ISSUES IN ANALYSIS OF SOIL-LANDSCAPE EFFECTS IN A LARGE REGIONAL YIELD MAP COLLECTION.

D. Brenton Myers*, Newell R. Kitchen, Kenneth A. Sudduth

Cropping Systems and Water Quality Research Unit
USDA-ARS
269 Ag Engineering Building, Columbia, MO.

ABSTRACT

Yield maps are commonly collected by producers and precision-agriculture service providers and are accumulating in warehouse scale data-stores. A key goal in analysis of yield maps is to understand how climate interacts with soil landscapes to cause spatial and temporal variability in grain yield. However, there are many issues that limit utilization of yield map data for this purpose including: i) yield-landscape inversion between climate years, ii) sensor system malfunction and inaccuracy, iii) poor data management practices and operator error, iv) field configuration and logistical limitations, v) spatial, temporal, and producer variability in agronomic management, and vi) incomplete target and predictor dataspace. Each of these issues requires a significant effort to understand and then address by the commercial and research precision agriculture community. A key goal of this investigation was to use a regional extent yield map data warehouse to model the effects of soil landscape properties on site specific mean yield and yield risk. Data mining technologies were used to examine relationships between yield map data and soil landscape attributes. Our initial results indicate challenges in training data mining algorithms to produce stable estimates when applied to independent testing data both within and across years. We found the above factors reduce the effectiveness of data mining approaches. To improve this situation, we propose a more stringent data cleansing procedure and a more agronomically complete yield map data model to better populate important predictive information in yield map databases.

Keywords: yield map, data mining, data warehouse, yield map errors, metadata

INTRODUCTION

Yield maps are an information-rich data source which can describe the integrated effects of climate and soil landscape properties on crop performance. They are now commonly collected by producers and precision agriculture service providers in warehouse scale data-stores and are multi-temporal due to yearly accrual. For over a decade now researchers have been searching for ways to use soil and landscape information to model variation in small collections of yield maps (Kravchenko and Bullock, 2000; Serele et al., 2000; Drummond et al., 2003; Irmak et al., 2006; Green et al., 2007; Norouzi et al., 2009). Successful and global models would be useful to quantitatively inform management decisions. However, the most common use of yield maps remains the qualitative appraisal of spatial and temporal yield variation. Warehouse scale collections of yield maps can be overwhelming for producers and agronomists to cope with or analyze. Large spatial and temporal variability can also challenge quantitative approaches to yield map use (Florin et al., 2009). Data mining algorithms offer a potential solution to this problem, but lack of suitable predictors, and errors and noise in yield map data may limit their utility.

Objectives

Our initial goal for this investigation was to use a regional yield map data warehouse to model the effects of soil landscape properties on yield and yield risk. This has led to a synthesis of present issues with the use of data mining algorithms on a large regional extent yield map data warehouse. From this we have developed an objective to propose some potential solutions for improving mining of yield map data. The specific objectives were to:

1. Model corn yield map variation with soil and landscape variables using a random forest algorithm.
2. Present a more agronomically complete yield map data model.

MATERIALS AND METHODS

Yield Map Data

Yield data were collected from producers in northeast Missouri (fig. 1) either directly or through their precision ag service providers. Yield data was received in native yield monitor file formats (e.g. *.yld, *.ilf, ,*gsy). Yield monitor files and the raw yield maps within them retained whatever settings, calibrations, and filters were applied in the field by the producer. Yield monitor files were imported into commercial yield mapping software (Ag Leader Technology, 2011) and fields/loads were processed and calibrations applied with all software filters disabled. Raw yield maps were exported in text format for cleaning. Exported yield maps were processed individually using Yield Editor 1.02 (Drummond and Sudduth, 2012). Key filters applied with this software included min-max, start and end pass delay, and thresher flow delay, all performed by the visual methods outlined in Sudduth and Drummond (2007). Obvious bad data points and transects were manually removed with the selection tools in Yield Editor. These may have included point rows with unknown width, edge rows of headlands, transects on terraces, and small isolated field areas. Yield maps were aggregated into rasters at 10 meter resolution by nearest neighbor averaging for data mining procedures. Training and testing datasets were selected from the yield maps by stratifying them into producer and year and randomly selecting 70% of these strata as training data with the remaining 30% as testing data. From within these groups we randomly selected 5% of the data points to train and test models. Table 1 lists some general information about the yield map warehouse. Figure 1 shows the spatial distribution of the yield maps within the counties of Northeast Missouri.

Table 1. General statistics describing the size of the yield map dataset used for this study.

Crop	Producers	Field Years	Acre Years
Corn	21	769	50,640
Soybean	21	1,525	63,500

Soil Landscape Data

Soil property and geomorphological variables were collected and/or calculated for use as predictors in data mining procedures. Digital terrain model (DTM) attributes were derived from the USGS National Elevation Dataset at both 10 and 30 m resolution. Extent of the DTM calculation procedures included the full area of all watershed basins intersected by the counties in Figure 1. Not all DTM attributes could be calculated at the 10 meter resolution due to processing limitations. Selected soil properties were joined from the SSURGO map unit attributes. See table 2 for a complete listing of the predictors and references.

Data Mining Analysis

For the purposes of this paper we limited our analysis to the corn yield map data from years 2002 to 2009. Random forest (Breiman, 2001) was chosen as a representative data mining procedure to demonstrate issues with modeling grain yield map data with soil landscape properties. Random forest is an ensemble learning technique whereby random selections of predictors and training observations are used to develop a population of regression trees. The predictions from the trees are combined to develop the final fitted model. We used a parallel implementation of the random forest algorithm (Liaw and Wiener, 2002) to manage the large size of the modeling task and to facilitate parameter tuning with multiple runs of the procedure. Random forest models were fitted for each year as well as across all years.

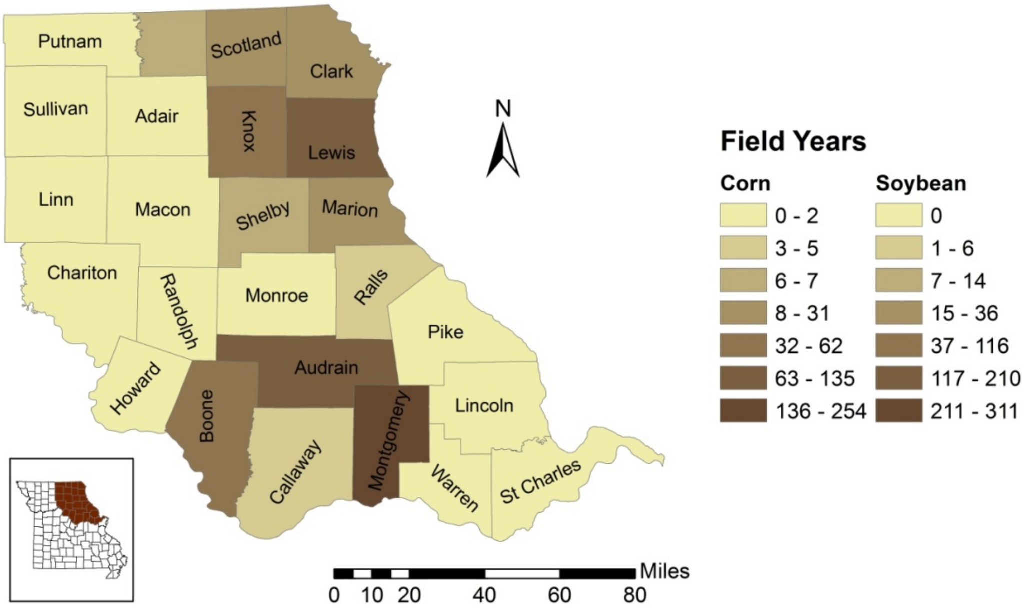

Figure 1. Location of the study area and approximate spatial distribution of yield maps aggregated by county.

Table 2. List of the predictor variables used to model soil-landscape effects on corn yield monitor data.

Type	Predictor	Detail	Procedure Reference
Digital Terrain Attributes	Elevation	10, 30 m	Gesch, 2007
	Slope	10, 30 m	Tarboton, 1997
	Avg., Planform and Profile Curvature	10, 30 m	Zevenbergen and Thorne, 1987
	Catchment Area and Height, Length of Slope, Flow Accum.	10, 30 m	Freeman, 1991
	Compound Topographic Index	10, 30 m	Boehner et al., 2002
	100 m Local Elev. and Deviation	10, 30 m	Boehner et al., 2009
	Baseline Subtracted Elev., Baseline Normalized Elev., Cell Balance	30 m	Boehner et al., 2009
	Downslope Distance Gradient	30 m	Hjert et al., 2004
	Terrain Ruggedness Index	30 m	Riley et al., 1999
	Multi-res. Valley Bottom Flatness, Multi-res. Ridge Top Flatness	30 m	Gallant and Dowling, 2003
	Diurnal Anisotropic Heating, Incoming Solar Rad., Diffuse Insolation, Direct Insolation	30 m	Boehner et al., 2009
Soil Map Unit Attributes	Slope, Flood Freq., Avail. Water (25, 50, 100, 150 cm), Drainage Class, Hydrologic Group	-	Soil Survey Staff, 2012

RESULTS AND DISCUSSION

Yearly statistics about the corn yield maps used in this study are shown in table 3. Number of yield maps varies from year to year due to increasing adoption of yield monitors throughout the study period. Several hundred thousand yield data points were collected for most years. Across all years the mean yield was 132.8 bu acre^{-1} and standard deviation of mean yield was 30.5 bu acre^{-1}. These results are typical of the Central Claypan Areas of Missouri and Illinois from where the data are collected. The soil landscape in this region has large variation in topsoil depth over an argillic horizon with 55 to 65 % clay content. The subsoil argillic horizons limit productivity due to early season wetness from a perched water table that can cause spatially variable de-nitrification, disease, and emergence problems. This condition occurred in 2002 when yields averaged 105 bu acre^{-1} for Missouri, and 121 bu acre^{-1} for the yield maps in our dataset. The high water tension imparted by the clay minerals in the subsoil also can exacerbate late season drought such as seen in 1999 when corn yields averaged 97 bu acre^{-1} for Missouri and 73 bu acre^{-1} for the yield maps.

Random Forest Data Mining

Random forest models exhibited excellent performance on the training data, but largely failed on the independent test data (table 4) indicating severe over-fitting to the training cases. This was true for models fitted within and across years (figure 2). Potential causes are pseudoreplication and spatial autocorrelation among the training observations, and an overall lack of information in the

Table 3. Yearly statistics from cleaned corn yield map data describing the mean yield, median minimum and maximum yields, and median interquartile range (IQR) in yield among the corn yield maps used for this study.

Year	Fields	n	Mean	St. Dev.	Min	Max	IQR
1996	2	35963	136.7	18.3	25.7	205.2	21.7
1997	12	153269	127.4	24.9	53.3	200.4	24.2
1998	30	283085	125.7	41.6	20.5	212.4	39.0
1999	56	412321	73.3	35.2	14.0	152.5	37.2
2000	52	329392	161.3	35.4	53.6	237.5	26.2
2001	34	313808	130.2	29.1	37.2	201.2	28.1
2002	50	378780	121.1	33.5	39.8	205.0	26.7
2003	83	719510	125.3	44.9	27.3	209.9	31.3
2004	50	392150	171.3	45.6	56.0	261.8	39.8
2005	54	400383	78.0	34.3	11.0	159.0	24.2
2006	93	756140	151.6	38.2	30.9	236.2	25.7
2007	95	603950	124.5	42.5	20.2	215.3	30.7
2008	77	601250	154.5	40.8	29.6	241.2	39.2
2009	83	429398	178.3	37.3	56.0	259.7	33.5

predictors relative to the noise in corn yield response. This suggests that the soil-landscape signal is not strong enough to drive the models, or that it is too unstable between fields and/or years. Autocorrelation is likely a significant contributor to over-fitting and a different modeling or sampling approach may be warranted. However, significant evidence exists for noise in the yield relationship to the soil-landscape signal. We plan to pursue the autocorrelation problem, but here we examine issues that may mask the soil-landscape effect on corn yield models developed from yield map collections. These issues can be considered in three general classes. First there are real landscape effects that lead to contradictory responses in different climate years. Second there are issues with yield map data that interfere with modeling. Third, yield maps generally do not carry important contextual details about the crop that could be used as predictors in an analytical model. Here we expand in some detail on the second and third points.

Table 4. Training and testing results of random forest models fitted from soil-landscape attributes.

	Train		Test	
Year	**R^2**	**RMSE**	**R^2**	**RMSE**
		bu acre^{-1}		bu acre^{-1}
2002	0.093	6.3	0.03	22.3
2003	0.96	6.2	-0.19	35.8
2004	0.97	6.1	-0.24	28.3
2005	0.94	5.9	-0.98	25.5
2006	0.95	5.9	0.39	22.8
2007	0.95	6.5	0.28	26.7
2008	0.94	6.7	0.35	26.8
2009	0.94	6.4	0.52	20.2
1998-2009	0.89	11.1	-0.02	35.9

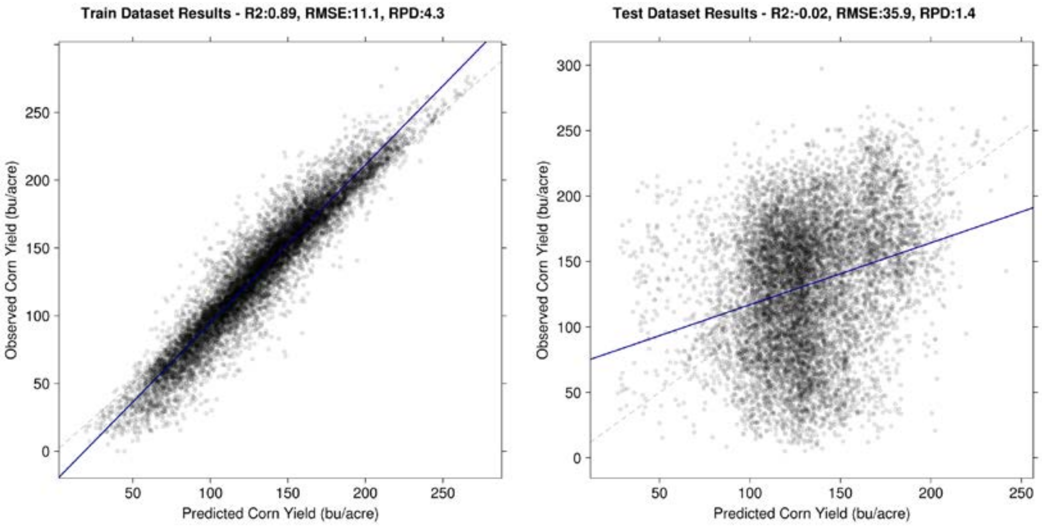

Figure 2. Training and testing results for random forest models of corn yield for years 1998 to 2009.

Yield Inversion Masks Soil Landscape Relationships

Soil landscapes in the study area have a well-documented effect on corn yield (Kitchen et al., 2003; Jung et al., 2006). Key factors are early season wetness and late season drought and how they interact with the landscape. Heat stress during the silking period can cause reduced pollination and shorter ears. Any of these conditions may or may not occur within a single growing season. Confounding these effects is the landform correlated variation in depth to claypan. The claypan is moderately deep (30 to 40 cm) on flat upland divides, more shallow on shoulders and upper backslopes (20 to 30 cm), minimal on eroded backslopes (0 to 20 cm), increases at lower backslope positions (20 to 30 cm), and is deep to very deep on footslope and depositional soils (30 to 100 cm). This variation leads directly to spatial variation in lateral water movement across the claypan as well as in available water capacity (Jiang et al., 2007).

These complicated landscapes cause a 'yield inversion' effect between climate years. For example, yield at backslope locations may be very poor in years with severe drought. Adjacent areas with deeper topsoil and greater water holding capacity and supply will produce a greater yield. However, in very wet years, yield from backslope locations may exceed yield from depositional and summit soils where de-nitrification severely limits productivity. Figure 3 demonstrates the year to year variation in corn response to general landscape positions. Soil-landscape yield inversions are evident comparing 2001 and 2009 corn yield. The relationship of corn yield to landscape position is unstable year-to-year.

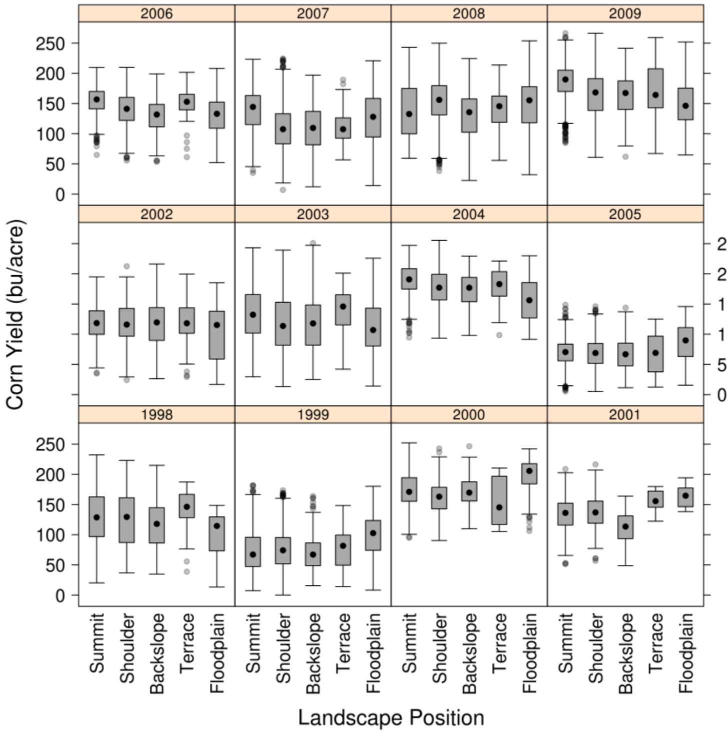

Figure 3. Yearly boxplots of corn yield for common landscape positions demonstrate the spatial variation of yield across landscape positions is not stable year-to-year.

Yield Map Issues that Interfere With Modeling

Due to our experience with collecting, cleaning, and analyzing a large collection of yield maps we have observed a wide range of issues that are not related to natural soil-landscape processes.

Yield Monitoring System Malfunction and Inaccuracy

Failures to the yield monitoring system commonly result in lost or poor quality yield data. This is demonstrated in the yield map in figure 4,a. Yield maps such as this can be a result of GPS system failure, a full storage card, an unplugged or broken cable, or in the case of this specific map, due to inadequate voltage supplied to the yield monitor as a consequence of alternator failure. Grain producers are taxed at harvest time to get their crops out in a timely fashion under optimal conditions and tend to stretch across larger acreages to benefit from economies of scale. Yield monitor failure does not prevent harvesting equipment from functioning and unless the combine itself is broken, the pressures of harvest usually compel producers to proceed without data collection. Missing fields are missing climate years in a temporal sequence of yield maps and can bias an analysis of soil-landscape effects.

Detailed management of GPS equipment, yield monitoring electronics, sensors, software, firmware, and sensor calibration is required in order to obtain accurate yield maps. Previous research has indicated problems with producer implementation of calibration procedures (Grisso et al., 2002). Points of failure are inadequate range of flow rates for calibration loads and no re-calibration as harvest conditions change. Incorrectly calibrated yield maps introduce biased yield measurements that can foil modeling of soil-landscape effects. The calibration quality of the yield maps in our collection is unknown.

Data Management Practices and Operator Error

An organized approach to data collection and management can prevent errors from getting into yield map databases and clouding analytical results. One key problem producers had was correctly inputting field names and crop types when harvesting new field areas. Figure 4,b demonstrates this problem. The two yield maps shown were harvested under the same field name and identified in the yield map as corn. However, they were planted with two different crops. The result is that a corn grain calibration was applied to the soybean field by the yield monitor before exporting. This field would become an outlier in a modeling procedure. Figure 4,c shows some additional outliers from our database of soybean fields. Data from multiple years are plotted against elevation. The data points circled in red are most likely corn fields that were harvested as soybean and processed with a corn calibration. These issues can be prevented if field names, crop types, and as-planted boundaries are pre-configured before the harvest season. If not screened, these errors can cripple the effectiveness of analytical procedures.

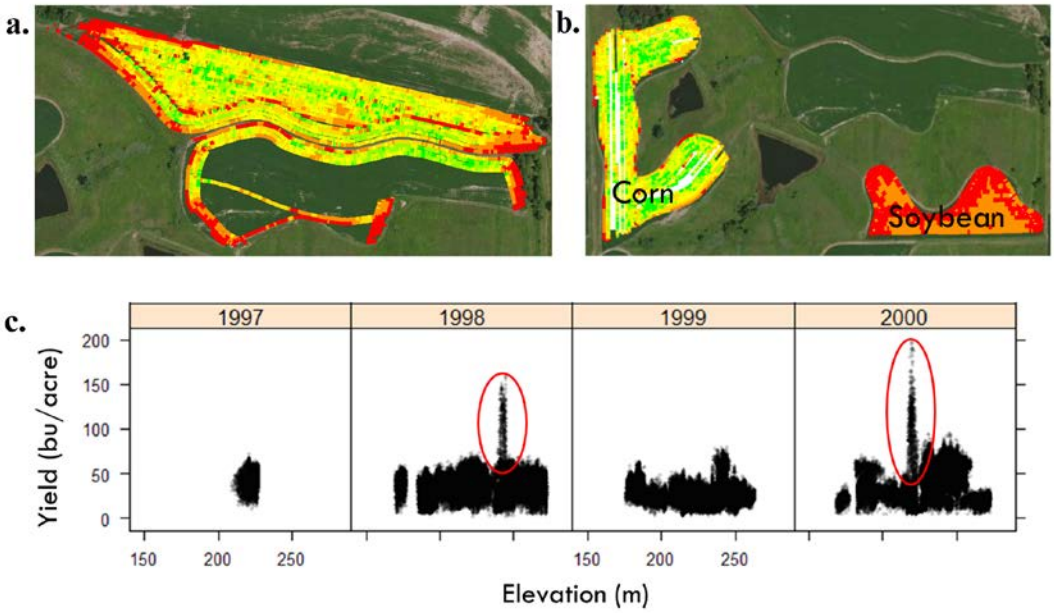

Figure 4. a) Yield monitor failure leads to missing data. b) These fields were harvested as a single corn field. The operator failed to change the crop when entering the soybean field causing the corn calibration to be applied. c) Circled data are from corn fields identified as soybean but with a corn calibration applied.

Field Configuration and Logistical Activity

Field configuration, modification, and logistical realities may be critical causes of difficulty in modeling soil landscape effects on yield. The natural productivity of a soil landscape can be masked by headlands, tree lines, and field entrances. These types of features are related to the field shape and can have markedly greater impact on landscapes where field sizes are smaller and more irregular. Irregular field shapes caused more overlapping during planting and inaccuracy in yield maps due to more frequent 'point rows'. Point rows lead to unknown swath sizes, and may cause inaccurately estimated yield because grain flow rates are at the extreme edge of the yield sensor's calibration range. Point rows are also more common in fields with extensive terracing. Terraces cause further unnatural effects on the soil-landscape interactions with yield due to physical changes in the soil and geomorphology as well as changes in water movement on the landscape. Field leveling, tiling, and ditching also confound the natural response of crops to soil variability.

Fields in our study area and in our yield data warehouse commonly demonstrate some of the issues described above. Field shapes often follow the contours of drainages, waterways, and rolling terrain. The yield map in figure 5,a depicts an extensively terraced field with yield transects having alternately very high (green) and low (red) yields. Either of these conditions may be different than the natural soil would have produced as soils are disturbed and compacted or

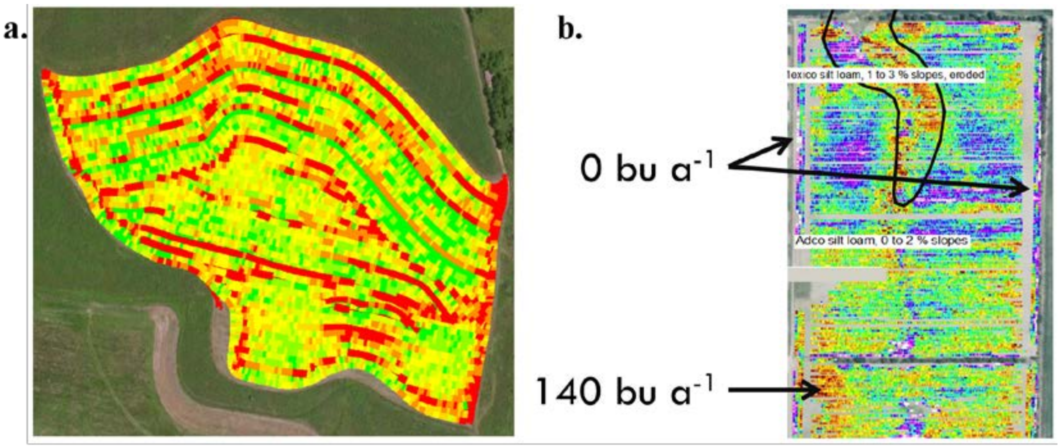
a.
b.
texico silt loam, 1 to 3 % slopes, eroded
0 bu a^{-1}
Adco silt loam, 0 to 2 % slopes
140 bu a^{-1}

Figure 5. a) Yield variation due to terracing. b) Yield variation between similar soil landscapes due to headlands.

drainage is modified due to the terrace. The terrace design and skill of the dozer operator can have a potentially large impact on this effect. Figure 5,b indicates three similar landscape positions with very different yields identified by arrows. The lower left corner of the field is an area of relatively flat deep topsoil where yields are around 140 bu a^{-1}. The left and right edges of this field are on similar landscape positions but are headlands with yield approaching 0 bu acre^{-1}. The yield measurements at these locations represent noise in the soil-landscape relationship to yield simply because of the logistical requirement to turn planting, spraying, and harvesting equipment at the field edge. Screening areas such as terraces and headlands may improve the data mining results presented here. Harvesting with high fidelity positioning systems including RTK correction and gyroscopic tilt compensation would improve the post-processing of yield monitor swath dimensions, allow better detection of overlapping, and provide more accuracy in yield maps with terracing and point rows.

Spatial, Temporal, and Producer Variability in Management

Variability in management techniques and timing both within and between producers introduce crop response effects not specifically related to soil-landscape properties. Some of these management activities are tillage, irrigation, variety, and pest management. Perhaps the largest problem with yield map data is a lack of detail about planting date or conditions. Soil conditions at planting can determine germination success and final plant population. Planting date can have a very significant impact on the climate interaction at a field specific level. For instance within the same climate year, later planting of corn may avoid early season emergence problems due to wet and cold conditions, while early planting may avoid hot dry weather during the short but sensitive pollination period. Planting dates can even be variable within the same field. Figure 6,a shows several areas harvested as a single field. They were planted on an initial date indicated by the red polygons (fig. 6,b), which also underlie the areas in green. The green polygons indicate areas that were then re-planted at a later date.

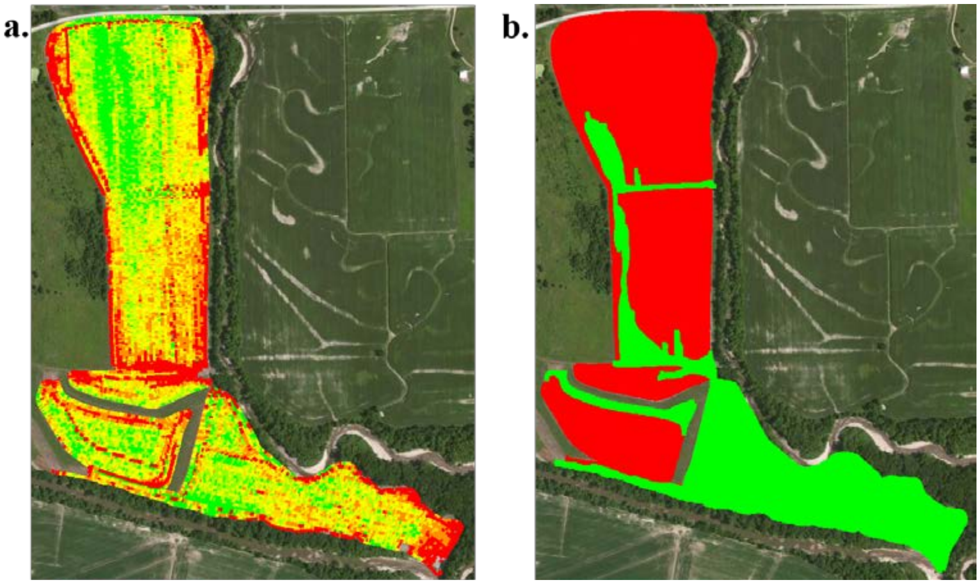

Figure 6. a) Yield maps harvested under a single field name and, b) as-planted maps for the same fields. These areas were planted on an initial date indicated by the red polygons also underlying all of the area in green. The green polygons indicate areas that were then re-planted at a later date.

Together, these images convey some critical data that are not present in the yield maps alone.

First, the soil landscape effects on yield are likely different between the red and green areas. A hiatus in planting activity or a replanting event may lead to a significant difference in the yield on similar adjacent landscape positions. Planting date determines the synchrony between the phonological stages of the crop development and the within season climate variation. None of our yield maps carry these critical pieces of information that might be predictive in a data mining algorithm. Second, the economics and production risk are worse for the replanted area. Site-specific profitability cannot be determined without data on replanting activity. Further, site-specific productivity risk cannot be accurately estimated by a model without considering the impact of whole field losses due to climate disasters. The later cases have no yield maps at all, yet have a major impact on risk assessment.

Incomplete Target and Predictor Dataspace

Analytical models require a connection or correlation to exist between the target variable and the predictors, but also require that the correct or full set of predictors is collected. The target and predictors must also be collected with adequate range. In the context of modeling soil landscape effects on yield maps this means collecting maps from all landscape positions in a representative set of climate years. The predictors should represent the full set of processes happening in the landscape or effectively stratify the population of yield maps into major response groups (e.g. irrigated versus non-irrigated). Some of the issues discussed above could be handled in a modeling procedure if information was present to be included in the model. For example, planting date may be the single most important information we lack in retrospective analysis of site specific productivity risk. Effects of climate interaction with soil-landscape might be better resolved if planting date was known because the climate year is initialized at the planting date. Another example of contextual data about a corn yield map that would be useful for modeling is data on the N application rate and timing. After soil moisture, nitrogen fertility is the single most critical determinant of corn yield. Models of soil-landscape effects on corn must assume adequate and persistent N availability for all fields equally in the absence of predictor variables that convey N status and explain variability due to that process.

Based on examination of the yield maps and unexplained variation after data mining we suggest that yield maps need an expanded yield map data model. The model should contain the management and climate information needed to better explain corn productivity and productivity risk due to soil-landscape properties for improved modeling. In the next section we provide some recommendations for the construction of yield data warehouses, and a yield map data model.

The Yield Map Data Model

The items that make up the yield map data model are a mix of point, polygon, raster and tabular data. Fusing these diverse data types into a comprehensive data model is complex but achievable within available software systems and with existing generalized data models such as the Field Operations Data Model (Map Shots, Inc, 2002). Initialization of the yield map data model occurs in spring before planting. It is fully populated by the events of the entire growing season. A proposed data dictionary for a yield map data model follows.

Field: The outermost bound of an area where crops are planted, important for planning and logistical purposes. Defining a field is a crucial step in preparing for the growing season in an organized manner. Field computers, planter controllers, and yield monitors should be initialized with the proper fields before operations. The field boundary is important in that it can attribute yield maps with uniform management information or for purposes of accounting, crop insurance, or other logistical directives. However, field is not always the unit of primary interest. Fields can commonly be split by hybrid, crop, or by hiatus in planting activity due to breakdown or weather factors. Replanting activities also commonly cover only sections of a field. Further, harvesting operations that do not cover a planted area in a short time frame may result in the need to analyze by different harvest dates due to crop losses that can occur between the harvest dates.

As-planted map: The spatial entity that attributes yield map data with crop, planting date, variety, seeding rate, starter fertilizer rate, planting conditions, down pressure, or other variable and non-variable attributes recordable at planting. The as-planted map can be a set of points or polygons representing the field area covered by sections of a planter. It could also be a set of line features representing the exact location of planted rows. This may soon be the exact location and attributes of a planted seed. The as-planted data should be considered an integral component of the yield map.

As-replanted maps: Similar to the as-planted map including as many as necessary for multiple dates with all of the same data in from as-planted maps.

Fertility status and or nitrogen application: Field specific or spatially variable information about plant available nutrients and N applications. The single most important agronomic consideration in analysis of corn yield data is the N status of the crop.

Pesticide application and effectiveness: As-applied maps or field boundary maps attributed with timing of fungicide, insecticide and herbicides applications. This might also include whole field ratings or spatial maps of pesticide effectiveness from field scouting. A low pesticide effectiveness rating might be useful as a screening variable or predictor in a data mining situation.

Climate/Irrigation data: Season long local or field specific measurements of rainfall, irrigation schedule, air and soil temperatures, soil moisture, and important climate events such as a hailstorm or frost damage. Climate metadata can be populated from a nearby climate network's weather station if planting date is available. The yield data model should also permit storage of field or farm specific daily weather data when it is recorded from a dedicated producer owned weather station. The climate in which crops develop is critical information to retrospectively model productivity risk from a yield map data warehouse.

Crop phenological timeline: Critical crop phonological stages such as emergence, major vegetative stages, major reproductive stages (e.g. tasseling, flowering, pod fill), and senescence should be tracked by as-planted areas. This could include maps showing spatial variability of key stages such as tasseling, flowering, or senescence obtained by remote imagery. The phonological timeline provides information about the length and progression of the growing season useful as in season predictive variables.

Yield map: The target variables, grain yield and moisture measurements recorded by the yield monitor along with appropriate sentences from the NMEA-0183 string (Trimble, 2004). The NMEA string includes the most important metadata, position, position accuracy, and time. A fuller NMEA string will allow post-processing and filtering of data points with poor accuracy.

Yield monitor settings: Applied by producers in the software of yield monitors such as swath width, antenna position, filters, and thresher delay.

Yield monitor calibration metadata: Calibration loads and measured weights, calibration models, accuracy and quality control statistics. Calibration metadata is particularly important at the data-warehouse stage as yield maps from a wide range of sources are aggregated. A few outlier maps with poor calibrations might have a large impact on some modeling procedures. Calibration quality might be used as a screening or weighting factor in a data mining procedure.

Raw data and version control: The raw data collected in the field is too often an ephemeral product that disappears when the yield monitor is cleared in the fall before or after harvest. Commonly, the producer will export the data from the monitor after application of some of the manufacturer provided data filters which may include GPS track correction, header up/down toggles, low and high pass filters, start and end pass delays, and (critically) a default or farmer specific thresher delay setting. Often these filters are selectable or have adjustable parameters which are not necessarily retained by the yield map data file after this point. Effectively the raw data is lost as only the processed yield map is retained. If the delay setting is wrong and start/end pass delays are enforced at export, then good data points are eliminated and correcting the delay leads to lost data at the start or end of a transect. This is further complicated if the producer or service provider knowingly or unknowingly applies additional filtering to the yield map in their desktop software. For these reasons raw uncorrected versions of the harvesting data should be retained. Processing, filtering and correcting of this data should be re-applied to the raw data as improvements are made, and version control should be implemented to maintain a history of events.

SUMMARY AND RECOMMENDATIONS

Algorithms are inefficient at extracting information from predictors when noise or variation not characterized by the predictors is present in the target variable. Algorithms are challenged to make valid estimates for future cases when predictors are not available to explain certain signals in the target, or when the data-space does not span the full range of possible outcomes, and especially when the measured target variable is in error due to mistakes, and systematic or random noise. Yield data warehouses are prone to five general issues that cause problems for data mining or statistical analysis of soil-landscape effects on yield and yield risk:

i) yield-landscape inversion between climate years
ii) sensor system malfunction and inaccuracy
iii) poor data management practices and operator error
iv) field configuration and logistical limitations
v) spatial, temporal, and producer variability in agronomic management
vi) incomplete target and predictor dataspace

Yield maps can be complex. They can have tens of thousands of data points with intricate spatial structure, yet even with all of this information, yield maps are difficult to analyze without fuller contextual details. How much rainfall was received? When did it fall? When was the planting date? These are just some of the more critical pieces of metadata that must be available to interpret the complex spatial and temporal patterns in individual yield maps. Producers may be able to manage these details for small collections of fields but as time passes, these details can become lost. As the number of yield maps increases, a dedicated yield map metadata system is needed. When the analyst of an entire population of yield maps is far removed from events in the field, contextual metadata about climate, crop phenology, and special issues must be organized and fully populated else quantitative interpretation and analysis are limited in their effectiveness. Adoption of a more agronomically complete yield map data model could improve analysis of yield map data warehouses.

REFERENCES

Ag Leader Technology, Inc.. 2011. Spatial Management System Basic. Ag Leader Technology, Inc., Ames, IA.

Böhner, J., and O. Antonić. 2009. Land-surface parameters specific to topo-climatology. p. 195–226. *In* Geomorphometry Concepts, Software, Applications. Elsevier.

Breiman, L. 2001. Random forests. Machine learning 45(1): 5–32.

Drummond, S.T., and K.A. Sudduth. 2012. Yield Editor Version 1.02 Beta. Available at http://www.ars.usda.gov/services/software/software.htm (verified 5/8/2012).

Drummond, S.T., K.A. Sudduth, A. Joshi, S.J. Birrell, and N.R. Kitchen. 2003. Statistical and neural methods for site-specific yield prediction. Trans. Am. Soc. of Agric. Engin.46(1): 5–16.

Florin, M.J., A.B. McBratney, and B.M. Whelan. 2009. Quantification and comparison of wheat yield variation across space and time. Eur. J. Agron. 30(3): 212–219.

Freeman, T.G. 1991. Calculating catchment area with divergent flow based on a regular grid. Comp. Geosci. 17(3): 413–422.

Gallant, J.C., and T.I. Dowling. 2003. A multiresolution index of valley bottom flatness for mapping depositional areas. Water Resour. Res. 39: 13 PP.

Gesch, D., M. Oimoen, S. Greenlee, C. Nelson, M. Steuck, and D. Tyler. 2002. The National Elevation Dataset. Photogram. Eng. Rem. Sens. 68(1):7–11

Green, T.R., J.D. Salas, A. Martinez, and R.H. Erskine. 2007. Relating crop yield to topographic attributes using spatial analysis neural networks and regression. Geoderma 139(1–2): 23–37.

Grisso, R.D., P.J. Jasa, M.A. Schroeder, and J. Wilcox. 2002. Yield monitor accuracy: Successful Farming magazine case study. Appl. Eng. Agric. 18(2): 147–152.

Hjerdt, K.N., J.J. McDonnell, J. Seibert, and A. Rodhe. 2004. A new topographic index to quantify downslope controls on local drainage. Water Resour. Res. 40: 6.

Irmak, A., J.W. Jones, W.D. Batchelor, S. Irmak, K.J. Boote, and J.O. Paz. 2006. Artificial neural network model as a data analysis tool in precision farming. Trans. ASABE 49(6): 2027.

Jiang, P., S.H. Anderson, N.R. Kitchen, K.A. Sudduth, and E.J. Sadler. 2007. Estimating plant-available water capacity for claypan landscapes using apparent electrical conductivity. Soil Sci. Soc. Am. J 71(6): 1902.

Jung, W.K., N.R. Kitchen, K.A. Sudduth, and S.H. Anderson. 2006. Spatial characteristics of claypan soil properties in an agricultural field. Soil Sci. Soc. Am. J. 70(4): 1387.

Kitchen, N.R., S.T. Drummond, E.D. Lund, K.A. Sudduth, and G.W. Buchleiter. 2003. Soil electrical conductivity and topography related to yield for three contrasting soil-crop systems. Agron. J. 95(3): 483.

Kravchenko, A.N., and D.G. Bullock. 2000. Correlation of corn and soybean grain yield with topography and soil properties. Agron. J. 92(1): 75.

Liaw, A., and M. Wiener. 2002. Classification and regression by randomForest. R News. 2(3): 18.

Map Shots, Inc. 2002. Field operations data model. Map Shots, Inc., Cumming, GA.

Norouzi, M., S. Ayoubi, A. Jalalian, H. Khademi, and A.A. Dehghani. 2009. Predicting rainfed wheat quality and quantity by artificial neural network using terrain and soil characteristics. Acta Agric. Scan., B - Soil & Plant Sci. 60(4): 341.

Riley, S.J., S.D. DeGloria, and R. Elliot. 1999. A terrain ruggedness index that quantifies topographic heterogeneity. Intermount. J. Sci. 5(1-4): 23.

Serele, C.Z., Q.H.J. Gwyn, J.B. Boisvert, E. Pattey, N. McLaughlin, and G. Daoust. 2000. Corn yield prediction with artificial neural network trained using airborne remote sensing and topographic data. p. 384 –386 vol.1. *In* Proceedings of the Geoscience and Remote Sensing Symposium.

Soil Survey Staff. Soil Survey Geographic (SSURGO) Database for Missouri. U.S. Department of Agriculture.

Sudduth, K.A., and S.T. Drummond. 2007. Yield Editor: software for removing errors from crop yield maps. Agron. J. 99(6): 1471.

Tarboton, D.G. 1997. A new method for the determination of flow directions and upslope areas in grid digital elevation models. Water Resources Research 33(2): 309.

Trimble. 2004. NMEA-0183 Messages Guide for AgGPS Receivers. Trimble Navigation Ltd., Overland Park, KS.

Zevenbergen, L.W., and C.R. Thorne. 1987. Quantitative analysis of land surface topography. Earth Surf. Proc. Landf. 12(1): 47–56.