标题

Raw Wind Data Processing: A Data-Mining Approach

问题

- The method proposed did not consider unnatural data.
- In real-world applications, artificial judgment is limited and inconvenient when the size of the database is large, and the wind farm operation state records are often unavailable.

贡献

- The contribution of this paper is to develop an empirical methodology for raw wind data preprocessing. The only information required for this methodology is the aggregated wind power output of the wind farm collected from the SCADA system, which is available at the dispatch center, and the wind speed magnitude data at the corresponding wind farm site.

II. Raw wind data properties

TABLE I
RAW DATA CLASSIFICATION

Category			Description
Valid			data with normal wind speed and power values, reflecting the natural properties of the wind power curve
Invalid	Incorrect	Missing	data without wind speed or wind power values
		Constant	data with constant wind speed and wind power values
		Exceeding	data with wind speed or wind power values exceeding physical limits
		Irrational	data with valid wind speed and wind power values but illogical in physics
	Unnatural		data with low wind power output at high wind speed periods

Table II

TABLE II
A SAMPLE OF INVALID WIND DATA

Category	Time	Wind Speed (m/s)	Wind Power (kW)
Missing	2012/11/21 12:00:00	4.2569	Not a Number
	2012/11/21 12:15:00	4.8924	Not a Number
	2012/11/21 12:30:00	5.3696	Not a Number
	2012/11/21 12:45:00	5.8469	Not a Number
	2012/11/21 13:00:00	6.3242	Not a Number
Constant	2013/01/31 20:00:00	10.0899	120150.5204
	2013/01/31 20:15:00	10.0899	120150.5204
	2013/01/31 20:30:00	10.0899	120150.5204
	2013/01/31 20:45:00	10.0899	120150.5204
	2013/01/31 21:00:00	10.0899	120150.5204
Irrational	2013/03/19 18:00:00	0.0000	133676.0860
	2013/03/19 18:15:00	0.0000	139648.4616
	2013/03/19 18:30:00	0.0000	144727.4801
Exceeding	2013/04/24 19:00:00	13.3746	999999.9999
	2013/04/24 19:15:00	14.0320	-999999.9999

Figure 1

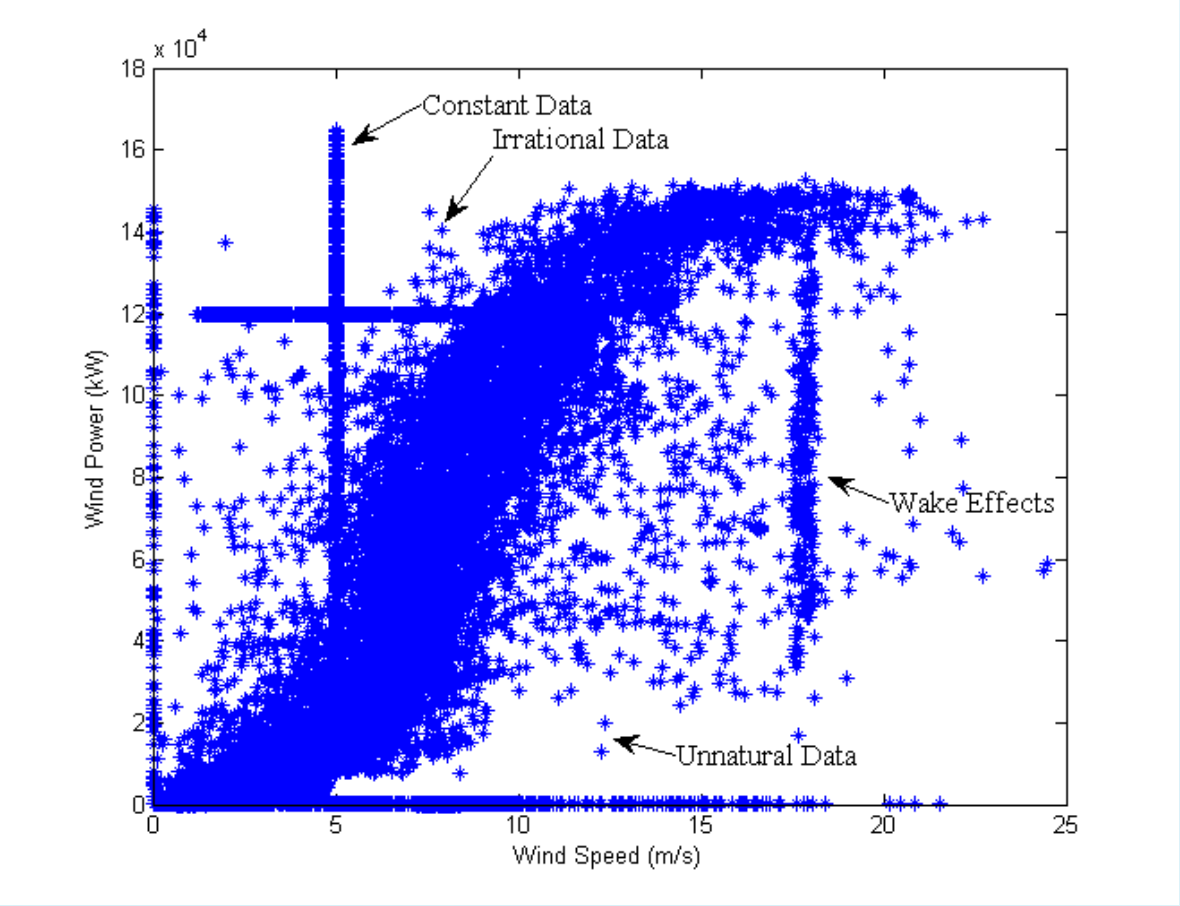

Figure 1. Raw scatter plot of wind farm output and wind speed

Figure 2

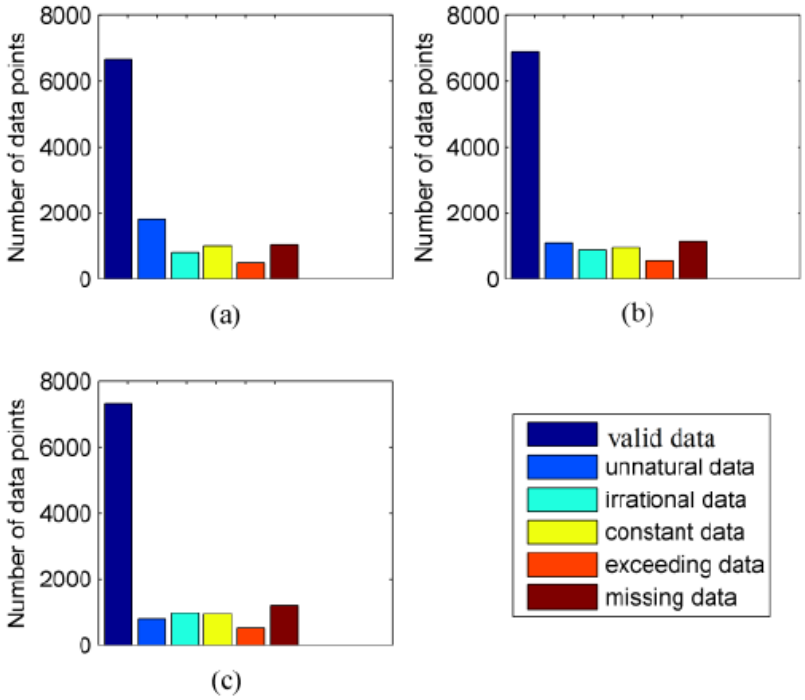

Fig. 2. Distribution of raw wind data: (a) period from 10/1/2010 to 1/31/2011; (b) period from 2/1/2011 to 5/31/2011; (c) period from 6/1/2011 to 9/30/2011.

Figure 2. Distribution of raw wind data

Section III

III. The Preprocessing Methodology

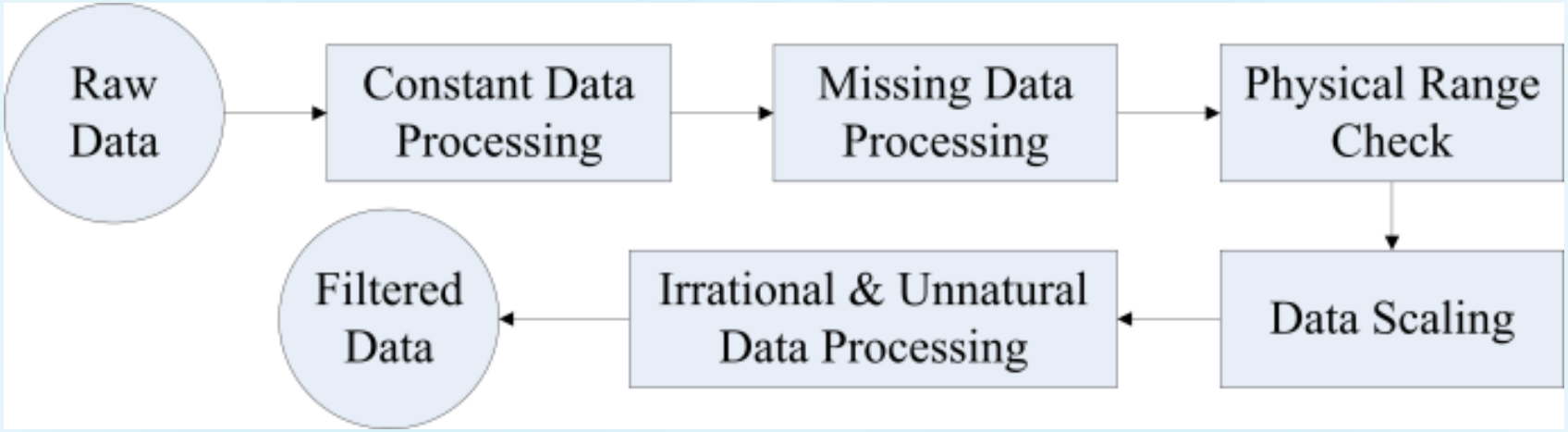

Figure 3. Structure of preprocessing system

Table III

TABLE III

CONSTANT DATA PROCESSING ALGORITHM

Algorithm: Constant data processing
Input: *R*, the raw wind database shown in Fig. 1, sorted by time stamp
Output: *MEIUV*, the database after excluding the constant data, consisting of the missing, exceeding, irrational, unnatural, and valid data
Method:

```
for (k=1; k≤length(R)-1; k++){
   if (data.speed[k]==data.speed[k+1] || data.power[k]==data.power[k+1])
   then flag[k]=0;
   else flag[k]=1;
}
for (k=1; k≤length(R); k++){
   if (flag(k)==0)
   then delete data[k];
   else add data[k] to MEIUV;
}
return MEIUV;
```

Table IV

TABLE IV
MISSING DATA PROCESSING ALGORITHM

Algorithm: Missing data processing
Input: *MEIUV*, the database after excluding the constant data
Output: *EIUV*, the database after excluding the constant and missing data, consisting of the exceeding, irrational, unnatural, and valid data
Method:

```
for (k=1; k⩽length(MEIUV); k++){
    if (data.speed[k]==NaN || data.power[k]==NaN)
    then delete data[k];
    else add data[k] to EIUV;
}
return EIUV;
```

Table V

TABLE V
EXCEEDING DATA PROCESSING ALGORITHM

Algorithm: Physical range check
Input: *EIUV*, the database after excluding the constant and missing data
Output: *IUV*, the database after excluding the constant, missing, and exceeding data, consisting of the irrational, unnatural, and valid data
Method:

```
for (k=1; k⩽length(EIUV); k++){
   if (data.speed[k] ∈ speed_range && data.power[k] ∈ power_range)
   then add data[k] to IUV;
   else delete data[k];
}
return IUV;
```

Figure 4

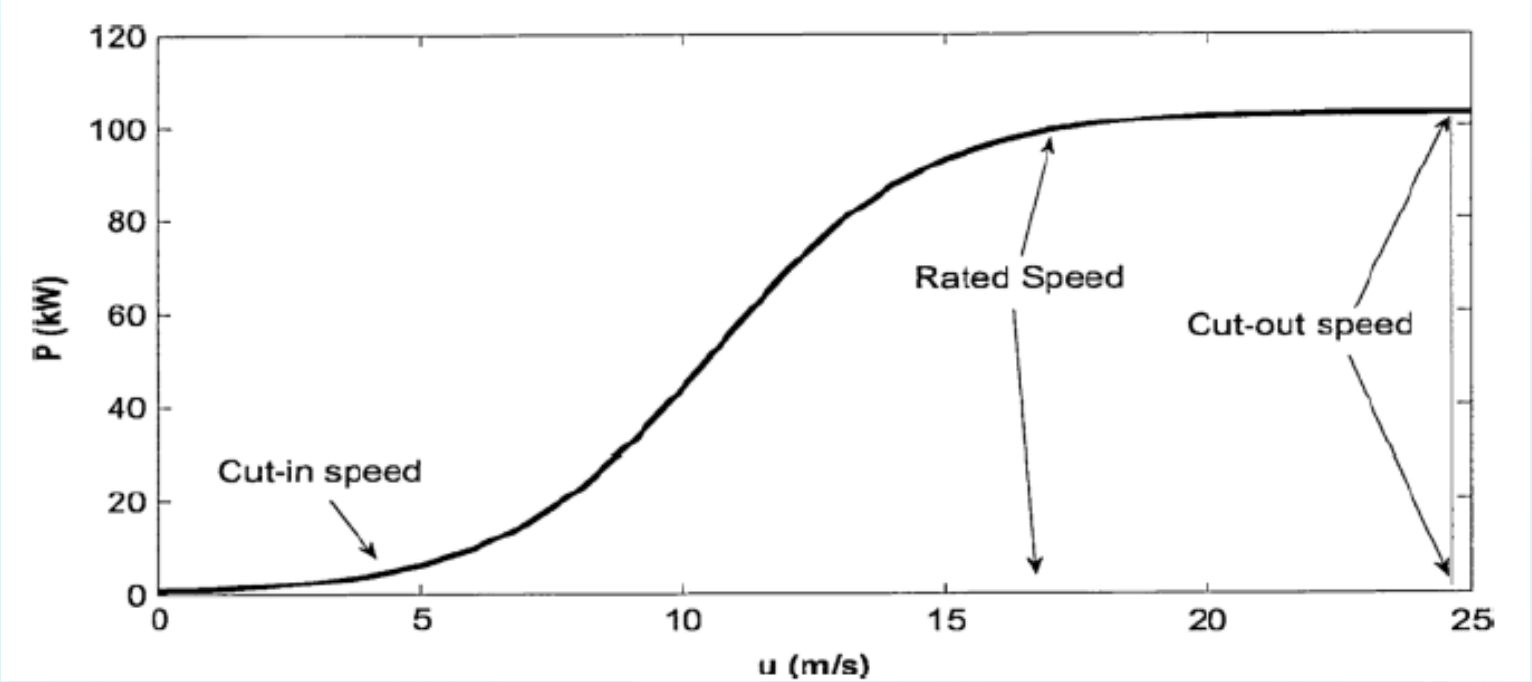

Figure 4. Wind turbine power curve

IV. Uncertainty Management

$$variance = \frac{(N_{cubic} - N_{common}) + (N_{linear} - N_{common})}{N_{common}} \times 100 \qquad (13)$$

Section V

V. Test Results and Discussion

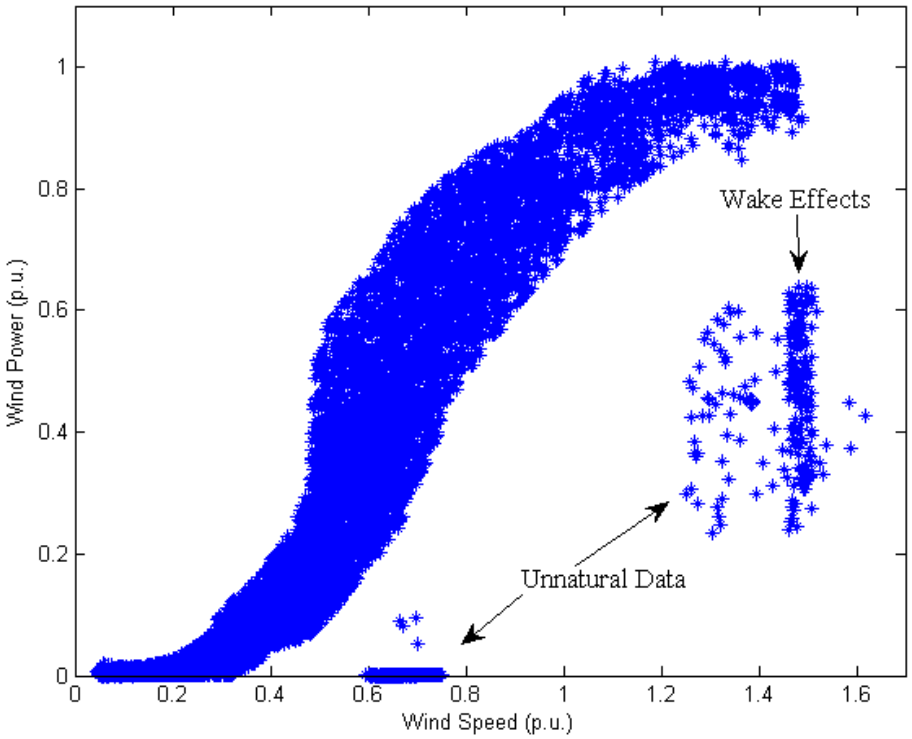

Figure 5. Filtered scatter plot of wind data with the Euclidean distance

Figure 6

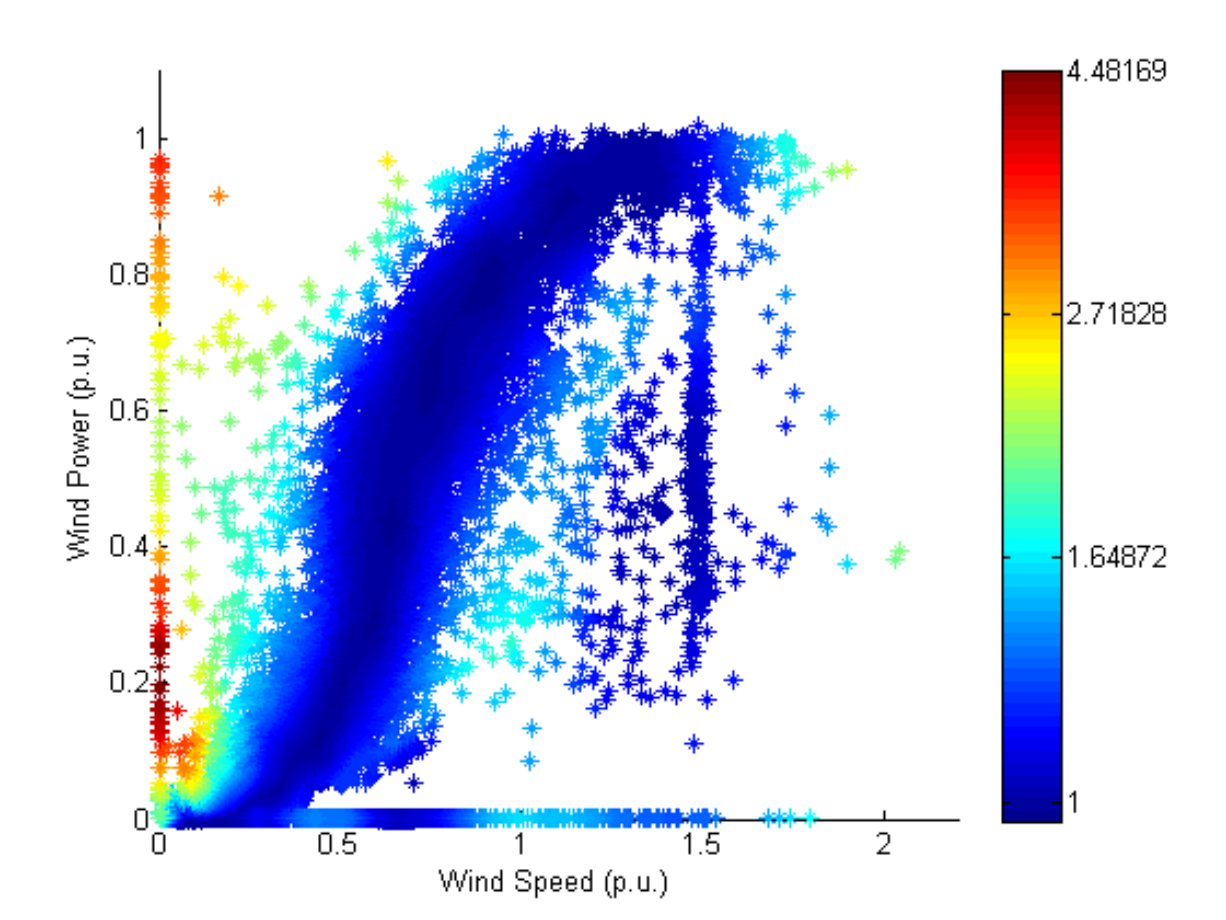

Figure 4. Outlier factors by LOF

Table VI

TABLE VI
RESULTS OF VARIOUS TUNING PARAMETERS

Tuning Parameter	Cubic Approximation	Linear Approximation	Common Data	Bias + Variance
0.5	Fail	Fail	-	∞
0.6	Fail	Fail	-	∞
0.7	9775	9661	9299	9.007
0.8	9702	9584	9186	9.958
1.0	9546	9408	8911	12.70
1.2	9331	9194	8608	15.21

Figure 7

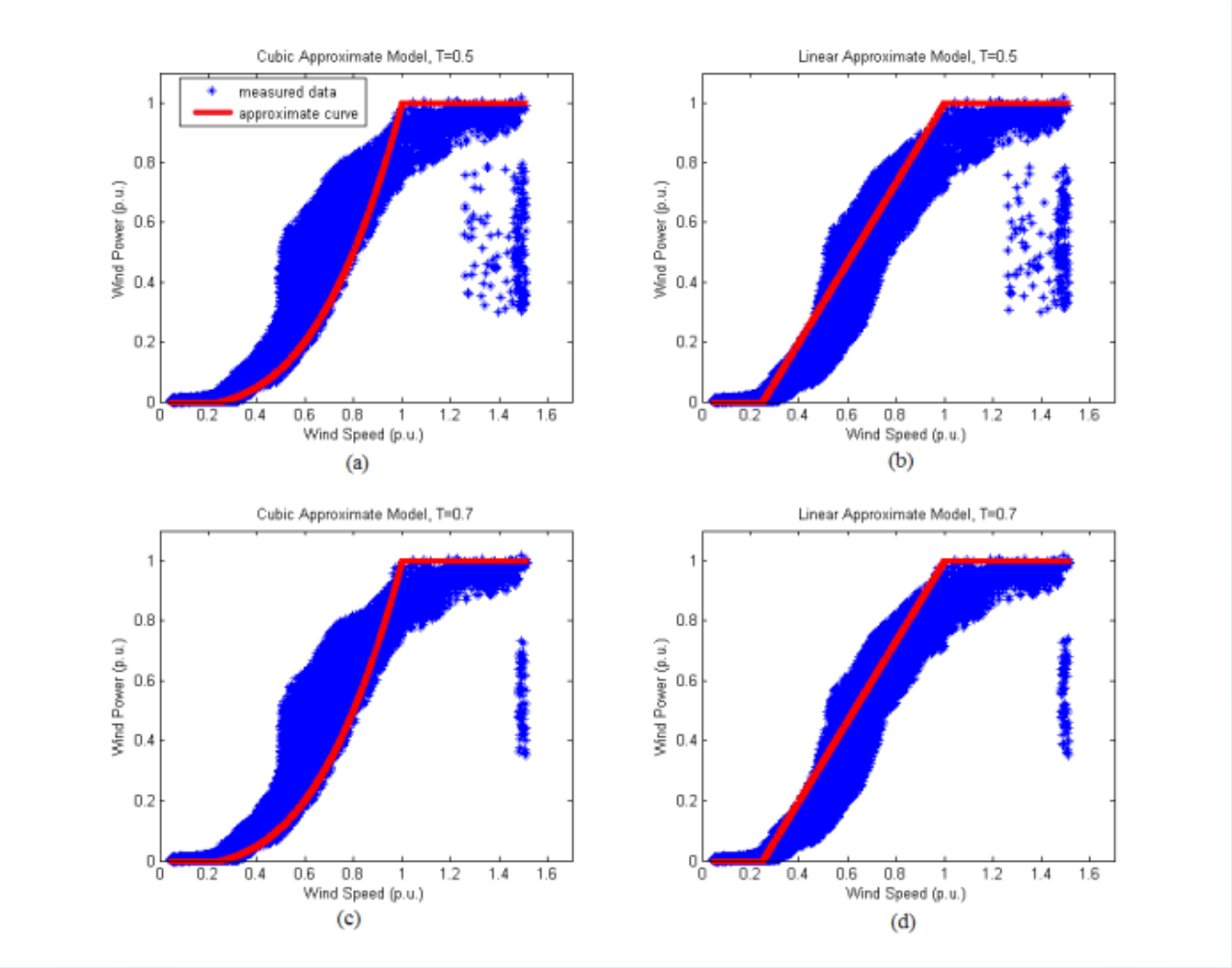

Figure 7. Filtered scatter plot of wind data with the weighted distance