Pseudo-réalisme et progressivité pour le tracé de rayons

Jean-Luc Maillot

ECOLE NATIONALE SUPÉRIEURE
DES MINES DE SAINT-ETIENNE

UNIVERSITÉ JEAN MONNET
DE SAINT-ETIENNE

N° d'ordre : **129ID**

THESE

Présentée par Jean-Luc MAILLOT

pour obtenir le titre de

Docteur

DE L'UNIVERSITE JEAN MONNET DE SAINT-ETIENNE ET DE
L'ECOLE NATIONALE SUPERIEURE DES MINES DE SAINT-ETIENNE

Spécialité Informatique, Synthèse d'images

Pseudo-réalisme et progressivité pour le tracé de rayons

Soutenue à Saint-Etienne le 12 Septembre 1996

Composition du Jury :

Monsieur	B. Peroche	Président
Messieurs	K. Bouatouch	Rapporteurs
	P. Guitton	
	L. Carraro	Examinateurs
	J. Azema	
	G. Drettakis	
	F. Sillion	

A mes parents,
pour leur amour et leurs sacrifices

Remerciements

Tous mes remerciements à MM. Bouatouch et Guitton, pour leur lecture attentive du manuscrit et leur remarques pertinentes tant sur la forme que sur le fond. Leurs commentaires ont permis de clarifier ce document et de rendre plus accessibles certains points difficiles de mes travaux.

Merci à MM. Drettakis, Sillion et Azema pour leur participation au jury et l'éclairage qu'ils ont apporté sur certains parallèles avec des travaux existants et insuffisamment traités dans le manuscrit original.

Merci à Bernard Peroche pour m'avoir accueilli dans son laboratoire et pour ses conseils.

Toute ma reconnaissance à Laurent Carraro pour son aide technique continue et polymorphe, ses conseils experts en la chose statistique (à moins qu'elle ne soit probabiliste, je saisirai peut-être un jour…), sa rigueur scientifique et son pragmatisme qui n'ont d'égal que son sens moral, ce qui n'est pas peu dire ! Nos interminables et multiples discussions, des fort déconcertantes pitreries lumineuses jusqu'à l'indispensable complétude des hypothèses en passant par quelques utopies et autres étrangetés du monde, ont nourrit cette thèse et l'ont fait grandir dans un esprit d'équipe où la confrontation a souvent été une occasion d'avancer (et donc d'avancées !). Enfin, son soutien de longue haleine a permis que mon travail aboutisse.

Mes remerciement à Pierre Igier et Olivier Breuil pour leurs efforts et leur patience dans la reproduction de cette thèse.

Merci à Françoise Sauzet pour son aide efficace dans la course poursuite finale et pour son discret soutien de longue date auquel j'ai été très sensible.

Toute mon affection à ma chère Loulou, dont la tendresse enfantine m'a réconforté maintes et maintes fois et fait fuir de grincheuses sapinous fort méchantes. Mes plus vives félicitations pour ma thèse à Anne et Georges car je l'ai accomplie par la grâce de leur amitié. Merci de m'avoir toujours supporté (dans les deux sens !). Nos longues soirées à refaire le monde, les délicieux tiramisus à double sens d'Anne, ses multiples attentions délicates et sensibles, son ravissement pétillant devant les merveilles enfantines, la profonde bonhomie de Georges, ses Chateaux Latour savoureux de joie de vivre, sa générosité inépuisable, sa simplicité de coeur, son attention affectueuse, leur disponibilité, leur accueil, leur bonté, leur tendresse m'ont profondément touché et m'ont permis de continuer à avancer malgré les difficultés.

Merci à Françoise pour ce que nous avons partagé et pour avoir fait pencher la balance sans le savoir.

Mes chers parents, Jacques, Marie-Paule, même s'il n'est pas toujours facile de comprendre les joies et les affres du thésard, votre amour m'a été d'un grand secours.

Table des matières

Notations i

Introduction 1

––– I. Les fondements physiques –––

1 Electromagnétisme 5
- 1 Les équations de Maxwell 5
- 2 Les ondes planes 6
- 3 Réflexion - réfraction. 9

2 Rugosité 13
- 1 Introduction 13
- 2 Modèles à facettes 16
- 3 Le modèle de Beckmann 19
  - Expression générale du champ réfléchi 19
  - Le coefficient de dispersion 22
  - Surfaces gaussiennes 25
  - Énergie réfléchie dans le modèle gaussien de Beckmann 26
- 4 Auto-ombrage des facettes 29
- 5 Résultats expérimentaux 30
  - Mesure de la rugosité 30
  - Modèle beckmannien comparé à l'expérience 31

3 Bilan 34

4 Radiométrie 35
- 1 Définitions 35
- 2 Réflectances bidirectionnelles 36

5 Photométrie 38
- 1 Égalisation en luminosité 38
- 2 Les lois de la photométrie 41
- 3 Unités de mesures 43

6 Colorimétrie 44

––– II. Modèles d’éclairement –––

1 Modèles d’éclairements locaux 51
1 Premiers modèles 51
2 Inspiration physique 53
3 Approche calculable 58
2 Modèles d’éclairements globaux 61
1 L’équation de rendu 61
2 Méthodes de Monte Carlo 64
Méthode brute 64
Stratification 64
Échantillonnage d’importance 65
Contrôle de variations 65
Applications 66
3 Un modèle pseudo-réaliste 69
1 Présupposés généraux 70
2 Hypothèses sur la surface 70
4 La question du réalisme 71
1 De la synthèse à l’inconscient 71
2 Conclusions 79

––– III. Résolution par Monte-Carlo –––

1 Le problème à résoudre 81
1 Positionnement 81
2 Processus markovien 84
2 Monte-Carlo ou la philosophie du caméléon 85
1 Principes 85
2 Méthode minimale 87
3 La stratégie de la poire 88
1 Partie diffuse 89
2 Cône spéculaire 90
3 Mélange diffus/spéculaire 91
4 Échantillonnage du pixel 93
5 Fiat lux 95
1 De l’attraction du tournesol ou comment tenir compte des sources 95
2 Combinaison lumière/matière 97
6 Échantillonnage a posteriori 97
7 Schéma global 100
8 De la nécessité de conserver l’énergie 103
9 Résultats partiels 105
10 Conclusions 107

––– IV. Tracé de rayons progressif –––

1 Concepts de base ... 109

2 Recherches apparentées ... 110

1 Échantillonnage de l'espace image ... 111
2 Critères et détection du crénelage ... 111
3 Reconstruction ... 112
4 Résultats ... 113
5 Bilan ... 113

3 Nouvelle approche ... 115

1 Progressivité ... 115
2 Analyse séquentielle ... 115
3 Homogénéité ... 117
4 Échantillonnage et reconstruction ... 118
5 Outil de validation ... 120

4 TSRV appliqué à l'image ... 124

1 Taille de la zone de tolérance ... 124
2 Proportion homogène ... 125
3 Structure du panel d'images ... 126
4 Définition du test ... 132
5 Test exact ... 133
Signification des risques ... 133
Région critique ... 134
Premières sorties du test ... 137
Risques réels ... 139
Test exact non biaisé ... 139
Test exact avec recyclage des échantillons ... 139
Test exact et moyenne approchée ... 141
Composition recyclage/moyenne approchée ... 144

5 Tracé de rayons progressif ... 146

1 Test inexact ... 146
Proportion inhomogène comme paramètre ... 146
L'intervalle incertain ... 147
Minimisation des risques ... 148
Décision express ... 148
Test inexact express ... 149
Test inexact express avec recyclage et moyenne approchée ... 150
2 Cohérence spatiale ... 153
3 Résultats ... 155
Performances ... 167
Qualité ... 168
4 Paramètres préconisés ... 169

6 Conclusions ... 169

––– Conclusions –––

1 Quelques mots sur l'implémentation 171

2 Bilan et prospectives 172

1 Méthodes de Monte-Carlo 172

2 Progressivité 173

Bibliographie 175

Crédits photographiques 181

Liste des tableaux 183

Index 185

Notations

$c = a + ib$	$i^2 = -1$	
	$\mathrm{Re}(c) = a$	Partie réelle
	$\mathrm{Im}(c) = b$	Partie imaginaire
	$\lvert c \rvert = \sqrt{a^2 + b^2}$	Module
	$c^* = a - ib$	Conjugué
v, **w** vecteurs	$\angle(\mathbf{v}, \mathbf{w})$	Angle entre **v** et **w**
	$\mathbf{v} \wedge \mathbf{w}$	Produit vectoriel
	$\mathbf{v} \cdot \mathbf{w}$	Produit scalaire
	$v = \lvert \mathbf{v} \rvert = \sqrt{\mathbf{v}_x^2 + \mathbf{v}_y^2 + \mathbf{v}_z^2}$	Module de $\mathbf{v} = [v_x, v_y, v_z]$.
	$\nabla \times \mathbf{v} = \begin{bmatrix} \mathbf{x} & \mathbf{y} & \mathbf{z} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ v_x & v_y & v_z \end{bmatrix}$	Nabla
	$\mathrm{div}\mathbf{v} = \frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z}$	Divergence
	$\mathrm{rot}\mathbf{v} = \begin{bmatrix} \frac{\partial v_z}{\partial y} - \frac{\partial v_y}{\partial z} \\ \frac{\partial v_x}{\partial z} - \frac{\partial v_z}{\partial x} \\ \frac{\partial v_y}{\partial x} - \frac{\partial v_y}{\partial x} \end{bmatrix}$	Rotationnel

	$\lvert a \rvert$	Valeur absolue de a
	e^x	Exponentielle
	$\mathrm{sinc}(x) = (\sin x)/x$	Sinus cardinal
	$\lfloor x \rfloor$	Plus grand entier inférieur ou égal à x
	$\lceil x \rceil$	Plus petit entier supérieur ou égal à x
	$a \gg b$	a très supérieur à b
	$a \ll b$	a très inférieur à b
	$\bar{x}$	Moyenne
	$\mathrm{card}(E)$	Cardinal de l'ensemble E
	$\mathbf{1}_E(x) = \begin{cases} 0 & \text{si } x \notin E \\ 1 & \text{si } x \in E \end{cases}$	Indicatrice
	$x \sim L$	x est une variable aléatoire de loi L
	$\Delta_{uv}(c, d)$	distance euclidienne dans l'espace L*u*v* entre la couleur c et la couleur d

		page
R_o	réflectance à incidence normale	12
$\overline{R}$	réflectance moyenne (dans le temps)	11
$R_{\parallel}$	composante parallèle de la réflectance	11
$R_{\perp}$	composante perpendiculaire de la réflectance	11
$\overline{T}$	transmittance moyenne (dans le temps)	10
$T_{\parallel}$	composante parallèle de la transmittance	10
$T_{\perp}$	composante perpendiculaire de la transmittance	10
$\hat{n} = n - ik$	indice de réfraction complexe	11
L	Luminance énergétique	36
ζ	angle zénithal i.e mesuré depuis la normale	20
f_r	FDRB	36

Unités

cd	candela
H	henry
Hz	hertz
F	farad
K	kelvin
lm	lumen
lx	lux
m	mètre
µm	micromètre (10^{-6} mètre)
nm	nanomètre (10^{-9} mètre)
rad	radian
S	siemens
s	seconde
sr	stéradian
V	volt
W	watt

Introduction

Le cadre de cette thèse est celui de la synthèse d'images dites réalistes dont le but ultime est de produire des images que l'on puisse confondre avec des représentations fidèles de la réalité. Pour réaliser ces images, nous allons proposer des techniques nouvelles basées sur le tracé de rayons, méthode bien connue consistant à simuler le trajet de la lumière entre l'oeil et les sources lumineuses. Nos travaux ont portés plus particulièrement sur deux aspects.

Le premier concerne les méthodes de Monte-Carlo qui permettent de rendre compte de tous les phénomènes d'inter-réflexion lumineuse entre objets. Nous proposerons une formulation précise du processus aléatoire que constitue les rebonds de la lumière sur les surfaces. A partir de cette approche, nous présenterons une méthode de tracé de chemins partant de l'oeil dont le but est d'accélérer la convergence traditionnellement lente de ces méthodes aléatoires.

La complexité de ces calculs, doublée des modèles d'éclairements physiques que nous utilisons, nous amènera au deuxième aspect de nos travaux, à savoir l'accélération du calcul d'une image. Cette accélération, basée sur la notion d'homogénéité d'image, consistera en un échantillonnage de l'écran qui permettra de partager l'image en zones homogènes et inhomogènes afin d'adapter cet échantillonnage à la complexité de ces zones.

L'ensemble de ces travaux est toute entier centré sur la notion de réalisme. Ce thème a parcouru et inspiré l'ensemble de cette thèse, souvent en filigrane, très indistinctement à l'origine, pour progressivement devenir un leitmotiv lancinant. Comment représenter fidèlement la réalité ? Telle est la question récurrente à laquelle nous avons tenté de fournir quelques éléments de réponse.

Nous avons débuté nos recherches sur le thème de la progressivité en nous demandant comment calculer progressivement une image de la façon la plus économe et la plus rapide possible. Cette question admet de multiples réponses, chacune dans des domaines différents du tracé de rayons. Il est en effet envisageable de rendre progressif, entre autres, le processus d'intersection, la génération des points de l'écran ou encore, le calcul du rendu. Nous avons choisi, comme un point de départ logique parce que conditionnant tous les autres, de nous attacher à la génération des points de l'écran. Les réponses que nous avons apportées à ce premier problème, nous ont amené à considérer l'origine de la couleur que nous manipulions et c'est ainsi que nous avons naturellement étudié les modèles d'éclairement.

Comprendre ce qu'est la couleur et donner un sens aux calculs qui y mènent s'est alors imposé à nos réflexions. De là, il n'y avait qu'un pas vers la notion fondatrice de ce sens, à savoir, le réalisme. Le réalisme passe ainsi, pour nous, par la dou-

ble compréhension des phénomènes physiques qui président à la naissance de la lumière et à l'aspect des choses d'une part et des «mécanismes» psycho-visuels qui organisent les sensations visuelles pour leur donner une signification d'autre part.

À notre sens, la synthèse d'image est une science triplement expérimentale. Elle l'est par l'utilisation de modèles physiques qui doivent être adaptés aux capacités de calcul de nos machines et nécessitent donc l'élaboration de méthodes numériques particulières. Elle l'est également par la vérification de ses résultats numériques qui passe par la comparaison entre théorie appliquée et expérimentation physique, une démarche classique en science. Et enfin, ce qui nous paraît un trait de caractère tout à fait spécifique que l'on ne retrouve guère dans d'autres sciences, l'image de synthèse est également expérimentale de par sa destination finale : l'être humain. L'observateur ultime possède ses propres références au réel, son propre système d'interprétation, dépassant tous deux, nous en donnerons quelques aperçus, la «simple» «somme» de ces fonctions physiologiques. C'est une caractéristique que l'on retrouve dans les arts picturaux, à la différence près qu'en synthèse, la fidélité au réel prime sur l'expression pure[†]. C'est ce parcours du physique au psychologique, quelques-unes de ses étapes disons, que nous proposons au lecteur à travers les quatre chapitres de cette thèse.

Le premier chapitre est consacré aux fondements physiques nécessaires au passage des ondes aux sensations visuelles. C'est un vaste programme traversant de nombreux domaines très variés et nous ne ferons qu'imparfaitement effleurer ces sujets fort intéressants, que le lecteur veuille bien nous en excuser. Nous espérons néanmoins donner ainsi quelques jalons indispensables à la modélisation d'une réalité très complexe et inciter le lecteur à s'interroger sur ce que la théorie décrit du réel. Au cours de ce chapitre nous expliquerons quelques aspects importants de l'interaction lumière/matière et nous détaillerons certains développements théoriques permettant de les décrire. Nous nous attacherons également à relier les quantités énergétiques véhiculées par les ondes aux sensations visuelles qu'elles engendrent ce qui nous amènera à faire quelques incursions dans la photométrie et la colorimétrie.

Dans le deuxième chapitre, nous détaillerons quelques-uns des modèles utilisés en synthèse d'image pour décrire une certaine réalité et calculer une image. À la lumière du chapitre précédent, nous examinerons les hypothèses, souvent implicites, qui sous-tendent ces modèles. Nous nous interrogerons sur la validité de certaines d'entre elles et des résultats obtenus. Les techniques de résolution attachées à ces modèles seront également considérées et plus particulièrement les méthodes de Monte-Carlo. Ceci nous conduira à choisir une fonction de transfert d'énergie qui nous permettra de proposer une nouvelle méthode de résolution globale des échanges énergétiques, objet du troisième chapitre. Parce qu'imparfait, nous qualifierons ce modèle de transfert de pseudo-réaliste, ce qui nous conduira à proposer des pistes de réflexions sur ce qu'est le réalisme. Une question peu abordée en synthèse bien qu'elle nous paraisse essentielle, ce n'est d'ailleurs pas un hasard si elle a cette place centrale dans notre thèse.

Nous présenterons ensuite, dans le troisième chapitre, une méthode de résolution globale des transferts énergétiques basée sur les méthodes de Monte-Carlo. Partant d'une formulation nouvelle de ces transferts sous forme de processus aléatoires, nous proposerons des méthodes d'échantillonnage, progressives à leur manière, et adaptées aux différents phénomènes gouvernant le comportement des flux lumineux. Ces méthodes dirigent l'échantillonnage de sorte qu'il soit proportionnel d'une part, à l'énergie reçue et réfléchie en chaque point de la scène et à la contribution reçue par l'oeil d'autre part.

†. Ce qui ne préjuge pas de l'utilisation des images synthétiques comme forme d'expression

Le quatrième et dernier chapitre traitera, comme nous l'avons dit ci-dessus et pour boucler la boucle, d'une nouvelle méthode de tracé de rayons progressif qui utilise, elle aussi, une théorie statistique encore inemployée en synthèse. Le problème du calcul progressif sur l'espace image s'apparente à celui de l'anti-crénelage, bien connu en synthèse. Les méthodes utilisées jusqu'à présent pour sa résolution, polarisent généralement leur attention sur les développements mathématiques de leur solution au détriment de l'aspect psycho-visuel du problème. L'ensemble conduit à des vitesses de convergence très lentes utilisant des critères insuffisamment pertinents, voire inadaptés. Nous proposerons des solutions à convergence rapide par l'utilisation de l'analyse séquentielle et de critères fondés sur le visuel.

I

Les fondements physiques

Un tour d'horizon de différents aspects physiques sur lesquels nous fonderons nos réflexions pour développer, dans le prochain chapitre, un modèle de rendu. Ce petit tour au pays des ondes, indices de réfraction, lumens et autres watts, ne prétend pas à l'exhaustivité, loin s'en faut. En d'autres termes, nous n'allons pas inventer ce que les physiciens ne savent pas, mais nous essayerons de poser les bases minimum raisonnables à la modélisation d'une réalité incomplète mais suffisamment générale pour décrire la plupart des phénomènes lumineux dont nous sommes les témoins quotidiens.

I.1 Electromagnétisme

I.1.1 Les équations de Maxwell

Les équations de Maxwell décrivent les phénomènes électromagnétiques de la façon suivante [Four 79] :

$$\mathrm{rot}\,\mathbf{E} = \frac{\partial \mathbf{B}}{\partial t} \qquad \mathrm{div}\mathbf{B} = 0$$

$$\mathrm{div}\mathbf{D} = \varsigma \qquad \mathrm{rot}\,\mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t}$$

E est le champ électrique
D est l'induction électrique
B est l'induction magnétique
H est le champ magnétique
J est la densité macroscopique de courant
ς est la densité macroscopique de charge électrique

Ces équations générales s'appliquent à tous les milieux. Pour tenir compte des propriétés de la matière, trois autres équations sont nécessaires :

- l'une pour les caractéristiques diélectriques (ou encore isolantes) : $\mathbf{D} = \varepsilon\mathbf{E}$
- la seconde pour les caractéristiques magnétiques : $\mathbf{B} = \mu\mathbf{H}$
- et la dernière, approximation vérifiée pour tous les cas pratiques que nous aurons à traiter : $\mathbf{J} = \gamma\mathbf{E}$. Cette relation est plus connue comme étant la *loi d'Ohm*. Elle caractérise des milieux dits linéaires ou ohmiques qui sont isotropes, isothermes et uniformes. Des milieux non-linéaires sont, par exemple, les semi-conducteurs. Dans la suite nous ne nous intéresserons qu'à des milieux isotropes et homogènes.

où ε désigne la *permittivité* (F/m), $\varepsilon = \varepsilon_o(1 + \chi_e)$
χ_e est la *sensibilité* électrique (sans dimension)
μ représente la *perméabilité* magnétique (H/m), $\mu = \mu_o(1 + \chi_m)$
χ_m est la *susceptibilité* magnétique (sans dimension)
γ désigne la *conductivité* (S/m)
$\varepsilon_o = 1/(\mu_o c_o^2) \approx 8854\times10^{-15}$ et $\mu_o \approx 1257\times10^{-9}$, tous deux relatifs au vide, tout comme $c_o \approx 2998\times10^{5}$ qui représente la vitesse de propagation d'une onde (m/s).

A partir des expressions précédentes, on peut définir **S** le vecteur de Poynting comme étant :

$$\boxed{\mathbf{S} = \mathbf{E} \wedge \mathbf{H}}$$

Ce vecteur revêt une importance toute particulière car son flux représente l'énergie d'une onde électromagnétique. Elle nous intéresse au premier chef puisque c'est elle qui est à l'origine de l'excitation des cellules de l'oeil et donc de la vision comme nous le verrons plus tard.

I.1.2 Les ondes planes

Nous utiliserons exclusivement des ondes planes dans toute cette thèse. On peut démontrer que ce type d'onde caractérise valablement la propagation des phénomènes électromagnétiques (le lecteur intéressé pourra consulter l'excellent ouvrage de G. Fournet [Four 79]), mais ce n'est pas la seule forme capable de le faire. Par exemple, des ondes sphériques sont également solutions des équations de Maxwell. Cette limitation aux ondes planes comme modélisation du monde physique qui nous entoure nous semble cependant suffisante pour les phénomènes que nous traiterons sans que nous en ayons trouvé de preuve formelle.

Une grandeur G se propageant par onde plane est caractérisée par le fait qu'elle possède, à un instant donné, le même état sur tous les points d'un plan perpendiculaire à une direction fixe. Si de plus elle est sinusoïdale ou encore harmonique (l'adjectif sera désormais sous-entendu), elle a la forme générale suivante :

$$G = Ae^{i(\omega(\mathbf{q})t - \mathbf{q}\cdot\overrightarrow{OQ})}$$

où A est une constante dépendant des conditions initiales ou aux limites, Q est un point quelconque dans un repère (O, x, y, z), t désigne le temps et **q** est la direction de propagation de l'onde

Dans le cas d'un milieu isotrope isolant ($\mathbf{J} = 0$, i.e $\gamma = 0$), le champ électrique d'une onde plane s'exprime par :

$$\mathbf{E} = \mathrm{Re}(\mathbf{E}_o e^{i(\omega t - \mathbf{q}\cdot\overrightarrow{OQ})})$$

> ω est la fréquence angulaire ou *pulsation* (rad/s), $\omega = 2\pi f = 2\pi/T$
> $\mathbf{E}_o$ est le vecteur représentant la magnitude (V/m) et la direction maximum de $\mathbf{E}$
> f est la *fréquence* (Hz), T la *période* (s)
> $\|\mathbf{q}\| = q = \omega\sqrt{\varepsilon\mu}$

Sans perte de généralité on peut choisir d'orienter le système d'axes du repère pour que, par exemple, Ox soit colinéaire à $\mathbf{q}$, l'expression précédente s'écrit alors :

$$\boxed{\mathbf{E} = \mathrm{Re}(\mathbf{E}_o e^{i\omega t} e^{iqx})}$$

On peut également définir la *longueur d'onde* (m) $\lambda = 2\pi/q = c/f$, c étant la vitesse de propagation dans le milieu, $c = \varepsilon\mu^{-1/2}$ (m/s). Dans ces conditions, le champ magnétique est en phase avec le champ électrique et peut s'exprimer sous la forme :

$$\mathbf{H} = \mathrm{Re}\left(\frac{1}{\omega\mu}\mathbf{q}\wedge\mathbf{E}\right) \qquad (1)$$

Si le milieu est résistant, autrement dit $\mathbf{J} = \gamma\mathbf{E}$, il faut alors introduire un vecteur de propagation complexe tel que :

$$q = \alpha - i\beta \qquad \text{et} \qquad q^2 = \omega^2\varepsilon\mu - i\omega\mu\gamma$$

ce qui entraîne :

$$\alpha^2 - \beta^2 = \omega^2\varepsilon\mu \qquad \text{et} \qquad 2\alpha\beta = \omega\mu\gamma$$

α et β sont réels ce qui conduit à :

$$\alpha^2 = \frac{\omega^2\varepsilon\mu}{2}\left(\sqrt{1+\gamma^2(\omega\varepsilon)^{-2}}+1\right)$$

$$\beta^2 = \frac{\omega^2\varepsilon\mu}{2}\left(\sqrt{1+\gamma^2(\omega\varepsilon)^{-2}}-1\right)$$

la solution retenue, en choisissant α et β positifs, s'écrit alors sous la forme :

$$\mathbf{E} = \mathrm{Re}(\mathbf{E}_o e^{-\beta x} e^{i(\omega t - \alpha x)})$$

le champ magnétique n'est alors plus en phase et s'exprime par :

$$\mathbf{H} = \mathrm{Re}\left(\frac{\alpha - i\beta}{\omega\mu}\mathbf{q}_u\wedge\mathbf{E}\right) \text{ où } \mathbf{q}_u \text{ est unitaire.}$$

Bien entendu lorsque $\gamma = 0$, β est nul et l'on retrouve (1). Dans le cas d'un milieu isolant le vecteur de Poynting est donné par :

$$\mathbf{S} = \mathbf{E}\wedge\mathbf{H} = \left(\mathbf{E}_o\wedge\left(\frac{\mathbf{q}\wedge\mathbf{E}_o}{\omega\mu}\right)\right)\cos^2(\omega t - \mathbf{q}\cdot\overrightarrow{OQ})$$

$$= \frac{\mathbf{q}}{q}\sqrt{\frac{\varepsilon}{\mu}}\mathbf{E}_o^2\cos^2(\omega t - \mathbf{q}\cdot\overrightarrow{OQ})$$

C'est sa valeur moyenne dans le temps qui nous intéressera :

$$\bar{\mathbf{S}} = \overline{\mathrm{Re}(\mathbf{E}) \wedge \mathrm{Re}(\mathbf{H})} = \mathrm{Re}(\mathbf{E} \wedge \mathbf{H}^*)/2$$

ce qui donne, pour un isolant

$$\bar{\mathbf{S}} = \frac{1}{2}\sqrt{\frac{\varepsilon}{\mu}}\mathbf{E}_o^2 \tag{2a}$$

et pour un conducteur :

$$\bar{\mathbf{S}} = \frac{\alpha}{2\omega\mu}\mathbf{E}_o^2 e^{-2\beta x} \tag{2b}$$

Dans ce dernier cas, on note que la puissance transmise diminue au fur et à mesure de la propagation de l'onde. Ce phénomène est dû à la déperdition d'énergie par effet Joule. Pour caractériser cette diminution, la notion d'*épaisseur de peau* δ est introduite : $\delta^2 = \omega\mu\gamma/2$. Lorsque $\gamma \gg \omega\varepsilon$, on a alors $\alpha = \beta = 1/\delta$ et on peut interpréter δ comme la distance pour laquelle la valeur maximum du champ est divisée par e, soit environ 37%. Le tableau ci-dessous donne quelques exemples de valeurs de δ pour différentes matières.

eau	32
verre	29
graphite	6.10^{-6}
or	15.10^{-5}

Tableau 1: épaisseur de peau, δ, en cm pour $l = 589.3\ nm$

La puissance électro-magnétique P, transmise dans un volume V délimité par une surface Σ, s'exprime donc par le flux de **S** à travers cette surface de normale unitaire $\mathbf{n}_u$:

$$P = \oint_\Sigma \mathbf{S} \cdot \mathbf{n}_u \, d\sigma$$

La conservation de l'énergie à travers ce volume a été établie sous forme de théorème par Poynting en 1884 et s'exprime par [Stra 61] :

$$\oint_\Sigma \mathbf{S} \cdot \mathbf{n}_u \, d\sigma = -\int_V \left(\mathbf{E} \cdot \frac{\partial \mathbf{D}}{\partial t} + \mathbf{H} \cdot \frac{\partial \mathbf{B}}{\partial t} \right) dv - \int_V \mathbf{E} \cdot \mathbf{J} \, dv$$

Autrement dit, le flux sortant de Σ est égal à la diminution des énergies électrique et magnétique à l'intérieur de V moins l'énergie perdue à l'intérieur du volume par transformation en une autre forme d'énergie (effet Joule notamment).

C'est du *théorème de Poynting* que le sens physique de **S** a été déduit. On notera cependant que si la validité de ce théorème est tout à fait générale, l'interprétation de **S** comme étant la densité de puissance en chaque point n'est pas sans ambiguïté. En effet, on peut très bien ajouter un vecteur arbitraire **S'** à **S**, pour peu que l'intégrale de **S'** sur Σ soit nulle, ce qui ne change pas l'égalité précédente. Rien n'empêche ensuite de choisir **S** + **S'** comme densité de puissance : la distribution de cette densité n'est donc pas établie. Néanmoins jusqu'à présent, aucune expérience ne contredit ces interprétations.

I.1.3 Réflexion - réfraction.

Lorsqu'une onde plane passe d'un milieu i à un milieu t séparés par une interface plane, il est possible de prouver que deux nouvelles ondes se forment. L'une réfléchie dans le milieu i, ($\mathbf{E}_r$, $\mathbf{H}_r$), l'autre transmise dans le milieu t, ($\mathbf{E}_t$, $\mathbf{H}_t$). La résolution des équations de Maxwell établit que ces deux ondes ont la même pulsation et que l'onde réfléchie est également plane. En revanche, les phénomènes régissant l'onde transmise sont plus complexes et cette dernière n'est pas toujours plane.

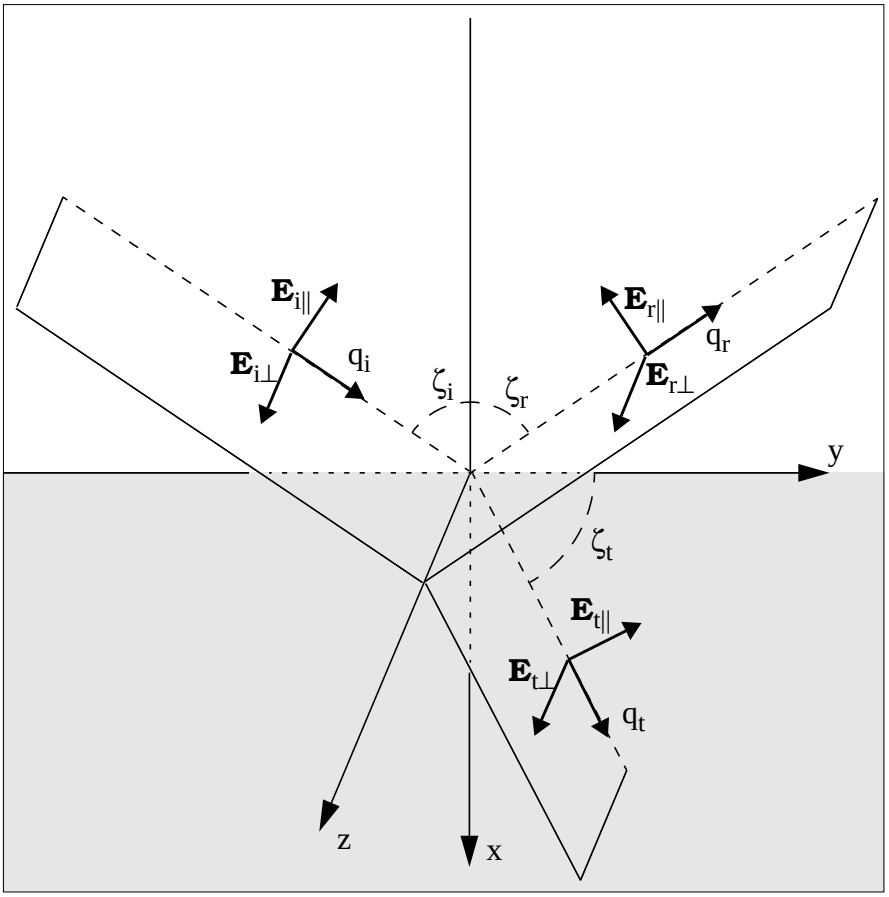


Figure 1: Réflexion et réfraction d'une onde sur une surface plane

Dans le repère de la figure ci-dessus, l'onde incidente est décrite par :

$$\boxed{\mathbf{E}_i = \mathrm{Re}(\mathbf{E}_{oi} e^{-i(x q_i \cos\zeta_i + y q_i \sin\zeta_i - \omega t)})}$$

où $\mathbf{q}_i$ s'exprime par $\mathbf{q}_i = \begin{bmatrix} q_i \cos\zeta_i & q_i \sin\zeta_i & 0 \end{bmatrix}$ avec $q_i = \omega\sqrt{\varepsilon_i \mu_i}$. On peut démontrer [Four 79] que l'onde réfléchie a alors la forme suivante :

$$\boxed{\begin{aligned} \mathbf{E}_r &= \mathrm{Re}\left(\mathbf{E}_{or} e^{-i(\mathbf{q}_r \cdot \overrightarrow{OQ} - \omega t)}\right) \\ \mathbf{q}_r &= \begin{bmatrix} -q_i \cos\zeta_i & q_i \sin\zeta_i & 0 \end{bmatrix} \end{aligned}}$$

L'onde transmise peut être de deux formes différentes selon l'angle d'incidence :

- si $q_t > q_i \sin\zeta_i$ on a

$$\begin{cases} \mathbf{E}_t = \mathrm{Re}\left(\mathbf{E}_{ot} e^{-i(\mathbf{q}_t \cdot \overrightarrow{OQ} + \omega t)}\right) \\ \mathbf{q}_t = \left[q_t \sin\zeta_t \quad q_t \sin\zeta_t \quad 0\right] \qquad q_t = \omega\sqrt{\varepsilon_t \mu_t} \end{cases}$$

où l'électromagnétisme retrouve la relation de Snell-Descartes $q_i \sin\zeta_i = q_t \sin\zeta_t$ plus connue sous la forme $n_i \sin\zeta_i = n_t \sin\zeta_t$ où $n_i = q_i / q_o = \sqrt{\varepsilon_i \mu_i / \varepsilon_o \mu_o}$, et $n_t = \sqrt{\varepsilon_t \mu_t / \varepsilon_o \mu_o}$, désignent les *indices de réfraction* de chacun des milieux. On peut remarquer que cette onde est également plane.

- en revanche si $q_t < q_i \sin\zeta_i$ (ou encore $n_t < n_i \sin\zeta_i$, ce qui ne peut se produire que si $n_t < n_i$) le vecteur de propagation doit s'écrire sous la forme :

$$\boxed{\mathbf{q}_t = \left[-i\sqrt{q_i^2 \sin\zeta_i - q_t^2} \quad q_i \sin\zeta_i \quad 0\right] \qquad q_t = \omega\sqrt{\varepsilon_t \mu_t}}$$

Il n'est alors pas possible de parler d'onde plane.

Outre l'aspect géométrique qui permet de déterminer la direction des ondes produites, le deuxième aspect primordial est l'échange énergétique qui en résulte. Il faut donc s'intéresser aux différents vecteurs de Poynting associés : $\mathbf{S}_i$, $\mathbf{S}_r$, $\mathbf{S}_t$. En $x = 0$ (c'est-à-dire sur la surface de séparation des deux milieux) et en régime stable, seule la composante x de ces vecteurs est non nulle et telle que $(S_x = E_y H_z + E_z H_y)_{x=0}$. On peut alors définir un coefficient énergétique de transmission $T = S_{tx} / S_{ix}$, la *transmittance*, dont seule $\bar{T}$, la moyenne dans le temps, nous intéressera compte tenu des temps de réaction très lents de l'oeil comparés à la rapidité des phénomènes vibratoires mis en jeu. On a alors

$$\bar{T} = T_{\parallel} \sin^2\alpha_i + T_{\perp} \cos^2\alpha_i \qquad (3)$$

où $\alpha_i = \angle(\mathbf{z}, \mathbf{E}_{oi})$

$$T_{\parallel} = \frac{4\mathrm{Re}(s_{\parallel})}{|1 + s_{\parallel}|^2} \qquad T_{\perp} = \frac{4\mathrm{Re}(s_{\perp})}{|1 + s_{\perp}|^2}$$

et $s_{\perp} = \dfrac{\mu_i q_{tx}}{\mu_t q_i \cos\zeta_i} \qquad s_{\parallel} = \dfrac{\mu_i q_t^2 \cos\zeta_i}{\mu_t q_{tx} q_i}$

Les indices $\perp$ et $\parallel$ se réfèrent au champ électrique

Lorsque $\mathbf{E}_i$ est parallèle au plan d'incidence (plan contenant $\mathbf{q}_i$ et la normale à la surface de séparation des deux milieux) on a $\bar{T} = T_{\parallel}$, par contre lorsqu'il est perpendiculaire à ce plan, $\bar{T} = T_{\perp}$. Si $\mathbf{E}_i$ est linéairement polarisé, c'est-à-dire si α_i est constant sur un

intervalle de temps, la formule (3) est à prendre telle quelle, par contre lorsque l'onde ne montre pas de polarisation particulière, α_i varie de façon aléatoire et il faut intégrer, ce qui donne :

$$\bar{T} = \frac{T_{\parallel} + T_{\perp}}{2}$$

La conservation de l'énergie entraîne que la *réflectance* $\bar{R}$, rapport du flux d'énergie réfléchi sur le flux incident, est égale à 1 - $\bar{T}$. Après un certain nombre de manipulations trigonométriques on obtient finalement :

$$\begin{aligned} T_{\parallel} &= \frac{\sin(2\zeta_i)\sin(2\zeta_t)}{\sin^2(\zeta_i+\zeta_t)\cos^2(\zeta_i-\zeta_t)} & T_{\perp} &= \frac{\sin(2\zeta_i)\sin(2\zeta_t)}{\sin^2(\zeta_i+\zeta_t)} \\ R_{\parallel} &= \frac{\tan^2(\zeta_i-\zeta_t)}{\tan^2(\zeta_i+\zeta_t)} & R_{\perp} &= \frac{\sin^2(\zeta_i-\zeta_t)}{\sin^2(\zeta_i+\zeta_t)} \end{aligned} \qquad (4)$$

où l'on reconnaît les formules dérivées par Fresnel en 1823. Si sa théorie de la lumière élastique était fausse, en revanche la multitude d'expériences qu'il a menées d'arrache-pied lui ont permis de dériver quantité de résultats qui restent tout à fait valides aujourd'hui encore.

Dans le cas où le milieu t est conducteur, on peut établir la réflectance et la transmittance en fonction des indices des deux milieux et obtenir [Shul 53] :

$$R_{\perp} = \frac{a^2+b^2-2a\cos\zeta_i+\cos^2\zeta_i}{a^2+b^2+2a\cos\zeta_i+\cos^2\zeta_i} \qquad R_{\parallel} = R_{\perp}\frac{a^2+b^2-2a\sin\zeta_i\tan\zeta_i+\sin^2\zeta_i\tan^2\zeta_i}{a^2+b^2+2a\sin\zeta_i\tan\zeta_i+\sin^2\zeta_i\tan^2\zeta_i}$$

$$\begin{aligned} a^2 &= \frac{\sqrt{(n_t^2-k_t^2-n_i^2\sin^2\zeta_i)^2+4n_t^2k_t^2}+n_t^2-k_t^2-n_i^2\sin^2\zeta_i}{2n_i^2} \\ b^2 &= \frac{\sqrt{(n_t^2-k_t^2-n_i^2\sin^2\zeta_i)^2+4n_t^2k_t^2}-n_t^2+k_t^2+n_i^2\sin^2\zeta_i}{2n_i^2} \end{aligned} \qquad (5)$$

l'indice de réfraction $\hat{n}$ n'étant plus réel mais complexe, $\boxed{\hat{n} = n - ik}$ où l'on a [Humm 93]

$$\boxed{\begin{aligned} n^2 &= (\sqrt{\varepsilon^2+(2\gamma/f)^2}+\varepsilon)/2 \\ k^2 &= (\sqrt{\varepsilon^2+(2\gamma/f)^2}-\varepsilon)/2 \end{aligned}}$$

k est l'*indice d'amortissement*. Ce terme peu usité a été choisi pour éviter toute confusion. Cet indice est en effet souvent appelé coefficient d'extinction ou indice d'atténuation, malheureusement certains auteurs définissent $\hat{n}$ par $\hat{n} = n(1 + i\kappa)$ où κ porte les mêmes dénominations...

Au passage on notera que la transmittance ne décrit les phénomènes qu'à l'interface des milieux en présence, ce qui, dans le cas des conducteurs, n'indique pas quelle énergie ressortira finalement du milieu t. Pour ce faire, il faut considérer l'énergie à une certaine distance de la surface comme nous l'avons fait pour l'effet de peau.

Les formules (5) couvrent bien sûr le cas du milieu t isolant, c'est-à-dire lorsque k = 0. Le cas particulier de l'incidence normale (ζ_i=0) est intéressant car souvent utilisé par les physiciens pour la détermination des «constantes» optiques :

$$R_{\parallel} = R_{\perp} = \frac{k_t^2 + (n_t - n_i)^2}{k_t^2 + (n_t + n_i)^2} \text{ et donc } \bar{R} = (R_{\perp} + R_{\parallel})/2 = R_{\perp} = \bar{R}_o,\ \bar{T} = 1 - \bar{R}$$

dans le cas diélectrique, k = 0, et on obtient $\bar{R} = (n_t - n_i)^2 (n_t + n_i)^{-2}$.

Il faut noter que ces relations ne sont valables que lorsque les matériaux sont parfaitement lisses et qu'elles sont spectrales. Les quantités ε, γ, μ et par conséquent n, k, R et T dépendent donc de la pulsation de l'onde incidente. Le terme de constantes optiques est ainsi assez impropre et ne doit pas faire oublier cette dépendance.

Une autre interdépendance, qui ne peut apparaître dans les formules ci-dessus, concerne les inter-relations entre les parties réelle et imaginaire de l'indice de réfraction complexe sur l'ensemble du domaine fréquenciel. Cette interdépendance s'exprime par les *relations de dispersion* de Kramers-Kronig [Hoc 85] :

$$\alpha(\omega_j) = \frac{2\omega_j}{\pi} \int_0^{\infty} (\ln\sqrt{\bar{R}_o(\omega)} - \ln\sqrt{\bar{R}_o(\omega_j)})/(\omega_j^2 - \omega^2)\, d\omega$$

$$n(\omega) = \frac{1 - \bar{R}_o(\omega)}{1 - 2\sqrt{\bar{R}_o(\omega)}\cos\alpha(\omega) + \bar{R}_o(\omega)}$$

$$k(\omega) = \frac{2\bar{R}_o(\omega)\sin\alpha(\omega)}{1 - 2\sqrt{\bar{R}_o(\omega)}\cos\alpha(\omega) + \bar{R}_o(\omega)}$$

où α représente la différence de phase entre l'onde incidente et l'onde réfléchie.

Ces relations sont souvent utilisées en physique pour la détermination de n et k, malheureusement leur application exigent une connaissance de $\bar{R}_o(\omega)$ sur des spectres très étendus (du lointain infra-rouge jusqu'aux rayons gamma) avec une très bonne précision puisqu'une erreur sur une fréquence quelconque se répercute sur l'ensemble de la région spectrale; d'autre part, les données sur de telles étendues ne sont pas très répandues pour les matériaux qui nous intéressent.

Il faut noter également que ces quantités dépendent de la température, ce qui n'apparaît pas dans les équations ci-dessus. Une discussion sur le sujet nous emmènerait trop loin. De plus, nous ne nous placerons qu'à température constante, c'est pourquoi nous n'en avons pas parlé.

L'élément indispensable est donc l'indice de réfraction complexe, ce en quoi nous rejoignons la thèse de Callet [Calle 93]. Un certain nombre de tables répertorient les indices de différents matériaux pour différentes longueurs d'ondes. De bonnes tables ([Hoc 85] par exmple) devraient indiquer tous les paramètres susceptibles de modifier les mesures :

- spécification du matériau : préparation, forme (solide, gazeux, film, ...), état de surface, composition.
- méthode de mesure et méthode d'ajustement.
- conditions d'expérience : température, degré d'humidité.
- précision des nombres fournis.

I.2 Rugosité

I.2.1 Introduction

Nous n'avons traité jusqu'à présent que de surfaces parfaitement planes mais la nature est autre comme nous le montrent ces quelques exemples de surfaces photographiées avec un microscope de type Nomarski [Benn et al 89].

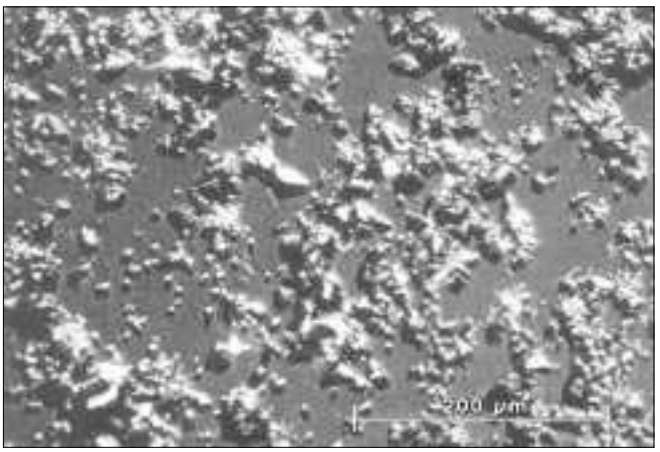


Figure 2: Verre poli

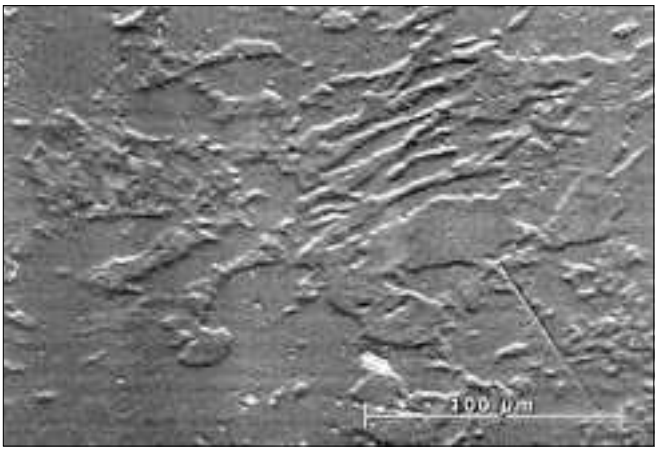


Figure 3: Molybdène poli

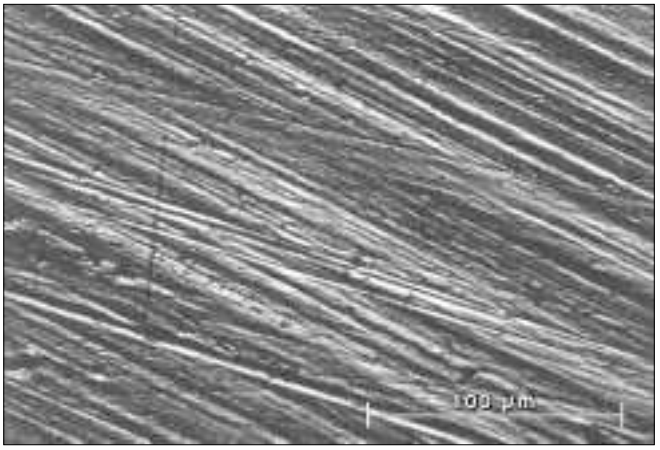


Figure 4: Feuille d'aluminium où les marques du laminoir sont clairement visibles

D'autre part la nature ondulatoire de la lumière produit des effets inexplicables sur des surfaces planes. Considérons une «surface» mono-dimensionnelle composée d'aspérités régulières de hauteur h sur laquelle un rayon de longueur d'onde λ est incident avec un angle ζ.

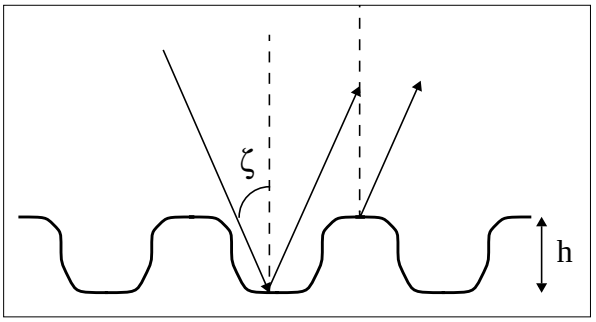


Figure 5: Interférences sur surface rugueuse

La différence de marche maximale entre 2 rayons réfléchis s'exprime par $\Delta d = (2h)/(\cos\zeta)$ tandis que la différence de phase maximale est alors $\Delta\varphi = 2\pi\Delta d/\lambda = 4\pi h/(\lambda\cos\zeta)$. Lorsque cette différence est faible, les deux rayons sont pratiquement en phase et l'interférence est donc constructive. Elle l'est de moins en moins au fur et à mesure que $\Delta\varphi$ grandit jusqu'à l'annulation du rayonnement réfléchi par interférences destructives lorsque $\Delta\varphi = \pi$. La conservation de l'énergie aidant, l'énergie portée par ces rayonnements doit se retrouver dans d'autres directions et donc conduire à un rayonnement diffus. Afin de caractériser les états de surface selon le type de réflexion qu'ils provoquent, Lord Rayleigh a proposé un critère, un peu artificiel, qui depuis porte son nom. Ainsi il retint la valeur médiane de $\Delta\varphi = \pi/2$ pour séparer les deux états. Une surface serait donc lisse si $h < \lambda\cos\zeta / 8$ et rugueuse dans le cas contraire. Comme on le voit, la notion de rugosité ne peut être que relative sauf à considérer des surfaces fractales, ce qu'explore la théorie actuelle (voir par exemple Jakeman [Jake 82]).

Si le partage entre rugueux et lisse par le critère de lord Rayleigh est un peu trop abrupt, on peut néanmoins en retenir que deux conditions peuvent faire considérer une surface comme lisse : soit $h/\lambda \to 0$, soit $\zeta \to \pi/2$. La seconde condition peut paraître étonnante puisqu'elle stipule que, quel que soit l'état d'une surface, celle-ci est réfléchissante pour des angles d'incidence rasante, mais ceci est prévu par l'expression des coefficients de Fresnel (voir (4) §I.1.3 lorsque $\zeta_i = \pi/2$) et l'on peut d'ailleurs en faire l'expérience avec une simple feuille de papier tenue bien à plat devant les yeux et dirigée

vers une source de lumière.

Plus formellement, un des paramètres les plus utilisés pour décrire une surface est l'*écart-type de la rugosité* (root-mean-square (rms) roughness en anglais) souvent abrégé en *rugosité* et notée σ ou σ_m. Si on note z_i les variations de hauteurs d'une surface décrite par $z = h(x,y)$ par rapport à sa surface moyenne $\bar{h} = 0$, sa rugosité s'écrit :

$$\sigma = \left(\frac{1}{M} \sum_{i=1}^{M} z_i^2 \right)^{-1/2}$$

où les z_i sont pris à intervalles réguliers

Cette description statistique des surfaces utilisée par les physiciens, sauf longue habitude, n'est pas facilement compréhensible par le néophyte. Nous donnons donc ci-dessous quelques exemples de distributions de hauteurs représentatives de divers profils. Dans ces représentations, des courbes gaussiennes «équivalentes» ont été calculées pour avoir la même aire et le même écart-type que les distributions.

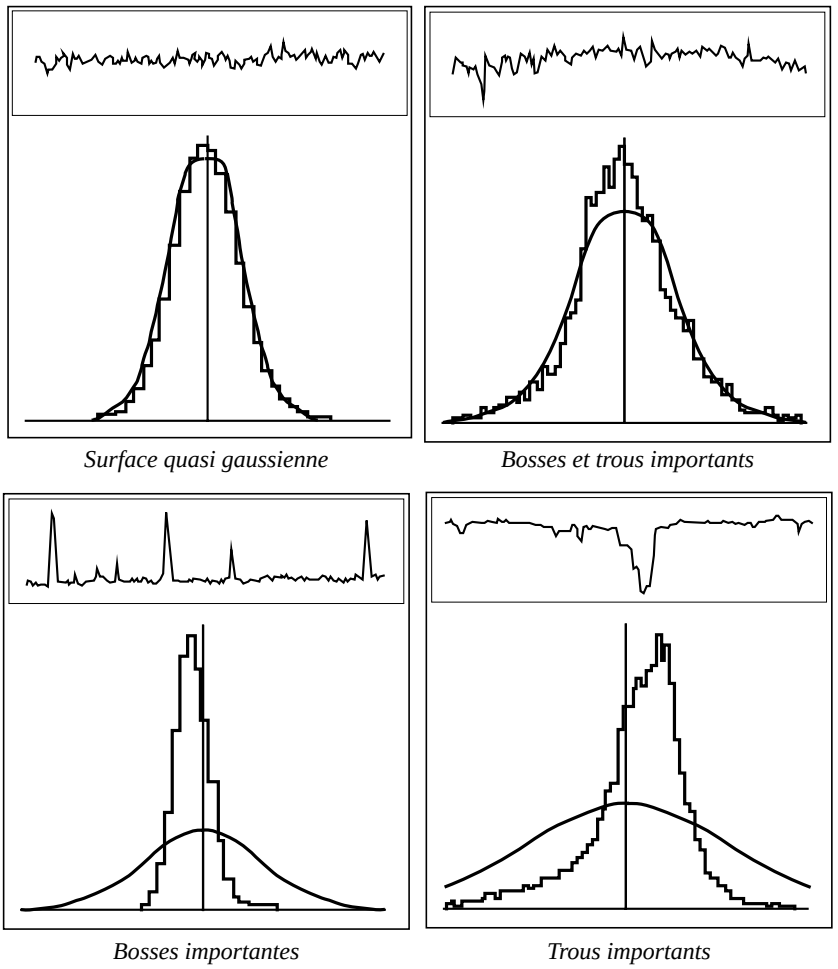


Figure 6: Exemples de profils de surfaces avec les distributions de hauteurs correspondantes

Ci-dessous le profil d'une surface de silicone avec la distribution correspondante ainsi que la distribution des hauteurs d'une surface de cuivre polie au diamant

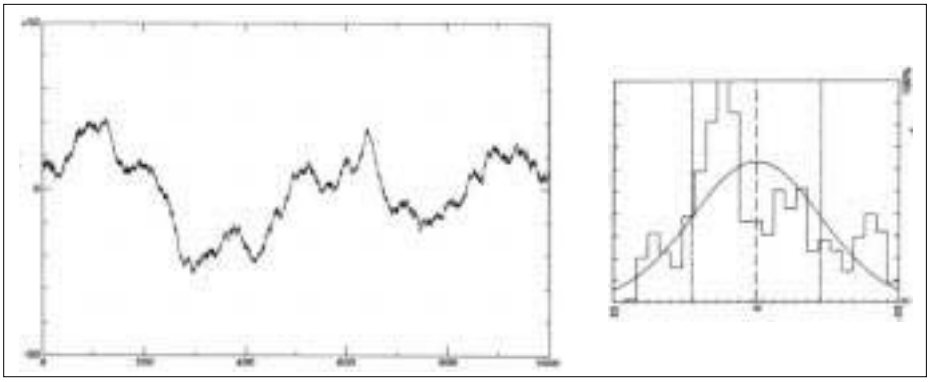


Figure 7: Profil et distribution des hauteurs d'une surface de silicone

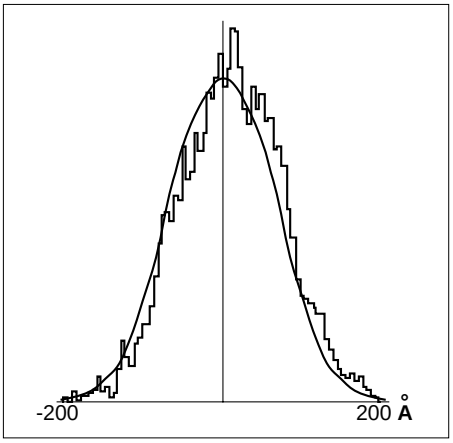


Figure 8: Distribution des hauteurs d'une surface de cuivre polie au diamant

I.2.2 Modèles à facettes

Une des premières idées concernant la modélisation de la réflexion sur surfaces rugueuses a consisté à supposer toute surface comme constituée de facettes planes réfléchissant spéculairement la lumière. Cette idée est en fait très ancienne puisque Bouguer fut sans doute le premier à la proposer en … 1760. Ce modèle a ensuite été affiné notamment par Torrance et Sparrow [Torr et al 67]. Trois hypothèses de base président à l'élaboration de leur modèle :

- Seules les lois de l'optique géométrique interviennent dans le phénomène de réflexion. Pour cela, il faut que la rugosité σ soit plus grande que la longueur d'onde incidente ($\sigma \gg \lambda$)
- la réflexion provient de deux phénomènes additifs, l'un diffus, l'autre spéculaire.
- Les réflexions multiples à l'intérieur des cavités formées par les facettes ne sont pas prises en compte (l'hypothèse est qu'elles se retrouvent dans la composante diffuse)

La seconde hypothèse conduit à ce qu'une source de petite taille réfléchie par un élément de surface produit une luminance réfléchie L_r qui peut s'exprimer par une somme de deux composantes :

$$L_r(\zeta_i, \zeta_r, \alpha_r) = L_{rs}(\zeta_i, \zeta_r, \alpha_r) + L_{rd}(\zeta_i)$$

La composante diffuse se comporte selon la loi de Lambert, autrement dit, pour une surface unitaire éclairée par une source de luminance isotrope L_i on a :

$$L_{rd}(\zeta_i) = m_1 L_i \cos\zeta_i$$

m_1 étant une constante dépendant du matériau

Si la répartition des hauteurs de la surface suit une gaussienne invariante par rotation autour de la normale moyenne **N**, sa distribution est donnée par :

$$p(\zeta_{\mathbf{n}}) = m_2 e^{-m_3^2 \zeta_{\mathbf{n}}^2}$$

m_2 et m_3 dépendent du matériau
$\zeta_{\mathbf{n}}$ est l'angle d'inclinaison local de **n** , la normale d'une facette, avec **N**

Le nombre de facettes inclinées de telle sorte que leur normale appartienne à un angle solide $d\omega_{\mathbf{n}}$ sur un élément de surface dA s'exprime par :

$$p(\zeta_{\mathbf{n}}) d\omega_{\mathbf{n}} d\mathrm{A}$$

dont l'aire vue par la source, si chaque facette à une aire a, est :

$$p(\zeta_{\mathbf{n}}) d\omega_{\mathbf{n}} d\mathrm{A}\, a \cos\psi_{\mathbf{n}}$$

$\psi_{\mathbf{n}}$ est l'angle d'incidence local avec les facettes bien orientées, c'est-à-dire le demi-angle entre la direction incidente et la direction d'observation

Le flux d'énergie qui est donc reçu par ces facettes bien orientées est alors :

$$d\phi_i = d\omega_i L_i p(\zeta_{\mathbf{n}}) d\omega_{\mathbf{n}} d\mathrm{A}\, a \cos\psi_{\mathbf{n}}$$

$d\omega_i$ est l'angle solide sous-tendu par la source lorsqu'elle est vue depuis un point de dA

Le flux qu'elles réfléchissent spéculairement s'exprimerait par :

$$d\phi_r = R(\psi_{\mathbf{n}}) d\phi_i$$

si le rayonnement incident arrivait entièrement aux facettes bien orientées. Ce n'est pas le cas puisque le rayonnement peut être masqué. Toutefois ce masquage est relativement simple car le modèle de facettes retenu par Torrance et Sparrow est très contraignant†, ses caractéristiques sont les suivantes :

- Chaque facette appartient à une cavité symétrique.
- L'axe longitudinal des cavités est parallèle au plan de la surface moyenne.
- Les orientations des cavités sont équiprobables.
- Les bords supérieurs des facettes sont tous dans le même plan.

† Il n'est d'ailleurs pas prouvé que les conditions posées par Torrance et Sparrow soient compatibles avec l'hypothèse gaussienne. Sans en apporter de preuves, nous pensons même que ces deux hypothèses sont exclusives l'une de l'autre.

A partir de ces hypothèses, Torrance et Sparrow expriment l'illumination† partielle des facettes par un facteur géométrique G :

$$G(\zeta_{ip}, \zeta_{rp}) = 1 - l_m / l_f$$

ζ_{ip} est l'angle que fait la direction incidente projetée sur le plan (**n** , **N**) avec la normale moyenne, de même ζ_{rp} est l'angle que la direction d'observation projetée sur le même plan fait avec **N**
l_f représente la longueur d'une facette dans le plan (**n** , **N**) et l_m la longueur ombrée d'une facette.

Dans ces conditions le flux effectivement réfléchi s'exprime par :

$$d\phi_r = G(\zeta_{ip}, \zeta_{rp}) R(\psi_{\mathbf{n}}) d\omega_i L_i p(\zeta_{\mathbf{n}}) d\omega_{\mathbf{n}} dA a \cos\psi_{\mathbf{n}}$$

flux aussi égal à :

$$d\phi_r = L_{rs}(\zeta_i, \zeta_r, \alpha_r) d\omega_r dA \cos\zeta_r$$

$d\omega_r$ est l'angle solide sous-tendu par un récepteur dans la direction d'observation lorsqu'il est vu depuis un point de *d*A

En égalisant ces deux dernières expressions et sachant que

$$d\omega_{\mathbf{n}} = \frac{d\omega_r}{4\cos\psi_{\mathbf{n}}}$$

il vient finalement

$$L_{rs}(\zeta_i, \zeta_r, \alpha_r) = \frac{G(\zeta_{ip}, \zeta_{rp}) R(\psi_{\mathbf{n}}) d\omega_i L_i p(\zeta_{\mathbf{n}}) a}{4\cos\zeta_r} \qquad (6)$$

Trowbridge et Reitz [Trow et al 75] ont généralisé le modèle à facettes pour en faire un modèle à cuvettes où les surfaces des facettes ne sont plus planes mais courbes. Ce modèle pose un certain nombre d'hypothèses simplificatrices que voici :

- Les réflexions multiples et la diffraction par les arêtes sont négligées.
- Les plus petites courbures des cuvettes ainsi que leurs plus petites dimensions sont plus grandes que la longueur d'onde incidente.
- Chaque cuvette est optiquement lisse.
- La surface a des propriétés statistiques uniformes et indépendantes de la direction sur la surface.
- L'observation se fait à grande distance.
- La densité du flux énergétique incident sur chaque facette est approximée par la densité incidente moyenne sur la surface.
- Le nombre de cuvettes éclairées est approximé par le nombre de celles éclairées en incidence normale auquel est adjoint une dépendance inverse au cosinus de l'angle d'incidence (c'est-à-dire proportionnelle à l'aire éclairée si la surface était plate).

† G est un facteur d'illumination et non de masquage comme souvent présenté puisqu'il représente la portion de facette éclairée, non celle qui est ombrée

La réflectance bidirectionnelle (Cf §I.4.2) à laquelle Trowbridge et Reitz aboutissent avec un modèle de cuvettes faites de morceaux d'ellipsoïdes de révolution répondant aux critères ci-dessus est la suivante :

$$f_r(\zeta_i, \zeta_r) = \frac{R(\psi_{\mathbf{n}})}{4\pi \cos\zeta_i \cos\zeta_r}\left(\frac{e^2}{e^2\cos^2(\zeta_{\mathbf{n}}) + \sin^2(\zeta_{\mathbf{n}})} \right)^2$$

e est l'excentricité moyenne des ellipsoïdes composant la surface

I.2.3 Le modèle de Beckmann

La théorie de la réflexion sur surfaces rugueuses s'est vraiment développée dans les années 50 avec l'étude de Davies [Davi 56] portant sur la dispersion d'onde radar sur l'océan. Etudes étendues ensuite dans le domaine optique par Bennett et Porteus en 1961 [Port et al 61] apparemment en même temps que Beckmann [Beck et al 63]. Ce dernier traite de façon approfondie la réflexion sur des surfaces rugueuses et ses résultats ont été amplement repris depuis.

L'approche de Beckmann et de ses successeurs est très sensiblement différente de celle adoptée pour les modèles à facettes. Alors que Beckmann cherche à décrire le comportement moyen du champ réfléchi par une surface, les tenants des modèles à facettes se placent d'emblée dans le cas géométrique pour s'intéresser au comportement moyen de la surface duquel ils déduisent le champ réfléchi.

I.2.3.1 Expression générale du champ réfléchi

Deux hypothèses préalables sont nécessaires à l'élaboration de ce modèle :

- Les effets d'ombrages et de réflexions multiples sont ignorés.
- La diffraction par les arêtes de la surface est négligée.

Partant de là, $\mathbf{E}_r$, le champ réfléchi en un point P quelconque au-dessus d'une surface Σ, s'exprime par l'*intégrale de Helmholtz* classique :

$$\mathbf{E}_r(P) = \frac{1}{4\pi}\int_{\Sigma}\left(\mathbf{E}_{\Sigma}\frac{\partial\psi}{\partial\mathbf{n}_u} - \psi\left\{\frac{\partial\mathbf{E}}{\partial\mathbf{n}_u}\right\}_{\Sigma}\right) ds \qquad (7)$$

où $\mathbf{E}_s$ désigne le champ électrique sur Σ

$\mathbf{n}_u$ est la normale unitaire locale sur Σ (étant issue du théorème de la divergence, l'intégrale ci-dessus doit utiliser une normale de norme un)

$\psi = e^{iq_r d}/d$, q_r étant la norme de $\mathbf{q}_r$, le vecteur de propagation du champ électrique réfléchi $\mathbf{E}_r$

d est la distance d'observation : distance entre P et un point de Σ

La forme $\{\partial\mathbf{E}/\partial\mathbf{n}_u\}_{\Sigma}$ désigne le gradient normal du champ sur Σ

De façon également classique, la dépendance temporelle est supprimée puisque nous ne nous intéresserons pas aux variations temporelles instantanées. Le champ électrique incident $\mathbf{E}_i$ s'exprime donc par :

$$\mathbf{E}_i = \mathbf{E}_{oi} e^{i \mathbf{q}_i \cdot \mathbf{p}} \quad (8)$$

où $\mathbf{p} = x\mathbf{i}+y\mathbf{j}+z\mathbf{k}$, (x,y,z) étant un point quelconque de l'espace au-dessus de la surface

Les normes des vecteurs de propagation sont égales : $q = |\mathbf{q}_r| = |\mathbf{q}_i| = 2\pi/\lambda$, λ la longueur d'onde des champs. Avec ces conventions et le système d'axes suivant

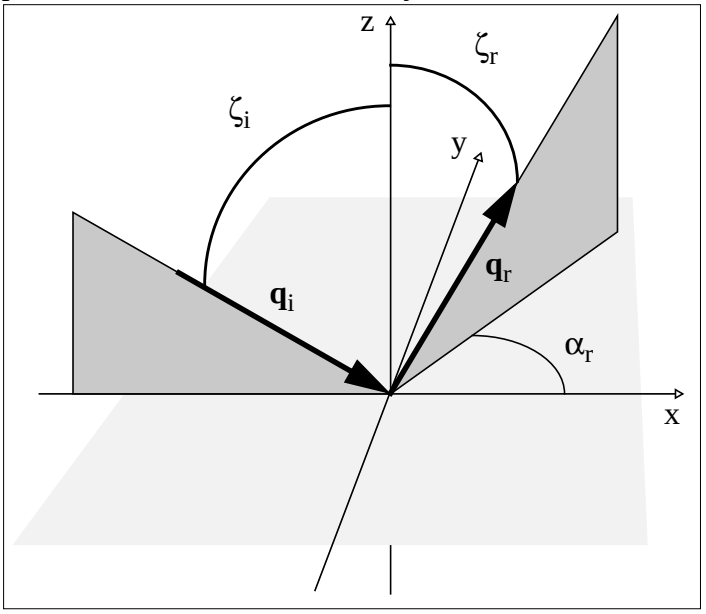


Figure 9: Système d'axes pour l'expression du champ réfléchi

les vecteurs de propagation s'expriment par :

$$\mathbf{q}_i = (\mathbf{i}\sin\zeta_i - \mathbf{k}\cos\zeta_i)q_i$$

$$\mathbf{q}_r = (\mathbf{i}\sin\zeta_r\cos\alpha_r + \mathbf{j}\sin\zeta_r\sin\alpha_r + \mathbf{k}\cos\zeta_r)q_r$$

Afin de ne considérer que des ondes planes, l'observation se fait à grande distance ($d \to +\infty$). Dans ces conditions si d_o est la distance à l'origine, on a

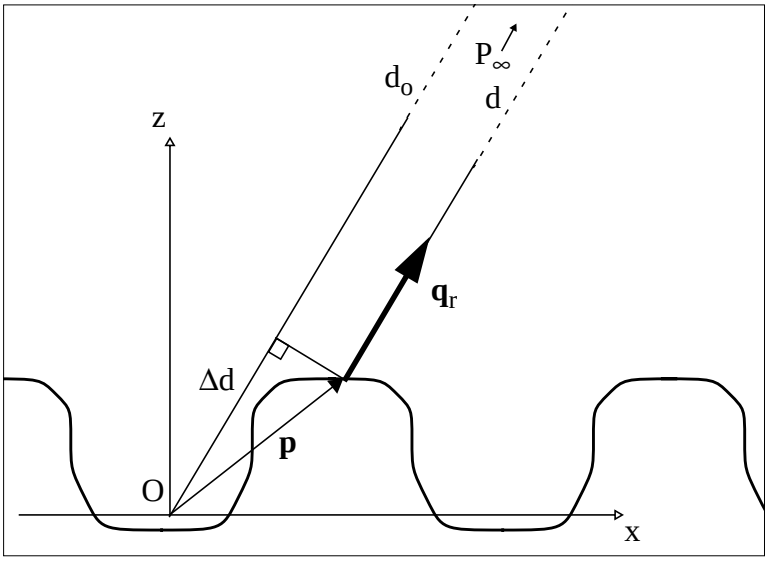


Figure 10: Observation à grande distance du champ réfléchi

$d = d_o - \Delta d$ et par conséquent $q_r d = q_r(d_o - \Delta d) = q_r d_o - \mathbf{q}_r \cdot \mathbf{p}$. D'autre part comme P est très éloigné, d et d_o sont grands par rapport aux dimensions de la surface. Le dénominateur de ψ peut être approximé par d_o et l'on obtient :

$$\psi = e^{i q_r d} / d \approx e^{i q_r d_o - i \mathbf{q}_r \cdot \mathbf{p}} / d_o$$

En supposant que le comportement du champ en un point de Σ est le même que celui qu'il aurait en remplaçant la surface autour de ce point par le plan qui lui est tangent, hypothèse connue sous le nom d'*approximation de Kirchhoff* (et souvent abrégée en KA par les anglo-saxons), on peut écrire :

$$\mathbf{E}_\Sigma = (1 + R_{\zeta_{li}})\mathbf{E}_i$$

$$\left\{\frac{\partial \mathbf{E}}{\partial \mathbf{n}_u}\right\}_\Sigma = i(1 - R_{\zeta_{li}})\mathbf{q}_i \cdot \mathbf{n}_u \mathbf{E}_i \qquad (9)$$

$R_{\zeta li}$ est le coefficient de Fresnel qui dépend donc de l'angle d'incidence local ζ_{li} et de la polarisation de $\mathbf{E}_i$

En incluant ces égalités dans (7) il vient

$$\mathbf{E}_r(P) = \frac{1}{4\pi}\int_\Sigma \left((1 + R_{\zeta_{li}})\mathbf{E}_i \frac{\partial \psi}{\partial \mathbf{n}_u} - i\psi(1 - R_{\zeta_{li}})\mathbf{q}_i \cdot \mathbf{n}_u \mathbf{E}_i \right) ds \qquad (10)$$

Le produit scalaire $\mathbf{q}_r \cdot \mathbf{p}$ s'exprimant par

$$\mathbf{q}_r \cdot \mathbf{p} = q(x \sin\zeta_r \cos\alpha_r + y \cos\zeta_r \sin\alpha_r + z \cos\zeta_r)$$

La dérivée de ψ s'écrit donc

$$\frac{\partial \psi}{\partial \mathbf{n}_u} = \begin{bmatrix} \frac{\partial \psi}{\partial x} \\ \frac{\partial \psi}{\partial y} \\ \frac{\partial \psi}{\partial z} \end{bmatrix} \cdot \mathbf{n}_u \approx \frac{e^{i q_r d_o}}{d_o} \begin{bmatrix} -i e^{-i \mathbf{q}_r \cdot \mathbf{p}} q \sin\zeta_r \cos\alpha_r \\ -i e^{-i \mathbf{q}_r \cdot \mathbf{p}} q \sin\zeta_r \sin\alpha_r \\ -i e^{-i \mathbf{q}_r \cdot \mathbf{p}} q \cos\zeta_r \end{bmatrix} \cdot \mathbf{n}_u = \frac{-i e^{i q_r d_o - i \mathbf{q}_r \cdot \mathbf{p}}}{d_o} \mathbf{q}_r \cdot \mathbf{n}_u = -i\psi \mathbf{q}_r \cdot \mathbf{n}_u$$

en remplaçant dans l'intégrale (10) on obtient

$$\mathbf{E}_r(P) \approx \frac{1}{4\pi}\int_\Sigma \mathbf{E}_i [-i(1 + R_{\zeta_{li}})\psi \mathbf{q}_r \cdot \mathbf{n}_u - i(1 - R_{\zeta_{li}})\psi \mathbf{q}_i \cdot \mathbf{n}_u] ds$$

$$\approx \frac{i}{4\pi}\int_\Sigma \mathbf{E}_i \psi [(1 + R_{\zeta_{li}})\mathbf{q}_r + (1 - R_{\zeta_{li}})\mathbf{q}_i] \cdot -\mathbf{n}_u ds$$

L'exponentielle de ψ en $q_r d_o$ étant constante sur la surface, elle peut sortir, et en incluant (8) il vient :

$$\mathbf{E}_r(P) \approx \frac{i e^{i q_r d_o}}{4\pi d_o} \mathbf{E}_{oi} \int_\Sigma e^{i \mathbf{q}_i \cdot \mathbf{p}} e^{-i \mathbf{q}_r \cdot \mathbf{p}} [(1 + R_{\zeta_{li}})\mathbf{q}_r + (1 - R_{\zeta_{li}})\mathbf{q}_i] \cdot -\mathbf{n}_u ds \qquad (11)$$

En explicitant le premier terme du produit scalaire on a

$$(1+R_{\zeta_{li}})\mathbf{q}_r + (1-R_{\zeta_{li}})\mathbf{q}_i = q\begin{bmatrix}(1+R_{\zeta_{li}})\sin\zeta_r\cos\alpha_r + (1-R_{\zeta_{li}})\sin\zeta_i\\(1+R_{\zeta_{li}})\sin\zeta_r\sin\alpha_r\\(1+R_{\zeta_{li}})\cos\zeta_r - (1-R_{\zeta_{li}})\cos\zeta_i\end{bmatrix} = q\begin{bmatrix}a\\c\\b\end{bmatrix}$$

En supposant de plus, sans perte de généralité, que Σ est décrite par $z = h(x,y)$, la normale locale s'écrit alors :

$$\mathbf{n}_u = \begin{bmatrix}-h'_x & -h'_y & 1\end{bmatrix} n^{-1} \qquad n = \sqrt{h'^2_x + h'^2_y + 1}$$

où h'_x et h'_y désignent les dérivées de h par rapport à x et y respectivement.

Enfin, en posant $\mathbf{v} = \mathbf{q}_i - \mathbf{q}_r = \mathbf{i}v_x + \mathbf{j}v_y + \mathbf{k}v_z$ et en remplaçant ces deux dernières expressions dans (11) on obtient finalement :

$$\mathbf{E}_r(P) \approx \frac{iqe^{iq_r d_o}}{4\pi d_o}\mathbf{E}_{oi}\int_\Sigma (ah'_x + ch'_y - b)e^{i\mathbf{v}\cdot\mathbf{p}}n^{-1}ds \qquad (12)$$

I.2.3.2 Le coefficient de dispersion

Beckmann cherche ensuite à calculer ρ, *coefficient de dispersion* qui est le rapport du champ réfléchi par la surface en P sur le champ réfléchi dans la direction spéculaire ($\zeta_r=\zeta_i$) par une surface parfaitement plane de conductivité infinie ($R_{\zeta li} = 1, \forall\ \zeta_{li}$) lorsque le champ incident est polarisé perpendiculairement au plan d'incidence :

$$\rho = \frac{\mathbf{E}_r}{\mathbf{E}_{rs}} \qquad \overline{|\rho|^2} = \overline{|\mathbf{E}_r|^2}/\overline{|\mathbf{E}_{rs}|^2}$$

$\mathbf{E}_{rs}$ peut être calculé à partir de (12) §I.2.3.1 puisque dans ce cas

$$h = h' = 0, \qquad R_{\zeta_{li}} = 1, \qquad \zeta_i = \zeta_r \qquad \text{et } \alpha_r = 0 \qquad (13)$$

D'autre part l'intégration sur la surface projetée selon l'axe des z, A_p, plutôt que sur Σ est réalisée par le changement de variables ad hoc :

$$ds = \sqrt{h'^2_x + h'^2_y + 1}\,dx\,dy = n\,dx\,dy$$

On notera que dans les conditions (13), n = 1, il vient donc :

$$\mathbf{E}_{rs} \approx \frac{iqe^{iq_r d_o}}{4\pi d_o}\mathbf{E}_{oi}\int_{-X}^{X}\int_{-Y}^{Y} -be^{i0}\,dy\,dx = -\frac{iqe^{iq_r d_o}}{4\pi d_o}\mathbf{E}_{oi}2X2Yb_s$$

où b_s est le cas particulier de la composante b dans les conditions (13), i.e. $b_s = 2\cos\zeta_i$, ce qui donne finalement :

$$\mathbf{E}_{rs} \approx -\frac{i2qe^{iq_r d_o}}{\pi d_o}\mathbf{E}_{oi}XY\cos\zeta_i \qquad (14)$$

Egalité pour le moins curieuse puisqu'elle rend l'amplitude du champ électrique proportionnelle à l'aire de la surface, ce qui, a priori, permet d'avoir un champ électrique aussi grand que l'on veut puisqu'il suffit de prendre Σ suffisamment grande. C'est néanmoins l'expression qu'obtient Beckmann au signe près (équation (32) p.23 pour le cas monodimensionnel [Beck et al 63]).

En utilisant cette dernière expression et avec l'aide de (12) §I.2.3.1, ρ s'exprime par :

$$\rho \approx -\frac{1}{8XY\cos\zeta_i}\int_{\Sigma}(ah'_x + ch'_y - b)e^{i\mathbf{v}\cdot\mathbf{p}}n^{-1}ds$$

Cette intégrale n'est pas exprimable analytiquement dans le cas général puisque $R_{\zeta li}$ et donc a,b et c dépendent de la position sur **Σ**. Beckmann fait donc l'hypothèse d'une conductivité infinie pour $\mathbf{E}_r$, ce qui permet l'intégration en rendant le vecteur [a, b, c] constant que l'on notera dans ce cas $[a_s, b_s, c_s]$:

$$I = \int_{-X}^{X}\int_{-Y}^{Y}(a_s h'_x + c_s h'_y - b_s)e^{i\mathbf{v}\cdot\mathbf{p}}nn^{-1}dydx = \int_{-X}^{X}\int_{-Y}^{Y}a_s h'_x e^{i\mathbf{v}\cdot\mathbf{p}}dydx + \int_{-X}^{X}\int_{-Y}^{Y}c_s h'_y e^{i\mathbf{v}\cdot\mathbf{p}}dydx - b_s\int_{-X}^{X}\int_{-Y}^{Y}e^{i\mathbf{v}\cdot\mathbf{p}}dydx$$

Le premier terme peut s'intégrer par parties comme suit :

$$\int_{-X}^{X}\int_{-Y}^{Y}a_s h'_x e^{i(xv_x + yv_y + h(x,y)v_z)}dydx = a_s\int_{-Y}^{Y}e^{iyv_y}\int_{-X}^{X}\underbrace{e^{ixv_x}}_{u}\underbrace{h'_x e^{ih(x,y)v_z}}_{v'}dxdy$$

$$= a_s\int_{-Y}^{Y}e^{iyv_y}\left(e^{ixv_x}\frac{e^{ih(x,y)v_z}}{iv_z}\Big|_{-X}^{X} - \int_{-X}^{X}iv_x e^{ixv_x}\frac{e^{ih(x,y)v_z}}{iv_z}dx\right)dy$$

$$= \frac{a_s}{iv_z}\int_{-Y}^{Y}e^{iyv_y}\left(e^{ixv_x + ih(x,y)v_z}\Big|_{-X}^{X}\right)dy - \frac{a_s v_x}{v_z}\int_{\Sigma}e^{i\mathbf{v}\cdot\mathbf{p}}n^{-1}ds$$

le second terme de I étant de la même forme on peut écrire :

$$I = \frac{a_s}{iv_z}\int_{-Y}^{Y}e^{iyv_y}\left(e^{ixv_x + ih(x,y)v_z}\Big|_{-X}^{X}\right)dy - \frac{a_s v_x}{v_z}\int_{\Sigma}e^{i\mathbf{v}\cdot\mathbf{p}}n^{-1}ds$$
$$+ \frac{c_s}{iv_z}\int_{-X}^{X}e^{ixv_x}\left(e^{iyv_y + ih(x,y)v_z}\Big|_{-Y}^{Y}\right)dx - \frac{c_s v_y}{v_z}\int_{\Sigma}e^{i\mathbf{v}\cdot\mathbf{p}}n^{-1}ds$$
$$- b_s\int_{\Sigma}e^{i\mathbf{v}\cdot\mathbf{p}}n^{-1}ds$$

En remarquant que $i^{-1} = -i$ et en regroupant les termes ayant même facteur on obtient finalement :

$$I = -\left[\left(b_s + \frac{a_s v_x + c_s v_y}{v_z}\right)\int_{\Sigma} e^{i\mathbf{v}\cdot\mathbf{p}} n^{-1} ds + \frac{i a_s}{v_z}\int_{-Y}^{Y} e^{i\mathbf{v}\cdot\mathbf{p}}\Big|_{-X}^{X} dy + \frac{i c_s}{v_z}\int_{-X}^{X} e^{i\mathbf{v}\cdot\mathbf{p}}\Big|_{-Y}^{Y} dx\right] \quad (15)$$

Selon Beckmann, les deux derniers termes de cette dernière expression sont physiquement négligeables par rapport au premier lorsque $XY \gg \lambda^2$. Les termes en question représenteraient en effet la contribution des bords de la surface au champ réfléchi, contribution quasi nulle lorsque les dimensions de la surface sont bien supérieures à la longueur d'onde. Dans ces conditions (12) §I.2.3.1 devient :

$$\mathbf{E}_r(P) \approx -\frac{i q e^{i q_r d_o}}{4\pi d_o}\mathbf{E}_{oi}\left(b_s + \frac{a_s v_x + c_s v_y}{v_z}\right)\int_{\Sigma} e^{i\mathbf{v}\cdot\mathbf{p}} n^{-1} ds \quad (16)$$

Reste à simplifier le facteur en a_s, b_s, c_s en remplaçant les composantes de **v** par leur valeur sans oublier que $R_{\zeta_{li}} = 1$:

$$v = q\left[s_i - s_r c_{ar} \quad -s_r s_{ar} \quad -c_i - c_r\right]$$

$$\left[a \; c \; b\right] = \left[a_s \; c_s \; b_s\right] = q\left[2 s_r c_{ar} \quad 2 s_r s_{ar} \quad 2 c_r\right]$$

$$\frac{b_s v_z + a_s v_x + c_s v_y}{v_z} = \frac{2q(-c_r\,(c_i + c_r) + s_r c_{ar}(s_i - s_r c_{ar}) - (s_r s_{ar})^2)}{-q\,(c_i + c_r)}$$

$$= \frac{-2(-c_i c_r - c_r^2 + s_i s_r c_{ar} - s_r^2)}{c_i + c_r}$$

$$= \frac{2(1 + c_i c_r - s_i s_r c_{ar})}{c_i + c_r}$$

où, pour simplifier, **s** et **c** sont en lieu et place de sin et cos, les indices i, r et ar tenant lieu de ζ_i , ζ_r , α_r respectivement

En remplaçant cette expression dans (16), il vient :

$$\mathbf{E}_r(P) \approx -\frac{i 2(1 + \cos\zeta_i\cos\zeta_r - \sin\zeta_i\sin\zeta_r\cos\alpha_r) q e^{i q_r d_o}}{4\pi d_o(\cos\zeta_i + \cos\zeta_r)}\mathbf{E}_{oi}\int_{\Sigma} e^{i\mathbf{v}\cdot\mathbf{p}} n^{-1} ds$$

Utilisé avec (14) on peut finalement exprimer ρ comme suit

$$\rho \approx \frac{1 + \cos\zeta_i\cos\zeta_r - \sin\zeta_i\sin\zeta_r\cos\alpha_r}{4XY\cos\zeta_i(\cos\zeta_i + \cos\zeta_r)}\int_{\Sigma} e^{i\mathbf{v}\cdot\mathbf{p}} n^{-1} ds$$

$$\approx \frac{F}{A_p}\int_{\Sigma} e^{i\mathbf{v}\cdot\mathbf{p}} n^{-1} ds \quad (17)$$

avec $F = (1 + \cos\zeta_i\cos\zeta_r - \sin\zeta_i\sin\zeta_r\cos\alpha_r)/(\cos\zeta_i(\cos\zeta_i + \cos\zeta_r))$ et $A_p = 4XY$ l'aire de Σ projetée sur le plan xy

L'ensemble des conditions de validité de cette expression est rappelé ci-dessous :

- L'observation se fait à grande distance de la surface (zone de diffraction de Fraunhofer) ⇒ traitement en ondes plane.

- L'approximation du plan tangent nécessite des rayons de courbure faibles. Beckmann suivant Brekhovskikh formalise cette condition par $4\pi r_c \cos\zeta_{li} \gg \lambda$ où r_c désigne le plus petit rayon de courbure de la surface ⇒ validité de la première égalité dans (9) §I.2.3.1.
- Il faut également que l'aire projetée soit suffisamment grande : $A_p \gg \lambda^2$ ⇒ rendre les deux derniers terme de (15) négligeables.
- Pour la même raison que le point précédent, les effets de diffraction par les bords de la surface sont négligés (et la plupart du temps négligeables dans les conditions qui nous intéressent).
- La surface a une conductivité infinie ⇒ facteur de réflexion constant sur l'ensemble de Σ.
- Le champ électrique incident est linéairement polarisé (perpendiculairement ou parallèlement au plan d'incidence).
- Les effets d'ombrage et de réflexions multiples par les aspérités de la surface sont négligés.

I.2.3.3 Surfaces gaussiennes

Malgré les simplifications déjà réalisées pour obtenir l'intégrale finale représentant le coefficient de dispersion au paragraphe précédent il faut d'autres hypothèses, notamment sur la surface, pour pouvoir l'intégrer analytiquement. C'est ce que fait Beckmann en considérant deux grands types de surfaces : celles à profil périodique (sinusoïdal, en dent de scie, etc…) et celles générées par un processus aléatoire. Parmi ces dernières Beckmann considère les surfaces gaussiennes, retenues pour le cas d'école (statistique!) qu'elles représentent plus que par leur représentativité physique dont nous parlerons brièvement plus tard.

Pour ce type de surfaces, h(x,y) représente toujours les hauteurs des aspérités de la surface mais cette fois-ci, possède une distribution w(z) isotrope en x et y de moyenne $\bar{h} = 0$ et d'écart-type σ :

$$w(z) = \left(\sigma\sqrt{2\pi}e^{z^2/(2\sigma^2)}\right)^{-1}$$

La spécification totale de la surface implique la nécessité d'adjoindre un coefficient de corrélation $C(\tau)$ à la distribution :

$$C(\tau) = e^{-\tau^2/T^2}$$

T représentant la distance d'*auto-corrélation* pour laquelle $C(\tau)$ tombe à la valeur e^{-1}

Lorsque σ tend vers 0, w(z) tend vers une distribution de Dirac et la surface se comportera comme si elle était plane. A l'inverse, quand σ est grand, la surface sera très perturbée avec de grandes différences de hauteurs. Lorsque T est grand, la surface ondule amplement et légèrement; en revanche, T petit décrit une surface très discontinue. Ce dernier cas peut conduire à la violation des conditions de validité de l'approximation de Kirchhoff nécessaires, comme nous l'avons vu précédemment, aux développements de Beckmann. Pour que l'approximation soit valide il faut que Σ respecte :

$$T \gg \lambda \qquad \text{et} \qquad T^2 \ll A_p$$

Dans ces conditions et après des traitements statistiques relativement standards, Beckmann obtient l'expression générale suivante :

$$\overline{|\rho|^2} \approx e^{-g}\left(\rho_s^2 + \frac{\pi T^2 F^2}{A_p}\sum_{m=1}^{\infty}\frac{g^m}{m!\,m}e^{-v_{xy}^2 T^2/(4m)}\right) \tag{18a}$$

avec $g = v_z^2\sigma^2$, $v_{xy} = \sqrt{v_x^2 + v_y^2}$ et $\rho_s = \mathrm{sinc}(X v_x)\mathrm{sinc}(Y v_y)$

Deux simplifications supplémentaires sont possibles selon que la surface est légèrement ou fortement rugueuse (g « 1 ou g » 1 respectivement) en ne considérant que les contributions les plus significatives de la série ([Beck et al 63] p.86-87) :

$$\overline{|\rho|^2}_{g\,«\,1} \approx e^{-g}\left(\rho_s^2 + \frac{g\pi T^2 F^2}{A_p}e^{-v_{xy}^2 T^2/4}\right)$$
$$\overline{|\rho|^2}_{g\,»\,1} \approx \frac{\pi T^2 F^2}{A_p g}e^{-v_{xy}^2 T^2/(4g)} \tag{18b}$$

I.2.3.4 Énergie réfléchie dans le modèle gaussien de Beckmann

La quantité intéressante pour nous reste l'énergie qui est proportionnelle au carré du champ électrique comme nous l'avons vu (équations (2a) et (2b) §I.1.2). On peut l'exprimer à l'aide du coefficient de dispersion comme suit :

$$\overline{|\mathbf{E}_r|^2} = \overline{|\rho|^2}\,\overline{|\mathbf{E}_{rs}|^2} \tag{19}$$

En remplaçant l'exponentielle de (14) §I.2.3.2 par sa forme en cosinus on obtient :

$$\begin{aligned}
\overline{|\mathbf{E}_{rs}|^2} = |\mathbf{E}_{rs}|^2 &\approx \left|-\frac{i2q}{\pi d_o}\mathbf{E}_{oi}XY\cos\zeta_i(\cos(q_r d_o) + i\sin(q_r d_o))\right|^2 \\
&\approx \left|\frac{2q}{\pi d_o}\mathbf{E}_{oi}XY\cos\zeta_i\sin(q_r d_o) - \frac{i2q}{\pi d_o}\mathbf{E}_{oi}XY\cos\zeta_i\cos(q_r d_o)\right|^2 \\
&\approx \left(-\frac{2q}{\pi d_o}\mathbf{E}_{oi}XY\cos\zeta_i\right)^2\cos^2(q_r d_o) + \left(\frac{2q}{\pi d_o}\mathbf{E}_{oi}XY\cos\zeta_i\right)^2\sin^2(q_r d_o) \\
&\approx \left(\frac{2q}{\pi d_o}\mathbf{E}_{oi}XY\cos\zeta_i\right)^2
\end{aligned}$$

En remplaçant q par sa valeur, $2\pi/\lambda$, il vient :

$$|\mathbf{E}_{rs}|^2 \approx \left(\frac{4\pi}{\pi d_o\lambda}\mathbf{E}_{oi}XY\cos\zeta_i\right)^2 = \frac{A_p^{\,2}}{d_o^2\lambda^2}\cos^2(\zeta_i)\mathbf{E}_{oi}^2$$

En reportant cette valeur dans (19) et avec l'aide de (18a) et (18b) §I.2.3.3 pour les différents cas de g on obtient :

$$\left|\mathbf{E}_r\right|^2_{g\approx 1} \approx e^{-g}\frac{A_p^{\ 2}}{d_o^2\lambda^2}\cos^2(\zeta_i)\mathbf{E}_{oi}^2\left(\rho_s^2 + \frac{\pi T^2F^2}{A_p}\sum_{m=1}^{\infty}\frac{g^m}{m!m}e^{-v_{xy}^2T^2/(4m)}\right) \tag{20a}$$

$$\left|\mathbf{E}_r\right|^2_{g\ll 1} \approx e^{-g}\frac{A_p^{\ 2}}{d_o^2\lambda^2}\cos^2(\zeta_i)\mathbf{E}_{oi}^2\left(\rho_s^2 + \frac{g\pi T^2F^2}{A_p}e^{-v_{xy}^2T^2/4}\right) \tag{20b}$$

$$\left|\mathbf{E}_r\right|^2_{g\gg 1} \approx \frac{\pi T^2F^2}{gd_o^2\lambda^2}\cos^2(\zeta_i)A_p\mathbf{E}_{oi}^2e^{-v_{xy}^2T^2/(4g)} \tag{20c}$$

En remarquant d'autre part que

$$\begin{aligned}\frac{v^2}{q^2} &= \frac{v_x^2+v_y^2+v_z^2}{q^2} = (s_i - s_rc_{ar})^2 + (s_rs_{ar})^2 + (c_i + c_r)^2\\ &= s_i^2 - 2s_ic_{ar}s_r + c_{ar}^2s_r^2 + s_r^2s_{ar}^2 + c_i^2 + 2c_ic_r + c_r^2\\ &= \underbrace{s_i^2 + c_i^2}_{1} - 2s_ic_{ar}s_r + \underbrace{s_r^2(c_{ar}^2 + s_{ar}^2) + c_r^2}_{1} + 2c_ic_r\\ &= 2(1 - s_ic_{ar}s_r + c_ic_r)\end{aligned}$$

on a

$$F = \frac{1}{\cos\zeta_i}\frac{v^2}{2q^2}\frac{q}{v_z} \qquad \frac{F^2}{g} = \frac{v^4}{\cos^2\zeta_i\ 4q^2v_z^4\sigma^2} \tag{21}$$

En notant que

$$\frac{v^4}{v_z^4} = \frac{(v_{xy}^2+v_z^2)^2}{v_z^4} = \frac{v_{xy}^4 + 2v_{xy}^2v_z^2 + v_z^4}{v_z^4} = \frac{v_{xy}^2}{v_z^2}\left(\frac{v_{xy}^2}{v_z^2} + 2\right) + 1$$

et en appelant β l'angle entre $\mathbf{v}$ et l'axe des z, on a :

$$\tan\beta = \frac{v_{xy}}{v_z}$$

que l'on peut inclure dans le rapport précédent pour obtenir

$$\begin{aligned}v^4/v_z^4 &= \tan^2\beta(\tan^2\beta + 2) + 1 = 1 + \tan^4\beta + 2\tan^2\beta\\ &= (1+\tan^2\beta)^2 = \left(\frac{\cos^2\beta + \sin^2\beta}{\cos^2\beta}\right)^2 = \cos^{-4}\beta\end{aligned}$$

soit en reportant dans (21) et en introduisant $\tan\beta_o = 2\sigma/T$:

$$\frac{F^2T^2}{g} = \frac{\cos^{-4}\beta}{\cos^2\zeta_i\ q^2}\frac{T^2}{4\sigma^2} = \frac{\cos^{-4}\beta}{\cos^2\zeta_i\ q^2}\tan^{-2}\beta_o$$

Finalement en utilisant le rapport précédent dans (20c) avec la valeur de q il vient

$$\left|\mathbf{E}_r\right|^2_{g\gg 1} \approx \frac{A_p}{4\pi d_o^2\tan^2\beta_o\cos^4\beta}\mathbf{E}_{oi}^2e^{-\tan^2\beta/\tan^2\beta_o}$$

Ici encore malheureusement, comme dans l'expression de $\mathbf{E}_{rs}$, on retrouve une dépendance sur l'aire projetée.

Les surfaces à conductivité finie ont été traitées de façon plus qualitative que quantitative par Beckmann. Le problème possède en effet un degré de complexité supérieur puisque, comme nous l'avons déjà remarqué, le coefficient de réflexion dépend alors de l'incidence locale en chaque point de la surface ainsi que de la polarisation du champ incident, ce qui interdit les simplifications menant à (16) §I.2.3.2. Beckmann propose donc de ne pas considérer le coefficient de réflexion local mais plutôt sa moyenne $\overline{R}$ sur la surface ce qui permet de remplacer le vecteur [a, b, c] dans (11) §I.2.3.1 par le vecteur moyen [$\bar{a}$, $\bar{b}$, $\bar{c}$]. Néanmoins, pour bien faire, il faudrait évaluer :

$$\overline{R} = \int_{-\infty}^{\infty} p(h')R(h')dh'$$

$p(h')$ étant la distribution de h'

Intégrale encore une fois trop complexe à évaluer, une approximation supplémentaire est donc introduite, approximation qui consiste à assimiler la valeur moyenne du coefficient de réflexion à la valeur de ce coefficient pour un angle égal à l'angle d'incidence par rapport à la surface moyenne :

$$\overline{R(h')} \approx R(h' = \bar{h}') = R(\zeta_{li} = \zeta_i)$$

ce qui permet de considérer [$\bar{a}$, $\bar{b}$, $\bar{c}$] et donc de réaliser les mêmes développements que pour la conductivité infinie en remplaçant R par $R(\zeta_{li} = \zeta_i)$. Cette conjecture a été confirmée à l'aide d'une approche différente par Stogryn [Stog 67] mais uniquement lorsque le champ réfléchi est dans le plan d'incidence ($\alpha_r = 0$) pour des surfaces fortement rugueuses. Dans ce cas, Stogryn obtient :

$$\left|\mathbf{E}_{r_p}\right|^2_{g \gg 1, \alpha_r = 0} \approx R_p(\zeta_m)\left|\mathbf{E}_r\right|^2_{g \gg 1}$$

où p désigne la polarisation perpendiculaire ($\perp$) ou parallèle ($\parallel$) des champs électriques (Stogryn note qu'un champ incident d'un certain type de polarisation ne contribue qu'au champ réfléchi de même polarisation).
ζ_m est le demi-angle entre $-\mathbf{q}_i$ et $\mathbf{q}_r$

Hors du plan d'incidence, mais toujours pour des surfaces isotropiques très rugueuses, Stogryn obtient les expressions du champ réfléchi total (polarisations réfléchies confondues) suivantes :

$$\left|\mathbf{E}_r\right|^2_{g \gg 1, I = \perp} \approx \left|\mathbf{E}_r\right|^2_{g \gg 1} \frac{R_\perp(\zeta_m)(\mathbf{q}_{i\parallel} \cdot \mathbf{q}_r)^2 + R_\parallel(\zeta_m)(\mathbf{q}_{i\perp} \cdot \mathbf{q}_r)^2}{(\mathbf{q}_{r\parallel} \cdot \mathbf{q}_i)^2 + (\mathbf{q}_{r\perp} \cdot \mathbf{q}_i)^2}$$

$$\left|\mathbf{E}_r\right|^2_{g \gg 1, I = \parallel} \approx \left|\mathbf{E}_r\right|^2_{g \gg 1} \frac{R_\parallel(\zeta_m)(\mathbf{q}_{i\parallel} \cdot \mathbf{q}_r)^2 + R_\perp(\zeta_m)(\mathbf{q}_{i\perp} \cdot \mathbf{q}_r)^2}{(\mathbf{q}_{r\parallel} \cdot \mathbf{q}_i)^2 + (\mathbf{q}_{r\perp} \cdot \mathbf{q}_i)^2}$$

$\mathbf{q}_{i\parallel} = \cos\zeta_i \mathbf{i} + \sin\zeta_i \mathbf{k}$ et $\mathbf{q}_{i\perp} = -\mathbf{j}$
L'indice en $I = \parallel$ ou $I = \perp$ indique que la composante réfléchie est issue de la polarisation $\parallel$ ou $\perp$ du champ incident

Qui ne permettent pas de déduire les composantes du vecteur réfléchi (voir l'équation (29) §II.1.2).

I.2.4 Auto-ombrage des facettes

Les modèles élaborés par Stogryn ou Beckmann font toujours l'hypothèse que l'ensemble de la surface éclairée participe à la réflexion. En fait, toute surface rugueuse porte ombrage sur elle-même. C'est pour combler cette lacune que Beckmann a proposé un modèle d'ombrage par une surface gaussienne en 65 [Beck 65]. Modèle insuffisant et vivement critiqué par Brokelmann et Hagfors [Broc et al 66] ce qui conduisit Smith à développer une approche plus satisfaisante [Smit 67]. Il cherche la probabilité que le rayon incident soit intercepté avant d'atteindre un point donné d'une surface supposée gaussienne. Pour ce faire, il cherche à évaluer la probabilité $g(\tau)\Delta\tau$ que la surface dans un intervalle $\Delta\tau$ intersecte un rayon sachant que ce rayon n'a pas été intersecté jusqu'à τ. Cette probabilité nécessite cependant de connaître la fonction d'auto-corrélation ainsi que la distribution précise de la surface. En dehors du fait que le calcul serait assez lourd, la connaissance même de ces caractéristiques surfaciques est souvent hors de portée.

Aussi Smith simplifie-t-il le problème en négligeant la corrélation entre pente et hauteur de la surface aux points d'intersections. Avec ces hypothèses, sans doute trop simplificatrices, il établit la probabilité $I(\zeta_i)$ qu'un point soit illuminé indépendamment des pentes et de la hauteur :

$$I(\zeta_i) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P_1(h)P_2(p_1)P_2(p_2)S(\zeta_i, h, p_1, p_2)dh\,dp_1\,dp_2$$

$P_1(h)\Delta h$ est la probabilité de trouver une déviation de hauteur dans l'intervalle Δh autour de h
$P_2(p)P_2(q)$ est la probabilité conjointe des pentes p et q de la surface
$S(\zeta_i, h, p, q)$ est la probabilité qu'une demi-droite dont le point d'appui a une hauteur h sur la surface et de pente p et q soit illuminé par un rayon incident faisant un angle ζ_i avec la normale moyenne

Probabilité que Smith trouve égale à :

$$I(\zeta_i) = \frac{1 - \mathrm{erfc}(c/(\sigma_p\sqrt{2}))/2}{\Lambda(c) + 1}$$

$$\Lambda(c) = \left(\sqrt{\frac{2}{\pi}}\frac{\sigma_p}{c}e^{-(c^2/(2\sigma_p^2))} - \mathrm{erfc}(c/(\sigma_p\sqrt{2}))\right)/2$$

$c = \cot\zeta_i$ et σ_p est l'écart-type des pentes

$$\mathrm{erfc}(x) = 1 - \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$$

I.2.5 Résultats expérimentaux

I.2.5.1 Mesure de la rugosité

La mesure du profil de surfaces peut être réalisée selon deux grandes familles de méthodes : celles par contact et celles à distance.

Les premières utilisent un stylet à pointe de diamant qui suit les variations de hauteurs soit par son déplacement propre soit par déplacement de l'échantillon. Les mouvements du stylet convertis en impulsions électriques actionnent un traceur ou sont passés à un convertisseur analogique-numérique. Avant toute mesure, l'instrument doit être calibré sur une surface de profil connu et la surface à mesurer doit être dépourvue de toutes poussières ou particules étrangères pouvant fausser les résultats. De plus, il est recommandé d'isoler l'installation de mesure de toutes vibrations

Les résultats fournis par ce type d'instrument ont deux inconvénients majeurs. Le premier provient de la charge appliquée sur le stylet. Vu la dureté du diamant, il peut facilement endommager l'échantillon ce qui rend impossible la répétition des mesures. Le second inconvénient est que les résultats fournis ne décrivent pas directement la surface mais plutôt sa convolution par le stylet et il n'est généralement pas possible de déconvoluer le signal résultant. Les mesures sont donc plus représentatives des fréquences spatiales du profil que des variations de hauteurs elles-mêmes.

Le second grand type de mesures utilise la lumière comme senseur et regroupe en fait de multiples méthodes différentes. Certaines détectent le changement de focalisation réalisée par la surface pendant que d'autres s'attachent à la détection du changement de direction du rayon lumineux après réflexion. La plupart d'entre elles utilisent un interféromètre allié à un microscope.

Contrairement aux méthodes par contacts, les méthodes à distance ne nécessitent pas de calibration préalable. La propreté de l'échantillon est bien entendu aussi indispensable que précédemment. La précision des mesures dépend essentiellement de la résolution du microscope et des éléments optiques adjoints ainsi que de la longueur d'onde éclairant l'échantillon.

Si ces méthodes sont non-destructives contrairement aux premières, elles ont néanmoins quelques inconvénients elles aussi. Tout d'abord, ici encore le vrai profil n'est pas directement accessible mais seulement sa moyenne sur une portion de surface (voir figure ci-dessous). D'autre part la résolution latérale est beaucoup plus grossière qu'avec un stylet (de l'ordre d'un facteur 10). Enfin certaines méthodes dépendent également de la précision des constantes optiques du matériau analysé.

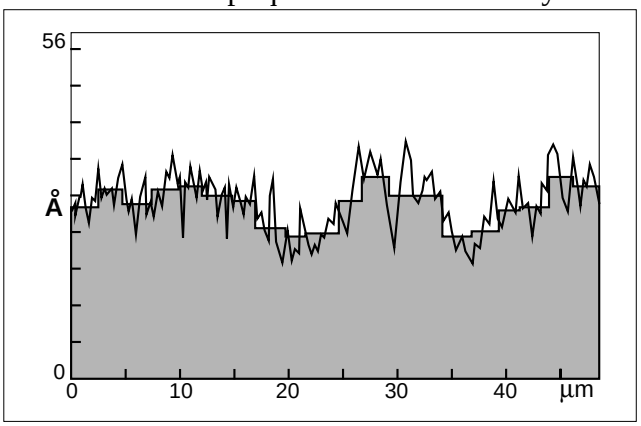


Figure 11: Profil moyen mesuré par un profilomètre optique comparée au profil réel

I.2.5.2 Modèle beckmannien comparé à l'expérience.

Un certain nombre d'expériences ont eu lieu afin de vérifier l'adéquation de la théorie de Beckmann avec des mesures de laboratoire. O'Donnell et Mendez [Mend et al 87] sont sans doute les premiers à avoir voulu vérifier la théorie sur des surfaces ayant des profils bien déterminés. Ils fabriquent donc des surfaces ayant des distributions gaussiennes par construction. Sans entrer dans les détails, disons seulement qu'elles sont ensuite recouvertes d'une fine couche d'or ou d'aluminium. Les échantillons peuvent être éclairés par trois lasers différents, deux dans le visible et le dernier dans l'infrarouge, à des longueurs d'onde de 0,633 μm (He-Ne), 0,514 μm (Argon-ion) et 10,6 μm (dioxyde de carbone) sur des portions de 20 mm de diamètre. L'énergie mesurée est uniquement celle réfléchie dans le plan d'incidence. Le calcul se fait à partir de l'équation (17) §I.2.3.2 mais avec une autre forme pour F, facteur qui semble avoir été beaucoup critiqué dans la littérature, où le $\cos\zeta_i(\ldots)$ est remplacé par $\cos\zeta_r(\ldots)$ selon des calculs effectués par Nieto-Vesperinas [Niet 82].

Leurs résultats montrent que les calculs théoriques sont en accord avec l'expérience lorsque la surface est très rugueuse et pour des angles d'incidence inférieurs à 50°. Les résultats d'une surface quasi-gaussienne, de rugosité σ = 2,27 μm et dont la fonction d'auto-corrélation suit également une gaussienne (T = 21,9 μm), éclairée par un faisceau incident de 0,633 μm faisant un angle de 20° avec la normale moyenne sont repris ci-dessous.

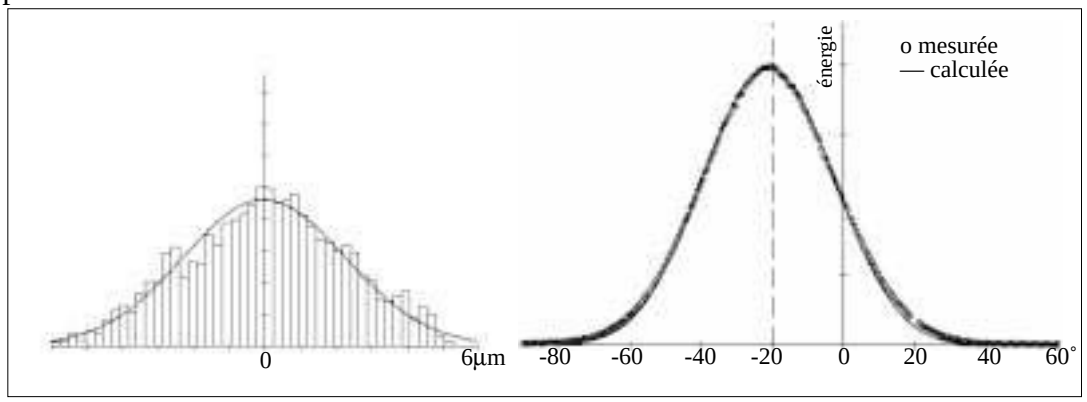


Figure 12: Distribution des hauteurs (à gauche) et énergie réfléchie (à droite). Le trait vertical en pointillé montre l'angle de réflexion spéculaire

Lorsque l'incidence augmente jusqu'à 70°, il n'y a plus convergence entre théorie et mesures comme le montre la figure suivante :

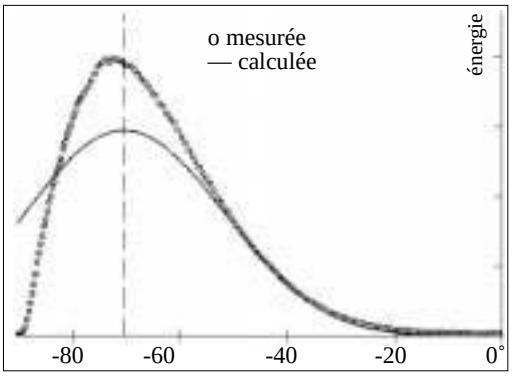


Figure 13: Énergie réfléchie pour une incidence de 70°

Plus encore, les calculs prédisent une dispersion derrière la surface. Les auteurs pensent que ceci est essentiellement imputable à l'absence des effets de masquage et de réflexions multiples dans le modèle. D'autres expériences impliquant des surfaces beaucoup moins rugueuses ont été menées par les auteurs. Une de leurs conclusions est qu'aucune théorie n'explique quantitativement les résultats obtenus même si certaines décrivent quelques traits marquants de façon qualitative. L'un de leurs résultats le plus problématique provient d'une surface dont les pentes sont très fortes et de rugosité très faible, à tel point que la rugosité et la distance d'auto-corrélation n'ont pu être mesurées par leur profilomètre mécanique. Ils ont estimé ces paramètres à l'aide de leur processus de fabrication des surfaces, $\sigma \approx 1\ \mu m$ et $T \approx 1{,}4\ \mu m$. Eclairé à 10° avec une longueur d'onde de 0,633 µm les composantes parallèles et perpendiculaires réfléchies sont reprises ci-dessous :

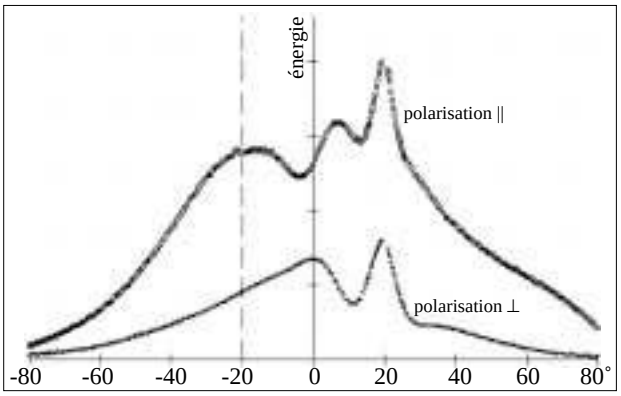


Figure 14: Énergie réfléchie pour une incidence de 10°. Le trait vertical en pointillé montre l'angle de réflexion spéculaire.

On remarque notamment qu'il n'y a pas de composante spéculaire avec cette longueur d'onde incidente située dans le visible. D'autre part, l'intensité maximale se situe dans la direction de «rétro-dispersion» (backscattering en anglais), i.e. la direction d'incidence. Dernière remarque, la dispersion ne peut s'expliquer par un phénomène additif entre diffus et spéculaire.

Marx et Vorburger se sont intéressés aux mesures effectuées sur des surfaces moyennement rugueuses [Marx et al 90]. A partir de dix échantillons d'acier inoxydable dont les surfaces suivent assez bien des distributions gaussiennes, ils mesurent l'énergie diffuse réfléchie par ces surfaces dans l'ensemble de l'hémisphère entourant la surface. Ils confrontent ensuite ces mesures avec les résultats de calculs effectués avec une forme de l'équation (17) §I.2.3.2 dans le cas monodimensionnel et en prenant une expression de F légèrement différente :

$$F = (1 + \cos\zeta_i \cos\zeta_r + \sin\zeta_i \sin\zeta_r) / (\cos\zeta_i (\cos\zeta_i + \cos\zeta_r))$$

> On pourra remarquer que l'absence du $\cos\alpha_r$ peut s'expliquer par le fait que le faisceau réfléchi est peu étendu selon les auteurs (environ 4°). En revanche la transformation de signe du produit des sinus est plus curieuse et inexpliquée.

Les rugosités (σ) des surfaces ont été mesurées sur des profils de 800 µm avec un appareil à stylet et s'étalent de 0,08 à 0,48 µm, les distances d'auto-corrélation (T) varient entre 1,52 et 4,07 µm. Les échantillons ont été éclairés sur une surface d'environ 2x3 mm par un laser He-Ne de longueur d'onde 0,6328 µm (région visible médiane).

Pour exemple, les deux graphiques suivants représentent les calculs et les mesures pour un angle d'incidence (ζ_i) de 54° et pour deux surfaces, l'une de rugosité 0,484 µm (T = 3,14 µm), l'autre de rugosité 0,08 µm (T = 1,52 µm).

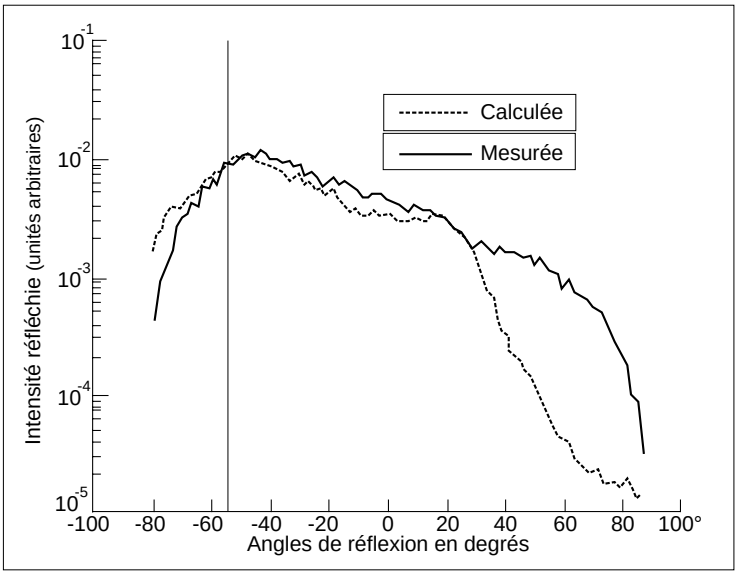


Figure 15: Dispersion par la surface de rugosité 0,484 µm

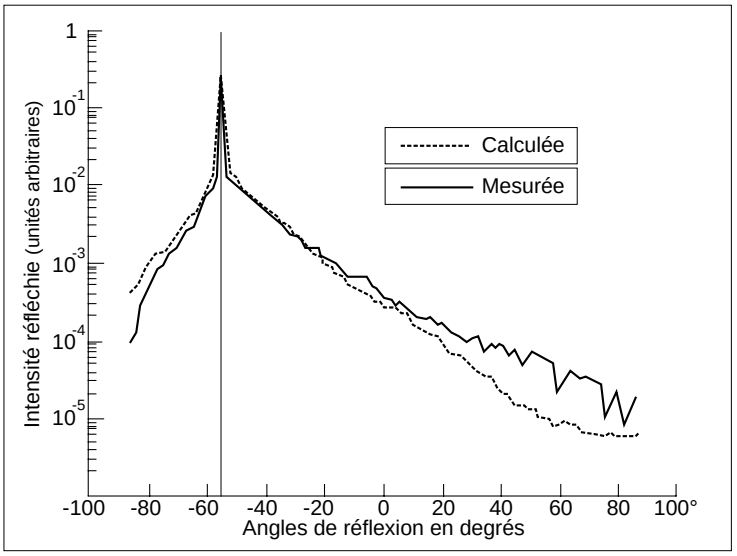


Figure 16: Dispersion par la surface de rugosité 0,08 µm

A noter que ces figures ont une échelle logarithmique pour l'énergie. Les auteurs interprètent leurs résultats comme étant en bon accord avec la théorie. Ils expliquent la discordance pour des grands angles par rapport à la direction spéculaire par la non prise en compte des réflexions multiples et de l'auto-ombrage ainsi que par la violation de l'approximation de Kirchhoff. Ils font également ressortir que la dispersion n'est qualitativement que moyennement sensible aux deux paramètres de la surface σ et T.

I.3 Bilan

La théorie électromagnétique est incontournable à la description des phénomènes lumineux les plus habituels. Sans l'aspect ondulatoire, comment expliquer ne serait-ce que la transmission ?… Cette approche nécessite non seulement de décrire finement les propriétés de la matière et des sources émettrices d'ondes mais également, et non moins finement, l'état des surfaces et leurs intéractions avec ces ondes. Le cas, apparemment fort simple, des surfaces planes est en fait déjà complexe et nous a amené à quelques simplifications du réel (absence de phénomènes de résonnance, matériaux amagnétiques, indices complexes constants sur la surface, etc). Le passage à des surfaces réelles (rugueuses) possède un degré de complexité supérieur, tant du point de vue de la description des surfaces (beaucoup d'hypothèses parfois trop simplificatrices ou invérifiées) que du comportement électromagnétique à leur contact. L'approche de Beckmann nous paraît nettement préférable aux hypothèses géométriques (Torrance-Sparrow), parce que moins contraignante pour le réel et mieux à même de décrire les échanges énergétiques dans leurs singularités ondulatoires. Malgré les développements de Stogryn, prolongeant les travaux de Beckmann en considérant des surfaces à conductivité finie indispensables dans le visible pour des matériaux courants, il n'y a pas encore, à notre connaissance, de modèle physique décrivant la réflexion des champs électromagnétiques par une surface rugueuse sous forme vectorielle, la seule capable de décrire complètement le phénomène.

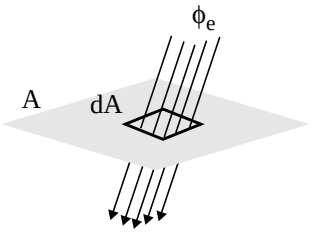
I.4 Radiométrie

I.4.1 Définitions

Le flux énergétique est l'énergie qui traverse par unité de temps, une surface d'aire finie dans un sens donné. Il est noté ϕ_e et se mesure en watt. Il est relié aux quantités électromagnétiques que nous avons définies au début de ce chapitre de la façon suivante :

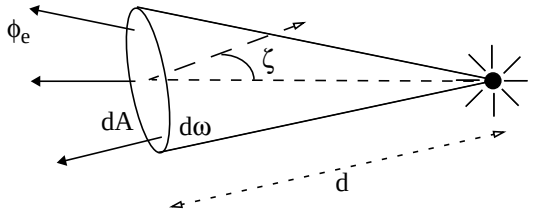
$$\phi_e = \frac{1}{2}\sqrt{\frac{\varepsilon}{\mu}}\mathbf{E}_o^2 = \bar{\mathbf{S}}$$

Si on considère une aire élémentaire dA sur cette surface, on obtient l'éclairement énergétique de la surface sur cette aire élémentaire [Afe 91] :



$$E_e = \frac{d\phi_e}{dA} \text{ en } \mathtt{W.m^{-2}}$$

Selon que l'on considère A comme une surface émettrice ou réceptrice on parlera d'*exitance* ou d'*éclairement* énergétique respectivement. L'*intensité* énergétique permet de considérer l'énergie dans une direction donnée ou encore pour une source ponctuelle :

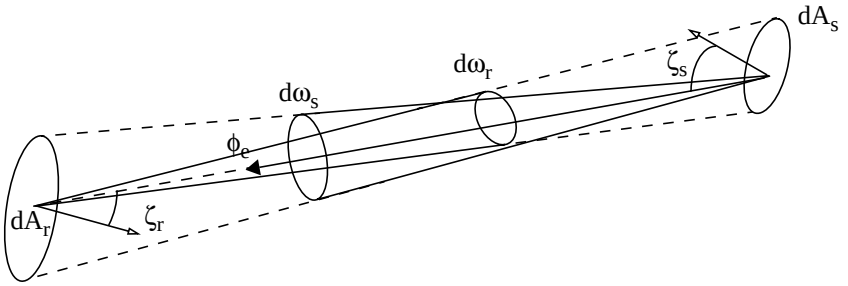


$$I_e = \frac{d\phi_e}{d\omega} \text{ en } \mathtt{W.sr^{-1}}$$

où $d\omega$ est l'angle solide contenant la direction considérée.

Dans le cas où $d\omega$ représente un angle solide élémentaire, c'est à dire un cône d'angle suffisamment petit pour pouvoir assimiler sa surface d'appui à une surface plane, $d\omega$ s'écrit : $d\omega = dA\cos\zeta / d^2$

La dernière quantité restant à définir est la *luminance* énergétique en un point d'une surface et dans une direction donnée :

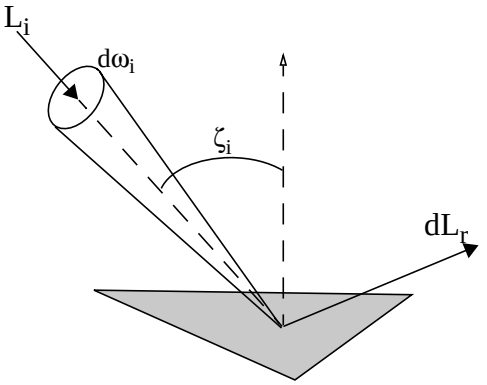


$$L_e = \frac{d^2\phi_e}{d\omega_s dA_s \cos\zeta_s} = \frac{d^2\phi_e}{d\omega_r dA_r \cos\zeta_r} = \frac{dE_e}{\cos\zeta_r d\omega_r} \text{ en } W.sr^{-1}.m^{-2}$$

I.4.2 Réflectances bidirectionnelles

À partir des différentes distributions de luminance, il est possible de définir la *fonction de distribution de réflectance bidirectionnelle* (FDRB ou BRDF pour l'acronyme anglais) comme le rapport de la luminance réfléchie sur l'éclairement incident pour des directions données [Nico et al 77] :

$$f_r(\lambda, \zeta_i, \alpha_i, \zeta_r, \alpha_r) = \frac{dL_r(\lambda, \zeta_i, \alpha_i, \zeta_r, \alpha_r)}{L_i(\lambda, \zeta_i, \alpha_i)\cos\zeta_i d\omega_i} = \frac{dL_r(\lambda, \zeta_i, \alpha_i, \zeta_r, \alpha_r)}{dE_i(\lambda, \zeta_i, \alpha_i)} \text{ en } sr^{-1} \qquad (22)$$



À partir de ce concept, on peut retrouver une expression générale de la réflectance :

$$\rho(\lambda, \Omega_i, \Omega_r) = \frac{d\phi_r}{d\phi_i} = \frac{\int_{\Omega_r}\int_{\Omega_i} f_r(\lambda, \zeta_i, \alpha_i, \zeta_r, \alpha_r) L_i(\lambda, \zeta_i, \alpha_i)\cos\zeta_i \cos\zeta_r d\omega_i d\omega_r}{\int_{\Omega_i} L_i(\lambda, \zeta_i, \alpha_i)\cos\zeta_i d\omega_i}$$

Une foule d'adjectifs (conique, biconique, hémisphérique, etc) peuvent accompagner le nom de réflectance selon les diverses configurations géométriques des faisceaux incidents et réfléchis. $\overline{R}$ n'est donc qu'un cas particulier pour des surfaces parfaitement planes où les angles d'incidence et de réflexion sont symétriques et égaux, tout comme les angles solides considérés dont l'étendue tend vers 0.

De la même façon, on peut définir une *fonction de distribution de transmittance bidirectionnelle* (FDTB) :

$$f_t(\lambda, \zeta_i, \alpha_i, \zeta_t, \alpha_t) = \frac{dL_t(\lambda, \zeta_i, \alpha_i, \zeta_t, \alpha_t)}{L_i(\lambda, \zeta_i, \alpha_i)\cos\zeta_i d\omega_i} = \frac{dL_t(\lambda, \zeta_i, \alpha_i, \zeta_t, \alpha_t)}{dE_i(\lambda, \zeta_i, \alpha_i)}$$

Ainsi qu'une transmittance générale

$$\tau(\lambda, \omega_i, \omega_t) = \frac{d\phi_t}{d\phi_i} = \frac{\int_{\Omega_t}\int_{\Omega_i} f_t(\lambda, \zeta_i, \alpha_i, \zeta_t, \alpha_t) L_i(\lambda, \zeta_i, \alpha_i)\cos\zeta_i \cos\zeta_t d\omega_i d\omega_t}{\int_{\Omega_i} L_i(\lambda, \zeta_i, \alpha_i)\cos\zeta_i d\omega_i}$$

La CIE† définit la *réflectivité* $\rho_\infty(\lambda)$ comme étant la réflectance d'un matériau d'épaisseur infinie (telle que l'épaisseur n'ait pas d'influence sur cette propriété), la *transmittivité* $\tau_{i,o}(\lambda)$ comme étant la transmittance à une distance unitaire de l'interface et enfin l'*absorptivité* comme le rapport du flux absorbé à une distance unitaire de l'interface sur le flux incident.

Un certain nombre de cas particuliers sont intéressants à détailler :

- Le *diffuseur parfait* ou surface *lambertienne* est tel que la luminance réfléchie est constante quel que soit l'angle de réflexion pour une luminance incidente donnée :

$$\rho(\lambda, \omega_i, 2\pi) = \frac{L_r(\lambda, \zeta_i, \alpha_i, \zeta_r, \alpha_r)}{E_i}\int_{2\pi}\cos\zeta_r d\omega_r = f_r(\lambda, \zeta_i, \alpha_i, \zeta_r, \alpha_r)\pi$$

Un *diffuseur idéal* est tel que $f_r(\lambda, \zeta_i, \alpha_i, \zeta_r, \alpha_r) = \pi^{-1}$

- Une surface *spéculaire* est décrite par

$$f_r(\lambda, \zeta_i, \alpha_i, \zeta_r, \alpha_r) = 2\frac{L_r(\lambda, \zeta_r, \alpha_r, \zeta_r, \alpha_r \pm \pi)}{L_i(\lambda, \zeta_r, \alpha_r \pm \pi)}\delta(\sin^2\zeta_r - \sin^2\zeta_i)\delta(\alpha_r + \alpha_i \pm \pi)$$

où δ est une fonction de Dirac

une surface *spéculaire idéale* est définie par

$$f_r(\lambda, \zeta_i, \alpha_i, \zeta_r, \alpha_r) = 2\delta(\sin^2\zeta_r - \sin^2\zeta_i)\delta(\alpha_r + \alpha_i \pm \pi)$$

A noter qu'aucun corps physique ne se comporte comme ces corps parfaits. Plus encore, pratiquement aucun corps ne se comporte comme une combinaison linéaire de ces deux états. Pour exemple, le comportement de l'aluminium est loin de cet idéal comme le montre la figure ci-après même si elle représente un état semi-diffus du matériau.

† Commission Internationale de l'Eclairage

Figure 17: Sculpture représentant la distribution de la luminance réfléchie par unité de luminance incidente produite par un faisceau de lumière collimatée (à gauche) incident à 33,2˚ sur un morceau d'aluminium. Cette sculpture est faite de pans verticaux dont la découpe suit la variation de la distribution. L'échelle logarithmique «azimutale» utilisée pour reporter les valeurs de f_r accentue les variations pour les valeurs peu élevées

Une propriété très intéressante de la FDRB est la *réciprocité* valable en l'absence de polarisation et de champ magnétique et découlant de la réciprocité d'Helmohltz

$$f_r(\lambda, \zeta_i, \alpha_i, \zeta_r, \alpha_r) = f_r(\lambda, \zeta_r, \alpha_r, \zeta_i, \alpha_i)$$

I.5 Photométrie

I.5.1 Égalisation en luminosité

Jusqu'à présent nous n'avons parlé que d'ondes. Chacun sait que l'ensemble du domaine fréquenciel ne provoque pas de réponse visuelle. Nous ne voyons, et c'est bien dommage, ni l'infra-rouge, ni les rayons X, ni les ondes radio; l'oeil n'est en effet sensible qu'à une plage bien restreinte d'ondes : la lumière. Pour tenir compte de cette sensibilité visuelle, la photométrie, science qui s'intéresse aux rayonnements dans leurs manifestations visuelles, définit deux observateurs standard et un certain nombre de lois. L'oeil est bien loin de n'être qu'un appareil de mesure dont la réponse est immuable dans le temps pour un même individu et encore moins d'un individu à l'autre.

Les deux observateurs standard retenus par la photométrie ont été obtenus

après de multiples expériences sur de nombreux individus. Les résultats de ces expériences ont ensuite été moyennés et lissés pour en faire des données facilement manipulables éliminant les variations locales. Dans ce sens, on peut dire que la photométrie ne se préoccupe que de résultats moyens, qui peuvent bien évidemment s'éloigner des sensations individuelles mais permettent d'avoir une bonne idée du passage des quantités énergétiques physiquement mesurables aux quantités visuelles.

L'opération fondamentale en photométrie est l'*égalisation visuelle* ou *égalisation en luminance visuelle*. Elle consiste à juger si deux luminances produisent la même sensation. Plus précisément, la notion de luminance visuelle peut être définie comme l'attribut d'une sensation visuelle selon laquelle un stimulus visuel apparaît plus ou moins intense ou encore selon laquelle la région contenant le stimulus émet plus ou moins de lumière.

Plusieurs procédés d'égalisation existent; on peut citer l'égalisation par papillotement, l'égalisation pas à pas ou encore la distinction minimale de la limite de séparation. Pour exemple, l'égalisation par papillotement consiste à proposer à un observateur une plage éclairée alternativement par deux sources, puis à lui faire ajuster la puissance de la seconde jusqu'à ce que le papillotement soit minimal (éventuellement nul). Les différentes méthodes ne mènent pas toutes à des résultats équivalents [Kowa 80], néanmoins la méthode par papillotement et celle pas à pas concordent et leurs résultats sont suffisamment stables pour divers observateurs. Le principe d'égalisation, si les deux sources sont monochromatiques et de même longueur d'onde, revient à établir $v(\lambda)$ et K_m tels que [Guth 70] :

$$L_v(\lambda) = K_m v(\lambda) L_e(\lambda)$$

$L_v(\lambda)$ est une grandeur caractérisant la *luminance visuelle*
K_m est le *coefficient d'efficacité lumineuse*, un facteur de proportionnalité dépendant des unités choisies.
$v(\lambda)$ représente la sensibilité de l'oeil
$L_e(\lambda)$ est la luminance énergétique

Il aurait été possible de choisir une autre grandeur que la luminance énergétique mais les expériences montrent que la sensation de luminosité est liée à celle de luminance, toutes choses étant égales par ailleurs.

Si les deux lumières sont de longueur d'onde différentes λ_1, λ_2, on cherchera à établir une relation du type :

$$L_v(\lambda_1) = K_m v(\lambda_1) L_e(\lambda_1)$$

$$L_v(\lambda_2) = K_m v(\lambda_2) L_e(\lambda_2)$$

où les luminosités seront identiques lorsque l'utilisateur signalera une égalisation :

$$v(\lambda_1) L_e(\lambda_1) = v(\lambda_2) L_e(\lambda_2) = L_v$$

On a pu constater que, pour une longueur d'onde de 555 nm, $L_e(555)$ était minimum ce qui signifie que ce rayonnement est particulièrement efficace. Il a donc été décidé de normaliser $v(\lambda)$ et de créer ainsi $V(\lambda)$, la *courbe spectrale d'efficacité relative* telle que :

$$V(555) = 1 \text{ et } V(\lambda) = L_e(555) / L_e(\lambda)$$

Les conditions d'expériences menant à l'établissement de cette courbe sont très strictes afin de bien isoler le phénomène. Une discussion détaillée se trouve dans l'excellent ouvrage de Stiles et Wyszecki [Stil et al 82]. Nous retiendrons notam-

ment que V(λ) est représentative d'une vision centrale avec des champs de 2° à 4° (voir figure ci-après) où les cellules en cônes de l'oeil sont presqu'exclusivement sollicitées, les luminances utilisées étant souvent inférieures à 10 $cd.m^{-2}$.

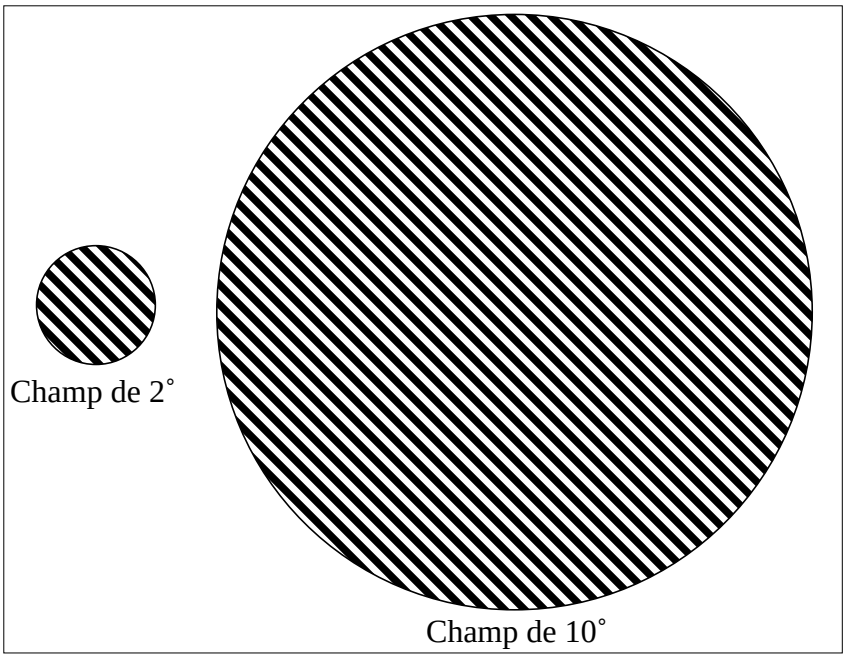
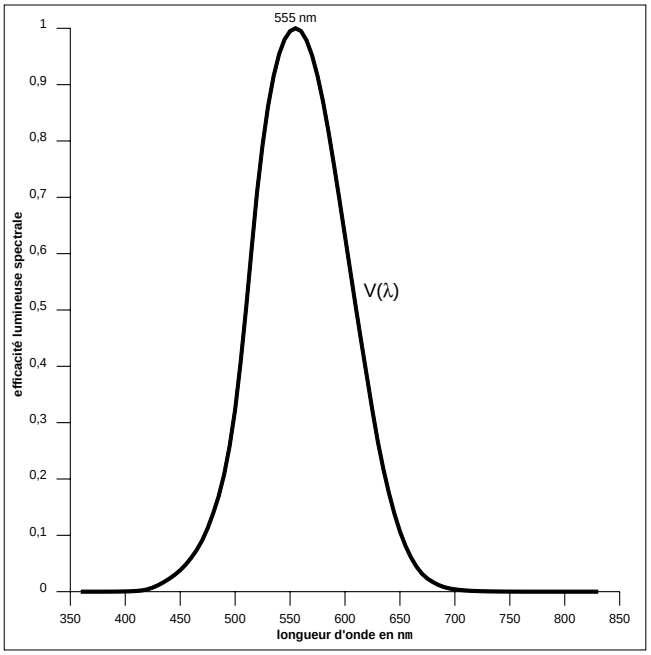


Figure 18: Vus à 45 cm ces deux cercles sont représentatifs de champ de vision de 2° (à gauche) et de 10° (à droite) retenus pour les observateurs standards 1931 et 1964 respectivement.

Ce type de vision, caractéristique d'une vision diurne, est dit *photopique* et a été adopté en 1924 par la CIE et en 1933 par le CIPM† pour une courbe de 380 à 780 nm dans l'air standard tous les 10 nm. Puis en 1971 une nouvelle courbe a été adoptée par la CIE (1972 par le CIPM) définie cette fois de 360 à 830 nm par pas de 1 nm, c'est cette dernière qui doit être utilisée. La définition de V(λ) implique donc celle du *visible photopique* : ϑ = [360, 830].



† Commission Internationale des Poids et Mesures

Figure 19: Éfficacité relative spectrale photopique

Une autre courbe, V'(λ) (figure ci-dessous) a été adoptée par la CIE en 1951 et en 1976 par le CIPM. Les expériences comportaient l'observation sous des angles d'au moins 5° par rapport à la fovéa de personnes de moins de 30 ans préalablement habituées à l'obscurité. Cette vision dite *scotopique* (vision nocturne) est déterminée de façon prédominante par les cellules en bâtonnets, l'efficacité maximum étant alors située à 507 nm†.

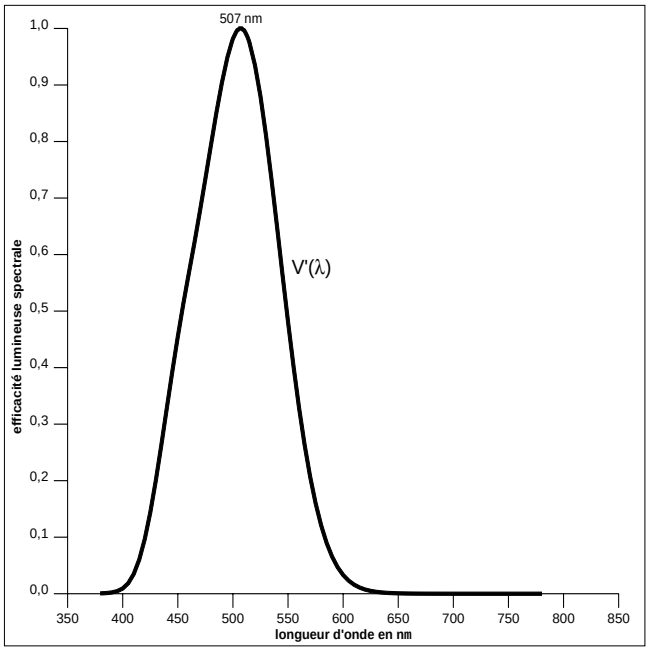


Figure 20: Éfficacité relative spectrale scotopique

Un troisième type de vision est dit *mésopique* lorsque les conditions de vision sont situées entre photopiques et scotopiques. Aucune courbe de sensibilité n'a été adoptée pour ce dernier type.

I.5.2 Les lois de la photométrie

L'égalisation en luminance visuelle doit vérifier un certain nombre de postulats que voici :

- Symétrie
  Si un stimulus A égalise un stimulus B alors B égalise A
- Transitivité
  Si A égalise B et B égalise C alors A égalise C
- Proportionnalité
  Si A égalise B alors αA égalise αB, où α est un facteur positif par lequel la puissance des stimuli est augmentée ou diminuée, la composition spectrale restant inchangée
- Additivité
  Si A égalise B, C égalise D et que A+C égalise B+D alors A+D égalise B+C

† Il a pu être vérifié expérimentalement que V'(λ) correspond à la courbe d'absorption spectrale de la rhodopsine, l'un des éléments biologiques essentiels à la vision.

Une égalisation hétérochrome en luminance visuelle (mais pas nécessairement en couleur) pour l'observateur photopique standard entre deux distributions de puissances spectrales $\{P_\lambda d\lambda\}$ et $\{P'_\lambda d\lambda\}$ sera réalisée lorsque :

$$\int_\vartheta P_\lambda V(\lambda) d\lambda = \int_\vartheta P'_\lambda V(\lambda) d\lambda$$

On remarquera que si la loi d'additivité est strictement vérifiée pour les grandeurs énergétiques, elle n'est pas parfaitement suivie par le système visuel humain. De nombreuses expériences montrant la faillite de la loi d'additivité ont été réalisées (voir [Stil et al 82]), l'une d'entre elles a été réalisée par Guth en 1970 [Guth 70]. L'expérience consistait à présenter à un observateur une surface divisée en deux : l'une des moitiés provoque un stimulus W de couleur fixe "blanche" (2800 K) pendant que l'autre provoque un stimulus C(λ) monochromatique. L'observateur doit faire varier la luminance énergétique du stimulus chromatique jusqu'à ce qu'il juge les deux moitiés également lumineuses. Lorsque cela est réalisé, la moitié de la puissance du rayonnement W est ajoutée au rayonnement chromatique et l'observateur cherche à égaliser ce nouveau stimulus sur la moitié du champ avec l'autre moitié produisant toujours le stimulus W. Il cherche donc à modifier C(λ) d'un coefficient α afin que

$$\alpha C(\lambda) + 0,5W \stackrel{L}{=} W$$

Si la loi d'additivité était strictement vérifiée α devrait être sensiblement égal à 0,5 Les résultats obtenus sur neuf observateurs sont repris dans la figure ci-dessous.

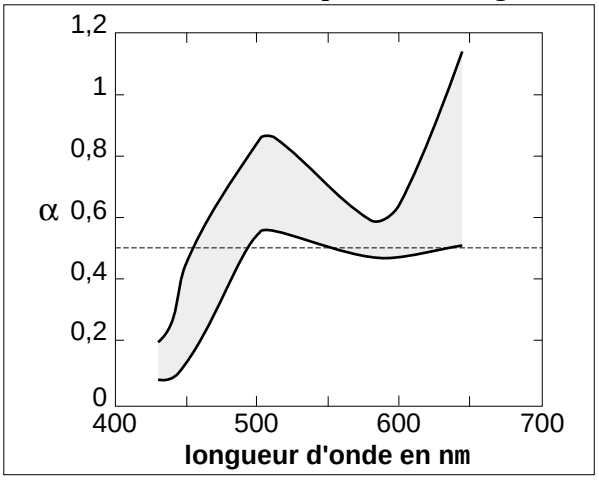


Figure 21: La zone grisée représente les variations de α tel que donné par les neuf observateurs. La ligne discontinue indique α=0,5, la zone grisée devrait coïncider avec cette ligne si la loi d'additivité était valide.

Les résultats obtenus sont loin de ceux attendus... La loi d'additivité est néanmoins admise pour les grandeurs visuelles. L'ensemble des lois ci-dessus sera intensivement utilisé par la suite mais on gardera à l'esprit que, compte tenu de la remarque précédente, il est préférable, sinon indispensable, de manipuler le plus longtemps possible des grandeurs énergétiques avant de passer aux grandeurs visuelles, sous peine d'accumuler des erreurs peu maîtrisables.

I.5.3 Unités de mesures

Reste à définir une unité pour les grandeurs photométriques. Du choix de ces unités dépendra la valeur du coefficient K_m d'efficacité lumineuse photopique (ou scotopique K'_m). Il est à noter que le choix de ces unités dépend de considérations historiques aussi bien que pratiques. Historiques, car elles ont d'abord été établies empiriquement avant que leurs liens ne soient établis avec les grandeurs énergétiques. Pratiques ensuite, car l'unité implique la réalisation d'étalons stables et aussi exactement reproductibles que possible. Ceci explique que les valeurs aient évolué de façon sensible. La définition la plus actuelle de la *candela*, adoptée en 1979 par la CGPM†, est la suivante :

> *La candela (cd) est l'intensité lumineuse dans une direction donnée, d'une source qui émet un rayonnement monochromatique de fréquence 540.10^{12} Hz et dont l'intensité dans cette direction est 1/683 $W.sr^{-1}$.*

Les définitions des quantités radiométriques décrites précédemment sont également utilisables pour les grandeurs photométriques. On retrouvera ainsi l'éclairement visuel (ou lumineux), l'intensité visuelle, etc. Chacune de ces quantités a ses unités propres, toutes définissables à partir de la candela, on trouve ainsi le lumen (lm) pour le flux visuel (ϕ_e) et le *lux* (lx) pour l'éclairement visuel :

$$I_v = \frac{d\phi_v}{d\Omega} \qquad \text{cd}$$

$$E_v = \frac{d\phi_v}{dA} \qquad \text{lm.m}^{-2} \equiv \text{lx}$$

$$L_v = \frac{dE_v}{d\Omega_r} \qquad \text{cd.m}^{-2}$$

Le *lumen* est le flux visuel émis dans un angle solide unité par une source ponctuelle uniforme ayant une intensité lumineuse de 1 candela.

Par définition l'efficacité lumineuse photopique d'un flux énergétique de fréquence 540.10^{12} Hz, c'est à dire à une longueur d'onde de 555,016 nm dans l'air standard, est de 683 $lm.W^{-1}$, d'où la valeur de K_m = 683 $lm.W^{-1}$. En fait, V(555,016) a pour valeur 0,999998 alors que l'on devrait avoir tout juste 1, ceci conduit à K_m = 683,002 $lm.W^{-1}$. Pratiquement cependant K_m est arrondi à 683 sans conséquence dans les mesures courantes. Le coefficient d'efficacité lumineuse scotopique K'_m est arrondi pour des raisons identiques à 1700 $lm.W^{-1}$.

† Conférence Générale des Poids et Mesures

A partir de ces unités on peut passer des grandeurs énergétiques aux grandeurs visuelles [Stil et al 82] :

$$\phi_v = K_m \int_{\vartheta} \phi_{e,\lambda} V(\lambda) d\lambda$$

$$L_v = K_m \int_{\vartheta} L_{e,\lambda} V(\lambda) d\lambda$$

$$E_v = K_m \int_{\vartheta} E_{e,\lambda} V(\lambda) d\lambda$$

I.6 Colorimétrie

La photométrie avait à l'origine pour but de mesurer l'intensité des sources lumineuses; elle a nécessairement conduit à introduire le concept de couleur et cherché le moyen de le quantifier objectivement. C'est donc tout naturellement qu'elle a donné naissance à la colorimétrie dont c'est l'objet. Ainsi de nouvelles lois doivent être introduites pour rendre compte de la couleur. L'*égalisation couleur* doit suivre également les lois de symétrie, de transitivité, de proportionnalité et d'additivité. Elle stipule de plus que tout stimulus provoqué par un rayonnement peut être produit par une combinaison linéaire de trois *stimuli primaires*.

Ces primaires peuvent être choisies selon différents objectifs. En 1931, la CIE a adopté le système **XYZ** qui fixe les primaires **X**, **Y** et **Z**. A partir de ces stimuli primaires, trois fonctions d'égalisation couleur sont définies : $\bar{x}(\lambda)$, $\bar{y}(\lambda)$, $\bar{z}(\lambda)$ à partir desquelles l'égalisation de deux puissances spectrales $\{P_{1\lambda}\, d\lambda\}$ et $\{P_{2\lambda}\, d\lambda\}$ s'exprime par [Stil et al 82] :

$$\int_{\vartheta} P_{1\lambda} \bar{x}(\lambda) d\lambda = \int_{\vartheta} P_{2\lambda} \bar{x}(\lambda) d\lambda$$

$$\int_{\vartheta} P_{1\lambda} \bar{y}(\lambda) d\lambda = \int_{\vartheta} P_{2\lambda} \bar{y}(\lambda) d\lambda$$

$$\int_{\vartheta} P_{1\lambda} \bar{z}(\lambda) d\lambda = \int_{\vartheta} P_{2\lambda} \bar{z}(\lambda) d\lambda$$

Les deux stimuli couleurs sont alors dits *métamères*.

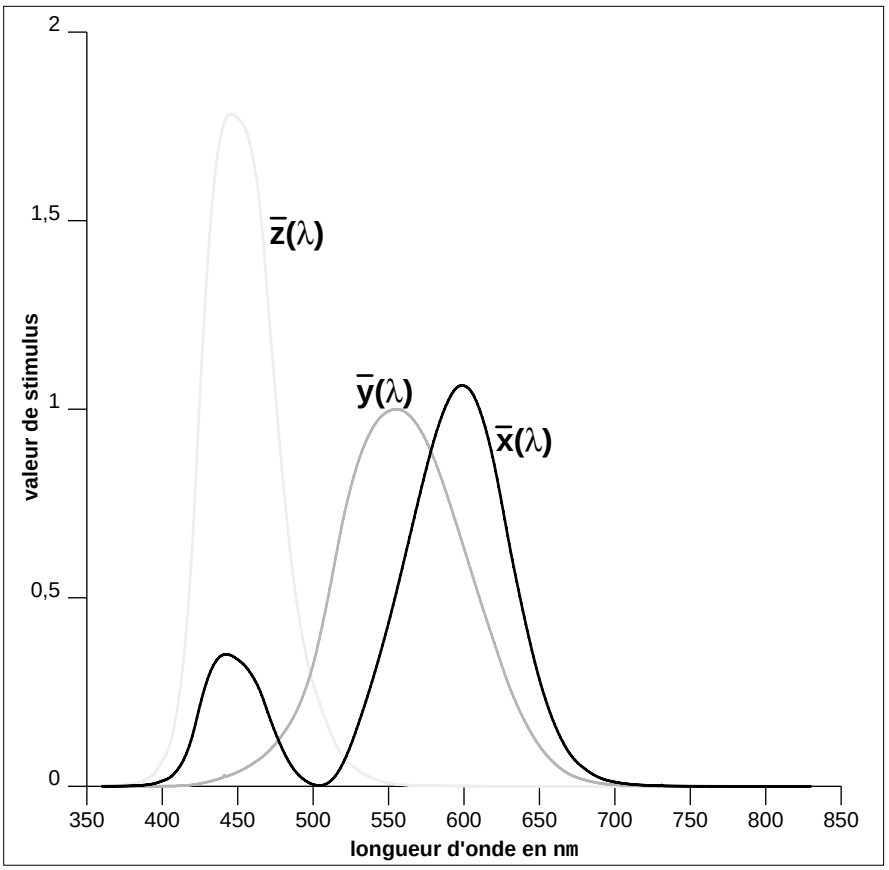


Figure 22: Observateur standard CIE 1931

Les trois courbes d'égalisation couleur (figure ci-dessus) définissent l'observateur standard 1931 dont on a parlé précédemment. Le système précédent de la CIE, **RGB**, était basé sur d'autres primaires mais avait le désavantage que la courbe $\bar{r}(\lambda)$, d'égalisation couleur du rouge, avait des parties négatives, ce qui obligeait à quelques gymnastiques pour l'intégration. D'autre part, le passage à la luminance lumineuse exigeait une autre intégration avec $V(\lambda)$ ou une combinaison linéaire des trois courbes. Les courbes $\bar{x}(\lambda)$, $\bar{y}(\lambda)$, $\bar{z}(\lambda)$ sont toutes positives. On peut également noter que ces primaires ne sont pas physiquement réalisables et que l'ensemble de leurs combinaisons ne donne pas systématiquement de stimuli réels, en d'autres termes une couleur observable.

L'autre avantage du système **XYZ** est qu'il est directement lié à la courbe de sensibilité $V(\lambda)$ car par définition :

$$\bar{y}(\lambda) \equiv V(\lambda)$$

Le deuxième observateur, l'observateur standard CIE 1964, a été défini par trois autres courbes, $\bar{x}_{10}(\lambda)$, $\bar{y}_{10}(\lambda)$, $\bar{z}_{10}(\lambda)$ pour tenir compte de champs visuels de plus de 4°.

A partir du système **XYZ**, la CIE définit un *système trichromatique* qui permet de représenter une couleur dans un espace à trois dimensions par ses coordonnées (X, Y, Z) :

$$X = k\int_{\vartheta} P_\lambda \bar{x}(\lambda)d\lambda$$

$$Y = k\int_{\vartheta} P_\lambda \bar{y}(\lambda)d\lambda$$

$$Z = k\int_{\vartheta} P_\lambda \bar{z}(\lambda)d\lambda$$

k étant un facteur de normalisation, qui, lorsqu'il est égal à K_m , permet d'interpréter Y comme une luminance visuelle ($lm.sr^{-1}.m^{-2}$) si P_λ est une luminance énergétique ($W.sr^{-1}.m^{-2}$).

Lorsque P_λ désigne le rayonnement d'un objet éclairé par une source définie par $S(\lambda)$ et une réflectance spectrale lambertienne $\beta(\lambda)$, la CIE définit la couleur de l'objet par :

$$X = k\int_{\vartheta} \beta(\lambda)S(\lambda)\bar{x}(\lambda)d\lambda$$

$$Y = k\int_{\vartheta} \beta(\lambda)S(\lambda)\bar{y}(\lambda)d\lambda$$

$$Z = k\int_{\vartheta} \beta(\lambda)S(\lambda)\bar{z}(\lambda)d\lambda$$

où $k = 100 / \int_{\vartheta} S(\lambda)\bar{y}(\lambda)d\lambda$

de telle sorte que Y désigne le *facteur de luminance* et soit égal à 100 dans le cas du diffuseur parfait, ($\beta(\lambda) = 1, \forall\lambda$). Ces recommandations de la CIE n'ont pas valeurs de référence absolue, k peut très bien être choisi de telle sorte que Y soit, par exemple, égal à 1 pour le diffuseur parfait, pourvu que ce facteur soit clairement spécifié et constant pendant toute la durée des conversions.

A partir des coordonnées (X, Y, Z), une autre représentation a été introduite, les coordonnées de *chromaticité* : x, y, z définies par :

$$x = \frac{X}{X+Y+Z} \qquad y = \frac{Y}{X+Y+Z} \qquad z = \frac{Z}{X+Y+Z}$$

Les coordonnées x et y sont souvent utilisées en conjonction avec Y pour spécifier une couleur dans l'industrie. Le couple (x, y) désigne en effet la chromaticité de la couleur, en d'autres termes la teinte, alors que Y désigne la «luminosité» de celle-ci.

Il est tentant d'associer une distance euclidienne à l'espace **XYZ** mais, malheureusement, si cette distance a un sens mathématique, elle n'est pas perceptuellement uniforme. Autrement dit, la même distance entre deux couleurs dans le système XYZ ne représente pas le même écart perceptuel selon l'endroit de l'espace où elle se situe. Depuis cinquante ans, de nombreuses tentatives pour trouver un espace de couleur uniforme ont été effectuées. Aucune n'a abouti à ce jour. La CIE a néanmoins fait des recommandations au sujet de deux espaces de couleurs basés sur le système **XYZ**, qui, sans être perceptuellement uniformes, donnent une meilleures représentation des écarts visuels.

L'espace L*a*b* date de 1976 et définit le système CIELAB comme suit :

- Si $X/X_n > 0{,}008856$, $Y/Y_n > 0{,}008856$ et $Z/Z_n > 0{,}008856$

$$\begin{aligned} L^* &= 116\sqrt[3]{\frac{Y}{Y_n}} - 16 \\ a^* &= 500\left(\sqrt[3]{\frac{X}{X_n}} - \sqrt[3]{\frac{Y}{Y_n}}\right) \\ b^* &= 200\left(\sqrt[3]{\frac{Y}{Y_n}} - \sqrt[3]{\frac{Z}{Z_n}}\right) \end{aligned} \qquad (23)$$

où (X_n, Y_n, Z_n) désigne les coordonnées d'un blanc pris comme référence pour un illuminant donné de telle sorte que $Y_n = 100$ lorsque cet illuminant éclaire le diffuseur idéal.

- Dans le cas contraire

$$L^* = 903{,}3\,Y/Y_n$$

et il faut remplacer chaque racine cubique par $7{,}787(C/C_n) - 16/116$ dans la formule (23) si $C/C_n \leq 0{,}008856$, C désignant indifféremment X, Y ou Z.

Une métrique est associée à cet espace afin de fournir une mesure approximative de l'écart perceptuel ΔE^*_{ab} entre deux couleurs (L_1^*, a_1^*, b_1^*) et (L_2^*, a_2^*, b_2^*) :

$$\Delta E^*_{ab} = \sqrt{\Delta L^{*2} + \Delta a^{*2} + \Delta b^{*2}}$$

avec $\Delta C = C_1^* - C_2^*$, $C \in \{L, a, b\}$

Datant de 1976 également, l'espace L*u*v* a été proposé par la CIE en même temps que L*a*b*, sans qu'elle ne recommande l'utilisation de l'un plus que l'autre tant tous deux sont imparfaits. Le système CIELUV est défini par :

- Si $Y/Y_n > 0{,}008856$

$$\begin{aligned} L^* &= 116\sqrt[3]{\frac{Y}{Y_n}} - 16 \\ u^* &= 13L^*(u' - u_n') \\ v^* &= 13L^*(v' - v_n') \end{aligned} \qquad (24)$$

$$u' = \frac{4X}{X + 15Y + Z} \qquad v' = \frac{9X}{X + 15Y + Z}$$

$$u_n' = \frac{4X_n}{X_n + 15Y_n + Z_n} \qquad v_n' = \frac{9X_n}{X_n + 15Y_n + Z_n}$$

- Si $Y/Y_n \leq 0{,}008856$

$$L^* = 903{,}3Y / Y_n$$

La différence de couleur associée est la suivante :

$$\Delta E_{uv}^* = \sqrt{\Delta L^{*2} + \Delta u^{*2} + \Delta v^{*2}} \qquad (25)$$

Nous la noterons plus simplement Δ_{uv} dans la suite.

La CIE a fait un certains nombre de recommandations relatives aux conditions d'étude des deux systèmes de couleurs [Kowa 80] :

- Angle de vision supérieur ou égal à 4°.
- Champ environnant uniforme
- Luminance du champ environnant de 100 à 1000 cd/m^2
- Chromaticité du champ environnant équivalent à l'illuminant D entre 5 500 et 7 500 K.
- Luminance de l'échantillon de 5 à 500 % de la luminance du champ environnant.

La distance à partir de laquelle deux couleurs ne peuvent être confondues ne semble pas avoir été clairement spécifiée, les documents que nous avons pu consulter n'en font pas mention. Une allusion dans *Color Science* (p.171) indique qu'au delà de 2 ou 3, la différence est vraiment sensible pour un même objet vu sous des illuminants différents. Une conversation avec Mme F. Viénot de la CIE a néanmoins confirmé qu'un écart inférieur ou égal à une unité, que ce soit en L*a*b* ou en L*u*v*, n'est pas discernable.

Dans les années 70 on dénombrait pas moins de vingt formules différentes couramment utilisées pour quantifier des écarts de couleurs [Robe 77]. Parmi les meilleures, on peut citer celles de Adams-Nickerson [Mcla 70], Sauderson-Milner [Saul et al 46], Scofield [Scof 43] et L*u*v*. Les études menées sur ces différentes formules sont nombreuses. On peut citer celles de Morley et al [Morl et al 75] qui a porté sur la comparaison de 555 paires d'échantillons entre l'écart visuel jugé par un observateur et celui calculé par les différentes formules. Les coefficients de corrélation entre les deux méthodes sont reportés dans le tableau ci-dessous :

Formule	Coefficient de corrélation
Adams-Nickerson	0,72
Sauderson-Milner	0,72
Scofield	0,72
CIELAB	0,72
CIELUV	0,71

Tableau 2: Coefficient de corrélation entre les jugements visuels et les écarts calculés par les différentes formules

Au vu de ces résultats, aucune formule n'est significativement meilleure que les autres et n'offre d'accord très précis avec le jugement visuel.

II

Modèles d'éclairement

Après avoir passé en revue les lois physiques présidant à l'aspect du monde, nous nous intéressons ici aux modèles utilisés couramment en synthèse d'image et s'inspirant peu ou prou de celles-ci. Nous verrons qu'il y a plusieurs approches possibles lorsque l'on se pose la question de la visualisation de scènes synthétiques. Nous serons amené à en retenir certaines au détriment d'autres ce qui nous conduira à la proposition d'un modèle, qui, sans être idéal, s'approche de l'idée que nous nous faisons d'une synthèse que nous qualifierons de pseudo-réaliste.

II.1 Modèles d'éclairements locaux

II.1.1 Premiers modèles

L'un des premiers modèles d'éclairements utilisé en image de synthèse fut introduit par Bui Tuong Phong en 75 [Phon 75]. Le but principal de son modèle était de tenir compte des effets visuels engendrés par les reflets directs d'une source sur un objet afin de dépasser l'aspect diffus auquel l'infographie de l'époque était limitée. Bui Tuong Phong propose donc une expression améliorée de $I_r(\lambda)$, l'«intensité†» d'une source ponctuelle à l'infini, réfléchie en un point d'une surface plane :

$$I_r(\lambda) = k_a C_r(\lambda) + (1 - k_a) C_r(\lambda) \max(0, \cos\zeta_L) + W(\zeta_L) \cos^n(\angle \mathbf{V}, \mathbf{R}_L)$$

$C_r(\lambda)$: cœfficient de réflexion au point d'intersection de la surface
k_a : cœfficient environnemental de réflexion diffuse
$W(\zeta_L)$: rapport de la lumière spéculairement réfléchie sur l'incidente (de ~ 10 à 80%)
ζ_L : angle d'incidence du rayon issu de la source
n : facteur modélisant la spécularité de la surface (de ~ 1 à 10)
$\mathbf{V}$: le vecteur formé par l'œil et le point d'intersection
$\mathbf{R}_L$: le vecteur réfléchi de la source L en considérant la surface comme spéculaire

† Un mot très usité, mais rarement dans son sens physique, désignant plutôt la valeur d'une composante RVB.

Le cosinus trouve sa justification dans la constatation empirique que si l'œil se trouvait dans la direction de réflexion spéculaire de la source sur la surface, la lumière renvoyée aurait une intensité maximale si la surface était parfaitement réfléchissante. De fait, plus l'œil s'éloigne de cette position, moins il reçoit de lumière. L'élévation à la puissance accélère cette décroissance au fur et à mesure de l'éloignement de la position de reflet maximum. L'effet obtenu en modulant ce paramètre est la variation de taille de la tache de reflet spéculaire. Si les objets de la scène sont suffisamment diffus, la variation de ce paramètre d'un objet à l'autre ne choque en général pas notre sens visuel. En revanche, des objets spéculaires ayant des valeurs du paramètre très différentes, donnent l'impression que chaque objet n'est pas éclairé par les mêmes sources : les taches spéculaires sont davantage interprétées comme des tailles de sources différentes que comme des effets de matière. Le cœfficient de réflexion environnemental, selon la terminologie de Bui Tuong Phong, sert essentiellement à donner de la couleur aux parties non éclairées des objets qui seraient totalement noires sans lui. Son inconvénient majeur, outre que l'éclairement ambiant n'est en général pas constant sur l'ensemble d'une scène, est que, passée la séparation entre illumination et ombrage, la couleur est constante quelle que soit la courbure de l'objet.

Partant de ce modèle, Blinn en développa un nouveau en 77 en s'inspirant davantage de résultats physiques [Blin 77]. C'est le premier à avoir proposé le modèle à facettes dans une décomposition désormais célèbre que l'on pourrait appeler *DGF* pour Distribution/Géométrie/Fresnel. Son modèle comporte toujours trois parties distinctes, à savoir ambiante, diffuse et spéculaire, dont la somme donne l'intensité réfléchie :

$$I_r(C) = C(k_a + k_d p_d) + k_s p_s$$

C : une des trois composantes RVB de l'objet
k_a : proportion de lumière ambiante
$k_d = \max(0, \cos\zeta_L)$
p_d : proportion de réflexion diffuse
p_s : proportion de réflexion spéculaire ($p_d + p_s = 1$)

Il proposa un certains nombre d'expressions pour k_s, inspirées, l'une de Bui Tuong Phong, les autres de travaux de physiciens. La formulation générale qu'il donne de ce facteur suppose un modèle à micro-surfaces (facettes ou cuvettes) dont chacune aurait un comportement spéculaire :

$$k_s = \frac{DFG}{\cos\zeta_{\mathbf{V}}}$$

$D(\zeta_{\mathbf{n}})$ représente la proportion de micro-surfaces faisant un angle $\zeta_{\mathbf{n}}$ avec la normale moyenne $\mathbf{N}$
G est la fonction d'illumination des facettes résultant de leur auto-ombrage. Partant de l'expression qu'en donnent Torrance et Sparrow, Blinn la trouve égale à $\min(2\cos\zeta_{\mathbf{n}}\cos\zeta_{\mathbf{L}} / \cos\psi_{\mathbf{n}}, 1, 2\cos\zeta_{\mathbf{n}}\cos\zeta_{\mathbf{V}} / \cos\psi_{\mathbf{n}})$.
$\zeta_{\mathbf{V}}$, l'angle entre l'œil et $\mathbf{N}$, $\psi_{\mathbf{n}}$, l'angle bissecteur entre $\mathbf{L}$ et $\mathbf{V}$, plus souvent rencontré sous la forme $\angle(\mathbf{H}, \mathbf{L}) = \angle(\mathbf{H}, \mathbf{V})$, $\mathbf{H} = \mathbf{L} + \mathbf{V}$
$F(\psi_{\mathbf{n}})$ est le cœfficient de Fresnel pour une lumière non-polarisée

Blinn propose trois fonctions de distribution des micro-surfaces :

- à partir de Bui Tuong Phong : $D(\zeta_{\mathbf{n}}) = \cos^n\zeta_{\mathbf{n}}$, une expression bien différente de celle de son inspirateur et fournissant d'étranges résultats puisqu'elle est invariante par rotation de l'observateur autour de **V** (voir [Fish et al 94] pour une discussion plus détaillée)
- à partir de Torrance et Sparrow [Torr et al 67] :

$$D(\zeta_{\mathbf{n}}) = e^{-(\zeta_{\mathbf{n}}\sigma)^2}$$

- ou encore de Trowbrige et Reitz [Trow et al 75] :

$$D(\zeta_{\mathbf{n}}) = (\varepsilon^2 / [1 + (\varepsilon^2 - 1)\cos^2\zeta_{\mathbf{n}}])^2$$

On notera que l'auto-ombrage n'est cohérent qu'avec la distribution de Torrance-Sparrow et que, quelle que soit l'expression de k_s, elle n'est pas normalisée, ce qui peut conduire à réfléchir plus d'«énergie» qu'il n'y en a d'incidente. Ceci provient entre autre du fait qu'aucune expression de D n'est une distribution au sens statistique du terme.

Il faudra attendre trois ans de plus pour qu'un modèle tenant compte des réflexions secondaires et des réfractions, apparaisse en 80 avec la technique du *tracé de rayons* de Whitted [Whit 80]. Aux composantes ambiantes et diffuses, il en adjoint deux autres, représentant les contributions spéculairement réfléchies et transmises menant à l'observateur :

$$I_r = I_a + k_s S + T k_t + k_d \sum_{j=1}^{n_l} \cos\zeta_{\mathbf{L}_j}$$

k_s, k_d et k_t sont respectivement les cœfficients de diffusion, de réflexion spéculaire et de transmission spéculaire
n_l est le nombre de sources lumineuses ponctuelles de la scène
S est l'intensité provenant de la direction $\mathbf{R_V}$ de réflexion spéculaire par rapport à l'observateur.
T est l'intensité provenant de la direction $\mathbf{T_V}$ de transmission spéculaire par rapport à l'observateur

Whitted propose de perturber la normale au point d'intersection afin de simuler une réflexion non parfaitement spéculaire. Néanmoins, comme cette technique est assez coûteuse, il l'utilise avec parcimonie et préconise l'emploi du modèle de Bui Tuong Phong modifié par Blinn pour les reflets spéculaires directement générés par une source.

II.1.2 Inspiration physique

Un an plus tard, Cook publie sa thèse† [Cook 81] puis deux articles avec Torrance [Cook et al 81] et [Cook et al 82] où le modèle DGF de Blinn est adapté aux résultats de Torrance et Sparrow. A partir de la définition de la FDRB et des travaux de ces physiciens, on a :

$$f_r = \frac{dL_r}{dE_i} = \frac{dL_r}{d\phi_i / dA} = \frac{dL_r}{d\omega_{\mathbf{L}} L_i \cos\zeta_{\mathbf{L}}}$$

† Master's thesis

La FDRB de la luminance spéculairement réfléchie par les facettes bien orientées peut être obtenue avec (6) §I.2.2 :

$$f_{rs} d\omega_{\mathbf{L}} L_i \cos\zeta_{\mathbf{L}} = dL_{rs} = \frac{G(\zeta_{ip}, \zeta_{rp}) R(\psi_{\mathbf{n}}) d\omega_{\mathbf{L}} L_i p(\zeta_{\mathbf{n}}) a}{4\cos\zeta_{\mathbf{V}}}$$

d'où l'on tire :

$$f_{rs} = \frac{G(\zeta_{ip}, \zeta_{rp}) R(\psi_{\mathbf{n}}) p(\zeta_{\mathbf{n}}) a}{4\cos\zeta_{\mathbf{V}} \cos\zeta_{\mathbf{L}}} \tag{26}$$

Une expression que Cook généralise un peu rapidement sous la forme :

$$f_{rs} = \frac{FDG}{\pi \cos\zeta_{\mathbf{V}} \cos\zeta_{\mathbf{L}}} \tag{27}$$

D désigne ici une distribution des pentes de la surface dont Cook propose plusieurs expressions :

$$D_1 = c e^{-(\zeta_{\mathbf{n}}/m)^2}$$

$$D_2 = \left(e^{(\tan\zeta_{\mathbf{n}}/m)^2} m^2 \cos^4\zeta_{\mathbf{n}} \right)^{-1}$$

$$D_3 = \sum_j w_j D(m_j)$$

c est une constante «arbitraire»
m : écart-type des pentes
w_j : poids $\sum w_j = 1$

La transition de la distribution des hauteurs à celle des pentes ne pose pas de problèmes mathématiques particuliers mais plutôt physiques. Comme le signalent Bennett et Mattson dans leur panorama déjà cité [Benn et al 89], la mesure de la distribution des pentes est beaucoup plus sensible au type de profilomètre utilisé que ne l'est celle de la distribution des hauteurs. La mesure de m peut varier en effet dans un rapport de 1 à 50.

On notera que D_1 n'est une distribution que par un choix approprié de c. En revanche D_2, présentée comme issue des travaux de Beckmann, n'est pas une distribution et encore moins celle de Beckmann. Ce dernier utilise une distribution des hauteurs gaussienne comme nous l'avons vu, si h(x) est la hauteur au point x, sa dérivée désigne la pente en ce point et sa fonction de corrélation s'exprime par :

$$B(\tau) = \sigma^2 e^{-\tau^2/T^2}$$

Or la distribution de la dérivée d'un processus gaussien est, elle aussi, gaussienne de moyenne nulle et de variance m^2 :

$$m^2 = -B''(0)$$

$$B''(\tau) = 2\sigma^2(2\tau^2 - T^2)T^{-4} e^{-\tau^2/T^2} \quad \Rightarrow \quad m^2 = 2\sigma^2 T^{-2}$$

La densité de h' s'exprime donc par :

$$P(h' < u) = \int_{-\infty}^{u} (m\sqrt{2\pi})^{-1} e^{-z^2/(2m^2)} dz$$

en faisant le changement de variable $\tan\zeta_{\mathbf{n}} = z$ on obtient la distribution recherchée :

$$P(h' < u) = \int_{-\pi/2}^{\arctan u} \left(m\sqrt{2\pi} e^{\tan^2 \zeta_n / (2m^2)} \cos^2 \zeta_n \right)^{-1} d\zeta_n = \int_{-\pi/2}^{\arctan u} D(\zeta_n) d\zeta_n \qquad (28)$$

La présence de π dans (27) ne s'explique que si F est la réflectance hémisphérique à incidence normale et non le facteur de Fresnel. Enfin, on ne peut que constater les différences entre l'expression (26) et celle de Cook.

A notre connaissance, aucun modèle à facettes n'applique le raisonnement réalisé sur la réflexion spéculaire à la transmission spéculaire. Tous ces modèles peuvent être étendus pour inclure les contributions secondaires à la manière de Whitted. L'expression générale est une somme de composantes ambiante, diffuse, spéculaire directe, spéculaire réfléchie et spéculaire transmise.

En 91, He et al. ont proposé un modèle d'éclairement [He et al 91] basé sur les développements de Beckmann, de Stogryn et de Smith déjà cités. Les résultats de Beckmann sont atténués par une adaptation des fonctions d'illumination de Smith tandis que la démarche de Stogryn est utilisée pour tenir compte des effets de polarisation. Deux formulations sont proposées selon que le rayonnement incident est polarisé ou non. En l'absence de polarisation l'expression de la FDRB est la suivante :

$$f_r = \frac{\rho_s}{\cos\zeta_i d\omega_i} \mathbf{1}_{d\omega_{sp}}(\zeta_r) + |R|^2 \frac{GDS}{\cos\zeta_i \cos\zeta_r} + a(\lambda)$$

$$\rho_s = S|R|^2 e^{-g}$$

$$|R|^2 = (R_\perp^2(\psi_n) + R_\parallel^2(\psi_n))/2$$

$$G = \frac{v^4}{v_z^2} \frac{((s_r \cdot q_i)^2 + (p_r \cdot q_i)^2)((s_i \cdot q_r)^2 + (p_i \cdot q_r)^2)}{|q_r \times q_i|^4}$$

$$S = S(\zeta_i) S(\zeta_r)$$

$$S(\zeta) = \frac{1 - \mathrm{erfc}(T \cot\zeta / (2\sigma_p))/2}{\Lambda(\cot\zeta) + 1}$$

$$D = \frac{\pi T^2}{4\lambda^2} \sum_{m=1}^{\infty} \frac{g^m e^{-g}}{m! m} e^{-v_{xy}^2 T^2/(4m)}$$

$$g = v_z^2 \sigma^2$$

$$\sigma = \sigma_o (1 + (z_o/\sigma_o)^2)^{-1/2}$$

$$z_o = \frac{\sigma_o (K_i + K_r)\sqrt{2}}{4\sqrt{\pi}} e^{-2(z_o/(2\sigma_o))^2}$$

$$K_d = \mathrm{erfc}\left(\frac{T}{2\sigma_o \tan\zeta_d} \right) \tan\zeta_d \qquad d \in \{i, r\}$$

s_r, p_r sont les vecteurs unitaires de polarisation respectivement perpendiculaire et parallèle

En l'absence de la thèse de He, à laquelle nous n'avons pu avoir accès, nous ne soulignerons donc que quelques points qui semblent incorrects. La fonction d'auto-ombrage, basée sur les travaux de Smith, fournit une approximation de la proba-

bilité d'illumination $S(\zeta_i)$ d'un point d'une surface gaussienne pour un angle d'incidence donné. He étend ces résultats au calcul de la probabilité qu'un rayon incident soit réfléchi dans une direction donnée par la simple loi conjointe $S(\zeta_i)S(\zeta_r)$. En clair, He considère comme indépendants l'un de l'autre, l'événement «le rayon incident n'est pas masqué» et l'événement «le rayon réfléchi n'est pas masqué», ce qui n'est pas correct. Ces deux événements sont liés et la manière correcte de les traiter serait d'évaluer la probabilité de l'événement : «le rayon réfléchi n'est pas masqué *sachant* que le rayon incident ne l'est pas non plus»; ce qui est d'une difficulté beaucoup plus grande.

Une autre erreur statistique du modèle provient de l'expression de σ issue des travaux de Beckmann [Beck 65] déjà cités (§I.2.4). Cette expression est déduite de l'expression de la probabilité qu'un point h(0) d'une surface ne soit pas ombré dans un intervalle [t, t + dt] sachant qu'il n'est pas ombré en h(t). Alors que la probabilité réellement recherchée est la probabilité qu'un point d'une surface ne soit pas ombré dans un intervalle [t, t + dt] sachant qu'il n'est ombré par *aucun point* de 0 à t. Cette erreur de raisonnement ainsi que d'autres ont fait l'objet de commentaires par Shaw [Shaw 66] peu après la parution de l'article de Beckmann.

Enfin, rappelons la démarche de Stogryn qui établit γ_{ab}, le coefficient de dispersion différentiel par unité d'aire :

$$\gamma_{ab}(\mathbf{q}_i, \mathbf{q}_r, \Sigma) = \lim_{A_p \to \infty} \lim_{d \to \infty} \frac{4\pi d^2 I_r}{A_p \cos\zeta_i I_i}$$

où les indices a et b désignent respectivement la polarisation incidente et celle réfléchie

définition générale que l'on peut rendre explicite en fonction du champ électrique pour chaque polarisation réfléchie :

$$\gamma_{ah}(\mathbf{q}_i, \mathbf{q}_r, \Sigma) = \lim_{A_p \to \infty} \lim_{d \to \infty} \frac{4\pi d^2}{A_p \cos\zeta_i} \frac{|\mathbf{h} \cdot \mathbf{E}_r|^2}{E_{oi}^2}$$

$$= \lim_{A_p \to \infty} \frac{1}{4\pi A_p \cos\zeta_i} \left| E_{oi}^{-1} \int_\sigma e^{-i(\mathbf{q}_r \cdot \mathbf{r})} (iq\mathbf{v}_r \cdot (\mathbf{E}_r \times \mathbf{n}) + \mathbf{h}_r \cdot ((\nabla \times \mathbf{E}_r) \times \mathbf{n})) d\sigma \right|^2$$

$$\gamma_{av}(\mathbf{q}_i, \mathbf{q}_r, \Sigma) = \lim_{A_p \to \infty} \lim_{d \to \infty} \frac{4\pi d^2}{A_p \cos\zeta_i} \frac{|\mathbf{v} \cdot \mathbf{E}_r|^2}{E_{oi}^2}$$

$$= \lim_{A_p \to \infty} \frac{1}{4\pi A_p \cos\zeta_i} \left| E_{oi}^{-1} \int_\sigma e^{-i(\mathbf{q}_r \cdot \mathbf{r})} (iq\mathbf{h}_r \cdot (\mathbf{E}_r \times \mathbf{n}) + \mathbf{v}_r \cdot ((\nabla \times \mathbf{E}_r) \times \mathbf{n})) d\sigma \right|^2$$

h, **v** vecteurs unitaires de polarisation horizontale ($\perp$) ou verticale ($\parallel$) respectivement

Dans ces équations cependant, le champ réfléchi $\mathbf{E}_r$ dépend toujours implicitement de la polarisation du champ incident. La notation retenue par Stogryn et reprise ci-dessus est d'ailleurs source de confusion. Pour pouvoir utiliser ces résultats, il est donc indispensa-

ble de pouvoir établir, à chaque interaction lumière/matière, les modules des composantes verticales et horizontales du champ électrique réfléchi, i.e. :

$$\begin{aligned}
\left|\mathbf{E}_{rv}\right|^2 &= \left|\mathbf{v}(\mathbf{v}\cdot\mathbf{E}_{h\to r}+\mathbf{v}\cdot\mathbf{E}_{v\to r})\right|^2 \\
&= \left|\mathbf{v}\cdot\mathbf{E}_{h\to r}\right|^2+\left|\mathbf{v}\cdot\mathbf{E}_{v\to r}\right|^2+2(\mathbf{v}\cdot\mathbf{E}_{h\to r})(\mathbf{v}\cdot\mathbf{E}_{v\to r}) \\
\left|\mathbf{E}_{rh}\right|^2 &= \left|\mathbf{h}(\mathbf{h}\cdot\mathbf{E}_{h\to r}+\mathbf{h}\cdot\mathbf{E}_{v\to r})\right|^2 \\
&= \left|\mathbf{h}\cdot\mathbf{E}_{h\to r}\right|^2+\left|\mathbf{h}\cdot\mathbf{E}_{v\to r}\right|^2+2(\mathbf{h}\cdot\mathbf{E}_{h\to r})(\mathbf{h}\cdot\mathbf{E}_{v\to r})
\end{aligned} \tag{29}$$

$\mathbf{E}_{h\to r}$ et $\mathbf{E}_{v\to r}$ désignant les portions du champ réfléchi issues de la polarisation horizontale et verticale du champs incident

Si l'on peut établir les deux premiers termes des membres droits de ces égalités à partir des coefficients de dispersion :

$$\left|\mathbf{b}\cdot\mathbf{E}_{a\to r}\right|^2 = \frac{\gamma_{ab}(\mathbf{q}_i,\mathbf{q}_r,\Sigma)A_p\cos\zeta_i\left|E_{oia}\right|^2}{4\pi d^2}$$

a étant la composante verticale ou horizontale du champ incident, b pour le champ réfléchi

en revanche, les produits vectoriels finaux ne peuvent en aucune façon être déduits de l'expression des coefficients de dispersion, puisque celle-ci est uniquement modulaire et non vectorielles comme il serait nécessaire. A fortiori, il est donc impossible d'en déduire l'intensité de chacune des composantes réfléchies puisque celle-ci est proportionnelle à leur module ($|\mathbf{E}_{rv}|^2$ et $|\mathbf{E}_{rh}|^2$) et donc toujours la somme de deux intégrales et non pas d'une seule comme He et al. le supposent dans leur définition de l'intensité :

$$dI_{rh} = (A_p\cos\zeta_i(4\pi)^2)^{-1}\left\langle\left|\int_\sigma e^{-i(\mathbf{q}_r\cdot\mathbf{r})}(iq\mathbf{v}_r\cdot(\mathbf{E}_r\times\mathbf{n})+\mathbf{h}_r\cdot((\nabla\times\mathbf{E}_r)\times\mathbf{n}))d\sigma\right|^2\right\rangle$$

$$dI_{rv} = (A_p\cos\zeta_i(4\pi)^2)^{-1}\left\langle\left|\int_\sigma e^{-i(\mathbf{q}_r\cdot\mathbf{r})}(iq\mathbf{h}_r\cdot(\mathbf{E}_r\times\mathbf{n})-\mathbf{v}_r\cdot((\nabla\times\mathbf{E}_r)\times\mathbf{n}))d\sigma\right|^2\right\rangle$$

< > désigne la moyenne sur la fonction de distribution caractérisant σ

Contrairement à la quasi-totalité des modèles précédents, He a été l'un des premiers à vouloir confronter son modèle à l'expérience. La comparaison entre les résultats théoriques et expérimentaux a été réalisée sur quatre matériaux différents : de l'aluminium, de l'oxyde de magnésium, du papier de verre ainsi que du plastique. Les échantillons sont éclairés par certaines longueurs d'onde : trois dans le visible (à 550, 500 et 460 nm), une dans l'infrarouge (à 2000 nm). La comparaison porte surtout sur la

FDRB relative, rapport de la FDRB dans une direction quelconque sur la FDRB dans la direction spéculaire. Les résultats sont donnés pour des directions dans le plan d'incidence et ne concordent guère pour le visible.

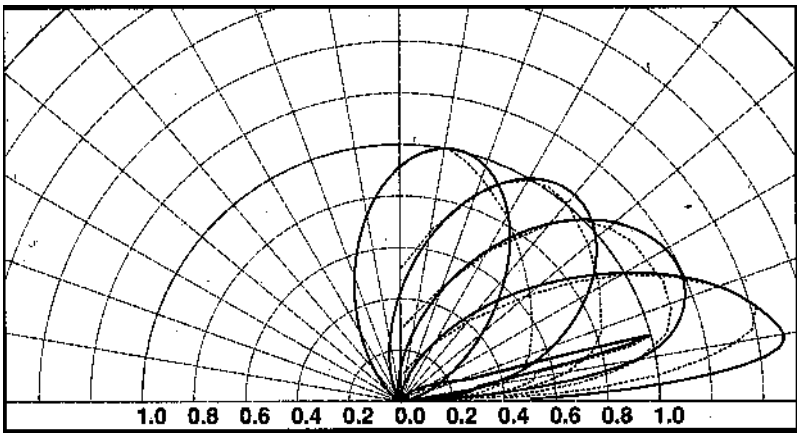


Figure 23: Exemple des résultats de He et al. pour de l'aluminium. La FDRB relative calculée (en traits pleins) comparé à celle mesurée (pointillés) pour une longueur d'onde incidente de 0,5 µm, σ_0 = 0,28 µm

La mesure de la FDRB absolue n'est fournie que dans le cas du plastique, ce qui est dommage car elle permet une comparaison plus précise. De même, on peut regretter que seuls les résultats dans le plan d'incidence soient fournis. Enfin ce type de comparaisons devrait également inclure la comparaison des distributions de pentes mesurées avec la distribution gaussienne du modèle. Une dernière remarque au sujet de curieux effets sur la photo 14 calculée avec le modèle : le reflet du cylindre sur le sol paraîtrait plus correct s'il était verticalement inversé, de même le cube semble parfaitement diffus sur la face verticale avant alors que le sommet est très réfléchissant.

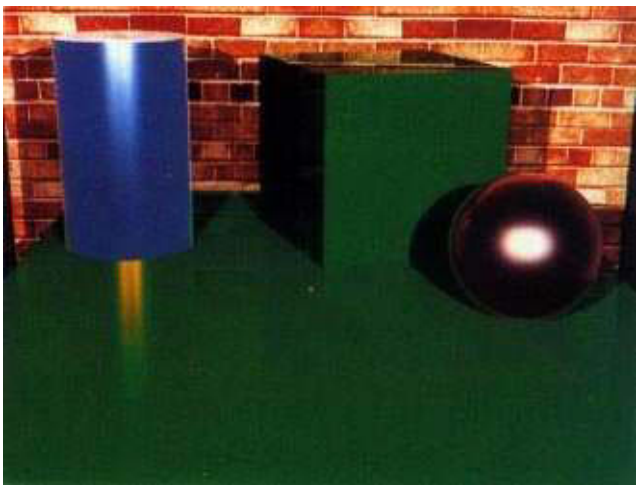


Figure 24: Photo 14 de l'article de He et al.

II.1.3 Approche calculable

Deux tendances s'affrontent en synthèse d'images quant à la proposition de nouveaux modèles. La première, illustrée par les derniers modèles que nous venons de voir, tend à considérer des modèles physiques pour les adapter à la synthèse. Devant la complexité croissante engendrée par cette approche, la deuxième tendance consiste à privilégier la rapidité de calcul sur la tentative de concordance avec les modèles de la physique.

C'est notamment le cas de Schlick qui se propose de remplacer les fonctions couramment utilisées par des développements en fractions rationnelles, équivalentes en un certain sens [Schl 92]. Cette équivalence est réalisée sur la base de valeurs particulières des fonctions à simplifier que les fractions doivent reproduire au mieux. La détermination des paramètres des fractions se fait sur une comparaison des valeurs obtenues par la vraie fonction avec celles données par la fraction pour les mêmes valeurs des paramètres. La comparaison est répétée aléatoirement un million de fois d'où est calculée une erreur relative moyenne - au passage on peut s'interroger sur cette méthode, le but final étant ni plus ni moins que de calculer une intégrale, des méthodes classiques tout à fait fiables auraient pu remplacer avantageusement cette évaluation un peu lourde. Le meilleur compromis précision/coût est retenu pour le choix final des paramètres de la fraction.

A titre d'exemple citons la recherche d'une expression simplifiée de D_2, la «distribution des pentes» que Cook attribue à Beckmann. Les propriétés remarquables de cette fonction retenues par Schlick sont, d'une part, l'invariance par rotation autour de la normale et, d'autre part, les deux valeurs particulières : $D_2(0) = 0$ et $D_2(1) = m^{-2}$. Une seule fraction rationnelle n'approximant pas suffisamment bien D_2, une fonction par morceaux a été retenue :

$$\tilde{D}_2 = 0 \qquad t \in [0, 1 - 2m^2]$$

$$\tilde{D}_2 = (t - 1 + 2m^2)^3 / (8tm^8) \qquad t \in]1 - 2m^2, 1]$$

La procédure de test donne une erreur relative moyenne de 1,8 % alors que le rapport du temps d'exécution entre D_2 et $\tilde{D}_2$ est de 29. Un gain très appréciable pour une erreur qui semble modeste.

Le point faible de la méthode nous semble cependant résider justement dans l'évaluation de l'erreur commise. L'erreur moyenne est effectivement importante mais l'écart-type l'est tout autant : une erreur moyenne faible ne préjuge pas d'un étalement étroit des valeurs autour de cette moyenne, ce que l'on recherche pour une «bonne» méthode. De même l'erreur maximale serait une indication précieuse. D'autre part, les erreurs numériques sont une chose, l'erreur perceptuelle en est une autre. D'autant plus que l'erreur numérique considérée ne renseigne pas sur le cumul des erreurs produit par la multiplication des fractions en un point d'intersection lorsque, non seulement D, mais également G et F sont approximées. Plus encore, l'erreur globale résultant des réflexions successives n'est pas considérée. L'étude du pire des cas au niveau perceptuel, ou même numérique, d'abord locale et ensuite, plus difficilement, globale serait une bonne indication de la fiabilité de la méthode. Si ce test est sans doute sévère, il faut néanmoins se rappeler la sensibilité de l'œil aux détails. Qu'un seul pixel d'une image soit franchement «défectueux» et il accrochera le regard de façon incontournable. Plus subtilement, le phénomène de bande de Mach provoqué par la discontinuité du signal sur des pixels adjacents montre combien il faut être prudent quant au choix des critères que doit satisfaire une approximation. Celui retenu par Schlick, s'il assure une certaine robustesse numérique locale, ne garantit pas la concordance finale globale au modèle de référence.

Une autre illustration de la recherche de modèles «calculables» est donnée par les développements récents de Ward [Ward 92]. Utilisant un goniophotomètre développé par le Lawrence Berkeley Laboratory, il propose un modèle empirique qu'il con-

fronte aux mesures de cet appareil. Ce modèle a deux formes selon que la surface à modéliser est ou non isotrope. La forme isotrope est la suivante :

$$\rho(\zeta_i, \alpha_i, \zeta_r, \alpha_r) = \frac{\rho_d}{\pi} + \rho_s \frac{e^{-\tan^2\zeta_n/\sigma_p^2}}{4\sigma_p^2\sqrt{\cos\zeta_i \cos\zeta_r}}$$

la forme anisotrope en est dérivée :

$$\rho(\zeta_i, \alpha_i, \zeta_r, \alpha_r) = \frac{\rho_d}{\pi} + \rho_s \frac{e^{-\tan^2\zeta_n(\cos^2\zeta_{np}/\sigma_x^2 + \sin^2\zeta_{np}/\sigma_y^2)}}{4\pi\sigma_x\sigma_y\sqrt{\cos\zeta_i \cos\zeta_r}}$$

Ward revendique une validité physique de ce modèle en arguant de sa symétrie et du respect de la conservation de l'énergie ainsi que de la comparaison aux mesures expérimentales. Les deux premières qualités sont malheureusement insuffisantes à l'attribution du label «physique», même si elles sont absolument nécessaires. Leur respect ne garantit pas, en soi, la fidélité aux phénomènes physiques. Si le modèle de Phong ne possède pas ces deux qualités comme le souligne Ward, il nous semble qu'il n'est guère plus éloigné de la physique que ne l'est celui de Ward eu égard aux multiples aspects de l'interaction lumière/matière qui ne se résument pas à ces deux propriétés. Reste donc la confrontation à l'expérience. Ward compare des mesures de la FDRB effectuées dans le plan d'incidence sur un aluminium brossé avec les résultats de son modèle paramétré pour suivre au mieux les variations expérimentales. Voici des mesures où l'accord entre modèle et expérience varie beaucoup selon les cas :

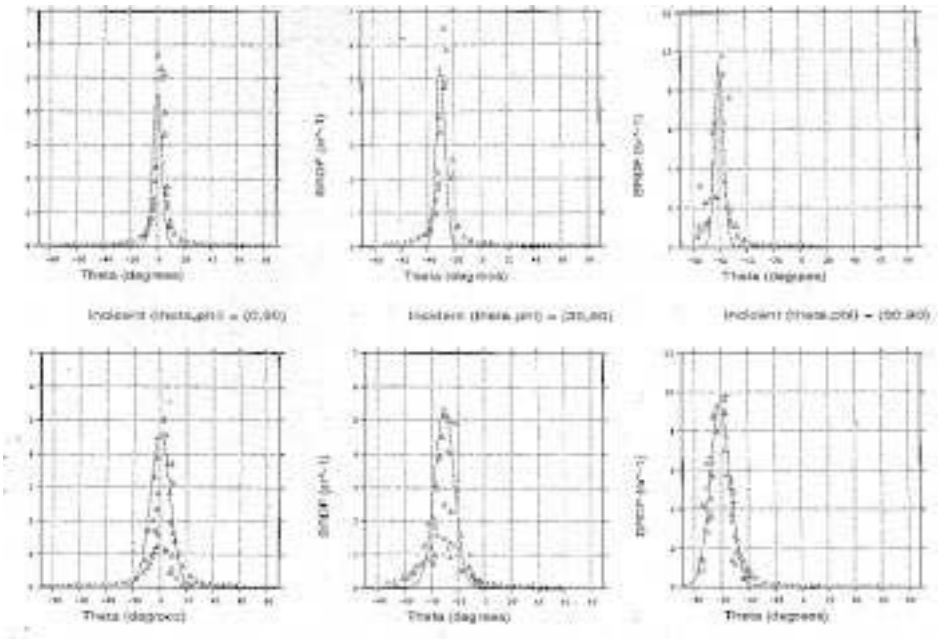


Figure 25: Mesures (triangles) et calculs (pointillés) de Ward pour la FDRB d'un aluminium brossé

Compte tenu de la connaissance assez précise de ρ_d et ρ_s, on peut s'étonner que l'accord ne soit pas bien meilleur. Comme le signale Ward en conclusion, de nombreux matériaux échappent totalement au modèle ce qui, au vu de sa construction, n'étonne guère.

II.2 Modèles d'éclairements globaux

Tous les modèles précédents sont considérés comme locaux ou semi-locaux, en ce sens qu'ils ne prennent pas en compte l'ensemble de l'énergie qui arrive en un point mais seulement celle provenant d'une ou deux directions privilégiées. Seul l'éclairement direct est pris en compte dans les modèles de Phong ou de Blinn. L'extension de Whitted permet de capturer les effets indirects spéculaires idéaux réfléchis ou transmis. Les autres contributions indirectes sont grossièrement approximées par les composantes diffuses et ambiante. Avec ces modèles, les surfaces fortement spéculaires sont relativement bien rendues tandis que les surfaces diffuses et leurs effets sur le reste de l'environnement n'ont qu'une représentation médiocre.

Devant ces insuffisances, des extensions ont été proposées pour élargir la gamme des effets lumineux capturés par l'élaboration de modèles globaux ou semi-globaux. Les deux premiers auteurs a avoir exploré cette voie en image de synthèse furent Cook et Goral qui proposèrent, tous deux en 84, de nouvelles méthodes.

La méthode introduite par Goral et al. [Gora et al 84] s'inspire des travaux réalisés en physique dans les années 50 pour le calcul d'échanges de chaleur. Considérant toutes les surfaces comme des diffuseurs parfaits, Goral et al proposent de calculer les échanges de *radiosités* entre toutes les surfaces d'une scène par la résolution d'un système d'équations représentant ces échanges. La radiosité a fourni une alternative au tracé de rayons en permettant de visualiser les interréflexions diffuses, phénomènes qui échappaient totalement à cette dernière technique.

Cook et al [Cook et al 84] proposèrent de rendre des phénomènes de flous avec la méthode du *tracé de rayons distribués* consistant à échantillonner plusieurs rayons au lieu d'un seul. Différents effets peuvent être ainsi produits selon l'origine de l'échantillonnage. Un flou de profondeur de champ est obtenu en échantillonnant plusieurs rayons par pixels simulant l'effet de focalisation d'une lentille. Des effets de réflexions ou de transmission floues sont obtenus par échantillonnage des rayons transmis ou réfléchis autour de la position spéculaire idéale. Des pénombres sont produites par échantillonnage des angles solides sous-tendus par des sources surfaciques. Enfin, des flous de bougé sont simulés par échantillonnage du temps. Le tracé de rayons distribués n'établit pas encore un modèle global puisque l'échantillonnage est limité à des cônes d'angles solides faibles et qu'il n'y a pas de notion de réduction de variance.

II.2.1 L'équation de rendu

Il faudra attendre le formalisme de Kajiya en 86 [Kaji 86] pour que la notion d'éclairement global prenne vraiment corps avec son *équation de rendu* :

$$I(x, x') = g(x, x')\left(\varepsilon(x, x') + \int_{\bigcup \Sigma_i} \rho(x, x', x'')I(x', x'')dx''\right) \quad (30)$$

$I(x, x')$ est l'intensité provenant du point x' et arrivant au point x
$g(x, x')$ est un terme représentant la visibilité et l'atténuation entre les points x et x'. Si x ne «voit» pas x', le terme est nul. Dans le cas contraire, il exprime l'atténuation qui existe dans l'espace séparant x de x'.
$\varepsilon(x, x')$ est l'émission propre du point x' vers le point x.
$\rho(x, x', x'')$ fournit le rapport entre l'intensité émise depuis x'', réfléchie en x' et arrivant en x.
$\cup\Sigma_i$ est l'ensemble de toutes les surfaces de la scène.

Cette équation représentant l'équilibre énergétique, Kajiya l'explicite en fonction du tracé de rayons et de la radiosité. En restant au niveau des principes, il passe en revue un certain nombre de méthodes d'échantillonnage qu'il qualifie de *hiérarchiques*.

La première est un échantillonnage séquentiel uniforme assurant la couverture complète du domaine de variation des paramètres. Les points échantillonnés sont stockés dans les feuilles d'un arbre. En une dimension, l'algorithme permettant cet échantillonnage est le suivant :

```
ChoisirUnNœud(Nœud)

  SI Nœud est une feuille ALORS
        ChoisirUnNœud ← Nœud
  SINON SI Niveau(Gauche(Nœud)) < Niveau(Droit(Nœud)) ET Equilibré(Gauche(Nœud)) ALORS
        ChoisirUnNœud ← ChoisirUnNœud(Gauche(Nœud))
  SINON SI Niveau(Gauche(Nœud)) > Niveau(Droit(Nœud)) ET Equilibré(Droit(Nœud)) ALORS
        ChoisirUnNœud ← ChoisirUnNœud(Droit(Nœud))
  SINON
        Choisir par tirage uniforme le nœud gauche ou droit
  FINSI
```

Cette stratégie est en fait un schéma de *stratification* (voir §II.2.2.2) où le seul critère de découpage des strates est l'effectif de la population dans chaque strate : l'échantillon est pris dans la strate ayant le moins d'échantillons et le plus grand effectif. Ce principe est l'un des plus mauvais qui soit eu égard à la vitesse de convergence de l'évaluation de (30). Il considère en effet sur un même pied d'égalité, les individus qui ont une forte contribution à l'intégrale et ceux qui n'y contribuent que peu ou pas.

Un schéma adaptatif est également proposé : l'*intégration hiérarchique adaptative*. Elle consiste à choisir un nœud selon un certain nombre de seuils ϕ_i valant zéro ou un selon que la règle i associée au seuil conduit au choix du nœud gauche ou droit de l'arbre. Ces seuils sont linéairement combinés avec des poids w_i selon l'importance accordée à la règle i. Un nouveau seuil ϕ est donc formé par :

$$\phi = \sum \phi_i w_i \qquad \sum w_i = 1,\ w_i \geq 0$$

Un nœud est alors choisi par tirage uniforme dans [0, 1] d'un nombre et comparaison de celui-ci avec ϕ. Sans être très précis, Kajiya propose quelques seuils : la différence d'intégrale entre deux sous-nœuds, un historique des changements dans les sous-nœuds (sous forme de variance par exemple) ou encore des fonctions a priori pouvant prévoir les éclairements les plus forts. Kajiya dit ne pas avoir trouvé de critère satisfaisant et l'implémentation de son échantillonnage n'est pas adaptative.

Le schéma général d'échantillonnage de ce que Kajiya appelle le *tracé de chemins* (path tracing) n'est malheureusement pas assez détaillé pour en déduire son fonctionnement réel. Le choix des rayons secondaires selon la probabilité médiane est notamment assez obscure. Le principe de base du tracé de chemins consiste en tous cas à n'échantillonner qu'un seul rayon par hémisphère.

Cette façon de procéder part de la constatation qu'au fur et à mesure du développement d'un arbre de rayons, le coût des nouveaux rayons est de plus en plus élevé tandis que la contribution à l'intégrale est, en général, de plus en plus faible. En d'autres termes, beaucoup de travail pour un résultat qui en vaut de moins en moins la peine selon Kajiya qui prend le contre-pied de cette tendance en choisissant de multiplier les chemins sans fourche.

Ce raisonnement est cependant partiel comme Kirk et Arvo l'ont bien montré [Kirk et al 90]. En effet, dans tous les cas où un certain nombre de matériaux spéculaires précèdent un matériau diffus, la stratégie de l'éparpillement sur le diffuseur est plus efficace que celle du rebond unique parce que la probabilité d'avoir une contribution représentative avec cette dernière technique est très faible. De façon plus formelle et suivant la démarche de Kirk et Arvo, une mesure μ de l'efficacité de ces méthodes peut être donnée par :

$$\mu = 1/(\sigma^2\tau)$$

où σ^2 est la variance de l'estimateur et τ le coût d'échantillonnage d'un chemin complet d'une particule. Plus μ est petit, moins bonne est la méthode puisque soit son coût soit sa variance sont élevés. Si N surfaces sont parfaitement spéculaires et que le $N + 1^{ème}$ rebond aboutit à un diffuseur parfait, l'efficacité de l'estimateur à rebond unique, s'il conduit à une variance σ_1^2, vaudra :

$$\mu_1 = 1/(\sigma_1^2(N\chi + \tau))$$

χ étant le coût moyen de lancé d'un rayon. Si, par contre, la méthode par éparpillement a une variance inférieure σ_*^2, avec m chemins de coût τ, son efficacité sera :

$$\mu_* = 1/(\sigma_*^2(N\chi + m\tau))$$

En observant que

$$\lim_{N \to \infty} \frac{\mu_*}{\mu_1} = \frac{\sigma_1^2}{\sigma_*^2} > 1$$

on constate que l'éparpillement peut être plus efficace si un nombre de réflexions spéculaires suffisant intervient avant une diffusion. Un autre cas ayant les mêmes conséquences se produit lorsque, quel que soit le point choisi sur un chemin, si m échantillons de coûts égaux produisent une variance σ_*^2 telle que $\sigma_*^2 < \sigma_1^2 / m$ alors l'éparpillement est plus efficace.

L'ennui majeur de l'éparpillement, comme l'a souligné Kajiya, est bien sûr le coût exponentiel de son développement. Aussi , Kirk et Arvo proposent un moyen efficace d'en limiter la profondeur. La troncature classique proposée par Hall [Hall et al 83] consiste à tracer un rayon en considérant son atténuation cumulée au fur et à mesure de ses rebonds successifs jusqu'à ce que cette atténuation passe en dessous d'un seuil fixé à l'avance. Une fois le seuil atteint, l'arbre des rayons est tronqué et seule la contribution obtenue à ce stade est considérée dans la somme finale. La troncature est indispensable pour deux raisons : non seulement pour limiter le coût du développement de l'arbre, mais également pour éviter les réflexions infinies, cas rarissime mais incontournable. La troncature adaptative de Hall, si elle limite la profondeur de l'arbre, introduit un biais systématique dans l'évaluation de l'éclairement. Kirk et Arvo proposent donc une troncature adaptative sans biais par la méthode dite de la *roulette russe*. Lorsque la contribution du chemin tombe en-dessous d'un seuil S donné, cette technique statistique consiste à tirer au hasard sa poursuite selon une probabilité P. En-dessous du seuil, un nombre u est tiré uniformément dans [0, 1], si u est inférieur à P, le chemin est tronqué. Dans le cas contraire, la contribution du rayon est pondérée par 1 / (1 - P) et un nouveau rayon est relancé jusqu'à troncature. On peut montrer que l'espérance du processus sans troncature est la même que celle du processus tronqué : il n'y a pas de biais. Le choix de P doit être réalisé de façon judicieuse pour que la variance de l'estimateur ne

soit pas trop grande (P ne doit pas être trop proche de 1) mais il faut également ne pas choisir P trop faible, sans quoi la longueur moyenne des chemins risque d'être démesurée.

II.2.2 Méthodes de Monte Carlo

Il y a finalement peu de temps que l'équation de rendu a pris un intérêt autre que théorique. Les années 90 ont en effet vu fleurir un grand nombre de méthodes d'échantillonnage permettant l'évaluation de l'éclairement autrement que par la force brute. Toutes ces techniques sont dérivées des méthodes statistiques dites de *Monte Carlo* dont le développement a débuté dans les années 1944 pour la mise au point de la bombe atomique. Leur champ d'application est très vaste car elles sont en fait très génériques. Nous donnons ci-après un certain nombre de techniques couramment utilisées dont on trouvera une discussion plus détaillée dans l'exposé de Rubinstein [Rubi 81].

II.2.2.1 Méthode brute

Soit à évaluer une intégrale du type

$$F = \int_0^1 f(x)dx \qquad f \in L^2(0, 1)$$

Si $u_1,\ldots,u_n$ sont des nombres aléatoires indépendants tirés uniformément dans [0, 1], l'espérance

$$\bar{F} = \frac{1}{n}\sum_{i=1}^{n} f(u_i)$$

est un estimateur sans biais de F dont la variance s'exprime par

$$\frac{1}{n}\int_0^1 (f(x) - F)^2 dx = \frac{\sigma_1^2}{n}$$

l'erreur standard étant sa racine carrée. Le facteur en racine de n de cette erreur indique que pour la réduire de moitié par exemple, il faudrait prendre quatre fois plus d'observations.

II.2.2.2 Stratification

La stratification consiste à découper la population en strates, dont l'union décrit l'ensemble du domaine d'intégration, et à échantillonner chaque strate séparément comme dans la méthode brute. L'estimateur sans biais de F a alors la forme suivante :

$$\tilde{F}_1 = \sum_{j=1}^{k}\sum_{i=1}^{n_j} \frac{s_j - s_{j-1}}{n_j} f(s_{j-1} + (s_j - s_{j-1})u_{ij})$$

où $s_1 = 0 < s_2 < \ldots < s_n = 1$ délimitent chaque strate
n_j : nombre de points échantillonnés dans la strate j

La variance de cet estimateur sans biais étant

$$\sigma_1^2 = \sum_{j=1}^{k} \frac{s_j - s_{j-1}}{n_j} \int_{s_{j-1}}^{s_j} f^2(x)dx - \sum_{i=1}^{n_j} \frac{1}{n_j} \left(\int_{s_{j-1}}^{s_j} f(x)dx \right)^2$$

elle peut être plus faible que $\sigma_1{}^2$ à condition que les différences entre les valeurs moyennes de f dans les strates soient plus grandes que ses variations dans ces mêmes strates. La partition la plus simple est le découpage en intervalles de longueur égale. Un découpage plus efficace consiste à choisir les s_j de façon à ce que $s_j / \sum s_j$ soit proportionnel à la valeur de F dans chaque strate.

II.2.2.3 Échantillonnage d'importance

Cette méthode conduit à privilégier les parties les plus importantes de f (i.e. celles contribuant le plus à F) en échantillonnant selon une fonction g ressemblant à f

$$F = \int_0^1 \frac{f(x)}{g(x)} g(x)dx = \int_0^1 \frac{f(x)}{g(x)} dG(x) \qquad G(y) = \int_0^y g(x)dx,\ G(1) = 1$$

$$\tilde{F}_2 = \frac{1}{n} \sum_{i=1}^{n} \frac{f(\eta_i)}{g(\eta_i)} \qquad \eta_i \sim G[0, 1]$$

$\tilde{F}_2$ est un estimateur sans biais de variance

$$\sigma_2^2 = \int_0^1 \left(\frac{f(x)}{g(x)} - \tilde{F} \right)^2 dG(x)$$

Il est clair que si f / g est constant la variance sera nulle. Cette situation idéale revient à savoir intégrer f analytiquement et donc à ne pas avoir besoin de Monte Carlo… Néanmoins une fonction g mimant f, même grossièrement, peut permettre une réduction de variance.

II.2.2.4 Contrôle de variations

Cette dernière technique requiert également l'intégration analytique d'une fonction v et utilise la même idée que la méthode précédente :

$$F = \int_0^1 v(x)dx + \int_0^1 (f(x) - v(x))dx$$

dont l'estimateur sans biais est :

$$\tilde{F}_3 = \frac{1}{n} \sum_{i=1}^{n} (f(u_i) - v(u_i)) + \int_0^1 v(x)dx$$

II.2.2.5 Applications

Les travaux utilisant ces résultats en synthèse se sont développés selon trois grands axes chacun impliquant des stratégies spécifiques d'échantillonnage. Le premier axe cherche à échantillonner des FDRBs assez simples. Le second s'intéresse à l'échantillonnage des sources et le dernier utilise une approche qui démarre des sources et non plus de l'oeil.

Lange est sans doute la première a avoir utilisé l'échantillonnage de la FDRB [Lang 91]. Partant d'un modèle de Phong composé d'une partie diffuse et d'une partie spéculaire, elle propose une méthode mélangeant deux échantillonnages pour chacune de ces composantes. L'échantillonnage de la partie diffuse consiste en un tirage aléatoire uniforme décrivant l'angle solide sous-tendu par l'hémisphère autour de chaque point d'intersection (Cf §III.2.2). La partie spéculaire est échantillonnée selon le pic spéculaire du modèle de Phong par une distribution basée sur le $\cos^n$ du modèle. Le mélange de ces deux échantillonnages de rayons réfléchis prend deux formes selon que le rayon incident est primaire ou secondaire. Le nombre de rayons, N, générés depuis l'oeil est fixé au démarrage par l'utilisateur, Lange en utilise 40 dans son implémentation. Les rayons réfléchis au premier point d'intersection sont générés aléatoirement dans la proportion k_sN avec l'échantillonnage spéculaire et dans la proportion k_dN avec l'échantillonnage diffus, k_s et k_d étant les coefficients classiques du modèle pour la surface intersectant le rayon primaire ($k_s+k_d = 1$). Il semble que les rayons secondaires soient échantillonnés aléatoirement, moitié de façon diffuse, moitié de façon spéculaire, pour fournir la contribution indirecte de l'éclairement. La contribution directe est systématiquement calculée et ajoutée à l'indirecte dans des proportions variant selon que l'échantillonnage retenu est diffus ou spéculaire. Cette stratégie a toutes les chances de mal fonctionner pour les rayons secondaires sur les matériaux spéculaires puisqu'ils ont une chance sur deux d'être échantillonnés selon le schéma uniforme de la partie diffuse.

Complémentaire de cette démarche, Shirley et al. [Shir et al 96] se sont intéressés à l'échantillonnage des sources et donc de l'éclairement direct que Lange ne prend pas vraiment en compte. Différentes formes de sources, en majorité planaires, sont considérées (source circulaire, rectangulaire, triangulaire, …) ainsi que les échantillonnages correspondants. Ces échantillonnages consistent essentiellement à échantillonner uniformément l'angle solide sous-tendu par chaque type de source en incluant souvent le cosinus lié à la FDRB. Shirley et al. s'intéressent ensuite aux scènes contenant plusieurs sources en proposant une méthode pour choisir celle qui sera échantillonnée. L'idée est de partitionner la scène par un octree où chaque feuille contient les lumières ayant les contributions les plus puissantes (au-dessus d'un certain seuil) moyennant quelques simplifications : une source est visible de tous les objets de la feuille et la FDRB de ces objets est constante et maximale. Échantillonner une source consiste alors à la choisir aléatoirement selon sa contribution, pondérée par la somme des contributions des sources les plus fortes, plus une approximation de la contribution des sources les plus faibles. Ceci conduit à échantillonner les sources qui a priori contribuent le plus à l'énergie réfléchie.

Veach et Guibas proposent une méthode générale permettant de combiner différents types d'échantillonnages [Veac et al 95]. Classiquement, combiner ces échantillonnages consiste à choisir d'utiliser l'un d'entre eux de façon aléatoire selon une proportion que l'on se donne. Partant de ces proportions, les auteurs proposent de les combiner de telle façon que la variance du nouvel estimateur avec cette combinaison soit inférieure ou égale à celle de l'estimateur avec la combinaison de base, c'est ce qu'ils appellent l'*heuristique de la balance*. Celle retenue par les auteurs consiste en un

mélange équitable entre un échantillonnage de la FDRB (non explicité) et l'échantillonnage uniforme des sources. Ils appliquent ensuite leur nouvelle combinaison à ce mélange. Ils proposent ensuite des variantes permettant une réduction de variance dans certains cas.

S'inspirant de Kajiya, d'autres travaux ont portés sur une nouvelle approche pour le tracé de chemins en lançant des rayons, non plus de l'œil seulement, mais également des sources. Parmi ceux-ci, on peut citer ceux de Lafortune et Willems qui proposent un tracé de chemins bidirectionnel (bi-directional path tracing)[Lafo et al 94], une idée également utilisée par Veach et Guibas à la même époque [Veac et al 94]. Ils construisent ainsi un estimateur du flux de la façon suivante :

$$\tilde{\Phi} = \sum_{i=0}^{N_l} \sum_{j=0}^{N_e} w_{ij}\phi_{ij}$$

ϕ_{ij} est une estimation du flux après i réflexions depuis les sources, j réflexions depuis l'œil.
w_{ij} est le poids de la contribution ϕ_{ij}.

Il semble que toutes les interactions possibles soient considérées par l'envoi de rayons d'ombres entre les points constituant le chemin depuis la source et ceux présents sur celui issu de l'œil. Trois cas peuvent être considérés selon la longueur des différents chemins :

➟ Contribution directe des sources sur l'œil : i = j = 0

$$\phi_{00} = GL_e(y_0, \theta_{y_0})$$

$$G = \int_A \int_{\Omega_y} g(y, \theta_y)\left|\theta_y \cdot N_y\right| d\omega_y d\mu_y$$

➟ Tracé de rayons classique : i = 0, j > 0

$$\phi_{0j} = GL_e(x_0, \theta_{x_0 \to y_{j-1}})L'f_r(y_{j-1}, \theta_{x_0 \to y_{j-1}}, \theta_{y_{j-1}})\frac{\left|\theta_{x_0 \to y_{j-1}} \cdot N_{x_0}\right|\left|\theta_{x_0 \to y_{j-1}} \cdot N_{y_{j-1}}\right|}{\left\|x_0 - y_{j-1}\right\|^2}\mathbf{1}_{x_0 \text{ voit } y_{j-1}}$$

$$L' = L \Big/ \left(\int_{\Omega_y} L_e(x_0, \theta_{x_0})\left|\theta_x \cdot N_{x_0}\right| d\omega_x\right)$$

➟ Cas général : i > 0, j > 0

$$\phi_{ij} = LGf_r(y_{j-1}, \theta_{x_0 \to y_{j-1}}, \theta_{y_{j-1}})f_r(y_{j-1}, \theta_{x_0 \to y_{j-1}}, \theta_{y_{j-1}})\frac{\left|\theta_{x_i \to y_{j-1}} \cdot N_{x_i}\right|\left|\theta_{x_i \to y_{j-1}} \cdot N_{y_{j-1}}\right|}{\left\|x_i - y_{j-1}\right\|^2}\mathbf{1}_{x_0 \text{ voit } y_{j-1}}$$

Le dernier volet de la méthode concerne le choix des pondérations. Celui-ci est assez libre pourvu que la condition suivante soit respectée :

$$\sum_{i=0}^{N} w_{i,N-i} = 1, \forall N$$

Par exemple, $w_{0j} = 1$ et $w_{i \neq 0\, j} = 0$ conduit à une forme du tracé de chemins de Kajiya.

Lafortune et Willems proposent un schéma légèrement différent :

$$w_{i0} = 0$$

$$w_{0\,j\neq 0} = \prod_{k=0}^{j-2} W_k$$

$$w_{i\neq j\neq 0} = \left(\prod_{k=0}^{j-2} W_k\right)(1 - W_j)$$

où les W_j sont choisis proportionnels au degré de spécularité de la surface au point y_j , ce qui permet de préférer la génération d'un rebond spéculaire à la liaison par rayon d'ombre ou inversement selon la matière. Dans l'implémentation qu'ils ont faite, les auteurs utilisent un modèle de Phong modifié pour le rendre réciproque et conservateur d'«énergie». Ils utilisent également une troncature du type roulette russe ainsi qu'une forme d'échantillonnage stratifié non explicitées. Un suréchantillonnage d'au moins 20 rayons par pixel est réalisé pour établir la contribution finale arrivant au pixel. Cette technique est soumise aux mêmes inconvénients que ceux signalés par Kirk et Arvo. Des techniques semblables de tracé de chemins sont proposées par les mêmes auteurs dans un autre article [Lafo et al 94], techniques utilisées en conjonction avec une réduction de variance par contrôle de variation. Cette technique n'a d'intérêt que lorsque v est analytiquement intégrable et proche de f pour qu'une réduction de variance ait effectivement lieu. Lafortune et Willems proposent la forme suivante :

$$I = \int f_r(x, \theta_y, \theta_x)(L_d(y, \theta_y) + L_i(y, \theta_y) - L_a(y, \theta_y))\left|\theta_y \cdot N_x\right| d\omega_y + \int f_r(x, \theta_y, \theta_x)L_a(y, \theta_y)\left|\theta_y \cdot N_x\right| d\omega_y$$

Dans leurs hypothèses, le second membre est analytiquement intégrable ce qui permet de le remplacer par $\rho(x, \theta_x)L_a$. L_a peut varier selon les objets intersectés mais les auteurs le choisissent constant sur l'ensemble de la scène. Les essais des auteurs montrent que la variance peut être réduite pour certaine valeur de L_a, mais elle peut également augmenter par rapport à la méthode classique du tracé de chemins selon la valeur de L_a. Le choix entre une bonne et une mauvaise valeur est intuitif et guère formalisable. En pratique, de plus, pour que la méthode apporte un gain réel, il faudrait que L_a soit différent en chaque point d'intersection. Le deuxième inconvénient majeur est la nécessité de l'intégration préalable de v. En fait, l'intégration de toutes les FDRB des différents matériaux de la scène selon toutes les incidences possibles. Une opération réalisable avec des FDRB très simples n'ayant qu'un vague rapport avec les modèles physiques.

L'idée de mélanger les méthodes partant de l'œil avec celles partant des sources a également été explorée par Veach et Guibas [Veac et al 94]. Ils adoptent une classification des méthodes selon le nombre d'étapes réalisées depuis la source ou depuis l'œil. Ainsi un chemin de longueur k peut donner lieu a une évaluation selon k méthodes : celle réalisant k - 1 brins depuis la source et joignant le dernier point d'intersection à l'œil, celle réalisant k - 2 brins depuis la source, générant un rayon depuis l'œil et joignant les deux points d'intersection, etc. Nous qualifierons de méthodes réceptives, les méthodes partant de l'œil, tandis que celles démarrant des sources seront dites émissives. Une méthode (o, l) génère un chemin de longueur o depuis l'œil et un chemin de longueur l depuis la source. Une contribution n'est bien sûr retenue que si les points d'aboutissement des deux chemins sont visibles l'un de l'autre. Certaines configurations

de scène ne peuvent être correctement traitées qu'en démarrant de la source, par exemple la probabilité d'échantillonner une source ponctuelle par une méthode réceptive aveugle est nulle. Partant de cette constatation, les auteurs donnent une heuristique permettant de choisir la méthode émissive ou réceptive la plus adaptée.

Se plaçant dans un schéma d'échantillonnage d'importance où la génération d'un chemin x de longueur k est conditionné par les différentes probabilités $p_i(x)$ pour chaque méthode i, l'heuristique du maximum est la suivante :

> *POUR CHAQUE* méthode i
> Choisir un chemin x avec la méthode
> *SI* $p_i(x) \geq p_j(x)\ \forall\ j \neq i$ *ALORS* $I \leftarrow I + f(x) / p_i(x)$

Le but est de minimiser les variations de la variance en choisissant pour chaque chemin la probabilité de réalisation la plus forte. Néanmoins, il n'est pas certain que l'algorithme ci-dessus réalise bien cette idée. Tout d'abord la contribution à l'intégrale finale n'est pas considérée, la réduction de variance agit donc localement mais ne garantit pas une réduction globale. Ensuite le choix systématique de la probabilité la plus grande n'assure pas un bon comportement moyen, un choix probabiliste basé sur ces probabilités serait bien meilleur. Enfin la limitation du nombre de chemins à exactement k est assez arbitraire et n'éliminera sûrement pas le bruit final.

II.3 Un modèle pseudo-réaliste

Comme nous venons de le voir, un modèle global comporte en fait deux parties distinctes et liées. La première est un modèle local de transfert d'énergie, la seconde une méthode de résolution globale des échanges énergétiques entre ces fonctions de transfert. Il nous faut désormais choisir un modèle local de transfert d'énergie qui soit aussi physiquement fondé que possible. Nous exposerons les méthodes de résolution dans le prochain chapitre.

Parmi les modèles que nous avons examinés, il nous semble que celui de Stogryn est le plus satisfaisant de par la généralité de ses traitements et son approche électromagnétique. Malheureusement, comme nous l'avons montré, son modèle ne permet pas d'établir les composantes du vecteur électrique réfléchi, nous ne pouvons donc utiliser cette approche. Faute de mieux, nous nous contenterons de l'approche géométrique de Torrance et Sparrow (Cf §I.2.2). La quantité essentielle que nous voulons établir est la luminance atteignant l'oeil. Pour ce faire, il nous faut pouvoir établir la luminance réfléchie en fonction de la luminance incidente à chaque interaction lumière/matière. Les calculs que nous avons menés ((6) §I.2.2 ainsi que (28) §II.1.2) nous donnent la réponse :

$$
\begin{gathered}
L_r = k_d L_{rd} + k_s L_{rs} \\
L_{rd} = L_i \cos\zeta_i \qquad L_{rs}(\zeta_i, \zeta_r, \alpha_r) = \frac{G(\zeta_{\mathbf{V}}, \zeta_{\mathbf{L}}, \zeta_{\mathbf{n}}) R(\psi_{\mathbf{n}}) d\omega_i L_i p(\zeta_{\mathbf{n}})}{4\cos\zeta_r} \\
G(\zeta_{\mathbf{V}}, \zeta_{\mathbf{L}}, \zeta_{\mathbf{n}}) = \min\left(1, \frac{2\cos\zeta_{\mathbf{n}}\cos\zeta_{\mathbf{L}}}{\cos\psi_{\mathbf{n}}}, \frac{2\cos\zeta_{\mathbf{n}}\cos\zeta_{\mathbf{V}}}{\cos\psi_{\mathbf{n}}}\right) \\
p(\zeta_{\mathbf{n}}) = \left(m\sqrt{2\pi}e^{\tan^2\zeta_{\mathbf{n}}/(2m^2)}\cos^2\zeta_{\mathbf{n}}\right)^{-1} \\
R(\psi_{\mathbf{n}}) = (R_{\parallel} + R_{\perp})/2
\end{gathered}
\tag{31}
$$

Nous soulignerons de nouveau les présupposés et les limitations de cette approche.

II.3.1 Présupposés généraux

Torrance et Sparrow se situent d'emblée dans le cas de l'optique géométrique : la surface se comporte donc localement comme un miroir. La conséquence de cette hypothèse est que la lumière réfléchie ne résulte que du seul phénomène spéculaire simplifié en deux états : l'un direct pour les portions de surface bien orientées, l'autre indirect, diffus, résultat des inter-réflexions entre toutes les portions mal orientées. On notera que cette séparation artificielle a le gros inconvénient de ne pas relier la partie spéculaire à la partie diffuse ce qui permet d'avoir, par exemple, une surface en même temps très diffuse aussi bien que spéculaire. Si le phénomène peut être limité en utilisant la conservation de l'énergie, il n'en reste pas moins que ces deux comportements restent indépendants sur le fond, alors qu'ils devraient être corrélés. L'hypothèse géométrique implique aussi que la rugosité soit supérieure à la longueur d'onde incidente.

II.3.2 Hypothèses sur la surface

La surface est en même temps décrite par un processus gaussien et par un certain nombre de contraintes :

- La surface est constituée de facettes planes
- Chaque facette appartient à une cavité symétrique
- L'axe longitudinal des cavités est parallèle au plan de la surface moyenne
- Les orientations des cavités sont équiprobables
- Les bords supérieurs des facettes sont tous dans le même plan

Ces contraintes sont à la fois très fortes et insuffisantes dans le sens où la définition de la surface qu'elles entraînent est incomplète. En effet le seul type de surface qui pourrait convenir serait composé de pyramides renversées (en fait des cônes symétriquement facettisés) : une fois fixé la première d'entre elles, les suivantes se construisent de proche en proche avec au moins deux côtés fixés par la pyramide initiale. Néanmoins les critères ci-dessus ne disent rien quant à la taille des facettes et à leur succession (l'auto-corrélation n'est pas spécifiée). Il résulte de cette indétermination qu'une infinité de surfaces fort différentes ne seront représentées que par un seul et unique comportement, ce qui n'est pas vraisemblable.

Nous ne connaissons aucune étude qui ait tenté de vérifier l'hypothèse gaussienne pour des surfaces réelles (encore moins celle des facettes symétriques). Étant donné son usage très répandu en physique, des vérifications expérimentales ont certainement été entreprises, il serait indispensable de les étudier pour connaître son degré de validité.

Compte tenu de toutes ces limitations, nous ne prétendons pas que ce modèle puisse rendre compte de la complexité de la réalité dans tous ses détails. Nous pensons néanmoins qu'il est mathématiquement cohérent et physiquement validé par les expériences de Torrance et Sparrow. Pour toutes ces raisons, nous le qualifions de *pseudo-réaliste* en lui reconnaissant toutes les limites inhérentes à sa conception que nous venons de rappeler. Le terme «réaliste» utilisé ici nécessite cependant des éclaircissements tant il est galvaudé et souvent usité sans définition préalable. Cette absence, d'autant plus curieuse que le concept est au coeur même de la synthèse, nous semble très préjudiciable à une approche complète des buts et moyens de la synthèse, aussi nous y attarderons-nous en guise de conclusion.

II.4 La question du réalisme

II.4.1 De la synthèse à l'inconscient

Le cadre dans lequel nous nous situons depuis le début de cette thèse est celui de la *synthèse réaliste*. Ces termes prennent en général deux acceptations différentes selon les auteurs. La première peut se résumer sous le vocable de *photoréalisme*, la seconde sous celui de *physicoréalisme*. Si le but communément admis par ces deux démarches est de produire des images indistinguables de celles produites par un appareil photographique, en revanche la mise en oeuvre de leur vérification est radicalement différente. Encore faut-il corriger quelque peu cette première définition. La quasi-totalité des techniques utilisées aboutissent à des images vidéographiques, la référence au support papier et au procédé argentique qu'implique la photographie est donc impropre. Il est plus approprié de dire que le but de la synthèse d'images réalistes est de produire des images qu'un observateur ne saurait distinguer de celles de scènes réelles équivalentes reproduites par le même média.

Une image sera donc déclarée valide lorsque le sens visuel commun la jugera acceptable. Si cette conclusion convient aux photoréalistes, les physicoréalistes exigent en plus que le résultat final de la simulation soit également physiquement vérifiable. Ceci implique non seulement que les modèles infographiques soient fidèles aux lois élaborées par la physique, mais également que le résultat des calculs puisse être confronté favorablement à des mesures de phénomènes réels. En étant quelque peu caricatural, nous disons que Phong est un photoréaliste tout comme Ward, alors que Meyer a fourni un des meilleurs exemples de démarche physicoréaliste.

Meyer cherche à valider le modèle de la radiosité par un schéma expérimental complet [Meye et al 86]. Partant d'une scène réelle composée de trois cubes et d'une source - dont les caractéristiques physiques ont été mesurées (dimensions, réflectance, puissance émise) - Meyer et al. en réalisent un modèle numérique discret facétisé. Des images sont alors calculées par la méthode classique de radiosité [Gora et al 84] ainsi que par l'extension de l'*hemi-cube* de Cohen et al. [Cohe et al 85]. Les calculs ont fait l'objet de vérification par comparaison aux mesures réalisées avec un photomètre. Ils ont également été soumis au jugement psycho-visuel de 20 observateurs auxquels étaient présentées deux vues fournies par des caméras, l'une filmant la scène réelle, l'autre un moniteur affichant l'image calculée.

A la question de déterminer quelle était l'image calculée, 45 % des personnes fournit la mauvaise réponse, un résultat à peu près aussi bon qu'un choix aléatoire. Dans l'ensemble, les observateurs jugèrent assez bon la correspondance entre modèle et simulation. Les comparaisons numériques sur les différentes configurations

donnèrent des écarts-type entre mesures et calculs variant entre 7 et 4 % pour les mesures absolues et aux alentours de 3 % pour des comparaisons relatives. La méthode de l'hémi-cube fournissant de meilleurs résultats que l'approche classique.

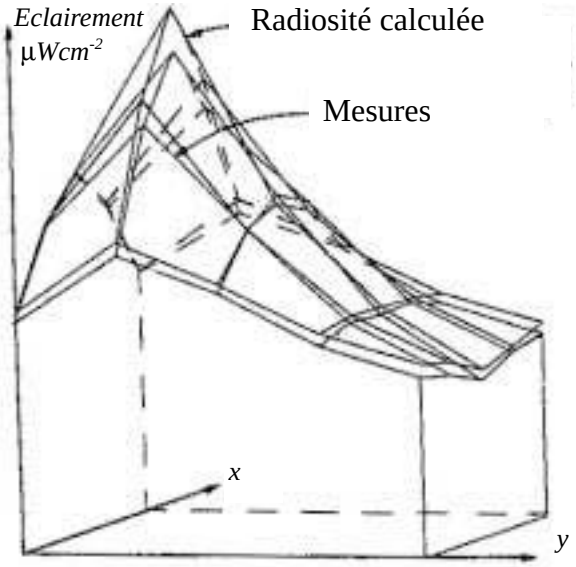


Figure 26: Calculs et mesures pour l'expérience de Meyer sur une scène composée d'un cube blanc ouvert sur le devant (là où sont faites les mesures) contenant un plus petit cube blanc

A l'opposé de cette démarche, on peut se demander s'il est bien indispensable de vouloir strictement se conformer aux modèles physiques, sachant que les résultats finaux manipulés sont à valeurs discrètes dans un intervalle «ridicule» de 0 à 255. Ce à quoi l'on peut rétorquer que les effets d'une approximation locale peuvent être amplifiés, démultipliés par l'application récursive de celle-ci comme c'est le cas en tracé de rayons. Les phénomènes en jeu sont, la plupart du temps, de nature multiplicatifs et une sous-évaluation ne pourra jamais être compensée par une sur-évaluation (modèle à conservation d'énergie en l'absence de surfaces émissives). On peut également s'interroger sur le comportement global d'un modèle sur l'ensemble d'une image. Une image qui ne choque pas notre sens visuel est-elle signe d'un «bon» modèle comme le sous-tend la démarche photoréaliste ? Si c'est une condition souvent nécessaire, elle n'est sûrement pas suffisante. Pour illustrer notre propos nous donnons ci-après quelques exemples d'images issues de divers domaines. Aucune d'entre elles ne choque notre sens visuel, aucune n'est incompréhensible. Toutes ont été créées à la main avec des techniques de dessin plus ou moins traditionnelles et peuvent être qualifiées de réalistes.

Figure 28: exemple 2

Figure 27: exemple 1

Figure 29: exemple 3

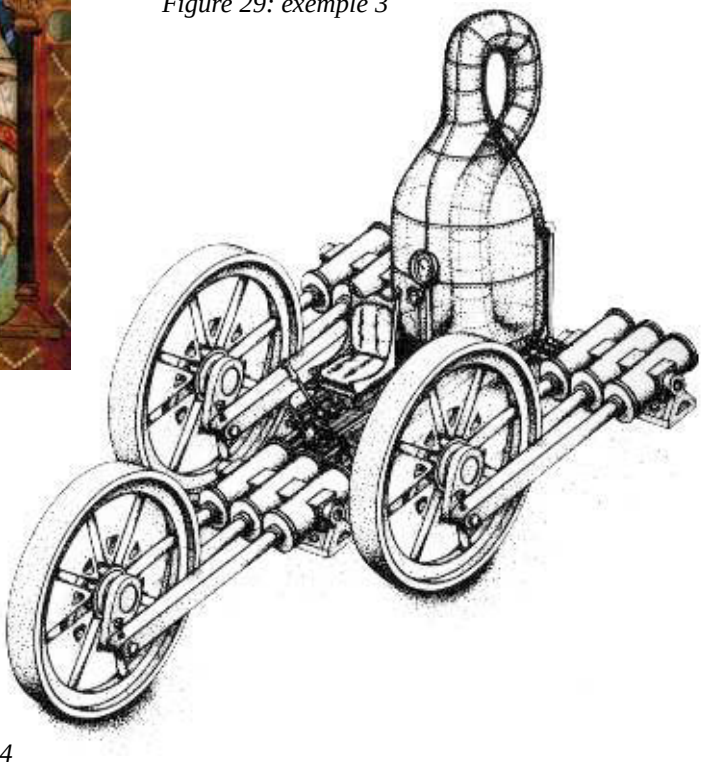


Figure 30: exemple 4

Un autre exemple, issu cette fois de la synthèse, avec des effets produits par la variation des paramètres de spécularité du modèle de Phong. Comme nous l'avons déjà dit, avec un peu d'habitude, il est en effet possible de simuler des effets de reflets qui seraient provoqués par des sources sphériques étendues (en faisant abstraction des ombres !) alors que le modèle n'utilise que des sources ponctuelles. Dans cet exemple comme dans les précédents se pose le problème du contrôle des effets obtenus dont la cohérence réside dans «l'intelligence» visuelle de l'opérateur et sa capacité à maîtriser les possibilités de son outil (pinceau, aérographe ou logiciel) et non dans une valeur intrinsèque du modèle. Ainsi, dans le *Bar des Folies Bergères* (figure [28] de la page précédente), Manet a-t-il volontairement déformé les lois de la réflexion et de la perspective pour nous montrer les reflets de la serveuse et du client dans le miroir en arrière-plan. Ces reflets devraient être pratiquement confondus avec les silhouetttes des deux personnages comme le montrent la réflexion des bouteilles et du comptoir conformes aux lois de l'optique.

La vision est un processus très complexe. Les illusions d'optiques, telles celles qui suivent, montrent combien il est ... illusoire de croire qu'elle permet une préhension «objective» et quasi-mécanique du réel.

Figure 31: croquis 1

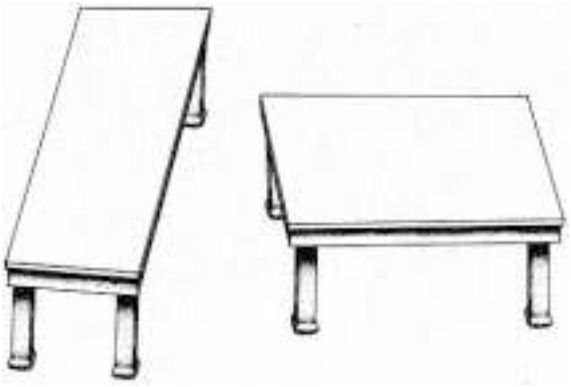
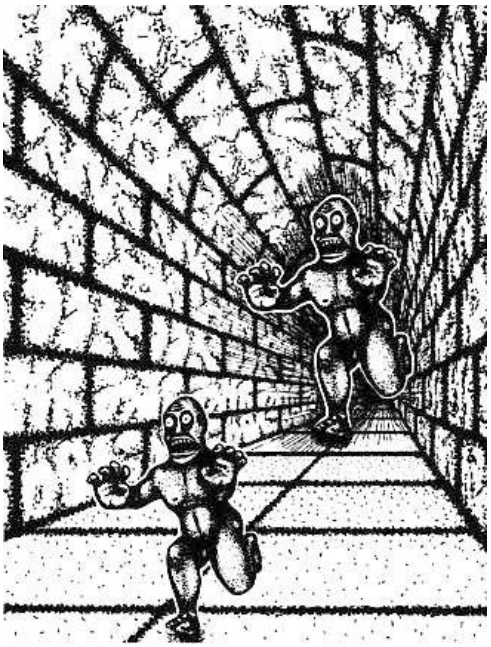


Figure 32: croquis 2



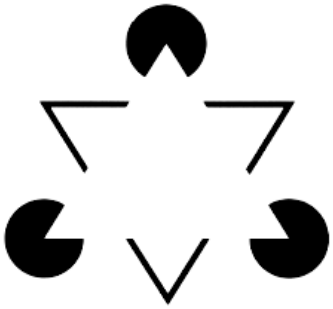
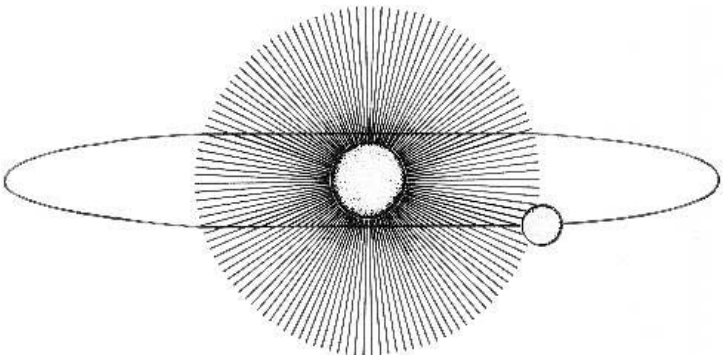


Figure 33: croquis 3

Figure 34: croquis 4



Ainsi dans le croquis 1, les plateaux des deux tables ont exactement la même forme et la même taille, c'est également le cas des deux personnages du croquis 2. La différence de taille que l'on attribue à ces deux éléments pourtant identiques est due à

l'interprétation volumique que nous donnons à ces deux dessins. Les lignes de fuites du croquis 2 et la perspective cavalière du croquis 1 incite à interpréter ces scènes en trois dimensions et donc à juger de leurs tailles selon nos habitudes du réel. Dans le croquis 3 nous percevons nettement les contours d'un triangle blanc (triangle de Kanizsa) alors que les éléments le suggérant sont très peu nombreux. Il semblerait que ce contour illusoire soit dû à deux aires du cerveau dont les cellules de l'une (aire V1) plus spécialisées dans la détection de formes ne perçoivent pas de contours pendant que l'autre (aire V2), plus analytique, «déduit» la présence de lignes [Zeki 92]. L'aire V2 provoquerait une rétroaction vers l'aire V1 et le cerveau homogénéiserait le résultat par l'adjonction du contour. Enfin, sans que nous ayons pu en trouver l'explication, les parties de l'ellipse du croquis 4 dans les rayons du «soleil» sont en fait parfaitement droites.

Voici un autre exemple plus subtil mais tout aussi révélateur du système visuel humain.

Figure 35: Une coquille monumentale dans cette édition de Libération qui a survécu aux relectures du rédacteur en chef, des secrétaires de direction aussi bien que des correcteurs

Il est en fait beaucoup plus simple que l'on ne croit souvent de contenter notre système visuel (l'exemple 3 de la page 73 par exemple est une impossibilité physique que l'on juge habituellement très convaincante, tout comme le *Bar des Folies Bergères* déjà évoqué). Comme l'explique Edholm [Edho 93], nous avons l'impression que notre vision est une perception directe du monde alors que ce que l'on voit résulte d'un processus très complexe de conjectures réalisées par le système visuel, conjectures qui mènent à une interprétation échappant bien souvent à la conscience. En l'occurrence, les exemples ci-dessus montrent bien à quel point le système visuel recherche un sens à toute scène qui lui est présentée, quitte à inventer des éléments qui font défaut à une certaine cohérence. Depuis 1920, un certain nombre de disciplines se sont développées pour tenter de comprendre l'élaboration de la perception. On peut citer, parmi les premières à avoir défriché le terrain, la *psycho-physiologie* qui considérait la perception comme une succession d'actes élémentaires dépendant uniquement de substrats physiologiques. Conceptions que la *Gestalthéorie* a contestées en montrant la nécessité d'une approche plus globale par la mise en évidence de l'émergence de «formes» c'est-à-dire la construction du sens par la perception de la totalité d'une situation qui dépasse la simple superposition précédente. Les psychologues de la forme ont ainsi dégagé un certain nombre de caractéristiques de l'organisation perceptive (voir l'ouvrage de référence de Guillaume [Guil 37] pour de plus amples détails). En voici quelques une :

- la proximité : des éléments géométriquement proches se regroupent (voir figure A)
- la ressemblance : des éléments ayant des caractéristiques communes ont tendance à être regroupés

- la symétrie : les figures symétriques sont perçues plus spontanément que les autres (voir figure B)
- la clôture : des figures qui semblent proches de formes connues simples ont tendance à être complétées (voir figure C).

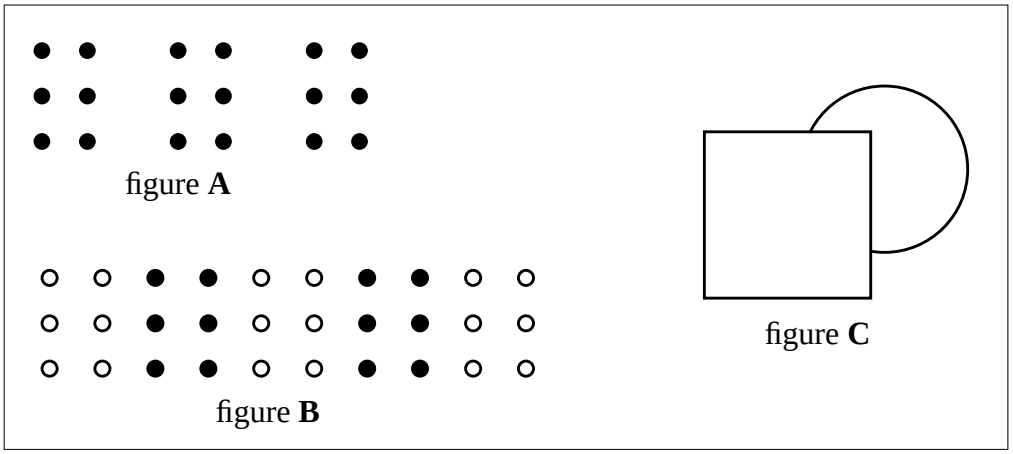


Figure 36: la figure A illustre la règle de proximité : nous percevons spontanément trois colonnes plutôt que trois lignes de 6 points. La figure B montre la règle de ressemblance où le regard tend à regrouper des éléments semblables (points noirs d'une part, points blancs d'autre part). La figure C illustre la clôture où l'on perçoit plutôt un carré sur un cercle alors qu'une multitude d'autres possibilités existent.

La *psychologie cognitive* a également investi le champ de la perception depuis une quarantaine d'années. Son approche est complémentaire de la gestalthéorie en considérant la perception dans une perspective plus globale comme un processus en trois étapes :

- une phase physiologique où seuls interviennent les stimulus et les caractéristiques du système sensoriel, ce dernier codant les informations de façon automatique.
- une deuxième phase où les informations sont regroupées en entités plus globales.
- la troisième phase interprète les informations sensorielles des deux premières phases pour identifier des objets ou événements à l'aide de connaissances antérieures.

Pour exemple des recherches menées dans le domaine de la perception visuelle, nous citerons les travaux de Ramachandran [Rama 88]. Cette étude s'attache à expliciter la façon dont nous percevons le relief à partir d'images bidimensionnelles : une gageure que l'oeil réalise quotidiennement. Elle montre tout d'abord que le système visuel «présuppose» habituellement que la lumière sensée éclairer la scène provient d'une source unique et que cette source est située en haut de la scène. Ensuite, lorsque les éléments d'ombrage sont ambigus, ce sont les éléments de contour qui sont utilisés pour résoudre l'ambiguïté. La frontière étant définie comme un changement abrupt de luminance. En d'autres termes, un changement de couleur mais de luminance uniforme n'aide pas à la résolution d'une ambiguïté. La reconstruction de la forme d'un objet résulte donc de la combinaison de l'ombrage et des contours. Le système visuel extrait un certain nombre de caractéristiques élémentaires (bords orientés, couleurs, direction de mouvement) dans les premières étapes de la vision. Il a été suggéré que le regroupement

d'objets dont nous avons parlé ci-dessus, ne peut avoir lieu que sur la base de ces caractéristiques élémentaires, regroupement qui permet, entre autre, de distinguer une forme d'un fond comme c'est le cas de façon assez spectaculaire dans l'image suivante.

Figure 37: Regroupements permettant la distinction forme/fond

Selon cette étude il semblerait que l'ombrage soit également une caractéristique élémentaire, groupable comme le montre la figure suivante. Sur cet exemple l'oeil ne distingue d'abord rien de particulier puis, après quelques secondes, des sphères émergent pour former un groupe clairement séparé du reste du dessin.

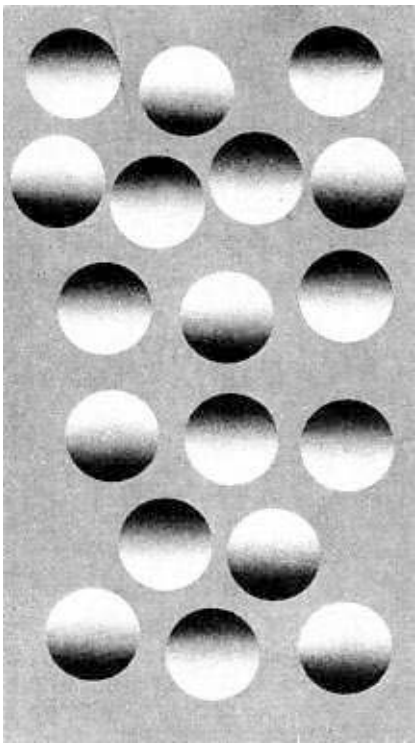


Figure 38: Regroupements d'objets selon l'ombrage

Le même phénomène s'observe sur la figure suivante : il semble que rien n'apparaît de prime abord. Pourtant pivotez la figure de 90 degrés et très rapidement des sphères apparaîtront sur un fond d'objets concaves (ou plats selon l'observateur) groupés en triangles montrant ici encore l'hypothèse de la source unique effectuée par le système visuel. Lorsque la règle de la source unique ne peut être satisfaite, une autre règle semble entrer en jeu qui tend à assimiler des formes d'orientations identiques à des surfaces parallèles.

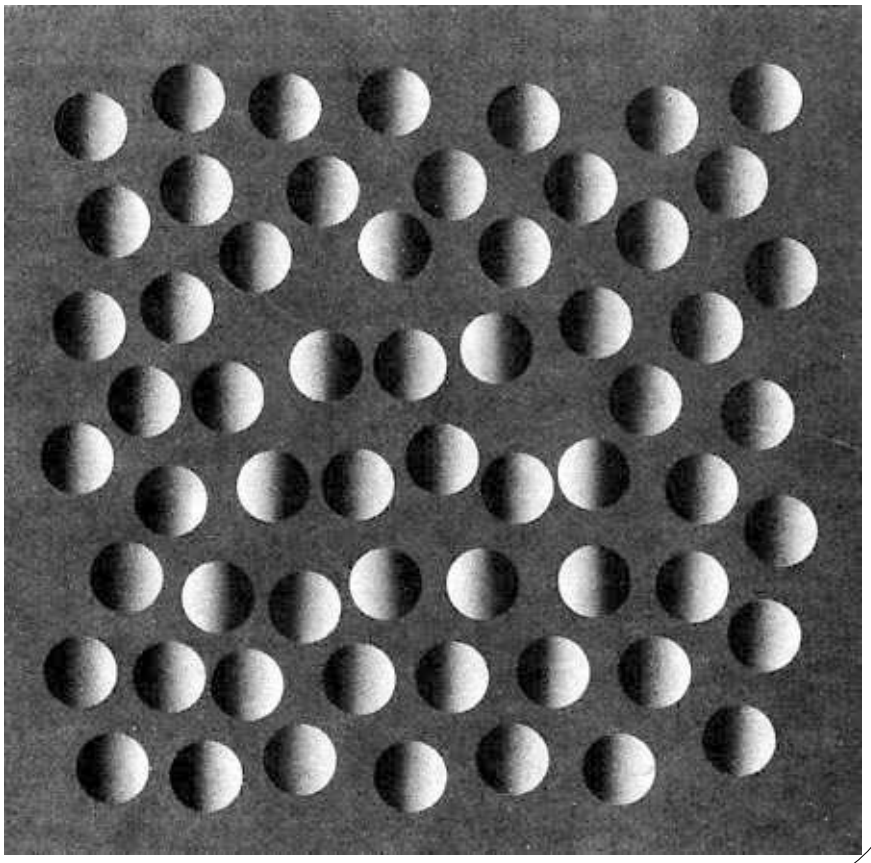


Figure 39: Phénomène de regroupement dû à l'hypothèse de la source unique

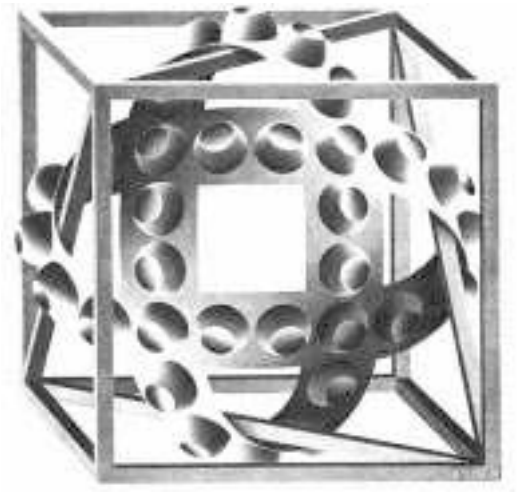


Figure 40: Perception équivoque du relief

Enfin une dernière constatation provenant de cette étude est que l'interprétation d'une image n'est pas toujours univoque, loin s'en faut, comme l'a magistralement saisi Escher dans l'oeuvre ci-contre. En effet, si les parties saillantes du ruban aux quatre points cardinaux ne prêtent guère à confusion, en revanche les éléments de la partie centrale peuvent-être interprétés comme des creux ou des bosses selon que l'on se convainc que la lumière provient plutôt du haut ou du bas.

Plus troublant encore ce dessin que vous percevrez sans doute comme une scène d'intérieur d'une pièce dont une fenêtre donne sur l'extérieur alors que des Africains de l'Est y voient plutôt une scène d'extérieur où une femme porte un bidon sur la tête. La perception visuelle dépasse donc de loin les aspects purement physiologiques déjà fort complexes pour dépendre également de la culture, des représentations sociales (on juge par exemple des différences de tailles entre objets ou personnes relativement à l'importance qu'on leur donne presqu'indépendamment de leur dimension) ainsi que de l'inconscient.

Figure 41: Différentes interprétations possibles selon la culture d'origine de l'observateur

II.4.2 Conclusions

Une image correcte pour le système visuel ne signifie donc rien quand à sa cohérence avec la réalité, de même d'ailleurs qu'une image incorrecte, tant il est facile d'introduire des paramètres incohérents dans le meilleur des modèles. Pire peut-être, certaines images parfaitement en accord avec la réalité peuvent à tel point choquer notre sens visuel qu'elles seront rejetées parce que jugées comme incohérentes, irréalistes, comme c'est le cas dans cette photographie d'une scène réelle reprise ci-dessous.

Figure 42: Impossible réalité

L'astuce de cette image réside dans l'angle de vue très particulier sous lequel ont été pris, non pas un objet, comme nous l'interprétons spontanément, mais deux objets disjoints. Que l'on se souvienne également de la réaction d'incompréhension et de rejet devant certaines représentations de visages aux yeux verticalement décalés réalisées par Picasso alors qu'elles correspondent bien à la vision qu'ont deux amoureux de leur visage lorsqu'ils s'embrassent.

La démarche purement photoréaliste, finalement assez répandue, fait, la plupart du temps, l'économie de toutes les études et théories psychovisuelles. Ce faisant elle ne se donne pas les moyens du contrôle de ses résultats et de leur vérification, en somme de la compréhension profonde de son fonctionnement et donc de sa maîtrise. En forçant quelque peu le trait, adopter cette démarche conduit sans doute à inventer une nouvelle forme de peinture mais sûrement pas une science. Il nous semble donc que la validité d'un modèle ne peut résider que dans sa cohérence interne, intrinsèque, avec la physique et les mathématiques et que les images qu'il peut produire ne sont finalement un gage de rien d'autre que de la capacité de notre système visuel à lui donner une interprétation «sensée» ou non. Ce qui ne veut pas dire que l'on puisse, in fine, se contenter de la concordance physique. Au contraire, l'apport des sciences cognitives, des neurosciences, de la biologie et, avec une égale importance, des arts, est indispensable à une compréhension globale de tous les phénomènes à l'oeuvre. Qui plus est, non seulement nous serons mieux à même de réaliser des images fidèles (ou infidèles en connaissance de cause, si bon nous semble), mais nous gagnerons au passage des simplifications et des gains en rapidité que la physique et les mathématiques ne pourront jamais nous apporter d'eux-mêmes. La synthèse ne peut, en fait, se passer de la compréhension psychovisuelle. Il est, à notre sens, indispensable et urgent, de développer des programmes de recherches trans-disciplinaires qui concilient les deux domaines en commençant par inclure à part égale dans l'enseignement de base de l'infographie avec les matières traditionnelles (algorithmique, physique et mathématiques), la psychologie et la biologie recoupant le domaine.

Le physico-réalisme restera néanmoins la base de la synthèse parce qu'il est le seul moyen de préhension du réel extérieur à nos sens et constitue, de ce fait, la base incontournable par laquelle on peut comparer les résultats de nos calculs à la réalité et obtenir une référence à partir de laquelle on peut juger d'approximations et de simplifications.

Résolution par Monte-Carlo

La nécessité du hasard, tel pourrait être le titre de ce chapitre tant nous allons user, pour ne pas dire abuser, des propriétés de l'aléatoire. La complexité cachée de l'équation de rendu, et plus encore la nature hautement probabiliste des phénomènes naturels qu'elle décrit, est en effet telle, que l'approche statistique nous semble incontournable. Les méthodes de Monte-Carlo, au coeur des solutions que nous proposons, nous permettront d'appréhender cette complexité et fourniront quelques outils robustes pour la circonvenir. La méthode que nous allons présenter est un tracé de chemins depuis l'oeil vers les sources. Caméléon, poire et tournesol sur fond de frayeurs russes sont au menu, bon appétit !

III.1 Le problème à résoudre

III.1.1 Positionnement

Comme nous l'avons vu précédemment les méthodes de Monte-Carlo sont très générales et peuvent s'appliquer à des domaines très variés. Leur efficacité dépend donc de façon cruciale de leur adaptation spécifique au problème précis que l'on souhaite résoudre. Aussi allons-nous détailler ici le problème qui nous préoccupe.

Tout d'abord, les scènes que nous souhaitons modéliser se composent d'objets lumineux et d'objets opaques. La surface de ces derniers suit les critères que nous avons établis pour définir notre fonction de transfert d'énergie. Nous verrons un peu plus loin que nos méthodes de résolution ne sont pas limitées à ces fonctions particulières, mais qu'elles sont tout à fait applicables à d'autres modèles pour peu qu'ils aient des caractéristiques de forme semblables. Les objets lumineux seront exclusivement constitués par des sources lumineuses sphériques à émission uniforme.

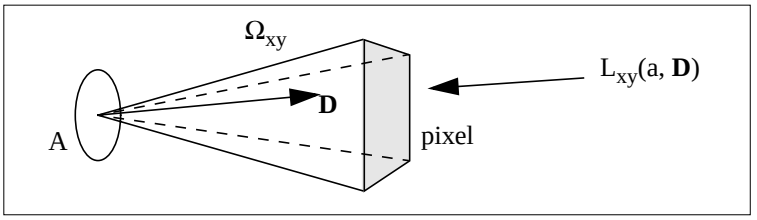
Ceci exclut, pour l'instant, les objets transparents d'une part, ainsi que les objets phosphorescents ou luminescents d'autre part, ces derniers plus pour une question de modélisation que de validité de la méthode. Le seul objet transparent autorisé pour

l'instant est le milieu ambiant, supposé d'indices optiques constants, où sont plongés tous les autres objets. Les milieux participatifs sont donc exclus.

Ces caractéristiques étant posées, nous souhaitons déterminer ce que voit l'oeil au travers de chaque pixel. Comme nous l'avons déjà dit, l'oeil est sensible à la luminance, c'est donc la luminance moyenne traversant un pixel et atteignant l'oeil que l'on cherche à calculer. Formellement cette luminance, $\bar{L}_{xy}$, s'exprime par :

$$\bar{L}_{xy} = \frac{1}{A\Omega_{xy}} \int_A \int_{\Omega_{xy}} L_{xy}(a, \mathbf{D}) d\omega da$$

A est l'aire de la pupille projetée perpendiculairement à la direction **D**.
D est une direction définie par un point de la pupille et un point du pixel
da aire projetée élémentaire
Ω_{xy} est l'angle solide sous-tendu par le pixel (x, y)
$L_{xy}(a, \mathbf{D})$ est la luminance suivant la demi-droite (a, **D**), a étant un point de la pupille



Nous supposerons que l'on peut faire l'approximation suivante :

$$\int_A \int_{\Omega_{xy}} L_{xy}(a, \mathbf{D}) d\omega da \approx A \int_{\Omega_{xy}} L_{xy}(O, \mathbf{D}) d\omega$$

O est le centre de la pupille

d'où l'on tire : $\bar{L}_{xy} \approx \frac{1}{\Omega_{xy}} \int_{\Omega_{xy}} L_{xy}(O, \mathbf{D}) d\omega = \frac{\angle(O, X, Y)}{\Omega_{xy}}$

Chacune des luminances élémentaires composant cette intégrale, arrive de la scène et s'exprime de deux façons différentes selon leur origine. Cette luminance est soit issue d'une source, ce que l'on exprimera par :

$$L_{xy}(O, \mathbf{D}) = S(O, \mathbf{D})$$

$S(O, \mathbf{D})$ représente la luminance dans la direction —**D** de la première source intersectée par la demi-droite (O, **D**)

, soit provient de la luminance renvoyée par le premier point d'intersection de (O, **D**) avec la scène, c'est-à-dire, en utilisant la définition de la FDRB ((22) §I.4.2) :

$$L_{xy}(O, \mathbf{D}) = \int_{\Omega_P} f_r(\mathbf{D}, \mathbf{R}) L(P, \mathbf{R}) \cos\angle(\mathbf{D}, \mathbf{N}) d\omega$$

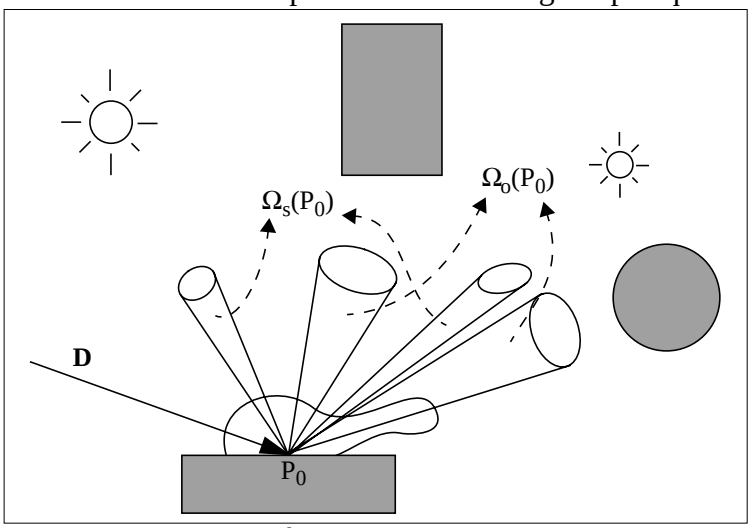
P est le premier point d'intersection entre la demi-droite (O, **D**) et la scène
Ω_P est l'hémisphère entourant P
R est une direction issue de P et dirigée dans Ω
$f_r(\mathbf{D}, \mathbf{R})$ est la fdrb au point P
N est la normale au point P

Ce que nous écrivons plus simplement :

$$L_{xy}(O, \mathbf{D}) = \int_{\Omega_P} \tau(\mathbf{D}, P, R) L(P, \mathbf{R}) d\omega$$

τ est la *fonction de transfert* au point P : $\tau(\mathbf{D}, P, \mathbf{R}) = f_r(\mathbf{D}, \mathbf{R}) \cos\angle(\mathbf{D}, \mathbf{N})$

Où l'on retrouve une expression classique très proche de l'*équation de rendu*. Néanmoins cette formulation cache la complexité de cette intégrale puisqu'elle est récursive.



Ainsi si on développe jusqu'au deuxième rebond, obtient-on :

$$L_{xy}(O, \mathbf{D}) = \int_{\Omega_s(P_0)} \tau(\mathbf{D}, P_0, \mathbf{R}) S(P_0, \mathbf{R}) d\omega + \int_{\Omega_o(P_0)} \tau(\mathbf{D}, P_0, \mathbf{R}_1) d\omega_0 \int_{\Omega_{P_1}} \tau(\mathbf{R}_1, P_1, \mathbf{R}_2) L(P_1, \mathbf{R}_2) d\omega_1$$

$\Omega_s(P)$ désigne l'ensemble des directions issues de P qui intersectent une source
$\Omega_o(P)$ désigne l'ensemble des directions issues de P qui intersectent un objet opaque

Et ainsi de suite, pour finalement obtenir une somme potentiellement infinie d'intégrales de taille quelconque :

$$L_{xy}(O, \mathbf{R}_0) = \sum_{n \geq 0} \int_{\Omega_o(P_0)} d\omega_0 \int_{\Omega_o(P_1)} d\omega_1 \ldots \int_{\Omega_s(P_{n-1})} d\omega_n \tau(\mathbf{R}_0, P_0, \mathbf{R}_1) \ldots \tau(\mathbf{R}_{n-1}, P_{n-1}, \mathbf{R}_n) S(P_{n-1}, \mathbf{R}_n) \quad (32)$$

Une remarque s'impose d'ores et déjà au sujet de la forme même de cette équation. Même en simplifiant à outrance les fonctions de transfert, l'apparition des fonctions d'illumination (S) est totalement imprévisible, même pour des scènes assez simples. Il en résulte que la complexité même de cette somme est quasiment incompressible. La seule restriction qu'il y ait à sa taille provient des fonctions de transfert qui ne renvoient jamais l'intégralité de l'énergie qu'elles reçoivent. Aussi, dans toute scène un tant soit peu réelle, ne rencontrerons-nous jamais une taille d'intégrale infinie. Ceci étant, même en tenant compte d'une profondeur maximale au-delà de laquelle le calcul ne serait plus nécessaire, il est illusoire de penser que l'on pourra obtenir plus qu'une approximation et que celle-ci puisse être bon marché…

III.1.2 Processus markovien

Les méthodes de Monte-Carlo sont toutes basées sur le calcul d'une espérance. L'expression-même de l'espérance, pourrait-on dire, induit la méthode. Aussi, pour calculer l'équation (32), devons-nous d'abord établir l'espérance correspondante. Cette étape est indispensable et, pour ce faire, nous aurons recours aux chaînes de Markov comme un passage obligé. En effet, l'équation (32) décrit un système continu formé par les rebonds d'une espèce de photon «géométrique» dont l'énergie diminue progressivement au fil du temps. On considère chaque rebond comme un état du système dont la succession définit une échelle de temps discret : le premier point d'impact définit le premier instant, le deuxième impact définit le second, etc. Dans cet esprit la luminance n'est qu'une fonction particulière de ces états à un instant bien spécifique déterminé par l'émission du photon par une source. Le comportement de ce système forme un processus aléatoire markovien.

Par définition, une suite de variables aléatoires $\{x_n\}$ est une *chaîne de Markov* si elle est caractérisée par une *loi initiale* :

$$\boxed{P(x_o \in dx) = i(x)dx}$$

et si la *probabilité de transition* de l'état x_n à l'état x_{n+1} peut s'écrire sous la forme d'une probabilité conditionnelle ne faisant intervenir que ces deux variables i.e :

$$\boxed{P(x_{n+1} \in dy / x_n = x) = t(x, y)dy}$$

en d'autres termes :

$$P(x_{n+1} \in dy / x_n = x_n, \ldots, x_0 = x_0) = P(x_{n+1} \in dy / x_n = x_n)$$

Plus prosaïquement, il n'est pas nécessaire de connaître l'ensemble des termes de la suite $(x_0,\ldots,x_{n+1})$ pour déterminer la probabilité de passage entre l'avant dernier et le dernier état. Seule la connaissance de ces deux états est nécessaire et suffisante.

Une chaîne de Markov est dite *homogène* si la probabilité de transition est indépendante de l'instant auquel l'observation est faite. Formellement :

$$P(x_{n+1} \in dy / x_n = x) = P(x_1 \in dy / x_0 = x)$$

Une propriété très intéressante concerne les fonctions de suite markovienne. Si $\psi(x_0,\ldots,x_n)$ est intégrable, son espérance s'exprime par :

$$\boxed{E(\psi(x_0, \ldots, x_n)) = \int i(x_0)dx_0 \int t(x_0, x_1)dx_1 \ldots \int t(x_{n-1}, x_n)dx_n \psi(x_0, \ldots, x_n)} \quad (33)$$

Dans notre cas, la suite des couples formés des directions et des points d'impacts, $\{(\mathcal{R}_n, \mathcal{P}_n)\}$, est une chaîne de Markov homogène. La loi initiale, $i(\mathbf{R}_0, P_0)$, décrit les directions formant l'angle solide sous lequel l'oeil voit un pixel donné. Les probabilités de transition, $t((\mathbf{R}_i, P_i), (\mathbf{R}_{i+1}, P_{i+1}))$, sont uniquement fonction de la direction incidente $\mathbf{R}_i$ au point P_i d'une part et de la direction réfléchie $\mathbf{R}_{i+1}$ et du point d'intersection P_{i+1} de la demi-droite $(P_i, \mathbf{R}_{i+1})$ avec la scène d'autre part. Ces probabilités sont bien sûr directement conditionnées par les fonctions de transfert d'énergie. On remarquera au passage que la donnée de $(P_i, \mathbf{R}_{i+1})$ détermine entièrement $P_{i+1} = f(P_i, \mathbf{R}_{i+1})$, la probabilité de transition s'écrira donc :

$$P((\mathcal{R}_{n+1}, \mathcal{P}_{n+1}) \in d\omega dq / (\mathcal{R}_n, \mathcal{P}_n) = (r, p)) = t((r, p), (s, q))d\omega d\delta_{f(p, s)}(q)$$

Comme nous l'avons souligné en introduction, la luminance est une fonction de la suite $\{(\mathcal{R}_n, \mathcal{P}_n)\}$ que l'on impose d'arrêter à un instant aléatoire N lorsque $\mathbf{R}_N \in \Omega_S(P_{N-1})$, c'est-à-dire lorsque l'on rencontre une source. Si l'on note

$L((\mathbf{R}_{0xy}, P_{0xy}),\ldots,(\mathbf{R}_N, P_N))$ la luminance «élémentaire» obtenue après N rebonds en partant du point (x, y) de l'écran, la quantité que l'on cherche à calculer $\angle(O, X, Y)$ a la forme suivante :

$$\angle(O, X, Y) = \sum_{N\geq 0} \int_{\Omega_{XY}} \int_{\Omega_o(P_1)} \cdots \int_{\Omega_o(P_{N-2})} \int_{\Omega_s(P_{N-1})} d\omega_{XY} d\omega_1 \ldots d\omega_A L((\mathbf{R}_{0XY}, P_{0XY}), (\mathbf{R}_1, P_1), \ldots, (\mathbf{R}_N, P_N))$$

L'idée que nous avons est d'utiliser un échantillonnage d'importance qui permette de suivre au mieux les variations de cette somme, nous introduisons donc cet échantillonnage sous la forme suivante :

$$\angle(O, X, Y) = \sum_{N\geq 0} \int_{\Omega_{XY}} i(\mathbf{R}_{0XY}, P_{0XY}) d\omega_{XY} \int_{\Omega_o(P_1)} t((\mathbf{R}_{0XY}, P_{0XY}), (\mathbf{R}_1, P_1)) d\omega_1 \ldots$$
$$\ldots \int_{\Omega_S(P_{N-1})} d\omega_N t((\mathbf{R}_{N-1}, P_{N-1}), (\mathbf{R}_N, P_N)) \frac{L((\mathbf{R}_{0XY}, P_{0XY}), (\mathbf{R}_1, P_1), \ldots, (\mathbf{R}_N, P_N))}{i(\mathbf{R}_{0XY}, P_{0XY}) t((\mathbf{R}_{0XY}, P_{0XY}), (\mathbf{R}_1, P_1)) \ldots t((\mathbf{R}_{N-1}, P_{N-1}), (\mathbf{R}_N, P_N))}$$

Expression que l'on va pouvoir exprimer sous forme d'espérance par l'utilisation de (33) :

$$\angle(O, X, Y) =$$
$$\sum_{N\geq 0} E\left(\frac{L((\mathbf{R}_{0XY}, P_{0XY}), (\mathbf{R}_1, P_1), \ldots, (\mathbf{R}_N, P_N))}{i(\mathbf{R}_{0XY}, P_{0XY}) t((\mathbf{R}_{0XY}, P_{0XY}), (\mathbf{R}_1, P_1)) \ldots t((\mathbf{R}_{N-1}, P_{N-1}), (\mathbf{R}_N, P_N))} \mathbf{1}_{\Omega_{XY}}(\mathbf{R}_{0XY}, P_{0XY}) \ldots \mathbf{1}_{\Omega_S(P_{N-1})}(\mathbf{R}_N, P_N) \right)$$
$$= E\left(\frac{L((\mathcal{R}_{0XY}, \mathcal{P}_{0XY}), (\mathcal{R}_1, \mathcal{P}_1), \ldots, (\mathcal{R}_N, \mathcal{P}_N))}{i(\mathcal{R}_{0XY}, \mathcal{P}_{0XY}) t((\mathcal{R}_{0XY}, \mathcal{P}_{0XY}), (\mathcal{R}_1, \mathcal{P}_1)) \ldots t((\mathcal{R}_{N-1}, \mathcal{P}_{N-1}), (\mathcal{R}_N, \mathcal{P}_N))} \right)$$

L'idéal serait que le dénominateur it…t soit égal, à un facteur constant près, au produit des fonctions de transferts auquel cas la variance de l'estimation serait nulle s'il n'y avait qu'une source. Les fonctions de transferts étant trop complexes, nous allons approcher ce produit par différentes méthodes inspirées de Monte-Carlo.

III.2 Monte-Carlo ou la philosophie du caméléon

III.2.1 Principes

L'idée de base des méthodes de Monte-Carlo (Cf §II.2.2) consiste à construire un échantillonnage qui mime les variations du processus à évaluer, ici une fonction à intégrer. Bien entendu, plus ce mimétisme est conforme à la fonction de base, plus forte sera la réduction de variance résultante, plus rapide sera la convergence de l'approximation vers la valeur exacte. Un échantillonnage, même grossièrement ressemblant, fournira toujours un meilleur résultat que la méthode brutale, c'est sur ce principe que nous avons construit notre méthode de résolution.

La fonction que nous souhaitons intégrer est le produit d'une fonction de transfert par une luminance incidente. Ces deux composantes sont en fait de nature assez différentes. La fonction de transfert est connue au moment de l'intersection d'un rayon avec la scène alors que la luminance n'est évaluable que lorsque le processus s'arrête sur une source. Les fonctions de transfert sont potentiellement accessibles dès la modélisation de la scène, mais elles restent d'une grande complexité puisqu'elles dépendent, dans notre cas, de pas moins de cinq paramètres. Il n'est donc guère envisageable d'en tirer quelque enseignement que ce soit à ce moment-là. En revanche, dès l'intersection réalisée, les paramètres libres sont au nombre de trois (deux sans la longueur d'onde), il est alors possible d'en dégager un certain nombre de caractéristiques. La luminance, quant à elle, n'est guère abordable qu'au travers de l'éclairement direct. Pour un point donné de la scène, on peut déterminer quelles sont les sources potentiellement visibles et celles qui

ne peuvent l'être. De ces constatations, nous allons déduire le comportement local probable de l'intégrale et construire une méthode d'échantillonnage mimant ce comportement.

La méthode de Monte-Carlo retenue est l'échantillonnage d'importance qui consiste à approximer l'intégrale $I = \int_0^1 f(x)dx$ en tirant une variable aléatoire x selon une densité p (aussi appelée *fonction de poids*) et à réaliser l'évaluation suivante :

$$\tilde{I} = \frac{1}{n}\sum_{i=1}^{n}\frac{f(x_i)}{p(x_i)} \qquad x_i \sim p(x)$$

> $\tilde{I}$ est une approximation de I qui converge vers la vraie valeur au fur et à mesure que le nombre d'échantillons, n, augmente, c'est pourquoi nous la notons avec un tilde
> $x_i \sim p(x)$ signifie que x_i est une réalisation de la variable aléatoire x de densité p

Dans notre cas, l'intégrale en un point s'exprime par :

$$L(\zeta_i, \alpha_i) = \int_\Omega \tau(\omega)L(\omega)d\omega = \int_0^{2\pi}\int_0^{\pi/2} \tau(\zeta_i, \alpha_i, \zeta_r, \alpha_r)L(\zeta_r, \alpha_r)\sin\zeta_r d\zeta_r d\alpha_r$$

La fonction de poids la plus simple et la plus naturelle consiste à prendre le sinus comme fonction de base. La densité associée correspond en effet exactement à l'échantillonnage uniforme de l'angle solide représenté par l'hémisphère. Cet échantillonnage représentera pour nous la méthode de Monte-Carlo de base à laquelle nous pourrons comparer d'autres méthodes plus sophistiquées. La fonction de base choisie, reste à la transformer en *densité* dont nous rappelons la définition :

$$g:\mathbb{R}^n \to \mathbb{R} \text{ est une densité si} \begin{cases} g \geq 0 \\ g \text{ est intégrable} \\ \int_{\mathbb{R}^n} g(x)dx = 1 \end{cases}$$

On notera G(x) sa *fonction de répartition* (dans le cas où g est fonction d'une seule variable) :

$$G(x) = \int_{-\infty}^{x} g(u)du$$

qui vérifie

$$\begin{cases} G \text{ est non décroissante} \\ G \text{ est continue} \\ \lim_{x \to -\infty} G(x) = 0 \\ \lim_{x \to +\infty} G(x) = 1 \end{cases}$$

Échantillonner X_i selon la densité g_x revient à tirer une variable aléatoire γ uniformément sur [0, 1] et à inverser sa fonction de répartition. Autrement dit :

$$\gamma \sim \mathcal{U}[0,1] \to X_i = G_x^{-1}(Y_i)$$

L'échantillonnage d'une variable aléatoire selon une *fonction de base* g_b se décompose donc en quatre étapes :

① Intégrer analytiquement g_b pour en déduire $G_b(x) = \int_{-\infty}^{x} g_b(u)du$

② Normaliser g_b pour construire sa densité associée g : $g(x) = \frac{g_b(x)}{G_b(+\infty)}$

③ Établir sa fonction de répartition $G_\chi(x)$ et son inverse $G_\chi^{-1}(y)$

④ Tirer Y_i uniformément et en déduire la réalisation X_i correspondante : $X_i = G_\chi^{-1}(Y_i)$

La première étape de la méthode ainsi que l'inversion de la fonction de répartition sont les deux points les plus délicats et limitent son application à des fonctions relativement simples. La dimension de la fonction de base peut également compliquer les calculs sauf lorsque les variables aléatoires $x_1,\ldots,x_n$ sont indépendantes. Dans ce cas en effet on a :

$$\begin{aligned} P(x_1 \le t_1, \ldots, x_n \le t_n) &= \prod_{i=1}^{n} P(x_i \le t_i) \\ &= \prod_{i=1}^{n} P(\gamma_i \le G_{x_i}(t_i)) \\ &= \prod_{i=1}^{n} P(G_{x_i}^{-1}(\gamma_i) \le t_i) \end{aligned} \qquad (34)$$

Ce qui permet d'échantillonner les X_i indépendamment les unes des autres via chacune de leur fonction de répartition.

III.2.2 Méthode minimale

La méthode d'échantillonnage la plus simple consiste donc à échantillonner uniformément l'angle solide décrit par l'hémisphère autour du point d'intersection. Ce qui conduit à choisir la fonction de base, $\tilde{m}_b$, suivante :

$$\tilde{m}_b(\zeta, \alpha) = \mathbf{1}_{[0,2\pi]}(\alpha)\sin\zeta$$

Sa normalisation se calcule facilement, ce qui permet d'obtenir la fonction de poids de cet échantillonnage $\tilde{m}(\zeta, \alpha)$:

$$\int_0^{2\pi}\int_0^{\pi/2} \tilde{m}_b(\zeta, \alpha)d\alpha d\zeta = 2\pi[-\cos\zeta]_0^{\pi/2} = 2\pi \Rightarrow \tilde{m}(\zeta, \alpha) = \frac{\tilde{m}_b(\zeta, \alpha)}{2\pi} \qquad (35a)$$

Les deux angles α et ζ étant indépendants, on peut les échantillonner indépendamment à l'aide de deux variables aléatoires uniformes, z et $\mathcal{A}$, elles-mêmes indépendantes. La fonction de répartition, $\tilde{M}_{z,\mathcal{A}}(\zeta, \alpha)$ s'écrit donc :

$$\tilde{M}_{z,\mathcal{A}}(\zeta, \alpha) = \tilde{M}_z(\zeta)\tilde{M}_{\mathcal{A}}(\alpha)$$

$$\tilde{M}_z(\zeta) = \int_0^{\zeta} \sin u du = 1 - \cos\zeta \qquad \tilde{M}_{\mathcal{A}}(\alpha) = \frac{1}{2\pi}\int_0^{\alpha} dt = \frac{\alpha}{2\pi}$$

d'où leur inverse :

$$\boxed{\begin{aligned} \tilde{M}_{\mathcal{Z}}^{-1}(z) &= \operatorname{acos}(1-z) = \zeta \\ \tilde{M}_{\mathcal{A}}^{-1}(a) &= 2\pi a = \alpha \end{aligned} \qquad (z, a) \sim \mathcal{U}[0,1] \times \mathcal{U}[0,1]} \tag{35b}$$

L'espérance estimée résultante est un estimateur sans biais et la luminance recherchée, L, et s'écrit :

$$L = E = \frac{1}{n}\sum_{r=1}^{n} \frac{\tau(\zeta_i, \alpha_i, \zeta_r, \alpha_r) L(\zeta_r, \alpha_r) \sin\zeta_r}{\tilde{m}(\zeta_r, \alpha_r)} = \frac{1}{n}\sum_{r=1}^{n} \frac{\tau(\zeta_i, \alpha_i, \zeta_r, \alpha_r) L(\zeta_r, \alpha_r) \sin\zeta_r}{\frac{\sin\zeta_r}{2\pi}}$$

Pour résumer, échantillonner selon la stratégie que nous venons de présenter consiste à choisir une direction d'échantillonnage en tirant aléatoirement ζ_r et α_r selon les fonctions de répartition inverses, $\tilde{M}^{-1}$, calculer la luminance réfléchie depuis cette direction, la pondérer par la fonction de poids $\tilde{m}$ et ajouter cette luminance pondérée aux autres contributions déjà calculées. Enfin diviser la somme des contributions par le nombre d'échantillons. On calcule donc :

$$\boxed{E = \frac{2\pi}{n}\sum_{r=1}^{n} \tau(\zeta_i, \alpha_i, \zeta_r, \alpha_r) L(\zeta_r, \alpha_r) \qquad \begin{cases} \zeta_r = \tilde{M}_{\mathcal{Z}}^{-1}(z) \\ \alpha_r = \tilde{M}_{\mathcal{A}}^{-1}(a) \end{cases}}$$

Cet échantillonnage uniforme est identique à celui utilisé par Lange ([Lang 91]) et constitue la méthode de base dont nous pourrons comparer les résultats avec les méthodes plus fines que nous allons élaborer dans les paragraphes suivants.

III.3 La stratégie de la poire

Même si la majorité des matériaux n'a pas un comportement qui se résume à une combinaison linéaire entre diffus et spéculaire, on peut néanmoins adopter cette décomposition pour caractériser schématiquement le fonctionnement de la FDRB. Ainsi, la plupart des rayonnements incidents seront réfléchis plus ou moins uniformément alors qu'une minorité, appartenant à un cône plus ou moins large, seront réfléchis plus fortement. Schématiquement donc, la FDRB se comporte comme une poire plus ou moins aplatie ou allongée selon la force relative des comportements diffus et spéculaire. Le sommet de la poire étant plus ou moins penché selon l'angle d'observation.

La forme de cette poire fournira en fait la densité avec laquelle la vraie FDRB sera échantillonnée. Décrire la densité d'une poire n'est pas une mince affaire, encore moins la fonction de répartition correspondante, aussi avons-nous simplifié cette forme en deux éléments : une partie diffuse d'une part et un cône spéculaire d'autre part, échantillonnés chacun séparément. Le mélange de ces deux distributions se faisant aléatoirement en respectant leur importance respective.

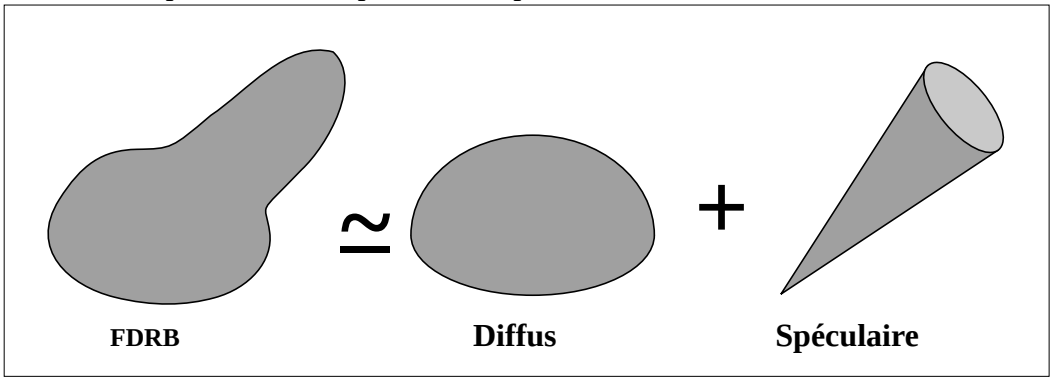


Figure 43: Approximation pour l'échantillonnage de la FDRB : la stratégie de la poire

III.3.1 Partie diffuse

Rappelons la fdrb que nous avons retenue précédemment ((31) §II.3) :

$$f_r(\zeta_i, 0, \zeta_r, \alpha_r) = k_d\frac{R_o}{\pi} + k_s\frac{GRP}{4\cos\zeta_i\cos\zeta_r} = k_d\rho_d + k_s\rho_s$$

La luminance réfléchie, L_r, s'exprimant par :

$$L_r = \int_{\Omega_p}\left(k_d\frac{R_o}{\pi} + k_s\frac{GRP}{4\cos\zeta_i\cos\zeta_r}\right)L(\zeta_r, \alpha_r)\cos\zeta_r d\omega$$

La luminance produite par la partie diffuse de notre modèle s'écrit donc :

$$\frac{k_d R_o}{\pi}\int_0^{2\pi}\int_0^{\pi/2} L(\zeta, \alpha)\cos\zeta\sin\zeta d\alpha d\zeta$$

La variation de cette luminance provient du produit cos x sin d'une part et de la luminance incidente $L(\zeta, \alpha)$ d'autre part. La luminance incidente étant trop difficile à évaluer a priori, c'est donc le produit que l'on retient comme base de notre échantillonnage :

$$\tilde{d}_b(\zeta, \alpha) = \mathbf{1}_{[0,2\pi]}(\alpha)\cos\zeta\sin\zeta = \mathbf{1}_{[0,2\pi]}(\alpha)\frac{\sin 2\zeta}{2}$$

Sa normalisation s'écrit :

$$\frac{1}{2}\int_0^{2\pi}\int_0^{\pi/2}\sin 2\zeta d\alpha d\zeta = \pi\left[\frac{-\cos 2\zeta}{2}\right]_0^{\pi/2} = \pi$$

La densité, autrement dit la fonction de poids, s'exprime simplement par :

$$\tilde{d}(\zeta, \alpha) = \mathbf{1}_{[0,2\pi]}(\alpha)\frac{\sin 2\zeta}{2\pi} \qquad (36a)$$

Les variables aléatoires étant indépendantes, la fonction de répartition $\tilde{D}_{z,\alpha}$ peut être scindée en deux :

$$\tilde{D}_z(\zeta) = \frac{1}{1}\int_0^{\zeta} \sin 2u\,du = \frac{1-\cos 2\zeta}{2}$$

$$\tilde{D}_\alpha(\alpha) = \frac{1}{2\pi}\int_0^{\alpha} dt = \frac{\alpha}{2\pi}$$

d'où l'on tire

$$\begin{aligned} &\tilde{D}_z^{-1}(z) = \frac{\arccos(1-2z)}{2} = \zeta \\ &\tilde{D}_\alpha^{-1}(a) = 2\pi a = \alpha \end{aligned} \qquad (z, a) \sim \mathcal{U}[0,1] \times \mathcal{U}[0,1] \qquad (36b)$$

Pour exemple d'application de cette méthode à un cas simple, voici l'espérance avec un matériau totalement diffus et donc la luminance réfléchie selon cet échantillonnage :

$$E = L_r = \frac{k_d R_o}{n} \sum_{i=1}^{n} L(\zeta_r, \alpha_r) \qquad \begin{cases} \zeta_r = \tilde{D}_z^{-1}(z) \\ \alpha_r = \tilde{D}_\alpha^{-1}(a) \end{cases}$$

III.3.2 Cône spéculaire

La luminance issue de la partie spéculaire de notre modèle est, grosso modo, une fonction décroissante de l'écart entre la direction réfléchie et la direction de réflexion spéculaire idéale. Nous approchons cette distribution par un cône plus ou moins englobant et fonction de l'ouverture du pic spéculaire. Contrairement aux apparences, l'échantillonnage d'un cône penché est assez complexe, un échantillonnage équivalent mais incomparablement plus simple, consiste à échantillonner un cône droit puis à réaliser une rotation sur les directions obtenues pour les diriger dans la direction spéculaire.

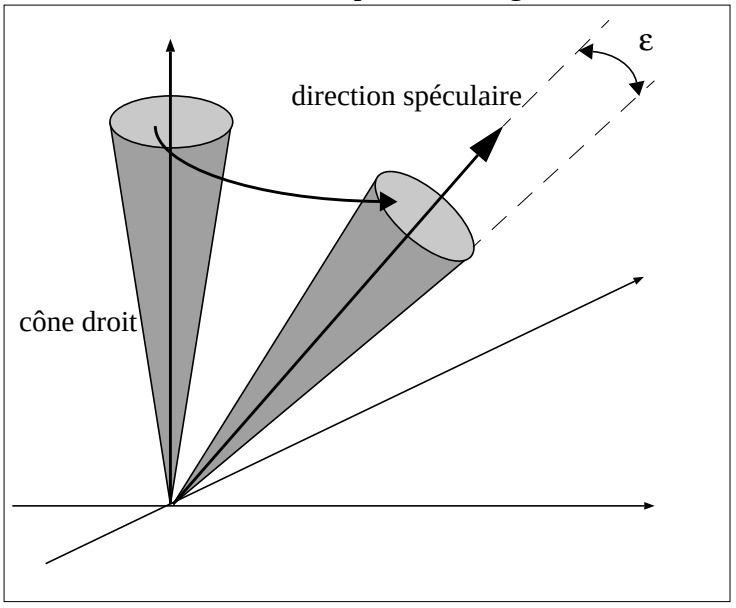


Figure 44: Échantillonnage équivalent d'un cône droit et d'un cône penché

Ces deux façons de procéder sont équivalentes puisqu'elles reviennent toutes deux à décrire l'angle solide sous-tendu par le cône qui est invariant par rotation. Le seul paramètre de cet échantillonnage est donc ε, la demi-étendue du cône. D'où la fonction de base décrivant ce cône :

$$\tilde{s}_b(\zeta, \alpha) = \mathbf{1}_{\text{Cône}}(\zeta, \alpha)\sin\zeta = \mathbf{1}_{[0,\varepsilon]}(\text{Rotation}(\zeta, \alpha))\sin\zeta$$

La normalisation correspondante est alors aisément calculée :

$$\int_0^{2\pi}\int_0^{\varepsilon} \sin\zeta d\alpha d\zeta = 2\pi(1-\cos\varepsilon)$$

La fonction de poids s'écrit donc : $\tilde{s}(\zeta, \alpha) = \dfrac{\sin\zeta}{2\pi(1-\cos\varepsilon)}$ (37a)

La fonction de répartition s'écrit, par la vertu de l'indépendance :

$$\tilde{S}_{\mathcal{Z}}(\zeta) = \frac{1}{1-\cos\varepsilon}\int_0^{\zeta} \sin u du = \frac{1-\cos\zeta}{1-\cos\varepsilon} \Rightarrow \tilde{S}_{\mathcal{Z}}^{-1}(z) = \text{acos}(1 - z(1-\cos\varepsilon)) = \zeta$$

$$\tilde{S}_{\mathcal{A}}(\alpha) = \frac{\alpha}{2\pi} \Rightarrow \tilde{S}_{\mathcal{A}}^{-1}(a) = 2\pi a = \alpha \qquad (37b)$$

À titre d'exemple, l'application de cette méthode sur un matériau uniquement spéculaire fournirait l'espérance suivante :

$$E = L = \frac{2\pi(1-\cos\varepsilon)}{n}\sum_{i=1}^{n} k_s\rho_s L(\zeta_r, \alpha_r)\cos\zeta_r \qquad \begin{cases} \zeta_r = \tilde{S}_{\mathcal{Z}}^{-1}(z) \\ \alpha_r = \tilde{S}_{\mathcal{A}}^{-1}(a) \end{cases}$$

III.3.3 Mélange diffus/spéculaire

Lorsqu'un matériau n'est ni parfaitement diffus, ni parfaitement spéculaire, comme c'est le cas en général, il faut que l'échantillonnage résultant mélange ces deux comportements dans les bonnes proportions pour traduire la force relative des deux composantes. Le mélange doit également bien sûr représenter une densité.

Si l'échantillonnage d'une fonction f était parfait, les deux fonctions, f et $\tilde{f}$, seraient égales à un facteur constant β près : $f = \beta\,\tilde{f}$. Ici, on a en fait :

$$s \approx \beta_s \tilde{s}_b \qquad \text{et} \qquad d \approx \beta_d \tilde{d}_b$$

$s = k_s\rho_s L(\zeta_r,\alpha_r)\cos\zeta_r\sin\zeta_r$ qui est la partie spéculaire élémentaire de notre modèle
$d = k_d\rho_d L(\zeta_r,\alpha_r)\cos\zeta_r\sin\zeta_r$ la partie diffuse élémentaire

ce qui implique :

$$\int_\Omega s(\omega)d\omega \approx \beta_s\int_\Omega \tilde{s}_b(\omega)d\omega \Rightarrow \beta_s \approx \int_\Omega s(\omega)d\omega \Big/ \int_\Omega \tilde{s}_b(\omega)d\omega$$

$$\int_\Omega d(\omega)d\omega \approx \beta_d\int_\Omega \tilde{d}_b(\omega)d\omega \Rightarrow \beta_d \approx \int_\Omega d(\omega)d\omega \Big/ \int_\Omega \tilde{d}_b(\omega)d\omega$$

Le mélange diffus/spéculaire permettant de maintenir la bonne proportion entre échantillonnage diffus et spéculaire conduit donc à la fonction de base

$$\tilde{b}_b = \beta_s \tilde{s}_b + \beta_d \tilde{d}_d$$

d'où la densité correspondante :

$$\tilde{b} = \frac{\beta_s \tilde{s}_b + \beta_d \tilde{d}_d}{\int_\Omega s(\omega) d\omega + \int_\Omega d(\omega) d\omega} = c_s \frac{\tilde{s}_b}{\int_\Omega \tilde{s}(\omega) d\omega} + c_d \frac{\tilde{d}_d}{\int_\Omega \tilde{d}(\omega) d\omega}$$

Pour que le choix entre l'échantillonnage diffus et l'échantillonnage spéculaire reflète le poids relatif des deux composantes, il faut et il suffit de choisir aléatoirement le diffus dans la proportion c_d et le spéculaire dans la proportion c_s. Bien sûr, les intégrales de s et de d sont inaccessibles, mais, ici encore, ce sont leur ordre de grandeur plus que leur valeur exacte qui nous intéresse. On pourrait envisager de les intégrer numériquement sur l'ensemble de leur domaine et d'utiliser le résultat de ces deux intégrations pour établir c_d et c_s mais l'opération est trop coûteuse, on choisit donc de ne considérer que leur moyenne dans le plan spéculaire (celui défini par la direction incidente et la normale à la surface) :

$$\int_\Omega d(\omega) d\omega \text{ est comparable à } \frac{k_d R_o}{\pi} \int_0^{\pi/2} \cos\zeta \sin\zeta d\zeta = \frac{k_d R_o}{2\pi} = m_1$$

$$\int_\Omega s(\omega) d\omega \text{ est comparable à } \frac{1}{n} \sum_{j=1}^{n} k_s \rho_s \cos\zeta_j \sin\zeta_j = m_2 \qquad \text{avec} \qquad \zeta_j = j\pi/n$$

d'où l'on tire :

$$c_s = \frac{m_2}{m_1 + m_2} \qquad c_d = \frac{m_1}{m_1 + m_2} \tag{38}$$

La fonction de poids de cet échantillonnage s'exprime donc simplement par :

$$c_s \tilde{s}(\zeta_r, \alpha_r) + c_d \tilde{d}(\zeta_r, \alpha_r) \tag{39}$$

L'espérance du mélange s'écrit alors :

$$E = \frac{1}{n} \sum_{i=1}^{n} \frac{\tau(\zeta_i, \alpha_i, \zeta_r, \alpha_r) L(\zeta_r, \alpha_r) \sin\zeta_r}{c_s \tilde{s}(\zeta_r, \alpha_r) + c_d \tilde{d}(\zeta_r, \alpha_r)}$$

$$\begin{cases} (\zeta_r, \alpha_r) = (\tilde{S}_z^{-1}(z), \tilde{S}_a^{-1}(a)) \text{ si } u \sim \mathcal{U}[0,1] < c_s \\ (\zeta_r, \alpha_r) = (\tilde{D}_z^{-1}(z), \tilde{D}_a^{-1}(a)) \text{ sinon} \end{cases}$$

En d'autres termes l'échantillonnage, soit selon la partie diffuse (eqs. (36a)-(36b)), soit selon le cône spéculaire (eqs (37a)-(37b)), est choisi aléatoirement selon les poids c_d et c_s.

III.4 Échantillonnage du pixel

Comme nous l'avons dit en début de chapitre, c'est la luminance moyenne sur le pixel, $\bar{L}_{xy}$, qui nous intéresse. C'est en effet à partir de cette luminance que l'on peut calculer la couleur du pixel :

$$X = k\int_{\vartheta}\bar{L}_{xy}(\lambda)\bar{x}(\lambda)d\lambda \qquad Y = k\int_{\vartheta}\bar{L}_{xy}(\lambda)\bar{y}(\lambda)d\lambda \qquad Z = k\int_{\vartheta}\bar{L}_{xy}(\lambda)\bar{z}(\lambda)d\lambda$$

Le facteur de normalisation, k, doit être fixé en fonction des caractéristiques de l'écran. Cette normalisation est basée sur la couleur blanche fournie lorsque les phosphores de l'écran sont excités au maximum. Les moniteurs sont en théorie calibrés sur des illuminants standard de la CIE pour que cette lumière blanche ait la même répartition spectrale que l'illuminant en question. Les données de base nécessaires à la conversion en XYZ et fournies par les constructeurs de moniteurs, sont constituées par la valeur de Y du blanc de référence, Y_b, et par l'illuminant de référence (D6500 par exemple) dont la répartition spectrale sera notée $E(\lambda)$. La normalisation consiste donc en :

$$Y_b = k\int_{\vartheta}\bar{L}_{xy}(\lambda)\bar{y}(\lambda)d\lambda = \frac{k}{A\Omega_{xy}}\int_{\vartheta}\int_{A}\int_{\Omega_{xy}}E_{xy}(\lambda,\mathbf{D})\bar{y}(\lambda)d\omega da d\lambda$$

$$= \frac{k}{A\Omega_{xy}}\int_{\vartheta}E_{xy}(\lambda,\mathbf{D})\bar{y}(\lambda)d\lambda\int_{A}\int_{\Omega_{xy}}d\omega da = k\int_{\vartheta}E_{xy}(\lambda,\mathbf{D})\bar{y}(\lambda)d\lambda$$

d'où l'on tire k :

$$\boxed{k = Y_b \Big/ \int_{\vartheta}E_{xy}(\lambda,\mathbf{D})\bar{y}(\lambda)d\lambda}$$

La couleur du pixel peut alors être calculée par :

$$Y = k\int_{\vartheta}\bar{L}_{xy}(\lambda)\bar{y}(\lambda)d\lambda = \frac{k}{A\Omega_{xy}}\int_{\vartheta}\int_{A}\int_{\Omega_{xy}}L_{xy}(\lambda,\mathbf{D})\bar{y}(\lambda)d\omega da d\lambda$$

$$\approx \frac{k}{\Omega_{xy}}\int_{\vartheta}\bar{y}(\lambda)\int_{\Omega_{xy}}L_{xy}(\lambda,\mathbf{D})d\omega d\lambda$$

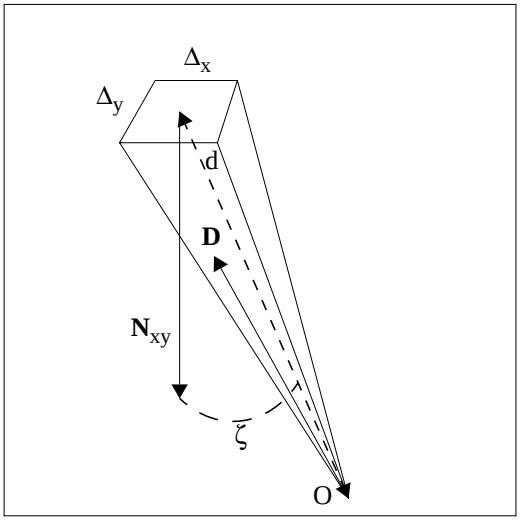
De même pour les composantes X et Y :

$$X \approx \frac{k}{\Omega_{xy}}\int_{\vartheta}\bar{x}(\lambda)\int_{\Omega_{xy}}L_{xy}(\lambda,\mathbf{D})d\omega d\lambda \qquad Z \approx \frac{k}{\Omega_{xy}}\int_{\vartheta}\bar{z}(\lambda)\int_{\Omega_{xy}}L_{xy}(\lambda,\mathbf{D})d\omega d\lambda$$

Nous cherchons donc à évaluer

$$\angle(O, X, Y) = \int_{\Omega_{xy}} L_{xy}(\lambda, \mathbf{D}) d\omega$$

ce que l'on peut écrire



$$\angle(O, X, Y) = \int_{\Delta_x}\int_{\Delta_y} L_{xy}(\lambda, \mathbf{D}) \frac{\cos\angle(\mathbf{D}, \mathbf{N}_{xy})}{d^2} dx dy$$

Les angles solides sous-tendus par les pixels sont extrêmement petits et l'on peut considérer que le rapport cos / d^2 est constant, d'où :

$$\angle(O, X, Y) = \frac{\cos\zeta}{d^2} \int_{\Delta_x}\int_{\Delta_y} L_{xy}(\lambda, \mathbf{D}) dx dy$$

Pour évaluer cette intégrale il suffit donc d'échantillonner uniformément l'aire du pixel et de calculer :

$$\angle(O, X, Y) = \frac{\Delta_x \Delta_y \cos\zeta}{d^2 n} \sum_{i=1}^{n} L_{xy}(\lambda, \mathbf{D}_i)$$

$$= \frac{\Omega_{xy}}{n} \sum_{i=1}^{n} L_{xy}(\lambda, \mathbf{D}_i)$$

On peut ainsi revenir aux composantes trichromatiques et constater que le calcul de l'angle solide n'est pas nécessaire :

$$X \approx \frac{k}{n} \int_{\vartheta} \bar{x}(\lambda) \left(\sum_{i=1}^{n} L_{xy}(\lambda, \mathbf{D}_i)\right) d\lambda \qquad Y \approx \frac{k}{n} \int_{\vartheta} \bar{y}(\lambda) \left(\sum_{i=1}^{n} L_{xy}(\lambda, \mathbf{D}_i)\right) d\lambda$$

$$Z \approx \frac{k}{n} \int_{\vartheta} \bar{z}(\lambda) \left(\sum_{i=1}^{n} L_{xy}(\lambda, \mathbf{D}_i)\right) d\lambda$$

III.5 Fiat lux

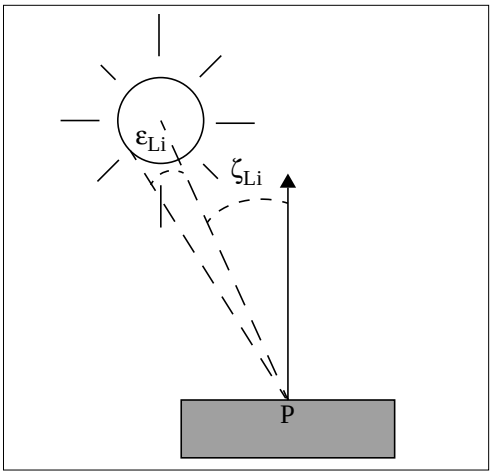
III.5.1 De l'attraction du tournesol ou comment tenir compte des sources

La stratégie de la poire approche la FDRB autant que faire se peut. Au passage, on peut noter que la méthode est extensible à d'autres modèles de transfert d'énergie pour peu qu'ils aient des parties diffuse et spéculaire suffisamment prononcées. Les seules parties spécifiques à notre modèle interviennent en effet dans le mélange diffus/spéculaire sous forme numérique, ce qui permet une adaptation facile à d'autres fonctions.

La variance de l'intégrale finale résulte de la forme de la FDRB conjuguée à la luminance incidente, il est donc indispensable d'en tenir compte. Au premier abord cette luminance ne peut être connue que par l'expérience, néanmoins et toujours dans l'optique des méthodes de Monte-Carlo, il est possible d'orienter l'échantillonnage vers des zones de l'hémisphère qui sont plus susceptibles que d'autres de contribuer fortement à l'intégrale finale.

Les fonctions de transfert sont des atténuateurs d'énergie, leur produit successif diminue progressivement l'énergie reçue initialement des sources. De plus, pour un hémisphère quelconque, la plus grosse contribution à l'énergie réfléchie provient, avec de fortes probabilités, de l'éclairement direct. Les exceptions à ce comportement sont au moins au nombre de deux. La première intervient lorsque le point d'intersection est à l'ombre d'une source et la seconde lorsque la source est en incidence rasante et ne coïncide pas avec le pic spéculaire.

Ces deux cas mis à part, la probabilité de contribution de la source à l'énergie réfléchie est d'autant plus grande que la source approche du zénith du point d'intersection et que le matériau est diffus. Ceci est d'autant plus vrai que la source est puissante et étendue (vue du point d'intersection).



L'échantillonnage doit donc tenir compte de l'étendue et de la puissance des sources ainsi que de leur position par rapport au point d'intersection.

Pour synthétiser l'ensemble de ces paramètres, nous avons choisi de rendre l'échantillonnage des sources proportionnel à l'expression suivante :

$$c_l = \frac{1}{2N_l \underset{j}{\mathrm{Max}}(P_{L_j})} \sum_{i=1}^{N_l} P_{L_i} (\mathrm{Max}(0, \cos\zeta_{L_i}) + \mathrm{Max}(0, \cos(\zeta_{L_i} - \varepsilon_{L_i})) \quad (40)$$

ζ_{Li} est l'angle que fait la normale avec le vecteur défini par P et le centre de la source i
ε_{Li} est la demi-étendue de la source i
N_l est le nombre de sources dans la scène
$P_{L_i} = \int_{\vartheta} P_i(\lambda) d\lambda$ est la puissance de la source i sur le visible

Ainsi c_l est d'autant plus important que les sources sont à l'aplomb du point d'intersection ($Max(0, \cos\zeta_i)$) et étendues ($Max(0, \cos(\zeta_{Li}-\varepsilon_{Li}))$) et qu'elles sont puissantes (multiplication par P_{Li}). Le choix du maximum permet de ne prendre en compte que les sources potentiellement visibles depuis le point d'intersection et d'écarter celles qui ne sont pas dans l'hémisphère autour du point. Le rapport avant la somme normalise c_l pour qu'il appartienne à [0, 1].

Ce choix a le triple avantage de rendre compte de la puissance, de l'étendue et de l'orientation des sources. L'échantillonnage des sources sera donc choisi proportionnellement à c_l, le choix d'une source particulière se faisant alors selon la même idée que précédemment, à savoir : privilégier les sources potentiellement les plus intéressantes. Le choix d'une source particulière se fait donc aléatoirement selon la densité suivante :

$$l_j = \frac{(Max(0, \cos\zeta_{L_j}) + Max(0, \cos(\zeta_{L_j} - \varepsilon_{L_j})))P_{L_j}}{\sum_{i=1}^{N_l}(Max(0, \cos\zeta_{L_i}) + Max(0, \cos(\zeta_{L_i} - \varepsilon_{L_i})))P_{L_i}}$$

Sa fonction de répartition s'écrivant simplement :

$$\mathcal{L}(j) = \sum_{i=1}^{j} l_j$$

Choisir une source selon cette densité revient à retenir la source k telle que :

$$\mathcal{L}(k-1) < u \le \mathcal{L}(k),\ u \sim \mathcal{U}[0,1],\ \mathcal{L}(0) = 0 \qquad (41)$$

La source étant choisie, nous échantillonnons une direction telle qu'elle soit uniformément répartie dans l'angle solide de cette source. Ce tirage s'apparente exactement à celui mis au point pour l'échantillonnage spéculaire, le cône spéculaire étant simplement remplacé par le cône sous-tendu par la source que l'on peut établir en connaissant la position et le rayon de la source.

Cette méthode tend à diriger l'échantillonnage vers la contribution lumineuse la plus forte, comme un tournesol s'oriente vers le soleil, d'où la dénomination de la méthode. On notera qu'avec cette méthode, aucun compte n'est tenu de l'absence de visibilité éventuelle de la source par le point (sauf pour les sources en-dessous de l'horizon). Cet inconvénient ne peut être corrigé qu'avec l'expérience, ce que nous présenterons dans un prochain paragraphe. En attendant de l'obtenir, la méthode du tournesol utilise au mieux les connaissances que l'on peut avoir a priori sur la scène, dès que le point d'intersection est connu.

III.5.2 Combinaison lumière/matière

Combiner l'échantillonnage des sources à celui de la FDRB revient à juger de l'importance de l'éclairement direct sur l'indirect. Ceci est indispensable mais n'est vraiment possible qu'a posteriori lorsque l'énergie incidente est connue, ce que nous ferons dans le paragraphe suivant. Dans un premier temps, il est néanmoins souhaitable d'orienter l'échantillonnage. Ainsi lorsque le comportement du matériau est dominé par son caractère spéculaire est-il préférable de choisir en priorité les directions spéculaires, quelle que soit la répartition des sources. Au contraire, lorsque le comportement diffus domine, il est préférable de considérer la conjonction source/fonctions de transfert la plus forte. Le schéma d'échantillonnage retenu est donc le suivant :

ChoixLumièreMatière

$u_1 \sim \mathcal{U}[0,1]$
SI $u_1 < c_s$ *ALORS* choisir l'échantillonnage spéculaire (eqs (37a)-(37b) §III.3.2)
SINON
    $u_2 \sim \mathcal{U}[0,1]$
    SI $u_2 < c_1$ *ALORS* choisir l'échantillonnage des sources (eq (41) §III.5.1)
    SINON choisir l'échantillonnage diffus (eqs (36a)-(36b) §III.3.1)

III.6 Échantillonnage a posteriori

Quel que soit le degré de sophistication de l'échantillonnage *a priori*, il ne peut que très imparfaitement tenir compte de la luminance incidente connue a posteriori. Il est donc indispensable de modifier l'échantillonnage a priori que nous avons présenté dans les paragraphes précédents pour tenir compte de l'énergie réellement réfléchie. Le but de cette modification étant de favoriser les directions incidentes qui contribuent le plus à l'énergie réfléchie.

Revenons à la somme infinie que nous avons définie en début de chapitre. Cette somme peut être décomposée de plusieurs façons et notamment selon la taille de ses constituants en groupant les termes ayant même longueur :

$$L(O, \mathbf{D}) = \int_{\Omega_s(P_0)} \tau(\mathbf{D}, P_0, \mathbf{R}) S(P_0, \mathbf{R}) d\omega + \int_{\Omega_o(P_0)} \tau(\mathbf{D}, P_0, \mathbf{R}_1) \int_{\Omega_s(P_1)} \tau(\mathbf{R}_1, P_1, \mathbf{R}_2) S(P_1, \mathbf{R}_2) d\omega_1 d\omega_0 + \ldots$$
$$= T_1 + T_2 + \ldots$$

Chaque T_n représente la luminance «directe», c'est-à-dire provenant de l'éclairement direct, réfléchie au $n^{ème}$ rebond. Le premier terme, T_1, désigne donc l'éclairement direct réfléchi après un rebond, le second l'éclairement direct renvoyé après deux rebonds et ainsi de suite. Nous parlerons ainsi de *niveaux de contribution* à l'intégrale finale.

Si le premier point d'intersection voit la source et est situé sur un objet plutôt diffus, la contribution de la luminance terminale T_1 à la somme finale a de fortes

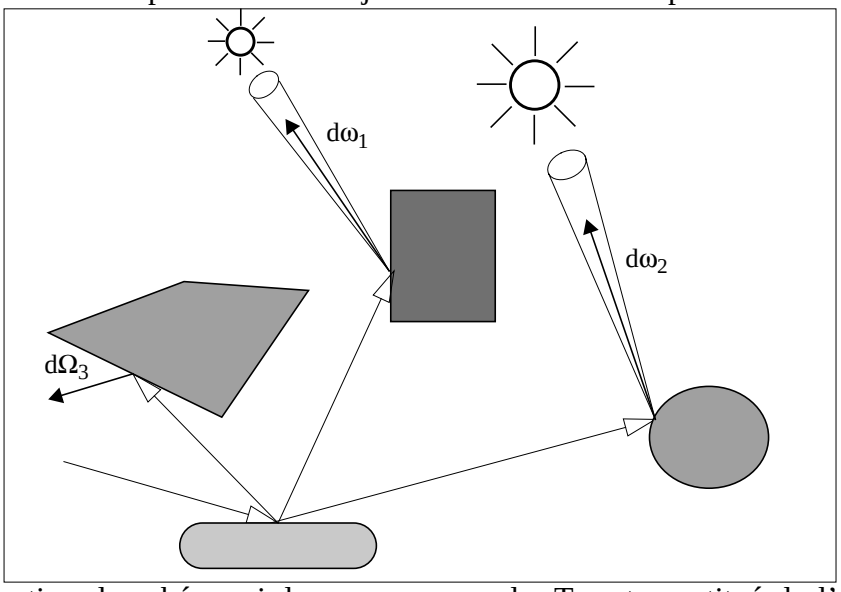
chances d'être dominante. En revanche, si le premier point d'intersection est à l'ombre ou situé sur un matériau spéculaire, la contribution la plus forte proviendra des niveaux supérieurs, probablement de T_2, etc. Le poids relatif de chaque niveau dans la somme finale fournit donc le meilleur échantillonnage possible. En effet, une contribution importante du niveau n dans la somme finale, traduit le fait que l'éclairement direct à ce niveau est prépondérant et donc qu'en moyenne l'échantillonnage à ce niveau doit privilégier les sources. Plus formellement, au rebond n, l'échantillonnage des sources doit être proportionnel à la densité N_n :

$$N_n = \frac{\int_{\Omega_o(P_0)} \dots \int_{\Omega_s(P_{n-1})} \tau(\mathbf{D}, P_0, \mathbf{R}_1)\dots\tau(\mathbf{R}_{n-1}, P_n, \mathbf{R}_n)S(P_n, \mathbf{R}_n)d\omega_0\dots d\omega_n}{L(O, \mathbf{D})}$$

$$= \frac{T_n}{L(O, \mathbf{D})}$$

$$= T_n / \sum_{j \geq 0} T_j$$

Il faut remarquer que ce raisonnement n'est valable qu'en moyenne car chaque T_i représente la contribution de l'ensemble des chemins de longueur i + 1. L'ensemble de ces chemins peut très bien porter sur des objets et des sources multiples.



Dans la situation du schéma ci-dessus par exemple, T_2 est constitué de l'éclairement direct provenant de $d\omega_1$ et $d\omega_2$ et a de fortes chances d'être prépondérant dans la somme finale puisque les sources sont masquées au premier rebond. S'il est prépondérant, notre stratégie consiste à échantillonner préférentiellement les sources au deuxième rebond. Il est clair que nous allons renforcer l'échantillonnage des sources sur $d\omega_1$ et $d\omega_2$, mais également sur $d\Omega_3$ si jamais l'échantillonnage sur le premier point d'intersection se dirige vers cette portion d'espace. En moyenne nous allons donc privilégier la contribution du niveau 2 et parfois nous perdrons peut-être du temps à échantillonner de «mauvaises» directions. Cette perte de temps n'est pas systématique puisque l'échantillonnage a priori que l'on utilise en conjonction avec cette stratégie de niveau ne choisira jamais d'échantillonner une source sur $d\Omega_3$ car c_1 est nul dans ce cas.

Le schéma d'échantillonnage que nous proposons est donc basé sur la densité N_n, que nous approximons au fil des calculs par ce que nous appelons un *estimateur de niveau*. Au niveau i, cet estimateur s'exprime par

$$\tilde{N}_i = \sum_{j=1}^{n_i} \tilde{T}_j^{(i)} \Big/ \sum_{k=1}^{n} \sum_{j=1}^{n_k} \tilde{T}_j^{(k)}$$

$\tilde{T}_j^{(i)}$ est la $j^{ième}$ luminance élémentaire provenant de l'éclairement direct au $i^{ème}$ rebond et atteignant l'observateur. Luminance calculée selon les méthodes déjà décrites.

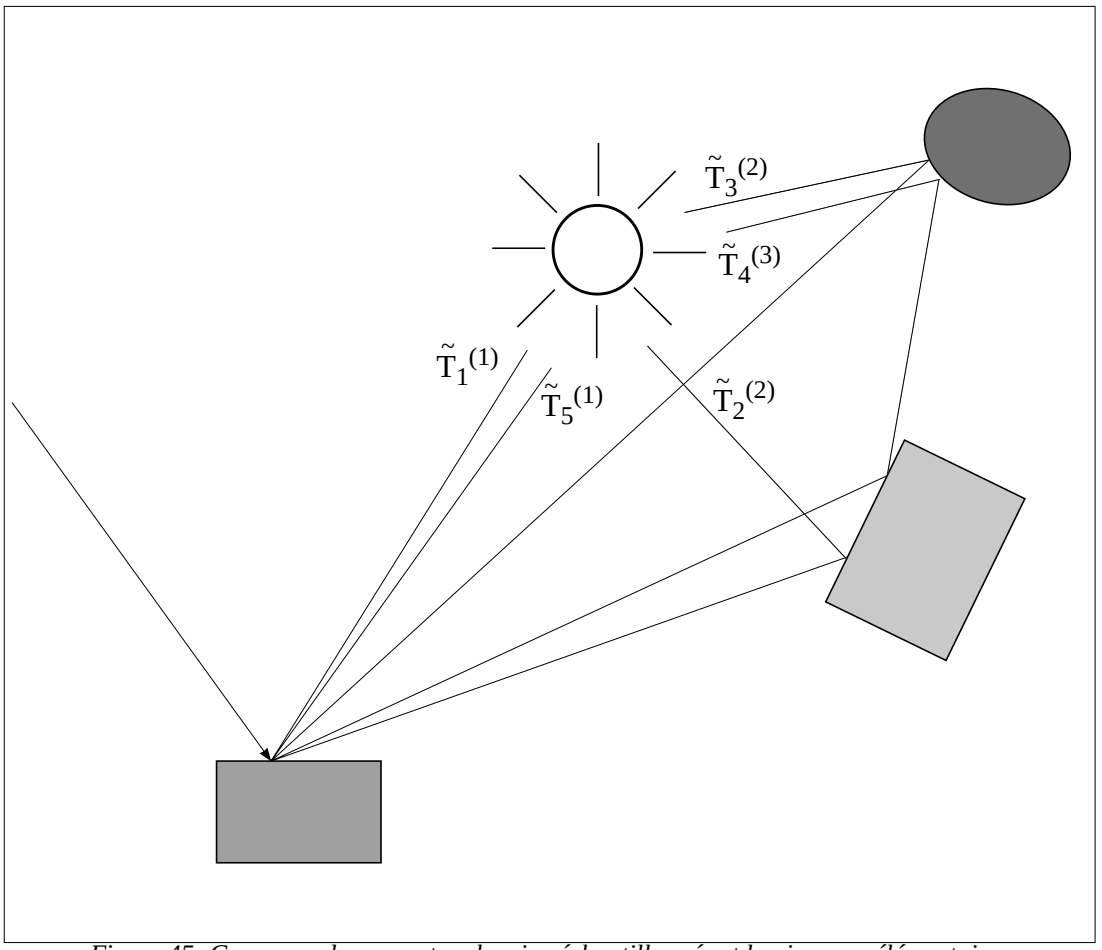


Figure 45: Correspondances entre chemins échantillonnés et luminances élémentaires

Sur l'exemple ci-dessus, nos estimateurs de niveau seraient les suivants :

$$\tilde{N}_1 = (\tilde{T}_1^{(1)} + \tilde{T}_5^{(1)}) \Big/ \sum_{k=1}^{3} \sum_{j=1}^{n_k} \tilde{T}_j^{(k)}$$

$$\tilde{N}_2 = (\tilde{T}_2^{(2)} + \tilde{T}_3^{(2)}) \Big/ \sum_{k=1}^{3} \sum_{j=1}^{n_k} \tilde{T}_j^{(k)}$$

$$\tilde{N}_3 = \tilde{T}_4^{(3)} \Big/ \sum_{k=1}^{3} \sum_{j=1}^{n_k} \tilde{T}_j^{(k)}$$

Cet échantillonnage a posteriori ne peut avoir lieu que progressivement, au fur et à mesure que la précision sur l'estimation par niveau devient fiable. Nous augmentons donc la fréquence d'utilisation de cet échantillonnage en fonction de l'effectif n_k du niveau k considéré, c'est-à-dire du nombre de chemins terminaux de longueur k+1. D'autre part, plus le chemin échantillonné s'allonge, plus il est souhaitable qu'il s'arrête, puisque son coût s'accroit en fonction de la profondeur. Il est donc souhaitable de privilégier les sources d'autant plus que le chemin actuel est long. Ce que l'on réalise en rendant le choix des sources proportionnel à :

$$t_k = \sum_{j=1}^{n_k} \tilde{T}_j^{(k)} \Big/ \sum_{i \geq k}^{n} \sum_{j=1}^{n_k} \tilde{T}_j^{(i)}$$

L'ensemble de ces caractéristiques a été réalisé par le schéma suivant au rebond k :

```
ChoixAPosteriori
    u1 ~ U[0,1]
    SI u1 < 1 - c_p^nk ALORS
        u2 ~ U[0,1]
        SI u2 < tk ALORS
            choisir l'échantillonnage des sources (eq (41) §III.5.1)
        SINON
            choisir l'échantillonnage de la poire (eq (39) §III.3.3)
        FINSI
    SINON
        choisir l'échantillonnage poire/tournesol (ChoixLumièreMatière)
    FINSI
```

c_p désigne la proportion d'utilisation de l'échantillonnage a priori par rapport à l'échantillonnage a posteriori et est fixé par l'utilisateur. Plus il est élevé (proche de 1), plus il faut prendre d'échantillons selon le schéma a priori avant que le schéma a posteriori ne s'enclenche. Ce paramètre n'a guère de sens en dehors de l'intervalle [0,9; 1[car alors il est peu probable que l'estimation des $\tilde{T}_j^{(i)}$ soit fiable.

III.7 Schéma global

L'estimation de (32) §III.1.1 que nous avons élaborée regroupe donc l'ensemble des schémas que nous avons présentés. À ceux-ci, il faut en ajouter deux autres.

Le premier est celui de la roulette russe dont nous avons déjà parlé. Le principe consiste à tronquer les chemins dont la contribution est négligeable, c'est-à-dire inférieure à un seuil C_o. Cette contribution est normalement calculée à partir des valeurs réelles de l'atténuation en chaque fourche du chemin (le produit des FDRBs). La nature du tracé de rayon que nous utilisons conduit à ne calculer cette atténuation qu'au cours de la remontée récursive et non pendant la descente (i.e durant l'élaboration des rebonds successifs). Utiliser la contribution réelle nous est donc impossible au moment opportun, nous utilisons donc le poids de chaque échantillon en lieu et place de leur contribution pour réaliser la roulette russe. Ces poids sont plutôt un majorant qu'un minorant de la

contribution réelle, la troncature que nous effectuons tend donc plutôt à poursuivre des chemins moins intéressants que leur poids ne le laisse paraître. Cette situation est préférable à l'inverse qui conduirait à tronquer injustement des chemins intéressants.

Le second schéma à ajouter à ceux que nous avons présentés, consiste à prévoir un degré de liberté minimum pour… l'imprévisible. Malgré toutes les précautions que nous avons prises pour que l'échantillonnage soit aussi fidèle que possible à l'équation (32) §III.1.1, il reste de faibles probabilités pour que des phénomènes marginaux échappent à notre méthode. Pour qu'il existe quelques chances de les capturer, nous instaurons un seuil minimal, c_{bruit}, en-dessous duquel l'échantillonnage uniforme de l'hémisphère est activé. De cette façon, si, contre toute attente, des contributions importantes sont obtenues de cette manière, la stratégie de niveau devrait être capable de les détecter pour leur donner une juste représentation.

Le schéma final que nous utilisons est donc la réunion de l'ensemble des stratégies que nous avons mises au point : poire, tournesol, stratégie de niveau, roulette russe et échantillonnage uniforme. L'algorithme de cette méthode au $k^{ième}$ rebond est donc le suivant :

ÉchantillonnageGlobal1

SI $u_1 < c_{bruit}$ *ALORS*
  $(\zeta, \alpha) \leftarrow$ échantillonnage uniforme (eq (35a)-(35b) §III.2.2)
SINON
  SI $u_2 < c_p^{n_k}$ *ALORS*
    SI $u_3 < c_s$ *ALORS*
      $(\zeta, \alpha) \leftarrow$ échantillonnage spéculaire (eqs (37a)-(37b) §III.3.2)
    SINON
      SI $u_4 < c_l$ *ALORS*
        $(\zeta, \alpha) \leftarrow$ échantillonnage des sources (eq (41) §III.5.1)
      SINON
        $(\zeta, \alpha) \leftarrow$ échantillonnage diffus (eqs (36a)-(36b) §III.3.1)
  SINON
    SI $u_6 < t_k$ *ALORS*
      $(\zeta, \alpha) \leftarrow$ échantillonnage des sources (eq (41) §III.5.1)
    SINON
      SI $u_7 < c_s$ *ALORS*
        $(\zeta, \alpha) \leftarrow$ échantillonnage spéculaire (eqs (37a)-(37b) §III.3.2)
      SINON
        $(\zeta, \alpha) \leftarrow$ échantillonnage diffus (eqs (36a)-(36b) §III.3.1)
  SI Poids(ζ, α) * ancien poids < C_o *ALORS*
    SI u_8 < proba d'arrêt *ALORS* tronquer le chemin actuel
    SINON nouveau poids = Poids(ζ, α) * ancien poids * (1 - proba d'arrêt)
  SINON nouveau poids = Poids(ζ, α) * ancien poids

avec $u_i \sim \mathcal{U}[0,1]$

Rappelons que les seuls paramètres à fixer par l'utilisateur sont c_{bruit}, pour lequel nous préconisons une valeur maximale de 0,1 , ainsi que c_p ($\in$ [0,9; 1[) et C_o. L'ensemble des autres est à calculer selon les méthodes décrites plus haut.

Le poids de l'échantillon (ζ, α) par lequel il faut pondérer chaque luminance à chaque rebond s'écrivant :

$$\begin{aligned}\text{Poids}(\zeta, \alpha) \;=\; & c_{\text{bruit}}\tilde{m}(\zeta, \alpha) + \\ (1 - c_{\text{bruit}})\Bigg[& c_p^{n_k}\Big\{ c_s\tilde{s}(\zeta, \alpha, \varepsilon_s) + c_d\Big[c_l \max_i \tilde{s}(\zeta, \alpha, \varepsilon_{L_i}) + (1 - c_l)\tilde{d}(\zeta, \alpha)\Big]\Big\} + \\ & (1 - c_p^{n_k})\Big\{ t_k \max_i \tilde{s}(\zeta, \alpha, \varepsilon_{L_i}) + (1 - t_k)[c_s\tilde{s}(\zeta, \alpha, \varepsilon_s) + c_d\tilde{d}(\zeta, \alpha)]\Big\}\Bigg]\end{aligned}$$

En regroupant les termes et expressions similaires, l'algorithme final s'écrit plus simplement comme suit :

ÉchantillonnageGlobal2

u ~ $\mathcal{U}[0,1]$
SI u < c_{bruit} *ALORS*
    u = u / c_{bruit}
    (ζ, α) ← échantillonnage uniforme
SINON SI u < c_{sources} *ALORS*
    u = (u - c_{uniforme}) / (c_{sources} - c_{uniforme})
    (ζ, α) ← échantillonnage des sources
SINON SI u < $c_{\text{spéculaire}}$ *ALORS*
    u = (u - c_{sources}) / ($c_{\text{spéculaire}}$ - c_{sources})
    (ζ, α) ← échantillonnage spéculaire
SINON
    u = (u - $c_{\text{spéculaire}}$) / (1 - $c_{\text{spéculaire}}$)
    (ζ, α) ← échantillonnage diffus
SI Poids(ζ, α) * ancien poids < C_o *ALORS*
    SI u < proba d'arrêt *ALORS* tronquer le chemin actuel
    SINON nouveau poids = Poids(ζ, α) * ancien poids * (1 - proba d'arrêt)
SINON nouveau poids = Poids(ζ, α) * ancien poids

et ne nécessite l'utilisation que de deux nombres aléatoires. L'un pour u ~ $\mathcal{U}[0,1]$ d'où l'on déduit α :

$$\alpha \;=\; 2\pi \times \begin{cases} u / c_{\text{bruit}} & \text{si } u < c_{\text{bruit}} \\ (u - c_{\text{bruit}}) / (c_{\text{source}} - c_{\text{bruit}}) & \text{si } c_{\text{bruit}} \le u < c_{\text{source}} \\ (u - c_{\text{source}}) / (c_{\text{spéculaire}} - c_{\text{source}}) & \text{si } c_{\text{source}} \le u < c_{\text{spéculaire}} \\ (u - c_{\text{spéculaire}}) / (1 - c_{\text{spéculaire}}) & \text{si } c_{\text{spéculaire}} \le u \end{cases}$$

, l'autre pour ζ, calculé selon chaque méthode.
De même le calcul du poids s'écrit simplement :

$$\text{Poids}(\zeta, \alpha) \;=\; K\Big(c_{\text{bruit}}\tilde{m}(\zeta, \alpha) + q_{\text{source}} \max_i \tilde{s}(\zeta, \alpha, \varepsilon_{L_i}) + q_{\text{spéculaire}}\tilde{s}(\zeta, \alpha, \varepsilon_s) + q_{\text{diffus}}\tilde{d}(\zeta, \alpha)\Big)$$

III.8 De la nécessité de conserver l'énergie

La méthode que nous avons présentée peut être étendue à d'autres modèles comme nous l'avons déjà mentionné. Néanmoins, il est essentiel que ces modèles soient conservateurs d'énergie. Toute violation de cette règle de base entraînerait inévitablement la production d'une image presque complètement blanche. La stratégie de niveaux implique, en effet, la recherche du niveau le plus énergétique et conduirait à favoriser l'excès d'énergie à chaque fois qu'il se produit. L'effet cumulatif de la récursion provoquerait alors rapidement une saturation complète de la couleur finale.

Il est à noter que, outre l'aspect précédent, la conservation de l'énergie est également et avant tout, une loi fondamentale de la physique. Nombre de résultats en électromagnétisme n'ont put être établis que par l'application de cette propriété : les lois de la réflexion et de la réfraction en sont une des illustrations les plus frappantes.

De nouvelles conditions sont donc nécessaires sur notre modèle pour que la conservation soit garantie. Cette propriété s'exprime par le fait que l'énergie réfléchie est au plus égale à l'énergie incidente :

$$\int_{\Omega_r} L_r(P, \mathbf{R}) \cos\zeta d\omega \le \int_{\Omega_i} L_i(P, \mathbf{R}) \cos\zeta d\omega$$

L_r représente la luminance réfléchie et L_i l'incidente

Ce qui est exprimable uniquement en fonction de la luminance incidente :

$$\int_{\Omega_r} \cos\zeta_r d\omega_r \int_{\Omega_i} f_r(\mathbf{D}_r, P, \mathbf{R}_i) L_i(P, \mathbf{R}_i) \cos\zeta_i d\omega_i \le \int_{\Omega_i} L_i(P, \mathbf{R}) \cos\zeta d\omega$$

Expression que l'on peut écrire plus utilement sous la forme :

$$\int_{\Omega_i} \underbrace{\left(\int_{\Omega_r} r(\omega_r, \omega_i) \cos\zeta_r d\omega_r \right)}_{T(\omega_i)} l(\omega_i) d\omega_i \le \int_{\Omega_i} l(\omega) d\omega$$

$$\int_{\Omega} T(\omega) l(\omega) d\omega \le \int_{\Omega} l(\omega) d\omega \qquad (42)$$

Cette condition est en fait équivalente à :

$$T \le 1$$

En effet, l'équation (42) devant être vraie pour toute fonction l, elle l'est pour :

$$l(\zeta, \alpha) = \mathbf{1}_{[\zeta_0, \zeta_0 + z]}(\zeta) \mathbf{1}_{[\alpha_0, \alpha_0 + a]}(\alpha)$$

ce qui entraîne

$$\frac{1}{az} \int_{\alpha_0}^{\alpha_0 + a} \int_{\zeta_0}^{\zeta_0 + z} T(\zeta, \alpha) \sin\zeta d\alpha d\zeta \le \frac{1}{az} \int_{\alpha_0}^{\alpha_0 + a} \int_{\zeta_0}^{\zeta_0 + z} \sin\zeta d\alpha d\zeta$$

Lorsque a et z tendent vers zéro, la limite s'écrit :

$$\forall \zeta_0, \alpha_0 \qquad T(\zeta_0, \alpha_0)\sin\zeta_0 \le \sin\zeta_0$$

ce qui entraîne

$$T \le 1$$

Dans l'autre sens

$$T \le 1, l \ge 0 \Rightarrow T(\omega)l(\omega) \le l(\omega)$$

$$\Rightarrow \int_\Omega T(\omega)l(\omega)d\omega \le \int_\Omega l(\omega)d\omega$$

Pour peu que l soit positive ou nulle, ce qui est le cas avec la luminance, la conservation de l'énergie se traduit donc par :

$$\boxed{\int_\Omega T(\omega)l(\omega)d\omega \le \int_\Omega l(\omega)d\omega \Leftrightarrow T \le 1}$$

Avec notre fonction de transfert d'énergie, il faut donc :

$$\boxed{k_d R_o + k_s \underbrace{\int_\Omega \rho_s(\mathbf{D}_r, P, \mathbf{R}_i)\cos\zeta_i \sin\zeta_i d\zeta_i d\alpha_i}_{I} \le 1} \qquad (43)$$

Ainsi par exemple, dans le cas d'un matériau uniquement diffus, faut-il veiller à ce que k_d vérifie :

$$k_d \le \frac{1}{\underset{\lambda \in \vartheta}{\text{Max}}\, R_o}$$

Dans tous les autres cas, il est nécessaire d'évaluer numériquement I et d'en déduire les valeurs de k_d et k_s qui assurent la conservation. Ceci peut être tabulé, au moins pour les valeurs maximum de ρ_s, avec une indexation sur la valeur de m (Cf eq (31) §II.3).

Comme on le voit, la signification de k_d et k_s n'est pas aussi simple que dans les modèles locaux où il suffit en général que leur somme soit unitaire pour qu'il y ait conservation. Ici, si (43) était prise comme une égalité et non comme une inégalité, le matériau modélisé renverrait l'ensemble de l'énergie qu'il reçoit et se comporterait comme une espèce de corps «blanc» sans effet Joule. De la même façon, la valeur relative de ces deux paramètres n'indique guère le comportement du matériau modélisé. Le seul moyen de construire ce comportement consiste à considérer le cas d'une luminance incidente uniforme, ce qui permet de comparer la composante diffuse avec la spéculaire :

$$k_d R_o \quad \begin{matrix} < \\ ? \\ > \end{matrix} \quad k_s \int_\Omega \rho_s(\mathbf{D}_r, P, \mathbf{R}_i)\cos\zeta_i \sin\zeta_i d\zeta_i d\alpha_i$$

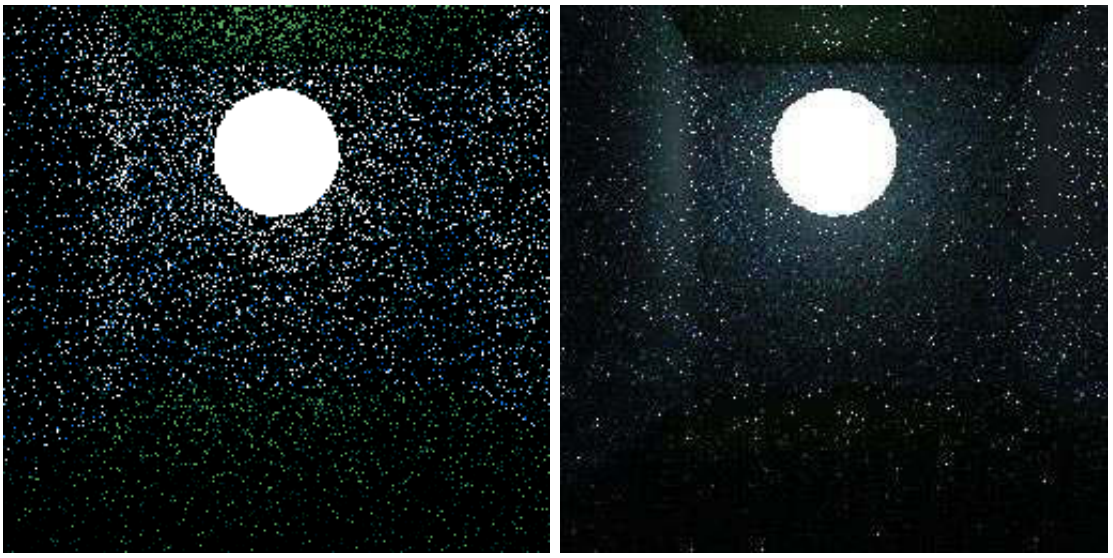
pour obtenir un comportement diffus lorsque $k_d \underset{\lambda}{\text{Min}}\, R_o > k_s \text{Max} I$ ou spéculaire dans le cas contraire. Au sujet de la conservation d'énergie pour des modèles d'éclairements plus simples, on pourra consulter Lewis [Lewi 93].

III.9 Résultats partiels

Un certains nombre de problèmes techniques liés à un changement de machines en fin de thèse nous empêchent de présenter une évaluation exhaustive de nos méthodes à l'heure de la conclusion de ce manuscrit. Bien que l'ensemble de ce que nous avons présenté dans ce chapitre ait été implémenté, les résultats que nous allons présenter sont, à notre grand regret, très partiels et ne nous permettent pas de faire le bilan des avantages et inconvénients de nos méthodes. Nous espérons que le lecteur nous excusera de ne pas lui fournir tous les éléments nécessaires à sa réflexion et saura néanmoins voir l'intérêt de nos travaux.

Les images qui suivent représentent une pièce faite de six murs totalement diffus éclairés par une source en son milieu. Afin de comparer nos résultats, ces images ont été calculées avec le schéma global que nous avons présenté (Cf §III.7 ÉchantillonnageGlobal2) ainsi qu'avec notre méthode minimale (Cf §III.2.2), c'est-à-dire en échantillonnant uniformément chaque hémisphère, augmentée d'une troncature de type roulette russe. Ces images ont été calculées avec 1, 10 et 100 rayons par pixel.

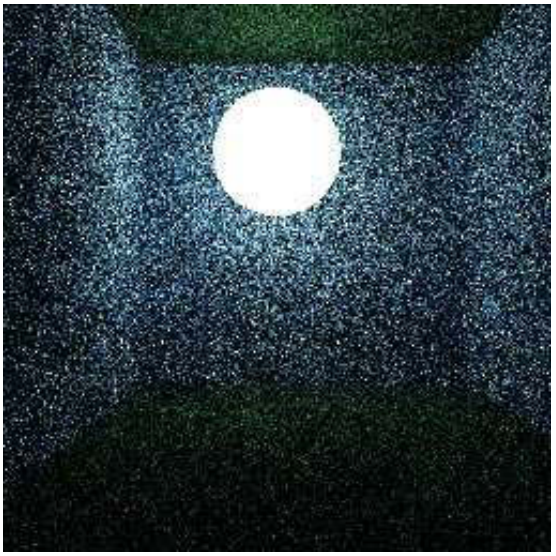
1 rayon par pixel



Méthode minimale
Temps de calcul : 9m18s56c
Rayons générés : 1 009 479

Schéma global
Temps de calcul : 1m22s96c
Rayons générés : 37 799

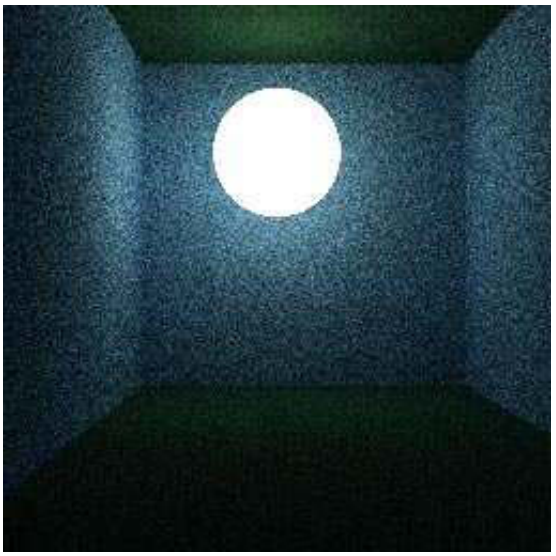
10 rayons par pixel



Méthode minimale
Temps de calcul : 1h30m55s19c
Rayons générés : 10 123 046

Schéma global
Temps de calcul : 14m00s35c
Rayons générés : 377 907

100 rayons par pixel



Méthode minimale
Temps de calcul : 14h32m05s69c
Rayons générés : 101 356 791

Schéma global
Temps de calcul : 2h23m34s93c
Rayons générés : 3 912 282

La différence de temps entre méthode minimale et schéma global peut paraître étonnante de prime abord. Elle résulte de l'échantillonnage uniforme qui génère des chemins extrèmement longs puisque la probabilité d'atteindre la source, et donc d'arrêter le chemin pour renvoyer une énergie, est très faible. Ceci entraîne la production d'un nombre très élevé de rayons et donc d'un temps de calcul en conséquence. Les points noirs de ces images sont dus à la troncature de la roulette russe sur ces chemins très longs. Cette troncature entraîne en effet l'arrêt du chemin et sa non-contribution dans la luminance moyenne du pixel. Alors que la troncature porte sur environ 49 % des

rayons de la méthode minimale (avec un nombre maximal de 26 rebonds), le schéma global ne tronque au plus que 2 rayons (le nombre de rebonds maximal variant de 12 à 16 selon le nombre de rayons par pixel).

Cette différence dans la troncature montre l'efficacité de notre méthode en ce qui concerne le choix entre l'échantillonnage des sources et celui de la FDRB. Les critères que nous avons retenus permettent en effet de choisir d'échantillonner les sources quand elles ont a priori une chance de contribuer fortement à l'énergie réfléchie.

L'un des caractères les plus importants dans toutes méthodes de Monte-Carlo est, bien sûr, la réduction du bruit dans l'image finale. Comme il est normal dans la méthode minimale, le bruit est très important et reste élevé même à 100 rayons par pixel. Le schéma global, quant à lui, fournit un résultat beaucoup plus stable dès le premier rayon lancé et continue à se stabiliser au fur et à mesure que la densité de rayons par pixel augmente. Cette caractéristique très importante fournit une confirmation des méthodes que nous avons introduites afin, précisément, de réduire la variance à chaque fois que nous en avions l'occasion.

III.10 Conclusions

Les méthodes de résolution que nous avons proposées dans ce chapitre partent toutes d'une formulation unifiée sous forme de chaînes de Markov homogènes. Si la décomposition explicite en somme infinie que nous avons utilisée n'est pas nouvelle (elle est aussi présente chez Shirley [Shir 90]), le traitement que nous en faisons sous forme d'espérance n'avait curieusement jamais été explicitée auparavant. De même, si, exceptionnellement, il est fait mention de chaînes de Markov ([Schl 94]), c'est sans explicitations ni conséquences particulières.

Notre approche de l'évaluation de la luminance reçue est, à notre connaissance, le premier schéma d'échantillonnage complet et adaptatif pour les modèles globaux. Complet en ce sens qu'il prend en compte toutes les variations de flux possibles, celles provenant des sources comme celles contenues dans la fonction de transfert, en considérant leur importance relative.

L'utilisation des informations a priori permet l'élaboration d'un échantillonnage de base ayant de bonnes caractéristiques que nous affinons au fur et à mesure que la luminance réelle est connue. Les informations a priori et cette adaptation progressive a posteriori laissent espéèrer une convergence plus rapide et un bruit plus faible que les méthodes actuellement utilisées. Faute d'une évaluation exhaustive de nos méthodes nous allons développer la comparaison avec les techniques existantes dans les paragraphes suivants.

La méthode minimale (§III.2.2) que nous avons adoptée a déjà été utilisée par de nombreux auteurs ([Lang 91], [Jens 95] par exemple) qui l'utilise curieusement pour l'échantillonnage de la partie diffuse. L'échantillonnage diffus que nous réalisons prend en plus en compte le cosinus en provenance de la fdrb ce que nous n'avons bizarrement rencontré nulle part, de même pour l'échantillonnage spéculaire qui exploite en général le modèle de Phong. L'échantillonnage de sources sphériques a été proposé par Shirley ([Shir et al 96]) dans un but de mise au point où il échantillonne l'intégralité de la sphère. Nous utilisons une forme modifiée de cet échantillonnage (§III.3.2) où nous n'échantillonnons que le cône sous-tendu par la source et non toute sa surface.

Les mélanges qui ont été proposés jusqu'à présent pour la FDRB ne tiennent pas compte du poids relatif des composantes diffuses et spéculaires et ne traitent

que de modèles d'éclairement très simplifiés. La stratégie de la poire (§III.3) permet, quant à elle, de traiter une vaste gamme de modèles d'éclairements par un traitement numérique très général moyennant quelques caractéristiques peu contraignantes sur ces modèles. La stratégie du tournesol que nous avons présentée (§III.5.1), ressemble à la méthode linéaire de Shirley([Shir et al 96]). Cette dernière choisie les sources à échantillonner à partir de directions particulières pour lesquelles la luminance est calculée en supposant la source visible. Notre méthode utilise la position et la puissance des sources mais également leur étendue, ce dont ne tiennent pas compte les critères de Shirley. De plus amples comparaisons seraient nécessaires pour évaluer les avantages de chaque approche. Par contre, dans une scène où de nombreuses sources sont présentes, la technique de Shirley mettant en oeuvre des octrees sera sans doute plus rapide que la nôtre.

Les stratégies adaptatives qui existent actuellement sont en majorité basées sur des notions de *cartes de photons* (photon map voir par exemple [Jens 95]) et sont utilisées dans des méthodes partant des sources. Une carte de photons représente en deux dimensions l'hémisphère paramétré selon les angles azimutal et zenithal et permet de stocker l'énergie incidente ou réfléchie selon diverses directions. Cette approche est sans doute envisageable avec des méthodes partant de l'oeil mais implique une mise en oeuvre très lourde en général et plus encore dans un contexte CSG. Notre estimation de niveau est beaucoup plus simple à mettre en place et énormément plus économe en mémoire. Les cartes de photon ont néanmoins une plus grande précision que nos estimateurs. Il serait d'ailleurs envisageable de marier les deux techniques : nos estimateurs de niveau peuvent être utilisés, comme nous l'avons fait, pour déterminer le niveau de contribution le plus important mais au lieu d'échantillonner les sources si le niveau est fort, on choisirait d'échantillonner selon la carte de photons.

D'autres extensions à nos méthodes sont déjà envisageables. Nous en voyons au moins deux facilement implémentables. La première consiste à utiliser la combinaison d'échantillonnages préconisée par Veach et Guibas [Veac et al 95]. Cette combinaison peut permettre quelques gains au niveau de la vitesse de convergence de nos méthodes.

La deuxième extension, qui est d'ailleurs applicable à toute méthode de Monte-Carlo adaptative partant de l'oeil, consiste à profiter des informations accumulées sur un pixel pour les utiliser sur un pixel adjacent. Cette application du paradigme progressif que nous allons présenter dans le chapitre suivant permettrait un nouveau mode d'échantillonnage a priori qui serait combiné à ceux que nous avons présenté. Cette combinaison pourrait être fixe pour l'ensemble de l'image. Le plus simple étant d'échantillonner pour moitié avec l'échantillonnage a priori (poire / tournesol) et pour moitié avec les informations du pixel adjacent, l'échantillonnage a posteriori étant toujours utilisé quand la précision est devenue suffisante. Une autre possibilité pourrait être basée sur la classification intuitive de l'image par l'utilisateur, selon l'une des trois distributions types que nous avons mis à jour (Cf §IV.4.3), ce qui fournirait trois mélanges possibles : priorité au schéma a priori, priorité aux informations du pixel adjacent ou équirépartition entre les deux. La combinaison pourrait également être variable d'un pixel à l'autre et fonction d'heuristiques pouvant «prédire» la ressemblance entre pixels. Cette dernière solution étant sans doute préférable mais beaucoup plus difficile. Cette approche progressive dans les méthodes de Monte-Carlo n'a pas encore été explorée mais promet des résultats intéressants.

Tracé de rayons progressif

Les millions d'opérations impliquées dans un tracé de rayons sont très fortement liées entre elles. La notion de cohérence spatiale et colorimétrique sous-tend, en effet, l'ensemble du procédé. Que ce soit l'unité géométrique des objets, le comportement réflexif semblable de points proches, le lien «topologique» entre objets regroupant des rayons par famille ou encore la cohésion colorimétrique d'une image, la cohérence est omniprésente. Il n'y a qu'un pas à franchir pour tirer partie de cette caractéristique et utiliser ses propriétés pour accélérer les calculs. Nous nous proposons de franchir la frontière pour quelques-uns de ces pas qui nous mèneront vers un tracé de rayons progressif.

IV.1 Concepts de base

L'idée de base à l'origine de ces travaux est le *paradigme progressif*. Il consiste à s'approcher de la solution d'un problème par étapes successives, chacune d'entre elles réutilisant au mieux les résultats des étapes antérieures pour éviter les efforts déjà consentis et progresser vers la solution en acquérant le plus de connaissances nouvelles possibles dans la limite des moyens impartis. En d'autres termes, utiliser la mémoire et l'astuce plutôt que la force brute pour concentrer les efforts sur l'acquisition d'éléments indispensables et encore inconnus. Une synthèse entre esprit économe et paresse efficace.

Le tracé de rayons, même avec des modèles de transferts d'énergie semi-globaux, est encore une technique très coûteuse en temps machine : plusieurs jours de calculs pour des scènes complexes est monnaie courante. Les outils de modélisation géométrique et la complexité croissante des modèles d'éclairements ne permettent guère d'avoir une idée claire et précise de l'image finale qui sera générée. Il n'est pas rare de devoir faire plusieurs allers-retours entre modélisation et visualisation avant d'être satisfait du résultat. D'où l'idée de départ de notre approche inspirée par la progressivité : il est inutile de faire de longs calculs pour avoir une excellente image des erreurs de modélisation, de défauts de modèle d'éclairement, etc. On peut en effet essayer d'obtenir rapidement une image grossière que l'on affinera petit à petit par étapes successives.

L'objectif étant que l'utilisateur s'aperçoive rapidement des erreurs commises : des objets de la scène sont mal positionnés, des sources mal orientées, des effets de matière farfelus,… Une image de plus en plus complète et précise est donc générée à chaque étape. L'utilisateur pouvant, soit poursuivre le processus, s'il est satisfait du résultat partiel, soit l'interrompre s'il décèle des anomalies. Dans le cas où toutes les étapes lui conviennent, le résultat final doit être autant que possible à la hauteur de ses attentes.

La deuxième idée forte découlant de l'approche progressive provient du constat que toute image possède des zones homogènes et inhomogènes. Les zones homogènes correspondent à des régions de faible contraste, des dégradés de couleurs doux que l'on trouve par exemple sur des portions d'objets diffus dont la normale varie peu. Les zones inhomogènes apparaîssent dans des zones à forts contrastes lors de changement brutaux de couleur, par exemple à la limite de passage entre ombre et lumière, sur un bord d'objet, sur un reflet ou encore sur un objet texturé. Il est clair que peu de points sont nécessaires à la description des zones homogènes alors que les inhomogènes en requièrent bien davantage. Notre but est donc de trouver une méthode qui sache déterminer d'«elle-même» l'homogénéité d'une région de l'image et le nombre de points le plus faible possible nécessaire à la représentation d'une région. Cette méthode est hautement désirable car toute économie réalisée sur la génération d'un point se répercutera de facto sur l'ensemble du processus. Un seul rayon primaire non généré et c'est un grand nombre de rayons secondaires, de calculs de rendu et d'intersections supprimés…pour peu que cette non-génération ne coûte pas plus cher que sa génération, évidemment. C'est donc comme une première étape que ce travail à porté sur la progressivité de cette opération.

Déterminer le type d'homogénéité qui nous est nécessaire ainsi que la détection de cette homogénéité sont les deux problèmes que nous avons à résoudre. Les méthodes d'anti-crénelage sont proches de nos préoccupations aussi allons-nous nous y intéresser.

IV.2 Recherches apparentées

De nombreux auteurs se sont penchés sur la question de la génération des points de l'écran, principalement pour résoudre le problème du crénelage (aliasing) engendré par le tracé de rayons. Ce phénomène est bien connu en image de synthèse et se manifeste conformément au théorème de Shannon [Shan 49] à chaque fois que la fréquence de variation d'une fonction est au moins deux fois plus rapide que la fréquence d'échantillonnage de cette fonction. Cette situation se produit dans trois grandes situations :

- lors d'un changement brutal de couleur provoqué soit par la modélisation géométrique (bords d'objets essentiellement) soit par l'éclairement (limite d'ombre).
- lorsque la projection d'un objet sur l'espace image a une taille inférieure au pixel.
- lorsque la couleur d'un objet varie plus d'une fois par pixel (avec des textures par exemple).

Toutes les méthodes proposées pour résoudre le crénelage se décomposent en deux aspects, parfois trois. Le premier est une méthode d'échantillonnage, le second est un test sur les échantillons pour déterminer s'il y a ou non crénelage et le troisième est un filtrage.

IV.2.1 Échantillonnage de l'espace image

L'un des premiers a avoir abordé le problème est sans doute Whitted [Dipp et al 85] lorsqu'il proposa la méthode du tracé de rayons. L'échantillonnage qu'il proposait consiste à lancer un rayon primaire de chaque coin d'un macro-pixel de taille prédéfinie. C'est également l'échantillonnage retenu par Jansen [Jans 83] et plus récemment encore par Akimoto [Akim et al 91].

L'approche de Dippé [Dipp et al 85] est plus générale et considère diverses formes d'échantillonnages aléatoires permettant d'approximer de façon discrète et presque sans crénelage une fonction continue. Il étudie notamment l'échantillonnage de Poisson qui consiste à échantillonner les points de telle sorte que lorsque l'espace échantillonné est partitionné en régions disjointes, le nombre de points de chaque région suive une loi de Poisson d'espérance proportionnelle à la population totale de cette région. Une méthode dérivée est l'échantillonnage de Poisson selon un disque (Poisson disk sampling) ou encore selon une distance minimale qui impose, en plus de la définition précédente, que deux points quelconques ne soient pas plus proches qu'une certaine distance. Ce dernier procédé est inspiré d'une étude du biologiste Yellott [Yell 83] portant sur la répartition des cellules en cônes dans l'oeil de singes rhésus, répartition qui suit une loi de Poisson à distance minimale. Selon Yellott, cette distribution minimise, voire supprime totalement le crénelage en le remplaçant par un bruit contenant plus de hautes fréquences que de basses/moyennes fréquences auxquelles l'oeil est plus sensible. Mitchell [Mitc 87] s'est également intéressé à cet échantillonnage avec l'algorithme dit du *lancé de flèchettes* (dart-throwing algorithm) approchant ce type de distribution. Devant le coût de la méthode, il propose un deuxième algorithme plus abordable mais aussi plus éloigné de la répartion de Poisson selon un disque appelé *diffusion de points* (point diffusion algorithm). Dippé analyse également l'échantillonnage généré par la perturbation aléatoire d'une grille régulière (jittered sampling) qui ressemble à une distribution de Poisson selon un disque mais se comporte moins bien en ce qui concerne l'anti-crénelage.

L'approche de Lee [Wald 48] utilise un échantillonnage stratifié classique et uniquement basé sur l'effectif de chaque strate.

Painter [Morl et al 75] utilise un échantillonnage uniforme construit sur l'échantillonnage séquentiel uniforme de Kajiya (Cf §II.2.1) représenté sous forme de k-D tree.

IV.2.2 Critères et détection du crénelage

Chaque auteur s'intéressant à l'anti-crénelage cherche un critère et une méthode qui en assure la détection. Les premiers critères retenus étaient assez frustes. Celui de Whitted consiste à comparer les intensités des quatre coins du macro-pixel. Si elles diffèrent peu et qu'aucun petit objet n'est présent dans la région qu'ils définissent, cette dernière est supposée anti-crénelée. Semblablement, Jansen teste la différence d'intensité entre un pixel et ses trois voisins. Quant à Akimoto, il définit plusieurs niveaux de similarité entre les quatre coins d'un macro-pixel. A partir de l'arbre des rayons stocké en chaque point, quatre échelons de complexité sont définis :

① Des objets différents apparaissent dans les rayons transmis
② Les rayons transmis sont identiques mais les rayons d'ombres diffèrent
③ Ombres et transmission sont identiques mais les arbres de textures sont différents ou l'écart maximum d'une composante avec l'intensité moyenne

dépasse un seuil S donné :

$$\underset{i \in [1, 4]}{\text{Max}} (|\bar{R} - R_i|, |\bar{V} - V_i|, |\bar{B} - B_i|) > S$$

④ Aucune des trois conditions précédentes n'est vérifiée

Selon le niveau de similarité détecté différentes méthodes d'interpolations ou de calcul des pixels manquants sont proposées.

La notion de contraste est un autre critère également utilisé, notamment par Mitchell [Mitc 87] sous la forme suivante pour chaque composante RVB :

$$C = \frac{I_{max} - I_{min}}{I_{max} + I_{min}}$$

I_{min} et I_{max} représentent respectivement la valeur minimum et maximum d'une composante sur une région donnée

La majeure partie des critères qui suivirent sont statistiques. Lee utilise la variance comme mesure pertinente du crénelage. Painter, quant à lui, utilise l'intensité moyenne et Brown, une variance pondérée par la moyenne.

La détection du critère se fait selon deux méthodes, l'une régulière, l'autre aléatoire. La régulière est simplement liée à l'échantillonnage régulier de l'écran ([Dipp et al 85], [Jans 83], [Akim et al 91]). Lorsque le critère de crénelage est vérifié, la région de l'image testée est subdivisée et un nouveau test a lieu. Dans le cas contraire, différentes stratégies de reconstruction sont mises en oeuvre par chaque auteur. Les méthodes aléatoires découlent bien sûr de l'échantillonnage du même nom, où le critère de crénelage est détecté par des tests statistiques ou des tests à seuil simple. Les détections à seuil simple demandent en général un taux d'échantillonnage minimum de un rayon par pixel. Si le crénelage est détecté, un deuxième taux fixe est utilisé pour suréchantillonner le pixel ([Mitc 87]). Les tests statistiques sont peu nombreux : Lee utilise un test du χ^2 sur la variance et Painter un intervalle de confiance sur la moyenne avec un test de Student pour chaque composante RVB. Painter met en oeuvre deux stratégies. La première porte sur les noeuds situés au-dessus du niveau du pixel et affine en priorité les bords d'objets et les noeuds ayant le produit aire × variance le plus fort. Cette première étape conduit à une espèce d'ordonnancement des noeuds pour affiner les plus importants en priorité. Le découpage et l'échantillonnage s'arrêtent lorsque la moyenne estimée se situe à $1/255^{ème}$ de sa vraie valeur selon le test statistique.

IV.2.3 Reconstruction

Lorsque l'échantillonnage est stoppé, soit parce qu'un taux maximum a été atteint, soit parce que le test l'a décidé, la région doit être reconstruite à partir de ses échantillons. Cette reconstruction est réalisée par filtrage si la densité de points est suffisante, c'est-à-dire si l'image est connue sur l'ensemble du support du filtre. Lorsque ce n'est pas le cas, le filtrage doit être précédé d'une interpolation pour donner une couleur aux points non-échantillonnés. L'interpolation n'a pas été beaucoup étudiée; dans tous les articles déjà cités, elle consiste simplement en une interpolation constante (Painter par exemple remplace les cellules terminales de l'arbre par la couleur du point qu'elles contiennent). Le filtrage a fait l'objet de plus d'études et l'ensemble de ceux utilisés par ces auteurs sont décrits par Brown. Ce dernier préconise l'utilisation du filtre de Mitchell/Netravali [Mitc et al 88] ayant la forme suivante :

$$f(u) = \begin{cases} 7|u|^3 - 12|u|^2 + 16/3 & |u| < 1 \\ \frac{7}{3}|u|^3 + 12|u|^2 - 20|u| + 14 & 1 \leq |u| < 2 \\ 0 & \text{sinon} \end{cases}$$

$$F(x, y) = f(x)f(y)$$

IV.2.4 Résultats

Les résultats obtenus par les auteurs de ces méthodes sont difficilement comparables. Les scènes de tests sont différentes, les modèles d'éclairement jamais explicités et le type de tracé de rayons utilisé pas toujours clairement affiché (un tracé de rayons distribués aura nécessairement besoin de beaucoup plus de rayons qu'un tracé classique). Néanmoins voici quelques-uns de ces résultats. La méthode de Dippé permet de ne calculer qu'un pixel sur deux sur l'image d'un échiquier (les exemples et les chiffres manquent pour en dire beaucoup plus), soit un gain approximatif de 50 % sur le nombre de rayons primaires générés par la méthode classique. Painter obtient une majorité de pixels nécessitant un suréchantillonnage inférieur à 10 rayons (en moyenne 2 par pixel), une portion non négligeable des pixels ne nécessitent aucun calcul (environ 20 %). Malheureusement, sur les deux images présentées, la projection de la scène sur l'écran représente 50 % de sa surface dans un cas, 25 % environ dans l'autre. Les 20 % de pixels non échantillonnés ne sont donc pas représentatifs même s'ils semblent indiquer que la méthode peut détecter ce genre de zone homogène. Mitchell ne fournit guère de résultats chiffrés, il indique simplement qu'il y a au moins un rayon par pixel voire beaucoup plus. Brown indique que le nombre d'échantillons peut être réduit d'un facteur 8 ou 9 par rapport à un suréchantillonnage systématique de 16 rayons par pixel. Les résultats d'Akimoto sont de natures très différentes. On peut simplement dire que 9 à 13 % des points sont calculés par tracé de rayons classique, 3 à 41 % par tracé de rayon «allégé» et enfin 45 à 88 % par interpolation, l'accélération du temps de calcul variant de 2 à 9. Compte tenu des remarques faites un peu plus haut sur la comparaison des résultats, il semblerait, malgré tout, qu'Akimoto et Brown obtiennent les meilleurs résultats.

IV.2.5 Bilan

Une méthode progressive qui utiliserait un échantillonnage régulier, fut-il adaptatif, devrait gérer d'importants problèmes de crénelage. Problèmes guère solubles autrement que par un suréchantillonnage massif. D'autre part, la régularité de ces échantillonnages est un gros handicap, s'ils peuvent avoir un bon comportement pour certaines images, il est certain que leurs performances se dégraderont pour d'autres. L'échantillonage aléatoire élimine ces défauts par une adaptabilité beaucoup plus grande et un comportement moyen bien meilleur. Une qualité qui minimise son inconvénient majeur qui est bien sûr, son coût plus élevé.

Utiliser des tests statistiques fournit un cadre beaucoup plus rigoureux que les tests à seuils simples, pour peu que la démarche soit suivie jusqu'au bout. On évitera par exemple de tester la variance par un test du χ^2 qui suppose que la répartition locale des valeurs de la fonction suit, au moins asymptotiquement, une loi normale (ou au moins que l'effectif sur lequel porte le test est assez grand), encore plus pour un test de

Student qui ne peut porter que sur une répartition strictement normale. Comme nous l'avons dit, la variance a des comportements très différents selon les zones de l'image où elle est mesurée, dans ces conditions on ne peut dire que la couleur suive une loi normale sur des bords d'objets aussi bien que sur des objets texturés ou des zones de contrastes faibles.

Ceci étant, lorsque la méthodologie est respectée, les tests de variance fournissent des résultats acceptables, en ce sens qu'ils détectent bien les zones de crenelage. Leur défaut principal réside dans leur vitesse de convergence lente, une propriété bien connue des statisticiens.

Parmi les échantillonnages aléatoires que nous avons présentés on notera particulièrement celui de Poisson selon un disque. Il posséde en effet de très bonnes propriétés, malheureusement, outre son coût élevé, il semble très difficile à implémenter sous forme adaptative. Une propriété intéressante du procédé reste néanmoins de garantir une bonne répartition des points dans l'espace échantillonné : les points couvrent l'espace sans former d'amas ni laisser de grands vides. Cette propriété est très intéressante car, quelle que soit la technique, la variation de couleur peut alors être mesurée avec un minimum de biais. L'échantillonnage selon une grille perturbée possède également, mais de façon moins forte, cette propriété. Il est peu coûteux et peut être mis en oeuvre sans trop de problèmes. Mais, lui aussi, se prête difficilement à un usage adaptatif : comment tenir compte des points déjà échantillonnés pour les réaligner sur une grille d'échantillonnage plus fin ? La meilleur méthode semble donc être celle de Kajiya utilisée par Painter avec l'échantillonnage séquentiel uniforme : fiable, assez rapide, assurant une bonne couverture de l'espace, adaptatif par nature et assez simple à mettre en oeuvre. Son seul défaut est l'utilisation quelque peu dispendieuse de la mémoire, défaut dont on s'accomode en regard de ses qualités.

En ce qui concerne les tests, outre leur vitesse de convergence assez lente, un autre facteur important à considérer dans le cadre progressif est la facilité de paramétrage et d'intervention dans la méthode. De ce point de vue, on peut dire que les tests de variance, même si leur contrôle peut être assez rigoureux n'en restent pas moins liés à l'expérience, pour ne pas dire l'intuition de l'utilisateur. Dippé, par exemple, choisit, entre autres, deux taux d'échantillonnage, un taux initial moyen et un taux maximum ainsi que l'erreur tolérable par l'utilisateur, critères que l'utilisateur doit adapter à chaque image. Autrement dit, il est relativement difficile de déterminer au départ la qualité de l'image que l'on souhaite obtenir. Brown, dont la technique reprend les travaux de Dippé pour une part et ceux de Painter d'autre part, essaye de déterminer l'influence des seuils de tolérance contrôlant son test de variance. Les seuils qu'il cite varient de 0,001 à 0,05 pour la variance externe (grosso modo la variance d'un contraste) et de 0 à 10^{-100} pour la variance interne (estimation de la variance de la moyenne), le suréchantillonnage étant enclenché lorsque la variance dépasse ces seuils. La plage recommandée pour le seuil de la variance interne est très restreinte sachant qu'une variance est toujours positive et on peut se demander si ce seuil a vraiment une signification réelle…

Un autre défaut des tests utilisés jusqu'à maintenant est le peu d'attention qu'ils attachent en général à l'aspect psycho-visuel des critères qu'ils manipulent. Les auteurs considèrent en effet davantage le problème sous sa forme mathématique, à savoir l'approche fournie par la théorie du signal (échantillonnage/reconstruction), que son interaction avec le sens visuel humain. C'est d'ailleurs sans doute une des raisons pour lesquelles la plupart des techniques proposées sont si gourmandes. Mitchell et Dippé s'intéressent à cet aspect du point de vue de l'échantillonnage (Poisson à disque) ainsi que de celui du critère (contraste) mais malheureusement sans études psycho-physiologiques à leur disposition. Dans cet optique on peut s'étonner que seul le système RVB soit

utilisé par tous ces auteurs alors que la notion de distance est au coeur de tous leurs critères et qu'elle n'a qu'une signification mathématique dans ce système.

En conclusion, on peut remarquer que les tests utilisés jusqu'à présent conduisent souvent à surévaluer plus ou moins fortement la complexité de l'image (à la nuance près que les images finales sont presque complètement anti-crenelées, ceci expliquant peut-être cela) et qu'ils ne donnent pas vraiment de contrôle sur la qualité finale.

IV.3 Nouvelle approche

IV.3.1 Progressivité

L'idée de progressivité exprimée en début de chapitre ne correspond pas vraiment à celle de l'anti-crénelage puisque l'on souhaite ne calculer progressivement que les points indispensables à la définition d'une image de qualité donnée. Cette qualité pouvant être inférieure à l'anti-crénelage si tel est le bon vouloir de l'utilisateur. Pour réaliser cette idée nous avons besoin de trouver un critère d'homogénéité pertinent pour l'oeil, ainsi qu'un test et une méthode d'échantillonnage qui lui soient adaptés. Sachant qualifier une région quelconque d'une image d'homogène ou d'inhomogène, l'image va être affinée petit à petit par étape successive chacune formant une *passe progressive*. L'homogénéité ne peut être qu'une notion locale, fonction de la distance d'observation, de la taille du détail observé, de la définition disponible ainsi que de la sensibilité de l'oeil à ces paramètres, en un mot de l'échelle. Sans prétendre concilier l'ensemble de ces critères, une étape de notre affinage progressif consistera à balayer toutes les régions inhomogènes pour les subdiviser compte tenu de la localité. Chacune des sous-régions ainsi créées sera testée et qualifiée à son tour. L'état homogène ou inhomogène de chaque sous-région sera ensuite répercutée sur la région mère, déclarée homogène si toutes ses filles le sont, inhomogène dans le cas contraire. Le processus se déroulant jusqu'à ce que toutes les régions soient homogènes, soit par terminaison homogène du test, soit par atteinte d'une taille de région minimale fixée par l'utilisateur.

Pour réaliser ce but, l'approche statistique, comme nous l'avons dit, est la plus souple et la plus susceptible d'un bon comportement moyen. Deux caractéristiques essentielles en regard de la diversité des situations et des complexités très variables tant d'une image à l'autre qu'à l'intérieur-même d'une image. Les techniques déjà utilisées semblent insuffisantes de par leur lenteur de convergence qui dépend en fin de compte du théorème de la limite centrale. Nous avons donc besoin d'un test statistique plus puissant qui sache contrôler de «lui-même» le nombre d'individus nécessaires à une décision avec une précision donnée. Ce contrôle du nombre d'échantillons est précisément au coeur de l'analyse séquentielle.

IV.3.2 Analyse séquentielle

L'*analyse séquentielle* est un ensemble de méthodes statistiques basées sur des procédures de test dont le nombre d'observations n'est pas fixé à l'avance contrairement aux tests statistiques classiques mais varie en fonction des résultats de l'expérience. Elle permet à chaque nouvelle observation de déterminer si le test doit être poursuivi avec un nouvel échantillon ou si, au contraire, on peut décider de son arrêt compte tenu des observations déjà faites.

La théorie de l'analyse séquentielle a principalement été développée par Wald pour aboutir au test séquentiel du rapport de vraisemblance (sequential probability ratio test) dont il est l'auteur [Wald 48]. Ce test (TSRV dans la suite) produit fréquemment un gain de 50 % sur le nombre d'observations réalisées par la meilleure procédure de test à nombre fixe d'observations et est toujours plus rapide que n'importe quel autre test séquentiel.

Nous aborderons le TSRV dans une présentation simplifiée. Pour de plus amples détails, tant pratiques que théoriques, se référer au très bon ouvrage de Wald déjà cité ainsi qu'à celui de Siegmund pour les développements théoriques plus récents [Sieg 85].

Soit une variable aléatoire x dont on veut classer chaque observation indépendante X_i en deux catégories disjointes c_0 et c_1 selon un critère donné. La proportion p d'observations de type c_1 est inconnue. On souhaite déterminer le type de la population et on dira qu'il est plutôt c_1 si $p = p_1$, une hypothèse que l'on nommera H_1, ou plutôt c_0 si $p = p_0$ (avec $p_0 < p_1$), ce sera l'hypothèse H_0. Dans tous les autres cas, aucune décision ne pourra être prise. À chacune de ces hypothèses est associé un risque : α pour H_0, β pour H_1. Le risque α représente la probabilité que l'on rejette l'hypothèse H_0 alors qu'elle est vraie, c'est-à-dire que l'on accepte H_1, tandis que la probabilité pour qu'on accepte H_0 alors que H_1 est vraie est de β ($\alpha+\beta<1$). Autrement dit la probabilité de déclarer que la population est de type c_1 alors qu'elle est de type c_0 est de α, celle de déclarer l'inverse est de β. La fonction de distribution de x est notée $f_p(x)$ et vaut $f_{p_0}(x)$ si H_0 est vraie, $f_{p_1}(x)$ lorsque H_1 est vraie. Puisque les observations sont indépendantes (voir l'équation (34) §III.2.1), la probabilité d'obtention de l'échantillon $(X_1,\ldots,X_k)$, également appelée *vraisemblance*, est donnée par :

$$p_1^{(k)} = f_{p_1}(x_1)\ldots f_{p_1}(x_k) \qquad \text{si } H_1 \text{ est vraie}$$

et par

$$p_0^{(k)} = f_{p_0}(x_1)\ldots f_{p_0}(x_k) \qquad \text{si } H_0 \text{ est vraie}$$

Le TSRV considère le ratio de ces deux probabilités, le *rapport de vraisemblance*, à chaque nouvel échantillon et se définit comme suit à l'étape k :

si $\dfrac{p_1^{(k)}}{p_0^{(k)}} \geq \dfrac{1-\beta}{\alpha}$ le test s'arrête par acceptation de l'hypothèse H_1

si $\dfrac{p_1^{(k)}}{p_0^{(k)}} \leq \dfrac{\beta}{1-\alpha}$ le test s'arrête par acceptation de l'hypothèse H_0

sinon aucune décision n'est prise et le test se poursuit par le tirage d'un nouvel individu

Wald prouve, et c'est une propriété essentielle de la méthode, que la probabilité d'arrêt du test est de 1, i.e il est certain qu'une décision pourra être prise. L'autre propriété indispensable réside dans la garantie des risques qui sont des majorations de la proportion de mauvaises décisions en toutes situations. Le TSRV est, de plus, valide, que les observations soient ou non indépendantes (la définition de $p_0^{(k)}$ et de $p_1^{(k)}$ est alors à reconsidérer) et peut utiliser un nombre quelconque d'hypothèses.

La grande flexibilité du test alliée à sa vitesse de terminaison rapide permettent d'espérer de meilleurs résultats que ceux fournis par les tests couramment utilisés en synthèse d'image. Pour pouvoir utiliser ces résultats, il nous faut maintenant établir un critère d'homogénéité (p_0, p_1, H_0, H_1) sur une image ainsi qu'une méthode d'échantillonnage nous permettant le calcul d'une probabilité.

IV.3.3 Homogénéité

Les critères retenus jusqu'ici pour l'anti-crénelage reposent sur la variance ou sur une notion de contraste et cherchent à quantifier la variation de couleur. Cette variation, telle que la variance ou le contraste en rendent compte, est très proche d'une notion d'étendue de valeurs. C'est particulièrement clair pour la variance qui mesure l'écart moyen entre l'ensemble des valeurs et leur moyenne. La notion d'étendue nous semble donc pertinente pour qualifier l'homogénéité d'une image, nous choisissons donc de la mesurer au travers de l'étendue des valeurs autour de leur moyenne. L'étendue renvoie immédiatement à la définition d'une distance, il est donc nécessaire d'en choisir une qui soit pertinente pour l'oeil. Dans ce contexte, la distance euclidienne sur l'espace L*u*v* (voir l'équation (24) §I.6) nous semble l'une des plus appropriées, en attendant d'en trouver une meilleure, étant donné son bon comportement psycho-visuel. Si l'on note C_y le triplet des coordonnées L*u*v* du point y de l'image, $\overline{C}_R$ le triplet des coordonnées moyennes d'une région R de cette image ($y \in R$) et si γ est une variable aléatoire uniforme sur R, alors connaître l'étendue de C autour de la moyenne revient à se donner l'*étendue* $\varepsilon > 0$ et à déterminer la probabilité suivante :

$$\boxed{p = P(\Delta_{uv}(\overline{C}_R, C_\gamma) \leq \varepsilon)} \qquad (44)$$

c'est-à-dire la probabilité que la distance colorimètrique entre un point quelconque de R et sa couleur moyenne se situe dans un intervalle centré sur cette moyenne, intervalle que nous appellerons désormais *zone de tolérance*. En première approche, une valeur pertinente de 2ε se situe aux alentours de 1 puisque c'est la distance juste discernable (just noticeable difference, JND) dans cet espace pour l'oeil. Dans ces conditions, si la majorité des points est située dans la zone de tolérance, la région sera jugée homogène et, à l'inverse, plus la probabilité d'appartenance est faible, plus la région est inhomogène. Nous définissons donc une probabilité p_h (pour *probabilité homogène*) au-delà de laquelle une région sera déclarée *homogène*, c'est-à-dire lorsque $p \geq p_h$. L'*homogénéité* ainsi définie, nous définissons l'*inhomogénéité* a contrario comme son absence. Ce dernier terme est retenu à dessein plutôt que celui d'hétérogénéité. L'homogénéité nous semble en effet une notion visuellement forte et précise. Il est assez facile de qualifier d'homogènes deux objets différents (par exemple une porte et un plafond de couleurs différentes mais unies) alors que l'inhomogénéité est un concept plus vague qui recouvre des situations fort différentes : par exemple si une table en bois et un sol carrelé en terre cuite peuvent être tous deux qualifiés d'inhomogènes, il est plus difficile de préciser s'ils le sont pareillement ou de «combien» ils diffèrent.

Cette notion d'homogénéité prend tout son sens sous l'aspect géométrique. Chaque point de R a une couleur représentable dans un espace a trois dimensions. Dans cet espace, L*u*v* pour nous, l'ensemble des couleurs de R forme un amas plus ou moins compact selon le degré d'homogénéité de la région. C'est cette compacité que nous mesurons à l'aide de notre critère puisque nous cherchons à quantifier la proportion de couleurs contenues dans une boule de rayon ε centrée sur la moyenne : la *sphère d'homogénéité*. Par rapidité de langage nous parlerons des points «dans la moyenne» pour signifier les points dont la couleur est contenue dans cette sphère.

IV.3.4 Échantillonnage et reconstruction

L'échantillonnage aléatoire du plan que l'on recherche doit être aussi homogène que possible au sens de sa répartition bidimensionnelle. Il ne doit pas laisser de grandes portions d'espaces sans point, ni créer d'amas, ni générer de motifs répétitifs. Il doit également permettre de prendre un nombre de points quelconque en une ou plusieurs fois, sans qu'un point puisse être échantillonné plusieurs fois, ceci pour éviter au maximum de refaire des calculs d'intersection et de rendu coûteux. Il est également souhaitable que le tirage d'un point soit aussi rapide que possible compte tenu de la fréquence d'utilisation de la procédure. Malheureusement, l'ensemble de ces contraintes idéales, comme souvent, est incompatible avec la nécessité d'être capable de calculer la probabilité d'apparition du caractère d'un point.

Le schéma que nous avons retenu consiste donc en un échantillonnage uniforme hypergéométrique simple. L'image est partitionnée selon un arbre quaternaire (quadtree) définissant les régions de l'image que l'on cherche à qualifier. À chaque noeud terminal de ce premier arbre est attaché un arbre d'échantillonnage également quaternaire dans lequel chaque feuille contient au plus un point. L'avantage de ces structures est l'accès rapide à une feuille. L'inconvénient majeur est la quantité de mémoire importante nécessaire à son stockage. Cherchant plus la rapidité que l'économie mémoire, c'est un inconvénient dont on s'accomode. À ces deux structures nous en avons ajouté une troisième pour accélérer certaines opérations de reconstruction, il s'agit d'une liste non triée constituée des points échantillonnés et accessible depuis l'arbre des régions. L'ensemble se présente sous la forme suivante :

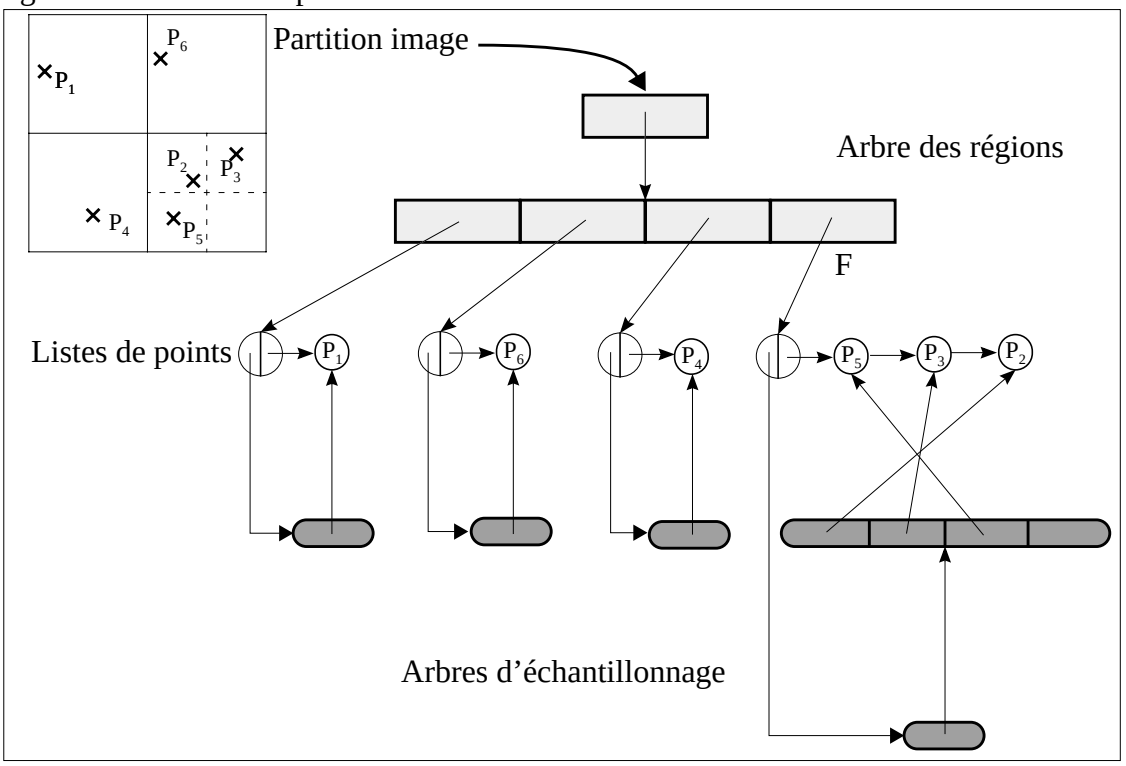


Figure 46: Structures de données pour test et échantillonnage

Dans cet exemple, l'arbre des régions indique que l'image a été partagée en quatre régions. Les trois premières ne contiennent qu'un point et les arbres d'échantillonnage correspondants sont réduits à un seul noeud. Dans le dernier noeud de l'arbre des régions, trois points ont été échantillonnés. Ils sont accessibles, dans l'ordre inverse de

leur échantillonnage, par la liste de points de la feuille de l'arbre des régions (F) à laquelle elle appartient. Ils sont également accessibles par le parcours entier de l'arbre d'échantillonnage propre à F.

L'algorithme d'échantillonnage d'un arbre d'échantillonnage est le suivant :

```
Tirer un point P dans le noeud N

    SI N est une feuille ALORS
        SI N est vide ALORS
            P ← Choix uniforme d'un point dans N
        SINON SI N n'est pas minimal ALORS {il contient un seul point}
            Diviser N en quatre sous-noeuds
            Attacher le point au bon fils
            i ← 1
            RÉPÉTER
                P ← Tirer un point dans le fils(i) de N
            JUSQU'À P ≠ vide OU i ≠ 4
    SINON SI N n'est pas plein ALORS
        i ← 1
        RÉPÉTER
            P ← Tirer un point dans le fils(i) de N
        JUSQU'À P ≠ vide OU i ≠ 4
    SINON
        P ← vide
```

Si la méthode ne fournit pas toutes les qualités souhaitées ci-dessus, notamment en ce qui concerne l'uniformité de répartition des points, elle est en revanche totalement adaptative, rapide et permet de calculer p_z la probabilité de tirage d'un nouveau point dans la zone Z sachant que n points y ont déjà été échantillonnés :

$$p_z = \frac{1}{n - \mathrm{card}(Z)}$$

où l'on reconnait une loi hypergéométrique.

La reconstruction que nous avons choisie prend deux formes différentes selon que la taille d'une feuille de l'arbre d'échantillonnage est supérieure ou strictement inférieure au pixel. La couleur d'une feuille de taille supérieure ou égale au pixel et possédant un point est la couleur de ce point. Lorsque la taille est inférieure, la couleur du pixel recouvrant les feuilles est la moyenne de leur couleur. Enfin, lorsqu'une feuille n'a pas de point, sa couleur est donnée par la couleur moyenne des feuilles à points ayant le même ancêtre. La moyenne s'entend, ici aussi, dans le système L*u*v*. Ce schéma d'interpolation / reconstruction a été retenu pour sa simplicité et sa rapidité plus que pour son adéquation visuelle. On notera en effet qu'il ne garantit pas la continuité de la couleur d'une feuille à l'autre, ce qui peut entraîner des effets de bandes de Mach sur l'image reconstruite. Cette propriété n'est du reste pas désirable sur l'ensemble d'une image (ni verifiée sur la plupart d'entre elles) mais seulement sur certaines portions difficiles à qualifier a priori. Ceci étant, ce défaut ne s'est révélé gênant sur aucune des images que nous avons pu réaliser par notre méthode.

IV.3.5 Outil de validation

L'introduction de notre critère d'homogénéité ainsi que la mise en oeuvre de notre processus progressif nécessite une méthode de validation qui puisse justifier ou infirmer nos choix. Pour ce faire, il est nécessaire de pouvoir comparer une image de référence, l'image complétement calculée, à l'image obtenue progressivement. La comparaison d'images est un problème très délicat qui a été peu abordé dans la littérature. Un certain nombre d'articles portent sur la différence de visualisation de plages uniformes [Rich et al 92], d'autres s'intéressent à des mesures globales de différence entre images comme Rushmeier et al [Rush et al 95] ou encore comme Dinet [Dine 94], ce qui est plus proche de nos préoccupations. L'approche de Dinet est intéressante parce que combinant des différences colorimétriques aussi bien que spatiales et tenant compte de l'acuité visuelle. Malheureusement, elle ne s'accomode pas des zones homogènes d'une image, dommage. La méthode utilisée par Rushmeier et al. compare des scènes réelles à des scènes calculées par l'utilisation de transformées de Fourier permettant de s'abstraire des problèmes d'alignements approximatifs entre modélisations géométriques et caractéristiques réelles. Nous n'avons pas, à proprement parler, de problème de calage géométrique entre image étalon et image progressive. Les régions dont la taille est celle du pixel ont la même couleur que l'étalon calculé à cette définition. Les différences n'interviennent éventuellement que pour des régions dont la taille est supérieure à la définition de l'image étalon. Sur ces régions, si la différence colorimétrique moyenne entre image référence et résultat progressif est inférieure à un certain seuil de tolérance visuelle, tout comme l'écart-type de cette différence, les deux images sont indiscernables. Concrètement, si $\overline{\Delta}_{uv}$, la distance L*u*v* moyenne sur une région, est nulle et si σ_{uv}, son écart-type, est inférieur à l'unité, il est impossible de distinguer une différence entre cette région et son homologue progressif. C'est en fait une conséquence naturelle de la définition de la différence L*u*v*. L'extension de notre mesure à l'ensemble d'une image est plus problèmatique dans le sens où une moyenne et un écart-type faibles ne signifient pas une similarité parfaite. En effet, cette mesure ne prend en compte ni les effets de contrastes simultanés, ni ceux provoqués par discontinuité (bande de Mach).

Cette mesure n'est donc pas une panacée pour comparer deux images quelconques. L'interprétation du couple moyenne/dispersion est aisée pour de petites valeurs, plus délicate pour des couples élévés puisqu'elle n'explique pas d'où vient la différence. De même, elle ne permet pas d'ordonnancer clairement des valeurs de couples proches pour une suite d'images. Dans ce cas en effet, il est difficile de dire quelle est l'image la plus proche de l'original. Il faudrait pour cela avoir une mesure de la répartition géométrique de l'erreur. Notre but avec ce procédé, n'est pas d'avoir une mesure absolue de la différence entre images, mais bien plutôt une indication relative qui nous permette de mettre au point les différents paramètres de notre test. Il s'agit plus d'une question de direction que de position : dans quel sens les variations de nos paramètres influencent-elles la différence à l'original. Malgré ces défauts, au vu des nombreux tests que nous avons effectués et compte tenu de notre processus progressif, nous sommes convaincu de la pertinence de cette distance pour notre problème particulier.

Notre méthode progressive n'est liée au tracé de rayons que par la fourniture d'une couleur en un point, il n'est donc pas indispensable pour sa mise au point pour peu que l'on sache fournir cette couleur. Nous allons donc valider notre méthode et ses paramètres sur un panel d'images test. Images de scènes réelles que nous avons choisies pour leur variété tant géométrique que colorimétrique : scènes d'extérieurs naturels, objets manufacturés, tableaux de peintres, images fabriquées (dessin animé, bande dessinée, image de synthèse).

Fig. 47: Feu d'artifice (39 502 couleurs différentes)

Fig. 50: Mongolfières (31 801 couleurs différentes)

Fig. 48: Bande dessinée (196 couleurs différentes)

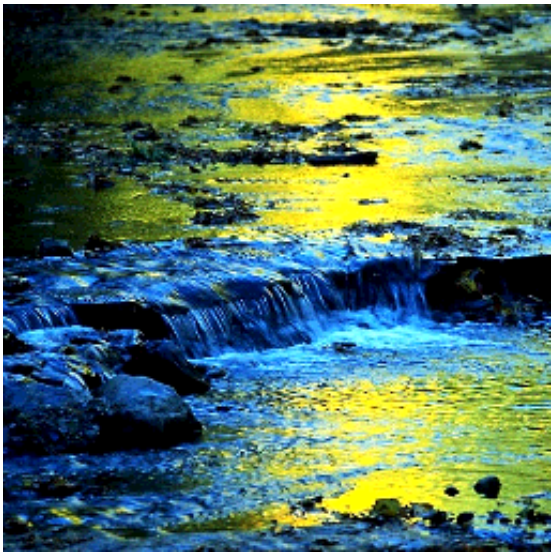


Fig. 51: Torrent (59 164couleurs différentes)

Fig. 49: Factory (64 coul. différentes - synthèse)

Fig. 52: Flour power (36 225 coul. différentes)

Fig. 53: Geyser (51 377 couleurs différentes)

Fig. 56: Horn-Fleig (245 coul. diff. - synthèse)

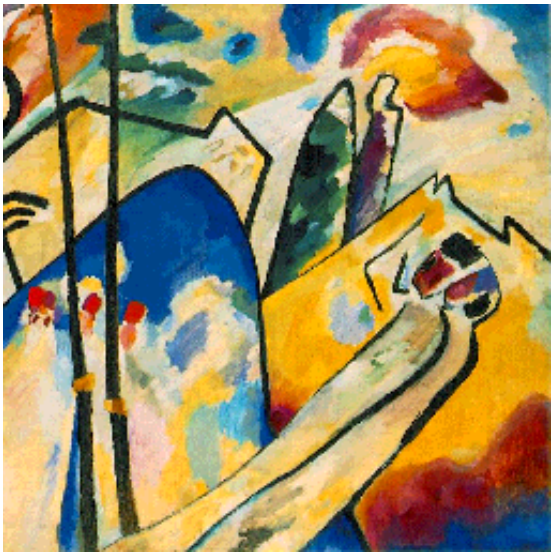


Fig. 54: Kandinski (59 480 couleurs différentes)

Fig. 57: Klee (61 261 couleurs différentes)

Fig. 55: Manet (51 711 couleurs différentes)

Fig. 58: Matisse (55 634 couleurs différentes)

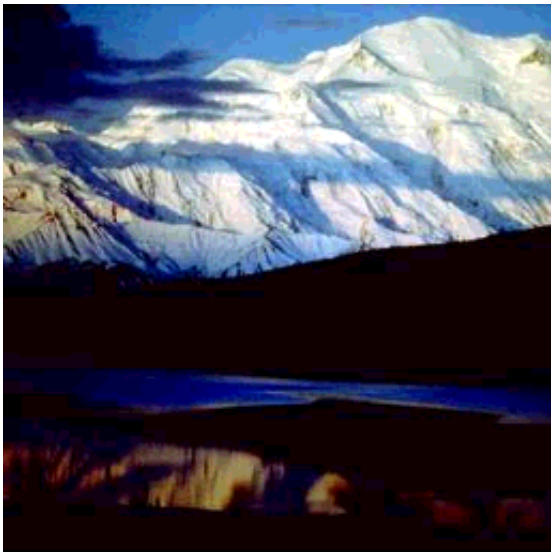


Fig. 59: McKinley (33 461 couleurs différentes)

Fig. 62: Noir et blanc (256 niveaux de gris)

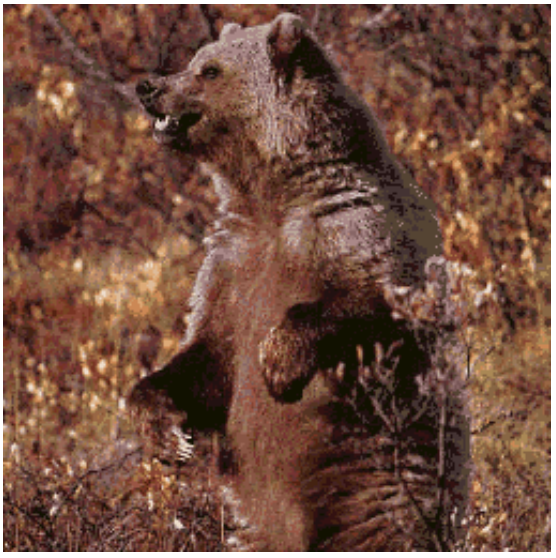


Fig. 60: Ours (54 couleurs différentes)

Fig. 63: Dessin animé (242 couleurs différentes)

Fig. 61: Roseaux (58 102 couleurs différentes)

Fig. 64: Tzigane (255 couleurs différentes)

Fig. 65: Van Gogh (60 060 couleurs différentes)

Fig. 67: Vitrail (38 009 couleurs différentes)

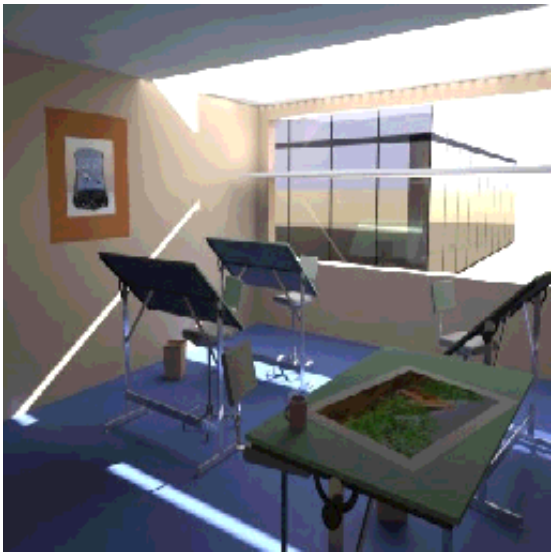


Fig. 66: Bureau (19 258 coul. diff. - synthèse)

IV.4 TSRV appliqué à l'image

IV.4.1 Taille de la zone de tolérance

Le premier paramètre que nous avons à déterminer est la taille de la sphère d'homogénéité. Pour ce faire, les images tests sont découpées jusqu'à l'homogénéité de l'ensemble de leurs régions. L'homogénéité que nous utilisons ici est une espèce d'homogénéité parfaite où l'on exige 100 % des points d'une région à l'intérieur de la sphère. Nous faisons donc varier l'étendue ε en exigeant pour toute région R :

$$\forall y \in R \qquad \Delta_{uv}(\overline{C}_R, C_y) \leq \varepsilon$$

La couleur des régions est donnée par la couleur moyenne des points qu'elles contiennent. Lorsque toutes les régions sont homogènes, l'image ainsi générée est comparée à l'original avec notre procédé. On s'affranchit ainsi de l'échantillonnage et l'on teste du même coup la pertinence de la moyenne. Nous avons calculé les différences d'images

pour des valeurs de ε de 0,5 à 10 par pas de 0,5. Les résultats ne sont donnés que pour les valeurs de ε maximales pour lesquelles la somme $\overline{\Delta}_{uv} + \sigma_{uv}$ est inférieure à 2,5, valeur jugée acceptable pour une bonne qualité d'image. La dernière colonne du tableau ci-après donne le pourcentage de points différents entre l'image étalon et l'image calculée par la méthode ci-dessus.

Image	Étendue ε	Distance moyenne	Écart-type	proportion de points différents
Mongolfières	4,5	1,093	1,204	60,33%
Farine	5	0,815	1,498	32,14%
Klee	5,5	0,826	1,399	32,40%
Kandinski	5,5	0,900	1,395	37,25%
Geiser	5,5	1,032	1,424	44,93%
Bureau	5,5	1,161	1,211	68,01%
HF	6	0,822	1,559	27,42%
Van Gogh	6	0,849	1,519	29,79%
Mc Kinley	6	0,913	1,548	34,71%
Dessin Animé	6,5	0,685	1,549	20,38%
Manet	6,5	0,796	1,498	29,40%
Factory	7	0,746	1,707	21,24%
Matisse	7,5	0,679	1,597	19,33%
Noir et blanc	7,5	0,819	1,597	32,32%
Ours	8	0,552	1,615	12,80%
Eau	8,5	0,645	1,705	15,51%
Roseau	9	0,599	1,748	12,80%
BD	10	0,314	1,272	7,57%
Vitrail	10	0,387	1,540	7,42%
Artifice	10	0,466	1,632	9,90%
Tzigane	10	0,473	1,567	11,82%

Tableau 3: Distance colorimétrique en fonction de la taille de la zone tolérance

La valeur maximale acceptable pour l'étendue devrait donc être inférieure ou égale à 4,5 pour une zone de tolérance pleine à 100 %. Dans les images du panel, on peut constater que cette limite est largement suffisante dans 95 % des cas.

IV.4.2 Proportion homogène

Un test similaire a été réalisé pour déterminer p_h, la *proportion homogène*, autrement dit la proportion de points d'une région appartenant à la zone de tolérance au-delà de laquelle cette région est dite homogène. Cette fois, on n'exige plus 100 % des points dans l'intervalle mais $100 \times p_h$ % seulement pour la classification homogène. Pour

établir les résultats ci-dessous, les tests ont été réalisés en faisant varier p_h de 0,7 à 0,95 par pas de 0,5 pour des valeurs de ε variant de 0,5 à 2,5 par pas de 0,5. Le tableau ci-après est un résumé de l'ensemble de ces résultats où l'on a reporté les images qui, à p_h et ε fixés, donnent la somme $\overline{\Delta}_{uv} + \sigma_{uv}$ maximale et inférieure à 2,5.

<table>
<tr><td></td><td></td><td colspan="10">ε</td></tr>
<tr><td></td><td></td><td colspan="2">0,5</td><td colspan="2">1</td><td colspan="2">1,5</td><td colspan="2">2</td><td colspan="2">2,5</td></tr>
<tr><td rowspan="10">p_h</td><td rowspan="2">0,95</td><td colspan="2">BD</td><td colspan="2">BD</td><td colspan="2">BD</td><td colspan="2">BD</td><td colspan="2">BD</td></tr>
<tr><td>0</td><td>0,38</td><td>0,02</td><td>0,8</td><td>0,05</td><td>1,4</td><td>0,09</td><td>1,88</td><td>0,15</td><td>2,33</td></tr>
<tr><td rowspan="2">0,9</td><td colspan="2">Mc Kinley</td><td colspan="2">BD</td><td colspan="2">BD</td><td colspan="2">BD</td><td colspan="2">BD</td></tr>
<tr><td>0,02</td><td>0,44</td><td>0,02</td><td>0,8</td><td>0,06</td><td>1,4</td><td>0,1</td><td>1,89</td><td>0,15</td><td>2,34</td></tr>
<tr><td rowspan="2">0,85</td><td colspan="2">Mc Kinley</td><td colspan="2">Factory</td><td colspan="2">BD</td><td colspan="2">BD</td><td colspan="2">BD</td></tr>
<tr><td>0,02</td><td>0,44</td><td>0,1</td><td>0,75</td><td>0,06</td><td>1,4</td><td>0,1</td><td>1,9</td><td>0,15</td><td>2,34</td></tr>
<tr><td rowspan="2">0,8</td><td colspan="2">Mc Kinley</td><td colspan="2">Mc Kinley</td><td colspan="2">Mc Kinley</td><td colspan="2">BD</td><td colspan="2" rowspan="6">$\overline{\Delta}_{uv} + \sigma_{uv} >$ 2,5</td></tr>
<tr><td>0,03</td><td>0,45</td><td>0,1</td><td>1,2</td><td>0,19</td><td>1,36</td><td>0,1</td><td>1,92</td></tr>
<tr><td rowspan="2">0,75</td><td colspan="2">Factory</td><td colspan="2">Mc Kinley</td><td colspan="2">Mc Kinley</td><td colspan="2">Mc Kinley</td></tr>
<tr><td>0,03</td><td>0,46</td><td>0,12</td><td>1,22</td><td>0,2</td><td>1,38</td><td>0,37</td><td>1,92</td></tr>
<tr><td rowspan="2">0,7</td><td colspan="2">N. et B.</td><td colspan="2">Mc Kinley</td><td colspan="2">BD</td><td colspan="2" rowspan="2">$\overline{\Delta}_{uv} + \sigma_{uv} >$ 2,5</td></tr>
<tr><td>0,07</td><td>0,65</td><td>0,18</td><td>1,31</td><td>0,12</td><td>1,9</td></tr>
</table>

Tableau 4: Distance colorimétrique en fonction de la proportion homogène

Comme il était prévisible, la valeur de ε maximale pour l'obtention d'une erreur acceptable a sensiblement diminué par rapport aux résultats obtenus pour $p_h = 1$. La taille maximale acceptable de la zone de tolérance est désormais inférieure ou égale à 2. Dans ces conditions, la proportion homogène doit être au minimum supérieure à 0,75.

IV.4.3 Structure du panel d'images

Pour chacune de nos images test, nous avons calculé p ((44) §IV.3.3), la proportion de points dont la couleur est située dans l'intervalle de tolérance, sur des régions de tailles décroissantes et pour une valeur raisonnable de ε (= 2). Le calcul se fait d'abord sur l'ensemble de l'image (où p est en général nulle), puis sur chacun des quarts, puis sur les huitièmes, etc, jusqu'au pixel où, bien sûr, p vaut 1. Les résultats de ces mesures nous ont quelque peu étonné puisqu'elles révèlent trois distributions types que nous ne nous attendions pas à découvrir.

La première est une distribution très inhomogène à toutes les échelles (celles du pixel mise à part) où la grande majorité des points se situe à l'extérieur de l'intervalle de tolérance. La valeur de p est en effet inférieure à 0,1 pour plus de 85 % des individus. Si l'on trace la fréquence d'apparition de p, on obtient des courbes du type suivant :

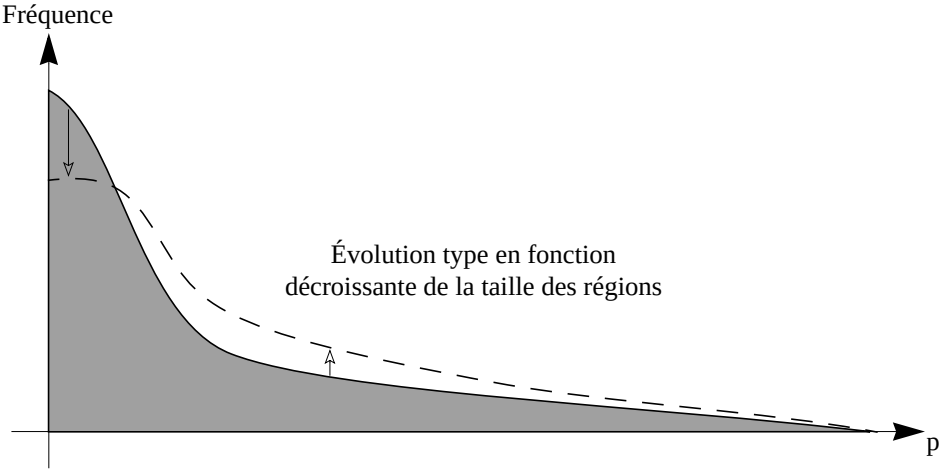
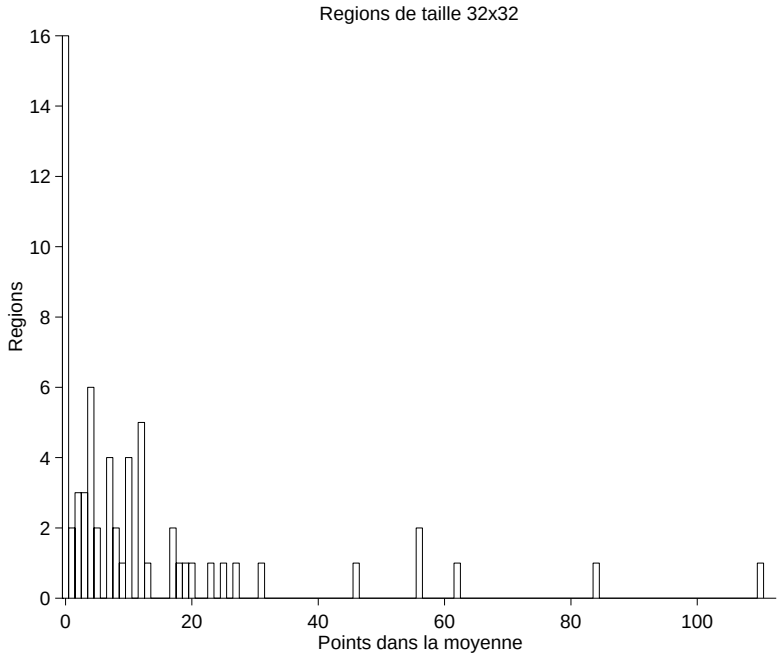
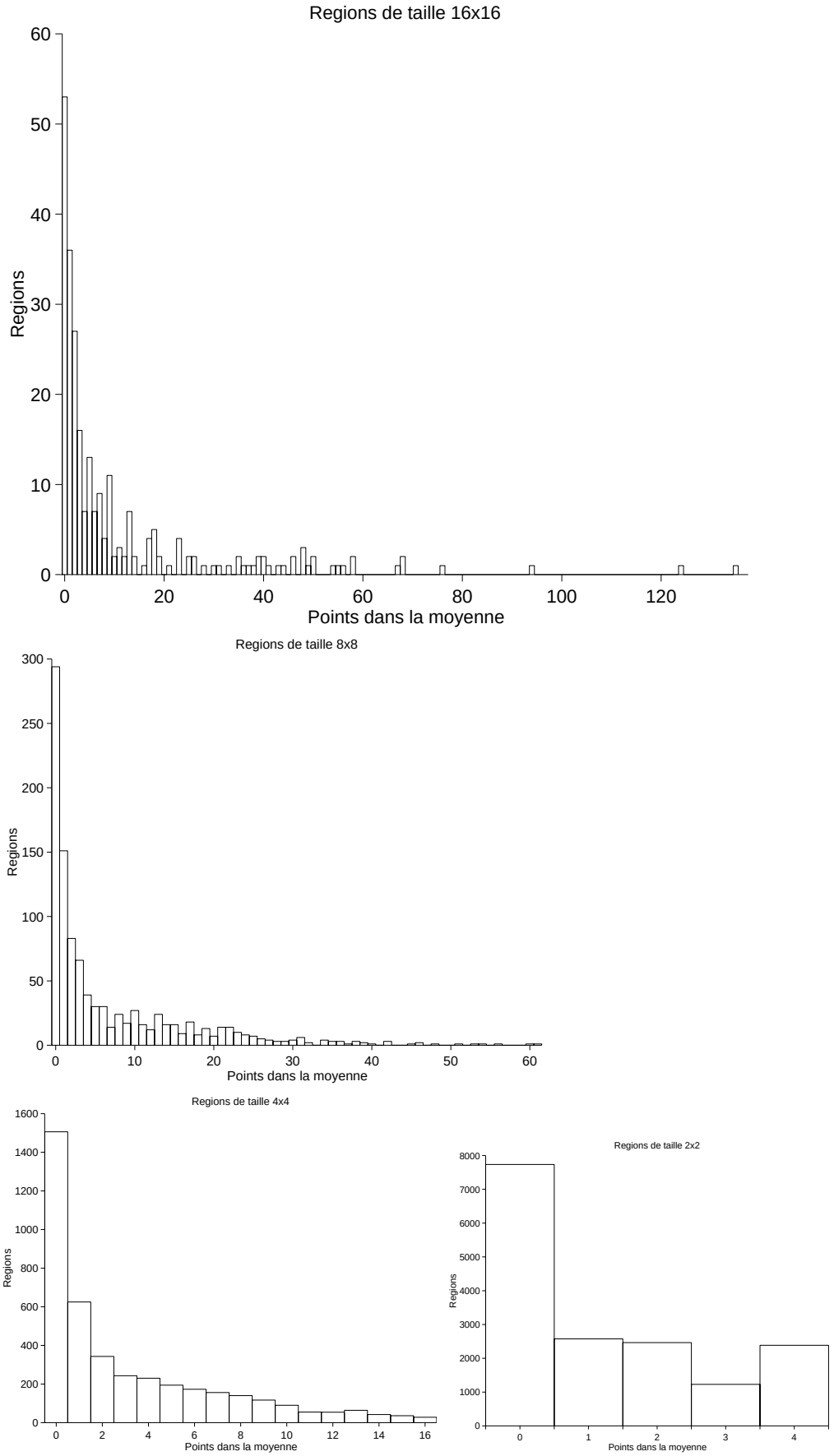


Figure 68: Distribution inhomogène

L'évolution de ces courbes en fonction de la taille est prévisible et commune en fait aux trois types de distributions. Elle consiste en un glissement progressif de l'inhomogène vers l'homogène : les effectifs les plus élevés décroissent pour se répartir sur les proportions intermédiaires. Dans ce premier type de distribution, ce glissement ne conduit jamais à une inversion de tendance, c'est-à-dire à l'émergence de pics homogènes forts. Une évolution typique de ces courbes en fonction de la taille des régions est donnée ci-dessous pour l'image du Geyser.




Regions de taille 16x16
Regions
Points dans la moyenne
Regions de taille 8x8
Regions
Points dans la moyenne
Regions de taille 4x4
Regions
Points dans la moyenne
Regions de taille 2x2
Regions
Points dans la moyenne

La seconde distribution type est mixte. Les distributions sur les régions de grande taille sont très inhomogènes comme il est normal et évoluent, à mesure que la taille diminue, vers des répartitions en U où la portion horizontale représente environ un tiers de la population :

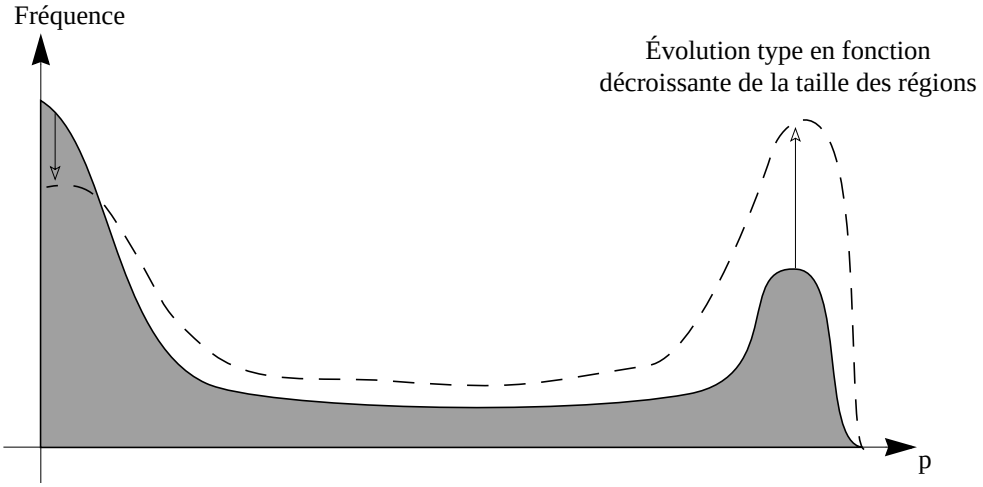
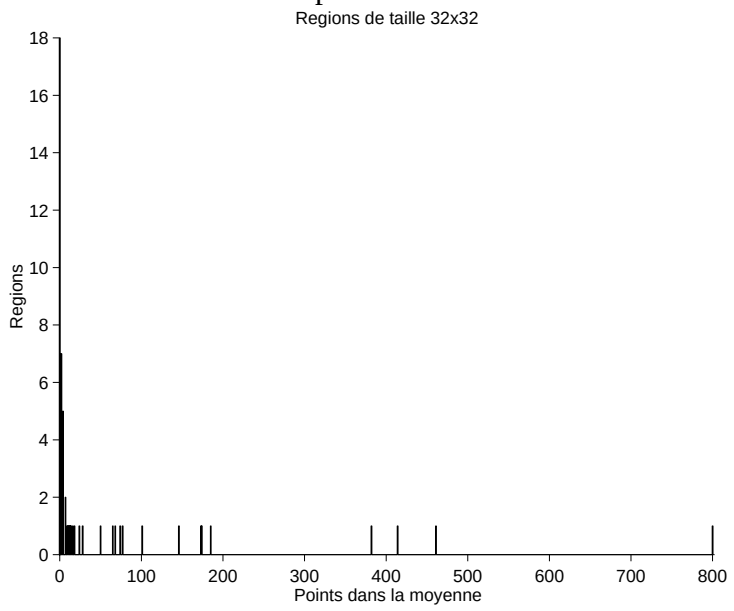
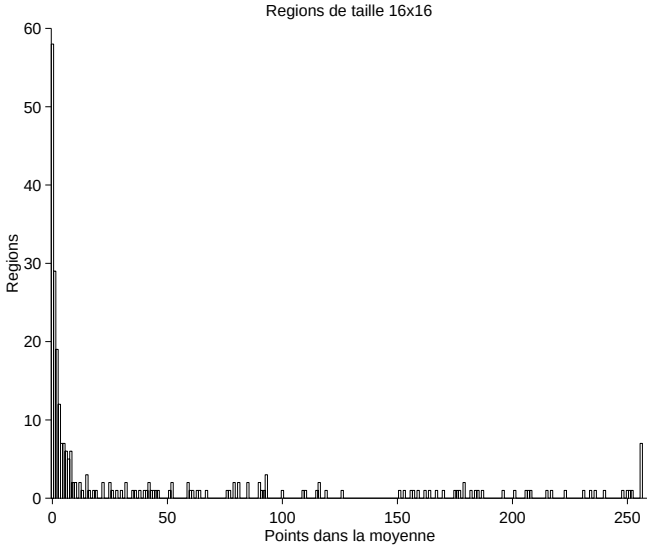
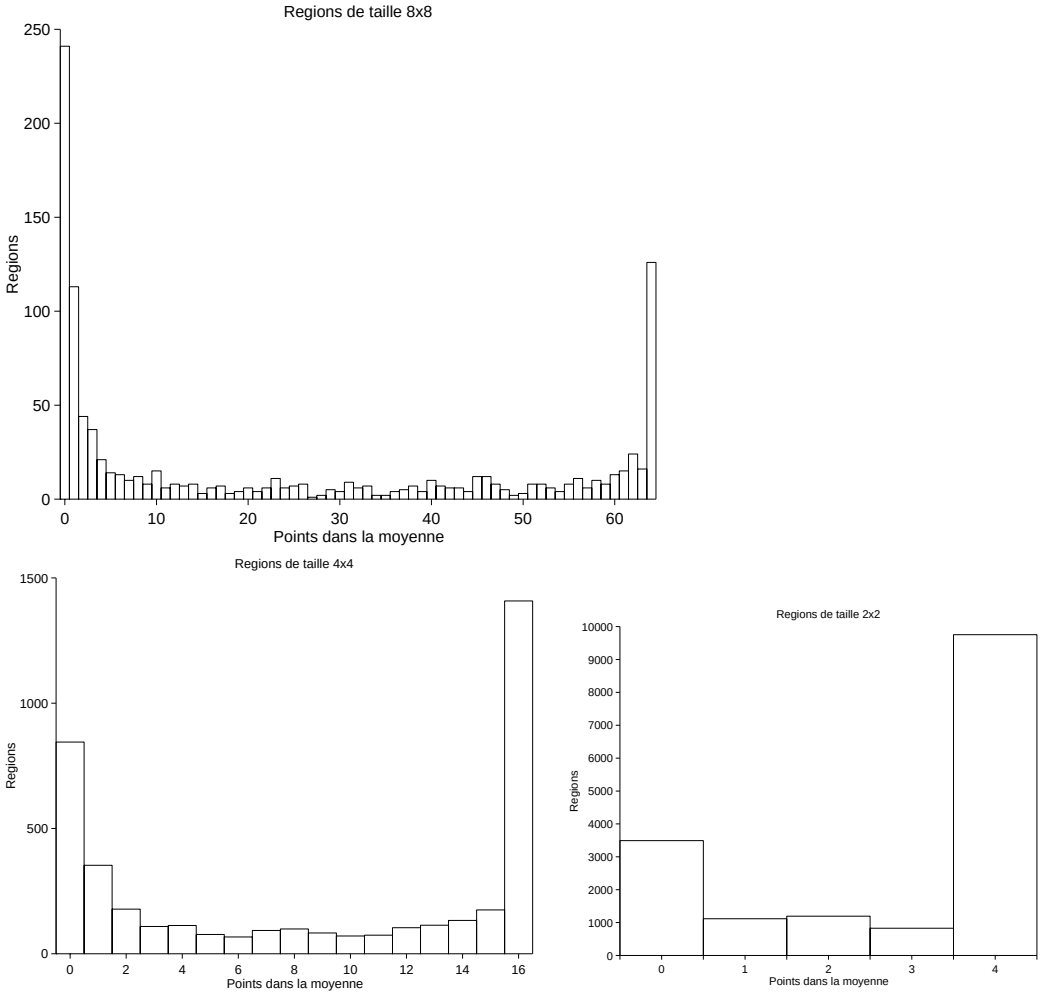


Figure 69: Distribution mixte

L'image du Bureau fournit un exemple de distribution mixte.







La dernière distribution type est intermédiaire entre les deux précédentes. Elle se caractérise par une forte représentation des valeurs centrales de p avec toujours un pic inhomogène important. Sur notre panel d'images, seule celle des mongolfières présente cette caractéristique mais il y a tout lieu de penser qu'elle est sans doute plus commune que ne le laisse penser le panel :

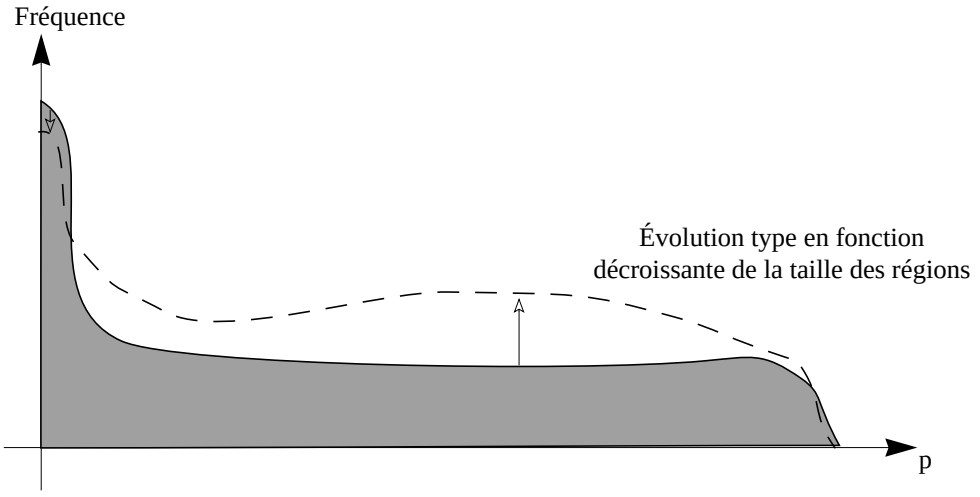
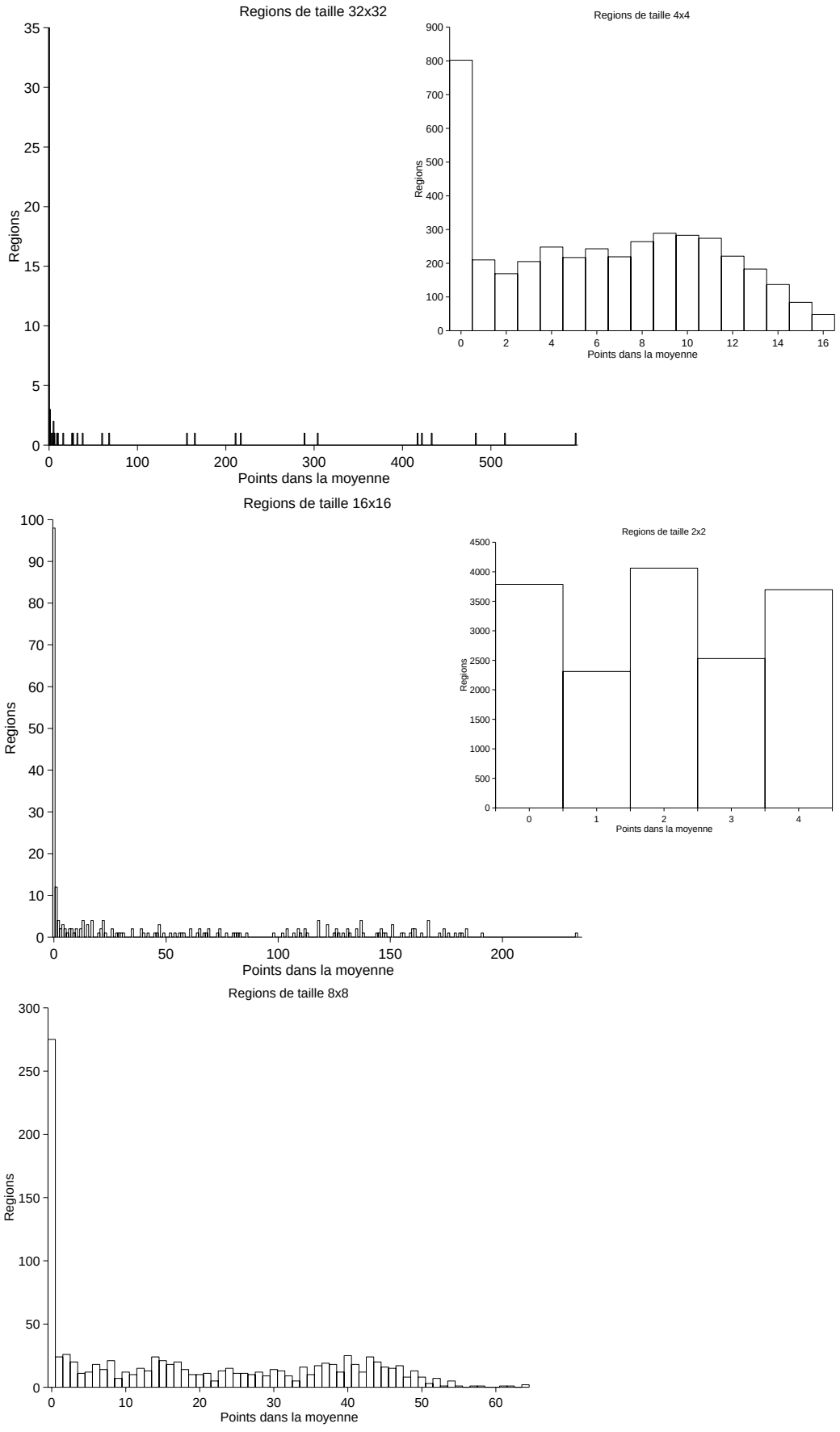


Figure 70: Distribution intermédiaire


Regions de taille 32x32
Regions
Points dans la moyenne
Regions de taille 4x4
Regions
Points dans la moyenne
Regions de taille 16x16
Regions
Points dans la moyenne
Regions de taille 2x2
Regions
Points dans la moyenne
Regions de taille 8x8
Regions
Points dans la moyenne

IV.4.4 Définition du test

Le TSRV requiert de pouvoir déterminer la probabilité d'apparition du caractère de chaque individu, c'est impossible avec p puisque sa loi est inconnue. Il est donc nécessaire de se ramener à des lois de probabilités dont la densité est connue, ce que nous réalisons en introduisant la variable aléatoire x telle que :

$$x = \mathbf{1}_{[\bar{c}_R - \varepsilon, \bar{c}_R + \varepsilon]}(C_\gamma)$$

x suit une loi de Bernouilli de paramètre p dont la fonction de distribution sera notée $f_x(p)$. La probabilité que x soit unitaire, c'est-à-dire que la couleur d'un point appartienne à la zone de tolérance, est alors donnée par :

$$f_h(p) = p$$

La probabilité que la couleur d'un point soit à l'extérieur de la zone de tolérance (x=0) s'écrit de même :

$$f_{\bar{h}}(p) = 1 - p$$

Si l'on note D_R le nombre de points de la région R dont la couleur est située dans la zone de tolérance, p s'exprime simplement par :

$$p = \frac{D_R}{n_R},\ D_R = \sum_R x_i = \sum_R \mathbf{1}_{[\bar{c}_R - \varepsilon, \bar{c}_R + \varepsilon]}(C_{y_i})$$

$n_r = \text{card}(R)$

L'échantillonnage hypergéométrique que nous avons choisi entraîne que les variables aléatoires $x_1,\ldots,x_k$, correspondant au tirage de k points, ne sont pas indépendantes. La probabilité d'apparition de x_2 est liée à celle de x_1, celle de x_3 est conditionnée par x_2 et x_1, etc. La distribution de la suite $(x_1,\ldots,x_k)$ s'écrit donc :

$$p^{(k)} = f_{x_1}\left(\frac{D_R}{n_R}\right) f_{x_2}\left(\frac{D_R - d_1}{n_R - 1}\right) \ldots f_{x_k}\left(\frac{D_R - d_{k-1}}{n_R - k + 1}\right)$$

$$d_j = \sum_{i=0}^{j} x_i \quad (d_k \le D_R), \text{ par convention } d_0 = 0$$

La première hypothèse, H, l'*hypothèse homogène*, est réalisée lorsque $p \ge p_h$ et donc lorsque R est homogène. L'hypothèse $\overline{H}$ est celle de l'inhomogénéité lorsque $p < p_h$. Les *hypothèses composites* telles celle que nous introduisons ($p \ge P$ contre $p < P$) sont généralement traitées dans la littérature sur l'analyse séquentielle par l'introduction de deux probabilités p_0 et p_1, $p_0 < p_1$, et la reformulation des hypothèses avec ces nouvelles quantités ($p \ge p_1$ contre $p \le p_0$). Reformulation sans déformation : la partition de l'espace du paramètre par les hypothèses doit rester complète, c'est-à-dire décrire l'ensemble de cet espace. Les résultats établis pour des hypothèses simples du type «$p = p_1$ contre $p = p_0$», comme nous les avons utilisées jusqu'à présent, sont alors directement applicables à des hypothèses composites si, de plus, les critères de Lehmann [Lehm 59] sont satisfaits, c'est-à-dire si le rapport de vraisemblance est monotone en la statistique, ce qui est le cas ici.

Sous l'hypothèse homogène (H), en notant D_h le nombre de points dans la zone de tolérance, $p_h = D_h / n_R$, la distribution précédente devient :

$$p_h^{(k)} = f_{x_1}\left(\frac{D_h}{n_R}\right) f_{x_2}\left(\frac{D_h - d_1}{n_R - 1}\right) \dots f_{x_k}\left(\frac{D_h - d_{k-1}}{n_R - k + 1}\right)$$

De la même façon sous l'hypothèse $\overline{H}$, on a :

$$p_{\bar{h}}^{(k)} = f_{x_1}\left(\frac{D_{\bar{h}}}{n_R}\right) f_{x_2}\left(\frac{D_{\bar{h}} - d_1}{n_R - 1}\right) \dots f_{x_k}\left(\frac{D_{\bar{h}} - d_{k-1}}{n_R - k + 1}\right)$$

D_h tel que : $p_{\bar{h}} = D_{\bar{h}} / n_R$

Le TSRV que nous utiliserons est donc défini comme suit à l'étape k pour la région R :

R est décrétée homogène avec le risque α si $p_h^{(k)} \geq \left(\frac{1-\beta}{\alpha}\right) p_{\bar{h}}^{(k)}$

R est inhomogène avec le risque β si $p_h^{(k)} \leq \left(\frac{\beta}{1-\alpha}\right) p_{\bar{h}}^{(k)}$

sinon le test se poursuit par le tirage d'un nouveau point

IV.4.5 Test exact

Nous avons supposé jusqu'ici que la couleur moyenne d'une région était connue, ce qui implique en fait de connaître l'image dans son intégralité. Ceci est impossible dans notre cas puisque cela équivaut à calculer l'image complétement. Laissant ce problème de côté pour l'instant, nous nous comporterons comme si nous la connaissions. De même, la définition rigoureuse de $p_{\bar{h}}$ implique que ce paramètre ne peut être choisi librement. D'une part, ce doit être une probabilité tout comme p_h, ce qui implique que la vraie valeur de p_h comme celle de $p_{\bar{h}}$ est différente pour chaque taille de région testée. Dans la suite nous parlerons uniquement de p_h, sa vraie valeur étant sous-entendue :

$$\text{vraie valeur de } p_h = \lfloor p_h n_R \rfloor / n_R$$

où $\lfloor x \rfloor$ dénote le plus grand entier inférieur ou égal à x

D'autre part, une fois p_h fixé, et parce que les hypothèses doivent être complémentaires, $p_{\bar{h}}$ ne peut qu'avoir la définition suivante :

$$p_{\bar{h}} = \frac{D_h - 1}{n_R} \quad \text{avec} \quad D_h = \lfloor p_h n_R \rfloor \tag{45}$$

Pour ces deux raisons, nous parlerons ici de *test exact* en nous intéressant à ses performances et à son comportement général.

IV.4.5.1 Signification des risques

Les risques associés à chaque hypothèse contrôlent la qualité du résultat final ainsi que la confiance que l'on peut lui accorder. Nous avons choisi de supporter un risque plus grand pour α que pour β. On préfère ainsi déclarer à tort l'homogénéité alors qu'il y a inhomogénéité (risque α) plus souvent que l'inverse, à savoir déclarer l'inhomogénéité alors qu'il y a homogénéité (risque β). La décision la plus coûteuse est en effet

celle de l'inhomogénéité puisqu'elle conduit à la subdivision et à la relance de tests et donc de rayons dans chacune des sous-régions. Nous souhaitons donc que l'effort consenti le soit à bon escient et conduise à la meilleure efficacité possible. Déclarer une région inhomogène alors qu'elle ne l'est pas est donc une perte de temps que l'on souhaite réduire au minimum. La situation opposée est moins génante dans le sens où l'utilisateur a la possibilité de subdiviser manuellement des régions qu'il juge inhomogènes pour y relancer un test. Ceci n'est tolérable que si l'erreur de décision, sur H comme sur $\overline{H}$, reste faible : les risques doivent de toute façon être petits.

IV.4.5.1.1 Région critique

Les valeurs de α et β déterminent la taille d'une *région critique* à l'intérieur de laquelle le test ne peut prendre aucune décision. Cette région correspond à la troisième alternative du TSRV lorsqu'un nouvel échantillon est nécessaire, situation dans laquelle on a :

$$\frac{\beta}{1-\alpha} < p_h^{(k)} / p_{\bar{h}}^{(k)} < \frac{1-\beta}{\alpha}$$

La taille de la région critique influe de facto sur le *temps d'arrêt* du test. Plus elle est grande, plus il faudra potentiellement de temps pour en sortir. Le déroulement du test est en fait une marche aléatoire à l'intérieur de cette région, marche qui s'arrête dès que l'une de ses frontières est franchie. Cette marche, comme il est visible dans la définition des probabilités, ne dépend que du nombre de points dans la moyenne (d_k) et du nombre d'échantillons (k).

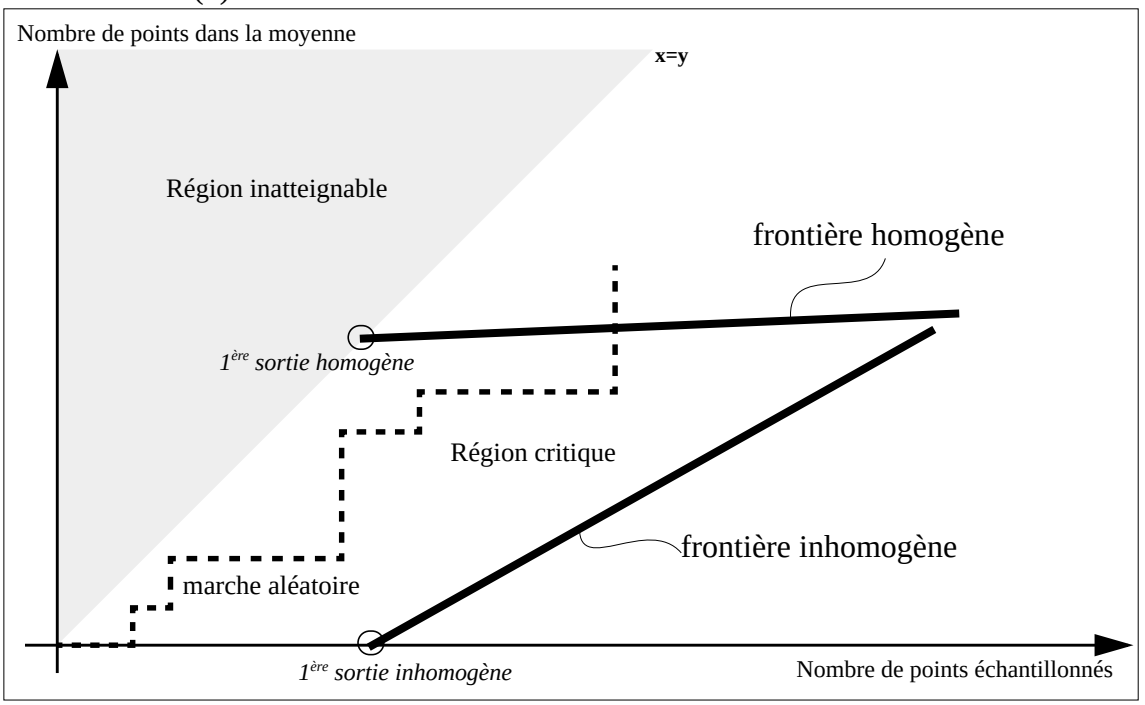


Figure 71: Marche aléatoire du TSRV dans la région critique

Les sorties possibles du test ne peuvent avoir lieu que sur certains points précis de deux frontières de la région critique. Ces points dépendent non seulement de α et β mais également et avec de très fortes implications, notamment sur les risques réels, de l'effectif de la région testée.

Pour les caractériser, on exprime $p_{\bar{h}}^{(k)}$ et $p_h^{(k)}$ sous la forme suivante :

$$p_{\bar{h}}^{(k)} = \frac{C_{D_{\bar{h}}}^{d} C_{n_R - D_{\bar{h}}}^{k-d}}{C_{n_R}^{k}} \qquad \text{où } d = d_{k-1}$$

$$p_h^{(k)} = \frac{C_{D_h}^{d} C_{n_R - D_h}^{k-d}}{C_{n_R}^{k}}$$

Dans le test exact, $D_{\bar{h}} = D_h - 1$, d'où le rapport de vraisemblance :

$$\frac{p_h^{(k)}}{p_{\bar{h}}^{(k)}} = \frac{C_{D_h}^{d} C_{n_R - D_h}^{k-d}}{C_{D_{\bar{h}}}^{d} C_{n_R - D_{\bar{h}}}^{k-d}} = \frac{D_h(n_R - D_h + 1 - k + d)}{(D_h - d)(n_R - D_h + 1)}$$

Une sortie homogène a lieu lorsque :

$$\frac{p_h^{(k)}}{p_{\bar{h}}^{(k)}} \geq \frac{1-\beta}{\alpha} \Leftrightarrow d \geq \frac{D_h(k + (n_R - D_h + 1)((1-\beta)/\alpha - 1))}{D_h + (n_R - D_h + 1)(1-\beta)/\alpha}$$

De même pour une sortie inhomogène :

$$\frac{p_h^{(k)}}{p_{\bar{h}}^{(k)}} \leq \frac{\beta}{1-\alpha} \Leftrightarrow d \leq \frac{D_h(k + (n_R - D_h + 1)(\beta/(1-\alpha) - 1))}{D_h + (n_R - D_h + 1)\beta/(1-\alpha)}$$

Lorsque le nombre d'échantillons, k, est fixé, on peut ainsi déterminer le nombre minimum de points dans la zone de tolérance nécessaires à une sortie homogène (d_h) ou inhomogène ($d_{\bar{h}}$) qui définissent les frontières discrètes de la région critique :

$$d_h = \left\lceil \frac{D_h(k + (n_R - D_h + 1)((1-\beta)/\alpha - 1))}{D_h + (n_R - D_h + 1)(1-\beta)/\alpha} \right\rceil, \; d_h \leq k$$

$$d_{\bar{h}} = \left\lfloor \frac{D_h(k + (n_R - D_h + 1)(\beta/(1-\alpha) - 1))}{D_h + (n_R - D_h + 1)\beta/(1-\alpha)} \right\rfloor, \; d_{\bar{h}} \leq k$$

où $\lceil x \rceil$ dénote l'entier supérieur ou égal à x

Pour une région de taille 2×2, la région critique a la forme suivante avec $D_h = 3$ ($p_h = 0{,}75$), $\alpha \leq 0{,}1$, $\beta \leq 0{,}1$:

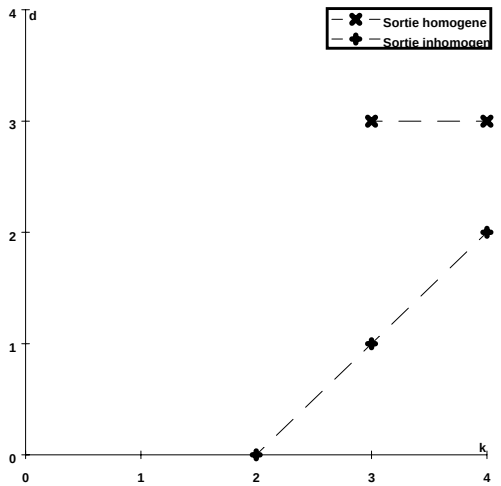


Figure 72: Région critique pour une région de taille 2×2

On constate que le test est ici parfaitement sûr puisqu'à décision certaine : pour qu'il y ait homogénéité, il exige 3 points dans la moyenne, pour l'inhomogénéité, il faut au moins deux points hors de la moyenne, ce qui implique qu'il ne peut y en avoir 3 dans la moyenne! Les risques des décisions sont donc, dans ce cas, nuls. On retrouve souvent la même situation pour les régions de taille 4×4 :

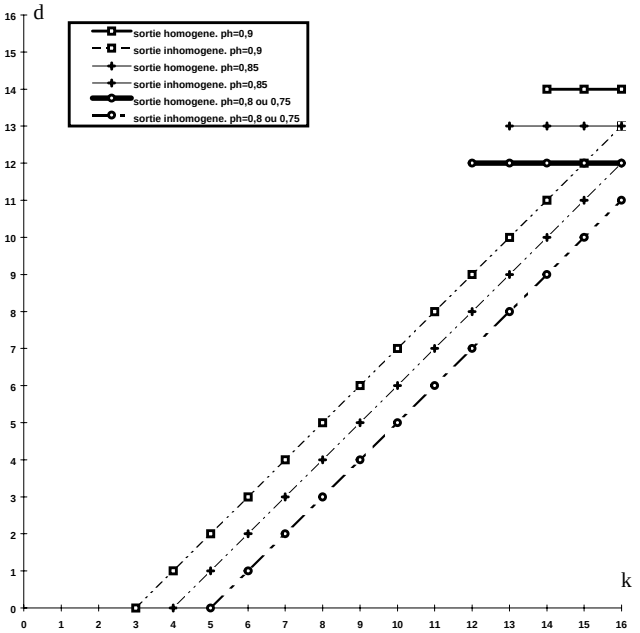


Figure 73: Région critique pour une région de taille 4×4.
$0{,}01 \le \alpha \le 0{,}06$, $0{,}01 \le \beta \le 0{,}1$, $0{,}9 \ge p_h \ge 0{,}75$

Le risque «homogène» (α) est en fait nul pour $\alpha \le 0{,}06$, $\beta \le 0{,}1$ et $p_h \ge 0{,}75$ parce qu'il faut 12 (= 0,75*16) points dans la moyenne ou 14 (= 0,9*16) pour que le test prenne la décision homogène. Pour $\alpha = 0{,}1$, le risque n'est pas nul mais très faible (11 points au lieu de 12, 13 au lieu de 14). Pour le risque «inhomogène» (β), le risque nul est également certain pour les même raisons.

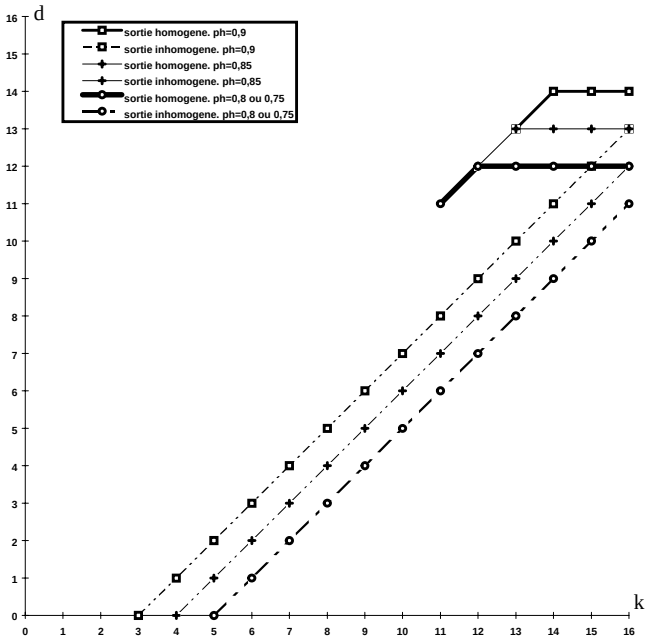


Figure 74: Région critique pour une région de taille 4×4.
$\alpha = 0{,}1$, $0{,}01 \le \beta \le 0{,}1$, $0{,}9 \ge p_h \ge 0{,}75$

Pour des tailles de régions de 8×8, le risque «inhomogène» est nul, le risque homogène n'est pas nul mais doit être très faible (de 1 à 4 points d'écart avec les points de sortie sûrs à 100 %).

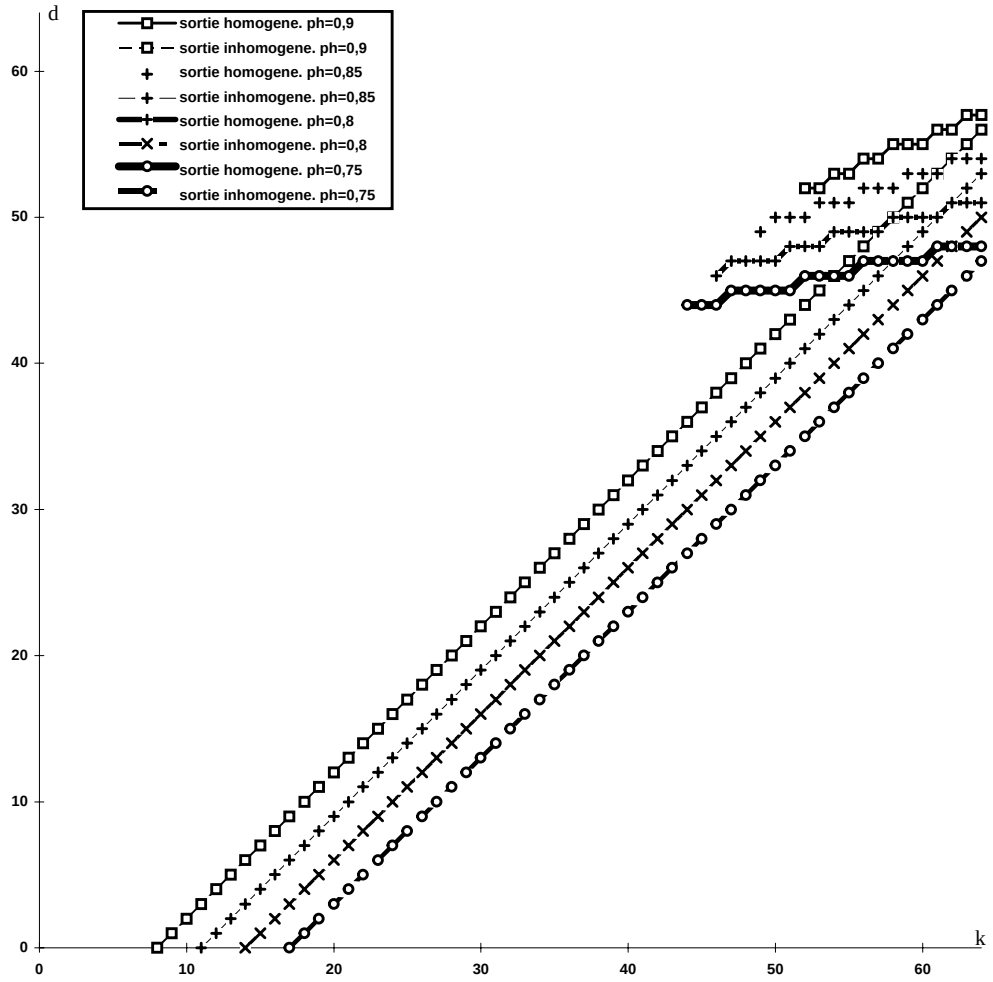


Figure 75: Région critique pour une région de taille 8×8.
$0,01 \leq \alpha \leq 0,1$, $\beta = 0,1$, $0,9 \geq p_h \geq 0,75$

Les mêmes constatations valent pour des tailles de 16×16 avec un écart plus sensible pour l'homogène ayant pour conséquence un risque sans doute plus proche du β théorique (entre 0,1 et 0,01).

IV.4.5.1.2 Premières sorties du test

Deux des points de sortie du test particulièrement intéressants sont les premiers instants de sortie homogène et inhomogène possibles. Ce sont de bonnes indications de la rapidité du test. La première sortie homogène possible a lieu lorsque tous les individus échantillonnés sont dans la zone de tolérance (section verticale de la marche aléatoire) et que le rapport de vraisemblance dépasse la frontière homogène. Le premier instant de sortie homogène possible, k_h, s'écrit alors :

$$k_h = \left\lceil D_h - 1 - \frac{\alpha D_h}{1-\beta} \right\rceil$$

Symétriquement la première sortie inhomogène correspond au fait que tous les individus sont extérieurs à la zone de tolérance (section horizontale de la marche aléatoire) et que le rapport de vraisemblance est inférieur à la frontière inhomogène. Le premier instant de sortie inhomogène possible, $k_{\bar{h}}$, s'exprime par :

$$k_{\bar{h}} = \left\lceil \bar{D}_h - \frac{\beta(\bar{D}_h + 1)}{1 - \alpha} \right\rceil$$

$$\bar{D}_h = n_R + D_h$$

En fixant α et β, on peut donc déduire les instants correspondants en fonction de D_h . Les risques fixés, les rapports $k_{\bar{h}} / \bar{D}_h$ et k_h / D_h le sont également. Nous avons calculé ces rapports pour un certains nombre de valeurs de α et β :

		α			
		0,01	0,05	0,1	0,15
β	0,01	99 % 99 %	95 % 99 %	90 % 99 %	85 % 99 %
	0,05	99 % 95 %	95 % 95 %	89 % 94 %	84 % 94 %
	0,1	99 % 90 %	94 % 90 %	89 % 89 %	83 % 88 %
	0,15	99 % 85 %	94 % 84 %	88 % 83 %	82 % 82 %

Tableau 5: Proportion de points à échantillonner avant la première sortie possible homogène (cadré à gauche) et inhomogène (cadré à droite). Valeurs approximatives pour des tailles de régions supérieures à 16 x 16 et pour $p_h \approx 0,9$

Comme il était prévisible compte tenu des décisions certaines, ces rapports sont très élevés. Les performances du test dans ces conditions seront assez faibles, i.e très lentes. En général comme D_h est très supérieur à $\bar{D}_h$, la détection inhomogène devrait être assez rapide, en revanche la décision homogène sera certainement beaucoup plus longue. Cette lenteur provient de la proximité de $p_{\bar{h}}$ et p_h ainsi que de l'hypergéométrique, les deux probabilités étant presque confondues, leur rapport tend très lentement hors de la région critique. La qualification totale d'une image dépend en fait très fortement du temps de qualification homogène puisque le processus progressif se termine par l'homogénéité de toutes les régions. En conséquence, même si la décision inhomogène a de grandes chances d'être rapide, la longueur de la durée de qualification homogène pénalisera fortement l'ensemble et aboutira à de piètres performances même si le résultat est souvent sûr à 100 %. Le nombre total de points nécessaires à l'aboutissement du processus sera en effet proche de la proportion k_h / n_R et sans doute supérieur à p_h.

IV.4.5.2 Risques réels

IV.4.5.2.1 Test exact non biaisé

Malgré la lenteur du test exact, l'examen des *risques réels*, α_r et β_r, i.e mesurés pendant l'expérience, peut fournir de précieux renseignements sur la qualité du procédé. Rappelons que la mesure des risques consiste à considérer la réalité, homogène ou inhomogène, d'une région et à dénombrer les mauvaises décisions prises compte tenu de cette réalité. α_r est donc le ratio des décisions homogènes lorsque la région est réellement inhomogène sur le nombre total des régions réellement inhomogènes, symétriquement pour β_r. La mesure a été réalisée sur l'ensemble des régions, pour chaque sortie du test et pour l'ensemble des passes progressives jusqu'à obtention de l'homogénéité de toutes les régions. Les points échantillonnés pendant une passe ne sont jamais réutilisés par les passes ultérieures afin d'éviter le biais que cela entraînerait, autrement dit, chaque test est effectué sur une région vierge de points.

Le test exact a été réalisé sur l'ensemble du panel avec $\varepsilon = 2$ pour des valeurs de p_h successivement égales à 0,8, 0,85 et 0,9 en faisant varier α et β pour chacune de ces valeurs entre 0,07 et 0,01 par pas de 0,02.

Quelles que soient les valeurs de paramètres employés, les risques globaux réels sont toujours nuls. En regard d'autres TSRV ou même de tests d'hypothèses plus classiques, cette situation est proprement extravagante. La nature aléatoire de tout test statistique contient intrinsèquement et de façon incontournable la notion de risques. Ceux-ci peuvent-être rendus aussi petits que nécessaire (avec un bémol sans doute pour les populations finies à petit effectif) mais la nullité parfaite est presque impossible. L'explication de ce comportement étrange réside, comme nous l'avons vu, dans les décisions certaines mais également dans la distribution du caractère mesuré. La structure des images est très contrastée et rassemble une énorme proportion de la population sur les valeurs extrêmes de p. Autrement dit, la valeur réelle du paramètre est très fréquemment très éloignée de p_h. En conséquence, le risque réel est largement inférieur au risque annoncé. Les situations les plus «risquées» pour le test sont celles où p est proche de p_h parce qu'alors il est difficile de discriminer le vrai caractère de p. On peut dire, de façon un peu excessive, que le risque réel est inversement proportionel à la distance entre p et p_h (ou peut-être même à quelque fonction exponentielle de celle-ci). α_r est ainsi nul par «définition», β_r l'est parce qu'en général p est très éloigné de p_h lorsqu'il y a inhomogénéité.

IV.4.5.2.2 Test exact avec recyclage des échantillons

Le fait de ne pas recycler les points déjà échantillonnés est assez pénalisant. Il implique de recalculer plusieurs fois un point donné ou de complexifier l'échantillonnage pour récupérer sa couleur si elle est déjà connue. Pour éviter ces désagréments, nous avons préféré adopter une autre stratégie qui consiste à recycler les échantillons déjà tirés. Lorsqu'une région est déclarée inhomogène, les tests de ses sous-régions commencent avec les échantillons du test de la région mère et se poursuivent éventuellement par le tirage de nouveaux individus. Le recyclage entraîne fatalement un biais qui se trouve n'influer que sur β_r, α_r restant en effet nul comme dans le test exact sans recyclage.

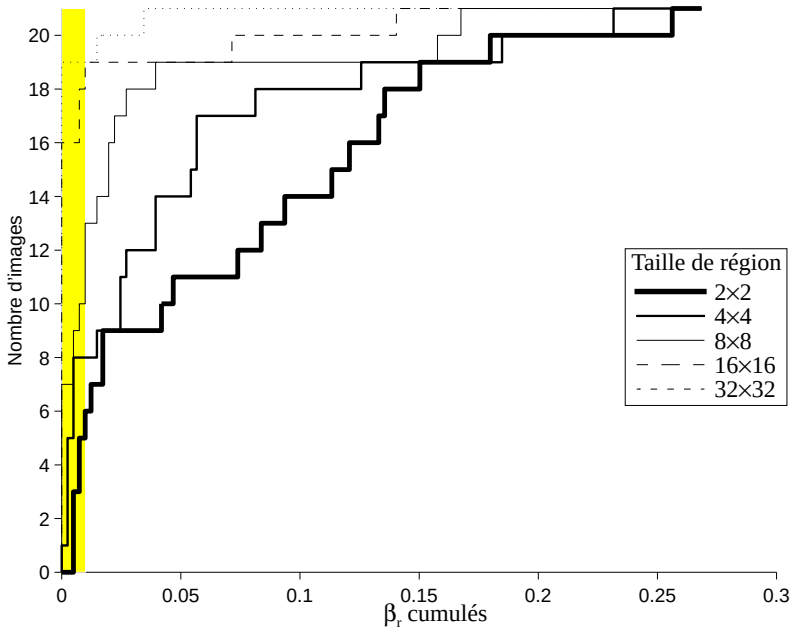


Figure 76: Fréquence cumulée de β_r pour le recyclage sur l'ensemble du panel en fonction de la taille des régions. α = 0,03, β = 0,01 (en grisé), p_h = 0,8

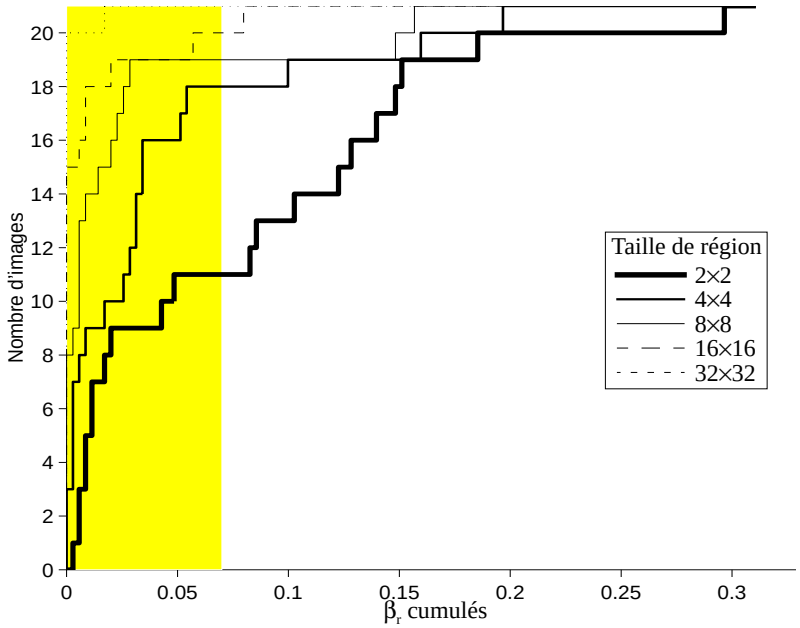


Figure 77: Fréquence cumulée de β_r sur l'ensemble du panel en fonction de la taille des régions. α = 0,07, β = 0,07 (en grisé), p_h = 0,9

Globalement, on constate que le biais est important, surtout pour les petites régions, et assez indépendant de α, β et que l'augmentation de p_h tend à faire augmenter β_r. La mauvaise détection de l'inhomogénéité est renforcée par le découpage récursif des régions. Une mauvaise décision est certainement due à une sur-représentation des échantillons extérieurs à la moyenne : comme ces échantillons sont réutilisés et que la moyenne ne varie guère avec le découpage du fait de l'homogénéité, l'effet de ce «mauvais» échantillonnage se retrouve et s'amplifie dans les régions filles. Cette amplification est d'autant plus forte que p_h est élevée puisque le nombre de points nécessaires à la décision inhomogène décroît lorsque p_h augmente.

IV.4.5.2.3 Test exact et moyenne approchée

Jusqu'à présent nous avons supposé la moyenne connue. Comme nous l'avons déjà signalé, ce n'est pas réaliste et, d'une manière ou d'une autre, il est indispensable de l'approximer. C'est ce que nous faisons en l'évaluant sur un certain nombre de points puis en réalisant le test exact sur d'autres points, l'échantillonnage étant toujours hypergéométrique dans les deux cas et sans recyclage d'une région à l'autre. Comment fixer le nombre de points impliqués dans l'évaluation de la moyenne? C'est une question «à vis (vice?) sans fin» pourrait-on dire. Pour faire un bon test, il faut une bonne moyenne, pour une bonne moyenne, il faut beaucoup de points et donc, globalement, un mauvais test (trop lent). Une mauvaise moyenne permet un bon test (rapide) mais mauvais quant à l'exactitude. Bref, il est difficile de gagner sur tous les tableaux et le résultat ne peut qu'être un compromis difficile et contraignant entre précision et rapidité.

Le nombre de points retenus pour le calcul de la moyenne conditionne la rapidité globale du test : il faut au minimum échantillonner ce nombre de points pour réaliser un test. D'autre part, ce nombre doit varier en fonction de l'effectif de la région à tester : représenter une région de taille 32×32 par 50 % de son effectif est une folie coûteuse et inutile alors que c'est tout juste suffisant pour un effectif total de 4 points. Un intervalle de confiance sur une distribution hypergéométrique est une bonne indication du nombre cherché en fonction de l'effectif. Une indication seulement, sans quoi le test séquentiel devient inutile : connaître la moyenne au seuil de 1 % revient à un test classique d'intervalle de confiance. Ce problème de représentativité sur de petites populations finies a été abondamment traité dans la littérature statistique, nous avons retenu l'approche de T. Wright [Wrig 91].

Si l'on note K, le nombre d'individus ayant une caractèristique particulière (homogène ou inhomogène ici) dans l'ensemble d'une population d'effectif N et que k représente les individus possédant ce caractère sur un tirage sans remise de n_m individus, il est possible de calculer l'intervalle maximal, $[K_i, K_s]$, encadrant la vraie valeur de K avec une probabilité γ donnée. À partir de là, nous avons cherché le nombre d'échantillons tel que la taille de cet intervalle soit suffisamment petite pour qu'elle n'englobe pas plus qu'une proportion μ de la population. Formellement, pour γ, μ, N que l'on se donne, on cherche n_m tel que :

$$\mathrm{Max}\left(\frac{K_s(\gamma, n_m) - K_i(\gamma, n_m)}{N}\right) \leq \mu$$

Le calcul avec $\gamma = 0{,}05$ mène aux valeurs de n_m suivantes :

Taille (N)	μ		
	0,05	0,1	0,2
2×2	4	4	4
4×4	16	15	13
8×8	61	53	36
16×16	212	139	59
32×32	71		

Tableau 6: Nombre d'échantillons (n_m) nécessaire à une «précision» donnée pour un tirage hypergéométrique

Comme on pouvait s'y attendre, les exigences de l'intervalle de confiance, uniquement basées sur les propriétés de l'hypergéométrique, sont très élevées, souvent trop pour l'usage que nous souhaitons en faire. Compte tenu de ce que nous avons dit plus haut le compromis suivant a été retenu :

Taille	2×2	4×4	8×8	16×16	≥ 32×32
n_m	3	9	36	59	71

Tableau 7: Nombre d'échantillons retenus, dans les tests, pour le calcul de la moyenne approchée

Les tests sur le panel ont donc été menés en évaluant la moyenne sur n_m points en fonction de la taille de la région testée puis en réalisant le test séquentiel sur d'autres points. Les résultats suivants, extraits des tests, montrent que l'approximation de la moyenne laisse β_r à peu près stable alors que α_r est très perturbé.

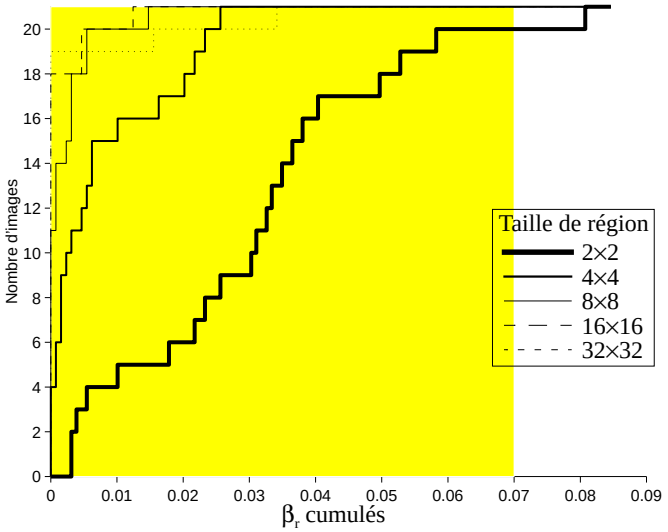


Figure 78: Fréquence cumulée de β_r pour la moyenne approchée sur l'ensemble du panel en fonction de la taille des régions. α = 0,07, β = 0,07 (en grisé), p_h = 0,8

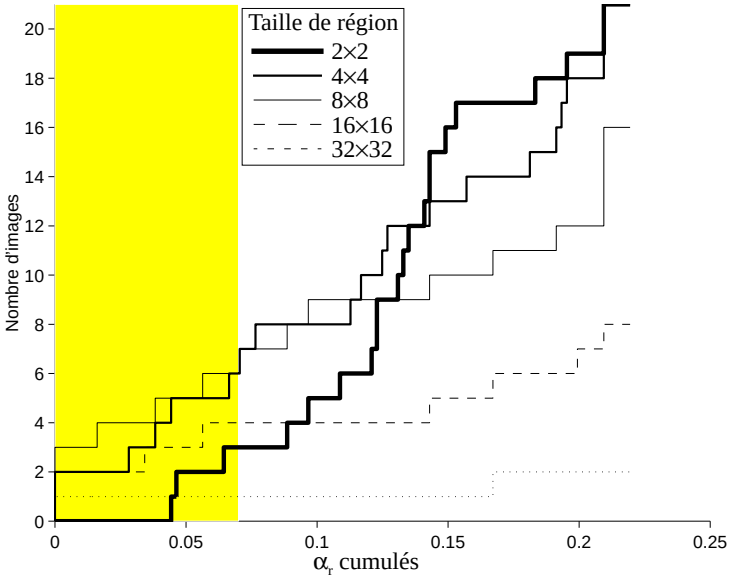


Figure 79: Fréquence cumulée de α_r pour la moyenne approchée sur l'ensemble du panel en fonction de la taille des régions. α = 0,07 (en grisé), β = 0,07, p_h = 0,8

La moyenne approchée est en fait représentative pour des régions homogènes : les points étant groupés, peu d'entre eux suffisent à une bonne approximation, prendre une région homogène pour une inhomogène est donc assez peu fréquent. Par contre, s'il y a inhomogénéité, la représentativité est très mauvaise, d'où une décision homogène facilement fausse lorsqu'il y a inhomogénéité. Le biais, ici aussi, est assez indépendant des valeurs de α annoncées et plus sensible sur les petites que sur les grandes régions ce qui est compréhensible compte tenu du petit nombre de points utilisés pour la moyenne sur celles-là. La valeur de β, quant à elle, influe bien sur le risque réel sauf pour les régions de taille 2×2.

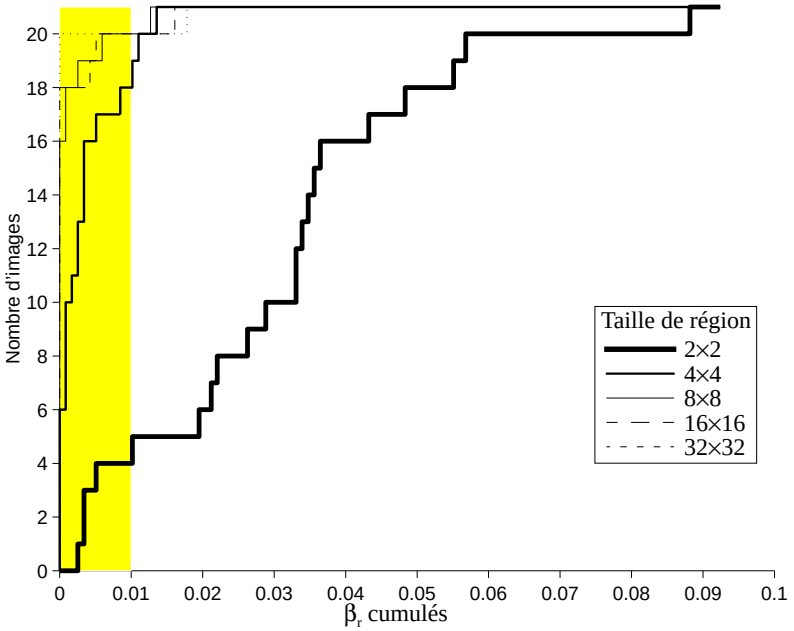


Figure 80: Fréquence cumulée de β_r pour la moyenne approchée sur l'ensemble du panel en fonction de la taille des régions. α = 0,03, β = 0,01 (en grisé), p_h = 0,9

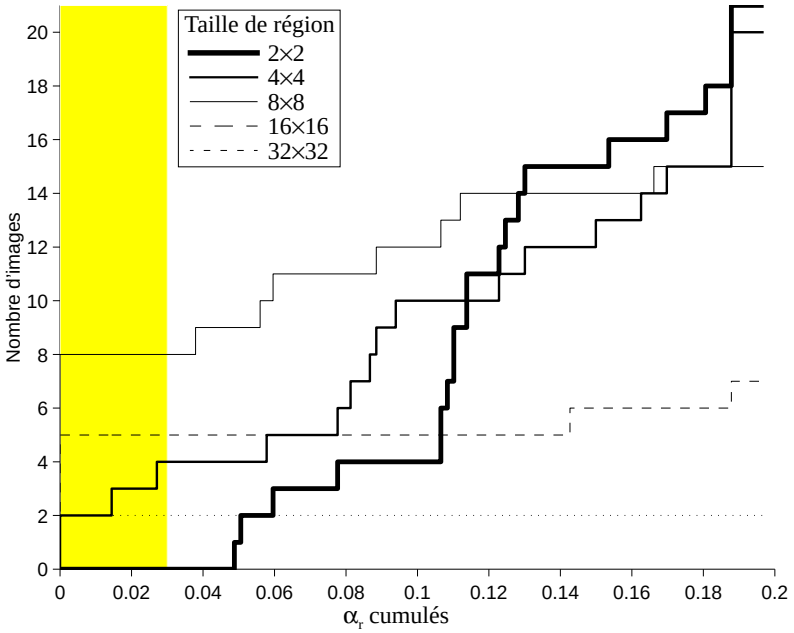


Figure 81: Fréquence cumulée de α_r pour la moyenne approchée sur l'ensemble du panel en fonction de la taille des régions. α = 0,03 (en grisé), β = 0,01, p_h = 0,9

IV.4.5.2.4 Composition recyclage/moyenne approchée

Après avoir testé séparément le recyclage et l'approximation de la moyenne, le test suivant consiste à mélanger les deux. La moyenne est évaluée sur n_m points comme précédemment. Le test séquentiel démarre en réutilisant ces points et continue par le tirage de nouveaux échantillons si les premiers sont insuffisants pour prendre une décision. De même, les régions filles réutilisent les points de la mère pour le calcul de la moyenne et éventuellement pour leur test séquentiel. De nouveaux tirages sont réalisés pour la moyenne ou les tests séquentiels si nécessaire.

Contrairement à ce que l'on pourrait croire a priori, le cumul des deux biais n'augmente pas l'incertitude mais, au contraire, diminue l'erreur commise, sur β_r davantage que sur α_r, même si elle reste relativement élevée. L'augmentation de p_h a tendance à diminuer l'erreur, tant sur β que sur α.

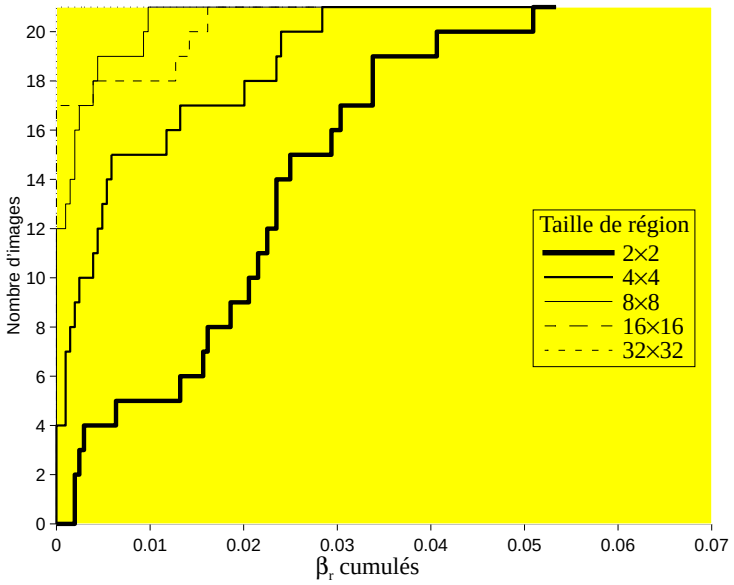


Figure 82: Fréquence cumulée de β_r pour recyclage et moyenne approchée sur l'ensemble du panel en fonction de la taille des régions. α = 0,07, β = 0,07 (en grisé), p_h = 0,8

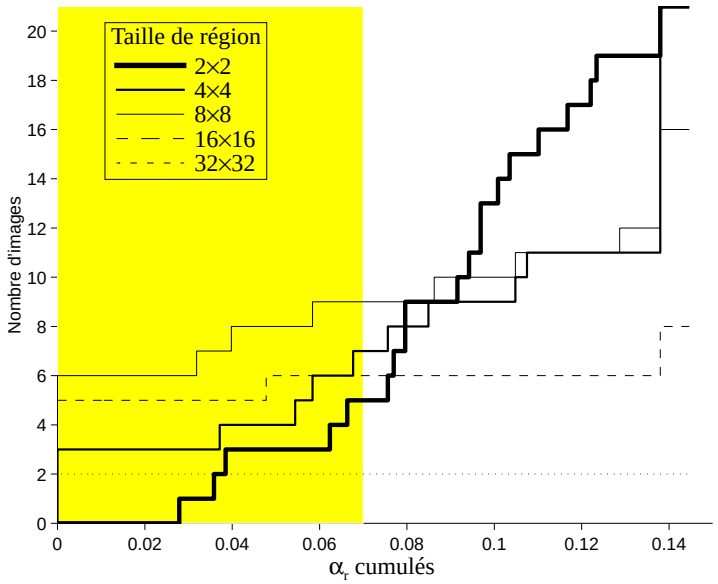


Figure 83: Fréquence cumulée de α_r pour recyclage et moyenne approchée sur l'ensemble du panel en fonction de la taille des régions. α = 0,07 (en grisé), β = 0,07, p_h = 0,8

La diminution de l'erreur s'explique en partie par la complémentarité de l'influence de chacun des biais : si le recyclage perturbe fortement β_r, l'approximation de la moyenne agit surtout sur α_r. Globalement néanmoins, le comportement de la moyenne prédomine sur celui du recyclage. En effet lorsque l'homogénéité est réelle, la moyenne a toutes les chances d'être bonne et par conséquent, les échantillons, représentatifs de la région. Déclarer l'inhomogénéité à tort dans ce cas, signifie une mauvaise moyenne et une certaine proportion de points à l'extérieur de celle-ci. Lors du découpage des sous-régions, toujours réellement homogènes, on peut penser que la répartition des points, à cause de la cohérence des images, va se faire en séparant les points très extérieurs à la moyenne des points intérieurs. Le calcul de la moyenne, en partie sur ces points, va probablement amener à un repositionnement de celle-ci plus au centre de chacun de ces groupes de points pour aboutir à une décision homogène. Il n'y aurait pas, ici, d'amplification de la décision inhomogène comme avec le recyclage seul.

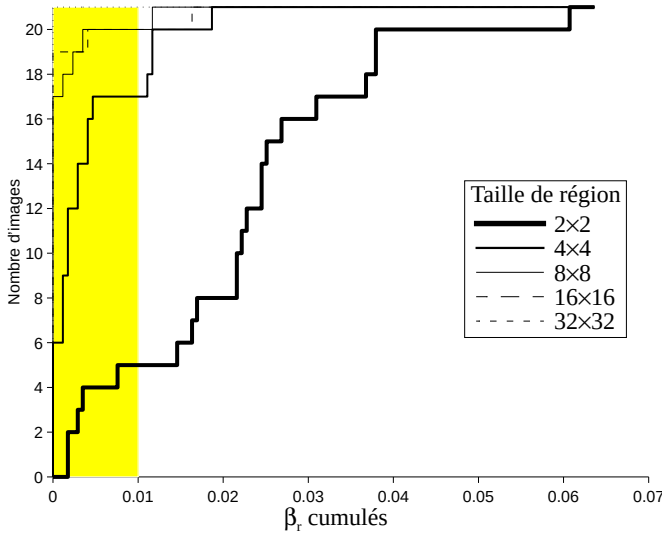


Figure 84: Fréquence cumulée de β_r pour recyclage et moyenne approchée sur l'ensemble du panel en fonction de la taille des régions. α = 0,03, β = 0,01 (en grisé), p_h = 0,9

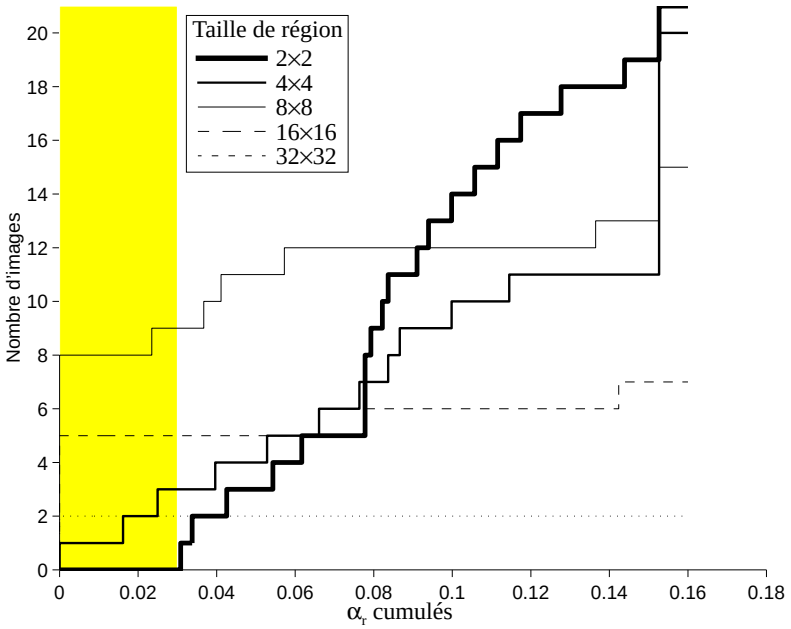


Figure 85: Fréquence cumulée de α_r pour recyclage et moyenne approchée sur l'ensemble du panel en fonction de la taille des régions. α = 0,03 (en grisé), β = 0,01, p_h = 0,9

IV.5 Tracé de rayons progressif

IV.5.1 Test inexact

IV.5.1.1 Proportion inhomogène comme paramètre

La lenteur du test exact et particulièrement de la décision homogène n'est pas envisageable dans un tracé de rayons. Même si l'échantillonnage et le test ne demandent pas beaucoup de temps de calcul, comme la proportion de points nécessaire à une image est proche de p_h, tout le bénéfice du test est perdu (rappelons qu'il est au moins aussi rapide que n'importe quel autre test, séquentiel ou non, et en général, plus rapide). Cette lenteur est due à la proximité des deux probabilités de base. En revanche, la sortie de la région critique est d'autant plus rapide que $p_{\bar{h}}$ et p_h s'éloignent l'une de l'autre. Pour s'en convaincre, il suffit de considérer les premiers instants de sorties possibles comme nous l'avons fait prédédemment (Cf §IV.4.5.1.2). À la différence du test exact, la définition de $p_{\bar{h}}$ n'est plus automatiquement dictée par p_h comme dans l'équation (45) §IV.4.5, au contraire, nous l'utiliserons comme un paramètre supplémentaire (sans oublier tout de même $p_{\bar{h}} < p_h$). La définition du test reste identique mais utilise D_h et $D_{\bar{h}}$ avec la nouvelle signification suivante :

$$\boxed{D_h = \lfloor n_R p_h \rfloor \qquad D_{\bar{h}} = \lfloor n_R p_{\bar{h}} \rfloor}$$

Le premier instant de sortie homogène, k_h, se produit alors lorsque :

$$\frac{D_h \ldots (D_h - (k_h - 1))}{D_{\bar{h}} \ldots (D_{\bar{h}} - (k_h - 1))} \geq \frac{1-\beta}{\alpha}$$

De même, le premier instant de sortie inhomogène, $k_{\bar{h}}$, a lieu quand :

$$\frac{(n_R - D_h) \ldots (n_R - D_h - (k_{\bar{h}} - 1))}{(n_R - D_{\bar{h}}) \ldots (n_R - D_{\bar{h}} - (k_{\bar{h}} - 1))} \leq \frac{\beta}{1-\alpha}$$

Les risques α et β fixés, la rapidité du test peut être estimée par les rapports k_h / D_h et $k_{\bar{h}} / D_{\bar{h}}$ comme précédemment (Cf §IV.4.5.1.2). Le tableau suivant donne ces rapports pour $\alpha = 0{,}02$ et $\beta = 0{,}02$:

	Taille				
	8×8	16×16	32×32	64×64	128×128
$p_h = 0{,}95$, $p_{\bar{h}} = 0{,}94$		72 % 4,58 %	32 % 1,97 %	9 % 0,55 %	2,34 % 0,14 %
$p_h = 0{,}95$, $p_{\bar{h}} = 0{,}9$	72 % 8,77 %	25 % 2,17 %	7,2 % 0,65 %	1,85 % 0,16 %	0,46 % 0,04 %
$p_h = 0{,}9$, $p_{\bar{h}} = 0{,}89$	98 % 14 %	73 % 9,25 %	32 % 3,84 %	9 % 1,07 %	2,33 % 0,28 %
$p_h = 0{,}9$, $p_{\bar{h}} = 0{,}85$	72 % 13 %	25 % 4,15 %	7,16 % 1,15 %	1,85 % 0,29 %	0,46 % 0,07 %

Tableau 8: Proportion de points à échantillonner avant la première sortie possible homogène (cadré à gauche) et inhomogène (cadré à droite) lorsque $p_{\bar{h}}$ varie.

On pourrait croire que nous venons de trouver la solution miracle puisque de très faibles écarts entre entre p_h et $p_{\bar{h}}$ conduisent à des gains de rapidité très appréciables. Hélas, hélas, chassez la théorie, elle revient au galop!… L'éloignement de ces deux probabilités, en effet, ne crée pas seulement une nouvelle hypothèse mais modifie également le sens de $\overline{H}$ et, plus grave peut-être, conduit à l'abandon de deux résultats de première importance de l'analyse séquentielle.

IV.5.1.2 L'intervalle incertain

Alors que la situation était confortablement bipolaire avec le TSRV initial, une troisième hypothèse s'immisce maintenant entre les deux premières puisque la proportion de points inconnue, p, peut désormais appartenir à l'intervalle $[p_{\bar{h}}, p_h]$, intervalle que nous qualifions d'*incertain*. De ce fait, $\overline{H}$ ne représente plus l'hypothèse complémentaire de H mais seulement une partie de cet espace complémentaire. L'hypothèse inhomogène change donc de sémantique et représente désormais la situation $p \leq p_{\bar{h}}$, contre $p < p_h$ dans le test exact, nous la noterons $\overline{H}$' pour marquer la différence. Le premier résultat théorique à tomber de ce fait, concerne la probabilité de terminaison du test qui ne peut plus être assurée. Que p vérifie $\overline{H}$' ($p \leq p_{\bar{h}}$) ou H ($p \geq p_h$) et le test se terminera à coup sûr. À l'inverse, si la troisième alternative est vraie, le test se terminera au pire lorsque l'intégralité de la région aura été échantillonnée et non par une sortie normale de la région critique. Si un tirage avec remise était utilisé au lieu d'un tirage sans remise, le test pourrait se poursuivre indéfiniment.

Le second résultat théorique à ne plus être vérifié concerne la validité du risque α : alors que ce risque est au pire atteint dans le test exact classique, lorsque $p_{\bar{h}}$ décroît, il n'y a plus de certitude. En effet, le nouveau risque α' dont on garde le contrôle, est celui de déclarer $p \geq p_h$ alors qu'en réalité $p \leq p_{\bar{h}}$. En dissociant p_h et $p_{\bar{h}}$, la seule chose qui puisse être dite désormais, est que α, le risque qui nous intéresse vraiment, est plus grand que α'. En d'autres termes, l'erreur commise lorsqu'une région est détectée homogène nous échappe. Le risque β' quant à lui, est en fait plus sévère que β car il cor-

respond à la probabilité de décréter $p \leq p_{\bar{h}}$ alors que $p \geq p_h$ est vraie. Il est donc, en quelque sorte, contenu dans β. Le contrôle de la décision inhomogène reste donc assuré et nous avons déjà souligné qu'il nous intéresse plus que celui de α.

La distance entre probabilité homogène et inhomogène n'est heureusement pas rédhibitoire. Pas totalement. Si l'intégralité de la population se situait hors de l'intervalle incertain, le test se comporterait parfaitement, c'est-à-dire avec sûreté et rapidité. Non seulement par absence de l'intervalle incertain mais également, selon toute vraisemblance, par éloignement du vrai paramètre par rapport à $p_{\bar{h}}$ et p_h : des risques doublement nuls. Par contre, s'il existe une proportion de la population $\mu_{\bar{h}h}$ entre $p_{\bar{h}}$ et p_h, le risque α' vaudra au pire $\alpha + \mu_{\bar{h}h}$, et puisqu'il n'y a pas de raison que l'on ait moins de chance que pour un pile ou face quelconque, disons qu'en moyenne α' vaudrait $\alpha + (\mu_{\bar{h}h} / 2)$.

IV.5.1.3 Minimisation des risques

IV.5.1.3.1 Décision express

Plus faible sera la proportion $\mu_{\bar{h}h}$, plus sûr sera le test, éventuellement même, tout à fait certain, mais… certainement, plus il sera lent aussi. Une autre vis sans fin[†]. Fort heureusement, les distributions du paramètre que nous avons mises à jour (Cf §IV.4.3) vont dans le bon sens : la population se répartit plus aux extrêmes qu'au centre, en général plus de 70 % de l'effectif se concentre effectivement sur les bords. En fixant p_h, nous avons donc cherché à déterminer $p_{\bar{h}}$ de telle sorte que $\mu_{\bar{h}h}$ soit inférieure à 5, 10, 15 ou 20 % pour chaque taille de région sur l'ensemble du panel. Autrement dit, nous avons déterminé les valeurs minimum de $p_{\bar{h}}$ pour lesquelles la taille de l'effectif de l'intervalle incertain est maximal sur le panel. Ce faisant, les situations les plus difficiles sont privilégiées, i.e les distributions les plus concentrées entre $[p_{\bar{h}}, p_h]$. Les valeurs de $p_{\bar{h}}$ ainsi déduites sont reprises dans le tableau suivant :

		$p_{\bar{h}}$		
p_h	$\mu_{\bar{h}h}$	Taille 16×16	Taille 8×8	Taille 4×4
0,95	5 %	0,7539	0,875	0,9375
	10 %	0,625	0,7344	0,875
	15 %	0,5313	0,6563	0,6875
	20 %	0,4668	0,625	0,75
0,9	5 %	0,7148	0,7344	0,875
	10 %	0,582	0,7031	0,8125
	15 %	0,5195	0,6563	0,75
	20 %	0,4531	0,625	0,6875

Tableau 9: Proportions inhomogènes retenues pour le test

† C'est décidé nous allons monter un magasin!

		$p_{\bar{h}}$		
p_h	$\mu_{\bar{h}h}$	Taille 16×16	Taille 8×8	Taille 4×4
0,85	5 %	0,7031	0,7969	0,8125
	10 %	0,582	0,7031	0,75
	15 %	0,5195	0,6563	0,75
	20 %	0,4531	0,6094	0,6875
0,8	5 %	0,6758	0,7188	0,75
	10 %	0,582	0,6875	0,75
	15 %	0,5195	0,6406	0,6875
	20 %	0,4531	0,6094	0,625
0,75	5 %	0,625	0,7344	0,625
	10 %	0,582	0,6719	0,6875
	15 %	0,5195	0,625	0,4375
	20 %	0,4531	0,5781	0,625

Tableau 9: Proportions inhomogènes retenues pour le test

Conformément à nos attentes et à la bonne répartition de p, l'éloignenemt entre $p_{\bar{h}}$ et p_h promet d'accélérer sensiblement le test séquentiel. Même lorsque $\mu_{\bar{h}h}$ est faible, la valeur de $p_{\bar{h}}$ correspondante est en général très éloignée de celle de p_h, ce qui augure de gains de rapidité importants. La question essentielle à laquelle il faut désormais répondre concerne les risques. La vélocité ne nous sert de rien s'ils explosent…

IV.5.1.3.2 Test inexact express

Nous avons donc mené nos tests habituels sur le panel pour évaluer les risques réels engendrés par la décision express. Le test reprend la même méthodologie que le test exact non biaisé (Cf §IV.4.5.2.1) à la différence que $p_{\bar{h}}$ n'est plus déterminé par p_h comme dans (45) §IV.4.5. Les tests ont été menés sur les valeurs de $p_{\bar{h}}$ du tableau précédent pour chaque $\mu_{\bar{h}h}$ avec p_h = 0,8, 0,85 et 0,95, ε = 1 et 2 et pour α' = 0,02, β' = 0,01. Pour les tailles de régions supérieures à 16×16, $p_{\bar{h}}$ vaut simplement $p_h - \mu_{\bar{h}h}$.

Sur l'ensemble de ces résultats, le risque «homogène» réel, α'_r , est presque toujours nul (sauf dans 4 cas sur 4032 représentants 4 mauvaises décisions homogènes sur les 755 805 que comptent les 4032 mesures de α'_r), le risque «inhomogène» réel, β'_r , est nul sauf dans 0,97 % des mesures où, sans l'être, il est toujours inférieur à β' (cela représente une cinquantaine de mauvaises décisions inhomogènes sur plus de 9,5 millions réalisées sur l'ensemble des mesures).

La décision express, bien que théoriquement biaisée, est pleinement justifiée par ces tests et se montre, ici, tout-à-fait fiable en pratique.

IV.5.1.4 Test inexact express avec recyclage et moyenne approchée

Reste à tester la réunion de toutes les techniques que nous avons exposées dans les paragraphes précédents. Rappelons l'ensemble. Pour chaque région, la moyenne est évaluée sur n_m points en fonction de sa taille, le test express est ensuite réalisé sur ces points et se poursuit éventuellement par des tirages supplémentaires si besoin. Les tests sur les sous-régions réutilisent les points échantillonnés dans la région mère, plus si nécessaire.

Les tests ont été menés avec p_h successivement égale à 0,8, 0,85 et 0,95, $\alpha' = 0{,}02$, $\beta' = 0{,}01$, $\mu_{\bar{h}h}$ égale à 0,05, 0,1, 0,15 ou 0,2 et $\varepsilon = 1$ ou 2. Les nombres de points retenus pour le calcul de la moyenne ont été augmentés vers la limite que nous préconisons comme maximale pour que les performances restent intéressantes :

Taille	2×2	4×4	8×8	16×16	≥ 32×32
n_m	3	12	39	66	71

Tableau 10: Nombre d'échantillons retenus pour le calcul de la moyenne approchée

On constate que β'_r est nul quels que soient les paramètres pour les régions de taille supérieure à 64×64. Pour les régions plus petites, β'_r est inférieur à 1 % dans la majorité des cas sauf pour les régions de taille 2×2 où il est assez souvent supérieur et toujours inférieur à 6 %. On constate globalement une tendance à la diminution lorsque p_h augmente. En l'occurence lorsque $p_h = 0{,}95$, $\varepsilon = 2$, $\mu_{\bar{h}h} = 0{,}05$, le risque réel est inférieur à 2 % (sauf pour les régions de taille 2×2).

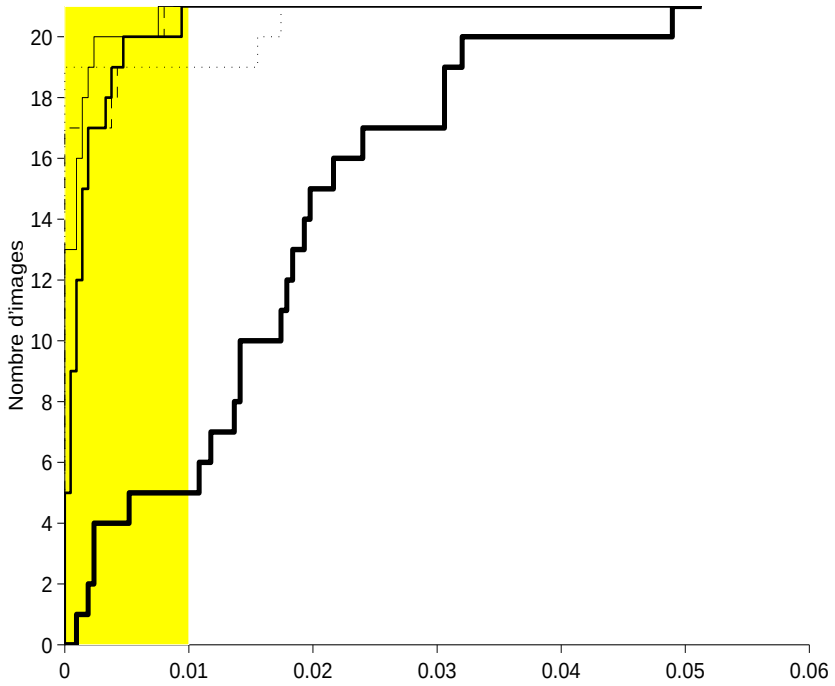


Figure 86: Fréquence cumulée de β'_r pour le test express sur l'ensemble du panel en fonction de la taille des régions. $\varepsilon = 2$, $\alpha' = 0{,}02$, $\beta' = 0{,}01$ (en grisé), $p_h = 0{,}95$, $\mu_{\bar{h}h} = 0{,}05$

Le risque réel de décision homogène lorsqu'il y a inhomogénéité, α'_r, est nul pour toutes les images dans 5 % des mesures pour des régions de taille 32×32. Globalement, il est inférieur au seuil fixé ($\alpha' = 0{,}02$) dans plus de 60 % des mesures. Ce pourcentage augmente au fur et à mesure que l'effectif des régions décroît. Le biais maximal pour ce risque apparaît lorsque ε est petit : la valeur maximum de α'_r pour $\varepsilon = 1$ est en effet de 22 % alors qu'elle ne dépasse pas 7 % pour $\varepsilon = 2$. Enfin, la proportion de points utilisés pour l'obtention des images finales varie entre 70 et 89 %, des performances intéressantes lorsque l'on sait la sur-représentation des images inhomogènes dans notre panel.

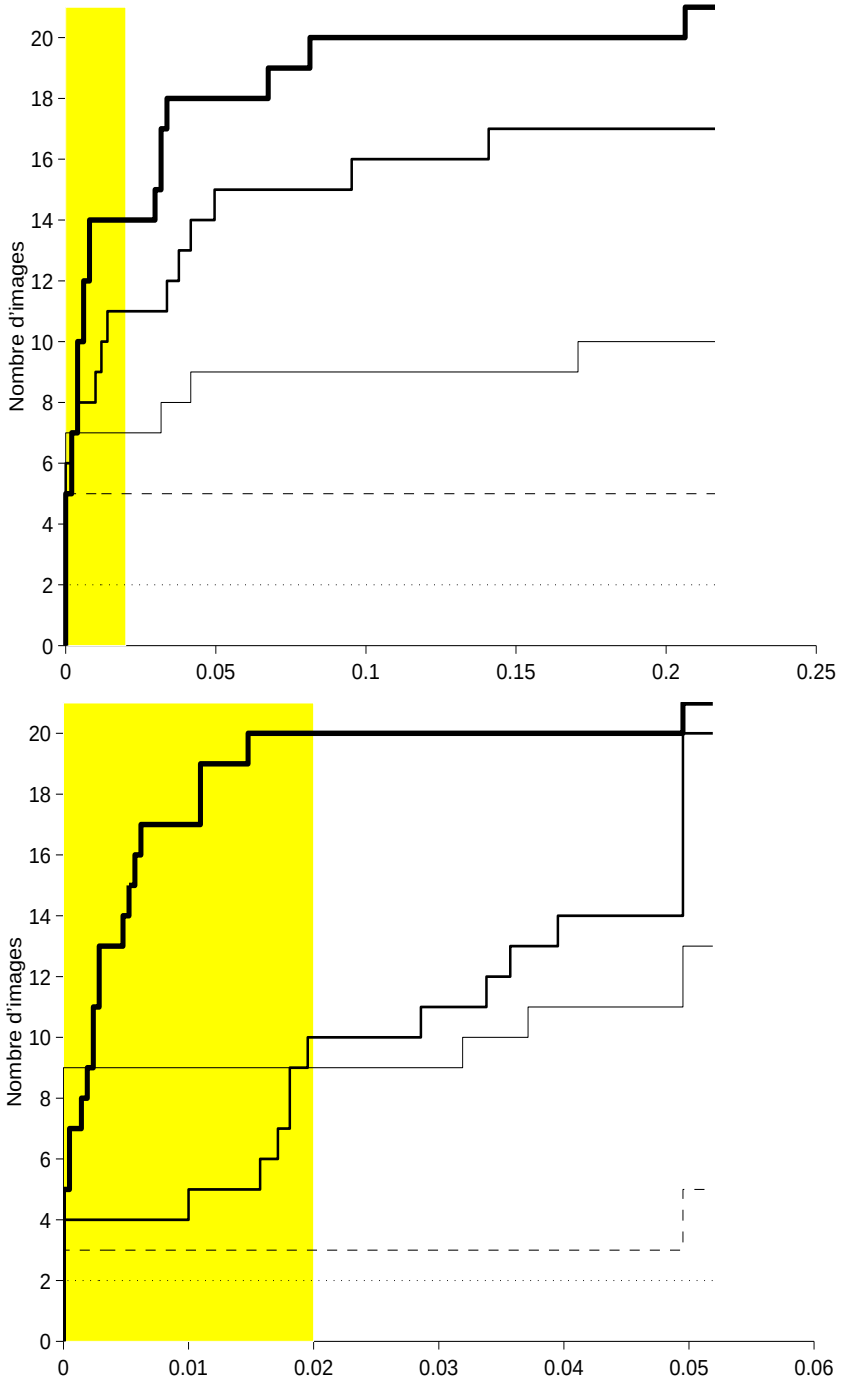


Figure 87: Fréquence cumulée de α'_r pour le test express sur l'ensemble du panel en fonction de la taille des régions. $\alpha'=0{,}02$ (en grisé), $\beta'=0{,}01$, $p_h=0{,}95$, $\mu_{hh}=0{,}05$. $\varepsilon=1$ (en haut) $\varepsilon=2$ (en bas). Noter la différence d'échelle

La distance colorimètrique entre les images résultantes et les références est le dernier élément qui permette de juger de la méthode. L'erreur perceptuelle est naturellement moindre lorsque ε est petit, p_h grand et $\mu_{\bar{h}h}$ faible. Pour ε = 1 et p_h = 0,95, la distance L*u*v* moyenne sur l'ensemble des images est de 0,09 (±0,1) et l'écart-type moyen de 0,73 (±0,99). Pour ε = 2 et avec la même probabilité homogène, la distance moyenne s'élève à 0,22 (±0,16) et l'écart-type moyen à 1,06 (±1,03).

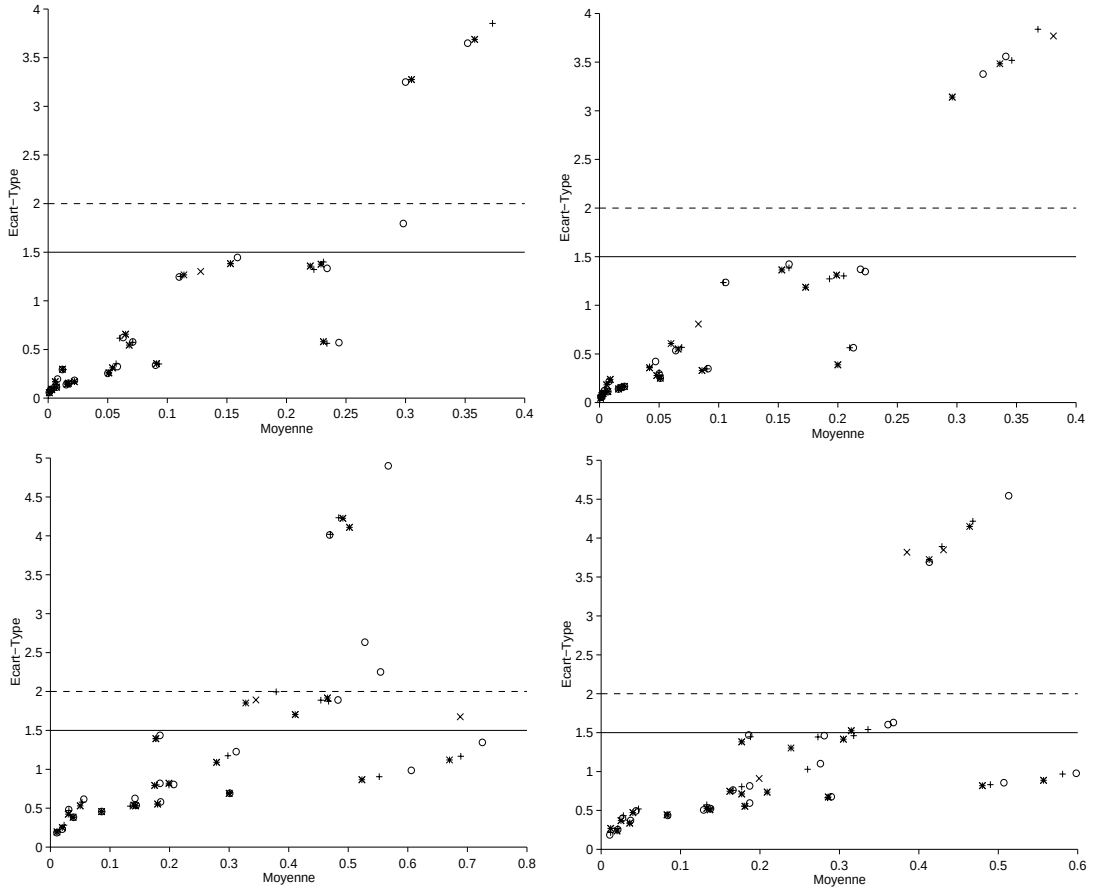


Figure 88: Distances colorimétriques pour les images du panel calculées par le test express.
ε = 1 (en haut), ε = 2 (en bas), p_h = 0,8 (à gauche), p_h = 0,95 (à droite),
$\mu_{\bar{h}h}$: ∘ : 0,2 + : 0,15 ∗ : 0,1 × : 0,05

En conclusion, le comportement moyen de cette méthode est tout à fait acceptable du point de vue perceptuel en produisant la plupart du temps des différences colorimètriques imperceptibles à l'oeil. La distance moyenne est toujours bien située (inférieure à 0,7 et, pour de bons paramètres, à 0,4), l'écart-type l'est également dans la grande majorité des cas (inférieur à 1,5) mais peut être parfois préoccupant (supérieur à 3) dans un petit nombre d'images. Au sujet des risques réels, s'ils sont difficilement contrôlables du fait des biais contradictoires, on peut dire qu'ils sont suffisamment faibles globalement (très souvent inférieurs à 5 %) pour rendre la méthode fiable. L'infériorité souhaitée de β'_r sur α'_r est effective (Cf §IV.4.5.1) et très appréciable : les efforts de calcul sont presque toujours réalisés à bon escient, i.e la détection de l'inhomogénéité est fiable. Une seule exception à cette constatation porte sur les régions de taille 2×2 où une mauvaise décision inhomogène peut être prise jusqu'à une fois sur trois. Guère mieux qu'un pile ou face, cette incertitude pourrait être intolérable pour des régions d'effectif supérieur, elle est ici bénigne car elle conduit simplement à une subdivision erronée. Ce supplément de calcul implique un surcoût quant à la durée du test mais ne conduit à

aucune erreur visuelle, comme c'est d'ailleurs toujours le cas lorsque β'_r n'est pas nul. Le risque «homogène» réel, α'_r, est par contre directement impliqué dans l'apparition de ces erreurs puisque seule une mauvaise décision homogène peut conduire à une «sous-représentation» des régions et donc à une mauvaise interpolation faute de points suffisamment représentatifs. Pour de «bons» paramètres (essentiellement $\varepsilon = 2$), α'_r peut être rendu faible (inférieur à 5 %) mais ces paramètres ne sont bons que pour la diminution de α'_r, pas pour celle de l'erreur perceptuelle puisque, pour elle, il est souhaitable que ε soit aussi petit que possible. Un autre effet ennuyeux sur α'_r est sa propension à être proportionnellement plus important sur les grandes régions que sur les petites résultant en erreurs visuelles plus préjudiciables.

Le paramètrage retenu jusqu'à présent pour chaque biais semble adapté à leur mélange mais il paraît difficile de le pousser davantage sans dégrader sérieusement la rapidité du test. Il est donc nécessaire de faire appel à des techniques extérieures au test et à la panoplie déjà développée pour diminuer le seul aspect réellement préjudiciable : l'insuffisance de la détection homogène.

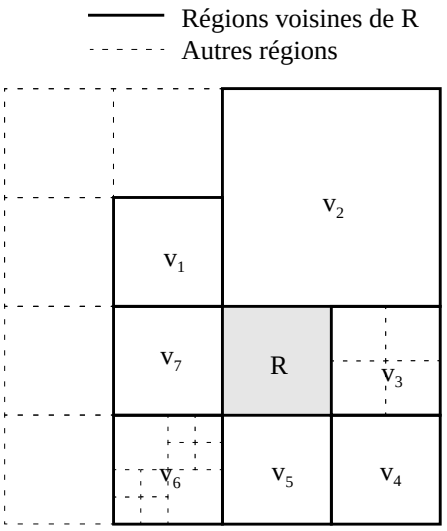
IV.5.2 Cohérence spatiale

Les informations générées par un tracé de rayon dépassent de loin la simple couleur du pixel. Les objets vus, la présence d'ombres ou de reflets, etc, constituent une grande richesse d'informations transportées par chaque rayon. La notion d'homogénéité que nous avons élaborée est un raccourci concomittant de certaines conjonctions de ces informations. Il y a en effet de fortes chances par exemple, pour qu'une région comportant plusieurs objets, ne puisse être homogène, de même si elle recouvre des ombres et des portions éclairées ou encore si elle contient des objets texturés. La détection homogène peut donc être fiabilisée par l'utilisation conjointe de ces informations. Ainsi l'homogénéité ne sera décrétée que s'il y a également *cohérence spatiale*. Dans notre implémentation, nous n'avons utilisé qu'un seul type de cohérence liée aux objets : une région n'est homogène que si elle comprend un seul objet. D'autres extensions sont possibles pour, par exemple, qu'une région soit homogène si elle ne contient qu'un objet éclairé ou un objet à l'ombre, etc.

Chaque point échantillonné contient donc l'identificateur de l'objet le plus proche de l'oeil. L'utilisateur peut avoir la charge de définir ces identificateurs au moment de la modélisation. Nous avons préféré automatiser cette tâche lors de la construction de notre scène CSG (constructive solid geometry) Un objet est une feuille de l'arbre CSG (une primitive géométrique) dont l'identificateur est un simple numéro donné lors du parcours de l'arbre. Cette notion d'objet pourrait être plus fine et inclure les informations de matière et d'opération booléenne pour qu'un objet puisse être un ensemble d'opérations booléennes ayant des caractéristiques communes. Nous nous sommes contenté de la numérotation des primitives.

Cette cohérence spatiale est liée à l'échantillonnage des régions. Lorsque l'aire projetée d'un objet sur l'espace image occupe une portion trop faible d'une région, il y a toutes les chances que cet objet ne soit pas échantillonné. Dans ces conditions, la vérification de l'homogénéité est insuffisante. Pour tenir compte du phénomène, nous avons étendu la cohérence spatiale aux voisines d'une région. Lorsque le test express détecte l'homogénéité d'une région et qu'elle ne contient qu'un seul objet, une ultime vérification est réalisée. Elle consiste à s'assurer qu'aucun objet n'a la possibilité de traverser plusieurs régions sans être échantillonné : toutes les régions voisines de celle tes-

tée sont parcourues et si un objet est commun à deux régions *opposées* et n'apparaît pas dans la région centrale, cette dernière est déclarée inhomogène. Deux régions voisines opposées par rapport à R sont telles qu'il existe au moins un segment coupant R dont les extrémités appartiennent à chacune de ces régions. Le voisinage retenu dans cette procédure est un voisinage V8 imparfait (à 8 voisins au plus) étendu à un arbre quaternaire. La voisine d'une région est une région d'aire supérieure ou égale ayant au moins un point de frontière commun avec elle :



Région	Opposées par rapport à R
v_1	v_3, v_4, v_5
v_2	v_3, v_4, v_5, v_6, v_7
v_3	v_1, v_2, v_5, v_6, v_7
v_4	v_1, v_2, v_7
v_5	v_1, v_2, v_3, v_7
v_6	v_2, v_3
v_7	v_2, v_3, v_4, v_5

Figure 89: Régions voisines et opposées par rapport à R

Cette vérification détecte bien les objets dont la projection est convexe sur l'écran mais peut conduire parfois à des découpages abusifs selon la forme projetée (i.e concave) des objets qui, bien qu'appartenant à deux régions voisines opposées, ne coupent pas la région centrale pour autant. Cette situation est cependant assez rare du fait des tailles généralement petites et proches des régions et de leurs voisines lors de cette vérification.

IV.5.3 Résultats

Le test express à cohérence spatiale a été réalisé sur cinq scènes dont les caractéristiques sont données ci-dessous :

Scène	Modèle d'éclairement	Primitives	Opérations booléennes	Description
The famous boules	Whitted (4 rebonds max.)	4	1	3 boules sur un plan
Oiseau	Whitted (2 rebonds max.)	20	24	un oiseau, 2 miroirs
Poissons	Whitted (2 rebonds max.)	55	51	5 poissons
Verre	Whitted (10 rebonds max.)	12	13	un verre, de l'eau, 2 échiquiers
Sphères	~ Cook/Blinn (4 rebonds max.)	42	1	41 sphères sur un plan

Tableau 11: Scènes de test

Les images de référence ont été calculées sur 512×512 pixels avec 1 ou 4 rayons par pixel (r/p ci-dessous) avec un tracé de rayons classique pour le modèle de Whitted et un tracé de rayons spectral pour le modèle de Cook/Blinn. Les paramètres utilisés pour notre tracé de rayons progressif dans ces tests sont les suivants :

$\varepsilon = 2$; $p_h = 0{,}9$; $\mu_{\bar{h}h} = 0{,}05$; $\alpha' = 0{,}02$; $\beta' = 0{,}01$

La moyenne étant calculée sur le nombre de points suivants :

Taille	2×2	4×4	8×8	16×16	≥ 32×32
n_m	3	12	39	66	71

Tableau 12: Nombre d'échantillons retenus pour le calcul de la moyenne approchée

Scène	r/p	Temps utilisateur				Pixels échantillonnés		Rayons générés		L*u*v*			Mémoire (en Mo)	
		Total	%	Test seul	%	Total	%	Total	%	$\overline{\Delta}_{uv}$	σ_{uv}	% points ≠		
The famous boules	1	2m43s52c				262 144		870 643						*référence*
		1m22s87c	*50,6*	5s90c	*7,1*	80 467	*30,7*	399 775	*45,9*	0,154	4,183	*2,1*	*6,6*	*sans cohérence*
		1m27s51c	*53,5*	8s14c	*9,3*	88 951	*33,9*	423 793	*48,7*	0,074	2,121	*2,05*	*7,2*	*avec cohérence*
	4	10m17s12c				1 048 576		3 482 656						*référence*
		4m04s28c	*39,6*	16s74c	*6,8*	220 748	*21*	1 167 428	*33,5*	0,108	1,846	*3,74*	*16,3*	*sans cohérence*
		4m44s87c	*46,2*	25s20c	*8,8*	243 223	*23,2*	1 250 508	*35,9*	0,091	1,71	*3,63*	*17,9*	*avec cohérence*
Oiseau	1	3m35s64c				262 144		1 147 028						*référence*
		1m47s24c	*49,7*	6s99c	*6,5*	85 512	*32,6*	415 968	*36,3*	0,258	4,304	*2,47*	*7*	*sans cohérence*
		1m56s47c	*54*	8s90c	*7,6*	92 090	*35,1*	448 414	*39,1*	0,222	3,837	*2,44*	*7,5*	*avec cohérence*
	4	14m36s95c				1 048 576		4 588 095						*référence*
		5m53s36c	*40,3*	22s36c	*6,3*	264 316	*25,2*	1 287 905	*28,1*	0,435	5,183	*4,2*	*19,2*	*sans cohérence*
		6m23s22c	*43,7*	28s27c	*7,4*	285 957	*27,3*	1 392 952	*30,4*	0,329	4,878	*3,89*	*20,8*	*avec cohérence*
Poissons	1	2m07s36c				262 144		412 888						*référence*
		1m21s47c	*64*	6s01c	*7,4*	87 083	*33,2*	203 616	*49,3*	0,107	3,286	*2,68*	*7,1*	*sans cohérence*
		1m28s56c	*69,5*	9s04c	*10,2*	102 197	*39*	223 921	*54,2*	0,057	2,494	*2,25*	*8,2*	*avec cohérence*
	4	8m36s68c				1 048 576		1 651 524						*référence*
		4m05s90c	*47,6*	20s26c	*8,2*	239 939	*22,9*	611 875	*37*	0,243	5,136	*4,97*	*17,7*	*sans cohérence*
		4m44s25c	*55*	32s84c	*11,5*	292 229	*27,9*	692 540	*41,9*	0,078	2,385	*4,33*	*21,4*	*avec cohérence*

Scène	r/p	Temps utilisateur				Pixels échantillonnés		Rayons générés		L*u*v*			Mémoire (en Mo)	
		Total	%	Test seul	%	Total	%	Total	%	$\overline{\Delta}_{uv}$	σ_{uv}	% points ≠		
Verre	1	34m59s22c				262 144		3 612 752						*référence*
		31m03s69c	*88,8*	12s20c	*0,6*	153 987	*58,7*	3 118 910	*86,3*	0,087	3,132	*2,81*	*12*	*sans cohérence*
		30m57s08c	*88,5*	14s74c	*0,8*	155 462	*59,3*	3 115 843	*86,2*	0,06	1,996	*3,01*	*12,2*	*avec cohérence*
	4	2h23m46s91c				1 048 576		14 454 436						*référence*
		1h53m28s89c	*78,9*	37s32c	*0,5*	444 509	*42,4*	11 163 551	*77,2*	0,092	1,707	*4,41*	*32,6*	*sans cohérence*
		1h57m25s10c	*81,7*	47s36c	*0,7*	454 869	*43,4*	11 226 389	*77,7*	0,071	1,264	*4,33*	*33,5*	*avec cohérence*
Sphères	1	9m31s79c				262 144		911 216						*référence*
		9m00s60c	*94,5*	19s08c	*3,5*	206 791	*79*	780 523	*85,7*	0,166	4,2	*3,08*	*16,5*	*sans cohérence*
		8m59s26c	*94,3*	20s23c	*3,7*	213 118	*81,3*	799 240	*87,7*	0,107	3,507	*2,88*	*17,1*	*avec cohérence*
	4	38m03s95c				1 048 576		3 647 483						*référence*
		32m43s04c	*85,9*	1m07s53c	*3,4*	716 615	*68,3*	2 822 249	*77,4*	0,185	2,455	*9,86*	*53,3*	*sans cohérence*
		34m48s23c	*91,4*	1m33s39c	*4,5*	748 579	*71,4*	2 919 164	*80*	0,124	1,556	*9,43*	*55,8*	*avec cohérence*

Les images qui suivent montrent les résultats de ces tests avec un ou quatre rayons par pixel selon les cas. Les images intitulées *différences* représentent les différences colorimétriques entre la référence et nos calculs progressifs. Les parties blanches de ces images sont strictement identiques à la référence, les points colorés indiquent des distances Δ_{uv} non nulles avec une échelle de couleur mentionnée sur le côté droit de ces images et graduée de 0 à 10. Une barre verticale à l'extrême droite situe la distance moyenne ($\overline{\Delta}_{uv}$) entre les images et la demie-étendue de cette barre a la taille de l'écart-type (σ_{uv}). Les différences indiscernables ($\overline{\Delta}_{uv} < 2$) sont représentées dans des tons de bleus (du foncé = 0 au cyan = 2), les tons dégradés entre le jaune et le rouge indiquent des différences comprises entre 2 et 10, le violet est réservé aux différences supérieures à 10. La bande horizontale supérieure quant à elle, représente la proportion de points différents dans chacune des 11 classes ([0, 1[; [1, 2[;…;]10, +∞[) et permet de juger de la répartition globale des erreurs. Les images des points échantillonnés avec 4 rayons par pixel ont la signification suivante : aucun point n'a été échantillonné dans les parties blanches, les points bleus signifient 1 point par pixel, les points jaunes correspondent à 3 points par pixel et les points rouges signifient 4 points par pixel.

The famous boules

Figure 90: Référence (1 rayon / pixel)

Figure 91: Sans cohérence spatiale

Figure 93: Avec cohérence spatiale

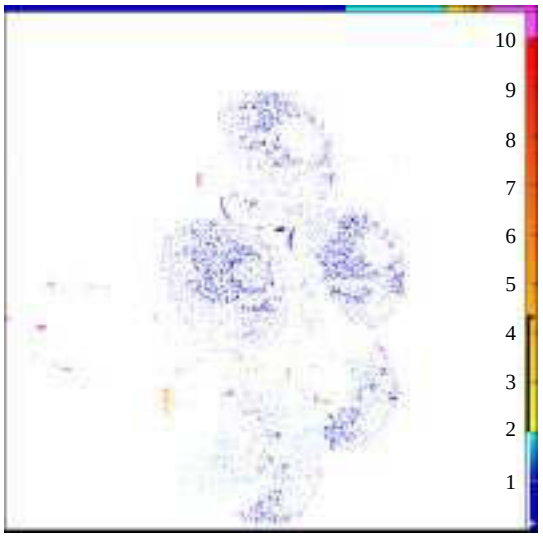


Figure 92: Différences

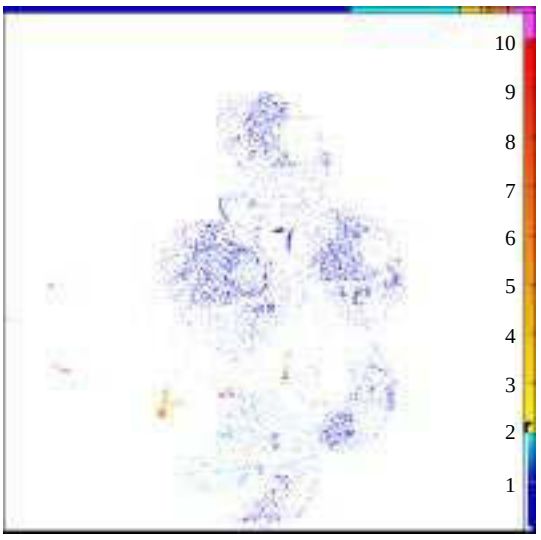


Figure 94: Différences

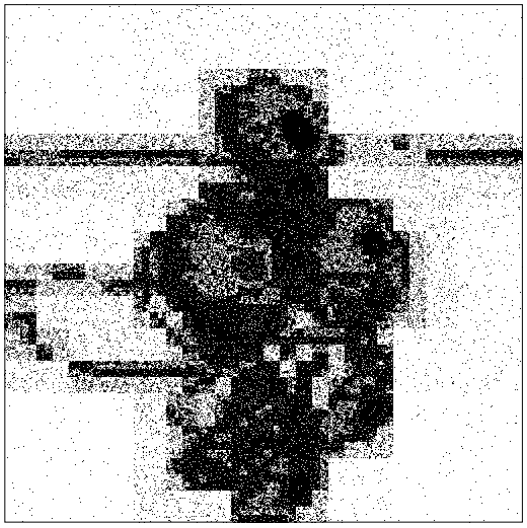


Figure 95: Points échantillonnés sans cohérence

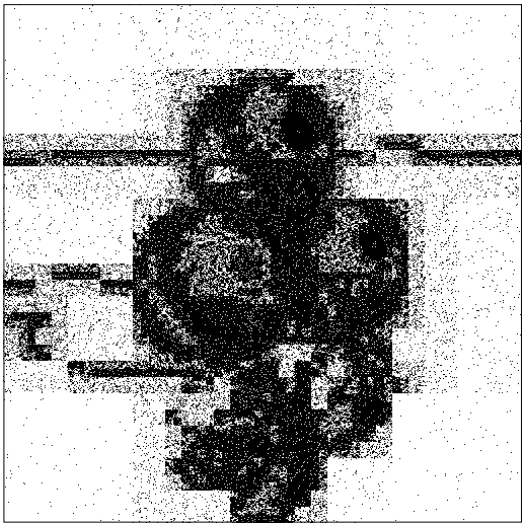


Figure 97: Points échantillonnés avec cohérence

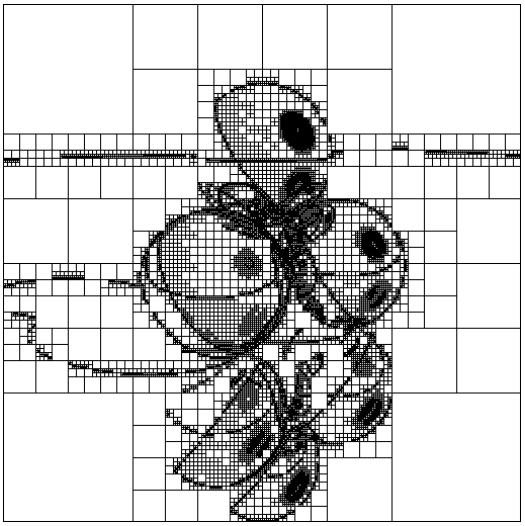


Figure 96: Régions homogènes sans cohérence

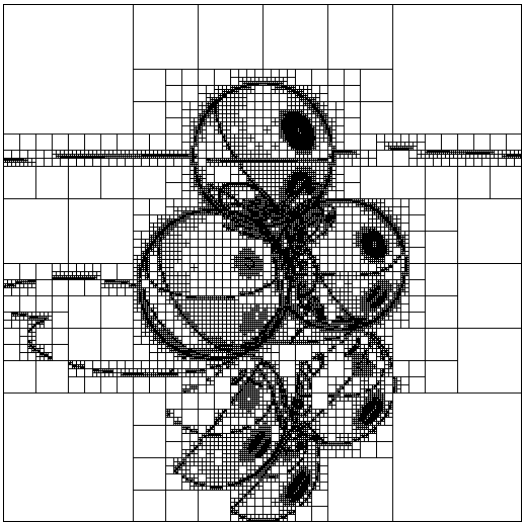


Figure 98: Régions homogènes avec cohérence

Oiseau

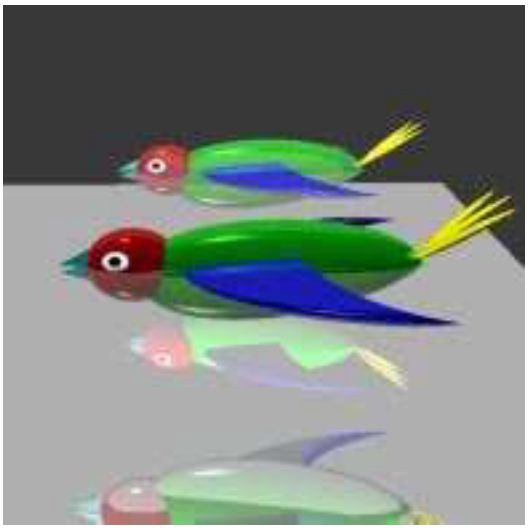


Figure 99: Référence (1 rayon / pixel)

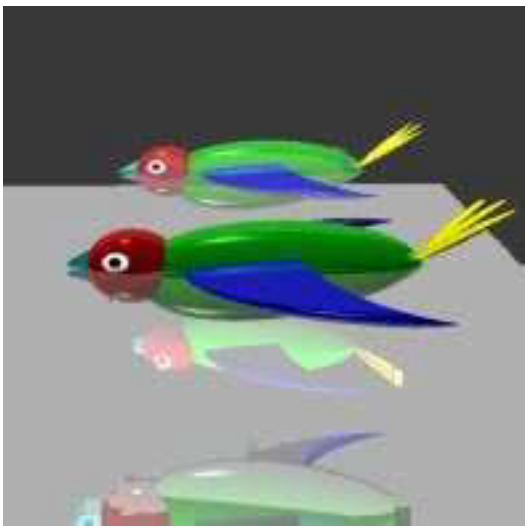


Figure 100: Sans cohérence spatiale

Figure 101: Différences

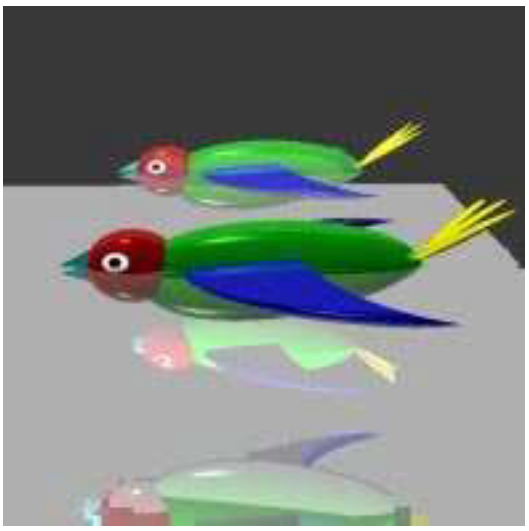


Figure 102: Avec cohérence spatiale

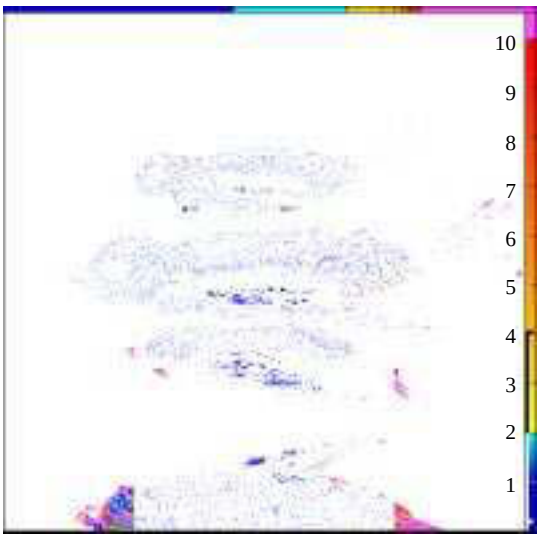


Figure 103: Différences

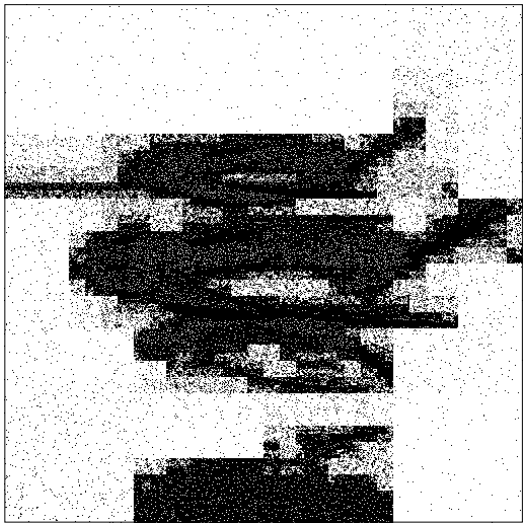


Figure 104: Points échantillonnés sans cohérence

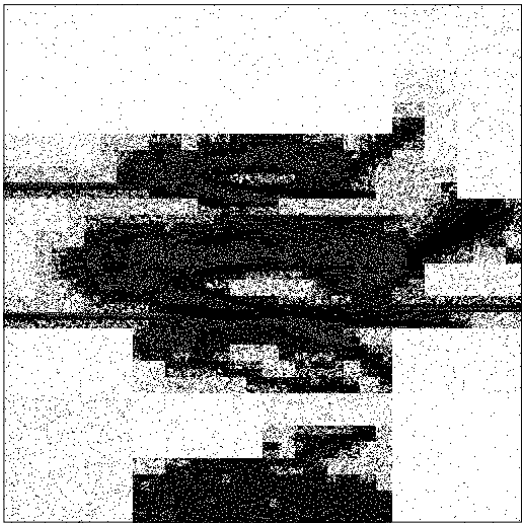


Figure 106: Points échantillonnés avec cohérence

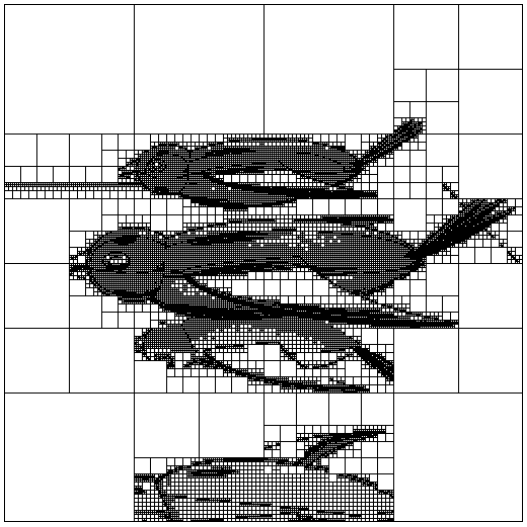


Figure 105: Régions homogènes sans cohérence

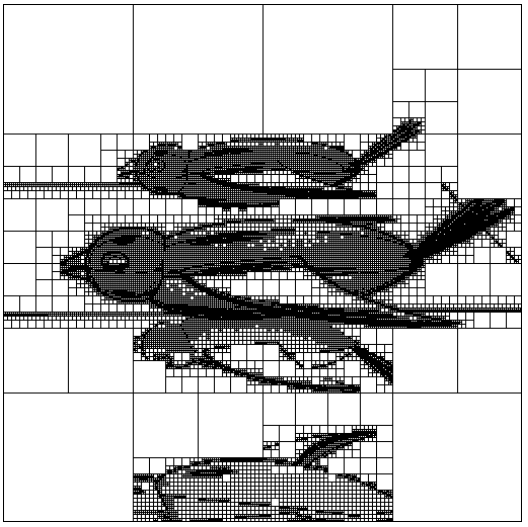


Figure 107: Régions homogènes avec cohérence

Poissons

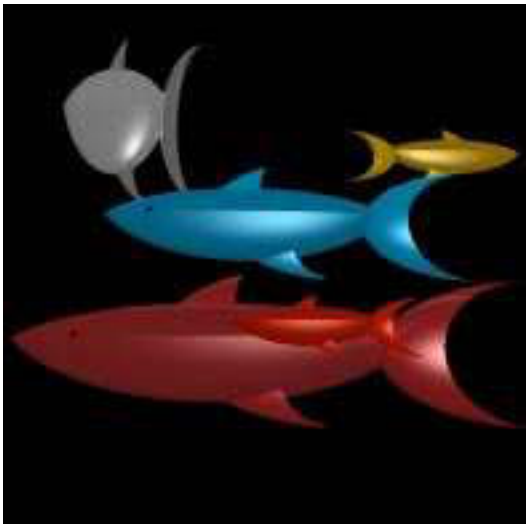


Figure 108: Référence (4 rayons / pixel)

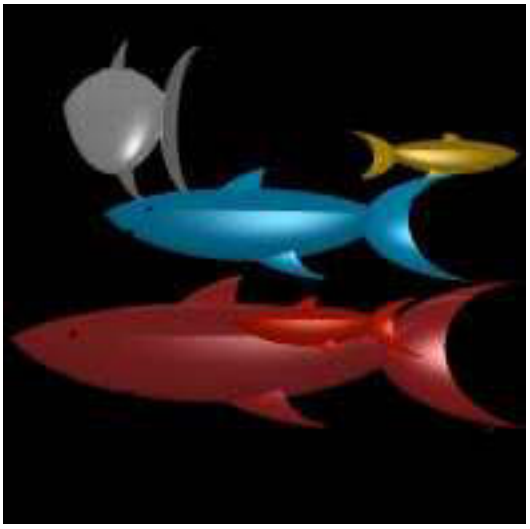


Figure 109: Sans cohérence spatiale

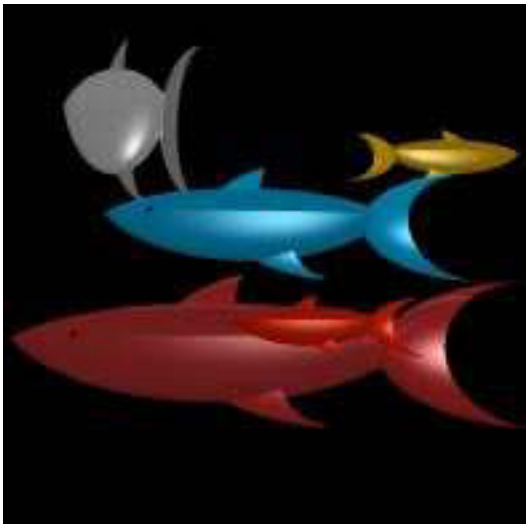


Figure 111: Avec cohérence spatiale

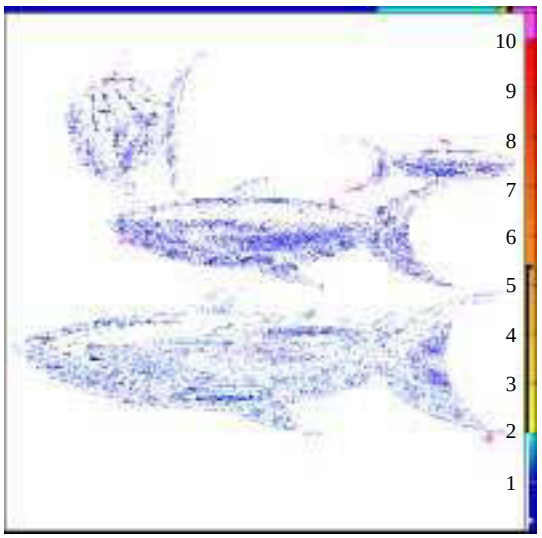


Figure 110: Différences

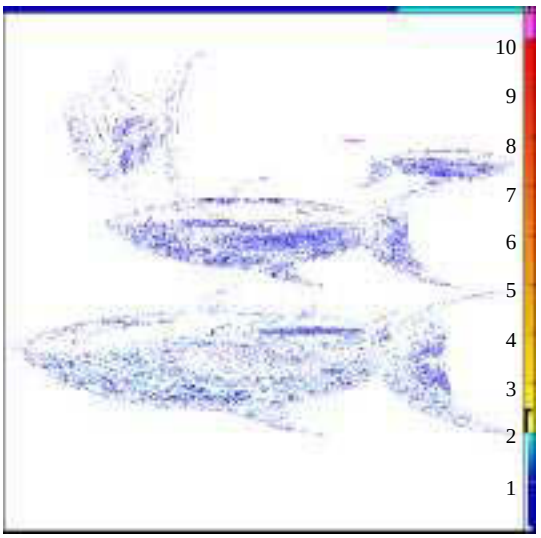


Figure 112: Différences

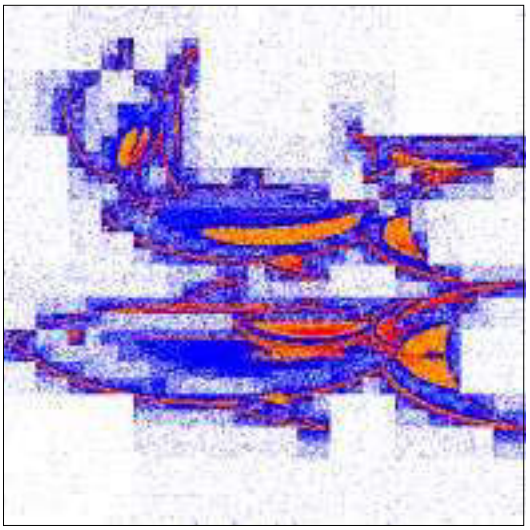


Figure 113: Points échantillonnés sans cohérence

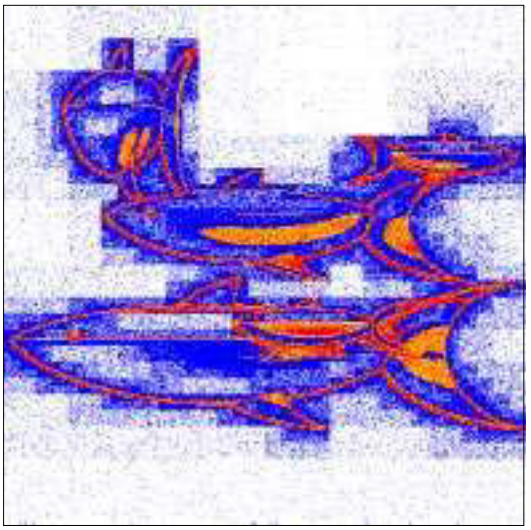


Figure 115: Points échantillonnés avec cohérence

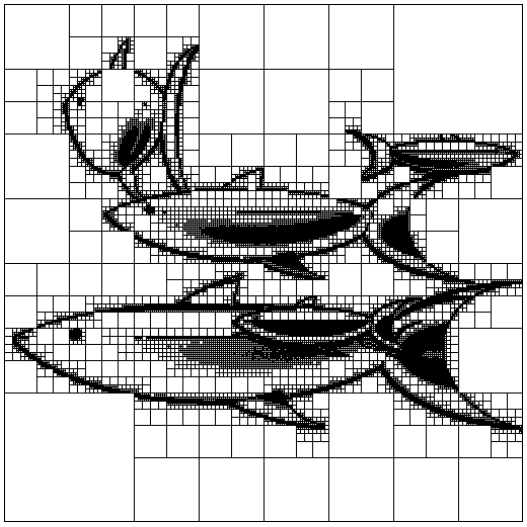


Figure 114: Régions homogènes sans cohérence

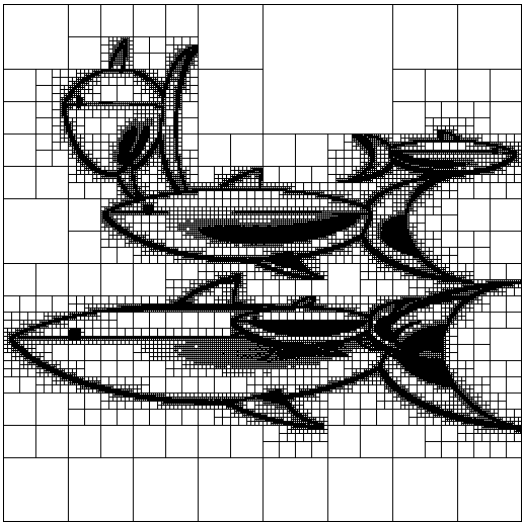


Figure 116: Régions homogènes avec cohérence

Verre

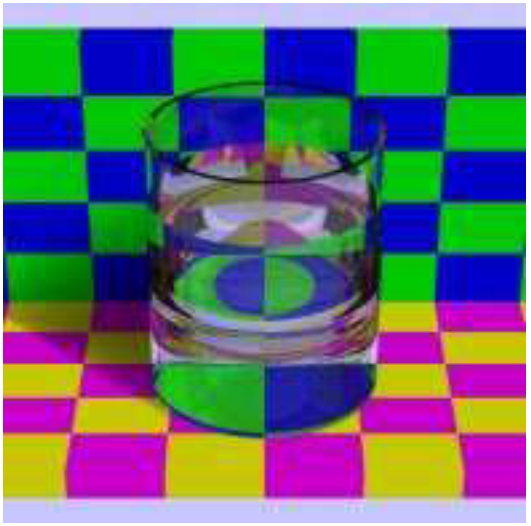


Figure 117: Référence (4 rayons / pixel)

Figure 118: Sans cohérence spatiale

Figure 120: Avec cohérence spatiale

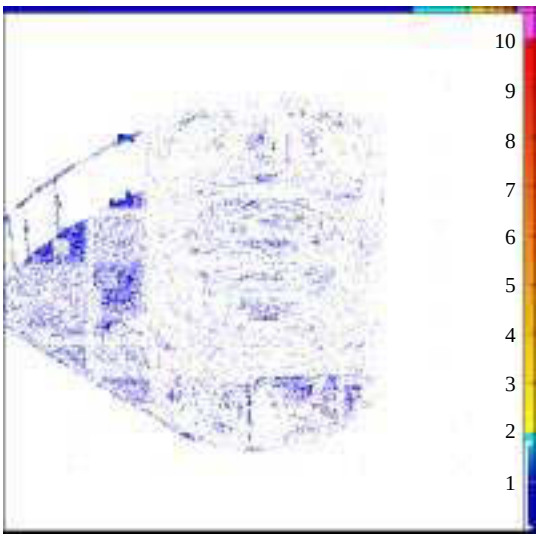


Figure 119: Différences

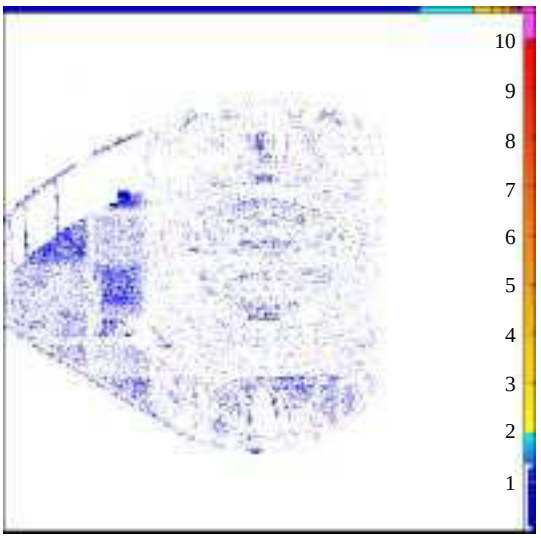


Figure 121: Différences

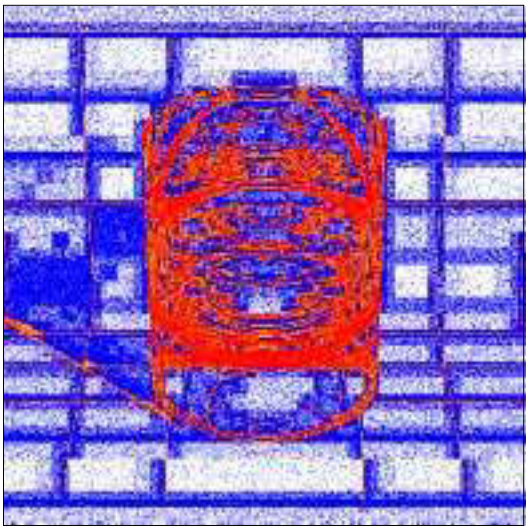


Figure 122: Points échantillonnés sans cohérence

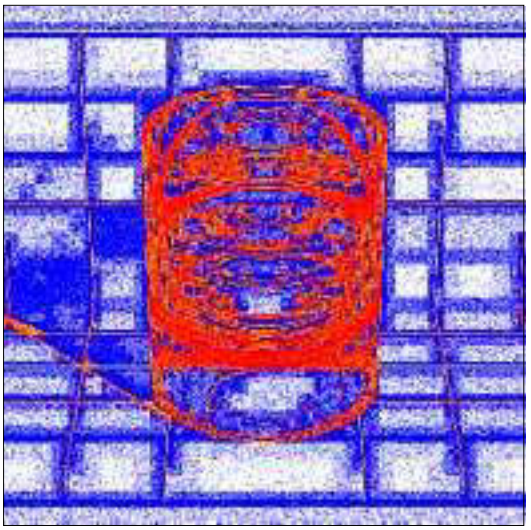


Figure 124: Points échantillonnés avec cohérence

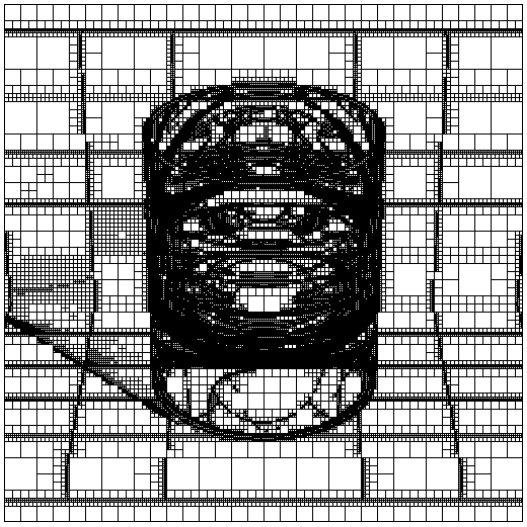


Figure 123: Régions homogènes sans cohérence

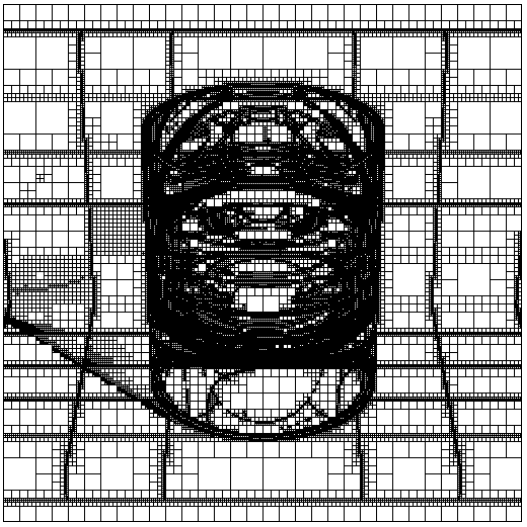


Figure 125: Régions homogènes avec cohérence

Sphères

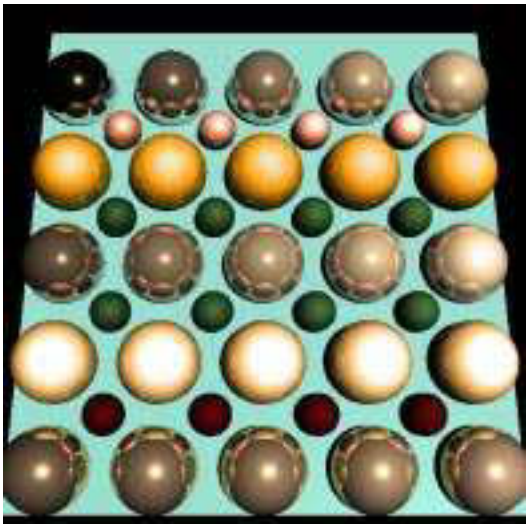


Figure 126: Référence - 1 rayon / pixel

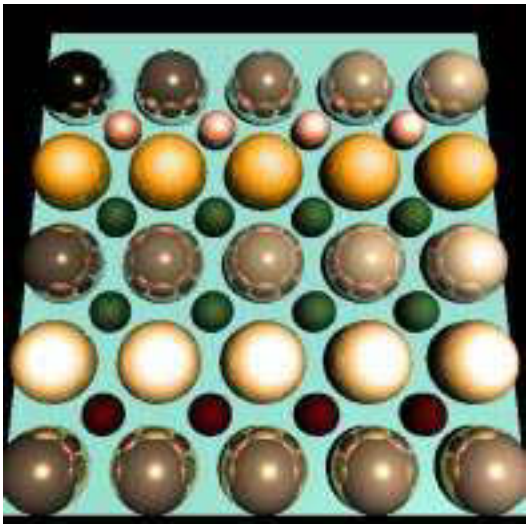


Figure 127: Sans cohérence spatiale

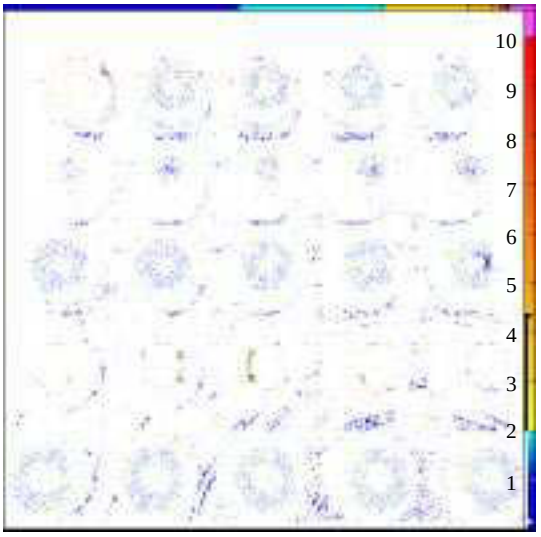


Figure 128: Différences

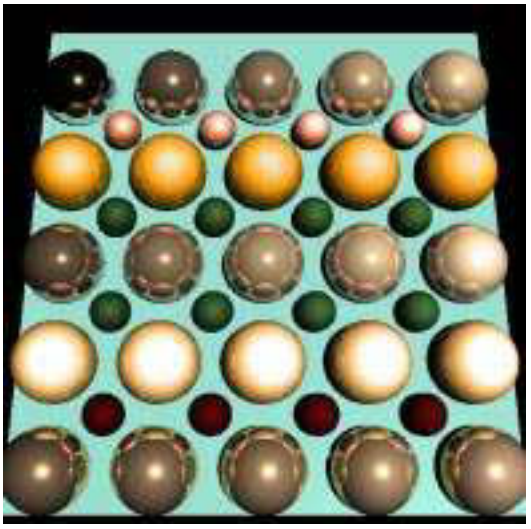


Figure 129: Avec cohérence spatiale

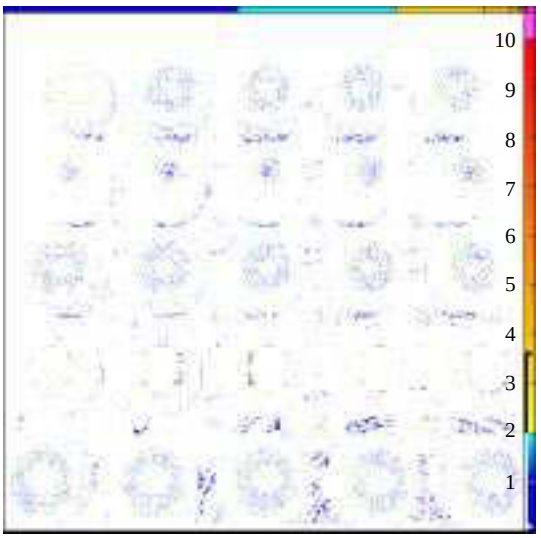


Figure 130: Différences

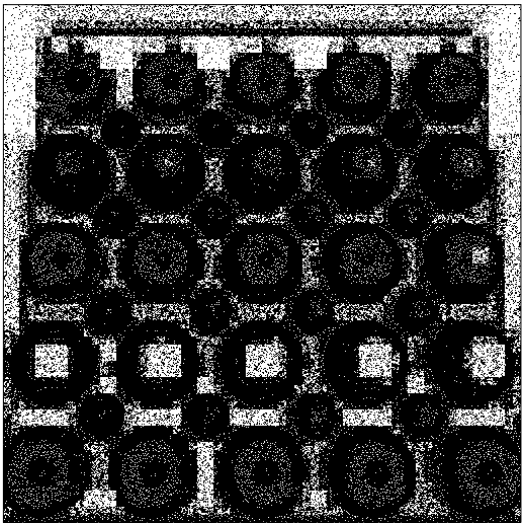


Figure 131: Points échantillonnés sans cohérence

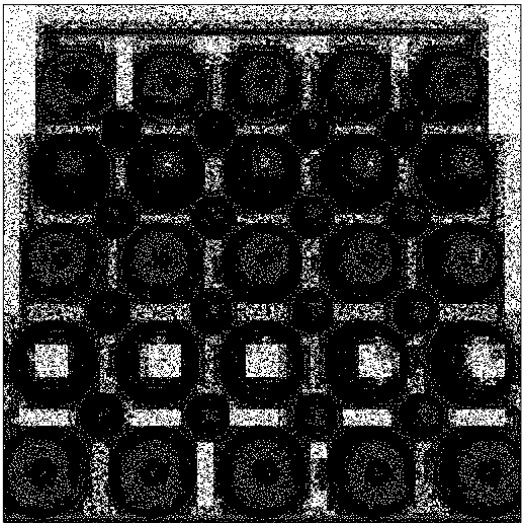


Figure 133: Points échantillonnés avec cohérence

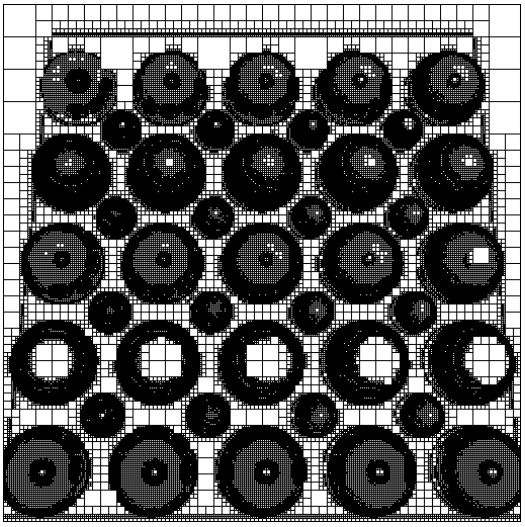


Figure 132: Régions homogènes sans cohérence

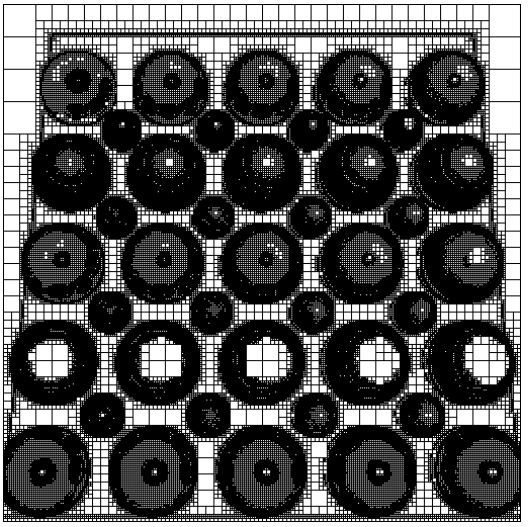


Figure 134: Régions homogènes avec cohérence

IV.5.3.1 Performances

Notre tracé de rayons progressif, avec ou sans cohérence spatiale, peut permettre un gain de temps important par rapport au tracé de rayons classique. Selon la complexité de la scène, ce gain varie de 60 à 5 % pour le test sans cohérence, de 56 à 5 % pour le test avec cohérence. Il arrive rarement que le processus progressif prenne plus de temps que la méthode classique. Le test séquentiel express ne représente pas plus de 9 % du temps de calcul total (souvent moins) ce qui est très modique et montre l'efficacité d'une bonne implémentation compte tenu des gains possibles et des dizaines de milliers de tests réalisés. La vérification de la cohérence est tout aussi performante puisqu'elle ne demande pas plus de 4 % du temps total.

Ce gain provient de l'économie de points et donc de rayons réalisée. La proportion de points nécessaire à une image peut-être très faible pour des images simples (jusqu'à un cinquième de la définition) et s'élever à 88 % pour une complexité importante. La proportion de rayons générés varie entre 30 et 88 % ce qui représente un gain

substantiel. L'augmentation de la définition de l'image conduit à un accroissement des gains en rapidité puisque les zones homogènes occupent une aire plus grande, une conséquence logique mais très appréciable. L'ensemble de ces remarques montre que les décisions hétérogènes, visuellement inappréciables, sont tout à fait correctes et conformes à ce qu'en prédisaient nos tests antérieurs, la rapidité du test en témoigne.

Comme il était prévisible, la mémoire utilisée pour le test est très importante : de 6 à 17 Mo sans suréchantillonnage, de 18 à 55 avec 4 rayons par pixel. Les deux tiers sont employés au stockage des listes de points, le reste est surtout consacré à l'arbre des régions et aux arbres d'échantillonnages. Les éléments mémorisés sont loin d'être tous indispensables à la bonne marche du test. Les structures d'arbres que nous avons utilisées, par exemple, ne servent qu'à ordonner les points et les régions, une liste convenablement triée pourrait parfaitement jouer le même rôle pour des coûts mémoires incomparablement moindres. Le temps d'accès et de construction de ces structures seraient plus élevés mais pourraient peut-être se compenser en partie par un swap moins important.

IV.5.3.2 Qualité

La distance perceptuelle que nous utilisons est en général acceptable pour le procédé avec cohérence (moins sans cohérence), tant pour l'écart-type que pour la moyenne. Mais, comme nous l'avions signalé, elle rend imparfaitement compte de la différence entre images. La répartition spatiale des erreurs est, en générale, très clairsemée. De ce fait, même lorsque l'écart-type est élevé, il est souvent réparti en petites entités sur de grandes portions de l'image comme noyées dans l'ensemble. Dans ce sens, on constate que l'incertitude sur les décisions homogènes pour les tailles de régions 2×2 est sans conséquence. Cette bonne répartition eu égard à la sensation visuelle est d'ailleurs due à la nature aléatoire de notre processus dont le comportement moyen est assuré. Les seules erreurs préjudiciables se produisent lorsque de gros amas de points sont mal représentés (voir l'image de l'oiseau). De ce point de vue, la vérification de la cohérence apporte une réelle amélioration aux décisions homogènes. Bien que portant sur peu de points (3 à 5 %), ces corrections se situent sur des endroits «stratégiques» que l'oeil détecte à coup sûr lorsqu'elles sont absentes. La distance moyenne et son écart-type sont d'ailleurs très significativement diminués par cette dernière passe de vérification pour un coût très modeste.

Il reste parfois d'assez importantes régions homogènes récalcitrantes à la détection. Il y a plusieurs raisons à cela. La première est contenue dans la nature aléatoire du procédé qui n'a pas que des avantages. Quels que soient les paramètres et les précautions, l'échantillonnage peut toujours avoir des ratés, et ce, avec une probabilité irréductiblement non nulle. C'est typiquement le cas pour le reflet manquant de la queue de l'oiseau sur le plan horizontal : aucun point extérieur à la moyenne n'a été échantillonné et le test ne peut évidemment pas inventer les informations qu'il ne possède pas. Ce type de risque est heureusement marginal et peut être résolu en partie par la cohérence. Il peut également être sensiblement diminué par l'adoption d'un autre type d'échantillonnage. La stratification notamment serait sans doute une bonne alternative. Malheureusement, toute nouvelle forme d'échantillonnage passe nécessairement par la capacité de calculer la probabilité d'apparition du caractère homogène : une difficulté souvent insurmontable. La seconde raison à la mauvaise représentation de certaines régions provient de l'interpolation utilisée. Lorsque les échantillons sont représentatifs, l'interpolation que nous avons retenue est justifiée. Lorsqu'ils ne le sont pas suffisament, on peut penser que d'autres interpolations pourraient être meilleures. Cette hypothèse a été explorée lors

d'un stage de DEA par F.Sauzet par l'utilisation de l'interpolation de Clough-Toucher sur une triangulation de Delaunay des échantillons. Les résultats sont malheureusement forts coûteux et visuellement plus gênants.

Les défauts résiduels peuvent être également corrigés par l'interactivité de notre implémentation qui permet à l'utilisateur de découper manuellement n'importe quelle région et de relancer un test spécifiquement sur celle-ci.

IV.5.4 Paramètres préconisés

Les paramètres que nous préconisons pour notre tracé de rayons progressif sont les suivants :

$$\varepsilon \in [1, 3]; p_h \in [0{,}75; 0{,}95]; \mu_{\bar{h}\bar{h}} \in [0{,}01; 0{,}2]; \alpha' \in [0{,}01; 0{,}05]; \beta' \in [0{,}01; 0{,}05]$$
$$+\leftarrow - \qquad - \rightarrow + \qquad +\leftarrow - \qquad +\leftarrow - \qquad +\leftarrow -$$

La flèche indique dans quel sens la qualité du résultat augmente. Ainsi par exemple, la qualité de l'image finale sera meilleure avec $\varepsilon = 1$ qu'avec $\varepsilon = 3$.

La moyenne étant calculée sur le nombre de points suivants :

Taille	2×2	4×4	8×8	16×16	≥ 32×32
n_m	3	[9, 12]	[36, 39]	[59, 66]	71
		$- \rightarrow +$	$- \rightarrow +$	$- \rightarrow +$	

Tableau 13: Nombre d'échantillons sur lesquels la moyenne approchée est calculée

IV.6 Conclusions

Partant de l'idée de progressivité, nous avons proposé une méthode originale de calcul incrémental de tracé de rayons. Cette démarche s'appuie sur un concept précis d'homogénéité colorimétrique mesurée par un ensemble de techniques spécialement adaptées, dont le coeur est constitué de méthodes statistiques puissantes encore inutilisées en images de synthèse. Nous nous sommes attachés à l'aspect mathématique de nos développements autant qu'à leur pertinence pour le système visuel humain. Nous nous sommes ainsi intéressés à la répartition de notre critère d'homogénéité perceptuelle sur un panel d'images «représentatives» et avons pu constater l'émergence de grandes familles caractéristiques insoupçonnées. Dans le même esprit, nous avons mesuré nos résultats à l'aune d'une distance colorimétrique, qui, sans être parfaite, nous a permis d'optimiser notre méthode en tenant compte de certains aspects de la vision.

L'article que nous avons publié en 92 sur notre tracé de rayon progressif [Mail et al 92], est très incomplet et peut induire en erreur. Tout au long de ce chapitre, nous nous sommes attachés à comprendre pourquoi notre méthode peut produire des résultats parfois surprenants de rapidité alors qu'elle possède trois biais majeurs : l'approximation de la moyenne, le recyclage des points et l'éloignement des probabilités de base. Nous espérons avoir fourni les éléments indispensables à la compréhension du phénomène et avoir démontré comment ces biais pouvaient se neutraliser et être neutralisés par la mise en oeuvre de contre-mesures adaptées.

L'ensemble de ces techniques forme un tout qui permet souvent des gains de rapidité très appréciables sur une méthode de tracé de rayons classique. Notre tracé de rayons progressif permet facilement des gains de 40 % pour une qualité souvent très bonne. Les erreurs de décision concernant l'homogénéité sont relativement rares et les

risques réels attachés à ces décisions, s'ils ne sont pas parfaitement contrôlables, restent pratiquement faibles. Ces erreurs passent souvent inaperçues compte tenu de leur taille mais pas toujours. Pour éliminer certaines de ces erreurs les plus génantes, nous avons proposé un critère de cohérence entre l'homogénéité de l'image et l'espace des objets de la scène. Critère extensible à d'autres caractéristiques sans trop de difficultés ni surcoût exhorbitant avec la garantie d'un meilleur résultat. Compte tenu de ces imperfections, notre tracé de rayons ne permet pas encore de générer des images parfaites à coup sûr mais constitue, en attendant, un très bon outil pour la mise au point d'images.

Nous voulons souligner la signification particulièrement forte de la moyenne pour l'oeil ainsi que la pertinence de notre critère d'homogénéité. Ce critère peut être affiné, notamment par la prise en compte de l'acuité visuelle, et il serait très intéressant de le confronter à des expériences visuelles plus poussées et plus systématiques pour en cerner l'essence exacte. Nous insistons encore sur l'absence et l'indispensable nécessité d'études psycho-visuelles approfondies permettant de caractériser des images par des paramètres significatifs.

Comme nous l'avions annoncé, ce travail est un point d'entrée dans la progressivité et l'exploitation des cohérences du tracé de rayons. De nombreuses voies restent ouvertes pour étendre le concept à d'autres domaines du tracé de rayons. Nous pensons particulièrement au processus d'intersection qui pourrait réutiliser les connaissances acquises lors de rayons précédents, que ce soit pour privilégier certaines boîtes englobantes ou pour simplifier la résolution d'équations réalisées par l'intersecteur. Mais aussi, pour utiliser des modèles d'éclairements de complexité croissante au cours du processus progressif ou encore dans le domaine de l'accélération du calcul d'animations.

Conclusions

1. Quelques mots sur l'implémentation

L'ensemble du tracé de rayons spectral, global et progressif que nous avons présenté tout au long de ces pages a fait l'objet d'une implémentation complète en C. Les modules que nous avons développés représentent plus de 15 000 lignes de code et ont été construits à partir d'un intersecteur créé par Marc Roelens qui a lentement et péniblement évolué vers un tracé de rayons. Les scènes sont modélisées en Castor, un langage de description de scène CSG propre à l'École des Mines.

Parallèlement à ces modules, nous avons construit une base de donnée spectrale de plus de 250 matériaux. La plupart sont représentés sous forme de réflectance spectrale à incidence normale sur une partie du visible photopique. Le reste (surtout des métaux) est décrit par des indices optiques complexes sur un certain nombre de longueurs d'ondes. Ces deux types de représentations sont manipulables par nos modules et reçoivent des traitements spécifiques lors du calcul de la réflexion ou de la transmission. Une dizaine d'illuminants courants font également partie de cette base de donnée ainsi que les courbes de sensibilité **XYZ** sur le domaine photopique par pas de 1 `nm`. Un outil indépendant de visualisation de spectres a également été conçu comme aide à la modélisation pour visualiser une surface plane d'une matière quelconque éclairée, en incidence normale, par une source ponctuelle de spectre quelconque.

Rien n'empêche de remplacer les courbes de sensibilité photopiques par d'autres. Des courbes de sensibilité scotopiques par exemple, pour simuler une vision nocturne ou, pourquoi pas, rêvons un peu, la sensibilité d'une plaque photographique pour radiographie médicale. Les modèles d'éclairements ou de transfert d'énergie que nous avons implémentés (Phong, Whitted, Schlick, Torrance avec des modèles d'auto-ombrage et de distribution de facettes divers et variés) sortiraient largement de leur domaine de validité dans ce cas mais leur liste n'est pas limitative et l'inclusion de nouveaux modèles est aisée, alors pourquoi pas? L'ensemble de nos modules sait gérer des matériaux et des lumières de composition spectrale quelconque et non uniforme sur l'ensemble d'une scène : le nombre de longueurs d'ondes portées par un rayon n'est pas fixé à l'avance mais varie en fonction des sources et des objets rencontrés.

Les objets transparents sont incomplétement traités dans notre implémentation. Si nous diffractons correctement un rayon spectral en provenance de l'oeil, les rayons issus des sources, par contre, ne provoquent pas de caustiques avec les modèles locaux. Notre tracé de rayons global permettrait de gérer ces effets lumineux avec un modèle de transfert d'énergie adapté, un processus d'échantillonnage a priori qui serait symétrique par rapport à la normale et un échantillonnage a posteriori modifié.

2. Bilan et prospectives

2.1 Méthodes de Monte-Carlo

Le modèle de transfert d'énergie que nous avons choisi pour notre tracé de rayons global est insuffisant comme nous l'avons souligné. La recherche d'une description fiable et précise du comportement des ondes électro-magnétiques sur des surfaces rugueuses est encore très actuelle en physique. Force est de constater qu'un modèle vectoriel tenant compte des auto-ombrages et des réflexions multiples ne semble pas encore prêt. On peut légitimement s'interroger sur le degré de précisions requis pour que les calculs «synthétiques» gardent une signification. À notre sens, il n'y a guère que la comparaison entre mesures physiques et simulations qui puisse apporter la réponse. Partant de là, il sera possible d'utiliser des approximations en connaissance de cause et de tirer partie des mécanismes psycho-visuels.

Nous regrettons le manque de données physiques relatives à ces domaines. En ce qui concerne la description des matériaux et notamment de leurs indices optiques, ce n'est pas tant que les études manquent mais plutôt qu'elles portent souvent sur des matières, sans doute très excitantes en physique, mais de peu d'usage pour nous : visualiser une fourchette en beryllium est d'un intérêt limité… Les descriptions sous forme de réflectance à incidence normale sont plus nombreuses et plus facilement utiles (bien que souvent insuffisamment précises) mais les indices optiques ne sont pas toujours reconstituables à partir de ces données, ce qui rend le facteur de Fresnel incalculable. La caractérisation de surfaces doit certainement avoir fait l'objet d'études. Il serait très intéressant de les confronter aux hypothèses des modèles utilisés. Enfin, l'interpolation nécessaire des données spectrales, des indices optiques comme des répartition énergétiques, reste un problème majeur jamais abordé en synthèse.

L'approche statistique que nous avons adoptée pour la résolution des échanges énergétiques est naturellement adaptée à l'essence aléatoire de la trajectoire des photons. Notre formulation nouvelle du phénomène, sous forme de processus markovien, ouvre la voie à des techniques propres à cette branche des probabilités et encore inemployées en synthèse. L'utilisation des méthodes de Monte-Carlo est récente en synthèse et a déjà été appliquée à l'échantillonnage de fonctions de réflexion simples (Phong) ou pour démarrer l'échantillonnage depuis les sources. Les techniques que nous proposons ont ceci d'original qu'elles s'adaptent, par construction, à la complexité de chaque flux lumineux.

La combinaison de nos échantillonnages a priori et a posteriori permet de favoriser les directions les plus porteuses d'énergie de prime abord pour progressivement équilibrer les contributions de l'éclairage direct et indirect en rapport avec leur importance relative. Les schémas d'échantillonnages a priori que nous avons proposés permettent de traiter des fonctions de transfert d'énergie variées pourvu qu'elles conservent l'énergie et qu'elles aient des «formes» diffuses et spéculaires suffisamment prononcées. Nous insistons sur la conservation de l'énergie qui n'est pas seulement une propriété physique requise mais également une nécessité, intrinsèque en un sens, de tout processus de Monte-Carlo adaptatif traitant ce problème.

Le traitement de la transparence n'est qu'une extension symétrique du mélange diffus/spéculaire et de la stratégie du tournesol pour l'échantillonnage a priori. Il implique également l'ajout d'un critère de choix supplémentaire dans l'échantillonnage a posteriori et ne modifie pas notre estimateur de niveau. Une fonction de transfert adaptée devrait bien sûr être employée. Ces extensions sont somme toute fort abordables, mais nous osons à peine imaginer le temps de calcul de telles images…

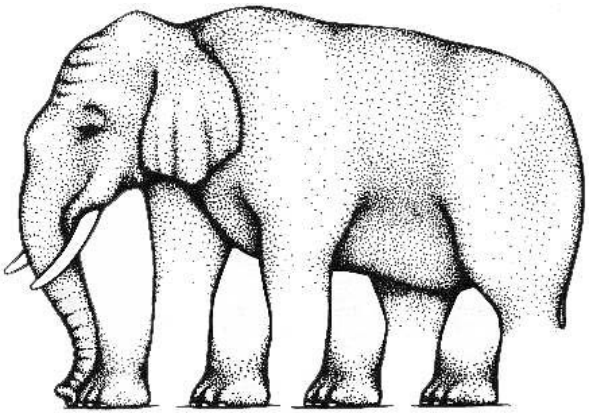
2.2 Progressivité

Le paradigme progressif que nous avons introduit pour l'échantillonnage des points de l'écran s'est révélé riche et fructueux. Notre tracé de rayons progressif permet en effet des gains de rapidité importants par rapport à une méthode classique avec une qualité d'image largement suffisante pour la mise au point d'une scène, parfois insuffisante pour une image parfaite. Les dernières lacunes de notre méthode peuvent être comblées à peu de frais par l'utilisateur à l'aide de notre implémentation interactive. Ceci peut être viable pour des images fixes mais sûrement pas pour des séquences animées. Les autres améliorations, facilement implémentables, portent sur l'extension de la cohérence entre l'homogénéité d'une région et l'ensemble objets/effets lumineux qu'elle contient. Cette extension permettrait de tirer pleinement profit de la très forte cohérence qui sous-tend profondément le tracé de rayons. Cette voie de recherche prenant l'idée de cohérence comme concept de base en soi, bien qu'encore peu exploitée (Akimoto est sur cette voie [Akim et al 91]) est prometteuse dans bien des domaines (rendu, intersection, animation).

La notion d'homogénéité que nous avons introduite contient une signification visuelle forte qui mérite d'être explorée davantage par de plus amples études. Dans ce sens, l'idée que l'homogénéité est plus facilement caractérisable que l'inhomogénéité semble un point d'importance. L'utilisation implicite de l'acuité visuelle que nous réalisons par l'adaptation de nos paramètres à la taille de région testée pourrait s'enrichir d'études dans ce domaine (voir par exemple Burt [Burt 88]).

La distance entre images que nous avons utilisée manque de précision mais nous a permis de discriminer certains paramètres de notre méthode. Au-delà, la nécessité d'une telle distance est tout à fait claire. Il est néanmoins peu probable qu'une telle mesure puisse être universelle étant données les attentes très diverses que peut recouvrir ce type de différence selon son domaine d'application. Il n'empêche, des recherches approfondies sont fortement désirables et ne peuvent être menées qu'en étroit partenariat avec des psycho-physiologues de la vision.

Dans le même esprit, il serait très intéressant de caractériser des images en général et plus spécialement celles que nous pouvons ou voulons produire. C'est ce que nous avons réalisé, à petite échelle, sur notre panel d'images pour les caractéristiques qui nous intéressaient. Nous avons pu constater, avec quelque surprise, que ce que nous pensions inorganisé et difficilement qualifiable, possédait en fait une structure sous-jacente. Ce type d'étude devrait être largement étendu, tant pour des images «naturelles» que «synthétiques». Nous sommes convaincus que des caractéristiques propres à chacun de ces types d'image émergeraient et pourraient éclairer des aspects du réalisme dont nous n'avons pas conscience et qui font souvent défaut, de façon insaisissable, aux images de synthèse actuelles.



Bibliographie

[Afe 91] *La Photomètrie en éclairage*
Association française d'Éclairage - Société d'édition Lux - 1991

[Akim et al 91] **T. Akimoto, K. Mase, Y. Suenaga**
Pixel-selected ray tracing
IEEE computer graphics and applications - 1991
vol. 11 - num. 4

[Beck et al 63] **P. Beckmann, A. Spizzichino**
The scattering of electromagnetic waves from rough surfaces
Pergamon press - 1963

[Beck 65] **P. Beckmann**
Shadowing of random rough surfaces
IEEE transactions on antennas and propagation - 1965
vol. AP-13 - p.384-388

[Benn et al 89] **J.M. Bennett, L. Mattsson**
Introduction to surface roughness and scattering
Optical Society of America - Washington D.C - 1989

[Broc et al 66] **R.A. Brockelman, T. Hagfors**
Note on the effect of shadowing on the backscattering of waves from a random rough surface
IEEE transactions on antennas and propagation - 1966
vol. AP-14 - num 5 - p. 621-629

[Blin 77] **J.F. Blinn**
Models of light reflection for computer synthesized pictures
Computer Graphics - 1977
vol. 11 - p. 192-198

[Brow 91] **W.T. Brown**
Filtering and adaptative supersampling for ray traced rendering
University of Illinois - Urbana-Champaign - Technical report GWRG-91-16 - 1991

[Burt 88] **P.J. Burt**
Attention mechanisms for vision in dynamic world
Proc. 9th ICPR - 1988
p. 977-987

[Calle 93] **P. Callet**
Contribution à la modèlisation de l'interaction lumière-matière pour la visualisation réaliste en synthèse d'image
Thèse de doctorat - École centrale de Paris - 1993

[Cohe et al 85] **M. Cohen, D.P. Greenberg**
The hemi-cube : a radiosity solution for complex environments
Computer Graphics - 1985
vol. 19 - num. 3 - p. 31-40

[Cook 81] **R.L. Cook**
A reflection model for realistic image synthesis
Master's thesis - Cornell University - Ithaca - 1981

[Cook et al 81] **R.L. Cook, K.E. Sparrow**
A reflectance model for computer graphics
Computer graphics - 1981
vol. 15, num. 3, p. 307-316

[Cook et al 82] **R.L. Cook, K.E. Sparrow**
A reflectance model for computer graphics
ACM Transactions on graphics
vol. 1, num. 1, p. 7-24

[Cook et al 84] **R.L. Cook, T. Porter, L. Carpenter**
Distributed ray tracing
Computer Graphics - 1984
vol. 18 - num. 3 - p. 137-144

[Chri 79] **J.S. Christie Jr**
Review of geometric attributes of appearance
Journal of coatings technology - 1979
vol. 51 - num 653 - p. 64-73

[Davi 56] **H. Davies**
The reflection of electromagnetic waves from a rough surface
Proc. inst. elec. eng IV - 1954
101 - p. 209-214

[Dine 94] **E. Dinet**
Contribution de modèles de vision humaine à la définition de méthodes multirésolutions en vision artificielle
Thèse de doctorat - Université de Saint-Étienne - 1994

[Dipp et al 85] **M.A. Dippé, E.H. Wold**
Antialiasing through stochastic sampling
Communication of the ACM - 1985
vol. 19 - num. 3 - p. 69-78

[Edho 93] **P.R. Edholm**
How to lie and confuse with visualization
Computer Graphics proceedings - 1993
p. 387-388

[Fish et al 94] **F. Fisher, A. Woo**
R.E versus N.H specular highlights
Graphics Gems - P.S. Heckert editor - 1994
p. 388-400

[Four 79] **G. Fournet**
Electromagnétisme à partir des équations locales
Masson ed. 1979

[Guil 37] **P. Guillaume**
La psychologie de la forme
Flammarion- Coll Champs - 1937 (1ère édition)

[Gora et al 84] **C.M. Goral, K.E. Torrance, D.P. Greenberg, B. Battaile**
Modeling the interaction of light between diffuse surfaces
Computer Graphics - 1984
vol. 18 - num. 3 - p. 213-222

[Guth 70] **S. L. Guth**
Photometric and colorimetric additivity at various intensities
AIC Proceedings Color 69 - Stockholm - Musterschmidt-Verlag, Gottingen. 1970
p. 172

[Hall et al 83] **R.A. Hall, D.P. Greenberg**
A testbed for realistic image synthesis
IEEE Computer Graphics and Applications - 1983
Nov. 83 - p. 10-20

[He et al 91] **X.D. He, K.E. Torrance, F.X. Sillion, D.P. Greenberg**
A comprehensive physical model for light reflection
Computer graphics - 1991
vol. 25 - num. 4 - p. 175-186

[Hoc 85] **E.D. Palik ed.**
Handbook of optical constants of solids - Tome I et II
Academic Press. 1985

[Humm 93] **R.E. Hummel**
Electronic properties of material
Springer. 1993

[Jake 82] **E. Jakeman**
Fresnel scattering by a corrugated random surface with fractal slope
Journal of the optical society of America. 1982
Vol 72 - Num 8 - p. 1034-1041

[Jans 83] **F.W. Jansen**
Fast previewing technique in raster graphics
Eurographics - 1983
p. 195-202

[Jens 95] **H.W. Jensen**
Importance driven path tracing using the photon map
Eurographics workshop on rendering - Dublin - 1995
p. 359-369

[Kaji 86] **J.T. Kajiya**
The rendering equation
Computer Graphics - 1986
vol. 20 - num. 4 - p. 143-150

[Kirk et al 90] **J. Arvo, D. Kirk**
Particle transport and image synthesis
Computer Graphics - 1990
vol. 24 - num. 4 - p. 63-66

[Kowa 80] **P. Kowaliski**
Vision et mesure de la couleur
Ed. Masson - 1980

[Lang 91] **B. Lange**
The simulation of radiant light transfer with stochastic ray-tracing
Second Eurographics workshop on rendering - Barcelone - 92
p. 31-44

[Lafo et al 94] **E.P. Lafortune, Y.D. Willems**
A theoretical framework for physically based rendering
Cmputer graphics forum - 1994
vol. 13 - num. 2 - p. 97-107

[Lafo et al 94] **E.P. Lafortune, Y.D. Willems**
The ambient term as a variance reducing technique for Monte Carlo ray tracing
5th Eurographics workshop on rendering - 1994
p. 163-171

[Leco 95] **J. Lecomte**
Comment nous percevons le monde
Sciences humaines - Avril 1995
num. 49 - p. 16-17

[Lee et al 85] **M.E. Lee, R.A. Redner, S.P. Uselton**
Statistically optimized sampling for distributed ray tracing
Communication of the ACM - 1985
vol. 19 - num 3 - p. 61-65

[Lehm 59] **E.L. Lehmann**
Testing statistical hypothesis - 1959
Wiley et Sons, New York - 1959

[Lewi 93] **R. R. Lewis**
Making shaders more physically plausible
4th Eurographics workshop on rendering - 1993
p. 47-62

[Marx et al 90] **E. Marx, T.V. Vorburger**
Direct and inverse problems for light scattered by rough surfaces
Applied optics - 1990
vol. 29 - num. 25 - p. 3613-3626

[Mail et al 92] **L. Carraro, J-L. Maillot, B. Peroche**
Progressive ray tracing
Third Eurographics workshop on rendering - Bristol - 1992

[Mcla 70] **K. McLaren**
The Adams-Nickerson colour difference formula
J. Soc. Dyers Colourists. 1970
vol 86 - p. 354-366

[Mend et al 87] **E.R. Mendez, K.A. O'Donnell**
Experimental study of scattering from characterized random surfaces
Journal of the optical society of America. 1987
vol. 4 - num. 7 - p. 1194-1205

[Merc 85] **B. Mercier-Miège**
Les aspects actuels de la photomètrie physique
Lux
n° 135 - Oct-Dec 1985

[Meye et al 86] **G.W. Meyer, H.E. Rushmeier, M.F. Cohen, D.P. Greenberg, K.E. Torrance**
An experimental evaluation of computer graphics imagery
ACM Transaction on graphics - 1986
vol. 5 - num. 1 - p. 30-50

[Mitc 87] **D.P. Mitchell**
Generating antialiased images at low sampling densities
Computer graphics - 1987
vol. 21 - num. 4 - p. 65-72

[Mitc et al 88] **D.P. Mitchell, A.N. Netravali**
Reconstruction filters in computer graphics
Computer graphics - 1988
vol. 22 - num. 4 - p. 221-228

[Morl et al 75] **D.I. Morley R. Munn, F.W. Billmeyer,Jr**
Small and moderate colour differences: II, The Morley data
J. Soc. Dyers Colourists. 1975
vol 91 - p. 229-242

[Mucc 95] **L. Mucchielli**
La Gestalt - L'apport de la psychologie de la forme
Sciences humaines - Avril 1995
num. 49 - p. 26-27

[Nico et al 77] **F.E. Nicodemus, J.C. Richmond, J.J. Hsia**
Geometrical considerations and nomenclature for reflectance
National bureau of standards - 1977

[Niet 82] **M. Nieto-Vesperinas**
Radiometry of rough surfaces
Optica acta - 1982
vol. 29 - num. 7 - p. 961-971

[Pain et al 89] **J. Painter, K. Sloan**
Antialiased ray tracing by adaptative progressive refinement
Communication of the ACM - 1989
vol. 19 - num. 3 - p. 61-65

[Phon 75] **P. Bui Tuong**
Illunmination for computer genrated pictures
Communication of the ACM - 1975
vol. 18 - num. 6 - p. 311-317

[Port et al 61] **J.O. Porteus, H.E. Bennett**
Relation between surface roughness and specular reflectance at normal incidence
Journal of the optical society of America. 1961
vol. 51 - p. 123-129

[Rama 88] **V.S. Ramachandran**
Perceiving shape from shading
Scientific American - 1988
Août 1988 - p. 76-83

[Rich et al 92] **D.C. Rich, D.L. Alston, L.H. Allen**
Psychophysical verification of the accuracy of color and color-difference simulations of surface samples on a CRT display
Color research and application - 1992
vol. 17 - num. 1 - p. 45-56

[Robe 77] **A.R. Robertson**
The CIE 1976 color-difference formulae
Color Res. and Appl. - 1977
vol 2 - num 1 - p. 7-11

[Rubi 81] **R.Y. Rubinstein**
Simulation and the Monte Carlo method
J.Wiley et Sons - 1981

[Rush et al 95] **H. Rushmeier, G. Ward, C. Piatko, P. Sanders, B. Rust**
Comparing real and synthetic images: some ideas about metrics
6th Eurographics workshop on rendering - Dublin - 1995
p. 213-222

[Saul et al 46] **J.L. Saulner, B.I. Milner**
Modified chromatic value color space
J. Opt. Soc. Amer. 1946
vol 36 - p. 36-42

[Schl 92] **C. Schlick**
Divers éléments pour une synthèse d'images réalistes
Thèse de l'université de Bordeaux I num.861 - 1992

[Schl 94] **C. Schlick**
An importance driven Monte-Carlo solution to the global illumination problem
5th Eurographics workshop on rendering - 1994
p. 173-183

[Scof 43] **F. Scofield**
A method for determination of color differences
Nat. Paint, Varnish, Lacquer Assoc, Sci. Circ - 1943
vol 664

[Shan 49] **C.E. Shannon**
Communication in the presence of noise
Proc. IRE - 1949
vol. 37 - p. 10-21

[Shaw 66] **L. Shaw**
Comments on «Shadowing of random surfaces»
IEEE Trans. antennas and propagation - 1966
vol. AP-14, p. 253

[Shep 92] **R.N Shepard**
L'oeil qui pense
Éditions du Seuil - 1992

[Shir 90] **P. Shirley**
Physically based lighting calculations for computer graphics
PhD Thesis - University of Illinois - Urbana Champaign - 1990

[Shir et al 96] **P. Shirley, C. Wang, K. Zimmerman**
Monte Carlo techniques for direct lighting calculations
ACM transactions on graphics - 1996
vol. 15 - num 1- p. 1-36

[Shul 53] **L.G. Shultz, F.R. Tangherlini**
Optical constants of silver, Gold, Copper and Aluminium II The index of refraction
Journal of the optical society of America. 1953
vol 44 - p. 362-368

[Sieg 85] **D. Siegmund**
Sequential analysis
Springer-Verlag - 1985

[Smit 67] **B.G. Smith**
Geometrical shadowing of a random rough surface
IEEE transactions on antennas and propagation - 1967
vol. AP-15 - num 5 - p. 668-671

[Stil et al 82] **W.S. Stiles, G. Wyszecki**
Color Science: Concept and Methods, Quantitative Data and Formulae
John Wiley et Sons. 1982

[Stog 67] **A. Stogryn**
Electromagnetic scattering from rough, finitely conducting surfaces
Radio Science - 1967
vol. 2 - num 4 - p. 415-428

[Stra 61] **J.A. Stratton**
Théorie de l'électromagnétisme
Dunod. 1961

[Torr et al 67] **E.M. Sparrow, K.E. Torrance**
Theory for off-specular reflection from roughened surfaces
Journal of the optical society of America. 1967
vol. 57 - num 9 - p. 1105-1114

[Trow et al 75] **K.P. Reitz, T.S. Trowbridge**
Average irregularity representation of a rough surface for ray reflection
Journal of the optical society of America. 1975
vol. 65- num 5- p. 531-536

[Veac et al 94] **E. Veach, L. Guibas**
Bidirectional estimators for light transport
5th Eurographics workshop on rendering - 1994
p. 147-162

[Veac et al 95] **E. Veach, L. Guibas**
Optimally combining sampling techniques for Monte-Carlo rendering
Computer graphics - 1995
p. 419-428

[Wald 48] **A. Wald**
Sequential analysis
Wiley et Sons, New York - 1948

[Ward 92] **G.J. Ward**
Measuring and modeling anisotropic reflection
Computer Graphics - 1992
vol. 26 - num 2 - p. 265-272

[Whit 80] **T. Whitted**
An improved illumination model for shaded display
Communication of the ACM - 1980
vol. 23 - num. 6 - p. 343-349

[Wrig 91] **T. Wright**
Exact confidence bounds when sampling from small finite universes
Springer-Verlag - 1991

[Yell 83] **J.I. Yellott Jr.**
Spectral consequences of photoreceptor sampling in the rhesus retina
Science - 1983
vol. 221 - p. 382-385

[Zeki 92] **S. Zeki**
Les images visuelles
Pour la science - 1992
Novembre 1992 - p. 60-68

Crédits photographiques

	page	
Fig. 1	9	Réflexion et réfraction d'une onde sur une surface plane
Fig. 2	13	Vue microscopique d'une surface de verre poli [Ben et al 89]
Fig. 3	13	Vue microscopique d'une surface de molybdène poli [Ben et al 89]
Fig. 4	14	Vue microscopique d'une surface d'aluminium [Ben et al 89]
Fig. 5	14	Interférences sur surface rugueuse
Fig. 6	15	Profils de surface et distributions de hauteurs correspondantes [Ben et al 89]
Fig. 7	16	Profil et distribution des hauteurs d'une surface de silicone [Ben et al 89]
Fig. 8	16	Distribution des hauteurs d'une surface de cuivre polie au diamant [Ben et al 89]
Fig. 9	20	Système d'axes pour l'expression du champ réfléchi
Fig. 10	20	Observation à grande distance du champs électrique réfléchi
Fig. 11	30	Mesures approximatives d'un profilomètre optique [Ben et al 89]
Fig. 12	31	Énergie réfléchie par une surface gaussienne - Incidence proche normale: bonne adéquation théorique [Niet 82]
Fig. 13	31	Énergie réfléchie par une surface gaussienne - Incidence presque rasante: mauvaise adéquation théorique [Niet 82]
Fig. 14	32	Énergie réfléchie par une surface gaussienne - Incidence presque normale: cas inexplicable par la théorie [Niet 82]
Fig. 15	33	Intensité réfléchie par une surface d'acier à rugosité forte. Discordance en incidence rasante [Mar et al 90]
Fig. 16	33	Intensité réfléchie par une surface d'acier à rugosité faible. Discordance en incidence rasante [Mar et al 90]
Fig. 17	38	Distribution de la luminance réfléchie par un morceau d'aluminium: absence du comportement diffus+spéculaire [Chri 79]
Fig. 18	40	Taille de champs de vision de 2˚ et 10˚
Fig. 19	40	Éfficacité relative spectrale photopique
Fig. 20	41	Éfficacité relative spectrale scotopique
Fig. 21	42	Échec de la loi d'additivité de l'égalisation en luminance visuelle [Guth 70]
Fig. 22	45	Observateur standard CIE 1931
Fig. 23	58	fdrb mesurée et calculée par He et al. pour de l'aluminium [He et al 91]
Fig. 24	58	Curieuse image de He et al. [He et al 91]
Fig. 25	60	Mesures et calculs de la fdrb par Ward [Ward 92]
Fig. 26	72	Mesures et calculs de radiosité par Meyer et al [Mey et al 86]
Fig. 27	73	Saint Marc - début IX^{ème} siècle
Fig. 28	73	Bar des folies bergères - Manet
Fig. 29	73	L'étrange machine - Shepard [Shep 92]
Fig. 30	73	Bande dessinée - Chienne de vie - M-A. Prado - 88
Fig. 31	74	Tables - Shepard [Shep 92]
Fig. 32	74	Tunnel - Shepard [Shep 92]

	page	
Fig. 33	74	Triangle de Kanizsa [Zeki 92]
Fig. 34	74	Soleil - Shepard [Shep 92]
Fig. 35	75	Une coquille (d'oeuf d'autruche!) dans Libé [Mucc 95]
Fig. 36	76	Règles de l'organisation perceptive selon les psychologues de la forme [Mucc 95]
Fig. 37	77	Groupement perceptuel permettant la distinction forme/fond [Lec 95]
Fig. 38	77	Groupement perceptuel selon l'ombrage [Lec 95]
Fig. 39	78	Groupement perceptuel illustrant l'hypothèse de la source unique [Lec 95]
Fig. 40	78	Perception équivoque du relief. M.C. Escher
Fig. 41	79	Interprétation variable selon la culture de l'observateur [Lec 95]
Fig. 42	89	Stratégie de la poire
Fig. 43	90	Échantillonnage équivalent d'un cône droit et d'un cône penché
Fig. 44	99	Correspondances entre chemins échantillonnés et luminances élémentaires
Fig. 45	118	Structures de données pour test et échantillonnage
Fig. 46	121	Image test: Feu d'artifice (Kodak)
Fig. 47	121	Image test: Bande dessinée - Broussaille par Frank et Bom
Fig. 48	121	Image test: Image de synthèse par M.Cohen et al
Fig. 49	121	Image test: Mongolfières (Kodak)
Fig. 50	121	Image test: Torrent (Kodak)
Fig. 51	122	Image test: Flour power par V. Gurganian - Oakland Press
Fig. 52	122	Image test: Midway geyser basin - Yellowstone (Kodak)
Fig. 53	122	Image test: Wassily composition IV - Kandinski - 1911
Fig. 54	122	Image test: Le chemin de fer - Manet - 1872-3
Fig. 55	122	Image test: Image de synthèse par R.Fleig et P.Horne (DesignWorkshop + Radiance)
Fig. 56	122	Image test: Jardins sud-tunisiens - Klee - 1919
Fig. 57	123	Image test: Robe violette et anémones - Matisse - 1937
Fig. 58	123	Image test: Mont McKinley - B. Mitchell (Hometown Newspapers)
Fig. 59	123	Image test: Ours - D. Mead
Fig. 60	123	Image test: Roseaux (Kodak)
Fig. 61	123	Image test: Livia - F. Sommer
Fig. 62	123	Image de synthèse: Dessin animé - Pocahontas - Walt Disney Inc.
Fig. 63	124	Image test: Tzigane - J-L. Maillot - Zenica - 1994
Fig. 64	124	Image test: La chambre de Vincent - Van Gogh - 1888
Fig. 65	124	Image test: Image de synthèse (Radiance)
Fig. 66	124	Image test: Vitrail de Notre Dame de Paris (Kodak)
Fig. 67	127	Distribution inhomogène type
Fig. 68	129	Distribution mixte type
Fig. 69	130	Distribution intermédiaire type
Fig. 70	134	Marche aléatoire du tsrv dans la région critique
Fig. 71	135	Région critique pour une région de taille 2×2
Fig. 72	136	Région critique pour une région de taille 4×4 (1)

	page	
Fig. 73	136	Région critique pour une région de taille 4×4 (2)
Fig. 74	137	Région critique pour une région de taille 8×8
Fig. 75	140	Fréquence cumulée de β_r pour le recyclage (1)
Fig. 76	140	Fréquence cumulée de β_r pour le recyclage (2)
Fig. 77	142	Fréquence cumulée de β_r pour la moyenne approchée (1)
Fig. 78	142	Fréquence cumulée de α_r pour la moyenne approchée (1)
Fig. 79	143	Fréquence cumulée de β_r pour la moyenne approchée (2)
Fig. 80	143	Fréquence cumulée de α_r pour la moyenne approchée (2)
Fig. 81	144	Fréquence cumulée de β_r pour recyclage et moyenne approchée (1)
Fig. 82	144	Fréquence cumulée de α_r pour recyclage et moyenne approchée (1)
Fig. 83	145	Fréquence cumulée de β_r pour recyclage et moyenne approchée (2)
Fig. 84	145	Fréquence cumulée de α_r pour recyclage et moyenne approchée (2)
Fig. 85	150	Fréquence cumulée de β'_r pour le test express
Fig. 86	151	Fréquence cumulée de α'_r pour le test express
Fig. 87	152	Distances colorimétriques pour les images du panel calculées par le test express
Fig. 88	154	Définition des régions voisines et opposées
Fig. 89	173	Eléphant - Shepard [Shep 92]

Liste des tableaux

	page	
Tab. 1	8	Épaisseur de peau
Tab. 2	49	Corrélation entre jugements visuels et écarts colorimétriques calculés
Tab. 3	125	Distance colorimétrique en fonction de la taille de la zone de tolérance
Tab. 4	126	Distance colorimétrique en fonction de la proportion homogène
Tab. 5	138	Proportion de points à échantillonner avant les premières sorties possibles dans le tsrv exact
Tab. 6	141	Nombre d'échantillons nécessaire à une «précision» donnée pour un tirage hypergéométrique
Tab. 7	142	Nombre d'échantillons retenus, dans les tests, pour le calcul de la moyenne approchée
Tab. 8	147	Proportion de points à échantillonner avant les premières sorties possibles lorsque la proportion inhomogène varie
Tab. 9	148	Proportions inhomogènes retenues pour le test
Tab. 10	150	Nombre d'échantillons retenus pour le calcul de la moyenne approchée
Tab. 11	155	Description des scènes de test
Tab. 12	155	Nombre d'échantillons retenus pour le calcul de la moyenne approchée dans le tracé de rayons progressif
Tab. 13	156	Résultats du tracé de rayons progressif sur les scènes test

Index

A

analyse séquentielle 115
approximation de Kirchhoff 21
auto-corrélation 25

C

candela 43
chaîne de Markov 84
Chaîne de Markov homogène 84
chromaticités 46
coefficient de dispersion 22
cohérence spatiale 153
conductivité 6, 22, 23, 25, 28
courbe spectrale d'efficacité relative 39

D

densité de probabilité 86
diffuseur idéal 37
diffuseur parfait 37
diffusion de points 111

E

échantillonnage hiérarchique 62
éclairement 35
égalisation couleur 44
égalisation visuelle 39
épaisseur de peau 8
équation de rendu 61, 83
estimateur de niveau 99
étendue 117
exitance 35

F

facteur de luminance 46
fonction de base 87
fonction de distribution de réflectance bidirectionnelle 36
fonction de distribution de transmittance bidirectionnelle 37
fonction de poids 86
fonction de répartition 86
Fonction de transfert 83
fréquence 7

G

Gestalthéorie 75

H

hemi-cube 71
homogénéité 117
hypothèse homogène 132
hypothèses composites 132

I

indice d'amortissement 11
indices de réfraction 10
inhomogénéité 117
intégrale de Helmholtz 19
intégration hiérarchique adaptative 62
intensité 35
intervalle incertain 147

L

lancé de flèchettes 111
loi d'Ohm 6
longueur d'onde 7
lumen 43
luminance 36
luminance visuelle 39
lux 43

M

Maxwell 5, 6, 9
mésopique 41
métamères 44
Monte Carlo 64

O

onde plane 6, 7, 9, 10, 20, 24

P

paradigme progressif 109
passe progressive 115
période 7
perméabilité 6
permittivité 6
photopique 40
photoréalisme 71
physicoréalisme 71
probabilité de transition 84
probabilité homogène 117
proportion homogène 125
psychologie cognitive 76
psycho-physiologie 75

pulsation 7

R

radiosités 61
rapport de vraisemblance 116
réciprocité 38
réflectance 11
Réflectance à incidence normale 12
réflectivité 37
région critique 134
région homogène 117
régions opposées 154
relations de dispersion 12
risques réels 139
roulette russe 63
rugosité 15

S

scotopique 41
sensibilité 6
sensibilité visuelle 38, 39, 41, 45
spéculaire 37
spéculaire idéale 37
sphère d'homogénéité 117
stimuli primaires 44
stratification 62
surface lambertien 37
susceptibilité 6
synthèse réaliste 71
système trichromatique 46

T

temps d'arrêt du test séquentiel 134
test exact 133
théorème de Poynting 8
tracé de chemins 62
tracé de rayons 53
tracé de rayons distribués 61
transmittance 10
transmittivité 37

V

vecteur de Poynting 6, 7
vraisemblance 116

Z

zone de tolérance 117