



Mise en évidence et utilisation de variables cachées

Bernard Beauzamy

avril 2021

On considère une Région donnée (par exemple la Région Bretagne, pour fixer les idées), pour laquelle on dispose des informations suivantes, pour un polluant donné (par exemple le dioxyde d'azote NO₂, pour fixer les idées) :

- Emissions : des séries temporelles, au pas de temps de la journée, géolocalisées, représentant les principaux émetteurs (industries, etc.), ainsi que le trafic routier, par zone. Disons, pour fixer les idées, que l'ensemble est fait à l'échelle des communes, représentées par un code postal.
- Mesures : des séries temporelles, au pas de temps de la journée, géolocalisées, représentant les mesures qui sont faites. Les stations de mesure ne sont pas régulièrement espacées ; pour des raisons de commodité, mais aussi par choix politique, elles sont situées à proximité des grandes villes : il y en a quelques dizaines.

On voudrait reconstituer une cartographie journalière, au moins approximative, de la concentration en polluant selon les zones.

Une difficulté est immédiate : une Région n'est pas une zone de l'espace bien close. Elle reçoit une certaine quantité de polluant venant de l'extérieur et renvoie une certaine quantité de polluant, selon la journée. Intellectuellement, on peut représenter ceci en faisant figurer deux zones supplémentaires, mais les quantités concernées sont inconnues. Il n'y a rien qui permette de supposer que les entrées/sorties sont négligeables devant les productions internes.

On peut commencer de la manière suivante : on découpe le territoire en morceaux, aussi homogènes que possible, chacun contenant une station de mesure et une seule ; on affecte la mesure à tout le morceau concerné. Faisant la différence d'un jour à l'autre, on obtient la variation de concentration, pour ce polluant, sur 24 h ; sommant sur toute l'année, on obtient la variation annuelle : elle est certainement très faible.

Les émissions étant par principe positives, la somme des émissions ne peut être égale à la variation annuelle. On s'aperçoit ainsi qu'il manque une variable, élément d'information essentiel, qui caractérise la disparition locale du polluant, soit par absorption, soit par transformation en d'autres composés chimiques. On notera A cette variable.

Le 1^{er} janvier, pour une zone particulière ou pour l'ensemble de la Région, on a une mesure M_1 de polluant. Les émissions ont été E_1 et les absorptions A_1 . Par conséquent :

$$M_1 = M_0 + E_1 - A_1$$

(On suppose ici les mesures faites en fin de journée ; M_0 est la mesure reportée du 31 décembre précédent)

Plus généralement, pour tout n :

$$M_n = M_{n-1} + E_n - A_n$$

On peut partir d'un jour donné (par exemple le 1^{er} janvier) et on en déduit par récurrence toutes les absorptions les jours suivants. La série temporelle A_n est généralement inconnue du public, qui ne soupçonne même pas qu'elle puisse exister. Il est donc particulièrement intéressant d'en faire une étude systématique.

On dispose des données nécessaires pour faire l'étude de la loi conjointe du couple (E_n, A_n) : émissions, absorptions. Cette loi conjointe permettra en particulier de savoir si les deux variables varient dans le même sens, si l'un est négligeable devant l'autre à certaines époques, etc. Comme les mesures ne peuvent pas tendre vers l'infini et sont plus ou moins cycliques, on a sur une année $\sum E_n \approx \sum A_n$.

On peut raisonnablement faire l'hypothèse suivante : il y a en permanence un "seuil d'équilibre", qui dépend de la température et donc de la saison. Lorsque la concentration excède ce seuil d'équilibre, la Nature met en œuvre des phénomènes d'absorption du polluant considéré ; il n'y a pas de raison de supposer le phénomène linéaire.