



Etude de l'impact des données manquantes sur l'analyse des variations spatiales des activités de pêche

Eva Dieudonné (LETG-Nantes) et Lise Bellanger (LMJL)

Session 1 : Caractériser



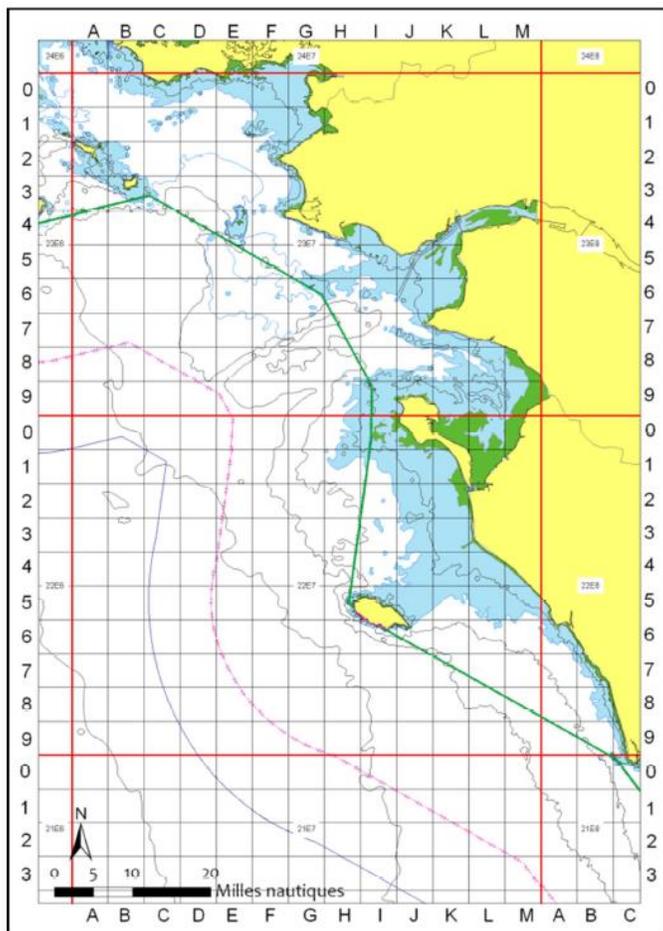
1. Introduction

Plan de la présentation :

1. Introduction
2. Les données Valpena : présentation générale
3. Le biais de non-réponse : origines et conséquences
4. Illustration du biais de non-réponse
5. Recommandations

- Les données de spatialisation des activités de pêche utilisées dans le cadre de la planification spatiale maritime ne sont pas épargnées par la problématique des **données manquantes**.
- Ignorer les données manquantes lors de l'analyse spatiale des activités de pêche peut conduire à des **résultats biaisés**
- A partir de la méthodologie développée par (Arbia *et al.*, 2016) et de l'exemple des **données Valpena**, l'objectif de cette présentation est de **sensibiliser** les utilisateurs de données de spatialisation des activités de pêche aux conséquences possibles de la présence de données manquantes sur les analyses

2. Les données Valpena : présentation générale



Extrait du maillage Valpena

- Objectif général du GIS Valpena : Proposer une méthodologie originale de collecte et de traitement de données de **spatialisation des activités de pêche maritime professionnelle**
- Protocole d'enquête : (Trouillet *et al.*, 2019) :
 - Entretiens semi-directifs auprès des pêcheurs
 - Renseignement pour chaque navire de la flotte des quadruplets (maille, mois, engin, espece) travaillés au cours de l'année
 - Campagnes d'enquêtes annuelles (alternance campagnes exhaustives et en plan d'échantillonnage)

Données fictives Valpena

| immatriculation | annee | mois | maille | engin | espece cible | cause non travail |
|-----------------|-------|------|----------|-------|--------------|-------------------|
| 123456 | 2022 | 1 | 2723E6B8 | OTB | SOL | |
| 123456 | 2022 | 2 | 2723E6B8 | OTB | SOL | |
| 123456 | 2022 | 2 | 2723E6B9 | OTB | SOL | |
| 123456 | 2022 | 3 | | | | AT |
| 456789 | 2022 | 1 | 2723F5A0 | LLS | BSS | |
| 456789 | 2022 | 1 | 2723F5A0 | GTR | BRB | |

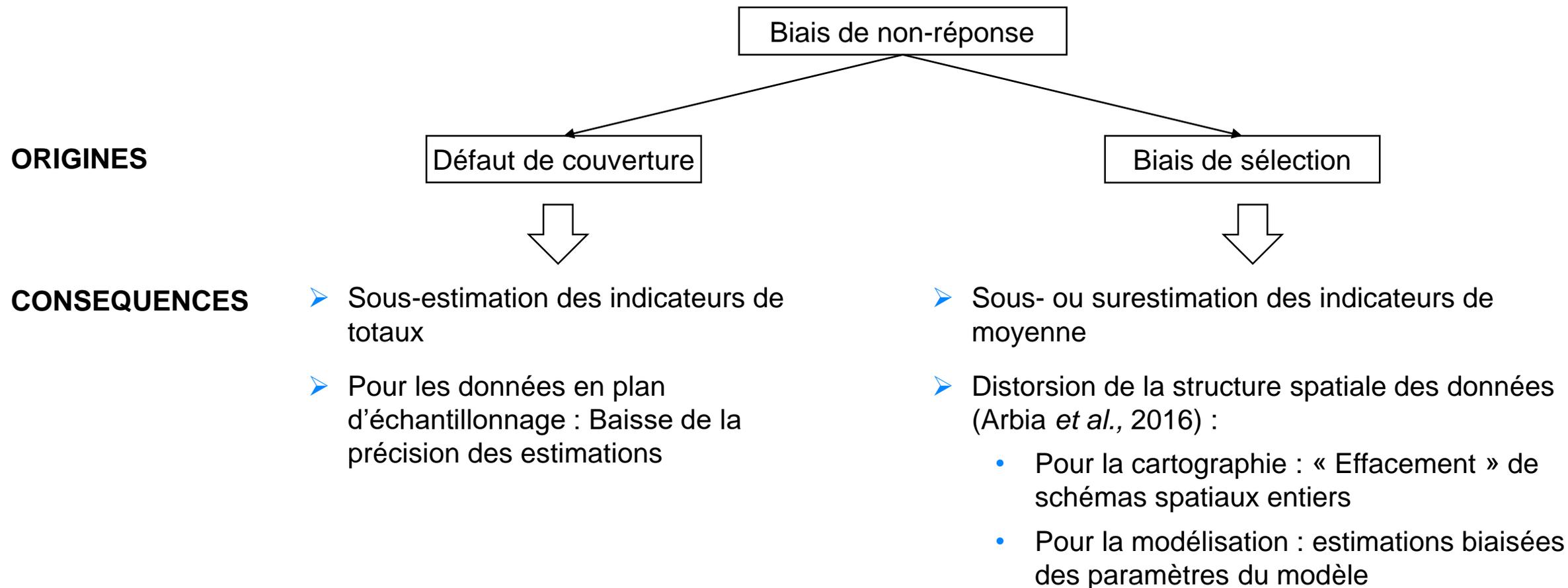
2. Les données Valpena : présentation générale

- Les données manquantes dans les enquêtes Valpena :
 - La présence de données manquantes dans les enquêtes Valpena est principalement lié au **refus** d'un pêcheur de répondre à l'enquête (taux de non-réponse moyen : 21%)
 - Il s'agit de **non-réponses totales** : aucune donnée sur l'activité du navire n'a été collectée

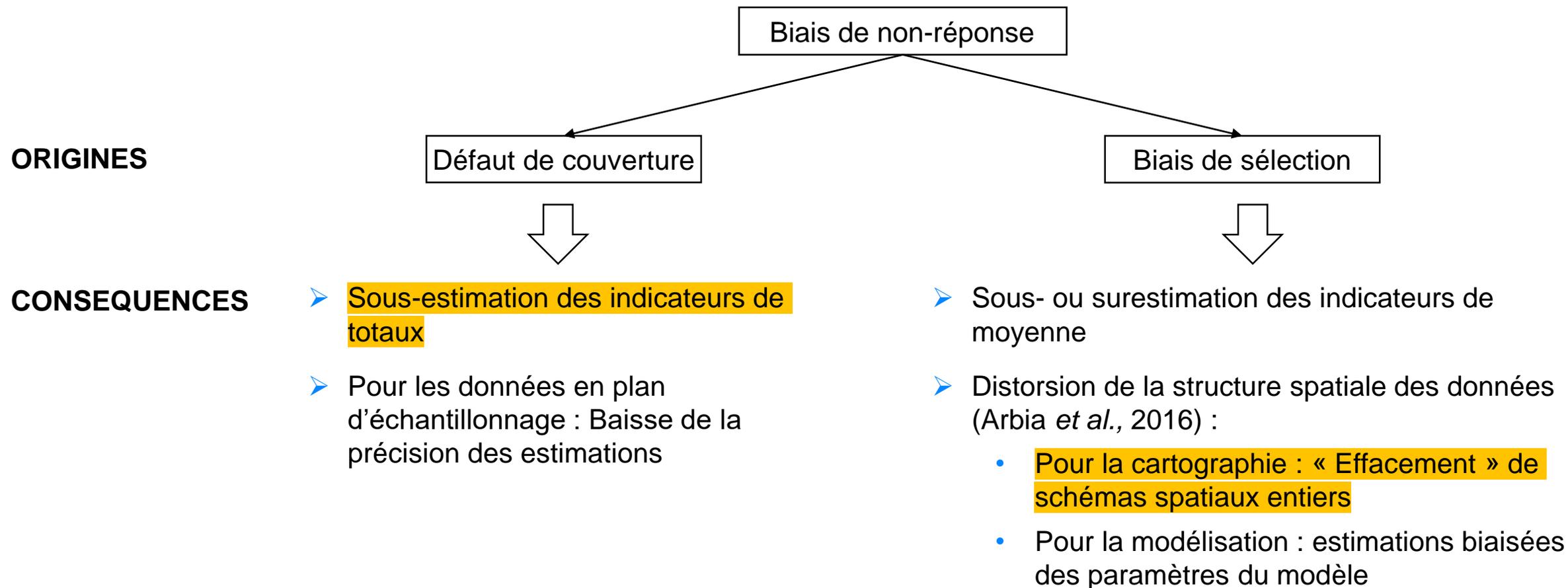
Exemple de données manquantes dans les enquêtes Valpena

| navire | mois | engin | espece | Maille A | Maille B | Maille C | ... | Maille K |
|--------|------|-------|--------|----------|----------|----------|-----|----------|
| 123456 | 1 | OTB | SOL | 0 | 1 | 1 | ... | 0 |
| 123456 | 2 | OTB | SOL | 0 | 0 | 1 | ... | 1 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 456789 | 11 | OTB | SOL | NA | NA | NA | NA | NA |
| 456789 | 12 | OTB | SOL | NA | NA | NA | NA | NA |

3. Le biais de non-réponse : origines et conséquences



3. Le biais de non-réponse : origines et conséquences



4. Illustration du biais de non-réponse

- Méthode : Simulations de Monte Carlo

A partir d'un jeu de données complet :

1. Génération de données manquantes

2. Calcul de l'indicateur d'intensité $\hat{II}_k = \sum_{i=1}^N \sum_{m=1}^{MO} \mathbb{I}_{nav_i \times maille_k \times mois_m}$

3. On répète 10 000 fois les étapes 1 et 2

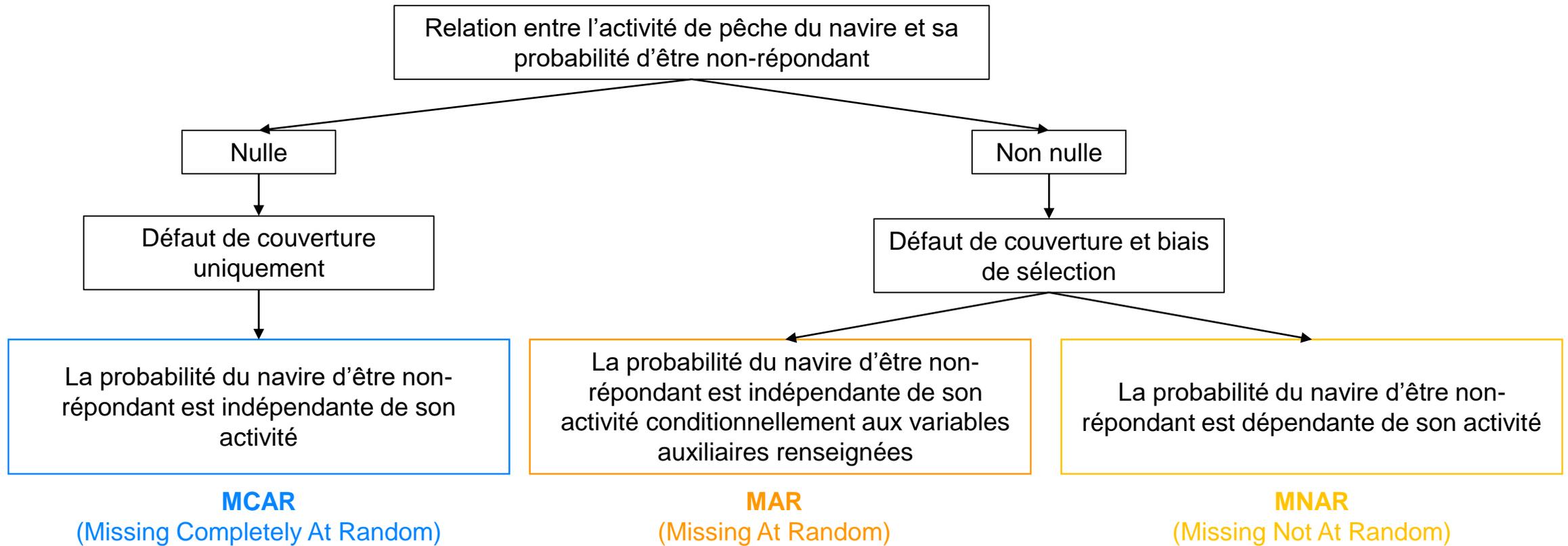
On calcule ensuite la valeur moyenne \bar{II}_k sur l'ensemble des simulations : $\bar{II}_k = \frac{1}{10000} \sum_{i=1}^{10000} \hat{II}_{k,i}$

- Calcul du biais :
$$PB = 100 \times \frac{|\bar{II}_k - II_k|}{II_k}$$

➤ **Objectif** : Quantifier et analyser la répartition spatiale du biais

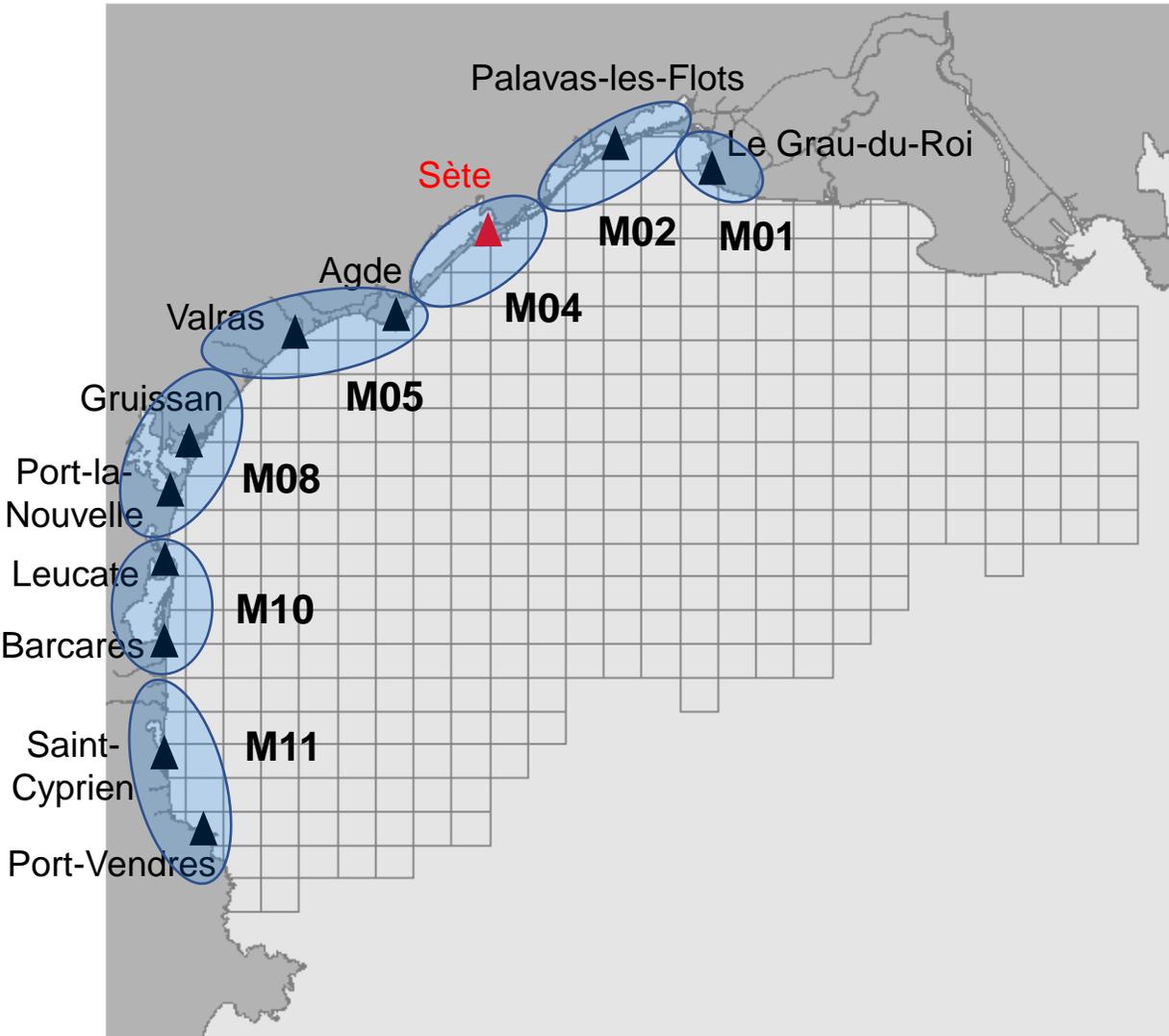
4. Illustration du biais de non-réponse

- Scénarios : 15 scénarios
 - 5 taux de non-réponse (10%, 20%, 30%, 40% et 50%)
 - 3 mécanismes de non-réponse (MCAR, MAR et MNAR)



Mécanismes de non-réponse (Rubin, 1976)

4. Illustration du biais de non-réponse



➤ **MAR** : Plus le port d'attache du navire est loin du siège du comité (Sète) et plus sa probabilité d'être non-répondant est grande

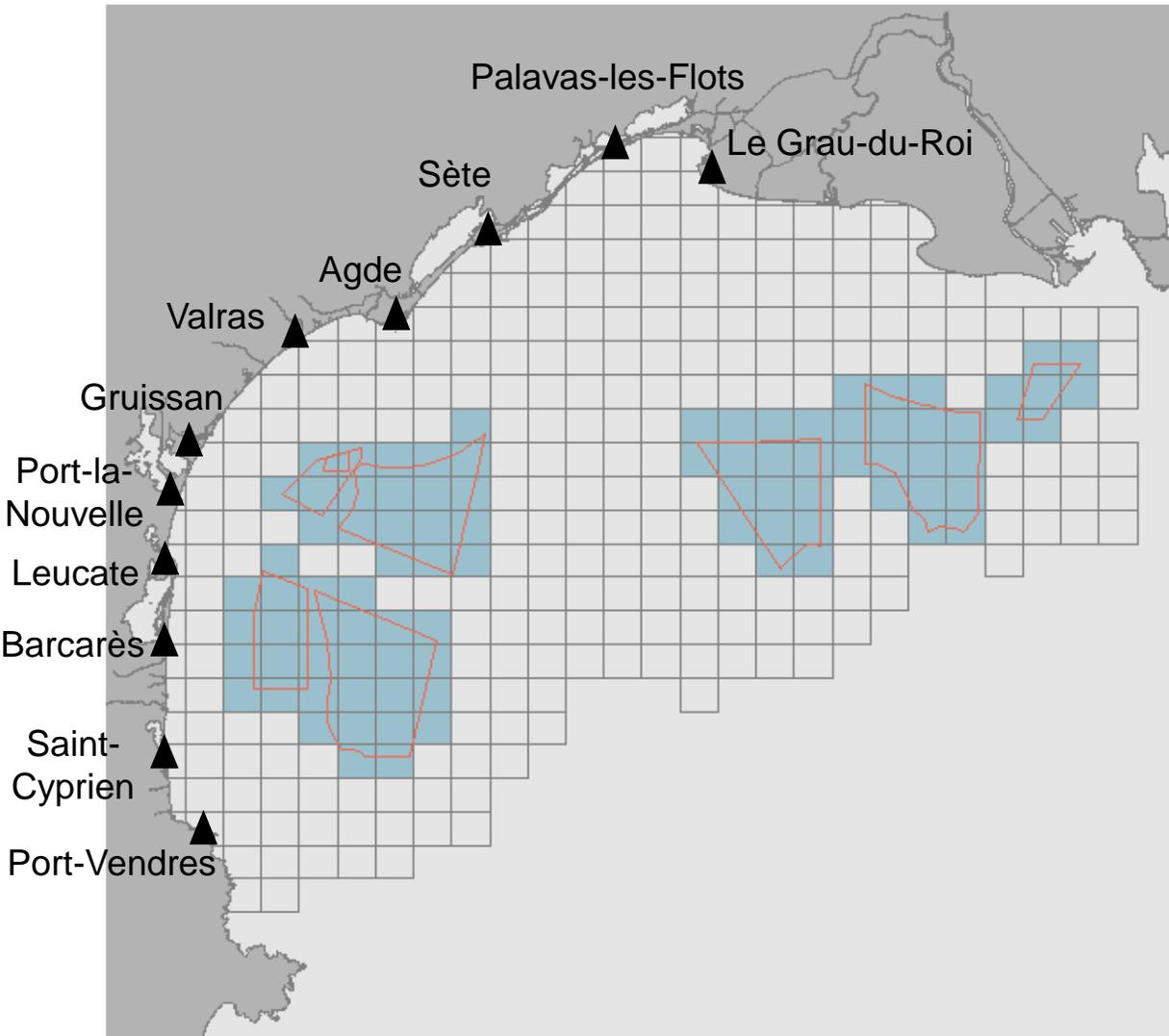
$$p_i = 1 - \frac{e^{\alpha + \beta d_i}}{1 + e^{\alpha + \beta d_i}}$$

Avec $\alpha = \begin{cases} 3.56 \\ 2.64 \\ 2 \\ 1.475 \\ 1 \end{cases}$ et $\beta = -0.25$

| | 10% | 20% | 30% | 40% | 50% |
|-----|------|------|------|------|------|
| M01 | 0.07 | 0.16 | 0.27 | 0.38 | 0.50 |
| M02 | 0.05 | 0.12 | 0.20 | 0.30 | 0.41 |
| M04 | 0.03 | 0.07 | 0.13 | 0.21 | 0.29 |
| M05 | 0.04 | 0.11 | 0.18 | 0.27 | 0.38 |
| M08 | 0.13 | 0.27 | 0.41 | 0.54 | 0.65 |
| M10 | 0.21 | 0.40 | 0.56 | 0.68 | 0.78 |
| M11 | 0.31 | 0.53 | 0.68 | 0.78 | 0.85 |

Probabilité de non-réponse du navire en fonction du taux de non-réponse du scénario et du port d'attache

4. Illustration du biais de non-réponse



- **MNAR** : Plus le navire réalise une grande part de son activité au sein de la zone d'étude et plus sa probabilité d'être non-répondant est faible

$$p_i = 1 - \frac{e^{\alpha + \beta \text{Activite}_i}}{1 + e^{\alpha + \beta \text{Activite}_i}}$$

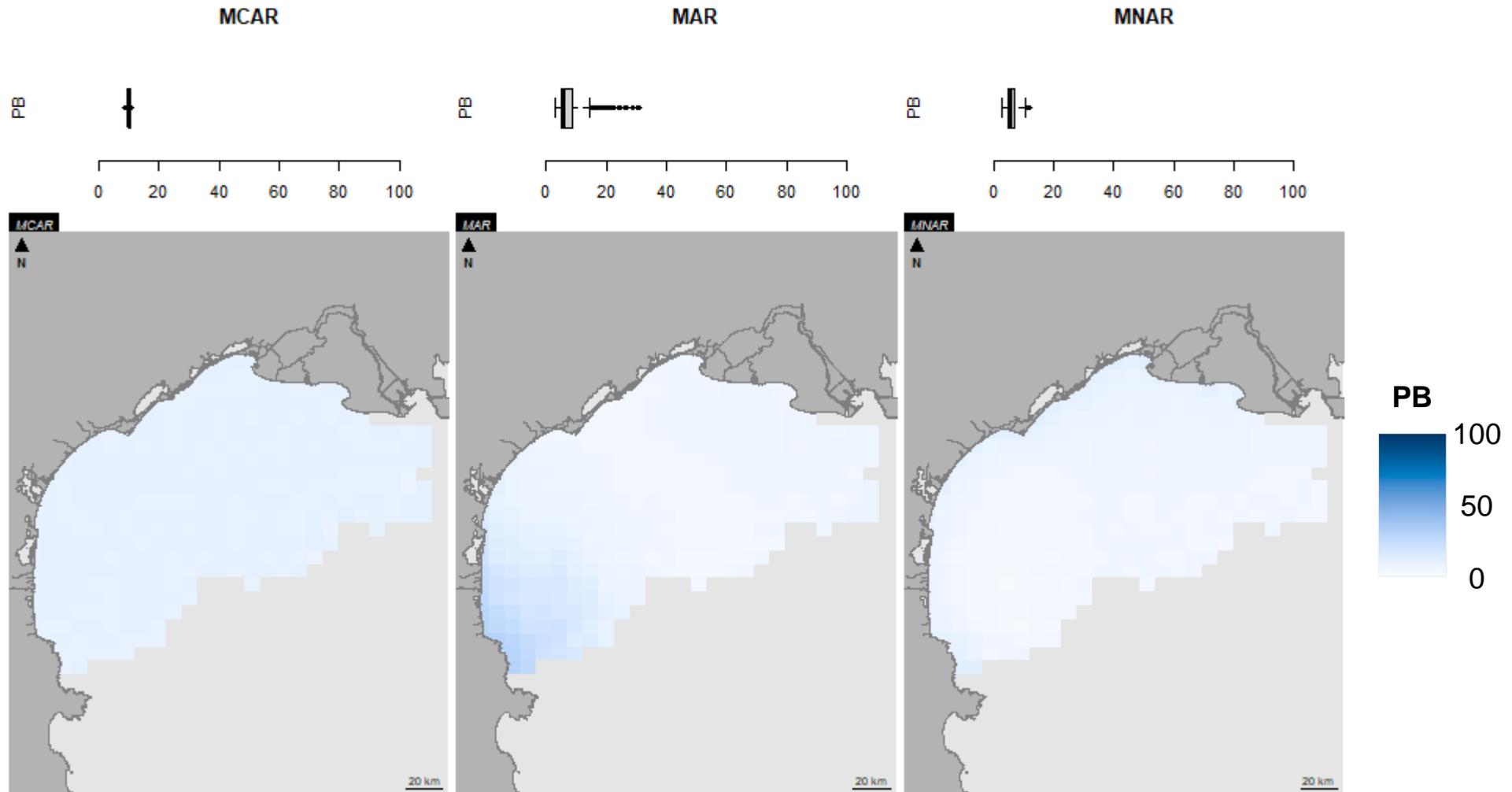
$$\text{Avec } \alpha = \begin{cases} 1.925 \\ 1.1 \\ 0.535 \text{ et } \beta = 4 \\ 0.07 \\ -0.36 \end{cases}$$

$$\text{Activite}_i = \frac{\sum_{k \in ZE} Y_{i,k}}{\sum_{k \in Z} Y_{i,k}}$$

$$Y_{i,k} = \sum_m \sum_e \sum_s I_{nav_i \times maille_k \times mois_m \times engin_e \times espece_s}$$

4. Illustration du biais de non-réponse

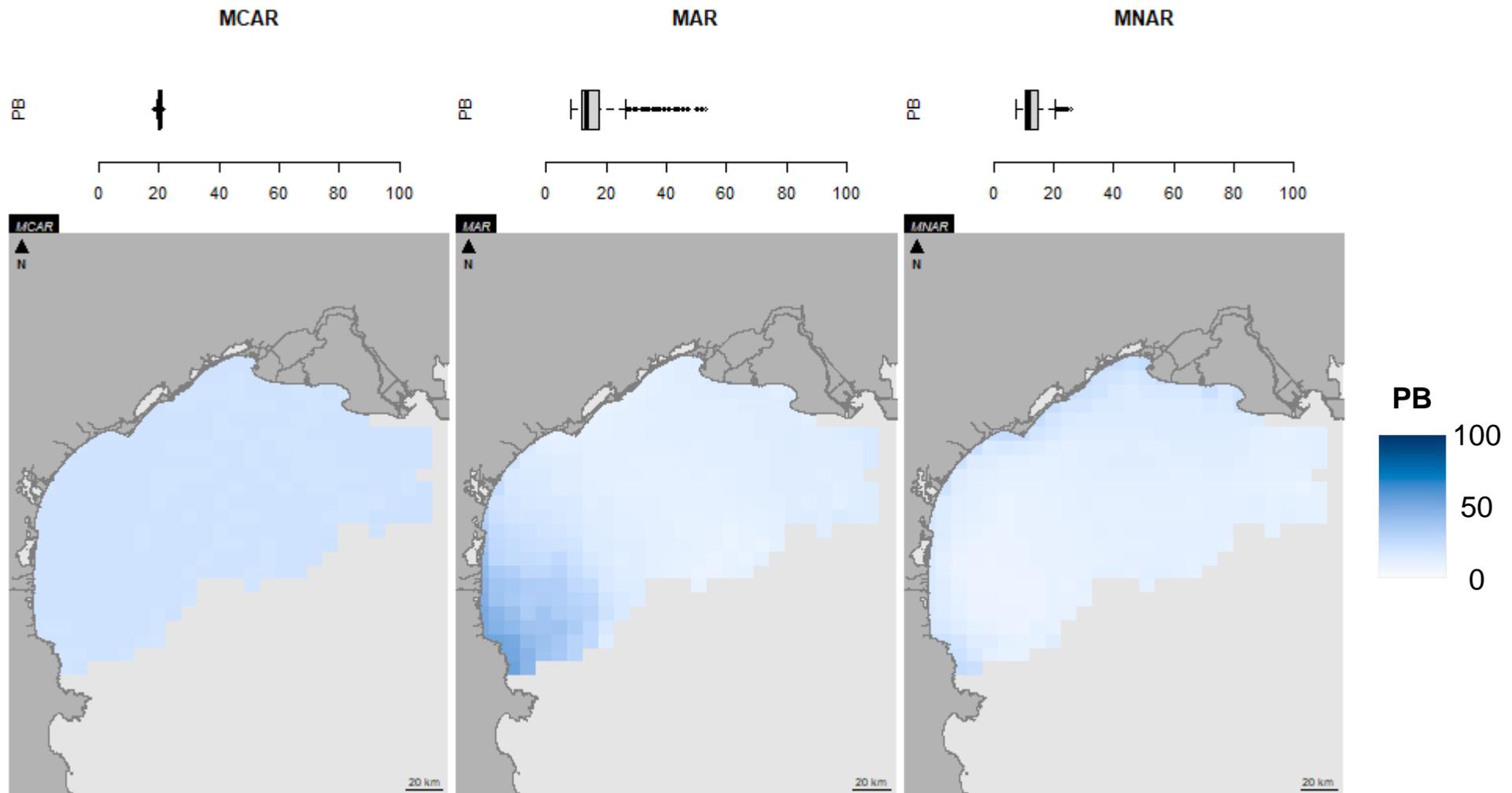
Taux de non-réponse : 10%



Biais (%) de l'indicateur d'intensité par maille

4. Illustration du biais de non-réponse

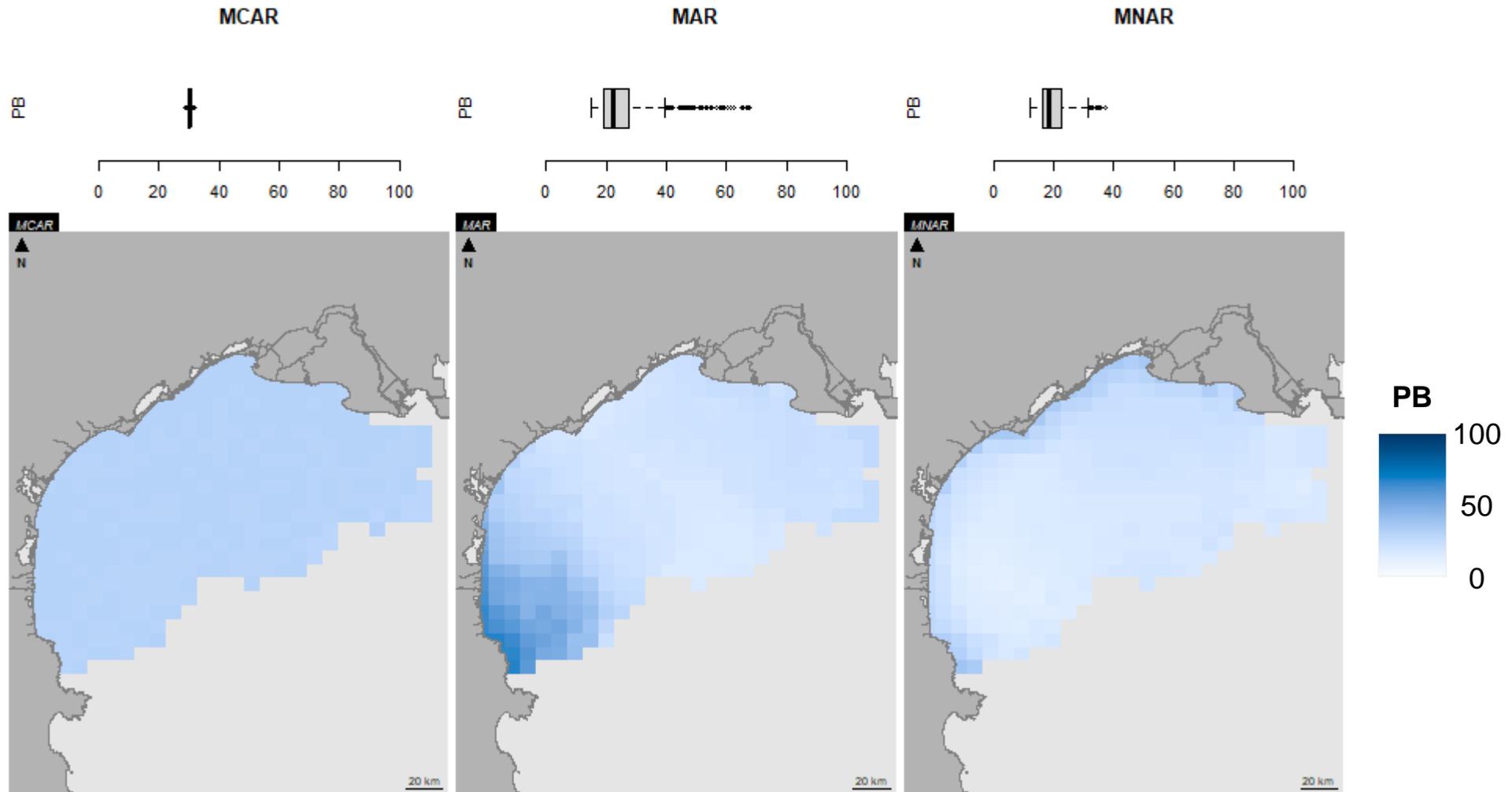
Taux de non-réponse : 20%



Biais (%) de l'indicateur d'intensité par maille

4. Illustration du biais de non-réponse

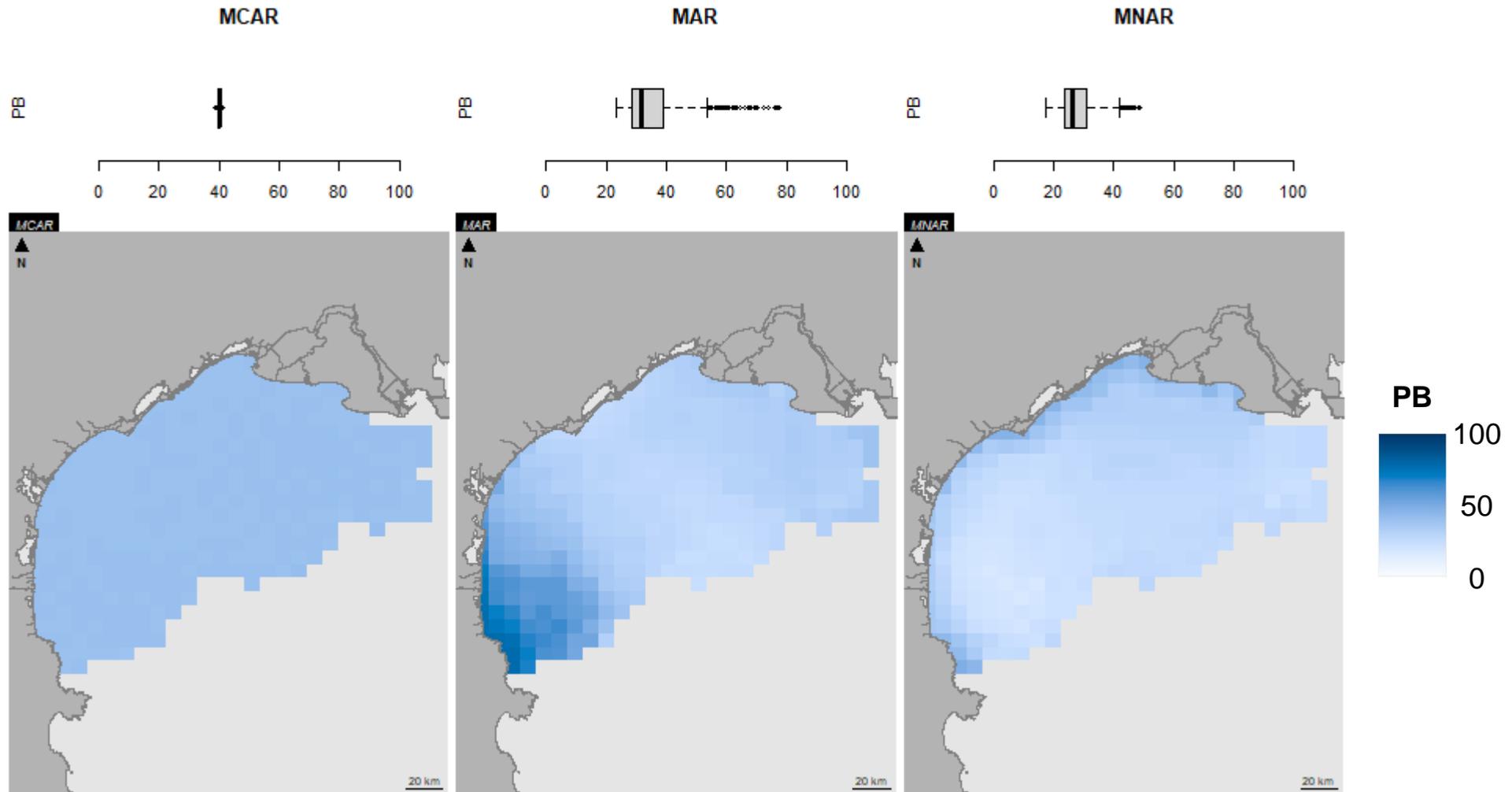
Taux de non-réponse : 30%



Biais (%) de l'indicateur d'intensité par maille

4. Illustration du biais de non-réponse

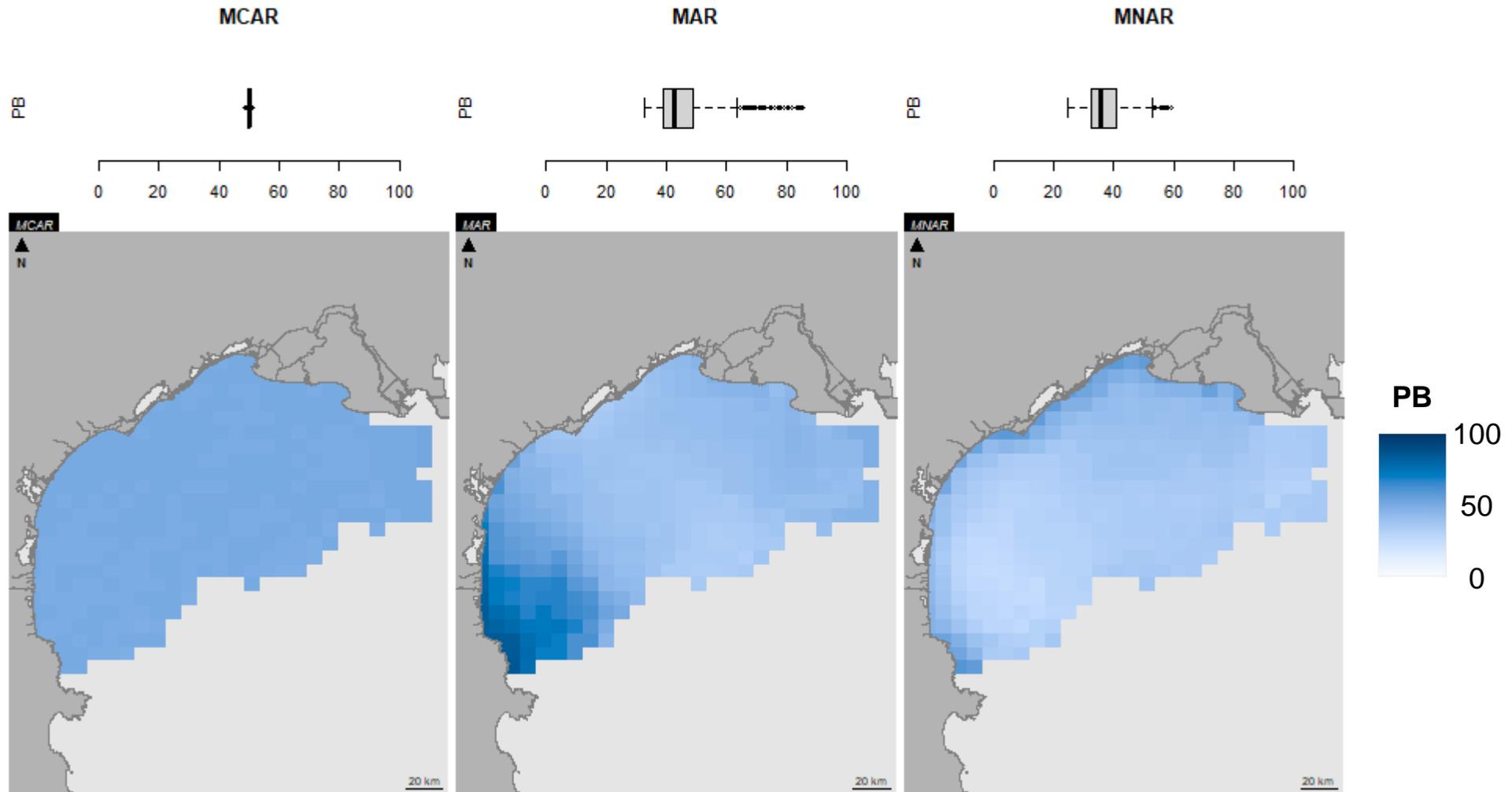
Taux de non-réponse : 40%



Biais (%) de l'indicateur d'intensité par maille

4. Illustration du biais de non-réponse

Taux de non-réponse : 50%



Biais (%) de l'indicateur d'intensité par maille

5. Recommandations

- Il est important de réaliser un **état des lieux des données manquantes** afin d'être en mesure de discuter de la représentativité des données collectées
- Cet état des lieux doit quantifier le **taux de non-réponse** et analyser la **répartition des données manquantes** au sein de la population d'étude en fonction de variables auxiliaires
- Pour cela, il est nécessaire de disposer d'une **base de données** permettant :
 - D'identifier l'ensemble des navires de la population d'étude
 - D'identifier les navires répondants et non-répondants
 - De disposer d'informations auxiliaires corrélées à l'activité des navires
- Lorsque le risque de biais est élevé, il est conseillé de mettre en œuvre une méthode de **traitement des données manquantes** (Little et Rubin, 2019)

Références

- Arbia, G., Espa, G., Giuliani, D., 2016. Dirty spatial econometrics. *Ann Reg Sci* 56, 177–189. <https://doi.org/10.1007/s00168-015-0726-5>
- Little, R.J.A., Rubin, D.B., 2019. *Statistical Analysis with Missing Data*, 3rd Edition. John Wiley & Sons, Ltd.
- Rubin, D.B., 1976. Inference and missing data. *Biometrika* 63, 581–592. <https://doi.org/10.1093/biomet/63.3.581>
- Trouillet, B., Bellanger-Husi, L., El Ghaziri, A., Lamberts, C., Plissonneau, E., Rollo, N., 2019. More than maps: Providing an alternative for fisheries and fishers in marine spatial planning. *Ocean & Coastal Management* 173, 90–103. <https://doi.org/10.1016/j.ocecoaman.2019.02.016>

Merci de votre attention