Actualisation de la classification des sites de mesure

Mathieu Joly

24 mars 2017

Table des matières

1 Extraction des indicateurs

1.1 Données utilisées

- Comme dans Joly et Peuch (2012), 8 années sont considérées : ici de 2007 à 2014.
- Données AirBase V8 jusqu'en 2012, puis AQeR pour 2013 et 2014.
- Comme dans Joly et Peuch (2012), pour les stations francaises, on complète certaines séries annuelles manquantes à l'aide des données validées du LCSQA.
- On considère un domaine géographique compris pour les latitudes entre les îles Canaries et le cap Nord, et pour les longitudes entre l'archipel des Açores et la frontière est de la Turquie.
- 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.
- Polluants considérés : O_3, NO_2, NO, SO_2, CO (en test), PM_{10}, et $PM_{2.5}$ (en test).

1.2 Filtrage des valeurs extrêmes

Les événements extrêmes sont trop sporadiques et difficiles à échantillonner pour nous aider à caractériser les sites de mesure de manière fiable. Par ailleurs, ces valeurs extrêmes sont susceptibles de dégrader les résultats de l'Analyse Discriminante. Dans Joly et Peuch (2012), les fortes valeurs des indicateurs étaient systématiquement rejetées. Désormais, nous avons décidé de filtrer aussi les fortes valeurs des séries temporelles en amont du calcul des indicateurs.

Pour cela, des seuils ont été mis en place à partir des PDF des polluants (tableau 1).

O_3	$300\ \mu g.m^{-3}$
NO_2	$500\ \mu g.m^{-3}$
NO	$1000\ \mu g.m^{-3}$
SO_2	$1000\ \mu g.m^{-3}$
CO	$15\ 000\ \mu g.m^{-3}$
PM_{10}	$1000\ \mu g.m^{-3}$
$PM_{2.5}$	$600\ \mu g.m^{-3}$

Tableau 1 – Seuils d'intérêt pour les données d'entrée.

1.3 Utilisation des métadonnées

La typologie simplifiée fondée sur les métadonnées de l'EEA a été légérement remaniée. Par ailleurs, pour la phase d'« apprentissage » la méthode va s'intéresser à un domaine géographique restreint, compris pour les latitudes entre le détroit de Gibraltar et le cercle polaire arctique, et pour les longitudes entre la côte ouest de l'Irlande, et la frontière est de la Finlande.

Type R : sites qualifiés *background* et *rural* (sans changement).

Type S : sites qualifiés *background* et *suburban* (sans changement).

Type U : sites qualifiés *background* et *urban* (sans changement).

Type T : sites qualifiés *traffic* et *urban*. Dans Joly et Peuch (2012), les stations trafic péri-urbaines étaient aussi prises en compte. Ici, nous préférons miser sur plus d'homogénéïté au sein de ce groupe, dont l'effectif semble suffisant. Cela devrait faciliter l'analyse.

Type O : toutes les autres stations, ainsi que les stations en dehors du domaine géographique restreint (mais à l'intérieur du domaine d'étude plus large défini au paragraphe 1.1). Ces stations ne seront pas prises en compte pour l'Analyse Discriminante, mais seront classifiées *a posteriori*.

La figure 1 montre l'hétérogénéïté du réseau de mesure selon le polluant. On retrouve le constat de Joly et Peuch (2012) : peu de mesures de $PM_{2.5}$, et gros déficit de stations rurales pour le CO. On voit par ailleurs le poids très important des stations urbaines (U & T), excepté pour l'ozone.

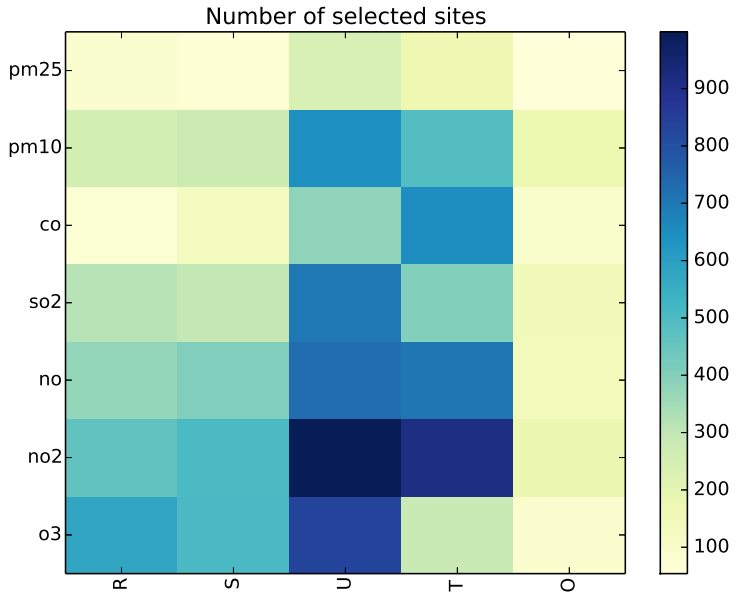

Figure 1 – Nombre de stations avec des données suffisantes, par type de métadonnée.

1.4 Nouveaux indicateurs

Dans Joly et Peuch (2012), l'Analyse Discriminante repose sur 8 indicateurs. Après réflexion et lecture de la bibliographie parue depuis, de nouveaux indicateurs sont ici testés pour affiner la classification. Au total, ce sont 25 indicateurs qui sont mis en concurrence.

1.4.1 Indice de Gini

Tapia et al. (2016) ont développé pour l'Espagne une classification des sites de mesure de l'ozone fondée sur la distribution en fréquence des données horaires. Selon les auteurs, l'indice Gini permet de distinguer 4 types de PDF d'ozone. Or la PDF de l'ozone est directement liée au rôle de titration joué par l'oxyde d'azote émis par les sources de combustion, et donc à l'environnement de la station.

À partir d'une représentation sous forme de distribution cumulée (courbe de Lorenz), l'indice de Gini est utilisé pour décrire l'inégalité de la répartition d'une variable dans une population donnée. De façon imagée, l'indice de Gini vaut 0 en cas d'égalité entre tous les individus, et 1 quand un individu a tout, et les autres rien.

Tapia et al. (2016) ne traitent que l'ozone à l'échelle de l'Espagne. Nous avons choisi d'implémenter le calcul des PDF, et de tester l'utilisation de l'indice Gini pour tous les polluants à l'échelle de l'Europe.

1.4.2 Paramètres de forme

L'indice de Gini n'étant pas nécessairement le plus approprié pour d'autres polluants que l'ozone, nous allons aussi tester l'utilisation des moments 3 et 4 de la distribution : respectivement *skewness* et *kurtosis*.

- La *skewness* mesure l'asymétrie de la distribution d'une variable. L'asymétrie d'une distribution est positive si la queue de droite (fortes valeurs) est plus longue ou grosse, et négative si la queue de gauche (faibles valeurs) est plus longue ou grosse.
- La *kurtosis* est une mesure de l'aplatissement, ou a contrario de la pointicité, de la distribution d'une variable aléatoire réelle, hors effet de dispersion (donnée par l'écart type). C'est le deuxième des paramètres de forme, avec la *skewness*.

1.4.3 Auto-corrélation

Barrero et al. (2015) ont choisi de classifier les sites de mesure du Pays Basque à partir d'une analyse de la variabilité des PM_{10}. Ils utilisent en particulier la fonction d'auto-corrélation des séries horaires et quotidiennes. Le corrélogramme permet de quantifier la persistance des variations et de détecter des périodicités dans les séries temporelles.

Pour notre classification, les périodicités diurnes et saisonnières sont déjà quantifiées par plusieurs indicateurs. Par contre, évaluer la persistance des anomalies mérite d'être testé. Pour cela, nous avons choisi comme indicateur l'auto-corrélation au $lag = 42$, qui se situe en dehors des périodicités 12h, 24h, 36h, 48h, etc, liées au cycle diurne. De surcroît, nous avons calculé cet indicateur à partir des séries filtrées du cycle diurne glissant (cf. Joly et Peuch, 2012).

1.4.4 Cycles diurne et annuel

Plutôt que l'écart *max - min* pour caractériser l'amplitude du cycle diurne, et le rapport *été / hiver* pour l'amplitude du cycle annuel, on va tester l'utilisation de l'écart-type, afin de distinguer la présence de plusieurs pics, ou de fluctuations aux inter-saisons.

Dans Joly et Peuch (2012), le cycle diurne est décrit par son amplitude et son maximum. Nous ajoutons ici le minimum quotidien qui, même s'il est certainement très corrélé aux autres indicateurs, possède une part d'information qu'il peut être utile de traiter indépendamment (par exemple, minimum quotidien d'ozone dû à la titration par les NO_x émis avant le lever du jour par les émissions du trafic).

1.4.5 Effet Weekend

Dans Joly et Peuch (2012), l'effet Weekend était détecté pour la moyenne, le maximum, et l'écart-type quotidien. Nous ajoutons ici, à titre d'essai, le minimum quotidien, pour les mêmes raisons qu'au paragraphe 1.4.4.

On teste par ailleurs un changement dans le mode de calcul. Plutôt que de calculer la moyenne des diagnostics du dimanche, et diviser par la moyenne des diagnostics du lundi, les ratios vont être calculés pour chaque couple de dimanche et lundi, puis l'indicateur sera la médiane.

1.5 Sélection des séries temporelles

Comme pour Joly et Peuch (2012), il faut au minimum 365x24 valeur horaires pour qu'une série temporelle soit prise en compte. Ensuite, pour le calcul des indicateurs, quelques critères ont évolué :

- Le quota de valeurs absentes (20%) toléré pour le calcul du cycle diurne sur 31 jours glissants était redondant avec le nombre minimal de valeurs (20 jours) pour calculer une valeur mensuelle. On utilise désormais un minimum de 20 valeurs dans les deux cas.
- Pour le calcul des valeurs quotidiennes (avant calcul d'une moyenne mensuelle), et pour le calcul de l'effet Weekend, pour Joly et Peuch (2012) il était nécessaire de disposer des 24 valeurs de la journée, sans aucune donnée absente. Il s'avère que dans les données AQeR récentes un nombre important des séries temporelles (jusqu'à un tiers pour certains polluants) ont une valeur absente « systématique » chaque jour. Il a donc fallu assouplir la règle, et tolérer désormais une valeur absente pour le calcul de ces valeurs quotidiennes.

1.6 Transformation des indicateurs

Dans Joly et Peuch (2012), une transformation logarithmique est appliquée aux indicateurs dont la distribution est très asymétrique. Le choix était subjectif, en fonction des PDF de chaque indicateur.

Afin d'automatiser cette étape de la procédure, et en raison du nombre plus important d'indicateurs ici testés (multipliés par le nombre d'espèce qui a augmenté), nous avons choisi d'utiliser la technique de Box et Cox (1964).

Noter qu'en cas de présence de valeurs négatives dans les valeurs de l'indicateur, une correction préalable est appliquée (soustraction du minimum de la distribution).

1.7 Propriétés des indicateurs

Dans certains cas (présence d'*outliers*), la technique de Box et Cox (1964) peut échouer à équilibrer la distribution. C'est ce que tente d'évaluer la figure 2.

- ☞ Les distributions de l'ozone et des PM ne posent aucun problème.
- ☞ La nouvelle façon de calculer l'effet Weekend entraîne des distributions plus « piquées » pour le SO_2 et le CO. L'indicateur ratiomin2 pose aussi problème pour le NO.

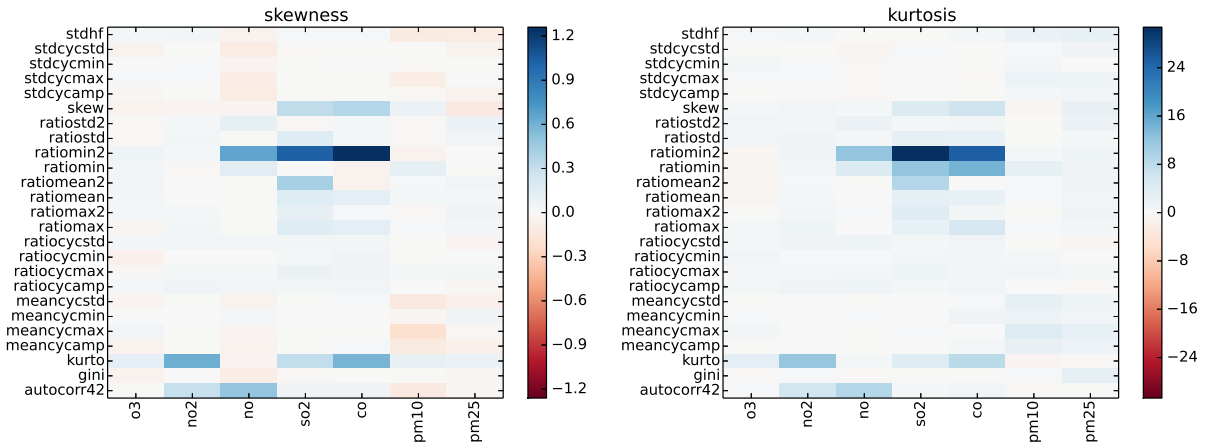

Figure 2 – Skewness et Kurtosis des distributions des indicateurs transformés.

☞ Les indicateurs Skewness et Kurtosis ont des distributions non normales pour le NO_2 et le CO, ainsi que l'autocorrélation pour le CO.

Cette analyse confirme la nécessité de mettre de côté les plus fortes valeurs des indicateurs avant l'Analyse Discriminante, comme dans Joly et Peuch (2012), pour que les distributions ne soient pas trop déséquilibrées (hypothèse de normalité...).

1.8 Performance des indicateurs

Les modifications apportées aux indicateurs définis par Joly et Peuch (2012) doivent être validées, et doivent améliorer *in fine* la classification. C'est ce qu'évaluent les figures 3 et 4.

Deux tests sont ici utilisés, car ils reposent sur des hypothèses différentes. Il sont appliqués aux binômes de populations *a priori* les plus difficiles à séparer : R & S, S & U, et U & T.

T-test de Welch : adaptation du T-test de Student, utilisé pour tester statistiquement l'hypothèse d'égalité de deux moyennes avec deux échantillons de variances et de tailles inégales.

U-test de Mann–Whitney : teste si deux échantillons proviennent de populations de même moyenne, sans supposer que les distributions sont normales.

☞ Les stations S et U sont les plus difficiles à séparer.
☞ Les deux tests donnent des résultats globalement cohérents, mais pour certains indicateurs, il y a des différences notables. Le U-test a l'air plus tolérant.

La figure 4 fait la synthèse des performances des différents indicateurs :

☞ La classification des $PM_{2.5}$ s'annonce délicate, contrairement au CO, pour lequel une majorité d'indicateurs sépare les différents groupes.
☞ La nouvelle méthode de calcul de l'effet Weekend (radio...2) est généralement plus efficace.
☞ La nouvelle façon de détecter les différences été/hiver est nettement plus pertinente, excepté pour les PM_{10}.
☞ Calculer l'amplitude du cycle diurne moyen ou son écart-type ne change pas significativement les résultats.

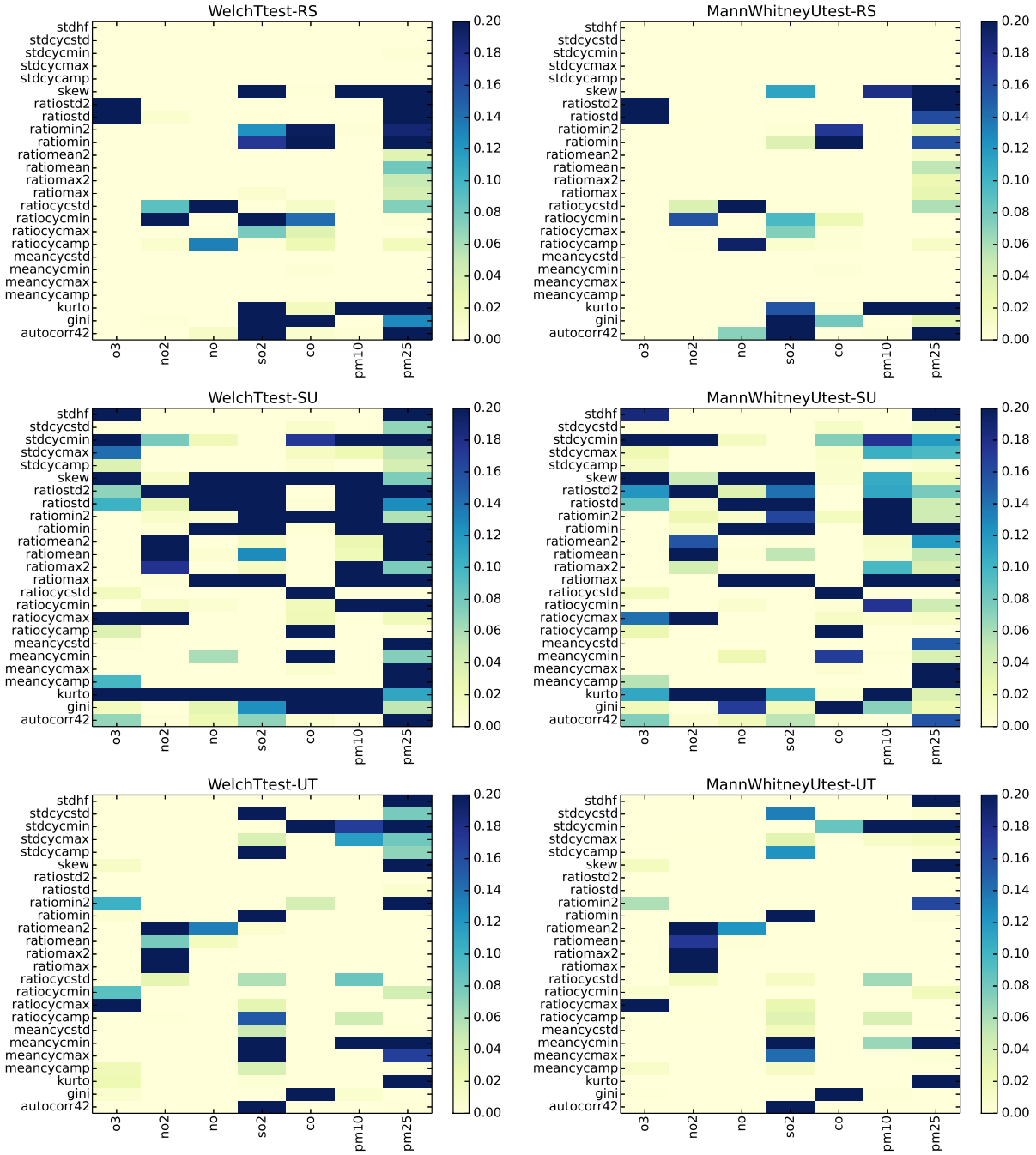

Figure 3 – P-value des tests de Welch (à gauche) et de Mann-Whitney (à droite) sur l'égalité des moyennes des groupes R & S (en haut), S & U (au milieu), et U & T (à droite).

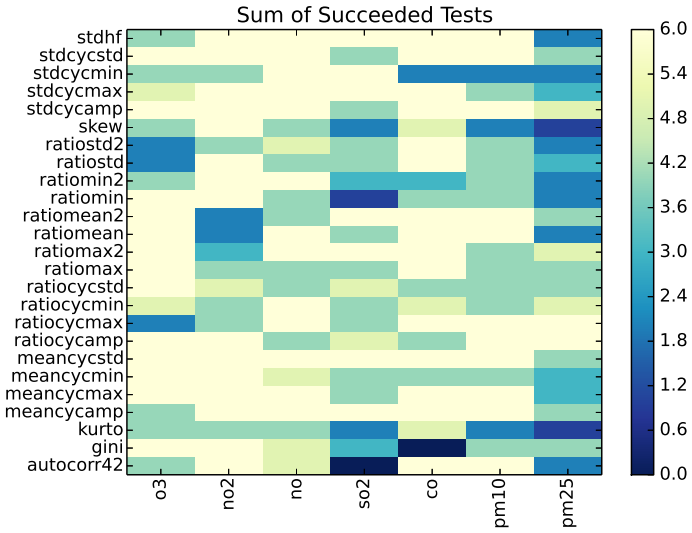

Figure 4 – Nombre de tests (Welch et Mann-Whitney) pour lesquels la P-value (pour les groupes R & S, S & U, et U & T) est inférieure ou égale à 0.05.

- ☞ Skewness et Kurtosis des distributions sont les indicateurs les moins performants, en particulier pour le SO_2 et les PM, et pour séparer les groupes S et U (figure 3).
- ☞ Si l'on somme les colonnes de la figure 4, et que l'on classe les indicateurs par performance, toutes espèces confondues, les nouvelles méthodes de calcul arrivent généralement devant les anciennes, à l'exception de stdcycmin, nettement moins bon que ratiocycmin pour le CO.

Nous avons vu au paragraphe 1.7 que la technique de Box et Cox (1964) échoue à équilibrer la distribution de l'indicateur ratiomin2 (effet Weekend sur le minimum quotidien), pour le NO, le SO_2, et le CO. La figure 4 montre que cet indicateur est peu performant pour séparer les groupes, en particulier pour les $PM_{2.5}$. Il semble donc raisonnable de l'abandonner par la suite.

1.9 Corrélations entre indicateurs

Avec 6 indicateurs en plus des 8 indicateurs de Joly et Peuch (2012), se pose la question de la redondance de l'information. En effet, il n'est pas souhaitable d'avoir en entrée de l'Analyse Discriminante des indicateurs systématiquement corrélés pour toutes les espèces. C'est ce qu'évalue la figure 5.

- ☞ L'ozone présente un comportement beaucoup moins lisible que pour les autres polluants, pour lesquels les indicateurs « de même famille » sont très corrélés entre eux (en particulier pour le NO).
- ☞ Les indicateurs Skewness et Kurtosis sont très corrélés entre eux, pour toutes les espèces sauf pour l'ozone (corrélation nulle). Pour les NO_x, Skewness et Kurtosis sont très anti-corrélés aux indicateurs du cycle diurne et à la variabilité haute-fréquence.

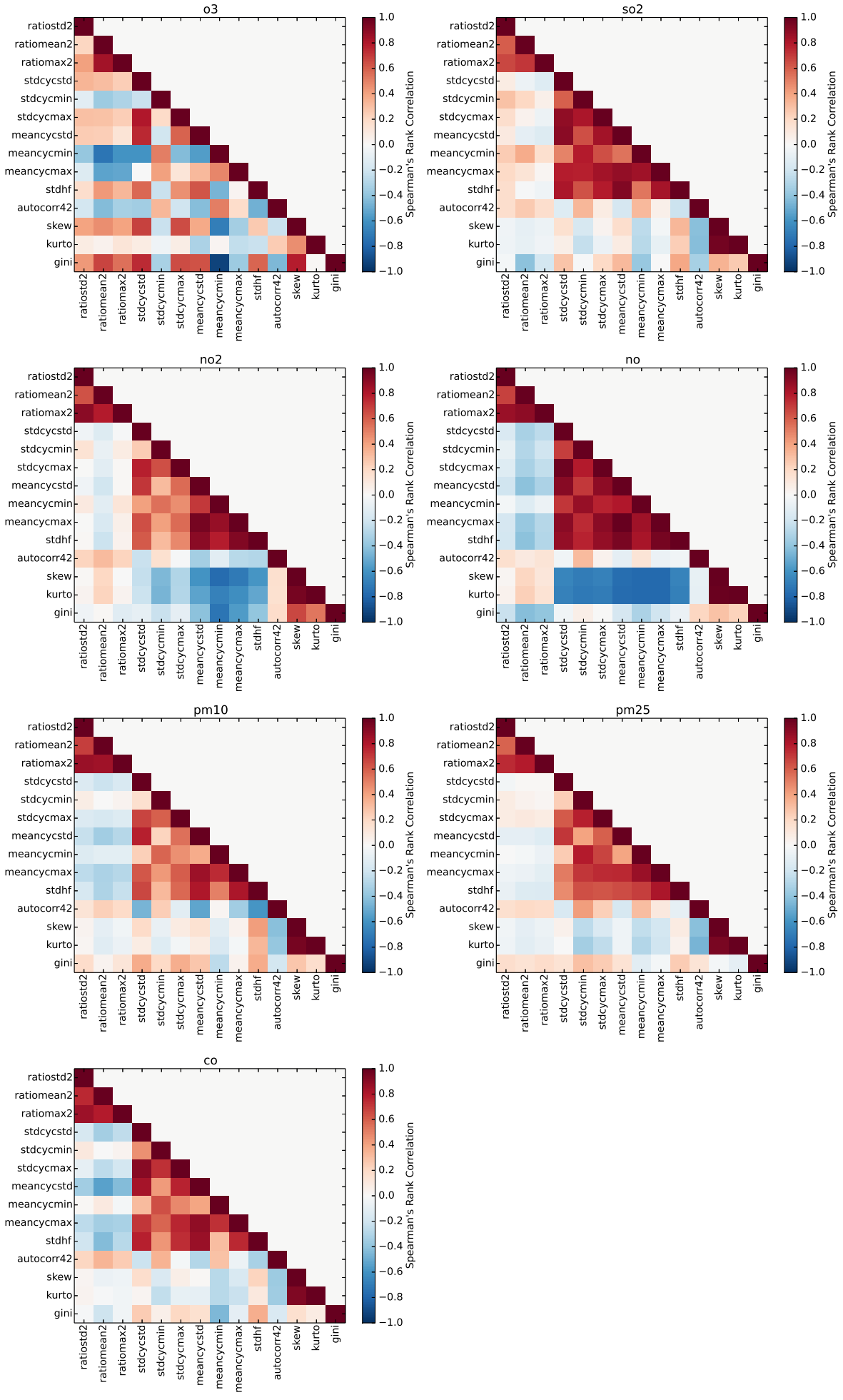

Figure 5 – Matrice des corrélations entre indicateurs.

2 Analyse Linéaire Discriminante

2.1 Indicateurs retenus

- Caractérisation de la distribution des valeurs (cf. § 1.4.1 et 1.4.2) :
  - **gini :** Gini index (inégalité dans la répartition).
  - **skew :** *skewness* (asymétrie) de la distribution.
  - **kurto :** *kurtosis* (aplatissement) de la distribution.

- Caractérisation de la variabilité (cf. § 1.4.3) :
  - **autocorr42 :** auto-corrélation pour un *lag* de 42h de la série filtrée du cycle diurne glissant.
  - **stdhf :** écart-type de la série filtrée du cycle diurne et de la basse-fréquence.

- Cycle diurne (cf. § 1.4.4) :
  - **meancycmax :** maximum du cycle diurne moyen.
  - **meancycmin :** minimum du cycle diurne moyen.
  - **meancycstd :** amplitude du cycle diurne moyen.

- Cycle annuel (cf. § 1.4.4) :
  - **stdcycmax :** variabilité du maximum du cycle diurne moyen.
  - **stdcycmin :** variabilité du minimum du cycle diurne moyen.
  - **stdcycstd :** variabilité de l'amplitude du cycle diurne moyen.

- Effet Weekend (cf. § 1.4.5) :
  - **ratiomax2 :** effet Weekend sur les maxima quotidiens.
  - **ratiomean2 :** effet Weekend sur les moyennes quotidiennes.
  - **ratiostd2 :** effet Weekend sur les écarts-types quotidiens.

2.2 Normalisation des indicateurs et détection des *outliers*

Les indicateurs, qui ont préalablement subi la transformation de Box et Cox (1964) pour se rapprocher de distributions normales, doivent aussi être normalisés avant d'appliquer une technique de réduction de dimension. Comme dans Joly et Peuch (2012) les variables sont centrées-normées en utilisant la médiane, et l'*inter-quartile range*.

Dans Joly et Peuch (2012) , la détection des *outliers* est très sommaire : les 1% de valeurs les plus fortes et les plus faibles sont systématiquement mises de côté pour chaque indicateur. Cela est justifié par le fait que pour les polluants considérés les données sont abondantes. Or ici, nous tentons d'appliquer la classification sur les données de $PM_{2.5}$ et de CO, qui sont moins nombreuses, en particulier pour les stations rurales. Une détection des *outliers* plus raffinée est donc nécessaire. En particulier, pour ensuite être en mesure de bien séparer les groupes de stations R, S, U, et T, il semble judicieux d'effectuer une détection d'*outliers* pour chacun de ces groupes séparément. Par ailleurs, des techniques statistiques évoluées sont désormais à disposition des scientifiques dans des librairies informatiques, ce qui nous a permis d'envisager une détection des *outliers* multi-variée (plutôt qu'indicateur par indicateur). Nous utilisons ici une méthode d'apprentissage supervisée de type *Support Vector Machines* (SVM).

Il s'agit d'un algorithme qui cherche à définir les distributions les plus probables d'un jeu de données. Il y a deux paramètres à spécifier :

gamma : peut être vu comme l'inverse du rayon de la zone d'influence de l'échantillon. Des faibles valeurs supposent un échantillon diffus, et de fortes valeurs supposent un échantillon compact

nu : peut être estimé par la formule $0.95 * f + 0.05$, où f est le ratio attendu d'*outliers* (entre 0 et 1).

Le ratio d'*outliers* f a été fixé arbitrairement à 0.05, et pour **gamma** des valeurs de 0.01, 0.05, et 0.1 ont été testées. La valeur 0.05 a été retenue car – comme l'illustre le tableau 2 – on obtient un nombre raisonnable d'*outliers* : autour de 10% (alors que f est fixé à 5%...).

	O_3	NO_2	NO	SO_2	CO	PM_{10}	$PM_{2.5}$
R	60 / 581	46 / 467	38 / 375	31 / 317	9 / 73	26 / 256	13 / 89
S	46 / 510	49 / 506	38 / 408	27 / 294	12 / 125	27 / 278	8 / 70
U	80 / 838	98 / 999	71 / 726	71 / 703	38 / 386	64 / 647	24 / 240
T	27 / 285	89 / 911	70 / 710	38 / 401	63 / 651	47 / 490	18 / 170

Tableau 2 – Nombre d'*outliers* comparé au nombre de stations sélectionnées.

2.3 Analyse en Composantes Principales

Nous avons vu au paragraphe 1.9 que pour la plupart des polluants, il y a des familles d'indicateurs très corrélés entre eux. Nous allons étudier la possiblité de réduire le nombre de dimensions à l'aide d'une Analyse en Composantes Principales, qui permet de trouver les directions qui maximisent la variance dans un jeu de données, et de projeter sur un sous-espace de plus petite dimension en retenant la majeure partie de l'information.

La figure 6 montre que les deux premières composantes principales expliquent près d'un tiers de la variance totale. Comme notre objectif est une réduction du nombre de dimensions du problème, en dégradant le moins possibles les données d'entrée de l'Analyse Discriminante, nous avons fixé arbitrairement à 7 le nombre de composantes retenues. Les 7 premières composantes principales permettent en effet d'expliquer 93% (pour l'ozone) à 98% (pour le NO) de la variance totale.

En projetant nos 14 indicateurs sur ce sous-espace à 7 dimensions, on a l'assurance d'avoir en entrée de l'Analyse Discriminante des variables moins redondantes.

2.4 Analyse Linéaire Discriminante

Tandis que l'ACP privilégie les directions qui maximisent la variance des données, la LDA privilégie les directions qui séparent le mieux différentes classes (algorithme dit « supervisé »).

Dans Joly et Peuch (2012), la LDA était réalisée entre les deux groupes : R et U+T. Une nouvelle amélioration est ici proposée : utiliser une LDA **multiple** pour séparer les 4 groupes R, S, U, et T.

La figure 7 illustre la projection sur les deux premiers axes de la LDA (qui expliquent la majeure partie de la variance totale). Les résultats pour le CO sont encourageants. La séparation des classes est manifestement plus difficile pour les $PM_{2.5}$, un peu comme pour le SO_2.

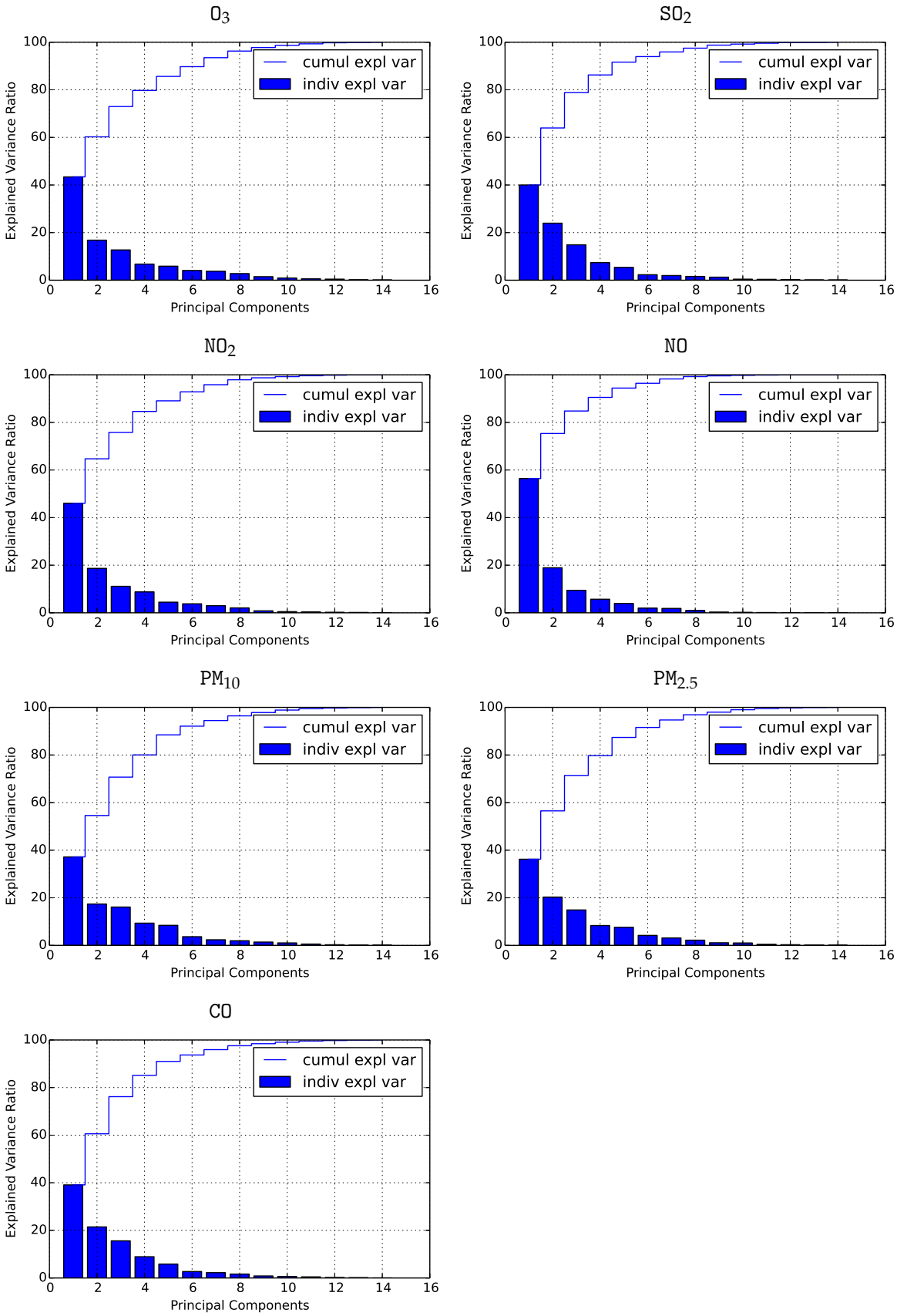

Figure 6 – Variance expliquée par les composantes principales de l'ACP.

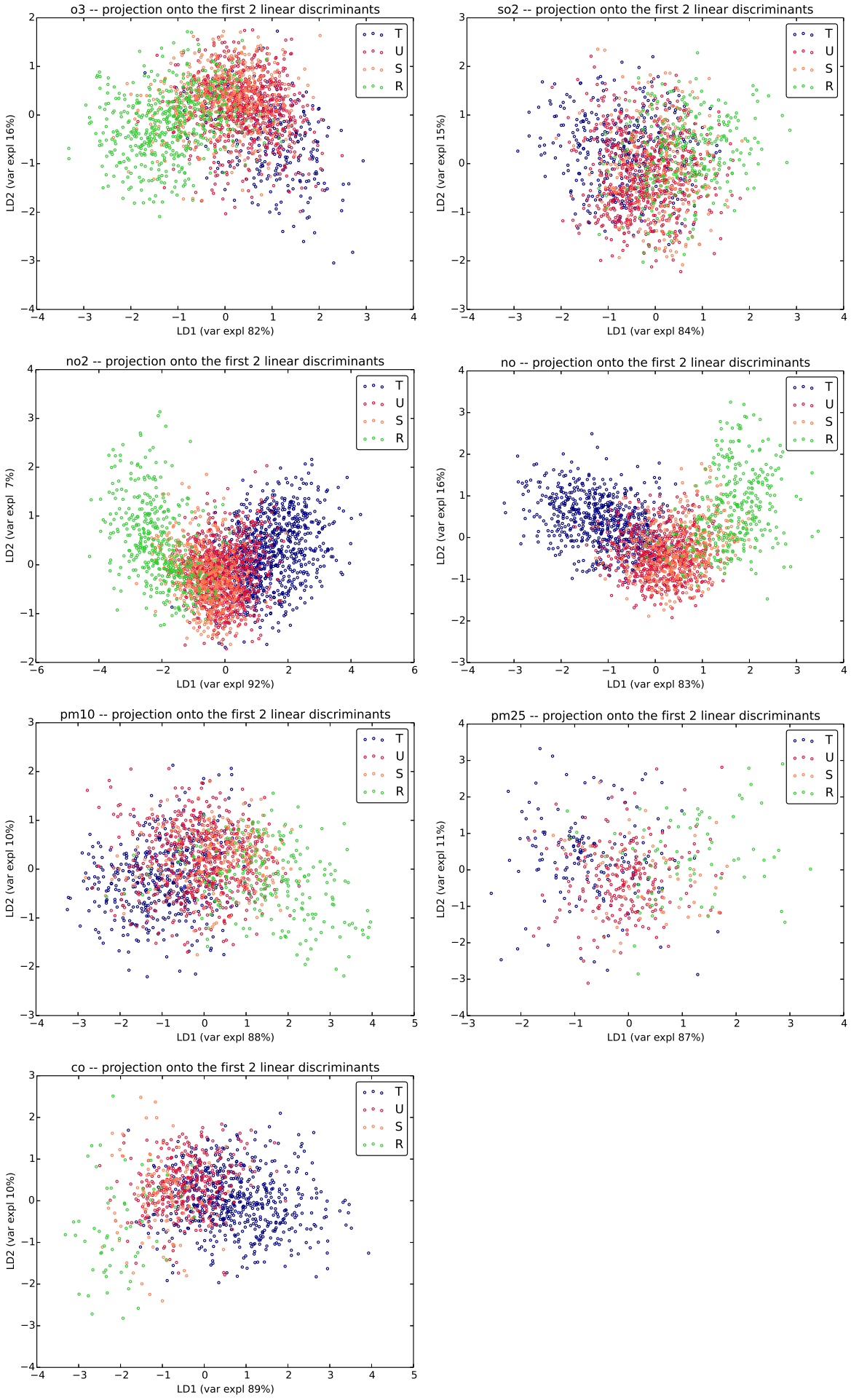

Figure 7 – Projection sur les 2 premiers axes de la LDA multiple.

Par ailleurs, comme au paragraphe 1.8, ce sont les stations S et U qui sont le plus difficile à distinguer.

La figure 8 illustre la projection sur le 1er axe, c'est-à-dire l'information que l'on va conserver pour la suite. On observe que la procédure fonctionne très bien pour le CO (qui n'était pas étudié par Joly et Peuch, 2012), ainsi que pour O_3, NO_2 et NO. Pour les autres espèces, SO_2, PM_{10}, et $PM_{2.5}$, on observe un « recouvrement » plus important des différents types (R et U en particulier).

2.5 Détermination des classes

Dans Joly et Peuch (2012), la projection sur l'axe de Fisher est « découpée » en classes selon les 9 déciles de la distribution des valeurs. Cela permet d'avoir des classes de même effectif. L'inconvénient, c'est que selon les polluants, la proportion des différents types de stations dans les données d'entrée varie fortement (cf. figure 1), ce qui rend délicate l'interprétation et la comparaison entre les classes obtenues pour les différents polluants.

Une amélioration est ici proposée, puisqu'on va tenir compte de la proportion des différents types de stations dans les données d'entrée de la manière suivante. Si dans les données on a : a% de R, b% de S, c% de U, et d% de T (en % de R+S+U+T), les bornes vont correspondre aux percentiles $\frac{15}{100}a$, $\frac{a}{2}$, a, $a+\frac{b}{2}$, $a+b$, $a+b+\frac{c}{2}$, $a+b+c$, $a+b+c+\frac{d}{2}$, $a+b+c+\frac{85}{100}$d.

Si la classification était parfaite, les stations R se retrouveraient toutes dans les classes 1-3, les stations S dans les classes 4-5, les stations U dans les classes 6-7, et les stations T dans les classes 8-10, **et ce quel que soit le polluant considéré**.

2.6 Validation croisée

La figure 9 montre :

1. L'intérêt de prendre en compte la composition des données de départ pour définir les bornes des classes. La partie rurale se trouve mieux mise en valeur, en particulier pour le SO_2. La partie urbaine est plus concentrée. L'interprétation et la comparaison entre polluants est facilitée.
2. La cohérence globale de la classification pour tous les polluants considérés.

Néanmoins, dans le détail les comportements sont variés :

Type R : une proportion importante des stations se retrouvent dans les classes 1-4. Les classes 2-3 sont généralement favorisées, mais c'est beaucoup plus « dilué » pour le SO_2 et les $PM_{2.5}$.

Type S : une proportion importante des stations se retrouvent dans les classes 3-7, avec un maximum pour la classe 6 pour toutes les espèces.

Type U : la procédure fonctionne assez bien puisque la majeure partie de ces stations se retrouvent comme prévu (par construction) en classes 6-7.

Type T : ces stations se retrouvent pour la plupart dans les classes 6-10, avec un regroupement 8-9 très net pour NO_2, NO et CO. Le comportement de l'ozone est atypique, et s'explique certainement par le fait que l'ozone étant transporté et titré par les NO_x, la pollution maximale ne se situe pas systématiquement à l'emplacement de l'émission des espèces primaires.

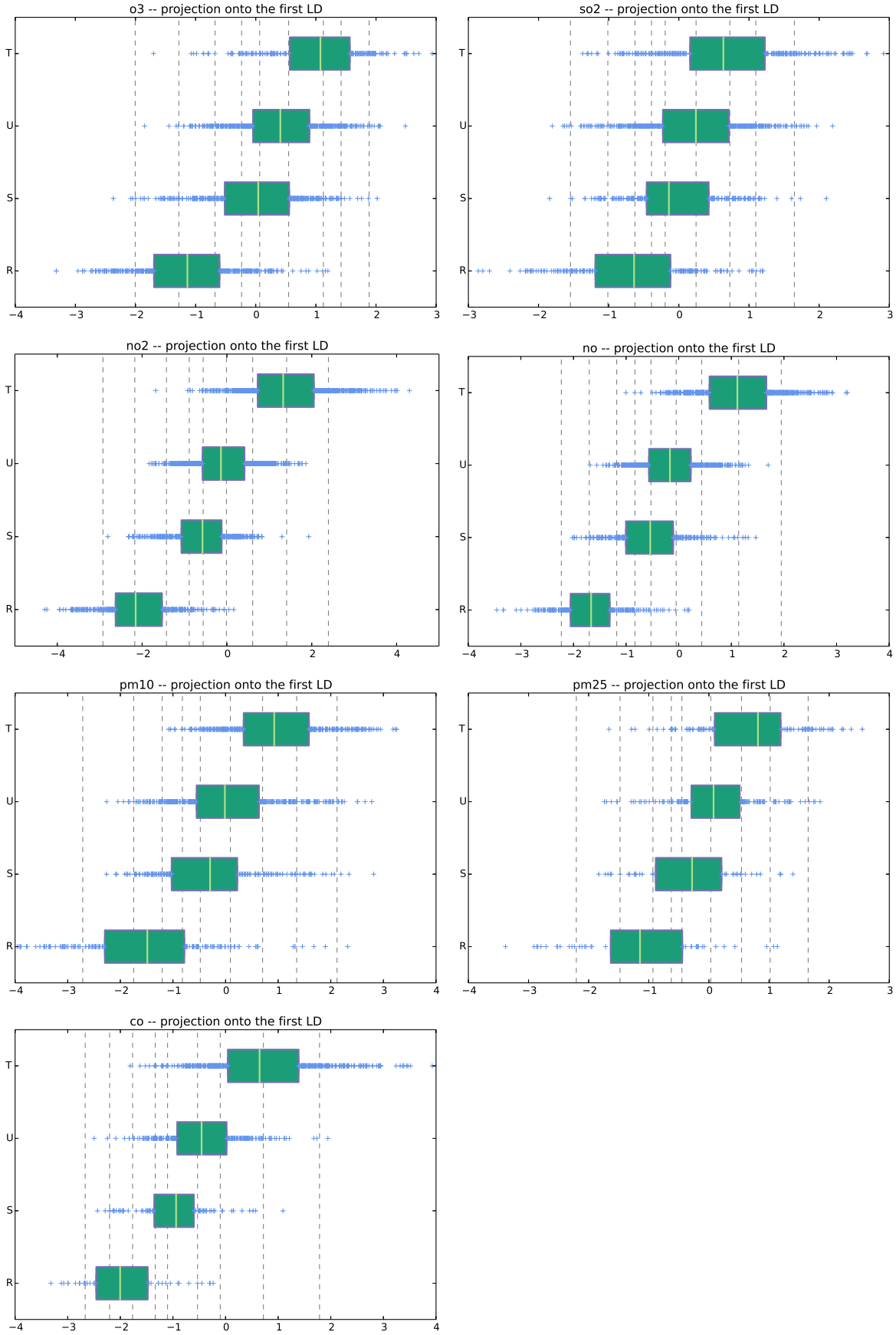

Figure 8 – Projection sur le 1er axe de la LDA multiple (après changement de signe pour certains polluants) : représentation des 3 quartiles, et des valeurs de part et d'autre. En tireté, les 9 bornes qui seront utilisées pour affecter dans les 10 classes.

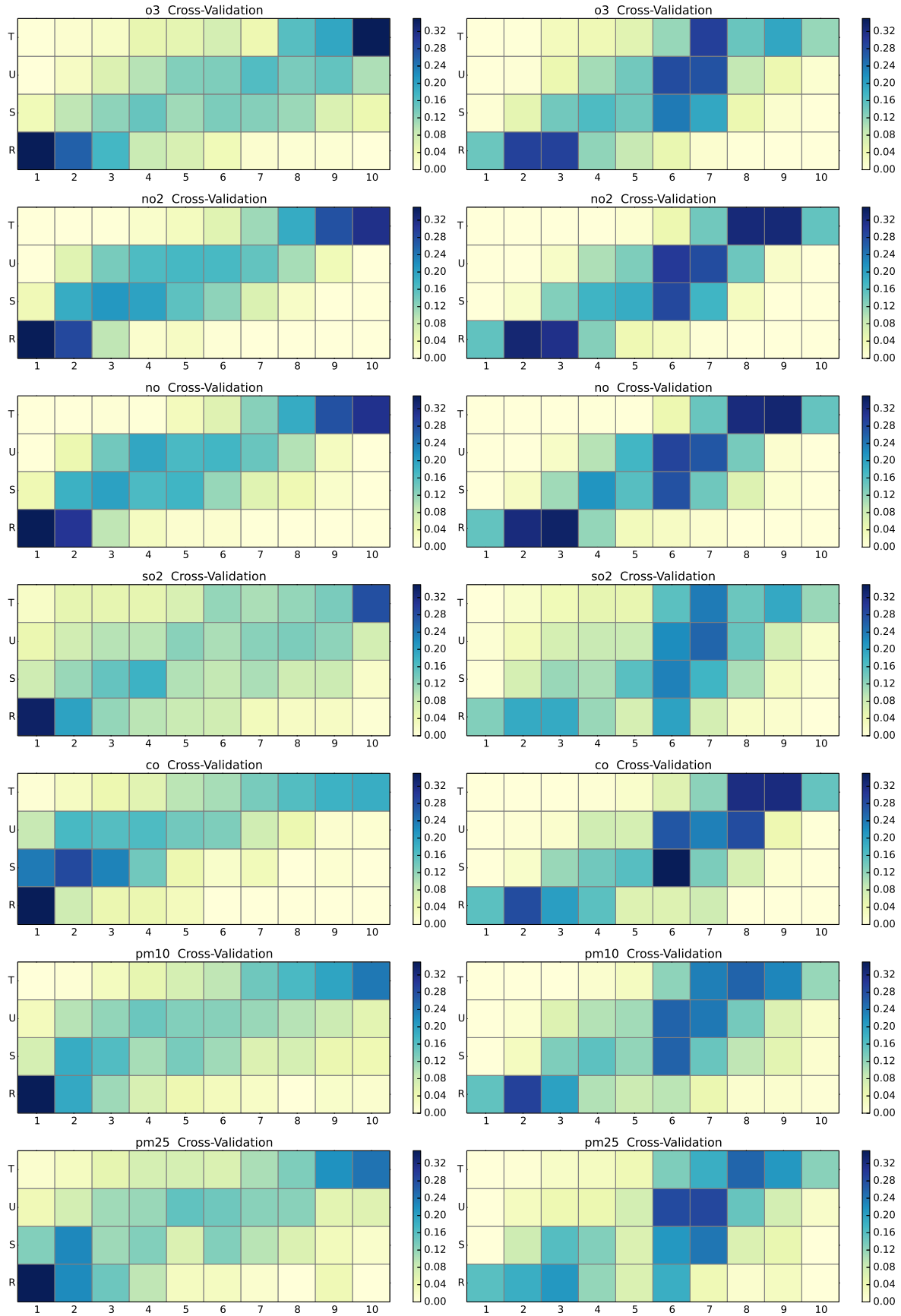

Figure 9 – Validation croisée : pourcentage dans chaque classe pour chaque type de station. A gauche, en déterminant les bornes à partir des déciles ; et à droite, avec la nouvelle méthode.

2.7 Quantification des anomalies

Une façon de synthétiser la figure 9 est de ne s'intéresser qu'aux comportements marginaux :

- 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	6%	4%	4%	31%	14%	17%	26%
S+U+T 1-3	10%	4%	4%	11%	2%	6%	8%

Tableau 3 – Pourcentage des anomalies (cf. paragraphe ci-dessus).

Le tableau 3 suggère la hierarchie suivante, en terme de performance de la classification : NO_2 et NO, O_3, CO, PM_{10}, $PM_{2.5}$, SO_2 (du plus au moins performant).

Les cartes 10 cartographient les anomalies du tableau 3. L'analyse est difficile, car il faudrait regarder localement la configuration de chacun de ces sites « douteux », et les sources de pollution environnantes. Cela n'est pas l'objet de cette étude.

2.8 Classification des autres stations

Les stations n'entrant pas dans les catégories R, S, U, ou T (cf. § 1.3), ou mises de côté au moment de l'analyse (cf. § 2.2), peuvent être *a posteriori* projetées sur le sous-espace issu de l'ACP et de la LDA. Néanmoins, Joly et Peuch (2012) ont montré que certaines stations ayant des comportements aberrants (par ex. changement d'unité au cours de la série temporelle) sont inutilisables. Pour éliminer ces *outliers* de façon automatique, nous allons utliser la même méthode de type *Support Vector Machines* (SVM) qu'au paragraphe 2.2, avec le même paramètre *gamma* et un ratio d'*outliers* cette fois fixé *a priori* à 1%. L'ensemble des stations est considéré (et non chaque type de station comme au § 2.2), mais seules les stations de type O pourront être rejetées. Le tableau 4 montre qu'au final environ 3% des stations sont marginales (parmi celles dont les données étaient suffisantes pour calculer les indicateurs), et ne doivent pas être classifiées.

O_3	NO_2	NO	SO_2	CO	PM_{10}	$PM_{2.5}$
68 / 2301	77 / 3061	74 / 2359	63 / 1864	44 / 1331	70 / 1844	28 / 623

Tableau 4 – Nombre de stations de type O rejetées, comparé au nombre total.

Les cartes de la figure 11 montrent qu'en sortie de l'algorithme, les stations éliminées sont relativement peu nombreuses par rapport à l'ensemble disponible (moins de 10%). On notera un certain nombre de *clusters* (sur l'Emilie-Romagne en Italie, la communauté valencienne en Espagne, les Pays-Bas et la frontière Pologne-Allemagne pour le SO_2, la région de Rome pour les PM, le Nord-Pas de Calais et la Sardaigne pour le CO, etc), dont certains s'expliquent : dans le cas de la région de Rome, par exemple, les données horaires de PM_{10} et $PM_{2.5}$ sont constantes tout au long de la journée (valeur quotidienne dupliquée). La responsabilité du réseau de mesure est souvent régionale en Europe, d'où des anomalies dans les jeux de données qui sont souvent régionales.

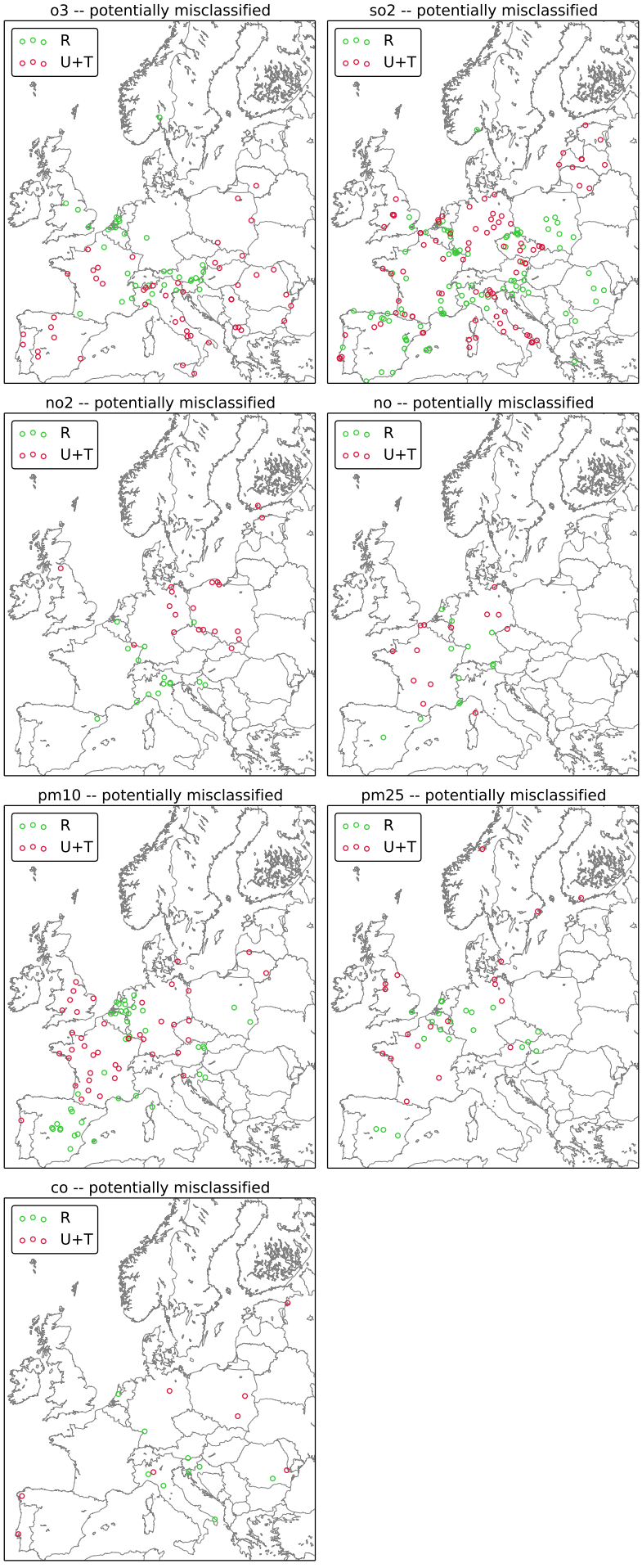

Figure 10 – Stations R qui se retrouvent dans les classes 6-10, et stations U et T qui se retrouvent dans les classes 1-3.

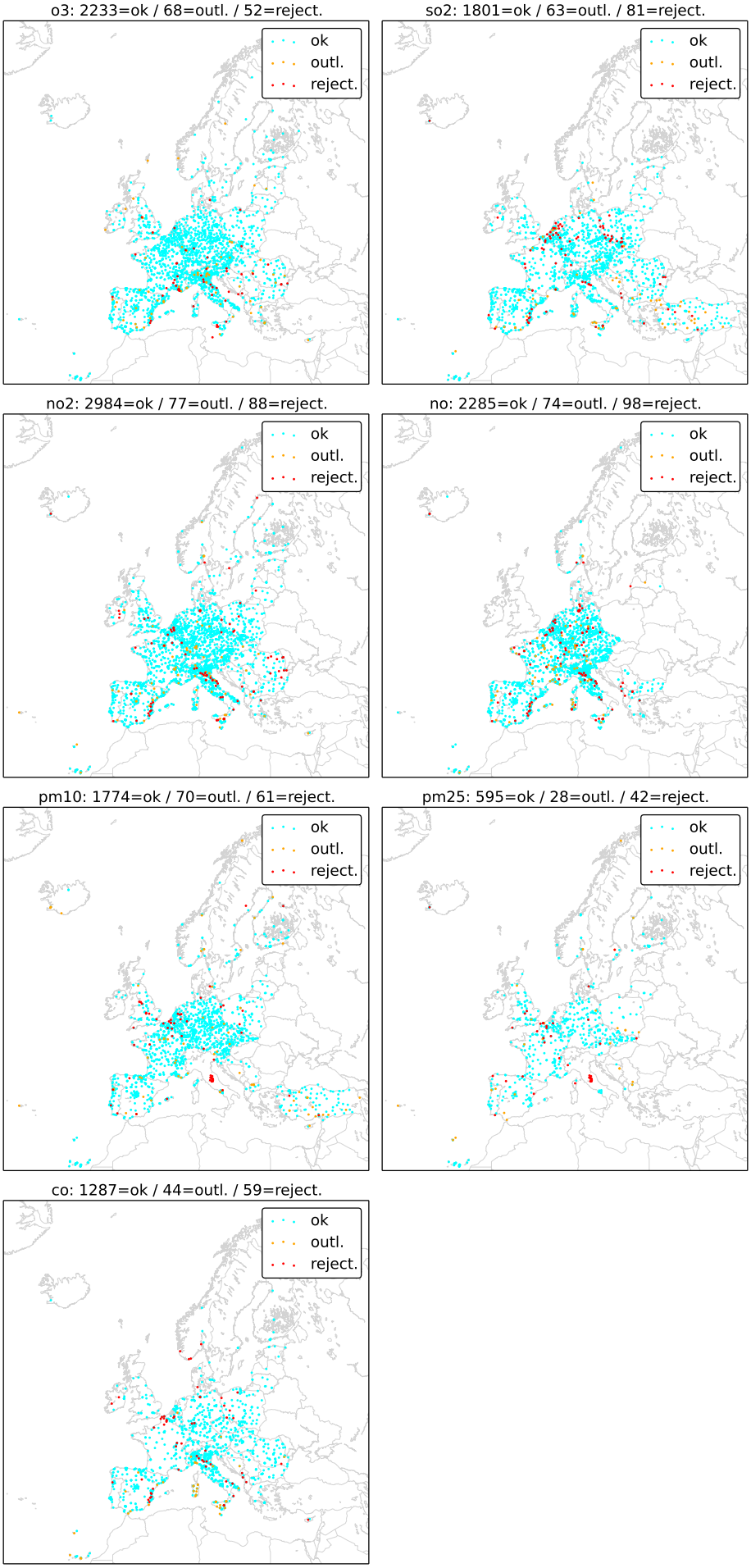

Figure 11 – Stations rejetées lors du calcul des indicateurs (rouge), ou lors de la classification (orange).

2.9 Cartographie des résultats

Les cartes 12 cartographient les classes obtenues pour chaque polluant.

- ☞ Les stations mesurant l'ozone obtiennent des classes plus faibles que pour le CO, par exemple, ce qui montre que la méthode reproduit bien les caractéristiques du réseau de mesure (cf. figure 1).
- ☞ Les classifications du NO_2 et du NO ont de fortes similitudes, en accord avec Joly et Peuch (2012).
- ☞ On observe des contrastes régionaux, avec par exemple des classes plus élevées pour les PM_{10} sur la Turquie. Par contre, les stations des îles Canaries ont pour tous les polluants des classes plutôt faibles (malgré les panaches de poussières désertiques probables pour ce qui concerne les particules).

3 Comparaison à l'ancienne classification

L'algorithme de Joly et Peuch (2012) est ici appliqué aux nouvelles données (2007 à 2014). Pour pouvoir comparer et interpréter plus facilement, quelques modifications ont malgré tout été prises en compte :

- ○ On considère un domaine géographique compris pour les latitudes entre les îles Canaries et le cap Nord, et pour les longitudes entre l'archipel des Açores et la frontière est de la Turquie.
- ○ Les stations trafic péri-urbaines sont exclues du type T (qui contient désormais uniquement les stations trafic urbaines), et se retrouvent dans le groupe O. Par ailleurs, les sites d'altitude supérieure à 1400 m, et les stations renseignées comme « industrielles » ne sont plus prises en compte.
- ○ Les bornes des classes sont désormais déterminées comme au paragraphe 1.3, en tenant compte de la proportion des différents types de stations dans les données d'entrée.

3.1 Comparaison des classes obtenues

La figure 13 compare les projections sur le 1er axe de l'analyse discriminante. La corrélation de Spearman (considérée ici du fait de la non linéarité de la relation) entre l'ancien et le nouvel algorithme est particulièrement forte pour NO_2 et NO, mais nettement plus faible pour le CO. Par ailleurs, ce sont les stations rurales qui s'écartent le plus de la droite $y = x$, avec des valeurs plus « regroupées » avec la nouvelle méthode (en particulier pour les NO_x).

Si maintenant on compare les classes obtenues, la figure 14 confirme la corrélation plus forte pour les NO_x.

3.2 Validation croisée

La figure 15 compare les « validations croisées » avec l'ancien et du nouvel algorithme. Les résultats sont globalement très cohérents, mais on notera une légère tendance à « concentrer » les stations dans certaines classes, avec le nouvel algorithme, ce qui est peut-être un effet de l'utilisation d'une analyse discriminante **multiple** (qui cherche à distinguer **chaque** groupe).

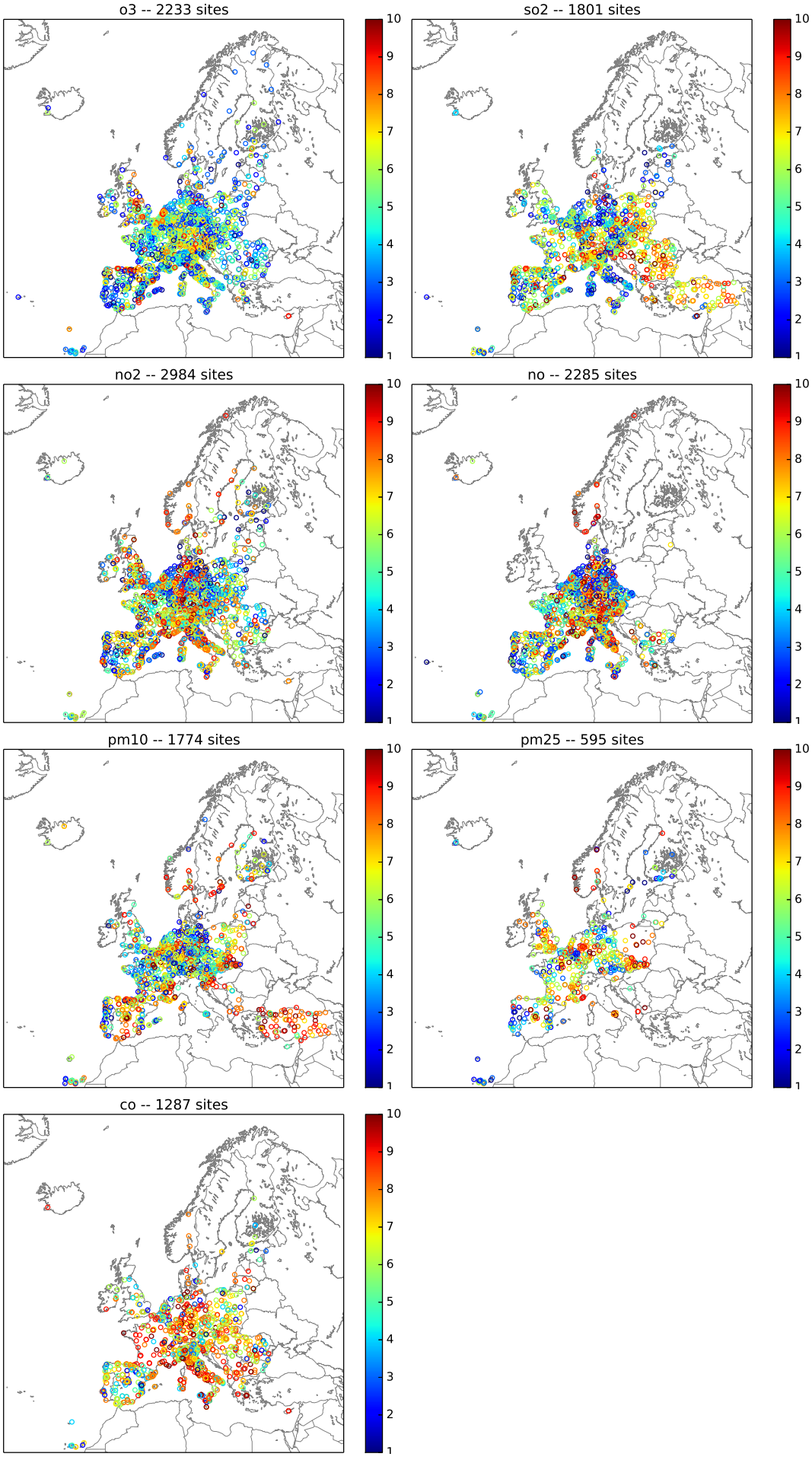

Figure 12 – Cartographie de la classification obtenue.

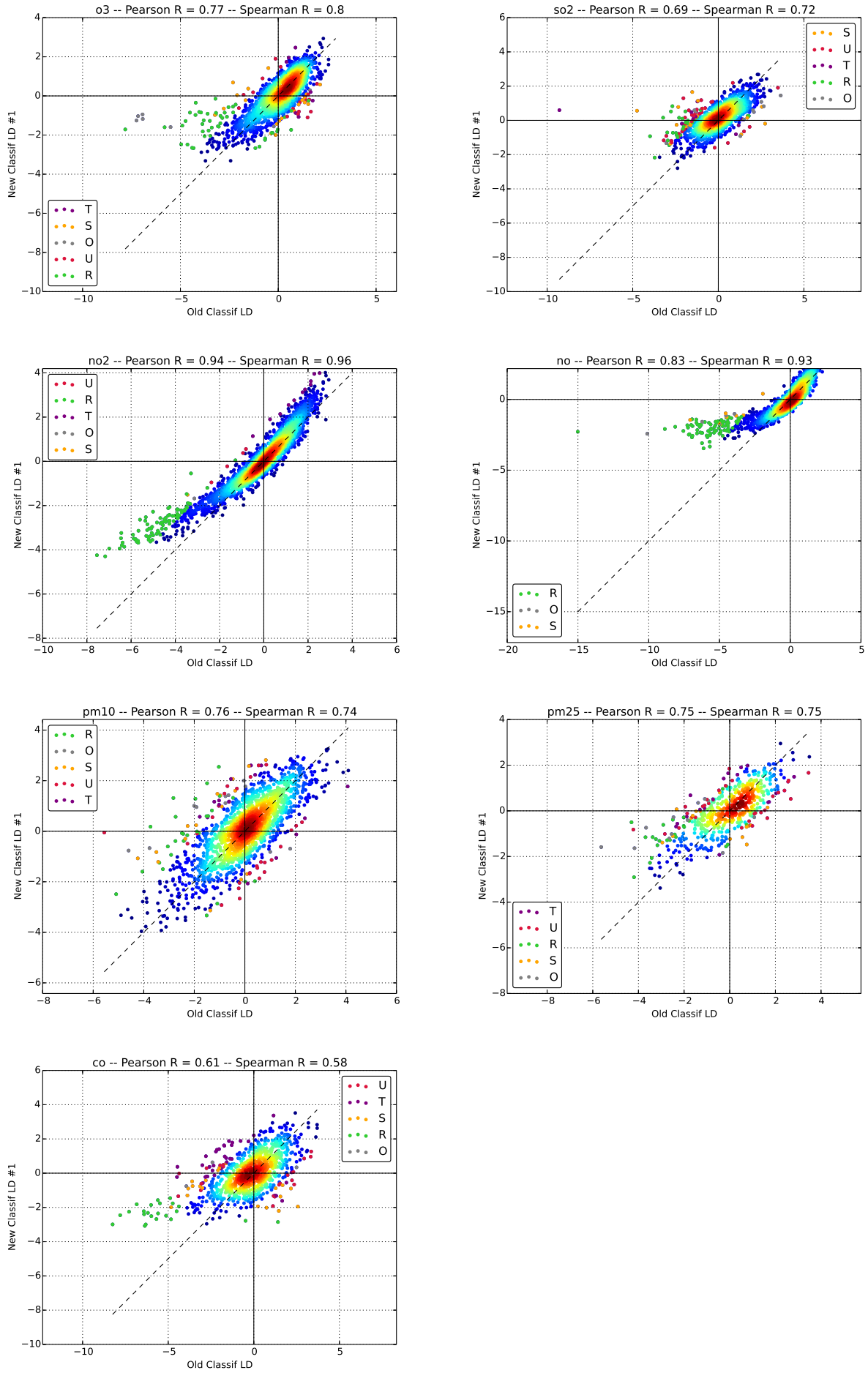

Figure 13 – Scatter Plot des projections sur le 1er axe de l'analyse discriminante. La couleur indique à la fois la densité des points, ainsi que le type de station pour les 100 stations les plus éloignées de la droite $y = x$.

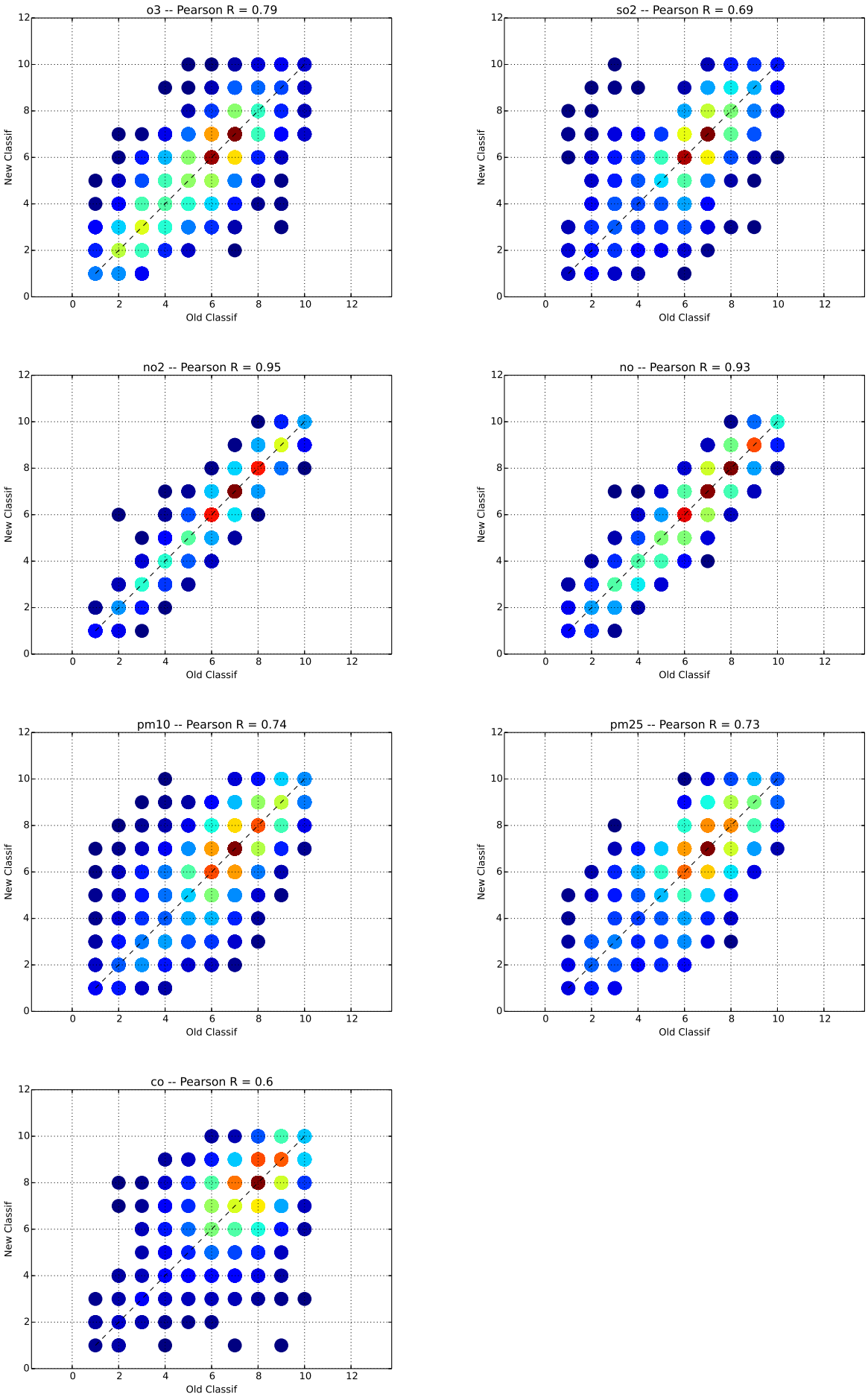

Figure 14 – Scatter Plot des classes obtenues avec l'ancien et le nouvel algorithme. La couleur indique la fréquence d'occurence.

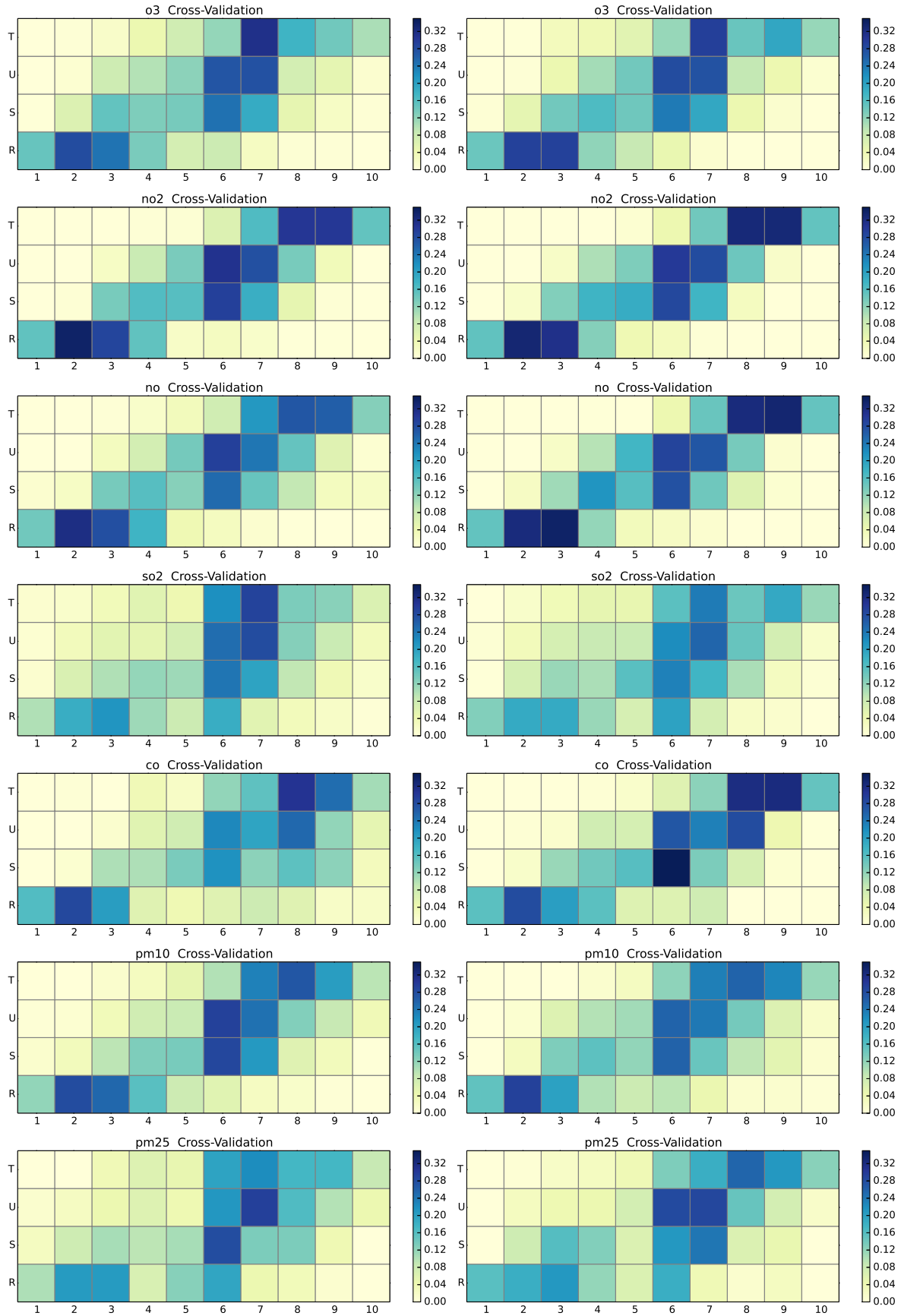

Figure 15 – Validation croisée : pourcentage dans chaque classe pour chaque type de station. À gauche, avec l'ancien algorithme ; et à droite, avec le nouvel algorithme.

Comme au paragraphe 2.7, nous allons nous intéresser aux comportements marginaux de la figure 15 :

- 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	12 → 6	5 → 4	4 → 4	30 → 31	24 → 14	11 → 17	29 → 26
S+U+T 1-3	12 → 10	4 → 4	5 → 4	10 → 11	2 → 2	6 → 6	9 → 8

Tableau 5 – Pourcentage des anomalies (cf. paragraphe ci-dessus). Évolution entre l'ancien et le nouvel algorithme (en vert pour une amélioration, en rouge pour une déterioration, et surligné de jaune quand plus de 5% des stations sont affectées).

Le tableau 5 montre une amélioration pour la plupart des polluants. Cette amélioration est nette pour l'ozone, et les stations rurales du CO et des $PM_{2.5}$. Par contre, la classification des stations rurales des PM_{10} est significativement dégradée, du fait d'un petit nombre de stations passant de la classe 4 à la classe 6 (cf. figure 13).

Références

Barrero, M., J. Orza, M. Cabello and L. Cantón (2015). Categorisation of air quality monitoring stations by evaluation of pm 10 variability. *Science of The Total Environment*, *524*, 225–236.

Box, G. E. and D. R. Cox (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 211–252.

Joly, M. and V.-H. Peuch (2012). Objective classification of air quality monitoring sites over Europe. *Atmospheric Environment*, *47*, 111–123.

Tapia, O., M. Escudero, Á. Lozano, J. Anzano and E. Mantilla (2016). New classification scheme for ozone monitoring stations based on frequency distribution of hourly data. *Science of The Total Environment*, *544*, 1–9.