Actualisation de la classification des sites de mesure

Mathieu Joly

11 février 2021

Table des matières

1 Données utilisées

- La période d'étude comprend 8 années, de 2013 à 2020, et comprend des données non validées pour 2020.
- 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*.

Dans la version 2020, les stations industrielles avaient été prises en compte par erreur, et se retrouvaient dans le « type O ». C'est corrigé dans cette nouvelle version.

Le CO et l'ozone font toujours figure d'anomalie, avec beaucoup de stations T dans un cas, et beaucoup de stations U et S dans l'autre. Le réseau de mesure s'étoffe doucement pour tous les polluants (on ne tiendra pas compte du « type O » gonflé par erreur avec les stations industrielles dans la précédente version).

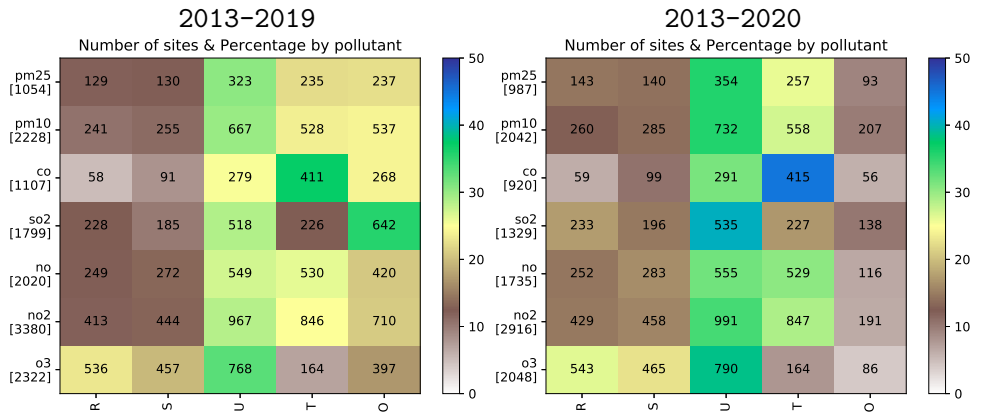

Figure 1 – Nombre de stations sélectionnées (données suffisantes), par type de métadonnée. Les couleurs correspondent au pourcentage par polluant.

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

en Italie, ou en Allemagne pour le NO. Les valeurs absentes sont trop nombreuses au sein de chaque journée. Le problème n'est pas corrigé par la prise en compte de l'année 2020 dans la nouvelle version.

3 Cartographie du résultat

Les figures 4 et 5 illustrent la classification obtenue pour chaque polluant. Le nombre de sites diminue du fait de l'exclusion des stations industrielles prises en compte par erreur dans la précédente version.

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	5 → 4	2 → 3	3 → 2	28 → 26	16 → 12	16 → 17	29 → 28
S+U+T 1-3	9 → 8	3 → 3	3 → 3	12 → 12	3 → 3	6 → 6	11 → 10

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 que confirme que la classification évolue peu, et plutôt favorablement pour CO et SO_2. Les plus fortes incohérences sont dénombrées pour les stations rurales de SO_2 et $SPM_{2.5}$ qui se retrouvelent régulièrement classées 6–10.

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.

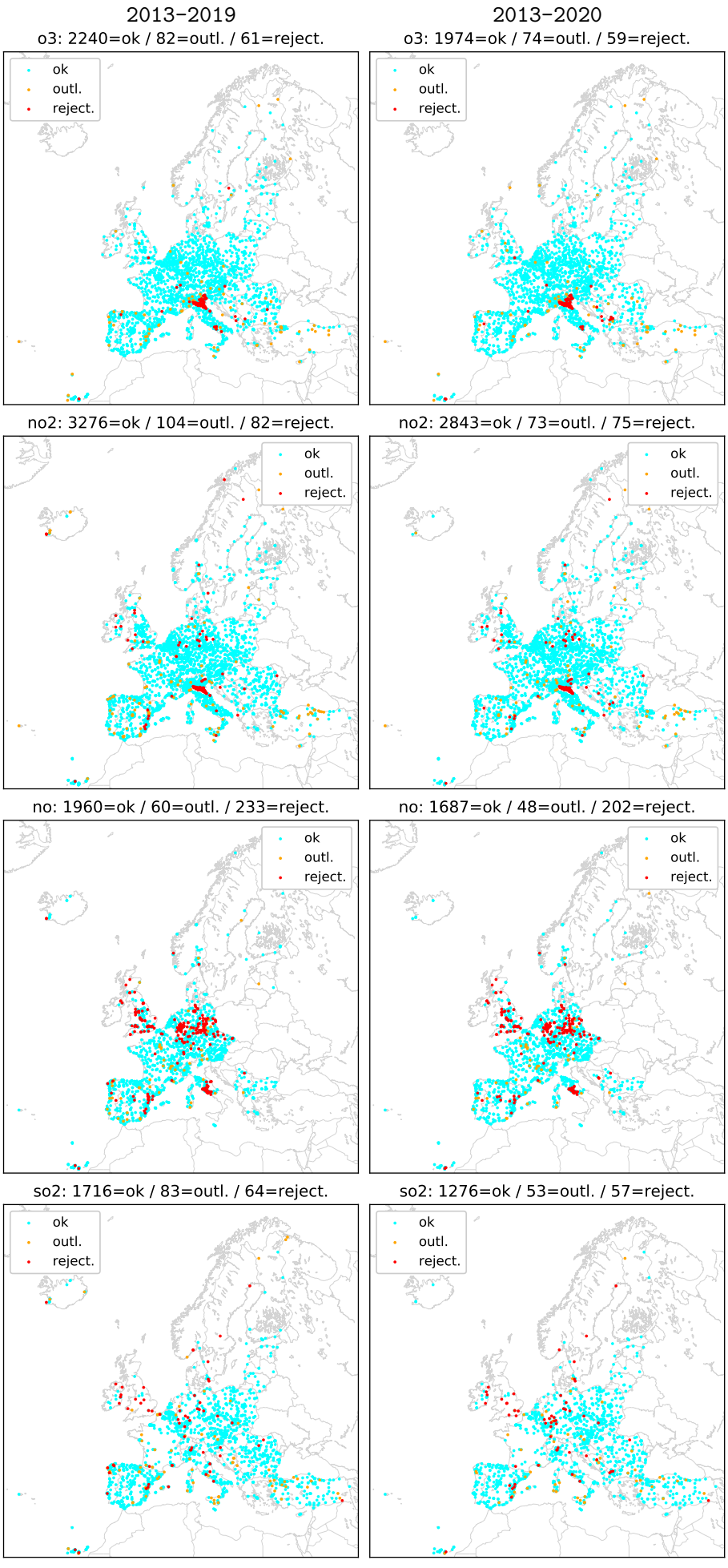

Figure 2 – Localisation des stations rejetées lors du calcul des indicateurs (*rejected*), ou lors de l'analyse (*outliers*). À gauche, pour la précédente classification ; et à droite, pour la nouvelle version.

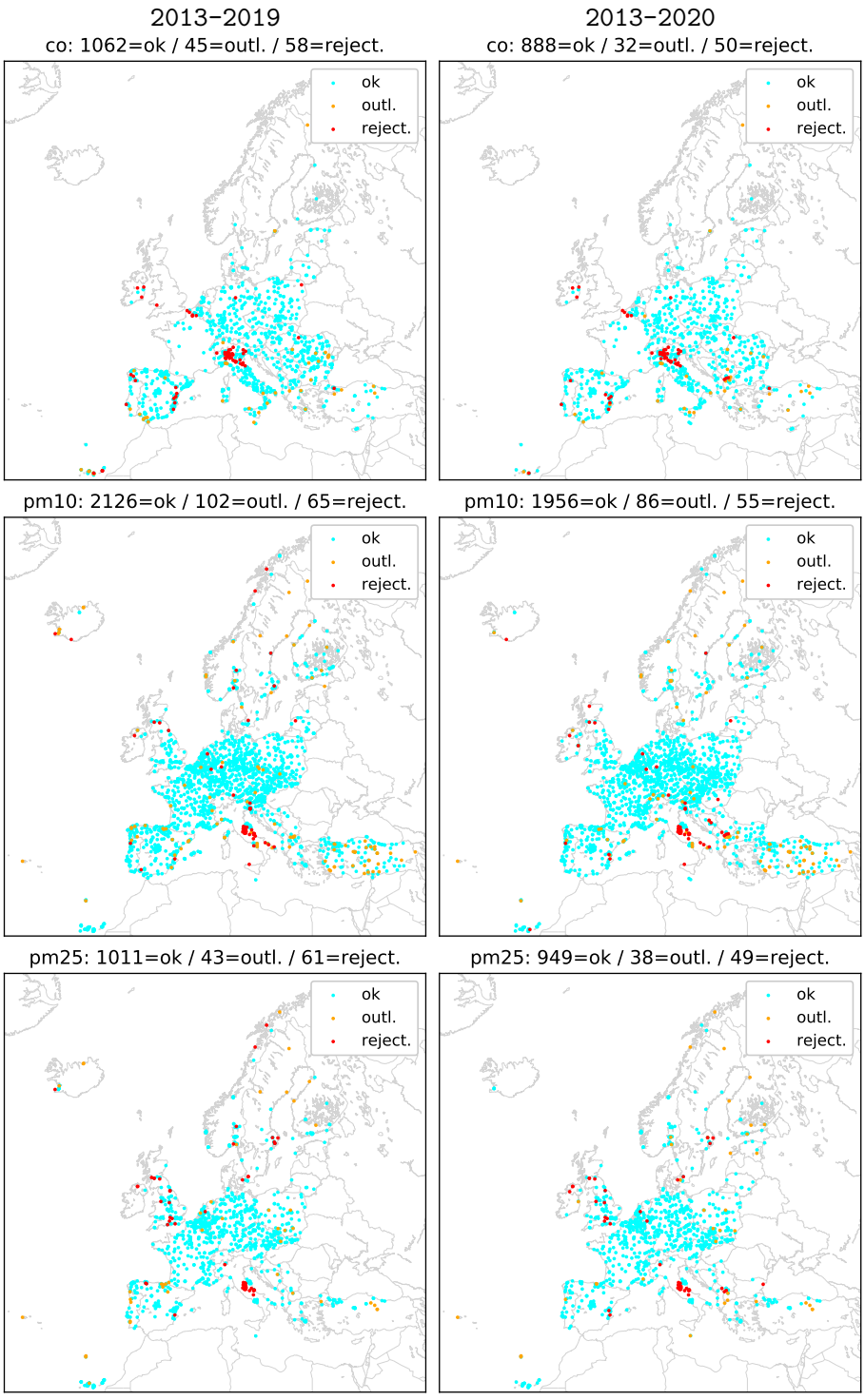

Figure 3 – Localisation des stations rejetées lors du calcul des indicateurs (*rejected*), ou lors de l'analyse (*outliers*). À gauche, pour la précédente classification ; et à droite, pour la nouvelle version.

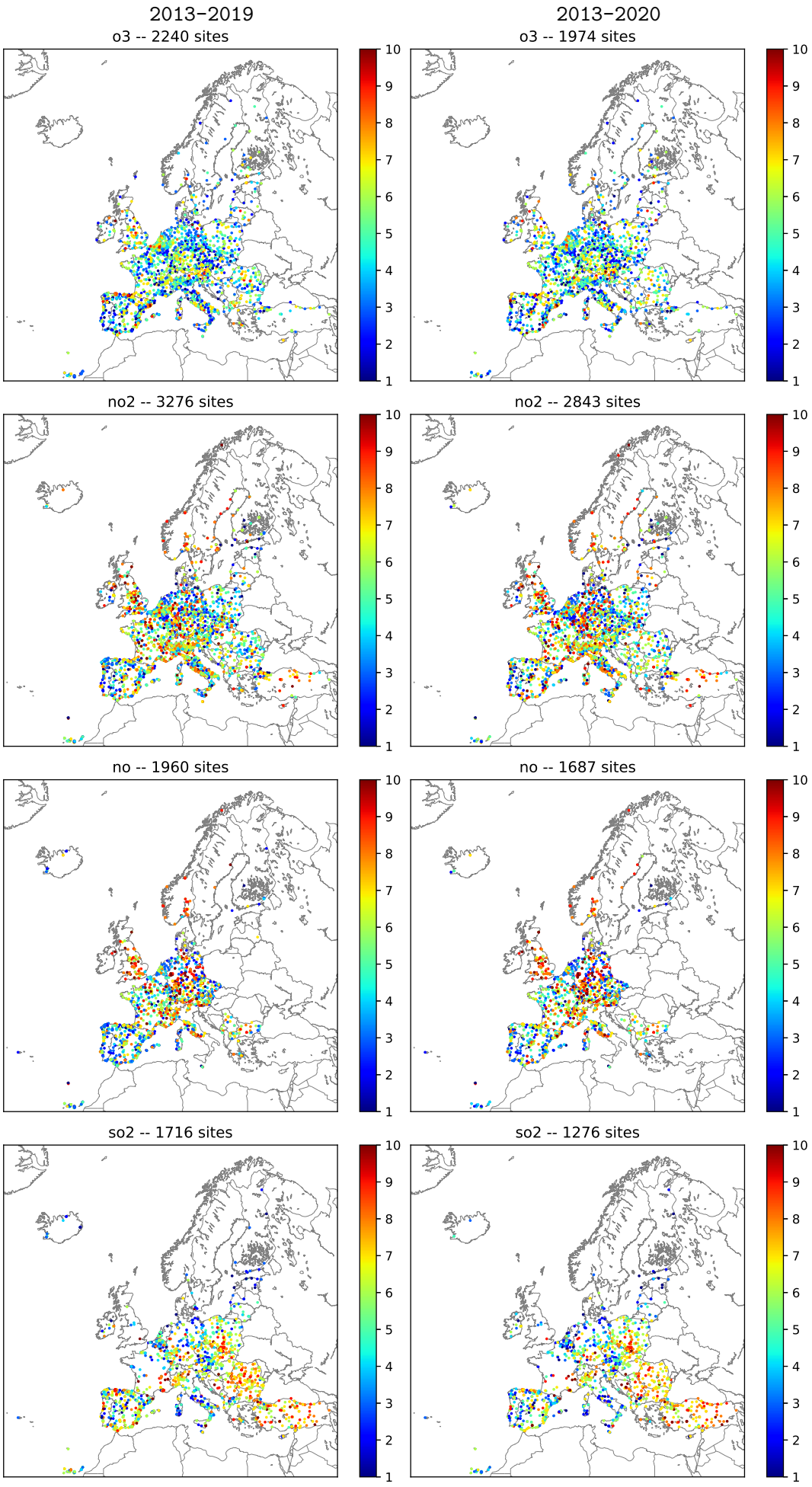

Figure 4 – Cartographie de la classification obtenue.

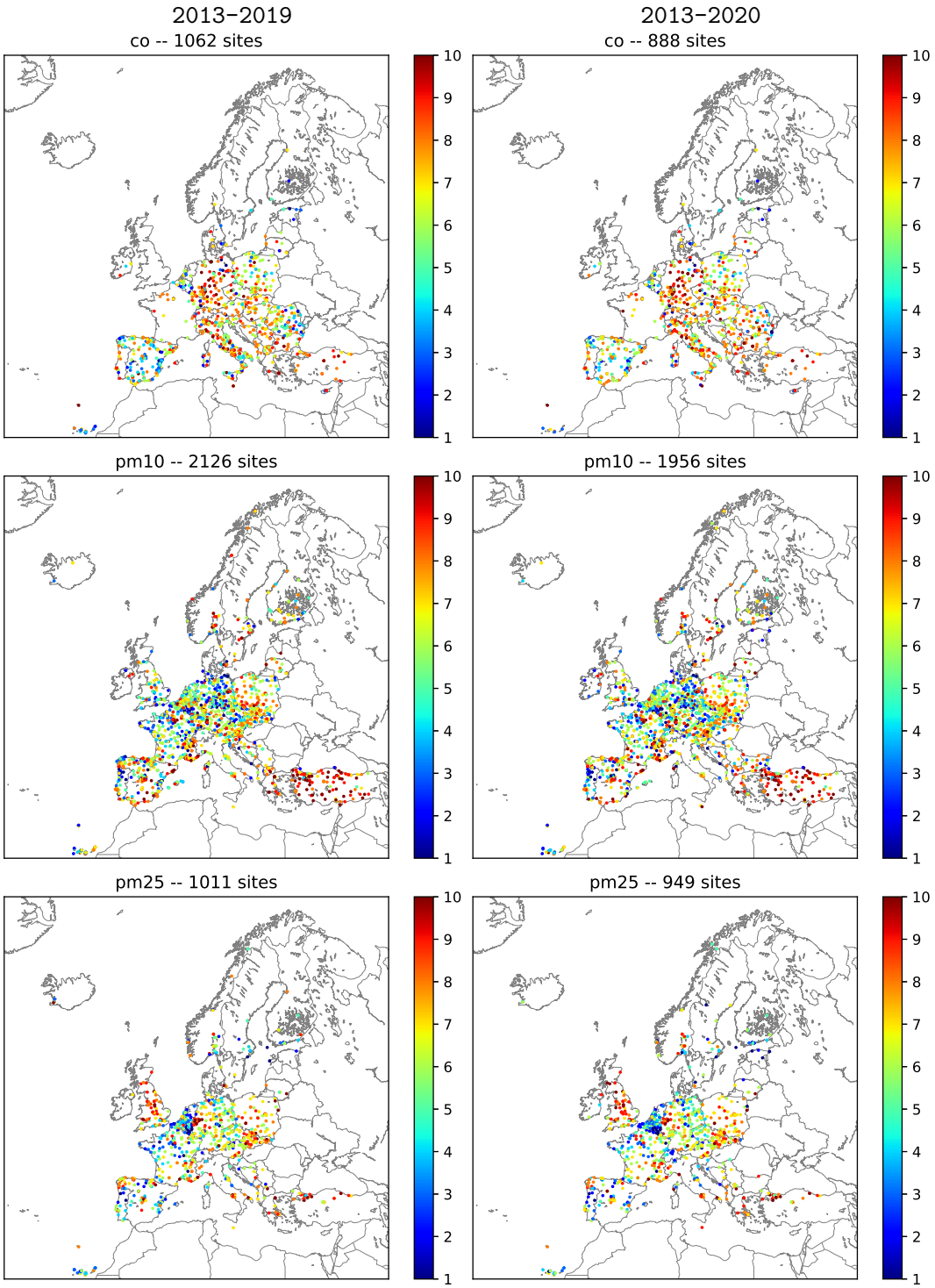

Figure 5 – Cartographie de la classification obtenue. À gauche, pour la précédente classification ; et à droite, pour la nouvelle version.

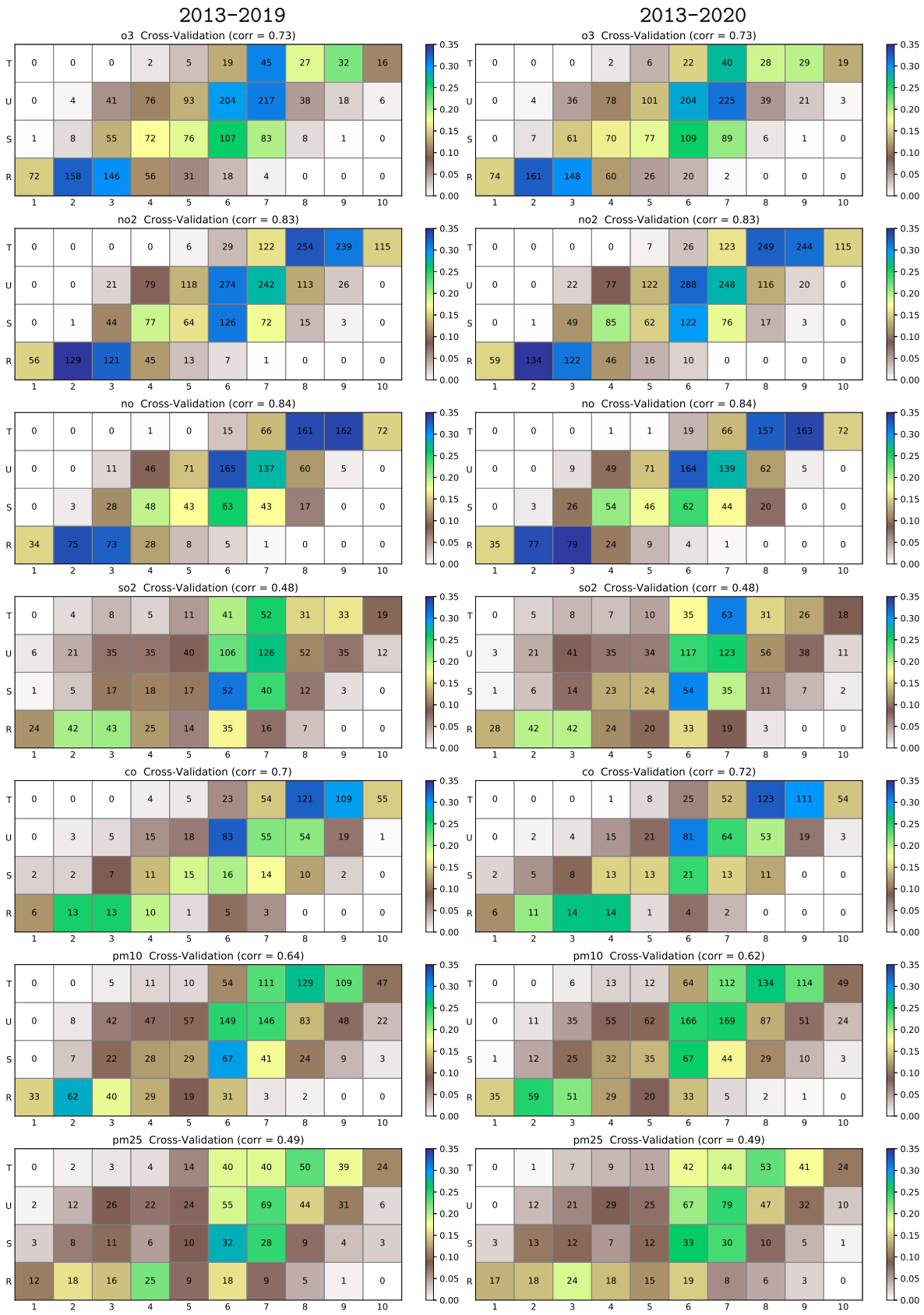

Figure 6 – Validation croisée : nombre et pourcentage (en couleur) dans chaque classe pour chaque type de station. À gauche, pour la précédente classification ; et à droite, pour la nouvelle version.

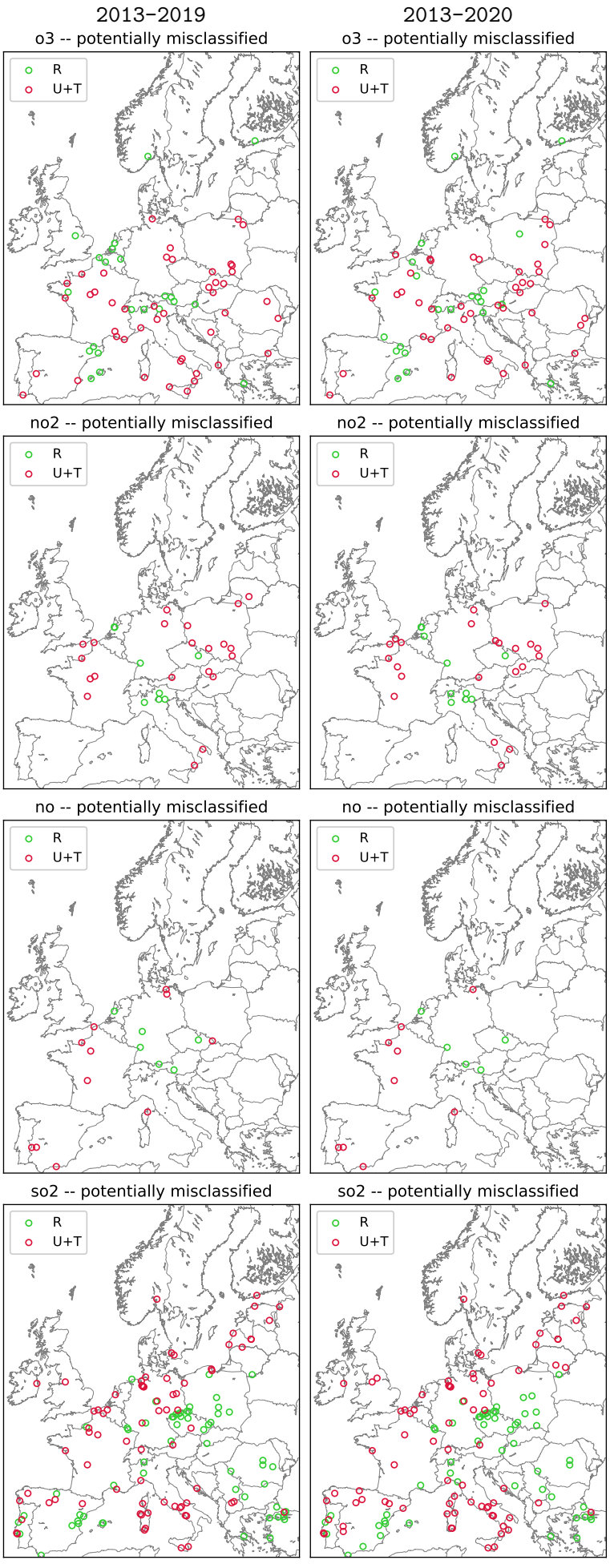

Figure 7 – Stations R qui se retrouvent dans les classes 6-10, et stations U et T qui se retrouvent dans les classes 1-3. À gauche, pour la précédente classification ; et à droite, pour la nouvelle version.

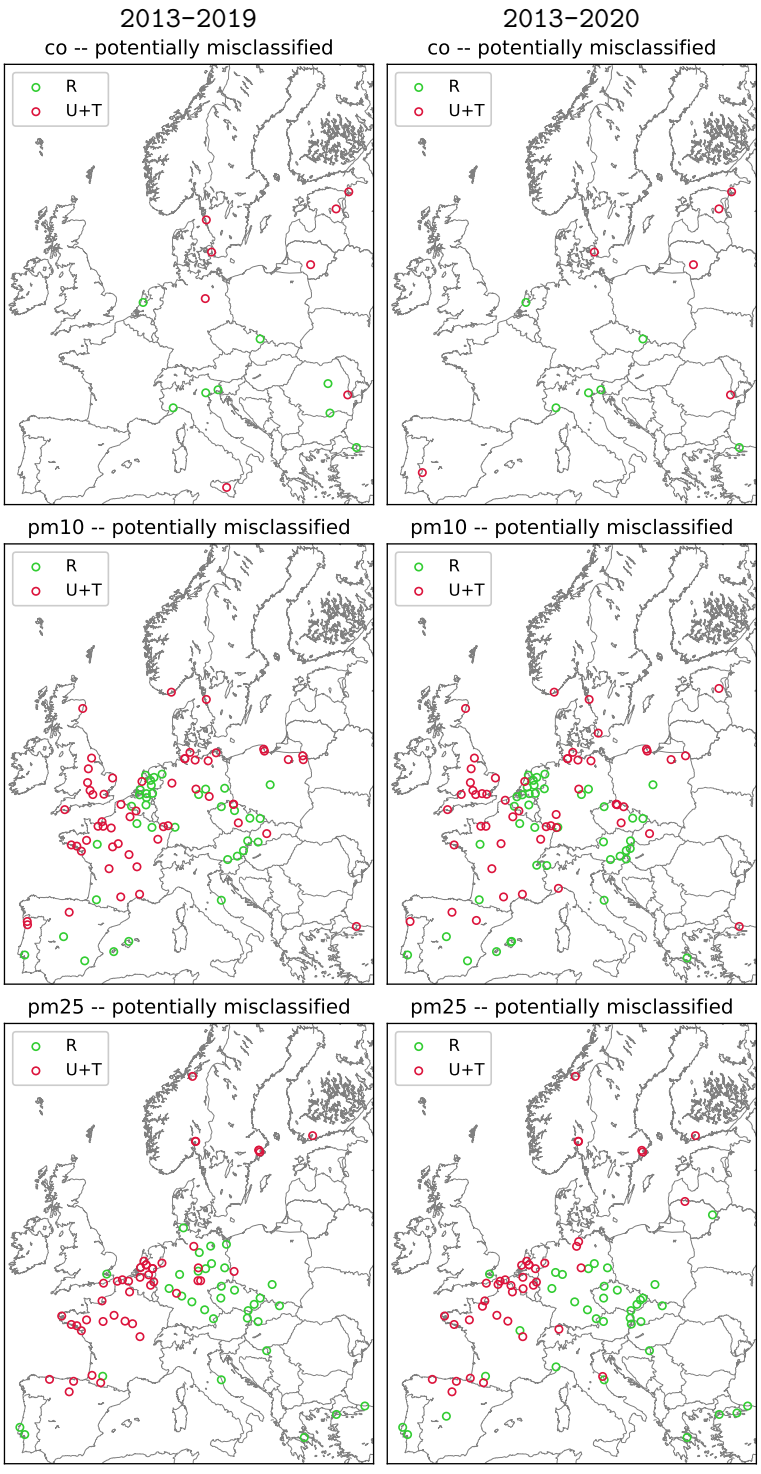

Figure 8 – Stations R qui se retrouvent dans les classes 6-10, et stations U et T qui se retrouvent dans les classes 1-3. À gauche, pour la précédente classification ; et à droite, pour la nouvelle version.

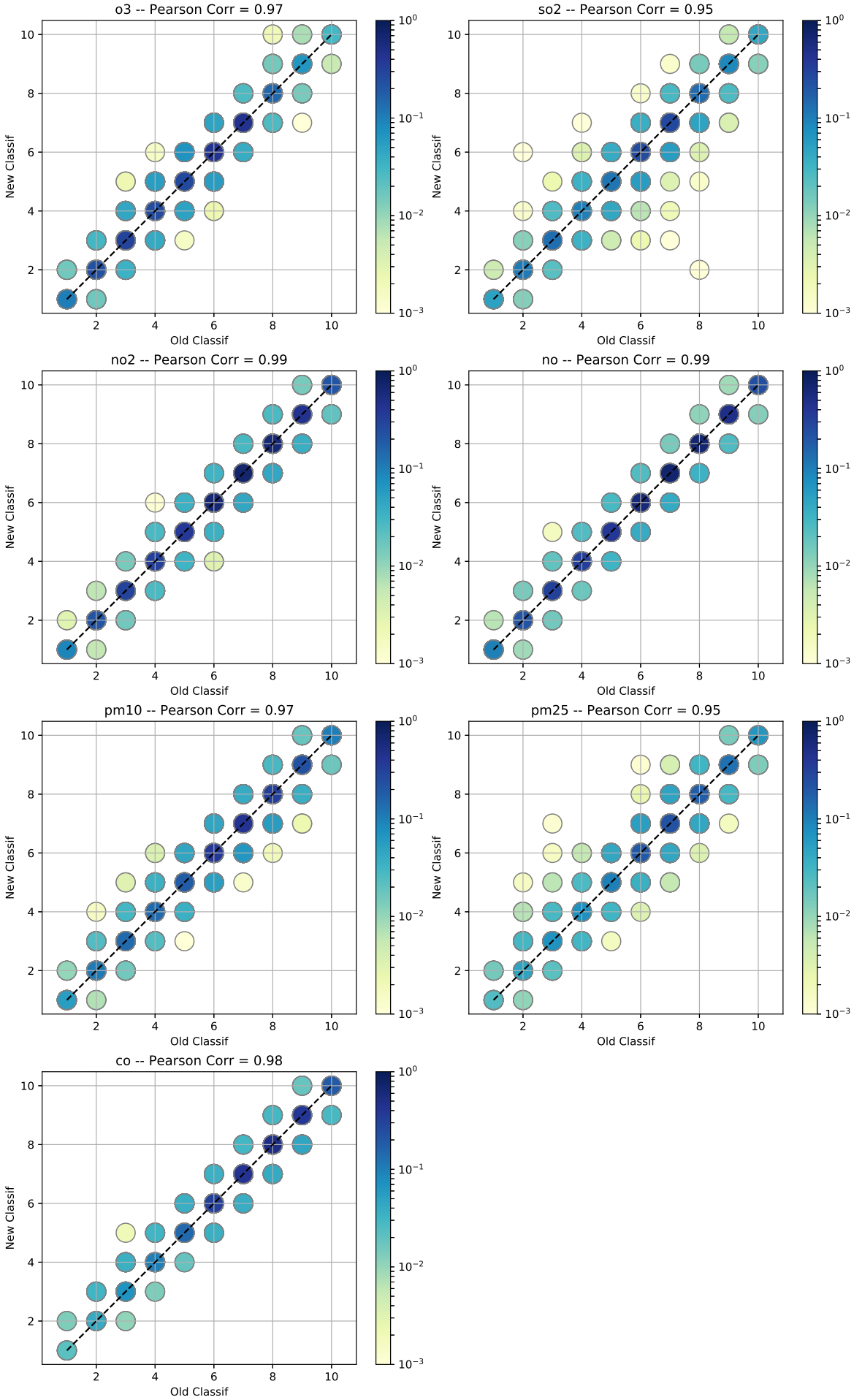

Figure 9 – Scatter Plot des classes obtenues avec l'ancienne et la nouvelle classification. La couleur indique la fréquence d'occurence.

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 (PM_{10} et SO_2), peut-être en conséquence de changements de code dans les métadonnées de l'EEA. Par contre, des stations disparaissent en Espagne, Belgique, Italie, et Roumanie.

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 2020.

- Le réseau de mesure s'est légèrement étoffé pour tous les polluants, sans toutefois que de nouveaux pays apparaissent. Il y a par contre un problème de gestion des codes de stations supérieurs à 7 caractères dans les métadonnées de l'EEA.
- Comme dans la version précédente, la qualité des séries temporelles est insuffisante en certaines régions d'Italie, ou en Allemagne pour le NO : les valeurs absentes sont trop nombreuses au sein de chaque journée.
- La cohérence entre les métadonnées et la classification objective s'améliore légèrement.

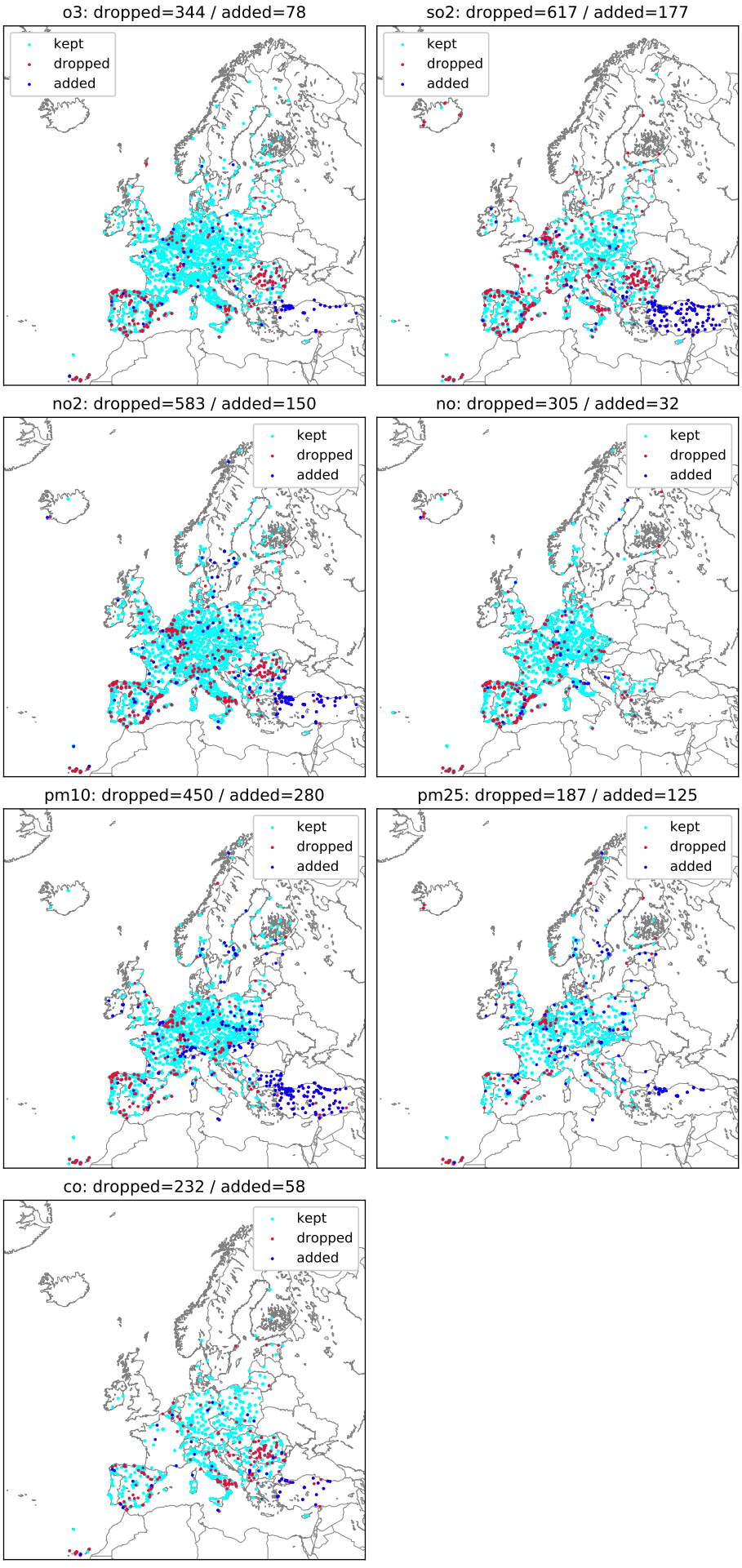

Figure 10 – Stations qui disparaissent (rouge), ou qui apparaissent (bleu) dans la nouvelle version.