To What Extent has Wind Energy's Need for a Fossil Fuel Backup Reduced its Net Benefit

Problem Definition: The project addresses the following question. *"Fact or Fiction: The intermittent nature of Renewable sources of energy, such as Wind and Solar, is such that the requirement to back them up imposes inefficiencies on existing fossil fuel plants, as measured by higher Heat Rates that negate most of the purported emission-reductions benefits from such Renewables*". The heat rate of a power plant is the amount of heat (Btu) from the combustion of fuel that is required to generate a standard kilowatt hour of power. This is important as the CO2 emissions are 100% directly related to the volume of fuel combusted – more fuel means more CO2!

Principal Finding: For the ERCOT electrical grid in Texas, 87% of the observed variability in the monthly heat-rates of the combined-cycle gas turbine power plants that constitute 93% of the gas-fired electrical power generation in Texas can be explained by a multivariate linear regression model that uses just six variables. These variables are, listed in order of their decreasing statistical significance, respectively:

1. The aggregate amount of electrical power generated by all gas-fueled power plants in ERCOT that month,
2. Time (since over time there has been a trend of adding new, and more efficient units),
3. Heating Requirements (specifically the number of Heating Degree Days that month),
4. The aggregate amount of Wind power generated by all wind turbine power plants in ERCOT that month,
5. The aggregate amount of nuclear power generated by all nuclear power plants in ERCOT that month, and
6. The month of the year (the colder Winter months having lower heat rates, the hotter Summer months having higher heat rates, and the lower-load months of March and April being subject to important seasonal maintenance of the nuclear plants.

The first four variables were each observed to be statistically significant at the 1% level (the likelihood of such a strong relationship being observed by chance being less than one percent), with the fifth variable (*nuclear power generation*) being significant at the 5% level. Of the twelve months of the year, eight were statistically significant at the 1% level. Recurring seasonal factors are clearly important in effecting the efficiency of power plants and need to be separately accounted for lest they act as a confounding variables.

While the amount of Wind Energy generated was observed to have a positive and highly statistically-significant relationship with the so-called Heat Rate of ERCOT's combined-cycle gas fired power plants, the relatively modest magnitude of this relationship implies that, when the added inefficiencies measured by the higher Heat Rates are factored in***, the imputed carbon footprint of an additional unit of wind generation is still approximately 80% lower than the average for the combined cycle gas-fired plants*** (and approximately [90]% lower than for ERCOT's coal-fired power plants). Importantly, these figures do not factor in the so-called "Peaker" single-cycle gas turbines, which although they constitute less than 10% of the gas-fired power plant capacity may be disproportionately used to back-up Wind generation on account of the speed with which they can be brought on line. This is a topic for potential future work. The point of this work is that at ERCOT, Renewables have been much more successful in reducing CO2 emissions than in Germany, for example.

Problem Significance: The question of whether Renewable forms of electrical power generation such as Wind and Solar truly convey reductions in aggregate CO2 emissions is highly controversial, with extreme views on either side. On the one hand, Green Groups fiercely claim that there is almost zero offsetting effect in terms of Wind generation making existing plants less efficient, and raising their heat rates. For an opposite view, see for example a recent piece by ASU's Professor Peter Rez entitled "*Why solar and wind won't make much difference to carbon dioxide emissions*". Available at https://blog.oup.com/2017/10/solar-wind-energy-carbon-dioxide-emissions/ . Just as cars are most fuel efficient when operated at a constant speed, so are power plants. Just a 20% drop in wind speed can cause a wind turbine to reduce its output by 50% or more, and so existing fossil fuel plants need to be able to ramp up and down their output within minutes. Power plants are usually most efficient when operated at full load, but the requirement to be able to increase output quickly means they may be obliged to operate at part load with a lower level of efficiency. So we have a possible so-called *"Fallacy of Composition"*: wind power itself emits no CO2 emissions, but the CO2 emissions of the supporting power plants may be materially increased.

Most fundamentally, Climate Change is generally perceived as an important problem for this and future generations. It seems that most of the world's largest cites located next to bodies of water may be subject to repeated flooding, and entire continents (such as Australia) may eventually become uninhabitable as the heat and heat humidity combine to make human habitation intolerable (according to the Nobel Prize-winning IPCC). Drought, crop failure, and refugee migration are already recognized from harsh experience as important problems. It is therefore of fundamental importance to society that inexpensive and genuine ways of reducing CO2 emissions are found and implemented. It is troubling (and perplexing) for many that Germany has spent more than $220 billion on Renewable forms of energy and related transmission infrastructure, yet aggregate emissions have simply not declined very much, as shown by charts published by ASU's Professor Peter Rez, as cited above.

Can modern Data Science techniques be applied in a rigorous manner to shed light on these questions? Specifically, if we look at the electric grid with the largest penetration of Renewables in the US (in Texas), can the comprehensive data sources freely available only in America be mined to infer clear statistical relationships? Moreover, if several thousand Megawatts of new Wind Turbine capacity are contracted for and have been financed, and are therefore likely to come on stream soon, can we actually predict with some accuracy what the net effect on aggregate CO2 emissions will be when the various interdependencies within the overall grid system are fully taken account of. This analysis should be of interest to all states and countries which are seeking to most economically reduce their CO2 emissions.

Personal Background: I have myself been aware of this sensitive and important question as to whether Renewables really do materially reduce CO2 emissions in aggregate since 2011, when I read a report by Peter Lang which purported to prove that they did not. I shared his detailed calculations with a former Cambridge University Professor, who responded that he could see no obvious errors in the calculations, but assured me that *The Department of Energy & Climate Change* of Her Majesty's government did not see it that way. He personally had no appetite for examining such a sensitive question. Over the years I have asked many distinguished experts this same question, including very senior and distinguished German energy experts such as Claudia Kemfert (http://www.claudiakemfert.de/), but not a single response has been forthcoming. That is why this effort is being made to seek to address this question using modern Data Science techniques, in a manner that does not appear hitherto to have been done (publicly, at least).

Methodology: The Heat Rate of a gas-fired power plant is the amount of natural gas fuel (measured in Btu's) that must be combusted to generate a given volume of electric power, specifically a standard kilowatt hour. There is a linear and proportionate relationship between the volume of fuel combusted and the derived CO2 emissions. So if we can calculate how Wind Energy may increase the heat rates of power plants, then our question can be readily answered.

I. Initial Exploratory Analysis:

Theoretical Background: the thermal efficiency (in converting heat to electrical power) of a modern combined-cycle gas turbine (CCGT) plant may reach 56%. However, it is physically possible for the efficiency of the very same unit to decline to just 28%, if it operates (a) in single-cycle mode (so the heat energy in the steam from the gas turbine's exhaust is not recovered from a steam cycle), and (b) it operates at a low load level, such as 40% of the capacity of the turbine. The reason why this may happen would be to allow the turbine to ramp up its power quickly from 40% of capacity to 100% in just ten minutes, for example to compensate from a decline in wind power due to a sudden drop in wind speed (which is very common). In contrast, a CCGT operating with its steam cycle may take as long as six hours to reach its full capacity. For this reason, Renewables could theoretically increase from zero percent to 50%, but if they were displacing the carbon-free steam cycle in the manner just outlined above, and also causing low load operation, then there could be no net decrease in the amount of natural gas combusted, and so too in the CO2 emitted.

An Empirical Question: But has anything like what is outlined above actually transpired at ERCOT, during the period in which electricity from Wind energy has increased more than tenfold to 35% or more? To examine this we identified a common set of CCGT plants that were operating in 2006, and also in 2016. With the same set of plants, obvious factors such as the introduction of new, large, modern highly efficient plants would be controlled for. Over this period, from 2006 to 2016, electric power generated from wind turbines had increased more than tenfold. So one approach is simply to look at the heat rates of plants in 2006, and then plot them against the heat rates in 2016. A line at a 45-degree angle would imply that the heat rate in 2006 and 2016 was identical. With 2006 data on the x-axis, and 2016 data for the y-axis, clusters of points above the line would suggest that heat rates had increased over that period, potentially supporting the notion that the growth in Wind energy could be a contributing factor. More formally, a statistical "Paired t-Test" could be conducted, which is a common and relatively "statistically-powerful" technique.

Precisely the above analysis was conducted, but the scatter plot had no obvious systematic tendencies. The heat rates for some plant had increased somewhat, while those for other plants had decreased. In fact, over the period on average there was a very small (just 0.1%) decrease in heat rates on average. In short, there was no evidence of an increase in Heat Rates due to the growth of Wind energy during the period 2006 to 2017.

However, other things could be going on that could disguise the effect. For example, there could still be effects due to capital improvements, such as the introduction of so-called "chillers", which chill the intake air used for combustion and so increase efficiency and lower the heat rate of a plant. These could have compensated for any increase in Heat Rate due to the intermittent nature of Wind Energy. An analysis is therefore required that looks at multiple factors that could confound the superficial aggregate analysis outlined above.

[slides pertaining to the above commentary are in the slide deck, entitled "Initial Exploratory Analysis"].

II. Subsequent Analysis using Multivariate Linear Regression

This analysis looks at multiple factors that could confound the superficial aggregate analysis outlined above. Although many variables were considered from diverse sources, it was found that just six variables can be used to explain 87% of the monthly variability in the Heat Rates of the subject plants in Texas, and each of these variables (using the STATSMODELS package) was very highly statistically significant – see below.

Executive Summary: while many independent variables were evaluated, the final model selected for ERCOT uses just six variables, namely (a) the amount of monthly power generated by Coal, (b) the amount of generation from Wind (and its square, so utilizing a non-linear relationship), (c) the Month of the year, and Time. Each of these variables was highly statistically significant, with the likelihood of the observed results having occurred by chance being less than 1% in each case (so each was significant at the 1% level). These variables were able to explain 87% of the aggregate variability in the so-called Heat Rate of the roughly forty combined-cycle gas-fired power plants that provide most of ERCOT's power.

Confounding Variables: it should be acknowledged at the outset that this analysis is flawed because it omits certain important variables which are not at all easy to observe (if at all) from publicly available data sets. The possible extraction of such data may be the subject of future work with appropriate sponsorship. Specifically, the intermittent nature of Wind Energy may effect (a) the average capacity utilization of plants (for example, lowering it), and (b) increased "cycling", when plants shut down and start-up. Far more than one hundred hours has already been spent on this Project, but there is still potential for clear improvements. At the same time, the work completed to date does offer a parsimonious model with a degree of explanatory power. Additional historical data is pending from different sources, and this may potentially be added when and if it becomes available. The integration (through so-called Joins) of even more data sources, such as additional EIA and EPA data may also be undertaken.

Cluster and Predict: another avenue for improving upon the 87% explanatory power of this model would be to cluster the approximately forty combined-cycle power plants into distinct groups, such as their actual output, their rated capacity, their capacity utilization, their flexibility to operate at low load, and other documented characteristics (including their location relative to the major wind generating areas, and related transmission lines). The heat rates of each of the cluster would then be predicted separately and aggregated. It is hard to believe that this would not improve the explanatory power, but this initiative would take months of data collection, data cleaning and wrangling, and finally model building.

A Note of Caution: one lesson from the much-discussed financial meltdown in 1998 of the *Long Term Capital Management,LLC* a prominent hedge fund headed by celebrated Nobel Laureates, was that sometimes financial models can be somewhat local in nature. They work well….until the outside world changes a lot, when they may work only very poorly (or not at all). In somewhat the same way, this model is a reasonable predictor for almost all months, but with a single notable outlier, in April of 2010 when two nuclear plants were taken offline at the same time, and spurred by the shortage of generation less efficient plants were brought into service and the average heat-rate of the subject plants surged very markedly. A more comprehensive model would model coal, nuclear, natural gas and wind quite separately, each with their own model. In essence, in the terminology of economics, a supply curve for each would be modelled, and these would be combined in the context of an aggregate demand model to solve for "market-clearing" values for all variables. This model is outside the scope of this initial analysis, which as discussed above is relative simple in its final form, and readily understandable to a non-expert.

Data Sources

The ERCOT (Electric Reliability Council of Texas) grid was chosen to evaluate this question because it has the highest penetration of Renewables of any of the major grids, nearly all in the form of Wind Generation. Specifically, power generated by Wind turbines has exceeded 45% of ERCOT's total generation at times, and more than 25% of the monthly total. ERCOT manages the flow of electric power on the Texas Interconnection that supplies power to 24 million Texas customers – representing 85 percent of the state's electric load. ERCOT is the first independent system operator (ISO) in the United States and one of nine ISOs in North America. Moreover, it is relatively independent and self-contained, with only minimal export and import of power to or from other grids. As a self-contained grid, it provides a basis of comparison with countries whose grids are also relatively self-contained.

ERCOT Monthly Data: from ERCOT's monthly "*Demand & Energy*" reports, the sixth tab (entitled "*Energy by Fuel Type*" was used to source the monthly generation data for electrical generation from Natural Gas, Coal, Nuclear, and Wind, respectively. In addition, average percentage Load data (specifically "*Net System Load Factors based on Hourly Demand*") was imported from the Excel Tab from the same report headed "Energy".

EIA Form-923, which is published annually be the Energy Information Agency. The U.S. Energy Information Administration (EIA) is a principal agency of the U.S. Federal Statistical System responsible for collecting, analyzing, and disseminating energy information to promote sound policymaking, efficient markets, and public understanding of energy and its interaction with the economy and the environment. EIA programs cover data on coal, petroleum, natural gas, electric, renewable and nuclear energy. EIA is part of the U.S. Department of Energy. The EIA's Form EIA-923 (and predecessor forms) provide monthly and annual data on generation and fuel consumption at the power plant and individual generator level. Specifically, EIA-Form 923 collects data on more than 12,000 power plants. The data may be accessed by the public via a downloadable Excel spreadsheet, each of which has multiple tabs, and as many as 100 columns per spreadsheet. This data was downloaded in a CSV format and read into a Pandas dataframe, and the fuel consumption and power generation for combined-cycle gas turbine plants in Texas was extracted on a monthly and annual; basis. This was not a straightforward process, and in itself took more than thirty hours and a good deal of help from Stack Exchange. Even experienced data science professionals have posted on Github seeking advice as to how to read these files correctly into a dataframe. In short, this data was used to compute the heat rates of the approximately forty combined cycle gas turbine power plants in ERCOT, and from this a composite monthly heat rate – the dependent variable in our models, or the variable whose behavior is to be explained.

Natural Gas Prices for Power Generation: The Energy Information Agency files were searched for relevant prices for natural gas, and two series were identified - natural gas prices specifically for electrical power generation in Texas, and another series for the USA (also for electric power generation) as a whole. Each variable is quite volatile from month to month. Given that low natural gas prices incentivize switching from coal to natural gas, but that such decisions may take time to implement, it might also be explored whether these price series should be lagged by one month, or even exponentially smoothed in some manner.

Data Wrangling: the challenges of importing the data from the EIA source Excel sheets are not to be underestimated. More than twenty hours was spent dealing with the common Excel practice of marking thousands by commas, compounded by data entered in inconsistent forms. Moreover, the presence of stray commas or periods strewn around the tables creates problems, as to change the type of a variable no missing entries can exist. This problem was addressed by a Python *dictionary* approach. Another issue is that the Combustion Turbine (coded "CT") and its associated steam turbine (coded "CA") are recorded as separate entries. So a separate dataframe was made for the CT's and the CA's, and these two dataframes were then merged together again in a reconstituted whole. Also, since new units are typically much more efficient than older units, this element of bias was in parts separated out by comparing the same subset of plants in 2006 and in 2016.

The Code for the following is set forth in the Jupyter Notebook entitled Reg1417July, as posted on Github:

Statistical Methodology: the model began with a multivariate linear regressions model from the Python STATSMODELS package. This package was chosen because it includes the "t-statistics" and other information commonly available from the R programming language (a successor to "S", essentially a statistical programming language). The first model included thirteen potential variables, namely:

month, being the month of the year

ng, being the Aggregate electrical generation from power plants powered by natural gas

coal, being the Aggregate electrical generation from power plants powered by coal

nuclear, being the Aggregate electrical generation from power plants powered by nuclear

wind, being the Aggregate electrical generation from power plants powered by wind

wind2, a non-linear variable, the square of the above item

windpct, being the percent contribution of wind to total generation

load , being the average hourly capacity for the system as a whole for that month

dheat , being the number of heating days

dcool , being the number of cooling days ("*Heating and cooling degree-days are indicators of how much energy a typical household or building will use for space heating or cooling. The more heating degree-days you have, the more energy it will take to heat the inside of the home or building. The more cooling degree-days you have, the more energy it will take to cool the inside of the home or building".)*

txgasprice , being the average natural gas price for gas supplied for power generation in Texas that month

usgasprice, being the national average natural gas price for gas supplied for power generation in that month

time, being the number of months since the start of the data set.

Commentary on Model Evolution: In the first model, with all variables included, apart from the single month of November, only three variables were statistically significant, namely wind, wind2 (being wind squared), and time. The R2 was 89%. Variables were then dropped stepwise in the following order, one at a time, and the model re-estimated (a standard statistical procedure for which stepwise packages are available in R). Specifically:

- After the first round, ***dcol*** was dropped as it was then the least significant remaining variable.
- After the second round, ***windpct*** was dropped as it was then the least significant remaining variable.
- After the third round, ***txgasprice*** was dropped as it was then the least significant remaining variable.
- After the fourth round, ***coal*** was dropped as it was then the least significant remaining variable.
- After the fifth round, ***usgasprice*** was dropped as it was then the least significant remaining variable.
- After the sixth round, ***load*** was dropped as it was then the least significant remaining variable.

At this point each of the remaining variables was statistically significant, so no further drops were made. Although six variables had been dropped, the R2 had declined by just 2%, from 89% to 87%.

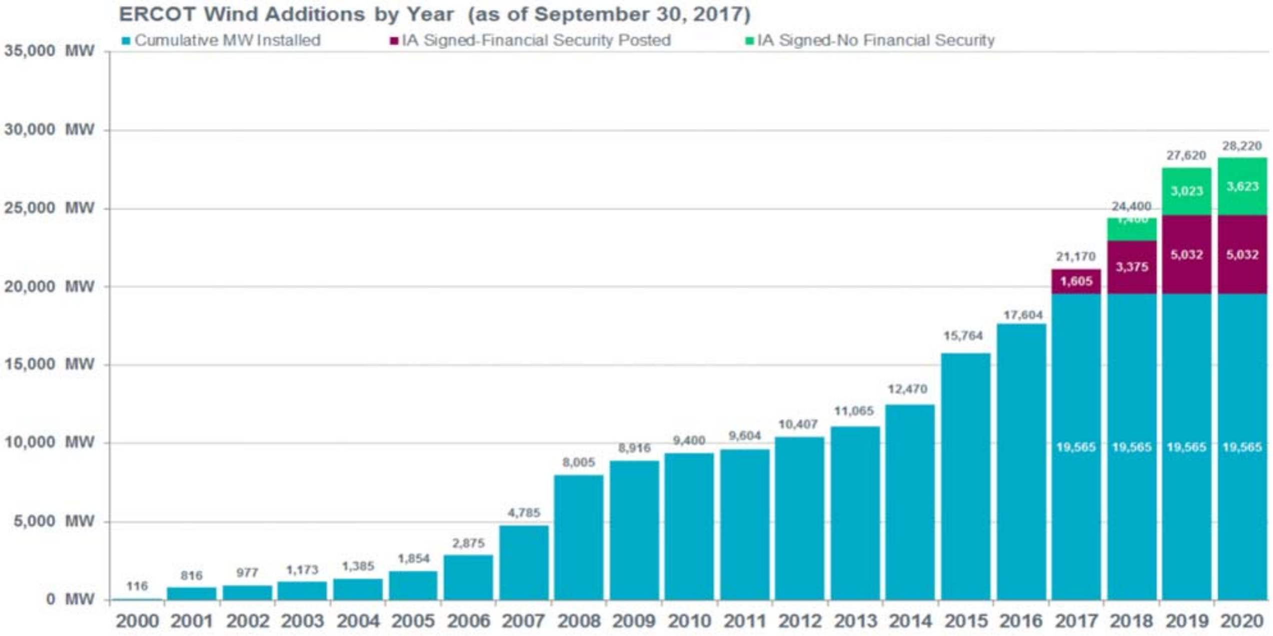
Of particular note in this context is the negative sign of the coefficient for the variable ***wind2***, being the variable ***wind*** raised to the second power. There has been commentary from Europe and Australia that suggests that as wind generation hits a particular threshold of say, 20% of total power generation, then the inefficiencies imposed on the fossil fuel plants become increasingly marked. If this were indeed the case at ERCOT, then this variable ***wind2*** should be of positive sign, and statistically significant. Instead, we observe the opposite, a negative coefficient which is statistically significant. How can this be? One possible explanation is that as Wind generation has grown, investments to de-bottleneck electric power transmission lines have been made, so that curtailments have actually declined in relative importance over time. In other words, when wind generation attains a certain critical mass, more investments are made in transmission to integrate it. Another possibility is that as the number of wind generation sites has grown, generation has become more diversified and volatility has become averaged out (this is a readily testable thesis). As a third alternative, as Wind generation has become more important it may have become more worthwhile to invest in the development of sophisticated models that cam more accurately predict wind generation levels within different time horizons (eg of minutes, hours, and days). This area is a suitable topic for future study.

[slides pertaining to the above commentary are set forth in the slide deck, entitled "Multivariate Regression Analysis of Plant Heat Rates").

Future Work: it would be of course desirable to complete the year 2017, and also to extend the model back to earlier years. A data request from AEP is pending for the variables *dcool* & *dheat*, and requests are also pending with ERCOT and the EIA. As more data points are added an "out-of-sample" R2 should also be computed, after first dividing the data into training and testing sets using the standard packages. More fundamentally, we should make the considerable extra effort to extract the data that is truly most physically relevant, namely the number of times per month that a plant starts-up and shuts down,

Python/Pandas Code: the code underlying this analysis currently encompasses twenty-two Jupyter Notebooks. Each year of EIA data from 2006 to 2017 was given its own Notebook, as problems associated with changes in labels and omitted data varied markedly by individual year. Currently under investigation is whether this can be reduced to just a single Notebook, for example by defining appropriate functions and, by using FOR loops, or list comprehension, iterating across files from different years.

Forecast: [*to come*]


ERCOT Wind Additions by Year (as of September 30, 2017)
Cumulative MW Installed
IA Signed-Financial Security Posted
IA Signed-No Financial Security
35,000 MW
30,000 MW
25,000 MW
20,000 MW
15,000 MW
10,000 MW
5,000 MW
0 MW