Study by International Food Policy Research Institute and World Resource Institute, Washington D.C.

[17] Zangwill, W.I. (1969): Nonlinear Programming, A Unified Approach. Prentice Hall, Englewoods Cliffs, New Jersey.

[18] Zenios, A., Zenious, S. (1992): Robust Optimization for Matrix Balancing with Noisy Data. Report 92-01-03, Dept. of Decision Sciences, Wharton School, Univ. of Pennsylvania.