# Méthodes de carroyage du recensement de la population dans les communes de 10 000 habitants et plus

*JMS 2022*

**M. Chevalier (DREES), G. Gallic (DDAR), C. Guillo (DMS), G. Guymarc (DR Normandie), C. Pilorge (DSDS)**

Insee

30/03/2022

# Plan de la présentation

# Le Carroyage du RP
## contexte et enjeux

Le carroyage du recensement de la population (RP)

- Contexte : le carroyage du RP est impulsé par une demande européenne (Census 2021)
  - Livrer différentes variables à l'échelle de carreaux de 1 km de côté (France métropolitaine) : population, population selon le sexe, l'âge (moins de 15 ans, 15-64 ans, 65 ans et plus), le lieu de naissance (en France, ailleurs en UE, hors UE) et le lieu de résidence un an auparavant (inchangé, ailleurs en France, à l'étranger)
  - Champ : France métropolitaine

# Le Carroyage du RP
## contexte et enjeux

- Enjeux : estimer la population au carreau sur les différents champs du RP

Logements ordinaires

| | Communes de 10 000 hab. et plus (grandes communes) | Communes de moins de 10 000 hab. (petites communes) | Communautés | Habitations mobiles et sans-abri |
|---|---|---|---|---|
| Recensement de la population | Par **sondage** (≈ 40 % des logements sur un cycle de 5 ans) | Exhaustif au cours d'un cycle de 5 ans | | |
| Adresses géolocalisées | Oui | Oui *(Gallic et Pagès, JMS 2022)* | Oui | Non (non demandé par Eurostat) |
| Carroyage du RP | Méthode à développer | Données au carreau = somme des données individuelles (estimations du RP) | | 1 carreau « virtuel » non géolocalisé |

# Le Carroyage du RP
## contexte et enjeux

- Enjeux : estimer la population au carreau sur les différents champs du RP

Logements ordinaires

| | Communes de 10 000 hab. et plus (grandes communes) | Communes de moins de 10 000 hab. (petites communes) | Communautés | Habitations mobiles et sans-abri |
|---|---|---|---|---|
| Recensement de la population | Par **sondage** (≈ 40 % des logements sur un cycle de 5 ans) | Exhaustif au cours d'un cycle de 5 ans | | |
| Adresses géolocalisées | Oui | Oui *(Gallic et Pagès, JMS 2022)* | Oui | Non (non demandé par Eurostat) |
| Carroyage du RP | **Méthode à développer** | Données au carreau = somme des données individuelles (estimations du RP) | | 1 carreau « virtuel » non géolocalisé |

OK

# Comment carroyer les données du RP dans les GC ?

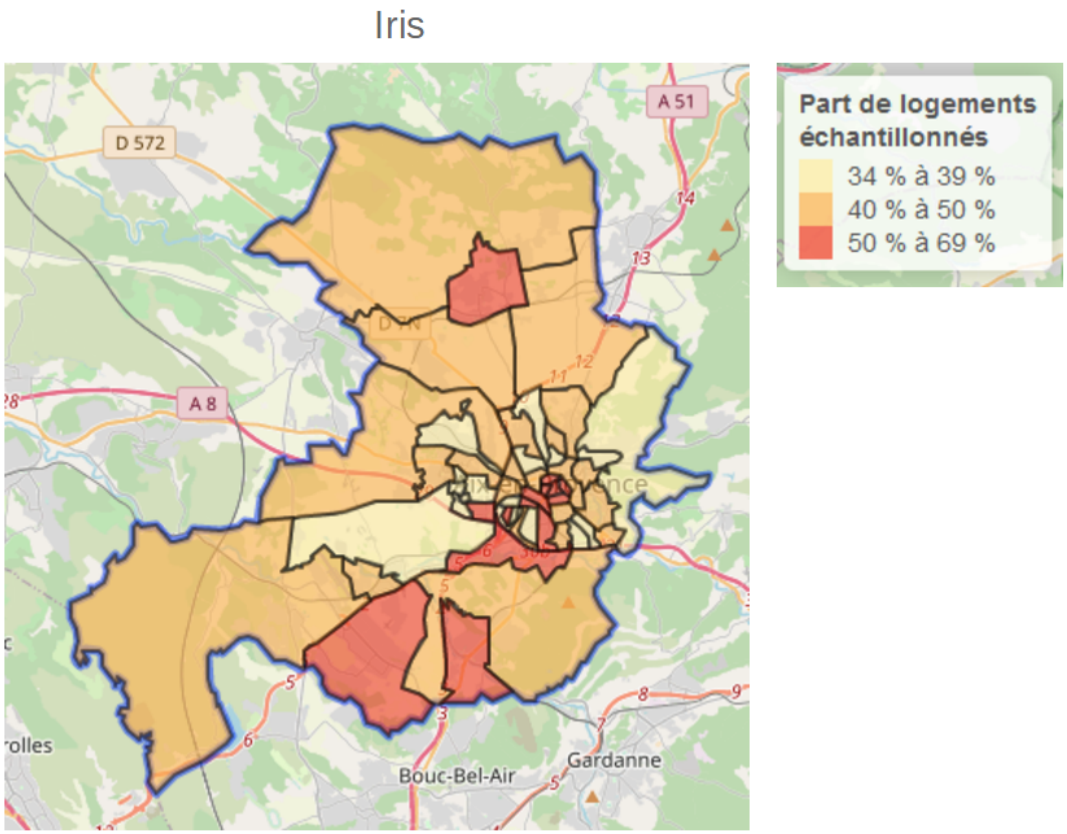
**Comment carroyer les données du RP dans les Grandes communes ?**

- Méthode « usuelle » d'estimation de la population ?
  - La population municipale est estimée à l'échelle des **Iris**.
  - Estimateur de Horvitz-Thompson : on fait la somme du nombre de personnes enquêtées à chaque adresse, pondérée par les poids de sondage.
  - Améliorer la précision : on cale sur le nombre de logements à chaque adresse, information exhaustive issue du répertoire des immeubles localisés (RIL).

# Comment carroyer les données du RP dans les GC ?

- Le tirage des adresses est équilibré sur le nombre de logements au sein de chaque Iris.

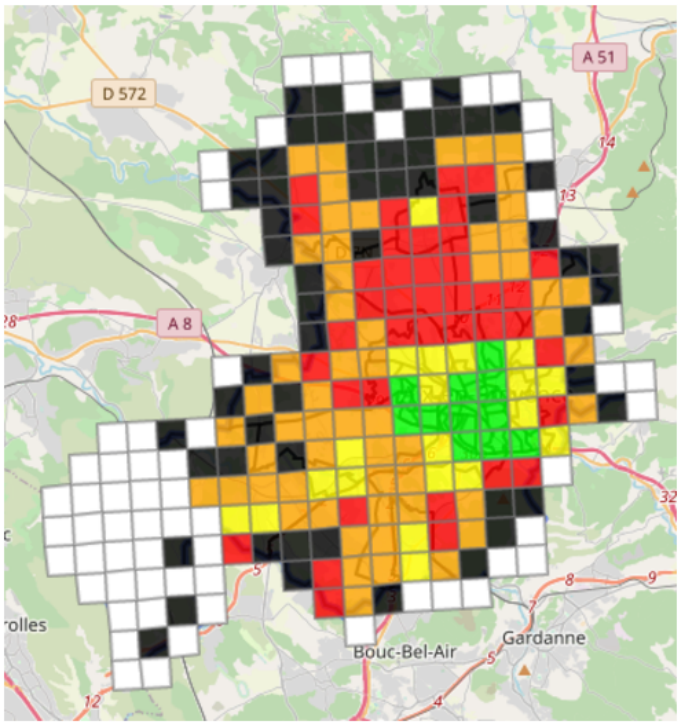
*Exemple pour la commune d'Aix en Provence :*

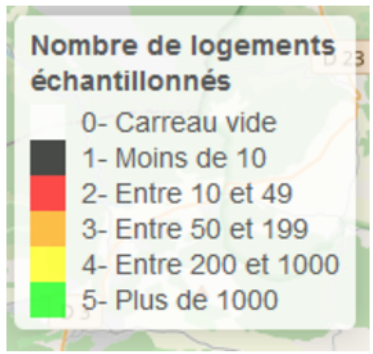


# Comment carroyer les données du RP dans les GC ?

- A l'échelle des carreaux : la méthode « usuelle » n'est pas déclinable.
- Plus d'1/3 des carreaux des GC ont moins de 10 logements échantillonnés (dont près de 30 % avec aucun logement échantillonné).

*Exemple pour la commune d'Aix en Provence :*



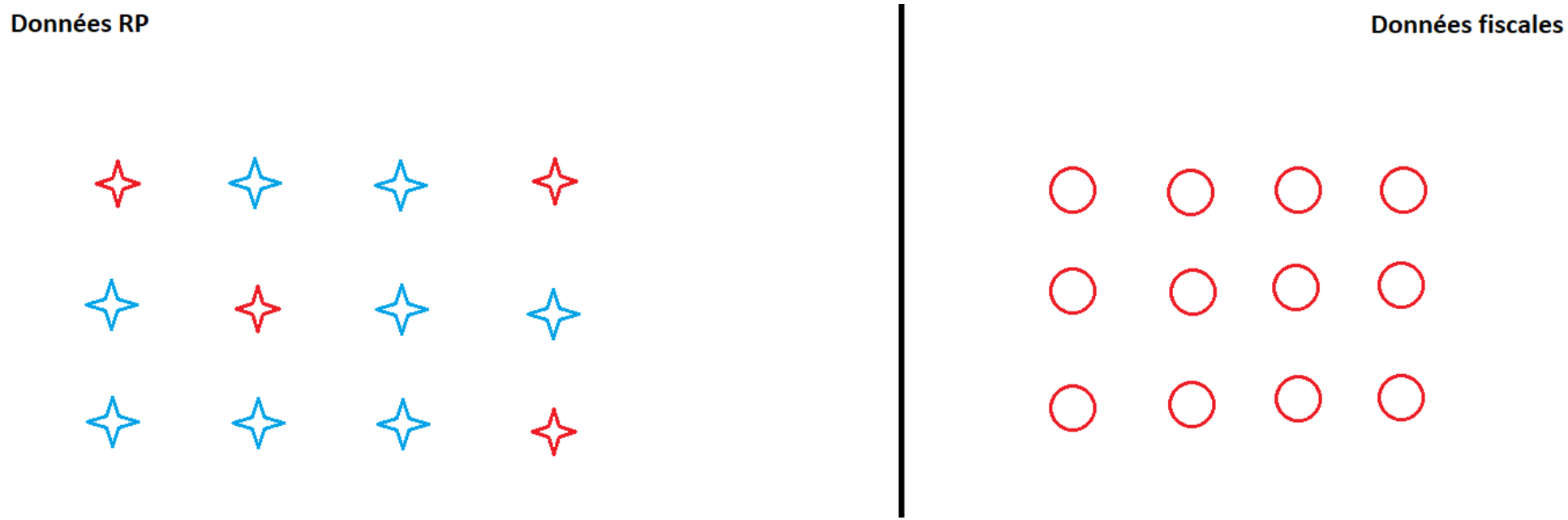

Carreaux de 1km²

# Comment carroyer les données du RP dans les GC ?

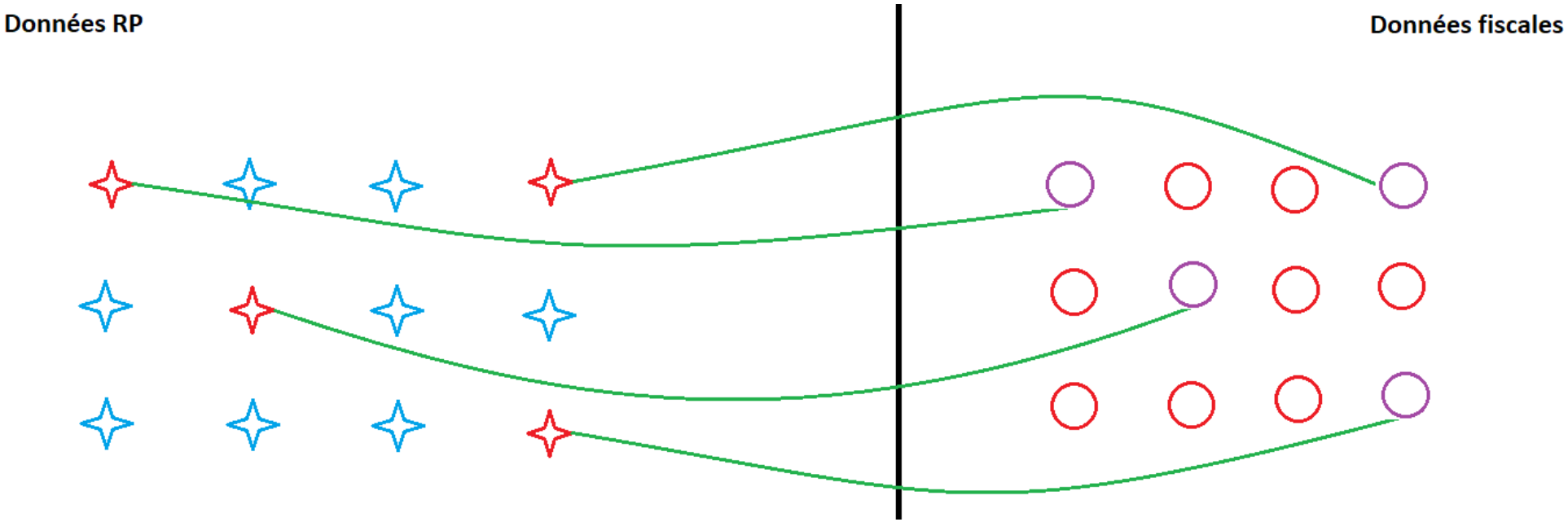
**Méthodes d'imputation de données à l'adresse :**

- *Idée* : imputer les caractéristiques de logements et d'individus à l'ensemble des adresses du RIL à partir des adresses enquêtées et d'informations auxiliaires d'origine fiscale $\longrightarrow$ *Reconstitution d'un RP exhaustif*
- *Trois méthodes étudiées* : une méthode par modélisation et deux méthodes par hot deck (une seule est présentée ici)

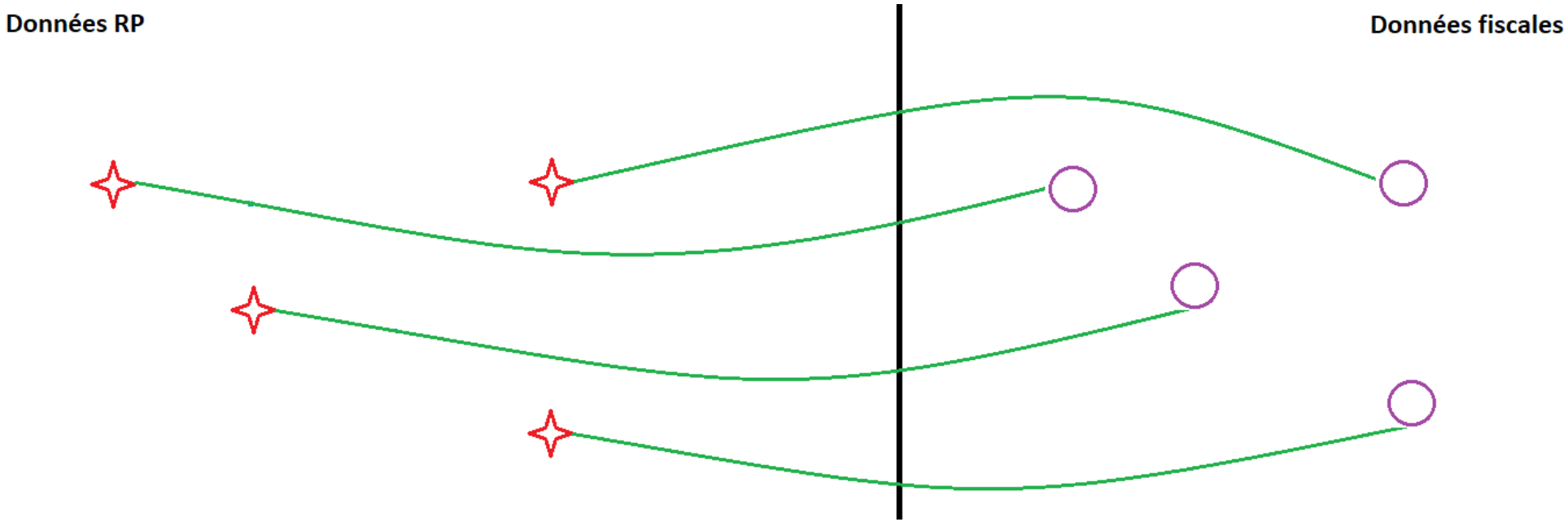
# Méthodes d'imputation : cadre général



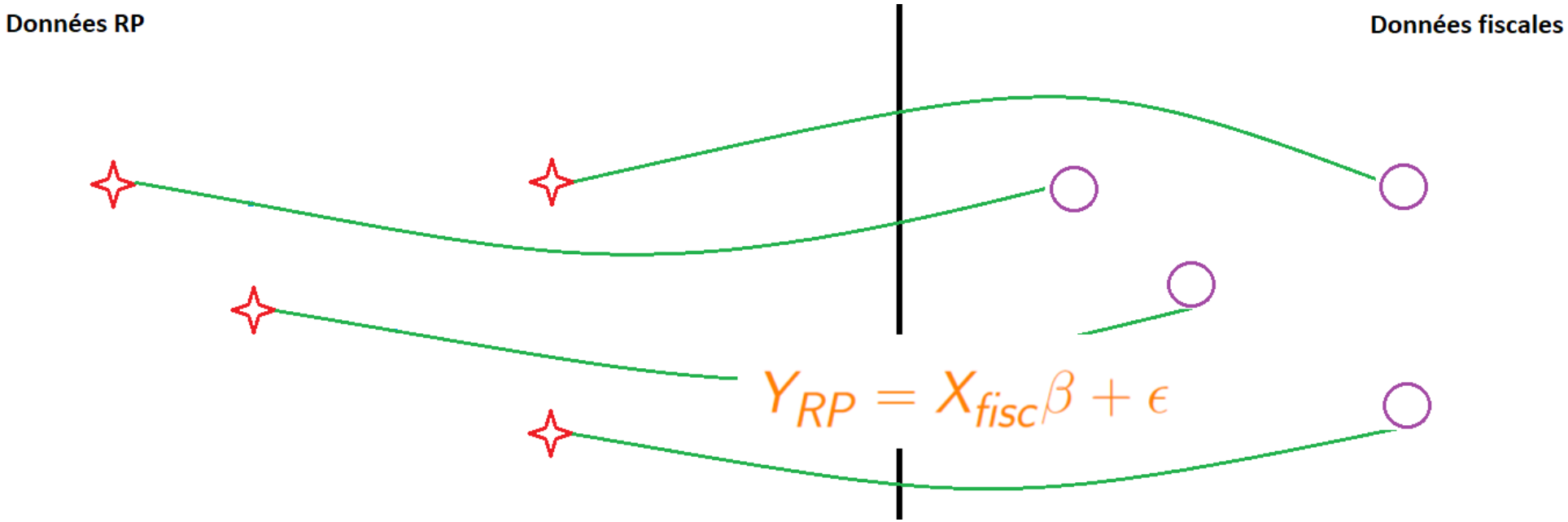
# Méthode d'imputation par modélisation : Principe



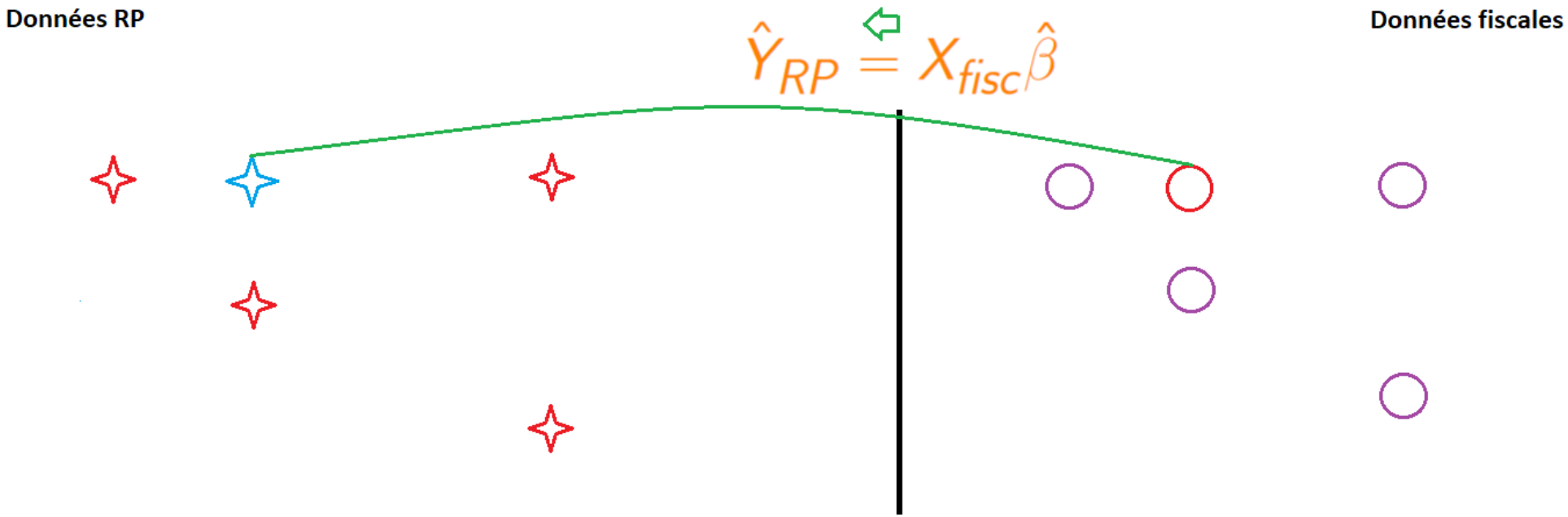
# Méthode d'imputation par modélisation : Principe



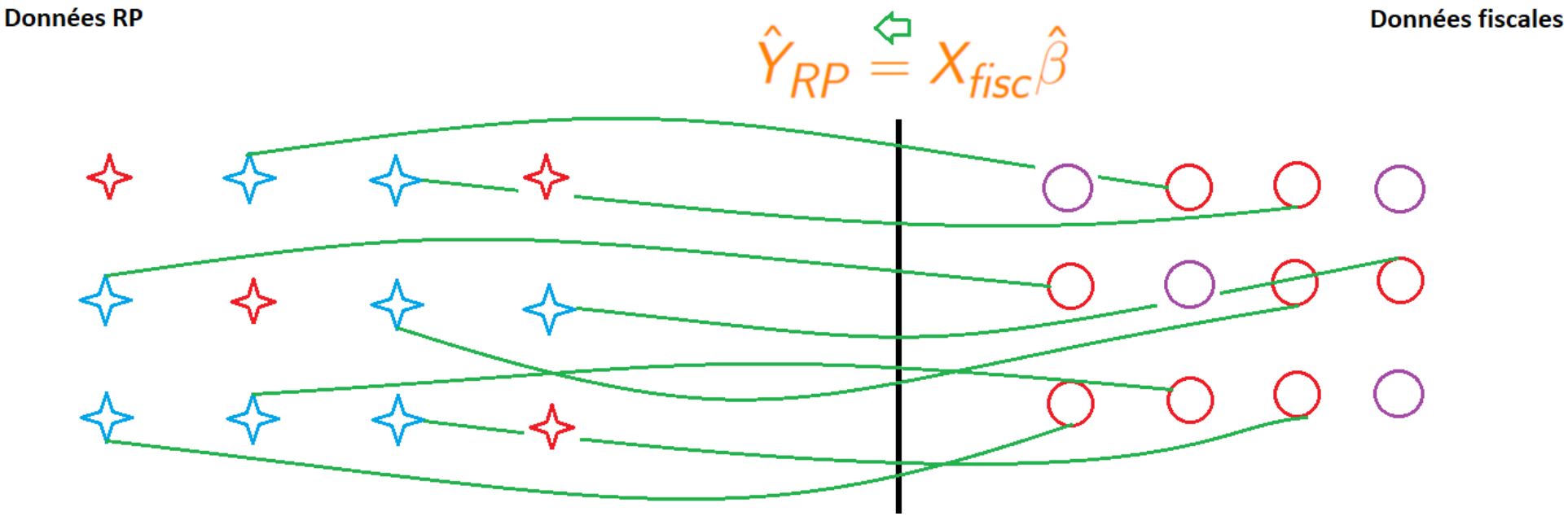
# Méthode d'imputation par modélisation : Principe



# Méthode d'imputation par modélisation : Principe



# Méthode d'imputation par modélisation : Principe



# Méthode d'imputation par modélisation : Les modèles

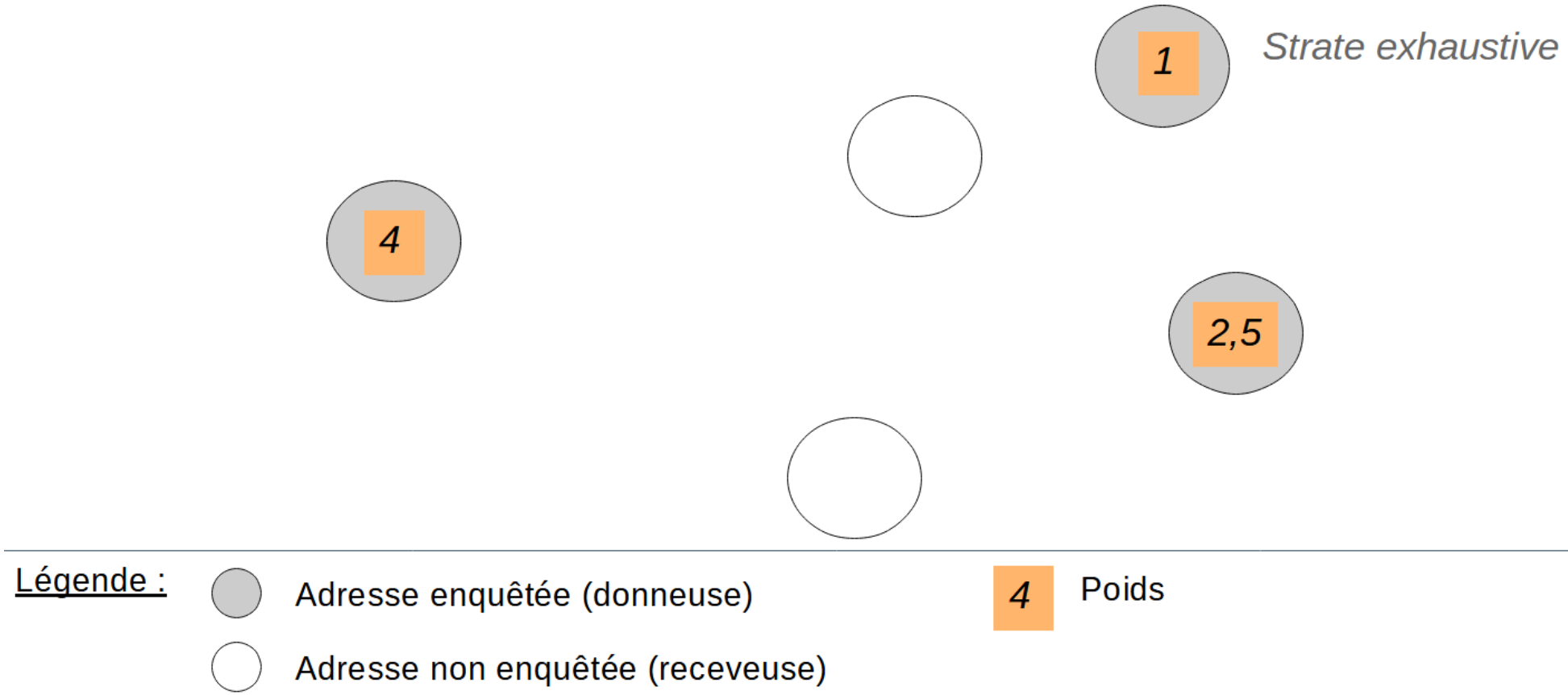
- Les modèles sont estimés commune par commune
- Deux types de modèle dépendant de la variable considérée :
  1. Pour les variables ayant leur homologue dans les données fiscales (population, population par sexe et par âge), les modèles « minimaux » sont retenus (sélection par validation croisée)
  2. Pour les autres variables d'intérêt, les modèles sont une combinaison linéaire des variables fiscales de population, pop. par âge et par sexe

# Méthode d'imputation par modélisation : mise en cohérence

- Chaque variable d'intérêt est estimée **indépendamment** : rien ne garantit une cohérence entre les variables au niveau de l'adresse
  $\longrightarrow$ étape de mise en cohérence :
  - *En amont des estimations* : description de toutes les relations d'ensemble entre les variables à estimer (ex. : nbre d'hommes + nbre de femmes = population totale)
  - *En aval* : calcul de la structure induite par les estimations pour chaque variable au niveau de l'adresse et application à l'estimation de population

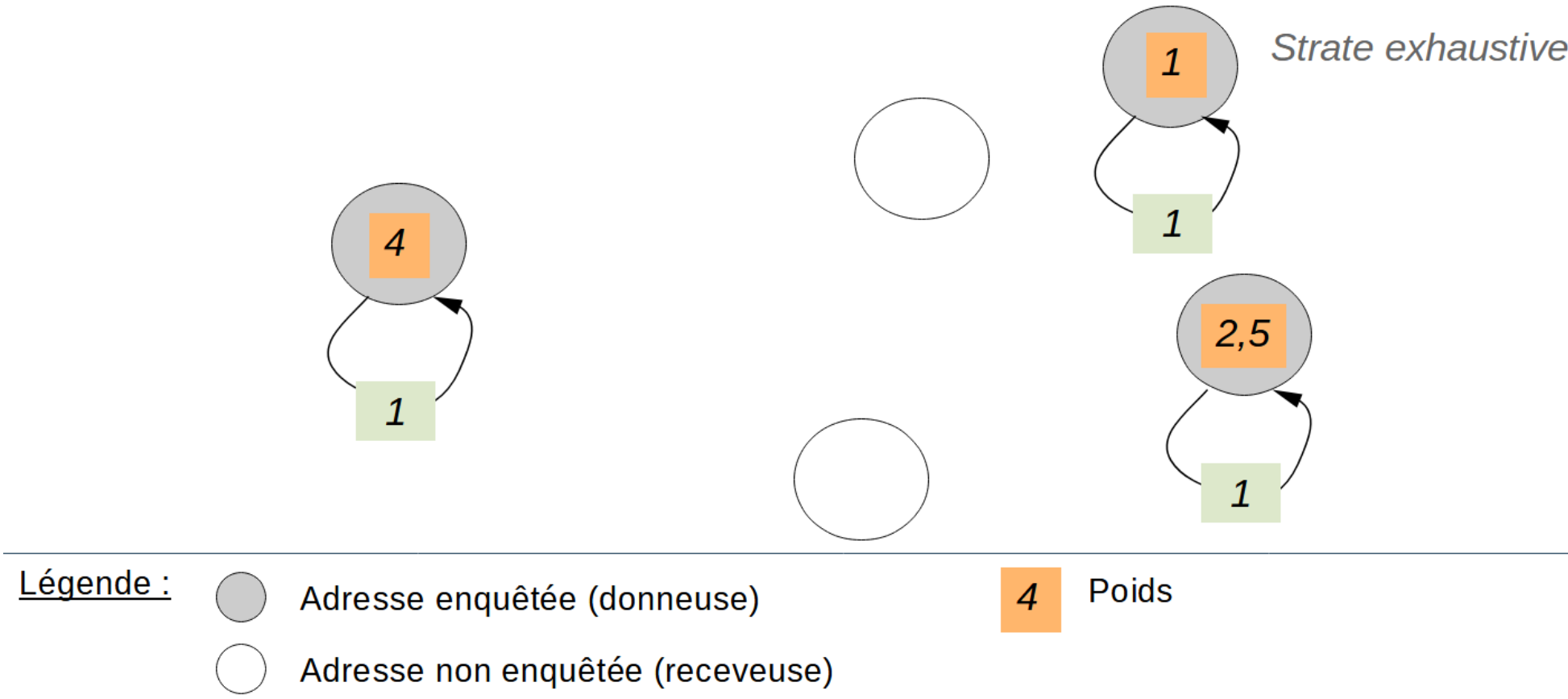
# Méthode par *HOT DECK*
## Schéma de la méthode

**Principe** : on distribue les caractéristiques des adresses enquêtées aux adresses non-enquêtées de sorte à reconstituer un RP exhaustif.



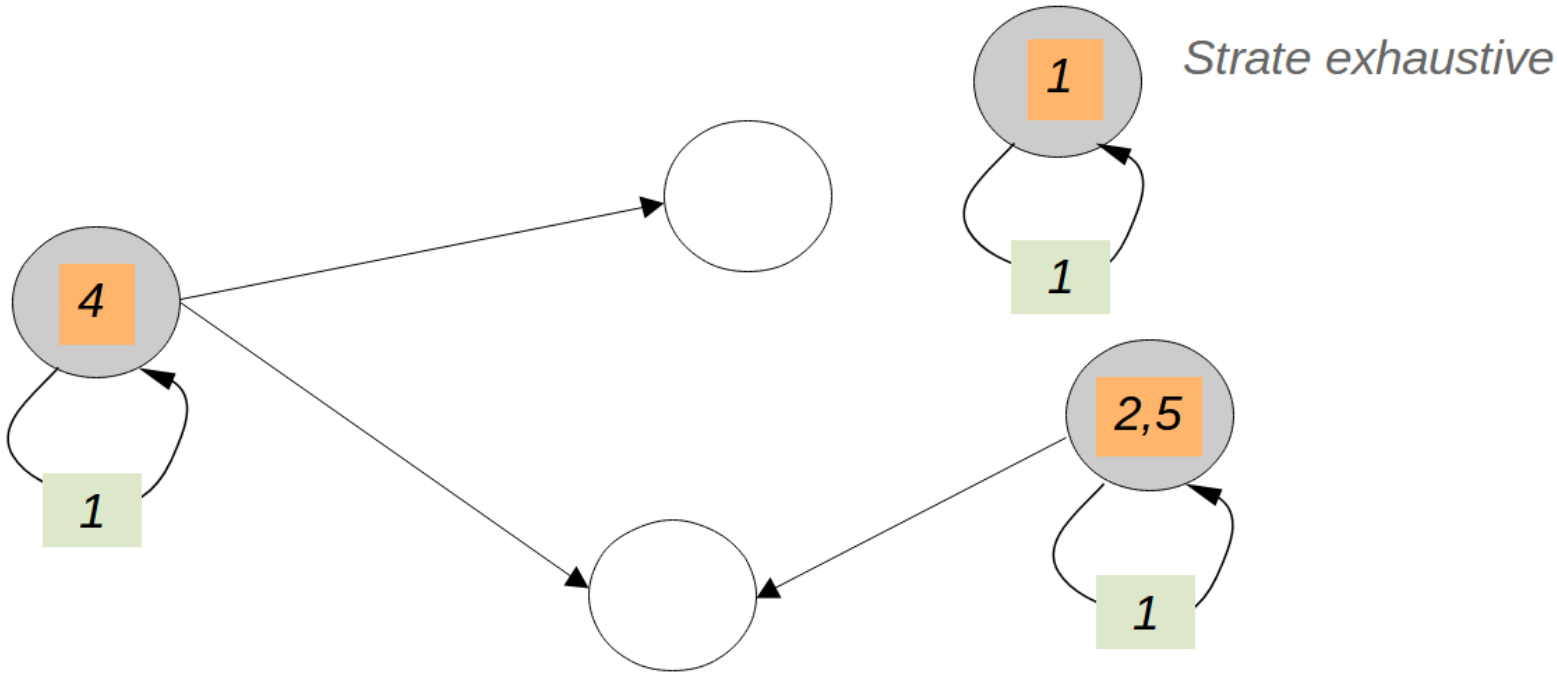
# Méthode par *HOT DECK*
## Schéma de la méthode

- Les **adresses enquêtées** distribuent leur poids :
  - Elles se donnent à elles-mêmes pour un poids de 1 (donneurs "stables")



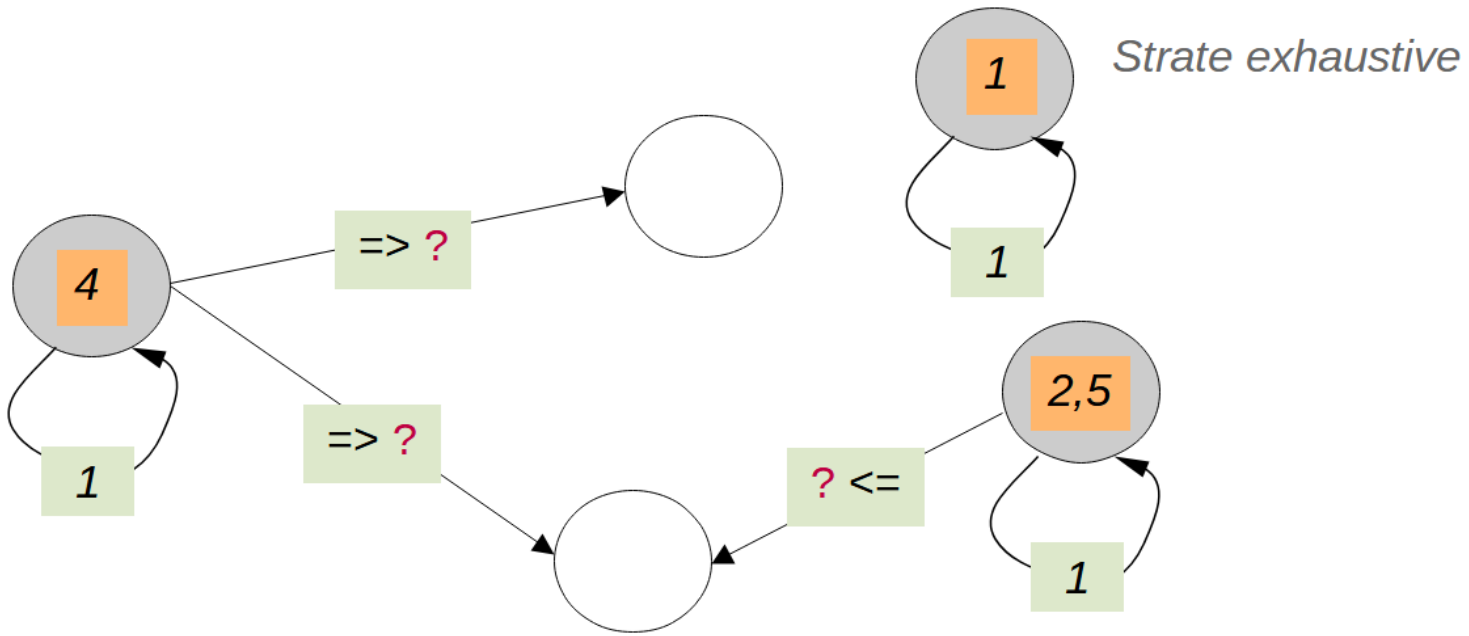
# Méthode par *HOT DECK*
## Schéma de la méthode

- Elles donnent le reste de leur poids à une ou plusieurs adresses non enquêtées, au sein du même **Iris**



# Méthode par *HOT DECK*
## Schéma de la méthode

- Les **adresses non enquêtées** reçoivent une fraction du poids d'une ou plusieurs adresses donneuses.
  $\longrightarrow$ **Comment ces poids vont-ils être répartis ?**




Légende :

- Adresse enquêtée (donneuse)
- Adresse non enquêtée (receveuse)
- 4 Poids
- => ? Poids distribué par l'adresse donneuse

# Méthode par *HOT DECK*
# Programme d'optimisation

**Programme de minimisation**

$$\min_{\lambda_{d,r}} \sum_{d,r} \lambda_{d,r} \overbrace{dist_{d,r}}^{\text{distance entre 2 adresses}}$$

→ Calcul à partir des données fiscales (+ distance géographique)

# Méthode par *HOT DECK*
# Programme d'optimisation

## Programme de minimisation

distance entre 2 adresses

$$\min_{\lambda_{d,r}} \sum_{d,r} \lambda_{d,r} dist_{d,r}$$

→ Calcul à partir des données fiscales (+ distance géographique)

indicateur d'appariement

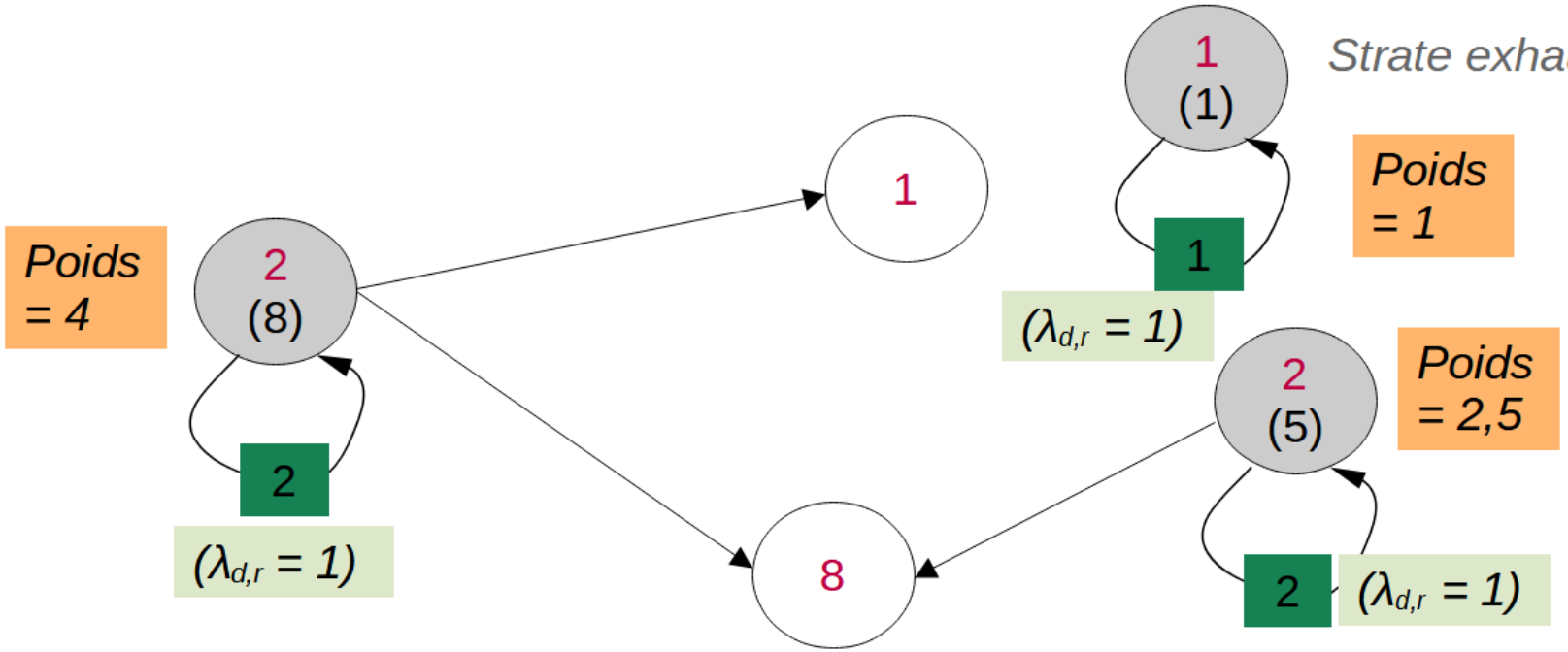
$$\text{s.c.} \quad \forall d \in L^{(D)} \sum_{r \in L^{(R)}} \lambda_{d,r} \leq w_d$$

Poids d'estimation de l'adresse

$$\forall r \in L^{(R)} \sum_{d \in L^{(D)}} \text{nb log RIL}_d \lambda_{d,r} = \text{nb log RIL}_r$$

$L^{(D)}$ est l'ensemble des adresses donneuses, $L^{(R)}$ l'ensemble des adresses receveuses.
*nb log RIL* = quantité cible à atteindre en « distribuant » les poids des adresses donneuses.

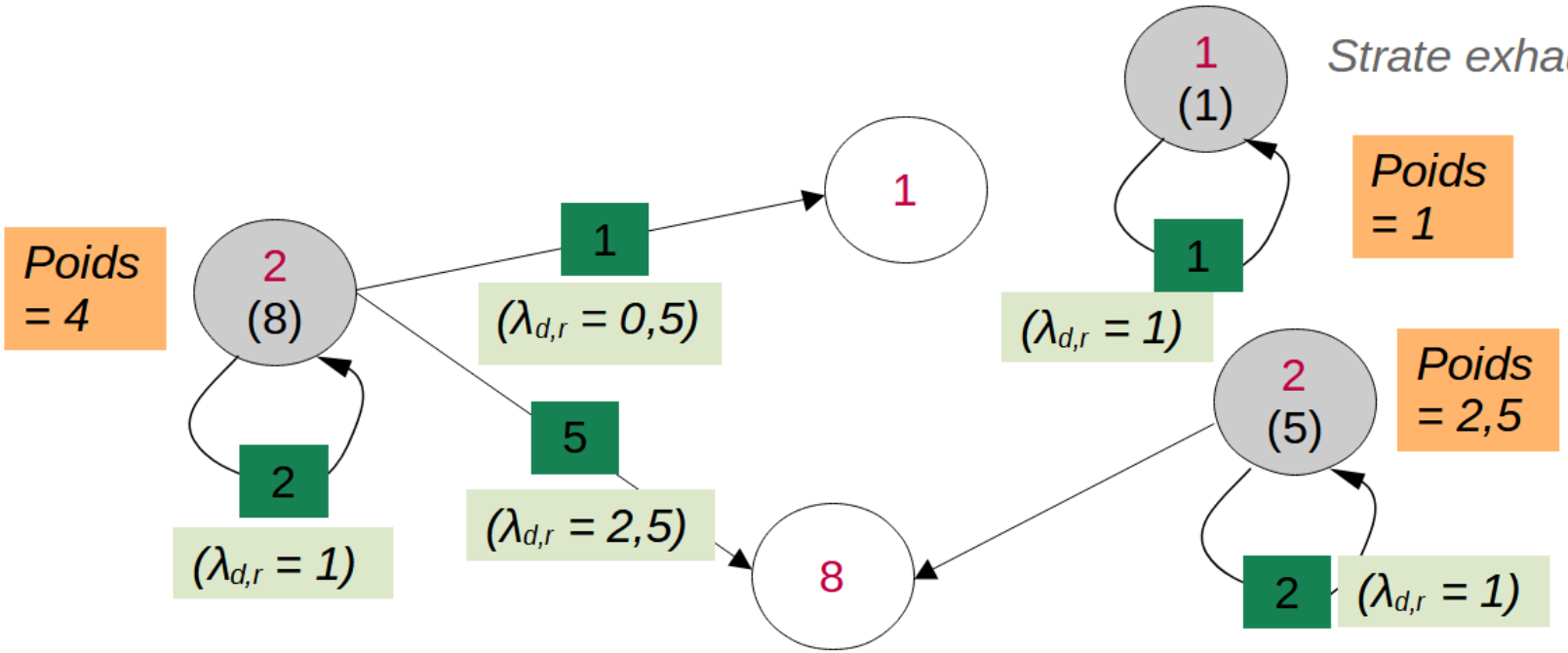
# Méthode par *HOT DECK*
## Schéma de la méthode avec cible de logements




Légende :

- Adresse donneuse
- Adresse receveuse
- 2 — Nbre de log donnés / reçus
- 2 — Nombre de log au RIL médian
- (8) — Nombre de log pondérés
- $\lambda_{d,r}$ — Indicateur d'appariement

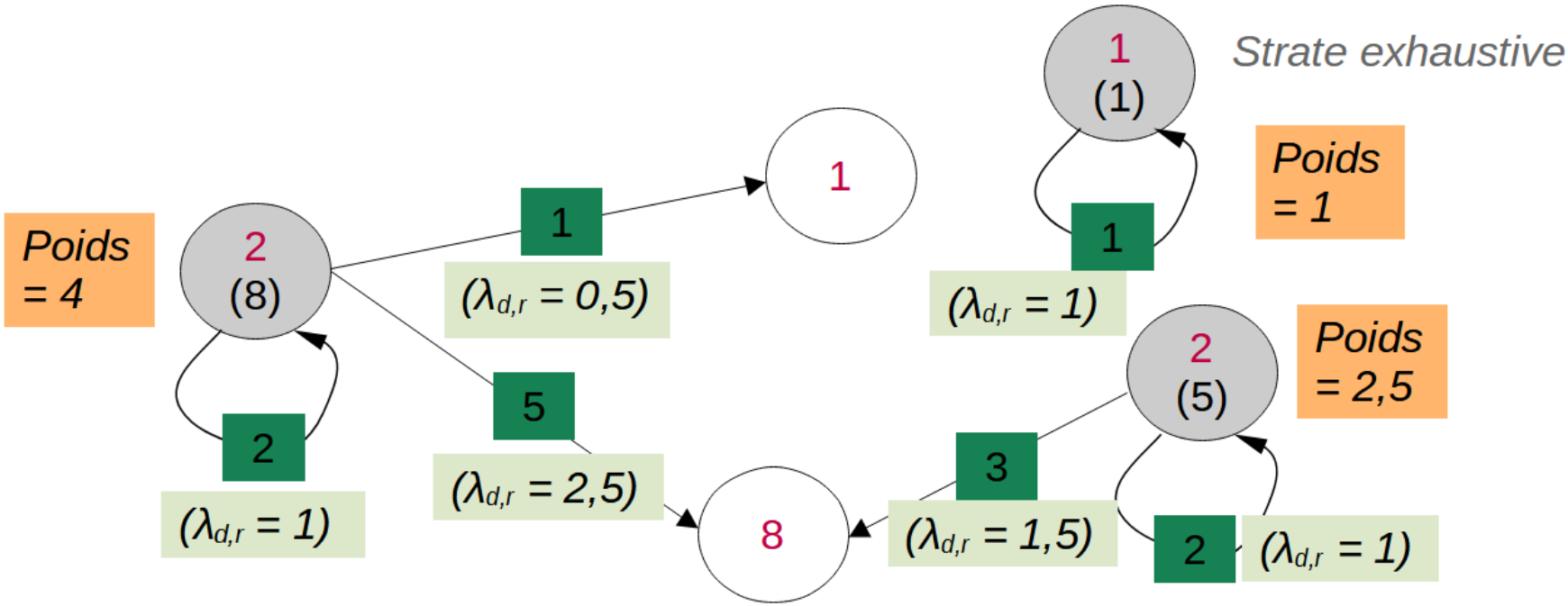
# Méthode par *HOT DECK*
## Schéma de la méthode avec cible de logements




Légende :

- Adresse donneuse
- Adresse receveuse
- 2 — Nbre de log donnés / reçus
- 2 — Nombre de log au RIL médian
- (8) — Nombre de log pondérés
- $\lambda_{d,r}$ — Indicateur d'appariement

# Méthode par *HOT DECK*
## Schéma de la méthode avec cible de logements




Légende :

- Adresse donneuse
- 2 — Nombre de log au RIL médian
- Adresse receveuse
- (8) — Nombre de log pondérés
- 2 — Nbre de log donnés / reçus
- $\lambda_{d,r}$ — Indicateur d'appariement

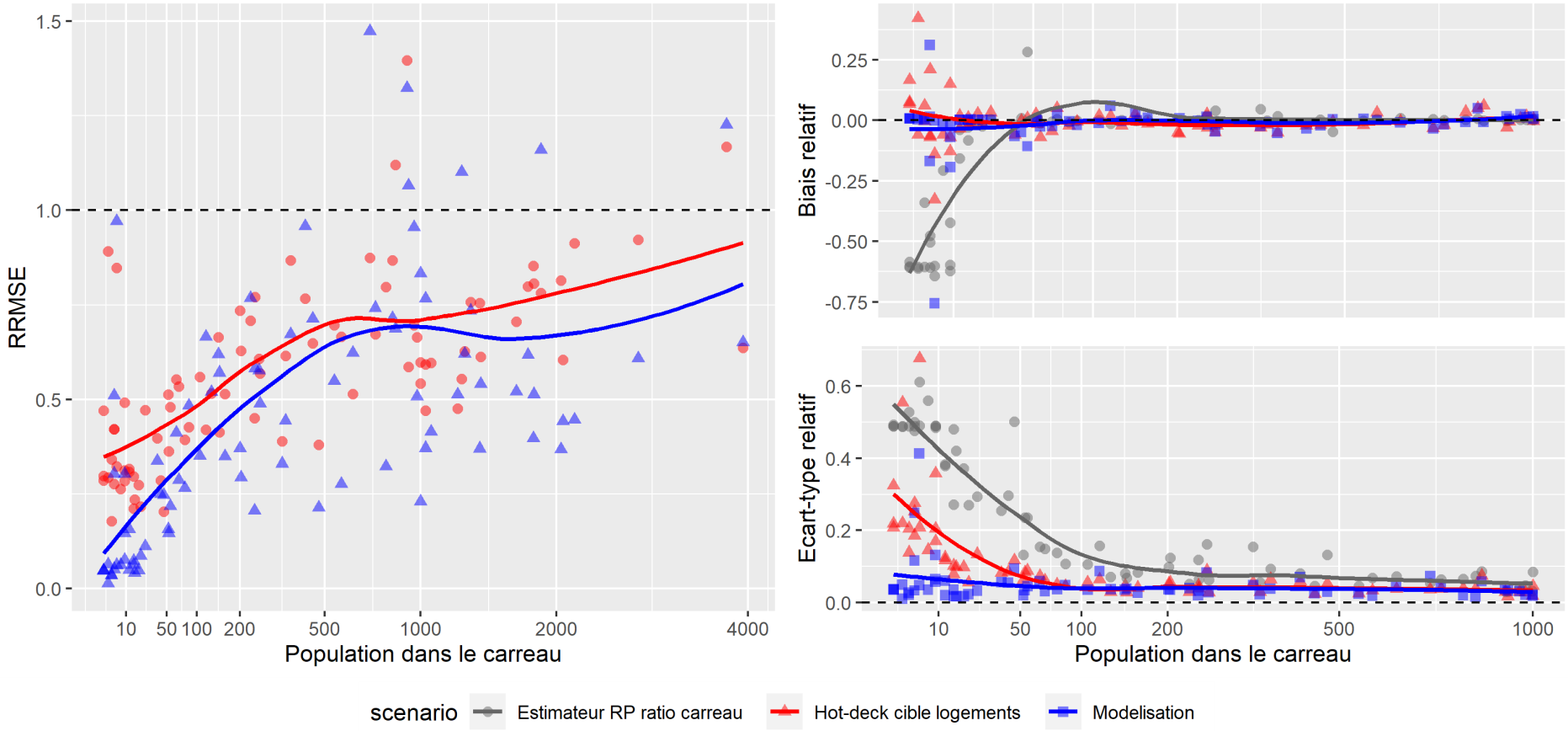
# Résultats : comparaison de la précision

**Comparaison de la précision des méthodes**

- Évaluation menée sur des communes avec un recensement exhaustif récent
  - Tirage d'un grand nombre d'échantillons d'adresses pour simuler les estimations au carreau, à partir de nos méthodes de carroyage et pour l'estimateur usuel du RP
  - Comparaison des estimations aux vraies valeurs
- Les deux méthodes permettent d'estimer la population au carreau de manière plus fiable que ne le fait l'estimateur par le ratio appliqué au carreau.
- La méthode par **modélisation** permet d'avoir les estimations les plus précises.

# Résultats : comparaison de la précision

## Exemple : précision obtenue pour l'estimation de la population



# Résultats : comparaison de la cohérence avec le RP

**Comparaison de la cohérence avec les données diffusées**

- La méthode *hot deck* avec cible de logements permet d'assurer une cohérence parfaite pour l'ensemble des variables, aux niveaux Iris et *supra*.
- La méthode par modélisation n'assure que la cohérence à l'échelle communale du nombre de logements au RIL et de la population légale, après calage.
  $\longrightarrow$ La cohérence avec les estimations à l'Iris (ou aux QPV) impose trop de contraintes

# Résultats : arbitrage entre les méthodes

## Quelle méthode retenir ?

| Critères de sélection | Méthode d'imputation par modélisation | Méthode d'imputation par hot deck avec cible de logements |
|---|---|---|
| 1. **Cohérence** avec les données diffusées du RP | Au niveau de la commune (nbre de logements au RIL et population) | **Totale au niveau Iris** |
| 2. **Précision** | **+++++** | +++(+) |
| 3. **Mise en production** | **L'ajout de variables supplémentaires peut être complexe** | **Toutes les variables du RP** sont **disponibles** |

- $\longrightarrow$ La cohérence est privilégiée à la précision des estimations.
- La facilité d'estimation de nouvelles variables au carreau est également prise en compte, ainsi que la moindre adhérence aux données fiscales.

# Travaux à venir

**Travaux à venir :**

- Mieux documenter la performance des méthodes : évaluation sur un plus grand nombre de communes
- Traitement de la confidentialité
- Extension du carroyage aux DOM ? Diffusion d'autres variables ? Tests sur des carreaux de 200m de côté ?

# Merci pour votre attention !

# Annexe
# Méthode par *HOT DECK*

**Illusatrion de la distribution des poids pour une sélection de 4 adresses au sein d'un Iris :**

## *Adresses donneuses*

| Adr | Com | Iris | Carreau | Nb log RIL médian | Poids | Pop | Nb F | Nb H | < 15 ans | 15-64 ans | 65 ans et + |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 01004 | 0201 | 03 | 2 | 4 | 4 | 3 | 1 | 2 | 1 | 1 |
| 9 | 01004 | 0201 | 04 | 2 | 2,5 | 3 | 1 | 2 | 0 | 3 | 0 |

## *Adresses du RIL*

| Adr | Com | Iris | Carreau | Nb log RIL médian | $\lambda_{d,r}$ | Pop | Nb F | Nb H | < 15 ans | 15-64 ans | 65 ans et + |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 01004 | 0201 | 02 | 2 | 1 | 4 | 3 | 1 | 2 | 1 | 1 |
| 2 | 01004 | 0201 | 03 | 1 | 0,5 | 2 | 1,5 | 0,5 | 1 | 0,5 | 0,5 |
| 3 | 01004 | 0201 | 02 | 8 | 2,5 | 10 | 7,5 | 2,5 | 5 | 2,5 | 2,5 |
| 3 | 01004 | 0201 | 02 | | 1,5 | 4,5 | 1,5 | 3 | 0 | 4,5 | 0 |
| 9 | 01004 | 0201 | 04 | 2 | 1 | 3 | 1 | 2 | 0 | 3 | 0 |