ELSEVIER

Agriculture, Ecosystems and Environment 87 (2001) 215–232

Agriculture Ecosystems & Environment

www.elsevier.com/locate/agee

Issues of scale for environmental indicators

A. Stein [a,*], J. Riley [b], N. Halberg [c]

[a] *Mathematical and Statistical Methods Group, Wageningen University, Dreijenlaan 4, 6703 HA Wageningen, The Netherlands*
[b] *Department of Statistics, IACR Rothamsted, Harpenden AL5 2JQ, UK*
[c] *Danish Institute for Agricultural Sciences, P.O. Box 50, DK-8830 Tjele, Denmark*

Received 10 October 2000; received in revised form 12 January 2001; accepted 19 January 2001

Abstract

The value of environmental indicators largely depends upon the spatial and temporal scale that they represent. Environmental indicators are dependent upon data availability and also upon the scale for which statements are required. As these may not match, changes in scales may be necessary. In this paper a geostatistical approach to analyse quantitative environmental indicators has been used. Scales, defined in terms of resolution and procedures, are presented to translate data from one scale to another: upscaling to change from high resolution data towards a low resolution, and downscaling for the inverse process. The study is illustrated with three environmental indicators. The first concerns heavy metals in the environment, where the zinc content is used as the indicator. Initially, data were present at a 1 km^2 resolution, and were downscaled to 1 m^2 resolution. High resolution data collected later showed a reasonable correspondence with the downscaled data. Available covariates were also used. The second example is from the Rothamsted's long-term experiments. Changes in scale are illustrated by simulating reduced data sets from the full dataset on grass cuts. A simple regression model related the yield from the second cut to that of the first cut in the cropping season. Reducing data availability (upscaling) resulted in poor estimates of the regression coefficients. The final example is on nitrate surpluses on Danish farms. Data at the field level are upscaled to the farm level, and the dispersion variance indicates differences between different farms. Geostatistical methods were useful to define, change and determine the most appropriate scales for environmental variables in space and in time.

Keywords: Spatial variability; Upscaling; Downscaling; Environmental indicators; Block kriging; Simulation; Zinc; Nitrate; Rothamsted

1. Introduction

Environmental indicators are of an increasing importance to relate the state of the environment to those who are interested in it, or responsible for it. To formulate indicators, information is necessary that is available at different scales. Environmental indicators are usually quantitative expressions measuring some particular condition in relation to an existing threshold value. Decision makers commonly use indicators to communicate with general public. It is therefore important to define and quantify indicators in a scientifically justifiable way. It has been recognised that the quality of indicators relies on the scale which they represent. The quality of the state of the environment at a provincial scale, for example, requires different information compared to the state of the environment at the scale of an urban region in the province. In this study, attention focuses on the relation between the scale of observation, and the scale at which information is required, usually the region, area or the system limits that are covered by an indicator. Although space and

* Corresponding author. Tel.: +31-317-483551; fax: +31-317-483554.
E-mail address: alfred.stein[a]mat.dpw.wau.nl (A. Stein).

0167-8809/01/$ – see front matter
PII: S0167-8809(01)00280-8

time are different dimensions, the time dimension in relation to scale issues is considered as well.

So far, the issue of scale for sustainability indicators has barely been addressed in the literature. Scale issues are considered to be of importance (Bierkens et al., 2000) and advantages have been reported in hydrology (Feddes, 1995) and soil science (Hoosbeek and Bouma, 1998; McBratney, 1998). The main components in terms of changing or matching scales are upscaling and downscaling. Upscaling is the process of aggregating information collected at a fine scale towards a coarser scale (Van Bodegom et al., 2002). Downscaling is the process of detailing information collected at a coarse scale towards a finer scale. In upscaling, the distinction is necessary between upscaling of processes and upscaling of data. The first concerns the existence of different processes that act at different scales. For up- and downscaling, therefore, these processes are to be matched, i.e. it has to be identified how these processes interact. Attention focuses on the second topic, and addresses useful ways of up- and downscaling for individual sets of data on indicators. In this paper up- and downscaling of processes is not considered, as it is primarily of a disciplinary nature.

The objective of this paper is to identify, quantify and apply methods for up- and downscaling of agricultural and environmental indicators and to demonstrate their degree of influence on policy making. For political decision making it is important to know the scale at which political statements are required. Thus, central to the paper are the questions whether it is possible to define indicators for use at different levels of scale, what are the associated problems and what are their advantages. For that reason, scale issues are illustrated with three examples, concentrating primarily on the specific aspects of indicators. The first example analyses zinc concentrations in groundwater as an indicator for environmental quality in relation to heavy metals. This indicator shows different behavior when considered at different scales (Stein et al., 1995). In particular, prior information is only useful at the coarsest scale. The second study analyses scale aspects in time, where yearly data on grass yields from a long-term Rothamsted's experiment are sampled at increasingly coarse scales. The third study concerns the use of nitrogen surplus as an indicator for N-losses in farming systems for intensive Danish farms.

2. Materials and methods

2.1. Choice and modelling of indicators

In this study an indicator is defined as a measure to describe or evaluate a particular system. Examples are *acidification*, *scorching* and *global change*. Commonly, threshold values are attached to an indicator. A particular indicator is usually defined through communication with general public and contains normative, political, economical and ethical aspects. Let therefore the indicator i be a model of a general concept I. Closeness between i and I must exist conceptually. However, often i is formulated in everyday language, leading to a rather fuzzy terminology, that has to be approximated in some sense by I (De Bruin, 2000). Modelling is done by a relation f depending on a range of variables $x_1, \ldots, x_n$: $i = f(x_1, \ldots, x_n)$ (Smith, 1999). At this stage, no requirement is put on the functional relation $f(\cdot)$. Data quality of indicators depends on the closeness between i and I, on the specific form of the relation $f(\cdot)$, for example whether it is qualitative or quantitative, on the amount and quality of the variables x_i and on the scale that both I and i represent. An obvious relation exists between the functional relation and quality of observations on x_i. If a good model is based on poor data, the quality of the indicator can be doubted.

Indicators require a scientific definition and a sound method of evaluation (Gaunt et al., 1997). Indicators should:

1. Explicitly relate to a problem or a question of interest for those who are either outside or part of the system to which the indicator applies.
2. Link the system with the problem owner in a transparent way, i.e. give the interested parties adequate knowledge concerning the performance of a system with regard to the relevant question.
3. Be applicable to various systems, and be able to show changes over time.
4. Be feasible to register or calculate at a reasonable cost. This permits multiple measurements to be made in various systems, and hence the monitoring of the state of the environment, the comparison of different sub-systems and the maintainance of the indicator as well at a political level.

The second and third demands require that indicators are suitable for the level of scale that matches the decision making they are intended for. For that reason, data might be aggregated from one level of scale to another. A very simple relation, depending on a single variable may be relevant for some concepts I, for others a more complex relation involving many variables at various levels may be required. Statistics and scale aspects are important when relating the indicator to a particular problem. Contributions from statistics are in summarizing data, interpolating from points to areas of land and design of optimal sampling schemes, but also in calibrating and validating models and in using available prior information.

2.2. Scale aspects

2.2.1. Representativeness

Environmental indicators at different scales are considered in relation to their spatial and temporal variation. Let A be the domain of an indicator, i.e. an area ($A \subset R^2$) or a time interval ($A \subset R^+$). An indicator varies at a coarse scale if a few evaluations within a relatively large area or at a relatively long time-span are sufficient to characterize its variation. Such coarse scale variation is always in contrast to variation at a finer scale where either a few evaluations within a small domain or a large number of evaluations in a large domain characterize its variation.

Let $I(s)$ for $s \in A$ be an environmental indicator defined within A. Evaluations $i(s_1), \ldots, i(s_n)$ characterize variation within A. These can be observations for some indicators, but may be modelled values as well. Suppose therefore that all evaluations of an indicator are quantitative. Such data are characterised by their scale. Spatial scale refers to their representativeness for single locations or for larger strata like mapping units or blocks. Temporal scale refers to daily or monthly averages as compared to individual evaluations. Also, scale is related to the distinction between replications (within-stratum variation) and representative measurements (between stratum variation). Suppose that A is divided into strata $A_1, \ldots, A_p$ and that the data are sufficiently dense to characterize these, i.e.:

- $n \geq p$.
- Each A_a contains at least one $i(s_j)$.

Depending on the ratio of p and n and the distribution of the observations over the strata it may be possible to quantify the difference between the within-strata and between-strata variation. If $p = n$ and each A_a contains one observation, then it is possible only to characterize the variation between strata. If each A_a contains at least three observations, standard deviations within the strata may characterize the within-strata variation and statistical testing becomes possible. If n further increases, the spatial structure of variation can be assessed, for example using the variogram.

A second level of strata may describe variation at a finer scale, i.e. there exists a second sub-division of A into substrata $B_1, \ldots, B_q$ with $q > p$. Often such a finer classification is a real sub-classification of the classification into $A_1, \ldots, A_p$, i.e. each B_b belongs to precisely one A_a, whereas each A_a may contain several B_bs, a so-called hierarchical classification. In practical studies, though, this may not necessarily hold. As for the A-classification, each B_b should contain at least one observation to be able to characterize the between-strata variation and an increasing number of observations gives an increasing amount of information. A hierarchical analysis may then relate the variation of the finer B-scale within the variation at the coarser A-scale.

Variation at the point level is slightly different from the previous partitions as the collection of individual points does not fully cover A. For that reason a mechanism should exist to relate the points to the area, and hence geostatistical procedures come into play. Each point is representative only for the location where it was measured. The simplest mechanism is where every point in A is assigned the value of the closest point. A somewhat more elaborate concept is where a partitioning exists where each point characterizes a single stratum. Many procedures exist to obtain values for larger areas. In this study, attention mainly focuses on geostatistical procedures, as these are the most general.

2.2.2. Changes in scale

To study changes in levels of scale consider upscaling from the B- to A-classification and from points to the A-classification. For convenience the highest level of classification is denoted as level I, whereas lower levels are given increasingly higher roman numbers. Alternatively, consider downscaling from A- to

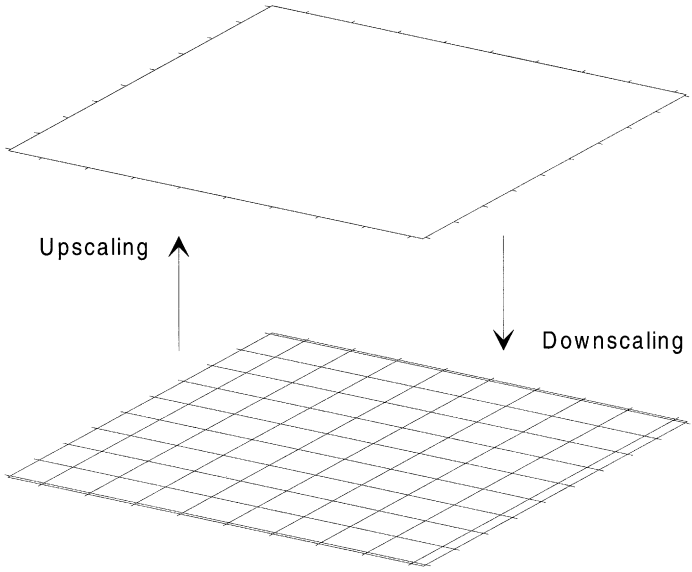

Fig. 1. Up- and downscaling processes in space, illustrated by a change in grid density.

B-classification and from A-classification to point data. The most common procedures for upscaling are averaging and block kriging procedures, whereas the most common procedures for downscaling are what is known in the statistical literature as the solution of the ecological fallacy problem, i.e. to reliably infer individual–level behaviour from aggregate data, leading to probabilistic statements. Additional information, say on a variable $Z(s)$ may be helpful to improve upon derived values.

These ideas can be applied to both spatial and temporal scales. The spatial case is two-dimensional (Fig. 1) where upscaling relates to changing from a fine to a coarse scale. In the temporal case upscaling relates to change in frequency of sampling and the estimation of indicators or processes at this new scale and their relationship with indicators or processes at the old scale. A sampling scheme at the maximal point frequency (say hourly data) can be reduced to a more sparse sampling frequency (monthly data, B-classification) and yet again to an even more sparse frequency (yearly data, A-classification).

2.2.3. The dispersion variance

Commonly variation is lost when data are upscaled. This is modeled by means of the dispersion variance which quantifies the amount of lost variance between two scales (Chilès and Delfiner, 1999). Consider the area A, centered on the point s_A and divided into p strata A_a, centered on the points $s_{A_a}, a = 1, \ldots, p$. An indicator is considered as a spatial (regional) variable $I(s)$ within A. For such a variable spatial variation is modeled as a function of the distance h by means of the variogram $\gamma_I(h)$.

The mean indicator within A equals $\bar{i}_A = 1/|A| \int_A i(s)\,\mathrm{d}s$, and similarly the mean indicator within each A_a equals $\bar{i}_{A_a} = 1/|A_a| \int_{A_a} i(s)\,\mathrm{d}s$.

Considering the two partitions as representations for two different scales, a deviation $[\bar{i}_A - \bar{i}_{A_a}]$ exists which corresponds to each of the p positions s_{A_a}. The dispersion variance of the p indicators $\bar{i}_{A_a}$ around their mean value $\bar{i}_A$ can therefore be characterised by the mean square deviation:

$$\sigma^2_{A|A_a} = \frac{1}{p} \sum_{a=1}^{p} [\bar{i}_A - \bar{i}_{A_a}]^2 \tag{1}$$

The dispersion variance $\sigma^2_{A|A_a}$ measures the dispersion of the values in the cells A_a. It can be viewed as the variance of the estimation of $I(A)$ by $I(A_a)$, the cell A_a being chosen at random among the p cells partitioning A. The problem is to estimate the two

main characteristics, $\bar{i}_A$ and $\sigma^2_{A|A_a}$. Estimation of the dispersion variance requires use of the variogram:

$$\sigma^2(A_a|A) = \frac{1}{|A|^2}\int_A\int_A \gamma_I(s'-s)\,\mathrm{d}s\,\mathrm{d}s' - \frac{1}{|A_a|^2}\int_{A_a}\int_{A_a}\gamma_I(s'-s)\,\mathrm{d}s\,\mathrm{d}s' \quad (2)$$

where the second integral is calculated according to an arbitrary A_a and $s'-s$ denotes a vector in two dimensions. In practical studies, integration is replaced by summation. A special case occurs when A_a reduces to a single point. Then the dispersion variance equals the average variogram for all distances within area A.

2.2.4. Geostatistical scaling procedures

Common procedures to address issues of scale can be found in the geostatistical literature. Here, attention focuses on two procedures: block kriging and block cokriging for upscaling, and point kriging and point cokriging for downscaling. Of importance is the presence of related covariables, that may be available at the finest scale when downscaling, or at the same scale as the indicator when upscaling.

Consider an m-variate environmental indicator $I(s)$, related to its position in space and/or time by the variable s. The indicator is either univariate ($m = 1$) or multi-variate ($m > 1$). A single indicator has $m = 1$, whereas higher values of m occur for multi-variate indicators, i.e. $m = 2$ if there is a single covariable and $m > 2$ for more than one covariable. Also a set of variograms $\gamma(h)$ exists, one for each component of $I(s)$ and one for each interaction between any two components. The variogram for each component measures its spatial dependence as a function of the distance between points where the indicator is defined, the variogram for interactions measures the strength of the interaction as a function of the distance between locations of components of the indicator. At this stage, the expectation is assumed to be constant and independent of location s for each of the components of $I(s)$, i.e. $E[I(s) - I(s+h)] = 0$, where 0 is a m-variate vector with all elements equal to 0. Generalisations can be found elsewhere in the literature (Stein et al., 1991).

The indicator $I(s)$ is defined throughout A, but usually only observed at a limited set of locations, and also our attention focuses on a limited set of points where it has to be evaluated. $I(s)$ is therefore restricted to a set of locations S. This set contains a subset S_O where $I(s)$ or one of its components is observed and another subset S_P where one of its components is to be predicted. Throughout this paper the index O denotes the observational part of a vector or matrix, and the index P denotes the part for which a prediction is made. The finite restriction of $I(s)$, arranged in the vector I, is partitioned into $\begin{pmatrix} I_\mathrm{O} \\ I_\mathrm{P} \end{pmatrix}$. For any linear combination of the P part, $\lambda'_\mathrm{P} I_\mathrm{P}$, a prediction is carried out with a linear combination of the O part, $\lambda'_\mathrm{O} I_\mathrm{O}$. The partitioned vector $\lambda = \begin{pmatrix} \lambda_\mathrm{O} \\ \lambda_\mathrm{P} \end{pmatrix}$ is the weight vector. For downscaling, the P part contains a single location and hence $\lambda_\mathrm{P} = 1$. This is a form of downscaling, as the information is detailed from a limited set of (observation) points to a larger set of (grid) points. For upscaling, the P-part consists of a set of locations and hence λ_P will be a vector of equal weights that sum to 1. Typically, the area destined for block kriging is discretised into a finite number n_P of individual locations, and each element of λ_P equals $1/n_\mathrm{P}$. In fact, block kriging is the limiting case for $n_\mathrm{P} \to \infty$, but in most practical studies a large value of n_P, say $n_\mathrm{P} = 100$ will not show much improvement of predictions when increasing (see Fig. 2). In practice, a single value is estimated at the center of the block and discretizing points are used for approximating the point-to-block and block-to-block covariance terms in the kriging system.

To make predictions, the linear model is used. The model matrix is denoted by X, partitioned into $X = \begin{pmatrix} X_\mathrm{O} \\ X_\mathrm{P} \end{pmatrix}$. Its number of columns equals m, the number of variables, and each column contains the values 0 and 1 only. Without covariables, i.e. for $m = 1$, $X = 1_{|\mathrm{O}+\mathrm{P}|}$, a vector of the size of the number of prediction and observation points, with all values equal to 1, that can be split as $X_\mathrm{O} = 1_{|\mathrm{O}|}$ and $X_\mathrm{P} = 1_{|\mathrm{P}|}$. If $m = 2$, X_O has two columns and has as many rows as the size of the finite restriction of indicator and covariable. Elements in the first column are equal to 1 if the finite restriction corresponds to the indicator and in the second column it corresponds to the covariable. The other elements are equal to 0. The sub-matrix X_P has values equal to 1 only in the first column, and values equal to 0 in the second column. For higher values of m, a similar structure applies. For the example, in

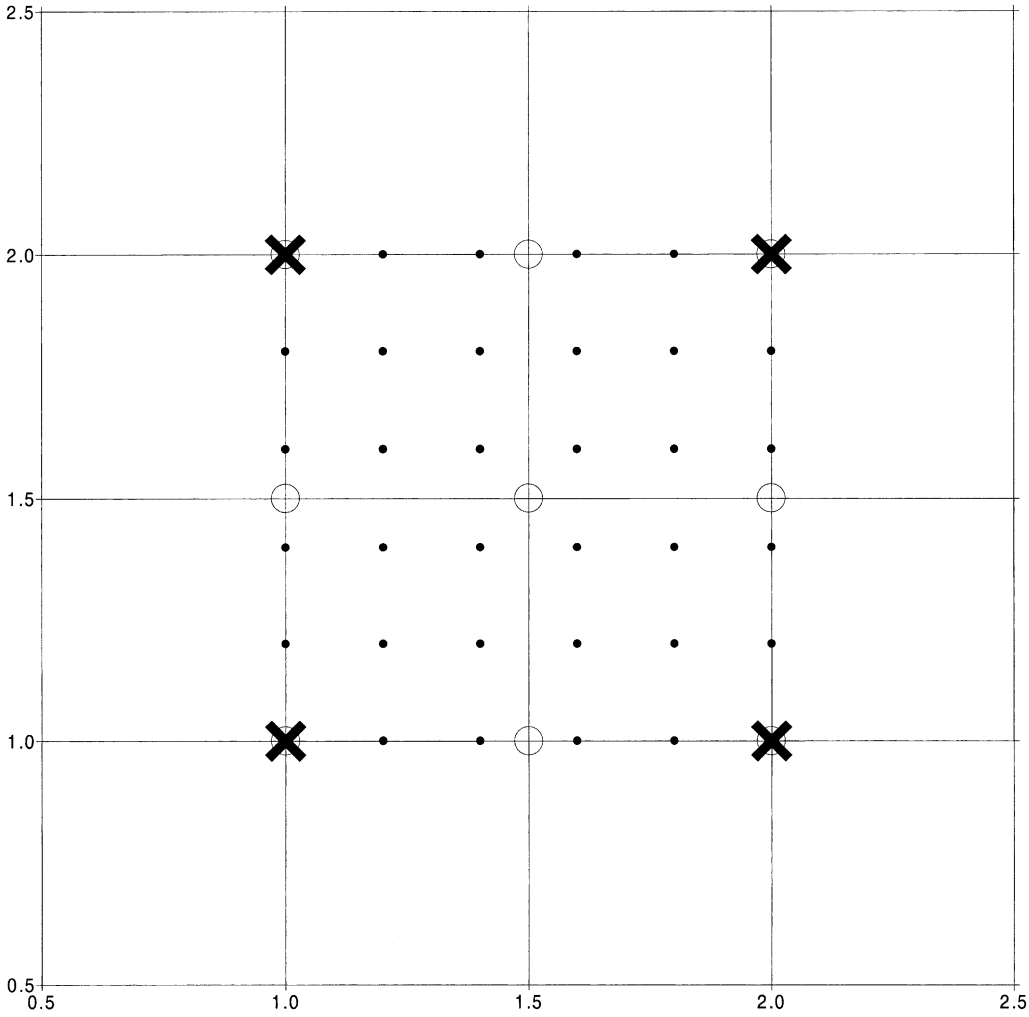

Fig. 2. Use of data for block kriging, where observations on variables (✖) and covariables (○) are used to a discretised block (●).

Fig. 2, four observations are made on the indicator, nine observations on a covariable and block kriging is carried towards an area that is discretised into 49 locations, X has two columns and 49 rows. The first four rows of X_O are equal to (1 0), the next nine rows of X_O are equal to (0 1), and all 36 rows of X_P are equal to (1 0). The distinction between upscaling and downscaling is therefore reflected in the number of rows of X_P, whereas the distinction between kriging and cokriging is reflected in the number of columns of X.

For any scaling procedure, linear unbiased (LU) predictors are used of $\lambda_P' I_P$: linear predictors $\lambda_O' I_O$ with the unbiasedness condition that $E[\lambda_P' I_P] = E[\lambda_O' I_O]$, which is equivalent to the prediction error having a zero expectation: $E[\epsilon' I] = 0$. The size of the prediction error is quantified by its mean squared error $E[\epsilon' I]^2 = \mathrm{Var}[\epsilon' I]$ and the best LU-predictor, the BLU-predictor, is the LU-predictor with minimum mean squared error. Because of these requirements, a prediction error $\epsilon' I$ satisfies $\epsilon' X = 0$ and its covariance structure is represented by the matrix Γ, parti-

tioned as:

$$\begin{pmatrix} \Gamma_{OO} & \Gamma_{OP} \\ \Gamma_{PO} & \Gamma_{PP} \end{pmatrix}$$

The matrix Γ_{OO} contains variogram values between the observations of the indicator, between those of covariables and those among them. The matrix Γ_{OP} ($=\Gamma'_{PO}$) contains the variogram values between observations and values actually not yet taken in the prediction locations, and the matrix Γ_{PP} contains the variogram values between the prediction locations. The general equation for the BLU predictor of $\lambda'_P I_P$ equals:

$$t = \lambda_P(X_P B + \Gamma_{PO} F) \tag{3}$$

with $B = VX'_O\Gamma_{OO}^{-1}Y_O$ and $F = \Gamma_{OO}^{-1}(Y_O - X_O B)$ (Cressie, 1993).

The mean square error of the BLU-predictor is obtained by using Eq. (3):

$$-\epsilon'\Gamma\epsilon = -\lambda'_P(\Gamma_{PP.O} + X_a V X'_a)\lambda_P \tag{4}$$

where $V = (X'_O\Gamma_{OO}^{-1}X_O)^{-1}$, $X_a = X_P - \Gamma_{PO}\Gamma_{OO}^{-1}X_O$ and $\Gamma_{PP.O} = \Gamma_{PP} - \Gamma_{PO}\Gamma_{OO}^{-1}\Gamma_{OP}$.

2.2.5. Upscaling: thematic contents comparison

Geostatistical procedures are useful when considering individual indicators, possibly related to a small number of covariables. They become intractable if the number of variables increases, or an indicator has in fact a multi-variate character. A set of procedures applicable under these circumstances is a thematic contents comparison. Multi-variate characterisation of individual strata are merged if the differences, e.g. expressed by means of the Mahalanobis distance, are not very large. An example is the quality of land, which can be expressed by a set of different variables, like economic, social, natural and physical variables. If this information is available at a set of points, any upscaling procedure should rely on all the measured variables, and hence a multi-variate procedure is appropriate.

2.2.6. Downscaling: the ecological fallacy

Downscaling has so far been addressed in the literature by probabilistic methods, also known as the ecological fallacy (Steel and Holt, 1996; King, 1997). Without taking into account spatial dependence consider p different strata at the coarsest level of scale, each containing n_{aj} observations on the jth variable, with mean and variance denoted by μ_{aj} and σ^2_{aj}, respectively. Typically, the first variable is the environmental indicator, the other variables are covariables. Let the overall means and variances be denoted by μ_{Aj} and σ^2_{Aj}, respectively. The between block variance is given by the matrix S, of which the entry $s_{j_1 j_2}$ is given by $\sum_{a=1}^{p} p/(p-1)(\mu_{aj_1} - \mu_{Aj_1})(\mu_{aj_2} - \mu_{Aj_2})$. The matrix S can be partitioned as:

$$S = \begin{pmatrix} S_{ii} & S_{iz} \\ S_{zi} & S_{zz} \end{pmatrix}$$

where the index i refers to the indicator that has to be downscaled, and z to the covariables. The relation between indicator and covariables is then given by $b_{zi} = S_{zz}^{-1}S_{zi}$, and its variances by the diagonal terms of the matrix S_{zz}^{-1}. Adjusted means, i.e. means for the indicator in stratum a corrected by co-information available at the finest level of scale, are obtained by $\tilde{\mu}_{ai} = \mu_{ai} + b_{zi}^{T}(\mu_{az} - \mu_{Az})$.

2.2.7. Downscaling: regression modelling

Another way of downscaling is entirely based on a dense set of measurements of covariables. At the highest level, a regression model $i = f(X, \beta)$ is developed to explain data i on the indicator that has to be downscaled by covariables X_I, yielding an estimated vector of coefficients $\hat{\beta}$. At the finest level, requiring downscaling, this model could be applied using observations on the covariable X_{II} and the estimated vector of coefficients $\hat{\beta}$. These modelled data could then serve as the environmental indicator at the finer scale level. Estimated indicator data at that level may then be interpolated to arrive at data at the required resolution. A well-recognised problem, though, is that this leads to the modifiable areal unit problem (MAUP), for example discussed in Fotteringham and Wong (1991). A second problem is that there need not be an obvious reason why the coefficients should be scale-invariant. Also, uncertainty about estimated indicator data is usually not taken into account in the interpolation. But as a first approximation it may give interesting results. Below it will be compared with other downscaling procedures.

3. Examples and results

3.1. Zinc data to address spatial resolution

The first practical case study deals with zinc (Zn) concentrations in the groundwater at different scales (Stein et al., 1995). Zinc serves as an indicator for the amount of heavy metal in the groundwater, and is as such an indicator for the quality of groundwater. It is a concentration that can be easily measured with a high precision. In the Dutch legislative system, three threshold values are associated with these values, being the target threshold T_T (65 mg l^{-1}), the intervention threshold T_I (800 mg l^{-1}) and an intermediate level T_M (433 mg l^{-1}). Soil with concentrations below T_T is considered to be clean, that with concentrations above T_I needs cleaning up, whereas T_M distinguishes heavily contaminated soil from low contaminated soil. Block cokriging is used for upscaling and point cokriging and regression modelling for downscaling.

In the study area in the southern Netherlands two levels of scale are distinguished:

L_I: The province of Noord-Brabant.
L_{II}: The city of Oss within L_I.

Effects of upscaling between the different levels are investigated. Prior information consists of a thematic mapper image and measurements on EC and pH at L_{II}. Concentrations of zinc in water samples were obtained from observation wells at 2–4 m depth. In total, 904 observation wells occurred at L_I which covers a much larger area than data at L_{II}, 86 wells occurred at L_{II}. Coordinates of the observation wells at L_I were in multiples of 1 km and were increased with 0.5 km to represent the actual center point of the groups of observations. This resulted in 152 groups of wells with different coordinates, each value being the average value of 1–25 individual wells. Within the city limits of Oss, 19 L_I observations occur. At L_{II} coordinates of a 1 m precision were available.

Maps for L_I and L_{II} are given in Fig. 3a and b. Fig. 3b shows much more variation, although the general picture of high values occurring at the center appears in both figures, due to averaging within blocks. Also, some blocks are left empty because of lack of observations.

The next stage deals with upscaling from L_I to L_{II}, and downscaling vice versa.

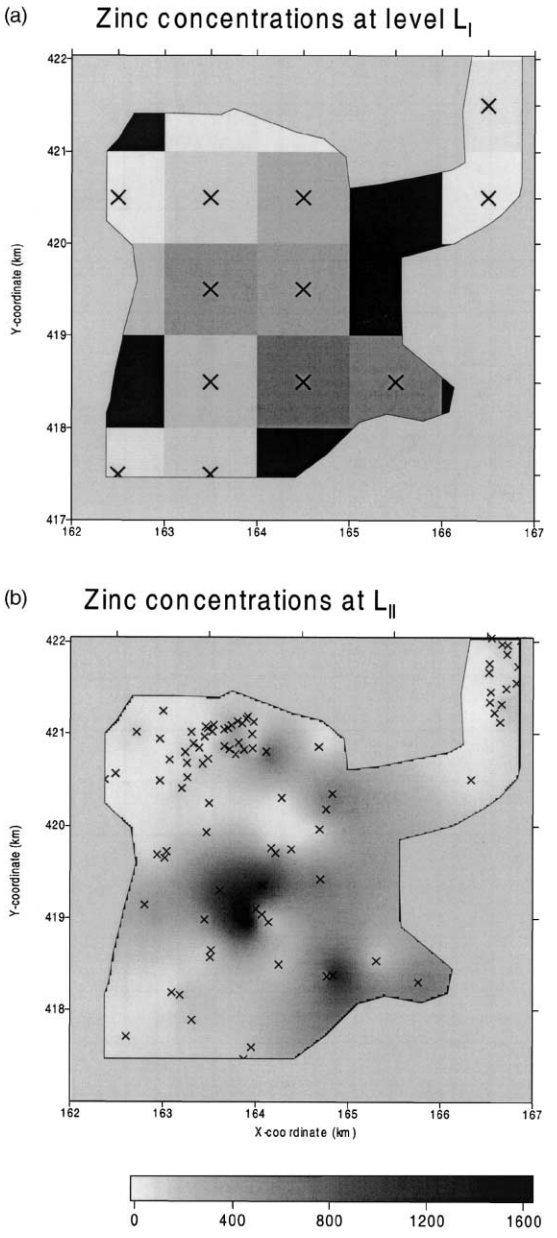

Fig. 3. Zinc concentrations at L_I (a) and L_{II} (b). Notice the larger blocks and the less pronounced greytones for the L_I map, caused by averaging data over the 1 km^2 grid cells.

3.1.1. Downscaling using probabilistic statements

Additional information consists of observations on pH and on electric conductivity (EC), available at both levels of detail. The vector b_{iz}, relating the indicator

Table 1
Descriptive statistics of [Zn], pH and EC at the two levels of scale, L_I and L_{II}

		N	Minimum	Maximum	*m*	*s*
L_I	Zn	15	18.8	767.7	279	231
	pH	14	4.9	6.9	6.19	0.65
	EC	15	289	1100	535	238
L_{II}	Zn	89	8.1	1650	264	366
	pH	63	5.0	7.8	6.34	0.64
	EC	63	84	1074	496	274

Zn to covariables pH and EC at L_I had as its coefficients −57.347 (0.182) and −0.775 (0.001), with the standard deviation given in brackets. Estimates for Zn within the 16 blocks of 1 km^2 are given in Tables 1 and 2 and shown graphically in Fig. 4. This shows some minor differences from Fig. 3b, although a correction was made for negative [Zn] values by equating these to zero.

3.1.2. Upscaling using block cokriging

Block cokriging served as an upscaling procedure, and allowed coverage of each grid cell, even if no observations were available (Fig. 5a). Notice that this may lead to a different pattern than the original L_I data, notably because values are also predicted in blocks without data. This was indicated as well by the high kriging standard deviations, with increased values for extrapolated data. Block kriging was also applied to cover the whole grid using the L_I data (Fig. 5b), including the standard deviation. This procedure yielded kriging standard deviations that do not show much difference between the blocks.

Table 2
Correction of [Zn] with EC and pH data towards the 16 blocks of 1 km^2

Stratum	x_C	y_C	[Zn]	$[Zn]_C$
1	162	417	86.000	−110.822
2	162	419	263.500	300.053
3	162	420	34.714	59.553
4	163	417	162.667	297.835
5	163	418	269.000	470.613
6	163	419	547.700	704.931
7	163	420	155.439	68.447
8	163	421	71.167	−157.098
9	164	418	767.667	881.228
10	164	419	441.300	583.574
11	164	420	355.258	309.107
12	164	421	18.750	−84.449
13	165	418	638.000	822.451
14	166	419	291.429	115.674
15	166	420	83.188	−29.175
16	167	417	29.500	8.819

3.1.3. Downscaling using regression modelling

A modelled indicator may show a large difference from the original indicator. A model for [Zn] at L_I was equal to:

$$[\text{Zn}] = 1976 - 246 \cdot \text{pH} - 0.37 \cdot \text{EC} \quad (5)$$

determined by regressing the L_I [Zn] data to the L_I pH and EC data. This model is applied at L_{II}, with pH and EC data available at that level (Fig. 6). The resulting map lead to differences from the original map, for example it showed less variation in the vicinity of points where no data were available on pH and [Zn], although zinc was observed at L_{II}. However, this double way of modelling (regression modelling, followed by interpolation) has as a risk that a much too smooth picture emerges, with smoothing caused by modelling on the one hand and interpolation on the other hand.

3.2. Grass data and the frequency of temporal sampling

In the second study attention focuses on scale issues in time. Choice of any sampling interval is usually dictated by some stage or event in a process and also by costs. Thus, annual harvests of wheat will dictate that the frequency is annual. On the other hand, estimates of numbers of plants lodged may be taken at various times during the growing season, but the number of times will be determined by costs. Baseline and follow-up surveys of farmer opinions must be done frequently but it is often difficult to decide upon the optimal frequency to detect changes that can be used to influence appropriate policy decision. It may be questioned whether different frequencies of observation give different impressions of the state of the environment. A simple example of quantitative data demonstrates some marked differences in a relationship between indicators as the scale of measurement in time is changed.

The Park Grass experiment at Rothamsted's Experimental Station was laid down in 1856 to examine

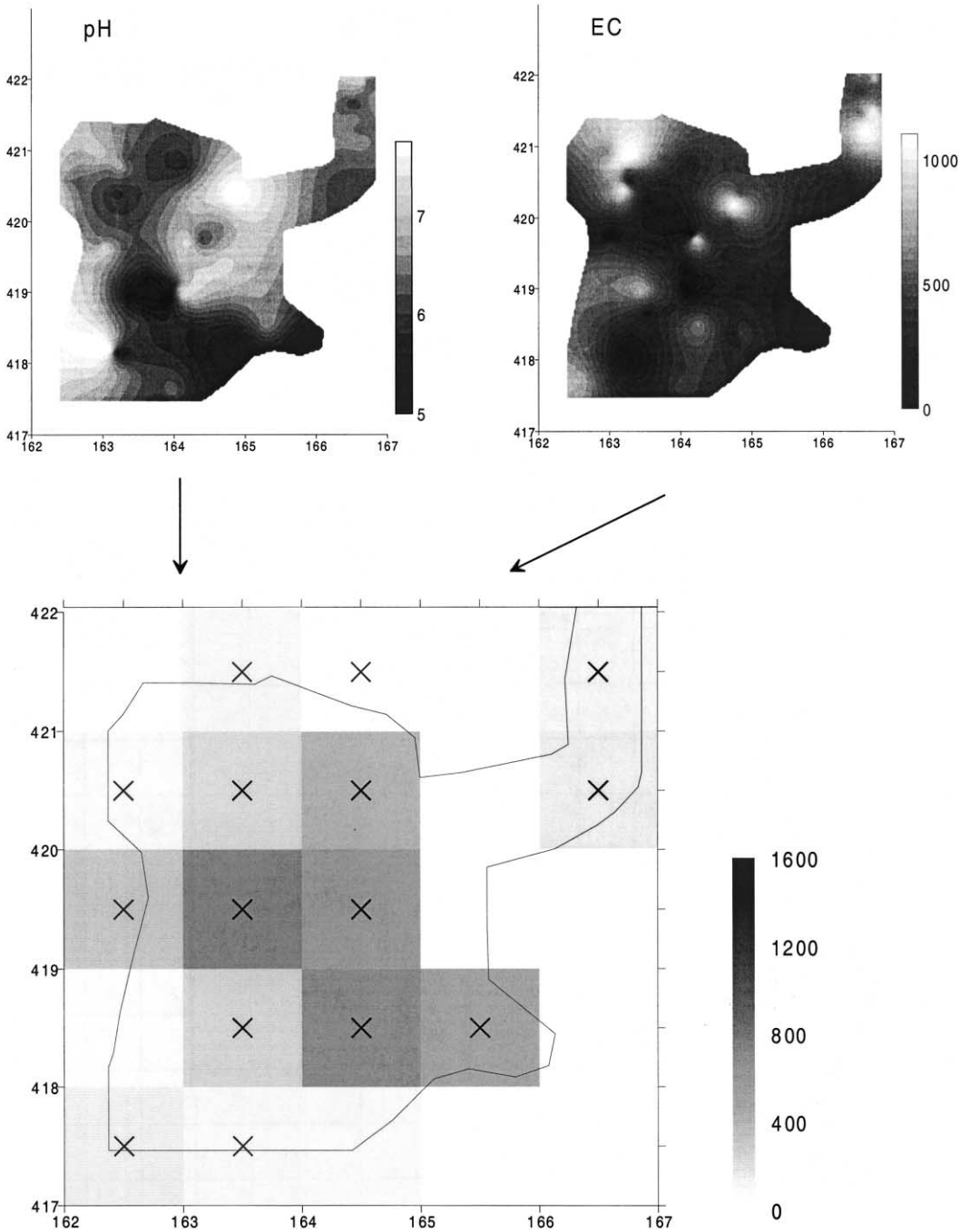

Fig. 4. Map of [Zn] at L_I using information of related pH and EC values.

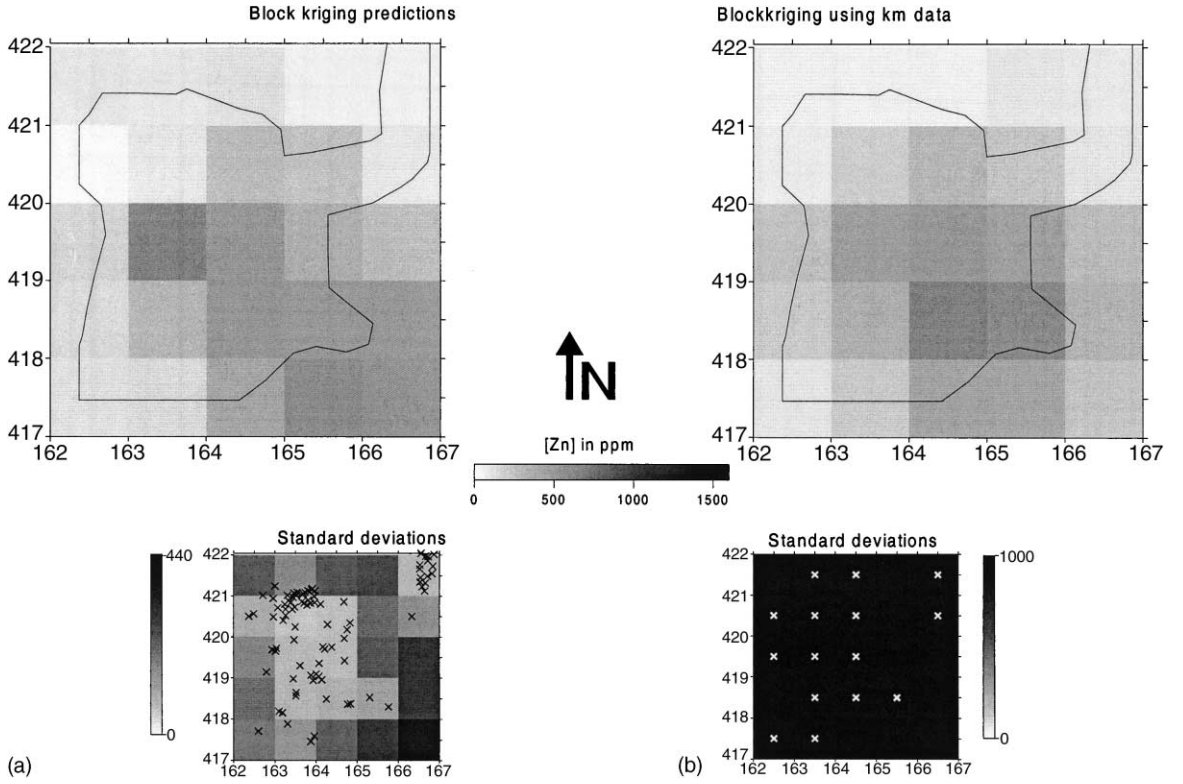

Fig. 5. Upscaling of [Zn] by means of block cokriging using data from L_{II} (a) and the uncertainty modelled by the block kriging error variance (b).

the long-term effects of different fertilizers on grassland. The plots have been sub-divided over the years to explore different inputs of lime and also other fertilizers. One-quarter plot is considered, plot 19–1 of 0.016 ha. Plot 19 currently receives 35 t ha^{-1} of farm yard manure every fourth year and the quarter sub-plot receives lime to maintain a pH of 7. The grass is cut twice during the growing season, at an interval of a few months, giving cut_1 and cut_2 for each year. A regression is done of cut_2 on cut_1 to find the relationship between the two. Changes in the models thus obtained are examined when different samples are selected, or when data are aggregated. Dry matter yields from the two cuts of this plot for the years 1920–1998 are used here to demonstrate the effects of changing sampling frequency in time. The number of points, 72, represents the maximum number of time points for which the treatment pattern was continuous, some data for some years being missing.

The following analyses were done:

S_1: The dry matter yields from cut_2 were regressed on those of cut_1 for each year, leading to a full data set of 72 paired observations.

S_2: Dry matter yields of cut_1 and cut_2 were chosen every other year, leading to a data set of 36 paired observations, and the process was repeated.

S_3: Dry matter yields of cut_1 and cut_2 were chosen every fourth year, leading to a data set of 18 paired observations, and the process was repeated.

S_4: Dry matter yields of cut_1 and cut_2 were chosen every eighth year, leading to a data set of nine paired observations, and the process was repeated.

S_5: Mean values of cut_1 and cut_2 were calculated for consecutive sets of four dry matter yields, with the first year value as the starting value for the subset choice, leading to a data set of 20 observations.

S_6: Similar to S_5, but with the second year as the starting value, giving 19 observations.

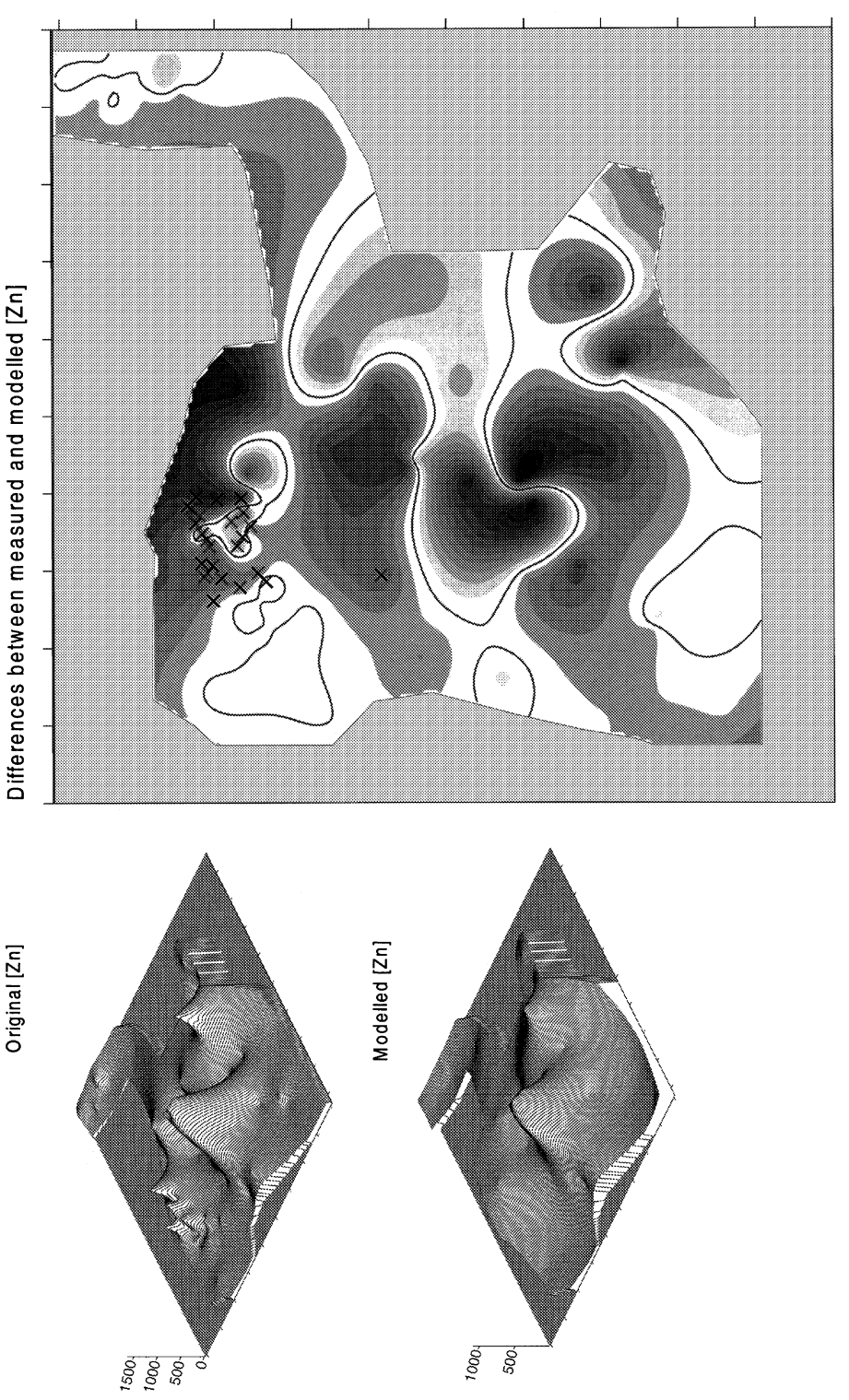

Fig. 6. Downscaling by means of modelling of individual [Zn] data is compared with the original data. Differences are large, ranging from −500 to +500 ppm. A larger variation is caused where modelling was not possible because of lack of data on the dependent variable.

Table 3
Estimated regression coefficients b_0 and b_1 of the second cut on the first cut with data sets at different scales

Dataset	N	b_0	S.E. (b_0)	b_1	S.E. (b_1)	t probability
S_1	72	0.061	0.425	0.561	0.117	<0.001
S_2	37	0.362	0.667	0.517	0.202	0.015
S_3	18	0.920	1.400	0.328	0.489	0.512
S_4	9	1.410	1.120	0.046	0.375	0.907
S_5	20	−0.391	0.584	0.677	0.166	<0.001
S_6	19	−0.732	0.764	0.785	0.223	0.003
S_7	19	−0.370	0.638	0.694	0.185	0.002
S_8	10	−0.913	0.325	0.829	0.093	<0.001

S_7: Similar to S_5, but with the third year as the starting value, giving 19 observations.

S_8: Mean values of cut_1 and cut_2 were calculated for consecutive sets of eight dry matter yields, with the first year value as the starting value for the subset choice, leading to a data set of 10 observations.

The number of points N in each data set and the estimates of the intercepts b_0 and slopes b_1 for each of the regression equations are in Table 3, with their precision and t probability for the slopes. None of the estimated intercepts was statistically different from zero. The slopes varied considerably, showing no great loss of significance when the data set was reduced by half but a large loss of significance when it was reduced to every four points and again when it was reduced by half to every eight points. The results reflect the greater relative dispersion of the points in S_3 and S_4. Data set S_5, using means of four values, has a similar slope to the complete data set and similar significance. S_6 and S_7 representing means of four values, shifted by one each time, show similar results to each other. There is clearly a difference in the dispersion of these three data sets as reflected by the standard error of b_1 for S_6, and the corresponding values of the significance levels. S_8 reduced to only 10 points, separated into two distinct groups in the lower left and upper right quarters of the graph with relatively little within-group dispersion. The fitted regression line therefore has the steepest slope and the smallest precision of estimation. Fig. 7 shows the first four data sets and the fitted regression lines.

The implications are that since the two cuts are the maximum that can be taken from the experiment, no more frequent measurements can be taken. This must be the 'maximal set'. Using this and measuring less frequently in time, the relationship is very different and not reliable. The fact that S_5 gave similar results is quite by chance, as is reflected by the different results obtained from choice of different starting points for the runs of four points. Therefore, quite different impressions are received if frequency of sampling becomes more widely spaced in time. If data are aggregated, greater relative dispersion in the data appears and the impressions change again, dependent on selection of aggregation zones. This situation commonly occurs when financial resources are limited. The scale of measurement is chosen to include the full time-span,

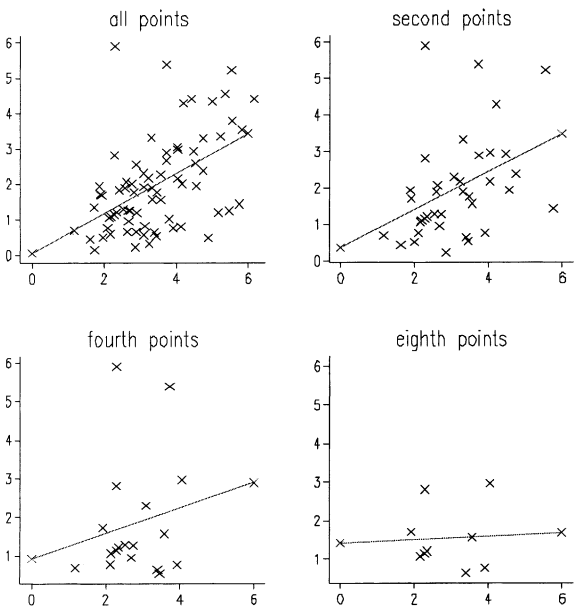

Fig. 7. Regression lines of fitting fertilizer contents at cut_2 on those at cut_1 at different temporal scales.

but can include only the maximum number of affordable samples. Whilst these may be spread evenly throughout the time-span, the estimated relationships, as shown here, may be quite different from that of the maximal set. Extreme points, or outliers, may not be included in the reduced sets and it is their presence which influences most the achieved relationship in the maximal set. Robinson and Tawn (2000) show that the observed extremes of a discrete time process depend on the process itself and the sampling frequency. The choice of scale of measurement — here frequency — is crucial to the picture presented to policy makers. A study of the influence of changes in time-scale for many commonly occurring relationships is therefore necessary.

3.3. Nitrogen surplus data as an indicator of N loss

The third example concerns N-surplus on 45 farms in Denmark. N-surplus per hectares is an indicator of the potential N-pollution from agricultural production occurring in a given area. It can be calculated for the field and farm level for farmer decision making (Halberg, 1999) and for comparison of different farming systems (Halberg et al., 1995). The farm balance is necessary for evaluation of the field balances. The average field balance should correspond to the farm level balance minus N-losses in stables. A high level of aggregation may correspond to high variation between the different strata at a lower level that may be caused by peak losses. Therefore, evaluation of the variation in field balances may be necessary for estimating the present losses and for discussing the potential to reduce the N-surplus at the farm level. On the other hand, evaluation of single field balances is not sufficient to evaluate the potential for reducing the losses since a reduction in one field might increase the loss in another field and thus not result in a decrease in the overall farm surplus. This especially applies to livestock farms with a relatively fixed amount

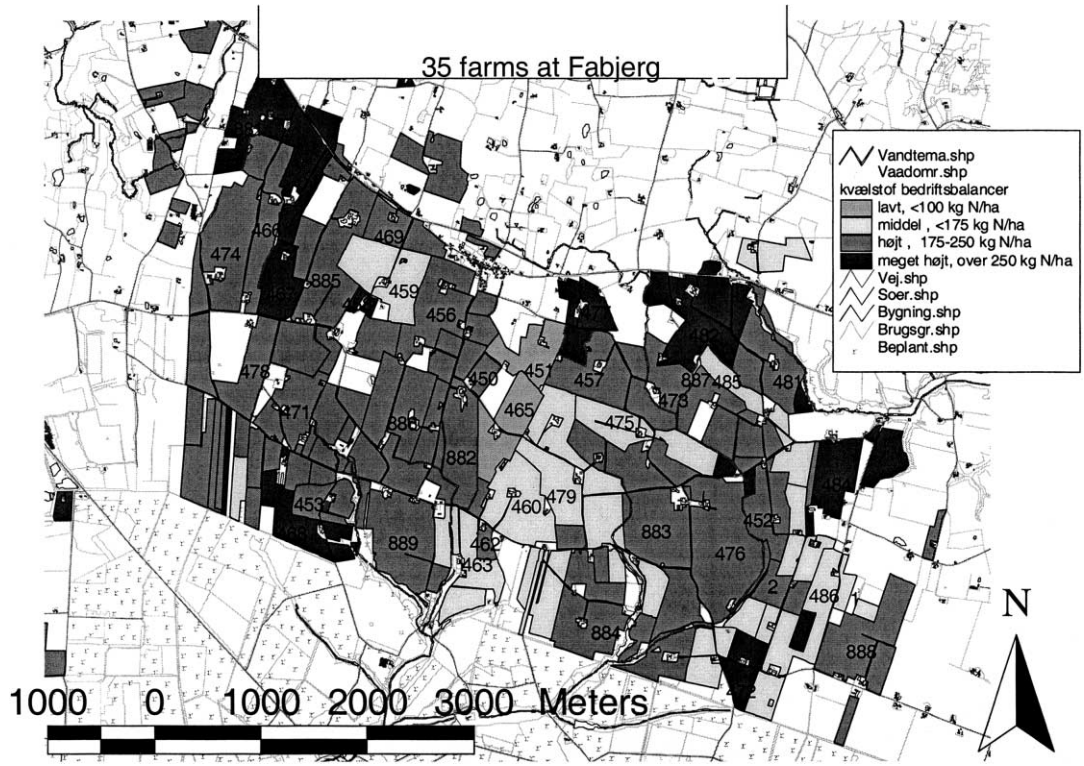

Fig. 8. Nitrogen surpluses measured at the farm scale.

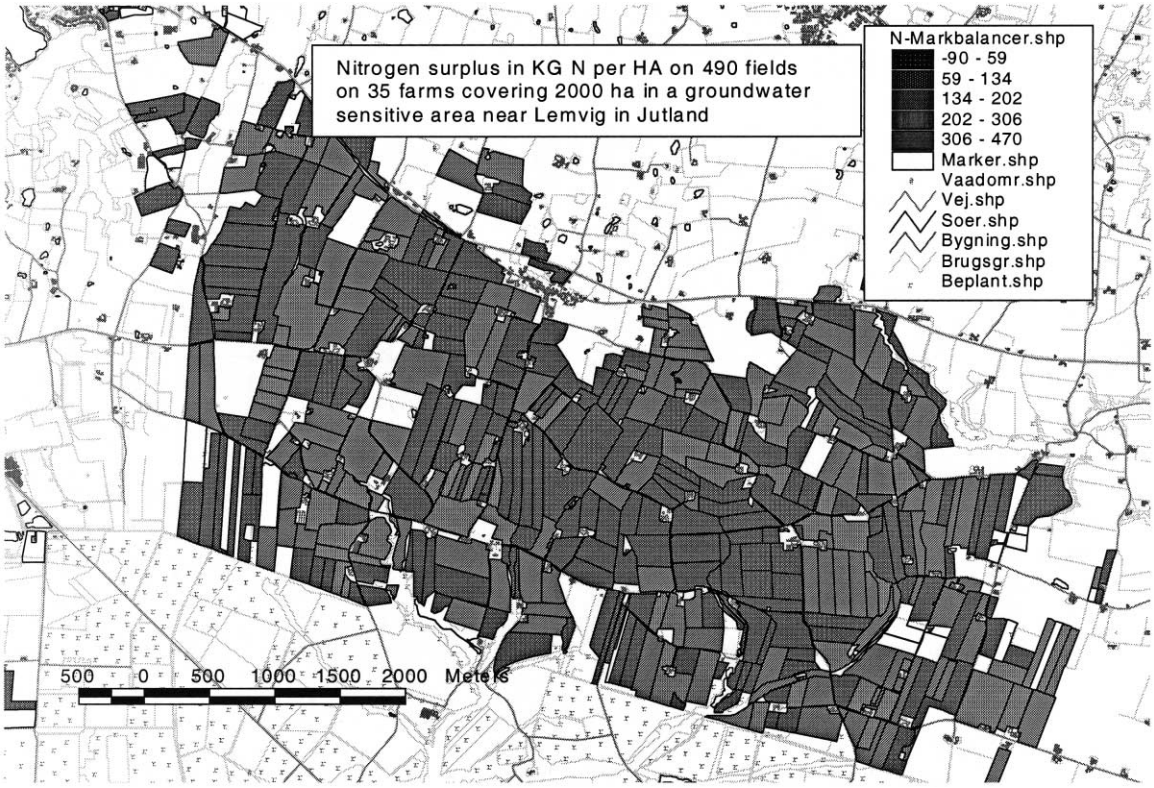

Fig. 9. Nitrogen surpluses measured at the field scale.

of manure. Also, on an aggregated scale the N-surplus might be useful to evaluate the total load of N in an area or the N-efficiency at country-level. Obviously the farm level is the most easy and secure level for calculations, but the field/crop level balances might give more information if properly checked. Differences between field-level N-balances depend on crop type, on amount of manure N supplied and probably on soil type. Therefore, the farm level balance covers systematic variation between individual fields. Different farm types may have different crop rotations. Some variation at the field level might be expressed by a combination of farm level N-surplus and farm type (i.e. dairy farm, pig farm, cash crop farm). The question is then, whether N-surplus at the farm level is sufficient to evaluate the average N-surplus (and potential N-loss) in an area (Fig. 8) or if it is necessary to know the variation in the field balances, which will be more costly to establish (Fig. 9). Clearly, Fig. 9 shows much more variation than Fig. 8.

To estimate the relations between the catchment level, the farm level and the field level consider the dispersion variance (Table 4, Fig. 10). The dispersion variance from catchment to farm level is only 41% of the dispersion variance from catchment to field level. On average, the dispersion variance from farm to field level is relatively small (2018). Therefore, the large increase has to be ascribed to some individual farms. This also emerges from Fig. 10, where a few farms are

Table 4
Dispersion variances of N-surpluses from the catchment (A) to the farm (A_a) level and from the catchment (A) to the field (B_b) level, and average dispersion variance from the farm (A_a) level to the field (B_b) level

Scales	Variance
From catchment to farm, $\sigma^2(A\|A_a)$	4910
From catchment to field, $\sigma^2(A\|B_b)$	11792
From farm to field, $\sigma^2(A_a\|B_b)$	2018

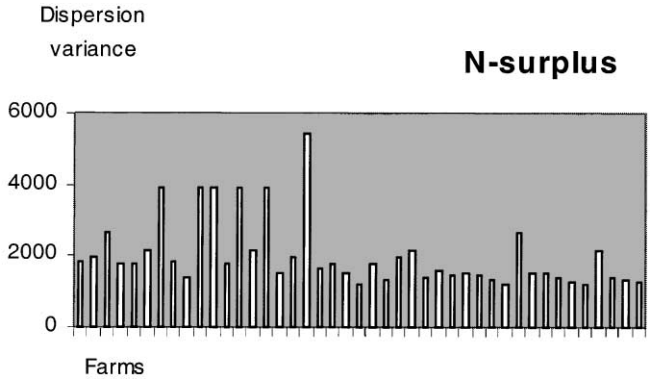

Fig. 10. Dispersion variance from the farm to the field level at each of the 45 different farms.

the main contributors to an increase in the dispersion variance.

4. Discussion

Scale issues are of primary importance for indicators at different scales different statements are obtained, for example in terms of statistics of parameters that describe these indicators. The zinc study clearly shows the different conclusions that are obtained at the two different scales. The grass cut data show clear effects of reducing data availability, and the nitrate data show how much information is lost when data at one scale are used at another scale.

From the examples in this study it is clear that upscaling has an effect on spatial and temporal variability. This could be both an advantage or a disadvantage. If interest centers on observing extremes in space and time, then upscaling is disadvantageous, as the small scale variation is smoothed away. But if policy making involves recognition of a general pattern and incidental extremes can be dealt with, then smoothing may possibly be advantageous. No policy should be made on single observations, i.e. single occurrences where thresholds may be exceeded. Here, also the support size is important, being the size of the area for which individual data are representative. If in spatial studies interest focuses on areas, then averages of multiple observations can be useful to characterize them. But, as was shown by the zinc data, that has a smoothing effect and hence influences the conclusions from such a study.

For decision support, models for leaching and transport are commonly used, e.g. in relation to spatial (geographical) information systems. Such models may show effects of possible scenarios. Often, interpolated data are being used for that purpose. Modern geostatistical models now increasingly consider spatial simulations. Such simulations provide a random field that reflects the statistical properties of the original data, like the same mean, variance and variogram as well as having the same observations as the original data. Although use of multiple simulations is in this way superior to the use of interpolated fields, the amount of work and computing time associated with it may be prohibitive.

Further, models may also be specific for various scales. It remains important to select a model and to validate its outcomes. These procedures are related to scales as well. Validating a point model at the square kilometer scale or validating a daily model at the weekly scale is obviously not very useful, as is validating models based on large regions at the point scale. For model validation, though, it is important to have the extremes properly modelled and to have a proper assessment of the variability.

For different multi-disciplinary indicators different time (and space) aggregation relationships may hold for different variables and policy makers should be made aware of this.

Indicators relate to dynamic systems. For that reason such a complex system should be monitored as well. In principle, therefore, this must be done continuously, requiring a continuous running of the

model and a continuous upgrading of input data. This might require a totally different approach towards collecting many data that are related to indicators, some that may not change for long periods of time.

Several tools are available to change scales on point data for indicators. Upscaling is not a problem, using geostatistics (block cokriging), downscaling is possible to a limited extent. For both up- and downscaling probabilistic statements, including confidence bounds appear to be useful. This requires care in interpretation, however, as environmental indicators primarily address public perceptions of the quality of the environment and which policy makers and the general public are also getting accustomed to. Modern information systems (GIS) are useful to visualize these effects and can contribute to better communication between politicians, decision makers and the general public.

5. Conclusions

On the basis of this study the following conclusions can be drawn:

- For upscaling, averaging, block kriging and block cokriging as geostatistical procedures can be applied. Averaging of data is the most global procedure, requiring little data, whereas more data are needed for block kriging and block cokriging.
- For downscaling, a probabilistic approach is most appropriate such as that provided by spatial simulations. Point kriging and results from the ecological fallacy principle can be applied for this purpose as well. Regression modelling seems for this purpose of less interest, mainly because the modelled relationships at one scale may not be applicable at another scale.
- Quantification was done in this study on three case studies. In the zinc study, scaling approaches worked nicely, showing the benefits of different approaches. In the temporal study, the variety of relations depending upon the scale of collected data was clearly visible. Finally, the nitrogen surplus study showed how the dispersion variance quantifies the additional uncertainty when changing from one scale to another.
- If the scaled relations are used for political decision making, different results are obtained for different scales. These differences can be quantified using the procedures presented in this study. However, the decision makers and public at large must be made aware of these differences.

Acknowledgements

Rothamsted Experimental Station is thanked for its permission to use the data in Section 3.2.

References

Bierkens, M.F.P., Finke, P.A., De Willigen, P., 2000. Upscaling and Downscaling Methods for Environmental Research. Kluwer Academic Publishers, Dordrecht.

Chilès, J.P., Delfiner, P., 1999. Geostatistics — Modelling Spatial Uncertainty. Wiley, New York.

Cressie, N.A.C., 1993. Statistics for spatial data. Wiley, New York.

De Bruin, S., 2000. Querying probabilistic land cover data using fuzzy set theory. Int. J. Geogr. Inform. Sci. 14, 359–372.

Feddes, R.A., 1995. Space and Time Variability and Interdependencies in Hydrological Processes. Cambridge University Press, Cambridge.

Fotteringham, A.S., Wong, D.W.S., 1991. The modifiable areal unit problem in multivariate statistical analysis. Environ. Plan. A 23, 1025–1044.

Gaunt, J.L., Riley, J., Stein, A., Penning de Vries, F.W.T., 1997. Requirements for effective modelling strategies. Agric. Syst. 54, 153–168.

Halberg, N., 1999. Indicators of resource use and environmental impact for use in a decision aid for Danish livestock farmers. Agric. Ecosyst. Environ. 76, 17–30.

Halberg, N., Kristensen, E.S., Kristensen, I.S., 1995. Nitrogen turnover on organic and conventional mixed farms. J. Agric. Environ. Ethics 8, 30–51.

Hoosbeek, M.R., Bouma, J., 1998. Obtaining soil and land quality indicators using research chains and geostatistical methods. Nutr. Cycling Agroecosyst. 50, 35–50.

King, G., 1997. A Solution to the Ecological Inference Problem. Princeton University Press, Princeton.

McBratney, A.B., 1998. Some considerations on methods for spatially aggergatinbg and disaggregating soil information. Nutr. Cycling Agroecosyst. 50, 51–62.

Robinson, M.E., Tawn, J.A., 2000. Extremal analysis of processes sampled at different frequencies. J. R. Stat. Soc. Ser. B 62, 117–135.

Smith, J., 1999. Models and scale: up- and downscaling. In: Stein, A., Penning de Vries, F.W.T. (Eds.), Data and Models in Action. Kluwer Academic Publishers, Dordrecht, pp. 81–98.

Steel, D.G., Holt, D., 1996. Analysing and adjusting aggregation effects: the ecological fallacy revisited. Int. Stat. Rev. 64, 39–60.

Stein, A., Van Eijnsbergen, A.C., Barendregt, L.G., 1991. Cokriging nonstationary data. Math. Geol. 23, 703–719.

Stein, A., Varekamp, C., Van Egmond, C., Van Zoest, R., 1995. Zinc concentrations in groundwater at different scales. J. Environ. Qual. 24, 1205–1214.

Van Bodegom, P.M., Verburg, P.H., Stein, A., Adiningsih, S., Denier van der Gon, H.A.C., 2002. Scaling effects on estimates of methane emissions from rice paddies. Environ. Ecol. Stat. Vol. 9 (1), in press.