Relating crop yield to topographic attributes using spatial analysis neural networks and regression

Timothy R. Green
Jose D Salas, Colorado State University
Ana Martinez, Colorado State University
Robert H Erskine

Available at: https://works.bepress.com/green/16/
Relating crop yield to topographic attributes using Spatial Analysis Neural Networks and regression

Timothy R. Greena,⁎, Jose D. Salasb, Ana Martinezb, Robert H. Erskinea

a USDA-ARS, Agricultural Systems Research Unit, Fort Collins, CO 80526, USA
b Colorado State University, Department of Civil Engineering, Fort Collins, CO 80526, USA

Received 17 January 2006; received in revised form 30 September 2006; accepted 15 December 2006
Available online 21 February 2007

Abstract

Land-surface topographic attributes can be useful for estimating stable spatial patterns of crop yield caused by spatial variability in soils and water availability. We present spatial analyses of grain yield for three fields of dryland winter wheat in northeastern Colorado using topographic attributes as explanatory variables. Topographic attributes including elevation, slope, aspect, curvature, specific contributing area, and wetness index were computed from a 10-m digital elevation model. A Spatial Analysis Neural Network (SANN) algorithm was used for joint spatial interpolation and yield prediction from the topographic attributes. SANN prediction errors were compared with the results of Multiple Linear Regression (MLR). SANN and MLR were assessed in terms of bias and relative root mean squared error (rRMSE) using validation data. SANN out-performed MLR in multivariate estimation, but not for the univariate cases. The greatest advantage of SANN was seen using four or more topographic attributes, whereas MLR showed diminishing efficiency with more than three explanatory variables. Prediction/interpolation errors within a given field were reduced substantially by using the spatial coordinates (latitude and longitude) in tandem with topographic attributes. The rRMSE value reached a minimum of 0.44 (model efficiency, E = 0.80) for interpolation with SANN on the West field. Using only topographic attributes as input, the minimum rRMSE values were 0.59 (E = 0.65) for SANN with 5 variables and 0.72 (E = 0.48) for MLR with 4 or 5 explanatory variables. Thus, this study demonstrated the utility of SANN with topographic attributes that contain implicit soil and water information for estimating spatial patterns of dryland crop yield.
© 2007 Elsevier B.V. All rights reserved.

Keywords: Topography; Crops; Soils; Regression analysis; Spatial data

1. Introduction

Geomorphologic, pedologic and hydrologic processes can cause pronounced space–time variability in soil hydraulic properties, surface and subsurface water movement, sediment erosion and deposition, nutrient availability, and water storage in the root zone. All of these factors vary spatially within agricultural fields and affect spatial patterns of crop grain yield in dryland (rain-fed) agriculture (e.g., Afyuni et al., 1993). However, such factors that control crop yield are rarely measured spatially in detail, but surface topography is often available and can serve as a surrogate for soil and agronomic factors.

Thus, high-resolution topographic attributes can be related to spatial crop yield at compatible scales.

An increasing number of investigators have quantified spatial variability in crop yield. Miller et al. (1988) and Mulla (1992) reported random variability (white noise) for distances less than 20 m and variogram nuggets greater than 30% of the total variance for wheat yields in California and Washington, respectively. Jaynes and Colvin (1997) analyzed the spatial statistics of crop data for a corn/soybean rotation in Iowa. They found that the spatial trends determined by “median polishing” (Cressie, 1991) comprised about 25% of the overall variance, the variogram ranges approached 150 m, and the nugget effect was negligible for all six years. However, the spatial patterns and structured variation changed from year to year, where water-logging in low areas was an important factor. Subsequently, Jaynes et al. (2003, 2005) used temporal cluster
analysis on six years of corn and five years of soybean yield data from transects within a 16-ha field. The resulting classifications were related to landscape topographic attributes using regression methods.

Variability in crop yield stems from nonlinear spatial interactions between numerous factors, including topographic features, precipitation, soil hydraulic properties, nutrients, organic matter and pH. Such factors may cause variability at different, nested scales that may give rise to self-similar patterns of variation described by fractal geometry (Mandelbrot, 1977; Burrough, 1981). Green and Erskine (2004) used simple fractals to identify spatial patterns in landscape topography similar to crop yield in the field sites studied here. Although there may be some changes from year to year, Green and Erskine (2004) identified temporal stability from 1997 to 1999 in the correlation structure of yield measured by variograms and fractal dimensions, even under different cropping systems. Furthermore, Andales et al. (in press) used 14 years of data from a crop rotation study on a nearby soil catena to quantify a strong effect of topographic position on both corn and wheat yield ($p < 0.0001$), but found no significant effect of crop rotation. The patterns of yield and soil water along the catena were temporally stable, as evidenced by strong relationships between slope positions. Linear regressions between slope positions explained 69–90% and 65–85% of the spatial-temporal variances in grain yields and profile soil water, respectively. Thus, computed topographic attributes in such environments should help explain and quantify spatial yield patterns.

Due to the limitations of linear regression, we explore a nonlinear method of spatial estimation called Spatial Analysis Neural Networks, SANN (Shin and Salas, 2000; Martinez et al., 2004). SANN is explicitly spatial, because it contains a kernel function that accounts for influence of neighboring points in the input space on the variable estimated at each point. Other neural network methods have been applied to a variety of estimation problems, including: crop yield (Drummond et al., 2003; Kitchen et al., 2003), soil water content (Altendorf et al., 1999), soil water retention (Pachepsky et al., 1996), soil hydraulic conductivity (Tamari et al., 1996), aquifer transmissivity (Mukhopadhyay, 1999; Shigidi and Garcia, 2003), chemical leaching (Starrett and Adams, 1997; Starrett et al., 1998), and crop diseases (Parmar et al., 1997; De Wolf and Francl, 2000). Neural network methods have been compared explicitly with least squares regression (Thai and Shewfelt, 1991; Pattie and Haas, 1996; Horimoto et al., 1997; Ramos Nino et al., 1997; Altendorf et al., 1999; Drummond et al., 2003; Kitchen et al., 2003), logistic regression (King et al., 2000; Leung and Tran, 2000), and principal component analysis (Horimoto et al., 1995, 1997). However, none of the neural network or comparative methods, other than SANN, explicitly accounts for spatial correlation.

Both multivariate linear regression (MLR) and SANN are empirical methods of multivariate correlation, which lack causative explanation for responses. Thus, the topographic attributes used as inputs or explanatory variables may be cross-correlated with other potentially causative factors. For example, it is likely that soil–water holding capacity is correlated with upslope contributing area, such that this topographic attribute contains soil information affecting crop yield. Spatial detail in soil factors is not available to identify such cross-correlation, and inference of causation is not intended here. Therefore, the present study provides a comparison of methods (SANN and MLR) for a spatial crop yield application using topographic attributes as explanatory variables. Application using direct soils data is limited only by data availability, and is left for future research.

The need for improved methods of spatial analyses remains, wherein crop yields can be estimated more accurately from topographic features within fields. Here, we address the following questions:

1. How well can one predict patterns of measured wheat grain yield from landscape topographic attributes?
2. How does the Spatial Analysis Neural Networks (SANN) method compare with Multiple Linear Regression (MLR) as a spatial predictor?
3. What advantages might SANN offer for interpolating yield values within a field?

2. Study area and methods

The agricultural fields used in this study are part of the Lindstrom farm located in eastern Colorado near the town of Sterling in Logan County (40.37 N, 103.13 W). The average pan evaporation is approximately 1000 mm per growing season, while the average annual precipitation is only 400 mm (Peterson et al., 2000). We collected and processed spatial yield data for winter wheat (*Triticum aestivum* L.) in 1997. Winter wheat has been the dominant dryland crop for much of the Great Plains. Although the findings here are directly applicable only to dryland winter wheat at this location, the same concepts should apply to other crops in a semi-arid climate like that of eastern Colorado.

2.1. Site landscape and soil properties

The terrain in northeastern Colorado is generally undulating due to aeolian deposition of predominantly silt-sized material mantling sedimentary rock (primarily sandstone) and fluvial deposits of the South Platte River Basin. Of the three fields in this study, the relief is most pronounced in the “North” field (21.3 m over 63 ha), where the elevation ranges from approximately 1361 m to 1382 m, with slopes exceeding 14%. The “South” field (72 ha) and “West” field (36 ha) have maximum relief values of 10.5 m and 6.7 m, respectively. The unconsolidated sediment and soils are relatively thick (at least 3 m) and with little or no surface expressions of groundwater or perched water in the root zone. Thin calcareous horizons have been sampled at depths of 20–50 cm, but soil horizons are otherwise not very pronounced.

The soil classes in these three fields have been mapped by the USDA Natural Resources Conservation Service (http://websoilsurvey.nrcs.usda.gov/app/). In general, the soils are all deep loams that are well drained. Table 1 shows the soil unit names and taxonomic classes with percentages of areas covered by each mapped unit. Although the soils are similar in all three
Table 1

<table>
<thead>
<tr>
<th>Soil unit name</th>
<th>Taxonomic class</th>
<th>Land area (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Region</td>
</tr>
<tr>
<td>Weld loam</td>
<td>Fine, smectitic, mesic Aridic Argiustolls</td>
<td>26.5</td>
</tr>
<tr>
<td>Rago loam</td>
<td>Fine, smectitic, mesic Pachic Argiustolls</td>
<td>16.8</td>
</tr>
<tr>
<td>Wagonwheel/Colby/S Stoneham association</td>
<td>Coarse-silty/fine-silty/fine-loamy, mixed, mesic Aridic Calciusterts/Ustorthents/ Haplustalfs</td>
<td>12.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10.2</td>
</tr>
<tr>
<td>Kuma loam</td>
<td>Fine-silty, mixed, mesic Pachic Argiustolls</td>
<td>9.8</td>
</tr>
<tr>
<td>Total area covered (%)</td>
<td></td>
<td>74.7</td>
</tr>
</tbody>
</table>

Percent of land areas are given for the five major soil units in the “Region” encompassing the three study fields: “North”, “South” and “West”.

fields (e.g., all are mesic), soils in the South and West fields are most similar to each other. Within the North field, soil hydraulic properties have been shown to vary significantly along a transect (Green et al., 2003). Explicit use of such data is left for further investigation. For the purposes of this study, soil differences between fields are not considered, but within field variability is captured to the extent to which controlling soil factors are correlated with topographic attributes.

2.2. Computing topographic attributes from digital elevation data

Elevation data were collected using a dual-frequency, real-time kinematic Global Positioning System (GPS) mounted to an all-terrain vehicle. This method of differential GPS allows for centimeter-level accuracy, both horizontally and vertically. Parallel transects, spaced 10 m apart or less, were driven and a GPS position was recorded less than every 10 m along each transect. (Note: Raw data from the North field were collected at approximately 5 m by 5 m, but data were filtered to approximately 10 m by 10 m for consistency with the other two fields. All coordinates were projected using the North American Datum (NAD) 1983 and Universal Transverse Mercator (UTM) system for Zone 13. Next, elevation data were interpolated to a 10-m by 10-m grid Digital Elevation Model (DEM), corresponding to the yield data resolution, using kriging methods in GS+ (Gamma Design Software LLC, Plainwell, Michigan, USA). A Gaussian variogram model was fit to the experimental variogram, out to a separation distance of 15 m. Cross-validation of this interpolation method resulted in root mean squared error (RMSE) in elevation values ranging 4 to 7 cm across the three fields (Erskine, 2005).

Topographic attributes can be computed efficiently using DEMs and geographic information system (GIS) software. Estimates of slope and aspect were calculated using the Spatial Analyst extension in ArcView 3.2 (Environmental Systems Research Institute, Inc., Redlands, California, USA), where aspect (\(\alpha\)) is defined as the angle between the direction of steepest slope and due north going clockwise from north. Because this representation of aspect \((0° \leq \alpha \leq 60°)\) is discontinuous, we use a cosine-transformed aspect, \(A = \cos(\alpha)\), where values vary smoothly. The cosine transformation differentiates between north and south, but not between east and west. This neglects factors associated with time-of-day effects (e.g., cloud shadowing) and directional soil formation factors (e.g., deposition or erosion associated with a prevailing wind direction). Curvature \((C)\) was estimated using a central finite-difference scheme centered on a three by three square grid window (Gerald, 1980, pp. 206–207). Specific contributing area (SCA), defined as the upslope contributing area per unit contour length, was computed outside of ArcView using TARDEM (Tarboton, 2000), a suite of FORTRAN programs for DEM analyses. The \(D_s\) method for flow direction was implemented for computing SCA (Tarboton, 1997), with and without sink filling. Finally, the SCA values were log-transformed, as per Western et al. (1999), to remove most of the extreme skewness in the frequency distribution. Fig. 1 shows two log-transformed distributions of SCA with and without sink filling, which are very similar except that the distribution with sinks filled displays a set of higher values in the upper tail. That is, the tail does not taper off monotonically. Both distributions remain skewed to the right due to truncation at a minimum value of 2.3 in this case.

The topographic wetness index (WI) is then computed from SCA and slope \((S=\tan \beta)\) as,

\[
WI = \ln \left( \frac{SCA}{S} \right) = D - \ln(S) 
\]  

(1a)

where \(D=\ln(SCA)\). Furthermore, \(D_L\) and \(D_s\) denote \(D\) computed with local sinks filled \((D_L)\) and with sinks remaining \((D_s)\). We used either \(D_L\) or \(D_s\) as inputs, but WI was computed using only \(D_L\). This was based on preliminary tests and confirmed as a reasonable approach below (i.e., yield is correlated more strongly with \(D_L\) than \(D_s\) or the two are equivalent). Thus, Eq. (1a) becomes more specifically,

\[
WI = D_L - \ln(S).
\]  

(1b)

For grid cells in which the estimated slope is less than 0.0005 m/m, we set \(S=0.0005\) m/m due to the DEM accuracy.
limitation. This potentially truncated the upper limit of WI, but
the approximation has minimal impact on the results due to the
small number of cells affected (1, 4, and 9 cells in the North,
West, and South fields, respectively). An alternative is to
remove values outside of a prescribed range, which requires a
non-matrix implementation of SANN.

Table 2 gives the basic statistics of the gridded topographic
attributes corresponding to the harvested areas on all three
fields. The North field has the highest standard deviations of
elevation, S, C, D_k, and WI, which are consistent with the
highest standard deviation in grain yield. Although the
topography in the North field is not unusual, the South and
West fields may be more representative of the region’s undu-
lating terrain. In all cases, the mean and median values of A
are negative, because more areas are facing south with milder
slopes, while the north-facing slopes tend to be steeper. Spatial
distributions of S, C, and D_k are shown in Fig. 2, along with
crop yield.

2.3. Spatial yield monitoring

The field was harvested in July 1997 using a 9-m wide
combine head. The yield was monitored with a GrainTrak™
yield monitor from MicroTrak Systems Inc. linked to a GPS
with satellite (OmniStar™) differential correction. The yield
spatial resolution is approximately 9 m by 9 m. Yield data were
calibrated to match bin weights using a weigh wagon with load
cell to within 1.0% of the bin weights, and the manufacturers
default lag time was used. The major sources of measurement
error and uncertainty are related to the load cell calibration to
actual grain flow, and transmission and storage-related dif-
ferences between the positions at which the grain was harvested
and to where the tractor had moved by the time the grain
reached the sensor. Artificially high and low values were filtered
out, where the combine either stopped (zero speed) while
continuing to thresh, or was driven over the same area twice.

Values from irregular yield data were interpolated using Kriging
onto a 10-m by 10-m grid, corresponding to the DEM reso-
lution. The interpolation errors between the raw and gridded
data were negligible (see Green and Erskine, 2004). The
resulting crop yield maps are shown in Fig. 2(a–c).

2.4. Spatial Analysis Neural Networks (SANN) algorithm

The basic concepts of the Spatial Analysis Neural Networks
(SANN) algorithm are summarized in this section. Shin and
tested the sensitivity of SANN to its internal parameters
(explained below) using grain yield data from the North field
only. The model is implemented here using MATLAB™ (The
MathWorks, Inc., Natick, MA, USA) scripts (.m files).

Suppose a spatial variable v(x) is measured at N locations
in a 2-dimensional domain, i.e., [x_n | n = 1, …, N], where x=
[x_1, x_2] represent the geographical coordinates. We want to
estimate the value of the variable v at any arbitrary location x.
The optimal estimator of the random variable v(x) given x is
the conditional expectation given by (Bishop, 1995):

\[ E[v(x)|x] = \frac{\int_{-\infty}^{\infty} v(x)p(x,v)dv}{\int_{-\infty}^{\infty} p(x,v)dv} \]  (2)

where p(x,v) is the joint probability density function of x and
v. This function may be defined by using a kernel density
function such as the multivariate Gaussian kernel density
estimator (Specht, 1991):

\[ p(x,v) = \frac{1}{(2\pi)^{N/2}} \sum_{n=1}^{N} \frac{1}{\sigma_{x,n} \sigma_v} \exp \left[ -\frac{d_{x,n}^2}{2\sigma_{x,n}^2} - \frac{d_{v,n}^2}{2\sigma_v^2} \right] \]  (3)

where \( \sigma_{x,n} \) is the Gaussian kernel width associated with the
vector x for the n-th observation, \( \sigma_v \) is the Gaussian kernel

## Table 2

<table>
<thead>
<tr>
<th>Field name</th>
<th>Statistic</th>
<th>Elevation (m)</th>
<th>Slope (m/100 m)</th>
<th>Curvature (m^-1)</th>
<th>Area=cos(aspect)</th>
<th>D_1</th>
<th>D_2</th>
<th>WI</th>
<th>Grain yield (Mg ha^-1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>North <em>n=6118</em></td>
<td>Minimum</td>
<td>1360.72</td>
<td>0.05 ^a</td>
<td>-0.015200</td>
<td>1.00</td>
<td>2.30</td>
<td>2.30</td>
<td>4.61</td>
<td>1.15</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>1381.68</td>
<td>14.35</td>
<td>0.014300</td>
<td>1.00</td>
<td>9.98</td>
<td>11.48</td>
<td>17.76</td>
<td>118.84</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>1368.26</td>
<td>4.20</td>
<td>0.000100</td>
<td>-0.62</td>
<td>3.97</td>
<td>4.00</td>
<td>7.07</td>
<td>27.82</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>1368.98</td>
<td>4.58</td>
<td>0.000028</td>
<td>-0.32</td>
<td>4.10</td>
<td>4.23</td>
<td>7.57</td>
<td>31.85</td>
</tr>
<tr>
<td></td>
<td>Standard deviation</td>
<td>4.33</td>
<td>2.87</td>
<td>0.002696</td>
<td>0.65</td>
<td>1.23</td>
<td>1.53</td>
<td>1.87</td>
<td>17.70</td>
</tr>
<tr>
<td>South <em>n=6869</em></td>
<td>Minimum</td>
<td>1356.57</td>
<td>0.05</td>
<td>-0.012700</td>
<td>1.00</td>
<td>2.30</td>
<td>2.30</td>
<td>4.87</td>
<td>8.11</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>1367.07</td>
<td>7.81</td>
<td>0.017100</td>
<td>1.00</td>
<td>9.30</td>
<td>10.36</td>
<td>17.92</td>
<td>70.76</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>1360.54</td>
<td>1.39</td>
<td>0.000000</td>
<td>0.47</td>
<td>4.05</td>
<td>4.11</td>
<td>8.30</td>
<td>34.06</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>1360.92</td>
<td>1.65</td>
<td>0.000035</td>
<td>-0.18</td>
<td>4.15</td>
<td>4.28</td>
<td>8.62</td>
<td>34.44</td>
</tr>
<tr>
<td></td>
<td>Standard deviation</td>
<td>2.05</td>
<td>1.12</td>
<td>0.001415</td>
<td>0.76</td>
<td>1.27</td>
<td>1.44</td>
<td>10.61</td>
<td></td>
</tr>
<tr>
<td>West <em>n=3317</em></td>
<td>Minimum</td>
<td>1381.60</td>
<td>0.05</td>
<td>-0.007800</td>
<td>1.00</td>
<td>2.30</td>
<td>2.30</td>
<td>5.56</td>
<td>14.88</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>1388.32</td>
<td>5.80</td>
<td>0.012000</td>
<td>1.00</td>
<td>8.92</td>
<td>9.03</td>
<td>14.25</td>
<td>79.46</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>1385.33</td>
<td>1.55</td>
<td>0.000000</td>
<td>-0.68</td>
<td>3.91</td>
<td>3.94</td>
<td>8.08</td>
<td>47.05</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>1385.16</td>
<td>1.72</td>
<td>-0.000035</td>
<td>-0.43</td>
<td>4.06</td>
<td>4.13</td>
<td>8.34</td>
<td>46.72</td>
</tr>
<tr>
<td></td>
<td>Standard deviation</td>
<td>1.66</td>
<td>0.91</td>
<td>0.001228</td>
<td>0.60</td>
<td>1.23</td>
<td>1.29</td>
<td>1.35</td>
<td>13.37</td>
</tr>
</tbody>
</table>

\_D_1 and D_2\_ are ln(SCA) with sinks remaining and filled, respectively, where SCA is specific contributing area. WI is the topographic wetness index, and \_n\_ is the number of cells per field with available wheat grain yield data.

\_a Minimum slopes were set to 0.0005 m/m for computation of WI.\_
width associated with the spatial variable \( v \), \( d_{x,x_n} = [(x - x_n)^T (x - x_n)]^{1/2} \) is the Euclidean distance between \( x \) and \( x_n \), and \( d_{v,v} = [(v(x) - v(x_n))]^T [v(x) - v(x_n)][1/2] \) is the Euclidean distance between \( v(x) \) and \( v(x_n) \), respectively.

By substituting Eq. (3) into Eq. (2), the point estimator of the conditional mean of \( v \) given \( x \) becomes

\[
\hat{v}(x) = \frac{\sum_{n=1}^{N} v(x_n) \left( \frac{1}{\sigma_{x,x_n}^2} \right) \exp \left[ -\frac{d_{x,x_n}^2}{2\sigma_{x,x_n}^2} \right]}{\sum_{n=1}^{N} \left( \frac{1}{\sigma_{x,x_n}^2} \right) \exp \left[ -\frac{d_{x,x_n}^2}{2\sigma_{x,x_n}^2} \right]}.
\]

Similarly conditional higher order moments can be obtained (Shin and Salas, 2000).

### 2.4.1. Spatial Analysis Neural Networks (SANN) structure

The computational algorithm used herein consists of four layers in which the neurons or nodes between layers are interconnected in the feed-forward direction, namely: 1) input layer, 2) Gaussian Kernel Function (GKF) layer, 3) summation layer, and 4) estimation layer. Consider the case of a spatial region \( R \) in a 2-dimensional domain and assume that \( N \) sample observations of the spatial variable \( v \) are available in the form of a set \( S \) (Sasowsky et al., 1992) consisting of pairs \( \{x_n, v(x_n)\} \).
neighbors, where \( x = [x_1, x_2] \) is the input vector and \( v(x) \) is the corresponding output vector. We want to estimate the values of \( v \) at any arbitrary point \( x \), i.e., \( \hat{\nu}(x) \). In this case, the input layer has 2 nodes, each representing the coordinates of the vector \( x = [x_1, x_2] \). The input layer passes the input coordinate vector to the GKF layer without any weighting.

The GKF layer consists of \( N \) GKF nodes that represent the receptive field or influence region of each observed vector \( x_n \) for \( n = 1, \ldots, N \). This vector (also called the center of the GKF node) uses the Gaussian function

\[
a_n = \exp \left[ -\frac{d_{x,n}^2}{2\sigma_{x,n}^2} \right], \quad n = 1, \ldots, N
\]

as the transfer or activation function. The smoothing parameter \( \sigma_{x,n} \) depends on two other parameters: the number of nearest neighbors, \( P \), and a control factor, \( F \). First, the root mean square distance (RMSD) between the center \( x \) and its \( P \) nearest neighbors is determined for each GKF node. The \( \sigma_{x,k} \) values have an important effect upon the accuracy of the estimation (Hayken, 1994), where the width \( \sigma_{x,k} \) is determined by

\[
\sigma_{x,k} = \text{RMSD}_k / F
\]

and \( F \) is a control factor that determines the density of the network. Here, optimal values of \( (P, F) = (9, 2.5) \) were used based on previous analyses (Martinez et al., 2004).

The summation layer in this example consists of 2 nodes. The outputs of the GKF nodes are passed to the summation layer through the weighted connections

\[
G_1 = \sum_{n=1}^{N} \left( \frac{1}{\sigma_{x,n}} \right) a_n, \quad \text{(7a)}
\]

and

\[
G_2 = \sum_{n=1}^{N} \left( \frac{1}{\sigma_{x,n}} \right) v(x_n) a_n. \quad \text{(7b)}
\]

Finally, the outputs from the summation nodes are passed with unit weights to the estimator layer, where

\[
\hat{\nu}(x) = G_2 / G_1. \quad \text{(8)}
\]

The SANN algorithm used herein has two operational modes: a training mode and an interpolation mode.

2.4.2. Network training and validation

In the training mode, the model structure and the parameters are determined in a two-stage procedure. Consider that \( N^* \) observations of \( v(x) \) are sampled randomly from the available data in a field. The total sample of size \( N^* \) is divided into two sets of equal size (i.e. \( N = N^*/2 \)) for training and validation. Previous work (Martinez et al., 2004) showed that 900 points each for calibration and validation provided reproducible results and appropriate spatial detail. Here, we set \( N = N^*/2 = 1000 \). The overall procedure may be summarized in the following stages:

2.4.2.1. Training stage.

a) \( N \) GKF nodes are selected randomly from the observation set of \( N^* \) nodes.

b) The centers of the GKF nodes are set equal to the observed vectors \( x_k \), and the activation widths \( \sigma_{x,k} \) of the GKF nodes are calculated.

c) For the given values \( (P, F) \) and the calculated width \( \sigma_{x,k} \) one can calculate the estimated value \( \hat{\nu}(x_k) \) from Eq. (8) and the estimation error from

\[
e_{\nu}^5 = \left( \frac{1}{N} \sum_{k=1}^{N} [v(x_k) - \hat{\nu}(x_k)]^2 \right)^{1/2}. \quad \text{(9)}
\]

d) The estimation error is minimized to determine the optimal weighting factors in the training stage.

2.4.2.2. Validation stage.

e) In this stage, another measure of the estimation error of the spatial variable \( v(x) \) is determined by using the other \( N^* - N \) observed points that were not used as GKF nodes in the first stage. Then, the root mean square error (RMSE) between the observed values \( v(x) \) and the estimated values \( \hat{\nu} \) is determined by

\[
e_{\nu}^n = \left( \frac{1}{N^* - N} \sum_{n=1}^{N^* - N} [v(x_n) - \hat{\nu}(x_n)]^2 \right)^{1/2}. \quad \text{(10)}
\]

Here, \( N^* = 2 \times 2000 \) sample points within each wheat field.

Thus, we compute RMSE values for both training and validation. The optimal parameter set is found by minimizing the estimation error (9) in the training stage. Only validation errors (10) are reported below, because validation provides the best measure of the model estimation error.

2.4.3. Joint spatial interpolation and yield estimation from topographic attributes

SANN contains a multivariate Gaussian kernel (Eq. (3)), which includes Euclidean distances and widths in geographical and topological spaces. In most cases, yield values do not vary in a consistent manner with easting and northing, but there may be large, nonlinear spatial patterns within a field that can be explained partially by the \((x,y)\) coordinates using SANN. These patterns cannot be generalized for prediction across different farm fields, so adding \((x,y)\) to the inputs is a means of interpolation within a given field. In the case of pure interpolation using only \((x,y)\) inputs, the separation distances used to compute the activation functions (Eq. (5)) and widths (Eq. (6)) take on geographical meaning. It is also possible to use SANN for joint interpolation (i.e., using \((x,y)\) only) and estimation based on topographic attributes.
2.5. Multiple Linear Regression (MLR) application

Regression analyses were performed in an equivalent manner to SANN in terms of the input data used and validation on a split set of points not used for calibration/training. First, half of the sample set (N=1000) were used for fitting a univariate or multivariate linear regression equation to observed yield data versus the discrete topographic attributes at each yield location. Second, the prediction error was computed using a separate set of 1000 observed values. This is not the conventional method, wherein coefficients of determination are normally based on only one set of data. Here, the goodness-of-fits statistics are better measures of predictive capability within each field.

Linear regression models were fitted to the calibration data by minimizing the sum of squared errors between observed and estimated values without any trimming of extreme values. Fitted parameters were saved and used for validation against data not used for fitting. Thus, prediction errors (Eq. (10)) may be greater than errors computed from the data used for model fitting (Eq. (9)).

Both SANN and linear regression models were calibrated (“trained”) and validated in a stepwise procedure starting with one topographic attribute at a time, followed by all paired combinations (dropping $D_f$), and progressing with selected attributes up to two combinations of the top six attributes (using either $D_f$ or WI). Finally, we added the normalized spatial coordinates ($X,Y$) under the category of 7 inputs, which essentially added interpolation of any spatial trend within a field.

2.6. Model prediction errors and statistics

Model prediction errors are measured using a relative root mean squared error (rRMSE), which is simply the RMSE for validation (Eq. (10)) divided by the sample standard deviation. A rRMSE value of zero represents a perfect prediction, and a value of unity signifies a prediction equivalent to using the mean value of the data at all spatial locations. In the case of an extremely poor spatial model, it is possible to obtain rRMSE>1. We also use a measure of model efficiency (Nash and Sutcliffe, 1970) defined as:

$$E = 1 - \frac{\text{rRMSE}^2}{1}.$$  \hspace{1cm} (11)

This is analogous to the coefficient of determination in terms of quantifying the fraction of variance explained up to a maximum value of 1.0, but it is possible to obtain $E<0$.

In all of the above methods, multiple sets of split calibration/validation data (30 for each set of inputs) were sampled randomly with replacement from the full data set in each field. In this way, we were able to compute a mean and standard deviation of each error statistic (i.e., rRMSE and bias). Thus, a measure of the uncertainty of the prediction error was estimated along with the mean error.

3. Results and discussion

3.1. Spatial patterns of measured crop yield

Crop grain yield varies spatially by an order of magnitude within each of the three fields with both ordered patterns and seemingly random variability. Fig. 2(a–e) shows measured wheat grain yield overlaid on the topography of our three fields, where the yield on these dryland fields varies from less than 1 Mg ha$^{-1}$ to over 7 Mg ha$^{-1}$ within small portions of each field. All three fields, but particularly the North field (a) with the most pronounced relief, show patches of higher yields in the swales and lower yields on summits and hillslope areas that are convex up. Associated patterns of topographic slope (d–f), total curvature (g–i), and log-transformed specific contributing area (j–l) are shown below the yield maps for visual comparison. Each topographic attribute highlights a different (but interrelated) characteristic of the landscape that may be expected to affect water, soil, and chemical/nutrient movement, as well as micro-meteorological conditions. For example, contributing area emphasizes watersheds with potential divergence (light) or convergence (dark) of flow, even delineating drainage lines (ephemeral streamlines) and valleys, where crop yield tends to be greatest.

Soil variability is partially correlated with topographic variability, such that use of topographic attributes for explanatory variables implicitly captures some fraction of the soil variability. A detailed study of soil formation in these fields is not available, but soil maps and field observations indicate that the soil types are fairly homogeneous at the field scale. However, soil hydraulic properties affecting water storage and flow rates can vary significantly based on soil cores taken along a transect on the North field (Green et al., 2003). Soil nutrients also vary with topography, but the present approach does not discriminate between soil factors that may be correlated with topography. In future work, one could use any soil property as an explicit explanatory variable in SANN or regression analyses, if such data were available spatially. Because this is the exception, the focus remains on topographic attributes here.

3.2. Linear regression versus SANN analyses

All of the detailed results are given in Table 3 and illustrated in Figs. 3 (MLR) and 4 (SANN). Figs. 3 and 4 show the detailed results in terms of the rRMSE statistic. Mean values of rRMSE are plotted with different symbols for different numbers of input variables, and error bars reflect the uncertainty in these values based on the 30 randomly generated sample sets. The variance among realizations is often small, such that the error bars may not be visible in some cases. Results for each number of input variables are ordered from best (lowest mean rRMSE) to worst (from $a$ to $p$ in Figs. 3 and 4), and a subset of the 10 best input sets ($a$ to $j$) are shown beneath each plot. In Fig. 3a, for example, slope ($S$), wetness index (WI), and elevation ($Z$) are the top three univariate explanatory variables for MLR in the North field.
3.2.1. Univariate results

Before exploring the full results, it is helpful to review the univariate results for both MLR (the $M$ is kept for notation, even though these results are univariate) and SANN. In Table 3, the mean relative bias (relative to the total sample standard deviation) is always very small, and the standard deviation of the relative bias ($s_{rb}$) is about an order of magnitude greater than the absolute value of the mean relative bias. The $s_{rb}$ values for SANN are generally greater than $s_{rb}$ for MLR (univariate). In all cases, the bias is substantially less than the RMSE and within the standard error of the RMSE values. Thus, we focus primarily on RMSE (or rRMSE) values in the remainder of the discussion.

Looking at rRMSE values for SANN versus MLR in the univariate cases, we discover a different ordering of input/explanatory variables in terms of mean prediction errors using...
In all three fields, regression resulted in lower prediction errors than SANN using only one topographic attribute (Table 3). Thus, point-to-point univariate regression out-performed SANN, which smoothed predictions over the input space due to the influence of the activation function with \( P = 9 \). We did not try to adjust \( P \) and \( F \) values to optimize for the univariate case, because optimal multivariate prediction was the primary goal. See Martinez et al. (2004) for the basis of selecting the \( P \) and \( F \) parameter values used here. Thus, SANN offered no advantage over regression in the univariate cases studied.

Comparing MLR across fields, not only is the ordering of input variables different, but the shapes of the univariate curves differ. In the North field (Fig. 3a) the first few variables are fairly evenly stratified in terms of rRMSE, whereas the first three input variables for the West field (3b) yield similar rRMSE values, which are markedly better than the other variables. In the West field, the three best explanatory variables \( (D_s, D_t, W_I) \) offered no advantage over regression in the univariate cases studied.

### Table 3 (continued)

<table>
<thead>
<tr>
<th>Number of inputs</th>
<th>Input variable(s)</th>
<th>Relative RMSE</th>
<th>Relative bias</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean Standard deviation</td>
<td>Mean Standard deviation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SANN MLR</td>
<td>SANN MLR</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean Standard deviation</td>
<td>Mean Standard deviation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean Standard deviation</td>
<td>Mean Standard deviation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean Standard deviation</td>
<td>Mean Standard deviation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean Standard deviation</td>
<td>Mean Standard deviation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean Standard deviation</td>
<td>Mean Standard deviation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean Standard deviation</td>
<td>Mean Standard deviation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean Standard deviation</td>
<td>Mean Standard deviation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean Standard deviation</td>
<td>Mean Standard deviation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean Standard deviation</td>
<td>Mean Standard deviation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean Standard deviation</td>
<td>Mean Standard deviation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean Standard deviation</td>
<td>Mean Standard deviation</td>
</tr>
</tbody>
</table>

b. See caption of Table 3a for notation

(continued on next page)
are all based on contributing area (see Eq. (1a)). Interestingly, slope ($S$) provided the worst predictions in the West field (3b), while standing out in the top two for the other fields (3a,c). Looking at Fig. 2 and Table 2, the West field has the least variability in slope, which may explain the relative weighting of $S$ in that case. It is worth noting that the strength of WI may be due to either $D_f$ or $S$.

Comparing SANN with MLR using only one input, MLR always produced mean values of rRMSE less than 1.0, while SANN exceeded 1.0 in the worst cases. Even for the best input sets, SANN yielded significantly greater errors than MLR, as discussed above. Both methods tended to order the inputs similarly (e.g., $S$, WI, $Z$ in the North field), but in the West field, $S$ went from 3rd in MLR to 7th in SANN. Furthermore, the cosine of aspect ($A$) had more explanatory power in SANN than in MLR for all cases, indicating nonlinear responses.

### 3.2.2. Multivariate analyses using topographic attributes only

For two or more inputs, the ordering of sets of input variables was less distinct, as indicated by the relatively flat curves.
Fig. 3. Multiple Linear Regression (MLR) validation results showing the mean values (symbols) and one standard deviation around each value of the relative Root Mean Squared Error (RMSE) of spatial predictions. Each value and standard error is computed from 30 sets of random samples of 1000 points not used to develop the regression relationship. *Inputs* or explanatory variable sets are given in tabular form below each graph in order from least to greatest error.
Fig. 4. Spatial Analysis Neural Network (SANN) validation results showing the mean values (symbols) and one standard deviation around each value of the relative Root Mean Squared Error (RMSE) of spatial predictions, as per MLR results in Fig. 3.
compared with the univariate curves in Figs. 3 and 4. This may have caused non-uniqueness in the selection of the sets of input variables used for prediction. However, the selection of the final 5 topographic inputs differed only by the use of either $D_f$ or WI (computed from $D_f$). Thus, differences between fields in terms of explanatory variables diminished as more inputs were used.

Overall results of the best mean predictions for each method and field are shown in Fig. 5. The main differences between the results for MLR (Fig. 5a) and SANN (5b) are evidenced by the steepness of the curves and the number of inputs at which the curves plateau. The steeper SANN curves indicate significant improvement in predictive power was achieved with each additional input variable up to 4 or 5, whereas MLR predictions did not improve much after only 3 inputs. This latter result is commonly observed in multiple regression studies of natural systems, where explanatory variables are often cross-correlated. Nonlinearities in the yield responses are captured by SANN using more inputs.

Using five of our main topographic attributes as inputs, SANN consistently out-performed MLR. In fact, Fig. 5 shows that, for the North and West fields, SANN out-performed MLR using two or more inputs, while model efficiencies were equivalent on the South field using 3 inputs. The improvement with SANN over MLR using five topographic attributes ranged from 10 to 20% of the variance explained (Table 4), and optimal model efficiencies using topographic attributes ranged from $E=0.39$ to 0.65.

3.3. Spatial yield interpolation/prediction within a field

By adding the normalized spatial coordinates $(X,Y)$ as input variables, one can improve the predictions within a given field by using interpolation along with the previous topographic estimators. The resulting geometric patterns are not expected to be transferable among different fields with distinct topographies. Such interpolation did not offer much benefit using MLR, which is limited to linear trends in space that are minimal at these scales of several hundred meters. The West field was an exception, where $E$ increased from 0.36 to 0.44 due to linear trends within this field that were not captured by the topographic attributes.

SANN, on the other hand, provides intrinsic local-scale interpolation among neighbors (here, $P=9$) via the activation function, which weights nearest neighbors the most. Thus, there was a pronounced jump in model efficiency values for all cases when we add $(X,Y)$ as inputs. The model efficiency for the West field increased by 0.24 with SANN versus 0.08 with MLR, indicating the importance of nonlinear trends. The maximum increase of 0.27 for the South field narrowed the gap between $E$ values of the three fields using topography alone. Overall, SANN with interpolation (adding $X,Y$) provided a marked improvement over MLR with the same inputs; the $E$ values were 29 to 36% higher (Table 4). Furthermore, this indicates that there is significant spatial autocorrelation in crop yield, which was not captured by the topographic inputs alone. That is, the residuals were not completely random. The SANN method is useful for identifying and predicting such behavior.

Finally, the ordering of fields in terms of model efficiency is consistent between methods. This indicates that it is easier to explain more of the spatial variability in some fields than others. This is not surprising in the case of the North field, which has the greatest topographic relief and variability in crop yield (Fig. 2). The difference between West and South fields, however, is less predictable a priori. In fact, the West field is flattest of the three (Table 2).
4. Summary and Conclusions

In this paper, we explored the use of landscape topographic attributes derived from DEMs of three farm fields in eastern Colorado to help explain/predict the spatial variability of winter wheat grain yield within each field. The elevation data used to generate the DEMs and the yield data were collected at approximately the same spacing (∼10 m). Data in two fields comprised over 6000 spatial points. We randomly sampled 30 sets of 2000 points from each of the three fields (with replacement), using half (1000 observations) for calibration/training and the other half for validation of each model. Model prediction errors and efficiencies were quantified in terms of a mean and standard deviation of the error statistics for each input set.

Using the Spatial Analysis Neural Networks (SANN) method, much of the variability in measured wheat grain yield within fields can be predicted from landscape topographic attributes. Sets of five topographic inputs (Z, S, A, C, and either Dr or WI) provided maximum model efficiencies of 0.39, 0.56 and 0.65 for the three fields using the SANN method. The best prediction (E = 0.65) using topographic attributes corresponds with the North field, which has the most pronounced topographic features affecting crop yield. The worst prediction (E = 39%) corresponds to the South field, where yield patterns were less correlated with topography.

Univariate linear regression out-performed SANN using only 1 input as a spatial predictor, but SANN consistently out-performed Multiple Linear Regression (MLR) with 3 to 5 topographic attributes as inputs. MLR experienced little benefit of adding the 4th and 5th inputs (explanatory variables), whereas the model efficiency of SANN climbed up to 4 inputs, and the 5th input provided 1% improvement in two of the three fields. Nonlinearity in the response and implicit smoothing by the SANN activation function may explain the improved predictions compared with MLR. More could be done with the SANN algorithm to preserve the spatial variance of each field, in addition to minimizing the sum of square errors.

In addition to improving predictions with the topographic inputs, SANN is a useful spatial interpolator. By simply adding the spatial coordinates (X,Y), SANN model efficiencies improved from 12 to 27% (Fig. 5, see 7 inputs), and the overall model efficiency (E) was 29 to 36% better than MLR (Table 4). This is due to the implicit local smoothing in SANN via the activation function, which takes on spatial significance using (X,Y). MLR predictions, on the other hand, improved by 0, 1 and 8%, where only the West field displayed significant linear trends over the field. Neither method is able to transfer these improved predictions between fields, which is why this approach is viewed as interpolation within a field. However, the results provide insights into the amounts of residual spatial patterns that could not be predicted using topography alone. Looking at Fig. 2b–c, for example, there are patches of high yield in unexpected topographic locations, like summits. Yet, there is enough spatial autocorrelation in these patches for SANN to fit that behavior through interpolation.

The methods proposed and tested here for estimating spatial patterns of crop yield (or similar agronomic variables) from topographic attributes can guide mapping efforts for sampling and variable-rate applications. The complexity of the SANN approach was rewarded with significant improvements over MLR in predictions of spatial crop yield patterns. Future applications of SANN may use direct soil information, given detailed spatial data for inputs.

Acknowledgments

Data were collected from the Lindstrom Farm in cooperation with Gilbert Lindstrom. Technical support from Mike Murphy is also appreciated. This work was performed as part of a specific cooperative agreement between the USDA-ARS and Colorado State University: “Quantifying Space–Time Variability in Agricultural Landscapes” administered with Dr. Laj Ahuja.

References


