Imputation Methodology for FAOSTAT Production Domain

Michael. C. J. Kao

Food and Agriculture Organization
of the United Nation

Outline

Why do we need imputation?

The agricultural production domain is integral to the compilation of Food Balance Sheets. In particular to estimate consistent food supplies, imputation is required to ensure that data are non-sparse. Owing to the potential impact of imputation when often data are missing, accuracy and reliability of food estimates cannot be compromised.

The relationship of production and its components can be expressed as:

$$P_t = A_t \times Y_t \tag{1}$$

Where P_t, A_t and Y_t denotes production, area harvested and yield, respectively, at time t.

Outline for section 1

The presently applied methodology aims to capture the variation of relevant commodity and/or geographic characteristics through the application of aggregated growth rates. a five-hierachy was designated represented by:

1. Same country/commodity aggregate
2. Sub-region aggregate/same commodity
3. Sub-region aggregate/commodity aggregate
4. Regional aggregate/same commodity
5. Regional aggregate/commodity aggregate

In short, the aggregation imputation method computes the commodity/regional aggregated growth of both area and production, the growth rate is then applied to the last observed value. The formulae of the aggregated growth can be expressed as:

$$r_{s,t} = \sum_{c \in \mathbb{S}} X_{c,t} / \sum_{c \in \mathbb{S}} X_{c,t-1} \tag{2}$$

The imputation can then be computed as:

$$\hat{X}_{c,t} = X_{c,t-1} \times r_{s,t} \tag{3}$$

There are several shortcomings of the current methodology,

- Divergence of area and production, there are mainly two reasons for this.
  1. Due to missing values, the aggregated growth can be heavily biased.
  2. The basket used to compute the aggregated growth rate is not the same over time and between area and production.
- Assumes perfect correlation between group and country series.
- Cannot support and incorporate additional information.

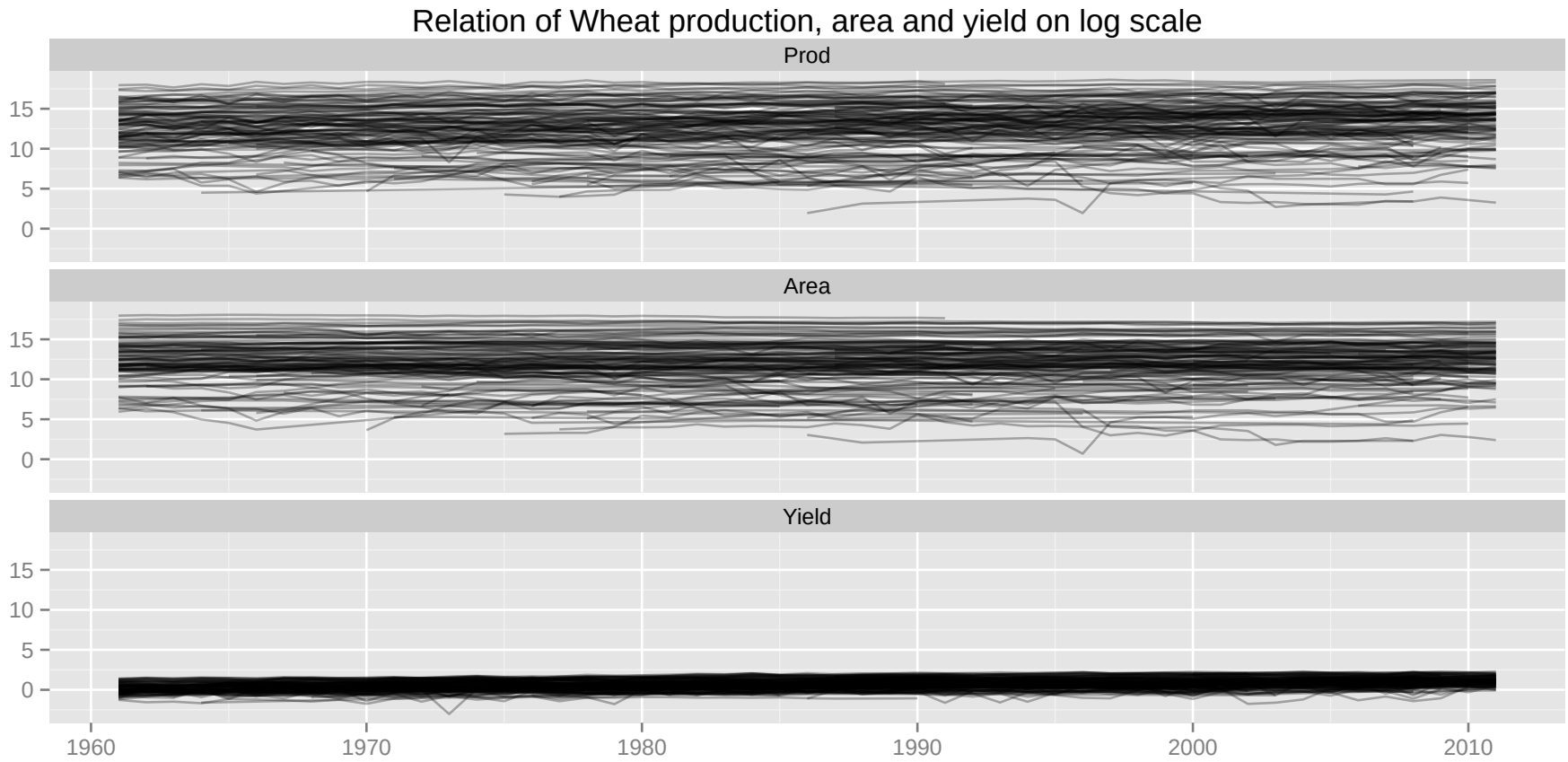
Outline for section 2

To illustrate the proposed imputation, we have chosen Wheat from the crop group and Cassava from the root group as case studies. Both are important agricultural products but Wheat is much more commercialised and traded than is Cassava.

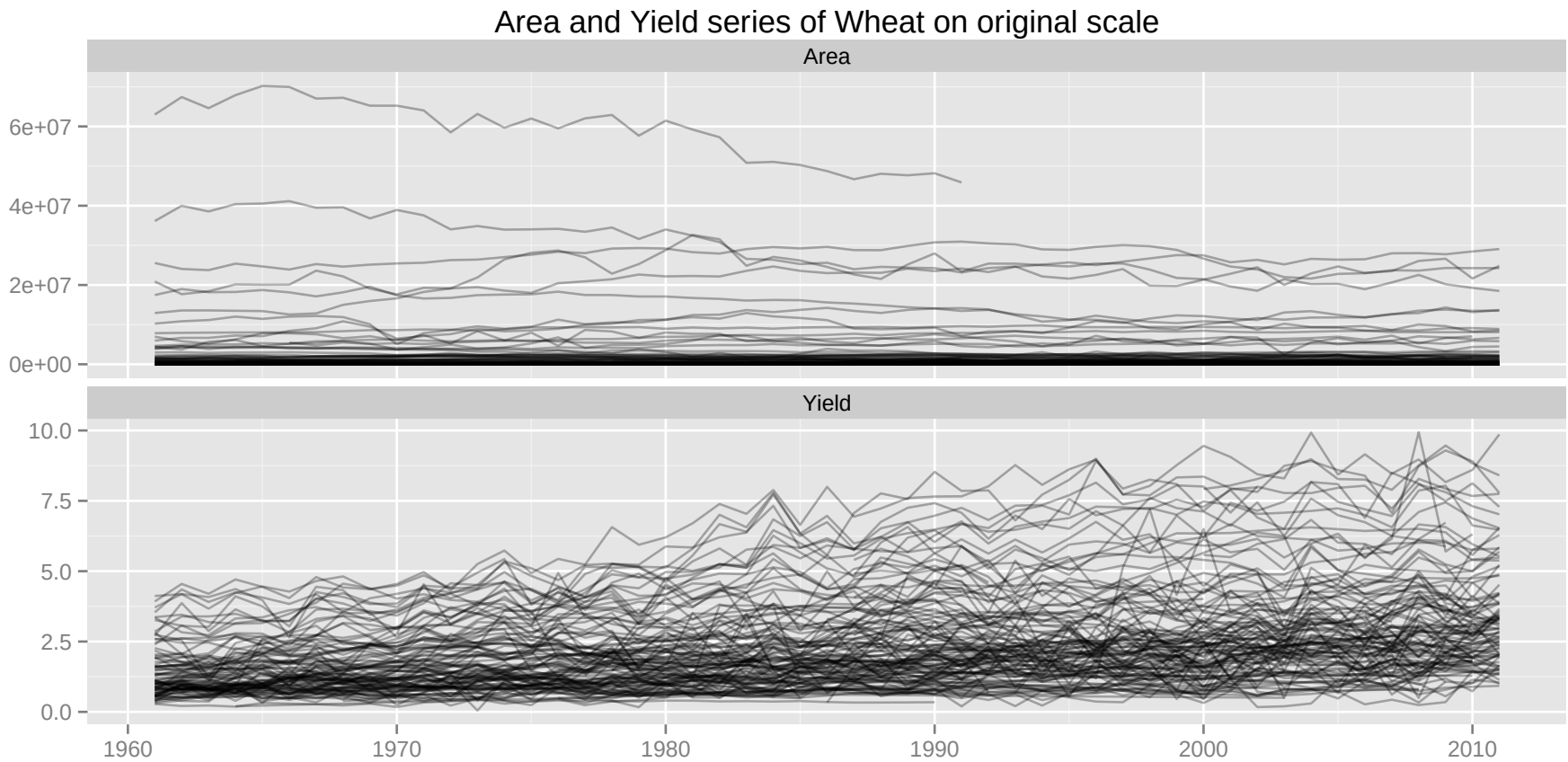
We have log-transformed the data to make the relationsihp an additive one, so that the production can be decomposed.

$$\log(P_t) = \log(A_t) + \log(Y_t) \tag{4}$$

Area dictates the level and changes in the production



Let us dig deeper



What is the data telling us?

What the data have shown is that the level, trend of the production is mainly determined by a smooth monotonic area occasionally affected by shocks, while the yield generates the variation from year-to-year reflecting climate or economic conditions.

This leads to the proposed methodology to estimate the year-to-year variation of yield while a stable method for area.

Outline for section 3

First of all, we propose to impute the yield and area, and along with the restriction of the new model this almost guarantees that area and production will not diverge.

Second, instead of applying the changes directly, the model estimates the relationship between the country and the aggregated series and applies the factors accordingly.

Finally the proposed model allows incorporation of additional information such as prices, vegetation indices and other data that may improve the accuracy of the imputation.

Linear Mixed Model

To capture the co-movement of yield and model sub-regional differences, we have proposed to model the yield with a Linear Mixed Model (LME), which can be expressed as follows in matrix notation:

$$\begin{aligned} \mathbf{y_i} &= \mathbf{X_i}\beta + \mathbf{Z_i}\mathbf{b_i} + \epsilon_i \\ \mathbf{b_i} &\sim \mathbf{N_q}(\mathbf{0}, \mathbf{\Psi}) \\ \epsilon_i &\sim \mathbf{N_{ni}}(\mathbf{0}, \sigma^2\mathbf{\Lambda_i}) \end{aligned} \tag{5}$$

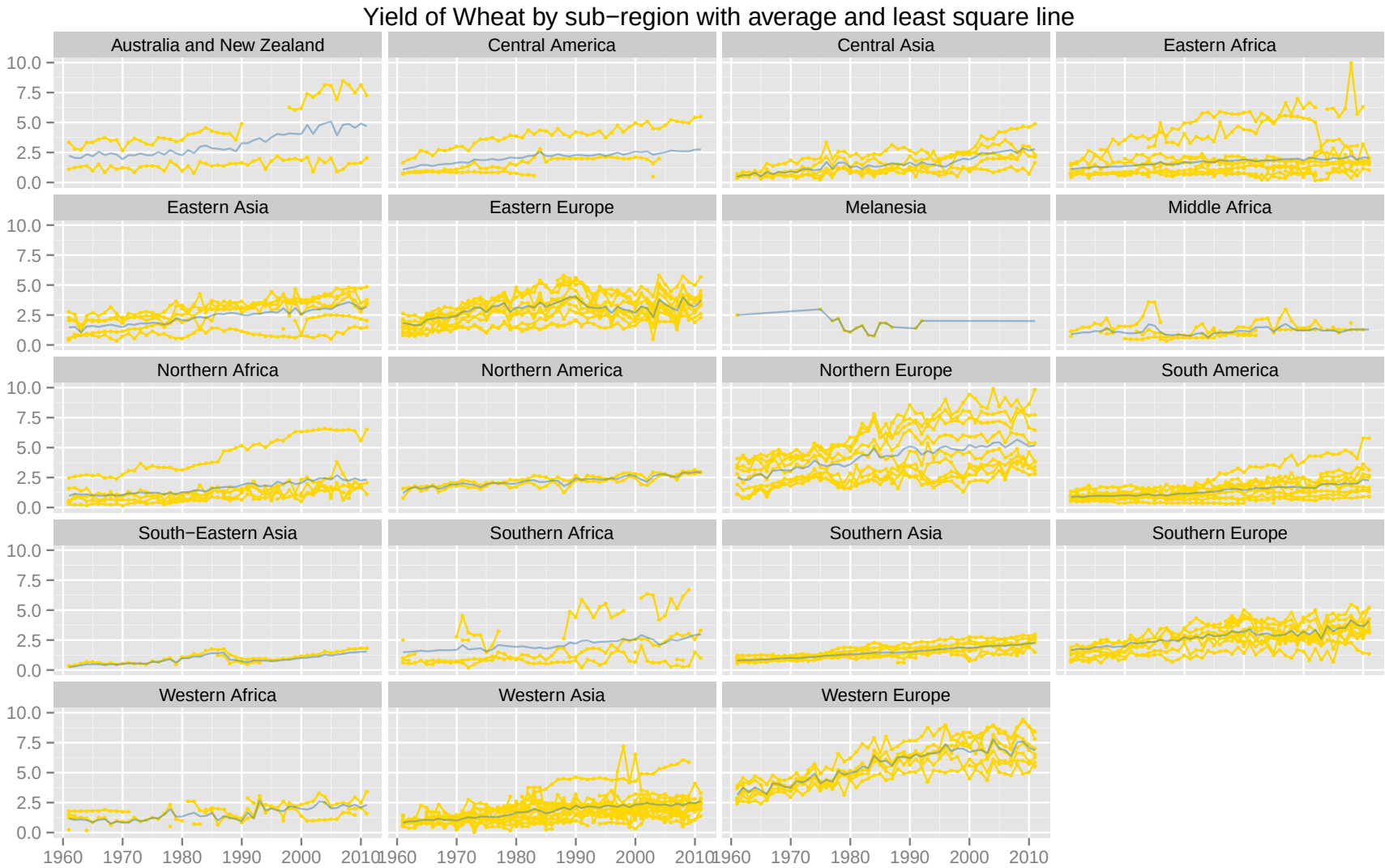
More specifically, the equation for the imputation is

$$Y_{i,t} = \overbrace{\beta_{0j} + \beta_{1j} t}^{\text{Fixed effect}} + \overbrace{b_{0,i} + b_{1,i} t + b_{2,i,t} \bar{Y}_{j,t}}^{\text{Random effect}} + \epsilon_{i,t} \tag{6}$$

The average yield can be calculated as follow,

$$\bar{Y}_{j,t} = \sum_{i \in j} \omega_i Y_{i,t} \tag{7}$$

Which acts as a proxy to reflect the change in climatics conditions and other factors which can simultaneously affect multiple countries. However, due to missing values, this quantity is not computed directly from the raw data. The EM-algorithm is implemented for the estimation for the unbiased average.


Yield of Wheat by sub−region with average and least square line
Australia and New Zealand
Central America
Central Asia
Eastern Africa
Eastern Asia
Eastern Europe
Melanesia
Middle Africa
Northern Africa
Northern America
Northern Europe
South America
South−Eastern Asia
Southern Africa
Southern Asia
Southern Europe
Western Africa
Western Asia
Western Europe
10.0
7.5
5.0
2.5
0.0
1960
1970
1980
1990
2000
2010

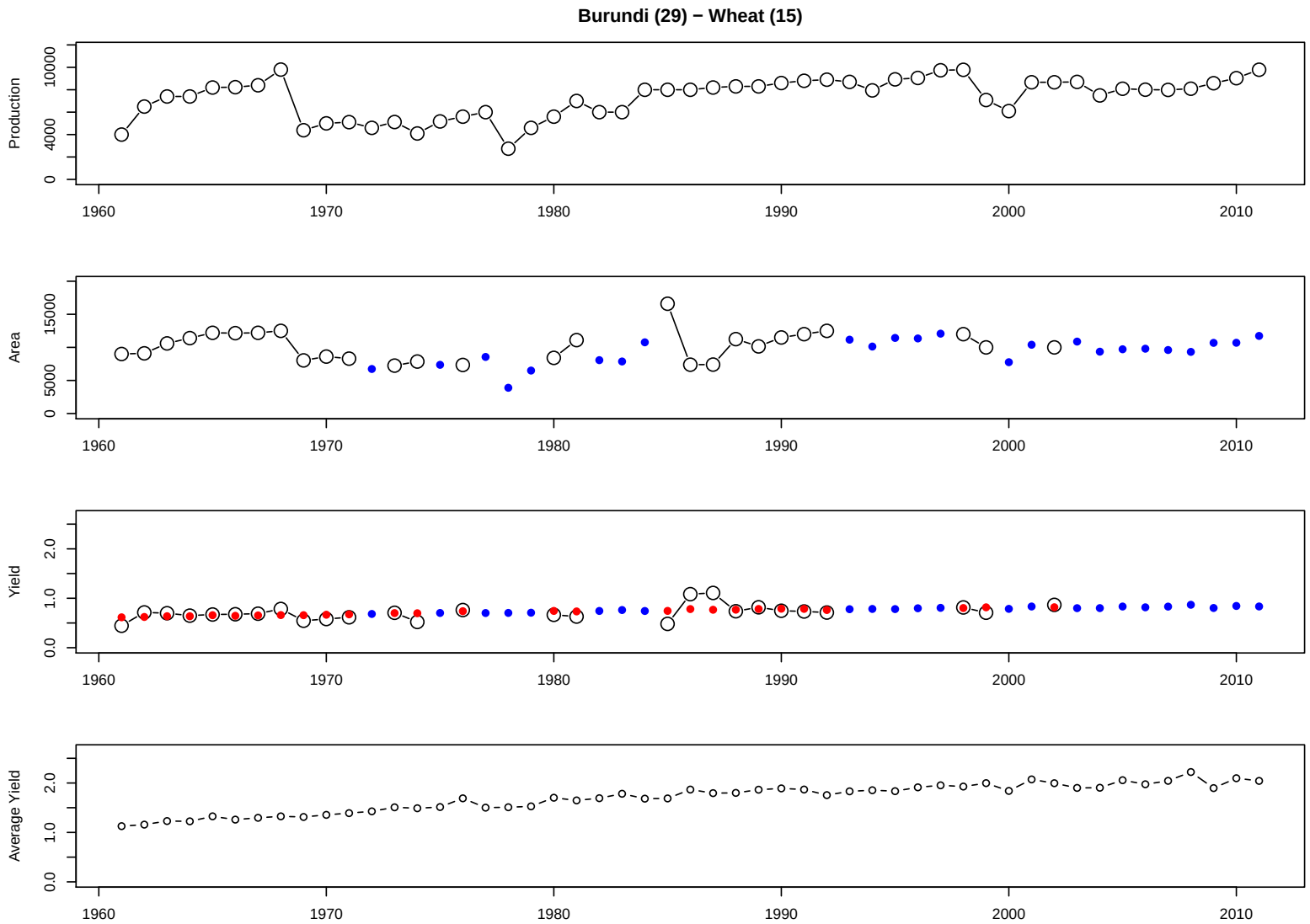
Currently we have adopted **linear interpolation** and **last observation carry forward** to impute area harvestd.

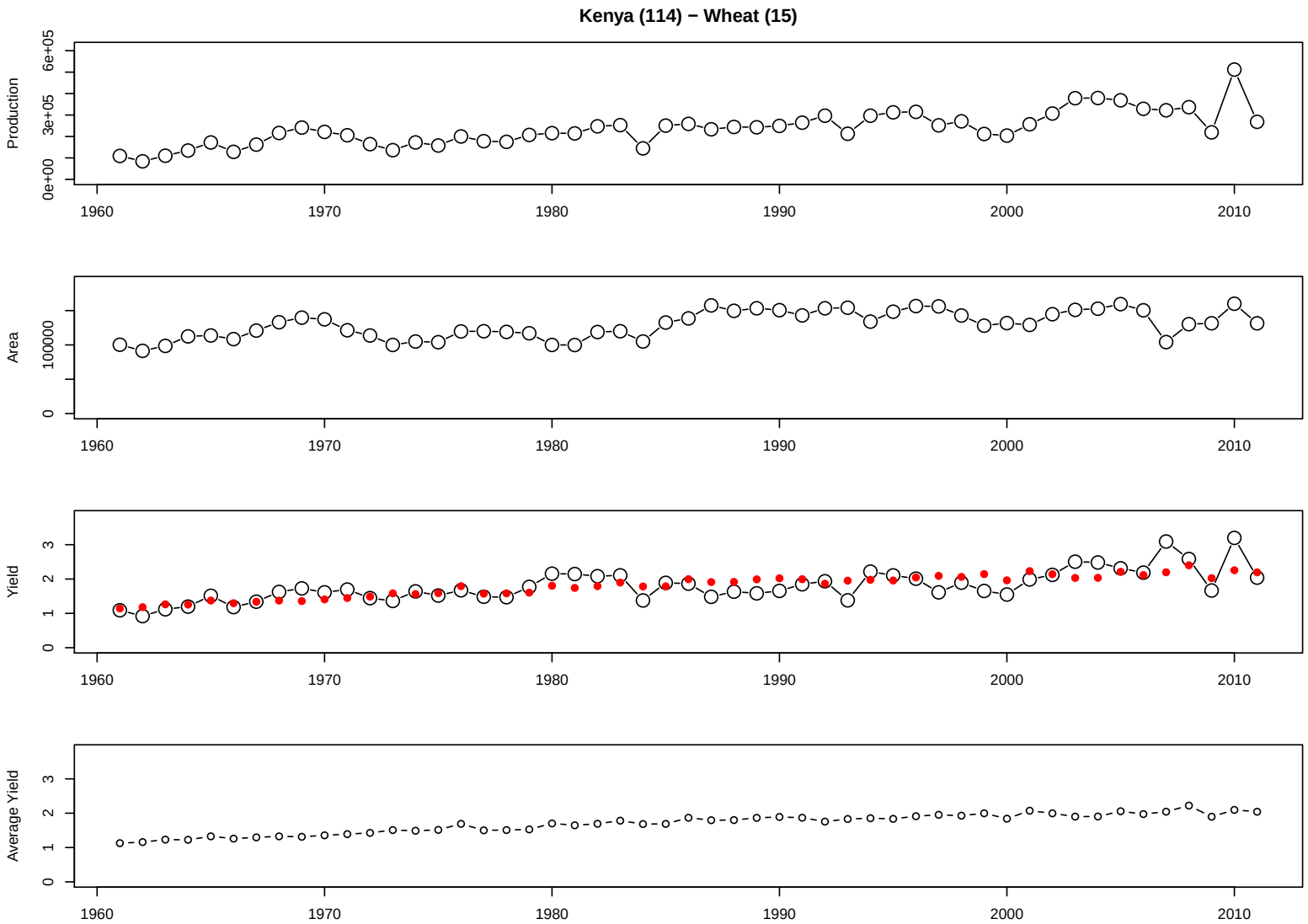
First the area harvested displays close to monotonic behaviour and much little year-to-year fluctuation and thus linear interpolation is suitable. Furthermore, a previous simulation study has shown linear interpolation gives best result.

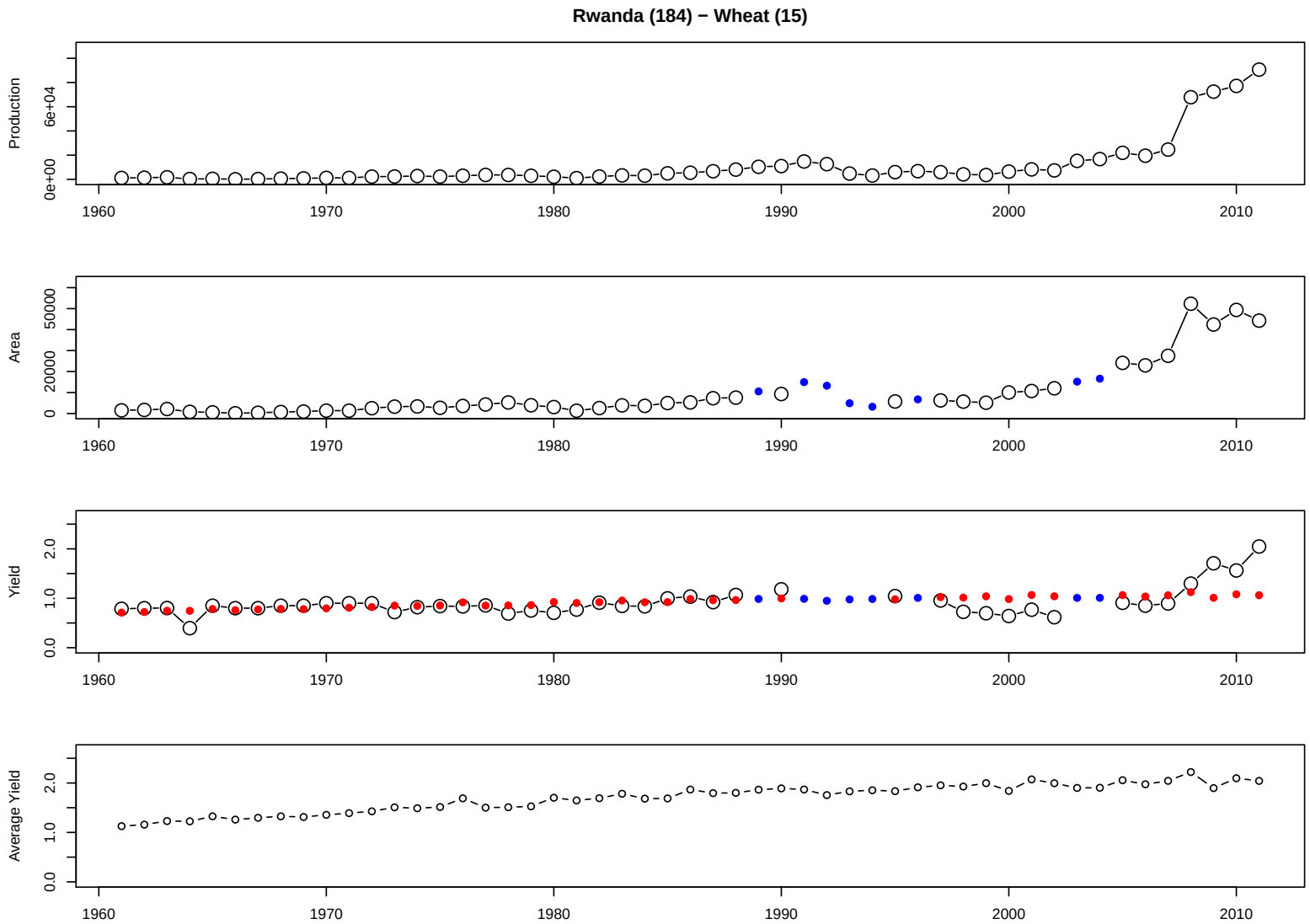
While "last observation carry forward" is useful when the last observed value is a true zero, we will not impute a positive value.

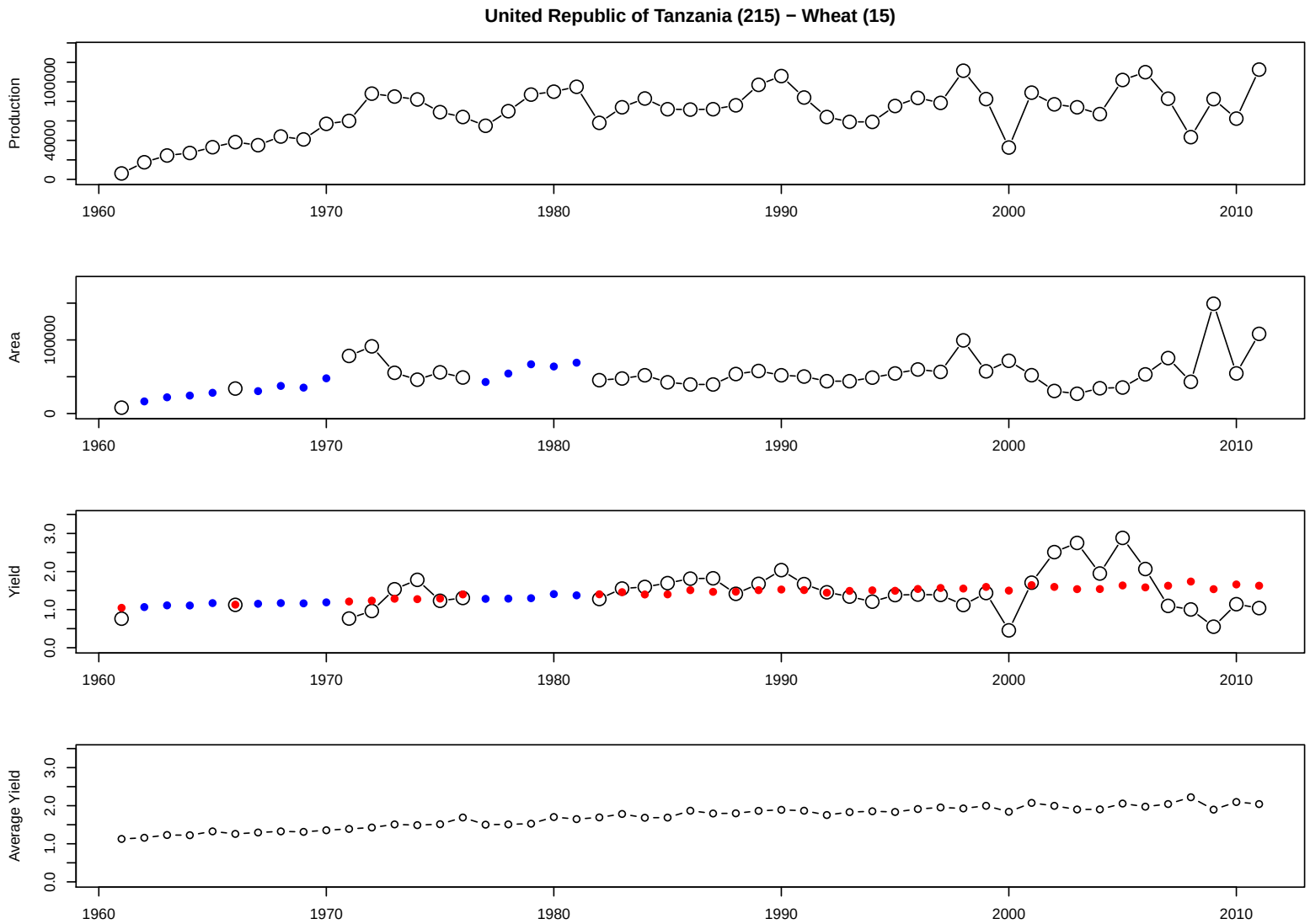
We are, however, investigating improvements in the imputation of area harvested using information such as area sown.

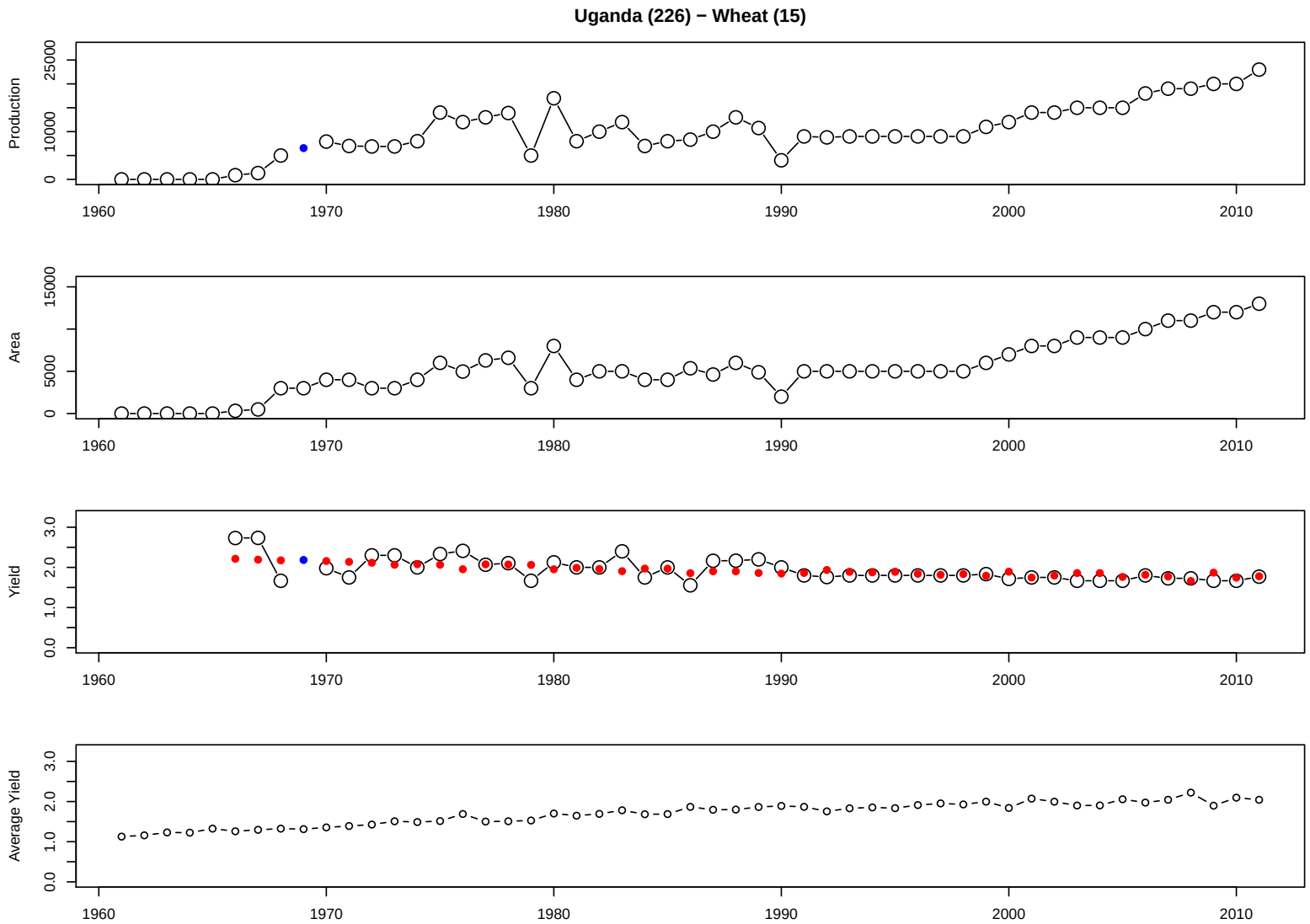
Outline for section 4

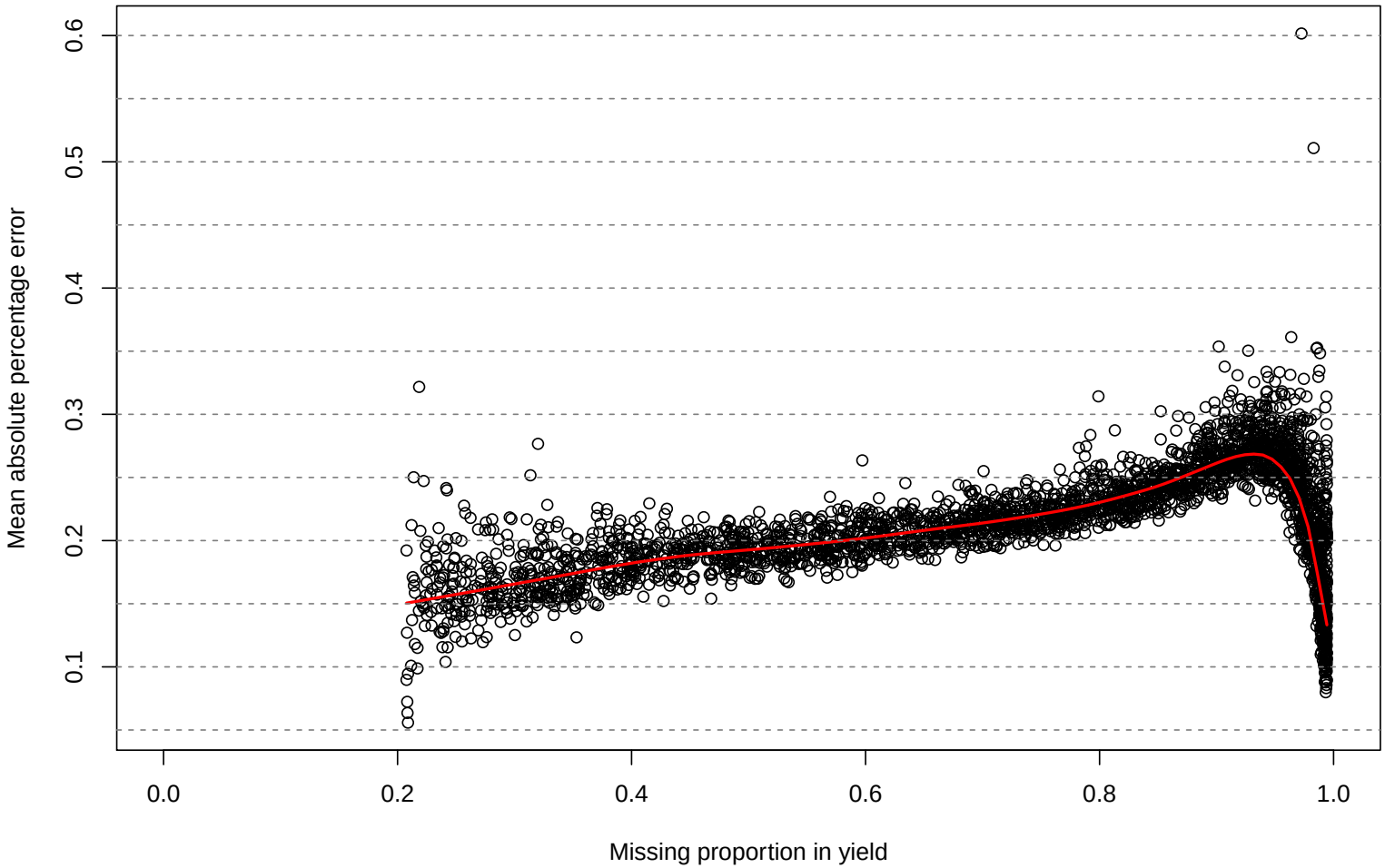

Burundi (29) − Wheat (15)
Production
0
4000
10000
Area
0
5000
15000
Yield
0.0
1.0
2.0
Average Yield
0.0
1.0
2.0
1960
1970
1980
1990
2000
2010


Kenya (114) − Wheat (15)
Production
0e+00
3e+05
6e+05
Area
0
100000
Yield
0
1
2
3
Average Yield
0
1
2
3
1960
1970
1980
1990
2000
2010


Rwanda (184) − Wheat (15)
Production
0e+00
6e+04
Area
0
20000
50000
Yield
0.0
1.0
2.0
Average Yield
0.0
1.0
2.0
1960
1970
1980
1990
2000
2010


United Republic of Tanzania (215) − Wheat (15)
Production
0
40000
100000
Area
0
100000
Yield
0.0
1.0
2.0
3.0
Average Yield
0.0
1.0
2.0
3.0
1960
1970
1980
1990
2000
2010


Uganda (226) − Wheat (15)
Production
Area
Yield
Average Yield
1960
1970
1980
1990
2000
2010


Mean absolute percentage error
0.1
0.2
0.3
0.4
0.5
0.6
0.0
0.2
0.4
0.6
0.8
1.0
Missing proportion in yield

Outline for section 5

- Incorporation additional information for imputation.
- Develope a better grouping classification.

Outline for section 6

The newly proposed methodology demonstrates the ability to resolve issues in the current methodology and extended to incorporate additional information.

We welcome any information which can enhance the performance of the imputation.

Thank you for your time and attention.