# Vers une stéganalyse certifiée pour des images JPEG

Etienne Levecque, Jan Butora, John Klein, Patrick Bas

# Vers une stéganalyse certifiée pour des images JPEG

Etienne LEVECQUE, Jan BUTORA, John KLEIN, Patrick BAS

Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France
prenom.nom@univ-lille.fr

**Résumé –** La stéganalyse consiste à détecter des images contenant un message caché. Nous cherchons ici à développer une méthode de stéganalyse qui permette de certifier la stéganalyse en contrôlant la probabilité de faux positifs associée. Pour cela nous nous focalisons sur des images compressées en JPEG avec un facteur de qualité de 100. Pour ce facteur de qualité la distribution des erreurs d'arrondies après décompression de ces images, lorsqu'elles ne contiennent pas de message caché, est modélisable de manière précise comme un mélange de 60 distributions uniformes indépendantes et de 8 distributions dépendantes. Nous utilisons ensuite cette modélisation pour construire un test de Kolmogorov-Smirnov, qui nous permet de maîtriser les probabilités de faux positifs associées. Nos résultats sur la base d'images ALASKA illustrent l'importance de la sélection des blocs et des pixels associés pour pouvoir contrôler le taux de faux positifs. La comparaison avec d'autres méthodes de stéganalyse met également en évidence le compromis qu'il existe entre la volonté de maîtriser le taux d'erreur et la puissance du détecteur construit.

**Abstract –** In this paper our goal is to develop a certified steganalysis scheme by controlling its associated false positive rate. In order to do so, the steganalysis scheme targets JPEG images compressed using quality factor 100. This is due to the fact that with this quality factor, the distribution of the rounding errors after decompression of the Cover images can be precisely modeled by a mixture of 60 uniforms and independent distributions plus 8 dependent distributions. This knowledge is thereafter used to build a Kolmogorov-Smirnov test which is used both to compute the false positive rate and to perform steganalysis. Our results on the ALASKA dataset highlight the role of carefully selecting the adequate blocks and samples in order to control the false positive rate. Moreover, the comparison with other steganalysis schemes shows the trade-off between the will to control the false positive rate and the power of the build detector.

## 1 Introduction

La stéganographie consiste en l'insertion de messages cachés dans des contenus anodins comme des images et elle cherche à être indétectable. Au contraire la stéganalyse vise à analyser des images afin de détecter la présence de messages cachés, et elle peut être en pratique utilisée afin de détecter des communications sensibles. Pour aller jusqu'à la décision, le stéganalyste doit pouvoir avoir des garanties sur les taux d'erreurs de son détecteur. Dans ce papier, nous cherchons à calculer théoriquement le taux de faux positifs d'un détecteur en stéganalyse afin d'aller vers une stéganalyse certifiée, c'est-à-dire qui offre des garanties en termes de probabilités d'erreurs.

Le problème est d'importance puisque les schémas de stéganalyse de l'état de l'art [6] souffrent d'une très grande sensibilité des performances par rapport à la base d'entraînement utilisée. À titre d'exemple, il a été montré dans [4] qu'un détecteur entraîné à partir d'une source constituée d'une base d'images développées avec un logiciel donné sera efficace sur une base test également développée avec ce même logiciel. Il pourra cependant devenir complètement inefficace si les images testées sont développées avec un autre logiciel, alors que le contenu visuel entre les deux bases est extrêmement proche. Ce problème (appelé CSM en anglais pour *Cover Source Missmatch*) est due à l'extrême diversité des signaux faibles (tels que le bruit photonique) présents dans les images, et à leur dépendance à la chaîne de développement utilisée. Ces schémas de stéganalyse, aussi performants soient-ils lorsque la source testée est identique à la source utilisée lors de l'apprentissage, deviennent alors inopérants lorsque la source diffère. Si dans certains cas il est possible d'estimer les probabilités d'erreurs de ces schémas pour une source donnée via des simulations de Monte-Carlo, ces estimations sont impossibles lorsque la source est différente.

De surcroît, le stéganalyste est souvent sujet à un problème pratique lorsque le taux de faux positif visé est trop faible puisque le seul moyen d'estimer de ce taux d'erreur est d'utiliser un estimateur de Monte Carlo qui potentiellement nécessite d'acquérir un très grand nombre d'images pour obtenir une estimation fiable.

L'objectif de ce papier est de proposer une méthode de stéganalyse qui permette de garantir *a priori* un taux de faux positif donné. Pour cela il est nécessaire de trouver un domaine de représentation des images analysées qui remplisse 3 conditions : (i) il doit être insensible aux changements de sources, (ii) il doit être sensible à l'insertion stéganographique quelque-soit la stratégie d'insertion, (iii) il doit être associé à un test statistique dont la fonction de répartition est connue.

Ce papier propose de répondre à ces trois objectifs, il se rapproche de la contribution [3] qui est cependant spécifique à la stéganographie $\pm 1$ et dont le test par rapport de vraisem-

blance généralisé est sensible à la source. A notre connaissance, il s'agit de la première contribution qui attaque de front ces trois contraintes. Nous présentons dans un premier temps le test utilisé pour contrôler le taux d'erreur (section 2), puis la façon dont nous modélisons les images Cover dans la section 3. Cette modélisation est testée sur des images naturelles dans la section 2. Enfin la section 4 présente les résultats obtenus sur la base ALASKA en termes de contrôle de taux de faux positifs, mais aussi de stéganalyse.

**Notations** : dans cet article, les images cover sont notées $\mathbf{c}$ et celles stégo $\mathbf{s}$. Ces ensembles de données sont notés $\mathcal{C}$ et $\mathcal{S}$ respectivement. Chaque pixel d'une image est indexé par une paire d'entier $(i,j)$, i.e. $\mathbf{c} = (c_{ij})_{(i,j)=(1,1)}^{(H,W)}$ où $H$ et $W$ correspondent à la taille de l'image.

# 2 Test de Kolmogorov-Smirnov

Le test de Kolmogorov-Smirnov (KS) est l'instrument choisi pour assurer un contrôle de la probabilité de fausse alarme (PFA) et s'appuie sur une grandeur $x$ tirée d'une image. Pour des raisons qui seront présentées dans la section suivante, cette grandeur correspondra dans notre cas à des erreurs d'arrondi après décompression JPEG. Le test KS est un test d'adéquation entre un échantillon et une loi de probabilité (test simple) ou entre deux échantillons (test double). La statistique de ce test est la borne supérieure de la distance verticale entre les fonctions de répartition (empirique dans le cas d'un échantillon). Dans le test simple, $D_n = \sup_x |F_n(x) - F(x)|$ et dans le test double, $D_{n,m} = \sup_x |F_n(x) - F_m(x)|$ avec $n$ et $m$ la taille des échantillons.

Notons $\mathcal{H}_0$ et $\mathcal{H}_1$ l'hypothèse nulle (cover) et alternative (stego) respectivement. La PFA $\alpha$ est la probabilité d'accepter $\mathcal{H}_1$ alors que la vérité est $\mathcal{H}_0$ : $\alpha = P_{\mathcal{H}_0}(\mathcal{H}_1)$. En choisissant judicieusement l'hypothèse nulle, pour tout échantillon $(x_i)_{1\leq i\leq n}$, on peut obtenir une statistique $D_n$ et appliquer la règle de classification suivante :

$$(x_i)_{1\leq i\leq n} \in \begin{cases} \mathcal{C} \text{ si } \sqrt{n}D_n \leq K_\alpha, \\ \mathcal{S} \text{ si } \sqrt{n}D_n > K_\alpha. \end{cases}$$

Si $n$ est très grand, le seuil $K_\alpha$ peut être déterminé de manière asymptotique à l'aide de la fonction de répartition de la loi de Kolmogorov. Si $K$ est une variable suivant cette loi, on définit $K_\alpha$ tel que $P(K \leq K_\alpha) = 1-\alpha$. Pour $n$ plus petit, il existe des approximations de la fonction de répartition [5] afin d'obtenir $K_\alpha$ tel que $P(\sqrt{n}D_n \leq K_\alpha) = 1-\alpha$.
Par construction, la PFA de ce classifeur sera au plus $\alpha$. Dans le cas du test double, il est aussi possible de construire un classifieur avec une PFA fixe. Pour cela il faut fixer un échantillon de référence et comparer les échantillons $(X_i)_{1\leq i\leq n}$ à cet échantillon de référence et la règle de décision pour $\mathcal{C}$ s'écrit $\sqrt{\frac{n\times m}{n+m}}D_{n,m} \leq K_\alpha$.

# 3 La stéganalyse RJCA revisitée

Cette méthode de stéganalyse modélise l'erreur d'arrondi des pixels d'images après décompression JPEG à un facteur de qualité 100. Dans un premier temps, il convient de modéliser la distribution de cette erreur à partir de l'erreur de quantification des coefficients DCT produite lors de l'étape de compression.

## 3.1 Retours sur le modèle RJCA

On suppose que le stéganalyste a accès à $\tilde{\mathbf{c}}$ ou $\tilde{\mathbf{s}}$ qui sont des images encodées en JPEG avec un facteur de qualité de 100. Dans ce cas, chaque image est divisée en blocs de taille $8\times 8$. Chaque bloc subit indépendamment une transformée en cosinus discrète : $\tilde{\mathbf{c}} = \text{DCT}\{\mathbf{c}\}$, on se place donc dans un bloc quelconque. Finalement, toujours avec un facteur de qualité de 100, le coefficient JPEG s'obtient par $[\tilde{\mathbf{c}}]$ où $[.]$ la fonction d'arrondi à l'entier le plus proche. On ignorera à ce niveau l'opération de seuillage des entiers entre 0 et 255.

Une hypothèse de travail très raisonnable pour la suite est de supposer que l'erreur d'arrondi dans le domaine DCT suit une loi uniforme sur l'intervalle $\left[-\frac{1}{2};\frac{1}{2}\right]$ :

$$u_{ij} = \tilde{c}_{ij} - [\tilde{c}]_{ij} \sim \mathcal{U}_{\left[-\frac{1}{2};\frac{1}{2}\right]}, \tag{1}$$

et ce pour tout entier $i$ et $j \in \{1,..,8\}$ et tout bloc. En outre, l'image décompressée, notée $\hat{\mathbf{c}}$, s'obtient par transformée inverse sur chaque bloc selon $\hat{\mathbf{c}} = \text{DCT}^{-1}\{[\tilde{\mathbf{c}}]\}$. Si on note $\mathbf{v} = \text{DCT}^{-1}\{\mathbf{u}\}$, il s'ensuit que :

$$\mathbf{v} = \mathbf{c} - \hat{\mathbf{c}}. \tag{2}$$

Le raisonnement formulé dans [1] conduit à présent à faire une hypothèse d'indépendance entre toutes les variables $u_{ij}$ et à exploiter le théorème centrale limite (CLT) afin d'affirmer qu'une très bonne approximation de la distribution d'un $\hat{v}_{ij}$ sera une gaussienne $\mathcal{N}(0, s_{ij})$ avec la variance $s_{ij} = \frac{1}{12}\sum_{m,n=1}^{8}\left(w_{mn}^{ij}\right)^2$ où les $w_{mn}^{ij}$ sont les coefficients utilisés dans le calcul de la DCT. Qui plus est, comme les $c_{ij}$ sont déterministes et entiers, cela revient aussi à affirmer que pour un pixel décompressé, on a $\hat{c}_{ij} \sim \mathcal{N}(c_{ij}, s_{ij})$. Finalement, l'erreur $e_{ij}$ d'arrondi dans le domaine pixel après décompression peut s'écrire :

$$\begin{aligned} e_{ij} &= \hat{c}_{ij} - [\hat{c}_{ij}], && (3)\\ &= c_{ij} - v_{ij} - [c_{ij} - v_{ij}], && (4)\\ &= [v_{ij}] - v_{ij}. && (5) \end{aligned}$$

Ainsi, les erreurs $e_{ij}$ correspondent à une erreur d'arrondi d'une variable gaussienne de distribution connue. Or, la loi d'une telle erreur possède elle-même une distribution connue, appelée « gaussienne repliée », qu'on notera $\mathcal{N}_r(0, s_{ij})$.

Dans [1], les auteurs notent que le même raisonnement appliqué aux images stégo aboutit au même modèle de gaussienne repliée pour l'erreur d'arrondi, mais que la variance de cette distribution est plus important de par la présence des modifications $\pm 1$ effectuées. Un simple test sur la variance des erreurs d'arrondi s'avère très discriminant.

Afin d'obtenir un test certifié, on peut penser exploiter le modèle des lois uniformes indépendantes dans le cadre d'un test KS tel que présenté dans la section précédente. Malheureusement, en visant des taux de faux positifs très faibles, il apparaît que l'hypothèse d'indépendance est trop approximative dans notre cas.

### 3.2 Modèle synthétique

La première contribution de cet article est de proposer un modèle plus précis de la distribution des erreurs d'arrondis après compression JPEG contrôlée à un facteur de qualité de 100. Le compresseur JPEG analysé est une implémentation Python de la DCT-II sous forme matricielle. Une observation approfondie des $u_{ij}$ relève qu'ils ne sont pas tous uniformes et pas non plus tous indépendants.

- Les modes $u_{00}, u_{04}, u_{40}, u_{44}$ suivent une loi uniforme discrète à 8 valeurs possibles : $\left\{\frac{i}{8},\, i \in [\![-4;4]\!]\right\}$. De plus, ces 4 modes sont dépendants puisque la somme des 4 modes est forcément un multiple de 0.5.

- Les modes $u_{22}, u_{66}$ suivent chacun une loi uniforme continue sur $]-0.5;0.5]$ et $u_{22}+u_{66}$ suit une loi uniforme discrète à valeur dans les multiples de 0.25.

- Les modes $u_{26}, u_{62}$ suivent chacun une loi uniforme continue sur $]-0.5;0.5]$ et $u_{26}-u_{62}$ suit une loi uniforme discrète à valeur dans les multiples de 0.25.

- Finalement, tous les autres modes suivent une loi uniforme continue sur $]-0.5;0.5]$ et semblent indépendants.

Une fois que les modes $u_{ij}$ sont correctement construits, on peut prendre la DCT inverse pour obtenir les distributions des $e_{ij}$. Afin d'utiliser le test KS simple, on a besoin de connaître une version analytique de la fonction de répartition des erreurs d'arrondis. Cependant, ce calcul est très complexe puisqu'il nécessite la convolution de 56 lois uniformes avec des supports différents plus la convolution des lois jointes.

### 3.3 Validation par test KS double

On peut toutefois échantillonner des $e_{ij}$ à partir de ce modèle et les comparer à des erreurs d'arrondis issues d'une vraie compression JPEG. Pour cela, on génère une image cover aléatoire $\mathbf{c}$, tel que pour toute paire d'entiers $(i,j)$, $\mathbf{c}_{ij} \overset{iid}{\sim} \mathcal{U}([\![0;255]\!])$. En passant ces images dans le compresseur JPEG à un facteur de qualité 100, on génère les erreurs d'arrondis du compresseur. Ces échantillons, notés $e^{ref}$ servent de référence dans le test KS double. D'autre part, on génère des erreurs d'arrondis à l'aide du modèle synthétique, notées $e^{model}$ et des erreurs d'arrondis suivant la normale repliée, notées $e^{nr}$. La DCT induit des distributions différentes pour chacune des 64 positions du bloc $8\times 8$ donc il y a 64 test à faire. La figure 1 montre que le modèle synthétique tangente l'erreur relative nulle et son intervalle de confiance comprend l'erreur nulle, ce qui n'est pas le cas de la loi normale repliée. La divergence vient de la variance qui croit avec $\alpha$ et de la somme de valeurs absolues qui sanctionne les erreurs relatives négatives.

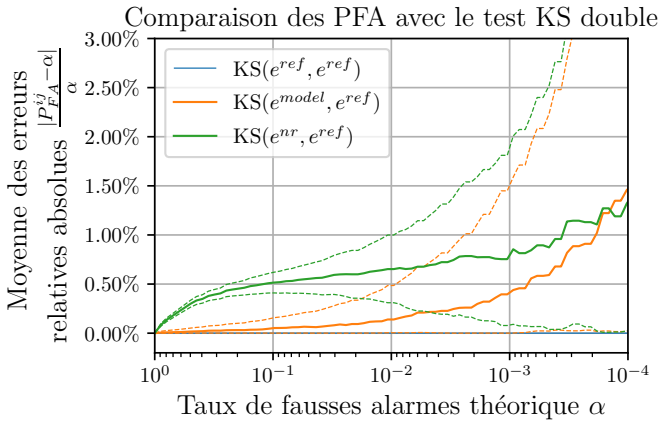


FIGURE 1 – PFA du test KS double pour différent modèles d'erreur d'arrondi.Les pointillés représentent l'intervalle de confiance à 95% construit sur 60 séries utilisants $5.10^5$ échantillons.

## 4 Evaluation sur des images naturelles

### 4.1 Evaluation du modèle

Le modèle synthétique de la partie précédente semble correct dans le cas théorique, il faut maintenant le valider sur des images naturelles. Pour cela on s'appuie sur la base ALASKA [2], composée de 80k images de taille $256\times256$. Ces images sont disponibles en uint8 ce qui nous permet d'utiliser le même compresseur JPEG que dans nos expériences. Si on suppose que la seule information inconnue du stéganalyste est la clé d'insertion, cette hypothèse de travail reste conforme. Le but ici est d'évaluer ce modèle en présence de contenu dans l'image.

Le test KS double montre clairement que le contenu de l'image (jusque là ignoré) a un impact sur la distribution des erreurs d'arrondis. Cependant, il est possible de sous-échantillonner l'image pour casser la dépendance locale induite par le contenu. Cependant, plus on utilise d'échantillons, meilleur est le test KS. Il y a donc un compromis à trouver entre la performance de détection et la fiabilité théorique du modèle.

Parmi les stratégies de sélection essayées, le plus important limiter toute dépendance entre les échantillons est de n'utiliser qu'un seul pixel par bloc DCT. Ensuite, pour la sélection des blocs, deux stratégies semblent fonctionner : soit prendre 1 bloc sur 2, soit prendre les blocs avec la plus grande variance. En effet, plus la variance du contenu est élevée, plus le contenu local sera variant et ainsi les erreurs d'arrondis moins dépendantes. La Figure 2 montre que deux stratégies apportent des garanties de PFA valables pour $\alpha \leq 10^{-3}$.

### 4.2 Stéganalyse

On construit les images stégo à partir des images cover de la base ALASKA en utilisant l'algorithme J-UNIWARD avec des

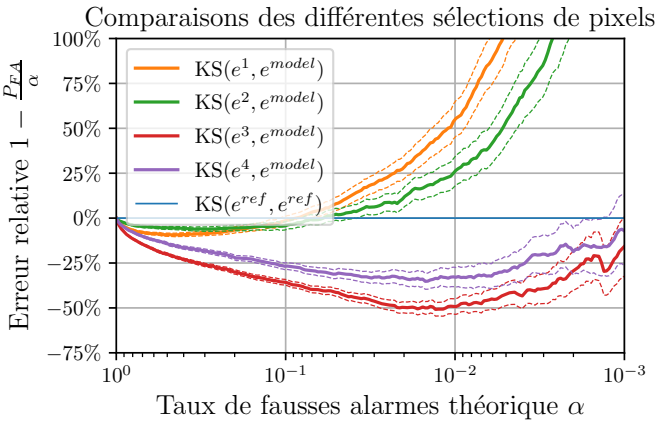


FIGURE 2 – Performances de différentes stratégies de sélection des pixels. Les différents échantillons sont construits comme suit, $e^1$ : tous les blocs DCT, 10 pixels par bloc. $e^2$ : tous les blocs, 1 pixel par bloc. $e^3$ : 1 bloc sur 4, 1 pixel par bloc. $e^4$ : les 500 blocs à variance de contenu élevée, 1 pixel par bloc.

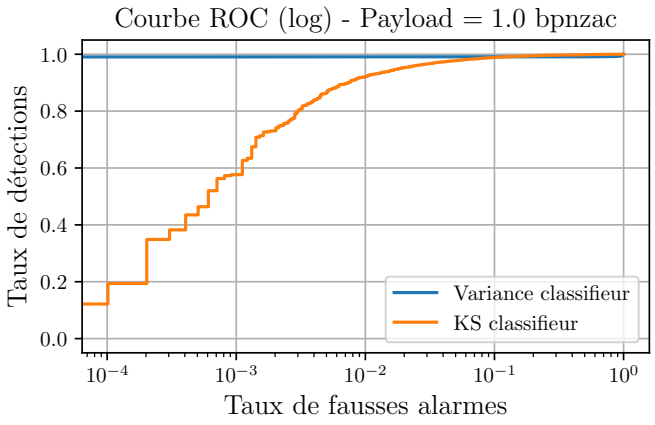


FIGURE 3 – Courbe ROC du classifieur de variance d'erreurs d'arrondis (orange) et du classifieur avec le test KS double sur une position aléatoire dans le bloc $8 \times 8$ à 1.0 bpnzac.

payloads variant de 0.1 à 1.0 bpnzac (bit par non-zéro AC DCT coefficients). A titre de comparaison, nous avons utilisé un classifieur de variance des erreurs d'arrondis de l'image [1], cet attribut étant très discriminant. Notre classifieur construit sur le test KS peut être entraîné à l'aide de quelques images cover afin de construire un fonction de répartition de référence mais il est aussi possible de construire cette fonction de référence à l'aide du modèle synthétique et ainsi de n'utiliser aucune forme d'entraînement. C'est l'approche que nous privilégions dans la suite. Pour sélectionner les pixels des images à classifier, on utilise la stratégie 3 : 1 bloc sur 4 et 1 pixel par bloc. Elle a la meilleure garantie de PFA jusque $\alpha \leq 10^{-3}$. La figure 3 montre que les performances de notre classifieur ne sont pas aussi bonnes que le classifieur de variance. Cependant on observe des taux de détection non nul à des PFA guaranties ($\alpha \leq 10^{-3}$) par les résultats préliminaires de la section 4. Cela signifie que la classification n'est pas triviale.

## 5 Conclusions et perspectives

Ces premiers travaux montrent qu'il est possible de construire des classifieurs dont la PFA est garantie pour un facteur de qualité 100. Pour ce faire on utilise les erreurs d'arrondis insensibles au changement de source (i), sensible à l'insertion stéganographique (ii) qu'on classifie à l'aide du test KS dont on connaît la fonction de répartition (iii).

Certes les performances n'atteignent pas l'état de l'art, mais il reste beaucoup de marge d'amélioration. Premièrement, connaître la formule analytique de la fonction de répartition de l'erreur d'arrondi permettrait d'utiliser le test KS simple et d'être plus discriminant. Deuxièmement, les 64 tests KS pourrait être agrégés pour obtenir un test unique pour chaque image. Finalement, une étude approfondie du lien entre contenu et erreur d'arrondi permettra de trouver des meilleures stratégies de sélection des pixels et ainsi d'avoir des modèles plus discriminant (en utilisant plus de pixels) et davantage de guaranties pour les faibles PFA.

## 6 Remerciements

Les travaux présentés dans ce papier ont également reçu un financement du programme H2020 de l'Union Européenne, accord de financement No 101021687, projet "UNCOVER".

## Références

[1] Jan Butora and Jessica Fridrich. Reverse jpeg compatibility attack. *IEEE Transactions on Information Forensics and Security*, 15 :1444–1454, 2019.

[2] Rémi Cogranne, Quentin Giboulot, and Patrick Bas. The ALASKA Steganalysis Challenge : A First Step Towards Steganalysis "Into The Wild". In *ACM IH&MMSec*, Paris, France, July 2019.

[3] Rémi Cogranne and Florent Retraint. An asymptotically uniformly most powerful test for lsb matching detection. *Information Forensics and Security, IEEE Transactions on*, 8(3) :464–476, 2013.

[4] Quentin Giboulot, Rémi Cogranne, Dirk Borghys, and Patrick Bas. Effects and Solutions of Cover-Source Mismatch in Image Steganalysis. *Signal Processing : Image Communication*, August 2020.

[5] Jan Vrbik. Deriving cdf of kolmogorov-smirnov test statistic. *Applied Mathematics*, 11(3) :227–246, 2020.

[6] Yassine Yousfi, Jan Butora, Jessica Fridrich, and Clément Fuji Tsang. Improving efficientnet for jpeg steganalysis. In *Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security*, pages 149–157, 2021.