2005

Processing of Yield Map Data

J. L. Ping
University of Nebraska-Lincoln

Achim R. Dobermann
University of Nebraska-Lincoln, adobermann2@unl.edu

Published in *Precision Agriculture* 6 (2005), pp. 193–212.

Processing of Yield Map Data

J. L. Ping and A. Dobermann

Department of Agronomy and Horticulture, University of Nebraska–Lincoln, Lincoln, NE 68583-0915, USA. *Correspondence:* adobermann2@unl.edu

Abstract

Yield maps reflect systematic and random sources of yield variation as well as numerous errors caused by the harvest and mapping procedures used. A general framework for processing of multi-year yield map data was developed. Steps included (1) raw data screening, (2) standardization, (3) interpolation, (4) classification of multi-year yield maps, (5) post-classification spatial filtering to create spatially contiguous yield classes, and (6) statistical evaluation of classification results. The techniques developed allow more objective mapping of yield zones, which are an important data layer in algorithms for prescribing variable rates of production inputs.

Keywords: yield data screening, yield mapping, spatial classification, yield zones

Introduction

Yield mapping is one of the most widely used precision farming technologies. As more yield monitors are used and multiple-year yield data are accumulated, there is increasing need for robust data processing and interpretation techniques. Yield monitor data contain systematic and random sources of measured yield variation, including (i) more stable yield variability related to climate and soil-landscape features, (ii) variable management-induced yield variability, and (iii) measurement errors associated with the yield mapping process itself (Stafford *et al.*, 1996; Lark *et al.*, 1997; Blackmore and Moore, 1999; Arslan and Colvin, 2002b). Management-induced variation includes random events that typically occur in small patches, such as planter skips, poor crop establishment, non-uniform fertilizer application, herbicide damage, lodging or pest damage. Measurement errors include grain flow and moisture sensor errors, errors due to geo-referencing and combine movement, operator errors, and data processing errors (Blackmore and Moore, 1999; Arslan and Colvin, 2002b).

Although a single-year yield map is useful for posterior interpretation of possible causes of yield variation, it is of limited value for strategic site-specific management decisions over medium to long-term periods. With multiple years of geo-referenced yield data, repeating yield patterns and their natural causes can be separated from management-or measurement-induced random yield variation in each year.

Our primary goal was to develop methodological guidelines for creating maps of yield classes in irrigated continuous maize (*Zea mays* L.) and maize-soybean (*Glycine max.* [L] Merr) systems, which must represent zones with different yield expectation within a field. Figure 1 illustrates a recently proposed flow diagram of yield data processing. In previous publications, we have reported on specific methodologies for each of the data processing steps shown in Figure 1, i.e., a new algorithm for raw data screening (Simbahan *et al.*, 2004), the use of remote sensing imagery

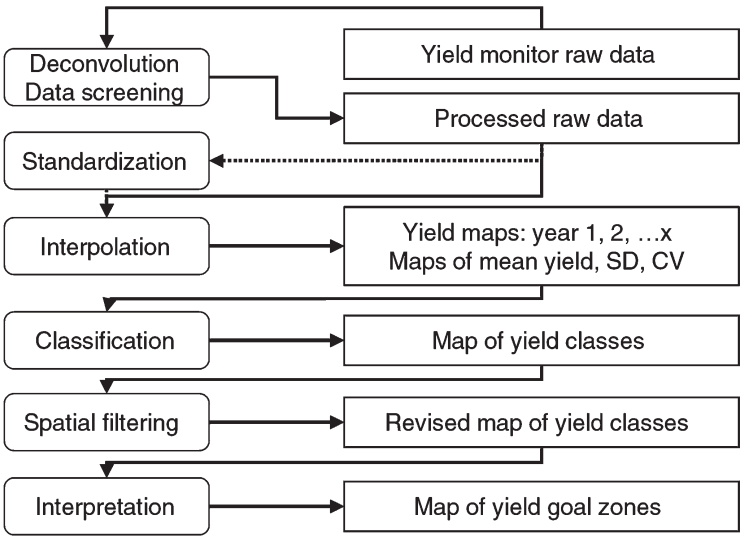

Figure 1. Proposed flowchart for post-processing of yield monitor data. Modified from Ping and Dobermann (2003).

for improving yield data interpolation (Dobermann and Ping, 2004), and techniques for spatial classification of multi-year sequences of yield maps into classes of different yield performance (Dobermann *et al.*, 2003; Ping and Dobermann, 2003). The objective for the present study was to apply the complete approach as a whole at a site which had not been used in previous studies on developing these methodologies, i.e., to validate the conclusions drawn for other sites with similar environmental and management characteristics.

Material and methods

Study site and yield data collection

Yield monitor data were collected from an irrigated continuous maize field near Bellwood, Nebraska, USA (41.3267° N, 97.3356°W) from 1997 through 2002. This field was 68.3 ha in size and included four major soil series (Soil Survey Staff, 1999): Thurman loamy fine sand (mesic Udorthentic Haplustolls), Muir silt loam (superractive, mesic Cumulic Haplustolls), Ovina-Thurman coarse-loamy sand (mixed, mesic Fluvaquentic), and Brocksburg sandy loam (mixed, mesic Pachic Argiustolls). Thurman soils with low fertility and yield potential mainly occurred across the southwest to northeast field parts, where elevation and slope changed from 3% to 6%. The rest of the field was flat with slopes in the 0–3% range. A drainage ditch crossed the whole field from north to south in the western half of the field (Figure 2).

Maize was planted around April 20 each year at a density of about 75,000 plants ha^{-1} and combine-harvested around October 15 using a calibrated Ag Leader™ PF3000 yield monitor with elevator mounted moisture sensor (Ag Leader® Technology, Inc., Ames, IA, USA) and a differential Global Positioning System (GPS) re-

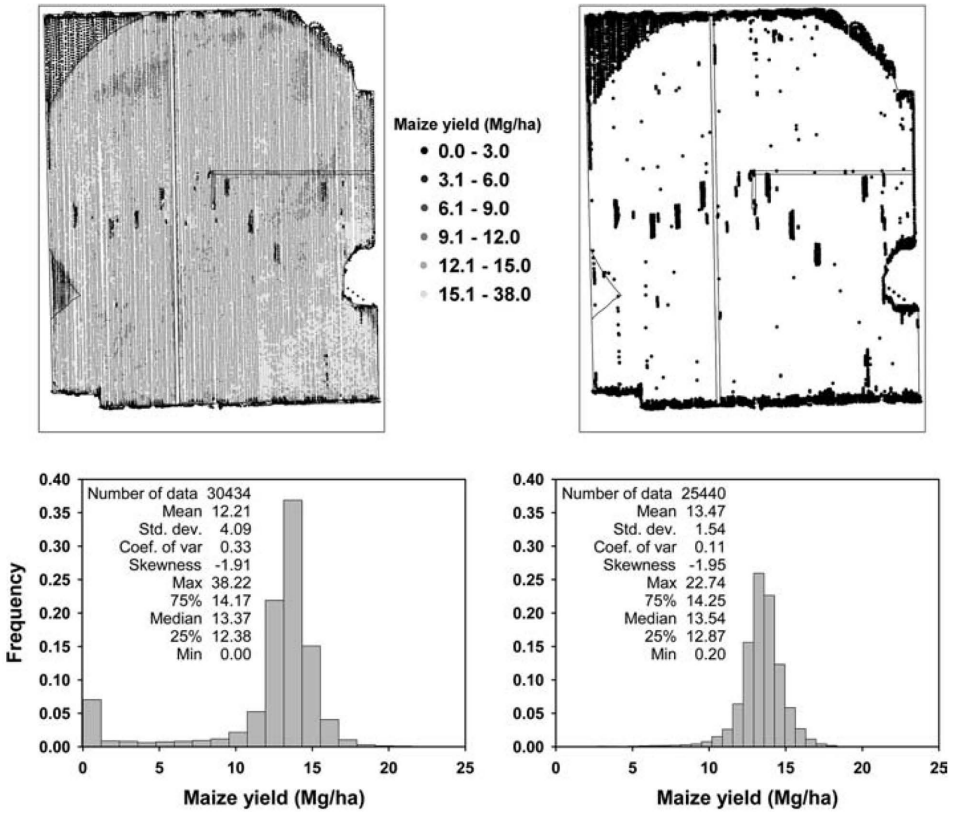

Figure 2. Yield monitor data measured in 2002 (top left), all data points removed by the yield screening algorithm (top right), and the frequency distribution of maize grain yield before (bottom left) and after screening (bottom right).

ceiver. The combine harvested with a swath width of 6.1 m (eight rows) and yields were recorded at 2 or 3 s logging intervals. Grain yields were expressed in Mg ha^{-1} with moisture content of 0.155 g H_2O g^{-1}.

Note that changes in water management, tillage and row direction probably influenced the yield patterns. From 1997 through 2000, only one centre-pivot irrigation system was used (indicated by the circle in Figure 4), while 14.8 ha on the southern end received furrow irrigation. In 2001, a second half-pivot was installed in the south and both pivots covered 67.8 ha since then. The field was managed as a ridge till system with 0.91 m row spacing until 2000, but with 0.76 m row spacing in 2001 and 2002. The southeast corner had maize rows in east–west direction before 2001, whereas all other areas were planted in north–south row direction. The whole field was planted in north south row direction in 2001 and 2002.

Screening and interpolation of yield data

Yield monitors are sensitive to changes in grain yield but a time delay exists because the grain flow through a combine resembles a diffusive process (Arslan and

Colvin, 2002b). In a first data processing step, grain flow delay correction and conversion of annual yield monitor raw data (.yld files) to advanced text file format were done using SMS Basic 1.01 (Ag Leader® Technology, Inc., Ames, IA, USA). Optimal grain flow shift settings were identified using the procedure proposed by Beal and Tian (2001). This procedure assumes that incorrect grain flow shift would result in a large ratio of the surface area of a 3-dimensional plot of the yield monitor data (yield plotted as *z*-variable versus geographical coordinates) to the 2-dimensional projected area of the upper surface (equivalent to the harvested whole field area). Optimum selection of grain flow shift would occur when this ratio is at a minimum. For all four fields, we estimated this ratio for different grain flow shifts ranging from 6 to 18 s and plotted it against the delay time. Based on the results, a value of 10 s was found to be optimal grain flow shift for maize, confirming results obtained at similar other sites (Simbahan *et al.*, 2004).

Following this, yield monitor data were screened to eliminate common errors. A sequential screening algorithm was applied, which screens for and deletes six types of erroneous or uncertain values (Simbahan *et al.*, 2004): (1) combine header status is up, (2) start-and end-pass delays for both headlands and stop-and-go segments within the field, (3) frequency distribution outliers of distance traveled, grain flow, and grain moisture, (4) yields outside user-defined minimum and maximum biological yield limits, (5) small patches or narrow strips with extremely low or high yields that are not closely related to immediate neighbors, and (6) short segments and co-located yield records.

Steps (1) and (2) remove technical errors that are always associated with yield monitor operation (Blackmore and Moore, 1999; Arslan and Colvin, 2002b). Step 1 eliminates erroneous data values that are recorded while the combine header is up. Step 2 removes yield points recorded after the header has been lowered but grain flow has not started or has not stabilized yet (start-pass delay), as well as values at the end of harvest segments, when cutting has stopped but the header has not been raised yet (end-pass delay). Settings for start-and end-pass delays may differ among crops and harvest combines due to differences in swath width, harvest speed, and grain flow through a combine. To obtain location-specific settings, grain flow measured during a short time period after start of a new harvest segment or before the end of a harvest pass was plotted versus time for numerous different harvest passes in the field (data not shown). Based on this, 12 and 6 s were selected as default settings for start-and end-pass delay, respectively, which was different from the 8 and 4 s delays found in the study of Simbahan *et al.* (2004), respectively.

Steps 3 through 6 attempt to remove other erroneous yield records caused by combine operation and yield sensing, as well as uncertain values due to localized, extreme yield variation. A combination of statistical tests and empirical criteria is used for this screening. In step 3, an outlier test is performed for the variables grain flow, grain moisture, and distance traveled, based on the global means and standard deviations (SD) of these three measurements. Values outside the mean ± 3 SD range are deleted. In step 4, the user must provide an estimate of the expected biologically possible yield range. The value for the maximum possible yield should represent the crop yield potential (Evans, 1993), whereas the minimum value should be a number close to the minimum value the combine harvester can measure accurately. Default values in our studies were 0.01 for lower

and 22 Mg ha^{-1} for upper yield limits in maize (155 g kg^{-1} moisture content). The maximum value represents the estimated yield potential for this site, based on crop simulations done using long-term weather data and the Hybrid Maize model (Yang *et al.*, 2004).

Step 5 attempts to remove yield variability that often occurs in small patches or strips. Following the movement of the combine through the field, a local neighborhood test is performed for each location for which a yield monitor value has been recorded (Simbahan *et al.*, 2004). Using inverse distance interpolation, grain yield is estimated for each location from all values within a moving window that includes the three preceding and three succeeding yield records in the same swath as well as yield records within a band perpendicular to the tangent of the path traveled, crossing three adjacent harvest passes on both the left and right sides of the path traveled. The confidence interval of the estimate is obtained (default value: 95% or 2 SD). If the measured yield is outside this interval, the yield value is considered a spatially uncorrelated outlier and discarded. The rationale for this definition is that yield at any location is likely to be spatially correlated to its immediate neighbors, irrespective of the direction of the combine movement. If that is not the case, a random event must have caused an unusually high or low yield value recorded at the location being tested, either due to yield monitor error or due to specific crop management events that occur in very small patches. The former may include sudden changes in speed or grain flow (Arslan and Colvin, 2002a,b), whereas the latter may be caused by planter skips, poor crop establishment, non-uniform fertilizer application, herbicide damage, lodging, pest damage, or other events. Conceptually, the local outlier test performed in Step 5 is similar to the H-method proposed by Noack *et al.* (2003), but both the definition of the local neighbors and the statistical outlier test differ from it.

Step 6 removes short segments caused by combine stop-and-go events within the field and data points that were recorded with the same geographical coordinates. Short segments are considered unreliable because most data points in them are affected by start-or end-pass delays. As a default, segments with less than 12 yield monitor points were identified as short segments and deleted. Co-located data points can be caused by GPS error or overlapping harvest passes.

To eliminate yield variation caused by different management systems, screened annual yield data were normalized by dividing the measured values by the average of the corresponding irrigation method for a given field and year. The resulting relative yields were the relative percentage yields as used by Blackmore (2000) and indicate how the yield at each point differs relative to the mean of the field. Normalized point yield data were interpolated to a 4 m × 4 m grid using ordinary block kriging (Minasny *et al.*, 2002). Maps for each year as well as maps of the mean yield and its standard deviation across all 6 years were used for cluster analysis.

Spatial classification of yield variability

Ward's minimum variance method (SAS Institute Inc., 1999) was used for hierarchical cluster analysis, while non-hierarchical clustering was done using the fuzzy *k*-means method (Minasny and McBratney, 2003). Following the approach

described by Dobermann *et al*. (2003), input variables for the cluster analysis were either average relative yield (MY, univariate classification), average relative yield and SD of yield (MS, bivariate classification) or all individual years of yield maps (AY, multivariate classification). The number of classes ranged from 2 to 10.

Many management decisions require that maps of yield classes contain relative large, homogeneous and spatially contiguous units. Two approaches for creating maps of spatially more contiguous yield classes were evaluated (Ping and Dobermann, 2003). In the first approach, prior-classification interpolation (PCI), it was assumed that larger grid sizes may result in maps of yield classes with less spatial noise and better suitability for management. To evaluate this, the square grid size of the annual yield maps was increased from 4 m (16 m^2 cells), 8 m (64 m^2), 16 m (256 m^2), 32 m (1024 m^2), to 64 m (4096 m^2) using ordinary block kriging and the global variogram option in VESPER (Minasny *et al.*, 2002). The resulting interpolated yield maps were then used for cluster analysis to map yield classes.

In the second approach, post-classification filtering (PCF), block kriging interpolation of annual yield maps was done at 4 m (16 m^2 cells) grid size, followed by cluster analysis and applying image filtering techniques to the map of yield classes (spatial clusters) to smoothen the map units. That process involved a sequence of applying Focal Analysis, Clump, and Eliminate functions in Erdas Imagine 8.5 (Leica Geosystems, Atlanta, GA, USA) to the original 4-m maps of the yield classes created by cluster analysis (Ping and Dobermann, 2003). Square window sizes in this filtering were varied to be equivalent to the grid sizes of 8, 16, 32, and 64 m used in PCI. Focal Analysis is a smoothing process, which uses a moving window to replace the value of a cell (= center point of a moving window) based on a set of surrounding cells. We used window medians to replace the center point from the neighbors of 3 × 3, 5 × 5, 9 × 9, and 17 × 17 cells, which simulated the corresponding 4, 8, 16, 32, and 64 m grid sizes used in the PCI approached. Next, in clumping analysis, each cell was assigned to contiguous groups and the resulting images were then processed through the ELIMINATE function, which removes small clumps by replacing the values of pixels in these clumps with the value of nearby larger clumps. The software applies a focal majority filter on the input file in an iterative fashion, so that the data values of large clumps overwrite the data values of small clumps. The iteration continues until all the small clumps have been completely removed. The final clumps are then recoded using the ''Original Value'' attribute so that the output values of the remaining clumps are in the same range as the values in the original file (Leica Geosystems, 2003).

Evaluation of classification performance

To compare the effectiveness of the different methods in explaining the yield variance in each year *j*, we used the complement of the relative variance (Webster and Oliver, 1990):

$$RV_j = 1 - s_W^2 / s_T^2 \quad (1)$$

where s_W^2 is the within-class variance and s_T^2 is the total variance, both estimated by post-classification analysis of variance for a particular year j. An RV_j value was computed for each individual yield map year and an average value (RV_c) was computed across years. An ideal classification method would have a RV_c close to one and a small range of the RV_j among individual years. An analysis of variance of the standardized mean yields among different classes was conducted to test for differences in mean relative yield among the yield classes.

Results

Yield data screening

In studies with irrigated and rain-fed maize and soybean at sites in a similar environment, the yield screening algorithm removed 13–20% of the original yield monitor data (Simbahan *et al.*, 2004). Similar ranges were found for irrigated maize grown at the Bellwood site. As an example, Figure 2 shows the result of the yield data screening in 2002. In this case, 16.4% (4992) of the original yield data were removed, which greatly improved the frequency distribution of grain yield. The frequency distribution of grain yield computed from the original yield monitor data was negatively skewed, including many zero values, but also some extreme yields which exceeded the known biological yield limit of about 22 Mg ha^{-1} for this site. Data removal mainly occurred in the non-irrigated, northwest field corner and near headlands, but also around stop-and-go segments within the field as well as locations dispersed throughout the entire area (Figure 2).

Stepwise removal of yield monitor data indicated that 71% of all data removal took place in the first two steps of data screening, whereas the remaining 29% were removed in steps three through six of the screening algorithm (Figure 3). These proportions were similar to observations made at other sites for both maize and soybean (Simbahan *et al.*, 2004). Data removal in the first two steps mostly included zero or very low yields, but few extremely high values were also removed. Erroneous yield points due to header-up status (1941) and start/end-pass delays (1581) in the yield monitor operation were mostly removed in the headland areas, but also included stop-and-go locations inside the fields. Step 3 removed 1007 yield data points, mostly located in the northwest field corner and around field edges. Most of these locations were outliers in the grain flow outlier test because maize in that non-irrigated corner suffered from severe drought during the year 2002 growing season, resulting in nearly complete crop failure. Step 4 removed data points in the same area as well as some other locations scattered throughout the field based on the empirically defined yield limits. Because many raw data values that would cause outliers in the computed grain yield were removed in the preceding step (3), only 86 additional points were deleted in step 4.

Step 5 removed 371 yield points that were identified as local outliers within the moving local neighborhood (Figure 3). Such outliers included most of the remaining yield points in the dry northwest corner as well as locations that were widely dispersed across the field. The latter included locations at which spikes or sudden

Figure 3. Sequential removal of erroneous or uncertain yield data points in the six screening steps, shown for the yield monitor data measured in 2002.

drops in yield occurred due to localized management problems or sudden shifts in combine speed. Step 6 deleted six points that had repeated records of yield for the same locations.

Overall, these results confirm that the yield screening algorithm used was robust in detecting major errors or extremes in yield monitor data that may be caused by the yield mapping process, by management or by natural events. The proportions of data removed in various steps were similar to previously analyzed fields. In

Table 1. Summary statistics of maize grain yield at Bellwood, Nebraska. Values shown refer to the yield monitor data remaining after data screening

Year	Mean	Median	Min	Max	SD	Skewness	CV
			Mg ha^{-1}				%
1997	11.9	12.1	0.7	20.0	1.64	–1.61	14
1998	12.7	12.7	1.7	21.8	1.64	–0.91	13
1999	12.3	12.7	1.5	20.1	1.70	–1.66	14
2000	11.5	11.9	1.3	17.0	1.68	–1.59	15
2001	12.0	12.2	2.4	21.0	1.30	–1.49	11
2002	13.5	13.5	0.2	21.7	1.54	–1.95	11

particular, it should be noticed that the total number of points removed in screening steps 3 through 6 accounted for less than one third of all removed yield data, but their removal significantly improved the modeling of semivariograms of grain yield which, at other sites, led to a relative increase of the precision of interpolated yield maps by about 4% to 5% (Simbahan *et al.*, 2004).

Spatial and temporal yield variability

Average maize yields after data screening ranged from 11.5 to 13.5 Mg ha^{-1} during the 1997–2002 period, with maximum yields ranging from 17.0 to 21.8 Mg ha^{-1} (Table 1). Relative spatial yield variability in each year was modest, with CVs ranging from 11% to 15%. In all years, the highest yields occurred in the southeast corner, whereas the lowest yields consistently occurred in northwestern and central-east areas of the field (Figure 4). Linear correlation coefficients of grain yields between different years ranged from 0.37 to 0.69 (Table 2), slightly less than at two other irrigated maize sites with similar management (Dobermann *et al.*, 2003).

Temporal variation in crop response to soil and climate among years, changes in crop management, and remaining artifacts in the yield maps probably caused the variations in correlations among yields measured in different years. In most years, small patches of randomly low or high yields remained even after data screening, especially in 1998 and 1999 (Figure 4). Although the exact causes of these ''speckles'' were not identified, it is likely that many of them were due to random events that caused gaps in the canopy or other crop damage. In some instances, undetected yield monitor errors such as surges in grain flow may have remained as well.

Temporal variability was affected by water management as indicated by relatively small correlation coefficients before 2001 as compared to after 2001, when the second pivot was installed (Table 2). Except for the edge areas where tillage and row direction changed, high yielding areas tended to have low standard deviation, whereas low yielding areas, headlands, and areas with changing irrigation management tended to have high standard deviation of relative yield across years (Figure 4).

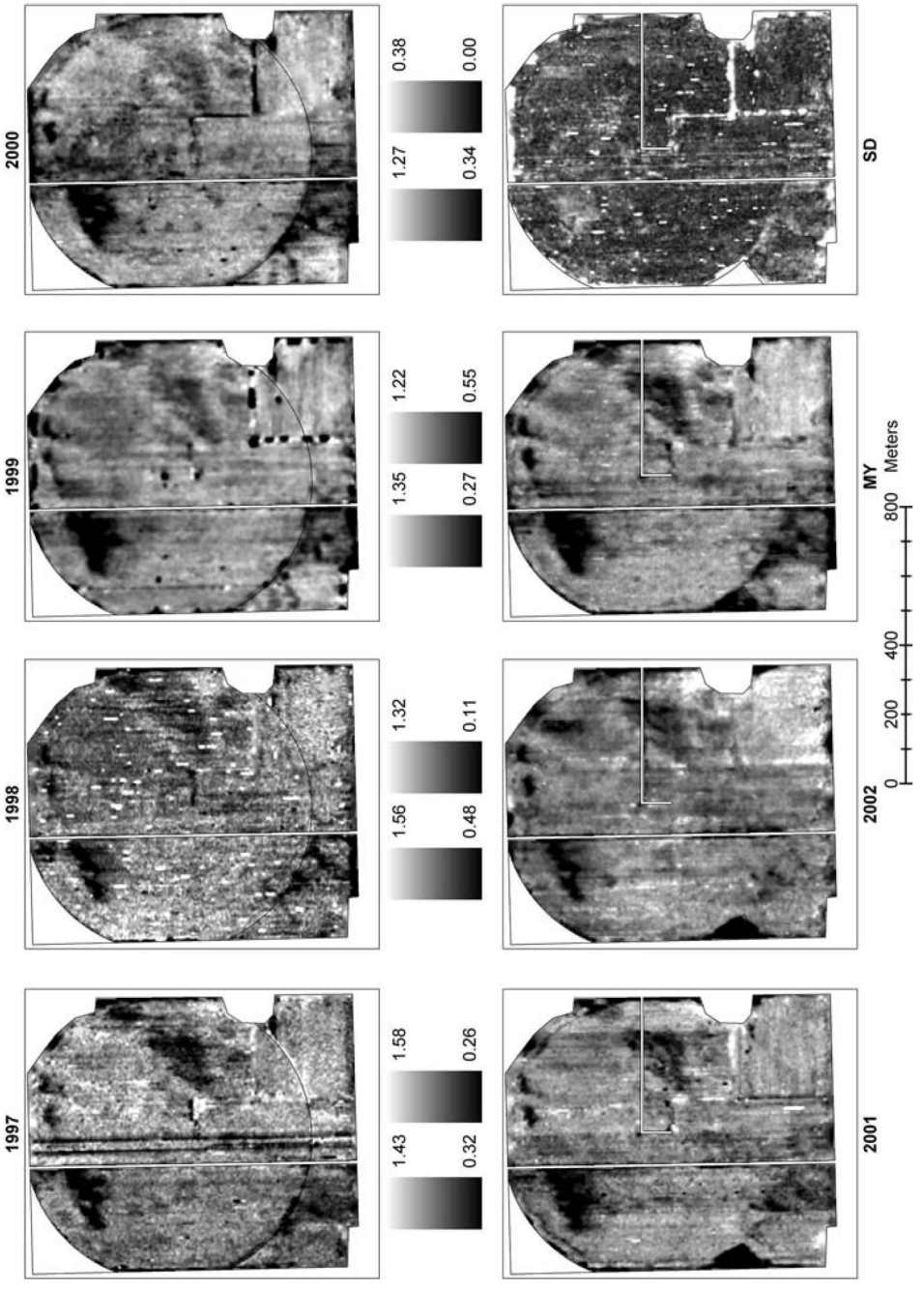

Figure 4. Kriged maps of relative grain yield at Bellwood from 1997 through 2002, mean yield (MY), and the standard deviation (SD) of grain yield across all years. Within each column, the left legend refers to the map shown above, the right legend to the map shown below. Relative yields were calculated as the relative difference to the field mean (%/100), i.e., a value of 1 represents the field mean yield for a particular year. Maps for 2001–2002 also show the pivot access road constructed in 2001.

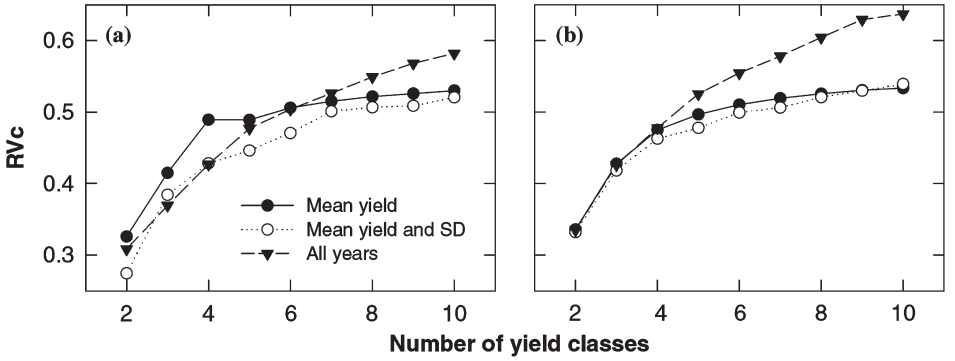

Figure 5. Average yield variability accounted for by the classification of multi-year yield map data (RV_c) as a function of data sources used and the number of classes selected. (a): hierarchical cluster analysis using Ward's methods; (b): non-hierarchical fuzzy-*k*-means cluster analysis.

Yield classification as affected by classification method, data source, and class number

Yield classification across years was affected by clustering method, data source, and the number of yields classes chosen (Figure 5). When mean yield (MY) or mean yield plus standard deviation (MS) were used as inputs for the cluster analyses, both Ward and fuzzy *k*-means methods showed similar results, in which RV_c increased with class number increasing from 2 to 7, but leveling off thereafter (Figure 5). Furthermore, MY resulted in better performance than did MS, as indicated by somewhat larger RV_c values. Using MY data and seven yield classes, both clustering methods accounted for 52% of the overall yield variation observed across all 6 years. This level of yield variation accounted for was similar to RV_c values of 0.60–0.66 for six to seven yield classes (MY-based) obtained at two other irrigated maize sites in Nebraska (Dobermann *et al.*, 2003). The slightly lower RV_c at the Bellwood site was probably due to the management changes occurring during the 1997–2002 period.

When yield maps of all individual years were used as input variables for multivariate cluster analysis (AY), the fuzzy *k*-means method resulted in significantly greater RV_c than those obtained with the Ward method. Moreover, with both clustering algorithms, RV_c kept increasing with increasing class number particularly for

Table 2. Linear correlation coefficients between maize grain yields in different years

Year	1997	1998	1999	2000	2001
1998	0.37				
1999	0.45	0.45			
2000	0.42	0.42	0.59		
2001	0.51	0.39	0.44	0.42	
2002	0.42	0.38	0.39	0.39	0.69

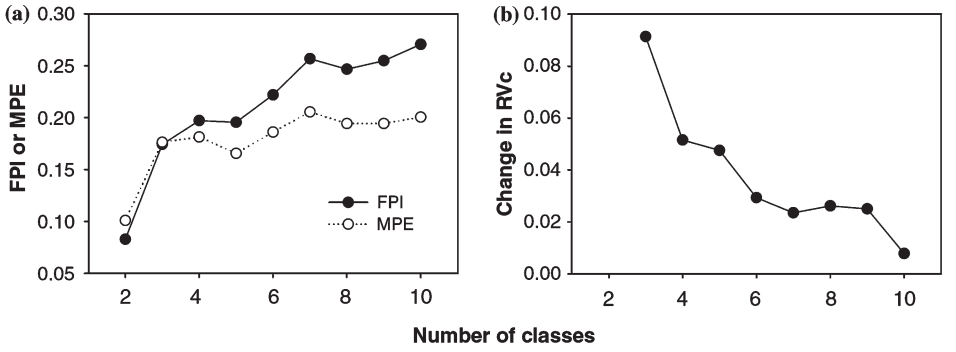

Figure 6. Fuzziness performance index (FPI), modified partition entropy (MPE), and the rate of change of the average yield variance accounted for by the classification (RV_c, %/100) as function of the number of yield classes chosen. Values refer to fuzzy-*k*-means clustering using individual years (AY) as input data.

the fuzzy-*k*-means method (Figure 5). Maximum RV_c achieved was 0.64 with fuzzy-*k*-means, using AY data, and 10 yield classes (Figure 5b).

There is a tradeoff between class number and classification performance. A large number of classes results in increased RV_c, but at the cost of increased map fragmentation, which may cause difficulties for implementing site-specific input management (Boydell and McBratney, 2002). In fuzzy *k*-means clustering analysis, two indices, fuzziness performance index (FPI) and modified partition entropy (MPE), can be used to determine the optimum class number (Roubens, 1982). The FPI estimates the degree of membership sharing among classes and ranges from 0 to 1, where a higher value indicates strongly sharing membership and 0 means crisp classes. The MPE estimates the degree of disorganization of classes and ranges from 0 to 1, higher MPE indicates strong disorganization and 0 indicates superior organizations.

Typically, FPI and MPE decline with increasing number of classes number and the optimum classification is reached at near minimum of both FPI and MPE. However, the relationships between FPI and MPE with class number do not always show such expected patterns (Boydell and McBratney, 2002). In our case study, both FPI and MPE increased with increasing class number (Figure 6). Lark and Stafford (1997) suggested to select the optimal class number k when $\mathrm{MPE}_{k-1} \rightarrow \mathrm{MPE}_k \approx \mathrm{MPE}_{k+1}$. Following this criterion, the optimal number of classes in our case was seven (Figure 6).

An additional criterion that we propose is the rate of change in RV_c with increasing number of yield classes. As the number of classes increases, RV_c gradually rises to a maximum value (Figure 5), but the rate of its increase declines sharply (Figure 6). At the Bellwood site, further gain in RV_c became small once class numbers exceeded six and the change in RV_c remained nearly constant for 7 to 9 classes. Thus, seven classes from the fuzzy *k*-means clustering of AY data appeared to be a reasonable solution for this site.

Spatial yield classification as affected by aggregation method and filtering of yield classes

Irrespective of the method chosen, yield classification based on small grid cells often results in fragmented maps that include much noise such as single pixels or small patches embedded within larger areas (Dobermann *et al.*, 2003). The original yield classification method—fuzzy-*k*-means with AY data and seven yield classes mapped for 4m × 4 m grid cells—resulted in 1178 individual patches (or potential management units), averaging just 0.058 ha per patch (Figure 7, top left). The clustering procedures focused on maximizing the variance between classes and minimizing the variance within classes, without constraints to form spatially contiguous patches that are large enough for management. Map fragmentation in yield classes may be caused by numerous small patches remaining in annual yield maps, even after intensive screening (Figure 4).

In general, map fragmentation decreased as interpolation grid sizes in the PCI method or spatial filtering window sizes in the PCF method were increased (Figure 7). However, important differences occurred between these two methods. In the PCI procedure, increase in interpolated grid size caused a significant loss of map quality, as evidenced by a steep decline in RV_c (Figure 8) and maps that did not accurately depict spatial yield patterns (Figure 7, top row). Choosing a coarse resolution of more than 8 m × 8 m for yield interpolation prior to classification (PCI method) resulted in significant loss of information. Misclassifications mainly occurred near yield transitions, there was poor agreement with the original map, and statistical separation of mean yields among classes was poor (Table 3).

In contrast, applying spatial filtering techniques to maps of yield classes that were created by cluster analysis of yield maps with a 4 m × 4 m resolution (PCF method) greatly improved the suitability of the map of yield classes for site- specific management. Post-classification spatial filtering removed map fragmentation and map unit contamination due to erroneous data, thereby creating maps of yield classes that were composed of smoother, spatially contiguous map units (Figure 7, bottom row). The original map resolution was maintained, little loss of the yield variability accounted for occurred (only small decrease in RV_c, Figure 8), and high spatial agreement with the original 4-m map was maintained. Average relative yields were significantly different among all yield classes (Table 3), indicating that good separation of yield zones was maintained. For example, seven yield classes at the original, unfiltered grid size of 4 m were mapped as 1178 patches with a mean patch size of 0.058 ha, accounting for 58% of the yield variance. With a 64 m filtering window size used in PCF, the number of patches decreased to 39 (mean size 1.75 ha per patch) with 42% of the yield variance accounted for. In comparison, increasing the interpolated grid size in PCI to 64 m yielded 112 patches (mean patch size 0.61 ha), but only 12% of the yield variance accounted for. The same PCF approach with 64-m filtering window resulted in 19 patches per field, patch sizes averaging 3.3 ha, and 53–57% of yield variance accounted for at two other irrigated maize sites in Nebraska, which had similar environmental conditions and crop management (Ping and Dobermann, 2003).

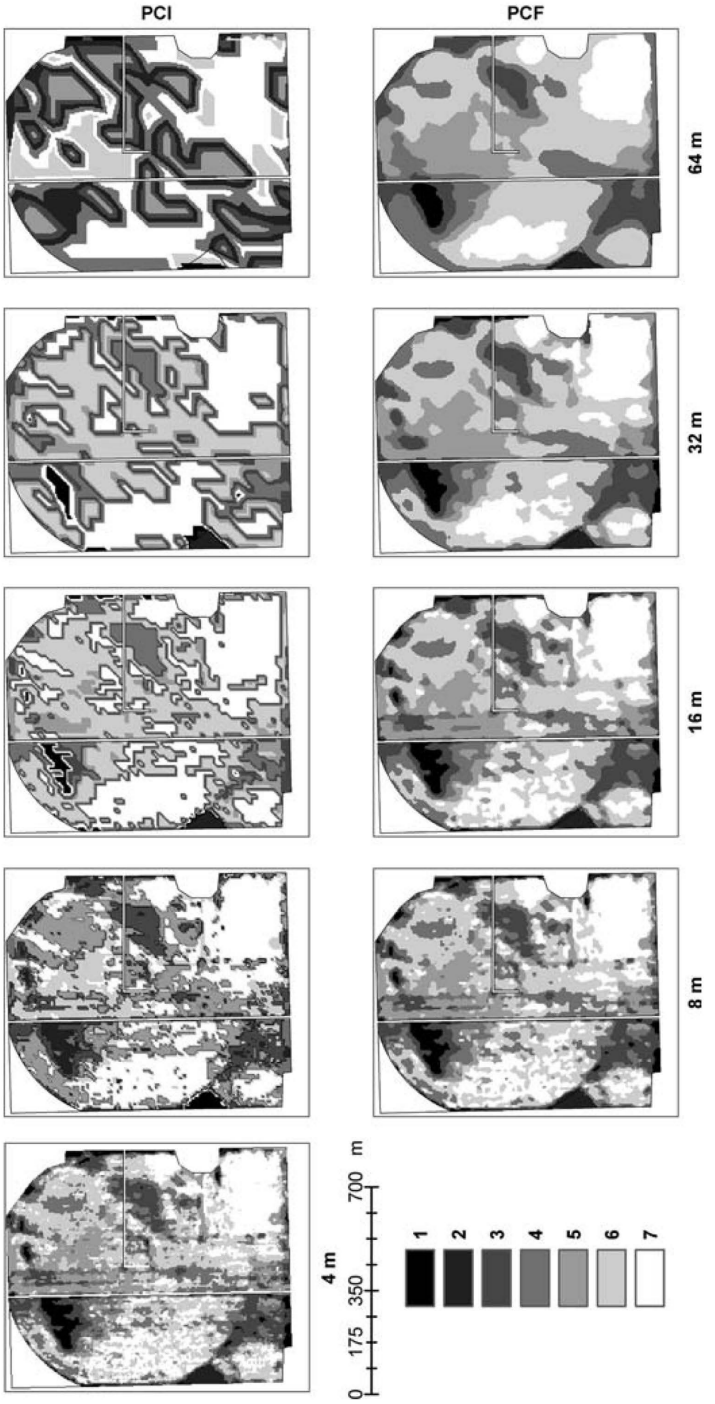

Figure 7. Maps of yield classes as affected by the spatial aggregation method. PCI: prior-classification interpolation with grid cell size ranging from 4 to 64 m. PCF: post-classification filtering with grid cell size 4 m and filtering window size ranging from 4 to 64 m. All maps were derived from fuzzy *k*-means cluster analysis of individual years (AY) as input data.

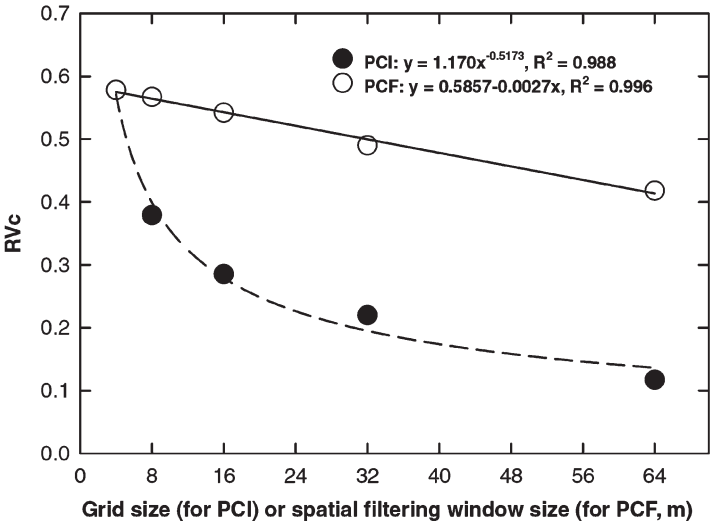

Figure 8. Effect of different spatial aggregation techniques on the average yield variance accounted for by the classification (RV_c, %/100). PCI: prior-classification interpolation with grid cell size ranging from 4 to 64 m. PCF: post-classification filtering with grid cell size 4 m and filtering window size ranging from 4 to 64 m. Values refer to fuzzy-*k*-means clustering using individual years (AY) as input data.

Table 3. Effect of post-classification filtering with 64-m window size on the mean relative yield, standard deviation (SD), coefficient of variation (CV, %), and the proportional area (% of whole field) of yield classes. Different letters show significant differences of the means of yield classes based on Duncan's multiple range test

Method	Class	Mean	SD	CV	Area
Base (4-m grid)	7	1.066 A	0.023	2.2	22.0
	6	1.022 B	0.016	1.6	29.2
	5	1.002 C	0.026	2.6	13.1
	4	0.972 D	0.020	2.1	19.5
	3	0.901 E	0.028	3.2	11.8
	2	0.836 F	0.063	7.6	1.1
	1	0.775 G	0.059	7.6	3.3
PCI (64-m grid)	7	1.030 A	0.052	5.0	29.6
	6	0.994 B	0.054	5.4	9.3
	5	0.992 B	0.034	3.4	8.7
	4	0.992 B	0.063	6.3	21.0
	3	0.982 C	0.061	6.3	13.0
	2	0.961 D	0.092	9.5	17.6
	1	0.783 E	0.063	8.1	0.8
PCF (4-m grid with 64-m filtering)	7	1.061 A	0.033	3.1	15.1
	6	1.023 B	0.034	3.4	36.3
	5	0.999 C	0.035	3.5	14.0
	4	0.970 D	0.048	5.0	21.3
	3	0.899 E	0.059	6.6	10.6
	2	0.834 F	0.062	7.5	1.1
	1	0.768 G	0.074	9.7	1.7

Discussion

The goal of the procedure outlined in Figure 1 is to map patterns of yield variation that are relatively consistent over time. Like many previous attempts, the proposed methods for yield data processing are not perfect, but they have worked well in the case studies conducted so far. The study at the Bellwood site largely confirmed the results obtained for other sites (Cairo, Clay Center, and Mead) with similar conditions for irrigated maize-soybean production in Nebraska (Dobermann *et al.*, 2003; Ping and Dobermann, 2003; Simbahan *et al.*, 2004), allowing us to draw several more general conclusions.

Errors associated with the harvest process and extreme short-distance yield variation caused by random seasonal events should be filtered out if the objective is to perform a multi-year analysis of yield patterns. The latter may be caused by measurement error, crop management or other events that are not related to the broader spatial patterns of crop yield variation. Often it remains uncertain whether extremes in yield data reflect true low or high yields or artifacts due to field management and harvest. This presents a challenge for any screening program. The raw data screening algorithm used here eliminated erroneous yield values based on a logical, sequential order of data screening, with a minimum of empirical limits imposed. It is likely to be robust enough to obtain more accurate yield maps that better illustrate the major spatial patterns of yield variation. The algorithm provided consistent results in terms of (i) what data were removed, (ii) the proportions of the different screening steps, and (iii) improvement in yield map precision (Simbahan *et al.*, 2004). Whether the algorithm removed all erroneous or uncertain yield data cannot be fully assessed because selecting accurate validation criteria remains a challenging issue. Further improvements of the yield data screening algorithm are possible by improving methods for specifying grain flow delays, changing the configuration or criteria used in the local neighborhood search, or adding criteria for detecting errors due to varying swath width or overlap of harvest passes (Beck *et al.*, 1999). The latter was not an issue in the row crops used in our studies. Using more sophisticated methods for grain flow delay correction to reduce the amount of yield smoothing that occurs when a crop is combine-harvested are of particular interest for studies in which very accurate yield measurements at fine spatial resolution are required. Instead of using fixed lag times for grain flow delay correction (as in most commercial yield mapping software), impulse response models can be used to reverse the smoothing behavior that is typical for yields measured with combines (Whelan and McBratney, 2000; Whelan and McBratney, 2002; Lark and Wheeler, 2003).

Spatial variation in crop yield data is mainly a function of climate, indigenous variation in soil productivity, field management and measurement error. If climate variation has less effect on crop growth (e.g., in irrigated agriculture), management is consistent and mapping errors are small and mostly random, only few years (perhaps about 5 years) of yield maps are required for a reliable yield classification. In that case, using mean relative yield (MY) in combination with any clustering method is often a reasonable choice for yield classification (Dobermann *et al.*, 2003). However, fuzzy *k*-means clustering tends to be more sensitive to the

choice of input data than, for example, hierarchical cluster analysis methods, which may be beneficial in environments with greater yield variability. Where management changes more frequently or greater climatic variation affects crop yield variability within a field and from year to year, longer time series of yields maps (perhaps 5–10 years) are required and yield classification should be done based on fuzzy *k*-means classification of AY data to properly account for this variability. Such conditions are typical for rain-fed agriculture, but they also occurred at the irrigated Bellwood site because changes in irrigation methods, row direction and tillage practices caused significant inter-annual variation in yield patterns (Figure 4).

In general, the optimal number of yield classes will vary among sites, depending on the determinants of yield variation and the purpose of yield mapping and classification. At the Bellwood site and two other irrigated maize fields in similar environments of Nebraska (Dobermann *et al.*, 2003), six to seven yield classes established by cluster analysis provided sufficient resolution of the spatio-temporal yield variability observed. This may be a typical range for relatively flat, irrigated fields of this size (about 65 ha), in which soil variation as a key yield determinant is largely overwritten by sufficient water and nutrient supply. The optimal number of yield classes may vary more widely in different environments, but RV_c as estimated here is a useful evaluation criterion for comparing different classification choices.

Yield zones used for site-specific management should display larger, spatially contiguous areas, which reflect major, and consistent differences in attainable yield, not noise introduced by annual factors and artifacts in a yield map. A first option for achieving this is to create yield maps at a relatively coarse grid size. Depending on the mapping purpose and scale a farmer wishes to manage, square grid sizes used of interpolated yield maps are often in the 10–50 m range (Lark and Stafford, 1998; Taylor *et al.*, 2001). However, the results obtained at Bellwood (Figures 7 & 8) and two other sites (Ping and Dobermann, 2003) suggest that unrealistic patterns of yield zones may result if the actual clustering is done using maps with large cell sizes. Likewise, many variable rate management technologies do not require large rectangular shapes for accurate equipment performance. Therefore, we recommend that the primary annual yield mapping be done at relatively fine spatial resolution (e.g., grid sizes of about 5 m or even less), followed by spatial classification and/or filtering to create smoother and perhaps more irregularly shaped map units.

Spatially weighted clustering techniques (Oliver and Webster, 1989) could be used for creating maps of yield classes that contain little noise and in which map units are spatially contiguous. However, these methods are difficult to implement for large datasets of spatially dense information.

Instead, we propose to map and classify yields at fine spatial resolution, followed by post-classification spatial filtering using the PCF algorithms applied in our studies (Ping and Dobermann, 2003) or other smoothing techniques. We recommend that window sizes for spatial filtering of yield maps should be in the 30–60 m range. Depending on the filtering technique, the location-specific nature of yield variation, how much loss of information is acceptable and how large the desired

yield zones should be, a window size of more than 60 m may result in significant loss of information (Figure 8). A window size smaller than about 30 m may result in too many patches in the map (Figure 7), including many smaller ones that cannot be managed effectively.

While the filtering techniques used in our studies significantly reduced spatial fragmentation in yield maps, the method is not optimal in a statistical sense because only the class membership is used, whereas individual properties pertaining to a group are ignored. Lark (1998) provided a more objective approach for post-classification smoothing. After performing a fuzzy classification of the data, the membership values in each class for each individual were subjected to spatial smoothing. The fuzzy membership values were replaced by a spatially weighted average of the membership values within a local neighborhood. The weight ascribed to the membership of individual j when computing for the smooth membership for individual i was proportional to their spatial dependence. Larks' method results in spatially less fragmented classes and it is suited for large datasets, but the choice of the size of the neighborhood is critical since a larger neighborhood will have a stronger smoothing effect. In future research, Lark's method should be compared with the PCF technique proposed by us.

Conclusions

Spatially varying yield goals are used in many site-specific management prescriptions. Because yield zones should mainly represent the stable site yield potential, they should be delineated as larger, spatially contiguous areas within a field. The sequence of procedures shown in Figure 1 allows doing this and it has proven to be robust, yielding similar results at several irrigated sites in Nebraska. Post-classification spatial filtering of maps of yield categories established by cluster analysis removed map fragmentation, thereby creating maps of yield classes that were composed of contiguous map units. The original map resolution was maintained and little loss of the yield variance accounted for occurred. In contrast, interpolating yield maps to a coarse grid size before the classification leads to erroneous maps of yield classes and significant loss of information.

More testing with other crops, in other environments, and with various yield monitor brands should be conducted. Further improvements of the yield screening algorithm could be possible by improving methods for specifying grain flow shifts, changing some of the test criteria used, or adding other criteria for detecting errors due to varying swath width or overlap of harvest passes. Guidelines for spatial yield classification should be established for other key environments and cropping systems. Inclusion of other data layers that affect yield (soil, topography, etc.) is needed for developing crop management zones and making management decisions. Procedures such as those shown in Figure 1 should be better implemented in commercial farm software.

Acknowledgments

We thank Jerry Mulliken and Duane Siffring for providing the yield monitor data used in this study. This material is based on research supported by the Hatch Act, the USDA-CSREES/NASA program on Application of Geospatial and Precision Technologies (AGPT, Grant No. 2001-52103-11303) and the U.S. Department of Energy: a) EPSCoR program, Grant No. DE-FG-02-00ER45827 and b) Office of Science, Biological and Environmental Research Program (BER), Grant No. DE-FG03-00ER62996. A contribution of the University of Nebraska Agricultural Research Division, Lincoln, NE. Journal Series No. 14409.

References

Arslan, S. and Colvin, T. S. 2002a. An evaluation of the response of yield monitors and combines to varying yields. *Precision Agriculture* 3, 107–122.

Arslan, S. and Colvin, T. S. 2002b. Grain yield mapping: yield sensing, yield reconstruction, and errors. *Precision Agriculture* 3, 135–154.

Beal, J. P. and Tian, L. F. 2001. Time shift evaluation to improve yield map quality. *Applied Engineering in Agriculture* 17, 385–390.

Beck, A. D., Roades, J. P. and Searcy, S. W. 1999. Post-process filtering techniques to improve yield map accuracy. ASAE Paper No. 99-1048. (ASAE, St. Joseph, USA).

Blackmore, B. S. 2000. The interpretation of trends from multiple yield maps. *Computers and Electronics in Agriculture* 26, 37–51.

Blackmore, B. S. and Moore, M. 1999. Remedial correction of yield map data. *Precision Agriculture* 1, 53–66.

Boydell, A. and McBratney, A. B. 2002. Identifying potential within-field management zones from cotton-yield estimates. *Precision Agriculture* 3, 9–23.

Dobermann, A. and Ping, J. L. 2004. Geostatistical integration of yield monitor data and remote sensing improves yield maps. *Agronomy Journal* 96, 285–297.

Dobermann, A., Ping, J. L., Adamchuk, V. I., Simbahan, G. C. and Ferguson, R. B. 2003. Classification of crop yield variability in irrigated production fields. *Agronomy Journal* 95, 1105–1120.

Evans, L. T. 1993. *Crop Evolution, Adaptation, and Yield* (Cambridge University Press, Cambridge, UK), p. 500.

Lark, R. M. 1998. Forming spatially coherent regions by classification of multivariate data. *International Journal Geographical Information Science* 12, 83–98.

Lark, R. M., Bolam, H. C. and Stafford, J. V. 1997. Limits to the spatial resolution of yield mapping systems for combinable crops. *Journal of Agricultural Engineering Research* 66, 183–193.

Lark, R. M. and Stafford, J. V. 1997. Classification as a first step in the interpretation of temporal and spatial variation of crop yield. *Annals of Applied Biology* 130, 111–121.

Lark, R. M. and Stafford, J. V. 1998. Information on within-field variability from sequences of yield maps: Multivariate classification as a first step of interpretation. *Nutrient Cycling in Agroecosystems* 50, 277– 281.

Lark, R. M. and Wheeler, H. C. 2003. A method to investigate within-field variation of the response of combinable crops to an input. *Agronomy Journal* 95, 1093–1104.

Leica Geosystems 2003. *ERDAS IMAGINE® Field Guide,* 7th edn. (Leica Geosystems, Atlanta, GA, USA).

Minasny, B. and McBratney, A. B. 2003. *FuzME version 3.0.* (Australian Centre for Precision Agriculture, The University of Sydney, Sydney). http://www.usyd.edu.au/su/agric/acpa

Minasny, B., McBratney, A. B. and Whelan, B. M. 2002. *VESPER version 1.5.* (Australian Centre for Precision Agriculture, The University of Sydney, Sydney). http://www.usyd.edu.au/su/agric/acpa

Noack, P. H., Muhr, T. and Demmel, M. 2003. An algorithm for automatic detection and elimination of defective yield data. In: *Precision agriculture. Proceedings of the 4th European Conference in Precision Agriculture,* edited by J. V. Stafford and A. Werner (Wageningen Academic Publishers, Wageningen, The Netherlands), p. 445–450.

Oliver, M. A. and Webster, R. 1989. A geostatistical basis for spatial weighting in multivariate classification. *Mathematical Geology* 21, 15–35.

Ping, J. L. and Dobermann, A. 2003. Creating spatially contiguous yield classes for site-specific management. *Agronomy Journal* 95, 1121–1131.

Roubens, M. 1982. Fuzzy clustering algorithms and their cluster validity. *European Journal of Operational Research* 10, 294–301.

SAS Institute Inc. 1999. *SAS System Version 8.0* (SAS Institute Inc., Cary, NC, USA).

Simbahan, G. C., Dobermann, A. and Ping, J. L. 2004. Screening yield monitor data improves grain yield maps. *Agronomy Journal* 96, 1091–1102.

Soil Survey Staff 1999. *Soil Taxonomy: A Basic System of Soil Classification for Making and Interpreting Soil Surveys* (USDA-NRCS, Washington, DC, USA), p. 869.

Stafford, J. V., Ambler, B., Lark, R. M. and Catt, J. 1996. Mapping and interpreting the yield variation in cereal crops. *Computers and Electronics in Agriculture* 14, 101–119.

Taylor, R. K., Kluitenberg, G. J., Schrock, M. D., Zhang, N., Schmidt, J. P. and Havin, J. L. 2001. Using yield monitor data to determine spatial crop production potential. *Transactions of the ASAE* 44, 1409– 1414.

Webster, R. and Oliver, M. A. 1990. *Statistical Methods in Soil and Land Resource Survey* (Oxford Univ. Press, Oxford, UK), p. 316.

Whelan, B. M. and McBratney, A. B. 2000. An approach to deconvoluting grain-flow within a conventional combine harvester using a parametric transfer function. *Precision Agriculture* 2, 389–398.

Whelan, B. M. and McBratney, A. B. 2002. A parametric transfer function for grain-flow within a conventional combine harvester. *Precision Agriculture* 3, 123–134.

Yang, H. S., Dobermann, A., Lindquist, J. L., Walters, D. T., Arkebauer, T. J. and Cassman, K. G. 2004. Hybrid-Maize—a maize simulation model that combines two crop modeling approaches. *Field Crops Research* 87, 131–154.