Estimation des émissions polluantes d'une installation industrielle par une approche de type « bottom-up »

Pollutant emission estimation for an industrial installation using a "bottom-up" approach

Anda IONESCU*, Yves CANDAU*

Résumé

La législation environnementale impose aux grandes installations industrielles de déclarer leurs émissions rejetées dans l'atmosphère. La mesure de ces émissions étant onéreuse et techniquement difficile, plusieurs méthodes d'estimation de ces émissions sont acceptées comme alternatives à la mesure directe, l'une d'entre elles étant celle de corrélation. La méthode de corrélation nécessite une campagne spéciale de mesure de ces émissions, à partir de laquelle on établit un modèle empirique par rapport aux variables caractérisant le fonctionnement de l'installation. Cette étude est consacrée au développement d'une méthode de corrélation basée sur un modèle linéaire (régression multiple) ou non linéaire, de type réseaux de neurones, pour estimer les émissions de CO_2 et NO_2 d'un four de réchauffage d'un laminoir de l'industrie sidérurgique. Les résultats obtenus sont très satisfaisants, de l'ordre de grandeur de l'erreur de mesure pour les deux polluants en utilisant les réseaux de neurones, et assez satisfaisants pour le CO_2 dans le cas du modèle linéaire. Les résultats intégrés sont comparés à la valeur globale obtenue en utilisant la méthode du facteur d'émission de l'installation.

Mots clés

Émissions polluantes. NO_2. CO_2. Méthode de corrélation. Réseaux de neurones artificiels. Régression linéaire multiple. Facteur d'émission. Four de réchauffage d'un laminoir.

Abstract

Environmental legislation requires air pollutant monitoring of large industrial installations emissions. Most of the available measurement techniques being expensive and technically difficult, some other estimation methods are accepted, alternatively, one of them being the correlation one. The correlation method requires a special monitoring campaign of the emissions, which are further used to build an empirical model, in relation with the main process variables. This study is devoted to the development of a correlation method based on a linear model (multiple regression) or a non linear one, e.g. artificial neural networks, in order to estimate the CO_2 and NO_2 emissions of a re-heating furnace used in the steel industry. The performance of the neural networks is very good for both types of emissions, being comparable to the measurement error. For the CO_2 emissions, the linear model gives satisfactory results, too. The results are compared to the global value obtained using the emission factor method.

Keywords

Fume emissions. NO_2. CO_2. Steelworks process modelling. Correlation method. Artificial neural networks. Multiple linear regression. Emission factor.

* *Centre d'études et de recherche en thermique, environnement et systèmes – Université Paris Est Créteil – 61, avenue du Général de Gaulle – F-94010 Créteil Cedex – E-mail : ionescu@u-pec.fr – candau@u-pec.fr*

1. Introduction

Les plus importantes installations industrielles françaises sont tenues de déclarer leurs rejets atmosphériques ; elles sont assujetties à une taxe selon le principe « pollueur-payeur ». Une discussion plus ample est présentée dans le même numéro de cette revue [1].

La législation environnementale française offre quatre possibilités d'évaluation des émissions rejetées.

La première méthode est la mesure en continu des émissions : ces mesures sont cependant onéreuses et techniquement difficiles à cause des températures élevées de la fumée et de la difficulté d'accès.

Une première alternative à cette mesure est une méthode basée sur un facteur d'émission, caractéristique de chaque type d'installation et chaque type de polluant ; la méthode est simple d'application, mais assez grossière. Pour les installations à combustion, un bilan de masse, basé sur les réactions chimiques de combustion peut être également mis en œuvre [2].

La dernière possibilité est la méthode dite de corrélation, basée sur une campagne ponctuelle de mesure, permettant d'établir des corrélations entre les émissions polluantes et les paramètres des processus ayant lieu dans l'installation.

Cet article présente l'établissement d'une corrélation pour une installation sidérurgique française constituée d'un four de réchauffage d'un laminoir. Cette étude a été réalisée dans le cadre du projet européen AI/EX* [3].

2. Description du cas d'étude

Le four de réchauffage du laminoir est utilisé dans l'industrie sidérurgique pour réchauffer des billettes d'acier (600 × 12 × 12 cm³), à une cadence de 150 billettes × heure⁻¹ (grande capacité). Il est divisé en 3 zones de combustion (chacune à 16 brûleurs), résultant en une capacité totale de 30 MW. Le volume est de 24 × 7 × 1,5 m³, et la cheminée monte jusqu'à 35 m. L'efficacité de combustion est de 60 %. Le fonctionnement est discontinu, il dépend de la production des billettes. Le combustible utilisé est le gaz naturel de Groningen (Pays-Bas). Un schéma du four est présenté dans la Figure 1.

Le gaz arrive aux trois zones de combustion par le circuit G. L'air de combustion – circuit A – est de l'air frais réchauffé dans le récupérateur avant d'être distribué aux zones 3, 2 et 1. La fumée – circuit F – collectée des trois zones, passe d'abord par le récupérateur, avant d'être rejetée dans l'atmosphère. Les billettes – circuit B – initialement à la température

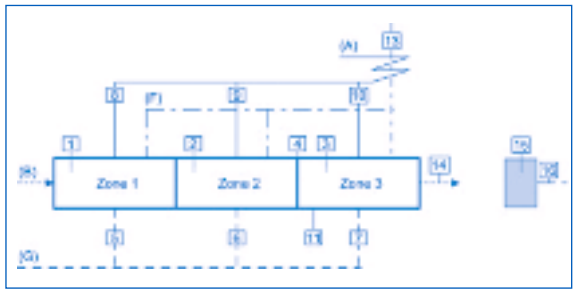

Figure 1.
Schéma du four de réchauffage du laminoir.
Billet re-heating furnace layout.

atmosphérique, passent successivement dans les trois zones de combustion, pour sortir du four à une température de 800-1 000 °C. La première zone de combustion est celle de préchauffage et la température varie entre 870 °C et 1 200 °C. La deuxième joue le rôle d'uniformisation de la température (à environ 1 200 °C), tandis que la troisième est celle de grands feux, la température étant fixée à 1 200 °C, avec une variation maximale de 40 °C. La combustion est contrôlée du point de vue stœchiométrique par un analyseur d'oxygène situé entre les zones 2 et 3, afin d'obtenir des rapports optimaux entre les débits d'air et de gaz. La combustion est effectuée avec de l'air en excès, c'est la raison pour laquelle une injection d'oxygène a été prévue.

Les paramètres du four sont présentés dans le Tableau 1. Ceux mesurés aux points 1-12 et 14-16 de la Figure 1 sont caractéristiques du fonctionnement et ils sont mesurés en continu pour le contrôle du processus (paramètres process).

Les fumées ne sont pas habituellement analysées à la sortie de la cheminée. Lors de la campagne de mesure effectuée dans le projet AI/EX [3] par le laboratoire LECES**, un analyseur de gaz a été posé à la sortie de la cheminée (point 13, Figure 1), couplé avec un système d'acquisition numérique des données. Ceci a permis de construire une nouvelle base de données, celle des émissions, les mesures étant des concentrations de O_2, SO_2, NO_2, CO, CO_2 dans les fumées (Tableau 1).

La base de données exploitée dans cette étude est formée, d'une part, des paramètres process, caractérisant le fonctionnement du four, et d'autre part, des concentrations de différents polluants mesurées dans les fumées, pendant la campagne de mesure spécialement conçue. Au total, on obtient une base de données de 25 variables, dont 21 caractérisant le process et 4 les émissions, mesurées avec un pas de temps de 1 minute. Ceci conduit à 25 540 enregistrements, pendant environ 3 semaines. La variable CO n'a pas été considérée dans la base, car il y a eu des erreurs d'enregistrement.

* Artificial Intelligence/EXpert systems for steelworks pollution controls, *ESCS-STEEL C, 7210-PR/076.*

** *LECES Environnement, Voie Romaine, Maizières-les-Metz, France.*

Tableau 1.
Paramètres du four de réchauffage et leurs principales statistiques [11].
Re-heating monitoring parameters and their main statistics [11].

		Statistiques				
		Symbole	Moyenne	Écart-type	Minimum	Maximum
1/2/3	Température à l'intérieur du four (°C) – zone 1/2/3	T_1	1 110	42,3	870	1 197
		T_2	1 199	30,3	1 017	1 278
		T_3	1 195	14,7	1 089	1 237
1/2	Pression à l'intérieur du four (mbar) – zone 1/2	P_1	9,64	1,13	1,30	14,32
		P_2	10,62	1,42	0,87	16,55
4	Concentration d'oxygène (%) – entre les zones 2–3	$O2_{23}$	2,73	2,18	0,05	14,24
	Température de l'oxygène (°C) – entre les zones 2–3	TO_2	1 209	28,3	1 070	1 276
5/6/7	Débits de gaz ($Nm^3\ h^{-1}$) zone 1/2/3	Qg_1	916	292,6	246	1 190
		Qg_2	961	435.8	304	1 785
		Qg_3	194	103,4	63	4 217
8/9/10	Débits d'air ($Nm^3\ h^{-1}$) zone 1/2/3	Qa_1	8 118	2 564	1 080	10 920
		Qa_2	8 541	3 841	2 580	16 880
		Qa_3	1 719	922	232	4 217
11	Température sous le four (°C) n° 1/n° 2	Tsf_1	87	12,5	63	128
		Tsf_2	112	19,3	82	191
12	Température de l'air de combustion après le récupérateur (°C)	Tac	294	48	153	384
	Température de la fumée après le récupérateur (°C)	Tse	391	35	159	554
13	Température de la fumée à la sortie du four (°C)	Tf_{sf}	689	61	499	857
	Concentration d'oxygène dans la fumée (%)	O_2	7,38	1,87	3,7	15,7
	Concentration de SO_2 dans la fumée ($mg\ Nm^{-3}$)	SO_2	13,19	15,05	1	122
	Concentration de NO_2 dans la fumée ($mg\ Nm^{-3}$)	NO_2	71,70	30,13	4,6	195,5
	Concentration de CO_2 dans la fumée (%)	CO_2	7,97	1,09	3,3	10,1
14	Cadence (entre deux billettes consécutives)	CAD	49,5	11,98	29	99
15	Intensité électrique du moteur	–				
16	Température en surface des billettes (°C)	TB	860	103	800	1 084

3. Prétraitement de données

3.1. Nettoyage de la base de données

Les données aberrantes doivent être repérées et enlevées de la base, car elles peuvent conduire à des artefacts. Dans cette étude, ont été considérées comme données aberrantes et supprimées de la base, les valeurs correspondantes à des périodes de calibrage des instruments de mesure (pics très importants) ou la dérive des instruments pendant le week-end, quand le four ne fonctionnait pas. Ces valeurs supprimées viennent se rajouter à d'autres valeurs manquantes, posant ainsi le problème de gestion des valeurs manquantes.

3.2. Traitement des données manquantes

C'est un problème général, auquel les auteurs proposent différentes solutions basées, à chaque fois, sur une modélisation des données [4-8].

Dans cette étude, afin de minimiser l'effet de certaines hypothèses sur les résultats, une discrimination a été faite entre les périodes longues et les périodes courtes de valeurs manquantes, selon la dynamique de chaque variable. Pour les périodes courtes, au maximum quatre valeurs consécutives manquantes, elles ont été remplacées par interpolation linéaire. Pour les périodes plus longues, tout l'enregistrement correspondant a été supprimé (l'intervalle a été ignoré de l'étude).

Après ce traitement, la base de données a été réduite de 25 540 enregistrements, à 13 947, ce qui correspond à une perte d'information de 45,4 %, avec une exception : la variable SO_2, pour laquelle, à la fin il ne restait plus que 6 763 valeurs.

4. Développement d'une méthode de corrélation

L'objectif est de trouver une relation entre les émissions du four et les paramètres process. Cette relation peut être appliquée ensuite aux paramètres process, qui sont mesurés en continu, afin de donner une estimation des émissions.

Plusieurs modèles peuvent être testés dans ce but, le plus simple étant le modèle linéaire.

4.1. Modèle linéaire (régression linéaire multiple)

La régression linéaire multiple [9, 10] a été appliquée pour estimer les concentrations d'émissions polluantes par la meilleure combinaison des variables process, du point de vue variance expliquée. Un bon indicateur de la qualité de ce modèle est l'erreur quadratique moyenne EQM définie par

$$EQM = \sqrt{\frac{1}{N} \cdot \sum_{i=1}^{N} (x_p - x_m)^2}$$

où x_p représente la valeur prédite par le modèle, x_m la valeur mesurée et N le nombre de mesures.

Le CO_2 peut être estimé de manière satisfaisante par la régression linéaire multiple (Figure 2a) : 83,4 % de sa variance est expliquée par la régression, et l'EQM est de 0,45 %, ce qui représente 5,58 % de la valeur moyenne.

Par contre, pour le NO_2 (Figure 2b) seulement 51,6 % de la variance est expliquée par la régression et l'EQM est de 20,55 µg.Nm^{-3} ce qui représente 28,66 % par rapport à la valeur moyenne.

Pour le CO_2, l'erreur d'estimation est en-dessous de la limite de sensibilité de l'appareil. Par contre, pour le NO_2 une modélisation non-linéaire s'avère nécessaire pour une meilleure estimation.

Les résultats précédents ont été obtenus en utilisant les 21 variables process. Lorsque l'on fait varier le nombre de variables utilisé, il est clair que tous les paramètres ne sont pas significatifs pour expliquer la variabilité du CO_2 ; à partir de 10 variables (les plus significatives), le coefficient de détermination reste presque constant. Avec seulement 5 paramètres (concentration d'oxygène $O2_{23}$, débit d'air dans la zone 1 Qa_1, débit de gaz dans la zone 3 Qg_3, température dans la première zone T_1, et température des fumées Tf_{sf}) on peut expliquer 80,6 % de la variance du CO_2. La température à la sortie du four (Tf_{sf}) peut à elle seule expliquer 66,9 % de la variance du CO_2, ce qui montre que c'est la température de la fumée qui caractérise au mieux le processus de combustion générant cette concentration de CO_2. Ce paramètre n'en est pas un de contrôle, mais le résultat final du processus de combustion ; néanmoins, il peut être considéré comme un indicateur brut de CO_2. Parmi les paramètres process, ceux qui sont les plus corrélés aux émissions de CO_2 sont $O2_{23}$, Qa_1, Qg_3 et T_1.

4.2. Modèle non-linéaire (réseaux de neurones artificiels)

4.2.1. Réseaux de neurones artificiels

Les réseaux de neurones artficiels (RN) peuvent modéliser des relations fortement non linéaires [7, 12,

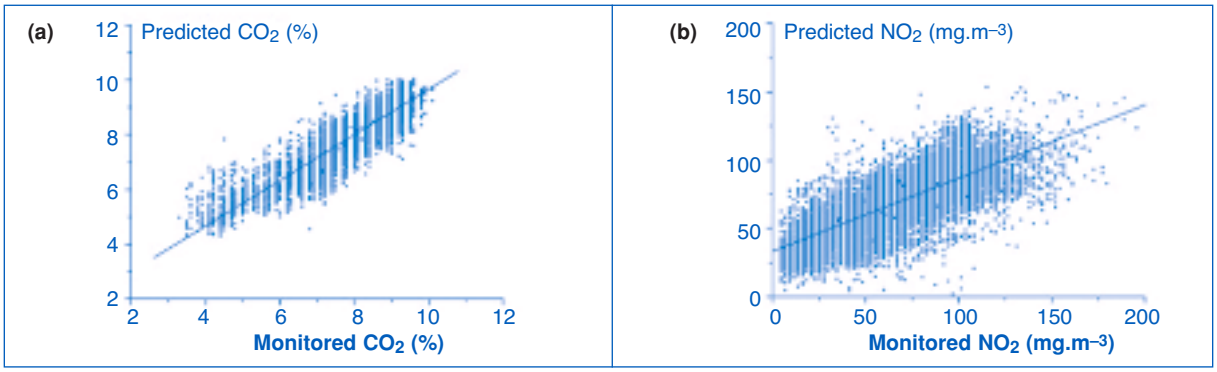

Figure 2.
(a) CO_2 estimé *(predicted CO_2)* par une régression linéaire multiple à partir de tous les paramètres process et le CO_2 mesuré *(monitored CO_2)*.
(b) NO_2 estimé *(predicted NO_2)* par une régression linéaire multiple à partir de tous les paramètres process et le NO_2 mesuré *(monitored NO_2)* [11].
(a) Predicted versus monitored CO_2. MLR result: predicted CO_2 is calculated *via* a linear combination of all the process parameters.
(b) Predicted versus monitored NO_2. MLR result: predicted NO_2 is calculated *via* a linear combination of all the process parameters [11].

13]. Le plus grand avantage d'un réseau de neurones est son potentiel à modéliser une relation non linéaire complexe sans aucune connaissance *a priori* sur sa nature [14].

Kalogirou [15] présente une synthèse de 22 applications de l'intelligence artificielle pour les systèmes de combustion, dont deux font recours aux RN : Tronci *et al.* [16] dans le cas des chambres à combustion et Ferretti et Piroddi [17] pour estimer les émissions de NO_x dans les centrales électriques. Une comparaison entre un modèle déterministe basé sur les équations de la mécanique des fluides et un modèle empirique RN a été faite par Zhou *et al.* [18] dans le cas d'une chaudière à charbon. Une conclusion générale de tous les articles cités est que les RN constituent un bon outil pour modéliser des installations industrielles complexes.

L'architecture la plus appropriée des RN pour l'approximation des fonctions est celle des perceptrons multicouches PMC *(multi-layer perceptron,* en anglais) [19-22].

La non-linéarité du PMC est obtenue en utilisant au moins une couche cachée dans son architecture, ainsi que par les fonctions d'activations non linéaires. Une bonne description du PMC (de son architecture et des équations) est présentée par Agirre-Basurko *et al.* [10]. La nature de la relation entre les entrées et les sorties est apprise pendant un processus supervisé d'apprentissage, directement à partir des données.

4.2.2. Algorithmes d'apprentissage

Pendant la procédure d'apprentissage supervisé, des séries d'entrées et les sorties associées sont présentées de façon répétée au réseau, afin qu'il apprenne à modéliser leur relation, ainsi qu'à pouvoir généraliser précisément, lorsqu'on lui présente des nouvelles données. Cette phase d'apprentissage correspond mathématiquement à l'optimisation d'une fonction coût dans l'espace des poids ; les poids caractérisent l'importance de la connexion entre les neurones des différentes couches, et ils représentent les paramètres à ajuster dans le modèle neuronal.

La fonction de coût est choisie selon les critères de performance du réseau. Si la performance est estimée en termes de précision de la prédiction, ceci correspond à la minimisation de la somme des carrés des erreurs, où l'erreur est définie comme différence entre la sortie désirée et la sortie réelle du réseau [5].

La précision de la prédiction est influencée par l'algorithme d'optimisation. Malheureusement, la surface de l'erreur est souvent complexe et contient plusieurs minima locaux [23], qui peuvent conduire à des modèles sous-optimaux. Les méthodes d'optimisation globale sont plus intéressantes de ce point de vue, mais leur convergence est lente [24] ; de plus, leur implémentation est assez difficile pour les cas complexes [25].

D'une manière générale, ce sont les méthodes d'optimisation locale qui sont préférées, même si le minimum global n'est pas atteint ; un bon minimum local est généralement considéré comme une solution acceptable [13].

4.2.3. Pouvoir de généralisation

Le pouvoir de généralisation est défini comme le potentiel du réseau à avoir un bon comportement avec des données qui n'ont pas été utilisées lors de son apprentissage [26].

Dans un but de prédiction, la propriété la plus importante d'un algorithme est sa capacité à généraliser et à filtrer le bruit. Lorsque le modèle apprend des détails du bruit dans les données d'apprentissage, on dit qu'il apprend par cœur et il a un pouvoir de généralisation assez réduit.

Afin d'éviter que le modèle n'apprenne par cœur, on peut limiter sa complexité en utilisant une technique de régularisation, comme celle de *"early stopping"* (ES). Pour appliquer cette technique, la base de données initiale est divisée en trois sous-ensembles [12]. Le premier ensemble est celui d'apprentissage et il est utilisé pour calculer le gradient et pour mettre à jour les paramètres du réseau. Le deuxième est celui de validation : pendant le processus d'apprentissage, l'erreur est calculée sur l'ensemble de validation ; normalement, cette erreur doit décroître. Lorsque le réseau commence à apprendre par cœur les données, l'erreur sur l'ensemble de validation commence à augmenter et dans ce cas l'apprentissage est arrêté et ce sont les paramètres du réseau obtenus lorsque l'erreur était minimale qui sont retenus. Finalement, la performance du réseau est testée sur un troisième ensemble, qui est appelé ensemble de test.

La régularisation bayésienne [27] est une méthode robuste pour éviter l'apprentissage par coeur du réseau. Dans ce cas, la base initiale de données est divisée en deux parties uniquement : l'ensemble d'apprentissage et l'ensemble de test. Un terme de régularisation est incorporé dans la fonction coût afin de pénaliser les modèles plus complexes [28] ; ce terme peut être la moyenne de la somme des carrés des poids et des biais du réseau. Par cette fonction, les poids et les biais du réseau sont plus petits, forçant la réponse du réseau à être plus lisse.

4.2.4. Prétraitement des données

Les variables ont des plages de variation assez différentes, des unités différentes. Afin de s'assurer qu'elles reçoivent le même poids pendant le processus d'apprentissage, elles doivent être normalisées [29]. De plus, les variables doivent être redimensionnées dans le domaine de variation de la fonction d'activation [20].

Les fonctions d'activation usuelles sont de type sigmoïdal, comme la fonction logistique ou la tangente hyperbolique, qui varient entre 0 et + 1, et – 1 et + 1, respectivement. Ces deux fonctions sont monotones croissantes et possèdent des dérivées simples. Très souvent, pour la couche de sortie, on

utilise la fonction identité, lorsqu'il est nécessaire de pouvoir extrapoler en dehors de la plage des données d'apprentissage [29].

Lorsqu'on utilise la sigmoïde logistique, il est recommandé de normaliser les variables entre 0 et +1. Les données peuvent être normalisées par la formule proposée par Elkamel *et al.* [6] :

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

La tangente hyperbolique (tanh) varie entre – 1 et + 1. Lorsqu'on utilise la tanh, Gardner et Dorling [12] proposent de calculer les variables normalisées par la formule :

$$x_{norm} = 2 \cdot \frac{x - x_{min}}{x_{max} - x_{min}} - 1$$

4.2.5. *Modélisation de NO_2 et CO_2 par des réseaux de neurones*

Dans cette étude, les données d'entrée et de sortie ont été normalisées par rapport à la moyenne et l'écart-type de la variable [24, 29] :

$$x_{norm} = \frac{x - \bar{x}}{\sigma}$$

La valeur maximum de x_{norm} est de 8 pour le NO_2. Chaque variable normalisée a été ensuite redimensionnée afin qu'elle varie dans un intervalle inclus dans [– 1 + 1] ; le redimensionnement a consisté en une division de chaque variable par une valeur S déterminée :

$$S = \alpha^{-1} \cdot \max\left(|x_{norm}|\right)$$

où α a été choisi à 0,8, inférieur à 1 afin d'éviter les valeurs proches de zéro de la dérivée de la fonction d'activation [11].

Il est important de noter que ce sont uniquement des transformations linéaires qui ont été appliquées aux données avant l'identification du modèle RN, et que ces transformations n'influent pas sur les résultats d'une régression non linéaire. Les données ont été transformées ensuite pour revenir aux unités d'origine, en utilisant les formules inverses correspondantes.

Les modèles sélectionnés pour la prédiction de NO_2 et du CO_2 sont des PMC à 3 ou 4 couches (1 ou 2 couches cachées). La première couche (celle d'entrée) contient un neurone pour chaque entrée (21 neurones au total), la dernière (la couche de sortie) ne contient qu'un seul neurone, qui correspond à la sortie (NO_2 ou CO_2), tandis que les couches cachées sont composées d'un nombre variable de neurones. La fonction d'activation tanh a été choisie pour les couches cachées, et la fonction identité (donc linéaire, non bornée) pour la couche de sortie. Plusieurs algorithmes d'apprentissage ont été testés, comme ceux de type gradient ou gradients conjugués, quasi-Newton, Levenberg-Marquardt [11]. Afin d'améliorer le potentiel de généralisation, l'ensemble initial de données, après avoir mis les enregistrements dans un ordre aléatoire et non chronologique, a été divisé en 3 sous-ensembles : apprentissage (60 %), validation (20 %) et test (20 %). L'ensemble de validation a été utilisé pour ES, mais la technique de régularisation bayésienne a été testée aussi.

Les meilleurs résultats obtenus pour le NO_2 présentés sur la figure 3, correspondent à une architecture formée de 21 neurones dans la couche d'entrée, 2 couches cachées avec 40 et 20 neurones respectivement, et une couche de sortie avec un neurone. L'algorithme d'apprentissage est basé sur la méthode d'optimisation de Levenberg-Marquardt. *Early stopping* (ES) a été utilisé afin d'éviter l'apprentissage par cœur. L'EQM et le coefficient de détermination obtenus pour cette simulation ont été de 7,48 mg.Nm^{-3} et R^2 = 0,94 pour l'ensemble d'apprentissage, 10,56 mg.Nm^{-3} et R^2 = 0,88 pour l'ensemble de validation, et de 10,39 mg.Nm^{-3} et R^2 = 0,88 pour celui de test. En termes d'erreur relative comparée à la valeur moyenne de NO_2, l'erreur d'estimation est de 10,43 % pour l'ensemble d'apprentissage, 14,72 % pour celui de validation et de 14,49 % pour celui de test. Ces erreurs sont comparables à l'erreur de mesure (10-12 %).

Pour le CO_2, les résultats obtenus par régression linéaire multiple étaient déjà comparables à l'erreur de mesure (5,6 %). Afin d'améliorer l'estimation du CO_2 estimation, le modèle de RN avec la même architecture que celle qui avait donné les meilleurs résultats pour le NO_2 a été utilisée et les résultats sont présentés dans la figure 3 (b, d, f). L'EQM et le coefficient de détermination ont été de : 0,24 % (erreur relative par rapport à la moyenne de CO_2 de 3 %) et R^2 = 0,95 pour l'ensemble d'apprentissage, 0,31 % (erreur relative de 3,9 %) et R^2 = 0,92 pour l'ensemble de validation, et 0,29 % (erreur relative de 3,6 %) et R^2 = 0,93 pour celui de test. Il en résulte que la différence de 3,6 % à 5,6 % correspond à la partie non-linéaire dans la variation du CO_2.

Le point le plus important à discuter est le potentiel et les avantages de l'utilisation de la méthodologie développée dans cette étude, ladite méthode de corrélation.

La modélisation par réseaux de neurones a conduit : (i) pour le NO_2, à une EQM d'environ 14,5 % par rapport à sa valeur moyenne (71 mg.Nm^{-3}), qui est légèrement supérieure à l'erreur de mesure 10-12 % ; (ii) pour le CO_2 à une EQM d'environ 3,6 % par rapport à sa valeur moyenne (3 %), sensiblement inférieure à l'erreur de mesure.

Malheureusement, le potentiel de généralisation de ces modèles n'a pas pu être testé sur d'autres bases de données. On peut supposer que si le régime de fonctionnement de l'installation ne change pas de façon significative, le modèle devrait conduire à des performances similaires. Il est nécessaire que la campagne de mesure utilisée pour l'identification du modèle soit conçue de façon à ce qu'elle englobe les différents régimes de fonctionnement de l'installation. Dans ces conditions, la modélisation basée sur

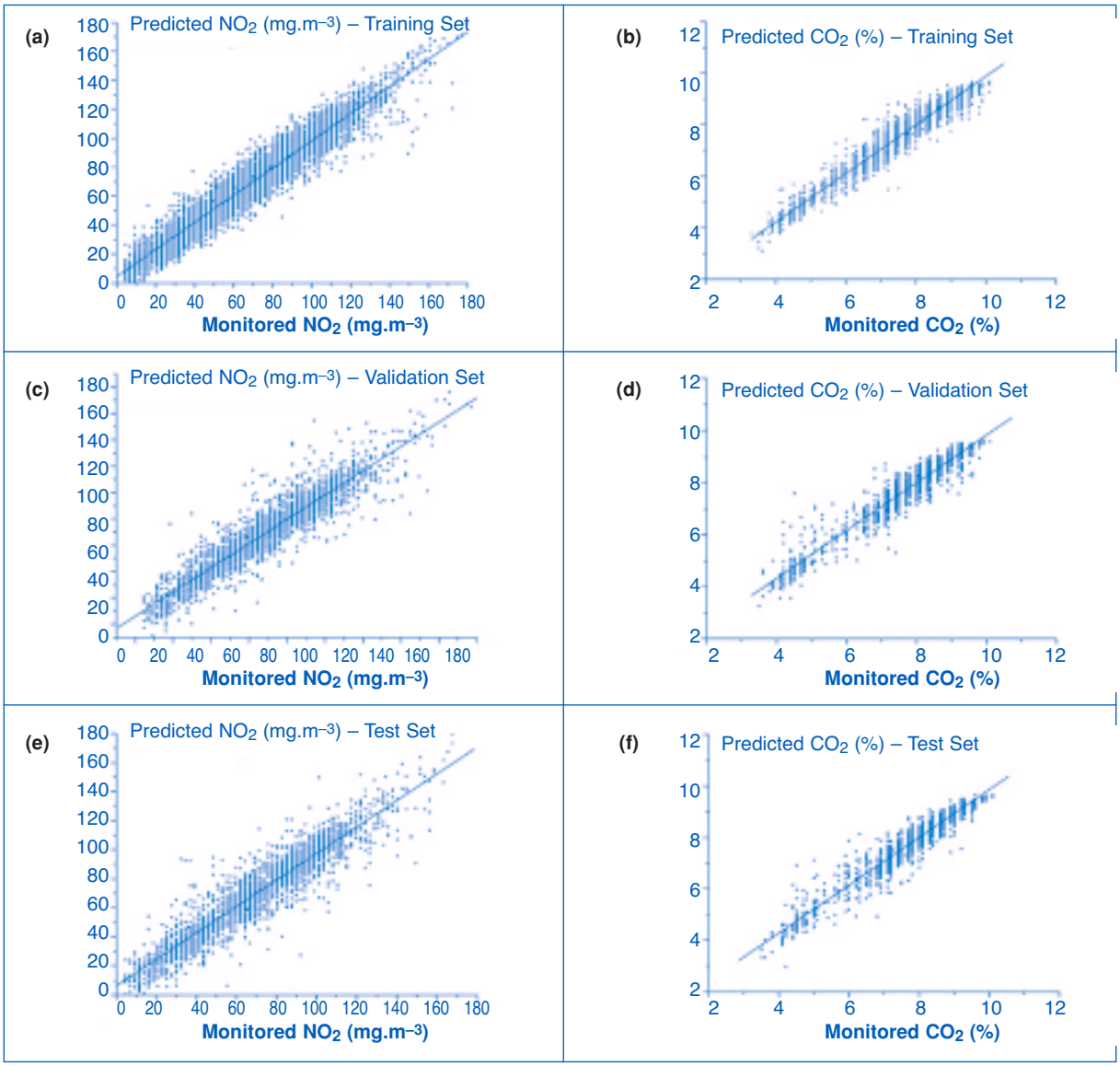

Figure 3.
NO_2 estimé *(predicted NO_2)* par réseaux de neurones et NO_2 mesuré *(monitored NO_2)* pour l'ensemble : (a) d'apprentissage, (c) de validation et (e) de test. CO_2 estimé *(predicted CO_2)* par réseaux de neurones et CO_2 mesuré *(monitored CO_2)* pour l'ensemble : (b) d'apprentissage, (d) de validation et (f) de test [11].
Monitored and predicted NO_2 for: (a) the training, (c) the validation and (e) the test set.
Monitored and predicted CO_2 for: (b) the training, (d) the validation and (f) the test set- results based on a neural network modelling [11].

les réseaux de neurones peut être considérée comme une méthode de corrélation fiable, aussi bien pour le NO_2 que pour le CO_2. De plus, dans le cas du CO_2, un simple modèle linéaire donne des résultats moins performants que les réseaux de neurones (5,6 %), mais ils sont encore comparables à l'erreur de mesure.

5. Conclusion

Par rapport à la méthode de corrélation, les deux autres méthodes acceptées par la législation environnementale française présentent quelques inconvénients : (i) le bilan global de toute l'installation (modélisation de toutes les réactions de combustion) est une méthode complexe, assez difficile à mettre en œuvre ; (ii) la méthode basée sur les facteurs d'émission fournit un résultat assez grossier.

On donne un exemple de comparaison entre les émissions de NO_2 mesurées et celles calculées à partir du facteur d'émission.

En utilisant les débits de gaz et les coefficients de combustion du gaz brûlé dans cette installation (Groningen), on calcule le débit de fumée, et ensuite on détermine la quantité totale de NO_2 en utilisant la concentration mesurée. On obtient environ 350 kg de NO_2 émis pendant les 233 heures de fonctionnement du four, que l'on peut considérer comme une valeur mesurée.

À partir de la cadence des billettes, on peut estimer la masse totale d'acier ; avec un facteur d'émission de 170 g NO_2/tonne d'acier [2], on obtient environ 1 800 kg de NO_2, ce qui représente 5 fois plus que la quantité mesurée.

À titre de comparaison, les émissions calculées avec le modèle réseau de neurones sur la même période, sont à 1 % près des mesures. Cette valeur est due au fait que les surestimations compensent les sous-estimations ; en effet, l'erreur algébrique moyenne entre la prédiction et la mesure est très basse : – 0,12 mg.Nm^{-3} pour l'ensemble d'apprentissage, – 0,37 mg.Nm^{-3} pour celui de validation, et 0,09 mg.Nm^{-3} pour celui de test. Le fait d'avoir une valeur très proche de zéro pour l'erreur moyenne algébrique montre que cette méthode peut être utilisée avec succès pour calculer des valeurs globales d'émissions, pour des périodes plus longues.

On peut remarquer que le modèle développé dans cette étude n'est pas parcimonieux. En effet, toutes les variables process ont été utilisées en entrée, car d'une part, elles sont toutes disponibles en permanence (pour le contrôle du processus), et d'autre part, le but principal a été d'estimer les émissions le plus précisément possible.

Le principal inconvénient réside dans le fait que la taille du réseau augmente artificiellement, et par conséquent le temps de calcul est plus long et, surtout, la quantité de données nécessaires pour estimer les poids des connexions du réseau efficacement devient plus importante.

La sélection des entrées est importante aussi pour trouver les variables les plus influentes sur les émissions et définir ainsi une stratégie de contrôle du processus plus efficace. Dans ce but, une analyse après l'identification du modèle peut être effectuée. En effet, on peut estimer l'importance de chaque variable en analysant les poids des connexions du réseau [7, 30, 31]. La sélection des variables d'entrée avant et après la modélisation est une perspective de cette étude.

Remerciements

Les auteurs tiennent à remercier tous les participants du projet AI/EX, et en particulier, M. Philippe Le Louër, de LECES Environnement, pour la collaboration fructueuse entre le LECES et le CERTES, ainsi que pour le support financier.

References

[1] Fontelle, 2010.

[2] BOMET. Circulaire du 24 décembre 1990 relative aux installations classées pour la protection de l'environnement. Taxe parafiscale sur la pollution atmosphérique (BOMET n° 588-91/14 du 20 mai 1991), http://aida.ineris.fr/cadre_rech.htm (page consultée en mai 2010).

[3] Schofield N, Le Louër P, Mirabile D, Hubner R. Primary Steelmaking. Artificial Intelligence/expert (AI/EX) systems for steelworks pollution control, Technical Steel Research, European Commission EUR 20 501, ISBN KI-NA-20501-EN-S, 2002 : 153 p.

[4] Kolehmainen M, Martikainen H, Ruuskanen J. Neural networks and periodic components used in air quality forecasting. *Atmospheric Environment* 2001 ; 35 : 815-25.

[5] Schlink U, Dorling S, Pelikan E *et al.* A rigorous inter-comparison of ground-level ozone predictions. *Atmospheric Environnement* 2003 ; 37 : 3237-53.

[6] Elkamel A, Abdul-Wahab S, Bouhamra W, Alper E. Measurement and prediction of ozone levels around a heavily industrialized area: a neural network approach. *Advances in Environmental Research* 2001 ; 5 : 47-59.

[7] Abdul-Wahab SA, Al-Alawi SM. Assessment and prediction of tropospheric ozone concentration levels using artificial neural networks. *Environmental Modelling & Software* 2002 ; 17 : 219-28.

[8] Andretta M, Eleuteri A, Fortezza F *et al.* Neural networks for sulphur dioxide ground level concentrations forecasting. *Neural Computing & Applications* 2000 ; 9 : 93-100.

[9] Saporta G. Probabilités, analyse des données et statistique. *Éditions Technip* 1990.

[10] Agirre-Basurko E, Ibarra-Berastegi G, Madariaga I. Regression and multilayer perceptron-based models to forecast hourly O3 and NO2 levels in the Bilbao area. *Environmental Modelling and Software* 2006 ; 21 : 430-46.

[11] Ionescu A, Candau Y. Air pollutant emissions prediction by process modelling - Application in the iron and steel industry in the case of a re-heating furnace. *Environmental Modelling & Software* 2007 ; 22 : 1362-71.

[12] Gardner MW, Dorling SR. Neural network modelling and prediction of hourly NO_x and NO_2 concentrations in urban air in London. *Atmospheric Environment* 1999 ; 33 : 709-19.

[13] Gardner MW, Dorling SR. Statistical surface ozone models: an improved methodology to account for non-linear behaviour. *Atmospheric Environment* 2000 ; 34 : 21-34.

[14] BuHamra S, Smaoui N, Gabr M. The Box-Jenkins analysis and neural networks: prediction and time series modelling. *Applied Mathematical Modelling* 2003 ; 27 (10) : 805-15.

[15] Kalogirou SA. Artificial intelligence for the modelling and control of combustion process: a review. *Progress in Energy and Combustion Science* 2003 ; 29 : 515-66.

[16] Tronci S, Baratti R, Servida A. Monitoring pollutant emissions in a 4.8 MW power plant through neural network. *Neurocomputing* 2002 ; 43 : 3-15.

[17] Ferretti G, Piroddi L. Estimation of NO_x emissions in thermal power plants using artificial neural networks. J Eng Gas Turbines Power Trans ASME 2001 ; 123 (2) : 465-71.

[18] Zhou H, Cen K, Fan J. Modelling and optimization of the NO_x emission characteristics of a tangentially fired boiler with artificial neural networks. *Energy* 2004 ; 29 (1) : 167-83.

[19] Abdi H. Les réseaux de neurones. Presse Universitaire de Grenoble 1994.

[20] Fausett L. Fundamentals of Neural Networks. Architectures, Algorithms and Applications. Prentice Hall, Englewood Cliffs, NJ 07632 1994.

[21] Bishop CM. Neural Networks for Pattern Recognition. Clarendon Press, Oxford 1995.

[22] Ripley BD. Pattern Recognition and Neural Networks, Cambridge University Press 1996.

[23] Comrie AC. Comparing neural networks and regression models for ozone forecasting. *Journal of the Air and Waste Management Association* 1997 ; 47 : 653-63.

[24] Maier HR, Dandy GC. The effect of internal parameters and geometry on the performance of back-propagation neural networks: an empirical study. *Environmental Modelling & Software* 1998 ; 13 : 193-209.

[25] Lu WZ, Fan HY, Lo SM. Application of evolutionary neural network method in predicting pollutant levels in downtown area of Hong Kong. *Neurocomputing* 2003 ; 51 : 387-400.

[26] Cheng B, Titterington DM. Neural networks: A review from a statistical perspective. *Statistical Science* 1994 ; 9 (1) : 2-54.

[27] Buntine WL, Weigend AS. Bayesian back-propagation. *Complex Systems* 1991 ; 5 : 603-43.

[28] Dorling SR, Foxall RJ, Mandic DP, Cawley GC. Maximum likelihood cost functions for neural networks models of air quality data. *Atmospheric Environment* 2003 ; 37 : 3435-43.

[29] Maier HR, Dandy GC. Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. *Environmental Modelling & Software* 2000 ; 15 : 101-24.

[30] Garson GD. Interpreting neural-network connection weights. *AI Expert* 1991 ; 6 (7) : 47-51.

[31] Goh ATC. Back-propagation neural networks for modelling complex systems. Artificial Intelligence in Engineering 1995 ; 9 : 143-51.