Coupling an Econometric Mixed Linear Model with an Autoregressive Moving Average Using Regional Variables Applied to Water Quality of the Ohio Basin

Table of Contents

Annotations

TRI	Toxic Release Inventory
Toxic $Release_{i,t+1}$	Natural Logarithm of one plus the total toxic chemical measure in thousands of pounds by all plans where i and t indexes firms and years respectively.
α_0	Intercept.
BC	Brand capital.
X_{it}	Control variables.
α_{year}	Fixed effects by year.
α_{ind}	Fixed effects by industry.
ε_{it}	Error term.
ESG	Environmental, social, and governance.
GDP	Gross Domestic Product.
RGDP	Real Gross Domestic Product.
Yit	Carbon Dioxide Emissions by region and over time.
δyi;t−1	Lagged value of Carbon Dioxide Emissions.

Zitβ	Vector of exogenous variables.
ηi	Fixed effects by time
U.S.	United Stated of America
R&D	Research and Development
ZIP	Zone Improvement Plan
EPA	Environmental Protection Agency
NAICS	North American Industry Classification System
$\ln TRI_{it}$	Natural logarithm of Toxic Release Inventory per county I and year y
δ	Regions
$\ln TRI_{i,t-1}$	Lagged value of the logarithm of Toxic Release Inventory per county and year
$lnCuanti_{it}$	Logarithm of EPA facilities
BEA	Bureau of Economic Analysis
FED	Federal Reserve System
IPI	Industrial Production Index
CC	Contaminants Concentration
FD	Flow discharge
HUCS4	Hydrologic Unit Code 4
USGS	United States Geological Survey

1. Theoretical Framework

We can refer to a study conducted by the University of Texas Rio Grande Valley, titled: "Does Brand Capital Influence Corporate Environmental Policy? Evidence from Toxic Release Inventory Data." This paper analyzes the TRI database, focusing on the pollutants released by various firms, and investigates whether emission levels are influenced by the company's brand capital. The study concludes that companies with high brand value tend to reduce their toxic emissions as part of their corporate environmental policy. "Brand capital" is defined as the customer's perception of the brand, encompassing subjective dimensions such as loyalty, trust, leadership, and overall brand reputation. In other words, companies that aim to project leadership or trustworthiness are more likely to adopt environmental practices that help to preserve their public image.

This study employs various econometric models, each designed to serve a specific purpose. The primary model is used to address the main research question, which is whether brand value is related to toxic releases. To do so, the authors propose the following equation:

$$\text{Toxic}\, Release_{i,t+1} = \alpha_0 + \alpha_1 BC + \alpha_K \; X_{it} + \; \alpha_{year} + \; \alpha_{ind} + \; \varepsilon_{it}$$

Similar to our model, this research uses the natural logarithm of toxic emissions as the dependent variable. It also includes an economic variable, referred to as BC (Brand Capital), which captures the brand value of the firm. Additional control variables are included to address potential endogeneity, specifically to account for the correlation between the independent variables and the error term. The model also features fixed effects for year and individual, a strategy we mirror using dummy variables to control for similar issues. Likewise, it includes a lagged dependent variable to help correct for autocorrelation in the residuals.

The model is inspired by the ESG literature, particularly by studies such as Kim et al. (2019) and Xu & Kim (2022), titled: "Financial Constraints and Corporate Environmental Policies", a highly cited work published in 2022. The study provides evidence that financial constraints can lead firms to increase their toxic emissions, as they often face a trade-off between the costs of reducing emissions and the potential legal liabilities associated with non-compliance.

In the previously mentioned studies, the analysis focuses on firms over time, examining their toxic emissions in relation to an economic variable. In our case, we follow a similar approach, but with a key difference: instead of analyzing firms, we focus on counties over time, linking them to an appropriate economic variable. While corporate studies typically use accounting or financial indicators, in our territorial approach, a comparable and relevant variable is the GDP at the county level. This variable allows us to consistently capture local economic activity across time.

In the study by Du, L., Wei, C., & Cai, S. (2012), titled "Economic Development and Carbon Dioxide Emissions in China: Provincial Panel Data Analysis", the authors examine the relationship between economic growth and CO_2 emissions at the provincial level. They use a panel data econometric model to evaluate how pollution levels change in response to economic development over time across Chinese provinces. The following equation represents the model they propose:

$$Yit = \delta yi;t-1 + Zit\beta + \eta i + \varepsilon it$$

In the proposed model, the dependent variable is carbon dioxide emissions by region (i) and over time (t), which is an environmental variable similar to the one used in our study. This variable is partly explained by its lagged value, allowing the model to capture persistence and dynamic effects in emission behavior. In addition, the model includes a vector of exogenous variables represented by Z_{it}, which accounts for factors such as GDP per capita, industrial composition, urbanization level, energy consumption structure, trade openness, and technological progress, among others. It is worth noting that all variables are expressed in natural logarithms, which enables the interpretation of coefficients as elasticities and simplifies comparison across variables with different measurement scales.

In our case, the goal is not to represent economic development in a broad sense, which encompasses aspects such as quality of life, equity, or social well-being, but rather to focus solely on capturing the level of economic production.

Another example that follows a similar research direction is the study by Wan-Jiun Paul Chiou, titled "Exploring the Impacts of Economic Policies, Policy Uncertainty, and Politics on Carbon Emissions." This study employs an econometric model similar to the one used in the Chinese case but applied to the United States. The dependent variable is the natural logarithm of monthly carbon dioxide emissions, while the independent variables include:

- The natural logarithm of GDP per capita,
- Changes in the average tax rate,
- The corporate interest rate,
- A dummy variable indicating whether the sitting president is a Democrat or Republican,
- In addition, a measure of the partisan composition of Congress.

This model captures both economic and political influences on emissions and provides valuable insights for studies focused on national-level dynamics.

Regarding the use of dummy variables, we identified certain similarities in production patterns across different regions, which led us to group the data by U.S. geographic regions, a structure that aligned well with the goals of our research. The use of dummy variables is a common practice in econometric modeling, as it allows researchers to account for category-specific or group-level effects. A relevant example is the study by Lianqun Sun, titled "Factors Influencing Seafood Sales in U.S. Retail Markets."

In that study, the author employs dummy variables in a way similar to ours—dividing the U.S. into geographic regions, as well as for other purposes such as identifying the originating store of each sale. This highlights the value of dummy variables in capturing unobserved heterogeneity within the model.

Another example that supports the use and effectiveness of dummy variables in econometric models is the study by Jens Horbach, titled "Determinants of Environmental Innovation: New Evidence from German Panel Data Sources." In this research, the author employs dummy variables to identify specific factors that influence environmental innovation, finding that improvements in technological capabilities, particularly through R&D investments, are crucial drivers of such innovation.

The use of dummy variables allows the model to control for structural or institutional differences across regions or sectors, highlighting their value in capturing unobserved heterogeneity within panel data settings.

In his econometric model, Jens Horbach use several dummy variables, each serving different analytical purposes. These include indicators for:

- Whether the firm introduced environmental innovations in the past two years,
- Whether overtime work was reported,
- Whether the firm is located in Eastern or Western Germany,
- Whether it received financial assistance,
- In addition, whether it belongs to specific economic sectors, coded as one if it does and zero otherwise.

In addition, the model incorporates continuous explanatory variables such as the age of the firm and the ratio of highly qualified employees, among others. This approach demonstrates how the combination of categorical and continuous variables can enhance the model's ability to capture the complexity of environmental innovation dynamics.

2. Context: Toxic Release Inventory

The main purpose of this research is to assess water quality of the Ohio River Basin, focusing on contamination concentration during different period of the year. To achieve this, various monitoring sites throughout the basin were previously analyzed. These sites measure variables such as water flow, as well as different physical and biological characteristics.

A database called TRI was found, which tracks the amount of chemicals released into the environment by companies. The database is broken by ZIP code, county, city, state, economic sector, etc. However, the main limitation is that measurements are reported annually, which poses a challenge for analysis that require higher temporal resolution. To address this problem, 2 approaches were considered: dividing the annual emissions by 12, assuming steady production through the year, or interpolating the data using a function or, in this case, a predictive model based on monthly level factors.

As a result, an econometric model was developed, where the dependent variable to be estimated is the amount of toxic emissions to the environment, based on a set of explanatory variables that will be described later in this document. When reviewing which variables could be used, the most evident one is economic activity. One of the key variables to represent this is the RGDP, which adjusts for inflation and reflects the true performance of the economy. This variable, therefore, captures the real fluctuations in economic activity and offers a more accurate representation of the actual production dynamics in the United States.

The variable CUANTI was also selected, representing the number of establishments (facilities) or firms registered with the EPA that hold permits to emit pollutants. It was verified that there is no multicollinearity with the RGDP variable. While it might seem intuitive that more firms would correlate with higher production, and thus with a higher GDP, but CUANTI captures a different aspect of industrial dynamics.

CUANTI helps illustrate structural changes in the market. Over time, markets often trend toward consolidation, reducing the number of active firms and evolving into oligopolies or even monopolies. However, this shift does not necessarily lead to a decrease in GDP, as the remaining firms tend to become more efficient and take over the market share left by others. Therefore, CUANTI can be seen as

a complementary indicator that reflects industry dynamics and structure, beyond the total volume of production.

Finally, a lagged version of the dependent variable was included in the model. Initially, this was done to correct for autocorrelation detected in the model. However, its inclusion also has strong theoretical justification. Companies often consider the amount of emissions recorded in previous periods to plan for current or future actions. This plan allows them to make informed decisions regarding their environmental policies, aiming to stay within the limits set by both their internal budgets and EPA regulations. In other words, emissions in the current period are typically correlated with those from the previous period, which supports the use of a lagged dependent variable in the model.

Before presenting the econometric model and its theoretical justification, it is important to clarify the direction of the analysis. Initially, the goal was to establish a direct connection between emissions reported in the TRI database and economic activity measured through the RGDP. However, conducting such a broad analysis at the national level proved impractical, especially considering that the final application was intended for a specific point within the U.S. territory. Therefore, the focus shifted to the smallest geographic unit with reliable economic data: counties. Each year, the federal government provides county-level Real GDP data, further broken down by economic sectors, ranging from agriculture to manufacturing, arts, and education.

Each sector is identified by a NAICS code, which serves to differentiate various economic activities. Similarly, the TRI data is also classified by economic sector, making it possible to align both datasets. This alignment enabled the integration of emissions data with sector-specific economic output, allowing us to focus on the sectors most relevant to the Ohio River Basin—specifically, mining sector and the nondurable goods manufacturing sector. These sectors are particularly significant in the region due to their substantial contribution to the total TRI emissions representing over an 80% of the total.

However, some clarifications are necessary as we delve into the model specification. All variables will be transformed using the natural logarithm. This approach serves two main purposes: first, to reduce the influence of outliers, and second, to harmonize the units of measurement, which originally differ. For example, TRI is reported in pounds or tons, while RGDP is expressed in thousands of dollars. Applying the natural logarithm allows us to work with these variables on a consistent relative scale and simplifies coefficient interpretation. It's also important to note that in log-log models—where both the dependent and independent variables are expressed in logarithmic form—the estimated coefficients represent elasticities. In other words, they indicate the percentage (%) change in the dependent variable resulting from a one-percent (%) change in the corresponding independent variable.

Dummy variables, or indicator variables taking values of 0 or 1 depending on the county or region, will also be included in the model. The purpose of these variables is to account for structural or contextual differences that are not explicitly captured by the main explanatory variables, which affect emission levels. These dummies help to control for unobserved factors such as differences in environmental policies at the local or regional level, as well as variations in production methods across different areas of the country. Incorporating these effects improves the accuracy of the model and allows for a more realistic representation of the diverse conditions present across the United States.

Finally, to summarize: two economic sectors were selected for this analysis. The first is Sector 21, which includes all types of mining activities (such as oil, gas, coal, etc.). The second sector is nondurable goods

manufacturing, which, according to the NAICS classification, includes codes 311–316 and 322–326. Given this sectoral segmentation, two separate econometric models will be developed—each estimating toxic emissions for its respective sector. The results of these models will then be summed to obtain the total estimated emissions for the combined sectors.

3. Model specifications

Let's check the econometric model for Sector 21:

$$\ln TRI_{it} = \delta_2 + \delta_3 + \delta_4 + \delta_5 + \delta_6 + \delta_7 + \ln TRI_{i,t-1} + \ln RGDP_{it} + \ln Cuanti_{it} + \varepsilon_{it}$$

As can be seen, the dependent variable is the natural logarithm of toxic emissions. The independent variables include the natural logarithm of Real Gross Domestic Product. Additionally, several control variables are included, such as the number of establishments registered with the EPA, the lagged variable of the logarithm of emissions, and finally, the dummy variables that serve as intercepts for groups of counties geographically clustered into regions.

Upon analyzing the data, significant similarities were noted between certain counties. As a result, the decision was made to geographically divide them as follows:

- δ_2= The Mid-Atlantic (It is the purple-colored area on the map, covering states such as New York, New Jersey, Pennsylvania, etc.)
- δ_3= The Southeast (It is the blue-colored area on the map, covering states such as West Virginia, Virginia, Tennessee, Florida, Arkansas, Alabama, etc.).
- δ_4= The Midwest (It is the gray-colored area on the map, covering states such as Kansas, Iowa, Illinois, Ohio, Michigan, Minnesota, etc.).
- δ_5= The Rocky Mountains (It is the red-colored area on the map, covering states such as Utah, Nevada, Idaho, Wyoming, and Montana).
- δ_6= The Southwest (It is the yellow-colored area on the map, covering states such as Texas, Oklahoma, New Mexico, and Arizona).
- δ_7= The Pacific Coast (It is the green-colored area on the map, covering states such as California, Alaska, Hawaii, Oregon, and Washington).

Figure 1 America Regions. Source: USA Welcome

By dividing the data into geographical zones, the model was run, yielding highly satisfactory results. The most important values, such as the lagged variable and Real Gross Domestic Product, showed strong statistical significance. Additionally, the control variables, such as the dummy variables and the number of establishments, were also statistically significant. More importantly, given our main goal of forecasting values, a high R^2 value was obtained, indicating that the model effectively explains the variability in toxic emissions, closely matching the actual value of TRI, or the predicted Y variable. This relationship is clearly shown in the next section.

$$\ln TRI_{it} = \delta_i + \ln TRI_{i,t-1} + \ln RGDP_{it} + \ln Cuanti_{it} + \varepsilon_{it}$$

This econometric model will be used to calculate the value of toxic emissions for the entire nondurable goods manufacturing sector. In the equation, we can observe that the natural logarithm of toxic emissions is our dependent variable. On the other side of the equation, we find the lagged variable and the natural logarithm of Real Gross Domestic Product as our main variables. Additionally, we include control variables, such as the natural logarithm of the number of establishments registered with the EPA, and, finally, a dummy variable for each county. Since it was difficult to identify common and statistically significant patterns between regions, we decided to treat each region separately in order to obtain a more accurate and realistic result (The results can be seen in the next section).

4.Predictions

One of our main goals is to generate predictions with the econometric model that closely reflect reality, so that eventually we can confidently forecast the TRI from companies on a monthly basis. As mentioned in the previous section, we now present two graphs comparing the actual data with the model's predictions during training. First we have the graph of the 21 industry:

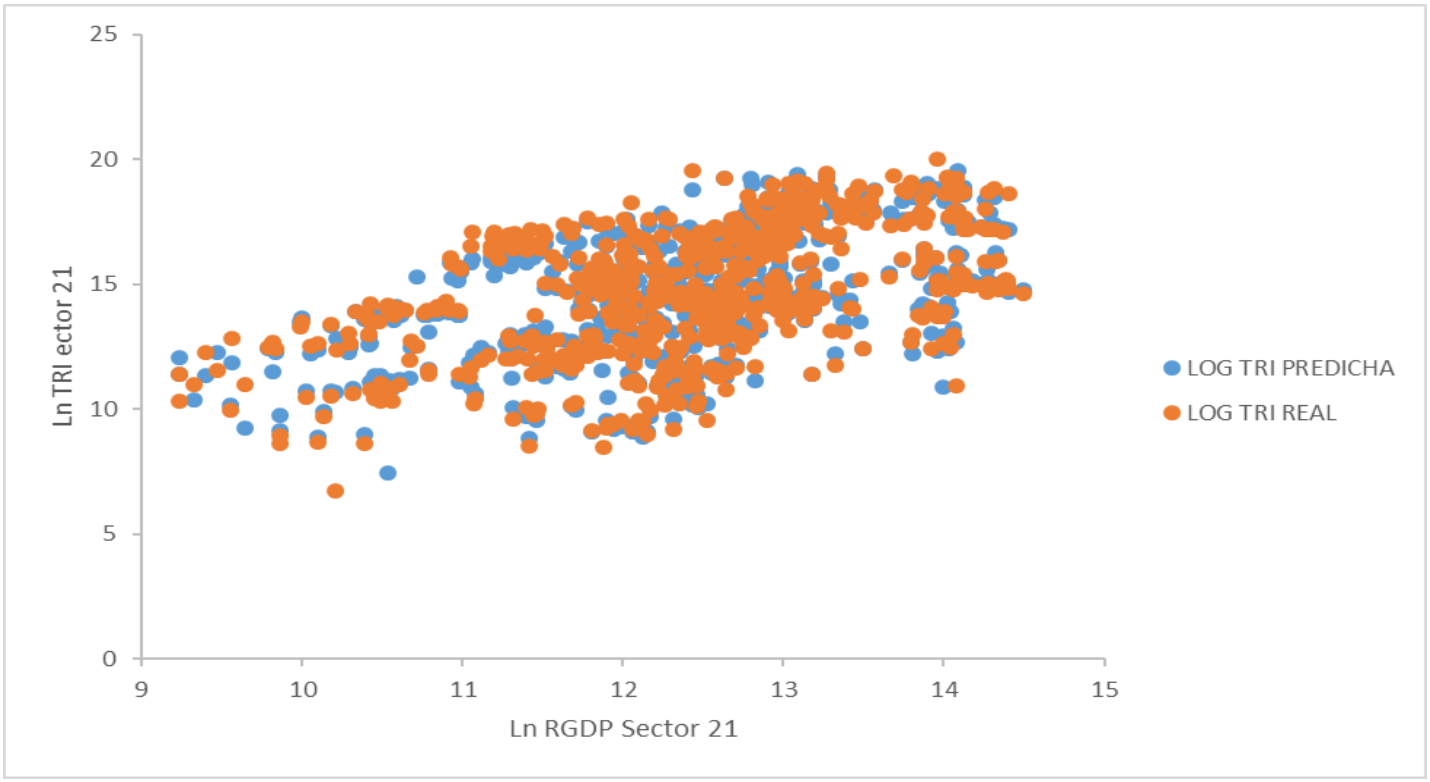

Figure 2 Natural Logarithm (ln) relationship between Real Gross Domestic Product (RGDP) and Toxic Release Inventory (TRI) in Mining sector. Source: U.S. EPA Toxics Release Inventory (2023) and Bureau of Economic Release GDP data

Then we have the graph for the Non-Durable Manufacturing:

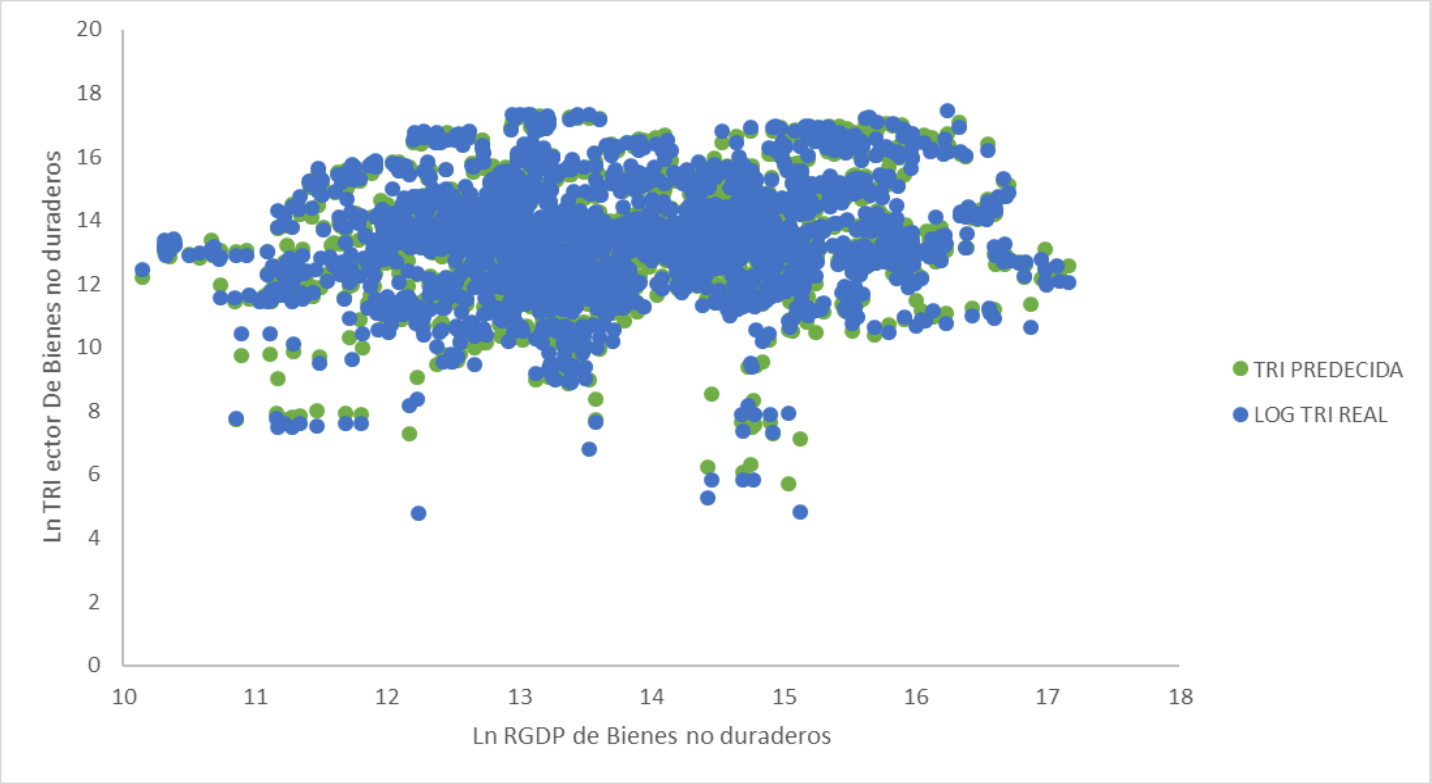

Figure 3 Natural Logarithm (ln) relationship between Real Gross Domestic Product (RGDP) and Toxic Release Inventory (TRI) in Non-Durable Manufacturing sector. Source: U.S. EPA Toxics Release Inventory (2023) and Bureau of Economic Release GDP data

5. Proportional Interpolation

To obtain monthly estimates of TRI, we relied on an economic variable with monthly frequency: RGDP. However, the BEA only publishes RGDP data annually at the county and industry level across the United States. To convert this annual data into a monthly format, we used the Industrial Production Index, which is published monthly by the FED. This index serves as a proxy for short-term economic activity and enables a more accurate interpolation of RGDP on a monthly basis.

The IPI is broken down by major industrial sectors, including the two key sectors considered in our study: mining and nondurable goods manufacturing. Using this index, we applied a proportional linear interpolation method to estimate monthly values of Real Gross Domestic Product for each industry. The interpolation was based on the monthly variations in the corresponding IPI, allowing us to construct a monthly RGDP time series that reflects the actual production dynamics of each sector.

6. Maps

Our main focus is the Ohio River Basin, where our objective is to analyze the evolution of pollutant concentration levels in relation to water, using this relationship as our key variable to determine whether water quality has improved or deteriorated. In the presentation of our project, we will display a total of five maps. The code includes the creation process for three of them, while the remaining maps were produced using specialized mapping software, due to the complexity of data extraction and processing. It is important to note that we also created these maps, even though a different platform was used to facilitate their generation.

As with any study that aims to analyze a specific region, it is essential to show its location to help the reader better, understand the context. For this reason, the first map we present will show the location of the Ohio River Basin within the North American territory. The map will highlight the area it covers and the U.S. states that fall within the basin. This also will serve as a starting point to introduce the significance of the region in the context of our study.

Figure 4 Ohio Basin Location. Source: Author's analysis using Python

The second map, first we need to understand that with the monthly economic variable and, subsequently, our main variable, the number of pollutants released into the environment, we can proceed to the next step of our research: analyzing the pollution concentration. To do this, we need to obtain monthly water flow data for the rivers previously selected, covering the majority of the Ohio River Basin territory. It is important to note that the drainage area delineation from the measuring point reported by the United States Geological Survey (USGS) was taken into account when collecting this data

$$CC = TRI/FD$$

In our formula, CA represents the contaminant concentration, while TRI corresponds to the level of pollutants previously calculated on a monthly basis. FD, on the other hand, refers to the water flow from the rivers, with data provided every 15 minutes by the United States Geological Survey (USGS). To convert this data into a monthly scale, we sum the flow values for each river throughout the month, obtaining the FD values for each river selected.

However, when summing the water flows from the rivers, and taking into account the delimitation of their respective watersheds, we observe that they follow a structure that allows the Ohio River Basin to be divided into smaller areas. These areas correspond to the hydrologic sub regions, known as HUCS4 , which is a standard classification used to divide watersheds.

The hydrologic sub regions will help us to analyze in more detail whether improvements of contaminants concentration have occurred in the selected years (2018-2023). Additionally, we have segmented each year into 4 quarters, providing a more precise view on a quarterly basis. Thus, the selected years (2018-2023) were divided into two subsets: 2018-2020 and 2021-2023, allowing us to compare trends and variations over this period.

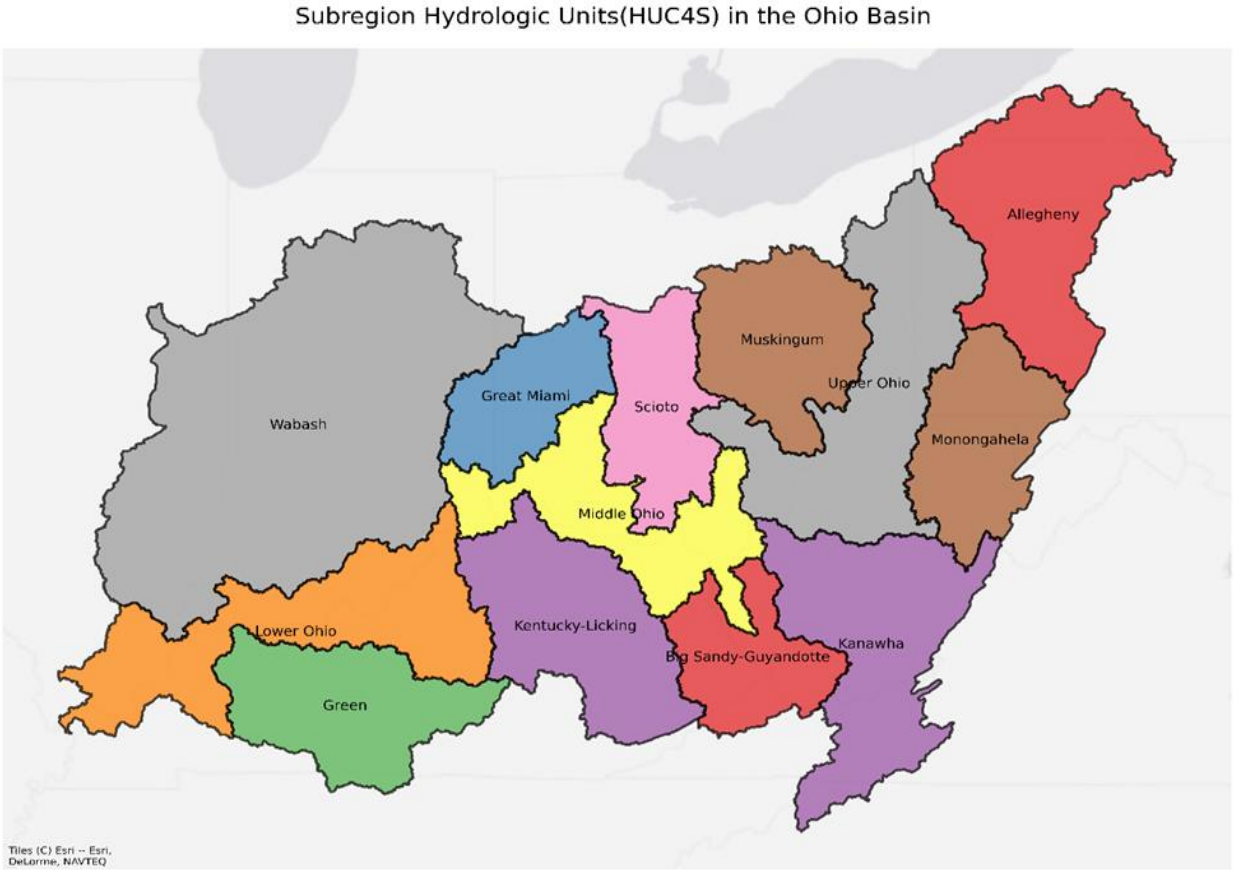

Figure 5 Hydrological Subregions on the Ohio Basin. Source: Author's analysis based on USGS dataset

The third map, explained in the code and serving as one of the key visualizations of our analysis, shows the growth rate in pollutant concentration across the previously analyzed hydrologic sub regions. This is

examined across the four quarters of the year, comparing water flow and pollutant levels between the periods 2018–2020 and 2021–2023. This approach allows us to identify temporal and spatial patterns in the evolution of water quality.

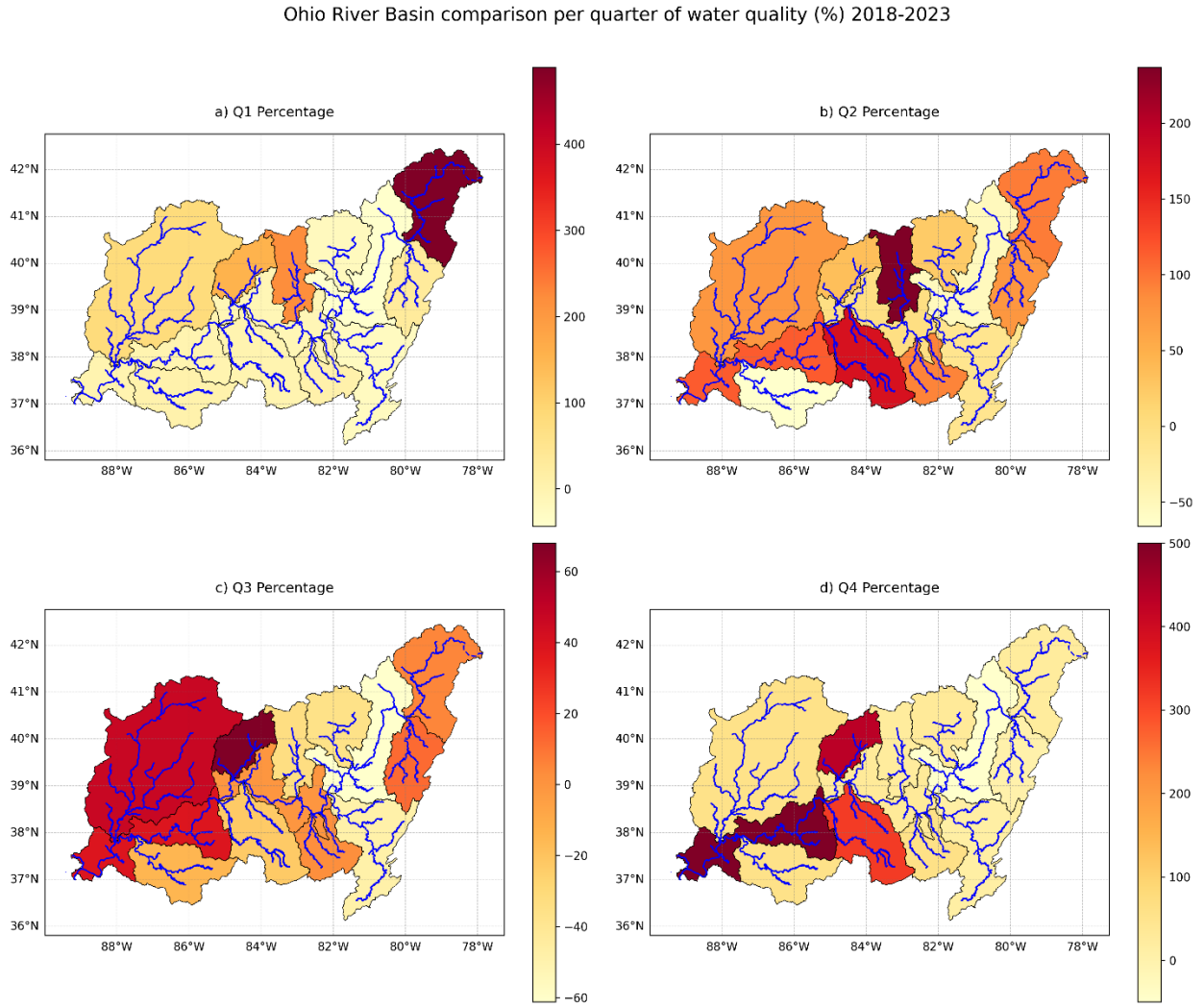

Figure 6 Comparative Map Analysis by Quarter (2018–2023). a) First Quarter (January–March); b) Second Quarter (April–June); c) Third Quarter (July–September); d) Fourth Quarter (October–December). Source: Author's work.

In the map, we can observe that during a) and b) quarters of the year, most hydrologic sub regions appear in light tones, indicating little or no change in water quality over the years analyzed. In contrast, during the b) and c), several regions show a decline in water quality. Interestingly, at least two sub regions display a noticeable improvement. It is important to note that a negative growth rate means the previous period (2018–2020) had higher pollution levels than the current one (2021–2023), indicating an improvement in those areas.

However, we are aware that there are certain gaps in our analysis. To address this, we created two additional maps outside of Python, using QGIS, a software specialized in geospatial mapping. The first of these supplementary maps displays the locations of facilities related to the non-durable manufacturing

industries within the Ohio River Basin. Additionally, it highlights areas where gas or oil wells and coalmines are located, facilities associated with Industry 21, which is one of the main sectors examined in our study.

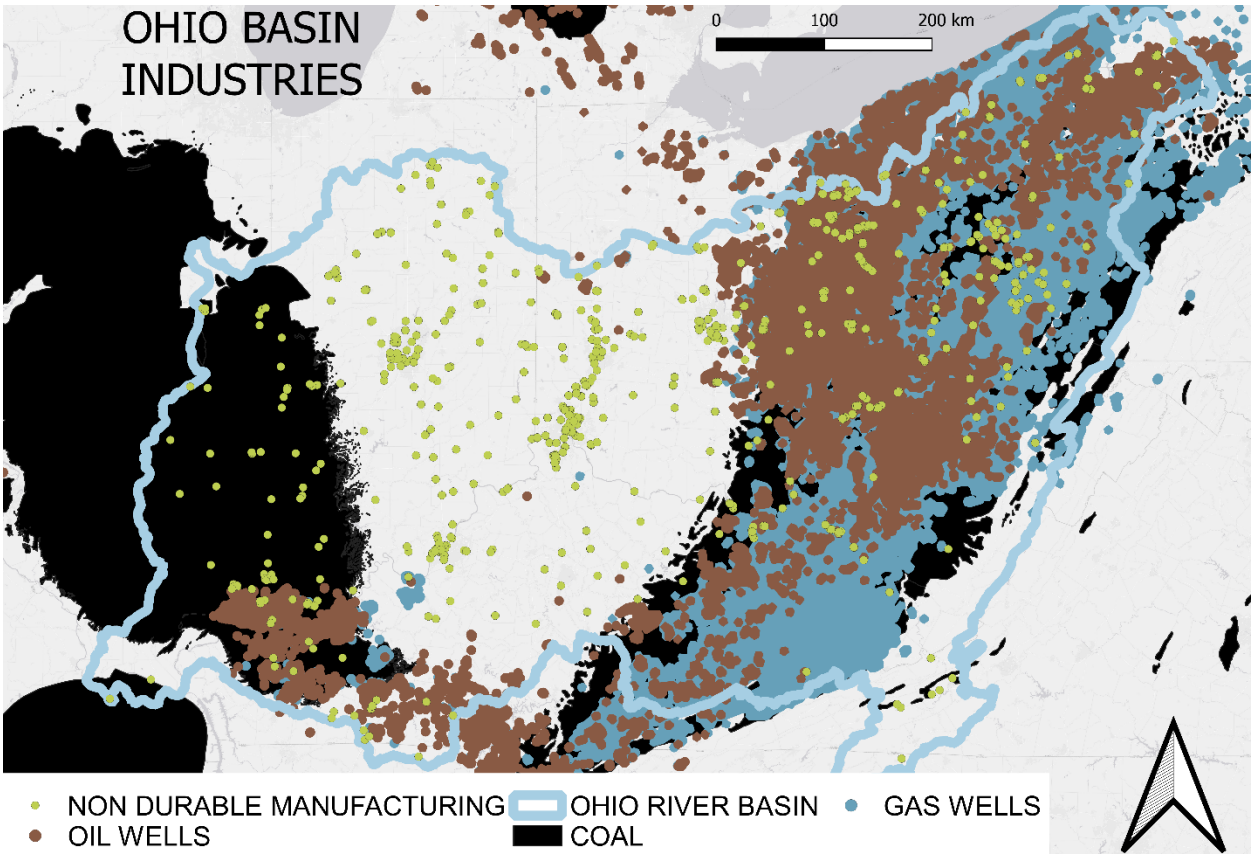

Figure 7 Oil, Carbon and Gas wells in the Ohio Basin. Source: Author's analysis based on U.S Energy Information Administration dataset

Finally, our last map presents the average Toxic Release Inventory (TRI) for the years 2018 to 2023 in the Ohio River Basin, broken down by county. As the reader may recall, our main database, the one used to train the model, is also organized at the county level.

This map helps us identify which counties in the basin, on average, are releasing more pollutants. Additionally, it provides context to better understand the differences observed in the comparative map of each hydrological sub region.

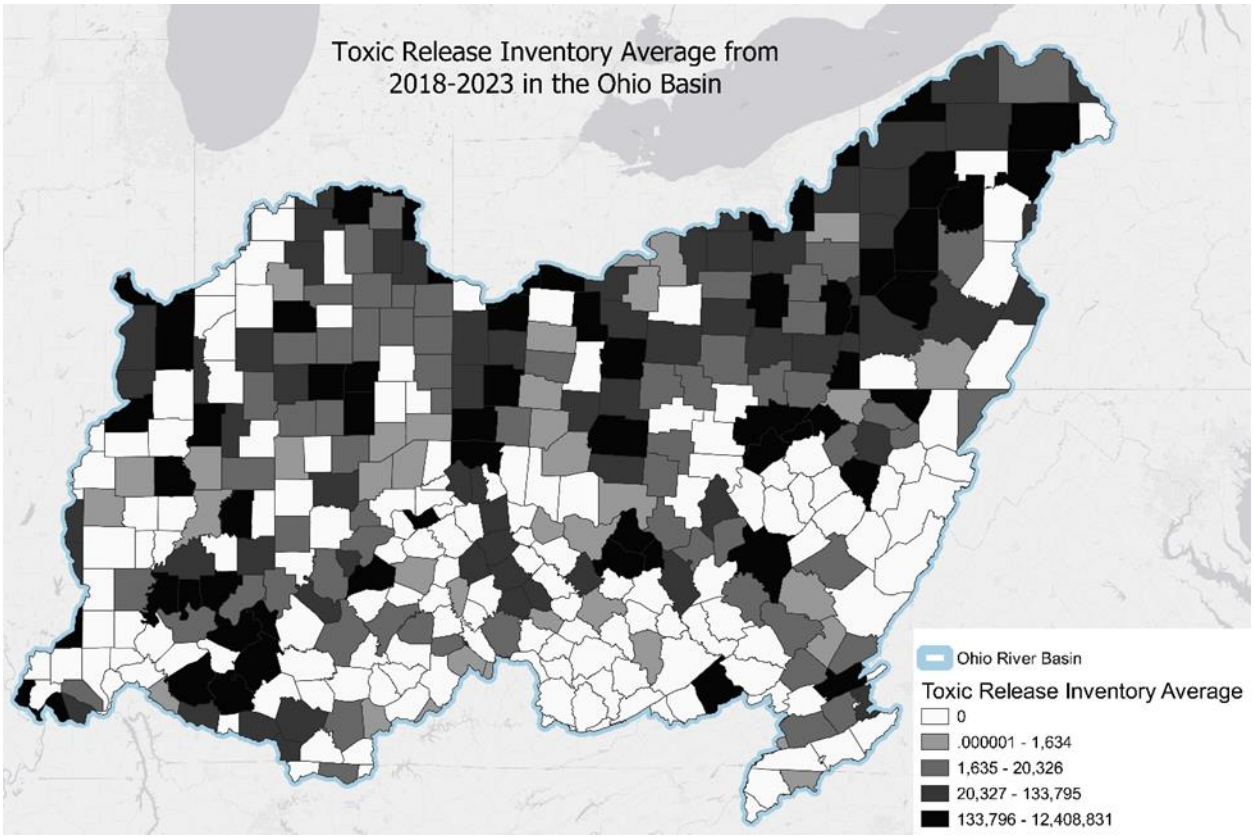

Figure 8 Toxic Release Inventory average from 2018 to 2023 in the Ohio Basin. Source: U.S. EPA Toxics Release Inventory (2023)

7. References

Sun, L., Kumar, G., & Engle, C. (2025). Factors influencing seafood sales in US retail markets. *Journal of the World Aquaculture Society, 56*(1), e70000. https://doi.org/10.1111/jwas.70000

Impact of urbanization on energy intensity in SAARC countries. (s. f.). South Eastern University of Sri Lanka. https://ir.lib.seu.ac.lk/bitstream/123456789/7262/1/Impact%20of%20urbanization%20on%20energy%20intensity%20%281%29.pdf

Chiou, W.-P. P., Fu, S.-H., Lin, J.-B., & Tsai, W. (2025). Exploring the impacts of economic policies, policy uncertainty, and politics on carbon emissions. *Environmental and Resource Economics, 88*(4), 895–919. https://doi.org/10.1007/s10640-025-00954-6

Haidar, Md Ismail and Kroll, Mark and Nguyen, Nam, Does Brand Capital Influence Corporate Environmental Policy? Evidence from Toxic Release Inventory Data. Available at SSRN: https://ssrn.com/abstract=4921687

Du, L., Wei, C., & Cai, S. (2012). Economic development and carbon dioxide emissions in China: Provincial panel data analysis. China Economic Review, 23(2), 371–384. doi:10.1016/j.chieco.2012.02

van den Broek, H., & van der Waa, J. (2022). Intelligent operator support concepts for shore control centres. Journal of Physics: Conference Series, 2311, 012032. https://doi.org/10.1088/1742-6596/2311/1/012032

Horbach, J. (2008). Determinants of environmental innovation—New evidence from German panel data sources. Research Policy, 37(1), 163–173. doi:10.1016/j.respol.2007.08

Garavaglia, S., & Sharma, A. (2021). *A smart guide to dummy variables: Four applications and a macro* [PDF]. UCLA Statistical Consulting Group. https://stats.oarc.ucla.edu/wp-content/uploads/2016/02/p046.pdf

Xu, Q., & Kim, T. (2022). Financial constraints and corporate environmental policies. Review of Financial Studies, 35(2), 576–635. https://doi.org/10.1093/rfs/hhab056

U.S. Environmental Protection Agency. (2023). *Toxics Release Inventory (TRI) data* [Data set]. https://www.epa.gov/toxics-release-inventory-tri-program/tri-data-and-tools

U.S. Bureau of Economic Analysis. (2023). *Real GDP by county* [Data set]. https://www.bea.gov/data/gdp/gdp-county-metro-and-other-areas

Board of Governors of the Federal Reserve System. (2023). *Industrial Production and Capacity Utilization (G.17)* [Data set]. https://www.federalreserve.gov/releases/g17

U.S. Geological Survey. (2024). [Data set]. USGS Water Data. https://waterdata.usgs.gov/nwis

U.S. Energy Information Administration. (2023). *Gas Wells* [Data set]. https://www.eia.gov/maps/maps.php

U.S. Energy Information Administration. (2023). *Oil Wells* [Data set]. https://www.eia.gov/maps/maps.php

U.S. Energy Information Administration. (2023). *Carbon Wells* [Data set]. https://www.eia.gov/maps/maps.php

USAWelcome.net. (s.f.). *Mapa de las regiones de Estados Unidos* [Foto]. https://www.usawelcome.net/es/explora/bueno-saber-/informacion-general/the-regions-of-the-united-states.htm