Introduction
Current Methodology
Exploratory Data Analysis
Proposed Methodology
Results
Implementation
Discussion

Imputation Methodology for FAOSTAT Production Domain

Michael. C. J. Kao

Food and Agriculture Organization
of the United Nation

July 12, 2013

Introduction
Current Methodology
Exploratory Data Analysis
Proposed Methodology
Results
Implementation
Discussion

Outline

Introduction
Current Methodology
Exploratory Data Analysis
Proposed Methodology
Results
Implementation
Discussion

Outline for section 1

The aim of this presentation to give an over view of the current status of the newly proposed imputation methodology for the FAOSTAT production domain.

The work presented is the current status at the date of the presentation, and is subject to further changes.

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.

However, it must be recognized that imputation should only be used as a last resort.

Introduction
Current Methodology
Exploratory Data Analysis
Proposed Methodology
Results
Implementation
Discussion

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/animal slaughtered and yield/carcass weight, respectively, at time t.

Introduction
Current Methodology
Exploratory Data Analysis
Proposed Methodology
Results
Implementation
Discussion

Outline for section 2

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

Introduction
Current Methodology
Exploratory Data Analysis
Proposed Methodology
Results
Implementation
Discussion

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:

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

The imputation can then be computed as:

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

Introduction
Current Methodology
Exploratory Data Analysis
Proposed Methodology
Results
Implementation
Discussion

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.

Introduction
Current Methodology
Exploratory Data Analysis
Proposed Methodology
Results
Implementation
Discussion

Outline for section 3

To give a brief demonstration, we have chosen Wheat to illustrate the properties and relationships between the time series.

The following graph illustrates the relationship between the production, area and yield.
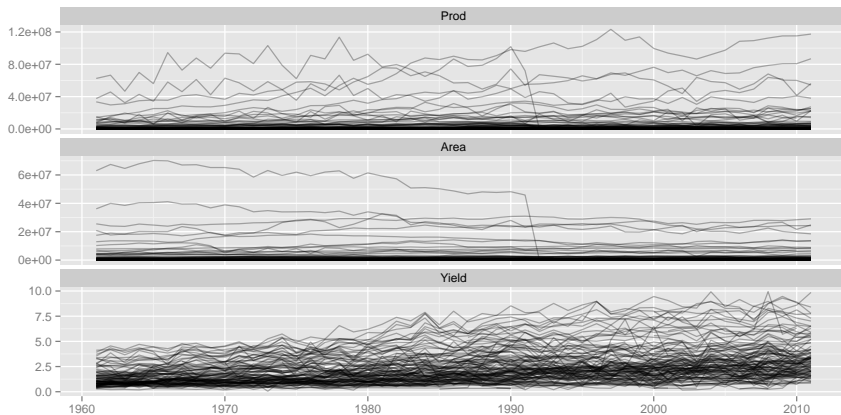
Now let us log transform the data so it becomes an additive relationshp.

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

Area dictates the level and changes in the production

Relation of Wheat production, area and yield on log scale

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 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.

Introduction
Current Methodology
Exploratory Data Analysis
Proposed Methodology
Results
Implementation
Discussion

Imputation for Area Harvested
Imputation for Yield
Imputation Procedure

Outline for section 4

Introduction
Current Methodology
Exploratory Data Analysis
Proposed Methodology
Results
Implementation
Discussion

Imputation for Area Harvested
Imputation for Yield
Imputation Procedure

First of all, we propose to impute the yield and area. This along with the restriction of the new model and the decomposition strategy 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, precipitation and other information that may improve the accuracy of the imputation.

Introduction
Current Methodology
Exploratory Data Analysis
Proposed Methodology
Results
Implementation
Discussion

Imputation for Area Harvested
Imputation for Yield
Imputation Procedure

Currently we have adopted **linear interpolation** and **last observation carry forward** to impute area harvestd or what we called the naive imputation.

First the area harvested and in particular carcass weight per animal and trees displays extremely smooth behaviour and little year-to-year fluctuation and thus linear interpolation is suitable. Furthermore, a previous simulation study has shown linear interpolation gives best result in some cases.

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

Introduction
Current Methodology
Exploratory Data Analysis
Proposed Methodology
Results
Implementation
Discussion

Imputation for Area Harvested
Imputation for Yield
Imputation Procedure

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:

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

Introduction
Current Methodology
Exploratory Data Analysis
Proposed Methodology
Results
Implementation
Discussion

Imputation for Area Harvested
Imputation for Yield
Imputation Procedure

More specifically, the equation for the imputation is

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

The average yield can be calculated as follow,

$$\bar{Y}_{j,t} = \sum_{i \in j} \omega_{i,t} Y_{i,t} \qquad (7)$$

Which acts as a proxy to reflect the change in climatics conditions
and other factors which can simultaneously affect multiple
countries.

Introduction
Current Methodology
Exploratory Data Analysis
Proposed Methodology
Results
Implementation
Discussion

Imputation for Area Harvested
Imputation for Yield
Imputation Procedure

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.
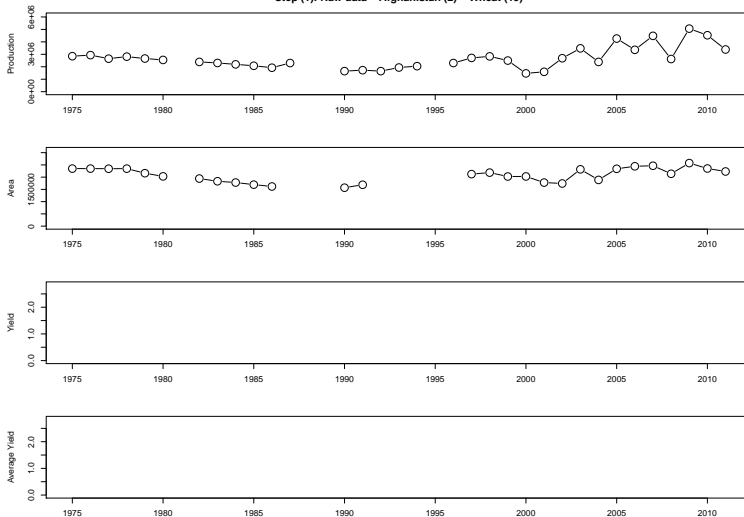
The average yield is only required if there are factors which can simultaneously affect several countries within the region. That is, the effect is only inlcluded when it improves the model.

The algorithm starts with the null model without the average effect, then **AIC** is used as the decision criteria as whether the model with average effect is needed.
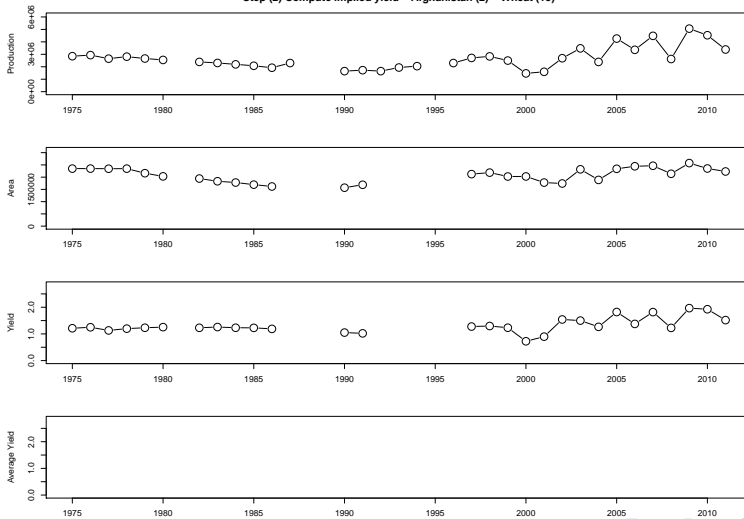
Introduction
Current Methodology
Exploratory Data Analysis
Proposed Methodology
Results
Implementation
Discussion

Imputation for Area Harvested
Imputation for Yield
Imputation Procedure

Yield of Wheat by sub-region with average and least square line

Introduction
Current Methodology
Exploratory Data Analysis
Proposed Methodology
Results
Implementation
Discussion

Imputation for Area Harvested
Imputation for Yield
Imputation Procedure

Step (1): Raw data – Afghanistan (2) – Wheat (15)

Introduction
Current Methodology
Exploratory Data Analysis
Proposed Methodology
Results
Implementation
Discussion

Imputation for Area Harvested
Imputation for Yield
Imputation Procedure

Step (2) Compute implied yield – Afghanistan (2) – Wheat (15)

Introduction
Current Methodology
Exploratory Data Analysis
Proposed Methodology
Results
Implementation
Discussion

Imputation for Area Harvested
Imputation for Yield
Imputation Procedure

Step (3) Impute the yield – Afghanistan (2) – Wheat (15)

Introduction
Current Methodology
Exploratory Data Analysis
Proposed Methodology
Results
Implementation
Discussion

Imputation for Area Harvested
Imputation for Yield
Imputation Procedure
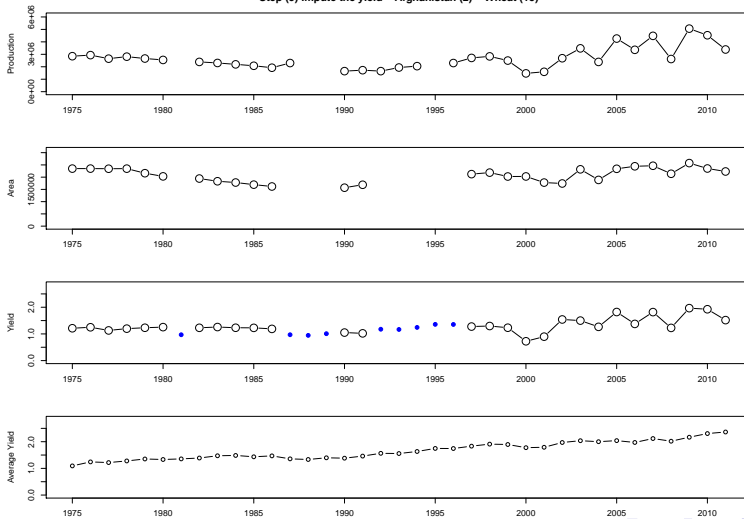
Step (4): Compute area and production where availableAfghanistan (2) – Wheat (15)

Introduction
Current Methodology
Exploratory Data Analysis
Proposed Methodology
Results
Implementation
Discussion

Imputation for Area Harvested
Imputation for Yield
Imputation Procedure
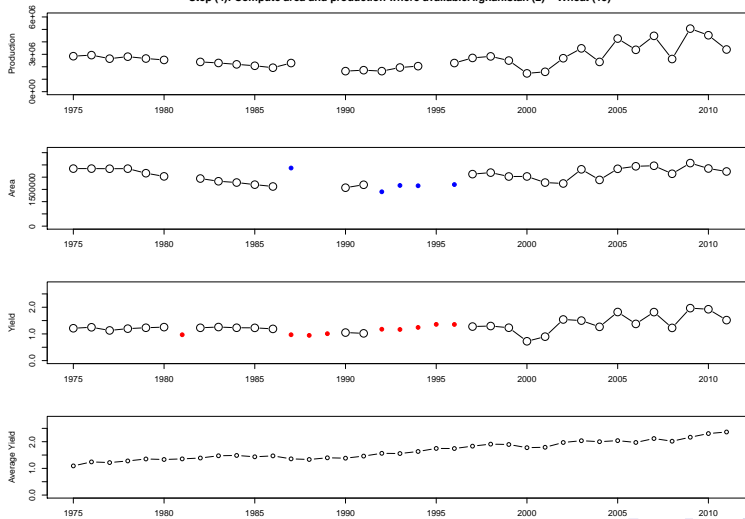
Step (5): Impute area with naive imputation – Afghanistan (2) – Wheat (15)

Introduction
Current Methodology
Exploratory Data Analysis
Proposed Methodology
Results
Implementation
Discussion

Imputation for Area Harvested
Imputation for Yield
Imputation Procedure
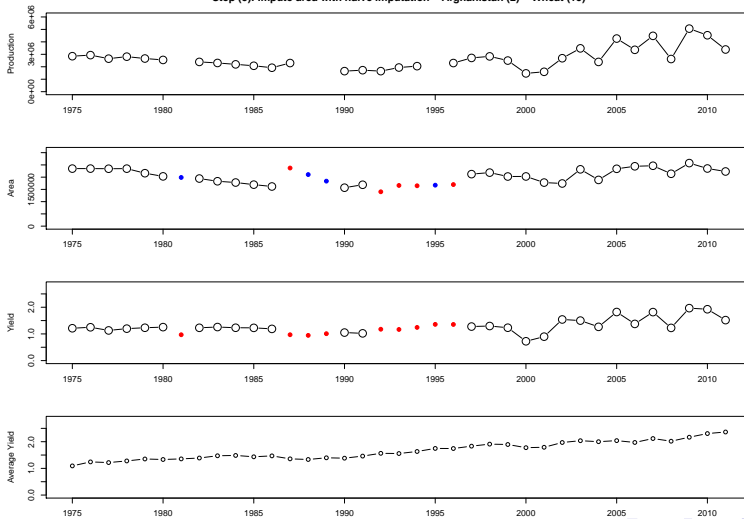
Step (6): Impute the production – Afghanistan (2) – Wheat (15)
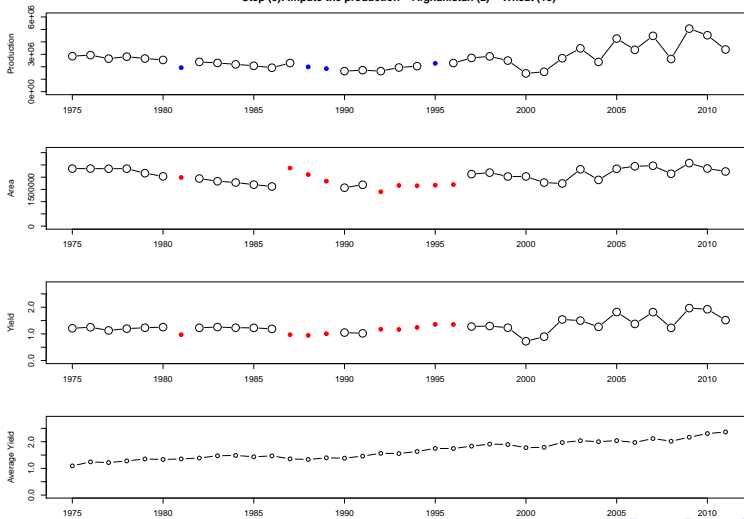
Introduction
Current Methodology
Exploratory Data Analysis
Proposed Methodology
Results
Implementation
Discussion

Individual imputation
Simulation Results

Outline for section 5

Burundi (29) – Wheat (15)

Introduction
Current Methodology
Exploratory Data Analysis
Proposed Methodology
Results
Implementation
Discussion

Individual imputation
Simulation Results

Cameroon (32) – Wheat (15)

Introduction
Current Methodology
Exploratory Data Analysis
Proposed Methodology
Results
Implementation
Discussion

Individual imputation
Simulation Results
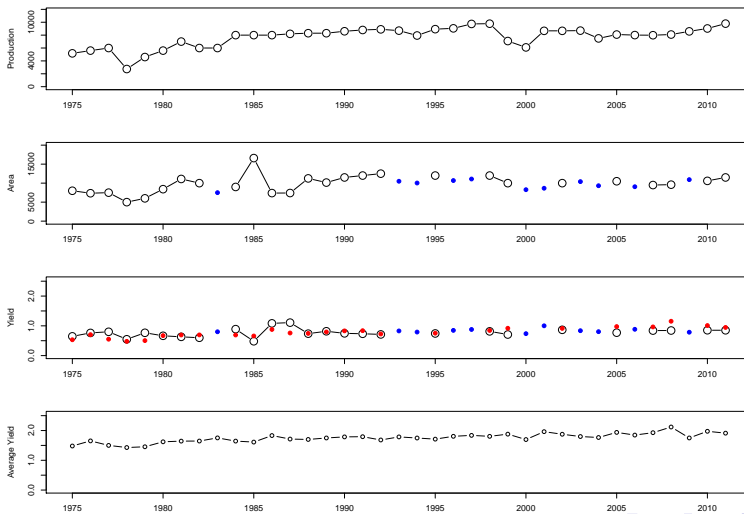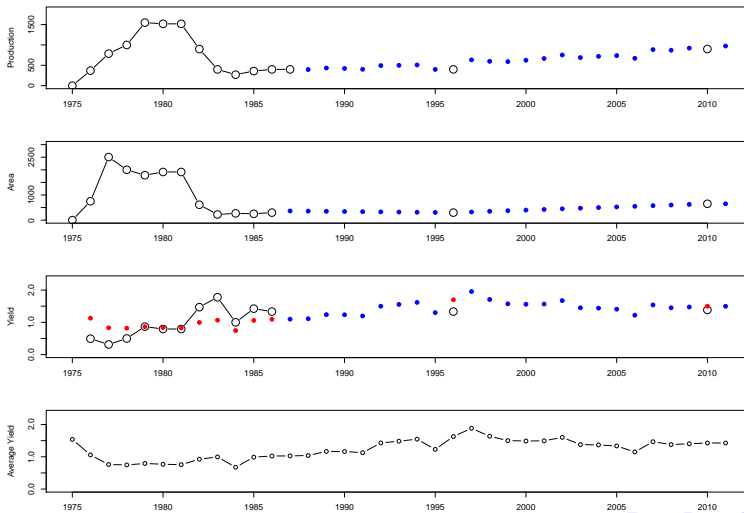
Democratic People's Republic of Korea (116) -- Wheat (15)
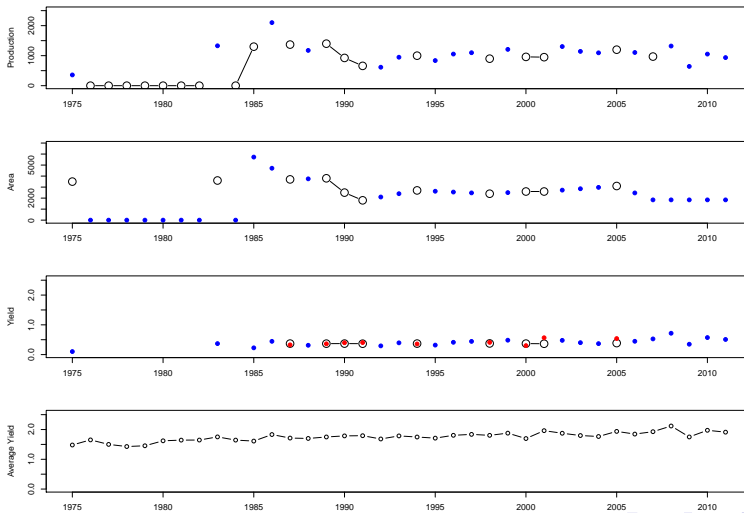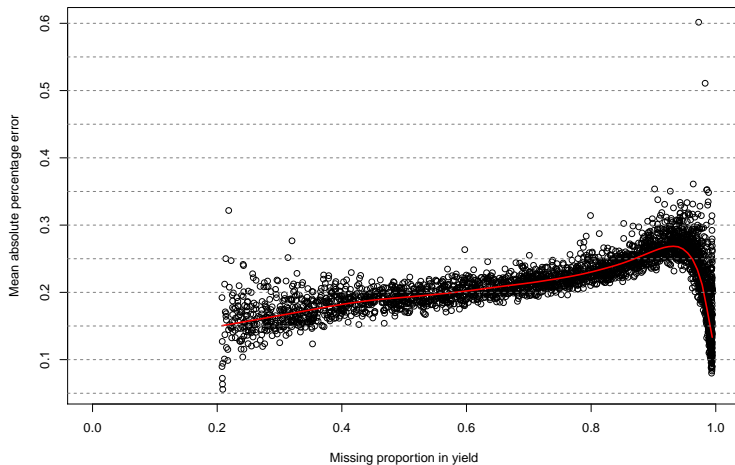
Introduction
Current Methodology
Exploratory Data Analysis
Proposed Methodology
Results
Implementation
Discussion

Individual imputation
Simulation Results

Libya (124) – Wheat (15)

Introduction
Current Methodology
Exploratory Data Analysis
Proposed Methodology
Results
Implementation
Discussion

Individual imputation
Simulation Results

Madagascar (129) – Wheat (15)

Introduction
Current Methodology
Exploratory Data Analysis
Proposed Methodology
Results
Implementation
Discussion

Individual imputation
Simulation Results

Somalia (201) – Wheat (15)

Introduction
Current Methodology
Exploratory Data Analysis
Proposed Methodology
Results
Implementation
Discussion

Outline for section 6

Introduction
Current Methodology
Exploratory Data Analysis
Proposed Methodology
Results
Implementation
Discussion

Following the methodology, we will now present how the methodology will be integrated into the Food Balance Sheet cycle. Imputation does not create information, and it should be used as a last resort. Model can only be as good as data.

1. Receive data from production questionaire, countrySTAT or other official/semi-official source.
2. Perform the imputation with data since 1980 to the most recent data, but only use the imputation for the years that are relevant to the preperation of the current Food Balance Sheet.
3. Conduct expert assessment, and revise the imputation.

Introduction
Current Methodology
Exploratory Data Analysis
Proposed Methodology
Results
Implementation
Discussion

Outline for section 7

Introduction
Current Methodology
Exploratory Data Analysis
Proposed Methodology
Results
Implementation
Discussion

Some on going work

- Investigate the performance of state-space model.

- Investigate ensemble models.

- Design a measure to quantify variance, in particular the frequency of fluctuation.

Introduction
Current Methodology
Exploratory Data Analysis
Proposed Methodology
Results
Implementation
Discussion

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.

You can find all the data, codes implementation, documentation, and the wiki page at the following repository. You can also install the package following the instruction on the page.

https://github.com/mkao006/Imputation