Documentation of QIBA PDF evaluation tool (underdevelopment)

* The choosing of evaluation standards.

The choosing of evaluation standards:

* (Graphics)**Scatter plot** to show the relationship of the calculated values with reference values
* (Graphics)**Box plot** to show the distribution of calculated value in each patch
* (Statistics)**Normality test.** To test whether a set of value follows normal distribution or not. This is the precondition of Student t-test.
  + If a set of data is not normally distributed, then this model is believed to work poorly with this set of K/V. A big minus for the final evaluation score.
  + If a set of data is normally distributed, then it can be described with mean value and deviation. Mean value can be used to evaluate the fitting with reference value. And the deviation can be a weighting factor for the evaluation score.
* (Statistics)**Student’s t-test**. For this application, one group of data will be tested, to see whether the mean is a certain value or not.
* (Statistics)**Mann-Whitney-U test**. It is a [nonparametric](http://en.wikipedia.org/wiki/Nonparametric_statistics) [test](http://en.wikipedia.org/wiki/Statistical_hypothesis_test) of the [null hypothesis](http://en.wikipedia.org/wiki/Null_hypothesis) that two populations are the same against an [alternative hypothesis](http://en.wikipedia.org/wiki/Alternative_hypothesis), especially that a particular population tends to have larger values than the other.
  + Maybe not so useful here? Because there’s no need to simply decide whether two group are the same or not, but which is better compared to the reference values in some standard.
* **Person correlation coefficient** is a value to measure the linear correlation between two groups of values, valued between -1 and 1.
  + Can be used to evaluate the correlation between calculated and reference value sets
* **Spearman’s rank correlation coefficient**. It assesses how well the relationship between two variables can be described with monotonic function.
  + This maybe not so useful. Because it’ll be too complicated with they so detailed processes. If linear regression is applied, then this is not going to be used.
* **Analysis of covariance**. It measures how much two variables change together and how strong the relationship is between them.
  + ANCOVA can be used to evaluate how much the performance (e.g. deviation of a patch) varies according to the parameters, i.e. Ktrans and Ve.(This might be meaningful to direct the user of a certain model to choose the reasonable parameters.)
* **Linear regression**.
  + Can be used to abstract the artificial slope and intercept of the model, with which the calculated data can be uniformed so that the evaluations have the same baseline.
  + The R value can be used to assess how well the data align linearly.
* **Analysis of variance**. It is used to analyze the differences between group means and their associated procedures.

The description of the evaluation score system:

* The system is to make sure that there’s only one score for one tested model (software package), so that the comparison and ranking among models are possible.
* The score system will be established relying on the statistics analysis of the calculated results.
* Different score standard should be available, so that diverse evaluation could be applied.

The basement of the score system:

* + The score goes down from a base score. Each influence factor can only lower the score. So finally the lower the score, the poorer the model performs.
  + Deviation for each patch. This can be used to evaluate how concentrated of the calculated parameter (factor a, changeable, optional).
  + Mean value (if normally distributed) or median value (if not) can be used to represent the patch. Whatever test we use, the point is to find out one value to represent the patch, so that the comparison dimension can reduce one.
  + For one parameter, e.g. Ktans = 0.01, there are 6 patches corresponding to the 6 parameters of Ve. For this set values, covariance analysis can be used to see the whether the calculated results have anything with the other parameter (factor b, changeable, optional).
  + Pick up one of these values to represent the column (or row). It can be mean value or median value. The idea is the same as for patches.
  + Linear regression of the columns (or rows). A very general standard to see how accurate the calculated values are (changeable, optional).

Score = sum of e^(each patch’s diffusion)\*a ,{each patch, as a basic score}

-sum of some operation with covariance\*b, {each row or column, covariance with the other parameter}

- sum of square of error\*c, {each column or row, how accurate the calculated results are}

The factor a, b, c are changeable to have different weighting.