Imputation of the Production Domain

Michael C. J. Kao

Food and Agriculture Organization
of the United Nations

Outline

1. Introduction
2. Yield
3. Production and Area Harvested

Outline for section 1

1 Introduction

2 Yield

3 Production and Area Harvested

Relationships

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

$$P_t := A_t \times Y_t \quad P_t \geq 0, A_t \geq 0, Y_t > 0 \tag{1}$$

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

Scope of the project

A total of **169** commodity just in the crop domain.

228 countries including obsolete classification and territories.

A total of **31,797** time series require imputation.

Percentage of missing value can be as high as **80%** (By commodity).

Outline for section 2

1 Introduction

2 Yield

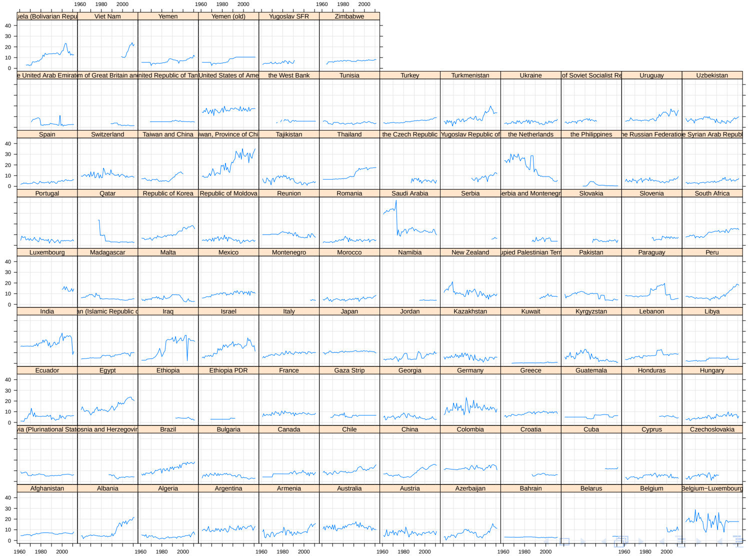
3 Production and Area Harvested

Linear mixed model

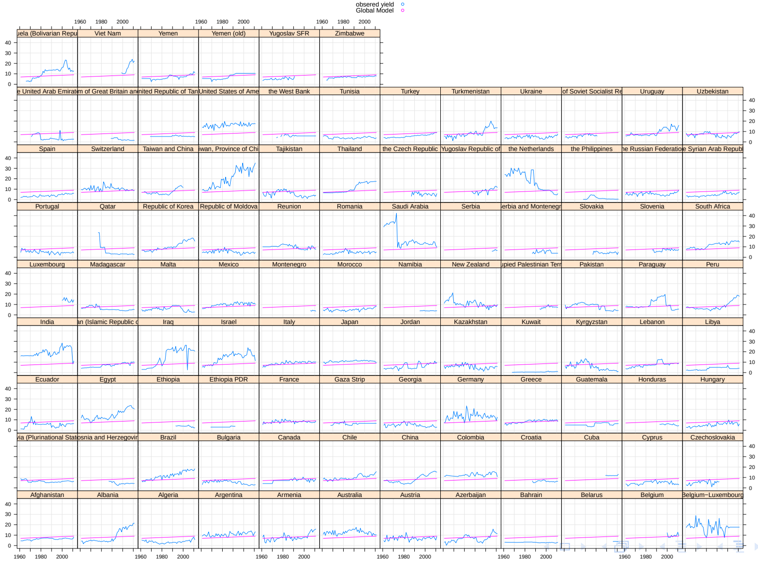
The model implemented in for the imputation of yield is the linear mixed model.

It is an extension of the simple linear regression, but enables the user to take advantage of any correlation structure exist.

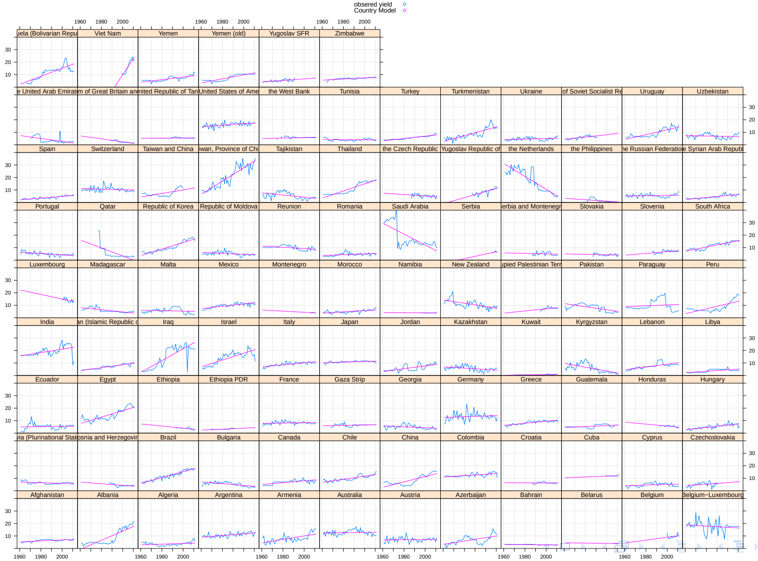
Yield of grape



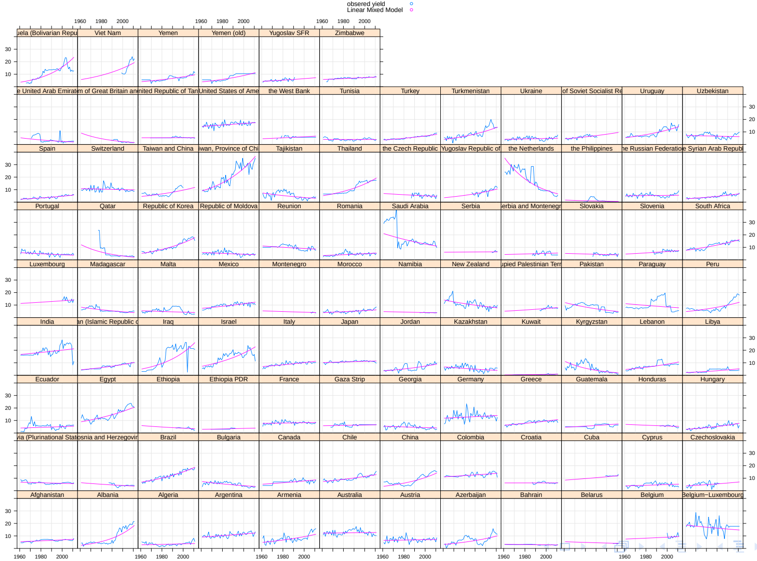
Global Model



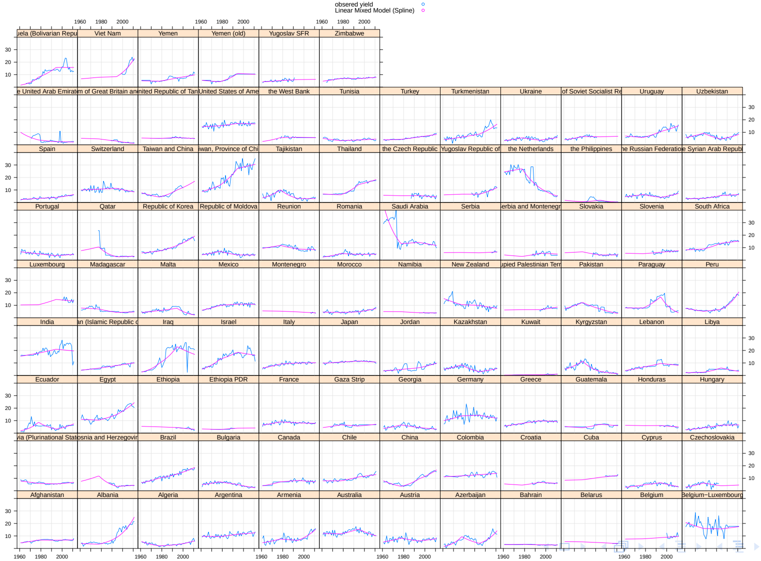
Country Model



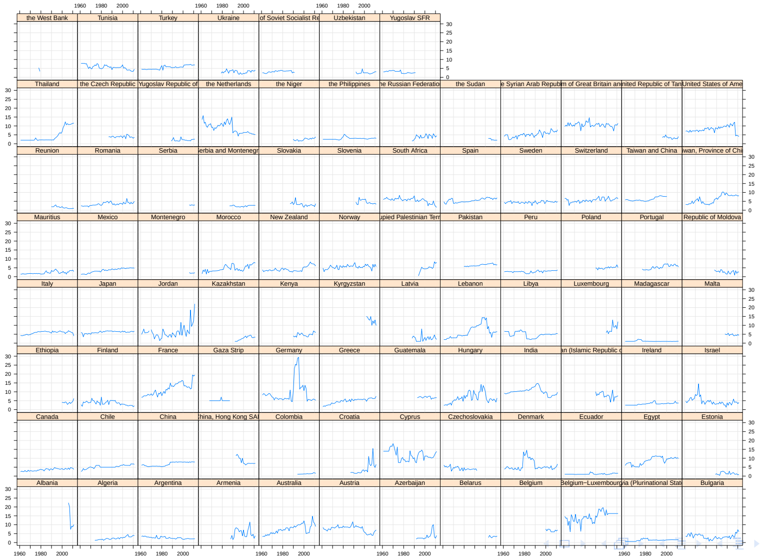
Linear Mixed Model



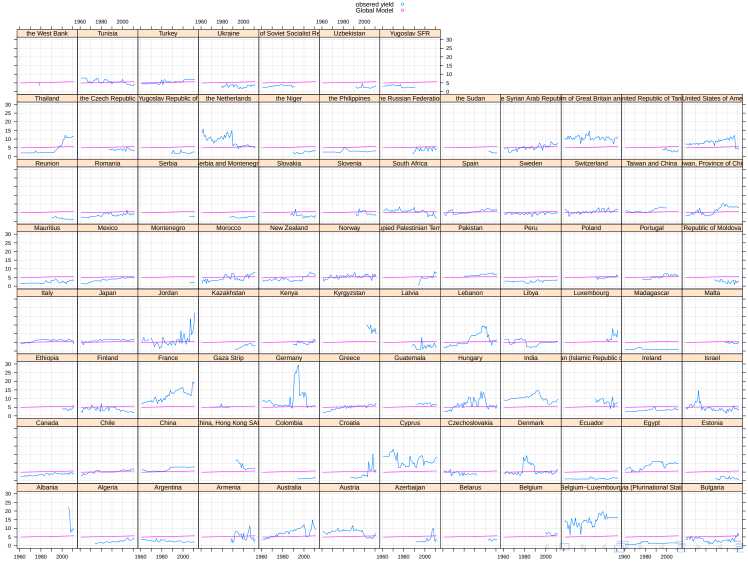
Linear Mixed Model with Splines



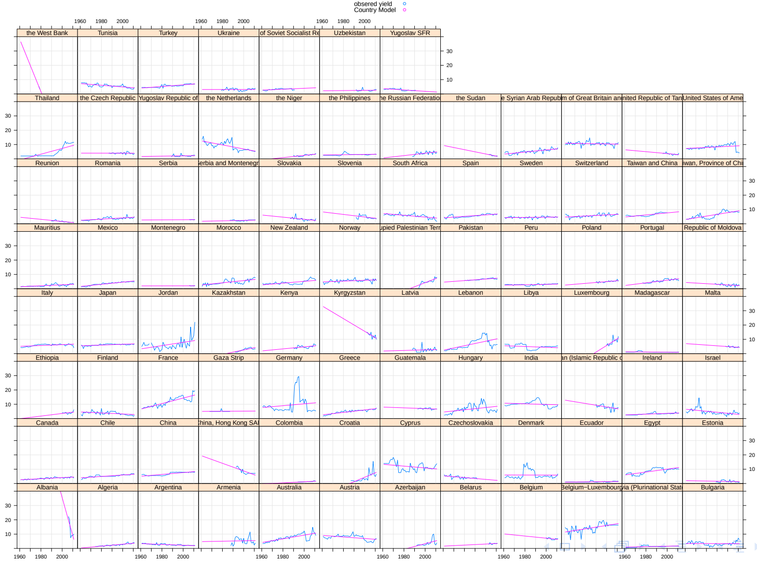
Yield of green peas



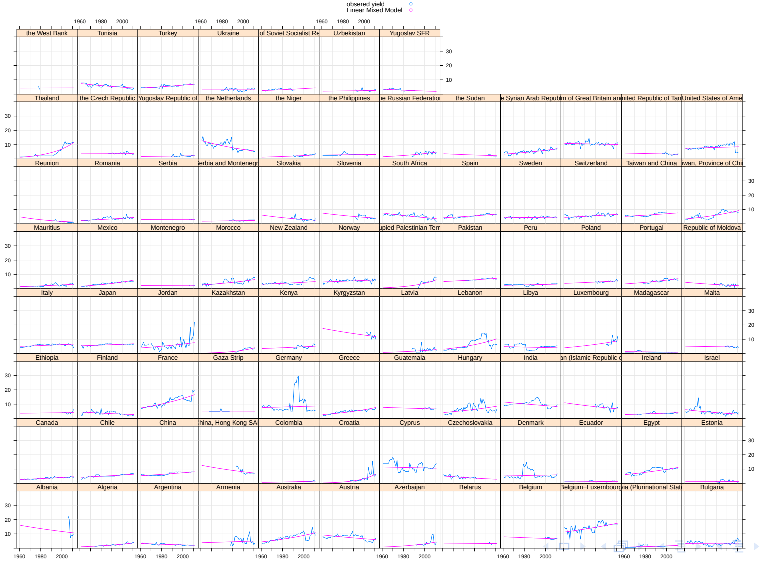
Global Model



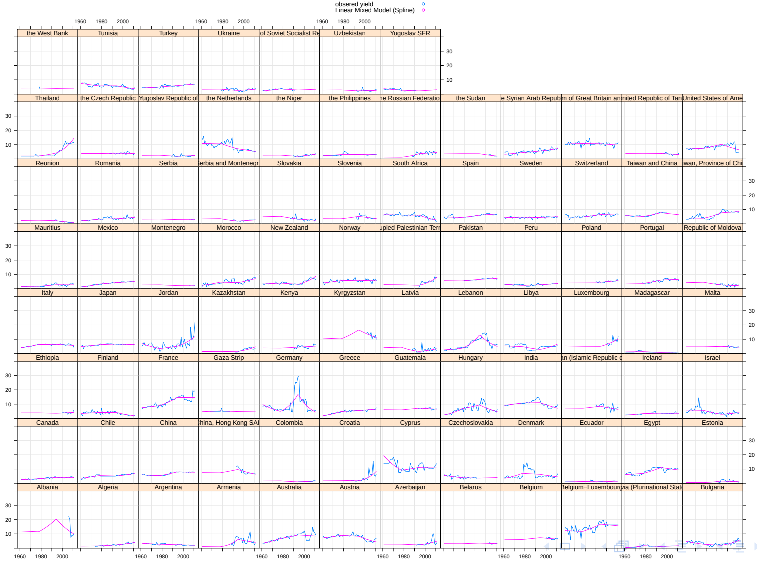
Country Model



Linear Mixed Model



Linear Mixed Model with Splines



There are several reasons why linear mixed model resolve many problems which are associated with fitting a global or country level model.

- It utlize cross-country information if it exists.
- It captures country specific pattern, but conforms to global constraints.
- This allow us to fit more flexible models without over-fitting.

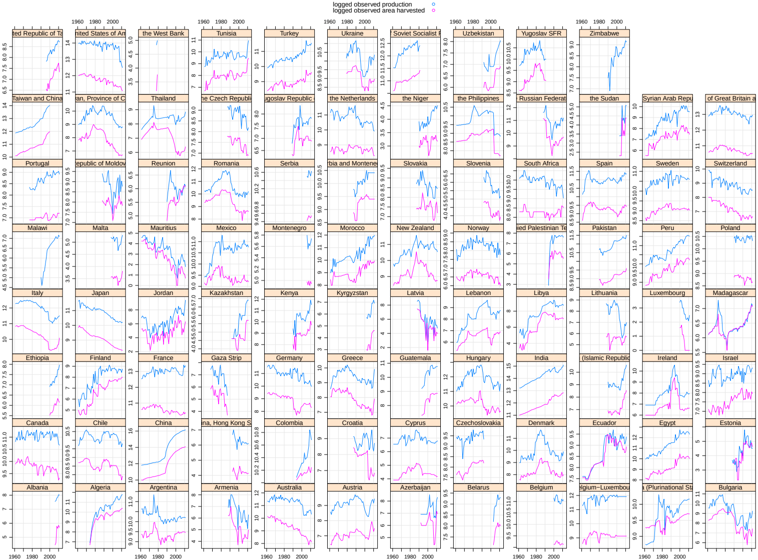
It makes a much more reasonable assumption to our data, countries are different but similar.

Outline for section 3

1. Introduction
2. Yield
3. Production and Area Harvested

After the imputation of the yield, we can balance the area harvested or production provided the counter-part exists.

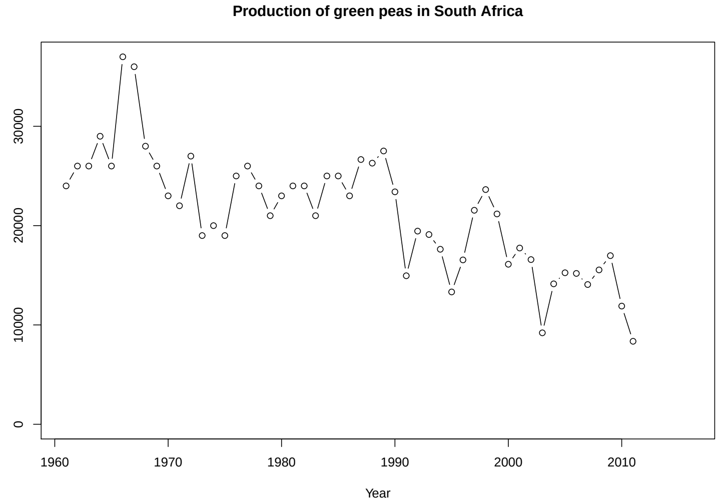
However, the problem is when both of them are missing.

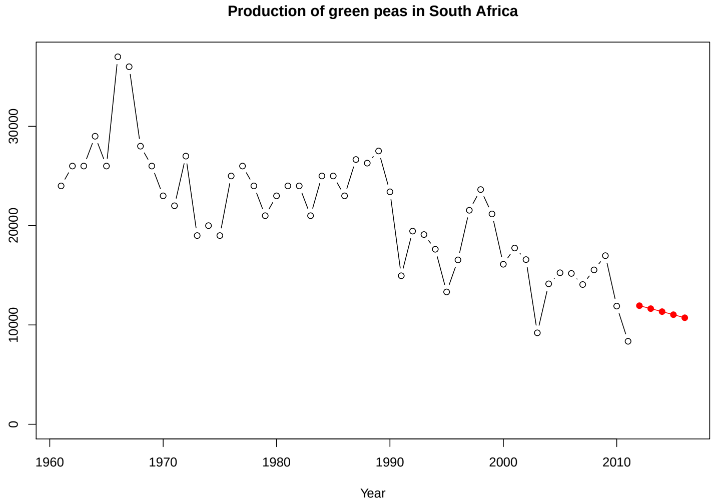

logged observed production
logged observed area harvested
1960 1980 2000
ted Republic of Ta
nited States of Am
the West Bank
Tunisia
Turkey
Ukraine
f Soviet Socialist F
Uzbekistan
Yugoslav SFR
Zimbabwe
Taiwan and China
an, Province of C
Thailand
he Czech Republi
ugoslav Republic
the Netherlands
the Niger
the Philippines
Russian Federat
the Sudan
Syrian Arab Repu
of Great Britain a
Portugal
epublic of Moldov
Reunion
Romania
Serbia
rbia and Montene
Slovakia
Slovenia
South Africa
Spain
Sweden
Switzerland
Malawi
Malta
Mauritius
Mexico
Montenegro
Morocco
New Zealand
Norway
ied Palestinian Te
Pakistan
Peru
Poland
Italy
Japan
Jordan
Kazakhstan
Kenya
Kyrgyzstan
Latvia
Lebanon
Libya
Lithuania
Luxembourg
Madagascar
Ethiopia
Finland
France
Gaza Strip
Germany
Greece
Guatemala
Hungary
India
(Islamic Republic
Ireland
Israel
Canada
Chile
China
ina, Hong Kong S
Colombia
Croatia
Cyprus
Czechoslovakia
Denmark
Ecuador
Egypt
Estonia
Albania
Algeria
Argentina
Armenia
Australia
Austria
Azerbaijan
Belarus
Belgium
elgium−Luxembou
a (Plurinational Sta
Bulgaria
1960 1980 2000

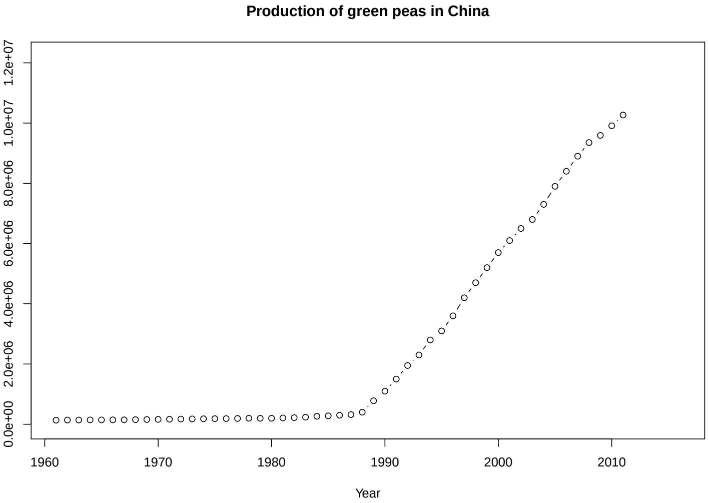
What is the solution?

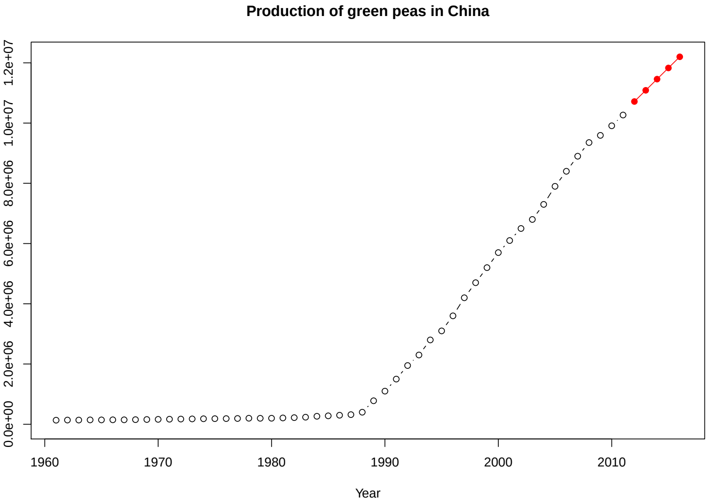
Oppose to yield, we have no clear cross-country information which we can utilize.

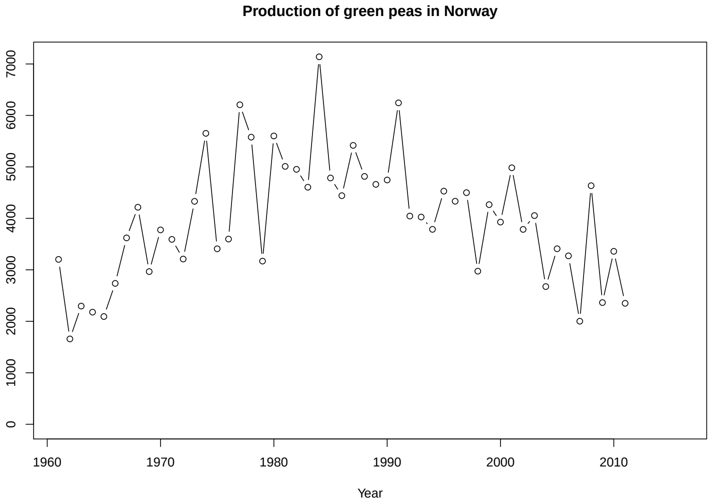
Furthermore, we actually have no idea about the trend, or even the shape of production!

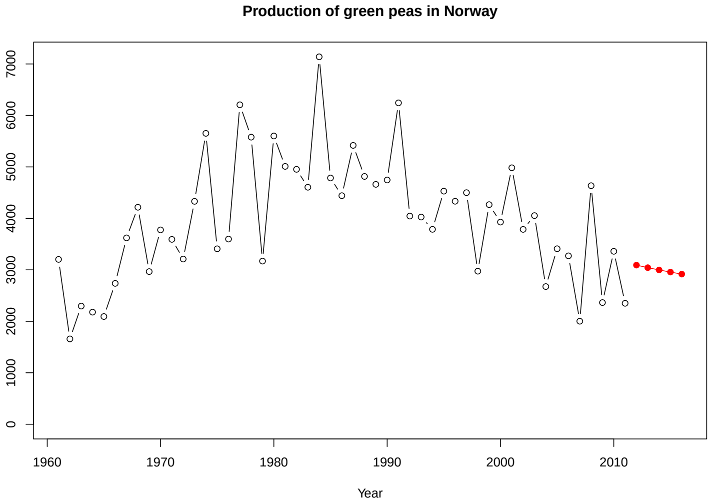

Production of green peas in South Africa
0
10000
20000
30000
1960
1970
1980
1990
2000
2010
Year

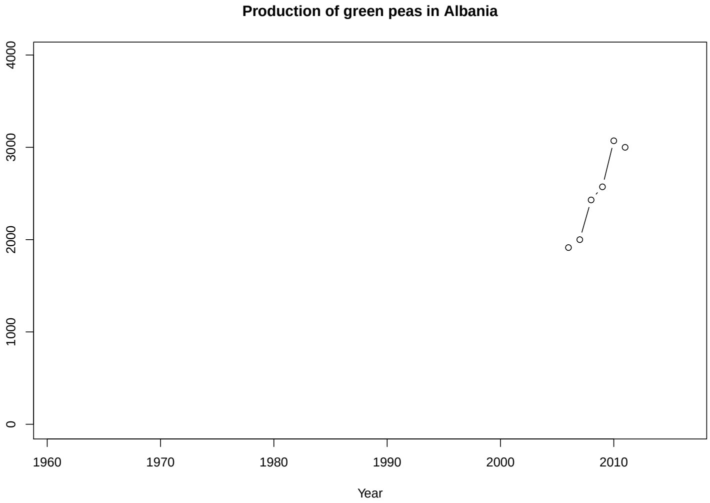

Production of green peas in South Africa
0
10000
20000
30000
1960
1970
1980
1990
2000
2010
Year

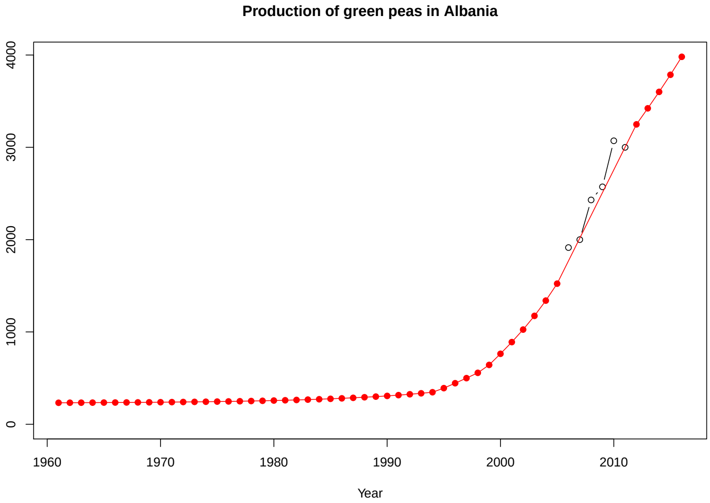

Production of green peas in China
0.0e+00
2.0e+06
4.0e+06
6.0e+06
8.0e+06
1.0e+07
1.2e+07
1960
1970
1980
1990
2000
2010
Year

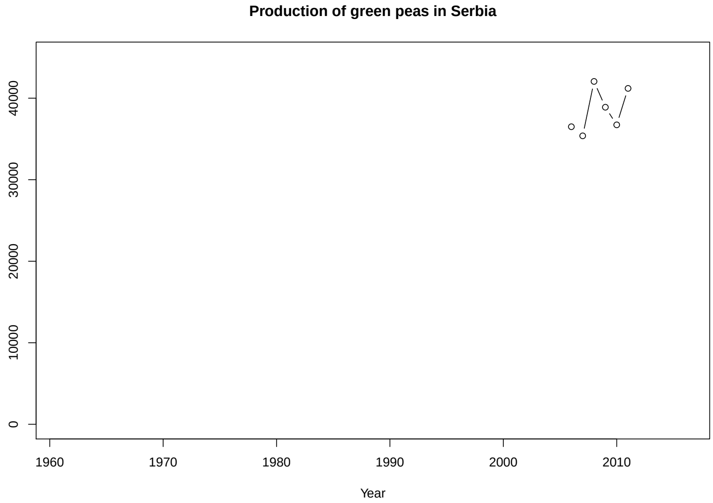

Production of green peas in China
0.0e+00
2.0e+06
4.0e+06
6.0e+06
8.0e+06
1.0e+07
1.2e+07
1960
1970
1980
1990
2000
2010
Year

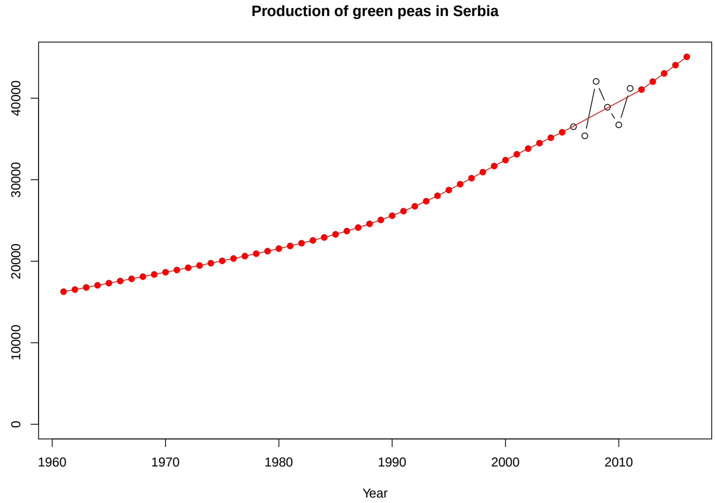

Production of green peas in Norway
0
1000
2000
3000
4000
5000
6000
7000
1960
1970
1980
1990
2000
2010
Year

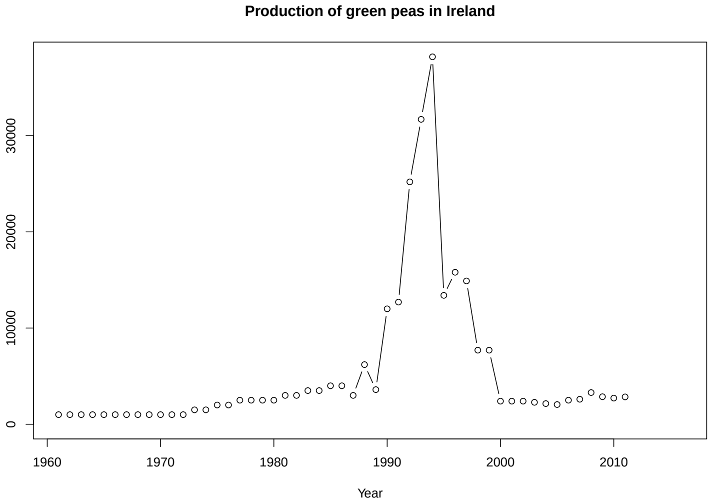

Production of green peas in Norway
0
1000
2000
3000
4000
5000
6000
7000
1960
1970
1980
1990
2000
2010
Year

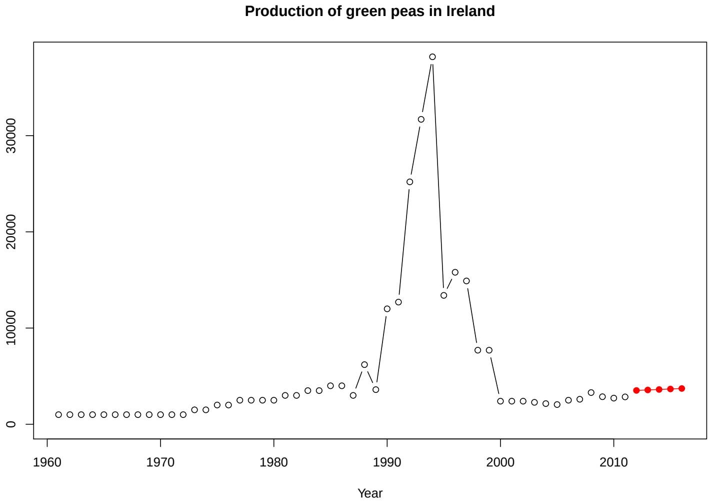

Production of green peas in Albania
0
1000
2000
3000
4000
1960
1970
1980
1990
2000
2010
Year

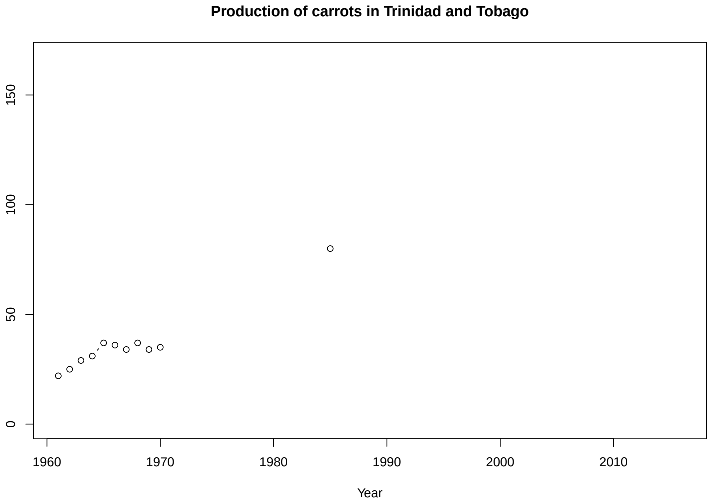

Production of green peas in Albania
0
1000
2000
3000
4000
1960
1970
1980
1990
2000
2010
Year

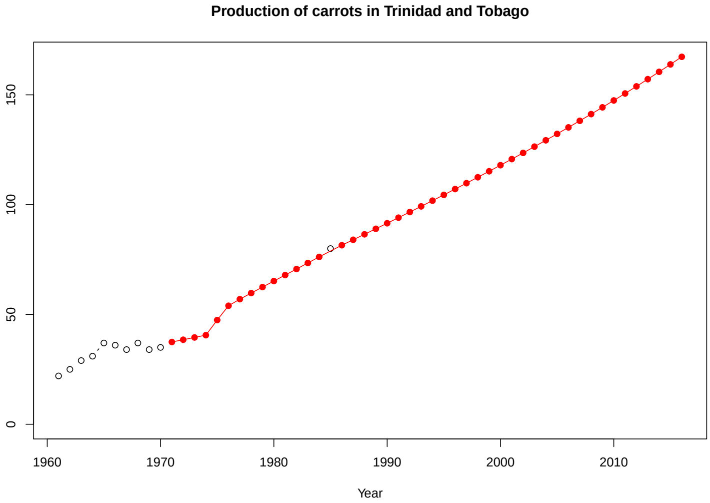

Production of green peas in Serbia
0
10000
20000
30000
40000
1960
1970
1980
1990
2000
2010
Year


Production of green peas in Serbia
0
10000
20000
30000
40000
1960
1970
1980
1990
2000
2010
Year


Production of green peas in Ireland
0
10000
20000
30000
1960
1970
1980
1990
2000
2010
Year


Production of green peas in Ireland
0
10000
20000
30000
1960
1970
1980
1990
2000
2010
Year


Production of carrots in Trinidad and Tobago
0
50
100
150
1960
1970
1980
1990
2000
2010
Year


Production of carrots in Trinidad and Tobago
0
50
100
150
1960
1970
1980
1990
2000
2010
Year

What is Ensemble Learning?

Ensemble learning in its simplest sense, is to build multiple models/learners and combine them to obtain the final model or prediction.

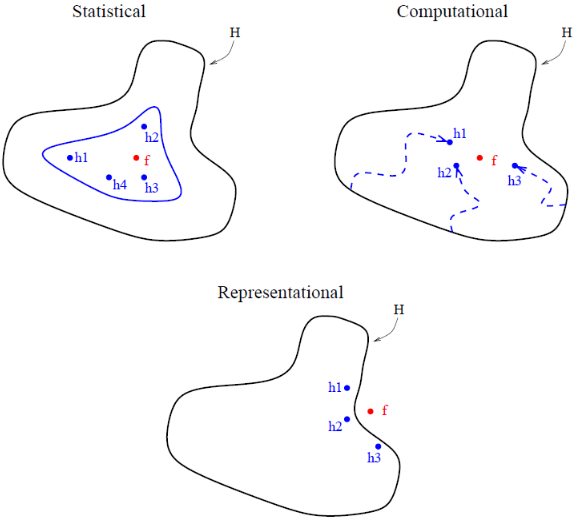
It consist of two components:

1. Building multiple models or learners.
2. Combine the models and predictions.

Why Ensemble Learning?

Ensemble as described by Dietterich (2000) can mitigate the following three issues.

Statistical:	Lack of data to identify an unique solution.
Computational:	Optimization
Representational:	Complex model



Statistical

Computational

Representational

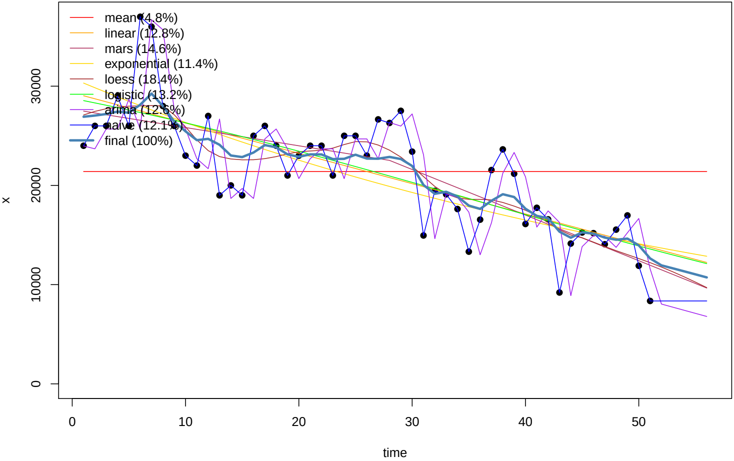
Implementation

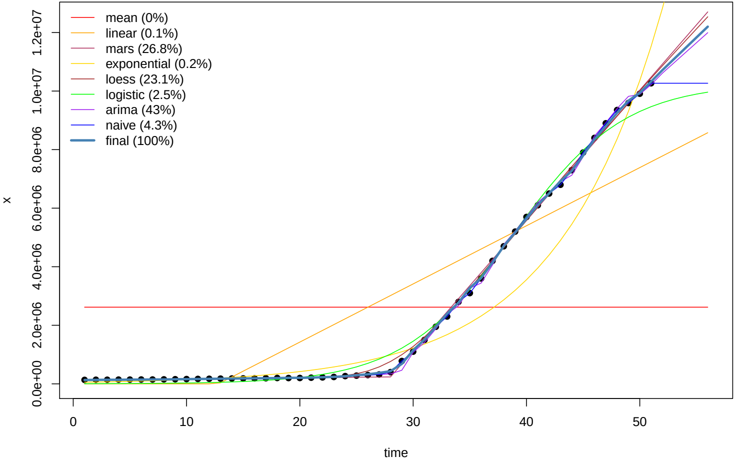
The details of the ensemble implemented is describe here:

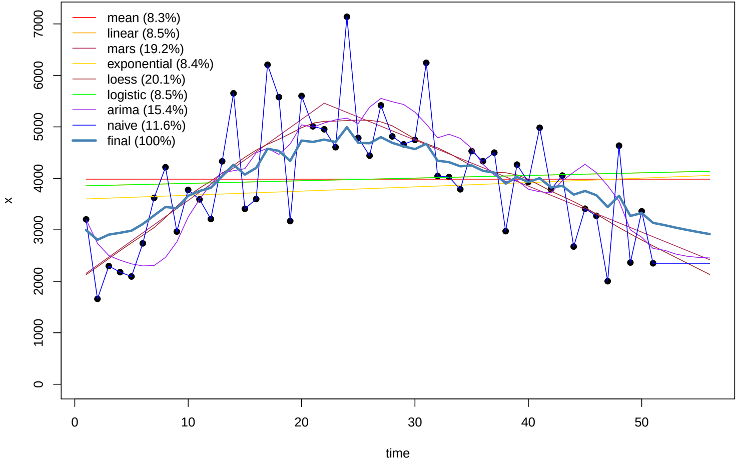
- **Base learners:**
  - mean
  - linear
  - MARS
  - exponential
  - locally smooth linear
  - logistic
  - ARIMA
  - naive
- **Combiner:** non-trainable algebraic combiner - Weighted sum rule

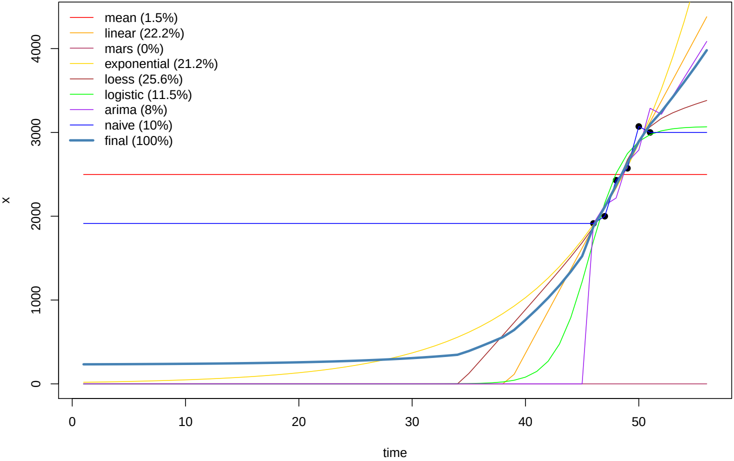
$$u_j(x) = \sum_{i=1}^{N} w_i d_{n,j}(x)$$

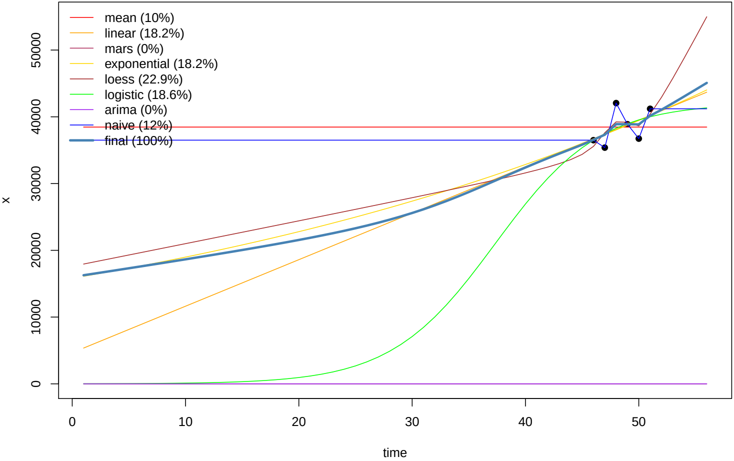
Where the weights depends on the fit on the available data.

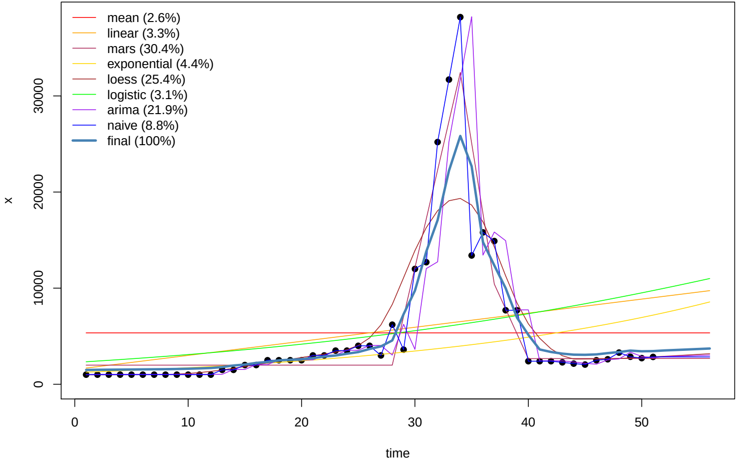

mean (4.8%)
linear (12.8%)
mars (14.6%)
exponential (11.4%)
loess (18.4%)
logistic (13.2%)
arima (12.6%)
naive (12.1%)
final (100%)
x
0
10000
20000
30000
time
0
10
20
30
40
50

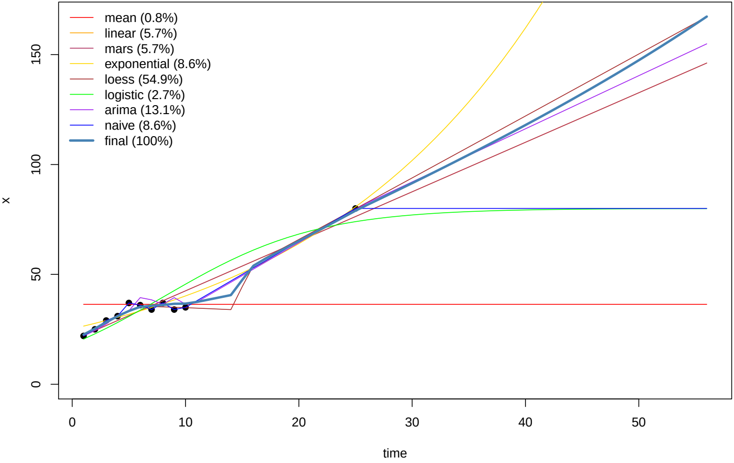

mean (0%)
linear (0.1%)
mars (26.8%)
exponential (0.2%)
loess (23.1%)
logistic (2.5%)
arima (43%)
naive (4.3%)
final (100%)
x
0.0e+00
2.0e+06
4.0e+06
6.0e+06
8.0e+06
1.0e+07
1.2e+07
0
10
20
30
40
50
time


mean (8.3%)
linear (8.5%)
mars (19.2%)
exponential (8.4%)
loess (20.1%)
logistic (8.5%)
arima (15.4%)
naive (11.6%)
final (100%)
x
0
1000
2000
3000
4000
5000
6000
7000
time
0
10
20
30
40
50


mean (1.5%)
linear (22.2%)
mars (0%)
exponential (21.2%)
loess (25.6%)
logistic (11.5%)
arima (8%)
naive (10%)
final (100%)
x
0
1000
2000
3000
4000
time
0
10
20
30
40
50


mean (10%)
linear (18.2%)
mars (0%)
exponential (18.2%)
loess (22.9%)
logistic (18.6%)
arima (0%)
naive (12%)
final (100%)
x
0
10000
20000
30000
40000
50000
time
0
10
20
30
40
50


mean (2.6%)
linear (3.3%)
mars (30.4%)
exponential (4.4%)
loess (25.4%)
logistic (3.1%)
arima (21.9%)
naive (8.8%)
final (100%)
x
0
10000
20000
30000
0
10
20
30
40
50
time


mean (0.8%)
linear (5.7%)
mars (5.7%)
exponential (8.6%)
loess (54.9%)
logistic (2.7%)
arima (13.1%)
naive (8.6%)
final (100%)
x
0
50
100
150
0
10
20
30
40
50
time

Practical benefits of adopting the ensemble model

From the examples, we can see that we have an extremely flexible model with almost no possibility of over-fitting since none of the model are itself complex.

We can continue to add in more models in which we consider appropriate.

The reason it is called ensemble learning is becaues it is a model that can learn. If a production that has been growing linearly in the past but suddenly at an exponential rate, the model will learn and shift the weight of the linear model to model which are more flexible.

Explanatory variables?

Why we don't use other explanatory variables:

- Can contain as many missing value or more than our production domain.
- Too many of them, hard to choose from the set. A simple set would contain precipitation and temperature, should we also include number of sunny days? and their interactions? What about CO_2/phosphorus concentraction? Number of bee hives? Fertilizer, pesticide consumption?
- Difficult to maintain and integrate into the system.