Using landscape characteristics to define an adjusted distance metric for improving kriging interpolations

Lyon, S W; Sørensen, R; Stendahl, J; Seibert, J


Postprint available at: http://www.zora.uzh.ch

Posted at the Zurich Open Repository and Archive, University of Zurich. http://www.zora.uzh.ch

Using landscape characteristics to define an adjusted distance metric for improving kriging interpolations

Steve W. Lyon\textsuperscript{1}, Rasmus Sorensen\textsuperscript{2}, Johan Stendahl\textsuperscript{3}, and Jan Seibert\textsuperscript{1,4}

\textsuperscript{1}Physical Geography and Quaternary Geology, Stockholm University, Stockholm, SWEDEN

\textsuperscript{2}Aquatic Science and Assessment, Swedish University of Agricultural Sciences, Uppsala, SWEDEN

\textsuperscript{3}Soil and Environment, Swedish University of Agricultural Sciences, Uppsala, SWEDEN

\textsuperscript{4}Department of Geography, University of Zurich, Zurich, SWITZERLAND

Correspondence to: Steve W. Lyon (steve.lyon@natgeo.su.se)
Physical Geography and Quaternary Geology
Stockholm University
106 91 Stockholm, SWEDEN
Phone: +46 (0)8 164888
Abstract

Interpolation of point measurements using geostatistical techniques such as kriging can be used to estimate values at non-sampled locations in space. Traditional geostatistics are based on the spatial autocorrelation concept that nearby things are more related than distant things. In this study additional information was used to modify the traditional, Euclidean, concept of distance into an adjusted distance metric that incorporates similarity in terms of quantifiable landscape characteristics, such as topography or land use. This new approach was tested by interpolating soil moisture content, pH and carbon to nitrogen (C:N) ratio measured in both the mineral and organic soil layers at a field site in central Sweden. Semivariograms were created using both traditional distance metrics and the proposed adjusted distance metrics to carry out ordinary kriging (OK) interpolations between sampling points. In addition, kriging with external drift (KED) was used to interpolate soil properties to evaluate the ability of the adjusted distance metric to incorporate secondary data into interpolations. The new adjusted distance metric typically lowered the nugget associated with the semivariogram thereby better representing small scale variability in the measured data compared to semivariograms based on a traditional distance metric. The pattern of the resulting kriging interpolations using KED and OK based on the adjusted distance metric were similar since they represented secondary data and, thus, enhanced small-scale variability compared to traditional distance OK. This created interpolations that agreed better with what is expected for the real-world spatial variation of the measured properties. Based on cross-validation error, OK interpolations using the adjusted distance metric better fit observed data than either OK interpolations using traditional distance or KED.

KEYWORDS: Geostatistics, Kriging, Semivariogram, Soil Moisture, Soil pH, C:N Ratio
1. Introduction

In environmental sciences, observations are often made at various locations in space. Usually, observations are spatially sparse compared to the heterogeneity found in natural systems due to constrains of sample collection and analysis. Still, the goal of such observations is often to infer continuous descriptions, or maps, of the property of interest.

Geostatistics offer methods to interpolate between point observations. At the heart of most geostatistical applications, semivariograms quantify spatial autocorrelation by evaluating to which degree samples collected near each other are more similar than those collected further from each other. To construct a semivariogram, the distances between all sampling locations, $d_{ij}$, are divided into lag bins of given distance intervals described by a mean distance, $d$. This allows defining the semivariance for each lag bin, $\gamma_s(x)$, as

$$\gamma_s(d) = \frac{1}{2N(d)} \sum_{(i,j)} (Y_i - Y_j)^2$$  \hspace{1cm} (1)

where, $N(d)$ is the number of pairs, $Y_i$ and $Y_j$ are the variable of interest at sampling point $i$ and $j$, respectively, with summation over all pairs $(i,j)$ which have a distance within a certain interval. Usually the average bin semivariance is plotted against the mean bin distance to create the sample semivariogram. This describes the expected variance between two sampling locations in space as a function of distance and is fitted by a function (also called a model) to create the semivariogram. The main parameters of the fitted semivariogram model are the nugget, the sill, and the range. The range provides a measure of the maximum distance over which spatial correlation affects the variable of interest. The sill represents the spatial variance of two distant measurements. The nugget gives the variance in the measurement due to the inherent variability of the sampling device and the occurrence of spatial patterns smaller than the sampling interval. The fitted semivariogram model provides a manner to interpolate the
variable of interest between sampling locations using kriging. Traditional kriging techniques, such as ordinary kriging (OK), generate predictions at unobserved locations by weighting the influence of neighboring sampled locations based on their distance and configuration. In land resource inventories, kriging and its variants have long been widely recognized as primary spatial interpolation techniques (Hengl et al., 2004).

In addition to traditional univariate kriging techniques, a number of interpolation techniques and variants exist to incorporate secondary information (See Goovaets, 1999 or McBratney et al., 2000 for detailed discussions). When measurements are sparse or poorly correlated in space, the estimation of the primary attribute of interest is generally improved by accounting for secondary information originating from other related categorical or continuous attributes (Goovaerts, 1999). For example, co-kriging uses the spatial information from primary observations along with spatial correlation between these primary observations and a secondary variable to make estimations at unobserved sites. More generally, however, practice has shown that co-kriging improves over kriging only when the primary observations are undersampled with regard to the secondary variables and those secondary data are well correlated with the primary value to be estimated (Journel and Huijbregts, 1978; Goovaerts, 1999). Kriging with external drift (KED) is similar to universal kriging, but it uses a secondary variable to represent trend in the primary data (Goovaerts, 1997). KED and other kriging with regression ‘hybrid’ interpolation techniques, such as regression kriging (e.g., Odeh et al., 1994), have become increasingly popular as remotely sensed data become more readily available for interpolations (especially for soil inventory) at the regional/catchment (20m – 2 km) scale (McBratney et al., 2000). These regression techniques are, however, rather complex and can be difficult to apply. There is, thus, a risk of incorrect applications, which
can result in worse estimates than straightforward OK since they rely on the correlation strength between observations and auxiliary variables (Hengl et al., 2004).

The aforementioned semivariogram models and kriging techniques are traditionally based on Euclidean distance metrics or the shortest path between observations defined using a Cartesian coordinate system. Often, this definition of distance is assumed by most geostatistics packages without consulting the user (Christakos, 2000). However, in many instances, the space separating two sampled points may represent a partial or complete barrier owing to biological or physical characteristics of the intervening space (Jensen et al., 2006). It is possible to introduce new metrics for distance in geostatistical calculations. Non-Euclidean distance metrics can incorporate physical properties of how the process under study has come to exist in space (Curriero, 2007). A good example is samples collected in stream networks where pathways connecting sampling locations are limited to in-stream paths. There is much recent research on the use of in-water distance measures honoring boundaries and flow patterns for kriging in stream networks and estuaries (e.g., Little et al., 1997; Rathbun, 1998; Gardner et al., 2003; Jensen et al., 2006; Peterson et al., 2007; Lyon et al., 2008). Recent work by Skøien et al. (2006) and Skøien and Blöschl (2007) provides a method (Top-kriging) which takes both the area and the nested nature of catchments into account to estimate streamflow-related variables in ungauged catchments. This concept focuses on manipulation of the semivariogram estimate and builds upon the early work of Gottschalk (1993a, 1993b) with extension by Sauquet et al. (2000) developing a method for calculating covariance along a river network to interpolate along the network. Directional trees corresponding to drainage network structure (i.e., channel width) have also been used to modify the geostatistical framework (Monestiez et al., 2005; Bailly et al., 2006). Monestiez et al. (1989) and Audergon et al. (1993) modeled variables measured on trees using isotropic models based on ultrametric
distances associated with hierarchical clustering on trees. Chokmani and Ouarda (2004) used a physiographical space-based kriging method incorporating physiographical and meteorological characteristics of stream gauging stations with multivariate analysis techniques to modify in-stream distance. For landscapes, Cressie et al., (1990) and Banerjee (2005) allowed for a spherical Earth by using geodetic distances on the Earth’s surface. Distance metrics derived from travel times are also possible (e.g., Krivoruchko and Gribov, 2004). Christakos (2000) and Christakos et al. (2000) focused on the structure of the spatiotemporal continuum to propose several possible metric definitions for distance usable in geostatistics. In general, while the goal is to characterize spatial dependence as best as possible, it may prove beneficial to consider possible non-Euclidean distances to describe proximity relationships among spatial data (Curriero, 2007).

By using a traditional, Euclidean distance metric, for geostatistical applications it is assumed that proximity of sampling locations will capture spatial autocorrelation to simulate processes in nature controlling the variable of interest. However, sampling at the appropriate scale (i.e., Skøien and Blöschl, 2006) to reflect processes may not be possible due to lack of time and money. It may be possible to reproduce variability between sampling locations at scales smaller than sampling distances by using secondary data which are available at high resolution (e.g., remotely sensed data) to adjust the distance used in geostatistical methods (e.g., semivariogram development and kriging interpolation). In this study, we focus on how distance between sampling locations is represented in a kriging interpolation and present a new approach incorporating secondary information to adjust distances between sampling points. This new adjusted distance metric builds on the conceptualization outlined by Lyon et al. (2008) and allows for an extension into the landscape.
2. Materials and Methods

2.1 Site description and data

Several properties were measured and landscape characteristics determined (Table 1) at the 20 km$^2$ Ovanmyra field site located in central Sweden (Figure 1). This site was located in a boreal coniferous forested landscape with Podzol soils mixed with some Histosols consisting of deep organic layers. There was clear differentiation between the organic soil layer and the mineral soil layer at Ovanmyra. A 5-meter resolution digital elevation model (DEM) was available for this region from light detection and ranging (LIDAR) data. At 100 sampling locations, composite soil samples were collected for both the organic and mineral soil layers from 5 sub-samples taken in each layer. From these composite samples, soil pH and C:N ratio were determined. The soil pH for both soil layers was measured in a water-soil suspension.

The total amounts of carbon and nitrogen for both layers were measured using a CNS-1000 analyzer (LECO Corporation, 2005) and C:N ratios were calculated for each sampling location. In addition, the soil moisture content was averaged from 5 measurements at each sampling location using a TRIME FM3 time domain reflectometry (TDR) probe (IMKO, 2004) with 16 cm long rods for the mineral soil layer and a ThetaProbe soil moisture sensor type ML2x (Delta-T Devices, 1999) with 6 cm long rods for the organic soil layer.

Landscape characteristics in this study were selected based on data availability and a basic understanding of the hydrology of the region. The topographic wetness index (TWI) (Beven and Kirkby, 1979) was computed from the DEM using a multiple flow-direction algorithm described in Seibert and McGlynn (2007). For this study, the downslope index of Hjerdt et al. (2004) was used instead of the traditional local slope. This index is based on the distance to be traveled downslope before a defined decrease in elevation is achieved. This vertical decrease was set to 5 meters for Ovanmyra as in Sørensen and Seibert (2007). Differing slightly from
local slope, the downslope index incorporates the occurrence of toe slopes and other
landscape characteristics into TWI calculations (Hjerdt et al., 2004). To fully investigate the
role of topographic characteristics both the downslope index value and the natural logarithm
of the upslope contributing area (i.e., the two main constituents of a TWI) were considered
separately as landscape characteristics. In addition to these topographic characteristics for the
Ovanmyra field site, the volume (per unit area) and age of the various forest stands covering
the area were included as landscape characteristics considered in defining an adjusted distance
metric. On average, each forest stand covers an area of about 1.2 ha.

2.2 Developing the adjusted distance metric

Consider two separate points \((i\) and \(j\)) in a spatial domain. Based on a traditional Euclidean
distance metric, the points are separated by a distance defined simply by a straight-line path
\((d_{ij})\) based solely on the coordinates of the points. Lyon et al. (2008) outlined how the
landscape characteristics of the contributing area to positions in a stream network can be used
to define attributes associated with that position in the stream network. Similarly, on the
landscape itself, points \(i\) and \(j\) also have certain local landscape characteristics, such as slope
or upslope accumulated area, associated with their position in the landscape. These landscape
characteristics for each point \((a_i\) and \(a_j\), respectively) can be used to define how similar or
different two positions in the domain are by the absolute difference in attribute \((a_{ij})\):

\[
a_{ij} = \left| a_i - a_j \right|
\]  

(2)

For an adjusted distance metric to incorporate information about both Euclidean distance and
landscape characteristics, it is necessary to use some combination of \(d_{ij}\) and \(a_{ij}\). Starting from
an approach similar to Lyon et al (2008), we can define an adjusted distance metric, \(h_{ij}\), in the
landscape. For given positions, this approach combines closeness in terms of traditional
distance \((d_{ij})\) and similarity in terms of a local landscape characteristic \((a_{ij})\) using a linear weighting:

\[
h_{ij} = (1 - \omega) \frac{d_{ij}}{d_{\text{median}}} + \omega \frac{a_{ij}}{a_{\text{median}}}
\]  

where \(\omega\) is a weighting factor varying from 0 to 1 and \(d_{\text{median}}\) and \(a_{\text{median}}\) are the median of all \(d_{ij}\) and \(a_{ij}\), respectively, for all pairs of sample points collected in a given spatial domain (i.e., all the points in a field campaign). The weighting factor allows us to adjust the relative importance of the physical distance between points and their similarity/dissimilarity. For \(\omega\) equal 0, the adjusted distance equals the traditional Euclidean distance between two points scaled by the median of all distance pairs. With a small value for \(\omega\), the traditional Euclidean distance between two points dominates, whereas with higher values the adjusted distance becomes more dominated by the differences of the landscape characteristic.

To be considered a true distance metric, the adjusted distance metric must satisfy the triangle inequality as outlined in Rathbun (1998). For the above mentioned adjusted distance metric (Eq. 3), this is obvious for the end-member cases where \(\omega\) equal 0 or 1 (the former case being a Euclidean metric and the latter being an attribute-space metric). For all other values of \(\omega\) considered in this study, the adjusted distance metric is a simple linear combination of these two valid distance metrics. It can be shown that the adjusted distance metric satisfies the triangle inequality for all values of \(\omega\) and, thus, conforms to the definition of a distance metric (see Appendix 1). Other formulations are possible for combining \(d_{ij}\) and \(a_{ij}\) than Eq. 3 such as non-linear combinations, but in this case caution needs to be taken that the resultant metric is indeed a valid distance metric (i.e., it satisfies the triangle inequality).
The linear combination in Eq. 3 can easily be generalized to consider more than one landscape characteristic as a linear combination. This could be formulated by defining an attribute distance $a_{n,ij}$ for each characteristic $n$ of the total set of characteristics $m$ considered using a formulation such as:

$$h_{ij} = (1 - \sum_{n=1}^{m} \omega_n) \frac{d_{ij}}{d_{median}} + \sum_{n=1}^{m} \omega_n \frac{a_{n,ij}}{a_{n,median}}$$  \hspace{1cm} (4)$$

where $\omega_n$ are weighting factors for each different attribute considered varying from 0 to 1 such that the sum of all $\omega_n$ is within the range 0 to 1 and distance $a_{n,median}$ is the median the attribute determined at all pairs of sample points in a given spatial domain.

The goal of this current study, however, was to investigate the merits of including individual landscape characteristics in defining a distance metric for geostatistical analysis within the landscape. While there are conceptual similarities between this current study and the previous work by Lyon et al (2008), it should be noted that the current study considers the local landscape position influence on how samples collected in the landscape relate to each other. This differs from Lyon et al. (2008) where the focus is on adjusting path-restricted, in-stream distances between stream water sampling positions based on the average composition of the drainage area to a given sampling point. The current study interpolates point observations of different properties in both the mineral and organic soil layers at the Ovanmyra field site in central Sweden (Figure 1) using both a traditional distance metric and the adjusted distance metric (Eq. 3). As the adjusted distance metric provides a method to incorporate secondary data into kriging interpolations, we also compare these results to KED interpolations. We evaluated the resulting interpolations by computing the cross-validation error associated with each method.
2.3 Geostatistical analysis using the adjusted distance metric

Each landscape characteristic (Table 1) can be used to define several unique adjusted distance metrics using Eq. 3 for different weighting values ($\omega$). Research is needed on how to define such a weighting value *a priori*; however, we can take a methodical approach to identify the best possible weighting value for a given landscape characteristic based on a given set of observations. This is similar to the method used in Lyon et al. (2008). We tested each of the five landscape characteristic with values for $w$ varying from zero to one in steps of 0.1 (here, we let $\omega$ step by increments of 0.1 to limit the number of possibilities). Each of these adjusted distance metrics was then used to produce a sample semivariogram from Eq. 1 to be used in a kriging interpolation. Sample semivariograms were fitted with an exponential semivariogram model of the form

$$
\gamma_e(h) = \sigma_0^2 + (\sigma_\omega^2 - \sigma_0^2) \left(1 - e^{-\frac{h}{\lambda}}\right)
$$

(5)

where $\gamma_e(h)$ is the fitted semivariogram model, $\sigma_0^2$ is the nugget, $\sigma_\omega^2$ is the sill and $\lambda$ is the correlation length.

A currently unresolved problem with using a landscape-based distance metric such as the adjusted distance metric for kriging is assuring the validity of the covariance matrix generated using the selected semivariogram model (Rathbun, 1998; Ver Hoef et al., 2006). While criteria for consistently valid combinations of semivariogram models and the adjusted distance metric are yet to be determined, possible candidate semivariogram models could be tested and rejected if they fail to meet the positive definite criterion. For the adjusted distance metric, the exponential semivariogram model is positive definite under the conditions considered in this study. This was checked using the simple test that the determinant of the correlation matrix was positive for the locations in this study (Rathbun, 1998). The validity of
other semivariogram models (e.g., Gaussian, spherical) is an unsolved problem and we have, thus, limited ourselves to an exponential function. The exponential semivariogram models were fit using an automated fitting procedure (Cressie, 1985; Cressie, 1991). Correlation lengths were divided by the maximum bin distance to scale distance making the traditional (see following section) and adjusted semivariograms comparable.

The fitted semivariogram models defined using the adjusted distance metrics were then used to perform OK interpolations for all observed properties. Note that the adjusted distance metric is used not only in the creation of the semivariogram model, but also in the distance definitions of the following kriging interpolations. To quantify the ability of each kriging interpolation to estimate the actual observed properties, we used a ‘leave-one-out’ cross validation methodology. This methodology omits a sampling location from the analysis and then estimates its value using the remaining sampling locations. After repeating for all sampling locations, a cross-validation error ($K_{RMSE}$) was then calculated as the root mean squared error from the differences between estimates and actual observed properties as

$$K_{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(E_i - Y_i)^2}{n}}$$

where $Y_i$ is the observed property at a given location, $E_i$ is the kriging estimated property at a given location, and $n$ is the number of sampling locations or the total number of samples.

We computed the leave-one-out cross validation error ($K_{RMSE}$) for each kriging interpolation based on each adjusted distance metric using combinations of landscape characteristic and $\omega$. Once the $K_{RMSE}$ for each adjusted distance metric was computed, we could identify the adjusted distance metric (i.e., the combination of landscape characteristic and $\omega$) that minimized $K_{RMSE}$. The combination of landscape characteristic and $\omega$ that resulted in the
lowest leave-one-out cross-validation $K_{\text{RMSE}}$ for each property was selected as the best performing adjusted distance metric for interpolating that particular property. This is similar to the methodology outlined by Lyon et al. (2008).

2.4 Comparison with traditional geostatistical methods

For comparison, OK interpolations and cross-validation were performed for each property using a traditional distance metric. For these interpolations, the sample semivariogram (Eq. 1) was modeled using the exponential semivariogram model (Eq. 5) with the fit optimized using the same automated fitting procedure (Cressie, 1985; Cressie, 1991) as above. It should be noted that it is possible to model a sample semivariogram by optimizing the cross-validation error results instead of optimizing the fit to the sample semivariogram. The cross-validation error (Eq. 6) was used to evaluate how well the OK interpolations fit to observations.

We also used KED (Goovaerts, 1997) to create an interpolation of each of the measured properties that incorporates secondary data. KED is comparable to universal kriging (kriging with a trend) and is based on an assumption that we know the shape of the trend. The kriging is then performed on the residuals while the trend parameters are implicitly estimated. In universal kriging, the trend is often a function coordinates while in KED the trend surface is based on a trend through the secondary data. For comparison purposes, we used the landscape characteristic associated with the best performing adjusted distance metric for each property as the secondary data in our KED analysis and use a linear regression to define the trend such that

$$Y_i = \alpha_0 + \alpha_1 A_i$$

where $Y_i$ is the observed values of the property at a given location and $A_i$ is the values of the secondary data at the same location. The coefficients ($\alpha_0$ and $\alpha_1$) for the trend are implicitly
estimated within a search neighborhood containing the 30 closest neighboring points. KED is
then performed using a fitted semivariogram model (Eq. 5) determined from a sample
semivariogram (Eq. 1) of the residuals of the observations from the trend. Cross-validation
error (Eq. 6) was used to evaluate how well the KED interpolations fit to the observed
properties.

3. Results

For each of the measured properties, the best performing adjusted distance metric was defined
based on a combination of landscape characteristic and $\omega$ (Table 2). Using these best
performing metrics, semivariograms were created based on both traditional and adjusted
distance metrics (Figure 2). Several of the semivariograms created using a traditional distance
metric had a poor structure with a lack of reduction in semivariance at smaller lag distances.
For the mineral soil C:N ratio, organic soil pH and C:N ratio at Ovanmyra, the
semivariograms based on traditional distance were best modeled with a pure nugget model
indicating no spatial structure in these properties at this sampling resolution. Using the
adjusted distance metric, the nuggets of the semivariogram models were reduced for all
properties except for mineral and organic soil moisture content. The properties modeled as
pure nugget models were better modeled with the exponential semivariogram model (Eq. 5)
when distance between sampling locations was defined using the best performing adjusted
distance metric. Sills for all properties stayed relatively the same when comparing the
traditional and adjusted distance metrics. This was expected because the overall variance in
the measured properties does not change with change in distance metric.

To visualize the influence of the adjusted distance metric, OK was conducted using both
traditional and adjusted distance metrics for Ovanmyra for the properties measured in both the
mineral (Figure 3) and organic (Figure 4) soil layers. KED was also performed using the landscape characteristic associated with the best performing adjusted distance metric for each property as the secondary data. Using traditional distance definitions, the OK interpolation showed islands occurring due to large nugget effects caused by the spatial scale of sampling. When the OK was based on the adjusted distance metric or when KED was used, the interpolations represented more the underpinning secondary data. For example, patterns of topographically convergent areas became visible for the interpolations of moisture content in the mineral soil layer. In general there were visible similarities between the KED maps and the OK maps based on the adjusted distance metrics. Summary statistics for each interpolation are provided in Table 3.

The ability of each interpolation in Figures 3 and 4 to fit the observed data was quantified using $K_{RMSE}$ (Table 4). The $K_{RMSE}$ was lowest using OK based on adjusted distance metrics compared to both OK using traditional distance and KED. Note that other criteria could be considered for comparison of the different methods and also to determine the ‘best’ performing adjusted distance metric. These include, for example, unbiasedness measured by the mean error or correlation between prediction errors and kriging variance quantified using mean square standardized residual (e.g., Wackernagel, 1998).

Kriging is a spatial predictor for dependant observations including prediction uncertainty. To give further visualization of the impact of the adjusted distance metric, we have mapped the kriging variance (e.g., Isaaks and Srivastava, 1989) for OK conducted using adjusted distance metrics for Ovanmyra for the properties measured in both the mineral and organic soil layers (Figure 5).
4. Discussion

The first step in the kriging of any dataset is the finding a sample semivariogram and fitting a theoretical model to it. Using the adjusted distance metric, semivariograms captured small-scale spatial structure better for most of the observed properties, which is indicated by decreases in the modeled nugget values (Table 2). Some of this reduction in modeled nugget may be attributed to the automatic fitting procedures used in this study. In addition, several pure nugget semivariograms could be transferred into semivariograms with increasing semivariance with distance modeled using exponential semivariograms (Figure 2, Table 2). In the case of pure nugget models, there is no advantage to applying kriging as an interpolation method over simple averaging as there is complete lack of spatial correlation at the spatial scale which the samples were collected. Changing pure nugget semivariograms to exponential semivariograms directly demonstrates the ability of the adjusted distance metric to incorporate variability at scales smaller than the sample spacing based on Euclidian distance.

More strikingly, the effect of the adjusted distance metrics is seen in a visual comparison of the maps generated by the different kriging interpolations (Figure 3 and Figure 4). Using the same sets of observations, different interpolation realizations are possible. By adjusting distance between sampling locations, small-scale variations in secondary data can be represented by incorporating landscape characteristics. Visually, this resulted in extremely different maps compared to those based on a traditional distance metric. While this is a subjective assessment, we would argue that the interpolated maps exhibiting significant small-scale variability look more realistic. In addition to ‘looking’ more realistic, the reduced cross-validation error using the adjusted distance quantifies that the interpolations better match the observed data. With respect to the kriging variance (Figure 5), which provides an estimation of the uncertainty associated with sampling, the spatial variations in kriging variance reflect
the spatial structure of the attribute used in the adjusted distance metric. This agrees with the concept that samples collected in regions of similar landscape characteristics will be related.

Interestingly, while the interpolations made using OK based on the adjusted distance metric look similar to those made using KED, there is a measured difference in the cross-validation error. In addition, KED results in, physically unreasonable, negative estimates for some locations (Table 3). For all properties in this study, interpolations using an adjusted distance metric lead to the lowest cross-validation errors while KED interpolations increased cross-validation errors relative to OK interpolations (Table 4). This identifies a possible strength of the adjusted distance metric and the methodology presented in the study. Typically, to incorporate secondary data into a kriging interpolation, there needs to be some correlation between the primary data being interpolated and the secondary data. Without an a priori understanding of the processes by which the secondary data control the spatial distribution of the primary data, we are limited to empirical, site-specific relations between the primary and the available secondary data. In some cases, it may be possible to explicitly correlate secondary data to the spatial distribution of samples, but in other cases this may be very difficult. The methodology presented in this study to identify the best performing adjusted distance metric does not require an explicit correlation between the primary and secondary data. In addition, we are not limited by incomplete process understanding as we allow the observed data to identify the best performing secondary data to improve interpolations. The interpolation of primary data can, of course, be improved upon with increased process understanding and subsequent collection of additional or more relevant secondary data. In real-world applications, however, it is often difficult to improve process understanding or gather such addition spatially continuous data.
When a sparse spatial sampling campaign is used to measure a given property, which is often the case given the limited resources available in environmental sciences, extreme variations in the properties can be measured at only a few sample points. This can especially be the case when the locations of sampling points differ in respect to their position in the landscape. Extreme variations in, for example, topography can also represent a partial or complete barrier of the intervening space. The use of traditional kriging is therefore limited in situations of complex terrain where the processes controlling the variable of interest are themselves complex (McBratney et al., 2000). If we use only traditional Euclidian distance, the physical properties of how the process under study has come to exist in space (Curriero, 2007) may not be properly represented resulting in kriging interpolations that are unrealistic. By incorporating landscape characteristics into the definition of a distance metric we were able to increase the information available for interpolation between sampling locations. In this way fewer sampling locations are required to obtain interpolations which show a realistic small-scale variability. There are, thus, both economic and logistic reasons for including secondary data in interpolations especially if the latter are more readily and cheaply available (McBratney et al., 2000).

In our test case, the adjusted distance approach resulted in smaller cross-validation errors than the KED method. A common critique of KED and other kriging with regression techniques is that they are complicated relative to OK (Hengl et al., 2007). In addition, these techniques often lack in off-the-shelf computation geostatistical software. The adjusted distance metric allows incorporation of secondary data in the definition of distance and use standard geostatistical mathematics to derive semivariograms and perform kriging interpolations. Thus, another advantage might be that the adjusted distance approach is conceptually simple and, thus, easy to implement in other studies.
5. Concluding remarks

A new metric for defining distances between point observations has been presented. This metric produces an adjusted distance by combining traditionally defined distance with landscape characteristics. Locations that have similar landscape characteristics and are spatially close influence each other more in semivariograms and kriging interpolations. The kriging interpolations based on the adjusted distance metric were able to reproduce variability occurring at scales smaller than sampled by including the influence of landscape properties.

6. Acknowledgements

This research was made possible with partial funding from the Swedish Research Council (grants 620-20001065/2001) and the Bert Bolin Centre for Climate Research, which is supported by a Linnaeus grant from VR and The Swedish Research Council Formas. The authors also thank the American-Scandinavian Foundation for funding that allowed for the initial collaboration.
References


Appendix 1: Triangle inequality

As given in Rathbun (1998), before adopting a distance metric, it must be satisfied that the metric fulfills the triangle inequality such that the length of a given side of a triangle measured with the metric must be less than the sum of the other two sides of the triangle but greater than the difference between the two sides. A brief proof is provided in this appendix with full proof given as supplemental material. With respect to the triangle inequality as it applies to Eq. 3 and the definition of an adjusted distance metric, we consider three cases:

(1) When $\omega=0$, Eq. 3 in the main text reduces simply to a scaled Euclidean linear distance. This is a trivial case with respect to triangle inequality.

(2) When $\omega=1$, Eq. 3 represents the scaled difference in absolute difference in attribute defined in Eq. 2. This is a definition of distance totally in the attribute space. Consider three separate points in the study region with given quantified attributes such that $n$ is the quantifiable attribute and $\alpha$ and $\beta$ are any real numbers so the quantified attribute value at each of the three points is defined as $n$, $n+\alpha$, and $n+\beta$. The basic triangle inequality can be constructed for these three points as:

$$\left| (n + \beta) - n \right| \leq \left| (n + \alpha) - n \right| + \left| (n + \beta) - (n + \alpha) \right|$$

(A1)

With three cases to evaluate:

(i) $\omega=\beta$. Here, both sides of Eq. A1 reduce to $\omega=\beta$ which is true by definition.

(ii) $\omega>\beta$. Eq. A1 can be shown to reduce to $|\beta|-|\alpha|<|\beta-\alpha|$ such that the inequality holds.

(iii) $\omega<\beta$. Here, Eq. A1 reduces to $|\beta|-|\alpha|\leq|\beta-\alpha|$ with $|\beta|-|\alpha|=|\beta-\alpha|$ by definition with $\omega<\beta$ such that the inequality holds.
(3) When $0<\omega<1$, we have a scenario of a linear combination of two valid distance metrics. It can be shown straight away that the triangle inequality holds for this case.

As such, it is shown that the distance metric defined using Eq. 3 satisfies the triangle inequality for all values of $\omega$ between 0 and 1 by definition.
Table 1: Summary statistics for the observed properties in both the mineral and organic soil layers at Ovanmyra and the landscape characteristics considered in this study.

<table>
<thead>
<tr>
<th>Property</th>
<th>Units</th>
<th>Average</th>
<th>Standard Deviation</th>
<th>Max</th>
<th>Min</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mineral Soil Layer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pH</td>
<td>-</td>
<td>4.41</td>
<td>0.34</td>
<td>6.17</td>
<td>3.34</td>
</tr>
<tr>
<td>Moisture Content %</td>
<td>%</td>
<td>30.5</td>
<td>23.0</td>
<td>100.0</td>
<td>10.4</td>
</tr>
<tr>
<td>C:N Ratio</td>
<td>-</td>
<td>17.2</td>
<td>4.9</td>
<td>36.8</td>
<td>7.2</td>
</tr>
<tr>
<td>Organic Soil Layer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pH</td>
<td>-</td>
<td>3.77</td>
<td>0.28</td>
<td>5.70</td>
<td>3.82</td>
</tr>
<tr>
<td>Moisture Content %</td>
<td>%</td>
<td>34.8</td>
<td>16.8</td>
<td>100.0</td>
<td>10.6</td>
</tr>
<tr>
<td>C:N Ratio</td>
<td>-</td>
<td>30.2</td>
<td>5.6</td>
<td>50.1</td>
<td>16.3</td>
</tr>
<tr>
<td>Landscape Characteristic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TWI</td>
<td>ln(m)</td>
<td>7.96</td>
<td>1.70</td>
<td>12.36</td>
<td>4.30</td>
</tr>
<tr>
<td>Slope</td>
<td>degree</td>
<td>0.06</td>
<td>0.05</td>
<td>0.35</td>
<td>0.01</td>
</tr>
<tr>
<td>Ln[Upslope Area]</td>
<td>ln(m²)</td>
<td>5.82</td>
<td>1.40</td>
<td>9.84</td>
<td>3.22</td>
</tr>
<tr>
<td>Forest Volume</td>
<td>m³/ha</td>
<td>138</td>
<td>87</td>
<td>305</td>
<td>0</td>
</tr>
<tr>
<td>Forest Age</td>
<td>years</td>
<td>53</td>
<td>29</td>
<td>100</td>
<td>1</td>
</tr>
</tbody>
</table>
Table 2: Best performing adjusted distance metric for each property along with the parameters from exponential model fitted to semivariograms based on both traditional and adjusted distance metrics.

<table>
<thead>
<tr>
<th>Property</th>
<th>Landscape Characteristic</th>
<th>Best performing adjusted distance metric</th>
<th>Semivariogram model using traditional distance metric</th>
<th>Semivariogram model using adjusted distance metric</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>ω</td>
<td>Nugget</td>
<td>Sill</td>
</tr>
<tr>
<td>Mineral Soil Layer</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pH</td>
<td>Slope</td>
<td>0.5</td>
<td>0.07</td>
<td>0.14</td>
</tr>
<tr>
<td>Moisture Content</td>
<td>TWI</td>
<td>1.0</td>
<td>44.5</td>
<td>633.6</td>
</tr>
<tr>
<td>C:N Ratio</td>
<td>Forest volume</td>
<td>0.2</td>
<td>20.2</td>
<td>20.2</td>
</tr>
<tr>
<td>Organic Soil Layer</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pH</td>
<td>Ln(Area)</td>
<td>0.3</td>
<td>0.09</td>
<td>0.09</td>
</tr>
<tr>
<td>Moisture Content</td>
<td>Forest volume</td>
<td>0.3</td>
<td>75.4</td>
<td>319.7</td>
</tr>
<tr>
<td>C:N Ratio</td>
<td>Ln(Area)</td>
<td>0.4</td>
<td>28.2</td>
<td>28.2</td>
</tr>
</tbody>
</table>
Table 3: Summary statistics for the interpolations in both the mineral and organic soil layers at Ovanmyra for OK using a traditional distance metric and the adjusted distance metric and for KED considered in this study.

<table>
<thead>
<tr>
<th>Property</th>
<th>Units</th>
<th>OK w/ Traditional Distance Metric</th>
<th>OK w/ Adjusted Distance Metric</th>
<th>KED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mineral Soil Layer</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pH</td>
<td></td>
<td>4.44</td>
<td>0.14</td>
<td>5.16</td>
</tr>
<tr>
<td>Moisture Content</td>
<td>%</td>
<td>33.3</td>
<td>13.8</td>
<td>90.8</td>
</tr>
<tr>
<td>C:N Ratio</td>
<td></td>
<td>17.2</td>
<td>1.6</td>
<td>23.9</td>
</tr>
<tr>
<td>Organic Soil Layer</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pH</td>
<td></td>
<td>3.75</td>
<td>0.10</td>
<td>4.45</td>
</tr>
<tr>
<td>Moisture Content</td>
<td>%</td>
<td>37.2</td>
<td>9.3</td>
<td>89.2</td>
</tr>
<tr>
<td>C:N Ratio</td>
<td></td>
<td>30.3</td>
<td>2.2</td>
<td>39.0</td>
</tr>
</tbody>
</table>
Table 4: Cross validation error ($K_{\text{RMSE}}$) for OK using a traditional distance metric and the adjusted distance metric and for KED.

<table>
<thead>
<tr>
<th>Property</th>
<th>OK w/ Traditional Distance Metric</th>
<th>OK w/ Adjusted Distance Metric</th>
<th>KED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mineral Soil Layer</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pH</td>
<td>0.32</td>
<td>0.30</td>
<td>0.34</td>
</tr>
<tr>
<td>Moisture Content</td>
<td>25.7</td>
<td>21.9</td>
<td>30.2</td>
</tr>
<tr>
<td>C:N Ratio</td>
<td>5.0</td>
<td>4.8</td>
<td>5.4</td>
</tr>
<tr>
<td>Organic Soil Layer</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pH</td>
<td>0.28</td>
<td>0.27</td>
<td>0.30</td>
</tr>
<tr>
<td>Moisture Content</td>
<td>16.2</td>
<td>15.1</td>
<td>16.2</td>
</tr>
<tr>
<td>C:N Ratio</td>
<td>5.7</td>
<td>5.4</td>
<td>5.8</td>
</tr>
</tbody>
</table>
**Figure Captions**

Figure 1: Location of the Ovanmyra field site in central Sweden. White dots on the field site maps show sampling locations used in this study.

Figure 2: Semivariograms and exponential models using both traditional and adjusted distance metrics at the Ovanmyra field site. For the mineral soil layer, a traditional distance metric was used to create semivariograms for pH (A), moisture content (C), and C:N ratio (E) while the best performing adjusted distance metric listed in Table 2 was used to created semivariograms for pH (B), moisture content (D), and C:N ratio (F). Similarly, for the organic soil layer, a traditional distance metric was used to create semivariograms for pH (G), moisture content (I), and C:N ratio (K) while the best performing adjusted distance metric listed in Table 2 was used to created semivariograms for pH (H), moisture content (J), and C:N ratio (L).

Figure 3: Kriging interpolations for the mineral soil layer at the Ovanmyra field site. OK with traditional distance metric was used for kriging pH (A), moisture content (D), and C:N ratio (G). OK with best performing adjusted distance metric listed in Table 2 was used for kriging pH (B), moisture content (E), and C:N ratio (H). KED was used for kriging pH (C), moisture content (F), and C:N ratio (I).

Figure 4: Kriging interpolations for the organic soil layer at the Ovanmyra field site. OK with traditional distance metric was used for kriging pH (A), moisture content (D), and C:N ratio (G). OK with best performing adjusted distance metric listed in Table 2 was used for kriging pH (B), moisture content (E), and C:N ratio (H). KED was used for kriging pH (C), moisture content (F), and C:N ratio (I).

Figure 5: Kriging variance for the OK interpolations using the adjusted distance metric for the mineral soil layer values of pH (A), moisture content (C), and C:N ratio (E) and organic soil layer values of pH (B), moisture content (D), and C:N ratio (F).
MINERAL SOIL LAYER

Traditional Distance Metric

Adjusted Distance Metric

Traditional Distance Metric

Adjusted Distance Metric

pH

Semivariance [^2]

Moisture Content [%]

Semivariance [^2]

C:N Ratio

Semivariance [^2]

Scaled Distance

Scaled Distance

Scaled Distance

Scaled Distance
Supplemental Material
With respect to the triangle inequality as it applies to Eq. 3 and our definition of an adjusted distance metric, we consider three cases:

(1) When \( \omega = 0 \), Eq. 3 in the main text reduces simply to a scaled Euclidean linear distance. This is a trivial case with respect to triangle inequality.

(2) When \( \omega = 1 \), Eq. 3 represents scaled difference in absolute difference in attribute defined in Eq. 2. This is a definition of distance totally in the attribute space. A basic proof of the triangular inequality for this space follows:

Consider three points in the study region with given quantified attributes such that \( n \) is the quantifiable attribute and \( \alpha \) and \( \beta \) are any real numbers:

The attribute distance between each point can be defined using absolute difference in attribute defined in Eq. 2 of the original manuscript. It is trivial to show that all distances are greater than or equal to zero (by definition of the absolute value operator) and that the distance from a point with attribute \( n \) to an arbitrary point with attribute \( n + \alpha \) is equal to the reversed distance from \( n + \alpha \) to \( n \) (again, by definition of the absolute value operator). This leaves the basic triangle inequality that needs to be satisfied (referencing the above figure) as:

\[
\frac{|(n+\beta) - n|}{|n+\alpha|} \leq \frac{|(n+\alpha) - n|}{|n+\alpha|} + \frac{|(n+\beta) - (n+\alpha)|}{|n+\alpha|} \quad (i)
\]

There are three cases to evaluate:

(1) \( \alpha = \beta \) (all three points have the same attribute value)
Here, both sides of Eq. (i) reduce and we are left with \( \alpha = \beta \) which is true by definition.

(2) \( \alpha > \beta \) (travel through a point with higher attribute value)
Eq. (i) becomes:

\[
|\beta| \leq |\alpha| + |\beta - \alpha|
\]

note that \( |\beta| - |\alpha| < 0 \) while \( |\beta - \alpha| > 0 \)
this gives \( |\beta| - |\alpha| < |\beta - \alpha| \) so the inequality holds.

(3) \( \alpha < \beta \) (travel through a point with lower attribute value)
Again, Eq. (i) becomes:

$$|\beta| \leq |\alpha| + |\beta - \alpha|$$

by definition $$|\beta| - |\alpha| = \beta - \alpha$$

also by definition $$\beta - \alpha = |\beta - \alpha|$$ in this case where $$\alpha < \beta$$

thus, the inequality holds.

The above simple proof shows that when $$\omega=1$$, the resulting scaled difference in absolute difference in attribute defines a valid distance metric.

(3) When $$0<\omega<1$$, we have a scenario with a linear combination of two valid distance metrics. Considering the following triangle setup of positions $$i, j,$$ and $$k$$ separated by scaled distances $$d_{ij}, d_{jk},$$ and $$d_{ik}$$ and by attribute distances $$a_{ij}, a_{jk},$$ and $$a_{ik}$$:

We setup the following triangle inequality:

$$\omega a_{ik} + (1-\omega)d_{ik} \leq (\omega a_{ij} + (1-\omega)d_{ij}) + (\omega a_{jk} + (1-\omega)d_{jk})$$

(ii)

With some straightforward manipulation of Eq. (ii)

$$\omega a_{ik} + (1-\omega)d_{ik} \leq (\omega a_{ij} + \omega a_{jk}) + ((1-\omega)d_{ij} + (1-\omega)d_{jk})$$

from the previous two cases we know

$$a_{ik} \leq a_{ij} + a_{jk} \Rightarrow \omega a_{ik} \leq \omega a_{ij} + \omega a_{jk}$$

$$d_{ik} \leq d_{ij} + d_{jk} \Rightarrow (1-\omega)d_{ik} \leq (1-\omega)d_{ij} + (1-\omega)d_{jk}$$

it follows that

$$\omega a_{ik} + (1-\omega)d_{ik} \leq (\omega a_{ij} + \omega a_{jk}) + ((1-\omega)d_{ij} + (1-\omega)d_{jk})$$

This creates a valid distance metric for all values of $$\omega$$ between 0 and 1 by definition.