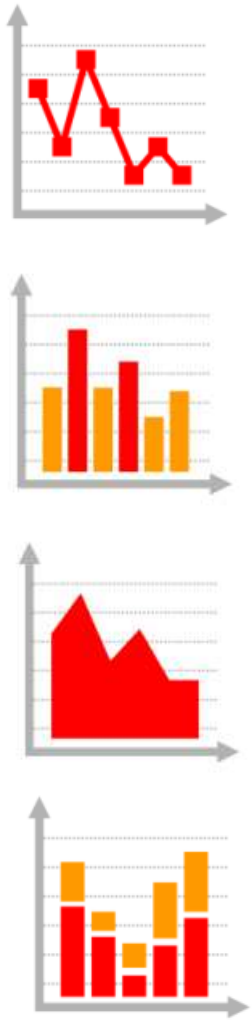
# Choisir de répondre au recensement par internet : un essai de modélisation

**Heidi Koumarianos**
**DMCSI**



22 juin 2017

# Plan

- Contexte et problématique
- Quel(s) modèles(s) ?
- Quelques résultats

# Contexte

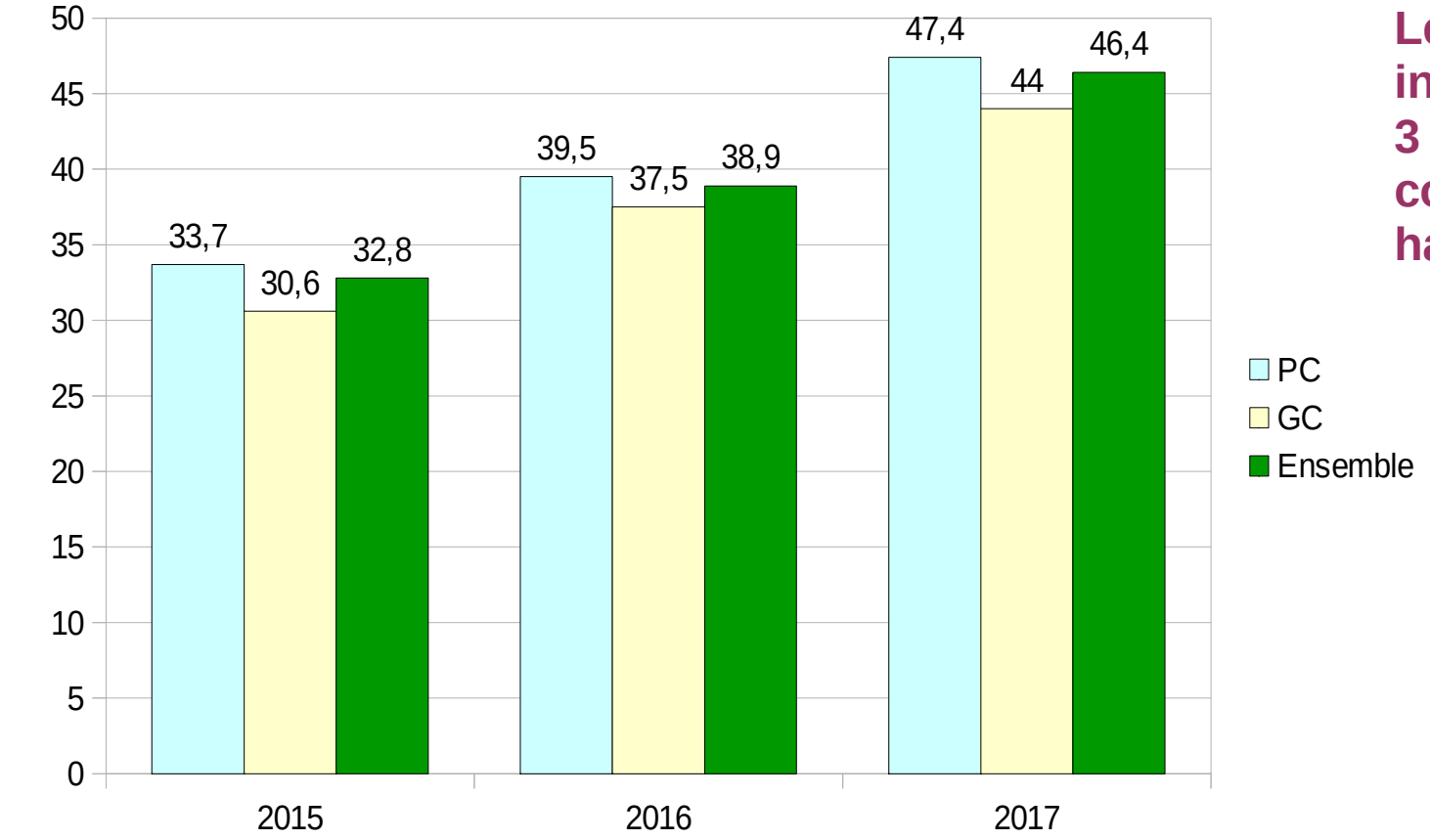
# La collecte du recensement (1)

- Depuis 2004, la collecte est annuelle, réalisée par les communes en partenariat avec l'Insee
- Chaque année, les communes de plus de 10 000 habitants (un peu moins de 1 000 communes) effectuent une collecte auprès de 8 % des logements
- Chaque année, 20 % des petites communes de moins de 10 000 habitants (environ 7 000 chaque année) effectuent un recensement exhaustif de leur population, à un rythme quinquennal
- Environ 24 000 agents recenseurs en 2017 (en moyenne 200 logements par agent recenseur)

# La collecte du recensement (2)

- Depuis 2015, chaque ménage peut choisir son mode de collecte (papier/internet)
- Les taux de réponse par internet sont très variables
  - Les communes et les zones de collecte diffèrent par leur composition socio-démographique
  - L'agent recenseur et la commune peuvent influencer le choix du mode de réponse, en privilégiant la proposition de l'un ou de l'autre.

# Le taux de réponse par internet a progressé de 14 points en 2 ans

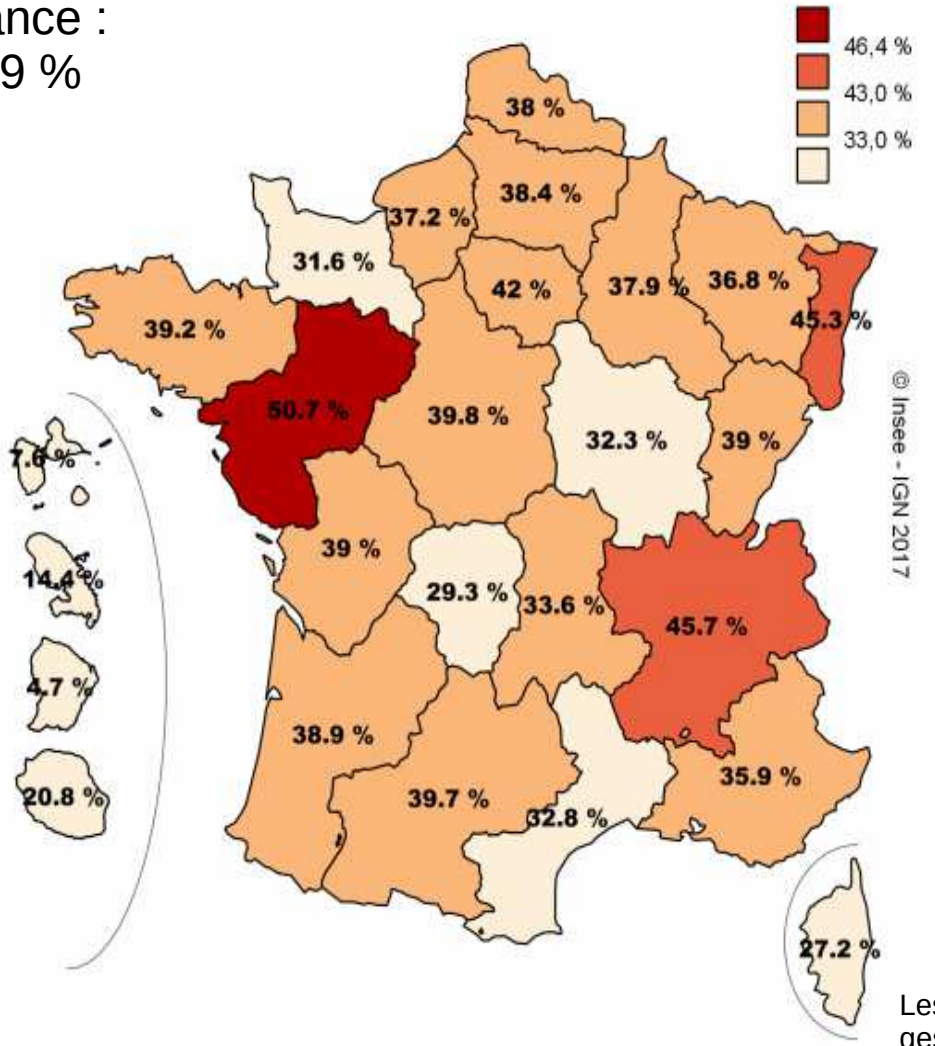



**Le taux de réponse par internet reste plus élevé de 3 points en « petite » commune (moins de 10 000 habitants) qu'en « grande »**

# Des disparités régionales dans les taux de réponse internet

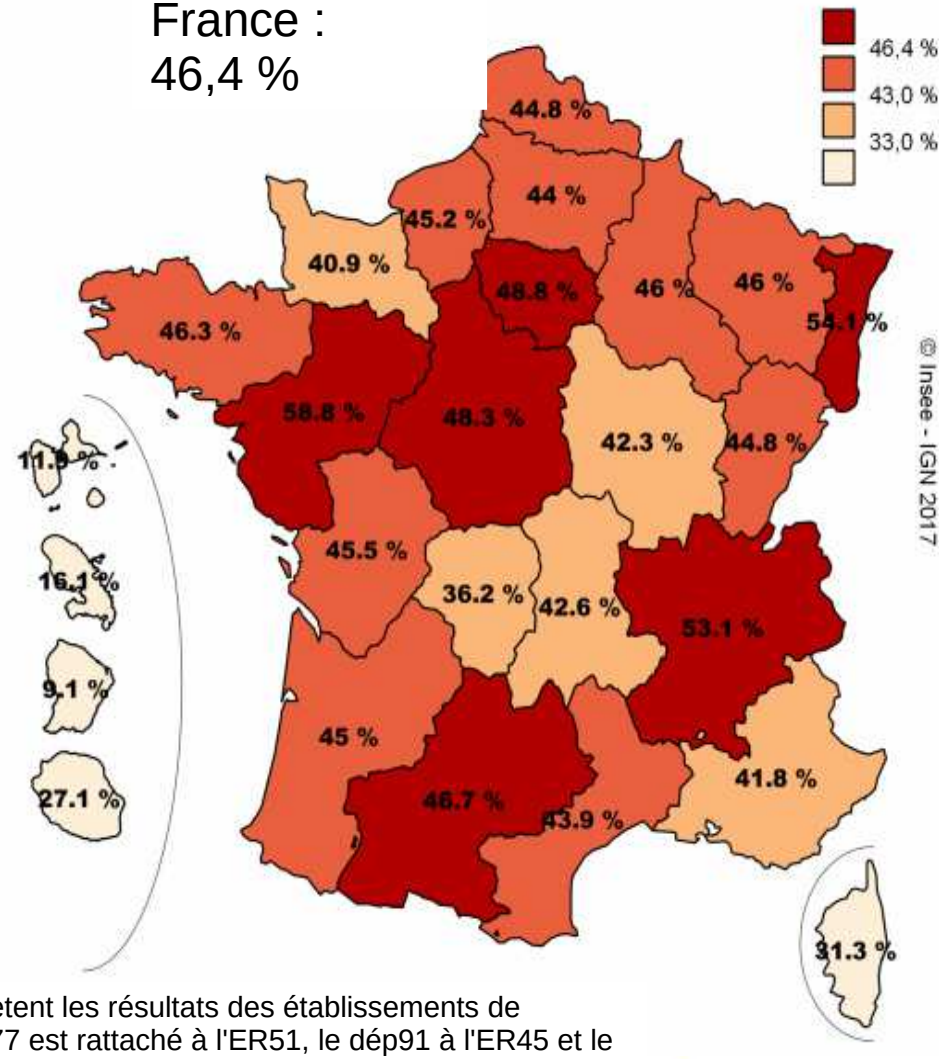
## EAR2016

France :
38,9 %




## EAR2017

France :
46,4 %




Les chiffres reflètent les résultats des établissements de gestion : le dép77 est rattaché à l'ER51, le dép91 à l'ER45 et le dép95 à l'ER76

# La problématique

# Pourquoi étudier la propension à choisir un mode de réponse dans le recensement ?

- C'est la plus grosse enquête de l'Insee (plus de 4 millions de ménages tous les ans), et elle est multimode
- La mesure des biais de sélection est nécessaire pour estimer d'éventuels effets de mode
- La mesure des biais de sélection est nécessaire pour mettre en place des protocoles de collecte différenciés pour des sous populations

# Les difficultés

- On constate de fortes disparités des taux de réponse internet par zone (zone de collecte, commune, région)

  => on souhaite neutraliser cet effet zone

- Les bases de données sont volumineuses

  => les choix de modélisation peuvent être contraints (4 millions de ménages, 24 000 zones de collecte, 8 000 communes et ... des outils limités)

# La modélisation

# Choix et construction des variables du modèle

- On dispose d'informations de niveau logement, ménage, et individus composant le ménage
  - statut d'occupation du logement (3 modalités)
  - taille du logement (surface par habitant, en tranche, 5 modalités)
  - type de ménage (5 modalités)
- Le choix du mode de collecte est fait pour un ménage : les informations de niveau individu doivent être agrégées.
  - moyenne d'âge des personnes majeures (en tranches, 6 modalités)
  - diplôme le plus élevé des personnes majeures (en 5 modalités)
  - présence d'au moins un actif occupé
- On peut construire des caractéristiques de zones, principalement les moyennes des caractéristiques individuelles, mais aussi :
  - le taux de collecte internet N-1 (en grande commune)
  - les revenus de la commune

# On peut modéliser :

- La propension d'un ménage à répondre par internet
  - en fonction de ses caractéristiques propres
  - en fonction des caractéristiques de sa zone de collecte
  - en intégrant le niveau zone de collecte
  - en intégrant le niveau commune
- Le taux de réponse par internet pour une zone de collecte
  - en fonction des caractéristiques de la zone de collecte
  - en intégrant le niveau commune

# Le modèle de Mundlak

- La probabilité qu'un ménage $m_i$ appartenant à la zone de collecte $ZC_j$ réponde par internet dépend entre autres :
  - De ses caractéristiques individuelles $X_{ij}$
  - De caractéristiques de sa zone de collecte, de façon générale les $Y_j$
  - Des moyennes des caractéristiques individuelles de la zone les $\overline{X}_{.j}$
  - D'une constante globale éventuelle K
  - D'une constante pour la zone éventuelle $k_j$

$$P\left(m_i \in ZC_j \text{ répond par internet}\right)$$
$$a_{ij}X_{ij} + b_j Y_j + c_{.j}\overline{X}_{.j} + K + k_j + e_{ij}$$

# Les premiers résultats

Les résultats présentés portent sur les grandes communes de métropole

# Quelles questions à ce stade ?

- Les effets commune et zone de collecte sont-ils avérés ?
- Y a-t-il des effets des moyennes socio-démographiques par zone ?
- Quelles sont les caractéristiques qui jouent le plus sur le choix du mode ?

# Les niveaux commune et ZC sont explicatifs du choix du mode

- Test d'un modèle « vide » : les niveaux commune et zone de collecte expliquent respectivement **9 % et 12 %** de la variance (coefficient de corrélation intra classe)
- Lorsqu'on introduit d'autres variables explicatives (de niveau ménage et moyennes de zone), les coefficients sont de 5 % (commune) et **15 %** (zone de collecte)
  - L'effet commune et l'effet zone de collecte sont en partie captés par l'introduction du taux internet de l'année précédente

# Limites du modèle à 3 niveaux

- Le modèle comprenant les niveaux ménage, zone de collecte et commune est trop important pour tourner sur la France entière
- Le niveau commune est bien moins explicatif que le niveau zone de collecte (estimation faite sur 5 régions)
- L'application aux petites communes est plus délicat

  => Par la suite on privilégie un modèle à deux niveaux (ménage et zone de collecte), que l'on peut estimer sur l'ensemble des communes (on distingue toutefois grandes et petites communes)

# Les caractéristiques socio-démographiques des zones sont peu explicatives

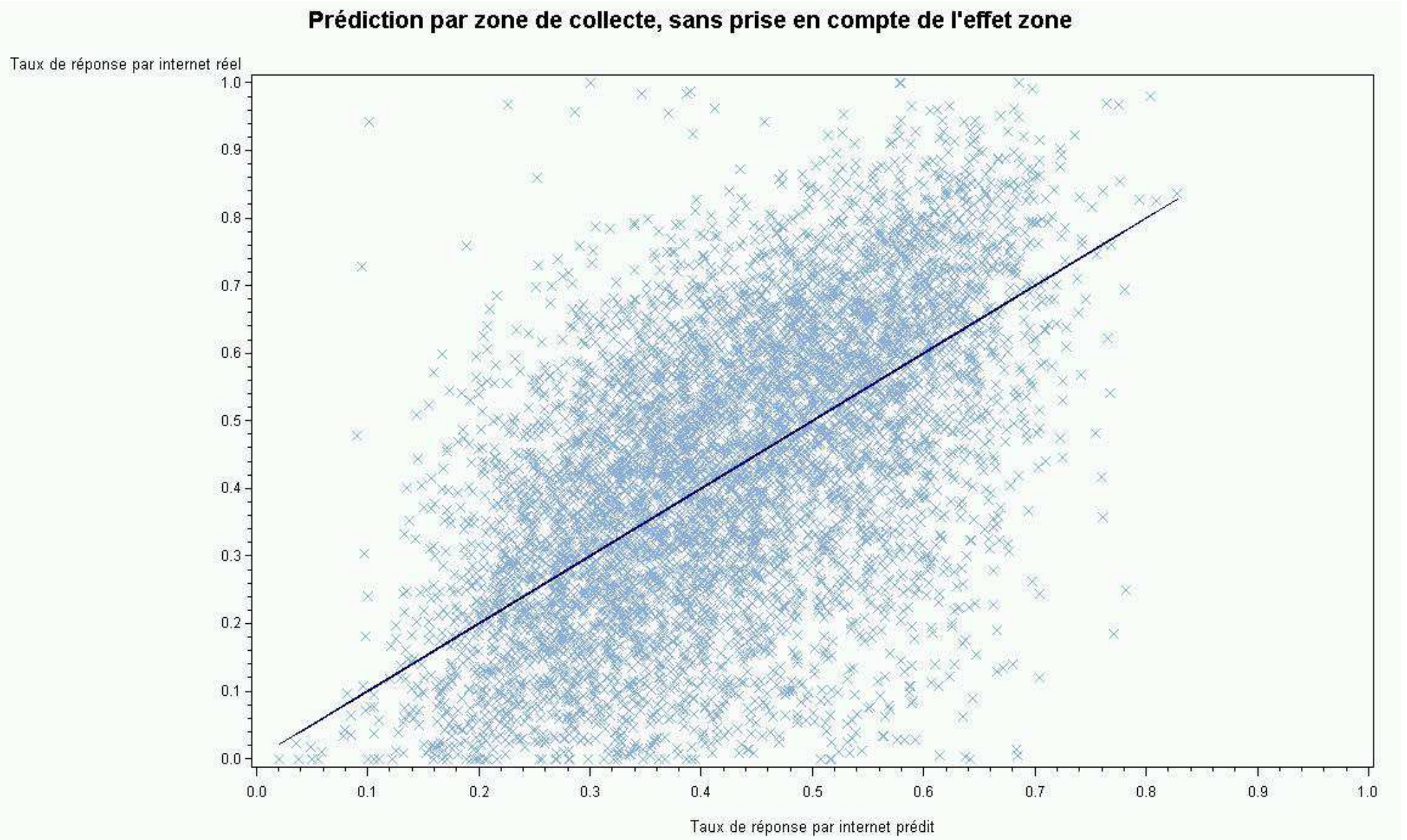
- Certaines caractéristiques ressortent parfois, lorsque le modèle porte sur un nombre plus réduit d'observations (une ou deux régions seulement)
- En général, il s'agit de caractéristiques de zones plutôt défavorisées : part de logements HLM, de ménages sans actif occupé, de ménages complexes
  - Hypothèse : il s'agirait de zones plus « difficiles » à collecter ?

# Des effets individuels connus

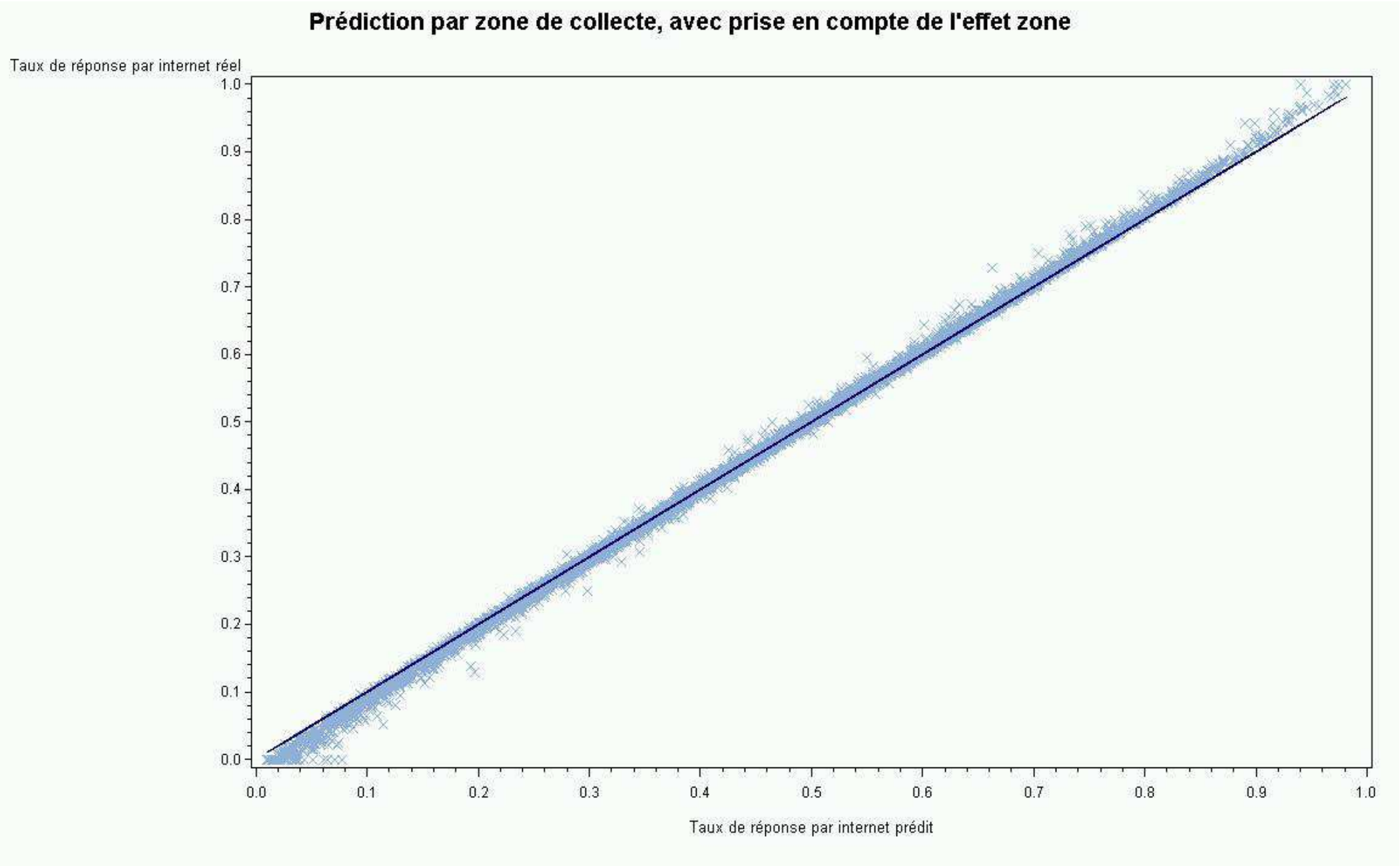
| | | Modalité de référence | Odd ratio |
|---|---|---|---|
| **Age moyen : 80 ans** | vs | Age moyen : 40 ans | 0,13 |
| **Pas de diplôme** | vs | Diplôme le plus élevé : baccalauréat | 0,42 |
| **Personne seule** | vs | Couple avec enfant | 0,53 |
| | | | |
| **Age moyen : 30 ans** | vs | Age moyen : 40 ans | 1,15 |
| **Diplôme le plus élevé : supérieur long** | vs | Diplôme le plus élevé : baccalauréat | 1,31 |

- Les diplômés et les jeunes choisissent plus souvent internet
- Les plus âgés, et sans diplôme préfèrent le papier

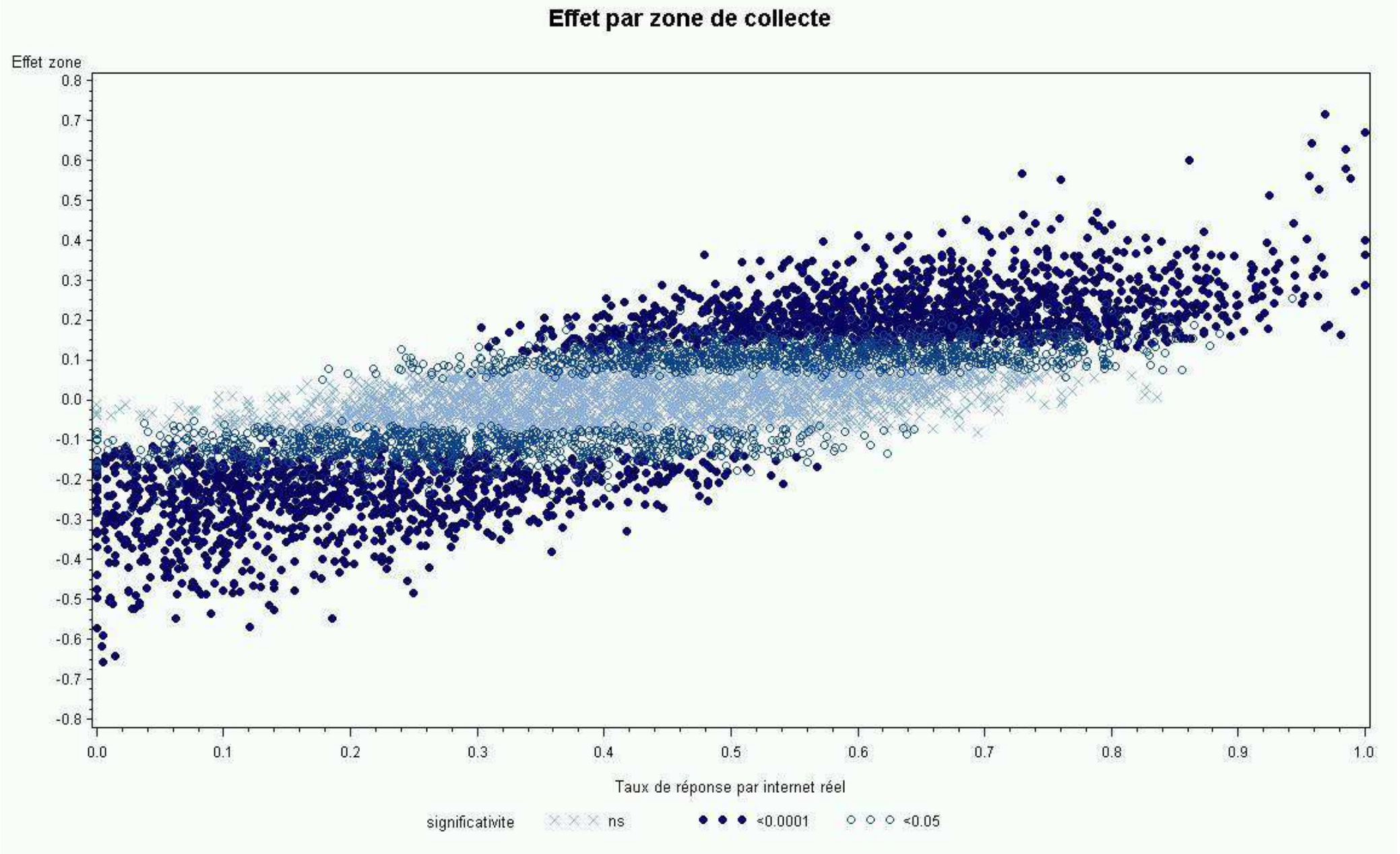
# Une prédiction … approximative



# Le modèle multi niveaux neutralise bien l'effet zone lorsqu'il est présent



# Un effet non significatif pour 38 % des zones de collecte



# Le modèle multi niveaux nous apprend :

- Qu'il y a bien un effet zone de collecte, explicatif du choix du mode de réponse
- Que cet effet est significatif pour plus de la moitié des zones de collecte
- Que cet effet est bien pris en compte en utilisant un modèle multiniveaux

# Suite des travaux

- Poursuivre l'analyse sur l'ensemble de la France pour les petites communes
- On peut alors estimer pour chaque ménage :
  - La probabilité de répondre par internet en prenant en compte l'effet zone de collecte
  - La probabilité de répondre par internet en neutralisant l'effet zone de collecte
- Cela permet d'essayer de quantifier les effets de mesure liés au mode
  - pour la non réponse partielle,
  - Ou pour le classement des individus en population municipale

# Références

- Diaporama de bilan de collecte 2017 – Pôle RP de Lyon
- Les modèles multiniveaux, Document de travail M2016/05, P. Givord et M. Guillerm

# Titre de la présentation

Merci de votre attention

Avez-vous des questions ?

**Insee**
**www.insee.fr**

**Heidi.koumarianos@insee.fr**