Maike Söderholm (e122856)

Forecasting company performance from space

Exploring the usability of the satellite based TROPOMI NO2 tropospheric column for company performance prediction

School of Accounting and Finance
Master's thesis in Finance
Master's Degree Program in Finance

Vaasa 2023

UNIVERSITY OF VAASA
School of Accounting and Finance
Author: Maike Söderholm (e122856)

Title of the Thesis: Forecasting company performance from space : Exploring the usability of the satellite based TROPOMI NO2 tropospheric column for company performance prediction

Degree: Master
Programme: Finance
Supervisor: Timothy King
Year: 2023 Number of Pages: 84

ABSTRACT:

Acknowledging the growing traction of non-traditional datasets in financial analysis, this study explores the potential of the satellite based TROPOMI NO2 tropospheric column as a predictive tool for assessing the financial performance and productivity of mining companies. The focus of this analysis lies on the investigation of the correlation between the TROPOMI NO2 tropospheric column and firms' financial performance measured as sales, net income, and ROA. Additionally, the relationship of the TROPOMI NO2 tropospheric column and firm productivity, measured as mine output is analysed. Utilizing the TROPOMI NO2 tropospheric column of the European Space Agency's Copernicus program, the correlation between NO2 concentration and firm performance is analysed employing univariate and multivariate regression. The results indicate a significant positive correlation between NO2 levels and sales, that is particularly pronounced among smaller firms, newly established firms, and those with less mines. For instance, the findings of the main analysis suggest that a 1% increase in NO2 concentration is associated with a 0.052% increase in sales. Conversely, the impact on net income and ROA is inconclusive, showing inconsistency across different analyses. The results of this thesis must be interpreted with caution due to data limitations and call for further investigation. Nevertheless, the findings uncover the potential value of satellite based NO2 data as a supplementary data source to enhance financial analysis, which is especially relevant for firms that are not publicly traded.

KEYWORDS: alternative data, satellite data, remote sensing, company performance

Contents

Figures

Tables

Abbreviations

EC	**– European Commission**
EEA	**– European Environment Agency**
ESA	**– European Space Agency**
FMI	**– Finnish Meteorological Institute**
NLP	**– Natural Language Processing**
ROA	**– Return on Assets**
SA	**– Seeking Alpha**
Sentinel-5-P	**– Sentinel-5-Precursor**
SX-EW	**– Solvent Extraction and Electrowinning**
TROPOMI	**– TROPOspheric Monitoring instrument**

1 Introduction

There has been a sharp increase in the amount of data during the information age, which refers to the time in which information is so widely distributed and readily available through the internet, that it has become a commodity (Demchenko et al., 2018). Although data has always been a key element in financial analysis, recently there has been an increase in data supply and use of particularly non-traditional datasets (Mitra et al., 2023). For example, the development of the internet has given rise to social media platforms on which individuals can generate a vast amount of user-generated content. This is one example of so-called "alternative data" a term that describes data that are not traditionally used in financial markets but is believed to contain useful information for investment purposes (Hansen & Borch, 2022). Consistent with the definition of data as a commodity Hansen and Borch (2022) argue that alternative data can be regarded as an asset once the raw data has been processed to make it usable for investment firms. Data processing means the transformation of unstructured data, such as textual data, into a structured format so that it can be inserted into financial models (Hansen & Borch, 2022).

Deloitte states in a 2017 report relating to alternative data for investment decisions that:

> "*Those firms that do not update their investment processes (…) could face strategic risks, and might very well be outmanoeuvred by competitors that effectively incorporate alternative data into their securities valuation and trading processes."(p.1)*

Today, alternative datasets are heavily used to support investment decisions in the financial sector (Mitra et al., 2023). For example, studies by Tang (2019) and Sheng (2022) find evidence that hedge funds are trading on information contained in X, formerly known as Twitter, or commentary and Glassdoor ratings respectively. Glassdoor is a website that offers company reviews, salary details, and job insights to provide transparency in the job market (Glassdoor, n.d.). While there is a hype around the topic, Hansen and Borch

(2022) were able to gain insight into the actual application of alternative data in the industry in their research, finding that while there is some hesitancy towards these datasets and alternative data is typically used as a supplementary data source, the exploration of alternative datasets is necessary to not fall behind the competition.

The value of alterative datasets is based in the assumption that alternative data precedes traditional data sources, such as a company's financial statements. The goal is to extract signals from alternative datasets that relate to traditional data and exploit the information advantage (Hansen & Borch, 2022). Furthermore, alternative data enables researchers to analyse for instance the impact of investor sentiment on the behaviour of capital markets, which are difficult to assess through traditional data sources (Teoh, 2018). Research related to the value of alternative data for investment purposes finds that it contains useful information to forecast a company's earnings, that is incremental to other, more traditional sources of information like analyst consensus (Jame et al., 2016; Bartov et al., 2018). Furthermore, signals extracted from different types of alternative data are predictive of stock prices and firm fundamental information (Chen et al., 2014; Tang, 2018; Hales et al., 2018; Katona et al., 2018; Kang et al., 2021; Feng and Fay, 2022). Moreover, researchers were able to formulate profitable trading strategies based on information obtained from various types of alternative data (Huang, 2018; Green et al., 2019; Katona et al., 2018; Sheng, 2022).

According to Denev and Saeed (2020), the decreasing returns the more heavily a dataset is used by investors, is another reason for the value of alternative datasets. Traditional data is usually easily accessible and is therefore utilized by many investors so that returns are fading away (Mitra et al., 2023). Alternative data on the other hand is associated with high acquisition, processing, and implementation costs, making it available to mostly sophisticated investors (Katona et al., 2018). Drawing on interview data with employees of firms that have specialized on both the sell- and buy-side of alternative data, Hansen and Borch (2022) find that tech- and data driven investment firms are always searching for new datasets to secure competitive advantages.

Satellite data is another promising source of alternative data. Anecdotal evidence suggests that information obtained from satellite imagery can be used to formulate profitable trading strategies (Partnoy, 2019). Previous research relating to satellite data mostly focuses on large-scale academic forecasting (Donaldson and Storeygard, 2016). However, recent studies find that parking lot traffic information from US retailers extracted from satellite imagery can be used to predict retailer performance (Katona et al., 2018; Kang et al., 2021; Feng & Fay; 2022).

Advances in the utilization of alternative dataset are always enabled through a combination of data availability and developments in machine-learning methods (Hansen & Borch, 2022). For example, the large amounts of textual data submitted through social media would make it impossible to retrieve a sentiment from the data without algorithms that are specialized in the processing of natural language. Likewise, it would not be feasible to utilize parking lot traffic without algorithms that automatically detect cars in satellite imagery.

Recent technological advancements in satellite data availability and machine-learning methods enable the measurement of NO2 concentrations from point sources at the surface of the Earth from space. The European Space Agency's (ESA) Copernicus program, specifically the Sentinel-5-P satellite, provides satellite data with a higher spatial resolution (Scheibenreif et al., 2019), meaning that the data is available for more precise locations. Scheibenreif et al. (2019) were able to develop a model that predicts NO2 emissions from point sources from only satellite data, for the first time. Furthermore, Martinez-Alonso et al. (2023) were able to establish a connection between NO2 concentrations derived from Sentinel-5P data and mine production in the Copperbelt region.

1.1 Purpose of the study

The purpose of this thesis is to analyse whether satellite based NO2 levels from copper mines can be used to predict performance and productivity metrics for companies in the mining industry. Specifically, the use of the "nitrogendioxide_tropospheric_column" of the TROPOMI NO2 level 2 data product from the Sentinel-5-P mission of the ESA's Copernicus program, to predict firms' financial performance measured as sales, net income, and return on assets (ROA) as well as productivity, measured as copper output in tonnes. To investigate the correlation between NO2 levels and financial performance, this study utilizes a sample of 224 firm-quarter observations from 14 distinct mining companies between 2019Q1 and 2022Q4. Furthermore, a sample of 208 firm-quarter observations for 13 distinct mining companies for the period from 2019Q1 until 2022Q4 is used to analyse the relationship between NO2 levels and mine-output.

The findings indicate a notable positive correlation between NO2 levels and sales, with subsequent impacts on sales outcomes in the following quarter, that is consistent for smaller firms, firms with less mines, and newly established firms. The results for the correlation between NO2 levels and net income as well as ROA did not yield consistent results. The correlation between NO2 levels and mine output showed initial significant differences in output between mines with high and low NO2 levels, yet the multivariate analysis did not confirm these outcomes, suggesting the need for further investigation. Overall, the results highlight the potential of the satellite based NO2 variable as a predictor of firms' financial performance, with caution due to data limitations. The findings provide a first exploration of the relationship, calling for careful interpretation and further research to generalize the results.

This thesis contributes to the literature related to the value of alternative data for investment purposes by analysing a new dataset. Signals obtained from the NO2 tropospheric column could be used as a supplementary source of data to inform trading strategies in the mining industry. Furthermore, the new dataset can be used to access information that is otherwise inaccessible as for example the financial performance of non-publicly traded companies.

1.2 Hypotheses

The focus of this thesis lies on the research question whether the satellite based NO2 variable can be used for the prediction of firms' financial performance. Additionally, the predictive capabilities of the NO2 variable for firm productivity is assessed. The following hypotheses are aimed to be tested in this thesis:

> H1: The financial performance of firms in the mining industry, measured as sales, net income, and ROA, can be predicted using satellite observed NO2 levels, measured as NO2 concentration in mol/cm^2.
> H2: The productivity of firms in the mining industry, measured as aggregated output from all mines, can be predicted using satellite observed NO2 levels, measured as NO2 concentration in mol/cm^2.

Regarding the second hypothesis (H2), Martinez-Alonso et al. (2023) find a strong correlation between NO2 emissions and mine-production in the Copperbelt region. The authors investigate the relationship between NO2 emissions from copper mines and mine production measured as the amounts of ore and waste, as well as total copper produced. Martinez-Alonso et al. (2023) are most interested in the energy consumption related to the production. Therefore, the authors argue that the total output amount is a less accurate proxy for energy consumption than for example the total amount of ore plus waste, because the total output is dependent on other factors in the mine's environment such as ore grade or fuel efficiency. However, for the purpose of this study, the use of total mine-output is reasonable, since the total output is closer related to the financial performance of the company which is the focus of this thesis.

The first hypothesis (H1) is based on the logical assumption that increased production, which is assumed to lead to higher NO2 emissions as suggested by Martinez-Alonso et al. (2023), should be reflected in better financial performance for the firms that own the respective mines. While this hypothesis is not derived from existing research that examines the relationship between the NO2 variable used in this thesis and company performance, it is supported by the fact that the NO2 variables in this research and the study by Martinez-Alonso et al. (2023) are linked. The authors derived NO2 emissions from the NO2 data that is used in this thesis, finding that NO2 emissions are positively related to mine output. Therefore, higher NO2 levels are expected to be positively correlated with firm performance indicators. Additionally, related research finds evidence that productivity is related to firm performance. For instance, Katona et al., (2018), Kang et al. (2021), and Feng and Fay (2022) analyse the relationship between parking lot traffic obtained from satellite imagery and retailer performance, finding that changes in parking lot traffic are positively associated with subsequent retailer performance. Although NO2 levels are different from parking lot traffic, both variables are indicators for the productivity of the respective business. However, regarding the association between higher NO2 levels and increased productivity, it is important to consider that higher emissions could also be a result of lower efficiency (Mehmood Mirza et al., 2022).

1.3 Limitations

This study is constrained by several limitations related to data availability and quality. Firstly, the NO2 concentration data contains numerous missing values, especially in areas with frequent could cover. Secondly, the size of the dataset is limited due to the period of available TROPOMI NO2 data (2019 - 2022). Additionally, the exclusion of firms with missing or incomplete financial data further reduced the sample size. Lastly, the variables included in the final dataset exhibit outliers and skewness.

Furthermore, the NO2 variable used in this thesis represents NO2 concentrations and not NO2 emissions. Concentration is the amount of NO2 present in the air, while emissions are the amount of NO2 released from a particular source. Beirle et al. (2019) point out some limitations of utilizing the NO2 tropospheric column used in this thesis. While plumes of strong point sources are observable for specific points in time, the NO2 patterns are “smeared out” when time-based averages are taken, which diminishes the advantage of the higher resolution (Beirle et al., 2019). This can impact the usefulness of this variable for certain types of analysis, like tracking the precise behaviour of emissions or estimating the actual emission levels from a particular source over time, as well as locating point sources of emissions. However, in the context of accessing the level of NO2 concentration from a known source and its relationship to company performance, the variable can still be useful. Since the sources are already known, the NO2 variable can be used to estimate the approximate level of NO2 emissions from this source. Even though the data might be "smeared out" when temporal averages are taken, it can still provide valuable insights into long-term trends and the relationship to performance and productivity. Overall, while NO2 emissions might be more effective when analysing the relationship between NO2 emissions and company performance, the satellite based NO2 variable should still be useful for this type of analysis.

Martinez-Alonso et al. (2023) find that the positive correlation between NO2 emissions and mine-production are mine-dependent and sensitive to changes in the environment of the mine such as ore grade and fuel efficiency. Since these parameters seem important in explaining the relationship between NO2 levels and mine output, it presents another limitation that these variables were not available in the dataset utilized in this thesis.

1.4 Structure of the study

The remainder of the thesis is organized as follows. The second chapter presents the theoretical background that forms the basis to understand the differentiation between different data types used in financial analysis and their relevance as well as their respective advantages and challenges. The chapter starts with alternative data and gets more specific to a description of satellite data, and the specific type of satellite data used in this thesis. The third chapter reviews previous literature related to the use of alternative data for company performance prediction, satellite data in financial research, and measuring NO2 emissions from point sources with satellite data. The fourth chapter describes the data and methodology that are used in this thesis. Chapter five presents the results of the empirical analysis and finally, the sixth chapter concludes the thesis.

2 Theoretical Background

This chapter provides a comprehensive description of the data sources used in financial analysis that are relevant for this thesis. The first subchapter focuses on the broader category of alternative data, followed by a more detailed insight into satellite data as one type of alternative data in the second subchapter. The first two subchapters comprise definitions, examples, as well as opportunities and challenges of the respective data sources. Lastly, the third subchapter describes the satellite program through which the data used in this thesis has been collected, as well as the NO2 variable itself.

2.1 Alternative Data

The types of data that belong to the category of alternative data are often explained through the differentiation to traditional data, as traditional and alternative datasets generally differentiate in their characteristics. Teoh (2018) refers to standardly used data sources as “traditional databases” which include rather numerical information, such as company disclosures, stock prices, and analyst reports. Mitra et al. (2023) characterize alternative data as unstructured, multidimensional, and available at a high volume and frequency (Mitra et al., 2023). Denev and Saeed (2020) define several characteristics of alternative data and consider a data source alternative data if it has one or more of these characteristics, which include less common use in financial markets, higher acquisition costs, typically outside of financial markets, shorter history and more challenging to use. The definitions overlap somewhat since the features of multidimensionality and unstructured form make the datasets more challenging to use.

There is not a strict definition of the term alternative data. The term is typically defined as data that do not belong to traditionally used data sources in the financial market, but which are presumed to provide valuable information for investors (Hansen & Borch, 2022). However, the scope of what is considered alternative data is constantly changing as new data sources are added when they become relevant for investment decisions and data that are commonly used in the industry are not considered alternative anymore

(Hansen & Borch, 2022). An example of a type of alternative data that went through this process is environmental, social, and governance (ESG) data, which nowadays are commonly used to support investment decisions (Hansen & Borch, 2022). The idea behind alternative data is to obtain signals that relate to variables that will end up on traditional data sources, such as the income statement, before they are released and trade on the information advantage (Hansen & Borch, 2022).

As the scope of alternative data is constantly changing it is difficult to obtain an exhaustive list of alternative data types and to categorize these. Additionally, companies are hesitant to provide information about the types of data or model specifications they use, as those parameters are the core of a firm's competitive advantage (Hansen & Borch, 2022). Denev and Saeed (2020) differentiate alternative data types by the source that generated them which can be individuals, institutions, and sensors. Variations of this categorization approach can be found in an online search for the term alternative data. For example, Refinitiv categorizes alternative data into the categories of individuals, business processes, as well as satellites and sensors. Furthermore, AlternativeData.org has adopted the categorization by generator of the data and provides and overview of the major types of alternative data for each category. However, the category of institutions is referred to as business processes and includes for example credit/debit card data and email/consumer receipts (AlternativeData.org, n.d.). According to Denev and Saeed (2020), these would belong to the category of individuals, as they are generated through purchases by individual customers. Following Denev and Saeed (2020) examples for alternative data produced by individuals include social media postings and web traffic, data generated by institutions includes corporate or government reports and sensor data includes satellite data.

Alternative data is typically not ready to use once it is retrieved. Denev and Saeed (2020) define the four stages of alternative data utilization as follows: raw data, processed data, signals, and strategy. This is in line with Hansen and Borch (2022), who argue that alternative data needs to go through a standardization process to make it useful for investment and trading firms. Hansen and Borch (2022) also call this standardization process “assetization” of data. Standardized data has properties of what is typically understood as an asset and thus can be viewed as such (Hansen & Borch, 2022). Hansen and Borch (2022) describe the processing stages by the example of social media for sentiment analysis. Sentiment analysis can be understood as the extraction of public sentiment from a large collection of social media posts or other textual data that are categorized into positive, neutral, and negative categories (Hansen & Borch, 2022). Firstly, tweets or other textual data are collected, duplicates are removed, and the remaining posts are categorized into negative, neutral, or positive posts. This results in a sentiment score that can be aggregated for different time-horizons to detect sentiment changes and base investment decisions on those. (Hansen & Borch, 2022). Such data is typically quite expensive and usually sold by companies that specialize in the cleaning of unstructured data (Mitra et al., 2023). The different data-science methos used for the standardization of alternative data, such as natural language processing (NLP) machine learning algorithms, are just as important as the data itself and they are regarded as the limiting factor of what is possible with the data (Hansen & Borch, 2022).

Alternative datasets come with challenges, some of which are also found in traditional datasets such as endogeneity issues or reversed causality (Teoh, 2018), but also some additional challenges that are mainly related to alternative data. The challenges and risks relating to alternative datasets can be categorized as risks associated with regulation, the use of alternative data, or general challenges with the processing of such data (Denev & Saeed, 2020). Legal issues comprise for example different data protection laws around the world or the retrieval of data from private websites or information behind a paywall that is accessed via web scraping (Denev & Saeed, 2020). Risks related to the use of alternative data include quality issues as for example the performance of algorithms, which depends on the image quality from satellite imagery (Teoh, 2018). Furthermore, social media sentiment is prone to manipulation and the derived sentiment is not necessarily correct (Hansen & Borch, 2022). Thus, the utilization of a signal obtained from certain sources of alternative data might not always be straightforward. On the one hand, the use of satellite imagery of parking lots for the prediction of retailer performance is straightforward. A higher parking lot fill rate is positively associated with performance indicators (Katona et al., 2018; Kang et al., 2021; Feng & Fay, 2022). On the other hand, sentiment data retrieved from social media can be used either to follow the aggregate investor opinion or bet against it, as it may not always be accurate (Hansen & Borch, 2022). Lastly, challenges regarding the processing of alternative data are related to the conversion of raw, unstructured alternative data into a structured form so that it can be inserted into financial models (Denev & Saeed, 2020). Additionally, the volume and complexity of alternative data are a challenge in the sense that it makes the use of such data expensive. Mitra et al. (2023) describe two options when using alternative data which are both associated with high costs. Firstly, a company can buy cleaned data at a high price since alternative datasets are often proprietary. Secondly, a firm can employ a team that takes care of the data processing which results in high payroll costs. Similarly, Katona et al. (2018) argue that high acquisition, processing, and implementation costs make alternative datasets only available to sophisticated investors which is increasing information asymmetry among market participants.

The amount of alternative data has been rapidly increasing as well as the use of these datasets (Denev & Saeed, 2020; Mitra et al., 2023). Major actors in the utilization of alternative data are firms operating in investment management, proprietary trading, data analytics, and securities exchanges (Hansen & Borch, 2022). However, it is important to note that alternative data is typically used as supplementary data next to traditional sources and investment decisions are not made solely on signals from alternative data (Hansen & Borch, 2022). Despite reservations due to current challenges and risks associated with the use of alternative data, exploring these data sources is necessary for companies to remain competitive (Denev & Saeed, 2020; Hansen & Borch, 2022).

2.2 Satellite Data

Satellite data is one type of alternative data that belongs to the category of sensor data (Gladilin et al., 2023). Sensor data is data that are collected via remote sensing, which can be defined as information collection from a distance (Nasa, n.d.). Sensor data includes for example satellite imagery, geolocation data, cameras, or weather and pollution sensors (Gladilin, 2023). Thus, satellite imagery, which are pictures of the surface of the earth taken from satellites, also belong to this category. Remote sensors can be placed on satellites or airplanes to collect data from the surface of the Earth from space (Nasa, n.d.). A satellite defines any object in an orbit, which is the path on which objects are moving around planets (Nasa, 2017). Thus, planets can also be satellites, however, in this thesis satellites are understood as man-made satellites. There are satellites in different orbits of the earth which differ in altitude and the specific path they follow, allowing the collection of data for different parts of the Earth's surface (Nasa, n.d.).

The method of remote sensing is based on the reflection of sunlight. Different surface characteristics, like roughness, determine the amount of sunlight that is absorbed and the amount that is reflected, which allows the identification of features on the surface (Nasa, n.d.). Sensors then measure the sunlight's energy that is reflected and depending on their source of illumination they are categorized into active and passive sensors (Nasa, n.d.). While passive sensors use sunlight as source of illumination, active sensors have their own illumination source, enabling them to sense through cloud covers for example (Nasa, n.d.).

One determinant of the quality of remote sensing data that is often mentioned is resolution. The resolution can be determined along four types, which are radiometric, spatial, spectral, and temporal (Nasa, n.d.). Radiometric and spatial resolution refer to the pixels in the data. While radiometric resolution determines the amount of information per pixel, spatial resolution defines the range of the Earth's surface that is contained in one pixel (Nasa, n.d.). Spectral resolution defines the fineness of wavelengths, also called bands, with narrower bands allowing for more exact distinctions between minerals or vegetation types for example (Nasa, n.d.). Lastly, temporal resolution defines the time needed for a satellite to complete its orbit (Nasa, n.d.).

Emissions are one of many data types that can be gathered from satellite data, like deforestation, soil disturbances, raw material stockpiles, or water stressors (El-Jourbagy & Gura, 2022). For emissions, there are different scopes of emissions that can be measured. Scope 1 emissions can be directly linked to point sources like smokestacks over manufacturing facilities or heavy industrial sites, scope 2 emissions are emissions that are measured through energy purchases, and scope 3 emissions are downstream emissions, like supply chains or customers (El-Jourbagy & Gura, 2022). Emissions are usually referred to as greenhouse (GHG) emissions, which include NO2 and CO2 among others (EPA, 2023).

As a subcategory of alternative data, satellite data share some of the same advantages and challenges but they can vary in magnitude. On the one hand, sensor data is more difficult to process, as such data typically has a higher volume than data from individuals and business processes and is usually unstructured (Gladilin et al., 2023). Furthermore, in their 2022 paper El-Jourbagy & Gura note that there are more commercial than governmental satellites in space. The authors express concerns that the proprietary technology could make this data inaccessible. These drawbacks make satellite data expensive through the high acquisition and processing costs. On the other hand, El-Jourbagy and Gura (2022) argue that the cost of satellite data is expected to decrease as competition increases. Moreover, unlike signals derived from social media data, satellite data is more accurate and less prone to manipulation. Furthermore, remote sensing can help address current issues with pollution tracking via pollution monitoring stations that are scarce in developing countries for example and can be manipulated (Donaldson & Storeygard, 2016). Further advantages of satellite data are the access to data that was unavailable before, a higher resolution, and wide geographic coverage (Donaldson and Storeygard, 2016).

2.3 TROPOMI Instrument and Sentinel Satellites

The sentinel satellites are part of the Copernicus program, which is Europe’s Earth observation program (European Space Agency, n.d.-a; GIS Geography, 2023). The

Copernicus program is a joint initiative by the European Commission (EC) and the European Space Agency (ESA) (European Space Agency, n.d.-a). While the EC takes care of the organizational parameters of the initiative, ESA handles the data delivery from the satellites (European Space Agency, n.d.-a).

Sentinels is the name of the so called "satellite family" that was developed by ESA to service the needs of the Copernicus program (European Space Agency, n.d.-a). The sentinel family provides free of charge data from space (GIS Geography, 2023). There are various sentinel missions, with each of them focusing on specific observations of land, ocean, and atmosphere (European Space Agency, n.d.-b; GIS Geography, 2023). A detailed description of the missions can be found on ESA's homepage.

The data used in this thesis is collected through the Sentinel-5-Precursor (Sentinel-5-P) mission, which is the first Copernicus mission to measure the atmosphere (European Space Agency, n.d.-b). The mission provides high spatio-temporal resolution data of trace gases and aerosols that are affecting air quality and climate (European Space Agency, n.d.-c). Measurements are taken with a TROPOspheric Monitoring Instrument (TROPOMI) that is attached to the satellite (European Space Agency, n.d.-b). In this context it is important to note that there is a distinction between NO2 emissions and NO2 concentrations. While NO2 concentration is the amount of NO2 present in the air, NO2 emissions is the amount of NO2 released from a particular source. It is not possible to measure emissions directly from space. However, with the help of atmospheric models NO2 concentrations can be transferred to emissions.

The TROPOMI instrument is a passive sensor, measuring the solar radiation that is reflected by and radiated from the earth at the top of atmosphere (European Space Agency, n.d.-d). Since the instrument uses a passive remote sensing technique it is not able to pierce through cloud covers. The measurements can be categorized into three themes including air quality, stratospheric ozone layer, and climate change monitoring (European Space Agency, n.d.-d). For instance, the NO2 data belong to the category of air quality.

TROPOMI data are provided in NetCDF format files, which can be accessed using data analysis packages like MATLAB or python for instance (European Space Agency, n.d.-d). The NetCDF files are multidimensional files that are organized with three key values, comprising time, latitude, and longitude. Variables are collected in these dimensions, containing a large amount of data including air conditions in the atmosphere, and quality indicators of the data for example. From this, variables of interest can be accessed and indicators of data quality are delivered as a so called “qua_value”, which is provided alongside the variables of interest. The European Space Agency provides an overview of the types of data products collected and user manuals for these data products, including the nitrogen dioxide level 2 product that is used in this thesis (European Space Agency, n.d.-e).

3 Literature Review

This chapter explains how this thesis is positioned within the financial literature and provides an overview of results from other research papers related to the topic. The central research question of this thesis is whether company performance can be predicted with the satellite based NO2 variable. Satellite data is a subcategory of alternative data, which has been emerging over the past years in the finance industry as well as in financial research. Therefore, this thesis is related to finance literature on alternative data and its value for company performance prediction, as well as research on the use of satellite data in financial research. Additionally, the thesis relates to literature that focuses on linking NO2 emissions obtained from satellite data to point sources on the Earth's surface.

3.1 Alternative Data and Company Performance

One type of alternative data that has already achieved a high level of development is sentiment data obtained from social media platforms (Hansen & Borch, 2022). Social media sentiment analysis can be defined as the extraction of public sentiment from a large collection of social media data, such as for example textual data from individual posts, that are categorized into positive, negative, and granular steps in between to obtain an aggregate opinion about a company (Hansen & Borch, 2022). This is also reflected by the financial literature, as many research papers relating to alternative data analyse the value of user-generated data online. The usability of an alternative dataset largely depends on the availability of suitable machine learning methods (Hansen & Borch, 2022).

Technological advances enabled access to new data sources that have potential value for investment decisions, since they have been inaccessible before. The development of the internet facilitated the rise of various online platforms through which retail-investors and other non-professionals can express their opinions as well as access those of others. For instance, traditionally, investors had to rely on middlemen, usually finance professionals, to obtain information for investment purposes, but in the information age investors have access to new information resources (Bartov et al, 2018; Sheng, 2022). Individuals increasingly turn to non-professional peers online for advice and this trend can also be observed in the financial context (Chen et al, 2014). Platforms like Seeking Alpha or Estimize aim to supplement or even disrupt traditional research done by finance professionals through the provision of crowd-sourced information (Jame et al., 2016).

Much of the scientific literature regarding alternative data and company performance is based on the wisdom of crowds. The wisdom of crowd concept was first introduced by Surowiecki in 2004 and is since often applied in the financial markets (Halton, 2022). The idea behind wisdom of crowds is that a group of non-experts can hold more accurate information than a single expert when their individual opinions are accumulated to an average (Green et al, 2019). By obtaining an average knowledge from aggregated individual opinions, potential biases, which could be held by a single expert, are expected to be eliminated (Halton, 2022). Green et al. (2019) further define a typical wisdom of crowds setting in the context of financial research as the attempt of individuals to predict the financial performance of a company. The availability of individual opinions about companies on social media, combined with machine learning methods that enable the analysis of a large sample of textual data, makes it possible to access the wisdom of crowds in scientific research.

Scientific research highlights several new opportunities that are facilitated by the availability of such data. For example, social media data could help to uncover information that does not need to be disclosed by corporations but could potentially hold relevant information for investment purposes (Tang, 2018). Furthermore, reviews written by employees could hold insider information that is not accessible through traditional data sources (Hales et al., 2018). Additionally, employer rating platforms like Glassdoor comprise information about employee satisfaction that are closely tied to the economic condition of a firm (Green et al., 2019). Based on the definition of the wisdom of crowds setting by Green et al. (2019), information on these platforms can be divided into such that is intended for investment purposes, like opinions about stock prices or future earnings, and information that is unrelated to investment purposes such as product or company reviews.

There are several studies that aggregate individual investment related opinions expressed online to analyse their correctness or predictive power for a firm's financial performance. For example, Chen et al. (2014) focus on the accuracy of "peer opinions" expressed on social media. The authors investigate whether investment advice submitted by non-professionals through social media holds reliable information. For their analysis the authors use Seeking Alpha (SA) articles and comments between 2005 and 2012. SA is a platform on which retail investors can express their investment opinions through either the submission of articles or by commenting below these articles. Using textual analysis, Chen et al. (2014) extract the fraction of negative words from both, articles, and commentary on SA and investigate the impact on abnormal returns. In their main analysis the authors find that a higher fraction of negative words in SA articles as well as SA commentary is associated with lower abnormal returns. While Chen et al. (2014) analyse textual data, Jame et al. (2016) instead use a more structured form of alternative data, investigating the value of forecasts on Estimize. Estimize is a platform that allows professional as well as non-professional individuals to submit earnings forecasts for companies. The authors analyse the value of crowdsourced earnings forecasts, by testing whether they are useful for forecasting earnings, to measure the market's expectations of earnings, and whether they help price discovery in the financial markets. The study uses a sample of 37,031 forecasts submitted by 2,835 contributors for 1,601 firms on Estimize in 2012 and 2013. They find that Estimize forecasts are a valuable supplementary data source for the prediction of earnings, measuring the market's expectation of earnings, and they are helping price discovery. Similarly, Bartov et al. (2018) analyse the predictive power of firm-specific investment-related information on X. The authors test whether aggregate opinion from individual tweets can be used to predict quarterly earnings as well as the stock price reaction to the quarterly earnings. Additionally, they investigate the role of the information environment of a firm in the predictive power of X. Analysing a large sample of tweets about 3,604 distinct Russel 3000 firms between January 1, 2009, and December 31, 2012, Bartov et al. (2018) find that the aggregate opinion from X is useful for the prediction of earnings and announcement returns. Furthermore, the authors find that the importance of X for the prediction of announcement returns is higher for firms in weak information environments, but the effect is still positive and significant for firms in strong information environments. The results of these studies provide evidence for the predictive power of alternative data for companies' financial performance and demonstrate the relevance of alternative data in finance.

Another strand of literature uses crowdsourced information where individuals do not try to predict a company's financial performance, but the information is still believed to be useful for investment purposes. For instance, Hales et al. (2019) argue that employee reviews on Glassdoor contain insider information that can be useful for the prediction of firm fundamentals. The authors investigate the relationship between employee outlook, expressed through social media and income statement line items. The study uses a dataset of 158,352 Glassdoor reviews for 1,265 firms out of the S&P 1500 firms. The focus of the study lies on the employees' opinion of the prospects of the firm. To assess employee outlook, the authors differentiate between the explicit outlook measured as the rating in the "firm outlook" variable on Glassdoor, as well as latent employee outlook obtained via factor analysis of other rating categories on Glassdoor. In their main research, Hales et al. (2018) find that explicit as well as latent employee outlook predicts firm performance metrics including sales, gross margins, operating income, and net income. The authors further find a positive relationship between both measures of employee outlook and earnings surprises, concluding that the information in employee outlook is incremental to other information. Additionally, the study finds that explicit employee outlook is positively related to 6-month stock returns. A closely related study is done by Sheng (2022), also analysing the usefulness of employees' opinions about prospects of their employer for stock markets. While Hales et al. (2018) focus on the predictive power of Glassdoor reviews for firm-fundamentals, the study by Sheng (2022) focuses on the prediction of stock returns. Sheng (2022) uses a similar sample consisting of employee reviews on Glassdoor for US public firms, specifically the business outlook variable, for the time between March 2012 and December 2018. Firstly, the author analyses the impact of employee expectation on stock returns by using a long-short trading strategy based on changes in the employee outlook variable. Sheng (2022) finds that such a strategy results in an abnormal return between 8% and 11%. Secondly, the author investigates whether the outlook variable is related to firm-fundamentals. The results show that employee outlook is positively associated with subsequent earnings and changes in profitability, confirming the results of Hales et al. (2018). Furthermore, Sheng (2022) finds that employee outlook precedes hedge fund trading. The results of these studies provide evidence for the usefulness of employee outlook obtained from Glassdoor for the prediction of the financial performance of a company that is not covered by other sources of data which are available at the same time, such as analyst consensus. Moreover, the studies prove that even social media data that is not directly investment related contains useful information for investment purposes.

While employee outlook could still be viewed as an attempt to predict a company's financial performance, other studies utilize user-generated information that is completely unrelated to investment advice. For example, Huang (2018) argues that consumers are believed to hold relevant information about the quality and value of the products of a company, which can be useful for investment purposes. Through a systematic approach the author aims to analyse the value of consumer opinions for investment purposes, by investigating the relationship between online customer reviews and stock prices. The study uses a sample of 14.5 million reviews of 269,957 products from 250 public companies, submitted by customers online on Amazon.com in the time between 2004 and 2015. Huang (2018) computes an average star rating and takes the difference between the monthly average rating and the average rating in the 12-month period before to obtain a measure of abnormal returns. Subsequently, the author sorts the stocks by their abnormal return rating and creates portfolios consisting of stocks with a low and high abnormal return rating respectively. Huang (2018) finds that, when compared to a passive benchmark, the portfolio consisting of stocks with high abnormal ratings outperform the benchmark by 0.53% while the portfolio consisting of stocks with low abnormal ratings underperform the benchmark. A portfolio that is short on stocks with a low abnormal rating and long on stocks with a high abnormal rating outperforms the benchmark by 0.56% to 0.73%. Furthermore, the author finds that abnormal ratings predict earnings surprises and are linked to net purchases by hedge funds, which is in line with the results of Sheng (2022), suggesting that sophisticated investors trade on information contained in alternative data. Tang (2019) also analyses the predictive power of X commentary but, unlike Bartov et al. (2018) who investigate tweets that contain investment advice, the author focuses on information about products and brands submitted through individual tweets. Tang (2019) analyses whether firm-level aggregated tweets about the products and brand of a company contain useful information for the prediction of sales growth and unexpected sales growth respectively. The variables of the main research are purchase intent and the ratio of positive over non-neutral tweets. The purchase intent variable is measured as the number of tweets that indicate a purchase or the intent to purchase, while the second variable aims to capture the customer satisfaction. The study finds that an increase in tweets that indicate purchase or purchase intent is associated with an increase in sales growth and unexpected sales growth. However, the relationship between the fraction of positive tweets and sales growth as well as unexpected sales growth is statistically insignificant. This is in line with the results of Huang (2018) since product ratings on Amazon are typically connected to a purchase of the respective item.

Green et al. (2019) also analyse the information value of employee reviews on Glassdoor, but while Hales et al. (2018) analyse the predictive power of employee opinion about the future prospects of a firm, Green et al. (2019) focus on the value of employee satisfaction for the prediction stock returns. By using a dataset of more than one million employee reviews submitted on Glassdoor for 3,906 firms between June 2008 and June 2016, Green et al. (2019) analyse the relationship between employee reviews and stock return. For their main research the authors construct portfolios based on quarterly changes in employee ratings. They find that firms with improvements in employee ratings outperform firms with declines in employee ratings. Furthermore, the authors analyse different dimensions of employee satisfaction, finding that the effect is stronger for changes in the categories career opportunities and senior management and unrelated to work-life balance. As the variables career opportunities and senior management are closer related to the future performance of a firm than the work-life balance, these results indicate that employee opinion is more valuable for investment purposes for variables that are closer linked to the prospects of a firm. Additionally, the study finds that changes in employee reviews are predictive of firm fundamentals and earnings surprises. The results of these studies provide evidence that alternative data which is unrelated to investment purposes also contains useful information for the prediction of firms' financial performance, but it depends on the kind of data. For instance, consumer opinions hold relevant information only if they are linked to purchases, and employee opinions are more relevant when they are closely linked to a firm's prospects.

3.2 Satellite Data in Financial Research

A growing literature is using different types of satellite data to measure economic activity or well-being. Early applications of satellite data are mainly related to large-scale economic forecasting like the prediction of crop yields (Donaldson & Storeygard, 2016). One example of satellite data that has already been established as a proxy for economic activity is nightlight data (Donaldson & Storeygard, 2016). For instance, a recent study by Otchia and Asongu (2021) uses nightlight data to assess the industrial development in Africa. Another type of data that has been used for a long time already is Landsat data, which are medium resolution satellite data that measure the reflection of sunlight at different spectral bands (Goldblatt et al, 2020). Landsat data are already available since 1972 and have mainly been used for measuring forest cover, urban land use, or mineral deposits (Donaldson & Storeygard, 2016). In a recent study that compares the usefulness of Landsat imagery and nightlight data for measuring economic activity, Goldblatt et al. (2020) find that Landsat data outperforms nightlight data in predicting enterprise counts, employment, and expenditure. Furthermore, Goldblatt et al. (2020) were able to identify objects from Landsat data based on their spectral signature on the ground.

Alternative approaches for the assessment of economic conditions from space use a combination of satellite imagery and machine learning methods to detect objects from satellite data (Engstrom et al, 2022; Minetto et al. 2021). To measure human and economic activity during the Covid-19 pandemic, Minetto et al. (2021) apply machine learning techniques to detect specific objects such as planes or cars from satellite images. Engstrom et al. (2022) use a similar approach to extract features like the number and density of buildings or number of cars from high-resolution satellite imagery which they use to measure economic well-being. The authors demonstrate how different objects can be identified from satellite imagery. Although their research relates to a larger scale, the object detection opens new possibilities to monitor the activities of individual companies.

Anecdotal evidence suggests that satellite data contains useful information to monitor the performance of retailers and that sophisticated investors utilize this information (Partnoy, 2019). There is a growing literature using car count data of parking lots as a measure of firm-level performance. While the studies differ in research focus, they commonly prove that parking lot traffic is a useful predictor of retailer-level performance and that the use of satellite data enables investors to formulate profitable trading strategies. Research by Katona et al. (2018) and Kang et al. (2021) is related to the unequal access to alternative data and the capital market implications of this information asymmetry among market participants. Both studies validate and use car count data retrieved from satellite imagery as a proxy for retailer performance. Feng and Fay (2022) further analyse the relationship between parking lot traffic and retailer performance and additionally investigate the impact of different factors on this relationship.

Katona et al. (2018) suggest that the high acquisition, processing, and implementation costs of alternative data make these datasets only accessible to sophisticated investors, which causes an unequal access to alternative data by market participants. The focus of their study is the analysis of the impact of the introduction of satellite data on information asymmetry and trading activity in the stock market. The authors obtain pre-processed parking lot information extracted from satellite imagery from RS Metrics, an alternative data provider. The final sample includes 3.4 million daily observations for 53,647 distinct store locations for 44 US retailers between 2011 and 2017. RS Metrics provides two variables, the number of cars parked and the total number of available parking spaces for each individual store. From this data Katona et al. (2018) compute the quarterly numbers of parked cars and parking spaces, aggregate them at the corporate level, and subsequently obtain a parking lot fill rate by dividing the number of cars parked by the total parking spaces available for each retailer. The final variable of interest is a year-over-year growth in same store parking lot fill rates. In their main research Katona et al. (2018) use a difference-in-difference method to study the impact of the introduction of satellite data for matched groups of retailers with and without satellite coverage. They find that the introduction of satellite data causes an increase in short-selling activity and a decrease in informed buying activity of individual investors as well as for stock liquidity ahead of quarterly reports. Furthermore, they find that the introduction of satellite data has no impact on the speed of price discovery.

Another analysis of the impact of unequal access to satellite data is provided in a study by Kang et al. (2021) investigating whether institutional investors have better information about the performance of stores in their proximity. While prior research finds proof that local information advantage exists, the study by Kang et al. (2021) is the first to identify what kind of information drives this advantage. Like Katona et al. (2018), the authors use store-level car count information, which they obtained in pre-processed form from Orbital Insight, another supplier of alternative datasets, providing the quarterly car counts for each individual store in a sample of 92,668 stores from 71 public US retailers. From this data Kang et al. (2021) construct one of their main variables of interest which is the change in retailer-level car counts. In their main research Kang et al. (2021) analyse the relationship between changes in the holdings of institutional investors and changes in retailer-level car counts. The results show that institutional investors are adjusting their holdings based on local store performance. Furthermore, the authors test the profitability of trades based on local store information by comparing the abnormal returns of investors that trade on local information advantage versus those that trade against it. The results indicate that investors who trade on their local information advantage earn higher returns.

While the focus of the study by Katona et al. (2018) lies on the implications of the introduction of satellite data for financial markets, the authors also verify that satellite data is predictive of retailer performance. To validate the usefulness of satellite data for anticipating retailer performance, Katona et al. (2018) regress same-store-sales growth on same-store parking lot fill rates and control variables. The results show that a one standard deviation increase in parking lot fill rate is associated with a 0.8% increase in same-store sales growth. This effect is more pronounced for decreases in parking lot traffic. Furthermore, the authors find that a portfolio consisting of stocks with abnormal increases of parking-lot fill rate outperforms the market by 1.66%, and a portfolio consisting of stocks with abnormal decreases of parking lot fill rates underperforms the market by -3.10%. These results confirm that the magnitude of returns is higher for abnormal decreases in parking lot fill rates. Furthermore, the effect remains significant after accounting for short-selling costs. Similarly, Kang et al. (2021) validate the use of car count data, measured as the change in retailer-level car counts, as proxy for retailer-level performance by regressing performance metrics such as changes in sales and net income on retailer-level car count changes, finding a positive association. The results of these studies show that satellite data is predictive of retailer performance and the results remain the same for different measures of parking lot traffic.

Feng and Fay (2022) further investigate the predictive power of parking lot information obtained from satellite imagery for retailer performance. The study contributes to the previous literature by investigating the relationship between parking lot traffic and forward-looking retailer performance, as well as moderating factors of this relationship. The study measures retailer performance as Tobin’s Q, which is defined as the market value divided by assets for the respective retailer. Feng and Fay (2022) use pre-processed parking lot traffic data obtained from Orbital Insight. The study uses an unbalanced sample of 402 retailer-quarter observations between 2011 and 2019 for 33,848 distinct stores from 15 retailers. The provided data contains the number of parked cars as well as the size of the parking lot in square kilometres for each individual store of the sample. Feng and Fay (2022) use the same calculation as Katona et al. (2018) to compute their main variable of interest, the parking lot fill rate, by aggregating the number of cars parked and parking lot sizes at the retailer level and dividing the number of cars by the size of the parking lot. In their main analysis, the authors regress retailer performance on the parking lot fill rate and find a positive association between the two variables. Moreover, Feng and Fay (2022) find that this relationship is positively moderated by comparable store sales and store management intensity and negatively moderated by industry concentration. Industry concentration measures to which extend a few companies are dominating the industry they are in (Feng & Fay, 2022). Although the study focuses on the retailing industry, the results provide evidence for the usefulness of satellite data for the prediction of firm-level performance, as well as moderating factors of this effect.

The studies by Katona et al. (2018), Kang et al. (2021), and Feng and Fay (2022) demonstrate how information obtained from satellite data can be used to generate competitive advantages for investors that can access this data. These results indicate that other types of satellite data can be used similarly.

3.3 Developments in NO2 Emission Monitoring

The observation of NO2 levels and subsequent identification of emission point sources from space is a new emerging topic that has been initiated by technological advances that enable higher resolution satellite data. While most current satellites provide air pollution coverage with a spatial resolution that makes it impossible to trace the pollution to the source on the surface, the TROPOMI instrument attached to the Sentinel-5P satellite provides NO2 data with a high spatial resolution of 5 x 3.5km (Scheibenreif et al., 2021). The higher spatial resolution makes it possible to measure NO2 over smaller areas like oil fields, cities, and power plants (Martinez-Alonso et al, 2023). Currently, measurements of air pollution rely on data from ground measuring stations that lack in spatial coverage and data from satellites that lack in spatial resolution (Scheibenreif et al., 2021). The TROPOMI data addresses the drawbacks of current methods by providing a higher frequency and resolution (Beirle, 2019). For example, plumes of point sources can be seen on a single overpass (Beirle et al., 2019). However, Beirle et al. (2019) point out that the NO2 data provided by the TROPOMI instrument are “smeared out” in time-based averages because of changing wind patterns for instance, which is diminishing the higher resolution.

Different methods to address the drawbacks of the TROPOMI data are explored in scientific literature related to the observation of air quality and identification of emission point sources. For example, a study by Scheibenreif et al. (2021) utilizes a deep learning approach using TROPOMI NO2 data to develop a model that can estimate NO2 emissions from point sources from space. Specifically, the study uses NO2 concentration data measured on the ground by air quality stations operated by the European Environment Agency (EEA), as well as satellite measures of air pollution provided by the Sentinel-2 and Sentinel-5-P satellites from ESA’s Copernicus program. With these data inputs Scheibenreif et al. (2021) train a model that predicts air pollution only with satellite data. The authors first train the model with satellite images from the Sentinel-2 mission and further improve their prediction model by adding Sentinel-5P data. The results show that the new deep learning-based approach can estimate NO2 concentrations on the surface of the Earth only using satellite data.

Another approach to derive NO2 emissions and point sources using TROPOMI NO2 data is the divergence method. This method is essentially used to estimate where emissions are released and how they are distributed by taking the movement of pollutants, for example through wind patterns, into consideration. A thorough explanation of the divergence method can be found in Dix et al. (2022). A study by Beirle et al. (2019) uses the divergence method to identify different point sources of NO2 emissions within the city of Riyadh. Furthermore, Dix et al. (2022) estimate NOx (NOx = NO+NO2) emissions of six oil and gas production areas in the United States and compare them to the fuel-based oil and gas inventory. The study provides evidence for the usefulness of the divergence method to estimate NOx emissions. Furthermore, Beirle et al. (2021) developed a global catalogue of point sources of NOx emissions, comprising of 451 locations of power plants, metal smelters, and other industrial areas. All the studies use the TROPOMI NO2 data product as input data.

Martinez-Alonso et al. (2023) also use Sentinel-5-P data products to analyse the impact of mining on the air quality in the Copperbelt region. The research objective is to identify point sources of NO2 emissions from satellite data and to understand the causes of varying emission levels from these point sources. In a first step, point sources are identified by calculating annual means from daily TROPOMI NO2 tropospheric vertical column density (VCD) data. Afterwards, the authors apply the divergence method to calculate TROPOMI derived NO2 emissions, which were averaged to annual means. To understand variations in NO2 emissions from point sources, Martinez-Alonso et al. (2023) study the relationship between TROPOMI derived NO2 emissions and mine production data. They find a positive correlation between TROPOMI derived NO2 emissions and mine production, which are mine dependent and are sensitive to changes in the environment of the mine such as ore grade and fuel efficiency. Furthermore, Martinez-Alonso et al. (2023) compare the TROPOMI derived NO2 emissions to inventory emissions, obtained from annual reports of publicly traded companies, finding that inventory emissions underpredict mine emissions by 61-91 %. The authors suggest that TROPOMI NO2 data products are useful for the identification and monitoring of NO2 emissions from point sources in fossil-fuel intensive industries like the mining industry, which are more accurate than other sources of information like inventory emissions. The study provides evidence for the value of NO2 emission data for the prediction of mine productivity. The results indicate that NO2 levels may be useful to assess information about firm performance in the mining industry.

4 Data and Methodology

The goal of this thesis is to test whether the satellite based NO2 variable can be used to predict company performance. This thesis draws on different data sources for the construction of the final sample. Firstly, the NO2 concentration data originates from the ESA's Copernicus program. Secondly, the output data and ownership structure of the mines in the sample, as well as the financial variables for the firms that own the mines are collected from Refinitiv. After aggregating mine NO2 concentration at the firm-level of the companies that operate the mines and excluding firms with missing or incomplete financial data in Refinitiv, the final sample for testing the relationship between NO2 emissions and company performance comprises 224 firm-quarter observations from 14 distinct companies for the period of 2019Q1 until 2022Q4.

4.1 Data

The NO2 variable used in this thesis originates from the "nitrogendioxide_tropospheric_column" of the TROPOMI NO2 level 2 data product from the Sentinel-5-P mission of the ESA's Copernicus program. The data retrieval and pre-processing were done by the Finnish Meteorological Institute (FMI) comprising several steps. First, the NetCDF file for the TROPOMI NO2 data product is downloaded from the Copernicus Access Hub. Secondly, the variable of interest "nitrogendioxide_tropospheric_column" is retrieved. Thirdly, quality filtering is done with a qa_value threshold of 0.8 to remove areas that are covered by clouds for example, as well as errors and otherwise problematic retrievals. The retrieval is based on difference, meaning that signals that are considered noise, like light reflection from the surface of the Earth, the clouds, or the atmosphere, are removed. The provided dataset comprises monthly NO2 data from 47 mines located in various parts of the world in pre-processed form. The NO2 data are the total concentration of NO2 molecules in the vertical column, which is the area between the surface and the satellite, during one satellite pass in the unit mol/m^2.

As explained in chapter 2.3 there is a distinction between NO2 emissions and concentration. While NO2 emissions would likely yield better results, the goal of this thesis is to test whether the satellite based NO2 variable, representing NO2 concentrations, can be used to forecast company performance. Concentration data is more easily accessible and can be measured directly from space. Furthermore, Alonso-Martinez et al. (2023) find evidence that NO2 emissions are linked to output and the NO2 emissions in their study are derived from the NO2 data used in this thesis. The variables are directly linked, and it can be expected that there is a relationship between NO2 concentrations and company performance.

In a first step, quarterly averages for each mine were calculated from the monthly NO2 data. The dataset contains several mine-month values that were negative or zero. The negative values are caused in the retrieval process when background noise is subtracted. If a specific area has a low value of concentrations, the subtraction can cause negative values. Since the TROPOMI instrument is a passive sensor, it cannot sense through cloud covers. Thus, the most likely explanation for zero values is that a reading was not possible for some areas due to cloud cover. Therefore, the quarterly values were calculated as follows: in the case that there is one value available for a quarter, this value was chosen. When a quarter had two values, the average of the two values was calculated and if all values were available the quarterly average was computed. Two mines were excluded due to a substantial number missing values. Furthermore, five mines had missing values in a maximum of one quarter per year. For these mines the missing values were filled using the average value of the remaining quarters for the respective year. The dataset includes quarterly NO2 data for 45 mines from 2019Q1 until 2022Q4 comprising 720 mine-quarter observations.

In a second step, the companies operating the mines were identified using the mine screener in Refinitiv and selecting the dominant company respectively. The mines in the sample always have a distinct operating company, but the ownership structure varies. Firstly, some mines are directly owned by the operating company and thus were appointed to this company. Secondly, there are mines in which the ownership is shared, but the operating company holds the majority shares. Therefore, the mine was appointed to the operating company as the majority shareholder. Lastly, in some cases different companies hold equal shares in the mine. In these cases, the mine was appointed to the operating company. An overview of the respective ownership structure for each mine can be found in Appendix 1.

The primary objective of this thesis is to test the usability of the satellite based NO2 variable as a predictor of company performance, in the context of in using alternative datasets for investment purposes. To obtain a comprehensive understanding of the relationship between NO2 concentration and company performance, various dependent variables are employed to analyse both financial and operational aspects. To achieve this objective, separate analyses are conducted using the three dependent variables including return on assets (ROA), sales, and net income. These variables are essential components of fundamental analysis to inform investment decisions and are used in literature related to firm performance. The use of these variables allows to explore the relationship between NO2 concentration and various aspects of a company's financial performance, focusing on revenue generation, profitability, and overall financial health. Sales provides a direct and easily interpretable financial indicator, reflecting a company's ability to generate revenue from its core operation. ROA measures a company's profitability and efficiency in asset utilization. Net Income provides an overview of a company's financial performance, considering both revenue and expenses. This approach enables a thorough exploration of the relationship between NO2 emissions and company performance.

In a third step, the beforementioned performance metrics and control variables for the appointed firms were retrieved from Refinitiv on a quarterly basis. Firms with missing or incomplete financial information for all quarters were excluded from the sample. Since the financial metrics are available at the firm-level, the mine-level NO2 data were aggregated at the firm-level as shown in equation 1,

$$NO2\ Concentration_{i,t} = \frac{\sum_{m=1}^{n_{i,t}} NO2\ Concentration_{i,m,t} - \sum_{m=1}^{n_{i,t}} NO2\ Concentration_{i,m,t-1}}{\sum_{m=1}^{n_{i,t}} NO2\ Concentration_{i,m,t-1}} x100. \tag{1}$$

where i denotes the firm, m denotes the mine, and t denotes the quarter. The final sample for the regression analysis comprises 224 firm-quarter observations from 14 distinct public and private firms to analyse the relationship between the NO2 variable and firm financial performance.

Lastly, the output data for the mines in the sample was obtained from Refinitiv on a quarterly basis for the time from 2019Q1 until 2022Q4. The output data is grouped by output from general mining activities and solvent extraction/electrowinning (SX-EW) plants respectively. A SX-EW plant is a facility where valuable metals like copper are extracted using a chemical process followed by an electrowinning step that uses electricity to retrieve the pure metal (KGHM, 2023). Since not all mines in the sample have output from SX-EW plants, the SX-EW output was excluded from the analysis. The reason to exclude the output data from these operations is motivated by the need to maintain consistency and comparability in the dataset. Excluding the output from SX-EW plants ensures that the relationship between NO2 concentration and output remains straightforward. Mines without available output data were excluded from the sample. Since the control variables are available at the firm-level, the output data was aggregated at the firm-level as shown in equation 2,

$$Output_{i,t} = \frac{\sum_{m=1}^{n_{i,t}} Output_{i,m,t} - \sum_{m=1}^{n_{i,t}} Output_{i,m,t-1}}{\sum_{m=1}^{n_{i,t}} Output_{i,m,t-1}} x100. \tag{2}$$

where i denotes the firm, m denotes the mine, and t denotes the quarter. The final sample includes 208 firm-quarter observations for 13 distinct firms for the period from 2019Q1 until 2022Q4, for the analysis of the relationship between the NO2 variable and firm productivity. An overview of all variables with their respective definitions is presented in Table 1.

Table 1. Variable Definitions.

Variable	Definition
Firm level	
$\ln\_sales$	Natural Logarithm of Sales
$\ln\_net\_income$	Natural Logarithm of Net Income (Net Income on Which Basic EPS is Calculated)
ROA	(Net Income Before Preferred Dividends + ((Interest Expense on Debt-Interest Capitalized) *(1-Tax Rate)))/Average of Last Year's and Current Year's Total Assets *100
$size$	Natural Logarithm of Total Assets
$leverage$	Total Assets Divided by Total Liabilities
$current\_ratio$	Current Assets Divided by Current Liabilities
$\ln\_firm\_age$	Natural Logarithm Firm Age in Years
Mine level	
$No2\_emissions$	NO2 Concentration in mol/m^2
$Output$	Total Copper Output in Tonnes
Other	
$\ln\_Copper\ Price$	Natural Logarithm of LME-Copper Grade A Cash U$/MT

Table 2 presents the descriptive statistics of the underlying data for the main analysis of the correlation between NO2 concentration and firms' financial performance. The total number of mines, obtained from the Refinitiv database, represents the rounded count of copper mines in which a specific firm holds ownership shares, out of all copper mines.

The figure denotes the firms' share of the total copper mines, accounting for varying ownership percentages across multiple mines.

Table 2. Descriptive Statistics Full Sample.

Variable	Unit	N	Mean	SD	Min	Median	Max
NO2	mol/cm^2	224	2.43E+15	2.82E+15	1.92E+14	1.46E+15	1.34E+16
Sales	million USD	224	4,546.76	5,719.35	67.01	2,682.75	2,6800.00
Net Income	million USD	224	482.63	1,084.52	-3,026.77	224.06	7,990.42
ROA	ratio	224	6.03	6.53	-7.14	5.02	29.16
Size	log TA	224	16.86	1.06	14.05	17.17	18.29
Leverage	ratio	224	2.06	0.52	1.35	1.92	3.59
Firm Age	years	224	54.93	34.89	9.00	41.00	132.00
Nr. of Mines	mines	224	9.86	4.42	2.00	8.50	17.00

4.2 Methodology

Given the nature of the data and research objective, this thesis uses a quantitative research design to analyse the relationship between NO2 concentration and company performance. This approach is in line with previous research, and it helps to find patterns and relationships. Furthermore, quantitative research enables the generalization of findings to a larger population, meaning that results may have broader applicability in the mining industry, and can help to formulate general insights that can inform decision making. Other advantages of a quantitative research are the replicability, and the objectivity of quantitative data minimizes the potential for bias.

To test whether NO2 levels can be used to predict performance, this thesis uses a deductive approach. The choice is based on the logical assumption that increased production, which may lead to higher NO2 emissions, should be reflected in higher sales. For example, Martinez-Alonso et al. (2023) find a positive relationship between TROPOMI derived NO2 emissions and mine-production for mines in the Copperbelt region. While the hypothesis that NO2 concentration is positively correlated with sales is not derived from research that directly examines the relationship between NO2 concentration and company performance, it is reasonable to explore this hypothesis for several reasons. First, the measure of NO2 emissions used in the research by Alonso-Martinez et al. (2023) is derived from the satellite based NO2 variable used in this thesis. Second, higher productivity is expected to lead to higher sales. Lastly, the NO2 variable is easier available than NO2 emissions. Therefore, investigating the relationship between NO2 concentration and sales can help to better understand their potential correlation. A significant relationship between the satellite based NO2 variable and performance metrics would provide an alternative data source that can be used as an additional variable to inform investment decisions.

The underlying dataset is a panel dataset, where observations are structured in a nested form, with some firms owning multiple mines (Appendix 2). The NO2 data are available at the mine-level. As mentioned in the previous chapter, the mine-level NO2 data are aggregated at the firm-level. As a result, the nested panel structure is transformed into a balanced panel structure where each firm is represented as a unique entity with aggregated NO2 data. The aggregation of mine-level concentration at the firm level has two advantages. Firstly, it ensures a more straightforward analysis with the firm as the primary unit of observation. Secondly, it simplifies the data structure as the nested structure becomes less relevant for the analysis, allowing for a more focused exploration of the relationship between NO2 concentration and company performance metrics at the firm-level.

For the analysis of the relationship between NO2 and firm performance, a fixed effects model with clustered standard error is selected. Given the panel data structure, a fixed effects or random effects regression model would be appropriate. Since the firms are located in various regions, it can be assumed that there may be firm-specific effects that do not change over time and that cannot be observed. Examples of such effects include the political regime or company culture. To support the theoretical reasoning and determine the most appropriate model for the analysis, a Hausman test was conducted. The Hausman test is a statistical tool employed to assess whether the fixed effects model is more suitable compared to the random effects model by testing whether the error term is correlated with the explanatory variables (Brooks, 2019). The results of the Hausman test indicate that the fixed effects model is a better fit for the underlying dataset (Appendix 3). Therefore, the fixed-effects model will be applied to account for the firm-specific factors. Clustered standard errors are used in this analysis to account for the potential correlation between observations within firms, making sure that the within-firm variation is appropriately captured in the standard errors.

As the dependent variables sales and net income as well as the NO2 variable exhibited a substantial number of outliers and skewness, a logarithmic transformation was applied to these variables. In preparation of the logarithmic transformation in the regression analysis, a strategy was employed to address the negative values in the net income variable. Specifically, all net income values were increased by a constant amount. For this thesis, all net income values were increased by 10,000,000 as the minimum net income value was -3,026,771. This adjustment ensured that all net income values, including the initially negative ones, were transformed into positive values, allowing for the subsequent logarithmic transformation. The implementation of this approach is appropriate as it maintained the relative differences between the original values during the transformation process.

Hypothesis one (H1) states that there is a positive relationship between NO2 concentrations obtained from satellite data and company performance. This hypothesis is tested by regressing the performance variables sales, net income, and ROA on the NO2 variable. Equation 3 states the regression model used in this thesis, which is based on related research by Alvarez (2012) who analyses the impact of CO2 variation on company performance,

$$\mathrm{P}erformance_{it} = \beta_0 + \beta_1 * \mathrm{NO2\ Concentration}_{it} + \beta_2 * Size_{it} + \beta_3 Leverage_{it} + \beta_4 * Current\ Ratio_{it} + \beta_5 Firm\ Age_{it} + a_i + \varepsilon_{it}. \tag{3}$$

where $\mathrm{P}erformance_{it}$ denotes the performance of firm i in quarter t and a_i denotes firm specific effects capturing time-invariant variables of the firm. The variable $Performance_{it}$ includes the variables $\ln\_Sales_{it}$, ROA_{it} , $\ln\_Net\ Income_{it}$, and $\mathrm{NO2\ Concentration}_{it}$ is the natural logarithm of the NO2 concentration. The control variables are commonly used in related literature.

In the analysis of the relationship between different types of alternative data and company performance, various modelling approaches are explored. For example, Kang et al. (2021) use changes in the car count variable and performance measures to assess the relationship between parking lot traffic and retailer performance. Therefore, a subsequent analysis of the impact of changes in NO2 concentration on changes in sales outcome is performed. Including this analysis provides insights into the dynamics of the relationship between NO2 levels and sales and ensures the robustness of the results of the main regression. The regression model specification is the same as in equation 3, but instead of the logarithmic values, changes in NO2 levels and the performance metrics sales, net income, and ROA are calculated as presented in equations 4 to 7 respectively, where i denotes the firm and t is the fiscal quarter.

$$\Delta NO2\ Concentration_{i,t} = \frac{NO2\ Concentration_{i,t} - NO2\ Concentration_{i,t-1}}{NO2\ Concentration_{i,t-1}}\ x\ 100. \tag{4}$$

$$\Delta Sales_{i,t} = \frac{Sales_{i,t} - Sales_{i,t-1}}{Sales_{i,t-1}}\ x\ 100. \tag{5}$$

$$\Delta Net\ Income_{i,t} = \frac{Net\ Income_{i,t} - Net\ Income_{i,t-1}}{Net\ Income_{i,t-1}}\ x\ 100. \tag{6}$$

$$\Delta ROA_{i,t} = \frac{ROA_{i,t} - ROA_{i,t-1}}{ROA_{i,t-1}}\ x\ 100. \tag{7}$$

$$\Delta Output_{i,t} = \frac{(Output\_firm_{i,t}) - (Output\_firm_{i,t-1})}{(Output\_firm_{i,t-1})}\ x\ 100. \tag{8}$$

Hypothesis two (H2) states that there is a positive relationship between the satellite based NO2 variable and mine-output. To test this hypothesis output, measured as the natural logarithm of the total copper output in tonnes, is regressed on NO2 concentrations. Additionally, changes in output are regressed on changes on NO2 concentration. Changes in output are calculated as the difference between firm-level output a of the firm mine in the previous quarter over the firm-level output in the previous quarter, as shown in equation 8. Equation 9 states the regression model,

$$\mathrm{Output}_{it} = \beta_0 + \beta_1 * \mathrm{NO2\ Concentration}_{it} + \beta_2 * Size_{it} + \beta_3 Leverage_{it} + \beta_4 * Current\ Ratio_{it} + \beta_5 Firm\ Age_{it} + a_i + \varepsilon_{it}..(9)$$

where Output_{it} denotes the copper output of firm i in quarter t and a_i denotes firm specific effects capturing time-invariant variables of the firm. The variable $Output_{it}$ includes the variables ln_Output_{it} and $\Delta\mathrm{Output}_{it}$. Furthermore, $\mathrm{NO2\ Concentration}_{it}$ includes $\mathrm{ln}\_NO2Concentration\ _{it}$ and $\Delta\mathrm{NO2\ Concentration}_{it}$. Control variables comprise the same variables as in the analysis of the correlation between NO2 levels and financial performance metrics.

To ensure the robustness and consistency of the findings from the main analysis regarding the correlation between NO2 concentration and sales, several subgroup analyses were conducted by dividing the dataset based on key firm-specific characteristics. The chosen firm characteristics encompass the age of the firm, the total number of mines owned by a firm, and the firm size, measured as the natural logarithm of total assets. The reason behind the subgroup analyses lies in the assumption that the relationship between NO2 concentrations and firm performance may differ for firms based on these characteristics. Additionally, the subgroup analyses allow for a more nuanced understanding of the relationship between NO2 levels and firm-level performance metrics.

5 Results

This chapter provides a comprehensive analysis of the relationship between NO2 levels and firm performance, as well as firm productivity. The chapter is structured into three subchapters that progressively delve into the data further. The first subchapter provides an overview of the general condition of the data, utilizing boxplots, histograms, and scatter plots to visualize the distribution of the dependent variables and their relationship with the independent variable NO2. The second subchapter conducts a first univariate analysis to explore the significance of variables and analyse whether there are statistically significant differences between firms with different levels of NO2 concentrations. Lastly, the third subchapter utilizes multivariate regression models to test whether the results from the univariate tests hold.

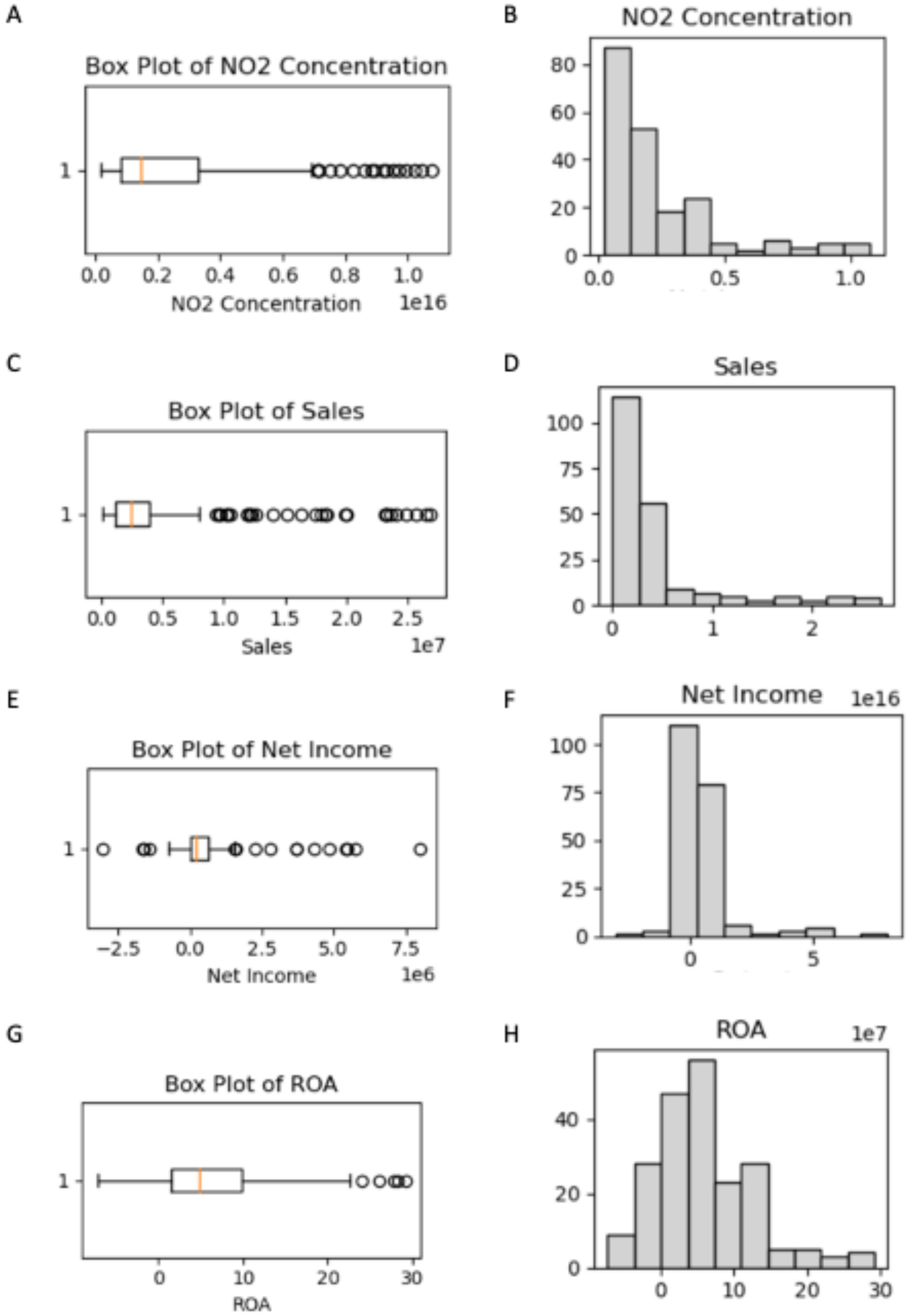
5.1 Data Exploration

Figure 1 displays boxplots and histograms of key variables used in the subsequent regression analyses. The boxplots enable the identification of central tendencies, variabilities, and outliers within the variables. In addition to the boxplots, histograms provide a detailed view of the distribution of each variable. Figure 2 presents scatter plots between the dependent variables and the independent variable NO2 concentration respectively, providing a first visualisation of the relationships that will be investigated further in the subsequent chapters. Together, these visualizations enhance the transparency of the findings by presenting the characteristics of the data and help to guide the subsequent data analysis.

Plots A and B in figure 1 show the distribution of NO2 concentration. The plots reveal that the variable is skewed to the right with outliers present at the higher end. Furthermore, the frequency of NO2 concentrations decreases as the value increases, indicating that there is a cluster of lower NO2 concentrations and a few large values which are scattered.

The sales variable is depicted in plots C and D of figure 1. The median is centred with an extended right whisker in the boxplot, together with a concentrated clustering of sales at the lower range in the histogram. The visualisations suggest that the distribution is skewed to the right with outliers on the higher end.

Figure 1.



Cont. Figure 1.

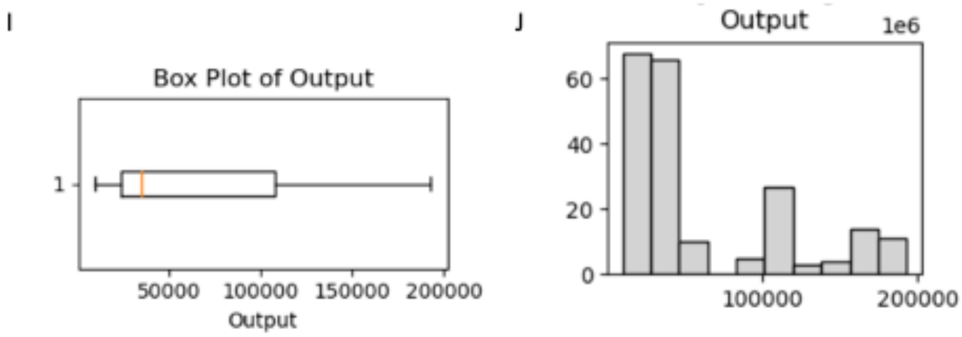


Figure 1. Boxplots and Histograms Dependent Variables, NO2 Concentration (Python).

Net income is presented in plots E and F of figure 1. The boxplot median is located to the left, and the histogram shows a notable peak on the left side. However, the boxplot shows multiple outliers to the left and slightly more to the right, indicating that most firms generate moderate net incomes.

Plots G and H of figure 1 capture the distribution of ROA. These plots show a more balanced distribution with fewer outliers. Despite a slight skewness to the right and a median that is slightly shifted to the left, the longer right whisker, and the long right tail in the histogram suggest that several firms achieve higher ROA values.

Finally, copper output is visualized in plots I and J of figure 1. The median in the boxplot is significantly shifted towards the left side, and the interquartile range is confirming the concentration of firms in the lower output range. The histogram shows multiple peaks, with the highest one on the left side. The peaks separate the histogram, indicating that there are subsets within the dataset with each cluster representing different scales of production levels. For subsequent analysis the distribution suggests that separate models for the clusters may be more appropriate. However, due to constraints in the sample size this is infeasible in this thesis.

In summary, the visualisations in figure 1 indicate various degrees of skewness across key variables, which can be challenging for linear regression analysis due to potential violations of the assumption of normality. Specifically, the variables NO2 concentration, sales, and net income are skewed to the right with a significant number of outliers at the higher end and for net income also at the lower end. ROA presents a moderately balanced distribution but with slight right skewness, and the variable output is also skewed to the right, containing multiple peaks indicating potentially distinct subgroups. In the subsequent analysis logarithmic transformation is applied to the variables NO2 concentration, sales, net income, and output, to address the skewness to the right and outliers of the variables.

Image A in figure 2 presents a scatter plot of NO2 concentration against sales. An apparent upward trend indicates that, on average, firms experience a gradual increase in sales corresponding to higher NO2 concentrations. However, the broad dispersion of data points around this trend suggests a relatively weak association overall. Notably, the data includes outliers that deviate from the general pattern of the rest of the data. These outliers correspond to a subset of observations of high sales figures despite lower NO2 concentrations, suggesting that there may be additional factors that could influence sales beyond the scope of NO2 levels. Nonetheless, for the majority of the sample, there remains a modest positive correlation between sales and NO2 concentrations, implying that higher levels of NO2 emissions may be linked with increased sales.

A scatter plot correlating NO2 concentration with net income is displayed in image B in figure 2. Initially, the datapoints suggest a trend, although upon closer inspection, the overall pattern is ambiguous. At first, a modest increase in net income is visible as NO2 levels increase, followed by a slight decline and subsequent rise again. These fluctuations indicate a complex relationship that may not be linear. Furthermore, at lower levels of NO2 concentrations, most datapoints cluster near an implied trend line with some variability. However, the consistency of these points diminishes with increasing levels of NO2 concentration. At higher NO2 levels, the data points spread out, indicating higher variability and a relationship that is departing from the general upward trend. Within the range of lower NO2 concentrations, some observations diverge significantly from the main cluster, suggesting that outliers are present. The relatively low number of these outliers suggest that they are probably caused by unique circumstances not reflected across the broader dataset.

Figure 2.

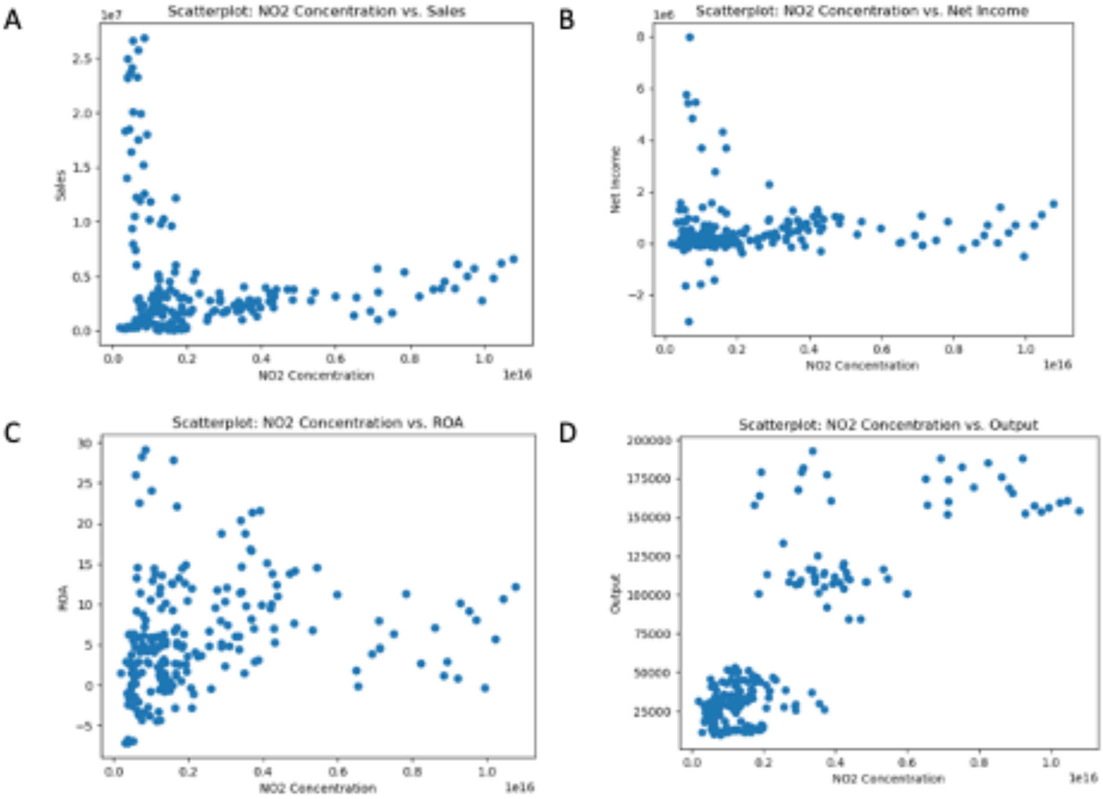


Figure 2. Scatterplots Dependent Variables with NO2 Concentration (Python).

Image C in figure 2 displays a scatter plot examining the relationship between NO2 concentrations and ROA. The plot reveals a tendency for ROA to increase alongside NO2 concentrations, although this trend is not consistent. At lower NO2 values, data points tend to cluster closely to the trend line, implying a stable relationship between NO2 concentration and ROA at these lower levels. However, as NO2 levels increase, the data starts to spread out more, indicating that the connection between higher NO2 concentration and ROA is not as strong. Most firms seem to have a moderate profit level when NO2 concentrations are not too high. Yet there is a small number of cases where NO2 levels are high, but ROA is still moderate. These are unusual and might not follow the general trend, suggesting that these particular cases exhibit unique circumstances that allow them to maintain profitability even with higher NO2 concentrations.

Figure 2, image D shows a scatter plot that relates NO2 concentrations to copper output. The plot indicates a noticeable trend suggesting that as NO2 concentration rises, so does the copper output. The data are aligned around this trend line and form three distinct clusters. The first, and most compact cluster, appears at the lower end of both NO2 concentrations and copper output. The other two clusters, though following the upward trend, are more spread out, suggesting greater variation in copper output at moderate and high NO2 levels. Additionally, there are a few outliers, especially at moderate NO2 levels, where copper output does not follow the general trend, indicating that other factors besides NO2 may be influencing output levels in those cases.

The scatter plots in figure 2 offer valuable insights into the relationships between NO2 concentrations and the variables of interest in the subsequent analysis. Firstly, higher NO2 levels generally seem to correlate with increased sales and output across firms. This is in support of the first hypothesis, stating that increased NO2 levels correlate with increased production levels which is reflected in higher sales. However, the widespread distribution of datapoints and the presence of outliers in the sales variable at low NO2 levels imply that NO2 may not be the only predictor of sales. Secondly, the scatter plot of NO2 and net income indicates a non-linear pattern. While there is an initial increase in net income with NO2 concentrations, the variability increases at higher NO2 levels, emphasizing the need for investigating other factors that might affect net income. Thirdly, the relationship between NO2 concentration and ROA seems to improve with higher NO2 concentrations up to a point after which the relationship weakens. The clustering of data points at the lower levels of NO2 suggests a more stable correlation in this range, but a spread at higher levels indicates possible diminishing returns or varying impacts of NO2 on firm profitability.

5.2 Univariate Analysis

Table 3 displays the descriptive statistics of the dataset for the analysis of the correlation between NO2 levels and financial performance, divided into two categories, low and high NO2 concentration, as divided by the median NO2 concentration. The table includes the dependent variables sales, net income, and ROA, as well as the control variables and the total number of mines owned.

Table 3. Descriptive Statistics Financial Performance Dataset.

Panel A: Low NO2 Concentration

Variable	Unit	N	Mean	SD	Min	Median	Max
NO2	mol/cm^2	112	7.28E+14	3.35E+14	1.92E+14	6.47E+14	1.44E+15
Sales	million USD	112	5,879.17	7,495.45	82.96	2,496.60	26,800.00
Net Income	million USD	112	495.52	1,374.41	-3,026.77	162.80	7,990.42
ROA	ratio	112	4.54	6.74	-7.14	3.93	29.16
Size	ln(total assets)	112	16.81	1.17	14.07	17.28	18.29
Leverage	ratio	112	2.08	0.63	1.35	1.90	3.59
Firm Age	years	112	48.95	27.99	9.00	43.50	106.00
Nr. of Mines	mines	112	7.74	4.02	2.00	8.00	15.00

Panel B: High NO2 Concentration

Variable	Unit	N	Mean	SD	Min	Median	Max
NO2	mol/cm^2	112	4.14E+15	3.16E+15	1.48E+15	3.04E+15	1.34E+16
Sales	million USD	112	3,214.36	2,439.71	67.01	2,705.65	12,200.00
Net Income	million USD	112	469.74	688.21	-1,415.09	339.54	4,305.17
ROA	ratio	112	7.52	5.97	-4.50	6.93	27.85
Size	ln(total assets)	112	16.91	0.93	14.05	17.16	18.14
Leverage	ratio	112	2.04	0.39	1.38	1.92	3.29
Firm Age	years	112	60.91	39.87	19.00	35.50	132.00
Nr. of Mines	mines	112	11.97	3.73	6.00	12.00	17.00

In panel A, which covers firms with low NO2 concentrations in a respective quarter, it can be observed that the average sales and net income are relatively high, at approximately 5,879.17 million USD and 495.54 million USD respectively. However, the standard deviations indicate substantial variability within this subgroup. Furthermore, firms in this category generally have lower ROA values and show a wide range of values as indicated by the high standard deviation and the range from the minimum to the maximum ROA. In comparison, panel B shows firms with high NO2 concentration having lower average sales (3,214.36 million USD) as well as a lower average net income (469.74 million USD) than those with low NO2 levels. However, the standard deviation for sales and net income is lower compared to firms with low NO2 concentrations. These differences in variation in the variables sales and net income between the subgroups reflect the spread-out data points that were observed for these variables in the scatter plots in the lower range of NO2 concentration. Interestingly, the ROA for firms in panel B is greater on average (7.52) than for those in panel A (4.54), suggesting a more favourable return on assets for the firms with higher NO2 values.

For further clarity on whether the differences between the subgroups are statistically significant, a t-test was conducted, with the results presented in Table 4. The t-test compares the means of the logarithmically transformed values for the variables sales and net income, and the untransformed ROA variable between the two groups.

Table 4. T-test: Differences Between Firms with High and Low NO2 Levels.

Variable	t_stat	p_value	mean1	mean2	sd1	sd2
Log Sales	0.289	0.773	14.660	14.610	1.499	1.054
Log Net Income	-0.212	0.832	16.159	16.162	0.114	0.062
ROA	-3.496	0.001	4.540	7.515	6.743	5.972

The t-test results for log sales and log net income show t-stat values of 0.289 and -0.212 with corresponding p-values of 0.773 and 0.832, respectively. These p-values are well above the conventional significance threshold of 0.05, indicating no statistical significance in the differences found between low and high NO2 firms with respect to sales and net income. This means that the variations observed in sales and net income in the descriptive statistics cannot be confidently attributed to the NO2 concentration levels.

However, the ROA presents a different picture. The t-test shows a t-stat value of -3.496 and a highly significant p-value of 0.001. This p-value, below the 0.05 threshold, indicates a statistically significand difference in ROA between firms with low and high NO2 intensity. This suggests that contrary to sales and net income, ROA is significantly affected by the firm's level of NO2 concentration.

As the standard deviations of the variables sales and net income revealed differences in variability for these variables between the subgroups, the assumption of equal variances, which underlies the standard t-test, may be violated. Therefore, an alternative approach was used to ensure the validity of the analysis. Specifically, the Welch's t-test was used, which is an adapted t-test that does not assume equal variation, was conducted. However, the results are the same in this alternative analysis.

Overall, while the descriptive statistics suggest lower sales and net income for firms with high NO2 concentrations compared to those with low NO2 concentrations, the t-test shows that only the differences in ROA are statistically significant. However, the results indicate that the outliers at low levels of NO2 concentration observed in the scatter plots of the previous chapter may affect the univariate analysis, even if the values are logarithmically transformed. In conclusion, these results point towards a more complex relationship between a firm's financial performance and their NO2 levels that may be influenced by additional factors not captured in the univariate analysis.

The correlation matrix, as presented in table 5, is included as an intermediate step before the regression analysis to identify potential relationships between key variables. Assessing the correlations between variables is necessary to anticipate potential multicollinearity issues and to inform the subsequent specification of the regression model.

Analysing the correlation matrix, sales shows a strong positive correlation with firm size (9.21, p<0.1), indicating that larger firms tend to have higher sales. Interestingly, sales correlates only modestly with NO2 concentration (0.125), suggesting that while there is some association between sales and NO2 concentration, it may not be strong enough to draw definitive conclusions about the nature of this relationship. Furthermore, net income displays a positive and strong correlation with ROA (0.654, p<0.1), indicating that firms with higher net income tend to also have a higher return on assets. This relationship is expected since both metrics are indicators of financial performance. Looking at the NO2 concentrations, the NO2 variable shows a very strong correlation with the total number of mines (0.613, p<0.1), which could indicate that firms with more mining sites tend to produce higher NO2 emissions. Therefore, this variable is relevant for the multiple regression model as it may have a significant role in explaining NO2 levels across firms.

Table 5. Correlation Matrix.

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
(1) Log Sales	1.000									
(2) Log NO2	0.125	1.000								
(3) Log Net Income	0.369*	0.056	1.000							
(4) ROA	0.317*	0.230*	0.654*	1.000						
(5) Size	0.921*	0.242*	0.326*	0.225*	1.000					
(6) Leverage	-0.424*	-0.058	-0.025	0.104	-0.337*	1.000				
(7) Current Ratio	-0.043	0.461*	0.081	0.237*	0.101	0.371*	1.000			
(8) Log Copper Price	0.145*	0.051	0.292*	0.404*	0.067	0.026	0.033	1.000		
(9) Log Firm Age	-0.235*	0.134*	0.162*	0.212*	-0.104	0.062	0.371*	0.035	1.000	
(10) Total Nr. of Mines	0.011	0.613*	-0.084	0.170*	0.131*	-0.169*	0.249*	0.000	0.148*	1.000

**** p<0.01, ** p<0.05, * p<0.1*

The negative correlation between leverage and sales (-0.424, p<0.1) suggests that firms with higher sales tend to have lower leverage ratios, which could be an indicator of better financial health or a different capital structure. Furthermore, the positive corelation between the current ratio and NO2 concentration (0.461, p<0.1) implies that firms with higher NO2 concentrations also maintain higher levels of liquidity. This could potentially serve as a safeguard in anticipation of environmental liabilities. Moreover, there is a significant link between the copper price and sales (0.145, p<0.1), net income (0.292, p<0.1), and ROA (0.404, p<0.1), which indicates that fluctuations in commodity prices can have an impact on firm performance. These results indicate that the copper price could be an important variable for the regression model. Lastly, the variable firm age shows diverse relationships with financial variables and could be an indirect indicator of a firm's maturity affecting its financial health and NO2 concentrations.

Table 6. Stepwise Regression.

	(1)	(2)	(3)	(4)	(5)
VARIABLES	model1	model2	model3	model4	model5
ln_no2	0.110**	0.113**	0.061**	0.052**	0.052**
	(0.048)	(0.049)	(0.024)	(0.020)	(0.020)
size	0.880***	0.877***	0.623***	0.657***	0.657***
	(0.136)	(0.134)	(0.100)	(0.093)	(0.093)
leverage	0.053	-0.034	-0.107	-0.098	-0.098
	(0.090)	(0.112)	(0.063)	(0.064)	(0.064)
current_ratio		0.060	0.046	0.049	0.049
		(0.041)	(0.032)	(0.032)	(0.032)
ln_copper			0.658***	0.761***	0.761***
			(0.150)	(0.105)	(0.105)
ln_firm_age				-0.992	-0.992
				(0.570)	(0.570)
o.nr_mines_total					-
Constant	-4.165*	-4.157*	-3.731**	-1.186	-1.186
	(2.105)	(2.106)	(1.425)	(1.602)	(1.602)
Observations	224	224	224	224	224
R-squared	0.427	0.433	0.655	0.668	0.668
Number of firm_id	14	14	14	14	14

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

In a stepwise regression analysis shown in table 6, control variables are introduced incrementally in each model. The dependent variable is always the natural logarithm of sales. Model 1 begins with the independent variables NO2 concentration, size, and leverage, with subsequent models adding current ratio (Model 2), log copper price (Model

3), and log firm age (Model 4). The variable total number of mines was omitted in the regression analysis, as indicated by the dash in Model 5.

Starting with model 1, it can be observed that NO2 concentration and size are significant predictors of sales. The results indicate that an increase in NO2 levels and size is associated with an increase in sales, which is consistent across all models. As additional controls are introduced in models 2 to 5, the values of the coefficient for NO2 diminishes slightly but remains at the same significance level. This suggests a genuine impact of NO2 concentration on the dependent variable sales. Furthermore, the variable size remains to have a strong positive influence that is significant and shows relatively stable coefficients. Surprisingly, leverage appears to be an insignificant predictor of sales, not confirming the results of the correlation matrix. The current ratio does not seem to have a significant impact on sales, which is in line with the results of the correlation matrix.

The copper price, introduced in model 4, emerges as a strongly significant variable, greatly improving the model's explanatory power, as reflected in a substantial increase in R-squared from 0.433 to 0.655. This indicates that the copper price should be included as a fixed effect to the regression, as it is the same for all companies and only varies over time. In line with the correlation matrix, the variable firm age shows a large negative coefficient in model 5, however it is not significant. The inclusion of firm age does not significantly change the R-square of the model, suggesting that firm age may not be a critical factor in the analysis of the relationship between NO2 concentration and sales. Overall, the models demonstrate an increasing fit, as indicated by the gradual increase in R-squared values from 0.427 in model 1 to 0.668 in model 4, highlighting an improved explanation of the variation in the dependent variable.

The descriptive statistics in table 7 outline the differences between two groups of firms categorized by their NO2 concentrations, following the same division as in table 3. It is important to note that the underlying dataset differs as one firm has been excluded due to missing copper output data. Panel A summarizes firms with low NO2 concentrations, averaging an output of 28,483 tonnes, while firms with high NO2 concentrations, shown in panel B, have a notably higher average output of 95,039 tonnes. Both groups are of comparable sizes based on the natural logarithm of total assets, with a marginal difference in mean values. Furthermore, firms with higher NO2 concentrations display a greater standard deviation in output, suggesting more variability within this subgroup.

Table 7. Descriptive Statistics Output Dataset.

Panel A: Low NO2 Concentration

Variable	Unit	N	Mean	SD	Min	Median	Max
Output	tonnes	104	28483.84	11310.20	9841.00	28762.00	53321.00
NO2	mol/cm^2	104	8.64E+14	3.43E+14	1.92E+14	8.30E+14	1.46E+15
Size	ln(total assets)	104	16.83	1.19	14.07	17.28	18.29
Leverage	ratio	104	2.13	0.60	1.37	1.98	3.59
Firm Age	years	104	54.95	32.04	9.00	52.00	109.00
Nr. of Mines	mines	104	7.40	3.82	2.00	8.00	15.00

Panel B: High NO2 Concentration

Variable	Unit	N	Mean	SD	Min	Median	Max
Output	tonnes	104	95039.00	58557.78	11366.00	108120.50	192862.00
NO2	mol/cm^2	104	3.95E+15	2.53E+15	1.48E+15	3.30E+15	1.08E+16
Size	ln(total assets)	104	16.86	0.98	14.05	17.12	18.14
Leverage	ratio	104	2.05	0.45	1.35	1.92	3.29
Firm Age	years	104	60.20	37.37	23.00	43.00	132.00
Nr. of Mines	mines	104	12.44	3.83	6.00	14.00	17.00

Leverage ratios are slightly lower on average for firms with high NO2 concentrations (2.05) compared to those with low NO2 concentration (2.13), indicating a moderate difference in financial risk. Moreover, firm age presents a higher mean for firm with high NO2 values (60.2 years) compared to those with low NO2 values (54.95 years), which indicates that established firms have higher NO2 emissions. Additionally, firms with higher NO2 concentration have a higher average number of total mines (12.44) than firms with lower NO2 concentration (7.40). This suggests that firms with higher emissions have a larger market share in the copper mining industry and this could correlate with higher overall production output, and consequently, higher NO2 emissions.

The values of the standard deviation of both groups are rather high, which could be caused by the wide distribution of the data points and outliers, as the actual values were used for the descriptive statistics. This is not surprising, considering the clusters in the scatter plot of NO2 concentration and copper output in the previous chapter. Nevertheless, there is a notable difference between the standard deviations for firms with low and high NO2 concentration, indicating differences in the variability of output between the two subgroups.

A univariate t-test was performed to test whether the differences in output between the averages of the two distinct groups are statistically significant. However, as the descriptive statistics revealed substantial differences in the variability of output between the two groups, the same approach as in the univariate analysis of the differences in the performance metrics was used. The results of the analysis indicate that the difference in copper output between firms with high and low NO2 concentrations is statistically significant (Appendix 4).

5.3 Multivariate Analysis

In this chapter the influence of NO2 concentration on company performance is analysed, employing a fixed-effects regression with clustered standard errors across all models. Table 8 presents the regression results for models 1 to 3, which explore the correlation between NO2 concentration and a firm’s financial performance, each with a different performance metric as the dependent variable. Specifically, the dependent variables examined in these models are sales, net income, and ROA respectively.

Model 1 demonstrates that there is a statistically significant relationship between NO2 concentration and sales, confirming the initial trend observed in the scatter plots between these variables. As the regression represents a log-log model, the coefficient can be interpreted as follows: a 1% increase in NO2 levels is associated with a 0.052% increase in sales. These results provide evidence that supports the first research hypothesis. Additionally, the firm size has a positive and significant effect on sales with a coefficient of 0.657 (P<0.01), suggesting that larger firms tend to have higher sales. Shifting the focus to net income, model 2 does not show a significant effect of NO2 concentration, suggesting that unlike sales, net income is not positively influenced by higher levels of NO2 concentration within this dataset. Lastly, model 3 shows that the relationship between NO2 concentration and ROA is positive and statistically significant with a coefficient of 1.144 (p<0.1), indicating that a 1% increase in NO2 levels is associated with a 1.144% increase in ROA. This result provides evidence to support the first hypothesis, indicating that higher NO2 emissions may correlate with an improved asset efficiency. Additionally, leverage is associated positively with ROA, suggesting that firms with a higher ratio of total debt to total assets might be earning higher returns on their assets.

Table 8. Regression of Sales, Net Income, and ROA on NO2 Concentration.

	(1)	(2)	(3)
VARIABLES	model1	model2	model3
ln_no2	0.052**	0.008	1.144*
	(0.020)	(0.013)	(0.610)
size	0.657***	-0.022	2.169
	(0.093)	(0.037)	(2.425)
leverage	-0.098	-0.020	5.531*
	(0.064)	(0.023)	(2.764)
current_ratio	0.049	0.020*	-0.486
	(0.032)	(0.010)	(1.137)
ln_copper	0.761***	0.134*	12.287***
	(0.105)	(0.074)	(4.027)
ln_firm_age	-0.992	-0.044	-10.412
	(0.570)	(0.079)	(6.542)
Constant	-1.186	15.242***	- 150.951***
	(1.602)	(0.380)	(27.644)
Observations	224	224	224
R-squared	0.668	0.142	0.352
Number of firm_id	14	14	14

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Furthermore, it is important to note that the copper price, as a time-fixed effect, has a positive and significant impact in all three models. These consistent findings suggest that copper prices can be considered a key driver of company performance, with higher prices translating into increased firm performance metrics. The results highlight the overall impact of this variable in the copper mining industry, where all firms are similarly affected by shifts copper prices.

The varying significance and magnitude of the NO2 coefficients across different performance metrics, indicate that the relationship between NO2 concentration and company performance is more complex than initially hypothesised. NO2 concentrations appear to be positively correlated with sales and ROA, while the effect on net income remains unclear. With R-squared values of 0.668, 0.142, and 0.352 for models 1, 2, and 3 respectively, the models explain a significant proportion of variability in sales and ROA but less for net income. In summary, the research hypothesis that NO2 concentration positively relates to company performance is supported in the case of sales and ROA but not conclusively for net income. Therefore, while NO2 could be considered a predictor of certain performance metrics, the evidence suggests that the relationship is contingent on the specific financial measure.

Table 9 shows the results of a further analysis of the relationship between changes in NO2 concentrations and changes in the performance metrics, after previously examining their log transformed values. In table 9, model 1 to 3 represent three separate regression models with the dependent variables being changes in sales, net income, and ROA, respectively. In each of these models the independent variables include changes in NO2 concentration along with other control variables. In model 4 to 6, the same dependent variables are shifted one period forward. This forward shift is aimed to analyse how future firm performance can be explained by the current period’s NO2 concentrations and control variables.

Table 9. Regression of Changes in Sales, NI, and ROA on Changes in NO2 Levels.

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	model1	model2	model3	model4	model5	model6
chg_no2	0.001	0.279	0.071	0.076***	-1.274*	-1.290
	(0.020)	(1.316)	(0.170)	(0.021)	(0.718)	(1.060)
size	-0.004	-0.068	-1.158*	-0.105*	-0.443	-2.040**
	(0.046)	(1.736)	(0.604)	(0.049)	(1.961)	(0.925)
leverage	-0.265***	1.041	-3.929**	-0.247	0.587	-3.219**
	(0.068)	(5.285)	(1.473)	(0.148)	(6.900)	(1.336)
current_ratio	0.044	-1.727	-0.277	0.038	-3.018	-1.502
	(0.038)	(2.353)	(0.576)	(0.048)	(3.076)	(0.935)
ln_copper	0.240***	3.101	2.509	-0.196*	-5.701	0.440
	(0.055)	(3.876)	(1.520)	(0.093)	(3.664)	(1.331)
ln_firm_age	-1.036	-19.302	-1.587	1.161***	13.157	5.738
	(0.588)	(17.106)	(3.649)	(0.361)	(10.831)	(6.683)
Constant	2.360*	48.520	11.939	-0.421	14.226	18.783
	(1.304)	(42.862)	(12.364)	(1.222)	(49.395)	(17.054)
Observations	210	210	210	196	196	196
R-squared	0.070	0.007	0.033	0.103	0.022	0.077
Number of firm_id	14	14	14	14	14	14

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The effect of changes in NO2 on all performance metrics is not significant in models 1 to 3 of table 9, which contradicts the positive relationship found when using the logarithmic values of sales (model 1) and ROA (model 3) of table 8. This indicates that while higher NO2 levels may be associated with higher sales and ROA, period-to-period fluctuations in NO2 do not have a clear predictive power for changes in sales and ROA. However, in models 4 through 6, where the dependent variables are shifted forward, NO2 concentration is positively associated with future sales (0.076) and this relationship is significant at the 1% level. These results further confirm the strong positive relationship between NO2 levels and sales. In contrast, NO2 levels are negatively associated with net income in the forward-shifted model, as presented in model 5, suggesting that current NO2 concentration has a negative impact on net income in the following period. The results indicate a lagged effect, where current NO2 concentrations impact future financial outcomes. The findings for net income provide evidence for the suggestion that increased

NO2 emissions could be a result of lower efficiency. Furthermore, size is a consistent negative predictor in the forward-shifted models, indicating that larger firms may not see immediate effects on financial performance compared to smaller firms.

The results of table 9 confirm the relevance of the NO2 variable as a potentially predictive variable for the future performance of firms when measured by sales, but do not provide consistent evidence for changes in net income or ROA. The negative impact of increased NO2 concentration on net income in the following quarter could indicate potential costs associated with higher emissions, such as increased operating costs or environmental fees. The results highlight the complexity of the relationship between NO2 concentrations and company performance. The results of tables 8 and 9 can be viewed as complementary, providing a comprehensive view of the impact of NO2 levels on firm performance, both in static and dynamic terms.

The robustness tests presented in table 10 are included to analyse the consistency of the correlation between sales and NO2 concentration, by dividing the dataset according to firm-specific characteristics such as age, size, and the total number of mines owned. While the categorisation by age and number of mines resulted in an equal number of firms for each group, the classification of firms as “small” and “large” is dynamic and can change over time. For example, a firm may be categorized as “small” in one period, but as the total assets increase, the same firm could be reclassified as “large” in a subsequent period. The firms were allowed to switch between size categories because the robustness tests are designed to determine whether the effect of NO2 concentration on sales varies with the firm’s size at a specific time, rather than if being categorized as “small” or “large” has a direct impact on the relationship. Furthermore, the analysis was strengthened by conducting a sensitivity analysis, which involved running regressions including only firms with a constant size classification throughout all quarters as shown in model 7 and 8.

Table 10. Robustness Tests.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	model1	model2	model3	model4	model5	model6	model7	model8
ln_no2	0.048**	0.070	0.077*	0.008	0.081**	-0.016	0.066*	0.062
	(0.018)	(0.049)	(0.034)	(0.056)	(0.026)	(0.043)	(0.028)	(0.050)
size	0.642***	0.687***	0.424***	1.069***	0.658***	0.326	0.443***	1.639**
	(0.123)	(0.130)	(0.045)	(0.224)	(0.089)	(0.243)	(0.042)	(0.370)
leverage	-0.079	-0.001	-0.210**	-0.081	-0.146**	0.047	-0.231**	-0.393***
	(0.101)	(0.124)	(0.064)	(0.157)	(0.052)	(0.190)	(0.073)	(0.077)
current_ratio	0.004	0.071	0.039	0.063	0.016	0.072	0.058	-0.026
	(0.051)	(0.041)	(0.035)	(0.065)	(0.026)	(0.083)	(0.042)	(0.052)
ln_copper	0.757**	0.827***	0.683***	0.656**	0.729***	0.769***	0.734***	0.699*
	(0.237)	(0.140)	(0.105)	(0.239)	(0.188)	(0.117)	(0.116)	(0.301)
ln_firm_age	-0.891	-2.480**	2.052**	-1.450**	-1.448***	0.985	1.809*	-1.167
	(0.569)	(0.710)	(0.715)	(0.466)	(0.172)	(1.600)	(0.783)	(0.664)
Constant	-1.559	3.923	-9.262***	-4.038	-0.065	-0.924	-8.828***	-16.951**
	(1.619)	(3.034)	(1.632)	(4.906)	(1.333)	(2.838)	(1.846)	(5.358)
Observations	112	112	112	112	112	112	96	80
R-squared	0.699	0.632	0.819	0.500	0.702	0.694	0.844	0.619
Number of firm_id	7	7	9	8	7	7	6	5

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Models 1 and 2 contrast the impact of NO2 concentration levels on sales between newly established and mature firms, respectively. The significant positive coefficient for NO2 in model 1 (0.048, p<0.05) underscores the primary analysis findings, suggesting that newly established firms experience an increase in sales at elevated NO2 levels, while the effect is statistically insignificant for mature firms. This could suggest that mature firms might have more established processes or infrastructure in place to mitigate the NO2 emissions from their operations.

When splitting the sample by the total number of mines, where model 5 represents firms that own fewer mines and model 6 includes firms with a greater number of mines, the outcomes diverge notably. For firms with fewer mines, NO2 concentration levels show a significant positive effect on sales (0.081, p<0.05). In contrast, NO2 levels have no effect on sales for firms that own more mines. This variation in results suggests that the impact of NO2 emissions on sales may be more pronounced for firms with less mines.

When dividing the sample by firm size, measured by the natural logarithm of total assets, a difference in the response to NO2 levels becomes apparent. Smaller firms, represented in model 3, show a statistically significant positive correlation between NO2 concentration and sales (0.077, p<0.1). This finding is consistent with the results of the main analysis. In contrast, larger firms, analysed in model 4, do not demonstrate a statistically significant correlation between NO2 levels and sales performance. This outcome is further supported by the sensitivity analysis performed in models 7 and 8, which include only firms with consistent size categories over time. The significant positive coefficients for NO2 concentration in these latter models reinforce the reliability of the previously observed effects.

In summary, the robustness tests validate the main analysis, confirming a positive correlation between NO2 concentration and sales, which is contingent on firm-specific characteristics. The high R-squared values across these models indicate that they explain a substantial portion of the variance in sales, highlighting the significance of NO2 concentration as a predictive variable.

Table 12 presents the results of regression analyses for models assessing the impact of NO2 concentration and other control variables on the copper output at the firm level. Model 1 employs a log-log specification, to analyse the relationship between NO2 concentration and copper output. The coefficient for NO2 is positive but not statistically significant. This suggests that there is no clear evidence to conclude that NO2 levels are associated with copper output in this model.

Table 11. Regression of Output on NO2 Concentration.

	(1)	(2)	(3)
VARIABLES	model1	model2	model3
ln_no2	0.057		
	(0.039)		
size	0.087	0.050	-0.015
	(0.064)	(0.066)	(0.035)
leverage	0.082	-0.190	-0.010
	(0.109)	(0.199)	(0.102)
current_ratio	0.002	0.104	0.039
	(0.040)	(0.059)	(0.023)
ln_copper	-0.013	0.046	-0.077**
	(0.106)	(0.102)	(0.035)
ln_firm_age	-0.254	-0.517	0.378*
	(0.312)	(0.815)	(0.186)
chg_no2		0.386	-0.002
		(0.236)	(0.015)
Constant	8.112***	0.896	-0.560
	(1.877)	(2.815)	(0.946)
Observations	208	207	194
R-squared	0.036	0.340	0.015
Number of firm_id	13	13	13

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Model 2 investigates the relationship between changes in copper output and changes in NO2 levels. The positive coefficient (0.386) of NO2 suggests that increases in NO2 are associated with increases in copper output. However, the significance level indicates that the results are not statistically strong enough to confirm this conclusion. Furthermore, model 3 is conceptually similar to model 2 but with the dependent variable shifted forward one period. This approach explores whether copper output is predicted by NO2 levels of the previous period. The NO2 coefficient is close to zero (-0.002) and not statistically significant.

When interpreting the results from table 12, it is important to consider several limitations of this study. Firstly, with data from only 13 firms the dataset is relatively limited in scope. Secondly, the observed clusters in the scatter plots between NO2 concentration and copper output suggest that different subgroups within the data might show different relationships between NO2 levels and copper output, complicating the specification of a model that fits for all firms. Thirdly, the moderate R-squared values across the models suggest that the variables included do not adequately explain the variation in the output variable. This could be an indication for the potential influence of unaccounted factors like ore grade or soil quality. Given these constraints, the results point toward certain trends that need further investigation but remains inconclusive in this analysis.

6 Conclusion

In light of the substantial increase of alternative datasets, there is a growing interest in leveraging such datasets for the prediction of company performance. While traditional data sources, such as financial reports, tend to be lagged, non-traditional datasets offer the potential to provide early insights into company performance. Consequently, there is a pressing need to explore alternative data sources to avoid falling behind the competition.

Notably, satellite data has emerged as a promising alternative data source. Although the predictive capability of some types of satellite data has been established, other available variables of satellite data remain largely unexplored. This thesis aims to investigate the potential of the satellite based NO2 variable in predicting company performance for firms in the mining industry. The specific focus is on assessing its impact on financial performance, with the goal of informing investment decisions. Additionally, the study seeks to assess the predictive power of the satellite based NO2 variable for mine output.

This analysis comprises two main components. Firstly, it involves investigating the relationship between NO2 levels, as indicated by the NO2 variable, and firm's financial performance metrics, such as sales, net income, and ROA. Secondly, the study explores the correlation between a firms NO2 levels and copper output. Consequently, this research aims to shed light on how satellite based NO2 data can be utilized to predict firm performance and inform investment decisions within the mining industry.

The scope of alternative data in financial markets is dynamic, continuously changing as datasets emerge or gain increased traction among investors. Alternative data typically comprises unstructured datasets that are not traditionally used in financial analysis but are assumed to offer valuable insights for investors. The "nitrogendioxide_tropospheric_column" from the TROPOMI NO2 level 2 data product, part of the European Space Agency's Copernicus program's Sentinel-5-P mission, falls within the category of alternative data.

Extensive literature exists on the exploration of alternative data sources to evaluate company performance, with much of it concentrating on sentiment analysis of text-based datasets from various social media platforms. Recent research has also uncovered the strong predictive power of satellite-derived car counts for assessing retailer performance. Leveraging higher spatial resolution of the TROPOMI instrument enables the identification of point sources of NO2 emissions, a feature that has been utilized in related literature to evaluate emissions from specific mines and other similar sources. These prior studies suggest that the NO2 variable holds promise for predicting firm performance.

The present study has yielded compelling evidence indicating the usefulness of the satellite based NO2 variable in predicting company performance. Multivariate analysis revealed a notable correlation between NO2 levels (measured as the quarterly average NO2 concentration) and sales, suggesting that a 1% increase in NO2 levels corresponds to a 0.052% increase in sales. Furthermore, the study demonstrated that changes in NO2 levels have a subsequent impact on sales outcomes in the following quarter. While a significant positive relationship between NO2 levels and ROA was observed, it was not consistently evident across all analyses. Additionally, the correlation between NO2 levels and net income was inconclusive. However, there is an indication that NO2 levels in the current quarter might have a negative impact on net income in the following quarter, needing further investigation. Robustness tests showed that the observed relationships were predominantly consistent for smaller firms (measured as the natural logarithm of total assets), firms with a smaller mine portfolio, and newly established firms.

The univariate analysis contrasting the copper output between firms with different NO2 levels revealed a significant difference, suggesting a correlation between higher NO2 levels and increased mine output. Nevertheless, the multivariate analysis did not yield definite outcomes, potentially attributed to constraints within this study. Therefore, the findings suggest certain trends in the relationship of NO2 levels and mine output that need further investigation.

The objective of this thesis was to test the correlation between NO2 levels and different financial performance metrics, as well as between NO2 levels and copper output. The first and second hypotheses of this thesis state that NO2 concentration is positively associated with a firm's financial performance and productivity respectively. These hypotheses were developed based on existing literature that analyses similar relationships. However, the usefulness of the satellite based NO2 variable has not been previously analysed.

The findings from this study indicate that the first hypothesis (H1) holds when considering specific performance metrics and accounting for firm-specific characteristics such as age, size, and number of mines owned. However, the nuanced nature of the relationship becomes evident as different types of performance metrics yield varying results. Given the relatively small dataset, the results for the performance metrics net income and ROA lack conclusiveness and warrant further investigation. Despite this, the results suggest the potential value of the NO2 variable as a supplementary data source for financial analysis. The results provide an initial exploration of the usefulness of the NO2 variable for the prediction of company performance, a research area that needs deeper exploration with a more comprehensive dataset.

The analysis of the correlation between NO2 levels and mine output did not conclusively address the second hypothesis (H2). While the initial analysis revealed a significant difference in outputs between mines with high and low NO2 levels, the multivariate analysis failed to confirm these results. This inconsistency between the univariate and multivariate analyses was unexpected, regarding the relationship of NO2 levels with both mine output and performance metrics. For instance, the univariate analysis of the differences between groups with high and low NO2 levels for the performance metrics indicated a statistically significant difference only for ROA. Possible explanations for these disparities could include the size of the dataset. Additionally, the inconclusiveness of the results of the multivariate analysis of NO2 levels and copper output could be explained by the impact of unobserved variables like ore grade and soil quality.

In conclusion, the findings of this thesis highlight the complexity of the relationship between NO2 levels and a company performance. Despite this complexity, the satellite based NO2 variable appears to be a significant predictor of firm performance, for certain financial metrics and contingent on firm-specific characteristics. These results provide insight into the potential use of the NO2 variable for predicting company performance, but caution is advised in making broad assumptions due to the data limitations of this study. Therefore, while these findings address the research question concerning the satellite based NO2 variable’s predictive capabilities, they call for careful interpretation and further investigation to generalize the results.

The research findings on the correlation between NO2 levels and company performance offer valuable practical implications for the finance industry. The demonstrated positive correlation between NO2 levels and sales, as well as the nuanced relationship with ROA, suggests potential applications of the NO2 variable for financial analysis. The evidence implies that utilizing satellite data, could enhance predictive models for firm performance, particularly for smaller and newly established firms. This could be especially useful to assess the performance of firms that are not publicly traded.

As the signal extracted from the NO2 variable exhibits an indication of relationships to several financial statement items, these insights have the potential to inform investment strategies and risk assessment processes within the finance industry. Additionally, the study’s inconclusive outcomes regarding the correlation between NO2 levels and mine output signal the need for further investigation, highlighting opportunities for innovative data-driven approaches in assessing firm performance in the mining industry.

From a policy perspective, these findings underscore the importance of considering alternative data sources in financial decision-making and regulatory frameworks. Policymakers could explore mechanisms to promote the incorporation of alternative data, such as satellite-based emission levels, into financial analysis and reporting practices. Additionally, efforts to enhance data quality and accessibility for such alternative data sources could further support the industry’s ability to leverage these insights effectively and reduce certain risks.

The results contribute to the current literature by investigating the predictive capability of a new alternative data source. This study provides a first exploration of the correlation between NO2 levels and firm performance. Thereby, it marks the first step toward a better understanding of this relationship, highlighting the need for further investigation. Future research can build upon these findings by incorporating larger datasets, additional variables, and diverse model specifications to deepen the understanding of this relationship.

In this study, several limitations are associated with data availability and quality. Firstly, the satellite derived NO2 variable measures NO2 concentration, while NO2 emissions from specific mines potentially offer a more accurate predictor of firm performance. However, NO2 emission data for copper mines was not available and obtaining NO2 emission data was not feasible, as deriving NO2 emissions from TROPOMI data through the divergence method requires the use of atmospheric models, extending beyond the scope of this thesis. Secondly, the NO2 data exhibited numerous missing values, particularly in areas with frequent cloud cover. Additionally, the variables in the dataset revealed a notable presence of outliers and skewness. Finally, the size of the dataset was limited due to the available time period of TROPOMI NO2 data and was further reduced by the exclusion of firms with missing or incomplete financial data.

These limitations underscore the need for caution when interpreting the study’s findings. The constraints related to data availability indicate potential challenges in fully capturing the complex relationship between NO2 levels and firm performance. As a result, the study’s ability to draw definite and generalizable conclusions may be limited.

References

AlternativeData.org. (n.d.). *Alternative Data*. AlternativeData.org. Retrieved September 26, 2023, from https://alternativedata.org/alternative-data/

Alvarez, I. G. (2012). Impact of CO $_2$ Emission Variation on Firm Performance. *Business Strategy and the Environment*, *21*(7), 435–454. https://doi.org/10.1002/bse.1729

Bartov, E., Faurel, L., & Mohanram, P. S. (2018). Can X Help Predict Firm-Level Earnings and Stock Returns? *The Accounting Review*, *93*(3), 25–57. https://doi.org/10.2308/accr-51865

Beirle, S., Borger, C., Dörner, S., Eskes, H., Kumar, V., De Laat, A., & Wagner, T. (2021). Catalog of NO $_x$ emissions from point sources as derived from the divergence of the NO $_2$ flux for TROPOMI. *Earth System Science Data*, *13*(6), 2995–3012. https://doi.org/10.5194/essd-13-2995-2021

Beirle, S., Borger, C., Dörner, S., Li, A., Hu, Z., Liu, F., Wang, Y., & Wagner, T. (2019). Pinpointing nitrogen oxide emissions from space. *Science Advances*, *5*(11), eaax9800. https://doi.org/10.1126/sciadv.aax9800

Brooks, C. (2019, March 28). *Introductory Econometrics for Finance*. Higher Education from Cambridge University Press; Cambridge University Press. https://doi.org/10.1017/9781108524872

Chen, H., De, P., Hu, Y. (Jeffrey), & Hwang, B.-H. (2014). Wisdom of Crowds: The Value of Stock Opinions Transmitted Through Social Media. *Review of Financial Studies*, *27*(5), 1367–1403. https://doi.org/10.1093/rfs/hhu001

Deloitte. (2017). *Alternative data for investment decisions. Today's innovation could be tomorrow's requirement* [Report]. Deloitte Center for Financial Services. https://www2.deloitte.com/tr/en/pages/financial-services/articles/fs-alternative-data-for-investment-decisions.html

Demchenko, Y., Los, W., & de Laat, C. (2018). *Data as economic goods: definitions, properties, challenges, enabling technologies for future data markets*. *2*.

Denev, A., & Amen, S. (2020). *The Book of Alternative Data: A Guide for Investors, Traders and Risk Managers*. John Wiley & Sons, Incorporated. http://ebookcentral.proquest.com/lib/tritonia-ebooks/detail.action?docID=6242905

Dix, B., Francoeur, C., Li, M., Serrano-Calvo, R., Levelt, P. F., Veefkind, J. P., McDonald, B. C., & De Gouw, J. (2022). Quantifying NO$_x$ Emissions from U.S. Oil and Gas Production Regions Using TROPOMI NO$_2$. *ACS Earth and Space Chemistry*, *6*(2), 403–414. https://doi.org/10.1021/acsearthspacechem.1c00387

Dix, B., Francoeur, C., Li, M., Serrano-Calvo, R., Levelt, P. F., Veefkind, J. P., McDonald, B. C., & de Gouw, J. (2022). Quantifying NOx Emissions from U.S. Oil and Gas Production Regions Using TROPOMI NO2. *ACS Earth and Space Chemistry*, *6*(2), 403–414. https://doi.org/10.1021/acsearthspacechem.1c00387

Donaldson, D., & Storeygard, A. (2016). The View from Above: Applications of Satellite Data in Economics. *Journal of Economic Perspectives*, *30*(4), 171–198. https://doi.org/10.1257/jep.30.4.171

El-Jourbagy, J., & Gura, P. P. (2022). In Space, No One Can Hear You're Green: Standardization of Environmental Reporting, the SEC 's Proposed Climate Change Disclosure Rules, and Remote Sensing Technology. *American Business Law Journal*, *59*(4), 773–820. https://doi.org/10.1111/ablj.12214

Engstrom, R., Hersh, J., & Newhouse, D. (2022). Poverty from Space: Using High Resolution Satellite Imagery for Estimating Economic Well-being. *The World Bank Economic Review*, *36*(2), 382–412. https://doi.org/10.1093/wber/lhab015

European Space Agency. (n.d.-a). *Europe's Copernicus programme*. European Space Agency. Retrieved September 21, 2023, from https://www.esa.int/Applications/Observing_the_Earth/Copernicus/Europe_s_Copernicus_programme

European Space Agency. (n.d.-b). *Sentinel missions*. European Space Agency. Retrieved, September 21, 2023, from https://sentinel.esa.int/web/sentinel/missions

European Space Agency. (n.d.-c). The Sentinel missions. European Space Agency. Retrieved, September 21, 2023, from https://www.esa.int/Applications/Observing_the_Earth/Copernicus/The_Sentinel_missions

European Space Agency. (n.d.-d). Sentinel-5P TROMOMI user guide. Sentinels Copernicus. Retrieved, November 04, 2023, from https://sentinels.copernicus.eu/web/sentinel/user-guides/sentinel-5p-tropomi

European Space Agency. (n.d.-e). S-5P Document library. Sentinels Copernicus. Retrieved, November 04, 2023, from https://sentinels.copernicus.eu/web/sentinel/user-guides/sentinel-5p-tropomi/document-library

Feng, C., & Fay, S. (2022). An empirical investigation of forward-looking retailer performance using parking lot traffic data derived from satellite imagery. *Journal of Retailing*, *98*(4), 633–646. https://doi.org/10.1016/j.jretai.2022.03.004

GIS Geography. (August 9, 2023). *Sentinel satellites of the Copernicus programme*. https://gisgeography.com/sentinel-satellites-copernicus-programme/

Gladilin, A., Hoang, K. T., Williams, G., & Sadik, Z. (2023). Sensors Data. In *Handbook of Alternative Data in Finance, Volume I*. Chapman and Hall/CRC.

Goldblatt, R., Heilmann, K., & Vaizman, Y. (2020). Can Medium-Resolution Satellite Imagery Measure Economic Activity at Small Geographies? Evidence from Landsat in Vietnam. *The World Bank Economic Review*, *34*(3), 635–653. https://doi.org/10.1093/wber/lhz001

Green, T. C., Huang, R., Wen, Q., & Zhou, D. (2019). Crowdsourced employer reviews and stock returns. *Journal of Financial Economics*, *134*(1), 236–251. https://doi.org/10.1016/j.jfineco.2019.03.012

Hales, J., Moon, J. R., & Swenson, L. A. (2018). A new era of voluntary disclosure? Empirical evidence on how employee postings on social media relate to future corporate disclosures. *Accounting, Organizations and Society*, *68–69*, 88–108. https://doi.org/10.1016/j.aos.2018.04.004

Halton, C. (2022, July 31). *Wisdom of crowds. Definition, theory, examples*. Investopedia. https://www.investopedia.com/terms/w/wisdom-crowds.asp

Hansen, K. B., & Borch, C. (2022). Alternative data and sentiment analysis: Prospecting non-standard data in machine learning-driven finance. *Big Data & Society*, *9*(1), 205395172110707. https://doi.org/10.1177/20539517211070701

Huang, J. (2018). The customer knows best: The investment value of consumer opinions. *Journal of Financial Economics*, *128*(1), 164–182. https://doi.org/10.1016/j.jfineco.2018.02.001

Jame, R., Johnston, R., Markov, S., & Wolfe, M. C. (2016). The Value of Crowdsourced Earnings Forecasts: THE VALUE OF CROWDSOURCED EARNINGS FORECASTS. *Journal of Accounting Research*, *54*(4), 1077–1110. https://doi.org/10.1111/1475-679X.12121

Kang, J. K., Stice-Lawrence, L., & Wong, Y. T. F. (2021). The Firm Next Door: Using Satellite Images to Study Local Information Advantage. *Journal of Accounting Research*, *59*(2), 713–750. https://doi.org/10.1111/1475-679X.12360

Katona, Z., Painter, M., Patatoukas, P. N., & Zeng, J. (Jieyin). (2018). *On the Capital Market Consequences of Big Data: Evidence from Outer Space* (SSRN Scholarly Paper 3222741). https://doi.org/10.2139/ssrn.3222741

KGHM Polska Miedz S.A. (n.d). *SX-EW (solvent extraction and electrowinning)*. KGHM Polska Miedz. Retrieved November 9, 2023, from https://kghm.com/en/our-business/processes/sx-ew

Martinez-Alonso, S., Veefkind, P., Dix, B. K., Gaubert, B., Theys, N., Granier, C., Soulié, A., Darras, S., Eskes, H., Tang, W., Worden, H. M., Gouw, J. A. D., & Levelt, P. F. (2023). *TROPOMI-derived NO2 emissions from copper/cobalt mining and other industrial activities in the Copperbelt (DRC and Zambia)* [Preprint]. Preprints. https://doi.org/10.22541/es-soar.168286778.87446295/v1

Mehmood Mirza, F., Sinha, A., Rehman Khan, J., Kalugina, O. A., & Wasif Zafar, M. (2022). Impact of energy efficiency on CO2 Emissions: Empirical evidence from developing countries. *Gondwana Research*, *106*, 64–77. https://doi.org/10.1016/j.gr.2021.11.017

Minetto, R., Segundo, M. P., Rotich, G., & Sarkar, S. (2021). Measuring Human and Economic Activity From Satellite Imagery to Support City-Scale Decision-Making During COVID-19 Pandemic. *IEEE Transactions on Big Data*, *7*(1), 56–68. https://doi.org/10.1109/TBDATA.2020.3032839

Mitra, G., Erlwein-Sayer, C., Hoang, K. T., Roman, D., & Sadik, Z. (Eds.). (2023). *Handbook of Alternative Data in Finance, Volume I*. Chapman and Hall/CRC. https://doi.org/10.1201/9781003293644

National Aeronautics and Space Administration (NASA). (2017, August 7). *What is an orbit*. https://www.nasa.gov/audience/forstudents/5-8/features/nasa-knows/what-is-orbit-58.html

National Aeronautics and Space Administration (NASA). (n.d.). *What is remote sensing*. NASA. Retrieved September 26, 2023, from https://www.earthdata.nasa.gov/learn/backgrounders/remote-sensing

Otchia, C., & Asongu, S. (2021). Industrial growth in sub-Saharan Africa: Evidence from machine learning with insights from nightlight satellite images. *Journal of Economic Studies*, *48*(8), 1421–1441. https://doi.org/10.1108/JES-05-2020-0201

Partnoy, F. (2019, April 16). *Stock Picks From Space*. The Atlantic. https://www.theatlantic.com/magazine/archive/2019/05/stock-value-satellite-images-investing/586009/

Royal Netherlands Meteorological Institute. (2022). *Sentinel-5P Level 2 Product User Manual for Nitrogen Dioxide* (Document No. S5P-KNMI-L2-0021-MA). Royal Netherlands Meteorological Institute, Ministry of Infrastructure and Water Management.

Scheibenreif, L., Mommert, M., & Borth, D. (2021). *Estimation of Air Pollution with Remote Sensing Data: Revealing Greenhouse Gas Emissions from Space* (arXiv:2108.13902). arXiv. http://arxiv.org/abs/2108.13902

Sheng, J. (2022). *Asset Pricing in the Information Age: Employee Expectations and Stock Returns* (SSRN Scholarly Paper 3321275). https://doi.org/10.2139/ssrn.3321275

Tang, V. W. (2018). Wisdom of Crowds: Cross-Sectional Variation in the Informativeness of Third-Party-Generated Product Information on X: WISDOM OF CROWDS. *Journal of Accounting Research*, *56*(3), 989–1034. https://doi.org/10.1111/1475-679X.12183

Teoh, S. H. (2018). The promise and challenges of new datasets for accounting research. *Accounting, Organizations and Society*, *68–69*, 109–117. https://doi.org/10.1016/j.aos.2018.03.008

United States Environmental Protection Agency (EPA). (2023, August 21). Scope 1 and scope 2 inventory guidance. https://www.epa.gov/climate-leadership/scope-1-and-scope-2-inventory-guidance

Appendices

Appendix 1. Ownership Structure of Copper Mines

mine_name	Operating Company	Mine/Plant Shareholder1	% owned1	Mine/Plant Shareholder2	% owned2	Mine/Plant Shareholder3	% owned3
Aitik	Boliden AB	Boliden AB	100	0	0	0	0
Aktogay	Kaz Minerals Ltd	Kaz Minerals Ltd	100	0	0	0	0
Almalyk	Almalyk Mining And Metallurgical Complex	Uzbekistan, Republic of (Government)	97.5	Various	2.5	0	0
Andina	Corporacion Nacional del Cobre (Codelco)	Corporacion Nacional del Cobre (Codelco)	100	0	0	0	0
Antamina	Compania Minera Antamina SA	Glencore PLC	33.75	BHP Group (UK) Ltd	33.75	Teck Resources Ltd	22.5
Antapaccay	Glencore PLC	Glencore PLC	100	0	0	0	0
Bagdad	Freeport-McMoRan Inc	Freeport-McMoRan Inc	100	0	0	0	0
Bingham Canyon	Kennecott Utah Copper LLC	Rio Tinto PLC	100	0	0	0	0
Bozshakol	Kaz Minerals Ltd	Kaz Minerals Ltd	100	0	0	0	0
Buenavista (Cananea)	Southern Copper Corp	Southern Copper Corp	88.91	Various	11.09	0	0
Caserones	Nippon Mining & Metals Co Ltd	Nippon Mining & Metals Co Ltd	75	Mitsui Mining and Smelting Co Ltd	25	0	0
Centinela (second concentrator)	Antofagasta PLC	Antofagasta PLC	70	Marubeni Corp	30	0	0
Cerro Verde	Freeport-McMoRan Inc	Freeport-McMoRan Inc	53.56	SMM Cerro Verde Netherlands BV	21	Compania de Minas Buenaventura SAA	19.58
Chapada	Lundin Mining Corp	Lundin Mining Corp	100	0	0	0	0
Cobre Panama	First Quantum Minerals Ltd	First Quantum Minerals Ltd	90	Korea Resources Corp	10	0	0
Collahuasi	Anglo American PLC	Glencore PLC	44	Anglo American PLC	44	Japan Collahuasi Resources BV	12
Constancia	Hudbay Minerals Inc	Hudbay Minerals Inc	100	0	0	0	0
Cuajone	Grupo Mexico SAB de CV	Grupo Mexico SAB de CV	88.91	Various	11.09	0	0

Dexing	Jiangxi Copper Co Ltd	Jiangxi Copper Co Ltd	100	0	0	0	0
Duobaoshan	Zijin Mining Group Co Ltd	Zijin Mining Group Co Ltd	100	0	0	0	0
El Abra	Freeport-McMoRan Inc	Freeport-McMoRan Inc	51	Corporacion Nacional del Cobre (Codelco)	49	0	0
Erdenet	Erdenet Mining Corp	Mongolia (Government)	51	Mongolian Copper Corp	49	0	0
Gibraltar	Taseko Mines Ltd	Taseko Mines Ltd	87.5	Cariboo Copper Corp	12.5	0	0
Grasberg	Freeport-McMoRan Inc	Indonesia Asahan Aluminium (Persero) PT	51.2	Freeport-McMoRan Inc	48.8	0	0
Highland Valley Copper	Teck Resources Ltd	Teck Resources Ltd	100	0	0	0	0
Kansanshi	Kansanshi Mining Plc	First Quantum Minerals Ltd	80	ZCCM Investments Holdings PLC	20	0	0
La Caridad	Southern Copper Corp	Southern Copper Corp	88.91	Various	11.09	0	0
Las Bambas	MMG Ltd	MMG Ltd	62.5	CNIC Corporation Ltd	22.5	CITIC Ltd	15
Los Bronces	Anglo American PLC	Anglo American PLC	50.1	Corporacion Nacional del Cobre (Codelco)	29.5	Mitsubishi Materials Corp	20.4
Los Pelambres	Antofagasta PLC	Antofagasta PLC	60	Nippon Mining & Metals Co Ltd	15.79	Mitsubishi Materials Corp	10
Lumwana	Barrick Gold Corp	Barrick Gold Corp	100	0	0	0	0
Mirador	TongLing Nonferrous Metals Group Holding Co Ltd	TongLing Nonferrous Metals Group Holding Co Ltd	70	China Railway Construction Corp Ltd	30	0	0
Mission	ASARCO LLC	Grupo Mexico SAB de CV	100	0	0	0	0
Morenci	Freeport-McMoRan Inc	Freeport-McMoRan Inc	72	Sumitomo Metal Mining Co Ltd	28	0	0
Mount Milligan	Centerra Gold Inc	Centerra Gold Inc	100	0	0	0	0
Ok Tedi	Ok Tedi Mining Ltd	Papua New Guinea, Independent State of (Government)	100	0	0	0	0

Oyu Tolgoi	Rio Tinto PLC	Rio Tinto PLC	66	Mongolia (Government)	34	0	0
Pinto Valley	Capstone Copper Corp	Capstone Copper Corp	100	0	0	0	0
Quebrada Blanca	Teck Resources Ltd	Teck Resources Ltd	90	Empresa Nacional de Mineria	10	0	0
Qulong	Zijin Mining Group Co Ltd	Zijin Mining Group Co Ltd	50.1	Others	49.9	0	0
Safford District	Freeport-McMoRan Inc	Freeport-McMoRan Inc	100	0	0	0	0
Salobo	Vale SA	Vale SA	100	0	0	0	0
Sierra Gorda	KGHM Polska Miedz SA	KGHM Polska Miedz SA	55	South32 Ltd	45	0	0
Sierrita	Freeport-McMoRan Inc	Freeport-McMoRan Inc	100	0	0	0	0
Toquepala	Grupo Mexico SAB de CV	Grupo Mexico SAB de CV	88.91	Various	11.09	0	0
Sentinel	First Quantum Minerals Ltd	First Quantum Minerals Ltd	100	0	0	0	0
Wunuketushan	China National Gold Group Co Ltd	China National Gold Group Co Ltd	100	0	0	0	0

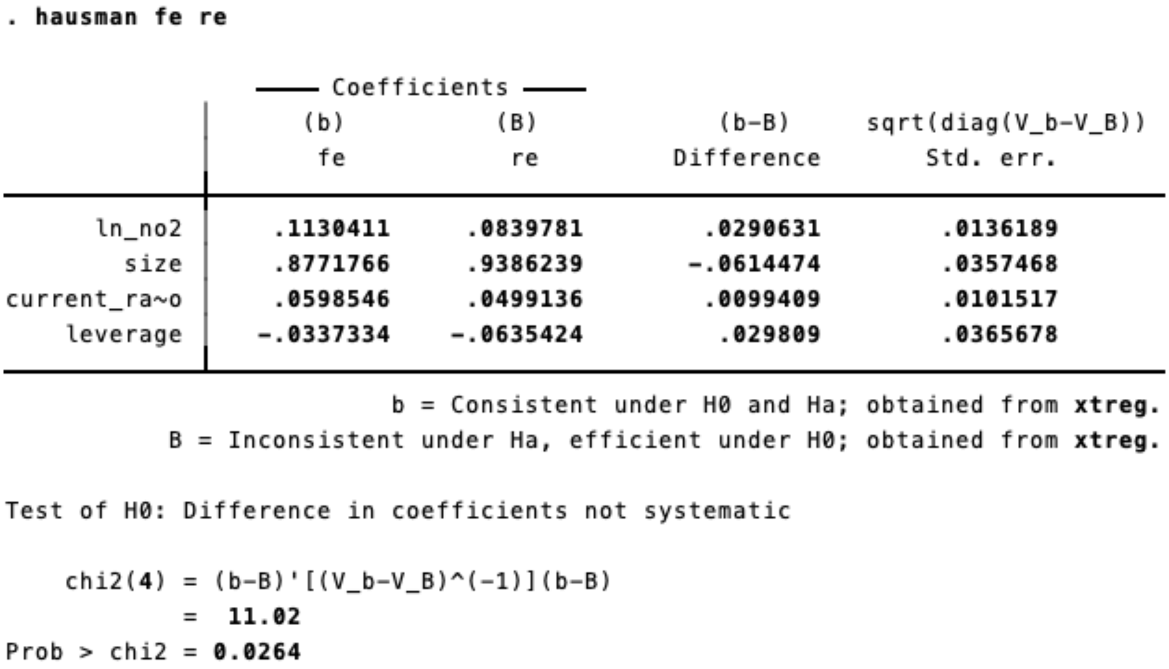
Appendix 2. Overview of Mines per Firm

Firm	Mines in Sample
Barrick Gold Corp	Lumwana
Capstone Copper Corp	Pinto Valley
Corporacion Nacional del Cobre (Codelco)	Andina
ENEOS Holdings Inc	Caserones
First Quantum Minerals Ltd	Cobre Panama
	Kansanshi
	Sentinel
Freeport-McMoRan Inc	El Abra
	Cerro Verde
	Sierrita
	Bagdad
	Safford District
	Morenci
Grupo Mexico SAB de CV	Cuajone
	Toquepala
	Mission
Hudbay Minerals Inc	Constancia
KGHM Polska Miedz SA	Sierra Gorda
Lundin Mining Corp	Chapada
Southern Copper Corp	La Caridad
	Buenavista (Cananea)
Teck Resources Ltd	Highland Valley Copper
	Quebrada Blanca
Vale SA	Salobo
Zijin Mining Group Co Ltd	Duobaoshan
	Qulong

Appendix 3. Hausman Test Results

Hausman (1978) specification test

	Coef.
Chi-square test value	11.016
P-value	.026



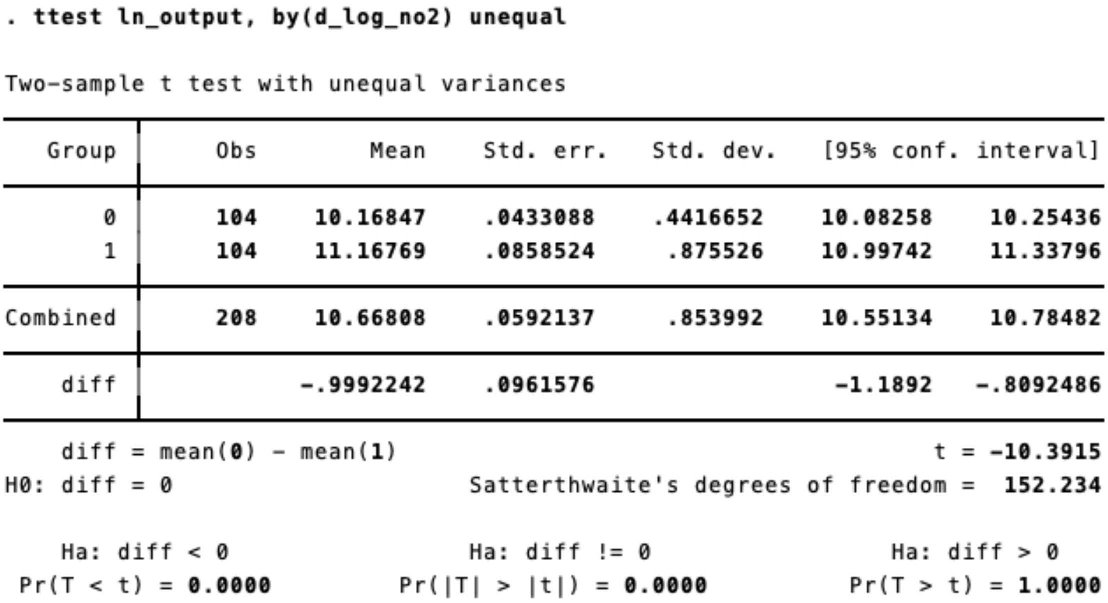
```
. hausman fe re

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |      fe           re          Difference       Std. err.
-------------+----------------------------------------------------------------
       ln_no2 |   .1130411     .0839781        .0290631        .0136189
         size |   .8771766     .9386239       -.0614474        .0357468
 current_ra~o |   .0598546     .0499136        .0099409        .0101517
     leverage |  -.0337334    -.0635424         .029809        .0365678
------------------------------------------------------------------------------
                          b = Consistent under H0 and Ha; obtained from xtreg.
           B = Inconsistent under Ha, efficient under H0; obtained from xtreg.

Test of H0: Difference in coefficients not systematic

    chi2(4) = (b-B)'[(V_b-V_B)^(-1)](b-B)
            =  11.02
Prob > chi2 = 0.0264
```

Appendix 4. Univariate T-test of Output by NO2 concentration



```
. ttest ln_output, by(d_log_no2) unequal

Two-sample t test with unequal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. err.   Std. dev.   [95% conf. interval]
---------+--------------------------------------------------------------------
       0 |     104    10.16847    .0433088    .4416652    10.08258    10.25436
       1 |     104    11.16769    .0858524     .875526    10.99742    11.33796
---------+--------------------------------------------------------------------
Combined |     208    10.66808    .0592137     .853992    10.55134    10.78482
---------+--------------------------------------------------------------------
    diff |           -.9992242    .0961576                -1.1892   -.8092486
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t = -10.3915
H0: diff = 0                     Satterthwaite's degrees of freedom =  152.234

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```