Conversion des données CASD

Loïc Poullain, Augustin Wenger

November 3, 2020

Abstract

Ce document présente les problématiques liées à l'utilisation des données fiscales, à leur pré-traitement et à leur calcul. Il prend comme exemple le calcul de l'impôt sur le revenu (IR) mais peut se généraliser à d'autres sujets.

1 Introduction

Grâce au *centre d'accès sécurisé aux données (CASD)*, nous avons accès à l'ensemble des déclarations fiscales des foyers fiscaux français. Ces données ne peuvent toutefois pas être utilisées en direct dans nos produits.

La première contrainte est imposée par le respect du secret statistique. Les utilisateurs, par leurs simulations, ne doivent pas être en mesure d'identifier une personne d'une quelconque manière.

La deuxième est due à des objectifs de performance, les simulations devant durer une minute toute au plus.

Pour pouvoir utiliser ces données et respecter ces contraintes, nous devons les convertir en un autre ensemble, plus petit, qui sera utilisé lors des simulations.

2 Formulation formelle du problème

Soit $E \subset \mathcal{F}(\mathbb{R}^n, \mathbb{R}^+)$ l'ensemble des modes de calcul (appelés aussi *réformes*) possibles de l'impôt sur le revenu qui, aux données fiscales d'un foyer fiscal,

associe le montant de l'impôt de ce même foyer.

Quelques exemples de fonction f:

- application d'un simple barème progressif ;
- application d'un barème proportionnel en fonction du nombre de parts ;
- etc.

Un exemple d'ensemble de départ $\mathbb{R}^n$:

- revenu des activités salariés et composition du foyer (veuf, enfants, célibataire, etc) ($n = 2$).

Nous pouvons agir sur n et la taille de E (degré de liberté des réformes simulées).

Soit $\theta : E \longrightarrow \mathbb{R}^+$ une fonction qui, à un mode de calcul de l'IR $f \in E$, associe les recettes totales perçues par l'Etat pour cet impôt [1].

En notant $\{R_p \in \mathbb{R}^n / p \in [\![1, nombredefoyersfiscaux = P]\!]\}$ l'ensemble des données fiscales de tous les foyers de France [2] , il vient :

$$\theta = \theta_P : f \longmapsto \sum_{p=1}^{P} f(R_p)$$

Lors du calcul de $\theta(f)$ pour $f \in E$, le simulateur n'a pas accès à $\{R_p\}$ mais à un ensemble plus petit $\{R_j^* / j \leq J \leq P\}$ qui respecte le secret statistique et que nous devons construire.

On recherche donc un ensemble $\{R_j^*\}$ et une fonction $\theta^* = \theta_J$ de la forme $\theta^*(f) = g(f, \{R_j^*\})$ tel que $\forall f \in E, |\theta(f) - \theta^*(f)|$ soit minime.

La fonction θ^* doit être rapide à calculer (1 minute maximum).

[1] En pratique, θ est à valeurs dans un espace multidimensionnel. Voir section Généralisation du problème.

[2] Dans les faits, le fichier exhaustif des déclarations contenant les feuilles d'impôt d'environ 38 millions de foyers fiscaux, on utilise $P \approx 38M$.

3 Résolution numérique et complexité

Dans le simulateur de *Leximpact IR*, les restrictions actuelles sont les suivantes :

- E décrit l'ensemble des réformes *paramétriques* de l'IR. Seules les valeurs numériques du texte de loi sont éditables.
- $n = 2$ et les données fiscales $\mathbb{R}^n = \mathbb{R} \times \prod_{i=1}^{n'} F_i$ (où les F_i sont des sous-ensembles de $\mathbb{R}$ ou $\mathbb{N}$) décrivent les révenus d'activité et la composition du foyer (situation maritale, nombre d'enfants à charge, caractéristiques des membres du foyer comme l'invalidité, etc).

On pose arbitrairement :

$$\theta^* : f \longmapsto \sum_{p=1}^{J=?} f(R_j^*) \cdot w_j$$

On recherche donc J et $\{R_j^*, w_j\}$ tels que $\forall f \in E, |\theta(f) - \theta^*(f)|$ soit minime.

Actuellement en pratique :

- Le J employé correspond au nombre de foyers fiscaux présents dans le fichier ERFS-FPR, qui nous permet d'utiliser ces données d'enquête comme base. Cette enquête contient les données de 50000 foyers fiscaux pondérés et représentatifs, ce qui représente une quantité permettant le calcul des effets d'une réforme dans openfisca en environ une minute, délai acceptable.
- Les fonctions f utilisées ne sont pas des fonctions calculant l'impôt sur le revenu mais des fonctions dont l'agrégation sur la population française est disponible publiquement (ou dans les données exhaustives) : "la part des foyers fiscaux ayant un revenu fiscal de référence supérieur à x" pour une trentaine de valeurs de x.
- Après la calibration des revenus apparaissant dans l'enquête ERFS-FPR pour représenter la véritable distribution des RFR par foyer fiscal, une transformation affine par morceaux des résultats est effectuée pour prendre en compte des éléments qui n'apparaissent pas dans les

données, principalement les crédits d'impôts. Cette transformation, qui s'effectue sur le serveur au moment du calcul, permet de faire correspondre les résultats obtenus par nos calculs avec le montant global d'impôts sur le revenu récolté par l'Etat.

4 Généralisation du problème

Ce problème se généralise à d'autres sujets. Par exemple, pour le calcul de la CVAE à l'échelle du pays, on a $\mathbb{R}^n = \mathbb{R}^2 = \{(CA, VA)\}$ où CA est le chiffre d'affaires et VA la valeur ajoutée.

Plus généralement, en pratique, les fonctions θ et θ^* sont à valeurs dans $\mathbb{R}^m$. Dans le cas de l'IR, par exemple, les valeurs de sortie sont le budget de l'Etat mais aussi le nombre de personnes nouvellement imposables ou nouvellement exonérés, etc.

De la même manière, pour la CVAE, on pourrait, en sortie, classer les CVAE payés en fonction du secteur d'activité, du type d'entreprises, etc.

5 Pistes

Plusieurs axes de recherche pour pouvoir supporter des n plus grands tout en ayant un θ^* proche de θ et qui se calcule en moins d'une minute:

- augmenter J et la puissance de calcul du serveur ;
- affiner la définition de E pour réduire la taille de l'ensemble et jouer sur la structure de f ;
- utiliser d'autres structures de θ^* ;
- trouver d'autres méthodes de calcul (efficaces notamment en complexité en temps) pour résoudre $\|\theta(f) - \theta^*(f)\|$ très petit.