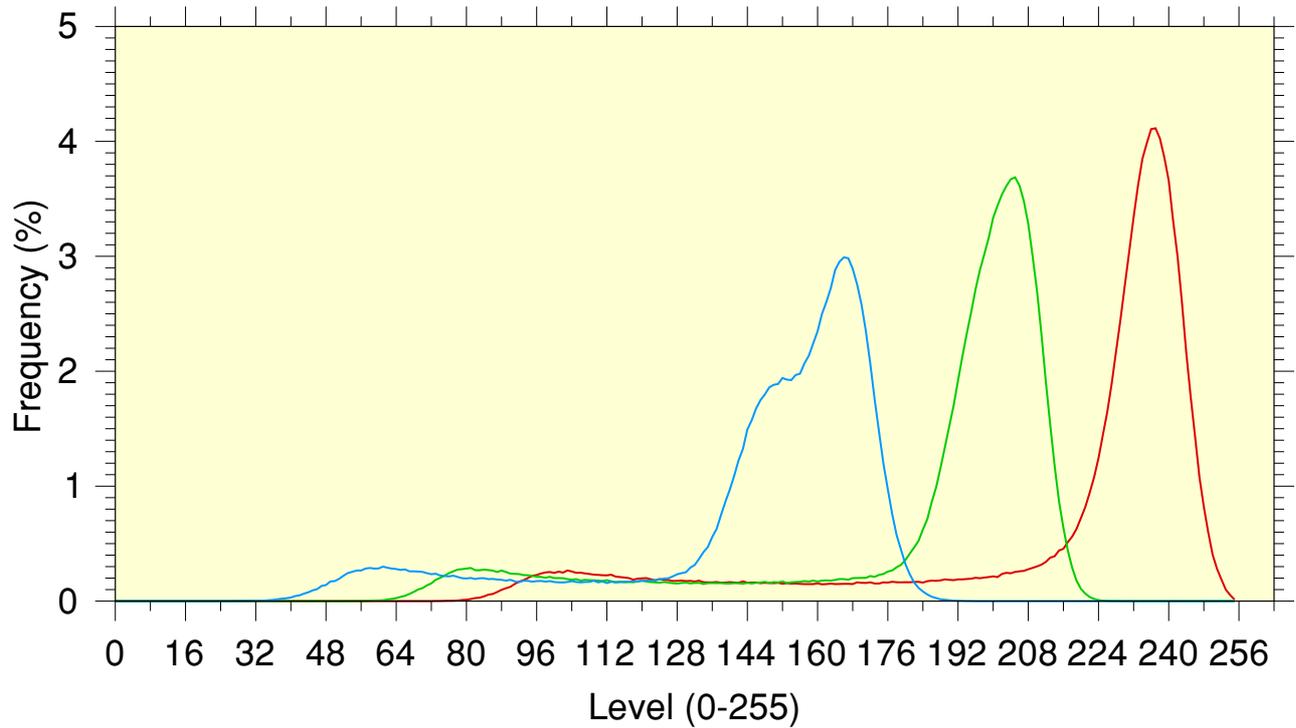


Example of the effect on the stats of RGB values of the oxidation of pulp paper

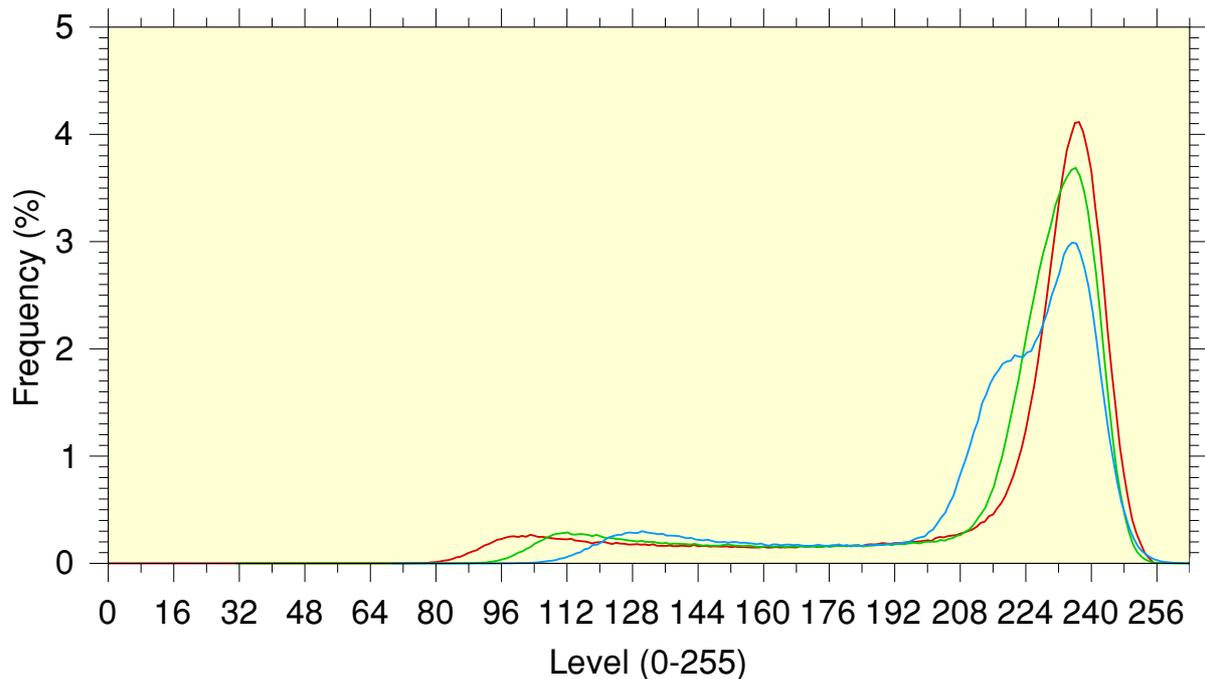


The above shows the results of plotting the frequency-of-occurrence of each possible RGB value in the pixels of a scanned image saved as a bitmap. The scanned page comes from an old book that used ‘pulp’ paper., produced some decades ago. Originally, the pages would have shown black ink on near-white paper. But pulp paper is prone to an oxidation process which over the years turns the paper brown, and can also make it become brittle.

The example used which gave this result was an ebook of scans of “The Best of John W. Cambell”. This shows a large area of oxidised paper.

Given clean black ink on white paper we would expect a scan of RGB values to essentially show a probability distribution which is greyscaled. i.e. the RGB distributions should be aligned. But the oxidation alters this, causing the RGB values to become misaligned. – as illustrated above.

Simply converting every pixel individually to greyscale by averaging loses this ‘signature’ – and thus discards the possibility that we might get better recovery of black text on a white page by applying a different approach to using the RGB distributions.

Example of the effect on the stats of RGB values of the oxidation of pulp paper

The above illustrates how we might be able to use the RGB distributions more effectively. Here they have been ‘re-aligned’ to have their patterns overlaid more realistically in terms of getting a “black on white” result. We can now speculate that the peaks at the ‘high’ end represent near-white paper pixels. We can also assume that white should mean that:

$$R = G = B = 255$$

Then perhaps assume that the aligned peak in the real values should be *at* 255. In addition we can guess that the blackest text should be at

$$R = G = B = 0$$

whereas in the scan, the lowest values when aligned are at about where the RGB values are around 80 to 100. Which implies we could ‘stretch’ the distribution to extend down to zero, and get blacker text that contrasts more clearly from the bulk of the paper!

Having performed a ‘re-alignment’ of the colour distributions the program could then decide how to use this initial information to convert to greyscale. The most obvious possibility being to ‘stretch’ the low-level (i.e. dark) section down towards ‘0’ and emphasise the text as being black.

