// NO.26-043 | 09/2026

DISCUSSION PAPER

// SEBASTIAN SCHMIDT, DEVIKA JAIN, JOHANNES DAHLKE,
JAN KINNE, ROBERT DEHGHAN, SHAILY GANDHI,
ROCKLI KIM, AND S.V. SUBRAMANIAN

Relating Corporate Sustainability Engagement with Environmental Quality: Empirical Evidence from India Using WebAI and Earth Observation Data

Leibniz Association

ZEW

Relating corporate sustainability engagement with environmental quality: Empirical evidence from India using WebAI and earth observation data

Sebastian Schmidt [a,b,c], Devika Jain [d], Johannes Dahlke [b,e,f], Jan Kinne [b,d,g], Robert Dehghan [b,h], Shaily Gandhi [c], Rockli Kim [i], S.V. Subramanian [j,k]

[a] *Department of Geoinformatics - Z_GIS, University of Salzburg, Austria*
[b] *ISTARI.AI, Germany*
[c] *GeoSocial Artificial Intelligence, Interdisciplinary Transformation University, Austria*
[d] *Center for Geographic Analysis, Harvard University, USA*
[e] *Entrepreneurship & Technology Management. University of Twente, Netherlands*
[f] *KOF D-MTEC, ETH Zurich, Switzerland*
[g] *Center for European Economic Research, Germany*
[h] *Center for SME Research and Entrepreneurship, University of Mannheim, Germany*
[i] *Division of Health Policy & Management, College of Health Science, Korea University, South Korea*
[j] *Harvard Center for Population and Development Studies, Harvard University, USA*
[k] *Department of Social and Behavioral Sciences, Harvard T. H. Chan School of Public Health, USA*

Abstract

Environmental pollution remains a critical challenge in developing countries. This paper thus investigates the relationship between corporate sustainability engagement and the environmental conditions of the districts in which firms operate across districts in India. Operationalising organisational websites, we apply deep learning methods to construct firm-level sustainability engagement scores and aggregate them to the district level. We combine these indicators with earth-observation measures of air pollution and vegetation and estimate district-level regressions using spatial lag models to account for spatial dependencies. We find that more polluted districts host more sustainability-engaged firms in absolute terms. Yet, this association is largely absorbed by industrial concentration, population density, and institutional quality once controls are included. The relative share of sustainability-engaged firms shows no significant relationship with air quality. The study highlights the potential of large-scale web text analysis combined with complementary geodata for examining environmental sustainability engagement. A key limitation is the low availability of firm-level data, highlighting the need for improved data infrastructure for environmental and sustainability research in emerging economies.

Keywords: WebAI, sustainability engagement, air pollution, spatial regression analysis
JEL codes: Q53, Q56, R11

1. Introduction

India is among the countries most severely impacted by environmental pollution and ecological degradation (Azimi and Rahman, 2024; Sehrawat et al., 2015). In particular, low air quality in Indian cities has been a persistent concern for decades due to its substantial impact on public health. This includes high concentrations of sulfur dioxide (SO_2) and nitrogen oxides (NO_x) in urban areas (Guttikunda et al., 2014), which are relevant

as precursor gases for particulate matter (Saikawa et al., 2017). An analysis from 2015-2019, based on Central Pollution Control Board (CPCB) and AirNow data, revealed that particulate matter (with a diameter of < 10 µm) (PM_{10}) and particulate matter (with a diameter of < 2.5 µm) ($PM_{2.5}$) dominate the pollution mix, especially in northern India, where levels exceed national standards by 100-150% (Sharma and Mauzerall, 2022). Influenced by its high population numbers, India is nevertheless one of the countries with the lowest per capita greenhouse gas emissions (Shahbaz et al., 2015).

Against the backdrop of vocal initiatives for climate protection measures (e.g. Fridays for Future), sustainability-related requirements such as Environmental, Social and Governance (ESG) reporting have become part of legislation in many countries. In this context, the external image of a company with regard to its commitment to sustainability is becoming increasingly important (Purwandani and Michaud, 2021). In India, this global momentum has materialised through a robust regulatory overhaul. In May 2021, Securities and Exchange Board of India (SEBI) replaced voluntary disclosures, with the formal Business Responsibility and Sustainability Reporting (BRSR) framework mandating standardised ESG reporting for the top 1,000 listed firms (Mohapatra and Seth, 2023). Building on India's nine National Voluntary Guidelines, BRSR aligns with leading global frameworks (e.g. Global Reporting Initiative (GRI)), requiring both essential and voluntary leadership indicators across core ESG domains. It introduces phased-in value-chain disclosure and third-party assurance requirements, scaling to all relevant companies by fiscal year 2026–27 (Upadhyay and Dugal, 2025). Empirical studies confirm a marked enhancement in the quality and breadth of ESG disclosures after BRSR's implementation, driven primarily by these new regulatory mandates (Nanjundaswamy et al., 2024). It also shows that India is targeting broader impact beyond its largest companies, acknowledging the need to steer its fast-paced industrial development. However, growing regulatory pressure does not automatically lead to proactive sustainability engagement of companies (Buysse and Verbeke, 2003). This suggests the need to adopt a broader institutional lens on how companies perceive and react to signals from their environment.

In this paper, we thus examine the alignment between green rhetoric and environmental reality. Specifically, we follow the WebAI paradigm (Dahlke et al., 2025) and use AI-based processing methods to generate information from a large sample of corporate websites to quantify the self-reported, ecological sustainability engagement of Indian companies. In a subsequent step, we aggregate this firm-level information to the district level and combine it with quantitative indicators on air quality and vegetational health derived from Earth Observation (EO) data, as well as socio-economic controls, to gain insights into the actual environmental quality of the districts in which these firms operate. We investigate these relationships through spatial regression analyses at the district level. Consequently, we seek to answer the following research questions:

- **RQ1**: Is there a correlation between firm density and environmental quality in India?
- **RQ2**: Is the environmental quality of a district associated with the sustainability engagement of the firms located there?

While there have been large-scale analyses using social media data in India (e.g. Kagan et al. (2015)), we are not aware of any nation-wide study based on corporate websites. Since research in environmental economics conducted on traditional data sources such as patents or surveys is generally limited in sample size (Kinne and Axenbeck, 2020), we address this gap for India with our paper. Additionally, there has generally been little

research so far combining textual web data with actual environmental measurements, particularly from EO data. Particularly for large-scale analyses, we argue that integrating these data sources has the potential to generate novel and valuable insights. Our study thus extends previous work by Schmidt et al. (2022) who only investigated the sustainability engagement of metal industry firms in the US and one air pollutant, namely SO_2, finding statistically significant relationships between web-based and satellite-based indicators.

2. Theoretical background

In the following, relevant literature from green innovation research, work on environmental issues in India, as well as previous WebAI studies (Dahlke et al., 2025) are summarised.

2.1. Green innovations

One of the most important, current challenges of our time is the *green transition*, i.e. the transformation of business models and processes into more sustainable and environmentally friendly forms (Martin, 2016). Consequently, corporate sustainability is a much-researched topic in economics and economic geography, as the implementation of green technologies has been identified as one of the primary instruments for reducing environmental pollution (Villanthenkodath and Mahalik, 2022). Chen et al. (2022) find that innovation in China considerably lessens firm-based pollution, conducting an analysis based on SO_2 emissions and patents. They also show that this impact varies depending on the sector and location. On a sample of 589 firms in Japan, Wang et al. (2021a) study the relation between innovation and carbon dioxide (CO_2) reduction, finding that particularly the manufacturing sector can generate ecological benefits. Liu et al. (2023) investigate the impact of the Chinese Pollution Charge System (PCS) on firm innovation, concluding that it strongly impedes small companies and businesses located in Western China. Bracht and Verhoeven (2025) find a negative effect of air pollution on innovation, investigating European regions. Similarly, Tan and Yan (2021) demonstrate that air pollution hampers corporate innovation in China by worsening financial constraints and adversely impacting human capital.

Following the induced innovation hypothesis established by Hicks (1932), firms tend to reorient their business models in response to external pressures, such as environmental constraints (e.g. stricter regulations, resource scarcity). Such pressures can stimulate innovation by increasing the incentives to develop more efficient technologies and production processes. Particularly, environmental regulations are widely recognised as a significant catalyst for (green) innovation (Melander, 2017; Zhu et al., 2012). In general, the transition to more sustainable business practices is shaped by a mix of external and internal drivers (Hojnik and Ruzzier, 2016). External factors include market-based incentives (Pellegrini et al., 2019) and the development of new markets (Beise and Rennings, 2005). Furthermore, government policy frameworks are implementing stringent environmental regulations in BRICS economies which have the potential to stimulate green technological innovation, contributing to reducing CO_2 emissions and environmental sustainability (Fatima et al., 2024). According to Porter and Linde (1995), such strict environmental regulations can trigger innovation that may even enhance firm competitiveness. This hypothesis has since then been differentiated into strong and weak versions (Jaffe and Palmer, 1997; Albrizio et al., 2014). Internal factors, by contrast, are typically rooted in a firm's technological capabilities and the expertise of its workforce (Horbach, 2008;

Pellegrini et al., 2019). Green investments can function as strategic signals of responsibility and differentiation, particularly in competitive or image-sensitive sectors, aligning with the signalling theory (Connelly et al., 2025). Despite these drivers, firms pursuing green innovation frequently encounter disincentives, such as the challenge of competing with firms that opt for cheaper, more polluting technologies (Rennings, 2000; Ramkumar et al., 2022). Consequently, Rodríguez-Pose and Bartalucci (2024) demonstrate significant regional variations in the impact of the green transition in Europe, highlighting peripheral regions as particularly vulnerable.

2.2. India and environmental quality

There have already been many studies on the topic of environmental quality in India, particularly with a focus on the impact of economic development. In studying the relationship between globalisation and CO_2 emissions through a cointegration approach, Shahbaz et al. (2015) confirm the Environmental Kuznets Curve (EKC) hypothesis, i.e. an inverted U-shape relationship between environmental pollution and economic growth (Kuznets, 1955), for India. They corroborate findings by Boutabba (2014) who identify a unidirectional causality between financial trends and carbon emissions. Using the EKC framework, Mohapatra and Giri (2009) highlight the importance of population density for air quality degradation and the positive impact of environmental spending in India. Villanthenkodath and Mahalik (2022) find that technological innovation has a negative impact on emissions in India. Tourism development has also been negatively related to environmental quality (Villanthenkodath et al., 2022). Nenavath (2022) shows that the implementation of green finance systems has a positive environmental impact on SO_2 emissions. Mahalik et al. (2021) find that foreign aid also decreases CO_2 emissions significantly. Investigating the link between socio-demographic factors and exposure to $PM_{2.5}$, deSouza et al. (2023) discover significant relationships with caste, religion and education levels. Similar findings by Chakraborty and Basu (2021) show considerable $PM_{2.5}$ increases in more rural districts with a high proportion of socially disadvantaged groups. A strong association between $PM_{2.5}$ exposure and neonatal mortality has also been shown (deSouza et al., 2022).

In the field of air pollution studies, satellite-based approximations have gained popularity as a data source in recent years due to their large spatial coverage and high temporal resolution (Yan et al., 2021; Holloway et al., 2021). Data from Sentinel-5P and its Tropospheric Monitoring Instrument (TROPOMI) has been used to analyse air pollution in India during the COVID-19 pandemic (Behera et al., 2022). Using Sentinel-5P data and a regression analysis approach, Reshi et al. (2022) identify an increase in ozone (O_3) levels despite decreasing nitrogen dioxide (NO_2) concentrations in Pune during lockdown. Mathew et al. (2024) find an overall increase in air pollutants from 2019 to 2022 with seasonal trends and different air pollution centres. For instance, SO_2 pollution in Kolkata is considerably higher than NO_2 pollution compared to Delhi, while Hyderabad is particularly affected by O_3 pollution. Muniraj et al. (2023) assess the concordance between Sentinel-5P and ground-based measurements between 83.5 and 94.0% for different pollutants.

H1: Firm density is positively correlated with local environmental pollution. Concentrated industrial activity typically elevates emissions and land use pressure, which deteriorates air quality and vegetational health.

2.3. Stakeholder theory and institutionalism

While there are many societal or even legal incentives for firms to increase their sustainability engagement, the relationship between such a commitment and the actual environment of the respective firm location remains unclear (Buysse and Verbeke, 2003). Stakeholder theory holds that a firm survives by creating value for every actor that can affect the firm or is affected by its activities (Freeman, 1984). Environmentally oriented corporate change is usually affected by a multitude of factors including regulators, customers, suppliers and local communities rather than by one homogeneous public. The weight of these stakeholder groups differs between places. For example, firms located in densely populated industrial districts may face more vocal non-governmental organisations, stricter municipal regulators and a watchful local press than firms in peripheral areas (Buysse and Verbeke, 2003; González-Benito and González-Benito, 2008; Ozen and Kusku, 2009). Companies are embedded in specific socio-institutional fabrics that shape how strongly external voices are considered. It is those place-based mechanisms that govern how shifting social norms diffuse inside geographical clusters, where firms monitor each other closely and imitate peers that introduce greener practices (Orsato, 2006). Given the place-based mechanisms, the diversity of stakeholder is facing a unifying exposure to the natural environment they operate in, which may affect their relations.

However, little is known about how the physical state of the ecological environment feeds back into managerial decisions. Two complementary mechanisms emerge from stakeholder theory and recent empirical studies: First, the firm population may be heterogeneously exposed to visible environmental issues. These impact visibility mechanisms may explain how deteriorating air quality or vegetation loss, which impose direct health and productivity costs on employees and residents, would lead organisations to respond in their economic activities. These costs heighten public concern, encourage regulators to tighten oversight and bring environmental issues onto the managerial agenda. Evidence from China shows that worsening air quality raises corporate environmental responsibility scores and triggers additional disclosure (Wang et al., 2021b; Lin and Ma, 2021). Second, legitimisation mechanisms might motivate firms to engage with sustainability more strongly. Firms operating in visibly polluted areas risk being labelled irresponsible. Investing in greener processes and communicating those efforts can ward off negative publicity and secure a social licence to operate (Buysse and Verbeke, 2003; Orsato, 2006). This leads us to hypothesise the following:

H2: Districts with poorer environmental conditions host a higher share of sustainability-engaged firms. Heightened stakeholder concern and the need to defend legitimacy raise the incentives for firms to adopt and communicate greener practices as they consider the interests of all their stakeholders. In addition, changing social norms could incentivise companies to establish more environmentally friendly processes and products.

2.4. Corporate websites and sustainability

In recent years, organisational websites have become a popular data source for economic analyses. This is due in particular to the liveness, scalability and semantic diversity of their content, i.e. their suitability for large-scale, up-to-date studies on various topics. The automatic scraping and subsequent Machine Learning (ML)-based analysis of content from organisational websites, has been referred to as the WebAI paradigm of innovation research (Dahlke et al., 2025). One of the first studies to fall under this research direction was Kinne and Lenz (2021) who develop a methodology to assess the innovativeness of a company by combining web scraping and deep learning, showing that their indicator is

comparable to traditional patent analyses. Other studies using data from organisational websites investigate topics related to specific emerging technologies, such as Artificial Intelligence (AI) (Dahlke et al., 2024; Pastor-Merino et al., 2026), 3D printing (Schwierzy et al., 2026), and blockchain (Gschnaidtner et al., 2024).

There have also been several studies on the topic of corporate sustainability engagement based on the WebAI paradigm. Studying the topic of widespread greenwashing in the US metal industry, Schmidt et al. (2022) combine a web-based sustainability engagement indicator with satellite-based air pollution measurements. While they find the topic of sustainability to be underrepresented in the metal industry and a correlation between firm locations and air pollution, they do not identify evidence for general greenwashing. Kinne et al. (2024) investigate the connection between blockchain knowledge and sustainability engagement in the US based on corporate website texts. Additionally, they consider location factors and network embedding to analyse what causes the sustainable use of blockchain technology, finding that network metrics are particularly relevant. Through analysing website texts and devising a sustainability indicator, Wildnerova et al. (2024) find that one-third of small and medium-sized enterprises (SME) in Europe are sustainability-engaged. Kriesch and Losacker (2024) analyse firm activities related to the bioeconomy based on website texts from CommonCrawl, revealing that these firms are primarily situated in rural areas near their biomass feedstocks. Auzepy et al. (2023) use a zero-shot approach to analyse climate-related financial disclosures from website texts. Schmidt et al. (2025) analyse the hyperlink networks of sustainability-engaged firms in the German-speaking countries, finding that these firms sustain denser network than the overall firm population and that spatial proximity still matters for collaboration. Dehghan et al. (2026) find an energy-reducing impact of AI through measuring adoption via website data and relating it to regional energy consumption in Germany. Riefolo et al. (2026) investigate family firms in Germany, finding that younger companies are more likely to report sustainability engagement on their websites. While WebAI studies have so far covered many different topics, their spatial focus has been on Europe and North America.

3. Methodology

In the following, we describe our employed methodology in detail. Figure 1 provides a high-level overview of our workflow.

We conducted our study for the entirety of India, which is the most populated country in the world (UN, 2023). Due to its enormous size, India consists of nine climate zones according to Köppen-Geiger's classification scheme (Yaduvanshi et al., 2021) and landscapes including sand deserts, marshlands, rain forests, and high-altitude cold deserts (Kale, 2014). For reasons of spatial contiguity, we excluded all firms in insular areas (e.g. Andaman and Nicobar islands).

3.1. Base data

For our analysis, we created a cross-sectional, single-year dataset. In a first step, we identified the Uniform Resource Locators (URLs) of companies in India by extracting all firms from the ORBIS database (Bureau van Dijk, 2022), a frequently used data source for firm-level economic analysis, that had a registered domain. This severely reduced the amount of firms from 15,234,271 to 433,963, representing merely 2.9%. This represented a major difference to previous studies using organisational website such as Kinne and

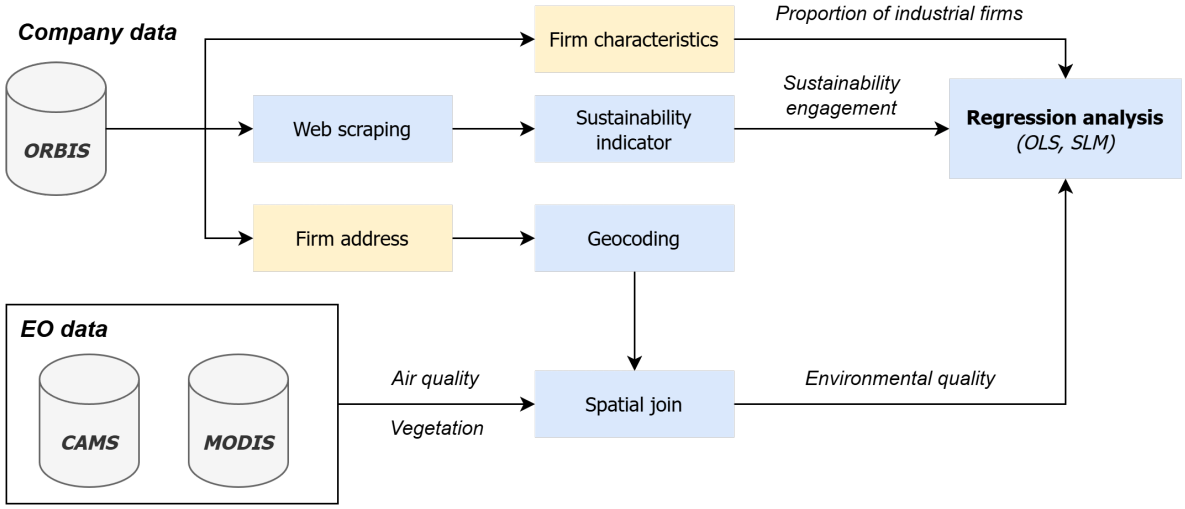


Figure 1: Methodology of study. Blue boxes represent methodological steps, yellow boxes are extracted information that has not been modified, grey boxes are data sources. Texts in italics are high-level descriptions of variables in the regression analysis.

Axenbeck (2020) who found that 46% of companies in Germany had their own website, while for some NACE sectors this number exceeded 70%.

We only kept firms in our analysis with a unique URL, since there were many entries in the ORBIS database that shared the same website (e.g. more than 5,000 firms pointing towards "*blogspot.com*"). Additionally, we also filtered based on the postal addresses. This was necessary, as some firms in the database only had information on city-level, which led to duplicate entries. For this filtering, we assumed that several firms could have the same address, keeping only firms where the address occurred ≤ 5 times in the entire dataset. These filtering steps further reduced our number of observations to 186,029.

Since the ORBIS database only contains addresses in string format, we had to perform geocoding to enable spatial joins with other data sources. For this, we used a two-step approach. First, we employed the geocoder Photon, developed by Komoot, which is based on OpenStreetMap (OSM). The geocoding returned coordinates for 108,549 firms (58.4%). The remaining firms were then geocoded using the Geocoding Service featured in ESRI's ArcGIS Online. In total, 117,622 firms (63.2%) could be geocoded. This further reduction of observations was caused by both erroneous addresses in ORBIS or formatting problems that the geocoders could not handle.

3.2. Web text mining

As a starting point for our semantic analysis, we then proceeded to scrape the corporate websites of all remaining observations, i.e. downloaded the corresponding HTML content and metadata in an automated manner. For each website, we considered the first 25 subpages. For this, we ordered the subpages based on the length of their respective URL, under the assumption that top-level pages are more likely to contain key information about a company profile (Kinne and Axenbeck, 2020). For our semantic analysis, we followed the approach presented by Wildnerova et al. (2024), Schmidt et al. (2022) and Kinne et al. (2024). To minimise computational effort, we first performed a keyword search on the extracted textual content to identify HTML paragraphs containing the following sustainability-related terms and related derivatives: *carbon capture*, *carbon*

dioxide, *carbon footprint*, *carbon neutral*, *carbon positive*, *circular*, *clean alternative*, *clean energy*, *cleantech*, *climate change*, *closed cycle*, *CO2 free*, *commitment to the community*, *compost*, *compostable*, *consistent*, *corporate responsibility*, *cradle to cradle*, *eco*, *efficient use of energy*, *electric car*, *electric vehicle*, *emission*, *emissions*, *energy*, *energy costs*, *energy efficient*, *energy transition*, *environment*, *environmental pollution*, *equal treatment*, *ESG reporting*, *ethically*, *food waste*, *footprint*, *good working conditions*, *hydrogen*, *local investments*, *made from plants*, *naturally*, *net zero*, *organic*, *photovoltaic*, *plastic pollution*, *rainwater*, *recycle*, *recycling*, *regenerative*, *renewable*, *reusable*, *reuse*, *save energy*, *sewage*, *sustainability*, *vegetable*, *vicinity*, *waste*, and *wind farms* (Wildnerova et al., 2024).

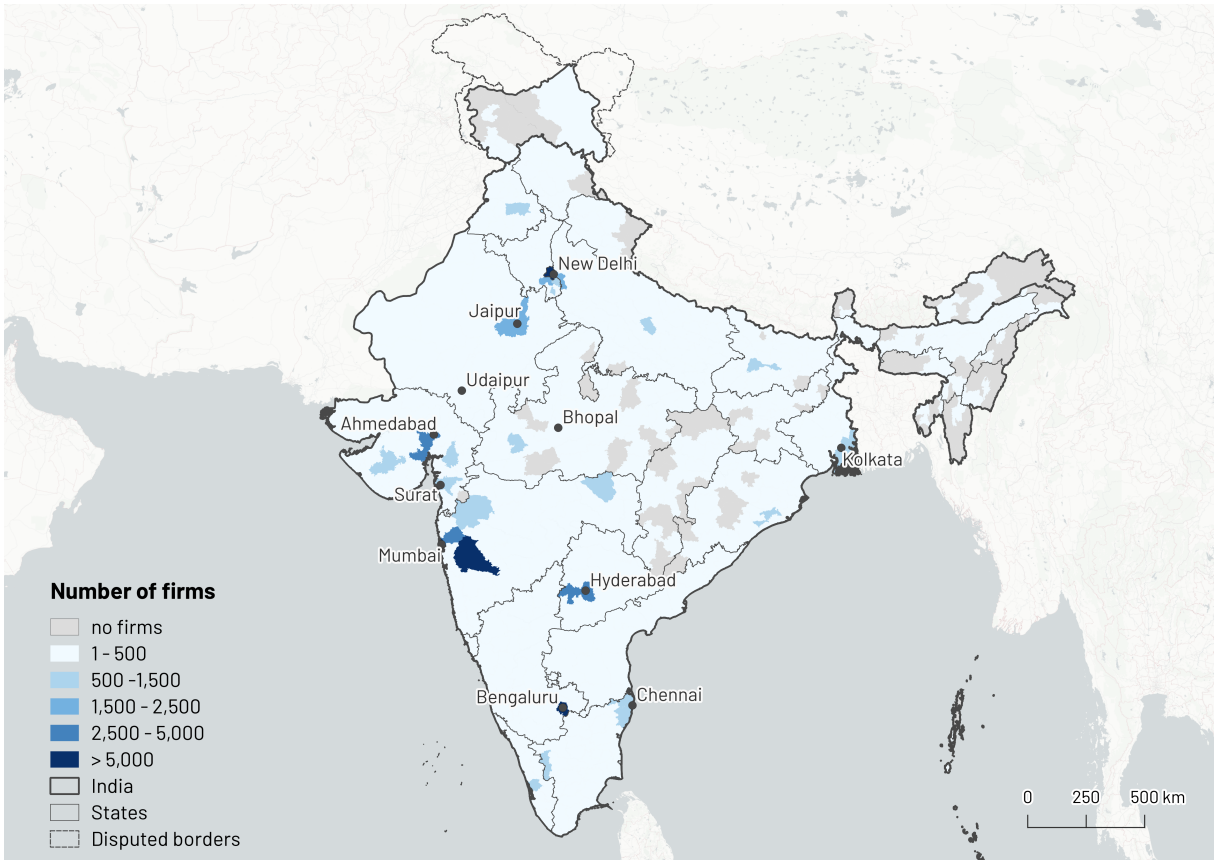


Figure 2: Distribution of firms with active website in India on district-level. The administrative areas used in all maps and analyses were acquired from GADM. For visualisation purposes only, disputed areas were added based on data from Survey of India. All maps are north-oriented.

These texts were then classified by a fine-tuned ensemble of Transformer models. The classification results on paragraph-level were then aggregated on website-level. For example, a company website could have 10 paragraphs with a sustainability-related keyword, but only 4 of them were assigned to the category *engaged* by our model. The number of identified sustainability-related paragraphs was then normalised by the website length in order to generate a comparable score on the importance of ecological sustainability on the respective website. This metric was based on the assumption that companies integrating sustainability into their core business are more likely to communicate about it proportionally more than less engaged firms.

Figure 2 shows the spatial distribution of firms in our dataset, aggregated on district-level. It clearly mirrors the distribution of population density in India with centres of high firm density in the large cities such as New Delhi, Kolkata, Bangalore, or Mumbai.

Consequently, the spatial distribution of firms in our dataset was quite uneven. The distribution of sustainability-engaged firms followed the same overall pattern. The only major difference was the much higher amount of districts without a single sustainability-engaged company.

3.3. Complementary data

In addition to our web-based indicators, we used openly accessible EO data sources to quantify environmental quality at firm locations. We decided to focus on data products derived from remote sensing data or spatial modelling instead of singular, ground-based measuring stations, due to their continuous nature and large spatial coverage.

To estimate the level of air pollution at a firm location, we used the GLOB-ANT dataset from the CAMS Global Emission Inventory, which is provided through European Space Agency (ESA)'s Copernicus programme. The data are not based on satellite measurements, such as those from the TROPOMI sensor, but are model-based estimates derived from activity data (e.g., fuel use, industrial output) combined with emission factors, compiled using national statistics and global datasets following a bottom-up methodology (Soulie et al., 2024). CAMS provides information on various air pollutants aggregated on a monthly basis with a spatial resolution of 0.1° (∼10 km). We calculated the median value over the year 2024 for each pixel and the following air pollutants: methane (CH_4), carbon dioxide (CO_2), nitrogen oxides (NO_x), carbon monoxide (CO), and sulfur dioxide (SO_2). Consequently, the dataset did not include any information on $PM_{2.5}$ concentrations, only on its precursors. Based on these information layers, we created a composite *air pollution score* (APS) by normalising the respective median air pollutant concentrations, summing them for each pixel and calculating the mean (cf. Equation 1). While the air pollutants under consideration have differing sources, they were generally quite highly correlated, with the smallest correlation being 0.94 between regional SO_2 and CH_4 concentrations. A similar air pollution score was proposed by Hasani and Ahamad (2026).

$$APS(x,y) = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} N_p(x,y), \qquad \mathcal{P} = \{\mathrm{CH_4}, \mathrm{CO_2}, \mathrm{NO_x}, \mathrm{CO}, \mathrm{SO_2}\} \tag{1}$$

where $N_p(x,y)$ is the min-max normalised median concentration of each pollutant at pixel (x,y).

Figure 3 shows the spatial distribution of our air pollution indicator across India, with darker shades highlighting regions of elevated pollutant concentrations. Two regional centres of high air pollution can be identified: The north of India, at the foot of the Himalayas between the metropolises of New Delhi and Kolkata, and the south. Regions close to other large cities, such as Ahmedabad and Mumbai, also stand out. The least air pollution was found in the border areas of Gujarat and Rajasthan with Pakistan and some rather sparsely populated, mountain regions such as Ladakh.

As a proxy for greenness and vegetational health, we used the Normalised Difference Vegetation Index (NDVI) which is calculated as

$$NDVI = \frac{NIR - R}{NIR + R} \tag{2}$$

where NIR is the near-infrared and R is the red band of the input imagery (Pettorelli et al., 2005). For the year 2024, we accessed monthly data from the MODIS satellite, which has a spatial resolution of 250 m, through Google Earth Engine (GEE) and calculated mean scores for each pixel. Figure 4 demonstrates the distribution of yearly median NDVI

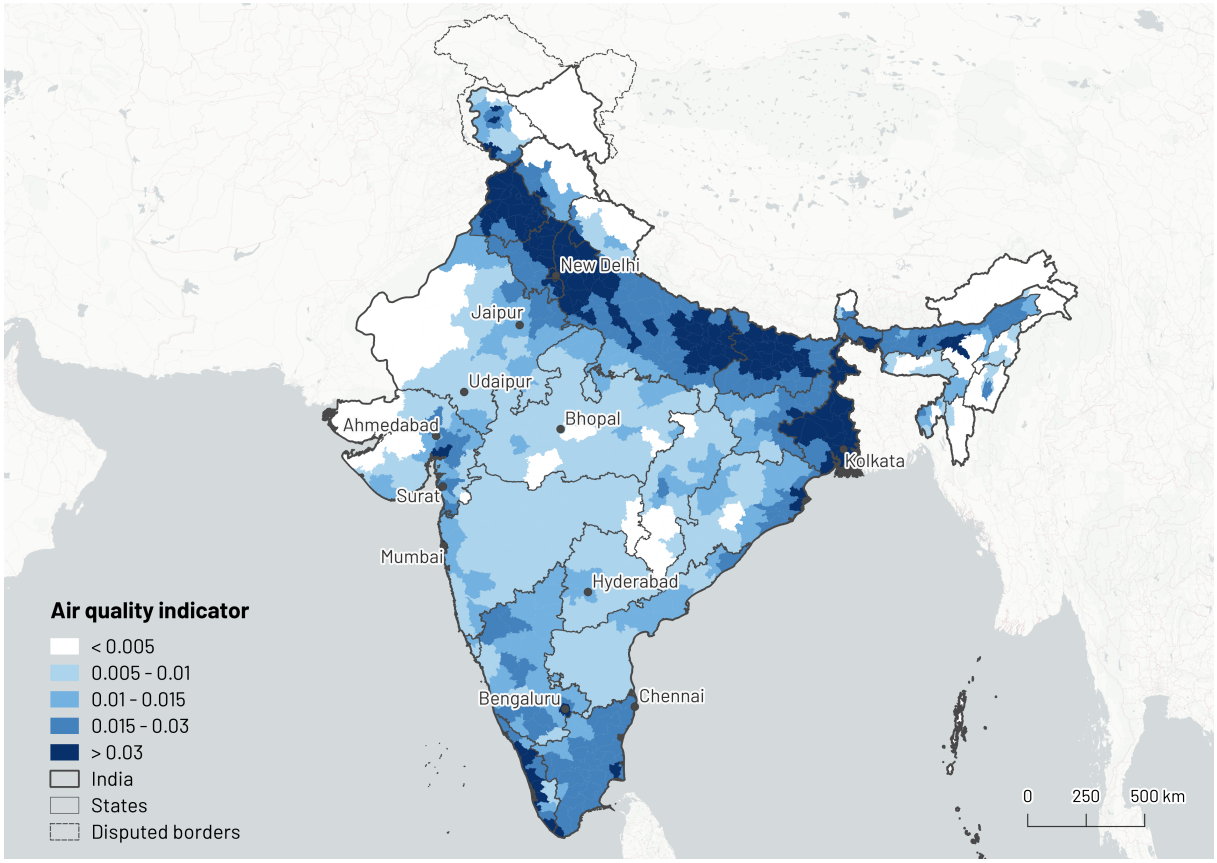


Figure 3: Air pollution indicator on district-level. The class distribution is inspired by quantiles.

values, i.e. an approximation of regional vegetational health. The higher the values, the more healthy vegetation can be found. Regions in the very east and coastal south-west of India had particularly high NDVI values, while especially the Western, rather arid parts of the country had comparatively low values.

Based on the coordinates derived from geocoding the postal addresses from the ORBIS database, we were able to perform a spatial join between the firm locations and the respective values from the continuous raster data sets on air pollution, vegetation, and population density. For the latter, we accessed data from 2020 with a spatial resolution of 30 arc seconds from CIESIN (2017). In addition to the URL and postal addresses, we also extracted information from the ORBIS database on the industrial sector of each firm, which we used to construct the district-level share of industrial firms (NACE sections from *B-F*) described below.

We conducted our analyses on regional level, i.e., districts (Kim et al., 2016; Jain et al., 2025). For this, we used the QGIS tool "Zonal statistics" to calculate mean and median values from each EO dataset for administrative areas. We additionally counted the number of firms, as well as the amount of sustainability-engaged firms, per region. To account for structural differences between the districts, we also used socio-economic data as controls. The data was acquired from the IPI dataset (Jain et al., 2025) and contained information on access to electricity, high school matriculation, and internet usage. Based on averaging the normalised values, we created the variable *institutional_ quality*, assuming that it would capture complementary dimensions of a district's level of socio-economic and institutional development.

For the regional analysis, we discarded agricultural firms (NACE section *A*) and

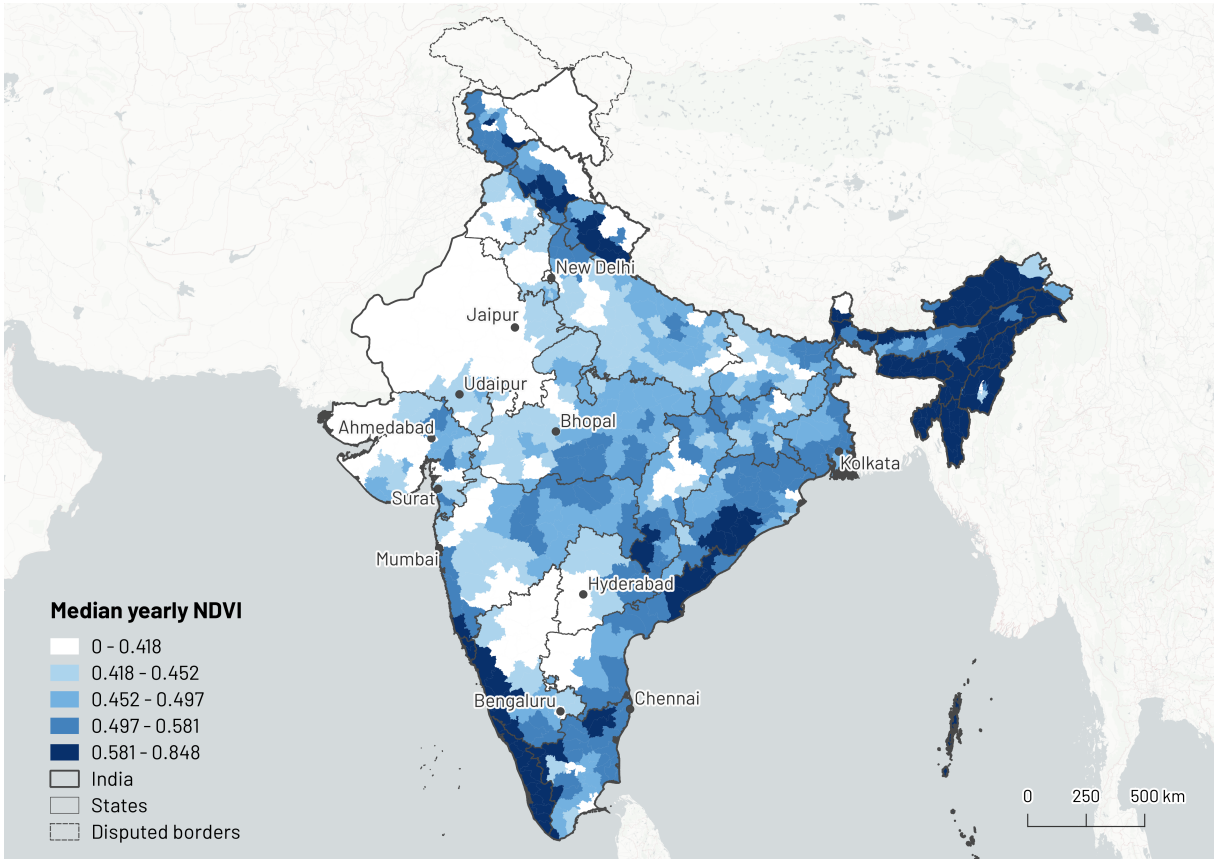


Figure 4: Median yearly NDVI on district-level. The class distribution is inspired by quantiles.

grouped the remaining NACE codes from the ORBIS database into industry (all NACE sections from *B-F*) and non-industry (all others), calculating the share of industrial companies per district (*industry_rel*).

3.4. Regression analysis

To investigate the relationship between corporate sustainability engagement and environmental conditions, we ran a district-level regression analysis on our spatially aggregated dataset. Table 1 gives an overview of all variables used in our models.

For our preliminary Ordinary Least Squares (OLS) models, we conducted a spatial autocorrelation analysis using the Global Moran's I metric (cf. Equation 3) to investigate the spatial characteristics of model residuals. Since the residuals of our preliminary OLS models showed significant spatial autocorrelation, we also calculated Spatial Lag Models (SLMs) to account for this inherent spatiality. SLMs incorporate the dependent variable values of neighbouring observations as an additional explanatory variable and are calculated according to Equation 4. To assess the suitability of a SLM, Lagrange Multiplier (LM) diagnostics were applied. The LM-Lag test examines the presence of spatial dependence in the dependent variable, whereas the LM-Error test evaluates spatial autocorrelation in the error term (Anselin, 1988).

For the conceptualisation of the spatial neighbourhood W in the SLMs, we employed a Queen contiguity approach which only considers directly spatially adjacent observations as neighbours. Following common practices in spatial modelling, we row-standardised our spatial weights matrix (Sutton and Sutton, 2024).

$$I = \frac{N}{W} \cdot \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{3}$$

where N represents the number of observations, W the sum of all weights, x_i and x_j are values of the variable at locations i and j, $\bar{x}$ is the mean of the values and w_{ij} is the spatial weight between locations i and j (Ord and Getis, 1995).

$$y = \rho W y + X\beta + \epsilon \tag{4}$$

where y is the dependent variable vector (size $n \times 1$), ρ is the spatial autoregressive parameter (scalar), W is the spatial weight matrix (size $n \times n$), X is the matrix of independent variables (size $n \times k$), β is the vector of coefficients (size $k \times 1$), and ϵ is the vector of error terms (size $n \times 1$) (Lam and Souza, 2020).

4. Results

In the following, we give an overview of our dataset and describe the results of our district-level regression analysis.

4.1. Descriptive statistics

Through our web-based analysis, we classified 46,468 firms as sustainability-engaged, i.e. their web-based indicator was > 0 (Schmidt et al., 2022). This corresponded to 39.5% of the 117,622 geocoded firms described in Section 3.1. As in previous research with web-based economic indicators (Kinne et al., 2024), the distribution of intensity values was highly right-skewed. These firm-level classifications were then aggregated to the district level to construct our dependent variables *n_sust_firms* (absolute number) and *n_sust_firms_relative* (relative number). Table 1 provides an overview of the value distribution of our district-level variables.

Table 1: Descriptive statistics of regression variables.

Variable	Definition	Source	Count	Mean	Min	Median	Max
n_sust_firms	Number of sustainability-engaged firms	WebAI	611	76.052	0.000	5.000	7,546.000
n_sust_firms_relative	Proportion of sustainability-engaged firms	WebAI	611	0.378	0.000	0.378	1.000
air_quality_indicator_log	Air quality indicator	CAMS	611	0.024	0.000	0.014	0.689
ndvi_year_mean	Mean yearly NDVI	MODIS	611	0.504	0.018	0.486	0.805
population_density_mean_log	Mean number of people per km^2	GPWv4	611	6.108	1.420	6.117	10.574
industry_rel	Proportion of industry firms	ORBIS	589	0.360	0.000	0.333	1.000
institutional_quality	Average indicator based on normalised access to electricity, internet usage, and high school matriculation	IPI	589	0.547	0.059	0.520	0.967

On a regional level (see Figure A.6), our analysis showed a strong positive bivariate correlation (r=0.592***) between the air pollution indicator and population density, and a positive correlation (r=0.583***) with the number of sustainability-engaged firms per district. There was a similar correlation (r=0.568***) between air quality and the overall firm number. In contrast, air quality was weakly negatively correlated with the median

NDVI (r=-0.177***). We also found a slightly negative correlation between NDVI and the number of sustainability-engaged firms (r=-0.172***). Institutional quality was moderately positively correlated with the absolute number of sustainability-engaged firms (r=0.250***) and to a lesser degree with the relative number (r=0.056). It should be stated that many regions only contained very few companies. Consequently, we ran all of our regression also for regions with > 10 firms as a robustness check.

4.2. Regression analysis

The results of five district-level regression models are presented in Tables 2 (dependent variable: *n_sust_firms_log*) and 3 (dependent variable: *n_sust_firms_relative*). The model specifications were estimated sequentially by gradually adding explanatory variables. The logarithmised air pollution indicator (*air_pollution_log*) served as explanatory variable in all specifications. The mean annual NDVI (*ndvi_year_mean*), the logarithmised mean population density (*population_density_mean_log*), the relative share of industrial enterprises (*industry_rel*), and the composite indicator for institutional quality (*institutional_quality*) were added successively. As the distribution of many variables was heavily skewed, we used logarithmised versions in the regressions, if the skewness was > 1 or the kurtosis > 3. For each specification, both an OLS model and a SLM were estimated in order to account for potential spatial dependence between neighbouring districts. As a robustness check, all specifications were also calculated only considering districts with > 10 firms to reduce the influence of sparsely populated areas. The results can be found in Table A.4 and A.5 in the appendix. We dropped regions with missing data for relevant variables for these specifications.

Across all model specifications, the air pollution indicator displayed a positive and mostly statistically significant association with the logarithmised number of sustainability-engaged firms. The magnitude of the coefficient declined substantially as additional explanatory variables were introduced. In the final specification, the coefficient was no longer statistically significant in either the OLS or the SLM, suggesting that the relationship between air pollution and the number of sustainability-engaged firms was largely absorbed by demographic, industrial, and institutional controls.

The mean NDVI was generally negatively associated with the number of sustainability-engaged firms and statistically significant in several models. In the fully specified Model 5, the relationship became strongly negative and highly significant in both the OLS and SLM estimations, indicating that districts with lower vegetation health tended to contain more sustainability-engaged firms.

Population density showed a consistently positive and highly significant relationship with the dependent variable across all specifications in which it was included. This finding suggests that sustainability-engaged firms were more concentrated in densely populated districts. The coefficient remained relatively stable across both OLS and spatial models. The relative share of industrial enterprises was positively associated with the number of sustainability-engaged firms. The coefficient was highly significant in all specifications, indicating that districts with a stronger industrial base tended to host more sustainability-engaged firms. Institutional quality, included only in Model 5, also exhibited a strong positive and highly significant association with the dependent variable. Districts characterised by higher institutional quality therefore appeared to contain substantially more sustainability-engaged firms.

Table 2: Regression results at the region level. The dependent variable is the logarithmised absolute number of sustainability-engaged firms per district.

	Model 1		Model 2		Model 3		Model 4		Model 5	
	OLS	SLM	OLS	SLM	OLS	SLM	OLS	SLM	OLS	SLM
CONSTANT	1.760***	0.682***	2.756***	1.161***	-0.62	-1.394***	-0.995*	-1.729***	-3.216***	-3.367***
APS_log	13.493***	8.085***	12.766***	7.830***	5.704***	2.584**	5.982***	2.631*	0.73	-0.42
W_n_sust_firms_log	-	0.567***	-	0.558***	-	0.526***	-	0.503***	-	0.332***
ndvi_year_mean	-	-	-1.942***	-0.902*	-1.640***	-0.73	-1.357**	-0.66	-3.301***	-2.463***
population_density_mean_log	-	-	-	-	0.556***	0.436***	0.546***	0.463***	0.603***	0.550***
industry_rel	-	-	-	-	-	-	0.877***	0.528**	0.953***	0.727***
institutional_quality	-	-	-	-	-	-	-	-	5.395***	4.378***
R^2	0.14	-	0.16	-	0.24	-	0.25	-	0.48	-
Adj. R^2	0.14	-	0.16	-	0.24	-	0.24	-	0.47	-
Pseudo R^2	-	0.02	-	0.02	-	0.05	-	0.05	-	0.13
AIC	2,260.28	2,106.38	2,252.25	2,105.44	2,188.70	2,053.94	2,106.82	1,989.69	1,832.96	1,781.82
Log-Lik	-1,128.14	-1,050.19	-1,123.12	-1,048.72	-1,090.35	-1,021.97	-1,048.41	-988.84	-910.48	-883.91
RMSE	1.53	1.30	1.52	1.30	1.44	1.25	1.44	1.26	1.20	1.13
F-test	103.43***	-	57.52***	-	65.53***	-	48.44***	-	102.72***	-
LM-Lag	183.08***	-	174.62***	-	174.86***	-	149.37***	-	62.70***	-
LM-Error	171.92***	-	163.62***	-	242.39***	-	194.71***	-	133.22***	-
VIF (max)	-	-	1.21	1.21	18.16	18.16	21.83	21.83	24.1	24.1
Multicollinearity	1.6	-	11.0	-	19.8	-	22.7	-	27.0	-
Moran's I	0.345***	-0.03	0.336***	-0.03	0.409***	0.037*	0.377***	0.03	0.319***	0.102***
Obs.	611	611	611	611	611	611	589	589	570	570

Dependent variable: *n_sust_firms_log*, * p<0.10, ** p<0.05, *** p<0.01.

Table 3: Regression results at the region level. The dependent variable is the relative number of sustainability-engaged firms per district.

	Model 1		Model 2		Model 3		Model 4		Model 5	
	OLS	SLM	OLS	SLM	OLS	SLM	OLS	SLM	OLS	SLM
CONSTANT	0.381***	0.348***	0.401***	0.365***	0.508***	0.469***	0.442***	0.394***	0.426***	0.397***
APS_log	-0.13	-0.13	-0.15	-0.14	0.07	0.07	0.21	0.20	0.15	0.14
W_n_sust_firms_relative	-	0.09	-	0.09	-	0.08	-	0.09	-	0.06
ndvi_year_mean	-	-	-0.04	-0.03	-0.05	-0.04	-0.03	-0.01	-0.06	-0.05
population_density_mean_log	-	-	-	-	-0.018*	-0.02	-0.022**	-0.021**	-0.025**	-0.024**
industry_rel	-	-	-	-	-	-	0.213***	0.208***	0.200***	0.196***
institutional_quality	-	-	-	-	-	-	-	-	0.103*	0.095*
R^2	0.00	-	0.00	-	0.01	-	0.07	-	0.07	-
Adj. R^2	-0.00	-	-0.00	-	0.00	-	0.07	-	0.07	-
Pseudo R^2	-	-0.01	-	-0.01	-	-0.03	-	-0.23	-	-0.22
AIC	-99.37	-99.53	-97.56	-97.66	-98.45	-98.29	-215.96	-216.25	-212.94	-211.94
Log-Lik	51.69	52.77	51.78	52.83	53.22	54.14	112.98	114.13	112.47	112.97
RMSE	0.22	0.22	0.22	0.22	0.22	0.22	0.20	0.20	0.20	0.20
F-test	0.48	-	0.33	-	1.18	-	11.85***	-	9.20***	-
LM-Lag	2.25	-	2.18	-	1.91	-	2.19	-	0.97	-
LM-Error	1.87	-	1.82	-	1.51	-	0.88	-	0.48	-
VIF (max)	-	-	1.21	1.21	18.16	18.16	21.83	21.83	24.1	24.1
Multicollinearity	1.6	-	11.0	-	19.8	-	22.7	-	27.0	-
Moran's I	0.036*	-0.01	0.036*	-0.01	0.032*	-0.01	0.03	-0.01	0.02	-0.01
Obs.	611	611	611	611	611	611	589	589	570	570

Dependent variable: *n_sust_firms_relative*, * p<0.10, ** p<0.05, *** p<0.01.

The spatial lag coefficient *W_n_sust_firms_log*, which represents the number of sustainability-engaged firms in neighbouring districts, was positive and highly significant across all SLM specifications. The coefficient declined with added controls, suggesting that part of the observed spatial dependence was explained by the inclusion of other explanatory variables, although significant spillover effects remained present.

Across all specifications, the SLMs consistently achieved lower Akaike information criterion (AIC) and Root Mean Square Error (RMSE) values than the corresponding OLS models, indicating improved model fit when spatial dependence was taken into account. The explanatory power of the OLS models increased considerably as additional controls were introduced, with the R^2 rising from 0.14 in Model 1 to 0.48 in Model 5. At the same time, the inclusion of further covariates increased multicollinearity, reflected in maximum Variance Inflation Factor (VIF) values rising to 24.1 in the most complex specifications. The Global Moran's I statistic indicated significant spatial autocorrelation in OLS residuals. In the corresponding SLMs, Moran's I values were strongly reduced, although some residual spatial dependence remained in the more complex specifications. This indicates that the spatial regression models captured a large share, but not all, of the spatial dependence present in the data.

For the models that discarded districts with < 10 firms, the results were overall quite similar, indicating a general robustness of our models. The air pollution indicator lost its significance with more control variables and, in contrast to the previous setup, even changed its sign in Model 5. The mean NDVI values remained negative across all specifications. However, statistical significance was observed primarily in the more comprehensive models, most notably in the fully specified Model 5, where the coefficient was strongly negative and highly significant in both estimation approaches. The positive and highly significant spatial lag coefficient persisted across all SLM specifications, indicating that spatial spillover effects remained robust after excluding districts with fewer than ten firms. The control variables and general evaluation statistics followed the same pattern as in the full sample, with SLM specifications consistently achieving lower AIC and RMSE values than the corresponding OLS models and explanatory power increasing substantially as additional controls were introduced.

When considering the relative number of sustainability-engaged firms per district as the dependent variable, the results differed substantially from those based on absolute firm counts. Overall, the explanatory power of the models was considerably weaker, and the environmental indicators lost statistical significance once the dependent variable was normalised by the total number of firms. Across all model specifications, the air pollution indicator showed no statistically significant relationship with the relative number of sustainability-engaged firms in either the OLS or the SLM estimations. Moreover, the sign of the coefficient changed across specifications, ranging from slightly negative in the simpler models to positive in later models, indicating no stable association between regional air pollution levels and the relative concentration of sustainability-engaged firms. Similarly, the mean NDVI consistently exhibited negative coefficients across all specifications, but these effects were never statistically significant. In contrast, several control variables displayed more robust relationships. Population density showed a small negative association with the dependent variable and became statistically significant in the more comprehensive model specifications. This suggests that although densely populated districts contained more sustainability-engaged firms in absolute terms, their relative share tended to be lower. The share of industrial firms exhibited a consistently positive and highly significant association in both the OLS and SLM specifications of Models 4 and 5.

Likewise, institutional quality was positively associated with the dependent variable and weakly statistically significant in both estimation approaches. The spatial lag coefficient was positive across all SLM specifications but never reached statistical significance. This suggests that, once the relative number of sustainability-engaged firms was considered, there was no robust evidence for spatial spillover effects between neighbouring districts.

As with the models based on absolute firm counts, the SLM specifications generally produced lower AIC values than the corresponding OLS models, with the exception of the fully specified Model 5, indicating a slightly better overall model fit. However, RMSE between the two approaches were practically identical, implying similar predictive performance. Moran's I values pointed to weak spatial autocorrelation in the OLS residuals, which was effectively mitigated in the SLM specifications. Nevertheless, the explanatory power of the models remained low overall with R^2 ranging from 0.000 to 0.07 and some adjusted R^2 even being slightly negative. Accordingly, the relative models explained substantially less variance than the models based on the absolute number of sustainability-engaged firms.

Restricting the sample to districts with > 10 firms yielded results that were broadly consistent with the full-sample specifications but with one notable difference: spatial spillover effects were positive and statistically significant across all SLM models, whereas in the full sample they were never significant. The main environmental results were largely unchanged: air pollution remained mostly insignificant, while NDVI showed unstable and non-robust effects. Population density continued to exhibit a negative and statistically significant association, and industrial structure and institutional quality remained positively related to the dependent variable. Model fit improved slightly compared to the full sample, while overall explanatory power remained moderate.

5. Discussion

In the following, we discuss and interpret our results. Additionally, we take a critical look at the data and methods used in our analysis, highlighting shortcomings and potential extensions of our work.

5.1. Discussion of results

For the district-level specifications based on the logarithmised absolute number of sustainability-engaged firms per district, our findings suggest that environmental and structural characteristics of districts were closely associated with the spatial distribution of sustainability-engaged firms. The initially strong positive association between air pollution and sustainability engagement supports the idea that firms may respond to local environmental issues. However, this relationship was not robust across all models and weakened as socio-economic controls were added. In the fully specified models, the coefficient for air pollution even became statistically insignificant, suggesting that the initial association was largely explained by demographic, industrial, and institutional characteristics rather than a direct relationship between pollution levels and sustainability engagement. For the absolute models, the negative association with vegetational health suggested that firms tended to cluster in less vegetated environments, which are often urban in nature. This was consistent with the positive association observed for population density. These patterns may reflect the concentration of sustainability-oriented business activity in more developed, industrial, and infrastructure-rich areas. This finding was broadly in line with previous studies that identify sustainability engagement as an urban topic in the US (Kinne et al., 2024), the German-speaking countries (Schmidt

et al., 2025) and China (Dagestani et al., 2024). However, it should also be noted that the relative number of companies with websites was probably significantly lower in rural regions, which lead to a bias in this evaluation.

The positive associations with industrial activity and institutional quality further support the relevance of regional capacity, e.g. through the availability of trained specialists and adequate infrastructure, in fostering sustainability engagement. However, it should be mentioned that controlling for institutional quality turned the effect of air pollution insignificant. This may suggest a close entanglement of air pollution and institutional quality in certain regions. Institutional quality showed a strong positive and statistically significant association with the number of sustainability-engaged firms, suggesting that governance quality was relevant for the regional concentration of such firms.

Finally, the consistent and significant spatial lag term, i.e. a strong spatial dependency, in the absolute model implied that sustainability engagement in one district was positively associated with similar engagement in neighbouring areas. This spatial dependency may reflect spillover effects due to shared policy frameworks, regional networks of environmental Non-Governmental Organisations (NGOs), common infrastructure, or competitive dynamics among neighbouring districts (Xie et al., 2025). However, the magnitude and significance of the spatial lag coefficient declined as additional controls were introduced, indicating that part of the spatial dependence was explained by observable regional characteristics. In the relative models, this spatial dependency was only detectable once the sample was restricted to districts with a sufficient number of firms, suggesting that the full sample was too noisy rather than that spatial structures were genuinely absent.

The regression models based on the relative number of sustainability-engaged firms per district differed considerably from their absolute counterparts. Most notably, the environmental indicators were not significant in any specification. Population density exhibited a small but significant negative effect, implying that more urbanised regions, despite hosting more sustainability-engaged firms in total, may have a lower proportion of them relative to the entire local firm population. The positive coefficients of *industry_rel* and *institutional_quality* indicated that regions with a higher industrial base and level of development tended to have a greater proportion of sustainability-oriented firms. Overall, this suggests that relative differences in sustainability engagement are only weakly explained by observable regional characteristics. While the direction of effects was partly consistent with the absolute models, the substantially lower explanatory power (with R^2 values close to zero in several specifications) suggested that proportional differences in sustainability engagement were only weakly captured by observable regional characteristics.

To check the validity of employing spatial regression models, we calculated LM tests for the OLS residuals. For the specification using *n_sust_firms_log*, both the LM-lag and LM-error tests were highly significant, indicating clear evidence of spatial dependence and motivating the use of a SLM. In contrast, when using *n_sust_firms_relative_log*, neither the LM-lag nor LM-error statistics were statistically significant across specifications. Future work could also investigate other types of spatial econometric models, including Spatial Error Models (SEMs) or Spatial Durbin Models (SDMs), to also account for spatial autocorrelation in model errors.

One aspect of our composite air pollution indicator warrants particular caution when interpreting the district-level results. The CAMS-GLOB-ANT inventory underlying our air pollution score is, as described in Section 3, a model-based estimate derived from

activity data such as fuel use and industrial output, rather than a direct satellite or ground-based measurement of ambient air quality (Soulie et al., 2024). Consequently, the strong bivariate correlation we observed between air pollution and both population density (r=0.59) and firm density (r=0.57) is, to some extent, expected by construction: all three quantities are partly derived from, or closely track, the same underlying signal of regional economic activity. Consequently, it can be assumed that the initial positive pollution association reflects the general concentration of economic activity within a district rather than a distinct environmental-response mechanism in which firms observe pollution and react to it. Put differently, what we labelled an environmental association may be better understood as a proxy relationship, since both the pollution indicator and the firm-count outcome are, in part, downstream measures of the same underlying industrial base. This distinction also matters for how the pollution variable relates to our stakeholder-based hypothesis (**H2**), which assumes that firms and their stakeholders perceive and react to perceivable environmental degradation. A district could score high on our composite indicator, for instance due to industrial CO_2 or NO_x emissions, without this being locally perceptible in the way our theoretical mechanism presumes, and vice versa. Gridded, activity-based emissions data are consequently better suited to describing the industrial footprint of a district than to proxying the perception of environmental degradation that our hypotheses are actually about. Comparing our results against a satellite-derived concentration product, such as Sentinel-5P/TROPOMI data, which several studies have used in India (Behera et al., 2022; Reshi et al., 2022; Mathew et al., 2024), would be a valuable robustness check for future work.

5.2. Implications

Our spatial models showed that absolute sustainability engagement was higher where air quality was poorer, and that engagement in one district might lead to spatial spillovers to its neighbours through significant spatial-lag effects in our absolute models, although their magnitude declined with additional controls. Indications of such patterns point to the emergence of nascent green clusters: geographically proximate firms that learn from, benchmark against and sometimes pressure each other to adopt cleaner practices (Porter and Linde, 1995; Orsato, 2006). However, this relationship was not fully stable across specifications: while air pollution was positively associated with sustainability engagement in simpler models, the effect diminished and became statistically insignificant once socio-economic, industrial, and institutional controls were included. This indicates that the initial pattern was largely driven by correlated regional characteristics rather than a direct pollution–response mechanism. In addition, for the relative models, no such stable spatial structure could be shown in the full sample.

India's recent policy developments align well with this geographic manifestation. The Perform, Achieve and Trade (PAT) scheme targets energy-intensive enterprises, while the Zero Defect Zero Effect (ZED) certification programme supports smaller manufacturers. Both encourage greener processes in the pollution-intensive districts where we observed higher absolute engagement (Bureau of Energy Efficiency, 2012; Ministry of Micro, Small & Medium Enterprises, 2016). The State Action Plan on Climate Change (SAPCC) and the Business Reforms Action Plan (BRAP) aim to raise institutional quality, which our models identified as crucial for district-level sustainability (Kumar, 2018). Finally, the mandatory BRSR framework increases transparency, providing the data infrastructure that underpins spatial spillovers and informed policy design (Joshi and Kansil, 2023). Collectively, these initiatives are well aligned with our findings that institutional capacity

and industrial structure are more strongly associated with sustainability engagement than environmental conditions alone.

At the same time, our results revealed a tension: while the number of sustainability-oriented firms grew in polluted, industrial districts, their share of the local business population did not. This may imply that no self-reinforcing, place-based momentum for coordinated sustainability transitions has taken hold. Consistent with resource-based theory, green pioneers must compete with conventional firms for mobile pools of finance, skilled labour and social capital. Only those able to bundle scarce inputs into distinctive environmental capabilities can sustain an advantage (Barney, 1991). Such resource competition tends to spur incremental greening among a few leaders rather than a systemic shift of the regional industrial base (Murillo-Luna et al., 2008; Orsato, 2006). Our findings were consistent with this interpretation, as the strongest and most stable predictors across specifications were structural factors such as industrial composition and population density, rather than environmental pressure alone.

For policy makers, this suggests that firm-level incentives should be complemented by collective instruments, such as dedicated green credit lines, shared clean-technology platforms and targeted training programmes to relax resource constraints and enable a broader set of local companies to engage meaningfully with sustainability. This implication is further supported by the strong effects that institutional quality of districts showed in our models.

5.3. Limitations

We initially also estimated firm-level regressions relating individual firms' sustainability engagement directly to the environmental conditions at their location, following the same general specification as our district-level models. We do not report these as a primary result, for two main reasons: First, our air pollution and vegetation indicators were derived from raster data with resolutions of 0.1° (∼10 km) and 250 m, respectively. Consequently, many firms located close to one another received near-identical values. This spatial clustering considerably reduces the number of effectively independent observations relative to the raw firm count, which would likely overstate the precision of firm-level standard errors computed under the usual independence assumption. Second, the bivariate correlations between our sustainability indicator and both environmental variables were close to zero at the firm level (cf. Figure A.5), and only reached statistical significance in multivariate specifications that combined many correlated controls (with VIFs up to approximately 19) and explained very little variance overall ($R^2 \leq 0.04$). Therefore, we considered our data and design better suited to characterising the district-level geography of sustainability engagement than for firm-level inference. One consequence of this choice is that our empirical test of **H2**, which was theorised as a firm-level mechanism, relied on the district-level distribution of engaged firms as a proxy.

For transparency, it is worth briefly summarising the substantive pattern of these exploratory firm-level models (see Table A.6). Across specifications, the air pollution indicator was associated with lower sustainability engagement at the firm level, the opposite sign to the positive association we found for the absolute number of engaged firms at the district level. The coefficient lost significance once broad industry-group controls were added, and only regained significance under more fine-grained sectoral controls, which points to considerable specification sensitivity rather than a stable relationship. Higher NDVI values were positively associated with engagement once basic controls were included. Firm size, proxied by revenue, was a robust positive predictor across all specifications, whereas firm age changed sign once additional controls were introduced, from

positive to negative. The clearest and most stable pattern was sectoral: manufacturing sub-sectors, such as chemicals, electrical equipment, and machinery, showed the strongest positive engagement, while service sectors such as IT, education, and healthcare were significantly less engaged. Despite this heterogeneity, the firm-level models explained very little variance overall ($R^2 \leq 0.04$).

According to data by STATISTA (2025), there were 1.6 million active companies in India in January 2024. While certainly not all of them have an active web presence, this still represents a large discrepancy with our dataset. The ORBIS database, which was the starting point of our analysis, could only provide a URL for 2.9% of all their Indian companies. Consequently, it should be expected that our sample was heavily biased towards more formalised and outward-facing firms that are perhaps more likely to engage in sustainability communication. To counter this potential bias, future studies should therefore try to capture more companies via other databases or domain searches, i.e. by automatically identifying URLs based on firm names and additional information via search engines (Dahlke et al., 2025). While other studies have shown that data on firm characteristics is already scarce in European countries (Kanilmaz et al., 2025), this issue was even more pronounced for our Indian firm population. Especially for the number of employees and revenue, there was very little available information from ORBIS, which further limited the feasibility of firm-level analyses discussed above. In our analysis, we were also unable to properly take into account the very large informal sector in India (Gururaja and Ranjitha, 2022), which is responsible for a considerable proportion of environmental pollution (Chakraborty et al., 2022).

One of the fundamental assumptions behind our study is the comprehensiveness of organisational web data (Dahlke et al., 2025), i.e. the wealth of information that can be sourced from their contents. Whether this assumption holds true for analyses in India can, however, be disputed. Many SME in India either have no dedicated website or maintain only a minimal web presence, relying instead on platform-based ecosystems (Chatterjee and Kumar Kar, 2020; Buteau, 2021). A lot of customer communication in India runs through different channels, e.g. social media such as Facebook pages, WhatsApp Business, Instagram, Google Business Profiles or online marketplaces like IndiaMART or Justdial. Consequently, company websites may not fully capture firms' sustainability activities or public communication. Future studies should thus consider extending the scope of web data under investigation.

It should also be noted that the quality of the addresses in the ORBIS database varied substantially, so that the assignment of coordinates via geocoding might be somewhat inaccurate for some companies. In addition, we were unable to geocode a large number of companies despite having their postal addresses available. This could be an indication that the quality of OSM, which was the data basis for most of our geocoding pipeline, in India is not comparable to Europe or North America. This aligns with findings that Nominatim exhibits a bias towards Europe (Serere et al., 2023). Addresses in India have also been described as unstructured, which further complicates geocoding (Singh et al., 2025). Furthermore, the address that is entered in the ORBIS database and generally also featured on the website is usually the headquarter of a company. For large companies in particular, however, this location does not necessarily have to be the same as the actual production site(s). Accordingly, the address used in our analysis might not coincide with the location where most of the potential pollution takes place.

We also acknowledge that there was a bias towards websites in English in our analysis, mainly due to the keyword-based pre-filtering step. Since India is a country with many

official languages, this was not necessarily thoroughly captured by our method. However, we decided to focus on English websites, since most Natural Language Processing (NLP) models work best for English (Søgaard, 2022).

In our analysis, we only considered aggregated values for air pollution and vegetational quality for an entire year. As Figure A.7 shows exemplarily for the NDVI, there is often temporal variation for environmental indicators. The highest NDVI values were generally in September, which corresponds to the end of the summer monsoon season (Gadgil, 2003). In contrast, there was relatively little variation in the first half of the year. We assumed that climatological variations that might affect the relationship between environmental quality and firm engagement were covered by our regional fixed effects and did therefore not include a temporal component into our analyses. Due to this overall cross-sectional, single-year design of our study, we were only able to identify associations rather than causal relationships. The latter would require longitudinal observations, additional temporal information, and appropriate causal modelling approaches.

In future studies, environmental quality could be described even more extensively, e.g. by including data on water and soil pollution. Further air pollutants not included in the CAMS-GLOB-ANT dataset, especially $PM_{2.5}$, or data on concrete air pollution sources (e.g. vehicle emissions) could also be considered. However, ensuring consistent availability, adequate spatial resolution, and quality of such data could be a challenge. Likewise, the inclusion of more socio-demographic or socio-economic data would also be desirable. Unfortunately, there is a lack of such openly available data in India. We decided against using data from the Indian census, since the last available dataset was from 2011. Other datasets containing relevant information (e.g. about the Gross Domestic Product (GDP)) were also outdated or only available for selected districts.

Previous web-based research by Kinne et al. (2024) and Schmidt et al. (2025) has shown the importance of network aspects for corporate online sustainability engagement. Following the approach proposed by Abbasiharofteh et al. (2023), hyperlinks between websites can be used to model inter-firm relationships. Investigating the impact of collaboration and knowledge spill-over in a sustainability context in India could be another interesting angle to consider. Due to the already limited size of our dataset, we, however, decided to omit this relationality dimension (Dahlke et al., 2025).

One issue that any web-based study on environmental sustainability encounters is greenwashing, i.e. the incorrect representation of environmental engagement by a firm (Schmidt et al., 2022). While we indirectly addressed this issue in our work by incorporating objectively measured environmental data, there might have still been a strong bias in our web-based indicators. The inclusion of quantitative data at company level, e.g. through ESG reporting, would lend itself to this. However, as such data does not exist for the majority of companies in India (Asokan, 2023; Mishra et al., 2024), we had no alternative but to interpret our results against the backdrop of an existing bias, i.e. an overrepresentation of sustainability engagement.

6. Conclusion

This study investigated whether there is a relationship between the sustainability engagement of Indian firms and the environmental quality of the districts in which they are located. For this, we employed a novel large-scale approach based on the WebAI paradigm (Dahlke et al., 2025), increasing the scale of previous environmental economic research in India.

We found a strong, statistically significant bivariate correlation between air quality and firm density on district-level, showing a clear positive relationship (**RQ1**) consistent with **H1**. However, this finding should be interpreted cautiously, as the air pollution indicator was itself constructed from economic activity data. Consequently, the observed relationship likely reflects that both firm density and modelled emissions are proxies for the same underlying concentration of economic activity, rather than providing evidence of an independent association. At the district level, air quality and vegetation were initially associated with the absolute number of sustainability-engaged firms, but these associations weakened substantially, and in the case of air pollution disappeared entirely, once population density, industrial structure, and institutional quality were included as controls. When the relative share of sustainability-engaged firms was considered instead, neither air pollution nor vegetation showed a significant relationship in any full-sample specification. Indicators of population density, industrial activity, and institutional quality proved to be the more consistent predictors throughout, suggesting that sustainability engagement is concentrated in economically developed districts rather than driven by local environmental pressure. The spatially lagged dependent variable showed evidence of spatial spillovers for the absolute number of sustainability-engaged firms throughout, and for the relative number once the sample was restricted to districts with a sufficient number of firms.

Taken together, we did not find robust support for **H2**: the geography of sustainability engagement in India appears to be shaped primarily by structural and institutional regional characteristics rather than by the environmental conditions firms are exposed to (**RQ2**). We consider this a substantively informative null result rather than a shortcoming of the analysis. It was obtained using a national-scale dataset that is the first of its kind for India, and the pattern held up consistently across absolute and relative specifications, OLS and spatial models, and a robustness check restricted to districts with a sufficient number of firms. It thus indicates that, at least at the scale we can observe with current data, the environmental-pressure channel implied by stakeholder theory is not what drives where sustainability-engaged firms locate in India.

The main limitation of our study was scarce data availability, particularly for URLs of corporate websites and firm-level control variables. Consequently, a key takeaway from our study is that classic data sources for econometric analyses, such as those generally used for studies in Europe or North America, can only be employed to a limited extent in other regions of the world. For future web-based economic research in India, it is therefore crucial that such data sets are created or made freely available.

Acknowledgements

We would like to thank Sacha den Nijs (TNO), Norman Metzinger (ISTARI.AI), Emilio Illán Fernández (University of Murcia), Martin Sudmanns, and Lorena Abad Crespo (University of Salzburg) for their helpful comments.

This paper received funding through the Austrian Research Promotion Agency (FFG) and the DigitalInnovationLayer project (FO999898902).

Data and code availability

Data and code can be found on GitHub.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used GPT-5 and Claude Sonnet 5 in order to identify spelling mistakes and improve grammar. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

References

Abbasiharofteh, M., Krüger, M., Kinne, J., Lenz, D., Resch, B., 2023. The digital layer: Alternative data for regional and innovation studies. Spatial Economic Analysis 18, 507–529. doi:10.1080/17421772.2023.2193222.

Albrizio, S., Botta, E., Kozluk, T., Zipperer, V., 2014. Do Environmental Policies Matter for Productivity Growth?: Insights from New Cross-Country Measures of Environmental Policies. OECD Economics Department Working Papers 1176. OECD. doi:10.1787/5jxrjncjrcxp-en.

Anselin, L., 1988. Lagrange Multiplier Test Diagnostics for Spatial Dependence and Spatial Heterogeneity. Geographical Analysis 20, 1–17. doi:10.1111/j.1538-4632.1988.tb00159.x.

Asokan, A., 2023. An overview of ESG reporting in India: Practices and challenges, in: Shrivastava, A., Bhusan, A. (Eds.), Sustainable Boardrooms. Springer Nature Singapore, Singapore, pp. 19–39. doi:10.1007/978-981-99-4837-6_2.

Auzepy, A., Tönjes, E., Lenz, D., Funk, C., 2023. Evaluating TCFD reporting—a new application of zero-shot analysis to climate-related financial disclosures. PLOS ONE 18, e0288052. doi:10.1371/journal.pone.0288052.

Azimi, M.N., Rahman, M.M., 2024. Unveiling the health consequences of air pollution in the world's most polluted nations. Scientific Reports 14, 9856. doi:10.1038/s41598-024-60786-0.

Barney, J.B., 1991. Firm resources and sustained competitive advantage. Journal of Management 17, 99–120. doi:10.1177/014920639101700108.

Behera, M.D., Mudi, S., Shome, P., Das, P.K., Kumar, S., Joshi, A., Rathore, A., Deep, A., Kumar, A., Sanwariya, C., Kumar, N., Chandrakar, R., Seshadri, S., Mukherjee, S., Bhattaram, S.K., Sirivella, Z., 2022. COVID-19 slowdown induced improvement in air quality in India: Rapid assessment using Sentinel-5P TROPOMI data. Geocarto International 37, 8127–8147. doi:10.1080/10106049.2021.1993351.

Beise, M., Rennings, K., 2005. Lead markets and regulation: A framework for analyzing the international diffusion of environmental innovations. Ecological Economics 52, 5–17. doi:10.1016/j.ecolecon.2004.06.007.

Boutabba, M.A., 2014. The impact of financial development, income, energy and trade on carbon emissions: Evidence from the Indian economy. Economic Modelling 40, 33–41. doi:10.1016/j.econmod.2014.03.005.

Bracht, F., Verhoeven, D., 2025. Air pollution and innovation. Journal of Environmental Economics and Management 130, 103102. doi:10.1016/j.jeem.2024.103102.

Bureau of Energy Efficiency, 2012. Perform, Achieve and Trade (PAT) scheme. URL: https://beeindia.gov.in/en/programmes/perform-achieve-and-trade-pat. accessed: 2025-07-04.

Bureau van Dijk, 2022. ORBIS - Data. URL: https://orbis.bvdinfo.com/.

Buteau, S., 2021. Roadmap for digital technology to foster India's MSME ecosystem—opportunities and challenges. CSI Transactions on ICT 9, 233–244. doi:10.1007/s40012-021-00345-4.

Buysse, K., Verbeke, A., 2003. Proactive environmental strategies: A stakeholder management perspective. Strategic Management Journal 24, 453–470. doi:10.1002/smj.299.

Chakraborty, J., Basu, P., 2021. Air quality and environmental injustice in India: Connecting particulate pollution to social disadvantages. International Journal of Environmental Research and Public Health 18, 304. doi:10.3390/ijerph18010304.

Chakraborty, P., Chandra, S., Dimmen, M.V., Hurley, R., Mohanty, S., Bharat, G.K., Steindal, E.H., Olsen, M., Nizzetto, L., 2022. Interlinkage Between Persistent Organic Pollutants and Plastic in the Waste Management System of India: An Overview. Bulletin of Environmental Contamination and Toxicology 109, 927–936. doi:10.1007/s00128-022-03466-x.

Chatterjee, S., Kumar Kar, A., 2020. Why do small and medium enterprises use social media marketing and what is the impact: Empirical insights from India. International Journal of Information Management 53, 102103. doi:10.1016/j.ijinfomgt.2020.102103.

Chen, F., Wang, M., Pu, Z., 2022. The impact of technological innovation on air pollution: Firm-level evidence from China. Technological Forecasting and Social Change 177, 121521. doi:10.1016/j.techfore.2022.121521.

CIESIN, 2017. Gridded population of the world, version 4 (GPWv4): Population density, revision 11. doi:10.7927/H49C6VHW.

Connelly, B.L., Certo, S.T., Reutzel, C.R., DesJardine, M.R., Zhou, Y.S., 2025. Signaling theory: State of the theory and its future. Journal of Management 51, 24–61. doi:10.1177/01492063241268459.

Dagestani, A.A., Chen, P., Du, L., Hu, J., Bilan, Y., 2024. The impacts of urban development orientation of resource-based cities on environmental information disclosure and greenwashing behavior of listed firms in China. Environment, Development and Sustainability 27, 21515–21540. doi:10.1007/s10668-024-04551-3.

Dahlke, J., Beck, M., Kinne, J., Lenz, D., Dehghan, R., Wörter, M., Ebersberger, B., 2024. Epidemic effects in the diffusion of emerging digital technologies: Evidence from artificial intelligence adoption. Research Policy 53, 104917. doi:10.1016/j.respol.2023.104917.

Dahlke, J., Schmidt, S., Lenz, D., Kinne, J., Dehghan, R., Abbasiharofteh, M., Schütz, M., Kriesch, L., Hottenrott, H., Kanilmaz, U.N., Grashof, N., Hajikhani, A., Liu, L., Riccaboni, M., Balland, P.A., Wörter, M., Rammer, C., 2025. The WebAI Paradigm of Innovation Research: Extracting Insight from Organizational Web Data Through AI. ZEW Discussion Paper .

Dehghan, R., Grashof, N., Schmidt, S., Kopka, A., Woywode, M., 2026. Thinking machines, shrinking footprints? The impact of Artificial Intelligence on energy consumption in regions. Energy Economics 161, 109526. doi:10.1016/j.eneco.2026.109526.

deSouza, P.N., Chaudhary, E., Dey, S., Ko, S., Németh, J., Guttikunda, S., Chowdhury, S., Kinney, P., Subramanian, S.V., Bell, M.L., Kim, R., 2023. An environmental justice analysis of air pollution in India. Scientific Reports 13, 16690. doi:10.1038/s41598-023-43628-3.

deSouza, P.N., Dey, S., Mwenda, K.M., Kim, R., Subramanian, S., Kinney, P.L., 2022. Robust relationship between ambient air pollution and infant mortality in India. Science of The Total Environment 815, 152755. doi:10.1016/j.scitotenv.2021.152755.

Fatima, N., Yanting, Z., Guohua, N., Khan, M.K., 2024. The dynamics of green technological innovation and environmental policy stringency for sustainable environment in BRICS economies. Natural Resources Forum , 1477–8947.12563doi:10.1111/1477-8947.12563.

Freeman, R.E., 1984. Strategic Management: A Stakeholder Approach. Pitman, Boston.

Gadgil, S., 2003. The Indian monsoon and its variability. Annual Review of Earth and Planetary Sciences 31, 429–467. doi:10.1146/annurev.earth.31.100901.141251.

González-Benito, J., González-Benito, Ó., 2008. A study of determinant factors of stakeholder environmental pressure perceived by industrial companies. Business Strategy and the Environment 17, 500–511. doi:10.1002/bse.612.

Gschnaidtner, C., Dehghan, R., Hottenrott, H., Schwierzy, J., 2024. Adoption and diffusion of blockchain technology. SSRN Electronic Journal doi:10.2139/ssrn.4775993.

Gururaja, B.L., Ranjitha, N., 2022. Socio-economic impact of COVID-19 on the informal sector in India. Contemporary Social Science 17, 173–190. doi:10.1080/21582041.2021.1975809.

Guttikunda, S.K., Goel, R., Pant, P., 2014. Nature of air pollution, emission sources, and management in the Indian cities. Atmospheric Environment 95, 501–510. doi:10.1016/j.atmosenv.2014.07.006.

Hasani, N.U., Ahamad, F., 2026. Machine learning based ozone sub-index estimation and composite AQI forecasting under the CPCB NAQI framework using Sentinel-5P. Discover Environment 4, 275. doi:10.1007/s44274-026-00782-5.

Hicks, J.R., 1932. The Theory of Wages. Palgrave Macmillan UK, London. doi:10.1007/978-1-349-00189-7.

Hojnik, J., Ruzzier, M., 2016. What drives eco-innovation? A review of an emerging literature. Environmental Innovation and Societal Transitions 19, 31–41. doi:10.1016/j.eist.2015.09.006.

Holloway, T., Miller, D., Anenberg, S., Diao, M., Duncan, B., Fiore, A.M., Henze, D.K., Hess, J., Kinney, P.L., Liu, Y., Neu, J.L., O'Neill, S.M., Odman, M.T., Pierce, R.B., Russell, A.G., Tong, D., West, J.J., Zondlo, M.A., 2021. Satellite Monitoring for Air Quality and Health. Annual Review of Biomedical Data Science 4, 417–447. doi:10.1146/annurev-biodatasci-110920-093120.

Horbach, J., 2008. Determinants of environmental innovation—new evidence from German panel data sources. Research Policy 37, 163–173. doi:10.1016/j.respol.2007.08.006.

Jaffe, A.B., Palmer, K., 1997. Environmental regulation and innovation: A panel data study. The Review of Economics and Statistics 79, 610–619. URL: http://www.jstor.org/stable/2951413, arXiv:2951413.

Jain, D., Kachinovsky, J., Rodriguez, G., Chen, J., Kim, R., Subramanian, S.V., 2025. India policy insights: A geospatial and temporal data science and visualization platform and architecture. SoftwareX 30, 102149. doi:10.1016/j.softx.2025.102149.

Joshi, S., Kansil, R., 2023. Looking at and beyond corporate governance in India, in: Corporate Governance in Emerging Markets. Springer, pp. 21–38. URL: https://link.springer.com/content/pdf/10.1007/978-981-99-3401-0.pdf. accessed: 2025-07-04.

Kagan, V., Stevens, A., Subrahmanian, V., 2015. Using Twitter sentiment to forecast the 2013 Pakistani election and the 2014 Indian election. IEEE Intelligent Systems 30, 2–5. doi:10.1109/MIS.2015.16.

Kale, V.S., 2014. Landscapes and Landforms of India. Number v.3 in World Geomorphological Landscapes Ser. 1st ed ed., Springer Netherlands, Dordrecht.

Kanilmaz, U.N., Schneidergruber, T., Schmidt, S., Dehghan, R., Scholz, J., 2025. Uncovering regional structures and collaboration hubs in Baden-Württemberg through clustering a company hyperlink network. AGILE: GIScience Series 6, 1–11. doi:10.5194/agile-giss-6-29-2025.

Kim, R., Mohanty, S.K., Subramanian, S., 2016. Multilevel geographies of poverty in India. World Development 87, 349–359. doi:10.1016/j.worlddev.2016.07.001.

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

Kinne, J., Dehghan, R., Schmidt, S., Lenz, D., Hottenrott, H., 2024. Location factors and ecosystem embedding of sustainability-engaged blockchain companies in the US. A web-based analysis. International Journal of Information Management Data Insights 4, 100287. doi:10.1016/j.jjimei.2024.100287.

Kinne, J., Lenz, D., 2021. Predicting innovative firms using web mining and deep learning. PLOS ONE 16. doi:10.1371/journal.pone.0249071.

Kriesch, L., Losacker, S., 2024. Bioeconomy firms and where to find them. REGION 11, 55–78. doi:10.18335/region.v11i1.523.

Kumar, V., 2018. Coping with climate change: An analysis of India's state action plans on climate change (Vol. II). URL: https://cdn.cseindia.org/attachments/0.40897700_1519110602_coping-climate-change-volII.pdf. accessed: 2025-07-04.

Kuznets, S., 1955. Economic growth and income inequality. The American Economic Review XLV, 1–28.

Lam, C., Souza, P.C., 2020. Estimation and selection of spatial weight matrix in a spatial lag model. Journal of Business & Economic Statistics 38, 693–710. doi:10.1080/07350015.2019.1569526.

Lin, B., Ma, X., 2021. Air pollution and corporate environmental responsibility: Evidence from Chinese listed firms. Ecological Economics 184, 106978. doi:10.1016/j.ecolecon.2021.106978.

Liu, G., Sun, W., Kong, Z., Dong, X., Jiang, Q., 2023. Did the pollution charge system promote or inhibit innovation? Evidence from Chinese micro-enterprises. Technological Forecasting and Social Change 187, 122207. doi:10.1016/j.techfore.2022.122207.

Mahalik, M.K., Villanthenkodath, M.A., Mallick, H., Gupta, M., 2021. Assessing the effectiveness of total foreign aid and foreign energy aid inflows on environmental quality in India. Energy Policy 149, 112015. doi:10.1016/j.enpol.2020.112015.

Martin, B.R., 2016. Twenty challenges for innovation studies. Science and Public Policy 43, 432–450. doi:10.1093/scipol/scv077.

Mathew, A., Shekar, P.R., Nair, A.T., Mallick, J., Rathod, C., Bindajam, A.A., Alharbi, M.M., Abdo, H.G., 2024. Unveiling urban air quality dynamics during COVID-19: A Sentinel-5P TROPOMI hotspot analysis. Scientific Reports 14, 21624. doi:10.1038/s41598-024-72276-4.

Melander, L., 2017. Achieving sustainable development by collaborating in green product innovation. Business Strategy and the Environment 26, 1095–1109. doi:10.1002/bse.1970.

Ministry of Micro, Small & Medium Enterprises, 2016. Zero Defect Zero Effect (ZED) certification scheme. URL: https://zed.msme.gov.in/. accessed: 2025-07-04.

Mishra, G., Patro, A., Tiwari, A.K., 2024. Does climate governance moderate the relationship between ESG reporting and firm value? Empirical evidence from India. International Review of Economics & Finance 91, 920–941. doi:10.1016/j.iref.2024.01.059.

Mohapatra, G., Giri, A., 2009. Economic development and environmental quality: An econometric study in India. Management of Environmental Quality: An International Journal 20, 175–191. doi:10.1108/14777830910939480.

Mohapatra, S., Seth, S., 2023. Navigating India's Transition to Sustainability Reporting. Technical Report. PwC India. URL: https://www.pwc.in/assets/pdfs/navigating-indias-transition-to-sustainability-reporting.pdf. introduces SEBI's May 2021 Business Responsibility and Sustainability Reporting framework mandatory from FY 2022–23.

Muniraj, K., Panneerselvam, B., Devaraj, S., Jesudhas, C.J., Sudalaimuthu, K., 2023. Evaluating the effectiveness of emissions reduction measures and ambient air quality variability through ground-based and Sentinel-5P observations under the auspices of COVID pandemic lockdown in Tamil Nadu, India. International Journal of Environmental Analytical Chemistry 103, 3109–3120. doi:10.1080/03067319.2021.1902997.

Murillo-Luna, J.L., Garcés-Ayerbe, C., Rivera-Torres, P., 2008. Why do patterns of environmental response differ? A stakeholders' pressure approach. Strategic Management Journal 29, 1225–1240. doi:10.1002/smj.701.

Nanjundaswamy, A., Suraj, N., M.S, D., 2024. Quality of ESG reporting and disclosure practices of Indian companies, in: ESG and Ecosystem Services for Sustainability. IGI Global, pp. 197–234. doi:10.4018/979-8-3693-3771-4.ch008.

Nenavath, S., 2022. Impact of fintech and green finance on environmental quality protection in India: By applying the semi-parametric difference-in-differences (SDID). Renewable Energy 193, 913–919. doi:10.1016/j.renene.2022.05.020.

Ord, J.K., Getis, A., 1995. Local spatial autocorrelation statistics: Distributional issues and an application. Geographical Analysis 27, 286–306. doi:10.1111/j.1538-4632.1995.tb00912.x.

Orsato, R.J., 2006. When does it pay to be green? California Management Review 48, 127–143. doi:10.2307/41166341.

Ozen, S., Kusku, F., 2009. Corporate environmental citizenship variation across economic sectors: A cross-country study. Business Strategy and the Environment 18, 180–199. doi:10.1002/bse.588.

Pastor-Merino, A., Martínez-Barbero, X., Vicente, M.R., Domenech, J., 2026. Does AI boost firm productivity? A web scraping and LLMs approach. Telecommunications Policy 50, 103138. doi:10.1016/j.telpol.2025.103138.

Pellegrini, C., Annunziata, E., Rizzi, F., Frey, M., 2019. The role of networks and sustainable intrapreneurship as interactive drivers catalyzing the adoption of sustainable innovation. Corporate Social Responsibility and Environmental Management 26, 1026–1048. doi:10.1002/csr.1784.

Pettorelli, N., Vik, J.O., Mysterud, A., Gaillard, J.M., Tucker, C.J., Stenseth, N.C., 2005. Using the satellite-derived NDVI to assess ecological responses to environmental change. Trends in ecology & evolution 20, 503–510.

Porter, M.E., Linde, C.V.D., 1995. Toward a new conception of the environment-competitiveness relationship. Journal of Economic Perspectives 9, 97–118. doi:10.1257/jep.9.4.97.

Purwandani, J.A., Michaud, G., 2021. What are the drivers and barriers for green business practice adoption for SMEs? Environment Systems and Decisions 41, 577–593. doi:10.1007/s10669-021-09821-3.

Ramkumar, S., Mueller, M., Pyka, A., Squazzoni, F., 2022. Diffusion of eco-innovation through inter-firm network targeting: An agent-based model. Journal of Cleaner Production 335, 130298. doi:10.1016/j.jclepro.2021.130298.

Rennings, K., 2000. Redefining innovation — eco-innovation research and the contribution from ecological economics. Ecological Economics 32, 319–332. doi:10.1016/S0921-8009(99)00112-3.

Reshi, A.R., Moniruzzaman, M., Tripathi, A., Tiwari, R.K., Rahaman, K.R., 2022. A remote sensing based study of tropospheric ozone concentration amid COVID-19 lockdown over india using Sentinel-5P satellite data. Geocarto International 37, 17145–17164. doi:10.1080/10106049.2022.2123957.

Riefolo, M., Grosskopf, S., Ahrens, J.P., 2026. The organization as an intelligent entity: Using artificial intelligence to explore the sustainability of family and non-family firms. Journal of Family Business Strategy 17, 100700. doi:10.1016/j.jfbs.2025.100700.

Rodríguez-Pose, A., Bartalucci, F., 2024. The green transition and its potential territorial discontents. Cambridge Journal of Regions, Economy and Society 17, 339–358. doi:10.1093/cjres/rsad039.

Saikawa, E., Trail, M., Zhong, M., Wu, Q., Young, C.L., Janssens-Maenhout, G., Klimont, Z., Wagner, F., Kurokawa, J.i., Nagpure, A.S., Gurjar, B.R., 2017. Uncertainties in emissions estimates of greenhouse gases and air pollutants in India and their impacts on regional air quality. Environmental Research Letters 12, 065002. doi:10.1088/1748-9326/aa6cb4.

Schmidt, S., Kanilmaz, U.N., Abbasiharofteh, M., Resch, B., 2025. Analysing the spatial manifestation of sustainability-engaged inter-firm networks in Germany, Austria and Switzerland. Spatial Economic Analysis doi:10.1080/17421772.2025.2573061.

Schmidt, S., Kinne, J., Lautenbach, S., Blaschke, T., Lenz, D., Resch, B., 2022. Greenwashing in the US metal industry? a novel approach combining SO2 concentrations from satellite data, a plant-level firm database and web text mining. Science of the Total Environment 835, 155512. doi:10.1016/j.scitotenv.2022.155512.

Schwierzy, J., Dehghan, R., Schmidt, S., Grashof, N., Hottenrott, H., Woywode, M., 2026. Mapping technology diffusion with AI: A web-based approach for tracking additive manufacturing adoption. International Journal of Information Management Data Insights 6, 100387. doi:10.1016/j.jjimei.2025.100387.

Sehrawat, M., Giri, A.K., Mohapatra, G., 2015. The impact of financial development, economic growth and energy consumption on environmental degradation: Evidence from India. Management of Environmental Quality: An International Journal 26, 666–682. doi:10.1108/MEQ-05-2014-0063.

Serere, H.N., Kanilmaz, U.N., Ketineni, S., Resch, B., 2023. A comparative study of geocoder performance on unstructured tweet locations. GI_Forum 1, 110–117. doi:10.1553/giscience2023_01_s110.

Shahbaz, M., Mallick, H., Mahalik, M.K., Loganathan, N., 2015. Does globalization impede environmental quality in India? Ecological Indicators 52, 379–393. doi:10.1016/j.ecolind.2014.12.025.

Sharma, D., Mauzerall, D., 2022. Analysis of air pollution data in India between 2015 and 2019. Aerosol and Air Quality Research 22, 210204. doi:10.4209/aaqr.210204.

Singh, U., Ravi Shankar, D., Bellala, G., Goel, V., 2025. Geo-spatially informed models for geocoding unstructured addresses, in: Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B.D., Schockaert, S., Darwish, K., Agarwal, A. (Eds.), Proceedings of the 31st International Conference on Computational Linguistics: Industry Track, Association for Computational Linguistics, Abu Dhabi, UAE. pp. 236–242. URL: https://aclanthology.org/2025.coling-industry.19/.

Søgaard, A., 2022. Should we ban english NLP for a year?, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates. pp. 5254–5260. doi:10.18653/v1/2022.emnlp-main.351.

Soulie, A., Granier, C., Darras, S., Zilbermann, N., Doumbia, T., Guevara, M., Jalkanen, J.P., Keita, S., Liousse, C., Crippa, M., Guizzardi, D., Hoesly, R., Smith, S.J., 2024. Global anthropogenic emissions (CAMS-GLOB-ANT) for the Copernicus Atmosphere Monitoring Service simulations of air quality forecasts and reanalyses. Earth System Science Data 16, 2261–2279. doi:10.5194/essd-16-2261-2024.

STATISTA, 2025. Number of registered active companies across India as of January 2025, by type. URL: https://www.statista.com/statistics/1008330/india-number-of-registered-active-companies-by-type/.

Sutton, J., Sutton, J., 2024. A road map to capture the spatial dependence underlying regions' economic resilience. Spatial Economic Analysis 19, 57–72. doi:10.1080/17421772.2023.2240405.

Tan, Z., Yan, L., 2021. Does air pollution impede corporate innovation? International Review of Economics & Finance 76, 937–951. doi:10.1016/j.iref.2021.07.015.

UN, 2023. India overtakes China as the world's most populous country. URL: https://www.un.org/development/desa/dpad/wp-content/uploads/sites/45/PB153.pdf.

Upadhyay, J.P., Dugal, I., 2025. India to review ESG disclosures for listed firms, market regulator says. URL: https://www.reuters.com/sustainability/boards-policy-regulation/india-review-esg-disclosures-listed-firms-market-regulator-says-2025-04-16/.

Villanthenkodath, M.A., Ansari, M.A., Shahbaz, M., Vo, X.V., 2022. Do tourism development and structural change promote environmental quality? Evidence from India. Environment, Development and Sustainability 24, 5163–5194. doi:10.1007/s10668-021-01654-z.

Villanthenkodath, M.A., Mahalik, M.K., 2022. Technological innovation and environmental quality nexus in India: Does inward remittance matter? Journal of Public Affairs 22, e2291. doi:10.1002/pa.2291.

Wang, J., Wang, L., Qian, X., 2021a. Revisiting firm innovation and environmental performance: New evidence from Japanese firm-level data. Journal of Cleaner Production 281, 124446. doi:10.1016/j.jclepro.2020.124446.

Wang, X., Li, X., Gao, Y., Lv, J., 2021b. Does air pollution affect corporate social responsibility? Evidence from China. Journal of Cleaner Production 287, 125011. doi:10.1016/j.jclepro.2020.125011.

Wildnerova, L., Menon, C., Dehghan, R., Kinne, J., Lenz, D., 2024. Which SMEs Are Greening?: Cross-Country Evidence from One Million Websites. OECD SME and Entrepreneurship Papers 60. OECD. doi:10.1787/ddd00999-en.

Xie, W., Qing, Y., Tao, L., Li, W., Wen, C., 2025. Spatial spillover effects of environmental regulation on ecological industrialization: Evidence from the upper reaches of the Yangtze River. International Review of Economics & Finance 98, 103862. doi:https://doi.org/10.1016/j.iref.2025.103862.

Yaduvanshi, A., Nkemelang, T., Bendapudi, R., New, M., 2021. Temperature and rainfall extremes change under current and future global warming levels across Indian climate zones. Weather and Climate Extremes 31, 100291. doi:10.1016/j.wace.2020.100291.

Yan, X., Zang, Z., Jiang, Y., Shi, W., Guo, Y., Li, D., Zhao, C., Husi, L., 2021. A Spatial-Temporal Interpretable Deep Learning Model for improving interpretability and predictive accuracy of satellite-based PM2.5. Environmental Pollution 273, 116459. doi:10.1016/j.envpol.2021.116459.

Zhu, Q., Cordeiro, J., Sarkis, J., 2012. International and domestic pressures and responses of Chinese firms to greening. Ecological Economics 83, 144–153. doi:10.1016/j.ecolecon.2012.04.007.

Appendix A.

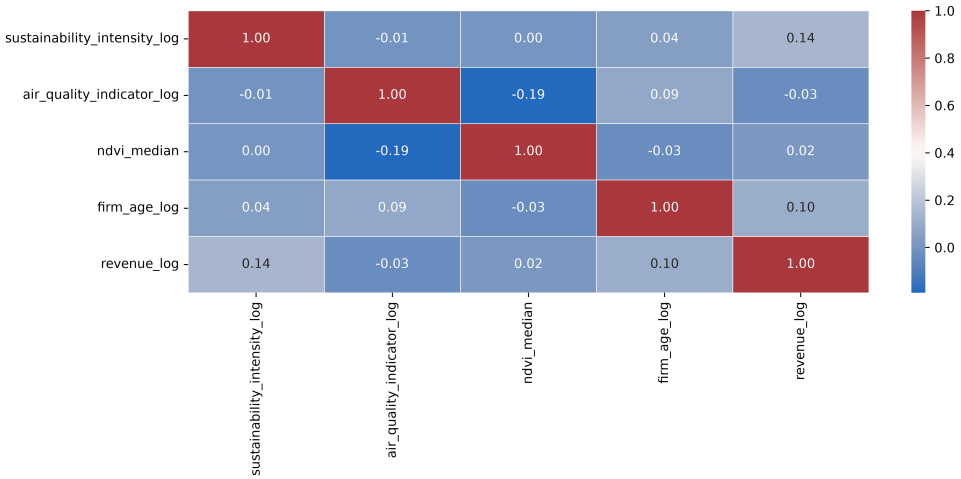


Figure A.5: Correlation matrix on firm-level.

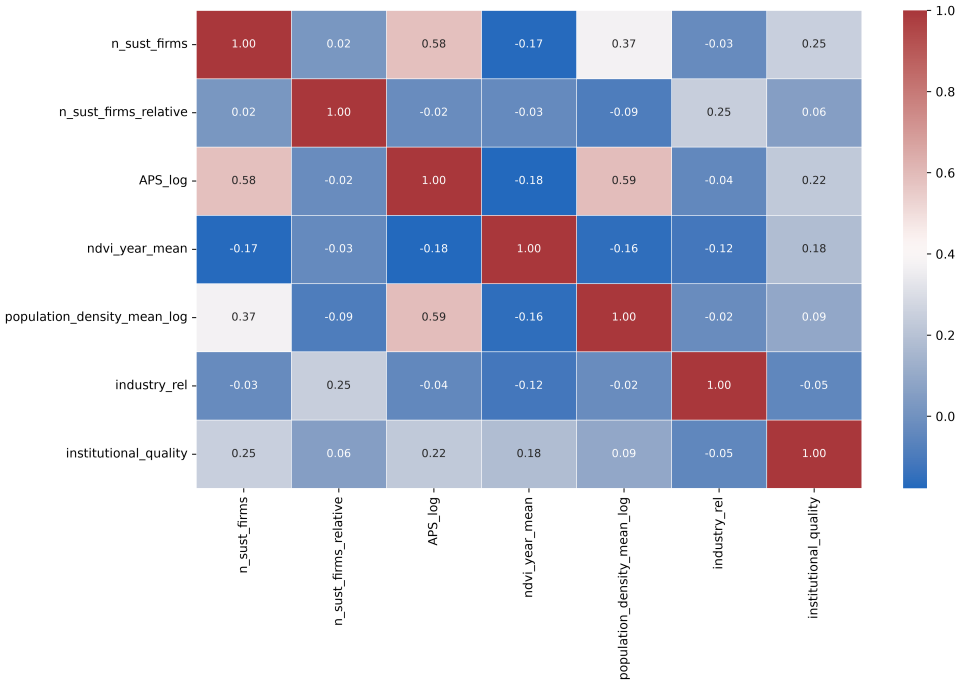


Figure A.6: Correlation matrix on district-level.

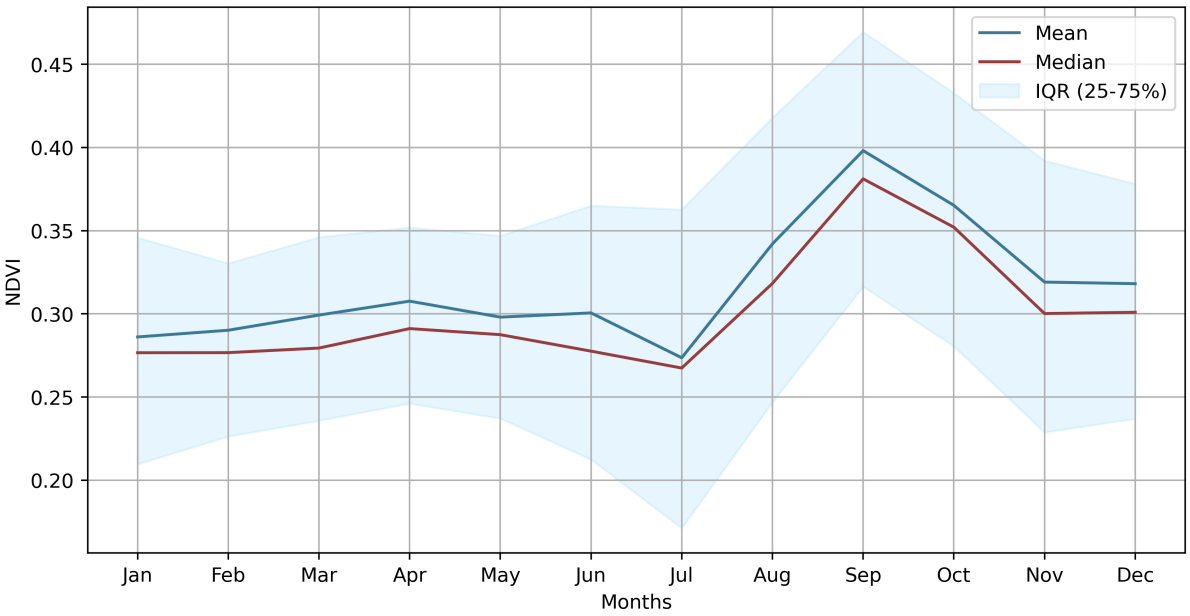


Figure A.7: Timeline of median NDVI values over the course of the year.

Table A.4: Regression results at the region level. The dependent variable is the logarithmised absolute number of sustainability-engaged firms per district. Only districts with >10 firms were considered.

	Model 1		Model 2		Model 3		Model 4		Model 5	
	OLS	SLM	OLS	SLM	OLS	SLM	OLS	SLM	OLS	SLM
CONSTANT	2.793***	1.836***	3.305***	2.148***	-0.39	-1.374**	-1.586*	-2.064***	-2.649***	-2.927***
APS_log	9.912***	7.653***	9.594***	7.487***	3.094*	1.28	3.200**	1.48	-0.11	-0.87
W_n_sust_firms_log	-	0.337***	-	0.334***	-	0.331***	-	0.308***	-	0.201***
ndvi_year_mean	-	-	-1.02	-0.60	-1.303*	-0.88	-0.64	-0.49	-3.357***	-2.965***
population_density_mean_log	-	-	-	-	0.629***	0.602***	0.682***	0.638***	0.651***	0.631***
industry_rel	-	-	-	-	-	-	1.379***	0.871**	1.328***	1.004***
institutional_quality	-	-	-	-	-	-	-	-	4.579***	4.165***
R^2	0.16	-	0.16	-	0.24	-	0.27	-	0.45	-
Adj. R^2	0.16	-	0.16	-	0.24	-	0.26	-	0.44	-
Pseudo R^2	-	0.03	-	0.03	-	0.06	-	0.07	-	0.14
AIC	1,211.24	1,176.13	1,211.67	1,177.50	1,179.47	1,143.53	1,170.48	1,140.76	1,045.74	1,032.53
Log-Lik	-603.62	-585.07	-602.83	-584.75	-585.73	-566.77	-580.24	-564.38	-516.87	-509.27
RMSE	1.36	1.27	1.36	1.27	1.30	1.21	1.28	1.20	1.11	1.08
F-test	66.41***	-	34.04***	-	36.80***	-	31.15***	-	53.97***	-
LM-Lag	38.65***	-	37.88***	-	42.00***	-	34.09***	-	16.34***	-
LM-Error	32.43***	-	32.94***	-	71.46***	-	53.07***	-	48.83***	-
VIF (max)	-	-	1.2	1.2	28.05	28.05	33.44	33.44	35.75	35.75
Multicollinearity	1.6	-	12.0	-	25.1	-	30.0	-	33.9	-
Moran's I	0.239***	-0.01	0.241***	-0.01	0.355***	0.093***	0.306***	0.082**	0.298***	0.146***
Obs.	349	349	349	349	349	349	349	349	340	340

Dependent variable: *n_sust_firms_log*, * p<0.10, ** p<0.05, *** p<0.01.

Table A.5: Regression results at the region level. The dependent variable is the relative number of sustainability-engaged firms per district. Only districts with >10 firms were considered.

	Model 1		Model 2		Model 3		Model 4		Model 5	
	OLS	SLM	OLS	SLM	OLS	SLM	OLS	SLM	OLS	SLM
CONSTANT	0.394***	0.305***	0.438***	0.327***	0.599***	0.474***	0.329***	0.268***	0.328***	0.273***
APS_log	-0.10	-0.10	-0.12	-0.11	0.16	0.13	0.182*	0.17	0.14	0.13
W_n_sust_firms_relative	-	0.236***	-	0.230***	-	0.216***	-	0.131***	-	0.114**
ndvi_year_mean	-	-	-0.09	-0.04	-0.07	-0.03	0.08	0.093*	0.03	0.05
population_density_mean_log	-	-	-	-	-0.027***	-0.024***	-0.016**	-0.014*	-0.018**	-0.016**
industry_rel	-	-	-	-	-	-	0.309***	0.293***	0.303***	0.290***
institutional_quality	-	-	-	-	-	-	-	-	0.075**	0.063*
R^2	0.00	-	0.01	-	0.04	-	0.28	-	0.29	-
Adj. R^2	0	-	0.00	-	0.03	-	0.27	-	0.28	-
Pseudo R^2	-	-0.00	-	-0.00	-	-0.02	-	-0.17	-	-0.17
AIC	-588.60	-607.57	-588.53	-606.03	-597.47	-612.74	-697.38	-702.84	-677.40	-681.07
Log-Lik	296.30	306.79	297.26	307.01	302.73	311.37	353.69	357.42	344.70	347.54
RMSE	0.10	0.10	0.10	0.10	0.10	0.10	0.09	0.09	0.09	0.09
F-test	1.13	-	1.53	-	4.71***	-	33.88***	-	27.38***	-
LM-Lag	22.36***	-	20.55***	-	18.13***	-	7.61***	-	5.78**	-
LM-Error	27.96***	-	26.00***	-	21.55***	-	6.58**	-	6.37**	-
VIF (max)	-	-	1.2	1.2	28.05	28.05	33.44	33.44	35.75	35.75
Multicollinearity	1.6	-	12.0	-	25.1	-	30.0	-	33.9	-
Moran's I	0.222***	0.052*	0.214***	0.051*	0.195***	0.05	0.108**	0.03	0.108***	0.04
Obs.	349	349	349	349	349	349	349	349	340	340

Dependent variable: *n_sust_firms_relative*, * p<0.10, ** p<0.05, *** p<0.01.

Table A.6: Regression results on firm-level.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11
CONSTANT	0.175***	0.174***	0.109***	0.060***	0.053***	0.112***	0.113***	0.080***	0.079***	0.132***	0.101***
APS_log	-0.037**	-0.036**	-0.074***	-0.098***	-0.115***	-0.02	-0.04	-0.118***	-0.142***	-0.04	-0.112***
ndvi_median	-	0.00	0.01	0.062***	0.078***	0.060***	0.065***	-	-	-	0.084***
firm_age_log	-	-	0.025***	-0.007**	-0.007**	-0.019***	-0.019***	-0.007**	-0.007**	-0.019***	-0.020***
revenue_log	-	-	-	0.012***	0.012***	0.011***	0.011***	0.012***	0.012***	0.011***	0.011***
state_Gujarat	-	-	-	-	0.018***	-	-0.00	-	0.013**	-	0.013**
state_Karnataka	-	-	-	-	0.00	-	0.00	-	0.00	-	0.012**
state_Maharashtra	-	-	-	-	0.026***	-	0.020***	-	0.025***	-	0.023***
state_Tamil Nadu	-	-	-	-	0.00	-	-0.01	-	0.00	-	0.01
state_Telangana	-	-	-	-	-0.016**	-	-0.020***	-	-0.018**	-	-0.01
state_Uttar Pradesh	-	-	-	-	0.00	-	-0.01	-	0.00	-	0.00
state_West Bengal	-	-	-	-	-0.018**	-	-0.022***	-	-0.016**	-	-0.017**
nacegroup_industry	-	-	-	-	-	0.065***	0.064***	-	-	0.065***	-
nacegroup_non_industry	-	-	-	-	-	-0.042***	-0.043***	-	-	-0.042***	-
nace_fine_20	-	-	-	-	-	-	-	-	-	-	0.169***
nace_fine_27	-	-	-	-	-	-	-	-	-	-	0.050***
nace_fine_28	-	-	-	-	-	-	-	-	-	-	0.048***
nace_fine_42	-	-	-	-	-	-	-	-	-	-	-0.01
nace_fine_46	-	-	-	-	-	-	-	-	-	-	0.016*
nace_fine_62	-	-	-	-	-	-	-	-	-	-	-0.097***
nace_fine_70	-	-	-	-	-	-	-	-	-	-	0.055***
nace_fine_82	-	-	-	-	-	-	-	-	-	-	0.025***
nace_fine_85	-	-	-	-	-	-	-	-	-	-	-0.102***
nace_fine_86	-	-	-	-	-	-	-	-	-	-	-0.104***
R^2	0.00	0.00	0.00	0.02	0.02	0.04	0.04	0.02	0.02	0.04	0.04
Adj. R^2	0.00	0.00	0.00	0.02	0.02	0.04	0.04	0.02	0.02	0.04	0.04
AIC	46,774.46	46,740.27	42,666.52	21,529.67	21,476.56	20,835.81	20,793.95	21,555.55	21,507.91	20,860.15	20,813.08
RMSE	0.29	0.29	0.30	0.32	0.32	0.32	0.32	0.33	0.32	0.32	0.32
F-test	5.31**	2.63	77.60***	201.99***	79.68**	254.64***	121.98***	265.45***	85.92**	303.53***	75.34***
VIF (max)	-	1.56	9.20	16.71	17.53	18.21	19.04	11.79	12.54	13.46	17.98
Multicollinearity	2.30	7.80	14.20	18.60	20.70	20.40	22.50	15.50	17.50	17.40	21.90
Obs.	117,622	117,537	103,831	36,673	36,673	36,673	36,673	36,704	36,704	36,704	36,673

Dependent variable: *sustainability_intensity_log*, * p<0.10, ** p<0.05, *** p<0.01.
NACE codes: 20 (Chemicals), 27 (Electrical equipment), 28 (Machinery), 42 (Civil engineering), 46 (Wholesale trade), 62 (IT services), 70 (Consulting), 82 (Business support), 85 (Education), and 86 (Healthcare)



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

Discussion Papers are intended to make results of ZEW research promptly available to other economists in order to encourage discussion and suggestions for revisions. The authors are solely responsible for the contents which do not necessarily represent the opinion of the ZEW.