What is the effect of Artificial Intelligence on the environment?

Master 1 Économétrie et Statistique

NAJI Anas

Toulouse School of Economics

Thesis advisor : *MISSIRIAN Anouch*

Writing period :

April - June 2025

Abstract

The aim of this project is to better understand the impact of artificial intelligence (AI) and, more precisely, large language models (LLMs) on the environment. An understanding of this impact led us to critically examine the limited responsibility assumed by companies in this domain. To address this issue, a new method was developed to estimate the environmental costs of AI more accurately.

Our findings revealed that existing indicators, such as Scopes while commonly used to assess firms' environmental footprints, suffer from a lack of transparency and completeness.

We then shifted our focus to the physical process behind AI and calculate the cater and electricity consumption per request.

Finally we focused on the impact of AI on the environment on a structural plan, by studying the localization of datacenter around the world. We identified patterns suggesting that companies strategically place their data centers near big cities, close to natural water sources, and in regions with low soil moisture. Then we attempted to evaluate the direct impact of data centers on local environments and urban ecosystems, but the insufficient data and the presence of confounding factors limited the scope of our conclusions.

Contents

I Introduction 7

II Scopes and the Example of CapGemini 9

1 An Existing Measure: Scope 9

1.1 What is a Scope? . . . 9

1.1.1 Scope 1 . . . 9

1.1.2 Scope 2 - Market / Location Based . . . 9

1.1.3 Scope 3 . . . 10

1.2 Example . . . 10

2 What About the Scopes? 10

3 Study of the firm 16

3.1 What About the Firm? . . . 16

3.1.1 Goal of the firm . . . 16

3.1.2 Evolution of the firm . . . 16

3.1.3 Greenhouse Gases Management Policy . . . 16

4 Forecasting 17

5 Scopes-Unreliable indicators? 23

5.1 Information Disclosure . . . 23

5.2 Lack of transparency . . . 24

III Understanding of an AI 25

6 How does an AI work 25

7 Calculus : Electricity Consumption 26

7.1 GPU & VRAM . . . 26
7.2 CPU . . . 26
7.3 Ram . . . 27
7.4 SSD . . . 27

8 Computation : Electricity consumption **29**
8.1 Step 1 . . . 29
8.2 Step 2 . . . 29
8.3 Step 3 . . . 29
8.4 Step 4 . . . 30
8.5 Step 5 . . . 31
8.6 Total . . . 31

9 Computation : Water consumption **32**

10 Calculus : Water cooling **32**

IV Data Centers **34**

11 Motivation **34**

12 Localization **34**
12.1 Mapping . . . 34
12.2 Variables . . . 36
12.2.1 Temperature . . . 36
12.2.2 Soil Moisture . . . 37
12.2.3 Earthquake . . . 37
12.2.4 Drought factor . . . 37
12.2.5 Proximity to a big city . . . 38
12.2.6 Proximity to oceans/seas/rivers/lakes . . . 38
12.3 Regression . . . 39

12.3.1 GLM 39
12.3.2 Logit 42

13 The issue of case studies **45**

V Conclusion **46**

VI Annexes **48**

List of Figures

1 Scope 1 over time in tonnes of CO_2. 11
2 Scope 2 LB vs MB over time in tonnes of CO_2. 12
3 Scope 3 over time in tonnes of CO_2. 13
4 Total Scopes over time in tonnes of CO_2. 14
5 Forecast of Scope 1 over time in tonnes of CO_2. 18
6 Forecast of Scope 2 LB over time in tonnes of CO_2. 19
7 Forecast of Scope 2 MB over time in tonnes of CO_2. 20
8 Forecast of Scope 3 over time in tonnes of CO_2. 21
9 Forecast of total Scopes over time in tonnes of CO_2. 22
10 Overview of the functioning of an AI. 25
11 Hex-bin map of data centers 35
12 Top 3 countries 36
13 Model Distance 39
14 Model proximity 40
15 Coefficient regression 41
16 Lasso plot 42
17 Optimal choice 43
18 Logit coefficients 44
19 Scope 1 over time in tonnes of CO_2. 48
20 Scope 2 LB vs MB over time in tonnes of CO_2. 49
21 Scope 3 over time in tonnes of CO_2. 50
22 Total Scopes over time in tonnes of CO_2. 51
23 Forecast of Scope 1 over time in tonnes of CO_2. 52
24 Forecast of Scope 2 LB over time in tonnes of CO_2. 53
25 Forecast of Scope 2 MB over time in tonnes of CO_2. 54
26 Forecast of Scope 3 over time in tonnes of CO_2. 55
27 Forecast of total Scopes over time in tonnes of CO_2. 56

28 Overview of the functioning of an AI. 57
29 Hex-bin map of data centers 58
30 Top 3 countries 58
31 Model Distance 59
32 Model proximity 59
33 Coefficient regression 60
34 Lasso plot 61
35 Optimal choice 61
36 Logit coefficients 62

Part I

Introduction

René Descartes (1956-1650), a French philosopher and mathematician, approached the idea of artificial intelligence in an interesting way. In *Le discours de la méthode* [Descartes, 1637] Descartes was already discussing the idea of the automates, stating that in addition to the fact that the human body may be compared to a complex machine, some simple human behaviors could be mechanically replicated. However he drew a line between humans and automates explaining that there is a clear distinction between the soul (consicent soul) and the body (mechanic and automatic) in *Les passions de l'âme* [Descartes, 1649].

Another pioneer of the idea of AI is Gottfried Wilhelm Leibniz (1646–1716) who created in 1673 the *machine à pensée*, a simple but innovative machine which was able to effectuate mathematic operations materializing ,implicitly, the idea that Descartes was right and that it is indeed possible to materialize the human thinking (calculus) into a machine.

John McCarthy, Marvin Minsky, Allen Newell or Herbert Simon were all able to develop AI thanks to people like Descartes and Leibniz. Just as in the age of the philosophers, it was questioning and critical thinking that allowed the idea of artificial intelligence to take root in people's minds and enabled scientists to develop it, we must now apply the same critical approach to its use and consider its impact on the environment.

Indeed, according to the whereabouts of our time, studying the ecological impact of new technologies is mandatory ([Zhuk, 2023]) and will surely be a key to the challenges related to climate issues.

This is in this spirit that we will try to tackle the issue and first study some already existing index like the scopes, see if its fitted in the AI field, find its shortcomings and see if, how and why firms may have a shady communication regarding those indexes. Then we will try to measure ourselves the impact of AI on the environment regarding LLM's. Finally we will study the impact of AI on it's environment on a global and structural way with the study of datacenter.

Part II

Scopes and the Example of CapGemini

1 An Existing Measure: Scope

1.1 What is a Scope?

Scopes were first introduced in 2001 by the Greenhouse Gas Protocol to represent and categorize greenhouse gas emissions. This concept is divided into three categories : Scope 1 , Scope 2 and Scope 3. Let's detail a bit what each scope embodies.

1.1.1 Scope 1

Scope 1 refers to all direct greenhouse gas (GHG) emissions resulting from sources that are owned or controlled by the company. This includes emissions from industrial facilities, warehouses, offices, and company-owned vehicle fleets. These emissions arise directly from the company's operational activities.

1.1.2 Scope 2 - Market / Location Based

Scope 2 encompasses indirect GHG emissions associated with the purchase and use of electricity, heat, steam, or cooling. Although these emissions occur outside the organization's direct operational boundaries, they are a direct consequence of its energy use.

The market based (MB) scope 2 reflects emissions based on the specific electricity contracts and sources chosen by the company.

The location based (LB) scope 2 calculates emissions using the average energy mix of the local grid where the electricity is consumed.

1.1.3 Scope 3

Scope 3 includes all other indirect emissions that occur in the company's value chain, both upstream and downstream. These can result from a wide range of activities, such as:

- emissions from suppliers and subcontractors
- business travel
- product used by customers
- end-of-life treatment of sold goods

1.2 Example

To illustrate the differences let's take the example of a truck owned and used by a firm A and manufactured by the firm B. Note that firm B also sells electricity.

Situation	**Scope for firm A**
Gas used by firm A to drive its own truck	Scope 1
Electricity purchased by firm A for its headquarters	Scope 2
Energy used by firm B to manufacture the truck	Scope 3

Table 1: Examples of emissions and their associated scopes for firm A

2 What About the Scopes?

We will analyze a dataset constructed from CapGemini's financial reports. It is made up of information concerning Scopes as well as financial information. Note that the Scopes are expressed in tons of CO^2 equivalent. First, let us display the value of the scopes over the year.

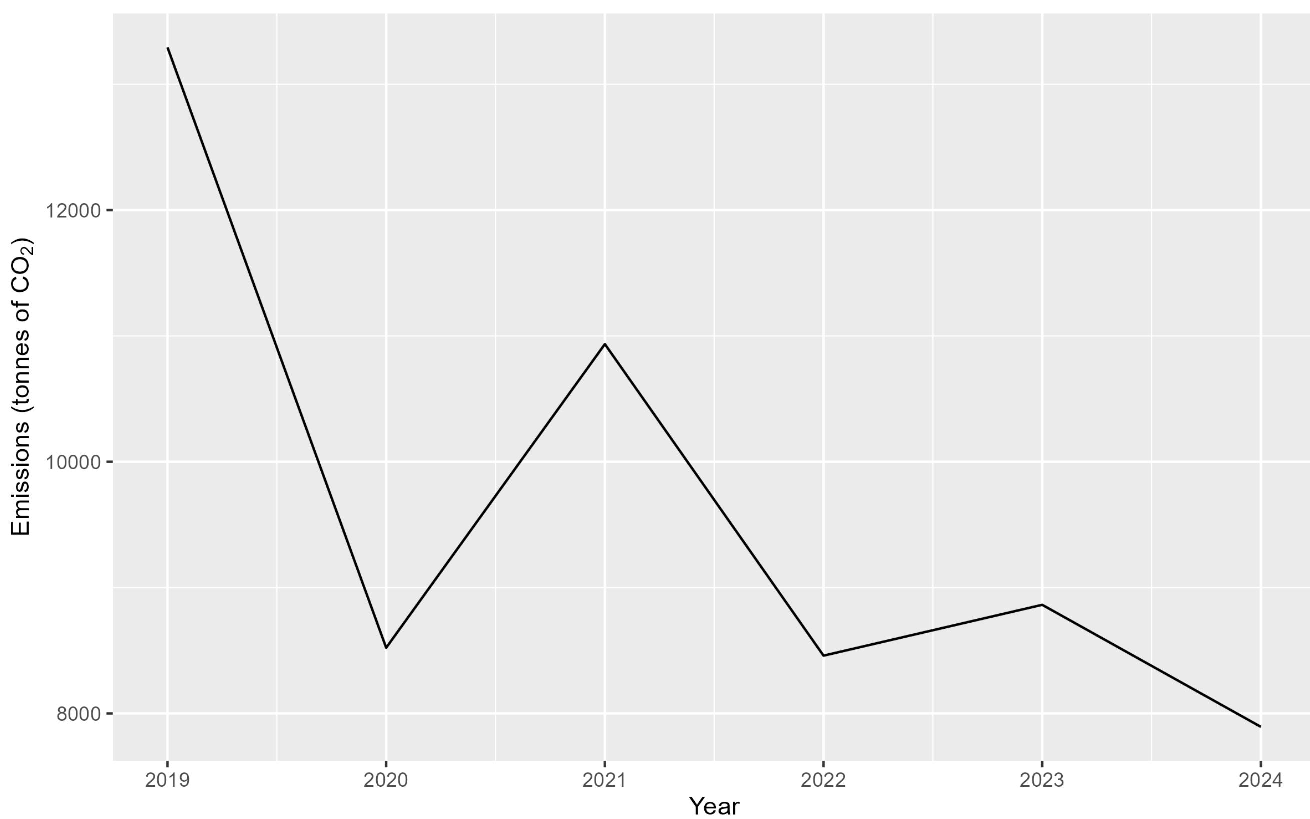

Figure 1: Scope 1 over time in tonnes of CO_2.

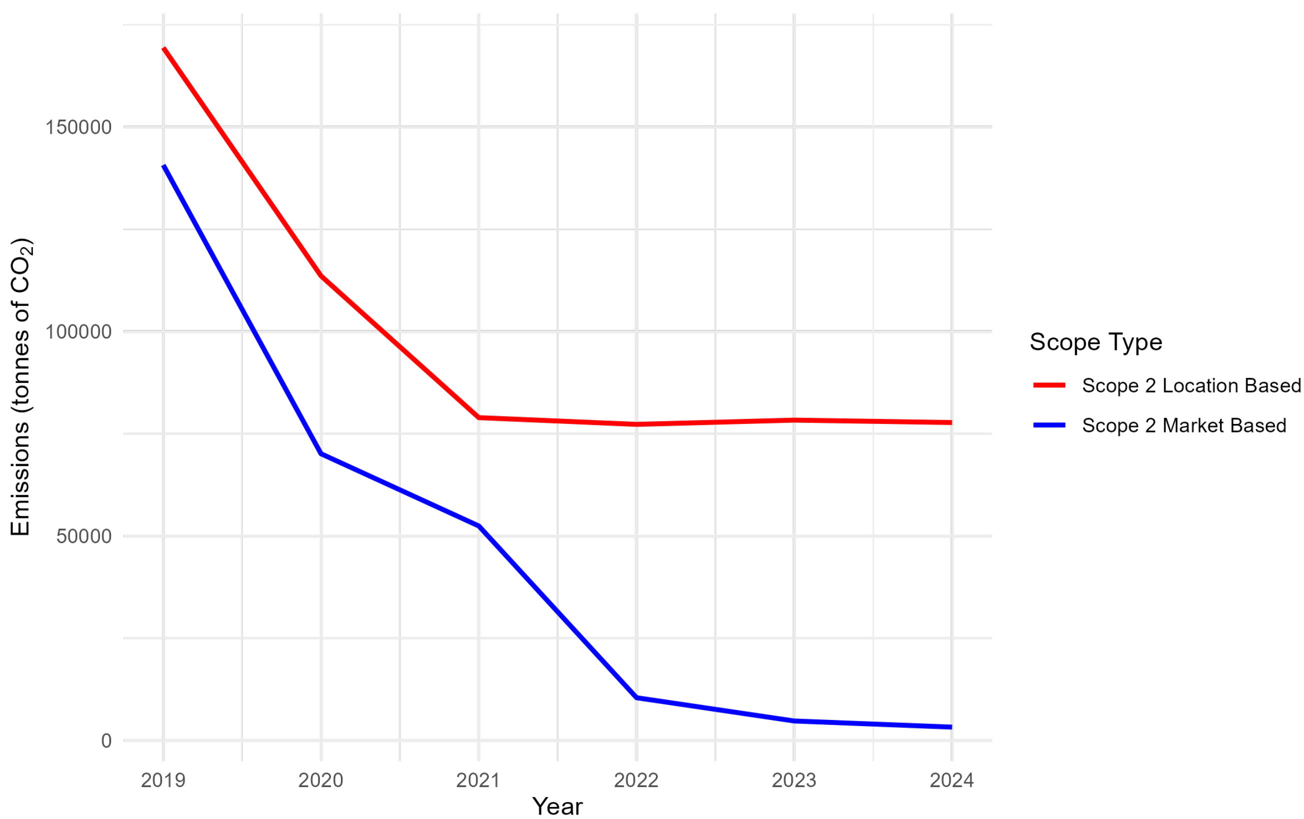

Figure 2: Scope 2 LB vs MB over time in tonnes of CO_2.

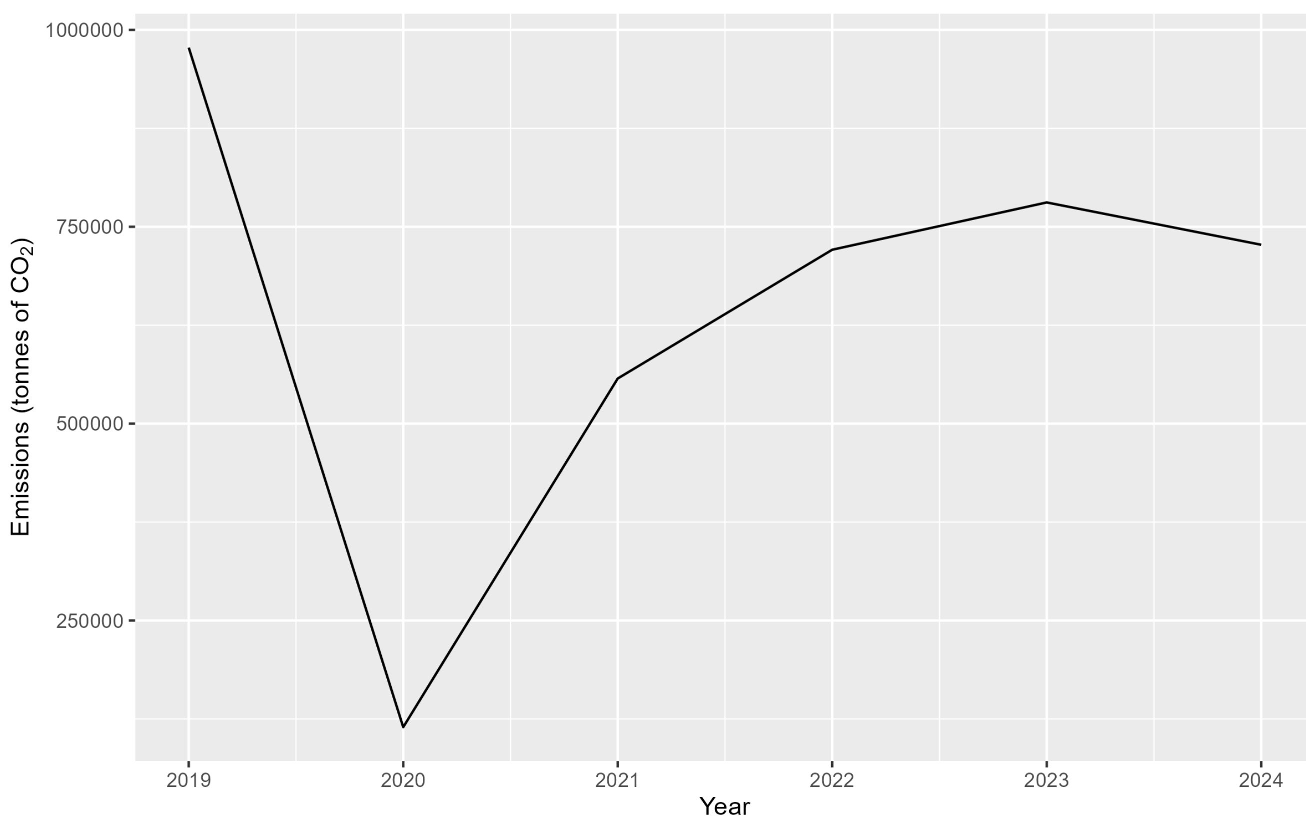

Figure 3: Scope 3 over time in tonnes of CO_2.

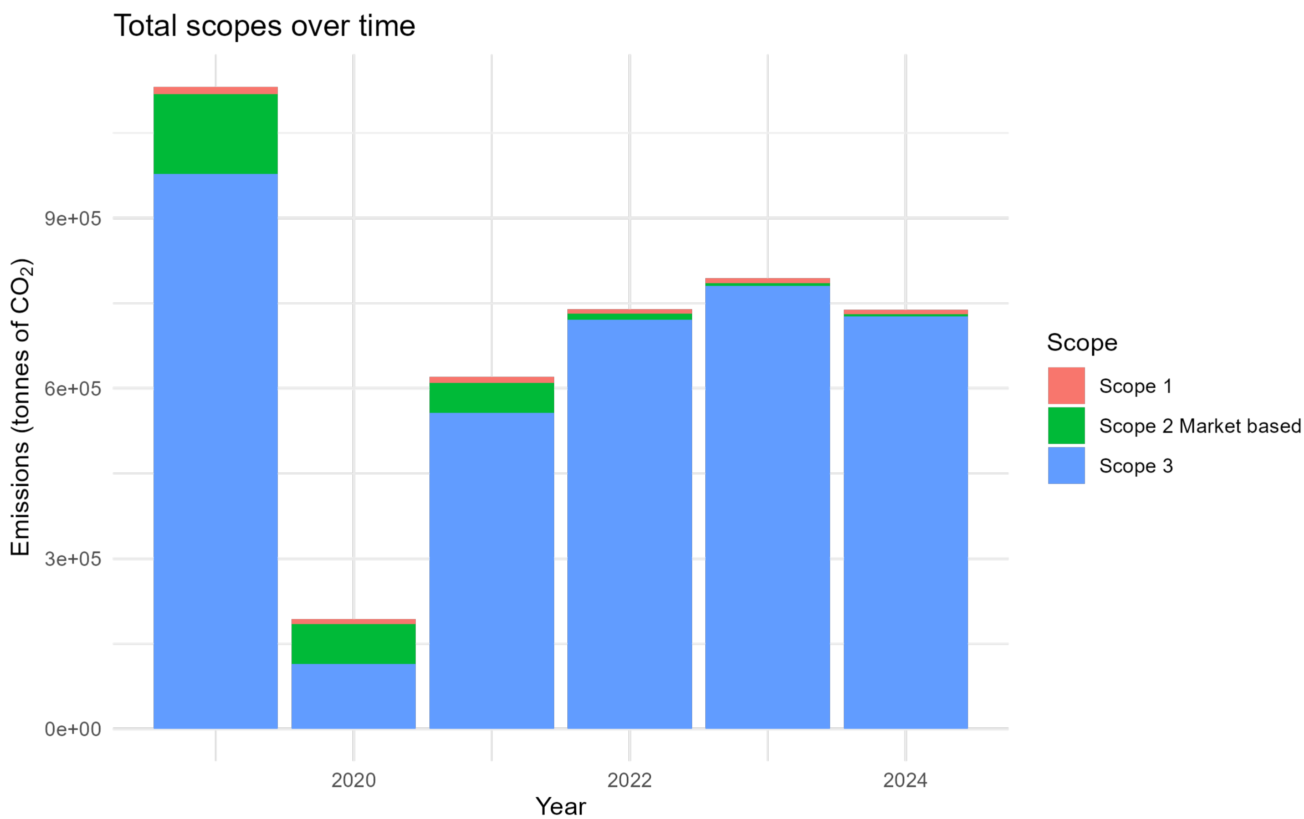

Figure 4: Total Scopes over time in tonnes of CO_2.

First we see that the scope 1 in Figure 19 is quite volatile : increasing in 2020 and 2022 and decreasing in 2019, 2021 and 2023. Furthermore, we see a huge decrease in 2020 corresponding to the pandemic which is quite clear to understand : there were no employees in the office meaning that all the offices, the industrial facilities or the company-owned vehicles were not running.

Then we see that Scope 2 MB and LB follows the same decreasing trend in Figure 20. Witnessing that the Scope 2 LB is always greater than the Scope 2 MB, we can suppose that CapGemini is trying to reduce its environement impact by buying electricity sourced from renewable energy.

Next, Scope 3 exhibits an atypical behavior in Figure 26. When analyzing the curve more closely, we observe a significant decrease between 2019 and 2020, similar to Scope 1 and for the same reasons. Following this decline, CapGemini appears to have partially recovered its past emission levels but at a lower magnitude, with emissions stabilizing over time.

Lastly, through the repartition of the Scopes in the total Scopes(that is computed with the scope 2 market based and not the location based) over time Figure 22 it seems that most of the total Scopes can be explained with Scope 3 (Scopes 1 and 2 are almost irrelevant regarding the total gas emissions of CapGemini over the year). This could also mean that policies targeting the reduction of direct emissions costs (Scopes 1 and 2) may be ineffective.

Now that we quickly took a look at the state of CapGemini, we will try to understand better this company and its GHG management policy.

3 Study of the firm

3.1 What About the Firm?

3.1.1 Goal of the firm

Capgemini is a French company founded in 1967 by Serge Kampf, originally under the name Sogeti. From its inception, Sogeti positioned itself as a key player in the technology sector by offering IT support and data processing services. The company adopted the name Capgemini following its merger with CAP and Gemini Computer Systems in 1970.

3.1.2 Evolution of the firm

Since then, Capgemini has steadily expanded its operations, with a strategic focus on cloud computing, cybersecurity, and more recently, artificial intelligence. The year 2018 marked a significant turning point with the launch of the AI@Scale initiative, which was designed to accelerate the large-scale adoption of AI solutions by its clients. This strategic shift was further reinforced in 2019 by the acquisition of Altran Technologies, a global leader in engineering and R&D services. This acquisition significantly strengthened Capgemini's positioning in the global AI ecosystem and marked the beginning of a notable rise in R&D investments.

3.1.3 Greenhouse Gases Management Policy

Capgemini aims to achieve carbon neutrality by 2040, with science-based targets validated by the SBTi. By 2030, it plans to reduce Scope 1 and 2 emissions by 80% and business travel emissions (Scope 3) by 55%, compared to 2019. The strategy also includes high-quality carbon offsetting, comprehensive emissions tracking, and recognition from CDP's A-list.

Knowing that Scope 3 accounts for the majority of total emissions and therefore of GHG emissions, these policies may be a positive signal. However, they cannot, on their own, guarantee a significant reduction in future GHG emissions. This reduction could also be attributed to other factors. Indeed, by dematerializing their activities, the company may reduce its Scope 1 emissions.

However, this shift would likely be reflected in an increase in Scope 3 emissions, which does not necessarily imply an overall reduction in total emissions.

4 Forecasting

Now that we have analyzed the significance and behavior of the scopes over time, it may be insightful to attempt forecasting their values. However, it is important to emphasize that, due to the limited amount of data, any forecasting results should be interpreted with caution. Given that we only have six time periods, an ARIMA model would likely lack robustness. Therefore, we will restrict ourselves to a linear model to capture general trends, and we will assess the reliability of the forecasts using confidence intervals. Let us precise that before running our prediction we will use the log values then turn them again in exponential to avoid some impossible negative values. We have the following results :

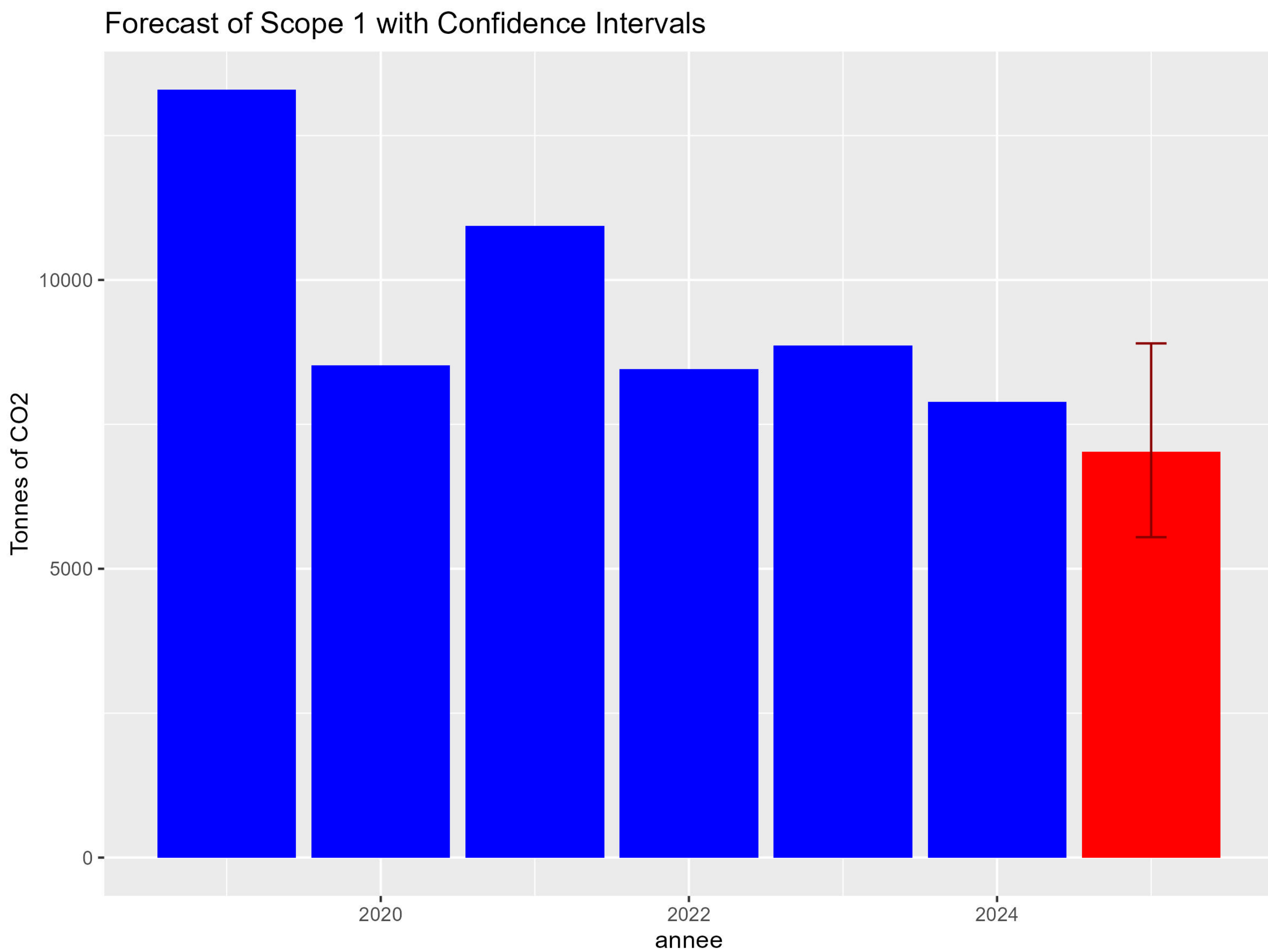

Figure 5: Forecast of Scope 1 over time in tonnes of CO_2.

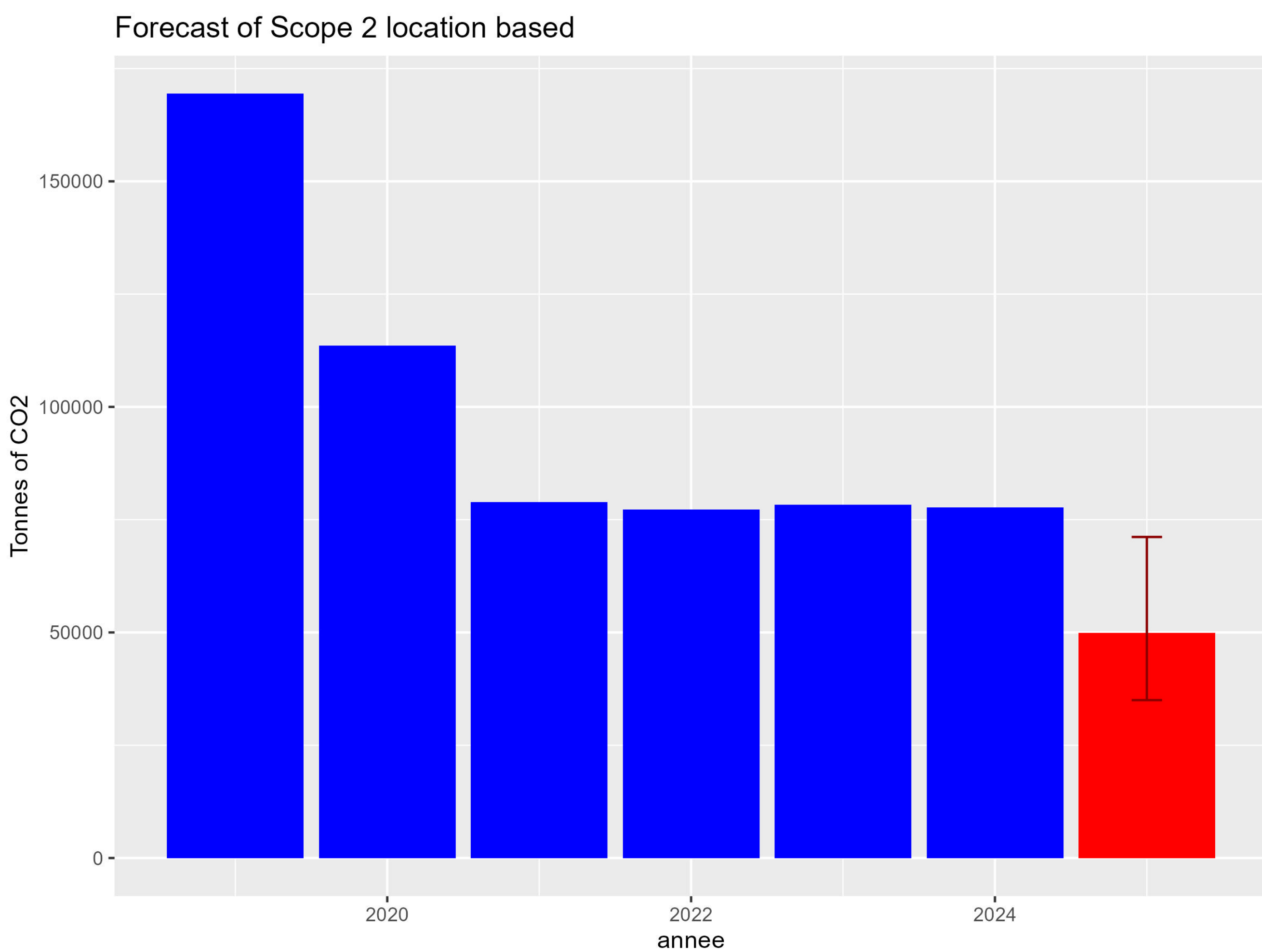

Figure 6: Forecast of Scope 2 LB over time in tonnes of CO_2.

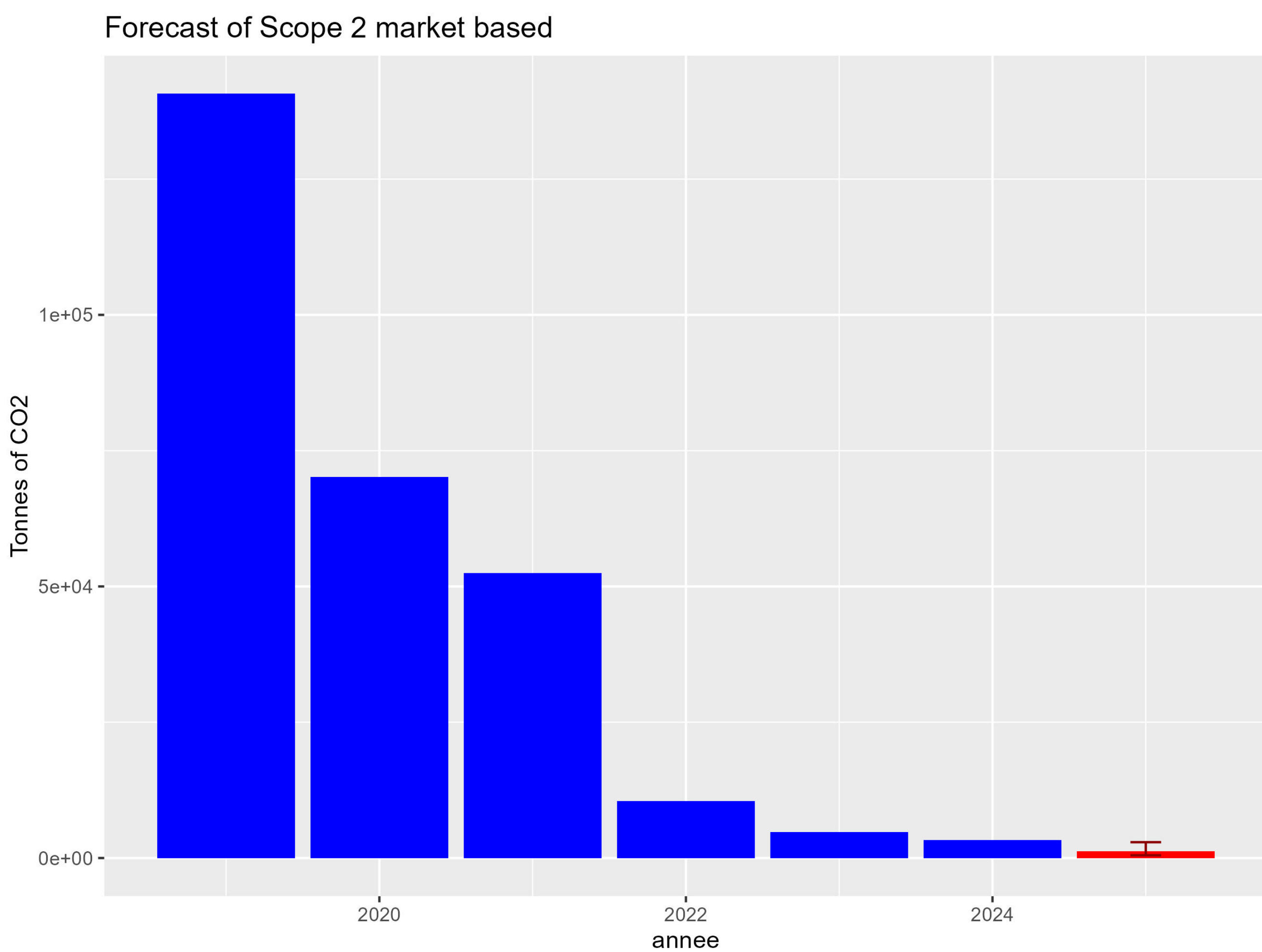

Figure 7: Forecast of Scope 2 MB over time in tonnes of CO_2.

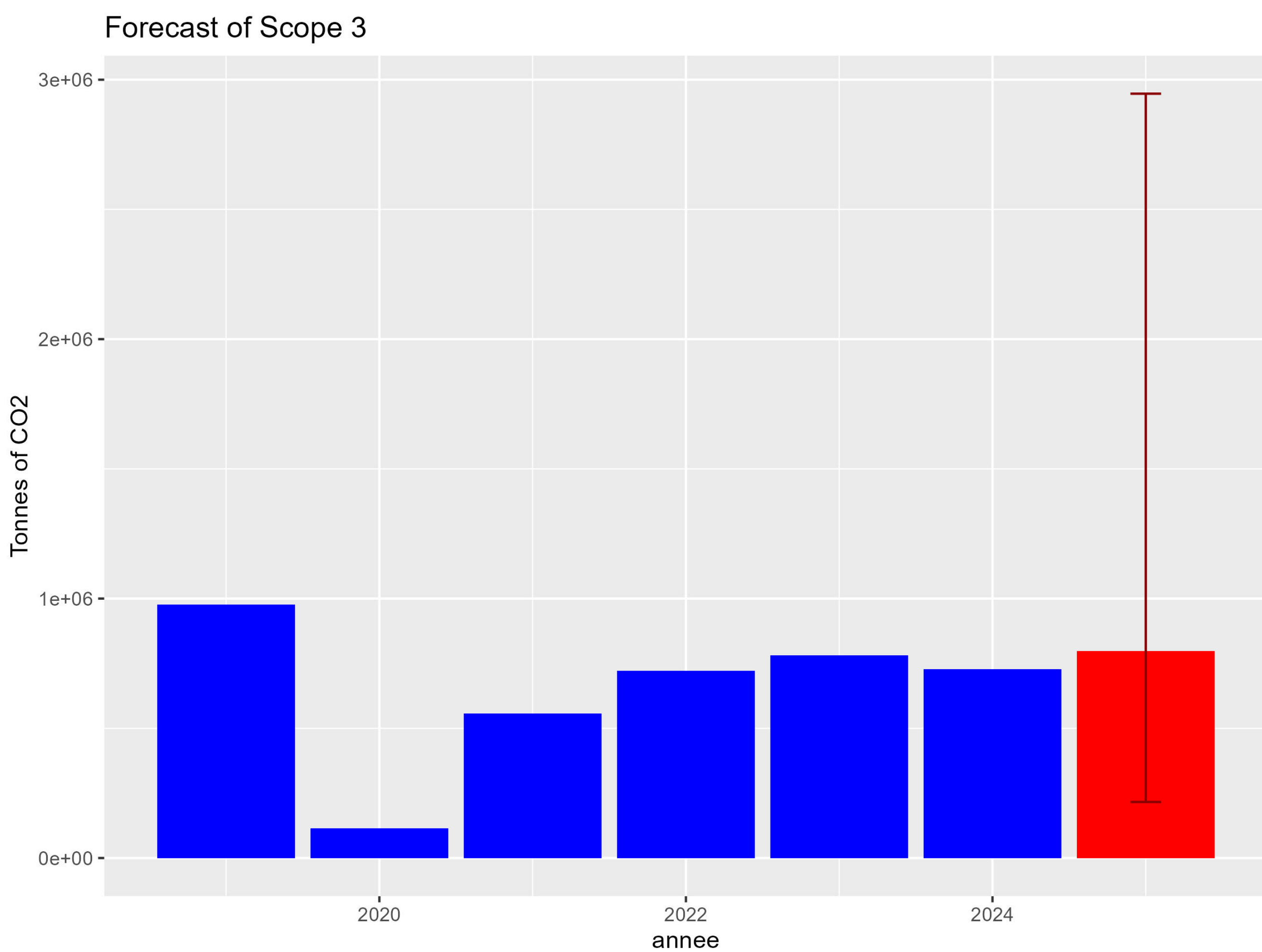

Figure 8: Forecast of Scope 3 over time in tonnes of CO_2.

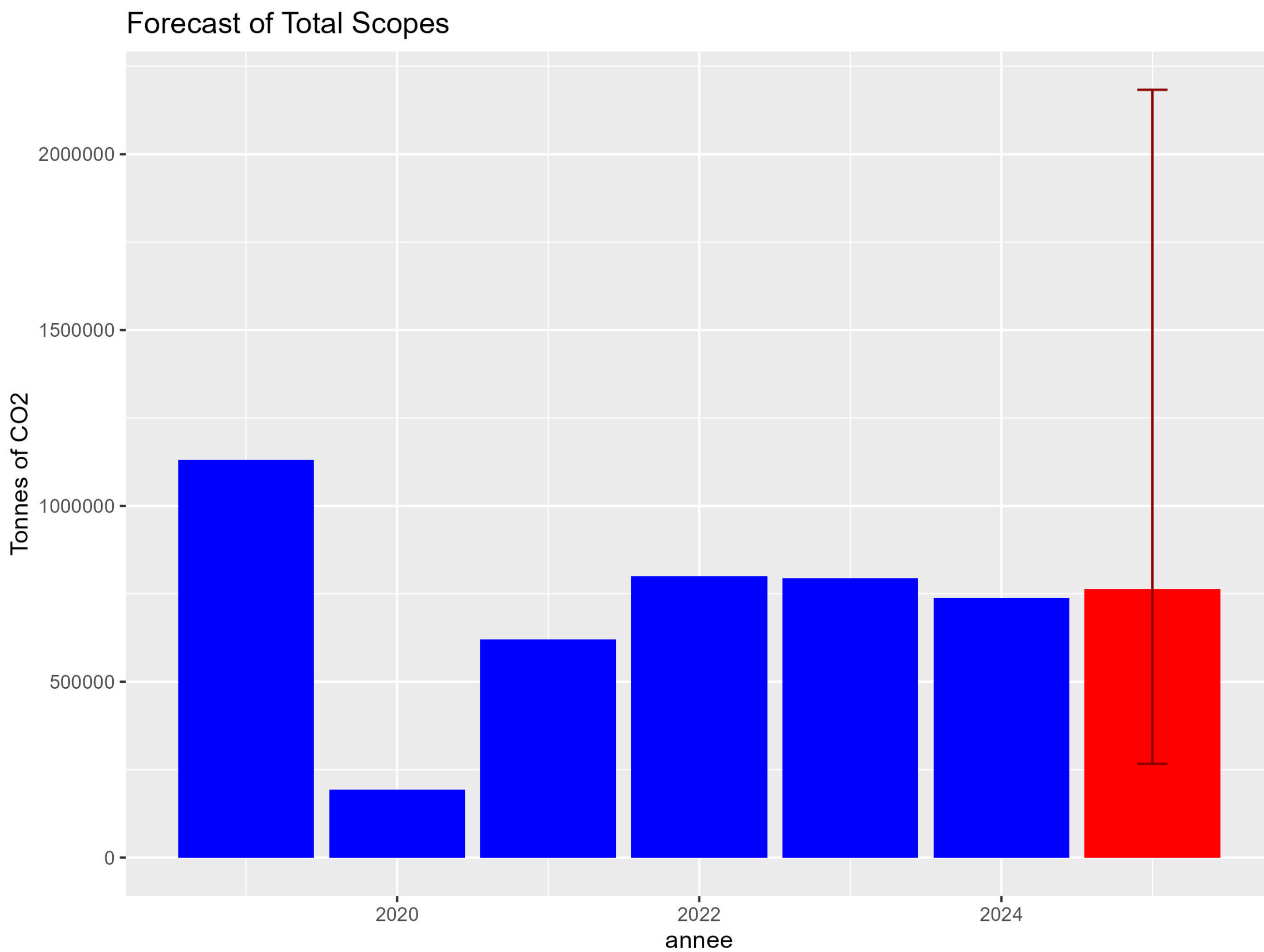

Figure 9: Forecast of total Scopes over time in tonnes of CO_2.

Only the trend indicated by the Scope 2 location-based emissions can be meaningfully interpreted. Because of the shape of the confidence interval we could say that in 2025 the Scope 2 location based should be at list slightly decreasing. From an econometric perspective, the linear forecasting model used here lacks robustness due to several limitations. Primarily, the short time series with insufficient periods undermines the reliability of the parameter estimates and limits the model's capacity to capture potential dynamic patterns or structural breaks. To overcome these shortcomings, a more comprehensive approach would require a longer dataset spanning multiple years. Such enhancements would improve the credibility and interpretability of the emission forecasts.

5 Scopes-Unreliable indicators?

The disclosure of the Scopes (particularly the 2 and 3) aren't as clear as we would like and the search of these data is quite hard. One question remains : why the firms aren't as open as we may think regarding the disclosure of these information?

5.1 Information Disclosure

First, we need to understand that the disclosure of such information is not always voluntarily undertaken by companies but is rather mandated, as is the case in certain countries within the European Union, such as the United Kingdom, where since 2013 companies have been required to report scopes 1 and 2 emissions, or France, which also obliges companies to disclose their scopes 1 and 2 emissions as well as partially scope 3. The objective of these countries is, by obliging firms to disclose these information, to reduce the GHG emissions [Chika Saka, 2014]. Indeed as stated by [Jianzu Wu, 2025] GHG emissions and its growing importance in our world could impact the consumption habits if the population and then impact negatively the revenue of firms. On one hand it could motivate firms to invest in new technologies to reduce their GHG emissions and water consumption so that they will improve their image in the eye of the public opinion, on the other hand they could keep their technology and don't change a thing but hide these "costs".

5.2 Lack of transparency

As we stated before finding and understanding the Scopes in the data provided by the companies is not easy. We also saw that they are able to move some values from the Scope 1 into the Scope 3. But as seen before not all companies are obliged to publish their Scope 3 based on their location. With this in mind we can suppose that some companies delocalize their activities in countries where the labor , the electricity and the water is cheaper which could be countries where they are not obliged of publish the Scope 3.

We analyzed the Scopes over time and tried to interpret it. Even if we depicted the lack of data as the sole issue in terms of forecasting, we also so that it is not always in the interest of the firms to display their GHG emissions and repartition. There is also a lack of transparency in terms of electricity and water consumption. This is what motivated the next part of our study : find a new method to modelise the electricity and water consumption related to AI.

Part III

Understanding of an AI

Before diving into the more technical aspects of the calculations, we need to have a clear understanding of how AI functions.

6 How does an AI work

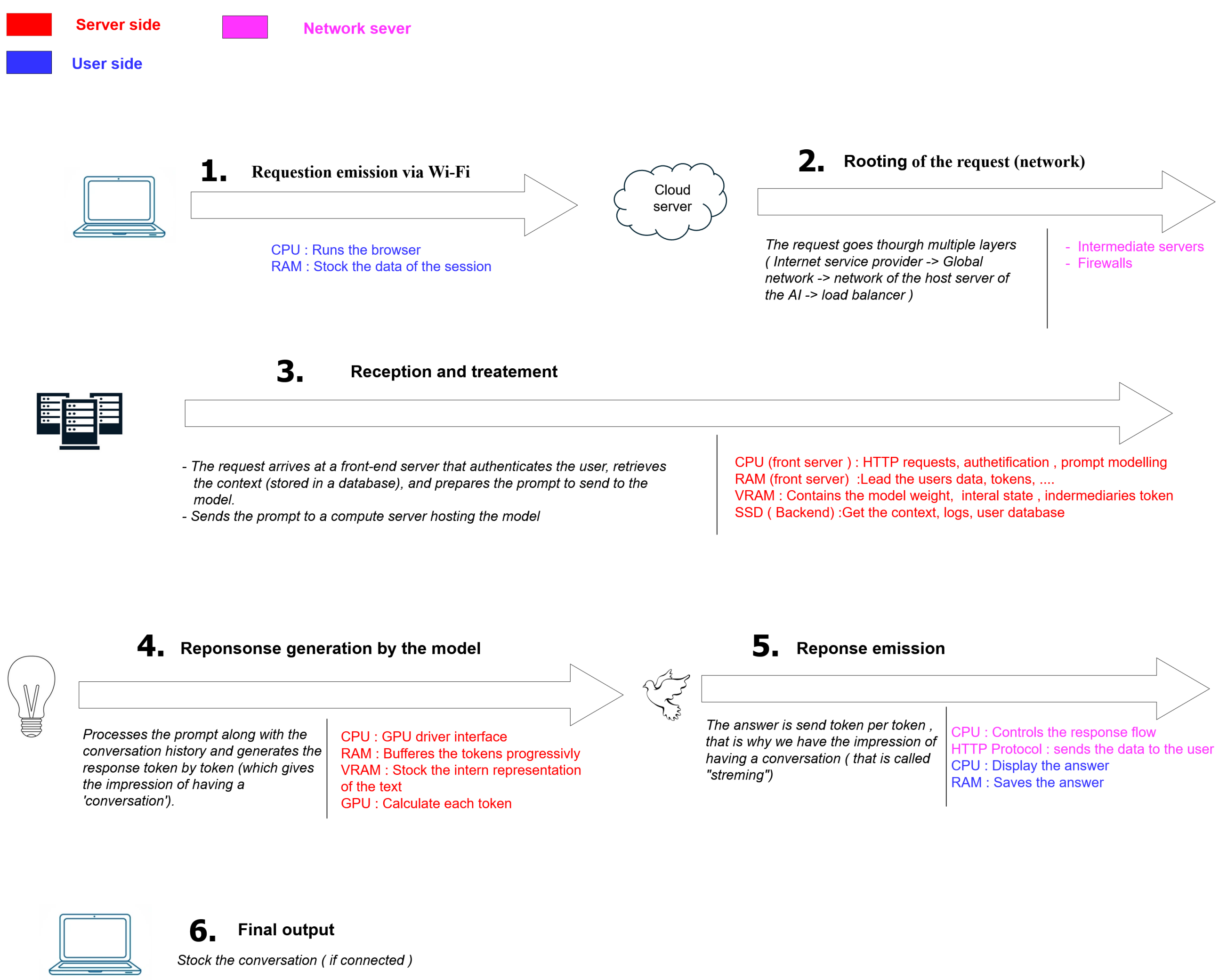

Figure 10: Overview of the functioning of an AI.

First the user at home sends a requests via WIFI to a cloud server. This request will go through multiple layers and then reach a datacenter where the model is stocked. Then the request arrive in a front server that will identify the user , give context to the request and will prepare the prompt that will be send to the server that contains the model. Next the server will treat the prompt regarding the historical record and answer to the request token by token (this is what simulates a "conversation"). Finally the answer is sent, again, token by token to the user. With everything in mind we broke every step of the operation and looked at which components may be involved. Now we will dig deeper into the calculus

7 Calculus : Electricity Consumption

7.1 GPU & VRAM

Because of the issues to monitor and estimate the VRAM consumption we will try to estimate it by playing with the variables. We need 3 information to assess the consumption of a GPU (Energie expressed in Watt per hour):

- The power of the GPU (P is in Watt)
- The usage time (T in seconds)
- The GPU utilization percentage (UP is in percentage)

The GPU utilization percentage will be manipulated to simulate the impact of the VRAM (100%) Then we will use the following formula :

$$Energie(Wh) = \frac{P \times T \times UP}{3600} \tag{1}$$

7.2 CPU

We need 6 information to assess the consumption of a CPU (Energie expressed in Watt per hour) :

- The usage time (T in seconds)
- Frequency (F in Hertz)

- Tension of the core (V in Volt)
- The CPU utilization percentage (UP is in percentage)
- Intern consumption capacity of the CMOS circuit (C in Farad)
- Number of cores of the CPU (n)

Then we use the following formula :

$$Energie(Wh) = \frac{\sum_{i=1}^{n} F_i \times V_i^2 \times UP_i \times C_i \times T}{3600} \tag{2}$$

7.3 Ram

We need 3 information to assess the consumption of a RAM (Energie expressed in Watt per hour) :

- The usage time (T in seconds)
- The capacity (C in Go)
- Power consumption (P in watts per gigabytes of RAM)

Then we use the following formula :

$$Energie(Wh) = \frac{C \times P \times T}{3600} \tag{3}$$

7.4 SSD

We need 3 information to assess the consumption of a SSD (Energie expressed in Watt per hour) :

- Average power for the state i (P_i expressed in Watt per hour)
- Percentage of time in the state i (t_i in percentage)
- Usage time (T in seconds)

Then we use the following formula :

$$Energie(Wh) = \frac{(P_{repos} \times t_{repos} + P_{lecture} \times t_{lecture} + P_{ecriture} \times t_{ecriture}) \times T}{3600} \tag{4}$$

Now that we obtained all the formula we need to know on which device we are going to "run" our request. This mean creating an average user and identifying the average components used in servers located inside the datacenter.

Concerning the customer side we identify the standard user as someone with a laptop with an intel Cleron N4500 as a CPU and a Ram DDR4 with a capcity of 8Go based on multiple sources [Canalys, 2025] [Nikolov, 2025] [VMR, 2025]

Concerning the server sides we will conceive a database[Google, 2025] [Amazon, 2025] [Microsoft, 2025] [OVH, 2025] based on the fact that CapGemini models are hosted in various datacenters like Google Cloud Platform, Amazon Web Services, Microsoft Azure or OVH Cloud. From this we draw a profile for the average components:

- a GPU with an average power of approximately 320.82 Watt
- a CPU named intel Xeon Platinium 8370C [AWS, 2025]
- A Ram type DDR4 of 128 Go
- A Samsung PM9A3 SSD

Now that with have the devices on which we want to estimate the consumption, let's head out to calculus.

8 Computation : Electricity consumption

8.1 Step 1

For the CPU, it is particularly difficult to obtain both the internal consumption capacity of the CMOS circuit and the core voltage. Therefore, we will use average estimated values based on various benchmarks: 0.85 volts[Notebookcheck, 2025] for the core voltage, and 50 pF[Instrument, 1997] for the load capacitance in a typical laptop. We also have that on average the usage time of the CPU when he is loading the browser lays between 0.1 and 0.3 seconds (based on multiple benchmark done with the Performance monitor) and that the CPU usage lays really close to 20% (also based on benchmarks). We obtain that the CPU consumes $3.179 * 10^{-12}$Wh.

We have that on average the Power consumption is between 0.3 and 0.5 Watt per Go for a DDR4 Ram and that the time a session with CapGemini or any AI model last on average between 4 and 5 minutes [TechnoLlama, 2025]. We obtain that the Ram consumes 0.24Wh.

With this, on average the Step 1 consumes 0.24Wh.

8.2 Step 2

The Step 2 concerns all the security check ups down before transmitting the request. We know that the request needs to go through firewalls but also intermediary servers. It is really hard to estimate these consumptions so we will suppose that they are really low (which is not an understatement based on the fact that these processes are not heavy).

8.3 Step 3

Regarding the GPU, an inference request to the model last more or less 2 seconds which means that the consumption will be on average of 0.18Wh.

Regarding the CPU, based on benchmark available on GitHubs and articles [Qi Fan, 2015] we have that:

- Receiving the request takes 15 ms and uses 20% of the CPU
- Authenticating the request takes 40 ms and uses 40%
- Prompt formatting takes 100ms and uses 60% of the CPU

Obtaining precise values for both the internal capacitance of the CMOS circuit and the core voltage is particularly challenging. Therefore, we adopt average estimated values based on various benchmarks: 0.9 volts for the core voltage and 200 pF for a server-grade processor based on sources [Intel, 2025] [Instrument, 1997]

We obtain that the receiving part consumes on average $4.04*10^{-11}$Wh , the authentication part consumes on average $2.43*10^{-10}$Wh and the prompt formatting consumes on average $8.09*10^{-10}$Wh. In total the CPU consummation comes down to an average of $1.09*10^{-9}$Wh.

Regarding the Ram we have that on average the Power consumption is on average between 0.3 and 0.4 Watt per Go for a DDR4 128 Ram and that the usage time is between 100 and 300ms (based on technical datasheet for this profile of Ram). We obtain that Ram consumes on average $0.28*10^{-2}$Wh.

Regarding the SSD (with the help of technical datasets [Samsung, 2025]) a SSD consumes on average $0.04*10^{-2}$Wh.

On average, step 3 consumed 0.18Wh.

8.4 Step 4

Regarding the CPU, to emulate its VRAM usage, we will suppose that the usage percentage of the GPU lays between 90 and 100%. Being still in the datacenter (as a local sever) we can use the same GPU as the one used previously inside datacenter. The issue remains in the time it takes a GPU to calculate each token, because it differs from one model to another. Based on different study and technical datasets[Nvidia, 2024] [Ramos, 2025] we can take as an average the fact that a GPU can calculate 40 token per second. The last issue we have is that there's no concrete answer to the question :" On average how many tokens are calculated per response", so to try to estimate that i took my request from Chatgpt and obtained an average of 148 tokens. We obtain that on average the GPU consumes 0.33Wh.

Regarding the CPU,To estimate the time needed we can take the time used per the GPU (the formatting and generation of the tokens are done simultaneously so we have the illusion to "speak" with someone). To try to emulate each of the steps we suppose that the percentage is of 100%. All these hypothesis comes from computations found on github [Github, 2023] and articles [Pujiang He, 2024]. We obtain that on average a CPU consumes $4.98 * 10^{-8}$.

Regarding the Ram consumption, the only missing information is the time of usage. We can suppose that the time is similar as the GPU one in this step because of the simultaneity of this step (to be more precise we can add a noise of 1 second). We obtain an average consumption of 0.07Wh.

In total, step 4 consumed on average 0.40Wh.

8.5 Step 5

As we have all the information we can go straight to calculus : the CPU consumes on average $3.18 * 10^{-12}$Wh.

8.6 Total

Thus, we find that, on average, a single request consumes approximately 0.82Wh of electricity. This result is consistent with findings from other researchers, such as [Nidhal Jegham, 2025], who report that electricity consumption per request lays between 0.43 Wh and 33 Wh, depending on the complexity of the model and the length of the request.

9 Computation : Water consumption

To estimate the cooling cost associated with each component, several key considerations must be taken into account. First, the heat generated by a component is closely related to its electrical consumption, since nearly all the energy consumed is converted into heat. Additionally, it is important to note that the vast majority of commercially available computers are not equipped with water-cooling systems, these are typically reserved for professional or gaming setups. Instead, cooling is generally achieved through air-based systems using fans, rather than thermodynamic processes.

Therefore, in order to estimate the amount of water required to cool each component for a single request, it is sufficient to rely on the electrical consumption observed within data centers and local servers.

10 Calculus : Water cooling

We have the following formula :

$$m = \frac{Q}{C * \Delta T} \tag{5}$$

with :

- The amount of water (m in kg)
- The amount of heat to dissipate (Q in Joules)
- The weight/thermal capacity of water (c which is approximately 4180 Joules/kg*C°)
- The allowable temperature rise in the water cooling circuit (ΔT in °C) (+-10) based on sources [Github, 2020]

We find that the total water consumption per request within a data center is approximately 0.05 kg. The associated costs can be broken down as follows:

- 0.04kg for the GPU cooling
- $4.39 * 10^{-9}$kg for the CPU cooling
- $6.00 * 10^{-3}$kg for the Ram cooling
- $3.73 * 10^{-5}$kg for the SSD cooling

We need to keep in mind that these results both, for electricity and water consumption represent the consumption for one request and not a "discussion". Based on our results, the cost of cooling a datacenter for 10 request is of a water bottle which is coherent with other studies [OpenAI, 2024].

Now that we have established a methodology to calculate the water and electricity consumption associated with the functioning of an AI system, we need to take a closer look at the infrastructure aspect of the problem: data centers.

Part IV

Data Centers

11 Motivation

Data centers are, among other things, the infrastructure where the LLMs used to power AI systems are stored and processed. Currently, there are approximately 6,000 data centers in operation, and this number is expected to rise to 8,000 by 2030. This increase reflects a growing demand for infrastructure that can be linked to the expansion of AI technologies.

In this section, we aim to examine the environmental impact of data center construction.

12 Localization

12.1 Mapping

Here we can see a hex-bin map of data centers in the world :

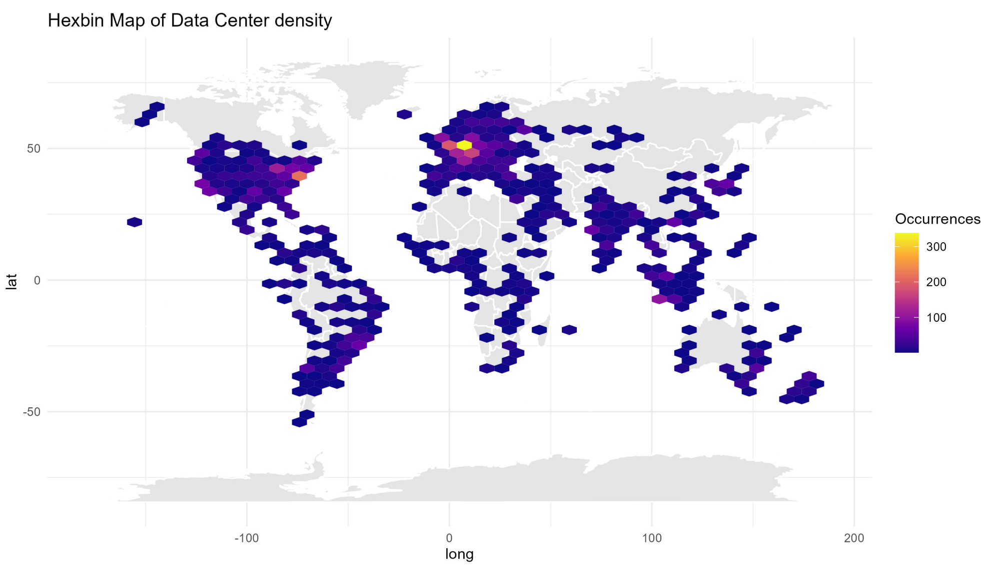

Figure 11: Hex-bin map of data centers

We can see that most of the data centers(Figure 29) are located in North America, Europe, and South America. This visual observation is backed out by the data, indeed 1505 data centers are located in the USA, 561 in Brazil and 459 in Germany.

	country	total_occurrence
1	United States	1505
2	Brasil	561
3	Deutschland	459

Figure 12: Top 3 countries

While the mean for a country is between 48 and 49 data centers, the top 3 countries seems to denote a little. We will try to understand which are the variables that are taken into account when a firm chooses the location of a future datacenter. To do this we will look at variables related to:

- the cooling process (temperature, soil moisture, drought factor, proximity to oceans/seas/rivers/lakes)
- the electricity consumption(Proximity to a big city)
- usual natural risks (Earthquake occurrences)

12.2 Variables

12.2.1 Temperature

We obtain the temperature data using the geodata package in R, which sources its information from the WorldClim database[WorldClim, 2020]. While seasonal variations naturally influence temperature, data centers do not change location throughout the year. Therefore, the key climatic indicators a firm may be interested in are the annual average temperature, the minimum temperature, and the maximum temperature.

We except that data centers needs to be located in areas where the average temperature is low.

12.2.2 Soil Moisture

We obtain the soil moisture data through the earthengine of google with NASA data [1]. The information we obtain are the soil moisture information from the 20th of june 2025 with an indicator that goes from 0 to 1, where 0 represents the lowest soil moisture (driest) and 1 the highest (wettest).

We can expect both signs for the coefficients of these variables. On one hand, low soil moisture appears necessary for constructing stable infrastructure; on the other hand, higher soil moisture can be an advantage for new data centers employing innovative cooling methods.

12.2.3 Earthquake

We obtained the earthquake data from the United States Geological Survey (USGS) [2]. Rather than considering only the exact locations of the earthquakes, we define a tolerance zone: if an earthquake has ever occurred within a 50 km radius of a data center, the corresponding variable is coded as "yes" in response to the question: "Has an earthquake ever occurred near this datacenter?"

We except that data centers shouldn't be located in areas where earthquake could occur.

12.2.4 Drought factor

The drought factor is a numerical indicator that quantifies the severity or intensity of drought conditions in a given region. We obtained the data from the National Centers for Environmental Information[3]. The variable used in our analysis is the Standardized Precipitation Index (SPI), a normalized metric that captures precipitation anomalies over various temporal scales. In this study, we use the annual average for the year 2024. The SPI is scaled from –3 to +3, with lower values indicating drier-than-normal conditions and higher values indicating wetter-than-normal conditions.

[1] Nasa Soil Moisture
[2] USGS Earthquake Database
[3] Drought Factor

We can expect both signs for the coefficient of the SPI. On one hand, it could be negative due to the need for dry air to optimize the cooling system; on the other hand, a positive coefficient may indicate that higher precipitation often corresponds to lower temperatures.

Values of the SPI	**Interpretation**
SPI > 2	Highly humid
$1.99 >$ SPI > 1	Humid
$0.99 >$ SPI > -0.99	Normal
$-1 >$ SPI > -1.49	Moderate drought
$-1.5 >$ SPI > -1.99	Severe drought
SPI < -2	Extreme drought

Table 2: Interpretation of SPI values.

12.2.5 Proximity to a big city

We define a large city as one with a population of at least one million inhabitants. This data is obtained from an R package (rnaturalearthdata), which sources its information from Natural Earth[4]. Based on this dataset, we derive two variables: the distance to a large city (a quantitative variable) and the proximity to a large city (a qualitative variable indicating whether the data center lies within a 50 km radius of a large city).

We expect that data centers prefer to be located near major cities to benefit from better connectivity, faster maintenance, and related advantages.

12.2.6 Proximity to oceans/seas/rivers/lakes

We load data on the locations of oceans, seas, rivers, and lakes from the same package used for proximity to large cities[5]. From this data, we create two variables: the distance to the nearest water body (a quantitative variable) and the proximity to a water body (a qualitative variable indicating whether the data center is located within a 10 km radius of a water body).

[4] Big city
[5] Water

We expect that data centers prefer to be located near water bodies to benefit from easier access to water supply for cooling their infrastructure.

12.3 Regression

12.3.1 GLM

First we will work with real distances to try to see if there is a tolerance for the distance to big cities and to water foot or if the relation is perfectly linear.

Distance We will work with a simple generalized linear model (GLM) and a Poisson family. We have the following equation:

$$\begin{aligned} \log(\text{occurences}_i) = \beta_0 + \beta_1 \times temperature_i + \beta_2 \times sm_i + \beta_3 \times matched_i \\ +\beta_4 \times spi_i + \beta_5 \times prox\_city\_dist_i + \beta_6 \times prox\_water\_dist_i + \epsilon_i \end{aligned} \tag{6}$$

We have the following results :

term	estimate	std.error	statistic	p.value
(Intercept)	1.3549	0.0472	28.733	0
temperature	-0.0031	0.0019	-1.6518	0.0986
sm	-0.4022	0.0733	-5.4858	0
earthquakeTRUE	-0.9194	0.5777	-1.5915	0.1115
SPI	0.0743	0.0138	5.3955	0
prox_water_dist	-7e-04	2e-04	-3.8796	1e-04
prox_city_dist	-0.001	1e-04	-16.1805	0
McFadden R2	0.0277			

Figure 13: Model Distance

We see that only the temperature and the earthquake coefficient are significant, both negative. We have that:

- Increasing the average temperature by 1°C decreases, on average, the expected number of data centers by approximately $1 - \exp(-0.0031) \approx 0.0031$, or 0.31%, ceteris paribus.

- Being near a location where earthquake happened, on average, decreases the number of data centers by approximately $1 - \exp(-0.9194) \approx 0.6012$, or 60.12%, ceteris paribus.

- The McFadden R^2 is approximately 2.77%, indicating that the model explains only 2.77% of the improvement in log-likelihood compared to the null model.

Proximity We will work with the qualitative variable regarding distances now. We have the following equation :

$$\begin{aligned}\log(\text{occurences}_i) = \beta_0 + \beta_1 \times temperature_i + \beta_2 \times sm_i + \beta_3 \times matched_i \\ + \beta_4 \times spi_i + \beta_5 \times prox\_city_i + \beta_6 \times prox\_water_i + \epsilon_i\end{aligned} \tag{7}$$

We have the following results :

term	estimate	std.error	statistic	p.value
(Intercept)	0.7346	0.0445	16.5165	0
temperature	-0.0088	0.0018	-4.8336	0
sm	-0.3859	0.0721	-5.3511	0
earthquakeTRUE	-0.429	0.578	-0.7423	0.4579
SPI	0.0329	0.0136	2.4147	0.0157
prox_waterTRUE	0.1871	0.0238	7.8751	0
prox_cityTRUE	0.7623	0.0251	30.3628	0
McFadden R2	0.0668			

Figure 14: Model proximity

The results are quite surprising and may induce that there is not link between data centers and the distance / proximity to big cities. But we need to take into account that we deal only with cities that have data centers, so the only conclusion we can draw is that being near a city /water foot is not enough to build more data centers, not that it is not relevant in the choice of finding an area to build data centers.

To assess this collinearity issue, we will use a Least Absolute Shrinkage and Selection Operator (LASSO) regression. First, we need to standardize the variable, then we run the regression :

	Coefficient	Variable
(Intercept)	0.5234329	(Intercept)
sm	-0.1446398	sm
prox_water	0.1509892	prox_water
prox_city	0.6522457	prox_city

Figure 15: Coefficient regression

The variables not retained by the model are temperature, earthquake, and SPI. It is plausible that soil moisture is influenced by temperature and may be correlated with SPI, which could explain their exclusion. Regarding the earthquake variable, it is reasonable to assume that companies investing in such costly infrastructure deliberately avoid locations prone to natural disasters.

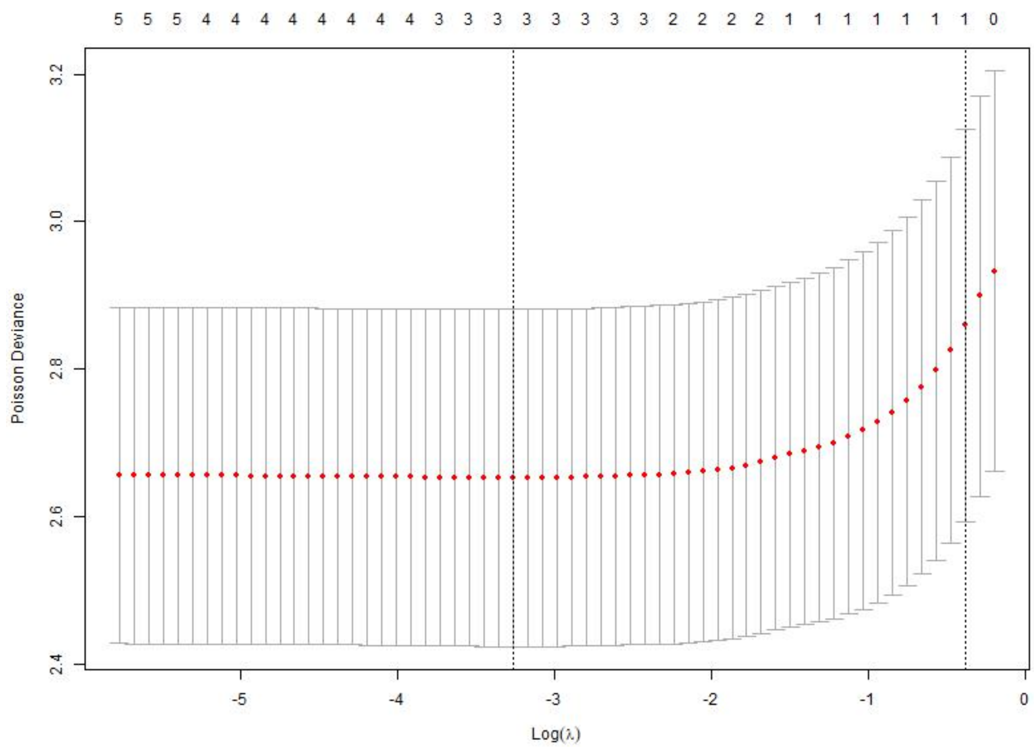

Figure 16: Lasso plot

12.3.2 Logit

The primary limitation of our dataset is the absence of cities without data centers. Addressing this issue may be feasible with more comprehensive datasets or greater computational resources. To mitigate this limitation, we will explore whether cities with a low number of data centers can be considered effectively as cities without any data centers.

To identify the optimal number of data centers for partitioning the dataset into two groups, we perform a one-dimensional thresholding procedure that minimizes the sum of within-group variances across the explanatory variables, similar to a single-split regression tree.

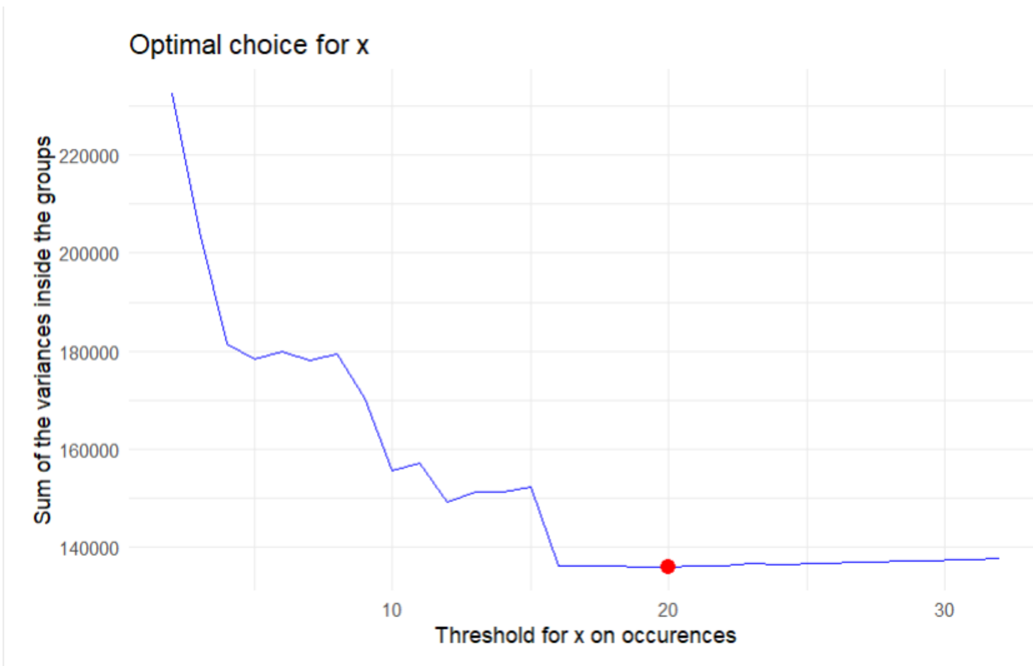

Figure 17: Optimal choice

We observe that the optimal choice would be to set the threshold at 20, suggesting that cities with fewer than 20 data centers are "sub-optimal" locations that could be considered as effectively lacking data centers. We will proceed by estimating a probit model using the variables selected through the lasso regression. The results are as follows:

term	estimate	std.error	statistic	p.value
(Intercept)	-7.1313	1.0842	-6.5774	0
sm	-0.9129	1.1359	-0.8037	0.4216
prox_cityTRUE	3.6453	1.0173	3.5832	3e-04
prox_waterTRUE	0.5942	0.3651	1.6278	0.1036
McFadden R2	0.1389			

Figure 18: Logit coefficients

First, let us examine the R^2. With a value of approximately 13.89%, it represents a significant improvement compared to the earlier models.

Furthermore, all variables are statistically significant, with coefficients exhibiting consistent signs: positive for the distance to water bodies and major cities, and negative for soil moisture.

However, these results are insufficient to fully identify all the factors influencing a firm's decision to build a data center. Additional variables of interest include the cost of electricity and water in each city, which could potentially enhance the explanatory power of the model.

Having established a clearer understanding of the determinants guiding a firm's choice of data center location, we will now investigate the effects of its implementation.

13 The issue of case studies

There are fourteen data centers registered in the Occitanie region, including four located in Toulouse. Initially, we considered studying the impact in Toulouse; however, the presence of multiple data centers and various confounding factors such as increased water consumption resulting from numerous construction projects makes the analysis quite hard.

For this reason, we focused on the city of Valbonne, near Nice, which hosts a single data center established in 2013, aiming to understand the effect of this data center's implantation on temperature, water, and electricity consumption. Valbonne has a population of approximately 13,000 inhabitants and covers an area of 19.15 square kilometers.

However, conducting a difference-in-differences analysis at the city level proves infeasible due to the lack of sufficiently granular data. At this scale, it is practically impossible to isolate the effect of the data center from other confounding influences, such as infrastructure developments, urban growth, or evolving socio-economic conditions that also impact consumption patterns. These confounders can bias the results and prevent clear identification of the causal effect.

Furthermore, identifying a truly comparable control city that has not experienced similar changes is difficult. Although Mougins was initially selected as a potential control city based on population and area, the presence of unobserved confounding factors undermines the validity of the difference-in-differences approach in this context.

Another issue is that this is still a recent phenomenon and that the Pandemic period had a huge impact on electricity consumption for example, so it is really hard to identify

Therefore, alternative methodologies or more localized data would be necessary to accurately assess the impact of the data center on local environmental and consumption variables.

However, thanks to the new methodology developed in the second part of our work, we are now able to estimate electricity and water consumption. According to information provided by Microsoft[6], an average data center processes between 10^9 and 10^{12} requests annually. Considering the typical size of data transfers, this corresponds to an estimated electricity consumption of approximately 820 MWh per year for an average data center. Meaning that the average water consumption per data center should be around 50,000 m^3 of water.

If we had access to the fitted data, we would be able to estimate the impact of such consumption on a city and subsequently draw conclusions using a difference-in-differences analysis. However, even in the absence of such data, it is reasonable to assert that these consumption levels are substantial, pertain to data centers of average size, and should raise concerns regarding the clear environmental impact.

Part V

Conclusion

First, we attempted to work with an existing indicator: Scopes. While it is generally a good measure of a firm's environmental impact, we identified a transparency issue. Indeed, disclosing greenhouse gas (GHG) emissions and their distribution is not always in the firm's best interest, for various strategic reasons.

To address this limitation, we returned to the fundamentals and investigated how AI systems operate. We found that each request to an AI model consumes between 0.43 and 33 Wh of electricity and approximately 0.05 kg of water. It is important to emphasize that these values represent minimum averages, based on usage by typical individual users. Firms, however, often employ

[6] https://azure.microsoft.com/en-us/blog/category/storage/

larger models and longer prompts, leading to significantly higher resource consumption.

Building on this, we examined the electricity and water requirements associated with data centers and sought to understand the geographical choices made by firms when establishing them. We found that data centers tend to be located near large urban areas, close to water sources (oceans, rivers, lakes), and in regions with low soil moisture. We also aimed to explore the direct impact of data centers on cities and their surroundings. However, due to a lack of granular data and the presence of multiple confounding variables, drawing robust conclusions proved impossible.

In conclusion, access to more detailed Scope-related data and finer-grained urban-level data would have significantly improved our analysis.

Part VI

Annexes

Figure 19: Scope 1 over time in tonnes of CO_2.

Values of the SPI	**Interpretation**
SPI > 2	Highly humid
$1.99 >$ SPI > 1	Humid
$0.99 >$ SPI > -0.99	Normal
$-1 >$ SPI > -1.49	Moderate drought
$-1.5 >$ SPI > -1.99	Severe drought
SPI < -2	Extreme drought

Table 3: Interpretation of SPI values.

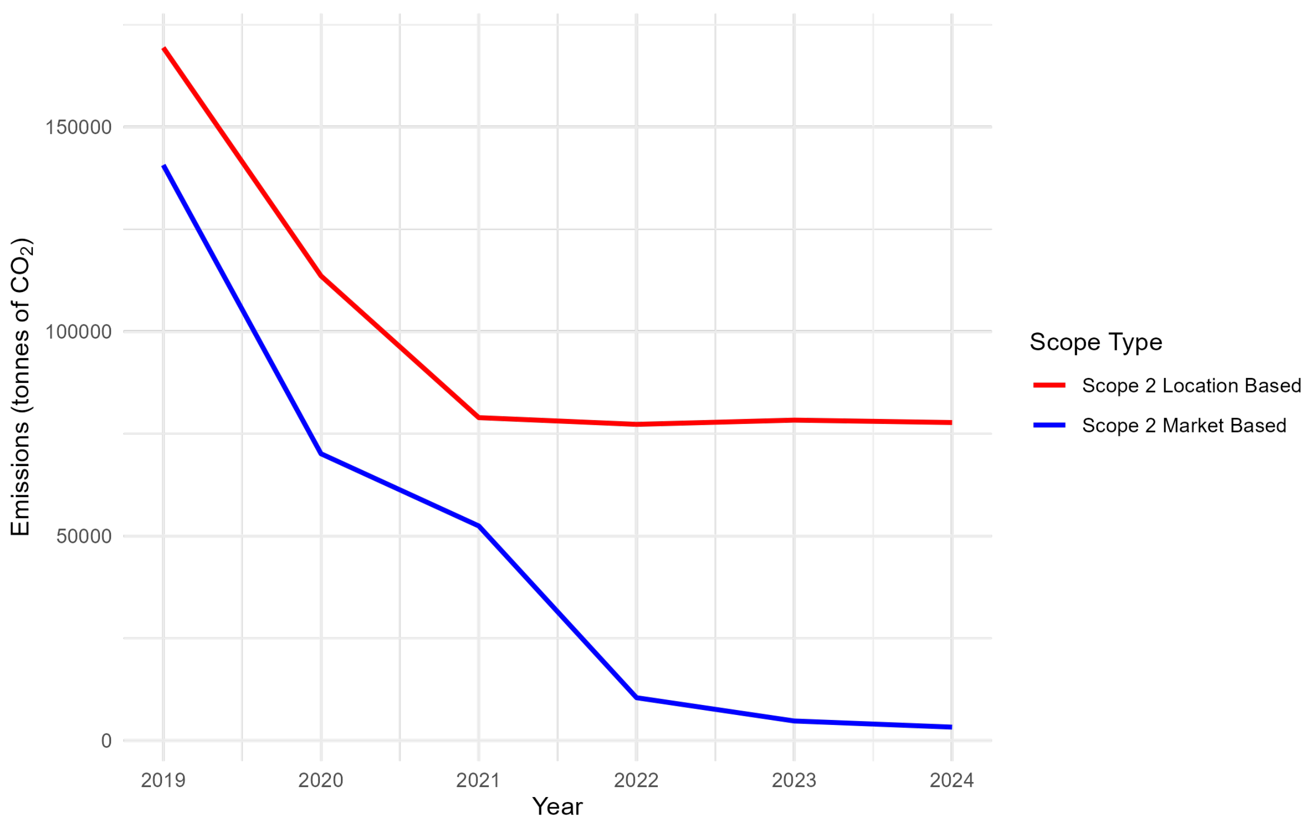

Figure 20: Scope 2 LB vs MB over time in tonnes of CO_2.

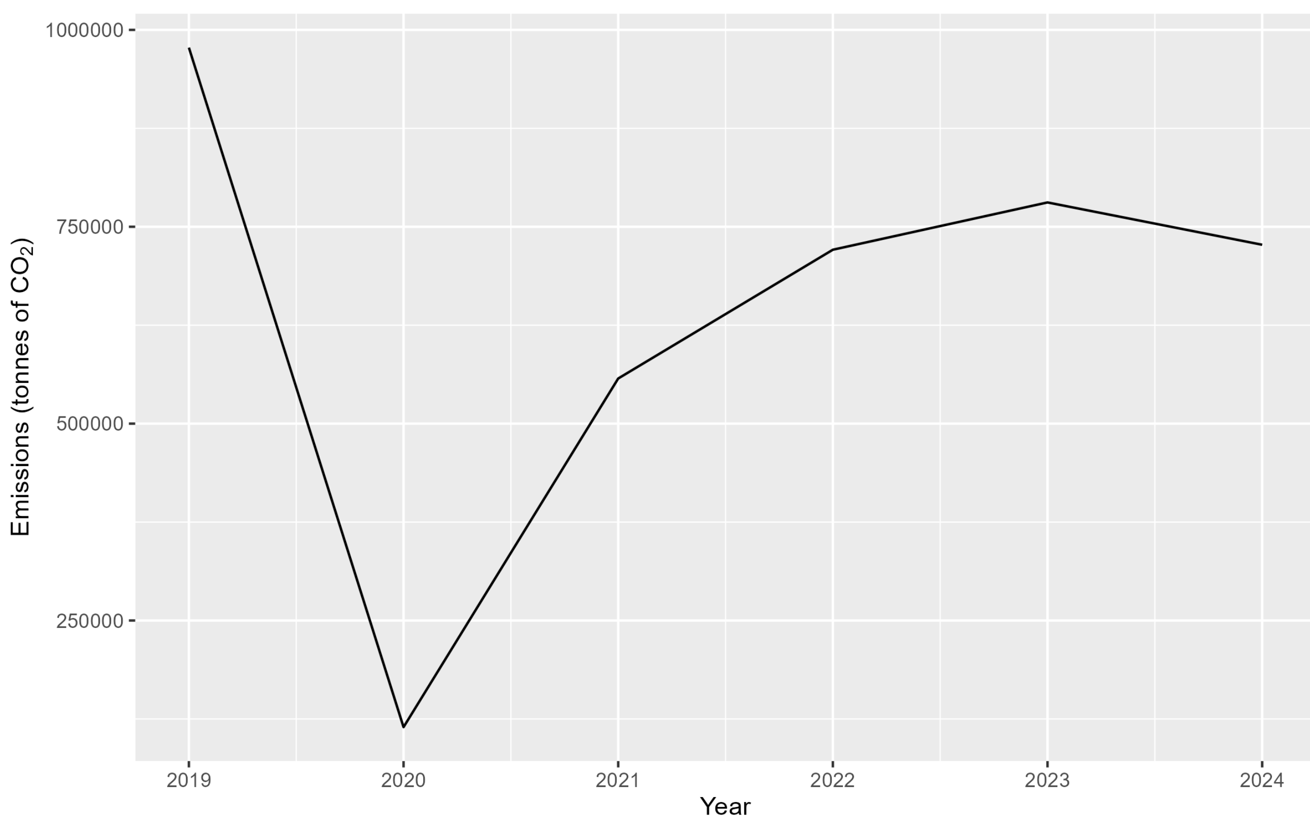

Figure 21: Scope 3 over time in tonnes of CO_2.

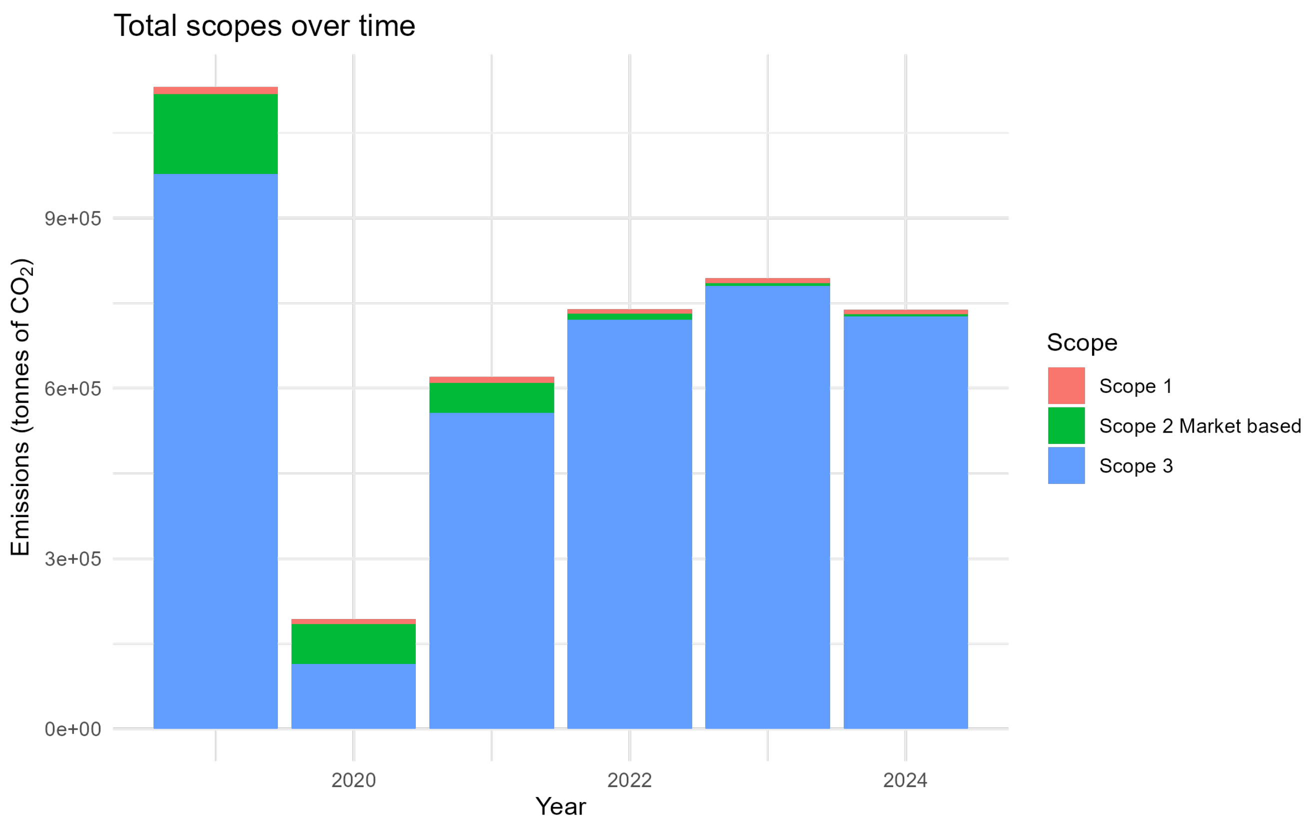

Figure 22: Total Scopes over time in tonnes of CO_2.

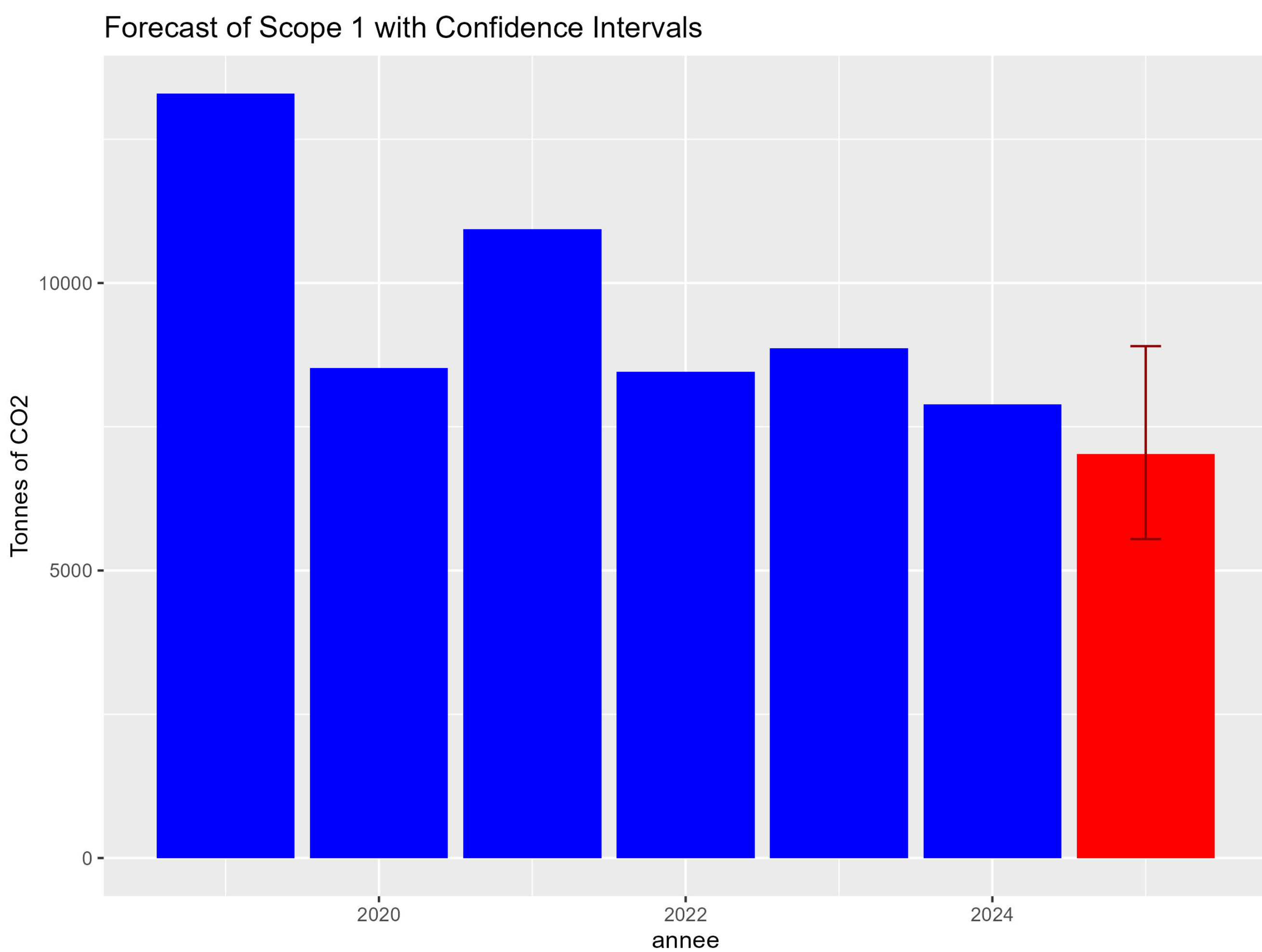

Figure 23: Forecast of Scope 1 over time in tonnes of CO_2.

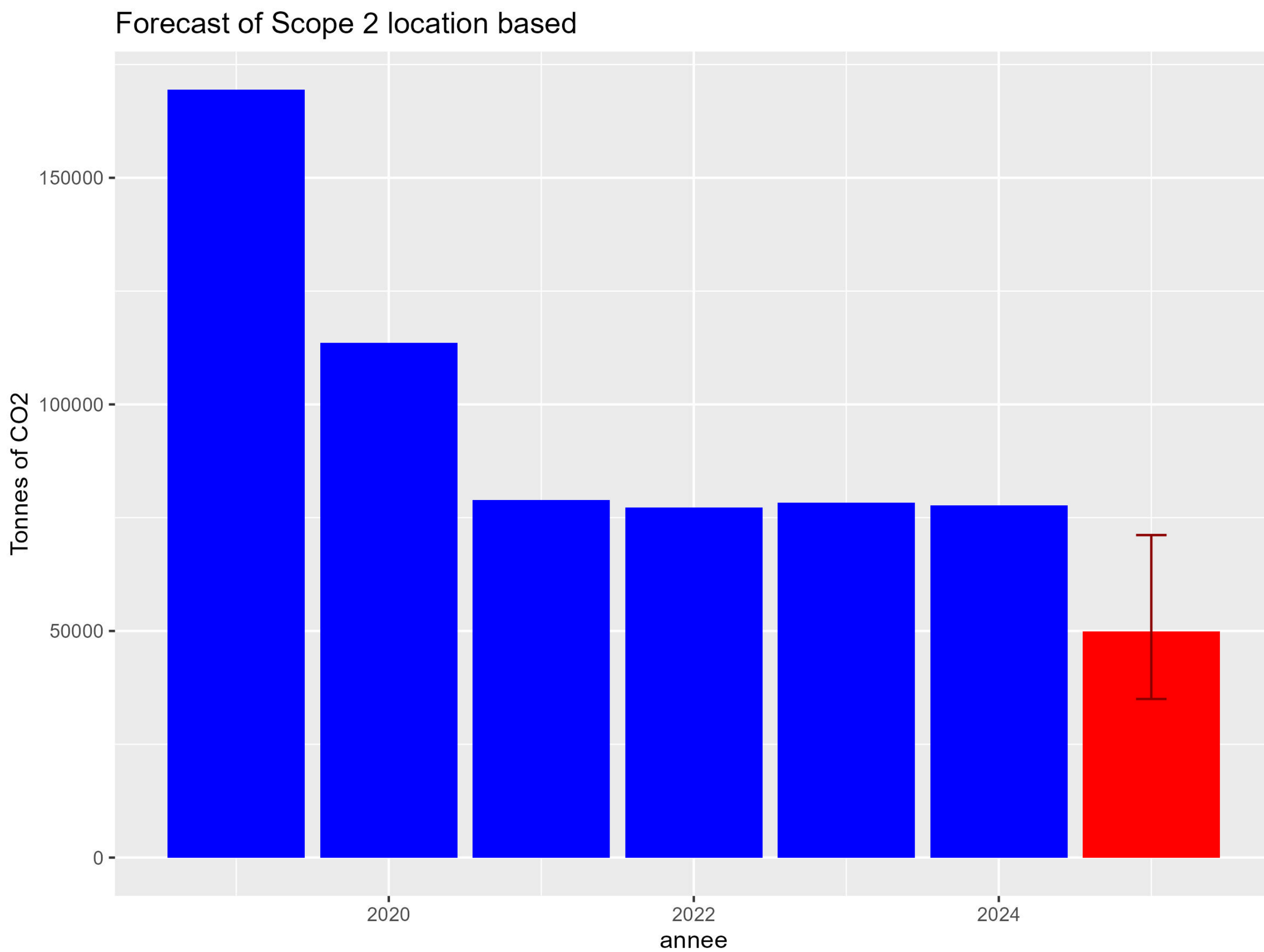

Figure 24: Forecast of Scope 2 LB over time in tonnes of CO_2.

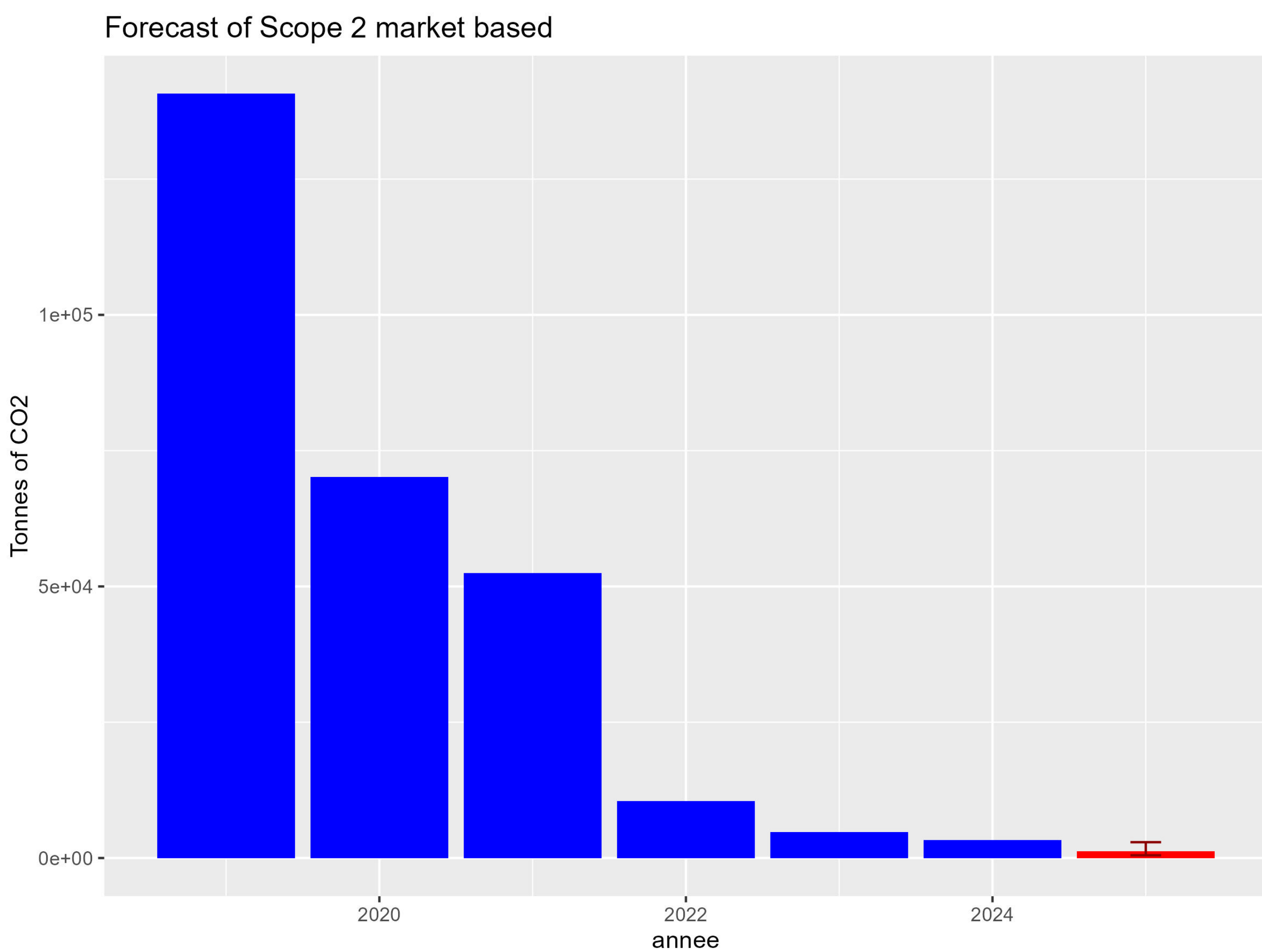

Figure 25: Forecast of Scope 2 MB over time in tonnes of CO_2.

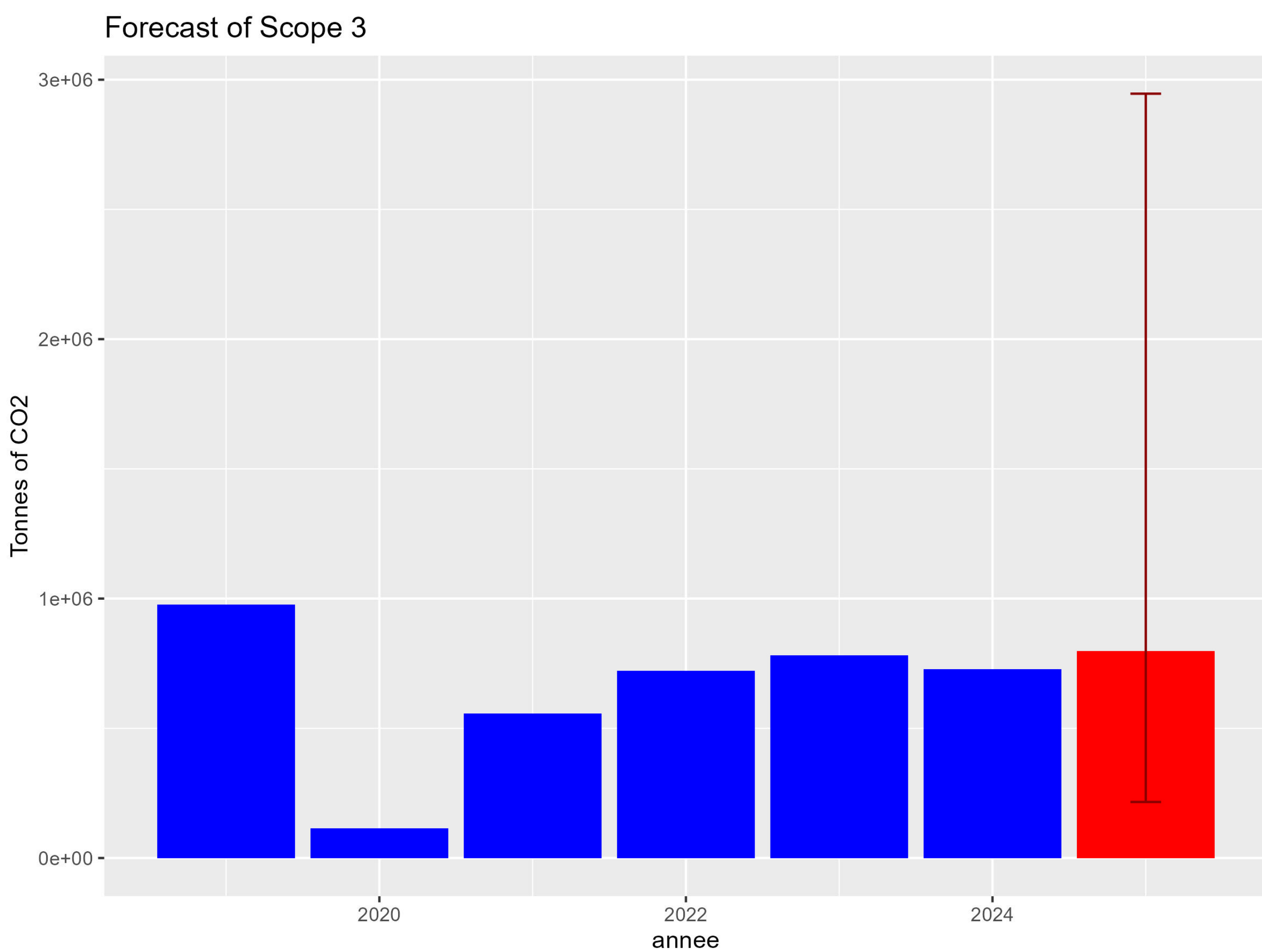

Figure 26: Forecast of Scope 3 over time in tonnes of CO_2.

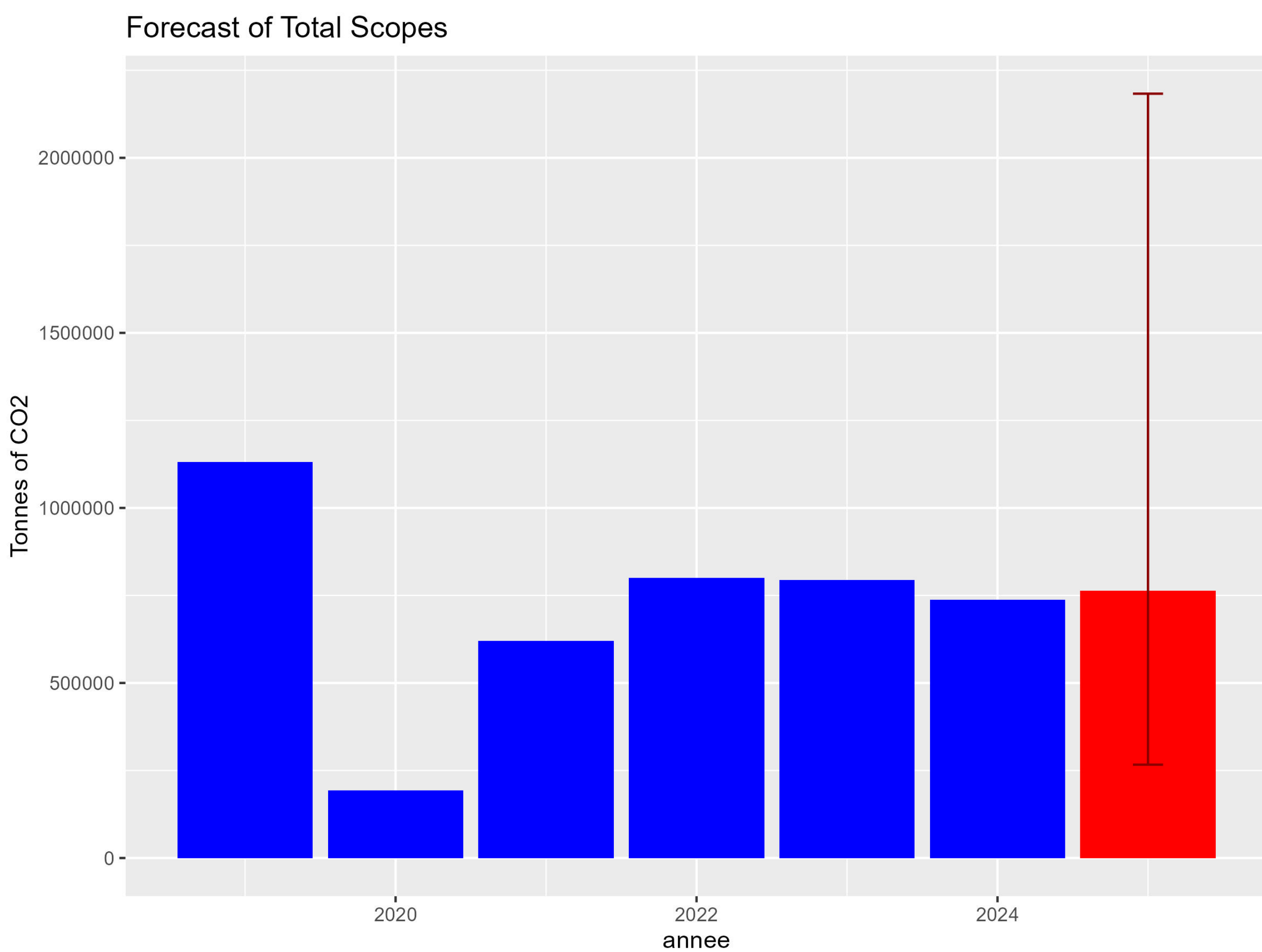

Figure 27: Forecast of total Scopes over time in tonnes of CO_2.

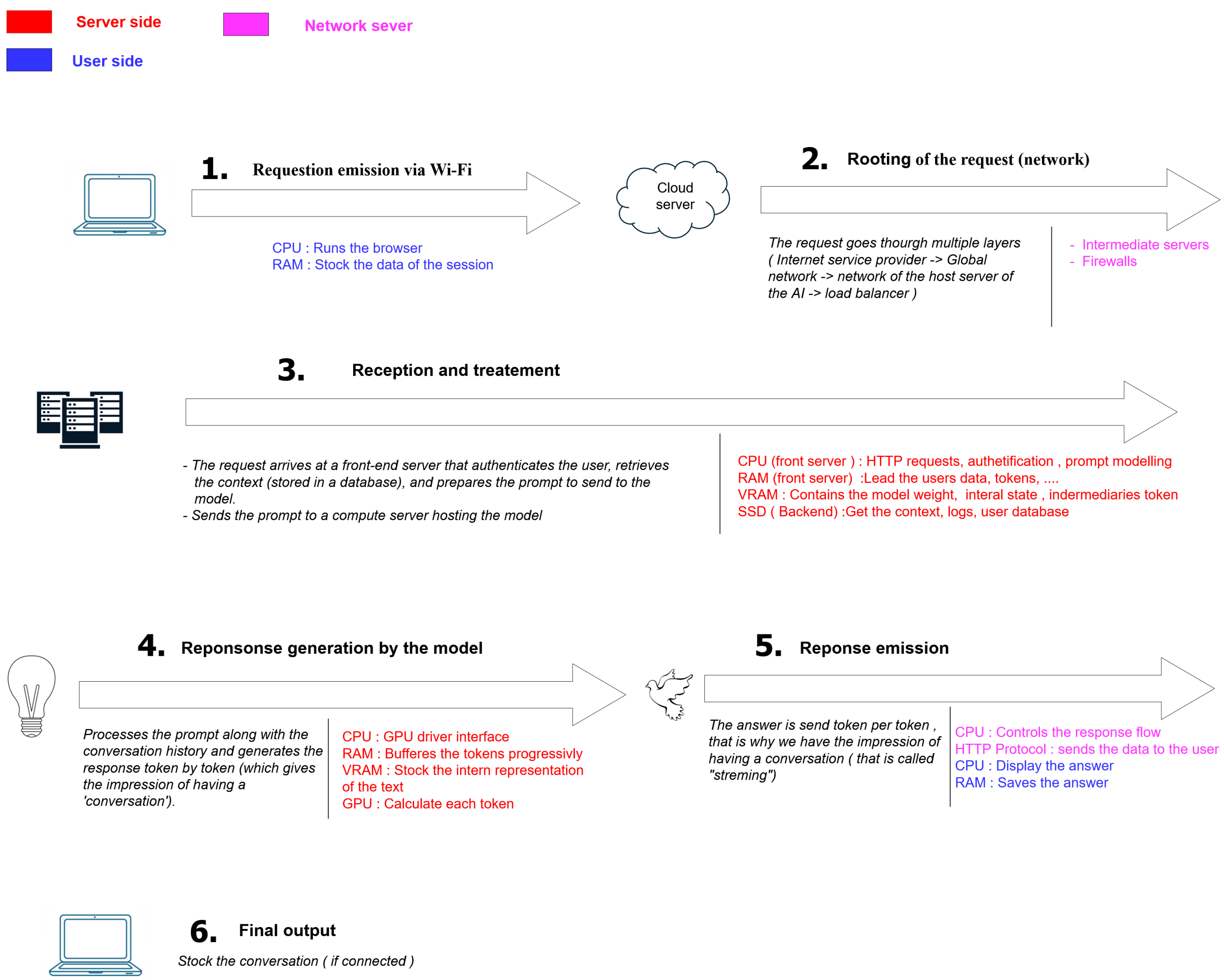

Figure 28: Overview of the functioning of an AI.

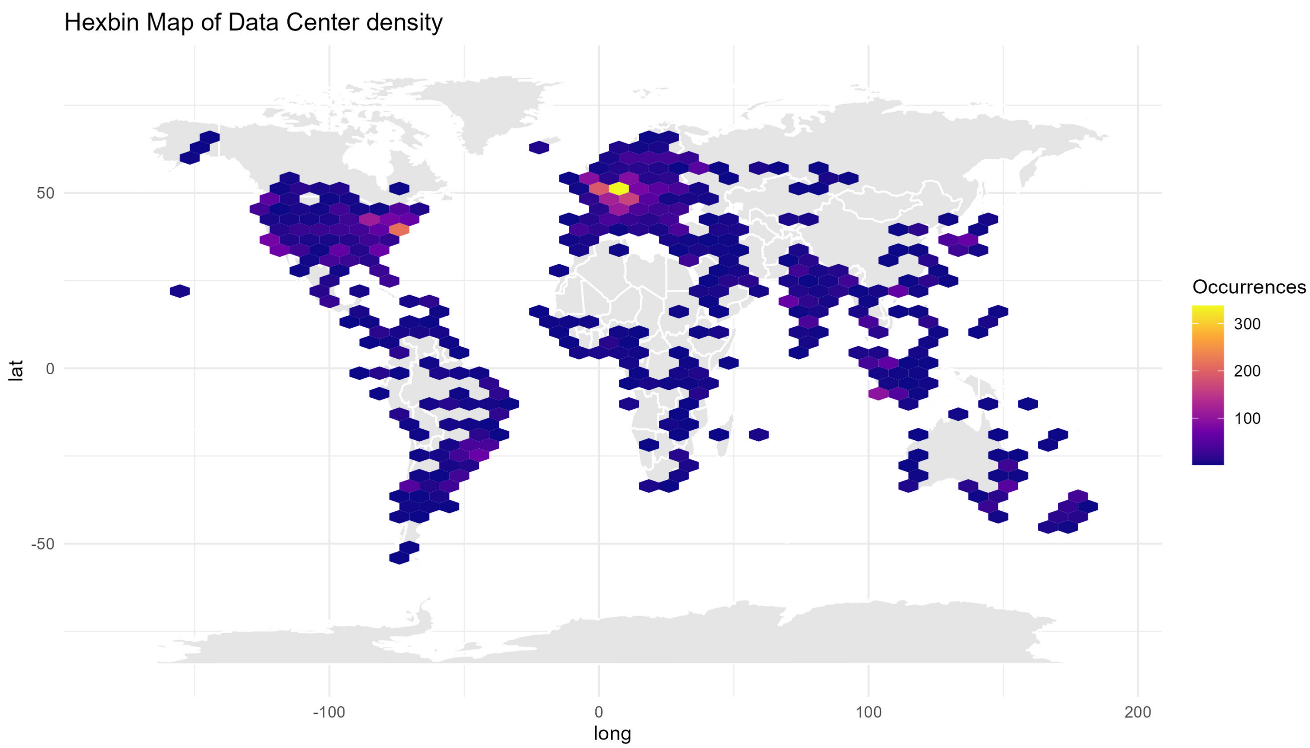

Figure 29: Hex-bin map of data centers

	country	total_occurrence
1	United States	1505
2	Brasil	561
3	Deutschland	459

Figure 30: Top 3 countries

term	estimate	std.error	statistic	p.value
(Intercept)	1.3549	0.0472	28.733	0
temperature	-0.0031	0.0019	-1.6518	0.0986
sm	-0.4022	0.0733	-5.4858	0
earthquakeTRUE	-0.9194	0.5777	-1.5915	0.1115
SPI	0.0743	0.0138	5.3955	0
prox_water_dist	-7e-04	2e-04	-3.8796	1e-04
prox_city_dist	-0.001	1e-04	-16.1805	0
McFadden R2	0.0277			

Figure 31: Model Distance

term	estimate	std.error	statistic	p.value
(Intercept)	0.7346	0.0445	16.5165	0
temperature	-0.0088	0.0018	-4.8336	0
sm	-0.3859	0.0721	-5.3511	0
earthquakeTRUE	-0.429	0.578	-0.7423	0.4579
SPI	0.0329	0.0136	2.4147	0.0157
prox_waterTRUE	0.1871	0.0238	7.8751	0
prox_cityTRUE	0.7623	0.0251	30.3628	0
McFadden R2	0.0668			

Figure 32: Model proximity

	Coefficient	Variable
(Intercept)	0.5234329	(Intercept)
sm	-0.1446398	sm
prox_water	0.1509892	prox_water
prox_city	0.6522457	prox_city

Figure 33: Coefficient regression

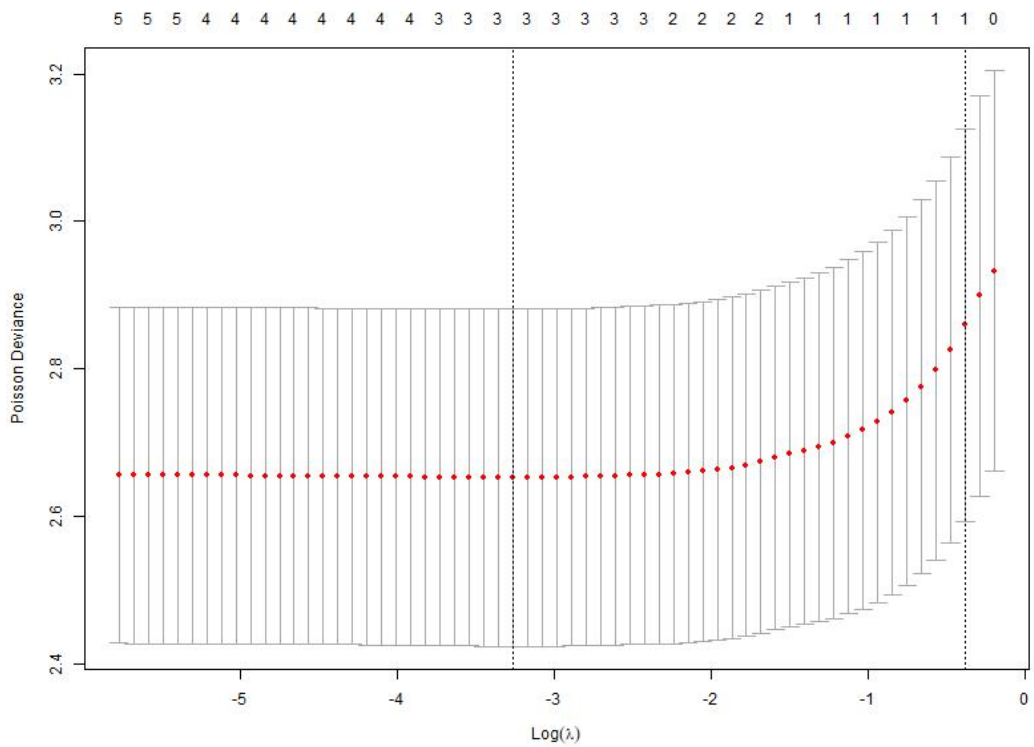

Figure 34: Lasso plot

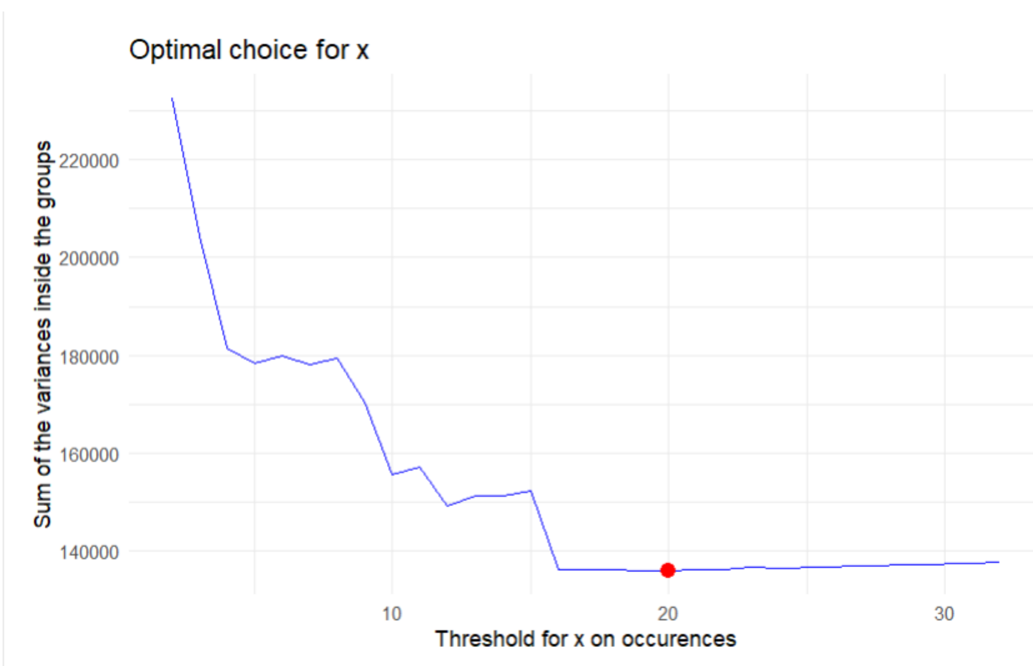

Figure 35: Optimal choice

term	estimate	std.error	statistic	p.value
(Intercept)	-7.1313	1.0842	-6.5774	0
sm	-0.9129	1.1359	-0.8037	0.4216
prox_cityTRUE	3.6453	1.0173	3.5832	3e-04
prox_waterTRUE	0.5942	0.3651	1.6278	0.1036
McFadden R2	0.1389			

Figure 36: Logit coefficients

References

[Amazon, 2025] Amazon (2025). Recommended gpu instances. https://docs.aws.amazon.com/dlami/latest/devguide/gpu.html.

[AWS, 2025] AWS (2025). Types d'instances amazon ec2. https://aws.amazon.com/fr/ec2/instance-types/.

[Canalys, 2025] Canalys (2025). Global pc shipments grew 3.9 https://www.canalys.com/newsroom/global-pc-shipments-q4-2024.

[Chika Saka, 2014] Chika Saka, T. O. (2014). Disclosure effects, carbon emissions and corporate value.

[Descartes, 1637] Descartes (1637). Discours de la méthode.

[Descartes, 1649] Descartes (1649). Les passions de l'âme.

[Github, 2020] Github (2020). Cooling. https://www.reddit.com/r/watercooling/comments/j6m7by/how_is_the_cooling_capacity_of_a_system_determined/.

[Github, 2023] Github (2023). Performance e-core. https://github.com/ggml-org/llama.cpp/issues/842.

[Google, 2025] Google (2025). Types de machines avec gpu. https://cloud.google.com/compute/docs/gpus?hl=fr.

[Instrument, 1997] Instrument, T. (1997). Cmos power consumption and cpd calculation. https://www.ti.com/lit/an/scaa035b/scaa035b.pdf?ts=1751145927920.

[Intel, 2025] Intel (2025). Processeur intel xeon platinum 8358p. https://www.intel.fr/content/www/fr/fr/products/sku/212308/intel-xeon-platinum-8358p-processor-48m-cache-2-60-ghz/specifications.html.

[Jianzu Wu, 2025] Jianzu Wu, Zhipiao Yu, R. W. (2025). The impact of public participatory environmental regulation on carbon emission intensity: a policy text analysis.

[Microsoft, 2025] Microsoft (2025). Virtual machine series. https://azure.microsoft.com/en-us/pricing/details/virtual-machines/series/?cdn=disable.

[Nidhal Jegham, 2025] Nidhal Jegham, Marwen Abdelatti, L. E. A. H. (2025). How hungry is ai? benchmarking energy, water, and carbon footprint of llm inference.

[Nikolov, 2025] Nikolov, S. (2025). Global ranking of best-selling laptop cpus on amazon – intel dominates, apple shines in premium, and amd holds steady. https://laptopmedia.com/highlights/q4-2024-global-ranking-of-best-selling-laptop-cpus-on-amazon-intel-dominates-apple-shin

[Notebookcheck, 2025] Notebookcheck (2025). Types d'instances amazon ec2. https://www.notebookcheck.net/Intel-Celeron-N4500-Processor-Benchmarks-and-Specs.514524.0.html.

[Nvidia, 2024] Nvidia (2024). Gpt results. https://docs.nvidia.com/nemo-framework/user-guide/24.07/llms/gpt/performance.html.

[OpenAI, 2024] OpenAI (2024). Water consumption. https://openai.com/index/carbon-emissions-metrics/.

[OVH, 2025] OVH (2025). Cloud gpu. https://www.ovhcloud.com/en/public-cloud/gpu/.

[Pujiang He, 2024] Pujiang He, Shan Zhou, C. L. W. H. W. Y. D. W. C. M. S. G. (2024). Distributed inference performance optimization for llms on cpus.

[Qi Fan, 2015] Qi Fan, Q. W. (2015). Performance comparison of web servers with different architectures: A case study using high concurrency workload.

[Ramos, 2025] Ramos, R. (2025). Measuring ai model performance: Tokens per second, model sizes, and inferencing tools. https://openmetal.io/resources/blog/ai-model-performance-tokens-per-second/.

[Samsung, 2025] Samsung (2025). Pcie 4.0 drives dynamic data traffics. https://semiconductor.samsung.com/ssd/datacenter-ssd/pm9a3/.

[TechnoLlama, 2025] TechnoLlama (2025). Global daily usage of generative ai: An assessment of scale and engagement (early 2025). https://www.technollama.co.uk/a-gemini-report-how-many-people-are-using-generative-ai-on-a-daily-basis-a-gemini-repor

[VMR, 2025] VMR (2025). Laptop memory (ram) market size, insights, dynamics forecast 2033. https://www.verifiedmarketreports.com/product/laptop-memory-ram-market-size-and-forecast/.

[WorldClim, 2020] WorldClim (2020). R package. https://worldclim.org/data/worldclim21.html.

[Zhuk, 2023] Zhuk, A. (2023). Artificial intelligence impact on the environment: Hidden ecological costs and ethical-legal issues.