

Yield gap decomposition: Theory & practice

João Vasco Silva (PhD)

j.silva@cgiar.org

Agronomy-at-scale Data Scientist

CIMMYT-Zimbabwe



Addis Ababa, May 24th 2023

<https://jvasco323.github.io/eia-yg-training>

Yield gap analysis

Search docs

Installing R

Concepts and definitions

Data collection tool

Stochastic frontier analysis

Workflow for Silva et al. (2017)

Introduction

Load required R packages

Farmer field data

Data manipulation

Descriptive statistics

Efficiency yield gap

Resource yield gap

Technology yield gap

Yield gap decomposition

Recommendations

Acknowledgments

Retrieve data from GYGA

Boundary line analysis

Random forest and Shapely values

Docs » Stochastic frontier analysis » Workflow for Silva et al. (2017) R code

Workflow for Silva et al. (2017)

João Vasco Silva, CIMMYT-Zimbabwe

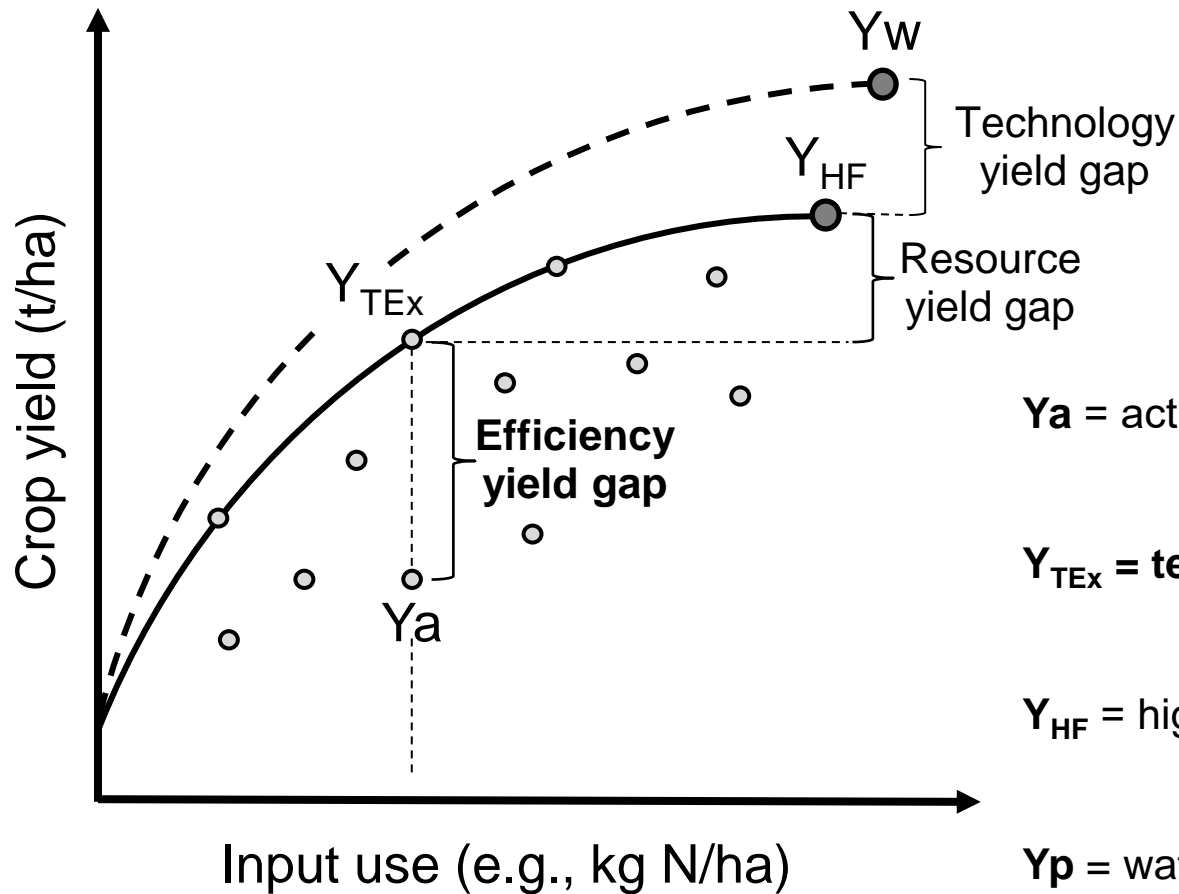
Introduction

Yield gap decomposition has been increasingly applied in agronomy to disentangle the impact of sub-optimal management on crop production and to identify agronomic measures to improve yields. To date, most applications refer to cereal crops (and some tuber and root crops) in a wide range of production systems worldwide, particularly in sub-Saharan Africa, South and Southeast Asia, and Northwest Europe. This notebook aims to formalize the R scripts used to decompose yield gaps across most of those applications making use of the framework introduced by [Silva et al. \(2017\)](#). Data collected by CIMMYT and EIAR for wheat in Ethiopia, previously used for yield gap analysis ([Silva et al., 2021](#)), are used here as an example. Before diving into the R scripts it is important to understand the key concepts and definitions involved in yield gap decomposition as these determine how the different yield levels and associated yield gaps are estimated.

The framework for yield gap decomposition described in this notebook considers four different yield levels ([Silva et al., 2017](#)). First, the **water-limited potential yield** (Y_w) refers to the maximum yield that can be obtained under rainfed conditions in a well-defined, and relatively homogeneous, biophysical environment ([van Ittersum et al., 2013](#)). Y_w can be simulated with crop growth models or derived from field trials with non-limiting levels of nutrients and pests, diseases, and weeds fully controlled. Second, the **highest farmers' yield** (Y_{HF}) refer to the maximum yields (e.g. average above the 90th percentile of actual farmers' yields) observed in a representative sample of farmers sharing



Efficiency yield gap and Y_{TEx}



- Y_a = actual farmers' yields
 - Farm household surveys
- Y_{TEx} = **technical efficient yields**
 - **Stochastic frontier analysis**
- Y_{HF} = highest farmers' yields
 - Top 10th percentile of Y_a
- Y_p = water limited potential yield
 - Global Yield Gap Atlas

Stochastic frontier analysis

- ❑ SFA is a **parametric** method for technical efficiency analysis that differentiates **two error terms** (random noise, v , and technical inefficiency, u).

$$\ln y_{it} = \alpha_0 + \sum_k^K \beta_k \ln x_{kit} + \frac{1}{2} \sum_k^K \sum_j^K \theta_{kj} \ln x_{kit} \times \ln x_{jit} + \delta T + \lambda T^2 + v_{it} - u_{it}$$

- ❑ **Technical efficiency** measures the effectiveness of converting inputs to outputs.
TE = 1 implies that maximum output is produced per unit input.
- ❑ **Econometric** method developed and widely applied by economists. But also used in many other disciplines.
- ❑ Applied to **individual firms** (e.g., farms, factories, banks, etc.) to understand the scope for improving performance and efficiency of production.

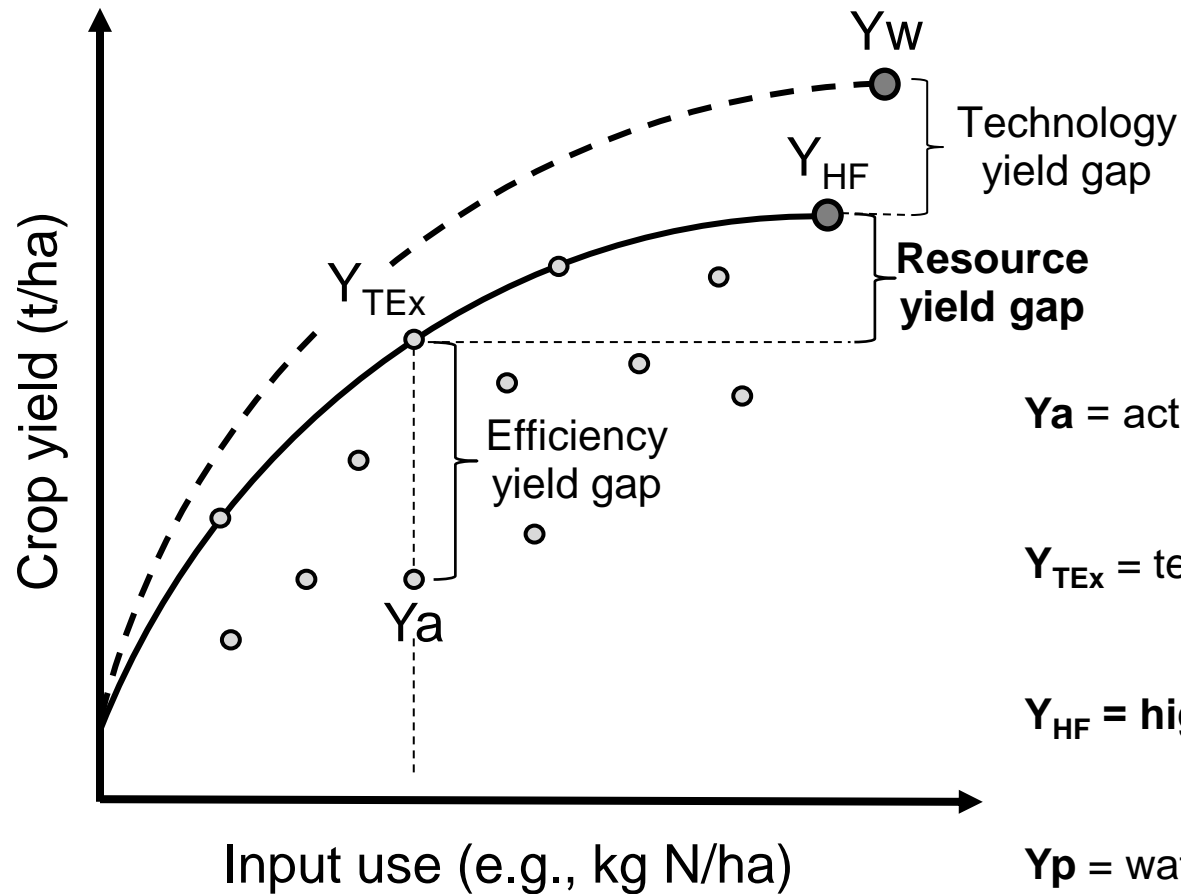


Stochastic frontier analysis in R

- ❑ Start **always** with `lm()` and check R^2 , residuals and multi-collinearity (VIF).
- ❑ `sfa()` function of the R package *frontier*.
- ❑ Battese & Coelli models:
 - Single-output multiple-input frontier;
 - 1992 model: Cross-section vs. panel data;
 - 1995 model: Single step estimation of frontier & inefficiency effects;
- ❑ Functional forms: Cobb-Douglas vs translog.
- ❑ Control for climate, soil, and varieties, so yield gaps are due to management only.
- ❑ Decide on variables for production frontier vs. inefficiency effects.

```
# fit cobb-douglas stochastic frontier
sfa_cd <-
  sfa(yield_tha ~
    season_year + gyga_gdd + gyga_tseas + seed_kgha + variety +
    gyga_ai + gyga_av_water + soil_depth + soil_fertility + waterlogging_yn + drought_yn + soilwatercons
    nfert_kgha + manure_yn + residues_yn + previous_crop + oxplough_freq_cat +
    herb_lha + handweeding_persdayha + weeding_yn + pesticide_yn + disease_incidence_yn + pest_incidence
    data=data_new)
```

Resource yield gap and Y_{HF}



- Y_a** = actual farmers' yields
 - Farm household surveys
- Y_{TEX}** = technical efficient yields
 - Stochastic frontier analysis
- Y_{HF}** = highest farmers' yields
 - Top 10th percentile of Y_a
- Y_p** = water limited potential yield
 - Global Yield Gap Atlas

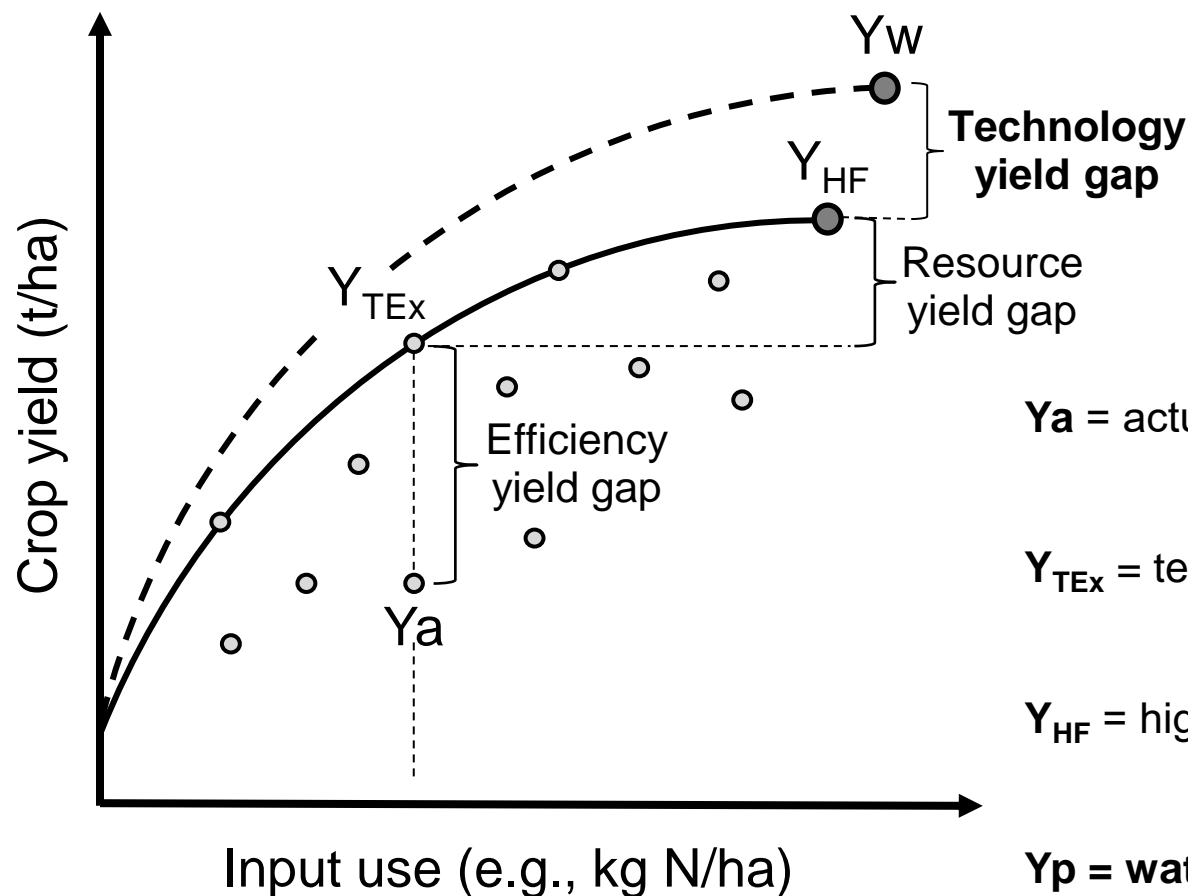
Quantiles within 'for loop' in R

```
# create an empty data frame
data_final <- data.frame()
#
# create loop per year
for(yr in unique(data$year)){
  subset_year <- subset(data, year == yr)
  #
  # create loop per climate zone
  for(cz in unique(subset_year$gyga_cz)){
    subset_cz <- subset(subset_year, gyga_cz == cz)
    #
    # create loop per soil type
    for(soil in unique(subset_cz$soil_fertility)){
      subset_soil <- subset(subset_cz, soil_fertility == soil)
      # create column with field class based on yield distribution
      subset_soil$field_class <- ifelse(subset_soil$yield_tha >= quantile(subset_soil$yield_tha, 0.90),
        'YHF', '')
      subset_soil$field_class <- ifelse(subset_soil$yield_tha <= quantile(subset_soil$yield_tha, 0.10),
        'YLF', subset_soil$field_class)
      subset_soil$field_class <- ifelse(subset_soil$yield_tha > quantile(subset_soil$yield_tha, 0.10) &
        subset_soil$yield_tha < quantile(subset_soil$yield_tha, 0.90),
        'YAF', subset_soil$field_class)

      #
      # subset highest yielding fields only
      yhf <- subset(subset_soil, field_class == 'YHF')
      #
      # add column with yhf in t/ha to data frame
      subset_soil['yhf_tha'] <- mean(yhf$yield_tha, na.rm=T)
      #
      # bind all individual fields into single data frame
      data_final <- rbind(data_final, subset_soil)
    }
  }
}
```



Technology yield gap and Y_w



- Y_a** = actual farmers' yields
- Farm household surveys
- Y_{TE}** = technical efficient yields
- Stochastic frontier analysis
- Y_{HF}** = highest farmers' yields
- Top 10th percentile of Y_a
- Y_p** = water limited potential yield
- Global Yield Gap Atlas

Global Yield Gap Atlas in R

Yield gap analysis

Search docs

Installing R

Concepts and definitions

Data collection tool

Stochastic frontier analysis

Workflow for Silva et al. (2017)

Retrieve data from GYGA

Introduction

Load required R packages

Access to GYGA data

Link to farmer field data

Export file with GYGA data

Final remarks

Boundary line analysis

Random forest and Shapely values

Docs » Stochastic frontier analysis » Retrieve data from GYGA

R code

Retrieve data from GYGA

- João Vasco Silva, CIMMYT-Zimbabwe
- Marloes van Loon, WUR

Introduction

This notebook complements an earlier notebook describing the methodology for yield gap decomposition. That earlier notebook makes use of water-limited yield data to decompose yield gaps. Such data were derived using the scripts documented in this notebook. The reader is referred to that earlier notebook for further information about the concepts and definitions considered in yield gap analysis. To make the approach fully reproducible, it is explained here how to retrieve the water-limited yield data from the Global Yield Gap Atlas (GYGA) using available APIs for acquiring such data. Also here an example is provided for for wheat in Ethiopia.

Load required R packages

First, the R packages needed to run this workflow are loaded.

```
# package names
packages <- c("dplyr", "tidyr", "httr", "jsonlite", "sf", "reshape2")
#
# install packages
installed_packages <- packages %in% rownames(installed.packages())
if (any(installed_packages == FALSE)){
```





**Thank you
for your
interest!**