In this document we show the full derivation of the MSPD equality shown in the paper “Generalising Random Forest Parameter Optimisation to Include Stability and Cost” by CHB Liu, BP Chamberlain, DA Little and Â Cardoso (2017).

The equality shows MSPD is a empirical estimate which captures the relationship between the variance of the random forest model, and the covariance of the predictions due to sampling of training data, as shown in the paper:

$$
\begin{aligned}
\mathrm{MSPD}(f) &= \frac{2}{R(R-1)} \sum_{j=1}^{R} \sum_{k=1}^{j-1} \left[\frac{1}{N} \sum_{i=1}^{N} \left[\left(\hat{y}_i^{(j)} - \hat{y}_i^{(k)} \right)^2 \right] \right] \\
&= \frac{2}{N} \sum_{i=1}^{N} \left[\frac{1}{R-1} \sum_{l=1}^{R} \left(\hat{y}_i^{(l)} - \mathbb{E}(\hat{y}_i^{(\cdot)}) \right)^2 \right. \\
&\qquad \left. - \frac{1}{R(R-1)} \sum_{j=1}^{R} \sum_{k=1}^{R} \left(\hat{y}_i^{(j)} - \mathbb{E}(\hat{y}_i^{(\cdot)}) \right) \left(\hat{y}_i^{(k)} - \mathbb{E}(\hat{y}_i^{(\cdot)}) \right) \right] \\
&= 2\mathbb{E}_{x_i}[\mathrm{Var}(f(x_i)) - \mathrm{Cov}(f_j(x_i), f_k(x_i))],
\end{aligned}
$$

where $\mathbb{E}_{x_i}$ is the expectation over all validation data, f is a mapping from a sample x_i to a label $\hat{y}_i$ on a given run, $\mathrm{Var}(f(x_i))$ is the variance of the predictions of a single data point over model runs, and $\mathrm{Cov}(f_j(x_i), f_k(x_i))$ is the covariance of predictions of a single data point over two model runs.

The Derivation

We begin with the definition of MSPD:

$$
\frac{2}{R(R-1)} \sum_{j=1}^{R} \sum_{k=1}^{j-1} \left[\frac{1}{N} \sum_{i=1}^{N} \left[\left(\hat{y}_i^{(j)} - \hat{y}_i^{(k)} \right)^2 \right] \right]. \tag{1}
$$

Rearranging the summations, and introducing two opposite signed expectations of $\hat{y}_i^{(\cdot)}$ (which cancels out each other) in the squared term:

$$
\frac{1}{N} \sum_{i=1}^{N} \left[\frac{2}{R(R-1)} \sum_{j=1}^{R} \sum_{k=1}^{j-1} \left[\left(\left(\hat{y}_i^{(j)} - \mathbb{E}(\hat{y}_i^{(\cdot)}) \right) - \left(\hat{y}_i^{(k)} - \mathbb{E}(\hat{y}_i^{(\cdot)}) \right) \right)^2 \right] \right] \tag{2}
$$

Expanding the squared term:

$$\frac{1}{N}\sum_{i=1}^{N}\left[\frac{2}{R(R-1)}\sum_{j=1}^{R}\sum_{k=1}^{j-1}\left[\left(\hat{y}_i^{(j)}-\mathbb{E}(\hat{y}_i^{(\cdot)})\right)^2 + \left(\hat{y}_i^{(k)}-\mathbb{E}(\hat{y}_i^{(\cdot)})\right)^2 - 2\left(\hat{y}_i^{(j)}-\mathbb{E}(\hat{y}_i^{(\cdot)})\right)\left(\hat{y}_i^{(k)}-\mathbb{E}(\hat{y}_i^{(\cdot)})\right)\right]\right] \tag{3}$$

Pushing the inner double summation into the expanded square expression:

$$\frac{1}{N}\sum_{i=1}^{N}\left[\frac{2}{R(R-1)}\sum_{j=1}^{R}\sum_{k=1}^{j-1}\left[\left(\hat{y}_i^{(j)}-\mathbb{E}(\hat{y}_i^{(\cdot)})\right)^2 + \left(\hat{y}_i^{(k)}-\mathbb{E}(\hat{y}_i^{(\cdot)})\right)^2\right] - \frac{2}{R(R-1)}\sum_{j=1}^{R}\sum_{k=1}^{j-1}\left[2\left(\hat{y}_i^{(j)}-\mathbb{E}(\hat{y}_i^{(\cdot)})\right)\left(\hat{y}_i^{(k)}-\mathbb{E}(\hat{y}_i^{(\cdot)})\right)\right]\right] \tag{4}$$

Notice the first inner double summation can be reindexed and collapsed into a single summation as follow:

$$\frac{1}{N}\sum_{i=1}^{N}\left[\frac{2}{R(R-1)}\sum_{l=1}^{R}\left[(R-1)\left(\hat{y}_i^{(l)}-\mathbb{E}(\hat{y}_i^{(\cdot)})\right)^2\right] - \frac{2}{R(R-1)}\sum_{j=1}^{R}\sum_{k=1}^{j-1}\left[2\left(\hat{y}_i^{(j)}-\mathbb{E}(\hat{y}_i^{(\cdot)})\right)\left(\hat{y}_i^{(k)}-\mathbb{E}(\hat{y}_i^{(\cdot)})\right)\right]\right] \tag{5}$$

Disassembling the $(R-1)$ term:

$$\frac{1}{N}\sum_{i=1}^{N}\left[\frac{2}{R(R-1)}\sum_{l=1}^{R}\left[R\left(\hat{y}_i^{(l)}-\mathbb{E}(\hat{y}_i^{(\cdot)})\right)^2\right] - \frac{2}{R(R-1)}\sum_{l=1}^{R}\left[\left(\hat{y}_i^{(l)}-\mathbb{E}(\hat{y}_i^{(\cdot)})\right)^2\right] - \frac{2}{R(R-1)}\sum_{j=1}^{R}\sum_{k=1}^{j-1}\left[2\left(\hat{y}_i^{(j)}-\mathbb{E}(\hat{y}_i^{(\cdot)})\right)\left(\hat{y}_i^{(k)}-\mathbb{E}(\hat{y}_i^{(\cdot)})\right)\right]\right] \tag{6}$$

Further simplifying the first inner summation, and notice the second inner summation and the inner double summation can be combined due to symmetry of the terms in the inner double summation:

$$\frac{1}{N}\sum_{i=1}^{N}\left[\frac{2}{(R-1)}\sum_{l=1}^{R}\left[\left(\hat{y}_i^{(l)}-\mathbb{E}(\hat{y}_i^{(\cdot)})\right)^2\right] - \frac{2}{R(R-1)}\sum_{j=1}^{R}\sum_{k=1}^{R}\left[\left(\hat{y}_i^{(j)}-\mathbb{E}(\hat{y}_i^{(\cdot)})\right)\left(\hat{y}_i^{(k)}-\mathbb{E}(\hat{y}_i^{(\cdot)})\right)\right]\right] \tag{7}$$

Notice the first term within the outer summation is two times the unbiased estimate of the sample variance, and the second term is two times the average sample covariance between all prediction pairs. The outer summation can also be seen as the expected value across all data points. We conclude by writing expression 7 as an estimate of:

$$2\mathbb{E}_{x_i}[\mathrm{Var}(f(x_i)) - \mathrm{Cov}(f_j(x_i), f_k(x_i))] \tag{8}$$