Eastern Wind Turbine Analysis and Reverse Engineering Data Sources

A case study in realizing open data

Steven Githens

IUPUI

sgithens@iupui.edu

Abstract

In order to be successful, the modern data scientist must be able to nimbly adapt to a wide array of data sources, schemas, and software tools. More than ever there is an abundance of freely available research data and numerous open source software packages on hand to analyze it. In this paper, we present a case study of how one might go about processing this data, and some of the tools that can be used. Additionally we talk about what can be done in difficult situations when there is only limited access to either the data or the tools used to produce it.

For our case study we use a set of data that was commissioned by the National Renewable Energy Laboratory [2] to simulate windfarm turbines in potential locations around the United States. The data was seperated into sets for the eastern and western sides of the country. The eastern data set was chosen simply because that happens to be where the author resides.

General Terms Statistics, Data, Visualization

Keywords R, Python, WRF, BatchGEO

1. Going Forth With Purpose

When faced with a new data set or corpus of material on a subject we wish to explore, there is always some purpose or reason in the back of our minds that we are equipped with to see us through the task. There are a few these in this study that the author has in mind which will be presented now.

The data set itself in this case is important as it deals with renewable energy, a key component is making sure our species does not become extinct in the new few centuries by the elimination of it's suitable habitat! In order to create the dataset, upwards of 7000 potential locations were chosen, both onshore and offshore of the atlantic seaboard and great lakes. These simulated wind turbine locations can be used as key aspects for affecting policy decisions and investments in energy in these areas.

The second motivation, that did not fully present itself until investigating the data, is having open access to data. In this case we do not mean just being able to freely download the data, but also being able to access the tools and methods necessary to reproduce the data, so that organizations can truly work with it in the future.

Lastly, it should be noted that this is in part an *educational* endeavor, so the goal is largely to demonstrate the usage and possibilities of the software packages involved. It would be untenable to use any of the results contained within for making serious desicions, however the approaches can be used as a foundation for conducting larger, more long-term studies.

2. About the Data

2.1 Overview

The data set being used is the Eastern Wind Dataset [1].

The raw data set consists of simulated wind data for roughly 7000 sites on the eastern half of the United States. These were selected based on varying criteria such as distance from airports, park land, and other issues. The wind speeds were then simulated using a proprietary weather model, with the output consisting of wind speed predictions every 10 minutes for 3 years including 2004, 2005, 2006. For various logistic reasons, this varies a bit for some sites (such as a few sites only include 1 year or some other fraction of the time), but in general the data set is very complete. Also included are intervals predicting the weather in 4, 6, and 24 hour forecasts. For this paper we will stick to the 10 minute simulated wind data.

Each selected geographical site contains 1 comma separated value (csv) file to hold it's contents. These files can become quite large, and the entire dataset is several gigabytes. For each of these files, the columns include:

- Date, separated into day and minute columns.

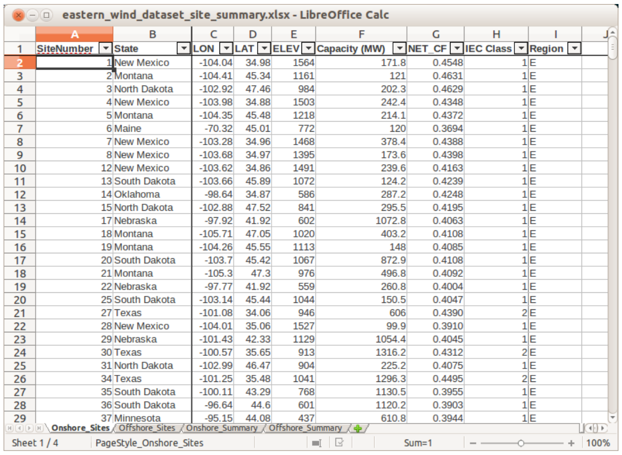
```
SITE NUMBER: 00001 RATED CAP:  171.8 IEC CLASS: 1 LOSSES (%): 14.2
SITE LATITUDE:   34.98420 LONGITUDE: -104.03971
DATE,TIME(UTC),SPEED80M(M/S),NETPOWER(MW)
20040101,0010,6.60,46.34
20040101,0020,6.77,48.00
20040101,0030,7.17,53.37
20040101,0040,7.84,62.71
20040101,0050,8.97,79.92
```

Figure 1: Raw data structure

- Wind Speed, in meters per second, measured at 80 meters above the ground.
- Netpower, in MegaWatts.

Additionally, the header lines of each file include important information we will scrape such as power class and rated power cap, latitutde and longitude, and site number. 1

The NREL also has some data in an excel sheet that aggregates the above data set into a summary. This includes information such as the site number, state, averaged power and capacity factors. 2



eastern_wind_dataset_site_summary.xlsx - LibreOffice Calc

	A	B	C	D	E	F	G	H	I
1	SiteNumber	State	LON	LAT	ELEV	Capacity (MW)	NET_CF	IEC Class	Region
2	1	New Mexico	-104.04	34.98	1564	171.8	0.4548	1	E
3	2	Montana	-104.41	45.34	1161	121	0.4631	1	E
4	3	North Dakota	-102.92	47.46	984	202.3	0.4629	1	E
5	4	New Mexico	-103.98	34.88	1503	242.4	0.4348	1	E
6	5	Montana	-104.35	45.48	1218	214.1	0.4372	1	E
7	6	Maine	-70.32	45.01	772	120	0.3694	1	E
8	7	New Mexico	-103.28	34.96	1468	378.4	0.4388	1	E
9	8	New Mexico	-103.68	34.97	1395	173.6	0.4398	1	E
10	12	New Mexico	-103.62	34.86	1491	239.6	0.4163	1	E
11	13	South Dakota	-103.66	45.89	1072	124.2	0.4239	1	E
12	14	Oklahoma	-98.64	34.87	586	287.2	0.4248	1	E
13	15	North Dakota	-102.88	47.52	841	295.5	0.4195	1	E
14	17	Nebraska	-97.92	41.92	602	1072.8	0.4063	1	E
15	18	Montana	-105.71	47.05	1020	403.2	0.4108	1	E
16	19	Montana	-104.26	45.55	1113	148	0.4085	1	E
17	20	South Dakota	-103.7	45.42	1067	872.9	0.4108	1	E
18	21	Montana	-105.3	47.3	976	496.8	0.4092	1	E
19	22	Nebraska	-97.77	41.92	559	260.8	0.4004	1	E
20	25	South Dakota	-103.14	45.44	1044	150.5	0.4047	1	E
21	27	Texas	-101.08	34.06	946	606	0.4390	2	E
22	28	New Mexico	-104.01	35.06	1527	99.9	0.3910	1	E
23	29	Nebraska	-101.43	42.33	1129	1054.4	0.4045	1	E
24	30	Texas	-100.57	35.65	913	1316.2	0.4312	2	E
25	31	North Dakota	-102.99	46.47	904	225.2	0.4075	1	E
26	34	Texas	-101.25	35.48	1041	1296.3	0.4495	2	E
27	35	South Dakota	-100.11	43.29	768	1130.5	0.3955	1	E
28	36	South Dakota	-96.64	44.6	601	1120.2	0.3903	1	E
29	37	Minnesota	-95.15	44.08	437	610.8	0.3944	1	E

Figure 2: Aggregated data for each site

2.2 Looking at a particular day

To get a better feeling of what a day of wind data looks like, we can use R to parse a file for a single day, and then pull just one days worth of data from that file. [1]

From the graphs of time vs speed 3 and time vs power 4 we can get a feel of how the speed affects power as the curves mimic each other to a certain extent.

2.3 Collecting data from all Site Files

In addition to working on single site csv files, we may want to collect and process header information from all of them. This would be necessary for reproducing the final aggregated data as part of the next section, but also for other visual means.

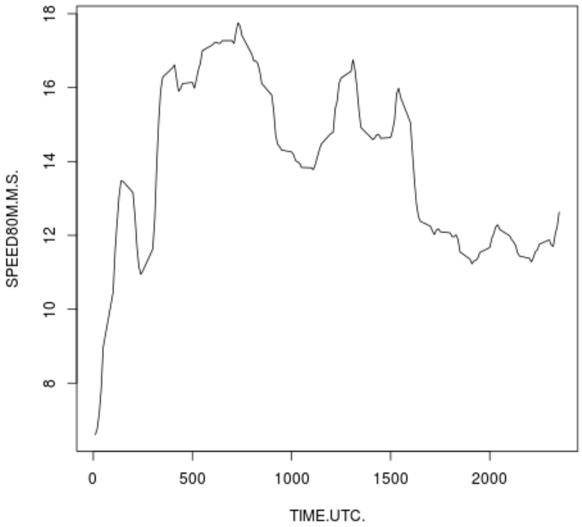


Figure 3: Time vs Wind Speed for a Single Day

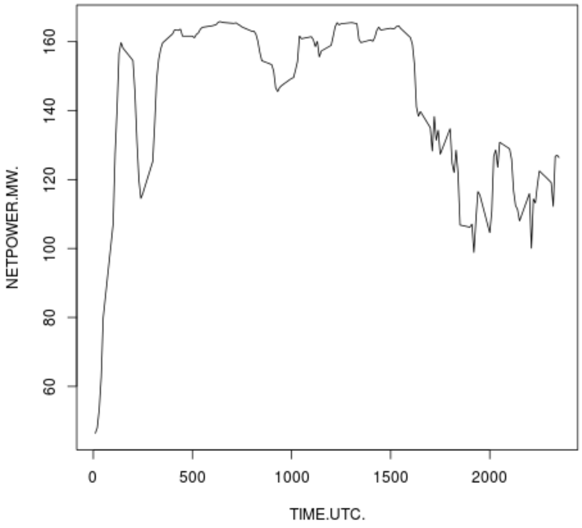


Figure 4: Time vs Power for a Single Day

The python source shows how we can parse all the site csv files we have from the simulated study, and then output them in a format suitable for creating a geographic visualization. In this example we are creating another csv file that can be used as input to BatchGEO, and popular Web 2.0 site that allows overlaying site data on a Google Map or Open Streetmap style mashup. 5 [2] Currently we are just pulling

[1] See the plotDay function in the R source

[2] The output from this data run is located at `http://batchgeo.com/map/c0bb3d2534dc11a0ac8e303b5cff809e`

out the site number, latitude, and longtitude, but any of the header items could be pulled out.

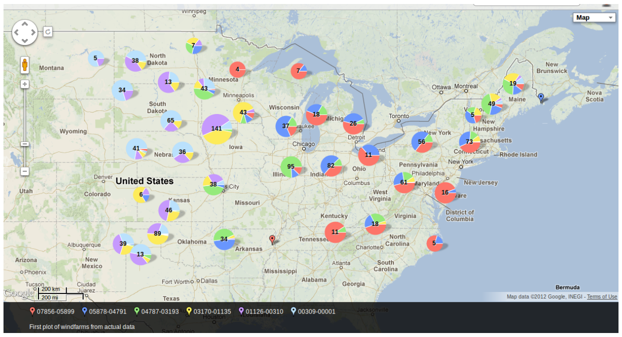


Figure 5: Interactive map of onshore sites

3. Reproducing the Data

3.1 A Small Excercise in Reverse Engineering Data

Next we turn to reassembling the data, and reproducing the study from scratch. This turns out to be interesting for us, because it allows us to look at some interesting statistical qualities of the data, and to explore more open source software packages.

The crux of the issue here is that a proprietary model was used to simulate the wind speeds, so that's no good for us. Luckily, the data set for the western side of the United States was simulated with an open source package, so there is hope that that can be used to fill in the gaps. This is a fairly commonly occuring phenomenon for the open data scientist: finding that certain parts of their progress are blocked by not having access to some of the required tools. Whatever the setback though, the challenge of working around them is always invigorating and usually introduces one to new communities of like minded scientists pursuing the same issues, and improving the global camaraderie among peers.

The second part of reproducing the raw data is determining a turbine model that can be used to create the megawatt output for each 10 minute interval based on the wind speed. The outputs in the raw dataset where produced by AWS [11] by combining information from 3 commercial turbine engines.

Given a current lack of access to civil or mechanical engineers we will bootstrap [3] our model by building it based off of the *current* data from the study. Given the incredibly large number of samples , this has turned out to be a reasonable approach.

With the above 2 requirements we are able to reproduce the raw output file format of the study. At that point, we would just need to aggregate them, and apply formulas to calculate the capacity factor based on the power outputs. For this iteration of the case study, one will notice that we are also leaving out the actual *selection* of the geographical sites. It's presumed that one could do this, given there are fairly available and open data sets for geographical data in the United States.

3.2 Simulating the Wind Speeds

The problem of simulating wind speeds can be solved by using an open software package that was used for generating the analogous data set for the western half of the United States. The Weather Research & Forecasting Model [5] is a collaborative project among a number of universities to build a set of tools for predictive analysis and simulation of weather and atmospheric conditions.

The project is written in C and Fortran and relatively straightforward to compile on a modern Linux distribution such as Ubuntu or Fedora. Being a rather large project, there is some learning curve, so for this study we did not get much further than downloading the source [7], compiling, and looking a few of the tutorials [6]. However, given adequate resource, the development of the configuration files and sample data for simulating wind speed samples on the eastern half of the continent is certainly feasible.

3.3 Looking for a Turbine Curve Model

We want to create a model for the turbine curve based off of wind speed. For sake of computation speed, we will use 1 days worth of data from site 1 for these figures. [4]

A beginning scatter plot shows that the data is a bit like a cubic on it's side, and we may want to see if we can still fit a linear model to it. It's clear that a first order model will be an awful fit, but we can come close with a cubic model. 6

This may be Ok, but it justs feels wrong. We *know* that this model was developed based on real turbine engines, and by doing a small bit research [3] we find that it's well known that these [5] can be modelled with Weibull distributions, so it makes sense to explore that.

We can do some exploration and manually fit a Weibull over our data by giving it the appropriate X and Y offsets, using a shape parameter of roughly 1.8, and keeping alpha at 1. 7 We can now turn to R's nls function for fitting non linear models. Unlike the linear model fitting in R, rather than providing a number of coeffients on the right hand side to fit, we provide an equation with certain unbound variable parameters, and these are what R will fit to solve the minimization problem.

Because NLS works by iterating along a gradient problem to find the minimization, it is necessary to provide starting values for the equation. This can be a very tricky problem, and as Peter Dalgaard describes, "Finding starting values

[3] It's important to note that in this scenerio, we use the term bootstrap in it's software engineering context, not in the context of statistical bootstrapping that invovles taking subsamples of your existing data, although we are doing work based on subsets of the data.

[4] Also because when plotting large amounts of the data, the curve becomes very dense from the points, making it hard to see other overlays and trends in the data.

[5] As well as many other other engineering problems

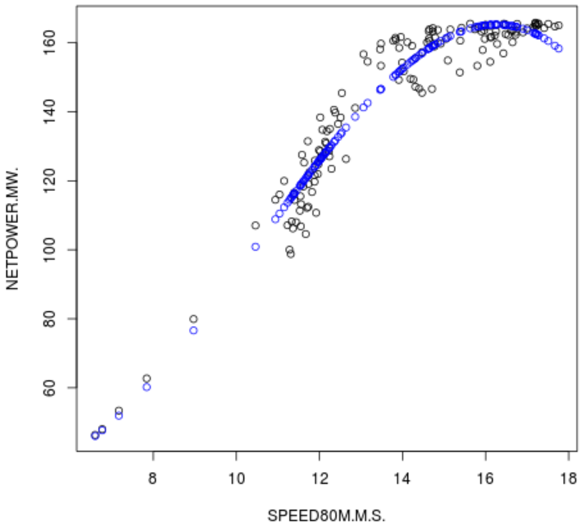


Figure 6: First attempt with cubic linear model

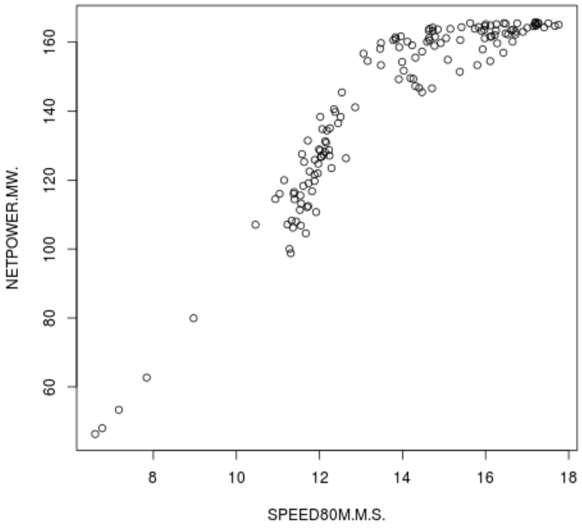


Figure 7: Manual inspection of a weibull to our data

is an art rather than a craft" [4] Given the author's current amatuer statistical abilities, we were unable to select starting values, but all is not lost! R ships with several *self starting* model functions that can be used in the nls function. When a self starting model is used, it is able to guess values to start with. In the case of this model, it works very well, with the fitted weibull 8 now modelling the turbine output curve quite well.

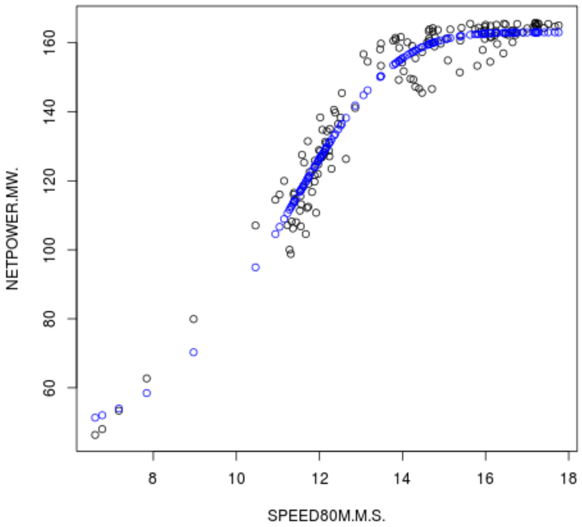


Figure 8: Nonlinear Weibull Model Fit

```
Formula: NETPOWER.MW. ~ SSweibull(SPEED80M.M.S., Asym, Drop, lrc, pwr)

Parameters:
     Estimate Std. Error t value Pr(>|t|)
Asym 163.0551     0.8997  181.23   <2e-16 ***
Drop 116.3540     4.3543   26.72   <2e-16 ***
lrc  -13.7634     1.0135  -13.58   <2e-16 ***
pwr    5.5966     0.4022   13.92   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 5.839 on 139 degrees of freedom

Number of iterations to convergence: 4
Achieved convergence tolerance: 4.451e-06
```

Figure 9: Fitted NLS Values

From the output 9 we can see the values of the parameter fits for the nonlinear model, and that it only took 4 iterations to minimize the problem using the self starting weibull function.

We can now take random samples from this resulting weibull distribution to model our output from the wind speeds.

4. Conclusions

We hope this brief survey of open source tools, visualization libraries, and statistical techniques has shown how it's possible to conduct import research and work to improve the public good even when portions of the necessary stack are closed or missing. Starting with a data set of great importance to society, namely of potential renewable energy sources, we've used several packages to analyse and find ways to reproduce the data. These techniques can now be used to further develop the data necessary to make informed policy decisions, and reproduce the study in other sections of the world, or repeat it in parts of the United States.

5. Running the code samples

Both the python and R code for this project can be run using the standard libraries included in their distributions and do not require extra libraries. In order to run the samples you will need to download at least the onshore_sites.zip file from the data download area [8] and unzip it's contents into a folder named 'actual' placed in the same directory as the python and R files. The different R routines are in the main function and can be uncommented separately to be run.

The aggregated spreadsheet [9] can also be found on the eastern wind data site.

The source code for the examples, as well as the ltx source for this paper can be found on the projects github site. [10]

References

[1] Eastern Wind Data Set, `http://www.nrel.gov/electricity/transmission/eastern_wind_methodology.html`

[2] National Renewable Energy Lab, `http://www.nrel.gov`

[3] Wind Statistics and the Weibull distribution, `http://www.wind-power-program.com/wind_statistics.htm`

[4] Peter Dalgaard: Introductory Statistics with R, Springer, 2008

[5] Weather Research & Forecasting Model Project `http://www.wrf-model.org`

[6] WRF Tutorial `http://www.mmm.ucar.edu/wrf/OnLineTutorial/index.htm`

[7] WRF Download `http://www.mmm.ucar.edu/wrf/users/download/wrf-regist.php`

[8] Eastern Wind Dataset Download `http://www.nrel.gov/electricity/transmission/eastern_wind_disclaimer.html?ftp`

[9] Aggregated Eastern Wind Dataset `http://www.nrel.gov/electricity/transmission/docs/eastern_wind_dataset_site_summary.xlsx`

[10] Project site on Github `https://github.com/sgithens/windfarm`

[11] Final Report from NREL Study `http://www.nrel.gov/electricity/transmission/pdfs/aws_truewind_final_report.pdf`