Document de travail : Methodologie d'imputation et de validation pour le domaine de production de FAOSTAT

Michael C. J. Kao
Food and Agriculture Organization
of the United Nations

Abstract

Ce Papier présente une nouvelle méthode d'imputation destinée au domaine de la production dans FAOSTAT. Cette méthode résout un nombre important de problèmes soulevés par l'approche actuelle, sa structure flexible permet d'incorporer de nouvelles informations et d'améliorer ses performances.

Nous examinons en premier lieu les facteurs déterminant des changements de production par produits, puis donnons un bref aperçu de la méthode actuelle et de ses limites. La nouvelle méthodologie est ensuite décrite, accompagnée d'une décomposition du modèle et de son explication.

Keywords: Imputation, Modèle linéaire mixte généralisé, Production Agricole, EM.

Avertissement

Ce document de travail ne représente pas les vues de la FAO. les points de vue exprimées dans ce document de travail sont celles de l'auteur et ne reflétent pas nÃl'cessairement celles de la politique de la FAO. Les Documents de travail décrivent les recherches en cours par l'auteur et visent à susciter commentaires et discussions.

1. Introduction

Les problèmes de données manquantes sont courants dans le domaine de la production agricole. Ils peuvent être dus a une absence de réponse de la part des entités pourvoyant les données ou une incapacité de celles-ci à obtenir les informations. Il est cependant de première importance, pour produire la balance alimentaire de pouvoir compter sur un domaine de production cohérent et le plus complet possible. Une imputation précise et fiable est donc un pré-requis essentiel.

Ce papier cherche à cerner et dépasser un certain nombre de limites de la méthodologie actuelle et à améliorer la précision de l'imputation en développant une nouvelle méthodologie.

La relation entre les variables du domaine de production peut être exprimée ainsi :

$$\mathrm{P}_t = \mathrm{A}_t \times \mathrm{Y}_t \tag{1}$$

Où P, A et Y représentent respectivement la production, la surface cultivée et le rendement, indexés par le temps t. Le rendement est inobservable et peut seulement être calculé quand la production et la surface sont disponibles. Pour certains produits la surface cultivable peut ne pas exister ou avoir une signification différente.

L'objectif de l'imputation est, en incorporant l'ensemble des informations fiables utilisables,

de fournir les meilleurs estimations de la quantite d'aliments disponible pour permettre le calcul de la balance alimentaire.

2. Contexte et revue de la méthodologie actuelle

Deux catégories de méthodologies ont été proposées par le passé pour évaluer les données manquantes dans le domaine de production. Les méthodologies appartenant à la première catégorie utilisent les séries historiques et appliquent des méthodes d'interpolation et de régression sur une tendance. Celles appartenant à la seconde catégorie basent l'imputation sur les taux de croissance des produits et/ou sur des agrégations par région. L'imputation est menée de manière indépendante à la fois sur la surface cultivée et sur la production, tandis que les rendements sont calculés de manière implicite.

Chacune de ces approches n'utilisent cependant qu'une dimension de l'information. De nombreuses améliorations peuvent être obtenues en combinant les différentes sources d'information et les méthodes citées plus haut.

De plus, ces méthodes ne permettent pas d'incorporer d'autres informations, comme les indices de végétation, de précipitations, ou de température qui peuvent apporter une information précieuse et aider a améliorer la précision de l'imputation.

Les résultats obtenus par les essais précédents indiquent que l'interpolation linéaire est une méthode stable et précise. Elle ne permet cependant pas d'utiliser des données transversales, ni d'extrapoler lorsque les points de connexion ne sont pas disponibles.

En conséquent, la méthode d'agrégation a été préférée car elle permet d'atteindre un taux de couverture élevé pour l'imputation, et semble extrêmement performante.

Dans un premier temps, cette méthode permet de calculer la croissance agrégée de la production et de la surface par produit et par région. Le taux de croissance est ensuite appliqué a la dernière valeur observée dans la série concernée. La formule est la suivante:

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

Où S réfère à l'ensemble des produits et pays appartenant aux groupes de produits et de la classification régionale concernée, après exclusion des données devant être imputées.

Par exemple, pour calculer la *croissance agrégée de la production céréalière* pour un pays dans le but d'imputer la production de blé, on additionne toute la production des produits appartenant au groupe de céréales d'un même pays en excluant le blé.

Pour imputer la production de blé à l'aide d'un (*indice régional de croissance agrégée*), les données de production du blé sont agrégées à l'intérieur du profile régional, à l'exception du pays concerné.

l'Imputation séffective donc de la manière suivante

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

Il y a un certain nombre de limites à cette méthodologie. Sa faiblesse principale vient du fait que la production et la surface sont estimées de manière indépendante. Des cas de divergence entre la production et la surface ont été observés, résultant en incohérences entre les tendances, ou en rendements bien trop élevés.

Ce problème prend sa source dans le calcul du taux de croissance agrégé.

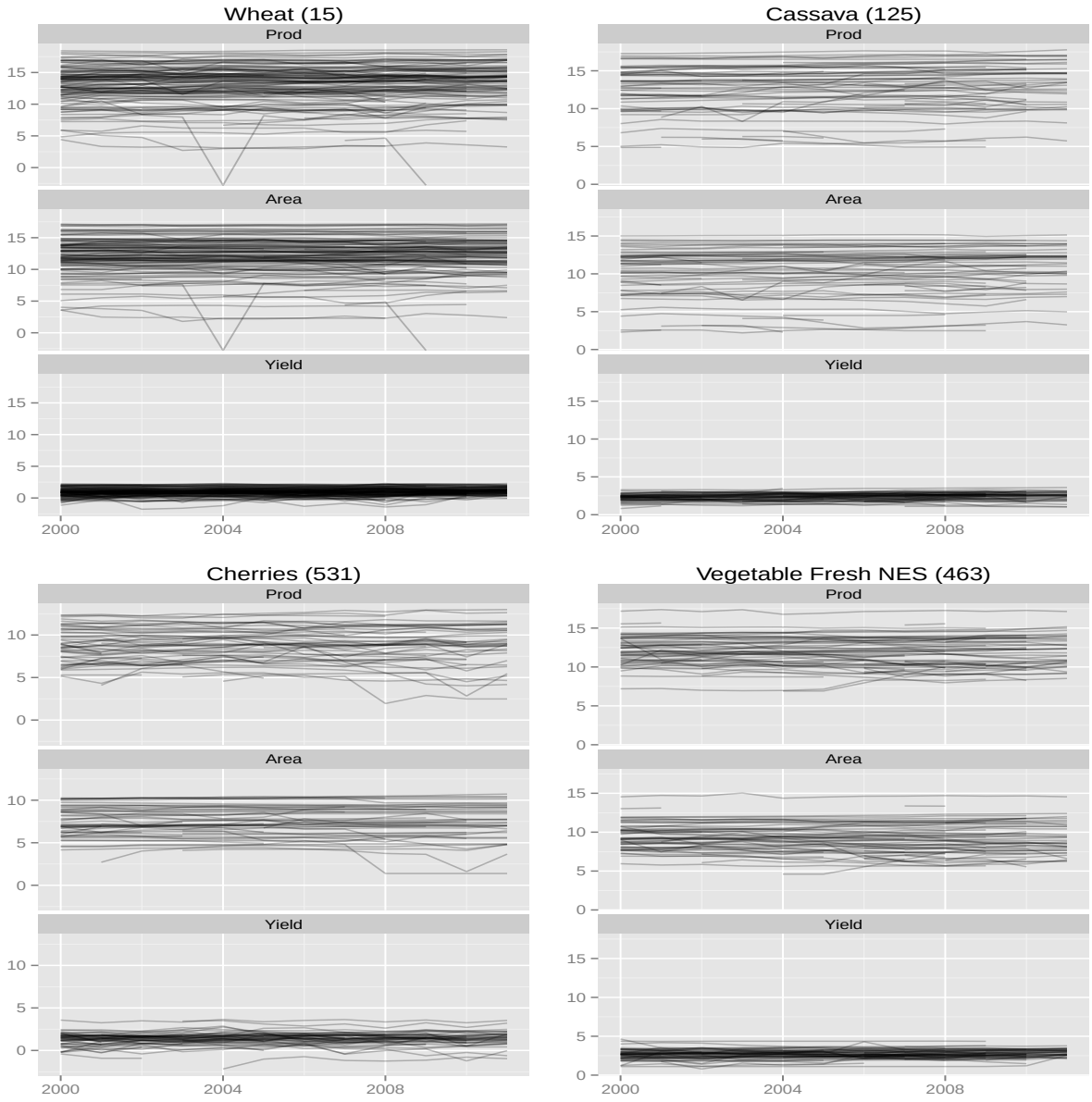
Du fait des données manquantes, le panier calculé peut ne pas être comparable au cours du temps, induisant ainsi des erreurs dans le calcul de la croissance de la production. De plus, les

paniers permettant de calculer les changements de production ou de surface cultivées peuvent être considérablement différents. Finalement, la méthodologie ne donne aucun aperçu des facteurs sous-jacents déterminant la production, qui sont pourtant nécessaires à une meilleure compréhension des phénomènes en jeux et donc à l'interprétation.

3. Première analyse de données

Avant qu'aucune modélisation ou analyse statistique ne soit faite, un aperçu des données est essentiel. Cette section est dédiée a l'exploration des données afin de comprendre la nature des séries et leurs déterminants. En premier lieu, nous explorerons la relation décrite par l'équation 1. Pour simplifier, nous avons appliqué aux données un logarithme, afin de transformer linéariser la relation.

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

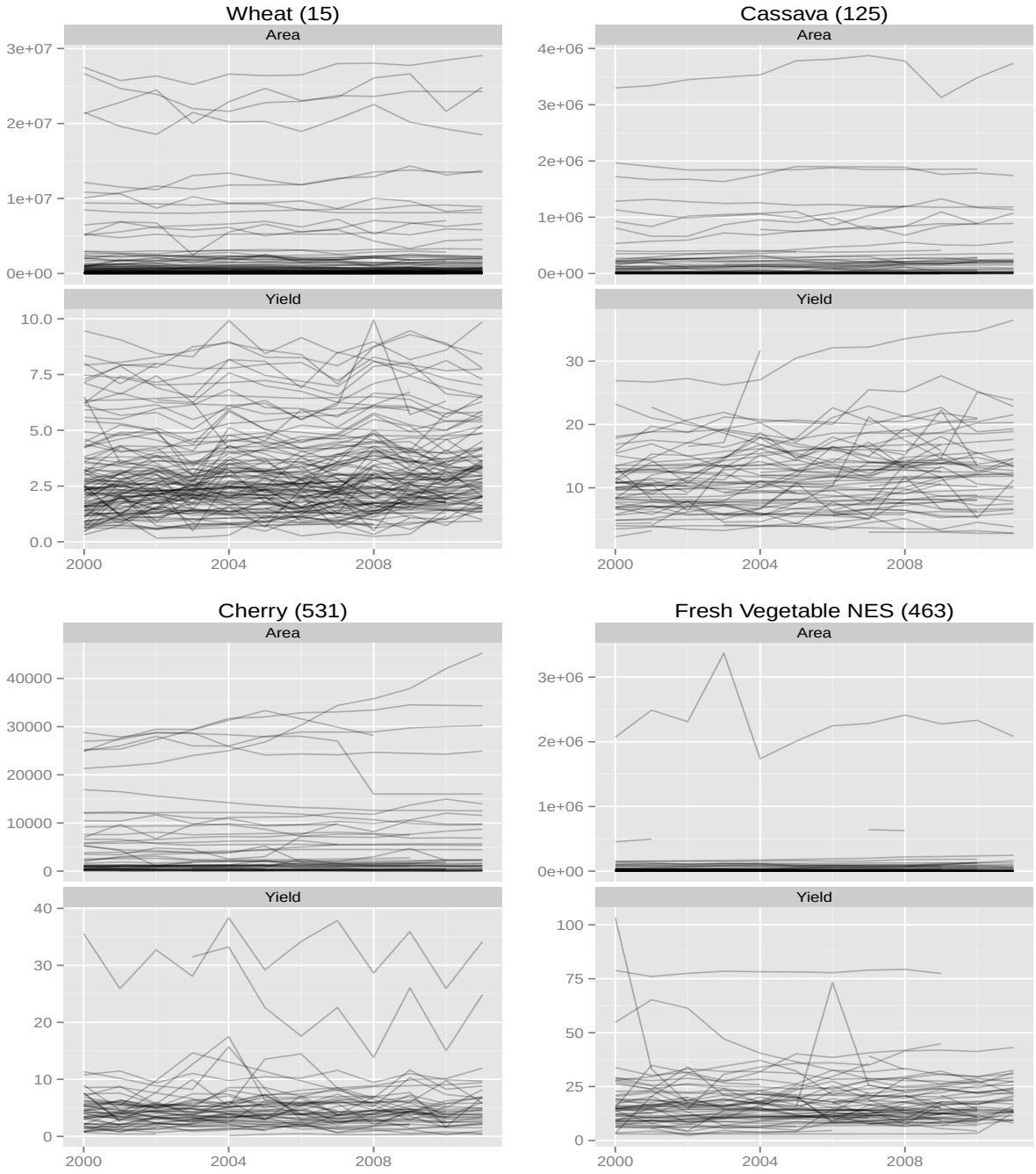


Sur les graphiques ci-dessus, les log de la production, surface et des récoltes d'un produit spécifique sont tracé par panel pour permettre la comparaison. Chaque ligne représente un

pays et la production est la somme de la surface et du rendement. Le premier aspect notable observé ici est que le niveau de production est principalement déterminé par le niveau de surface cultivée. Les chocs sur la production sont par ailleurs liés a des changements affectant la surface plus que les rendements. La surface cultivée est habituellement considérée comme stable et prévisible dans le temps, bien que vulnérable à des climatiques.

Le second aspect notable est que l'intervalle de variation du taux de rendement est petit en comparaison de celui de la surface. Ceci est en accord avec l'intuition qu'il existe des contraintes physique au rendement potentiel d'une récolte sur une surface donnée. Ces résultats ne varient pas selon les produits considérés.

Nous allons maintenant explorer plus en détail l'évolution du rendement et de la surface. Les graphiques ci-dessous représentent la surface et le rendement pour le même ensemble de produit, mais cette fois sans transformation des données.



Nous pouvons en premier lieu observer que les séries de la surface cultivée sont en général

plus stables et lisses que celles qui représentent le rendement. Le rendement fluctue d'une année sur l'autre tout en présentant une certaine corrélation, qui est plus durablement observé dans la série du blé. Ceci peut être expliqué par des facteurs sous-jacents, comme des facteurs climatiques, qui impacteraient les rendements de différents pays simultanément. Cependant cette caractéristique n'est pas observée dans la catégorie NES (non spécifiées ailleurs), ce qui suppose que l'impact de tels facteurs est fort au sein d'un type de production mais faible entre différentes productions.

Les données suggèrent que la tendance et le niveau de la production sont très largement déterminés par la surface cultivée, mais la variation d'une année à l'autre est en revanche déterminée par le rendement, qui peut être associé aux changements climatiques. L'analyse exploratoire des données nous éclaire sur la nature de la série temporelle. Elle soutient l'utilisation d'un modèle de décomposition de la variance qui attribuerait les fluctuations à la surface et aux rendements.

4. Méthodologie proposée

Afin d'éviter des problèmes d'identification, et de tenir compte de la corrélation des rendements entre différents pays, nous proposons d'imputer les rendements et la surface, et non la production et la surface. Le second avantage de cette approche et qu'associée a un system de validation, elle garantie que les séries ne divergent pas comme elles le font dans l'approche actuelle.

4.1. Imputation pour le rendement

Le modèle proposé pour estimer le rendement est un modèle linéaire mixte. L'usage de ce modèle permet d'incorporer à la fois l'information transversale et l'information historique. D'autres indicateurs, comme l'indice de végétation, la concentration en CO_2 peuvent aussi être l'indice testés et incorporés s´ils améliorent la prévision.

La forme générale du modèle peut être spécifiée de la manière suivante :

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

Où la composante fixe $\mathbf{X_i}\boldsymbol{\beta}$ désigne le niveau régional et la tendance , tandis que la composante aléatoire $\mathbf{Z_i}\mathbf{b_i}$ capture la variation spécifique du pays autour du niveau régional. Plus spécifiquement, le modèle proposé pour la production dans FAOSTAT est le suivant :

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

Où Y désigne le rendement, $\bar{Y}$ désigne le rendement moyen du groupe, i indique le pays, j le groupe régional, et t le temps. La moyenne du groupe est calculée de la manière suivante :

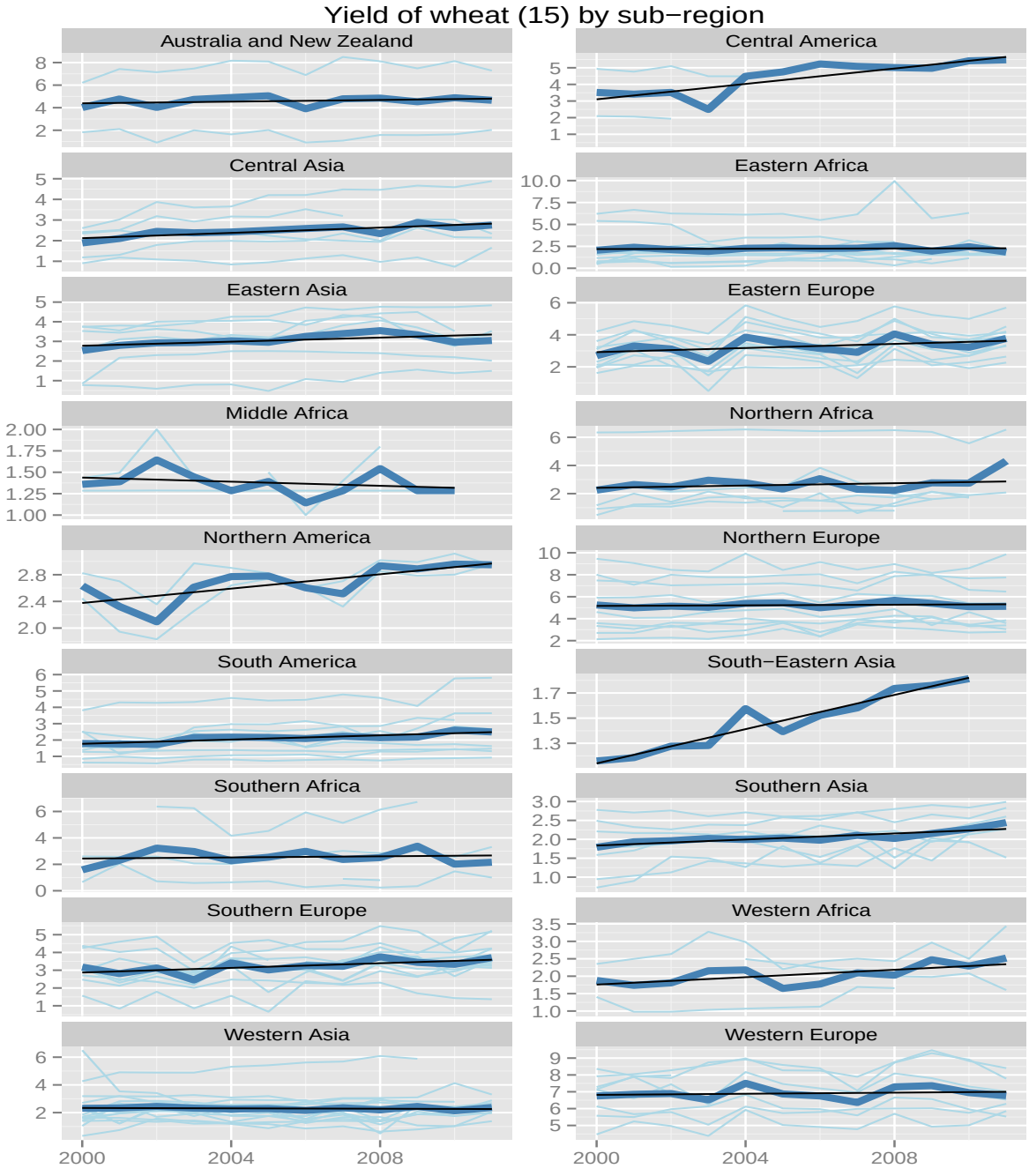
$$\bar{Y}_{j,t} = \frac{1}{N_i}\sum_{i \in j} \hat{Y}_{i,t} \tag{7}$$

Cependant, comme le rendement moyen du groupe est seulement partiellement observé compte tenu des données manquantes, le rendement moyen est estimé grâce a l'algorithme EM (maximisation de l'espérance).

L'estimation du rendement est basée sur le niveau spécifique du pays et sur la tendance historique régionale, tout en tenant compte de la corrélation entre les pays et des variations régionales.

Contrairement á la méthodologie précédemment utilisée, oú la variation était appliquée entièrement, la méthodologie proposée mesure le degré de la relation entre la série individuelle et les variations régionales pour estimer l'effet aléatoire du pays. Comme à la fois les données historiques et transversales sont utilisées, les données estimées présentent des caractéristiques stables tout en reflétant les changements climatiques.

Afin de mieux comprendre la méthodologie, nous présentons ci-dessous le niveau régional (ligne noire), et le rendement moyen des pays du groupe (ligne bleue foncée) sur le même graphique. Le modèle attribue à chaque série une tendance et un niveau régional (représenté par la ligne noire), et modélise la corrélation avec la série du rendement moyen régional représentée par ligne bleue foncée.



4.2. Estimation pour la surface cultivée

Après avoir estimé le rendement, calculé la surface cultivée et la production quand cela était possible, nous estimons la surface à l'aide d'une interpolation linéaire et répliquons la dernière observation quand la production et la surface ne sont pas disponibles.

D'après de précédentes recherches et nos études actuelles, l'interpolation semble appropriée car la surface cultivée est caractérisée par des séries extrêmement stables autour de leur tendance.

En dépit de cette stabilité, les chocs sont parfois observés dans les séries de la surface cultivée. Cependant, sans une compréhension plus grande de la nature et de la source de ces chocs, appliquer aveuglement le modèle n'améliorerait pas la performance de l'estimation. Nous avons choisi à ce stade de répliquer les dernières données disponibles lorsque l'interpolation linéaire n'est pas applicable. L'avantage principal de cette approche est que si la production cesse, les chiffres de la production et la surface s'établissant a zéro l'année précédente, nous n'imputerons pas une donnée positive.

$$\hat{A}_t = A_{t_a} + (t - a) \times \frac{A_{t_b} - A_{t_a}}{t_b - t_a} \tag{8}$$

Nous continuons néanmoins à explorer les données et à étudier des méthodes plus efficaces qui pourraient être appliquées à l'estimation des données manquantes pour la surface.

Pour les données manquantes que nous ne pouvons imputer à l'aide de l'interpolation linéaire, nous remplaçons par la dernière valeur disponible.

$$\hat{A}_t = A_{t_{nn}} \tag{9}$$

5. Conclusion et améliorations futures

Le but de ce papier est de réviser la méthodologie actuelle et de produire une méthodologie plus pertinente et plus performante.

Le modèle proposé permet de résoudre des problèmes posés par les séries de de production et de surface divergentes ou les biais dans le calcul croissance résultant des données manquantes. De plus, la proposition offre la possibilité d'incorporer l'information adéquate tout en maintenant un cadre souple permettant de tenir compte des informations supplémentaires.

Les équipes techniques continuent de collaborer afin d'améliorer le modèle et de mieux comprendre les données. Un modèle espace-état pourrait être un bon candidat à cette méthodologie car il permettrait à la production, l'espace et le rendement d'être imputés simultanément.

Remerciements

Ce travail a été superviseé par Adam Prakash, avec l'aide du Nicolas Sakoff, Onno Hoffmeister and Hansdeep Khaira essentielle pour le développment de la méthodologie. L' auteur voutrait aussi remercier les membres de l'équipe qui ont participé aux discussions. Nous remercions également Cécile Fanton et Franck Cachia pour la traduction en français.

Annexe 1: Classification Géographie

La classification géographique suit la classification UNSD M49 http://unstats.un.org/unsd/methods/m49/m49regin.htm. La définition est aussi disponible dans le `FAOregionProfile` du package R **FAOSTAT**.

Annexe 2: Code

Code et les données sont disponibles dans le fichier github https://github.com/mkao006/Imputation.

Algorithm 1: EM-Algorithm for Imputation

Initialization;
  $\hat{Y}_{i,t} \leftarrow f(Y_{i,t})$;
  $\mathcal{L}_{\text{old}} = -\infty$;
  $\mathcal{E}$ = 1e-6;
  n.iter = 1000;
begin
  for *i=1* **to** *n.iter* **do**
    E-step: Compute the expected group average yield;
      $\bar{Y}_{j,t} \leftarrow 1/N \sum_{i \in j} \hat{Y}_i$;
    M-step: Fit the Linear Mix Model in 6;
    if $\mathcal{L}_{new} - \mathcal{L}_{old} \geq \mathcal{E}$ **then**
      $\hat{Y}_{i,t} \leftarrow \hat{\beta}_{0j} + \hat{\beta}_{1j}t + \hat{b}_{0i} + \hat{b}_{1i}t + \hat{b}_{2j}\bar{Y}_{j,t}$;
      $\mathcal{L}_{\text{old}} \leftarrow \mathcal{L}_{\text{new}}$;
    end
    else
      break
    end
  end
end

Algorithm 2: Imputation Process

Data: Production (element code = 51) and Harvested area (element code = 31) data
Result: Imputation

Missing values are denoted $\emptyset$;

Initialization;
begin
  if $A_t = 0 \wedge P_t \neq 0$ **then**
    $A_t \leftarrow \emptyset$;
  end
  if $P_t = 0 \wedge A_t \neq 0$ **then**
    $P_t \leftarrow \emptyset$;
  end
end

Start imputation;
begin
  forall the *commodities* **do**
    (1) Compute the implied yield;
      $Y_{i,t} \leftarrow P_{i,t} / A_{i,t}$;
    (2) Impute the missing yield with the imputation algorithm 1;
    forall the *imputed yield* $\hat{Y}_{i,t}$ **do**
      if $A_t = \emptyset \wedge P_t \neq \emptyset$ **then**
        $\hat{A}_{i,t} \leftarrow P_{i,t} / \hat{Y}_{i,t}$;
      end
      if $P_t = \emptyset \wedge A_t \neq \emptyset$ **then**
        $\hat{P}_{i,t} \leftarrow A_{i,t} \times \hat{Y}_{i,t}$;
      end
    end
    (4) Impute area ($A_{i,t}$) with equation 8 then 9;
    forall the *imputed area* $\hat{A}_{i,t}$ **do**
      if $\hat{Y}_{i,t} \neq \emptyset$ **then**
        $\hat{P}_{i,t} \leftarrow \hat{A}_{i,t} \times \hat{Y}_{i,t}$;
      end
    end
  end
end

Affiliation:

Michael. C. J. Kao
Economics and Social Statistics Division (ESS)
Economic and Social Development Department (ES)
Food and Agriculture Organization of the United Nations (FAO
Viale delle Terme di Caracalla 00153 Rome, Italy
E-mail: michael.kao@fao.org
URL: https://github.com/mkao006/Imputation