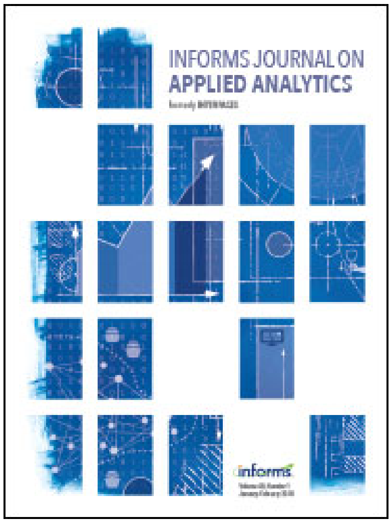

INFORMS Journal on Applied Analytics

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Advanced Analytics for Agricultural Product Development

Joseph Byrum, Craig Davis, Gregory Doonan, Tracy Doubler, David Foster, Bruce Luzzi, Ronald Mowers, Chris Zinselmeier, Jack Kloeber, Dave Culhane, Stephen Mack,

Interfaces
Vol. 46, No. 1, January–February 2016, pp. 5–17
ISSN 0092-2102 (print) | ISSN 1526-551X (online)

informs®
http://dx.doi.org/10.1287/inte.2015.0823
© 2016 INFORMS

THE FRANZ EDELMAN AWARD
Achievement in Operations Research

Advanced Analytics for Agricultural Product Development

Joseph Byrum, Craig Davis, Gregory Doonan, Tracy Doubler, David Foster,
Bruce Luzzi, Ronald Mowers, Chris Zinselmeier
Syngenta, Slater, Iowa 50244
{joseph.byrum@syngenta.com, craig.davis@syngenta.com, gregory.doonan@syngenta.com, tracy.doubler@syngenta.com,
david.foster@syngenta.com, bruce.luzzi@syngenta.com, ronald.mowers@syngenta.com, chris.zinselmeier@syngenta.com}

Jack Kloeber, Dave Culhane, Stephen Mack
Kromite LLC, Lambertville, New Jersey 08530
{jkloeber@kromite.com, dculhane@kromite.com, sbmack@kromite.com}

Syngenta, a leading developer of crop varieties (seeds) that provide food for human and livestock consumption, is committed to bringing greater food security to an increasingly populous world by creating a transformational shift in farm productivity. Syngenta Soybean Research and Development (R&D) is leading Syngenta's corporate plant-breeding strategy by developing and implementing a new product development model that is enabling the creation of an efficient and effective soybean breeding strategy. Key to the new strategy is the combination of advanced analytics and plant-breeding knowledge to find opportunities to increase crop productivity and optimize plant-breeding processes. Syngenta uses discrete-event and Monte Carlo simulation models to codify Syngenta Soybean R&D best practices, and uses stochastic optimization to create the best soybean breeding plans and strategically align its research efforts. As a result of using these new analytical tools, Syngenta estimates that it will save more than $287 million between 2012 and 2016.

Keywords: simulation; optimization; data quality; agriculture; breeding; soybean.
History: This paper was refereed.

The world's population is growing at a daily rate that exceeds 200,000; by 2050, the total population will increase by one-third to 9.6 billion (United Nations 2013). Thus, 2.4 billion more people will need to be fed; however, fewer water, land, and energy resources will be available to support the required increase in crop output.

Crop production can only increase by expanding the cultivated land area or by improving crop output on the existing land area. Expanding the cultivated land area is not environmentally sustainable or socially responsible. A more efficient solution, and one that does not require the use of additional water, energy, or land resources, is to improve crop output through plant breeding. However, with current breeding methodologies, the rate of increase in crop production is insufficient to meet today's food needs. Currently, 805 million people go to bed hungry each night (Food and Agriculture Organization of the United Nations 2014). To address this, plant breeding must undergo a fundamental, data-driven transformation within the next few decades to improve crop production; otherwise, millions more will go hungry.

Until now, the plant-breeding industry has not taken full advantage of the sophisticated data analysis that has fundamentally transformed other industries.

Syngenta, a leading developer of crop varieties (seeds) that provide food for human and livestock consumption, is changing that by applying operations research (OR) methods to enable better decisions, which result in reducing the time and cost required to develop higher-productivity crops. This data-based transformation in the breeding process makes a quantifiable contribution to Syngenta's commitment to meeting the world's growing food needs in an economically and environmentally sustainable way. Providing adequate food for the world's population is one of humanity's toughest challenges.

In September 2013, Syngenta announced the launch of its Good Growth Plan, consisting of six commitments designed to address the global food-security challenge (see http://www.syngenta.com/global/corporate/en/goodgrowthplan/home/Pages/homepage.aspx). This plan reflects Syngenta's belief that agricultural productivity must increase to feed a rising global population. One of its specific, ambitious, and measurable commitments is to increase the average productivity of the world's major crops by 20 percent without using more land, water, or other inputs (i.e., resources) (Syngenta 2015). As Mike Mack, Syngenta Chief Executive Officer, says, "The Good Growth Plan represents our collective commitment as a company to do things differently and better. We know we can't solve the challenges alone, which is why we are bringing together stakeholders from across the world to share our intent and to benefit from their input" (Syngenta 2013). It is in this spirit that Syngenta Soybean R&D has brought together OR tools and soybean breeding to help solve the global food-security challenge.

Syngenta Soybean R&D: A Strategy

Syngenta Soybean R&D developed a novel plant-variety development strategy that combines advanced analytics with sound scientific, mathematical, genetic, and breeding principles to improve the quality and quantity of the soybeans that farmers produce per acre. Soybean R&D demonstrated the successful application of this strategy by addressing the following objectives:

- Increase the frequency of favorable traits within the population of soybean plant varieties;
- Reduce the time required to develop new soybean plant varieties with favorable traits;
- Build a process to efficiently transfer favorable traits among soybean plant varieties;
- Improve data quality, prediction of variety performance, and characterization of environments;
- Make better decisions to positively impact the probability, cost, and timeline of developing a new soybean plant variety.

Soybean-breeding project teams are responsible for implementing strategy and for developing commercial soybean varieties. Each breeding project team consists of members with different functional skills and backgrounds. The project lead is a plant breeder by training and is responsible for making decisions that affect the development of new soybean varieties in his (her) breeding project.

Syngenta Soybean-Variety Pipeline

The soybean-variety pipeline (pipeline) produces commercial (elite) soybean varieties containing traits that enable it to thrive in different environmental conditions and against various diseases and pests, while providing higher yields than the varieties currently available. At any one time, the pipeline contains plants in each of five pipeline phases (Figure 1):

1. Variety design. During this initial phase, the project lead mates two soybean varieties (parent varieties) to produce progeny in which half of its approximately 46,000 genes are inherited from each parent. As such, each progeny contains a unique combination of genes, resulting in a unique combination of characteristics (traits). The project lead chooses the parent varieties for mating based on their performance relative to other soybean varieties in evaluation trials, or because they carry favorable traits. Desirable progeny selected by the project lead advance to the next phase.

2. Variety advancement. During this phase, progeny advance via self-pollination (mating of a plant with itself) for four generations if they have the appropriate combination of favorable traits. Self-pollination ensures that favorable traits become permanently fixed (homozygous) within the new variety. The selected progeny become experimental varieties and advance to the next phase.

3. Variety evaluation. During this phase, yield and other important traits are measured in each of three field-trial stages (stage 1, stage 2, and stage 3) that are

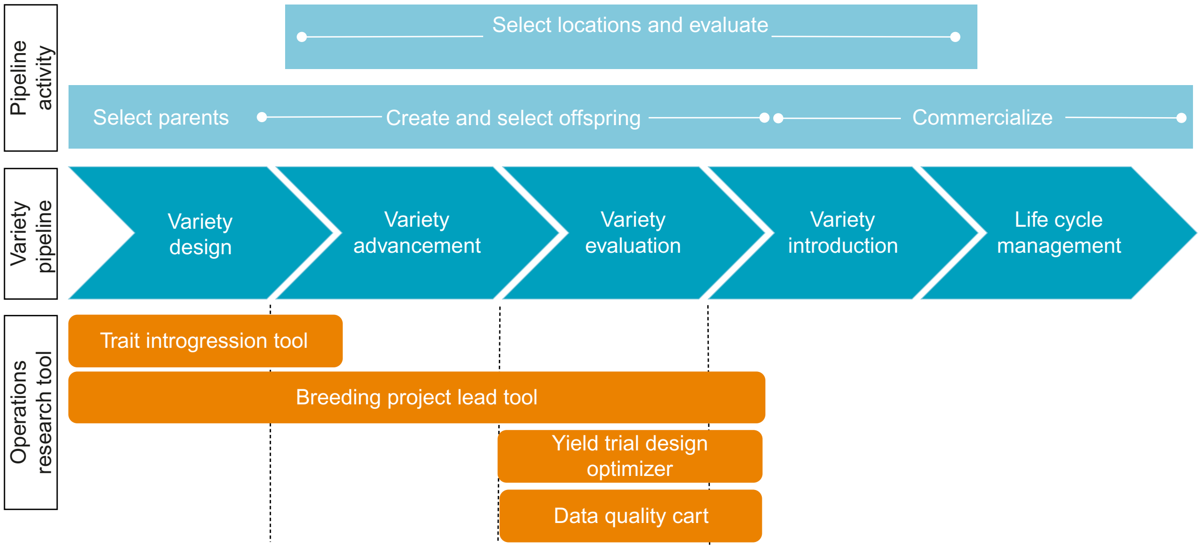

Figure 1: (Color online) We developed a suite of OR tools to support the development phases of the variety pipeline. Pipeline activities are aligned above each pipeline phase they support; OR tools are aligned below each pipeline phase they support.

replicated across and within different environments (locations). Only experimental varieties that outperform current commercial varieties become new commercial varieties.

4. Variety introduction. Experimental varieties that have been promoted to commercial status are evaluated during this phase in yield trials conducted across a range of environments, which represent those within the commercial market region.

5. Life cycle management. The life span of a commercial variety (the period of time during which it is sold commercially) is determined during this phase by evaluating its yield and sales performance relative to the performance of new varieties that could replace it.

Decisions that affect the cost, time, and success of developing a commercial soybean variety are made in phases 1, 2, and 3, and decisions relative to commercializing are made during phases 4 and 5. As such, the discussion of tool use will focus only on phases 1–3.

A project lead also makes decisions that improve the parent varieties used in the variety-design phase of the pipeline with a process we refer to as trait introgression (TI). During the TI process, the project lead uses the backcross breeding method (i.e., the mating of a plant with one of its parents) to transfer (introgress) favorable traits from one soybean variety (donor variety) into another soybean variety (elite variety). An elite variety is typically a current commercial variety. With the backcross method, the progeny of a prior mating between a donor variety and an elite variety are mated once again to the elite variety to increase the percentage of elite variety genetic background in the progeny. The percentage of elite-variety genetic background that is desired in the new variety dictates the number of backcross generations that are made. The TI process is complete when a new variety that contains the desired percentage of genetic background of the elite variety and the favorable traits of the donor variety is available for use in the variety-design phase of the pipeline.

Soybean R&D Challenge

Along the pipeline, some factors are uncontrollable (e.g., environmental conditions) and unpredictable (e.g., variety performance), thus making it a complex and dynamic process. While developing each pipeline, the project lead must consider approximately 200 binary decisions (with 1.6×10^{60} possible outcomes). These binary decisions determine, for example, which varieties to mate, which traits to select, and where, when, and how to evaluate and advance varieties.

Each decision can impact the pipeline design and the cost, time, and probability of successfully developing a commercial variety.

Historically, companies have managed such complexity by heavily resourcing and expanding the size of their development programs. Syngenta's challenge is to design a pipeline that allows it to effectively develop higher-yielding varieties with the most efficient use of resources and time, while also maintaining the rate of genetic gain of other companies.

It takes approximately six years from initiating a project in the pipeline (i.e., identifying parent varieties for crossing) to identifying a commercial variety. Most decisions made for a specific pipeline activity affect, and often limit, decisions made for downstream pipeline activities. Without modeling capabilities, the project lead must wait the full six years to realize the impact that a specific combination of decisions has on the success, cost, and time of the pipeline design; however, evaluating the impact of all possible decision options is impossible and inefficient. In addition, the learning curve for a project lead is such that becoming an expert typically requires 15–20 years of experience. Conversely, when the consequences of decisions made by the project lead can be predicted accurately, it is possible to rapidly evaluate and choose between designs, achieve an effective and efficient pipeline for developing commercial soybean varieties, and maintain a competitive industry rate of genetic gain at an optimal investment level.

Solutions

To evaluate the consequences of project lead decisions on the variety-development pipeline, Syngenta developed four OR tools to support the first three phases of the pipeline. The TI and breeding project lead (BPL) tools facilitate decisions related to the development of efficient and effective pipeline processes, and the yield trial design (YTD) optimizer and data quality cart (DQC) tools support decisions that impact the quality of variety selection and advancement toward commercialization.

Planning a Trait Introgression Program: The Trait Introgression Tool

The TI tool evaluates mating decisions that are made by a project lead during the variety-design phase of the pipeline. Specifically, it evaluates the time, cost, and probability of successfully transferring one or more traits from one soybean line into another based on project lead decisions. Two decisions that have a large impact the TI process are (1) number of donor and elite varieties used for mating and (2) number of traits being introgressed. Once the specific TI process has been determined, mathematical computations that are driven by outputs of the TI tool (e.g., number of progeny produced from mating, genetic composition of progeny) and by inputs provided by the project lead (e.g., genetic composition of parent plants, number of mating attempts) change the probability distributions for time, cost, and success.

The TI tool uses discrete-event simulation to model the flow of a process, the biology of creating progeny from mating two plants, and the calculations needed to track genetic segregation according to Mendel's principles of genetics (Sleper and Poehlman 2006). Mendel (Sleper and Poehlman 2006) developed probabilistic inheritance principles of a simple trait, given the genetic composition of each parent variety used in mating. Each variety has two alleles (genetic factors) for each trait, and each allele can have a positive (favorable) or negative (unfavorable) impact on a trial. Thus, for a given trait, a plant may contain two favorable alleles, two unfavorable alleles, or one favorable and one unfavorable allele. The number of a plant's favorable alleles for a given trait is modeled using the binomial distribution with parameters of two trials (because of two alleles) and the probability of a favorable allele determined as a function of a plant's genetic composition. Self-pollination increases the frequency of progeny that contain two unfavorable alleles or two favorable alleles.

The TI tool models genetically distinct varieties and evaluates different sequences of key processes chosen by the project lead, such as the specific generations in which self-pollination and cross-pollination will occur and whether the process requires one or two donors. Each process can affect the attrition rate of progeny and the genetic makeup of the progeny in a probabilistic way. The last step processed by the tool for each progeny generation is to determine whether a sufficient number of seeds is available from the preceding self- or cross-pollination activity with the desired combination of traits from the donor and elite parents to meet the needs of the next self- or cross-pollination activity.

Although cost and time are key outcomes of the TI tool, knowing both the expected number of seeds with the desired combination of traits and the uncertainty surrounding that expectation are critical to selecting the best breeding plan. Discrete-event simulation allows a project lead to plan each process step and view the consequences of the plan on cumulative cost, cumulative time, and genetic composition of the progeny at the end of each step in the TI process. If the cost, time, or probability of successfully transferring the desired traits (POS) of the planned TI process is unacceptable, the project lead can change one or more decisions and quickly rerun the simulation.

Prior to the availability of the TI tool, a project lead developed introgression plans aided by software programs that focused solely on the averages of distributions and selections, usually for one donor variety and one trait. The TI tool provides the project lead with a reliable estimate of cost, time, and POS with multiple donor varieties, different introgression pathways (sequences of self- and cross-pollination activities across progeny generations), and multiple traits. In addition, the model generates a resource-utilization report for planning and budgeting purposes.

For many parameters in the TI tool, the model relies on probabilistic calculations using value parameters that are uncertain and must be estimated from experts' experiences using uniform, normal, or triangular distributions. This is true for (1) plant growing rate, (2) seed yield of progeny from soybean mating and backcrossing, and (3) number of parent-variety plants needed to support the number of mating attempts requested by the project lead.

Implementing the Trait Introgression Tool

A project lead interfaces with the TI tool using a dashboard. All decisions that are required to determine the cost, time, and success of a TI process are made directly on the dashboard, which can access all outputs generated by the TI tool. The dashboard makes the complex TI process less error prone and more manageable. After entering decisions into the dashboard and activating the tool, the project lead executes a simulation run. The output contains a summary of the run results, including cost, time, and POS.

The TI tool uses project lead decisions regarding the introgression pathway and number of backcrosses to calculate the time required to develop a new variety for the pipeline. The cost of a TI process is based on the number of plants required to support the number of mating attempts that the project lead selected for the initial cross and each backcross. Greenhouse resources are required for each seed planted and each seed harvested. In developing a TI process, the project lead can indicate in the tool interface that genetic markers, a major expense within the TI process, will be used to identify plants containing the favorable traits.

The TI tool provides an estimate of the probability of successfully developing a new variety containing the desired combination of traits from the donor and elite parent varieties based on breeder inputs into the model. The number of favorable traits being transferred from a donor variety into an elite variety is the most significant factor impacting the probability that a given TI process will successfully produce a new variety. This number influences the number of mating attempts required to generate a population large enough for a reasonable probability of developing a new parent variety.

Table 1 shows the results of two TI designs on the success, cost, and time of a TI project. The goal of each design is to successfully develop up to 20 new varieties during the variety-design phase. The designs differ

	Design	
	#1	#2
TI decision		
No. of donors	1	1
No. of traits	5	5
Initial no. of mating attempts	20	200
No. of backcrosses	1	1
No. of backcross attempts	200	200
No. of elite parents	20	20
TI tool result		
Success (no. of new parents)	3	18
Cost per success	Base	−80%
No. of days	310	310

Table 1: The table illustrates the results of two alternative TI tool designs on the success, cost, and time of a trait introgression project.

only in the number of mating attempts (200 versus 20) requested for the initial cross and for each backcross generation. Although the times required to complete the TI processes are the same, Design 1 resulted in 18 new parents; Design 1 resulted in only three. In addition, the cost per successfully developed parent is 80 percent less for Design 2 than for Design 1. The decision to increase the number of mating attempts allows more progeny with the favorable traits to be available for the next step of the TI process. The cost of additional mating attempts is low relative to the positive impact on POS. Without the TI tool, the project lead would be unable to accurately estimate the frequency of success for either design.

Planning a Variety Development Pipeline: The Breeding Project Lead Tool

The BPL tool is designed to assist a project lead in planning and optimizing a soybean-variety development pipeline. It evaluates and determines the best use of available facilities, fields, and plant materials by using several OR analytical methods. The following factors drove the development of a discrete-event simulation platform: (1) discrete and continuous probability distributions, (2) elements interacting at different times, (3) continuous time tracking, and (4) variability of key planning values.

The need to accurately simulate the complexities of Syngenta's soybean breeding program inspired the development of the BPL tool. The breeding program is complex because it involves decisions affecting trial location(s) and population size, selection intensity and attrition rate, government and company regulations, traits and trait analyses, assay accuracy, time, and costs. The dependencies among these variables are critical for the process and add to its complexity. To ensure that the genetic composition of an individual seed or a population of seeds as it advances through multiple generations is realistic and accurate, the applicable principles of genetic segregation are modeled. In addition, genetic principles must be applied to the known genetic composition of the varieties to track multiple traits as they are assayed and selected across multiple generations of progeny.

The BPL tool contains a customized and evolutionary algorithm (optimizer) to support a project lead in designing the activities within each phase of the variety development pipeline. This algorithm enables the tool to make a large number of runs of a single design (each design represents a specific combination of activities across all phases of the pipeline), use complex combinations of aggregate statistics within the objective function, and compare results with other designs, while striving to identify the optimal pipeline design. The tool optimizer continues to maintain a population ($n = 10$) of primary solutions, giving the project lead both the optimal design and nine additional outstanding solutions. The objective function minimizes (1) the attrition rate in the variety design and variety advancement phases of the pipeline, (2) the average cost (weighted for attrition versus success), and (3) the development time (variety design through variety evaluation) of each design.

The BPL tool uses Time Visualizer (a Microsoft Excel-based graphics tool) and Expert System (an ExtendSim-based warning system) to identify and display pipeline delays and policy violations, respectively, caused by project lead decisions.

Implementation of the Breeding Project Lead Tool

The project lead uses the BPL tool dashboard to enter design inputs for simulations and to view all tool-generated outputs. The tool dashboard guides the project lead through the development phases of the pipeline, variety design through variety evaluation, by asking the lead to make decisions about each activity in the development process that affects the cost, time, and success of that activity and the overall pipeline.

Each set of BPL simulation runs refers to a pipeline design that begins in the variety design phase with a mating between two parent soybean varieties. The project lead defines the genetic composition and relative maturity of each parent variety, the date and geographic location for the mating, and the number of first-generation progeny desired. The chosen location must be capable of supporting plant growth during the selected date and the maturity of the varieties. The number of seeds of the progeny impacts the probability of developing a new soybean variety and the downstream costs.

The project lead chooses the location for growing each generation during the variety advancement phase. The choice of locations may be limited depending on the progeny generation, plant maturity, and seed availability. The location determines the cycle duration, cost per population, and seed yield per plant. At the second and third progeny generations, the project lead chooses genetic markers to assist with trait selection, makes selections in the third or fourth progeny generations, and chooses the number of plants to select and the location to be used for growing the plants and making selections. The location decision impacts the timeline, costs, and planting date.

The project lead also chooses an experimental design (e.g., number of replications and locations), the number of lines selected for advancement (e.g., selection intensity) from generation to generation, and molecular markers and (or) disease or pest assays to assist with variety selection. These choices significantly impact costs.

As the BPL tool conducts simulation runs, it generates a significant amount of data during each process step. At the conclusion of all simulation runs, the pipeline design is evaluated relative to its impact on aggregated time, cost, and success in identifying commercial varieties.

Table 2 shows an example of a project lead using the BPL tool. The lead enters an initial pipeline design to achieve the breeding objective, simulates the design using the BPL tool, and records the results. The tool's evolutionary algorithm is then used to find a better design, if one exists. The project lead compares the results of both designs relative to cost, time, and success. In this example, the BPL tool designed a pipeline that was as successful as the project lead's design but cost 23 percent less; however, it used a different set of locations and took 360 days longer to complete. Based on project lead input, the BPL tool weighted the importance of cost as higher than the importance of time. The project lead can change the weighting among cost, time, and success parameters according to the goals established for the pipeline.

	Pipeline designer	
Pipeline decision	Project lead	BPL tool
Cross location	United States	United States
1st progeny location	Puerto Rico	Hawaii
2nd progeny location	Puerto Rico	Puerto Rico
3rd progeny location	Puerto Rico	Puerto Rico
4th progeny location	Chile	Puerto Rico
5th progeny evaluation	United States	United States
Stage 1 location	Argentina	United States
Stage 2 location	United States	United States
Stage 3 location	United States	United States
BPL tool result	Base case	Difference from base
Total cost	Base cost	−23%
No. of successful progeny	Success rate	0%
No. of days	1,800	+360

Table 2: This example compares the results of project lead and BPL tool decisions on the success, cost, and time of a variety development pipeline design.

Designing Field Trials: The Yield Trial Design Optimizer

A project lead's objective is to economically and accurately identify and advance experimental varieties that meet commercial expectations with regard to variety performance in specific geographical areas and environmental conditions. To be successful, the lead must design trials with the optimal number of varieties, locations, and within-location replications of varieties. The YTD optimizer was developed specifically for this purpose. The experimental design the project lead selects includes a BPL tool scenario that identifies a pipeline design to maximize the probability of both introducing superior varieties into the variety evaluation phase and correctly identifying superior varieties for commercialization.

The YTD optimizer model examines the solution space (experimental designs) through a large randomization of feasible solutions across a specified range of design parameters and costs. The model takes historical data and uses these data to generate data vectors of statistics to fit the probability distributions needed to drive the simulation model, which replicates the processes in the variety evaluation phase. After validating the simulation model, the YTD optimizer creates and evaluates a large number of alternative experimental designs for performance and cost. The analysis identifies the best-yielding designs for a given set of resources.

The YTD optimizer model requires three probability distributions as the key descriptors of yield performance for a cohort of varieties. They represent the variability of intrinsic yield, variability effects on yield because of location (abiotic and biotic conditions), and variability effects on yield across replications of varieties at each location. Those distributions provide the parameters used in probability distributions embedded in the model to simulate yield at each trial stage, for each replication, and at each location. The data vectors are imported into Microsoft Excel, and the analysis fits probability distributions to these data vectors using the @Risk Excel add-in (Palisade 2014a). The data analyses produce a fitted distribution for each stochastic input parameter in the model: AvDiff, location variance, and replication variance. AvDiff is a normalized function of yield to account for the variability of average yield of varieties planted at different locations. Ideally, varieties with high intrinsic AvDiff and low variance are advanced in the variety evaluation phase. However, superior varieties can be masked by inferior varieties with lower AvDiff but higher variance that randomly deliver superior yields. The probability distributions in the simulation model replicate this stochastic phenomenon.

A simulation run generates a number of varieties, each with a mean AvDiff, location variance, and replication variance drawn from the appropriate probability distributions. In each trial stage, the model calculates mean yield for each variety after random draws across the number of locations and number of replications. The top mean-yielding varieties identified (the number of varieties varies depending on selection intensity) advance to the next trial stage. The AvDiff assigned to the top-yielding variety becomes the measure of outcome for each simulation run.

A team of project leads established a reasonable range for each of the nine design variables (combination of input parameters and development stages). A presimulation utility creates 1,800 simulation designs by randomly generating decision variable values within their reasonable ranges. Cost and other secondary variables also constrain the creation of feasible designs. An Excel Visual Basic routine quickly executes the randomization process and generates a design data set.

Implementing the Yield Trial Design Optimizer

A two-phase simulation approach keeps the execution times of an analysis within reasonable bounds. Phase 1 of the analysis simulates each of the 1,800 designs using 400 iterations. The objective of phase 1 is to identify an initial set of high-quality candidate designs across the entire cost range. Using an iterative Pareto algorithm, the YTD optimizer ranks the outputs from phase 1, identifying the top 10 percent (180) of the designs across the span of costs and advances them to phase 2. It again simulates these top designs using 2,000 iterations to reduce the confidence interval of the resulting means. Plotting the 180 designs reveals the AvDiff cost pairs that constitute the efficient frontier (Figure 2).

The YTD optimizer uses the ExtendSim platform and discrete-event simulation as the base model. The benefit of such a discrete-event formulation is that probability distributions are sampled only for varieties that advance to the next trial stage in the variety evaluation phase. Modeling a variable number of trial stages is also easier. In the formulation, the varieties are the discrete items and the events are the three trial stages. The randomly assigned distribution parameters used to simulate yield are item attributes. Varieties flow through model blocks, which generate the simulated yields at each trial stage, sort varieties by simulated AvDiff, and route the high-yielding varieties to the next trial stage, while deleting the remainders from the system.

The YTD optimizer integrates an Excel interface with the simulation model to permit both input of a large number of designs and output of the results for each design. In a postprocessing analysis, the neural network Excel-based software, NeuralTools (Palisade 2014b), assesses the relative contribution of each of the nine decision variables to the target output variable (AvDiff). Instead of analyzing only the few designs on the Pareto frontier, the analysis process uses NeuralTools to analyze the complete set of phase 1 designs, which includes both many poor and many efficient designs. This allows the neural network model to be more robust.

The deterministic meta-model quickly evaluates a given design and finds the optimal design within cost constraints with nearly the same accuracy. A variable impact analysis rank ordered the impact of key decision variables on AvDiff. The results were consistent across the data sets evaluated and confirmed the generalized recommendations for more efficient resourcing.

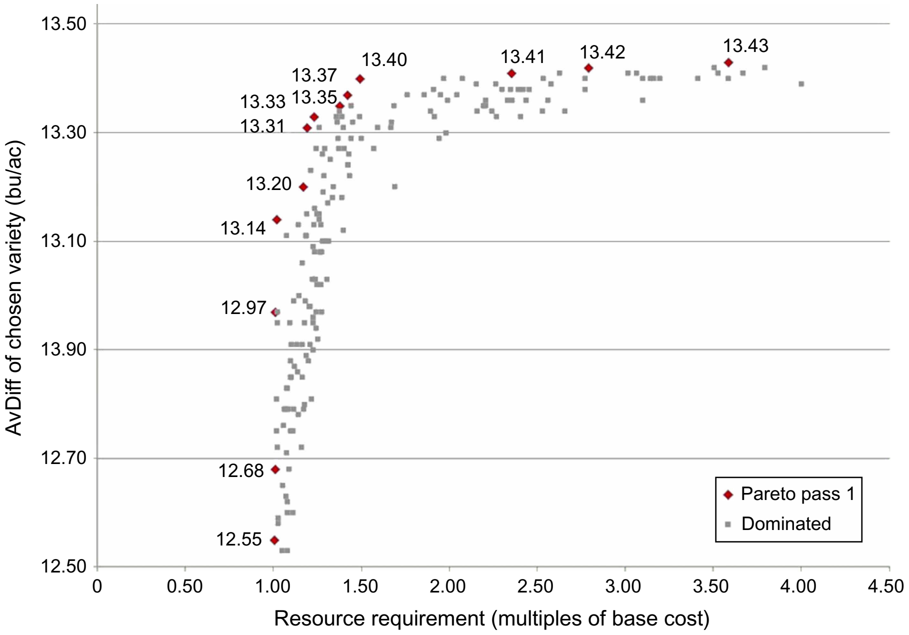

Figure 2: (Color online) The graph shows plotted AvDiff in bushels of soybean produced per acre (bu/ac) vs. trial design cost for data of varieties with a relative maturity between 2.0 and 2.5. The shape of the curve is similar for each variety data set evaluated.

Improving Data Quality: The Data Quality Cart (DQC) Tool

The environment, whether it is a controlled greenhouse environment or an uncontrolled and variable field environment, impacts nearly every activity that occurs across the pipeline. Unfortunately, the most resource-intensive and critical activities are conducted in the field environments. Therefore, extracting useful and reliable data is important to keeping a pipeline on its timeline and for maintaining cost effectiveness. If data from field locations are not usable, the organization loses the benefits of an optimal pipeline.

In a field situation, environmental variability is caused by an individual factor or, more likely, a combination of factors (e.g., soil type, topography, disease, insects, weather). During the growing season, the impact of environmental variation is sometimes evident from the within-field fluctuations in plant characteristics, such as height, color, and vigor. At other times, the impact of the environment is more subtle and difficult to recognize. In either case, the result has a significant impact on the data being collected. Because undetected factors cause data variability, the analysis and interpretation of those data lead to incorrect conclusions and, for a pipeline, can lead to the selection and advancement of inferior varieties. Realizing that the majority of the time, within-field variation will be visually undetected, Syngenta developed the data quality cart (DQC) tool. This tool uses OR methodology to improve data quality before the data undergo any statistical analyses or evaluation by a project lead during the development of a new variety.

The DQC tool conducts a residual analysis on data uploaded from stage 1, stage 2, and stage 3 trials. A residual is the calculated difference between a trait's observed value and its predicted value based on a statistical model. For example, we might define a model for a yield trial as mean + entry effect + location effect + error.

The residuals indicate shortcomings of the model and are a result of (entry × location) effects + systematic error + random error. Random error is often identified as an outlier. It may result from variability in data measurement as opposed to resulting from variability in the growth environment. Patterns that emerge from the data, which are not related to an entry (genotype) or treatment, identify systematic error. These patterns are often the result of nonrandom similarities or differences factors such as soil, weather, or disease, or they might result from using field equipment or chemical applications.

Implementing the Data Quality Cart Tool

All field-trial data are evaluated with the DQC tool to make variety advancement decisions for Soybean R&D breeding projects. One output of the DQC tool is a map-based display using residuals (Figure 3), which project leads use to assess the appropriateness of prediction models. A random distribution of residuals indicates that the impact of the environment on the data is relatively uniform. In contrast, groups of high and low residuals creating patterns in a map-based display of residuals indicate underlying field variation. After evaluating the residual pattern, the project lead can choose to exclude data from further analysis or make an adjustment to a specific user-defined set of data.

The DQC tool generates output in graphical and tabular formats. Each format provides insights into the presence of variability in the field and indicates if a correction should be made. This tool prepares statistical summaries of data for each trial and for each trait. Statistical summaries include a visual representation of data distributed by location and distributed across all locations. The tool also produces a correlation table evaluating in a pairwise fashion the agreement between data collected from any two locations. A location that correlates negatively with all other locations for one or more traits may be a location from which data should be excluded from further analysis.

The decision to correct or eliminate data is not trivial. Adjusting data for all variations detected in a field environment may result in loss of information that may be useful for another purpose, such as measuring the impact of microenvironmental effects. However, if a variation is recorded in field notes, the underlying pattern of variation is not related to entry or treatment, and the experimental design is robust (e.g., with a sufficient number of replications, the entries an (or) treatments were randomized), then correcting the data is justified.

The DQC tool output reveals patterns in the data that may not be discernable in the raw data. In the example in Figure 3, the DQC tool reveals a pattern that was traced to a manufacturing defect in the harvest machinery. If this pattern had been undetected, advancement decisions made by project leads, product quality, and the reputation and revenue of the machinery manufacturer would have been significantly and negatively impacted.

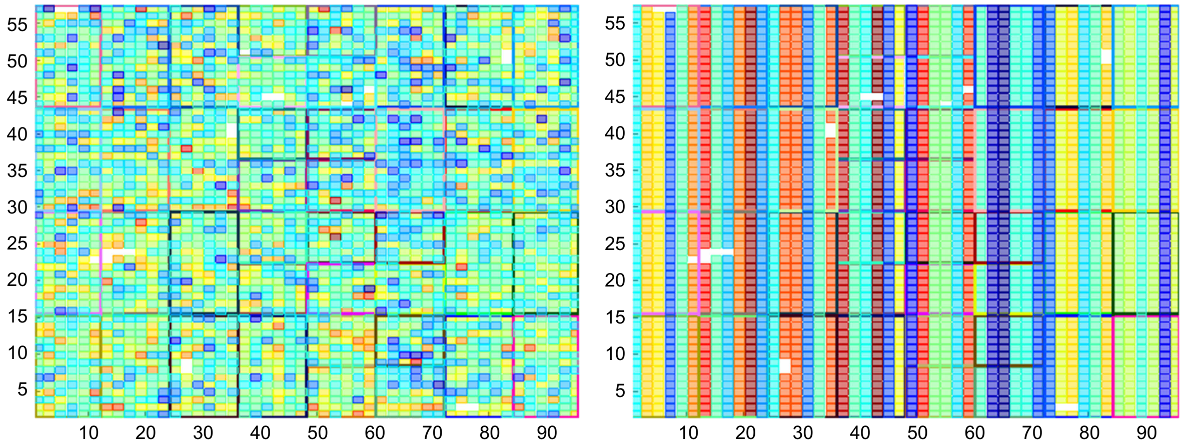

Figure 3: (Color online) This graphic illustrates DQC tool output for yield data collected from the same field of soybeans. On the left, we show the distribution of actual data values; on the right, we show the distribution of residuals, revealing the presence of systematic error. Each square represents an individual field plot and data point.

Benefits of Operations Research Tools to Soybean-Variety Development

The integrated set of tools brings myriad benefits to Syngenta Soybean R&D. The TI tool enables identification of a cost-effective TI design, ensuring that new parent varieties are developed rapidly and available in the variety-design phase of the pipeline. The BPL tool helps identify the most cost- and time-effective pipeline design, ensuring that genetic gain is maximized and the development and release of high-yielding varieties occur quickly. The YTD optimizer simulates the cost effectiveness of different three-year-long experimental designs by providing opportunities to improve pipeline effectiveness without increasing pipeline costs and to reduce pipeline costs without reducing pipeline effectiveness. The DQC tool improves the quality and value of field-trial data and decisions made based on those data by enabling systematic errors to be identified and corrected prior to analysis. When all the OR tools are used effectively, the Soybean R&D challenge of making the right decisions to design an optimal variety development pipeline is resolved.

We allocated the cost avoidance that Syngenta achieved for each tool approximately as follows: DQC (38 percent), YTD (38 percent), BPL (12 percent), and TI (12 percent). Additionally, the BPL and TI tools have significantly improved the probability of successfully delivering elite varieties to the commercial portfolio, improving the probability of success from 65 to 85 percent for the BPL tool and from 20 to 90 percent for the TI tool.

The OR tools also provide the following benefits to soybean product development:

- Increased speed and flexibility: The TI tool and YTD optimizer develop and simulate TI scenarios and experimental designs, respectively, prior to initiating any actual TI or yield trial activities. Both tools also allow the project lead to proactively evaluate the results of design changes, which may be required after initiating the TI process. Quick turnaround, reliability, and repeatability permit better planning and better resource allocation for TI projects and for the evaluation phase of all variety development projects. The BPL tool calculates the time required to complete activities within each phase of the pipeline and across the entire pipeline. It allows changes to one or more variables and evaluates the impact on time. If the options selected result in an unusually large delay, the BPL tool issues a warning and offers the project lead an opportunity to identify and change the decision that resulted in the delay.
- Understanding of the consequences of decisions: Each decision impacts cost, time, and POS. Using the TI tool, a project lead can decide between various TI process scenarios and manage trade-offs between risk, cost, and time. The BPL tool improves decision making by giving the project lead the ability to evaluate the results of altering any decision on the pipeline. It can also be used to evaluate impacts of using molecular markers and disease and (or) pest evaluations for making advancement decisions. The YTD optimizer provides immediate feedback (cost and yield) to the project lead regarding changes to the experimental design; the results of such feedback are often counterintuitive. Since Syngenta implemented the DQC tool in 2011, the coefficient of variation for yield in soybean field trials has decreased significantly, thus improving decision accuracy. Advancing a variety based on accurate data greatly increases POS and portfolio value.
- Efficient utilization of resources and increased success in the development of commercial soybean varieties: The TI tool simulates various strategies and estimates POS, cost, and time for each. Numerous simulations, each with a different set of TI process conditions, can be run within minutes. The project lead can focus on finding the optimal strategy—the strategy that provides Syngenta the highest probability of developing a new parent variety within specific cost and time constraints. Spending occurs only for TI processes that have a high POS, and costs are avoided for TI processes with low POS. The BPL tool designs a specific pipeline to achieve a specific product description regardless of cost and time. However, the optimizer feature in the BPL tool is used to identify a pipeline design that achieves the same product description using the least amount of resources in the shortest time. Pipelines for new varieties are designed once or twice a year, and for as many as six ongoing projects. Therefore, the benefits that the BPL tool provides increase across projects and across years. The BPL tool ensures that spending occurs only for pipelines that meet time and success metrics and avoids costs for pipelines that do not. The YTD optimizer indicates that the greatest impact on pipeline costs and effectiveness is the design of stage 1 trials. The model output (Pareto table) and

the neural net variable impact analysis indicate that (1) increasing the number of stage 1 varieties has the largest positive impact on the probability of identifying superior experimental varieties, and (2) Syngenta can obtain sufficient selection accuracy using four or five stage 1 trial locations (50 percent reduction). In stage 2, the Pareto analysis suggests that moderate resourcing is sufficient to provide good selection accuracy and that some designs may reduce the total number of plots by 50 percent without significant loss of selection accuracy. The second-greatest impact on pipeline costs and effectiveness is stage 3 trial resourcing. The analysis strongly supports using more than 30 locations for stage 3 trials and shows that the addition of a new location provides greater discrimination than adding a replication to an existing location. The number of plots in some stage 3 designs should be nearly doubled to maximize the effectiveness of the variety evaluation phase.

Operations Research Tools: Deployment and Training

The deployment and utilization of OR tools is mandated within the soybean product development program. Training on the tools is accomplished via face-to-face and online meetings. The tools are designed for ease of use; project leads do not need to learn a programming language to obtain the full benefits. For those with prior knowledge of breeding and the variety development process, the tools are an opportunity to (1) better understand the results of project lead decisions, (2) improve tool functionality, and (3) optimize the pipeline design. For a new project lead, the tools are a means to learn about plant breeding and the variety development process; the result is a more effective Syngenta employee.

Concluding Remarks

Advanced analytical and OR methods have transformed industries throughout the world, increasing productivity and saving billions of dollars. Until now, these revolutionary methods have not been applied to R&D programs in plant agriculture.

In support of Syngenta’s Good Growth Plan, the Soybean R&D group developed and implemented a strategy that combines advanced analytics with sound scientific, mathematical, genetic, and breeding principles to find opportunities for increasing genetic gain and optimizing the soybean seed variety development pipeline without increasing the use of land, water, or energy. The strategy resulted in the development of a suite of OR tools that increases the probability of successful development and decreases the pipeline timeline, thereby resulting in significant cost avoidance.

In addition, these OR tools increase the confidence that project leads have in the data, provide better training for breeders, and offer standardized comparisons of alternative plans; the total cost avoidance from 2012 to 2016 directly attributable to these four tools and analyses is $287 million across Syngenta seeds product development.

Syngenta recognizes the impact these tools have made within its Soybean organization. In 2014, R&D leaders initiated a project to customize and launch similar tools across Syngenta’s crop platform. This across-crop initiative is underway and scheduled for completion in 2018.

References

Food and Agriculture Organization of the United Nations (2014) World hunger falls, but 805 million still chronically undernourished. Accessed August 1, 2015, http://www.fao.org/news/story/en/item/243839/icode/.

Palisade (2014a) @Risk 6. Accessed July 1, 2015, http://www.palisade.com/risk/.

Palisade (2014b) NeuralTools6. Accessed July 1, 2015, http://www.palisade.com/neuraltools/.

Sleper D, Poehlman J (2006) *Breeding Field Crops*, 5th ed. (Wiley-Blackwell, Ames, IA).

Syngenta (2013) Syngenta launches the good growth plan. Accessed July 1, 2015, http://www.syngenta.com/global/corporate/en/news-center/news-releases/Pages/130919.aspx.

Syngenta (2015) Taking action, one planet, six commitments. Accessed July 1, 2015, http://www.syngenta.com/global/corporate/en/goodgrowthplan/commitments/Pages/commitments.aspx.

United Nations (2013) World Population Prospects: The 2012 Revision. Report, United Nations, New York.

Joseph Byrum leads soybean breeding, trait research, and product development at Syngenta, helping deliver improved genetics and trait technologies for soybean growers. Dr. Byrum has extensive global experience in plant genetics and biotechnology being actively involved in soybean research for 20 years. Since joining Syngenta in 2006, he has held a variety of leadership roles. Dr. Byrum obtained his MBA from the Ross School of Business at the University of Michigan. He holds Bachelor of Science and Master of Science degrees from Michigan State University. He earned his doctorate in genetics from Iowa State University. Additionally, he is a fellow at the Aspen Institute. Most recently, he became an INFORMS Edelman Fellow and was the chief architect of "Good Growth through Advanced Analytics," a competition presentation that won the 2015 INFORMS Edelman Award, which encompasses advanced mathematics and the management sciences.

Craig Davis has led the Soybean effort to increase the rate of genetic gain in Syngenta's breeding program since 2010. Breeding cycle times decreased by 40 percent while increasing development and testing of experimental lines in both North and South America. Beginning in 2012, Craig has served as a senior technical advisor to Syngenta's development and deployment of analytical tools that improve program efficiency and effectiveness and enable breeders to improve decisions by basing them on advanced mathematics and modeling. His current efforts focus on improving breeder parent line selection and choice of combinations between parents by utilizing the entire suite of analytical and statistical tools available to breeders.

Greg Doonan brings broad knowledge of genetics and product development experience to the role of genetic project lead.

For the last five years, Greg held roles in Soybean Trait Introgression and Development. From 2009 to 2012, he led the soybean trait introgression program in Clinton, Illinois. There he oversaw the successful completion and outfitting of Syngenta's state-of-the-art growth room facility. Recently he has worked with the business, crop head, and product development leads to develop and implement soybean's global strategy. His primary focus was executing on Soybean's trait platform strategies.

He holds a master's degree in plant breeding from Iowa State University.

Tracy Doubler is head of North American Soybean Breeding Projects at Syngenta Seeds based in Slater, Iowa. He has held several positions including genotyping lab manager, soy genetic information manager, and global head of soy genetic projects before transitioning to his current role in January 2013. Tracy received his Project Management Professional (PMP) certification in 2010. He earned his BS and MS degrees from Southern Illinois University at Carbondale and is pursuing his PhD in agronomy (plant breeding) at Iowa State University.

David Foster works in Syngenta's Biological Data Analytics group as crops DA manager, leading a team of R&D specialists in supporting crop (trialing) data collection, analysis, reporting, visualizing, and archiving. He recently co-developed an interactive data visualization tool that is becoming widely used within Syngenta by field scientists for the removal of outliers and the correction of systematic error in trialling data.

Bruce Luzzi brings considerable experience and expertise in the genetic improvement of soybean, canola, and rice as soybean seeds project lead. Since joining Syngenta, Bruce has taken a lead role in the development and adaptation of analytical tools to improve the efficiency and effectiveness of the soybean product development pipeline. In addition, Bruce is the manager of multiple project teams responsible for the identification of favorable soybean characteristics that can be used in the soybean product development pipeline.

Bruce earned a PhD in agronomy from the University of Georgia.

Ronald Mowers is Syngenta North and South America statistics lead with responsibility for statistical consulting, statistics methods research, and management of the Syngenta Americas Biometrics team. His areas of expertise include statistics and quantitative genetics, improvement of genetic gain (e.g., corn strategy algorithms in 2014), genotype $*$ environment methods and most recently simulation studies of data aggregation methods for variety comparison testing.

Ron holds a PhD in statistics and agronomy from Iowa State University, MS in plant and soil science from Southern Illinois University, and MS and BS in mathematics from the University of Illinois.

Chris Zinselmeier is a trait project lead for the Global Soybean Research and Development Team. He is responsible for leading key trait projects important to the Syngenta Soybean Team and bringing winning innovation to the market. He earned a PhD in agronomy and physiology at the University of Minnesota.

Jack Kloeber is a retired U.S. Army lieutenant colonel. During his military career Jack taught mathematics at West Point, and graduate level decision analysis, systems simulation, and technology selection at the Air Force Institute of Technology. After retiring from the U.S. Army, he became head of portfolio management for Bristol-Myers Squibb and later at J&J Pharma Services His work has supported decisions for various DoE superfund sites, many technology organizations within the DoD, and pharmaceutical and agriculture R&D organizations. Jack joined KROMITE LLC as partner in 2007 and has been the Principal since 2012.

He received his PhD in economic decision analysis from Georgia Institute of Technology and a master's degree in industrial engineering from Lehigh University. He is a Founding Fellow and current president of the Society of Decision Professionals and is a member of the Decision Analysis Society and INFORMS.

Dave Culhane has 20 years of experience in the agribusiness, pharmaceutical, nutritionals, and consumer products industries. He joined KROMITE LLC in 2011 and is now the lead consultant in pharmaceutical decision analysis. He has a BS in chemical engineering and an MBA in finance from Lehigh University. He is a member of the Society of Decision Professionals.

Stephen Mack is a decision management consultant with over 25 years of experience. He specializes in integrating both quantitative and qualitative decision support methodologies as part of a holistic decision management process. Steve has an MS in OR from the George Washington University and a BS in chemistry from Drexel University.