Actualisation de la classification des sites de mesure

Mathieu Joly & Valentin Petiot

8 février 2022

Table des matières

1 Données utilisées **2**

2 Traitement des données **2**

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 2014 à 2021, avec des données non validées pour 2021.
- 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.

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. Le réseau de mesure s'étoffe doucement (+5%), en particulier pour les PM (+12%). Le nombre de stations diminue légèrement pour le SO_2.

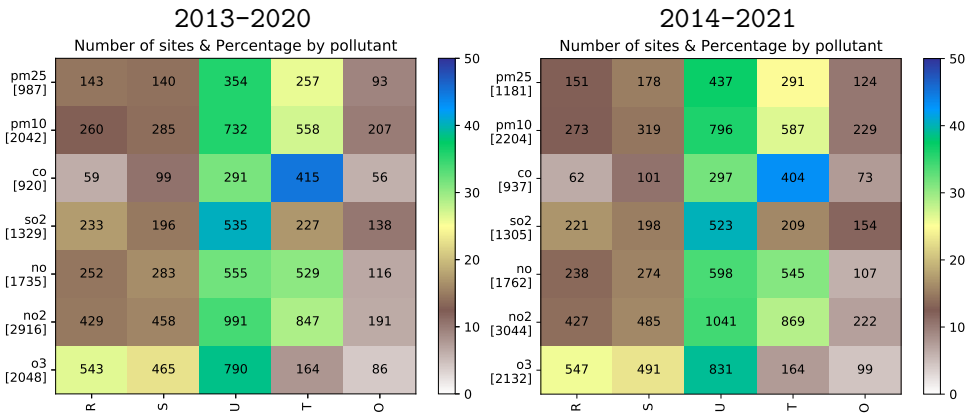

Figure 1 – Nombre de stations sélectionnées (données suffisantes), par type de métadonnée. Les couleurs correspondent au pourcentage par polluant.

2 Traitement des données

Les figures 2 et 3 montrent que les séries temporelles sur certaines régions, bien que suffisantes en quantité de données, ne permettent pas de calculer tous les indicateurs. C'est le cas pour une région d'Italie, ainsi qu'en Allemagne et Grande Bretagne pour le NO et le SO_2. Pour ces stations, 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 faibles

valeurs, autrefois mises à zéro, aient été supprimées de la base de données, d'où un taux de valeurs absentes trop élevé. Les autres pays n'ont pas procédé à un tel néttoyage des valeurs proches de zéro.

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 décroît légèrement pour le SO_2 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 → 4	3 → 1	2 → 3	26 → 32	12 → 15	17 → 17	28 → 23
S+U+T 1-3	8 → 9	3 → 3	3 → 3	12 → 13	3 → 3	6 → 7	10 → 9

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 confirme que la classification comporte un peu plus d'anomalies que l'année précédente, excepté pour le NO_2 et les $PM_2.5$ qui s'améliorent significativement. Les plus fortes incohérences sont dénombrées pour les stations rurales de SO_2 et $PM_{2.5}$ qui se retrouvelent régulièrement classées 6–10. Ce nombre d'anomalies augmente pour le SO_2.

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.

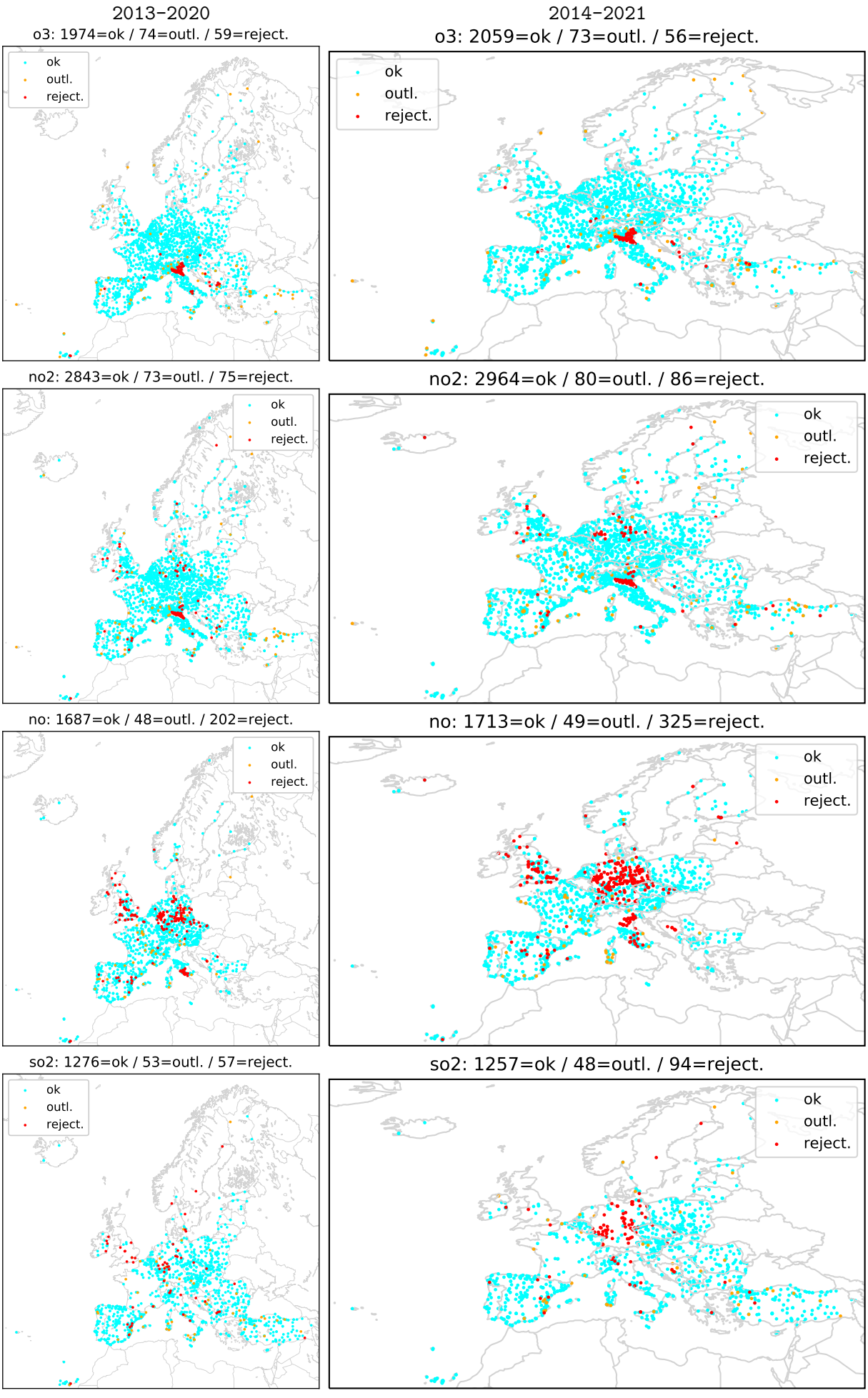

Figure 2 – Localisation des stations rejetées lors du calcul des indicateurs (*rejected*), ou lors de l'analyse (*outliers*). À gauche, pour la précédente classification ; et à droite, pour la nouvelle version.

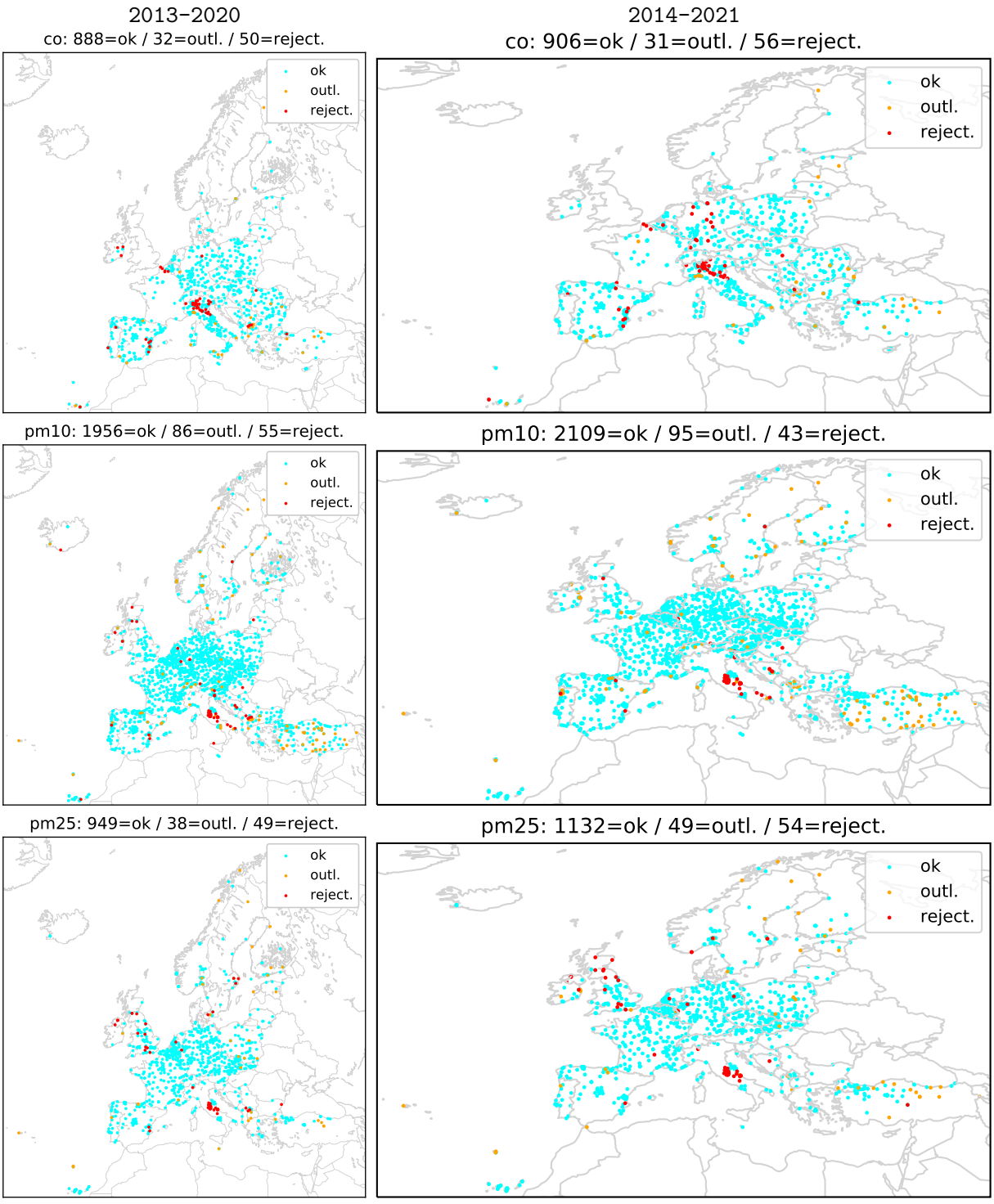

Figure 3 – Localisation des stations rejetées lors du calcul des indicateurs (*rejected*), ou lors de l'analyse (*outliers*). À gauche, pour la précédente classification ; et à droite, pour la nouvelle version.

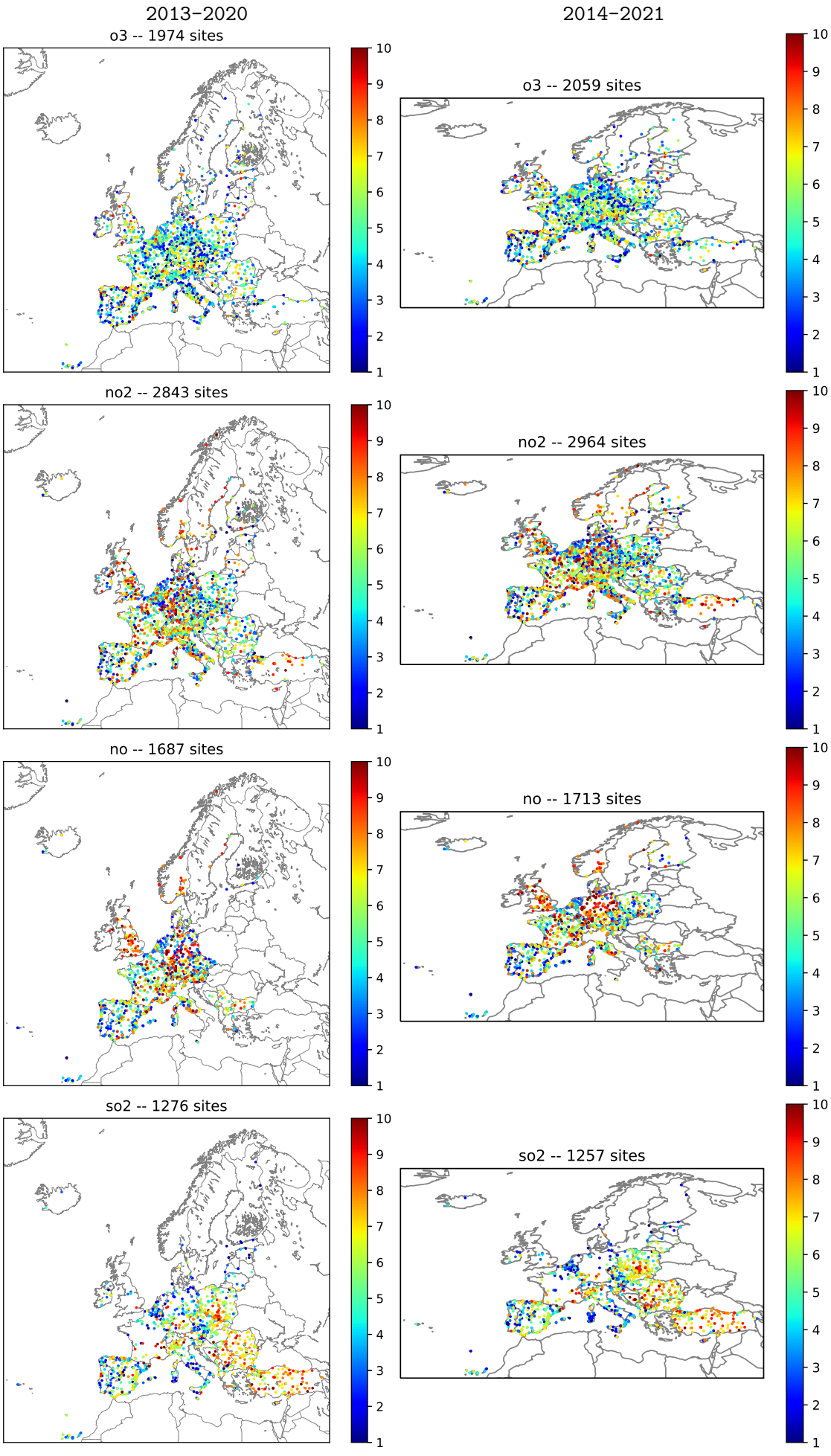

Figure 4 – Cartographie de la classification obtenue.

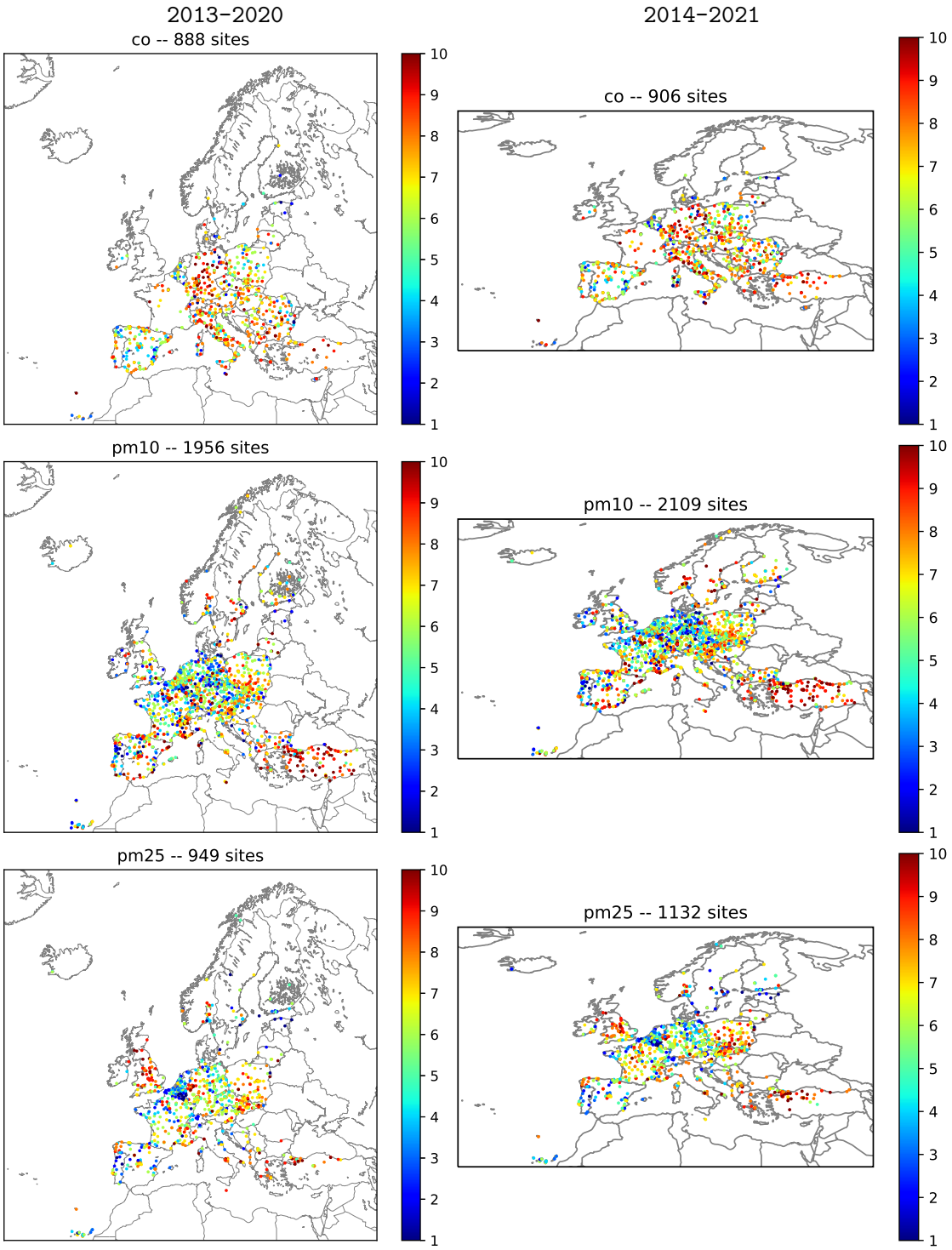

Figure 5 – Cartographie de la classification obtenue. À gauche, pour la précédente classification ; et à droite, pour la nouvelle version.

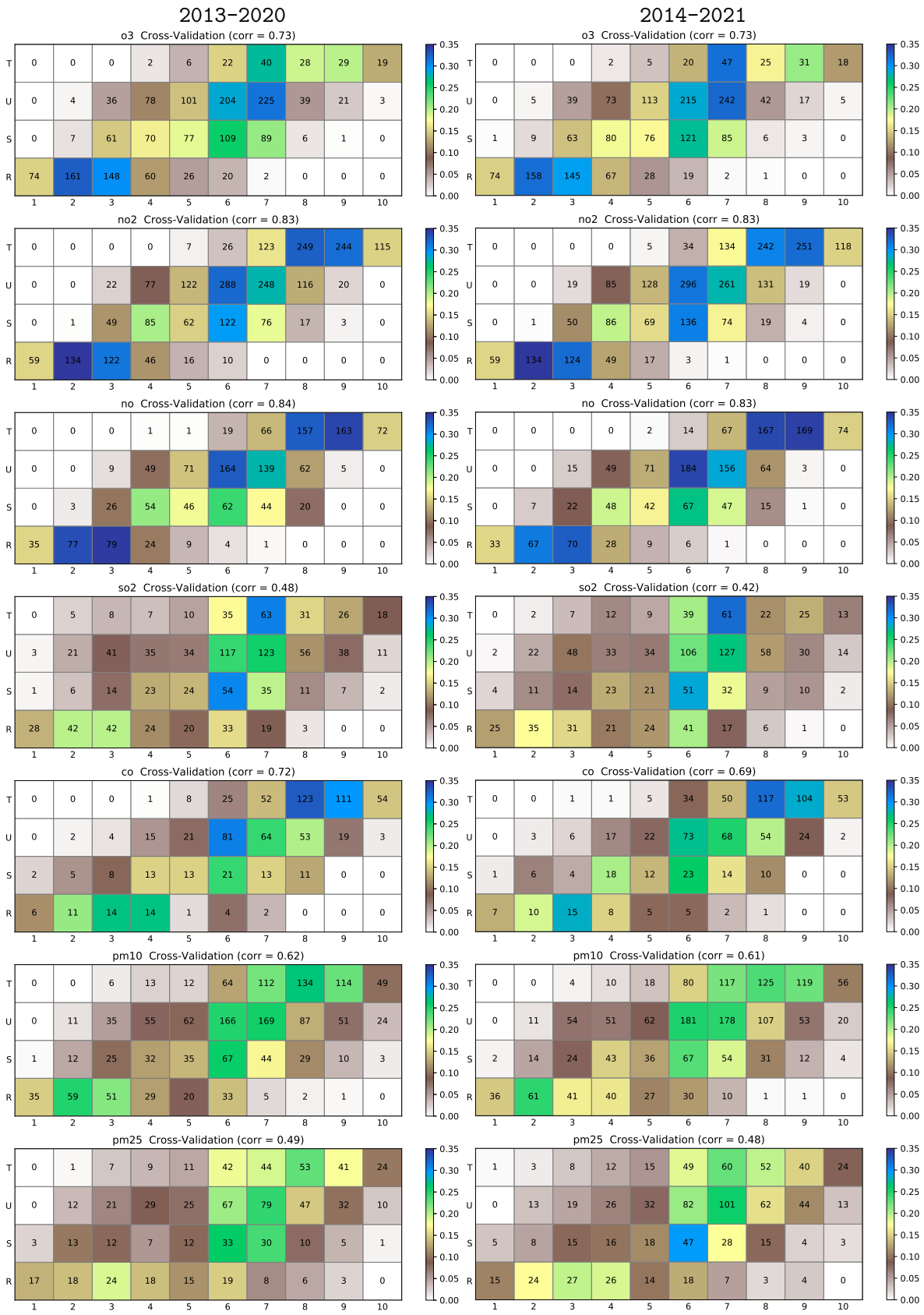

Figure 6 – Validation croisée : nombre et pourcentage (en couleur) dans chaque classe pour chaque type de station. À gauche, pour la précédente classification ; et à droite, pour la nouvelle version.

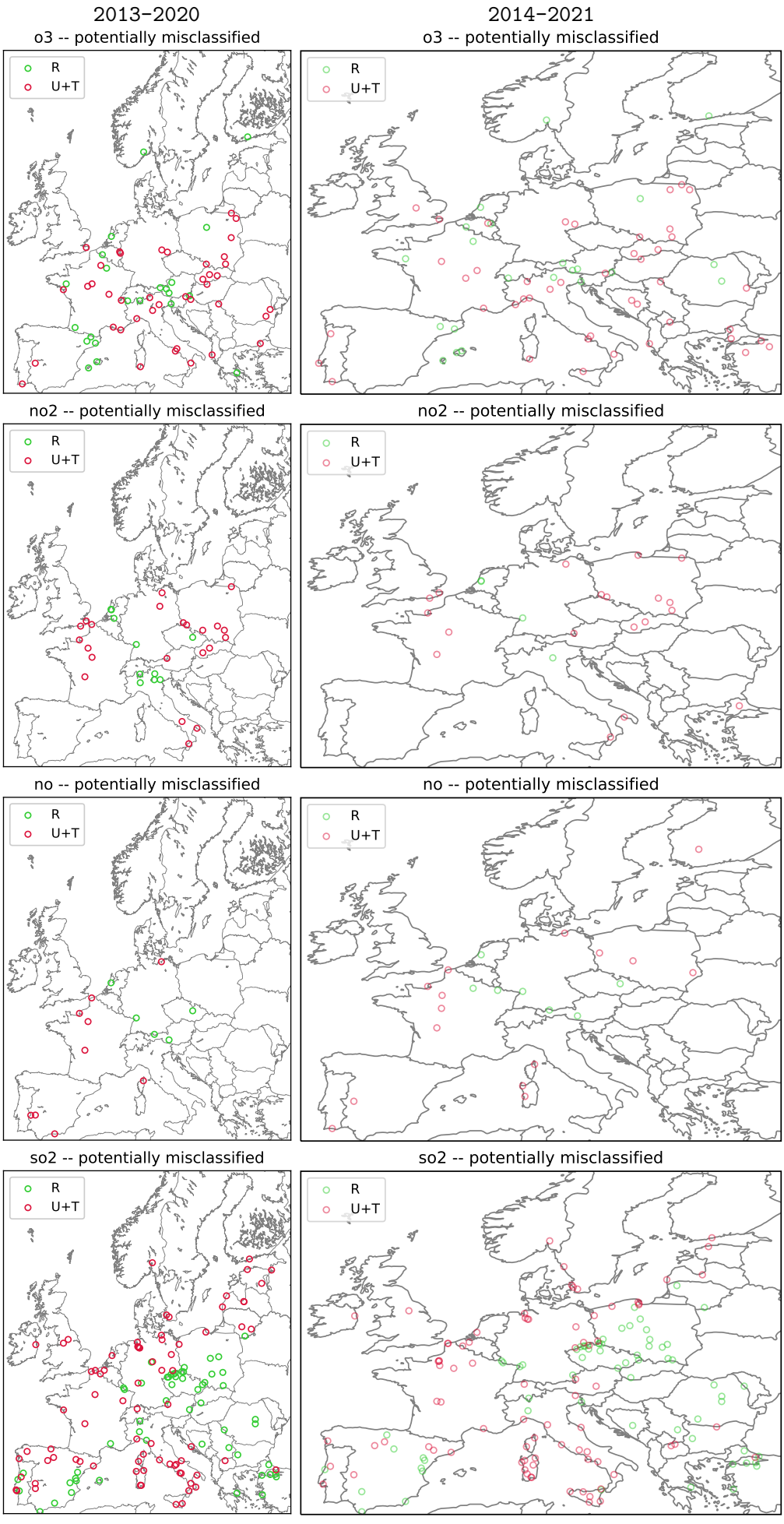

Figure 7 – Stations R qui se retrouvent dans les classes 6-10, et stations U et T qui se retrouvent dans les classes 1-3. À gauche, pour la précédente classification ; et à droite, pour la nouvelle version.

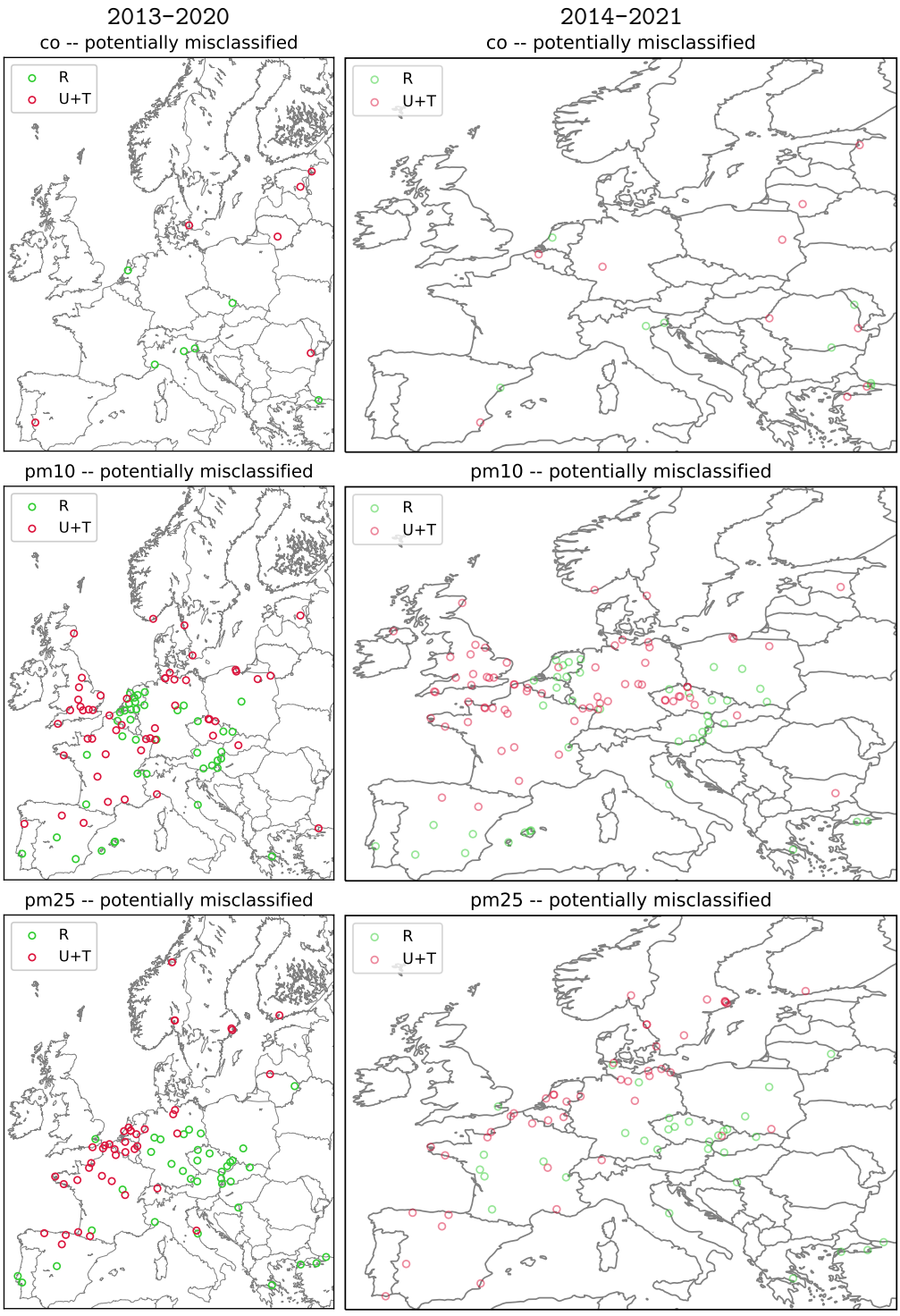

Figure 8 – Stations R qui se retrouvent dans les classes 6-10, et stations U et T qui se retrouvent dans les classes 1-3. À gauche, pour la précédente classification ; et à droite, pour la nouvelle version.

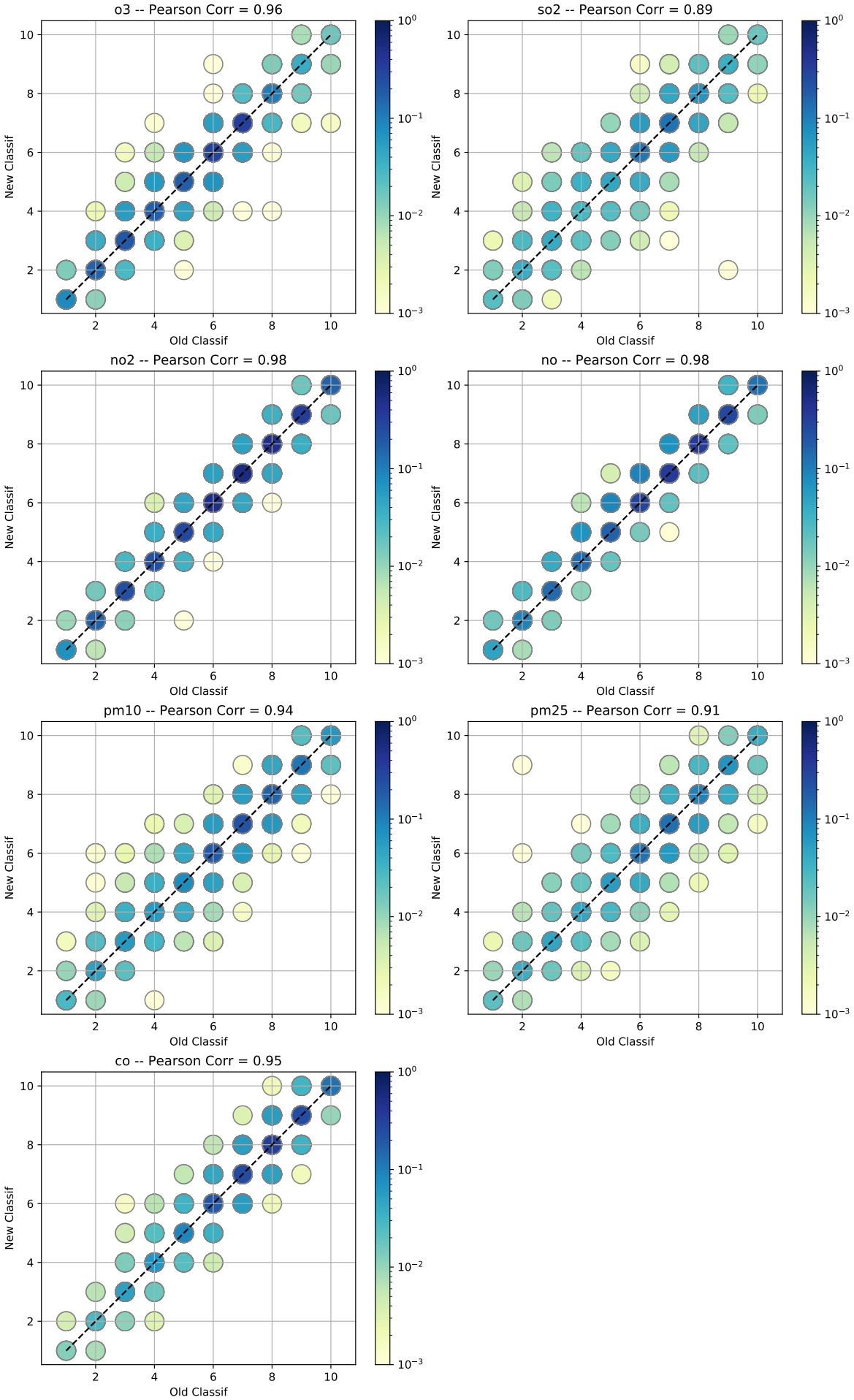

Figure 9 – Scatter Plot des classes obtenues avec l'ancienne et la nouvelle classification. La couleur indique la fréquence d'occurence.

7 Évolution du jeu de stations classifiées

La figure 10 permet de suivre l'évolution du jeux de données classifiées. On notera l'apparition de nouvelles stations en Turquie et en Pologne. Par contre, des stations disparaissent en Allemagne pour le NO et le SO_2, pour les raisons évoquées au paragraphe 2.

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 2021.

- Le réseau de mesure s'est légèrement étoffé, en particulier sur la Turquie et la Pologne. Il y a par contre un problème de gestion des codes de stations supérieurs à 7 caractères dans les métadonnées de l'EEA (renommage manuel).
- Comme dans la version précédente, la qualité des séries temporelles est insuffisante en certaines régions d'Italie, ainsi qu'Allemagne et Grande Bretagne pour le NO et le SO_2 : les valeurs absentes sont trop nombreuses au sein de chaque journée.
- La cohérence entre les métadonnées et la classification objective diminue légèrement en raison de quelques cas particuliers.

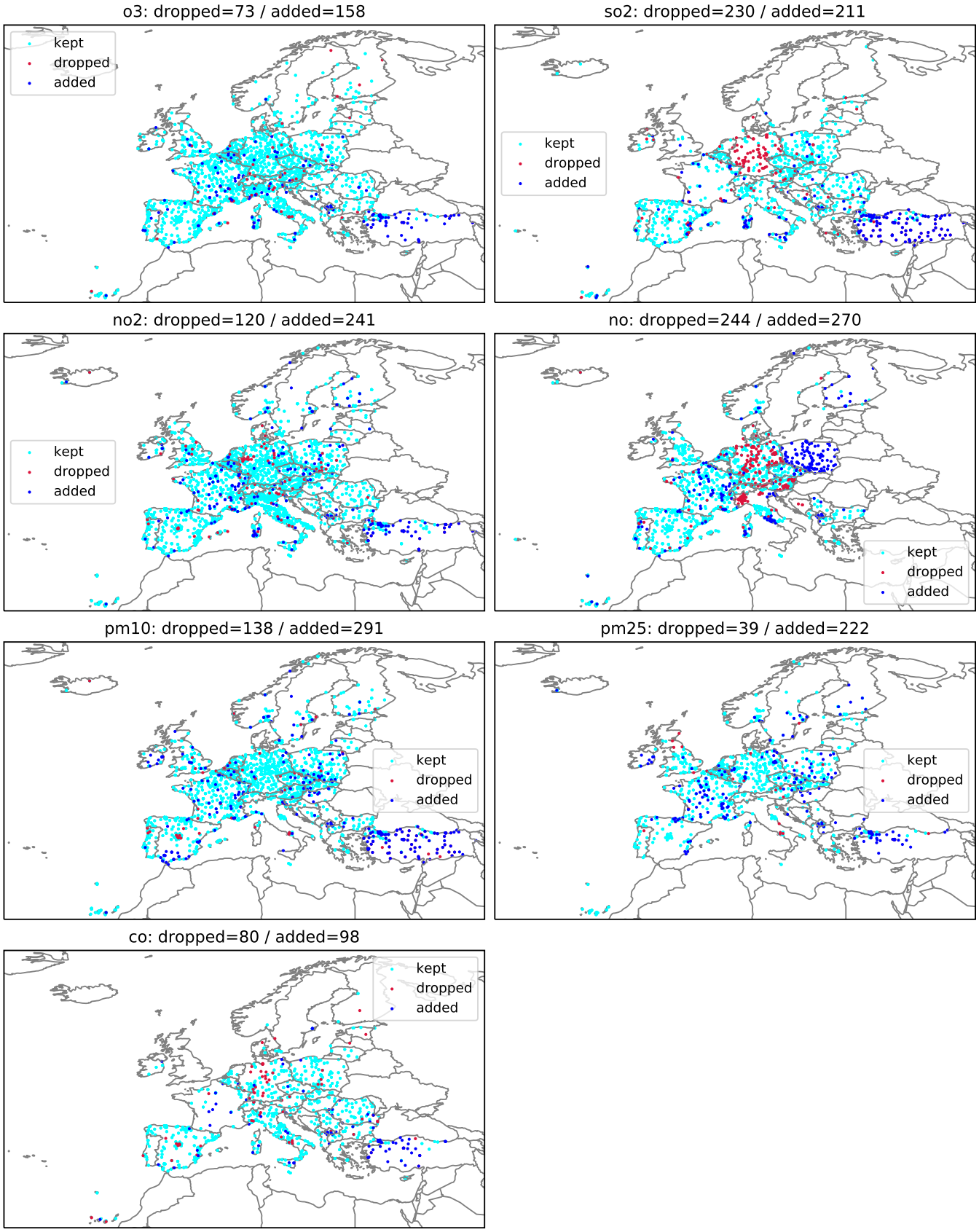

Figure 10 – Stations qui disparaissent (rouge), ou qui apparaissent (bleu) dans la nouvelle version.