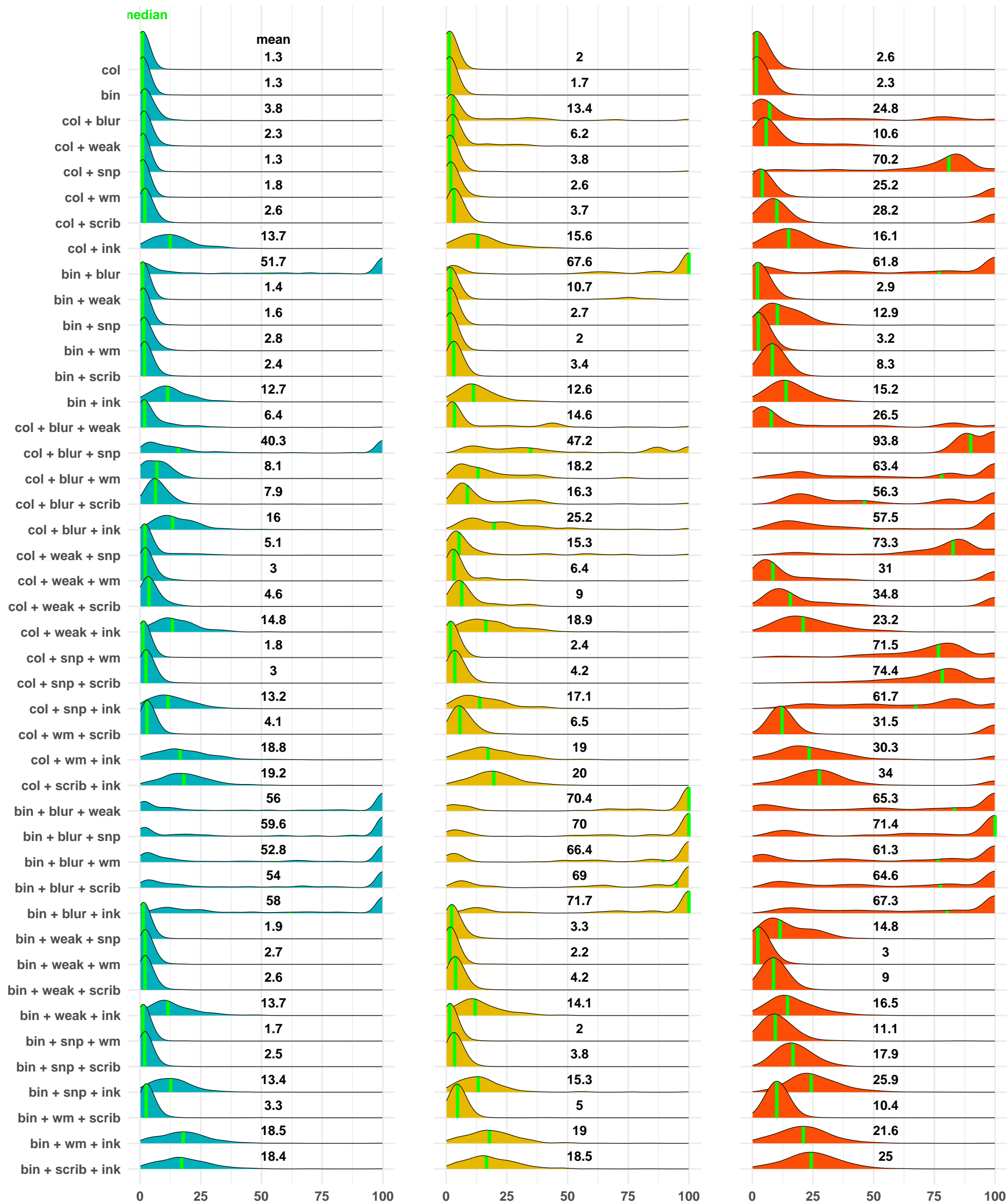


Data: Single-column text in historical book scans with noise added artificially (n=42,504; 322 per engine and noise type).
Noise codes: 'col'=colour, 'bin'=binary, 'blur'=blur, 'weak'=weak ink, 'snp'=salt&pepper, 'wm'=watermark, 'scrib'=scribbles, 'ink'=ink stains.

Google Document AI

Amazon Textract

Tesseract



Word error (percent)