Low Probability Events

Cristian Vava, PhD

This white paper will discuss some important issues related to the prediction of low and very low probability events and possible solutions Innovatorium has developed based on Heuristic Analytics.

To avoid some daily language inaccuracies we'll adopt the convention that a low probability event has a chance of occurrence of less than 1%. When necessary we may even distinguish the very low probability events as having a chance of occurrence of less than 0.1%.

The most familiar examples of low probability events are lottery winning, economic depression, and insurable events like flooding or earthquakes. However we are surrounded by many low probability events that we choose not to care about since their effects are irrelevant to us. In practice we usually want to know the statistical distribution of these events and also a cost function which describes their effects.

Analyzing low probability events is difficult since we need huge volumes of data and to use statistical distributions different from the ones we use in common applications. Moreover estimating parameters is also a very complex task at least because statistical parameters are usually derived based on the maximum likelihood method which leads to unreliable results if there is a significant mismatch between the selected and the empirical distribution.

Below we'll use as example the flooding of a small Floridian community. From historic records during the 12/01/1930 to 12/01/2010 period we found 14 cases of flooding. This represents a probability of flooding of 0.049% or an average likelihood of once every 2060 days. From this number of occurrences it is practically impossible to build a reliable predictor and collecting more data is not an option since it may take over 1500 years, a perfect case of unknowable.

From the Heuristic Analytics perspective the solution to this dilemma is to build an Observer with predictive power. What we have to do is to find other measurable parameters and a formula to combine the values of these parameters to describe the parameter(s) of interest. Obviously the new observer has predicting power if it correctly predicts all known events and doesn't falsely predict inexistent events. An Observer represents an indirect measurement and is used in many fields from engineering, medicine, to social studies. From medicine a well known example of observer was developed by Dr. Lee Goldman from Harvard for quick selection of patients with heart attack. A much more exotic example is the Facial Action Coding System (FACS) known as the taxonomy of facial expressions.

Today every field of activity is dominated by the use of observers and a good measure of advancements in that field is given by the complexity of available observers. The main reason for using observers is to replace the hard to obtain direct value with a set of easier and more accurate direct measurements plus a formula or process to convert the known values into a result with predictive power.

In the particular case of our study building the basis for an observer was an easy task since we have available historic measurements of the water level in a nearby creek. And since flooding and the water in the creek had the same source it is easy to build a mental connection justifying the observer. In general to build the flooding observer you need to analyze very large volumes of data and know very well the local geography.

Innovatorium ©2011

The next step involves choosing a statistical distribution and finding parameters to match empirical data thus validating the observer. Although it is not very difficult to search among various distributions until one matches close enough the empirical distribution, finding its parameters could be a very challenging task. The main reason is that standard formulas are based on maximum likelihood which is unable to correctly predict the parameters when the empirical distribution is far from the assumed one. Figure 1 at the end of this paper shows the empirical and the Pearson IV Probability of Distribution Functions (PDF) with parameters derived using the maximum likelihood method.

At Innovatorium we have built other estimators based on the Best Fit generating acceptable parameters irrespective of the empirical distribution. Figure 2 shows a comparison of the empirical distribution and several other distributions like Weibull, Log-Normal, Pearson IV, and Chi Square with parameters based on our algorithms. These approximations are much closer to the empirical distribution but are still not very good at predicting the very low probability events we care about. While for common applications predicting errors of less than 0.25% are very good for low probability events it is totally unacceptable. Figure 3 exemplifies this by showing a detail of the same distributions as above but restricted to water levels of 14 ft and over.

At Innovatorium we have designed statistical distributions of higher order able to better fit the needs of these very low probability events. Figure 4 shows a detail within the same range of 14 ft and over using a proprietary distribution labeled Optimal Estimator. Using the same algorithm we may identify similar estimators describing also the expected limits. These estimators predict that flooding occurs when water level is over 19.5 ft which happens with a probability of 0.045% close enough to the true value of 0.049% thus giving enough predicting power to our observer. On average the water level is 19.5 ft and over every 2219 days which again is close enough to the true value of 2060 days. This estimator predicts a 0.00206% probability for water level to be within 19.5-19.75 ft and a 0.00071% probability for the range 22.0-22.25 ft, both at orders of magnitude higher than what normal distribution predicts for these extreme cases.

Mathematically the whole model is defendable. For example the existence of these bounds was predicted by the Gartner-Ellis theorem of large deviations for weakly dependent sequences. On the other side from the Predictive Analytics the most common solutions are based on Poisson processes relying on the hard to validate assumption of independence.

By having such a detailed description of the most likely probability and its expected range we could have a much cleaner description of the risk therefore we could optimize insurance premiums. The insurance agency could really differentiate customers that have implemented mitigation measures involving for example raising the ground even at the inch level thus providing incentives to reduce the cost of the entire system by reducing the losses.

Conclusions

Low probability events can be predicted conveniently and accurately enough using observers built from background knowledge of the real events and data analysis.

Innovatorium has designed techniques to build convenient observers, special statistical distributions, and formulas to estimate their parameters capable to reach accuracies many times an order of magnitude better than traditional ones.

Figure 1

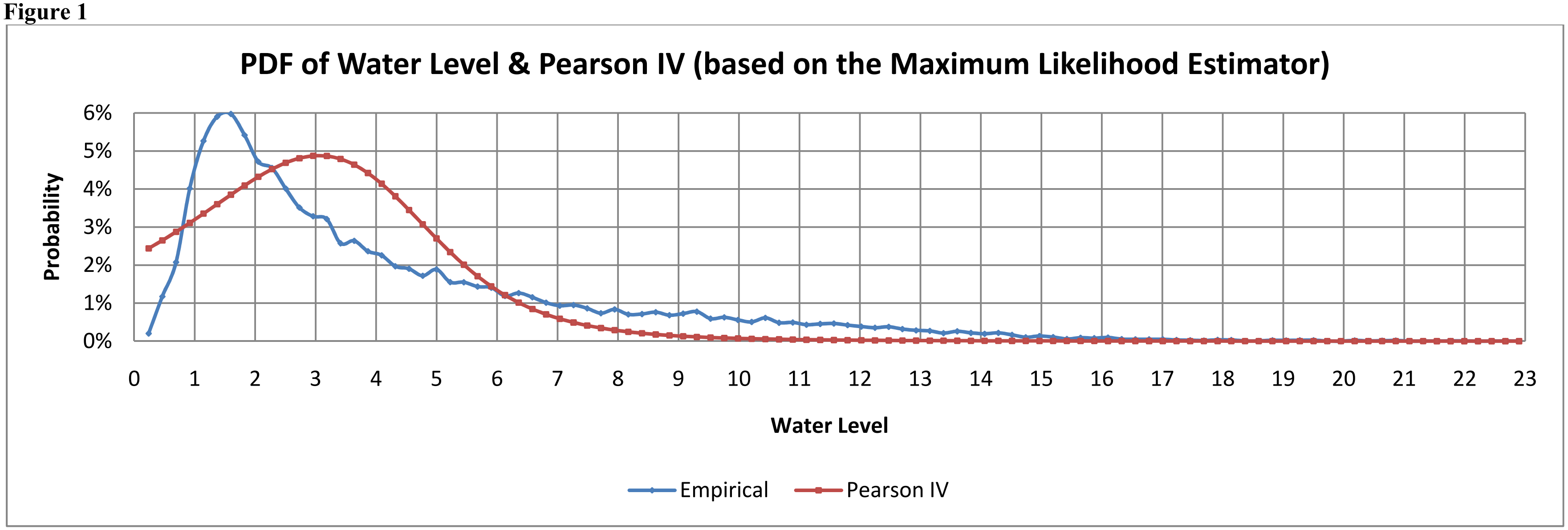


Figure 2

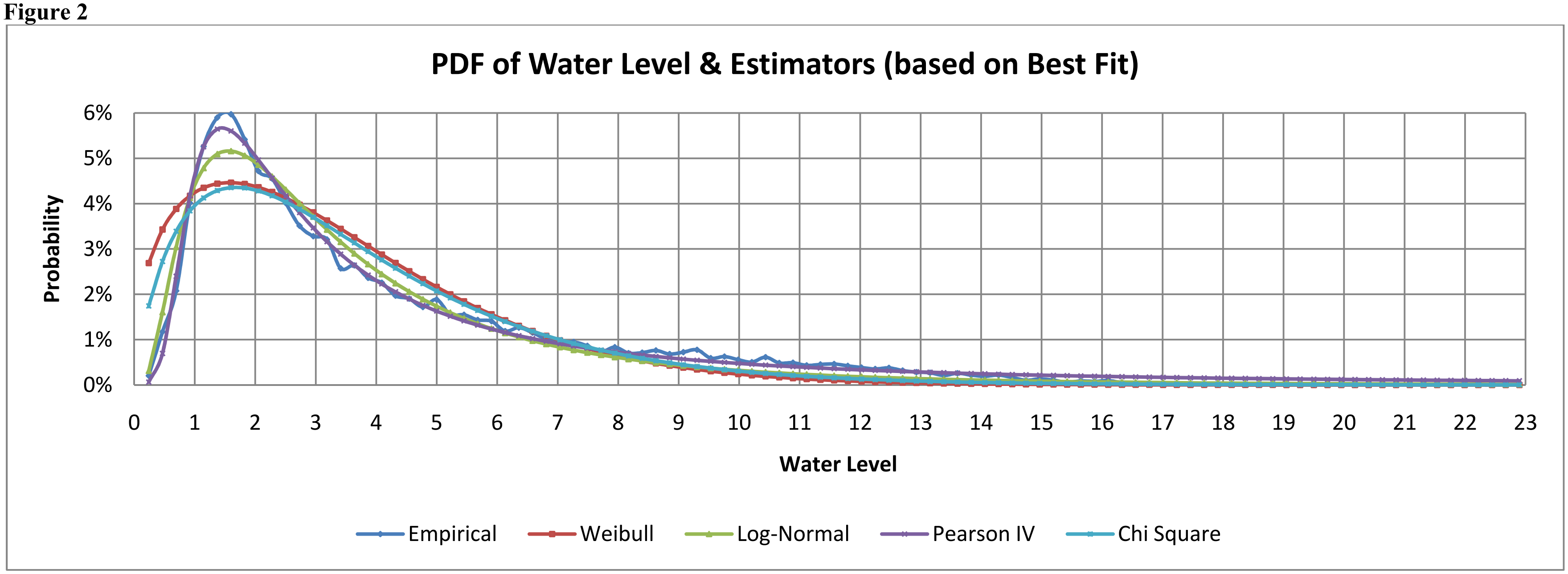


Figure 3

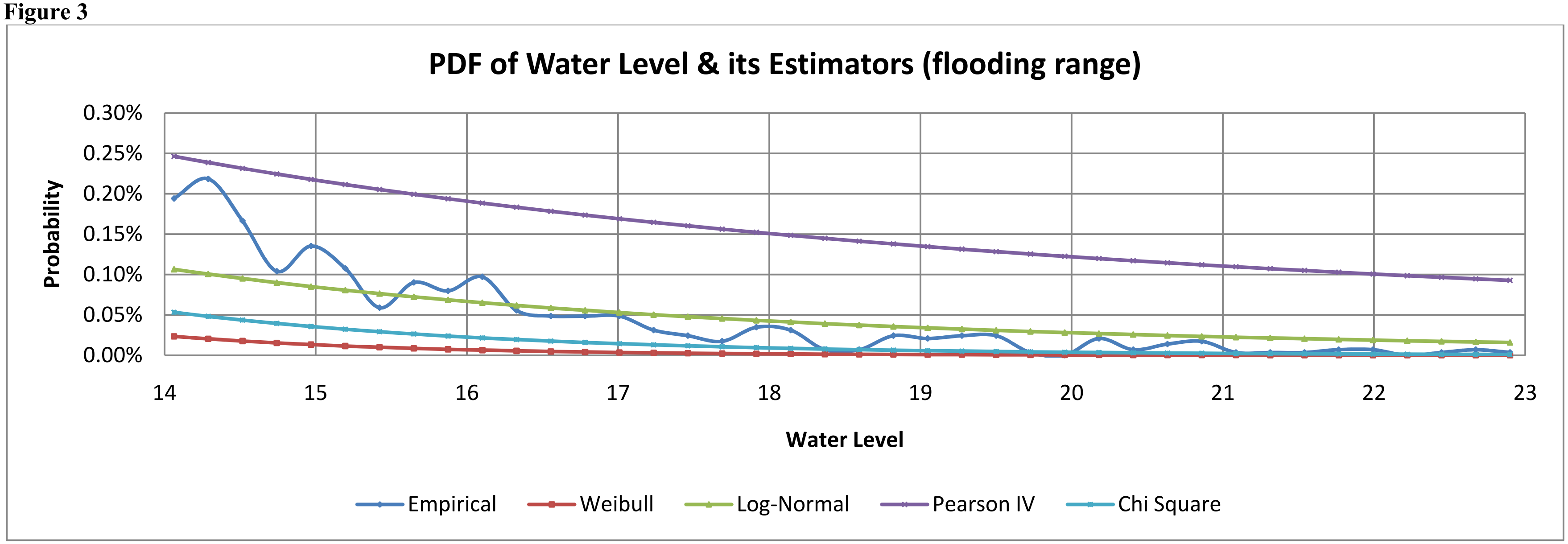
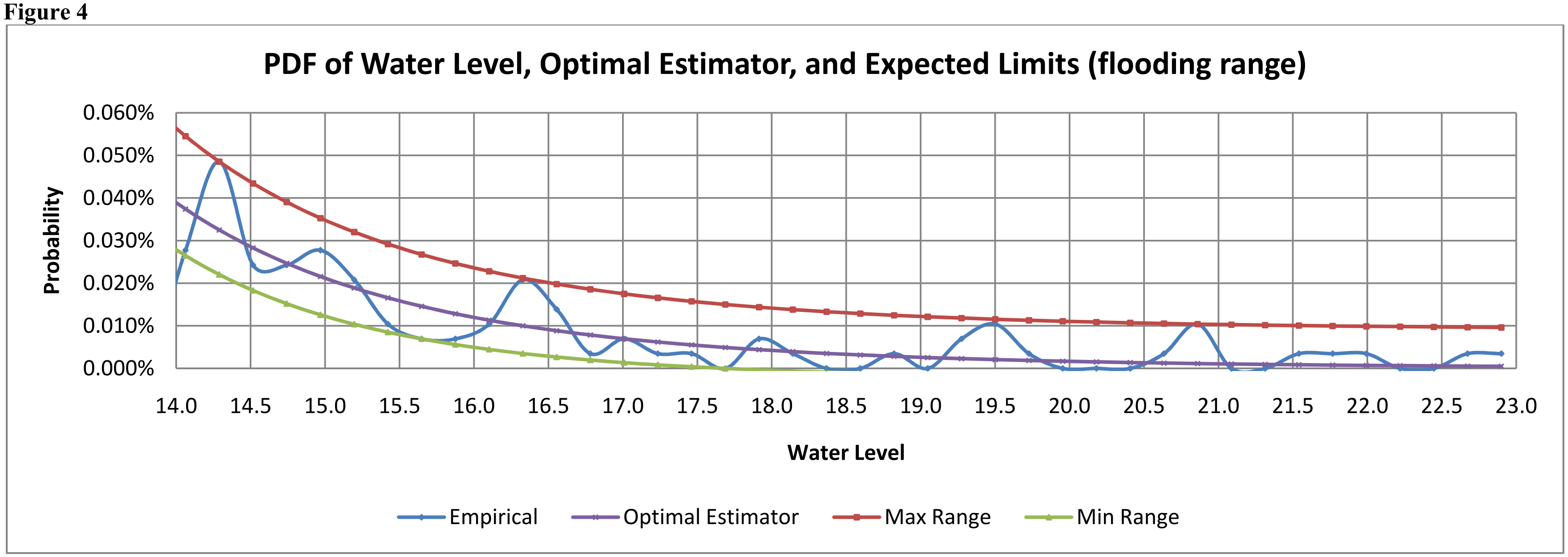


Figure 4


PDF of Water Level, Optimal Estimator, and Expected Limits (flooding range)
Probability
Water Level
Empirical
Optimal Estimator
Max Range
Min Range