2020/2021

INCORPORATING CLIMATE RISK INTO FINANCIAL ACTIVITY OF BANKS

SHERMAN ALINE
HAMZA OUAMMOU
CAMILLE PORTES
SUNNY WANG

ADAMANTIA ADVISOR :
DAVID ROUX

TSE ADVISORS :
ANNE VANHEMS AND ABDELAATI DAOUIA

ADAMANTIA

Toulouse School of Economics

Statistical Consulting Report *

Sherman Aline Hamza Ouammou Camille Portes Sunny Wang

April 1, 2021

Abstract

This report provides a detailed account of the statistical consulting project between the Toulouse School of Economics and Adamantia Consulting. While the overall theme is climate risk, the project is split into two parts due to the peculiarities of the data: the first establishes the regular credit risk model, and the second extends said model to incorporate climate risk.

1 Introduction

Despite the growing number of functions a modern bank undertakes, its primary role remains that of a financial intermediary. A core facet of this is lending money to businesses. Before doing so, a bank has to accurately assess whether the business can reliably repay their loans. This is done mainly through credit risk modelling, the goal of which is to predict a potential borrower's probability of default using relevant financial indicators.

With the widespread consensus by the scientific community on the drastic effects that climate change will have on our future personal and commercial lives, it is paramount to incorporate climate risk when assessing a borrower's ability to repay their loans in the future. Our project can be split into two distinct yet complementary parts. The first is to build a statistical model to predict the credit risk of companies in a data set given to us. Because the data set was presented to us in a manner where it was impossible to directly incorporate climate indicators, we had to deal with climate risk separately in a second part. We do this by building a basic application that allows non-technical staff at Adamantia to model climate risk *in the future*, when the data becomes available to them.

The rest of our report is outlined as follows: section 2 presents some motivation and empirical evidence on why it's important to treat climate risk seriously, while section 3 discusses some of the

*We are grateful to Anne Vanhems and Abdelaati Daouia for their guidance and support in this project, particularly for giving valuable technical advice.

current literature on climate risk models. In section 4, we provide a detailed description of the data set provided to us, its properties, and the difficulties we faced in dealing with it. Section 5 lays out the statistical model that we've used, while section 6 presents the results we've obtained after fitting our model. Finally, the issue of climate risk is dealt with in section 7, where we showcase and describe our client-friendly solution that allows the company to model credit risk when the data becomes available to them in the future.

2 Motivation

According to the Intergovernmental Panel on Climate Change (IPCC), the leading authority on assessing the impacts of climate change, human activities (from pre-industrial times till the present) are estimated to have caused an increase of approximately 1.0 °C. This global warming from anthropogenic emissions will persist for centuries, leading to drastic changes in our current climate system. Some examples include hot extreme temperatures in the most inhabited regions, heavy precipitation, droughts, and a rise in sea levels. In addition to affecting the marine biodiversity and our natural ecosystems, these ecological changes will affect our livelihoods, food security, water supply, and economic growth. For example, crops of agricultural firms might get wiped out with the increased prevalence of natural disasters such as floods and droughts, petroleum companies will likely be subject to heavy carbon taxes as governments seek to reduce emissions with policies, and producers might experience a change in their consumer demand due to shifting consumer preferences (for example, customers might prefer to buy environmentally friendly products) or a shift in production prices (for instance when non-green options become more expensive).

Since these physical and transition risks can have dramatic impacts on a company's financial health, it is important to take them into account when performing credit risk modelling. This is all the more pertinent when one takes a long-term view, since global warming is likely to accelerate in the future. With the current rate of emissions, it is estimated that global warming is likely to reach 1.5 °C between 2030 and 2052 [1], exacerbating the consequences mentioned earlier.

Unfortunately, climate risk modelling remains a novel (though burgeoning) approach to credit risk, and most banks treat it as part their corporate social responsibility campaigns instead of making it a standard practice. Because of this, good data for climate indicators remains difficult to come by, with complete data sets of lending history with company-level climate indicators proving even rarer.

Because Adamantia did not yet have a basic credit risk model for their data, we first dealt with this task by estimating a model using a variety of statistical methods (after the appropriate data transformations and cleaning). Thereafter, we decided to deal with climate risk separately by

[1] According to IPCC's special report on the impacts of global warming of 1.5 above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty

building a web application which allows the user to upload climate data when it becomes available instead of attempting to incorporate it into our data set which does not contain enough information for us to do so.

3 Literature Review

Before we began our project, we performed a literature review on climate risk, in order to get some idea on the models currently used by institutions. However, due to the structure of the data, we are unable to directly apply these models into our current context. Thus, in this section, we mostly focus on why the current models in the literature *can't be applied* to our project. By doing so, it also highlights the peculiarities of our data set and the consulting project.

We looked at research on the link between environmental variables and financial risks. In Berenguer, Cardona, and Evain (2020), the authors outline a comprehensive list of environmental risks and discuss possible solutions. There are three types of financial risk: physical risk, transition risk, and liabilities.

- Physical risks result from natural disasters, which harm physical assets and decrease cash flows by temporarily lowering productivity. Some examples include property owners in coastal areas suffering damages to their buildings from floods and agricultural firms experiencing crop losses due to droughts.

- Transition risks concern the cost of transitioning from current energy sources to a low-carbon model. Transition risks primarily affects asset values, especially fossil fuels. An example would be the increased cost to manufacturers as they transition to greener raw materials due to government mandate.

- Liabilities are the costs related to government regulations penalising industries which are not green. There is currently a lack of research on the impact of these costs.

Governments employ two approaches to deal with financial risks from the environment. The first integrates climate risks into capital requirements for banks, which will in turn incentivize banks to consider environmental risks in their models. This approach will suppress brown assets, which are high carbon assets that contribute to climate change instead of mitigating it. The second approach involves Economic Policy, where the principal goal is to move credit in a way which softens the transition to a low-carbon economy through incentivizing green investments. This is more likely to support green assets.

For our goal of developing a risk model, the risk approach is more relevant. There are some challenges emphasized in Berenguer, Cardona, and Evain (2020) when it comes to modelling risk from environmental data. There is a lack of historical data to develop risk measures, as measurements of environmental data, especially for individual companies, have been incorporated only

recently. Additionally, standard risk models rely on a short term horizon, but this conflicts with the long term effects of climate impact.

In 2014, the EU introduced the Small and Medium Entrprise (SME) and Infrastructure Supporting factor. The goal was to increase credits given to these two sectors, as increased regulation had decreased loans. However, there is not enough evidence to show that it is effective. Many other countries have also developed their own models incorporating climate risk into financial regulation. In the risk approach there are several perspectives.

- Green Supporting Factor (GSF)
- Brown Penalising Factor (BPF)
- Environmental Risk Weighted Asset
- Green Weighting Factor

Another useful example is the Transition Score developed by Credit Agricole. Their model relies on

- Energy Transition Score provided by Vigeo (no greater details available)
- Intended Nationally Determined Contribution for the asset's sector and geographic location, normalised for the specific year

Many of the models utilised in the industry are trade secrets, and knowledge of what covariates are used is difficult to come by.

In Allen et al. (2020), they model climate risk by performing stress-testing using a macroeconomic framework. The key idea is to account for the macro-financial impacts of adverse shocks and study their spillover effects across markets and countries. Since this is a complex exercise, the authors combine a suite of models instead of developing a single tool that can encompass everything.

The modelling architecture is a sequential one - they first use standard models (e.g Integrated Assessment models that the IPCC uses to quantify mitigation scenarios) to predict key outputs such as carbon prices and greenhouse emissions, before feeding them as inputs into a country leve macroeconomic model (e.g National Institute Global Econometric Model used by policy makers used for economic forecasting). The predicted macroeconomic outputs such as inflation and unemployment are then fed into an in-house sectoral model, which subsequently provides indicators such as turnover for 55 different sectors. Finally, these sector level outputs are then fed into a rating into a rating model to predict probability of defaults. While this top-down approach seems like an attractive one in modelling climate risk where only high-level variables are known, the final probability of defaults (response variable) require firm level information such as CO2 emissions,

which we do not possess in our current data set. Thus, if we were to adopt this approach, we would be forced to assign the climate variable on an industry level. This will not add anything useful to our regression model, since everything will just be scaled by a constant.

4 Data

4.1 Description of Data

A financial institution provided our data set, which comes in the the form of a data frame containing 32443 rows and 20 variables which are described below. The two response variables are *Financial rating* and *Qualitative Rating*. The first three variables are descriptive variables for each company. The remaining variables are explanatory variables which are used to compute the two ratings.

A key issue we have faced with this data was to determine its underlying structure. Many rows are similar, which is very suspicious. Indeed, the probability that several companies posses exactly the same amount of *Assets*, *Liability* or *Turnover* is in our opinion 0. This discovery led us to think more about the type of data we're dealing with. Since we couldn't ask for confirmation, we first assumed that the rows represented a unique company (which is unlikely), and then treated the data as if each entry corresponds to a loan request, which can cause data duplication. In the latter case, we removed the exact duplicates and we got 19928 rows.

- `ID` - Unique user ID
- `Status` - Legal statutes of the company
- `Sector of Activity` - INSEE activity code
- `Financial Rating` - Grade chosen by the bank, this grade is more important than Qualitative Rating
- `Qualitative Rating` - Grade checked by the bank in case of doubts after looking at the Financial Rating
- `Qualitative rating about transparency` - Grade given by the bank based on documentation provided by the company during the loan request
- `Qualitative rating about shareholders contribution` - Grade given on the position of shareholders toward the company. If the grade is above 10, the shareholders are able to support the company, if below 10, shareholders can withdraw dividends even if the company is in a bad position
- `Favorable economic market` - Grade given by the bank on the position of the company on the market

- `Sector will increase` - Grade on the prospects for the future of the sector
- `Management quality` - Grade on the experience of the manager
- `Hold by a bigger company` - 1 if the company is part of a bigger firm, 0 otherwise
- `CEO involved` - Consistency between the president (highest responsible of shareholders) and the director (management part)
- `Help from the group on legal` - Legal help given by the group that hold the company
- `Assets` - What the company possess
- `Liability` - What the company owe
- `Turnover` - Total revenues earned from clients
- `EBITDA` - Earnings before interest, taxes, depreciation of its value, and amortisation
- `Debt on equity` - Amount that the company owes to the bank
- `Gross operating surplus/global costs` - Ratio between the surplus and the costs
- `(Gross operating surplus / Turn over) * 100` - Ratio between the surplus and the Turn over

4.2 Exploratory Analysis

In this subsection, we will provide an exploratory analysis of our data. We focus on financial rating as the response variable, since the other response variable (qualitative rating) is a categorical variable, which isn't as interesting to look at in terms of descriptive statistics. Finally, we will discuss the data transformations we've performed on each covariate to increase the performance of our model. With regards to the data transformations, we focus our attention only on a few specific groups instead of all of them.

4.2.1 Global Covariate Analysis

Since the structure of the covariates are quite similar, we will focus our analysis on the common themes, leaving most of the graphs to the Appendix to avoid repetition.

We use Figure 1, which represents a scatter plot of Turnover against Financial Rating to elucidate the problematic structure of our data.

We can see that the values of Turnover, our co-variate are extremely skewed with an overwhelming concentration around zero. This over-representation of zero values is a first indication that the Turnover possibly comes from a zero-inflated model, a mixture distribution with a part

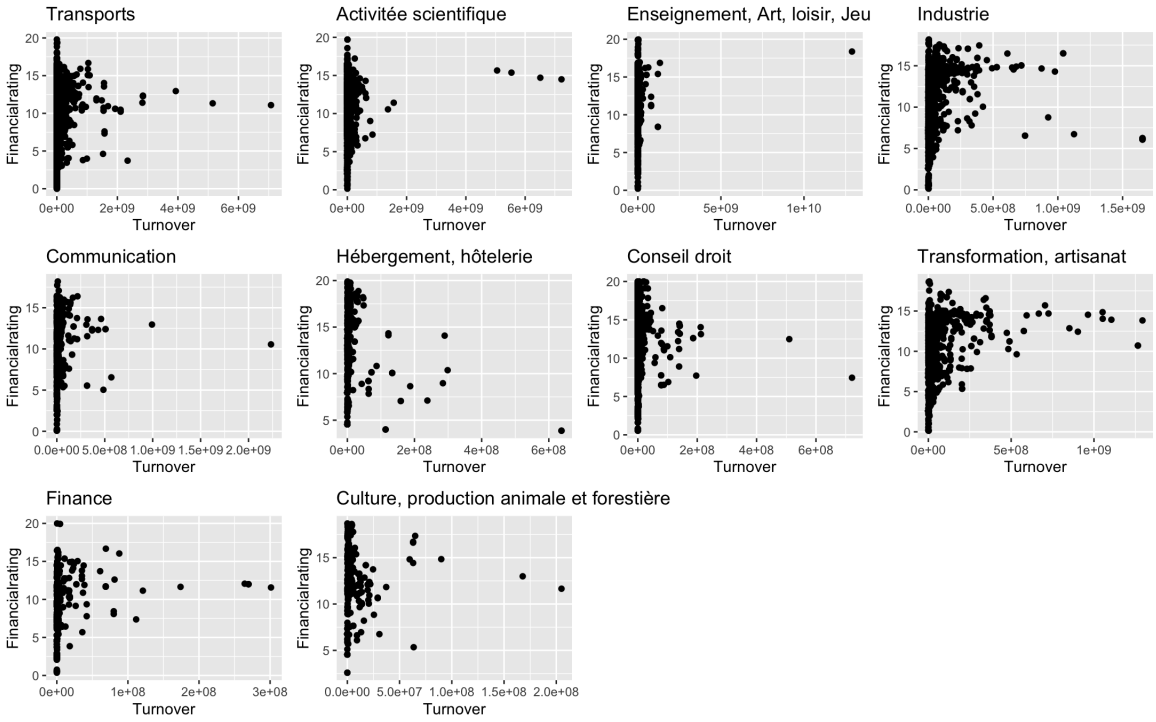

Figure 1: Scatter Plots for Turnover on Financial Rating for all groups

containing a point mass around zero along with another part for positive values. Since the zero-inflation is in the covariate and not the response, we will not use a zero-inflated regression model, but rather specify a point mass around zero using a dummy variable (we postpone this discussion to the following sections, since we are focused on exploratory analysis here).

We can also see extreme values of Turnover present across all groups, with values reaching the hundreds of millions. While we are not surprised that Turnover for some companies can be very large, particularly if they are big conglomerates, we did not expect to see this amount of concentration around the two extremes, with a lot of sparsity present when it comes to mid-levels of Turnover. In our opinion, this calls for a deeper probe into how the data was constructed, since it's much more logical to expect that we have more small and medium sized enterprises (SMEs) - according to the OECD, they account for 99.9% of all enterprises in France. We have two hypotheses about what is really going on with the data - the first is that these SMEs are actually being masked as zero values, perhaps because of reporting issues (such as not submitting a tax declaration on time), or the fact that the bank didn't collect this information (which seems highly unlikely if you were to consider them for a loan). The other is that the bank has a very unique base of clientele that consists only of businesses with extreme Turnover values. Similarly, we find this quite unlikely, since if this were the case, we should expect to see a much bigger concentration of positive values around the upper end of the distribution (i.e the bank is only focused on lending money to very large corporations with high Turnover), instead of companies with zero turnover. It is far too risky to even consider lending money to these zero turnover businesses, and even for a

bank with a high risk appetite, they should only constitute a small proportion of potential clients.

From a quick glance at the scatter plots, we can see that the groups with the best looking "distributions" are `Industrie` and `Transformation, artisanat`. For `Industrie`, this is a common theme - it appears as the best looking group across all covariates, and thus it is one of the groups we focus on when performing our analysis for Financial Rating. The other group we chose to focus on is `Conseil Droit` - not because it has the best scatter plots, but because the client wanted us to look at another group that weren't going to be as dramatically affected by climate change, since `Industrie` will constitute one of them.

Although we focused our analysis only by referring to the scatter plots of Turnover on Financial Rating, the common problems of over-concentration around extreme values that we see are universal across all groups, and we refer the reader to the Appendix should they be interested in the other plots.

4.2.2 Data Transformations

In order to address the significant irregularities of our data, we performed various data transformations in order to prepare them for modelling. We focus only on two groups - `Industrie` and `Conseil droit`. As briefly discussed earlier, the basis for choosing `Industrie` is because the plots look the best out of all the covariates. The latter was chosen since we were told to study another group that will not be negatively affected by climate change in a significant way. For the sake of brevity, we only discuss the transformations for `Industrie` here to avoid repetition - the transformations performed for `Conseil Droit` are almost identical, and more details can be found in the appendix. Proceeding similarly to the previous subsection, we do not expound on every single coviarate, since most of the transformations performed are almost identical. We will focus our energy only on two different sets of covariates where the transformation was done differently across groups.

The justification for the transformations are rather qualitative in nature - we simply looked at the scatter plots, decided what the best course of action was, and experimented with the results until we achieved a satisfactory one.

4.2.2.1 Turnover

We first removed the extreme values for Turnover. Specifically, we removed all values above the 95th and below the 5th quantiles to make the data less skewed, but also to get a more fine-grained view of the non-extreme values.

A log transformation was then performed on all strictly positive values to re-scale the data. After doing so, we obtained the scatter plot in Figure 2.

We can clearly see two different "clusters" - a cloud of points showing a somewhat positive relationship for log values centred around 15, and another cluster with a point mass around zero.

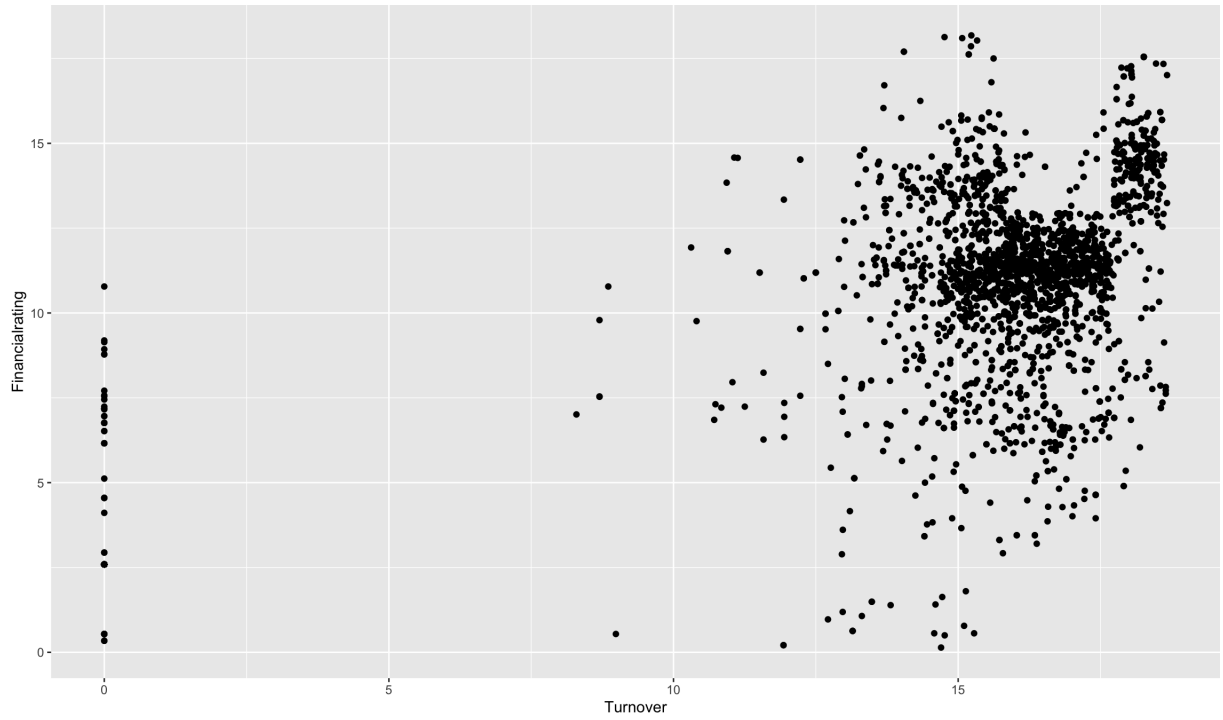

Figure 2: Transformed Scatter Plot - Turnover on Financial Rating

Clearly, if we want to capture this bimodal relationship, we'll need to treat these two cloud of points differently in our regression. To complete our data preparation for our regression later, we created a new binary variable, which takes the value 1 if the Turnover is zero.

4.2.2.2 All Other Covariates

For all other co-variates, the process was very similar to the one for Turnover - we first removed the extreme values above the 95th and below the 5th quantiles, created new binary variables for each covariate if they take the value zero. The main difference between these other set of co-variates and Turnover is that we did not perform a log transformation. The reason is simply that after removing the extremes, the values already looked a lot better and well spread out, and we did not see a strong justification for performing a log transformation here except for interpretation purposes. Since we're not entirely sure about what the data really represents and are thus hesitant about giving any interpretation at all, we chose to omit the log transformation.

Here, we use the transformed scatter plot for EBITDA on Financial Rating as an illustration, which is depicted in Figure 3.

We can again see two "clusters" of points, one containing strictly positive values and another with a point mass around zero. To treat these zero-inflated values similarly, we created a new binary variable as mentioned above.

Since the results are rather similar, we leave the rest of the transformed scatter plots to the appendix. What we do want to point out is that although all these variables had zero-inflated

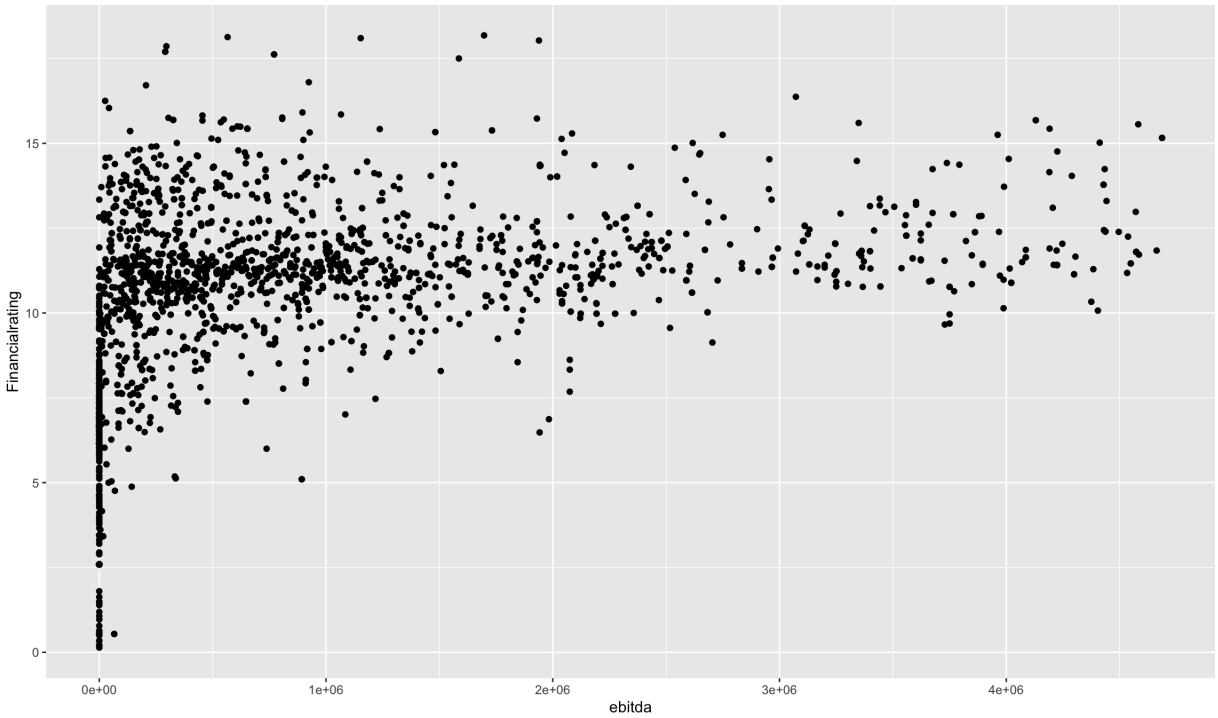

Figure 3: Scatter Plot for EBITDA on Financial Rating

values, they showed different types of relationships against our response (either positive or negative). However, just from the cloud of points alone, we can see that most of the relationships of the covariates with the response for the strictly positive values are rather weak.

5 Statistical Methodology

We fitted two different models for each of our variables response - for qualitative rating, we used a simple linear model which worked surprisingly well, while for financial rating, we used a generalised additive model. We first turn our attention on the financial rating model, since it is less familiar to most.

5.1 Financial Rating

The generalised additive model builds on the generalised linear model, which has two fundamental assumptions - the first is a distributional assumption that the response variable has a probability density which belongs to the exponential family, while the other is a structural assumption that the conditional mean is related to the covariates by some link function.

For completeness sake, we write out our distributional assumption explicitly as

$$f(y|\theta,\phi) = \exp\{w\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\} \tag{1}$$

where w are weights, θ is the location parameter, ϕ is the dispersion parameter, $b(\theta)$ is a known function such that $b^{'}(\theta) = \mathbb{E}(Y)$ and $c(y, \phi)$ is a known function serving as a constant.

For the generalised additive model, the distributional assumption remains the same, but the structural one is different, taking the form

$$g(\mathbb{E}(Y|X_1, ..., X_m)) = \beta_0 + \sum_{i=1}^{m} f_i(X_i) \tag{2}$$

where the functions f_i are smooth functions to be estimated non-parametrically, and the various x_i's are our covariates for financial rating. We used the built-in R function `gam` in the `mgcv` package which estimates each of these functions using Thin-Plate splines. Using the specification above, the conditional mean for our response will thus be given by

$$\mathbb{E}(Y|X_1, ..., X_m) = g^{-1}(\beta_0 + \sum_{i=1}^{m} f_i(X_i)) \tag{3}$$

where g^{-1} denotes the inverse image of our link function g.

The justification for using the GAM is firstly that is that it encompasses a rather broad class of models. Since we do not have prior knowledge of the distribution or the relationship between our variables, we thought that it is important to let the data speak for itself. We did not resort to "pure" non-parametric regression methods, since our data set isn't very large (especially after filtering down to specific groups), and we were afraid of sparsity issues which can make these non-parametric methods inaccurate.

5.2 Qualitative Rating

While a linear model seems overly simplistic and too strong an assumption, we chose it for qualitative rating simply because it's the first model we tried and it worked *surprisingly well*. We postpone the details on the performance to the next section - here we simply give a quick reminder on the generic form of the regression equation. We are fitting the following model (in matrix form):

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon \tag{4}$$

The conditional mean, which we are interested in predicting, will thus be given by $\mathbb{E}(\mathbf{Y}|\mathbf{X}) = \mathbf{X}\beta$. The specific covariates are provided in the next section.

6 Analysis of Results

6.1 Metrics of Interest

In the process of building our qualitative model, we separated our data into train and test sets of 80% and 20% respectively. By using this cross-validation setup, we can evaluate the fit of the model while also checking for over-fitting at the same time.

When building the Financial models we rely on the score of Generalised Cross-Validation (GCV) and R-squared to evaluate the in-sample fit of our model, and they are given by the following formulas

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-m-1} \tag{5}$$

where m is the number of covariates, and R^2 is given by

$$R^2 = 1 - \frac{SSR}{SST} \tag{6}$$

where SSR and SST represents sum of squared residuals and sum of squared totals respectively, and

$$GCV(\lambda) = \frac{n \times SSR(\lambda)}{(n - tr(H))^2} \tag{7}$$

where $tr(H)$ is the trace of our hat matrix.

The algorithm which fits the generalised additive models attempts to minimise the GCV score, with the ideal GCV score being zero and the ideal adjusted R-squared being one. Finally, we visually inspect the graphs of each function in our model to check for any overfitting.

6.2 Qualitative Model

After some data exploration and analysis, we deduced that the best model is to fit a linear model. Qualitative rating is a quantitative variable (i.e numeric / continuous), while all its covariates are qualitative variables (i.e factors / discrete).

We'll fit the linear model using the following covariates: "SC", "FEM" , "SWI", "MQ", "HBF", "CEO", "LEG".

$$QR_i = \beta_0 + \beta_1 SC_i + \beta_2 FEM_i + \beta_3 SWI_i + \beta_4 MQ_i + \beta_5 HBF_i + \beta_6 CEO_i + \beta_7 LEG_i + \epsilon_i$$

Where

- `SC` - Qualitative rating about shareholder's contribution.
- `FEM` - Favourable Economic Market.

MSPE	RMSPE
0.1131564	0.3363873

Table 1: Prediction errors

- `SWI` - Sector will increase.
- `MQ` - Management Quality.
- `HBF` - Hold by a bigger firm.
- `CEO` - CEO Involved.
- `LEG` - Help from the group on legal.

We tried to perform some predictions using our qualitative rating model to check it's performance. We randomly sampled 80 % of our data to be the training set and 20 % to be the test set.

It was difficult to see how the prediction performs from the plots since we have too many observations. Instead, we looked at the mean-squared prediction error, given by:

$$MSPE = \mathbb{E}[(Y_i - \hat{Y}_i)^2]$$

where Y_i are the observed values of our response variable, while $\hat{Y}_i$ represents the predicted values given our covariates. We got the scores given in Table 1.

We focus on the mean-prediction error, which is $\sqrt{MSPE}$, since they are easier to interpret. Our interpretation is that on average, our prediction is "off" by about 0.34, which seems like a reasonable number since qualitative rating ranges from 0 to 20.

6.3 Financial Model

We developed two models based on different groups: `Industrie` and `Conseil droit`. Both these models are fitted using on general additive models, as described in section 5.

6.3.1 Industrie Model

Coefficients:

	edf	Ref.df	F	p-value
s(Turnover)	6.63	7.51	34.92	0.00
s(ebitda)	6.50	7.64	3.45	0.00
s(Debtonequity)	4.87	5.92	53.00	0.00
s(grossoperatingsurplusglobalcosts)	8.38	8.89	2.35	0.01
s(grossoperatingsurplusTurover100)	6.22	7.40	10.20	0.00

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	11.58	0.93	12.44	0.00
zero_tover	-8.80	14.93	-0.59	0.56
zero_ebitda	-1.87	0.24	-7.87	0.00
zero_doe	-3.33	0.17	-19.82	0.00
zero_gos	0.09	0.27	0.33	0.75
zero_gos_100	1.65	1.07	1.55	0.12

R-sq.(adj) = 0.762	Deviance explained = 76.6%	
GCV = 2.1323	Scale est. = 2.0898	n = 1939

Table 2: Regression Output for GAM - Industrie

For this model we have a GCV score of 2.13 and an adjusted R^2 of 0.76.

6.3.2 Conseil Droit Model

	edf	Ref.df	F	p-value
s(Turnover)	7.46	8.09	4.70	0.00
s(ebitda)	7.86	8.63	3.03	0.00
s(Debtonequity)	8.50	8.93	19.30	0.00
s(grossoperatingsurplusglobalcosts)	8.07	8.74	6.19	0.00
s(grossoperatingsurplusTurover100)	6.42	7.56	6.85	0.00

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	21.23	27.03	0.79	0.43
zero.Turnover	-24.49	54.54	-0.45	0.65
zero.ebitda	-2.78	0.32	-8.67	0.00
zero.Debtonequity	0.53	0.24	2.20	0.03
zero.grossoperatingsurplusglobalcosts	1.10	0.26	4.24	0.00
zero.grossoperatingsurplusTurover100	4.64	1.20	3.85	0.00

R-sq.(adj) = 0.725		Deviance explained = 73.1%
GCV = 6.543	Scale est. = 6.3837	n = 1820

Table 3: Regression Output for GAM - Conseil Droit

For this model we have a similar adjusted R^2 but a much higher GCV score. However, cross-validation scores are relative and cannot be compared between different models.

7 Extensions - Part II

The previous parts deal with creating a suitable model to fit the data and perform predictions, in an attempt to uncover what the data generating processes of our main response variables are, together with their reduced form. While they are useful in loan assessments in the short run, they do not yet take into account climate variables. Since one of the main goals of the project is to incorporate climate risk, we had to look for ways to extend our previous models to do so.

Unfortunately, as mentioned in earlier sections, the provided datasets pay no mind to climate variables, since the collection and incorporation of climate data for risk assessment in banks for risk assessment is still at a very early stage in the banking sector. In order to tackle this problem, we decided to build a web application to enable the client to incorporate climate risk should these data sets become available to them in the future.

The purpose and functionality of the proposed web application is straightforward - when armed with a new data set containing the relevant climate variables, they will be able to upload it into our web application, choose some variables to perform exploratory analysis, and fit the model constructed in earlier sections to perform predictions based on their new data set. Since we've performed all our statistical consulting tasks in R and its purpose is specific enough, we decided to go for a Shiny application, which allows us to build interactive web applications using only R, reducing the need to use an entire stack of development tools that is more typical of full-blown web applications with a wide range of functionality.

This section will be more of a tutorial and tour of the web application rather than a detailed analysis, since that is already done in the preceding sections, and we are using the exact same models and tools. We will separate our exposition into the different tabs of the web application for simplicity. Before we go into each tab separately, it's nice to take a look at what it looks like at the outset as shown in Figure 4 below.

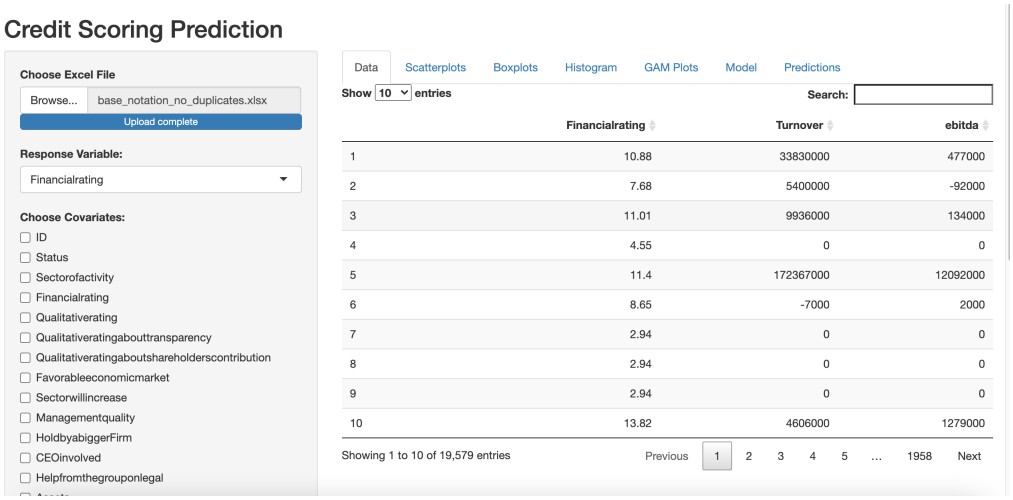

Figure 4: Snapshot of the Web Application

7.1 Scatterplot Tab

The Scatterplot tab allows the user to see scatter plots between the covariates selected and the response variable. These types of plots are useful to have a representation of the relationships between the independent variables and the depeneant variables. It can help the user to visualise the outliers and ascertain if the values are concentrated around some values. Thus the user can decide how to treat the data separately before applying a regression model. These scatter plots are interesting for quantitative variables.

7.2 Boxplots Tab

The Boxplots tab is another tool to visualise the quantitative data. It gives additional information such as the 1st quantile, the mean and the 3rd quantile. As the different boxplots are represented on the same picture, there could be some scale issues. We put a button which allows the user to logscale the data to obtain a better representation.

7.3 Histogram Tab

The histogram tab shows a histogram for the dependant variables. It is mainly there to be used with the variable Financial Rating as it is the response variable in the GAM model, which adds special significance to its distribution. By looking at the distribution on this histogram, the user is able to choose between the different exponential families to apply the GAM model. The histogram for the Qualitative Rating can be used to determine if the linear model we've used is suitable for the new data set. If that is not the case, the company will have to separately fit their own model.

7.4 Model Tab

The model tab outputs the results of the model fitted on the uploaded data based on the response variable and co-variate that has been selected by the user. What is being printed is basically identical to what a user will see when they use the "summary" function on a model in R. We can see that the model is outputting the coefficients from the regression (depending on whether the user has selected Financial Rating or Qualitative Rating as the response, it is either fitting the Generalised Additive or Ordinary Linear model). For the GAM, the user can choose the exponential family the response variable belongs to. The outputs from the other tabs which performs exploratory analysis will be helpful for the user to judge which exponential family is most suitable. We only included families that were shown to be of relevance in the literature, as the exponential family is too wide to exhaustively cover.

7.5 GAM Plots Tab

The GAM Plots tab displays one plot for each covariate. Each graph shows the marginal plots of our multivariate regression function, and the sum of all these graphs would correspond to the full model. Along with the GAM, a rug plot is displayed to give an indication of the placement of individual data points.

7.6 Prediction tab

This tab displays the first 20 predicted values based on the model (described above) fitted on the selected variables. For the full set of predictions, the download button can be used to output a csv file for the user.

The predictions are performed by cross-validation - the data uploaded by the user is first split into training and test sets (80% train, 20% test) of the observations, where the model is estimated based on the training values and predicted based on the covariate values available in the test set.

Finally, the RMSPE is displayed at the end, which is calculated based on the formula provided in section 6.2.

8 Conclusion

In this statistical consulting project, our main goal was twofold: first to find a suitable model to perform credit risk scoring based on covariates provided to us in a data set by Adamantia, and then attempt to extend the model we've built by incorporating climate risk.

In order to carry out the first task, we did a rigorous analysis of the data set using advanced statistical methodologies and performed various transformations to improve the performance of said methodologies. We used the ordinary linear model for the qualitative rating model and a generalised additive model for the financial rating model, and evaluated the suitability of these models using various metrics. We managed to achieve an adjusted R-squared of around 76%, and a root mean squared prediction error of around 0.34 (out of a scale of 20).

With regards to the incorporation of climate risk, we did not have access to any relevant data, so we built a Shiny application to allow the client to put our model to good use once they have access to climate variables in the future. Our hope is that the application will be a solid foundation for the company to build on in the future, where they can perhaps extend it to a wider range of models that encompasses most cases typically studied in climate risk models.

To conclude, we thought it'll be fitting to briefly discuss what we've taken away from an academic year of working on this project:

- We experienced first hand how to deal with "real-world" data, which is often messy, unorganised and requires a lot of data cleaning and processing before being fed into any statistical

model.

- We learned how to come up with creative solutions when the current situation didn't allow us to achieve the initial goals.
- We learned how to deploy a functional web application on Shiny and the many tools used in development, such as Github.
- We sharpened our technical skills and learned to apply and choose between the many different models we have studied in class.

9 Appendix

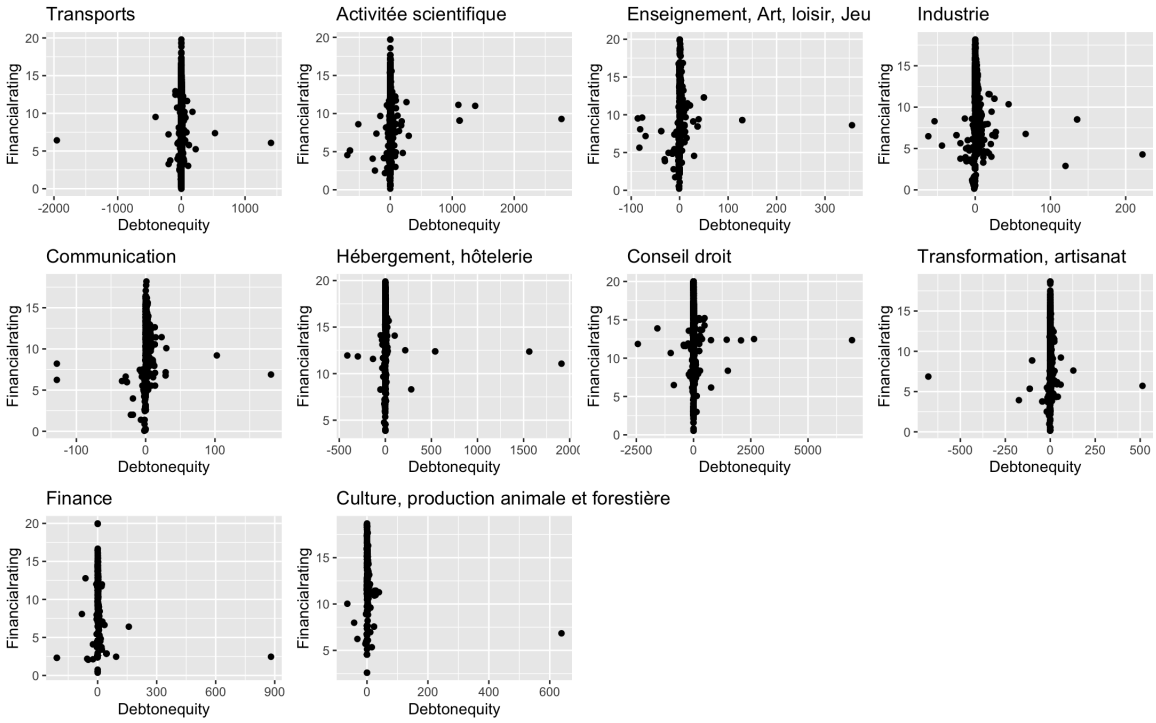

Figure 5: Scatter Plot - Debt on Equity on Financial Rating

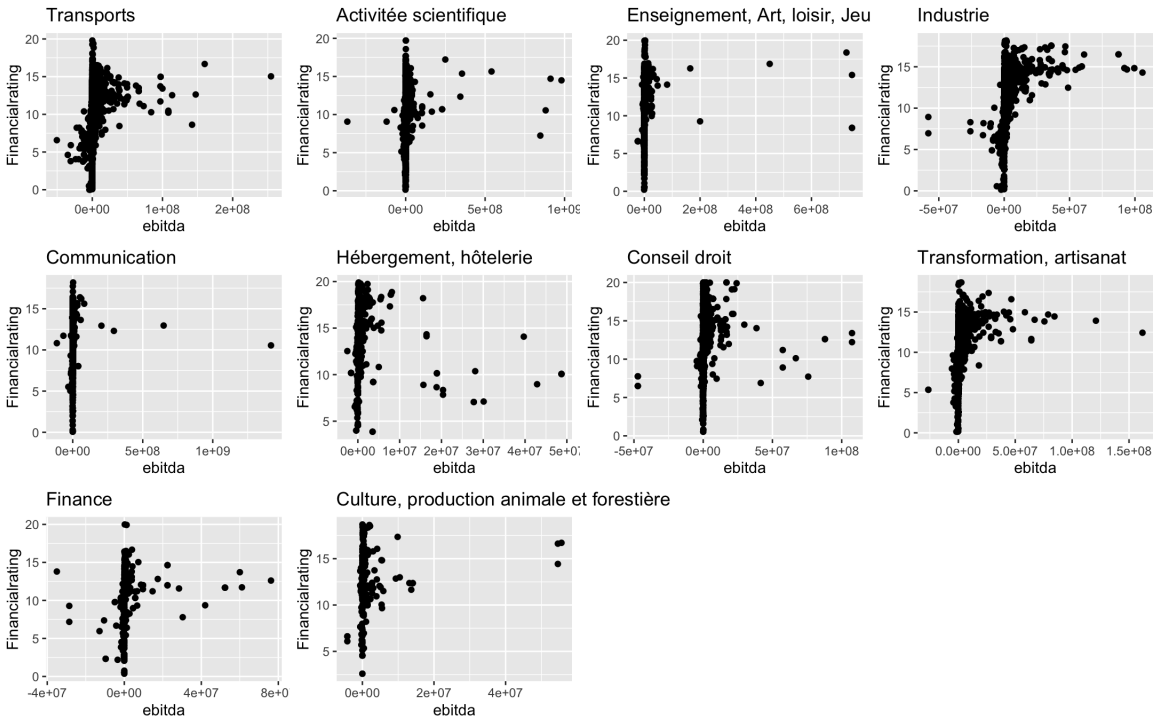

Figure 6: Scatter Plot - EBITDA on Financial Rating

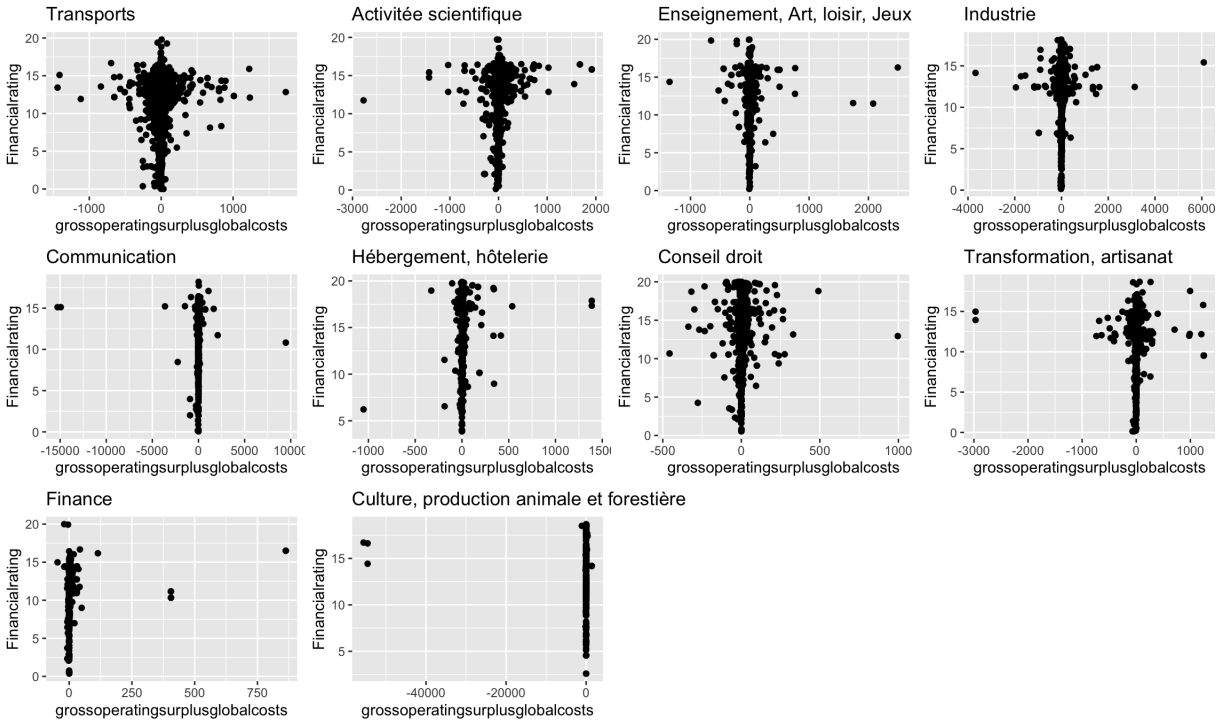

Figure 7: Scatter Plot - Gross Operating Surplus on Financial Rating

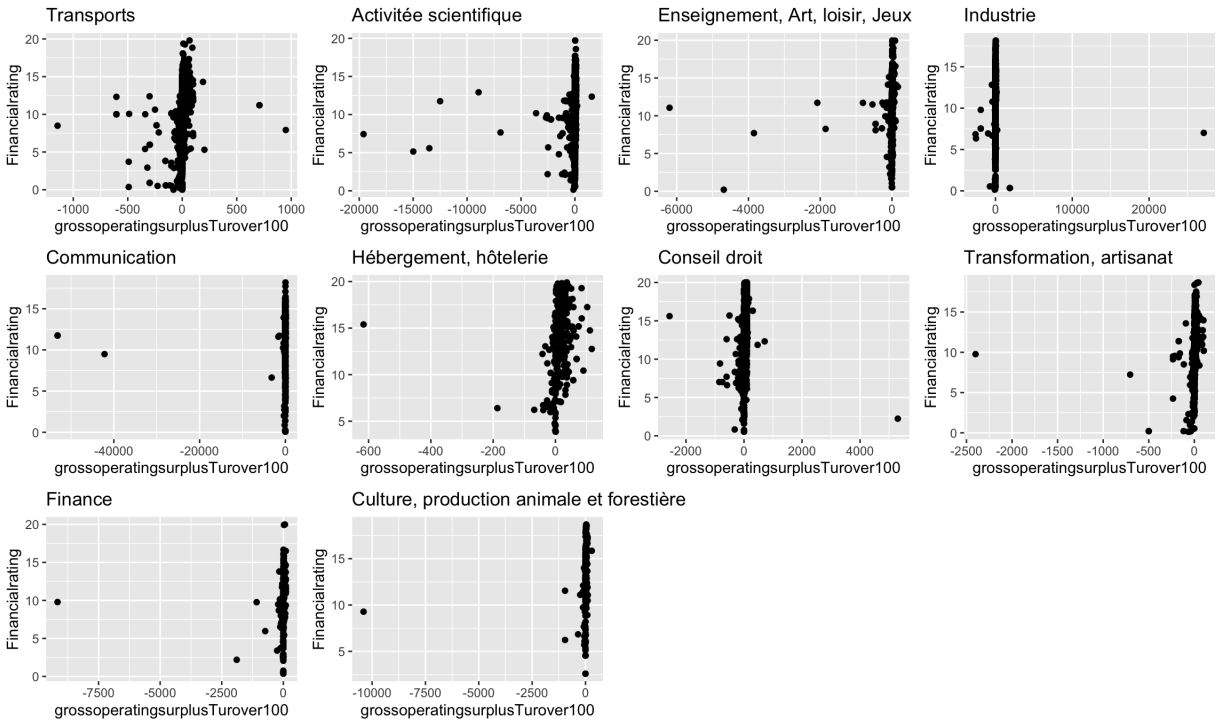

Figure 8: Scatter Plot - Gross Operating Surplus 100 on Financial Rating

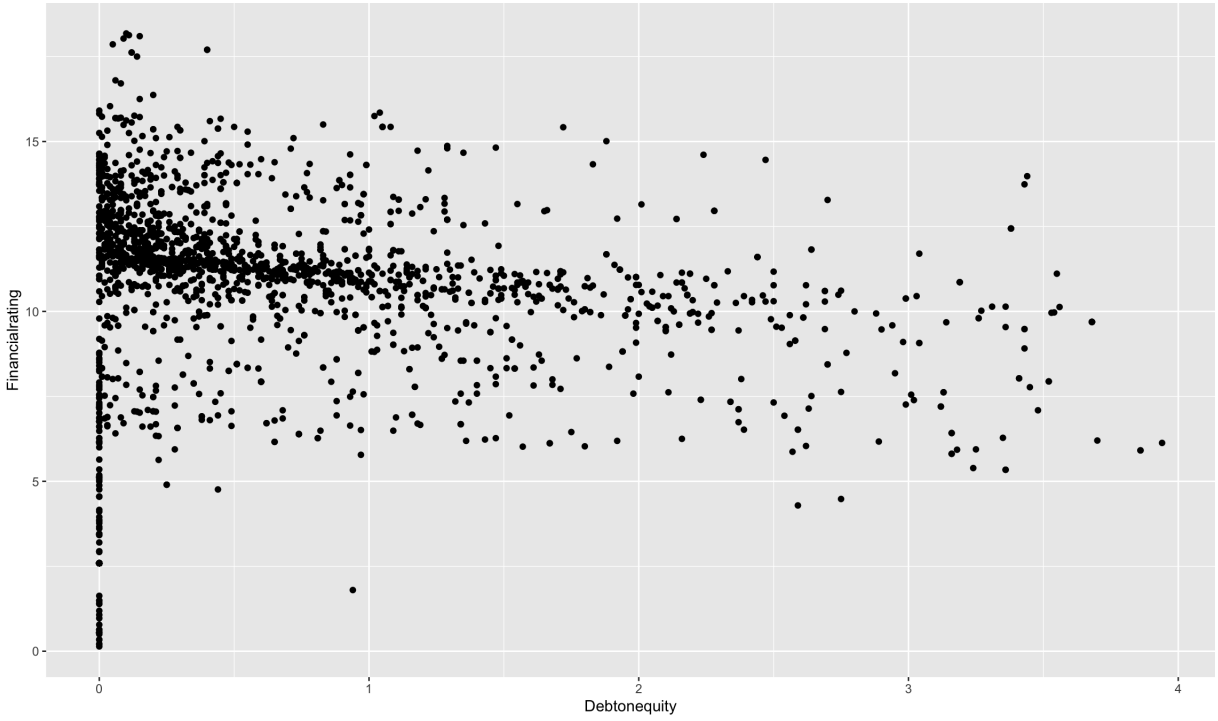

Figure 9: Scatter Plot - Debt on Equity on Financial Rating

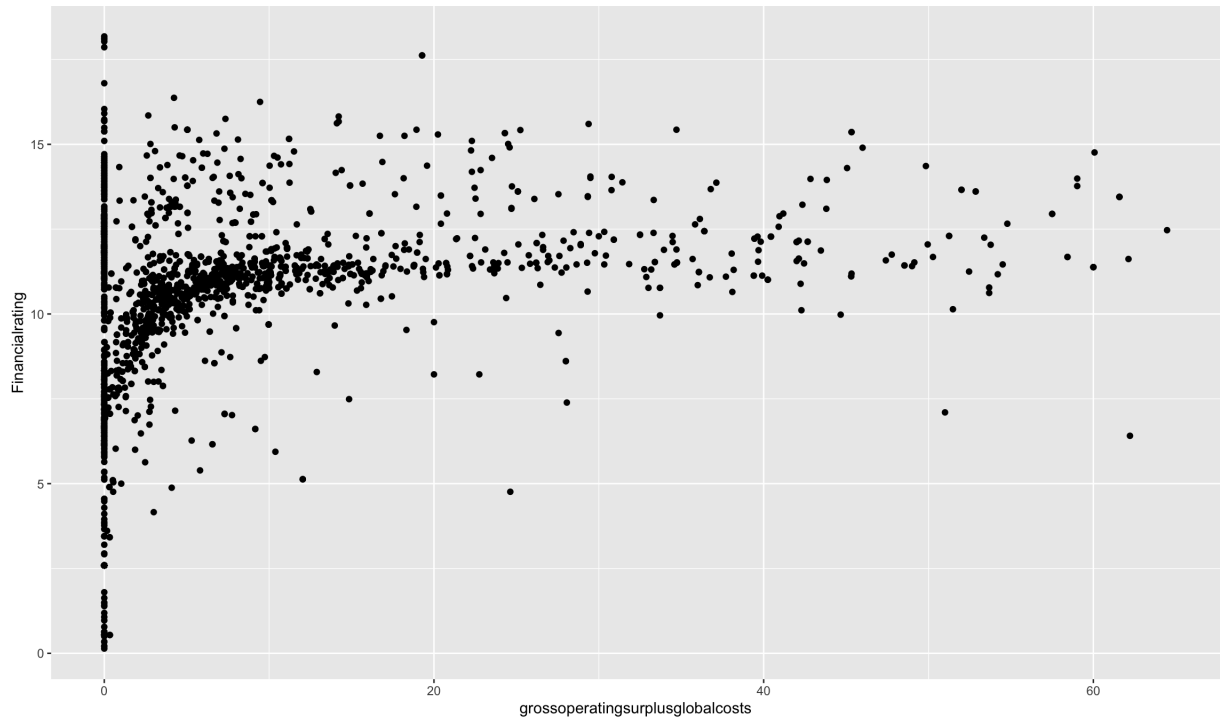

Figure 10: Scatter Plot - Gross Operating Surplus

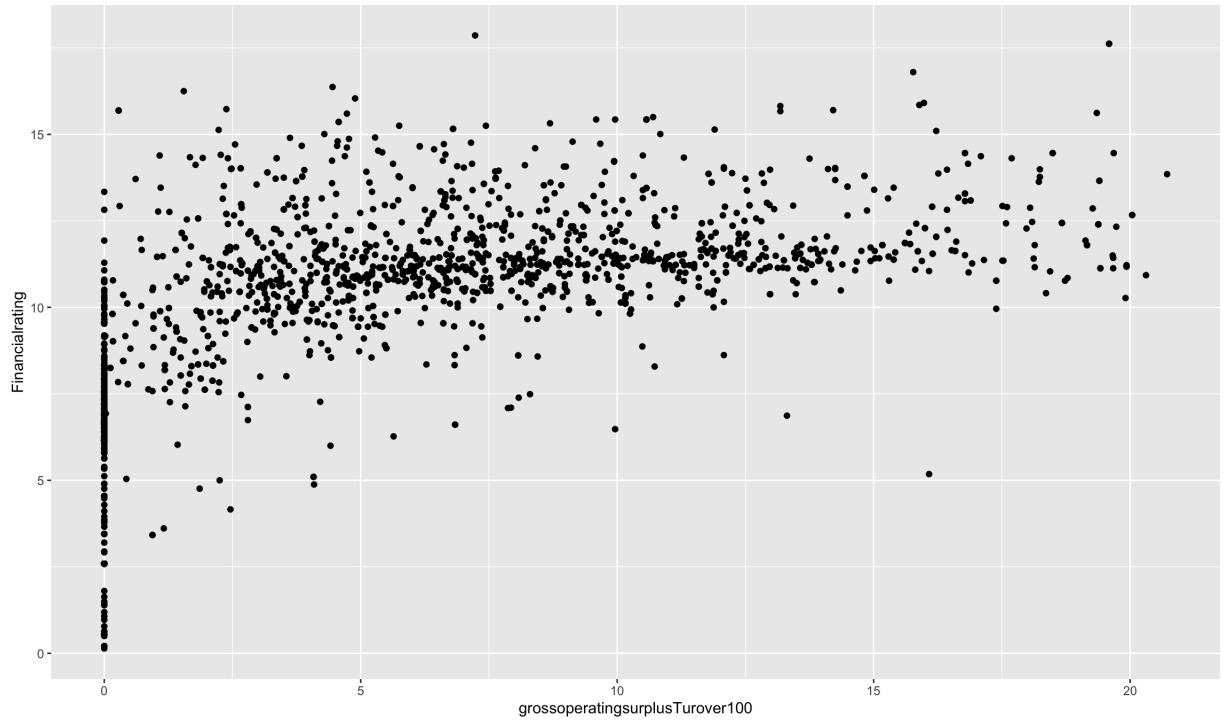

Figure 11: Scatter Plot - Gross Operating Surplus Turnover 100

References

Acce, Blog (2020). "Lire les autres articles de cette série : Accompagnement de la transition énergétique , engagement pour une finance verte et prise en compte des nouveaux risques liés au réchauffement et aux aléas climatiques figurent aujourd ' hui à l ' agenda de la maje". In: pp. 1–5.

Accentur, Blog (2020). "Les objectifs de la supervision bancaire peuvent se décliner en trois volets : Principales avancées". In: pp. 1–6.

Accenture, Blog and Banque Page (2020). "Lire les autres articles de cette série: 1". In: pp. 1–9.

Allen, Thomas et al. (2020). "Climate-Related Scenarios for Financial Stability Assessment: An Application to France". In: *SSRN Electronic Journal*. ISSN: 1556-5068. DOI: 10.2139/ssrn.3653131.

Berenguer, Maria, Michel Cardona, and Julie Evain (2020). "Intégrer les risques liés au climat dans les exigences de fonds propres des banques". In:

Craven, Peter and Grace Wahba (1978). "Smoothing noisy data with spline functions - Estimating the correct degree of smoothing by the method of generalized cross-validation". In: *Numerische Mathematik*. ISSN: 0029599X. DOI: 10.1007/BF01404567.

De Cara, Stephane (2007). "Economie du Changement Climatique". In: *Futuribles*, pp. 25–42.

ECB (2020). "Guide on climate-related and environmental risks risk management and disclosure". In: May, pp. 1–52. URL: https://www.bankingsupervision.europa.eu/press/pr/date/2020/html/ssm.pr201127~5642b6e68d.en.html.

Energie, Chaire (2017). "climatique pour le secteur bancaire : Sommaire". In: pp. 1–26.

Finance for Tomorrow (2019). "Le risque climatique en finance". In: URL: https://financefortomorrow.com/actualites/publication-du-rapport-le-risque-climatique-en-finance/.

Godard, Olivier (2007). "Le Rapport Stern sur l'économie du changement climatique était-il une manipulation grossière de la méthodologie économique ?" In: *Revue d'économie politique* 117.4, p. 475. ISSN: 0373-2630. DOI: 10.3917/redp.174.0475.

Hastie, T. J. and R. J. Tibshirani (2017). *Generalized additive models*. DOI: 10.1201/9780203753781.

IPCC (2018). "Proposed outline of the special report in 2018 on the impacts of global warming of 1 . 5 ° C above pre-industrial levels and related global greenhouse gas emission pathways , in the context of strengthening the global response to the threat of climate cha". In: *Ipcc - Sr15* 2.October, pp. 17–20. URL: www.environmentalgraphiti.org.

"Modalités techniques pour l ' exercice pilote climatique – groupes bancaires" (n.d.). In: ().

Secr, Anne-lise Bontemps-chanel (2019). "Changement climatique : quels risques pour le secteur financier français ?" In:

Trésor, D G (2020a). "Effets économiques du changement climatique". In:

— (2020b). "Impact Économique Du Changement Climatique : Revue Des Méthodologies D ' Estimation , Résult Ats Et Limites Climatique : Revue Des Méthodologies". In:

Wood, Simon N. (2003). *Thin plate regression splines*. DOI: 10.1111/1467-9868.00374.