ISSN: 2281-1346

Department of Economics and Management

DEM Working Paper Series

Proposed Coal Power Plants and Coal-To-Liquids Plants: Which Ones Survive and Why?

Dean Fantazzini
(Moscow School of Economics)

Mario Maggi
(Università di Pavia)

# 82 (07-14)

Via San Felice, 5
I-27100 Pavia
http://epmq.unipv.eu/site/home.html

July 2014

Proposed Coal Power Plants and Coal-To-Liquids Plants: Which Ones Survive and Why?

Dean Fantazzini[*], Mario Maggi[†‡]

July 2014

Abstract

The increase of oil and natural gas prices since the year 2000 stimulated the planning and construction of new coal-fired electricity generating plants and coal-to-liquids plants in the US. However, a large number of these projects have been canceled or abandoned since 2007. Using a set of 145 proposed coal power plants and 25 coal-to-liquids plants, we examine the main determinants that influence the decision to abandon a project or to proceed with it. In case of coal power plants, the number of searches performed on Google relating to coal power plants and the prices of alternative fuels for electricity generation are the main factors. As for coal-to-liquids plants, the political affiliation of the state governor is the most important factor across several model specifications. An out-of-sample exercise confirms these findings. These results hold also with robustness checks considering alternative Google search keywords and the potential effects of the recession in the years 2008-2009.

JEL CLassification: C25, C52, C53, L94, Q40, Q41

Keywords: Coal, Coal plants, Coal-to-Liquids, Logit, Probit, Training, Validation, Forecasting, Model Confidence Set, Google, Google Trends, Second Great Contraction, Global Financial Crisis

1 Introduction

The first decade of the 21st century witnessed a strong increase in oil prices due mainly to the growing demand by China and India, as well as to a growing difficulty to increase oil production worldwide, with the notable exception of North America (see Fantazzini et al. (2011), for a recent review). Similarly, US natural gas prices showed a growing trend, reaching the level of 13 \$/MMBtu in June 2008. As a consequence, the increasing power demand in the US at that time sparked a renewed interest in using coal for power generation, stimulating the planning and/or construction of almost 150 coal-fired electricity generating plants by 2007 (of Energy (2007)). Moreover, coal-to-liquids plants became an interesting alternative for producing oil liquids (see Höök et al. (2014) for a recent review of hydrocarbon liquefaction as a peak oil mitigation strategy). However, since 2007-2008 the energy landscape has changed substantially: the advent of shale gas has reduced considerably the price of natural gas in the US, reaching a minimum of 1.9 \$/MMBtu in April 2012. Besides, the costs of renewables have fallen, while the growth in electricity demand

[*] Moscow School of Economics, Moscow State University, Russia. E-mail: fantazzini@mse-msu.ru

[†] Department of Economics and Management, University of Pavia, Italy. E-mail: maggma@eco.unipv.it

[‡] This work has been partly financed by PRIN project MISURA (cod. 2010RHAHPL).

has slowed due to the economic crisis; the construction costs for coal plants have increased considerably, as well as coal prices (see Freese et al. (2011) and Fleischman et al. (2013) for recent reviews). Moreover, the US Environmental Protection Agency (EPA) has began regulating greenhouse gases (GHGs) from mobile and stationary sources of air pollution under the Clean Air Act since 2011. Furthermore, there has been an increasing awareness about the health risks posed by pollutants from power plants as synthesized by recently available Google data (more below). As a result, more than 100 coal plants projects were either cancelled or abandoned (see e.g. the Sierra Club database Club (2014) and the Coal-Swarm database for Media and Democracy (2014)), and the Energy Information Administration (EIA) expects that very few new coal plants will be built through 2040 (EIA – Annual Energy Outlook 2014 EIA (2014b)).

Even though coal is still the main source for US electricity power production, the average age of the plants is rather high. In 2011, the capacity weighted average age of coal-fired plants was 36 years, whereas it was only 18 for natural gas-fired plants (and 35 for oil-fired plants[1]), see Table 1. Refitting these coal plants to comply with the recent stricter emission standards is very expensive, so that many of them will face retirement in the upcoming years (Fleischman et al. (2013)).

Fuel type	**Coal**	**Natural Gas**	**Petroleum**
average size (MW)	245.54	85.65	15.39
average age (years)	36.34	17.88	35.16
25% built before	1967	1981	1970
50% built before	1974	2001	1972
75% built before	1981	2003	1978
CO2/capacity (Million Metric Tons/MWh)	0.9931	0.3972	0.8689

Table 1: Capacity weighted distribution of electricity power production plant by different fuel. 2011 data from `http://www.eia.gov`.

Given this background, we analyze the main determinants that influenced the decision to abandon or to proceed with a coal project using a unique dataset of 145 coal-power plants projects and 25 coal-to-liquids plants projects, observed between 2004 and 2013.

There are several reasons why investors, industry professionals and scholars may find this issue relevant. First, the amount of money and time required for planning a plant (not to mention building one) is substantial. Prior knowledge of the main factors influencing the viability of a plant project is fundamental for successful strategy and policy making. To our knowledge, this is the first study that analyzes these factors after the advent of US shale gas and the global economic crisis in 2008-2009. Moreover, this is the first study that deals with coal-to-liquids plants. We remark that even though we focus on coal plants, our findings are not limited to them: Ansolabehere and Konisky (2009) examined the 2008 MIT Energy Survey to measure public support for and opposition to the local siting of power plants and they found that "*attitudes about plant siting depend heavily on perceptions of the environmental harm and costs of specific facilities; the effects of these attributes are similar across different types of fuel sources, suggesting that there is a common underlying structure to an individual's attitude.*"

Second, a vast body of the literature has found that the public attitude toward the location of environmentally hazardous facilities is a major determinant of siting costs, which can increase quickly when the local community agreement is missing (see Ansolabehere and Konisky (2009) and Garrone and A. (2012) for extensive reviews). Measuring these attitude is not easy, can be costly and standard energy surveys may be already "old" when they are fi-

[1] The old age of oil-fired plants is also due to the fact that in US oil has a small and decreasing weight in electricity production.

nally compiled and delivered. To solve this problem, we propose the use of Google search data to measure public attitudes towards coal plants and environmental issues in general. In this regard, Google has offered since 2004 a tool called Google Trends that provides information of users' relative interest for a particular search query at a given geographic region and at given time (the data are available on a weekly or even a daily basis). In recent years, researchers worldwide have started to use online search data to forecast data in real time when data from official releases are published with a time lag (i.e. nowcasting), or simply as an additional variable for forecasting purposes (see Choi and Varian (2009), Askitas and Zimmermann (2009), Suhoy (2009), Ginsberg et al. (2009), Da et al. (2011), D'Amuri and Marcucci (2013) and Fantazzini and Fomichev (2014) for some recent applications).

Third, even though cheap shale gas and falling prices for renewables have started to reduce the coal share of US electricity generation, the most recent data seem to show that there is a future for coal: since the minimum in 2012, the price of natural gas has more than doubled due to a production slowdown that has entered a plateau (EIA Natural Gas Monthly - May 2014 EIA (2014d)). As a consequence, coal-fired electric generation increased from 3405 GWhD in March 2012 to 4227 GWhD in March 2013 and to 4419 GWhD in March 2014, while gas-fired electric generation decreased from 2984 GWhD in March 2012 to 2733 GWhD in March 2013, and to 2500 GWhD in March 2014 (EIA Electric Power Monthly - May 2014 EIA (2014c)), notwithstanding the 18 GW of coal units that were retired between 2011 and 2013 (Fleischman et al. (2013)). This recent loss in natural gas market share in favor of coal was due to the natural gas prices spiking during the winter time and the general upward trend in prices since 2012. If natural gas price continues to increase, it is likely that there will be a renewed interest in coal, given the US abundant resources. Moreover, carbon capture and storage technologies and plants using integrated gasification combined cycle (IGCC) are being developed in response to the stricter regulations by the EPA (see the Clean Coal Research – US Department of Energy (2014) EIA (2014a) for more details). Therefore, getting rid of coal-fired power plants altogether may be definitely premature.

Finally, we perform an out-of-sample forecasting comparison with a set of competing models, together with several robustness checks to verify that our results hold also with different settings. While this approach is rather standard in the medical, economic and financial literature (see e.g. Diebold (2006), Carney et al. (2010), Danielsson (2011), Hansen et al. (2011)), this is not common with studies dealing with power plants, with the notable exception of Young et al. (2011) who examined the factors influencing the location of bioenergy and biofuel plants and performed an out-of-sample analysis with alternative competing models.

The paper is organized as follows. Section 2 describes the data and methods used in our work while the empirical analysis is performed in Section 3. Robustness checks are discussed in Section 4, while Section 5 includes a brief conclusion.

2 Data and Methods

2.1 Data

The National Energy Technology Laboratory (NETL), a division of the Department of Energy, maintained a database of all new projects of coal-fired electricity generating plants, but ceased providing project-specific information as of May 2007. Since then, the *Coal*

Issues Portal on SourceWatch (a project of CoalSwarm and the Center for Media and Democracy) has maintained a dataset of the proposed coal plants in the United States and their latest status. We separated the variable "status" into two groups: one collecting all plants that are active/upcoming/operating and another group with all plants that were cancelled/abandoned or have an uncertain status[2].

Even though the Coal Issues portal contained some information about the coal projects, like the US state location and in some cases also the total capacity (in MW for power plants and bbl/day for coal-to-liquids), this information was not sufficient and was augmented by an extensive online search for each coal project. Unfortunately, this search was not successful for several plants, for which budget costs, capacity, carbon dioxide (CO2) emissions, project beginning year and project duration were not available. Therefore, the initial dataset was filtered and the final dataset consisted of 145 coal-power plant projects and 25 coal-to-liquids plant projects, observed between 2004 and 2013, whose names are reported in Tables 18-19 in Appendix. The dataset of coal-power plants projects consists of 97 plants that were cancelled/abandoned and 48 plants that are active/operating/upcoming, for a total of 574 yearly data samples. The dataset of coal-to-liquids projects consists of 17 plants that were cancelled/abandoned and 8 plants that are active/upcoming, for a total of 94 yearly data samples. We discarded the (few) projects that were either operative or cancelled before 2004, since those early projects in the NETL database had very different economics from subsequent projects (see Höök and Aleklett (2014), Fantazzini et al. (2011) and Höök et al. (2014)).

The past literature has identified four main groups of variables that can influence the plant location choice. First, R. (1960) suggested that site-specific environmental externalities should be the main determinants of location choices. In this regard, a profit-maximizing firm will try to find an agreement with the community that suffers the least damage, other things being equal. However, Hamilton (1993), Hamilton (1995) and Jenkins et al. (2004) questioned this hypothesis and advanced the idea that physical and demographic characteristics of local community can influence externalities costs: communities that show a stronger opposition are less likely to host a plant, or any environmentally hazardous facility. Therefore, any model trying to explain the location of a (coal) plant should consider a group of "voice" indicators. A third group of variables should include traditional industrial location factors like infrastructure, construction and labor costs, see Anderton et al. (1994), Been and Gupta (1997), Arora and Cason (1998), Wolverton (2009), Garrone and A. (2012). More recently, given the falling prices of renewables and natural gas, several authors have started comparing the economics of these alternative sources of electricity generation with the economics of coal plants to determine the best choice and location, see Freese et al. (2011), Cleetus et al. (2012), Tierney (2012), Fleischman et al. (2013) and Pratson et al. (2013). Given this background, Table 2 illustrates the regressors that we used to explain the status of a coal plant project.

We used the state population in millions and the CO2 output in tons to measure the externalities costs a state can suffer given that the larger the population and the CO2 amount the larger the perception of the expected environmental damage and the smaller the probability a site will be located there (see Hamilton (1993), Boer et al. (1997), Garrone and A. (2012))[3].

[2] An online search allowed us to find that all plants with an uncertain status were either cancelled or abandoned. They had no related news for years.

[3] We tried the population density in place of the population data, as done by Garrone and A. (2012), but this resulted in worse in-sample results, models' residuals and out-of-sample results. Therefore, we used the population data instead.

Variables	**Description**	**Sources**
Externalities costs		
CO2(TONS)	Carbon Dioxide output in tons	Carbon Monitoring for Action (CARMA) database
POPULATION	Population by US state in millions	U.S. Department of Commerce: Census Bureau
Awareness and ability to pay for environmental quality		
INCOME	Median Household Income by US state	U.S. Department of Commerce: Census Bureau
LFP	Labor Force Participation by US state	U.S. Department of Labor: Bureau of Labor Statistics
UR	Unemployment Rate by US state	U.S. Department of Labor: Bureau of Labor Statistics
GI(JOBS)	Google index for the keyword "jobs"	Google Trends
Awareness and voice factors		
GI (COAL)	Google index for the keyword "coal"	Google Trends
GI(COAL POWER +COAL PLANT)	Google index for the keywords "coal power+coal plant"	Google Trends
GI(COAL-TO-LIQUIDS + CTL COAL)	Google index for the keywords "coal-to-liquids+ctl coal"	Google Trends
GI(POLLUTION)	Google index for the keyword "pollution"	Google Trends
GOVERNOR	Binary variable that is 1 if Republican and 0 otherwise	www.rulers.org
Traditional industrial location factors		
COST	Plant cost estimate (billion $)	CMD / Google search
COAL PRICE	US Central Appalachian coal spot price ($/ton)	BP Statistical Review of World Energy 2013 / US EIA
RAIL	Rail miles by US state	Association of American Railroads
CAPACITY(MW)	Plant capacity expressed in MW for coal power	NETL-US DOE / CMD / Google search
CAPACITY(BBL/DAY)	Plant capacity expressed in bbl/day for CTL plants	
ELECTRICITY	Average electricity price by US state($/Kwh)	US Energy Information Administration (EIA)
Economics of alternative energy sources		
WIND PRICE	Average levelized long-term wind power purchase agreement prices ($/Mwh)	US Department of Energy / Energy Analysis and Environmental Impacts Department - Lawrence Berkeley National Laboratory
SOLAR PRICE	Installed price of residential and commercial solar photo-voltaics ($/W)	US Department of Energy (DOE) / Lawrence Berkeley National Laboratory
NG PRICE	US Henry Hub natural gas price ($/MmBtu)	BP Statistical Review of World Energy 2013 / US EIA
Additional indicators		
DURATION	The number of years that has passed at time t since the project started	The National Energy Technology Laboratory (NETL) The Center for Media and Democracy (CMD) Google search

Table 2: Regressors: description and source

Four indicators were used to represent the awareness of local residents and their ability to pay for environmental quality: the median household income, the labor force participation, the unemployment rate and the Google Index (GI) for the keyword "jobs." The GI is computed as the ratio of the search queries for a specific keyword (or group of keywords) relative to the total number of searches performed in the selected region at a given point of time and then standardized between 0 and 100 (where the standardization is done over the whole time period and all the considered searches). The data were collected for each US state for the period January 2004 through December 2013. The Google data have a weekly frequency and they were converted to a yearly frequency by taking average values to match coal plant data. D'Amuri and Marcucci (2013) found this GI to be the best predictor for the US unemployment rate.

To measure awareness and the ability to organize protests against coal plant projects as well as to ask for compensation (the so-called "voice" factors), we used the GI for the keyword "coal," the GI for the keywords "coal plant+coal power," the GI for the keywords "coal-to-liquids+ctl coal" and the GI for the keyword "pollution." In this regard, the analysis of Google data showed that several searches for the previous keywords included and/or were related also to "legal action," "protest," "stop," etc., which clearly highlights that separating awareness from voice factors may not be immediate[4]. Following Ansolabehere and Konisky (2009), we also used the political affiliation of the state governor as an important voice factor.

Five indicators were used to consider traditional industrial location factors: the plant cost estimate, the plant capacity, the coal price, the available rail miles (which is important for coal transportation) and the average electricity price. We remark that the latter can also be interpreted as a measure of (past) profitability. Moreover, the plant cost estimates

[4] Alternative keywords for Google search with smaller search volumes will be analyzed in Section 4 dealing with robustness checks.

were updated each year using the Chemical Engineering Plant Cost Index (CEPCI), which is a dimensionless number used to update the capital cost required to build a chemical plant from a past date to a later time. This index is widely accepted and consists of subcomponents dealing with equipment, labour costs, buildings, engineering, supervision and other parameters affecting costs. Kreutz et al. (2008) provide a comparison of the CEPCI with the Marshall and Swift index, the US GDP deflator and the Handy-Whitman Total Plant-All Steam Generation Index, while Höök et al. (2014) used this index to compute the economics of coal-to-liquids and gas-to-liquids plants.

The competition with alternative energy sources was measured by using the average levelized long-term wind price, the average price of residential and commercial solar photovoltaics, and the Henry Hub natural gas price.

Finally, we considered also an indicator to measure how many years have passed since the coal plant project started: we noted that the more time the project spends in its planning phase the less probable will be its full development. Given our dataset, we found two reasons for this phenomenon: strong cost escalations and a prolonged legal battle between the local communities and the plant developers. Often, these two reasons were interconnected: the legal battle delayed the coal project to such an extent that the new price environment was no longer profitable due to cost escalations and falling prices of energy alternatives (see the Coal Issues Portal and the history of each coal plant reported there). Moreover, this phenomenon also confirms again that separating the different indicators in clear-cut categories is not always possible.

All data had yearly frequency or were converted to a yearly frequency to match the coal plants data, the only exception being the plant capacities that were held constant for the period of observation. Moreover, all data were transformed into logs, except for the *duration* indicator and the binary variable *governor*. However, in the subsequent section, devoted to the out-of-sample forecasting analysis, we considered a wide set of models, including models with data in levels, that is without any transformation.

The first analysis that we performed after collecting the data was to compute the correlation among regressors (see Figures 1 and 2), as well as the Variance Inflation Factors[5] (VIF) for each regressors (see Tables 3 and 4), where we differentiated between coal power plants and coal-to-liquids plants[6].

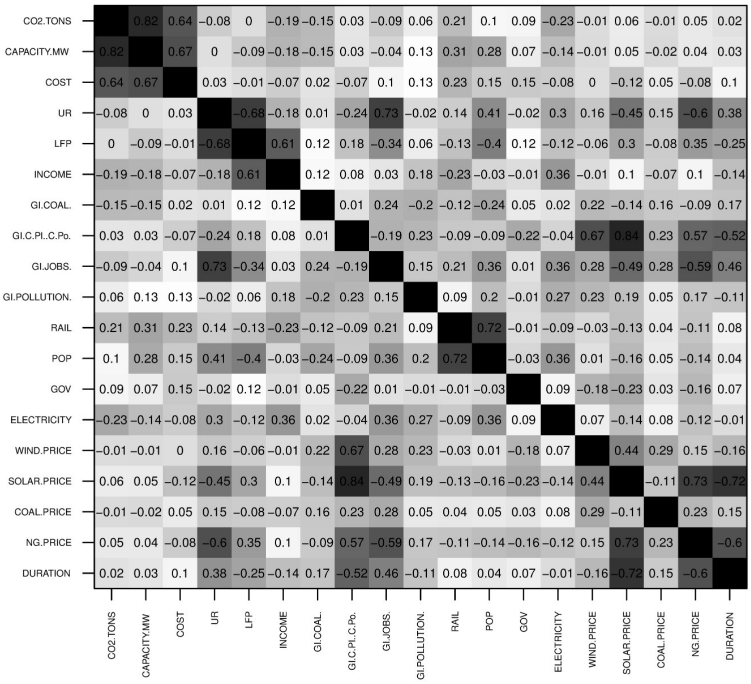
CO2(TONS)	3.60	RAIL	5.57
CAPACITY(MW)	4.19	POPULATION	7.95
COST	2.18	GOVERNOR	1.29
UR	5.17	ELECTRICITY	2.16
LFP	6.91	WIND PRICE	3.03
INCOME	4.22	**SOLAR PRICE**	**19.83**
GI (COAL)	1.67	COAL PRICE	4.09
GI(COAL PLANT + COAL PLANT)	**10.18**	NG PRICE	7.68
GI(JOBS)	4.78	DURATION	2.55
GI(POLLUTION)	1.52		

Table 3: Variance Inflation Factors (VIF): regressors for coal power plants. VIFs higher than 10 are in bold font.

Figures 1 and 2 and Tables 3 and 4 show that some indicators display a high degree of collinearity, even though this result was expected for many of them. Some examples are the strong correlation between cost, capacity and CO2 output, and the correlation

[5] Variance Inflation Factors are used to measure the degree of collinearity among the regressors in a linear equation. They can be computed by dividing the variance of a coefficient estimate with all the other regressors included, by the variance of the same coefficient estimated from an equation with only that regressor and a constant.

[6] The CO2 output was not considered for coal-to-liquids plants, given the very few plants for which this data was available.



	CO2.TONS	CAPACITY.MW	COST	UR	LFP	INCOME	GI.COAL.	GI.C.PI..C.Po.	GI.JOBS.	GI.POLLUTION.	RAIL	POP	GOV	ELECTRICITY	WIND.PRICE	SOLAR.PRICE	COAL.PRICE	NG.PRICE	DURATION
CO2.TONS		0.82	0.64	−0.08	0	−0.19	−0.15	0.03	−0.09	0.06	0.21	0.1	0.09	−0.23	−0.01	0.06	−0.01	0.05	0.02
CAPACITY.MW	0.82		0.67	0	−0.09	−0.18	−0.15	0.03	−0.04	0.13	0.31	0.28	0.07	−0.14	−0.01	0.05	−0.02	0.04	0.03
COST	0.64	0.67		0.03	−0.01	−0.07	0.02	−0.07	0.1	0.13	0.23	0.15	0.15	−0.08	0	−0.12	0.05	−0.08	0.1
UR	−0.08	0	0.03		−0.68	−0.18	0.01	−0.24	0.73	−0.02	0.14	0.41	−0.02	0.3	0.16	−0.45	0.15	−0.6	0.38
LFP	0	−0.09	−0.01	−0.68		0.61	0.12	0.18	−0.34	0.06	−0.13	−0.4	0.12	−0.12	−0.06	0.3	−0.08	0.35	−0.25
INCOME	−0.19	−0.18	−0.07	−0.18	0.61		0.12	0.08	0.03	0.18	−0.23	−0.03	−0.01	0.36	−0.01	0.1	−0.07	0.1	−0.14
GI.COAL.	−0.15	−0.15	0.02	0.01	0.12	0.12		0.01	0.24	−0.2	−0.12	−0.24	0.05	0.02	0.22	−0.14	0.16	−0.09	0.17
GI.C.PI..C.Po.	0.03	0.03	−0.07	−0.24	0.18	0.08	0.01		−0.19	0.23	−0.09	−0.09	−0.22	−0.04	0.67	0.84	0.23	0.57	−0.52
GI.JOBS.	−0.09	−0.04	0.1	0.73	−0.34	0.03	0.24	−0.19		0.15	0.21	0.36	0.01	0.36	0.28	−0.49	0.28	−0.59	0.46
GI.POLLUTION.	0.06	0.13	0.13	−0.02	0.06	0.18	−0.2	0.23	0.15		0.09	0.2	−0.01	0.27	0.23	0.19	0.05	0.17	−0.11
RAIL	0.21	0.31	0.23	0.14	−0.13	−0.23	−0.12	−0.09	0.21	0.09		0.72	−0.01	−0.09	−0.03	−0.13	0.04	−0.11	0.08
POP	0.1	0.28	0.15	0.41	−0.4	−0.03	−0.24	−0.09	0.36	0.2	0.72		−0.03	0.36	0.01	−0.16	0.05	−0.14	0.04
GOV	0.09	0.07	0.15	−0.02	0.12	−0.01	0.05	−0.22	0.01	−0.01	−0.01	−0.03		0.09	−0.18	−0.23	0.03	−0.16	0.07
ELECTRICITY	−0.23	−0.14	−0.08	0.3	−0.12	0.36	0.02	−0.04	0.36	0.27	−0.09	0.36	0.09		0.07	−0.14	0.08	−0.12	−0.01
WIND.PRICE	−0.01	−0.01	0	0.16	−0.06	−0.01	0.22	0.67	0.28	0.23	−0.03	0.01	−0.18	0.07		0.44	0.29	0.15	−0.16
SOLAR.PRICE	0.06	0.05	−0.12	−0.45	0.3	0.1	−0.14	0.84	−0.49	0.19	−0.13	−0.16	−0.23	−0.14	0.44		−0.11	0.73	−0.72
COAL.PRICE	−0.01	−0.02	0.05	0.15	−0.08	−0.07	0.16	0.23	0.28	0.05	0.04	0.05	0.03	0.08	0.29	−0.11		0.23	0.15
NG.PRICE	0.05	0.04	−0.08	−0.6	0.35	0.1	−0.09	0.57	−0.59	0.17	−0.11	−0.14	−0.16	−0.12	0.15	0.73	0.23		−0.6
DURATION	0.02	0.03	0.1	0.38	−0.25	−0.14	0.17	−0.52	0.46	−0.11	0.08	0.04	0.07	−0.01	−0.16	−0.72	0.15	−0.6	

Figure 1: Correlation among regressors: Coal power plants

CAPACITY (BBL/DAY)	9.58	**RAIL**	**15.26**
COST	8.74	**POPULATION**	**16.87**
UR	6.71	GOVERNOR	1.86
LFP	7.91	ELECTRICITY	5.58
INCOME	5.71	**WIND PRICE**	**18.85**
GI(COAL)	3.76	**SOLAR PRICE**	**38.53**
GI(COAL-TO-LIQUIDS+CTL)	**11.37**	COALPRICE	3.13
GI(JOBS)	6.04	NG PRICE	7.89
GI(POLLUTION)	2.06	DURATION	3.36

Table 4: Variance Inflation Factors (VIF): regressors for coal-to-liquids plants. VIFs higher than 10 are in bold font.

between solar and wind prices, not to say the correlation between population and rail miles. Classical "rules of thumbs" to get rid of collinearity are to eliminate those variables with a VIF higher than 10 or to eliminate one of the two variables with a correlation higher than 0.7-0.8 (in absolute value) (see O'Brien (2007) and Dormann et al. (2013) for extensive reviews of collinearity and methods to deal with it). Unfortunately, eliminating variables can be a solution worse than the initial problem, as clearly shown by O'Brien (2007) and Dormann et al. (2013). Therefore, we preferred to follow a less aggressive approach that considers economic and financial logic: in the case that two variables have a correlation coefficient (in absolute value) higher than 0.8, we took the first one and the ratio between the first and the second one, where this ratio should have an economic and/or financial meaning whenever possible. As the consequence we considered the following ratios:

- CO2/capacity in place of CO2 output (for coal power plants only);
- cost/capacity in the place of cost;

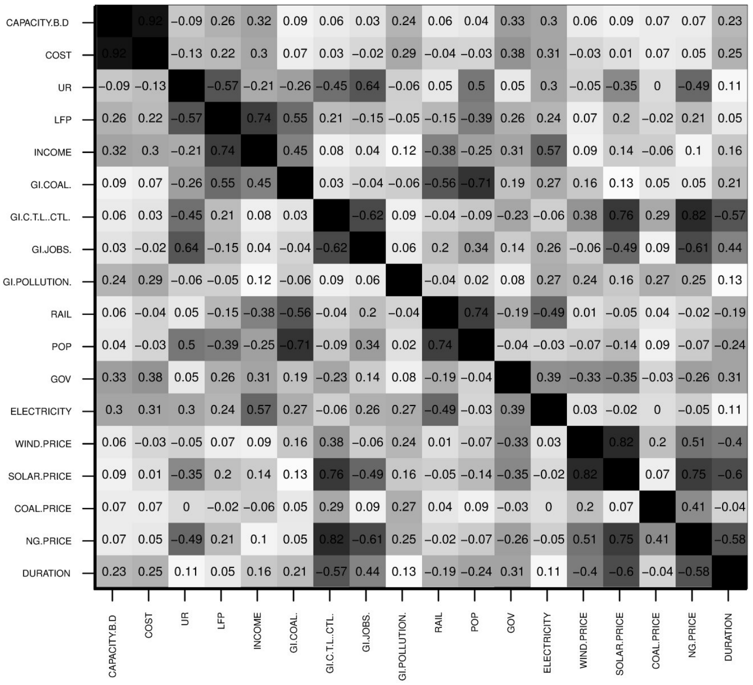


Figure 2: Correlation among regressors: Coal-to-liquids plants

- rail/population in place of rail miles;
- solar price/natural gas price in place of solar price (for coal power plants only);
- solar price/wind price in place of solar price (for coal-to-liquids plants only);
- the Google index for the keywords "coal-to-liquids+ctl" divided by the natural gas price, in place of the initial GI (for coal-to-liquids plants only).

The last one is the only ratio without an immediate economic and/or financial meaning. However, considering that hydrocarbon liquefaction can be implemented either using coal or using natural gas, the previous ratio can be roughly interpreted as a ratio between the interest for coal-to-liquids plants and gas-to-liquids plants.

The next step was to check whether our data are stationary. Given the moderate size of our dataset in case of coal power plants and the small size for coal-to-liquids plants, we employed a battery of panel unit root tests: the test by Levin et al. (2002), the test by Im et al. (2003), the Fisher-type tests using ADF and PP tests by Maddala and Wu (1999) and Choi (2001), and the Hadri test (Hadri (2000)). The first three of them test the null hypothesis of a unit root, whereas the latter tests the null of stationarity. Panel unit root tests tend to have higher power than unit root tests based on individual time series. For the Hadri test we also considered bootstrap critical values that allow for potential cross-sectional dependence: this extension was proposed by Carrion-i Silvestre et al. (2005) and represents an example of second-generation panel unit root test. See (Baltagi, 1961, chapter 23) for a detailed review of panel unit root tests. Given the very short time dimension of our dataset, we could not implement any panel unit root test with structural breaks[7].

The results in Tables 5-6 show that our data are stationary.

Unit Root Tests	Statistic	P-values
Null: Unit root (assumes common unit root process)		
Levin, Lin and Chu t-stat	-33.07	0.00
Null: Unit root (assumes individual unit root process)		
Im, Pesaran and Shin W-stat	-46.98	0.00
ADF - Fisher Chi-square	1596.10	0.00
PP - Fisher Chi-square	1712.77	0.00
Null Hypothesis: Stationarity		
Hadri test (homogeneity)	-2.92	1.00
Hadri test (heterogeneity)	0.27	0.40
Bootstrap critical values (Bartlett kernel)	5% C.V.	1% C.V.
Hadri test (homogeneity)	4.55	8.99
Hadri test (heterogeneity)	2.70	4.68

Table 5: Panel unit root tests: coal power plants.

Unit Root Tests	Statistic	P-values
Null: Unit root (assumes common unit root process)		
Levin, Lin and Chu t-stat	-12.56	0.00
Null: Unit root (assumes individual unit root process)		
Im, Pesaran and Shin W-stat	-19.76	0.00
ADF - Fisher Chi-square	333.10	0.00
PP - Fisher Chi-square	291.10	0.00
Null Hypothesis: Stationarity		
Hadri test (homogeneity)	-2.94	1.00
Hadri test (heterogeneity)	-1.94	0.97
Bootstrap critical values (Bartlett kernel)	5% C.V.	1% C.V.
Hadri test (homogeneity)	4.28	7.52
Hadri test (heterogeneity)	3.04	5.18

Table 6: Panel unit root tests: coal-to-liquids plants.

2.2 Methods

We first introduce some unifying notation that we will use throughout our work. For observation i, $(i = 1, \dots, n)$, time t, $(t = 1, \dots, T)$, let Y_{it} denotes the response variable that indicates whether a coal project is active/upcoming $(Y_{it} = 0)$ or abandoned/canceled $(Y_{it} = 1)$, while X_{it} denote a $p \times 1$ vector of regressors.

We are interested in predicting the expectation of the response variable as a function of the regressors. The expectation of a simple binary response is just the probability that the response is 1:

$$E(Y_{it}|X_{it}) = \pi(Y_{it} = 1|X_{it})$$

In linear regression, this expectation is modelled as a linear function $\beta' X_{it}$ of the regressors. For binary responses, as in our case, this approach may be problematic because the probability must lie between 0 and 1, whereas regression lines are not limited. Instead, a nonlinear regression is specified in one of two ways:

$$\pi(Y_{it} = 1|X_{it}) = h(\beta' X_{it}),$$

or

$$g\{\pi(Y_{it} = 1|X_{it})\} = \beta' X_{it} = \nu_i,$$

where ν_i is referred to as the *linear predictor*. These two formulations are equivalent if the function $h(\cdot)$ is the inverse of the *link function* $g(\cdot)$. We have introduced two components of

[7] However, in Section 4 dedicated to robustness checks, we will examine the potential effect of the global financial crisis using a dummy variable for the recession in the US that took place in the years 2008 and 2009.

a generalized linear model: the linear predictor and the link function. The third component is the distribution of the response given the regressors. For binary response, this is always specified as a Bernoulli distribution (π_i). Typical choices for the link function g are the logit or probit links. The logit link is appealing because it produces a linear model for the log of the odds, $\log\left\{\frac{\pi(Y_{it}=1|X_{it})}{1-\pi(Y_{it}=1|X_{it})}\right\}$, implying a multiplicative model for the odds themselves (for more details, see e.g. A. (2002), Rabe-Hesketh and Skrondal (2004) and Rabe-Hesketh and Skrondal (2005)). We remark that the logistic regression model can also be viewed as a latent response model, which assumes that underlying the observed dichotomous response Y_{it} there is a continuous response Y^*_{it}: if the latter is greater than 0, the observed response is 1. A linear regression model is specified for this latent response and the error term in the regression model can follow a normal or a logit distribution:

$$\begin{aligned} Y_{it} &= \begin{cases} 1 & \text{if} \quad Y^*_{it} > 0 \\ 0 & \text{otherwise} \end{cases} \\ Y^*_{it} &= \beta' X_{it} + \varepsilon_{it} \end{aligned}$$

See W. (2011) for more details about this alternative interpretation. To relax the assumption of conditional independence among the coal plants given the regressors, we can include a plant-specific random intercept $\varsigma_i \sim N(0, \psi)$ in the linear predictor:

$$g\{\pi(Y_{it} = 1|X_{it})\} = \beta' X_{it} + \varsigma_i$$

This last model with a logit or a probit link, is called a random effects panel logit/probit model in the econometric literature, see Fantazzini et al. (2009) for a recent application with credit risk data. Similarly, we can add a state-specific random intercept $\varsigma_j \sim N(0, \varphi)$ and/or random coefficients. However, all models with random intercepts and/or random coefficients either they did not converge numerically, or they showed variances ψ, φ that were not statistically different from zero (these results are not reported but they are available from the authors upon request). Moreover, we point out that fixed effects panel logit models were not considered since they would have implied working with abandoned/canceled projects only, that is with the only binary data having sequences different from 0,0,0,. . . – see Cameron and Trivedi (2005) for more details. Therefore, in the following discussion, we will only consider simple (pooled) logit and probit models.

2.3 Model Evaluation

The intensive widespread use of computational methods has led to the development of intensive model selection criteria, see e.g. Giudici and Figini (2009) for a review of model comparison. In particular, we will report the standard Akaike and Schwartz information criteria (AIC and SIC, respectively) for each model. Moreover, we will also compute the Ljung-Box Ljung and G. (1979) test statistic for testing the absence of autocorrelation up to order k in the models' standardized residuals and residuals squared, as well as the BDS test by W. et al. (1996), to test whether the standardized residuals are independent and identically distributed (iid). This test is robust against a variety of possible deviations from independence, including linear dependence, non-linear dependence, or chaos.
Given that we work with binary data, we will focus on the results coming from the predictive classification table known as *confusion matrix* (Kohavi and Provost (1998)). A confusion matrix contains information about actual and predicted classifications obtained through a classification system, and the performance of a model is commonly evaluated

$Observed/Predicted$	$EVENT$	$NON - EVENT$
$EVENT$	a	b
$NON - EVENT$	c	d

Table 7: Theoretical confusion matrix. Number of: a true positive, b false positive, c false negative, d true negative.

using the data in the matrix. Table 7 shows the confusion matrix for a two class classifier (which is our case with binary models).

In the specific context of our analysis, the entries in the confusion matrix have the following meaning:

a is the number of correct predictions that a project is abandoned/canceled,
b is the number of incorrect predictions that a project is abandoned/canceled,
c is the number of incorrect predictions that a project is active/upcoming, and
d is the number of correct predictions that a project is active/upcoming.

Given a confusion matrix the following conditional frequencies have a relevant role:

- *sensitivity*: $a/(a+b)$ proportion of correctly predicted events (*hit rate*);
- *specificity*: $d/(d+c)$ proportion of correctly predicted non events;
- *false positive rate* (or *False Alarm Rate*): $c/(c+d)$ or equivalently, $1 - specificity$, proportion of non events predicted as events (type II error);
- *false negative rate*: $b/(a+b)$ or equivalently, $1 -$ *sensitivity*, proportions of events predicted as non events (type I error).

A classifier is said *cut-off dependent* if the classification depends on a discrimination threshold (the cut-off) applied to the score produced by the underlying model (i.e. the estimated $\pi(Y_{it} = 1|X_{it}) = h(\beta' X_{it})$ in a logit regression). For a cut-off dependent classifier a common performance evaluation tool is the ROC (Receiver Operating Characteristic) curve by Metz and Kronman (1980), Goin (1982) and Hanley and McNeil (1982). The ROC curve is obtained by plotting, for any given cut-off level, the sensitivity (y-axis) with respect to the false positive ratio (x-axis). Each point in the curve corresponds therefore to a particular cut-off, so that the ROC curve can also be used to select a cut-off point, trading-off sensitivity and specificity. In terms of model comparison, the best curve is the one that is leftmost, the ideal one coinciding with the y-axis (see Krzanowski and Hand (2009) and Fantazzini et al. (2009) for a recent application). However, while the ROC curve is independent of class distribution or error costs (Provost et al. (1998)), nevertheless it is cut-off dependent. Recent literature proposed the calculation of the Area Under the ROC-Curve (AUC) as a cut-off independent measure of predictive performance, see e.g. Buckland and Augustin (1997). The AUC is always between zero and one and the closer it is to one, the more accurate the rating system is. We will report the AUC for all competing models.

Even though the AUC is one of the most common measures to evaluate the discriminative power of a predictive model for binary data, it has also some drawbacks, as recently reviewed by Figini and Maggi (2014) and references therein. In particular, as stated in (Krzanowski and Hand, 2009, p. 108), "*one can easily conjure up examples in which the AUC for classifier 1 is larger than the AUC for classifier 2, even though classifier 2 is superior to classifier 1 for almost all choices of the classification threshold.*"

Therefore, we also computed the *Model Confidence Set* (MCS), proposed by Hansen et al. (2011) and extended by Figini and Maggi (2014) to binary models, to assess the prediction power of the competing models. Following Hansen et al. (2011), the MCS procedure selects the best model and computes the probability that other models are undistinguishable from the best one using an evaluation rule based on a *loss function.* In general, the more the data are informative, the smaller the MCS will be. In this work, we computed the MCS following the procedure set up by Hansen et al. (2011)[8], adopting the χ^2 test for the model elimination rule and using different loss functions[9].

In our analysis, we measure the predicting performances of the competing models evaluating the average loss $\frac{1}{n}\sum_{i=1}^{n} L\left(P_{it}, Y_{it}\right)^2$, where $P_{it} = P\left[Y_{it} = 1 \mid X_{it}\right]$, and n is the validation sample size. We will consider the following loss functions:

$$
\begin{array}{ll}
\text{Square loss} & L(P,Y) = (P-Y)^2 \\
\text{Spherical loss} & L(P,Y) = \begin{cases} 1 - \dfrac{P}{\sqrt{P^2+(1-P)^2}}, & \text{if } Y=1 \\ 1 - \dfrac{1-P}{\sqrt{P^2+(1-P)^2}}, & \text{if } Y=0 \end{cases} \\
\text{Logarithmic loss} & L(P,Y) = \begin{cases} -\log(P), & \text{if } Y=1 \\ -\log(1-P), & \text{if } Y=0 \end{cases} \\
\text{Asymmetric quadratic loss} & L(P,Y) = \begin{cases} k\left(1 - \frac{(1-c)^2-(1-P)^2}{T(c)}\right), & Y=1 \\ k\left(1 - \frac{c^2-P^2}{T(c)}\right), & Y=0 \end{cases} \\
 & \text{where } c \in [0,1], \quad T(c) = \begin{cases} (1-c)^2, & P \geq c \\ c^2, & P < c \end{cases} \\
 & \text{and} \quad k = \begin{cases} \frac{c}{2}, & c \leq \frac{1}{2} \\ \frac{1-c}{2}, & c > \frac{1}{2} \end{cases}
\end{array}
\tag{1}
$$

We remark that the first three loss functions in (1) are symmetric around $\frac{1}{2}$; this means that the magnitude of the corresponding value of the loss function does not depend on the sign of the error: $L(P, Y \mid Y = 0) = L(1 - P, Y \mid Y = 1)$. Following Winkler (1994), it is possible to generalize symmetric loss functions, assigning different weights to the different kinds of errors. The asymmetric quadratic loss has the ability to take this fact into account and it directly generalizes the square loss, which can be obtained setting $c = \frac{1}{2}$, see Figini and Maggi (2014) for further details.

3 Results

3.1 In-sample Analysis

We report the estimated models for coal power plants and coal-to-liquids plants in Tables 8-9, respectively, where the left columns show the results with all the regressors, while the right columns the restricted models with only the regressors that were significant at the 5% level (for coal power plants) and at the 10% level (for coal-to-liquids plants)[10].

[8] The results are obtained running the Ox package Mulcom 2.00 (`http://mit.econ.au.dk/vip_htm/alunde/mulcom/mulcom.htm`). This package can be run with the Ox console (version 6.2), which is free for academic research, study and teaching purposes (`http://www.doornik.com/download.html`).

[9] Other tests can be applied: for instance, the F statistic or other statistics built on the t-statistic that do not require the computation of the model covariance matrix. In our applications, the F statistic and other t-statistics delivered similar results to the χ^2. However, the t-statistics are much more demanding in terms of computing time and are convenient when the number of models is large relative to the sample, which is not our case.

[10] We used a higher probability level for coal-to-liquids plants due to the small size of the dataset.

	LOGIT		PROBIT		LOGIT restricted		PROBIT restricted	
	Coef.	*P-value*	*Coef.*	*P-value*	*Coef.*	*P-value*	*Coef.*	*P-value*
CO2/CAPACITY	0.95	0.21	0.48	0.26				
GI(COAL)	0.09	0.62	0.06	0.57				
COAL PRICE	**-1.32**	**0.04**	**-0.71**	**0.04**	**-1.59**	**0.00**	**-0.86**	**0.00**
GI(COAL PLANT + COAL POWER)	**13.17**	**0.00**	**7.35**	**0.00**	**14.18**	**0.00**	**7.85**	**0.00**
COST/CAPACITY	-0.53	0.55	-0.28	0.59				
DURATION	**0.22**	**0.03**	**0.13**	**0.03**	**0.22**	**0.02**	**0.12**	**0.02**
ELECTRICITY	-0.01	0.99	-0.01	0.98				
GOVERNOR	-0.29	0.26	-0.17	0.23				
INCOME	1.36	0.40	0.84	0.37				
GI(JOBS)	1.06	0.38	0.63	0.36				
LFP	-3.22	0.37	-1.96	0.34				
NG PRICE	**-10.32**	**0.00**	**-5.72**	**0.00**	**-10.22**	**0.00**	**-5.72**	**0.00**
GI(POLLUTION)	0.02	0.94	0.02	0.88				
POPULATION	-0.25	0.15	-0.14	0.16				
RAIL/POPULATION	-4.19	0.16	-2.25	0.16				
SOLAR PRICE/NG PRICE	**-14.78**	**0.00**	**-8.13**	**0.00**	**-15.34**	**0.00**	**-8.53**	**0.00**
CAPACITY (MW)	0.51	0.08	0.26	0.11				
UR	-1.26	0.08	-0.74	0.06				
WIND PRICE	0.24	0.80	0.03	0.95				
CONSTANT	-16.63	0.22	-9.51	0.21	**-14.77**	**0.00**	**-8.20**	**0.00**
Information criteria and AUC								
AIC		516.52		515.44		501.22		**500.12**
SIC		603.58		602.50		527.33		**526.23**
AUC		71.02%		71.08%		67.08%		67.11%
Residual tests								
Ljung-Box(50) res.[p-val]		0.21		**0.03**		0.36		0.51
Ljung-Box(50) res.sq.[p-val]		0.29		0.34		**0.00**		0.50
BDS(dim=2) [p-val]		0.09		0.33		0.14		0.09
BDS(dim=6) [p-val]		0.56		0.91		0.56		0.57

Table 8: Coal Power Plants: Model Estimation Results. Smallest information criteria and p-values smaller than 5 % are reported in bold font.

	LOGIT		PROBIT		LOGIT restricted		PROBIT restricted	
	Coef.	*P-value*	*Coef.*	*P-value*	*Coef.*	*P-value*	*Coef.*	*P-value*
GI(COAL)	1.59	0.31	0.92	0.30				
COAL PRICE	0.31	0.88	0.18	0.87				
GI(COAL-TO-LIQUIDS+CTL)/NG PRICE	0.57	0.28	0.32	0.29				
COST/CAPACITY	-1.83	0.40	-1.08	0.37				
DURATION	0.45	0.25	0.26	0.18				
ELECTRICITY	3.20	0.28	1.86	0.18				
GOVERNOR	**-2.28**	**0.02**	**-1.32**	**0.01**	**-1.27**	**0.03**	**-0.76**	**0.02**
INCOME	0.81	0.87	0.43	0.86				
GI(JOBS)	-4.70	0.38	-2.74	0.31				
LFP	-8.11	0.41	-4.69	0.36				
NG PRICE	-0.22	0.92	0.00	1.00				
GI(POLLUTION)	-0.54	0.37	-0.34	0.32				
POPULATION	1.32	0.14	0.74	0.10				
RAIL/POPULATION	2.20	0.86	1.12	0.85				
SOLAR PRICE /WIND PRICE	*-54.47*	*0.08*	*-31.10*	*0.07*	**-20.32**	**0.03**	**-11.80**	**0.02**
CAPACITY (BBL/DAY)	0.06	0.88	0.04	0.85				
UR	-3.91	0.12	-2.21	0.12				
WIND PRICE	1.83	0.49	0.86	0.54				
CONSTANT	42.34	0.28	25.55	0.21	*8.60*	*0.06*	**4.98**	**0.04**
Information criteria and AUC								
AIC		109.51		108.99		87.44		**87.07**
SIC		157.84		157.31		95.07		**94.70**
AUC		79.83%		80.21%		70.51%		70.51%
Residual tests								
Ljung-Box(50) res.[p-val]		0.42		0.62		0.20		0.47
Ljung-Box(50) res.sq.[p-val]		0.21		0.91		0.06		0.44
BDS(dim=2) [p-val]		0.41		0.53		**0.00**		0.85
BDS(dim=6) [p-val]		0.54		0.95		0.07		0.93

Table 9: Coal-to-Liquids Plants: Model Estimation Results. Smallest information criteria and p-values smaller than 5 % are reported in bold font.

In case of coal power plants, as expected, the longer the planning period the higher is the probability that the project will be abandoned/cancelled: expensive legal battles, cost escalations due to project modifications required to meet new regulations and/or legal orders, can easily erase the profitability of a new coal project. Moreover, the lower the price for natural gas and the lower the price for solar photovoltaics with respect to natural gas, the higher is the probability that the project will be abandoned. Probably, the most interesting result is that the higher the Google search volumes about coal plants and/or coal power, the higher is the probability the coal project will be abandoned. Therefore, an increasing number of people looking for information about coal plants on the web highlights a future growing opposition to coal projects. The only (partially) unexpected result is the significant negative coefficient for coal price, which indicates that a higher coal price will increase the probability that a coal plant will be fully developed. A potential explanation of this result could be the strong commercial relationships between coal mining companies and coal power companies, so that high coal prices can still be economically viable. Given the very sketchy information about the business structure of the companies involved in coal projects (particularly for abandoned projects), we leave this issue as an interesting avenue of further research.

As for coal-to-liquids plants, only two regressors were found to be significant at the 10% level: the political affiliation of the state governor and the ratio between solar and wind prices. A Republican governor will increase the likelihood that the coal plant will be build, whereas a lower price for solar photovoltaics with respect to wind price will increase the probability that the project will be abandoned. Given the greater technical complexity and the higher costs of coal-to-liquids plants (see Höök et al. (2014)), the importance of the governor political affiliation is not a surprise: a coal-to-liquids project will have a probability to succeed only with a strong political support at the level of the local state government, otherwise it will be better not to proceed further. The population size and the unemployment rate had coefficients whose significance level is almost close to 10% and with the expected signs: a higher population increases the probability of project failure, while the reverse is true in case of higher unemployment rate. However, the restricted models with also these variables included did not reject the null hypothesis that their coefficients were zero (with high p-values) and the information criteria were higher, so that we did not report them.

In general, probit models fared better than logit models, showing lower information criteria and better residuals properties. Restricted models showed lower information criteria, but full models had higher AUC values.

3.2 Out-of-Sample Forecasting Analysis

To better evaluate the predictive performance for each model, we also implemented a cross-validation procedure. We divided our dataset into two parts of equal size: the first one was used as the training set, while the second one as the validation set. Moreover, similarly to what performed by Young et al. (2011), we compare a set of alternative models whose characteristics are reported in Table 10. We considered both logit an probit models, models with all the regressors, as well as restricted models with only significant parameters at the 5% level; models with data transformed in logs and models with data in levels without any transformation; models without Google indexes and models with only Google indexes. In case of coal-to-liquids plants, due to the very small sample size of the training and validation sets, we only considered restricted models with Google indexes only and without Google indexes.

The estimated AUC for all previous models are reported in Table 11. The restricted probit model with data in logs was the best for coal power plants projects, while the probit model with data in logs and no Google indexes was the best for coal-to-liquids projects, thus confirming previous in-sample results.

We then employed the MCS approach developed by Hansen et al. (2011) and discussed in Section 2.3 to test for statistically significant differences in the forecast performances among the competing models. We remark that the MCS procedure will yield a set containing the best forecasting models at a given confidence level, see also Fantazzini and Fomichev

COAL POWER PLANTS				COAL-TO-LIQUIDS PLANTS			
Model	*Data transformation*	*All regressors/ restricted model*	*Google data*	*Model*	*Data transformation*	*All regressors/ restricted model*	*Google data*
LOGIT	log	All	YES	LOGIT	log	restricted	NO
PROBIT	log	All	YES	PROBIT	log	restricted	NO
LOGIT	log	restricted	NO	LOGIT	log	restricted	ONLY
PROBIT	log	restricted	NO	PROBIT	log	restricted	ONLY
LOGIT	log	restricted	NO	LOGIT	levels	restricted	NO
PROBIT	log	restricted	NO	PROBIT	levels	restricted	NO
LOGIT	log	restricted	YES	LOGIT	levels	restricted	ONLY
PROBIT	log	restricted	YES	PROBIT	levels	restricted	ONLY
LOGIT	log	restricted	ONLY				
PROBIT	log	restricted	ONLY				
LOGIT	levels	All	YES				
PROBIT	levels	All	YES				
LOGIT	levels	restricted	YES				
PROBIT	levels	restricted	YES				
LOGIT	levels	restricted	NO				
PROBIT	levels	restricted	NO				
LOGIT	levels	restricted	NO				
PROBIT	levels	restricted	NO				
LOGIT	levels	restricted	ONLY				
PROBIT	levels	restricted	ONLY				

Table 10: List of forecasting models

Models: COAL POWER PLANTS	**AUC**	**Models: COAL-TO-LIQUIDS**	**AUC**
Logit log	59.48%	Logit log (no Google)	60.74%
Probit log	60.13%	**Probit log (no Google)**	**61.85%**
Logit log (no Google)	57.14%	Logit log (only Google)	52.04%
Probit log (no Google)	57.74%	Probit log (only Google)	52.78%
Logit log (no Google) restricted	58.23%	Logit levels (no Google)	60.37%
Probit log (no Google) restricted	58.25%	Probit levels (no Google)	59.07%
Logit log restricted	63.91%	Logit levels (only Google)	53.89%
Probit log restricted	**64.13%**	Probit levels (only Google)	55.74%
Logit log (only Google)	48.18%		
Probit log (only Google)	48.08%		
Logit levels	60.13%		
Probit levels	60.00%		
Logit levels restricted	62.35%		
Probit levels restricted	63.87%		
Logit levels (no Google)	57.46%		
Probit levels (no Google)	58.01%		
Logit levels (no Google) restricted	60.17%		
Probit levels (no Google) restricted	60.17%		
Logit levels (only Google)	48.11%		
Probit levels (only Google)	48.33%		

Table 11: A.U.C. for each forecasting model. The best model is reported in bond font.

(2014) and Rossi and Fantazzini (2014) for recent applications in financial forecasting. The p-values for the test statistics were obtained by using the stationary block bootstrap with a block length of 2 years and 10000 re-samples: if the p-value was lower than a defined threshold level α, the model was not included in the MCS and viceversa. We set $\alpha = 0.10$ as in Hansen et al. (2011). The results of the MCS procedure are reported in Table 12. In case of coal power plants, the restricted probit model with data in logs is the model with the lowest loss for almost all loss functions considered, thus confirming the previous results. Moreover, models with Google data represent the majority of models included in the MCS at the 10% level, while models without Google data are seldom included, thus confirming the important information that this type of data can provide. As for coal-to-liquids plants, the logit models with data in levels without Google data is the one that has the lowest loss across a spectrum of loss functions. However, almost all models are now included in the MCS, which highlights that the validation set is not very informative (which was expected given its small size).

4 Robustness checks

We wanted to verify that our previous results hold also with alternative data setups. Therefore, we performed a series of robustness checks: we considered alternative keywords

COAL	***LOGARITHMIC***			***SPHERICAL***			***MAD***			***ASYMMETRIC QUADRATIC***														
										c=0.1			c=0.25			c=0.5			c=0.75			c=0.9		
																(mse)								
Model	loss	p-val.		loss	p-val.		loss	p-val.		loss	p-val.		loss	p-val.		loss	p-val.		loss	p-val.		loss	p-val.	
Logit log	0.515	0.01		0.347	0.05		0.264	0.08	*	0.547	0.00		0.347	0.81	*	0.149	0.09		0.157	0.09		0.160	0.09	
Probit log	0.537	0.01		0.345	1.00	*	0.262	1.00	*	0.544	0.00		0.344	0.88	*	0.148	0.09		0.156	0.09		0.159	0.09	
Logit log NO GI	0.527	0.00		0.350	0.02		0.264	0.06		0.596	0.00		0.352	0.29	*	0.149	0.09		0.157	0.09		0.159	0.09	
Probit log NO GI	0.558	0.00		0.347	0.60	*	0.262	0.94	*	0.596	0.00		0.349	0.81	*	0.148	0.09		0.156	0.09		0.159	0.09	
Logit log NO GI res	0.464	0.01		0.376	0.00		0.276	0.00		0.451	0.00		0.333	0.88	*	0.145	0.09		0.154	0.09		0.157	0.09	
Probit log NO GI res	0.464	0.01		0.375	0.00		0.276	0.00		0.452	0.00		0.333	0.88	*	0.145	0.09		0.154	0.09		0.157	0.09	
Logit log res	0.449	0.75	*	0.366	0.00		0.270	0.06		0.436	0.02		0.317	0.88	*	0.141	0.39	*	0.151	0.39	*	0.155	0.39	*
Probit log res	0.448	1.00	*	0.364	0.00		0.269	0.06		0.434	0.21	*	0.315	1.00	*	0.141	1.00	*	0.151	1.00	*	0.155	1.00	*
Logit log only GI	0.462	0.01		0.387	0.00		0.281	0.00		0.433	0.22	*	0.348	0.29	*	0.144	0.09		0.153	0.09		0.157	0.09	
Probit log only GI	0.461	0.47	*	0.386	0.00		0.281	0.00		0.432	0.93	*	0.347	0.81	*	0.144	0.23	*	0.153	0.23	*	0.156	0.23	*
Logit lev	0.472	0.01		0.365	0.00		0.274	0.00		0.439	0.01		0.338	0.88	*	0.150	0.09		0.157	0.09		0.160	0.09	
Probit lev	0.471	0.01		0.364	0.00		0.274	0.00		0.444	0.00		0.338	0.88	*	0.149	0.09		0.157	0.09		0.159	0.09	
Logit lev res	0.450	0.75	*	0.365	0.00		0.270	0.04		0.428	0.93	*	0.320	0.88	*	0.142	0.23	*	0.152	0.23	*	0.155	0.23	*
Probit lev res	0.448	0.83	*	0.363	0.00		0.269	0.06		0.428	1.00	*	0.318	0.88	*	0.142	0.39	*	0.152	0.39	*	0.155	0.39	*
Logit lev NO GI	0.474	0.01		0.366	0.00		0.272	0.00		0.491	0.00		0.341	0.88	*	0.147	0.09		0.156	0.09		0.158	0.09	
Probit lev NO GI	0.474	0.01		0.365	0.00		0.272	0.00		0.494	0.00		0.340	0.88	*	0.147	0.09		0.155	0.09		0.158	0.09	
Logit lev NO res	0.455	0.47	*	0.373	0.00		0.273	0.00		0.447	0.00		0.327	0.88	*	0.142	0.23	*	0.152	0.23	*	0.155	0.23	*
Probit lev NO res	0.454	0.75	*	0.373	0.00		0.273	0.00		0.447	0.00		0.326	0.88	*	0.142	0.39	*	0.152	0.39	*	0.155	0.39	*
Logit lev only GI	0.467	0.01		0.390	0.00		0.283	0.00		0.447	0.00		0.353	0.29	*	0.145	0.09		0.154	0.09		0.157	0.09	
Probit lev only GI	0.467	0.01		0.390	0.00		0.283	0.00		0.446	0.00		0.353	0.29	*	0.145	0.09		0.154	0.09		0.157	0.09	

CTL	***LOGARITHMIC***			***SPHERICAL***			***MAD***			***ASYMMETRIC QUADRATIC***														
										c=0.1			c=0.25			c=0.5			c=0.75			c=0.9		
																(mse)								
Model	loss	p-val.		loss	p-val.		loss	p-val.		loss	p-val.		loss	p-val.		loss	p-val.		loss	p-val.		loss	p-val.	
Logit log NO GI	3.203	0.03		0.405	0.76	*	0.360	0.22	*	1.152	0.26	*	0.577	0.45	*	0.289	0.05		0.473	0.01		0.762	0.06	
Probit log NO GI	4.034	0.03		0.396	0.76	*	0.343	0.52	*	1.145	0.28	*	0.560	0.88	*	0.260	0.23	*	0.366	0.02		0.707	0.06	
Logit log only GI	0.949	0.05		0.389	0.98	*	0.342	0.25	*	1.378	0.26	*	0.640	0.22	*	0.265	0.10	*	0.267	0.02		0.267	0.09	
Probit log only GI	0.960	0.05		0.388	0.98	*	0.339	0.52	*	1.375	0.26	*	0.634	0.45	*	0.261	0.23	*	0.264	0.07		0.265	0.23	*
Logit lev NO GI	1.651	0.05		0.359	1.00	*	0.300	1.00	*	1.086	0.32	*	0.528	1.00	*	0.213	1.00	*	0.232	1.00	*	0.231	1.00	*
Probit lev NO GI	2.752	0.05		0.361	0.98	*	0.306	0.52	*	1.095	0.32	*	0.534	0.88	*	0.223	0.23	*	0.248	0.07		0.234	0.57	*
Logit lev only GI	0.825	0.49	*	0.390	0.76	*	0.337	0.52	*	0.961	1.00	*	0.630	0.31	*	0.254	0.23	*	0.260	0.07		0.261	0.23	*
Probit lev only GI	0.819	1.00	*	0.387	0.98	*	0.333	0.52	*	0.987	0.32	*	0.620	0.45	*	0.250	0.23	*	0.257	0.07		0.259	0.32	*

Table 12: Model Confidence Set results for Coal Power Plants (upper table) and coal-to-liquids (lower table). Different loss functions are evaluated: logarithmic, spherical, absolute deviation (MAD) and asymmetric quadratic, where the last is computed with different values of the parameter c. The case $c = 0.5$ corresponds to square errors (MSE). The reported loss value should be multiplied by 10^3. * indicates the model is included in the MCS at 10% confidence level.

for Google search, and we evaluated the effect on our estimates of the global financial crisis in 2008 and 2009.

4.1 Alternative Keywords

One of the regressors in our analysis was the Google index for the search term "pollution." While this keyword is very general and should include all possible searches related to environmental hazards, it may be well be too way general and not related to coal plants: for example, two of the top rising searches for this term in the US were "*pollution in china*" and "*china pollution*." In this regard, Google Trends provides also the search trends for specific categories, which include all searches related to the chosen category according to some internal selection algorithms. The closest category related to pollution and environmental hazards is *Business and Industrial / Energy and utilities / Waste Management*. Similarly, we also downloaded the GI related to the keywords "*coal power+coal plant*" and "*coal-to-liquids+ctl coal*," but restricted to the category *Business and Industrial / Energy and utilities*. The estimated coefficients for the models including these two alternative GIs in the place of the initial ones are reported in Tables 13-14 for coal power plants and coal-to-liquids plants, respectively.

	LOGIT		**PROBIT**	
	Coef.	*P-value*	*Coef.*	*P-value*
CO2/CAPACITY	1.01	0.19	0.52	0.24
GI(COAL)	0.04	0.82	0.03	0.81
COAL PRICE	-0.18	0.78	-0.12	0.71
GI(COAL PLANT + COAL POWER)	*2.02*	*0.06*	*0.95*	*0.05*
COST/CAPACITY	-0.64	0.53	-0.40	0.45
DURATION	*0.21*	*0.07*	*0.12*	*0.05*
ELECTRICITY	0.09	0.91	0.06	0.88
GOVERNOR	-0.31	0.24	-0.18	0.20
INCOME	1.55	0.34	0.93	0.32
GI(JOBS)	1.65	0.13	1.01	0.12
LFP	-3.33	0.37	-2.07	0.33
NG PRICE	**-3.69**	**0.03**	**-1.87**	**0.02**
GI(WASTE)	-0.49	0.66	-0.26	0.67
POPULATION	-0.24	0.20	-0.14	0.19
RAIL/POPULATION	-4.01	0.17	-2.12	0.18
SOLAR PRICE/NG PRICE	**-4.63**	**0.03**	**-2.35**	**0.03**
CAPACITY (MW)	*0.56*	*0.05*	*0.30*	*0.06*
UR	*-1.28*	*0.06*	**-0.78**	**0.04**
WIND PRICE	0.91	0.29	0.52	0.27
CONSTANT	-8.11	0.57	-4.50	0.57
Information criteria and AUC				
AIC		528.377		527.580
SIC		615.430		614.632
AUC		68.41%		68.54%
Residual tests				
Ljung-Box(50) res.[p-val]		0.13		0.12
Ljung-Box(50) res.sq.[p-val]		**0.01**		**0.01**
BDS(dim=2) [p-val]		0.15		0.15
BDS(dim=6) [p-val]		0.69		0.69

Table 13: Coal Power Plants: Model Estimation Results with alternative Google Indexes. P-values smaller than 5 % are reported in bold font.

In case of coal power plants, the results do not change much in terms of signs and significance with respect to the baseline case in Table 8: the GI for "waste management" is not significant like the GI for "pollution" was in the baseline case, whereas the restricted GI for the related to the keywords "*coal power+coal plant*" is now significant only at the 10% level for the logit model. All parameters that were statistically different from zero in the baseline case keep on being significantly different from zero in this setup and with the same signs, even though in some cases only at the 10% level, like for the *duration* indicator. The only major difference is that now we do not reject that the *coal price* has a zero coefficient with p-values higher than 70%. The information criteria (AIC and SIC) are higher than in the baseline case while the AUCs are lower. Moreover, the residuals

	LOGIT		***PROBIT***	
	Coef.	*P-value*	*Coef.*	*P-value*
GI(COAL)	2.26	0.11	1.45	0.08
COAL PRICE	2.07	0.36	1.06	0.33
GI(COAL-TO-LIQUIDS+CTL COAL)/NG PRICE	**2.02**	**0.03**	**1.10**	**0.01**
COST/CAPACITY	-1.36	0.58	-0.76	0.55
DURATION	0.40	0.35	0.24	0.23
ELECTRICITY	3.99	0.30	1.93	0.23
GOVERNOR	**-3.08**	**0.01**	**-1.74**	**0.00**
INCOME	0.47	0.94	-0.10	0.97
GI(JOBS)	*-11.54*	*0.09*	*-6.50*	*0.04*
LFP	-7.37	0.51	-3.46	0.53
NG PRICE	-6.92	0.13	*-3.60*	*0.09*
GI(WASTE)	-0.12	0.97	0.00	1.00
POPULATION	*1.87*	*0.09*	*1.00*	*0.06*
RAIL/POPULATION	-2.54	0.83	-2.57	0.65
SOLAR PRICE /WIND PRICE	**-103.99**	**0.01**	**-57.39**	**0.00**
CAPACITY (BBL/DAY)	-0.14	0.79	-0.05	0.87
UR	**-7.61**	**0.03**	**-4.13**	**0.01**
WIND PRICE	3.26	0.27	1.67	0.26
CONSTANT	96.16	0.04	56.34	0.03
Information criteria and AUC				
AIC		105.88		105.78
SIC		154.20		154.10
AUC		81.67%		82.43%
Residual tests				
Ljung-Box(50) res.[p-val]		0.75		0.62
Ljung-Box(50) res.sq.[p-val]		1.00		0.99
BDS(dim=2) [p-val]		0.48		0.52
BDS(dim=6) [p-val]		0.98		0.98

Table 14: Coal-To-Liquids Plants: Model Estimation Results with alternative Google Indexes. P-values smaller than 5 % are reported in bold font.

tests highlights some small misspecification in the squared residuals. Therefore, in general, this robustness check confirms the previous results but with a worse fit than the baseline case.

In case of coal-to-liquids plants, the main findings of the baseline case are also confirmed, but there are now some interesting differences. The indicator *governor* keeps on being a strong significant variable (now even at the 1% level) and with the same sign as in the baseline case. Similarly, lower solar prices -with respect to wind prices- will increase the probability that the project will be abandoned. Moreover, the GI for "waste management" is not significantly different from zero, as it was the case for the GI for the keyword "pollution." However, the ratio of the GI for the keywords "*coal-to-liquids+ctl coal*" and the natural gas price is now significant at the 5% level with a positive coefficient: the more people look for information about coal-to-liquids plants -with respect to natural gas prices-, the higher the probability the project will be abandoned. Moreover, differently from the baseline case in Table 9, the lower the unemployment rate and the lower the number people looking for "jobs," the higher is the probability that the coal project will be abandoned/canceled. Furthermore, an higher population will decrease the odds that the coal plant will be built. The information criteria are now slightly lower, the AUCs are slightly higher, while the residuals tests do not highlight any particular misspecification. In general, restricting the selection criteria for Google data seems to be beneficial for the analysis of coal-to-liquids plants by eliminating too many unrelated searches and highlighting additional significant factors beyond the political affiliation of the state governor and renewable prices, which still remain the most important factors[11].

[11] Eliminating these variables results in steep increases of information criteria and a worse AUC, much more than the other regressors.

4.2 The Recession in the Years 2008-2009

The second robustness check was to evaluate the effect on our estimates of the global financial crisis in 2008 and 2009. Given the small temporal dimension of our dataset, we used a dummy variable for the years 2008 and 2009, in correspondence to the official NBER recession for the US. The estimated models including this dummy variable are reported in Tables 15-16 for coal power plants and coal-to-liquids plants, respectively.

	LOGIT		**PROBIT**	
	Coef.	*P-value*	*Coef.*	*P-value*
CO2/CAPACITY	0.93	0.22	0.47	0.28
GI(COAL)	0.09	0.63	0.06	0.57
COAL PRICE	*-1.57*	*0.07*	*-0.88*	*0.07*
GI(COAL PLANT + COAL POWER)	**11.79**	**0.01**	**6.47**	**0.01**
COST/CAPACITY	-0.55	0.59	-0.29	0.57
DURATION	*0.21*	*0.06*	**0.12**	**0.04**
ELECTRICITY	-0.02	0.98	-0.02	0.96
GOVERNOR	-0.30	0.26	-0.17	0.23
INCOME	1.35	0.41	0.83	0.37
GI(JOBS)	1.20	0.32	0.71	0.30
LFP	-3.24	0.36	-1.96	0.33
NG PRICE	**-9.75**	**0.00**	**-5.39**	**0.00**
GI(POLLUTION)	0.02	0.96	0.02	0.91
POPULATION	-0.26	0.13	-0.14	0.14
RAIL/POPULATION	-4.17	0.16	-2.24	0.16
SOLAR PRICE/NG PRICE	**-14.25**	**0.00**	**-7.84**	**0.00**
CAPACITY (MW)	*0.51*	*0.08*	0.26	0.10
UR	*-1.25*	*0.08*	*-0.73*	*0.06*
WIND PRICE	0.01	0.99	-0.10	0.86
DUMMY(2008-2009)	0.35	0.65	0.23	0.62
CONSTANT	-11.20	0.52	-5.91	0.56
Information criteria and AUC				
AIC		518.34		517.22
SIC		609.74		608.63
AUC		71.18%		71.19%
Residual tests				
Ljung-Box(50) res.[p-val]		0.21		0.17
Ljung-Box(50) res.sq.[p-val]		0.28		0.33
BDS(dim=2) [p-val]		0.08		0.08
BDS(dim=6) [p-val]		0.51		0.53

Table 15: Coal Power Plants: Model Estimation Results with a dummy variables for the recession in 2008-2009. P-values smaller than 5 % are reported in bold font.

	LOGIT		**PROBIT**	
	Coef.	*P-value*	*Coef.*	*P-value*
GI(COAL)	1.54	0.30	1.01	0.24
COAL PRICE	*6.43*	*0.08*	*3.43*	*0.06*
GI(COAL-TO-LIQUIDS+CTL COAL)/NG PRICE	*1.62*	*0.07*	**0.87**	**0.04**
COST/CAPACITY	-0.86	0.74	-0.49	0.71
DURATION	0.41	0.37	0.25	0.23
ELECTRICITY	5.97	0.13	3.10	0.07
GOVERNOR	**-2.91**	**0.02**	**-1.62**	**0.01**
INCOME	2.25	0.74	1.03	0.74
GI(JOBS)	-12.16	0.10	**-6.70**	**0.04**
LFP	-13.09	0.25	-6.96	0.21
NG PRICE	-6.64	0.18	-3.34	0.15
GI(POLLUTION)	-1.01	0.07	-0.60	0.07
POPULATION	**2.02**	**0.04**	**1.10**	**0.02**
RAIL/POPULATION	0.78	0.95	-0.36	0.95
SOLAR PRICE /WIND PRICE	**-84.36**	**0.04**	**-46.35**	**0.01**
CAPACITY (BBL/DAY)	-0.26	0.61	-0.12	0.66
UR	**-8.83**	**0.02**	**-4.76**	**0.01**
WIND PRICE	**12.53**	**0.04**	**6.65**	**0.02**
DUMMY(2008-2009)	**-4.65**	**0.03**	**-2.49**	**0.02**
CONSTANT	38.50	0.33	23.70	0.29
Information criteria and AUC				
AIC		106.16		105.98
SIC		157.02		156.85
AUC		83.12%		82.66%
Residual tests				
Ljung-Box(50) res.[p-val]		0.66		0.52
Ljung-Box(50) res.sq.[p-val]		1.00		0.98
BDS(dim=2) [p-val]		0.46		0.46
BDS(dim=6) [p-val]		0.95		0.98

Table 16: Coal-To-Liquids Plants: Model Estimation Results with a dummy variables for the recession in 2008-2009. P-values smaller than 5 % are reported in bold font.

In case of coal power plants, the dummy variable is not statistically different from zero across all model specifications, and the coefficients of all other parameters are very close to the baseline case reported in Table 8. Instead, the results for coal-to-liquids projects are partially different: the coefficients for the political affiliation of the state governor and for the ratio between solar and wind prices are again significantly different from zero, with the same signs and similar magnitudes as reported in Table 9 for the baseline case. However, other variables are now significant at the 5% level and 10% level: the lower the unemployment rate and the lower is the number of people looking for "jobs", the higher is the probability that the coal project will be abandoned/canceled. Moreover, the higher is the number of people looking for "*coal-to-liquids* or "*ctl coal*" in Google -with respect to natural gas prices- and the higher is the coal price, the higher is the probability of project failure. Besides, the higher is the population, the higher is the probability that the project will be abandoned. In this regard, all these indicators have signs that conform to the past literature, and they are very similar to those found restricting the GIs in the previous section. The wind price has a significant positive coefficient, which means that higher wind prices increase the probability of project failure. This result may seem unexpected at a first glance: however, given the strong correlation between solar and wind price, it has to be examined together with the solar/wind ratio, where an increase in wind prices strongly decreases the probability of project failure. Probably, the most interesting new result is the significant negative coefficient for the dummy variable for the years 2008 and 2009: those two years witnessed very high prices for premium oil liquids (which represent the main output of a coal-to-liquids plant), and this fact may have sparked a strong interest in this type of coal plants. However, given that these results are very close to those in Table 14 using restricted Google Indexes, we also estimated a model including both the dummy variable for 2008 and 2009 and the alternative Google search data (see Table 17).

	LOGIT		***PROBIT***	
	Coef.	*P-value*	*Coef.*	*P-value*
GI(COAL)	2.06	0.17	1.37	0.10
COAL PRICE	3.14	0.33	1.58	0.34
GI(COAL-TO-LIQUIDS+CTL)/NG PRICE	**1.92**	**0.04**	**1.05**	**0.02**
COST/CAPACITY	-1.09	0.66	-0.64	0.62
DURATION	0.39	0.37	0.24	0.24
ELECTRICITY	4.13	0.29	1.96	0.23
GOVERNOR	**-2.99**	**0.01**	**-1.72**	**0.00**
INCOME	0.26	0.97	-0.29	0.92
GI(JOBS)	*-11.81*	*0.09*	**-6.62**	**0.04**
LFP	-6.61	0.54	-3.00	0.58
NG PRICE	-7.46	0.13	-3.88	0.10
GI(POLLUTION)	-0.63	0.86	-0.17	0.93
POPULATION	*1.93*	*0.09*	*1.02*	*0.05*
RAIL/POPULATION	-2.97	0.79	-2.87	0.60
SOLAR PRICE /WIND PRICE	**-97.66**	**0.02**	**-54.38**	**0.01**
CAPACITY (BBL/DAY)	-0.23	0.69	-0.08	0.78
UR	**-7.90**	**0.03**	**-4.28**	**0.01**
WIND PRICE	4.53	0.25	2.33	0.26
DUMMY(2008-2009)	-0.84	0.59	-0.40	0.63
CONSTANT	86.97	0.10	*52.38*	*0.07*
Information criteria and AUC				
AIC		107.58		107.55
SIC		158.44		158.4164
AUC		82.05%		82.20%
Residual tests				
Ljung-Box(50) res.[p-val]		0.75		0.60
Ljung-Box(50) res.sq.[p-val]		1.00		1.00
BDS(dim=2) [p-val]		0.57		0.57
BDS(dim=6) [p-val]		1.00		0.98

Table 17: Coal-To-Liquids Plants: Model Estimation Results with a dummy variables for the recession in 2008-2009 and with alternative Google Indexes. P-values smaller than 5 % are reported in bold font.

Once the restricted Google data are included, the coefficient for the dummy variable is no more significant across all model specifications, whereas all other results remain basically the same. Therefore, the global financial crisis seems not to have influenced significantly the fate of coal-to-liquids projects, similarly to what we found for coal power plants.

5 Conclusions

The construction of new coal plants has become an issue of great relevance in the US, given the large fleet of old coal power plants that should be replaced. Besides, over the past 30 years the overall US coal consumption displayed a (weak) upward trend. Moreover, even though the European consumption has constantly decreased (principally during the '90s), Asian coal demand rapidly boosted and more than compensated this decrease[12], thus confirming a continued interest in this source of energy worldwide. Furthermore, the recent price tensions in oil and natural gas markets renewed the interest in the use of coal for electricity production. On the other hand, the increasing environmental awareness about the hazards posed by coal plants has lead to an increase in the opposition by local communities against the installation of coal plants in their region.

In this context, the analysis of the determinants that influence the success or failure of coal plants projects may be relevant for both energy policy making and project planning. In this paper we analyzed 145 coal plants and 25 coal-to-liquid plants that have been proposed in US in the period 2004-2013 and we investigated the decision to settle the plant or abandon the project using several variables. Beside common industrial explanatory variables (size, input, output and labour costs, substitute costs, infrastructure), we also considered measures of social and environmental awareness using Google search data. After controlling for collinearity, stationarity and robustness, we performed an extensive model specification, comparison and selection.

In case of coal power plants, we found that the project duration, the prices of energy substitutes for electricity generation, the coal price and the awareness about the coal projects and its hazards are the main factors. More specifically, the longer the planning period the less likely the project will be implemented: expensive legal disputes, and costly project modifications to meet new requirements, can vanish the plant profitability. Moreover, the lower the price for natural gas and the lower the price for solar photovoltaics with respect to natural gas price, the higher is the probability that the project will be abandoned. Interestingly, we found that the higher the coal price, the higher the probability that a coal plant will be built. This result is partially puzzling. A possible explanation may rely on the strong commercial relationships between coal mining companies and coal power companies, so that the increase in coal prices will not weaken the interest toward the coal plant. Finally, the awareness by local communities as measured by the Google search volumes about coal plants and/or coal power plants increases the probability that the project will be abandoned.

As for coal-to-liquids plants, we found that the state governor's political affiliation, the ratio between solar and wind prices, the population size, the unemployment rate and the job searches as measured by Google data are the main drivers (however, the latter three are only weakly significant). Particularly, coal-to-liquids plants are more likely to be completed in conservative states, where we can presume that there is stronger political support for heavy industry projects. Besides, the lower price of solar photovoltaics with respect to wind price, the higher the probability that the project will be abandoned. Larger

[12] In 2012, China is by far the first coal consumer (49% of World consumption), followed by US (11%) and India (9%) (source `www.eia.gov`).

populations make these projects less likely, as expected, while higher unemployment rates and job searches increase the probability of successful implementation.
Even though we considered a large set of alternative model specifications, we had to restrict the potential range of models to keep the empirical analysis computationally tractable. An avenue of future research would be twofold: on one hand we would integrate additional web-based data, exploring the recent social network data; on the other hand we would consider additional models like Bayesian models, non-parametric methods and generalizations of the standard logit-probit model.

References

A., D., 2002. Introduction to Generalised Linear Model. Chapman and Hall.

Anderton, D., Anderson, A., Oakes, J., Fraser, M., 1994. Environmental equity: the demographics of dumping. Demography 31 (2), 171–192.

Ansolabehere, S., Konisky, D., 2009. Public attitudes toward construction of new power plants. Public Opinion Quarterly doi: 10.1093/poq/nfp041.

Arora, S., Cason, T., 1998. Do community characteristics influence environmental outcomes?: Evidence from the toxics release inventory. Journal of Applied Economics 1 (2), 413–453.

Askitas, N., Zimmermann, K., 2009. Google econometrics and unemployment forecasting. Applied Economics Quarterly 55, 107–120.

Baltagi, B., 1961. Econometric Analysis of Panel Data, 5th Edition. Wiley.

Been, V., Gupta, F., 1997. Coming to the nuisance or going to the barrios? A longitudinal analysis of environmental justice claims. Ecology Law Quarterly 24 (1), 1–56.

Boer, J., Pastor, M., Sadd, J., Snyder, L., 1997. Is there environmental racism? The demographics of hazardous waste in los angeles county. Social Science Quarterly 78 (4), 793–809.

Buckland, S.T. Burnham, K., Augustin, N., 1997. Model selection: An integral part of inference. Biometrics 53, 603–618.

Cameron, C., Trivedi, P., 2005. Microeconometrics: Methods and Applications. Cambridge University Press.

Carney, P., Sickles, E., Monsees, B., Bassett, L., Brenner, R., Feig, S., Smith, R., Rosemberg, R., Bogart, T., Browning, S., Barry, J., Kelly, M., Tran, K., Miglioretti, D., 2010. Identifying minimally acceptable interpretative performance criteria for screening mammography. Radiology 255 (2), 354–361.

Carrion-i Silvestre, J., Del Barrio, T., Lopez-Bazo, E., 2005. Breaking the panels. An application to the GDP per capita. Econometrics Journal 8, 159–175.

Choi, H., Varian, H., 06 2009. Predicting the present with Google trends. Working paper, Google Inc.

Choi, I., 2001. Unit root tests for panel data. Journal of International Money and Finance 20, 249–272.

Cleetus, R., Clemmer, S., Davis, E., Deyette, J., Downing, J., Frenkel, S., 2012. Ripe for retirement: The case for closing America's costliest coal plants. Working paper, Cambridge, MA: Union Of Concerned Scientists.

Club, S., 2014. Proposed coal plant tracker. http://content.sierraclub.org/coal/environmentallaw/plant-tracker.

Da, Z., Engelberg, J., Pengjie, G., 2011. In search of attention. Journal of Finance 5, 1461–1499.

D'Amuri, F., Marcucci, J., 11 2013. The predictive power of Google searches in forecasting unemployment. Temi di discussione (Economic working papers) 891, Bank of Italy.

Danielsson, J., 2011. Financial Risk Forecasting. Wiley Finance.

Diebold, F., 2006. Elements of Forecasting, 4th Edition. Cengage Learning.

Dormann, C., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., García Marquéz, J., Gruber, B., Lafourcade, B., Leitao, P., Münkemüller, T., McClean, C., Osborne, P., Reineking, B., Schröder, B., Skidmore, A., Zurell, D., Lautenbach, S., 2013. Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. Ecography 36, 27–46.

EIA, 2014a. Clean coal research – US department of energy. http://energy.gov/fe/science-innovation/clean-coal-researc

EIA, 2014b. EIA Annual Energy Outlook 2014. http://www.eia.gov/forecasts/aeo/.

EIA, 2014c. EIA Electric Power Monthly – May 2014. http://www.eia.gov/electricity/monthly/.

EIA, 2014d. EIA Natural Gas Monthly – May 2014. http://www.eia.gov/naturalgas/monthly/.

Fantazzini, D., De Giuli, M., Figini, S., Giudici, P., 2009. Enhanced credit default models for heterogeneous SME segments. Journal of Financial Transformation 25, 31–39.

Fantazzini, D., Fomichev, N., 2014. Forecasting the real price of oil using online search data. International Journal of Computational Economics and Econometrics 4 (1-2), 4–31.

Fantazzini, D., Höök, M., Angelantoni, A., 2011. Global oil risks in the early 21st century. Energy Policy 39 (12), 7865–7873.

Figini, S., Maggi, M., 2014. Performance of credit risk prediction models via proper loss functions. Working Paper 64, Department of Economics and Management, University of Pavia.

Fleischman, L., Cleetus, R., Clemmer, S., Deyette, J., Frenkel, S., 2013. Ripe for retirement: An economic analysis of the U.S. coal fleet. The Electricity Journal 26 (10), 51–63.

for Media, C. C., Democracy, 2014. Proposed coal plants in the United States. http://www.sourcewatch.org/index.php/Category:Proposed_coal_plants_in_the_United_States.

Freese, B., Clemmer, S., Martinez, C., Nogee, A., March 2011. A risky proposition – The financial hazards of new investments in coal plants. Working paper, Union Of Concerned Scientists.

Garrone, P., A., G., 2012. Siting locally-unwanted facilities: What can be learnt from the location of Italian power plants. Energy Policy 45, 176–186.

Ginsberg, J., Mohebbi, M., Patel, R., Brammer, L., Smolinski, M., Brilliant, L., 2009. Detecting influenza epidemics using search engine query data. Nature 457, 1012–1014.

Giudici, P., Figini, S., 2009. Applied Data Mining for Business and Industry, 2nd Edition. Wiley.

Goin, J., 1982. ROC curve estimation and hypothesis testing: Applications to breast cancer detection. The Journal of the Pattern Recognition Society 15, 263–269.

Hadri, K., 2000. Testing for stationarity in heterogeneous panel data. Econometric Journal 3, 148–161.

Hamilton, J., 1993. Politics and social costs: Estimating the impact of collective action on hazardous waste facilities. The RAND Journal of Economics 24 (1), 101–125.

Hamilton, J., 1995. Testing for environmental racism: prejudice, profits, political power? Journal of Policy Analysis and Management 14 (1), 107–132.

Hanley, A., McNeil, B., 1982. The meaning and use of the area under a receiver operating characteristics (ROC) curve. Diagnostic Radiology 143, 29–36.

Hansen, P., Lunde, A., Nason, J., 2011. The model confidence set. Econometrica 79 (2), 453–497.

Höök, M., Aleklett, K., 2014. A review on coal to liquid fuels and its coal consumption. International Journal of Energy Research 34 (10), 848–864.

Höök, M., Fantazzini, D., Angelantoni, A., Snowden, S., 2014. Hydrocarbon liquefaction: viability as a peak oil mitigation strategy. Philosophical Transactions of the Royal Society A forthcoming.

Im, K., Pesaran, M., Shin, Y., 2003. Testing for unit roots in heterogeneous panels. Journal of Econometrics 115, 53–74.

Jenkins, R., Maguire, K., Morgan, C., 2004. Host community compensation and municipal solid waste landfills. Land Economics 80 (4), 513–528.

Kohavi, R., Provost, F., 1998. Glossary of terms. Machine Learning 30 (2-3), 271–274.

Kreutz, T., Larson, E., Liu, G., Williams, R., 2008. Fischer-Tropsch fuels from coal and biomass. In: Proceedings of the 25th Annual International Pittsburgh Coal Conference. 29 September to 2 October, Pittsburgh, USA, available from: `http://web.mit.edu/mitei/docs/reports/kreutz-fischer-tropsch.pdf`.

Krzanowski, W., Hand, D., 2009. ROC curves for continuous data. CRC/Chapman and Hall.

Levin, A., Lin, C., Chu, C., 2002. Unit root tests in panel data: Asymptotic and finite-sample properties. Journal of Econometrics 108, 1–24.

Ljung, G., G., B., 1979. On a measure of lack of fit in time series models. Biometrika 66, 265–270.

Maddala, G., Wu, S., 1999. A comparative study of unit root tests with panel data and a new simple test. Oxford Bulletin of Economics and Statistics 61, 631–652.

Metz, C., Kronman, H., 1980. Statistical significance tests for binormal ROC curves. Journal of Mathematical Psychology 22, 218–243.

O'Brien, R., 2007. A caution regarding rules of thumb for variance inflation factors. Quality and Quantity 41 (5), 673–690.

of Energy, T. N. E. T. L. D., 2007. www.netl.doe.gov.

Pratson, L., Haerer, D., Patino-Echeverri, D., 2013. Fuel prices, emission standards, and generation costs for coal vs natural gas power plants. Environmental Science and Technology 47 (9), 4926–4933.

Provost, F., Fawcett, T., Kohavi, R., 1998. The case against accuracy estimantion for comparing induction algorithms. In: Shavlik, J. (Ed.), Proceedings of the Fifteenth International Conference on Machine Learning. Morgan Kaugfman, San Francisco, CA, pp. 445–453.

R., C., 1960. The problem of social cost. The Journal of Law and Economics 3, 1–23.

Rabe-Hesketh, S., Skrondal, A., 2004. Generalized Latent Variable Modeling. Chapman and Hall.

Rabe-Hesketh, S., Skrondal, A., 2005. Multilevel and Longitudinal Modeling Using Stata,. STATA press.

Rossi, E., Fantazzini, D., 2014. Long memory and periodicity in intraday volatility. Journal of Financial Econometrics forthcoming.

Suhoy, T., 2009. Query indices and a 2008 downturn. Discussion Paper 6, Bank of Israel.

Tierney, S., 2012. Why coal plants retire: Power market fundamentals as of 2012. `http://www.analysisgroup.com/uploadedFiles/News_and_Events/News/2012_Tierney_WhyCoalPlantsRetire.pdf`.

W., B., D., D., J., S., B., L., 1996. A test for independence based on the correlation dimension. Econometric Reviews 175 (3), 197–235.

W., G., 2011. Econometric Analysis, 7th Edition. Prentice Hall.

Winkler, R., 1994. Evaluating probabilities: Asymmetric scoring rules. Management Science 40 (11), 1395–1405.

Wolverton, A., 2009. Effects of socio-economic and input-related factors on polluting plants' location decisions. The B.E. Journal of Economic Analysis and Policy 9 (1), Article 14, available at: `http://www.bepress.com/bejeap/vol9/iss1/`.

Young, T., Zaretski, R., Perdue, J., Guess, F., X., L., 2011. Logistic regression models of factors influencing the location of bioenergy and biofuels plants. BioResources 6 (1), 329–343.

Appendix: Coal power plants and Coal-To-Liquids plants

US State	Plant name	US State	Plant name
AK	Cook Inlet Region Inc. Project	MS	Mississippi Power Kemper IGCC
AK	Healy Clean Coal Plant	MT	Highwood Generating Station
AK	Nuvista - Bethel Power Plant	MT	Nelson Creek (aka Circle)
AK	Western Arctic Coal Project	MT	Otter Creek (Bechtel / Kennecot Project)
AK	Kenai Blue Sky Project	MT	Roundup Power Project
AR	Plum Point I	NC	Cliffside Plant
AR	Hempstead (AEP)	ND	South Heart Coal
AR	Plum Point II	ND	Spiritwood Station
AZ	Springerville Generating Station Unit 3	ND	Gascoyne 175 Generating Station
AZ	Springerville Generating Station Unit 4	ND	Gascoyne 500 Generating Station
CA	Hydrogen Energy California	NE	OPPD's Nebraska City 2
CO	Buick Coal and Power Project	NJ	PurGen One
CO	Comanche Generating Station Unit 3	NJ	West Deptford Project
CO	Ray D. Nixon Power Plant	NM	Desert Rock
CO	Xcel Energy IGCC plant	NV	Ely Energy Center
CO	Lamar Light & Power/Arkansas River Power	NV	Toquop Energy Project
DE	Indian River Expansion (IGCC)	NV	White Pine Energy Station
FL	Orange County IGCC Plant	NV	Newmont coal plant
FL	Seminole 3	NY	Huntley (NRG)
FL	Taylor Energy Center - Alternative Proposal	NY	BPU Jamestown plant
FL	Polk Power Station 6	OH	American Municipal Power Generating Station
FL	Glades - Florida Power and Light	OH	Dominion Conneaut
GA	Ben Hill Plant	OH	Great Bend IGCC
GA	Longleaf Plant	OH	Lima Energy Station
GA	Washington County Power Station	OH	Irontron Energy Center
IA	Council Bluffs Energy Center Unit 4	OK	AES Shady Point II
IA	Sutherland Generating Station Unit 4	OK	Sallisaw Electric Generating Plant
IA	LS Power Elk Run Energy Station	OK	Red Rock Generating Facility
ID	Idaho Power company IGCC proposal	OR	Lower Columbia Clean Energy Center
ID	Power County Advanced Energy Center	PA	Beech Hollow Energy Project
IL	Dynegy/Illinois Power - Baldwin Energy Complex	PA	Good Spring Plant
IL	Franklin County Power Plant	PA	Greene Energy Resource Recovery Project
IL	FutureGen	PA	River Hill Power Project
IL	FutureGen 2.0	PA	Sithe Shade Township Project
IL	Madison Power Corp.	SC	Cross Generating Station Unit 3
IL	Prairie State/Peabody	SC	Cross Generating Station Unit 4
IL	Southern Illinois University at Carbondale Plant	SC	Pee Dee Facility
IL	Springfield- Dallman Unit 4	SD	Hyperion Energy Center
IL	Taylorville Energy Center	SD	Milbank / Big Stone City
IN	Duke Energy's Edwardsport plant	SD	NextGen Energy Facility
IN	Indiana Gasification	TX	ConocoPhillips Sweeny
IN	Purdue University's Wade Power Plant	TX	Freeport Plant
KS	Holcomb / Tri-State	TX	TXU Oak Grove Plant
KS	Westar Energy Kansas Plant	TX	Las Brisas Energy Center
KY	Cash Creek IGCC	TX	Spruce Unit 2
KY	Estill County Energy Partners	TX	Summit Power -Texas Clean Energy Project
KY	Kentucky Mountain Power (EnviroPower)	TX	Twin Oaks Power Unit 3
KY	BELL Sky Energy	TX	Sandy Creek Plant
KY	Smith Station	TX	White Stallion Energy Center
KY	Spurlock Power Station Unit 4	TX	Limestone III
KY	Thoroughbred Generating Station	TX	Coleto Creek Expansion
KY	Trimble County Generating Station 2	TX	Nueces IGCC Plant
LA	Big Cajun II Unit 4	TX	Tenaska - Trailblazer Energy Center
LA	Little Gypsy refit	UT	NEVCO (Sevier Plant)
LA	Rodemacher Power Station (unit 3)	UT	Green River Plant
LA	Big Cajun I	UT	Hunter 4 Power Plant (PacifiCorp)
MA	Somerset Generating Station	VA	Cypress Creek Power Station
MD	Sparrows Point	VA	Virginia City Hybrid Energy Center
ME	Twin River Energy Center	WA	Wallula Energy Resource Center
MI	Northern Michigan University Ripley Heating Plant	WA	Energy Northwest-Pacific Mountain Energy Center
MI	Wolverine Power Plant	WI	Cassville/Nelson Dewey III
MI	Lansing Board of Water and Light	WI	Oak Creek Unit 1
MI	Consumers Energy Plant	WI	Oak Creek Unit 2
MI	Escanaba Power Plant	WI	Wausau / Weston 4
MI	Midland power plant	WV	Greenbrier County / Rainelle (Western Greenbrier)
MI	James De Young Station	WV	Longview plant/ Monongalia County
MI	Great Lakes Energy and Research Park / Alma	WV	New Haven (Mason County) / Mountaineer Plant
MI	TES Filer IGCC	WY	Jim Bridger unit 5
MN	Mesaba Energy Project	WY	Wygen Unit 2
MN	New Ulm Boiler #4	WY	Wygen Unit 3
MO	Associate Electric Cooperative's Norborne	WY	Basin / Dry Fork Station
MO	Iatan 2 (KCPL)	WY	Two Elk Energy Park Unit 1
MO	Southwest Power Station Unit 2		

Table 18: Coal power plants names

US State	Plant name
AK	Alaska Natural Resources-to-Liquids Plant
AK	Fairbanks Coal-to-Liquids
IL	Drummond CTL
IL	Decatur Plant
IN	Clean Coal Refining CTL plant
KY	Fuel Frontiers plant
KY	Clean Coal Power Operations Coal-to-Liquids Plant
KY	Buffalo Creek Energy CTL
KY	Secure Energy Paducah Plant
MT	Roundup Coal-to-Liquids
MT	Malmstrom Air Force Base Coal-to-Liquids
MT	Ambre Energy plant
MT	Many Stars Coal-to-Liquids
ME	Twin River Energy Center
MS	Belwood Coal-to-Liquids Project (Natchez)
ND	American Lignite Co's Coal-to-Liquids plant
OH	Ohio River Coal-to-Liquids plant
PA	Gilberton Coal-to-Clean-Fuels and Power Project
PA	EmberClear CTL
TN	Freedom Energy Diesel
TX	Hunton "Green Refinery" CTL
WV	Benwood project (Marshall County Industrial Park)
WV	Mingo County Project
WV	TransGas Development Systems CTL
WY	Medicine Bow plant

Table 19: Coal-To-Liquids plants names