# Actualisation de la classification des sites de mesure

Mathieu Joly & Valentin Petiot

4 mars 2025

## Table des matières

**1 Données utilisées** **2**

**2 Traitement des données** **3**

**3 Cartographie du résultat** **3**

**4 Validation croisée** **3**

**5 Étude des anomalies** **3**

**6 Comparaison à la précédente version** **3**

**7 Évolution du jeu de stations classifiées** **12**

**8 Conclusion** **12**

# 1 Données utilisées

- La période d'étude comprend 8 années, de 2017 à 2024, avec des données non validées pour 2023 et 2024.
- Ne sont pas pris en compte les sites d'altitude supérieure à 1400 m (altitude à partir de laquelle le nombre de stations diminue fortement). En Europe, ces stations sont peu nombreuses, mais ne peuvent pas être confondues avec les sites de plaine pour l'analyse.
- Les stations renseignées comme « industrielles » ne sont pas prises en compte. La variabilité temporelle de ce type de mesure est très difficile à caractériser, et la méthode n'est pas suffisamment robuste pour appréhender le comportement potentiellement erratique des indicateurs calculés.
- Les mesures de CO à très faibles résolution (valeurs discrètes multiples de 100) sont éliminées en amont. Ces stations ne sont de toutes façons pas utilisées en aval par CAMS2_40.

À partir des métadonnées, on dérive la typologie simplifiée suivante :

**Type R :** sites qualifiés *background* et *rural*.

**Type S :** sites qualifiés *background* et *suburban*.

**Type U :** sites qualifiés *background* et *urban*.

**Type T :** sites qualifiés *traffic* et *urban*.

**Type O :** toutes les autres stations, ainsi que les stations en dehors du sous-domaine, qui ne seront pas prises en compte pour l'Analyse Discriminante, mais qui seront classifiées *a posteriori*.

Le CO et l'ozone font toujours figure d'anomalie, avec beaucoup de stations T dans le premier cas, et beaucoup de stations *background* (R, S, et U) dans l'autre. Pour le $SO_2$ les stations T sont peu nombreuses.

Avec l'utilisation du nouveau flux EEA (fichiers « parquet »), le nombre de stations prises en compte augmente (+6%), en particulier pour le NO (+8%), pour les $PM_{2.5}$ (+10%), et pour les $PM_{10}$ (+14%).

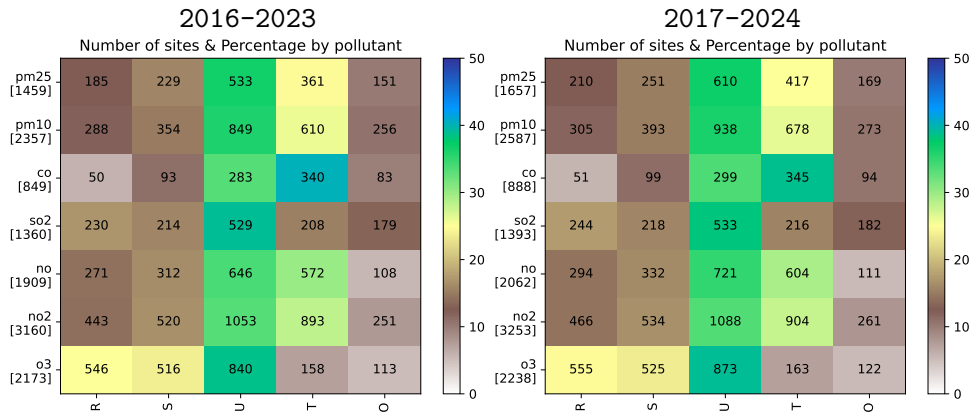


**Figure 1 –** Nombre de stations sélectionnées (données suffisantes), par type de métadonnée. Les couleurs correspondent au pourcentage par polluant.

## 2 Traitement des données

Les figures 2 et 3 montrent les stations pour lesquelles les séries temporelles, bien que suffisantes en quantité de données, ne permettent pas de calculer tous les indicateurs. L'utilisation des fichiers « parquet » de l'EEA permet de récupérer certaines régions d'Italie qui posaient problème auparavant. Malgré tout, pour le NO, il reste un nombre de stations important pour lesquelles les valeurs absentes sont trop nombreuses au sein de chaque journée, ce qui empêche le calcul d'indicateurs quotidiens. Pour l'Allemagne, il semble que les valeurs inférieures à un seuil de détection instrumental aient été supprimées de la base de données.

## 3 Cartographie du résultat

Les figures 4 et 5 illustrent la classification obtenue pour chaque polluant.

## 4 Validation croisée

La figure 6 compare les « validations croisées » par rapport aux types dérivés des métadonnées. La cohérence entre les classifications subjective (métadonnées) et objective est stable par rapport à la précédente version.

## 5 Étude des anomalies

Nous allons nous intéresser aux comportements marginaux de la figure 6 :

- le pourcentage des stations R qui se retrouvent dans les classes 6-10.
- le pourcentage des stations S, U et T qui se retrouvent dans les classes 1-3.

| | $O_3$ | $NO_2$ | NO | $SO_2$ | CO | $PM_{10}$ | $PM_{2.5}$ |
|---|---|---|---|---|---|---|---|
| R 6-10 | 4 → 5 | 2 → 2 | 4 → 3 | 28 → 30 | 18 → 21 | 19 → 20 | 24 → 22 |
| S+U+T 1-3 | 9 → 9 | 3 → 3 | 4 → 4 | 13 → 14 | 3 → 3 | 7 → 7 | 8 → 8 |

**Tableau 1 –** Pourcentage des anomalies (cf. paragraphe ci-dessus). Évolution entre l'ancienne et la nouvelle classification (en vert pour une amélioration, en rouge pour une déterioration, et surligné de jaune quand plus de 2% des stations sont affectées).

Le tableau 1 montre que pour les stations rurales, la classification comporte des anomalies moins nombreuses que l'année précédente pour les $PM_{2.5}$. Par contre, pour le CO et le $SO_2$ les anomalies sont plus nombreuses. Pour les autres polluants, les différences sont faibles.

Les cartes 7 et 8 cartographient les anomalies du tableau 1. L'analyse est difficile, car il faudrait regarder localement la configuration de chacun de ces sites « douteux », et les sources de pollution environnantes. La nouvelle version ne modifie pas beaucoup la localisation de ces anomalies.

## 6 Comparaison à la précédente version

Pour les stations en commun dans les deux classifications, la figure 9 compare les classes obtenues. La classification est généralement très stable, à l'exception de quelques anomalies isolées. Pour les $PM_{2.5}$ la cohérence s'est améliorée ces dernières années.

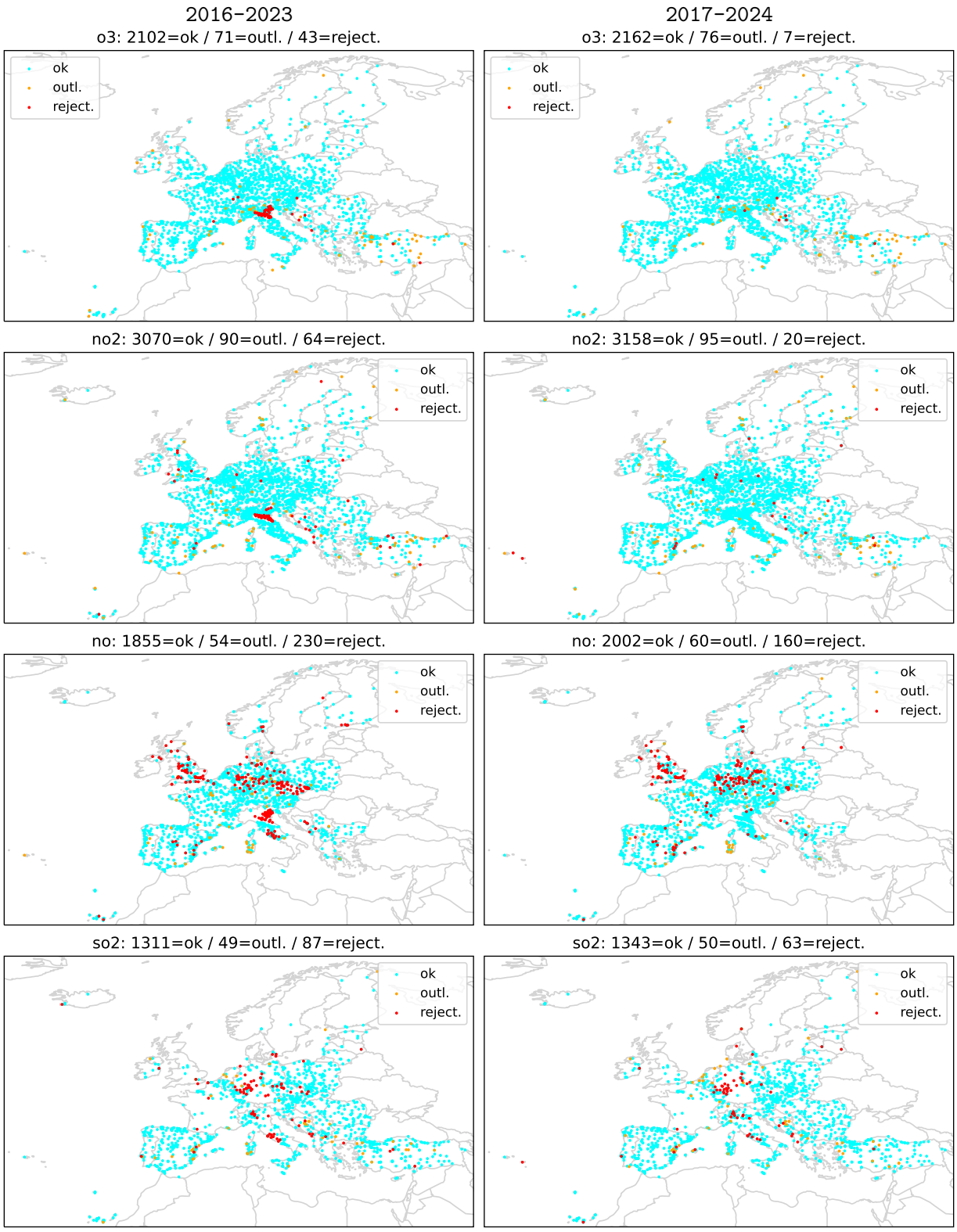


**Figure 2 –** Localisation des stations rejetées lors du calcul des indicateurs (*rejected*), ou lors de l'analyse (*outliers*). À gauche, pour la précédente classification ; et à droite, pour la nouvelle version.

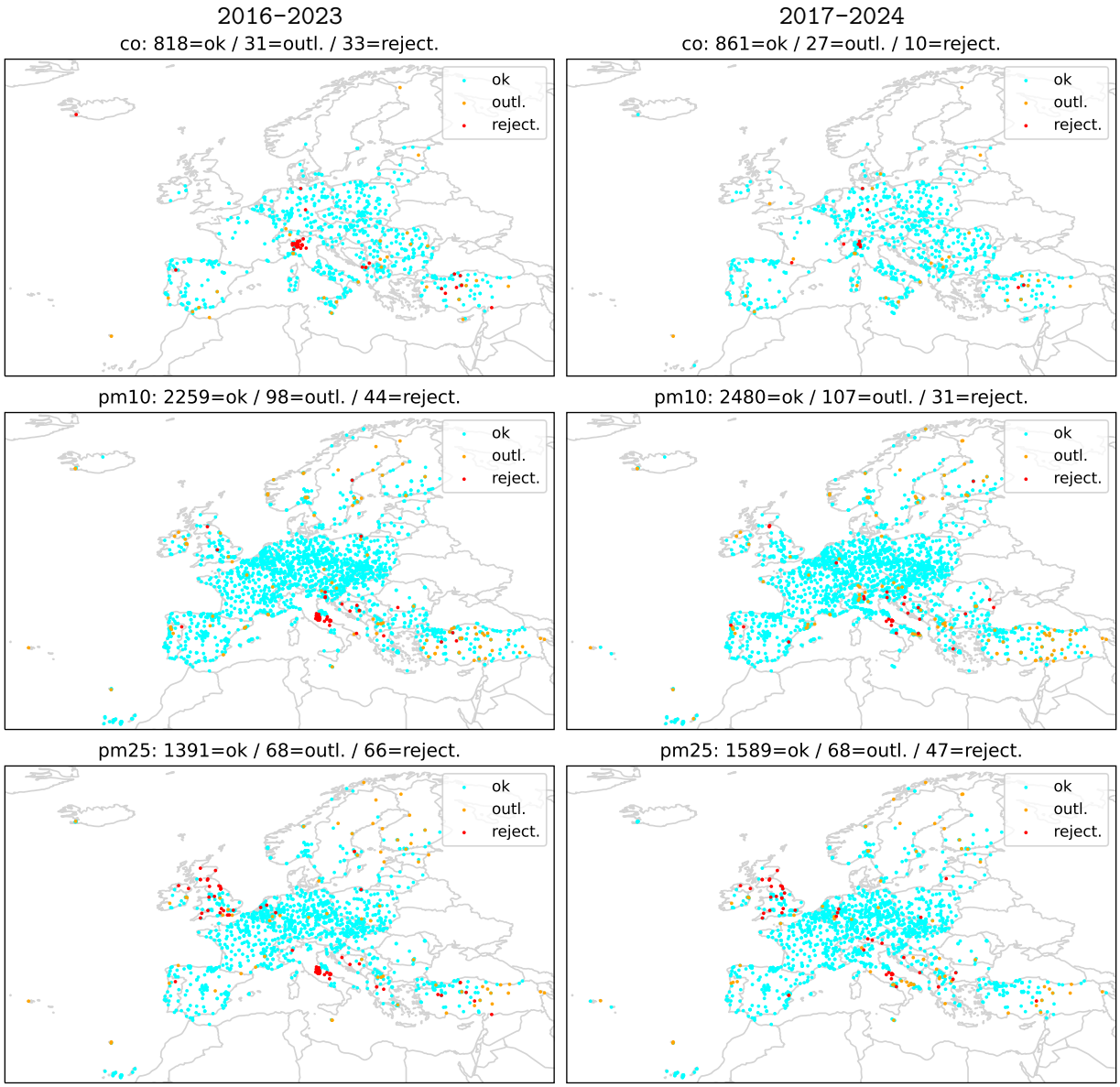


**Figure 3 –** Localisation des stations rejetées lors du calcul des indicateurs (*rejected*), ou lors de l'analyse (*outliers*). À gauche, pour la précédente classification ; et à droite, pour la nouvelle version.

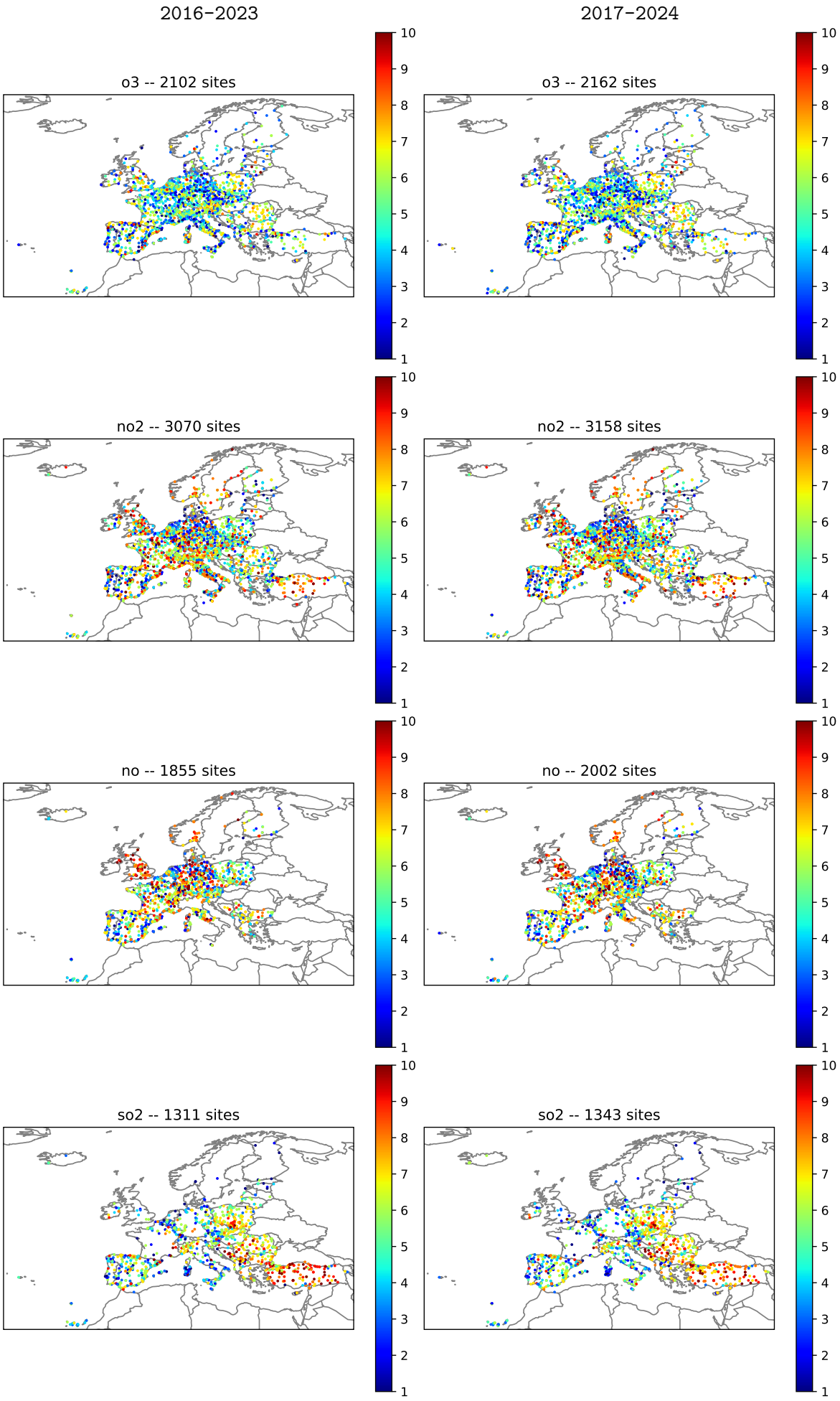


**Figure 4 –** Cartographie de la classification obtenue.

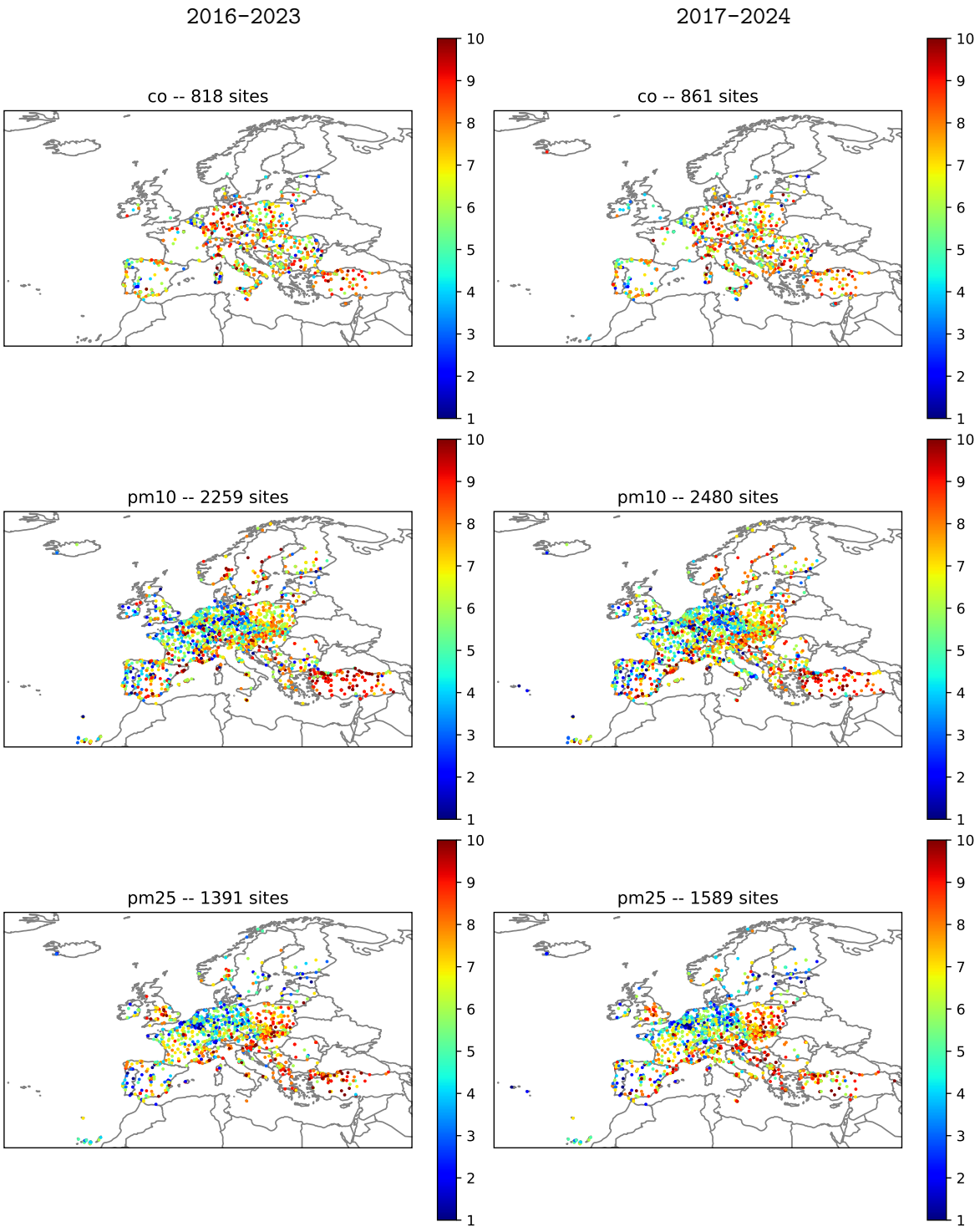


**Figure 5 –** Cartographie de la classification obtenue. À gauche, pour la précédente classification ; et à droite, pour la nouvelle version.

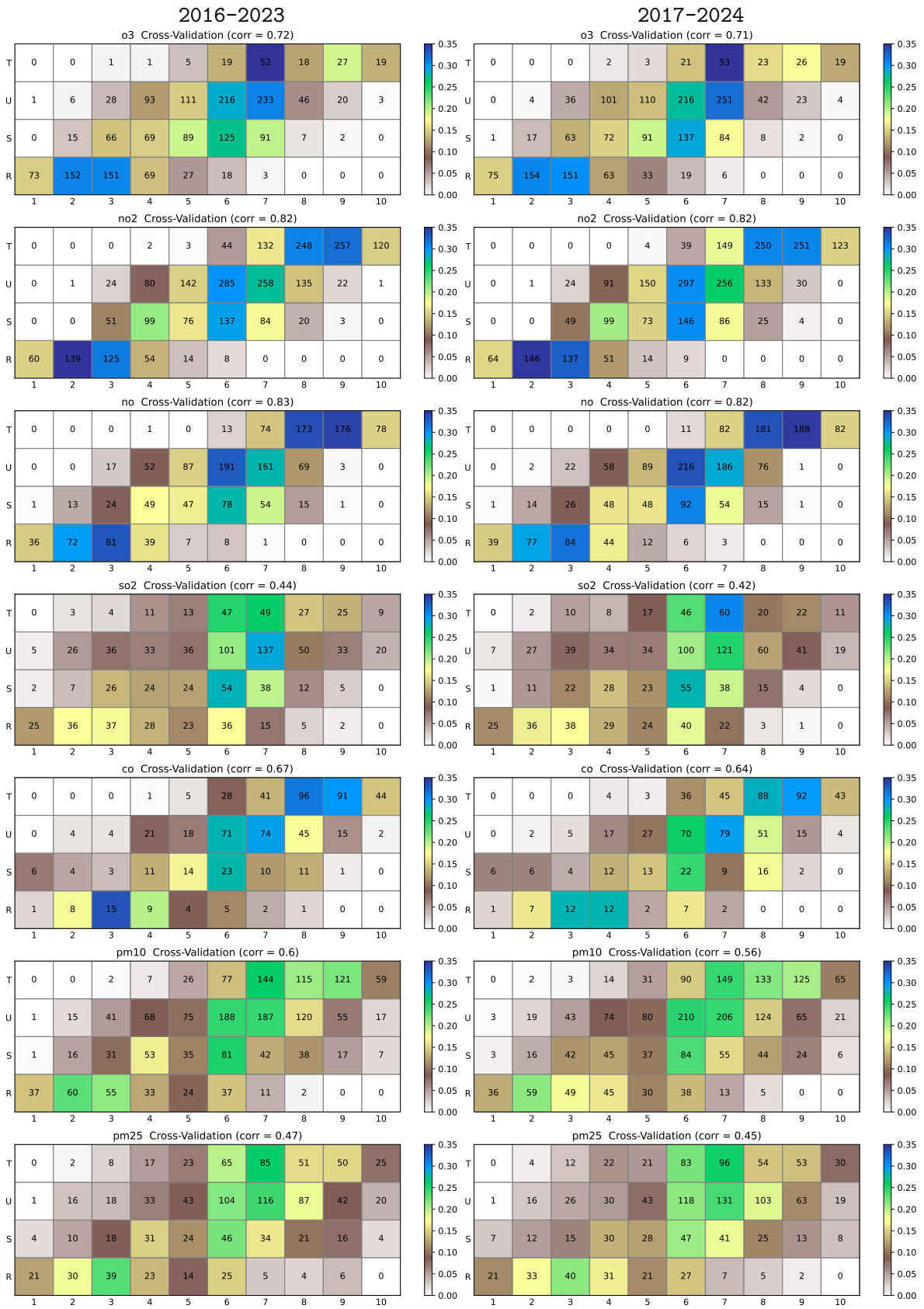


**Figure 6 –** Validation croisée : nombre et pourcentage (en couleur) dans chaque classe pour chaque type de station. À gauche, pour la précédente classification ; et à droite, pour la nouvelle version.

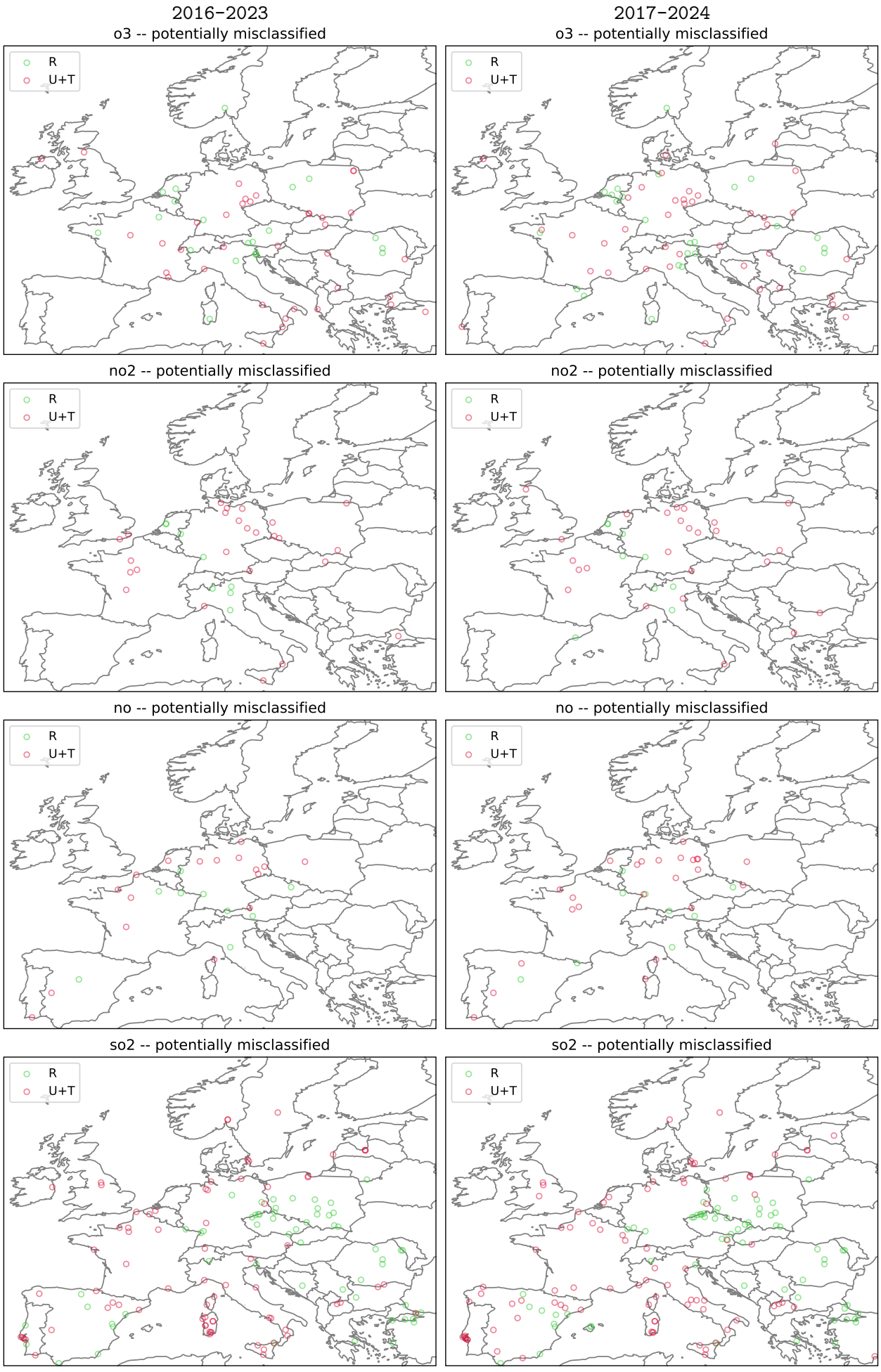


**Figure 7 –** Stations R qui se retrouvent dans les classes 6-10, et stations U et T qui se retrouvent dans les classes 1-3. À gauche, pour la précédente classification ; et à droite, pour la nouvelle version.

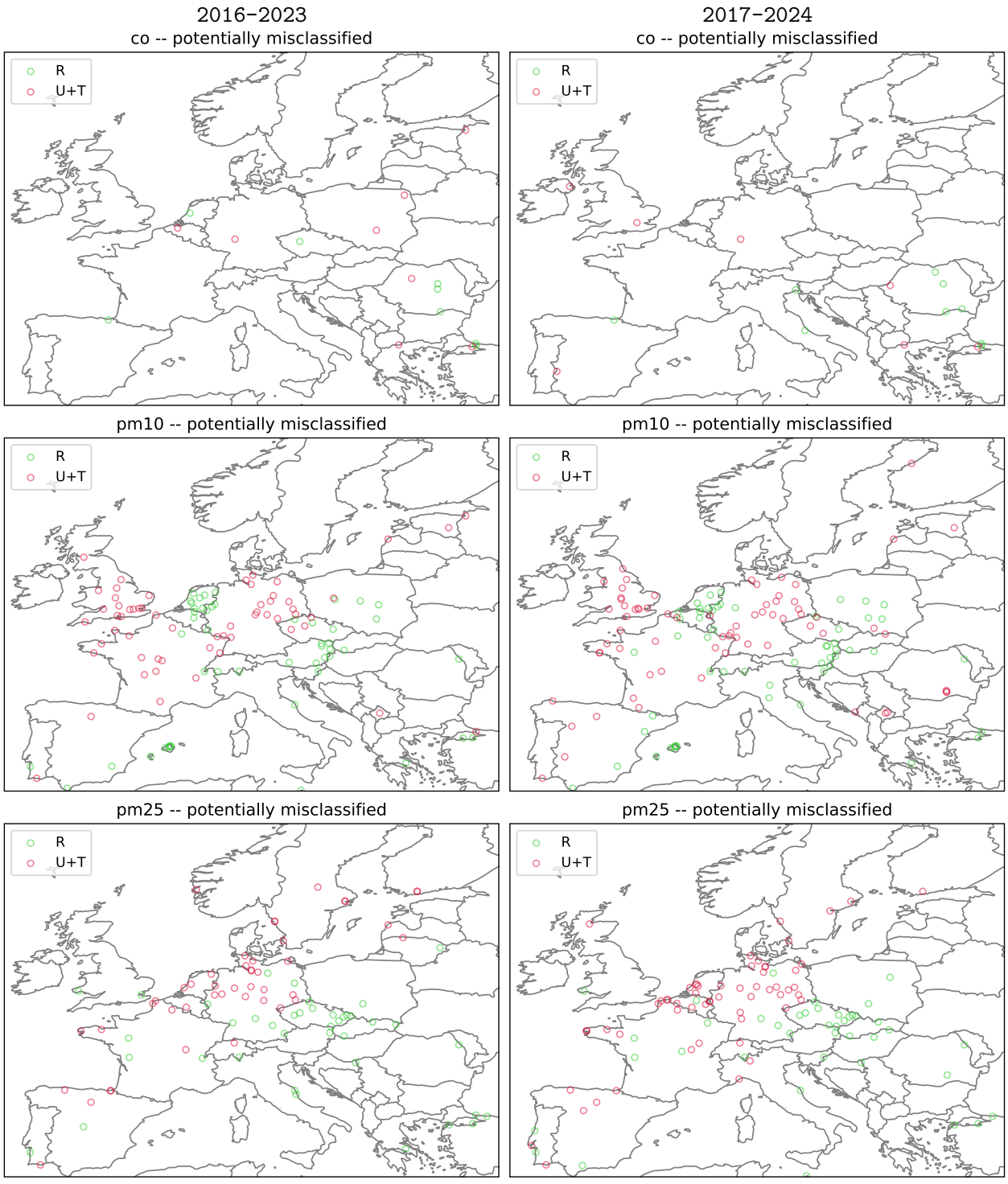


**Figure 8 –** Stations R qui se retrouvent dans les classes 6-10, et stations U et T qui se retrouvent dans les classes 1-3. À gauche, pour la précédente classification ; et à droite, pour la nouvelle version.

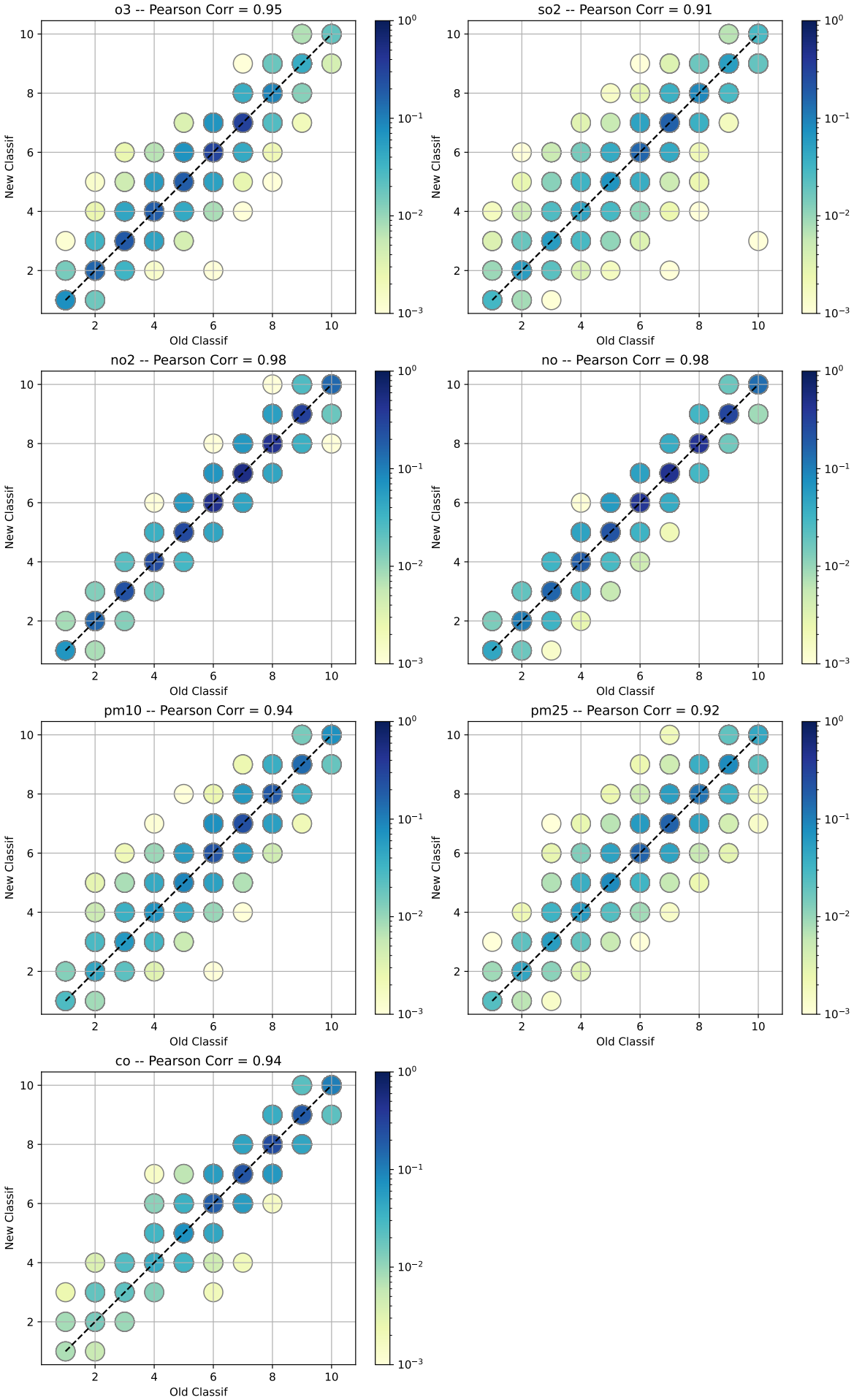


**Figure 9 –** Scatter Plot des classes obtenues avec l'ancienne et la nouvelle classification. La couleur indique la fréquence d'occurence.

# 7 Évolution du jeu de stations classifiées

La figure 10 permet de suivre l'évolution du jeux de données classifiées. Le bilan est positif, puisque l'on gagne plus de stations que l'on en perd. L'Italie et la Turquie en particulier, sont mieux représentées.

# 8 Conclusion

Cette version utilise le flux de l'EEA mis en place dans le cadre de CAMS. La période d'étude comprend 8 années, avec des données non validées pour 2023 et 2024.

- L'utilisation des fichiers « parquet » de l'EEA pour 2023 et 2024 a permis de récupérer davantage de données pour certaines régions. Le nombre de stations classifiées est donc en augmentation avec cette version, pour tous les polluants (et tout particulièrement les PM avec environ 200 stations supplémentaires).
- La cohérence entre les métadonnées et la classification objective est globalement stable.

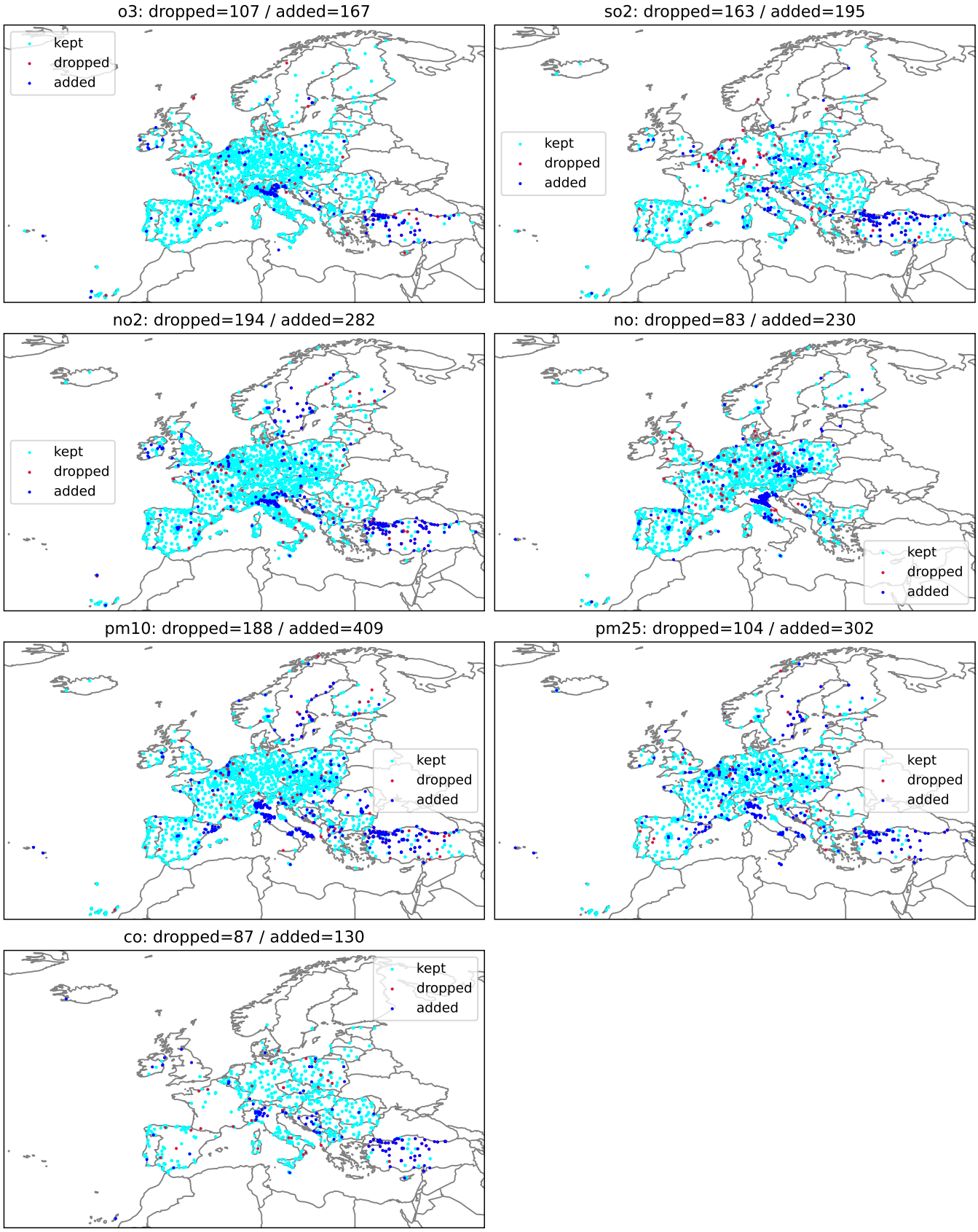


**Figure 10 –** Stations qui disparaissent (rouge), ou qui apparaissent (bleu) dans la nouvelle version.