Cheat sheet on how to calculate *realized* genetic gains

CGIAR | Excellence in Breeding Platform

Optimizing breeding schemes

In this document we provide recommendations on how to set experiments, organize the data and analyze it in order to provide accurate estimates of realized genetic gain.

Contact: **Giovanny Covarrubias-Pazaran**
g.covarrubias@cgiar.org

Version: 05/16/2022

There are three important aspects to be considered when calculating the realized genetic gain:

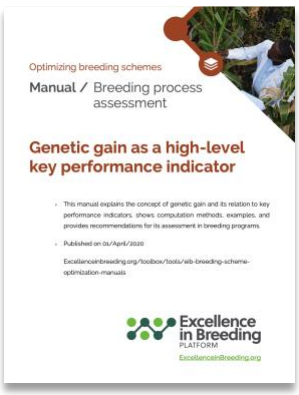
1. The experimental design of the trials (e.g., number of locations, connectivity strategy, etc.).
2. The extraction and organization of the data to be analyzed (stage, pipelines, etc.).
3. The statistical analysis of the data (traits, model, quality control, etc.).

In this document, it is assumed that the desired future state is a program that has an experimental design that allows for temporal and spatial adjustments (e.g., with a solid check replacement strategy and TPE coverage), where pipelines serve concrete market segments (target) so that data can be easily retrieved by pipeline, where data is analyzed in a way so that connectivity allows to account for year (temporal) effects, and where genotype means are accurately adjusted for all other nuisance factors (e.g., spatial effects).

It is also assumed that the calculation of realized genetic gain uses trial data generated by the regular stage-gate process and no additional trials are run (i.e., era trials) but instead, analysis of historical data is the standard.

The future and current state may differ. For example, a program may not be following the recommendations in this document yet (e.g., not using a check strategy, the proper design, etc.) but can use this as reference to know where to go in the future and how to analyze the current data available.

For a more thorough review of the topics discussed in this document, see the EiB manual on *Genetic gains as a key performance indicator*.



Annex 1. Recommendations for a proper experimental design

A regular breeding program follows a stage-gate approach when it comes to testing for population improvement (early stages) and advancement of products (late stages) (**Figure 1**). Below, we present a protocol on how to maximize connectivity in a trial run for a particular stage to maximize connectivity.

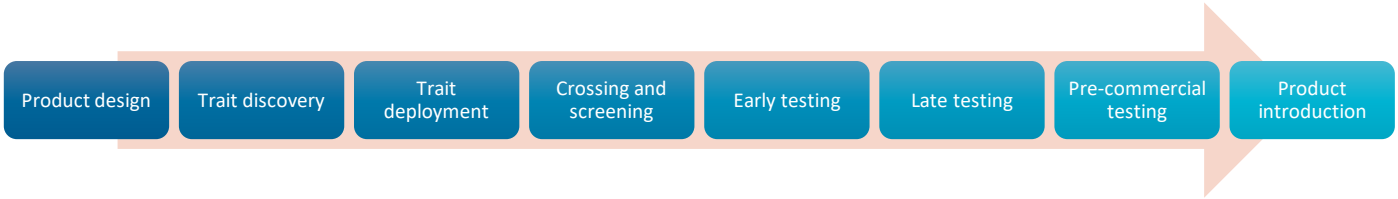


Figure 1. Graphical representation of an example stage-gate approach used in a plant breeding program.

1. Generate the list of materials (entry list) to be tested in each stage and extracted from the database (inventory). The entry list should include:
    a. The list of *new materials* (current cohorts) to be tested.
    b. At least two *temporal checks* that are as much "dynamically stable" as possible (genotypes to be present in as many years as possible years, and replicated as any other tested entry).
    c. At least (but not limited to) three to four *spatial checks* to accounting for no more than 5% of the check plots spread in the field (to be renewed at the rate of ~ one per year in a stair-like pattern to strengthen the connectivity).
    d. The temporal and spatial checks can act as commercial, local or any other contrast purpose (e.g., performance) but should respect the replacement guidelines in points b and c.
2. Generate the field books for each testing stage making sure you do the following:
    a. Move the entry list including the new materials and all types of checks mentioned in point 1 to the field book.
    b. Generate the experimental design required for the testing stage ensuring that temporal checks have at least 2 replicates and spatial checks replicated as needed without exceeding the 5% of plots in the trial. [We recommend augmented and p-rep designs for early and intermediate generation variety trials and alpha designs for late-stage trials].
    c. Try to cover at least 6 locations of the target population of environments (TPE) every year by spreading your seed packets using sparse testing methodologies if the seed supply is limited.
3. Run the trials and store the phenotypic data collected in the database for posterior analysis.

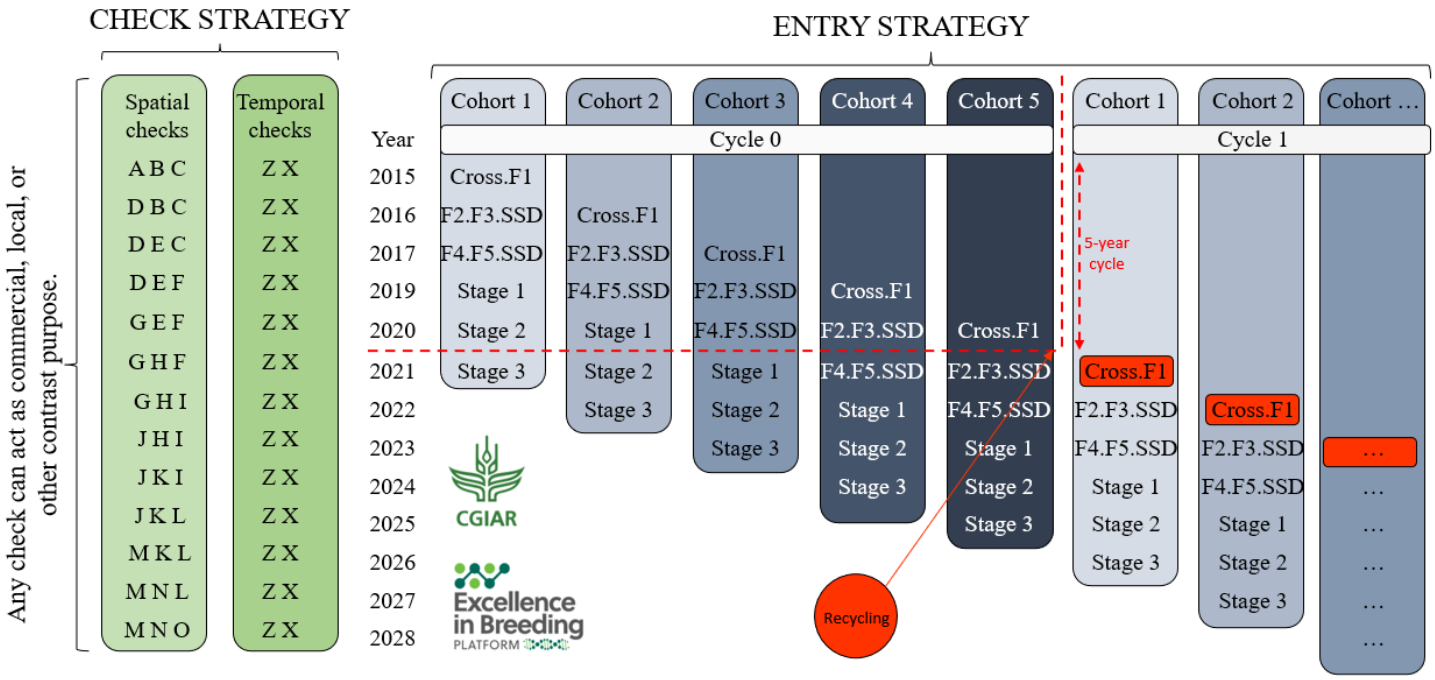


Figure 2. Graphical representation of an optimal check and entry strategy to maximize temporal and spatial connectivity among breeding trials and entries across years. On the right side, an example of a line crop with a five-year recycling time program (recycling at stage 2) is shown which leads to the existence of ~ five cohorts (columns) across multiple years (rows), and stages within a year. On the left, the proposed check strategy is displayed, where *spatial checks* normally used as benchmark varieties for product replacement purposes are proposed to be retired in a stair-like pattern (e.g., renew one at the rate of one per year) but not expanded beyond the requirements of the program, *temporal checks* (present in all years) are used to connect the data and estimate properly the year effect.

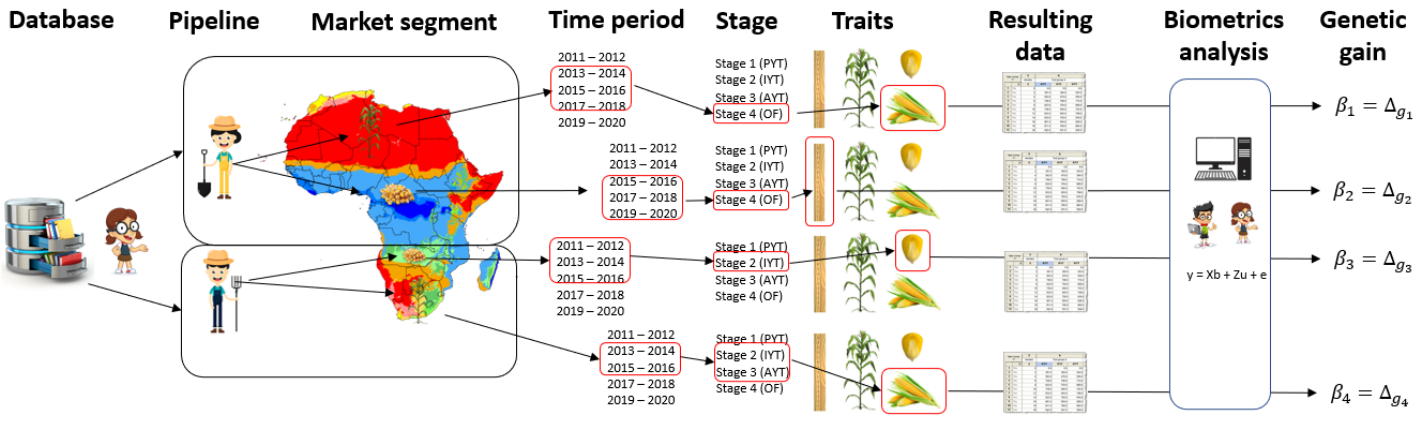


Figure 3. Graphical representation of major steps to calculate realized genetic gain using historical data. It is critical to split the data in a meaningful way to interpret properly the resulting value of genetic gain.

Annex 2. Recommendations for a proper organization and extraction of the data

The data produced for genetic gain calculation can come from the regular stage-gate approach or trials explicitly run for this purpose (e.g., era trials). Below, we provide some recommendations for users to know how to structure and deal with the data before moving to the analysis.

1. Data Generation and structure

Analysis of historical data

1. Approach your database manager.
2. Gather data records from the pipeline of interest.
3. Split the data by market segment targeted (defined by region, product features, agroecological region and production system).
4. Keep the split data for the time period of interest (e.g., last 5-10 years).
5. Keep the split data for a given germplasm stage(s) of interest (e.g., late-stage testing).[1]
6. Keep the split data only for the trait(s) that are being improved according to the product profile (e.g., yield).
7. Hand the batches of data to your Biometrics support and request a realized genetic gain analysis.
8. Request support from your Biometrician to fit a one-stage or two-stage multi environment trial (MET) analysis with a mixed model to account for nuisance (spatial, location, year), and genotype (genotype, genotype interactions) effects. Extract adjusted means for all genotypes. Merge the year of origin of the genotypes to their adjusted means and fit a linear model of the form adjusted.mean~year.origin.

Analysis of era trial data

1. Go to your germplasm bank manager.
2. Gather the list of materials for the pipeline you are interested in.
3. Split the list of materials by market segment targeted (defined by region, product features, agroecological region and production system).
4. Keep the list of materials for the time period of interest (e.g., last 5-10 years).
5. Keep the list of materials for a given germplasm stage(s) of interest (e.g., late-stage testing).[1]
6. Decide the traits to phenotype in the trials and if will be considered by separate or in an index (if weights are available).
7. Hand the list of materials to your Biometrics support to run an experimental design.
8. Plant the experiment and harvest the phenotypes and do points 7 and 8 of HD.

[1] From product development perspective use late-stage trials (e.g., Stage3 & On-Farm) to estimate genetic gain. From a population improvement perspective use early-stage trials (e.g., Stage 1 & Stage 2)

Annex 3. Recommendations for phenotypic analysis to calculate genetic gain

Even when trial data is produced and stored properly, proper statistical analysis is required to ensure an accurate estimate of realized genetic gain and trend. Below, we propose the following steps to perform such statistical analysis.

1. Retrieve the trial data as described in **Annex 2**. That is, extract the data for a specific pipeline targeting one or more market segments, and subset the data for specific targets, stages of testing (preferably Stage 3 and/or Stage 4) and time period (minimum 5 years). Decide the trait to focus on for the analysis.
2. Perform a single year single location analysis to remove outliers and typos in the data and identify trials with low H^2. Remove experiments with entry-mean H^2 lower than 0.2. Cullis et al. (2006) H^2 is a good option.
3. With the data cleaned perform a one-stage or a two-stage analysis to fit all genetic and nuisance terms in order to obtain across-year adjusted means for all genotypes. In ASReml-R language the model has the form:

```
model <- asreml(data= yourData,

fixed = Trait ~ YearAsFactor+ Genotype
            + Genotype:YearAsFactor,

random =    ~ Location:YearAsFactor +
            + Genotype:Location +
            Genotype:YearAsFactor:Loca
            tion + ExpDesign[2],

residuals=~dsum(~units|Year:Location),

)
```

   Genotypes should be fit as fixed to properly calculate the genetic value. Year (as factor) will account for the non-genetic trend. The random terms will ensure that interactions are accounted for and that across-years-locations predictions are well adjusted.

4. Obtain across-year adjusted means for all genotypes in the historical dataset. In ASReml language that is:

```
predictions <- predict(model,
classify= "Genotype")$pvals
```

5. Merge the adjusted means (predictions) with the year of origin and fit a model to calculate the genetic gain. In ASReml language that is:

```
finalModel <- asreml(data=
predictions,

fixed = predicted.value ~
          yearOfOrigin      )
```

6. The slope from the final model can be interpreted as the rate of response to selection or genetic gain.

*Using pedigree/markers can increase the connectivity of the data but will underestimate genetic gain and the de-regression method is needed. Sample scripts can be found at: gitlab.com/excellenceinbreeding/module2.

[2] From Alternatively, the experimental design noise can be taken into account with a two-stage analysis.