

Data requirements for yield gap analysis

João Vasco Silva (PhD)

j.silva@cgiar.org

Agronomy-at-scale Data Scientist

CIMMYT-Zimbabwe



Addis Ababa, May 23rd 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



Minimum data requirements

- 1) Delineation of relevant **biophysical units** to ensure differences between farm/field performance are only attributed to management practices.
- 2) **Farmer field data** to estimate Y_a , Y_{Tex} and Y_{HF} , and associated efficiency and resource yield gaps.
 - a) Stochastic frontier analysis
 - b) Actual yield percentiles
 - c) What were the reasons behind the selection of these sites?
- 3) **(Simulated) yield ceilings** from crop growth simulation models to estimate technology yield gaps.



Farmer field data

GPS coordinates, water regime, variety, sowing & harvest dates, yield

Crop management (water, nutrient, pest, disease, weed), socio-economics





**Thank you
for your
interest!**