

A Two-Factor Error Model for Quantitative Steganalysis

Rainer Böhme^a and Andrew D. Ker^b

^aTechnische Universität Dresden, Institute for System Architecture, 01062 Dresden, Germany;

^bOxford University Computing Laboratory, Parks Road, Oxford OX1 3QD, United Kingdom

ABSTRACT

Quantitative steganalysis refers to the exercise not only of detecting the presence of hidden stego messages in carrier objects, but also of estimating the secret message length. This problem is well studied, with many detectors proposed but only a sparse analysis of errors in the estimators. A deep understanding of the error model, however, is a fundamental requirement for the assessment and comparison of different detection methods. This paper presents a rationale for a two-factor model for sources of error in quantitative steganalysis, and shows evidence from a dedicated large-scale nested experimental set-up with a total of more than 200 million attacks. Apart from general findings about the distribution functions found in both classes of errors, their respective weight is determined, and implications for statistical hypothesis tests in benchmarking scenarios or regression analyses are demonstrated. The results are based on a rigorous comparison of five different detection methods under many different external conditions, such as size of the carrier, previous JPEG compression, and colour channel selection. We include analyses demonstrating the effects of local variance and cover saturation on the different sources of error, as well as presenting the case for a relative bias model for between-image error.

Keywords: Steganalysis, LSB steganography, regression analysis, steganalytic security metrics, benchmarking

1. INTRODUCTION AND MODEL

Quantitative steganalysis refers to the exercise not only of detecting the presence of hidden stego messages in carrier objects, but also of estimating the secret message length. Although quantitative methods have been applied against different embedding functions in various domains,¹ this paper is confined to detectors of LSB replacement steganography in the spatial domain of mono-channel images. This problem is well studied, with many detectors proposed²⁻⁹ but only a sparse analysis of errors in the estimators. Therefore it is valuable to study the source of these errors, and to determine their influencing factors on the basis of empirical data.

Similar efforts have been made in a recent workshop paper,¹⁰ where the accuracy of two quantitative attacks has been evaluated by means of generalised non-linear regression models. This work goes beyond the previous research in three respects: first, we show that the overall error distribution, which was previously assumed to be Cauchy, should actually be described as a compound distribution with at least two parts. Second, we produce more focussed results showing how the two error components are influenced by other factors, including embedding rate and cover image properties. Third, we have substantial confidence in the results, because of a wider experimental basis consisting of over 200 million attacks, five detection algorithms, and a more representative set of cover images.

The reason for a two-factor model for errors in estimation comes from the estimation processes themselves. In each of the quantitative steganalysis algorithms considered here^{2-4,8} there are two critical parts to derivation of the method. First, there is an assumption that the hidden message, embedded by replacement of least significant bits, is random in that a) the selection of bits for replacement is random and b) the message itself is uncorrelated with the cover bits replaced. Second, there is some assumption about cover images, specific to the steganalysis method being used. In the literature it is generally acknowledged that the second assumption (the cover image assumption) may not be completely accurate, and that this leads to deviation in the message length estimate. But

Further author information:

R. Böhme: E-mail: rainer.boehme@inf.tu-dresden.de, Telephone: +49 351 463 37918

A. D. Ker: E-mail: adk@comlab.ox.ac.uk, Telephone: +44 1865 283530

(SPIE 2006: Security, Steganography, and Watermarking of Multimedia Contents VIII; updated January 2006)

the first assumption (the hidden message assumption) is generally disregarded, although it should not be: even if the embedded message is truly random, some messages will, by chance, have correlations with the replaced cover bits, and this will affect the result. As we shall see, the magnitude of error introduced by the hidden message assumption can, in some cases, be larger than that caused by the cover image assumption.

Therefore we propose a two-factor model for errors. Instead of considering the error between the actual proportionate length of hidden message $p_{i,j}$ and estimated length $\hat{p}_{i,j}$, when message j is embedded in cover image $c^{(i)}$, as a realization of a single distribution, we describe it as the sum of two different sources of error.

$$\hat{p}_{i,j} = p_{i,j} + X(\mu_i, \sigma_i) + Z(\zeta, \theta) \quad (1)$$

The random variable X models the error caused by random correlations between the hidden message and cover. This we call the *within-image* distribution. Its parameters μ_i and σ_i (and in general there may be more) depend on the cover image $c^{(i)}$ and on $p_{i,j}$. The random variable Z describes the error caused by inaccuracy in the cover image assumption; we call this the *between-image* distribution. Its parameters ζ (for location) and θ (for higher-order properties) may also be influenced by $p_{i,j}$, and also by general characteristics of the cover images. For a given sample of cover images, we assume these parameters to be fixed and thus can be approximated from empirical measurements. In the context of sampling theory, the parameters can be generalized to the population from which the sample has been drawn at random, and the precision of the parameter estimates depends on the sample size.

We will see, in Sect. 3, that the distribution of X is well-modelled by the Gaussian (normal) family, but that the distribution of Z is heavy-tailed. We will show that the latter can be modelled fairly well by Student’s “ t distribution” family, and compare the magnitudes of the random variables X and Z . Models of influences on the parameters of X and Z are presented in Sect. 4. Finally, Sect. 5 briefly concludes the paper.

2. EXPERIMENTAL SETUP

The entire NRCS Photo Gallery* was downloaded, and some corrupt and reduced-colour images removed. This amounted to just over 3000 very large TIF files (around 2100×1500 pixels in size), apparently scanned from film. A random selection of 800 equal-sized images was made, and the images were reduced (using Photoshop’s bicubic resampling method) to 640×458 pixels. Finally, the luminescence component (for grayscale tests) and the red colour component (for colour image tests) were extracted.

We implemented five quantitative steganalysis algorithms:

- (i) The method of RS,² using horizontal groups of 4 pixels and the mask $[0, 1, 1, 0]$, as suggested in the original publication.
- (ii) The method known as Weighted Steganalysis (WS).⁶ We used the simple pixel predictor, and weighting method, given in this publication ($\alpha = 1$), but did not include the so-called “flat pixel correction” because we found that it reduced detector accuracy.
- (iii) The method of Sample Pairs (SPA),³ making use of all horizontally and vertically adjacent pairs of pixels, as in the original publication.
- (iv) The Least Squares variation on Sample Pairs,⁴ which we denote SPA/LSM.
- (v) The Triples method⁸ making use of all horizontal groups of three pixels, and using only the case of “Parity Symmetry” when both indices are odd (see Ref. 8 for an explanation of this terminology).

In contrast with other benchmarking experiments in the literature, we repeat the embedding-detection cycle with 200 random messages m_j , for each image $c^{(i)}$ and each embedding rate $p \in \{0.01, 0.05, 0.1, 0.2, 0.4, 0.6, 0.8, 1\}$. This allows us to separate the within-image errors, caused by the message, from between-image errors and the influence of p . It amounts to some 6.4 M attacks on 800 images. The experiments were repeated with both grayscale and the red component of colour images, both with and without prior JPEG compression (at “quality factor” 80), for a base of 25.6 M attacks. Finally, in order to examine the effect of cover image size we created new sets of 800 cover images from the originals, by cropping down to random regions with 75 %, 50 %, and 25 %

*<http://photogallery.nrcs.usda.gov>

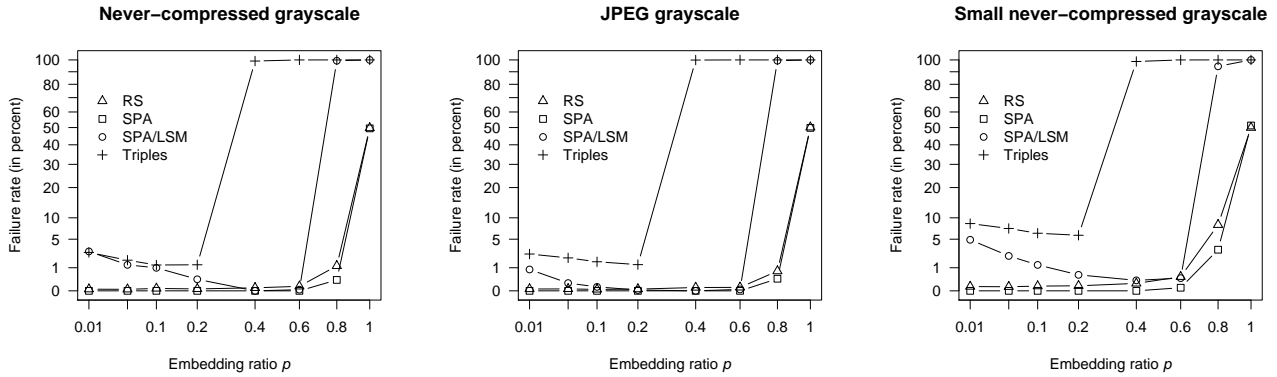


Figure 1. Failure rates for different detectors, as a function of the embedding ratio p . Results from 15.4 M attacks.

of the original number of pixels, and also by resampling down to smaller images with 75 %, 50 %, and 25 % of the original number of pixels.

Of the 5 detectors implemented, only the WS method is guaranteed to produce an answer. The methods of RS and SPA can fail because they both involve solving a quadratic equation; if the equation has no roots then the method does not make an estimate for p . It is common practice to return an estimate $\hat{p} = 1$ in such circumstances, but we decide to treat such situations as genuine failures and exclude all such cases from our subsequent analyses. The frequency of failure depends on both the cover image and the length of embedded data: there are negligibly few failures when no data is embedded, and the failure rate rises to approximately 0.5 as p approaches 1.

The methods of SPA/LSM and Triples have additional failure modes. Although not noted in Ref. 4, the SPA/LSM detector is prone to wild inaccuracies when p is close to 1. In practice we would expect to screen this method by first applying a different detector, for example the SPA estimator, and not proceeding to the SPA/LSM method if p appears close to 1. In our experiments we did exactly this, and caused the SPA/LSM method to fail to give an estimate when the SPA estimate was greater than 0.75 or failed. The Triples method is even more prone to failure, suffering from inaccuracies even for moderate values of p . We censored the Triples results whenever the SPA estimate was greater than 0.3.

We plot the observed failure rates of the detectors (excluding WS, which never fails) in Fig. 1 as p varies, for three of the sets of covers (full size grayscale with and without JPEG compression, and the never-compressed grayscale covers subject to reduction to 25 % pixels before embedding). We observe that the Triples method has some failures even at $p = 0$ and that both SPA/LSM and especially Triples jump up to total failure as p goes above a certain threshold. This is expected behaviour, because of the screening which is designed to force failure (rather than allow an inaccurate result) in such cases. But the fact that the failure rate of Triples and SPA/LSM somewhat *decreases* as p initially increases from 0 is very surprising. RS and SPA remain at very low failure rates until $p = 1$, but in the smaller images we also see significant failures at $p = 0.8$. We postpone further examination of failure rate, and its influences, to future work.

3. SEPARATION OF ERROR SOURCES

The large body of experiments, including 200 repetitions of each embedding (with different messages) allows us to separate the within-image error (caused by the message) from the between-image error (caused by inaccurate cover assumptions). We call each set of 200 repetitions a *cell*, in which all conditions other than the hidden message are kept constant (cover image, detection algorithm, embedding ratio) and it represents a sample from a within-image error distribution. Note that there may be fewer than 200 results in each cell, if the steganalysis algorithm fails for some of the repetitions. By taking the mean of the values in each cell we expect to remove (or reduce to insignificance) the effect of within-image error, leaving the between-image error exposed.

Figure 2 shows a particular case (the RS detector, messages of length $p = 0.2$ embedded in the never-compressed grayscale covers) to illustrate the appearances of the within- and between-image error. It displays

Table 1. Summary of 763 k tests for normality of the within-image error distribution. The percentages reflect the proportion of cells that passed the test at the threshold indicated. If the normality assumption is true, we would expect a pass rate of 90 %.

Test	Emb. ratio	Detector				
		RS	WS	SPA	SPA/LSM	Triples
Shapiro-Wilk: $P(\alpha) > 0.1$						
	0.01	90.2 %	79.7 %	90.1 %	88.3 %	86.8 %
	0.05	90.2 %	87.8 %	90.0 %	88.2 %	87.5 %
	0.10	90.2 %	89.2 %	90.3 %	88.6 %	87.6 %
	0.20	89.6 %	89.4 %	90.3 %	89.3 %	87.4 %
	0.40	89.6 %	90.4 %	89.8 %	88.8 %	78.7 %
	0.60	87.5 %	90.1 %	88.1 %	85.0 %	79.3 %
	0.80	59.9 %	89.9 %	55.8 %	82.6 %	
	1.00	81.8 %	90.4 %	83.8 %		
Number of tests:		179.2 k	179.2 k	179.2 k	140.8 k	85.1 k

histograms obtained by (respectively) pooling cells of this type, taking just one cell (i.e. fixing the cover image), and the mean of each cell (removing intra-cell deviation). The middle histogram has a best-fit Gaussian density superimposed.

It is apparent from the histograms that the within-image error is short-tailed but that the between-image error is heavy-tailed. Further evidence is provided by the quantile-quantile plots below each histogram in Fig. 2, where the tail probabilities are compared (on a log-log scale) to that predicted by the Gaussian distribution and by the Student- t distribution with 2.5 degrees of freedom (of which more later).

3.1. Shape of within-image error

Each cell is a sample, of size up to 200, from the within-image error distribution (plus some constant, determined by the otherwise constant factors such as cover image and detector type). We claim that the parent distribution is Gaussian.

Evidence is provided in Tab. 1. Here, we have taken every cell of our experimental data, and performed a separate normality test on each: using each of the 28 sets of covers (grayscale and colour; with and without prior JPEG compression; full size, cropped to one of three smaller sizes, and resampled to one of three smaller sizes), with 800 covers in each set, and each embedding rate from 0.01 to 1 gives rise to 896000 cells. A few cells, for the SPA/LSM and Triples detectors, were discarded because they contained fewer than 50 data points, leaving approximately 763000 normality tests. We set a significance level of 90% and report, in Tab. 1, the proportion of cells passing the test, broken down by detector type and embedding rate. If the within-image errors are truly Gaussian, we would expect that 90% of cells would pass the test. The normality test used was the Shapiro-Wilk test,¹¹ chosen for its good power in discriminating Gaussian distributions from heavy-tailed alternatives, although we observed similar results with other normality tests.

The table confirms that, in a very wide variety of circumstances, the within-image error has a Gaussian distribution. There are a few anomalous figures in the table, however: the relatively low rate of passes for WS at low embedding rates comes from small images sizes. The gap at $p = 0.80$ for RS, SPA is difficult to explain, but we found that JPEG compressed images pass the test more frequently (about 62%) than never-compressed images (58%). This discrepancy is even stronger for SPA (61% and 51%, respectively). We have no particular explanation for these results, but in any case a pass rate of over 50% indicates that departures from normality are not severe.

We conclude that the claim of normality of within-image error is well supported. This is important for the rest of our analysis, as the Gaussian short tails (which we could also verify using quantile-quantile plots, but have omitted from this paper) mean that there is rapid convergence of cell mean to the true mean. The average

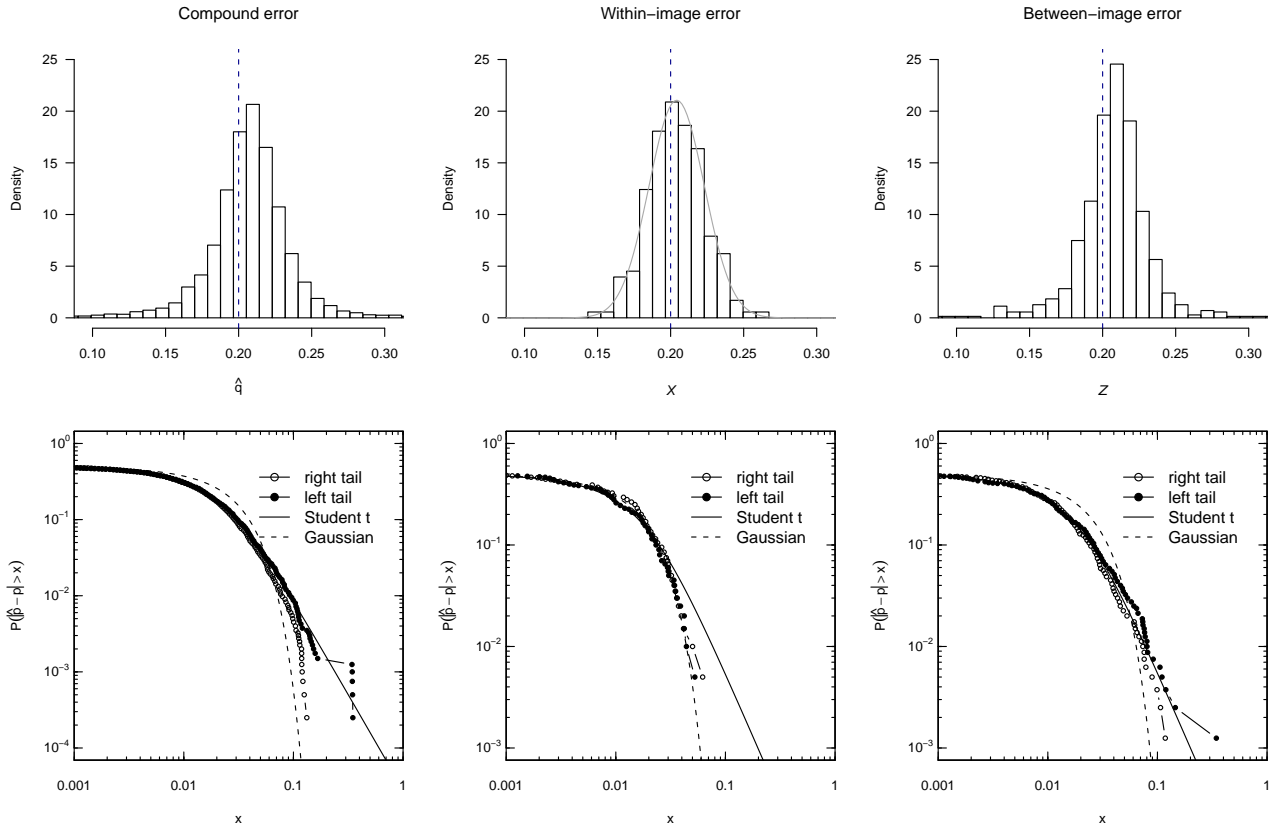


Figure 2. Histograms and QQ-plots of error components for RS analysis. The compound error distribution across all images and messages (left) can be decomposed as a sum of a Gaussian within-image error distribution (middle, for one selected image) and a heavy-tailed between-image distribution (right). Data from never-compressed grayscale images with embedding ratio $p = 0.2$. Degrees of freedom parameter $\nu = 2.5$.

of 200 data points should reduce the within-image error by a factor of $\sqrt{200}$, which will make it small enough to ignore. In this way the between-image error can be separated.

As an additional benefit, we can now replace the millions of attacks with a smaller data set: the sample size, mean and standard deviation of each cell—this suffices to describe the data contained in it. It helps to reduce the complexity of subsequent analyses.

3.2. Shape of between-image error

As strong as is the evidence that within-image errors are short-tailed Gaussian, so is the evidence that between-image errors are not. Given the preceding section, we take the means of cells to indicate estimates with within-image error removed, and perform normality tests for these values as they vary over the 800 cover images, for each detector and embedding rate separately (approximately 1000 such tests). It is not interesting enough to merit a table, because all fail with p-values below 10^{-4} .

The briefest examination reveals that the between-image errors are heavily tailed. In Ref. 10 it was proposed that the compound error (within- and between-image errors not separated) could be modelled by a Cauchy distribution. Our observations indicate that the between-image error is not quite as heavily tailed as the Cauchy distribution (where the cumulative density F satisfies $1 - F(x) \sim x^{-1}$) but has a tail index closer to 2 (i.e. the cumulative density satisfies $1 - F(x) \sim x^{-2}$). We propose that the between-image error can be well-modeled using the Student- t family of distributions.¹²

The Student- t family is parameterised by a *degrees of freedom* parameter ν , as well as standard scale and location parameters. The distributions are defined for $\nu > 0$, are all symmetric about their mean, and smaller

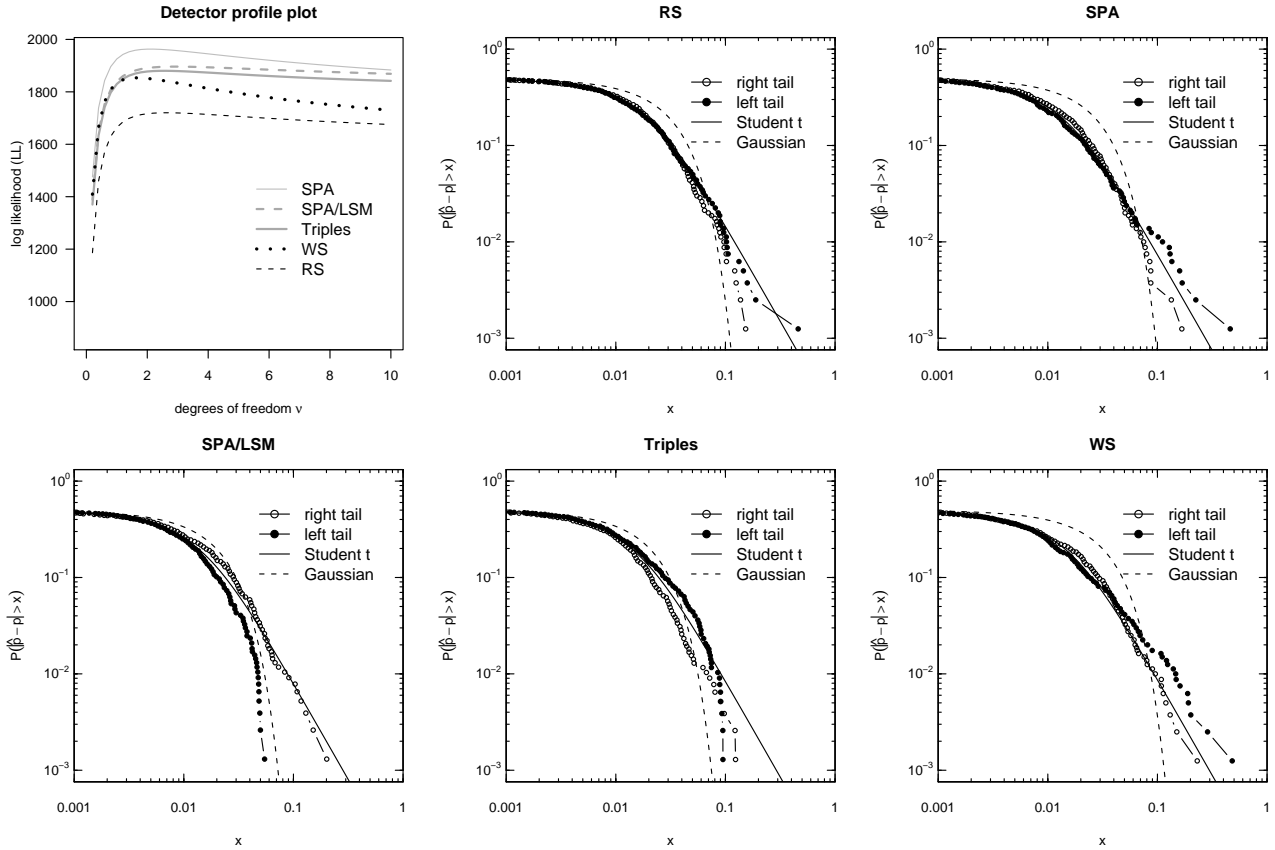


Figure 3. Log-likelihood profile and QQ-plots for Student- t ($\nu = 2$) fit of between-image distribution for different detectors. Results from 800 never-compressed grayscale images with $p = 0$ (i.e. distribution of $\hat{p}^{(0)}$).

ν indicates heavier tails. When $\nu = 1$ the distribution is Cauchy, and as $\nu \rightarrow \infty$ it approaches the Gaussian distribution. The cumulative density satisfies $1 - F(x) \sim x^{-\nu}$ and this means that only the moments strictly greater than ν are finite. In particular, when $\nu \leq 2$ the distribution has infinite variance.

Figure 3 compares between-image errors (considering only the case $p = 0$ and with never-compressed full-size grayscale covers, i.e. 800 samples from a single between-image distribution) with the Student t -distribution family. The first chart shows the log-likelihood profile for each detector as ν varies—we observe that $\nu \approx 2$ seems to be the best fit. The other charts plot the sample tail probabilities against those predicted by both Gaussian and t_2 distributions (we use t_2 as a shorthand for the Student- t distribution with 2 degrees of freedom). It is clear that the t_2 distribution is a much better fit for the observed data than the Gaussian distribution (perhaps less so for the Triples detector, which may have lighter tails than the others). We caution the reader against lending too much significance to the last few data points, because they represent only the highest and lowest few observations in the sample and, like all extreme value statistics, are prone to outliers. The left and right tails have been plotted separately, and there does appear to be discrepancies between them, for the SPA/LSM and Triples detectors. We should investigate this asymmetry further, in future work.

The charts in Fig. 3 have been restricted to a single cover image set, with fixed p . This is to control for cover properties and embedding rate, both of which will surely affect the error distribution; we study these influences in Sect. 4. Similar charts, with different choices for the cover set and p , also show that t_2 is a reasonably good fit for the between-image error.

In the preceding analyses we required an estimation of the location and scale parameters of the t_2 distribution so as to fit the quantile-quantile plots correctly. The Maximum Likelihood Estimators (MLE) for scale and location (and, later, for the parameter ν) were used, with the scoring method of Ref. 12 to accelerate convergence.

Table 2. Magnitude of error components measured in interquartile ranges $Q_{75} - Q_{25}$; breakdown by detector and embedding ratio.

Detector	Embedding ratio							
	$p = 0.01$		$p = 0.1$		$p = 0.4$		$p = 0.8$	
	within-image $\Delta Q(X)$	between-image $\Delta Q(Z)$	within-image $\Delta Q(X)$	between-image $\Delta Q(Z)$	within-image $\Delta Q(X)$	between-image $\Delta Q(Z)$	within-image $\Delta Q(X)$	between-image $\Delta Q(Z)$
Never-compressed grayscale images								
RS	0.38	2.92	1.18	2.62	2.18	1.71	3.56	0.63
WS	0.23	2.02	0.70	2.37	1.27	3.26	1.58	2.14
SPA	0.20	1.92	0.61	1.78	1.18	1.19	2.24	0.44
SPA/LSM	0.19	2.10	0.59	1.95	1.22	1.31	–	–
Triples	0.27	2.17	0.89	2.10	–	–	–	–

Base: 2.7 M attacks on 800 source images; smaller interquartile ranges denote better accuracy

This reference also demonstrates that the MLEs are consistent, at least in the univariate case. However the degrees of freedom parameter is notoriously difficult to estimate accurately, as indeed the reader can already see from the first diagram of Fig. 3, where the likelihood function has a rather a gentle slope as ν varies. The approximate value of $\nu = 2$ seems at first sight problematic, because $\nu = 2$ is the critical value, only above which variance is finite. Finite variance would allow the use of many traditional statistical methods, including the Central Limit Theorem. However, even if the true value of ν turns out to be a little above 2, any such asymptotic results will have extremely slow convergence and should not be relied upon.

We note that there is no theoretical justification for the t_2 distribution (or any other t -distribution) of between-image errors and we do not seriously advance the distribution as an exact model. However we have demonstrated that it is a much more accurate model than the alternatives, particularly the Gaussian distribution. It has consequences for subsequent analyses, particularly the regressions of Sect. 4 for which the traditional least-squares minimization methods will not be appropriate.

3.3. Comparison of error magnitudes

We wish to compare typical magnitudes of within- and between-image error, obtained from our experiments. Unfortunately it is not obvious how this magnitude should be defined: the usual measures of distributional spread (standard deviation, average absolute error) are not suitable for heavy-tailed distributions, and it is hard to find a sensible like-for-like comparison between one heavy-tailed and one light-tailed distribution. We settled on use of the interquartile range (which we denote as $Q_{75} - Q_{25}$ with explicit quantiles and abbreviate ΔQ) as a robust measure of distributional spread, while noting that it takes no account of the far tails.

For now taking only the full-size never-compressed grayscale covers, and $p \in \{0.01, 0.1, 0.4, 0.8\}$, we compute the interquartile range both of within-image error (the distribution X) and of between-image error (Z), tabulating the results in Tab. 2. It illustrates vividly that the between-image error is fairly insignificant for small messages (when $p = 0.01$ the between-image errors are about 10 times larger than within-image error) but for medium-length messages the two sources of error are comparable (although the different shape of tails should be born in mind here) and for messages as long as $p = 0.8$ the within-image error is the more substantial. Curiously, this is not the case for the WS detector, for which the between-image error remains the larger.

It is certainly conceivable that different detectors may be more prone to one type of error or the other. Further, we might expect that the two types of error depend on, for example, the cover image size in different ways. Figure 4 shows how both types of error depend on the embedding rate p , in a variety of sets of cover images. Note how the between-image error decreases, but the within-image error increases, as p increases. The exception here is the WS detector, for which both sources of error increase until, at $p = 1$, the between-image error suddenly jumps to zero. Although performing rather poorly for p in the range 0.1–0.8 (especially with

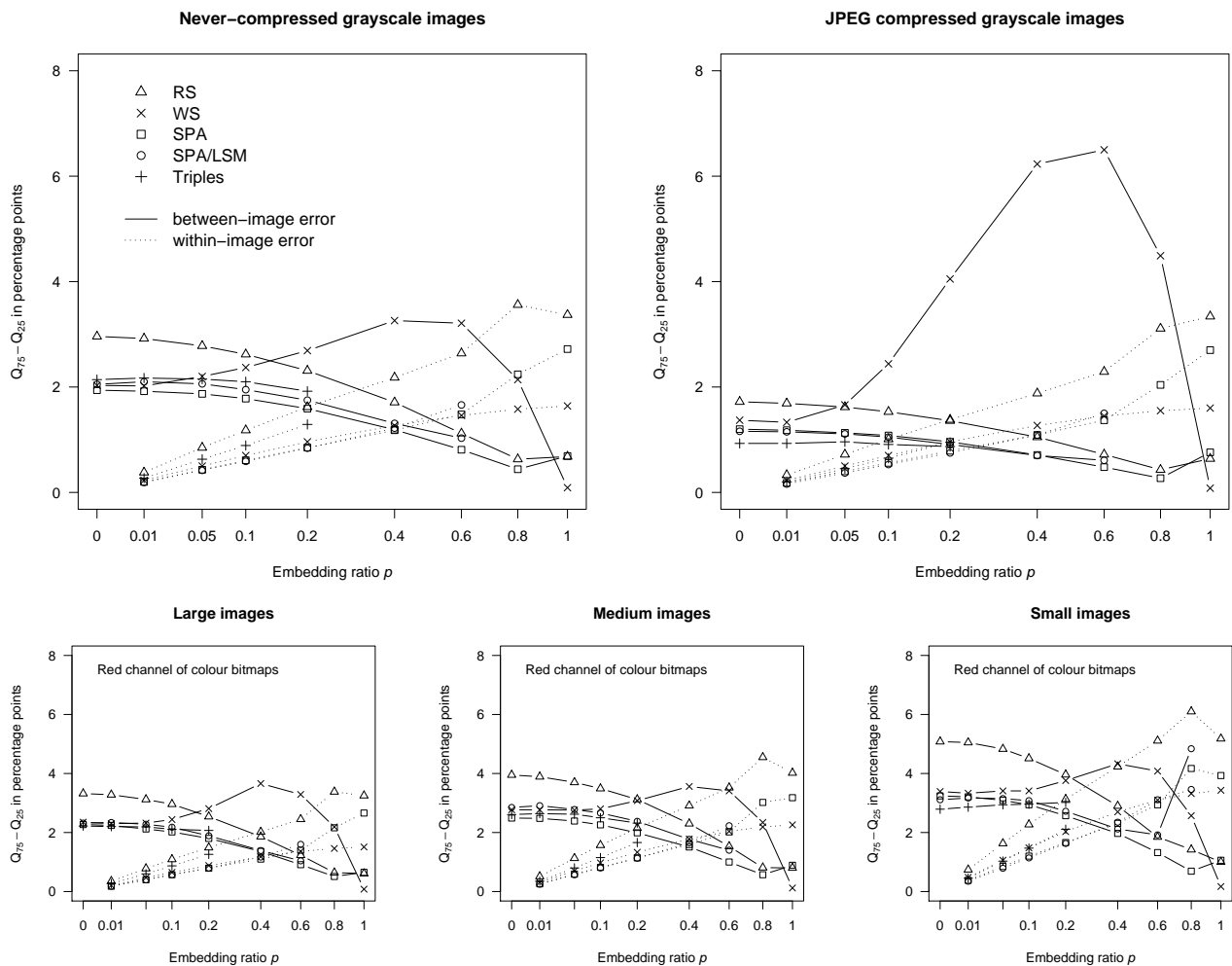


Figure 4. Relative size of within- and between-image errors measured in interquartile ranges to cope with the different shapes of the error distributions. Smaller ranges denote better detector accuracy. Above left, for grayscale never-compressed bitmaps at full size. Above right, grayscale bitmaps subject to JPEG precompression. Below, colour never-compressed bitmaps at three different sizes: (left to right) 640×458 , 452×324 , and 320×229 .

JPEG covers), the WS detector is in fact the most reliable right at $p = 1$. In the lower part of Fig. 4 we see how image size affects error: both types of error increase in smaller images, for all detectors.

The visual presentation of Fig. 4 is intuitive, but a more rigorous assessment is needed to properly distinguish the influences of p , image size, and cover image. A multi-factor analysis is presented in the next section.

3.4. Location of between-image error

After removing within-image error by taking cell means, we can also strip out between-image error by taking the median of the 800 cell means, for the 800 cover images in each set (median rather than mean for better robustness). What is left is a detector-specific bias, which also depends on the cover image set and the embedding rate p . For the grayscale covers, with and without prior JPEG compression, these biases are plotted in Fig. 5. Particularly notable features include that a) WS apparently follows a different pattern to the other detectors, b) there is a sudden downward bias at $p = 1$, particularly for SPA and RS, and c) possible image pre-processing steps change the magnitude and, for some detectors, the structural relationship with p .

We reserve further analysis of this bias for future work, while noting that the negative bias at $p = 1$ represents a substantial error in the detector output: there is a systematic underestimate of around 0.05, with the SPA and RS detectors.

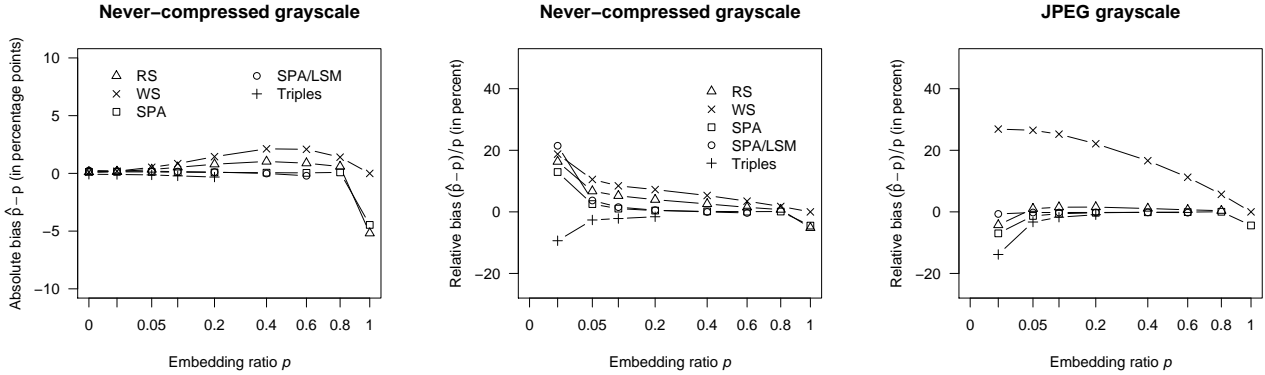


Figure 5. Absolute and relative estimation bias for varying embedding rates p . Absolute (left) and relative (middle) bias for grayscale never-compressed bitmaps, and relative bias for JPEG compressed grayscale bitmaps (right).

4. FACTORS INFLUENCING DETECTION ACCURACY

Regression models offer an analytical framework to assess the accuracy of quantitative steganalytic methods conditional on influencing factors.¹⁰ The decomposition of the estimation error into a within-image and a between-image component allows a more fine-grained analysis of the effects of influencing factors. For both types of error, the dispersion of the error distribution is considered as an indicator of *accuracy* whereas the deviation of the location is referred to as *bias*.

4.1. Models of image-specific bias

In the comparison of error magnitudes (Sect. 3.3, see also Fig. 4) it was evident that the within-image error converges to zero for very small embedding rates, whereas the between-image error remains at a considerable level even when no message is embedded. This is particularly unfortunate because it sets an upper limit on the operating threshold of a stego-detector, and thus a lower limit for the minimum reliably detectable stego message. We use regression models to show the presence of a non-negligible image-specific bias μ_i that can be approximated by $\hat{p}_i^{(0)}$, the detector outcome if nothing is embedded ($p = 0$) and appears as a persistent additive error across all embedding rates.

Therefore we first formulate a **default model** without the image-specific part.

$$\bar{p}_i = a_0 + a_1 \cdot p + \varepsilon, \quad \varepsilon \sim t_2(0, \lambda) \quad (2)$$

with the dependent variable \bar{p}_i being the mean detector result over all 200 random messages per image, to eliminate within-image error from the data. That is, we are working on the level of aggregated cells and consider the mean \bar{p}_i (here) and the empirical standard deviation $\hat{\sigma}_i$ (later below) as dependent variables which describe the between- and within-image errors, respectively. Vector $\mathbf{a} = (a_0, a_1, a_2)$ contains the coefficients for the location model, and p the true embedding rate used in the simulation experiments.

The coefficients in Tab. 3 have been estimated separately for each detector with a ML approach, considering the t_2 residual distribution. As expected, the coefficients a_0 are close to zero and a_1 close to 1. The interquartile ranges $Q_{75} - Q_{25}$ of the residuals correspond to the levels shown in Fig. 4.

The **absolute bias model** assumes the presence of an image-specific bias μ_i which is constant for all embedding ratios and can be approximated by $\hat{p}_i^{(0)}$. We fix a_0 at zero to keep to a two-parameter model.

$$\bar{p}_i = a_1 \cdot p + a_2 \cdot \hat{p}_i^{(0)} + \varepsilon, \quad \varepsilon \sim t_2(0, \lambda) \quad (3)$$

There is clear evidence for an image-specific bias as, for all detectors, the absolute bias model fits better than the default model. The superiority can be seen both in the goodness of fit ($-\log$ likelihood; accordingly, a likelihood ratio test (LRT) returns highly significant results for all detectors), and in the reduction of residual dispersion.

Table 3. Fitted coefficients of fixed effects model for between-image bias. Dependent variable $\hat{p}_{i,j}$ with fixed predictors p_i , $\hat{p}_i^{(0)}$, and interaction term. Within-image randomness is eliminated from the general model by using aggregated per-cell averages; data from never-compressed grayscale images only.

Parameter	Detector				
	RS	WS	SPA	SPA/LSM	Triples
Default model					
a_0	0.01 (0.000)	0.01 (0.000)	0.00 (0.000)	0.00 (0.000)	-0.00 (0.000)
a_1	0.98 (0.001)	1.00 (0.001)	0.99 (0.001)	0.99 (0.001)	0.99 (0.004)
Res. $Q_{75} - Q_{25}$	2.98	2.30	1.99	1.65	2.03
-log likelihood	14044	14620	15908	12027	7613
Absolute bias model					
a_1	1.01 (0.000)	1.01 (0.000)	1.00 (0.000)	1.00 (0.000)	0.99 (0.000)
a_2	0.79 (0.003)	0.83 (0.005)	0.82 (0.002)	0.89 (0.002)	0.95 (0.001)
Res. $Q_{75} - Q_{25}$	1.00	1.65	0.48	0.25	0.14
-log likelihood	17332	15781	20275	18235	14573
Relative bias model					
a_1	1.01 (0.000)	1.01 (0.000)	1.00 (0.000)	1.00 (0.000)	0.99 (0.000)
a_2	0.99 (0.003)	0.99 (0.007)	1.00 (0.000)	1.00 (0.000)	1.00 (0.001)
Res. $Q_{75} - Q_{25}$	0.49	1.55	0.06	0.06	0.08
-log likelihood	19422	16604	26352	23195	15569
Summary					
n	6400	6400	6400	4745	3100
Range of p	[0.01, 1.00]	[0.01, 1.00]	[0.01, 1.00]	[0.01, 0.80]	[0.01, 0.40]

Std. errors in brackets; all coefficients significant with $P(\alpha) < 0.001$

There are theoretical reasons, omitted here, for believing that the message length estimators should follow a relative, as opposed to absolute, bias model. The **relative bias model** assumes a constant image-specific bias at $p = 0$ that vanishes with increasing p . We can formalise it as

$$\bar{p}_i = \mu_i + p \cdot (1 - \mu_i) + Z(\theta) \quad (4)$$

Inserting $\hat{p}^{(0)}$ and rewriting as a regression formula yields a model with an interaction term (we keep one coefficient for the whole bracket to avoid estimating more than two parameters):

$$\bar{p}_i = a_1 \cdot p + a_2 \cdot (\hat{p}_i^{(0)} - p \cdot \hat{p}_i^{(0)}) + \varepsilon, \quad \varepsilon \sim t_2(0, \lambda) \quad (5)$$

The results in Tab. 3 clearly show that the relative bias model fits best for all detectors, with coefficients very close to the expected value 1 and remaining residual dispersion between 3% (SPA) and 67% (WS) of that in the default model. We conclude that there is solid evidence for relative bias in these detectors.

Moreover, it is interesting to note that the image-specific bias is apparently of similar origin for all detectors, even though some of them are constructed along completely different principles: Tab. 4 displays the pairwise correlations of $\hat{p}_i^{(0)}$ between detectors applied on the same set of carrier images. The correlation coefficients are in about the same range for never-compressed grayscale and red colour channel data (red not printed), and somewhat lower if the images were subject to JPEG compression before the embedding. We also computed the eigenvectors of the correlation matrices to explore the dimensionality of the underlying structure. It turned out that in all cases only one eigenvalue is larger than one, which leads us to the conclusion that there is just one latent image-specific factor (within the particular set of covers tested) that biases all detectors similarly.

Table 4. Correlation between detectors of image-specific bias $\hat{p}^{(0)}$.

	Never-compressed grayscale images					JPEG compressed grayscale images				
	RS	WS	SPA	SPA/LSM	Triples	RS	WS	SPA	SPA/LSM	Triples
RS	1.00	–	–	–	–	1.00	–	–	–	–
WS	0.64	1.00	–	–	–	0.40	1.00	–	–	–
SPA	0.76	0.86	1.00	–	–	0.55	0.75	1.00	–	–
SPA/LSM	0.60	0.86	0.89	1.00	–	0.31	0.71	0.75	1.00	–
Triples	0.45	0.66	0.64	0.70	1.00	0.18	0.40	0.30	0.46	1.00

Base: 8000 attacks; correlation coefficients estimated for multivariate Student t -distributions, see Ref. 13.

4.2. Influence of local variance

The local variance of an image has been identified as the most influential predictor for detection accuracy among a number of statistical image characteristics.¹⁰ Against the backdrop of a two-factor error model, the aim of this section is to assess the influence of local variance on both error components separately. Therefore the *local variance* $v_{\text{loc}i}$ of each source image i , represented as the $u \times v$ intensity matrix $c_{k,l}^{(i)} \in \{0, 1, \dots, 255\}$, has been computed as

$$v_{\text{loc}i} = \frac{1}{2uv - (u + v)} \left[\sum_{k=1}^{u-1} \sum_{l=1}^v (c_{k,l}^{(i)} - c_{k+1,l}^{(i)})^2 + \sum_{k=1}^u \sum_{l=1}^{v-1} (c_{k,l}^{(i)} - c_{k,l+1}^{(i)})^2 \right]. \quad (6)$$

In our sample of never-compressed grayscale images (sized 640×458), v_{loc} ranges from 3.0 to 1494.4 with a median at 250.5. As there is no theoretical reason to assume a particular parametric distribution for v_{loc} , we used descriptive and graphical tools to find a suitable working assumption. Surprisingly, the fourth root turned out to fit a Gaussian very well: it passes the Shapiro-Wilk test with $P(\alpha) > 0.1$ and the QQ-plot diagnostic shows no suspicious deviation (see Fig. 6, rightmost chart). So we decided to define $\ell_i = \sqrt[4]{v_{\text{loc}i}}$ as an independent variable in the following regression models. For the dependent variable, we found that a log transform centers the distribution well, so the first model to evaluate the influence of **local variance on the magnitude of the within-image error** is a log-linear relationship:

$$\log(\hat{\sigma}_i) = b_0 + b_1 \cdot \ell_i + \varepsilon, \quad \varepsilon \sim N(0, \sigma_\varepsilon) \quad (7)$$

The coefficient vector $\mathbf{b} = (b_0, b_1)$ has been estimated using ordinary least squares for each detector separately. The results are presented in Tab. 5 together with the estimated parameters for all following models in this section. The intercepts b_0 reflect the overall level of within-image error and the coefficients b_1 show a significant positive linear relationship between local variance and the magnitude of between-image errors for all detectors. A measure for the proportion of explained variance in the dependent variable by its predictors is given by the R^2 statistic.

$$R^2 = \frac{\text{var} \log \hat{\sigma}_i - \text{var} \varepsilon_i}{\text{var} \log \hat{\sigma}_i} \quad (8)$$

With values between 12% (Triples) and 39% (RS) we see a considerable influence of local variance and some noteworthy detector differences: the Triples estimator is least adversely affected by images with high local variance.

To conduct a similar analysis for the influence of **local variance on the dispersion of the between-image error**, we specify a heteroscedastic regression model as:

$$\begin{aligned} \bar{p}_i &= a_0 + \varepsilon_i, & \varepsilon_i &\sim t_2(0, \lambda) \\ \log(\lambda^2) &= b_0 + b_1 \cdot \ell_i + e, & e &\sim t_2(0, \cdot) \end{aligned} \quad (9)$$

An efficient method to compute estimates for the parameter vectors \mathbf{a} and \mathbf{b} is given by iterative reweighted least squares scoring which alternates the update steps for the location and scale models.¹² The results in

Table 5. Influence of local variance on estimation accuracy; estimated regression coefficients for never-compressed grayscale images at embedding rate $p = 0.05$.

Parameter	Detector				
	RS	WS	SPA	SPA/LSM	Triples
Local variance on std. dev. of within-image error					
b_0	-6.95 (0.079)	-7.19 (0.091)	-7.43 (0.084)	-7.04 (0.091)	-6.39 (0.089)
b_1	0.44 (0.020)	0.36 (0.022)	0.39 (0.021)	0.28 (0.022)	0.23 (0.022)
R^2	0.39	0.24	0.31	0.17	0.12
Local variance on dispersion of between-image error					
a_0	0.05 (0.001)	0.06 (0.001)	0.05 (0.001)	0.05 (0.001)	0.05 (0.001)
b_0	-10.76 (0.451)	-10.87 (0.451)	-10.73 (0.451)	-11.02 (0.457)	-10.40 (0.458)
b_1	0.63 (0.111)	0.56 (0.111)	0.45 (0.111)	0.57 (0.113)	0.43 (0.113)
R^2	0.03	0.01	0.02	0.01	0.02
$-\log$ likelihood	1779	1857	2012	1926	1873
Summary					
n	800	800	800	790	788

Std. errors in brackets; all coefficients significant with $P(\alpha) < 0.001$

Tab. 5 indicate a positive and significant relationship for all detectors; that is, larger local variance predicts wider between-image errors. However, the influence of the predictor as measured with the R^2 statistic is of the order of 10 times smaller than the influence on the within-image error. This relation remains stable across different embedding rates, with an exception for WS at $p = 1$ (see Fig. 6) so the embedding rate has been fixed for all parameter estimations of models in this section. Albeit to a lesser extent, local variance most affects the RS detector, whereas the WS detector is least influenced. A possible explanation for the latter is that the WS employs a weighting mechanism based on a local variance criterion.

For the model specified in (9), the R^2 statistic has been computed on $\text{var}(\log(\lambda^2))$ and $\text{var}(e)$ after estimating a_0 from data. Note that this is still a suboptimal measure for models with non-Gaussian and especially for those with heavy-tailed residuals. We opted for it despite these concerns because there is no clearly better alternative for the general problem of comparing dispersion between heavy- and short tailed distributions.

Possible effects of **local variance on the image-specific bias** have been captured with the following model:

$$\begin{aligned} \hat{p}_i^{(0)} &= a_0 + a_1 \cdot \ell_i + \varepsilon_i, & \varepsilon_i &\sim t_2(0, \lambda) \\ \log(\lambda^2) &= b_0 + e, & e &\sim t_2(0, \cdot) \end{aligned} \quad (10)$$

Since none of the estimated coefficients a_1 differ significantly from zero (all are below 10^{-2} in absolute value), we reject the hypothesis of the existence of a linear relationship between local variance and the image-specific bias $\hat{p}^{(0)}$. There is no evidence that high local variance is responsible for bias.

To conclude this section, we have validated the large influence of local variance on detection accuracy. Moreover, we have shown that the overall impact is largely due to the sensitivity of the within-image error and only to a much lesser extent to the sensitivity of the between-image error. Finally, there is no evidence for a linear relationship between local variance and image-specific bias.

4.3. Influence of saturation

The last analyses deal with the influence of saturation in images. There is anecdotal evidence that steganalysis, and detection of LSB overwriting in particular, profits from the existence of saturated areas in the carrier images. The influence of saturation on the accuracy of quantitative estimators of hidden message length, however, is likely

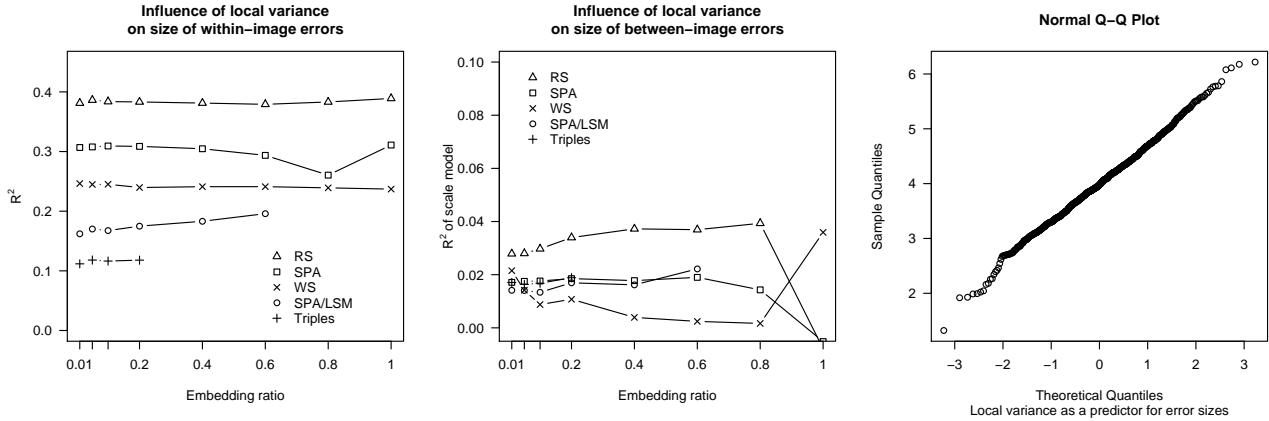


Figure 6. Local variance as predictor for estimation errors. Results from linear models with local variance as predictor for the size of within-image errors (left) and between-image errors (center) for different detectors and embedding rates. The R^2 statistic is a measure for the ratio of variance in the dependent variable explained by the model. Higher values imply better model fits. Surprisingly, local variance of our test data could be transformed to an almost normal distributed predictor variable by computing the fourth root (see QQ-plot, right). Data from never-compressed grayscale images only.

to be much more complicated. We use regression analysis to assess the influence of saturation on both types of errors in our two-factor model.

We define an independent variable *saturation* s_i as the ratio of pixels in carrier $c^{(i)}$ with intensity $c_{k,l}^{(i)} \in \{0, 255\}$. To assess the influence of **saturation on the magnitude of the within-image error**, we formulate a model similar to Eq. (7):

$$\log(\hat{\sigma}_i) = b_0 + b_1 \cdot s_i + \varepsilon, \quad \varepsilon \sim N(0, \sigma_\varepsilon) \quad (11)$$

Note that we analyse only the colour covers' red channel, because grayscale images might be affected from singular histogram effects due to saturation in individual colour channels. The estimated parameters are reported as model (a) in Tab. 6. The coefficients b_1 are significant and negative for all detectors, indicating that detection accuracy profits from large saturated errors. These improvements account for between 4% (WS) and 19% (Triples) of the variance in $\log(\hat{\sigma})$.

Modeling the **influence of saturation on dispersion of the between-image error** follows the methodology from Eq. (9).

$$\begin{aligned} \hat{p}_i^{(0)} &= a_0 + \varepsilon_i, & \varepsilon_i &\sim t_2(0, \lambda) \\ \log(\lambda^2) &= b_0 + b_1 \cdot s_i + e, & e &\sim t_2(0, \cdot) \end{aligned} \quad (12)$$

The results in Tab. 7, model (a), indicate positive and highly significant coefficients b_1 for all detectors except Triples. This implies that all but Triples suffer from large saturated areas in terms of between-image error, which might compensate (part of) the improvements of the within-image error. The R^2 statistics—where the same limitations apply as in Sect. 4.2—are somewhat lower than for the within-image error, but not seriously so.

To analyse the influence of **saturation on bias of the between-image error** we fit the following model:

$$\begin{aligned} \hat{p}_i^{(0)} &= a_0 + a_1 \cdot s_i + \varepsilon_i, & \varepsilon_i &\sim t_2(0, \lambda) \\ \log(\lambda^2) &= b_0 + e, & e &\sim t_2(0, \cdot) \end{aligned} \quad (13)$$

As all coefficients a_1 turn out to be positive and significant (see model (b) of Tab. 7), we conclude that saturation causes a tendency to overestimate the actual secret message length (this is a rather surprising result and we must temper the conclusion by the possibility that it is due to an artefact of the cover images we tested). Comparing the detectors, Triples reacts least sensitive to adverse influence from saturation, again.

Table 6. Influence of saturation on detection accuracy; estimated regression coefficients for $p = 0.05$ of red channel of never-compressed colour images; positive values b_1 denote a loss in accuracy when the cover has a higher proportion of saturated pixels.

Parameter	Detector				
	RS	WS	SPA	SPA/LSM	Triples
Ratio of saturated pixels on std. dev. of within-image error					
Model (a): Regular saturation at intensity range limits					
b_0	-5.24 (0.017)	-5.78 (0.017)	-5.90 (0.017)	-5.90 (0.016)	-5.46 (0.015)
b_1	-4.38 (0.490)	-2.96 (0.498)	-4.37 (0.483)	-4.59 (0.471)	-5.90 (0.435)
R^2	0.09	0.04	0.09	0.11	0.19
n	800	800	800	792	793
Predictor range	0.0–44.6 %	0.0–44.6 %	0.0–44.6 %	0.0–44.6 %	0.0–44.6 %
Model (b): Singular saturation measured as proportion of mode intensity					
b_0	-5.09 (0.019)	-5.67 (0.020)	-5.76 (0.018)	-5.76 (0.018)	-5.30 (0.016)
b_1	-5.86 (0.359)	-4.45 (0.379)	-5.74 (0.355)	-5.86 (0.352)	-6.60 (0.312)
R^2	0.25	0.15	0.25	0.26	0.36
n	800	800	800	792	793
Predictor range	0.6–47.0 %	0.6–47.0 %	0.6–47.0 %	0.6–47.0 %	0.6–47.0 %
Std. errors in brackets; all coefficients significant with $P(\alpha) < 0.001$					

Finally, we experimented with an alternative indicator for saturation artifacts that is robust to possible re-scaling of the intensity values. Therefore we re-estimate models (11), (12) and (13) with s_i replaced by the ratio of pixels at the mode intensity, s'_i .

The results are reported as model (b) in Tab. 6 and models (c) and (d) in Tab. 7 for influence of s'_i on within-image error, between-images error and bias, respectively. Regarding the coefficients, the alternative predictor shows the same direction and strength of linear relationship as the saturation measure s_i . The R^2 statistics, however, are considerably higher for the models on error magnitude, which indicates an improved explanatory power of the alternative predictor. Additional regression models including both s_i and s'_i at the same time revealed that both indicators are highly confounded and deliver similar information (models omitted here due to space limitations).

To conclude this section, saturation in carrier images first reduces the magnitudes of within-image errors, second increases the dispersion of between-images errors, and third causes overestimation of the actual message length. A robust measure of saturation has been proposed as the proportion of pixels at the mode intensity. This predictor shows similar influence pattern on bias and error magnitudes, but explains more of the deviation in the error magnitudes than saturation at intensities $\{0, 255\}$.

5. CONCLUSION

We have presented a rationale for a two-factor model for sources of error in quantitative steganalysis, and shown evidence from a nested design, using systematic experimental evidence in great quantity. We have found that the error caused by random hidden messages—which is usually disregarded in the benchmarking of steganalysis methods—has a Gaussian distribution, whereas the error caused by incorrect cover assumptions is long-tailed and well-modelled by the Student- t family. Further, the magnitude of deviation caused by the first error is significant when compared with the second, at least when the hidden message is of moderate length, so researchers should not continue to disregard it.

The relative performance of five quantitative steganalysis estimators has been evaluated. We have shown that some steganalysis attacks are more prone to one type of error than the other. There may be important

Table 7. Influence of saturation on detection accuracy; estimated regression coefficients for $\hat{p}^{(0)}$ of red channel of never-compressed colour images; positive values b_1 in models (a) and (b) denote a loss in accuracy when there are a higher proportion saturated pixels; positive significant values of a_1 in models (c) and (d) indicate overestimation.

Parameter	Detector				
	RS	WS	SPA	SPA/LSM	Triples
Ratio of saturated pixels on dispersion of between-image error					
Model (a): Regular saturation at intensity range limits					
a_0	0.00 (0.001) ***	0.00 (0.001) ***	0.00 (0.001) ***	0.00 (0.001) ***	0.00 (0.001)
b_0	-8.02 (0.082) ***	-8.64 (0.082) ***	-8.81 (0.082) ***	-8.62 (0.083) ***	-8.56 (0.082) ***
b_1	13.13 (2.361) ***	17.88 (2.361) ***	18.01 (2.361) ***	13.77 (2.372) ***	-0.08 (2.367)
R^2	0.01	0.02	0.02	0.01	0.00
-log likelihood	1646	1792	1895	1781	1851
n	800	800	800	782	785
Model (b): Singular saturation measured as proportion of mode intensity					
a_0	0.00 (0.001) ***	0.00 (0.001) ***	0.00 (0.001) ***	0.00 (0.001) ***	-0.00 (0.001)
b_0	-8.31 (0.099) ***	-9.25 (0.099) ***	-9.33 (0.099) ***	-9.11 (0.100) ***	-8.81 (0.100) ***
b_1	12.91 (1.904) ***	25.02 (1.904) ***	22.33 (1.904) ***	20.21 (1.951) ***	8.49 (1.914) ***
R^2	0.03	0.08	0.07	0.06	0.01
-log likelihood	1654	1831	1923	1806	1858
n	800	800	800	782	785
Ratio of saturated pixels on bias of between-image error					
Model (c): Regular saturation at intensity range limits					
a_0	0.00 (0.001) **	0.00 (0.001) **	0.00 (0.001) ***	0.00 (0.001) ***	-0.00 (0.001)
a_1	0.24 (0.025) ***	0.15 (0.019) ***	0.18 (0.018) ***	0.13 (0.019) ***	0.05 (0.019) **
b_0	-7.96 (0.079) ***	-8.54 (0.079) ***	-8.71 (0.079) ***	-8.55 (0.080) ***	-8.58 (0.080) ***
R^2	0.06	0.04	0.06	0.04	0.01
-log likelihood	1657	1793	1901	1787	1857
n	800	800	800	782	785
Model (d): Singular saturation measured as proportion of mode intensity					
a_0	-0.00 (0.001)	-0.00 (0.001)	0.00 (0.001)	0.00 (0.001)	-0.00 (0.001) **
a_1	0.16 (0.021) ***	0.15 (0.015) ***	0.13 (0.014) ***	0.13 (0.016) ***	0.09 (0.015) ***
b_0	-7.94 (0.079) ***	-8.53 (0.079) ***	-8.68 (0.079) ***	-8.54 (0.080) ***	-8.60 (0.080) ***
R^2	0.05	0.06	0.05	0.07	0.04
-log likelihood	1653	1795	1897	1791	1865
n	800	800	800	782	785

Std. errors in brackets; sig. levels: *** $P(\alpha) < 0.001$, ** $P(\alpha) < 0.01$, * $P(\alpha) < 0.05$

lessons, here, for the construction of future steganalysis estimators, as we gain a better appreciation of why the current estimators fail.

Further, we have examined how the different detectors' performance is affected by properties of cover images, including local variance and saturation. The findings presented here show the need for a careful assessment of both sources of error, and their respective moderating factors, when benchmarking steganalytic techniques. These, and more complex factor analyses, are the subject of further study. A practical application of this work is in the calculation of a confidence interval for steganalysis estimators, based on observed properties of the image under consideration. This should lead to more reliable steganalysis.

ACKNOWLEDGMENTS

The second author is a Royal Society University Research Fellow.

REFERENCES

1. J. Fridrich, M. Goljan, D. Hoge, and D. Soukal, "Quantitative steganalysis of digital images: Estimating the secret message length," *ACM Multimedia Systems Journal* **9**, pp. 288–302, 2003.
2. J. Fridrich, M. Goljan, and R. Du, "Detecting LSB steganography in color and grayscale images," *IEEE Multimedia* **8**(4), pp. 22–28, 2001.
3. S. Dumitrescu, X. Wu, and Z. Wang, "Detection of LSB steganography via sample pair analysis," *IEEE Trans. of Signal Processing* **51**, pp. 1995–2007, 2003.
4. P. Lu, X. Luo, Q. Tang, and L. Shen, "An improved sample pairs method for detection of LSB embedding," in *Information Hiding (6th International Workshop)*, J. Fridrich, ed., *LNCS 3200*, pp. 116–127, Springer Verlag, (Berlin Heidelberg), 2004.
5. J. Fridrich, M. Goljan, and D. Soukal, "Higher-order statistical steganalysis of palette images," in *Security, Steganography and Watermarking of Multimedia Contents V (Proc. of SPIE)*, E. J. Delp and P. W. Wong, eds., pp. 178–190, (San Jose, CA), 2003.
6. J. Fridrich and M. Goljan, "On estimation of secret message length in LSB steganography in spatial domain," in *Security, Steganography and Watermarking of Multimedia Contents VI (Proc. of SPIE)*, E. J. Delp and P. W. Wong, eds., (San Jose, CA), 2004.
7. A. D. Ker, "Improved detection of LSB steganography in grayscale images," in *Information Hiding (6th International Workshop)*, J. Fridrich, ed., *LNCS 3200*, pp. 97–115, Springer Verlag, (Berlin Heidelberg), 2004.
8. A. D. Ker, "A general framework for structural steganalysis of LSB replacement," in *Information Hiding (7th International Workshop)*, M. Barni et al., ed., *LNCS 3727*, pp. 296–311, Springer Verlag, (Berlin Heidelberg), 2005.
9. X. Zhang, S. Wang, and K. Zhang, "Steganography with least histogram abnormality," in *Computer Network Security (MMM-ACNS)*, V. Gorodetsky, L. J. Popyack, and V. A. Skormin, eds., *LNCS 2776*, pp. 395–406, Springer Verlag, (Berlin Heidelberg), 2003.
10. R. Böhme, "Assessment of steganalytic methods using multiple regression models," in *Information Hiding (7th International Workshop)*, M. Barni et al., ed., *LNCS 3727*, pp. 278–295, Springer Verlag, (Berlin Heidelberg), 2005.
11. P. Royston, "An extension of Shapiro and Wilk's test for normality to large samples," *Applied Statistics* **31**, pp. 115–124, 1982.
12. J. Taylor and A. Verbyla, "Joint modelling of location and scale parameters of the t distribution," *Statistical Modelling* **4**, pp. 91–112, 2004.
13. S. Demarta and A. J. McNeil, "The t copula and related copulas," *International Statistical Review* **71**(1), pp. 111–129, 2005.