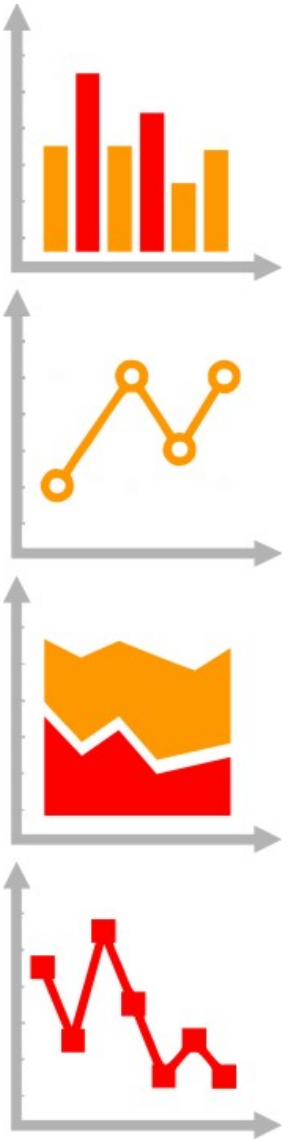
Données de caisse et webscraping

De nouvelles sources pour mesurer les prix à la consommation



Insee
Mesurer pour comprendre

L'indice des prix à la consommation aujourd'hui

- **L'indice des prix à la consommation mesure :**
  - l'évolution des prix à la consommation à qualité constante
  - en suivant un panier fixe de produits
  - renouvelé annuellement pour être représentatif de la consommation des ménages
  - en effectuant des ajustements qualité quand les produits disparaissent en cours d'année et qu'on les remplace

- **Pour calculer l'indice des prix à la consommation, on utilise chaque mois :**
  - 200 000 relevés effectués par des enquêteurs de l'Insee dans 30 000 points de vente
  - 190 000 prix collectés centralement : relevés sur internet, tarifs, bases de données administratives

De nouvelles sources de données...

- **Les données de transaction :**
  - Des données enregistrées au moment de la transaction avec enregistrement du prix, des quantités et un identifiant précis des produits
  - Exemples : les données de caisse des hyper et supermarchés (2020), les données des pharmacies pour les médicaments
- **La collecte automatisée de prix sur internet ou webscraping**
  - Essentiellement pour des services-formulaires mais dans d'autres pays également pour des biens
  - En mimant les clics des internautes ou en interrogeant des API
  - Exemples : le transport aérien, maritime, ferroviaire (à venir)

Données de transaction : les avantages

- **Une connaissance sur les quantités consommées**
  - Possibilité d'avoir des pondérations
  - Prise en compte des produits les plus représentatifs
- **L'exhaustivité des ventes sur un champ :**
  - Existence d'une base de sondage
  - Une plus grande précision pour des indices sur des segments particuliers, pour des indices géographiques
  - De nouvelles méthodes pour mesurer les ajustements qualité
- **L'économie de la collecte par les enquêteurs**

Les nouvelles questions que posent les données de transaction

L'accessibilité des données de transaction

- La loi numérique : de nouvelles possibilités...
- ...mais qui ne permettent pas de faire l'économie de contacts et d'échanges avec les entreprises...
- … et supposent que les données privées soient facilement extraites par les entreprises

Le volume de données à traiter

- Des données à traiter avec des architectures informatiques particulières
- Des traitements à automatiser : identifier les produits, les classer dans une nomenclature, les remplacer...

La finalité non statistique des données privées

- Des données qui s'appuient sur un système de gestion de l'entreprise (promotions, retour de produits…) sans toutefois que cela ne pose trop de problèmes pour les données de caisse des grandes enseignes.
- Des données que l'on n'observe que lorsqu'il y a transaction : quid des produits vendus à un rythme peu fréquent ?

Le webscraping : les avantages

- **L'accessibilité de l'information en apparence**
- **La possibilité de mimer le comportement du consommateur**
  - Être représentatif de la consommation sur internet
  - Pouvoir rendre compte des nouvelles politiques de prix : notamment yield management et classe d'antériorité
- **La possibilité d'augmenter sans coût marginal le volume de prix collectés… une fois que le robot est développé**

Les nouvelles questions avec le webscraping

- **La question de l'accessibilité**
  - La protection juridique des données librement accessibles comme les sites internet
  - Des difficultés techniques : robots bloqués, modifications de la structure des sites…
  - La transparence de la collecte par rapport à l'enquêté

- **Des problèmes similaires à la collecte actuelle dans les points de ventes physiques**
  - Liés à l'absence d'information sur les quantités vendues (quels sites interrogés, quels produits enquêtés...)
  - Qui peuvent être renforcés par la possibilité de suivre l'exhaustivité des produits (yc ceux qui sont marginaux)
  - Mais ne se posent pas dans la même mesure pour les produits vendus par un nombre restreint d'opérateurs

- **Une nécessité d'automatisation des traitements statistiques**
  - Pour l'identification des produits, leur classement dans la nomenclature, les remplacements des produits
  - Mais ne se pose pas dans la même mesure pour les services-formulaires

En conclusion,

- **Recourir à ces nouvelles sources de données**
  - Est une nécessité pour prendre en compte de nouveaux modes de consommation ou de nouveaux phénomènes de fixation des prix
  - Permet d'être plus précis et représentatif de la consommation des ménages

- **Qui pose un certain nombre de difficultés**
  - Accès sûr et durable aux données
  - Traitements statistiques nouveaux nécessitant des informations externes
  - Coût lorsqu'il y a de nombreuses enseignes / sites

- **Un choix de l'Insee prudent**
  - Ne pas modifier les concepts de l'indice des prix à la consommation
  - Tester et expérimenter avant de mettre en production

Données de caisse et webscraping

www.insee.fr
@InseeFr

Merci de votre attention

Avez-vous des questions ?

Insee
Mesurer pour comprendre

Marie Leclair
Chef de la division des prix à la consommation
DSDS
01.87.69.63.42
marie.leclair@insee.fr

Insee
DG