The importance of data availiability
Effects of missing or erroneous data

PO.ID
263

AQSystem
Remote sensing technique

Lasse Johansson

AQ System AB, Göteborg/Tyresö, Sweden

Abstract

To achieve certain knowledge we need correct and complete observations. If we have one year worth of such wind observations, we know the mean wind speed that year correctly, and that all other estimates of it have larger errors. In reality we are faced with non-perfect data and the *observers dilemma:*

- Erroneous observations give a bias and must thus be removed.
- Removed and missing observations give a bias.

Finding and removing erroneous observations is difficult. It is however necessary since erroneous observations will have a detrimental influence on wind resource estimates. Even when it can be done successfully, the error due to removing data must be estimated in order to assess if the remaining data can be used to make estimates of the wind resource.

We explain some features of bias and uncertainty here.

Data and model

The *de-facto* standard of ten minute wind speed averages are not independent of one another in a statistical sense. Let us therefore study chunks of wind speed long enough to be *reasonably* independent of one another and also of a size typical of data loss periods; months. Unheated anemometers in cold climate typically are down 10-30 % i.e. a few months per year. Remote sensors can have data losses of 1-20 % per year.

We assume the following simple description of the wind:

monthly wind = constant annual mean + seasonal part + random part.

This model is illustrated in Figure 1.

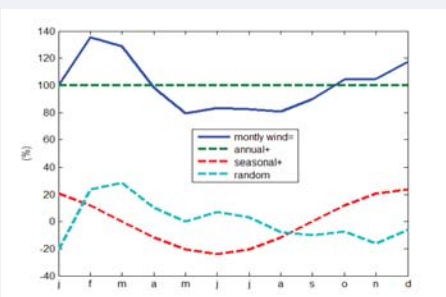

Figure 1. An example of one year in monthly means consisting of the annual mean + the seasonal part + the random part. The mean wind has been set = 100.

Seasonal bias

If the data loss is concentrated to a specific season there will be a bias to the annual mean. The bias can be estimated from the data if we have enough of it, at least 6-8 months. In Figure 2 we can see this seasonal bias due to missing data for our model wind. Clearly, the bias will develop differently depending on when the data hole starts. If the data holes are distributed evenly over the year, the seasonal bias is insignificant. Measurements extending through a part of a year can be viewed as seasonally biased. Seasonal bias is a case of statistical conditioning below.

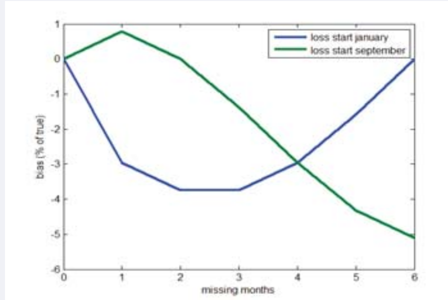

Figure 2. Bias due to data loss and the seasonal variation. If the data loss starts in a windy season the bias grows negative quickly with the length of the missing period.

Uncertainty

The random part of the wind results in an uncertainty in the annual mean. It can be expressed for instance as the standard deviation of the mean, $s = \sigma/\sqrt{12}$, if there are twelve months of measurements and σ is the standard deviation of the ensemble of monthly averages – preferably after the seasonal component has been subtracted. The value s is the lower limit of uncertainty; it is meaningless to specify the annual average claiming a higher precision than this. If one or more months of data are missing, the uncertainty will be greater. It will increase with data loss irrespective of how the loss is distributed; one long loss or spread out over the measurement period.

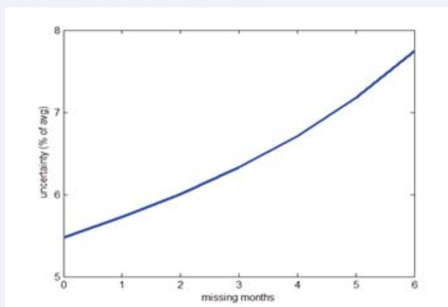

Figure 3. The uncertainty increases with number of months missing.

Conditioning

Conditioning occurs when the data loss depends on what is being observed. There is direct and indirect conditioning. An example of direct conditioning would be a rain gauge that only works in sunshine. It is strongly conditioned and produces a biased rain estimate. An example of indirect conditioning is an anemometer that freezes and stops at temperatures below zero. If the wind is as strong in cold as in warm weather, there is no bias. But if the wind speed is weaker in cold than in warm weather, the effect will be that the anemometer collects stronger wind and the average is biased. Conditioning can also appear through technical problems. A remote sensor which works poorly or stops in certain weather, be it rain, fog or snow can produce biased wind speed observations.

In Figure 4 we can see an example of how conditioning on temperature gives a bias in wind speed. At this location, there is more wind around 0ºC than at warmer and colder weather. Thus, if there is data loss when the temperature is a few degrees below zero, there is a pronounced bias. The bias decreases rapidly if the data loss is limited to below –10ºC.

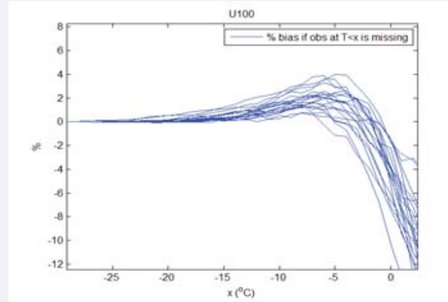

Figure 4. The bias if observations below a certain temperature are missing. Different curves describe 22 years. Example 1: If observations below 0ºC are missing, the mean wind speed will be underestimated by 1 to 11 %. Example 2: If observations below –10ºC are missing, the mean wind speed will be overestimated by 1 to 3 %.

If the erroneous data is not found and removed, the bias rapidly becomes catastrophic. In Figure 5, we can see the effect when erroneous zero-wind observations are not removed.
Zero or constant observations are easy to identify and delete. Other erroneous wind speed values can be very hard to find.

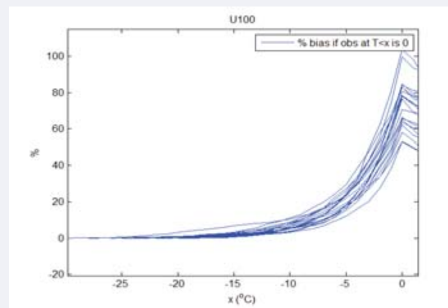

Figure 5. The bias if observations below a certain temperature are erroneously zero. Different curves describe 22 years. Example 1: If observations below 0ºC are missing, the mean wind speed will be overestimated by 50 to 100 %. Example 2: If observations below –10ºC are missing, the mean wind speed will be overestimated by 3 to 10 %.

Conclusions

- To achieve certain knowledge we need correct and complete observations. If we are not this fortunate, we need to quantify the problems.
- Erroneous and missing data can introduce bias and will increase the uncertainty of the wind resource estimate.
- Erroneous data are most dangerous. A small amount of them could ruin the quality of the measurement result.
- Generally, erroneous data is a greater problem than missing data. Thus, it is better to delete suspicious data and deal with the hole than to risk adverse effects of keeping bad data.
- A long data hole introduces a seasonal bias which can be estimated.
- Data loss causes increased uncertainty when it is concentrated to one episode as well as when it is spread out in many smaller holes. The increased uncertainty can be estimated.
- If the probability of obtaining an observation depends on the weather, for example through technical systems such as power supply, icing or air quality, there will be bias through statistical conditioning. This bias depends strongly on the circumstances during the particular measurement period. This bias should be estimated to ensure that it does not deteriorate the wind assessment reliability**.**