Applications mathématiques dans le domaine de la mesure d'audience des médias

Formation IPR "Mathématiques et sciences sociales"
Jeudi 29 septembre 2011

mediametrie

Plan

Médiamétrie

- Présentation de l'entreprise
- Outils mathématiques utilisés à Médiamétrie

Exemples d'application des outils mathématiques

- Calculs de précision des indicateurs d'audience
- Fusions de données d'enquêtes
- Hybridation de données
- Prévision du taux d'équipement TNT
- Probabilisation des contacts

Conclusion

Médiamétrie

L’entreprise et la place des mathématiques au sein de Médiamétrie

mediametrie

Médiamétrie

Vocation

- Assurer la mesure d'audience des médias audiovisuels et interactifs pour en faire un outil de référence professionnel unique avec le consensus général

Organisation

- Des directions commerciales
  - En charge des relations avec les clients
  - Connaissance « métier » des résultats d'audience

- Des directions « support »
  - Statistique, informatique, technologie, terrain d'enquêtes
  - Expertise en termes de conception méthodologique et de production des résultats

La mesure d'audience des médias

Radio

- Enquête par sondage réalisée par téléphone auprès de la population des 13 ans et plus
- 400 interviews par jour en continu sur l'année
- Résultats produits sur des cumuls bimestriels ou trimestriels

Télévision

- Mesure automatique des chaînes regardées sur un panel de 4300 foyers représentatif des foyers français équipés TV
- Recrutement par téléphone
- Résultats produits quotidiennement et livrés tous les jours à 9h

La mesure d'audience des médias

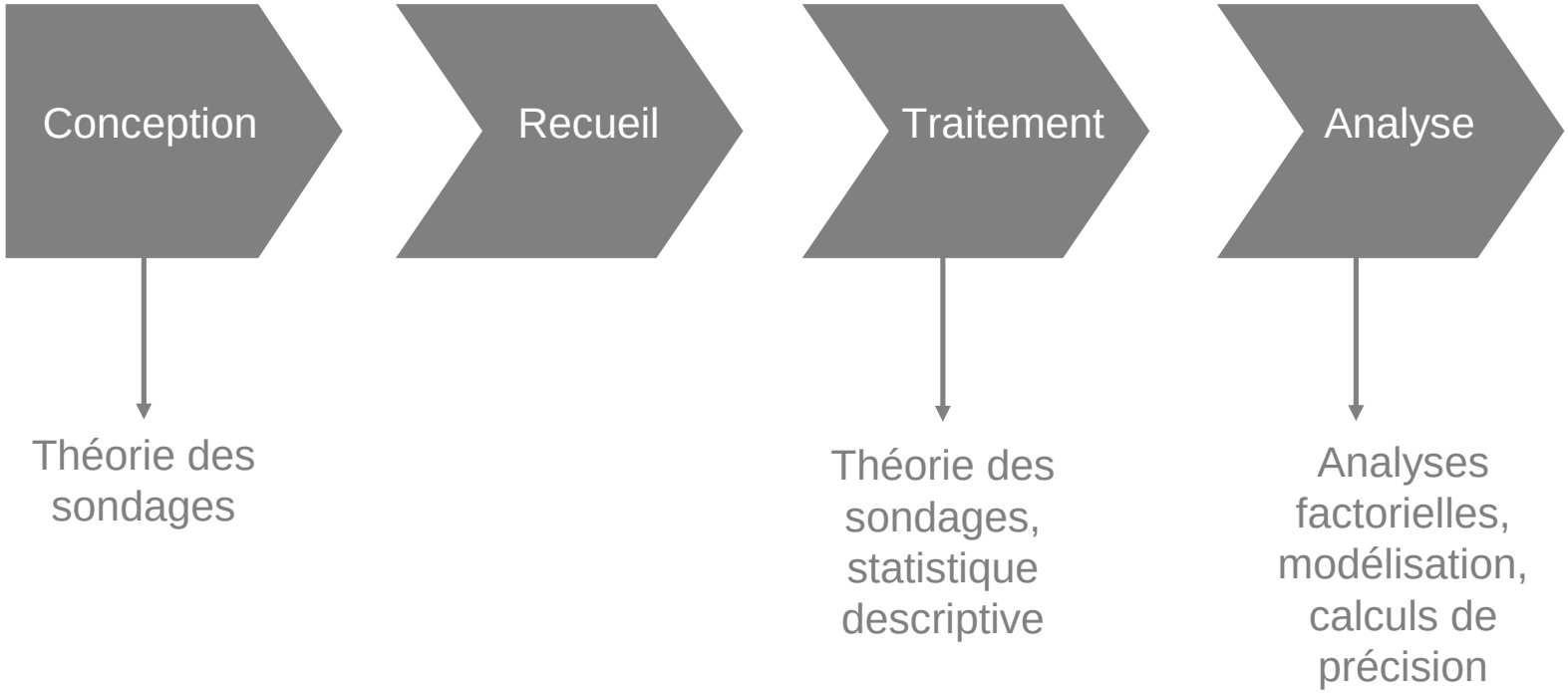
Internet

- Mesure automatique des sites Internet visités sur un panel de 25 000 individus
- Recrutement par téléphone et par Internet
- Résultats produits mensuellement

Internet Mobile

- Collecte exhaustive des logs des trois principaux opérateurs français
- Un panel de 10 000 individus de 11 ans et plus équipés Internet Mobile
- Recrutement par téléphone et par Internet
- Résultats produits trimestriellement (mensuellement à partir de 2012)

Les outils mathématiques associés à la réalisation d'enquêtes par sondage



mediametrie

Exemples d’application des outils mathématiques

Calculs de précision des indicateurs d’audience

mediametrie

La durée d'écoute

Définition :

- Variable de base (Vu/Pas vu):
  - X(i,c,s) = 1 si l 'individu i regarde la chaîne c à la seconde s
  - X(i,c,s) = 0 sinon.

- Poids de redressement :
  - On note $\pi(i)$ le poids de l 'individu i

- Durée d'écoute de l'individu i
  - Temps passé par l 'individu i devant la chaîne c durant la tranche horaire T :

$$D(i,c,T) = \sum_{s \in T} X(i,c,s)$$

- Durée d'écoute individuelle (notée DEI) :
  - En moyenne (pondérée) sur le panel, temps passé devant le chaîne c durant la tranche horaire T :

$$D(c,T) = \frac{1}{\sum_i \pi(i)} \times \sum_i \pi(i) \times D(i,c,T)$$

Le Taux Moyen d'audience

Définition :

- On note Dd(T) la durée de la tranche horaire T
- Taux Moyen d'audience :
  - $$\boxed{TM(c,T) = \frac{D(c,T)}{Dd(T)}}$$
  - l'indicateur de référence en télévision
- Par la suite, on note N la taille pondérée du panel $\Sigma_i\ \pi(i) = N$

mediametrie

Le Taux Moyen d'audience

Interprétation :

- **2 écritures possibles :**

  - **(1)** $$TM(c,T) = \sum_i \frac{\pi(i) \times D(i,c,T)}{N \times Dd(T)} = \frac{1}{N} \sum_i \left[\pi(i) \sum_{s \in T} \frac{X(i,c,s)}{Dd(T)} \right]$$
  - **(2)** $$TM(c,T) = \frac{1}{Dd(T)} \sum_{s \in T} \left[\sum_i \frac{\pi(i) \times X(i,c,s)}{N} \right]$$

- **2 interprétations possibles :**

  - **(1)** : Moyenne, sur l'ensemble des individus, des proportions de vision de T :
    - l'individu moyen de la population étudiée a regardé TM(c,T) (en %) de la durée de T sur la chaîne c,
  - **(2)** : Moyenne calculée sur toute la durée de T exprimée en secondes, de la proportion de personnes regardant l'émission E à la seconde s
    - à la seconde moyenne de T, TM(c,T) (en %) de la population regardait la chaîne c

Méthode analytique

- Variable aléatoire étudiée :
  - Z variable aléatoire à valeurs dans [0, 1] représentant la proportion de vision de T sur la chaîne c

- Statistique étudiée :
  - Moyenne : m(Z) = E(Z)
  - TM(c,T) est un estimateur (sans biais) de m(Z)

- Observations :
  - $Z(i) = Z(i,c,T) = \dfrac{D(i,c,T)}{Dd(T)}$ proportion de vision de T sur la chaîne c par l'individu i

- Variance du Taux Moyen d'audience :
  - $$V(TM(c,E)) = \frac{1}{N^2}\left[\sum_i \pi^2(i)\, V(Z(i)) + \sum_i \sum_{j \neq i} \pi(i)\, \pi(j)\, Cov(Z(i), Z(j))\right]$$

Corrélation entre deux individus

mediametrie

Méthode analytique

Inconvénients

- Application numérique très lourde et donc difficile à mettre en oeuvre, même avec des hypothèses simplificatrices sur les corrélations

Besoins client

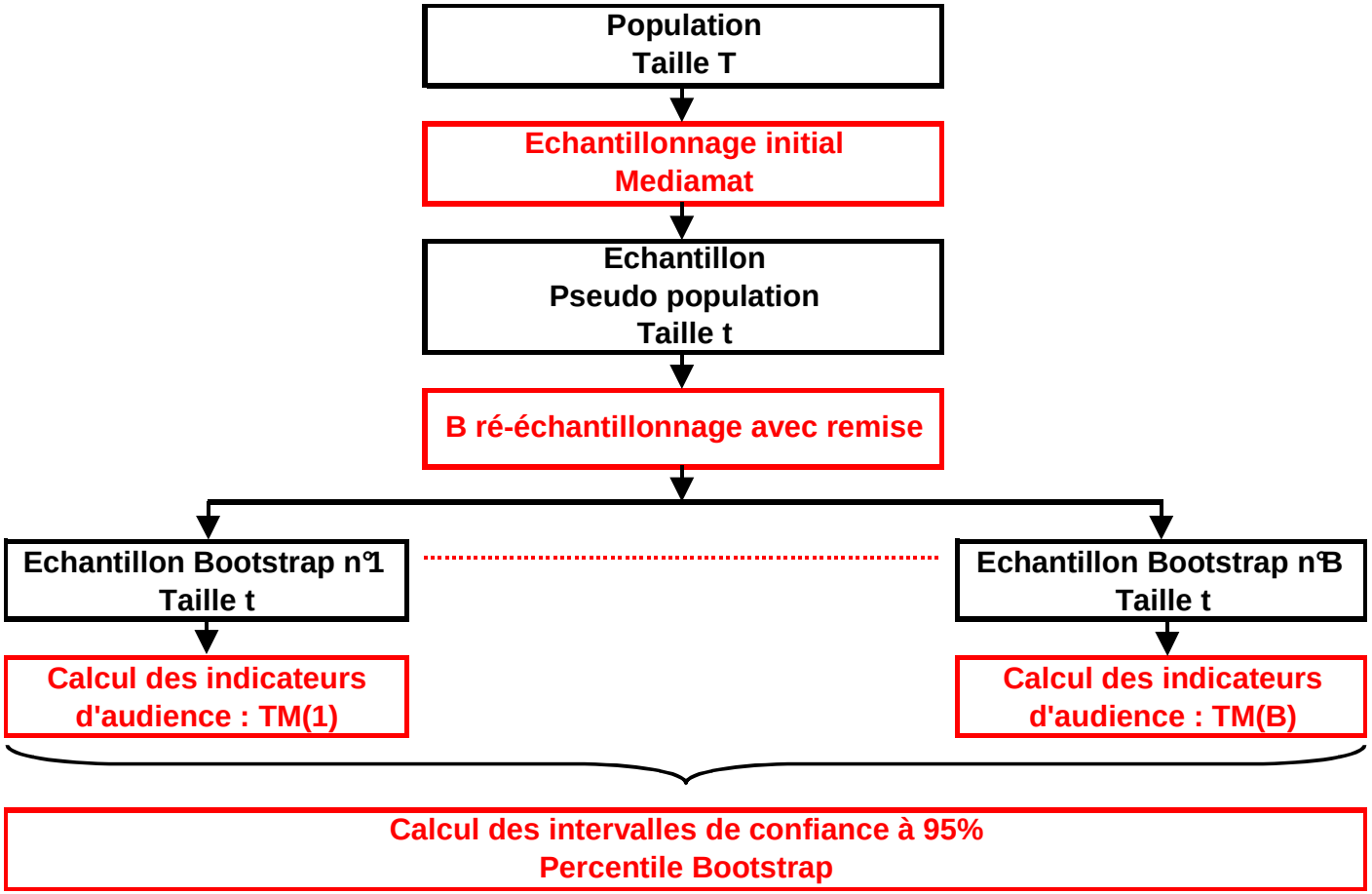
- Un intervalle de confiance pour chaque valeur de TM publiée (par cible, tranche horaire, chaîne, ...)
- Un besoin identique pour des indicateurs plus complexes comme la part d'audience (rapport de taux moyens)

Bilan

- Nécessité de la méthode empirique

Méthode empirique

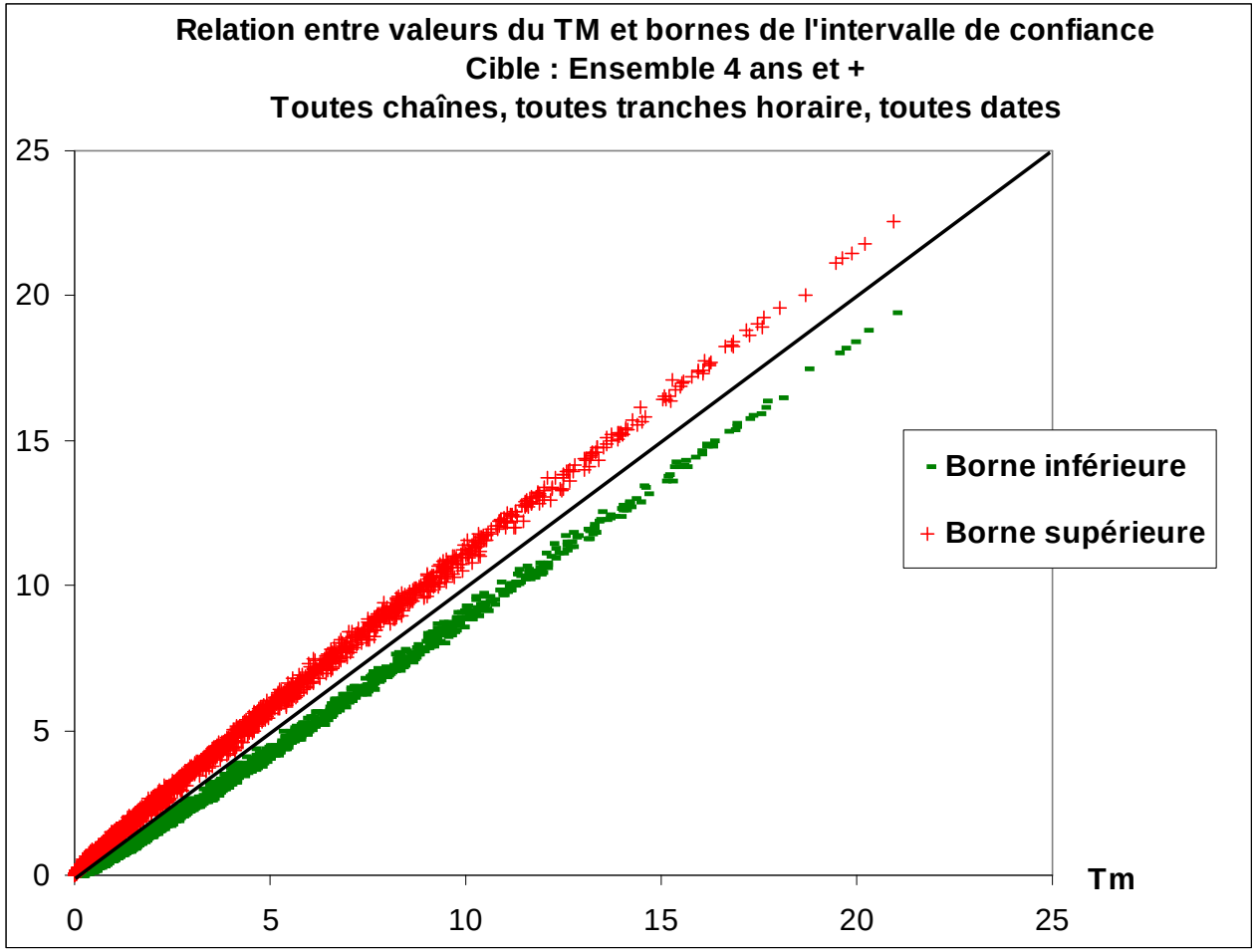
Description de la méthode



mediametrie

Méthode empirique

Résultats

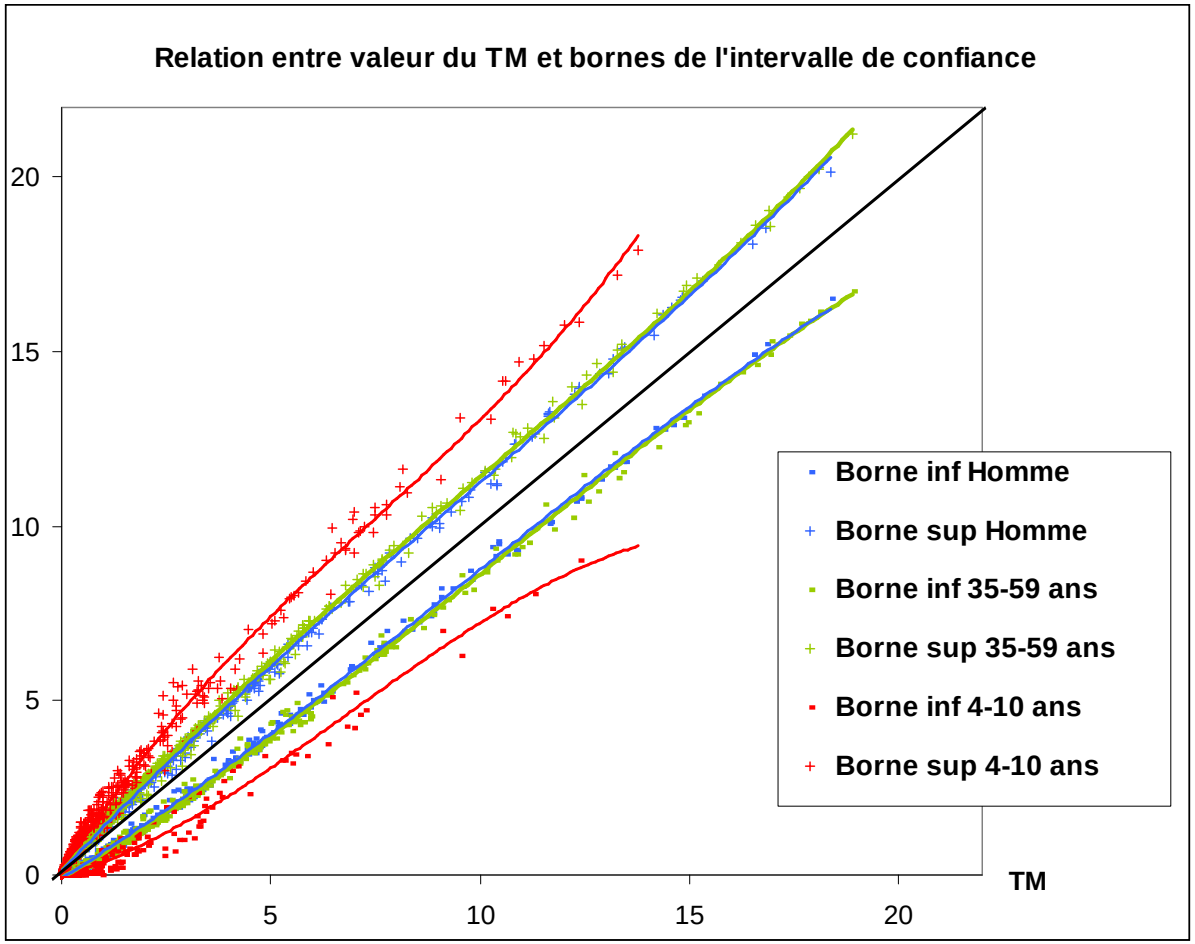


- Pas d'effet tranche horaire
- Pas d'effet date
- Pas d'effet chaîne

mediametrie

Méthode empirique

Résultats



Taille des cibles :

Homme	2655
35-59 ans	2670
4-10 ans	800

- Effet cible : lié uniquement à la taille de la cible

mediametrie

Méthode empirique

Résultats

- Les bornes de l'intervalle de confiance du Taux Moyen peuvent donc s'exprimer en fonction :
  - (1) de la valeur du Taux Moyen (TM)
  - (2) de la taille de la cible (**n_{cible}**)

- Modélisation des bornes de l'intervalle de confiance :

$$\text{Borne sup} = TM + A \quad \text{avec} \quad \log A = a_1 \times \log\left(\frac{TM}{\sqrt{n_{cible}}}\right) + a_2$$

$$\text{Borne inf} = TM - B \quad \text{avec} \quad \log B = b_1 \times \log\left(\frac{TM}{\sqrt{n_{cible}}}\right) + b_2$$

Exemples d’application des outils mathématiques

Fusions de données d’enquêtes

mediametrie

De problématiques mono-média à une problématique pluri-média

Chaque média dispose aujourd'hui d'un dispositif de mesure d'audience spécifique

- Enquête 126 000 Radio pour la Radio
- Panel Médiamat pour la TV
- Panel Médiamétrie//NetRatings pour Internet
- Mesure d'audience de l'internet mobile

Or, aujourd'hui, les stratégies deviennent pluri-média

- L'objectif est d'avoir une estimation de l'ensemble des contacts médias, tout support confondu

De problématiques mono-média à une problématique pluri-média

Comment répondre à cette problématique ?

- Option 1 : Constituer un panel sur lequel on recueillerait les comportements d'audience sur l'ensemble des médias
  - Difficilement concevable aujourd'hui (coûts élevés et taux de participation a priori faibles)

- Option 2 : Recueillir, auprès des mêmes individus, les habitudes de consommation des différents médias
  - Ne répond que partiellement à la problématique
  - En particulier, ne permet pas de mesurer de manière précise la duplication entre supports d'un même groupe média (par exemple, NRJ, NRJ12 et nrj.fr)

De problématiques mono-média à une problématique pluri-média

- Option 3 : Procéder à des fusions statistiques des différentes enquêtes
  - Principe : injecter au fichier dit « receveur » un certain nombre de variables présentes dans le fichier « donneur »

Base A Receveuse	i_1	$X(i_1)$	$Y(i_1)$
	i_2	$X(i_2)$	$Y(i_2)$
	…	…	…
	i_{nA}	$X(i_{nA})$	$Y(i_{nA})$
Base B Donneuse	j_1	$X(j_1)$	$Z(j_1)$
	j_2	$X(j_2)$	$Z(j_2)$
	…	…	…
	j_{nB}	$X(j_{nB})$	$Z(j_{nB})$

→

Base F Fusionnée	…	…	…	…
	f =(i,j)	X(i)	Y(i)	Z(j)
	…	…	…	…

mediametrie

Les fusions de données

Deux types d'approche

- Approche par modèle
  - Chaque variable spécifique est étudiée de manière indépendante
  - Limite : les corrélations observées dans le fichier donneur sont difficilement maintenues dans le fichier receveur

- Appariement
  - Détecter, pour chaque individu du fichier receveur, un ou plusieurs « sosies » au sein du fichier donneur

  ⇨ C'est cette approche qui a été choisie pour l'étude Cross-Médias

Les grandes étapes de l'appariement

Etape 1 : sélection des variables relais

- Lister ou construire les variables communes aux deux fichiers
- Les hiérarchiser en fonction de leur pouvoir explicatif sur les variables spécifiques :
  - certaines ne seront pas retenues comme variables relais du fait de leur faible pouvoir explicatif
  - d'autres seront considérées comme prioritaires / obligatoires : ce sont les variables de contrôle
- Au sein de chaque strate, donneurs et receveurs seront appariés en fonction de leur similitude au regard des variables relais

Les grandes étapes de l'appariement

Etape 2 : choix de la fonction de distance

- Distance de Manhattan : $\sum_{i=1}^{n} |x_i - y_i|$
- Distance euclidienne : $\sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$
- Distance de Levenshtein :
  - distance intégrant une fonction de coût permettant de pénaliser les différences en tenant compte de la probabilité d'apparition d'une valeur
- Distance de Jaccard :
  - distance permettant de mesurer la proximité entre deux jeux de variables binaires
- Et bien d'autres…

Les grandes étapes de l'appariement

Etape 3 : sélection des sosies

- Processus d'optimisation sur les critères suivants :
  - Distance moyenne minimale
  - Minimisation de la réplication d'un même donneur

Etape 4 : validation

- Validation de l'appariement
  - Distribution des distances entre donneur et receveur
  - Nombre d'apparitions des donneurs et des receveurs dans la base finale
- Cohérence avec les bases d'origine
  - Moyennes / distributions marginales
  - Test de Wilcoxon de comparaison des rangs
  - Test de Mantel de comparaison des corrélations

Les enjeux

Difficultés mathématiques

- Choix d'une fonction de distance
- Choix des critères sur lesquels on détermine la similarité entre 2 individus
- Conservation des résultats initiaux (moyennes des variables Y calculées sur A et des variables Z calculées sur B)

Limites en terme d'analyse

- Précautions à prendre sur l'analyse et l'interprétation des résultats

Exemples d’application des outils mathématiques

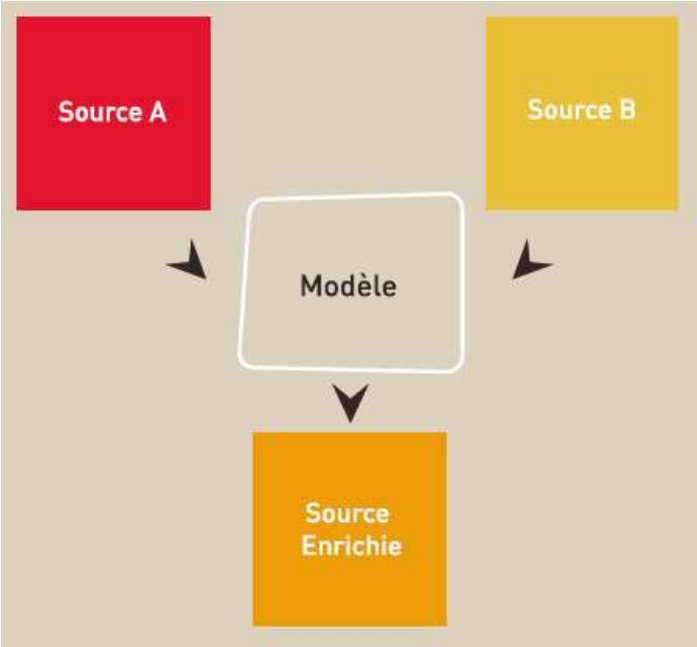
Hybridation de données

mediametrie

Le concept d'hybridation de données

Principe et objectif

- Mélanger deux sources d'information de natures ou de niveaux différents, les croiser pour en créer une troisième, plus fine ou plus riche



mediametrie

L'hybridation en pratique

Aujourd'hui, les voies de retour apportent des informations exogènes complémentaires aux enquêtes

- Sur Internet, on dispose d'une mesure exhaustive du nombre de pages vues et de visite par site
- Sur la TV par ADSL, on connaît, à chaque instant, le nombre de boîtiers allumés sur chaque chaîne

Comment utiliser ces informations complémentaires ?

- La question n'est pas tant celle de l'amélioration de la précision des enquêtes que celle de la mise en cohérence des résultats de sources différentes

L’hybridation en pratique

La méthode classique : le calage sur marges

- Le principe consiste à modifier les poids de sondage de manière à faire coïncider les totaux observés sur l’échantillon avec ceux calculés sur la population
- L’utilisation de la macro CALMAR, mise au point par O. Sautory (INSEE) sur l’approche générale formulée par J.C. Deville et C.E. Särndal (1992), permet de minimiser la distance entre les poids de redressement et les poids de sondage

Les difficultés

Le problème de la multiplicité des variables auxiliaires

- Combien de variables auxiliaires peut-on faire intervenir ?
- Au-delà de la question de la convergence de l'algorithme se pose celle de la dispersion des poids de redressement
- Dans le cas d'une trop grande dispersion des poids de redressement, l'estimateur redressé peut perdre son efficacité

⇨ **Trouver un compromis acceptable ou mettre au point des méthodes d'estimation différentes dans lesquelles l'enquête n'a plus le rôle central**

Un cas pratique : la mesure d'audience de l'internet mobile

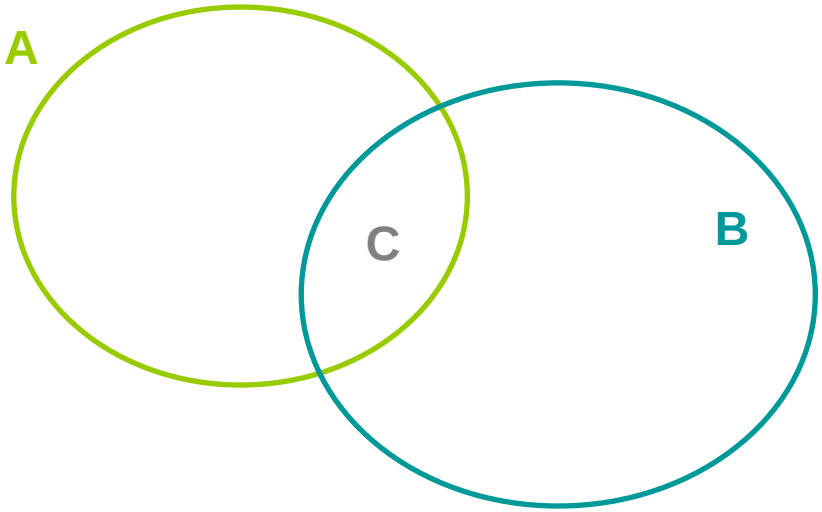
Deux sources d'information

- Les logs des opérateurs de téléphonie
  - Nombre d'individus connectés par site/application
  - Nombre de pages vues par site
- Un panel de 10 000 individus
  - Profil des visiteurs des sites/applications
  - Duplication entre sites/applications

Comment combiner ces deux sources ?

Un cas pratique : la mesure d'audience de l'internet mobile

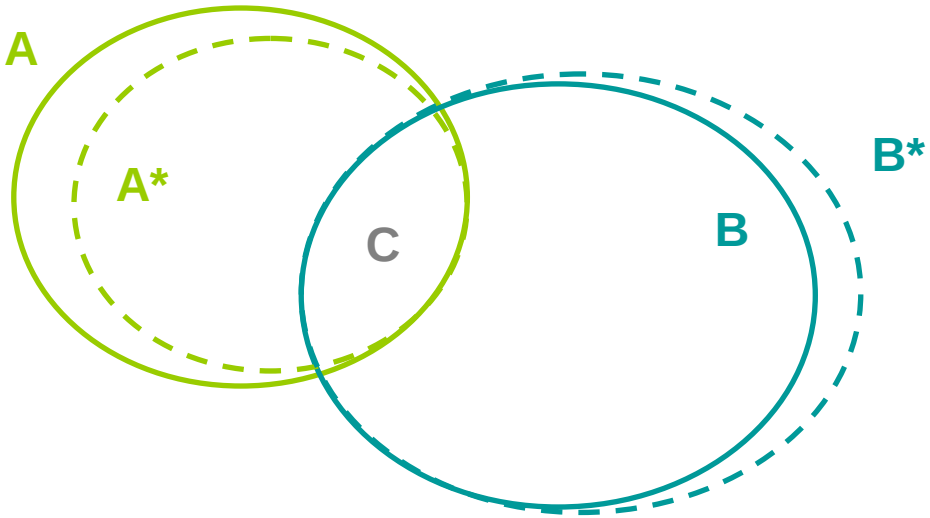
L'estimation de la duplication



- Notations :
  - A, B le nombre de visiteurs respectifs des sites A et B dans le panel
  - C le nombre de visiteurs communs aux sites A et B dans le panel

Un cas pratique : la mesure d'audience de l'internet mobile

L'estimation de la duplication



- Notations :
  - A*, B* le nombre de visiteurs respectifs des sites A et B dans la mesure exhaustive

- Question : estimation de C*

mediametrie

Un cas pratique : la mesure d'audience de l'internet mobile

Les hypothèses de départ

- Dans le cas où A et B sont indépendants, c'est-à-dire

$$C = \frac{A \times B}{N}$$, alors on veut que $$C^* = \frac{A^* \times B^*}{N} = C_{ind}$$

- Dans le cas où A est inclus dans B, alors

$$C^* = \min\left(\frac{C \times B^*}{B}, A^*\right) = C_{inc}$$

- Dans les cas intermédiaires, on est entre les deux…

Un cas pratique : la mesure d'audience de l'internet mobile

Notations

- $K = \min\left(\frac{C}{A \times B/N}, \frac{A \times B/N}{C}\right)$
- $L = \frac{C}{A}$
- Alors C* est estimée par : $C^* = \frac{C_{ind} \times (1-L) + C_{inc} \times (1-K)}{1-L+1-K}$

mediametrie

Exemples d’application des outils mathématiques

Prévision du taux d’équipement TNT

mediametrie

Prévision du taux d'équipement TNT

Contexte

- Dans le cadre de la mesure d'audience de la télévision, besoin d'un suivi au plus près de l'équipement audiovisuel des ménages français

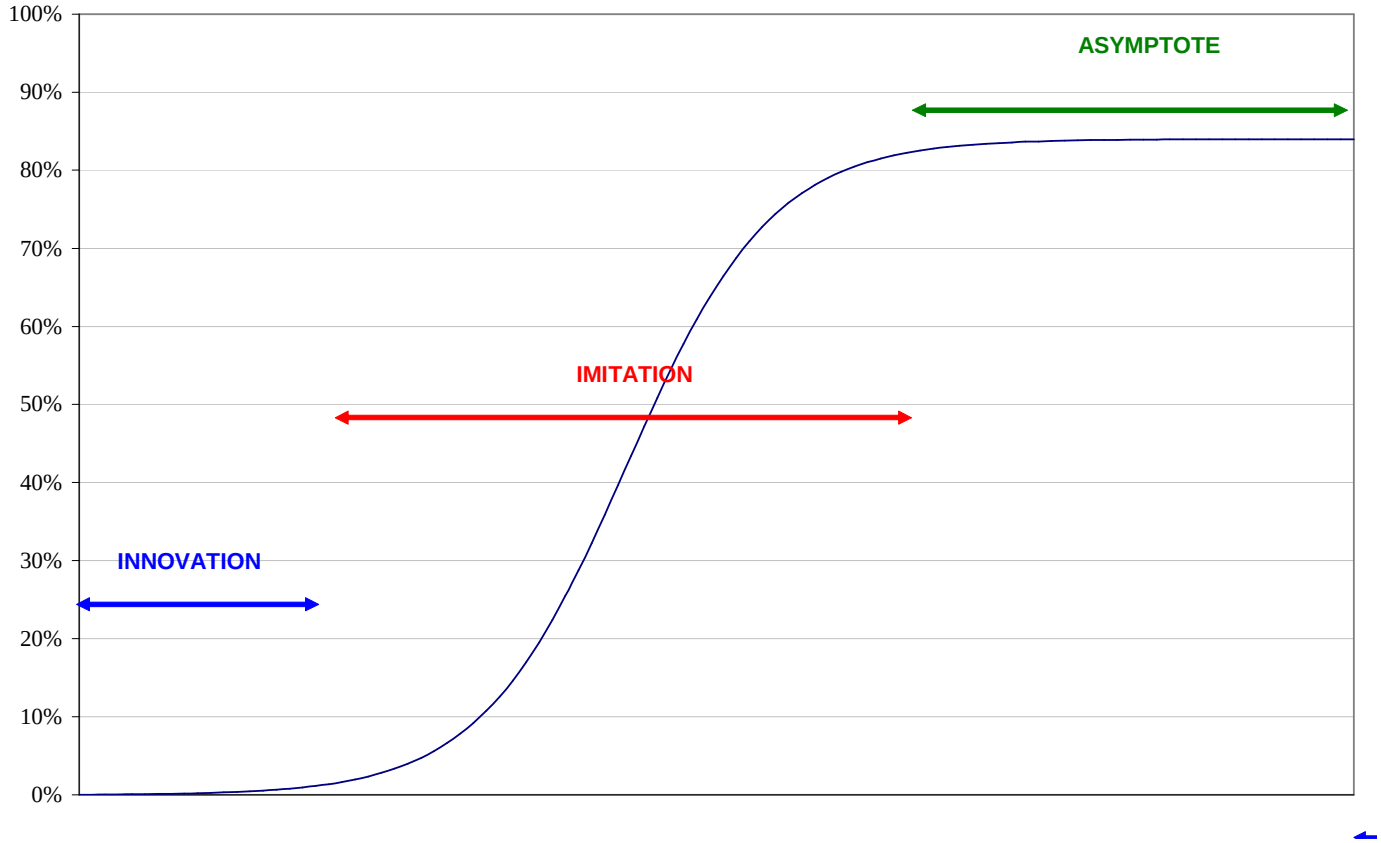
Positionnement du problème

- Suivi assuré à l'aide d'une enquête réalisée en face-à-face mais dont les résultats sont publiés trimestriellement
- Pour un bien d'équipement en forte progression, cette fréquence est insuffisante

⇨ Mise au point d'un modèle de prévision mensuel

Modèles de diffusion et d'adoption de nouveaux produits/technologies

Le taux d'adoption d'un nouveau produit ou d'une nouvelle technologie suit une courbe en S :



mediametrie

Modèles de diffusion et d'adoption de nouveaux produits/technologies

- Deux grands types de modèles :

  - **Le modèle logistique** : (du nom de la courbe utilisée)

$$N(t) = \frac{1}{1 + \mu \cdot e^{-\lambda t}}$$

    - ajustement du modèle aux données par transformation simple :
      - $\ln\left(\frac{N(t)}{1 - N(t)}\right) = \lambda t - \ln(\mu)$ (linéaire en t)
      - estimation de λ et μ par les Moindres Carrés Ordinaires (MCO)
      - Remarque : asymptote = 100%

mediametrie

Modèles de diffusion et d'adoption de nouveaux produits/technologies

- Deux grands types de modèles :

  - **Le modèle de Bass :** $$\frac{dN(t)}{dt} = [p + \frac{q}{m}N(t)] \cdot [m - N(t)]$$

    - m : potentiel du marché
    - N(t) : marché conquis à la date t
    - p : coefficient d'innovation
    - q : coefficient d'imitation

  - Solution sous la forme : $$N(t) = m \cdot \left[\frac{1 - e^{-(p+q)t}}{1 + \frac{q}{p} e^{-(p+q)t}} \right]$$

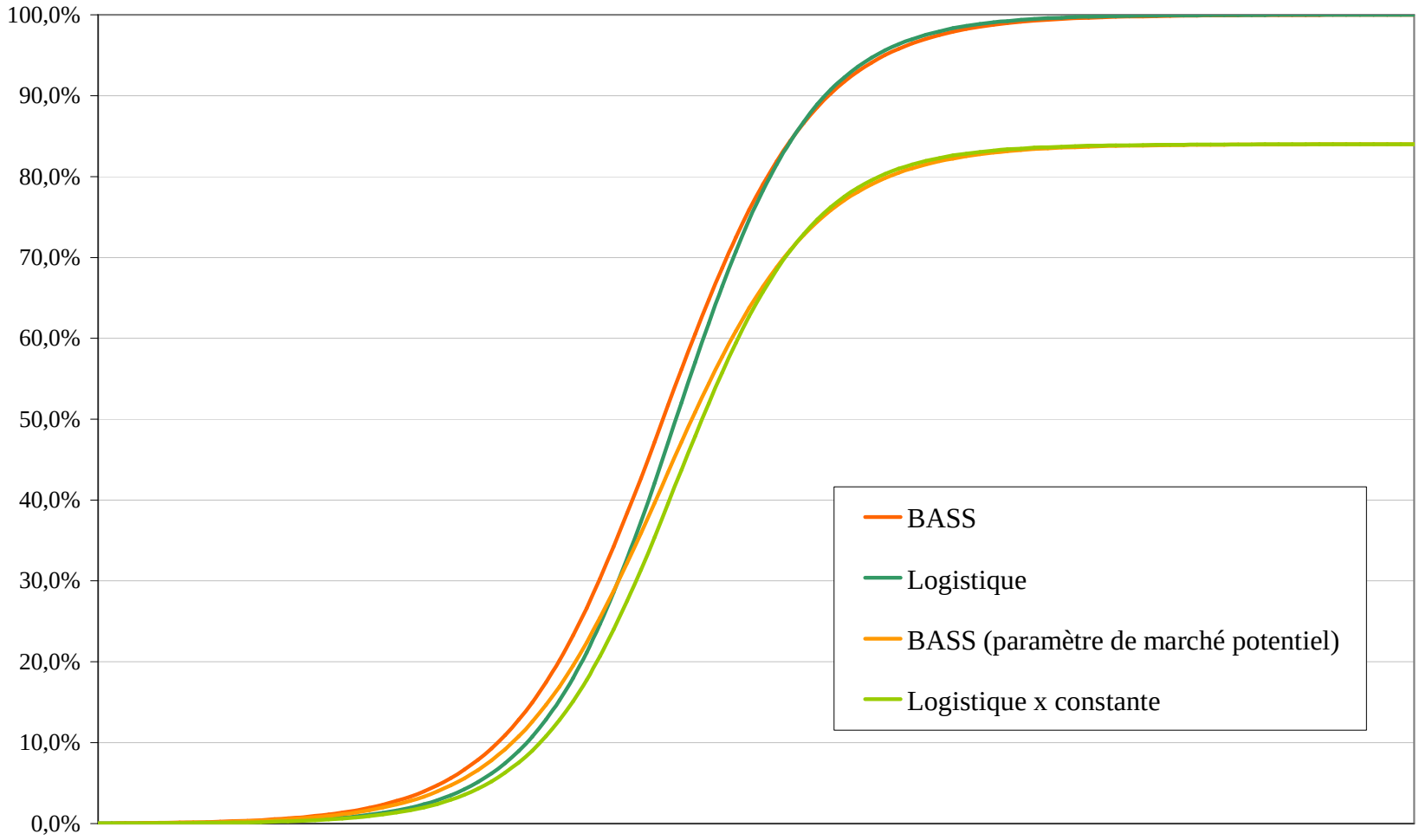
  - Ajustement aux données :

    - $N(t+1) - N(t) = pm + (q-p) \cdot N(t) - \frac{q}{m} \cdot (N(t))^2$ (linéaire en N(t) et $N^2(t)$)

    - estimation de m, p et q par les MCO

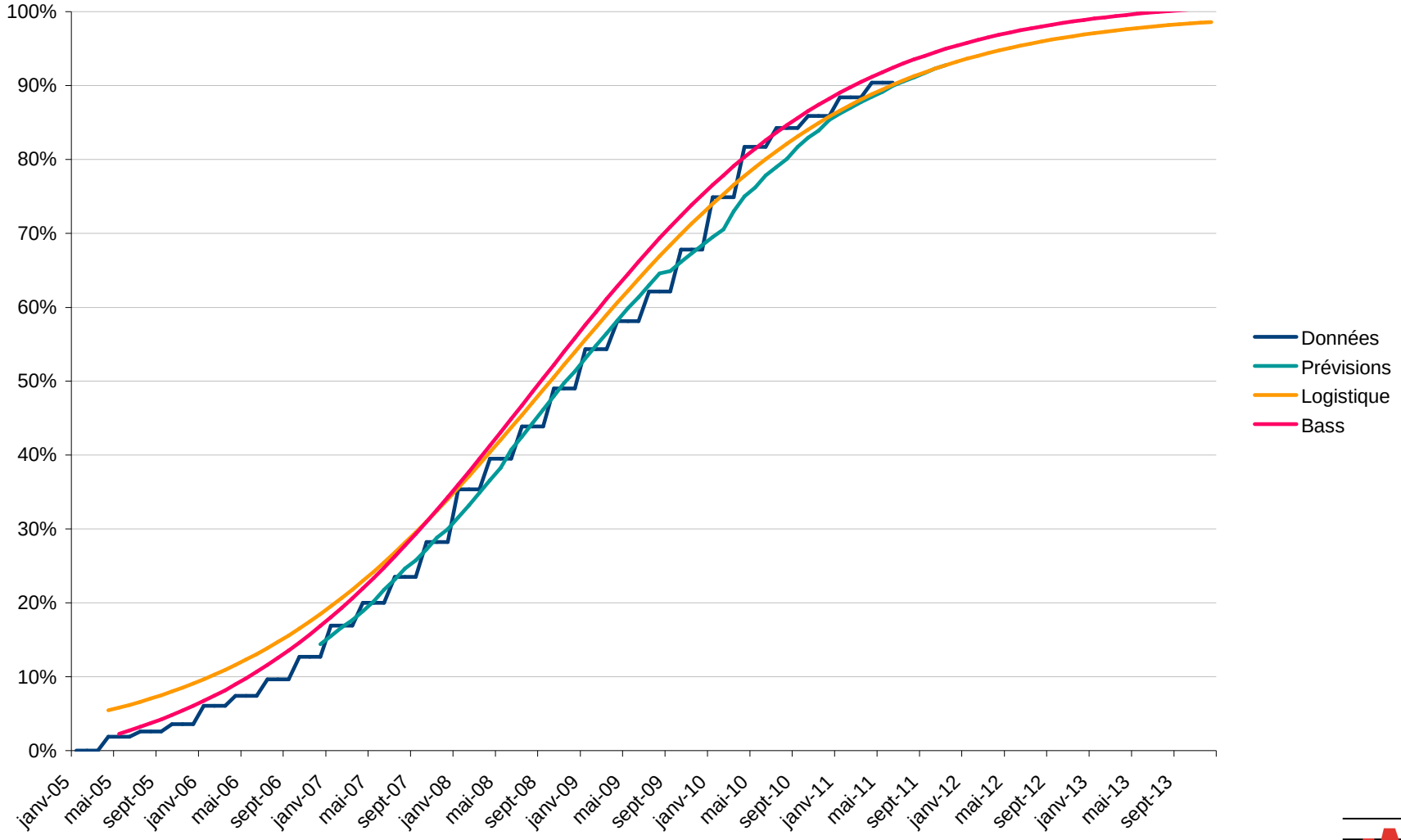
Modèles de diffusion et d'adoption de nouveaux produits/technologies

- Deux grands types de modèles :



mediametrie

L'équipement TNT



mediametrie

Exemples d’application des outils mathématiques

Probabilisation des contacts

mediametrie

La mesure d’audience : pourquoi ?

Les besoins des diffuseurs

- Constitution d’une grille de programmes
  - Émissions écoutées ou regardées
  - Transferts d’audience
  - Stabilité / fidélité de l’auditoire
- Analyse de la concurrence

Les besoins des publicitaires

- Choix du média
- Choix des supports
- Choix des emplacements
- Contrôle a posteriori de l’atteinte des objectifs

Médiaplanning

Définition

- Ensemble des opérations qui visent à étudier puis à sélectionner les différents media, puis les différents supports, pour y réserver l'espace nécessaire à la diffusion du message publicitaire la plus efficace possible et au meilleur coût
- Détermination de la combinaison optimale de médias et supports qui, compte tenu des moments de passage des messages, permet d'atteindre la majeure partie de la cible visée au moindre coût

Pour cela

- L'estimation des performances du plan est faite a priori (prévision) à partir de probabilités de contact d'un individu avec chaque support

⇨ **Méthode de probabilisation**

Probabilisation

Qu’est-ce qu’un fichier d’audiences probabilisées ?

- C’est un échantillon d’individus caractérisés, d’une part, par des informations sociodémographiques et, d’autre part, par un vecteur de probabilités de contact avec un support

Objectif de la probabilisation

- Affecter ces probabilités individuelles de contact avec chaque support de façon à être le plus proche possible des données effectivement mesurées
- La répartition effective des probabilités d’audience sur une population donnée est modélisée par un modèle de probabilités dit « bêta-binomial » (Chandon, 1976), avec loi bêta pour le paramètre sur]0, 1[et segments en 0 et 1

Probabilisation

Phase 1 :

- X est la v.a. dénombrant les contacts avec la campagne
- Détermination de la loi de X

Phase 2 :

- Hiérarchisation des individus selon une probabilité d'audience croissante
- Critères utilisés :
  - nombre de contacts observés
  - déclarations d'habitudes d'écoute du média, des supports…
  - profil socio-démographique

Phase 3 :

- Affectation des probabilités individuelles

Conclusion et perspectives

mediametrie

Conclusions et perspectives

Les mathématiques

- Au cœur de notre activité
- Des applications concrètes et opérationnelles

Et face à une concurrence de plus en plus forte

- Problématiques de plus en plus complexes
- Besoin d'analyse en croissance permanente

Autre dimension importante : la vulgarisation scientifique

Merci pour votre attention

Pour plus d'informations …

www.mediametrie.fr

Aurélie VANHEUVERZWYN

Directrice Analyses et Méthodes Scientifiques

avanheuverzwyn@mediametrie.fr