

# Using Data Intensive Plant Breeding and Genetic Gain as an Indicator for Agricultural Research and Development Will Speed Up Agriculture

Maria Martigoni\*

Department of Agricultural Research, University of Toronto, Toronto, Ontario, Canada

## Abstract

Growing on the availability of new data technologies, accelerating the pace of genetic gain has recently emerged as a fundamental goal in plant breeding for the global South, linking biological interest in crop development with economic interest in increasing the cost effectiveness of breeding programmes. The concept of genetic gain, the circumstances surrounding its emergence as an indicator of agricultural development and the broader implications of this development are all explained in this paper. A special focus is placed on how plant breeding's knowledge control regimes are changing, the social and political repercussions for smallholder farmers and climate adaptive agriculture. We examine how the relationship between development objectives and practise is impacted by the order in which the variables used to construct the indicator are prioritized when choosing agricultural policies. We come to the conclusion that, in the absence of data on other critical areas (such as agro biodiversity, seed systems and the varied impacts of climate change on soil, crops and communities), as well as tools to assess the benefits and drawbacks of the acceleration in seed selection, management and evaluation fostered by the adoption of genetic gain as a key indicator, genetic gain should not be taken into consideration as a primary indicator of agricultural development.

**Keywords:** Indicators; Quantitative genetics; Plant breeding; Agriculture; Climate change; Development; Molecular basis of plant breeding; Favorable allele; Fiber length; Gene based breeding

## Introduction

At practical, social and epistemological dimensions as well as technological ones, the genomics revolution in the biosciences has resulted in significant changes in the organization of biological research and the production of biotechnologies [1]. A lot of focus has been placed on the advancement and possible effects of transgenic technologies and genome editing methods like CRISPR for food and agriculture and plant science and its applied fields are no exception. These technologies, which are frequently referred to as "new breeding techniques," are frequently examined in terms of their novelty, disruptive potential and risk. At practical, social, epistemological and technological levels, the genomics revolution in the biosciences has resulted in significant changes in the way biological research is organized and how biotechnologies are developed. The development and potential effects of transgenic technologies and genome editing methods like CRISPR for food and agriculture have received considerable interest and plant science and its applied fields are no exception. The novelty, disruptive potential and risk of these technologies, which are commonly referred to as "new breeding techniques," are extensively examined [2]. Limited resources are a major barrier to the widespread adoption of cutting edge technologies like CRISPR for international agricultural research and breeding networks focusing on the global South; comparable concerns also apply to regions of the global North that are unable to reproduce the conditions needed for intensively managed crops or are remote from large scale processing infrastructure [3]. Agricultural research networks are being rebuilt in ways that integrate earlier statistics and more current data intensive breeding procedures, with extensive consequences for scientific research, breeding practise and agricultural systems. This process is less obvious but has a far larger influence. In this study, we examine the significance of the rate of genetic gain, a marker that is increasingly important for predicting the shape and course of such reorganizations.

The term "genetic gain" was originally used in the context of animal breeding at the beginning of the 20<sup>th</sup> century as a statistical measure of the genetic improvement of breeding populations [4]. Given the renewed interest in the opportunities provided by quantitative genetics and the accessibility of increasingly low cost and simple means for gathering, distributing and analyzing genomic data, genetic gain has taken on new significance for plant breeding. Breeders, researchers and funders from all over the world are actively promoting it as a crucial performance indicator for plant breeding, along with an active commitment to "accelerate" rates of genetic gain as a crucial policy objective [5].

The Consultative Group for International Agricultural Research (CGIAR), arguably the largest and most influential research network for agricultural research in the world and the Bill and Melinda Gates Foundation (BMGF), a current major funder of the CGIAR and of international research focused on development more generally are leading this commitment in plant breeding for the public domain and the global South [6]. Advanced commercial plant breeding programmes have well established goals that are similar to these ones. To concentrate on the specific adjustments to public plant breeding in

\*Corresponding author: Maria Martigoni, Department of Agricultural Research, University of Toronto, Toronto, Ontario, Canada; E-mail: martigonim@mun.ca

**Received:** 09-December-2022, Manuscript No. JPGB-22-82967; **Editor assigned:** 12-December-2022, PreQC No. JPGB-22-82967 (PQ); **Reviewed:** 26-December-2022, QC No. JPGB-22-82967; **Revised:** 31-March-2023, Manuscript No. JPGB-22-82967 (R); **Published:** 07-April-2023, DOI: 10.4172/jpgb.1000140

**Citation:** Martigoni M (2023) Using Data Intensive Plant Breeding and Genetic Gain as an Indicator for Agricultural Research and Development Will Speed Up Agriculture. J Plant Genet Breed 7: 140

**Copyright:** © 2023 Martigoni M. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution and reproduction in any medium, provided the original author and source are credited.

the international context, where resources are scarce and the structure of objectives is expressly oriented towards a wider range of development goals beyond commercial growth, we leave the latter aside in this work.

## Literature Review

Here, we use a process based model to illustrate the systematic characterization process for the development of TPEs. Engage breeders, obtain input data and carry out modelling simulations.

### Define the target area

Within a certain TPE, the analysis should ideally encompass at least 80% of the target crop's present growing region [7]. The best data to use is recent time series production data, which is frequently supplied by agricultural ministries, organizations, or institutions that provide web based tools like the District Level Database (DLD). Another type of data that can be used to specify target areas is satellite photography data. As these areas can be included in future TPEs, it is crucial to take into account places where there is a rising tendency in the area under cultivation for a particular crop. Crop geography over time can be examined using remote sensing data, census information, survey results, local expert interviews, and combinations of these data sources [8]. Without referring to specific cropping areas, TPEs can be defined geographically in the absence of data, such as within national or state boundaries.

### Get genotype parameters

Genotype parameters are coefficients that form the basis of mathematical equations that depict fundamental biological and plant production processes. A large portion of this is easily accessible in the literature or was taken from related mechanistic models. This information is a useful place to start if the value of a parameter has been determined in previously published research [9]. Parameters must be estimated using defined techniques if they are not already known and then converted into model coefficients. Since many of these factors are fixed in practise, just a few tens of parameters accurately reflect the genetic variability. Genotype parameters from unsuitable models should be avoided since they require a "calibration" procedure to better fit the model prediction to the observations.

If genotype parameters are not already known, they must be obtained through field trials with crops produced without nutritional restrictions or biotic challenges, together with weather, soil and management information. In order to ensure that genotype parameters are generated for genotypes that breeders attempt to enhance, such as well-known cultivars, discussion with breeders is crucial at that time. For parameterization, it is necessary to have at least records of various phenology stages, time series of aboveground biomass, leaf area index across time and grain yield. This can be used to estimate parameters for phenology prediction, leaf area development, biomass accumulation, harvest index and yield formation.

### Parameter

**Phenology phases:** Critical phenological stages must be predicted in all crop models using Thermal Time (TT) or Biological Days (BD). Based on cardinal temperatures, the daily average temperature and the time until the start of each phenological stage, these can be approximated. The phenological stages may vary depending on the

crop and the model. Recording the day to flowering and the day to maturity from sowing is particularly significant, as these two measurements can be used to define the TT or BD requirements for various phenological stages [10].

**Reaction to Photoperiod (PP):** Some genotypes or crops are sensitive to PP for a portion of their cycle. Some parameters with photoperiod sensitive genotypes need to be estimated. Critical photoperiods are frequently rather stable between cultivars of a species. Models employ a function that modifies the occurrence of phenology phases [11]. Iterative optimization techniques can be used to obtain the parameter for this function, Photoperiod sensitivity slope (PPsen), from growth chamber and/or field experiments (programs exist). Stages that are PP sensitive can be discovered in the literature or discovered through research that uses reciprocal transfers between photoperiod treatments.

**Leaf area development:** Using ceptometers or experimental observations of the crops, one can collect all leaf area development characteristics directly. By fitting a linear regression line to the main stem leaf/node number versus the temperature unit, the phyllochron can be determined. By fitting a power equation to the relevant data, *i.e.* plant leaf area vs. main-stem leaf number, one can determine the allometric connection coefficients between plant leaf area and main stem leaf number [12]. It is also possible to determine their relationship to plant density if different plant density data are available. A power equation is fitted to the data of each plant density in order to accomplish this. It is therefore possible to determine how the obtained coefficients and plant density are related. A typical exponential regression function explains LAI expansion as a function of normalized temperature units in various models, such as EPIC and SWAT. In this instance, growth circumstances and plant density may have an impact on the highest predicted LAI (LAIMX) value. A specific leaf area is estimated using the slope of an LAI plot versus leaf dry weight. Given the fragility of early stage leaves, data obtained at higher LAIs should be used to prevent overestimating SLA.

**Biomass accumulation:** Depending on the model chosen, several techniques and parameters are employed to forecast daily dry matter production; nevertheless, the majority of parameters is constant or is available in the literature. For instance, additional parameters may be required for models like ORYZA2000, CROPGRO and hybrid maize that separately simulate gross photosynthesis and respiration. The majority of CSMs also employ temperature adjusted Radiation Usage Efficiency (RUE) to transform captured sunlight from photosynthetically active organisms into dry matter. Dry matter production by a crop canopy must be quantified along with concurrent measurements of PAR interception by the canopy in order to estimate RUE experimentally. RUE can be constant across genotypes of a species and is discussed in works by Soltani and Sinclair, among others. The extinction coefficient is a further variable employed in the model's estimation of biomass production (KPAR). KPAR is a composite property that combines plant and canopy properties. It can be measured experimentally using crop LAI and PAR interception [13].

**Yield development:** In crop models, the rate of Harvest Index (HI) rise, the maximum predicted HI and the proportion of remobilizable dry mass are frequently used as three parameters to estimate final yield. The most important one to estimate is maximal HI, which can be acquired directly from experimental crop observations. As a gauge of reproductive effectiveness, the ratio of grain to total shoot dry matter rise can also be used to assess the rate of HI increase.

## Get environmental data

The environment, which includes the soil and climate, is the primary cause of yield variability according to studies on plant breeding. Such information ought to be gathered for the concerned TPE. There is a lack of high quality climate data and a high degree of soil variability in many agricultural areas. The development of global or continent specific databases of observed and generated datasets has recently advanced. These databases, many of which are freely accessible online, might allow for high resolution simulations. The choice of weather data can affect estimates of crop yield responses to climate variability and change. When possible, it is best to always validate the relevance of synthetic data sources against observed data when using them. This is especially crucial in areas with a lot of topographic variation. CSM can create a virtual plant and act as a quality indicator for such data by allowing researchers to compare the model's various outputs when using various sources of Measured (MWD) and gridded Weather Data. The correlation coefficient ( $r$ ), absolute Mean Error (ME), Root Mean Square Error (RMSE), normalised RMSE (RMSE n%) and coefficient of determination ( $r^2$ ) according to linear regression are typical measures used to evaluate agreement and biases between GWD and MWD for a given weather or crop parameter.

The considerable regional variation makes it more difficult to collect soil data. To simplify soil profiles based on three factors soil texture, root depth and organic carbon some soil databases, such as HC27 of IFPRI, have been created. Crop models respond best to these criteria. The HC27 generic soil map, which divides soils into 27 profiles, is the result of a partnership between FAO, IIASA, ISRIC, ISSCAS and JRC. To evaluate the quality of the HC27 soil information, a straightforward CSM was used to compare the model's output with actual soil data. It was determined from the comparison that HC27 can be used to simulate potential yield and water related factors because the results were statistically similar to observed soil data. The knowledge on typical soil types and efficient rooting depths for the target crop should be acquired through surveys and consultations with local experts if gridded or generic soil data is not accessible or reliable enough for a region.

## Get management data

Process based models require data on agricultural management techniques, including weed control, plant density, irrigation and fertilizer use. The model determines the level of detail in the management information. For instance, not all are intended to model weed management and competition. Such management knowledge can be acquired by closely working with regional breeders and agronomists or it can be discovered on international databases, platforms or scientific publications. The FAO crop calendar tool, which includes 130 crops and a global data set of monthly irrigated and rainfed growing areas, which includes 26 irrigated and rainfed crops, have both been developed as platforms to support assessments at the global level. However, management information other than the sowing date is not provided. While Sacks, et al., also offer compiled datasets of global agricultural planting and harvesting dates for 19 important crops, local crop management information is gathered *via* alternative platforms like the GYGA data sheet. There is a need for clarification for the predominant water regimes (*i.e.* rainfed, partially or fully irrigated) and the percentage of water availability, which may be found in FAO AQUASTAT, official country level irrigation statistics gathered by National Agricultural Research Centres (NARC)

(*e.g.* area under irrigation). You can also get recommended fertilizer rates from the NARC, seed firms, agriculture ministries or breeders' advice for variety management. To restate, since they frequently have direct contact with the many local agencies, close communication with breeding groups is essential in this situation.

## Evaluate the model

It is essential to thoroughly compare the model predictions (phenological stage, grain yield, aboveground dry matter, and crop evapotranspiration) with high quality data from diverse trials conducted in a range of settings and management techniques. Such similarities can be quantified using statistical methods. However, a realistic depiction of the temporal dynamics of the crop growth phases is just as crucial as a prediction of the final yield that is deemed to be acceptable. The crop model's sensitivity or its ability to accurately forecast crop performance under a variety of conditions is another crucial factor. For instance, it was discovered that one of the major factors affecting groundnut yield in India is water. The model was able to sensibly predict a fairly wide range of observed yields as a result. More precise evaluation is required when it comes to extreme events like frost or heat shocks. While the majority of models show a similar trend in simulated crop yields as temperatures rise, some models do not directly account for the impacts of heat stress, which could result in significant yield heterogeneity. Far from all CSM having been tested for all potential pressures or all combinations of stresses.

The relevance of model outcomes, on the other hand, is determined by the calibre of the observed data and the statistical analysis of robustness evaluations aids in describing the anticipated consistency of model assessment results. Crop models offer a special framework to capture consequences in untested situations, such as anticipated climates, when they are appropriately created, deployed and tested.

## Generate model runs across a target area

In order to produce an accurate estimate of growth and development characteristics as well as yield at a certain area once the model has been validated, crop model simulations must be run for a predetermined number of seasons. 10 years-30 years of daily meteorological data are necessary for reliable calculations of yield potential (irrigated) or yield potential (water limited, rainfed) and their variability; the more data, the better when year to year variability in water availability is large. The needed number of years is shorter than in unfavourable situations when there are fewer variables present, such as in favourable rainfed or irrigated areas. Statistical qualities equivalent to real data required at least 15 years of simulated weather data, according to Soltani and Hoogenboom's findings from 2003. Sinclair, et al., citing a number of studies, noted that it might take up to 30 seasons to fully account for weather variability's influence on simulation outcomes at each location.

## Discussion

While discussing the importance of indicators of (in her case, socioeconomic) development, Morgan asserts that these data "should not be regarded measurements of development" even though they undoubtedly contain information indicating aspects of growth. Genetic gain is one of a larger group of indicators that play a crucial role in the knowledge control regimes of agricultural research and development as a measure of agronomic and breeder performance. As an illustration, consider the metrics that make up goal 2 of the UN

sustainable development goals, zero hunger. Indicators more closely related to the worries and goals surrounding seed systems stated in the preceding section are also included. For instance, calculating the "weighted average age of varieties" sown across a specific geographic area is a crucial step in assessing varietal turnover. A question that needs further investigation is how such indicators are applied in practise to guide development interventions and how they are related to genetic gain data (since the two are closely related in the imperatives for "accelerated" breeding that we have discussed throughout the paper). In fact, it is a question that deserves special attention because the answer could have a big impact on how "the field" the object of intervention transforms from being a set of environmental variables linked to agronomic performance (similar to in-field analyses of genetic gain) to a physical setting populated by farmers and other stakeholders whose practises and decisions are being assessed and governed with regard to crop selection and cultivation.

What is evident is that, contrary to Morgan's suggestion, the overall picture painted by these sets of indicators does not include a measure of development. This has to do with the way that indicators are epistemic objects in and of themselves. As abstract figures derived from various data series, they can be put together in a variety of ways, and as we showed in the cases of genetic gain and climate adaptation, the particular ways in which they are put together are crucial to understanding their significance and utility. In addition, certain sets of concerns are constantly expressed in the formulation and selection of indicators. The so called green revolution, which has deep historical roots and has fully realized a particular vision of agricultural development, is linked to the special importance given to breeding speed and the commercialization of seed systems. Biodiversity is frequently seen in this way as a storehouse of crop varieties whose agronomic merits can be investigated, enhanced and ultimately approved through research and related commercialization. Fenzi and Bonneuil refer to this model as "resourcist" to emphasize how important it is to view plants as genetic "resources" from which value should be extracted. For instance, an additional proposal has been made that breeders' source material should be limited to elite germplasm, *i.e.* that from high performing (and frequently commercial) varieties, in addition to the methodological adjustments indicated in section 3. As a result, breeders can avoid engaging in time consuming backcrossing procedures with non-elite varieties like farmers' landraces, which immediately increases genetic gain. The wide range of values of crop biodiversity, including not only for breeding itself but also for health and wellbeing, cultural identity, ecology and sustainability, cannot be taken into account by calculations of genetic variance, no matter how advantageous they may be. Additionally, this suggestion adds new restrictions to the laws governing intellectual property that affect breeders. The range of material accessible for public breeding is actually just those elite varieties made available as public goods by CGIAR or other public breeders because the majority of elite varieties come from commercial breeding projects and are thus protected by intellectual property rights. It's possible that those made public by the CGIAR will also be governed by its own intellectual property policies.

## Conclusion

It does not lessen the value of genetic gain calculations as a practical element of plant breeding to point out their limitations and the knowledge control regime in which they are used. When

attempting to quantify and compare the results of breeding models and particular programmes, proponents have argued that determining the rate of genetic gain can be very helpful, especially when it comes to improving quantitative genetic traits and adjusting for the effects of genotype environment interactions. However, it is important to consider critically how these calculations are used to support agricultural growth and to lessen the effects of climate change on agriculture. Considering the significance of creating complementary indicators that take a variety of viewpoints on what constitutes agricultural development into account, it is important to debate whether genetic gain should be used as the primary performance indicator for plant breeding. In other words, we contend that an improved framing of the knowledge control regimes related to data intensive agriculture is required in order to take into account both the value of genetic gain as an indicator and its limitations, as shown above.

It is not within the scope of this paper to provide a thorough analysis of what such an enriched framing should entail; however, we have shown through our long standing collaboration with academics and institutions from a variety of fields involved in agricultural development that such elaboration requires transdisciplinary cooperation across many stakeholders in the agricultural system. We can offer some recommendations for future agricultural knowledge control regimes in the form of our final point. We support the adoption of metrics and infrastructures that are specifically geared towards agrobiodiversity and agroecological conservation, the diversification of seed systems and mitigating the differential impact that climate change is having on soil, crops and communities, in addition to the technologies supporting primary breeding indicators like genetic gain. Additionally, a dedication to gathering, preserving and disseminating data such as metadata and documentation on data standardization processes will help stakeholders break down indicators like genetic gain and comprehend their constituent parts and the various implications they pose, enhancing their capacity to critically engage with performance indicators for rapid breeding. The current investment in data intensive agriculture is affected in a real way by this. It calls for funding to: Strengthen the governance of data infrastructures through participatory engagement that facilitates farmers' input into the production and use of related resources and resulting indicators; identify and consider development goals beyond high yield agriculture, including the valorization of agrobiodiversity for food security and make the underlying data and models accessible and scrutinizable, which helps to disaggregate indicators and verify their provenance.

## References

1. Ababaei B, Chenu K (2019) Recent trends in drought, heat and frost induced yield losses across the Australian wheat belt. *Multidiscip Digit Pub Proc* 36: 5.
2. Barlow KM, Christy BP, O'leary GJ, Riffkin PA, Nuttall JG (2015) Simulating the impact of extreme heat and frost events on wheat crop production: A review. *F Crop Res* 171: 109-119.
3. Battisti R, Sentelhas PC (2019) Characterizing Brazilian soybean growing regions by water deficit patterns. *F Crop Res* 240: 95-105. Boote KJ, Jones JW, Hoogenboom G (2021) Incorporating realistic trait physiology into crop growth models to support genetic improvement. *In Silico Plant* 3.
5. Brown D, Bergh IVD, Bruin SD, Machida L, Etten JV (2020) Data synthesis for crop variety evaluation: A review. *Agron Sustain Dev* 40: 25.

- 
6. Ceccarelli S (1989) Wide adaptation: How wide? *Euphytica* 40: 197-205.
  7. Ceccarelli S, Grando S, Baum M (2007) Participatory plant breeding in water limited environments. *Exp Agric* 43: 411-435.
  8. Ceccarelli S, Grando S, Maatougui M, Michael M, Slash M, et al. (2010) Plant breeding and climate changes. *J Agric Sci* 148: 627-637.
  9. Chapman SC, Hammer GL, Butler DG, Cooper M (2000) Genotype by environment interactions affecting grain sorghum. III. Temporal sequences and spatial patterns in the target population of environments. *Aust J Agric Res* 51: 223.
  10. Chu X, Ilyas IF, Krishnan S, Wang J (2016) Data cleaning: Overview and emerging challenges. *Proc Int Conf Manag Data* 2201-2206.
  11. Haas D (2015) Wisteria: Nurturing scalable data cleaning infrastructure. *Proc VLDB Endow* 8: 2004-2007.
  12. Hodge V (2004) A survey of outlier detection methodologies. *Artif Intell Rev* 22: 85-126.
  13. Hurley M (2016) Credit scoring in the era of big data. *Yale JL Tech* 18: 148.