// NO.22-006 | 02/2022

DISCUSSION PAPER

// SEBASTIAN SCHMIDT, JAN KINNE, SVEN LAUTERBACH,
THOMAS BLASCHKE, DAVID LENZ, AND BERND RESCH

Greenwashing in the US Metal Industry? A Novel Approach Combining SO_2 Concentrations From Satellite Data, a Plant-Level Firm Database and Web Text Mining.

Leibniz Association

ZEW

Greenwashing in the US metal industry? A novel approach combining SO_2 concentrations from satellite data, a plant-level firm database and web text mining.

Sebastian Schmidt[1,2,*], Jan Kinne[2,3], Sven Lautenbach[4,5], Thomas Blaschke[1], David Lenz[2,6], Bernd Resch[1,7]

Abstract

This Discussion Paper deals with the issue of greenwashing, i.e. the false portrayal of companies as environmentally friendly. The analysis focuses on the US metal industry, which is a major emission source of sulfur dioxide (SO_2), one of the most harmful air pollutants. One way to monitor the distribution of atmospheric SO_2 concentrations is through satellite data from the Sentinel-5P programme, which represents a major advance due to its unprecedented spatial resolution. In this paper, Sentinel-5P remote sensing data was combined with a plant-level firm database to investigate the relationship between the US metal industry and SO_2 concentrations using a spatial regression analysis. Additionally, this study considered web text data, classifying companies based on their websites in order to depict their self-portrayal on the topic of sustainability. In doing so, we investigated the topic of greenwashing, i.e. whether or not a positive self-portrayal regarding sustainability is related to lower local SO_2 concentrations. Our results indicated a general, positive correlation between the number of employees in the metal industry and local SO_2 concentrations. The web-based analysis showed that only 8% of companies in the metal industry could be classified as engaged in sustainability based on their websites. The regression analyses indicated that these self-reported "sustainable" companies had a weaker effect on local SO_2 concentrations compared to their "non-sustainable" counterparts, which we interpreted as an indication of the absence of general greenwashing in the US metal industry. However, the large share of firms without a website and lack of specificity of the text classification model were limitations to our methodology.

Keywords: Sentinel-5P, air pollution, natural language processing, spatial regression
JEL Classification: Q53, Q56, R11

[1]Department of Geoinformatics — Z_GIS, University of Salzburg, 5020 Salzburg, Austria,
[2]ISTARI.AI, 68163 Mannheim, Germany,
[3]Department of Economics of Innovation and Industrial Dynamics, Centre for European Economic Research, 68161 Mannheim, Germany,
[4]Heidelberg Institute for Geoinformation Technology at Heidelberg University, 69118 Heidelberg, Germany,
[5]GIScience department, Heidelberg University, 69120 Heidelberg, Germany,
[6]Department of Statistics and Econometrics, Justus-Liebig-University, 35394 Giessen, Germany
[7]Center for Geographic Analysis, Harvard University, 9VGM+R8 Cambridge, USA,
[*]Corresponding author. E-mail: sebastian.schmidt@plus.ac.at

1 Introduction

Air pollution has been described as one of the most crucial global health risks (Chatkin et al., 2021). Mainly as a consequence of growing industries and populations, the concentrations of many air pollutants in the troposphere have risen over the past decades (Oxoli et al., 2020). Among the primary pollutants that are emitted directly into the atmosphere, sulfur oxides (SO_x) are considered some of the most harmful. One of these oxygen compounds is sulfur dioxide (SO_2), which is emitted by both natural and man-made sources, with emissions from transport and industry being especially prominent (Goudarzi et al., 2016; Theys et al., 2017). One of the main contributors to industrial SO_2 emissions is the metal industry, which was therefore the focus of this study (Garg et al., 2001).

Due to the high levels of industrial pollution, the topic of sustainability has become increasingly more important in recent years (Lee, 2017). The Brundtland Commission famously defined sustainable development as "development that meets the needs of the present generation without compromising the ability of future generations to meet their own needs" (Brundtland, 1987). Social pressure and the need for a "clean" image can lead companies to present themselves as particularly sustainable. At the same time, however, the same company may shy away from the costs associated with transforming its business model and thus settle for "empty words". This is commonly referred to as *greenwashing* (Delmas & Burbano, 2011). Arguably, the most important medium for the external presentation of a company to a broad audience today is the company's own website. In this article, we used the webAI web analytics tool developed by ISTARI.AI to evaluate the websites of 9,430 companies in the United States (US) metal industry with Natural Language Processing (NLP) in terms of how they present themselves with respect to sustainability. By contrasting these web-based self-representations and SO_2 measurements from satellite sensors, this study aimed to investigate whether greenwashing is prevalent in the US metal industry. Accordingly, this paper aspired to answer the following research questions:

- RQ 1: What conclusions can be drawn about sustainability in the US metal industry based on web text mining?
- RQ 2: Does the self-representation of metal industry companies regarding sustainability coincide with findings based on SO_2 remote sensing data?

In order to answer these research questions, non-spatial and spatial regression models were used, which were based on variables representing different natural and man-made influences on SO_2 concentrations. The most appropriate model in terms of model type, variable combination and spatial weights matrix was selected. In addition, metal industry companies were classified individually as *sustainable* or *non-sustainable* based on their websites (RQ 1). The results were then entered into the regression model in the form of weighted and aggregated variables to answer RQ 2. The effect of the sustainable metal industry on SO_2 concentrations according to the regression model was ultimately used to confirm or deny greenwashing. We are not aware of any study that utilised a similar data fusion of remote sensing data, a firm database and information from company websites. This paper therefore represents a first approach to measuring greenwashing at large scale.

2 Background

2.1 US metal industry

The US iron and steel sector provides jobs for approximately two million people and generates an annual output of over US$ 520 billion (AISI, 2020). Although production peaked in 1973 and has since been on the decline, the US is still the fourth largest producer of steel in the world (USGS, 2020; Hasanbeigi & Springer, 2019; Fenton, 2005). Current political upheavals surrounding restrictions on imports of "dirty steel" from China to the US in favour of domestic production show that the industry remains highly relevant (Financial Times, 2021).

While the metal industry is undeniably an important catalyst for economic development, it has a rather bad reputation for being "dirty and unsustainable" (Lee, 2017). Apart from its high energy consumption (Worrell et al., 2010), the mining and processing of metals has led to long-lasting consequences for many industrial districts. While some impacts are directly visible (e.g. changes in topography), many forms of persistent pollution are not, e.g. highly contaminated top soils or aquatic ecosystems (Johnson et al., 2016). Particularly, the impact of the metal industry on the atmosphere is often the focus of attention. With a global share of 24%, the iron and steel industry is a major industrial contributor of carbon dioxide (CO_2) emissions (Hasanbeigi & Springer, 2019). Other pollutants, such as SO_2, nitrogen oxides (NO_x), carbon monoxide (CO) and metal oxides, are produced in heating, smelting and sintering processes (Cirtina et al., 2016; Ma et al., 2012).

The term *metal industry* is used as a kind of hypernym in this paper, since we refer to all companies involved in the extraction and basic processing of metals. This corresponds with the first part of the anthropogenic metal cycle (Chen et al., 2016). We rely on the popular North American Industry Classification System (NAICS) to delineate the metal industry. The definition we apply is based on the *SIC Major Group: 33 - Primary Metal Industries*, using the respective NAICS codes (NAICS Association, 2018). Additionally, companies from metal ore mining, forging and stamping were included.

2.2 Air pollution from SO_2

Sulfur dioxide (SO_2) is a highly toxic, gaseous air pollutant and atmospheric trace gas (Metya et al., 2020; Zheng et al., 2018). Due to its high impact on morbidity and mortality rates, it is described as a *criteria air pollutant* by the United States Environmental Protection Agency (EPA), i.e. a particularly harmful substance for the environment and human health (He et al., 2016; EPA, 2021). It has a seasonally varying but relatively short residence time (12-78 hours) in the lower troposphere, which can be further decreased by precipitation due to the solubility of SO_2 (Hidy & Blanchard, 2016; Lee et al., 2011). Consequently, SO_2 concentrations tend to be particularly high in the emission region (Lewinschal et al., 2019).

SO_2 is quite reactive in the atmosphere, resulting e.g. in the formation of sulfate (SO_4^{2-}) (Hidy & Blanchard, 2016; Lee et al., 2011), which is the main component of particulate matter with a diameter $\leq$ 2.5 µm ($PM_{2.5}$) (Zhang et al., 2007). $PM_{2.5}$ has a considerable effect on global and local climates (Smith et al., 2011), in addition to being the main cause of smog (Cheng et al., 2017). SO_2 has also

been identified as the main trigger for the emergence of acidic deposition, more commonly called *acid rain*. Since the 1970s, it has been one of the most discussed ecological damages, particularly affecting the eastern US, Europe and China. Acidic deposition has manifold effects on soils, vegetation, and bodies of water (Menz & Seip, 2004). With regard to human health, air pollution is the reason for more than four million premature deaths worldwide each year (Lewinschal et al., 2019). Concerning SO_2, crucial health consequences include the worsening and even emergence of asthma, bronchitis, and lung carcinoma (Chatkin et al., 2021).

The emission sources of SO_2 can be both natural and anthropogenic and include volcanoes, industrial processes, transportation (cars, ships, airplanes) and energy generation (Kumar et al., 2018; Goudarzi et al., 2016). In fact, most SO_2 emissions come from the combustion of fossil fuels (coal, crude oil, natural gas) and smelting activities (Theys et al., 2017). The global emission peak of anthropogenic SO_2 was in the 1970s. Ever since, particularly coal-based emissions have declined in Europe and North America (Smith et al., 2011). In the US, the U.S. Clean Air Act of 1970 and its 1990 amendments have been named among the most important steps in the emission reduction process (Mitchell & Likens, 2011). The increased use of low-sulfur coal, desulfurization, emission abatement technologies, and the reduction of sulfur in metal ores used for smelting have contributed to decreasing the SO_2 emissions in the US (Smith et al., 2011; Mallik et al., 2019). A political measure was the introduction of a SO_2 Allowance Trading System (Schmalensee & Stavins, 2013). Nevertheless, significant industrial emitters still exist in the US, such as the steel industry (Zhong et al., 2020).

2.3 Remote sensing for air pollution monitoring

While the traditional ground-based measuring networks are quite extensive in most developed countries and deliver consistent data, they have several shortcomings, e.g. concerning the coverage of rural regions (Oxoli et al., 2020; Cromar et al., 2019). Therefore, satellite data has been used increasingly to monitor the concentration of certain gases in the stratosphere and troposphere since the launch of the Total Ozone Mapping Spectrometer (TOMS) in 1978. In particular, the launch of the Ozone Monitoring Instrument (OMI) in 2004 and the Tropospheric Monitoring Instrument (TROPOMI) in 2017 have been named as landmarks in this regard (Kaplan & Avdan, 2020). With their improved spatial resolution, they overcame a central limitation for the use of satellite data in air pollution monitoring (Cromar et al., 2019).

As part of the Sentinel-5 Precursor programme, TROPOMI is a push-broom imaging spectrometer that is able to record eight wavelength ranges from ultraviolet (UV) to short wavelength infrared (SWIR) with a swath width of 2,600 km on a low-Earth, early afternoon orbit. TROPOMI has an unprecedented spatial resolution of initially 7×3.5 km and, since August 2019, 5.5×3.5 km. Global measurements are possible for CO, formaldehyde (CH_2O), methane (CH_4), nitrogen dioxide (NO_2), ozone (O_3), SO_2, aerosol distribution and cloud coverage (Romahn et al., 2020; Hedelt et al., 2019; Veefkind et al., 2012).

While ground sensors actually measure the concentration of air pollutants directly, satellite sensors approximate the concentration based on spectral signatures, using numerical models (Oxoli et al., 2020). The information on SO_2 measured by TROPOMI is the so-called *vertical column density*, which describes the total of SO_2 molecules in an air column over a unit area (Fioletov et al., 2017). It can be calculated based on the strong absorption of solar radiation by SO_2 in the near UV spectral range (Wang et al., 2020; Theys et al., 2019). For this detection, a technique called Differential Optical Absorption Spectroscopy (DOAS) is applied, leading to three separate heights of vertical column density (1 km, 7 km, 15 km) (Romahn et al., 2020; Hedelt et al., 2019).

3 State of the Art

Due to the longer duration of its mission, more studies draw data from OMI than from TROPOMI. These mainly investigate the spatial and temporal distribution of different pollutants and their sources. For instance, a study in China uses TROPOMI data to show seasonal and agglomeration effects in the distribution of NO_2 (Zheng et al., 2019). Remote sensing data is also deemed a suitable substitute for ground-based measurements for studies on natural SO_2 emissions, as shown by research on volcanoes (Queißer et al., 2019). A greater focus in research, however, is on anthropogenic SO_2 emissions. At that, many studies deal with the analysis of major emitters, such as power plants (e.g. Song and Yang, 2014). A study in Mongolia finds that the daily and seasonal trends of SO_2 concentrations are not only influenced by man-made emissions, but also by meteorological factors (e.g. boundary layer height) (Zheng et al., 2018). In recent months, the global COVID-19 pandemic has given rise to new research topics. Several studies on air pollution are able to show a decrease in pollutants that can be attributed to reduced human activity (e.g. Hashim et al., 2021).

Several machine learning methods have been used in air pollution monitoring and prediction in the past few years, including deep neural networks (Karimian et al., 2019), extreme gradient boosting models (Xu et al., 2018), kriging (Li et al., 2019) and random forests (Brokamp et al., 2017). Some studies use non-spatial regression models for air pollution analyses, such as ordinary least squares (OLS) models (Zhao et al., 2019), quantile regression models (Xu & Lin, 2020) or land-use regression models (Han et al., 2020). However, since air pollution is an inherently spatial issue, many studies employ spatial regression models, such as a spatial lag model (SLM) (Ren & Matsumoto, 2020), a spatial error model (Zhou et al., 2018) or a spatial Durbin model (Chen et al., 2017). For instance, Yang et al., 2017a find that temperature and precipitation can have different regional effects on SO_2 concentrations by employing a SLM. Geographically weighted regression models have also been employed in numerous studies, e.g. by Zhao et al., 2020.

No web data was used in previous studies on air pollution, even though web-based methods generally open up new possibilities for research in economic geography. Nowadays, almost every company has its own website, where it discloses information about e.g. its goods, markets and customers (Gök et al., 2015). Many companies also use their website to build up their public image regarding environmental protection and sustainability. Since web-based methods make it possible to collect large amounts of data in a timely manner, web scraping is also seen as a cost-effective alternative for classic business surveys (Kinne & Axenbeck, 2020). Even though web data is generally freely available, it is of large-scale and in an unstructured form, which poses serious methodological challenges. However, relevant

advances in machine learning applications for this purpose have already been made, e.g. regarding deep neural networks (Kinne & Lenz, 2021) and general progress in NLP (Li, 2017). So-called *web text mining* has already been used to generate economic indicators in studies regarding firm-level innovation activities (Kinne & Lenz, 2021), R&D and collaboration (Beaudry et al., 2016), the impact of the COVID-19 pandemic (Kinne et al., 2020) and 3D printing diffusion (Schwierzy et al., 2022).

4 Methodology

4.1 Study area

The study area was the contiguous US, thus, encompassing 48 states (including Washington D.C.) and approximately 82.2% of the total area of the US (USCB, 2018). Due to its large extent, there is a high spatial variation concerning land cover, climate and population density. The research area was mainly chosen due to the size of the US metal industry, which is an important prerequisite for the feasibility of the analysis.

4.2 Data

TROPOMI data

The TROPOMI data - as well as all other remote sensing data and derived products - were obtained via the Google Earth Engine (GEE) (Gorelick et al., 2017). We used the georeferenced, orthorectified and pre-processed (Level 3) SO_2 Offline (OFFL) data product (Sentinel-5P, 2021; Romahn et al., 2020). Compared to the also available Near Real-Time (NRTI) product, it had the advantage that more elaborate air mass factors were used in the conversion process (Verhoelst et al., 2021). The vertical column density at ground level (1 km) was chosen for this analysis as emissions from metal industry plants were expected to be concentrated closer to the surface, whereas e.g. volcanic emissions reach higher layers of the troposphere (Hedelt et al., 2019). Data from this band was selected on the basis of a rectangular polygon. A large time period from 01.01.2019 to 31.12.2020 was chosen to capture continuous emission centres. For this purpose, the year 2019 was the first fully available year on the GEE. The resulting composite consisted of 10,318 images, for which the mean values were calculated, as it was already done by Kaplan et al., 2019. The spatial resolution of TROPOMI data on the GEE is 0.01 arc degrees. As it is not possible to access the data with the native pixel size and aspect ratio, downsampling to a spatial resolution of 7×7 km was employed. This resolution was used as the reference grid for the entire analysis.

Business data

The Infogroup US Historical Business Data contains annual, plant-level, geo-coded information of companies and organisations in the US, such as industry description, annual revenues, number of employees and year of foundation (Infogroup, 2018). For this analysis, we used the 2018 data set from which the relevant locations of the metal industry were filtered based on the NAICS codes (cf. Table 5). For multi-site companies, all locations marked as headquarters were excluded as they are usually office buildings and therefore do not contribute significantly to SO_2 emissions. A thorough,

positive manual validation was performed to check the validity of this filtering, using Bing Maps imagery and Google Street View. Besides the coordinates, the number of employees per location was another central information needed for the analysis. According to the data, 475,517 people were employed in 9,430 locations of the metal industry in 2018. If there was no information on the number of employees (< 1% of all locations), the missing value was set to the average value in the respective NAICS category. In order to operationalise a measure for the local presence of the metal industry, we created a weighted count of plant locations per raster cell based on their respective number of employees.

In addition to the metal industry, we also extracted companies from the manufacturing sector (all NAICS codes starting with 31-33, excluding the metal industry) and applied the same pre-processing procedure described above. The manufacturing sector could also be responsible for SO_2 emissions and was therefore used as a control variable in the subsequent regression analyses.

Additional data

In order to control for power plant SO_2 emissions, we used data released by the EPA as part of their CSAPR programme (EPA, 2019b). The data set includes the SO_2 emissions in tons for 1,535 power plants in the US for the year 2019 and was used to calculate the SO_2 emissions per cell (EPA, 2019a).

Data on 2017 vehicle emissions in the US was accessed from the Database of Road Transportation Emissions (DARTE), which releases a grid with annual CO_2 emissions in tons/km^2 with a spatial resolution of 1 km (Gately et al., 2019). While this data does not contain any direct information on SO_2 emissions, it can be seen as an approximation for traffic density and, therefore, also for vehicle-based SO_2 emissions.

Data on annual precipitation was acquired from the gridMET data set on the GEE for the year 2019. As it did not contain any information on precipitation over oceans, the Great Lakes and the Florida Keys, the SAGA (Conrad et al., 2015) interpolation tool *Close gaps* was used to fill these missing values. Data on average temperature was acquired based on the *temperature_2m* variable in the ERA5 data set on the GEE, which has a spatial resolution of 0.1 arc degrees (Muñoz Sabater, 2019). The year 2019 was chosen, as the data for 2020 was incomplete.

Additionally, data from the USGS National Land Cover Database (NLCD) from 2016 was used (Yang et al., 2018). We reduced the number of classes to seven by combining similar land use classes using QGIS (QGIS Development Team, 2021) (cf. Table 6). The distance of each grid cell to the closest major body of water was considered another sensible control variable for the analysis, as bodies of water play a central role in human activities and influence climate and local wind systems decisively. Data on commercially navigable waterways (NPMS, 2019), the major lakes and the two coastlines (USGS, 2011) were downloaded and converted to linestrings. The distance from each grid cell centroid to the next linestring was then calculated using *geopandas* (Jordahl et al., 2020). For observations that were assigned the land cover class *water*, the distance was set to 0. Furthermore, data on elevation was accessed from USGS, 2012 and data on population density in the year 2020 was acquired from WorldPop, 2020.

We followed Kaplan and Avdan, 2020 and downsampled all data to TROPOMI's spatial resolution, as it constitutes the central research object and the most coarse data set. For this, the *gdal.Warp*

function in Python was used, specifying a bilinear resampling algorithm and pixel alignment. Only for the land cover data a nearest neighbour algorithm was chosen. The data was then combined based on the coordinates of the grid cell centroids.

A land mask was created as a further pre-processing step for the regression analysis. This was necessary, as there was a lot of missing data for various variables for areas outside of the contiguous US. To achieve the filtering, all the cells without land cover information were dropped from the data set, reducing the number of cells from 440,840 to 216,058. Consequently, maritime areas as well as the Canadian and Mexican territories were excluded. However, inland waters within the contiguous US and a small coastline were preserved for the regression analysis. None of the remaining cells had any missing values.

4.3 Regression analysis

4.3.1 Variables

Table 1: Descriptive statistics of regression variables. The units of some variables were adjusted for better readability of the regression coefficients.

Variable	Description	Source	Unit	mean	min	median	max
SO2_19_20	SO_2 emissions	Sentinel-5P, 2021	mol/km^2	0.17	-0.02	0.16	0.62
cover	land cover classification	Yang et al., 2018	-	-	-	-	-
elev	elevation	USGS, 2012	m	762.78	-81.08	457.21	3,797.81
manuf_log	employees in manufacturing industry (excluding metal industry)	Infogroup, 2018	log(count)	0.65	0.00	0.00	11.19
metal_log	employees in metal industry	Infogroup, 2018	log(count)	0.08	0.00	0.00	8.52
pop_log	population density	WorldPop, 2020	log(count)	0.40	0.00	0.43	9.85
power_log	SO_2 power plant emissions	EPA, 2019a	log(t)	0.01	0.00	0.00	7.19
prec	precipitation	Abatzoglou, 2012	m/a	0.94	0.07	0.93	4.14
temp	average annual temperature	Muñoz Sabater, 2019	°C	11.41	-2.31	10.86	25.83
veh_log	CO_2 vehicle emissions	Gately et al., 2019	log(t/km^2)	9.30	0.00	9.59	17.61
water	distance to body of water	NPMS, 2019; USGS, 2011	km	64.69	0.00	46.59	359.98

Table 1 provides an overview and basic descriptive statistics for all variables used in the regression analysis. Economic variables frequently tend to be skewed due to the underlying complexity of anthropogenic phenomena (Xu & Lin, 2020). The particularly right skewed variables were, therefore, log-transformed.

Multicollinearity, as a common problem in regression analysis (Dormann et al., 2013), was checked by means of the Spearman correlation coefficient which is relatively robust to outliers and does not require a normal distribution of the data (de Winter & Gosling, 2016). For bivariate correlations > 0.7, one of the variables was excluded from the regression analysis. Due to the high correlation of 0.78

between vehicle emissions and population density, the latter was dropped, as vehicle emissions were considered to be a better estimation for SO_2 emissions. Furthermore, they might be able to portray emission sources in regions with a low population density better (e.g. industrial zones, where metal industry firms are often located). In addition, we used the Variance Inflation Factor (VIF) (Dormann et al., 2013) as a measure of muliticollinearity for the selected models.

4.3.2 Regression models

Different sets of predictors were tested in the regression analysis that always considered the emission sources (*metal_log*, *manuf_log*, *veh_log*, *power_log*) along with additional variables. The dependent variable for all specifications was the averaged SO_2 concentration for 2019 and 2020 from TROPOMI. First, only the total number of employees in the metal industry (*metal_log*) was used and then compared to a model that used the number of employees for the different sustainability categories from the web text mining analysis (see below).

Since atmospheric phenomena are innately spatial (Zhou et al., 2018), we expected spatial structures in the data. To address the problem of spatial autocorrelation, we compared different spatial regression models. We also calculated an OLS model as a non-spatial baseline model (equation 1):

$$y = \alpha\iota_n + X\beta + \epsilon \tag{1}$$

where y is the dependent variable, "ι_n is a $n \times 1$ vector of ones associated with the constant term parameter α, X denotes an $n \times K$ matrix of explanatory variables associated with the $K \times 1$ parameter vector β" and ϵ is an error term (Halleck Vega & Elhorst, 2015).

In order to take spatial effects into account in a regression model, topological relations have to be considered, which are represented by a spatial weights matrix W. Due to the large sample size and the resulting high computational demand, we refrained from using a distance based weights matrix. Instead, we used a contiguity matrix based on the queen neighbourhood definition (Chen & Ye, 2015), which seemed well suited for the grid-based data structure. The resulting spatial weights matrix did not have any so-called *islands*, i.e. all cells had at least one neighbour. All calculations were performed using the Python module *libpysal* (Rey & Anselin, 2007). A variance-stabilising transformation of the weights was performed for the spatial weights matrix. This transformation is seen as a sensible procedure to handle frequent problems such as heteroscedasticity (Bagnall et al., 2006; Tiefelsdorf et al., 1999).

In this study, the presence of global spatial autocorrelation was verified using the Moran's I measure, one of the most common specification tests for spatial autocorrelation (Anselin, 2001). The employed SO_2 concentrations from TROPOMI had a Moran's I of 0.901 (p=0.001), indicating a strong clustering. Moran's I was also calculated for the residuals of all regression models.

We considered two classes of spatial regression models: the spatial lag of X (SLX) model (equation 2), which uses spatially lagged predictors, and the SLM (equation 3), which includes a spatial lag on the dependent variable. The SLX model considers a $n \times n$ spatial weights matrix W which is used to generate lagged versions of all or a subset of the predictors of the design matrix X. Regression coefficients are fitted for both lagged (θ) and non-lagged (β) predictors (Halleck Vega & Elhorst, 2015).

In the SLM, W_y represents the spatially lagged dependent variable and ρ a spatial autoregressive coefficient (Anselin, 2001). As values of explanatory variables within the SLM are assumed dependent on their adjacent values, the resulting spatial lag term can be used as an additional predictor W (Wu et al., 2020). For the SLM, the two stage least squares estimation from the *spreg* library was used (Rey & Anselin, 2007).

$$y = \alpha\iota_N + X\beta + WX\theta + \epsilon \tag{2}$$

$$y = \rho W_y + X\beta + \epsilon \tag{3}$$

4.4 Web text mining

In order to differentiate metal industry companies with regard to their attitude towards sustainability, a complementary, sophisticated web-based analysis was carried out. In this context, the entirety of a firm's web presence is referred to as a *website* that may contain several *web pages*. The start page is called *main page*, while any other web page is referred to as *subpage* (Kinne & Axenbeck, 2020).

The Infogroup Business data set does not contain information about the web presence of each company. Therefore, a two-part, automated URL search was performed. In the first step, the company's name and address were sent as a search query to a web search engine. In the second step, the three top hits were validated by scraping the indexed websites and checking for both the searched company's name and address. Websites were classified as valid if the name and/or address of the searched company appeared on several sub-pages of the website. A manual validation of the resulting URLs was carried out for a random sample of 50 companies. Our search algorithm found the correct website for 44 firms (88%), while only two results were actually incorrect. For the remaining websites, it was unclear to the reviewer whether the result was valid.

In the next step, the company websites were scraped, which means that their respective HTML content and metadata were downloaded. The maximum number of subpages accessed per website was set to 125. The download priority of these subpages was ordered according to the length of their respective URL, as it is expected that the most important information can be found on the top-level webpages (e.g. main page), which usually have the shortest URLs (Kinne & Axenbeck, 2020). In addition, webpages in English were preferred.

In the downloaded content, text paragraphs were identified in which keywords on the topic of "sustainability" occurred (cf. Table 7). Since the model was developed for cross-sectoral applications, some of the featured keywords were not necessarily relevant for the metal industry, such as *bio* or *organic*. These paragraphs were then analysed and classified using ISTARI's webAI, which is based on an ensemble of several transformer style language models and transfer learning. The base models used in transfer learning are pre-trained on vast amounts of textual data and, thus, have a basic understanding of the structure of natural languages (Malte & Ratadiya, 2019). This means that a relatively small amount of training data is sufficient to achieve good classification results. In the training phase of the model, HTML paragraphs that contained at least one of the keywords were randomly extracted and then labelled manually (n=3,247). Five different categories were assigned (ordinal ranking: *frontrunner*, *enabler*, *engaged*, *information*, *not_engaged*; cf. Table 8), and the resulting label-text pairs were used for fine-tuning the models.

The final classification results of the models in the ensemble were then combined on the basis of a majority decision rule. For example, if six out of ten models classified a paragraph as *engaged*, this class was chosen (Dörr et al., 2021). The resulting paragraph-level classifications were then aggregated to the company-level. Thus, a website may contain several paragraphs with different classification outcomes, e.g. 15 paragraphs that were assigned to the category *engaged* and 5 paragraphs classified as *not_engaged*. For such mixed cases, the aforementioned ordinal ranking of categories was used to classify the entire website (i.e. the company). The highest ranked category was selected for a website, if it corresponded to at least 10% of its classified paragraphs. If this was not the case, the next highest category with more than 10% was chosen. For example, a website that had 5 paragraphs classified as *frontrunner*, 10 paragraphs as *engaged* and 25 paragraphs as *not_engaged* was, thus, categorised as *frontrunner*.

Websites with no keyword hits were classified as *not_engaged*, while companies for which no URL was found were given their own category *no_website*. Additionally, companies whose URL was not unique and was also assigned to more than five other companies were classified as *no_website*. The reason for this is that these were usually misidentified URLs, which e.g. point to online databases like *yellowpagesdirectory.com*. A second, aggregated classification was also carried out by combining the classes *frontrunner*, *enabler* and *engaged* to form the class *sustainable*, while *information* and *not_engaged* were aggregated to create the class *non_sustainable*. For the validation of the web-based classification, a random sampling (n=100) was carried out and manually assessed independently of the model predictions.

In order to take the results of the web-based classification into account in the regression, we split the population of metal industry companies into groups corresponding to their classification. Analogous to the original variable *metal_log*, the number of employees was again taken as a measure of the size of the metal industry and the variables were logarithmised due to their high skewness. The created web-based variables then replaced the variable *metal_log* in the regression model.

5 Results

5.1 Web text mining

The number of paragraphs that were assigned to the different sustainability categories by the text analysis model (cf. Table 2) was skewed in favour of the categories *not_engaged* (29.2%) and *engaged* (61.2%). On average, each website had 13.1 identified paragraphs, with the median being only 3.0. Accordingly, there was a wide range of identified paragraphs between 0 and 897 per website.

The classification of the entire website - based on the number of paragraphs per category - showed an unbalanced distribution across the five categories (cf. Table 2). Therefore, the categories were aggregated into two classes: 51.3% of the companies were classified as *non_sustainable* and 8.1% as *sustainable*. The remaining 40.6% did not have their own website, so that no statement could be made about how sustainable they described themselves.

The analysis of the random sample of websites for the validation of the web-based classification showed that approximately 87% of the companies were classified correctly (cf. Table 3). For three websites,

Table 2: Results of web text mining based classification of the company websites.

narrow classification category	paragraphs	number of companies		broad classification category
frontrunner	546	111		
enabler	429	48	760	sustainable
engaged	7,492	601		
information	1,492	62		
not_engaged	15,681	4,759	4,821	non_sustainable
no_website	25,640	3,822		no_website

manual classification could not be performed because the websites were not accessible at the time of validation. According to the f1-score of 0.86, the model performed well to very well overall. In particular, the precision (0.88) and recall (0.96) of the *non_sustainable* class were very good. However, the model showed weaknesses in the recall of the *sustainable* class (0.57), meaning that a high proportion of sustainable companies were not detected during classification (false negatives). At the same time, the precision in this class was good (0.81), i.e. if the model classified a company as sustainable, it was usually correct (true positives). For the purpose of this study, the model can thus be described as well suited, as the methodology was particularly dependent on the reliable identification of sustainable companies and hence a high degree of precision. Misclassification occurred e.g. when the keyword *environment* referred to a working atmosphere or with some keywords like *durable*, *waste* or *circular*, which can stand for sustainability in other contexts.

Table 3: Confusion matrix

	precision	recall	f1-score	support
non_sustainable	0.88	0.96	0.92	74
sustainable	0.81	0.57	0.67	23
accuracy			0.87	97

The relative densities of sustainable and non-sustainable metal industry locations showed spatial clusters (cf. Figure 1): Centres with high shares of non-sustainable metal industry locations were found e.g. in Detroit, Cleveland, New York, Seattle, Tulsa, Minneapolis and Los Angeles. On the other hand, Dayton, Toledo, Buffalo, Pittsburgh, Salt Lake City, Sacramento, Knoxville, Oklahoma City and particularly the northern outskirts of Chicago were identified as hubs of the sustainable metal industry. In the Western US, there were significantly fewer areas with high sustainable metal industry shares. In some cities there even seemed to be an internal division of the metal industry (e.g. Chicago, Houston, Dallas, St. Louis), meaning that the sustainable metal industry appeared to be concentrated in one part of the city, while environmentally uncommitted companies clustered in other districts. This can be illustrated by the example of Chicago, where sustainable businesses were mainly found in the northern part of the city and the southern neighbourhoods contained rather uncommitted metal industry companies (cf. Figure 2).

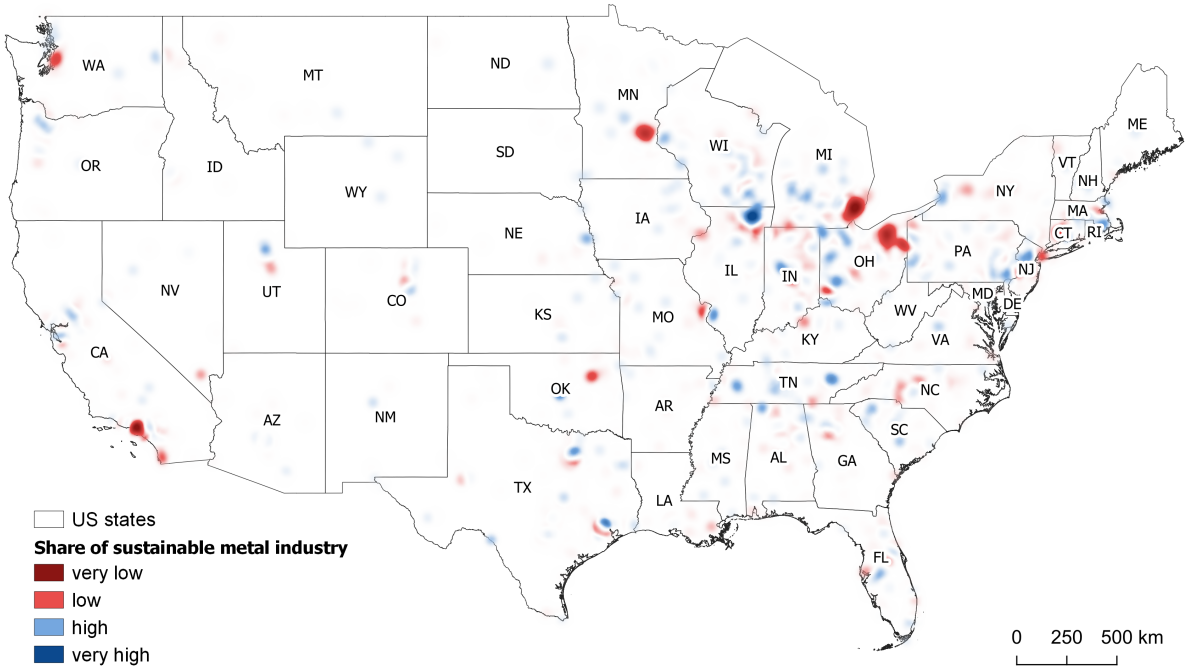

Figure 1: Heat map of the distribution of *sustainable* metal industry companies. The map was based on the combination of two separate bi-square kernel density maps with a pixel size of 1 km: one based on the locations of the *sustainable* and one for the *non-sustainable* metal industry. These were then normalised and subtracted from each other, creating a new raster with positive values in areas with a high (normalised) count of *sustainable* company locations.

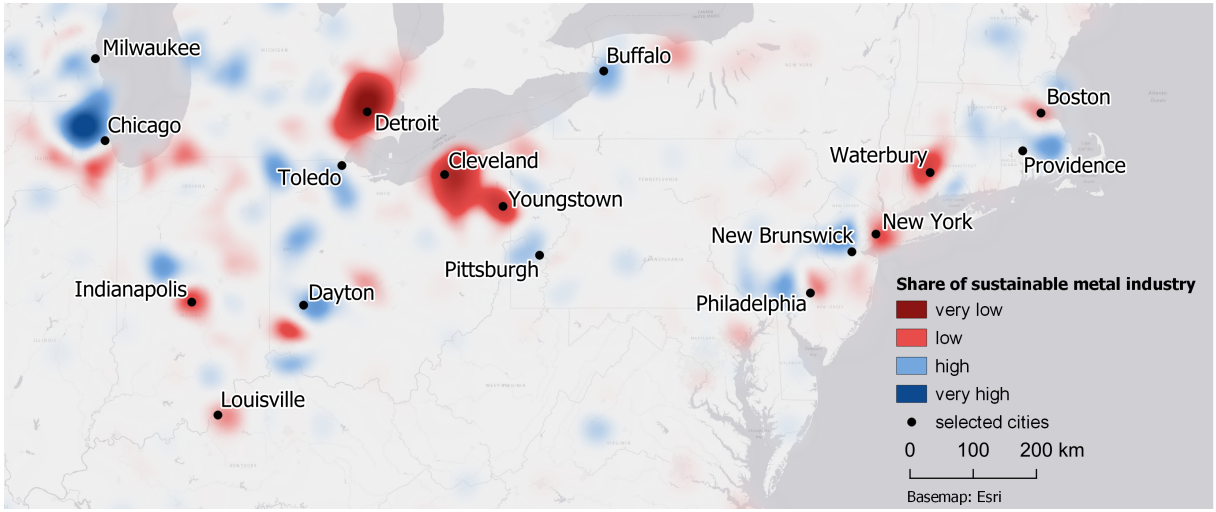

Figure 2: Heat map of the distribution of the *sustainable* metal industry on the East Coast. The same method as for Figure 1 was used.

Sustainable metal industry companies had more employees (average: 65.6, median: 30.0) than non-sustainable companies (average: 46.2, median: 18.0; t-test: 4.11 (p=0.000)). Companies without websites were also significantly smaller (average: 53.1, median: 14.0; t-test: 2.00 (p=0.045)). Concerning the average year of foundation, there was no major difference between the categories. However, since there was no year of foundation in the database for 48.4% of the firms, this statement is not too reliable.

5.2 Regression analysis

We calculated various specifications for OLS, SLX and SLM, considering the entire metal industry (*metal_log*). The results of the most suitable specifications can be found in Table 10. As variables with high VIF values should be removed from a model (Naughton et al., 2018), we had to exclude the land cover dummy variables, since half of the them were above the frequently used threshold of 10 (Altman & Krzywinski, 2016). The OLS model with the highest number of variables had the lowest Akaike information criterion (AIC) and showed strong positive spatial autocorrelation in its residuals (I_r: 0.861, p=0.001). The results of the SLX models were very similar and not trustworthy due to a likewise high Moran's I of the residuals (I_r: 0.862, p=0.001).

Based on the significant results of the standard and robust Lagrange Multiplier tests (cf. Table 9), a SLM seemed suitable (Ren & Matsumoto, 2020; Golgher & Voss, 2016). A SLM specification that included climatic conditions in addition to the basic variables was assumed the best-fitting model, as it was the specification with the lowest AIC in which the variable of interest (*metal_log*) was significant. The inclusion of a spatial lag of the dependent variable in the SLM clearly reduced the spatial autocorrelation of the residuals. The value of -0.013 (p=0.001) for I_r indicated that the residuals showed signs of a weak regular distribution. This negative autocorrelation presumably resulted in conservative, slightly too large standard errors.

Table 4: Regression coefficients and goodness of fit measures for both OLS and SLM. The results of this model were used to answer RQ 2. Statistical significance is expressed by asterisks according to the commonly used significance levels.

Model type	**OLS**		**SLM**	
variable	**coefficient**	**std.error**	**coefficient**	**std.error**
	constant			
constant	0.09813***	0.00060	0.01185***	0.00046
	variables of interest			
sustainable_log	0.00188**	0.00075	0.00016	0.00027
non_sustainable_log	0.00368***	0.00041	0.00036*	0.00015
no_website_log	0.00259***	0.00045	0.00047**	0.00016
	control variables			
manuf_log	0.00627***	0.00012	0.00072***	0.00005
veh_log	0.00072***	0.00006	-0.00044***	0.00002
power_log	0.00125	0.00095	-0.00016	0.00034
temp	0.00152***	0.00003	0.00017***	0.00001
prec	0.04512***	0.00033	0.00424***	0.00023
	spatially lagged dependent variable			
W (SO_2)	-	-	0.91569***	0.00432
R^2 / Pseudo R^2	0.16550		0.89476	
Spatial pseudo R^2	-		0.13102	
n	216,058		216,058	

Dependent variable: SO2_19_20, significance levels: * ≤ 0.05, ** ≤ 0.01, *** ≤ 0.001

In a further step, we calculated both an OLS and a SLM, for which the variable *metal_log* was split into three variables (*sustainable_log*, *non_sustainable_log*, *no_website_log*) based on our web classification. The Spearman correlation coefficients between these variables were well below the previously established threshold of 0.7. This was also confirmed by the VIFs for all variables, which were < 2, suggesting low multicollinearity.

In both models, the variable *sustainable_log* had a smaller coefficient estimate than both the variables non_sustainable_log and *no_website_log*) (cf. Table 4). However, it was only statistically significant in the OLS model. While it was the largest coefficient for the metal industry in the SLM, the coefficient for companies without a website fell between the coefficients for the web-based variables in the OLS model. Overall, the results of the two models were relatively similar. Only for the variables *veh_log* and *power_log* did the direction of the coefficients not match, the latter variable being non-significant in both models. There were also a few differences regarding the p-values.

6 Discussion

6.1 Interpretation

The results of the SLM, which considered the undifferentiated metal industry (*metal_log*), showed a clear correlation between the size of the metal industry and local SO_2 concentrations. The variable of interest had a significant coefficient of 0.00041, indicating that a 1% growth in the local number of employees in the metal industry would lead to an increase in SO_2 concentration of approximately 0.0004 mol/km^2.

For the specifications that differentiated the metal industry by sustainability, the results were not quite as clear. Although the coefficients between OLS and SLM were mostly similar, there were differences in terms of variable significance. This was especially true for one of the variables of interest, the size of the sustainable metal industry. If the self-proclaimed sustainable companies were indeed less polluting, we would expect the regression coefficient of *sustainable_log* to be non-significant (i.e. no effect at all on SO_2) or significantly smaller (i.e. less severe effect) than for the variables *non_sustainable_log* and *no_website_log*. In line with this expectation, *sustainable_log* was statistically significant in the OLS model and had a much smaller coefficient than *non_sustainable_log*. While this difference in coefficient also applied to the SLM, the variable *sustainable_log* was not significant in this model. Even when the class *sustainable* was delimited more strictly and only the categories *frontrunner* and *enabler* were considered, the results of the regression model hardly changed. Consequently, the size of the local *sustainable* metal industry did not have a statistically detectable influence on local SO_2 concentrations in this specification, while the *non-sustainable* metal industry had a measurable effect (i.e. increased the concentration of SO_2). The SLM also showed that companies without a website had an even stronger effect than non-sustainable companies. In addition, high SO_2 concentrations in the direct vicinity of the respective observation appeared to be of great importance, as the indirect effects in our model accounted for 91.6% of the total effects. This indicated that atmospheric SO_2 concentrations were strongly influenced by spillover emissions from the neighbourhood of a location. This might also have affected the significance of the variable of interest and thus explain the deviation between OLS and SLM.

While the results of the models differed somewhat, they did not contradict our expectations. In addition to our manual validation of the web-based sustainability variable, these results also suggested that the applied AI text analysis model was indeed able to correctly classify the web page content. If the classification had been random, identical effects for *sustainable_log* and *non_sustainable_log* would have been expected in the regression. Accordingly, we interpreted the results as follows:

a. Our web-based classification approach correctly classified texts of corporate websites into *sustainable* and *non-sustainable*.

b. We did not observe greenwashing. Companies presenting themselves as *sustainable* actually differed in their contribution to local SO2 concentrations from companies that did not present themselves as *sustainable*.

The positive coefficients of the variables *metal_log* and *manuf_log* were in line with Yang et al., 2017b, who find positive coefficients for the secondary sector share in their regression models on SO_2 levels in China. Contrary to our results, however, they demonstrate negative coefficients for precipitation and average temperatures. Other studies, such as Li et al., 2014, find a positive correlation between air pollutants and population density. Our SLM, on the other hand, provided a negative coefficient for vehicle emissions, which we considered an approximation of the population distribution. One possible, general explanation is that these deviations from other studies might be attributed to the variable structure (e.g. aggregation, resampling, average values). In the case of vehicle emissions, one conceivable explanation could be that high traffic emissions in the US occur mainly along the major highways and near downtown areas. Thus, they are not necessarily in spatial proximity to areas with high population density or to important SO_2 emitters such as power plants or industrial sites. For the climatic variables, it is also conceivable that other factors (e.g. humidity, air stratification, wind characteristics) have a higher importance than precipitation and temperature.

6.2 Limitations

Anthropogenic emissions often develop in a so-called *Environmental Kuznets Curve*, meaning that they increase until a certain threshold in the socio-economic development of a country and then begin to decline as a result of implemented environmental protection measures (Ru et al., 2018). As a consequence, there has been a shift in SO_2 emission centres from the industrial countries in North America and Europe to Asia. Around the turn of the millennium, China became the largest emitter of SO_2, but since 2016, the country has been overtaken by India (Li et al., 2017). As some researchers consider SO_2 pollution to be "well under control" in the US (Hidy & Blanchard, 2016), it would therefore be more sensible to conduct this research in Asia. However, the necessary business data was not available for this study, which consequently opens up the possibility of transferring the approach to another study area. For such an analysis, however, it would probably be advisable to decrease the size of the study area, which would allow taking more site-specific characteristics into account. However, due to the high spatial autocorrelation of the dependent variable, a stronger aggregation could also be sufficient for this. The results could perhaps also be improved by not relying on averaged data on SO_2 concentrations. Additionally, there was some temporal inconsistency between the different data sets, which should be avoided.

The validation of the web-based classification showed that some problems were caused by the fact that the model was originally intended for a broader definition of sustainability. For further analyses, a more specific model should be developed, trained on a more precise terminology concerning air pollution. Accuracy could also be increased by only considering websites that have been classified with high confidence by the model ensemble, expanding the category *no_website* to *unknown*.

6.3 Future research

To extend the web-based methodology, the inclusion of social media data might also be interesting, as networks such as Twitter or LinkedIn provide additional platforms for companies to report on their sustainability agenda. There has already been research on sustainability using social media data, e.g. by Chae and Park, 2018. For future research, it could also be sensible to consider natural SO_2 emissions, particularly if a study area with significant volcanic activity is investigated. Additional climatic factors, such as relative humidity (Ren & Matsumoto, 2020), as well as socio-economic variables (Zhou et al., 2018) could also be added. Other variables such as population density or the distance to the nearest harbour were excluded from our models due to multicollinearity. As this paper only deals with greenwashing in relation to SO_2 emissions, data for other air pollutants could also be included in subsequent studies.

Another aspect that was not considered in this study was the anisotropic influence of wind on the distribution of pollutants. The current spatial weights matrix handles all neighbours independently of their position in the neighbourhood. However, it might be beneficial to give a stronger weight to upwind areas. The inclusion of wind direction into the spatial weights matrix was not viable for the entire study area due to computational complexity. For such an analysis, it would probably be more sensible to employ daily SO_2 concentrations and wind data, which is possible due to the short revisit time of Sentinel-5P. Merk and Otto, 2020 already developed a suitable method for this.

7 Conclusion

Our analysis showed a positive and significant relationship between the size of the metal industry and SO_2 concentrations for the contiguous US. An essential and unique contribution of this study was the inclusion of web data to investigate greenwashing (RQ 1). We found that only 8.1% of the 9,430 companies in the US metal industry were classified as sustainable based on their online self-presentation. Sustainable firms had more employees on average and showed a tendency to be spatially clustered.

We interpreted the results of our OLS and SLM, which considered the web-based variables, as an indication of the absence of general greenwashing. Accordingly, we answered RQ 2 in the affirmative: The self-reporting of the companies regarding sustainability generally did seem to match the SO_2 concentrations obtained from remote sensing data. However, the high proportion of companies without a website (40.6%) was a clear limitation to these conclusions.

Acknowledgement

We want to thank Hannah Kemper, Theresa Keller, Tobias Hellmundt (University of Göttingen) and Dorian Arifi (University of Salzburg) for their helpful input. We are also grateful to Georg Licht (ZEW) for his support and helpful comments.

Funding sources

Sven Lautenbach acknowledges funding by the Klaus-Tschirra Stiftung.

CRediT author statement

Sebastian Schmidt: conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing - original draft, visualization; **Jan Kinne**: conceptualization, methodology, software, resources, writing - review & editing; **Sven Lautenbach**: conceptualization, methodology, writing - review & editing, supervision; **Thomas Blaschke**: conceptualization, methodology, writing - review & editing, supervision; **David Lenz**: software, resources; **Bernd Resch**: writing - review & editing, supervision

Data availability

Our data and script used for the regression analysis can be found in the following GitHub repository: https://github.com/cordoba27/US-metal_greenwashing.

Bibliography

Abatzoglou, J. T. (2012). 'Development of gridded surface meteorological data for ecological applications and modelling'. In: *International Journal of Climatology* 33(1), pp. 121-131. https://developers.google.com/earth-engine/datasets/catalog/IDAHO_EPSCOR_GRIDMET#description. Accessed February 25, 2021. DOI: 10.1002/joc.3413.

AISI (2020). *2020 profile of the American Iron and Steel Institute*. Accessed May 24, 2021. URL: https://www.steel.org/wp-content/uploads/2020/12/2020-AISI-Profile-Book.pdf.

Altman, N. and M. Krzywinski (2016). 'Regression diagnostics'. In: *Nature Methods* 13 (5), pp. 385–386. DOI: 10.1038/nmeth.3854.

Anselin, L. (2001). 'Spatial econometrics'. In: *A Companion to Theoretical Econometrics*. Ed. by B. H. Baltagi. Blackwell Publishing Ltd. Chap. 14, pp. 310–330. ISBN: 978-0-631-21254-6.

Bagnall, A., I. Whittley, M. Studley, M. Pettipher, F. Tekiner and L. Bull (2006). 'Variance stabilizing regression ensembles for environmental models'. In: *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, pp. 5355–5361. DOI: 10.1109/IJCNN.2006.247314.

Beaudry, C., C. Rietsch and M. Héroux-Vaillancourt (2016). 'Validation of a web mining technique to measure innovation in the Canadian nanotechnology-related community'. In: *Conference: CARMA 2016 - 1st International Conference on Advanced Research Methods and Analytics*. DOI: 10.4995/CARMA2016.2016.3140.

Brokamp, C., R. Jandarov, M. B. Rao, G. LeMasters and P. Ryan (2017). 'Exposure assessment models for elemental components of particulate matter in an urban environment: A comparison of regression and random forest approaches'. In: *Atmospheric Environment* 151, pp. 1–11. DOI: https://doi.org/10.1016/j.atmosenv.2016.11.066.

Brundtland, G. H. (1987). *Our common future: Report of the World Commission on environment and development*. Accessed July 12, 2021. URL: https://sustainabledevelopment.un.org/content/documents/5987our-common-future.pdf.

Chae, B. and E. Park (2018). 'Corporate social responsibility (CSR): A survey of topics and trends using Twitter data and topic modeling'. In: *Sustainability* 10 (7), p. 2231. DOI: 10.3390/su10072231.

Chatkin, J., L. Correa and U. Santos (2021). 'External environmental pollution as a risk factor for asthma'. In: *Clinical Reviews in Allergy & Immunology*, pp. 1–18. DOI: 10.1007/s12016-020-08830-5.

Chen, W., M. Wang and X. Li (2016). 'Analysis of copper flows in the United States: 1975–2012'. In: *Resources, Conservation and Recycling* 111, pp. 67–76. DOI: 10.1016/j.resconrec.2016.04.014.

Chen, X. and J. Ye (2015). 'When the wind blows: Spatial spillover effects of urban air pollution'. In: *Environment for Development. Discussion Paper Series*, pp. 1–32. URL: https://www.jstor.org/stable/resrep15025.

Chen, X., S. Shao, Z. Tian, Z. Xie and P. Yin (2017). 'Impacts of air pollution and its spatial spillover effect on public health based on China's big data sample'. In: *Journal of Cleaner Production* 142 (Part 2), pp. 915–925. DOI: 10.1016/j.jclepro.2016.02.119.

Cheng, Z., L. Li and J. Liu (2017). 'Identifying the spatial effects and driving factors of urban $PM_{2.5}$ pollution in China'. In: *Ecological Indicators* 82, pp. 61–75. DOI: 10.1016/j.ecolind.2017.06.043.

Cirtina, D., O. Chivu and L. M. Cirtina (2016). 'Assessment of air pollutants produced by industrial activity from an aluminium alloys foundry'. In: *Metalurgija* 55 (1), pp. 11–14.

Conrad, O., B. Bechtel, M. Bock, H. Dietrich, E. Fischer, L. Gerlitz, J. Wehberg, V. Wichmann and J. Böhner (2015). 'System for automated geoscientific analyses (SAGA) v. 2.1. 4'. In: *Geoscientific Model Development* 8 (7), pp. 1991–2007.

Cromar, K. R., B. N. Duncan, A. Bartonova, K. Benedict, M. Brauer, R. Habre, G. S. W. Hagler, J. A. Haynes, S. Khan, V. Kilaru, Y. Liu, S. Pawson, D. B. Peden, J. K. Quint, M. B. Rice, E. N. Sasser, E. Seto, S. L. Stone, G. D. Thurston and J. Volckens (2019). 'Air pollution monitoring for health research and patient care. An official American Thoracic Society workshop report'. In: *American Thoracic Society Documents* 16 (10), pp. 1207–1214. DOI: 10.1513/AnnalsATS.201906-477ST.

de Winter, J. C. F. and S. D. Gosling (2016). 'Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data'. In: *Psychological Methods* 21 (3), pp. 273–290. DOI: 10.1037/met0000079.

Delmas, M. A. and V. C. Burbano (2011). 'The drivers of greenwashing'. In: *California Management Review* 54 (1), pp. 64–87. DOI: 10.1525/cmr.2011.54.1.64.

Dormann, C. F., J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carré, J. R. García Marquéz, B. Gruber, B. Lafourcade, P. J. Leitão, T. Münkemüller, C. McClean, P. E. Osborne, B. Reineking, B. Schröder, A. K. Skidmore, D. Zurell and S. Lautenbach (2013). 'Collinearity: A review of methods to deal with it and a simulation study evaluating their performance'. In: *Ecography* 36 (1), pp. 27–46. DOI: 10.1111/j.1600-0587.2012.07348.x.

Dörr, J. O., J. Kinne, D. Lenz, G. Licht and P. Winker (2021). 'An integrated data framework for policy guidance in times of dynamic economic shocks'. In: *ZEW Discussion Paper No. 21-062*.

EPA (2019a). *1990 vs 2019 annual SO_2 comparisons. Acid rain program and cross-state air pollution rule emissions and changes at facilities*. https://www.epa.gov/sites/production/files/2020-03/ge_map_data_2019.xls. Accessed April 22, 2021.

– (2019b). *Power plant emissions trends*. Accessed April 22, 2021. URL: https://www.epa.gov/airmarkets/power-plant-emission-trends.

– (2021). *Criteria Air Pollutants*. Accessed July 27, 2021. URL: https://www.epa.gov/criteria-air-pollutants.

Fenton, M. D. (2005). *Mineral commodity profiles - Iron and steel: U.S. Geological Survey open-file report 2005-1254*. Accessed July 20, 2021. URL: https://pubs.usgs.gov/of/2005/1254/2005-1254.pdf.

Financial Times (2021). *Biden hails EU-US steel deal as chance to curb 'dirty' Chinese imports*. Accessed November 12, 2021. URL: https://www.ft.com/content/2c49dedf-6a32-484c-809a-14d00f2f9c57.

Fioletov, V., C. A. McLinden, S. K. Kharol, N. A. Krotkov, C. Li, J. Joiner, M. D. Moran, R. Vet, A. J. H. Visschedijk and H. A. C. Denier van der Gon (2017). 'Multi-source SO_2 emission retrievals and consistency of satellite and surface measurements with reported emissions'. In: *Atmospheric Chemistry and Physics* 17 (20), pp. 12597–12616. DOI: 10.5194/acp-17-12597-2017.

Garg, A., P. R. Shukla, S. Bhattacharya and V. K. Dadhwal (2001). 'Sub-region (district) and sector level SO_2 and NO_x emissions for India: Assessment of inventories and mitigation flexibility'. In: *Atmospheric Environment* 35 (4), pp. 703–713. DOI: 10.1016/S1352-2310(00)00316-2.

Gately, C., L. R. Hutyra and I. S. Wing (2019). *DARTE annual on-road CO_2 emissions on a 1-km grid, conterminous USA, V2, 1980-2017*. en. https://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=1735. Accessed March 24, 2021. DOI: 10.3334/ORNLDAAC/1735.

Gök, A., A. Waterworth and P. Shapira (2015). 'Use of web mining in studying innovation'. In: *Scientometrics* 102, pp. 653–671. DOI: 10.1007/s11192-014-1434-0.

Golgher, A. B. and P. R. Voss (2016). 'How to interpret the coefficients of spatial models: Spillovers, direct and indirect effects'. In: *Spatial Demography* 4, pp. 175–205. DOI: 10.1007/s40980-015-0016-y.

Gorelick, N., M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau and R. Moore (2017). 'Google Earth Engine: Planetary-scale geospatial analysis for everyone'. In: *Remote Sensing of Environment* 202, pp. 18–27. DOI: 10.1016/j.rse.2017.06.031.

Goudarzi, G., S. Geravandi, E. Idani, S. Ahmad Hosseini, M. Mehdi Baneshi, A. Reza Yari, M. Vosoughi, S. Dobaradaran, S. Shirali, M. Bagherian Marzooni, A. Ghomeishi, N. Alavi, S. Shaghayegh Alavi and M. Javad Mohammadi (2016). 'An evaluation of hospital admission respiratory disease attributed to sulfur dioxide ambient concentration in Ahvaz from 2011 through 2013'. In: *Environmental Science and Pollution Research* 23, pp. 22001–22007. DOI: 10.1007/s11356-016-7447-x.

Halleck Vega, S. and J. Elhorst (2015). 'The SLX Model'. In: *Journal of Regional Science* 55 (3), pp. 339–363. DOI: 10.1111/jors.12188.

Han, L., J. Zhao, Y. Gao, Z. Gu, K. Xin and J. Zhang (2020). 'Spatial distribution characteristics of $PM_{2.5}$ and PM_{10} in Xi'an City predicted by land use regression models'. In: *Sustainable Cities and Society* 61, pp. 102329–102345. DOI: 10.1016/j.scs.2020.102329.

Hasanbeigi, A. and C. Springer (2019). *How clean is the U.S. steel industry? An international benchmarking of energy and CO_2 intensities*. Accessed February 2, 2021. URL: https://aciercanadien.ca/files/resources/HowCleanistheU.S.SteelIndustry.pdf.

Hashim, B. M., S. K. Al-Naseri, A. Al-Maliki and N. Al-Ansari (2021). 'Impact of COVID-19 lockdown on NO_2, O_3, $PM_{2.5}$ and PM_{10} concentrations and assessing air quality changes in Baghdad, Iraq'. In: *Science of the Total Environment* 754, p. 141978. DOI: 10.1016/j.scitotenv.2020.141978.

He, H., K. Y. Vinnikov, C. Li, N. A. Krotkov, A. R. Jongeward, Z. Li, J. W. Stehr, J. C. Hains and R. R. Dickerson (2016). 'Response of SO_2 and particulate air pollution to local and regional emission controls: A case study in Maryland'. In: *Earth's Future* 4 (4), pp. 94–109. DOI: 10.1002/2015EF000330.

Hedelt, P., D. S. Efremenko, D. G. Loyola, R. Spurr and L. Clarisse (2019). 'Sulfur dioxide layer height retrieval from Sentinel-5 Precursor/TROPOMI using FP_ILM'. In: *Atmospheric Measurement Techniques* 12 (10), pp. 5503–5517. DOI: 10.5194/amt-12-5503-2019.

Hidy, G. M. and C. L. Blanchard (2016). 'The changing face of lower tropospheric sulfur oxides in the United States'. In: *Elementa: Science of the Anthropocene* 4. DOI: 10.12952/journal.elementa.000138.

Infogroup (2018). *Infogroup US Historical Business Data*. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/PNOFKI. Accessed July 6, 2021. DOI: 10.7910/DVN/PNOFKI.

Johnson, A. W., M. Gutiérrez, D. Gouzie and L. R. McAliley (2016). 'State of remediation and metal toxicity in the Tri-State Mining District, USA'. In: *Chemosphere* 144, pp. 1132–1141. DOI: 10.1016/j.chemosphere.2015.09.080.

Jordahl, K. et al. (2020). *geopandas/geopandas: v0.8.1*. DOI: 10.5281/zenodo.3946761.

Kaplan, G. and Z. Y. Avdan (2020). 'Space-borne air pollution observation from Sentinel-5p TROPOMI: Relationship between pollutants, geographical and demographic data'. In: *International Journal of Engineering and Geosciences* 5 (3), pp. 130–137. DOI: 10.26833/ijeg.644089.

Kaplan, G., Z. Y. Avdan and U. Avdan (2019). 'Spaceborne nitrogen dioxide observations from the Sentinel-5P TROPOMI over Turkey'. In: *Proceedings* 18 (1), pp. 1–6. DOI: 10.3390/ECRS-3-06181.

Karimian, H., Q. Li, C. Wu, Y. Qi, Y. Mo, G. Chen, S. Sachdeva and X. Zhang (2019). 'Evaluation of different machine learning approaches in forecasting PM2.5 mass concentrations'. In: *Aerosol and Air Quality Research* 19, pp. 1400–1410. DOI: 10.4209/aaqr.2018.12.0450.

Kinne, J. and J. Axenbeck (2020). 'Web mining for innovation ecosystem mapping: A framework and a large-scale pilot study'. In: *Scientometrics* 125, pp. 2011–2041. DOI: 10.1007/s11192-020-03726-9.

Kinne, J., M. Krüger, D. Lenz, G. Licht and P. Winker (2020). 'Coronavirus pandemic affects companies differently. A high-frequency website analysis of companies' reactions to the coronavirus pandemic in Germany'. In: *ZEW expert brief* 20-05. URL: https://ftp.zew.de/pub/zew-docs/ZEWKurzexpertisen/EN/ZEW_Shortexpertise2005.pdf.

Kinne, J. and D. Lenz (2021). 'Predicting innovative firms using web mining and deep learning'. In: *Plos One* 16 (4). DOI: 10.1371/journal.pone.0249071.

Kumar, D. S., S. H. Bhushan and D. A. Kishore (2018). 'Atmospheric dispersion model to predict the impact of gaseous pollutant in an industrial and mining cluster'. In: *Global Journal of Environmental Science and Management* 4 (3), pp. 351–358. DOI: 10.22034/gjesm.2018.03.008.

Lee, C., R. V. Martin, A. van Donkelaar, H. Lee, R. R. Dickerson, J. C. Hains, N. Krotkov, A. Richter, K. Vinnikov and J. J. Schwab (2011). 'SO_2 emissions and lifetimes: Estimates from inverse modeling using in situ and global, space-based (SCIAMACHY and OMI) observations'. In: *Journal of Geophyiscal Research: Atmospheres* 116 (D6). DOI: 10.1029/2010JD014758.

Lee, K.-H. (2017). 'Does size matter? Evaluating corporate environmental disclosure in the Australian mining and metal industry: A combined approach of quantity and quality measurement'. In: *Business Strategy and the Environment* 26 (2), pp. 209–223. DOI: 10.1002/bse.1910.

Lewinschal, A., A. M. L. Ekman, H.-C. Hansson, M. Sand, T. K. Berntsen and J. Langner (2019). 'Local and remote temperature response of regional SO_2 emissions'. In: *Atmospheric Chemistry and Physics* 19 (4), pp. 2385–2403. DOI: 10.5194/acp-19-2385-2019.

Li, C., C. McLinden, V. Fioletov, N. Krotkov, S. A. Carn, J. Joiner, D. Streets, H. Hao, X. Ren, Z. Li and R. R. Dickerson (2017). 'India is overtaking China as the world's largest emitter of anthropogenic sulfur dioxide'. In: *Scientific Reports* 7 (1). DOI: 10.1038/s41598-017-14639-8.

Li, H. (2017). 'Deep learning for natural language processing: Advantages and challenges'. In: *National Science Review* 5 (1), pp. 24–26. DOI: 10.1093/nsr/nwx110.

Li, Q., J. Song, E. Wang, H. Hu, J. Zhang and Y. Wang (2014). 'Economic growth and pollutant emissions in China: A spatial econometric analysis'. In: *Stochastic Environmental Research and Risk Assessment* 28, pp. 429–442. DOI: 10.1007/s00477-013-0762-6.

Li, R., L. Cui, Y. Meng, Y. Zhao and H. Fu (2019). 'Satellite-based prediction of daily SO_2 exposure across China using a high-quality random forest-spatiotemporal Kriging (RF-STK) model for health risk assessment'. In: *Atmospheric Environment* 208, pp. 10–19. DOI: 10.1016/j.atmosenv.2019.03.029.

Ma, S., Z. Wen and J. Chen (2012). 'Scenario analysis of sulfur dioxide emissions reduction potential in China's iron and steel industry'. In: *Journal of Industrial Ecology* 16 (4), pp. 506–517. DOI: 10.1111/j.1530-9290.2011.00418.x.

Mallik, C., P. S. Mahapatra, P. Kumar, S. Panda, R. Boopathy, T. Das and S. Lal (2019). 'Influence of regional emissions on SO_2 concentrations over Bhubaneswar, a capital city in eastern India downwind of the Indian SO_2 hotspots'. In: *Atmospheric Environment* 209, pp. 220–232. DOI: 10.1016/j.atmosenv.2019.04.006.

Malte, A. and P. Ratadiya (2019). 'Evolution of transfer learning in natural language processing'. In: *Computing Research Repository*. URL: http://arxiv.org/abs/1910.07370.

Menz, F. C. and H. M. Seip (2004). 'Acid rain in Europe and the United States: An update'. In: *Environmental Science & Policy* 7 (4), pp. 253–265. DOI: 10.1016/j.envsci.2004.05.005.

Merk, M. S. and P. Otto (2020). 'Estimation of anisotropic, time-varying spatial spillovers of fine particulate matter due to wind direction'. In: *Geographical Analysis* 52 (2), pp. 254–277. DOI: 10.1111/gean.12205.

Metya, A., P. Dagupta, S. Halder, S. Chakraborty and Y. K. Tiwari (2020). 'COVID-19 lockdowns improve air quality in the South-East Asian regions, as seen by the remote sensing satellites'. In: *Aerosol and Air Quality Research* 20 (8), pp. 1772–1782. DOI: 10.4209/aaqr.2020.05.0240.

Mitchell, M. J. and G. E. Likens (2011). 'Watershed sulfur biogeochemistry: Shift from atmospheric deposition dominance to climatic regulation'. In: *Environmental Science & Technology* 45 (12), pp. 5267–5271. DOI: 10.1021/es200844n.

Muñoz Sabater, J. (2019). *ERA5-Land monthly averaged data from 1981 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS)*. https://developers.google.com/earth-engine/datasets/catalog/ECMWF_ERA5_LAND_MONTHLY. Accessed February 25, 2021. DOI: 10.24381/cds.68d2bb30.

NAICS Association (2018). *Standard Industrial Code Divisions. Major group: 33—Primary metal industries*. Accessed July 6, 2021. URL: https://www.naics.com/standard-industrial-code-divisions/?code=33.

Naughton, O., A. Donnelly, P. Nolan, F. Pilla, B. D. Misstear and B. Broderick (2018). 'A land use regression model for explaining spatial variation in air pollution levels using a wind sector based approach'. In: *Science of the Total Environment* 630, pp. 1324–1334. DOI: 10.1016/j.scitotenv.2018.02.317.

NPMS (2019). *Commercially navigable waterway (CNW) data - Version 5*. https://www.npms.phmsa.dot.gov/CNWData.aspx. Accessed March 24, 2021.

Oxoli, D., J. R. Cedeno Jimenez and M. A. Brovelli (2020). 'Assessment of Sentinel-5P performance for ground-level air quality monitoring: Preparatory experiments over the COVID-19 lockdown period'. In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLIV-3/W1-2020, 2020 Gi4DM 2020 – 13th GeoInformation for Disaster Management conference, 30 November–4 December 2020, Sydney, Australia*, pp. 111–116. DOI: 10.5194/isprs-archives-XLIV-3-W1-2020-111-2020.

QGIS Development Team (2021). *QGIS Geographic Information System*. QGIS Association. URL: https://www.qgis.org.

Queißer, M., M. Burton, N. Theys, F. Pardini, G. Salerno, T. Caltabiano, M. Varnam, B. Esse and R. Kazahaya (2019). 'TROPOMI enables high resolution SO_2 flux observations from Mt. Etna, Italy, and beyond'. In: *Scientific Reports* 9. DOI: 10.1038/s41598-018-37807-w.

Ren, L. and K. Matsumoto (2020). 'Effects of socioeconomic and natural factors on air pollution in China: A spatial panel data analysis'. In: *Science of the Total Environment* 740, p. 140155. DOI: 10.1016/j.scitotenv.2020.140155.

Rey, S. J. and L. Anselin (2007). 'PySAL: A Python library of spatial analytical methods'. In: *The Review of Regional Studies* 37(1), pp. 5-27. DOI: 10.52324/001c.8285.

Romahn, F., M. Pedergnana, D. Loyola, A. Apituley, M. Sneep, J. P. Veefkin, N. Theys and P. Hedelt (2020). *Sentinel-5 precursor/TROPOMI Level 2 product user manual sulphur dioxide SO_2*. Accessed March 10, 2020. URL: https://sentinel.esa.int/documents/247904/2474726/Sentinel-5P-Level-2-Product-User-Manual-Sulphur-Dioxide.

Ru, M., D. T. Shindell, K. M. Seltzer, S. Tao and Q. Zhong (2018). 'The long-term relationship between emissions and economic growth for SO_2, CO_2, and BC'. In: *Environmental Research Letters* 13 (12). DOI: 10.1088/1748-9326/aaece2.

Schmalensee, R. and R. N. Stavins (2013). 'The SO_2 Allowance Trading System: The ironic history of a grand policy experiment'. In: *Journal of Economic Perspectives* 27 (1), pp. 103–122. DOI: 10.1257/jep.27.1.103.

Schwierzy, J., R. Dehghan, S. Schmidt, E. Rodepeter, A. Stömmer, K. Uctum, J. Kinne, D. Lenz and H. Hottenrott (2022). *Technology mapping using WebAI: The case of 3D printing*. https://arxiv.org/abs/2201.01125.

Sentinel-5P (2021). *Sentinel-5P OFFL SO2: Offline sulphur dioxide*. https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S5P_OFFL_L3_SO2#description. Accessed February 25, 2021.

Smith, S. J., J. van Aardenne, Z. Klimont, R. J. Andres, A. Volke and S. Delgado Arias (2011). 'Anthropogenic sulfur dioxide emissions: 1850–2005'. In: *Atmospheric Chemistry and Physics* 11 (3), pp. 1101–1116. DOI: 10.5194/acp-11-1101-2011.

Song, H. and M. Yang (2014). 'Analysis on effectiveness of SO_2 emission reduction in Shanxi, China by satellite remote sensing'. In: *Atmosphere* 5 (4), pp. 830–846. DOI: 10.3390/atmos5040830.

Theys, N., P. Hedelt, I. De Smedt, C. Lerot, H. Yu, J. Vlietinck, M. Pedergnana, S. Arellano, B. Galle, D. Fernandez, C. J. M. Carlito, C. Barrington, B. Taisne, H. Delgado-Granados, D. Loyola and M. Van Roozendael (2019). 'Global monitoring of volcanic SO_2 degassing with unprecedented resolution from TROPOMI onboard Sentinel-5 Precursor'. In: *Scientific Reports* 9. DOI: 10.1038/s41598-019-39279-y.

Theys, N., I. De Smedt, H. Yu, T. Danckaert, J. van Gent, C. Hörmann, T. Wagner, P. Hedelt, H. Bauer, F. Romahn, M. Pedergnana, D. Loyola and M. Van Roozendael (2017). 'Sulfur dioxide retrievals from TROPOMI onboard Sentinel-5 Precursor: Algorithm theoretical basis'. In: *Atmospheric Measurement Techniques* 10, pp. 119–153. DOI: 10.5194/amt-10-119-2017.

Tiefelsdorf, M., D. A. Griffith and B. Boots (1999). 'A variance-stabilizing coding scheme for spatial link matrices'. In: *Environment and Planning A: Economy and Space* 31 (1), pp. 165–180. DOI: 10.1068/a310165.

USCB (2018). *State area measurements and internal point coordinates*. Accessed May 24, 2021. URL: https://www.census.gov/geographies/reference-files/2010/geo/state-area.html.

USGS (2011). *North America rivers and lakes*. sciencebase.gov/catalog/item/4fb55df0e4b04cb937751e02. Accessed May 26, 2021.

– (2012). *USGS National Elevation Dataset 1/3 arc-second*. https://developers.google.com/earth-engine/datasets/catalog/USGS_NED?hl=en. Accessed February 25, 2021.

– (2020). *Iron and steel data sheet*. Accessed May 16, 2021. URL: https://pubs.usgs.gov/periodicals/mcs2020/mcs2020-iron-steel.pdf.

Veefkind, J. P. et al. (2012). 'TROPOMI on the ESA Sentinel-5 Precursor: A GMES mission for global observations of the atmospheric composition for climate, air quality and ozone layer applications'. In: *Remote Sensing of Environment* 120, pp. 70–83. DOI: 10.1016/j.rse.2011.09.027.

Verhoelst, T. et al. (2021). 'Ground-based validation of the Copernicus Sentinel-5P TROPOMI NO_2 measurements with the NDACC ZSL-DOAS, MAX-DOAS and Pandonia global networks'. In: *Atmospheric Measurement Techniques* 14 (1), pp. 481–510. DOI: 10.5194/amt-14-481-2021.

Wang, Z., P. Ma, L. Zhang, H. Chen, S. Zhao, W. Zhuo, C. Chen, Y. Zhang, C. Zhou, H. Mao, Y. Wang, Y. Wang, L. Zhang, A. Zhao, G. Weng and K. Hu (2020). 'Systematics of atmospheric environment monitoring in China via satellite remote sensing'. In: *Air Quality, Atmosphere & Health* 14, pp. 157–169. DOI: 10.1007/s11869-020-00922-7.

WorldPop (2020). *United States of America - Population density*. en. https://www.worldpop.org/geodata/summary?id=39730. Accessed March 23, 2020. DOI: 10.5258/SOTON/WP00674.

Worrell, E., P. Blinde, M. Neelis, E. Blomen and E. Masanet (2010). *Energy efficiency improvement and cost saving opportunities for the U.S. iron and steel industry. An ENERGY STAR® guide for energy and plant managers*. Accessed February 2, 2021. URL: https://www.osti.gov/servlets/purl/1026806.

Wu, Z., Y. Chen, Y. Han, T. Ke and Y. Liu (2020). 'Identifying the influencing factors controlling the spatial variation of heavy metals in suburban soil using spatial regression models'. In: *Science of the Total Environment* 717, p. 137212. DOI: 10.1016/j.scitotenv.2020.137212.

Xu, B. and B. Lin (2020). 'Investigating drivers of CO_2 emission in China's heavy industry: A quantile regression analysis'. In: *Energy* 206. DOI: 10.1016/j.energy.2020.118159.

Xu, Y., H. C. Ho, M. S. Wong, C. Deng, Y. Shi, T.-C. Chan and A. Knudby (2018). 'Evaluation of machine learning techniques with multiple remote sensing datasets in estimating monthly concentrations of ground-level $PM_{2.5}$'. In: *Environmental Pollution* 242 (Part B), pp. 1417–1426. DOI: 10.1016/j.envpol.2018.08.029.

Yang, L., S. Jin, P. Danielson, C. Homer, L. Gass, A. Case, C. Costello, J. Dewitz, J. Fry, M. Funk, B. Grannemann, M. Rigge and G. Xian (2018). 'A new generation of the United States National Land Cover Database: Requirements, research priorities, design, and implementation strategies'. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 146, pp. 108-123. https://developers.google.com/earth-engine/datasets/catalog/USGS_NLCD#bands. Accessed February 25, 2021.

Yang, X., S. Wang, W. Zhang, D. Zhan and J. Li (2017a). 'The impact of anthropogenic emissions and meteorological conditions on the spatial variation of ambient SO_2 concentrations: A panel study of 113 Chinese cities'. In: *Science of the Total Environment* 584-585, pp. 318–328. DOI: 10.1016/j.scitotenv.2016.12.145.

Yang, X., S. Wang, W. Zhang and J. Yu (2017b). 'Are the temporal variation and spatial variation of ambient SO_2 concentrations determined by different factors?' In: *Journal of Cleaner Production* 167, pp. 824–836. DOI: 10.1016/j.jclepro.2017.08.215.

Zhang, Q. et al. (2007). 'Ubiquity and dominance of oxygenated species in organic aerosols in anthropogenically-influenced Northern Hemisphere midlatitudes'. In: *Geophysical Research Letters* 34 (13), pp. 1–6. DOI: 10.1029/2007GL029979.

Zhao, R., L. Zhan, M. Yao and L. Yang (2020). 'A geographically weighted regression model augmented by Geodetector analysis and principal component analysis for the spatial distribution of $PM_{2.5}$'. In: *Sustainable Cities and Society* 56. DOI: 10.1016/j.scs.2020.102106.

Zhao, X., W. Zhou, L. Han and D. Locke (2019). 'Spatiotemporal variation in $PM_{2.5}$ concentrations and their relationship with socioeconomic factors in China's major cities'. In: *Environmental International* 133 (Part A), p. 105145. DOI: 10.1016/j.envint.2019.105145.

Zheng, C., C. Zhao, Y. Li, X. Wu, K. Zhang, J. Gao, Q. Qiao, Y. Ren, X. Zhang and F. Chai (2018). 'Spatial and temporal distribution of NO_2 and SO_2 in Inner Mongolia urban agglomeration obtained from satellite remote sensing and ground observations'. In: *Atmospheric Environment* 188, pp. 50–59. DOI: 10.1016/j.atmosenv.2018.06.029.

Zheng, Z., Z. Yang, Z. Wu and F. Marinello (2019). 'Spatial variation of NO_2 and its impact factors in China: An application of Sentinel-5P products'. In: *Remote Sensing* 11 (16), p. 1939. DOI: 10.3390/rs11161939.

Zhong, Q., H. Shen, X. Yun, Y. Chen, Y. Ren, H. Xu, G. Shen, W. Du, J. Meng, W. Li, J. Ma and S. Tao (2020). 'Global sulfur dioxide emissions and the driving forces'. In: *Environmental Science & Technology* 54 (11), pp. 6508–6517. DOI: 10.1021/acs.est.9b07696.

Zhou, C., J. Chen and S. Wang (2018). 'Examining the effects of socioeconomic development on fine particulate matter (PM2.5) in China's cities using spatial regression and the geographical detector technique'. In: *Science of the Total Environment* 619-620, pp. 436–445. DOI: 10.1016/j.scitotenv.2017.11.124.

Appendix

Table 5: NAICS codes for the metal industry.

Primary NAICS code	Description
2122	Metal Ore Mining
3311	Iron and Steel Mills and Ferroalloy Manufacturing
3312	Steel Product Manufacturing from Purchased Steel
3313	Alumina and Aluminum Production and Processing
3314	Nonferrous Metal Production and Processing
3315	Foundries
3321	Forging and Stamping
332811	Metal Heat Treating

Table 6: Reclassification of land cover data.

class	reclassified as	original classes
1	water	11 - open water 12 - perennial ice / snow
2	developed areas	21 - developed (open space) 22 - developed (low intensity) 23 - developed (medium intensity) 24 - developed (high intensity)
3	barren land	31 - barren land (rock / sand / clay)
4	forest	41 - deciduous forest 42 - evergreen forest 43 - mixed forest
5	scrub / grassland	52 - shrub / scrub 71 - grassland / herbaceous
6	agricultural	81 - pasture / hay 82 - cultivated crops
7	wetlands	90 - woody wetlands 95 - emergent herbaceous wetlands

Table 7: Keywords for text classification model. The same keywords were also translated to other Indo-European languages (e.g. German, Italian), as the model was developed for cross-language use.

Keyword		
Bio	Corporate responsibility	Good working conditions
Circular	Eco	Local Investments
Clean alternative	Emissions	Organic
Climate change	Environment	Sustainable
Closed-cycle	Equal treatment	Waste
CO2-free	ESG reporting	
Community engagement	Ethically	

Table 8: Classes for text classification.

class	definition
frontrunner	A company that sees sustainability as a central part of its business model, e.g. through carbon-neutral production.
enabler	A company that helps other companies to be more sustainable, e.g. a solar panel manufacturer.
engaged	A company that offers sustainable products or services, e.g. with eco-labels.
information	A company that only informs about sustainability, e.g. through blog posts.
not_engaged	A company that makes no or non-relevant statements on the topic of sustainability.

Table 9: Results of Lagrange Multiplier tests.

test	value	probability
Lagrange Multiplier (lag)	637,747.53	(0.00)***
Robust LM (lag)	4,387.59	(0.00)***
Lagrange Multiplier (error)	634,970.81	(0.00)***
Robust LM (error)	1,610.86	(0.00)***

Table 10: Regression analysis results using the aggregated employees of the metal industry as a predictor. Only the results of the most suitable model specification are included. The upper part presents goodness of fit measures as well as Moran's I as a measure of the global spatial autocorrelation of the residuals. For the SLM, R^2 is a Pseudo R^2. The lower part contains the estimates of the regression coefficients. W (metal_log) represents the effect of the spatially lagged predictor in the SLX model and W (SO_2) the effect of the lagged response variable in the SLM.

model	OLS	SLX	SLM
AIC	-557,275	-558,931	-976,115
RMSE	0.06663	0.06637	0.02528
Adjusted R^2	0.23920	0.24500	0.89480
I_r	0.861	0.862	-0.013
constant	0.18053***	0.18187***	0.01179***
metal_log	0.00342***	-0.00037	0.00041***
manuf_log	0.00481***	0.00339***	0.00072***
veh_log	0.00019**	-0.00010	-0.00043***
power_log	0.00074	0.00004	-0.00015
temp	0.00055***	0.00058***	0.00017***
prec	0.01234***	0.01244***	0.00421***
water	-0.00022***	-0.00022***	-
elev	-0.00003***	-0.00003***	-
W (metal_log)	-	0.02266***	-
W (SO_2)	-	-	0.91644***

Dependent variable: SO2_19_20, significance levels: * $\leq$ 0.05, ** $\leq$ 0.01, *** $\leq$ 0.001

Download ZEW Discussion Papers:

https://www.zew.de/en/publications/zew-discussion-papers

or see:

https://www.ssrn.com/link/ZEW-Ctr-Euro-Econ-Research.html
https://ideas.repec.org/s/zbw/zewdip.html

IMPRINT

ZEW – Leibniz-Zentrum für Europäische Wirtschaftsforschung GmbH Mannheim
ZEW – Leibniz Centre for European Economic Research

L 7,1 · 68161 Mannheim · Germany
Phone +49 621 1235-01
info@zew.de · zew.de