How To Address Data Gaps in Life Cycle Inventories: A Case Study on Estimating CO2 Emissions from Coal-Fired Electricity Plants on a Global Scale

ARTICLE *in* ENVIRONMENTAL SCIENCE & TECHNOLOGY · APRIL 2014

Impact Factor: 5.33 · DOI: 10.1021/es500757p · Source: PubMed

Zoran Steinmann
Radboud University Nijmegen

Mara Hauck
TNO

Aafke Schipper
Radboud University Nijmegen

Mark A J Huijbregts
Radboud University Nijmegen

Environmental Science & Technology

Article

pubs.acs.org/est

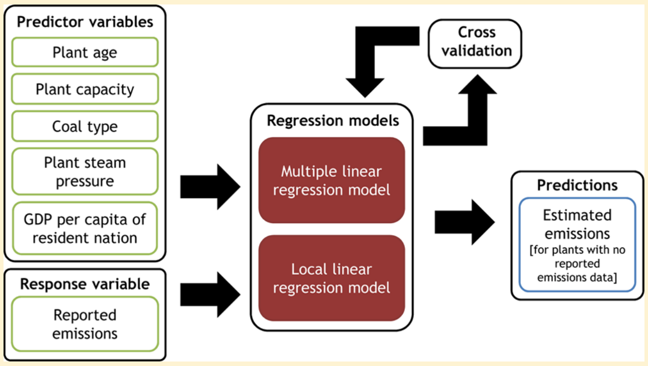
How To Address Data Gaps in Life Cycle Inventories: A Case Study on Estimating CO_2 Emissions from Coal-Fired Electricity Plants on a Global Scale

Zoran J. N. Steinmann,[†] Aranya Venkatesh,*[,‡] Mara Hauck,[†] Aafke M. Schipper,[†] Ramkumar Karuppiah,[‡] Ian J. Laurenzi,[‡] and Mark A. J. Huijbregts[†]

[†]Department of Environmental Science, Radboud University Nijmegen, Toernooiveld 1, 6525 ED Nijmegen, Netherlands

[‡]ExxonMobil Research and Engineering Company, 1545 Route 22 East, Annandale, New Jersey 08801-3059, United States

Supporting Information


Predictor variables
Plant age
Plant capacity
Coal type
Plant steam pressure
GDP per capita of resident nation
Response variable
Reported emissions
Cross validation
Regression models
Multiple linear regression model
Local linear regression model
Predictions
Estimated emissions [for plants with no reported emissions data]

ABSTRACT: One of the major challenges in life cycle assessment (LCA) is the availability and quality of data used to develop models and to make appropriate recommendations. Approximations and assumptions are often made if appropriate data are not readily available. However, these proxies may introduce uncertainty into the results. A regression model framework may be employed to assess missing data in LCAs of products and processes. In this study, we develop such a regression-based framework to estimate CO_2 emission factors associated with coal power plants in the absence of reported data. Our framework hypothesizes that emissions from coal power plants can be explained by plant-specific factors (predictors) that include steam pressure, total capacity, plant age, fuel type, and gross domestic product (GDP) per capita of the resident nations of those plants. Using reported emission data for 444 plants worldwide, plant level CO_2 emission factors were fitted to the selected predictors by a multiple linear regression model and a local linear regression model. The validated models were then applied to 764 coal power plants worldwide, for which no reported data were available. Cumulatively, available reported data and our predictions together account for 74% of the total world's coal-fired power generation capacity.

INTRODUCTION

Life cycle assessment (LCA) is a tool that is often used to quantify the environmental impact of products and services.[1] One of the major challenges in LCA is the availability and quality of life cycle inventory (LCI) data, i.e., the emissions of substances and use of resources during a product's life cycle. LCI databases, such as Ecoinvent,[2] GaBi,[3] and the U.S. Life Cycle Inventory Database,[4] provide information for a large number of processes but are at the same time far from complete. Approximations and assumptions are often made if appropriate data are not readily available. For instance, a common approach used to address data gaps in LCA is to derive estimates from similar processes in other regions of the world, as done in the Ecoinvent database.[2] These proxies introduce uncertainty into LCA results by neglecting regional differences. An alternative approach is to derive inventory data from more readily available information on process characteristics. For example, Wernet et al.[5] used neural networks to estimate the cumulative energy demand (CED) of chemicals based on their molecular properties. Caduff et al.[6,7] used scaling laws to estimate the fuel use of different size generators and the environmental impact of energy generation from wind. Venkatesh et al.[8] and Karras[9] used linear regression models based on characteristics of crude oil to estimate emissions associated with its refining. While these approaches take into account regional or process-specific differences, which are often not well-characterized in LCA, uncertainties in the model predictions need to be quantified.

Regression models can be developed using parametric or non-parametric approaches[10].[11] Non-parametric regression techniques, such as local linear regression, allow for the fitting of local models around a particular query point without assuming a global model structure that is valid for the entire data set, thus allowing for some flexibility in modeling the structure of functional relationships.[11,12] A drawback of the non-parametric models is that all data have to be stored in the model, so that it can be accessed for each new query. Parametric models, such as multiple linear regression models, on the other hand, assume an existing global data structure and use all data points to fit the model. These models have the advantage of being able to determine the most influential

Received: June 7, 2013
Accepted: April 7, 2014
Published: April 21, 2014

parameters easily, and they are easily transferred to other applications (only the model parameters need to be known).

Fossil energy demand constitutes one of largest sources of environmental impacts for many products and processes.[13,14] In particular, coal power plants are among the highest emitters of greenhouse gases (GHGs) per kWh of electricity compared to other sources of electricity.[15] Coal is also the dominant source of electricity generation; i.e., 42% of all electricity generated in the world in 2011 was produced by coal plants.[16] Coal power plant efficiencies and, therefore, their CO_2 emissions vary greatly across the world. For example, the average thermal efficiency of Japanese plants is around 42%, while plants from India have an average efficiency of about 30%.[17] Hence, approximations of power generation models using "typical" values from other contexts (for example, assuming Indian power plants have impacts similar to U.S. power plants) may significantly influence an assessment of environmental impacts over a product's life cycle. Quantifying emission factors of coal-fired electricity as accurately as possible could therefore significantly improve the performance of life cycle models that include electricity generation as a major component. A considerable effort has been made by Dones et al.,[18] who developed coal power plant emission factors for China, the United States, and 19 different European countries/regions based on reported data for coal use and electricity generation. However, coal-fired electricity generation is widespread, and emission data are not reported for many other countries in the world. Moreover, Steinmann et al.[19] demonstrated that there is considerable variability in life cycle GHG emissions between individual power plants, implying that there is added value in plant-based emission models compared to country-based approaches.

We are aware of one study by the Center for Global Development,[20,21] called Carbon Monitoring in Action (CARMA), where combustion-phase CO_2 emissions from thermal power plants across the world were estimated with a global regression model based solely on the design characteristics of the plants. As far as we know, no other efforts have been made to estimate power plant emissions on a global scale. We predict CO_2 combustion emissions per kWh (from here onward referred to as emission factors) from coal power plants on a global scale and quantify uncertainty in the predictions. We employ two predictive regression models: a parametric and a non-parametric model. After fitting and validating both models, we use these to predict CO_2 emission factors for 764 coal-fired power plants worldwide. These predictions can be used to supplement information in life cycle modeling efforts wherever the required data are unavailable.

Similar to the Center for Global Development approach, we include power plant design characteristics as explanatory variables within the regression models. There are, however, some notable differences between the two approaches. The quantification of statistical uncertainty is one of the advantages of our model framework. Quantifying the uncertainty that arises from using a predictive model is important because it can be propagated to the final output of any LCA that uses electricity from one of these coal plants as input. Another difference is that our model can distinguish between lignite and non-lignite plants, whereas the CARMA model cannot. Also, the differences between countries were made explicit in our model by including per capita gross domestic product (GDP) as a predictor. Finally, we explored various types of parametric and non-parametric model approaches and settings to find optimal model predictions.

METHODS

Regression models using reported emission data were developed and validated to predict CO_2 emission factors from coal-fired power plants across the world. In this section, we describe the data used to develop the regression models, model structures, metrics to evaluate the performance of the models, and application of the models.

Emission Data. Reported CO_2 emissions from individual power plants were obtained from databases published by governmental agencies, such as the U.S. Energy Information Administration (EIA), or from reports by private power companies. A more extensive discussion on how the CO_2 emissions per kWh were derived from reported data can be found in SI1 of the Supporting Information.

If emission data were available for multiple years, the most recent year was used. Only plants with a capacity greater than 100 MW were included in the analysis to ensure that the data set was not confounded by very small autonomous producers (for example, a power plant belonging to a paper mill). In total, emissions were reported for 444 plants with a total capacity of 494 GW (Table 1).

Table 1. Characteristics of Individual Coal-Fired Power Plants by Country Used for Model Validation

country	number of plants	sum capacity (MW)	year	source
United States	310	288689	2010	Steinmann et al.[19] based on EIA[38]
India	59	64623	2010−2011	Central Electric Authority[39]
Australia	24	27494	2010	AEMO[40]
South Africa	13	37678	2011	Eskom[41]
China (including Hong Kong)	2	5368	2011	CLP group[42]
Canada	4	4221	2010	OPG[43]
Bulgaria	1	908	2010	Enel[44]
Greece	4	3977	2006	Kavouridis[45] WWF[46]
England and Wales	8	17884	2006	WWF[46]
Germany	10	22324	2006	WWF[46]
Poland	4	10905	2006	WWF[46]
Czech Republic	1	1490	2006	WWF[46]
Italy	1	2640	2006	WWF[46]
Spain	1	140	2006	WWF[46]
Portugal	1	1250	2006	WWF[46]
Scotland	1	2400	2006	WWF[46]
total	444	493586		

Predictor Variables and Data. To estimate power-plant-specific CO_2 emission factors, a number of potentially important predictor variables were identified (Table 2). For our models to be applicable to coal-fired plants throughout the world, we selected predictors that were available from public data sources. Plant age, total capacity, steam pressure, and coal type were all expected to influence CO_2 emissions and were derived from the World Electric Power Plant (WEPP) database.[22] The plant age was calculated by taking the capacity-weighted average of the age of all active generators

Table 2. Range of Predictor Variables

variable	range used in model fitting	total range in WEPP database and IMF	notes
plant age	0−60 years	0−70 years	in the case of multiple generators, a weighted average age is used; the ages of the generators were calculated by subtracting the operation year from the year for which the data were reported; the weights were based on the generator capacity as a proportion of the total plant capacity
coal type	lignite or non-lignite[a]	lignite or non-lignite	plants that did not report their fuel type or that report a combination of fuel types, such as lignite/bituminous, were excluded from the analysis
steam pressure	35−293 bar	17−286 bar	the average steam pressure of all operational generators per plant
total capacity	100−4440 MW	100−5500 MW	total capacity of all operational generators per plant; only plants of >100 MW were included
GDP per capita	$3694−48387	$487−48387	GDP per capita in $ of PPP

[a] Non-lignite plants are modeled as 0, while lignite plants were assigned a value of 1.

at a plant. The GDP per capita of the resident nations of the power plants was also chosen as a predictor because we hypothesized that it correlates with the GHG emission policies and/or power plant maintenance in those nations, which, in turn, affects the efficiency (and therefore the CO_2 emissions) of the plants. The 2011 GDP per capita [in purchasing power parity (PPP)] for each country was obtained from the International Monetary Fund's World Economic Outlook.[23] Figure S1 of the Supporting Information shows how the emission factors and predictor variables relate to each other.

Model Fitting. Two different modeling approaches were employed, i.e., a multiple linear regression model and a local linear regression. Because the models derived in this study were used to predict CO_2 emission factors for a large number of unknown plants, a thorough assessment of the predictive power of the model is required. Therefore, prior to model fitting, the data set (444 plants) was split into a training set and a test set. The training set consisted of 311 of 444 plants (approximately 70%), while the remaining 133 plants in the test set were used for validation of the models.

Multiple Linear Regression. A multiple linear regression model can be used to predict the value of a response variable based on a linear relation to any number of predictor variables. The CO_2 emission factor associated with a power plant is inversely related to its generation efficiency. We hypothesize that the selected predictors are linearly related to the efficiency of the power plant and, therefore, inversely related to its CO_2 emission factor (eq 1)

$$\hat{z} = \frac{1}{\hat{y}} = \beta_0 + \beta X \tag{1}$$

where $\hat{y}$ is a vector representing an estimate of the emission factor for all plants, X is the matrix where each column corresponds to individual predictor variables (Table 2), β is a vector of coefficients, and β_0 represents the intercept of the model.

Ordinary least-squares (OLS) fitting was used to calculate the model coefficients. The need for log transformation (with 10 as a base) of the predictor variables, total capacity, plant age, steam pressure, and GDP per capita, was assessed, given the skewed distribution of the predictor variables (see SI2 of the Supporting Information). Prior to the log transformation, 1 year was added to the plant age because some plants that were less than 1 year old were assigned a plant age of 0. We used the package MuMIn in the statistical program R[24] to generate all possible models using any combination of the predictor variables (both log-transformed and non-transformed). To find the optimum between model complexity and the accuracy of the predictions, we calculated Akaike's Information Criterion (AIC) for all 162 possible combinations. AIC gives a bonus for the goodness of fit [the log likelihood function, $\ln(L)$] and a penalty for the number of predictors k (eq 2). The model with the lowest AIC value was considered to be the best model.

$$\text{AIC} = 2k - \ln(L) \tag{2}$$

The best model was subsequently checked for multicollinearity in its predictor variables by calculating variance inflation factors (VIFs). Typically, predictors with a VIF > 10 indicate multicollinearity in the inputs and need to be excluded from the model.[25] No predictor variables were excluded from our model based on this threshold because the highest observed VIF was below 2 (see Table S1 of the Supporting Information). Furthermore, 95% prediction intervals (as a function of the standard error in the model fit and the deviation of each predictor from its mean value) were calculated by the R function predict.lm.

Furthermore, leave-one-out cross-validation (LOOCV) was performed, which is a procedure in which a model is fitted with all power plants but one. The fitted model is then used to predict the emission factor of the power plant that was left out, and this procedure is repeated until a prediction is obtained for every single power plant. These estimates were used to assess the predictive power of the model.

To determine the relative importance of each of the predictor variables for the best model (that is, the model with the lowest AIC), we performed a separate analysis based on standardized predictor variables. Because all of the input variables are measured in different units (e.g., age in years, capacity in MW, etc.), they were first standardized to z scores. Regression coefficients resulting from the standardized fit directly reflect the relative importance of each predictor variable.

Local Linear Regression Model. In locally weighted polynomial regression, a low-degree polynomial is fitted locally around a query point using a weighted least-squares approach, where observations near the query point are assigned higher weights.[26] In this study, we used an implementation of a local first-order (linear) regression by Kalnins et al.,[27] developed by Jekabsons,[28] detailed as follows.

Assume that y^p is a vector that represents emission factors and X^p represents a matrix of d columns corresponding to the individual predictor variables; these represent the training set. The training set includes n observations (X_i^p, y_i^p); $i = 1, ..., n$. The predicted emission factor of a particular (query) plant j, with characteristics described by X_j^q, is estimated by performing a linear regression locally in the neighborhood of query plant j.

The linear model to be used locally can be represented as shown in eq 3.

$$F(X_i) = a_0 + \sum_{d}^{m=1} a_m X_{i_m} \tag{3}$$

The coefficients a in eq 3 are calculated using a weighted least-squares method, as shown in eq 4. In this method, the neighbors nearest to query point X_j^q are assigned higher weights.

$$a = \arg\min \sum_{n}^{i=1} w(X_j^q, X_i^p)(F(X_i^p) - y_i^p)^2 \tag{4}$$

The weighting function w used in eq 4 is a function of the distance between observations X^p and query point X_j^q. While a number of different weighting functions, such as the Epanechnikov quadratic weighting, tri-cube, and Gaussian weighting can be used, it has been shown that the choice of the weighting function does not significantly affect the results.[29] We used the Gaussian weighting function, implemented by Kalnins et al.[27], as follows. In this implementation, μ_i refers to the scaled distance from the query point j to the ith plant in the training set, as shown in eq 5.

$$\mu_i^j = \frac{\| X_j^q - X_i^p \|}{\| X_j^q - X_{\text{farthest}}^p \|} \tag{5}$$

The Gaussian weight function is then calculated as a function of the scaled distance μ_i and coefficient α, as shown in eq 6. This coefficient controls the linear approximation, by defining the extent of the neighborhood around a query point that is weighted strongly in the approximation.

$$w(X_j^q, X_i^p) = \exp(-\alpha \mu_i^j) \tag{6}$$

Before being used to estimate weights, all predictors in the training set (without log transformation) were first transformed to z scores. Furthermore, Gower's dissimilarity metric was used to transform the binary coal-type variable (by multiplying the z scores of the binary coal-type variable by a factor of $1/\sqrt{2}$, as suggested by Sigovini[30]). The coefficient α was tuned using LOOCV, suggested as a widely accepted method to measure the goodness of fit of the local model.[31] In this approach, for a specific value of α, a prediction was made for a single power plant based on a fitted model that included all other power plants in the training set. This was performed for each power plant in the training data set, resulting in emission factor estimates for every single plant that was used to evaluate the model mean square error for that specific value of α. The optimal value of the coefficient α was determined using a simple stepwise search algorithm presented by Kalnins et al.,[27] which resulted in the lowest LOOCV mean square error.

The optimal coefficient α, thus determined, was then used in conjunction with the training set data to estimate the emission factors of all plants in the test set.

Finally, bootstrap sampling was used to develop pointwise prediction interval estimates for all power plants, as shown by Aneiros-Pérez et al.[32] The training data residuals, calculated using the local linear regression model, were used to estimate 1000 bootstrapped residuals for each data point in the test set. The 95% prediction intervals for new plants were based on these 1000 bootstrapped residuals.

Model Evaluation. We analyzed the R^2 for the cross-validated training set (311 plants) and the R^2 for the 133 plants in the test set to obtain an indication of the predictive power of the models.[33] These R^2 metrics have values that are between $-\infty$ and 1. A value of 1 reflects a perfect prediction, while any model with a R^2 value below 0 has no added value compared to using the average of the data set.[34,35]

In addition to the R^2 metrics for the training and test sets, we calculated relative prediction errors for each power plant, as the difference between the reported and predicted emission factors, represented as a fraction of the reported emission factor. We also plotted relative prediction errors for all plants against the individual predictors to identify predictor ranges for which the estimates are particularly uncertain. To check the influence of the relatively large number of U.S. plants in the training data set (223 of 311) on the global multiple linear regression model, an additional multiple regression fit was performed where 135 of the U.S. plants were removed from the training data set to obtain a data set with an equal number of U.S. and non-U.S. plants (88 each).

Model Application. The multiple linear regression model with the lowest AIC value and the local linear regression model were applied to coal-fired power plants in the WEPP database, for which all predictors listed in Table 2 were available and for which no reported emission data were available. In total, the WEPP database includes 1974 coal plants with a capacity of >100 MW (cumulative capacity of 1687 GW) in 72 countries, with at least one operational generator in the year 2012. Emissions were reported for 444 of these plants (494 GW), contributing to 29% of the total capacity of the world (see the Emission Data section). Of the remaining 1540 plants, all predictor data were available for 764 plants with a total capacity of 760 GW (45% of the total capacity of the world).

RESULTS

Multiple Linear Regression Model. The AIC values of all possible multiple regression models (see Table S2 of the Supporting Information) showed that the model with the log-transformed values of the four continuous predictors and the categorical fuel type (lignite or non-lignite) was the best model. The coefficients for this model are shown in Table 3. The normalized model coefficients, indicating influence of each predictor, are shown in Table S3 of the Supporting Information. The log-transformed steam pressure has the highest normalized coefficient value, indicating that it is the most important predictor. The coefficients of the model that used an equal number of U.S. plants and non-U.S. plants are well within 50% of coefficients of the model fitted to the training data, for all predictors except for plant age, for which

Table 3. Coefficients of the Best Multiple Linear Regression Model

predictor name	coefficient	p value
intercept	-3.65×10^{-1}	9.7×10^{-5}
log[capacity (in MW)]	6.38×10^{-2}	9.8×10^{-6}
log[age + 1 (in years)]	-8.69×10^{-2}	2.7×10^{-4}
log[steam pressure (in bar)]	3.46×10^{-1}	8.8×10^{-14}
log[GDP (in \$PPP) per capita]	1.20×10^{-1}	3.7×10^{-17}
fuel type (1, lignite; 0, non-lignite)	-1.49×10^{-1}	3.5×10^{-17}

the coefficient almost doubles because of the removal of the U.S. plants (see Table S4 of the Supporting Information). The predictions from this model are consistent with the predictions from the model based on the entire training data set; on average, these differ by 1.5%.

The cross-validated R^2 value based on the training set fit is 0.53, while the R^2 based on the 30% test set is 0.49. Reported and predicted emission factors (with 95% prediction intervals) for the 133 power plants in the test set are displayed in Figure 1A. On average (over all 133 plants), the lower bound of the 95% prediction interval (2.5th percentile) is 14% lower than the mean value, while the upper bound (97.5th percentile) is 19% higher than the mean value.

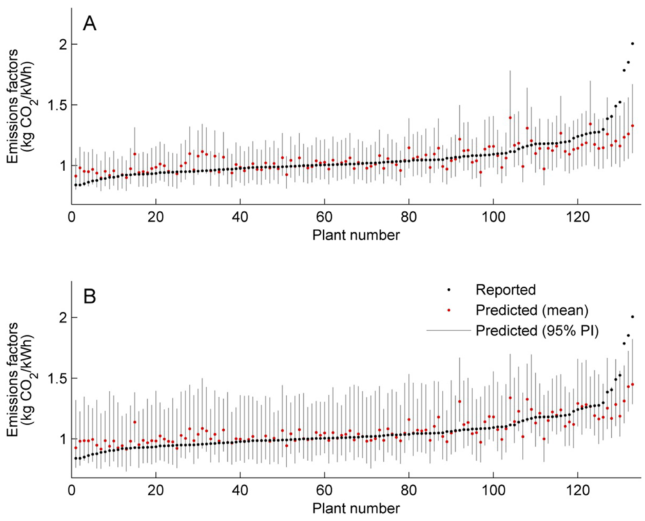


Figure 1. Reported and predicted CO_2 emission factors for the 133 plants in the test set with corresponding 95% prediction intervals (*y* axis) by an individual power plant (*x* axis) for the (A) multiple regression and (B) local linear regression models. Power plants are sorted from low to high by reported CO_2 emission factor.

The multiple regression model does not perform well for plants with emission factors exceeding 1.5 kg of CO_2/kWh (Figure 1A). Results of the relative errors plotted against the predictors (see SI3 of the Supporting Information and Figure S2) show that relative errors lower than −0.2 (i.e., plants for which the model underestimates emission factors by more than 20%) are mainly observed for plants that are older than 30 years, have a capacity <1000 MW, and/or a steam pressure below 125 bar.

Local Linear Regression Model. The optimal value of α, selected by minimizing the mean square error through LOOCV of the training set, was found to be 9.3. The R^2 value of the local linear model based on the training data set is 0.55, while this value based on the test set is 0.61. Reported and predicted emission factors (with 95% prediction intervals) for all 133 power plants in the test set are displayed in Figure 1B. On average (over all 133 plants), the lower bound of the 95% prediction interval (2.5th percentile) is 17% lower than the mean value, while the upper bound (97.5th percentile) is 31% higher than the mean value. The model performs slightly better than the global multiple regression model, as also observed from Figure 1.

Model Application. Both models were applied to predict the CO_2 emission factors of 764 coal-fired power plants worldwide. As an example, Figure 2 shows the predicted CO_2 emission factors for countries with >30 power plants, for which (almost) no reported emission data were available, Russia and China. For the Russian plants, the multiple regression model predicts an unweighted (i.e., with equal weight to every plant, regardless of the capacity of the plant) mean emission factor of 1.14 kg of CO_2/kWh (individual plant emission factors ranging between 0.95 and 1.48 kg of CO_2/kWh), while the local linear regression method predicts an unweighted mean emission factor of 1.25 kg of CO_2/kWh (0.97−1.57 kg of CO_2/kWh). The estimated emission factors of Chinese coal-fired power plants are typically lower than Russian plants, i.e., 1.03 kg of CO_2/kWh (0.90−1.49 kg of CO_2/kWh) for the multiple regression model and 1.04 kg of CO_2/kWh (0.88−1.46 kg of CO_2/kWh) for the local linear regression.

Estimates and the corresponding 95% prediction intervals for each of the 764 plants can be found in Table S5 of the Supporting Information. Unweighted mean CO_2 emission factors over 764 plants were estimated to be 1.08 kg of CO_2/kWh by the multiple regression model and 1.12 kg of CO_2/kWh by the local linear regression model. The range of mean predictions per power plant obtained by the multiple regression

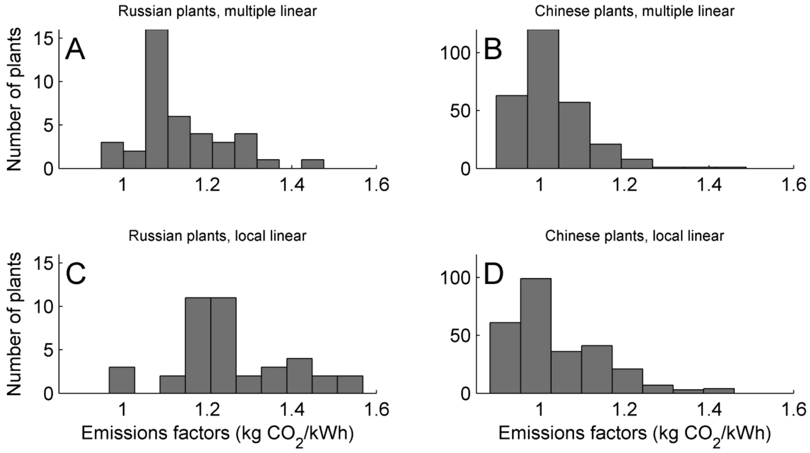


Figure 2. Histograms of predicted CO_2 emission factors (kg of CO_2/kWh) of coal plants in Russia and China for the (A and B) multiple regression model and (C and D) local linear regression model. The predicted emission factors (mean and 95% prediction intervals) for the corresponding individual plants are presented in Figure S4 of the Supporting Information.

model (0.84−2.34 kg of CO_2/kWh) is larger than the range for the local linear model (0.87−1.89 kg of CO_2/kWh). It should be noted, however, that the maximum value predicted by the multiple regression model refers to a plant with extreme characteristics (old plants with small capacity and low steam pressure) located in a country with low GDP per capita (Zimbabwe). The characteristics of this plant are outside the range of predictors used to train the model. When the 95% prediction intervals are considered, we observe a smaller range for the multiple regression model (0.90−1.42 kg of CO_2/kWh) compared to the range (0.90−1.57 kg of CO_2/kWh) found for the local linear regression.

■ DISCUSSION

Model Performance. We employed two regression techniques (multiple linear regression and local linear regression) to estimate the CO_2 emission factors of coal-fired power plants on a global scale. Predictions obtained by the local linear regression model show a slightly better performance, as demonstrated by the higher R^2 of the test set (0.61 for local linear regression versus 0.49 for the multiple linear regression). We also tested a number of alternate non-parametric regression approaches, including *k*-nearest neighbors and kernel regression (see Table S6 of the Supporting Information). However, the local linear regression was found to perform better than these methods based on the training and test set R^2 values.

Both regression models have a relative error of less than 20% for more than 95% of the power plants based on the test set with 30% of the plants. A direct comparison of these values to the findings by Ummel in the CARMA model[21] is impossible because the results were not provided separately for coal-fired plants; CO_2 emissions from all thermal power plants (including other generation types, such as natural gas plants) were estimated in that study. As noted, our multiple regression model, in particular, underestimates the highest reported emission factors (>1.5 kg of CO_2/kWh). A similar effect can be observed in the CARMA model.[21] The highest predictions (for any fuel type) from that study were approximately 1.5 kg of CO_2/kWh, while the highest reported emission factors were close to 2.0 kg of CO_2/kWh.

A possible limitation of our models is that they do not address temporal variability in the emission factors. Because of data limitations, it was not possible to obtain reported emissions and generation data for multiple years for all power plants in the data set. We did, however, assess the temporal variability in the emission factors of a subset of U.S. coal plants for which data were available (see SI4 of the Supporting Information). Even though the total annual CO_2 emissions and the net generation were found to fluctuate over the years (as presented in the CARMA report[21]), our results show that emission factors appear to be relatively stable. For the 306 plants with an emission factor of <1.5 kg of CO_2/kWh, the difference between extreme values observed over a 3 year time period was, on average, 3.1% of the mean value. Additionally, the highest temporal variability was found for plants with high CO_2 emission factors. The four plants with an average emission factor of >1.5 kg of CO_2/kWh show an average difference of 33% between the extreme values. This finding indicates that the CO_2 emission factors of the plants with relatively high CO_2 emission factors are inherently variable over time. This finding may also help explain why our models make less accurate predictions for plants with high emission factors.

Analysis of the relative errors in the training set and the test set (see SI3 of the Supporting Information) shows that the model fit is not as good for plants that are over 30 years old, have a capacity below 1000 MW, or especially, have steam pressures below 125 bar. Therefore, we suggest caution when applying our model to plants with characteristics in this range, especially if a plant has all three characteristics in the critical range.

Number of Predictors. A limited number of readily available predictors was included in our models, because the aim was to develop models to predict emission factors for a large number of power plants. On the basis of AIC, we found that the best model included all predictors, indicating that none of these predictors were redundant. However, several other predictors may influence plant efficiencies and, therefore, CO_2 emission factors, such as the cooling processes of the plant, the presence of SO_2 and NO_x control equipment, and plant capacity factor. Furthermore, the grid stability of a country may influence the performance of a power plant as well. If the grid lacks stability, plants have to stop and start relatively often, which lowers the average plant efficiency. We were, however, not able to add the regional grid stability as an extra predictor because of the lack of data.

Lam and Shiu[36] found that the capacity factor strongly influences power plant efficiency. As part of a sensitivity analysis, we included the plant capacity factor as an additional predictor (see SI5 of the Supporting Information). The predictive power of both models increased as a result of including the capacity factor as an additional predictor. However, both models were not able to make accurate predictions for plants with high emission factors, even when including the capacity factor as a predictor variable. It was also found that estimates for plants with capacity factors higher than 50%, in general, have lower relative errors (as shown in SI5 of the Supporting Information). It should be noted, however, that for the external application of our models, the capacity factor has no added value because net electricity generation is typically not known for plants that do not report CO_2 emissions.

Model Application. The applicability domain of a model refers to the range of predictor values in which a model can be applied. The observed ranges of predictors in the application set were similar to or within the range of the model development set. From this, we conclude that the plants in the application set were within the applicability domain of our models. One exception is the per capita GDP, which ranges between \$3700 and \$48 000 per capita in the model training set. A total of 11 plants in the application set are situated in countries with GDP below \$3700 per capita. Extrapolation outside the original range causes the predictions for these plants to be more uncertain, such as for the prediction for the coal plant in Zimbabwe, which should be interpreted with caution.

The uncertainties that are introduced by predictive modeling are around 8 times larger than the uncertainty that was found for measured emissions in the United States by Steinmann et al.[19] Their study showed that it is possible to reduce uncertainty in the emissions from individual power plants to a much smaller range with full disclosure of power plant fuel use and electricity generation data. In the absence of more data, however, a modeling approach for calculating the CO_2 emission factors of individual power plants could help address data gaps in life cycle inventories. With small adaptations to our model framework, it is possible to estimate power plant efficiencies directly. The estimated average efficiency of plants in a country could then be applied in a database, such as Ecoinvent, where the process of coal delivery to a plant is already coupled to the operational efficiency. To derive a country estimate from the individual power plants in that country, one needs to combine the estimates (and the uncertainty in the estimates) for every power plant. Uncertainty ranges may vary from plant to plant depending upon the number of predictors that are available for each plant. Monte Carlo simulations can be used to sample from the uncertainty range of each individual power plant. These samples can then be combined to generate an overall country estimate (with its own uncertainty interval).

The regression models presented here predict the combustion-phase CO_2 emission factors for coal-fired power plants. Although coal combustion typically contributes to 90% or more of the life cycle emissions from coal-fired electricity generation (see e.g. refs 19 and 37), upstream emission factors need to be assessed as well. The upstream emissions cannot be modeled easily, however, because they depend upon factors such as the type of coal mine, heat content of the fuel, transport type, and transport distance (as demonstrated by Steinmann et al.[19]). Except for the type of fuel, these data are not readily available for most power plants. Part of this information (country-specific transport distances and coal sources) is, however, already available in life cycle inventory databases. Coupling of our model estimates with this type of data can provide a better estimate (for data-scarce countries) of the total life cycle emissions of electricity generation.

Despite limitations in the regression models developed, we found that they provide an improvement compared to assigning averaged CO_2 emission factors to each power plant. Although we applied this methodology to a case study of coal-fired power plants, we expect that a similar approach can be used to estimate emission factors from biomass-, gas-, and oil-fired power plants.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Additional explanation for some methods used, intermediate and supplementary results, and references. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: aranya.venkatesh@exxonmobil.com.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was funded by ExxonMobil Research and Engineering, Corporate Strategic Research. The authors thank Niranjan Subrahmanya from ExxonMobil for providing expert advice that helped supplement this research. The authors also thank the two anonymous reviewers who helped to improve the quality of the paper.

■ REFERENCES

(1) International Organization for Standardization (ISO). *ISO 14040:2006. Environmental Management—Life Cycle Assessment—Principles and Framework*; ISO: Geneva, Switzerland, 2006.

(2) Frischknecht, R.; Jungbluth, N.; Althaus, H.-J.; Doka, G.; Dones, R.; Heck, T.; Hellweg, S.; Hischier, R.; Nemecek, T.; Rebitzer, G.; Spielmann, M.; Wernet, G. *Overview and Methodology*, Ecoinvent Report 1; Swiss Centre for Life Cycle Inventories: Duebendorf, Switzerland, 2007.

(3) PE International. *GaBi LCA Databases*; http://www.gabi-software.com/international/databases/gabi-databases/.

(4) National Renewable Energy Laboratory (NREL). *U.S. Life Cycle Inventory Database*; http://www.nrel.gov/lci/.

(5) Wernet, G.; Hellweg, S.; Fischer, U.; Papadokonstantakis, S.; Hungerbühler, K. Molecular-structure-based models of chemical inventories using neural networks. *Environ. Sci. Technol.* **2008**, *42* (17), 6717−6722.

(6) Caduff, M.; Huijbregts, M. A. J.; Althaus, H. J.; Hendriks, A. J. Power-law relationships for estimating mass, fuel consumption and costs of energy conversion equipments. *Environ. Sci. Technol.* **2010**, *45* (2), 751−754.

(7) Caduff, M.; Huijbregts, M. A. J.; Althaus, H. J.; Koehler, A.; Hellweg, S. Wind power electricity: The bigger the turbine, the greener the electricity? *Environ. Sci. Technol.* **2012**, *46* (9), 4725.

(8) Venkatesh, A.; Jaramillo, P.; Griffin, W. M.; Matthews, H. S. Uncertainty analysis of life cycle greenhouse gas emissions from petroleum-based fuels and impacts on low carbon fuel policies. *Environ. Sci. Technol.* **2011**, *45* (1), 125−131.

(9) Karras, G. Combustion emissions from refining lower quality oil: What is the global warming potential? *Environ. Sci. Technol.* **2010**, *44* (24), 9584.

(10) Everitt, B.; Dunn, G. *Applied Multivariate Data Analysis*; Wiley: Hoboken, NJ, 2001.

(11) Schaal, S.; Atkeson, C. G. Robot juggling: Implementation of memory-based learning. *Control Syst., IEEE* **1994**, *14* (1), 57−71.

(12) Hastie, T. J.; Tibshirani, R. J.; Friedman, J. H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: New York, 2009.

(13) Huijbregts, M. A. J.; Rombouts, L. J. A.; Hellweg, S.; Frischknecht, R.; Hendriks, A. J.; van de Meent, D.; Ragas, A. M. J.; Reijnders, L.; Struijs, J. Is cumulative fossil energy demand a useful indicator for the environmental performance of products? *Environ. Sci. Technol.* **2006**, *40* (3), 641−648.

(14) Huijbregts, M. A. J.; Hellweg, S.; Frischknecht, R.; Hendriks, H. W. M.; Hungerbühler, K.; Hendriks, A. J. Cumulative energy demand as predictor for the environmental burden of commodity production. *Environ. Sci. Technol.* **2010**, *44* (6), 2189−2196.

(15) Frischknecht, R.; Tuchschmid, M.; Faist Emmeneger, M.; Bauer, C.; Dones, R. *Strommix und Stromnetz*, Ecoinvent Report 6, Version 2.0; Paul Scherrer Institut Villingen, Swiss Centre for Life Cycle Inventories: Duebendorf, Switzerland, 2007; p 143.

(16) *Worldcoal Factsheet "Coal in the Global Energy Supply"*; http://www.worldcoal.org/bin/pdf/original_pdf_file/coal_matters_1_-_coal_in_the_global_energy_supply(16_05_2012).pdf (accessed June 30, 2013).

(17) Graus, W. H. J.; Voogt, M.; Worrell, E. International comparison of energy efficiency of fossil power generation. *Energy Policy* **2007**, *35* (7), 3936−3951.

(18) Dones, R.; Bauer, C.; Roeder, A. *Kohle*, Final Report; Paul Scherrer Institute Villingen, Swiss Centre for Life Cycle Inventories: Duebendorf, Switzerland, 2007; p 346.

(19) Steinmann, Z.; Hauck, M.; Karuppiah, R.; Laurenzi, I.; Huijbregts, M. A methodology for separating uncertainty and variability in the life cycle greenhouse gas emissions of coal fueled power generation in the United States. *Int. J. Life Cycle Assess.* **2014**, DOI: 10.1007/s11367-014-0717-2.

(20) Wheeler, D.; Ummel, K. *Calculating CARMA: Global Estimation of CO_2 Emissions from the Power Sector*; Center for Global Development: Washington, D.C., May 2008; Working Paper 145.

(21) Ummel, K. *CARMA Revisited: An Updated Database of Carbon Dioxide Emissions from Powerplants Worldwide*; Center for Global Development: Washington, D.C., 2012; Working Paper 304.

(22) Platts. *World Electric Power Plants (WEPP) Database*; http://www.platts.com/Products/worldelectricpowerplantsdatabase (accessed May 2012).

(23) International Monetary Fund (IMF). *World Economic Outlook*; http://www.imf.org/external/pubs/ft/weo/2012/01/weodata/download.aspx.

(24) R Foundation for Statistical Computing. *R Core Team R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2012.

(25) Field, A. *Discovering Statistics Using SPSS*; Sage Publications Limited: Thousand Oaks, CA, 2009.

(26) Cleveland, W. S. Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.* **1979**, *74* (368), 829−836.

(27) Kalnins, K.; Ozolins, O.; Jekabsons, G. Metamodels in design of GFRP composite stiffened deck structure. *Proceedings of 7th ASMO-UK/ISSMO International Conference on Engineering Design Optimization*; Association for Structural and Multidisciplinary Optimization in the UK (ASMO-UK): Bath, U.K., 2008; p 11.

(28) Jekabsons, G. *Locally Weighted Polynomials for Matlab*, 2010; http://www.cs.rtu.lv/jekabsons.

(29) Chu, C. K.; Marron, J. S. Choosing a kernel regression estimator. *Stat. Sci.* **1991**, *6* (4), 404−419.

(30) Sigovini, M. Multiscale dynamics of zoobenthic communities and relationships with environmental factors in the Lagoon of Venice. Doctoral dissertation, Università Ca'Foscari Venezia, Dorsoduro, Italy, 2011; http://dspace.unive.it/handle/10579/1092.

(31) Schaal, S.; Atkeson, C. G. Assessing the quality of learned local models. *Adv. Neural Inf. Process. Syst.* **1994**, 1−8.

(32) Aneiros-Pérez, G.; Cao, R.; Vilar-Fernández, J. M. Functional methods for time series prediction: A nonparametric approach. *J. Forecasting* **2010**, *30* (4), 377−392.

(33) Schüürmann, G.; Ebert, R.-U.; Chen, J.; Wang, B.; Kühne, R. External validation and prediction employing the predictive squared correlation coefficient—Test set activity mean vs training set activity mean. *J. Chem. Inf. Model.* **2008**, *48* (11), 2140−2145.

(34) Legates, D. R.; McCabe, G. J., Jr. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water Resour. Res.* **1999**, *35* (1), 233−241.

(35) Todeschini, R. *Tutorial 5, Useful and Unuseful Summaries of Regression Models*; http://www.moleculardescriptors.eu/tutorials/tutorials.htm.

(36) Lam, P. L.; Shiu, A. A data envelopment analysis of the efficiency of China's thermal power generation. *Util. Policy* **2001**, *10* (2), 75−83.

(37) Littlefield, J.; Bhander, R.; Bennett, B.; Davis, T.; Draucker, L.; Eckard, R.; Ellis, W.; Kauffman, J.; Malone, A.; Munson, R.; Nippert, M.; Ramezan, M.; Bromiley, R. *Life Cycle Analysis: Existing Pulverized Coal (EXPC) Power Plant*; National Energy Technology Laboratory (NETL): Pittsburgh, PA, 2010; 110809, p 112.

(38) U.S. Energy Information Administration (EIA). *Form EIA-923 Detailed Data*; http://www.eia.gov/electricity/data/eia923/.

(39) Central Electric Authority. *Baseline Carbon Dioxide Emissions from Power Sector*; http://www.cea.nic.in/reports/planning/cdm_co2/cdm_co2.htm (accessed June 2012).

(40) Australian Energy Market Operator (AEMO). *National Transmission Network Development Plan, Supply Input Spreadsheets*; http://www.aemo.com.au/Consultations/National-Electricity-Market/Closed/~/media/Files/Other/planning/0410-0029%20zip.ashx (accessed June 2012).

(41) Eskom. *CDM Calculations*; http://www.eskom.co.za/c/article/236/cdm-calculations/ (accessed June 2012).

(42) CLP Group. *Facility Performance Statistics for Fangchenggang and Castle Peak*; https://www.clpgroup.com/ourvalues/report/Pages/sustainabilityreport.aspx (accessed June 2012).

(43) Ontario Power Generation. *Sustainable Development Report 2010*; Ontario Power Generation: Toronto, Ontario, Canada, 2011.

(44) Enel. *Sustainability Report 2010*; http://www.enel.com/en-GB/doc/report2010/Sustainability_report_2010_30_06_2011.pdf (accessed June 2012).

(45) Kavouridis, K. Lignite industry in Greece within a world context: Mining, energy supply and environment. *Energy Policy* **2008**, *36* (4), 1257−1272.

(46) World Wide Fund for Nature (WWF). *Dirty Thirty, Ranking of the Most Polluting Power Stations in Europe*; http://wwf.panda.org/?100140/Europes-Dirty-30 (accessed July 2012).