YIELD MONITOR DATA ANALYSIS PROTOCOL: A PRIMER IN THE MANAGEMENT AND ANALYSIS OF PRECISION AGRICULTURE DATA

by

Terry W. Griffin[a], Jason P. Brown[b], and Jess Lowenberg-DeBoer[c]

[a] Assistant Professor and Extension Economist, Department of Agricultural Economics and Agribusiness, Cooperative Extension Service, Division of Agriculture, University of Arkansas

[b] Graduate Research Assistant, Department of Agricultural Economics, Purdue University

[c] Professor, Department of Agricultural Economics, and Associate Dean and Director, International Programs in Agriculture, Purdue University

Version 2 – June 2007

Department of Agricultural Economics and Agribusiness
Cooperative Extension Service
Division of Agriculture
University of Arkansas

Department of Agricultural Economics
College of Agriculture
Purdue University

Keywords: yield monitor data, spatial analysis, GIS, ArcView, ArcMap, precision agriculture

Table of Contents

List of Figures

Yield Monitor Data Analysis Protocol: A Primer in the Management and Analysis of Precision Agriculture Data

Please send any comments, suggestions and questions to:
Terry Griffin (spaceplowboy@gmail.com) 501.249.6360.

This document is available on-line and can be cited as:

Griffin, T.W., Brown, J.P., and Lowenberg-DeBoer, J. 2007. Yield Monitor Data Analysis Protocol: A Primer in the Management and Analysis of Precision Agriculture Data. Site Specific Management Center Publication. Purdue University, West Lafayette, Indiana.

Executive Summary

This document serves to share our techniques for managing the analysis of site-specific precision agriculture data for the purposes of analyzing field-scale on-farm trial experiments. The content of this document is the culmination of over a decade of on-farm trial and spatial analysis experience which continually expands. Working with precision agriculture data can be very frustrating even for those versed in GIS and programming. This protocol was written to assist other analysts of precision agriculture data to be able to follow our steps in an effort to reduce their frustration by making available techniques that have worked for us.

It has been our experience that researchers, farmers, and consultants have interest in performing yield monitor data analysis. Researchers may include agricultural economists, agronomists, pathologists, agricultural engineers, and other scientists. Farmers and their advisors have attended workshops meant to train participants in the use of yield monitor data for whole-farm decision making (Erickson, 2005; Nistor and Florax, 2007). This document is intended for those conducting field research, whether it is the farmer, consultant, or professional researcher.

This version of the protocol has been enhanced by expanding the description of the GIS steps previously conducted in ESRI ArcView GIS 3.3 to include ESRI ArcGIS 9.X in a separate chapter. There is a great deal of overlap between the two chapters and either one of the two are meant to be read, but not both unless the reader has interests in using both software packages. Most readers will favor one or the other software and will only want to read the respective chapter. Therefore, this second version may replace the first version by accommodating users of both ArcView 3.3 and ArcMap 9.X. With either GIS software, the final data needs to be in a Shapefile format for statistical analysis. This version also migrated between statistical software packages. The current version only refers to GeoDa and R for statistics and has omitted reference to SpaceStat for spatial statistical analysis. We suspect most new users will be using the open source and free software.

Another version of this protocol that does not rely upon advanced GIS software but upon farm level mapping software may be written once the procedures described in this document can be

feasibly performed using a single farm level software package. Recent developments from the software community have allowed farm-level software to perform many of the GIS tasks described in this document to manage yield monitor data in addition to shelling out to USDA-ARS Yield Editor.

This document also gives specifics to introduce the reader to spatial statistical analyses for analyzing site-specific yield monitor data rather than using traditional and albeit less efficient analysis. In the presence of spatially variable data, traditional forms of analysis such as non-spatial analysis of variance (ANOVA) and least squares regression are unreliable and should be avoided. To our knowledge, this document provides the most appropriate analysis methods for field-scale research with yield monitor data.

Much of the following text and examples are useful for a wide range of precision agriculture applications, but the overall thrust of this document is intended for analyzing planned field-scale experiments. To conduct spatial analyses of yield monitor data both 1) a good experimental design and 2) a planned comparison must be in place. A planned comparison can also be called a testable question or testable hypothesis. If there is no hypothesis, neither traditional nor spatial analysis can be conducted for valid inference.

Although we recommend not using spatial interpolation techniques to create explanatory variables for conducting inferential statistics, we do not make any statements on the use of these smoothing techniques for prescription maps, defining management zones or other common uses.

The authors assume the reader has a working knowledge of GIS, spreadsheets and farm-level mapping software. The steps in this document were conducted in Windows XP Pro operative system and MS Excel 2003. MS Excel 2007 does not support saving files in the *.dbf file format. The reader is invited to make suggestions and comments that may be incorporated into future versions of this document.

Dispelling Myths of Field-Scale Experimentation

A number of common fallacies exist and continue to be brought forth by field-scale researchers whether by farmers, researchers, or analysts. Common misconceptions or myths include:

Myth # 1) "Collecting more dense data prevents analysis problems from spatial variation"
Myth # 2) "If enough replications with split planter trials, variability will be negated"
Myth # 3) "Each yield monitor data point is a replication"
Myth # 4) "Small plot experimental designs and analysis are sufficient at field-scales"
Myth # 5) "Farmers do not see the value in on-farm field-scale experimentation"
Myth # 6) "Averages of yield monitor data by treatment gives the information needed"

Although each of the above-mentioned misconceptions will not be addressed in this protocol, it is important to have a valid understanding of the types of data associated with precision agriculture and what can and cannot be done. Griffin et al. (2005) address some of these issues in the December 2005 Site-specific Management Center Newsletter.

Preface: Overview of Spatial Analysis Steps

The following procedures describe the steps we take in data acquisition, management, and analysis. Chapter 1 describes the methods for data acquisition and data filtering. In nearly all cases, yield monitor data must be filtered before use in inferential analysis. Data assimilation and management with a geographical information system (GIS) is described with several specific treatments of the data illustrated in detail in Chapter 2 and Chapter 3. Data preparation for analysis is explained using standard spreadsheets in Chapter 4. The discussion on exploratory spatial data analysis (ESDA) precedes the discussion on spatial statistical analysis in Chapter 5 and Chapter 6, respectively. Finally, interpretation and economic analyses are described in remaining chapters.

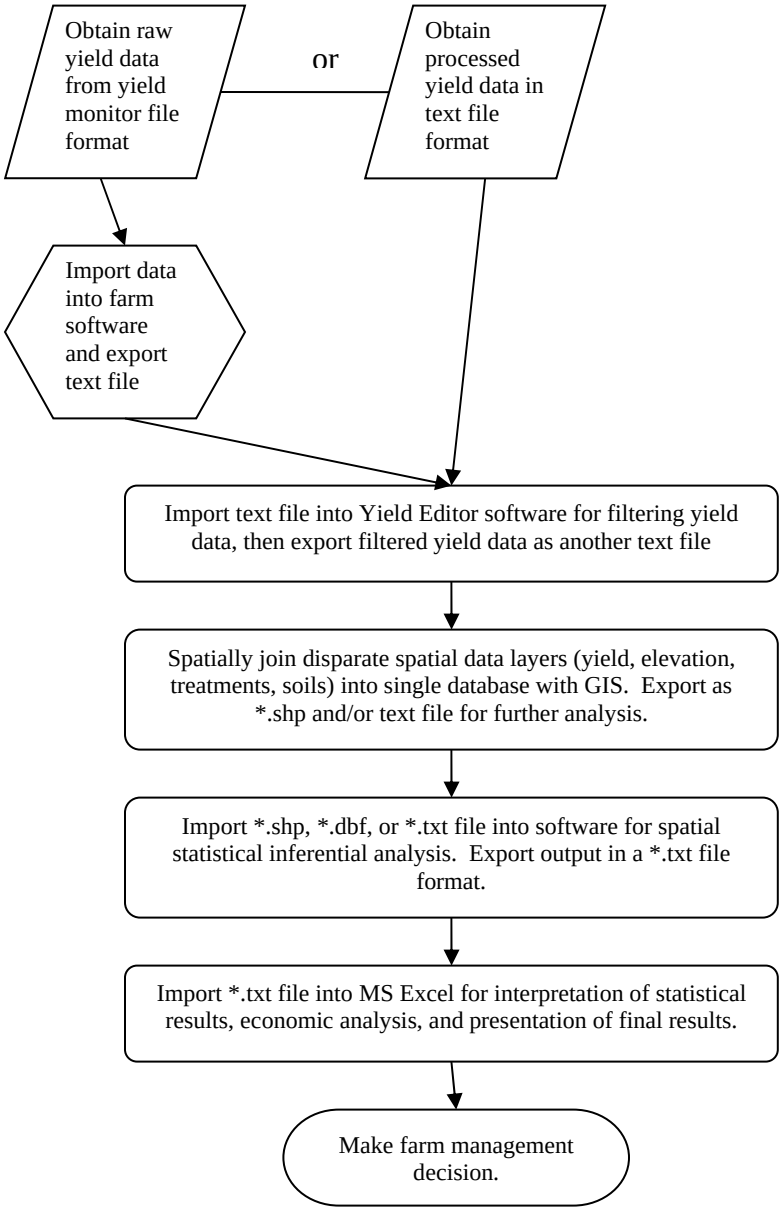


Figure 1. Flow chart of analysis steps in this protocol.

Chapter 1 Yield Monitor Data Preparation

GIGO = “Garbage In, Garbage Out”

"I think there is a world market for maybe five computers."
~ Thomas Watson, chairman of IBM, 1943

Chapter 1 describes the process of acquiring yield monitor data from the raw data and preparing the data for further processing. In most cases, the user should subject the yield monitor data to a filtering procedure to remove potentially erroneous data with software such as Yield Editor from USDA-ARS in Columbia, MO (Drummond, 2006). Yield Editor can be downloaded from the USDA-ARS website which can be found by conducting an Internet search for “ARS Yield Editor”. According to Drummond’s (2006) criteria for importing data, a few steps may need to be conducted to ensure the data is ready to be imported into Yield Editor. This step is easiest if using the yield monitor’s native software package, however this is not always possible especially for yield monitors from other than the major manufacturers. Both scenarios are described.

It should be noted that these data preparation procedures may be referred to as “data cleaning” or “data filtering” but actually are little more than adjusting the location of the observations and removing measurements that are known to be erroneous due to harvester machine dynamics and operator behavior. These data filtering procedures are by no means an unethical modification or manipulation of the data. It is expected that data filtering improves the quality of the dataset.

Discussion on using raw yield monitor data rather than filtering erroneous data
Removing observations from a dataset without some sort of protocol has not been a commonly accepted practice in statistics. Many analysts have omitted outliers by removing ±3 standard deviations of the data or by plotting the data on a scattergram and removing obvious erroneous data caused by factors such as human error, measurement error, or natural phenomena. With the case of instantaneous yield monitor data, it is widely known that many observations have erroneous yield values due to simple harvester machine dynamics. These erroneous observations can be identified by examining harvester velocity, velocity change, maximum yield, and other parameters. With harvester yield data, errors also arise from start and stop delays for beginning and ending of passes. The ramping up and ramping down effects of the harvester yield monitor has adverse effects on yield measurements. The flow delay caused by inaccurate assignment of yield measurement to GPS coordinate location is the effect of grain being harvested at one location, yield measured while harvester is in another location, and recorded with GPS coordinates at potentially another location. The flow delay must be corrected. If this error is not corrected, yield values that are otherwise “good” are at the wrong location. Allowing native software packages to impose the default processing such as 12 second delay may be a good average, but we have typically seen appropriate flow delays of six to 18 seconds but rarely exactly 12. Our assertion is that it is dangerous to use yield monitor data processed using default settings for analysis. Conscious decisions must be made as to the most appropriate handing of the data. Some researchers have argued that data filtering is unethical and prefer to accept data “as is” from the yield monitor, and thus from their farm-level software regardless of the default

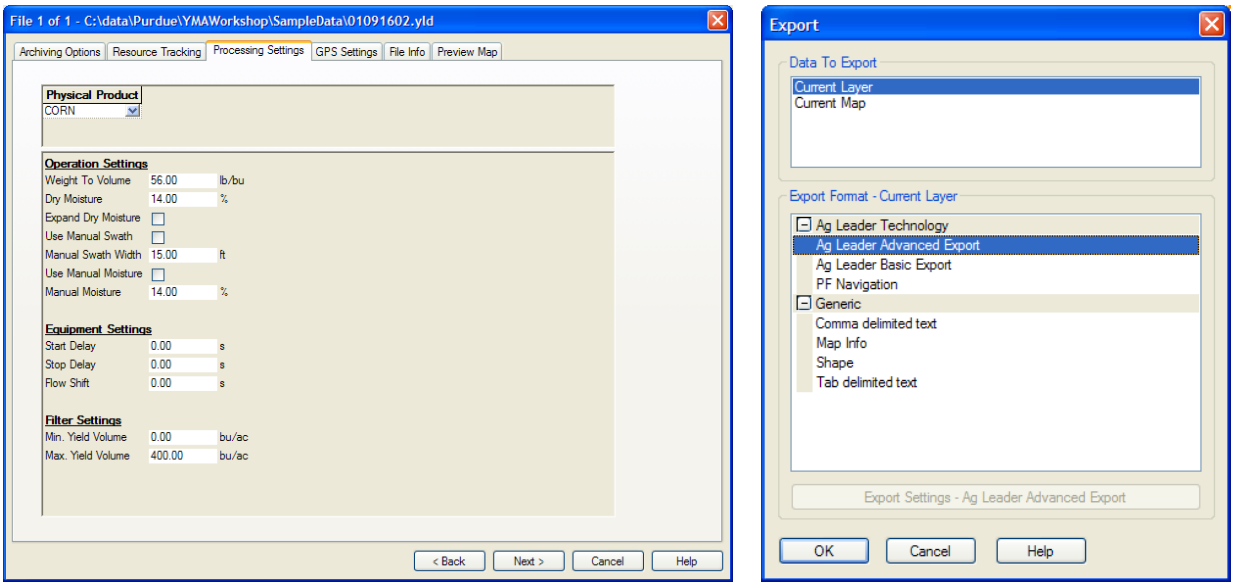
filtering settings. However, it is obvious that accepting this unprocessed "as is" data is not a sound practice.

Using the Farm-Level Mapping Software

We define the "farm-level mapping software" as those software packages intended to be used by farmers and field researchers. In packages such as JDOffice, AgLeader SMS, MapShots EasiSuite, Farmworks and others the default import settings for start delay, stop delay, and flow delay are preset to some predetermined expected average. These settings are typically 4, 4, and 12 for start delay, stop delay, and flow delay, respectively, although some variation between software packages exists. It is our practice to set these to zeros. If there is a minimum and maximum yield; we set these to zero and some value near the maximum physical measurement of the yield monitor, respectively. These settings are chosen so that the native software does not perform its own "filtering" or preprocessing procedures so that more complete control is possible during the filtering protocol. We do not perform any data manipulation procedures in these native software packages other than a simple import of the raw yield monitor data and export of the yield monitor data in a format usable by other software. SMS and JDOffice both have an automatic export function that exports yield monitor data in the appropriate format for Yield Editor, a template can be created in other software. The yield data should be exported twice, once as a text file for Yield Editor and again with all the variables included in a shape file.

AgLeader SMS Software

When importing the raw yield monitor data file, the default processing parameters should be set as shown in Figure 2.



Default settings changed Exporting as Ag Leader Advanced Export

Figure 2. Screen captures from AgLeader SMS during import and export procedures.

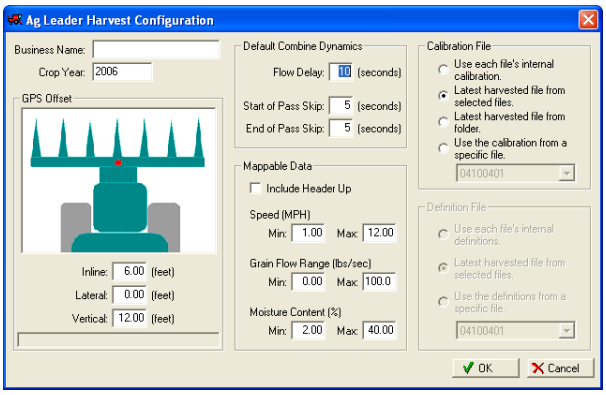
Once the yield data has been imported into SMS, export the data by File: Export: AgLeader Advanced Export (Figure 2). It should be noted that SMS does not have to be a registered installation and can be used after he evaluation period has expired in order to perform the necessary procedures.

JDOffice Software from John Deere

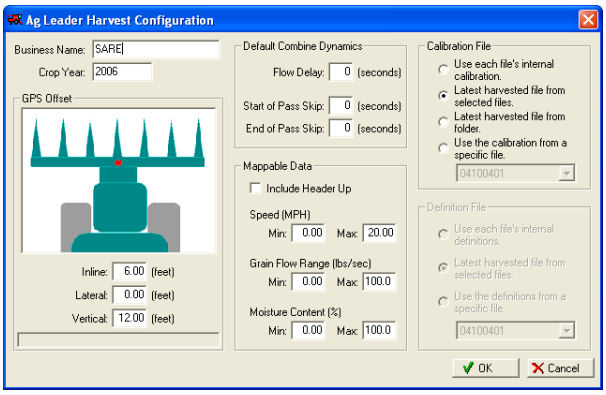
In order to export data in the appropriate format, a one time setting must be made. Go to File:Preferences:Export and click the radio button next to "Text (comma delimited)." This setting will cause the exported data to be in a text file format rather than the Shape file format. Once the yield layer of the field of interest is active, go to File:Export:Layer Data.

MapShots EASiSuite

Similar steps as with SMS and JDOffice can be conducted in MapShots EASiSuite (Figure 3).



Default settings



Default settings replaced by 0

Figure 3. Screen captures from MapShots EASiSuites.

Using Yield Monitor Data in absence of native yield monitor files

Whether the data is already in the ArcView Shapefile format, a georeferenced text file, or other file format, the data can usually be manually converted into the appropriate *.txt file for Yield Editor pending having all the necessary data (columns). The required data columns and arrangement are described in Drummond (2006).

Data that has already been exported

Using the *.dbf file portion of the Shapefile as exported from FarmWorks, SMS, JDOffice, EasiSuite MapShots or other software package has been successful. Care must be taken to know if the flow rates have been exported in kg per second or in the lbs per second as required by Yield Editor (Drummond, 2006). Other measurements with metric or English units must also be identified and converted to English units if necessary. Remaining data columns can be deleted.

Using the manual export features of farm-level software
We save an export template in SMS, FarmWorks, MapShots EASiSuite (others may work, however we do not have extensive experience with other farm-level software) when we export yield data so we can quickly and easily export yield data in the future for Yield Editor. This configuration can be saved as a template and loaded each time data is to be exported.

Removing Erroneous Measurements

If Yield Editor is not being used, the reader may skip directly to the section on GIS; however it is our experience that better farm management decisions are made with data cleaned with Yield Editor. Yield Editor 1.02 Beta (USDA-ARS) (Drummond, 2006) is used to remove erroneous data, i.e., filter the raw yield monitor data. Under a certain set of known harvester characteristics, the yield monitor is unable to make accurate measurements. It is under these conditions that we use Yield Editor to remove data points that are known to have been inaccurately measured. In five of seven on-farm trials evaluated in Griffin's (2006) Ph.D. dissertation differing farm management decisions would have been made based upon yield data subjected to the filtering process with Yield Editor and yield data subjected to the default processing procedure of the farm level mapping software.

Once a dataset is in the appropriate format as per the previous section, it can be imported into Yield Editor (Figure 4). A user-defined or other standard protocol for filtering data can be instated on the yield data but this is not recommended. The analyst's intuition, experience, and skill should guide the procedures. The data points are visually displayed so further manual deletions can be made or points added back into the dataset if needed; thus this is an example of where the analyst's intuition is useful. The data filtering protocol may be farmer or field specific. The best starting point is most likely zeros for all parameters, but this is dependent on how data was managed during the import process in the farm-level mapping software, i.e. if flow delays were allowed to be imposed on the data such as 4, 4, and 12 for start, stop, and flow delays, respectively. However, conscious decisions must be made as to whether the protocols are appropriate for the user's application. It is our experience that no single parameter setting structure is universally appropriate, even with the same harvester and operator. Adjusting flow delay, start pass delay, and end pass delay are the most difficult and may be the most important to the quality of the data. End row yield points should be similar to adjacent end row points. Differences are due to ramping up and down of the harvester at the beginning and end of rows. Field experience indicates that it may take as much as 100 feet of harvester travel before accurate yield measurements can be made. Adjust the delays until the analyst's intuition is satisfied. (Values for the delays will typically be: Flow Delay 8 to 24, Start Pass Delay 0 to 10, End Pass Delay 0 to 16; however variation occurs and parameters are set by trial-and-error plus intuition). Negative values are possible especially if the data were already subjected to processing by the farm-level mapping software. Setting the flow delay is easiest when the operator harvests three to eight passes in one direction and alternates the pattern across the field. This allows a visual reference wide enough to be seen on the Yield Editor map. Alternating direction between individual passes does not give the needed visual reference. In addition, fields with distinct variability such as center pivot irrigation tend to be among the easier fields with which to adjust flow delays.

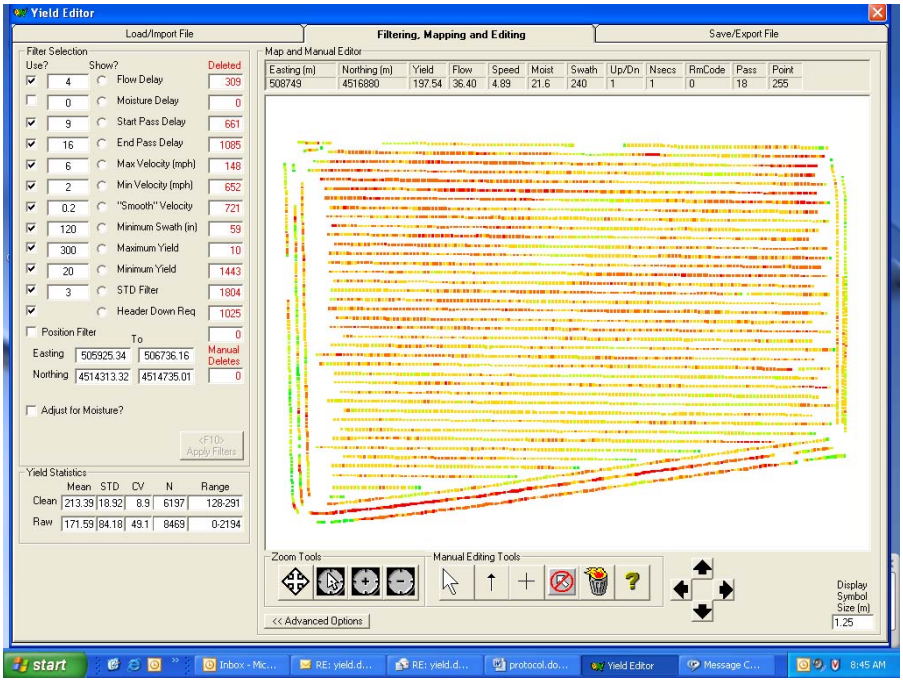


Figure 4. Screen capture of the filtering, mapping and editing tab in Yield Editor.

Once the analyst is satisfied with the data filtering process and has recorded the parameters either by saving the session or manually recording the parameter values in another document, the filtered data can be exported into one of a few file formats. We typically export the data as "space delimited ASCII" to facilitate less total steps before the import into ArcView GIS. When prompted we place a check next to longitude (DD), latitude (DD), and yield under the Save/Export File tab as in Figure 5. Some analysts choose to use UTM Easting (m) and UTM Northing (m) in meters instead of decimal degree coordinates. Other data fields can be selected.

The *.txt file exported from Yield Editor must have the proper column names prior to importing into the GIS. This can be accomplished by a variety of methods, and we describe two, one using a text editor and the second using a spreadsheet software. The user can choose the method that they prefer. The two examples assume that only the latitude, longitude and yield were included in the exported datasets, some users may opt to export additional data; thus additional column headings may be needed.

Option 1: using a text editor
Open the *.txt file with a text editor such as WordPad or NotePad. Add a blank line or row and name the column headings. The column heading names should be separated with only a space. For instance, the columns would read: *lat long yield.* Save this file as a tab delimited *.txt file.

Option 2: using a spreadsheet
Open the *.txt file with a spreadsheet software program and specify "space delimited" if prompted. Add a blank row and label the first, second, and third columns as lat, long, and yield, respectively. Save this file as a tab delimited *.txt file.

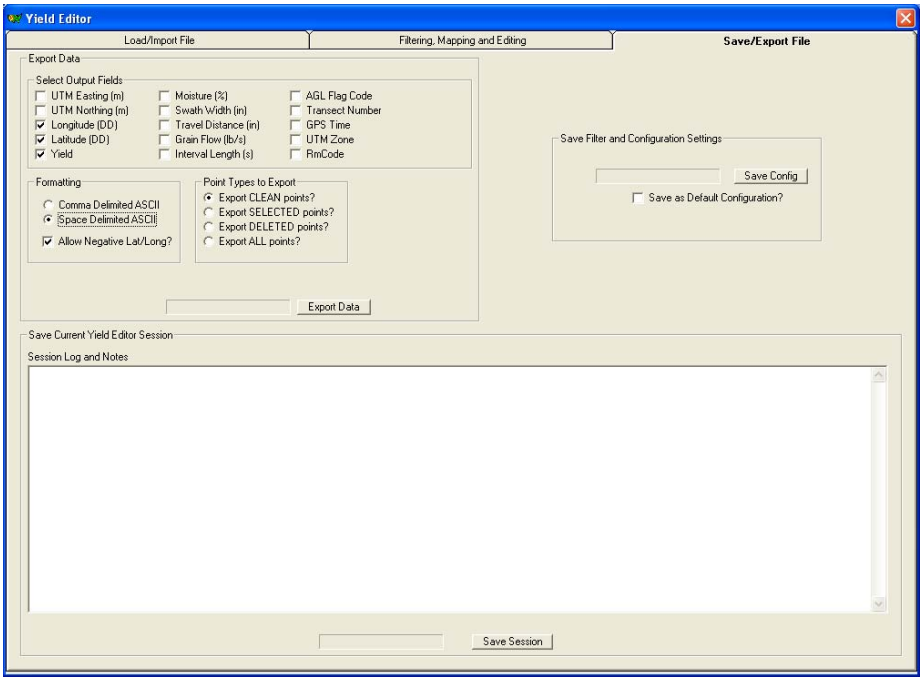


Figure 5. Screen capture of Yield Editor Save/Export File tab.

Separating Data into Different Columns in a Spreadsheet

When using spreadsheets such as MS Excel, we sometimes use a handy trick to convert delimited data into a spreadsheet format. For instance, when the user has Windows Explorer open and right clicks on the space delimited text file and then chooses "Open with" and then click on Microsoft Office Excel, all the data will come into the spreadsheet as one column. If the user selects all the data such that the first column is selected (by holding Shift and Control down on the keyboard and then hold down arrow), the user can go to Data on the main menu and click "Text to Columns . . .". This opens the "Text to Columns Wizard". The user can select the radio button next to "Delimited" and click Next, then put a check mark in the box for the type of delimiter used. In the case of bringing in data from Yield Editor, the choices were comma delimited and space delimited. By convention, we have opted to use space delimited. Then click on "Next" and then "Finish". The data should all be in the proper columns.

Chapter 2 Data Management in ESRI ArcView GIS 3.3

> "Everything is related to everything else, but near things are more related than distant things."
> ~ Tobler's First Law of Geography

Chapter 2 deals with managing the yield monitor and other site-specific data with geographical information systems (GIS) software. This chapter assumes the reader has access to and is a user of ESRI ArcView GIS 3.3. Other professional GIS software including other versions of ESRI software such as ArcGIS and ArcMap (the use of ArcMap 9.X for these procedures is described in Chapter 3) is capable of performing the same or similar tasks. Some farm-level mapping software may have incorporated enough GIS functionality to perform these tasks although are not described in this protocol.

Assimilate Data using ArcView GIS 3.3

Once ArcView GIS is open, add the *.txt file. From the Project Window in ArcView GIS, select Table and click on Add. Navigate to where the *.txt file is saved and select it. Go to the View to visualize the data and click View: Add Event Theme, specify the *.txt file and assign the X and Y fields. Now that the *.txt file is loaded into the GIS, make sure it appearscorrectly in the expected location with expected yield variation patterns similar to the variation in the final Yield Editor map window (**Figure 4**). Depending upon which column variables were selected in Yield Editor to export, your dataset will have differing pieces of data. At the very least, you will have X and Y coordinates and the yield. The *.txt file will need to be converted to a Shapefile format (Theme: Convert to Shapefile). Treatments, covariates, dummy variables and topographical information will need to be added to this Shapefile in the GIS.

Adding Disparate Spatial Data Layers with Spatial Joins in ArcView GIS 3.3

Once the yield data is in the Shapefile format and has been adjusted for spatial location and erroneously measured observations have been deleted, information from the original yield data file such as elevation can be added to the new yield data Shapefile. A spatial join is conducted to append the pertinent information from the original yield data Shapefile to the new yield Shapefile. The original yield data Shapefile was the data exported from the farm level mapping software package. The column fields that may be important to include in the final dataset may include information from the original yield data file or other site-specific data including elevation, treatment information, and covariates such as electrical conductivity.

Aggregating dense data to the least dense data in ArcView GIS 3.3

Rarely ever do the differing data layers share the same spatial resolution or density, so some sort of aggregation of the data is necessary. We have chosen the following process to minimize the interference of the statistical reliability. Yield data is typically the most dense, followed by soils such as electrical conductivity or other scouting information. Soil sampling for chemical analysis tends to be the most sparsely collected data, such that it may be too sparse to even be

included in the data. It has been our practice to keep the data in the format that it was original measured with the least dense dataset as the basis for the remaining data layers. We caution the analyst not to conduct spatial interpolations via kriging or other geostatistical methods to remedy the dilemma of spatially disparate spatial data layers. If spatial interpolation is used to convert the data points into a smooth surface, a systematic error is introduced into the data, causing a problem in deriving inference (Anselin, 2001), therefore we have avoided spatial interpolation when possible especially for soil characteristics measured at relatively sparse densities. Common spatial interpolation methods are not limited to: kriging, spline, inverse distance, and minimum curvature.

There are a number of ways to assimilate relatively denser data with relatively less dense data, i.e. yield data with soil sample data. Some sort of spatial polygon structure can be assigned to the dataset with each sparse soil data point being attributed to a single grid unit. Our preferred method is to create a polygon such as a circle with given radius with the less dense point as the center using the XTools (DeLaune, 2001) extension for Arcview GIS and is explained in the following section. A specialized form of grid cells known as Thiessen polygons can be created in GIS or GeoDa (University of Illinois) (Anselin, 2003) for the same purpose. GeoDa can be downloaded from: https://www.geoda.uiuc.edu/ and Thiessen polygons created by clicking Tools:Shape:Points to Polygons. Thiessen polygons are a form of nearest neighbor interpolation created by surrounding each input point with an areal unit such that any location within that area is closer to its original point than any other point. Thiessen polygons are sometimes called or very similar to Voronoi polygons, Delaunay Triangles, and Dirichlet Regions. A regular grid can also be used, but it is difficult to spatially align irregular spaced data in a one-to-one format.

Creating buffer areal units for sparse data in ArcView GIS 3.3

With the sparse data layer projects in the chosen distance units in ArcGIS, go to XTools: Buffer Selected Features and choose the measurement unit of your choice (Figure 6), choose the most sparse layer you intend to use (Figure 7), give the theme a name when prompted, choose Buffer Distance, assign a buffer distance in your units of choice (Figure 8), and select Noncontiguous (Figure 9). The buffer distance should be chosen as to 1) not overlap into areas of different treatments (or alternatively further processed to omit observations from different treatments), 2) be large enough to have at least one yield observation if possible, and 3) be small enough to only include yield data that are comparable with other yield data in buffered zone and are representative or affected b y the sparse data. A new Shapefile layer with circular areal units around each of the sparse points is ready for the dense data points to be added. These circular area units may overlap or even include the same dense data point in two different buffer areas, but that is not of concern.

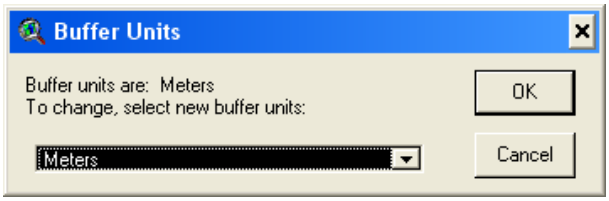


Figure 6. Screen capture of intermediate dialogue box to selected “meters” as the buffer unit.

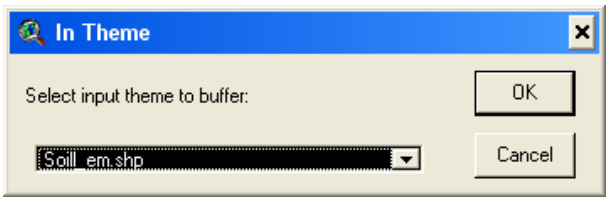


Figure 7. Screen capture of intermediate dialogue box to create a buffer.

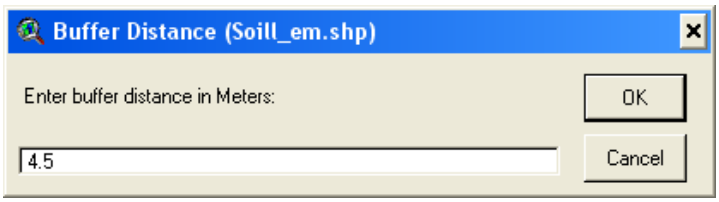


Figure 8. Screen capture of the buffer distance dialogue entered as 4.5 meters.

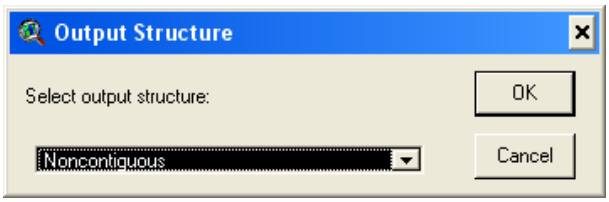


Figure 9. Screen capture of Output Structure with “Noncontiguous” selected.

Assigning dense yield data to sparse data points in ArcView GIS 3.3

Once polygon areal units have been created around the soils data, yield data can be assigned to the soil location. The USGS Point Stat Calc (Dombroski) extension for ArcView GIS is useful in simplifying this step. Select the dense yield data theme and the areal unit theme for the less dense soils data as described in the previous discussions on buffered zones, and make sure both themes are active in the View. Select the value of interest for the point data (yield) and select all the statistics you wish to use (**Figure 10**). We typically only use “Average”; however other descriptive statistics may give indication of the appropriateness of the buffered distance for the given spatial variation.

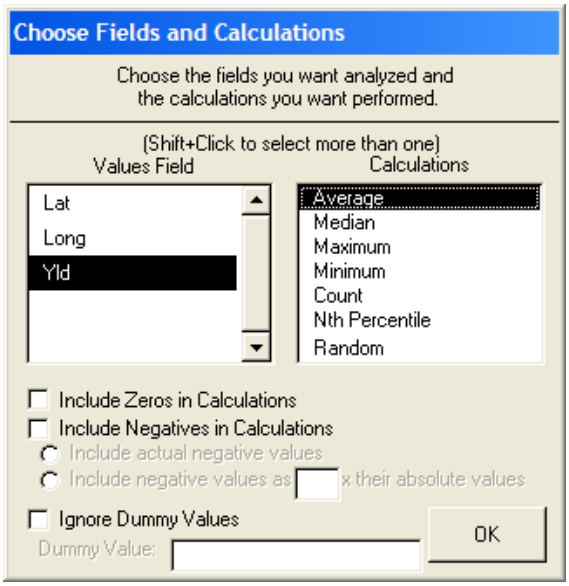


Figure 10. Screen capture of PointStatCalc for ArcView GIS.

After clicking "OK", you will be prompted to provide a file name and decide whether you want to create a new table or use the existing table. We typically accept all the default parameters. Click "OK" again when prompted. Processing may take several minutes to a few hours depending on the size and scale of the datasets and the computation power. Click "OK" when done. The new yield averages have been added to the buffer polygon theme. Similar steps will need to be conducted to append the soils data to the soils buffered polygon theme. These polygon areas need to be converted back into single points. This can be done either 1) by using the original coordinates or 2) adding the centroid X and Y coordinate to the dataset, opening the *.dbf of the buffered polygon theme with the Table command in the Project window as described earlier in the section on adding the *.txt file from Yield Editor to ArcView GIS. Now that the data from the dense and sparse data layers are in a single point data layer with the same resolution as the original sparse dataset, the data are useful for inferential statistical analysis.

Digression on the Modifiable Areal Unit Problem

The procedure discussed in the previous section leads to a digression on the Modifiable Areal Unit Problem (MAUP) which is common across many disciplines (Gotway and Young, 2002). The size of the buffer around the sparse data point must be chosen via a conscious decision and not arbitrarily. With the given mean and standard deviation descriptive statistics, a coefficient of variation (CV) can easily be calculated by dividing standard deviation by the mean. The CV, or even the standard deviation, can be used to graphically represent the variation on a map. This visual representation can be useful to decide if the variance is stable over space and time, assuming there are multiple years of data. The CV is useful when comparing yields from different crops across years in the same field.

Appending treatment information to the dataset in ArcView GIS 3.3

Treatment information may need to be added to the data file. If this information is not already present in the dataset, it can be added in a number of ways. For instance, if the treatments occur in blocks like tillage treatments or split field trials, polygons can be created and merged together to form the treatment polygon map. From this polygon, a specific treatment can be selected. Once the treatment is selected from the treatment polygon map, a Select by Theme can be done on the yield data points with respect to the selected portion of the treatment polygon map. Now that the yield data points associated with the treatment are selected, a dummy variable can be added using the SpaceStat Extension to ArcView GIS (TerraSeer) (Anselin, 1999). To add a dummy variable, click Data, Add Dummy and give an appropriate name. A "1" is added in this column for selected features and a "0" otherwise. These same steps can be done to add a dummy for soil series, other regions such as old feedlots, pastures, homesteads, and two existing fields were joined to be one large field. A dummy variable should be added for each categorical treatment, soil zone, and every measurable discrete factor to be included in the statistical model.

Adding the distance from a given attribute in ArcView GIS 3.3

In some cases, a distance variable may be useful to help describe variability from isotropic or anisotropic effects. In cases of furrow irrigation where plants near the water canal will surely get more water, albeit colder water, than plants at the other end of the row, differing yield responses are expected. Distances are also useful in modeling the isotropic effect of flood irrigated rice production where plants near the water source tend to have lower yields due to the colder temperature of the ground water near the well. The distance to the given attribute can be added to the dataset in a number of ways. One method involves using the Distance Matrix extension for ArcView GIS (Jenness, 2005a). The output from Distance Matrix can be joined into the existing dataset by the standard table joining techniques in ArcView GIS.

Elevation, slope, aspect and associated problems in ArcView GIS 3.3

Due to introduction of variability problems associated with geostatistical techniques (Isaaks and Srivastava, 1989) and imperfect information on proper parameters to assign to interpolation methods, spatial interpolation methods such as inverse distance weighting, kriging, and co-kriging have been avoided. However, if slope or aspect variables are desired, the elevation data must be interpolated into a surface. In addition, the elevation data must be collected at a resolution sufficient to describe the topography and with adequate accuracy. Tractors equipped with RTK automated guidance typically provide sufficient data during plating or other field operations. Coast Guard and WAAS DGPS do not always provide the needed accuracy. Additional data points and resolution are not substitutes for accuracy. An alternative to including slope derived from interpolated elevation surface is to use relative elevation as described in Lowenberg-DeBoer et al. (2006).

A slope, aspect, or other topographic surface can be calculated from the elevation surface. The slope surface can be converted into a contour line vector with base of zero and interval of 0.25 percent. The yield data can be appended with a value for slope by choosing the closest slope contour line by using the Spatial Join function in Geostatistical Wizard in ArcView GIS. The danger in spatial interpolation of a surface is the introduction of variability or in other words

introducing a systematic variable which causes problems with statistical inference (Anselin, 2001).

Removing duplicate points in ArcView GIS 3.3

It may be necessary to remove duplicate points in the data. For instance, GeoDa does not allow points with the same coordinate. If this is a problem, the Find Duplicate Shapes or Records extension (Jenness, 2005b) can be used in ArcView GIS. When using this extension, the analyst is asked to give the name of the theme and unique identifier (**Figure 11**), the criteria for defining duplicates (Figure 12), and is provided a report of the duplicated points and which points were removed (Figure 13). Adding a unique identifier is discussed later in the section on spreadsheets. If a variable is to have unique values such as the identifier, then the Clean Shapefile can be used with the SpaceStat extension to ArcView GIS (Anselin, 1999).

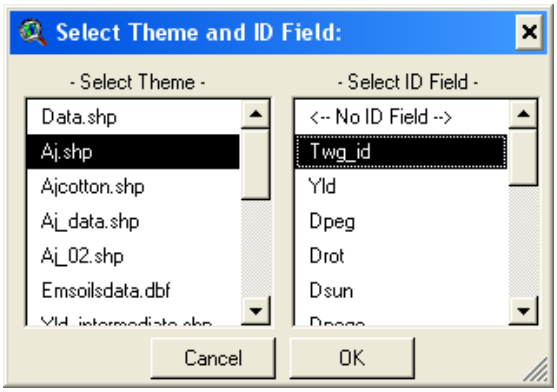


Figure 11. Screen capture of Find Duplicate Shapes or Records in ArcView GIS 3.3.

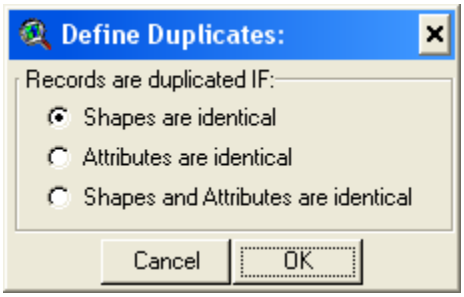


Figure 12. Screen capture of selecting duplicate criteria in ArcView GIS 3.3.

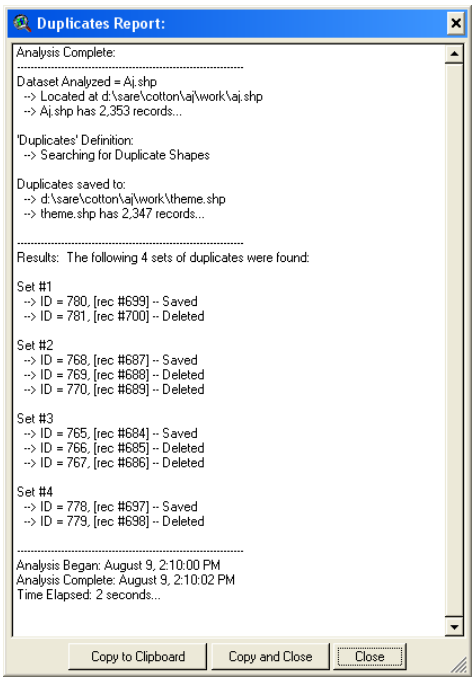


Figure 13. Screen capture of report on duplicates in ArcView GIS 3.3.

Chapter 3 Data Management in ESRI ArcMap 9.X

"Knowing where things are, and why, is essential to rational decision making"
~ Jack Dangermond, Environmental Systems Research Institute (ESRI)

Chapter 3 deals with managing the yield monitor and other site-specific data with geographical information systems (GIS) software. This chapter assumes the reader has access to and is a user of ESRI ArcGIS. Other professional GIS software including previous versions of ESRI software such as ArcView GIS 3.3 is capable of performing the same or similar tasks. Chapter 2 describes these procedures for ArcView 3.3. Some farm-level mapping software have incorporated enough GIS functionality to perform these tasks although are not described in this protocol.

Assimilate Data with ArcMap 9.X

The first step is to add the *.txt file to your GIS map. Once ArcMap is open, select File from the main menu and click on Add Data. Navigate to the directory where the *.txt file is saved. The *.txt file should be listed under Layers on the left hand side of the screen. Right click on the name of the *.txt file and select "Display X Y Data. . .". Under "Specify the files for the X and Y coordinates:", select Long (or the variable name for the longitude coordinate) from the drop down list next to "X Field:" and select the latitude variable from the drop down list next to the "Y Field:". Click on OK. Now that the *.txt file is loaded into the GIS, make sure it appears correctly on the screen and in the expected location with expected yield variation patterns similar to the variation in the final Yield Editor map window (**Figure 4**). Depending upon which column variables you selected in Yield Editor to export, your dataset will have differing pieces of data. At the very least, you will have X and Y coordinates and the yield. The *.txt file should be converted to the Shapefile format. Converting layers to Shapefiles can be performed by a variety of procedures, but we have opted to right click on the name of the displayed layer file, click on "Data" and then click on "Export Data" then navigate to where the new Shapefile is to be stored on the hard drive and give the Shapefile a new name. Click OK. Other data to be used including treatments, covariates, dummy variables and topographical information should be added to this Shapefile within the GIS.

Adding Disparate Spatial Data Layers with Spatial Joins with ArcMap 9.X

Once the yield data is in the Shapefile format and has been adjusted for spatial location and erroneously measured observations have been deleted, information from the original yield data file such as elevation can be added to the new yield data Shapefile. A spatial join is conducted to append the pertinent information from the original yield data Shapefile to the new yield Shapefile. The original yield data Shapefile was the data exported from the farm level mapping software package. The column fields that may be important to include in the final dataset may include information from the original yield data file or other site-specific data including elevation, treatment information, and covariates such as electrical conductivity.

Aggregating dense data to the least dense data using ArcMap 9.X

Rarely ever do the differing data layers share the same spatial resolution or density, so some sort of aggregation of the data is necessary. We have chosen the following process to minimize the interference of the statistical reliability. Yield data is typically the most dense, followed by soils such as electrical conductivity or other scouting information. Soil sampling for chemical analysis tends to be the most sparsely collected data assuming each location is analyzed separately, such that it may be too sparse to be included in the data. When several soil cores are taken from a grid and aggregated together for a composite sample, then the soil test information may be useful in the analysis of yield monitor data although this also assumes that the grid was small enough resolution. It has been our practice to keep the data in the format that it was original measured with the least dense dataset as the basis for the remaining data layers. We caution the analyst not to conduct spatial interpolation via kriging or other geostatistical methods to remedy the dilemma of spatially disparate spatial data layers. If spatial interpolation is used to convert the data points into a smooth surface, a systematic error is introduced into the data, causing a problem in deriving inference (Anselin, 2001), therefore we have avoided spatial interpolation when possible especially for soil characteristics measured at relatively sparse densities. Common spatial interpolation methods are not limited to: kriging, spline, inverse distance, and minimum curvature.

There are a number of ways to assimilate relatively denser data with relatively less dense data, i.e. yield data with soil sample data. Some sort of spatial polygon structure can be assigned to the dataset with each sparse soil data point being attributed to a single grid unit. Our preferred method is to create a polygon such as a circle with given radius with the less dense point as the center. This process can be accomplished by using the "Buffer Features (Retain Attributes)" function under "Vector Editing Tools" of Hawth's Analysis Tools extension for ArcMap (http://www.spatialecology.com/htools) and is explained in the following section. A specialized form of grid cells known as Thiessen polygons can be created in GIS or GeoDa (University of Illinois) (Anselin, 2003) for the same purpose. GeoDa can be downloaded from: https://www.geoda.uiuc.edu/ and Thiessen polygons created by clicking Tools:Shape:Points to Polygons. Thiessen polygons are a form of nearest neighbor interpolation created by surrounding each input point with an areal unit such that any location within that area is closer to its original point than any other point. Thiessen polygons are sometimes called or very similar to Voronoi polygons, Delaunay Triangles, and Dirichlet Regions. A regular grid can also be used, but it is difficult to spatially align irregular spaced data in a one-to-one format.

Creating buffer areal units for sparse data using ArcMap 9.X

With the data layers in the appropriate projection and distance units, go to HawthsTools then select "Vector Editing Tools" and then select "Buffer Features (Retain Attributes)". Select the layer to create the buffer from the drop down list next to "Feature layer to buffer:" and enter the distance which to create the buffer (remembering the map units). Give the new file a name under "Output shapefile:" and click OK. Once the process is complete, click OK. The buffer distance should be chosen as to 1) not overlap into areas of different treatments (or alternatively further processed to omit observations from differing treatments), 2) be large enough to have at least one yield observation if possible, and 3) be small enough to only include yield data that are comparable with other yield data in buffered zone and are representative or affected by the

sparse data. A new Shapefile layer with circular areal units around each of the sparse points is ready for the dense data points to be added. These circular area units may overlap or even include the same dense data point in two different buffer areas, but that is not of concern.

Assigning dense yield data to sparse data points using ArcMap 9.X

Once polygon areal units have been created around the least dense spatial data layer, denser data such as yield data can be assigned to the location of the areal unit around the sparse data layer. The Intersect (analysis) function in ESRI ArcMap is the first step to assign the average value from the dense data within the areal unit to the sparse data.

Make sure the data layers are projected in the proper system and the distance units are known. Double click on "Intersect (analysis)" (**Figure 14**). Under Input File, select the areal polygon created as a buffer around the sparse data points and select the point data from the dense dataset (**Figure 15**). Click OK. ArcMap may take a few moments to process the data. Open the attribute table of the resulting data file. Highlight the column of interest by left clicking on the column heading. Right click on the column heading and then click on "Summarize" (**Figure 16**). Under "Select a Field to Summarize" choose the name of the variable that came from the buffered polygon. This variable will usually have the letters "FID" with an underscore before the original file name of the buffer file and is usually at the top of the list (**Figure 17**). Then choose the variable to be summarized within the buffer polygon and select each of the sample statistics desired. We typically only use "Average"; however examining other descriptive statistics gives an indication of the appropriateness of the buffered distance for the given spatial variation (**Figure 18**). This step may take a few moments. When asked if you want to add the data to the map, click "Yes".

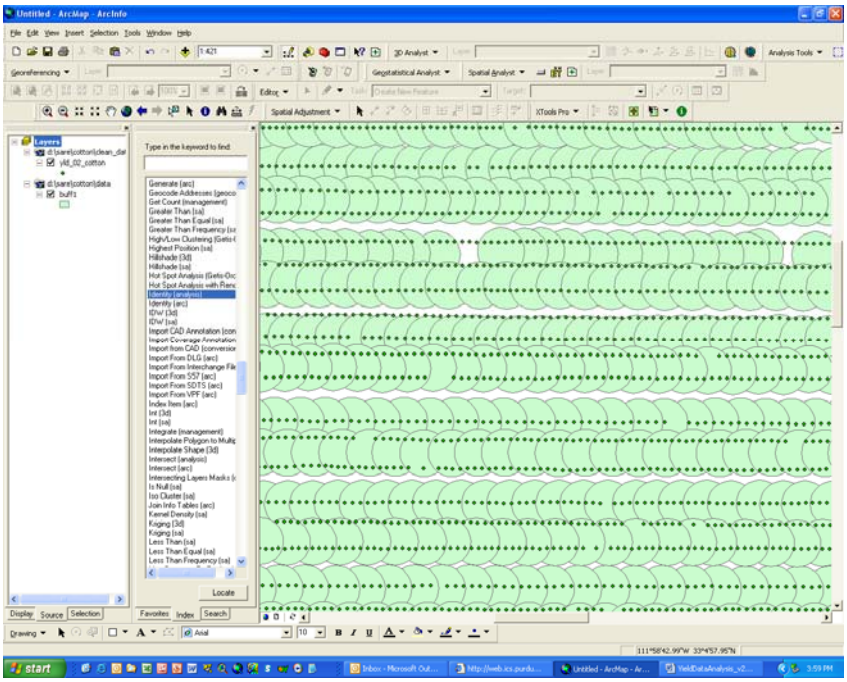


Figure 14. Screen capture of ESRI ArcMap indicating "Intersect (analysis)".

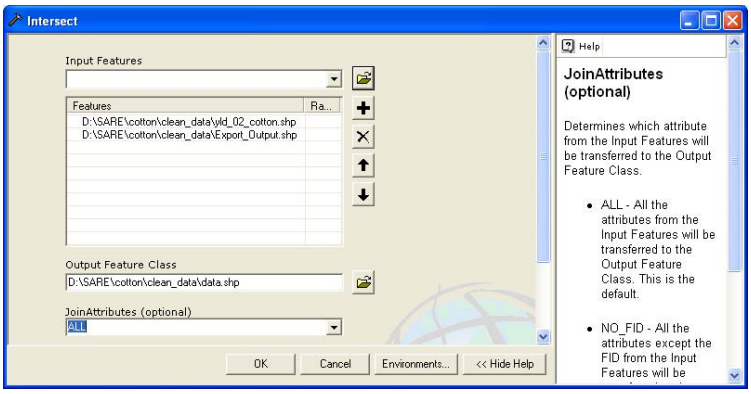


Figure 15. Screen capture of Intersect Box in ArcMap.

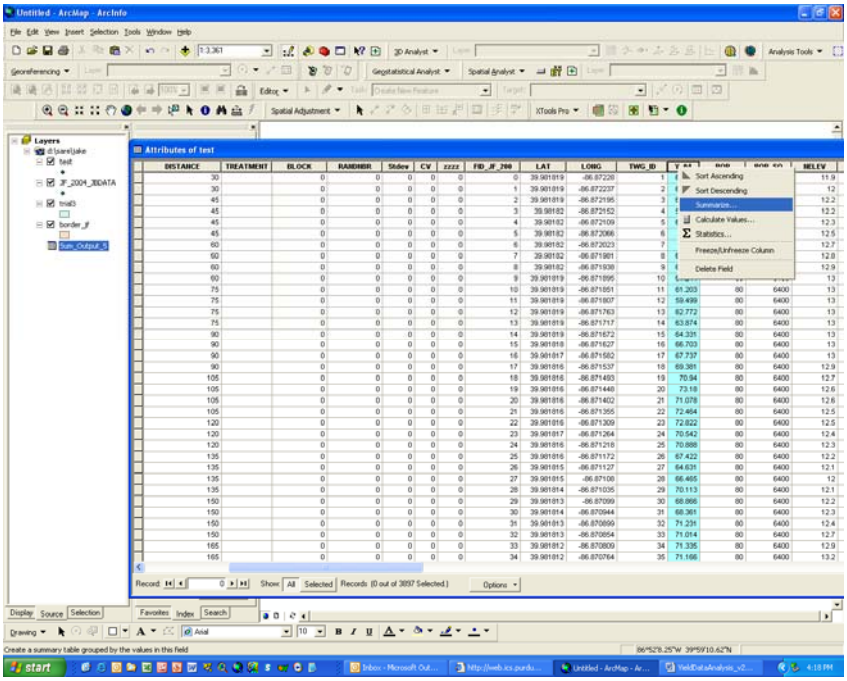


Figure 16. Screen capture of ArcMap attribute table.

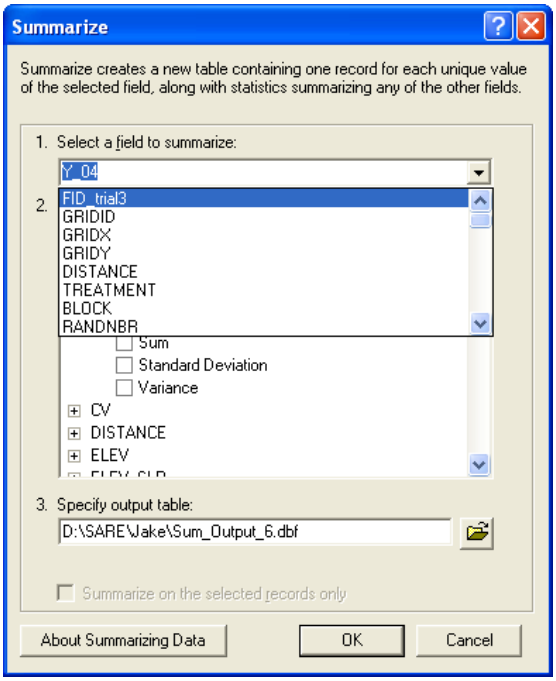


Figure 17. Screen capture of ArcMap selecting summarize variables.

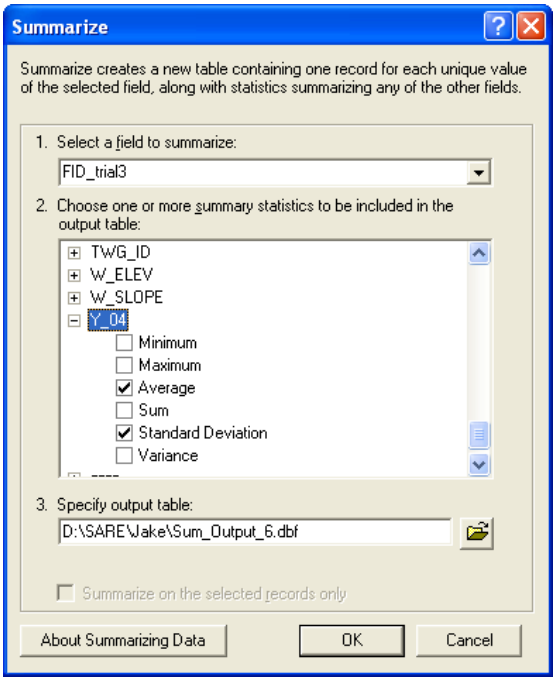


Figure 18. Screen capture of ArcMap selecting descriptive statistics.

Joining Data from Table to Point Data Shapefile using ArcMap 9.X

The summarized data can be joined to the appropriate Shapefile once the data has been summarized and added to the map. Right click on the name of the point data file that was considered the sparse data layer, then click on "Joins and Relates" and then click on "Join"

(Figure 19). Under "1. Choose the field in this layer that the join will be based on:" select the unique identifier used as the ID number when creating the buffered polygon. Under "3. Choose the filed in the table to base the join on:" select the unique identifier given to the data table based upon the buffer polygon. Click on OK. Check the attribute table to confirm that the data from the summarized data table were appended to the Shapefile. Now that the data from the dense and sparse data layers are in a single point data layer with the same resolution as the original sparse dataset, the data are useful for inferential statistical analysis

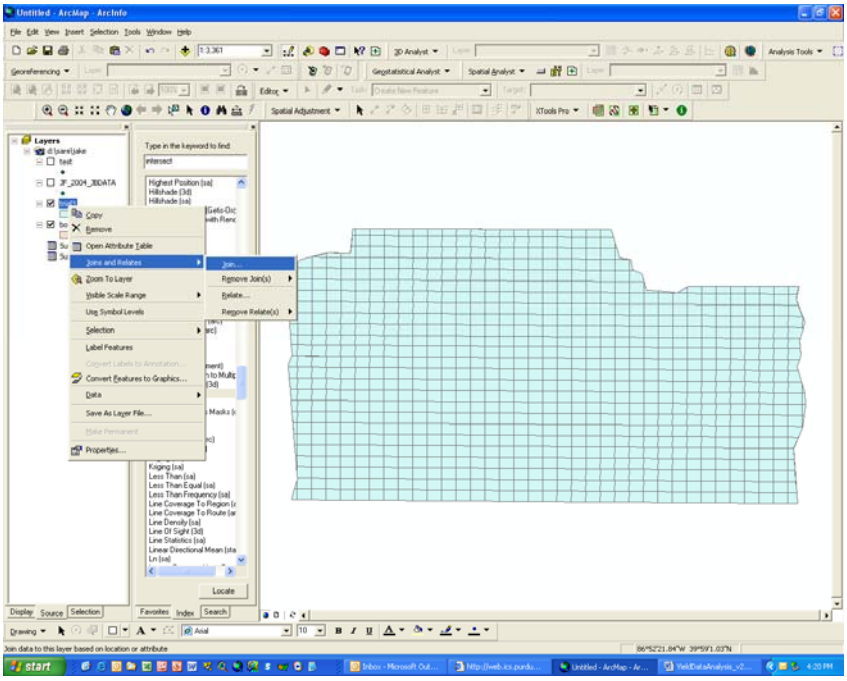


Figure 19. Screen capture joining data table to Shapefile in ArcMap.

Digression on the Modifiable Areal Unit Problem

The procedure discussed in the previous section leads to a digression on the Modifiable Areal Unit Problem (MAUP) which is common across many disciplines (Gotway and Young, XXXX). The size of the buffer around the sparse data point must be chosen via a conscious decision and not arbitrarily. With the given mean and standard deviation descriptive statistics, a coefficient of variation (CV) can easily be calculated by dividing standard deviation by the mean. The CV, or even the standard deviation, can be used to graphically represent the variation on a map. This visual representation can be useful to decide if the variance is stable over space and time, assuming there are multiple years of data. The CV is useful when comparing yields from different crops across years in the same field.

Appending treatment information to the dataset using ArcMap 9.X

Treatment information may need to be added to the data file. If this information is not already present in the dataset, it can be added in a number of ways. For instance, if the treatments occur

in blocks like tillage treatments or split field trials, polygons can be created and merged together to form the treatment polygon map. From this polygon, a specific treatment can be selected. Once the treatment is selected from the treatment polygon map, a Select by Location from Selection on the main menu can be done on the yield data points with respect to the selected portion of the treatment polygon map. Now that the yield data points associated with the treatment are selected, a dummy variable can be added using the Field Calculation. Right click the name of the data layer that has selected features and select "Open Attribute Table". Click on Options at the bottom of the table and click "Add Field". Give the new column a name and leave the type as "Short Integer". Right click the column heading created in the previous step and click "Field Calculator…". In the dialogue box, put a "1" and click OK. Confirm that the table includes "1" under the new variable for the selected features only and a "0" otherwise. These same steps can be done to add a dummy for soil series, other regions such as old feedlots, pastures, homesteads, and two existing fields were joined to be one large field. A dummy variable should be added for each categorical treatment, soil zone, and every measurable discrete factor to be included in the statistical model.

Adding the distance from a given attribute using ArcMap 9.X

In some cases, a distance variable may be useful to help describe variability from isotropic or anisotropic effects. In cases of furrow irrigation where plants near the water canal will surely get more water, albeit colder water, than plants at the other end of the row, differing yield responses are expected. Distances are also useful in modeling the isotropic effect of flood irrigated rice production where plants near the water source tend to have lower yields due to the colder temperature of the ground water near the well. The distance to the given attribute can be added to the dataset in a number of ways. One method involves using the Spatial Analyst Extension in ArcMap. From the Index, choose Euclidean Distance (sa). From the Euclidean Distance box, choose the data layer that the distance is to be measured from the drop down list under "Input raster o feature source data" and then click OK. This may take a few moments. From the index, double click Contour (sa). Under "Input raster" select the layer created in the previous step and assign a value for the "Contour interval" and then click OK. The value of the contour lines (i.e. distance from the chosen data points) can be joined to the main shapefile by the standard table joining techniques in ArcMap. Right click on the data layer that the distance values are to be appended, click Joins and Relates and then click Join. From the drop down list under "What do you want to join to this layer?" choose "Join data from another layer based on spatial location". Under "1. Choose the layer to join to this layer, or load spatial data from disk:" choose the contours created in the previous step. Under "2. You are joining: Lines to Points" click on the radio button next to "Each point will be given all the attribute of the line that is closest to it, and a distance field showing how close that line is (in the units of the target layer)." Click OK. The new data layer will have the distance data as a variable.

Elevation, slope, aspect and associated problems using ArcMap 9.X

Due to introduction of variability problems associated with geostatistical techniques (Isaaks and Srivastava, 1989) and imperfect information on proper parameters to assign to interpolation methods, spatial interpolation methods such as inverse distance weighting, kriging, and co-kriging have been avoided. However, if slope or aspect variables are desired, the elevation data must be interpolated into a surface. In addition, the elevation data must be collected at a resolution sufficient to describe the topography and with adequate accuracy. Tractors equipped

with RTK automated guidance typically provide sufficient data during plating or other field operations. Coast Guard and WAAS DGPS do not always provide the needed accuracy. Additional data points and resolution are not substitutes for accuracy. An alternative to including slope derived from interpolated elevation surface is to use relative elevation as described in Lowenberg-DeBoer et al. (2006).

From this elevation surface a slope, aspect, or other topographic surface can be calculated. The slope surface can be converted into a contour line vector with base of zero and interval of 0.25 percent. The yield data can be appended with a value for slope by choosing the closest slope contour line by using similar techniques as described in the section on *Adding the distance from a given attribute.* The danger in spatial interpolation of a surface is the introduction of variability or in other words introducing a random variable which causes problems with statistical inference (Anselin, 2001).

Chapter 4 Use of Spreadsheets

"There is no reason anyone would want a computer in their home."
~ Ken Olson, president of Digital Equipment Corp., 1977.

Once the dataset has all the necessary GIS work, a spreadsheet such as MS Excel is useful for calculating additional variables. These variables may include interaction terms, dummy variables of differing coding, squaring continuous explanatory variables, and a unique identifier field if one has not already been created. The unique ID field is required by GeoDa and many GIS functions. We typically add a column and name it with our initials, an underscore, and "ID" so "TWG_ID" may be used by Griffin. Some analysts use "POLYID" by convention. Then a sequential set of numbers are added to uniquely identify each row of data or record.

For the purposes of regression analysis, some variables must be squared, cubed, square root, natural log, or other transformation. For most studies, the original variable plus a squared term is sufficient. If the variable is a continuous experimental treatment such as a rate trial, squared, cubed and other higher order transformations may be needed depending upon the model.

Once all the main variables are created and exist in the spreadsheet, interaction terms of all the explanatory variables should be created that are intended to be used in the full model. The most important interaction terms are the linear factors with each other if there is more than one factor. Interaction terms of the factor with other variables such as elevation, soil zone, dummy variable, or other covariates are also useful.

For categorical treatments and supporting variables such as soil zones, hybrids, or other discrete choices, a binary or dummy variable should be created. For instance, any observation that is present in soil A has a "1" with other observations having a "0" as outlined in a previous section on *Assigning dense yield data to sparse data points* in either Chapter 2 or Chapter 3. To make the regression comparable to ANOVA and to have the coefficients presented as the difference from average conditions, a restriction on the dummy variables that they sum to zero can be imposed $\left(\sum d_{ij} = 0\right)$. This can be done when there are two or more categories. When there are three or more dummy variables, the convention is to select one treatment to be the reference and not include the reference in the full regression model. The process for assigning dummy variables is to subtract the value of the reference from the remaining categories. This method generates a "-1" if the observation is of the reference category, a "1" for an observation from the category in question and a "0" otherwise. When the regression is run, the reference category is omitted from the analysis and is captured in the intercept. When the dummy variables are coded this way, the coefficients are evaluated as differences from the mean condition.

Continuous covariates may be manipulated to allow estimated coefficients and subsequent economic analysis to be evaluated at meaningful levels. For instance, if absolute elevation is included in the model, the coefficients are estimated where elevation equals zero, or at sea level, rather than at the elevation the data were collected; however more meaningful coefficients can be estimated if a simple transformation is applied to the continuous elevation variable. To evaluated the coefficients at mean elevation, a new variable must be crated by subtracting the mean elevation from the elevation value for each observation. Similar transformations can be

made by subtracting the minimum, maximum, first quartile or similar values in the same manner. Although yield responses can be calculated after the fact, some analysts prefer to be able to evaluated estimated coefficients at the mean value from the study area. In addition, the standard errors will differ at different values of the continuous variable. Similar transformation can be made for other continuous variables.

Spreadsheet tips and tricks

When working with large spreadsheets having thousands of rows of data, using shortcut methods can save a lot of time. For instance, if the user wants to select a group of cells from an initial cell to the last row of data in the spreadsheet, select the initial cell then press and hold Control and Shift and then press the down arrow. Remember when using formulas to fill in data, that the formulas need to be saved as values so the resulting *.dbf or *.txt files operate properly. When working with a *.dbf and the user wants to create new columns, it is easiest to insert a new column in the middle of existing data columns so that there are data columns to the right of the new column. Otherwise, the file may not save the new columns if they are to the right of the existing data. In addition, using a *.dbf may not save the number of decimal places and revert to an integer, causing difficulties when dealing with many types of data or even coordinate systems. These data columns can be adjusted by selecting the data in the column, right clicking, click Format Cells…, select Number tab, select Number under Category and enter “6” next to Decimal places:. For these reasons, it is a good practice to first save the spreadsheet as the native *.xls file and then perform a “save as” to the *.dbf or *.txt so that a clean backup with formulas is available. The analyst should avoid sorting data within the spreadsheet software unless care is taken to sort the data in a specific manner to be able to resort the data to the original sequence of data rows. The best way to sort the data is to have a unique identifier column that has a sequential order. The whole dataset except for column headings must be sorted all at once. Before saving the dataset file, the whole dataset must be sorted back to the original sequence by using the unique identifier column. If the rows of data get arranged in an inappropriate manner, the GIS software still operates properly however the data does not match the appropriate shape, i.e. location. In other words, all the data is present, but is associated with the wrong location. Likewise, the analyst must not delete rows of data in the spreadsheet because the GIS software will not accept the Shapefile.

Chapter 5 Exploratory Spatial Data Analysis

> "If you put tomfoolery into a computer, nothing comes out of it but tomfoolery. But this tomfoolery, having passed through a very expensive machine, is somehow ennobled and no-one dares criticize it."
>
> ~ Pierre Gallois

"In exploratory spatial data analysis, one should not rigidly follow a prescribed sequence of steps but should, instead, follow one's instinct for explaining anomalies" (Isaaks and Srivastava, page 525). This leads to an underlying assumption in spatial analysis, that the analyst either has intimate knowledge of the field or is in close contact with a collaborator who does, i.e. the farmer. The results of exploratory spatial data analysis (ESDA), and steps the analyst takes to arrive at these results, are intended to give the analyst a better understanding of the spatial variation of the data.

Now that the entire dataset is in a single Shapefile, ESDA can be performed using GeoDa. Open a file using the standard icons and navigate to the folder where the Shapefile was saved. GeoDa asks that a unique identifier be assigned and is referred to as a key variable (**Figure 20**). The key variable is an unique identifier, typically a series of unique numbers; we typically add a column and with a sequential list of numbers starting with 1 that allows each observation to be identified such as the TWG_ID variable described in Chapter 4. To perform any ESDA, a weights matrix must be specified. This can be done by clicking Tools:Weights:Create. The resulting box (**Figure 21**) asks for an input file (which will probably be the same Shapefile), a name for the weights matrix (in this case W_min) and the key variable again. In this example we chose to have an Euclidean distance with a cutoff of 7.169765 meters, the minimum distance such that each observation has at least one neighbor which can be determined when the sliding bar is all the way to the left. If the data are in areal units or polygons rather than points, contiguity matrices using criteria such as queen is common.

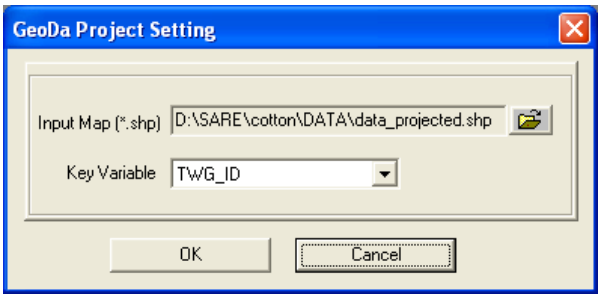


Figure 20. Screen capture of selecting a GeoDa project and assigning the key variable.

In order for GeoDa to display the distance in meters or any other specified unit, the Shapefile should be exported in some projection other than decimal degrees. This can be done in the GIS software. Whatever map units that the map is projected will be the units GeoDa displays. Otherwise if the Shapefile is exported without the projection, the units will be in decimal degrees and difficult to interpret.

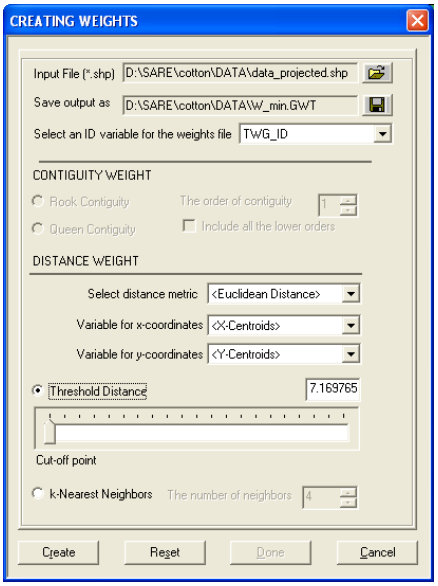


Figure 21. Screen capture of creating a spatial weights matrix in GeoDa.

One statistical measure of spatial variability is Moran's I (Anselin, 1988; Cliff and Ord, 1981). Moran's I is a global indicator of spatial autocorrelation. To calculate and plot the data for Moran's I, go to Space:Univariate Moran and select the variable you wish to explore (Figure 22). You will be asked to provide a weights matrix to use which was just created (Figure 23). The resulting Moran's I scatter plot and value (**Figure 24**) gives indication to the amount of spatial autocorrelation. In most site-specific data that we have used, we typically expect to have positive values and not negative or zero values for variables at field-scales.

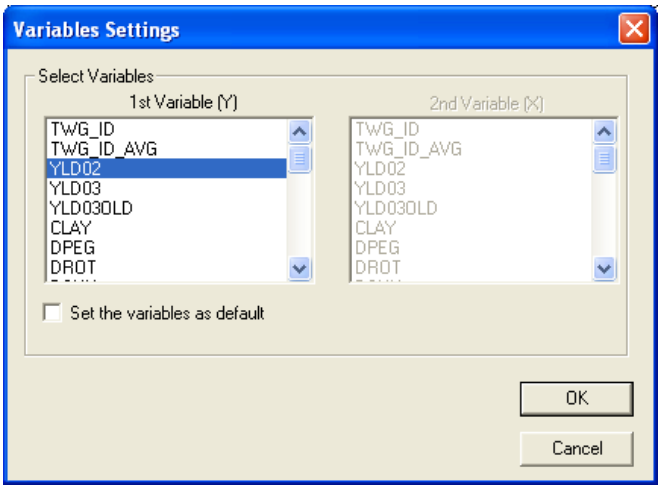


Figure 22. Screen capture of selecting the YLD02 variable to calculate a Moran's I.

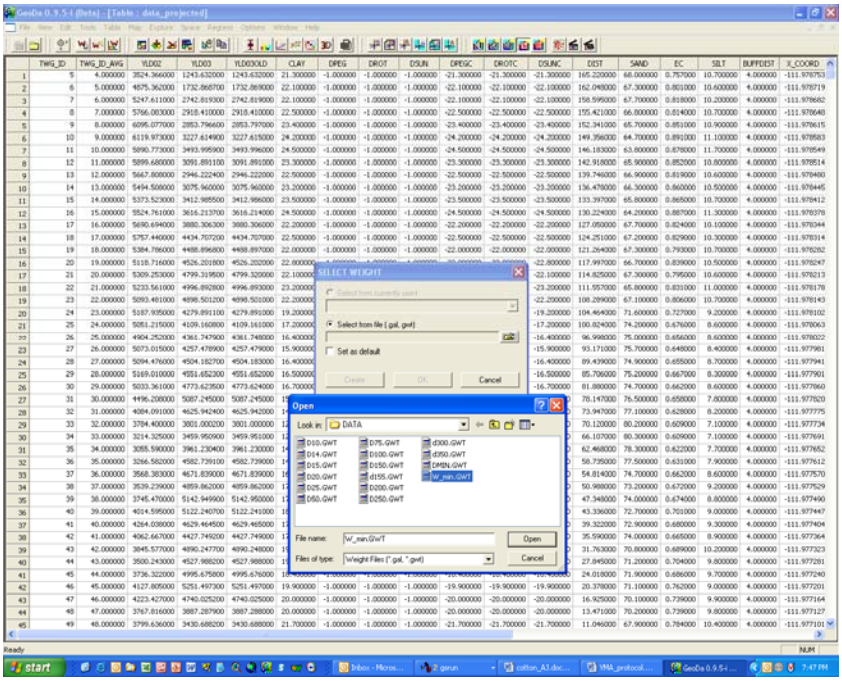


Figure 23. Screen capture assigning a spatial weights matrix in GeoDa.

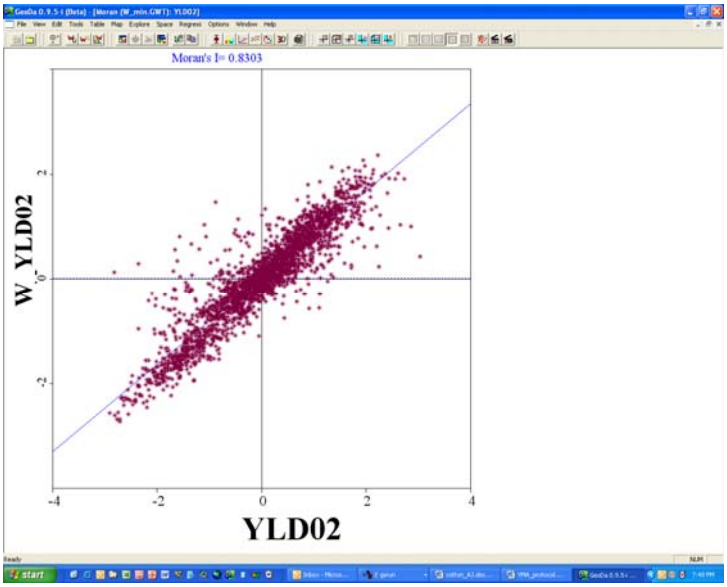


Figure 24. Screen capture of a univariate Moran's I scattergram for the YLD02 variable.

Spatial Correlogram

A spatial correlogram can be analogous to the semivariogram used in geostatistical analysis. The spatial correlogram plots the Moran's I value for each distance for which it is measured. The distance at which observations are no longer spatially autocorrelated is termed the spatial range and can be determined by the spatial correlogram or semivariogram. At this or greater distances between observations, the data can be analyzed as an non-spatial model. A spatial correlogram

can be constructed by running a series of Moran's I tests with GeoDa or by the sp.corrrelogram function in the spdep (Bivand, 2007) contributed package in R.

Other forms of ESDA can be conducted. Local indicators of spatial association (LISA) are useful in determining which geographic areas are spatially correlated either with its self or another variable. See Anselin (2003) for additional ideas and details on using GeoDa for ESDA.

Now that the analyst has a firm understanding of the spatial variation of the dataset, the analyst is ready to conduct statistical analyses.

Chapter 6 Spatial Statistical Analysis

"Torture numbers, and they'll confess to anything."
~Gregg Easterbrook

Traditional non-spatial analyses such as ANOVA and linear regression are unreliable when spatial effects such as spatial autocorrelation and spatial heteroskedasticity are present in the data. The classical assumptions of independent observations, normality, and identically and independently distributed (iid) errors are often violated. Spatial regression analysis is one methodology that overcomes these limitations of traditional analyses (see Anselin {1988} or Cressie {1993} for a thorough treatment of spatial statistical methodologies).

Definition of regression analysis
Regression analysis defined in the traditional sense can be thought of as a model-driven functional relationship between correlated variables that can be estimated from a given dataset. Regression can be used to predict values of one variable when given values of the others. Spatial statistics expands upon traditional regression to address the problems of spatial dependence, specifically in the form of spatial autocorrelation, and spatial heterogeneity (Anselin, 1988). Any appropriate statistical analysis of a spatial dataset can be thought of as spatial statistics.

GeoDa (Anselin, 2003) provides a non-spatial model estimated as ordinary least squares (OLS) and two spatial regression models both estimated with maximum likelihood (ML). Non-spatial regression is necessary for the purpose of conducting spatial diagnostics on the residuals to determine whether a spatial regression method is justified and which of the two methods available in GeoDa is the most appropriate. If the diagnostics of the residuals suggests a spatial method is appropriate, either a spatial error or a spatial lag model will be indicated that best describes the data. From our experience with field-scale on-farm data, the diagnostics indicate a spatial error model most of the time which is also expected from theory. GeoDa presents spatial diagnostics including Lagrange Multiplier (LM) values and Robust LM values for both spatial error and spatial lag. The diagnostic values with the largest LM and Robust LM values, or smallest probability levels, is the most appropriate to use (Anselin, 2003). In most cases, both the LM and Robust LM diagnostics indicate the same model. In addition, there is some conceptual evidence that the spatial error model is more appropriate than the spatial lag model; however some disagreement by researchers exists as described in the digression on spatial statistical methods below.

Digression on appropriateness of spatial error and spatial lag process models
Debates over which spatial model is most appropriate for site-specific data are still on-going between practitioners and theorists. It is our position that the spatial error model is conceptually the most appropriate for field-scale data. Conceptually, the spatial error model tends to be the most appropriate model when the spatial structure is explained in the residuals of the regression, or in other words due to omitting important spatially variable factors that explain the yield variability. In practice at field scales we are often unable to measure all the factors influencing yield and in particular yield variability. Yield variability at field-scales occurs for several factors, and most can not be feasibly measured and therefore are not included in the statistical model. When the statistical model is run without the factors causing yield variability, the unexplained

variability inevitably winds up in the residuals making the spatial error model the most appropriate. It is doubtful that researchers and farmers will collect the exact data at the resolution needed to overcome the omitted variable problem, even with relatively dense soil data such as electrical conductivity. Conversely, the spatial lag model is conceptually the most appropriate model when the spatial variability occurs in the predicted dependent variable itself, and in our case crop yield. In situations where the dependent variables affect each other directly instead of being affected by an underlying mechanism, the spatial lag model is appropriate. These situations may include any contagion such as disease spread and insect infestation. These factors affect and are affected by one another. It is counterintuitive to suggest that high crop yields in one location cause crop yields in adjoining locations to be high and vice versa. However, from statistical theory we know that the spatial lag model accounts for spatial autocorrelation in both the dependent variable and error terms. This has caused some theorists to suggest that the spatial lag model is most appropriate. This is an open debate and we welcome the thoughts and experiences of the reader on this topic.

Chapter 7 Interpretation of Statistical Results

Spatial regression techniques may someday become commonplace to the farmer or farm consultant, but currently university researchers are still developing the procedures and adapting the methodology. For the time being, spatial analysts who invest a portion of their time to teach the ultimate end user of this technology, the farm manager, to interpret analysis results rather than conduct the intricate details may have made considerable contributions to spatial analysis (Griffin and Lambert, 2005).

Goodness-of-fit measurements useful with spatial data

In traditional analyses, the R-squared statistic is a common measure of 'goodness-of-fit' or the adequacy of the model. The R-squared statistic ranges from zero, meaning it explains none of the variability in the data, to one, meaning the model explains 100% of the data. R-squared values somewhere between zero and one are expected. Although non-spatial models estimated as OLS report R-squared even with spatial data, the R-squared value are meaningless with spatial data. For instance, Griffin et al. (2004) and Griffin (2006) showed that non-spatial models were unable to adequately explain spatial datasets under simulation; however the R-squared values and F-statistics were very high. If spatial diagnostics on the OLS residuals indicated the presence of spatial autocorrelation, then the non-spatial model coefficients, standard errors, and goodness-of-fit statistics for non-spatial models should be ignored. In addition, R-squared values do not have the same interpretation with a spatial model as the non-spatial model and are normally assumed to be invalid (Anselin, 1988).

A better goodness-of-fit measurement is the maximized log-likelihood which can be used to calculate the information criterion. The use of traditional measures such as chi-squared and mean squared error provides misleading results with spatial models (Anselin, 1988). The Akaike Information Criterion (AIC) estimates the expected value of the Kullback-Leibler information criterion (KLIC) which has an unknown distribution (Anselin, 1988). The ranking of models by AIC is useful although the specific value has little meaning. The analyst should examine several goodness-of-fit measurements and not make judgments based on a single measure.

With spatial error models, the coefficients, standard errors, z-value, and probability has similar interpretation as non-spatial models, with the z-value corresponding to the t-value. Asymptotically, the absolute value of the z-value will be greater than or equal to 1.96 to be significant at the 5% confidence level. The probability level will be 0.05 for the 5% level meaning that we expect to be wrong 5% of the time. Although confidence levels such as 1%, 5%, and 10% are chosen by convention, the analyst is able to set their own requirements for confidence. The analyst should be cautioned that while the regression results from spatial error models can be directly compared to least squares and ANOVA, spatial lag model regression coefficients must be adjusted using an infinite series expansion adjustment.

A regression model can possess independent variables that are solely binary dummy variables. These models are commonly referred to as analysis of variance (ANOVA) models. If the ANOVA coding is used as described in Chapter 4 where the restriction that dummy variables sum to zero $\left(\sum d_{ij} = 0\right)$ is imposed, the analyst should be aware that the reported p-values represent the model at the average conditions, and not at the intercept. In the absence of other

continuous covariates, this is mathematically identical to ANOVA; however field-scale research typically has a wide range of soils, topography, and other yield influence factors. When ANOVA is used with small-plot experiments, the average condition of the plots is very similar to any given plot. At field-scales, the average condition probably does not closely describe the majority of locations in the field and the analyst must understand that the p-values may differ at differing locations in the field (i.e. soil clay content, organic matter level, elevation, etc.).

Chapter 8 Economic Analysis and Decision Making

"To err is human, but to really foul things up requires a computer."
~*Farmer's Almanac*, 1978

Many farmers and field researchers suggest that economic analysis is missing from their studies. In a large proportion of on-farm or field-scale trials, the economic analysis is straight forward even to non-economists although other types of trials require advanced techniques and calculus. Although the economic analysis of categorical trials may only include partial budgeting techniques, partial budgeting is also useful for economic analysis of rate trials.

Economic Analysis, Partial Budgeting and Presentation of Results

It is our practice to take the regression results and graph them so that the results can be easily communicated with decision makers. Once the regression output is available, copy and paste the output to a spreadsheet. It may be necessary to click Data: Text to Columns to nicely fit the data into the spreadsheet cells. From the coefficients, we calculate the dependent variable, typically yield, over a range of the covariates such as clay content, elevation, or other continuous variables for each treatment and/or other discrete categories such as soils. From these calculations, we create a XY(Scatterplot). These graphs are useful in discussing and interpreting the results of the planned comparison with the farm management decision-maker.

Categorical Trials and Partial Budgeting

For rudimentary economic analysis of side-by-side or categorical treatments, a partial budget is sufficent. A partial budget includes only the costs and revenues that differ between alternatives, while an enterprise budget is exhaustive. For field-scale experiments, the difference in revenue may only include the difference in revenue for each treatment, or $R = p_y y$ where R is revenue, p_y is price of crop, and y is the crop yield. The difference in costs may include the seed costs if a hybrid trial or the machinery costs if a tillage trial.

Rate Trials, Profit Maximization and Partial Budgeting

For rate trials such as nitrogen rates or seeding rates, the equation derived from the regression model is used. For instance with soybean seeding rates, the equation may be $y_s = pop + pop^2 + elev$ where y_s is soybean yield, pop and pop^2 are seeding population and population squared, and $elev$ is the elevation. Other transformations of pop are possible including pop^3 and the natural log of pop. Model specifications may be chosen *a priori* or tested. The model coefficients are used to calculate yield maximizing soybean population levels, or what is commonly known as agronomic maximum. However, yield maximized levels are not profit maximization levels unless the soybean seed is free, an unlikely situation. To calculate profit maximization levels, or economic optimal levels, the profit function must be used $\pi = R - C$ where π is profit, R is revenue and C is cost. The profit function can be expanded to

$\pi = p_y y - p_x x$ where p_x is the price of the input, x. So the equation for profit from a soybean population rate study may be $\pi_s = p_y\left(pop + pop^2 + elev\right) - p_s\left(pop\right)$ where π_s is profit from soybean and p_s is the price of soybean seed. Yield maximization and profit maximization levels can be found in the above examples by taking the first derivative and solving for the optimum level of input usage. For instance, the profit maximization level can be solved for the research factor from the above equation by $pop = \frac{p_s - p_y}{2p_y}$. It is a good practice to take the second derivative in order to assure the analyst of the shape of the curve so that the analyst does not inadvertently minimize profits or maximize costs. The above examples are only one of a large number of possibilities for models and research factors. Each planned comparison may have a completely different model, costs, and treatments and the analyst should be prepared to adjust their own protocol accordingly.

Farm Management Recommendations and Decision Making

Farm Management Recommendations and Decision Making was chosen for the title of this section because whether the analyst is the farmer or a third party, the farmer must make the farm management decision based upon evidence which may have arrived in the form of a farm management recommendation or from their own statistical and economic analysis.

We have stressed that the result of a spatial analysis is a production recommendation and not just a map. Some farmers, consultants, and researchers have concluded that precision agriculture is simply a map; potentially from services offering precision agriculture analysis but in reality only provide a map. Our conjecture is that analysis of precision agriculture data must result in a farm management recommendation that the farm manager can feasibly implement in a timely manner. We sometimes use maps for communication and validation purposes, but never as the ultimate end product of a spatial analysis.

Although appropriate spatial analysis is sometimes difficult and time consuming, it is imperative to provide the farm manager with a production recommendation in a timely manner. A “timely manner” may be defined in a variety of ways based upon the season and the input tested. For instance, corn hybrid seed are typically ordered near the end of harvest for the following year to secure early order discounts. If a production recommendation based upon spatial analysis of a corn hybrid trial was provided to the farm manger sometime in early spring or even late winter, the value of the recommendation has diminished.

References

Anselin, L. 1988. *Spatial Econometrics: Methods and Models,* Kluwer Academic Publishers, Drodrecht, Netherlands.

Anselin, L. 1992. SpaceStat Tutorial. University of Illinois, Urbana-Champaign, Urbana, IL 61801. http://www.terraseer.com

Anselin, L. 1999. Spatial Data Analysis with SpaceStat and ArcView Workbook 3rd Ed. Available on-line at: http://www.terraseer.com/products/spacestat/docs/workbook.pdf

Anselin, L. 2001. Spatial Effects in Econometric Practice in Environmental and Resource Economics, American Journal of Agricultural Economics 83, 705-710.

Anselin, L. 2003. GeoDa 0.9 User's Guide. Spatial Analysis Laboratory, University of Illinois, Urbana-Champaign, IL. http://sal.agecon.uiuc.edu/geoda_main.php

Anselin, Luc (1998). Interactive techniques and exploratory spatial data analysis, in P. Longley, M. Goodchild, D. Maguire and D. Rhind (eds.), Geographical Information Systems: Principles, Techniques, Management and Applications, pp. 251–264. New York, Wiley.

Anselin, Luc and Shuming Bao (1997). Exploratory spatial data analysis linking SpaceStat and ArcView, in M. Fischer and A. Getis (eds.), Recent Developments in Spatial Analysis, pp. 35–59. Berlin, Springer-Verlag.

Bivand, Roger. 2007. spdep: Spatial dependence: weighting schemes, statistics and models. R package version 0.4-2.

Cliff, A.D. and Ord, J.K. 1981. Spatial Processes, Models and Appplications. London: Pion.

Cressie, N. A.C. 1993. Statistics for Spatial Data. John Wiley & Sons: New York.

DeLaune, Mike. Guide To XTools Extension September 2003. Available on-line at: http://www.odf.state.or.us/divisions/management/state_forests/XTools.asp

Dombroski, Mathew. ESRI ArcView Extension: Point Stat Calc. Available on-line at: http://pubs.usgs.gov/of/of00-302/

Drummond, Scott. 2006. Yield Editor 1.02 Beta Version User's Manual. http://www.fse.missouri.edu/ars/ye/yield_editor_manual.pdf

Erickson, B. 2005. Workshop Helps Farmers Utilize One Of Their Key Resources: Information. November 2005 SSMC newsletter, Available on-line at: http://www.purdue.edu/ssmc

Gotway, C.A. and Young, L.J. 2002. Combining Incompatible Spatial Data. Journal of the American Statistical Association. June 2002, Vol. 97, No. 458.

Griffin, T.W. 2006. Decision-Making from On-Farm Experiments: Spatial Analysis of Precision Agriculture Data. Ph.D. Dissertation, Purdue University, West Lafayette, IN, USA.

Griffin, T.W., Florax, R.J.G.M., and Lowenberg-DeBoer, J. 2005. Yield Monitors and Remote Sensing Data: Sample Statistics or Population? Site-Specific Management Center December 2005 Newsletter. Available on-line at: www.purdue.edu/ssmc

Griffin, T.W., D.M. Lambert and J. Lowenberg-DeBoer, 2004. "Testing for Appropriate On-Farm Trial Designs and Statistical Methods for Precision Farming: A Simulation Approach." Forthcoming in 2005 Proceedings of the 7th International Conference on Precision Agriculture and Other Precision Resources Management, ASA/SSSA/CSSA, Madison, Wisconsin.

Griffin, Terry and Dayton Lambert. 2005. Teaching Interpretation of Yield Monitor Data Analysis: Lessons Learned from Purdue's 37th Top Farmer Crop Workshop. Journal of Extension 23(3).

Griffin, T.W., Fitzgerald, G, Lambert, D.M, Lowenberg-DeBoer, J., Barnes, E.M., and Roth, R. 2005a. Testing Appropriate On-Farm Trial Designs and Statistical Methods for Cotton Precision Farming, *Proceedings of the Beltwide Cotton Conference*, January 4 – 7, 2005, New Orleans, LA. Available at: http://www.cotton.org/beltwide

Isaaks, E.H. and Srivastava, R.M. 1989. An Introduction to Applied Geostatistics. Oxford University Press, Inc. New York, NY.

Jenness, J. 2005a. Distance Matrix (dist_mat_jen.avx) extension for ArcView GIS 3.3, v. 2. Jenness Enterprises. Available at: http://www.jennessent.com/arcview/dist_matrix.htm

Jenness, J. 2005b. Find Duplicate Shapes or Records (find_dupes.avx) extension for ArcView 3.x, v. 1.1. Jenness Enterprises. Available at: http://www.jennessent.com/arcview/find_dupes.htm

Littell, R.C., G.A. Milliken, W.W. Stroup, R.D. Wolfinger. 1996. SAS System for Mixed Models. The SAS Institute Inc., Cary, North Carolina.

Lowenberg-DeBoer, J., Griffin, T.W. and Florax, R.J.G.M. 2006. Local Spatial Autocorrelation in Precision Agriculture Settings Accounting for Micro-Scale Topography Differences" in Proceedings of the 8th International Conference on Precision Agriculture and Other Resource Management, Minneapolis Minnesota, July, 2006.

Nistor, A. and Florax, R.J.G.M. 2007. Farmers and Consultants Receive Training in Spatial Analysis of Yield Monitor Data. Site-Specific Management Center website April 2007 Newsletter. http://www.purdue.edu/ssmc

Appendix: Useful and Free Software and GIS Extensions

Useful (and free) software

The R Project for Statistical Computing http://www.r-project.org/

GeoDa https://www.geoda.uiuc.edu/

Useful (and free) extensions to ESRI ArcView GIS 3.3

PointStatCalc
By Matthew Dombroski
http://pubs.usgs.gov/of/2000/of00-302/

EFRA
Enhanced Farm Research Analyst
developed under direction of Dr. Don Bullock, University of Illinois

ET
Edit Tools
By Ianko Tchoukanski
http://www.ian-ko.com/

Stream Mode Digitizer
By Minnesota Department of Natural Resources
http://www.dnr.state.mn.us/mis/gis/tools/arcview/extensions.html

XTools
By Mike DeLaune
http://www.odf.state.or.us/divisions/management/state_forests/XTools.asp

Distance Matrix
By Jeff Jenness
http://www.jennessent.com/arcview/dist_matrix.htm

Find Duplicate Shapes or Records
By Jeff Jenness
http://www.jennessent.com/arcview/find_dupes.htm

Useful (and free) extensions to ESRI ArcMap 9.2

Hawth's Analysis Tools
By Hawthorne Beyer
http://www.spatialecology.com/htools

About the Authors

Terry Griffin is an Assistant Professor and Extension Economist in the Department of Agricultural Economics and Agribusiness at University of Arkansas and a former Graduate Research Assistant in the Department of Agricultural Economics at Purdue University. Griffin's Ph.D. Dissertation evaluated alternative field-scale experimental designs and inferential spatial statistical analysis methods for whole-farm decision making. This yield monitor data analysis document was developed in association with Griffin's Ph.D. dissertation research. Before pursuing the Ph.D., Griffin was a farm management specialist with University of Illinois Extension. Griffin holds M.S. and B.S. degrees in Agricultural Economics and Agronomy, respectively, from the University of Arkansas where he began research on precision agriculture.

Jason Brown is a Graduate Research Assistant in the Department of Agricultural Economics at Purdue University. Brown's M.S. thesis dealt with yield monitor data and uses GIS to accomplish analysis goals. Brown was also instrumental in converting these GIS techniques to the current methods presented in this document. Brown's M.S. thesis evaluated controlled drainage and the status quo as treatments for field-scale on-farm trials.

Jess Lowenberg-DeBoer is a Professor in the Department of Agricultural Economics and Associate Dean of International Programs in Agriculture at Purdue University. Dr. Lowenberg-DeBoer served as Griffin's major professor and Ph.D. committee chair as well as Brown's M.S. committee chair. Dr. Lowenberg-DeBoer has been conducting precision agriculture research for over a decade.

Acknowledgements

This protocol is the result of Terry Griffin's Ph.D. Dissertation research which was funded by a United States Department of Agriculture - Sustainable Agriculture Research and Education (USDA-SARE) for Graduate Student Grant Program project number GNC03-020 entitled "Development of Appropriate Participatory On-Farm Trial Designs for Sustainable Precision Agriculture Systems".

The authors wish to thank all those who made suggestions and comments. Special thanks to Zach Cain, former graduate student in the Department of Agricultural Economics at Purdue University, and Bruce Erickson, Department of Agricultural Economics at Purdue University.

Disclaimers

The purpose of this document is to provide a suggestion on using yield monitor data and spatial analysis methods in evaluation of treatments from field-scale on-farm trial experiments. The opinions and conclusions expressed here are those of the authors. Mention of specific suppliers of hardware and software in this manuscript is for informative purposes only and does not imply endorsement. This document is in a continued state of improvement; please forward any and all comments and suggestions to the authors for the next version.